2021-08

Genome scan for signatures of adaptive evolution in wild African goat (capra nubiana)

Chebii, Vivien

NM-AIST

https://doi.org/10.58694/20.500.12479/1368

GENOME SCAN FOR SIGNATURES OF ADAPTIVE EVOLUTION IN WILD AFRICAN GOAT (*Capra nubiana)*

Vivien Jepchirchir Chebii

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Life Sciences of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

August, 2021

ABSTRACT

Nubian ibex (*Capra nubiana*) is a wild goat species inhabiting the Sahara and Arabia deserts. *C. nubiana* thrives well in its habitat which is characterized by intense solar radiation, high temperatures, little feed, and water supply. The genetic basis of *C. nubiana* adaptation to its environment remains unknown. Adaptive signatures of evolution in *C. nubiana* genome were investigated using comparative genomics approaches. Paired-end sequence reads of three *C. nubiana* individuals and other publicly available genome data were used for comparative genomic analysis. Positive selection signals were detected from sequence alignment by comparing the rates of synonymous versus non-synonymous substitutions (dN/dS) in orthologous protein-coding genes shared by *C. nubiana* and related species using CodeML program in PAML package. Copy number variations were detected from the sequence data using read-depth method, with the domestic goat genome data acting as the reference. Genes involved in the skin barrier and hair follicle development, such as ATP binding cassette subfamily A member 12 and UV stimulated scaffold protein A, were found to be positively selected, suggesting that *C. nubiana* has evolved adaptive mechanisms to cope with solar radiation and temperature extremes in its environment. Additionally, a DNA repair gene (UV stimulated scaffold protein A) was reported to be under strong selection signals, further supporting the assertion that *C. nubiana* has acquired adaptive mechanisms to deal with possible DNA damages induced by prolonged exposure to solar radiation. Similarly, duplications of viral response genes such as UL16 binding protein 3, Cluster of Differentiation 48, Natural Killer Group 2D ligand 1-like, Bactericidal/permeability-increasing fold containing family A, member 1, and Natural Killer Group 2D ligand 4-like were reported in *C. nubiana*, indicating that it has acquired adaptive strategies to cope with viral stressors in its environment. Additionally, xenobiotic compounds metabolism genes involved in biotransformation (Cytochrome P450 2D6, carboxylesterase 1 and cytochrome P450 family 2 subfamily B member 6), conjugation (UDP Glucuronosyltransferase-2B7 and Glutathione S-transferase Mu 4), and transport (Multidrug resistance protein 4) of toxic compounds were found to be expanded in *C. nubiana*, suggesting possible adaptative mechanisms to desert diets that are affluent in toxic compounds. This work represents the first effort to understand the genomic of adaptations in *C. nubiana*, a wild African goat. The *C. nubiana* genomic information generated in this study is an important resource for researchers seeking to find genes of interest for breeding purposes among small ruminants of economic importance.

DECLARATION

I, **(Vivien Jepchirchir Chebii),** do hereby declare to the Senate of Nelson Mandela African Institution of Science and Technology that this dissertation is my own original work and that it has not been submitted nor being concurrently submitted for degree award in any other institution.

Vivien Jepchirchir Chebii Date: 14th July 2021

The above declaration is confirmed

Prof. Morris Agaba Date: 19th July 2021

Supervisor, School of Life Science and Bioengineering, NM-AIST, Tanzania

Dr. Emmanuel Mpolya Date: 19th July 2021

Supervisor, School of Life Science and Bioengineering, NM-AIST, Tanzania

Dr. Josiah Musembi Mutuku Date: 19th July 2021

Supervisor, The Biosciences eastern and central Africa - International Livestock Research Institute (BecA-ILRI) Hub, Nairobi, Kenya

COPYRIGHT

CERTIFICATION

The undersigned certify that they have read and hereby recommend for acceptance by the Nelson Mandela African Institution of Science and Technology a dissertation entitled: **Genome scan for signatures of adaptive evolution in *Capra nubiana*** by **Vivien Jepchirchir Chebii** in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Life Sciences of the Nelson Mandela African Institution of Science and Technology.

Prof. Morris Agaba

Supervisor 1 Date

Dr. Emmanuel Mpolya

Supervisor 2 **Date**

Dr. Josiah Musembi Mutuku

Supervisor 3 **Date**

ACKNOWLEDGEMENTS

First, much thanks to Almighty God for his protection and provision throughout my study period. I would like to express my special appreciation and thanks to my supervisor, Prof. Morris Agaba for his guidance throughout my studies; I owe most of my success to his persistent support. Thanks for introducing me to the field of evolutionary genomics, for being patient, for answering my emails even in the middle of the night, and for always believing in me, even when I had doubts about myself. I recognize and appreciate the invaluable support from my co-supervisors My sincere gratitude goes to my co-supervisors, Dr. Emmanuel Mpolya and Dr. Josiah Mutuku the invaluable support they accorded me throughout my research period.

I wish to thank my sponsors. The research was funded by Swedish International Development Cooperation Agency (SIDA) through grants to Biosciences eastern and central Africa-International Livestock Research Institute (BecA-ILRI Hub) (Grant number: UF2011/55504/UD/UP). My graduate fellowship was funded by the Deutscher Akademischer Austausch Dienst (DAAD) and was supplemented by BecA-ILRI Hub through Africa Biosciences Challenge Fund (ABCF) program. Much thanks to National Zoological Gardens of South Africa for providing the *C. nubiana* tissue samples. Next, I would like to thank BecA - ILRI for providing the computational infrastructure and research facilities. Sincere thanks to Jean-Baka Domelevo Entfellner and Samuel O. Oyola for their bioinformatics and genomics expertise advise. I am thankful to Joyce Njuguna, John Juma and Alan Orth for the bioinformatics and the computing cluster support. Thanks to the capacity building team; Jeniffer Kinuthia, Leah Symerkher, Valerian Oloo and Dr. Wellington Ekaya for your all rounded support while at ILRI.

My acknowledgement goes to the NM-AIST community for the overall support during my study. Special thanks to Dr. Edson Inshengoma for his worthy suggestions on how to improve my work. Much appreciation to my family for their support throughout my study life, despite the fact that I was not able to explain to them what it was I was doing to any degree of satisfaction, or what it was that I might end up doing afterwards. They certainly deserved more phone calls and visits than what I gave. Finally, I would like to thank my husband, Robert Joe Wendot for being supportive throughout my PhD journey.

DEDICATION

This work is dedicated to my lovely husband Robert Joe Wendot his love and care during my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND SYMBOLS

BAM	Binary Alignment Map
BEB	Bayes Emperical Bayes
BLAST	Basic Local Alignment Search Tool
Bp	Base Pair
BUSCO	Benchmarking Universal Single-Copy Orthologs
BWA	Burrows-Wheeler Aligner
CNV	Copy Number Variant
DNA	Deoxyribonucleic Acid
Fostes	Fork Stalling and Template Switching
Gbps	Giga Base Pairs
GO	Gene ontology
Indel	Insertion/Deletion
LRT	Likelihood Ratio Test
NAHR	Non-Allelic Homologous Recombination
NHEJ	Non-Homologous End-Joining
NGS	Next Generation Sequencing
OLC	Overlap/Layout Consensus
PCR	Polymerase Chain Reaction
SAM	Sequence Alignment Map
SNP	Single Nucleotide Polmorphism
SNV	Single Nucleotide Variant
VCF	Variant Call Format

CHAPTER ONE

INTRODUCTION

1.1 Background of the Problem

1.1.1 *Capra nubiana*

The Nubian ibex (*Capra nubiana*) is a wild desert goat belonging to the *Capra* genus. It inhabits rocky mountainsides of the hot desert regions with altitudes ranging from 400 meters below sea level to 1 500 meters above sea level (Shackleton, 1997). *Capra nubiana* is found in Israel, Saudi Arabia, Jordan, Egypt, Oman, Yemen, Eritrea and Sudan (Fig. 1). *Capra nubiana* depends on herbaceous and woody plants for their food. There is no population data for *C. nubiana*, but its number has been declining; hence, it is listed as an endangered species (Ross *et al.*, 2020). *Capra nubiana* is a sexually dimorphic animal with males being larger than females (Gross *et al.*, 1995). Males have beards and well pronounced semi-circular narrow horns with several knobs on the outer curves. In contrast, the females are beardless and have small horns that are slightly curved backward with few knobs (Habibi, 1997). Both sexes have black and white markings on the legs and a light, smooth tanned coat with a white underbelly (Fig. 2). In winter, the coat colour appears darker, while in summer, it's lighter. *C. nubiana* has a gestation period of 150 days and a life span of 17 years (Castello, 2016), and they segregate in small groups as an anti-predator strategy (Habibi, 1997). It is well adapted to intense solar radiation, extremely high temperatures and scarce feed and water supply in its desert environment (Baharav & Meiboom, 1981). In addition, *C. nubiana* thrives well in rough undulating terrains with mainly bare ground except for few xerophytic plants (Tadesse & Kotler, 2012). *Capra nubiana* is vulnerable to extinction because of habitat destruction by human beings, hunting by predators, poaching for aesthetic values, and hunting for trophies (Tadesse & Kotler, 2012). There have been efforts to conserve *C. nubiana*; however, individual countries hinder conservation efforts because of conflicting interests (Shackleton, 1997).

Figure 1: A map showing the worldwide distribution of *C. nubiana* [1] and other three species; *Capra walie* [2], *Capra aegagrus hircus* [3] and *Capra hircus* [4]

Figure 2: A photograph showing two male (M) and female (F) *Capra nubiana*

Photo credit: https://animals.sandiegozoo.org/animals/nubian-ibex.

1.1.2 Vertebrates adaptations to desert environments

Desert vertebrates such as goats, sheep, and camels must contend with harsh environmental conditions in their habitats. Desert vertebrates have various behavioural, morphological, physiological and genetic adaptive mechanisms to deal with diverse stressors in their environments (Berihulay *et al.*, 2019). Behavioural adaptations are activities that animals carry out to increase their survival in harsh conditions, such as diet selection, the timing of activities, migration, social behaviours and timing of reproduction (Gebreyohanes & Assen, 2017). A case of behavioural adaptation is showed by the desert mule deer and other ungulates; they prefer to rest in shaded, lower-temperature microhabitat during the hottest time of the day (Tull *et al.*, 2001).

Desert-dwelling animals also use morphological traits such as skin colour, body size, and fat deposition that aids in heat load reduction and water loss. *Capra nubiana* for instance, has a light, smooth, shiny coat that reflects solar radiation (Castello, 2016), while the camel has fat deposits in the hump that acts as an energy reserve during starvation (Guo *et al.*, 2019). In

addition, mammals living in deserts have evolved physiological adaptations to minimize water loss through urine, faeces, skin, and lactation. Some physiological mechanisms employed by desert species include; selective brain cooling, evaporative cooling, concentrated urine output, and low moisture faeces (Berihulay *et al.*, 2019). Arid-adapted mammals such as camels output little and concentrated urine as a way of minimizing water loss; this is linked to their kidney ability to produce concentrated urine owing to their long loop of Henle (Gebreyohanes & Assen, 2017). In addition to behavioural, physiological and anatomical adaptations, desert mammals have evolved genetic traits to survive in harsh environments. Genetic bases of adaptation refer to an evolutionary process that moulds genes of a species to adapt to a given environment. For instance, the camel has evolved adaptive strategies to deal with prolonged exposure to ultraviolet light, evidenced by strong selection signals in visual protection genes (*OPN1SW* and *CNTFR*) (Wu *et al.*, 2014).

