Mitochondrial and repetitive DNA defining the sheep genome landscape

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

By

Sarbast Ihsan Mustafa

(BSc, MSc)

2018

Department of Genetics & Genome Biology

Abstract

Mitochondrial and repetitive DNA defining the sheep genome landscape

Sarbast Ihsan Mustafa

Repetitive DNA sequences (tandemly repeated and dispersed elements) vary in abundance, composition and organization between individuals, breeds, and related species. Here, we aimed to define the repetitive DNA landscape in sheep (*Ovis aries*). Whole genome sequence reads (38Gbp; 10x genome coverage) from five Kurdistani sheep individuals were investigated by graph-based read clustering (RepeatExplorer), frequency analysis of short motifs (*k*-mers), alignment to reference genome assemblies, *de novo* assembly and fluorescent *in situ* hybridization. To show genes in the sequences, the scrapie locus was identified and found to be associated with intermediate susceptibility. Mitochondrial genomes of breeds Hamdani and Karadi were assembled and grouped with known sheep haplogroups. Notably, abundant nuclear mitochondrial DNA segments (*numts*) were found at centromeres of chromosomes, and included mitochondrial sequences from ancestral species. The tandemly repeated DNA satellite I sequence represented 6% of the genome and satellite II was 2%. Meiotic analysis showed a loose chromatin loop organization of satellite I, while satellite II sequences were tightly organized and attached to the synaptonemal complex along with telomere repeats. Novel species-specific tandem sequences (1% of the genome) were also found. Non-LTR retrotransposons including LINEs and derived SINEs represented more than 20% of the genome, while DNA transposons comprise a lower proportion (<0.05%). Complete genomes of endogenous beta-retroviruses (enJSRV) plus three classes of endogenous retroviruses (ERVs) were identified. In total, repetitive sequences covered 30% of the genome, with tandemly repeated sequences at centromeres, and non-LTR retroelements families showing a centromeric to dispersed distribution with some being amplified on sex or submetacentric chromosomes. ERVs showed centromeric to dispersed distribution. Our results provide informative DNA markers within Bovidae lineages. Rapidly evolving repetitive sequences allow us to study processes of chromosome or genome evolution, homogenization or diversification in sheep, and more broadly across the Bovidae.

Declaration

I hereby declare that no part of this thesis has been previously submitted to this or any other university as part of the requirements for a higher degree. The work presented in this thesis, unless otherwise acknowledged in the text or by reference, was conducted by the undersigned who is fully responsible.
The work was conducted in the Department of Genetics and Genome Biology, University of Leicester, during the period from September 2013 to September 2017.

Signed: Date:

Dedication

This thesis is dedicated to:

My beloved, brilliant and supportive wife ‘’**Paiman**’’ who never got tired of encouraging me and standing by me.

My lovely little guys, my sons ‘’ **Hudd** ‘’ and ‘’ **Haboon**‘’.

My mother and father; the symbol of love and giving.

My beloved **brothers (Nabas, Osama, Bassam, Zaid) and sisters (Sahira, Kavin, Zaman and Aveen).**

My lovely English friend ‘’ **Liane** ‘’ for being such a good cheerleader and GREAT tennis player.

All of the people in my life who have touched my heart.

Acknowledgements

First and foremost, thanks to ALLAH for giving me the opportunity, knowledge and courage to overcome challenges and achieve this work. There is no doubt about it that, without his blessings, this accomplishment would not have been possible.

I would like to thank and express my sincere appreciation to my supervisor **Prof. Pat Heslop-Harrison**, for his continuous support, guidance and encouragement. In addition to being an outstanding supervisor, he is a man of principles and has professional knowledge after conducting tremendous research. I appreciate all of his contributions of time, support and ideas. I would like to thank **Dr. Trude Schwarzacher** for her valuable suggestions, discussions and constructive comments to improve this work.

I would like to thank my sponsorship (HCDP; Human Capacity Development Program) from the **Kurdistan Regional Government** for funding my academic study and living expenses. I would like to thank the Peshmerga (military Kurdish forces) for defending and protecting our country and our families whilst I was studying PhD in the UK.

I would like to thank my colleagues from Department of Animal Production/University of Duhok for their help in collecting blood from local sheep breeds.

I would like to thank **Prof. Jaladet Mohammed Saleh** and Dilan Jasim Khalil who allowed me use their laboratory facilities in the Research Centre at the University of Duhok for DNA extraction. I also wish to thank Joseph Morris Butchers Ltd, Lutterworth, Leicestershire for providing sheep blood samples.

I owe everything to my wife and my family, so I wish to thank them for all their love and support to stand with me in this long journey.

I also wish to thank my kind friends in Lab 201.

Last, and by no means least, I want to thank the **Department of Genetics and Genome Biology** at the University of Leicester for their privileged research facilities and services. Many thanks to the technical and administration team, Ramesh and Heather, for their support and making my work easier.

Publications

Mustafa S., Schwarzacher T. & Heslop-Harrison J. (2016) **The repetitive DNA landscape in sheep**. In: *Chromosome Research*, pp. S39-S. SPRINGER VAN GODEWIJCKSTRAAT 30, 3311 GZ DORDRECHT, NETHERLANDS. 22nd International Colloquium on Animal Cytogenetics and Genomics. 2–5 July 2016, Toulouse – France. Communication. Abstract. Talk. https://doi.org/10.1007/s10577-016-9532-x.

Mustafa S.I., Schwarzacher T. & Heslop-Harrison J. (2018) **Complete mitogenomes from Kurdistani sheep: abundant centromeric nuclear copies representing diverse ancestors**. *Mitochondrial DNA Part A*, 1-14. A full version of the article can be found at https://doi.org/10.1080/24701394.2018.1431226.

Table of contents:

List of Abbreviations

Abbreviation	Meaning
%	Percentage
BAC	Bacterial artificial chromosomes
BCIP	5-bromo-4-chloro-3-indolyl-phosphate
bp	Base pairs
BSA	Bovine Serum Albumin
CL	Cluster
cm	Centimetres
CsCl	Cesium Chloride
DAPI	4',6-diamidino-2-phenylindole
Dig	Digoxigenin
DIRs	Dictyostelium intermediate repeat sequence
DNA	Deoxyribonucleic acid
dNTPs	Deoxy nucleotide triphosphates
EBI	European Bioinformatics Institute
EDTA	Ethylene diamine tetra-acetic acid
enJSRV	Endogenous beta-retroviruses
ERVs	Endogenous Retroviruses
FISH	Fluorescent *in situ* hybridization
FITC	Fluorescein Isothiocyanate
gm	Gram
Gbp	Giga base pairs
HCl	Hydrochloric acid
HPG	Haplogroup
hrs	Hour(s)
INT	2-(4-iodophenyl)-5-(4-nitrophenyl)-3-phenyltetrazolium chloride
ISGC	International Sheep Genomics Consortium
kbp	Kilo base pair
LINEs	Long Interspersed Nuclear Elements
Low Complexity	LC
LTR	Long terminal repeat
M	Molar
mg	Milligram(s)
MITE	Miniature inverted-repeat transposable elements
ml	Millilitre(s)
MYA	Million years ago
NCBI	National Center for Biotechnology Information
ng/ul	Nanogram/Microlitre
NGS	Next Generation Sequencing
o.	*Ovis*
^{o}C	Degrees celsius
ORFs	Open reading frames

PBS	Phosphate buffered saline
PCR	Polymerase Chain Reaction
rDNA	Ribosomal DNA
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic acid
RNase	Ribonuclease
rpm	Rotations per minute
RT	Room temperature
RTE	Retro transposable element
Sat	Satellite DNA
SC	Synaptonemal complex
Simple Repeat	SR
SINEs	Short Interspersed Nuclear Elements
SNP	Single Nucleotide Polymorphism
TAREAN	TAndem REpeat Analyzer
TE	Transposable element
TIR	Terminal inverted repeat
TPRT	Target primed reverse transcription
TSD	Target site duplication
U	Unit
UV	Ultra violet
v/v	Volume per volume
w/v	Weight per volume
WGS	Whole Genome Shotgun
µg	Microgram
µg/ml	Microgram per milliliter
µl	Microlitre
µM	Micro molar

Chapter 1 Introduction

1.1 Bovidae family

Bovidae is a mammalian family belonging to the order of Artiodactyla (even-toed ungulates), suborder of Ruminantia and comprises of more than 149 species distributed in 49 genera and seven subfamilies Hippotraginae, Alcelaphinae, Cephalophinae, Reduncinae, Antilopinae, Bovinae and Caprinae (Nowak 1999; Matthee & Davis 2001). In addition, Bibi (2013) reconstructed the phylogeny tree of the Bovidae family by analyzing the mitochondrial genome of more than 125 ruminants supported by the usage of 16 fossil calibration points. His results distributed the Bovidae family to nine major tribes including Hippotragini, Alcelaphini, Cephalophini, Reduncini, Antilopini, Tragelaphini, Boselaphini, Bovini and Caprini.

The Bovini tribe is part of the Bovinae subfamily and includes domestic cattle *Bos taurus*. While, the domestic sheep *Ovis aries* and goats *Capra hircus* are the main species of the Caprini tribe under the Caprinae subfamily (Bibi 2013). Domestic animals within this family include sheep, goat, cattle and water buffalo are important species that contribute significantly to the economic and food security in the world. In addition, species in this family are key targets to preserving the biodiversity. However, some species under this family are not protected and are facing extinction due to loss of habitat, smuggling and illegal hunting (Chaves *et al.* 2000a; Cai *et al.* 2011; Kopecna *et al.* 2012) see (http://www.iucnredlist.org).

The taxonomic distribution and classification of the Bovidae family is one of extreme controversy with species in different geographic regions and with diverse phenotypes and genotypes. Furthermore, the evolutionary and phylogenetic relationships among the species of this family requires more investigation to clarify the evolutionary events (Nowak 1999; Kopecna *et al.* 2012). Fossil records indicated that the subfamily Bovinae, including Bovini and Tragelaphini tribes are the oldest fossils of bovid of roughly 23 million years old and were discovered in France and sub-Saharan Africa (Vrba 1985).

Although morphological and cytogenetic data, allozymes, fossil evidence, serum immunology, and DNA sequencing analysis support the monophyly of most subfamilies among the Bovidae family, some areas of evolutionary and phylogenetic relationships remain unclear (Matthee & Davis 2001). Therefore, to increase the resolution of the phylogeny of the Bovidae family, more molecular data including ribosomal mitochondrial DNA sequences (12S and 16S rRNA genes) are required (Allard *et al.* 1992; Gatesy *et al.* 1992). For instance, data obtained from investigating the distribution of repetitive DNA sequences including satellite DNAs within Bovidae tribes, provide a more comprehensive picture of evolutionary events (Jobse *et al.* 1995; Modi *et al.* 1996; Chaves *et al.* 2000b; Chaves *et al.* 2005). Therefore, the presence of whole genome sequencing could provide remarkable information for better understanding the phylogenetic relationships between intra and inter species within the Bovidae family.

1.2 Species of the genus *Ovis* and their relations to the origin of domestic sheep

The discovery of the exact wild ancestor of the modern domestic sheep remains one of the most controversial issues, despite several attempts in genetic studies that have been carried out (Meadows *et al.* 2007; Meadows *et al.* 2011). The genus *Ovis* is one of the more complicated mammalian categories in terms of its systematics and evolution. It includes several species and subspecies from which one or some have been proposed to be the origin of the present day domestic sheep *Ovis aries*. One of the suggestions is an Asiatic origin of the genus *Ovis* which was migrated through the route of North-Eastern Asia and the Bering strait to North America; and also, diversified in Eurasia less than 300,000 decades ago which caused the genus *Ovis* to be passed through such evolutionary processes resulting in sequential speciation events happening along different routes of migration spreading from the ancestral zone (Rezaei *et al.* 2010; Lv *et al.* 2015). According to the recent nomenclature of taxonomic species, the genus *Ovis* has been classified into seven species (Festa-Bianchet 2000), from which the most likely ancestor of all domestic sheep is considered to be Asiatic mouflon *Ovis orientalis* (Hiendleder *et al.* 1998b; Hiendleder *et al.* 2002; Bruford & Townsend 2006; Rezaei *et al.* 2010). During the last two centuries, several classifications and revisions based on

geographical distribution and morphological criteria have been projected (Hiendleder *et al.* 2002). Although all wild sheep were suggested as polymorphic populations of a single species by Haltenorth (1963), they have since been recognized as belonging to seven species (Nadler *et al.* 1973). Wild sheep are also characterized by various phenotypic traits such as colour, body size, pattern of the coat and horn morphology (Fedosenko & Blank 2005). Furthermore, they vary in their geographical distribution (Rezaei *et al.* 2010) and also in their genetic makeup, as different diploid numbers of chromosome were found within several species of wild sheep (Nadler *et al.* 1973; Bunch *et al.* 2005).

The Argali (*Ovis ammon* 2n=56) exist in mountainous regions of central Asia. The European mouflon (*Ovis musimon* 2n=54) and the Asiatic mouflon (*Ovis orientalis* 2n=54) are established in the west of Asia and Europe. The Bighorn (*Ovis canadensis* 2n = 54) are found in the Rocky Mountains, covering from Canada to south wards Colorado and Mexico. The Snow Sheep (*Ovis nivicola* 2n=52) is commonly found in the North East of Asia; the Urial (*Ovis vignei* 2n=58) are widely settled in Asia Minor. The thin horn or Dall sheep (*Ovis dalli* 2n=54) live in the mountainous areas of Western USA and Canada see (Rezaei *et al.* 2010). Furthermore, different taxonomic categories of genus *Ovis* with overlapping geographical distributions could display intermediate chromosome numbers. For instance, this happens when *Ovis orientalis* (2n=54) and *Ovis vignei* hybrids (2n=58) in the central region of Iran hybridized and produced such a fertile subspecies. Thus, presence of species or subspecies with intermediate chromosome numbers supports the existence of a single 'moufloniform' species (*Ovis orientalis*) comprised of the Asiatic mouflon, the Urial and their hybrids (Nadler *et al.* 1971; Valdez *et al.* 1978; Valdez & Batten 1982). Moreover, (Shackleton & Lovari 1997); (Shackleton 1997) and the recent research have adopted the classification of both (Nadler *et al.* 1973; Valdez & Batten 1982) as one of the main reference.

Similarly, the phylogeography and classification of the wild *Ovis* species are well-documented and related to the Fertile Crescent (domestication centre) (Rezaei *et al.* 2010). Wild sheep are branched into three classes including Argaliforms (*Ovis ammon*), in the central Asian highlands; Pachyceriforms (*Ovis dalli*, *Ovis canadensis*, in Northern America; *Ovis nivicola* in North Russia; and Moufloniforms (*Ovis vignei*, in Aralo-Caspin

basin; *Ovis orientalis*, in Iran, Armenia, Turkey; *Ovis* musimon, Europe). Depending on geographical overlap and cytogenetics knowledge, *Ovis* orientalis from the Moufloniforms class have been suggested as the progenitor of the present day domestic sheep *Ovis aries*. Separating the genetic input of other *Ovis* species into the domestic sheep has been sophisticated, nonetheless, odd numbered karyotypes might be produced from known fertile cross-species hybridizations (Bunch *et al.* 2005). Investigation of maternal lineages of wild and domestic sheep species could reveal the origin and relationship between these species. The advent of next generation sequencing NGS data enable assembly of the complete mitochondrial genomes which could provide better understanding about the relationship between species of the genus *Ovis*. Furthermore, it could also provide insight into the ancestral sequences of mitochondrial genomes.

1.3 Characterization and distribution of Iraqi sheep breeds

The Kurdistan Region in the north of Iraq corresponds to the zone of the initial domestication of sheep which includes many native sheep breeds such as Hamdani, Karadi, and Awassi see section 3.1 and Appendix 3.1 (Zeder 2008). Sheep are farmed mainly for carpet-wool production, but the distinctive and morphologically well-defined fat-tailed breeds are also raised for meat and milk. The landraces are well adapted to poor grazing habitats, showing hardiness in diverse harsh environments. Iraqi sheep breeds have not been well characterized under farming conditions (Alkass & Juma 2005). The Iraqi sheep breeds are remarkable animal producers due to their ability to survive and reproduce under a variety of environmental challenges of desert and semi-desert life with adequate rainfall in spite of their slow growth, low production and fertility (Al-Rawi *et al.* 1996). The major sheep breeds are distributed in various geographical regions over the country. Both the Karadi and Hamdani breeds are mainly distributed and dominant in the Iraqi Kurdistan region while, the Awassi breed is distributed more in the middle and west of Iraq (Alkass & Juma 2005). Although the genetic diversity of some Kurdistani sheep breeds have been studied using microsatellite markers (Mohammed 2009; Al-Barzinji *et al.* 2011; Al-Barzinj & Ali 2013), their diversity in terms of maternal lineages is unknown. Thus, next generation sequences will allow assembly of their

complete mitochondrial genomes which could offer good evidence about origin and divergence time of fat tailed sheep breeds.

1.3.1 **Karadi sheep**

Karadi sheep are a fat-tailed sheep breed farmed on the high plains and in the mountains of the Kurdistan region of Iraq, mainly in Erbil, Sulaimaniya, Duhok and Nineveh. They have a yellowish-white fleece with a black face. The black colour sometimes, extends to the shoulders or to other parts of their body. Some of them are variegated. They have pendulous ears, but they are shorter than those of the Hamdani breed. The head of the ewe is somehow long and straight, while that of the ram is slightly convex in profile. Both sexes are polled. The main character they have is a large and heavy fat-tail. The thin end of their tail emerging from the fat lobes nearly reaches the ground. Karadi sheep are considered to have the coarsest wool properties from among all the sheep breeds in the country and the breed is known for its good milking potential (Juma & Alkass 2000). The average fleece weight is about 1.6 kg. The weight of a male is about 50kg and a female is 42 kg (Karam *et al.* 1976). They produce, on average roughly 126kg of milk during the 169 days suckling period. The fat percentage of the milk is about 7.3% (Juma & Alkass 2000; Alkass & Akreyi 2016).

1.3.2 **Hamdani sheep**

Hamdani sheep are considered to be the largest Iraqi sheep breed in terms of their size (Karam *et al.* 1976) and are also suggested to be a strain of the Karadi sheep breed. The fleece colour is usually white. The body has a long black and fine-boned head. Occasionally, two wattles are found under the throat of ewe. The face of the ewe is slightly less convex than that of the ram. Both rams and ewes are hornless. The distinctive feature of the Hamdani breed is the very pendulous ears which extend beyond the length of its head. Furthermore, the end of the ears is characterized by an outward-curving point. Hamdani ewes are accepted to be good milk producers by having well-developed udders with long and large teats. They produce, on average roughly 96.3 kg of milk during 145 days, and the milk contains about 7.04% fat (Hammodat 1985; Maarof *et al.* 1986). The body weight of an adult female and an adult male are on

average 65 kg and 80 kg respectively (Karam *et al.* 1976). The wool and fleeces are usually coarse and heavy (3 kg) and often have coloured fibres (Al-Azzawi 1977; Maarof *et al.* 1982). The Hamdani breed are mainly distributed in Erbil, Duhok and Nineveh (Al-Mourrani *et al.* 1980).

1.3.3 Awassi sheep

Awassi sheep are the most abundant sheep breed in Iraq. The fleece is commonly white with red to brown faces. The breed is rarely variegated. The head is big, with a noticeable convexity in rams which could camouflage its length. Ewes are polled with very rare cases of short horns. While, rams have long spiral horns. The ears look semi-pendulous. The legs are generally covered with white hair and rarely with coloured spots. The fat tail is medium in size, round and short extending only to their hocks. The Awassi breed have two uncharacterized strains called the Na'aimi and the Shefli (Iñiguez 2005). The breed produce, on average yield of 126L of milk which contains about 7.5% fat (Alkass & Akreyi 2016). The average body weight of an adult female and an adult male is 55 kg and 75 kg respectively. The Awassi breeds produce carpet quality wool, which is commonly utilized in the carpet industry. Their fleeces weigh about 2 kg with an intermediate fineness which lies between the Karadi sheep (coarsest) and the Na'aimi sheep (finest). The Awassi breeds are distributed in various regions of the northern and the southern areas of the Jazira desert and the western semi-desert (Al-Mourrani *et al.* 1980).

1.4 Repetitive DNA sequences of eukaryotic genomes

The eukaryotic genomes consist of two main components, i.e. unique and repetitive DNA sequences. Britten and Kohne (1968) established that large amounts of DNA sequences of the eukaryotic genome are made up of repetitive DNA sequences. They classified repetitive DNA sequences into ‘highly’ or ‘moderately’ repeated sequences, based on the degree of their repetition. The term ‘repetitive sequences’ (e.g. repetitive DNA; repeats; repetitive elements; and DNA repeats) refers to the fact that DNA sequences are constituted of identical (or almost identical) nucleotides copied in large numbers throughout the genome (Charlesworth *et al.* 1994; Schmidt & Heslop-Harrison

1998; Heslop‐Harrison & Schwarzacher 2011). Repetitive DNA sequences are generally categorized into tandem repeats (satellites DNA, very highly and/or moderately repetitive, minisatellite and microsatellite sequences) and transposable elements (moderately repetitive, dispersed and/or mobile repeats), based upon their organization and structure throughout the eukaryotic genome (Charlesworth *et al.* 1994; Biscotti *et al.* 2015b) Figure1.1). The repetitive DNA landscapes of sheep genome have not been characterized. Thus, in this study, the NGS data alongside recent efficient bioinformatics tools and *in situ* hybridization will be utilized in order to explore and characterize the major dispersed and tandemly repeated DNA elements in sheep genome.

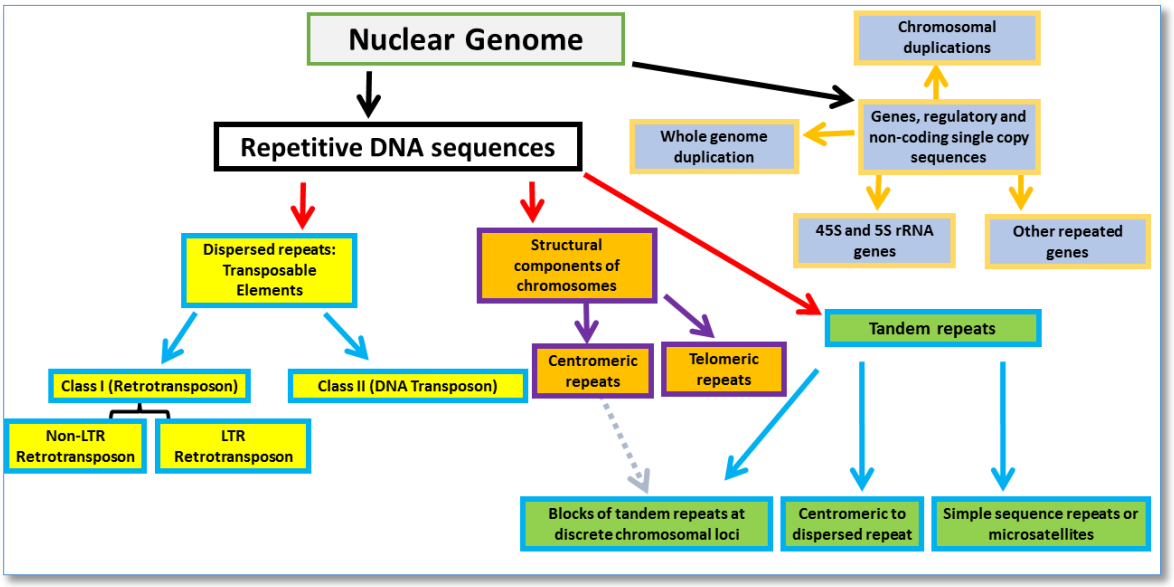


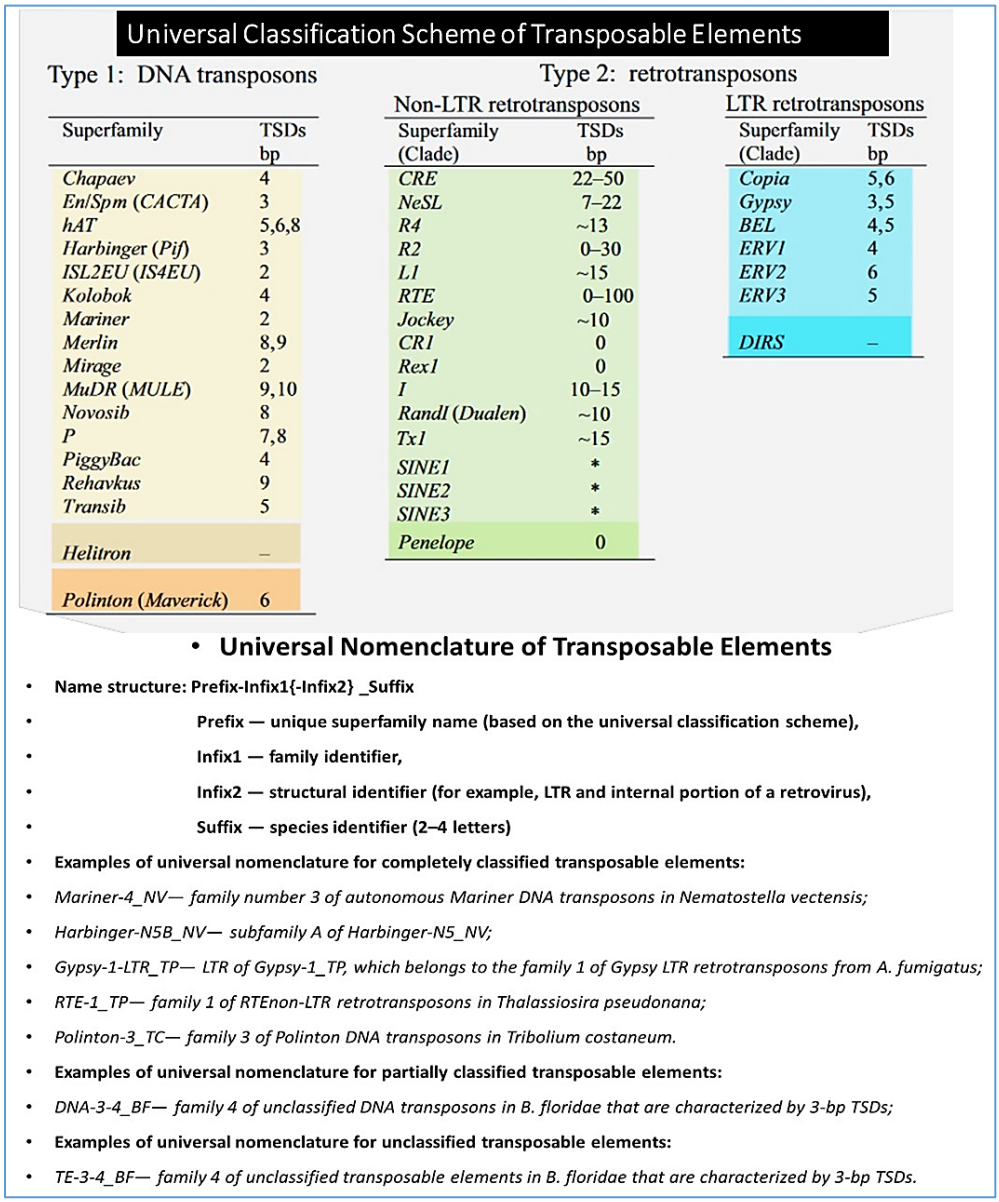
Figure 1.1 General overview shows the main divisions of repetitive DNA sequences in the eukaryotic genome. Taken from Biscotti *et al.* 2015.

1.5 Transposable Elements

Transposable Elements is a broad term which defines great diversity within mobile elements. They are highly ubiquitous DNA sequences which can transfer from one chromosomal location to another inside the same host genome. They are observed in most eukaryotic genomes (Jurka *et al.* 2007). Transposable elements are broadly split into two main classes [Class I elements retrotransposons and Class II elements DNA transposons], depending on whether their transposition or mobilization mechanism happens with the presence or absence of an RNA intermediate Figure 1.1. They include

further diverse subclasses, orders, and super-families (clades) Table 1.1 (Wicker *et al.* 2007).

The first classification scheme of transposable elements was suggested by Finnegan (1989). Kapitonov and Jurka (2008) implemented a universal classification and nomenclature of eukaryotic transposable elements in the Repbase databases Figure 1.2.



Universal Classification Scheme of Transposable Elements

Type 1: DNA transposons

Superfamily	TSDs bp
Chapaev	4
En/Spm (CACTA)	3
hAT	5,6,8
Harbinger (Pif)	3
ISL2EU (IS4EU)	2
Kolobok	4
Mariner	2
Merlin	8,9
Mirage	2
MuDR (MULE)	9,10
Novosib	8
P	7,8
PiggyBac	4
Rehavkus	9
Transib	5
Helitron	–
Polinton (Maverick)	6

Type 2: retrotransposons

Non-LTR retrotransposons

Superfamily (Clade)	TSDs bp
CRE	22–50
NeSL	7–22
R4	~13
R2	0–30
L1	~15
RTE	0–100
Jockey	~10
CR1	0
Rex1	0
I	10–15
RandI (Dualen)	~10
Tx1	~15
SINE1	*
SINE2	*
SINE3	*
Penelope	0

LTR retrotransposons

Superfamily (Clade)	TSDs bp
Copia	5,6
Gypsy	3,5
BEL	4,5
ERV1	4
ERV2	6
ERV3	5
DIRS	–

- **Universal Nomenclature of Transposable Elements**
- **Name structure: Prefix-Infix1{-Infix2} _Suffix**
- **Prefix — unique superfamily name (based on the universal classification scheme),**
- **Infix1 — family identifier,**
- **Infix2 — structural identifier (for example, LTR and internal portion of a retrovirus),**
- **Suffix — species identifier (2–4 letters)**
- **Examples of universal nomenclature for completely classified transposable elements:**
- *Mariner-4_NV— family number 3 of autonomous Mariner DNA transposons in Nematostella vectensis;*
- *Harbinger-N5B_NV— subfamily A of Harbinger-N5_NV;*
- *Gypsy-1-LTR_TP— LTR of Gypsy-1_TP, which belongs to the family 1 of Gypsy LTR retrotransposons from A. fumigatus;*
- *RTE-1_TP— family 1 of RTEnon-LTR retrotransposons in Thalassiosira pseudonana;*
- *Polinton-3_TC— family 3 of Polinton DNA transposons in Tribolium costaneum.*
- **Examples of universal nomenclature for partially classified transposable elements:**
- *DNA-3-4_BF— family 4 of unclassified DNA transposons in B. floridae that are characterized by 3-bp TSDs;*
- **Examples of universal nomenclature for unclassified transposable elements:**
- *TE-3-4_BF— family 4 of unclassified transposable elements in B. floridae that are characterized by 3-bp TSDs.*

Figure 1.2 shows the universal classification and nomenclature of eukaryotic transposable elements.

Different classes of transposable elements are differently coloured. Penelope and DIRS can be viewed as two additional classes of retrotransposons. An asterisk indicates that the lengths of the target-site duplications (TSDs) by short interspersed nuclear elements (SINEs) depend on non-LTR retrotransposons being involved in their transpositions. Taken from Kapitonov and Jurka (2008).

Inside this broader classification, transposable elements of class I and class II can be described as autonomous or non-autonomous, determined by whether they encode proteins required for their own transposition or they do not encode them. Autonomous elements are transposable elements which are capable of self-mobilization due to encoding the necessary proteins to perform transposition and moving from one chromosomal location to another. While non-autonomous elements are those not encoding the transposition machinery and thus depending on co-mobilization carried out by the enzymatic machinery of other autonomous transposable elements (Craig 2002; Jurka *et al.* 2007; Kapitonov & Jurka 2008; Richard *et al.* 2008; Piégu *et al.* 2015).

1.5.1 Class I elements retrotransposons or retroelements or retroposons

Class I element retrotransposons are transposed through an RNA intermediate also known as a mechanism of "copy and paste'' which includes reverse transcription and genomic integration. Following a recent classification (Wicker *et al.* 2007; Kapitonov & Jurka 2008), five types of eukaryotic Class I Elements retrotransposons can be categorized. (1) LTR retrotransposons (Long Terminal Repeat elements); (2) non-LTR retrotransposons or Long Interspersed Elements (LINEs); (3) DIRS-retrotransposons (Dictyostelium Intermediate Repeat Sequence); (4) (PLE) Penelope-Like retrotransposons; and (5) Short Interspersed Elements (SINEs). Within the class I elements both LTR and non-LTR retroelements are the most abundant and widespread transposable elements in eukaryotes. In animals, the LTR retrotransposons are less abundant. By contrast, in plants, the LTR retrotransposons are the most dominant type, comprising of 75-88% of the maize genome (Schnable *et al.* 2009; Jiang & Ramachandran 2013) and 55% of the sorghum genome (Paterson *et al.* 2008) which indicates a relationship between the quantity of LTR elements and the genome size.

It has been reported that the mouse genome has accumulated newer repetitive sequences than the human genome (Mouse Genome Sequencing 2002). Roughly, 46% of the human genome is composed of interspersed repeats. These probably resulted from the insertions of transposable elements which have been active in the last 150-200 million years. Because of the high degree of sequence divergence, it is not possible to identify fossils older than a certain age. If it were possible, the total fraction of the

human genome resulting from transposons might be significantly larger. It has been concluded that the lower activity of transposons since the divergence of both lineages human and mouse, caused the smaller size of the mouse genome. However, this is not always the case, as transpositions have been more active in the mouse lineage (Mouse Genome Sequencing 2002). Interspersed repeats could be subdivided into ancestral repeats and lineage-specific repeats. Ancestral repeats refer to repeats that exist in a common ancestor, while the lineage-specific repeats are repeats that were introduced by transposition after the divergence of human and mouse. Transposon-derived sequences have been increased in the mouse genome due to the transposon insertions since their divergence from humans. For example, approximately 32.4% of the mouse genome consists of lineage-specific repeats while in human genome they are about 24.4%. Furthermore, it is reported that the rate of transposition has been noticeably dropped in the human genome over the past 40 million years (Smit 1999; Lander *et al.* 2001).

Table 1.1 Proposed classification of eukaryotic transposable elements (Wicker *et al.* 2007).

Class	Order	Superfamily	Phylogenetic distribution
Class I Retrotransposons	LINEs	L1	Metazoans, Plants, Fungi
		RTE	Metazoans
		R2	Metazoans
		I	Metazoans, Plants, Fungi
		Jockey	Metazoans
	SINEs	5S	Metazoans
		7SL	Metazoans, Plants, Fungi
		tRNA	Metazoans, Plants, Fungi
	LTR	Copia	Metazoans, Plants, Fungi
		Bel-Pao	Metazoans
		Retrovirus	Metazoans
		ERV	Metazoans
	PLE	Penelope	Metazoans, Plants, Fungi
	DIRS	DIRS	Metazoans, Plants, Fungi
		Ngaro	Metazoans, Fungi
		VIPER	Trypanosomes
Class II DNA transposons *Subclass 1*	TIR	Tc1-Mariner	Metazoans, Plants, Fungi
		hAT	Metazoans, Plants, Fungi
		Mutator	Metazoans, Plants, Fungi
		Merlin	Metazoans
		Transib	Metazoans, Fungi
		P	Metazoans
		PiggyBac	Metazoans
		PIF-harbinger	Metazoans, Plants, Fungi
		CACTA	Metazoans, Plants, Fungi
	Crypton	Crypton	Fungi
Class II DNA transposons *Subclass 2*	Helitron	Helitron	Metazoans, Plants, Fungi
	Maverick	Maverick	Metazoans, Fungi

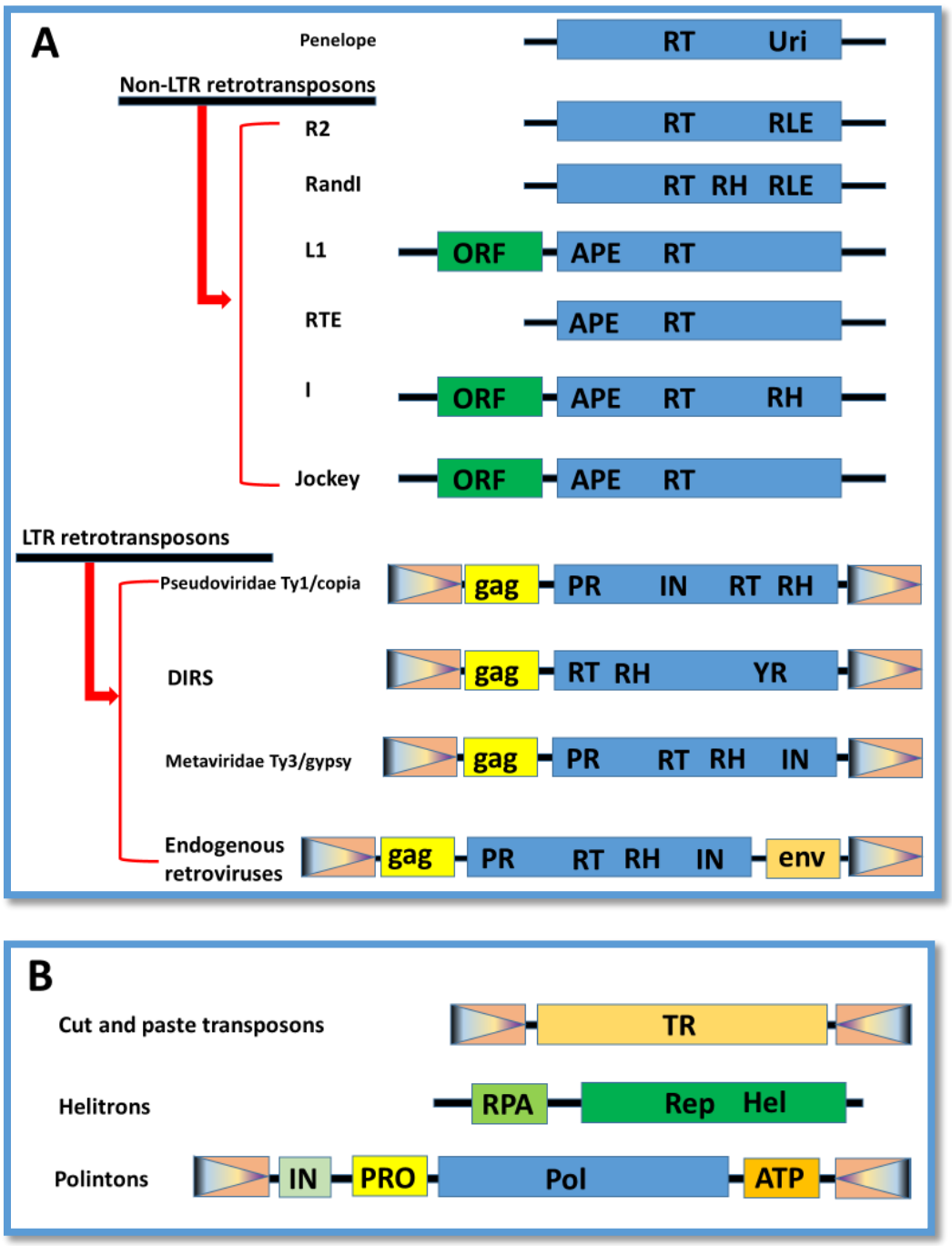


Figure 1.3 Schematic structure and classification of autonomous retrotransposons. Abbreviations used: APE; apurinic endonuclease; env; envelope gene; gag; gag gene; IN; integrase; ORF1; open-reading frame 1; PR; proteinase; RH; RNase H domain; RLE; restriction-like endonuclease; RT; reverse transcriptase; Uri; endonuclease domain with similarity for group I; introns; YR; tyrosine recombinase. The black lines indicate the non-protein coding regions of the retrotransposons. The boxes represent the open-reading frames and the boxed triangles represent the LTRs. (B) Schematic structure of autonomous class II transposons. The following abbreviations are used: ATP; ATPase; Hel; helicase; IN; integrase; Pol; polymerase; PRO; cysteine protease; Rep; replication initiation domain; RPA; replication protein A; TR; transposase. Boxed triangles represent the TIRS. Structures adapted from López-Flores and Garrido-Ramos (2012).

1.5.1.1 **Non-LTR retrotransposons**

1.5.1.1.1 **LINEs (Long Interspersed Nuclear Elements)**

LINE repeats are the most thoroughly studied transposable elements in the genomes of various organisms. For instance, 17% of human genomes are comprised of LINEs (Cordaux & Batzer 2009). Within the human LINE family, the LINE_L1 repeat (6-8kb) is the most abundant interspersed repeat which accounts for roughly 15%, and represents more than 500,000 copies of its genome (Lander *et al.* 2001).

LINEs are the most dominant repeats in most animals while they are less common in plants. The most abundant repeats within the transposable elements of birds and mammalian genomes are LINEs repeats, which account for approximately 50-60%. LINE-like elements called CR1 chicken repeat 1 are abundant in the chicken genome representing roughly 6.4% (Hillier *et al.* 2004). In mammalian genomes sequenced so far, the opossum genome contains about 29% LINEs while the dog genome harbours 18% LINEs repeats (Lindblad-Toh *et al.* 2005; Mikkelsen *et al.* 2007).

The LINEs can exhibit either a tandem repeat poly (A) tail or just an A-rich region. The LINEs contain several superfamilies L1, RTE, R2, Jockey and I Table 1.1. These superfamilies are classified into 28 various clades which can be attributed to one of the major types of non-LTR retrotransposons (Kapitonov *et al.* 2009). The LINEs superfamilies contain either one or two open reading frames. Enzymes encoded in the open reading frames of LINEs repeats contribute into the retrotransposition process. In general, the retrotransposon mRNA is produced from the transcription process which is then exported to the cytoplasm in order to translate into proteins. The translated proteins form complex be inserted to the RNA. This complex is then imported to the nucleus in order to insert in another genomic location. Firstly, the endonuclease encoded by the open reading frame 2 of the element produces a nick at target AT-rich regions on the bottom strand of the insertion site of the double stranded DNA. Then, the poly (A) tails of the 3' LINE RNA hybridizes to the poly (T) at the 5' nick. The 3' OH produced by the cleavage will then be used by the reverse transcriptase to synthesize the cDNA from the mRNA. This reaction is known as target-primed reverse transcription (TPRT) due to the occurring of the reverse transcription at the insertion sites. The

synthesis of the second strand of DNA involves making a nick a few nucleotides away from the first nick at the other target genomic DNA strand after degradation of the LINE RNA. This reaction will operate the 3' end as a primer for the synthesis of the second stranded DNA, which is then followed by ligation. Some LINEs are characterized by truncated 5’ ends due to the premature termination of reverse transcription, which is common in the mechanism of the target-primed reverse transcription. It has been suggested that elements belonging to the clades of L1, RTE, L2, & CR1 are mobilized by a mechanism like the TPRT reaction (Cost *et al.* 2002; Ichiyanagi & Okada 2008). More than 80% of all dispersed repeats in the chicken genome nearly 200,000 copies are non-LTR retrotransposons LINE like elements CR1 with a full-length 4.5 kb (Hillier *et al.* 2004). Similar to the mammalian L1 element, the chicken LINE, CR1 have a (G+C)-rich internal promoter region, followed by two open reading frames (Hillier *et al.* 2004).

In regard to their genomic distributions, the LINE elements prefer accumulation on sex chromosomes as in mouse and human genomes (Mouse Genome Sequencing Consortium 2002). For instance, the density of lineage-specific LINE (L1) copies is almost twice in X chromosome than in autosomes in mouse genome (28.5% on X in compare with 14.6% for the autosomes). Moreover, in humans, the sex chromosomes X and Y display more robust preference (18.0% on Y and 17.5% on X in compare with 7.5% for the autosomes). It can even be noticed that the L1 enrichment on the sex chromosomes is still extremely substantial after the commonly higher (A+T) content has been accounted for in the sex chromosomes (Mouse Genome Sequencing 2002)

In early hybridization experiments, high quantity of LINE L1 was observed on sex chromosomes (Korenberg & Rykowski 1988; Boyle *et al.* 1990). Thus, it has been proposed that LINE L1 copies could play an essential role in the inactivation of the X chromosome (Lyon 1998; Bailey *et al.* 2000). Probably, there are some reasons and explanations behind this privileged gathering of LINE L1 elements on the sex chromosomes. Due to the presence of poor gene regions in Y chromosome that could allow the LINE repeats for more insertions and accumulations. Regarding the X chromosome, the explanation behind enrichment of LINE L1 elements on the X chromosome is still not clear in human and mouse lineages.

1.5.1.1.2 **SINEs Short Interspersed Nucleotide Elements**

SINEs are non-autonomous retroposons that lack their own machinery for retrotranscription. Structurally, SINEs are short in length up to 0.5kbp. They mainly consist of two to three components. The first component is the head (5'-terminal) which contains the promoter (Pol III) by which could identify origin and classes of SINE repeats (tRNA, 7SL RNA and 5S rRNA). The second component is the body (internal region) which is very important due to its characteristics of having a LINE-related segment and thus it is variable in origin and family-specific. The last component is the tail 3'-terminal A-rich region. A-rich region is usually composed of degenerate repeats with different sequence lengths. Furthermore, SINEs can be found in more complex structures Figure 1.4 (Kramerov & Vassetzky 2011). SINEs encode no proteins and for their retrotransposition process, they borrow reverse transcriptase (RT) from a LINE-like element, which is capable of recognizing the sequences at the 3'of the SINE RNA. In other words, SINEs are derived from structural RNA genes, such as 7SL, and transfer RNA (tRNA) genes in the 5' promoter and transcribed by means of RNA polymerase III (Pol III) which is recognized by the LINE machinery in order to be mobilized (Sakamoto & Okada 1985; Okada 1991a, b; Kajikawa & Okada 2002; Hayashi *et al.* 2014).

Mammalian genomes contain considerable amounts of SINEs sequences. For instance, bovine genomes are comprised of three diverse families of SINEs. These including Bov-B, Bov-tA & Bov-A2 with genomic proportions 0.5%, 1.6% and 1.8% respectively. Bov-B is also called PstI repeat which consists of 560 bp in which the short region has similar sequences to the A element. Bov-tA consists of 115bp and 73bp representing a tRNA pseudogene united to an A element. Bov-A2 is a dimer of two monomer units each one is 115bp in addition to two short tandem repeats (Lenstra *et al.* 1993). tRNA derived SINEs are mostly present in vertebrates, invertebrates and plants. While, in primates, rodents and scadentians, 7SL RNA-derived SINEs are more dominant. However, SINEs originating from 5S RNAs were found in few mammals and some fishes (Deragon & Zhang 2006; Kriegs *et al.* 2007; Kramerov & Vassetzky 2011). Moreover, in humans, three diverse monophyletic families of SINEs including the most abundant and active Alu element, and the inactive Ther2/MIR3 and MIR were identified (Lander *et al.* 2001). For instance, Alu elements have been used as forensic tools, and their insertion and

polymorphisms have been used to examine the origin, population structure and demography of humans (Cordaux & Batzer 2009). Furthermore, Shimamura *et al.* (1997) identified two families and nine retropositional events of SINEs in the order of Cetacea and Artiodactyla where these SINEs were found more exceptionally in the genomes of ruminants, whales and hippopotamuses. Their data provides more evidence and phylogenetic resolution that ruminants, whales and hippopotamuses makeup a monophyletic cluster. Thereafter, Nijman *et al.* (2002) studied comparative sequencing of SINEs in ruminants numerous indels (deletions and insertions) were discovered. Thus, retrotransposition of SINEs may be used as one of the informative markers in studying and reconstructing phylogenetic relationships between ruminant species at different levels of classification. In other animals like bivalves, Nishihara *et al.* (2016) categorized eight novel superfamilies of SINEs that related to the V-SINEs, DeuSINEs, CORE-SINEs and the MetaSINEs. Such a structural information and broad distribution of SINEs could be useful for comparative analyses to explore why the SINE repeats have been reserved in metazoan genomes during their evolution.

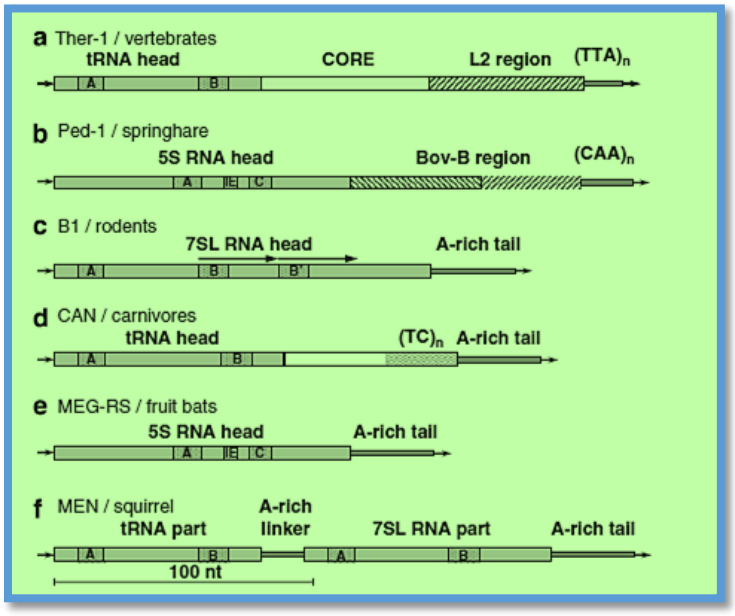


Figure 1.4 Shows examples of different SINE structures originated from various organisms. **a**. It is a tRNA head related CORE SINE of vertebrates represents the central specific sequence and comprises of LINE region and simple repeats (TTA)n. **b**. Another SINE structure from springhare contains 5S rRNA head, bipartite LINE related regions and multiple simple repeats (CAA)n. **c**. SINEs from rodents consist of 7SL RNA head and A-rich tail. **d**. SINEs from carnivores including tRNA head and fragments of multiple simple repeats (TC)n. **e**. SINE from fruit bats structured of 5S rRNA and A-rich region. **f**. SINEs from squirrel containing two parts of RNA; tRNA and 7SL RNA separated by A-rich linker. Taken from Kramerov and Vassetzky (2011).

1.5.1.1.3 **RTE repeats**

Like LINE_L1, LINE_RTE is one of the major clades of LINE repeats (Wicker *et al.* 2007; Ruggiero *et al.* 2017). RTE class is one of non-LTR retro transposable elements, which encodes an open reading frame comprising of two domains of reverse transcriptase (RT) and apurinic-apyrimidic endonuclease (APE). RTE were first identified in the genome of the *Caenorhabditis elegans* (Youngman *et al.* 1996). Malik and Eickbush (1998) indicated that the RTE repeat is the ancestor of many other SINEs and is commonly scattered in animal genomes. Several and different types of LINE_RTE repeat have been found in various genomes including mammals (bovine) and insects (mosquito) and others organisms (Ohshima & Okada 2005). Non-LTR LINE_RTE BovB repeat occupy substantial fractions of the total content of bovine interspersed repeats representing the clear majority of bovine specific repeats (Adelson *et al.* 2009). Bovine dimer-driven family (BDDF) was first referred to as a bovine LINE like retrotransposon Bov-B LINE which was related to the Alu-like sequences of bovine (Szemraj *et al.* 1995). This has been categorized as a member of the RTE-1 family (Malik & Eickbush 1998; Malik *et al.* 1999). Approximately, 50 to 270 thousand copy numbers of the 3' end of RTE-1 were estimated in the bovine genome (Lenstra *et al.* 1993; Kordiš & Gubenšek 1999). Gentles *et al.* (2007) identified more than four families of RTE in the genome of opossum *Monodelphis domestica*. In the genomes of ruminants and marsupials, Ohshima & Okada 2005 stated a symbiotic relationship between LINE RTEs and SINEs, in which RTE repeat encodes the machinery to transpose SINE repeats, including SINE BovA and SINE RTE. Moreover, Gentles *et al.* (2007) indicated that the several families of SINEs present in the genome of opossum Monodelphis domestica utilize RTE repeats for their mobilization. Furthermore, it has been believed that RTE LINEs are transmitted in a horizontal mode from reptiles to marsupials (Gentles *et al.* 2007) and to ruminants (Kordis & Gubensek 1998; Kordiš & Gubenšek 1999).

1.5.1.2 **LTR retrotransposons**

These elements belong to class I with length ranges from 100bp to 25kbp and are characterized by harbouring LTRS with a size of about few 100bp up to more than 5000bp (Wicker *et al.* 2007). The LTR retrotransposons can be subdivided into three major superfamilies including the Tyl/copia; the Bel-Pao; and the Ty3/gypsy. However,

Wicker *et al.* (2007) included two additional superfamilies known as the vertebrate retroviruses and the endogenous retroviruses ERVs.

The LTR retrotransposons comprise of the promoter sequences and transcription features. Furthermore, they contain many genes and each one encodes a different enzymatic domain. They contain a gag gene, which might participate in the reverse transcription process as it encodes a nucleic acid binding protein, and a pol gene that encodes other enzymatic domains: reverse transcriptase (RT), proteinase (PR), RNase H (RH), and integrase (INT).

The retrotransposition mechanism for the LTR retrotransposons begins by annealing a tRNA molecule to the primer-binding site at the 3' end of the retrotransposon RNA in order to prime the reverse transcription, which occurs in the cytoplasm. After the two DNA strands of the complementary DNA being produced, the cDNA is then moved to the nucleus in order to be integrated in the new genomic location. The percentage of LTR elements varies across different organisms.

1.5.1.2.1 **Endogenous retrovirus related repetitive elements**

Eukaryotic genomes contain considerable fragments of endogenous retroviruses (ERVs) (Kumar & Bennetzen 1999; Lander *et al.* 2001; Hillier *et al.* 2004; Mikkelsen *et al.* 2005). ERVs are considered as repetitive transposable elements and they are difficult to recognize due to the rapid and high mutations occurring in ERVs sequences (Sperber *et al.* 2007). Many algorithms have been established for identifying and searching sequences (Altschul *et al.* 1990), however, they are limited in the discovery of genomic related ERVs on a large-scale. In recent years, huge numbers of genomes of various organisms being sequenced. Thus, identification of retroviral sequences across a broad range requires an efficient way of detection, classification and annotation.

1.5.1.2.2 **DIRS elements**

These elements were first discovered in *Dictyostelium discoideum* and thereafter they were found in various species including animals, fungi and green algae. It has been proposed that these elements evolved from other type of LTR retrotransposon such as

a gypsy-like ancestral retrotransposon, although the recent LTR retrotransposons cannot be considered as being derivatives from DIRS elements. In contrast to the properties of LTR retrotransposons, DIRS elements contain a tyrosine recombinase domain (YR) and inverted LTRs and some of them have direct LTRs Figure 1.3. DIRS elements also have an internal complementary repeat due to the repetition of a piece of the LTR sequence along the element. These structures and properties give the DIRS elements a typical mechanism of integration, which is different from other integration modes that are present in the other types of retrotransposons (Jurka *et al.* 2007; Wicker *et al.* 2007; Eickbush & Jamburuthugoda 2008)

1.5.2 Class II element DNA transposons

DNA transposons are transposable elements that mobilize through a DNA-mediated transposition mode analogous to a mechanism called “cut and paste’’ as they cut their elements from one chromosomal location and integrate into another place within the genome without an RNA intermediate mode. DNA transposases which are encoded by autonomous DNA transposons are responsible for catalyzing DNA transpositions (Wilhelm & Wilhelm 2001; Craig 2002).

DNA transposons are a general group split into two main subclasses depending on the number of DNA strands, which are excised during the transposition mechanism. Both subclasses include four main orders (Wicker *et al.* 2007) Table 1.1.

DNA transposons are comprised of three major types: type 1 characterized by having two terminal inverted repeats (TIRs) also known as cut-and-paste elements. The elements of the second type are called rolling-circle DNA transposons (Helitrons), and the third type are known as self-synthesizing DNA transposons (Polintons) Figure 1.3. Most of the identified eukaryotic DNA transposons belong to the subclass 1 of cut-and-paste DNA transposons, currently represented by only 15 superfamilies (Kapitonov & Jurka 2008).

The cut and paste elements are abundantly distributed in all phyla, in particular, in bacteria where they are known as insertion sequences (Feschotte & Pritham 2007). Transposons of cut and paste mechanisms have a simple structure characterized by a

single open reading frame which encodes a transposase flanked within two terminal inverted repeats (TIRs). Terminal inverted repeats of the element are recognized by the transposase and then the transposons are excised and integrated somewhere else inside the same genome. The duplications of the target site are produced once the transposons are being integrated.

1.6 Tandemly repetitive DNA

Much of eukaryotic genomes constitute important components of repetitive DNA sequences, inside which a significant portion make up the non-coding DNA sequences that are tandemly repeated well known as satellite DNA sequences (Charlesworth *et al.* 1994; Elder Jr & Turner 1995; Schmidt & Heslop-Harrison 1998). In terms of experiments of cytogenetic mapping, satellite DNAs mainly exist in heterochromatic regions within chromosomes.

In general, in comparison to minisatellites and microsatellites, satellite DNAs have maximum monomers, maximum length and array, dominant and different genomic locations and different mechanisms contributed in their proliferation. However, it is not always the case, as some satellite DNAs such as those present in hermit crabs or Drosophila have simple and short repeat units (Bonaccorsi & Lohe 1991). Satellite DNAs refers to the identical DNA sequences that are highly repeated with the repeat unit size ranging from 100 to 1000 nucleotides. These are tandemly repeated in the genome, and exist at multiple copy numbers from 1000 to more than 100,000 copies. Satellite DNAs are organized in the form of repeat units also known as monomers. Monomers are commonly connected to each other in a head to tail arrangement forming a long array, which can include hundreds to thousands of copies of satellite fragment, each one with a different size and length generally over 200bp. They are the major constituent of heterochromatin blocks that are arranged into a long arrays, and are present mainly in centromeric and telomeric domains of chromosomes (Charlesworth *et al.* 1994; Meštrović *et al.* 2015; Utsunomia *et al.* 2017). Historically, when eukaryotic DNA was subjected to caesium chloride density gradient centrifugation, small DNA bands formed with a different density from the rest of the main bands of genomic DNA. These asymmetric bands called "Satellite" which usually have a lower density due to the high

proportion of AT-content. The percentage of satellite DNA varies among different organisms including animals, plants and prokaryotes from 1 to 65% (Pathak & Ali 2012). Satellite DNAs are varied in many aspects including repeat unit lengths, nucleotide sequences, and their genomic abundance or copy numbers. Repeat units of satellite DNAs are not strictly similar within a species instead they show polymorphisms in their sequences. However, when compared with DNA repeats of various species, it can be observed that they are similar in accordance with a pattern known as concerted evolution (Dover 1982).

Isolation of satellite DNA and their features have been described in two economically important domestic animals, which is in cattle and sheep within the Bovidae family in the order of Artiodactyla. In the bovine genomes, eight main types of highly repetitive unique satellite DNAs have been identified (Macaya *et al.* 1978). Satellite I DNA sequences (the 1.715 family) are tandem repeats that make up about 6-9% of the bovine genomic DNA (Kurnit *et al.* 1973). In addition, it forms the centromeric heterochromatin of all autosomal chromosomes except the sex chromosomes (Płucienniczak *et al.* 1982; Taparowsky & Gerbi 1982). However, the acrocentric X chromosome of *Bos taurus* which has satellite I at its centromeric region has been considered as the primitive conditions for the Bovidae (Chaves *et al.* 2005). Besides, to the satellite I, the bovine genome contains satellite II, III, and IV. These satellites are localized at the centromeric and pericentromeric parts of all autosomal chromosomes (Kurnit *et al.* 1973; Kopecka *et al.* 1978). None of these satellites is localized on sex chromosomes. Moreover, the satellite I, III and IV of bovine are always organized in the same order of autosomal chromosomes: p-ter-satIV-satI-satIII-q (Chaves *et al.* 2003a).

On the other hand, in ovine genomes, two satellite DNAs were recognized; satellite I (Buckland 1983; Reisner & Bucholtz 1983) and satellite II sequences (Buckland 1985). These two sheep satellites have been considered the main constituents of their centromeric and pericentromeric heterochromatin (Burkin *et al.* 1996; D'aiuto *et al.* 1997). The sheep satellite I DNA (the 1.714 family) has a monomer of 816-820bp. Similarly, to the cattle's satellite DNA sequences, it localizes at the centromeric heterochromatin of all autosomal chromosomes except the sex chromosomes (Buckland

1983; Burkin *et al.* 1996; Chaves *et al.* 2000a; Chaves *et al.* 2005). Nevertheless, the structural arrangement and abundance of sheep satellite l DNA vary within their autosomes. For example, the centromeres of the submetacentrics 2 and 3 contain higher amounts of satellite I sequences than in the largest pairs of submetacentrics 3. The sheep satellite II DNA has a repeat unit of 700bp and it localizes at the centromeric or pericentromeric heterochromatin of all sheep chromosomes but not on the Y chromosome (Burkin *et al.* 1996; D'aiuto *et al.* 1997). The chromosomal distribution of the sheep satellite II family is more variable than their satellite I DNA. For instance, it is mostly located at the centromeric regions of their autosomes, in particular to the submetacentrics and X chromosomes (Burkin *et al.* 1996).

The quantity of satellite DNAs is different depending on the host organism. For instance, they are highly abundant in genomes of some rodents and insects with percentage approximately 50% (Macas *et al.* 2010). However, in human genomes, satellites DNAs are present at around 5% (Lander *et al.* 2001), while in bovine genomes are represent about 23% (Gaillard *et al.* 1981). In plant genomes, they occupy a considerable part of around 20% (Macas *et al.* 2010). Nevertheless, it is worth noticing that the estimated amounts of satellite DNAs are probably lower than their total and real abundances that exist in various genomes. This depends on the methods that were utilized to quantify their percentages. Therefore, whole genome sequencing, proper bioinformatics tools and cytogenetics methods are required to investigate the quantity and physical organizations of satellite DNA sequences.

1.7 Identification of repetitive DNA sequences

The remarkable advances in fields of genomics, molecular biology, cytogenetics, bioinformatics and genetics in general, have unlocked many doors for the scientific community to investigate and understand the biology of repetitive DNA sequences. Furthermore, the availability of next generation sequencing methods has revolutionized the biological and informatics methods for the identification of the repetitive DNA sequences in various genomes. Thus, more and more bioinformatics and algorithms have been developed and provide new insights into the structure, dynamics and abundance of genomes in terms of their repetitive DNA motifs (Bergman & Quesneville

2007; Janicki *et al.* 2011). Recently, bioinformatics approaches for the detection of repetitive elements were broadly reviewed. Bergman and Quesneville (2007) revised some computational tools used for the identification and annotation of transposable elements. They concluded that advances in bioinformatics tools could improve the way of discovering repetitive elements. Although it has been investigated how repetitive DNA families play a versatile role in various genomes (see section on the importance of the repeat), bioinformatics tools and the algorithms for the identification of repetitive sequences are comparatively primitive in relation to those being used for genes investigation (Saha *et al.* 2008a). Nonetheless, due to the nature of repetitive sequences such as their complex structure, their lack of knowledge in terms of their function makes their identification such a challenge. Therefore, several *ab initio* tools have been established to categorize new repetitive landscapes within newly sequenced genomes (Saha *et al.* 2008a). For instance, Saha *et al.* (2008b) described and compared the most widely *ab initio* tools used for repetitive DNA sequences. Their findings indicated that each of the *ab initio* tools showed a different performance in identification of known repeats or novel repeats. For example, ReAS and RepeatFinder were efficient for identification of satellite sequences, while, RepeatScout was more effective for LTR and non-LTR retroelements.

Furthermore, recently, Lerat (2010) and Janicki *et al.* (2011) summarized and updated the bioinformatics approaches and databases that are used for the identification, analysis, classification, visualization and annotation of repetitive DNA sequences. They have classified all available bioinformatics tools and algorithms up to 2011 and divided them into many broad groups according to the usage of their categories and approaches.

The first group called homology-based methods that compare input read sequences with databases of known repetitive sequences. Some examples of libraries belonging to this method are Repbase; RepeatMasker; Transposon Express; PLOTREP, Greedier and TransposonPSI, (Herron *et al.* 2004; Jurka *et al.* 2005; Tóth *et al.* 2006; Li *et al.* 2007; Lorenzi *et al.* 2008). Another one, which utilizes NGS data as a computational pipeline (T-lex), is used to discover the annotated copies of such transposable elements (Fiston-Lavier *et al.* 2010). However, some of these tools have limitations. For instance, they are

less efficient in the detection of repetitive sequences that are characterized by the lack of conserved sequences between species such as non-autonomous class of transposable elements, MITEs. Some others like in T-lex require high coverage of sequences in order to be effective in the identification of transposable elements.

The second group is called signature-based methods which utilize the common structural landscapes of repetitive elements to locate novel DNA repeats by using known amino acids or nucleotides that are common to repetitive element classes of interest (Janicki *et al.* 2011). These approaches might also use libraries to classify newly identified repetitive elements into families (Saha *et al.* 2008a). A good example of a signature-based approaches would be for LTR retrotransposons, as Lerat (2010) indicated that both Find_LTR and LTRharvest are considered to be effective programs for the identification of LTR retrotransposons, although they have some drawbacks for instance the generating of high false positives (Rho *et al.* 2007; Ellinghaus *et al.* 2008). There are other algorithms and programs, which use strategies of signature-based approaches such as LTR_STRUC; LTR_par; RetroTector (McCarthy & McDonald 2003; Kalyanaraman & Aluru 2006; Sperber *et al.* 2007; Sperber *et al.* 2009). Furthermore, MITE-Hunter is another efficient program for signature-based methods which was developed for the identification and annotation of inverted-repeat transposable elements from genomic sequences which are not feasible by similarity-based approaches (Han & Wessler 2010).

The third group is termed *de novo* methods, which mainly based on the repetitive nature of transposable elements and other repeats in order to identify new families of repeats (Janicki *et al.* 2011). All repeat families are extremely assembled by *de novo* methods those meeting thresholds of their copy numbers. These methods are either build on *k*-mer frequency (the occurrence of small strings) or self-alignment based approaches. In contrast to the previous two methods, homology and signature approaches that begin with classification, *de novo* methods need classification information after discovering repetitive families (Kurtz *et al.* 2008; Janicki *et al.* 2011). Accordingly, many computational tools have been developed based on the characters of *de novo* based methods. Examples for self-alignment approaches are RECON (Bao & Eddy 2002), PILER (Edgar & Myers 2005) and others. While, programs using *k*-mer frequencies are based

on counting the occurrence of short identical motifs that are present in genome sequences in multiple copies. These including REPuter (Kurtz *et al.* 2001), ReAS (Li *et al.* 2005), RepeatScout (Price *et al.* 2005) and others. It seems that no single comprehensive bioinformatics tool could be capable of investigating all repetitive DNA sequences including tandem and dispersed elements (Bergman & Quesneville 2007; Saha *et al.* 2008a). However, in 2010 and 2017, other bioinformatics tools such as RepeatExplorer and TAREAN pipelines were launched. Because these tools are available for public to use and include properties of signature-based and *de novo* methods, we used them for investigation the NGS data of sheep in order to characterize major classes of repetitive DNA (see below).

1.7.1 **Graph-based clustering approach (RepeatExplorer)**

The advent of NGS (see section 1.14) enables rapid sequencing of large amounts of DNA, and thus could provide better insight into the landscapes of repetitive DNA elements between and within different genomes. Many bioinformatics tools for repeat identification are available (see section 1.7). Recently, Novák *et al.* (2010); (Novák *et al.* 2013) developed a graph-based read clustering approach implemented in the RepeatExplorer program on a Galaxy server (www.repeatexplorer.org). This approach uses the mgblast tool to perform an all-to-all sequence comparison of NGS reads. Then, mutual sequence similarities above 90% over at least 55% of the read length are charted and used in order to construct graphs using a De Bruijn graph approach, where nodes represent the sequence reads and edges (lines) between the nodes refer to a strong similarity between reads. Afterwards, according to the amount of similarity hits, these sequences are then assembled into contigs within each cluster by implementing the CAP3 program. The used threshold and similarity search against known repeat databases will characterize the size of each cluster, and thus will calculate its graphical layout. The algorithm Fruchterman and Reingold is also applied to calculate relationship between and within clusters. A schematic illustration of the RepeatExplorer components and workflow is shown in Figures 1.5 & 1.6.

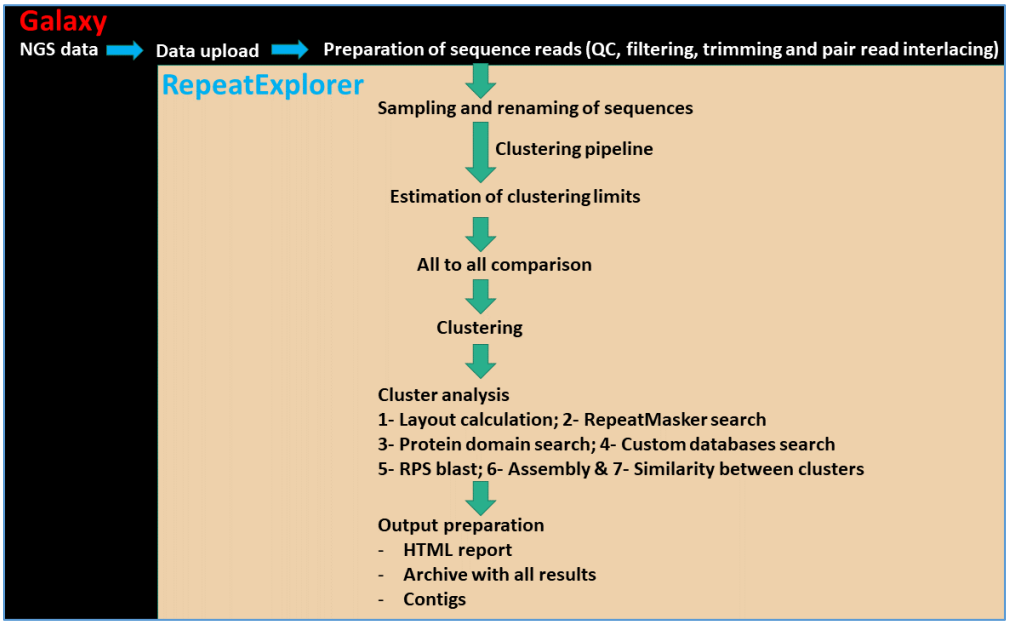


Figure 1.5 Shows workflow of RepeatExplorer. Fasta file of NGS data is uploaded and pre-processed; it uses a clustering pipeline to compare sequence reads all-to-all to find resemblances; to make repetitive assemblies by following graph-based read clustering; to compare these assemblies to conserved domains of repetitive related proteins and to available repetitive databases; and, after searching for mutual similarity, to calculate graph layouts of cluster (which can be used to classify and annotate repetitive sequences in clusters).

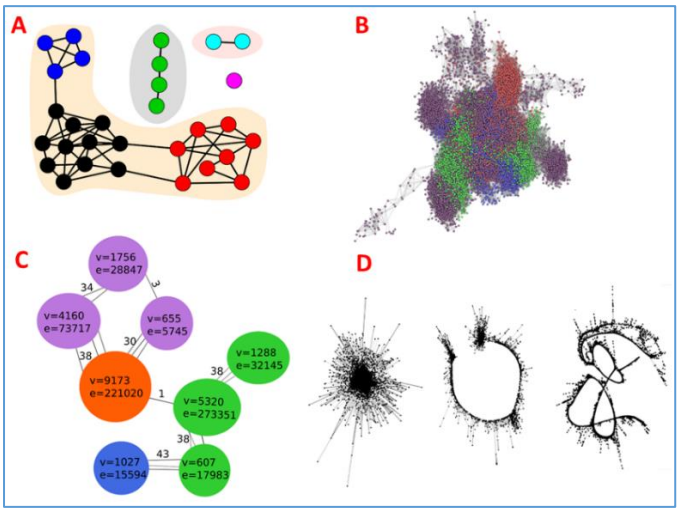


Figure 1.6 Organization of sequence reads in a graph structure.

A. Different types of clusters; reads are communicated with each other through sequence similarities forming clusters where nodes (vertices)=single reads, while edges=sequence overlaps
B. Overlay of read clusters where the hierarchical agglomeration algorithm used to identify and label different communities of reads (classes of repeats)
C. Colored circles of resulting clusters reflect relationship between reads using (V) vertices= number of single reads, while (e) edges= number of sequence overlaps use to find relationship between and within the clusters.
D. The Fruchterman and Reingold algorithm used to calculate graph layouts drawing sequence relationship in a 3D dimension. Different repeat classes produce different shapes of graph (see Appendix 1.1).

RepeatExplorer is mainly used to identify and characterize the genomic composition of repetitive regions. It can handle and investigate millions of unassembled raw reads and characterize their diversity and abundance in terms of their repetitive DNA sequences (Macas *et al.* 2011; Dodsworth *et al.* 2015; García *et al.* 2015; Feng *et al.* 2017). Thus, the well-known and novel tandemly and dispersed repeats can be discovered. Accordingly, we applied this tool to identify and quantify the major repetitive DNA classes in sheep genomes.

Although there is no research into the weaknesses of this program, Staton and Burke 2015 stated that RepeatExplorer needs a better integrated design and productive computation in order to efficiently investigate genomic reads. The ability to assemble repetitive communities into a single cluster and study the phylogenetic relationships between these communities would further improve the RepeatExplorer program. Furthermore, it is unclear why RepeatExplorer is able to produce more clusters in plant genomes than in animal genomes.

1.7.2 **TAndem REpeat ANalyzer (TAREAN)**

TAREAN is a computational pipeline running under the Galaxy environment available via the RepeatExplorer server (http://www.repeatexplorer.org/). TAREAN was developed by Novák *et al.* (2017). The workflow of this tool starts with analysing the graph-based read clustering principles identifying genomic repetitive DNA sequences from unassembled NGS raw reads (see above section 1.7.1). Putative tandemly repeated sequences are subsequently discovered by the presence of circular graph of clusters shown in Figure 1.7 (see also Appendix 1.1). Sequences of monomers and most consensuses are then reconstructed through identification and extraction the most frequent *k*-mers (oligomers of length k) by which TAREAN builds De Bruijn circular graphs representing the most tandemly repeated motifs. This tool also reports many characteristics about tandemly repeated monomers. These are characterized in nucleotide sequences, lengths, *k*-mer motifs, alternative variants present in each consensus or monomer. Additionally, graph and logo images which represent the most conserved part of the consensus sequence of each reconstructed monomer are also

produced and thus can be selected as a good candidate for the oligo probe in order to be used in further molecular cytogenetic techniques.

The strengths of this tool are its ability to identify previously characterized satellite repeats and discover novel tandemly repeated sequences by using millions of unassembled NGS reads to classify them as either high or low confidence putative satellites. Thus, this pipeline avoids the problem of assembling the satellite DNA sequences into their most consensuses. Another remarkable feature of TAREAN is its ability to estimate the genomic proportion of tandemly repeated sequences and group them into one cluster as one family (González *et al.* 2017; Li *et al.* 2017; Novák *et al.* 2017).

Although the weaknesses of TAREAN have not been reviewed yet, the construction of longer repeat units (higher order) of tandemly repeated DNA from NGS is still a big challenge for this pipeline. Furthermore, TAREAN still lacks the ability to analyse the phylogenetic relationship between and within the major and novel tandemly repeated consensuses. In this study, TAREAN will be utilized in order to identify tandemly repeated members in sheep genomes not based on prior knowledge.

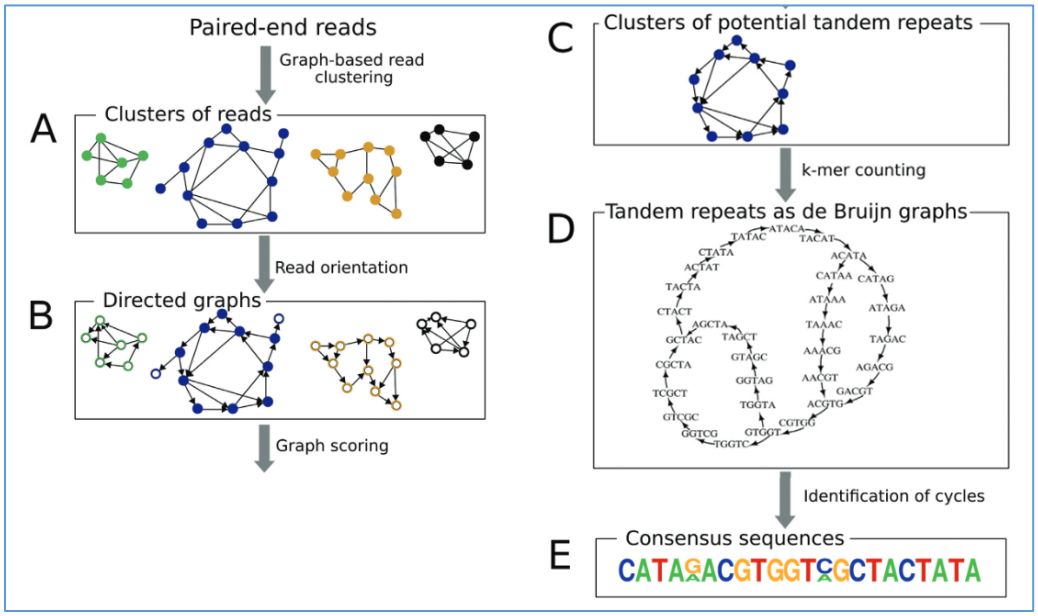


Figure 1.7 Graphic illustration of the TAREAN analysis workflow. Taken from Novák *et al.* (2017)

1.8 Importance of repetitive DNA sequences

A significant portion of eukaryotic genomes are comprised of repetitive DNA sequences characterized by a considerable degree of diversity and heterogeneity in their repeated elements (Charlesworth *et al.* 1994). Previously, most repetitive DNA sequences were regarded as 'Junk DNA' or 'Selfish DNA' (Schmidt & Heslop-Harrison 1998; Biémont & Vieira 2006) and the repeats were thus considered to be unusable or inactive elements. This leads to the fundamental question of why repetitive DNA families should be predominant in the genome. Thus, more recently, the concept of the value of these repetitive DNA sequences has changed. This is mainly because of the scientific revolution occurring in different fields such as the advent of next generation sequencing, the availability and accessibility of many sequenced genomes, the significant advances in molecular and cytogenetics techniques and the dramatic breakthrough of bioinformatics and algorithm tools for the classification and characterization of repetitive elements. Thus, over the last four decades, a huge number of studies have been carried out in several fields in the context of clarifying the possible biological importance of repetitive DNA sequences in various genomes. For instance, it has been indicated that several repeats are important for the functional and structural organization of the genome, mainly due to the finding of transcribed regions inside repetitive arrays and their participation in controlling gene expressions (Grewal & Jia 2007; Chuong *et al.* 2017a).

In addition, repetitive DNA sequences evolve rapidly, causing variation in their distribution and capacity, and influencing genome organization and function. These features make repetitive DNA sequences powerful tools for the analysis of phylogenetic and evolutionary relationships between species, and for comparative investigations of genomes (Chaves *et al.* 2000a; Heslop‐Harrison & Schwarzacher 2011). Furthermore, they also provide considerable amounts of information concerning a number of prominent aspects such as driving forces and the order related to the evolutionary processes of genomes, rendering them valuable markers for the investigation of the phylogenetic relationships of species, along with micro-evolutionary studies (Chaves *et al.* 2000a; Ugarković & Plohl 2002; Feliciello *et al.* 2005; Adega *et al.* 2006; Adega *et al.* 2007).

Moreover, tandem repeats including satellite sequences have characteristic features with functional properties. The centromeres (i.e. the primary constriction of chromosomes important for proper chromosomal segregation during cell division) contain large arrays of repetitive DNAs, including satellites and dispersed elements (Heslop‐Harrison & Schwarzacher 2011; Biscotti *et al.* 2015a; Garrido-Ramos 2015). The telomeres contribute to proper chromosome replication and perform an important function in chromosomal stability. Satellite DNAs are also involved in the formation of heterochromatic sections, which are critical for correct chromosomal behaviors in both meiosis and mitosis (Charlesworth *et al.* 1994; Csink & Henikoff 1998). In terms of tandemly repeated satellite DNA sequences, it has been proven that the centromeric equid satellite 37cen significantly contributed in the function of centromere and is transcriptionally active (Cerutti *et al.* 2016). This supports the hypothesis that centromeric transcripts could play a significant role in the function of the centromere.

It has been indicated from the initial studies carried out on mice, transposable elements (LINE-1 retrotransposons) are associated with determining the construction and expression of the transcriptomes(Han *et al.* 2004; Han & Boeke 2005). In mammals, mobile elements endorse the diversification and regulatory variation of genes with specialized functions (van de Lagemaat *et al.* 2003). Furthermore, a substantial fraction of regulatory sequences in human genomes originated from derived sequences of transposable elements, which occupy approximately 25% of analyzed promoter regions (Jordan *et al.* 2003). In addition, since the first suggestion by McClintock 1984 about the activation of transposable elements in response to the many challenges exposed to the genome, the release and importance of epigenetic regulation and the silencing of transposable elements has been stated in response to temperature, pathogen infection, UV exposure and others. For example, the reactivation of Tnt1 retrotransposon in response to tobacco infection and the activation of some other transposable elements in response to heat stress such as in Drosophila melanogaster (Slotkin & Martienssen 2007).

It has been clarified that DNA repeats play crucial roles in the evolutionary events in eukaryotes. In general, these include the influence of repetitive elements in

recombination, the gene expression and function, the maintenance of chromosome structure, the increase of genetic variation and diversity through mutation, the generating of novel genes by means of their transposition, the contribution to adaptation and evolution in plants and the association with chromatin modifications (Thornburg *et al.* 2006; Cohen *et al.* 2009; Lisch 2013; Krassovsky & Henikoff 2014; Chuong *et al.* 2017a).

1.9 Meiosis

The cell is the basic and functional unit of life due to its self-division and reproduction, which is considered as the most fundamental characteristics. Animal cell undergoes two main types of nuclear divisions known as mitosis and meiosis. Mitosis occurs in somatic cells in which parental cells generate two identical daughter cells having the same chromosome number. Mitosis requires replicated chromosomes in order to distribute equally in each new cell (Hartwell *et al.* 2011). Unlike mitosis in somatic cells, meiosis happens during gamete formation, e.g. Egg and sperm cells each being haploid and having only one set of chromosomes. This reduction in chromosome numbers by half during meiosis is of importance for sexual reproduction, ensuring that the next generation has a diploid number of chromosomes. Furthermore, it allows involvement in exchanging and combining the genetic materials of both parents and producing a genetically distinct generation carrying new combination of traits which are different from that of the parental individuals (Lee & Amon 2001; Hunt & Hassold 2002; Hartwell *et al.* 2011). Meiosis involves two sets of nuclear divisions known as meiosis I and meiosis II, but only one round of DNA replication that occurs prior to meiosis I (Cheng *et al.* 2006; Hartwell *et al.* 2011). Each meiotic division includes prophase, metaphase, anaphase, telophase and cytokinesis. Meiotic prophase I is the most important phase due to the establishment of the synaptonemal complex SC, the pairing of homologous chromosomes and the chiasmata (Lee & Amon 2001).

1.10 **The synaptonemal complex (SC)**

The synaptonemal complex (SC) was originally described 61 years ago, by Fawcett (1956) and Moses (1956) independently. SC is a tripartite proteinaceous scaffold that assembles during the meiotic prophase I and has been discovered in all sexually reproducing eukaryotic organisms. Ultrastructural analysis from transmission electron microscopy revealed that the SC is a zipper- or ladder- like structure composed of two lateral elements LE also called two chromosomes axes. The central element (CE) connects the two lateral elements upon synapsis over their entire length by fine fibrillary structures called transverse filaments (TFs), forming a tripartite organization. The SC maintains a presynaptic arrangement between the axes of the homologous chromosomes and supports crossing over and recombination events, creating physical connections between the chiasmata. Univalent achiasmatic chromosomes, arise when there is a failure to form chiasmata. Thus, risk of mis-segregation at the first meiotic division and aneuploid germ cells occur. Therefore, the proper formation of the SC is evolutionarily conserved assemblies that have an important contribution in many significant events during meiosis processes particularly in synapsis, recombination and chromosomal segregations (Zickler & Kleckner 1998; Page & Hawley 2004; Costa & Cooke 2007; Yang *et al.* 2008; Handel & Schimenti 2010) Figures 1.8 & 1.9.

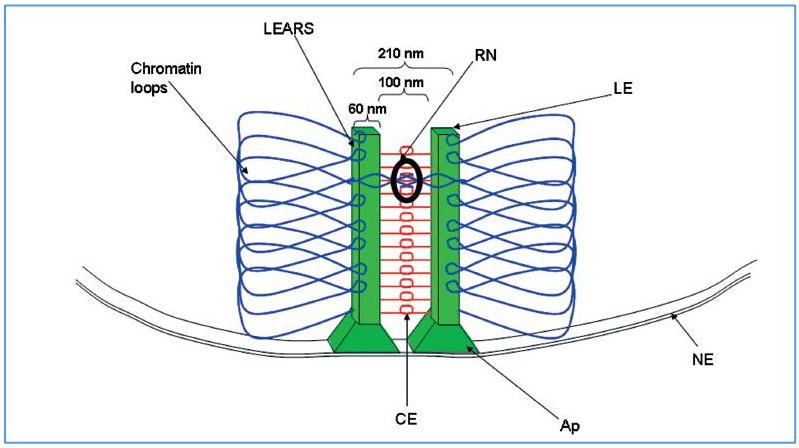


Figure 1.8 SC is a tripartite structure consisting of two lateral elements LE and a central element CE. The homologous chromosomes anchor to the lateral elements by Lateral Element-Associated Repeat Sequences (LEARS). The lateral elements are attached to the nuclear envelope (NE) at their ends through the adhesion plate (AP). The recombination nodule (RN), as a secondary specific structure, has the function of mediating recombination in meiosis and aids the generation of cross-overs between homologous chromosomes (Yuan et al., 2000). The thickness of the complex is 210nm including two LEs each with 60nm in width and a central space with 100nm. Taken from Hernández-Hernández *et al.* (2009).

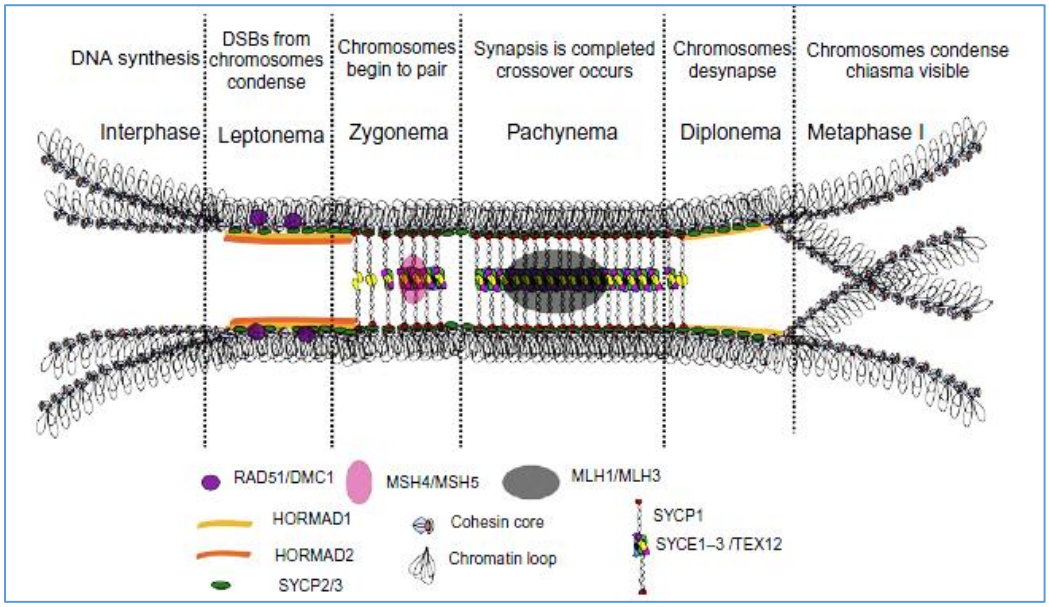


Figure 1.9 Schematic representation of the events occurring between homologous chromosomes during prophase I of the first meiotic division. Sub stages of prophase I and relative progression of synapsis and recombination are depicted by the spatiotemporal distribution of proteins involved in the SC formation and recombination. Taken from Bolcun-Filas and Schimenti (2012).

1.10.1 **Repetitive elements and synaptonemal complex (SC)**

During formation of the SC, chromatin at prophase I is organized in large loops that are attached to the SC Figure 1.8. It has been suggested that the chromosomal DNA binds to the SC through some proteins of the axial elements (Pelttari *et al.* 2001; Kolas *et al.* 2004). Johnson *et al.* (2013) used chromatin immunoprecipitation and DNA sequencing to reveal that the components of the mammalian SCs are characterized by some specific elements of repetitive DNA sequences including SINEs and other interspersed repeats. Furthermore, their observations implied that the most actively retrotransposing SINEs in the rhesus monkey AluY and in mice B1 might play several possible functions in the binding of axial elements of the synaptonemal complex. One possible function involves acting as a regulator or suppressor of retrotransposition, the other is as the anchoring point for the SC. Several studies from various organisms investigate whether the repetitive sequences play a crucial role in the earliest activities of the meiotic prophase. For examples, DNA repeats in the telomere simplify the interactions between homologous chromosomes through the establishment of the meiotic bouquet (Johnson *et al.* 2013). Thomas and McKee (2007) in Drosophila, Phillips *et al.* (2009) in their studies

of *C. elegans* show the association with other definite repetitive sequences in the synapsis and pairing process.

Biochemical studies have revealed that the DNA sequences associated with the SC are corresponding to the (GT/CA) sequences, such as the LINEs and SINEs (Karpova *et al.* 1989; Karpova *et al.* 1994). Pearlman *et al.* (1992) also identified a small subset of repetitive DNA sequences and demonstrated their association to the SC in rat and mouse. Dadashev *et al.* (2005) and Grishaeva *et al.* (2005) have suggested in their in silico studies of human that some DNA repeat sequences such as Alu elements might contribute to the SC by anchoring the chromosomes to their axial elements. In addition, they also verified that simple repeats such as (GT/CA)n sequences were the adjoining meiotic recombination sites.

Hernández-Hernández *et al.* (2008) performed chromatin immunoprecipitation (ChIP) assays using the major protein of the lateral elements (SYCP3) in rat spermatocytes. They have identified many genomic sequences, including 100 independent DNA sequences, representing various repetitive elements such as SINE, LINE, LTR, satellites, and simple repeats. They indicated in their bioinformatics analysis that these repetitive elements were highly abundant and amplified throughout the entire rat genome. Interestingly, their results of fluorescence *in situ* hybridizations combined with the immunolocalization of SYCP3 showed a clear association of isolated repetitive sequences to the lateral elements of the SC through the detection of signals over the chromatin nearby the SC and through protrusion of the small loops from the lateral elements into the central region. For example, signals of SINE-M9 and LINE-M8 probes are localized in the majority of the pachytene nuclei as thread-like structures corresponding to the SC. These results conclude that repetitive DNA sequences perform active role in linking the chromosomes to the protein scaffold of the SC at pachytene in rat spermatocytes. Moreover, Nin *et al.* (1993) and Ortiz *et al.* (2002) described the existence and distribution of DNA sequences within the lateral elements in rat, mice and guinea-pig. It has been demonstrated that some putative DNA binding motifs have been found in sequences analyzed in the major proteins of the SC including SYCP1 and SYCP2 (Meuwissen *et al.* 1992; Offenberg *et al.* 1998). It has also been found that repetitive

elements provide some features in their chromatin organization such as modifications of precise histone post-translation (Kondo & Issa 2003; Martens *et al.* 2005; Peng & Karpen 2007). Hence, repetitive elements involved with the lateral elements have such functional significance in conferring definite chromatin arrangement and attachment (Hernández-Hernández *et al.* 2008). However, the meiotic behavior of tandemly repeated sequences is not clear yet. Therefore, fluorescent *in situ* hybridization combined with immunostaining could show meiotic organization of major tandem repeats in sheep.

1.11 Satellite DNAs, Heterochromatin and Robertsonian translocations in the karyotype evolution

Heterochromatin was described in 1928 by Heitz as a fraction of chromatin that maintains its compact state during the interphase and even at the beginning of the prophase. On the contrary, the remaining part of the chromatin (the euchromatin) decondenses at the end of the telophase and re-condenses for the next division. Heterochromatin was classified by Brown (1966) into two main classes: constitutive and facultative heterochromatin. The first class refers to the more condensed and poorly expressed chromatin, which is prevalent in pericentric and near telomeric regions of chromosomes. The eukaryotic genome contains a considerable amount of constitutive heterochromatin which is composed mainly of repetitive elements, in particular, satellite DNA sequences (Brutlag 1980; D'aiuto *et al.* 1997; Sumner 2003). Furthermore, (Chaves *et al.* 2004) thought that constitutive heterochromatin regions are the 'hotspots' for structural chromosome rearrangements.

The facultative class, on the other hand, is variable and is a gene-containing chromatin fraction. However, it displays a suppression of gene expression, due to higher condensation and epigenetic mechanisms (Grewal & Jia 2007). The classic example being the inactive X chromosome in females.

The formation of heterochromatic short arms on acrocentric or telocentric chromosomes is common to chromosomal evolution events and can be observed by comparison closely related species. Additionally, Robertsonian translocations (the

formation of meta- or sub metacentric chromosomes through either the fusion of two acro- or telocentric chromosomes or the reverse fission process) are common features of mammalian chromosome evolution (Sumner 2003). Sheep chromosomes have been investigated to demonstrate these evolutionary events among the Bovidae family, as sheep has three pairs of large submetacentric chromosomes, which are corresponding to the banding patterns of the six acrocentric chromosomes that present in goats and cattle (see section 1.12). Centric fusion named Robertsonian translocations due to their first explanation by Robertson (1916). Centric fusion translocations are considered as prevailing mechanisms that include important chromosomal rearrangements that are significantly involved in the evolution of the karyotype of different species of mammals. Centric fusion translocations are carried out through the two main steps starting by the fusion of two non-homologous acrocentric autosomes over their centromeres. Dicentric biarmed chromosomes are the outcome of the first step and then transformed to the monocentric chromosomes when the chromosome reached its stable and final state as a result of the loss of one centromere or because of the loss of constitutive heterochromatin at one or both centromeres (Iannuzzi *et al.* 2009) due to the presence of satellite DNA repeats in constitutive heterochromatic regions and due to the loss of constitutive heterochromatin through the centric fusion translocations. This could be one of the main reasons that led to the loss of satellite I sequences in biarmed chromosomes of the sheep karyotype which means that satellite DNA could have played a role in centric fusions (Chaves *et al.* 2000b; Adega *et al.* 2009). It has been proposed that the rapid evolution of satellite DNA sequences and the dynamic behavior (intragenomic movements) of these satellite families amongst non-homologous chromosomes and between different chromosomal locations, including centromeres and telomeres, caused the promotion of chromosomal rearrangements (Wichman *et al.* 1991). Further to the well-known six centric fusions, another novel chromosomal translocation t(8;11) was found in sheep using chromosome painting and probes of both major satellites which provides a robust mark for an intermediate step in the evolution of the biarmed chromosomes (Chaves *et al.* 2003b).

In sheep, the lack or absence of satellite I signal in the biarmed chromosome pairs suggested that they have lost their sequences during formation of biarmed

chromosomes suggesting the loss of constitutive heterochromatin (Chaves *et al.* 2000b). Similarly, in cattle, the translocated chromosome rob (1; 29), unlike the acrocentric autosomes, showed no hybridization of α-satellite I sequence at its centromere, although it was present at the centromeres of acrocentric chromosomes including 1 and 29. This means that the translocated chromosome rob (1; 29) had lost the α-satellite I sequence at its centromeric region during translocation mechanisms (Chaves *et al.* 2000b). Thus, the satellite probes can be used as karyotypic markers of the evolution within the Bovidae family (Chaves *et al.* 2000a). Therefore, investigation of sheep genome in terms of their repetitive sequences could provide better understanding of the role of these repeats in the evolution of sheep karyotype.

1.12 **Karyotype evolution in sheep**

Setting up a standard karyotype of domestic animals is considered one of the major steps progressed in the history of animal cytogenetics in studying the chromosomal complement and structure. For example, the 'standard' G-banded karyotype of cattle and of other domestic animals was provided for the first time in a reading conference (Ford *et al.* 1980). The first investigation of karyotyping was carried out according to the size of chromosomes. Thereafter, and more recently, many banding techniques including G-; R-; C-; T- and Q- banding were developed in addition to many other cytogenetics methods in order to investigate and karyotype mammalian chromosomes.

Over three decades, the chromosomes of species of the Bovidae family including sheep, goats, and cattle have been studied. Analysis of karyotypes and cytogenetic maps are essential for understanding the chromosome evolution and genomic organization of domestic ruminants within the Bovidae family. Evans *et al.* (1973) introduced the first comparison of chromosomal bands of some members of the Bovidae family and also supported the hypothesis of Wurster and Benirschke (1968) who suggested that all bovids were originated from a common ancestral bovid. Later, several authors including Buckland and Evans (1978), Bunch and Nadler (1980), Berardino *et al.* (1981), Mensher *et al.* (1989), Iannuzzi *et al.* (1990), Hayes *et al.* (1991), Gallagher Jr and Womack (1992) used various banding techniques and observed banding homologies of chromosomes in a huge number of species of the Bovidae family. Banding patterns of the autosomal

chromosomes of sheep, goats, cattle, and buffaloes showed high similarity, which could be indicating considerable conservation of their karyotypes during evolution.

According to fossil records, the lineages of sheep and goats diverged 5-8 million years ago while, the lineages of Caprinae subfamily diverged from the Bovinae subfamily 17-20 million years ago (Maddox 2005). In spite of the different divergence times, the diploid number of domestic cattle and goat chromosomes is the same with 58 acrocentric autosomes. On the other hand, the karyotype of domestic sheep *Ovis aries*, has 2n=54 chromosomes of which 46 are acrocentrics, six are large biarmed autosomes and then the sex chromosomes X and Y.

Evans *et al.* (1973), Zartman and Bruere (1974) and Bunch *et al.* (1976) used banding techniques such as G-banding, and recognized that six pairs of acrocentric autosomes of goat (*Capra hircus*) were involved in the centric fusion translocations. Since the Reading Conference in 1976 that established standard karyotype of domestic animals including sheep and goats, many elongated chromosome-banding techniques have been used. Mensher *et al.* (1989) used high-resolution G-banded karyotype to reexamine the chromosome complements, to determine banding equivalence between sheep and goat karyotypes and to verify the previous results. Their results confirmed the previous identifications, indicating that the six acrocentric autosomes of goat 1/3, 2/8 and 5/11 were displayed in Robertsonian translocations forming three submetacentric chromosomes in the domestic sheep. Moreover, Kattanovskaya and Serov (1994) compared the chromosomes of sheep, goats and cattle with the use of high resolution GTG-banding patterns and they found the same results of Mensher *et al.* (1989) as described in Figures 1.10 & 1.12.

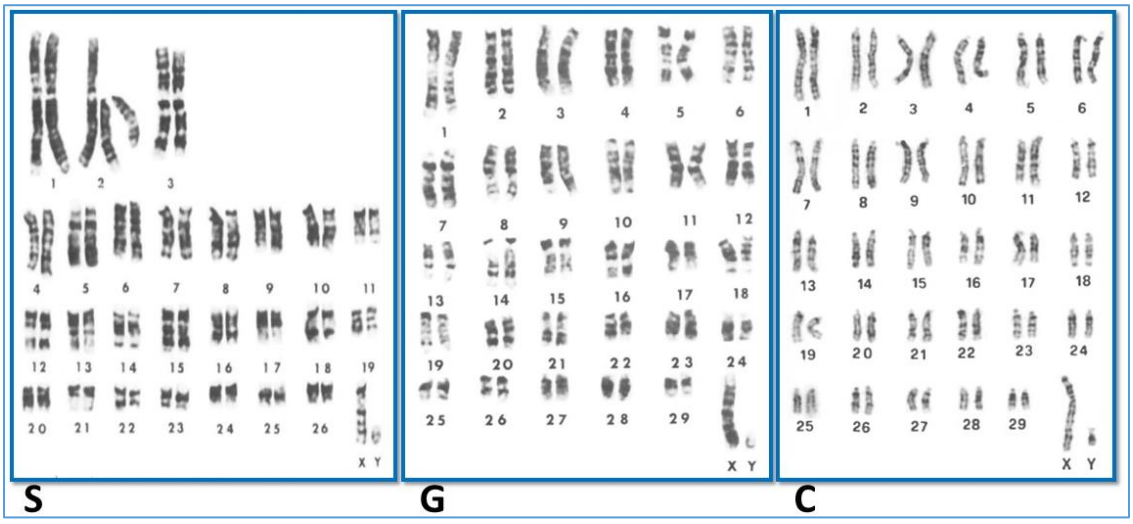


Figure 1.10 (S) High-resolution G-banded karyotype of the domestic sheep. Figure (G) Proposed standard high-resolution G-banded karyotype of the domestic goat. Figure (C) GTG-banded karyotype of cattle (*Bos taurus*). The sheep karyotype has 54 chromosomes, consisting of 3 pairs of submetacentric and 23 pairs of acrocentric autosomes, a large acrocentric X, and a very small metacentric Y. The goat karyotype has 60 chromosomes comprising of 29 pairs of acrocentric autosomes, a large acrocentric X, and a very small metacentric Y. Taken from Mensher *et al.* (1989); (Iannuzzi 1996).

In general, the genus *Ovis* is one of the best model for studying and understanding karyotype evolution within species of the Bovidae family at the chromosomal level. Phenotypically, the closest genus and species to study and analyze the Robertsonian translocations in the domestic and wild sheep is goats (*Capra hircus*, 2n=60). Thus, as a result of Robertsonian translocations and the evolution that occurred at the autosomal karyotypic level in goats, their diploid number reduced to 58 (*Ovis vignei*), 56 (*Ovis ammon*), 54 (*Ovis aries*) and other species, and 52 (*Ovis nivicola*). In other words, such classification for the members of *Ovis* genus might be set through grouping and arrangement the number of biarmed chromosomes in ascending order. The possible classification is as follows: (A)- One pair of biarmed chromosomes originating from the Robertsonian translocations of 1; 3 found in (*Ovis* vignei , 2n = 58); (B)- two pairs of biarmed chromosomes originating from the Robertsonian translocations 1; 3 and 2; 8 found in (*Ovis* ammon , 2n = 56); (C)- Three pairs of biarmed chromosomes originating from Robertsonian translocations 1; 3, 2; 8, and 5; 11 found in the domestic sheep (*Ovis aries* , 2n = 54) and also other species of *Ovis* such as *Ovis canadensis*, *O. dalli*, *O. musimon* and *O. orientalis*); and (D)- Four pairs of biarmed chromosomes originating from Robertsonian translocations of 1; 3, 2; 8, 5; 11 and 9; 19 found in Siberian sheep (*O. nivicola* , 2n = 52). These findings indicate that the Robertsonian translocations have

been the predominant mechanism in the evolution of the autosomes of bovids species (Bunch *et al.* 1976; Bunch & Nadler 1980; Gallagher *et al.* 1999; Bunch *et al.* 2005; Iannuzzi *et al.* 2009) Figure 1.11.

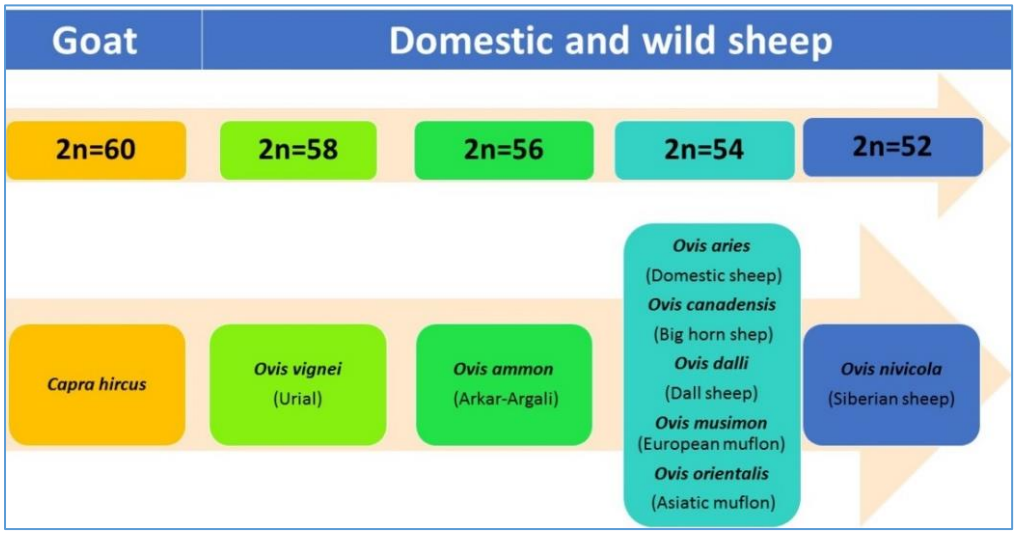


Figure 1.11 Diploid number in the genus of Capra (domestic goat) and *Ovis* (domestic and wild sheep)

The G- and R-banding techniques (Iannuzzi & Di Meo 1995), and comparative FISH-mapping (Di Berardino *et al.* 2001; Iannuzzi *et al.* 2001; Di Meo *et al.* 2007; Goldammer *et al.* 2009) techniques demonstrated that the origin of three large biarmed chromosome pairs in the present day domestic sheep is the Robertsonian translocations of the homologous chromosomes of goat and cattle 1/3, 2/8 and 5/11 Figure 1.12.