1.1.3 The genomics of adaptations

Recent advances in genomic technologies have provided opportunities for understanding the adaptive signatures seen in diverse species. Genomics studies of desert species such as camel and tortoise have provided substantial evidence of the adaptive signatures behind the endurance to harsh desert environments. For example, genome sequence analysis of the camel has shown that energy production and storage genes are rapidly evolving, suggesting an adaptative trait to food scarcity in deserts (Bactrian Camels Genome Sequencing and Analysis Consortium *et al.*, 2012; Wu *et al.*, 2014). Similarly, arachidonic pathway genes are under strong selection signals in camels and sheep inhabiting desert environments (Bactrian Camels Genome Sequencing and Analysis Consortium *et al.*, 2012; Yang *et al.*, 2016). The arachidonic pathway regulates water re-absorption and retention in the kidney by modulating reno-vascular tone changes (Miyata & Roman, 2005). Positive selection of ultraviolet radiation-related genes has been reported in desert animals such as tortoises (Tollis *et al.*, 2017). Genome sequence comparison of related species inhabiting diverse biomes provides clues into the genetic bases of adaptations. The availability of the reference genome of *C. hircus* (domestic goat) (Bickhart *et al.*, 2017) and other related species such as cow (Zimin *et al.*, 2009), sheep (Jiang *et al.*, 2014), and yak (Qiu *et al.*, 2012), provide a promising avenue for studying the genomics of adaptations in *Capra* species.

1.2 Statement of the problem

Goats are a source of milk, meat, and wool for many households across the globe. Their ability to adapt to diverse environments makes them an important source of livelihood for the people living in low-input production systems. There are about 1 billion goats globally, with the largest percentage (59.38%) being in Asia, followed by Africa (35%), with the remaining small percentage being distributed in America, Oceania, Europe and the Caribbean (Gurgul *et al.,* 2019). Given the goats' economic importance, genomics initiatives such as the International Goat Genome Consortium (IGGC) have been established to facilitate in-depth goat genome biology research. The release of the reference genome of the (*C. hircus*) domestic goat (Bickhart *et al.,* 2017) adds to the pool of goat genomic resources, which are key tools in understanding the goat genome. Despite the appreciable efforts made in goat genomics, much emphasis has been put on the domesticated goat species. Wild goats have rich genetic resources, yet their genome is unexplored. The availability of diverse goat genetic resources provides new avenues for understanding wild goat genome biology through comparative genomic analyses.

1.3 Ratinale of the study

Capra nubiana thrives well in an inhospitable environment marked byl little feed resources, solar radiation and temperatures extremes and limited water. On the other hand, the domestic goats inhabit various agro-ecosystems in Africa and Asia continents. While *C. nubiana* is endemic to hot deserts, a substantial percentage of the domesticated goats are found in arid and semi-arid regions where the climate is expected to become hotter and drier due to global warming effects (Henry *et al.,* 2018). Climate change due to global warming and other factors has a considerable impact on animals, including livestock species (Henry *et al.,* 2018). The influence of climate change is particularly pronounced in extreme environments such as arid and semi-arid regions. I hypothesized that *C. nubiana* had evolved adaptive strategies to survive harsh conditions, and; robust analysis of its genome should reveal footprints of its adaptations. Therefore, it is paramount to know the genetic basis of well-adapted species such as *C. nubiana* for developing appropriate goat breeding programs in anticipation of future climate change scenarios.

1.4 Research Objectives

1.4.1 General objective

The objective of this study was to identify adaptive signatures of evolution in the *C. nubiana*, a wild African goat. The adaptive signatures will serve as selection markers in goat breeding programs to improve their resilience to challenging environments.

1.4.2 Specific objectives

(i) To generate *Capra nubiana* genome sequence data.

(ii) To identify positive selection signatures in *Capra nubiana* genome.

(iii) To detect copy number variable regions in *Capra nubiana* genome.

1.5 Research question

(i) What is the genetic basis of *Capra nubiana* adaptation to hot desert environments?

1.6 Significance of the study

The data generated in this study provides new genomic data for an important *Capra* species (*C. nubiana*), a threatened wild goat; this will open avenues for the conservation of their biodiversity. The adaptive signatures detected could be used as selection markers for designing goat breeding programs in view of the rapid global effects of climate change.

1.7 Delineation of the study

Climate change due to global warming and other factors has a considerable impact on animals, including livestock species (Henry *et al.*, 2018). The influence of climate change is particularly pronounced in extreme environments such as arid and semi-arid regions. Therefore, the present study focused on the investigation of genetic basis of well-adapted species such as *Capra nubiana* for developing appropriate goat breeding programs in anticipation of future climate change scenarios.

CHAPTER TWO

LITERATURE REVIEW

2.1 Comparative genomics as a tool for understanding adaptive evolution

Comparative genomics is a research field in which genome sequences of different closely related species are compared. Genome sequence comparisons explain what distinguishes different species from each other at the molecular level and the possible genetic basis of their adaptations. This chapter reviews genomics of adaptations, various comparative genomics approaches for detecting adaptive evolution and finally, a brief overview of *Capr*a species and the present status of goat genomics.

2.2 Comparative genomics analysis: insights into vertebrates environmental adaptations

Genomes are shaped by random genetic drift, selection or both, which could, in turn, result in variation in phenotypic traits. Natural selection leaves signatures in a genome that can be used to identify the genes underlying a given phenotype (Bamshad & Wooding, 2003). Genome comparison of related species provides information on the possible selection signatures contributing to phenotypic traits seen in species. For instance, genome comparison of the domestic yak and related species revealed that hypoxia tolerance genes (Arginase 2 and Matrix Metallopeptidase 3) were evolving in the yak, possibly suggesting an adaptive traits to high altitudes (Qiu *et al.,* 2012). While comparison of camelid genomes showed that the dromedary camels have evolved genetic mechanisms such as ability to metabolize high salt in their diet and endurance to prolonged exposure to solar radiations in its desert environment (Wu *et al.,* 2014). For instance, genes involved in fat metabolisms (e.g., acetyl-CoA carboxylase 2 and Diacylglycerol Kinase Zeta), oxidative stress response (e.g., Nuclear Factor, Erythroid 2 Like 2 and Microsomal Glutathione S-Transferase 2) and salt metabolism (e.g., Nuclear Receptor Subfamily 3 Group C Member 2 and Insulin receptor substrate 1) are under accelerating evolution in dromedary camels (Wu *et al.,* 2014). Comparative genomic analysis of giraffe and related species showed that blood pressure regulation gene (Fibroblast Growth Factor Receptor Like 1) is positively selected in giraffe and is associated with hypertension endurance (Agaba *et al.,* 2016). Comparative genomics of desert tortoise has shown that it is enriched for genes involved in response to ultraviolet radiation (e.g. DNA excision repair protein ERCC-6) and regulation of urine volume (e.g. Hyaluronidase-2); that are the key genetic basis of

adaptations to arid environments (Tollis *et al.,* 2017). Similarly, genome sequences comparison of ruminants showed that Alcelaphni species have evolved adaptive strategies to its grasslands environment has evidenced by selection signals in cursorial genes (erythropoietin and angiotensin I converting enzyme), crucial for endurance (Chen *et al.,* 2019). At the same time, adaptive evolution of Period Circadian Regulator 2 gene in reindeer is associated a circadian arrhythmicity a survival mechanism to Arctic environments (Lin *et al.,* 2019).

2.3 Genome sequence generation

Comparative genomics studies rely solely on genome sequence data of the species of interest and reference genome. The generation of genome sequences for any given species is now possible due to sequencing technologies' advances. Next-generation sequencing (NGS) technologies invented in early 2000 have become methods of choice for genome sequencing experiments due to their ability to generate large volume of data at a reasonable cost than Sanger sequencing method (Goodwin *et al.,* 2016). Several next-generation sequencing technologies available in the market such as Illumina (Solexa) HiSeq and MiSeq sequencing, Roche 454 pyrosequencing, and Ion Torrent.

The general workflow for next-generation sequencing entails DNA extraction, library preparation and amplification, clonal formation, sequencing, and quality control analysis (Kchouk *et al.,* 2017). Sequencing experiments yield millions of short DNA fragments known as raw sequence reads. The raw sequence data are then subjected to a series of quality control analysis procedures to inspect contaminants or low-quality bases using Bioinformatics algorithms such as FASTQC (Andrews, 2010). The FASTQC program output summaries of the data quality; based on the quality control analysis, pre-processing procedures such as adapter and poor-quality sequence trimming may be carried out using programs such as Trimmomatic v.39 (Bolger *et al.,* 2014). Assembly of the 'clean' sequence reads follows the quality control check phase.

2.4 Genome assembly approaches

There are two main genome assembly approaches *de novo* and reference-based assembly. The reference-based genome assembly approach reconstructs the genome of interest using a genome of a closely related species as the reference. It involves aligning the sequence reads to a reference genome and reconstructing the genome by taking the consensus call for a given

base. Reference-based genome assembly depends on the availability of a high-quality reference genome sequence. It offers a convenient way of detecting genomic variations and species evolution without carrying out *de novo* assembly, which is a challenging and complicated in terms of cost, computational resources and time. *De novo* assembly, on the other hand, seeks to reconstruct a genome from sequence reads without using any reference genome (Miller *et al.,* 2010). In brief, *de novo* assembling is merging sequence reads to long continuous stretches of sequences known as contigs, which share identical nucleotide sequences as the sequenced DNA template (Paszkiewicz & Studholme, 2010). Contigs are unordered and have gaps. Two or more contigs ordered and joined together using read-pair information, with gaps filled with the consecutive letter 'N' denoting regions of uncertainty forms scaffolds (Yandell & Ence, 2012). *De novo* assembly is widely used to generate a draft genome of species when the reference genome is unavailable. There are two major *de novo* assembly approaches; overlap-layout-consensus (OLC) (Li *et al.,* 2012) and de-brujin-graph (DBG) (Miller *et al.,* 2010).

The OLC works by finding the overlaps (O) in all reads, then it lays out (L) the reads, and overlap information in a graph and finally infers the consensus (C) sequence from the multiple sequence alignments (Li *et al.,* 2012). On the other hand, a de Bruijn graph is a graph representing a homogeneous overlap between sequences (Jackman *et al.,* 2017). The DBG works by first fragmenting the sequence reads into shorter sequences of defined lengths called k-mers used to construct de Brujin graphs (Ekblom & Wolf, 2014). The DBG based approach is appropriate for high throughput data since it doesn't do all-against-all pair-wise read comparison like OLC; hence it is not computationally expensive (Miller *et al.,* 2010). With the rapid generation of high throughput data for large complex genomes, DBG based algorithms have become the most preferred assembly tools. Examples of DBG based programs include; By Jackman *et al.* (2017), and Soapdenovo2 (Luo *et al.,* 2012). Although modern assemblers can handle repetitive and heterozygous issues partially, short sequence reads often lead to fragmented assemblies (Bao *et al.,* 2014). *De novo* assembled genomes can be improved using the reference-assisted assembly approach and other techniques such as generating more sequence data of different lengths. Reference assisted *de novo* genome assembly involves using a reference genome of a closely related species to guide the extending and joining of *de novo*-assembled contigs (Lischer & Shimizu, 2017). Some reference-assisted *de novo* assembly algorithms such as AlignGraph significantly improve de novo assembled genomes (Bao *et al.,* 2014).