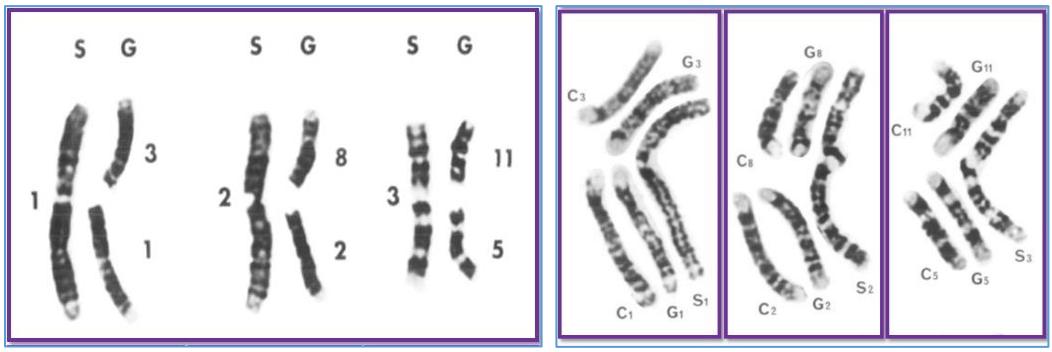


Figure 1.12 Combined haploid karyotype of GTG-banded chromosomes of cattle (C), goat (G), and sheep (S). Figure shows involvement of six acrocentric chromosomes of each of cattle and goat through Robertsonian translocations (centric fusions) forming three submetacentric chromosomes in sheep. C= Cattle acrocentrics, G= Goat acrocentrics and S= Sheep submetacentrics. Taken from Mensher *et al.* (1989); (Kattanovskaya & Serov 1994).

The sex chromosomes differ in size (amount of constitutive heterochromatin) and shape (centromere location). In cattle, the X-chromosome is submetacentric while in sheep and goat X-chromosomes are acrocentric. Similarly, the Y-chromosome varies in size and shape, as it is submetacentric in cattle, while in sheep and goat it is very small metacentric Y-chromosome Figure 1.13. Minute short arms were observed in most of the acrocentric autosomes.

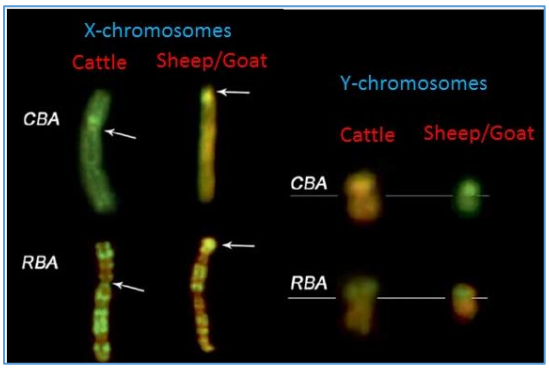


Figure 1.13 CBA- and RBA-banding on X and Y-chromosomes of cattle and sheep/goat. It shows also the submetacentric Y chromosome of cattle is longer than the small Y metacentric sheep chromosome.

The presence of 29 acrocentric autosomes in the Bovinae subfamily as in domestic cattle provide a key indicator of the ancestral karyotype for other species within Bovidae family (Chaves *et al.* 2005). The Bovid ancestral karyotype (BAK) with 2n=60 and the fundamental number= 60 derived as a result of two levels of evolutionary events of chromosomes Figure 1.14. The first level was the transition which happened between the ancestral Cetartiodactyla karyotype (CAK) with 2n = 52 and the ancestral pecoran karyotype (PAK) with 2n = 58. This transition was characterized by five fusions and nine fissions. While, in the second evolutionary level, a single transition occurred from pecoran ancestral karyotype (PAK) with 2n=58 to Bovid ancestral karyotype (BAK). The karyotypes of the Bovinae subfamily nowadays are almost identical to the Bovid ancestral karyotype (BAK). This resemblance between these two karyotypes is supported by clear evidence of the presence of acrocentric chromosomes, including the sexual chromosomes. For example, chromosomes 9, 14 and an X chromosome of Bovinae were observed in the Bovid ancestral karyotype (BAK).

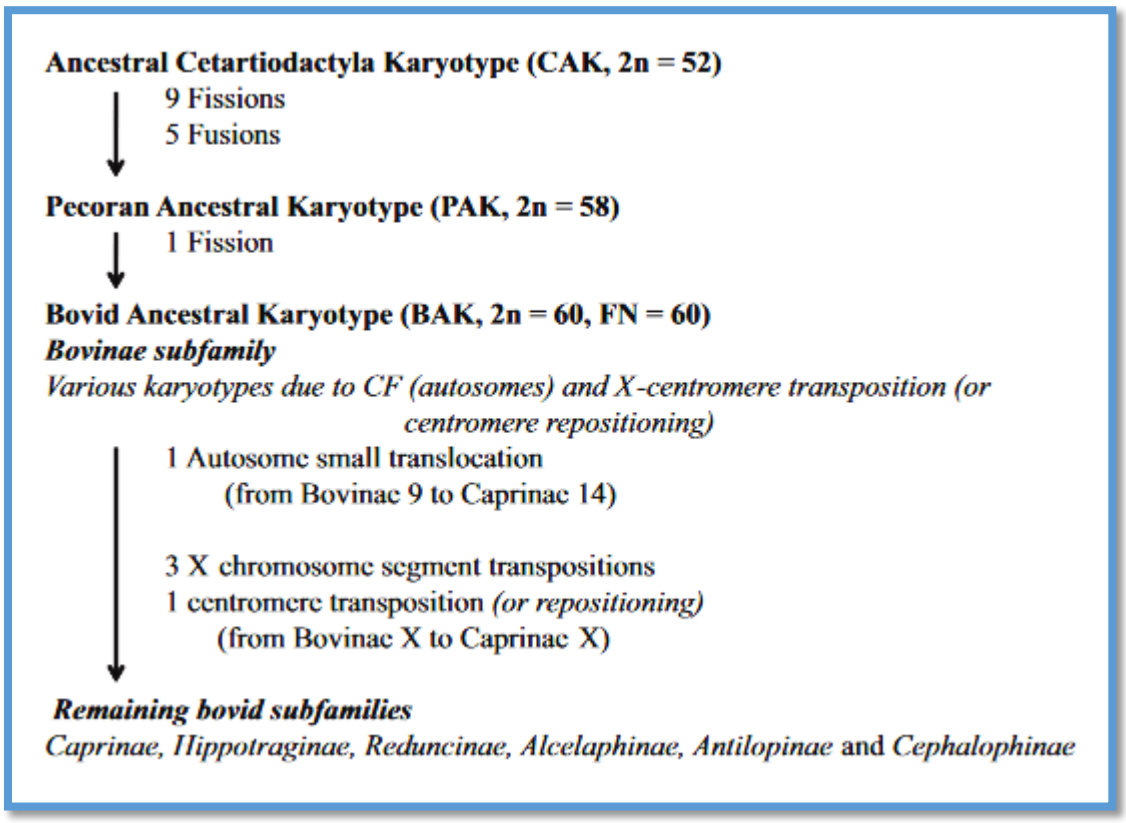


Figure 1.14 Schematic evolution of bovids from order Cetartiodactyla and advanced pecorans. Taken from (Balmus *et al.* 2007); Iannuzzi *et al.* (2009).

On the other hand, interestingly, various karyotypes of the remaining species of the Bovidae subfamilies including the Caprinae subfamily were derived from the karyotype of the Bovinae subfamilies. This is predominantly due to the occurrence of translocations, in particular the centric fusion between different autosomal chromosomes of Bovinae and Caprinae subfamilies Figure 1.14. (Balmus *et al.* 2007; Iannuzzi *et al.* 2009).

1.13 Fluorescence *in situ* hybridization (FISH)

In situ hybridization is a common and powerful method used to localize DNA sequences on their "morphologically preserved cytological specimens" such as genomes, chromosomes, interphase cell nuclei, and extended chromatin fibres (Haaf 2000; Schwarzacher & Heslop-Harrison 2000; Schwarzacher 2003). Before the arrival of fluorescence *in situ* hybridization in biological fields, the chromosomal distribution of target DNA sequences was one of great challenge in cytogenetic studies. The early radioactive methods of hybridization depended on radiolabelled probes (Gall & Pardue

1969; John *et al.* 1969). These were relatively expensive and time-consuming and suffered from several drawbacks, including unstable probes; limited resolution; and hazardous materials. This encouraged the development in the 1980s of a new powerful technique, known as fluorescence *in situ* hybridization, which made substantial progress in the safety, resolution, speed and localization of simultaneous multiple targets (Schwarzacher & Heslop-Harrison 2000; Levsky & Singer 2003; Schwarzacher 2003). As a result, a number of major scientific breakthroughs in cytogenetic fields were achieved, establishing the localization of targeted DNA sequences on chromosomes by FISH. Furthermore, the approach of this technique initiated the 'molecular cytogenetic era' and the 'phylogenomic era,' integrating cytology with genomic DNA sequence data. This promotes the integration of the physical location of DNA sequences with their molecular information on chromosomes, whole genomic DNA, or even parts of chromosomes (Schwarzacher 2003; O'connor 2008; Chen & Chen 2013). This is thus a powerful means of discovering the distribution and abundance of repetitive DNA families, and of establishing their physical map positions.

The basic principles of FISH experiments include the following: sourcing of probe DNAs; probe labelling; chromosome slide preparations; pre-treatment of slides; denaturation and hybridization of probe and target sequences; washing; detection; and interpretation (Haaf 2000; Schwarzacher & Heslop-Harrison 2000). Firstly, the DNA sequence to be used as a probe needs to be complementary to the target sequences of interest. FISH probes used to analyze genomes are available in a variety of sequence sources: clones with short or large inserts containing genes; unique sequences or repetitive DNA sequences, as well as total genomic; whole chromosome or chromosome arm specific DNA sequences used for chromosome paints (Schwarzacher & Heslop-Harrison 2000). These can be used to identify chromosomes, chromosome segments and characterize chromosomal rearrangements. Of interest in relation to this work are specific probes of repetitive DNA sequences to investigate centromeric or pericentromeric regions (Tsuchiya 2011). Secondly, labelling of the probe can be performed either by nonradioactive indirect labels or by direct fluorophore labels. In indirect labelling, hapten labels are applied in combination with antibody conjugates. Widely used and commercially available options for hapten labels include biotin, digoxigenin, and

fluorescein (FITC). These molecules are linked to nucleotides and incorporated in the probe by different techniques, including random primer labelling, nick translation and PCR-based amplification. Visualization then uses an antibody on the hapten, linked to commercially available fluorophores, which produce many colours under the fluorescence microscope by using appropriate excitation (Schwarzacher & Heslop-Harrison 2000; O'connor 2008). A further important step for this technique is the pre-treatment of slides, which is used both to decrease the background and improve the FISH probe penetration by removing surplus RNA and proteins, and also to fix chromosomes and nuclei, thus avoiding their loss from slides during post steps (Schwarzacher & Heslop-Harrison 2000). Following this, during the denaturation and hybridization processes, both the target DNA sequences of chromosome, and the labelled probe are exposed to heat source using a hybridization machine to denature them and enable them to become single strands. These are then cooled to allow the labelled probes to anneal to the target sequences of interest. The slides then undergo the post hybridization washes to remove weakly hybridized and unbound probes, thereby decreasing the background signals. During this process, it is important to control the elements of stringency, including temperature, salt concentration and duration of washing. This is because the presence of a high temperature (even for few seconds) has the potential to remove target DNA probes, while, background signals could increase if reduced stringency washes are employed (Schwarzacher & Heslop-Harrison 2000).

In order to visualize the probe target hybrids thus formed, a detection step is required that depends on the labelled probes used. Indirect labels (such as biotin and digoxigenin) need to be detected immunohistochemically by a fluorophore-tagged antibody (normally red or green fluorescing, respectively). However, direct fluorophore labels do not require this detection step. Finally, chromosomes are counterstained with suitable fluorochrome (such as DAPI (fluorescing blue)), mounted in antifade to prevent any fading of the fluorescence. The preparations are then analyzed, using an epifluroescent microscope with suitable filter sets to excite and visualize the different fluorochromes, enabling images to be captured and processed. The observed signals identify the loci of hybridized probes with target DNA sequences on chromosomes. Overall, the progress and powerful of *in situ* hybridization technique has revolutionized a broad spectrum of

experimental and diagnostic applications in many research areas such as cytogenetics, genomics, tumor biology and others (Haaf 2000; Schwarzacher & Heslop-Harrison 2000; Schwarzacher 2003; O'connor 2008; Chen & Chen 2013). Thus, FISH technique will be used in this study to characterize the genomic distribution and abundance of repetitive elements and ancestral mitochondrial sequences in sheep genome.

1.14 **Next Generation Sequencing**

Next Generation Sequencing is a general and comprehensive term used to state several miscellaneous modern sequencing technologies that enable outputting a huge amount of sequence data through sequencing thousands to millions of DNA molecules in a single run at the same time. Thus, this powerful platform has revolutionized the study of many fields including genomics, molecular biology and beyond (Mardis 2008; Metzker 2010; Hui 2012; Van Dijk *et al.* 2014).

Historically, over the past five decades, several technologies of DNA sequencing have been developed. In 1970, the first generation (classical sequencing) was established by Sanger and Coulson. This is commonly referred to as the Sanger sequencing method which performs enzymatic DNA sequencing that utilizes DNA polymerase (Sanger *et al.* 1977). In 1975, Sanger and Coulson launched another method of sequencing called (Plus and Minus) in addition to the sequence of bacteriophage (Sanger & Coulson 1975). Thereafter, Sanger *et al.* (1977), developed a more efficient DNA sequencing method called the “Chain termination method’’. They used either *in vitro* or *in vivo* to produce DNA templates to be sequenced. Nowadays, Sanger sequencing is still used broadly for routine applications of DNA sequencing and to authenticate the data of next generation sequencing. In 1977, another classical method of non-enzymatic DNA sequencing was developed by Maxam and Gilbert known as “Maxam and Gilbert sequencing’’. Although this method, compared to Sanger sequencing, requires less complicated preparation of samples for sequencing, its maximum reads are short (100bp). Furthermore, it includes the usage of hazardous chemicals for sequencing processes, which, in itself, is technically difficult (Maxam & Gilbert 1977). There are some drawbacks of the application of first generation sequencing. One of the main drawbacks is their low throughput. For instance, Sanger sequencing can read sequences of templates up to 1kbp in each run.

The second limitation is their high cost per base in comparison with modern technologies. Another limitation of classical sequencing is its difficulty to detect polymorphic sites that are present at low frequency due to the high levels of backgrounds (Morey *et al.* 2013).

NGS is also known as Second-generation sequencing-SGS. Due to the high cost and low throughput of classical Sanger sequencing, the first equipment for NGS were developed and became available in 2004 and their commercial tools were launched in 2005 (Schadt *et al.* 2010; Morey *et al.* 2013). These tools are able to sequence many DNA molecules in parallel which generates a huge amount of data output. For instance, more than 300 gigabases of throughput can be generated in a single run using Illumina's HiSeq 2000 instrument.

In recent years, many sequencing chemistries and platforms of second-generation sequencing have been developed such as 454 Pyrosequencing; sequencing-by-synthesis (Qiagen-intelligent bio-systems); sequencing-by-synthesis (Illumina), Danaher Motion Polonator, Sequencing-by-ligation (Applied Biosystems SOLiD); Ion-torrent semiconductor sequencing and DNA nanoball sequencing. Although different machines for next generation sequencing are available and can be characterized by various technical approaches, they all share some common features including sample preparation, sequencing machines and data output. However, the differences between several platforms of NGS are primarily seen in the technical details of the sequencing reactions (Metzker 2010; Morey *et al.* 2013; Van Dijk *et al.* 2014).

Third Generation Sequencing (TGS), another new group of technologies, are commencing to appear with the capability to sequence single molecules of DNA with no clonal amplifications needed before sequencing, thus preventing the introduction of PCR artefacts and requiring less sample manipulation as well (Levene *et al.* 2003). Although third generation sequencing includes sequencing-by-synthesis chemistries, their detection techniques are not based on the detection of chemical incorporation but rather are mainly constructed on the physical recognition of nucleotides in an original DNA strand (Pettersson *et al.* 2009). Furthermore, the sequencing reactions in the third-generation sequencing are not stopped for the steps of 'wash and scan' once each base

has been incorporated (Schadt *et al.* 2010). However, as these newer technologies generate a different kind of data; this will be the next challenge for bioinformatics analysis. Morey *et al.* (2013) reported in their mini-review six approaches of third generation sequencing including Nanopore sequencing, real time sequencing of single molecules and others.

The output per sequencing technology platforms such as Roche, LifeTechnologies, Illumina/solexa, Pacific Biosciences and Helicos are summarized in terms of read length, reads and output per run using either single or paired reads (Buermans & Den Dunnen 2014). Furthermore, Goodwin *et al.* (2016) classified and compared platforms of second and third generation sequencing in more details with respect to their read length, throughput, accuracy and cost. Each platform has its own advantages and disadvantages (Metzker 2010; Van Dijk *et al.* 2014; Goodwin *et al.* 2016). For instance, the Pacific Biosciences RS II instrument is widely used for generating longer read lengths, which is ideal for applications of *de novo* assembly (Schadt *et al.* 2010). However, the Pacific Biosciences have some limitations such as high rate of indel errors. Although several providers of NGS technologies are available, the Illumina instruments are increasingly being used in the research of next generation sequencing (Goodwin *et al.* 2016). In general, Schadt *et al.* (2010) compared the primary features of first, second and third generation sequencing to indicate their weaknesses and strengths.

Next generation sequencing technologies provide many important applications such as "whole genome resequencing" (Suzuki *et al.* 2011), "targeted resequencing'' (Hedges *et al.* 2011), "*de novo* sequencing" (Ghosh *et al.* 2011), "whole transcriptome analysis with gene expression analysis''(Alter *et al.* 2008), "small RNA sequencing'' (Schopman *et al.* 2011), "methylation analysis'' (Zeschnigk *et al.* 2009), "ChIP-Seq'' (Mokry *et al.* 2010). In recent years, these applications have been used to carry out studies in several fields including the diagnosis of genetic diseases; microbiological studies; mitochondrial genome studies; evolutionary and population studies, such as personal forensics/identification (Irwin *et al.* 2011). Therefore, in this study, the whole sequencing sheep genome (NGS) generated by Illumina approaches was utilized in order to assemble an entire mitochondrial genome and analyze their ancestral sequences.

Furthermore, NGS data alongside recent bioinformatics tools will also be used for investigation of repetitive landscapes in sheep genome.

1.15 Progress and status of sheep genome project

The main aims of sequencing the complete genome of sheep, *Ovis* aries, is to provide a complete and accurate sequence consisting of billions of base pairs that will constitute the sheep genome covering coding (genes) and non-coding DNA regions. Like other mammalian genomes, such as cattle and human, the assembly project of the complete sheep genome is crucial because it uses DNA as a source of genetic information to advance new methods of medical treatment and prevention of genetic diseases, to provide a better understanding for biological research, such as by characterizing the structure and function of genes. Recent bioinformatics tools require a complete reference genome to be used in their analysis such as comparing sheep genomes with others in order to study chromosomal rearrangements and karyotype evolution. Mapping of the sheep genome will have a significant impact on the rural economy by applying the research findings in order to guide farmers on how to improve the major sheep products of milk, meat and wool. Furthermore, assembly and better understanding of the genetic make-up will lead to the development of more effective management and breeding strategies so as to produce healthier breeds and more productive generations. Studying complete sheep genomes will reveal the consequences of domestication and selection (Bourque *et al.* 2005; Archibald *et al.* 2010; Rubin *et al.* 2010; Church *et al.* 2011; Dong *et al.* 2015; Fuentes-Pardo & Ruzzante 2017).

International Sheep Genomics Consortium (ISGC) (Archibald *et al.* 2010), including 26 institutions across eight countries, carried out several genetic studies in order to sequence the complete genome of domestic sheep. After eight years of work, researchers have completed the first sequencing of the entire sheep genome.

In the last few decades, several submitters started to sequence the *Ovis aries* genome. Table 1.2 demonstrates all submitters that focus on producing the reference sheep genome at both chromosomal and scaffold levels (https://www.ncbi.nlm.nih.gov/genome/83; http://www.sheephapmap.org/; https://www.ensembl.org/*Ovis* aries/Info/Annotation).

Table 1.2 Shows sheep genome representation of the latest version status of all assemblies starting from 2010 till 2017. So far, the version 4 is used as a representative RefSeq category of sheep genome. All assemblies were produced at chromosomal level except Oori1 of *Ovis aries* musimun from European Bioinformatics Institute (EBI) (https://www.ncbi.nlm.nih.gov/genome/83).

Organism	Name	Submitter	Date	Assembly level
***Ovis aries* (sheep)**	*Ovis*_aries_1.0 Synonyms: oviAri1	International Sheep Genomics Consortium	25/02/2010	Chromosome
***Ovis aries* (sheep)**	Oar_v3.1 Synonyms: oviAri3	International Sheep Genome Consortium	24/09/2012	Chromosome
***Ovis aries* musimon (mouflon)**	Oori1	EBI; European Bioinformatics Institute	15/07/2014	Scaffold
***Ovis aries* (sheep)**	Oar_v4.0	International Sheep Genome Consortium	20/11/2015	Chromosome
***Ovis aries* (sheep)**	Oar_rambouillet_v1.0	Baylor College of Medicine Human Genome Sequencing Center	02/11/2017	Chromosome

In 2002, the ISGC began informally with the construction of a high-quality BAC library. More recently, in 2015, they released the updated version of the genome assembly of sheep named Oar_v4.0 [assembly accession of GenBank (GCA_000298735.2)].

In this assembly, they used Illumina technology GAII, 454 and PacBio RSII for sequencing the DNA sampled from the Texel breed (single ram and single ewe) in which the assembly is built on ewe data while the ram data set is provided to fill in the gaps. They assembled the sheep genome at chromosomal level using assembly methods SOAPdenovo v. 1.03 and PBJelly2 v. 14.9.9. Up till now, they assembled all sheep chromosomes except the Y chromosome. The size, GC%, proteins, tRNA, other RNA, genes and pseudogenes of each chromosome are described in Appendix 1.2 (https://www.ncbi.nlm.nih.gov/genome?term=*Ovis*%20aries). As a result of sequencing the sheep genome, ISGC assembled 2.6 billion base pairs as a total sequence length (see section for comparison). The global statistics of the current sheep genome assemblies are summarized in Appendix 1.3.

Furthermore, the web based genome browser of the University of California Santa Cruz (UCSC) added the first (ISGC *Ovis*_aries_1.0/oviAri1) and third (ISGC Oar_v3.1/oviAri3) assembly of sheep which provide public access to genomic databases including the chromosomal sequences of sheep genomes (Rosenbloom *et al.* 2014).

The exploration of the landscape of repetitive DNA sequences, the abundance of *numts* and the infrastructure of the Y chromosome is still missing from the chromosome reference assemblies. Studies of the Y chromosome present a particular challenge in the assembling processes. Hence, this study will estimate the sheep genome size using an unbiased *k-mer* method based on calculating the frequency of short motifs. Additionally, this current work will also characterize the quantity and genomic organizations of the major dispersed and tandemly repeated sequences. It will also identify specific sequences for the Y chromosome to be used as molecular markers for the sheep genome, thus providing better insights into its structure which could be inserted in reference sheep genome. Overall, this work will provide some of the missing data in the reference assemblies, such as the repetitive landscapes of the sheep genome, in order to give a better understanding about its structure.

1.16 Aims and objectives

Aim: The overall aim of this thesis is to characterize the nature, variation, and genomic distribution of repetitive DNA families in the sheep genome with a focus on Kurdistani sheep breeds from the centre of diversity. The implications of the results for genome evolution, sequence diversification and homogenization will be considered. Molecular, cytogenetic and bioinformatics methods (including sequencing, PCR analysis, and fluorescent DNA *in situ* hybridization) are used to address the following objectives:

1- Sequence complete mitochondrial genomes from local breeds of Iraqi Kurdistani sheep using whole genome sequencing (NGS) data and investigate sequence phylogenies, variation and origins including the presence of nuclear-mitochondrial sequences (*numts*). How do mitochondrial sequences relate to the geographical diversity of sheep and what is the origin of *numts*? See Chapter 3.

2- Identify repetitive tandem DNA families in sheep using whole genome sequence data, and characterize their chromosomal location using fluorescent *in situ* hybridization. Are there any novel satellites or tandem repeats or is there any chromosome-specific repeat? And are there any specific tandem repeats in the sheep genome? See Chapter 4.

3- Investigate the meiotic behaviour of major satellite sequences and relate this to chromosome condensation and positions using immunostaining combined with *in situ* hybridization experiments. Are there any associations between repeats and SC (synaptonemal complex) or between SC themselves? See Chapter 4.

4- Identify repetitive dispersed DNA families in sheep whole-genome sequencing data and investigate their chromosomal distribution using fluorescent *in situ* hybridization (FISH). Are most repeats transposable element (TE) -related, and are TEs dispersed, localized or on specific chromosomes in the genome? See Chapter 5.

5- Assemble and investigate the phylogenetic relationship of complete genome of endogenous Jaagsiekte sheep retrovirus utilizing the whole genome sequences of local breeds of the Iraqi Kurdistan region. Identify all classes of endogenous retroviruses related elements. Do Kurdistani sheep breeds have integrated

complete virus, and are all ERV repeats dispersed or present in specific domains of chromosome? See Chapter 6.

6- Investigate genotypes and polymorphism of the ovine prion protein (PrP) gene in the local sheep breeds of Iraqi Kurdistan region. Do Kurdistani sheep breeds have prion disease resistance? See Chapter 7.

Together, the results will lead to understanding of the long-range organization of repetitive sequences and the consequences of variation and homogenization during the evolution of these DNA families in sheep.

Chapter 2 Materials and methods.

2.1 Materials

2.1.1 Blood

Sheep blood was obtained from Joseph Morris (Joseph Morris Butchers Ltd, Lutterworth, Leicestershire, UK), and collected from freshly slaughtered commercial sheep in sterile 15ml centrifuge tubes containing 150µl of 0.5M EDTA for DNA extraction and 150µl of 1.5-2 units/µl heparin for short-term culture, respectively. Extracted DNA was used for PCR amplification of repetitive DNA sequences. Blood samples were also collected with K2E Vacutainers (Becton Dickinson) from the jugular vein of sheep, including Hamdani, Karadi and Awassi breeds. Blood was sampled from flocks representing different locations of the Iraqi Kurdistan Region (Duhok, Erbil and Sulaymaniyah Governorates; see Appendix 2.1.

2.1.2 Standard Solutions

Table 2.1 Standard solutions prepared and used in experiments

Experiments	Solution	Constitutions
Gel electrophoresis	6x Gel loading dye	60% (v/v) glycerol (Fisher Scientific); 0.25% (w/v) bromophenol blue (Fisher Scientific); 0.25% (w/v) xylene cyanol FF (Fisher Scientific); Diluted to 1x in 50% (v/v) glycerol.
	50x TAE (**t**ris-**a**cetate-**E**DTA) buffer	2 M Tris-HCl (Sigma-Aldrich); 50 mM EDTA (**e**thylene**d**iamine**t**etra**a**cetic acid; pH 8; Sigma-Aldrich) 5.71% (v/v) glacial acetic acid (Acros Organics) Diluted to 1x in deionised H2O.
	Ethidium Bromide (10 mg/ml)	1g Ethidium bromide, 100ml of sterile distilled water. No autoclaving and stored at 4°C.
Cloning	LB (lysogeny broth) agar plates	2.5% LB broth (Melford); 1.5% agar (ForMedium); 100 µg/ml ampicillin (Sigma-Aldrich); 80 µg/ml x-gal (Sigma-Aldrich); 0.5 mM IPTG (**i**sopro**p**yl β-D-1-**t**hio**g**alactopyranoside; Sigma-Aldrich); pH 7.2
	LB solution	2.5% LB Broth (Melford); 40µg/ml ampicillin (Sigma-Aldrich); pH 7.2
	SOB (**s**uper **o**ptimal **b**roth) medium	2% tryptone (Oxoid); 0.5% yeast extract (Oxoid); 8.5 mM NaCl (Fisher Scientific); 2.5 mM KCl (Fisher Scientific) 100 mM MgCl2 (Fisher Scientific); pH 7
Colourimetric	Buffer 1	100 mM Tris-HCl (tris(hydroxymethyl)aminomethane), pH 7.5 (Sigma-Aldrich); 15 mM NaCl (Fisher Scientific)
	Buffer 2	0.5% (w/v) blocking reagent (Roche); prepared in buffer 1.
	Buffer 3	100 mM Tris-HCl, pH 9.5 (Sigma Aldrich); 100 mM NaCl (Fisher Scientific); 50 mM MgCl2 (Fisher Scientific)

FISH	20x SSC (**S**aline-**S**odium **C**itrate)	- 175.3g NaCl (3M) (Fisher Scientific), 88.2g Trisodium citrate $Na_3C_6H_5O_7$ (0.3M) (Fisher Scientific) per litre. The pH was adjusted to 7.0 by adding few drops 14N solution of HCl. The solution was autoclaved before using. Stock was diluted to make 2x SSC & 0.1x SSC.
	Detection buffer	4x SSC; 0.2% (v/v) Tween (Sigma Aldrich)
	20% SDS	2g, Sodium dodecyl sulfate (SDS) with 8ml water Not autoclaved
	Salmon sperm DNA	1µg/ml salmon sperm DNA (Sigma-Aldrich); Stock salmon sperm 10mg/ml diluted in TE buffer (0.5M EDTA;1M Tris); pH 8
	50% Dextran sulphate	50 gm Dextran sulfate with 100 ml distilled water, Filter sterilized and stored at -20°C.
	McIlvaine's buffer	0.1 M citric acid (Fisher Scientific); 0.2 M di-sodium hydrogen phosphate (Fisher Scientific); pH 7
	DAPI solution	Stock solution of 100µg/ml diluted in water. Final concentration of 4µg/ml was diluted with McIlvaine's buffer. Store at -20°C.
	Blocking solution	5% (w/v) BSA in 4xSSC with 0.2% (v/v) tween 20.

2.2 Methods

2.2.1 Genomic DNA extraction

Total genomic DNA was extracted from whole blood using the Wizard Genomic DNA Purification kit (Promega. Firstly, 900µl of Cell Lysis Solution was transferred to a sterile 1.5ml Eppendorf tubes. Then 300µl of blood added to each tube. After adding the blood, the tubes were inverted 5-6 times to mix blood with lysis solution. Afterwards, the mixture incubated for 10 minutes at room temperature (RT) and the tubes were inverted 2-3 times during the incubation time. The cell lysis solution was used to destroy the red blood cells. After incubation, the mixture was centrifuged at 13,000-16,000×g* for 1 minute at RT. The supernatant was discarded and the tubes were vigorously subjected to vortex for 10-15 seconds to resuspend the visible white pellet. Then, 300µl of the Nuclei Lysis Solution added to the tubes containing the resuspended cells and then pipetting the mixture 5-6 times to lyse the white blood cells. The mixture was incubated at 37°C for 10 minutes to make the mixture more viscous and to avoid the cells become clumps. To get better quality of genomic DNA, 1.5µl of RNase solution was added to the nuclear lysate, and mixed by inverting the tube 2-5 times and then incubated at 37°C for 15 minutes. After incubation, 100µl of Protein Precipitation Solution added to the nuclear lysate, and the mixture subjected to vortex vigorously for 20 seconds. In this step, small protein clumps could be visible. After centrifugation, the mixtures at 13,000-16,000×g* for 3 minutes, the supernatant was transferred to a clean 1.5ml Eppendorf

tubes containing 300µl of RT isopropanol. The solution was mixed gently by inversion until the white thread-like strands of DNA form a visible mass. To collect the genomic DNA, tubes centrifuged again at 13,000-16,000×g* for 1 minute at room temperature. The supernatant then discarded and 300µl of 70% RT ethanol was added to the DNA pellet. The tubes were gently inverted several times to wash the DNA. After centrifugation, the tubes at 13,000-16,000×g* for 1 minute, the ethanol was carefully aspirated using either a drawn Pasteur pipette or a sterilized pipette tip. The tubes containing the genomic DNA pellet were left at RT for 10-15 minutes to dry the pellet. Finally, 40-100µl of sigma water was added to each tube and stored overnight at room temperature or at 4°C. Then DNA samples were frozen at -20°C for longer time.

2.2.2 **PCR amplification**

Firstly, genomic DNA was diluted to 30-40 ng/µl in 300µl tube with sterile water (H_2O) (Sigma-Aldrich). Similarly, primers and the dNTP mix were also diluted. Primers and annealing temperature are designated in the corresponding result chapters. Different types of repetitive DNA elements (Tandem, dispersed and endogenous retroviruses related DNA repetitive elements) were amplified from sheep total genomic DNA. Each primer combination was tested on a Tprofessional Gradient Thermocycler (Biometra) with different annealing temperatures to optimize the amplification condition to observe specific band. PCR amplifications were set up either in 25µl or 50µl total volume reaction mixture containing ddH_2O (Sigma-Aldrich) (18.4µl or 36.8µl), 10x Buffer A (Kapa Biosystems) (2.5µl or 5µl), 10mM dNTP Mix (1µl or 2µl), 10 µM primers forward and reverse primers; Sigma-Aldrich (each 0.5µl or 1µl) and 5U/µL KAPA Taq DNA Polymerase (Kapa Biosystems) (0.1µl or 0.25µl). This master mix was added to 80-120ng of DNA samples. Then, the PCR cycling conditions followed the program consisting of 3 min initial denaturation at 95°C, followed by 35 cycles of denaturation (95°C, 0.5min), annealing (Tm-5°C, 0.5 min) and primer extension (72°C, 8min). The final cycle was the 1 min final extension at 72°C followed by indefinite hold time between 4-16 ◦C.

2.2.2.1 **PCR reaction of telomeric tandem repeats**

For telomeric repeats, a slight modification was used: the reaction was set up in 50µl total volume containing 38.7µl ddH_2O, 5µl 10 x Buffer A (Kapa Biosystems), 2µl 10mM

dNTP Mix, 2µl of each 10µM telomeric forward (TTAGGGTTAGGGTTAGGG) and reverse (CTAACCCTAACCCTAACC) primers and 0.3µl of 5U/µL KAPA Taq DNA Polymerase (Kapa Biosystems). The PCR cycling conditions used a 'touch down' programme and consisted of 3 minutes initial denaturation at 95°C, followed by 7 cycles of denaturation (94°C, 2 minutes), annealing (66°C, 0.5min decreased 1°C during each of the six consecutive cycles until it reaches 60°C) and primer extension (72°C, 0.75min). The final step was a 2 minutes extension at 72°C followed by indefinite hold time between 4-16°C.

2.2.2.2 Direct purification of PCR products

For direct purification, amplified PCR products were purified and other PCR constituents removed from the PCR reaction mixture using either the NucleoSpin® Gel and PCR Clean-up kit (MACHEREY-NAGEL) or E.Z.N.A.® Cycle Pure Kit (Omega Bio-tek) following the manufacturer's instructions of each kit.

2.2.2.3 Quantity and quality of genomic DNA and PCR products

The purified PCR products and genomic DNA were assessed spectrophotometrically using a Nanodrop® ND-2000 spectrophotometer (Thermo Scientific) at a wavelength of 260 and 280 nm. The spectrophotometer was first blanked using 1µl of ddH_2O (Sigma-Aldrich). 1µl of each PCR product or genomic DNA was then used for Nanodrop. The quantity (ng/µl) and purity ratios (A260/A280; A260/A230) of purified PCR products and genomic DNA were measured and results showed the OD260/280 nm value (1.80-2.2) regarded as a high quality. The purified PCR products and genomic DNA were then stored at -20°C till use.

2.2.3 Agarose gel electrophoresis

Amplified DNA fragments were separated by agarose gel electrophoresis. Using a microwave oven, the standard agarose gels 1% (w/v) were prepared by boiling agarose (Molecular Grade, Bioline) in 1x TAE (Table 2.1). The boiled gel was allowed to cool down before adding the ethidium bromide (final concentration of 0.5µg/ml) inside a fume hood. Gel combs were placed in sealed gel trays to make wells. Then, the agarose was poured into tray and left at RT to solidify. Then, the combs were removed and the gel placed in an electrophoresis tray ensuring the surface of the gel is covered with 1x TAE.

PCR products along with the proper quantity of loading buffer were loaded on the wells of the gel. Simultaneously, 5µl of DNA marking ladder (HyperLadder™ 1 kb (Bioline)) was loaded in either left or right well of the gel to assess the size of amplified PCR products. Then the gel was run for 60-90 minutes at 75-100 voltages. Finally, the gel was visualized with gel documentation system (Gene Flash (Syngene Bio imaging)) and the result of amplification and the size of bands were then determined. Agarose gel electrophoresis was also used to assess the molecular weight and quality of genomic DNA samples.

2.2.3.1 Excision and purification of amplified PCR products from gels

25-50µl of each PCR product was loaded along with 5-10µl of loading buffer into two juxtaposed wells onto a standard agarose gel 1% (w/v). PCR products were left running at 75 V for 60-90 minutes. Next, after all materials, including the surface of the UV apparatus, scalpels and forceps were sterilized using ethanol and flame. The targeted bands were excised into 1.5 Eppendorf tubes from the gel with the help of a UV transilluminator (UVP). Finally, the excised bands were washed and purified from the remaining agarose gel either with the NucleoSpin® Gel and PCR Clean-up kit (MACHEREY-NAGEL) or E.Z.N.A.® Gel Extraction Kit (Omega Bio-tek) following the manufacturer's instructions of each kit.

2.2.4 Cloning of PCR products

Purified PCR amplicons were cloned by insertion into pGEM®-T Easy vectors, following the manufacturer's protocol of pGEM®-T Easy Vector System I kit (Promega).

2.2.4.1 Ligation and transformation of competent *E. coli* cells

Ligation of purified PCR products into multiple cloning sites (MCS) of the plasmid vector was performed in a small 300µl tube containing 7µl of 2X Rapid Ligation Buffer [60mM Tris-HCL pH 7.8, 20mM MgCl2, 20mM DTT, 2mM ATP, 10% PEG (Promega)], 0.9µl of the pGEM®-T Easy Vector (50ng/µl), 1.2µl of T4 DNA Ligase (3 Weiss units/µl) and 5.4 µl of purified PCR products, made up to final volume 15µl with ddH_2O. After the ligation reactions being mixed were incubated at RT for 1 hr and then left at 4°C overnight. To obtain an appropriate amount of target DNA (insert) from cloning, Insert: Vector Molar

Ratios were optimized and calculated according to the following equation set up by Promega (http://www.promega.com/).

$$\frac{\text{ng of vector} \times \text{kb size of insert}}{\text{kb size of vector}} \times \text{insert:vector molar ratio} = \text{ng of insert}$$

Thereafter, transformation into competent E. coli cells was set up by adding 5µl of the aforementioned ligation mixtures to a new Eppendorf tube containing 50µl of the thawed competent E. coli cells (α -Select Bronze Efficiency, Bioline). Once the transformation tube being gently flicked, it was then incubated on ice for 20 minutes. Then, cells were subjected to a heat shock in a water bath for 45-60 secs at 42°C and returned immediately to ice for further 10 minutes. Next, 700µl of pre-warmed Super Optimal Broth media SOB medium were added to the tube of the transformants and Incubated for 3 hrs at 37°C in an orbital incubator with shaking 230rpm to promote the growth of transforming E. coli cells. Following the transformation process and the last incubation, 50µl, 100µl and 150µl of transforming E. coli cell cultures were plated out by spreading onto pre-prepared LB agar plates containing 100µg/ml ampicillin, 40µg/ml (X-Gal) and 500µM (IPTG). Finally, after the transformed cell cultures being left at room temperature for 20 minutes, the plates were incubated overnight at 37°C.

2.2.4.2 **Screening and Isolation of Recombinant clones**

Recombinant cells were identified by colour screening of blue-white colonies. Basically, recombinant clones interrupt the coding sequence of β-galactosidase and lack the active β-galactosidase activity thus white colour colonies being produced. In contrast, non-recombinant colonies characterized by functional β-galactosidase activity, leading to produce blue colonies because of breaking down the X-gal substrate by β- galactosidase enzyme. Only white colonies were picked with a sterile toothpick and subcultured in a tube containing 5ml LB medium. Finally, the culture was incubated at 37°C overnight with shaking 230rpm in an orbital shaker.

2.2.4.3 **Amplification of colony PCR products**

To determine whether inserts were successfully cloned or not and to confirm their sizes. PCR reaction mixture containing ddH_2O (31.2µl), 10 x Buffer A (5µl), 10mM dNTP Mix

(4µl), MgCl2 (3.3µl), 10µM custom primers forward and reverse primers; (each 3µl) and 5U/µL KAPA Taq DNA Polymerase (0.5µl) were prepared and added to the PCR tube containing 2-3µl of recombinant cultures. PCR cycling conditions were followed section 2.2.2. The amplified inserts were then visualized on a 1% (v/w) standard agarose gel, alongside HyperLadder™ 1kb following section 2.2.3.

2.2.4.4 **Isolation of plasmid DNA**

Successfully recombinant plasmid DNAs were purified from cell cultures using the NucleoSpin® Plasmid kit (Machery-Nagel), following the manufacturer's instructions.

2.2.5 **Labelling of probes using random priming**

The concentration of both purified PCR products and purified cloned products was measured in terms of ng/µl following section 2.2.2.3. Purified PCR products and clones were labelled with biotin-16-dUTP (Roche) or digoxigenin-11-dUTP (Roche) using the Bioprime® Array CGH random priming kit (Invitrogen). The reactions protocol was described by Schwarzacher & Heslop-Harrison (2000).

Biotin probes reactions were set up with the BioPrime® DNA Labelling System (Invitrogen). Approximately 400-800ng of purified products together with 2.5x random primer solution were denatured either using water bath or PCR machine at 95°C for 5 minutes and then left on ice for 5 minutes. After denaturation, 5µl of the 10X dNTP Mix and 1µl of Klenow Fragment of *E. coli* DNA polymerase I (40U) were added to the denatured DNA mixtures and mixed gently but thoroughly and incubated at 37°C for 2 hrs. After incubation, the Biotin probe reactions were finished by adding 5µl of Stop Buffer. Furthermore, reactions of Biotin probes were also performed with BioPrime® Array CGH Genomic Labelling Module (Invitrogen) following the same protocol of Digoxigenin probes with exception of using biotin-16-dUTP (1mM) (Roche) instead of digoxigenin-11-dUTP (Roche).

Digoxigenin probes reactions were performed with BioPrime® Array CGH Genomic Labelling Module (Invitrogen). In like Biotin probes, roughly 400-800ng of purified products together with 2.5X random primer solution were denatured either using water

bath or PCR machine at 95°C for 5 minutes and then left on ice for 5 minutes. After denaturation of the mixtures, on ice, 3µl of the 10X dUTP Nucleotide Mix, 1.8µl of digoxigenin 11-dUTP (1mM) (Roche) and 0.8µl of Exo-Klenow Fragment (40U) were added to the denatured DNA mixtures and mixed gently but thoroughly and spun down for 5-15 seconds. The reactions were incubated for 2 hrs at 37°C. The reactions were then stopped by addition 5µl of stop buffer (0.5 M EDTA pH 8.0) provided by the kit.

2.2.5.1 **Purification of labelled probes**

The labelled probes obtained from the above reactions were cleaned using either the NucleoSpin® Gel and PCR Clean-up kit (MACHEREY-NAGEL) or the BioPrime® Purification Module (Invitrogen), following the manufacturer's instructions of each kit. Then, the labelled probes were stored at -20°C, in order to stable for a longer time.

2.2.6 **Dot Blot Test (testing of labelled probes)**

The efficiency of labelled probes (biotin-16-dUTP or digoxigenin-11-dUTP) before their using in FISH and immunostaining experiments was tested by a colourimetric dot-blot test, following the protocol of Schwarzacher & Heslop-Harrison (2000).

Firstly, 1µl of probe samples was applied onto a bit of prewashed charged nylon membrane (Hybon- N+, Amersham Biosciences). Then, the membrane was soaked twice in buffer 1 and buffer 2 for 1 min and 30 minutes respectively. Antibody-AP mixture- 1:500 dilution of streptavidin conjugated to alkaline phosphatase (Roche) and 1:5000 dilution of anti-digoxigenin conjugated to alkaline phosphatase (Roche) in buffer 1 are used. After incubation at 37°C for 30 minutes, the membrane washed twice in buffer 1 and buffer 3 for 15 min and 2 min respectively. The incorporation and strength of probes were then detected by detection reagent (50mg/ml INT/BCIP (Roche) in Buffer 3. Finally, positively labelled probes appear brown with precipitate. For solutions see Table 2.1.

2.2.7 DNA sequencing

2.2.7.1 Sanger sequencing

Purified PCR and plasmid DNA amplicons were sequenced commercially using two different Sanger sequencing companies. The first company was *Source Bioscience*, 20µl (15-30ng/µl) of purified PCR or 20µl (50-80ng/µl) of plasmid DNAs with 20µl (1µM) of either forward or reverse primers were put into a separated Eppendorf tube. The second company was German sequencing (*GATC Biotech*), 5µl (15-30ng/µl) of purified PCR products or 5µl (50-80ng/µl) of plasmid DNA were premixed with 5µl (5µM) of either forward or reverse primers added together into an Eppendorf tube. After sequencing, all sequenced results were analyzed using Geneious 8.0 ((Kearse *et al.* 2012) http://www.geneious.com).

2.2.7.2 Whole genome sequencing (Next Generation Sequencing; NGS)

Five samples of genomic DNA were sequenced using the Illumina NextSeq500 mid-throughput 2x150bp cycle system with barcoded/multiplexed total DNA samples (in two runs at the University of Florida Interdisciplinary Centre for Biotechnology Research) giving 43 to 60 million reads (2-3X coverage of the sheep genome) with 5 to 6 Gb total sequence for each DNA sample. Samples Hamdani breed (HamM) and Karadi breed (KarM) were sequenced with six non-sheep samples in the first run, and the others (HamJ1; HamJ2 & KarJ) with four non-sheep samples in the second run Table 2.2.

Table 2.2 Total numbers, coverage and GC content of raw reads sequenced in each sample of sheep genome.

DNA samples (Male/Female)	Total number of sequenced reads per each sample	One-fold of sheep Genome /read length(3Gbp/150bp)	Number of fold per each sequenced genome (Coverage X)	GC% content
HamJ1(M)	52048068	20000000	2.60X	45.6
HamJ2(M)	56220882	20000000	2.81X	46
HamM(M)	43596654	20000000	2.18X	42.8
KarJ(F)	60605648	20000000	3.03X	45
KarM(M)	44933034	20000000	2.25X	43

2.2.8 **Phylogenetic analysis**

Firstly, MEGA6 (Tamura *et al.* 2013) was used to find the best model of mitochondrial and the complete enJSRV genomes alignments using the maximum likelihood criteria. Then, the Geneious software was used for alignment with default parameters and optimized manually. Bayesian phylogeny inference was used for analysis with MrBayes 3.2.6 (Huelsenbeck & Ronquist 2001) within the Geneious and largely default parameters based on the best substitution models identified by MEGA6; **G**eneral **T**ime **R**eversible (GTR) for mitochondrial genomes and **H**asegawa-**K**ishino-**Y**ano (HKY) for enJSRV genomes were selected. Furthermore, invariant gamma rate variation with four gamma categories, a burn-in length of 15,000 and chain length of 20,000 were applied.

2.2.9 **Lymphocyte culture preparation for sheep chromosomes**

Peripheral sheep blood was collected from freshly slaughtered commercial sheep (Joseph Morris Butchers Ltd, Lutterworth, Leicestershire, UK) in sterile 50ml tubes containing heparin. Lymphocyte short term medium contained 43.5ml of RPMI medium 1640 (1X) (Gibco), 6ml of foetal calf serum and 0.5ml of HyClone (Antibiotic antimycotic solution containing 10,000 U/ml penicillin G, 10,000µg/ml of streptomycin, and 25µg/ml of Amphotericin B). 0.5ml or 0.75ml of blood were added to 7ml medium containing 10-30 µg/ml phytohemagglutinin (PHA) in 5% CO2 incubator at 37°C for 3-5 days. Metaphases were arrested by adding 50-90µl of Demecolcine solutions [10µg/mL in HBSS, ACF Qualified, BioXtra (Sigma Aldrich)] and for further 1.5-2 hrs at 37°C. Metaphase chromosome preparations were then made as described by (Schwarzacher & Heslop-Harrison 2000) using hypotonic treatment with 0.075M KCl and fixation in absolute methanol to glacial acetic acid 3:1. For chromosomes dropping, few drops of fixed cell suspension were dropped from a different height of 10-30cm onto slides. Slides were left to air dry at room temperature overnight. Before the slides being stored at -20°C, the quality of chromosome spreads and their cell densities were checked under a phase contrast microscope.

2.2.10 **Cattle chromosome preparations**

Standard somatic chromosome preparations from short-term lymphocyte cultures of *Bos taurus* male Malaysian Brakmas (crossbred of Kedah-Kelantan (KK) and Brahman bulls) were provided by Dr Trude Schwarzacher (University of Leicester, United Kingdom, 2014). Slides had been made in 2004, after colcemid arrest and hypotonic treatment with 0.075M KCl by fixation and dropping in 100% methanol: glacial acetic acid 3:1. The slides were stored at -80°C until use.

2.2.11 **Fluorescent *in situ* Hybridization (FISH)**

Fluorescent *in situ* hybridization was carried out following the protocol generated by Schwarzacher and Heslop-Harrisson (Schwarzacher & Heslop-Harrison 2000).

2.2.11.1 **Pre-hybridization**

After the slides being scanned under the microscope, slides with well metaphase spread were selected and re-fixed by immersion in fresh fixative of absolute ethanol: glacial acetic acid in volume 3:1 for 30 minutes. Then, for dehydration, slides were rinsed in absolute ethanol 100% twice for 5 minutes each and then they were left to air-dry. Next, 200µl of RNase solution (100µg/ml) diluted in 2xSSC was applied to the marked area of each slide and covered with a large plastic slip and incubated at 37°C in a humid chamber for 1 hour. After incubation, the slides were washed twice with 2xSSC for 2 minutes and then for 10 minutes. Coverslips were removed carefully during the first wash. Slides were then re-fixed in 4% (w/v) paraformaldehyde at room temperature for 10 minutes under the fume hood. The slides were washed again in 2xSSC for 2 minutes and 10 minutes. Slides were then dehydrated in a series of 70%, 85% and 100% absolute ethanol for 2 minutes each. Slides were left to air-dry. Slides were re-scanned under the phase contrast microscope to assess the possible loss or damage to cells that could have happened during the previous treatments.

2.2.11.2 **Hybridization**

The hybridization mixture (34µl per slide) was prepared in an Eppendorf tube containing a different components with final concentration (50% (v/v) formamide, 20% (w/v)

dextran sulphate, 2x SSC, 0.025µg of salmon sperm DNA and 0.125%SDS (sodium dodecyl sulphate), 0.125mM EDTA (ethylenediamine-tetraacetic acid) as well as an appropriate amount of probes (30-120ng) was added.

Hybridization mixture including probe was denatured in a water bath or in PCR machine at 80°C for 10 minutes and immediately cooled on ice for a further 10 minutes. The hybridization mixture was applied to the marked areas of each slide covered with a small plastic cover slip, and then the slides placed in a thermal cycler (Hybaid Omniblock). In this cycler, slides heated to70°C for 7 minutes to denature both probes and chromosomal DNA together. Finally, the cycler was set to slowly cool to 37°C which is the hybridization temperature to allow probe and chromosomal DNA to anneal, and the slides were left on thermal cycler at 37°C overnight (16-20 hrs). The temperature of denaturation, the formamide concentration and Na+ ion amount in SSC limits the hybridization stringency. The salmon sperm DNA and blocking DNA reducing or removing the non-specific hybridization. The dextran sulfate used to increase the volume of mixture without decreasing the concentration of probe. SDS improves the penetration of probe while the EDTA stopovers nucleases (Schwarzacher and Heslop-Harrison 2000). The concentrations of salt and formamide permitted the sequences with homology 75-80% to form duplexes.

2.2.11.3 **Post hybridization washes**

After hybridization of slides overnight, the slides were subjected to post hybridization washes which including 2XSSC and 0.1XSSC washes. Firstly, slides were submerged into 2XSSC at 35-40°C to float off the coverslips and forceps were used to take out the coverslips. Post hybridization were includes either low or high stringency washes. Regarding the low stringency, slides were washed with 0.1x SSC twice, each for 5 minutes followed by another wash with 0.1x SSC for10 min. Then slides were washed with 2XSSC for 5 minutes and slides were then cooled to room temperature. For a high stringency, the slides were washed with 2x formamide (25%) followed by one wash with 0.1x SSC and then slide were cooled at room temperature in 2x SSC for 5 min. Post hybridization washes are essential for decreasing the background signal through removing weakly bounded or non-specific probes and other hybridization mixtures.

2.2.11.4 **Detection of hybridization sites**

One of important steps of *in situ* hybridization is the detection of hybridization sites which allows the visualization of probes. Therefore, after slides being left to cool at room temperature, slides were incubated in detection buffer (4X SSC, 0.2% (v/v) tween-20) for 5 minutes. Next, blocking solution (5% (w/v) BSA in 4xSSC with 0.2% (v/v) tween-20) was applied to slides under a plastic cover slip and incubated at 37°C for 30 minutes. This solution was used to block non-specific sites that could react with detection reagents. To detect the hybridization sites 50 - 60µl of 2µg/ml streptavidin conjugated (1 mg/ml stock, Sigma) to Alexa594 (Molecular probes) and 4µg/ml FITC anti-digoxigenin (fluorescein isothiocyanate, 200 mg/ml stock, Roche Diagnostics) were prepared in the blocking solution (5%BSA) and then added to slides under a plastic cover slip. Slides were then incubated in a humid chamber for 1 hrs, followed by two washes with the detection buffer at 42°C each 8 minutes to remove the extra antibodies.

2.2.11.5 **Counterstaining and mounting of slides**

After the slides being washed in detection buffer, they were incubated with 100µl of DAPI solution (4µg/ml DAPI) diluted in McIlvaine's buffer (0.1M citric acid, 0.2M di-sodium hydrogen phosphate) under a plastic cover slip at room temperature for 30 minutes in the dark. Slides were quickly rinsed in detection buffer and anti-fade solution (Citifluor, Agar Scientific) was added to the marked area of slides before a large glass cover slips (No. 0.24x40 mm) were placed on them. Glass cover slips were squashed smoothly, but thoroughly with a filter paper to remove the extra anti-fade. Finally, the slides were stored in the cold room at 4°C in the dark overnight. This permits the antifade solution to bind to the fluorophores which could stabilizes the fluorescence when the slides viewed under the microscope. Sometimes, slides were simultaneously counterstained and mounted by applying the mixture containing [Dapi; 100 µg/ml (6µl) + antifade (97µl) + ddH_2O (97µl)] onto the hybridized cells, and a large plastic coverslip was placed on top. Then, the slides were stored at 4°C in the dark cool room till use (more than 24 hrs.).

2.2.11.6 **Microscopy, photography and image processing**

The florescent *in situ* hybridization slides were scanned using a Nikon Eclipse N80i fluorescent microscope equipped with a DS-QiMc monochromatic camera (Nikon, Japan). Firstly, optical investigation was carried out to find chromosome metaphases using a DAPI filter to visualize the chromosomes. Each metaphase was captured in two different filter sets: blue excitation for FITC or green excitation for Alexa594 and UV excitation for DAPI. Images were falsely coloured (red for the probe and cyan for DAPI), overlaid and the contrast adjusted with NIS-Elements BR3.1 software (Nikon) using only cropping, and functions affecting the whole image equally. Furthermore, Adobe Photoshop CC was used with only functions of contrast and brightness adjustment that affect the whole area of the image equally.

2.2.12 **Immunocytochemistry and fluorescent *in situ* hybridization**

2.2.12.1 **Reagents**

Solutions and antibodies specific for immunostaining and FISH are given in Tables 2.3 & 2.4.

Table 2.3 Immunostaining and FISH solutions

Mammalian ringer solution	- Molecular Weight of (NaCl; 58.44g), (KCl; 74.56g), ($CaCl_2$; 111.02g) & (NaHCO3; 84.01g)
	- 9.2g NaCl, 0.4g KCl, 0.24g $CaCl_2$, 0.15g $NaHCO_3$ per 900ml distilled water. The solution was filled up to 1000ml & autoclaved before using.
0.2M or 0.4M Sucrose	- Molecular Weight of Sucrose; (342.3g)
	- 0.2M Sucrose: 6.85g Sucrose per 100 ml distilled water.
	- 0.4M Sucrose: 13.69g Sucrose per 100 ml distilled water.
	- The solution was not autoclaved before using.
0.6 % (v/v); 0.4 % (v/v); 0.2 % (v/v) Triton-*x* in 0.2M Sucrose solution	- 0.6 % (v/v) Triton-*x*: 0.2M Sucrose + 600 µl Triton-x per 100 ml
	- 0.4 % (v/v) Triton-*x*: 0.2M Sucrose + 400 µl Triton-x per 100 ml
	- 0.2 % (v/v) Triton-*x*: 0.2M Sucrose + 200 µl Triton-x per 100 ml
	- The solution was not autoclaved before using.
Formaldehyde fixation	- 4% formaldehyde solution (Fisher Scientific)
	- Molecular Weight of (NaH_2P0_4; 120g), (Na_2HPO_4; 141.96g)

10XPhosphate buffered Saline (PBS)	- 2.28g NaH_2PO_4 (0.038M), 11.5g Na_2HPO_4 (0.162M), 43.84g NaCl were dissolved in 450 ml distilled water & pH was adjusted to 7.4 by adding dilute1N HCl if necessary. Final volume was completed to 500 ml & autoclaved before use. Stock was diluted (1:10) to make 1x final solution prior to use.
0.4% (v/v) Photo-flo	- Kodak film wetting agent. 400 µl of 100 % photo-flo in 100 ml distilled Water
100 % dimethyl sulfoxide (DMSO)	- (Sigma-Aldrich)
Detection buffer (1x PBS+0.5 % Tween-20)	- 950ml of 1x PBS + 5 ml Tween-20.
Immunostaining Blocking solution	- 5 % (w/v) BSA (Bovine serum albumin) in (1x PBS+0.5 % Tween-20).
Acetic acid fixative	- 3:1 preparation for 100 % ethanol & glacial acetic acid respectively

Table 2.4 Primary and secondary antibodies and dilutions used.

Antibody Name	**Abbreviation**	**Company**	**Dilution used**
Anti- digoxigenin-fluorescein, Fab fragments (200µg/ml)	Anti-dig FITC	Roche Life Science (11207741910)	1:100 in 1xPBS/0.5%Tween-20 or 4xSSC/0.2%Tween
Streptavidin- Alexa Fluor® 594 conjugate (200µg/ml)	Streptavidin Alexa 594	Invitrogen/Molecular Probes (S11227)	1:200 in 1xPBS/0.5%Tween-20 or 4xSSC/0.2%Tween
Streptavidin- Alexa Fluor® 647 conjugate (200µg/ml)	Streptavidin Alexa 647	Invitrogen/Molecular Probes (S21374)	1: 50 in 1xPBS/0.5%Tween-20 or 4xSSC/0.2%Tween
Anti-SCP1 Polyclonal Antibody	SCP1	Abcam (ab15087)	1:120 in 1xPBS+0.5%Tween-20
Alexa Fluor® 594 $F(ab')_2$ fragment of goat anti-rabbit IgG (H+L)	Anti-rabbit IgG Alexa 594	Invitrogen/Molecular Probes (A11072)	1:150 in 1xPBS+0.5%Tween-20

2.2.12.2 **Materials**

Whole testes were collected from freshly slaughtered 6-8 months old ram lambs at Joseph Morris Butcher Ltd (South Kilworth, Leicestershire; UK). Testicles were stored at 4°C for a maximum of 2 days. For long-term storage, small amounts of testicular samples

were placed in cryo tubes containing 400µl dimethyl sulfoxide (DMSO) and 3.6ml of mammalian ringer solution. Then, samples were stored at -80°C until further use.

2.2.12.3 **Mammalian synaptonemal complex spreading**

Synaptonemal complex (SC) spreads were prepared following the protocol described by Schwarzacher *et al.* (1984) with few modifications. A few and small pieces of fresh or frozen testicular tissues (seminiferous tubules) were chopped with a razor blade or scalpel in a glass petri dish containing ringer solution to obtain meiotic cell suspensions. Then 10-30µl of serum containing released meiotic cells were added to 10-30µl of 0.2M Sucrose + 0.2 or 0.4 % (v/v) Triton-x onto a slide. To make different densities of cells, different volumes of serum and Triton were tested. Then, the cells were distributed over the slides to make a smear and left to dry on a 40°C hot plate and then overnight at room temperature. Slides were fixed by placing in a jar 100ml of 4% paraformaldehyde containing 3.6% (w/v) sucrose for 10 minutes at room temperature under the fume hood. Slides were rinsed with distilled water twice for two minutes each followed by a rinse in 0.4% (v/v) Photo-flo, and left to air dry. Finally, slides were checked under the phase contrast microscope for cell density and quality of the spread. Areas of well spread cells were marked with a diamond pen and then stored at -20°C up to several months until their use for immunostaining experiments.

2.2.12.4 **Fluorescent *in situ* hybridization combined with immunostaining**

Slides with synaptonemal complex spreads were subjected to FISH combined with immunostaining following the protocols modified from Schwarzacher and Heslop-Harrison (2000).

Probes labelling and *in situ* hybridization steps (pre-hybridization, hybridization and post-hybridization washes) were followed sections 2.2.5 and 2.2.11. Slides with SC spreads were then washed in PBS/Tween buffer for 5 minutes. 200µl of blocking solution was added to each slide, covered with large plastic cover slips and incubated at 37°C for 30 minutes. Afterwards, the cover slips were removed and 50-70µl of the primary antibodies (Table 2.4) were applied onto each slide, covered once again with large plastic cover slips, and incubated at 37°C for 3-4 hrs in a humid chamber. After

incubation, the slides were washed two times in 1xPBS/Tween buffer, for 5 minutes each. Then, antibodies for detection of FISH probe signals were combined with the secondary antibodies for immunostaining in different combinations. The first combination included anti-rabbit IgG Alexa 594 (red), Anti-dig FITC (green) and Streptavidin Alexa 647 (far red, and the second combination included anti-rabbit IgG Alexa 488 (green) and Streptavidin Alexa 594 (red). 60-70µl of the prepared secondary antibodies combination was applied to each slide. Slides were incubated at 37°C for 1 hour in a humid chamber. After incubation, slides were washed in 1x PBS/0.5% Tween-20 three times for 5 minutes each. Finally, slides were counterstained with DAPI and mounted as described in section 2.2.11.

2.2.12.5 **Scanning of FISH and FISH-Immunostaining slides**

FISH-Immunostaining slides were scanned using a Nikon 80i epifluorescent microscope. equipped with single band filters as follows: filter set 31023 (yellow excitation and far red emission) for streptavidin linked to Alexa 647; filter set 31002 (green excitation, red emission) for antibodies and streptavidin linked to Alexa 594; filter set 31001 (blue excitation and green emission) for digoxigenin conjugated to FITC, and secondary antibodies linked to Alexa 488 and filter set 31000v2 (UV excitation and cyan/blue emission) for DAPI Slides were scanned using 20x lense and DAPI excitation and selected cells captured using 100x lense and the appropriate filter sets depending on probe and antibody combinations in order of decreasing excitation wave length (far-red, red, green and finally DAPI). 5-15 cells were captured from each slide. Nikon N IS-Elements 4.0 imaging software was used to overlay images, and adjust the signal levels using only those functions that treat all pixel of the image.

2.2.13 **Identification and quantification of repetitive DNA landscapes**

Whole genome sequencing raw reads NGS (Table 2.2) were used to identify major families of tandemly, dispersed and endogenous retroviruses related DNA repetitive elements.

2.2.13.1 **Graph-based read clustering (RepeatExplorer)**

Graph-based approaches (http://www.repeatexplorer.org; (Novák *et al.* 2010; Novák *et al.* 2013)) were performed for the clustering analysis of whole genome sequencing raw reads. Parameters of RepeatExplorer were mostly set on default; minimum overlap length in nucleotides of similarity hits for clustering was at least 55% of the read length; over 82 bases. Both 0.01% and 0.001% were used as thresholds for building of linked reads. Mammalian databases were set for similarity hits of RepeatMasker. A schematic representation including components and analysis workflow of RepeatExplorer were described in section 1.7.1.

2.2.13.2 ***k*-mer frequency tool (Jellyfish)**

The program Jellyfish version 2 was used on an Ubuntu Linux computer to enumerate *k*-mers (short motifs) and their occurrence counts. Jellyfish is a command-line tool used for parallel reading and counting of occurrences of short motifs (*k*-mers) (a substring of length k) of DNA sequences within multi-FASTA files. Based upon the usage of different commands, Jellyfish produces *k*-mer accounts in the form of a binary format and then transforms its *k*-mer accounts into a text format which can be read by human (Marçais & Kingsford 2011). Three commands were used to numerate *k*-mers and their occurrence counts. The first command used for counting all *k*-mers of the whole paired end raw reads was (jellyfish count -m K -s 500M -t 8 -C), where a substring of DNA sequence length is (k); length of mer is (-m); hash size is (-s); number of threads is (-t) and canonical pairing of both strands refers to (-C). The second subcommand (jellyfish dump mer_counts.jf) was used to output all the counts for all the *k*-mer s presented inside the output file (mer_counts.jf) of the first command. Finally, grep subcommand (grep -a -A 1 --no-group-separator '>.\{N\}w' *k*-mer _counts_dumps.fa > *k*-mer sGTN.fa) and different thresholds of 10, 100, 1000, 10,000 presented in (N) were used against dump output file (mer_counts_dumps.fa) to generate final fasta file containing all counted short DNA sequences depending on the length (k) and the value of threshold selected in commands. The final output of *k*-mer counting was imported into the Geneious program version 8.0 (Kearse *et al.* 2012) (http://www.geneious.com) in order to generate longer contigs representing overlapping *k*-mers from short abundant repeat motifs. Thereafter, for identification of repetitive DNA landscapes, the consensus of

longer assembled contigs, representing large numbers of abundant, overlapping or similar contigs was blasted against available databases see section 2.2.14.2.

2.2.13.3 **Tandem Repeat Analyzer (TAREAN)**

TAREAN (TAndem **RE**peat **AN**alyzer) is a computational pipeline running under Galaxy server available via RepeatExplorer (https://repeatexplorer-elixir.cerit-sc.cz/; Novák *et al.* (2017). This tool is based on graph-based clustering principles mainly used for identification and characterization of genomic tandemly repeated sequences from unassembled raw reads. The tool explores NGS data for recognition of tandemly repeated sequences and reconstruction of their monomers presenting high and low confidence putative satellites.

Fasta file of whole sequencing paired raw reads from the male HamJ1 and female KarJ sheep genomes were subjected to TAREAN analysis using the workflow as described by (Novák *et al.* 2017) (see section 1.7.2). Input sequences were preprocessed using the utility 'Preprocessing of fastq paired reads' which includes trimming, quality and cut-adapt filtering and then interlacing while broken read pairs were excluded. Then, interlaced raw reads were then investigated in order to identify known and novel tandemly repeated DNA sequences. Recognition of satellite DNA sequences is performed on detection of circular constructions in the graphs of resulting clusters. In regard with the monomer reconstruction of tandem repeats, TAREAN build de Bruijn graphs utilizing the most frequent short motifs *k*-mer. Neither read assembly nor alignments are involved in this tool.

2.2.13.4 ***De novo* assembly**

The whole sequencing raw reads (NGS) of the five individuals were subjected to *de novo* assembly using assembler of Geneious software version 8.0 (Kearse *et al.* 2012) (http://www.geneious.com). Firstly, the bi-directional sequence reads were grouped together as paired end reads. After that, Geneious assembler was set to use 20% of the input NGS data and generates 100 contigs. Each contig was composed of many overlapping assembled reads with consensus representing the most frequent base calls.

The sequence consensus of each contig was then compared with databases of repetitive sequences mentioned in section 2.2.14.2.

2.2.13.5 **Reference mapping**

Using Geneious 8 software (Kearse *et al.* 2012) the 'Map to reference' was performed to compare and assemble the whole sequencing raw reads of NGS data to the sequences representing tandemly repeated monomers, dispersed repeats, endogenous retroviruses related repeats and to the complete genome of mitochondria and enJSRV. In all cases of map to reference, one contig, one consensus sequence representing the most frequent overlapped paired reads and one report were resulted. The total assembled raw reads presented in the report were then used to estimate copy numbers, genomic proportions and coverage of target sequences following two Mathematical equations;

1- **Copy numbers**= number of assembled reads*150(read length)/size of PCR product or reference sequence.
2- **Genomic proportion**= number of assembled reads/ total number of whole sequencing raw reads (Table 2.2)*100.

2.2.13.6 **Dot plots (graphic matrix)**

A dot plot inside the Geneious software version 8.0 (Kearse et al. 2012) (http://www.geneious.com) was used to compare two sequences against each other in order to find similar regions and also to determine whether a similarity between the two sequences is over-all or local. When a sequence is compared to itself, the dot plot shows regions of self-complementarity, direct repeats, and palindromic subsequences. However, when two different sequences are compared with each other, the dot plot draws regions of sequence similarity as a straight line, while regions with variant nucleotides will be drawn in the form of dashed lines (Church & Helfman 1993; Sonnhammer & Durbin 1995). Thus, a dot plot was used to detect tandemly repeated motifs which could be present in the consensuses of clusters or contigs of RepeatExplorer, *k*-mer frequency tool and *de novo* assemblies.

2.2.14 Bioinformatics approaches

2.2.14.1 Computational analysis using Geneious

Geneious is one of the powerful available bioinformatics software used to perform many valued functions. This versatile platform is characterized by many features such as its capability to combine molecular biology with computational science performing many purposes. Geneious software 8.0 ((Kearse *et al.* 2012) http://www.geneious.com) was used for various analysis and organization of data. These were including import and export of sequences with different format, pairwise and multiple alignments of DNA sequences, design and analysis of primers, building and viewing of phylogenetic trees, editing and analyzing of cloned sequences and other sequenced PCR products, viewing sequence logo, paired ends of whole sequencing raw reads, map referencing, *de novo* assembly, navigation of SNP and polymorphisms, annotation of genome and sequences, submission of genomes and DNA sequences to the GenBank databases.

2.2.14.2 Blast databases

NCBI databases were used to access and pick up the DNA sequences and genome assemblies. For Example, downloading available complete genome or different repetitive DNA families of the species among the Bovidae family. Secondly, program of sequence comparison (BLAST); Basic Local Alignment Search Tool (Altschul *et al.* 1990) was used to compare DNA sequences of proposed study with a library of sequences, and identify sequences that are highly like the query sequence of the same or different organisms. Similarly, BLAT (BLAST-like alignment tool) (Kent 2002) was also used to search and allocate DNA sequences of present study over chromosomes of *Ovis aries* (see section 1.15). Other databases such as (Repbase; (Jurka *et al.* 2005; Bao *et al.* 2015) (http://www.girinst.org/repbase/), (RepeatMasker; Smit *et al.* (2013-2015) http://www.repeatmasker.org) and (TEclass; (Abrusán *et al.* 2009) http://www.compgen.uni-muenster.de/teclass/index.hbi) were mainly used for identification of repetitive DNA landscapes. Furthermore, Tandem Repeats Finder (Benson 1999) (https://tandem.bu.edu/trf/trf.html) and dot plot (Self) (see section 2.2.13.6) of Geneious software were also used as an indicator of finding tandemly repeated DNA sequences.

2.2.15 **Estimation of genome size**

The Jellyfish *k*-mer count results were used for estimation of sheep genome size. Approximately 257404286 unassembled Illumina paired end raw reads (2X150bp) representing combined NGS data of five sequenced DNA samples (Table 2.2), which correspond to 38.59Gbp total base pairs were analyzed (Marçais & Kingsford 2011) (Figure 2.1). Different *k*-mer sizes (15, 16, 17, 18, 19, 20, 21 and 32 mer) were analyzed and estimated. The plot of frequency distribution of short motifs (*k*-mer frequency) allowed estimation of sheep genome size. For an applied example see section 5.5; Chapter Five.

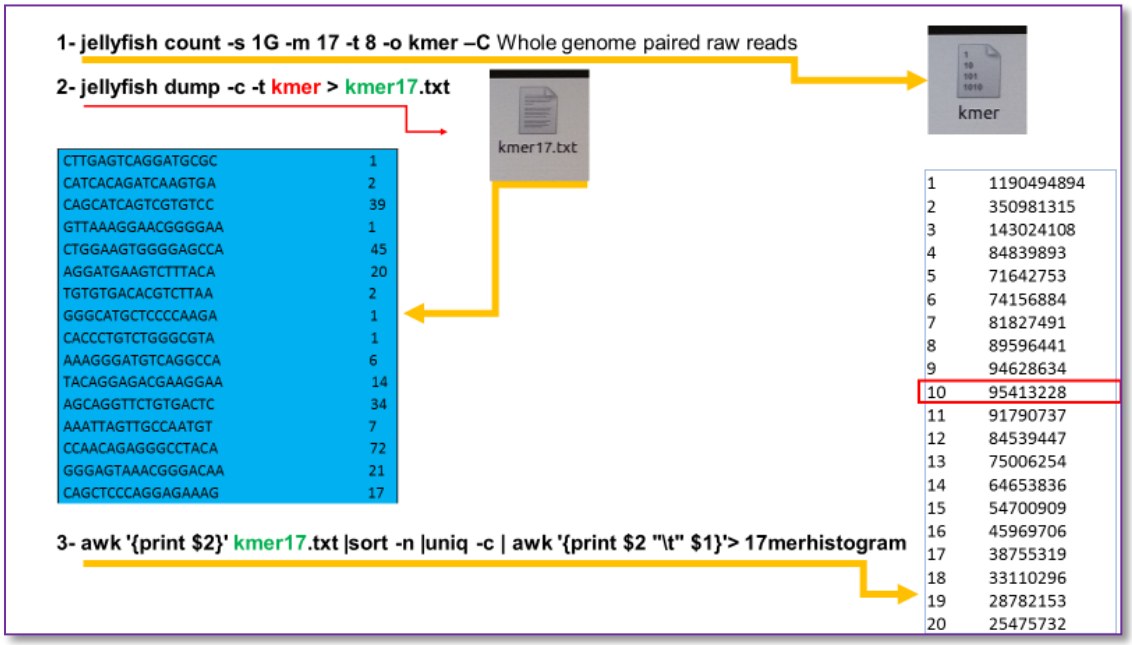


Figure 2.1 Shows three commands including the outcome of each one used to assess the sheep genome size.

Chapter 3 Mitogenomes in Kurdistani sheep: abundant centromeric nuclear copies representing diverse ancestors

3.1 Introduction

Sheep were among the first group of livestock species to be domesticated, with archaeological and genetic studies showing they were farmed 8000-11000 years BP in the Fertile Crescent, covering parts of Western Asia and north-east Iraq (Ryder 1983; Ryder 1984, 1991; Zeder 2008). The Kurdistan Region in the north of Iraq corresponds to the zone of initial domestication of sheep (Figure 3.1A) and still many native sheep breeds such as Hamdani, Karadi, and Awassi are kept by farmers, mainly for carpet-wool production, but the distinctive and morphologically-well-defined fat-tailed breeds (Figure 3.1B; 3.1C & Appendix 3.1) are also raised for meat and milk. Sheep breeding is one of the important sources of income for smallholder farmers in the region (Alkass & Juma 2005), and the landraces are well adapted to poor grazing habitats, showing hardiness in diverse harsh environments, and domestic sheep (*Ovis aries*) are in the Caprini tribe, Caprinae subfamily, of family Bovidae in the order Artiodactyla (even-toed ungulates). The phylogeography and classification of the wild *Ovis* species is extensively discussed (Rezaei *et al.* 2010), and includes the species or subspecies *O. orientalis, O. vignei, O. musimon* and *O. ammon* that occur in or close to the Kurdistan region, where the sheep sampled in this thesis were collected.

Domestic sheep are reported to derive from two subspecies (Ryder 1983), with approximately five independent domestication events giving rise to modern breeds, supported by archaeological and genetic evidence (Meadows *et al.* 2007; Zeder 2008). Like all domesticated animals, prerequisites for domestication include changes in a number of behavioural and physiological traits such as easy management (e.g. lack of aggressiveness and fight response), environmental tolerance (hardiness and disease resistance), and productivity (for sheep, initially meat and rapid reproduction, later including wool and milk). Given their role in energy metabolism, mitochondrial genes contribute to many adaptive and productivity characters, and mitochondrial genome

variation has been associated with phenotype, including hardiness, disease tolerance and resistance, milk production and fertility (Hiendleder *et al.* 2008) and see also MITOMAP.org in human. The definition of these specific diagnostic mutations has also improved phylogenetic information. Complete mitochondrial sequences provide a basis not only for identifying polymorphisms which may relate to energy metabolism, but also polymorphisms that are not selectively neutral and may lie outside regions chosen for genotyping studies using universal primers.

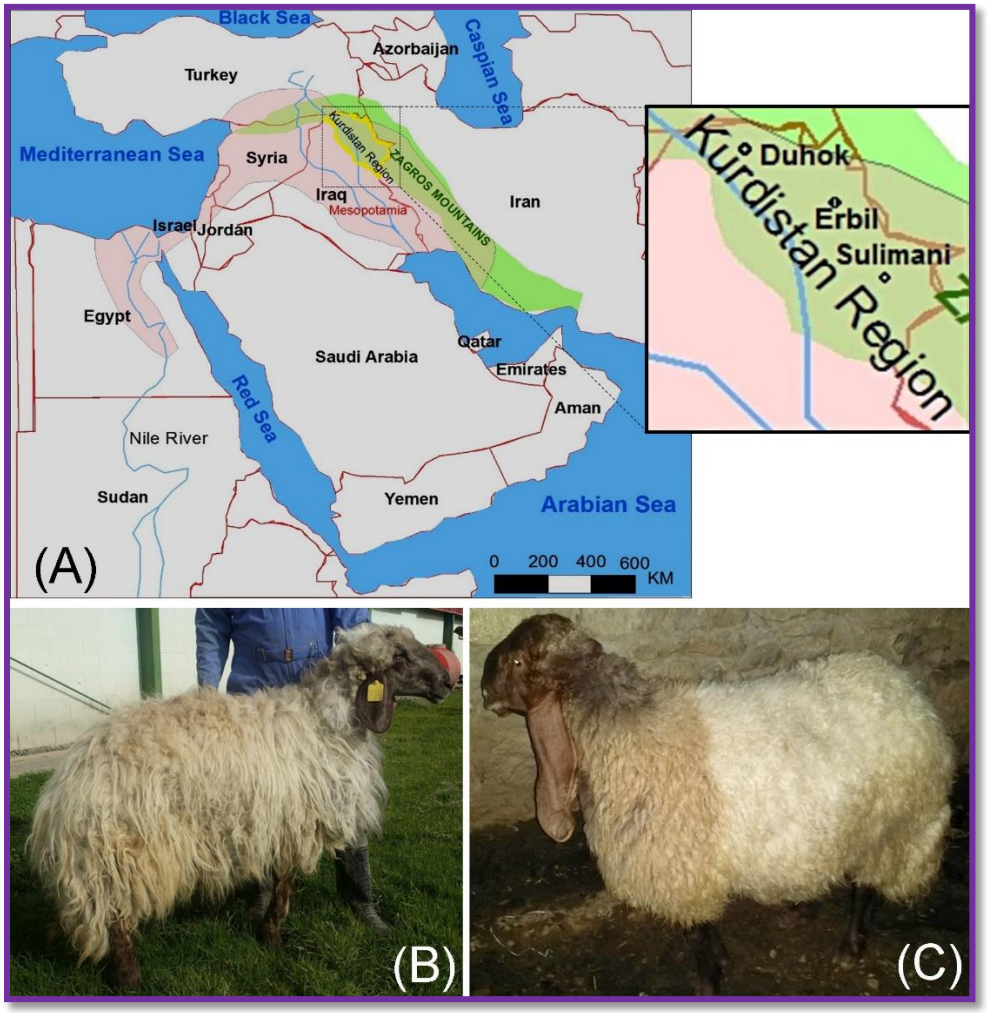


Figure 3.1 Locations of sampling and breed characteristics of Kurdistani sheep.

A. DNA was sampled from local sheep breeds in the Iraqi Kurdistan region (Duhok, Erbil and Sulaymaniyah governorates) in the 'Fertile Crescent' (pink shading) with high species and genotype diversity and the of many domestication events. B. and C. Breed characteristics of Kurdistani sheep include fat tail, long wool and long ears, Roman nose and specific coloration. Karadi sheep (B, showing ram K279-P) tend to have yellowish very coarse wool, black faces and long ears while Hamdani sheep (C, showing ram Hb4) are larger and have longer ears than Karadi sheep. Tails are almost reach the ground; their fleece is more whitish but often speckled. The Awassi breed (Appendix 3.1) are commonly white with red to brown faces and they are often horned.

The population and evolutionary biology of mitochondrial DNA (mtDNA) sequences have been extensively studied in many animals (e.g. bats, Dool *et al.* (2016). Nucleotide polymorphisms within mtDNA show diversity in maternal lineages and their phylogenetic relationship in different domesticated animals can be deduced or correlated with geographical locations or to resolve origins and ancestry (Kimura *et al.* 2010) in donkey and (Yang *et al.* 2017) in dogs. Due to their universal application, mitochondrial DNA markers are used for identification of species in food testing and archaeology. In sheep, Meadows *et al.* (2007) amplified fragments of the mitochondrial control region, tRNAPhe, and 12S rRNA while many authors (Zhao *et al.* 2011; Demirci *et al.* 2013; Mariotti *et al.* 2013) amplified D-loop sequences to identify mtDNA diversity and phylogenetic relationships. Altogether variation in mtDNA sequences has identified multiple maternal lineages by which the main haplogroups of mitogenome diversity have been classified (Hiendleder *et al.* 1998a; Pedrosa *et al.* 2005; Meadows *et al.* 2007; Meadows *et al.* 2011): the five major identified haplogroups, HPGA, HPGB, HPGC, HPGD and HPGE, are geographically wide-ranging, some being dominant and more specific to particular regions. HPGA and HPGB are most common and have been widely observed in Asia. In Europe, HPGB is considered the main maternal lineage while both HPGC and HPGE haplogroups and the less frequent HPGD have been described in Turkey, the Caucasus and China. Mitogenomes of the well-defined wild taxa (*O. ammon* and *O. vignei*) have also been sequenced (Meadows *et al.* 2007; Jiang & Ramachandran 2013).

It is notable that there has been minimal sampling of sheep from the south and eastern parts of the Middle East and specifically the Kurdistan region despite its central location in the Fertile Crescent (Lv *et al.* (2015) reporting a meta-analysis of sheep mitogenomes). Although the genetic diversity based on microsatellite and genetic markers of some sheep breeds of the Kurdistan region have been studied (Mohammed 2009; Al-Barzinji *et al.* 2011; Al-Barzinj & Ali 2013), their maternal diversity is unknown and no reports of genetic diversity of the fat-tailed sheep breeds nor their genetic significance and distinctiveness are available (Rocha *et al.* 2011). Many mitochondrial diversity studies have used PCR amplification of a limited number of genome regions, but reduced costs and the availability of next generation sequences allows complete mitochondrial genome sequences to be obtained that provide a reference to study all variation in a

species, identify haplogroups, and can be exploited to develop PCR-based markers to target polymorphisms informative at species, population and accession levels.

Mitochondrial genomes are normally considered to show strictly matrilineal or maternal inheritance. Polymorphisms may be found in sequence data because of either heteroplasmy – the occurrence of more than one mitochondrial variant (mtDNA sequence) – or the presence nuclear mitochondrial DNA segments, *numts*. Mitochondrial DNA is known to insert into the nuclear genome as *numts* (Vaughan *et al.* 1999), where they could visualize incorporation of mitochondrial sequences in the chromosomes of the nuclear genome by fluorescent *in situ* hybridization (Zhang & Hewitt 1996). Du and Qin (2015) identified more than 200 *numts* in the nuclear genome of the honeybee with the length of most *numts* less than 1kbp and identities of 75-90% to the mtDNA fragments. In mammals, Hazkani-Covo and Graur (2006) identified 452 and 469 mostly short *numts* in both human and chimpanzee respectively including 391 orthologous *numts* in both genomes. *Numts* were present in variable size and generally found to be highly fragmented, rearranged and distributed among and within nuclear genomes with different degrees of homology to their mtDNA sequence fragments (Zhang & Hewitt 1996; Woischnik & Moraes 2002). The abundance and mitochondrial coverage of *numts* in most species are still unknown.

3.2 Aims and objectives

The current study aimed to

1- Assemble the complete mtDNA genome of the two main Kurdistani sheep landraces, Hamdani and Karadi.
2- Analyze variants (SNPs) across the whole mitochondrial genomes and ascertain the maternal haplogroups of a larger panel of individuals including Awassi breeds using PCR-RFLP.
3- Identify and quantify the ancestral sequences (*numts*) of mitochondrial genome in NGS data, and characterize their presence and genomic location using DNA *in situ* hybridization.

3.3 Materials and methods

3.3.1 Assembly of complete mitochondrial genomes

Five samples of genomic DNA were sequenced (see section 2.2.7.2). For each sample, paired end reads were assembled against the complete mtDNA genome of Oxford Down *Ovis aries* (KF938359) as a reference using low stringency (maximum mismatches per read 10%). The five complete mitogenomes were then annotated using Geneious 8.0 (Kearse *et al.* 2012) (http://www.geneious.com). The five complete mitogenomes are available in GenBank under accession numbers (MF004242-6) see Appendix 3.2.

3.3.2 Data analysis and relationships

The five complete mitochondrial genomes of Hamdani and Karadi sheep breeds were aligned with published genomes from 10 domestic, 6 wild sheep species and 2 *Ovis musimon* species samples (see Appendix 3.2 for accessions and references) representing the five main sheep haplogroups, HPGA, HPGB, HPGC, HPGD and HPGE (Hiendleder *et al.* 2002; Meadows *et al.* 2011) and used to construct a Bayesian tree. Phylogenetic trees were built for the entire mitochondrial genomes following section 2.2.8. All analyses were carried out using the mitochondrial goat genome *Capra hircus* as the outgroup.

3.3.3 PCR-RFLPS (CAPS) for surveying mitochondrial sequence variation

After alignment between the consensus of the main haplogroups HPGA, HPGB, HPGC, HPGD and HPGE, there were polymorphic sites enable restriction enzymes to cut one haplogoup not the others. Thus, each of these polymorphisms occurring between Hamdani, Karadi and the reference mitochondrial haplogroup sequences was evaluated for generating a polymorphism in a restriction nuclease recognition site that would distinguish haplogroups.

Six primer pairs were designed for PCR-RFLP or [CAPS (Cleavage Amplification Polymorphisms); Table 3.1 to span polymorphic restriction sites, encompassing three different parts of the mtDNA genome, the ND1 gene (2850-3341nt), Cox1 gene (5437-6024nt) and CYTB gene (14786nt-15208nt. Genomic DNA from 5 sequenced and 26

additional sheep (Table 3.4) was amplified using the primers in 50µl (see section 2.2.2). Amplified fragments (2-3µl) were digested with one of eight restriction enzymes in the appropriate buffer (Table 3.3) and digested PCR products were separated by gel electrophoresis (2% w/v agarose) in 1x TAE buffer following section 2.2.3.

3.3.4 **Sanger sequencing of polymorphic regions**

Four primer pairs were designed for Sanger sequencing (Table 3.1) spanning variant related positions of the ND1 gene (2876-3552) and Cox1 gene with part of tRNA Ser (6370-6916). The amplified PCR products were spanning the same region showing heterogeneity in base calls Figure 3.4 and Appendix 3.18 at the same positions of polymorphisms.

Table 3.1 Primers used for PCR amplification of different regions of the mitochondrial genome for A.) Determination of haplogroups and heterozygosity, and B) resequencing by Sanger sequencing. C) Primers for FISH experiments

Primer name	Primer sequence (5'-3')	Primer Position (nt)	Region
A- For PCR-RFLP analysis (see details of fragments in Table 3.3			
2850F_OaMit	TCGAAAAGGCCCAAACGTTG	2850-2869	ND1 gene
3341R_OaMit	GGTGCTCGGTTTGTTTCTGC	3322-3341	ND1 gene
5437F_OaMit	AAGCCTACTAATTCGCGCCG	5437-5465	Cox1 gene
6024R_OaMit	ATAGGGTCTCCTCCTCCTGC	6005-6024	Cox1 gene
14786F HPGCE	ACCCCACAGGAATTCCATCG	14786-14805	CYTB gene
15208R HPGCE	TGTAGGGGTGTTCAACTGGC	15189-15208	CYTB gene
B- For amplification of selected gene regions to confirm SNPs by Sanger sequencing			
2876F.SPGA	GGCTTACTTCAACCCATCGC	2876-2895	ND1 gene
3574R.SPGA	GGATGCTCGGATTCATAGGAAGG	3574-3552	ND1 gene
6370F.CT.region	TTCTTTTCACAGTCGGAGGC	6370-6389	Cox1 gene/ tRNA (Ser)
6935R.CT.region	ATAGTGGCTATGGTGTTGGC	6935-6916)	Cox1 gene/ tRNA (Ser)
C- For amplification of selected gene regions for *in situ* hybridization (FISH)			
297 F(12s)	TGGTAAATCTCGTGCCAGCC	297-316	12S rRNA
931 R(12s)	TACTTGAGGAGGGTGACGGG	912-931	12S rRNA
1,335 F(16s)	TAACCCGAAACCAGACGAGC	1335-1354	16S rRNA
2,094 R(16s)	AGTAAAACCCTCGTGTGGCC	2075-2094	16S rRNA
2,848 F(nd1)	AAAAGGCCCAAACGTTGTAGG	2848-2869	ND1 gene
3,649 R(nd1)	GCATAGGGCTAGTGTTAGGGG	3629-3649	ND1 gene
4,048 F(nd2)	AAAAGCACAACCCACGAGCC	4048-4067	ND2 gene

4,771 R(nd2)	AGTGCTGTAATTGCTATGAGGG	4750-4771	ND2 gene
7,022 F(COX2)	TGGCATATCCCATACAACTAGGC	7022-7044	COX2 gene
7,620 R(COX2)	GCAAATTTCTGAGCATTGACCG	7599-7620	COX2 gene
14,166 F(CYTB)	AACATCCGAAAAACCCACCC	14166-14185	CYTB gene
15,201 R(CYTB)	TGTAGGGGTGTTCAACTGGC	15182-15201	CYTB gene
15570F(CR)	GAAAAGCACAACCACCCACC	15570-15589	Control region
16001R(CR)	TGTACGGTCAAGCAGTTTAATAT	15979-16001	Control region

3.3.5 Variant frequencies of mitochondrial genomes

After assembly, it was evident that some sites had a substantial proportion of reads with alternative bases to the consensus. Thus, the programme Geneious (Kearse *et al.* 2012) (8.0 http://www.geneious.com) was used to call variants across the whole mitochondrial genomes. Following the assembly to generate complete mitochondrial genomes, the raw reads from the complete mitochondrial genome were reassembled to the annotated genome. Disagreements (Variant frequencies) were configured by setting appropriate features such as minimum variant frequency of 0.005%, default sets of variant P-value, strand bias P-value and genetic code of mitochondrial vertebrates.

The variant frequencies were investigated separately in selected regions of mitochondrial genomes. Variant frequency, synonymous, non-synonymous, transition and transversions were calculated for within CDS and within non-CDS. Hence, SNPs within variant frequencies of (2.5%– 7.5%) were selected and analyzed to identify percentage of polymorphic reads within the assembled reads. While, the SNPs with variant frequencies below 2.5% and above 7.5% were discarded. Moreover, the top 10% of SNPs with extreme strand bias were also deleted. The SNP resulted were added to the reference sequence as an annotated track Table 3.5.

3.4 Results

3.4.1 The mitochondrial genome in Kurdistani sheep and relationships

Total genomic DNA samples of two Hamdani, one Karadi and one of each breed with some intermediate characters (Table 3.2; Figure 3.1 & Appendix 3.1) were sequenced. From each of the 5 to 6 Gb of paired end reads, the complete consensus mitochondrial genomes, HamJ1, HamJ2, HamM, KarM and KarJ (Figure 3.2 & Appendices 3.3-3.6) were extracted by mapping to a reference *Ovis aries* mtDNA genome (KF938359 from Oxford Down; Lv *et al.* (2015)). The total lengths of the consensus mitogenomes were 16617, 16618 or 16619bp. The sequencing gave coverage of 120-308 times and is equivalent to 56 to 105 mitochondrial genomes per nuclear genome; the coverage was on average 1.46 times greater in the female than male samples.

Table 3.2 Breed identity based on phenotypic appearance and mitochondrial genome assembly data of the five Kurdistan sheep using Illumina NextSeq500 of total genomic DNA.

Sample code	Phenotypic characteristic /sex	Mitogenome & accession no.	Mitogenome size (bp)	Maternal haplogroup	Assembled reads (no.)	Coverage
Hb4	Hamdani /M	HamJ1_ MF004243	16618	HPGA	20866	188
H115-P	Hamdani mixed with Karadi /M	HamJ2_ MF004242	16619	HPGA	25715	232
H369-P	Hamdani /M	HamM_ MF004244	16618	HPGA	13269	120
K279-P	Karadi /M	KarM_ MF004246	16617	HPGB	34117	308
5546	Karadi mixed with Awassi/F	KarJ_ MF004245	16617	HPGB	16905	153

A complement of 37 genes was found (Figure 3.2 & Appendices 3.3-3.6), consisting of 22 tRNA genes, 2 rRNA genes (12S rRNA and 16S rRNA), 13 protein-coding genes (CDS), and 1 control region (D-loop), and the GC content averaged 38.9%. Variants between the five mitogenomes of the Kurdistani sheep breeds were tabulated in accordance with the sequence variant descriptions recommended by the HGVS nomenclature (Dunnen *et al.* 2016) generated by Mutalyzer (Appendix 3.7). An additional tandem repeat which is found in the wild species was not present in the Kurdistani mitochondrial genomes.

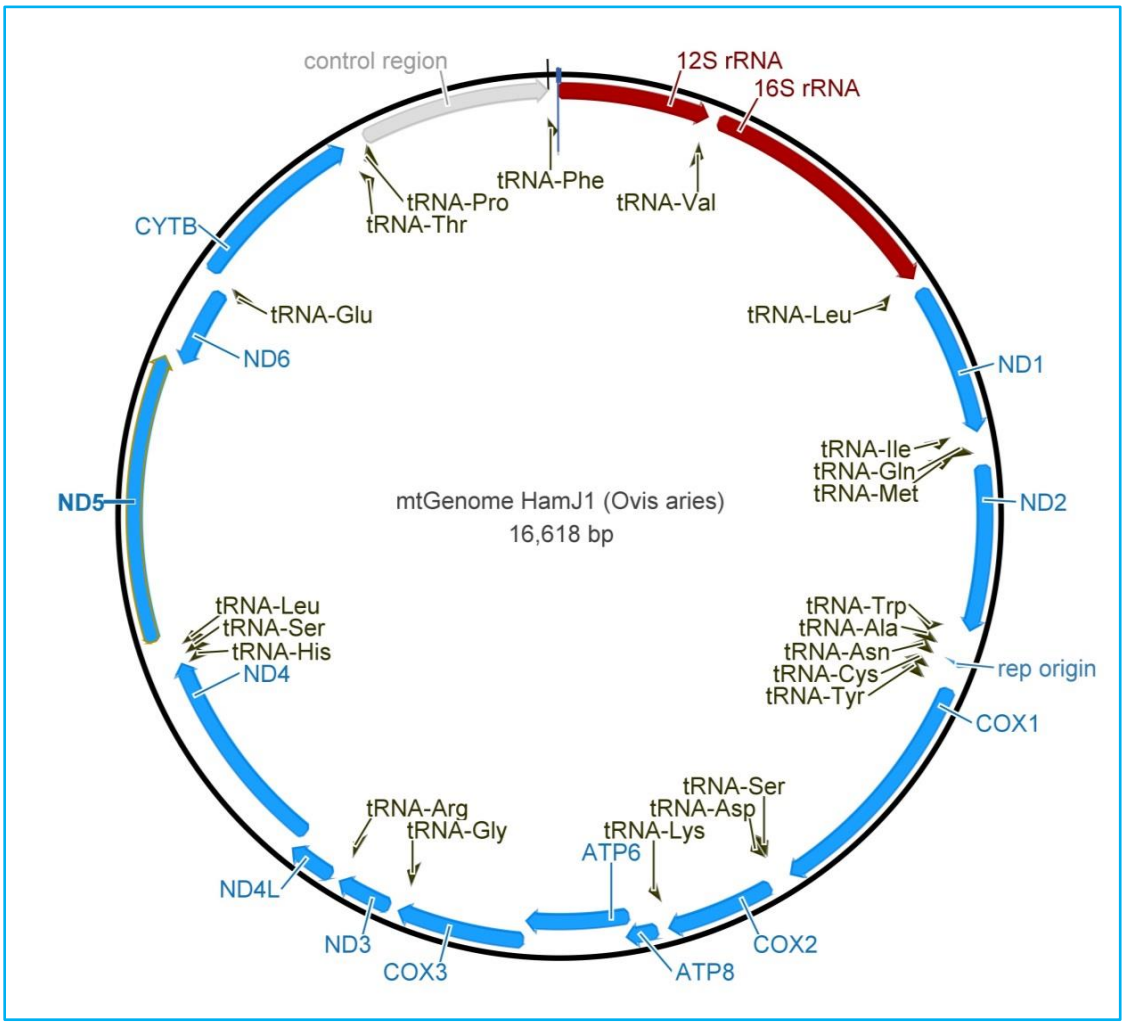


Figure 3.2 **Kurdistani sheep mitogenome map.** The assembled mitogenome HamJ1 (16,617bp) of *Ovis aries* Hamdani landrace animal Hb4, GenBank accession number (MF004243) with major features: there are 13 protein-coding genes (light blue bars, with the arrow pointing in the transcription directions), 22 tRNA genes (black triangles), the 12S and 16S rRNA genes (dark red) and the D-loop control region (grey). Equivalent maps of the four other mitogenomes assembled are shown in appendices 3.3-3.6.

The phylogenetic position of the five assembled Kurdistani mitochondrial genomes was established by Bayesian tree analysis including as reference genomes published haplogroups of domestic sheep, wild sheep *O. musimon, O. vignei*, *O. ammon*, *O. canadensis*, and as an outgroup, goat, *Capra hircus* (Hassanin *et al.* 2010) (Figure 3.3; Appendix 3.2). The consensus mitochondrial sequences from Kurdistan were placed on branches with the recognized sheep haplogroup HPGA (three animals of Hamdani breed) and HPGB (two animals of Karadi breed), while the other three known haplogroups HPGC, HPGD and HPGE (Meadows *et al.* 2011) were on separate branches. Two of the reference samples of *O. musimon* (Mouflon - HM236184, HM236185) were sisters to the haplogroup HPGB within the same subclade as the Karadi mitochondria.

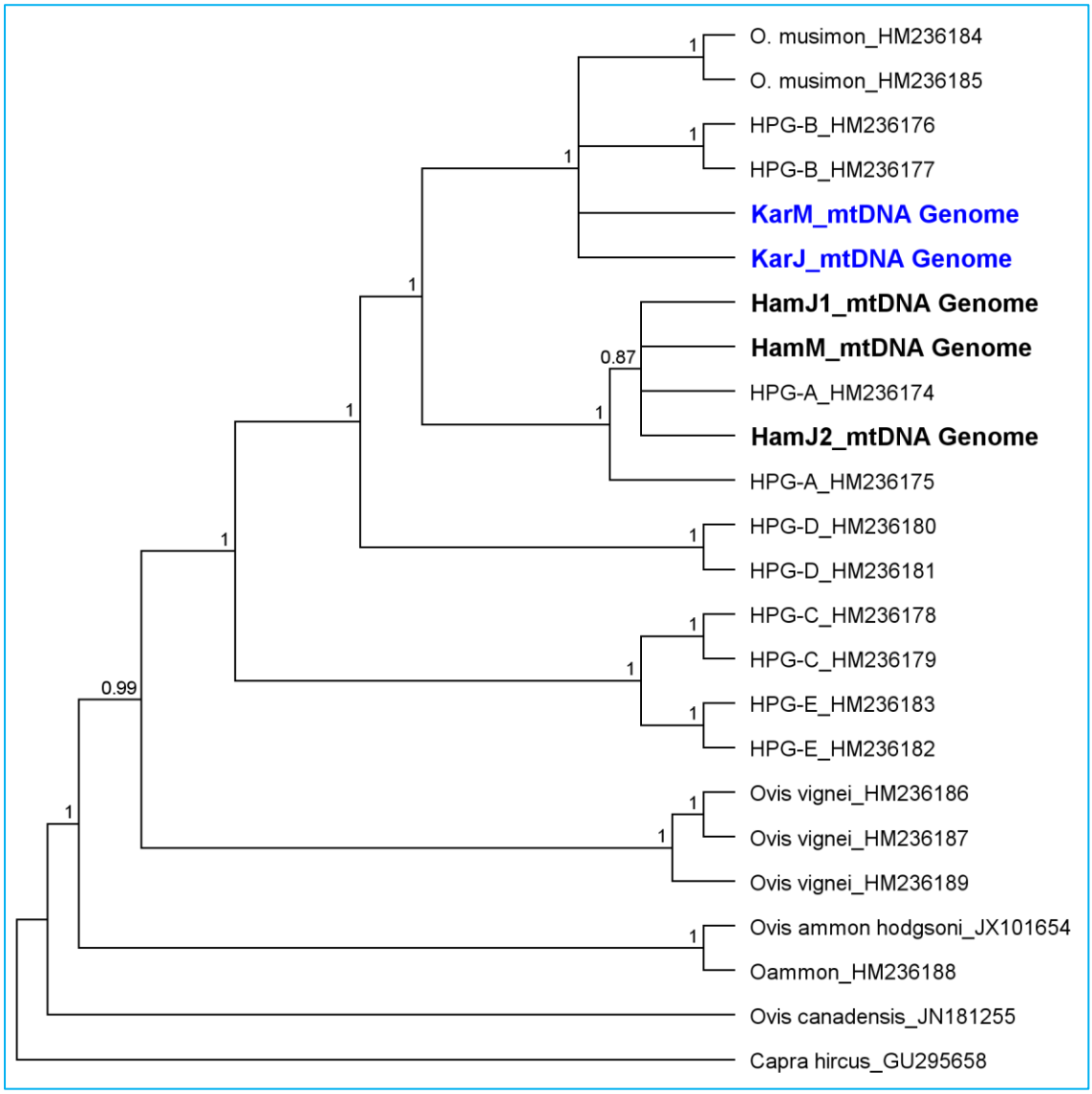


Figure 3.3 **Phylogenetic relationship of sheep mitogenomes.** Bayesian tree showing positions of Hamdani and Karadi mitogenomes in relation to major haplogroups of domestic *O. aries*, two of *O. musimon,* and six wild sheep (*O. ammon, O. vignei* and *Ovis canadensis*). The tree was derived from Bayesian (MrBayes) analysis of alignments of the whole mitochondrial genome sequences. Nodes are labelled with posterior probabilities. The mitochondrial genome of *Capra hircus* used as outgroup. Appendix 3.2 gives accession numbers and references for sequences included in the tree.