2.5 Evaluation of genome assembly quality

Assessment of assembled genome quality in terms of contiguity and accuracy is essential. Several metrics such as N50 contigs size, k-mer statistics, percentage of mapped reads and genome completeness in terms of gene contents are among the key indicators used to assess the genome assembly quality. The N50 statistics is the median contig size of a given genome used to measure the assembly contiguity (Yandell & Ence, 2012). The N50 is obtained by ranking all the contigs from the longest to the smallest, then adding them starting from the longest, until the sum just exceeds 50% of the total length of all the contigs present (Paszkiewicz & Studholme, 2010). The number and lengths of contigs and scaffolds give an overview of the assembly contiguity. Longer and fewer contigs and scaffolds reflect a good genome assembly.

The total assembled genome size relative to the expected genome size is another standard measure for assembly evaluation. Assembled genome size is calculated by summing the lengths of all contigs together (Gurevich *et al.*, 2013). Expected genome size is commonly inferred using k-mer frequency-based approaches (Vurture *et al.,* 2017). Theoretically, the total assembled size should reflect the estimated size of the target genome, but due to the genome's complex nature arising from repeats and heterozygosity, some genomes have slight variation but not significant. For a good assembly, the estimated genome size should be close to the expected genome size. Programs such as QUAST tool could be used to generate assembly statistics, scaffolds lengths and size, genome size and N50 statistics (Gurevich *et al.,* 2013). Expected gene content based on conserved genes of related species is another metric used to check the completeness of a genome. Gene content is assessed by checking the presence of single-copy orthologous genes across diverse species (Waterhouse *et al.,* 2011), using Benchmarking Universal Single-Copy Orthologue (BUSCO) assessment tool (Simao *et al.*, 2015). The BUSCO determines the completeness of a genome by evaluating a set of conserved single-copy orthologs that are expected to be present in any mammalian kingdom against the assembled genome. A qualitative measure is generated when the BUSCO tool identifies single-copy orthologs genes as complete, duplicated, fragmented or missing (Simao *et al.,* 2015). A genome with high completeness will have a higher percentage of complete single-copy orthologs and few missing genes.

The final metric for assessing the genome assembly completeness is by utilizing K-mer plots. K-mer plots generated by K-mer Analysis Toolkit (KAT) (Mapleson *et al.,* 2017)

provide reliable metrics to evaluate assembly accuracy, sequence biases and contaminants. The KAT plots give information on how much kmer content from the sequence reads are in the assembly. Missing sequence reads in the assembly are illustrated as black sections below the main red peak, and sequencing errors are represented as black bars along the graph's y-axis (Fig. 3). The histogram bars in red color represent sequence reads in the assembly once, indicating a good assembly. Peaks, which are not black or red are indicative of duplications in the assembly. A poor assembly will have many k-mers missing in the assembly **(**Fig 3A; black histogram below the main red peak. In ideal situations, a good assembly should have most if not all the k-mers accounted for in the assembly once (red histogram bars) (Fig. 3B).

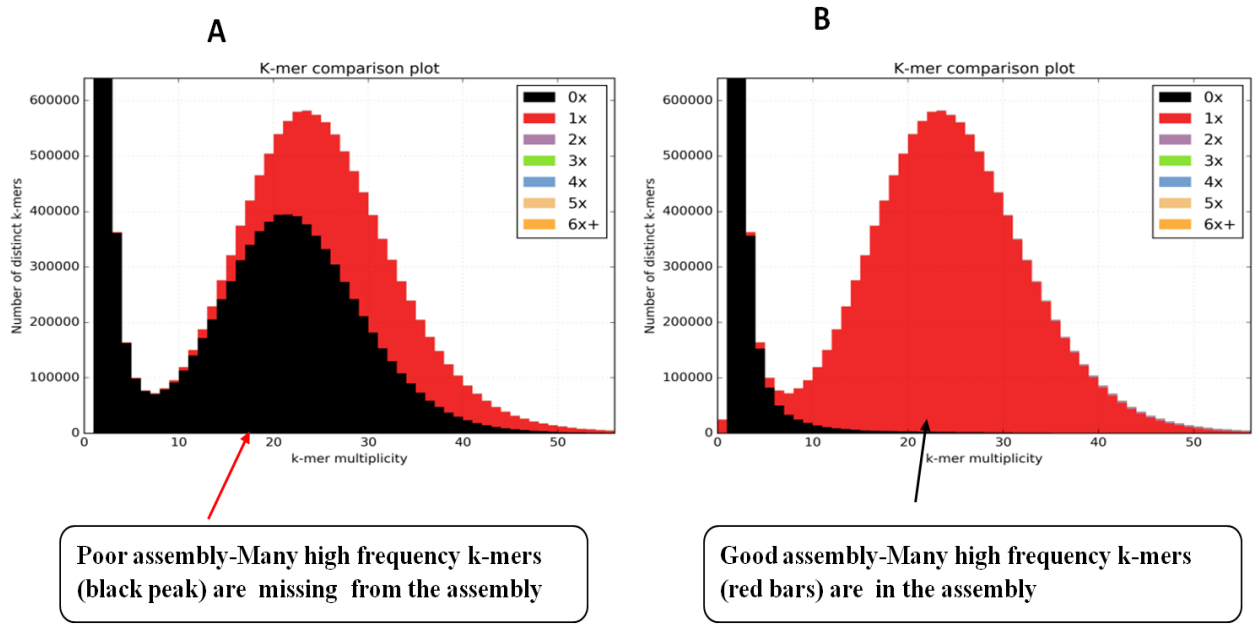


Figure 3: Kmer-spectra plot generated using Kmer analysis tool (KAT) showing motif and copy number representation. Coloured plots show how many times fixed-length words (k-mers) from the reads appear in the assembly, frequency of occurrence (multiplicity; x-axis) and the number of distinct k-mers (y-axis). Black represents sequence reads missing in the assembly; red, sequence reads that appear once in the assembly; green, twice. A, the black distribution between kmer multiplicity 10 and 40 represents sequence reads that are not in the assembly. B: All sequence content is present in the assembly once (shown by the red distribution). The KAT plots used here are part of KAT tool documentations (Mapleson *et al.*, 2017)

2.6 Genome annotation

Genome annotation is a way of linking biological information to sequence data (Ekblom & Wolf, 2014). The first step of genome annotation involves identification and masking of repeats. The first step of genome annotation involves the identification and masking of repeats. Repetitive elements are nucleotide sequences that occur in multiple copies throughout the genome and are composed of low-complexity sequences and mobile elements (Yandell & Ence, 2012). Repeats are either tandem or interspersed repeats (Huang *et al.,* 2016). Repeat masking programs such as RepeatMasker (Hoff *et al.,* 2019). Repeat masking programs such as RepeatMasker (Smit *et al.,* 1996) are commonly used for repeat masking purposes.

There are two main genome annotation types: structural and functional annotation. Structural annotation entails the prediction of protein-coding genes, RNAs, repetitive elements and regulatory motifs. Conversely, functional annotation is the discovery of gene biological, molecular and cellular functions. While genome annotation involves the characterization of several biological elements such as non-coding RNAs, much focus has been put on protein-coding genes owing to their functional roles. Gene prediction methods are classified into de novo and similarity-based approaches. *De novo* approaches use statistical models such as Hidden Markov Model Programming and Neural networks to predict gene structures such splice sites, start codons and stop codons (Wang *et al.,* 2004). Some of the commonly used de novo gene predictors include; Augustus (Stanke *et al.,* 2006) and Genescan (Burge & Karlin, 1997). On the other hand, similarity-based approaches infer gene from similar sequences of related species by comparing the target and the reference genome (Yandell & Ence, 2012). Similarity-based approaches assume that gene exons are conserved; hence, the similarity between certain genomic regions in the target genome and the reference can be used to infer the gene structure of that region (Wang *et al.,* 2004). The BLAST is among the prominent similarity-based gene prediction tool (Altschul *et al.,* 1990). A combination of de novo and similarity-based methods is recommendable for high confidence gene predictions.

2.7 Genomic variations and its detections

Genomes evolve by accumulating variations at a single base level or large-scale rearrangements like copy number variations. Genomics variations are differences in DNA sequence between two or more genomes, and they play essential roles in adaptation and diversification. Hence, the detection of genomic differences provides insight into the

phenotypes associated with adaptation of a given species to diverse habitats and processes (Andersson & Georges, 2004).

2.7.1 Single nucleotide variants (SNV), Insertions/deletions (InDels), and detection

Single nucleotide variants (SNVs) are single nucleotide substitutions; SNVs with a frequency of >1% in a population are called single nucleotide polymorphisms (SNPs) (Vignal *et al.,* 2002). The SNVs are of two types; transitions (swap of purine or pyrimidine) or transversions (swap of purine with pyrimidine or vice versa). Transitions are found twice compared to transversions in a genome (Rosenberg *et al.,* 2003). Single nucleotide mutations arise due to error in DNA replication, oxidant damage by reactive oxygen species, ionizing radiation or due to chemical mutagens (Spencer *et al.,* 2015). The SNVs contribute to the phenotypic variation seen in species; for instance, a substitution of arginine to cysteine in coatomer protein complex, subunit alpha gene is linked to a striking Dominant Red phenotype in Holstein cattle (Dorshorst *et al.,* 2015). While, amino acid substitution of Alanine to glycine at position 69 of heat shock protein family B (small) member 7 gene is suggested to contribute to heat tolerance in cattle (Zeng *et al.,* 2019). The SNVs may occur in the coding or non-coding regions of the genome. Non-coding SNVs overlaps with 5' and 3' UTRs, introns and intergenic regions, while coding SNV overlaps with coding sequence region. Coding SNVs are either synonymous or non-synonymous SNV. Non-synonymous SNV changes the amino acid sequence while synonymous SNV does not.

Non-synonymous SNV may change the protein function by altering its structure, while non-coding SNV may alter the gene expression by affecting transcription binding and splicing regulation (Zhang *et al.,* 2012). Non-synonymous SNVs alter the amino acid sequence; these changes significantly impact the protein structure and function (Dakal *et al.,* 2017). The severity of the non-synonymous SNV varies depending on the region of the protein affected and the nature of the change. If an amino acid is changed into amino acid with similar chemical characteristics (e.g. hydrophobic-hydrophobic or hydrophilic-hydrophilic), its effects on protein function might have a less severe impact compared to a case where the amino acid is changed to an amino acid with a different chemical composition (e.g. hydrophobic-hydrophilic) (Spencer *et al.,* 2015). Non-synonymous SNVs are termed as nonsense mutation when it introduces a premature stop codon by deleting or adding a stop codon to a sequence (Haraksingh & Snyder, 2013). Generally, non-synonymous SNVs (nsSNVs) have been shown to affect the protein function by destabilizing its structure or by affecting physicochemical

properties; hence they are very important genetic variants (Katsonis *et al.,* 2014). Owing to SNVs significant role in genetic diversity in mammals, several studies have attempted to carry out in-depth studies on these genetic markers. For example, in the cattle genome, several studies have sought to identify SNVs (Bovine HapMap Consortium, 2009; Medeiros de Oliveira Silva *et al.,* 2017), similar to sheep (Naval-Sanchez *et al.,* 2018; Yang *et al.,* 2016). In goats, appreciable efforts have been made to identify SNVs and characterize them. For example, in their studies (Benjelloun *et al.,* 2015) identified approximately 24 million SNVs in Moroccoian domestic goats, while Guo *et al.* (2019) identified 15 million SNVs in six domestic goat breeds. Several other efforts to generate SNV data in domestic goats have been reported (Kijas *et al.,* 2013; Nicoloso *et al.,* 2015; Tosser-Klopp *et al.,* 2014); however, none has been reported in wild goats.