Based on the consensus mitochondrial sequences, primers were designed to amplify fragments of the genome encompassing identified polymorphisms in parts of ND1, Cox1 and CYTB genes (Table 3.1). PCR-RFLPs (CAPs) (Table 3.3; Figure 3.4; Appendix 3.18 & Appendices 3.8-3.16) were then used to identify the major mitochondrial haplogroup in another 26 male and female Hamdani, Karadi and Awassi sheep making it a total of 31 sheep analysed in this study (Table 3.4). About half were HPGA with the remaining divided equally between HPGB and HPGC. As further confirmation of the presence of the three haplogroups, a primer pair was designed to amplify a second part of the ND1 gene

and PCR products were sequenced and confirmed HPGA (H-390-P, H1a, 1Aw), HPGB (K279-P, H-368-P) and HPGC (3K 00454, K5-SUL, H-364-P).

Table 3.3 Primer combinations, restriction enzyme cleavage sites and expected products differentiating mitochondrial haplogroups (gel electrophoresis images see Appendices 3.8-3.16)

Primers Forward & Reverse **(expected PCR products)**	Target part of mtDNA genome	Restriction enzymes	Cleavage site of Restriction Enzymes on mtDNA genome (nt)	Cut Haplogroups (HPG)		Uncut Haplogroups (HPG)
				Estimated size of restriction fragments (bp)		Estimated size of restriction fragments (bp)
2850F_OaMit 3341R_OaMit **(492bp)**	ND1 gene (2850-3341nt)	*Bam*HI	3224	HPGA/HPGC/HPGD/HPGE		HPGB
				379+113		492
		*Ava*II	3225	HPGB	HPGC/HPGD/HPGE	HPGA
				379+113	465+27	492
		*Alu*I	2972	HPGA		HPGB/HPGC/HPGD/HPGE
				368+124		492
		*Mse*I	2910	HPGA/HPGB/HPGD	HPGC/HPGE	-------
				273+63+67+89	272+71+150	-------
5437F_OaMit 6024R_OaMit **(588bp)**	Cox1 gene (5437- 6024nt)	*Bgl*II	5768	HPGA/HPGB/HPGD		HPGC/HPGE
				331+257		588
		*Bst*NI	5793	HPGA/HPGC/HPGD/HPGE		HPGB
				356+232		588
		*Hinf*I	5569	HPGB/HPGC/HPGD/HPGE		HPGA
				452+136		588
		*Acc*I	5767	HPGC/HPGE	HPGA/HPGB/HPGD	-------
				285+257+45	303+285	-------
14786F 15208R **(423bp)**	CYTB Gene (14786-15208nt)	*Alu*I	15155	HPGC		HPGE
				369+54		423

Table 3.4 Maternal haplogroups (A) and their frequencies (B) of 31 Kurdistani sheep determined by genomic NGS data and genotyping by PCR-RFLP.

A

Breed[a)]	**Sample location** [b)]	**Sample code/ Mitogenome c)**	**Sex**	**Haplogroup** [d)]
Hamdani	Duhok	H-364-P	F	HPGC
Hamdani	**Erbil**	**H-369-P/HamM** [c)]	**M**	**HPGA**
Hamdani	Duhok	H-368-P	F	HPGB
Hamdani	Duhok	H-390-P	F	HPGA
Hamdani	Duhok	H-374-P	F	HPGB
Hamdani	Duhok	H1a	M	HPGA
Hamdani	Erbil	Hb2-a	F	HPGC
Hamdani	Erbil	Hb1-B	F	HPGA
Hamdani	**Duhok**	**H115-P/HamJ2** [c)]	**M**	**HPGA**
Hamdani	**Erbil**	**Hb4/HamJ1** [c)]	**M**	**HPGA**
Karadi	**Erbil**	**K279-P/KarM** [c)]	**M**	**HPGB**
Karadi	Duhok	K972-P	F	HPGA
Karadi	Duhok	K970-P	F	HPGB
Karadi	Duhok	K680-P	F	HPGC
Karadi	Duhok	K688-P	F	HPGC
Karadi	Duhok	K1a	M	HPGB
Karadi	Sulaymaniyah	K5-SUL	M	HPGC
Karadi	Sulaymaniyah	K6-SUL	M	HPGC
Karadi	Sulaymaniyah	K7-SUL	M	HPGB
Karadi	Sulaymaniyah	K8-SUL	M	HPGA
Karadi	Duhok	KB5-B	F	HPGA
Karadi	Duhok	KB3	F	HPGC
Karadi	Duhok	1K	M	HPGA
Karadi	Duhok	2K-5350	F	HPGB
Karadi	Duhok	3K-00454	F	HPGC
Karadi	Duhok	4K-5530	F	HPGA
Karadi	**Duhok**	**5546/KarJ** [c)]	**F**	**HPGB**
Awassi	Duhok	1Aw	F	HPGA
Awassi	Duhok	2Aw	F	HPGA
Awassi	Duhok	4Aw	F	HPGB
Awassi	Duhok	5Aw	F	HPGA

B

		Haplogroup[d)]				
Breed[a)]	**Sample size**	**HPGA**	**HPGB**	**HPGC**	**HPGD**	**HPGE**
Hamdani	9	5	2	2	0	0
Karadi	18	6	6	6	0	0
Awassi	4	3	1	0	0	0
Total	31	14	9	8	0	0
Frequency %		**45.2**	**29**	**25.8**	**0**	**0**

a) Based on predominant phenotypic characteristics.

b) Coordinates for Duhok Governorate are 36.8679N, 42.9488W; Erbil Governorate 36.1911N, 44.0091W; Sulaymaniyah Governorate (35.5641N, 45.3756W)

c) Mitogenomes were assembled from 5 target sheep DNA samples; see Table 3.2 and section 2.2.7.2.

d) Haplogroups as defined by Meadows et al (2011), determined by genomic NGS data [in bold, see c)] or PCR-RFLP (see Table 3.3 and Appendices 3.8-3.16)

3.4.2 Nuclear-mitochondrial DNA sequences (*numts*)

3.4.2.1 Presence of variant mitochondrial sequences

The most frequent base in the overlapping reads had been used as the consensus for each mitochondrial genome. However, multiple sites with variant base calls were found in the sequencing reads with each animal having two or more bases (>1%) represented at some positions of the sequence reads mapped to the consensus assembly. In order to understand these variations and to distinguish between presence of heteroplasmy or *numts*, as well as excluding sequencing and assembly errors, the types and frequencies of SNPs were analyzed for the mitogenome HamJ1 (see section 3.4.2.2) and selected variant regions were amplified by PCR followed by Sanger sequencing; mitogenome sequences were searched in whole nuclear genome assemblies of sheep (Figure 3.5) and used as probes for fluorescent *in situ* hybridization to mitotic chromosomes.

3.4.2.2 Characterization of variant mitochondrial regions

Within the assembly of HamJ1, a total of 394 SNPs was present in multiple, but relatively small proportion of reads (between 2.5% and 7.5% of the reads), including 262

synonymous and 23 non-synonymous SNPs within coding sequences (Table 3.5). Most (363/394) were transitions while transversions were rare. Polymorphisms were present but more limited inside the control region, with 10 SNPs. Some 31 SNPs were found in each of the 12S and 16S rRNA, and 37 SNPs were found in all tRNA regions.

Figure 3.4 gives an example where fragments of 17 of 436 raw read sequences mapped to a 151bp region of the ND1 gene and included multiple variants. Another example is shown in Appendix 3.18. PCR primers were designed to span selected polymorphic regions (Table 3.1) and DNA was amplified from the animals used for DNA sequencing and the other individuals. The PCR products were sequenced by Sanger sequencing; polymorphic base calls (stacked peaks from two bases) were reported only at same locations as in the Illumina shotgun sequence reads (Figure 3.4 & Appendix 3.18) indicating that NG sequencing and assembly errors can be excluded.

3.4.2.3 **Nature of variant mitochondrial regions**

To assign variant reads to mitochondrial or nuclear genomes, raw reads of regions identified to contain SNPs were assembled against appropriate regions or whole consensus mitochondrial genome sequences at low stringency, allowing up to c. 10% mismatches. All the reads were then extracted and re-assembled at high stringency against the consensus mitochondrial genome. Unassembled polymorphic reads (variant 150bp reads; different by more than 1% from the consensus) were extracted and used to make *de novo* assemblies. They were then compared with the GenBank database, and the highest similarity was found to mitogenomes of several species including *O. canadensis*, *O. ammon, O. vignei* and genus *Capra*, and some sequences reported as nuclear including regions of *O. canadensis* chromosome 26 and *O. aries* chromosome X. The detailed results of these comparisons, including coding and control regions and the complete genome are shown in Appendix 3.17. After finding the similarity with *O. ammon* and *O. vignei*, the raw reads from the Kurdistani sheep were assembled to the mitochondrial genomes of the wild *Ovis* species; 1000 to 2000 reads showed 100% similarity.

Table 3.5 Total SNP frequency, synonymous and non-synonymous, transitions and transversions over the whole mitogenome of HamJ1.

Mitochondrial regions/Total length (bp)	**Total SNPs**	**frequency of SNP**	**Synonymous**	**Non synonymous**	**% synonymous**	**Transitions**	**Transversions**	**Tran/Trav**
ND1/957	28	2.90%	26	2	92.90%	28	0	0
ND2/1044	41	3.90%	37	4	90.20%	39	2	0.05
COX1/1545	42	2.70%	41	1	97.60%	41	1	0.02
COX2/684	24	3.50%	24	0	100.00%	23	1	0.04
ATP8/201	7	3.50%	6	1	85.70%	7	0	0
ATP6/681	11	1.60%	11	0	100.00%	10	1	0.1
COX3/804	23	2.90%	22	1	95.70%	21	2	0.1
ND3/357	14	3.90%	14	0	100.00%	13	1	0.08
NDL4/297	7	2.40%	7	0	100.00%	7	0	0
ND4/1395	13	0.90%	12	1	92.30%	12	1	0.08
ND5/1821	24	1.30%	19	5	79.20%	23	1	0.04
ND6/528	11	2.10%	8	3	72.70%	11	0	0
CYTB/1140	40	3.50%	35	5	87.50%	36	4	0.11
Total CDS/11454	**285**	**2.50%**	**262**	**23**	**91.90%**	**271**	**14**	**0.05**
rRNA (12S)/959	31	3.20%				31	0	0
rRNA (16S)/1575	31	2.00%				23	8	0.35
total rRNA/2534	62	2.40%				54	8	0.15
All tRNA/1514	37	2.40%				30	7	0.23
Control region/1181	10	0.80%				8	2	0.25
Whole genome/16618	**394**	**2.40%**	**262**	**23**		**363**	**31**	**0.09**

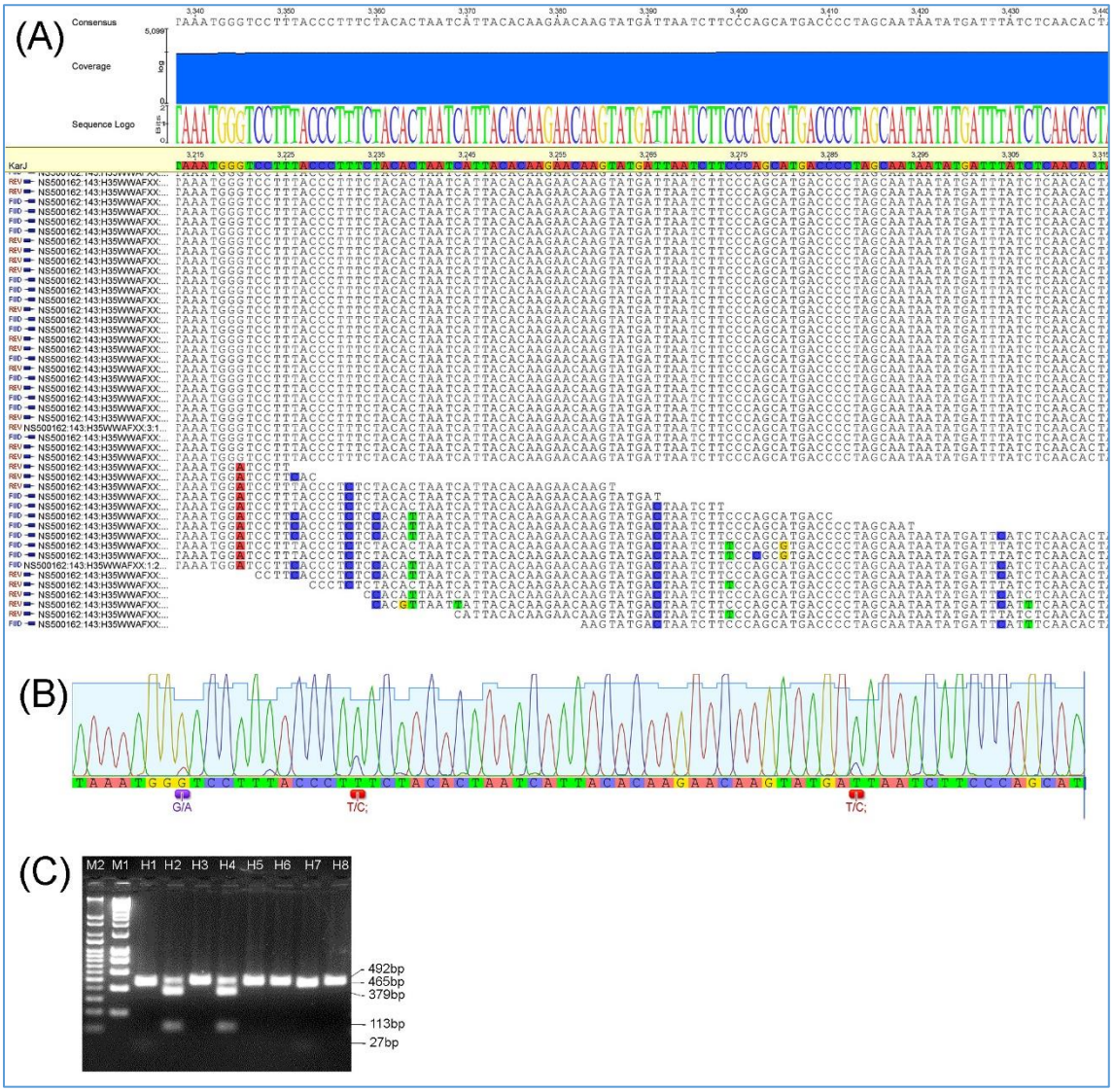


Figure 3.4 **Polymorphisms in mitogenome sequence assembly.** A) Raw 150bp reads assembled to consensus showing some sites with polymorphisms highlighted with boxes. B) Sanger sequencing trace of PCR products spanning the same region showing heterogeneity in base calls (boxes) at the same positions of polymorphisms. C) PCR-RFLP patterns of the ND1 mitochondrial gene digested with restriction enzyme AvaII distinguishing three different haplogroups. Haplogroup HPGB with 2 bands (379bp /113bp) and HPGC/HPGD/HPGE with another 2 bands (465/27) and the uncut band represent the haplogroup HPGA. Lanes 2 and 4 show heterogeneity with HPGA as the major haplogroup, and presence of another haplogroup (uncut). M1 '1kb ladder' from 200bp. M2 Q-Step2 DNA ladder marker from 100bp.

To further elucidate the origin of the variant sequences, the new sheep mitochondrial genome sequences were compared with the nuclear chromosome assemblies of *O. aries Oar_v4.0* databases. 211 nucleotide sequence fragments were found with different lengths with high similarity (total length 236,434 bp) indicating that they are potential nuclear mitochondrial sequences (*numts*) When the fragments were aligned back to the complete mitochondrial sequence, coverage was seen to be relatively equal over all

regions with less over the control region Figure 3.5; only the X chromosome (12681bp) and chromosome 3 (8273bp & 7355bp) had two long assemblies covering nearly the all parts of the mitochondrial genome, and 85% detailed number of the chromosome assemblies had less than 2kb of homology. The 236kb of *numts* in the chromosome assemblies is 7-fold less than the total length of sequence (1,643 kb or 0.055% of all raw reads) with mitochondrial homology but different from *O. aries*, found in the raw reads.

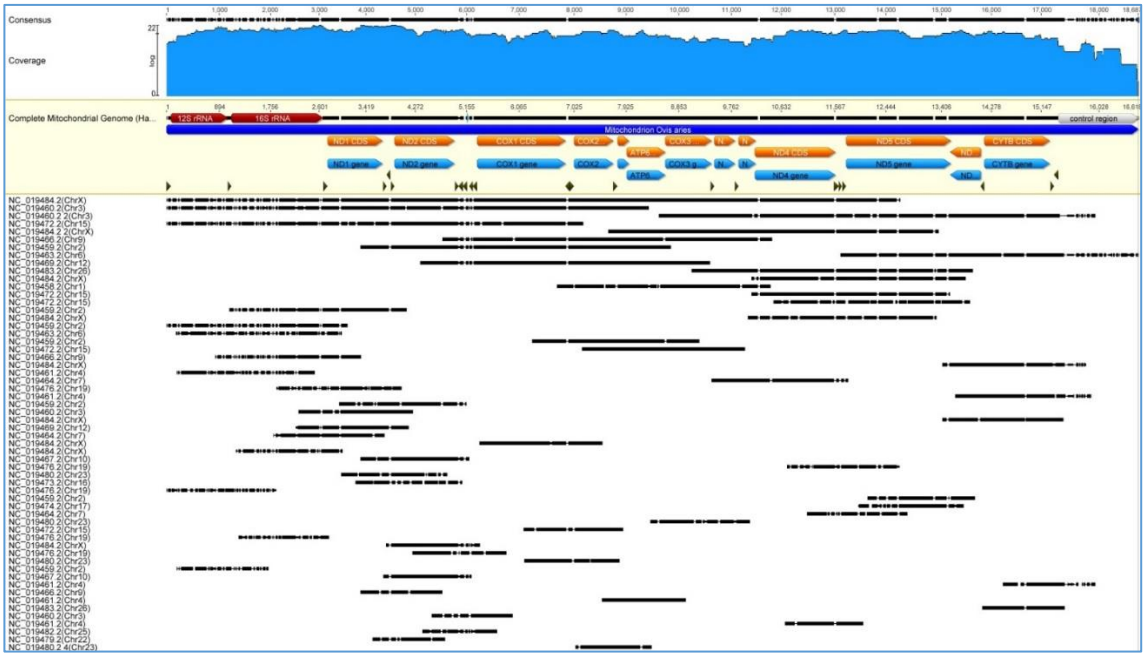


Figure 3.5 **Mitogenome sequences found in nuclear DNA assemblies**. Alignment of selected fragments from the *Ovis aries* whole genome assembly *Oar v4.0* to the sheep mitogenome.

3.4.2.4 **Chromosomal localization of *numts***

We used fluorescent *in situ* hybridization to male metaphase sheep chromosomes with probes from seven regions of the mitochondrial genome, including coding, rRNA, and control regions, to first prove the presence of *numts*, second to localize the *numts* on the chromosomes, and third to indicate their abundance Figures 3.6 & 3.7. All seven probes gave strong *in situ* hybridization signal at the centromeres of the 23 autosomal chromosome pairs with equal to variable strength while additional signal at intercalary and subtelomeric positions could be seen occasionally Figure 3.7B. This indicates that the sequences homologous to essentially all of the mitochondrial genome are integrated and amplified, potentially with degeneration as FISH conditions allow up to 15% mismatch.

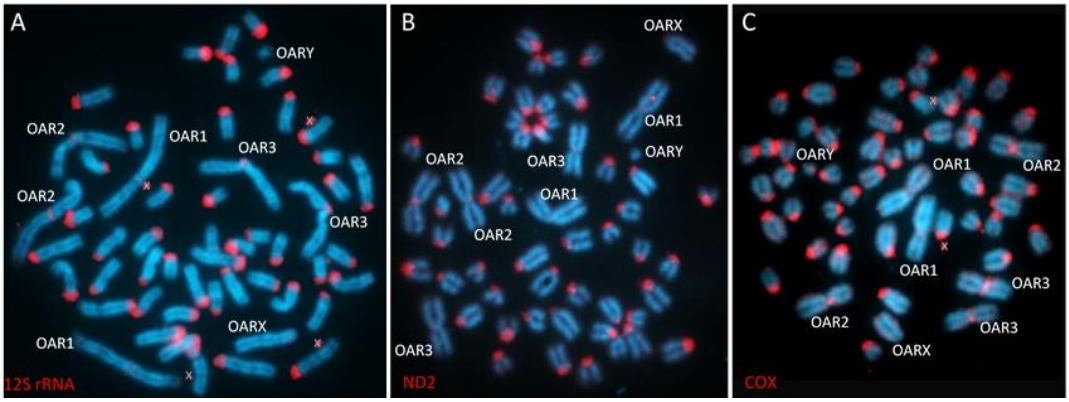


Figure 3.6 **Mitogenome sequences are detected on chromosomes**. Fluorescent *in situ* hybridization of probes (detected in red) to metaphase chromosomes of sheep (2n=54; stained blue with DAPI). Probes were amplified by PCR primers spanning domains of the mitochondrial sequences. OAR chromosome identifications are indicated by numbers. A) 12S rRNA probe; B) ND2 probe; C) COX2 probe. All probes give strong signal at the centromeres of all 23 pairs of autosomal acrocentric chromosomes. Hybridization strength to the centromeres of the three pairs of submetacentric chromosomes and the X and Y chromosomes, always weaker, differs between probes. Figure 3.7 shows hybridization results from four additional mitochondrial probes.

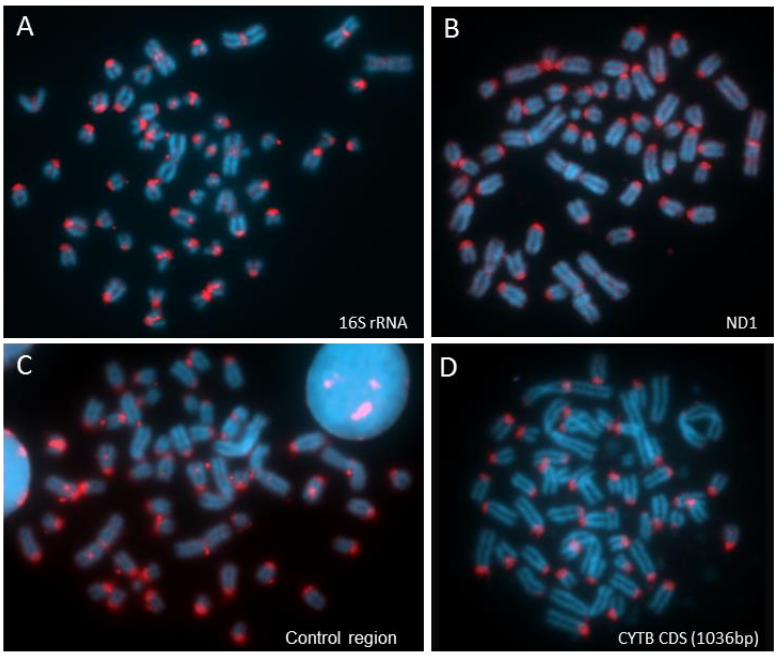


Figure 3.7 Fluorescent *in situ* hybridization of probes (detected in red) to metaphase chromosomes of sheep (2n=54; stained cyan with DAPI). Probes were amplified by PCR primers spanning domains of the mitochondrial sequences A) 16S rRNA probe; B) ND1 probe; C) Control region and D) CYTB probe (see Table 3.1C for probe description). All probes give strong signal at the centromeres of all 23 pairs of autosomal acrocentric chromosomes. Hybridization strength to the centromeres of the three pairs of metacentric chromosomes and the X and Y-chromosomes, always weaker, differs between probes.

3.5 Discussion

3.5.1 Mitochondrial sequences of Kurdistani sheep

3.5.1.1 Complete sequences and abundance

Results of our study fill a gap in geographical coverage (Lv *et al.* 2015), with previous sheep samples from Turkey and Israel (proximate to Kurdistan) identifying multiple maternal haplogroups including HPGA, HPGB, HPGC, HPGD and HPGE (Meadows *et al.* 2007; Demirci *et al.* 2013; Rafia & Tarang 2016). Our study indicates that sheep breeds from the Kurdistan Region of Iraq originate from multiple maternal lineages within known consensus diversity, notwithstanding the geographic location of Kurdistan near the centre of sheep diversity and domestication.

Complete mitochondrial genome sequences were assembled from five sheep of two most widespread breeds, Hamdani and Karadi (Figure 3.1 and Appendix 3.1). These fitted to the known sheep HPGA and HPGB haplogroups (Table 3.2 and Figure 3.3), but had additional previously undescribed SNPs.

The blood lymphocyte DNA used here was suitable for extracting and assembling whole mitochondrial sequences. As reported by Ding *et al.* (2015) in human, the female samples had more copies of mitochondria per nuclear genome cell than male cells Table 3.2. While there is only limited energy metabolism in blood lymphocytes, once growth has finished, females, because of pregnancy and lactation, may have a higher energy requirement and hence additional mitochondria, and female mammals usually have longer lifespan.

3.5.1.2 Kurdistani breed haplogroups

To genotype a larger set of animals (in total 31; Table 3.4), PCR primers were designed to span polymorphic restriction sites enabling PCR-RFLP (Tables 3.3 & 3.5). The mitochondrial consensus haplogroups of the three Kurdistani sheep breeds sampled here fitted within three of the five main *Ovis aries* haplogroups known to occur in European and Asian breeds, HPGA, HPGB and HPGC; no HPGD and HPGE were found in the samples analyzed. Half of the Kurdistani sheep were HPGA, with the others HPGB and HPGC

(Table 3.4B). In all breeds surveyed previously, the most common Asian haplogroup has been reported as HPGA, while most European sheep have HPGB (Wood & Phua 1996; Hiendleder *et al.* 1998a), consistent with our findings in Kurdistan (west Asia). Lineage HPGC, found here, has been observed in fat-tail Asian and Middle Eastern sheep breeds (Pedrosa *et al.* 2005) and is frequent in sheep from Southeast Anatolia (Demirci *et al.* 2013). Rafia and Tarang (2016) looked at Iranian breeds suggesting gene flow and intermixing.

With respect to separation of lineages, HPGB represents the mouflon/*Ovis musimon* domestication event while HPGA originates from the *O. aries* lineage Hiendleder *et al.* (1998b) estimates that these European- and Asian-types of mitochondria separated 375,000 to 750,000 years ago. Based on the polymorphisms seen here, the separation between our haplogroup sequence variants HPGA and HPGB (analyzed by MEGA6 (Tamura *et al.* 2013)) is estimated as occurring similarly 400,000 to 800,000 years ago. Loftus *et al.* (1994) show a similar time of separation of zebu and taurine cattle (0.2 to 1 million years ago; 8% different in hypervariable mitochondrial region), both events being well before domestication (c. 10,000 years ago). No additional haplogroups, not described for domestic sheep before were found, despite proximity to wild sheep species. This is different of the situation in cattle where of eleven bison herds in the US, one has cattle mitochondria (Halbert & Derr 2006), an intergeneric transfer of mitogroups. The sheep here, from the interface of Europe and Asia, are unsurprisingly mixtures of the haplogroups.

3.5.2 Nuclear mitochondrial sequences, *numts*

Examination of the raw read assembly showed that bases were different to the consensus mitochondrial sequence (Figure 3.4 and Appendix 3.18). Many such differences are ascribed to sequencing errors which are removed by the high coverage. However, sites were noted where multiple raw reads, from both forward and reverse directions, showed the same alternative bases. This could arise from artefacts including 1) systematic instrument/chemistry/base calling errors; 2) wrong assembly of duplications in the mitochondrial sequence; 3) mapping of nuclear copies of homologous sequences to the mitochondrial genome; or 4) systematic of chance

accumulations of sequencing errors. Continuous improvement of base calling algorithms, chemistry or protocols, read-lengths, and instrument software mean that the rates and nature of errors in Illumina sequence calls are continuously changing and not systematically documented, so reference to error rates in published works such as Wall *et al.* (2014) is of little value. The sequencing here was carried out in two multiplex NextSeq500 mid-throughput 2x150bp runs, with two sheep and three samples of plants, and three sheep and three plant samples. The plant reads gave c. 40-fold coverage of the cytoplasmic chloroplast genome of *Taraxacum* (Salih *et al.* 2017) which was assembled with a similar approach to that described here. The overall error rate in the sequence calls (<1%) was identical to that in the sheep data omitting the alternative calls, and in particular, no systematic alternative reads were seen in other high-coverage assemblies, suggesting that the multiple reads of different sequences here are not sequencing artefacts.

Whole genome assembly algorithms are not optimized to identify nuclear vs mitochondrial sequences, nor to assemble *numts* in chromosomes, so further analysis of whole genomes reverse-analyses the assembly algorithm. BAC sequencing (now largely historical) would include the whole 16kbp mitogenomes within their typical 100kbp (although multiple tandemly arranged copies, and recombination or chimerism would be hard to rule out). Our whole genome read analysis shows the abundance of non-*O.aries* mitochondria (0.05% of the whole nuclear genome). It will not detect which reads of the *O.aries*-type come from nuclear copies vs mitochondria. As shown in the BLAST searches, the whole genome assembly does not include even non-*O. aries* type *numts*: only a few mitochondrial sequences are assembled on 3 or 4 chromosomes, far from the signal from multiple copies of the mitogenome on every one of the acrocentric autosomes after FISH to metaphase chromosome preparations (see Figures 3.6 & 3.7), or from the high level of polymorphisms seen in the PCR results. In the analysis, here, 0.055% of all reads, representing 1.6Mb of DNA per genome (or 100 mitochondrial genome copies), were homologous to mitochondrial sequences other than the consensus type. The strong signal on all the 23 pairs of autosomes chromosomes, from all domains of the mitochondrial genome, suggests more than an average of 4 copies of the mitochondrial genome per chromosome, so it is likely that some of the consensus

reads also originate from nuclear copies. Hazkani-Covo and Graur (2006) have suggested there are 300 to 400 *numts* in human and chimpanzee; cats also have been reported to have *numts*, and we have detected abundant mitochondrial sequences on chromosomes in the grasshopper *Chorthippus* (Vaughan *et al.* 1999). Amplification of large mtDNA fragments longer than 2kbp as *numts* have been found in many metazoans in short fragments (less than 1000bp) (Bensasson *et al.* 2001).

Fluorescent *in situ* hybridization results show very little *numts* signal on sub-metacentric chromosome pairs, with only some weak signal on chromosome 1. We know submetacentrics and acrocentrics differ in satellite organization (see Chapter Four), and it will be interesting to determine how satellite sequences are organized with respect to *numts* in centromeric regions of acrocentrics and the evolutionarily fused autosomal arms in sheep submetacentric chromosomes, including species of Bovidae family with different numbers of submetacentric chromosomes. Miraldo *et al.* (2012) shows how *numts* were transferred before the separation of the extant species of lizard. In our data, the sheep include mitochondrial sequences similar to other species of *Ovis* and even *Capra* (Appendix 3.17, suggesting ancient transfer before separation of the species of genera.

Evolutionarily, there is a trend for organellar genes to be transferred to the nucleus where the gene is functional in encoding proteins, and normally the gene acquires features of the nuclear genome such as the nuclear gene code including via exon shuffling, promoters and transit peptides (Wischmann & Schuster 1995; Gunbin *et al.* 2017). The *numts* recognized here retained the sequences of other species and all polymorphisms corresponded to sites previously reported to vary between *Ovis* sequences; it is unknown whether they have any transcriptional activity. In previous studies, however, both copy number and the fragment lengths of *numts* have been lower than seen here.

There are sporadic reports of mitochondrial heteroplasmy in vertebrates arising from mutations (e.g. in human; (Wallace & Chalkia 2013; Stewart & Chinnery 2015) and from biparental inheritance (e.g. in great tit; (Kvist *et al.* 2003)). In sheep, Zhao *et al.* (2004) reported paternal as well as maternal (biparental) inheritance of mitochondria, although

this has not been found in normal offspring of other mammals including human (Pyle *et al.* 2015). The mitochondrial DNA composition of seven foetuses and five lambs cloned from foetal fibroblasts showed heteroplasmy in seven of 12 clones tested (Burgstaller *et al.* 2007). Meadows *et al.* (2011) quoting and extending (Hiendleder *et al.* 2002) removed the repeat unit of the mitochondrial control region from phylogenies because of its known heteroplasmic behaviour, and Meadows *et al.* (2007) state "Others have noted that low-frequency mtDNA haplogroups such as HD and HE [HPGD and HPGE] may in fact be nuclear mitochondrial pseudogenes (*numts*) (Parr *et al.* 2006))" but consider *numts* from these haplogroups extremely unlikely by analyzing control regions. Nevertheless, in human, Parr *et al.* (2006) were able to clone the full-length mitochondrial genome from nuclear DNA, like our data showing the whole mitochondrial genomes from other sheep species can be found in *O. aries*.

3.6 Conclusions

Analysis of the mitochondrial DNA sequences of fat-tailed Kurdistani sheep, sampled from the centre of diversity and domestication, were found to have the three of the five major haplogroups of the domestic species. The presence of both the Asian and European types support the multiple domestication events, and the diversity suggests ongoing gene flow, presumably occurring as sheep are traded and move, on top of the historical waves of distribution giving a complex history of domestication. The HPGC haplogroup found is likely to represent recent introgression from wild species rather than independent domestication as HPGC is linked to geographic range of fat-tailed breeds (Lv *et al.* 2015). While no major haplogroups were identified in this study, nor was there evidence for ongoing introgression, some SNP variants may represent important genotypes for conservation of genetic diversity in sheep and be resources for geneticists and breeders of the strong and robust fat-tailed types.

The whole genome sequencing results showed presence of substantial numbers of copies of the essentially the whole mitochondrial genomes of other species of *Ovis* and *Capra*. *In situ* hybridization showed the mitochondrial genome was present on nuclear chromosomes, particularly at the centromeres of the acrocentric autosomes. These *numts* were presumably introgressed before separation of the modern species, over 4

MYA, but the strong *in situ* signals indicate presence of the *O. aries* mitochondrial sequences too, although the chromosome assemblies show very few *numts*. Our results emphasize the need for considering *numts* in analysis of phylogeny (and heteroplasmy or identification of mt disease variants), and also the need to improve genome assembly algorithms to account for repetitive sequences, including mitochondrial-related *numts*.

Chapter 4 Tandemly repetitive sequences: their abundance, diversity and chromosomal distribution during mitosis and meiosis in sheep.

4.1 Introduction

Eukaryotic genomes constitute a large amount of repetitive DNA sequences (in assembled vertebrate genomes make up about 4-60% (Sotero-Caio *et al.* 2017)), and amongst them, a significant portion is made up of non-coding DNA sequences that are tandem repeats, also known as satellite DNA sequences (Charlesworth *et al.* 1994; Elder Jr & Turner 1995; Schmidt & Heslop-Harrison 1998). Isolation of repetitive DNA sequences has been reported, including satellite DNAs from species within the Bovidae family in the order of Artiodactyla, particularly in the two important domestic animals, cattle and sheep. For many years, repetitive DNA sequences have been thought to be junk DNA, but increasingly important chromatin organizations essential for the functioning of the genome and chromosomes during division have been associated with repetitive DNA sequences (Heslop-Harrison & Schwarzacher 2011), and as Grewal and Jia (2007) indicated that the association of repetitive DNA sequences in the formation of heterochromatin is essential for functional organization of centromere and telomeres.

Karyotypes in the Bovidae have been studied extensively by G-banding in the 1980/90s (Gallagher Jr & Womack 1992; Iannuzzi & Di Meo 1995) and they showed that the main mode of chromosome rearrangements involve centric fusions and fissions, so called Robertsonian translocations. Cattle only have acrocentric chromosomes, while sheep and pig have submetacentric, metacentric and acrocentric chromosomes. Some satellite sequences were isolated in cattle and sheep (Chaves *et al.* 2000a; Chaves *et al.* 2000b; Adega *et al.* 2006) and deletion of certain centromeric sequences were shown after chromosome fusion events and formation of metacentric chromosomes of cattle in the frequent 1;29 translocation found in some breeds (Chaves *et al.* 2000b). In pig, two distinctive satellite families are present that are located at the centromeres of metacentric or acrocentric chromosomes respectively. The acrocentric DNA sequences

are AT-rich and show high sequence homology between members, while the metacentric are GC-rich and more diverse (Jantsch et al. 1990). Interestingly, pig chromosome 1 that constitutes two chromosomes in the related peccary has both the acrocentric and metacentric sequences. This was proposed to be a consequence of differential meiotic behaviour and tight association of acrocentric centromeres during pachytene when recombination takes place and would allow sequence to become similar through DNA exchange and non-reciprocal recombination (Schwarzacher *et al.* 1984; Jantsch *et al.* 1990).

During meiotic prophase, the synaptonemal complex (SC), a proteinous structure is formed that promotes homologous chromosome synapsis and recombination. Chromatin is substantially reorganized to attach to the SCs in loops of different, density and species-specific size (Capilla *et al.* 2016). The association of repetitive DNA sequences in meiosis in context of synaptonemal complex has been investigated using different methods. Repetitive DNA elements might contribute crucial role in the earliest activities of meiotic prophase, and might be involved with the components of the synaptonemal complex to perform a functional role in conferring definite chromatin arrangement and attachment (Hernández-Hernández *et al.* 2008; Schwarzacher 2008; Johnson *et al.* 2013).

In recent years, the International Sheep Genomic Consortium (ISGC) generated the fourth version (Oar.V4) of draft assembly of the sheep genome (https://www.ncbi.nlm.nih.gov/genome?term=*Ovis*%20aries) see also (Archibald *et al.* 2010). Nevertheless, repetitive DNA sequences, in particular satellite DNAs are often missing or are badly annotated from the whole genome sequences as programmes mask them to achieve assembly, and features such as genomic proportion, structure, amplification mechanism, diversification and homogenization of these repeats remain largely unexplored.

Therefore, alternative methods are required for analysis of tandemly repetitive sequences. Here, we aimed to use raw unassembled reads from next generation sequencing data utilizing available bioinformatics algorithms and cytogenetic

techniques to investigate abundance, diversity, organization, specificity, meiotic behaviour and evolution of major tandemly repeated sequences in *Ovis aries* genome.

4.2 Aims and objectives

The current study aimed to

1- Characterize the major tandemly repetitive DNA sequences of the sheep genome using whole sequencing raw reads of four male genomes and one female genome employing a range of bioinformatics resources.

a- Identify new tandem repeats including specific tandem repeat for sheep genome or for any individual chromosome.

b- Investigate the structure and monomers of identified repeats and confirm novelty using Tandem Repeat Analyzer (TAREAN) server.

c- Investigate the abundance, sequences diversity and genomic proportions of tandemly repetitive DNA sequences and compare it to the sheep whole genome assemblies.

2- Characterize the chromosomal locations and organization of each identified tandemly repetitive DNA sequence family using *in situ* hybridization experiments and compare with bioinformatics results.

3- Investigate meiotic behaviour of major satellite sequences in relation to the synaptonemal complex by means of *in situ* hybridization combined with immunostaining.

4.3 Materials and Methods

4.3.1 Primers, PCR amplification and cloning

Primers are listed in Table 4.1. Those to amplify known repeats where either as described in the literature such as [Satellite I (SatI); GenBank: X01839.1 (Reisner & Bucholtz 1983); Satellite II (SatII); GenBank: X03117.1 (Buckland 1985)]; sequences of satellite I and II clones (Chaves *et al.* 2000a; Chaves *et al.* 2003b; Chaves *et al.* 2005); sequences of bovine satellite I (AJ293510; Chaves *et al.* (2000b)) and bovine satellite IV (X00979; Skowronski *et al.* (1984) and (Nijman & Lenstra 2001)) or designed from identified sequences in my own analysis using Primer3 of Geneious software version 8.0 (Kearse *et al.* 2012) (http://www.geneious.com). Novel repeat sequences identified within the NGS data (see Results) used the identified monomer to design primers; sequences (repeats or probes) derived from RepeatExplorer are denoted with CL and the number gives the identified Cluster. *k-mer* analysis gave repeats as contigs. Sequences derived from *k*-mer are denoted with the number of *k*-mer that used for frequency of short motifs. Parameters and cycling conditions of PCR amplification were carried out following section 2.2.2. PCR products containing satellite I and II sequences were purified, cloned and sequenced following sections 2.2.2.2; 2.2.3.1; 2.2.4 and 2.2.7.1. Identification of tandemly repetitive DNA sequences was followed section 2.2.13.

Table 4.1 List of primer sequences, annealing temperatures and expected product size of PCR products used for amplification of tandemly repetitive sequences.

Repeat/Probe name	Primers [Sequence (5'-3')]	Expected Product Size (bp)	Annealing temp.
CL4_SatI	F= CAAAGCAGCGAAAGGAACCC	432	56,58
	R= ACATGTCAGCACTGGAGAGG		
CL3C3_SatI	F= ATGTGCTGTTTCCACTCC	514	56
	R= TCCCTAACATACCCATCTCC		
32merC3_SatI	F= CAACTCGAGACAATCCAGG	400	56
	R= GAATGGACTGACACATCTGG		
CL7C71_SatII	F= CTAGGTTCAAAGGCAAGGG	469	54
	R= CATGACAGACGTAGGTTCC		
CL6_SatII	F= CCTACCGCTCTCCTCCCC	548	56,58
	R= CTAACCCTTCCCTTTCCTCGC		
Cloned_SatI_639	F= CTCTCGAGTGGAGACGGGTA	639	60
	R= AGTGTCCCCCAGATGTCTCA		

Cloned_SatII_535	F= GTTGCACATCCAAGGGCTCC	535	58-62
	R= CCGGCAGAGCAGCCTCGC		
Cloned_SatI_791	F= CCCTCATCTCGAGCTACGAGGCG	791	64
	R= GAATTCCAGGCGTTCCCGTCGCA		
Telomeric_Tndm	F= TTAGGGTTAGGGTTAGGG	Smear to little bands	66-60
	R= CTAACCCTAACCCTAACC		
CL22C4_Sat	F= TTGAGGTGTGACAGGAACGC	461	58,60
	R= TAACACTGACGTTCCAGCCG		
ERV2	F= ATTGTGGGGATGCAGAGCC	591	60
	R= TGCATCCCAAGAGATCAGCC		
Putative Sat_716bp_NSBL	F= CTTTCCCTGTGAGTGTGGGG	616	62
	R= TGTGTGAACCTGCATACCCG		
Putative Sat_716bp_NSR8	F= GCCAACAATTTTCCTGAAAGACC	270	62
	R= TAGAGGCTTGCGAACACACC		
32merC16_Sat_CRC	F= TCATGACATCCAAGCAAGCG	508	62
	R= ATTCGCATTTCTTTCCCGCG		
CL66_TND_Ychr	F= ACTTTCGTCTTCACACTCCCC	450	50-58
	R= CATATGGACAAGTTTCCTCTCAGGG		
SatI-2_Bovine	F= ACTCGAGATTCCCGCCG	550	62
	R= GTGACAGGCCGCTTGTCGAG		
SatI-4_Bovine	F= CGACAAGCGGCCTGTCACCC	900	66
	R= GAGTTGCGGCGGGAATCTCG		
SatIV_Bovine	F= AAGCTTGTGACAGATAGAACGAT	603	58
	R= CAAGCTGTCTAGAATTCAGGGA		
SatI_after junction SatI_AJ	F= ACCAGATCGAGTCCCTGAGG	539	58
	R= TCTTTCCCACGAGGCTTTCC		
SatII_before junction*	F= TTGGGGAGGGCTTAGGGG	598	58
	R= GGAAGCCTAGCTGTGAGGC		
Junction between SatI and SatII*	F= TTCAGTTTTGTTTGGCCCCC	527	66
	R= CCTCCTCTTGAGATGCGACG		
69bp_ Junction	AGTTTTGTTTGGCCCCCCAAACTCACCTCTGCACTAGGAGCACCAGGCTCCCTG GGGCTGTGGATGAGG		
Novel Tandem_44bp	GCCCCACCCGGAAATCACGTGGGCCCCACGG		

Notes; * only used for PCR amplification not for probes.

4.4 Results

4.4.1 Identification of tandemly repeated DNA sequences using graph-based clustering (RepeatExplorer)

Tandemly repeated DNA sequences were identified from whole sequencing raw reads of two different samples of HamJ1_male and KarJ_female using graph-based read clustering (see section 1.7.1) (Novák *et al.* 2010; Novák *et al.* 2013). The outcomes of RepeatExplorer are generated in form of clusters; each cluster being composed of different numbers of reads that are aligned into contigs. RepeatExplorer classifies clusters by homologies to known protein sequences of retro-elements and DNA transposons, and DNA homology to a few satellite, microsatellites and rDNA sequences using RepeatMasker (Novák *et al.* 2010; Novák *et al.* 2013); percentage of hits are given for each cluster and allow the identification of the class of repeats present in each cluster. Many are composed of a single main repeat class, but others are composites probably indicating degeneration and rearrangements within and between repeat classes; some are classed as 'low complexity', really meaning 'not identified' or 'other' retro-element and DNA transposon related sequences are easy to identify by their protein hits, show semi-linear or large circular clusters.

On the other hand, satellite and tandem repeats are generally represented in RepeatMasker and do not have known protein domains and hence are not automatically characterized by RepeatExplorer. Therefore, several additional methods were applied to identify and describe tandem repeats in the sheep genome. Read sequences of each putative tandem repeat cluster were compared with Repbase databases (Jurka *et al.* 2005; Bao *et al.* 2015). Additional clusters were identified by their circular graph layouts patterns, e.g. CL12_HamJ1 and CL15_KarJ (Table 4.2). Read sequences of these clusters were blasted against Tandem Repeats Finder (Benson 1999) using default parameters to identify short tandem motifs. A summary of RepeatExplorer clusters with tandem or satellite repeats identified is shown in Figure 4.1 and Table 4.2; the major satellites I, II and two novel satellite sequences as well as other satellite like sequences were identified and showed significant abundance within the top 94 clusters Figure 4.1.

Satellite I occupied the second highest position of sheep genome clusters following dispersed non LTR retrotransposon repeats (see Chapter Five) and was distributed over four clusters, each one with different genomic proportions Figure 4.1. On the other hand, satellite II sequences were only specific to one cluster within the top ten clusters. Similarly, Novel Tandem_44bp repeat, was assigned to one cluster amongst the 20th top clusters, while, the Putative Sat_716bp and other unknown satellite like sequences were found with very low genomic proportions distributed over several clusters Figure 4.1.

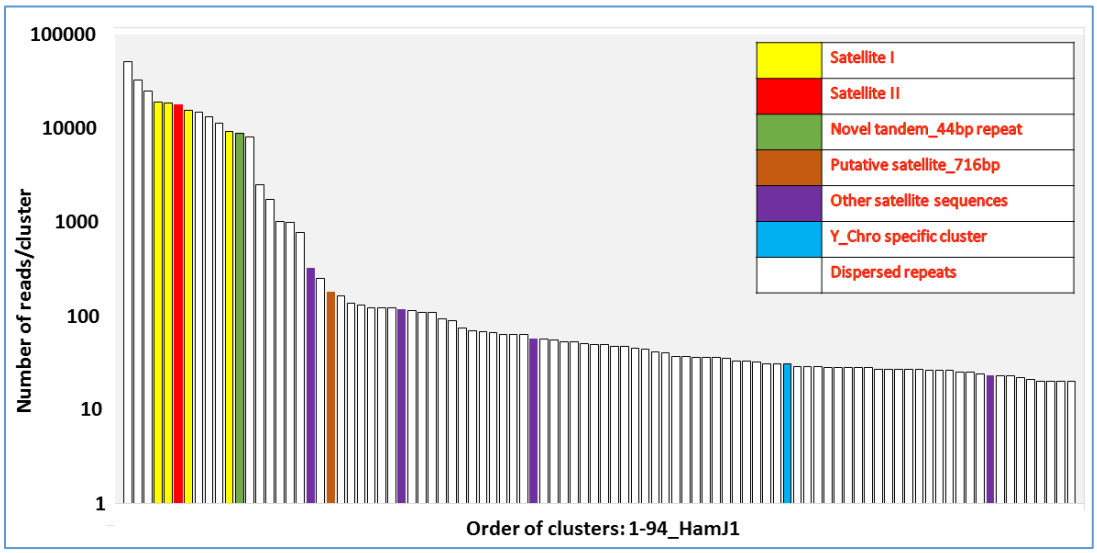
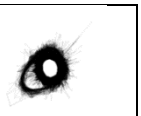
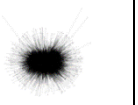
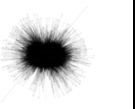
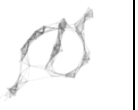
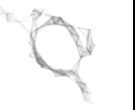


Figure 4.1 Size distribution and composition of different tandem repeat within the main clusters CL1-CL94 of HamJ1 genome as identified by RepeatExplorer. Each chart bar represents one cluster that is coloured according to the main class of tandemly repeated sequences found from the reads of each cluster. The number of reads in each cluster diagrammed the height. White bars represent dispersed repeats see Chapter Five, other colours as described in the figure.

Table 4.2 Genomic proportions, sequence hits and graphic layout of tandemly repeated clusters. Clusters containing similar or identical sequence families were found within the two samples (HamJ1 and KarJ), but due to random selection of raw reads and real differences between samples; cluster numbers that are attributed by hit ranking are not the same. Construction of graph layouts were explained in section 1.7.1.

Tandem repeats & their clusters	Total length [bp]	Number of reads	Genome proportion[%]	RepeatMasker	Graph Layout
Satellite I **CL4_HamJ1** **CL3_KarJ**	2845107	18882	2.26	Satellite.centr (18841hits, 98.2%) LINE.RTE.BovB (3hits, 0.0064%) Satellite (1hits, 0.00295%)	

Satellite II **CL6 _HamJ1** **CL7C71_KarJ**	2665850	17692	2.11	Satellite.centr (17646hits, 97.1%) Low_complexity (6hits, 0.00735%)	
Novel Tandem_44bp **CL12_HamJ1**	1322786	8778	1.05	Low_complexity (1802hits, 5.5%) Simple_repeat (976hits, 2.7%) Satellite.centr (1hits, 0.00386%)	
Novel Tandem_44bp **CL15_KarJ**	1154749	7666	0.7800	Low_complexity (1346hits, 4.52%) Simple_repeat (943hits, 3.04%) Satellite.centr (2hits, 0.0109%) Satellite (1hits, 0.00476%)	
Putative Sat_716 (NSR8 & NSBL) **CL21_HamJ1**	26962	179	0.0214	Satellite (20hits, 6.87%) Satellite.centr (5hits, 2.19%) DNA.hAT.hAT5 (6hits, 1.85%) Simple_repeat (12hits, 1.75%)	
Putative Sat_716 (NSR8 & NSBL) **CL31_KarJ**	23332	155	0.0158	Satellite (38hits, 15.4%) Satellite.centr (12hits, 6.37%) Simple_repeat (6hits, 1.06%)	
32merC16_Sat_CRC **CL17_HamJ1**	150277	998	0.119	LTR.ERV1 (463hits, 40.2%) Satellite (129hits, 9.06%) Satellite.centr (47hits, 3.93%)	
32merC16_Sat_CRC **CL20_KarJ**	135938	903	0.0919	LTR.ERV1 (471hits, 42.6%) Satellite (155hits, 12.4%) Satellite.centr (42hits, 3.81%)	
CL22C4_Sat **CL22**	23332	155	0.017	Satellite.centr (145hits, 78%) Satellite (18hits, 8.38%) LTR.ERVK (2hits, 0.501%)	
CL66_TND_Ychr **CL66_HamJ1**	4670	31	0.0037	LTR.ERVL.MaLR (3hits, 2.93%)	

Genomic proportions of satellite sequences were estimated by summation of genomic proportion of related clusters and showed 6-7% for satellite I; 1.7-2.1 % for satellite II; about 1% for the Novel Tandem_44bp repeat and less than 0.07% for other repeats, in all cases the female genome showed less repeats of each family than the male genome analyzed Table 4.3.

Table 4.3 Genome proportion of major tandemly repeated sequences identified in unassembled raw reads of sheep genome using graph-based read clustering.

Tandem repeats	**Genomic proportions %**	
	HamJ1_male	**KarJ_female**
Satellite I	7.42	5.77
Satellite II	2.11	1.69
Novel tandem_44bp repeat	1.05	0.78
Novel putative satellite_716bp	0.0214	0.0158
Other satellite sequences	0.062	0.05

4.4.2 Male-specific DNA repeats on Y-chromosome of *Ovis aries*

Using RepeatExplorer with low threshold 0.001%, led to generate clusters specific to single chromosome. Sequences of CL66_HamJ1 showed high sequence identities 95% to two accessions of *Ovis aries* Y chromosome repeat regions including OY9 DNA sequence (U30306.1_Length:10063bp) and OY4 DNA sequence (U30378.1_Length:6710bp). Genomic proportions of CL66 and *Ovis aries* Y chromosome repeat region OY4 DNA sequence (U30378.1_Length:6710bp) were about 0.005% and 0.017% respectively. While, interestingly, CL66 sequences were excluded from the whole paired raw reads of female genome KarJ Table 4.6. The repeat and probe named CL66_TND_Ychr.

4.4.2.1 Assembly of whole genome sequences of *Ovis Canadensis* to CL66_TND_Ychr

The whole raw reads of *Ovis canadensis* genome sequences (accession; SRR1752652.sra) (Miller *et al.* 2015) were downloaded from the DRASearch (https://trace.ddbj.nig.ac.jp/DRASearch/). Firstly, the format of file (SRA) was converted to the fasta file using Galaxy (Goecks et al, 2010) and Geneious software. Then, the fasta file of 312,261,788 raw reads with sequence length of 50bp was produced.

To investigate the Y-specific sequences in the related wild bighorn sheep, whole sequencing 312,261,788 raw reads of *Ovis canadensis* were mapped to the domestic sheep Y chromosome repeat region (U30306.1; 10063bp) containing CL66 sequences, and 289,689 reads (about 0.1%) were assembled. The assembled reads of *O. canadensis* generated a complete consensus sequence of 10231bp. The alignment between Y repeat consensus of *O. aries* and *O. canadensis* showed sequence identities 97.5% including conserved regions and polymorphic sites (transition and transversions).

4.4.3 Structure and abundance of satellite I and satellite II

4.4.3.1 Copy number of satellite I and II.

In order to confirm the copy number estimates from RepeatExplorer and TAREAN analysis, the whole paired raw reads were assembled to monomer of each satellite following section 2.2.13.5. More than 128 thousand copies of satellite I and nearly 42 thousand copies of satellite II, (6.86% and 1.96%) respectively were estimated for the male haploid genome, while, lower copy numbers of satellite I and II monomer approximately 100 and 34 thousand copies (5.23% and 1.59%) respectively were assessed for the female haploid genome Table 4.4.

Table 4.4 Genomic proportion, copy numbers of satellite I and II in each of male and female genomes. Whole unassembled paired raw reads were mapped to their monomers, and their copy numbers were estimated per diploid and haploid sheep genomes.

Satellite DNA monomer-Sex	Whole paired raw reads (Coverage)	Assembled reads	Genomic proportion %	Copies of satellites/one-fold	Copies of satellites/haploid genome
SatI- 803bp HamJ1-M	52048068 (2.6x)	3570149	**6.86**	256263.9	**128131.95**
SatI- 803bp KarJ-F	60605648 (3.03x)	3172661	**5.23**	195576.32	**97788.16**
SatII-702bp HamJ1-M	52048068 (2.6x)	1019978	**1.96**	83747.21	**41873.6**
SatII-702bp KarJ-F	60605648 (3.03)	961278	**1.59**	67782.88	**33891.44**

4.4.3.2 Identification of junction region between satellite I and satellite II

The sequences of CL6_HamJ1 (homologous to CL7_KarJ) initially identified to contain SatII sequences Table 4.2. However, the blast results with Repbase databases indicated that the cluster sequences were highly similar not only to satellite II, but also contained

a region with hitherto unknown sequence. NCBI results showed high similarity to the putative junction sequence in pericentromeric region between satellite I & II previously identified by D'aiuto *et al.* (1997) (accession; U62384.1). Hence, we think that CL6/CL7 contain junctions between satellite I and II. Therefore, the whole genome sequencing paired end reads were mapped against CL6 sequences in order to obtain a full assembly of the junction region including both satellite I and II. Two monomers of satellite I and three monomers of satellite II were found. Numbers of raw reads assembled to junction sequence were about 45-110 reads Figure 4.2 & Appendix 4.1.

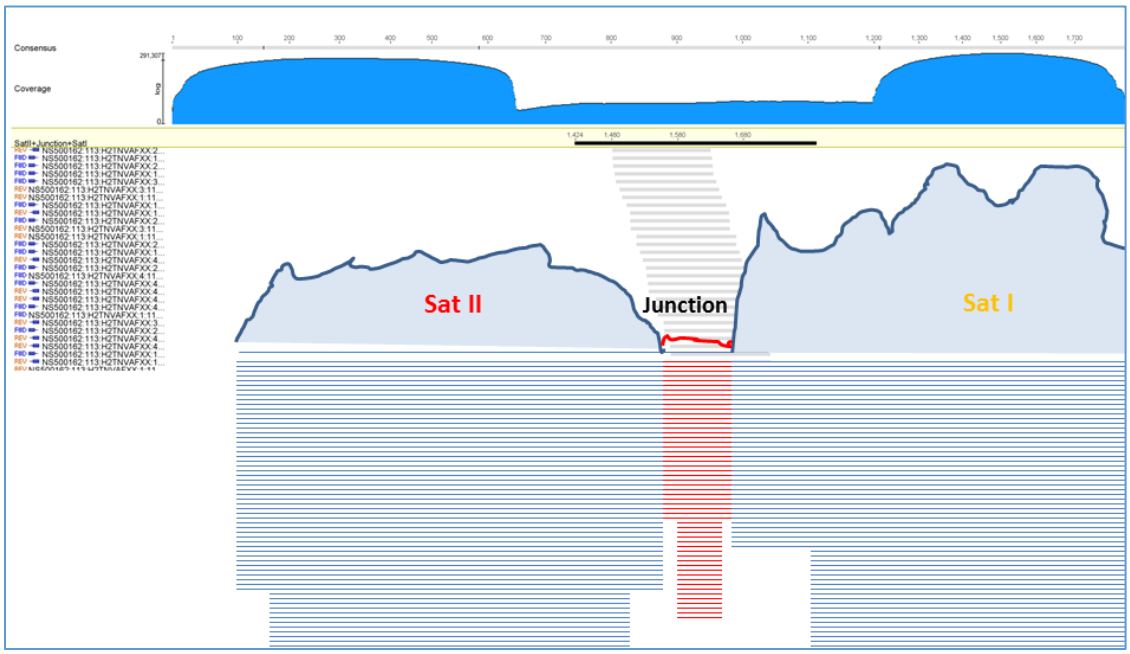


Figure 4.2 Assembly of raw reads (white lines) to junction fragment of satellites (black line). Reads cover the whole sequence with different coverages (blue). *Illumina* sequencing gave good shotgun coverage (equal forward FWD and reverse REV) reads, and matched left and right paired-end reads to the sequence (shown by symbol after REV/FWD and before read code NS500).

4.4.3.3 ***De novo* assembly of whole sequencing raw reads**

In order to find contigs with long satellite sequences, the whole sequencing raw reads were *de novo* assembled using Geneious assembler. Only 20% of the input NGS data were used. As results, approximately 13% of the used NGS data were assembled to generate more than 400000 contigs with different sizes (Appendices 4.2 and 4.3). Consensus sequences of the first top 100 contigs with various lengths (1000-16641bp) were compared with the Repbase and NCBI databases, and most of them were highly

similar to the satellite related sequences. In addition, other satellites were also identified.

4.4.3.4 Assembly of whole sequencing raw reads to the satellite sequences of other species of Bovidae family

The whole sequencing raw reads of sheep were assembled to the satellite I sequences of wild sheep and related species including *O. ammon*, *O. dalli, O. canadensis, O. aries musimon, Oryx gazella*, *Ammotragus lervia*, *Cephalophus natalensis* and *Bos taurus*, and good coverage was produced except the *Bos taurus* satellite I DNA (V00124; (Gaillard *et al.* 1981)) where no reads were assembled. Regarding the satellite II sequences, the whole paired raw reads of sheep were assembled over the complete sequence of goat satellite II DNA fragment with good coverage. However, no reads were assembled to satellite II sequences of bovine and buffalo Table 4.5. Sequence identities between the major satellite sequences of sheep and the corresponding satellites of other species were estimated as shown in Table 4.5. The alignment sequences between sheep satellite I and the other species were used for phylogenetic relationship. As expected from 'map to reference' analysis, satellite I of domestic and wild sheep grouped within the same clade. While other species were more distant (Appendix 4.4).

Table 4.5 Assembly of whole sequencing raw reads of sheep genome to satellites I and II of other species of Bovidae family. No reads were assembled to satellite II sequences of bovine and buffalo.

Satellite I of domestic sheep	**Satellite I of other species (Accessions)**	**Sequence similarities to sheep satellite I**	**No. of assembled reads of sheep raw reads to each satellite**	**Satellite percentage in sheep raw reads NGS data**
The monomer (816bp) sequence of TAREAN	*Bos taurus* 1.715 satellite DNA (V00124)	55.50%	**No reads assembled**	**0%**
	Ovis ammon 1.714 satellite DNA (X96873)	97.30%	3,534,476	6.8
	Ovis canadensis DNA repeat region (X58076)	93.50%	3,548,678	6.8
	C.hircus 1.714 satellite DNA (X57335)	86.10%	3,585,769	6.9
	Ovis dalli 1.714 satellite DNA (X59242)	86.60%	3,166,796	6.1
	Ovis aries musimon satellite I sequence (KM272303 (reversed)	99.50%	3,565,425	6.9
	Oryx gazella clone 2 satellite I sequence (KF787926)	78.70%	3,600,696	6.9

	Ammotragus lervia clone 2 satellite I sequence (KF787909)	88.40%	3,552,724	6.8
	Cephalophus natalensis clone 2 satellite I sequence (KF787907)	74.3	2,515,347	4.8
Satellite II of domestic sheep	**Satellite II of other species (Accessions)**	**Sequence similarities to sheep satellite II**	**No. of assembled reads of sheep raw reads to each satellite**	**Satellite percentage in sheep raw reads NGS data**
The monomer (702bp) sequence of TAREAN	*Bos taurus* satellite II DNA repeat unit 686bp (X03116)	66.60%	**No reads assembled**	**0%**
	Bubalus bubalis clone pDS4.2 satellite sequence 673bp (AY960121)	67.90%	**No reads assembled**	**0%**
	Goat satellite II DNA repeat (240bp) (X03118)	84.4	757,976	1.45%

4.4.4 Identification of novel and minor repeats

In addition to identification of the major two satellites in sheep genome, other unique families of tandemly repeated DNA sequences were identified from the NGS data using graph-based read clustering and other methods Figure 4.1 & Table 4.2.

4.4.4.1 Novel Tandem_44bp repeat of sheep

One of these families was identified in RepeatExplorer CL12_HamJ1 and CL15_KarJ Figure 4.3. It was identified by its unique star like RepeatExplorer graphic Figure 4.3B and D and the self-dot blot of extracted contig 146 Figure 4.3 A and C identifies many tandem repeats of 22 or 44bp monomers within the 3kbp sequence. The consensus sequence shows that this repeat has a higher order structure where two 22bp monomers with 91-100% sequence homology form a higher order 44bp monomer Figure 4.3C. This repeat is the new tandemly repeated member of the domestic sheep genome and does not show homology in the databases. The genomic proportion of this tandem repeat that we called Novel Tandem_44bp was about 0.87% in female and 1.05% in male genome Table 4.3. This novel repeat was also found in the *k*-mers analysis (64mer GT1000).

In order to check the uniqueness of the Novel Tandem_44bp repeat, the *O. canadensis* whole sequencing raw reads (accession; SRR1752652.sra) (Miller *et al.* 2015) were mapped to CL12C164 (3332bp) containing the 22-44bp tandem repeat and to the most

abundant 44bp monomer, but only 150-200 reads were assembled. Another analysis was carried out using a command (grep -A 1 'ATTTCCCCGTGGGGCCCACGTG' Fasta file | wc –l). In this analysis, approximately 611328 and 456532 copies of the 22bp monomer were found in male and female sheep genomes respectively, but only 70 copies of the same monomer length from sheep were found in the whole sequencing raw reads of *O. canadensis*.

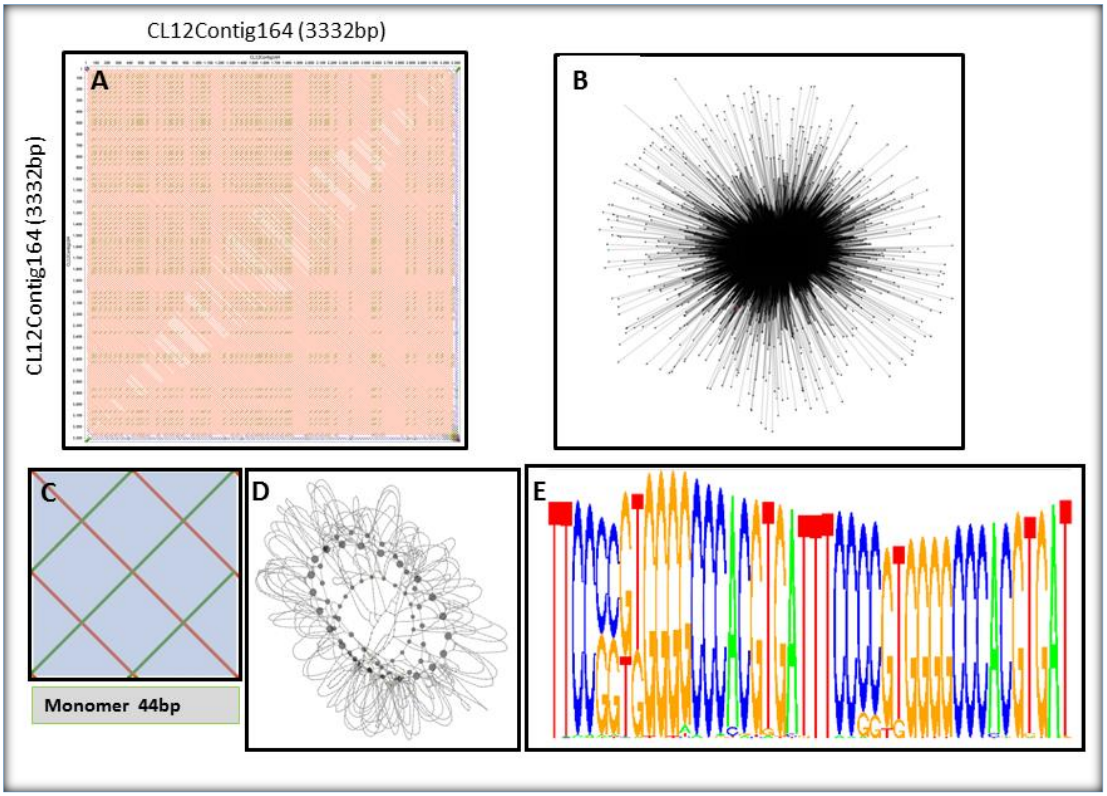


Figure 4.3 Novel tandem_44bp repeat

A) Self dot blot of extracted contig 164 of CL12

B) Repeat explorer cluster graph of CL12

C) Self dot blot of subset of contig 164 showing the monomer of the repeat in an A/A* where two 22bp sub monomers form a higher order structure of a 44bp monomer

D) TAREAN cluster blot showing large circles indicating the 44bp monomer and smaller circles representing the 22bp sub-monomer

E) Consensus Logo representation showing the larger 44bp monomer and the two 22bp sub-repeats monomers

4.4.4.2 **Satellite sequences combined with endogenous retroviruses_ERV2 sequences**

CL22_HamJ1 included 21 contigs with genomic proportion of 0.017%. It corresponded to the CL23_ KarJ analysis. Sequences of CL22 were blasted against NCBI and Repbase

databases, respectively while, lower sequence homology of less than 50% were found to sheep satellite II sequences. Furthermore, RepeatExplorer identified some low components of endogenous retroviruses like sequences in CL22/CL23 Table 4.2. When CL22 sequences were searched against the *Ovis aries* assembly, they were found in *Oar_v4.0* Unplaced Scaffold_004083318.1. Accordingly, NGS raw reads were mapped to the scaffold (6289bp), and consensus sequence with more than 7kbp was resulted Figure 4.4. Repbase comparison of the consensus 7kbp showed two different repeat elements: satellite like sequences (CL22C4_Sat) and endogenous retroviruses_ERV2 Figure 4.4.

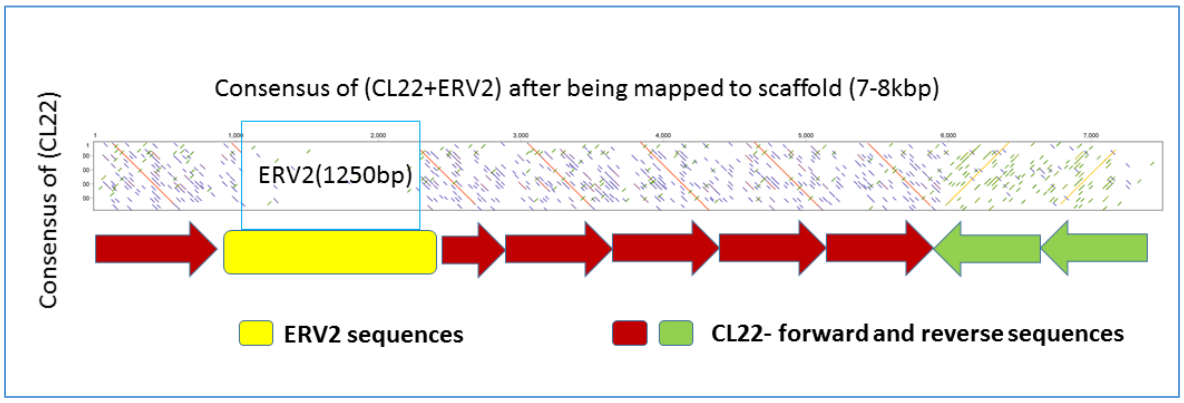


Figure 4.4 Organization of satellite like sequence on RepeatExplorer CL22 and endogenous retroviruses ERVs demonstrated by using dot plot (self) of sequences mapped to *de novo* assembled scaffold.

4.4.4.3 **Putative Sat_716bp repeat**

The sequences of CL21C1_HamJ1 and corresponding CL31C1_KarJ Table 4.2 of graph-based read clustering outcome were compared with NCBI databases and to the *Oar_v4.0*; high sequence similarities were found to the *O. aries* non-LTR-retrotransposon-like sequences and nuclear sequences of chromosomes 20 and 13. However, blasting CL21/CL31 sequences against Repbase databases, sequence identities to satellite families of Bovidae and Cervidae species were found. Furthermore, from alignment of cluster sequences to the corresponding satellite sequences, approximately 50-60% sequence identities were found to satellite II sequences of sheep and satellites of centromeric repetitive DNA of Cervidae species. Showing significant variation and polymorphic bases. Furthermore, when sequences of both clusters CL21/CL31, were aligned to the satellite I sequence of sheep, only 45% sequence identities were found. Genomic proportions of these clusters were about 0.0158% and 0.0214% for KarJ and HamJ1 respectively Table 4.2. Copy numbers of probes NSR8 & NSBL representing this

putative satellite show a higher proportion indicating its sequences are also present in other clusters Figure 4.5 & Table 4.6.

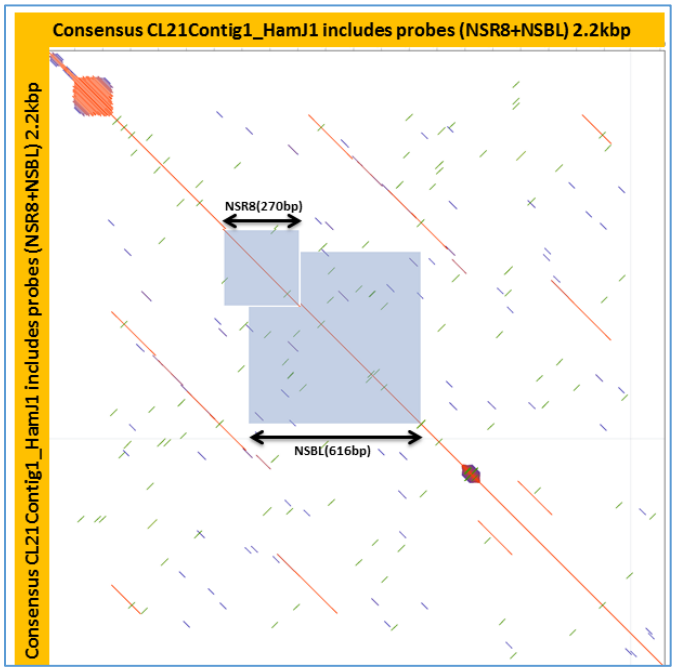


Figure 4.5 Dot plot (self) of putative satellite_716bp repeat showing positions of probes representing this repeat.

4.4.5 Identification of tandem repeats using *k*-mer frequency tool

Tandemly repeated DNA sequences including the major satellites and novel putative satellites were identified using *k*-mer frequency tool (Jellyfish). Substrings of DNA sequence were identified from the whole sequencing raw reads using different values of canonical motifs (K) (16, 22, 32, 56, 64 & 128) and that are repeated 100 to 100000 times (see Chapter Five). Then, all identified short motifs were assembled into longer contigs using Geneious assembler. The assembly of short motifs produced many thousand contigs. Sequence of only the first top 100 contigs were compared with NCBI, Repbase databases and Tandem Repeats Finder. Satellite I sequences were more abundantly found than satellite II sequence in accordance with their genomic proportion. The Novel Tandem_44bp repeat and another new putative satellite 32merC16_Sat_CRC (see below) was identified each in a single separate contig of *k*-mer assemblies.

4.4.5.1 **New putative satellite like sequences 32merC16_Sat_CRC**

k-mer contig 16 resulted from assembly of motifs 32merGT1000 of KarJ was compared with Repbase databases, and the highest similarity 60% was found to the centromeric repetitive DNA from Cervidae species such as *Muntiacus muntjak vaginalis* (AY064466-AY064469). Thus, the sequence with length 508bp was named 32merC16_Sat_CRC. Sequences of all RepeatExplorer clusters of HamJ1 and KarJ were mapped to 32merC16_Sat_CRC and identified clusters CL17C18_HamJ1 and CL20C16_KarJ1 (see Table 4.2). Therefore, the whole paired raw reads were mapped to the longest sequence of CL17C18, and a 5kbp consensus including four copies of the 32merC16_Sat_CRC was found. Consensus of CL17C18 was nearly 60% homology to the sheep satellite II and to the centromeric satellite sequence_*Muntiacus muntjak vaginalis*_AY064466.1 Figure 4.6. Copy numbers of probe 32merC16_Sat_CRC representing this satellite are shown in Table 4.6.

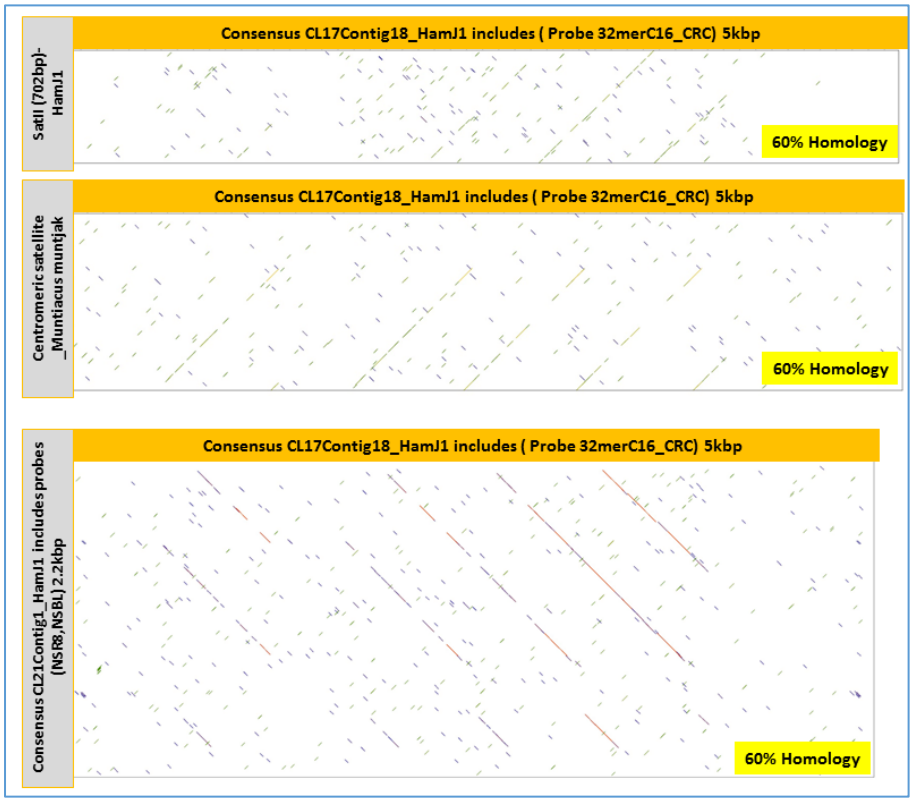


Figure 4.6 Dot plot (self) of satellite like sequences 32merC16_Sat_CRC compared to satellite II sequences of sheep and Cervidae species and to cluster containing sequences of both probes NSR8 and NSBL.

4.4.6 Identification of major and putative satellites in sheep genome using TAREAN: TAndem REpeat ANalyzer

TAREAN tool was used for identification and reconstruction of repeat units which accomplished by estimating the frequency of *k*-mers in the NGS data creating clusters containing tandemly repeated sequences representing the most conserved regions (Novák *et al.* 2017). Here, we used TAREAN to find additional copies of identified repeats to improve the putative monomer consensus sequences, and to further investigate their abundance and structure within the sheep genome. An example of the HTML output of TAREAN results is provided in (Appendix 4.5). Workflow and methods of using TAREAN were stated in sections 1.7.2 & 2.2.13.3.

4.4.6.1 **Satellite I monomers**

TAREAN identified all NGS raw reads related to satellite I sequences and merged them into one cluster, CL2 (Appendix 4.6); the most abundant consensus sequence of satellite I monomer was characterized with length of 803bp and is assumed the most confident putative satellite sequence. Within the cluster of satellite I, consensuses with different lengths were identified based on the analysis of *k*-mers frequency of repeat variability. Additionally, the genomic proportions of satellite I estimated by TAREAN analysis were 7.4% and 5.7% of male and female genome respectively Figure 4.7.

4.4.6.2 **Satellite II monomers**

TAREAN analysis resulted in the specific cluster CL4 related to the sequences of satellite II (Appendix 4.7) which was characterized by the most abundant monomer with a length of 702bp. The report of satellite II cluster contained the most abundant consensus sequences generated from the analysis of different length of *k*-mer frequencies. The genomic proportions of satellite II sequences were 2.2% and 1.7% of male and female sheep genomes respectively Figure 4.7.

4.4.6.3 **Novel Tandem_44bp repeat monomer**

TAREAN analysis identified as a low confidence putative satellite a unique cluster CL7 representing the novel tandem repeat characterized by the most frequent consensus monomer with length 22-44bp. Within the cluster, 15 consensuses with different

lengths were reported (Appendix 4.8). The genomic abundance of the Novel Tandem_44bp repeat in male and female genomes was 1.1% and 0.77% respectively Figure 4.7.

4.4.6.4 **Putative Sat_716bp repeat**

Not only tandem repeats with high genomic abundance were identified by TAREAN, but also putative satellites with low genomic proportion 0.019, such as the Putative Sat_716 repeat found in TAREAN cluster CL16 as a low putative satellite sequence. Within this cluster, 19 consensus sequences with different lengths were classified (Appendix 4.9).

4.4.7 Genomic proportions of major satellites in sheep genome presented in three different methods

Genomic proportions of satellites were estimated and compared utilizing three different methods of bioinformatics analysis, assembly/map to reference, RepeatExplorer and TAREAN Figure 4.7. In general, the three methods estimated similar genomic proportions in particular the map to reference gave lower values for both satellite I and II. The female and male genomes also showed different abundances of the major satellites with the male genome having slightly higher proportions than the female genome in all three analyses. In male genome, there was nearly 7.4% satellite I and 2.2% satellite II. In contrast, the female genome contained lower percentages 5.7% satellite I and 1.7% satellite II Figure 4.7.

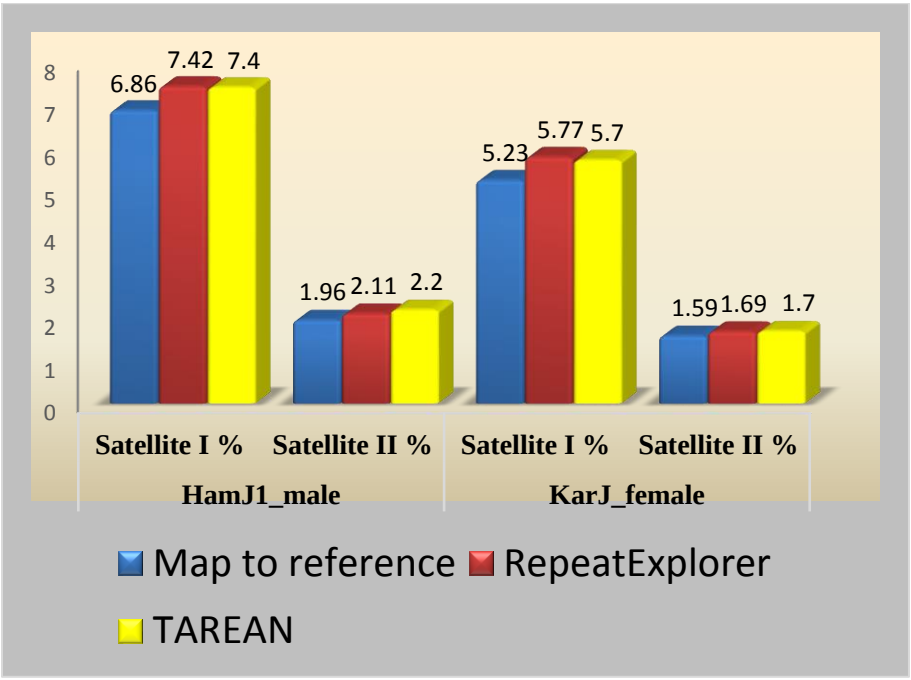


Figure 4.7 Comparison of genomic proportions of major satellites in sheep genomes resulted from using three different methods of bioinformatics (map to reference; RepeatExplorer and Tandem Repeat Analyzer (TAREAN). Male sheep genome is dominated by satellites families over female genomes.

4.4.8 Bioinformatics abundances of probes used for FISH

Table 4.6 Copy numbers and genomic proportion of each probe representing different tandemly repeated sequences used in this chapter were estimated following section 2.2.13.5. Whole genome sequencing raw reads (NGS data) used to assess copy numbers of probes were 52048068 reads for HamJ1 and 60605648 reads for KarJ (see section 2.2.7.2).

		HamJ1_Male			KarJ_Female		
Used probes	**PCR product bp**	**Assembled reads**	**Copies of probe**	**Genomic proportion%**	**Assembled reads**	**Copies of probe**	**Genomic proportion%**
CL4_SatI	432	3479601	1208195	6.6854	3172263	1101480	5.23427
CL3C3_SatI	514	3560785	1039140	6.8413	2489321	726456	4.1074
32merC3_SatI	370	3570253	1447400	6.8595	3142190	1273861	5.1846
CL7C71_SatII	469	1016208	325013	1.9524	717257	229400	1.1835
CL6_SatII	548	998664	273357	1.9187	942125	257881	1.55452
CL22C4_Sat	461	3267	1063	0.0063	3211	1045	0.0053
ERV2	214	1142	800	0.0022	1261	884	0.0021
Putative Sat_716bp_NSBL	616	31629	7702	0.0608	8959	2182	0.0148
Putative Sat_716bp_NSR8	270	10613	5896	0.0204	6744	3747	0.0111
32merC16_Sat_CRC	508	5000	1476	0.0096	6000	1772	0.0099
CL66_TND_Ychr	450	2718	906	0.0052	0	0	0.0000
Repeat region	**Length bp**	**HamJ1_Male**			**KarJ_Female**		
		Assembled reads	**Copies of probe**	**Genomic proportion%**	**Assembled reads**	**Copies of probe**	**Genomic proportion%**
Y chromosome repeat region OY4 (U30378)	6710	8959	200	0.0172	0	0	0.0000
CL17Contig18 _CRC	3935	451130	17197	0.8668	167569	6388	0.2765
CL21Contig_NSR	1456	191687	19748	0.3683	224954	23175	0.3712

4.4.9 Chromosomal organization of major satellite DNA sequences in sheep

4.4.9.1 Satellite I and Satellite II on somatic metaphase chromosomes

Single and double probe FISH was carried out to determine the chromosomal distribution of satellite I (SatI) and satellite II (SatII) of sheep on somatic male sheep metaphase chromosomes and in some cases also on cattle chromosomes. Probes were labelled with biotin and digoxigenin and detected by red and green fluorescence respectively. Chromosomes were stained with DAPI (seen in blue).

Probes of satellite I included, CL4_SatI, CL3C3-SatI and 32merC3_SatI and all showed slightly different hybridization intensities. In general, probes hybridized to all acrocentric chromosomes resulting in signals on their centromeres while no signal was detected in the sex chromosomes X and Y. Signals of SatI were seen on the two smaller pairs of submetacentric chromosomes, but no or only very weak signal was seen on the largest pair of submetacentrics Figure 4.8.

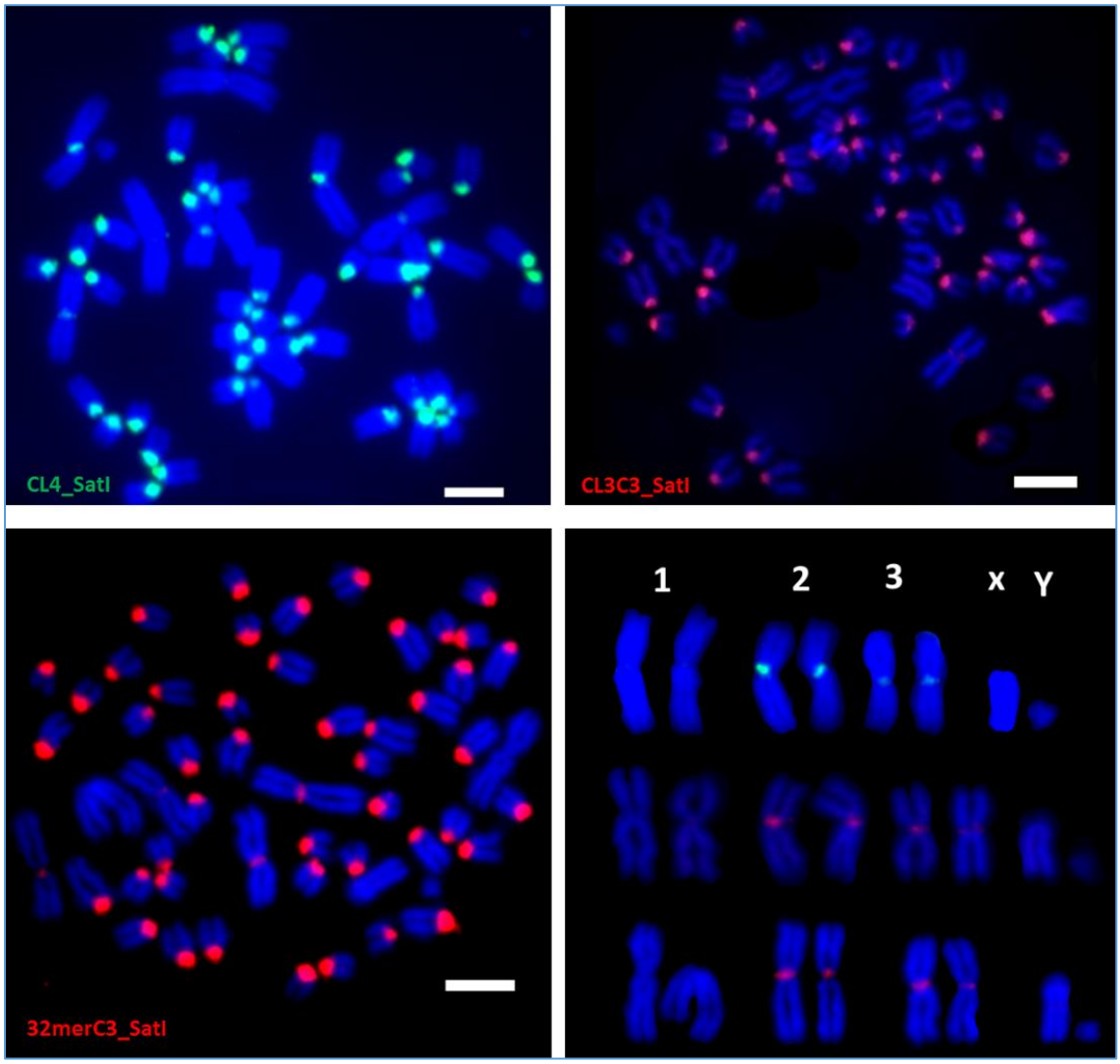


Figure 4.8 FISH of probes CL4_SatI, CL3C3-SatI and 32merC3_SatI hybridized to the metaphase spreads of male sheep chromosomes (*Ovis aries*; 2n=54, XY). Satellite I 32merC3_SatI amplified from contig consensus of the *k*-mers frequency showed stronger and broader signals over centromeric regions of chromosomes in comparison to probes that resulted from RepeatExplorer or from publication databases. Karyotype of three pairs of submetacentric and the sex chromosomes shows signals on second and third pairs of submetacentric, while no signals were shown on the first submetacentrics and sex chromosomes. Scale bar equals 5µm.

Probes of satellite II sequences CL7C7_SatII, CL6_SatII and cloned_SatII_535 hybridized with all acrocentric sheep chromosomes showing slightly variable signals on their centromeres. In contrast to satellite I, signal of satellite II was detected on the X chromosome, but also not on the Y chromosome. Satellite II signals were seen on all three pairs of submetacentric chromosomes Figure 4.9.

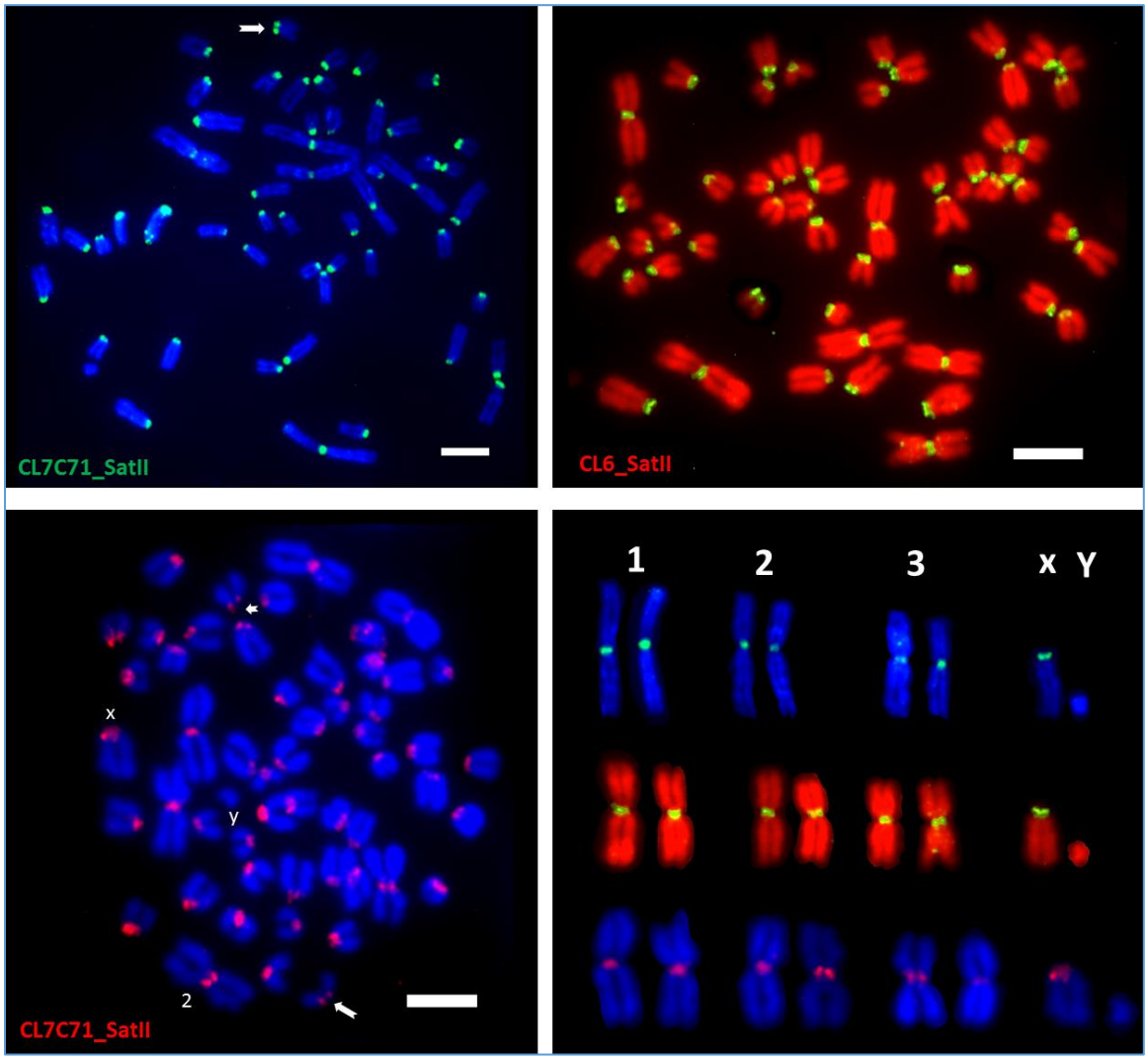


Figure 4.9 FISH of probes CL7C7_SatII, CL6_SatII and cloned_SatII_535 hybridized to the metaphase spreads of male sheep chromosomes (*Ovis aries*; 2n=54, XY). Two separate signals of SatII can be seen on some acrocentrics and submetacentrics (see arrows for acrocentrics and no.2 for submetacentrics. Scale bar equals 5µm.

Both satellites cloned_SatI_639 and cloned_SatII_535 were used for dual probe hybridization experiments to establish their physical relationship on the chromosomes overlapping but variable strength signal was observed on the acrocentric and submetacentrics, with the SatII sometimes stronger and SatI stronger in other cases; probe SatII appears more distal then SatI (Figure 4.10C & 4.11). Furthermore, CL4_SatI and CL6_SatII were also used for *in situ* hybridization against cattle chromosomes, but no hybridization signals were found in all cattle chromosomes including the sex chromosomes.

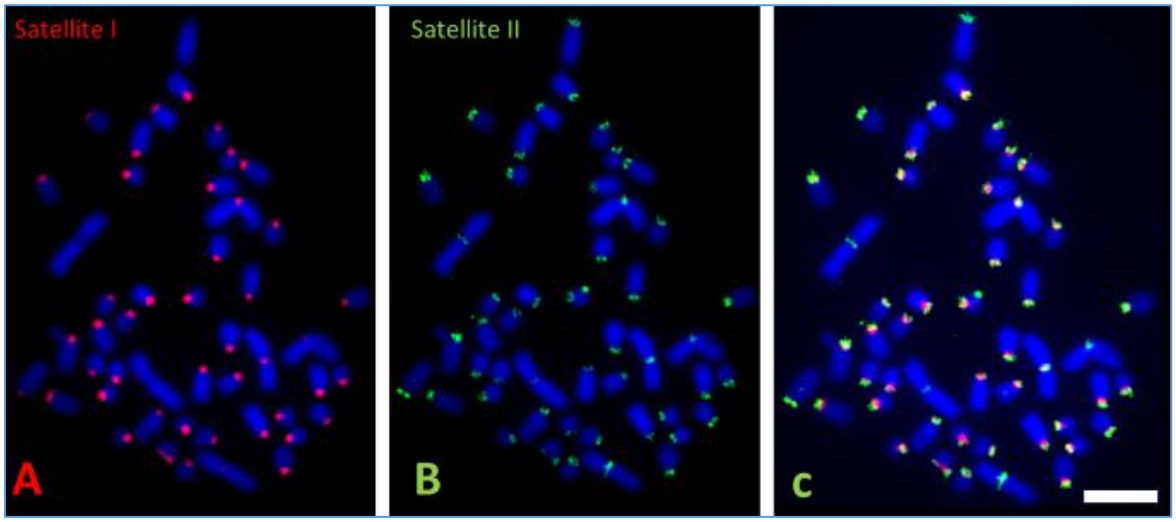


Figure 4.10 FISH of probes cloned_SatI_639 and cloned_SatII_535 to metaphase spreads of male sheep chromosomes (*Ovis aries*; 2n=54, XY). Scale bar equals 5µm. Both satellites show inconstant signals. Both satellites locate different position as cloned_SatII_535 seems to be closer to acrocentric ends than SatI sequences. Scale bar equals 5µm.

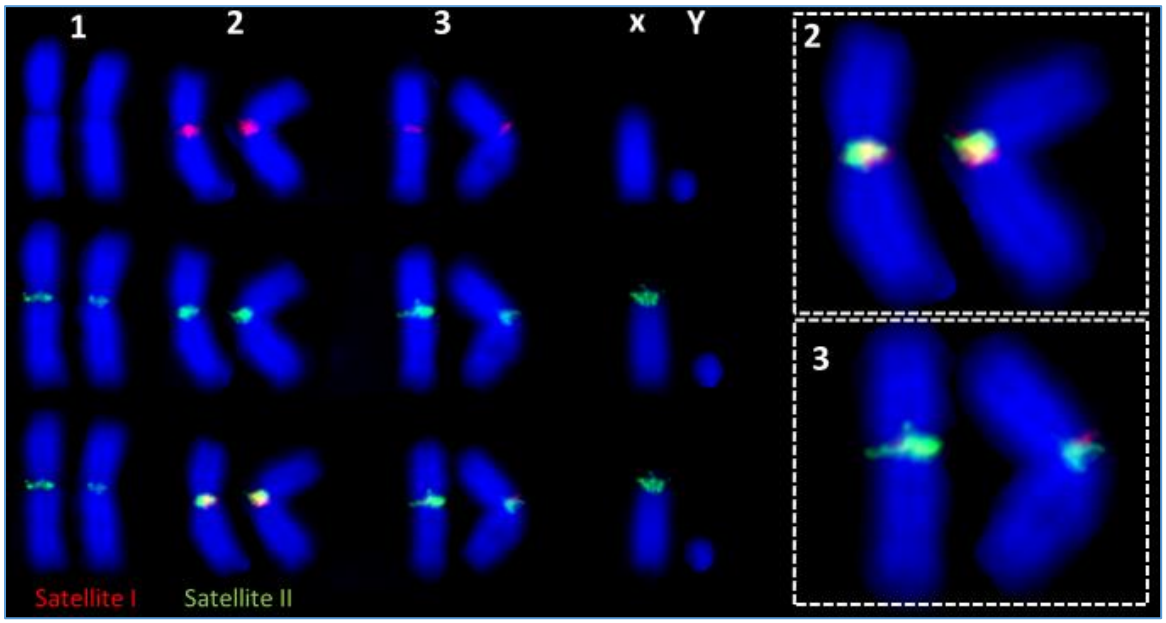


Figure 4.11 Karyotype of three pairs of submetacentric and the sex chromosomes hybridized with dual probes representing cloned SatI and SatII. It shows overlay of red and green signals showing in yellow where signal overlap.

4.4.9.2 **Major satellite organization at meiotic prophase of sheep**

FISH and immunostaining experiments were combined to identify the synaptonemal complexes (SCs) and investigate meiotic behaviour of major tandemly repetitive DNA sequences of sheep. Detection of four colours were developed using blue, green, red and far red fluorescence. All preparations were stained with DAPI in order to make sure that only single nuclei were analyzed (Appendix 4.10). Different combinations for FlSH and immunostaining were used such as anti-digoxygenin FITC (green) and streptavidin-

Alexa 647 (far red) for the detecting the FISH signals of satellite DNA sequences and anti-rabbit Alexa 594 (red) to visualize SCPI antibody; the major component of the transverse filaments of SCs.

The findings of these experiments indicate that satellite I sequences cover a larger area and show a looser chromatin loop organization (Figure 4.12A&D). Satellite II sequences, on the other hand, are tightly organized and are attached to the SC at a more distal position than satellite I sequences at the end of SCs of acrocentric chromosomes (Figure 4.12B&D).

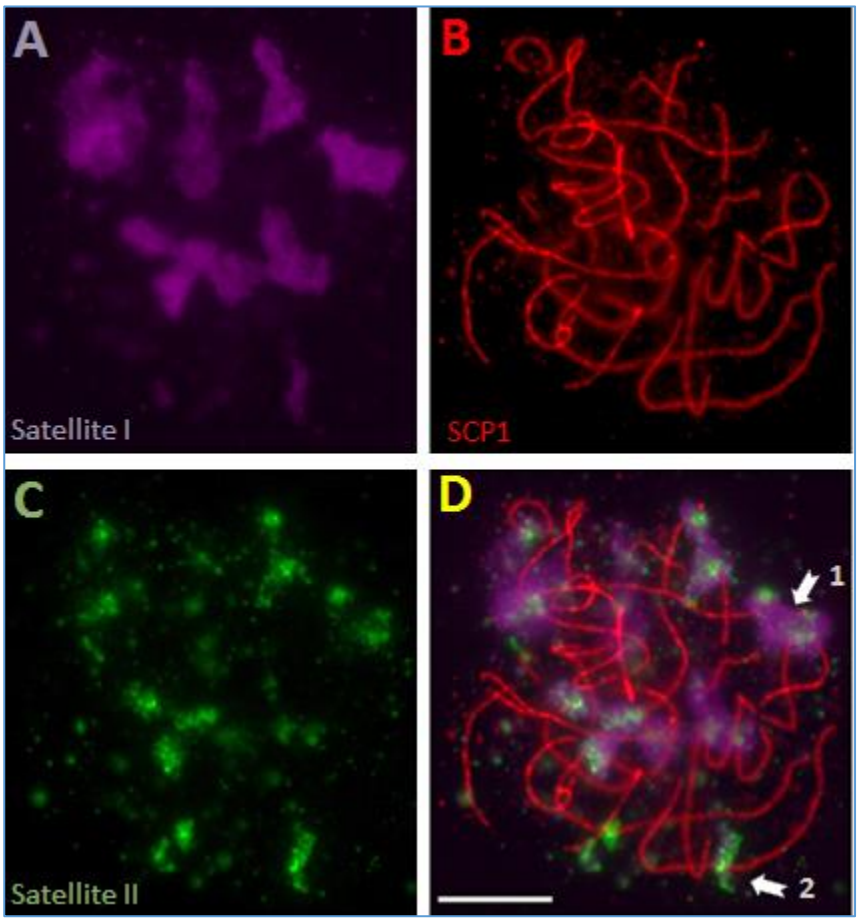


Figure 4.12 FISH and Immunostaining results of sheep SC spreads at pachytene stage using SCP1 rabbit antibody and probes of satellite DNA I and II of sheep. Probe cloned_SatI_791 was labelled with biotin-16-dUTP detected by Alexa 647 Streptavidin (Far red signal, in A) and probe cloned_SatII_535 was labelled with digoxigenin 11-dUTP detected by fluorescein-isothiocyanate (FITC; green signal in C). SCP1 signal detected with Antirabbit Alexa 594 (B). Overlay of the signals including SCPI in red, satellite I-Bio in far red and satellite II-Dig in green (D). 1. The heterochromatin of several chromosomes forms large satellite I labelled clusters with 3 synaptonemal complexes attached. 2. SatII is seen in the middle of submetacentric SCs. Scale bar equals 5µm.

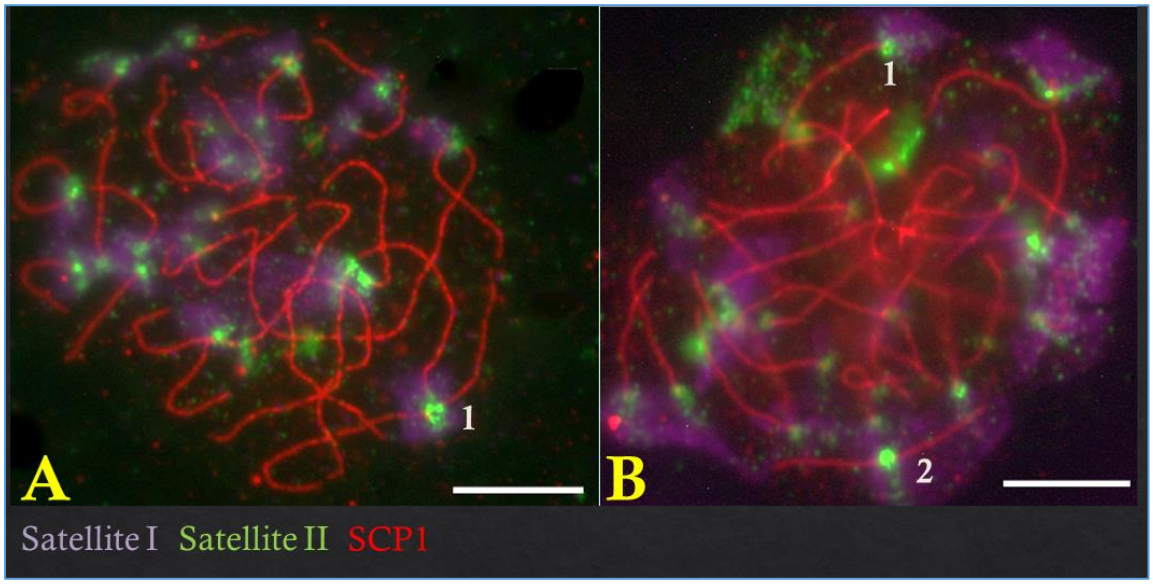


Figure 4.13 A1. Association of SatII sequences together and covered by SatI sequences and both satellites form clusters with 3 SCs. B1. SatII sequences are present at the end of acrocentric SCs while, B2. Prominent FISH signal of SatII is seen in the middle of submetacentric SCs. Furthermore, in general, SatII signals are either separate or fused within the SatI clusters. Scale bar equals 5µm.

4.4.9.3 **Telomeric repeats and relation to satellite I and II**

The (TTAGGG)n telomeric sequence amplified from PCR was used as probe Telomeric_Tndm and hybridized to the ends of acrocentric and submetacentric sheep chromosomes Figure 4.14.

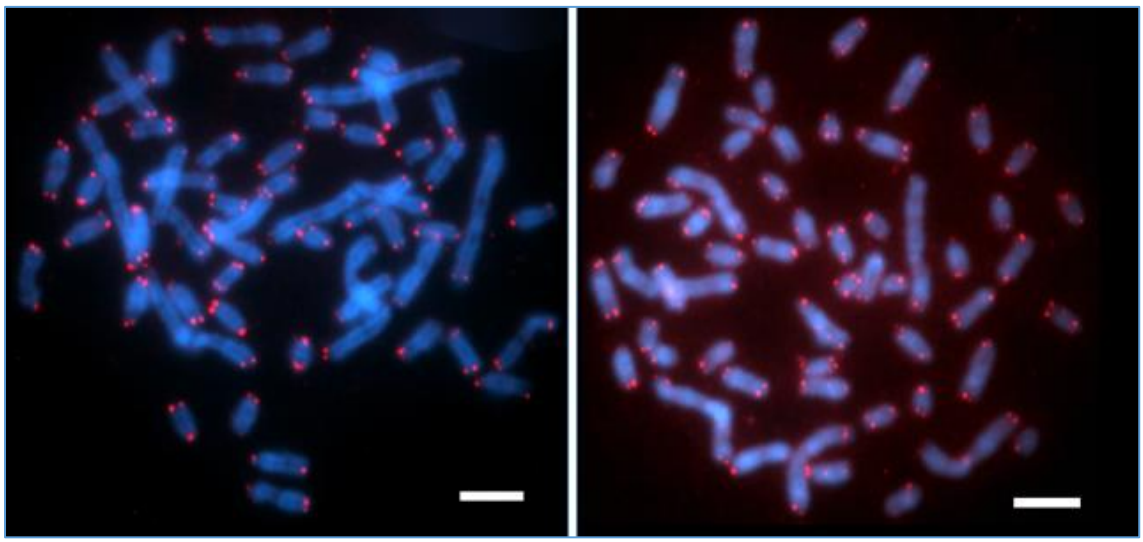


Figure 4.14 Shows that most of acrocentric and submetacentrics in mitosis have double dots at their end, but sometimes only visible on one end. Scale bar equals 5µm.

The probe Telomeric_Tndm was also investigated in meiosis for its relation to the synaptonemal complex. The results indicated the signals of telomeric probes are present at the very ends of both acrocentric and submetacentrics SCs instead of dispersed over

the chromatin loops Figure 4.15. Sometimes one end shows stronger signal than the other end and sometimes only one end shows signal. Interestingly, likewise satellite I and II DNA sequences, telomeric repeat sequences form clusters and are associated with several SCs Figure 4.15B.

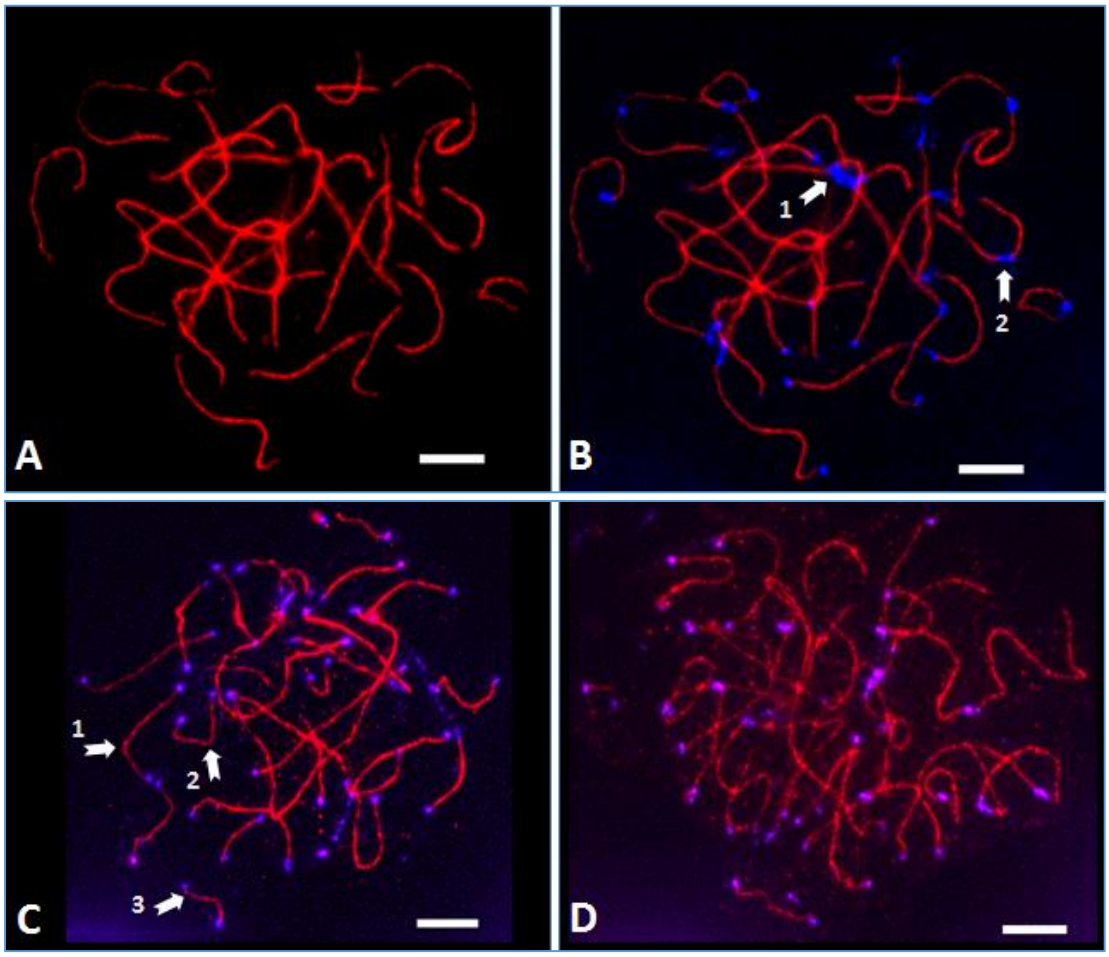


Figure 4.15 FISH and Immunostaining results of sheep SC spreads at pachytene stage using SCP1 rabbit antibody and probe of telomeric sequences Telomeric_Tndm of sheep genome. (A) SCP1 signal detected with Antirabbit Alexa 594 (Seen in red). (B, C&D) Overlay of the signals including SCPI in red and hybridization patterns of the telomeric probe labelled with biotin-16-dUTP detected by Alexa 594 Streptavidin (seen in blue and pink). B1&2 Association of telomeric sequences together and form like clusters with 2-5 SCs. C1, 2&3C Telomeric repeat sequences are seen at the ends of SC. Scale bar equals 5µm.

Furthermore, telomeric probes were also used simultaneously with each of satellite I and II to investigate the relation of the telomeric sequences to the satellite sequences. SatI sequences were as describe above dispersed over chromatic loops and showed no clear association with telomeric sequences Figure 4.16.

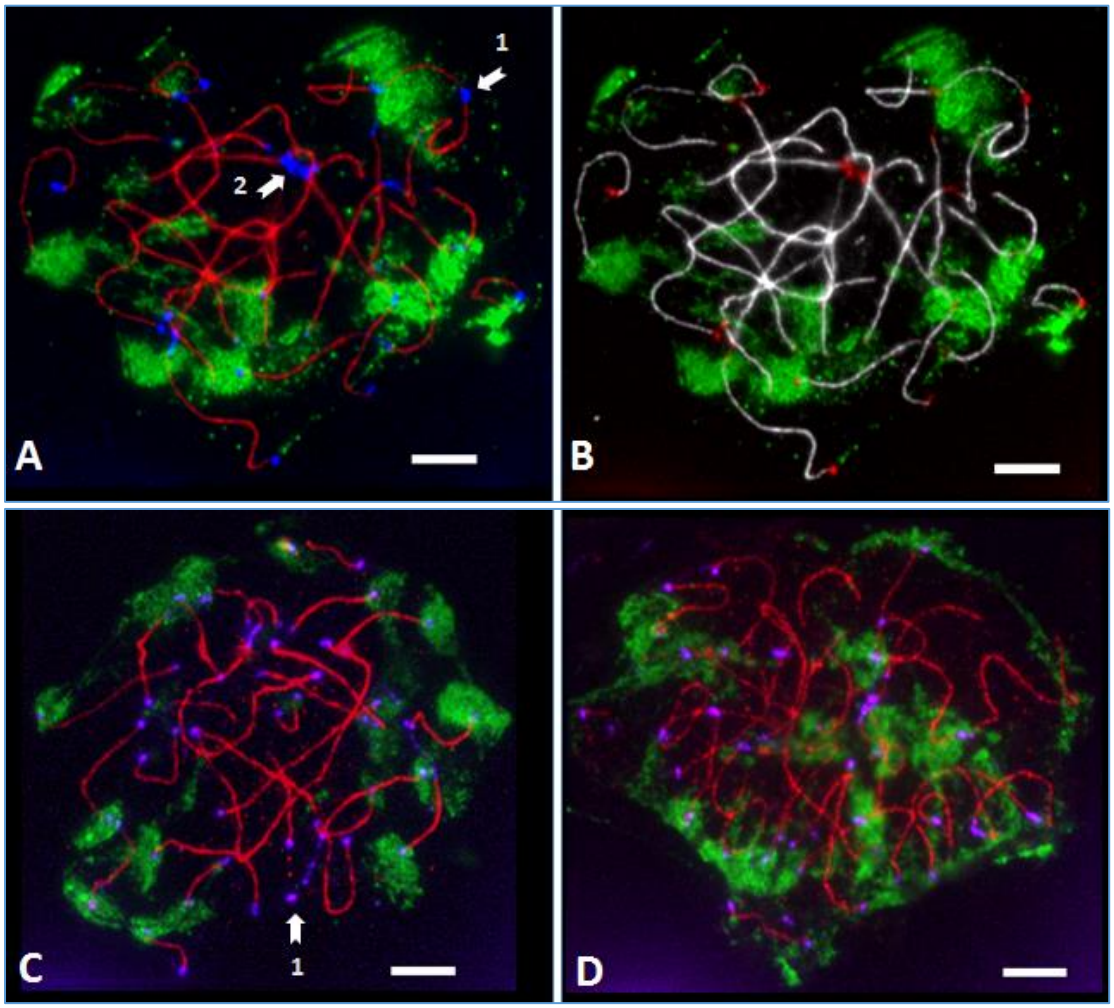


Figure 4.16 FISH and Immunostaining results of sheep SC spreads at pachytene stage using SCP1 rabbit antibody and probe of telomeric Telomeric_Tndm (seen in blue or pink) and satellite I sequences (seen in green). A2 Hybridization patterns of the telomeric probe form cluster and associated with each other between the numbers of acrocentric SCs or present at the end of SCs (A1). While, hybridization patterns of the SatI are dispersed and not associated neither with the SCs nor with telomeric motifs. Scale bar equals 5µm.

However, satellite II sequences were hybridized more strongly to the end of SCs and are associated closer to the telomeric repeat sequences than satellite I sequences Figure 4.17.

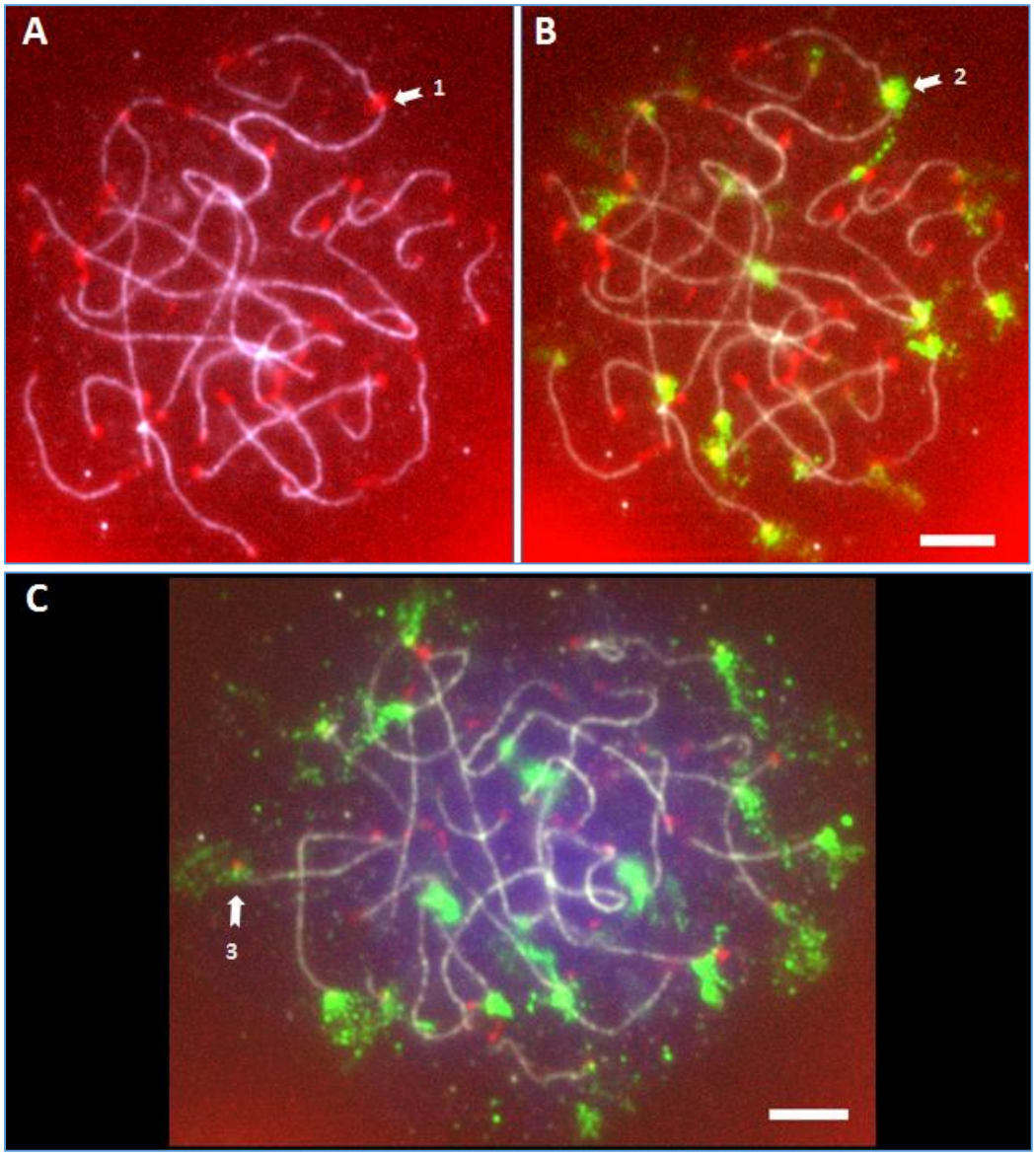


Figure 4.17 FISH and Immunostaining results of sheep SC spreads at pachytene stage using SCP1 rabbit antibody and probe of telomeric Telomeric_Tndm (seen in red) and satellite II sequences cloned_SatII_535 (seen in green). A, B&C Overlay of the signals including SCPI (seen in white). A1&B2 Association of two telomeric sequences together with SatII. C3 Telomeric repeat sequences at the end of acrocentric SC next to SatII sequences. Scale bar equals 5µm.

4.4.10 **Chromosomal organization of putative novel satellites**

4.4.10.1 **Chromosomal characterization of Novel Tandem_44bp repeat**

The Novel Tandem_44bp repeat was identified from RepeatExplorer, *k*-mer frequency and TAREAN analysis. A 31bp long conserved part (GCCCCACCCGGAAATCACGTGGGCCCCACGG) was designed from the reconstructed monomers and a direct oligo probe labelled with 6-fluorescein amidite (6-FAM) (Sigma-Aldrich), used for FISH. The probe hybridized to the centromeres of approximately three

quarters about 40 chromosomes of male metaphase sheep. Signals were strong on some acrocentrics, while very weak or no signal was seen on others. Probe Novel Tandem_44bp neither hybridized to the sex chromosomes X and Y nor to the three pairs of submetacentrics Figure 4.18. The probe was also used against cattle chromosomes and no signals were detected.

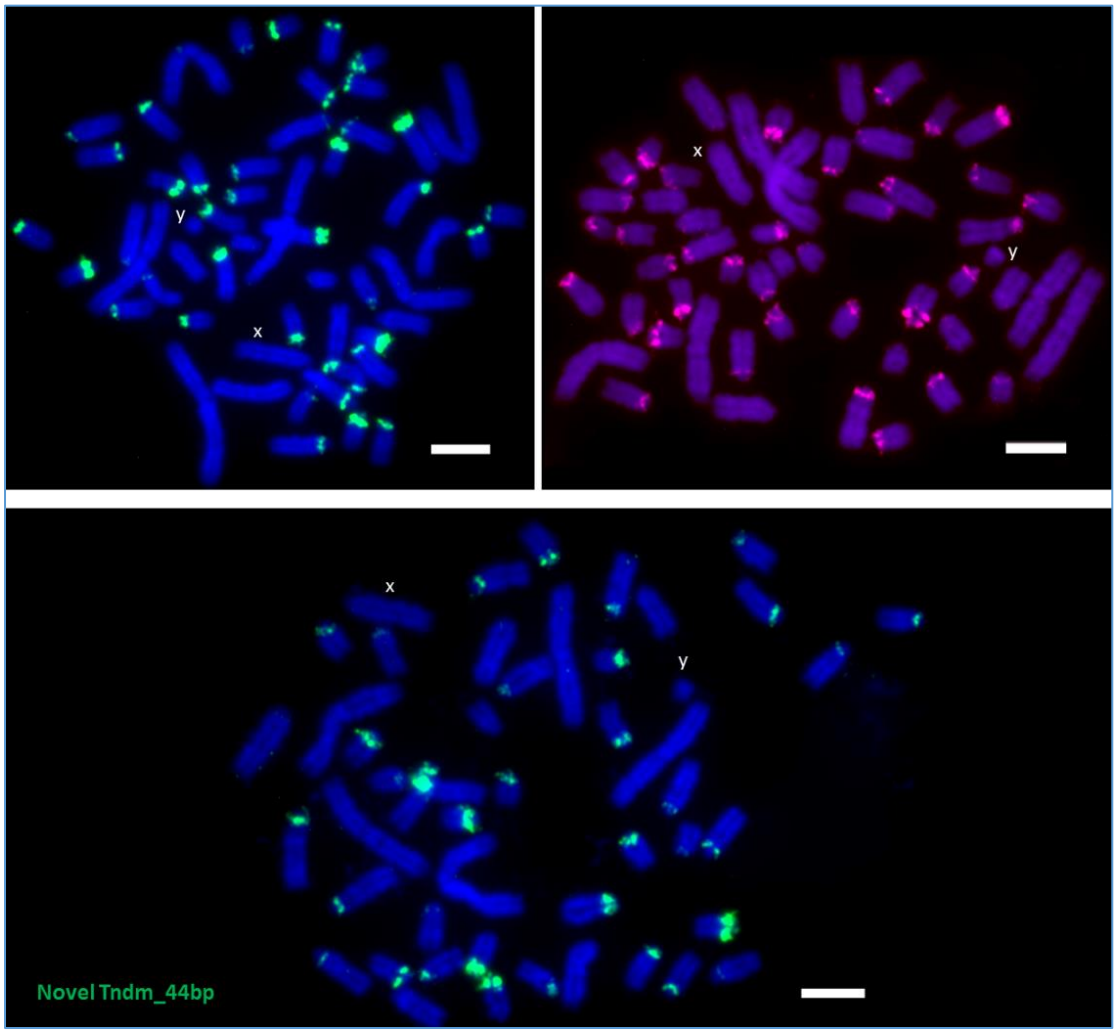


Figure 4.18 FISH of probe Novel Tndm_44bp to metaphase spreads of male sheep chromosomes (*Ovis aries*; 2n=54, XY). Chromosomes were stained with DAPI (seen in blue). Hybridization patterns of the probe Novel Tndm_44bp (seen either in green or pink). Some acrocentrics show stronger signals than others. Two separate signals can also be seen at the end of some acrocentrics. Scale bar equals 5µm.

4.4.10.2 **4.4.7.2 Chromosomal characterization of CL22C4_Sat repeat and ERV2 sequences**

The purified PCR products of both sequences CL22 repeat and EVR2 were labelled and used as probes for *in situ* hybridization. The results confirmed the assembly and

chromosomal organizations of both repeats. Probe CL22C4_Sat hybridized to the centromeres of all acrocentrics and one pairs of submetacentrics, but not to the sex chromosomes. The probe ERV2 was dispersed over all chromosomes including the sex chromosomes X and Y with stronger signal at the centromeres of most acrocentrics and less to the submetacentrics. Overlay of both probes indicated that their sequences are very close to each other at the centromeres Figure 4.19.

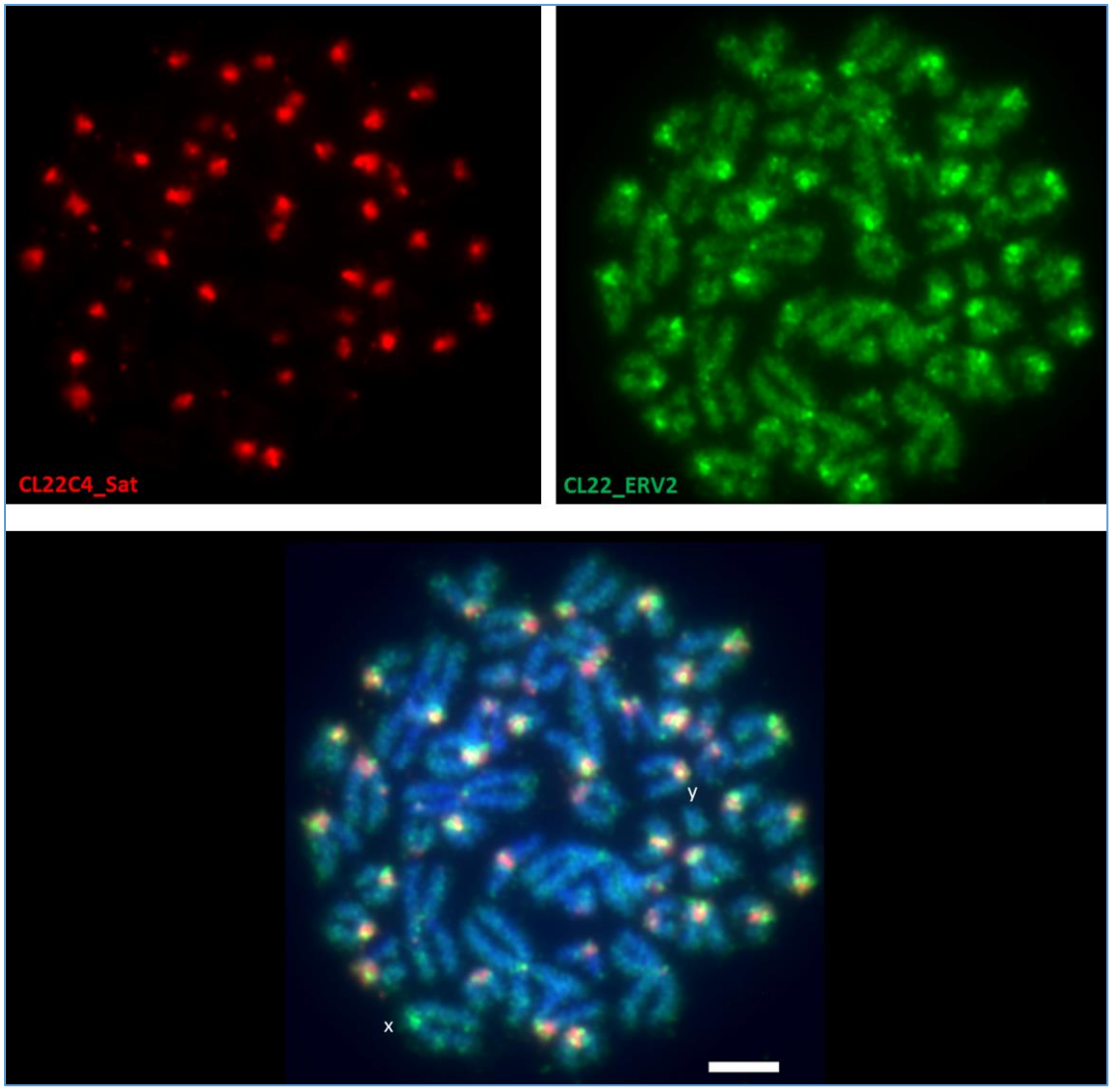


Figure 4.19 FISH of probes CL22C4_Sat (seen in red) and CL22_ERV2 (seen in green) to metaphase spreads of male sheep chromosomes (*Ovis aries*; 2n=54, XY). Overlay of red and green signals showing hybridization of dual probes indicating close position of sequences of both satellites and ERVs. Sex chromosomes X and Y having ERVs but not satellite sequences. Scale bar equals 5µm.

4.4.10.3 **Chromosomal characterization of Putative Sat_716bp repeat**

Two probes of this repeat named Putative Sat_716bp_NSR8 (270bp) and Putative Sat_716bp_NSBL (616bp) were used for FISH. Probe Putative Sat_716bp_NSR8

hybridized to the centromeres of all acrocentric chromosomes except two of them. In addition to centromeric locations, signal was dispersed over submetacentrics mostly on one arm. Signals were undetectable on the Y but dispersed on X chromosome Figure 4.20. While, probe Putative Sat_716bp_NSBL showed slightly additional intercalary signals, plus the centromeric signals at the acrocentric and one pair of submetacentrics. Interestingly, Putative Sat_716bp_NSR8 also hybridized to cattle chromosomes but signals were dispersed over all chromosomes including the sex chromosomes while excluded from the centromeric domains.

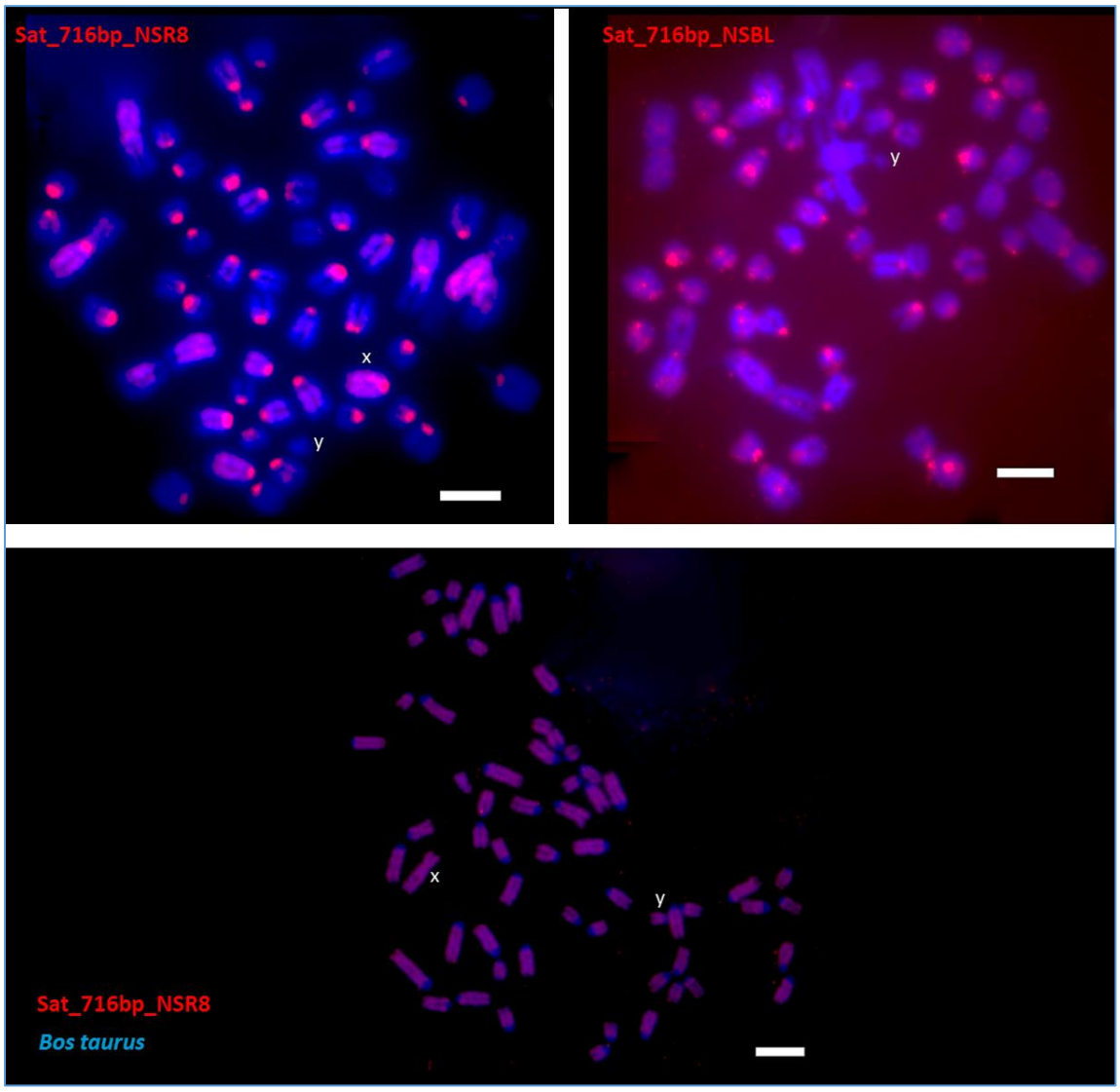


Figure 4.20 FISH of probe Putative Sat_716bp_NSR8 and Putative Sat_716bp_NSBL (seen in red) to metaphase spreads of male sheep and cattle chromosomes (*Ovis aries*; 2n=54, XY). Signals of both probes were centromeric on acrocentrics to dispersed on arms of submetacentrics and X chromosome. However, no centromeric signals were found in hybridization of probes with *Bos taurus* chromosomes. Scale bar equals 5µm.

4.4.10.4 **Chromosomal characterization of satellite like sequences 32merC16_Sat_CRC**

Probe 32merC16_Sat_CRC (508bp) representing the consensus sequence of contig 16 of *k*-mer frequency analysis was used to investigate their abundance on sheep chromosomes. The probe hybridized to the centromeres of all acrocentric chromosomes and showed dispersed bands along the chromosomes. Probe was dispersed over submetacentrics and slightly enhanced at centromeres. Interestingly, the probe signals were centromeric and dispersed along the sex chromosomes X and Y Figure 4.21.

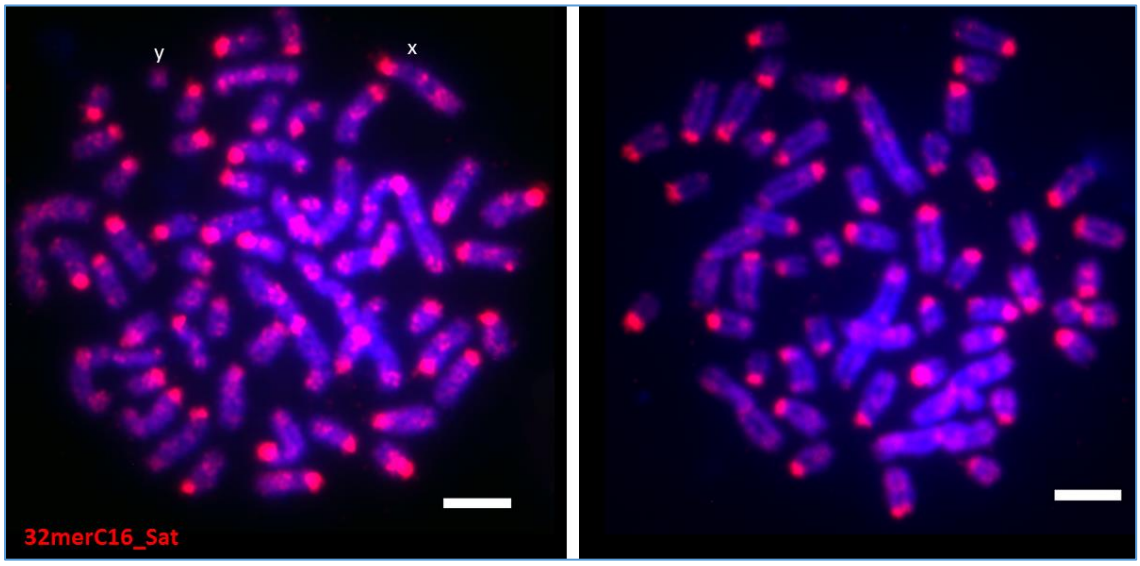


Figure 4.21 FISH of probe 32merC16_Sat to metaphase spreads of male sheep chromosomes (*Ovis aries*; 2n=54, XY). In general, signals are centromeric to dispersed present as 2-3 bands over most chromosomes. Scale bar equals 5µm.

4.4.10.5 **Chromosomal characterization of CL66_TND_Ychr**

From RepeatExplorer analysis specific repeat to *Ovis aries* Y chromosome was discovered in the sequences of CL66_HamJ1 male genome. Sequences of CL66 spanning the tandem repeat region (800bp) of accession (U30306.1) of *Ovis aries* Y chromosome OY9 DNA sequence Figure 4.22 were amplified and labelled as probe CL66_TND_Ychr in order to use for *in situ* hybridization. Probe CL66_TND_Ychr was only and strongly abundant on the Y chromosome of male sheep metaphases Figure 4.23. The probe signal was excluded from the rest of all chromosomes even after extended exposure to capture the images. Furthermore, specificity of CL66_TND_Ychr to male genome was confirmed using bioinformatics analysis. The whole paired raw reads of the KarJ female were

mapped to the sequences of CL66_TND_Ychr and to the *Ovis aries* Y chromosome repeat region OY4 and no reads were assembled Table 4.6.

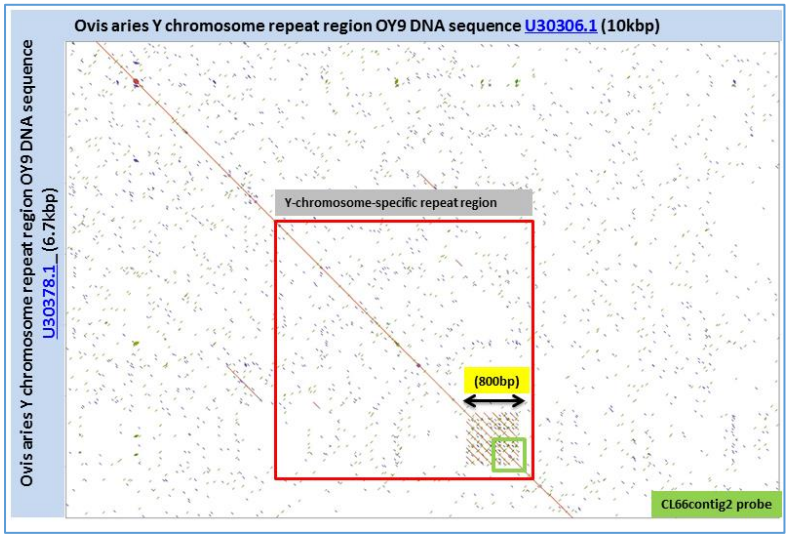


Figure 4.22 Dot plot (self) of scaffold of Y chromosome repeat region OY9 DNA sequence containing CL66 motifs. It shows location of probe CL66_TND_Ychr used for *in situ* hybridization experiment (Green box).

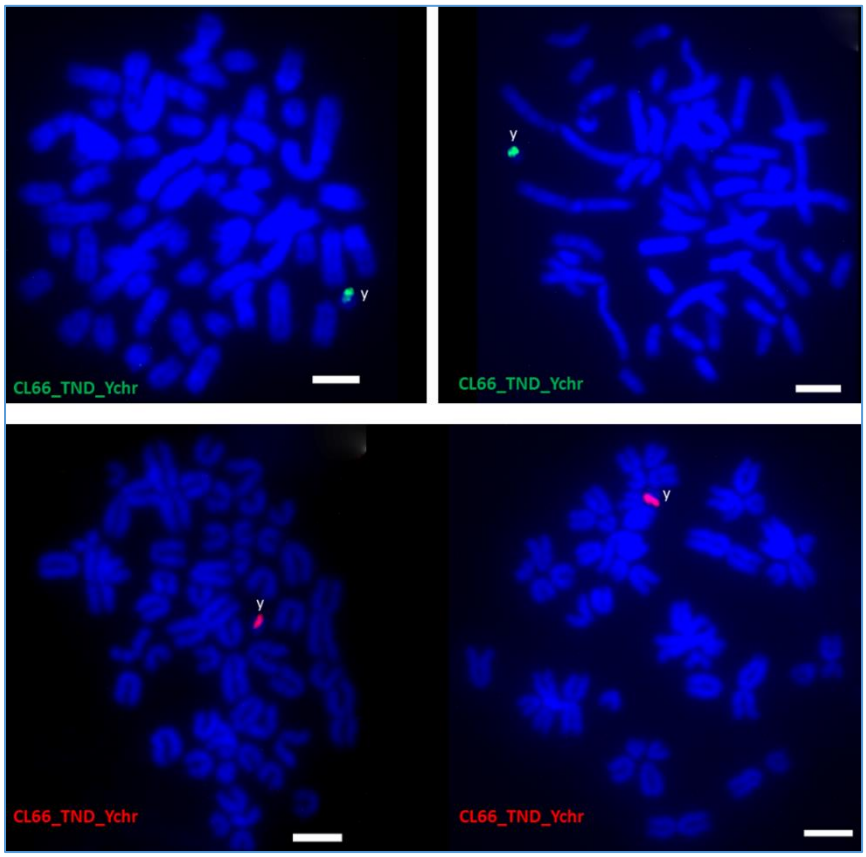


Figure 4.23 FISH of probe CL66_TND_Ychr hybridized to metaphase spreads of male sheep chromosomes. Chromosomes were stained with DAPI (seen in blue). Probe CL66_TND_Ychr sequences was labelled with either biotin-16-dUTP (red signal) or digoxigenin-11-dUTP (green signal). Signal was absent from all chromosomes except Y chromosome. Scale bar equals 5µm.

4.5 Discussion

4.5.1 Major satellite sequences in sheep

4.5.1.1 Chromosomal distribution of satellites I and II

Chromosomal localization of major sheep satellites I and II and bovid satellites I and IV on sheep and cattle chromosomes using different parts of the repeats and differently generated consensus sequences showed variable abundance and distribution of the two satellites in the centromere of autosomal and sex chromosomes (for sheep see Figures 4.8-4.11; for cattle see Appendices 4.11-4.13). Sheep satellite I and II sequences were amplified using different primer sequences based on RepeatExplorer, *k-mers* frequency or published database. Satellite I hybridized with the centromere of all acrocentrics, but with varying intensity. Their signals were undetectable in the largest first pair of submetacentrics, while the other two submetacentric chromosome pairs, have satellite sequences.

Similar to satellite I, satellite II probes hybridized to the centromeres of all acrocentric and submetacentric chromosomes. In contrast to satellite I, satellite II were present on the X chromosome while no signal was observed on the Y chromosome. FISH results of satellite I and II sequences reported in this study agree with the findings of previous studies (Burkin *et al.* 1996; D'aiuto *et al.* 1997; Chaves *et al.* 2003b). Furthermore, more recently, Nieddu *et al.* (2015) observed positive hybridization between satellite I repeat isolated from *Ovis orientalis musimon* and chromosomes from Caprini members.

In this study probes of satellite I and II showed different hybridization patterns and signal intensity over acrocentric, submetacentrics and sex chromosomes. This could be interpreted as each chromosome having a different number of monomer or dimer copies of satellites or that satellite variants are present on different chromosomes that show variable degree of sequence similarities to the probe used for FISH. The later hypothesis is supported by the fact that different bioinformatics analysis tools showed slightly different consensus sequences that gave slightly different FISH signal intensities (Figure 4.24; 32merC3_SatI). Novák *et al.* (2017) designed oligonucleotide probe from the most conserved part of reconstructed monomer of satellite sequences of ***Vicia faba***

based on *k*-mers frequency, and strong signals were detected on chromosomes. Probes of satellite II showed either one strong broader signal or appeared more like two separate bands especially on acrocentric chromosomes (Figure 4.9). D'aiuto *et al.* (1997) also found two independent signals of satellite II sequences on acrocentrics. Submetacentrics also showed different intensities of probe labelling but all less than the acrocentric chromosomes. In agreement to the findings of Burkin *et al.* (1996) who reported that there is different quantities of satellite I sequences on different submetacentrics in sheep as pairs no. 1 and 3 contain lower satellite copies than submetacentric pairs no. 2 (see Figures 4.8-4.11). FISH results found here agree with the findings of Dávila-Rodríguez *et al.* (2009) who found that all pericentromeric regions of acrocentric chromosomes have greater amounts of constitutive heterochromatin than submetacentrics or sex chromosomes.

A likely interpretation of the low copy number of satellite sequences on the submetacentrics would be rearrangement or loss of heterochromatin following Robertsonian translocations. Regarding sex chromosomes, the noticeable and interesting fact is the absence of satellite I signals in both sheep and cattle. However, this is not always the case as satellite I is found on the centromere of X chromosomes in other tribes such as Reduncini, Tragelaphini and Hippotragini (Chaves *et al.* 2005). Thus, satellite I distinguishes Caprini and Bovini from other related tribes. Furthermore, Kopecna *et al.* (2012) indicated significant variation in arrangement and sequence composition of the sex chromosomes between tribes and between species within the same tribe as well.

4.5.1.2 **Genomic proportion of satellite DNA sequences**

Genomic proportion of satellite sequence content in sheep *Ovis aries* had not been estimated so far based on analysis of NGS data. Therefore, the whole sequencing raw reads of genomic DNA from five individuals of sheep breeds were used and allowed us to estimate the abundance of satellite families as a percentage using variety of computational tools such as RepeatExplorer and TAREAN. In regards with satellite I sequences, their genomic proportions in sheep genome was about 5.7-7.4%. Curtain *et al.* (1973) estimated the satellite I sequences as up to 12 % of the sheep total DNA

content. This percentage seems an overestimation in comparison to our results. Possibly, the reasons of different estimation could be the low efficiency of the methods that were used by Curtain. We belief that the advent of next generation sequencing and availability of more advanced bioinformatics tools as used in this study produce more robust analysis and results and that our estimates (Table 4.3, 4.4 & Figure 4.7) are more accurate.

Genomic proportions of the satellite I DNA in cattle comprises about 6-9% (Kurnit *et al.* 1973) and is similar to our sheep estimates and can be explained that in both species satellite constitutes the centromeric heterochromatin of all autosomes except the sex chromosomes XY. Satellite II content, on other hand, showed lower percentages 1.7-2.2% and copy numbers approximately 34-42 thousand copies of female and male sheep (Table 4.3, 4.4 & Figure 4.7). None of the previous research on satellite II sequences estimated their content in sheep genome.

4.5.1.3 **Monomer and array organization of major satellites**

Repeat unit of satellite sequences were described (Buckland 1983; Reisner & Bucholtz 1983; Buckland 1985) and later (Burkin *et al.* 1996; D'aiuto *et al.* 1997; Chaves *et al.* 2000b; Chaves *et al.* 2003b). These studies mainly investigated lengths and abundances of satellite sequences using restriction enzymes and clones. However, in recent years, availability of NGS data alongside the development of bioinformatics and *de novo* assembly tools has improved the analysis of genomic organization of DNA sequences (Novák *et al.* 2017; Utsunomia *et al.* 2017). Thus, in this study, to further understand the structure and organization of the sheep genome of repeat units of satellite sequences, we have investigated whole genome sequencing raw reads to identify monomers and dimers of tandemly repetitive DNA sequences using Tandem Repeat Analyzer (TAREAN) pipeline. In case of sheep satellite I, the most abundant consensus sequence of satellite I monomer was 803bp (Appendix 4.6). Within the resulted cluster of satellite I, several consensuses with different lengths were identified, some up to 816bp Indicating different variants of satellite I. Similar to satellite I, the most abundant consensus of satellite II sequences was 702bp, but seven other consensus sequences with different lengths were also included (Appendix 4.7). Burkin *et al.* (1996) isolated monomers of

satellite I and II fragments by cleaving genomic DNA with restriction enzymes. For example, satellite I fragments produced by EcoRI were 800-830bp and BamHI 430-470bp long, while satellite II fragments isolated by SstI were 420-480bp. Our results do not coincide with Burkin *et al.* (1996) findings because monomers of satellites reconstructed by TAREAN were mainly based on the analysis of *k*-mer frequency of the most frequent short motifs involved in the structure of each monomer. Bioinformatics approaches utilizing *k*-mers frequency such as TAREAN are more appropriate and efficient for monomer reconstruction of tandemly repeated sequences from whole sequencing raw reads and thus an alternative way to using restriction enzymes with genomic DNA (Macas *et al.* 2010; Macas *et al.* 2011; Torres *et al.* 2011; Novák *et al.* 2017; Ribeiro *et al.* 2017)

4.5.1.4 **Junction region between satellite I and II**

In addition to identification of the major ovine satellites, we also identified a putative junction region between satellite I and II in the sheep. As a result of mapping the whole paired end reads to the putative junction sequence, the assembled consensus sequence of the junction region included two monomers of satellite I and three monomers of satellite II (Figure 4.2 and Appendix 4.1). The junction region itself is only 69-99bp long, highly diverged and was characterized by low abundance of 45-110 reads and did not show signals when used as probes for FISH. This means that junction sequences are relatively rare in the genome potentially only occurring one or two times per chromosome.

D'aiuto *et al.* (1997) isolated a phage clone containing putative sequence of junction region including fragments of both ovine satellite I and II sequences. Furthermore, Vissel *et al.* (1992) reported two components of satellite sequences (satellite III & α satellite) in human chromosome separated by direct repeats.

We designed primers (Table 4.1) spanning junction regions and we could amplify sequence across the junction between satellite I and II suggesting contiguous connection between blocks of both satellites I and II (Appendix 4.14). Further studies are needed to clarify whether these junction regions play functional or structural role in mechanisms

that involved in the phenomenon of Robertsonian translocations of ancestral centromere.

4.5.1.5 **Comparison of sheep satellites and satellites from other species**

Both cattle satellite I and IV throughout the bovine autosomal chromosomes displayed signals varying in intensity. Satellite I sequences are found on all acrocentric autosomal cattle chromosomes, while satellite IV probes only hybridized to about two-thirds of the cattle chromosomes (Appendices 4.11-4.13). In our results, no homologous sequences of bovine satellite IV were found in the sheep NGS data. The results of Adega *et al.* (2006) determined that satellite IV sequences on the autosomal and sex chromosomes could be independent in terms of their evolutionary pathways in the species of Tragelaphini and Bovini. Hybridization strength of two probes SatI-2 and SatI-4 cover different parts of the cattle satellite I repeat unit was similar between chromosomes indicating that the repeat unit was mainly amplified as a whole during the evolution of cattle despite showing different sequence similarities along the repeat unit to other species within Bovidae family. Nieddu *et al.* (2015) observed positive hybridization between the Bovini satellite I sequences and chromosomes from members of the Bovini tribe. In terms of cross hybridization, the results of Kopecna et al (2012) showed negative hybridization signals of cattle satellite I clone- BTREP15 (581bp) against sheep metaphase when they used two different post-hybridization washes; low and high stringency. Our results confirm the findings of Kopecna *et al.* (2012) as no raw reads of the NGS data of sheep genome mapped to the entire 1402bp satellite I sequences of cattle (Table 4.5). Furthermore, no FISH signals were detected with sheep satellite I and II sequences against cattle chromosomes as also reported by (Nieddu *et al.* 2015). Comparison of satellite I and II sequences of sheep and other two species cattle and buffalo showed an overall 55-66% sequence identity confirming the negative FISH results at stringencies of 70-75% used here.

In this study, the whole sequencing raw reads of sheep were assembled to satellite I and II sequences of other species in4cluding *Ovis ammon*, *O. dalli*, *O. canadensis*, *O. aries musimon*, *Oryx gazelle*, *Ammotragus lervia*, *Cephalophus natalensis and Capra hircus*. However, no reads were assembled to satellite I and II sequences of bovine and buffalo

(Table 4.5). As expected from the map to reference, phylogenetic relationship between satellite I sequences of sheep and the other species demonstrated coherent results with our findings of mitogenome haplogroups (see Chapter Three). For instance, domestic and wild sheep grouped within the same clade. While, bovine and antelope fitted away from sheep (Appendix 4.4). Chaves *et al.* (2000a) and Kopecna *et al.* (2014) concluded that satellite DNA sequences could be used as valuable markers of phylogenetic relationships between species originated from different tribes of Bovidae family.

4.5.2 Meiotic behaviour of major satellite sequences and telomeres in sheep

Synaptonemal complex (SC) spreads were prepared using detergents and surface tension to distribute the chromatin and allow access to the proteins of the SCs (Schwarzacher *et al.* 1984). There are several advantages over hypotonically spread pachytene and somatic metaphase chromosomes: higher spreading forces are applied and therefore the individual chromosomes are well separated and the chromatin is relatively de-condensed and free of debris making SC spreads highly accessible for FISH probe penetration. Furthermore, important facts about meiotic chromatin organization can be studied. SCP1 is one of the main proteins of the SC establishing the transverse filaments that connect the lateral elements; it, thus, plays a critical role in the assembly of the SC and is indispensable for synapsis (Yuan *et al.* 2000; de Vries *et al.* 2005). Using SCP1 antibodies, different meiotic prophase stages could be identified and in particular pachytene as SCs are entirely synapsed with continuous CEs and hence string like SCP1 signals (Figures 4.12 & 4.13). FISH combined with immunostaining of SCP1 was used to investigate the meiotic behaviour, organization and association of satellites I and II in respect to the SC using testicular materials of rams.

4.5.2.1 Relation of satellite sequences to the synaptonemal complex

A significant difference of repeat sequence location was observed for satellites I and II. FISH signals of satellite I were dispersed and mainly located in the chromatin loops that radiate out from the SC. While satellite II were closely associated with the SCs and often seen within the cloud of satellite I signal (Figures 4.12 & 4.13). Finding different association to the SC is not unusual as Moens and Pearlman (1990) showed for the

mouse where they found that the major satellite hybridized to the chromatin loops while the minor satellite sequences were mainly localized at the SC of the centromeric region. In rats, satellite I sequences and in human classical satellites 1qh&9qh and centromeric alpha-satellites, were also found in the loops and only associated with SCs at their bases (Moens & Pearlman 1989; Barlow & Hultén 1996). Furthermore, Hernández-Hernández *et al.* (2008) performing, FISH combined with immunolocalization of SYCP3, showed clear association of satellite repeats to the lateral elements of SC. Further evidence of repetitive DNA sequences playing an important role in homologous pairing (Schwarzacher 2008) and synapsis through DNA-protein bindings sites that anchor the chromatin loops to the lateral elements comes from studies in Dropsophila (Noreen *et al.* 2007) and *C. elegans* (Phillips *et al.* 2009).

4.5.2.2 **Relation of satellite sequence and telomere**

Satellite II sequences were clearly attached to the acrocentric SCs at a more distal position than satellite I supporting the results from FISH to somatic metaphase chromosomes (Figure 4.13B *vs* Figure 4.10C; and (D'aiuto *et al.* 1997; Chaves *et al.* 2003b)). Therefore, we tested the relation of both probes to the telomere sequence (TTAGGG)n and tight signals of telomeric probes were present at both the ends of acrocentric and submetacentrics SCs (Figure 4.14 & 4.15). Our results are compatible with the findings of Moens and Pearlman (1990) showing that the telomere sequences were situated at the ends of each SC in the mouse and were not distributed in the chromatin loops. No association of the telomeric sequences with the dispersed signal of satellite I sequences were observed in sheep (Figure 4.16), but as expected for their more distal location, satellite II sequences hybridized strongly at the end of SC and were associated closely to the telomeric repeat sequences (Figure 4.17). Similarly, to the sheep, Santos *et al.* (2004) the major satellite DNA family of the domestic cat (FA-SAT) being co-localized to the telomeric regions of cat chromosomes and in the mouse, Kipling *et al.* (1991) found that the minor satellite DNA family was physically associated to the proximal telomere. Satellite II sequences could therefore form a specific class of telomere associated repetitive DNA sequences (TAS) that have been described for plants and animals and that form a bridge between the proper telomere and distal chromatin often containing degenerate TTAGGG repeats (see Contento *et al.* (2005).

Association of telomeric sequences with the SC possibly reflects interaction between the SC components and the telomere-associated proteins and attachment to the nuclear envelope. For example, ring chromosomes in male mouse meiosis require telomere repeats in order to localize successfully to the nuclear periphery of spermatocytes (Voet *et al.* 2003). Similarly, the repetitive telomeric DNA sequences in human spermatocytes were found tightly associated with the SCs suggesting their kinetic properties in relation to the nuclear membrane (Barlow & Hultén 1996). Scherthan (2007) indicated contribution of telomere repeats in clustering of meiotic telomeres that occurs universally in many organisms at the zygotene bouquet stage of meiotic prophase (see Sepsi *et al.* (2017).

4.5.2.3 **Association of acrocentric heterochromatin during meiosis**

Association of acrocentric chromosomes was observed during meiotic prophase as evidenced by few and large signals for satellite probes. Several SCs can be seen to be associated with the same large signal (Figures 4.12 & 4.13), but it is notable that the SCs themselves do not seem to associate with each other. The satellite sequences are part of the pericentromeric heterochromatin, and it is likely that it is the repeated nature of these sequences that is responsible for the heterologous chromosome associations. This was reported for the domestic pig where a large single chromocentres was observed for the centromeric heterochromatin of the acrocentric chromosomes, but not the metacentric chromosomes (Schwarzacher *et al.* 1984; Jantsch *et al.* 1990). The phenomenon was explained by the clustering of telomeres at the meiotic bouquet stage and the resulting proximity of acrocentric centromeres that facilitated DNA sequence homogenization of the satellite sequences of the acrocentric heterochromatin. A similar mechanism can be expected in the acrocentric chromosomes of sheep; thus, meiotic arrangements play crucial role in homogenization events and explain the relatively high similarity scores for satellite I and II sequences (see Results). However, some sheep satellite sequence variants and nucleotide differences in consensus sequences using different bioinformatics tools were found with variable distribution amongst chromosomes and chromosome specific variation is expected as indicated by not all acrocentric chromosomes being involved in a single chromocentre as observed in pig. It is notable that the satellite DNA sequences of sheep have undergone several diverging

and homogenization events since the not so distant evolutionary split of *Ovis aries* and *Bos taurus* and cooperates the notion that concerted evolution is common and crucial for genome evolution as is leads to homogenization of tandemly repeated sequences (see Kuhn *et al.* (2011).

4.5.2.4 **Submetacentric chromosomes at meiosis**

Satellite II sequences, but not satellite I sequences were seen at the larger submetacentric chromosomes roughly in the middle (Figures 4.12 & 4.13). The presence of only satellite II agrees with the origin of submetacentric chromosomes of *Ovis aries* that were produced by centric fusion of acrocentric chromosomes. As described above the proximity of acrocentric centromeres at meiotic prophase could facilitate such Robersonian translocations and the contained repetitive DNA sequences enhance the associations and recombination of heterologous chromosomes (Pathak *et al.* 1982). Signals of satellite I sequences, on other hand, were undetectable at the submetacentric SCs. Robertsonian translocations could be one of the main reasons that caused to the loss of the satellite I sequences at the largest submetacentric chromosomes (Chaves *et al.* 2003b).

4.5.3 **Novel repeats**

Novel putative satellites were discovered using RepeatExplorer, *k*-mer, and TAREAN pipeline. TAREAN has been proved to be highly reliable computational pipeline due to its ability to identify previously characterized satellite repeats and novel tandem repeats from high throughput sequencing raw reads (Novák *et al.* 2017). Alkan *et al.* (2011) defined candidate monomers of satellite repeats in some mammalian species using RepeatNet an *ab initio* algorithm for detection of centromeric sequence and demonstrated their centromeric locations on chromosomes.

4.5.3.1 **Novel Tandem_44 repeat**

According to our results, this repeat was discovered with a genomic proportion 1.1% and 0.77% in male and female genomes respectively. FISH results indicated that probe hybridized mostly to 40 acrocentric chromosomes (Figure 4.18) but not the sex chromosomes and the three pairs of submetacentric. The repeat has a very interesting

structure (Figure 4.3): it is relatively short but present in many copies in the genome and its sequence analysis indicates that two shorter similar but not identical monomers of 22bp have formed a higher order structure of 44bp in an A-A* structure that then was tandemly amplified. Similar repeat structures have been found in Drosophila (Kuhn *et al.* 2008) and are probably due to rolling circle amplifications.

The Novel Tandem_44bp repeat was not found in cattle chromosomes. Thus, it is very likely that it is a specific satellite DNA to the sheep genome. It is not unusual to find species specific satellite DNAs and Kopecna *et al.* (2012) described it for *Oreotragus oreotragus* of Antilopinae subfamily that was absence in other bovid species. We have also found that the middle and the last parts of satellite I of *Cephalophus natalensis* are more specific to its own genome as none of whole raw reads of sheep were assembled against them. We also investigated the abundance of sheep Novel Tandem_44bp repeat in whole sequencing raw reads of *Ovis canadensis*, and only very few hits were found. Hence, we concluded that the Novel Tandem_44bp repeat is common to the *Ovis* genus, but has been evolved since the split of *O. aries* from *O. canadensis*.

Further studies of *in situ* hybridization of this novel repeat over chromosome of wide range of wild sheep and goat species would expand the knowledge about evolutionary events such as diversification and homogenization of this tandem repeat.

4.5.3.2 **Satellite sequences CL22 harbouring endogenous retroviruses_ERV2**

Combined sequences of satellite and ERVs were found in CL22. Probe CL22C4_Sat hybridized to the centromeres (Figure 4.19). Due to the incorporation of ERV2 within the CL22 repeat and because of low sequence identities to the major sheep satellites, it is more likely that CL22C4_Sat sequences are a new putative satellite family in sheep genome. Incorporation of endogenous sequences within tandem repeats is not uncommon. Alkan *et al.* (2011) identified a 528bp centromeric satellite in the gray short-tailed opossum, which is an endogenous retrovirus repeat, and hybridized at the centromere of homologous chromosomes of opossum. This discovery of ERV elements at the centromeres of opossum might indicate the integration as a mammalian ancestral state (see Chapter Six).

4.5.3.3 **Novel Putative Sat_716bp repeat and 32merC16_Sat_CRC**

Graph-based read clustering indicated that some clusters with low genomic proportions have sequence similarities to satellite DNA families. Only 45-50% sequence identities were found between sequence of this novel satellite and satellite I and II sequences of sheep. However, it showed higher sequence similarities of 60% to centromeric repetitive DNA of Cervidae species. The Cervidae family including species of elk and deer is supposed to have diverged from Bovidae family about 25.5-27.8 million years ago, (Wu *et al.* 2012). Similarly, satellite I, II and III DNA sequences with different repeat unit have been identified in several species of Cervid families such as in Roe deer, Caribou, Red deer, White tailed deer and Indian muntjac (Scherthan 1991; Lee *et al.* 1994; Qureshi & Blake 1995; Lee *et al.* 1997; Buntjer *et al.* 1998; Li *et al.* 2000a; Li *et al.* 2000b). They observed extremely high sequence conservation among the satellite DNA clones produced from four deer species of Cervid family. However, in our study the sequence consensus of monomers and dimers of the Novel Putative Sat_716bp repeat was more diverged. FISH signals were localized over centromeric domains of all acrocentrics except two of them were dispersed over submetacentrics with some intercalary positions (Figure 4.20). Signals were undetectable on the Y but more likely dispersed on X chromosome. Interestingly, one of the probes used for FISH also hybridized to the cattle chromosomes but signals were dispersed over all chromosomes including the sex chromosomes and excluded from the centromeric regions (Figure 4.20). This is due to the low sequence identities 50-56% of the 716bp repeat to both satellite I and II sequences of bovine. We speculate that the sequence has evolved before the split of the Bovidae and Cervidae families, but have diverged significantly since and also have been incorporated into dispersed elements.

Another satellite like sequence 32merC16_Sat_CRC was identified using *k*-mers. In comparison to Repbase databases, 32merC16_Sat_CRC showed highest sequence similarities 60% to the centromeric repetitive DNA from Cervidae species and to sheep satellite II sequences. Sequences of 32merC16_Sat_CRC hybridized to the centromeres of all acrocentric and the sex chromosomes and also showed dispersed like bands along the euchromatin (Figure 4.21). (Kuhn *et al.* 2011) investigated that short arrays of satellite DNA III sequences (1.688) of Drosophila melanogaster can be dispersed over

euchromatic domains. Furthermore, it has been stated some satellite sequences that found in transposable elements such as retrotransposons are originated through amplification of short arrays of tandem repeats. Similarly, (Macas *et al.* 2009; Novák *et al.* 2017) indicated that satellite repeats could be present as dispersed short motifs in genome. From sequence and chromosomal characterization, we could assume such events as amplification and homogenization contributed in this satellite sequence.

4.5.3.4 **Specific tandem repeat to male genome of sheep CL66_TND_Ychr**

Next generation sequencing technologies, bioinformatics tools and cytogenetic techniques allowed us to identify and localize a specific repeat to Y chromosome of sheep. This repeat was discovered in the RepeatExplorer analysis of male genome. We found CL66_TND_Ychr probe was only abundant and strongly hybridized to Y chromosome of male sheep metaphase (Figure 4.23). Furthermore, *in situ* results were confirmed by mapping the whole paired raw reads of the KarJ female sheep genome to the consensus of CL66 and no reads were assembled (Table 4.6). Our results are in agreement to findings of Pertile *et al.* (2009) where they found specific sequences to Y centromere of house mouse. Y chromosomes have been found to have different structures to autosomal chromosomes because of the small pairing region with the X chromosomes and different recombination events and gene content see Graves *et al.* (2006). Hence, our findings in sheep confirm that mammalian Y chromosomes are characterized by a different model from autosomes and different satellite repeats accumulate.

Interestingly, when we investigated the abundance of Y chromosome scaffold repeat region containing CL66_TND_Ychr sequences of sheep (*Ovis aries*) in whole sequencing raw reads of *O. canadensis*, we were able to assemble consensus of Y chromosome repeat region of *O. canadensis* including highly conserved sequences with sequence identities 97.5%. These results indicate that Y chromosomes of domestic and wild sheep are more likely share the same common parental ancestor. Presence of SNPs indicates rapid evolution since their speciation time. Moreover, these high sequence similarities between *O. aries* and *O. canadensis* could be used as an indicator to estimate divergence time between domestic and wild sheep. Thus, to understand structure, organizational

architecture for the sheep Y chromosome, a comparative analysis of Y chromosome sequences between sheep and related species within Bovidae family is required. This could provide insights into the evolution and divergence of Y chromosome in closely related species of Bovidae family.

4.6 Conclusions

The combination of detailed DNA sequence using novel bioinformatics tools of NGS raw reads with analysis of chromosomal distribution at mitosis and meiosis allowed a detailed description of the repeat landscape of the sheep genome. Classical major satellite sequences and novel repeats were described. Sequence similarity, polymorphisms (transitions & transversions) and indels were found between nucleotide sequences of monomers and dimers of each major satellites and novel satellites. Consensus sequences of monomers and dimers of satellite I, II and Novel Tandem_44 repeat were most conserved with sequence identities 90-100%. However, consensus sequences of monomers and dimers of another novel putative satellite_716bp were more diverged with sequence similarities of 50-100%. Repeat unit array of the major satellites were characterize by different lengths and copies of monomers along each of acrocentric, submetacentrics and sex chromosomes and is likely to include chromosome specific variants and could be one of the main reasons about presence, absence and intensity of signals of major and novel satellites on each chromosome observed. Burkin *et al.* (1996) suggested that centromeres of each of acrocentric, submetacentrics and sex chromosomes appeared to contain different subfamilies of satellite I and II sequences. Furthermore, in terms of monomer activities in sheep genome, some monomers could perform vital functions in centromere domains. Li *et al.* (2002) identified new satellites in species of Cervidae family and their sequences were localized with centromeric proteins at the kinetochore, suggesting functional role of such satellite sequence due to its close association with kinetochores. Cerutti *et al.* (2016) isolated three satellite sequences in horse genome and proved that satellite 37cen (221bp) is transcriptionally active using the ChIP-seq methodology.

High sequence identities between the consensus of monomers of major and novel satellites indicate strong homogenization events that we speculate are facilitated by the

association of acrocentric centromeres at meiotic prophase when chromosomes undergo reciprocal recombination and similar repeats on heterologous chromosomes can exchange sequences. Different amplification and homogenization mechanisms including unequal crossing over and rolling circle mechanisms giving rise to higher order repeat structures, often act together on tandemly repeated sequences and therefore lead to differences in their genomic organization impacting on diversification and concerted sequence evolution and in turn speciation (Kuhn *et al.* 2008; Plohl *et al.* 2008; Richard *et al.* 2008; Kuhn *et al.* 2010; Kuhn *et al.* 2011; Kopecna *et al.* 2014; Garrido-Ramos 2015) and is demonstrated here in sheep.

Chapter 5 Transposable elements and dispersed repetitive DNA sequences in the sheep genome: their nature, diversity and distribution

5.1 Introduction

In mammalian genomes, the repetitive component contains a high proportion of transposable elements, comprising class I elements (retrotransposons amplifying through an RNA intermediate) which commonly represent about half of the genome, and the less abundant class II elements (DNA transposons) (Wicker *et al.* 2007; Elsik *et al.* 2009; Pagán *et al.* 2012; Biscotti *et al.* 2015b). Both classes are typically widespread in the genome, with regions along chromosomes of higher and lower abundance. Identification, quantification and characterization of all dispersed repeats, mostly consisting of transposable elements, is challenging.

Genes show allelic variation between individuals, breeds or related species and may be duplicated or deleted. Various repetitive DNA motifs can show diversity at different taxonomic levels (e.g. in Drosophila, Kuhn *et al.* (2008)), providing data to characterize evolutionary and diversification events, perhaps leading to isolation and speciation. There are clear differences in large-scale genome organization between plants and animals (Heslop‐Harrison & Schwarzacher 2011; Biscotti *et al.* 2015b), not least in chromosome-specific fractions enabling chromosome painting in animals (Ferguson-Smith *et al.* 2005) and small numbers of highly abundant centromeric tandem repeats (see Chapter Four). There are relatively few studies of the nature and relevance of other types of DNA repeats in animal genomes (Adelson *et al.* 2009; Alkan *et al.* 2011; Gouveia *et al.* 2017).

Repetitive DNA elements have been studied in several mammalian genomes with a variety of methods, including identification of clones with high-copy sequences, or targeted PCR amplification based on conserved motifs. However, these approaches were not efficient to identify all types of repetitive sequences. High-throughput whole genome sequencing included many transposable element reads, but most analyses aim

to remove the repetitive sequences and focus on the low copy fractions including genes and regulatory elements. There has been a need for bioinformatics tools for investigation of DNA repeat families in large numbers of short reads (<300bp). Novák *et al.* (2010) developed an important graph-based approach to cluster raw reads with sequence similarity, then comparing consensus regions of the clusters with databases to identify any homologies to known repeat sequences. A second approach to repeat identification involves measuring the frequency of each short DNA motif *k* bases long (*k*-mers) in raw reads, and analyzing the motifs which are repeated most frequently.

5.2 Aims and objectives

The current study aimed to

1- Identify dispersed repeats from whole genome sequencing raw reads (NGS data) of the sheep genomes using the complementary approaches of graph-based clustering and *k*-mer frequency analysis.
2- Characterize the nature, measure the abundance, and find sequence diversity of dispersed repeats.
3- Characterize chromosomal locations and organization of major dispersed repeat elements using *in situ* hybridization.
4- Explore features of selected repeats including those with chromosome-specific amplification and relate repetitive elements to genome or chromosomal evolution.

5.3 Materials and methods

5.3.1 Primer design and PCR amplification

Consensus assemblies of contigs resulting from RepeatExplorer and *k*-mer frequency analyses were selected for designing primers Table 5.1 for PCR amplification. Amplified PCR products were purified, labelled and used as probes for *in situ* hybridization experiments. Parameters and cycling conditions of PCR amplification, probe labelling and *in situ* were carried out following sections 2.2.2, 2.2.3, 2.2.5 and 2.2.11. Identification of dispersed repetitive DNA sequences was followed section 2.2.13. The dispersed repetitive DNA sequences will be submitted to the Repbase databases.

Table 5.1 Primer sequences used for PCR amplification of consensus of RepeatExplorer clusters and *k*-mer contigs. Amplified PCR products were labelled and used as probes for *in situ* experiments.

Probe names	Name of primers [Sequence (5'-3')]	Product size (bp)	Annealing temp.
CL5C418_RTE	F= ATTGCACTCATCTCACATGC	500	56
	R= GTATCATCCGTGTATCTGAGG		
32merC15_RTE	F= TTTTTGAACTGTGGTGTTGG	324	56
	R= CATGAATTGCAGCACACC		
CL5C464_RTE	F= TTAGAAAAGGCAGAGGAACC	607	56
	R= AGTTAGTTGTGATCCACACC		
CL8C129_RTE	F= GCCTTCTTATCTCTCCTTGC	481	56
	R= CTCTACACATGGACATCACC		
CL10C107_RTE	F= TGATCCACACAGTCAAAGGC	450	64
	R= TCAAGTGGGCCTTAGGAAGC		
32merC12_RTE	F= AATAGATGGGGAAACAGTGG	183	56
	R= ACTAGATGGACCTTTGTTGG		
CL8C95_RTE	F= AATTCGTACCTTGAGAACCC	562	54
	R= GTCAACTTCAGCTTCTTTGG		
CL7C43_RTE	F= GAGGTTCGTGACATTGTACAGG	470	64
	R= GCCCACTTGACTTCACATTCC		
32merC31_RTE	F= CAGCAAACTAAAATGGACTGG	258	56
	R= TAGAACCGTTCAACTTCAGC		
CL4C63_RTE	F= TGCAGTGATTTCCAAGCCCC	489	64
	R= GTGTGGATCACAAGAAACTGGG		
CL12C16_L1-3	F= TAACTCCCAGTAACCATTGC	543	56
	R= GCACCATTTATTGAAGAGGC		
CL12C27_L1-3	F= GAGGTTACAACAGACAATGC	362	56
	R= GTTGGTATCAGGGTGATGG		
CL12C2_L1	F= AGAAAACAGGCATAAAAGGG	303	54
	R= GCTCTAGTAATTTTCTGGTGG		
CL25C6_L1-2	F= AGGATAGTTAGGGAATTTGGAATGG	180	64

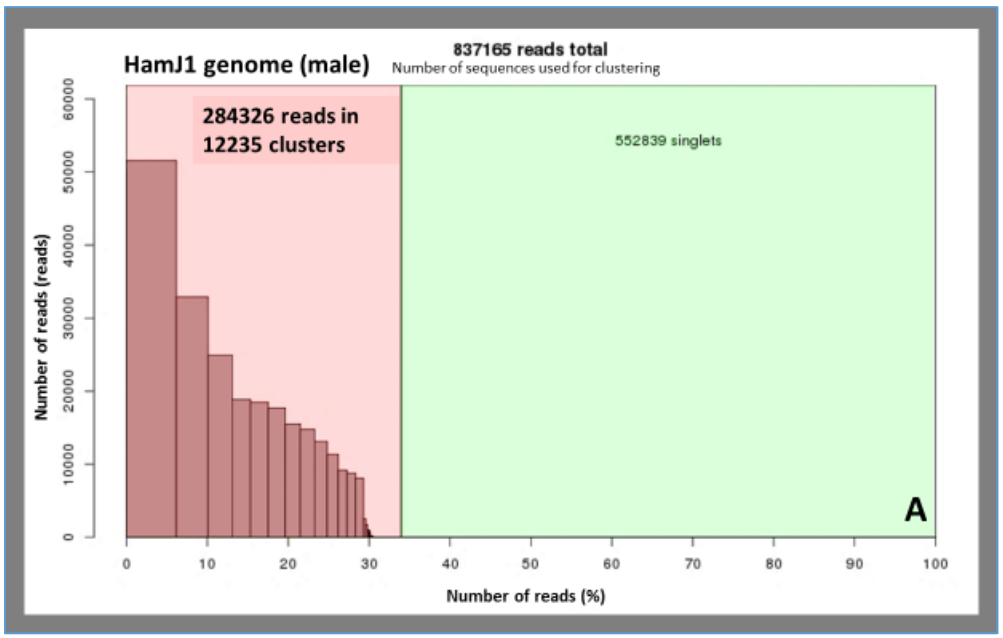
	R= AACAGTGAAGGGACTCAGCC		
CL9C194_L1	F= TGGAAAAGACAAGTGTACCCG	520	62
	R= AGTTCCCAATCCTGAACCCC		
CL17C5_L1-3	F= CAAAGATGAACCCCAGACAAACGG	398	62,64
	R= GTTTGCTTTTCCTACTGTTCTTTTCCC		
CL26C9_L1-3	F= CACAGTTCACCTCACCTCCC	296	62
	R= CACAATTGAAACCCAGGGGC		
32merC16-L1	F= CAATGCAATCCCTATCAAGC	304	56
	R= CCATTGTATATTCTTGCCTCC		
CL11C6_L1	F1= GGTGGGGTTCTATCAGTGGC	576	64
	R1= CTGGCTAGCTTGCTCTCTCC		
CL28C8_LINE1	F= TGCACACCAGGAGACCCC	285	58,60
	R= CCTTTGCTCCTTTCTCTTGGG		
CL2C941_SINE	F2= GACGCCATAGACGGTAGCC	400	64
	R2= GGAAGATACACCGACCTGCC		
CL2C1043_SINE	F3= GGTTGCCATTTCCTTCTGCG	269	64
	R3= ACTCCAGTGTTCTTGCCTGG		
CL1C2_SINE Ruminant	F= CAGGAGATATGGGTTTGATTCC	149	56
	R= CTTTGTATTGTTGCCCAGC		
CL36_SINE.BovA	F= GTGTAAACAGCGTCACAGGC	436	62
	R= GGCAAAGATACTGGAGGGGG		
CL31C1_SINE.tRNA	F= GTTCAAGGACACTGAAAGAACCC	624	64
	R= GCTTTCTCAGACCTCCTCCC		
CL94_SINE.MIR	F= ACTTTGCCTGCTATGTGGGG	450	62
	R= ACCAGTTGCGTCTTCTTGGG		
CL78_SINE.MIR	F= TGACGCTGTATTTCCCTACCC	417	62
	R= TCAAGACAGACTCTGCTTGG		
CL30C2_Low_complexity & SR	F= CCATCTACGCACCCAAACCC	585	62
	R= ACTGACAGAAGGCAGAAGGC		
CL27_Low complexity & SR	F= GTGCCTGTCTGCCTTTTTGC	556	62
	R= CAAGAGCTCGTCGAGGAGG		
CL40_Low complexity & SR	F= GTCGGGAACTTCAGGCTCC	535	62
	R= TATCCAGCTCCAAACGCTCG		
CL19_Low complexity & SR	F= CAGCATGTCGAGGAGGGC	468	62
	R= CTCGACGAGCTCTTGGCG		
CL43_DNA transposons	F= ACATAAATTACCCATCGGTTTCCC	581	62
	R= CATAAGCCTGGTGAAATGCCC		
CL46_DNA.hAT.Charlie	F= GAGACTTCCAGCAGCACAGG	545	62
	R= ATTGCTCAGCTCACCAAACG		
CL50_Interspersed repeat	F= GCTCAGCTTCAAAACCAGTCG	628	62
	R= TTTGTCCGTCCACTTTCCCC		
CL85_Non LTR	F= TCTCTGTGCACCATTCCAGG	403	64
	R= GTGGTTTCAGCTGACACTTAGC		
32merC7_ncRNA	F= GGCTACGTACAATCCATTCGG	608	60
	R= ACAGCTTGTTTCCTCCTGCC		

5.4 Results

5.4.1 Graph-based repeat identification and classification

The whole genome sequencing produced 43 (Karadi) to 60 million raw reads (Hamdani), a coverage of 2 to 3X of the sheep genome (see section 2.2.7.2). The focus of Chapter 5 is on the identification of dispersed repetitive elements without prior knowledge of their nature, before characterization by comparison with databases of known elements. The major tandemly repeated satellite DNA sequences with repeated motifs <1kb are discussed in Chapter 4. The rDNA sequences were left in the analysis since they provided a reference of an abundant and well-known repeat class.

Graph-based clustering of reads (see section 1.7.1) with default parameters for RepeatExplorer gave 12235 clusters (HamJ1_Male), 34% of the analyzed data (total of 837165 reads) or 16419 clusters (KarJ_Female), also 34% of the data (982736 reads analyzed). Each cluster consisted of 2 up to 51000 reads Figure 5.1A&B. Different thresholds for abundance were used: a 0.01% threshold identified 31 to 34 top clusters; using a 0.001% threshold, there were 94 clusters of HamJ1 to 127 clusters of KarJ, including specific clusters with male/female differences, or (see below) amplified on one chromosome. 97% of the reads clustered by RepeatExplorer (30% of the whole genome) were allocated to the most abundant 20 clusters (see cumulative frequency graph Figure 5.2).



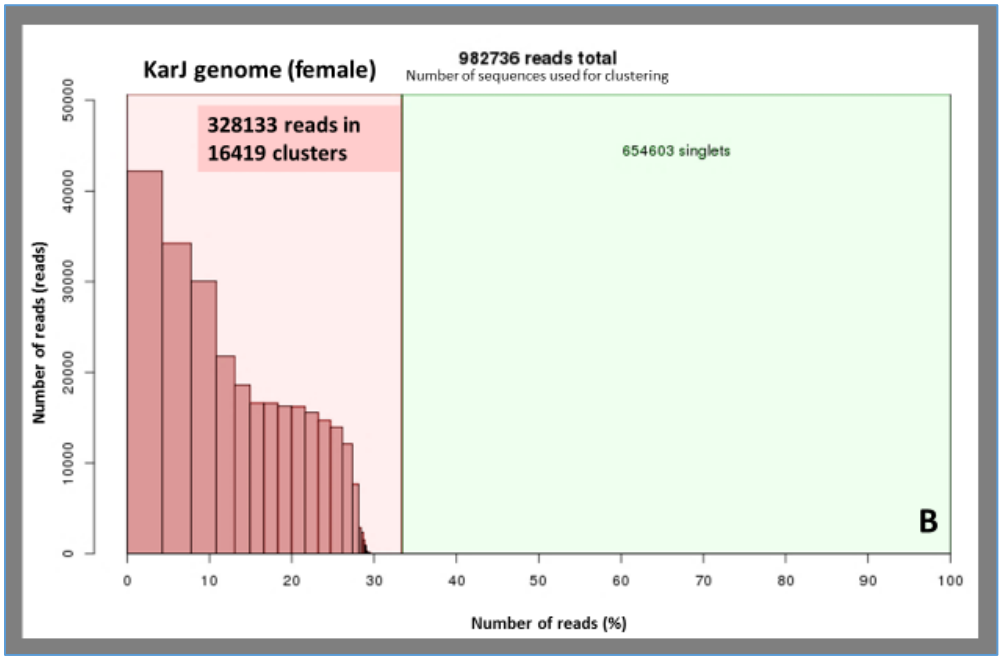


Figure 5.1 Distribution of clusters resulted from analysis of RepeatExplorer using 800 to 900 thousands of randomly selected raw reads from each of HamJ1_Male (A) and KarJ_Female (B) genomes. Clusters are ordered based upon the abundance of repeat sequences within each cluster. The most 10 common clusters are greatly more abundant than following clusters. The analyzed clusters represented 34% of all repeat clusters resulted from analysis of HamJ1 and KarJ genomes. The number of raw reads in each cluster diagramed the height, while width indicated the genomic proportion of the chart bars. Furthermore, unclustered raw reads called (singlets) or single-copy are exposed to the right of the vertical bar.

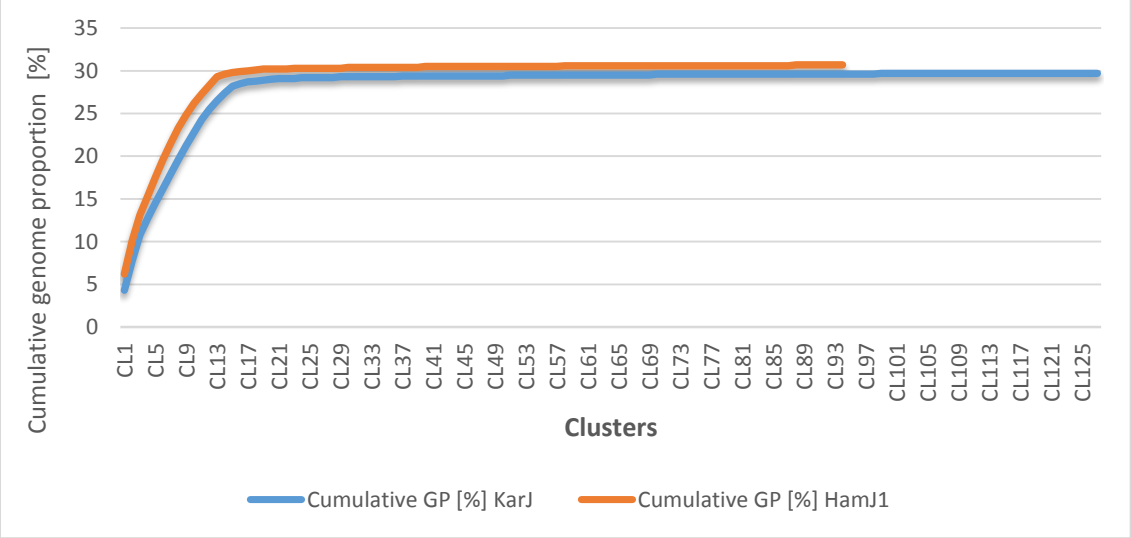


Figure 5.2 Cumulative abundance of the most common repeat families from sheep HamJ1_Male and KarJ_Female. The Y-axis shows the cumulative content of the reads in the clusters as an indicator of most abundant repeats; and the X-axis shows only the first 125 largest clusters. The graph shows slight differences in cumulative genome proportion of male and female. Graph also reflect that the majority of repetitive DNA landscapes including dispersed and tandemly repeated DNA sequences were populated in the first 20 clusters which mostly reflect that 29% of female and 31% male sheep genomes include repetitive families.

5.4.2 **Identification of dispersed elements using RepeatExplorer**

Candidate dispersed element sequences were identified in RepeatExplorer clusters by comparison with Repbase (Jurka *et al.* 2005; Bao *et al.* 2015) and TEclass (Abrusán *et al.* 2009)) databases, in particular searching for sequences with similarities to transposable elements or sequences with no known similarities. The mammalian RepeatMasker hits shown in the graph-based cluster analysis were also used for repeat identification Table 5.2. In HamJ1, 94 clusters were identified as LINEs, SINEs, satellites & novel tandem repeats, endogenous retroviruses, rDNA, and unidentified simple repeats or unclassified DNAs sequences Figure 5.3 and (Appendix 5.1). The various cluster graphs (where each read is represented by a vertex (node) and sequence overlaps by edges) showed distinct shapes (graph layouts; Tables 5.2), assisting classification and annotation of repeats in each cluster (see section 1.7.1). Non-LTR retrotransposons either show star-shaped (like satellite tandem repeats, see Chapter Four) or linear graphs, while other repeats showed linear or regular arc shapes.

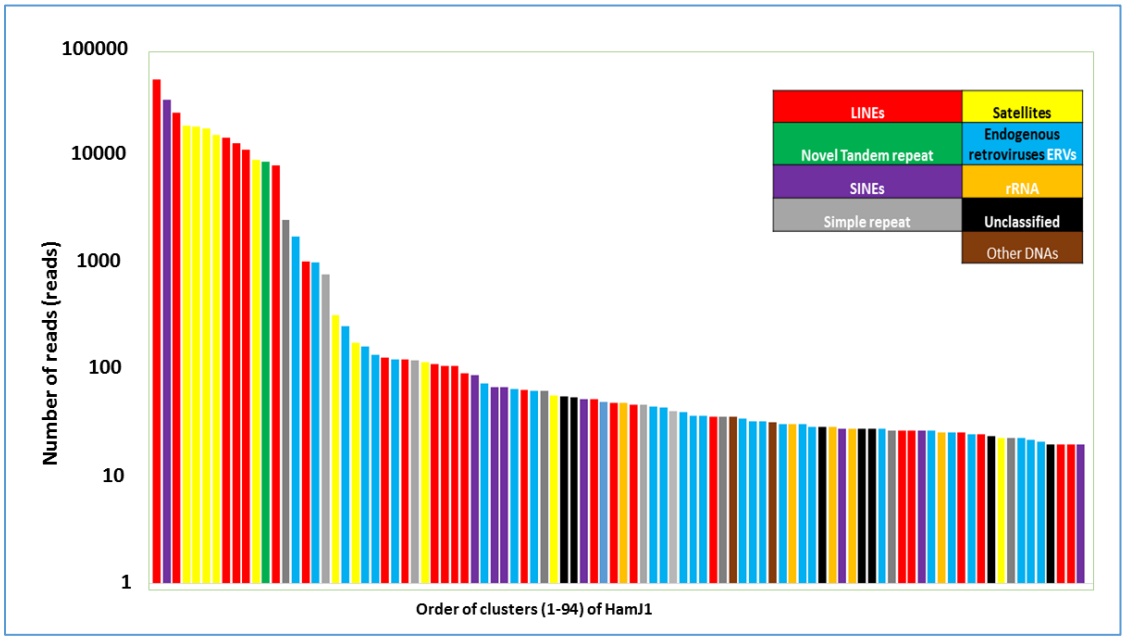
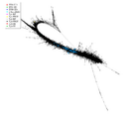
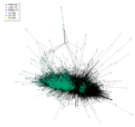
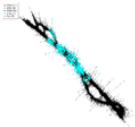
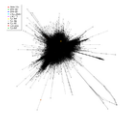
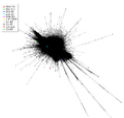
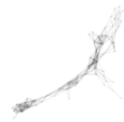
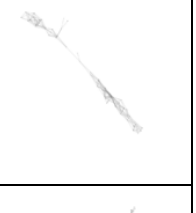
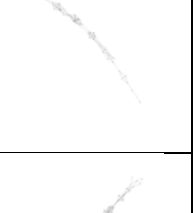
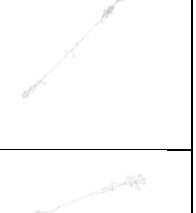
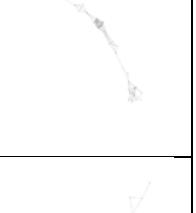
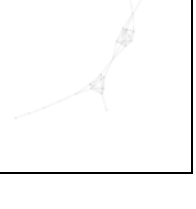


Figure 5.3 Shows annotation of 94 clusters including dispersed and tandemly repetitive elements. The number of raw reads in each cluster diagramed the height, while width indicated the genomic proportion of the chart bars (Different colours indicate different classes of repetitive DNA sequences that represented in several clusters resulted from analysis of NGS data by RepeatExplorer (see section 1.7.1).

Table 5.2 Shows genomic proportions, number of reads, total lengths, hits% of homologous sequences and graphic layout of clusters representing different classes of dispersed repeats which selected used as probes. Construction of graph layouts were explained in section 1.7.1

Clusters	Total length	Number of reads	Genome proportion[%]	RepeatMasker	Layout
CL8C95_RTE **CL8C129_RTE**	2108709	14017	1.610	LINE.RTE.BovB (14084hits, 93%) SINE.MIR (536hits, 1.58%) LINE.L1 (229hits, 1.04%) SINE.tRNA.Core.RTE (368hits, 0.949%) SINE.Core.RTE (43hits, 0.0994%) LTR.ERV1 (18hits, 0.0488%)	
CL10C107_RTE	2166707	14401	1.600	LINE.RTE.BovB (15018hits, 84.8%) LTR.ERVK (4003hits, 11.1%) SINE.Core.RTE (44hits, 0.106%) LINE.L1 (40hits, 0.0926%) SINE.tRNA.Core.RTE (25hits, 0.0553%)	
CL7C43_RTE	2903946	19305	2.150	LINE.RTE.BovB (19434hits, 94.1%) SINE.MIR (519hits, 1.12%) LINE.L1 (232hits, 0.704%) SINE.tRNA.Core.RTE (372hits, 0.682%) SINE.Core.RTE (71hits, 0.122%) Simple_repeat (41hits, 0.0331%)	
CL4C63_RTE **CL5C418_RTE** **CL5C464_RTE**	3385765	22503	2.500	LINE.RTE.BovB (22711hits, 97.1%) LINE.L1 (72hits, 0.112%) SINE.Core.RTE (48hits, 0.0696%) SINE.tRNA.Core.RTE (39hits, 0.0559%) LTR.ERV1 (16hits, 0.0269%) Simple_repeat (34hits, 0.0252%)	
CL12C2_L1 **CL12C16_L1-3** **CL12C27_L1-3**	2314166	15383	0.994	LINE.L1 (13856hits, 91.4%) Satellite (1049hits, 5.81%) LINE.RTE.BovB (70hits, 0.16%) SINE.tRNA.Core.RTE (27hits, 0.0573%) SINE.Core.RTE (20hits, 0.0456%) SINE.MIR (11hits, 0.0229%)	
CL25C6_L1-2	20772	138	0.015	LINE.L1 (139hits, 90.4%) SINE.Core.RTE (2hits, 0.428%) SINE.tRNA (2hits, 0.424%) Simple_repeat (1hits, 0.12%) Low_complexity (1hits, 0.106%)	
CL9C194_L1	2198525	14609	1.620	LINE.L1 (13434hits, 83.1%) Satellite (1562hits, 8.58%) LTR.ERVK (286hits, 0.737%) Low_complexity (175hits, 0.305%) LINE.RTE.BovB (124hits, 0.265%) Simple_repeat (95hits, 0.136%) <="" .....=""">	
CL26C9_L1-3	19567	130	0.014	LINE.L1 (123hits, 85.8%) SINE.tRNA.Core.RTE (16hits, 8.23%) SINE.Core.RTE (3hits, 0.721%) LINE.RTE.BovB (1hits, 0.394%)	

CL11C6_L1	1479436	9834	1.090	LINE.L1 (9809hits, 96%) Low_complexity (41hits, 0.135%) Simple_repeat (47hits, 0.0934%) LINE.RTE.BovB (26hits, 0.0792%) SINE.Core.RTE (18hits, 0.055%) SINE.tRNA.Core.RTE (11hits, 0.0312%)<.....	
CL28C8_LINE1 Dolphin	17017	113	0.013	LINE.L1 (69hits, 47.4%) LINE.RTE.BovB (1hits, 0.347%)	
CL2C941_SINE **CL2C1043_SINE**	5336988	35470	3.940	SINE.tRNA.Core.RTE (23957hits, 55%) SINE.Core.RTE (13250hits, 29.1%) SINE.MIR (968hits, 1.16%) LINE.L1 (453hits, 0.431%) LINE.RTE.BovB (461hits, 0.407%)	
CL1C2_SINE Ruminant	6529741	43392	3.050	SINE.tRNA.Core.RTE (27862hits, 53.6%) SINE.Core.RTE (16500hits, 30.8%) SINE.MIR (904hits, 0.81%) LINE.L1 (533hits, 0.439%) LINE.RTE.BovB (472hits, 0.348%) SINE.tRNA (235hits, 0.228%)	
CL36_SINE.BovA	10241	68	0.0081	SINE.BovA (17hits, 18.8%)	
CL31C1_SINE.tRNA	14440	96	0.011	SINE.tRNA (6hits, 3.19%) SINE.tRNA.Core.RTE (1hits, 0.589%)	
CL94_SINE.MIR	3006	20	0.0024	SINE.MIR (2hits, 5.56%) Simple_repeat (1hits, 0.765%)	
CL78_SINE.MIR	4064	27	0.0032	SINE.MIR (7hits, 11.9%) Simple_repeat (1hits, 0.689%)	
CL30C2_Low_complexity & SR	15832	105	0.012	Simple_repeat (1hits, 0.619%) Low_complexity (3hits, 0.436%)	

CL27_Low complexity & SR	18396	122	0.0146	Low_complexity (31hits, 6.78%) Simple_repeat (2hits, 0.348%)	
CL40_Low complexity & SR	9496	63	0.0075	Low_complexity (13hits, 3.96%) Simple_repeat (1hits, 0.263%)	
CL19C3_Low complexity & SR	29720	197	0.022	Low_complexity (63hits, 8.37%) Simple_repeat (4hits, 0.485%)	
CL44_SINE.rRNA	7984	53	0.0063	SINE.tRNA.Glu (10hits, 9.99%) LTR.ERV1 (4hits, 4.03%) Simple_repeat (1hits, 0.388%)	
CL43_DNA transposons	8279	55	0.0066		
CL46_DNA.hAT.Charlie	7526	50	0.0060	DNA.hAT.Charlie (7hits, 8.25%) Simple_repeat (1hits, 0.797%)	
CL50_Interspersed repeat	7067	47	0.0056	Unknown (3hits, 5.01%) Simple_repeat (1hits, 0.269%)	
CL85_Non LTR	3609	24	0.0029		

5.4.3 Abundance of repetitive DNA sequences analyzed by RepeatExplorer

5.4.3.1 Non-LTR Retrotransposons

Non-LTR retrotransposons including LINEs_L1, LINEs_RTE and SINEs from the RepeatExplorer analysis were found to be the most abundant repetitive elements within the whole sequencing raw reads of sheep genome in comparison to the abundance of other repeats. Within the non-LTR retrotransposons, the LINE_RTE elements constituted the most abundant genomic proportion about 12.34% and 11.87% of HamJ1_Male and KarJ_Female respectively. SINE repeat sequences come in the second position in context of their genomic proportion, about 4.15% for HamJ1_Male and 4.9% for KarJ_Female genomes. The least genomic percentage, and hence less abundant repeats, of non-LTR retrotransposons was represented in the LINE_L1 family. LINEs_L1 elements occupied about 2.78% and 2.70% of HamJ1_Male and KarJ_Female genomic sequencing raw reads respectively. In total, the non-LTR-retrotransposons were found to be predominant repetitive component, occupying nearly 20% of the sheep genome Table 5.3. Regarding numbers of RepeatExplorer clusters, LINEs_RTE were highly abundant and distributed into more than eight clusters within the top 100 clusters. LINEs_L1 repeats were also assigned to at least eight clusters with different contributions in each cluster. While, the SINEs elements were more specific to one cluster although they contributed their reads in other clusters with lower genomic proportions. Furthermore, some clusters were connected to each other due to presence of sequence similarities between reads of these clusters. However, some clusters were more independent and specific to only one type of repeats (see above; Figure 5.3).

Table 5.3 Genome proportion of major groups of repetitive sequences identified in unassembled raw reads of sheep genome using utilities of RepeatExplorer. Only transposable elements and dispersed repeat classes are considered in Chapter 5.

Repetitive Classes	**Repetitive elements**	**Genomic proportions %**	
		HamJ1 (male)	**KarJ (female)**
Non-LTR retrotransposons	LINEs_RTE	12.34	11.87
	LINEs_L1	2.78	2.70
	SINEs	4.15	4.90
	Total %	**19.27**	**19.47**
LTR retrotransposons	Endogenous retroviruses related repetitive elements	0.55	0.54
DNA transposons	DNA.TcMar.Mariner	0.02	0.011
	DNA.hAT		
	DNA.PiggyBac		
	RC.Helitron		
Tandem repeats	Satellite I	7.42	5.91
	Satellite II	2.11	1.69
	Novel tandem repeat	1.05	0.78
	Other Satellites	0.04	0.04
	Total%	**10.62**	**8.42**
rDNA	18S rDNA and other ribosomal seqeunces	0.02	0.028
Unclassified sequences	Simple_repeats	0.30	0.29
	Low_complexity	0.06	0.06
	Others	0.03	0.06
	Total %	**0.39**	**0.41**
Total%		**30.87**	**28.88**

5.4.3.2 **DNA transposons**

Repetitive elements including DNA.TcMar.Mariner, DNA.hAT, DNA.PiggyBac and RC.Helitron of DNA transposons class were also discovered from analysis of the outcome of RepeatExplorer. In comparison to the genomic proportion of all other repetitive elements, DNA transposons were found less frequently as repeats in whole sequencing raw reads. In other word, RepeatExplorer, discovered very rare amounts of DNA transposons with genomic proportions about 0.02% and 0.011% of HamJ1_Male and KarJ_Female sheep genomes Table 5.3 & Appendices 5.2 & 5.3.

5.4.3.3 **Unidentified repeats (automated annotation as "low complexity", "unknown" or "simple repeats")**

Some clusters of RepeatExplorer consisted of unidentified or unclassified sequences, sometimes with automated annotation as "simple repeat", "low complexity" as well as "unknown". CL14_HamJ1 and CL16_KarJ were examples of unidentified sequences, and furthermore, no similarities to the Repbase databases were found. Genomic proportion and graph layouts of these clusters were about 0.29-0.30% Figure 5.4. Sequences of these two clusters were not used for further analysis. However, other clusters like CL12_HamJ1 and CL15_KarJ were found to be highly abundant tandem repeats (see Chapter Four). Probes representing some other "unidentified" clusters were used for *in situ* hybridization and named [**L**ow **C**omplexity (LC) & **S**imple **R**epeat (SR)].

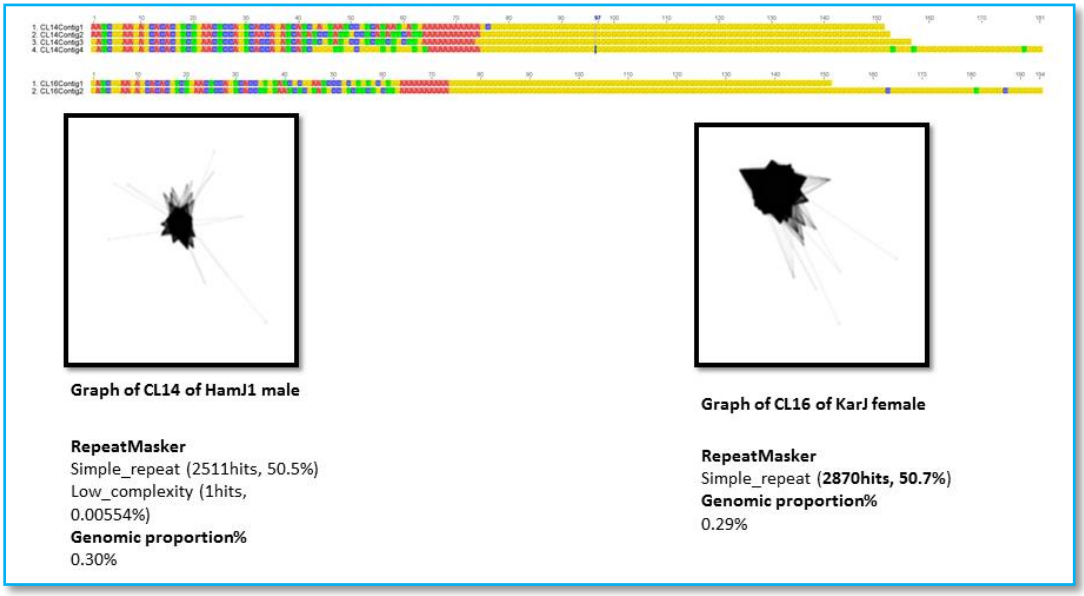


Figure 5.4 Sequences, genomic proportions, sequence hits and graphic layout of clusters of Low Complexity (LC) & Simple Repeat (SR). These clusters were not used for FISH.

5.4.3.4 **rDNA gene sequences**

Analysis of whole sequencing raw reads by RepeatExplorer enabled discovery of gene sequences related to ribosomal RNA repeat unit. Different parts of rDNA sequences such as 18S rDNA and other sequences were classified. Sequences of rDNA gene with different lengths were distributed over several clusters see above; Figure 5.3. Overall, genomic proportions of rDNA gene in whole sequencing raw reads were about 0.02%

and 0.028% of each HamJ1_Male and KarJ_Female genome respectively Table 5.3 & Appendices 5.2 & 5.3.

5.4.4 *k*-mer analysis in sheep genome sequencing raw reads

k-mer analysis involves counting the frequency of each short sequence motif *k* bases long in whole genome sequencing raw reads. This method is devoid of assembly algorithms and is an unbiased method to identify repetitive DNA motifs. Several values of *k* (from 12-144bp) were used to assess the repetitive landscapes in unassembled raw reads of the sheep genome. A slope representing the frequency of short bases of sequences (mer) starting from 12 until 16 bases long showed presence of larger repetitive motifs approximately two-thirds times the frequency compared to values starting after 16mers. In other words, short motifs less than 32bp mer long bases were found most frequently as sequence motifs from *k*-mer analysis Figure 5.5. To identify repetitive composition, higher values of *k* were also used, for instance, 32mers.

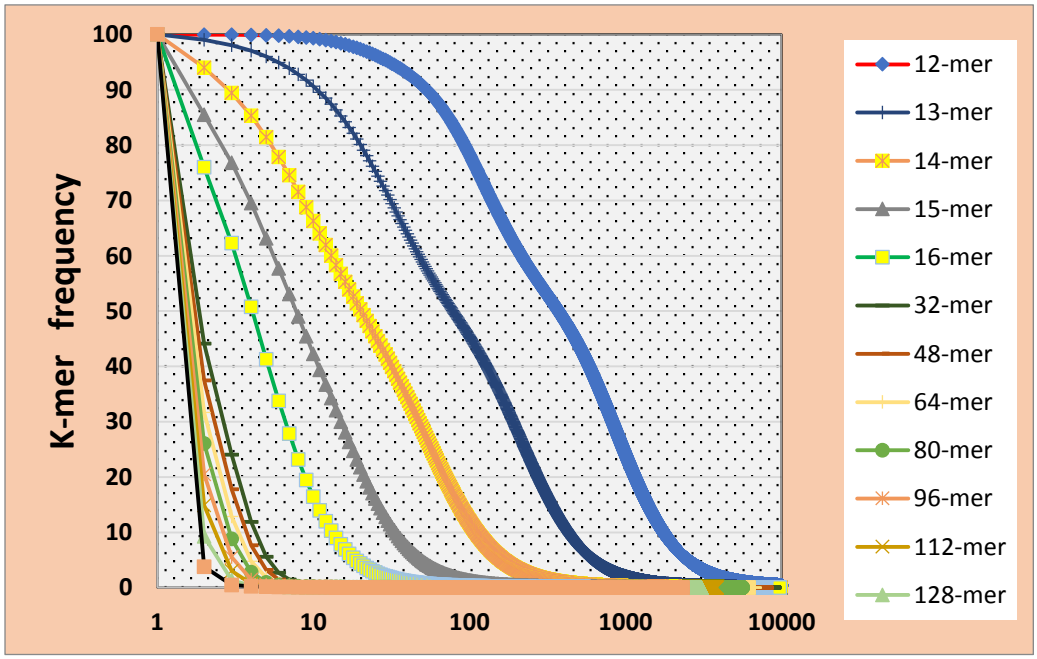


Figure 5.5 Whole sequencing raw reads analysed by *k*-mer frequency. Variety of *k*-mer values was used to show the frequency of short motifs in the raw reads of sheep genomes

5.4.5 Identification of dispersed repeats found as abundant *k*-mers

Dispersed DNA elements were investigated using the *k*-mer frequency tool (Jellyfish). *k*-mers with different values of *k* (22, 32, 56, 64 & 128 mers) repeated 100 to 100000 times as appropriate were used for assembly, giving contigs with various lengths. Consensus sequences representing large numbers of reads (typically in the top 100 contigs) were compared with Repbase, TEclass and NCBI databases. Many contigs of *k*-mer frequency were similar to known dispersed repeat motifs including LINEs (L1 & RTE) and SINEs elements.