Insertions and deletions (InDels) mutations lead to the gain or loss of one or more nucleotides in a genomic region and include events less than 1kb in length (Sehn, 2015). Many processes such as replication slippage, imperfect double-stranded DNA break repairs, and recombination generate InDels (Boschiero *et al.,* 2015). The InDels in coding sequence regions may lead to frameshift or non-frameshift mutations. Frameshift changes the reading frame from the site of insertion/deletion, leading to a change in the protein sequence (Lin *et al.,* 2017). Non-frameshift mutations, on the other hand, introduce the insertion or deletion of one or more amino acids while keeping the rest of the protein sequence unaltered (Sehn, 2015). The InDels, just like SNVs, contribute to phenotypic variations seen in species; for example, amino acid deletion of Fanconi anaemia pathway genes in African wild dog is suggested to contribute to specialised cursoriality ability through reduction of digits on the forepaws that allows for increased speed and capture of prey (Chavez *et al.,* 2019). Similarly, amino acid deletion in CREB Binding Protein might have contributed to hypercanivory ability in African wild dogs (Chavez *et al.,* 2019). The InDels identification and characterization lagged behind for several years since there were no well-established methods for detecting them (Montgomery *et al.,* 2013; Stafuzza *et al.,* 2017). However, the recent reduction of sequencing cost has sparked many studies aimed at identifying and characterising InDels in genomes (Lee *et al.,* 2016; Stafuzza *et al.,* 2017). The SNVs and InDels are detected from sequencing data using mapping-based approaches (Li, 2012). Mapping based approaches entail aligning sequence reads to a reference genome, and the differences are identified (variant calling) (Olson *et al.,* 2015). Detections of SNVs and InDels from sequence data follow three analysis steps: sequence reads quality analysis, alignment of the sequence reads to suitable reference genome and

variant calling. Following quality control analysis, sequence reads are aligned to the reference genome using aligners such as Burrow Wheele Aligner (Li & Durbin, 2010) or Bowtie (Langmead & Salzberg, 2012). The variants are then called using variant callers; samtools mpilep (Li *et al.,* 2009) and Genome Analysis Tool Kit HaplotypeCaller (GATK-HC) (McKenna *et al.,* 2010). The list of variants detected are then annotated; this entails attaching some information such as gene symbol, amino acid, exonic function and genomic feature using bioinformatics tools such as Variant effect predictor (VEP) (McLaren *et al.,* 2016) and ANNOVAR (Wang *et al.,* 2010).

2.7.2 Copy number variations (CNVs) and detection methods

Previously SNVs and InDels were considered the only variants contributing to the genomic variation observed in genomes. However, other structural variations such as copy number variations (CNVs) that affect thousands of base pairs of the genome also contribute to a significant percentage of variations seen in species. Copy number variants are structural variations in which a genomic fragment greater than 1000 base pairs are lost or gained, leading to copy number differences in specific genomic regions between genomes (Alkan *et al.*, 2011). The CNVs overlap with large stretches of genomic regions as a result they affect several functional genes and fitness in organisms (Schrider & Hahn, 2010). Two primary cellular mechanisms, i.e. non-allelic homologous recombination (NAHR) and non-homologous end joining (NHEJ), both of which are intended for maintaining the DNA integrity by repairing DNA double-stranded breaks (DSBs) are contributing factors behind CNVs formations (Arlt *et al.,* 2012; Hastings *et al.,* 2009). The NAHR arise during meiosis or mitosis when recombination occurs between non-allelic homologous DNA sequences from different chromosomal positions (Arlt *et al.,* 2012). Unequal crossing-over events of these genomics regions result in gain or loss of copies (Stankiewicz & Lupski, 2010). Segmental duplications and low copy repeat regions are often substrates of CNVs formation via NAHR (Stankiewicz & Lupski, 2010). The NHEJ is another mechanism used to repair double-stranded DNA (DSBs) breaks, which results from ionizing radiation or other DNA damaging agents. Double-stranded DNA breaks prompt NHEJ DNA repair mechanisms to modify broken ends to fit each other and finally ligate them together (Chiruvella *et al.,* 2013). If the NHEJ DNA repair process is erroneous, it leads to a gain or loss of genomic regions. The distinguishing feature between the two DNA repair mechanisms is that NAHR significant sequence identity is not required. Other mechanisms that have been implicated in CNVs formation include fork stalling and template switching (FoSTeS) and mobile element insertion (MEI) (Hastings *et al.*, 2009).

The FOSTES occurs when the DNA replication complex stalls due to DNA lesions of the nucleoside bases as a result the lagging strand of DNA associates with a different region of the genome with high sequence similarity leading to CNV formation (Hastings *et al.,* 2009; Zhang *et al.,* 2009). Copy number variants contribute to phenotypic variations by changing transcript structure, gene dosage or by regulating gene expression (Bickhart & Liu, 2014). A CNV overlapping with coding sequence region may alter the gene expression level by changing the number of functional copies or altering the structure of regulatory regions (De Smith *et al.,* 2008). The CNVs have important fitness effects in individuals which are either beneficial or detrimental (Buchanan & Scherer, 2008). Many copy numbers (>4) of the beta-defensin gene, for example, is linked to Crohns' disease (Bentley *et al.,* 2010). At the same time, many copies of the *Rhg1* gene in soybean contribute to resistance to nematodes (Cook *et al.,* 2012). The CNV affecting genes also contributes to phenotypic variations observed in different livestock species. For example, gain of a copy of KIT Proto-Oncogene, Receptor Tyrosine Kinase gene contributes to dominant white colour observed in swine (Giuffra *et al.,* 1999), while copy gain of the Agouti Signaling Protein gene contributes to white coat colour in sheep (Fontanesi *et al.,* 2011). Owing to the CNVs phenotypic significance, several studies have been undertaken in livestock species such as cattle (Bickhart *et al.,* 2012; Gao *et al.,* 2017), sheep (Jenkins *et al.,* 2016; Yang *et al.,* 2018), pig (Paudel *et al.,* 2013), and yak (Qiu *et al.,* 2012). In goats, some CNV studies have been reported (Di Gerlando *et al.,* 2020; Fontanesi *et al.,* 2010; Jenkins *et al.,* 2016) as well.

Traditionally, SNP array and array comparative genomic hybridization (aCGH) methods were used to detect CNVs. Array CGH approach involves co-hybridization of differently fluorescent-labelled genomic DNA from the test and reference sample to an array of probes (oligonucleotides or bacterial artificial chromosome clones); it then compares the ratio of the fluorescence signals to infer copy number variations (Carter, 2007). A high-intensity signal signifies gain of copy events, while low intensity indicates loss of copy events. The SNP array also involves hybridization but relies on the use of single nucleotide polymorphisms (SNP) array to assess genomic regions of high or low probe intensity, indicative of gain or loss of copy events (Alkan *et al.,* 2011). Although SNP arrays and array CGH methods were the most used approaches to detect CNVs in the past, they suffered limitations such as hybridization noise, low resolution, limited genome coverage, and sensitivity (Zhao *et al.,* 2013). Advances in genomic technologies spurred the decelopment of next-generation sequence (NGS) based methods such as Read Pair (paired-end mapping), Split Read, Read Depth (Depth of coverage)

and Assembly based approach (Zhao *et al.,* 2013). Read Pair method is based on sequence information from both ends of a DNA segment (paired-end reads) and is only used to identify CNVs from paired-end reads (Tattini *et al.,* 2015). In a paired-end sequencing experiment, the DNA fragments in a given library are prepared such that the insert size has a specific distribution (size). The paired-end mapping approach detects CNVs by determining discordantly mapped paired-end reads whose distances are inconsistent from the predetermined insert size (Korbel *et al.,* 2007). If read pairs map further apart than the predetermined insert size, this suggests a loss of copy event. In contrast, a gain of copy event is detected when read pairs appear in reversed order with differences in their span though the orientation is maintained. The main drawback of the read pair approach is that it cannot detect exact breakpoints and CNVs in repetitive regions (Pirooznia *et al.,* 2015).

The split-read approach aligns paired-end reads to reference genome, CNV event is inferred if one read pair fails to map to the reference genome (Zhang *et al.,* 2011). The unmapped reads are split into multiple fragments, then the first and the last fragment of each split read are re-aligned to the reference genome (Zhao *et al.,* 2013). The re-alignment step provides precise start and end positions of CNVs events. Split read approaches perform well in detecting CNVs; however, they perform poorly in regions with low complexity (Pirooznia *et al.,* 2015).

On the other hand, assembly-based approaches assemble genome sequences into contigs and scaffolds using *de novo* assembly approach. The resulting assembly is then aligned to a suitable reference genome; a CNV event is inferred if the assembly is inaccurate (Pirooznia *et al.,* 2015). Assembly based methods can detect all forms of variations such as; deletions, duplication, transversions, translocations, and inversions. However, they perform poorly in repeat-rich regions, and it is computationally expensive (Alkan *et al.,* 2011).

Finally, depth of coverage (read-depth based) approaches assume Poisson distribution in mapping read depth, any divergence from this distribution is indicative of CNVs event in the sequenced sample (Abyzov *et al.,* 2011; Xie & Tammi, 2009). This approach detects CNVs based on sequence coverage variations; when a sequenced individual exhibits more copies than the reference genome, the event is classified as a gain of a copy. In contrast, few copies in the sequenced genome, compared to the reference genome, are termed loss of copy event (Xie & Tammi, 2009). Read depth approaches calculates exact copy numbers and CNVs in genomic loci (Teo *et al.,* 2012), unlike the other methods. However, one drawback with read-depth techniques is their inability to distinguish various duplication events (Alkan *et al.,* 2011).

Despite this drawback, depth of coverage method remain the most robust approach for identifying CNVs since it accurately predicts exact copy numbers and CNV in complex genomic regions such as segmentally duplicated regions, unlike PEM/SR which only detects the position (Zhao *et al.*, 2013).

Read depth approaches rely on the variation of normalized read depth to estimate copy number variations. More copies of a locus in the sequenced individual as compared to the reference genome are indicative of gain of copy events, while few copies of a locus in the sequenced genome relative to the reference are indicative of loss of copy events (Abyzov *et al.*, 2011; Xie & Tammi, 2009). Due to its robustness, the read-depth approach has been used in several between species CNV studies, for instance, in the detection of CNV between wild and domestic yak (Zhang *et al.*, 2016). Similarly, read depth approaches have been used in CNV studies between river buffalo and cattle (Li *et al.*, 2019) and studies between the gray wolf and dhole genomes (Wang *et al.*, 2019).