5.4.6 Identification of non-coding RNA sequences using *k*-mer frequency and RepeatExplorer

Following comparison, the contigs resulted from assembly of short motifs of 32mers to databases of NCBI, consensus of contig7 was matched to non-coding RNA sequences of *Ovis aries* uncharacterized LOC101104348 (LOC101104348), ncRNA. Furthermore, sequences of all clusters resulted from the outcome of graph-based clustering method analysis of HamJ1_Male and KarJ_Female genomes were mapped to the contig7, and clusters such as CL17C13/CL20C9 were contained the same sequences of non-coding RNA. Sequences of ncRNA in cluster CL17C13 were incorporated within two different types of repetitive elements; LINEs and ERVs.

5.4.7 Assembly of 18S rDNA genes

The 22mer frequency analysis and assembly of 22mers repeated more than 100 times from HamJ1 generated a contig matching the 18S ribosomal RNA gene sequence of Bubalus bubalis and *Bos taurus*. Thus, contig 3 of 22mers with 2044bp was representing sequences of 18S rDNA gene in sheep Figure 5.6A. The second method of 18S rRNA gene assembly was by mapping the whole paired reads of sheep genome to a reference sequence; the 18S rRNA gene of *Bos taurus* following section 2.2.13.5. Then, the consensus of the assembled 18S rRNA gene was extracted, reassembled and annotated using the Geneious program. As well *k*-mer frequency and map to reference, the complete consensus of the 18S rRNA gene was assembled from mapping sequences of several clusters of RepeatExplorer outcome of HamJ1_Male and KarJ_Female to the

annotated *Ovis aries* 18S rRNA gene, and mostly 10 clusters were assembled Figure 5.6B. Genomic proportion of 18S rRNA gene in female was about 0.011% while in male was 0.0081% Table 5.4. The complete sequence (1869bp) of *Ovis aries* 18S ribosomal DNA is available in GenBank under accession number KY129860.

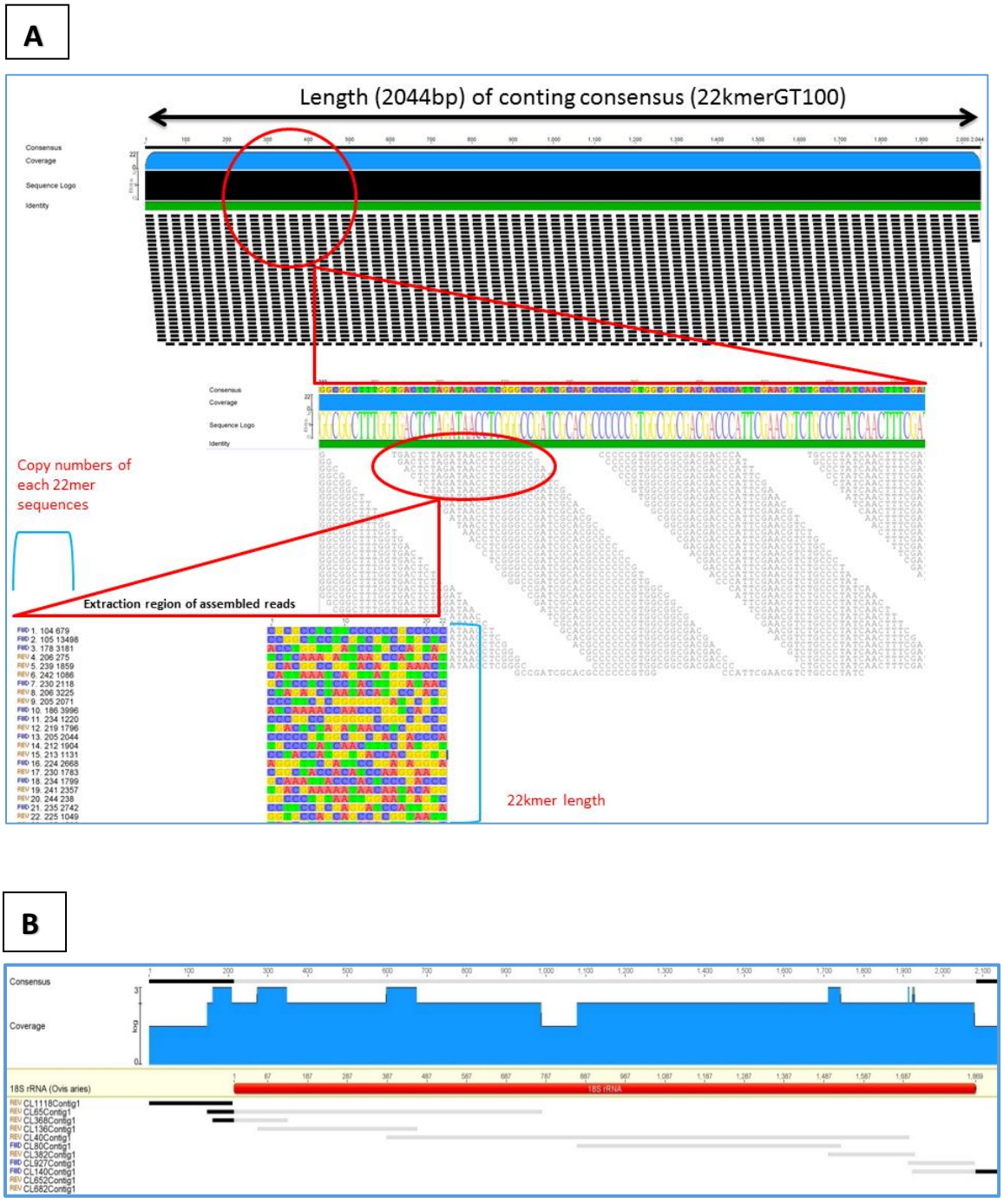


Figure 5.6 A. Assembly of 22mers using *k*-mer frequency analysis. B. Assembly of clusters of graph-based read clustersing to complete gene of 18S rRNA. See section 2.2.13.2.

5.4.8 Assembly of 5S ribosomal RNA gene from RepeatExplorer

Sequences of some RepeatExplorer clusters as well as contigs of 44mersGT100 were highly similar to the sequences of 5S ribosomal RNA gene; Syrian hamster 5S ribosomal RNA gene; J00063 (Hart & Folk 1982). The 5S ribosomal RNA consensus was also matched SINE-like 5S-derived retro pseudogene from guinea pig (138bp) from Repbase databases. Furthermore, the whole paired reads were mapped to the 5S ribosomal RNA gene in order to extend the gene consensus and also estimate its copy numbers and genomic proportion. Approximately, 16 thousand copies of 5S ribosomal RNA gene were estimated per haploid sheep genome. Dot plot (self) consensus of 5S ribosomal RNA gene is shown in Appendix 5.4.

5.4.9 Non-LTR retrotransposons - LINEs

5.4.9.1 Assembly of sheep RTE non-LTR retrotransposon - a consensus

Many contigs from different clusters of RepeatExplorer outcome of HamJ1 (CL1, CL3, CL8 and CL10) and KarJ (CL2, CL3, CL6, CL8, CL10 and CL11) were representing sequences of non-LTR retrotransposon RTE repeat. The consensus of RTE non-LTR retrotransposon in sheep with total length 4.6-50Kbp was assembled from mapping the whole paired raw reads to the longest contig of LINEs_RTE (Appendix 5.5). Genomic proportions of RTE non-LTR retrotransposon consensus in each of HamJ1_Male and KarJ_Female genomes were about 6.4% (107876 copies) to 6.7% (133207.3 copies) respectively Table 5.4. Alignment between RTE non-LTR retrotransposon consensus of *Bos taurus* (3847bp) (Repbase databases) and *Ovis aries* RTE non-LTR retrotransposon consensus showed 97% sequence identities including 50 SNPs. However, less than 50% sequence identities were found between consensus of RTE and L1 sequences of sheep.

5.4.9.2 Assembly of L1; Non-LTR Retrotransposon; Transposable Element; L1-3-Sheep consensus

Several clusters resulted from analysis of RepeatExplorer of HamJ1 (CL13, CL24, CL38, CL47, CL49, and CL77) and KarJ (CL13, CL22, CL27, CL28, CL33, CL43, CL47, and CL88) were found highly similar to non-LTR retrotransposon LINE L1 repeat. To assemble the whole consensus of Non-LTR Retrotransposon L1-3-Sheep, the whole paired raw reads

were mapped to the longest contig containing sequences of LINEs_L1 repeat. As a result, a consensus with total length 8467-8846bp was assembled (Appendix 5.6). Genomic proportion and copy numbers of non-LTR retrotransposon L1 consensus in each of HamJ1_Male and KarJ_Female genomes were about 0.96% (8873 copies) to 1.2% (13126 copies) respectively Table 5.4. Alignment between consensus of Non-LTR Retrotransposon L1-3 of *Bos taurus* (8468bp) and *Ovis aries* (8467-8846bp) showed 96% sequence identities including several polymorphisms 400 SNPs.

5.4.9.3 Assembly of L1 Non-LTR Retrotransposon; Transposable Element; L1-Sheep consensus

Several clusters of HamJ1 (CL13 and CL16) and KarJ (CL13 and CL18) were matched the L1; Non-LTR Retrotransposon repeats. The consensus of Non-LTR Retrotransposon L1-Sheep with total length 8800-9600bp was resulted from assembly the whole paired raw reads of sheep genome to the longest contig (Appendix 5.7). Its genomic proportion and copy numbers in each of HamJ1_Male and KarJ_Female genomes were about 1.47% (13000 copies) to 1.59% (16000 copies) respectively Table 5.4. L1; Non-LTR Retrotransposon; Transposable Element; L1-BT of *Bos taurus* (8390bp) from Repbase was aligned to the sheep non-LTR retrotransposon L1 consensus. 91% sequence identities including huge polymorphic sites 700 SNPs were found between *Ovis aries* and *Bos taurus* (Appendix 5.8). Furthermore, in comparison to the consensus of L1 Non-LTR Retrotransposon of *Bos taurus*, the tandem region 880bp was not identified in the sheep consensus (Appendix 5.7; yellow box). Tandem region (yellow box) was not assembled from mapping to reference either. On other hand, sequence identities between L1 and L1-3 of sheep consensus were about 75% (Appendix 5.9).

5.4.9.4 Assembly of L1 Non-LTR Retrotransposon; Transposable Element; L1-2 Sheep consensus

Contigs of CL19_HamJ1 were highly homologous to sequences of L1; Non-LTR Retrotransposon; Transposable Element; L1-2. Therefore, the whole paired raw reads of sheep genome were mapped to the longest contig, and the consensus of L1; Non-LTR Retrotransposon L1-2 Sheep with total length 2041bp was resulted (see dot plot; Appendix 5.10). Its genomic proportion in each of HamJ1_Male and KarJ_Female genomes was about 0.57% (22166 copies) to 0.62% (27801) respectively Table 5.4. In

comparison to cattle, L1; Non-LTR Retrotransposon; Transposable Element; L1-2_BT of *Bos taurus* (2015bp) was downloaded from Repbase and aligned to the L1-2 Sheep consensus, and 87% sequence identities were found. Furthermore, sequence identities between L1-2 Sheep consensus to each of L1-3 and L1 of sheep genome was about 86%. Consensus of L1-2 is corresponding to the last part of sheep consensus of each of L1-3 and L1.

Table 5.4 Genomic proportion and copy numbers of Non-LTR Retrotransposon LINEs_RTE, LINEs_L1 and 18S rRNA consensuses.

		HamJ1_Male				KarJ_Female			
Repeats	**Length bp**	**Assembled reads**	**Whole genome raw reads**	**Copy numbers**	**Genomic proportion %**	**Assembled reads**	**Whole genome raw reads**	**Copy numbers**	**Genomic proportion %**
Sheep-RTE	4606	3312537	52048068	107876.8	6.36	4090353	60605648	133207.3	6.74
Sheep-L1-3	8467	500831	52048068	8872.64	0.96	740892	60605648	13125.52	1.22
Sheep-L1	8800	768794	52048068	13104.44	1.47	967874	60605648	16497.85	1.59
Sheep-L1-2	2041	301606	52048068	22166.04	0.57	378289	60605648	27801	0.62
18S rRNA	1869	4235	52048068	339.88	0.0081	7010	60605648	562.6	0.0115

5.4.9.5 Assembly of L1; Non-LTR Retrotransposon from (L1-1_Ttr) of dolphin species (*Tursiops truncates*)

Following blasting of RepeatExplorer clusters against Repbase databases, sequences of CL28C8 were found highly similar to sequences of LINE repeat of dolphin species. Thus, complete consensus of LINE (6kbp) from dolphin was downloaded from Repbase and used as reference. Whole paired raw reads of sheep genome mapped to the LINE consensus of dolphin, and 549,912 reads (2X150bp) were assembled generating nearly the complete consensus of LINE repeats about 5kbp. Genomic proportion was about 1%. The sequence identities between consensus of dolphin LINE and the assembled LINE of sheep were more than 85%.

5.4.10 Non-LTR retrotransposon derivatives: SINEs

5.4.10.1 SINE clusters

Many sequences reads from SINEs will be clustered with the reads of their parent element; solo-LTRs for LTR retroelements also would cluster with their parent elements. SINE repeats were assigned to several RepeatExplorer clusters, many with low genomic proportions. CL2C941, CL2C1043, CL31C1, CL1C2, CL44, CL78, CL94 and CL36 matched

SINE repeats from different species of mammal. Some SINEs clusters were selected to use them as probes in order to characterize their chromosomal abundance see copy numbers and genomic proportions in Table 5.5.

5.4.10.2 **Comparison of *Ovis aries* SINEs with *Bos taurus* and ancestral SINEs from *Ruminantia***

A subset including nine SINE repeats of *Bos taurus* (BCS, BOVA2, BOVTA, BTALUL1, CHR-2_BT, CHRL1_BT, SINE2-1_BT, SINE2-2_BT and SINE2-3_BT) and three ancestral SINEs of *Ruminantia* (Bov-tA1, Bov-tA2, Bov-tA3) were downloaded from Repbase databases (Jurka *et al.* 2005; Bao *et al.* 2015). These SINE repeats, which belong to SINE2/tRNA class, were then compared with the sheep RepeatExplorer clusters. The three ancestral SINEs, five *Bos taurus* SINEs (BCS, BOVA2, BOVTA, BTALUL1 and SINE2-3_BT) were matched the contig sequences of RepeatExplorer clusters of sheep genome. The other bovine SINEs sequences (CHR-2_BT, CHRL1_BT, SINE2-1_BT and SINE2-2_BT) were not found. Following section 2.2.13.5, the whole sequencing raw reads of sheep were assembled to three ancestral SINEs of *Ruminantia* and to the other five SINEs of *Bos taurus* with genomic proportion 0.30%-0.85% respectively. High similarity upto 95% including indels and SNPs were estimated from the alignment between SINE sequences of sheep and *Bos taurus*.

5.4.10.3 **Comparison of SINEBase database with sheep whole genome sequencing**

A subset including 221 fragments of SINE sequences from SINEBank, six fragments of COREBank, and seven fragments of LINEBank including SINE sequences were downloaded from SINEBase: a database and tools for SINE analysis (Vassetzky & Kramerov 2012). All SINE sequences of SINEBase were used as reference, and compared to whole sheep genome sequencing raw reads following section 2.2.13.5. Only Bov-tA, and Bov-tA2 were found in sheep genome suggesting that such SINE sequences are conserved among *Ruminantia*, while the remaining SINEs repeats are more likely species-specific.

5.4.11 **Bioinformatics abundances of probes used for FISH**

Copy numbers and genomic proportion of each probe representing different dispersed repetitive DNA sequences used in this chapter were estimated following section 2.2.13.5 Table 5.5. Whole genome sequencing used for estimation the probe copy numbers were 52048068 reads for HamJ1 and 60605648 reads for KarJ (see section 2.2.7.2).

Table 5.5 Abundances of probes representing different classes of dispersed repeats used for *in situ*.

		HamJ1_Male			KarJ_Female		
Used probes	**Product Size (bp)**	**assemlbed reads**	**Copies of probe**	**Genomic proportion%**	**KarJ-assembled reads**	**Copy numbers of probes**	**Genomic proportion %**
CL5C464_RTE	607	557957	137881	1.072	590000	145799	0.9735
CL5C418_RTE	500	982819	294846	1.8883	971807	291542	1.6035
CL8C129_RTE	481	393496	122712	0.756	490684	153020	0.8096
CL8C95_RTE	562	257678	68775	0.4951	319680	85324	0.5275
32merC12_RTE	183	358132	293551	0.6881	400000	327869	0.66
32merC15_RTE	324	1406285	651058	2.7019	400000	185185	0.66
32merC31_RTE	258	416868	242365	0.8009	334579	194523	0.5521
CL4C63_RTE	489	733546	225014	1.4094	933150	286242	1.5397
CL7C43_RTE	470	480516	153356	0.9232	703491	224518	1.1608
CL10C107_RTE	450	700000	233333	1.3449	600000	200000	0.99
CL1C1022_RTE	514	1094678	319459	2.1032	964604	281499	1.5916
CL12C16_L1-3	543	108027	29842	0.2076	108096	29861	0.1784
CL12C27_L1-3	362	44264	18341	0.085	53900	22334	0.0889
CL9C194_L1	520	210243	60647	0.4039	250000	72115	0.4125
CL12C2_L1	303	86514	42829	0.1662	107143	53041	0.1768
CL11C6_L1	576	55151	14362	0.106	47684	12418	0.0787
CL11C80_L1-3	520	17151	4947	0.033	21250	6130	0.0351
CL17C5_L1-3	398	3800	1432	0.0073	4000	1508	0.0066
CL25C6_L1-2	180	2700	2250	0.0052	3000	2500	0.005
CL26C9_L1-3	296	3000	1520	0.0058	3000	1520	0.005
CL28C8_LINE1 Dolphin	285	800	421	0.0015	464	244	0.0008
CL50_Interspersed repeat	628	4842	0.0093	1157	29	7	0.00005
32merC36_Capra	586	6000	1536	0.0115	6500	1664	0.0107
32merC7_ncRNA	609	6028	1485	0.0116	4262	1050	0.007
CL43_DNA transposons	581	4848	1252	0.0093	36	9	0.00006
CL85_Non LTR	403	3644	1356	0.007	0	0	0
CL46_DNA.hAT.Charlie	545	1700	468	0.0033	0	0	0
CL36_SINE.BovA	436	5362	0.0103	1845	5370	1847	0.00886
CL31C1_SINE.tRNA	624	6807	0.0131	1636	3956	951	0.0065
CL94_SINE.MIR	450	1990	0.0038	663	23	8	0.00004
CL78_SINE.MIR	417	2031	0.0039	731	9543	3433	0.01575
CL44_SINE.tRNA	553	3414	0.0066	926	0	0	0
CL2C1043_SINE	269	50000	0.0961	27881	40000	22305	0.066
CL1C2_SINE Ruminant	149	12	0	12	10	10	0
CL40_Low complexity & SR	535	4841	0.0093	1357	2683	752	0.00443
CL30C2_Low_complexity & SR	585	7254	0.0139	1860	7651	1962	0.0126
CL19C3_Low complexity & SR	468	3500	0.0067	1122	2700	865	0.0045

5.5 Assessment of sheep genome size from NGS data

The size of the sheep genome is well known (and similar to nearly all other mammalian genomes). Here, the genome size was re-estimated using the *k*-mer analysis as a validation following section 2.2.15. Several *k*-mer size were investigated, read depth was calculated and the peak *k*-mer depth was then extracted. Peak *k*-mer depth was calculated by counting *k*-mer frequencies using the paired raw reads following section 2.2.15. Two formulae were then applied to quantify genome size: here is an example of *k-mer* size 17.

[Read depth= Peak *k*-mer depth*Read length / (Read length-*k*-mer size+1)];

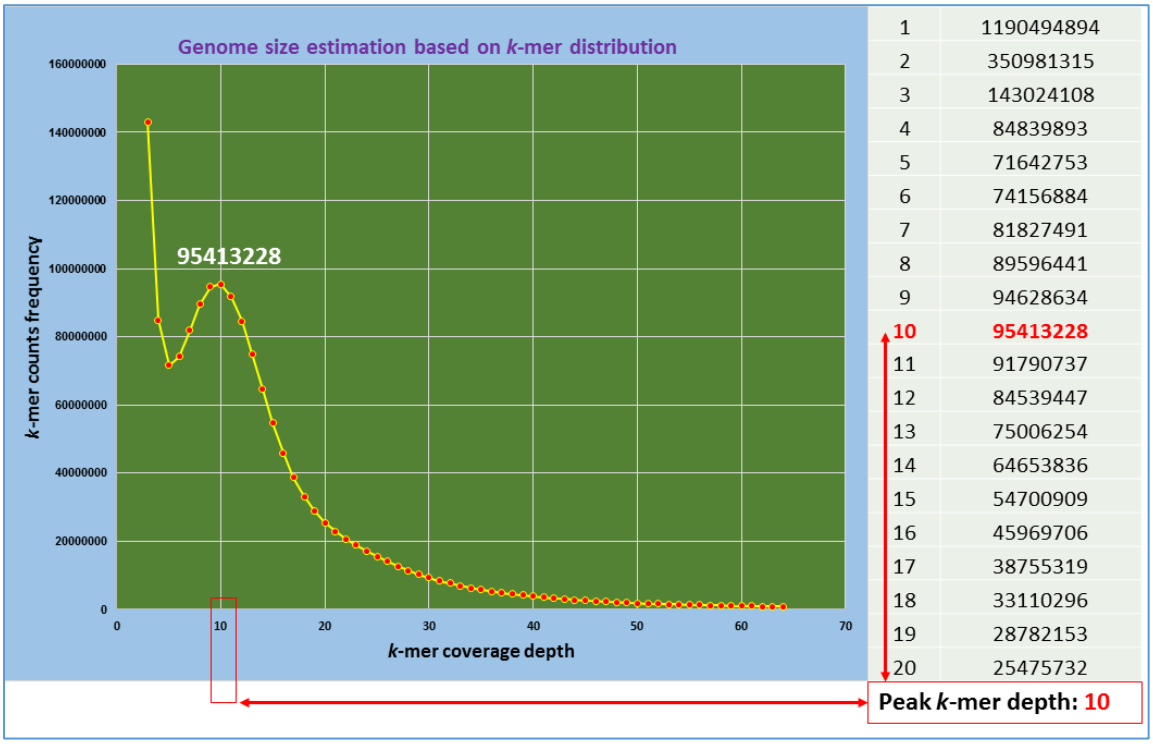
Read depth =10*150 / (150-17+1) = 11.194

And

[Genome size= Total base pairs/read depth]= 38.59/11.194= 3.45

Thus, the sheep genome size based on 17mers = 3.45Gbp

k-mer count frequency and *k*-mer depth resulted from jellyfish histo were used for drawing the slope representing peak *k*-mer depth Figure 5.7. The optimum *k*-mer lengths with sharpest single-copy peaks were between 15 and 18 mers and gave a genome size of 3,000,000,000 base pairs, similar to that from assemblies and microdensitometer measurements Figure 5.8.



1	1190494894
2	350981315
3	143024108
4	84839893
5	71642753
6	74156884
7	81827491
8	89596441
9	94628634
10	**95413228**
11	91790737
12	84539447
13	75006254
14	64653836
15	54700909
16	45969706
17	38755319
18	33110296
19	28782153
20	25475732

Peak *k*-mer depth: 10

Figure 5.7 Frequency and depth of *k*-mer counts generated from jellyfish histo to draw the slope representing peak *k*-mer depth see section 2.2.15.

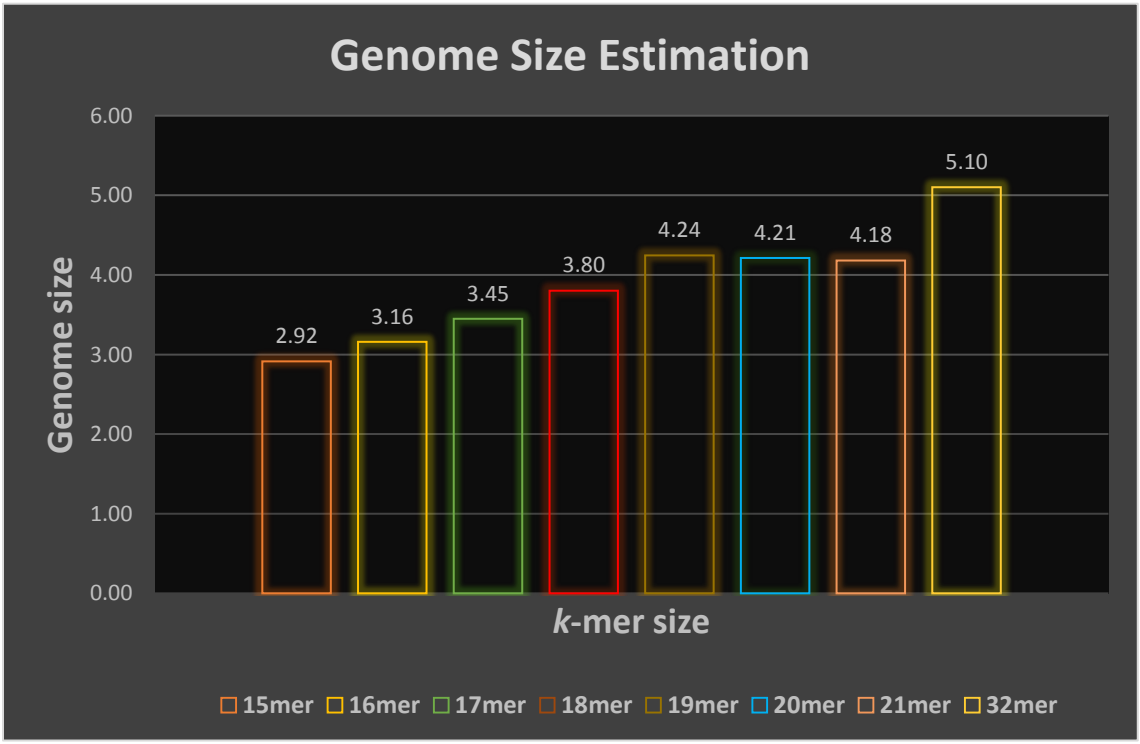


Figure 5.8 Different short motifs (mers) by *k*-mer freuquency estimated different sizes of sheep genome. The optimum *k*-mer lengths with sharpest single-copy peaks were between 15 and 18.

5.6 Characterization of chromosomal locations by *in situ* hybridization

5.6.1 Repeat elements, primer design and labelling

PCR primers spanning consensuses of cluster and contig were designed with various product sizes Table 5.1. After checking that single bands of the expected size were amplified, PCR products of these clusters and contigs were labelled and used as a probe for *in situ* hybridization experiments to investigate their genomic distributions on sheep chromosomes. In some cases, probes were also hybridized to metaphases from *Bos taurus*. The abundances, homologies and graph layouts of RepeatExplorer clusters of dispersed repeats used as probes are shown in Table 5.5. While, the genomic proportions of probes are mentioned in section 5.4.11. *In situ* figures show DAPI counterstained chromosomes and *in situ* hybridization with labelled probes detected with red or green fluorescence; the images are shown separately and as overlays, but in some cases uneven illumination is evident so single-colour images were always examined for several metaphases from each probe.

5.6.2 Non-LTR retrotransposon RTE

LINEs_RTE related repeat fragments from RepeatExplorer clusters CL5C418_RTE, CL8C129_RTE, CL5C464_RTE, CL8C95_RTE, CL7C43_RTE, CL10C107_RTE, CL4C63_RTE and contigs 32merC15_RTE, 32merC12_RTE and 32merC31_RTE from the *k*-mer analysis were used as probes Figures 5.9-5.12. All showed a widespread distribution over the genome, but there were notable differences in uniformity and strength of hybridization along and between individual chromosomes, particularly related to the submetacentric, X and Y sex chromosomes, and the centromeric regions. Copy number analysis showed these LINEs_RTE sequences represented between 0.5% and 1.75% of the genome Table 5.5; *in situ* hybridization as used here is at best semi-quantitative, but there was no suggestion that differences in hybridization pattern related to the relatively small, <4-fold, range in copy number.

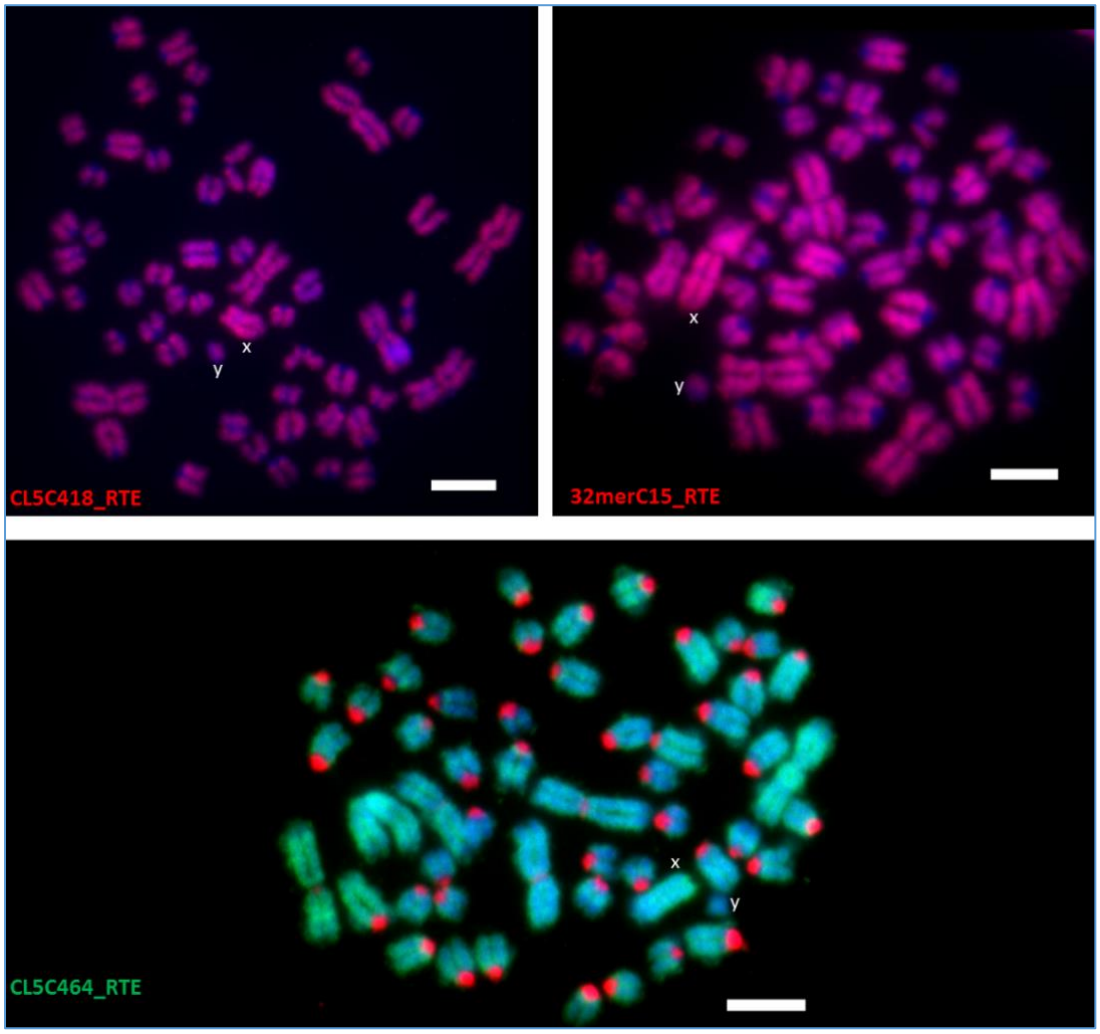


Figure 5.9 Probe CL5C418_RTE hybridized relatively equally to all autosomes and the Y chromosome, and the probe signal was stronger on the X chromosome. Signals were dispersed and excluded from centromeric domains of all acrocentric chromosomes. Similar hybridization results were obtained from *k*-mer probe 32merC15_RTE. Signals of CL5C464_RTE were found over all chromosomes, but with slightly stronger hybridization to the X and three submetacentric chromosome pairs.

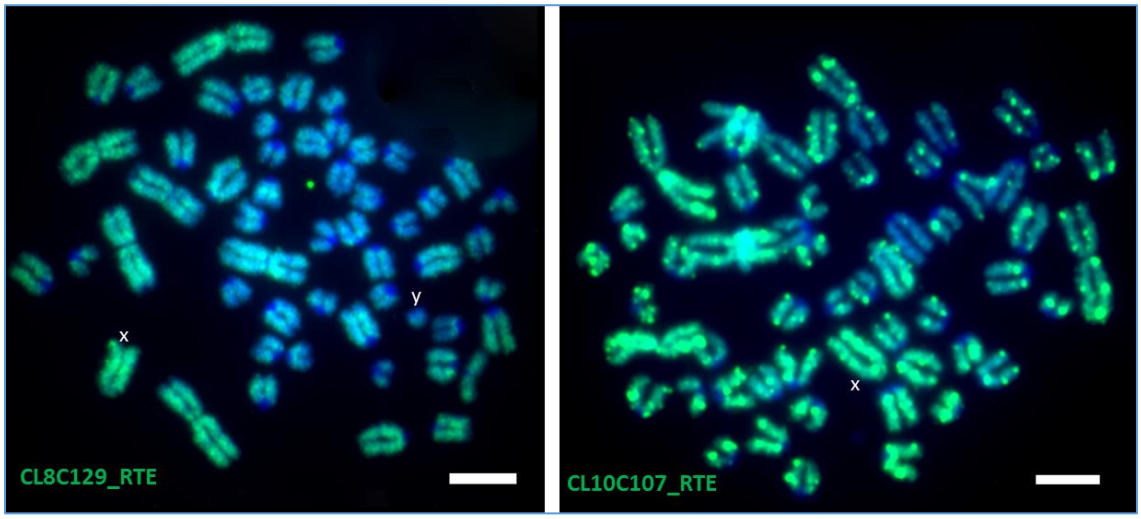


Figure 5.10 Signals of CL8C129_RTE were found over all chromosomes, but with slightly stronger hybridization to the X and three submetacentric chromosome pairs. Probe CL10C107_RTE showed bands on many acrocentric chromosomes and was in low abundance on two acrocentric pairs. There was a more uniform and stronger signal on two pairs of submetacentrics and the X chromosome. Signals were absent at centromeric domains of all acrocentric chromosomes.

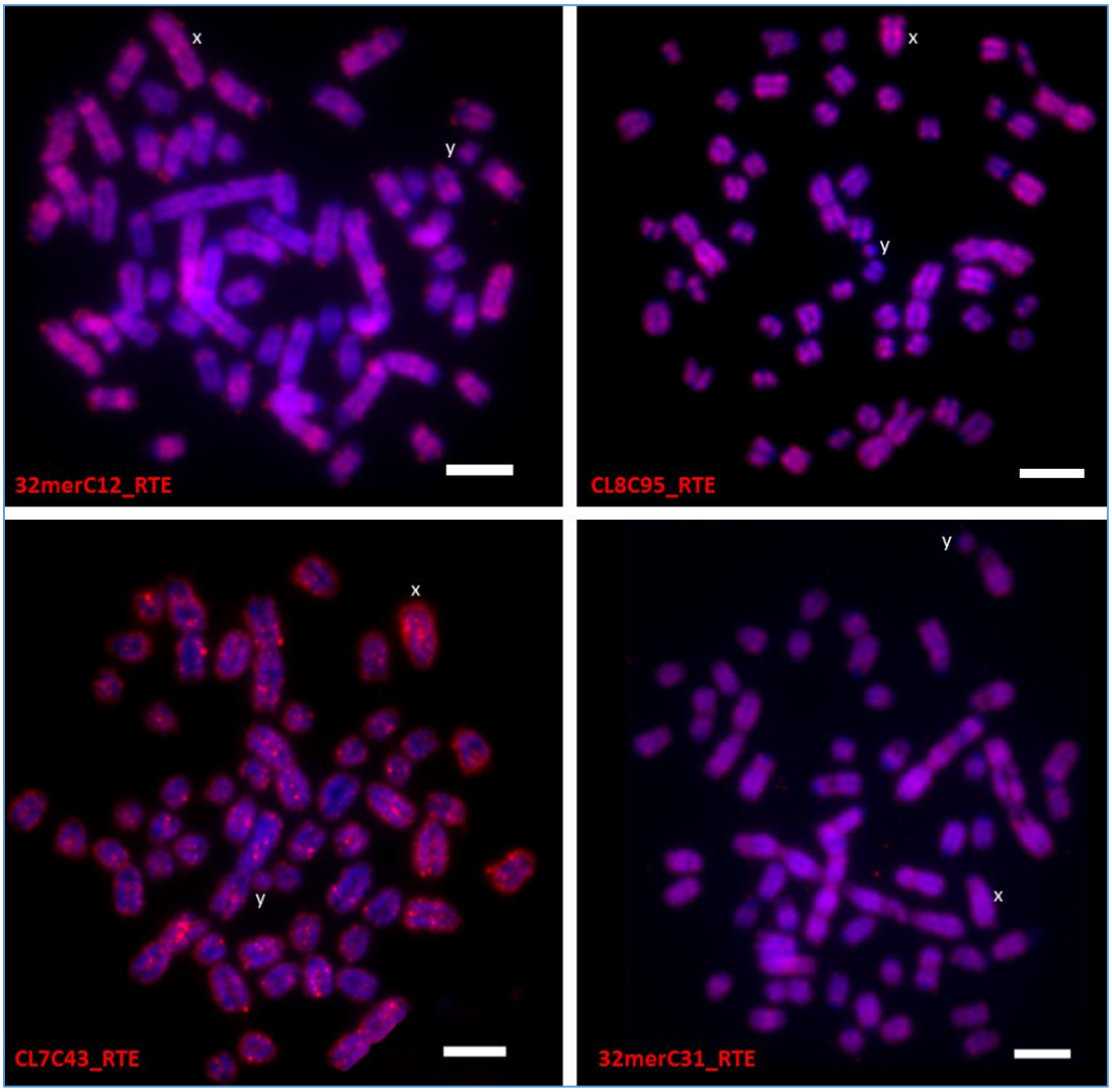


Figure 5.11 32merC12_RTE showed more variable hybridization strengths, with some acrocentrics showing little hybridization, and some stronger bands on many chromosomes. Probes of k-mer 32merC12_RTE, CL8C95_RTE, CL7C43_RTE and 32merC31_RTE from RepeatExplorer clusters were widespread throughout chromosomes except centromeric domains. Signals of the 32merC31_RTE probe were found on some chromosomes with bands and gaps, as in DAPI DNA staining. Half of Y chromosome was labelled with probe of CL8C95_RTE.

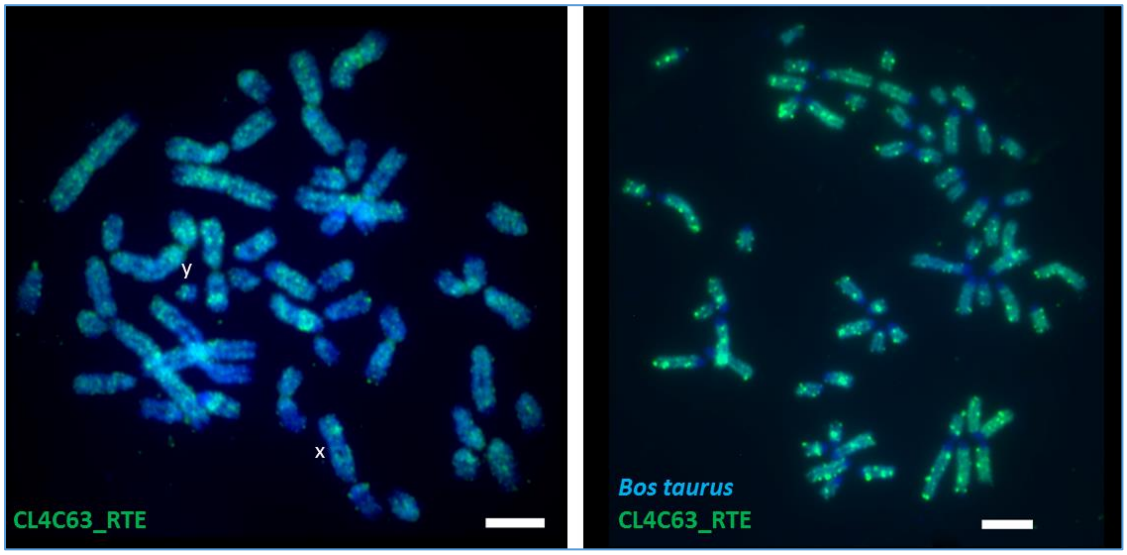


Figure 5.12 Probe CL4C63_RTE was hybridized with sheep and cattle metaphases. Signals were dispersed over all chromosomes including sex chromosomes of both sheep and cattle. Signals were apparent over centromeres of sheep chromosomes, but were excluded from centromeric domains of cattle chromosomes. Some small dots were seen on cattle chromosomes.

5.6.3 **Non-LTR retrotransposon LINEs_L1 repeats.**

The genomic distribution of LINEs_L1 repeats on sheep chromosomes was investigated using probes from several clusters of RepeatExplorer and one contig of *k*-mer results Figures 5.13-5.15.

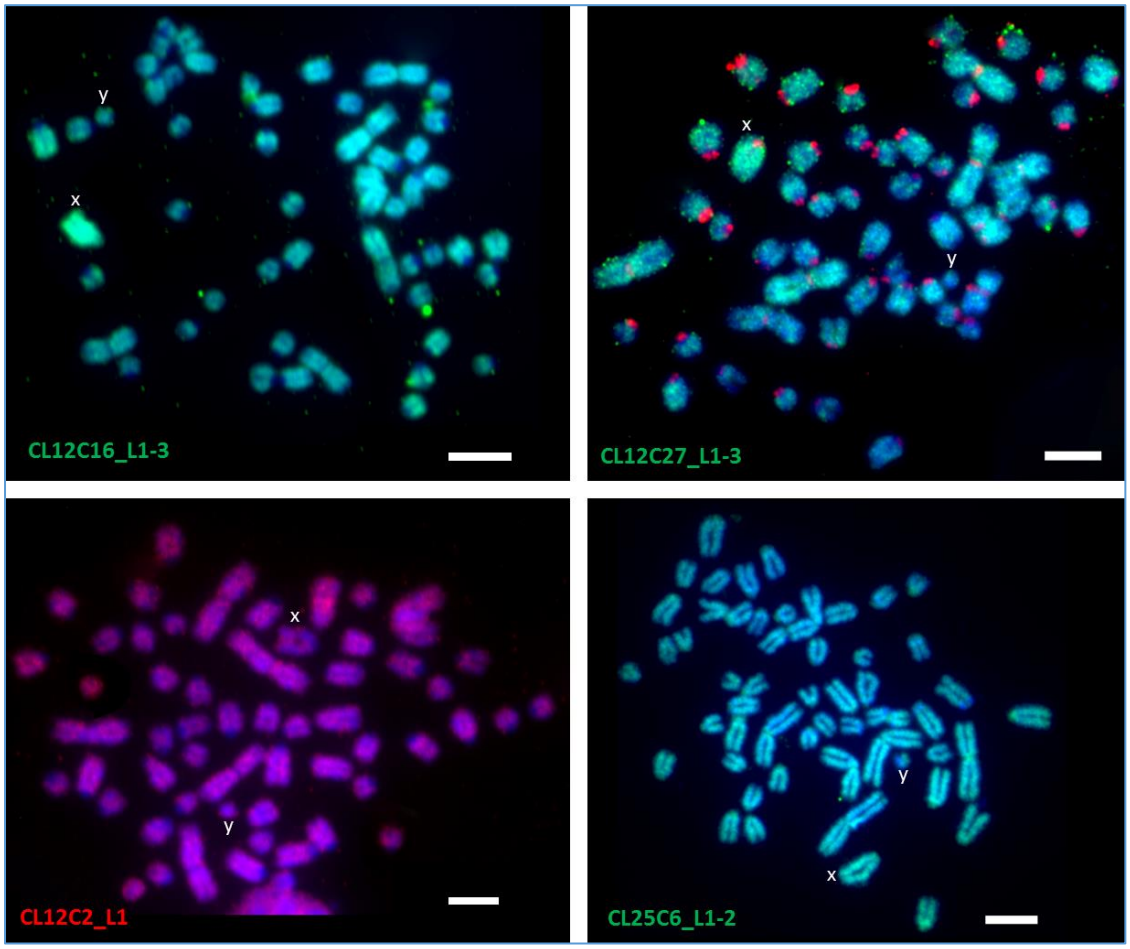


Figure 5.13 Three probes from CL12 such as CL12C16_L1-3, CL12C27_L1-3 and CL12C2_L1 showed dispersed signals distributed over all chromosomes including the Y chromosome with less centromeric hybridization. Relative to autosomes, probes CL12C16_L1-3, CL12C27_L1-3 were amplified on X chromosomes, while probe CL12C2_L1 showed weaker hybridization to X. While, the low abundance CL25C6_L1-2 (0.005% of the genome; Table 5.5) was relatively strong and uniform over all chromosomes.

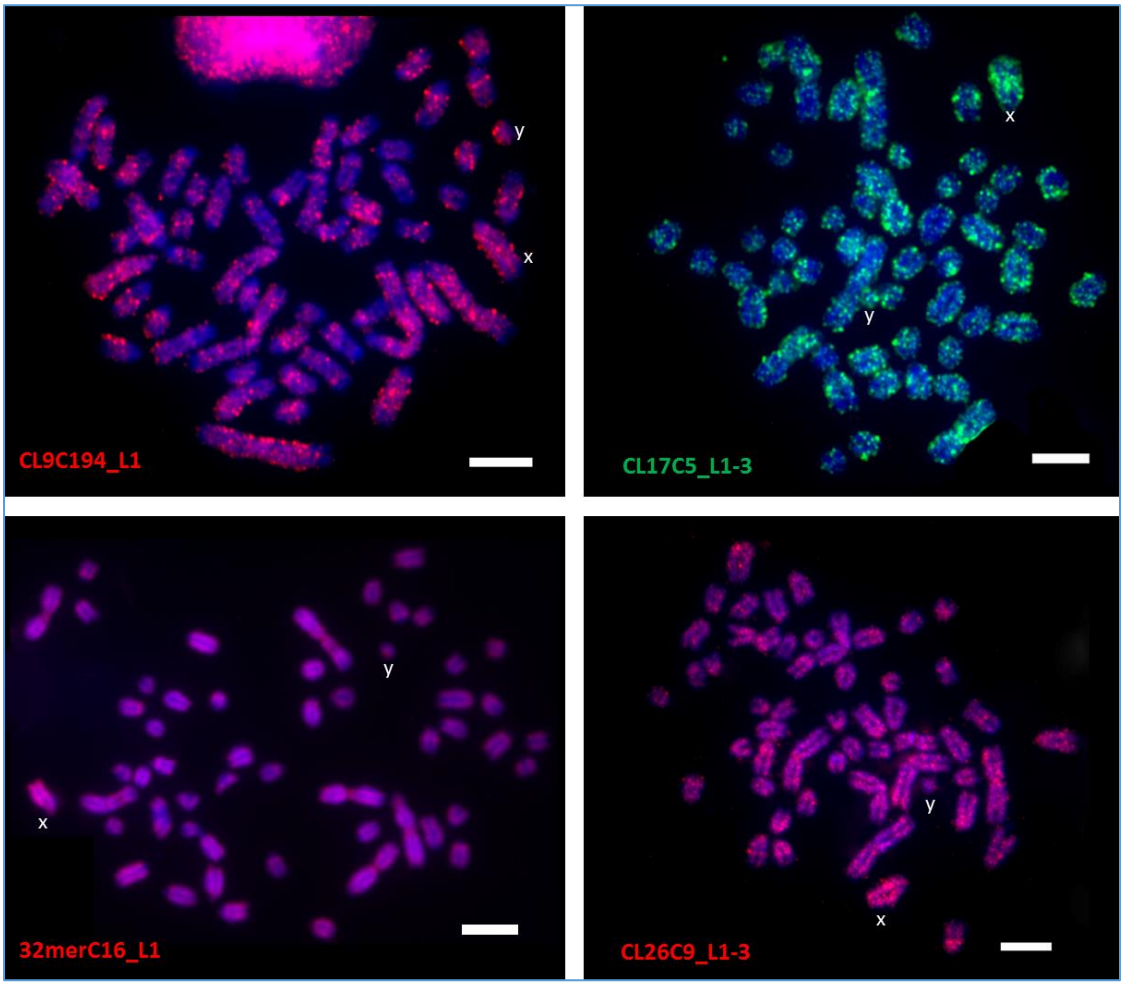


Figure 5.14 Probes from other LINEs related clusters such as CL9C194_L1, CL17C5_L1-3, CL26C9_L1-3 of RepeatExplorer and 32merC16_L1 of *k*-mer analysis showed generally similar dispersed hybridization patterns, with only small differences detectable on X and Y chromosomes. CL9C194_L1 has some additional sites on telomeric domains while centromeric areas of other chromosomes were nearly unlabeled. CL17C5_L1-3 was scattered over all chromosomes showing multiple small dots of signals. In general, lower abundance sequences showed more dots rather than uniform dispersed hybridization patterns (compare also LTR-elements).

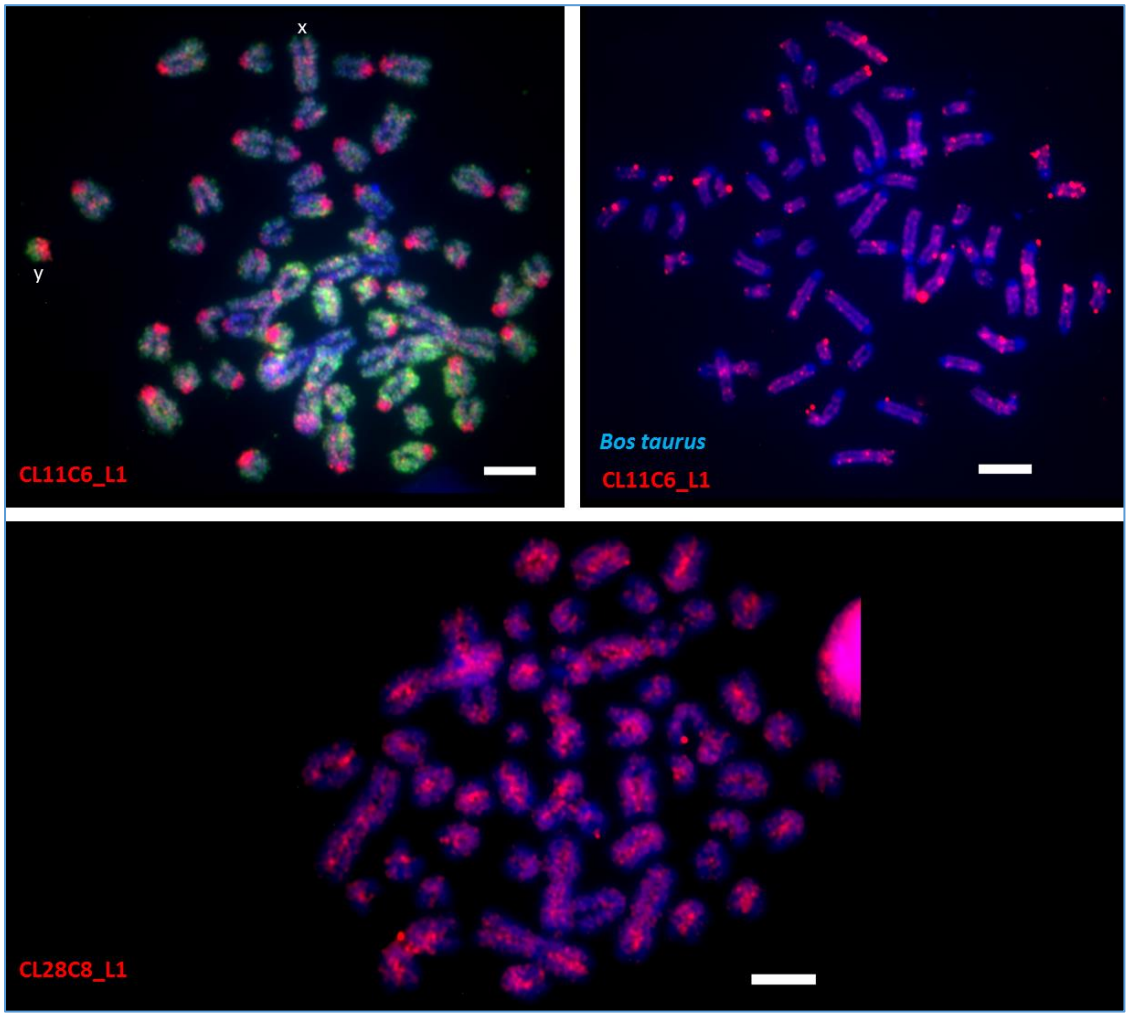


Figure 5.15 Probe CL11C6_L1 from RepeatExplorer was used for *in situ* hybridization against sheep and cattle chromosomes. In sheep, CL11C6_L1 hybridized in broad centromeric pattern with slightly dispersed over all chromosomes. In contrast, in cattle, the same probe CL11C6_L1 was present at intercalary bands on most chromosomes with strong signals at some but not all centromeric domains. Signals were dispersed over sex chromosomes of both sheep and cattle. Probe CL28C8_LINE1 that matched high sequence similarities to LINEs related dolphin species (see section 5.4.9.5) showed dispersed intercalary hybridization patterns.

5.6.4 **SINE repeats**

Seven probes representing different RepeatExplorer clusters showing homology to SINEs were used for characterization of their chromosomal distribution Figures 5.16 & 5.17. Different *in situ* hybridization patterns of dispersed to centromeric signals of probes were observed.

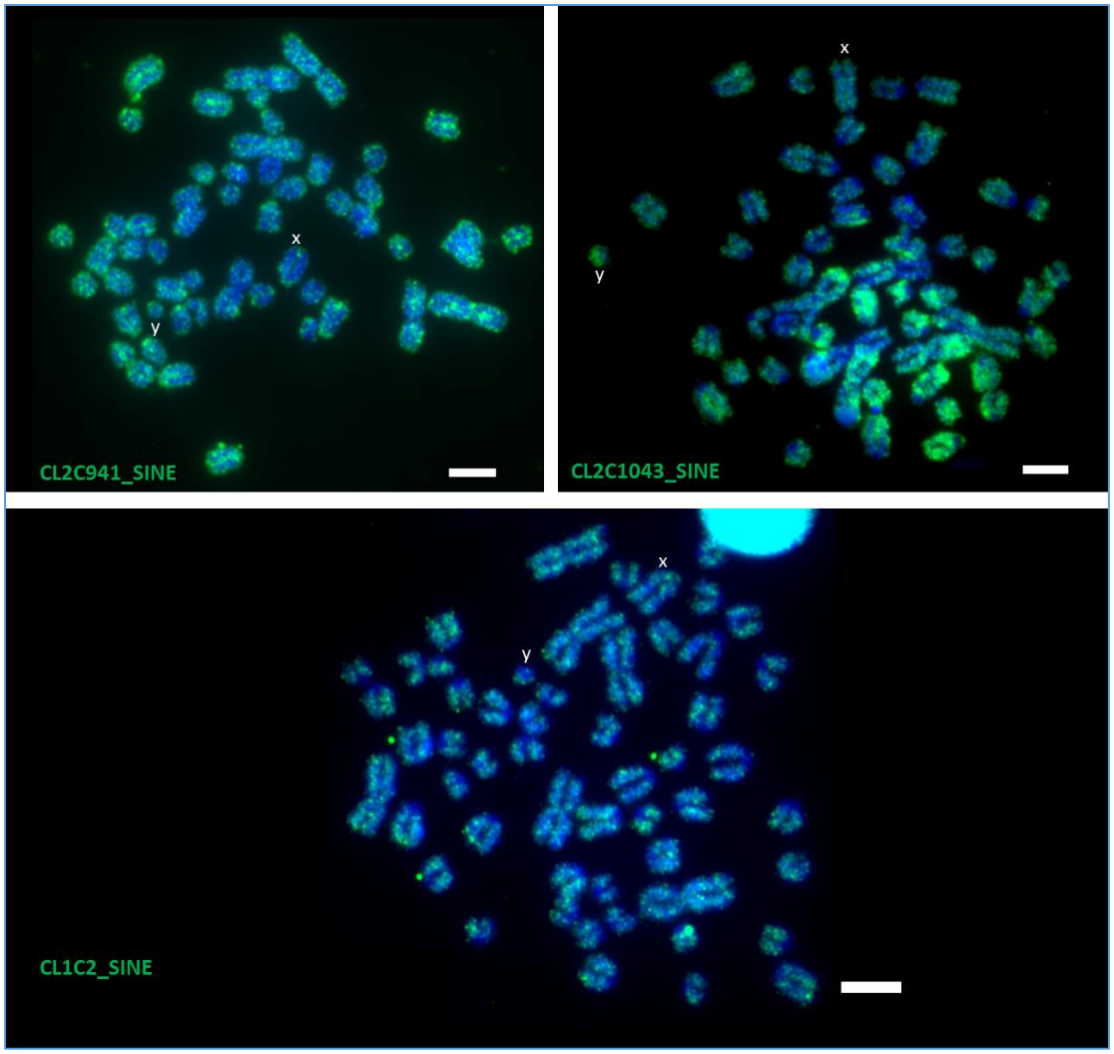


Figure 5.16 Signals of three probes CL2C941_SINE, CL2C1043_SINE and CL1C2_SINE were diffused over all chromosomes including sex chromosomes X and Y. In contrast, other four probes (see below; Figure 5.17) showed mostly signals at or near centromeric domains.

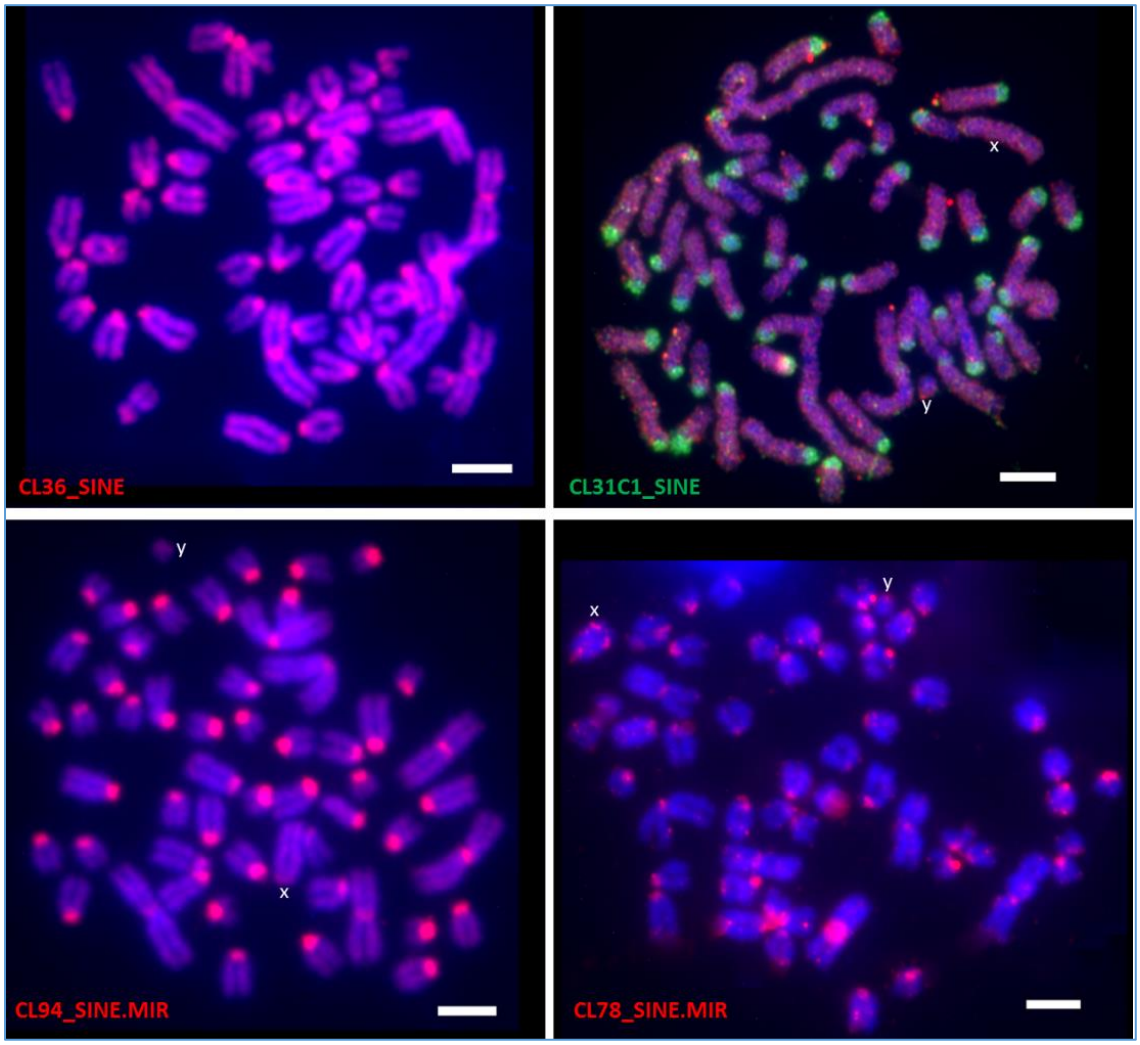


Figure 5.17 Signals of probes CL36_SINE.BovA, CL31C1_SINE.tRNA and CL94_SINE.MIR were strongly hybridized to the centromeric area of all sheep chromosomes except sex chromosomes where they have more dispersed signals. In comparison to the other SINE probes, probe CL78_SINE.MIR showed weaker centromeric signals over all chromosomes.

5.6.5 **Repeats with no known homology**

Outcomes of RepeatExplorer were characterized by some clusters of unclassified sequences were described as “low complexity or “simple repeat” and not found in Repbase, RepeatMasker and NCBI. Genomic proportion of these clusters was low compared to the other abundant repeats found in whole sequencing raw reads of sheep genome Table 5.1. Clusters of low complexity and simple repeats were enriched of GC

content about 64-74%. The genomic abundance of “low complexity or “simple repeat” were characterized by using four probes for *in situ* hybridization Figure 5.18.

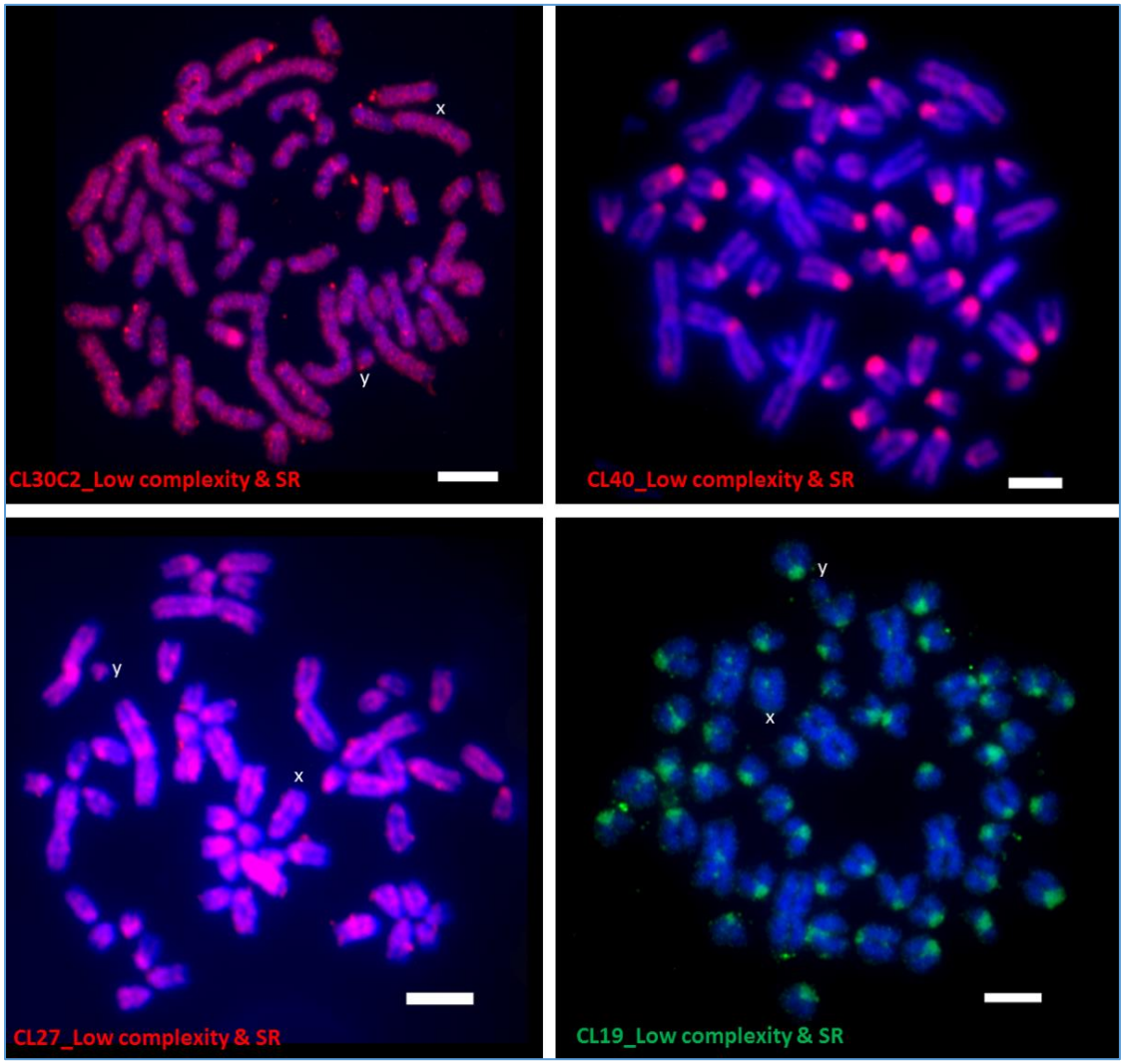


Figure 5.18 Signals of probe CL30C2_Low_complexity & SR and CL27_Low complexity & SR were widespread distribution over all chromosomes including sex chromosomes. In contrast, probes CL40_Low complexity & SR and CL19_Low complexity & SR were hybridized mostly to the centromeric regions of acrocentric chromosomes while slight centromeric to dispersed signals were found in submetacentric and sex chromosomes.

5.6.6 **Other dispersed repeats**

Sequence of other clusters CL43_DNA transposons, CL46_DNA.hAT.Charlie, CL50_Interspersed repeat and CL85_Non LTR was also unknown in databases of Repbase, RepeatMasker and NCBI. Alternatively, consensus of each cluster was blasted

against TEclass tool (Abrusán et al, 2009). TEclass tool classified each cluster into different classes of transposable elements considered CL43 as a DNA transposons, CL46 as a DNA.hAT.Charlie, CL50 as a interspersed repeat and CL85 as a Non-LTR repeats (Figures 5.19 & 5.20). Probe representing each cluster produced different patterns of hybridization.

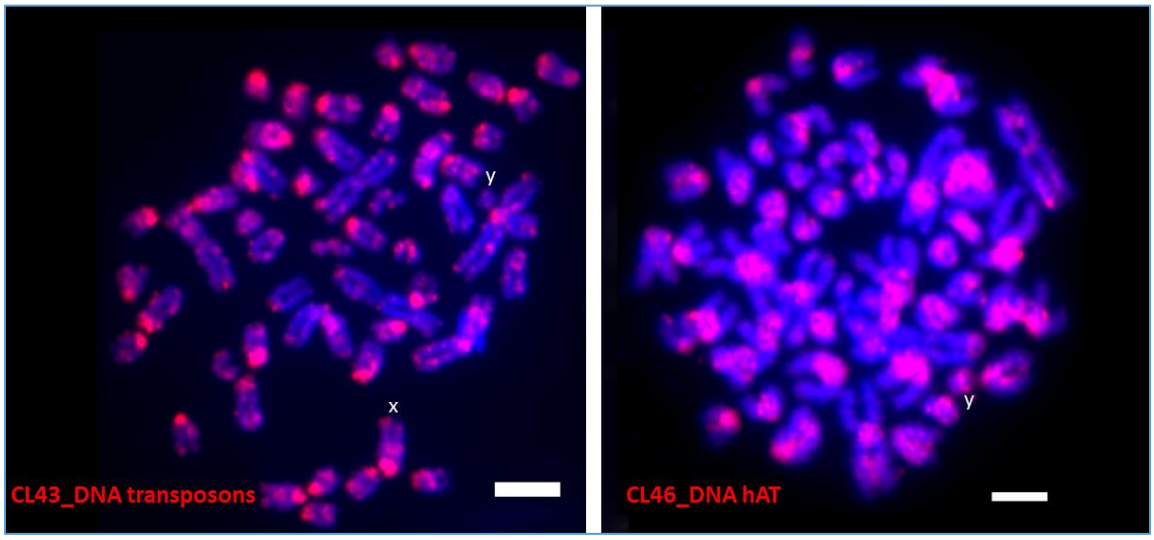


Figure 5.19 Probe CL43_DNA transposons was hybridized to all chromosomes dispersed as dots on some but others have centromeric signals. Signals were detected in both X and Y chromosomes. Probe CL46_DNA.hAT.Charlie was labelled broadly at or near centromeric area and strong signal was present over Y chromosomes.

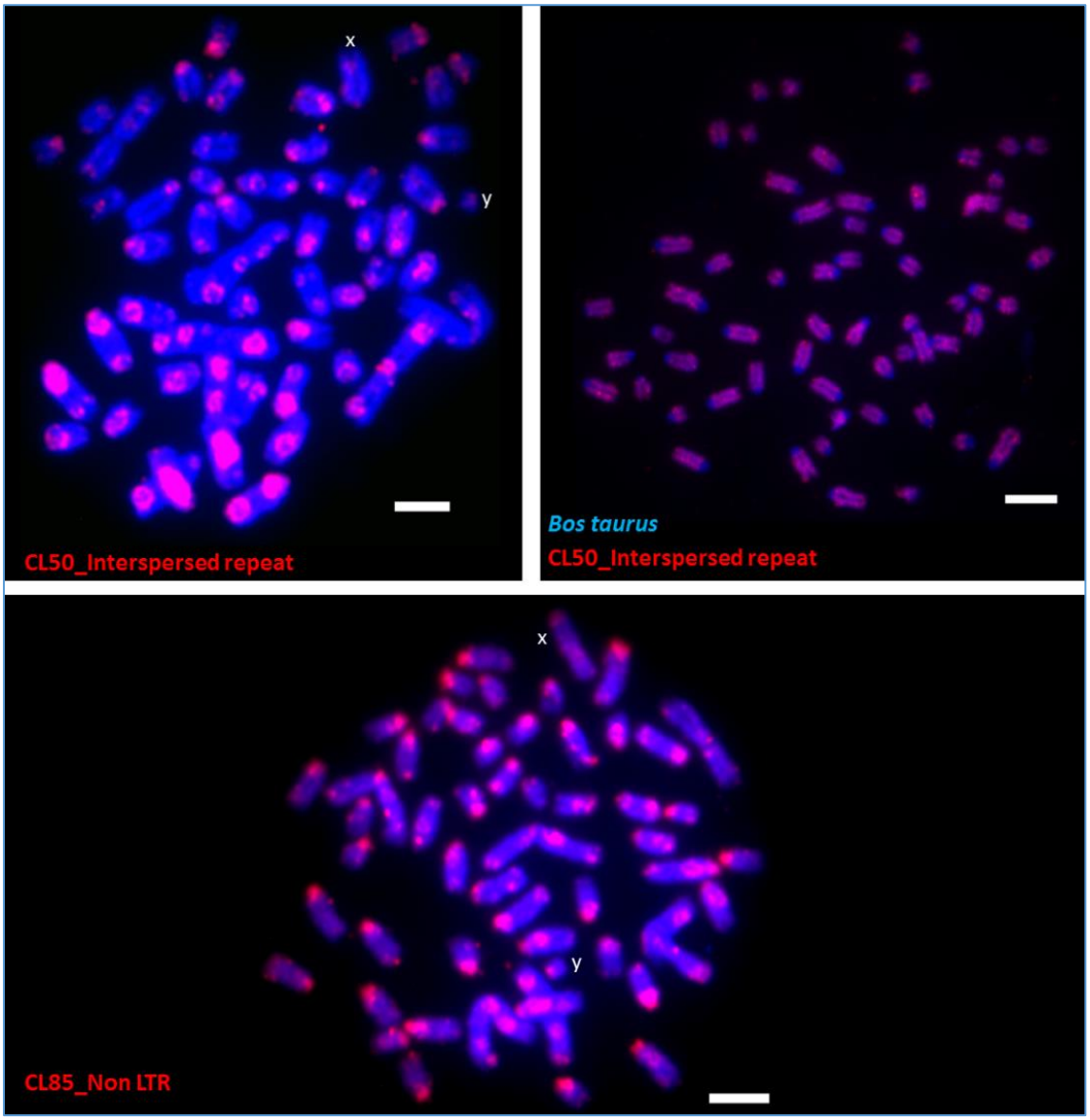


Figure 5.20 Probe CL50_Interspersed repeat was used against sheep and cattle chromosomes. In sheep chromosomes, signals were apparently centromeric to telomeric position on some chromosomes. Over submetacentrics two to four dots were seen. While, in cattle, probe was dispersed over all chromosomes with exception of signals from centromere of all chromosomes.

Similar to CL50_Interspersed repeat, probe CL85_Non LTR was hybridized to all chromosomes distributed as dots over centromeric to telomeric domains of sheep chromosomes. Signals were also slightly centromeric to X chromosome while dispersed over Y chromosome.

5.6.7 **Non-coding RNA**

Consensus of contig7 resulted from *k*-mer analysis was highly homologous to non-coding RNA sequences of *Ovis aries* (uncharacterized LOC101104348, ncRNA). Thus, to investigate their genomic distribution in sheep chromosomes, probe 32merC7_ncRNA was used for *in situ* Figure 5.21.

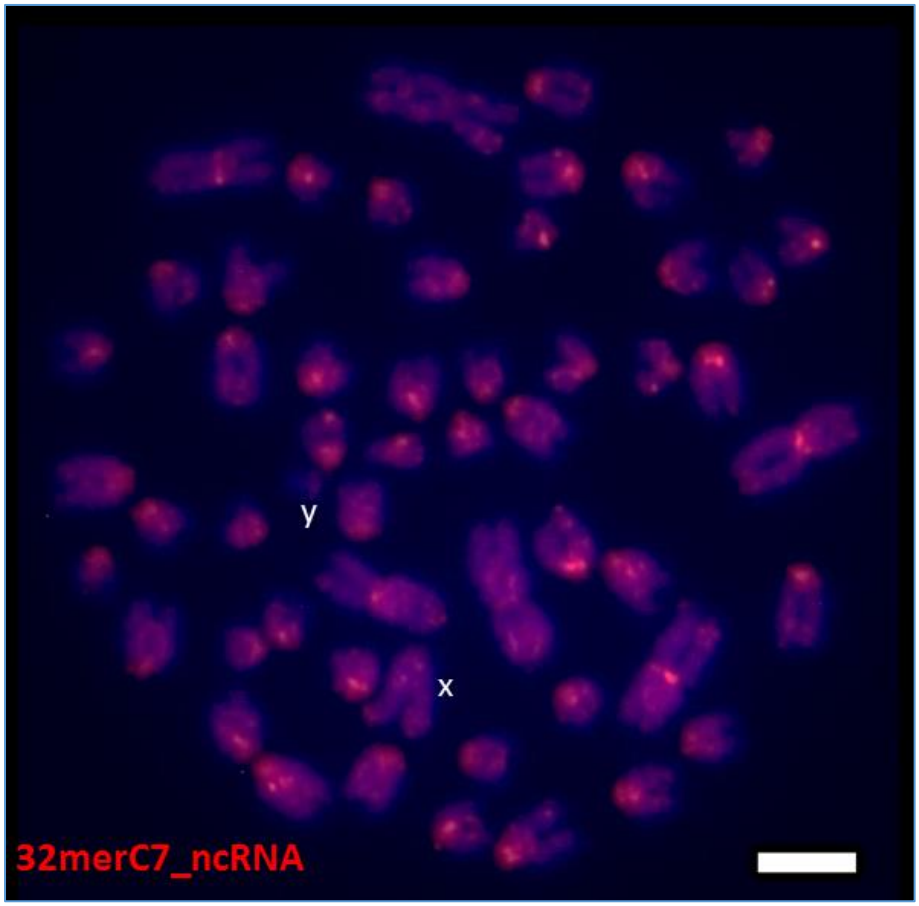


Figure 5.21 Centromeric signals with some diffused patterns over all chromosomes were detected. Signals were dispersed but devoid from centromeric domains of the largest pairs of submetacentric chromosomes while, apparently centromeric signals were seen in the other two pairs of submetacentric chromosomes. In terms of sex chromosomes, probe was dispersed on X while slight signal was detected on Y chromosome.

5.7 Discussion

5.7.1 Abundance of dispersed repeats in sheep genome

Graph-based read clustering and *k*-mer frequency analysis of unassembled raw reads from high-throughput DNA sequencing showed that a major part of the sheep genome was represented by repeated DNA sequence motifs, as found in all mammals (Biscotti *et al.* 2015b). Graph-based clustering was an efficient approach for repeat identification (Novák *et al.* 2010; Pagán *et al.* 2012; Novák *et al.* 2013). The repetitive DNA, including dispersed and tandemly repeated DNA sequences, occupied about 30% of sheep genomes in the samples analyzed Table 5.3 and Appendices 5.2 & 5.3. In total, 34% of sheep NGS data were representing repetitive DNA population see section 5.4.1. This repeat abundance is similar to that reported in dog and mouse genomes (35% to 38%), but somewhat lower than the repetitive proportions of human and bovine genomes (45% to 46.5%) using alternative techniques often involving assembly (Lander *et al.* 2001; Mouse Genome Sequencing 2002; Lindblad-Toh *et al.* 2005; Adelson *et al.* 2009).

The most abundant repetitive elements within the sheep genome were represented by non-LTR class I retrotransposons (LINEs_L1, LINEs_RTE and SINEs). These non-LTR retrotransposons accounted for approximately 20% of the sheep genome. The LINEs_L1 repeats were less abundant and occupied 2.75%, while the LINEs_RTE class were more abundant and engaged 12%. SINEs occupied 4.5% of the sheep genome Table 5.3 and Appendices 5.2 & 5.3. Lenstra *et al.* (1993) described three diverse families of SINEs (Bov-B, Bov-tA & Bov-A2) with a genomic proportion about 3.9% in total. While each study uses a different analytical approach, super families of LINE and SINE repeats make different contributions to various mammalian genomes. Adelson *et al.* (2009) found that both LINE and SINE elements constituted 25% of all repeat composition in the bovine genome and 17% of the human genome comprises of LINE repeats (LINE-1; L1; out of a total non-LTR retrotransposon proportion of 30%) (Penzkofer *et al.* 2017). In other mammalian genomes sequenced so far, 29% of repeat content of opossum genome and 18% of the dog genome were comprised of LINE repeats (Lindblad-Toh *et al.* 2005; Mikkelsen *et al.* 2007), and in chicken genome, they are 6.4% (Hillier *et al.* 2004).

Only 0.015% of the sheep genome was identified as class II transposable elements (DNA transposons; DNA.TcMar.Mariner, DNA.hAT, DNA.PiggyBac and RC.Helitron), notably less than in human (3%), bovine (2%) and mouse (1%) genomes (Lander *et al.* 2001; Mouse Genome Sequencing 2002; Adelson *et al.* 2009).

As expected, other abundant motifs (identified in RepeatExplorer and *k*-mer analyses) were related to tandem satellite repeats (see Chapter Four) or rDNA genes. The complete sequence (1869bp) of *Ovis aries* 18S ribosomal RNA gene could be assembled with a genomic proportion of 0.01%, c. 150 copies per genome, which is somewhat lower than the 300–400 copies reported in human genomes (Henras *et al.* 2015). Assembly of the sheep 5S ribosomal RNA gene gave 16000 copies. SINEs originating from 5S RNAs has been reported in a few mammals and some fish (Deragon & Zhang 2006; Kriegs *et al.* 2007; Kramerov & Vassetzky 2011).

Notably, in our analysis of the repetitive DNA of sheep, one abundant cluster (genomic proportion 0.30%) had no homology to characterized sequences: others were satellite DNA (including rDNA) or transposable element-related. It could be speculated that this unidentified cluster is a retroelement LTR (or solo-LTR) with minimal numbers of its parental (or ancestral) retroelements and no characteristic sequence features.

5.7.2 Chromosomal localization of repeats

Fragments from a representative range of the repetitive clusters were labelled and used as a probe for *in situ* hybridization to confirm their abundance and examine their genomic distribution in sheep chromosomes. In some cases, probes were also hybridized in cattle chromosomes. In most cases, only one probe from each cluster was used.

The abundant (1.7% to 0.5% of the genome) clusters of the RTE retroelements showed small but distinct differences in the hybridization patterns (Figures 5.9-5.12). With a dispersed distribution, less abundant non-LTR retroelements (0.4 to 0.005% of the genome) showed less distinctive patterns.

For CL12, a non-LTR retroelement, three probes were used which gave closely similar hybridization patterns (Figures 5.13), although CL12C2_L1 showed slightly less abundance on the X chromosome than the other two probes.

In general, lower abundance sequences showed more dots rather than uniformly dispersed hybridization patterns (compare also LTR-elements), although the low abundance CL25C6_L1-2 (0.005% of the genome; Table 5.5) was relatively strong and uniform on the chromosomes; it would be possible for a cluster domain to be common between two sequences but this would firstly collapse the two clusters into one, and secondly the high abundance of the domain would be reflected in the 'map-to-reference' (with the probe sequence as the reference) approach used to measure copy numbers.

5.7.3 Genomic distribution of non-LTR retrotransposon LINEs

Probes of LINEs_RTE repeats (Figures 5.9-5.12) and LINEs_L1 repeats (Figures 5.13-5.15) were used for *in situ* hybridization, and variable patterns of dispersed repeats over sheep chromosomes were seen. Genomic distributions of LINEs_RTE and LINEs_L1 repeats were dispersed throughout whole chromosomes including the sex chromosomes. These results reflect the bioinformatics analysis where LINE repeats, including RTE and L1, were found to be the most abundant repeat in sheep genome (Table 5.3). Interestingly, some probes of both RTE and LINEs, such as CL5C418_RTE; 32merC15_RTE; CL8C129_RTE; CL5C464_RTE; CL12C16_L1-3 and 32merC16-L1 (Figures 5.9, 5.10, 5.13 and 5.14), were strongly hybridized in the X chromosomes showing intense hybridization signals. Furthermore, signals from all of the probes were detected in Y chromosomes. It has been found that LINE elements prefer accumulation on sex chromosomes in both mice and human genomes (Mouse Genome Sequencing 2002). For example, in the mice genome, the density of lineage-specific LINE (L1) copies in the X chromosome is almost twice that found in the autosomes. In humans, the sex chromosomes also display more robust preference for LINE repeats with 18.0% on Y and 17.5% on X in comparison with only 7.5% on the autosomes. It has been proposed that LINE elements could play an essential role in the inactivation of the X chromosomes (Lyon 1998). Bailey *et al.* (2000) pointed out that "the non-random properties of LINE

distribution on the X chromosome provides strong evidence that LINE elements may serve as DNA signals to propagate X inactivation along the chromosome", which is perhaps the case here. One of the possible explanations is that the poverty of genes of the Y chromosomes allows for a bigger insertion and accumulation of LINE repeats. Failure and disability of recombination in purification of deleterious mutations could be one of the reasons. Additionally, the presence of a three-fold greater intensity of full-length L1 copies in the Y chromosome could be eliminated somewhere else in the genome (Boissinot & Furano 2001). Furthermore, more recently, Chow *et al.* (2010) indicated that "LINEs may facilitate X chromosome inactivation at different levels, with silent LINEs participating in assembly of a heterochromatic nuclear compartment induced by Xist RNA coats, and active LINEs participating in local propagation of X chromosome inactivation into regions that would otherwise be prone to escape".

Different patterns of hybridization of RTE repeat probes, like CL4C63_RTE, were found when signals were dispersed on sheep and cattle chromosomes. For sheep chromosomes, signals were concentrated on or near centromeric regions whereas, probe signals were excluded from the centromere of cattle chromosomes (Figures 5.12). Although bioinformatics analysis showed high sequence identities (97% between LINEs_RTE consensus of *Ovis aries* and *Bos taurus)*, their genomic structure was different: it seems that the genomes of sheep and cattle evolved but conserved their RTE sequences to some degree since the time of their speciation. It has been suggested that LINEs_RTE are transmitted horizontally from reptiles to marsupials (Gentles *et al.* 2007) and to ruminants (Kordis & Gubensek 1998; Kordiš & Gubenšek 1999), so occasional transfer cannot be ruled out.

Sequences of CL28C8_LINE1 from RepeatExplorer were matched to the sequences of LINE repeat (L1-1_Ttr) of dolphin species (*Tursiops truncates*). Its probe CL28C8_LINE1 was dispersed over all chromosomes, including sex chromosomes, plus some additional intercalary signals within submetacentric chromosomes (Figure 5.15). Several studies investigated the genetic distances and evolutionary rates between artiodactyl and cetacean species. Based on phylogenetic relationships of mitochondrial genomes, sheep and goat from the ruminants were found to be a sister group of the fin and blue whale

(Hiendleder *et al.* 1998a; Arnason *et al.* 2000). Furthermore, Graur and Higgins (1994) analyzed data of protein and mitochondrial sequences of different species from both cetaceans and artiodactyls, and indicated that cetaceans are closely related to members of the suborder Ruminantia of artiodactyls and are within the phylogenetic tree of Cetartiodactyla or artiodactylamorpha. This discovery of LINE element sequences with high similarity between sheep and cetaceans supports their common ancestry and conservation of the element over 50 MYA.

Non-LTR retrotransposon consensus of L1-sheep were assembled from whole sequencing raw reads corresponding to the non-LTR Retrotransposon (L1-BT) of *Bos taurus* (8390bp). Probes of *k*-mer frequency 32merC16_L1 and RepeatExplorer clusters CL12C2_L1, CL9C194_L1 and CL11C6_L1 were representing non-LTR retrotransposon L1-sheep. Probes 32merC16_L1, CL12C2_L1 & CL9C194_L1 showed different hybridization patterns of dispersed signals, while another probe CL11C6_L1 demonstrated more centromeric to dispersed signals on all sheep chromosomes (Figures 5.13, 5.14 and 5.15). However, the same probe CL11C6_L1 was dispersed but excluded from centromeric area of cattle chromosomes (Figure 5.15). These results indicated the different genomic organization of LINE repeats, although high sequence identities of 91% were found between L1-sheep and L1-BT of cattle. In Normande cattle, Girardot *et al.* (2006) found that sequences of non-LTR retrotransposon (L1-BT) inserted in the 5'-genomic sequence of their Agouti gene stimulated variable expressions of alternative transcripts, which are directed by the promoter of LINE repeat (L1-BT). This could be one of the main reasons why the Normande cattle have the brindle colour coat pattern. Furthermore, in mice, Michaud *et al.* (1994) and Argeson *et al.* (1996) demonstrated that the insertion of LTR retrotransposon in Agouti alleles caused a range of mosaic phenotype. The expression of the Agouti gene has been demonstrated in diverse tissues of bovine (Girardot *et al.* 2005) and adipocytes of bovine (Sumida *et al.* 2004). Moreover, Girardot *et al.* (2006) suggested that the Agouti protein is over-expressed in Normande cattle and could be involved in the synthesis of fatty acid. In sheep, the association of polymorphisms of the agouti signaling protein (ASIP) gene with the coat colour has been analyzed (Norris & Whan 2008; Han *et al.* 2015). However, the expression and relationship of the agouti signaling protein (ASIP) gene has not been investigated in

other sheep tissues. Thus, further investigations are required to determine if the expression of the Agouti gene in sheep has correlation with the production of milk and meat, in particular the fat content of milk. Here, it was found that the full length of non-LTR retrotransposon L1-sheep consensus contained two open reading frames which originated from Iraqi Kurdistan sheep breeds (Appendix 5.11). However, it is not clear whether the LINE repeats identified here could have an impact on the coat colour or fat content of milk in sheep breeds.

In mammals, centromeric accumulations of non-LTR retrotransposons like probe CL11C6_L1 in the chromosomes of sheep are unusual. However, some cases have been reported. For example, centromeric distributions of LINE-1 were found in the mammalian karyotypes of Afrotheria and Xenarthra (Waters *et al.* 2004). Similarly, Sotero-Caio *et al.* (2015) demonstrated the centromeric distribution of LINE-1 in the chromosomes of the phyllostomid bats. Recently, de Souza *et al.* (2017) *in situ* hybridized LINE-1 to the chromosomes of A. planirostris, and centromeric positions were found in most autosomes collocated with pericentromeric blocks of heterochromatins.

It has been suggested that such a centromeric distribution of LINE-1 could involve in chromosomal reorganization. Shi *et al.* (2010) and de Souza *et al.* (2017) pointed out that although there is no clear reason why LINE-1 repeats are highly abundant at centromeres, other factors such as gene conversion could promote massive colonization of LINEs at centromeric regions.

5.7.4 **Genomic distribution of non-LTR retrotransposon SINE repeats**

For *in situ* hybridization, seven probes of different classes of SINE repeats including SINE (BOVA2), SINE.tRNA, SINE.MIR, SINE2 from ruminants (Bov-tA3) and SINE.BovA. Different *in situ* hybridization patterns showing dispersed to centromeric signals were observed (Figures 5.16 & 5.17). It seems that SINE organization in sheep chromosome is different from one class to another. For instance, diffused signals were found of three probes CL2C941_SINE, CL1C2_SINE Ruminant and CL2C1043_SINE while, four other probes (Figure 5.22; CL36_SINE.BovA, CL31C1_SINE.tRNA, CL94_SINE.MIR and CL78_SINE.MIR were hybridized mostly at or near to centromeric locations. It seems that

the hybridization results of SINE repeats are compatible with their genomic abundance, as SINE repeats populated more than 4% of the sheep genome. Mammalian genomes contain considerable amounts of SINEs sequences (Lenstra *et al.* 1993; Lander *et al.* 2001; Vassetzky & Kramerov 2012). However, probes with less genomic proportion were found with strong signals of hybridization.

As expected but showing the strength of the analysis of whole sequencing raw reads using graph-based clustering method, association was observed between sequences of LINEs_RTE and SINE repeats as both were found in the same cluster (Table 5.3). Similarly, Ohshima & Okada 2005 investigated ruminants and marsupial's genomes, and they stated that there was a symbiosis relationship between SINEs and LINEs_RTE in which the RTE repeat encodes the machinery to transpose SINE repeats including SINE BovA and SINE RTE. Likewise, Gentles *et al.* (2007) indicated that several families of SINEs present in the genome of opossum *Monodelphis domestica* utilized RTE repeats for their mobilization. Thus, it is more likely that SINEs in sheep genome were mobilized by LINEs_RTE repeats.

From the bioinformatics analysis, similar sequences of SINEs elements were found originating from *Bos taurus* and from ancestral ruminants. In this study, three ancestral SINEs and five *Bos taurus* SINE repeats were present in the sheep genome (see section 5.4.10.2). However, the other SINEs sequences of *Bos taurus* (CHR-2_BT, CHRL1_BT, SINE2-1_BT and SINE2-2_BT) were not identified in the sheep genome. Genomic proportions of assembled reads of the ancestral SINEs of *Ruminantia* and *Bos taurus* were about 0.30% and 0.85% respectively. It seems that the sheep genome have conserved some amounts of SINE repeats from their ancestor until present. This comparison indicated that although *Ovis aries* genome matched the ancestral SINE sequences from *Ruminantia*, its genome lacks some other SINEs that originated from *Bos taurus*. Thus, different species of mammals, including closely related animals, could contain different or unique SINE repeats. Various types of SINE have been found specific to such order like SINEC in Carnivora and Alu elements in primates (Kramerov & Vassetzky 2011).

Furthermore, inside Cetartiodactyla, another abundant family of SINE repeats named CHR has been discovered in Cetacea, Hippopotamidae and *Ruminantia* (Shimamura *et al.* 1999). Thus, these SINE repeats including BOVA2, Bov-tA and CHR have been widely used as valuable phylogenetic markers to study relationships within Cetartiodactyla (Shimamura *et al.* 1997; Nomura *et al.* 1998; Nikaido *et al.* 1999; Nikaido *et al.* 2001; Nijman *et al.* 2002; Nilsson *et al.* 2012; Gallus *et al.* 2015).

Likewise, insertion and polymorphism of Alu elements have been exploited for investigation the origin, population structure and demography of humans (Cordaux & Batzer 2009). Additionally, Shimamura *et al.* (1997) identified two families and nine retropositional events of SINEs in the order of Cetacea and Artiodactyla, where these SINEs found more exceptionally in the genomes of ruminants, whales and hippopotamuses. Nijman *et al.* (2002) demonstrated numerous indels, including deletions and insertions, from studying comparative sequencing of SINEs in ruminants. Thus, the retrotransposition of SINEs may be used as an informative marker in order to study and reconstruct phylogenetic relationships between ruminant species at different levels of classification. From the bioinformatics analysis, we found specific SINE sequences such as CL44_SINE.tRNA to the male sheep genome. Gallus *et al.* (2015) suggested that the genomic landscape of transposable elements can rapidly change in any lineage.

5.7.5 Genomic distribution of non-coding RNA sequences

Consensuses of contig7 of *k*-mer frequency CL17C13/CL20C9 HamJ1_Male and KarJ_Female genomes were matched to non-coding RNA sequences (ncRNA) of *Ovis aries*. Probe 32merC7_ncRNA (Figure 5.21) was used for *in situ* hybridization and signals were centromeric but devoid from centromeric domains of the largest pairs of submetacentrics, while apparently centromeric signals were seen in the other two pairs of submetacentrics. In terms of the sex chromosomes, the probe was dispersed over X while only a slight signal was seen on the Y chromosome. Combined sequences of ncRNA and ERV1 were found in the same cluster CL17C13. Accordingly, centromeric signals of ncRNA are in agreement with the genomic distribution of ERV1 Probes (see Chapter Six) where their signals were present at centromeric domains.

In eukaryotes, thousands of long and short non-coding RNAs (ncRNAs) sequences including microRNAs (miRNAs), pseudogenes and circular RNAs (circRNAs) and long non-coding RNAs (lncRNAs) have been found in their genomes. These were found to be involved as key regulators in several cellular processes such as apoptosis, proliferation, and differentiation (Li *et al.* 2013). Moreover, Mattick and Makunin (2006) pointed out that ncRNAs could perform several functions such as splicing, translational inhibition, RNA editing, mRNA destruction and control of chromosome dynamics.

5.8 Conclusions

The analysis of unassembled sequence reads covering the whole sheep genome showed that almost all abundant repeated motifs were recognizable as tandemly repeated satellite sequences or derivatives of transposable elements. Most individual motifs identified had characteristic distributions.

Notably, apart from the sex chromosomes and rDNA, no near-chromosome-specific repeats were found. This suggests that the pools of probes used for chromosome painting and evolutionary studies include mid- to low-copy non-coding sequences, the chromosome-specific non-coding DNA, rather than arising from the expansion and homogenization of repetitive sequences within one chromosome.

As well as the differences between sex chromosomes and autosomes, particularly with respect to LINE abundance (see above), there were some differences in repeat abundance between centromeres of submetacentric and acrocentric autosomes. Potentially, the recombination and homogenization events, involving the removal of dispersed repeats (see also Chapter Four), may be different between acrocentric and submetacentric centromeres. Chaves *et al.* (2005) showed that in cattle and other Bovidae, satellite sequences showed changes following evolutionary or more recent chromosome fusion and fission events.

Chapter 6 Assembly and characterization of the complete endogenous betaretroviruses and endogenous retroviruses related repetitive elements

6.1 Introduction

Jaagsiekte sheep retrovirus (JSRV) is a pathogenic and exogenous retrovirus (Palmarini & Fan 2001) which is the causative agent of major infectious disease in small ruminants called ovine pulmonary adenocarcinoma (OPA); a transmissible lung cancer of sheep (Sharp & Angus 1990; Palmarini *et al.* 1996; DeMartini & York 1997; York & Querat 2003; Arnaud *et al.* 2007; Murcia *et al.* 2007). The family Retroviridae is fairly diverse, but all infectious members contain at least four main genes in the order 5'-gag-pro-pol-env-3' (Bannert & Kurth 2006). The JSRV genome is characterized by simple genetic structure and replication-competent betaretroviruses composing of retroviral genes gag, pro, pol and env that present in the form of canonical structures (York *et al.* 1991; York *et al.* 1992; Palmarini *et al.* 1999; Palmarini *et al.* 2004). In addition to the encoded genes, the JSRV contains two long terminal repeats 5'LTRs and 3'LTRs having the viral enhancers and promoter which are potentially dynamic in differentiated lung cells and interact with transcription factors specific to lung (Palmarini *et al.* 2000a; McGee-Estrada *et al.* 2002). Endogenous retroviruses ERVs share to some extent a similar genetic structure as exogenous retroviruses including the two LTRs surrounding the four-basic internal coding retroviral (Goff 2007).

Endogenous retroviruses have been identified in all vertebrates (Benveniste & Todaro 1973, 1977; Boeke & Stoye 1997; Vargiu *et al.* 2016). Co-existence of both endogenous and exogenous retroviruses in some cases such as koala retrovirus (KoRV) or the mouse mammary tumor virus (MMTV) have been found in their hosts (Baillie *et al.* 2004; Tarlinton *et al.* 2008). Furthermore, *in vitro*, it has been concluded that JSRV Env alone could expressed and enable to induce cell transformation (Maeda *et al.* 2001; Palmarini *et al.* 2001; Chow *et al.* 2003; Danilkovitch-Miagkova *et al.* 2003). *Ovis aries* genome and goat *Capra hircus* comprise roughly 15-20 copies of endogenous betaretroviruses and

type D retroviruses which are highly related to the JSRV and accordingly referred as enJSRVs (York *et al.* 1992; Hecht *et al.* 1994; Hecht *et al.* 1996; Palmarini *et al.* 2000b).

Sheep betaretroviruses offer a unique model system to study the complex interaction between retroviruses and their host. Some proviruses such as enJSRV-10 and enJSRV-6 are shared across various genera, suggesting their integration before the split of the genus *Ovis* (sheep-like species) from the genus *Capra* (goat-like species) that is estimated to have happened between 5 and 7 MYA Figure 6.1 (Irwin *et al.* 1991; Randi *et al.* 1991; Fernández & Vrba 2005; Arnaud *et al.* 2007).

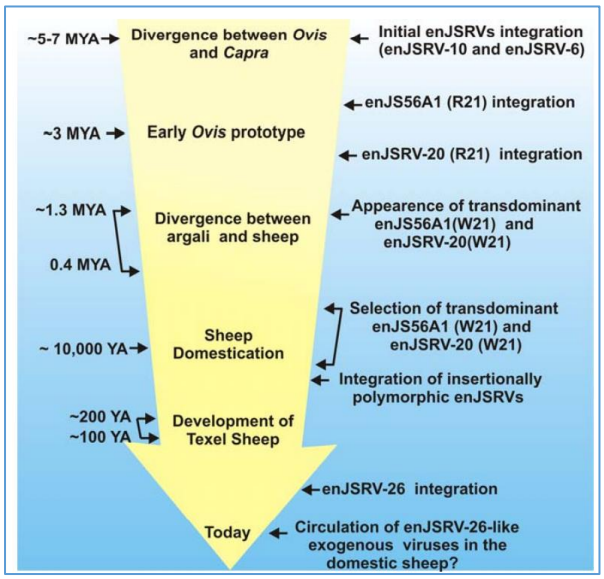


Figure 6.1 Integration of enJSRVs before and after sheep domestication

Furthermore, Arnaud *et al.* (2007) identified ten proviruses e.g. enJS56A1 was common between the domestic sheep and members of the genus *Ovis*, including bighorn sheep (*O. canadensis*), Dall sheep (*O. dalli*) and argali (*O. ammon*). The origin of the *Ovis* genus is estimated to have occurred approximately 3 MYA (Bunch *et al.* 2005; Fernández & Vrba 2005). The enJSRVs genome is 90 to 98% identical to the JSRV genome at their amino acid level which indicates a high relationship between genomes of both enJSRVs and JSRV (Bai *et al.* 1996; Palmarini *et al.* 1996; Bai *et al.* 1999; Palmarini *et al.* 2000a; Rosati *et al.* 2000). Three complete genome of enJSRV proviruses enJS56A1, enJS5F16 and enJS59A1 have been isolated and sequenced. The three genomes derived from genomic DNA phage library of sheep containing open reading frame encoding structural genes. Nucleotide differences in some genomic regions including long terminal repeat, gag region and envelope transmembrane found to be a good marker in differentiating

between the exogenous pathogenic form of JSRV and enJSRV (Palmarini *et al.* 2000a). enJSRVs can block the replication of JSRV through novel mechanism (Palmarini *et al.* 2004; Arnaud *et al.* 2007; Murcia *et al.* 2007).

Integration of viral genomes into their host cell genomes is considered one of the main features of the life cycle of retroviruses forming the provirus Figure 6.1. Such integration into the genetic information of a germ cell open up the way for retroviruses to settle in the germ line of their hosts. Where they can stay active for multiple generations in the form of stable integrated proviruses so called endogenous retroviruses (ERV). A huge number of retroviruses and retroviral-like elements have been discovered through analysis such genomic DNA. This demonstrating that the reverse transcription products have played an important role in eukaryotic genome structure. Thus, in general, genome of vertebrates including animals and humans have been colonized by retroviruses and their inheritance happen either vertically such in endogenous retroviruses following Mendelian model. While, others are horizontally transmitted so called exogenous retroviruses having no power in infection of germ line of their hosts (Boeke & Stoye 1997; Coffin *et al.* 1997; Patience *et al.* 1997; Löwer 1999; Fan *et al.* 2003).

In terms of endogenous retroviruses related repetitive elements, the eukaryotic genomes contain some considerable fragments of ERVs (Kumar & Bennetzen 1999; Lander *et al.* 2001; Hillier *et al.* 2004; Lindblad-Toh *et al.* 2005). ERVs are considered as repetitive transposable elements and they are difficult to recognize due to the rapid and high mutations occurring in their sequences (Sperber *et al.* 2007). Many algorithms have been established for searching and comparing sequences to databases (Altschul *et al.* 1990). However, they are limited in the discovery of several and different classes of repetitive DNA sequences, in particular genomic endogenous retroviruses at large-scale. In recent years, huge numbers of genomes from various organisms have been totally sequenced. Thus, identification of retroviral sequences at a broad range requires an efficient way of detection, classification and genomic distribution.

The convenience of next generation sequencing and bioinformatics tools have progressed the wide analyses of several mammalian genomes in terms of detection of transposable elements including endogenous retroviruses related sequences. Such

programs as RepeatMasker (RepeatMasker; Smit *et al.* (2013-2015) http://www.repeatmasker.org), Repbase (Jurka *et al.* 2005; Bao *et al.* 2015), BLAST-based searches (Tristem 2000; Villesen *et al.* 2004), LTR_STRUC (McCarthy & McDonald 2003) and Retrotector (Sperber *et al.* 2007; Sperber *et al.* 2009) have been widely utilized for identification of endogenous retroviruses related repetitive sequences. Thus, in this study, we analyzed whole sequencing raw reads of sheep genome using map to reference, graph-based read clustering and *k*-mer approaches (see section 2.2.13) to generate an overview of the population of endogenous retroviruses as repeats and as complete genomes in Kurdistani sheep breeds. This is the first time and type of results providing insight into their genomic structures, genomic proportions, and phylogenetic relationship to the known enJSRV over the world.

6.2 Aims and objectives

The current study aimed to

1. Assemble the complete genome of the endogenous betaretroviruses; enJSRV of sheep breeds from Iraqi Kurdistan region.
2. Relate enJSRV sequences from Kurdistani sheep to those found in worldwide sheep breeds.
3. Identify the major repetitive DNA families of endogenous retroviruses from unassembled genomic sequences using graph-based clustering and *k*-mer frequency, measuring their abundance and sequence diversity.
4. Investigate the genomic distribution and chromosomal organization of endogenous retroviruses related repetitive elements using *in situ* hybridization.
5. Understand the amplification (genomic proportions) and evolution (enJSRV and ERV families, sequences and chromosomal localization), to show their role in speciation, disease, and domestication.

6.3 Materials and Methods

6.3.1 Assembly of the complete genome of the endogenous betaretroviruses enJSRV

Whole sequencing paired raw reads obtained from sequencing five samples of genomic DNA representing two main sheep breeds (Karadi and Hamdani) from Kurdistan region (see section 2.2.7.2) were aligned against the complete genome of the endogenous betaretroviruses (enJSRV) of the Inner Mongolia (accession no. DQ838493; (Wang *et al.* 2008)) following section 2.2.13.5. The five-complete endogenous betaretroviruses enJSRV genomes (HamJ1, HamJ2, HamM, KarJ and KarM) were then assembled and annotated using Geneious 8.0 (Kearse *et al.* 2012) (http://www.geneious.com). Sequences are available in GenBank under accession numbers (MF175067, MF175068, MF175069, MF175070 and MF175071).

6.3.2 Data analysis and relationships

The five complete genomes of the endogenous betaretroviruses enJSRV of Hamdani and Karadi sheep breeds were aligned with published genomes from various sheep breeds originated from geographically different locations. Phylogentic tree was built for the entire endogenous betaretroviruses enJSRV genomes following section 2.2.8. All analyses were done using EF680305 *Ovis aries* strain genome as the outgroup.

6.3.3 Amplification of endogenous retroviruses repetitive related sequences

Consensus sequences of ERV related contigs resulting from RepeatExplorer and *k*-mer frequency were selected and used for PCR primer designing in order to amplify sequences representing different classes of endogenous retroviruses repeats Table 6.1. Amplified PCR products were purified, labelled and used as probes for *in situ* hybridization. Parameters and cycling conditions of PCR amplification were carried out following section 2.2.2. The endogenous retroviruses repetitive related elements will be submitted to the Repbase databases. Identification of endogenous retroviruses related repetitive elements was followed section 2.2.13.

Table 6.1 Primer sequences, PCR products, and probe names used for amplification and *in situ* hybridizations.

Probe names	Name of primers [Sequence (5'-3')]	Expected Product Size (bp)	Annealing Temp.
CL14C75_ERV2	F= GGTGATTTACATCATCTTCTGGCC	505	62
	R= AGCTTGCCTAACAGGTTCCC		
CL18C5_ERV1	F= ATCTTGGCTGAGCGATGCG	246	62
	R= GGGCTCTTGTCTAACACTCGG		
CL20C5_ERV1	F= TGTGTTGCCATGACCACTCC	574	62
	R= TGCCAGCATTCTTGGACTCC		
CL23C4_ERV1	F= CAAGGAATTTGGAGTGGTGGG	195	62
	R= TCGGTGGTCCTGTTGTAGCC		
CL27C1_ERV1+ERV3	F= GCAGGTCGGTGTATCTTCCC	619	62
	R= GGGAACTTGCAAGAGTGGGG		
32mer_ERV1-RE	F= GGTTTTAGATGGGACCGGGC	564	62
	R= TCTTCCTGCCATTCGAAGGC		
32mer_ERV1.T3	F= TGCTTCTTTTCAACGCACCC	541	64
	R= CTTGATGGAGCCAGGTACCC		
CL25_ERV1	F= TGTCATCTGGTCACTGCTGC	402	62
	R= AGGGAGTTTGCAGGATGTGG		
32mer_ERV1+ERV3	F= CTTGCAAGAGTGGGGAAAGC	615	62
	R= GCAGGTCGGTGTATCTTCCC		
CL37_ERV2	F= TGTCTTTTCCTCTCCTCGGC	488	62
	R= CATGCTTATGTCTGGGCTGC		
32mer_ERV1+CRC	F= TACAGAGCAAAGGGGATGGG	468	60
	R= TGGTTGTTTCTTTCCACCATTCC		
22mer_ERV1.A.RE	F= CACTCTTTTGCCCAATCCGG	545	60
	R= CAGCTACTTTTCGAGCTGCC		
OuttopCL_ERV2	F= AAAGGTCACGAGGATGAGGC	555	60
	R= AGGACAAAGGTGCAGTGGG		
CL67_ERV3	F= ATTCAATCTCCTAATATTCCCACCC	315	58
	R= GTTAGTAGTCAAGCTTTTGTCTGGC		

6.4 Results

6.4.1 The complete genome of enJSRV in Kurdistani sheep breeds and their phylogenetic relationships

The complete consensus endogenous betaretroviruses enJSRV genome of three Hamdani and two Karadi sheep breed were extracted from mapping whole sequencing paired raw reads of each breed to a reference (Inner Mongolia; DQ838493) using map to reference (see section 2.2.13.5) Figure 6.2. The total lengths of the consensus endogenous betaretroviruses enJSRV genomes of each target sheep breed were 7941.

A complement of four open reading frames was found, each one was corresponding to the gag, pro, pol and env gene Figure 6.2. The GC content averaged 41.6%.

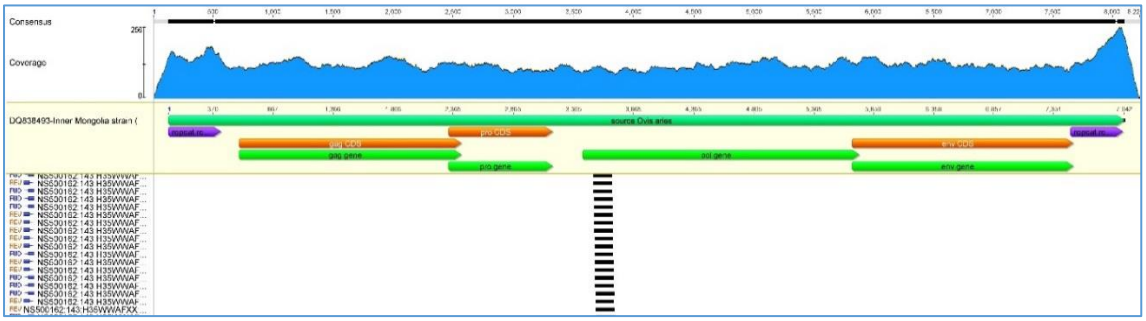


Figure 6.2 Assembly of raw reads (black lines) to reference complete genome of enJSRV DQ838493 showing the 5' and 3' LTRs (purple) and the four gene open reading frames, ORFs (brown for coding sequence and genes for green). Reads cover the whole sequence with an average depth of c. 120x (blue, see Table 6.2) and increased depth in the LTRs. Illumina sequencing gave good shotgun coverage (equal forward FWD and reverse REV) reads, and matched left and right paired-end reads to the sequence (shown by symbol after REV/FWD and before read code NS500162…).

The phylogenetic position of the five assembled Kurdistani endogenous betaretroviruses enJSRV genomes was established by Bayesian tree analysis (see section 2.2.8) including published reference genomes, and as an outgroup, Jaagsiekte sheep retrovirus strain (enJSRV-5; EF680305) was used. The consensus sequences of endogenous betaretroviruses (enJSRV) from the large fat-tailed sheep breeds from the Kurdistan region was placed with the recognized *Ovis aries* of AF153615, EF680302 and DQ838493 Jaagsiekte sheep retrovirus. The Kurdistani subclade was sister to the subclade including different strains of Jaagsiekte sheep retrovirus sampled from geographically different locations Figure 6.3.

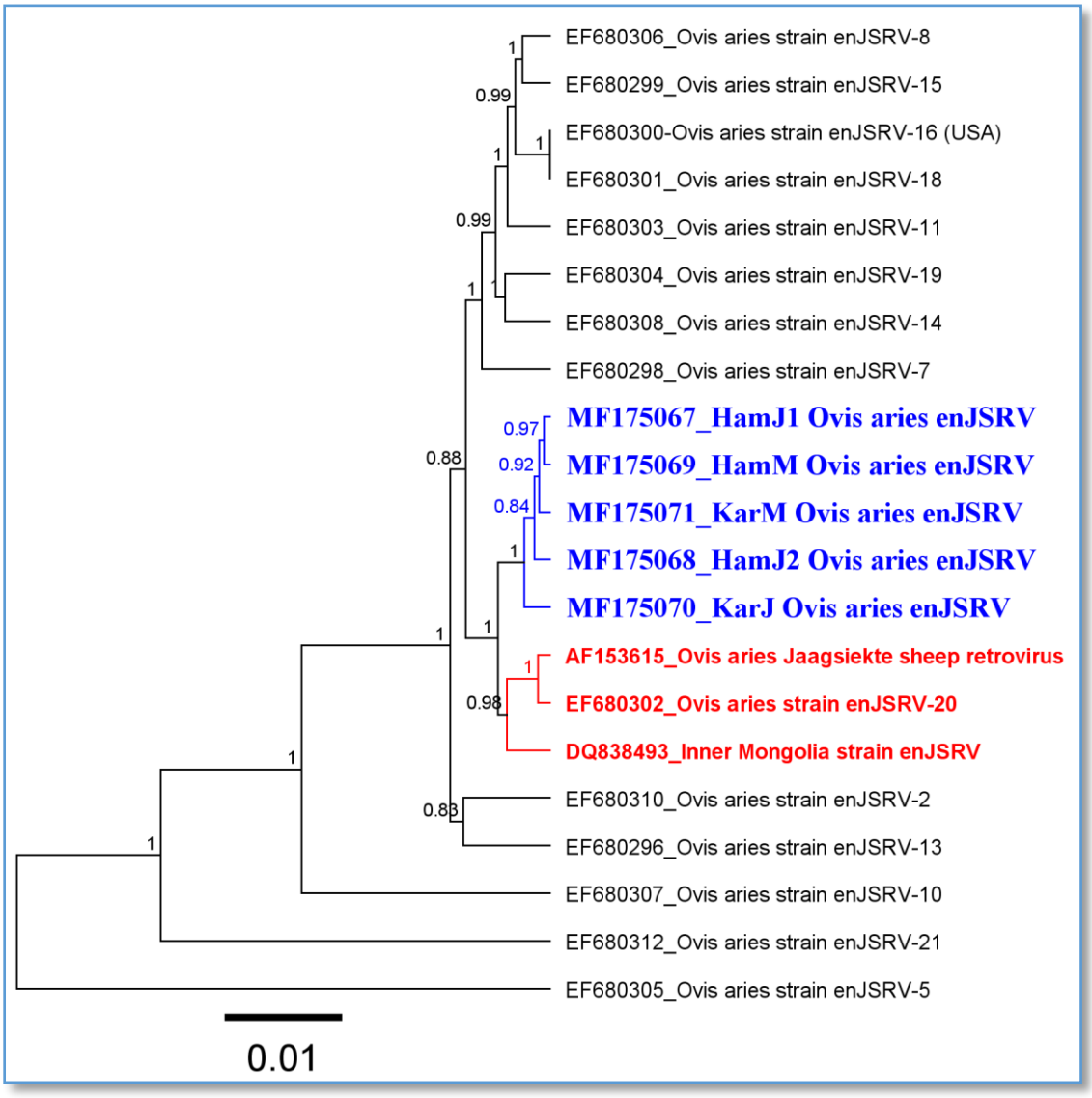


Figure 6.3 Phylogenetic relationship showing position of complete genomes of endogenous betaretroviruses (enJSRV) of Kurdistani breeds (Hamdani and Karadi) in relation to other sheep breeds originating from geographically different locations. The five-complete genome enJSRV are available in GenBank under accession numbers **MF175067, MF175068, MF175069, MF175070 & MF175071**. The tree was built for the entire endogenous betaretroviruses (enJSRV) genomes and derived from Bayesian (MrBayes) analysis. All analyses were done using the complete enJSRV genome of the more polymorphic (diverged) EF680305 *Ovis aries* strain as outgroup. Nodes are labelled with posterior probabilities. Notably, the five Kurdistani enJSRV genomes form a well-supported cluster among other enJSRV sequences.

6.4.2 **Coverage and genomic proportions of enJSRV**

Whole sequencing raw reads of five individual of sheep breeds were mapped to the consensus of endogenous betaretroviruses (enJSRV) following section 2.2.13.5 to estimate copy numbers and genomic proportions of enJSRV genome per each individual

breed sample. As a result, genomic proportions of enJSRV were 0.0087% to 0.0118% (71 to 124 copies) in the sheep genome Table 6.2.

Table 6.2 Copy numbers and genomic proportion of complete genomes of the endogenous betaretroviruses enJSRV integrated in the main sheep breeds of Iraqi Kurdistan region.

Breeds	Complete enJSRV genome bp	Assembled reads	Total reads of each genome (coverage X)	Genomic proportion %	Copies of enJSRV (assembled reads*150/7941)	Copies of enJSRV per one fold genome
enJSRV_HamJ1	7941	5390	52,048,068 (2.6)	0.0104	101.81	39.12
enJSRV_HamJ2	7941	6606	56,220,882 (2.81)	0.0118	124.78	44.39
enJSRV_HamM	7941	3809	43,596,654 (2.18)	0.0087	71.95	33.01
enJSRV_KarM	7941	4846	44,933,034 (2.25)	0.0108	91.54	40.74
enJSRV_KaJ	7941	5386	60,605,648 (3.03)	0.0089	101.74	33.57

6.4.3 Estimation of integration time of endogenous betaretroviruses enJSRV

The complete genome of endogenous betaretroviruses enJSRV contained two long terminal repeats (LTR); 5' and 3' LTR. Each LTR subdivided into three regions of sequences (U3 region, R region and U5 region). Each of 5' and 3' LTR sequence was composed of 446bp located at the first and last part of each enJSRV genome Figure 6.2 & Table 6.3. The number of variant or polymorphic sites between the sequences of 5' and 3' LTR were used by previous sequencing projects to estimate the integration time of endogenous betaretroviruses in sheep genome (Appendix 6.1;(Arnaud *et al.* 2007). Accordingly, the sequences of the 5' and 3' LTR of each enJSRV genome identified from each individual sheep breed sampled in this study were aligned against each other and the number of polymorphic sites were assessed. Four complete genomes HamJ1, HamM, KarJ and KarM of endogenous betaretroviruses enJSRV were characterized by presence of nucleotide differences between their 5' and 3' LTR and, thus, several polymorphic sites were recorded except genome HamJ2 where alignment of 5' and 3' LTR repeat region showed 100% sequence identities. Different polymorphic sites were found in each genome [HamJ1 (6 SNPs); HamJ2 (0 SNPs); HamM (5 SNPs); KarM (2 SNPs) and KarJ (5SNPs)] Tables 6.3 & 6.4. These polymorphic sites were including transitions and transversions Table 6.4. Based on the number of nucleotide differences (polymorphic sites) found here, the integration time of endogenous betaretroviruses

enJSRV into the genomes of Kurdistani sheep breeds was estimated and dated back to recent integration less than 0.45 up to 6.5 million years ago Table 6.3.

Table 6.3 Integration time of the endogenous betaretroviruses (enJSRV) estimated based on length and nucleotide differences at 5' and 3' LTR sequences (see Appendix 6.1).

Breeds	Gender	LTR length bp		Nucleotide differences	Estimated integration (MYA)
		5'UTR	3'UTR		
enJSRV_HamJ1	Male	446	446	6	3.0 - 6.5
enJSRV_HamJ2	Male	446	446	0	0 - 0.45
enJSRV_HamM	Male	446	446	5	2.2 - 4.9
enJSRV_KarM	Male	446	446	2	0.9 - 1.9
enJSRV_KaJ	Female	446	446	5	2.2 - 4.9

Table 6.4 Polymorphic sites found between the enJSRV 3'UTR and the 5'UTR sequences.

SNP positions	71	105	114	183	307	342
Breeds	3'UTR/5'UTR	3'UTR/5'UTR	3'UTR/5'UTR	3'UTR/5'UTR	3'UTR/5'UTR	3'UTR/5'UTR
Variants (SNPs)	C/T	G/A	A/G	T/C	T/C	G/T
enJSRV_HamJ1	C/T	G/A	A/G	T/C	T/C	G/T
enJSRV_HamJ2	None					
enJSRV_HamM	C/T	G/A	A/G	None	T/C	T/G
enJSRV_KarM	None	None	None	None	T/C	T/G
enJSRV_KarJ	C/T	G/A	A/G	T/C	C/T	None

6.4.4 Identification of endogenous retroviruses related repetitive elements using graph based read clustering

The outcome of RepeatExplorer was investigated to identify and classify LTR retrotransposons. Several clusters matching the ERV repeats, were distributed over the top clusters each with different genomic proportions Table 6.5. As results of comparing the cluster sequences to Repbase and RepeatMasker databases, all classes of endogenous retroviruses within the LTR retrotransposons, including ERV1, ERV2 & ERV3 were identified. Manual inspection of many other repeats and database comparisons (Chapter Five) did not reveal further clusters with similarity to ERVs, only to transposable elements. Inspection of the ERV domains matched by sequences in each cluster confirmed the reads had abundant ERV-related sequences except for CL23 which was retained in the analysis as an outlier Table 6.5. In comparison to non-LTR Retrotransposons (Chapter Five), sequences of the three ERV classes measured were mostly less abundant. The total genomic proportions of all classes of endogenous

retroviruses were estimated based on the abundance of ERVs in each cluster and some 0.50% of all reads represented ERV components. The genomic proportion of merged clusters was also analyzed by the TAREAN tool (Chapter Four) to classify related clusters based on the presence of ERV class. The ERV2 class was found to be more abundant than other classes of ERV repeats in the five analyzed genomes. Some clusters representing ERV repeats were apparently gender-specific ERV, identified in male but not in female sheep e.g. CL67_ERV3; Table 6.6.

6.4.5 Exploration of endogenous retroviruses DNA repeats using *k*-mer frequency

Repetitive classes of endogenous retroviruses were identified using *k*-mer frequency (Jellyfish) tool. Firstly, short motifs of DNA sequence with abundance of different motifs in lengths 22mers, 32mers, and 44mers were accounted from analysis of whole sequencing raw reads. Then, the short motifs that repeated more than 100 times were assembled using Geneious assembler. Several thousand contigs were produced from assembly of short motifs and the consensus of the top 100 contigs were compared with Repbase databases. Different classes of endogenous retroviruses related repeats such as ERV class1, ERV class2 and ERV class3 were found.

Table 6.5. Graph-based clusters with similarity to ERV from RepeatExplorer analysis and RepeatMasker comparisons. Construction of graph layouts were explained in section 1.7.1.

Clusters	Total Length	Number of reads	Genome proportion [%]	RepeatMasker database similarities	Graph layout
CL14C75_ERV2	**309579**	**2057**	**0.229**	LTR.ERVK (1906hits, 79.8%) LINE.RTE.BovB (12hits, 0.181%) Low_complexity (4hits, 0.0701%) SINE.tRNA.Core.RTE (4hits, 0.0656%) Simple_repeat (5hits, 0.053%)	
CL18C5_ERV1	**32673**	**217**	**0.024**	LTR.ERV1 (217hits, 88.5%) LINE.RTE.BovB (3hits, 0.435%) SINE.tRNA.Core.RTE (3hits, 0.404%) LINE.L1 (1hits, 0.162%)	
CL20C5_ERV1	**28460**	**189**	**0.021**	LTR.ERV1 (125hits, 59.5%) LINE.L1 (18hits, 6.38%) SINE.tRNA (12hits, 4.62%) SINE.tRNA.Core.RTE (1hits, 0.158%) Simple_repeat (1hits, 0.158%) LINE.RTE.BovB (1hits, 0.116%)	

CL23C4_ERV1	**21077**	**140**	**0.016**	LTR.ERV1 (3hits, 0.859%) SINE.Core.RTE (3hits, 0.688%) LINE.L1 (1hits, 0.289%) SINE.tRNA.Core.RTE (1hits, 0.152%) SINE.tRNA (1hits, 0.152%) Simple_repeat (1hits, 0.109%)	
CL27C1_ERV3+ERV1	**17746**	**118**	**0.013**	LTR.ERV1 (54hits, 27.3%) Satellite.centr (2hits, 1.08%)	
CL17	**150277**	**998**	**0.1190**	LTR.ERV1 (463hits, 40.2%) Satellite (129hits, 9.06%) Satellite.centr (47hits, 3.93%) SINE.tRNA.Glu (18hits, 1.34%) LINE.L1 (13hits, 0.949%) SINE.BovA (9hits, 0.282%)	
CL25_ERV1	**18501**	**123**	**0.0147**	LTR.ERV1 (35hits, 15.7%) Satellite.centr (19hits, 12.1%) SINE.BovA (3hits, 0.611%)	
CL37_ERV2	**9926**	**66**	**0.0079**	LTR.ERVK (66hits, 96.8%) LINE.RTE.BovB (2hits, 1.17%) SINE.BovA (1hits, 0.443%)	

6.4.6 Identification of ERV repeats following map to reference

Whole genome sequencing (52048046 paired raw reads) of HamJ1 were mapped to concatenated ERVs related sequences; 105 ancestral sequences (145kbp) and 90 sequences (173kbp) of *Bos taurus*. Accordingly, 10518 reads (2*150bp) were assembled to ancestral ERV sequences. These ancestral reads were compared to Repbase databases, and matched ERV3 sequence with high similarities 70-88%. While, in case of mapping NGS data to ERV sequences of *Bos taurus*, 238182 reads were assembled. Genome proportions of ancestral and *Bos taurus* related ERVs found in the whole sequencing raw reads of sheep genome were about 0.02% and 0.45% respectively. The ERV sequences reconstructed here will be submitted to the Repbase databases.

6.4.7 Identification of ERV classes from *de novo* assembly of raw reads

Whole sequencing paired raw reads were subjected to *de novo* assembly using Geneious assembler following section 2.2.13.4. As a result, several thousand contigs with different sizes were assembled (see section 4.4.3.3). Then, 1769433 contigs of *de novo* assembly

of HamJ2 were mapped to the ERV related sequences of ancestral and *Bos taurus* (see above section 6.4.6). 57 sequences out of 105 ERVs related ancestral sequences were assembled with sequence identities 70-85%. Majority of endogenous retroviruses found in ancestral sequences were characterized by ERV3 related repeats. However, most ERV sequences of *de novo* contigs that were homologous to ERVs related *Bos taurus* were belonged to the classes I and II ERVs.

6.4.8 **Bioinformatics abundances of probes used for FISH**

Copy numbers and genomic proportion of each probe representing different classes of endogenous retroviruses related DNA repetitive elements used in this chapter were estimated following section 2.2.13.5 Table 6.6. Whole genome sequencing used for estimation the probe copy numbers were 52048068 reads for HamJ1 and 60605648 reads for KarJ (see section 2.2.7.2).

Table 6.6 Copy numbers of various ERV related fragments used for *in situ* hybridization.

Probes of endogenous retroviruses ERVs	PCR product bp	HamJ1_Male genome			KarJ_Female genome		
		Assembled reads	Copies of probe	Genomic proportion%	Assembled reads	Copies of probe	Genomic proportion%
CL14C75_ERV2	505	5294	1572	0.0102	6943	2062	0.0115
CL15C6_ERV1)	426	3483	1226	0.0067	4024	1417	0.0066
CL18C5_ERV1	246	5973	3642	0.0115	7525	4588	0.0124
CL20C5_ERV1	574	3124	816	0.006	3676	961	0.0061
CL23C4_ERV1	195	2157	1659	0.0041	2000	1538	0.0033
CL27C1_ERV1+ERV3	619	5907	1431	0.0113	5595	1356	0.0092
32mer_ERV1-RE	564	3753	998	0.0072	3426	911	0.0057
44mer_ERV1-T1	511	5710	1676	0.011	3762	1104	0.0062
32mer_ERV1.T3	541	1709	474	0.0033	2554	708	0.0042
22mer_ERV1.T2	400	1720	645	0.0033	2348	881	0.0039
CL25_ERV1	402	6900	2575	0.0133	4364	1628	0.0072
CL67_ERV3	315	1000	476	0.0019	0	0	0
32mer_ERV3+ERV1	615	5500	1341	0.0106	5300	1293	0.0087
CL37_ERV2	488	2113	649	0.0041	2338	719	0.0039
32mer_ERV1+CRC	468	5700	1827	0.011	5760	1846	0.0095
22mer_ERV1.A.RE	545	3452	950	0.0066	3008	828	0.005
CL39_ERV1	270	3756	2087	0.0072	3053	1696	0.005
OuttopCL_ERV2	555	1299	351	0.0025	1177	318	0.0019

6.4.9 Genomic organization and abundance of endogenous retroviruses related repetitive elements on chromosomes

6.4.9.1 Probes for chromosomal *in situ* hybridization

Amplified PCR products representing various endogenous retroviruses DNA repetitive related elements were labelled and hybridized to the male sheep metaphase chromosomes following sections 2.2.5 and 2.2.11. This enabled analysis of the distribution, organization and abundance of endogenous retrovirus. Raw reads from sheep genome were mapped against the consensus of each probe (ERVs related consensuses) in order to estimate their copy numbers and genomic proportions see section 6.4.8. RepeatExplorer graph layouts, numbers of reads, total lengths and genomic proportions of each ERVs related clusters are given in Table 6.5.

6.4.9.2 Endogenous retroviruses class1_ ERV1

Several probes of ERV1 related repeats were investigated for their chromosomal distribution in sheep and cattle chromosomes.

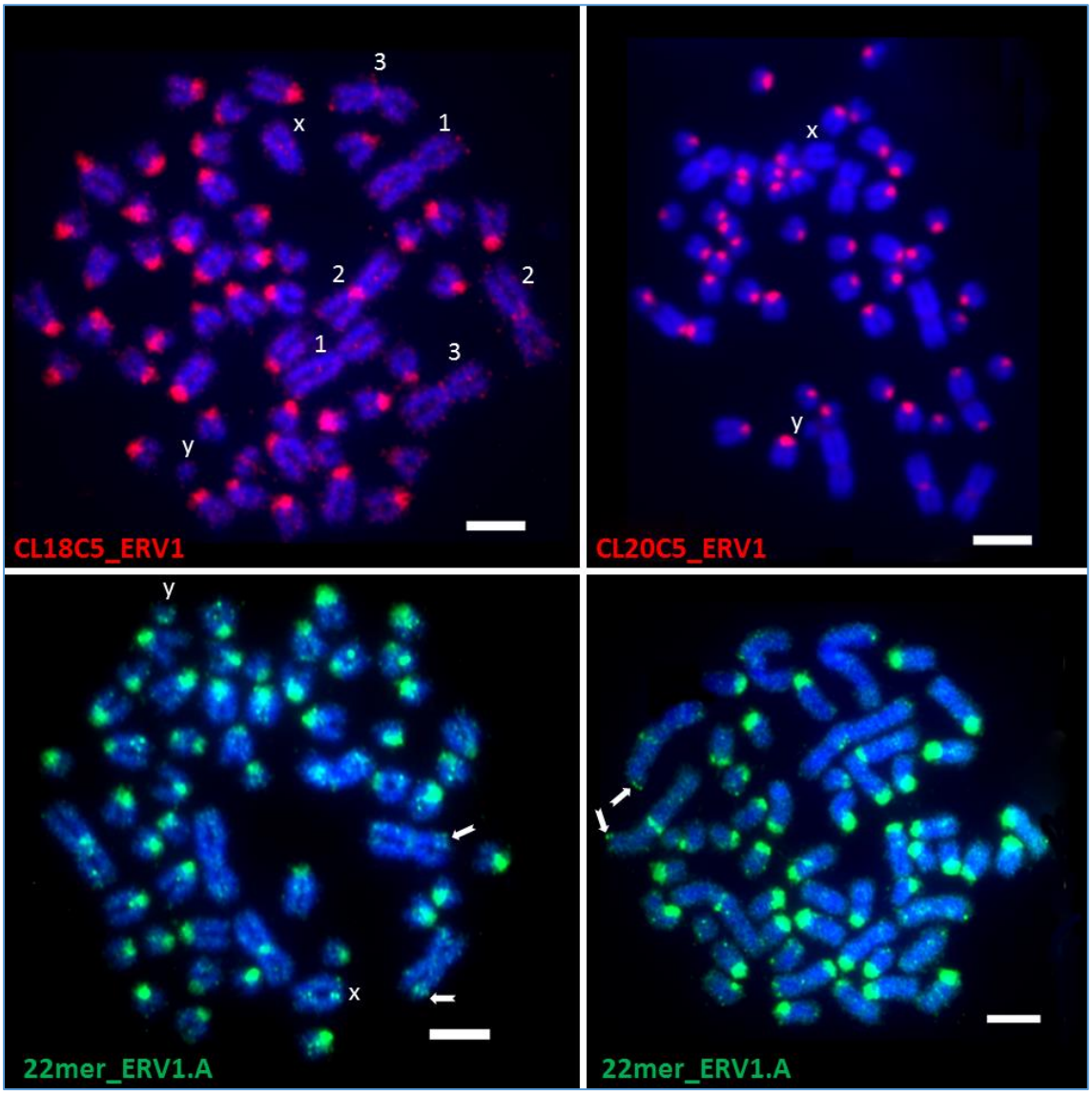


Figure 6.4 Probe CL18C5_ERV1 produced strong signals on the centromeres of acrocentrics. Pair of submetacentric has strong centromeric signals, while, the other two pairs have weaker signals. Both X and Y chromosomes have weak but noticeable signals. Sequences of probe CL20C5_ERV1 observed specific signals to the centromeres of acrocentrics, weaker signals present on submetacentrics. No signals were seen on pairs of largest submetacentrics and sex chromosomes. Probe 22mer_ERV1.A from contigs of 22mers GT100 was labelled to the centromeric regions of all acrocentrics. It showed small dots over the sex chromosomes X and Y. Weak signals and very small dots were seen at centromeric and subtelomeric regions of submetacentrics.

In comparison to *in situ* results of the above ERV1 probes (Figure 6.4), probes CL23C4_ERV1, CL25_ERV1, 32mer_ERV1 and 32mer_ERV1.T3 showed different genomic distributions Figure 6.5.

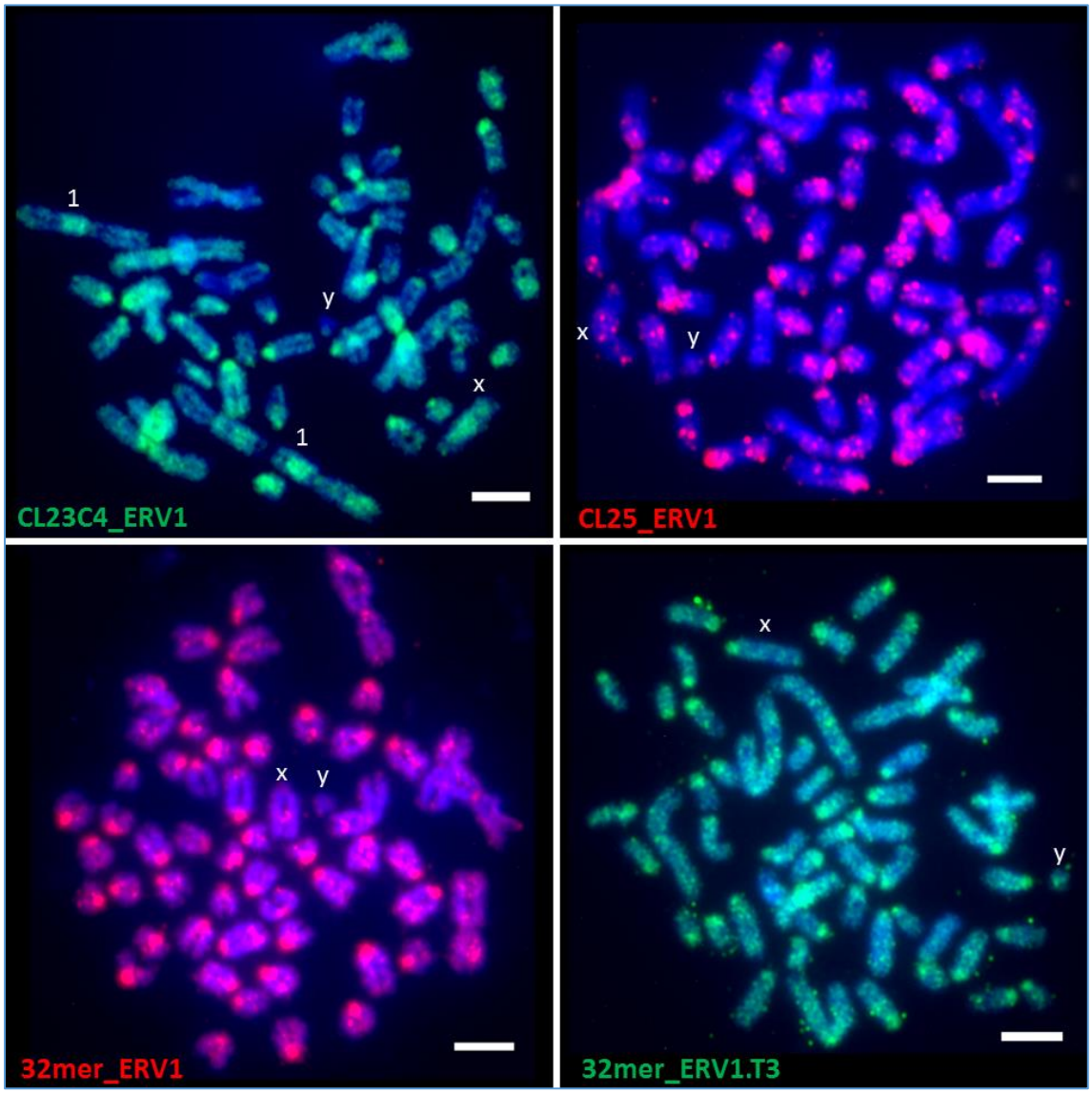


Figure 6.5 Signals OF probe CL23C4_ERV1 presented on a few centromeres but about half of the acrocentrics; then a lot of dispersed signals on some but not all chromosomes or arms in the submetacentrics. There were slight signals on Y chromosome but stronger on X chromosome. Probe CL25_ERV1 showed variable signals more likely dotted and slightly dispersed over all chromosomes including the sex chromosomes. Signals were close to the centromeric domains of few acrocentrics. While, probe 32mer_ERV1 from 32mers GT100 was rather dispersed with concentration at centromeres. There are some gaps on submetacentrics. Notable signals were also seen in X and Y chromosomes. The

in situ results of probe 32mer_ERV1.T3 from 32mers GT100 showed centromeric and dispersed dots over all chromosomes, mostly rather uniform. Signal was dispersed on Y chromosome, while more centromeric to X chromosome.

6.4.9.3 **Endogenous retroviruses class2_ ERV2**

In contrast to ERV1, chromosomal abundance of probes OuttopCL_ERV2, CL14C75_ERV2 and CL37_ERV2 containing ERV2 related sequences showed different patterns Figure 6.6.

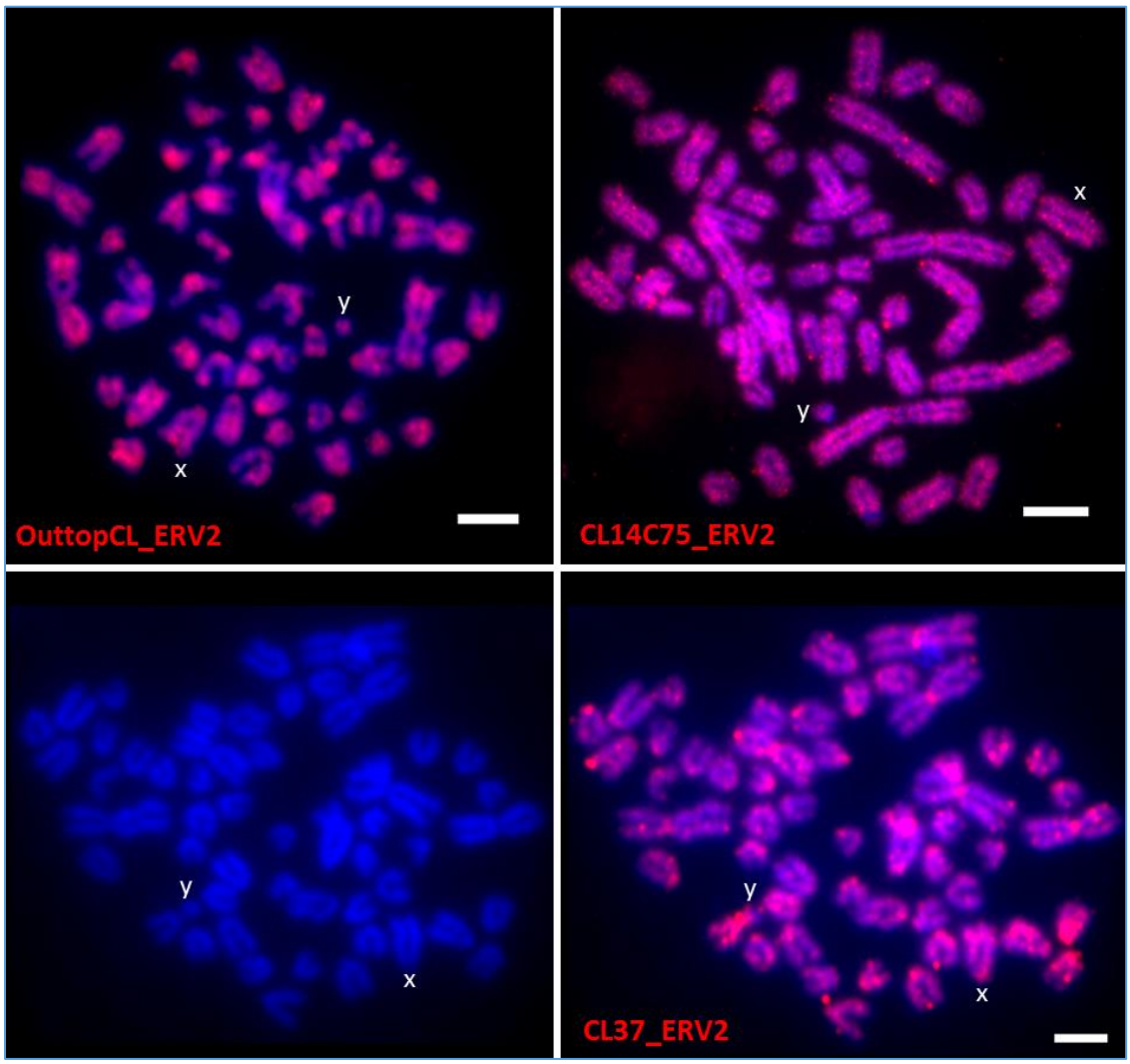


Figure 6.6 Probe OuttopCL_ERV2 showed signals at both centromeric and telomeric domains of acrocentric chromosomes. Signals were more like banding patterns on submetacentrics. There were strong signals on X and Y chromosomes. Signals of probe CL14C75_ERV2 were broadly dispersed on all chromosomes while some DAPI gaps can

be seen on some chromosomes. Probe was also hybridized to X and Y chromosomes. Probe CL37_ERV2 showed signals distributed over centromeric regions of some acrocentric and submetacentrics, while signals were apparently undetectable on other centromeres. Signals were dispersed on all chromosomes including the sex chromosomes.

6.4.9.4 **Endogenous retroviruses class3_ ERV3**

Cluster 67 was matched the ERV3 related sequences. This cluster was found in the RepeatExplorer outcomes of male genome, but was absent in the female. The whole sequencing raw reads of Female_KarJ were mapped to the consensus of CL67, and no raw reads were assembled. Different metaphases showed centromeric signals with some intercalary dots on one metaphase Figure 6.7.

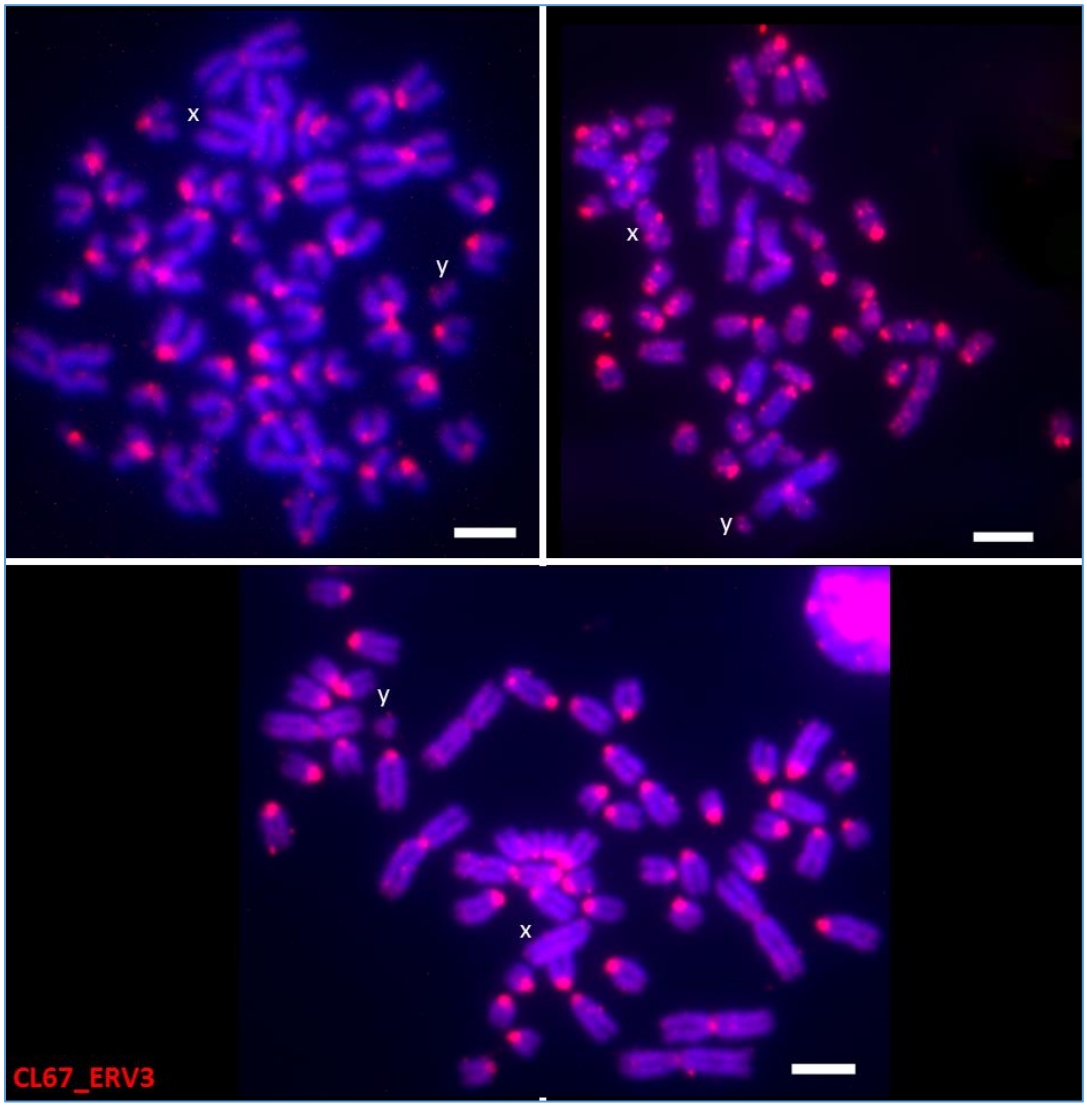


Figure 6.7 Probe CL67_ERV3 showed much more abundant signals at centromeres of all acrocentrics, and one submetacentric pair, while weaker signals were seen on the other two pairs of submetacentrics. A few weak bands or double dots or slight signals were seen throughout chromosome. Signals were also present on sex chromosomes.

6.4.9.5 **Endogenous retroviruses ERV1+ERV3**

In addition to three classes of ERVs (see above), contig sequences of RepeatExplorer and *k*-mer frequency were also including combined sequences of two classes ERV1 and ERV3.

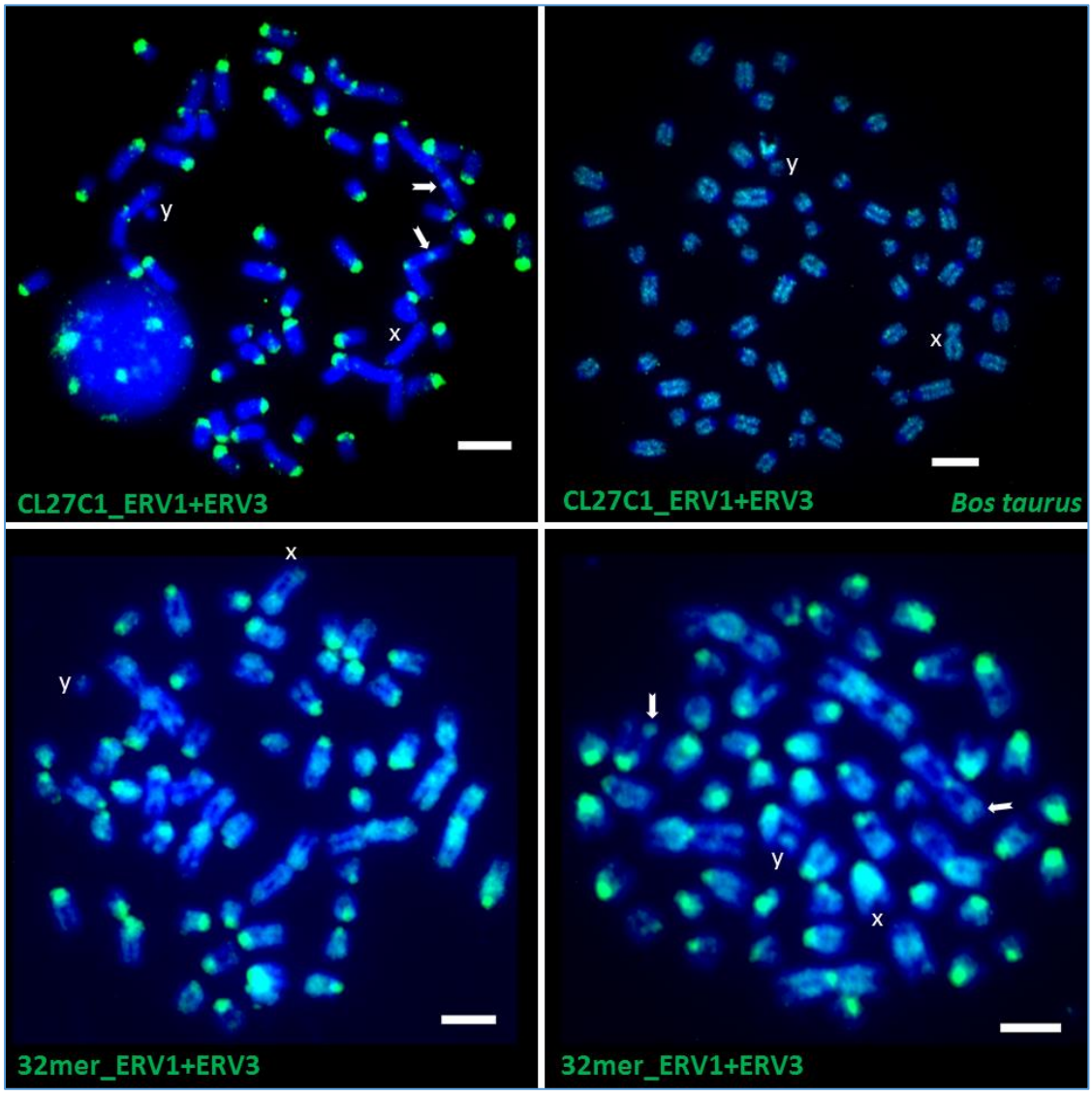


Figure 6.8 Probe CL27C1_ERV1+ERV3 was strongly hybridized to the centromeres of all acrocentrics. While weak signals can be seen at centromeric and subtelomeric regions of some of submetacentrics. Signals were not seen on sex chromosomes. Interestingly, the same probe CL27C1_ERV1+ERV3 was hybridized to the cattle chromosomes and

signals were dispersed on all chromosomes except centromeric domains. Furthermore, from *k*-mer assembly of 32mers GT100, probe 32mer_ ERV1+ERV3 showed centromeric signals with some bands or broader sites along arms of chromosomes. Telomeric signals on some acrocentrics and submetacentrics were present. Signals were also incorporated in X and Y chromosomes.

6.4.9.6 **Endogenous retroviruses and satellite like sequences ERV1+CRC**

Combined sequence of ERV1 and satellite like repeats 32merC16_Sat_CRC (see section 4.4.5.1) was found in the contig 61 consensus of 32mers GT100. Accordingly, probe was named 32mer_ERV1+CRC Figure 6.9. Furthermore, combined ERV1+CRC sequences and three copies of 32merC16_Sat_CRC satellite like sequences were found in the same consensus of CL15C14 (4191bp) of RepeatExplorer (Appendix 6.4). Probes of ERVs and 32merC16_Sat_CRC were used for *in situ* hybridization separately, and centromeric signals were observed which confirms their chromosomal organization next to each other.

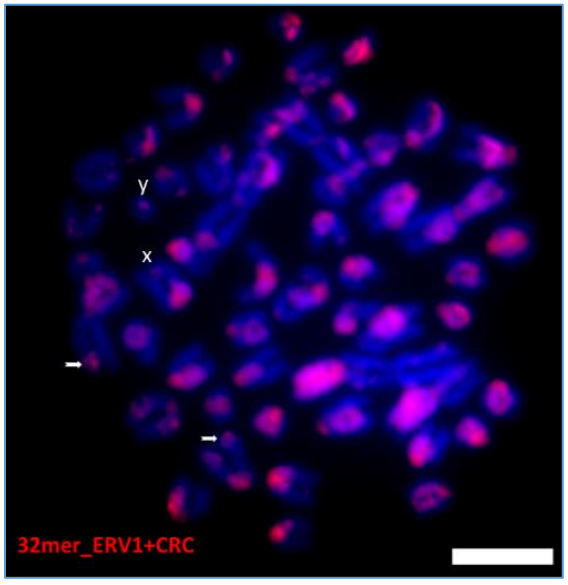


Figure 6.9 Probe 32mer_ERV1+CRC showed different signals, some were present on centromeric locations, while one arm of submetacentrics have broader signals. There were some dots on sex chromosomes. Single was also found at the telomeric regions of acrocentric and submetacentrics.

6.5 Discussion

6.5.1 The endogenous betaretroviruses (enJSRV) genomes in Kurdistani sheep

Some 0.01% of shotgun sequence reads from the two sheep breeds were homologous to the endogenous betaretroviruses genome sequences (enJSRV) Table 6.2. The reads gave uniform coverage over the whole genome Figure 6.2, allowing reconstruction of enJSRV from five individuals from Hamdani and Karadi sheep from the Kurdistan region of Iraq, (accession numbers in Figure 6.3). Analysis of coverage suggested that 33 to 45 enJSRV copy numbers were present in the sheep genome Table 6.2. Chessa *et al.* (2009) reported at least 27 strains of endogenous retroviruses (ERVs) related to the enJSRV retrovirus, noting that they are informative genetic markers.

6.5.2 Phylogenetic relationships and polymorphisms of enJSRV sequences

Endogenous retroviruses can be categorized through comparison of sequences and generation of a phylogeny (Jern & Coffin 2008). The phylogenetic position of the five assembled Kurdistani endogenous retroviruses enJSRV genomes was established by Bayesian tree analysis including as reference genomes published complete genomes of enJSRV in sheep see section 2.2.8. The complete enJSRV genomes of breeds from Kurdistan were placed on branches with three complete genomes of the enJSRV proviruses in sheep (enJS56A1; enJSRV-20 and enJSRV-NM). Although the five complete genomes of enJSRV grouped together in their phylogenetic relationship, several polymorphic sites (nucleotide differences) were found between the five enJSRV genomes of Hamdani and Karadi breeds (Appendix 6.5 and 6.6). Wang *et al.* (2008) cloned the complete betaretroviruses (enJSRV-NM) from sheep sampled from Inner Mongolia. Our study indicates that enJSRV sequences from the Kurdistan Region (with very close relationship of Hamdani and Karadi breeds) are most like those in sheep breeds from China and South Africa.

Arnaud *et al.* (2007) isolated and characterized 27 proviruses (enJSRV) integrated in the genomes of different species of genus *Ovis* within Caprinae subfamily including (*Ovis*

aries, Ovis ammon, Ovis canadensis and *Ovis dalli*) (Figure 6.1). They found that enJSRV proviruses (enJS56A1 and enJSRV-20) entered the host genome within the last 3 million years (before and during speciation within the genus *Ovis*), characterized by transdominant phenotype able to block late replication steps of related exogenous retroviruses. These results fit well the times calculated for the Kurdistani breeds (Tables 6.3 & 6.4): we found high sequence identities of the enJSRV genomes of Kurdistan sheep to these enJSRV proviruses (enJS56A1 and enJSRV-20) which refers to the fixation of enJSRV proviruses in the host genome before or around the time of sheep domestication. It has been suggested that endogenization and selection of ERVs performed as restriction factors used by the host to fight retroviral infections. Proviruses like HamJ2_enJSRV could escape from transdominant enJSRV. Viruses escaping the transdominant enJSRV loci have recently emerged (Arnaud *et al.* 2007). Therefore, endogenization of these retroviruses may still be occurring today (the HamJ2 sequence here; Tables 6.3 & 6.4).

Tracking the evolutionary history of proviruses in sheep genome will uncover events highlighting the host-virus relationship or ‘‘struggle’’ over several million years. In order to understand the nature of the association of endogenous retroviruses with their host, their evolutionary history has been investigated and their age can be estimated by assessing the sequence divergence between the proximal and distal LTRs within the same provirus, as it can be assumed that they were identical at the moment of proviral integration. The divergence accumulated between the LTRs over time can be used as a molecular clock (~ 2.3–5 X 10-9 substitutions per site per year) Table 6.3 and Appendix 6.1 (Johnson & Coffin 1999; Arnaud *et al.* 2007). Accordingly, in this study, the sequences of the 5’ and 3’ LTR of each enJSRV genome assembled from whole sequencing raw read of each individual sheep breed were aligned against each other and the number of polymorphic sites between the sequences of the 5’ and 3’ LTR were assessed Table 6.4. The nucleotide differences between 5’ and 3’ LTR of enJSRV_HamJ2 genome were 100% identical, which displays as an indicator of recent integration of endogenous retroviruses in the genome f HamJ2. However, the other four complete enJSRV genomes (HamJ1, HamM, KarJ and KarM) were characterized by having nucleotide differences between their proximal and distal LTRs, as several polymorphic sites were recorded (Tables 6.3 &

6.4). Based on the number of polymorphic sites found here, the integration time of endogenous betaretroviruses (enJSRV) into the genomes of Kurdistani sheep breeds was estimated and dated back from recent (with no polymorphisms, 0 to 0.45 MYA), to old integration up to 6.5 million years ago Table 6.3. Our results were consistent with the findings of Arnaud *et al.* (2007) who found sheep proviruses either with identical or with several polymorphic sites (1-8 SNPs), resulting in an estimated time of integration of these ERVs from less than 450,000 year ago to around 8 million year ago (MYA). This period spans most of the evolutionary history of the subfamily Caprinae (Fernández & Vrba 2005) (Figure 6.1 & Appendix 6.1). Sistiaga-Poveda and Jugo (2014) observed that enJSRV diversity was quite different among the species of *Ovis aries*, *Ovis* musimon and Rupicapra pyrenaica, among individuals within the species.

6.5.3 Major families of endogenous retroviruses in Kurdistani sheep

Endogenous retroviruses (ERVs) have been characterized into three classes based on their phylogenetic relationship to the established exogenous retroviruses genera; class I ERVs are related to the Gammaretrovirus and Epsilonretrovirus genera; class II ERVs are related to the Alpharetrovirus, Betaretrovirus, Deltaretrovirus, and Lentivirus genera while class III ERVs are related to the genus Spumavirus (Gifford *et al.* 2005). Various endogenous retroviruses from different genera have been characterized from a variety of mammalian species (Garcia-Etxebarria & Jugo 2010).

In this study, from mapping whole sequencing raw reads of Kurdistani sheep breeds to the references of genera Deltaretrovirus (*Bos taurus*), Lentivirus (*Ovis aries*) of class II and Spumavirus (*Bos taurus*) of class III retroviruses, no reads were assembled indicating no integration (or indeed exogenous sequences co-purified with genomic DNA) of such retroviruses in Kurdistani sheep breeds.

Following map to reference, in this study, we found that 0.02% of whole sequencing raw reads of sheep genome have similar sequences of ancestral ERV repeats and majority of them were matched the ERV3 class. However, about 0.45% of NGS reads were matched ERVs related *Bos taurus*. This percentage is covering mostly the total genomic

proportions of all classes of endogenous retroviruses that estimated by RepeatExplorer (see aslo section 5.4.3).

The current study is the first work investigating copy numbers and genomic proportions of endogenous betaretroviruses (enJSRV) using whole sequencing raw reads of five individual of sheep breeds. Behind map to reference, we found that 0.0087% (72 copies) to 0.0118% (125 copies) of NGS data were comprised genomic sequences of enJSRV. In other way, copy numbers of complete endogenous betaretroviruses (enJSRV) were 30-45 copies per sheep genome Table 6.2. Klymiuk *et al.* (2003) analyzed the retroviral *pro-pol* sequences of two retroviral families (B/D-type) and (C-type) and estimated their copy numbers in several sheep breeds using Southern blot analysis. They have found 5 up to 100 copies including 25 copies of B type ERVs. Our results of copy numbers are consistent with these results as we found approximately 30 to 45 copies of complete enJRSV per one fold genome.

6.5.4 Genomic distribution and chromosomal organization of ERV repeats

Based upon the findings of *in situ* hybridization, represented sequences from each class of ERVs showed quite differences and diversity in terms of their distribution patterns and their abundance over all sheep chromosomes particularly on the sex chromosomes.

Zahn *et al.* (2015) identified a new member of human endogenous retrovirus type-K distributed at multiple loci of the pericentromeric locations of several human chromosomes. Such abundant signals of ERVs at the centromere of sheep chromosomes demonstrated here that endogenous retroviruses were amplified during sheep evolution (see section 6.4.9). The second interpretation would be copying of retroviral sequences have occurred as a results of recombination events of the centromere of various chromosomes during evolutionary events of sheep genome. In kangaroo genomes, Ferreri *et al.* (2011) indicated that amplification of their endogenous retrovirus happened in a lineage-specific fashion which is limited to the centromeres of chromosomes. Although the centromeric and pericentromeric domains are gene poor (Lomiento *et al.* 2008), it is not clear whether such abundant ERVs in centromeric regions

of sheep chromosomes are involved in karyotypic rearrangement and the evolutionary fusion of six ancestral acrocentric chromosomes to the three submetacentric chromosomes in sheep. Ferreri *et al.* (2011) suggested that Kangaroo endogenous retroviruses (KERV) may be associated or involved with rearrangements targeted to centromere regions that characterize this group of mammals.

In this study, we found some other probe representing different classes of ERVs dispersed over all chromosomes including sex chromosomes of sheep (see Figure 6.6). Transposable elements including endogenous retroviruses have been proposed to facilitate regulatory network evolution as they comprise regulatory elements and can amplify in number and/or move throughout the genome (Wang *et al.* 2007; Feschotte 2008; Chuong *et al.* 2016). However, we suggest that the identification of one dispersed sequence as an ERV in sheep is not supportable, since the similarity is low.

Sistiaga-Poveda and Jugo (2014) characterized copies (types) of enJSRVs and their integration sites in domestic and wild species of the sheep lineage. enJSRVs copies were detected by amplifying the env-LTR region by PCR, and 103 enJSRV sequences were produced across 10 individuals and enJSRV integrations were found on 11 of the 28 sheep chromosomes. Our results reported here are applied to their findings as they proposed that the integration sites of some enJSRVs are not only different at classification hierarchy but also the geographical regions of species.

In this study, we compared the complete genome of enJSRV betaretroviruses with the nuclear chromosome assemblies of *O. aries Oar_v4.0* databases. Some 294 nucleotide sequence fragments were found with different lengths (37bp-7899bp; total length 136kbp) with high similarity (70-100%) indicating nuclear related sequences. Garcia-Etxebarria and Jugo (2010) analyzed the cow genome *Bos taurus*, and 928, 4487, 9698 ERVs related sequences were detected using three different methods, BLAST-based searches, LTR_STRUC and Retrotector, respectively. Furthermore, they found positive correlation between numbers of ERVs and size of chromosomes while negative correlation was noticed with chromosomal GC content. ERVs were not homogeneously distributed across chromosomes.

6.5.5 Evolution of ERV families and their role in speciation and domestication

The results here show that the Kurdistani sheep from the Fertile Crescent (near the centre of diversity and domestication) have similar abundances of enJSRV betaretroviruses and more generally endogenous retroviruses, ERVs (Table 6.2), to other sheep. While the ERV sequences in most mammals accumulate at the centromeres, the *in situ* hybridization results show some differences between probes, maybe allowing insight into the structural organization of the sequences in the centromere in the future as with the centromeric satellite analysis (Chapter Four & (Chaves *et al.* 2005)). At least, the evolutionary history of X and Y, and submetacentric chromosomes, is different to most autosomal acrocentrics: submetacentrics show loss of much of centromere-located sequences.

The ERV sequences identified here fall within the range of diversity previously reported, but form a distinct group (Figure 6.3), suggesting that there is either homogenization of the sequences (including potentially through gene conversion), or loss and replacement with new copies; however, unlike many retroelements, it is notable that the copy number is the same across all sheep breeds studied. Our genomic proportion results in sheep (some 0.50% of the genome), using raw reads without assembly artefacts, agree closely with others using Southern or dot-blot hybridization, quantitative polymerase chain reaction (qPCR), or assemblies to compute copy number. Our *in situ* results suggest greater abundance given the hybridization strengths on nearly all centromeres, perhaps because the other methods require longer stretches of homology and/or look for homology with higher stringency. In human, approximately 8% of the genome has been reported to consist of retroviral origin sequences, considered a result of continuous infections of the germ line of the host lineage by ancient viruses over millions of years of evolution (Lander *et al.* 2001; Paces *et al.* 2002). (Mouse Genome Sequencing 2002) reports the dissimilar evolutionary history and activity of ERVs between mouse and human.

For several decades, the biological significance of ERVs has been argued and sometimes considered to be “junk DNA” (Bock & Stoye 2000). However, ERVs are now considered

to have a variety of beneficial roles in their host genome contributed to genome plasticity (Jern & Coffin 2008; Varela *et al.* 2009; Kurth & Bannert 2010). In plants, the endogenous pararetrovirus sequences incorporated in the genome, first discovered in banana by *in situ* hybridization (Harper *et al.* 1999), are now thought to protect the host via RNAi mechanisms (Noreen *et al.* 2007). It is possible that the inactivation history of the viral sequences is different between mammals as suggested by the contrasting sequence variants and abundance between human, mouse and sheep but more work using identical analytical tools is needed.

Chapter 7 Genotyping and polymorphism of the ovine prion protein (PrP) gene in the Kurdistani sheep breeds

7.1 Introduction

Scrapie, the well-known animal fatal neurodegenerative disease, belongs to a cluster of disorders known as transmissible spongiform encephalopathy (TSE) or prion disease. It is the oldest known TSE disease, first described in sheep breeds in the United Kingdom in 1732 and then in Germany in 1750. Transmissible and genetic neurodegenerative diseases including scrapie in sheep and goats, bovine spongiform encephalopathy BSE in cattle, and Creutzfeldt-Jakob disease CJD in human are most caused by prions (Prusiner 1991). Prions are proteins which perform a vital role in these types of diseases in various species of mammals (Heaton *et al.* 2003). Although the primary cause of these disorders are poorly understood, prions are believed to be the predominant causative agent and it is thought that the disease caused by atypical virus due to their long incubation periods (Brown 2005; Schneider *et al.* 2008).

In sheep, the prion gene (PrP) is located on chromosome 13 which consists of three exons separated by two introns. The open reading frame (ORF) of the PrP gene is positioned on exon three which encodes 256 amino acids long protein (Goldmann *et al.* 1990; Lee *et al.* 1998; Tranulis 2002). It has generally been accepted that susceptibility to scrapie disease are most likely linked to the polymorphisms at the three well studied codons of the prion gene (Hunter 1997). The codons 136 (Alanine/Valine A/V), codon 154 (Histidine/Arginine H/R) and the codon 171 (Arginine/Glutamine R/Q) are the main candidates that linked to the resistance and susceptibility of scrapie disease. Alleles (A/H/R) are associated with resistance, while the alternative alleles (V/R/Q) are associated to susceptibility (Laplanche *et al.* 1993; Hunter *et al.* 1994; Westaway *et al.* 1994; Ikeda *et al.* 1995; Hunter 1996; O'Rourke *et al.* 1997; Hunter 2007). In sheep, five haplotypes ARR, AHQ, ARH, ARQ and VRQ are common and result from the presence of different combinations at the three codons of prion gene (Goldmann *et al.* 1994; Hunter 1996). The genotype ARR/ARR are highly resistant to susceptibility of classical scrapie, while VRQ/VRQ are the most susceptible genotype. The remaining genotypes are

associated with the intermediate susceptibility to disease. The five common alleles have been ranked in a descending order of resistance to susceptibility (Detwiler & Baylis 2003). Accordingly, a total of 15 allelic variants categorized in to the main five risk groups R1 to R5 where R1 genotype are related to low level of risks, while R5 genotypes are linked to high susceptibility to the disease (Dawson *et al.* 1998) Table 7.1.

Table 7.1 Scrapie resistance and susceptibility based on the classification of PrP genotypes.

Classification	R1	R2	R3	R4	R5
			ARQ/ARQ		
		ARR/ARQ	AHQ/ ARQ		VRQ/ARH
Genotypes	ARR/ARR	ARR/ARH	AHQ/AHQ	ARR/VRQ	VRQ/AHQ
		ARR/AHQ	AHQ/ARH		VRQ/ARQ
			ARH/ARH		VRQ/VRQ
			ARH /ARQ		

However, in the USA, the resistant and susceptible sheep to the scrapie disease were studied based on the investigation of polymorphisms at just codons 136 and 171, while the codon 154 is in minor role. Polymorphism in the PrP gene has been investigated and reported in many sheep breeds sourced from geographically different locations over the world.

The Kurdistan Region in the north of Iraq corresponding to the zone of initial domestication of sheep (see section 3.1 & Appendix 3.1) includes several fat-tailed sheep breeds such as Hamdani, Karadi, and Awassi. However, these local sheep breeds have not previously been investigated in terms of polymorphisms of the PrP gene.

7.2 Aims and objectives

The current study aimed to

1- Identify polymorphisms and genotypes of the ovine prion protein (PrP) gene in the local sheep breeds of Iraqi Kurdistan region using PCR sequencing.
2- Characterize the genetic aspects of scrapie disease in Kurdistani sheep breeds and relate that to the susceptibility and resistant to that disease.

7.3 Materials and methods

7.3.1 PrP gene amplification and sequencing

Two primer pairs (F.Scrp= TCCTGGTTCTCTTTGTGGCC and R.Scrp= GGTGAAGTTCTCCCCCTTGG) were designed from reference (*Ovis aries* prion protein (PrP) gene, complete cds; GenBank: M31313.1) using Primer3 of Geneious software. Primer pairs were used to amplify regions of PrP gene spanning allelic variants (covering three codons). Genomic DNA samples (Appendix 2.1) from local sheep breeds were used for PCR amplification in 50µl following section 2.2.2. After confirmation and separation of the PCR products (578bp) by gel electrophoresis (see section 2.2.3), PCR products were purified and allelic variants of PrP gene were sequenced by Sanger sequencing (see section 2.2.7.1).

7.4 Results

Four allelic variants (ARR/ARQ/ARH and ARK) were observed at the main candidate codons 136, 154 and 171 of the PrP gene in sheep breeds Hamdani, Karadi and Awassi that relate to susceptibility and resistance of scrapie disease. No variations were found at either codon 136 and 154. As a result, four different genotypes were found ARR/ARQ, ARQ/ARQ, ARH/ARH and ARK/ARK. Combination of all alleles and genotypes of PrP gene found in each DNA sample of local sheep breeds are shown in Table 7.2. Genotypes ARR/ARQ, ARQ/ARQ and ARH/ARH were identified in both of Karadi and Awassi breeds. Genotypes ARR/ARQ, ARQ/ARQ and ARK/ARK were found in Hamdani breed.

Table 7.2 Combination of all alleles and genotypes of PrP gene found in each DNA sample of local sheep breeds

DNA Samples code	Codons			Genotypes		
	136	154	171	3 codons	2 codons (136 & 171)	Risk group
Hamdani	AA	RR	RQ	**ARR/ARQ**	**AR/AQ**	**R2**
Hamdani	AA	RR	RQ	**ARR/ARQ**	**AR/AQ**	**R2**
Hamdani	AA	RR	RQ	**ARR/ARQ**	**AR/AQ**	**R2**
Hamdani	AA	RR	RQ	**ARR/ARQ**	**AR/AQ**	**R2**
Hamdani	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Hamdani	AA	RR	RQ	**ARR/ARQ**	**AR/AQ**	**R2**
Hamdani	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Hamdani	AA	RR	RQ	**ARR/ARQ**	**AR/AQ**	**R2**
Hamdani	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Hamdani	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Hamdani	AA	RR	KK	**ARK/ARK**	**AK/AK**	**R3**
Karadi	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Karadi	AA	RR	HH	**ARH/ARH**	**AH/AH**	**R3**
Karadi	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Karadi	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Karadi	AA	RR	RQ	**ARR/ARQ**	**AR/AQ**	**R2**
Karadi	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Karadi	AA	RR	HH	**ARH/ARH**	**AH/AH**	**R3**
Karadi	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Karadi	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Karadi	AA	RR	RQ	**ARR/ARQ**	**AR/AQ**	**R2**
Karadi	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Karadi	AA	RR	RQ	**ARR/ARQ**	**AR/AQ**	**R2**
Karadi	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Karadi	AA	RR	HH	**ARH/ARH**	**AH/AH**	**R3**
Karadi	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**
Awassi	AA	RR	HH	**ARH/ARH**	**AH/AH**	**R3**
Awassi	AA	RR	RQ	**ARR/ARQ**	**AR/AQ**	**R2**
Awassi	AA	RR	RQ	**ARR/ARQ**	**AR/AQ**	**R2**
Awassi	AA	RR	QQ	**ARQ/ARQ**	**AQ/AQ**	**R3**

Different combinations of genotypes and allelic variants resulted. Overall, in the three sheep breeds, the most frequent genotype was ARQ/ARQ (3 codons); AQ/AQ (2 codons) with frequency 46.66%. The second highest frequency was for genotypes ARR/ARQ and AR/AQ with frequency 36.67%. The frequency of the other genotypes ARH/ARH, AH/AH and ARK/ARK, AK/AK was about 13.33% and 3.33% respectively Table 7.3.

Table 7.3 Frequency of different combinations of genotypes and allelic variants in each breed

Genotypes					sheep breeds						
					Hamdani		Karadi		Awassi		Total
3 codons			2 codons		N	%	N	%	N	%	%
136	154	171	136	171							
ARR/ARQ			AR/AQ		6	20	3	10	2	6.67	36.67
ARQ/ARQ			AQ/AQ		4	13.33	9	30	1	3.33	46.66
ARH/ARH			AH/AH		0	0	3	10	1	3.33	13.33
ARK/ARK			AK/AK		1	3.33	0	0	0	0	3.33

In terms of allelic variants, in total, 65% was for alleles ARQ (3 codons) and AQ (2 codons). The second frequent allelic variants were represented in ARR and AR with frequency 18.33%. The frequency of the other allelic variants ARH, AH and ARK, AK was about 13.33% and 3.33% respectively Table 7.4.

Table 7.4 Frequency of allelic variants at either two or three codons analysed in each breed

Allelic Variants					sheep breeds						
3 codons			2 codons		Hamdani		Karadi		Awassi		Total
136	154	171	136	171	N	%	N	%	N	%	%
ARR			AR		6	10	3	5	2	3.33	18.33
ARQ			AQ		14	23.33	21	35	4	6.67	65
ARH			AH		0	0	6	10	2	3.33	13.33
ARK			AK		2	3.33	0	0	0	0	3.33

In regard to the risk groups of scrapie disease, the most frequent genotypes belonged to the risk group R3 with 63%, followed by the risk group R2 with 37%. The risk group R3 is considered to be moderate susceptibility to scrapie disease, while the genotypes of R2 risk group has a lower level of susceptibility. The genotypes of R1, R4 and R5 were not found in any studied sheep breed.

Two additional polymorphisms were also found at different codons 127 and 146 of PrP gene. At the first position of codon 127, nucleotide substitution G>A was found by which the amino acid glycine G changed to serine S. The G127S polymorphism was found in both Hamdani and Awassi sheep breeds but not in Karadi breed. The other

polymorphism was at the second positon of codon 146 where nucleotide substitution A>G was present. This substitution changed the amino acid asparagine N to serine S. The N146S was found only in Karadi breed not in other breeds.

7.5 Discussion

This is the first study highlighting the genetic aspects of scrapie disease in Kurdistani sheep breeds around the centre of initial domestication of species. The relative frequencies of all potential combinations of genotype of PrP gene in terms of polymorphisms in the main candidate codons that relate to susceptibility and resistant to scrapie were described here. No polymorphic variations were found at either codon 136 or 154. Similarly, no polymorphism was found at 136 in many Pakistani sheep breeds (Babar *et al.* 2008; Babar *et al.* 2009; Hussain *et al.* 2011). Although the effect of allelic variants of codon 154 is in a minor role, it has generally been suggested that (H) at this position promotes the resistance to scrapie disease in some sheep breeds (Dawson *et al.* 1998; Elsen *et al.* 1999; Thorgeirsdottir *et al.* 1999). The allele of (H) at 154 codon has been identified in some of Asian (Gombojav *et al.* 2003; Lan *et al.* 2006), European (Hanušovská *et al.* 2003; Eglin *et al.* 2005; Holko *et al.* 2005), Pakistani (Babar *et al.* 2008) and Iranian sheep breeds (Karami *et al.* 2011).

All allelic variants at codon 154 in Karadi, Hamdani and Awassi were predominantly representing (R) while, no (H) polymorphism was found. Similarly, in Chinese sheep breeds of Xinjiang region, high allelic frequency of R was found at codon 154 which suppose that their sheep are at risk of scrapie (Lan *et al.* 2006). This does not mean that Kurdistani sheep breeds are free of scrapie disease because even in considered free countries of scrapie disease like Australia and New Zealand, cases of disease have been found (Hunter & Cairns 1998; Bossers *et al.* 1999).

On the other hand, alleles and genotypes at codon 171 were highly polymorphic. The genotypes ARQ/ARQ (3 codons) AQ/AQ (2 codons) and allelic variants ARQ (3 codons), AQ (2 codons) were the most frequent in Kurdistani sheep breeds. The same results have previously been reported in other sheep breeds in many studies (Elsen *et al.* 1999; DeSilva *et al.* 2003; Gombojav *et al.* 2004; Zhang *et al.* 2004; Gama *et al.* 2006; Ün *et al.*

2008). ARQ allele has been suggested to be ancestral allele of the PrP gene. The frequency of ARQ and AQ alleles was about 65% Tables 7.3 and 7.4. This allele is belonging to the risk group R3 which is moderate susceptibility to scrapie disease. Very low scrapie risks are associated to genotypes of R1 group while the high susceptibility to disease were linked to genotypes of R5 group (Table 7.1) (Dawson *et al.* 1998). The most frequent genotypes were belonging to the R3 groups including ARQ/ARQ and ARH/ARH with frequency 63% which considered to be the moderate level of risks. 37% of genotypes were found in R2 group which has less risks than the level R3. On the other hand, no genotypes were found in R1, R4 and R5 groups. In British sheep, Baylis et al, 2004 reported that the genotypes VRQ/VRQ ARH/VRQ and ARQ/VRQ were found within the greatest scrapie risk, while the genotype ARQ/ARQ was ranked as the next greatest risk. Animals with the genotype ARR/ARR were reported extremely resistant to scrapie, at least within their commercial lifespan (Jeffrey *et al.* 2014). Thus, neither complete resistant genotypes ARR/ARR (R1) nor extreme susceptibility genotypes R4 & R5 were found in Kurdistani sheep breeds. Furthermore, to date, no cases of scrapie have been recorded in Kurdistani sheep breeds. This is due to some possible reasons such as lack of suitable diagnosis screening systems, culling of animals at reasonably earlier ages and long incubation period of scrapie disease.

Additional polymorphisms at PrP gene at other than the main codons were found in the Kurdistani breeds. Other polymorphisms have been investigated in previous studies. However, there is no evidence of the role of these polymorphisms in scrapie disease. As with sheep breeds globally, it is necessary to know the genetic aspects of scrapie disease in the Iraqi Kurdistan Region. This enables improvement in sheep health and potential reduction in zoonoses (Cassard *et al.* 2014), as in other countries where long-term plans of selection and breeding programs increase the resistant genotypes of scrapie disease see (Hunter 2007).

Chapter 8 General discussion

In this study, analysis of genome composition and chromosomal characterization of dispersed, tandem and endogenous retrovirus-related repetitive DNA elements and other nuclear sequences such as *numts* has provided useful insights into their organization, homogenization and diversification in the sheep genome. This in turn may give further insight into possible functions of major repetitive sequences either dispersed throughout the chromosomes or co-localized in specific regions such as centromeres or telomeres.

Repetitive DNA families at centromeres play vital roles in their organization and evolution (Garrido-Ramos 2017). This study has shown that satellites, endogenous retroviruses, LINEs, SINEs and *numts* have all been amplified and are integrated components of centromeres or pericentromeric regions of sheep chromosomes. Likewise, pericentromeric regions of human chromosomes are dominated by alpha satellites but also intruded by other tandemly and dispersed DNA repeats LINEs, SINES and LTRs (Plohl *et al.* 2012). Indeed, satellite DNA sequences have been identified at the centromeres of almost all mammalian orders (see section 1.6). Centromeric abundance of ERVs in human (Zahn *et al.* 2015) indicates their amplification was limited to this region. Shi *et al.* (2010) and de Souza *et al.* (2017) note gene conversion could lead to massive amplification of LINEs at centromeres.

At meiosis, we see repetitive DNA sequences including centromeric and telomeric tandem sequences associating with each other. Thus, this physical association of SCs could play a crucial role in homogenization events of tandemly repeated DNA. The explanation behind presence of transposable elements such as LINEs and ERVs at the centromeres of sheep chromosome could be the outcome of different mechanisms including rolling cycle amplification, transposition, unequal crossing over and possibly other as yet unknown factors, which cause these repeats to amplify in some chromosomal regions and then jump into the centromeres. As inbreeding produces new tandemly repeated sequences at the centromeres of maize (Schneider *et al.* 2016),

similar studies on the impact of breeding on the diversity of DNA repeats at the centromeres of domesticated animals may be worthwhile.

Dispersed repeats like LINEs and SINEs can overlap coding regions, which in turn can lead to a change in or induce expression of some genes (Kidwell & Lisch 1997; van de Lagemaat *et al.* 2003; Chuong *et al.* 2017b; Hirsch & Springer 2017). One example is the presence of the brindle colour coat pattern in Normande cattle due to the insertion of non-LTR retrotransposon sequences in Agouti gene (Girardot *et al.* 2006). It would therefore be interesting to investigate the *in silico* relationship between repetitive elements with other components of sheep genome such as coding sequences. However, this will require a complete reference sheep genome (see section 1.15).

It has been hypothesized that some tandemly repeated sequences might have originated directly from the sequences of transposable elements (retrotransposons and/or transposons) either by the amplification of pre-existing internal repetitive sequences or by misaligned recombination. Similarly, conversion of transposable elements to tandem repeats is also possible as long as satellite DNA can be present as integral part of transposable elements (Biscotti *et al.* 2015b; Meštrović *et al.* 2015).

The tools used in this study allow analysis of unassembled sequence reads without the need for a reference genome to identify and quantify tandemly and dispersed repetitive DNA landscapes of many animal and plant species (Weiss-Schneeweiss *et al.* 2015; Ruiz-Ruano *et al.* 2016) see also sections 1.7.1 & 1.7.2. The bioinformatics pipelines used are thus helping to launch a new era in discovering the abundance and variability of different repeat families within genomes and thus contribute to the development of comparative genomics and phylogenomics. This will increase insight into the role of repeat sequences in processes that shape genome structure and contribute to evolution (Novák *et al.* 2010; Novák *et al.* 2013; Novák *et al.* 2017).

These methods allowed us to develop some repetitive-based molecular markers, namely Y chromosome-specific repeats and sheep-specific tandem repeats. NGS data of domestic sheep were assembled to satellite sequences of wild sheep, and satellite DNA sequences could be used as phylogenetic markers. Satellite sequences are not only

similar between human and other primates (Plohl *et al.* 2012), but also between human and sheep. We found that a 53bp portion of probe CL19_Low complexity and SR (see section 5.6.5) is highly similar to the beta satellite core sequence in *Homo sapiens* (Meneveri *et al.* 1993; Winokur *et al.* 1994) and a strong centromeric distribution was also confirmed by FISH (see Figure 5.18).

Different classes of repetitive DNA sequences show substantial variation between individuals, animal lineages and over longer timescales associated with animal domestication, chromosomal and genome evolution. In comparison to the maternal lineage of sheep breeds, repetitive DNA sequences along with cytological methods exposed differences among intra and inter-species such as sheep and cattle genomes. For example, this study found dispersed and tandemly repetitive DNA sequences in the sheep genome are highly heterogeneous with respect to copy number and thus contribute to variation in genome size and karyotypic diversity. Also, using FISH, a variable genomic (dispersed to centromeric) distribution of LINEs and SINEs were seen on sheep and cattle chromosomes in this study. Thus, repetitive DNA sequences can be used in comparative analysis as a good marker to distinguish between species within the Bovidae family and to utilize them for breeding and selection. The recent advances in NGS projects, bioinformatics analysis and molecular cytogenetics have greatly increased our understanding of the nature and behavior of major component of genome.

Chapter 9 General conclusions

Taken together, the results in this thesis show that using modern sequencing strategies, bioinformatics algorithms and *in situ* hybridization, we can deeply explore genome structure and organization (see also Figure 9.1). The results show evolution, and evolutionary mechanisms, of chromosomes and DNA sequences, during longer and shorter evolutionary periods. The data also provide informative markers – mitochondrial and repetitive DNA – for specific species or tribes within the Bovidae family, as well as landraces and breeds of sheep. Further studies should be carried out to extend the studies to mitochondrial, dispersed and tandemly repeated DNA in the genomes of wild sheep or other species within Bovidae family. Additionally, *in situ* hybridization of repeats over chromosome of wide range of wild sheep and goat species would expand the knowledge about evolutionary events such as diversification and homogenization of sequences and restructuring of the genome.

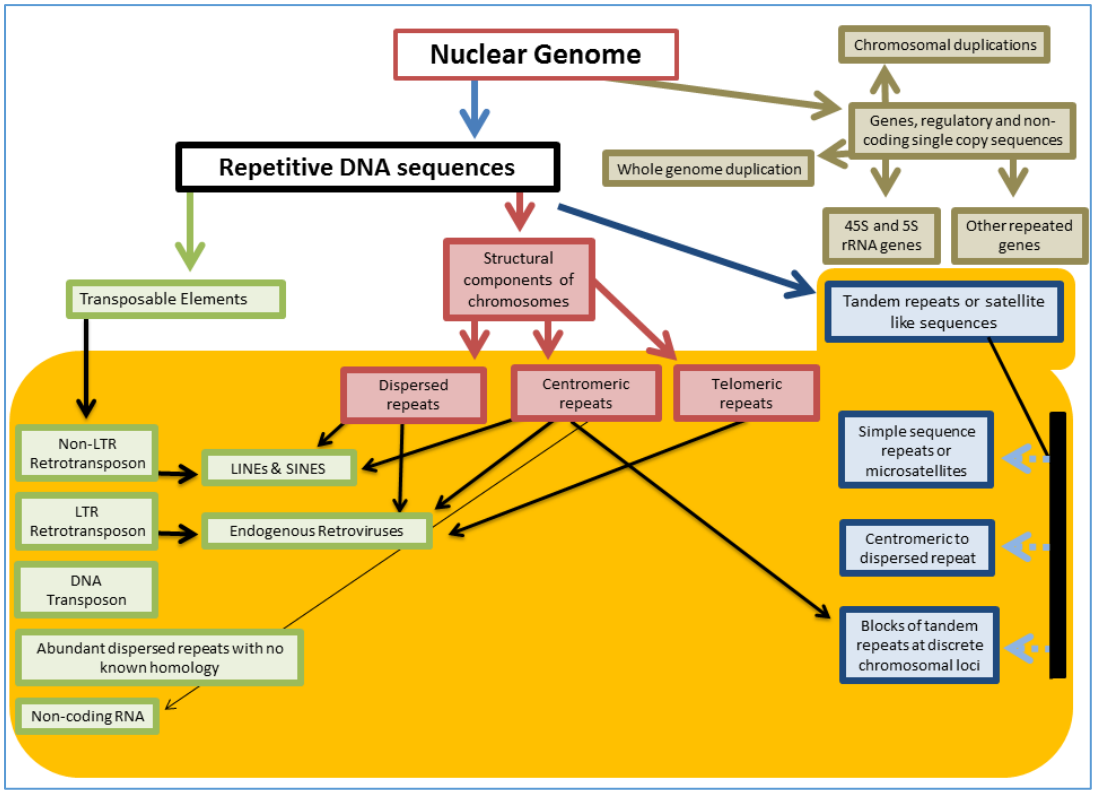


Figure 9.1 DNA sequence component of the nuclear genome of sheep including modified part (Orange background) of repetitive landscapes according to the findings of this thesis. Modified for sheep genome from overview of repetitive landscapes see section 1.4 for comparison (Biscotti *et al.* 2015b).

References

Abrusán G., Grundmann N., DeMester L. & Makalowski W. (2009) TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329-30.

Adega F., Chaves R. & Guedes-Pinto H. (2007) Chromosomal evolution and phylogenetic analyses in Tayassu pecari and Pecari tajacu (Tayassuidae): tales from constitutive heterochromatin. *Journal of genetics* **86**, 19-26.

Adega F., Chaves R., Guedes-Pinto H. & Heslop-Harrison J. (2006) Physical organization of the 1.709 satellite IV DNA family in Bovini and Tragelaphini tribes of the Bovidae: sequence and chromosomal evolution. *Cytogenetic and Genome Research* **114**, 140.

Adega F., Guedes-Pinto H. & Chaves R. (2009) Satellite DNA in the karyotype evolution of domestic animals–clinical considerations. *Cytogenetic and Genome Research* **126**, 12-20.

Adelson D.L., Raison J.M. & Edgar R.C. (2009) Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proceedings of the National Academy of Sciences* **106**, 12855-60.

Al-Azzawi W.A. (1977) A comparative study of fleece characteristics in Iraq sheep. In: *Faculty of Agriculture*. University of Cairo., Egypt.

Al-Barzinj Y.M.S. & Ali M.K. (2013) Genetic Diversity among Some Sheep Breeds in Sulaimani Governorate Using RAPD-PCR Technique. *Journal of Life Sciences* **7**, 971.

Al-Barzinji Y., Lababidi S., Rischkowsky B., Al-Rawi A., Tibbo M., Hassen H. & Baum M. (2011) Assessing genetic diversity of Hamdani sheep breed in Kurdistan region of Iraq using microsatellite markers. *African Journal of Biotechnology* **10**, 15109-16.

Al-Mourrani W., Mahamoud A.K. & Al-Wahab R.M. (1980) *Animal Management*, Baghdad, Iraq. In Arabic.

Al-Rawi A., Al-Haboby A. & Salman A. (1996) Small Ruminants Breeding, Reproduction Physiology Research as Technology Transfer I Iraq. *ICARDA West Asia Regional Program, Amman, Jordan*.

Alkan C., Cardone M.F., Catacchio C.R., Antonacci F., O'Brien S.J., Ryder O.A., Purgato S., Zoli M., Della Valle G. & Eichler E.E. (2011) Genome-wide characterization of centromeric satellites from multiple mammalian genomes. *Genome research* **21**, 137-45.

Alkass J.E. & Akreyi I.A.I. (2016) Full Length Research Paper Milk Production of Awassi and Karadi Ewes Raised Under Farm Conditions.

Alkass J.E. & Juma K.H. (2005) Small ruminant breeds of Iraq. *Characterization of Small Ruminant Breeds in West Asia, North Africa* **1**, 63-101.

Allard M.W., Miyamoto M.M., Jarecki L., KRAuS F. & Tennant M.R. (1992) DNA systematics and evolution of the artiodactyl family Bovidae. *Proceedings of the National Academy of Sciences* **89**, 3972-6.

Alter M.D., Rubin D.B., Ramsey K., Halpern R., Stephan D.A., Abbott L. & Hen R. (2008) Variation in the large-scale organization of gene expression levels in the hippocampus relates to stable epigenetic variability in behavior. *PloS one* **3**, e3344.

Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990) Basic local alignment search tool. *Journal of molecular biology* **215**, 403-10.

Archibald A., Cockett N., Dalrymple B., Faraut T., Kijas J., Maddox J., McEwan J., Hutton Oddy V., Raadsma H. & Wade C. (2010) The sheep genome reference sequence: a work in progress. *Animal genetics* **41**, 449-53.

Argeson A.C., Nelson K.K. & Siracusa L.D. (1996) Molecular basis of the pleiotropic phenotype of mice carrying the hypervariable yellow (A hvy) mutation at the agouti locus. *Genetics* **142**, 557-67.

Arnason U., Gullberg A., Gretarsdottir S., Ursing B. & Janke A. (2000) The mitochondrial genome of the sperm whale and a new molecular reference for estimating eutherian divergence dates. *Journal of Molecular Evolution* **50**, 569-78.

Arnaud F., Caporale M., Varela M., Biek R., Chessa B., Alberti A., Golder M., Mura M., Zhang Y.-p. & Yu L. (2007) A paradigm for virus–host coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses. *PLoS Pathogens* **3**, e170.

Babar M., Farid A., Benkel B., Ahmad J., Nadeem A. & Imran M. (2009) Frequencies of PrP genotypes and their implication for breeding against scrapie susceptibility in nine Pakistani sheep breeds. *Molecular biology reports* **36**, 561-5.

Babar M., Farid A., Benkel B., Ahmad J., Sajid I., Imran M., Hussain T. & Nadeem A. (2008) Genetic variability at seven codons of the prion protein gene in nine Pakistani sheep breeds. *Journal of genetics* **87**, 187-90.

Bai J., Bishop J.V., Carlson J.O. & DeMartini J.C. (1999) Sequence comparison of JSRV with endogenous proviruses: envelope genotypes and a novel ORF with similarity to a G-protein-coupled receptor. *Virology* **258**, 333-43.

Bai J., Zhu R., Stedman K., Cousens C., Carlson J., Sharp J. & DeMartini J. (1996) Unique long terminal repeat U3 sequences distinguish exogenous jaagsiekte sheep retroviruses associated with ovine pulmonary carcinoma from endogenous loci in the sheep genome. *Journal of virology* **70**, 3159-68.

Bailey J.A., Carrel L., Chakravarti A. & Eichler E.E. (2000) Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proceedings of the National Academy of Sciences* **97**, 6634-9.

Baillie G.J., van de Lagemaat L.N., Baust C. & Mager D.L. (2004) Multiple groups of endogenous betaretroviruses in mice, rats, and other mammals. *Journal of virology* **78**, 5784-98.

Balmus G., Trifonov V.A., Biltueva L.S., O'Brien P.C., Alkalaeva E.S., Fu B., Skidmore J.A., Allen T., Graphodatsky A.S. & Yang F. (2007) Cross-species chromosome painting among camel, cattle, pig and human: further insights into the putative Cetartiodactyla ancestral karyotype. *Chromosome Research* **15**, 499-514.

Bannert N. & Kurth R. (2006) The evolutionary dynamics of human endogenous retroviral families. *Annu. Rev. Genomics Hum. Genet.* **7**, 149-73.

Bao W., Kojima K.K. & Kohany O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11.

Bao Z. & Eddy S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research* **12**, 1269-76.

Barlow A.L. & Hultén M. (1996) Combined immunocytogenetic and molecular cytogenetic analysis of meiosis I human spermatocytes. *Chromosome Research* **4**, 562-73.

Bensasson D., Zhang D.-X., Hartl D.L. & Hewitt G.M. (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in ecology & evolution* **16**, 314-21.

Benson G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573.

Benveniste R.E. & Todaro G.J. (1973) Homology between type-C viruses of various species as determined by molecular hybridization. *Proceedings of the National Academy of Sciences* **70**, 3316-20.

Benveniste R.E. & Todaro G.J. (1977) Evolution of primate oncornaviruses: An endogenous virus from langurs (Presbytis spp.) with related virogene sequences in other Old World monkeys. *Proceedings of the National Academy of Sciences* **74**, 4557-61.

Berardino D.D., Iannuzzi L., Bettini T. & Matassino D. (1981) Ag-NORs variation and banding homologies in two species of Bovidae: Bubalus bubalis and Bos taurus. *Canadian Journal of Genetics and Cytology* **23**, 89-99.

Bergman C.M. & Quesneville H. (2007) Discovering and detecting transposable elements in genome sequences. *Briefings in bioinformatics* **8**, 382-92.

Bibi F. (2013) A multi-calibrated mitochondrial phylogeny of extant Bovidae (Artiodactyla, Ruminantia) and the importance of the fossil record to systematics. *BMC Evolutionary Biology* **13**, 166.

Biémont C. & Vieira C. (2006) Genetics: junk DNA as an evolutionary force. *Nature* **443**, 521-4.

Biscotti M.A., Canapa A., Forconi M., Olmo E. & Barucca M. (2015a) Transcription of tandemly repetitive DNA: functional roles. *Chromosome Research* **23**, 463-77.

Biscotti M.A., Olmo E. & Heslop-Harrison J.P. (2015b) Repetitive DNA in eukaryotic genomes. *Chromosome Research* **23**, 415-20.

Bock M. & Stoye J.P. (2000) Endogenous retroviruses and the human germline. *Current opinion in genetics & development* **10**, 651-5.

Boeke J. & Stoye J. (1997) Retrotransposons, endogenous retroviruses, and the evolution of retroelements.

Boissinot S. & Furano A.V. (2001) Adaptive evolution in LINE-1 retrotransposons. *Molecular Biology and Evolution* **18**, 2186-94.

Bolcun-Filas E. & Schimenti J.C. (2012) 5 Genetics of meiosis and recombination in mice. *International review of cell and molecular biology* **298**, 179.

Bonaccorsi S. & Lohe A. (1991) Fine mapping of satellite DNA sequences along the Y chromosome of Drosophila melanogaster: relationships between satellite sequences and fertility factors. *Genetics* **129**, 177-89.

Bossers A., Harders F. & Smits M. (1999) PrP genotype frequencies of the most dominant sheep breed in a country free from scrapie. *Archives of virology* **144**, 829-34.

Bourque G., Zdobnov E.M., Bork P., Pevzner P.A. & Tesler G. (2005) Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome research* **15**, 98-110.

Boyle A.L., Ballard S.G. & Ward D.C. (1990) Differential distribution of long and short interspersed element sequences in the mouse genome: chromosome karyotyping by fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences* **87**, 7757-61.

Britten R.J. & Kohne D.E. (1968) Repeated sequences in DNA. *Science* **161**, 529-40.

Brown D.R. (2005) *Neurodegeneration and prion disease*. Springer Science & Business Media.

Brown S.W. (1966) Heterochromatin. *Science* **151**, 417-25.

Bruford M.W. & Townsend S.J. (2006) Mitochondrial DNA diversity in modern sheep. *Documenting domestication: New genetic and archaeological paradigms*, 306-16.

Brutlag D.L. (1980) Molecular arrangement and evolution of heterochromatic DNA. *Annual review of genetics* **14**, 121-44.

Buckland R. & Evans H. (1978) Cytogenetic aspects of phylogeny in the Bovidae. *Cytogenetic and Genome Research* **21**, 42-63.

Buckland R.A. (1983) Comparative structure and evolution of goat and sheep satellite I DNAs. *Nucleic acids research* **11**, 1349-60.

Buckland R.A. (1985) Sequence and evolution of related bovine and caprine satellite DNAs: identification of a short DNA sequence potentially involved in satellite DNA amplification. *Journal of molecular biology* **186**, 25-30.

Buermans H. & Den Dunnen J. (2014) Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1842**, 1932-41.

Bunch T., Foote W. & Spillett J. (1976) Translocations of acrocentric chromosomes and their implications in the evolution of sheep (Ovis). *Cytogenetic and Genome Research* **17**, 122-36.

Bunch T. & Nadler C. (1980) Giemsa-band patterns of the tahr and chromosomal evolution of the tribe Caprini. *Journal of Heredity* **71**, 110-6.

Bunch T.D., Wu C., Zhang Y.-P. & Wang S. (2005) Phylogenetic analysis of snow sheep (Ovis nivicola) and closely related taxa. *Journal of Heredity* **97**, 21-30.

Buntjer J.B., Nijman I.J., Zijlstra C. & Lenstra J.A. (1998) A satellite DNA element specific for roe deer (Capreolus capreolus). *Chromosoma* **107**, 1-5.

Burgstaller J.P., Schinogl P., Dinnyes A., Müller M. & Steinborn R. (2007) Mitochondrial DNA heteroplasmy in ovine fetuses and sheep cloned by somatic cell nuclear transfer. *BMC developmental biology* **7**, 141.

Burkin D.J., Broad T.E. & Jones C. (1996) The chromosomal distribution and organization of sheep satellite I and II centromeric DNA using characterized sheep-hamster somatic cell hybrids. *Chromosome Research* **4**, 49-55.

Cai Y., Zhang L., Shen F., Zhang W., Hou R., Yue B., Li J. & Zhang Z. (2011) DNA barcoding of 18 species of Bovidae. *Chinese Science Bulletin* **56**, 164-8.

Capilla L., Caldés M.G. & Ruiz-Herrera A. (2016) Mammalian meiotic recombination: a toolbox for genome evolution. *Cytogenetic and Genome Research* **150**, 1-16.

Cassard H., Torres J.-M., Lacroux C., Douet J.-Y., Benestad S.L., Lantier F., Lugan S., Lantier I., Costes P. & Aron N. (2014) Evidence for zoonotic potential of ovine scrapie prions. *Nature communications* **5**, 5821.

Cerutti F., Gamba R., Mazzagatti A., Piras F.M., Cappelletti E., Belloni E., Nergadze S.G., Raimondi E. & Giulotto E. (2016) The major horse satellite DNA family is associated with centromere competence. *Molecular cytogenetics* **9**, 35.

Charlesworth B., Sniegowski P. & Stephan W. (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**, 215-20.

Chaves R., Adega F., Heslop-Harrison J., Guedes-Pinto H. & Wienberg J. (2003a) Complex satellite DNA reshuffling in the polymorphic t (1; 29) Robertsonian translocation and evolutionarily derived chromosomes in cattle. *Chromosome Research* **11**, 641-8.

Chaves R., Adega F., Wienberg J., Guedes-Pinto H. & Heslop-Harrison J.S. (2003b) Molecular cytogenetic analysis and centromeric satellite organization of a novel 8; 11 translocation in sheep: a possible intermediate in biarmed chromosome evolution. *Mammalian Genome* **14**, 706-10.

Chaves R., Guedes-Pinto H., Heslop-Harrison J. & Schwarzacher T. (2000a) The species and chromosomal distribution of the centromeric α-satellite I sequence from sheep in the tribe Caprini and other Bovidae. *Cytogenetic and Genome Research* **91**, 62-6.

Chaves R., Guedes-Pinto H. & Heslop-Harrison J.S. (2005) Phylogenetic relationships and the primitive X chromosome inferred from chromosomal and satellite DNA analysis in Bovidae. *Proceedings of the Royal Society of London B: Biological Sciences* **272**, 2009-16.

Chaves R., Heslop-Harrison J. & Guedes-Pinto H. (2000b) Centromeric heterochromatin in the cattle rob (1; 29) translocation: α-satellite I sequences, in-situ MspI digestion patterns, chromomycin staining and C-bands. *Chromosome Research* **8**, 621-6.

Chaves R., Santos S. & Guedes-Pinto H. (2004) Comparative analysis (Hippotragini versus Caprini, Bovidae) of X-chromosome's constitutive heterochromatin by in situ

restriction endonuclease digestion: X-chromosome constitutive heterochromatin evolution. *Genetica* **121**, 315-25.

Chen A.Y. & Chen A. (2013) Fluorescence in situ hybridization. *The Journal of investigative dermatology* **133**, e8.

Cheng C.-H., Lo Y.-H., Liang S.-S., Ti S.-C., Lin F.-M., Yeh C.-H., Huang H.-Y. & Wang T.-F. (2006) SUMO modifications control assembly of synaptonemal complex and polycomplex in meiosis of Saccharomyces cerevisiae. *Genes & development* **20**, 2067-81.

Chessa B., Pereira F., Arnaud F., Amorim A., Goyache F., Mainland I., Kao R.R., Pemberton J.M., Beraldi D. & Stear M.J. (2009) Revealing the history of sheep domestication using retrovirus integrations. *Science* **324**, 532-6.

Chow J.C., Ciaudo C., Fazzari M.J., Mise N., Servant N., Glass J.L., Attreed M., Avner P., Wutz A. & Barillot E. (2010) LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141**, 956-69.

Chow Y.-H.J., Alberti A., Mura M., Pretto C., Murcia P., Albritton L.M. & Palmarini M. (2003) Transformation of rodent fibroblasts by the jaagsiekte sheep retrovirus envelope is receptor independent and does not require the surface domain. *Journal of virology* **77**, 6341-50.

Chuong E.B., Elde N.C. & Feschotte C. (2016) Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083-7.

Chuong E.B., Elde N.C. & Feschotte C. (2017a) Regulatory activities of transposable elements: from conflicts to benefits. *Nature reviews. Genetics* **18**, 71.

Chuong E.B., Elde N.C. & Feschotte C. (2017b) Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics* **18**, 71.

Church D.M., Schneider V.A., Graves T., Auger K., Cunningham F., Bouk N., Chen H.-C., Agarwala R., McLaren W.M. & Ritchie G.R. (2011) Modernizing reference genome assemblies. *PLoS biology* **9**, e1001091.

Church K.W. & Helfman J.I. (1993) Dotplot: A Program for Exploring Self-Similarity in Millions of Lines of Text and Code. *Journal of Computational and Graphical Statistics* **2**, 153-74.

Coffin J., Hughes S. & Varmus H. (1997) Retroviruses. Woodbury. New York, USA: Cold Spring Harbor Laboratory Press.

Cohen C.J., Lock W.M. & Mager D.L. (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* **448**, 105-14.

Contento A., Heslop-Harrison J. & Schwarzacher T. (2005) Diversity of a major repetitive DNA sequence in diploid and polyploid Triticeae. *Cytogenetic and Genome Research* **109**, 34-42.

Cordaux R. & Batzer M.A. (2009) The impact of retrotransposons on human genome evolution. *Nature reviews. Genetics* **10**, 691.

Cost G.J., Feng Q., Jacquier A. & Boeke J.D. (2002) Human L1 element target-primed reverse transcription in vitro. *The EMBO journal* **21**, 5899-910.

Costa Y. & Cooke H.J. (2007) Dissecting the mammalian synaptonemal complex using targeted mutations. *Chromosome Research* **15**, 579-89.

Craig N.L. (2002) Mobile DNA: an introduction. In: *Mobile DNA II* (pp. 3-11. American Society of Microbiology.

Csink A.K. & Henikoff S. (1998) Something from nothing: the evolution and utility of satellite repeats. *TRENDS in Genetics* **14**, 200-4.

Curtain C., Pascoe G. & Hayman R. (1973) Satellite DNA in the sheep and goat. *Biochemical genetics* **10**, 253-62.

D'aiuto L., Barsanti P., Mauro S., Cserpan I., Lanave C. & Ciccarese S. (1997) Physical relationship between satellite I and II DNA in centromeric regions of sheep chromosomes. *Chromosome Research* **5**, 375-81.

Dadashev S.Y., Grishaeva T. & Bogdanov Y.F. (2005) In Silico identification and characterization of meiotic DNA: AluJb possibly participates in the attachment of chromatin loops to synaptonemal complex. *Russian Journal of Genetics* **41**, 1419-24.

Danilkovitch-Miagkova A., Duh F.-M., Kuzmin I., Angeloni D., Liu S.-L., Miller A.D. & Lerman M.I. (2003) Hyaluronidase 2 negatively regulates RON receptor tyrosine kinase and mediates transformation of epithelial cells by jaagsiekte sheep retrovirus. *Proceedings of the National Academy of Sciences* **100**, 4580-5.