The present study used read depth (depth of coverage) approach, implemented using CNVnator to detect CNV between the domestic goat and *C. nubiana*. The CNV calling using depth of coverage method involves a number of steps: alignment of sequenced species data to the reference genome, reference binning, read depth count, normalization and variant calling. The first step into CNV calling using depth of coverage method involves aligning the sequenced genome to the reference genome. Then the reference genome is divided into different non-overlapping bins (windows) of equal size. The window size is determined by the sequence data coverage, read length, and data quality (Abyzov *et al.*, 2011). Basing on their analysis (Abyzov *et al.*, 2011) suggested an optimal bin size of 30 bp for 100x coverage, 100 bp for 20-30x coverage and 500 bp for 4-6x coverage; determined by calculating the ratio of read-depth signal to its standard deviation (4-5). The number of reads mapped in each bin is then counted to obtain initial read depth signals. Depth of coverage approaches assumes that the sequencing process is uniform, hence mapped reads follow the Poisson distribution and is proportional to the number of copies. Genome sequencing technologies are prone to biases such as guanine-cytosine (GC) content and mappability biases. The GC biases in whole genome sequences such as sequences generated using Illumina technology result from the polymerase chain reactions (PCR) amplifications procedures used during library preparation and cluster amplification on the flow cells (Oyola *et al.*, 2012).

The GC content bias occurs when read coverage varies depending on the GC content (low/high) of the genome region, while mapping biases occur because the genome contains many repetitive regions leading to ambiguous mapping in those regions (Abyzov *et al.*, 2011). Several studies using Illumina sequencing technology have shown a strong correlation between read depth and GC content; hence GC content biases significantly affect CNV detections (Abyzov *et al.*, 2011; Magi *et al.*, 2012). Read-depth based algorithms mitigate systematic biases that might influence CNV detection by employing normalization (bias correction) procedures. The GC bias is corrected by binning the genomic regions by GC content and then adjusting the average read depth of each bin to the average read depth of the genome.

In any whole genome sequence mapping experiment, each mapped read is assigned mapping quality value which is a measure of the confidence that a sequence read actually comes from the position to which it is mapped (Li *et al.*, 2008). In CNVnator, when sequence read pairs map to two or more locations mapping quality assigned zero (*q0* filter), and to handle the mapping bias, one is randomly chosen (Abyzov *et al.*, 2011). Depth of coverage signals are processed after correction using the fragment segmentation method; in CNVnator, this is achieved using the mean shift algorithm (Abyzov *et al.*, 2011). Statistical hypothesis testing is then carried out for the read depth signals in each bin.

2.8 Adaptive signatures of evolution

Comparative genomics provides a reliable way of detecting selection signatures from sequence data; it uses statistical approaches such as maximum likelihood to identify the sequence-altering mutation (selection signatures). This is achieved by estimating the ratio of synonymous and non-synonymous substitution rates (dN/dS); the rate of fixation of these two types of mutations provides a powerful tool for understanding the sequence evolution (Nielsen & Yang, 1998). A significantly high non-synonymous versus synonymous ratio may indicate adaptive evolution at the molecular level and is an invaluable resource for understanding genetic mechanisms behind adaptation to diverse environments by living organisms (Yang *et al.*, 2000).

2.8.1 Nonsynonymous/synonymous substitution (dN/dS) analysis for detecting adaptive signatures of evolution

Non-synonymous single nucleotide substitution is single base changes, which alters the amino acid sequences of a protein, while synonymous substitution is single base changes that do not alter amino acid sequences (Yates & Sternberg, 2013). Non-synonymous/synonymous ratio (ω(omega) = *d*N/*d*S) identifies sequence altering mutation by estimating the *d*N and *d*S ratio (Utsunomiya *et al.,* 2013). Where omega (ω) of less than one reflects negative, omega (ω) equal to one indicates neutral evolution, while omega (ω) greater than 1 is evidence of adaptive evolution (Yang *et al.*, 2000). The Maximum Likelihood method based on comparative genomics (Nielsen & Yang, 1998; Yang *et al.,* 2000) is a valuable tool for estimating the dN/dS ratio as well as identifying positively selected genes. Probability theory of the Markov process that describes substitution between 61 sense codons and forms basis of maximum likelihood estimation of *d*N/*d*S ratio. Markov process is used to describe substitutions between the sense codons, where substitution is either transversion or transition.

Markov codon models is characterized by rate generator matrix $Q = \{qij\}$, where $\{qij\}$ is the substitution from sense codon i to sense codon j (Bielawski & Yang, 2003). It is described as follows:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position,} \\ \mu\pi_j, & \text{for synonymous transversion,} \\ \mu\kappa\pi_j, & \text{for synonymous transition,} \\ \mu\omega\pi_j, & \text{for nonsynonymous transversion,} \\ \mu\omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases}$$

Where πj is the codon frequency, κ is the transition/transversion rate, and w is the $dN=dS$. k parameter accounts for transition/transversion bias, codon frequency accounts for codon bias while ω accounts for synonymous and non-synonymous bias. This forms the basis for more complicated models of evolution and it is specified by model=0, Nssites=0 explained in subsequent sections.

In summary, Markov codon models allow the computation of the probability that a given protein sequence evolves into some other sequence over a certain amount of time. Working

hand in hand with Markov codon models are the likelihood ratio tests (LRT) which offer a means of testing assumptions (model parameters) through comparison of the null models (H0) and the alternative models (H1). Bayes empirical Bayes (BEB) method is used to test sites that are positively selected (Nielsen, 2005).

Markov codon models are classified into three models; branch models, site models and branch-site models (Yang *et al.,* 2000; Yang & Nielsen, 2002). The branch model was the first model to be developed, and it allows the (ω) to vary among branches in a phylogeny; hence it used to detect selection affecting lineages (Nielsen & Yang, 1998). One drawback with branch models is that ω values are averaged overall positions in the alignment; that is an unrealistic assumption since not all sites in a protein alignment are similar. As a result, we should not expect positive selection to act on all the protein sequence; this led to the development of site models (Yang *et al.,* 2000). The site model allows (ω) to vary among sites (codons) and it assumes three different classes of sites: $\omega < 1$, $\omega = 1$, and $\omega > 1$. Like the branch models, site models too has its limitation in that it is conservative for many genes since the test is only significant if the average $\omega > 1$ holds for all the sites (Reis & Yang, 2011). In most cases, positive selection affects only specific sites in specific branches; to overcome the weakness of previous models, a more robust model combining branch and site models was developed (Yang & Nielsen, 2002).

Branch-site models estimate different dN/dS among sites and branches (Yang & Nielsen, 2002). In branch-site models, a branch that is hypothesized to evolving is labelled as the foreground, while the unlabelled branches act as the background. Background branches share the same distribution of omega (ω) value among sites, whereas different values can apply to the foreground branch (Yang & Nielsen, 2002). To test for significance branch site model (alternate model) is compared with neutral site model M1a using LRT. The limitation of this model is that any instances of relaxation of purifying selection on the pre-specified branch could be interpreted as a positive selection (Zhang, 2004). To handle the drawback of the model, an improved Branch-site model A was developed by (Zhang *et al.,* 2005). Branch-site model A allows three sites in the foreground branch ($\omega = 1$, $\omega > 1$ and $\omega < 1$) and two classes of sites ($\omega = 0$, $\omega = 1$) in the background branches. It assumes four site classes; Class 0 - Codons are under purifying selection in all branches with $0 < \omega < 1$. Class1- codons evolving under neutral selection in both foreground and background branches ($\omega = 1$). Class 2a -codons may be evolving under positive selection ($\omega > 1$) on the foreground branch but under purifying selection ($\omega < 1$) on background branches. Class 2b- codons may be under positive selection ($\omega > 1$) on

the foreground branch but under neutral evolution (ω =1) on background branches. In branch-site model, model A which has four parameters: p_0, p_1, ω_0 and ω_2 under ω (omega) distribution serves as the altenate model, while null model is modified version of model A where $\omega 2 = 1$ fixed (Zhang *et al.,* 2005). Improved Branch-site model A has its weakness in that saturation over long evolutionary times might occur (Gharib & Robinson-Rechavi, 2013). Once a given model has been used to identify positive selection signatures, the resulting outputs are further subjected to a series of statistical analyses. For instance, the likelihood ratio test compares nested probabilistic models; the null model does not allow $\omega > 1$ and the alternate model which does. Significant LRT and $\omega > 1$ in at least one of the models indicate positive selection in a given gene. Examples of these nested models include; M1 versus M2 (alternate), M7 versus M8 (alternate) and M8a versus M8 (alternate) site models; for branch models, it includes one ratio model versus two ratio models (alternate) and one ratio model versus free ratio models. Finally, there is a branch-site model, which includes Branch model A versus the null model A; for this model, the parameters are similar except that in the null model $\omega 2 = 1$.

The LRT for various models is computed as follows; LRT= 2 (lnL1-lnL0), where lnL1 is the alternate model, while lnL0 is the null model. The p-value is then computed following chi-square distribution, donated as p-value = chi-square (2*delta lnL, df), where 2* delta lnL= 2 (lnL1-lnL0) =LRT and degree of freedom is the difference in the number of parameters. When the LRT of a gene is significant (i.e., <0.05 or <0.01). Positivel selected sites in candidate genes are detected using Bayes empirical Bayes approach (BEB) (Yang *et al.,* 2005). Basing on the series of statistical analysis, one can conclude that a gene is evolving if (ω > 1 and a significant LRT result) and that specific sites of a gene are the target of selections if posterior probability (pp) is >0.50. Sites with high posterior probabilities >0.9 are normally interpreted as a sign of strong positive selection at that a given site.

(i) Strengths and limitations of maximum likelihood approaches

Maximum likelihood approaches are considered the most reliable tools for detecting adaptive signatures since they consider molecular biases such as transition-transversion, codon usage bias, and they don't rely on ancestral reconstruction (Anisimova *et al.,* 2002). Despite its strengths it also has its limitations; maximum likelihood methods rely on the accuracy of sequence alignment and phylogeny tree to detect adaptive signatures; unreliable trees and alignments could lead to false-positive results (Schneider *et al.,* 2009). The power of the maximum likelihood test is affected by sequence length; longer sequences have a high

probability of detecting adaptive evolution, unlike short sequences (Anisimova *et al.,* 2001). Similarly, the prediction of positive genes is unreliable when the number of taxa used is small and when highly conserved or diverged sequences are used (Anisimova *et al.,* 2002). Maximum likelihood approaches assume there is no recombination event, any increase in recombination events leads to the unreliability of the LRT test since it interferes with phylogeny tree quality by introducing long tree lengths (Anisimova *et al.*, 2003).

2.9 Functional impact of amino acid substitutions

Non-synonymous single nucleotide polymorphism alters the amino acid sequence leading to change in protein structure and function and structure. Computational approaches of predicting the impact of non-synonymous substitution have been developed, which are grouped into; sequence conservation-based approaches, structure analysis-based approaches, combined (both sequence and structure information) and meta-prediction (combines multiple predictors) (Tang & Thomas, 2016).

Sequence conserved-based approach such as Sorting Intolerant From Tolerant (SIFT) (Kumar *et al.,* 2009) infers the functional implications of substitutions based on sequence homology and physical properties of amino acids. The SIFT predicts the degree of amino acid conservation residues in sequence alignments of closely related sequences and is assigned a tolerance index score ranging from 0 to 1. The SNPs with a tolerance index score of <0.05 is considered deleterious (intolerable) and tolerated if the score is >0.05 (Kumar *et al.,* 2009).

PolypPhen-2 (Polymorphism Phenotyping v2), on the other hand, employs a combination of sequence-based and structural predictive features to predict the possible functional consequences of an amino acid change on the structure and function of a protein (Adzhubei *et al.,* 2013). It uses query protein sequence in FASTA format as input and estimates the influence of a particular amino acid variant at a given position in the query sequence. Position-specific independent count (PSIC) score for every variant and the score difference between variants is then calculated. A score <0.2 predict a benign variant, a score of 0.2 to 0.85 predicts a possibly damaging variant, while 0.85 to 1 predict a probably damaging variant (Adzhubei *et al.,* 2013). A benign variant is considered neutral hence does not have any functional impact; damaging variants are considered deleterious since they change the protein function. Probably damaging variant has a more confident prediction that it causes a change in amino acid functions. In contrast, a possibly damaging variant has a less confident prediction that it causes a change in amino acid functions and structure. Since these two approaches rely on multiple sequence

alignment, alignment quality might affect the prediction accuracy; hence there is a need to generate high-quality alignments (Tang & Thomas, 2016). The SIFT has a false positive rate of 20%, while Polyphen-2 false positivity rate is 9% (Ng & Henikoff, 2006).

2.10 Goat and its genomics status

The *Capra* species which includes; *C. aegagrus, C. falconeri, C. sibirica, C. cylindricornis, C. caucasica, C. walie, C. nubiana* and *C. hircus* are classified based on the horn morphology and the shape of the cross-section of the horn sheaths of adult males (Fig. 4) (Manceau *et al.,* 1999; Pidancier *et al.,* 2006). For example, the *C. pyrenaica* has horns curved like a lyre, while the *C. falconeri* horns are twisted. *C. nubiana, C. walie, Capra ibex* and *C. sibirica* have sickle-shaped horns that have a flat anterior surface broken by transverse ridges (Fig. 4). The wild and the domestic goat species have distinct physical and behavioural characteristics, which distinguishes them from each other. For instance, domestic goats are easily distinguished from their wild counterparts due to their docility, small body size, horns, ears and diverse coat colour (Du *et al.,* 2014). On the other hand, the wild goats are aggressive; they have a large body size, large horns of different shapes, depending on the subspecies, and uniform coat colour, which is tanned for the majority of them (Parrini *et al.,* 2009).

Capra species are hardy species; they survive in wide range of biomes such as extreme cold deserts (*C. sibirica*), high altitude ranges (*C. walie*), temperate environments (*C. aegagrus*) and hot deserts (*C. nubiana*). Cases of *Capra* species hybridization have been reported; for instance, domestic goats hybridize with *C. ibex, C. pyrenica, C. falconeri, C. sibirica, C. caucasica, C. cylindricornis* and *C. nubiana* with most hybrids being fertile (Alasaad *et al.,* 2012; Giacometti *et al.,* 2004; Hammer *et al.,* 2008; Lightner, 2006; Stuwe & Grodinsky, 1987).

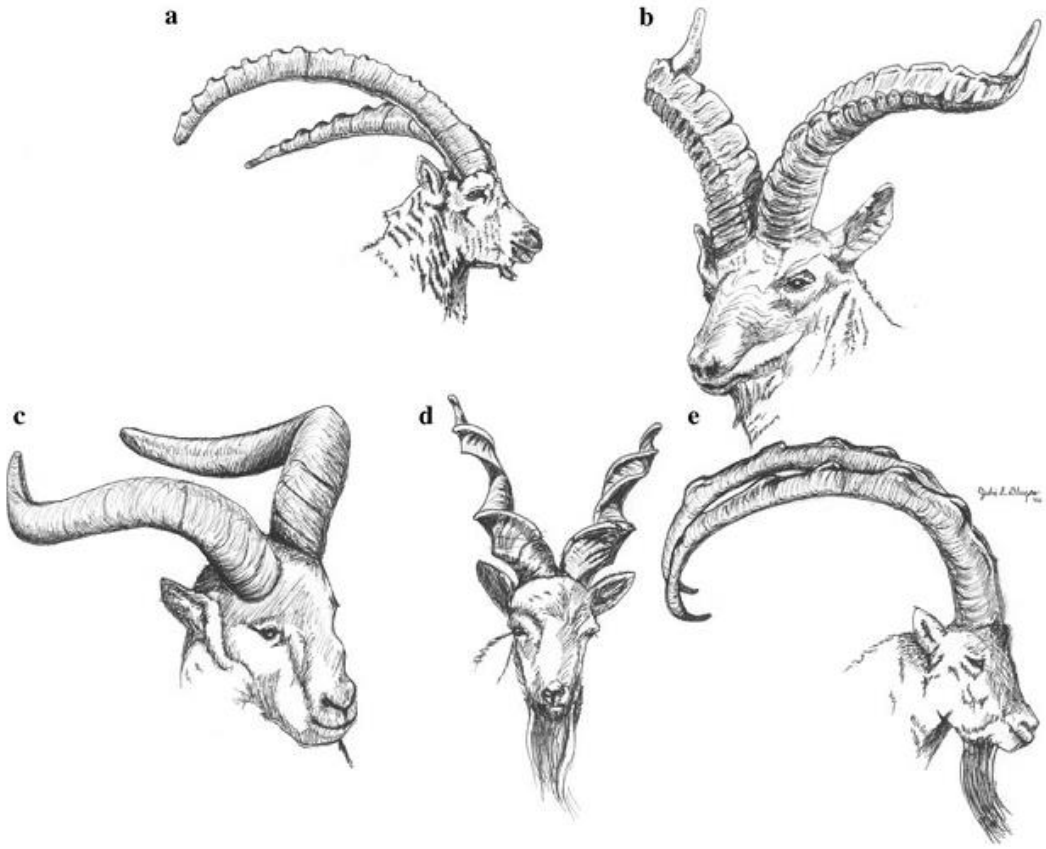


Figure 4: Capra species horn morphology. The major morphotypes include: (a) the generalized ibex-type (African ibex, C. sibirica, C. caucasica (b) C. pyrenaica (c) *C. cylindricornis*, (d) *C. falconeri*, and (e) *C. aegagrus* - Artwork (Pidancier *et al.*, 2006)

Goats are a source of meat, skin, fibre and milk; hence they are reared worldwide. In addition to being a source of livelihood to farmers across the globe, goats are potential animal models for investigating the genetic basis of complex traits such as adaptation to diverse environments. Due to goats economic and evolutionary importance, there has been growing interest in studying goat genomics. In the dawn of the genomics era, highly sophisticated genomics tools have contributed immensely to the advancement of livestock genomics studies (Groenen *et al.*, 2012; Jiang *et al.*, 2014; Qiu *et al.*, 2012; Zimin *et al.*, 2009). However, goat genomics work proceeded at a relatively slower pace than other livestock species such as cattle and sheep. In the year 2010, there was a concerted effort by the International community to promote goat genomic work; this gave birth to the International Goat Genome Consortium (IGGC), which facilitated the generation of a 52K SNP chip and sequencing of domestic goat genome (Du *et al.*, 2014; Tosser-Klopp *et al.*, 2014).

A year later, the wild progenitor Bezoar (*C. aegagrus*) genome was sequenced (Dong *et al.*, 2015). Availability of these initial draft genomes and SNP array data have facilitated several

studies geared towards understanding the goat genome and possibly pinpointing beneficial genetic traits (Benjelloun *et al.,* 2015; Dong *et al.,* 2015; Onzima *et al.,* 2018; Rahmatalla *et al.,* 2017). For instance, a genome comparison study between the domestic goat and its wild progenitor showed that genes associated with behavioural traits (5-hydroxytryptamine receptor 3A) and production traits (Fatty acid synthase and Low-density lipoprotein receptor-related protein 1) were under selection in domestic goats; a possible adaptation to domestications (Dong *et al.,* 2015). A genome study of three Moroccan indigenous goat populations showed that genes involved in fatty acids and lipids metabolism (Carnitine Palmitoyltransferase 1A, glyceronephosphate O-acyltransferase) and cellular stress response (TNF Receptor Associated Protein 1) were under positive selection (Benjelloun *et al.,* 2015). A genome scan of eight morphologically or geographically specific Chinese domestic goat populations showed that genes associated with coat colouration (agouti signalling protein), body size (T-Box 15), cashmere traits (Lim-Homeobox gene 2), and hypoxia adaptation (NADPH oxidase activator 1) were under strong selection pressures (Wang *et al.,* 2016).

The availability domestic goat reference genome and other goat project such as the AdaptMap (Stella *et al.,* 2018) have facilitated several other domestic goat genomic research in recent years (Bertolini *et al.,* 2018; Di Gerlando *et al.,* 2020; Guo *et al.,* 2018; Liu *et al.,* 2019). By utilizing large-scale data provided by AdaptMap (Stella *et al.,* 2018), and *C. hircus* genome, a CNV map for worldwide goats populations, was generated, which offers excellent resources for goat evolutionary studies (Liu *et al.,* 2019). More recently, a draft genome of wild goat (*C. ibex*) was released, forming additional goat genomics resources (Chen *et al.,* 2019). Generation of *C. nubiana* sequence data in this study adds to the available genomics resources for studying *Capra* species adaptive evolution (Chebii *et al.,* 2020).

CHAPTER THREE

MATERIALS AND METHODS

3.1 *Capra nubiana* genome sequence

The *C. nubiana* genome sequence and a draft assembly was generated in this first part of the research.

3.2.1 Samples

Liver tissue samples were collected from male *C. nubiana* from the National Zoological Garden biobank. Animal tissue samples collection procedures and ethical clearances were approaved by relevant South African governmental authorities (NZG/P14/13). The *C. nubiana* genomic DNA was isolated from the liver tissue samples using the standard phenol/chloroform extraction method (available in appendix 1). Briefly, DNA extraction using phenol/chloroform protocol involves; cell lysis, precipitation of proteins, removal of RNA, and precipitation of DNA. The DNA was quantified using Nanodrop and assessed for quality using gel electrophoresis in 1.5% agarose gel. The purified DNA was sequenced using the Illumina platform.

3.2.2 DNA extraction, libraries construction, and sequencing

DNA library was constructed using a TruSeq nano library prep kit following the manufacturer's protocol (https://support.illumina.com). Briefly, 200 ng genomic DNA was fragmented using a Covaris M220 instrument into 450 bp fragments. The fragments overhangs were end-repaired, adenylated, and ligated to DNA adapter sequences. The DNA fragments were hybridized into flow cells and enriched using Polymerase chain reaction (PCR) to amplify the amount of DNA library. Paired-end sequence reads, 125 bp in length were generated using Illumina Hiseq 2500 platform. Quality control analysis of the raw sequence reads were carried out using FastQC v.0.10.065 (Andrews, 2010). Poor quality reads and PCR duplicates were trimmed off using Trimmomatic v.0.32 (Bolger *et al.,* 2014) with parameters set to; minimum sequence length: 70, Require quality:15, the minimum quality required for 5' and 3' end: 14, clip seed mismatches: 2, clip threshold: 30.

3.2.3 Genome size estimation

The genome size was estimatied using GenomeScope (Vurture *et al.,* 2017) with parameters set to; K-mer size 77, read length 125 and maximum kmer coverage 10 000. A variety of k-mer lengths to be used as GenomeScope input was generated using Kmergenie.v.1.7044 (Chikhi & Medvedev, 2014), and the optimal k-mer length was selected.