Dávila-Rodríguez M., Cortés-Gutiérrez E., López-Fernández C., Pita M., Mezzanotte R. & Gosálvez J. (2009) Whole-comparative genomic hybridization in domestic sheep (Ovis aries) breeds. *Cytogenetic and Genome Research* **124**, 19-26.

Dawson M., Hoinville L., Hosie B. & Hunter N. (1998) Guidance on the use of PrP genotyping as an aid to the control of clinical scrapie. *Veterinary Record* **142**, 623-5.

de Souza É.M.S., Gross M.C., Silva C.E.F., Sotero-Caio C.G. & Feldberg E. (2017) Heterochromatin variation and LINE-1 distribution in Artibeus (Chiroptera, Phyllostomidae) from Central Amazon, Brazil. *Comparative Cytogenetics* **11**, 613.

de Vries F.A., de Boer E., van den Bosch M., Baarends W.M., Ooms M., Yuan L., Liu J.-G., van Zeeland A.A., Heyting C. & Pastink A. (2005) Mouse Sycp1 functions in synaptonemal complex assembly, meiotic recombination, and XY body formation. *Genes & development* **19**, 1376-89.

DeMartini J.C. & York D.F. (1997) Retrovirus-associated neoplasms of the respiratory system of sheep and goats: ovine pulmonary carcinoma and enzootic nasal tumor. *Veterinary clinics of North America: food animal practice* **13**, 55-70.

Demirci S., Baştanlar E.K., Dağtaş N.D., Pişkin E., Engin A., Özer F., Yüncü E., Doğan Ş.A. & Togan İ. (2013) Mitochondrial DNA diversity of modern, ancient and wild sheep (Ovis gmelinii anatolica) from Turkey: new insights on the evolutionary history of sheep. *PloS one* **8**, e81952.

Deragon J.-M. & Zhang X. (2006) Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Systematic biology* **55**, 949-56.

DeSilva U., Guo X., Kupfer D., Fernando S., Pillai A., Najar F., So S., Fitch G. & Roe B. (2003) Allelic variants of ovine prion protein gene (PRNP) in Oklahoma sheep. *Cytogenetic and Genome Research* **102**, 89-94.

Detwiler L. & Baylis M. (2003) The epidemiology of scrapie. *Revue scientifique et technique (International Office of Epizootics)* **22**, 121-43.

Di Berardino D., Di Meo G., Gallagher D., Hayes H. & Iannuzzi L. (2001) International system for chromosome nomenclature of domestic bovids (ISCNDB 2000). *Cytogenetic and Genome Research* **92**, 284.

Di Meo G., Perucatti A., Floriot S., Hayes H., Schibler L., Rullo R., Incarnato D., Ferretti L., Cockett N. & Cribiu E. (2007) An advanced sheep (Ovis aries, 2n= 54) cytogenetic map and assignment of 88 new autosomal loci by fluorescence in situ hybridization and R-banding. *Animal genetics* **38**, 233-40.

Ding J., Sidore C., Butler T.J., Wing M.K., Qian Y., Meirelles O., Busonero F., Tsoi L.C., Maschio A. & Angius A. (2015) Assessing mitochondrial DNA variation and copy number in lymphocytes of~ 2,000 Sardinians using tailored sequencing analysis tools. *PLoS genetics* **11**, e1005306.

Dodsworth S., Chase M.W., Kelly L.J., Leitch I.J., Macas J., Novák P., Piednoël M., Weiss-Schneeweiss H. & Leitch A.R. (2015) Genomic Repeat Abundances Contain Phylogenetic Signal. *Systematic biology* **64**, 112-26.

Dong Y., Zhang X., Xie M., Arefnezhad B., Wang Z., Wang W., Feng S., Huang G., Guan R. & Shen W. (2015) Reference genome of wild goat (capra aegagrus) and sequencing of goat breeds provide insight into genic basis of goat domestication. *BMC genomics* **16**, 431.

Dool S.E., Puechmaille S.J., Foley N.M., Allegrini B., Bastian A., Mutumi G.L., Maluleke T.G., Odendaal L.J., Teeling E.C. & Jacobs D.S. (2016) Nuclear introns outperform mitochondrial DNA in inter-specific phylogenetic reconstruction: lessons from horseshoe bats (Rhinolophidae: Chiroptera). *Molecular phylogenetics and evolution* **97**, 196-212.

Dover G. (1982) Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111-7.

Du W. & Qin Y. (2015) Distribution of mitochondrial DNA fragments in the nuclear genome of the honeybee. *Genetics and Molecular Research* **14**, 13375-9.

Dunnen J.T., Dalgleish R., Maglott D.R., Hart R.K., Greenblatt M.S., McGowan-Jordan J., Roux A.F., Smith T., Antonarakis S.E. & Taschner P.E. (2016) HGVS recommendations for the description of sequence variants: 2016 Update. *Human mutation* **37**, 564-9.

Edgar R.C. & Myers E.W. (2005) PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, i152-i8.

Eglin R., Warner R., Gubbins S., Sivam S. & Dawson M. (2005) Frequencies of PrP genotypes in 38 breeds of sheep sampled in the National Scrapie Plan for Great Britain. *The Veterinary Record* **156**, 433-7.

Eickbush T.H. & Jamburuthugoda V.K. (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus research* **134**, 221-34.

Elder Jr J.F. & Turner B.J. (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quarterly Review of Biology* **70**, 297-320.

Ellinghaus D., Kurtz S. & Willhoeft U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics* **9**, 18.

Elsen J.-M., Amigues Y., Schelcher F., Ducrocq V., Andreoletti O., Eychenne F., Khang J.T., Poivey J.-P., Lantier F. & Laplanche J.-L. (1999) Genetic susceptibility and transmission factors in scrapie: detailed analysis of an epidemic in a closed flock of Romanov. *Archives of virology* **144**, 431-45.

Elsik C.G., Tellam R.L. & Worley K.C. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522-8.

Evans H., Buckland R.a. & Sumner A. (1973) Chromosome homology and heterochromatin in goat, sheep and ox studied by banding techniques. *Chromosoma* **42**, 383-402.

Fan H., Palmarini M. & DeMartini J. (2003) Transformation and oncogenesis by jaagsiekte sheep retrovirus. In: *Jaagsiekte Sheep Retrovirus and Lung Cancer* (pp. 139-77. Springer.

Fedosenko A.K. & Blank D.A. (2005) Ovis ammon. *Mammalian Species*, 1-15.

Feliciello I., Picariello O. & Chinali G. (2005) The first characterisation of the overall variability of repetitive units in a species reveals unexpected features of satellite DNA. *Gene* **349**, 153-64.

Feng R., Wang X., Tao M., Du G. & Wang Q. (2017) Genome size and identification of abundant repetitive sequences in Vallisneria spinulosa. *PeerJ* **5**, e3982.

Ferguson-Smith M., Yang F., Rens W. & O'Brien P. (2005) The impact of chromosome sorting and painting on the comparative analysis of primate genomes. *Cytogenetic and Genome Research* **108**, 112-21.

Fernández M.H. & Vrba E.S. (2005) A complete estimate of the phylogenetic relationships in Ruminantia: a dated species-level supertree of the extant ruminants. *Biological reviews* **80**, 269-302.

Ferreri G.C., Brown J.D., Obergfell C., Jue N., Finn C.E., O'Neill M.J. & O'Neill R.J. (2011) Recent amplification of the kangaroo endogenous retrovirus, KERV, limited to the centromere. *Journal of virology* **85**, 4761-71.

Feschotte C. (2008) The contribution of transposable elements to the evolution of regulatory networks. *Nature reviews. Genetics* **9**, 397.

Feschotte C. & Pritham E.J. (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* **41**, 331-68.

Festa-Bianchet M. (2000) A summary of discussion on the taxonomy of mountain ungulates and its conservation implications. In: *Workshop on Caprinae taxonomy, Ankara, Turkey*.

Finnegan D.J. (1989) Eukaryotic transposable elements and genome evolution. *TRENDS in Genetics* **5**, 103-7.

Fiston-Lavier A.-S., Carrigan M., Petrov D.A. & González J. (2010) T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic acids research* **39**, e36-e.

Ford C., Pollock D. & Gustavsson I. (1980) Proceedings of the First International Conference for the Standardisation of Banded Karyotypes of Domestic Animals University of Reading Reading, England 2nd-6th August 1976. *Hereditas* **92**, 145-62.

Fuentes-Pardo A.P. & Ruzzante D.E. (2017) Whole-genome sequencing approaches for conservation biology: advantages, limitations, and practical recommendations. *Molecular ecology*.

Gaillard C., Doly J., Cortadas J. & Bernardi G. (1981) The primary structure of bovine satellite 1.715. *Nucleic acids research* **9**, 6069-82.

Gall J.G. & Pardue M.L. (1969) Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proceedings of the National Academy of Sciences* **63**, 378-83.

Gallagher D., Davis S., De Donato M., Burzlaff J., Womack J., Taylor J. & Kumamoto A. (1999) A Molecular Cytogenetic Analysis of the Tribe Bovini (Artiodactyla: Bovidae: Bovinae) with an Emphasis on Sex Shromosome Morphology and NOR Distribution. *Chromosome Research* **7**, 481-92.

Gallagher Jr D. & Womack J. (1992) Chromosome conservation in the Bovidae. *Journal of Heredity* **83**, 287-98.

Gallus S., Kumar V., Bertelsen M.F., Janke A. & Nilsson M. (2015) A genome survey sequencing of the Java mouse deer (Tragulus javanicus) adds new aspects to the evolution of lineage specific retrotransposons in Ruminantia (Cetartiodactyla). *Gene* **571**, 271-8.

Gama L., Carolino M., Santos-Silva M., Pimenta J. & Costa M. (2006) Prion protein genetic polymorphisms and breeding strategies in Portuguese breeds of sheep. *Livestock Science* **99**, 175-84.

Garcia-Etxebarria K. & Jugo B.M. (2010) Genome-wide detection and characterization of endogenous retroviruses in Bos taurus. *Journal of virology* **84**, 10852-62.

García G., Ríos N. & Gutiérrez V. (2015) Next-generation sequencing detects repetitive elements expansion in giant genomes of annual killifish genus Austrolebias (Cyprinodontiformes, Rivulidae). *Genetica* **143**, 353-60.

Garrido-Ramos M.A. (2015) Satellite DNA in plants: more than just rubbish. *Cytogenetic and Genome Research* **146**, 153-70.

Garrido-Ramos M.A. (2017) Satellite DNA: an evolving topic. *Genes* **8**, 230.

Gatesy J., Yelon D., DeSalle R. & Vrba E. (1992) Phylogeny of the Bovidae (Artiodactyla, Mammalia), based on mitochondrial ribosomal DNA sequences. *Molecular Biology and Evolution* **9**, 433-46.

Gentles A.J., Wakefield M.J., Kohany O., Gu W., Batzer M.A., Pollock D.D. & Jurka J. (2007) Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica. *Genome research* **17**, 992-1004.

Ghosh W., George A., Agarwal A., Raj P., Alam M., Pyne P. & Gupta S.K.D. (2011) Whole-genome shotgun sequencing of the sulfur-oxidizing chemoautotroph Tetrathiobacter kashmirensis. *Journal of bacteriology* **193**, 5553-4.

Gifford R., Kabat P., Martin J., Lynch C. & Tristem M. (2005) Evolution and distribution of class II-related endogenous retroviruses. *Journal of virology* **79**, 6478-86.

Girardot M., Guibert S., Laforet M.P., Gallard Y., Larroque H. & Oulmouden A. (2006) The insertion of a full-length Bos taurus LINE element is responsible for a transcriptional deregulation of the Normande Agouti gene. *Pigment Cell & Melanoma Research* **19**, 346-55.

Girardot M., Martin J., Guibert S., Leveziel H., Julien R. & Oulmouden A. (2005) Widespread expression of the bovine Agouti gene results from at least three alternative promoters. *Pigment Cell & Melanoma Research* **18**, 34-41.

Goff S.P. (2007) Host factors exploited by retroviruses. *Nature reviews. Microbiology* **5**, 253.

Goldammer T., Di Meo G., Lühken G., Drögemüller C., Wu C., Kijas J., Dalrymple B., Nicholas F., Maddox J. & Iannuzzi L. (2009) Molecular cytogenetics and gene mapping in sheep (Ovis aries, 2n= 54). *Cytogenetic and Genome Research* **126**, 63-76.

Goldmann W., Hunter N., Foster J.D., Salbaum J.M., Beyreuther K. & Hope J. (1990) Two alleles of a neural protein gene linked to scrapie in sheep. *Proceedings of the National Academy of Sciences* **87**, 2476-80.

Goldmann W., Hunter N., Smith G., Foster J. & Hope J. (1994) PrP genotype and agent effects in scrapie: change in allelic interaction with different isolates of agent in sheep, a natural host of scrapie. *Journal of general virology* **75**, 989-95.

Gombojav A., ISHIGURO N., HORIUCHI M., SERJMYADAG D., BYAMBAA B. & SHINAGAWA M. (2003) Amino acid polymorphisms of PrP gene in Mongolian sheep. *Journal of veterinary medical science* **65**, 75-81.

Gombojav A., ISHIGURO N., HORIUCHI M. & SHINAGAWA M. (2004) Unique amino acid polymorphisms of PrP genes in Mongolian sheep breeds. *Journal of veterinary medical science* **66**, 1293-5.

González M.L., Chiapella J.O. & Urdampilleta J.D. (2017) Characterization of some satellite DNA families in Deschampsia antarctica (Poaceae). *Polar Biology*, 1-12.

Goodwin S., McPherson J.D. & McCombie W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333-51.

Gouveia J.G., Wolf I.R., Vilas-Boas L.A., Heslop-Harrison J.S., Schwarzacher T. & Dias A.L. (2017) Repetitive DNA in the catfish genome: rDNA, microsatellites, and Tc1-mariner transposon sequences in Imparfinis species (Siluriformes, Heptapteridae). *Journal of Heredity* **108**, 650-7.

Graur D. & Higgins D.G. (1994) Molecular evidence for the inclusion of cetaceans within the order Artiodactyla. *Molecular Biology and Evolution* **11**, 357-64.

Graves J.A.M., Koina E. & Sankovic N. (2006) How the gene content of human sex chromosomes evolved. *Current opinion in genetics & development* **16**, 219-24.

Grewal S.I. & Jia S. (2007) Heterochromatin revisited. *Nature reviews. Genetics* **8**, 35.

Grishaeva T., Dadashev S.Y. & Bogdanov Y.F. (2005) Identification and Characterization in silico of Meiotic DNA. *Russian Journal of Genetics* **41**, 563-6.

Gunbin K., Peshkin L., Popadin K., Annis S., Ackermann R.R. & Khrapko K. (2017) Integration of mtDNA pseudogenes into the nuclear genome coincides with speciation of the human genus. A hypothesis. *Mitochondrion* **34**, 20-3.

Haaf T. (2000) Fluorescence in situ hybridization. *Encyclopedia of analytical chemistry*.

Halbert N.D. & Derr J.N. (2006) A comprehensive evaluation of cattle introgression into US federal bison herds. *Journal of Heredity* **98**, 1-12.

Haltenorth T. (1963) Klassifikation der lebenden Säugetiere. *Handb. d. Zool. Artiodactyla I*.

Hammodat S.G. (1985) Studied in milk production and growth of suckling lambs in Hamdani sheep. In: *College of Agriculture,*. University of Salah Eldin, Erbil, Iraq. In Arabic.

Han J., Yang M., Yue Y., Guo T., Liu J., Niu C. & Yang B. (2015) Analysis of agouti signaling protein (ASIP) gene polymorphisms and association with coat color in Tibetan sheep (Ovis aries). *Genet Mol Res* **14**, 1200-9.

Han J.S. & Boeke J.D. (2005) LINE-1 retrotransposons: Modulators of quantity and quality of mammalian gene expression? *Bioessays* **27**, 775-84.

Han J.S., Szak S.T. & Boeke J.D. (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429**, 268.

Han Y. & Wessler S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic acids research* **38**, e199-e.

Handel M.A. & Schimenti J.C. (2010) Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nature reviews. Genetics* **11**, 124.

Hanušovská E., Novák M., Arvayová M. & Mikula I. (2003) The PrP genotype of sheep of the improved Valachian breed. *Folia microbiologica* **48**, 269-76.

Harper G., Osuji J.O., Heslop-Harrison J.P. & Hull R. (1999) Integration of banana streak badnavirus into theMusagenome: molecular and cytogenetic evidence. *Virology* **255**, 207-13.

Hart R.P. & Folk W.R. (1982) Structure and organization of a mammalian 5 S gene cluster. *Journal of Biological Chemistry* **257**, 11706-11.

Hartwell L., Leroy Hood , Michael L. Goldberg , Ann E. Reynolds & Silver L.M. (2011) *Genetics, from genes to genomes*, New York: McGraw-Hill.

Hassanin A., Bonillo C., Nguyen B.X. & Cruaud C. (2010) Comparisons between mitochondrial genomes of domestic goat (Capra hircus) reveal the presence of numts and multiple sequencing errors. *Mitochondrial DNA* **21**, 68-76.

Hayashi Y., Kajikawa M., Matsumoto T. & Okada N. (2014) Mechanism by which a LINE protein recognizes its 3' tail RNA. *Nucleic acids research* **42**, 10605-17.

Hayes H., Petit E. & Dutrillaux B. (1991) Comparison of RBG-banded karyotypes of cattle, sheep, and goats. *Cytogenetic and Genome Research* **57**, 51-5.

Hazkani-Covo E. & Graur D. (2006) A comparative analysis of numt evolution in human and chimpanzee. *Molecular Biology and Evolution* **24**, 13-8.

Heaton M.P., Leymaster K.A., Freking B.A., Hawk D.A., Smith T.P., Keele J.W., Snelling W.M., Fox J.M., Chitko-McKown C.G. & Laegreid W.W. (2003) Prion gene sequence variation within diverse groups of US sheep, beef cattle, and deer. *Mammalian Genome* **14**, 765-77.

Hecht S.J., Carlson J.O. & Demartini J.C. (1994) Analysis of a type D retroviral capsid gene expressed in ovine pulmonary carcinoma and present in both affected and unaffected sheep genomes. *Virology* **202**, 480-4.

Hecht S.J., Stedman K.E., Carlson J.O. & DeMartini J.C. (1996) Distribution of endogenous type B and type D sheep retrovirus sequences in ungulates and other mammals. *Proceedings of the National Academy of Sciences* **93**, 3297-302.

Hedges D.J., Guettouche T., Yang S., Bademci G., Diaz A., Andersen A., Hulme W.F., Linker S., Mehta A. & Edwards Y.J. (2011) Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PloS one* **6**, e18595.

Henras A.K., Plisson-Chastang C., O'Donohue M.F., Chakraborty A. & Gleizes P.E. (2015) An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdisciplinary Reviews: RNA* **6**, 225-42.

Hernández-Hernández A., Rincón-Arano H., Recillas-Targa F., Ortiz R., Valdes-Quezada C., Echeverría O.M., Benavente R. & Vázquez-Nin G.H. (2008) Differential distribution and association of repeat DNA sequences in the lateral element of the synaptonemal complex in rat spermatocytes. *Chromosoma* **117**, 77-87.

Hernández-Hernández A., Vázquez-Nin G., Echeverría O. & Recillas-Targa F. (2009) Chromatin structure contribution to the synaptonemal complex formation. *Cellular and Molecular Life Sciences* **66**, 1198-208.

Herron P.R., Hughes G., Chandra G., Fielding S. & Dyson P.J. (2004) Transposon Express, a software application to report the identity of insertions obtained by comprehensive transposon mutagenesis of sequenced genomes: analysis of the preference for in vitro Tn 5 transposition into GC-rich DNA. *Nucleic acids research* **32**, e113-e.

Heslop-Harrison J. & Schwarzacher T. (2011) Organisation of the plant genome in chromosomes. *The Plant Journal* **66**, 18-33.

Hiendleder S., Kaupe B., Wassmuth R. & Janke A. (2002) Molecular analysis of wild and domestic sheep questions current nomenclature and provides evidence for

domestication from two different subspecies. *Proceedings of the Royal Society of London B: Biological Sciences* **269**, 893-904.

Hiendleder S., Lewalski H. & Janke A. (2008) Complete mitochondrial genomes of Bos taurus and Bos indicus provide new insights into intra-species variation, taxonomy and domestication. *Cytogenetic and Genome Research* **120**, 150-6.

Hiendleder S., Lewalski H., Wassmuth R. & Janke A. (1998a) The complete mitochondrial DNA sequence of the domestic sheep (Ovis aries) and comparison with the other major ovine haplotype. *Journal of Molecular Evolution* **47**, 441-8.

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

Hillier L.W., Miller W., Birney E., Warren W., Hardison R.C., Ponting C.P., Bork P., Burt D.W., Groenen M.A. & Delany M.E. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716.

Hirsch C.D. & Springer N.M. (2017) Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1860**, 157-65.

Holko I., Novackova A., Holkova T. & Kmet V. (2005) PrP genotyping of sheep breeds in Slovakia. *Veterinary Record* **157**, 628.

Huelsenbeck J.P. & Ronquist F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-5.

Hui P. (2012) Next generation sequencing: chemistry, technology and applications. In: *Chemical Diagnostics* (pp. 1-18. Springer.

Hunt P.A. & Hassold T.J. (2002) Sex matters in meiosis. *Science* **296**, 2181-3.

Hunter N. (1996) Prion protein (prnp) genotypes and natural scrapie in closed flocks of Cheviot and Suffolk sheep in Britain. In: *Court L, Dodet B, editors. Transmissible subacute spongiform encephalopathies: prion diseases, Paris: Elsevier*, pp. 47-50.

Hunter N. (1997) PrP genetics in sheep and the implications for scrapie and BSE. *Trends in microbiology* **5**, 331-4.

Hunter N. (2007) Scrapie—Uncertainties, biology and molecular approaches. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1772**, 619-28.

Hunter N. & Cairns D. (1998) Scrapie-free Merino and Poll Dorset sheep from Australia and New Zealand have normal frequencies of scrapie-susceptible PrP genotypes. *Journal of general virology* **79**, 2079-82.

Hunter N., Goldmann W., Smith G. & Hope J. (1994) The association of a codon 136 PrP gene variant with the occurrence of natural scrapie. *Archives of virology* **137**, 171-7.

Hussain A., Babar M.E., Imran M., Haq I.U. & Javed M.M. (2011) Detection of four novel polymorphisms in PrP gene of Pakistani sheep (Damani and Hashtnagri) and goats (Kamori and Local Hairy) breeds. *Virology journal* **8**, 246.

Iannuzzi L. (1996) G-and R-banded prometaphase karyotypes in cattle (Bos taurus L.). *Chromosome Research* **4**, 448-56.

Iannuzzi L., Di Meo G., Perucatti A. & Ferrara L. (1990) A comparison of G-and R-banding in cattle and river buffalo prometaphase chromosomes. *Caryologia* **43**, 283-90.

Iannuzzi L., Di Meo G., Perucatti A., Schibler L., Incarnato D. & Cribiu E. (2001) Comparative FISH mapping in river buffalo and sheep chromosomes: assignment of forty autosomal type I loci from sixteen human chromosomes. *Cytogenetic and Genome Research* **94**, 43-8.

Iannuzzi L. & Di Meo G.P. (1995) Chromosomal evolution in bovids: a comparison of cattle, sheep and goat G-and R-banded chromosomes and cytogenetic divergences among cattle, goat and river buffalo sex chromosomes. *Chromosome Research* **3**, 291-9.

Iannuzzi L., King W. & Di Berardino D. (2009) Chromosome evolution in domestic bovids as revealed by chromosome banding and FISH-mapping techniques. *Cytogenetic and Genome Research* **126**, 49-62.

Ichiyanagi K. & Okada N. (2008) Mobility pathways for vertebrate L1, L2, CR1, and RTE clade retrotransposons. *Molecular Biology and Evolution* **25**, 1148-57.

Ikeda T., Horiuchi M., Ishiguro N., Muramatsu Y., Kai-Uwe G.D. & Shinagawa M. (1995) Amino acid polymorphisms of PrP with reference to onset of scrapie in Suffolk and Corriedale sheep in Japan. *Journal of general virology* **76**, 2577-81.

Iñiguez L. (2005) *Characterization of small ruminant breeds in West Asia and North Africa*.

Irwin D.M., Kocher T.D. & Wilson A.C. (1991) Evolution of the cytochrome b gene of mammals. *Journal of Molecular Evolution* **32**, 128-44.

Irwin J., Just R., Scheible M. & Loreille O. (2011) Assessing the potential of next generation sequencing technologies for missing persons identification efforts. *Forensic Science International: Genetics Supplement Series* **3**, e447-e8.

Janicki M., Rooke R. & Yang G. (2011) Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. *Chromosome Research* **19**, 787.

Jantsch M., Hamilton B., Mayr B. & Schweizer D. (1990) Meiotic chromosome behaviour reflects levels of sequence divergence inSus scrofa domestica satellite DNA. *Chromosoma* **99**, 330-5.

Jeffrey M., Martin S., Chianini F., Eaton S., Dagleish M.P. & Gonzalez L. (2014) Incidence of infection in Prnp ARR/ARR sheep following experimental inoculation with or natural exposure to classical scrapie. *PloS one* **9**, e91026.

Jern P. & Coffin J.M. (2008) Effects of retroviruses on host genome function. *Annual review of genetics* **42**, 709-32.

Jiang S.-Y. & Ramachandran S. (2013) Genome-wide survey and comparative analysis of LTR retrotransposons and their captured genes in rice and sorghum. *PloS one* **8**, e71118.

Jobse C., Buntjer J.B., Haagsma N., Breukelman H.J., Beintema J.J. & Lenstral J.A. (1995) Evolution and recombination of bovine DNA repeats. *Journal of Molecular Evolution* **41**, 277-83.

John H., Birnstiel M. & Jones K. (1969) RNA-DNA hybrids at the cytological level. *Nature* **223**, 582-7.

Johnson M.E., Rowsey R.A., Shirley S., VandeVoort C., Bailey J. & Hassold T. (2013) A specific family of interspersed repeats (SINEs) facilitates meiotic synapsis in mammals. *Molecular cytogenetics* **6**, 1.

Johnson W.E. & Coffin J.M. (1999) Constructing primate phylogenies from ancient retrovirus sequences. *Proceedings of the National Academy of Sciences* **96**, 10254-60.

Jordan I.K., Rogozin I.B., Glazko G.V. & Koonin E.V. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *TRENDS in Genetics* **19**, 68-72.

Juma K. & Alkass J. (2000) Sheep in Iraq.

Jurka J., Kapitonov V.V., Kohany O. & Jurka M.V. (2007) Repetitive sequences in complex genomes: structure and evolution. *Annu. Rev. Genomics Hum. Genet.* **8**, 241-59.

Jurka J., Kapitonov V.V., Pavlicek A., Klonowski P., Kohany O. & Walichiewicz J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**, 462-7.

Kajikawa M. & Okada N. (2002) LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell* **111**, 433-44.

Kalyanaraman A. & Aluru S. (2006) Efficient algorithms and software for detection of full-length LTR retrotransposons. *Journal of bioinformatics and computational biology* **4**, 197-216.

Kapitonov V.V. & Jurka J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics* **9**, 411-2.

Kapitonov V.V., Tempel S. & Jurka J. (2009) Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* **448**, 207-13.

Karam H., Markotie B. & Al-Maali a.H.N.A. (1976) Iraqi breeds of sheep. (ed. by 12 TRN), Development of Livestock production in Northern Iraq.

Karami M., Amirinia C., Kashan N.E., Amirmozafari N., Chamani M. & Banabazi M.H. (2011) Polymorphisms of the prion protein gene Arabi sheep breed in Iran. *African Journal of Biotechnology* **10**, 15819-22.

Karpova O., Penkina M., Dadashev S., Mil'shina N., Hernandes J., Radchenko I. & Bogdanov I. (1994) Features of the primary structure of DNA from the

synaptonemal complex of the golden hamster. *Molekuliarnaia biologiia* **29**, 512-21.

Karpova O., Safronov V., Zaitseva S. & Bogdanov Y.F. (1989) Some properties of DNA isolated from mouse synaptonemal complex fraction. *MOLECULAR BIOLOGY* **23**, 448-56.

Kattanovskaya H. & Serov O. (1994) High-resolution GTG-banded chromosomes of cattle, sheep, and goat: a comparative study. *Journal of Heredity* **85**, 395-400.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S. & Duran C. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-9.

Kent W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome research* **12**, 656-64.

Kidwell M.G. & Lisch D. (1997) Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences* **94**, 7704-11.

Kimura B., Marshall F.B., Chen S., Rosenbom S., Moehlman P.D., Tuross N., Sabin R.C., Peters J., Barich B. & Yohannes H. (2010) Ancient DNA from Nubian and Somali wild ass provides insights into donkey ancestry and domestication. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20100708.

Kipling D., Ackford H.E., Taylor B.A. & Cooke H.J. (1991) Mouse minor satellite DNA genetically maps to the centromere and is physically linked to the proximal telomere. *Genomics* **11**, 235-41.

Klymiuk N., Müller M., Brem G. & Aigner B. (2003) Characterization of endogenous retroviruses in sheep. *Journal of virology* **77**, 11268-73.

Kolas N., Yuan L., Hoog C., Heng H., Marcon E. & Moens P. (2004) Male mouse meiotic chromosome cores deficient in structural proteins SYCP3 and SYCP2 align by homology but fail to synapse and have possible impaired specificity of chromatin loop attachment. *Cytogenetic and Genome Research* **105**, 182-8.

Kondo Y. & Issa J.-P.J. (2003) Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells. *Journal of Biological Chemistry* **278**, 27658-62.

Kopecka H., MACAYA G., CORTADAS J., THIÉRY J.P. & BERNARDI G. (1978) Restriction enzyme analysis of satellite DNA components from the bovine genome. *The FEBS Journal* **84**, 189-95.

Kopecna O., Kubickova S., Cernohorska H., Cabelova K., Vahala J., Martinkova N. & Rubes J. (2014) Tribe-specific satellite DNA in non-domestic Bovidae. *Chromosome Research* **22**, 277-91.

Kopecna O., Kubickova S., Cernohorska H., Cabelova K., Vahala J. & Rubes J. (2012) Isolation and comparison of tribe-specific centromeric repeats within Bovidae. *Journal of applied genetics* **53**, 193-202.

Kordis D. & Gubensek F. (1998) Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes. *Proceedings of the National Academy of Sciences* **95**, 10704-9.

Kordiš D. & Gubenšek F. (1999) Molecular evolution of Bov-B LINEs in vertebrates. *Gene* **238**, 171-8.

Korenberg J.R. & Rykowski M.C. (1988) Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* **53**, 391-400.

Kramerov D. & Vassetzky N. (2011) Origin and evolution of SINEs in eukaryotic genomes. *Heredity* **107**, 487.

Krassovsky K. & Henikoff S. (2014) Distinct chromatin features characterize different classes of repeat sequences in Drosophila melanogaster. *BMC genomics* **15**, 105.

Kriegs J.O., Churakov G., Jurka J., Brosius J. & Schmitz J. (2007) Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *TRENDS in Genetics* **23**, 158-61.

Kuhn G.C., Küttler H., Moreira-Filho O. & Heslop-Harrison J.S. (2011) The 1.688 repetitive DNA of Drosophila: concerted evolution at different genomic scales and association with genes. *Molecular Biology and Evolution* **29**, 7-11.

Kuhn G.C., Schwarzacher T. & Heslop-Harrison J.S. (2010) The non-regular orbit: three satellite DNAs in Drosophila martensis (buzzatii complex, repleta group) followed three different evolutionary pathways. *Molecular genetics and genomics* **284**, 251-62.

Kuhn G.C., Sene F.M., Moreira-Filho O., Schwarzacher T. & Heslop-Harrison J.S. (2008) Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the Drosophila buzzatii cluster. *Chromosome Research* **16**, 307-24.

Kumar A. & Bennetzen J.L. (1999) Plant retrotransposons. *Annual review of genetics* **33**, 479-532.

Kurnit D.M., Shafit B.R. & Maio J.J. (1973) Multiple satellite deoxyribonucleic acids in the calf and their relation to the sex chromosomes. *Journal of molecular biology* **81**, 273IN1279-278IN2284.

Kurth R. & Bannert N. (2010) Beneficial and detrimental effects of human endogenous retroviruses. *International journal of cancer* **126**, 306-14.

Kurtz S., Choudhuri J.V., Ohlebusch E., Schleiermacher C., Stoye J. & Giegerich R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic acids research* **29**, 4633-42.

Kurtz S., Narechania A., Stein J.C. & Ware D. (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC genomics* **9**, 517.

Kvist L., Martens J., Nazarenko A.A. & Orell M. (2003) Paternal leakage of mitochondrial DNA in the great tit (Parus major). *Molecular Biology and Evolution* **20**, 243-7.

Lan Z., Wang Z., Liu Y. & Zhang X. (2006) Prion protein gene (PRNP) polymorphisms in Xinjiang local sheep breeds in China. *Archives of virology* **151**, 2095-101.

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M. & FitzHugh W. (2001) Initial sequencing and analysis of the human genome.

Laplanche J., Chatelain J., Westaway D., Thomas S., Dussaucy M., Brugere-Picoux J. & Launay J. (1993) PrP polymorphisms associated with natural scrapie discovered by denaturing gradient gel electrophoresis. *Genomics* **15**, 30-7.

Lee B. & Amon A. (2001) Meiosis: how to create a specialized cell cycle. *Current opinion in cell biology* **13**, 770-7.

Lee C., Cho C., Haslett J.L. & Lin C.-C. (1997) Higher-order organization of subrepeats and the evolution of cervid satellite I DNA. *Journal of Molecular Evolution* **44**, 327-35.

Lee C., Ritchie D. & Lin C. (1994) A tandemly repetitive, centromeric DNA sequence from the Canadian woodland caribou (Rangifer tarandus caribou): its conservation and evolution in several deer species. *Chromosome Research* **2**, 293-306.

Lee I.Y., Westaway D., Smit A.F., Wang K., Seto J., Chen L., Acharya C., Ankener M., Baskin D. & Cooper C. (1998) Complete genomic sequence and analysis of the prion protein gene region from three mammalian species. *Genome research* **8**, 1022-37.

Lenstra J., Boxtel J., Zwaagstra K. & Schwerin M. (1993) Short interspersed nuclear element (SINE) sequences of the Bovidae. *Animal genetics* **24**, 33-9.

Lerat E. (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* **104**, 520.

Levene M.J., Korlach J., Turner S.W., Foquet M., Craighead H.G. & Webb W.W. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682-6.

Levsky J.M. & Singer R.H. (2003) Fluorescence in situ hybridization: past, present and future. *Journal of cell science* **116**, 2833-8.

Li J.-H., Liu S., Zhou H., Qu L.-H. & Yang J.-H. (2013) starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic acids research* **42**, D92-D7.

Li R., Ye J., Li S., Wang J., Han Y., Ye C., Wang J., Yang H., Yu J. & Wong G.K.-S. (2005) ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS computational biology* **1**, e43.

Li S.-F., Su T., Cheng G.-Q., Wang B.-X., Li X., Deng C.-L. & Gao W.-J. (2017) Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. *Genes* **8**, 290.

Li X., Kahveci T. & Settles A.M. (2007) A novel genome-scale repeat finder geared towards transposons. *Bioinformatics* **24**, 468-76.

Li Y.-C., Lee C., Hsu T.-H., Li S.-Y. & Lin C. (2000a) Direct visualization of the genomic distribution and organization of two cervid centromeric satellite DNA families. *Cytogenetic and Genome Research* **89**, 192-8.

Li Y.-C., Lee C., Sanoudou D., Hsu T.-H., Li S.-Y. & Lin C.-C. (2000b) Interstitial colocalization of two cervid satellite DNAs involved in the genesis of the Indian muntjac karyotype. *Chromosome Research* **8**, 363-73.

Li Y., Lee C., Chang W., Li S.-Y. & Lin C. (2002) Isolation and identification of a novel satellite DNA family highly conserved in several Cervidae species. *Chromosoma* **111**, 176-83.

Lindblad-Toh K., Wade C.M., Mikkelsen T.S. & Karlsson E.K. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803.

Lisch D. (2013) How important are transposons for plant evolution? *Nature reviews. Genetics* **14**, 49.

Loftus R.T., MacHugh D.E., Bradley D.G., Sharp P.M. & Cunningham P. (1994) Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Sciences* **91**, 2757-61.

Lomiento M., Jiang Z., D'Addabbo P., Eichler E.E. & Rocchi M. (2008) Evolutionary-new centromeres preferentially emerge within gene deserts. *Genome biology* **9**, R173.

López-Flores I. & Garrido-Ramos M. (2012) The repetitive DNA content of eukaryotic genomes. In: *Repetitive DNA* (pp. 1-28. Karger Publishers.

Lorenzi H., Thiagarajan M., Haas B., Wortman J., Hall N. & Caler E. (2008) Genome wide survey, discovery and evolution of repetitive elements in three Entamoeba species. *BMC genomics* **9**, 595.

Löwer R. (1999) The pathogenic potential of endogenous retroviruses: facts and fantasies. *Trends in microbiology* **7**, 350-6.

Lv F.-H., Peng W.-F., Yang J., Zhao Y.-X., Li W.-R., Liu M.-J., Ma Y.-H., Zhao Q.-J., Yang G.-L. & Wang F. (2015) Mitogenomic meta-analysis identifies two phases of migration in the history of eastern Eurasian sheep. *Molecular Biology and Evolution* **32**, 2515-33.

Lyon M.F. (1998) X-chromosome inactivation: a repeat hypothesis. *Cytogenetic and Genome Research* **80**, 133-7.

Maarof N.N., Chakmakchey A.M. & Moustafa K.S. (1982) Preliminary Studies on the wool quality traits of Hamdani ewes. In: *In proceedings, 6th International Conference on Animal and Poultry production*, pp. Page 51-5. , Zagazig University, Zagazig, Egypt.

Maarof N.N., Juma K.H., Arafat E.A. & and Chakmakchey A.M. (1986) Evaluation of factors affecting birth and weaning weights and milk production in Hamdani sheep. *World Review of Animal Production 22:51-55.*

Macas J., Kejnovský E., Neumann P., Novák P., Koblížková A. & Vyskot B. (2011) Correction: Next Generation Sequencing-Based Analysis of Repetitive DNA in the Model Dioceous Plant Silene latifolia. *PloS one* **6**, 10.1371/annotation/4ccaacb2-92d7-445a-87da-313cedf18feb.

Macas J., Koblížková A., Navrátilová A. & Neumann P. (2009) Hypervariable 3' UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. *Gene* **448**, 198-206.

Macas J., Neumann P., Novák P. & Jiang J. (2010) Global sequence characterization of rice centromeric satellite based on oligomer frequency analysis in large-scale sequencing data. *Bioinformatics* **26**, 2101-8.

Macaya G., CORTADAS J. & BERNARDI G. (1978) An Analysis of the Bovine Genome by Density-Gradient Centrifugation. *The FEBS Journal* **84**, 179-88.

Maddox J.F. (2005) A presentation of the differences between the sheep and goat genetic maps. *Genetics Selection Evolution* **37**, S1.

Maeda N., Palmarini M., Murgia C. & Fan H. (2001) Direct transformation of rodent fibroblasts by jaagsiekte sheep retrovirus DNA. *Proceedings of the National Academy of Sciences* **98**, 4449-54.

Malik H.S., Burke W.D. & Eickbush T.H. (1999) The age and evolution of non-LTR retrotransposable elements. *Molecular Biology and Evolution* **16**, 793-805.

Malik H.S. & Eickbush T.H. (1998) The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. *Molecular Biology and Evolution* **15**, 1123-34.

Marçais G. & Kingsford C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-70.

Mardis E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387-402.

Mariotti M., Valentini A., Marsan P.A. & Pariset L. (2013) Mitochondrial DNA of seven Italian sheep breeds shows faint signatures of domestication and suggests recent breed formation. *Mitochondrial DNA* **24**, 577-83.

Martens J.H., O'Sullivan R.J., Braunschweig U., Opravil S., Radolf M., Steinlein P. & Jenuwein T. (2005) The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *The EMBO journal* **24**, 800-12.

Matthee C.A. & Davis S.K. (2001) Molecular insights into the evolution of the family Bovidae: a nuclear DNA perspective. *Molecular Biology and Evolution* **18**, 1220-30.

Mattick J.S. & Makunin I.V. (2006) Non-coding RNA. *Human molecular genetics* **15**, R17-R29.

Maxam A.M. & Gilbert W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences* **74**, 560-4.

McCarthy E.M. & McDonald J.F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**, 362-7.

McGee-Estrada K., Palmarini M. & Fan H. (2002) HNF-3β is a critical factor for the expression of the Jaagsiekte sheep retrovirus long terminal repeat in type II pneumocytes but not in Clara cells. *Virology* **292**, 87-97.

Meadows J., Hiendleder S. & Kijas J. (2011) Haplogroup relationships between domestic and wild sheep resolved using a mitogenome panel. *Heredity* **106**, 700.

Meadows J.R., Cemal I., Karaca O., Gootwine E. & Kijas J.W. (2007) Five ovine mitochondrial lineages identified from sheep breeds of the near East. *Genetics* **175**, 1371-9.

Meneveri R., Agresti A., Marozzi A., Saccone S., Rocchi M., Archidiacono N., Corneo G., Valle G.D. & Ginelli E. (1993) Molecular organization and chromosomal location of human GC-rich heterochromatic blocks. *Gene* **123**, 227-34.

Mensher S., Bunch T. & Maciulis A. (1989) High-Resolution G-Banded Karyotype and Idiogram of the Goat: A Sheep–Goat G-Banded Comparison. *Journal of Heredity* **80**, 150-5.

Meštrović N., Mravinac B., Pavlek M., Vojvoda-Zeljko T., Šatović E. & Plohl M. (2015) Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosome Research* **23**, 583-96.

Metzker M.L. (2010) Sequencing technologies--the next generation. *Nature reviews. Genetics* **11**, 31.

Meuwissen R., Offenberg H.H., Dietrich A., Riesewijk A., van Iersel M. & Heyting C. (1992) A coiled-coil related protein specific for synapsed regions of meiotic prophase chromosomes. *The EMBO journal* **11**, 5091.

Michaud E.J., Van Vugt M.J., Bultman S.J., Sweet H.O., Davisson M.T. & Woychik R.P. (1994) Differential expression of a new dominant agouti allele (Aiapy) is correlated with methylation state and is influenced by parental lineage. *Genes & development* **8**, 1463-72.

Mikkelsen T.S., Hillier L.W., Eichler E.E. & Zody M.C. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69.

Mikkelsen T.S., Wakefield M.J., Aken B., Amemiya C.T., Chang J.L., Duke S., Garber M., Gentles A.J., Goodstadt L. & Heger A. (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**, 167.

Miller J.M., Moore S.S., Stothard P., Liao X. & Coltman D.W. (2015) Harnessing cross-species alignment to discover SNPs and generate a draft genome sequence of a bighorn sheep (Ovis canadensis). *BMC genomics* **16**, 397.

Miraldo A., Hewitt G.M., Dear P.H., Paulo O.S. & Emerson B.C. (2012) Numts help to reconstruct the demographic history of the ocellated lizard (Lacerta lepida) in a secondary contact zone. *Molecular ecology* **21**, 1005-18.

Modi W.S., Gallagher D.S. & Womack J.E. (1996) Evolutionary histories of highly repeated DNA families among the Artiodactyla (Mammalia). *Journal of Molecular Evolution* **42**, 337-49.

Moens P.B. & Pearlman R.E. (1989) Satellite DNA I in chromatin loops of rat pachytene chromosomes and in spermatids. *Chromosoma* **98**, 287-94.

Moens P.B. & Pearlman R.E. (1990) Telomere and centromere DNA are associated with the cores of meiotic prophase chromosomes. *Chromosoma* **100**, 8-14.

Mohammed A. (2009) Genetic Diversity in Some Iraqi Sheep Breeds Using Molecular Techniques In: *Council of the College of Education*, University of Duhok.

Mokry M., Hatzis P., De Bruijn E., Koster J., Versteeg R., Schuijers J., van de Wetering M., Guryev V., Clevers H. & Cuppen E. (2010) Efficient double fragmentation ChIP-seq provides nucleotide resolution protein-DNA binding profiles. *PloS one* **5**, e15092.

Morey M., Fernández-Marmiesse A., Castiñeiras D., Fraga J.M., Couce M.L. & Cocho J.A. (2013) A glimpse into past, present, and future DNA sequencing. *Molecular genetics and metabolism* **110**, 3-24.

Mouse Genome Sequencing C. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520.

Murcia P.R., Arnaud F. & Palmarini M. (2007) The transdominant endogenous retrovirus enJS56A1 associates with and blocks intracellular trafficking of Jaagsiekte sheep retrovirus Gag. *Journal of virology* **81**, 1762-72.

Nadler C., Hoffmann R. & Woolf A. (1973) G-band patterns as chromosomal markers, and the interpretation of chromosomal evolution in wild sheep (Ovis). *Cellular and Molecular Life Sciences* **29**, 117-9.

Nadler C., Lay D. & Hassinger J. (1971) Cytogenetic analyses of wild sheep populations in northern Iran. *Cytogenetic and Genome Research* **10**, 137-52.

Nieddu M., Mezzanotte R., Pichiri G., Coni P.P., Dedola G.L., Dettori M.L., Pazzola M., Vacca G.M. & Robledo R. (2015) Evolution of satellite DNA sequences in two tribes of Bovidae: A cautionary tale. *Genetics and molecular biology* **38**, 513-8.

Nijman I.J. & Lenstra J.A. (2001) Mutation and recombination in cattle satellite DNA: a feedback model for the evolution of satellite DNA repeats. *Journal of Molecular Evolution* **52**, 361-71.

Nijman I.J., van Tessel P. & Lenstra J.A. (2002) SINE retrotransposition during the evolution of the Pecoran ruminants. *Journal of Molecular Evolution* **54**, 9-16.

Nikaido M., Matsuno F., Hamilton H., Brownell R.L., Cao Y., Ding W., Zuoyan Z., Shedlock A.M., Fordyce R.E. & Hasegawa M. (2001) Retroposon analysis of major cetacean lineages: the monophyly of toothed whales and the paraphyly of river dolphins. *Proceedings of the National Academy of Sciences* **98**, 7384-9.

Nikaido M., Rooney A.P. & Okada N. (1999) Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales. *Proceedings of the National Academy of Sciences* **96**, 10261-6.

Nilsson M., Klassert D., Bertelsen M., Hallström B. & Janke A. (2012) Activity of ancient RTE retroposons during the evolution of cows, spiral-horned antelopes, and Nilgais (Bovinae). *Molecular Biology and Evolution* **29**, 2885-8.

Nin G.H.V., Flores E., Echeverría O.M., Merkert H., Wettstein R. & Benavente R. (1993) Immunocytochemical localization of DNA in synaptonemal complexes of rat and mouse spermatocytes, and of chick oocytes. *Chromosoma* **102**, 457-63.

Nishihara H., Plazzi F., Passamonti M. & Okada N. (2016) MetaSINEs: broad distribution of a novel SINE superfamily in animals. *Genome biology and evolution* **8**, 528-39.

Nomura O., Lin Z.-H., Wada Y. & Yasue H. (1998) A SINE species from hippopotamus and its distribution among animal species. *Mammalian Genome* **9**, 550-5.

Noreen F., Akbergenov R., Hohn T. & Richert-Pöggeler K.R. (2007) Distinct expression of endogenous Petunia vein clearing virus and the DNA transposon dTph1 in two Petunia hybrida lines is correlated with differences in histone modification and siRNA production. *The Plant Journal* **50**, 219-29.

Norris B.J. & Whan V.A. (2008) A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. *Genome research* **18**, 1282-93.

Novák P., Ávila Robledillo L., Koblížková A., Vrbová I., Neumann P. & Macas J. (2017) TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic acids research*.

Novák P., Neumann P. & Macas J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC bioinformatics* **11**, 378.

Novák P., Neumann P., Pech J., Steinhaisl J. & Macas J. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792-3.

Nowak R.M. (1999) Order artiodactyla. *Walker's mammals of the world. John Hopkins University Press, London*.

O'connor C. (2008) Fluorescence in situ hybridization (FISH). *Nature Education* **1**, 171.

O'Rourke K.I., Holyoak G., Clark W., Mickelson J., Wang S., Melco R., Besser T. & Foote W. (1997) PrP genotypes and experimental scrapie in orally inoculated Suffolk sheep in the United States. *Journal of general virology* **78**, 975-8.

Offenberg H.H., Schalk J.A., Meuwissen R.L., van Aalderen M., Kester H.A., Dietrich A.J. & Heyting C. (1998) SCP2: a major protein component of the axial elements of synaptonemal complexes of the rat. *Nucleic acids research* **26**, 2572-9.

Ohshima K. & Okada N. (2005) SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenetic and Genome Research* **110**, 475-90.

Okada N. (1991a) SINEs. *Current opinion in genetics & development* **1**, 498-504.

Okada N. (1991b) SINEs: short interspersed repeated elements of the eukaryotic genome. *Trends in ecology & evolution* **6**, 358-61.

Ortiz R., Echeverría O., Ubaldo E., Carlos A., Scassellati C. & Vázquez-Nin G. (2002) Cytochemical study of the distribution of RNA and DNA in the synaptonemal complex of guinea-pig and rat spermatocytes. *European journal of histochemistry: EJH* **46**, 133.

Paces J., Pavlícek A. & Paces V. (2002) HERVd: database of human endogenous retroviruses. *Nucleic acids research* **30**, 205-6.

Pagán H.J., Macas J., Novák P., McCulloch E.S., Stevens R.D. & Ray D.A. (2012) Survey sequencing reveals elevated DNA transposon activity, novel elements, and

variation in repetitive landscapes among vesper bats. *Genome biology and evolution* **4**, 575-85.

Page S.L. & Hawley R.S. (2004) The genetics and molecular biology of the synaptonemal complex. *Annu. Rev. Cell Dev. Biol.* **20**, 525-58.

Palmarini M., Cousens C., Dalziel R., Bai J., Stedman K., DeMartini J. & Sharp J. (1996) The exogenous form of Jaagsiekte retrovirus is specifically associated with a contagious lung cancer of sheep. *Journal of virology* **70**, 1618-23.

Palmarini M., Datta S., Omid R., Murgia C. & Fan H. (2000a) The long terminal repeat of Jaagsiekte sheep retrovirus is preferentially active in differentiated epithelial cells of the lungs. *Journal of virology* **74**, 5776-87.

Palmarini M. & Fan H. (2001) Retrovirus-induced ovine pulmonary adenocarcinoma, an animal model for lung cancer. *Journal of the National Cancer Institute* **93**, 1603-14.

Palmarini M., Gray C.A., Carpenter K., Fan H., Bazer F.W. & Spencer T.E. (2001) Expression of endogenous betaretroviruses in the ovine uterus: effects of neonatal age, estrous cycle, pregnancy, and progesterone. *Journal of virology* **75**, 11319-27.

Palmarini M., Hallwirth C., York D., Murgia C., De Oliveira T., Spencer T. & Fan H. (2000b) Molecular cloning and functional analysis of three type D endogenous retroviruses of sheep reveal a different cell tropism from that of the highly related exogenous jaagsiekte sheep retrovirus. *Journal of virology* **74**, 8065-76.

Palmarini M., Mura M. & Spencer T.E. (2004) Endogenous betaretroviruses of sheep: teaching new lessons in retroviral interference and adaptation. *Journal of general virology* **85**, 1-13.

Palmarini M., Sharp J.M., De Las Heras M. & Fan H. (1999) Jaagsiekte sheep retrovirus is necessary and sufficient to induce a contagious lung cancer in sheep. *Journal of virology* **73**, 6964-72.

Parr R.L., Maki J., Reguly B., Dakubo G.D., Aguirre A., Wittock R., Robinson K., Jakupciak J.P. & Thayer R.E. (2006) The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation. *BMC genomics* **7**, 185.

Paterson A.H., Bowers J.E., Bruggmann R., Grimwood J., Gundlach H., Haberer G., Hellsten U., Mitros T., Poliakov A. & Schmutz J. (2008) The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**.

Pathak D. & Ali S. (2012) Repetitive DNA: A tool to explore animal genomes/transcriptomes. In: *Functional genomics* (Intech.

Pathak S., Van Tuinen P. & Merry D. (1982) Heterochromatin, synaptonemal complex, and NOR activity in the somatic and germ cells of a male domestic dog, Canis familiaris (Mammalia, Canidae). *Cytogenetic and Genome Research* **34**, 112-8.

Patience C., Takeuchi Y. & Weiss R.A. (1997) Infection of human cells by an endogenous retrovirus of pigs. *Nature medicine* **3**, 282-6.

Pearlman R.E., Tsao N. & Moens P.B. (1992) Synaptonemal complexes from DNase-treated rat pachytene chromosomes contain (GT) n and LINE/SINE sequences. *Genetics* **130**, 865-72.

Pedrosa S., Uzun M., Arranz J.-J., Gutiérrez-Gil B., San Primitivo F. & Bayón Y. (2005) Evidence of three maternal lineages in Near Eastern sheep supporting multiple domestication events. *Proceedings of the Royal Society of London B: Biological Sciences* **272**, 2211-7.

Pelttari J., Hoja M.-R., Yuan L., Liu J.-G., Brundell E., Moens P., Santucci-Darmanin S., Jessberger R., Barbero J.L. & Heyting C. (2001) A meiotic chromosomal core consisting of cohesin complex proteins recruits DNA recombination proteins and promotes synapsis in the absence of an axial element in mammalian meiotic cells. *Molecular and Cellular Biology* **21**, 5667-77.

Peng J.C. & Karpen G.H. (2007) H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. *Nature cell biology* **9**, 25-35.

Penzkofer T., Jäger M., Figlerowicz M., Badge R., Mundlos S., Robinson P.N. & Zemojtel T. (2017) L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic acids research* **45**, D68-D73.

Pertile M.D., Graham A.N., Choo K.A. & Kalitsis P. (2009) Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. *Genome research* **19**, 2202-13.

Pettersson E., Lundeberg J. & Ahmadian A. (2009) Generations of sequencing technologies. *Genomics* **93**, 105-11.

Phillips C.M., Meng X., Zhang L., Chretien J.H., Urnov F.D. & Dernburg A.F. (2009) Identification of chromosome sequence motifs that mediate meiotic pairing and synapsis in C. elegans. *Nature cell biology* **11**, 934.

Piégu B., Bire S., Arensburger P. & Bigot Y. (2015) A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. *Molecular phylogenetics and evolution* **86**, 90-109.

Płucienniczak A., Skowroński J. & Jaworski J. (1982) Nucleotide sequence of bovine 1.715 satellite DNA and its relation to other bovine satellite sequences. *Journal of molecular biology* **158**, 293-304.

Plohl M., Luchetti A., Meštrović N. & Mantovani B. (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero) chromatin. *Gene* **409**, 72-82.

Plohl M., Meštrović N. & Mravinac B. (2012) Satellite DNA evolution. In: *Repetitive DNA* (pp. 126-52. Karger Publishers.

Price A.L., Jones N.C. & Pevzner P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351-i8.

Prusiner S.B. (1991) Molecular biology of prion diseases. *Science* **252**, 1515-22.

Pyle A., Hudson G., Wilson I.J., Coxhead J., Smertenko T., Herbert M., Santibanez-Koref M. & Chinnery P.F. (2015) Extreme-depth re-sequencing of mitochondrial DNA

finds no evidence of paternal transmission in humans. *PLoS genetics* **11**, e1005040.

Qureshi S.A. & Blake R. (1995) Sequence characteristics of a cervid DNA repeat family. *Journal of Molecular Evolution* **40**, 400-4.

Rafia P. & Tarang A. (2016) Sequence Variations of Mitochondrial DNA Displacement-Loop in Iranian Indigenous Sheep Breeds. *Iranian Journal of Applied Animal Science* **6**, 363-8.

Randi E., Fusco G., Lorenzini R., Toso S. & Tosi G. (1991) Allozyme divergence and phylogenetic relationships among Capra, Ovis and Rupicapra (Artyodactyla, Bovidae). *Heredity* **67**, 281-6.

Reisner A. & Bucholtz C. (1983) Apparent relatedness of the main component of ovine 1.714 satellite DNA to bovine 1.715 satellite DNA. *The EMBO journal* **2**, 1145.

Rezaei H.R., Naderi S., Chintauan-Marquier I.C., Taberlet P., Virk A.T., Naghash H.R., Rioux D., Kaboli M. & Pompanon F. (2010) Evolution and taxonomy of the wild species of the genus Ovis (Mammalia, Artiodactyla, Bovidae). *Molecular phylogenetics and evolution* **54**, 315-26.

Rho M., Choi J.-H., Kim S., Lynch M. & Tang H. (2007) De novo identification of LTR retrotransposons in eukaryotic genomes. *BMC genomics* **8**, 90.

Ribeiro T., Marques A., Novák P., Schubert V., Vanzela A.L., Macas J., Houben A. & Pedrosa-Harand A. (2017) Centromeric and non-centromeric satellite DNA organisation differs in holocentric Rhynchospora species. *Chromosoma* **126**, 325-35.

Richard G.-F., Kerrest A. & Dujon B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews* **72**, 686-727.

Rocha J., Chen S. & Beja-Pereira A. (2011) Molecular evidence for fat-tailed sheep domestication. *Tropical animal health and production* **43**, 1237-43.

Rosati S., Pittau M., Alberti A., Pozzi S., York D., Sharp J. & Palmarini M. (2000) An accessory open reading frame (orf-x) of jaagsiekte sheep retrovirus is conserved between different virus isolates. *Virus research* **66**, 109-16.

Rosenbloom K.R., Armstrong J., Barber G.P., Casper J., Clawson H., Diekhans M., Dreszer T.R., Fujita P.A., Guruvadoo L. & Haeussler M. (2014) The UCSC genome browser database: 2015 update. *Nucleic acids research* **43**, D670-D81.

Rubin C.-J., Zody M.C., Eriksson J., Meadows J.R., Sherwood E., Webster M.T., Jiang L., Ingman M., Sharpe T. & Ka S. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587.

Ruggiero R.P., Bourgeois Y. & Boissinot S. (2017) LINE insertion polymorphisms are abundant but at low frequencies across populations of Anolis carolinensis. *Frontiers in genetics* **8**.

Ruiz-Ruano F.J., López-León M.D., Cabrero J. & Camacho J.P.M. (2016) High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific reports* **6**, 28333.

Ryder M. (1984) Sheep. *Evolution of domesticated animals* **1**, 63-84.

Ryder M. (1991) Domestication, history and breed evolution in sheep. *World Animal Science (Netherlands)*.

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

Saha S., Bridges S., Magbanua Z.V. & Peterson D.G. (2008a) Computational approaches and tools used in identification of dispersed repetitive DNA sequences. *Tropical Plant Biology* **1**, 85-96.

Saha S., Bridges S., Magbanua Z.V. & Peterson D.G. (2008b) Empirical comparison of *ab initio* repeat finding programs. *Nucleic acids research* **36**, 2284-94.

Sakamoto K. & Okada N. (1985) Rodent type 2 Alu family, rat identifier sequence, rabbit C family, and bovine or goat 73-bp repeat may have evolved from tRNA genes. *Journal of Molecular Evolution* **22**, 134-40.

Salih R.H.M., Majeský Ľ., Schwarzacher T., Gornall R. & Heslop-Harrison P. (2017) Complete chloroplast genomes from apomictic Taraxacum (Asteraceae): Identity and variation between three microspecies. *PloS one* **12**, e0168008.

Sanger F. & Coulson A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* **94**, 441IN19447-446IN20448.

Sanger F., Nicklen S. & Coulson A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463-7.

Santos S., Chaves R. & Guedes-Pinto H. (2004) Chromosomal localization of the major satellite DNA family (FA-SAT) in the domestic cat. *Cytogenetic and Genome Research* **107**, 119-22.

Schadt E.E., Turner S. & Kasarskis A. (2010) A window into third-generation sequencing. *Human molecular genetics* **19**, R227-R40.

Scherthan H. (1991) Characterisation of a tandem repetitive sequence cloned from the deer Capreolus capreolus and its chromosomal localisation in two muntjac species. *Hereditas* **115**, 43-9.

Scherthan H. (2007) Telomere attachment and clustering during meiosis. *Cellular & Molecular Life Sciences* **64**.

Schmidt T. & Heslop-Harrison J. (1998) Genomes, genes and junk: the large-scale organization of plant chromosomes. *Trends in Plant Science* **3**, 195-9.

Schnable P.S., Ware D., Fulton R.S., Stein J.C., Wei F., Pasternak S., Liang C., Zhang J., Fulton L. & Graves T.A. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-5.

Schneider K., Fangerau H., Michaelsen B. & Raab W.H.-M. (2008) The early history of the transmissible spongiform encephalopathies exemplified by scrapie. *Brain research bulletin* **77**, 343-55.

Schneider K.L., Xie Z., Wolfgruber T.K. & Presting G.G. (2016) Inbreeding drives maize centromere evolution. *Proceedings of the National Academy of Sciences* **113**, E987-E96.

Schopman N.C., Willemsen M., Liu Y.P., Bradley T., van Kampen A., Baas F., Berkhout B. & Haasnoot J. (2011) Deep sequencing of virus-infected cells reveals HIV-encoded small RNAs. *Nucleic acids research* **40**, 414-27.

Schwarzacher T. (2003) DNA, chromosomes, and in situ hybridization. *Genome* **46**, 953-62.

Schwarzacher T. (2008) Chromosomes, recombination and proteins at meiosis–A tribute to Peter Moens (1931–2008). *Chromosome Research* **16**, 679-82.

Schwarzacher T. & Heslop-Harrison P. (2000) *Practical in situ hybridization*. BIOS Scientific Publishers Ltd.

Schwarzacher T., Mayr B. & Schweizer D. (1984) Heterochromatin and nucleolus-organizer-region behaviour at male pachytene of Sus scrofa domestica. *Chromosoma* **91**, 12-9.

Sepsi A., Higgins J.D., Heslop-Harrison J.S.P. & Schwarzacher T. (2017) CENH3 morphogenesis reveals dynamic centromere associations during synaptonemal complex formation and the progression through male meiosis in hexaploid wheat. *The Plant Journal* **89**, 235-49.

Shackleton D. & Lovari S. (1997) Classification adopted for the Caprinae survey. *Wild sheep and goats and their relatives. Status survey and conservation action plan for Caprinae*, 9-14.

Shackleton D.M. (1997) *Wild sheep and goats and their relatives*. IUCN.

Sharp J. & Angus K. (1990) Sheep pulmonary adenomatosis: studies on its aetiology. In: *Maedi-Visna and related diseases* (pp. 177-85. Springer.

Shi J., Wolf S.E., Burke J.M., Presting G.G., Ross-Ibarra J. & Dawe R.K. (2010) Widespread gene conversion in centromere cores. *PLoS biology* **8**, e1000327.

Shimamura M., Abe H., Nikaido M., Ohshima K. & Okada N. (1999) Genealogy of families of SINEs in cetaceans and artiodactyls: the presence of a huge superfamily of tRNA (Glu)-derived families of SINEs. *Molecular Biology and Evolution* **16**, 1046-60.

Shimamura M., Yasue H., Ohshima K. & Abe H. (1997) Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature* **388**, 666.

Sistiaga-Poveda M. & Jugo B. (2014) Evolutionary dynamics of endogenous Jaagsiekte sheep retroviruses proliferation in the domestic sheep, mouflon and Pyrenean chamois. *Heredity* **112**, 571-8.

Skowronski J., Plucienniczak A., Bednarek A. & Jaworski J. (1984) Bovine 1.709 satellite: recombination hotspots and dispersed repeated sequences. *Journal of molecular biology* **177**, 399-416.

Slotkin R.K. & Martienssen R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nature reviews. Genetics* **8**, 272.

Smit A., Hubley R. & Green P. (2013-2015) RepeatMasker Open-4.0. URL http://www.repeatmasker.org.

Smit A.F. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current opinion in genetics & development* **9**, 657-63.

Sonnhammer E.L. & Durbin R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**, GC1-GC10.

Sotero-Caio C.G., Platt R.N., Suh A. & Ray D.A. (2017) Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome biology and evolution* **9**, 161-77.

Sotero-Caio C.G., Volleth M., Hoffmann F.G., Scott L., Wichman H.A., Yang F. & Baker R.J. (2015) Integration of molecular cytogenetics, dated molecular phylogeny, and model-based predictions to understand the extreme chromosome reorganization in the Neotropical genus Tonatia (Chiroptera: Phyllostomidae). *BMC Evolutionary Biology* **15**, 220.

Sperber G., Lövgren A., Eriksson N.-E., Benachenhou F. & Blomberg J. (2009) RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences. *BMC bioinformatics* **10**, S4.

Sperber G.O., Airola T., Jern P. & Blomberg J. (2007) Automated recognition of retroviral sequences in genomic data—RetroTector©. *Nucleic acids research* **35**, 4964-76.

Stewart J.B. & Chinnery P.F. (2015) The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nature reviews. Genetics* **16**, 530.

Sumida T., HINO N., Kawachi H., Matsui T. & Yano H. (2004) Expression of agouti gene in bovine adipocytes. *Animal Science Journal* **75**, 49-51.

Sumner A.T. (2003) Euchromatin and the Longitudinal Differentiation of Chromosomes. *Chromosomes: Organization and Function*, 117-32.

Suzuki S., Ono N., Furusawa C., Ying B.-W. & Yomo T. (2011) Comparison of sequence reads obtained from three next-generation sequencing platforms. *PloS one* **6**, e19534.

Szemraj J., Płucienniczak G., Jaworski J. & Płucienniczak A. (1995) Bovine Alu-like sequences mediate transposition of a new site-specific retroelement. *Gene* **152**, 261-4.

Tamura K., Stecher G., Peterson D., Filipski A. & Kumar S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* **30**, 2725-9.

Taparowsky E.J. & Gerbi S.A. (1982) Sequence analysis of bovine satellite I DNA (1.715 gm/cm 3). *Nucleic acids research* **10**, 1271-81.

Tarlinton R., Meers J. & Young P. (2008) Endogenous retroviruses. *Cellular and Molecular Life Sciences* **65**, 3413-21.

Thomas S.E. & McKee B.D. (2007) Meiotic pairing and disjunction of mini-X chromosomes in drosophila is mediated by 240-bp rDNA repeats and the homolog conjunction proteins SNM and MNM. *Genetics* **177**, 785-99.

Thorgeirsdottir S., Sigurdarson S., Thorisson H.M., Georgsson G. & Palsdottir A. (1999) PrP gene polymorphism and natural scrapie in Icelandic sheep. *Journal of general virology* **80**, 2527-34.

Thornburg B.G., Gotea V. & Makałowski W. (2006) Transposable elements as a significant source of transcription regulating signals. *Gene* **365**, 104-10.

Torres G.A., Gong Z., Iovene M., Hirsch C.D., Buell C.R., Bryan G.J., Novák P., Macas J. & Jiang J. (2011) Organization and evolution of subtelomeric satellite repeats in the potato genome. *G3: Genes, Genomes, Genetics* **1**, 85-92.

Tóth G., Deák G., Barta E. & Kiss G.B. (2006) PLOTREP: a web tool for defragmentation and visual analysis of dispersed genomic repeats. *Nucleic acids research* **34**, W708-W13.

Tranulis M.A. (2002) Influence of the prion protein gene, Prnp, on scrapie susceptibility in sheep. *Apmis* **110**, 33-43.

Tristem M. (2000) Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *Journal of virology* **74**, 3715-30.

Tsuchiya K.D. (2011) Fluorescence in situ hybridization. *Clinics in laboratory medicine* **31**, 525-42.

Ugarković Đ. & Plohl M. (2002) Variation in satellite DNA profiles—causes and effects. *The EMBO journal* **21**, 5955-9.

Ün C., Oztabak K., Özdemir N., Tesfaye D., Mengi A. & Schellander K. (2008) Detection of bovine spongiform encephalopathy-related prion protein gene promoter polymorphisms in local Turkish cattle. *Biochemical genetics* **46**, 820-7.

Utsunomia R., Ruiz-Ruano F.J., Silva D.M., Serrano É.A., Rosa I.F., Scudeler P.E., Hashimoto D.T., Oliveira C., Camacho J.P.M. & Foresti F. (2017) A glimpse into the satellite DNA library in Characidae fish (Teleostei, Characiformes). *Frontiers in genetics* **8**, 103.

Valdez R. & Batten J. (1982) *The wild sheep of the world*. Wild Sheep & Goat International.

Valdez R., Nadler C. & Bunch T. (1978) Evolution of wild sheep in Iran. *Evolution* **32**, 56-72.

van de Lagemaat L.N., Landry J.-R., Mager D.L. & Medstrand P. (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *TRENDS in Genetics* **19**, 530-6.

Van Dijk E.L., Auger H., Jaszczyszyn Y. & Thermes C. (2014) Ten years of next-generation sequencing technology. *TRENDS in Genetics* **30**, 418-26.

Varela M., Spencer T.E., Palmarini M. & Arnaud F. (2009) Friendly viruses. *Annals of the New York Academy of Sciences* **1178**, 157-72.

Vargiu L., Rodriguez-Tomé P., Sperber G.O., Cadeddu M., Grandi N., Blikstad V., Tramontano E. & Blomberg J. (2016) Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology* **13**, 7.

Vassetzky N.S. & Kramerov D.A. (2012) SINEBase: a database and tool for SINE analysis. *Nucleic acids research* **41**, D83-D9.

Vaughan H., Heslop-Harrison J. & Hewitt G. (1999) The localization of mitochondrial sequences to chromosomal DNA in orthopterans. *Genome* **42**, 874-80.

Villesen P., Aagaard L., Wiuf C. & Pedersen F.S. (2004) Identification of endogenous retroviral reading frames in the human genome. *Retrovirology* **1**, 32.

Vissel B., Nagy A. & Choo K. (1992) A satellite III sequence shared by human chromosomes 13, 14, and 21 that is contiguous with α satellite DNA. *Cytogenetic and Genome Research* **61**, 81-6.

Voet T., Liebe B., Labaere C., Marynen P. & Scherthan H. (2003) Telomere-independent homologue pairing and checkpoint escape of accessory ring chromosomes in male mouse meiosis. *The Journal of cell biology* **162**, 795-808.

Vrba E. (1985) African Bovidae: evolutionary events since the Miocene. *South African Journal of Science* **81**, 263-6.

Wall J.D., Tang L.F., Zerbe B., Kvale M.N., Kwok P.-Y., Schaefer C. & Risch N. (2014) Estimating genotype error rates from high-coverage next-generation sequence data. *Genome research* **24**, 1734-9.

Wallace D.C. & Chalkia D. (2013) Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harbor perspectives in biology* **5**, a021220.

Wang T., Zeng J., Lowe C.B., Sellers R.G., Salama S.R., Yang M., Burgess S.M., Brachmann R.K. & Haussler D. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences* **104**, 18613-8.

Wang Y., Liu S.-y., Li J.-y., Han M. & Wang Z.-l. (2008) Cloning and sequence analysis of genome from the Inner Mongolia strain of the endogenous betaretroviruses (enJSRV). *Virologica Sinica* **23**, 15-24.

Waters P.D., Dobigny G., Pardini A.T. & Robinson T.J. (2004) LINE-1 distribution in Afrotheria and Xenarthra: implications for understanding the evolution of LINE-1 in eutherian genomes. *Chromosoma* **113**, 137-44.

Weiss-Schneeweiss H., Leitch A.R., McCann J., Jang T.-S. & Macas J. (2015) Employing next generation sequencing to explore the repeat landscape of the plant genome. *Next Generation Sequencing in Plant Systematics. Regnum Vegetabile* **157**, 155-79.

Westaway D., Zuliani V., Cooper C.M., Da Costa M., Neuman S., Jenny A.L., Detwiler L. & Prusiner S.B. (1994) Homozygosity for prion protein alleles encoding glutamine-171 renders sheep susceptible to natural scrapie. *Genes & development* **8**, 959-69.

Wichman H., Payne C., Ryder O., Hamilton M., Maltbie M. & Baker R. (1991) Genomic distribution of heterochromatic sequences in equids: implications to rapid chromosomal evolution. *Journal of Heredity* **82**, 369-77.

Wicker T., Sabot F., Hua-Van A., Bennetzen J.L., Capy P., Chalhoub B., Flavell A., Leroy P., Morgante M. & Panaud O. (2007) A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**, 973-82.

Wilhelm M. & Wilhelm F.-X. (2001) Reverse transcription of retroviruses and LTR retrotransposons. *Cellular and Molecular Life Sciences CMLS* **58**, 1246-62.

Winokur S.T., Bengtsson U., Feddersen J., Mathews K.D., Weiffenbach B., Bailey H., Markovich R.P., Murray J.C., Wasmuth J.J. & Altherr M.R. (1994) The DNA rearrangement associated with facioscapulohumeral muscular dystrophy involves a heterochromatin-associated repetitive element: implications for a role of chromatin structure in the pathogenesis of the disease. *Chromosome Research* **2**, 225-34.

Wischmann C. & Schuster W. (1995) Transfer of rps10 from the mitochondrion to the nucleus in Arabidopsis thaliana: evidence for RNA-mediated transfer and exon shuffling at the integration site. *FEBS letters* **374**, 152-6.

Woischnik M. & Moraes C.T. (2002) Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome research* **12**, 885-93.

Wood N. & Phua S. (1996) Variation in the control region sequence of the sheep mitochondrial genome. *Animal genetics* **27**, 25-33.

Wu H., Tong C., Li E. & Luo T. (2012) Insight into gene evolution within Cervidae and Bovidae through genetic variation in MHC-DQA in the black muntjac (Muntiacus crinifrons). *Genetics and Molecular Research* **11**, 2888-98.

Wurster D. & Benirschke K. (1968) Chromosome studies in the superfamily Bovoidea. *Chromosoma* **25**, 152-71.

Yang F., Gell K., Van Der Heijden G.W., Eckardt S., Leu N.A., Page D.C., Benavente R., Her C., Höög C. & McLaughlin K.J. (2008) Meiotic failure in male mice lacking an X-linked factor. *Genes & development* **22**, 682-91.

Yang H., Wang G., Wang M., Ma Y., Yin T., Fan R., Wu H., Zhong L., Irwin D.M. & Zhai W. (2017) The origin of chow chows in the light of the East Asian breeds. *BMC genomics* **18**, 174.

York D. & Querat G. (2003) A history of ovine pulmonary adenocarcinoma (jaagsiekte) and experiments leading to the deduction of the JSRV nucleotide sequence. *Jaagsiekte Sheep Retrovirus and Lung Cancer*, 1-23.

York D., Vigne R., Verwoerd D. & Querat G. (1992) Nucleotide sequence of the jaagsiekte retrovirus, an exogenous and endogenous type D and B retrovirus of sheep and goats. *Journal of virology* **66**, 4930-9.

York D.F., Vigne R., Verwoerd D. & Querat G. (1991) Isolation, identification, and partial cDNA cloning of genomic RNA of jaagsiekte retrovirus, the etiological agent of sheep pulmonary adenomatosis. *Journal of virology* **65**, 5061-7.

Youngman S., van Luenen H.G. & Plasterk R.H. (1996) Rte-1, a retrotransposon-like element in Caenorhabditis elegans. *FEBS letters* **380**, 1-7.

Yuan L., Liu J.-G., Zhao J., Brundell E., Daneholt B. & Höög C. (2000) The murine SCP3 gene is required for synaptonemal complex assembly, chromosome synapsis, and male fertility. *Molecular cell* **5**, 73-83.

Zahn J., Kaplan M.H., Fischer S., Dai M., Meng F., Saha A.K., Cervantes P., Chan S.M., Dube D. & Omenn G.S. (2015) Expansion of a novel endogenous retrovirus throughout the pericentromeres of modern humans. *Genome biology* **16**, 74.

Zartman D. & Bruere A. (1974) Giemsa banding of the chromosomes of the domestic sheep (Ovis aries). *Canadian Journal of Genetics and Cytology* **16**, 555-64.

Zeder M.A. (2008) Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proceedings of the National Academy of Sciences* **105**, 11597-604.

Zeschnigk M., Martin M., Betzl G., Kalbe A., Sirsch C., Buiting K., Gross S., Fritzilas E., Frey B. & Rahmann S. (2009) Massive parallel bisulfite sequencing of CG-rich DNA fragments reveals that methylation of many X-chromosomal CpG islands in female blood DNA is incomplete. *Human molecular genetics* **18**, 1439-48.

Zhang D.-X. & Hewitt G.M. (1996) Nuclear integrations: challenges for mitochondrial DNA markers. *Trends in ecology & evolution* **11**, 247-51.

Zhang L., Li N., Fan B., Fang M. & Xu W. (2004) PRNP polymorphisms in Chinese ovine, caprine and bovine breeds. *Animal genetics* **35**, 457-61.

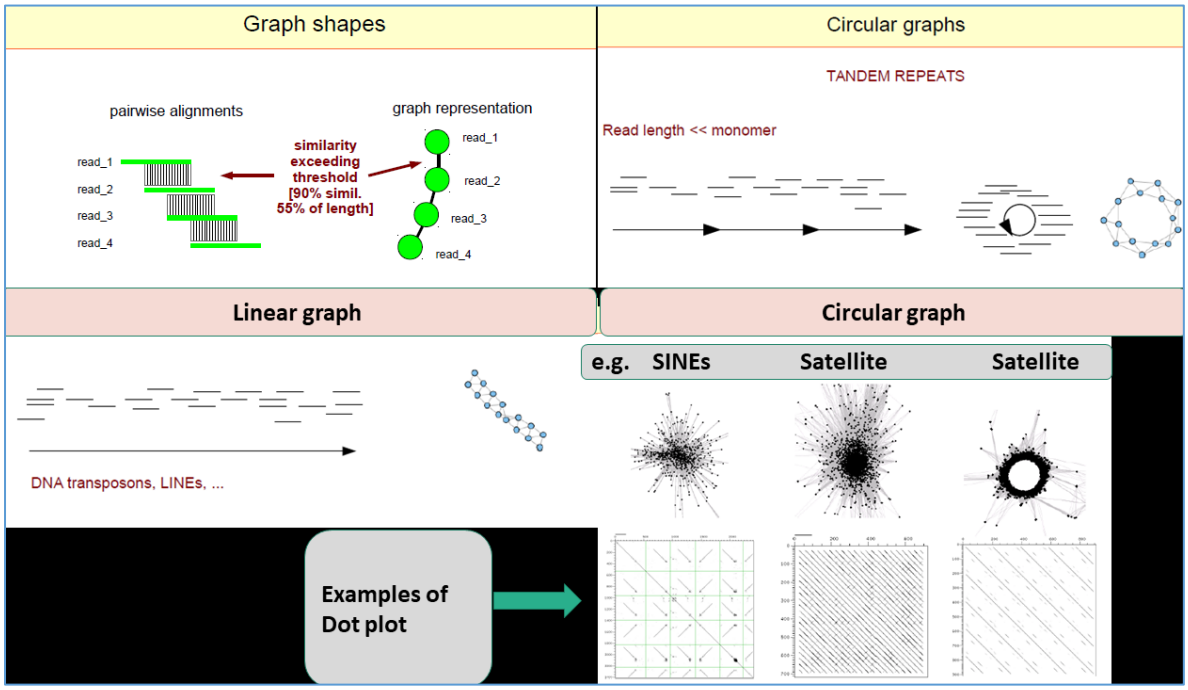
Zhao X., Li N., Guo W., Hu X., Liu Z., Gong G., Wang A., Feng J. & Wu C. (2004) Further evidence for paternal inheritance of mitochondrial DNA in the sheep (Ovis aries). *Heredity* **93**, 399.

Zhao Y., Zhao E., Zhang N. & Duan C. (2011) Mitochondrial DNA diversity, origin, and phylogenic relationships of three Chinese large-fat-tailed sheep breeds. *Tropical animal health and production* **43**, 1405.

Zickler D. & Kleckner N. (1998) The leptotene-zygotene transition of meiosis. *Annual review of genetics* **32**, 619-97.

Appendices

Appendix 1.1



Appendix 1.1 Show conversion of sequence relationship into a 3D dimension. Different repetitive classes yield different graphic shapes. For instance, LINEs repeats shape linear graphs while tandemly repeated sequences (see dot plot) produce circular shapes of graph. The resulted shapes reflect sequence similarity between reads. RepeatExplorer utilize the Fruchterman and Reingold algorithm to calculate graph layouts.

Appendix 1.2 shows the size, GC%, proteins, tRNA, other RNA, genes and pseudogenes of each chromosome assembled so far in sheep genome.

Chromosome No.	RefSeq	Size (Mb)	GC%	Protein	tRNA	Other RNA	Gene	Pseudogene
Chr_1	NC_019458.2	275.41	40.9	4,492	200	593	2,813	331
Chr_2	NC_019459.2	248.97	41	3,655	111	555	2,139	264
Chr_3	NC_019460.2	224	42.4	4,287	127	550	2,739	295
Chr_4	NC_019461.2	119.22	40.6	1,372	51	287	975	121
Chr_5	NC_019462.2	107.84	42	2,204	65	285	1,616	174
Chr_6	NC_019463.2	116.89	40	1,361	52	206	823	120
Chr_7	NC_019464.2	100.01	41.6	1,972	66	268	1,197	114
Chr_8	NC_019465.2	90.62	40.2	1,028	42	184	646	80
Chr_9	NC_019466.2	94.58	41.1	1,001	38	191	690	81
Chr_10	NC_019467.2	86.38	40.4	748	31	160	521	64
Chr_11	NC_019468.2	62.17	46.3	2,083	77	264	1,449	91

Chr_12	NC_019469.2	79.03	42.8	1,240	37	221	872	85
Chr_13	NC_019470.2	82.95	43.8	1,572	34	252	967	80
Chr_14	NC_019471.2	62.57	45.8	1,955	45	302	1,446	106
Chr_15	NC_019472.2	80.78	42	1,682	40	176	1,212	153
Chr_16	NC_019473.2	71.69	41.1	631	20	139	464	68
Chr_17	NC_019474.2	72.25	42.4	1,080	42	184	739	72
Chr_18	NC_019475.2	68.49	43.1	955	34	410	747	81
Chr_19	NC_019476.2	60.45	43.6	1,206	20	307	685	39
Chr_20	NC_019477.2	51.05	43.6	1,203	184	186	1,022	75
Chr_21	NC_019478.2	49.99	44.3	1,027	22	161	824	87
Chr_22	NC_019479.2	50.78	42.9	841	19	96	473	42
Chr_23	NC_019480.2	62.28	41.9	689	26	134	423	42
Chr_24	NC_019481.2	41.98	47.3	1,133	69	133	891	50
Chr_25	NC_019482.2	45.22	42.4	744	20	124	456	73
Chr_26	NC_019483.2	44.05	41.8	461	20	102	324	33
Chr_X	NC_019484.2	135.19	40.5	1,765	75	220	1,228	263

Appendix 1.3 shows global statistics of the current *Ovis aries* genome assemblies (NCBI, 2015)

Total sequence length	2,615,516,299
Total assembly gap length	28,000,626
Number of scaffolds	5,466
Scaffold N50	100,009,711
Scaffold L50	8
Number of contigs	48,482
Contig N50	150,472
Contig L50	5,008
Total number of chromosomes and plasmids	28

Appendix 2.1 (a) Genomic DNA of 31 Kurdistani sheep breeds. (b) Geographical locations. (c) DNA samples used for Next Generation Sequencing see section 2.2.7.2. F; female and M; male.

Breed[a)]	**Sample location** [b)]	**Sample code/ Mitogenome** [c)]	**Sex**
Hamdani	Duhok	H-364-P	F
Hamdani	**Erbil**	**H-369-P/HamM** [c)]	**M**
Hamdani	Duhok	H-368-P	F
Hamdani	Duhok	H-390-P	F
Hamdani	Duhok	H-374-P	F
Hamdani	Duhok	H1a	M
Hamdani	Erbil	Hb2-a	F
Hamdani	Erbil	Hb1-B	F
Hamdani	**Duhok**	**H115-P/HamJ2** [c)]	**M**
Hamdani	**Erbil**	**Hb4/HamJ1** [c)]	**M**
Karadi	**Erbil**	**K279-P/KarM** [c)]	**M**
Karadi	Duhok	K972-P	F
Karadi	Duhok	K970-P	F
Karadi	Duhok	K680-P	F
Karadi	Duhok	K688-P	F
Karadi	Duhok	K1a	M
Karadi	Sulaymaniyah	K5-SUL	M
Karadi	Sulaymaniyah	K6-SUL	M
Karadi	Sulaymaniyah	K7-SUL	M
Karadi	Sulaymaniyah	K8-SUL	M
Karadi	Duhok	KB5-B	F
Karadi	Duhok	KB3	F
Karadi	Duhok	1K	M
Karadi	Duhok	2K-5350	F
Karadi	Duhok	3K-00454	F

Karadi	Duhok	4K-5530	F
Karadi	**Duhok**	**5546/KarJ** [c)]	**F**
Awassi	Duhok	1Aw	F
Awassi	Duhok	2Aw	F
Awassi	Duhok	4Aw	F
Awassi	Duhok	5Aw	F

Appendix 3.1 breed characteristics of Kurdistani sheep

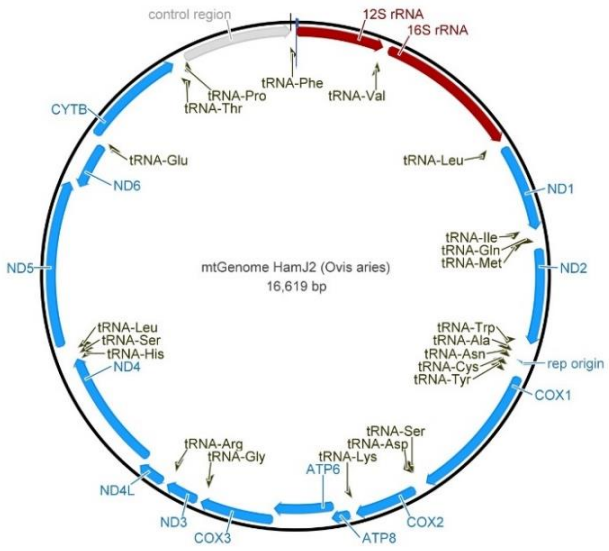
Appendix 3.1 Breed characteristics of Kurdistani sheep include fat tail, long wool and long ears, Roman nose and specific coloration. Karadi sheep tend to have yellowish very coarse wool, black faces and long ears while Hamdani sheep are larger and have longer ears than Karadi sheep. Tails are almost reach the ground; their fleece is more whitish but often speckled. The Awassi breed are commonly white with red to brown large faces and produce carpet quality wool; they are often horned.
A: Hamdani mixed with Karadi (H115-P), male

B: Hamdani (H369-P), male

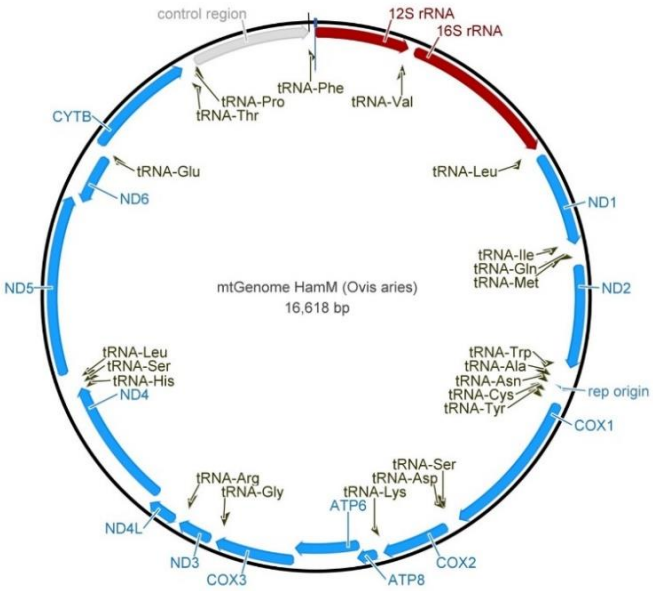
C, D: Karadi mixed with Awassi (5546), female

Appendix 3.2 Mitogenome sequences, their database accession numbers and references used for phylogenetic analysis shown in Figure 3.3).

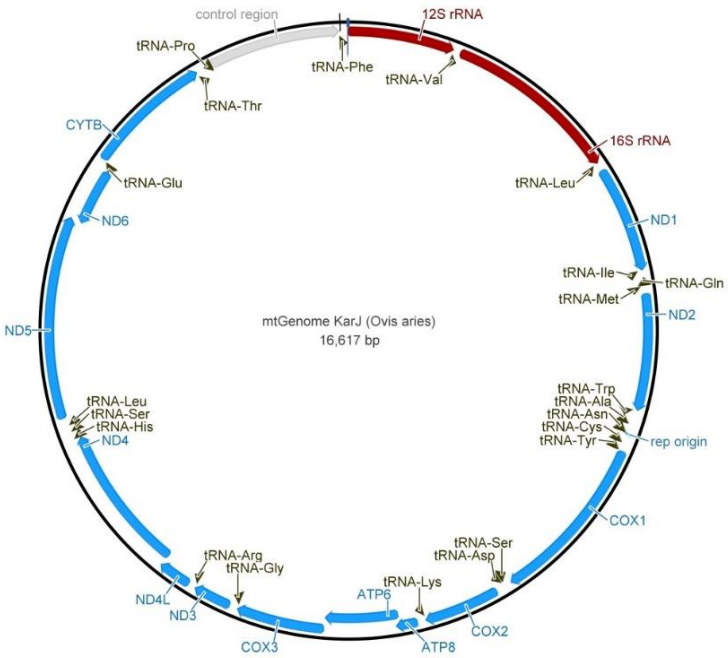
Accession	Sample name	Haplogroup	References
HM236174	*O. aries* cl122	HPGA	Meadows et al. (2011)
HM236175	*O. aries* r359	HPGA	Meadows et al. (2011)
HM236176	*O. aries* kk1	HPGB	Meadows et al. (2011)
HM236177	*O. aries* kk2	HPGB	Meadows et al. (2011)
HM236178	*O. aries* kk12	HPGC	Meadows et al. (2011)
HM236179	*O. aries* mk4	HPGC	Meadows et al. (2011)
HM236180	*O. aries* mk3	HPGD	Meadows et al. (2011)
HM236181	*O. aries* mk9	HPGD	Meadows et al. (2011)
HM236182	*O. aries* aw25	HPGE	Meadows et al. (2011)
HM236183	*O. aries* tj6	HPGE	Meadows et al. (2011)
HM236184	*O. musimon* h1	HPGB	Meadows et al. (2011)
HM236185	*O. musimon* h2	HPGB	Meadows et al. (2011)
MF004242	*O. aries* HamJ2	HPGA	Present Study
MF004243	*O. aries* HamJ1	HPGA	Present Study
MF004244	*O. aries* HamM	HPGA	Present Study
MF004245	*O. aries* KarJ	HPGB	Present Study
MF004246	*O. aries* KarM	HPGB	Present Study
HM236188	*O. ammon* Argali h77	wild sheep	Meadows et al. (2011)
HM236186	*O. vignei* Urial h75	wild sheep	Meadows et al. (2011)
HM236187	*O. vignei* Urial h76	wild sheep	Meadows et al. (2011)
HM236189	*O. vignei* Urial h78	wild sheep	Meadows et al. (2011)
JN181255	*O. canadensis*	bighorn sheep	Miller et al. (2012)
JX101654	*O. ammon hodgsoni*	wild sheep	Jiang et al. (2013)
GU295658	*Capra hircus*	goat	Hassanin et al. (2010)



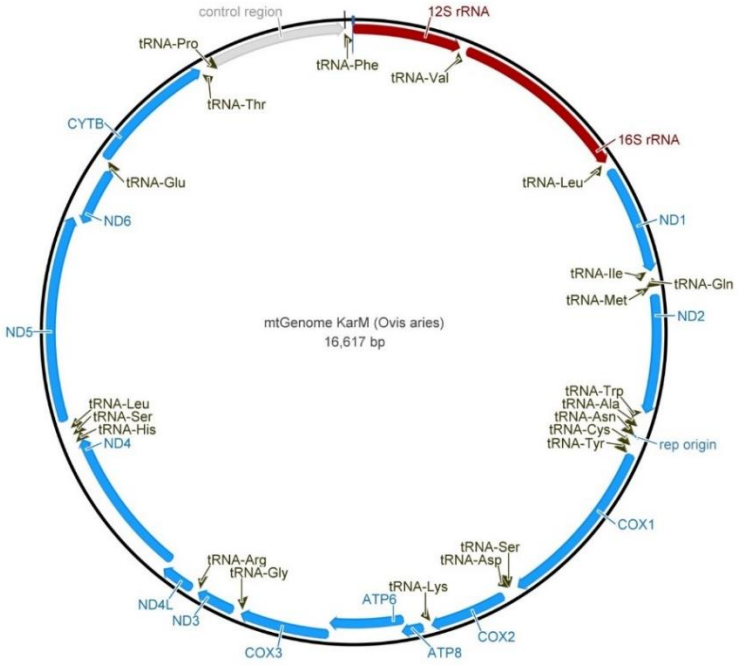
Appendix 3.3. The assembled mitogenome HAMJ2 (16,618bp) of *Ovis aries* Hamdani landrace animal H115-P, Genbank accession number (MF004242) with major features: there are 13 protein-coding genes (light blue bars, with the arrow pointing in the transcription directions), 22 tRNA genes (black triangles), the 12S and 16S rRNA genes (dark red) and the D-loop control region (grey). The GC content is 38.9%. For assembly data see table 1.



Appendix 3.4. The assembled mitogenome HamM (16,618bp) of *Ovis aries* Hamdani landrace animal H369-P, Genbank accession number (MF004244) with major features: there are 13 protein-coding genes (light blue bars, with the arrow pointing in the transcription directions), 22 tRNA genes (black triangles), the 12S and 16S rRNA genes (dark red) and the D-loop control region (grey). The GC content is 38.9%. For assembly data see table 1.



Appendix 3.5. The assembled mitogenome KarJ (16,617bp) of *Ovis aries* Karadi landrace animal 5546, Genbank accession number (MF004245) with major features: there are 13 protein-coding genes (light blue bars, with the arrow pointing in the transcription directions), 22 tRNA genes (black triangles), the 12S and 16S rRNA genes (dark red) and the D-loop control region (grey). The GC content is 38.9%. For assembly data see table 1.



Appendix 3.6. The assembled mitogenome KarM (16,617bp) of *Ovis aries* Karadi landrace animal K269-P, Genbank accession number (MF004246) with major features: there are 13 protein-coding genes (light blue bars, with the arrow pointing in the transcription directions), 22 tRNA genes (black triangles), the 12S and 16S rRNA genes (dark red) and the D-loop control region (grey). The GC content is 38.9%. For assembly data see table 1.

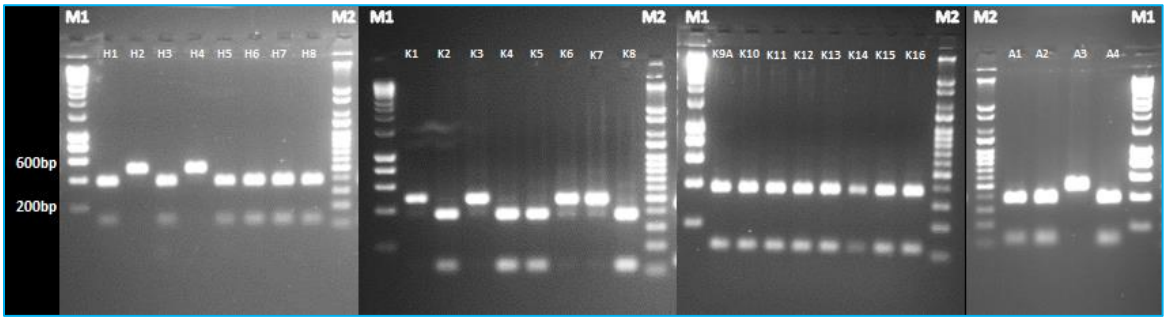
Appendix 3.7 Mutalyzer report of SNPs between published and our assembled **HPG**A and **HPG**B mitogenomes respectively, as well as between our **HPG**A and **HPG**B reference mitogenomes.

Position and variants	Type	**HM236174 (HPGA)**	**HamJ1 (HPGA)**	Position and variants	Type	**HamJ1 (HPGA)**	**HamJ2 (HPGA)**
59T>C	subst	T	C	59C>T	subst	C	T
569dup	dup		A	1513T>C	subst	T	C
1512C>T	subst	C	T	1663C>T	subst	C	T
1662T>C	subst	T	C	2224C>T	subst	C	T
2225G>A	subst	G	A	3611C>T	subst	C	T
4180T>C	subst	T	C	3673T>C	subst	T	C
7128C>T	subst	C	T	4181C>T	subst	C	T
7727A>G	subst	A	G	7129T>C	subst	T	C
8126G>A	subst	G	A	8304A>G	subst	A	G
8176G>A	subst	G	A	9557G>C	subst	G	C
8303G>A	subst	G	A	14597T>C	subst	T	C
11033C>T	subst	C	T	14612A>G	subst	A	G
12644C>T	subst	C	T	14880T>C	subst	T	C
14879C>T	subst	C	T	15156T>C	subst	T	C
15000del	del	T		15462G>A	subst	G	A
15001_1insT	ins		T	15581T>C	subst	T	C
15462A>G	subst	A	G	15649G>A	subst	G	A
15581C>T	subst	C	T	15681C>T	subst	C	T
15681T>C	subst	T	C	16057G>A	subst	G	A
15745C>T	subst	C	T	16149C>T	subst	C	T
15820C>T	subst	C	T	16474dup	dup		T
15895C>T	subst	C	T				
15974G>A	subst	G	A				
16057A>G	subst	A	G				
16149T>C	subst	T	C				

Position and variants	Type	**HamJ1 (HPGA)**	**KarJ (HPGB)**	Position and variants	Type	**KarJ (HPGB)**	**KarM (HPGB)**
59C>T	subst	C	T	5690A>G	subst	A	G
291T>C	subst	T	C	6692C>T	subst	C	T
538G>A	subst	G	A	7816T>C	subst	T	C
1100T>A	subst	T	A	8125C>T	subst	C	T
1113C>T	subst	C	T	9058C>T	subst	C	T
1513T>C	subst	T	C	13008T>C	subst	T	C
1663C>T	subst	C	T	13014T>C	subst	T	C
2445T>C	subst	T	C	14303A>C	subst	A	C
2776T>C	subst	T	C	15461G>A	subst	G	A
2968T>C	subst	T	C	15688A>G	subst	A	G
3220A>G	subst	A	G	15763A>G	subst	A	G
3433A>G	subst	A	G	15838A>G	subst	A	G
3664A>G	subst	A	G	15905T>C	subst	T	C

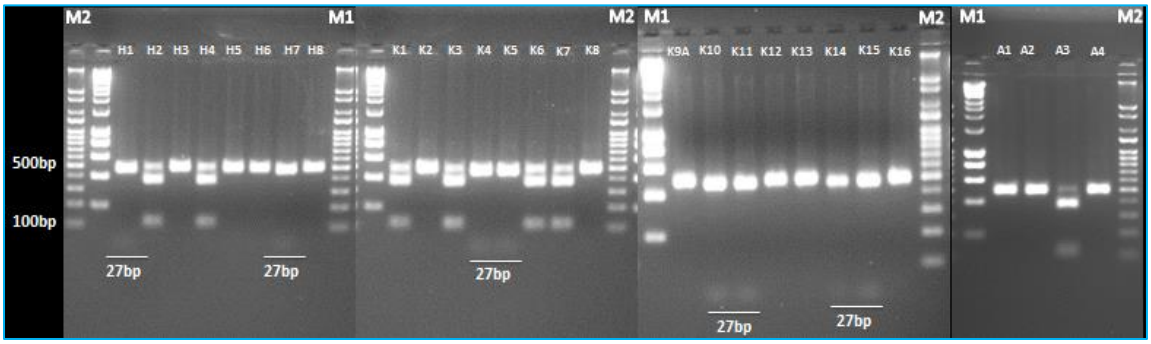
4181C>T	subst	C	T	15924A>G	subst	A	G
4184C>T	subst	C	T	15935C>T	subst	C	T
4217T>C	subst	T	C	15978G>A	subst	G	A
4445C>T	subst	C	T	16603T>C	subst	T	C
4841T>C	subst	T	C				
4917T>C	subst	T	C				
4937T>C	subst	T	C				
5567T>C	subst	T	C				
5690G>A	subst	G	A				
5786C>T	subst	C	T				
6269T>C	subst	T	C				
6512C>T	subst	C	T				
6557G>A	subst	G	A				
6630T>C	subst	T	C				
6728G>A	subst	G	A				
7129T>C	subst	T	C				
7143C>T	subst	C	T				
7218T>C	subst	T	C				
7437T>C	subst	T	C				
7721G>T	subst	G	T				
7816C>T	subst	C	T				
8041G>A	subst	G	A				
8123C>T	subst	C	T				
8150A>G	subst	A	G				
8258C>T	subst	C	T				
8304A>G	subst	A	G				
8378T>C	subst	T	C				
9058T>C	subst	T	C				
9130A>G	subst	A	G				
9190G>A	subst	G	A				
9286G>A	subst	G	A				
9758T>C	subst	T	C				
9998T>C	subst	T	C				
10120A>G	subst	A	G				
10551G>A	subst	G	A				
10785T>C	subst	T	C				
10854A>G	subst	A	G				
11025A>G	subst	A	G				
11319G>A	subst	G	A				
11484C>T	subst	C	T				
11493A>G	subst	A	G				
11608T>C	subst	T	C				
11784T>C	subst	T	C				
11835C>T	subst	C	T				
11847C>T	subst	C	T				
12024C>T	subst	C	T				
12288C>T	subst	C	T				
13008C>T	subst	C	T				

13014C>T	subst	C	T				
13098A>G	subst	A	G				
13173C>T	subst	C	T				
13437C>T	subst	C	T				
13577T>C	subst	T	C				
13838C>T	subst	C	T				
13856C>T	subst	C	T				
14303C>A	subst	C	A				
14468T>C	subst	T	C				
14654A>G	subst	A	G				
14880T>C	subst	T	C				
15460_15462delinsCGA	delins	TAG	CGA				
15485A>G	subst	A	G				
15548A>G	subst	A	G				
15581T>C	subst	T	C				
15584T>C	subst	T	C				
15598C>T	subst	C	T				
15636_15647delins	delins	GAATGTGCTAAG	AAACATGCAAA				
15658A>G	subst	A	G				
15681C>T	subst	C	T				
15710C>T	subst	C	T				
15716G>A	subst	G	A				
15733A>G	subst	A	G				
15785C>T	subst	C	T				
15791G>A	subst	G	A				
15808A>G	subst	A	G				
15860C>T	subst	C	T				
15866G>A	subst	G	A				
15883A>G	subst	A	G				
15941G>A	subst	G	A				
15946G>A	subst	G	A				
15958_959inv	inv	CA	TG				
15960C>T	subst	C	T				
15974A>G	subst	A	G				
15980C>T	subst	C	T				
15984G>A	subst	G	A				
16022G>A	subst	G	A				
16024C>T	subst	C	T				
16038A>G	subst	A	G				
16044C>T	subst	C	T				
16050C>T	subst	C	T				
16057G>A	subst	G	A				
16098_16099inv	inv	TG	CA				
16149C>T	subst	C	T				
16211T>C	subst	T	C				
16219T>C	subst	T	C				
16443C>T	subst	C	T				
16456A>G	subst	A	G				

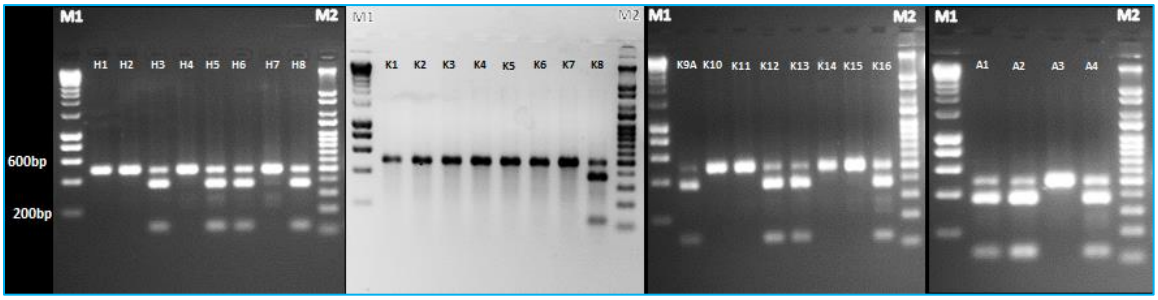


Appendix 3.8. RFLP Patterns of mtDNA ND1 gene digested with restriction enzyme BamHI distinguishing haplogroups HA/HC/HD/HE from the haplogroup HB. The two cut bands mean haplotypes HA/HC/HD/HE while the uncut bands mean haplogroup HB. M1 lane is the 1 kb ladder marker starting from 200bp. M2 lane is the Q-Step2 DNA ladder marker starting from 100bp.