3.2.4 De novo genome assembly

The paired-end sequence reads were assembled using using Soapdenovo2 v. r240 (Luo *et al.,* 2012) with following parameters: SOAPdenovo-127mer all -s configFile -p 6 -K 77 -R -F -o output_file and ABySS v.2.1.4 (Jackman *et al.,* 2017) with parameters set to: abyss-pe name=Ibex k=77 -j 12 np=24 v=-v in='R1.fastq R2.fastq'. The resulting assembly was improved using reference assisted *de novo* approach implemented using AlignGraph (Bao *et al.,* 2014) with parameters set to AlignGraph --read1 Read1.fa --read2 Read2.fa --contig Ibex-scaffolds.fa --genome c.hirucs.fa --distanceLow 375 --distanceHigh 1375 –extendedContig /Nubian-blat-extend.fa –remainingContig /Nubian-blat-remain.fa

3.2.5 Assessment of the genome assembly completeness

The genome assembly summary statistics such as the number of scaffolds and size, N50 statistics, and total assembly length were computed using QUAST v.4.3 (Gurevich *et al.,* 2013). Genome assembly coverage was assessed by remapping the sequence reads to the assembly using BWA-mem v. 0.7.15 (Li & Durbin, 2010), while Qualimap. v2.2.1 (García-Alcalde *et al.,* 2012) was used to extract alignment information. Kmer analysis tool (KAT).v.2.3.4 spectra-cn (copy number spectra) program was used to compare kmers in the assembly versus kmers in the sequence reads as a way of assessing assembly completeness in terms of kmers contents (Mapleson *et al.,* 2017). The BUSCO (benchmarking universal single-copy orthologs, version 2.0) tool was used to assess gene space content (Simao *et al.,* 2015). The tool inspected the assembled genome by searching for 4104 mammalian BUSCO gene groups expected to be present in all mammalian species.

3.2.6 Gene features prediction and function annotation

The assembled genome was masked for repetitive sequences using RepeatMasker v.4.0.5 (Smit *et al.,* 1996) with parameters set to RepeatMasker -qq -noint -par 8 -species mammal. Gene prediction was carried out using Augustus v.3.3 *complete mode* (Stanke *et al.,* 2006) with

a model trained using human datasets. The results were exported in text format in the general feature format (GFF). The predicted coding sequences were extracted from the GFF file and were used to create a protein fasta formatted database. A blastp search of the *C. nubiana* fasta formated database was carried out using domestic goat protein sequences with parameters set to; e-value 1e-6, the maximum number of hits were set to 1, coverage >70% and percentage identity >50%.

Gene functions were assigned according to best match of alignments using Blastp.v.2.2.30+ (Altschul *et al.,* 1990) against SwissProt database (Bairoch & Apweiler, 2000); blastp parameters were set to; e-value 1e-6, percentage identity>70% and coverage>70%. The motifs and domains of genes were determined by InterProScan.v. 5.25-64.0 with parameters -goterms and --pathway (Jones *et al.,* 2014), against protein databases including ProDom, PRINTS, Pfam, SMART, PANTHER and PROSITE, and corresponding Gene Ontology (GO) ID were obtained (Bairoch & Apweiler, 2000). The KEGG Orthology-Based Annotation System (KOBAS) v.3.0 (Xie *et al.,* 2011) was used to annotate gene sequences with KEGG; orthology terms by mapping them to known pathways in the KEGG pathway database (Kanehisa *et al.,* 2017).

3.2.7 Orthologs identifications

Protein-coding genes for cow, sheep, domestic goat, and horse were downloaded from Ensembl v.97 (Zerbino *et al.,* 2018). Bezoar (assembly CapAeg_1.0) and Alpine ibex (assembly IBX) were downloaded from Genbank (Clark *et al.,* 2016) and the protein-coding genes were predicted from genome data using similar protocols used in *C. nubiana* gene predictions. Single-copy gene orthologs shared among cattle, sheep, yak, domestic goat, horse, Alpine ibex, Bezoar and *C. nubiana* were then identified using Reciprocal-best-BLAST-Hits (RBH) (Wall *et al.,* 2003) using blastn with parameters set to 1e-10, coverage>70 and percentage id>50% (Altschul *et al.,* 1990). Pairwise orthologs were derived between domestic goat protein-coding genes, and each of the other eight species and an intersection across all the pairs was taken to construct combined gene orthologs across the nine species.

3.2.8 Phylogenetic analysis and divergence time estimation

Single-copy orthologous genes with length>500 bp shared by the nine species were aligned using PRANK program (Loytynoja, 2014). Aligned sequences were trimmed to remove unreliable alignment using BMGE (Block Mapping and Gathering with Entropy) (Criscuolo

& Gribaldo, 2010) using default parameters with option -t set to codons. The resulting alignments for each gene group were then concatenated into one supergene for each species and used as input for building the phylogenetic tree with GTR model using PhyML.v.3.3.2 (Guindon *et al.,* 2010). The branches' reliability was assessed using 1000 bootstrap replicates.

The *C. nubiana* divergence time was estimated using Reltime ML program (Tamura *et al.,* 2012) in Molecular Evolutionary Genetics Analysis (MEGA X) program (Kumar *et al.,* 2018); the horse was used as the outgroup. The time tree was computed using 5 calibration constraints; *C. nubiana* and *C. hircus* (1.17-6.65 million years ago), *C. nubiana* and *C. ibex* (1.25-5.61 mya), cow and sheep (22.17-29 mya), sheep and *C. hircus* (8.53-12.04 mya) and cow and Yak (2.72-6.46 mya). The reference divergence times used for calibrations were obtained from the Time tree database (Kumar *et al.,* 2017).

3.2 Detection of positive selection signatures in *C. nubiana* genome

The second experiment in this research study focused on SNV variant calling and detection of positively selected genes in *C. nubiana* genome.

3.2.1 Data sources

Genome sequence data for one *C. nubiana* was generated in sections 3.1.2 and 3.1.3, while for two others sampled from Sinai, Egypt and Howtat, Saudi Arabia (Grossen *et al.,* 2020) were downloaded from NCBI (NCBI Resource Coordinators, 2016) under accession number SRR8437789 and SRR8437792. The genome sequence data for *C. hircus* (domestic goat) was obtained from Ensembl (Zerbino *et al.,* 2018).

3.2.2 Single Nucleotide Variants and indels calling

The paired-end sequence reads were aligned to the *C. hircus* (domestic goat) genome using BWA-MEM v.0.7.15 default parameters (Li & Durbin, 2010). The SNVs and indels were called using SAMtools *mpileup* (Li *et al.,* 2009) with parameters set to *-Q 30 –q 30*. The *mpileup* output file was converted into Variant Call Format format using the BCFtools view program (Li *et al.,* 2009). The variant calls were filtered using vcfutils.pl varFilter module with the minimum and maximum read depths set to 6 and 100 reads, respectively (Li *et al.,* 2009). The transition-to-transversion (Ti/Tv) ratio was calculated using vcftools v.0.1.15 (Danecek *et al.,* 2011). The SNV sites common in the genomes of the three studied *C. nubiana* were obtained using a shell script.

3.2.3 Variant annotation (SNVs and InDels)

The variants were annotated using Variant Effect Predictor (VEP v.96) (McLaren *et al.*, 2016). Variants were classified based on their location within the genomic region (intergenic, intronic, untranslated regions and splice region variants) and the effect (synonymous or non-synonymous) they exact on the amino acid sequence of protein-coding genes. Effect types and functional classes were not mutually exclusive; for example, some variants were classified as both intronic and 5'-UTR.

3.2.4 The *C. hircus* and related species coding DNA Sequences (CDS) used in positive analysis

The *C. hircus* CDS were obtained from Ensembl BioMart (Kinsella *et al.*, 2011). The *C. nubiana* CDS, on the other hand, were generated by performing consensus mapping using *C. hircus* CDS; and subsequently substituting bases in the *C. hircus* CDS with the corresponding *C. nubiana* alleles in SNVs positions identified in Section 3.2.2. Pairwise alignments of the *C. nubiana* CDS and the corresponding *C. hircus* CDS were carried out, and a visual inspection was performed to confirm that the *C. hircus* alleles at the SNVs sites were correctly substituted with *C. nubiana* alleles.

The CDS for other species (cattle, goat, sheep, pig, yak, panda, bison, donkey, dog, horse, cat, and tiger) used as background data in positive selection analysis were acquired from Ensembl v.97 (Zerbino *et al.*, 2018). The water buffalo and Tibetan antelope CDS were extracted from the Genbank file (Clark *et al.*, 2016). Refer to Appendix 2 for data sources for all the species used for this experiment.

3.2.5 Single gene copy ortholog identification

The coding DNA sequences described in section 3.2.4 for *C. nubiana* and related species (cattle, goat, sheep, yak, bison, donkey, horse, donkey, cat, tiger, dog, pig, panda, water buffalo, dog and Tibetan antelope) were used for single gene copy ortholog identifications. Orthologous gene pairs were determined using the reciprocal best hit method, implemented in blastn program (Altschul *et al.*, 1990; Ward & Moreno-Hagelsieb, 2014). Blastn parameters were specified to: e *-evalue 1e-10 -perc_identity 60 -qcov_hsp_perc 70*. Pairwise orthologs between *C. nubiana* and each background species were derived using bash script, and then intersection across all the pairs was used to create a combined single-copy gene set. Gene

orthologs present in a minimum of seven species, including the *C. nubiana* and *C. hircus* were used for positive selection analysis.

3.2.6 Positively selected genes identification

The single-copy gene orthologs CDS (described in section 3.2.5) were translated into polypeptides using the mod_translate program (Wernersson & Pedersen, 2003). Poor quality sequences, including those with internal stop codons, were excluded. The polypeptides sequences were then aligned using the MUSCLE program v. 3.8.1551 (Edgar, 2004); the alignments were then used to guide CDS alignments using the RevTrans program version 1.4 (Wernersson & Pedersen, 2003). The CDS alignments were used to generate phylogenetic trees implemented in PhyML package v. 3.0 (Guindon *et al.*, 2010). The *C. nubiana* branch (foreground branch) labelling in each of the phylogenetic trees was implemented in ETE toolkit, v. 3.1.2 (Huerta-Cepas *et al.*, 2016). The branches for the other species were left unlabelled (background branches).

Multiple sequence alignment and the corresponding phylogenetic tree of each single-copy gene ortholog were used as input for positive selection analysis. Revised branch-site model A (Yang & Dos Reis, 2010) in CodeML program (Yang, 2007) was used to identify evolving genes in *C. nubiana* branch. Prior to positive selection analysis, the branches in the phylogeny tree were subdivided into the foreground (*C. nubiana*) and background branches (other related species; see section 3.2.4). The *C. nubiana* branch (foreground branch) was hypothesized to be having rapidly evolving sites, while sites in the background branches were evolving under negative or balancing selection. The CodeML control files for the alternative and null model used are provided in Appendix 4. A gene was considered to be under adaptive evolution if the Likelihood Ratio Test values are significant (p values<0.05) based on chi-square and if the omega (ω) >1. The sensitivity of the branch-site model in detecting selection signals is dictated by the taxa sample size (Anisimova *et al.*, 2002), for this reason, the initial candidate positively selected genes were reinvestigated using more even-toed ungulates data (each gene set had between 10-19 CDS including *C. nubiana* and *C. hircus*). The additional even-toed ungulates CDS data were obtained Genbank database. Positively selected sites under selection were identified using the Bayes Empirical Bayes (BEB) algorithm (posterior probability>0.8). The SNV sites detected in the three analyzed *C. nubiana* were compared to ascertain if the positively selected sites are *C. nubiana* specific.

3.2.7 Functional annotation and impact analysis of rapidly evolving genes

The biological processes associated with positively selected genes were downloaded from Ensembl Biomart (Kinsella *et al.,* 2011). While, enriched biological terms in positively selected genes were obtained from the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis *et al.,* 2003). Additionally, the functional consequences of the mutations in positively selected genes were investigated using Polymorphism Phenotyping-2 (Polyphen-2) program (Adzhubei *et al.,* 2013). Amino acid mutations with a score < 0.2 were considered benign, while scores between 0.2–01 were considered possibly or probably damaging.