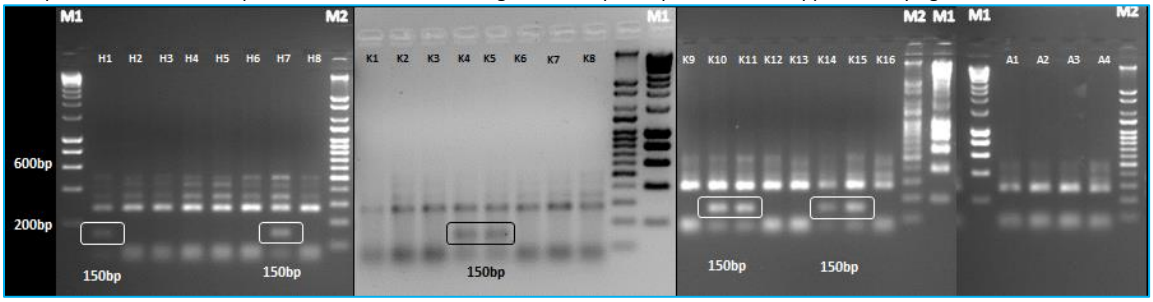
Codes refer to DNA samples as follows: (H1) H-364-P, (H2) H-368-P, (H3) H-390-P, (H4) H-374-P, (H5) Hb1-B, (H6) H-369-P, (H7) Hb2-a, (H8) H1a; (K1) K279-P, (K2) K972-P, (K3) K970-P, (K4) K680-P, (K5) K688-P, (K6) 2K – 5350, (K7) K7-SUL, (K8) 1K; (K9A) H115-P, (K9B) K1a, (K10) K5-SUL, (K11) K6-SUL, (K12) K8-SUL, (K13) KB5-B, (K14) KB3, (K15) 3K-00454, (K16) 4K-5530; (A1) 1Aw, (A2) 2Aw, (A3) 4Aw, (A4) 5Aw.



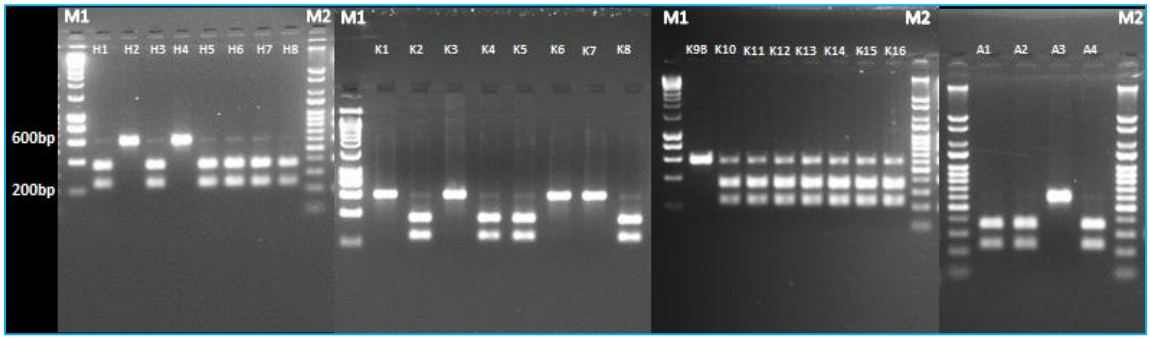
Appendix 3.9. RFLP Patterns of mtDNA ND1 gene digested with restriction enzyme AvaII distinguishing three different haplogroups. Haplogroup HB with 2 bands (379bp /113bp) and HC/HD/HE with another 2 bands (465/27) but the uncut band represent the haplogroup HA. M1 lane is the 1 kb ladder marker starting from 200bp. M2 lane is the Q-Step2 DNA ladder marker starting from 100bp. Sample codes see supplementary fig. 6.



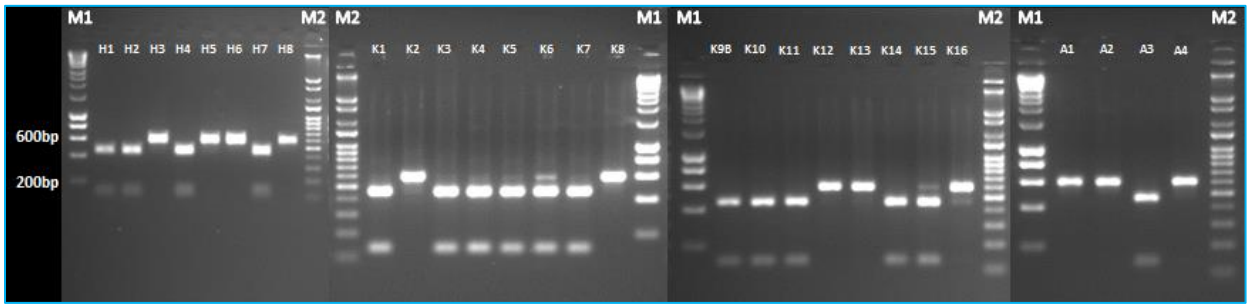
Appendix 3.10. RFLP Patterns of mtDNA ND1 gene digested with restriction enzyme AluI distinguishing the haplogroup HA with 2 obtained bands (368bp/124bp) from the rest haplogroups (uncut band – 492bp). M1 lane is the 1 kb ladder marker starting from 200bp. M2 lane is the Q-Step2 DNA ladder marker starting from 100bp. Sample codes see supplementary fig. 6.



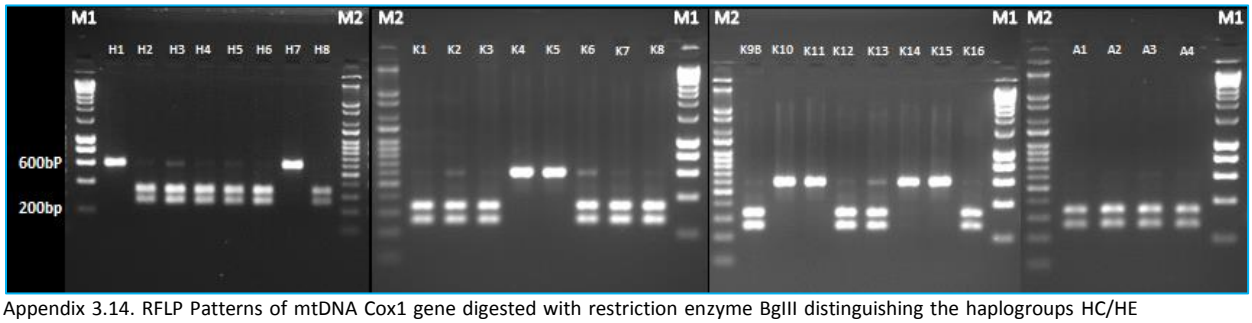
Appendix 3.11. RFLP Patterns of mtDNA ND1 gene digested with restriction enzyme MseI identifying the both haplogroups HC/HE by showing 150bp band. M1 lane is the 1 kb ladder marker starting from 200bp. M2 lane is the Q-Step2 DNA ladder marker starting from 100bp. Sample codes see supplementary fig. 6.



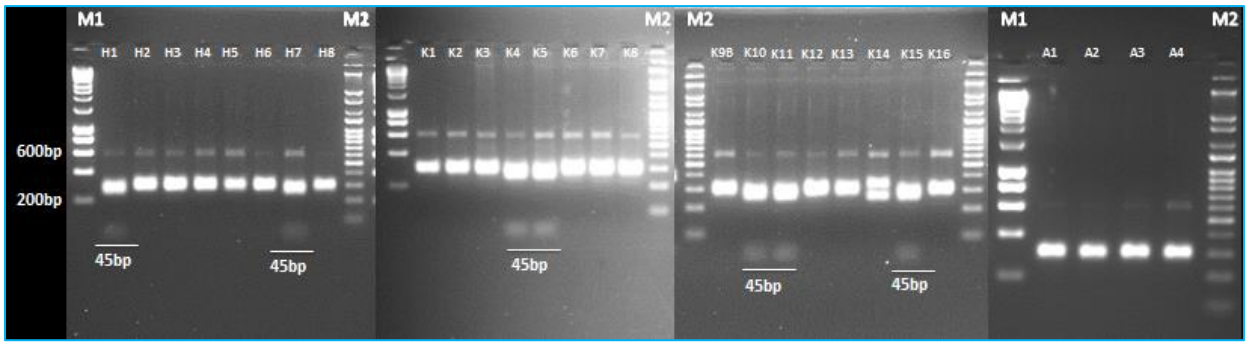
Appendix 3.12. RFLP Patterns of mtDNA Cox1 gene digested with restriction enzyme BstNI distinguishing the haplogroup HB (uncut band-588bp) from the rest haplogroups HA/HC/HD/HE (354bp/271bp). M1 lane is the 1 kb ladder marker starting from 200bp. M2 lane is the Q-Step2 DNA ladder marker starting from 100bp. Sample codes see supplementary fig. 6.



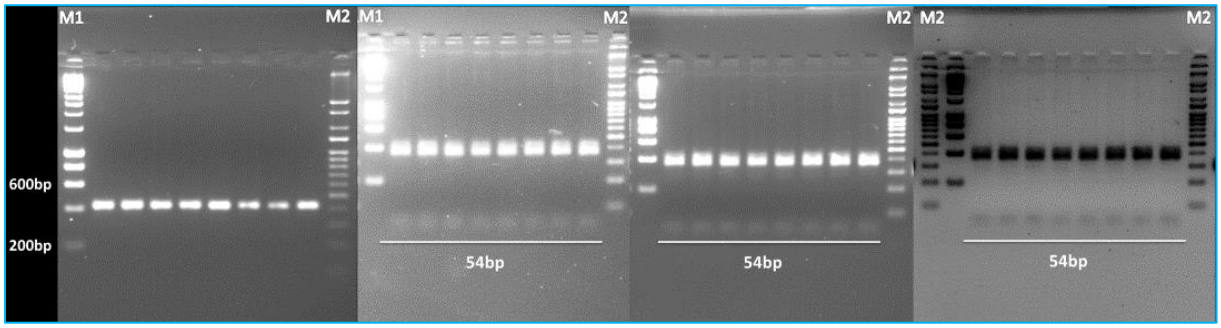
Appendix 3.13. RFLP Patterns of mtDNA Cox1 gene digested with restriction enzyme HinfI distinguishing the haplogroup HA (uncut band-588bp) from the rest haplogroups HB/HC/HD/HE (452/136). M1 lane is the 1 kb ladder marker starting from 200bp. M2 lane is the Q-Step2 DNA ladder marker starting from 100bp. Sample codes see supplementary fig. 6.



Appendix 3.14. RFLP Patterns of mtDNA Cox1 gene digested with restriction enzyme BglII distinguishing the haplogroups HC/HE together (uncut band-588bp) from the rest haplogroups HA/HB/HD/ (331/257). M1 lane is the 1 kb ladder marker starting from 200bp. M2 lane is the Q-Step2 DNA ladder marker starting from 100bp. Sample codes see supplementary fig. 6.



Appendix 3.15. RFLP Patterns of mtDNA Cox1 gene digested with restriction enzyme AccI distinguishing the haplogroups HC/HE together (285/257/45) from the rest haplogroups HA/HB/HD/ (303/285). M1 lane is the 1 kb ladder marker starting from 200bp. M2 lane is the Q-Step2 ladder marker starting from 100bp. Sample codes see supplementary fig. 6.



Appendix 3.16. RFLP Patterns of mtDNA CYTB-CDS (CYTB Gene) digested with restriction enzyme AluI distinguishing the haplogroup HC (369/54) from the haplogroups HC/HE (423bp). The first left image reverse before digestion while the rest images are representing 8 samples of haplogroups HC. M1 lane is the 1 kb ladder marker starting from 200bp. M2 lane is the Q-Step2 DNA ladder marker starting from 100bp.

Appendix 3.17. NCBI search results of de novo assembly of unused reads. NCBI search results of contigs of assembled 'unused reads' covering the whole mtGenome, coding region and control region only.

Contigs covering whole mtGenome		Contigs covering coding regions	
#	main NCBI blast result	#	main NCBI blast result
1	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)	1	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)
2	mtGenome of domestic sheep breeds such as (Turfan Black,Baerchuke, Qula Tibetan, Aland and Yecheng)	2	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Kirghiz)
3	mtGenome ofdomestic sheep breeds such as (Gala, Turfan Black,Baerchuke, Lop, Tan and Yecheng)	3	mtGenome of domestic sheep breeds such as (Turfan Black,Baerchuke, Qula Tibetan, Aland and Yecheng)
4	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Kirghiz)	4	mtGenome of domestic sheep breeds such as (Turfan Black,Baerchuke, Qula Tibetan, Aland and Yecheng)
5	*Ovis* canadensis chromosome sequences (X); mtGenome of Capra hircus breeds	5	*Ovis* canadensis chromosome sequences (X & 2); mtGenome of domestic goat
6	mtGenome of domestic sheep breeds such as (Qinhai Tibetan, Aland, Tashkurgan, Tan and Qira Black)	6	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)
7	mtGenome of domestic sheep breeds such as (Gala, Turfan Black,Baerchuke, Lop, Tan and Yecheng)	7	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)
8	mtGenome of domestic sheep breeds such as (Sunite, Turfan Black,Baerchuke, Lop, Tan and Yecheng)	8	mtGenome of domestic sheep breeds such as (Sunite, Turfan Black,Baerchuke, Lop, Tan and Yecheng)
9	mtGenome of domestic sheep breeds such as (Sunite, Turfan Black,Baerchuke, Lop, Tan and Altay)	9	mtGenome of domestic sheep breeds such as (Sunite, Turfan Black,Baerchuke, Lop, Tan and Altay)
10	mtGenome of wild sheep such as (*Ovis* vignei and *Ovis* ammon darwini)	10	mtGenome of wild sheep such as (*Ovis* ammon hodgsoni, O. Vignei, O. ammon and some domestic sheep)
11	MtGenome of wild sheep such as (*Ovis* ammon hodgsoni); mtGenome of capra hircus and mtGenome of Ovibos moschatus.	11	mt genome of (*Ovis* ammon hodgsoni, capra hircus, and Ovibos moschatus)
12	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)	12	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)
13	mtGenome of domestic sheep breeds such as (Turfan Black,Baerchuke, Qula Tibetan, Aland and Yecheng)	13	mtGenome of wild sheep such as (*Ovis* ammon hodgsoni, O. Vignei, O. ammon and some domestic sheep)
14	mtGenome of domestic sheep breeds such as (Duolang, Minxian Black Fur)	14	mtGenome of Ammotragus lervia mitochondrion, Capra pyrenaica, Capra aegagrus and capra hircus)
15	mtGenomes of (Ammotragus lervia mitochondrion0; mtGenome of wild and domestic goat (Capra pyrenaica, Capra aegagrus and capra hircus)	15	mtGenome of domestic sheep such as (Finnsheep, Oxford down); mtGenome of Pseudois nayaur
16	mtGenoome of (Naemorhedus goral, Capra hircus and som domestic sheep breeds)	16	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Kirghiz)
17	mtGenome of domestic sheep breeds such as (Sunite, Turfan Black,Baerchuke, Lop, Tan and Yecheng)	17	mtGenome of domestic sheep breeds such as (Sunite, Turfan Black,Baerchuke, Lop, Tan and Yecheng)
18	mtGenome of domestic sheep breeds such as (Oparino, Turfan Black,Baerchuke, Lop, Tan and Yecheng)	18	mtGenome of domestic sheep breeds such as (Oparino, Turfan Black,Baerchuke, Lop, Tan and Yecheng)
19	mtGenome of (Budorcas taxicolor, Pseudois nayaur, Ammotragus lervia, Capra and O. vignei)	19	mtGenome of (Budorcas taxicolor, Pseudois nayaur, Ammotragus lervia, Capra and O. vignei)

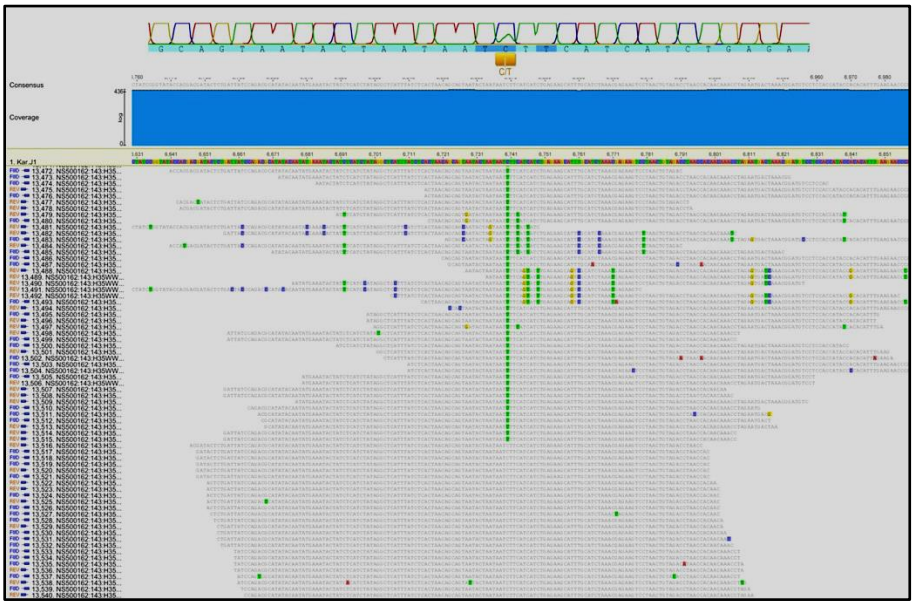
20	mtGenome of (Capra hircus, *Ovis* ammon, *Ovis* orientalis and oxford breed down)	20	mtGenome of (Capra hircus, *Ovis* ammon, oxford breed down and *Ovis* orientalis)
21	mtGenome of wild sheep such as (*Ovis* ammon hodgsoni, O. Vignei, O. ammon and some domestic sheep)	21	mtGenome of (*Ovis* ammon hodgsoni, O. Vignei and O. ammon)
22	mtGenome of wild sheep such as (*Ovis* ammon hodgsoni, O. Vignei, O. ammon and some domestic sheep)	22	mtGenome of wild sheep such as (*Ovis* ammon hodgsoni, O. Vignei, O. ammon and some domestic sheep)
23	mtGenome of domestic sheep breeds such as (Gala, Turfan Black,Baerchuke, Lop, Tan and Yecheng)	23	*Ovis* canadensis chromosome sequences (2); mtGenome of domestic sheep breeds
24	*Ovis* canadensis chromosome sequences (2); mtGenome of Capra hircus breeds	24	mtGenome of (*Ovis* canadensis, *Ovis* vignei,O ammon and some domestic sheep)
25	*Ovis* canadensis chromosome sequences (X & 2); mtGenome of wild sheep breeds	25	*Ovis* canadensis chromosome sequences (6); mtGenome of domestic sheep breeds
26	mtGenome of domestic sheep breeds such as (Duolang, Minxian Black Fur)	26	mtGenome of Hemitragus jayakari mitochondrion
27	mtGenome of (Hemitragus jayakari)	27	mtGenome of domestic sheep (Capra hircus)
28	mtGenome of (Capra hircus, ovibos moschatus, *Ovis* vignei, and Naemorthedus goral0	28	mtGenome of (Capra hircus, ovibos moschatus, *Ovis* vignei Naemorthedus goral)
29	Capricornis milneedwardsii mitochondrion, *Ovis* canadensis mitochondrion, Rupicapra rupicapra, *Ovis* vignei and *Ovis aries*	29	mtGenome of (Capricornis milneedwardsii, *Ovis* canadensis , Rupicapra rupicapra, *Ovis* vignei and domestic sheep)
30	mtGenome of wild sheep such as (*Ovis* ammon hodgsoni, O. Vignei, O. ammon and some domestic sheep)	30	mtGenome of wild sheep such as (*Ovis* ammon hodgsoni, O. Vignei, O. ammon and some domestic sheep)
31	mtgenome of (*Ovis* ammon hodgsoni, Rupicapra rupicapra, *Ovis* vignei and Capra)	31	*Ovis* canadensis chromosome sequences (26); mtGenome of domestic sheep)
32	mtGenome of wild and domestic goat (Capra pyrenaica, Capra aegagrus, capra hircus) and wild sheep (O. ammon)	32	mtGenome of (*Ovis* ammon hodgsoni, Rupicapra rupicapra, *Ovis* vignei and Capra)
33	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)	33	mtGenome of (*Ovis* ammon hodgsoni, Rupicapra rupicapra, *Ovis* vignei and Capra)
34	mtGenome of domestic sheep breeds such as (Sunite, Turfan Black,Baerchuke, Lop, Tan and Yecheng)	34	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)
35	mtGenome of domestic sheep breeds such as (Assaf, Turfan Black,Baerchuke, Lop, Tan and Yecheng)	35	mtGenome of domestic goat (Capra hircus)
36	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)	36	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)
37	mtGenome of domestic goat (Capra hircus)	37	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)
38	mtGenome of domestic sheep breeds such as (Kail, Aland, Tashkurgan, Tan and Qira Black)	38	mtGenome of (Capra hircus and Pseudois nayaur nayaur0
39	*Ovis* canadensis chromosome sequences (9); mtGenome of Cephalophus silvicultor , Boselaphus and Tragelaphus)	39	*Ovis* canadensis chromosome sequences (9); mtGenome of Cephalophus silvicultor , Boselaphus and Tragelaphus)
40	mtGenome of wild sheep such as (*Ovis* ammon hodgsoni, O. Vignei, O. ammon and some domestic sheep)	40	mtGenome of wild sheep such as (*Ovis* ammon hodgsoni, O. Vignei, O. ammon and some domestic sheep)
41	*Ovis* canadensis chromosome sequences (9); mtGenome of (Budorcas taxicolor, Hemitragus hylocrius and Capra)	41	mtGenome of (Budorcas taxicolor, Hemitragus hylocrius and Capra)
42	*Ovis* canadensis chromosome sequences (26&X); mtGenome of Naemorhedus swinhoei and Capra hircus)	42	mtGenome of (Naemorhedus swinhoei, Capra, *Ovis* canadensis)
		43	mtGenome of domestic sheep breeds such as (Awang, Qula Tibetan, Aland, Tashkurgan and Tan)

Contigs covering the control region	
#	**main NCBI blast result**
1	mtGenome of domestic sheep breeds such as (Sunite, Turfan Black,Baerchuke, Lop, Tan and Yecheng)
2	mtGenome of domestich and wild goat (Capra hircus & Capra aegagrus); and mtGenome of Hemitragus jemlahicus

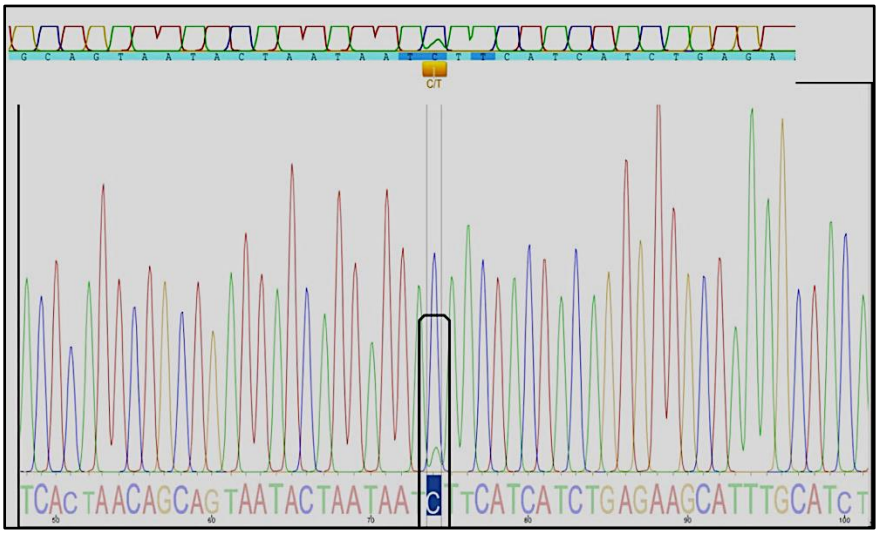
3	mtGenome control region of domestic sheep breeds such as (Turfan Black,Baerchuke, Qula Tibetan, Aland and Yecheng)
4	mtGenome ofdomestic sheep breeds such as (Gala, Turfan Black,Baerchuke, Lop, Tan and Yecheng)
5	mtGenome control region of domestic sheep breeds such as (Turfan Black,Baerchuke, Qula Tibetan, Aland and Yecheng)
6	mtGenome control region of domestic sheep breeds such as (Turfan Black,Baerchuke, Qula Tibetan, Aland and Yecheng)

Appendix 3.18

(A)

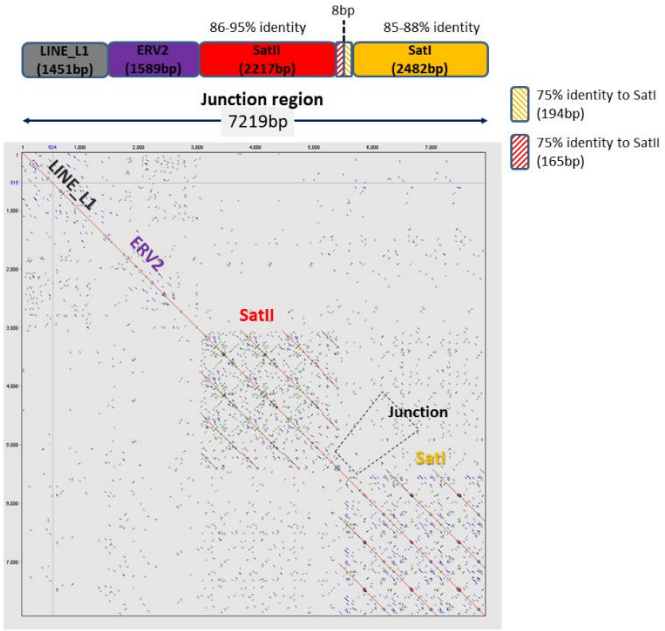


(B)



Appendix 3.18: Polymorphisms in sequence assembly. (A) Raw reads assembled to consensus of last fragment of Cox1 gene and first fragment of tRNA Ser of mitochondria DNA genome showing some sites with polymorphisms, coloured boxes. (B) Sanger sequencing trace showing heterogeneity in base calls (highlighted).

Appendix 4.1 Structure of junction region, including both sheep satellite I and II repeat sequences assembled from map to reference. The coloured boxes show length of repeats and their positions. Percentages above refer to sequence identities to the corresponding satellite monomers. Next to satellite II, it shows other than satellite repeats including (Endogenous retroviruses and LINE-L1).



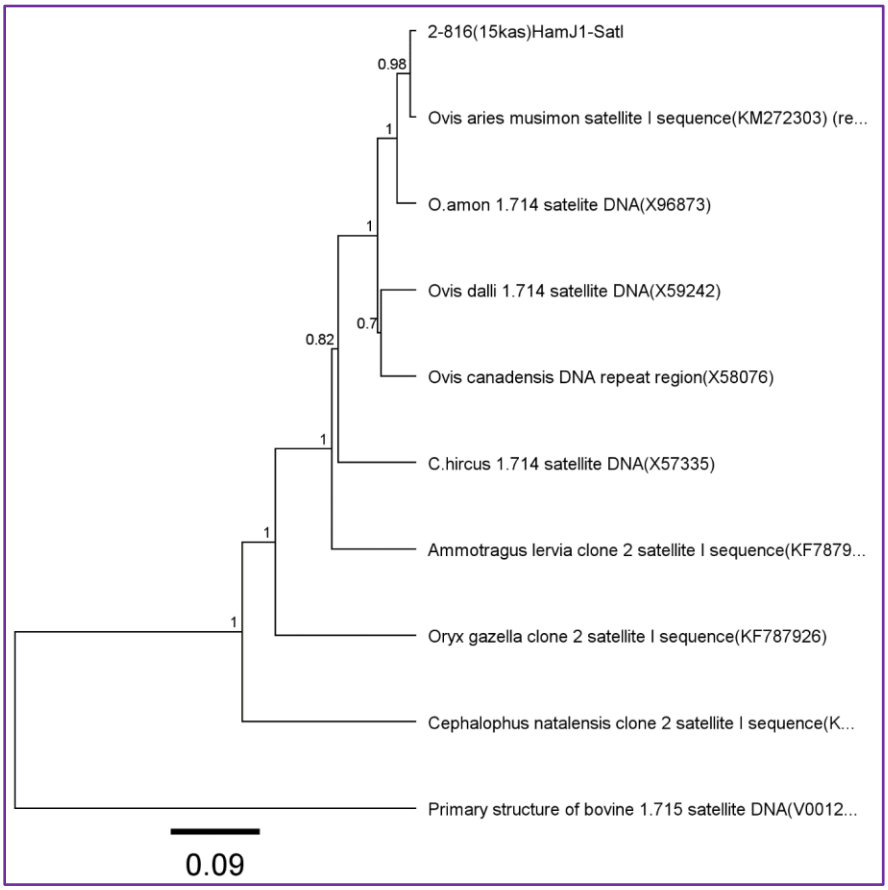
Appendix 4.2 NGS data used for de novo assembly of four sets (L1-L4) of raw reads HamJ2. It shows percentage of raw reads used for de novo assembly resulted in several thousands of contigs.

NGS (Input)	Whole paired reads as input (A)	Used reads by assembler (B)	Total assembled reads (C)	All contigs resulted	% NGS data used for assembly (B/A*100)	% assembled reads out of 20% (C/B*100)	% assembled reads out of input (C/A*100)
L1	14,192,114	2,838,422	1,836,220	441,044	20	64.7	12.9
L2	13,597,746	2,719,549	1,761,485	436,488	20	64.8	13
L3	14,639,056	2,927,811	1,918,003	465,571	20	65.5	13.1
L4	13,791,966	2,758,393	1,774,977	426,330	20	64.3	12.9

Appendix 4.3 Example of assembly report outcome of using set (L3) of raw reads of HamJ2.

Statistics	Unused Reads	All Contigs	Contigs >=100 bp	Contigs >=1000 bp
Number of	1,009,808	465,571	465,533	1,666
Min Length (bp)	35	35	100	1,000
Median Length (bp)		265	265	1,190
Mean Length (bp)	150	295	295	1,444
Max Length (bp)	151	16,641	16,641	16,641
N50 Length (bp)		277	277	1,307
Number of contigs >= N50		171,846	171,842	589
Length Sum (bp)	151,949,244	137,385,685	137,383,121	2,407,234

Appendix 4.4 Phylogenetic relationships between domestic and wild sheep and other species of Bovidae family based on neighbour joining unrooted tree using satellite I DNA sequences.



Appendix 4.5

Tandem Repeat Analyzer
Run statistics:
Number of input sequences: 58792112; Number of analysed sequences: 663509; Threshold for cluster merging: 0.2; Proportion of sequences in analysed clusters: 29 %

Consensus files - fasta format:

Putative satellites (high confidence) - total 2 found

Putative satellites (low confidence) - total 2 found

Documentation

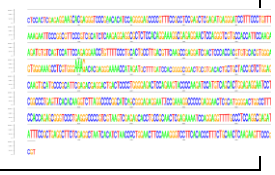
Putative satellites (high confidence)

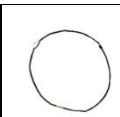
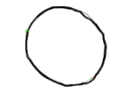
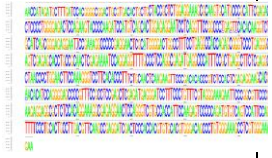
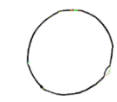
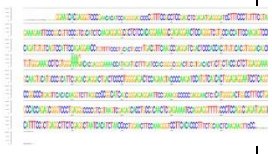
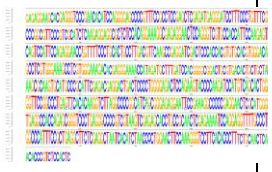
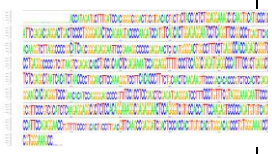
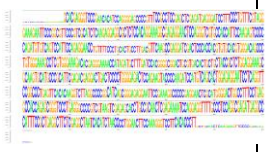
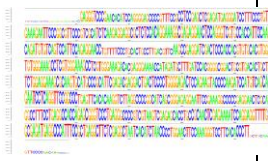
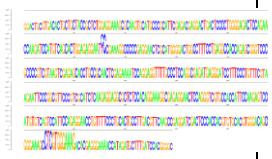
Cluster	Genome Proportion [%]	Size real	Satellite probability	Consensus length	Consensus	Kmer analysis	Graph layout	Connected component index C	Pair completeness index P	Kmer coverage
CL2	5.7	37854	0.993	803	CCGTCTCCACTCGAGAG GAAGCACGAGGGTCCCG AACACATCCAGGGGAGC CCC--------- CGAGGCTAATCACATCT AACCCCTGGAACTTCCA AAGGGTCCTTCACACCC TTTCTGCAACTCAAGAA GTTCCCGACA CAC	report		0.988	0.965	0.736
CL4	1.7	11133	0.993	702	AGCCTGCTCCTTCCTGG GAAGGAAGCCTAGCTGT GAGGCAAGTCTGGGCAA GGCTCG--------- AAGGCAAGGCAGGAGCG GAGCTGAGGTCAGGAAG GCGCCAGTGGCGTCTCC CAGCCGCGCGC	report		0.994	0.963	0.754

Putative satellites (low confidence)

Cluster	Genome Proportion [%]	Size real	Satellite probability	Consensus length	Consensus	Kmer analysis	Graph layout	Connected component index C	Pair completeness index P	Kmer coverage
CL7	**0.770**	**5139**	**0.0140**	**44**	GGGCCCCACGGGGAAATCACGTG GGCCCCACGGGGAAATCACGT	report		0.707	0.761	0.489
CL16	**0.019**	**125**	**0.0308**	**716**	AAGCTCAGGGATAACTGCCTCTAA GAAAGGAATGGCAAGAAGCGCTA ATCTTGGAGGTTGGGACAGGGAG TT--------- GTGGGGAAAGAAAGACTTCTTCG GGGAAGAAGAGGCAACCACCAGC AAAGGGGATAGAGGAGATGTATG GAGGAGTCTTTGCCTTTCCCTGTG AGTGTGGGGCTACTCCCTACGGCC CC	report		0.760	0.689	0.545

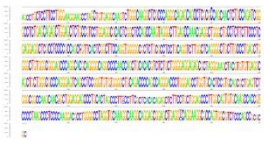
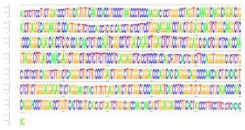
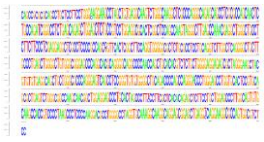
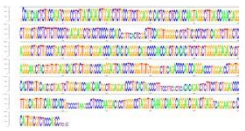
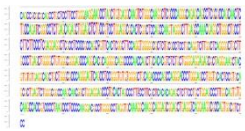
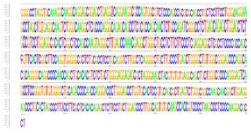
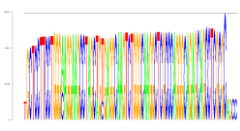
Report of Satellite I

≠	kmer	variant	TS	Monomer_length	score_bn	consensus	graph_image	logo_image
1	**15**	**2**	**0.7356**	**803**	**6.4e-04**	CCGTCTCCACTCGAGAGGAAGCACGAGGGTCCCGAACACAT CCAGGGGAGCCCCGTTTCCGCCTCCGAGCTCGAGATGAG GGATCCTTTCCCTGTTTCGTAGGGAAAGAATTCCCGGCGTT CCCGTCGCATCTCAAGAGGAGGCGCTCTCCACAGGAAAG --------- CGAGAGCACCTGCCGCAACTCGAGAAAATCCAGGAGGTTTT GCCCTCCAGGCGAGATGAGGCCCATTTCCGCTGAGGCTT CTCGAGGCTAATCACATCTAACCCCTGGAACTTCCAAAGGG TCCTTCACACCCTTTCTGCAACTCAAGAAGTTCCCGACA CAC		
2	**19**	**2**	**0.6931**	**803**	**6.2e-04**	CTCCACTCGAGAGGAAGCACGAGGGTCCCGAACACATCCAG GGGAGCCCCGTTTCCGCCTCCGAGCTCGAGATGAGGGAT --------- AGGCTAATCACATCTAACCCCTGGAACTTCCAAAGGGTCCT TCACACCCTTTCTGCAACTCAAGAAGTTCCCGACACACC CGT		

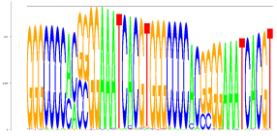
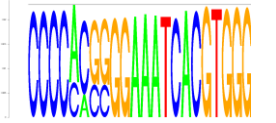
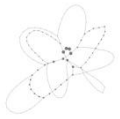
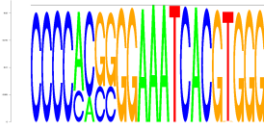
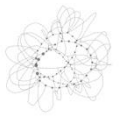
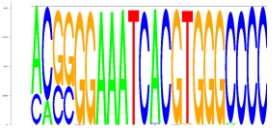
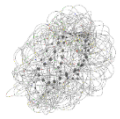
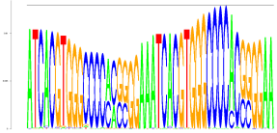
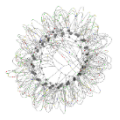
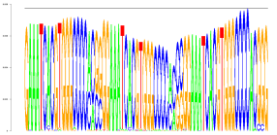
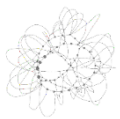
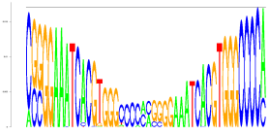
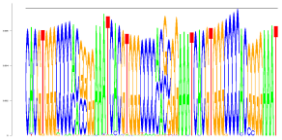
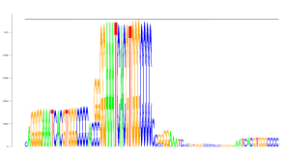
3	15	3	0.6891	804	1.0e-04	GTCTCCACTCGAGAGGAAGCACGAGGGTCCCGAACACATCC AGGGGAGCCCCGTTTCCGCCTCCGAGCTCGAGATGAGGG --------- AGAGCACCTGCCGCAACTCGAGAAAATCCAGGAGGTTTTGC CCTCCAGGCGAGATGAGGCCCATTTCCGCTGAGGCTTCT CGAGGCTAATCACATCTAACCCCTGGAACTTCCAAAGGGTC CTTCACACCCTTTCTGCAACTCAAGAAGTTCCCCGACAT ACCC		
4	23	2	0.6534	803	6.0e-04	AACCATAGATGCTTTGATCCACGGGGCGGACTGCGTGACAC TGCTGCTACCGCTCTGGAGGAAAGCGCAAGTGCATGCCC --------- AGGGTCGTGCCACCATTCCAAGAGTCCCCCAGATGTGTCAG TCCATTCCAGAGGAACCTGTTTTCCCTGCACTGCCTTGA CGTTCAAGCCGAGGGATCGACTCCCACCACGTGTGCACGTGG GACAGCCCTGTGGGAAAGCCTCGTGGGAAAGACACGAGG GAA		
5	19	3	0.6410	804	8.5e-05	TCTCCACTCGAGAGGAAGCACGAGGGTCCCGAACACATCCA GGGGAGCCCCGTTTCCGCCTCCGAGCTCGAGATGAGGGA --------- GAGCACCTGCCGCAACTCGAGAAAATCCAGGAGGTTTTGCC CTCCAGGCGAGATGAGGCCCATTTCCGCTGAGGCTTCTC GAGGCTAATCACATCTAACCCCTGGAACTTCCAAAGGGTCC TTCACACCCTTTCTGCAACTCAAGAAGTTCCCCGACATA CCCG		
6	27	1	0.6341	816	5.7e-04	GAGAGGAAGCACGAGGGTCCCGAACACATCCAGGGGAGCCC CGTTTCCGCCTCCGAGCTCGAGATGAGGGATCCTTTCCC TGTTCGTAGGGAAAGAATTCCCGGCGTTCCCGTCCGCATCT CAAGAGGAGGCGCTCTCCACAGGAAAGGCGAGAGGAACT CCAGGGTCGTGCCACCATTCCAAGAGTCCCCCAGATGTGTC AGTCCATTCCAGAGGAACCTGTTTTCCCTGCACTGCCTT --------- TCTCGAGGCTAATCACATCTAACCCCTGGAACTTCCAAAGG GTCCTTCACACCCTTTCTGCAACTCAAGAAGTTCCCGAC ACACCCGTCTCCACTC		
7	23	4	0.6109	816	6.4e-05	GGGAAAGACACGAGGGAAAACCATAGATGCTTTGATCCACG GGGCGGACTGCGTGACACTGCTGCTACCGCTCTGGAGGA AAGCGCAAGTGCATGCCCGCATTCGAGACGAGGACTGACTC CCCTGGGGAGACTCCAGAAGTACCCCAAGATCCATGTCA --------- TGAGGGATCCTTTCCCTGTTTCGTAGGGAAAGAATTCCCGG CGTTCCCGTCGCATCTCAAGAGGAGGCGCTCTCCACAGG AAAGGCGAGAGGAACTCCAGGGTCGTGCCACCATTCCAAGA GTCCCCCAGATGTGTCAGTCCATTCCAGAGGAACCTGTT TTCCCTGCACTGCCTTGACGTTCAAGCCGAGGATCGACTCC CACCACGTGTGCACGTGGGACAGCCCTGTGGGAAAGCCT CGTGGGAAAGCCCGT		
8	23	3	0.5905	804	7.4e-05	TCTCCACTCGAGAGGAAGCACGAGGGTCCCGAACACATCCA GGGGAGCCCCGTTTCCGCCTCCGAGCTCGAGATGAGGGA --------- GGCCCCCGAGGAACTCGCATGGGGACTGGCCTTTCCTGAGG CCACCAGAGCGGGTCCCTGAGGGCCCCGTCGTAAGTCGA GAGCACCTGCCGCAACTCGAGAAAATCCAGGAGGTTTTGCC CTCCAGGCGAGATGAGGCCCATTTCCGCTGAGGCTTCTC GAGGCTAATCACATCTAACCCCTGGAACTTCCAAAGGGTCC TTCACACCCTTGCTGCAACTCAAGAAGTTCCCCGACATA CCCG		
9	27	2	0.5667	817	6.0e-05	TCTCCACTCGAGAGGAAGCACGAGGGTCCCGAACACATCCA GGGGAGCCCCGTTTCCGCCTCCGAGCTCGAGATGAGGGA --------- CGTCGTAAGTCGAGAGCACCTGCCGCAACTCGAGAAAATCC AGGAGGTTTTGCCCTCCAGGCGAGATGAGGCCCATTTCC GCTGAGGCTTCTCGAGGCTAATCACATCTAACCCCTGGAAC TTCCAAAGGGTCCTTCACACCCTTGCTGCAACTCAAGAA GTTCCCCGACATACCCG		
10	11	5	0.5535	558	6.9e-04	GGACTGCGTGACACTGCTGCTACCGCTCTGGAGGAAAGCGC AAGTGCATGCCCGCATTCGAGACGAGGACTGACTCCCCT --------- ATGTGTCAGTCCATTCCAGAGGAACCTGTTTTCCCTGCACT GCCTTGACGTTCAAGCCGAGGATCGACTCCCACCACGTG TGCACGTGGGACAGCCCTGTGGGAAAGCCTCGTGGGAAAGA CACGAGGGAAAACCATAGATGCTTTGATCCACGGGGC		
11	11	3	0.1750	183	6.6e-04	ACACCCGTCTCCACTCGAGAGGAAGCACGAGGGTCCCGAAC ACATCCAGGGGAGCCCCGTTTCCGCCTCCGAGCTCGAGA TGAGGCCCATTTCCGCTGAGGCTTCTCGAGGCTAATCACAT CTAACCCCTGGAACTTCCAAAGGGTCCTTCACACCCTTT CTGCAACTCAAGAAGTTCCCGAC		
12	11	4	0.1563	184	1.2e-04	GTCTCCACTCGAGAGGAAGCACGAGGGTCCCGAACACATCC AGGGGAGCCCCGTTTCCGCCTCCGAGCTCGAGATGAGGC CCATTTCCGCTGAGGCTTCTCGAGGCTAATCACATCTAACC CCTGGAACTTCCAAAGGGTCCTTCACACCCTTTCTGCAA CTCAAGAAGTTCCCCGACATACCC		

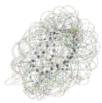
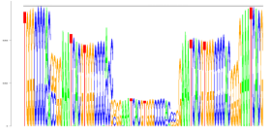
13	11	2	0.0617	62	8.9e-04	ATCCTCAGGTTCCGGCCCTGAGTTCACACAAGGTCTTAGGC CCCGGCATCAGCGGGAGAGGA		
14	11	1	0.0230	13	1.0e-03	CCTCGTGGGAAAG		
15	15	1	0.0179	13	7.9e-04	GTGGGAAAGCCTC		
16	19	1	0.0134	13	7.0e-04	TGGGAAAGCCTCG		
17	23	1	0.0093	13	6.8e-04	GAAAGCCTCGTGG		

Report of Satelllite II

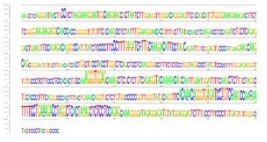
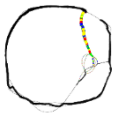
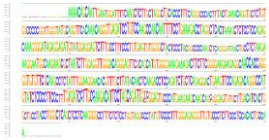
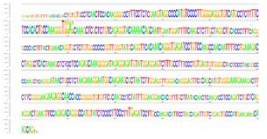
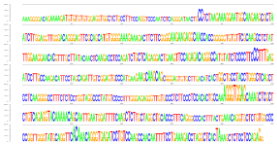
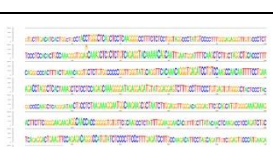
≠	kmer	variant	total_ score	monomer_ length	score_bn	consensus	graph_image	logo_image
1	11	3	0.754	702	0.00065	AGCCTGCTCCTTCCTGGGAAGGAAGCCTAGC TGTGAGGCAAGTCTGGGCAAGGCTCGCCCAG GGCAGAGCAGCCTCGCGC --------- CTCCGTGCAGCGCTGAGTCTGGGCGGCCAAG CAACGGCTGGAGGAGCCCTGCACTAGCCCTT TCCTTCGCTCACACAGAG CTCTTCCTGCTGGAGCCCTTGGCAGTGTGCA AGCCACCGTCCCCTAAGCCCTCCCCAAGGAC GCCTGGGGCCTAGGTGCA AAGGCAAGGCAGGAGCGGAGCTGAGGTCAGG AAGGCGCCAGTGGCGTCTCCCAGCCGCGCGC		
2	15	1	0.712	702	0.00071	AGCTCTTCCTGCTGGAGCCCTTGGCAGTGTG CAAGCCACCGTCCCCTAAGCCCTCCCCAAGG ACGCCTGGGGCCTAGGTG --------- CCCCAACCAGCTGCACACGTCTCTGCTGGGG AGGAGAGCGCTAGGGAAGCTACGTGTGTGAG GCAGCTGCTGGGCGCCCA GGGAGTCAGCCTCCGGGTGTGTGGGGCTGCA GACCCAGCCAGGGACCCTGGGAGCCTGGTG CACTCCGTGCAGCGCTGA GTCTGGGCGGCCAAGCAACGGCTGGAGGAGC CCTGCACTAGCCCTTTCCTTCGCTCACACAG		
3	19	1	0.667	702	0.00070	CAGCCGCGCGCAGCCTGCTCCTTCCTGGGAA GGAAGCCTAGCTGTGAGGCAAGTCTGGGCAA GGCTCGCCCAGGGCAGAG --------- AGCCTGGTGCACTCCGTGCAGCGCTGAGTCT GGGCGGCCAAGCAACGGCTGGAGGAGCCCTG CACTAGCCCTTTCCTTCG CTCACACAGAGCTCTTCCTGCTGGAGCCCTT GGCAGTGTGCAAGCCACCGTCCCCTAAGCCC TCCCCAAGGACGCCTGGG GCCTAGGTGCAAAGGCAAGGCAGGAGCGGAG CTGAGGTCAGGAAGGCGCCAGTGGCGTCTCC		
4	11	2	0.638	622	0.00030	GCCAGCAACTCTGCCAGATCGGGCCTCTGAG CAGAGCTGGAGCTCTGCCGTCCTGGAGGCGC ACTCGGCTCCAGCCAAGT --------- CACTCCGTGCAGCGCTGAGTCTGGGCGGCCA AGCAACGGCTGGAGGAGCCCTGCACTAGCCC TTTCCTTCGCTCACACAG AGCTCTTCCTGCTGGAGCCCTTGGCAGTGTG CAAGCCACCGTCCCCTAAGCCCTCCCCAAGG ACGCCTGGGGCCTAGGTG CAAAGGCAAGGCAGGAGCGGAGCTGAGGTCA GGAAGGCGCCAGTGGCGTCTCCCAGCCTCGC		
5	23	1	0.629	702	0.00064	CAGCCGCGCGCAGCCTGCTCCTTCCTGGGAA GGAAGCCTAGCTGTGAGGCAAGTCTGGGCAA GGCTCGCCCAGGGCAGAG --------- CTCACACAGAGCTCTTCCTGCTGGAGCCCTT GGCAGTGTGCAAGCCACCGTCCCCTAAGCCC TCCCCAAGGACGCCTGGG GCCTAGGTGCAAAGGCAAGGCAGGAGCGGAG CTGAGGTCAGGAAGGCGCCAGTGGCGTCTCC		
6	27	1	0.597	702	0.00065	GGGGCCTAGGTGCAAAGGCAAGGCAGGAGCG GAGCTGAGGTCAGGAAGGCGCCAGTGGCGTC TCCCAGCCGCGCGCAGCC --------- GTGCAGCGCTGAGTCTGGGCGGCCAAGCAAC GGCTGGAGGAGCCCTGCACTAGCCCTTTCCT TCGCTCACACAGAGCTCT TCCTGCTGGAGCCCTTGGCAGTGTGCAAGCC ACCGTCCCCTAAGCCCTCCCCAAGGACGCCT		
7	11	1	0.091	80	0.00025	TGCTCCTTCCTGGGAAGGAAGCCTAGCTGTG AGGCAAGTCTGGGCAAGGCTCGCCCAGGGCA GAGCAGCCTCGCGCAGCC		

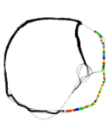
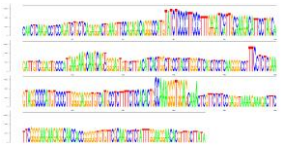
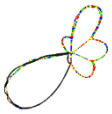
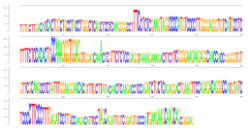
Report of Novel Tandem_44bp repeat

≠	kmer	variant	Total score	monomer_ length	score_bn	consensus	graph_image	logo_image
1	23	4	0.48917	44	8.5e-03	GGGCCCCACGGGGAAATCA CGTGGGCCCCACGGGGAAA TCACGT		
2	11	1	0.47029	22	2.0e-02	CCCCACGGGGAAATCACGT GGG		
3	15	1	0.45403	22	2.0e-02	CCCCACGGGGAAATCACGT GGG		
4	19	1	0.43331	22	1.9e-02	ACGGGGAAATCACGTGGGC CCC		
5	27	3	0.42643	44	7.1e-03	ATCACGTGGGCCCCACGGG GAAATCACGTGGGCCCCAC GGGGAA		
6	23	5	0.42223	66	4.1e-03	GAAATCACGTGGGCCCCAC GGGGAAATCACGTGGGCCC CACCCGGAAATCACGTGGG CCCCACGGG		
7	19	2	0.41495	44	3.2e-03	CGGGGAAATCACGTGGGCC CCACGGGGAAATCACGTGG GCCCCA		
8	27	4	0.40252	66	4.8e-03	CACGTGGGCCCCACGGGGA AATCACGTGGGCCCCACGGC GGAAATCACGTGGGCCCCA CGGGGAAAT		
9	23	6	0.30673	88	1.3e-04	CAGGGGAAATCACGTGGGC CCCACGGGGAAATCACGTG GGCCCCCCGGGGAAATCCC GTGGGCCCCCCGGGGAAAT CACG TGGGCCCC		

10	27	5	0.30497	88	6.3e-04	TGGGCCCCACGGGGAAATC ACGTGGGCCCCACCGGGAA ATCACGTGGGCCCCACCCG GAAATCACGTGGGCCCCAC GGGG AAATCACG		
11	23	3	0.10836	43	1.1e-05	ACGTGGGCCCCACCCGGAA ATCACGTGGGCCCCACGGG AAATC		
12	27	2	0.08898	43	1.3e-05	GGGAAATCACGTGGGCCCC ACGGGGAAATCACGTGGGC CCACG		
13	23	1	0.01255	21	1.4e-05	GTGGGCCCACGGGGAAATC AC		
14	27	1	0.00365	22	7.9e-06	GGAAATCACGTGGGCCCCA CCG		
15	23	2	0.00072	22	9.2e-06	GCACGTGGGCCCCCAGGGG AAA		

Report of of Putative Sat_716

≠	kmer	variant	Total score	monomer _length	score_bn	consensus	graph_image	logo_image
1	15	2	0.545	716	4.7e-04	AAGCTCAGGGATAACTGCCTCTAAGAAAGGAATG GCAAGAAGCGCTAATCTTGGAGGTTGGGACAGGG AGTTCGCAGCAT ---------- TTTTCAAGCTCTTGCTAGGCTGCACCCTTTCAGG GCCCACTTACTGAAACAGGTGCTCTGTGGCCCC CGGTTGGGTATG CAGGTTCACAAACAGGGTGAGATCCTGTCCAAGC CAACAATTTTCCTGAAAGACCTAGGCTCACTAAA GCTCTCCTCCAG AGCAAGGGGATAGAGGAGATGTATGGAGGAGTC TTTGCCTTTCCCTGTGAGTGTGGGGCTACTCCCT ACGCCCC		
2	15	3	0.503	723	4.7e-04	AAGCTCAGGGATAACTGCCTCTAAGAAAGGAATG GCAAGAAGCGCTAATCTTGGAGGTTGGGACAGGG AGTTCGCAGCAT ---------- GGGTATGCAGGTTCACAAACAGGGTGAGATCCTG TCCAAGCCACACAATTTTCCTGAAAGACCTAGGCT CACTAAAGCTCT CCTCCAGAGCAAAGGGGATAGAGGAGATGTATGG AGGAGTCTTTGCCTTTCCCTGTGAGTGTGGGGCT ACTCCCTACGGC CCC		
3	27	2	0.497	716	6.5e-05	TCGCTCTGTCAGAGGTACAAAAACAGCAATTGAA TGGATTTTCAAGCTCTTGCTAGGCTGCACCCTTT CAGGGCCCACTT---------- CATGAAGAAGCAAGCAGCGGGAGTGTGCTTGACA TCACTGGCTGCCTACCTGGGCTCAGCTCCTCAAG GGGCCTTTCTCT CCTGGTAGGCCCTATGTCCCCTTTGGGAGAGGGT TGTGCCCTCTTCCCTCCACACTGCCAACGGGTGC TGCAAAGC		
4	19	1	0.491	707	2.4e-04	GGAGTGTGTTTGACACGACTGTGTGCCTCAACTC CACAAGGGGCCTTTCCTCTCGAAGTAGGCCCCAT GTCCCCTTGGGG ---------- ATGTGGGGAAAGAAAGACTTCTTCGGGGAAGAAG AGGCAACCACCGGGGGTGTGTTCGCAAGCCTCTA TTTGGAAGGAAC ACGTTTGCTTATACAACTCAAGACCTCCAGATCT GCTCAGAGGACTGAAGTTCCAGAGACAGGGCCAT GTATCTCCCCTT CCCTTTGAGATCCTTGCCAAGACATTCCTAGCAG ATTGTCGGAGTGCCCATGAAGAAGCAAGCAGTA		
5	19	2	0.481	716	4.3e-04	GGAGTGTGCTTGACATCACTGGCTGCCTACCTGG GCTCAGCTCCTCAAGGGGCCTTTCTCTCCTGGTA GGCCCTATGTCC ---------- GGAAGGAACACGTTTGCTTATACAACTCAAGACC TCCAGATCTGCTCAGAGGACTGAAGTTCCAGAGA CAGGGCCATGTA TCTCCCCTTCCCTTTGAGATCCTTGCCAAGACATT CCTAGCAGATTGTCGGAGTGCCCATGAAGAAGC AAGCAGCG		
6	15	1	0.476	696	4.1e-04	AAAGGGGACAGAAAAGATGTGTGTGTGGAGGTGG CTCTGCCTTTCCAGGTCCCAATCTCAGGGATAAC TGCCTCTAAGAA AGGAATGGCAAGAAGCGCTAATCTTGGAGGTTGG GACAGGGAGTTCGCAGCATGTGGGGAAAGAAAGA CTTCTTCGGGGA ---------- AGGGTGGAGCAAAGCTCGCTCTGTCAGAGGTACA AAAACAGCAATTGAATGGATTTTCAAGCTCTTGC TAGGCTGCACCC TTTCAGGGCCCACTTTACTGAAACAGGTGCTCTG TGGCCCCGGTTGGGTATGCAGGTTCACAAACAG GGTGAGATCCTG TCCAAGCCAACAATTTTCCTGAAAGACCTAGGCT CACTAAAGCTCTCCTCCAGAGG		
7	23	1	0.467	716	3.8e-04	GTGCTTGACATCACTGGCTGCCTACCTGGGCTCA GCTCCTCAAGGGGCCTTTCTCTCCTGGTAGGCCC TATGTCCCCTTT ---------- CACGATGTGGGGAAAGAAAGACTTCTTCGGGGAA GAAGAGGCAACCACCGGGGGTGTGTTCGCAAGCC TCTATTTGGAAG GAACACGTTTGCTTATACAACTCAAGACCTCCAG ATCTGCTCAGAGGACTGAAGTTCCAGAGACAGGG CCATGTATCTCC CCTTCCCTTTGAGATCCTTGCCAAGACATTCCTA GCAGATTGTCGGAGTGCCCATGAAGAAGCAAGCA GCGGGAGT		

8	27	1	0.440	707	6.5e-05	ACTGAAGTTCCAGAGACAGGGCCGTGTATCTCCC CTTCCCTTTGTGATCATTGCAAAGACATTCCTAG CTGATAGCGGGA GGGCCCATGAAGAAGCAAGCAGTAGGAGTGTGTT TGACACGACTGTGTGCCTCAACTCCACAAGGGGC CTTTCCTCTCGA---------- AGTGTGGGGCTACTCCCTACGGCCCCAAGCTCAG GGATAACTGCCTCTAAGAAAGGAATGGCAAGAAG CGCTAATCTTGG AGGTTGGGACAGGGAGTTCGCAGCATGTGGGGAA AGAAAGACTTCTTCGGGGAAGAAGAGGCAACCAC CGGGGGTGTGTT CGCAAGCCTCTATTTGGAAGGAACACGTTTGCTT ATACAACTCAAGACCTCCAGATCTGCTTAGACG		
9	11	9	0.309	372	5.2e-04	CAACTCAAGACCTCCAGATCTGCTCAGAGGACTG AAGTTCCAGAGACAGGGCCATGTATCTCCCCTTC CCTTTGAGATCC TTGCCAAGACATTCCTAGCAGATTGTCGGAGTGC CCATGAAGAAGCAAGCAGCGGGAGTGTGCTTGAC ATCACTGGCTGC CTACCTGGGCTCAGCTCCTCAAGGGGCCTTTCCT CTCGAAGTAGGCCCCATGTCCCCTTGGGGAGGGT GTCATGCCTCTT TTCTCCACACTGCCAAAGGGTGGAGCAAACTCGT TCTCTCCGAGGTAAAAGAAAGACTTCTTCGGGGA AGAAGAGGCAAC CACCGGGGGTGTGTTCGCAAGCCTCTATTTGGAA GGAACACGTTTGCTTATA		
10	11	11	0.292	378	4.6e-04	TGTGCTTGACATCACTGGCTGCCTACCTGGGCTC AGCTCCTCAAGGGGCCTTTCCTCTCGAAGTAGGC CCCATGTCCCCT TGGGGAGGGTGTCATGCCTCTTTTCTCCACACTG CCAAAGGGTGGAGGTGGAGCAAACTCGTTCTCTC CGAGGTAAAAGA AAGACTTCTTCGGGGAAGAAGAGGCAACCACCGG GGGTGTGTTCGCAAGCCTCTATTTGGAAGGAACA CGTTTGCTTATA CAACTCAAGACCTCCAGATCTGCTCAGAGGACTG AAGTTCCAGAGACAGGGCCATGTATCTCCCCTTC CCTTTGAGATCC TTGCCAAGACATTCCTAGCAGATTGTCGGAGTGC CCATGAAGAAGCAAGCAGCGGGAG		
11	11	2	0.284	243	4.0e-04	GGGACAGAAAAGATGTGTGTGTGGAGGTGGAGCA AAGCTCGCTCTGTCAGAGGTACAAAAACAGCAAT TGAATGGATTTT ---------- TTCACAAACAGGGTGAGATCCTGTCCAAGCCAAC AATTTTCCTGAAAGACCTAGGCTCACTAAAGCTC TCCTCCAGAGGA AAG		
12	11	7	0.283	363	4.6e-04	ACGACTGTGTGCCTCAACTCCACAAGGGGCCTTT CCTCTCGAAGTAGGCCCCATGTCCCCTTGGGGAG GGTGTCATGCCT ---------- AGAGGACTGAAGTTCCAGAGACAGGGCCATGTAT CTCCCCTTCCCTTTGAGATCCTTGCCAAGACATT CCTAGCAGATTG TCGGAGTGCCCATGAAGAAGCAAGCAGTAGGAGT GTGTTTGAC		
13	11	10	0.281	373	4.6e-04	TGTGCTTGACATCACTGGCTGCCTACCTGGGCTC AGCTCCTCAAGGGGCCTTTCCTCTCGAAGTAGGC CCCATGTCCCCT ---------- AAGACATTCCTAGCAGATTGTCGGAGTGCCCATG AAGAAGCAAGCAGCGGGAG		
14	11	8	0.280	371	4.6e-04	TGTGCTTGACATCACTGGCTGCCTACCTGGGCTC AGCTCCTCAAGGGGCCTTTCTCTCCTGGTAGGCC CTATGTCCCCTT ---------- GACATTCCTAGCAGATTGTCGGAGTGCCCATGAA GAAGCAAGCAGCGGGAG		
15	11	5	0.266	344	4.6e-04	GGCCCCAAGCTCAGGGATAACTGCCTCTAAGAAA GGAATGGCAAGAAGCGCTAATCTTGGAGGTTGGG ACAGGGAGTTCG ---------- TGAAAGACCTAGGCTCACTAAAGCTCTCCTCCAG AGCAAAGGGGATAGAGGAGATGTATGGAGGAGTC TTTGCCTTTCCC TGTGAGTGTGGGGCTACTCCCTAC		
16	11	6	0.265	345	1.7e-04	GTGCTTGACATCACTGGCTGCCTACCTGGGCTCA GCTCCTCAAGGGGCCTTTCCTCTCGAAGTAGGCC CCATGTCCCCTT ---------- GACATTCCTAGCTGATAGCGGGAGT		
17	11	4	0.237	324	4.0e-04	GGGACAGAAAAGATGTGTGTGTGGAGGTGGCTCT GCCTTTCCAGGTCCCAATCTCAGGGATAACTGCC TCTAAGAAAGGA ---------- GTTCACAAACAGGGTGAGATCCTGTCCAAGCCAA CAATTTTCCTGAAAGACCTAGGCTCACTAAAGCT CTCCTCCAGAGG AAAG		
18	11	3	0.192	244	4.0e-04	GGGACAGAAAAGATGTGTGTGTGGAGGTGGAGCA AAGCTCGCTCTGTCAGAGGTACAAA---------- ---- GTTCACAAACAGGGTGAGATCCTGTCCAAGCCAA CAATTTTCCTGAAAGACCTAGGCTCACTAAAGCT CTCCTCCAGAGG AAAG		
19	11	1	0.019	27	1.7e-04	CCCATGAAGAAGCAAGCAGCGGGAGTG		

Appendix 4.6

#	Satellite I (CL2)	Consensus length (bp)	GC%
1	11_1_sc_0.0229839_l_13	13	61.5
2	15_1_sc_0.0178866_l_13	13	61.5
3	19_1_sc_0.0133994_l_13	13	61.5
4	23_1_sc_0.00929624_l_13	13	61.5
5	11_2_sc_0.061697_l_62	62	61.3
6	11_3_sc_0.175018_l_183	183	57.9
7	11_4_sc_0.156253_l_184	184	57.6
8	11_5_sc_0.553542_l_558	558	59.3
9	15_2_sc_0.735627_l_803	803	59
10	19_2_sc_0.693086_l_803	803	59
11	23_2_sc_0.653412_l_803	803	59
12	15_3_sc_0.689101_l_804	804	59
13	19_3_sc_0.641034_l_804	804	59
14	23_3_sc_0.590496_l_804	804	59.1
15	27_1_sc_0.634101_l_816	816	59.1
16	23_4_sc_0.610854_l_816	816	59.2
17	27_2_sc_0.56669_l_817	817	59.1

Note; TAREAN code (15_2_sc_0.735627_l_803) = Kmer_variant_total score_monomer length;

Appendix 4.7

#	Satellite II (CL4)	Consensus length (bp)	GC%
1	11_1_sc_0.0911919_l_80	80	66.3
2	11_2_sc_0.638027_l_622	622	68.2
3	11_3_sc_0.753882_l_702	702	68.1
4	15_1_sc_0.712055_l_702	702	68.1
5	19_1_sc_0.667101_l_702	702	68.1
6	23_1_sc_0.629048_l_702	702	68.1
7	27_1_sc_0.597494_l_702	702	68.1

Appendix 4.8

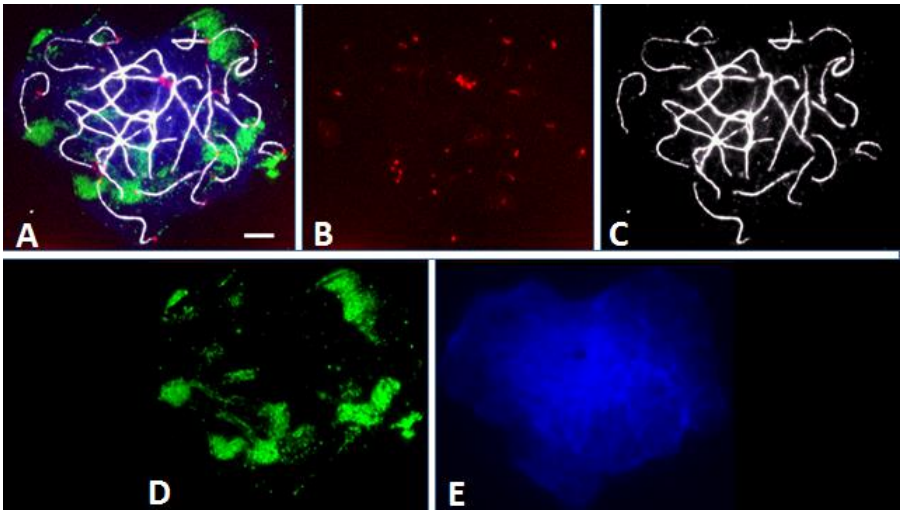
#	Novel tandem_44 repeat (CL7)	Consensus length (bp)	GC%
1	23_1_sc_0.0125539_l_21	21	66.7
2	11_1_sc_0.470287_l_22	22	68.2
3	15_1_sc_0.454034_l_22	22	68.2
4	19_1_sc_0.433314_l_22	22	68.2
5	27_1_sc_0.00365483_l_22	22	68.2
6	23_2_sc_0.000717193_l_22	22	72.7

7	23_3_sc_0.108365_l_43	43	67.4
8	27_2_sc_0.088984_l_43	43	67.4
9	23_4_sc_0.489174_l_44	44	68.2
10	27_3_sc_0.42643_l_44	44	68.2
11	19_2_sc_0.414954_l_44	44	68.2
12	23_5_sc_0.422227_l_66	66	68.2
13	27_4_sc_0.402522_l_66	66	68.2
14	23_6_sc_0.306728_l_88	88	71.6
15	27_5_sc_0.304965_l_88	88	68.2

Appendix 4.9

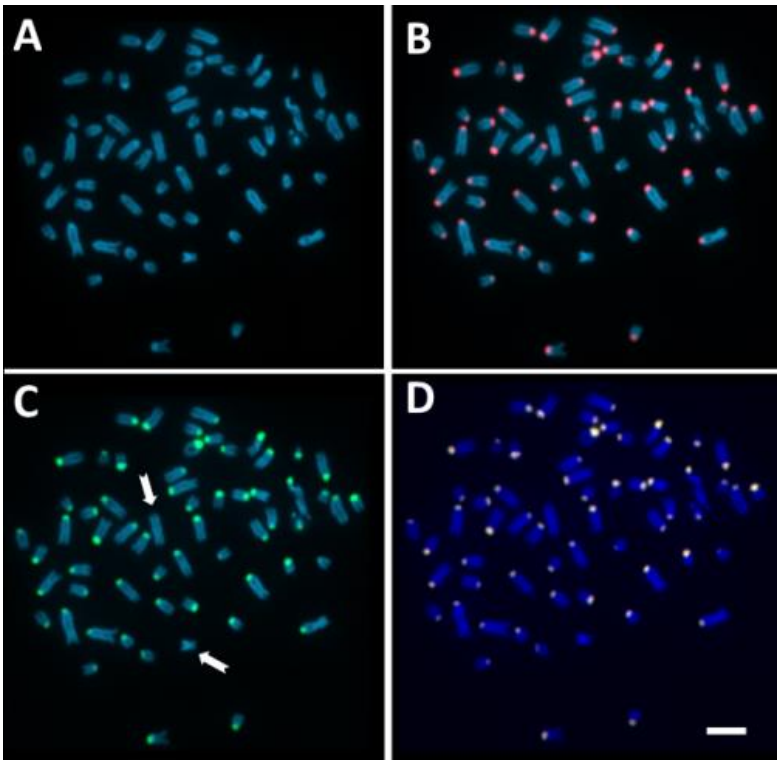
#	**Putative satellite_716bp (CL16)**	**Monomer length (bp)**	GC%
1	11_1_sc_0.0194481_l_27	27	59.3
2	11_2_sc_0.284436_l_243	243	51.4
3	11_3_sc_0.191613_l_244	244	51.2
4	11_4_sc_0.237278_l_324	324	51.9
5	11_5_sc_0.266479_l_344	344	52.6
6	11_6_sc_0.265389_l_345	345	53.6
7	11_7_sc_0.283002_l_363	363	52.3
8	11_8_sc_0.279904_l_371	371	53.6
9	11_9_sc_0.308646_l_372	372	53.8
10	11_10_sc_0.281338_l_373	373	54.2
11	11_11_sc_0.292181_l_378	378	54
12	15_1_sc_0.475577_l_696	696	52.7
13	19_1_sc_0.490658_l_707	707	52.2
14	27_1_sc_0.439959_l_707	707	52.2
15	15_2_sc_0.545331_l_716	716	53.1
16	27_2_sc_0.496598_l_716	716	53.2
17	19_2_sc_0.48092_l_716	716	53.1
18	23_1_sc_0.4667_l_716	716	53.1
19	15_3_sc_0.502569_l_723	723	53.4

Appendix 4.10



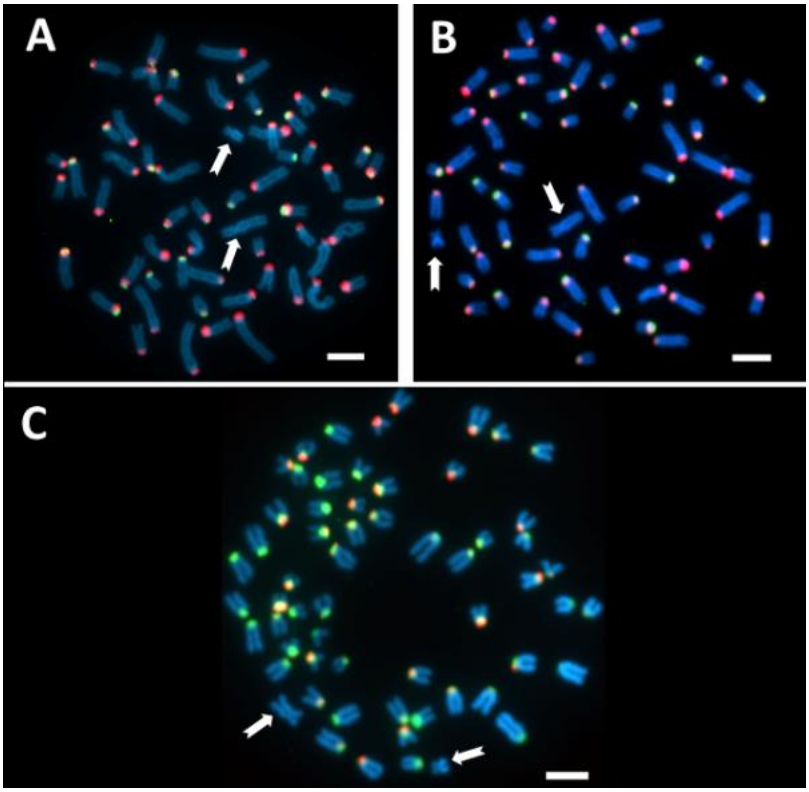
Appendix 4.10 FISH and Immunostaining results of sheep SC spreads at pachytene stage using SCP1 rabbit antibody and probe of satellite I sequences SatI_AJ of sheep. (A) Overlay of the signals including SCPI in white and SatI_AJ in green. (B) Hybridization patterns of the telomeric probe Telomeric_Tndm labelled with biotin-16-dUTP detected by Alexa 594 Streptavidin. (C) SCP1 signal detected with Antirabbit Alexa 594 (Seen in white). (D) Hybridization patterns of the probe satellite I SatI_AJ labelled with digoxigenin-11-dUTP detected by fluorescein-isothiocyanate (FITC; green signal). (E) Dapi stained nucleus (indicate presence of only single cell). Bar equals 5µm.

Appendix 4.11



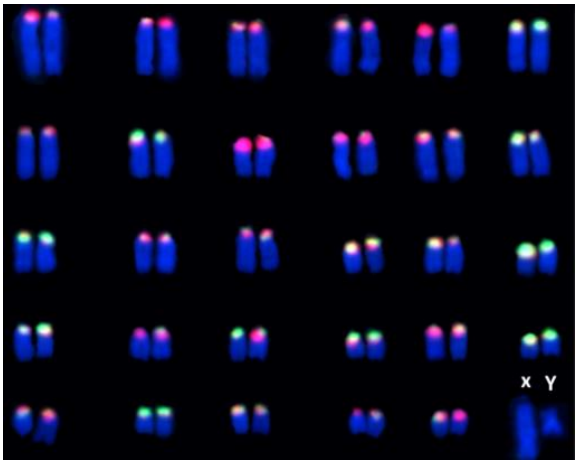
Appendix 4.11 FISH of satellite I probes (Sat1-2 and Sat1-4) to metaphase spreads of male cattle chromosomes (Bos taurus; 2n=60, XY). (A) Chromosomes were stained with DAPI (seen in blue). (B) Probe SatI-2 was labelled with biotin-16-dUTP detected by Alexa 594 Streptavidin (red signal). (C) Probe SatI-4 was labelled with digoxigenin-11-dUTP detected by fluorescein-isothiocyanate (FITC; green signal). (D) Overlay of red and green signal in D showing in yellow where signals overlap. The sex chromosomes are indicated by arrows. Bar equals 5µm.

Appendix 4.12



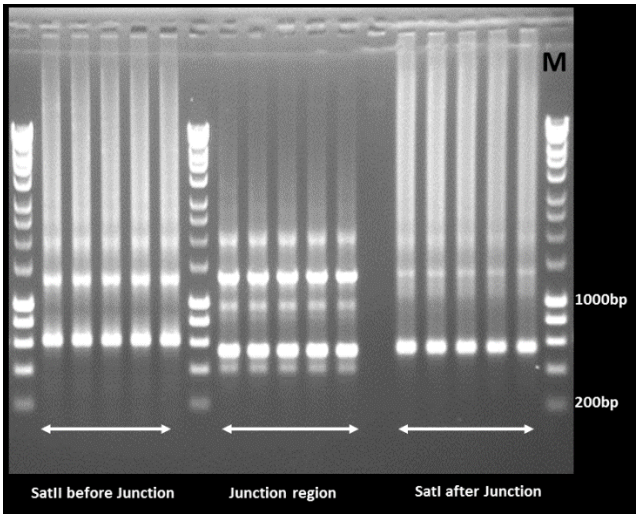
Appendix 4.12 FISH (A&B) cattle satellite I and satellite IV probes (SatI-4 and SatIV-5) to metaphase spreads of male cattle chromosomes (*Bos taurus*; 2n=60, XY). Chromosomes were stained with DAPI (seen in blue). Probe SatI-4 was labelled with biotin-16-dUTP detected by Alexa 594 Streptavidin (red signal). While, probe SatIV-5 was labelled with digoxigenin-11-dUTP detected by fluorescein-isothiocyanate (FITC; green signal). (C) Cattle satellite I and satellite IV probes (SatI-2 and SatIV-5) to metaphase spreads of male cattle chromosomes (*Bos taurus*; 2n=60, XY). Probe SatI-2 was labelled with digoxigenin-11-dUTP detected by fluorescein-isothiocyanate (FITC; green signal). While, probe SatIV-5 was labelled with biotin-16-dUTP detected by Alexa 594 Streptavidin (red signal). The sex chromosomes are indicated by arrows. Bar equals 5µm.

Appendix 4.13



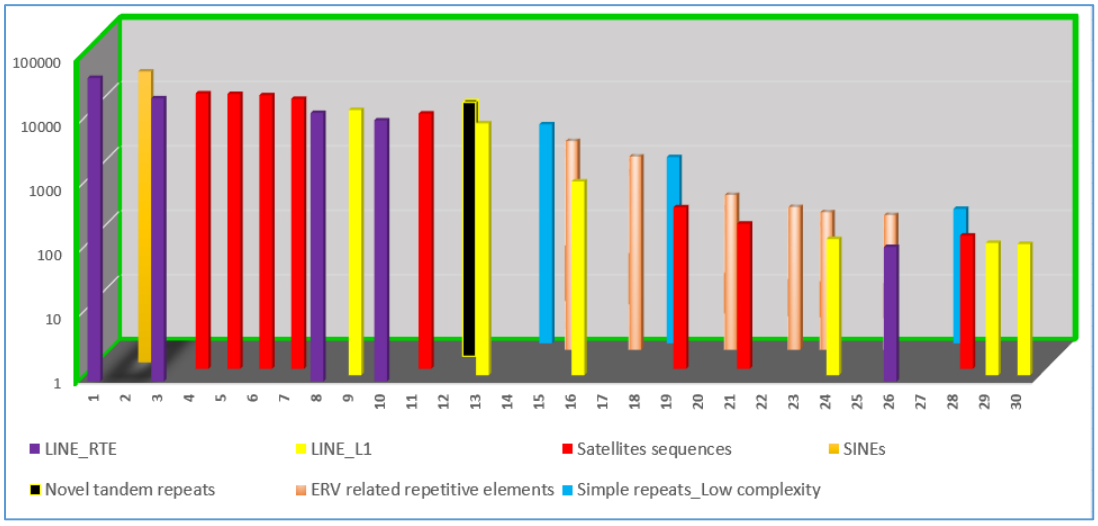
Appendix 4.13 FISH Karyotype representing cattle satellite I and satellite IV probes (SatI-4 and SatIV-5) hybridized to metaphase spreads of male cattle chromosomes (Bos taurus; 2n=60, XY). Chromosomes were stained with DAPI (seen in blue). Probe SatI-4 was labelled with biotin-16-dUTP detected by Alexa 594 Streptavidin (red signal). While, probe SatIV-5 was labelled with digoxigenin-11-dUTP detected by fluorescein-isothiocyanate (FITC; green signal). Overlay of red and green signalS showing in yellow where signals overlap. The sex chromosomes are indicated by (X & Y). Bar equals 5µm.

Appendix 4.14



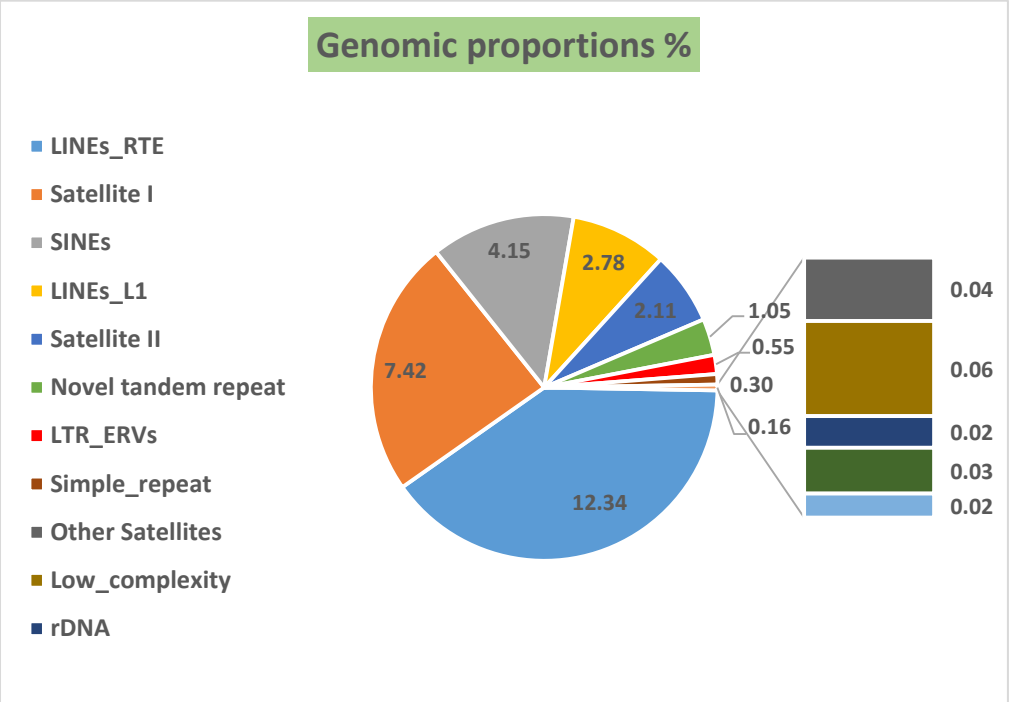
Appendix 4.14 Gel image of PCR products of satellite I and II sequences before and after junction region.

Appendix 5.1

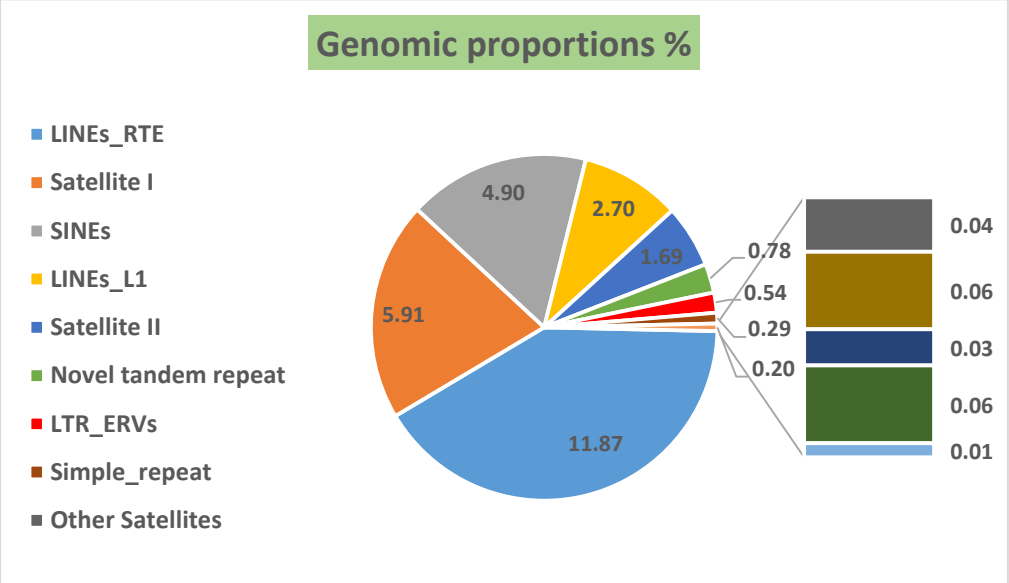


Appendix 5.1 Figure shows size distribution and composition of different repetitive classes within the top 30 clusters. Graphs are coloured according to the available repeat type in each cluster investigated depending on the similarity searchers of Repbase databases.

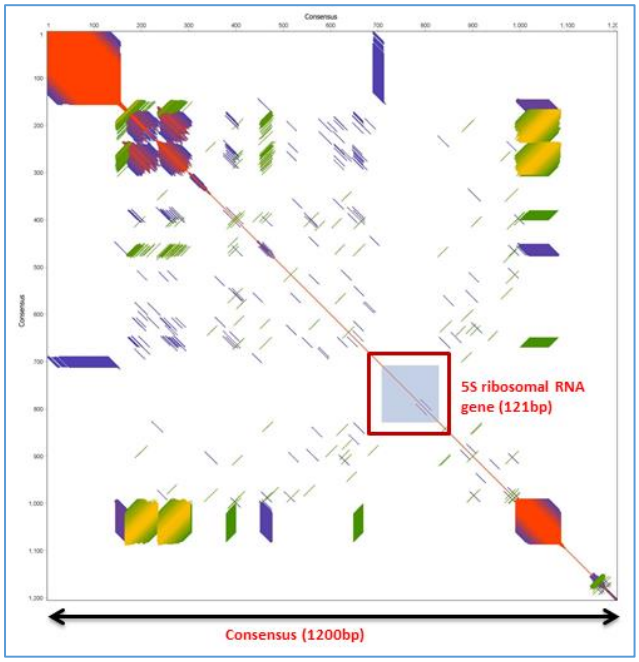
Appendix 5.2 and 5.3



Appendix 5.2 Genomic proportion of major groups of repetitive sequences identified in unassembled raw reads of male sheep genome HamJ1 using RepeatExplorer

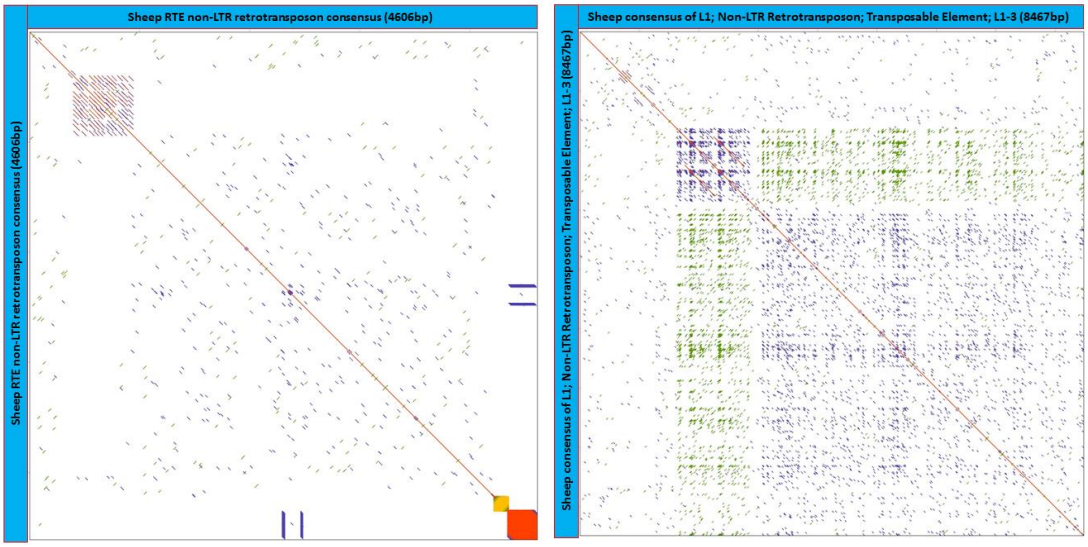


Appendix 5.3 Genomic proportion of major groups of repetitive sequences identified in unassembled raw reads of female sheep genome KarJ using RepeatExplorer

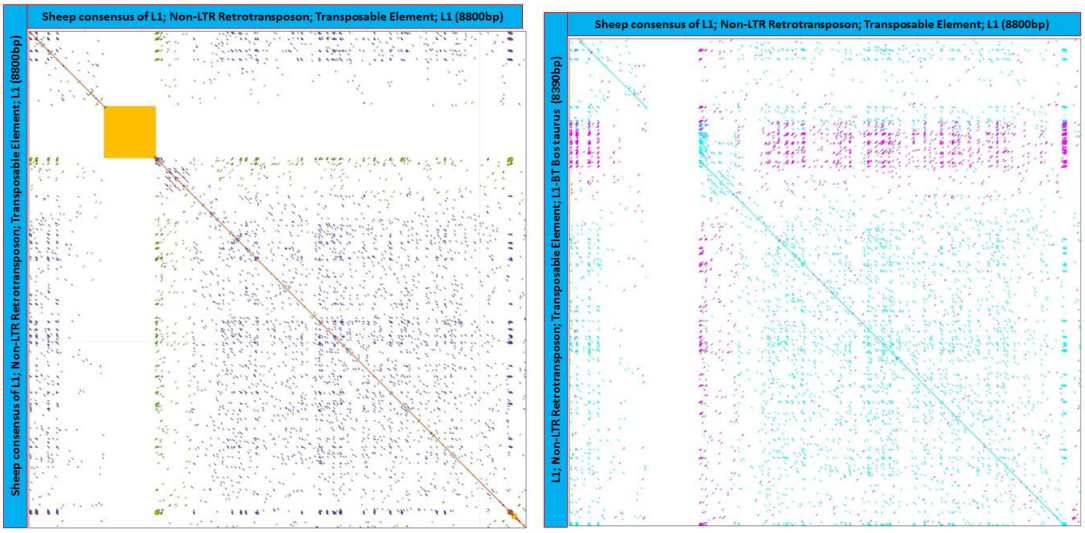


Appendix 5.4 Dot plot (self) consensus of 5S ribosomal RNA gene assembled from mapping of whole paired reads of KarJ genome.

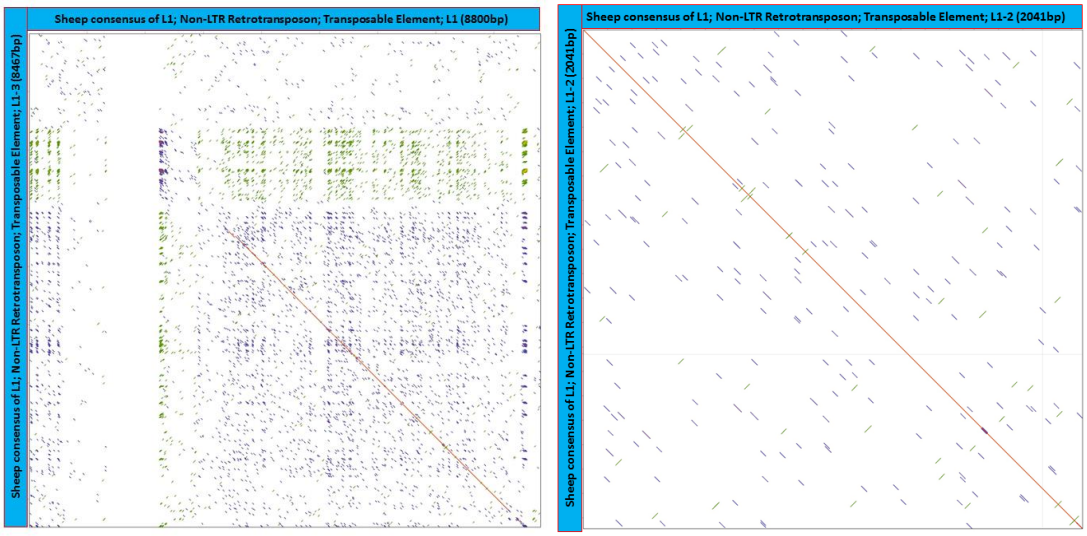
Appendix 5.5 and Appendix 5.6



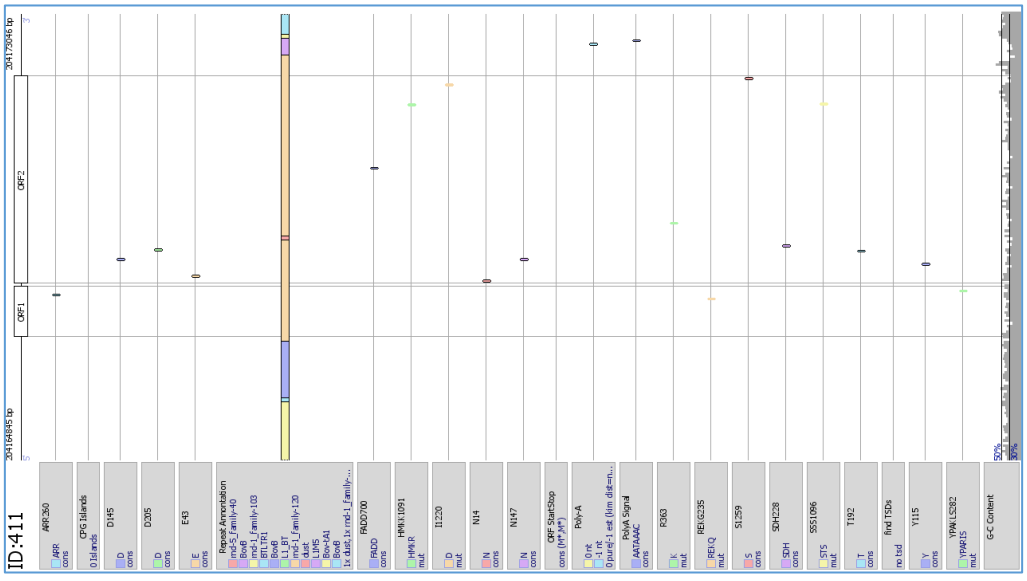
Appendix 5.7 and Appendix 5.8



Appendix 5.9 and Appendix 5.10



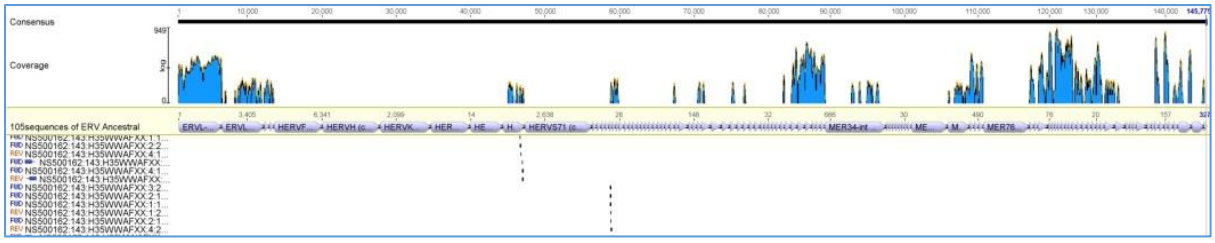
Appendix 5.11 Open reading frames of LINE



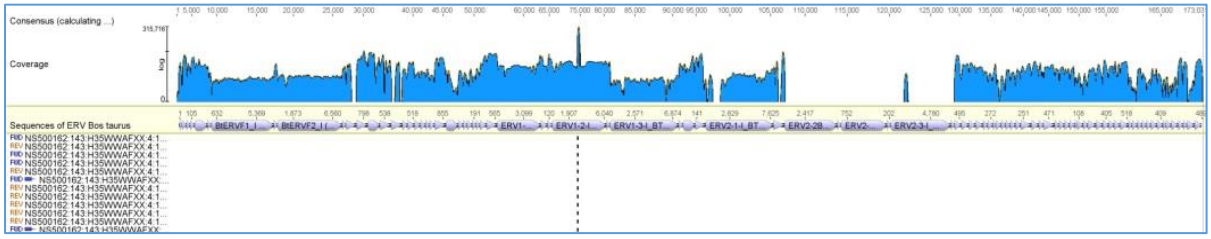
Appendix 6.1 Estimated integration time for each enJSRV provirus based on the differences between the 5’ and 3’ LTR. Arnud et al. 2007

Provirus*	LTR length	nt differences	Estimated Integration (MYA)§
enJSRV-26	413	0	< 0.45
enJSRV-15	447	0	< 0.45
enJSRV-16	446	0	< 0.45
enJSRV-18	446	0	< 0.45
enJSRV-11	447	0	< 0.45
enJSRV-19	446	0	< 0.45
enJSRV-8	446	0	< 0.45
enJS5F16	446	0	< 0.45
enJSRV-20	446	1	0.4 - 0.9
enJSRV-14	446	1	0.4 - 0.9
enJS56A1	446	2	0.9 - 1.9
enJSRV-21	446	3	1.3 - 2.9
enJSRV-6	444	3	1.3 - 2.9
enJS59A1	429	4	1.9 - 4.0
enJSRV-25	446	4	1.8 - 3.9
enJSRV-10	444	5	2.3 - 4.9
enJSRV-2	446	5	2.2 - 4.9
enJSRV-7	446	5	2.2 - 4.9
enJSRV-1	404	6	3.0 - 6.5
enJSRV-9	446	7	3.1 - 6.8
enJSRV-13	446	8	3.6 - 7.8
enJSRV-5	441	8	3.6 - 7.9
enJSRV-3	403	8	4.0 - 8.6

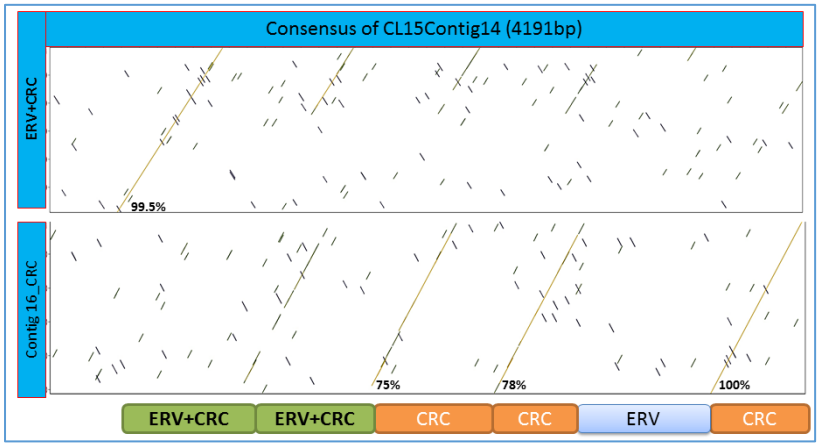
Appendix 6.2 Mapping of whole sequencing of sheep to ancestral sequences of ERVs



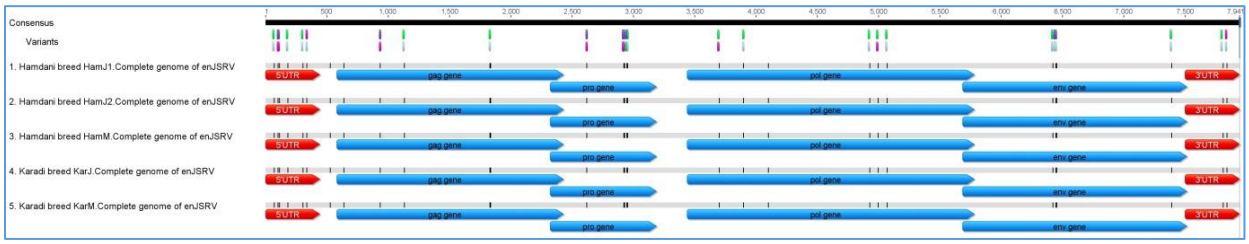
Appendix 6.3 Mapping of whole sequencing of sheep to *Bos taurus* sequences of ERVs



Appendix 6.4 Consensus of CL15C14 (4191bp) of RepeatExplorer including combined ERV1+CRC sequences, three copies of 32merC16_Sat_CRC satellite like sequences (see section 4.4.5.1) and ERVs.



Appendix 6.5 Distribution of 50 SNPs along the complete genome of enJSRV of five samples of Hamdani and Karadi breeds.



Appendix 6.6 Mutalyzer report of SNPs between different combinations of complete genomes of enJSRV of five samples of sheep. HamJ1; HamJ2; HamM; KarJ and KarM

#	positon and variants	HamJ1	HamJ2	positon and variants	HamJ1	HamM	positon and variants	HamJ1	KarJ
1	69T>C	T	C	181C>T	C	T	305C>T	C	T
2	103A>G	A	G	340T>G	T	G	527A>T	A	T
3	112G>A	G	A	935G>A	G	A	642A>C	A	C
4	181C>T	C	T	1130T>C	T	C	1832A>G	A	G
5	340T>G	T	G	2920G>A	G	A	1835T>C	T	C
6	935G>A	G	A	3897T>C	T	C	2620A>G	A	G
7	2620A>G	A	G	4923T>C	T	C	2920G>A	G	A
8	2920G>A	G	A	4989G>C	G	C	2926A>G	A	G
9	2926A>G	A	G	7387T>C	T	C	2943A>C	A	C
10	2953T>C	T	C	7837G>T	G	T	2953T>C	T	C
11	4095T>A	T	A				5067C>T	C	T
12	6445T>A	T	A				6415T>C	T	C
13	7387T>C	T	C				6445T>A	T	A
14	7802T>C	T	C				7387T>C	T	C
15							7802T>C	T	C
16							7837G>T	G	T

#	positon and variants	HamJ1	KarM	positon and variants	KarJ	KarM	positon and variants	HamJ2	KarM	positon and variants	HamM	KarM
1	69T>C	T	C	69T>C	T	C	2953C>T	C	T	69T>C	T	C
2	103A>G	A	G	103A>G	A	G	3699G>C	G	C	103A>G	A	G
3	112G>A	G	A	112G>A	G	A	4095A>T	A	T	112G>A	G	A
4	181C>T	C	T	181C>T	C	T	6415T>C	T	C	1130C>T	C	T
5	340T>G	T	G	305T>C	T	C	6438A>T	A	T	2620A>G	A	G
6	935G>A	G	A	340T>G	T	G	6445A>T	A	T	2926A>G	A	G
7	2620A>G	A	G	527T>A	T	A	7802C>T	C	T	3699G>C	G	C
8	2920G>A	G	A	642C>A	C	A	7837G>T	G	T	3897C>T	C	T
9	2926A>G	A	G	935G>A	G	A				4923C>T	C	T
10	3699G>C	G	C	1832G>A	G	A				4989C>G	C	G
11	6415T>C	T	C	1835C>T	C	T				6415T>C	T	C
12	6438A>T	A	T	2943C>A	C	A				6438A>T	A	T
13	7387T>C	T	C	2953C>T	C	T						
14	7837G>T	G	T	3699G>C	G	C						
15				5067T>C	T	C						
16				6438A>T	A	T						
17				6445A>T	A	T						
18				7802C>T	C	T						

#	positon and variants	HamJ2	KarJ	positon and variants	HamM	KarJ	positon and variants	HamJ2	HamM
1	69C>T	C	T	181T>C	T	C	69C>T	C	T
2	103G>A	G	A	305C>T	C	T	103G>A	G	A
3	112A>G	A	G	340G>T	G	T	112A>G	A	G
4	181T>C	T	C	527A>T	A	T	1130T>C	T	C
5	305C>T	C	T	642A>C	A	C	2620G>A	G	A
6	340G>T	G	T	935A>G	A	G	2926G>A	G	A
7	527A>T	A	T	1130C>T	C	T	2953C>T	C	T
8	642A>C	A	C	1832A>G	A	G	3897T>C	T	C
9	935A>G	A	G	1835T>C	T	C	4095A>T	A	T
10	1832A>G	A	G	2620A>G	A	G	4923T>C	T	C
11	1835T>C	T	C	2926A>G	A	G	4989G>C	G	C
12	2943A>C	A	C	2943A>C	A	C	6445A>T	A	T
13	4095A>T	A	T	2953T>C	T	C	7802C>T	C	T
14	5067C>T	C	T	3897C>T	C	T	7837G>T	G	T
15	6415T>C	T	C	4923C>T	C	T			
16	7837G>T	G	T	4989C>G	C	G			
17				5067C>T	C	T			
18				6415T>C	T	C			
19				6445T>A	T	A			
20				7802T>C	T	C			