3.3 Copy number variable genes identifications

The third insilico experiment in this study involved the investigation of copy number variations in *C. nubiana.*

3.3.1 Whole-genome sequence data

The *C. nubiana* data sources described in section 3.2.1 were in this experiment. Briefly, the data comprise of paired-end sequence reads generated in sections 3.1.2 and 3.1.3 and sequence data for two additional *C. nubiana* and the domestic goat were acquired from public databases described previously.

3.3.2 Copy number variants (CNV) calling

Binary Alignment Map (BAM) files generated by aligning sequence data for the three *C. nubiana* individuals in section 3.22. Were used as input data for CNV calling. The CNVs were called from the genomes of the three each of the *C. nubiana* individuals using CNVnator, a depth of coverage based algorithm (Abyzov *et al.,* 2011) by comparing it with the reference genome of *C. hircus* (domestic goat). The bin sizes of the CNVnator were set to 100 bp and 200 bp, respectively, with other parameters set to default.

The results were filtered such that only CNVs calls with a fraction of reads mapped with p-value < 0.05 and that > 1 kb in size were retained. To distinguish deletions from duplications events in *C. nubiana* and *C. hircus* genome following parameters were further applied to the filtered CNVs: A call with q0 > 0.7 and normalized read depth < 0.7 was regarded as duplication in *C. hircus* genome. Similarly, CNV with q0 > 0.7 and normalized read depth > 1.20 was deemed duplication CNV in *C. hircu*s and *C. nubiana*; however, the *C. hircus* has

more copies than *C. nubiana*. The CNV region with q0 < 0.2 and normalized read depth >1.5 was inferred as duplication in *C. nubiana*, while CNV with p- q0 < 0.2 and normalized read depth < 0.7 was considered as deletion in *C. nubiana*. Additional CNV events shared across the three analyzed *C. nubiana* with more than 50% overlap were considered for subsequent analysis.

Duplication events in *C. hircus* (domestic goat) reference genome were then validated by using dot plot analysis and blast (Altschul *et al.,* 1990). Briefly, DNA sequences corresponding to the candidate CNV regions were extracted using Bcftools –get fasta program; then aligned using an online NCBI dot plotter (https://www.ncbi.nlm.nih.gov/) to check if the CNV event is a tandem repeat. Similarly, CNV events were determined if they are segmental duplication using blastn (parameters set to evalue 1e-10, coverage >65% and identity>80%) (Altschul *et al.,* 1990). Genomic coordinates for CNVs discovered earlier on in goat populations (Di Gerlando *et al.,* 2020; Guan *et al.,* 2020) were compared with CNVs sites in *C. nubiana* using bedtools intersect (Quinlan & Hall, 2010). The CNV loci in *C. nubiana* with greater than 10% overlap with those in the domestic goat genomes were excluded from further analysis.

3.3.3 Evaluation of CNVnator sensitivity using artificial copy number variations

The CNVnator has been used mainly in CNV calling within species; hence the interpretation of copy numbers when carrying out CNV between species is unclear; for instance, using its default parameters, a read depth of less than 0.5 is interpreted as a deletion in the test genome; however, it could also be other genotypes like duplication in the reference genome. Insilico simulation experiment was conducted to find out appropriate cut-offs for CNVs and to assess CNVnator sensitivity in calling duplications in the reference genome. Briefly, two genomic coordinates per chromosome in copy number neutral regions in the *C. hircus* were duplicated. One of the simulated sites reflects two tandem duplications, while the other site represents four tandem duplications. The CNVs were called using CNVnator as described in section 3.3.2, with the artificially generated *C. hircus* genome acting as the reference.

3.3.4 Copy number variants sequence annotations

The CNVs sequence annotations were carried out using VEP v.96 (McLaren *et al.,* 2016) relative to the reference genome of the *C. hircus* (domestic goat). The CNV events were classified based on their respective location in the genome: non-coding, coding, downstream, upstream, intergenic and (UTR) sequences regions. The biological processes for CNV-

associated genes were acquired from Ensembl Biomart (Kinsella *et al.*, 2011) and literature. Enriched biological terms in CNV-associated genes were obtained from DAVID v. 6.8 (Dennis *et al.*, 2003). Since domestic goat genes are not yet available in DAVID, domestic goat Ensembl gene IDs were converted to the corresponding orthologous human Ensembl gene ID using Biomart (https://www.ensembl.org/biomart/). The human Ensembl gene IDs were used for gene enrichment analysis as described above.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Results

4.2.1 *Capra nubiana* sequence data

(i) Genome data

High-quality DNA with the ratio of absorbance at 260 nm to absorbance at 280 nm being 1.8-2 were isolated from *C. nubiana* liver tissue samples. Whole-genome sequencing was performed using genomic DNA from one individual goat using Illumina Hiseq 2500 sequencing platform. The sequencing process generated 912 929 400 of 125 base pairs Illumina raw paired-end reads with sequence coverage of 43.9x and an insert size of 450 bp. Following quality trimming, 781 955 700 high-quality reads were retained for subsequent analysis. The genome sequence data is available at NCBI (Bioproject accession: PRJNA674751). A summary statistics of the sequence reads data, estimated genome coverage before and after trimming is provided in Table 1.

Table 1: *Capra nubiana* sequence data summary statistics

Sequence reads information	Read length(bp)	Total number of reads	Number of bases (bp)	Sequence coverage
Paired-end raw sequence reads	125	912 929 400	114 116 175 000	43.9x
Paired-end trimmed sequence reads	125	781 955 700	92 871 389 869	37.6x

Coverage was estimated using the formula C=L*N/G (where G is haploid genome length), L is read length and N is the number of reads. Coverage was calculated using estimated *C. nubiana* genome size (2.6 Gbps).

(ii) Estimated genome size

The estimated *C. nubiana* genome size based on GenomeScope (Vurture *et al.*, 2017) was 2.6 Gbps as shown in Fig. 5.

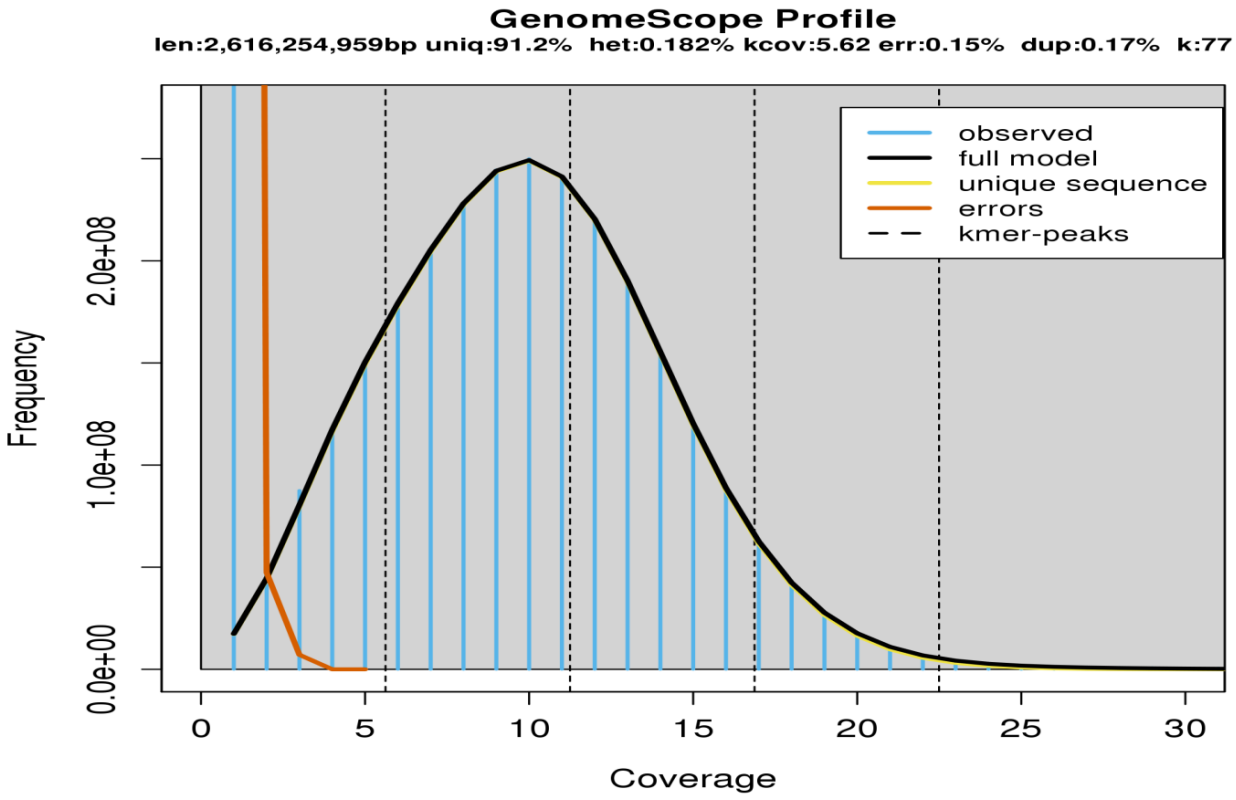


Figure 5: GenomeScope K-mer profile plot of *Capra nubiana* genome. The abbreviation 'len' is inferred genome length in base pairs

(iii) Genome assembly

Capra nubiana genome was *denovo* assembled using two different de bruijn graph assemblers; Soapdenovo2 (Luo *et al.*, 2012) and ABySS (Jackman *et al.*, 2017). The *denovo* assemblies statitistics summary is shown in Table 2. The ABySS genome assembly performed better since it had a larger N50 of 13,812 bp than the Soapdenovo2 assembly, which had an N50 value of 9368. The genome length of the ABySS assembly was 2,813,437,185 bp, while the Soapdenovo2 assembly was 2 308 354 460 bp. Genome assembly assessment using KAT plots and BUSCO analysis (not provided for Soapdenovo2) showed that ABySS assembler performed better; hence, subsequent analysis was carried out using ABySS assembled genome.

Reference-*denovo* assisted assembly was attempted to improve the assembly generated using ABySS, however it yielded poor assembly (results not included). For example, the largest contig produced by the aligngraph algorithm was 150 046 bp and N50 was 10 597 bp.

Table 2: Genome assembly's statistics based on ABySS and Soapdenovo2 genome assemblers

Assembly	ABySS : Ibex-scaffolds	Soapdenovo2: Ibex-scaffolds
Contigs (>= 0 bp)	3 592 580	984 977
Contigs (>= 1000 bp	251 601	348 150
Total Length (>= 0 bp)	2 813 437 185	2 308 354 460
Total Length (>= 1000 bp)	2 296 515 906	2 183 005 339
# Contigs	363 780	403 679
Largest Contig	189 973	135 758
Total Length	2 375 851 108	2 223 687 367
GC content (%)	41.88	41.82
N50	13 812	9 368

(iv) Genome assembly assessment

Remapping the sequence reads to the assembly showed that 99.91% of the sequence reads mapped back. KAT spectra-cn (copy number spectra) (Mapleson *et al.*, 2017) analysis showed that *C. nubiana* assembly had all the sequence reads contents represented in the assembly (i.e; red histogram bars) (Fig. 6), an indication of a good assembly. The genome assembly was mostly homozygous as shown by red bars appearing in one copy in Fig. 6.

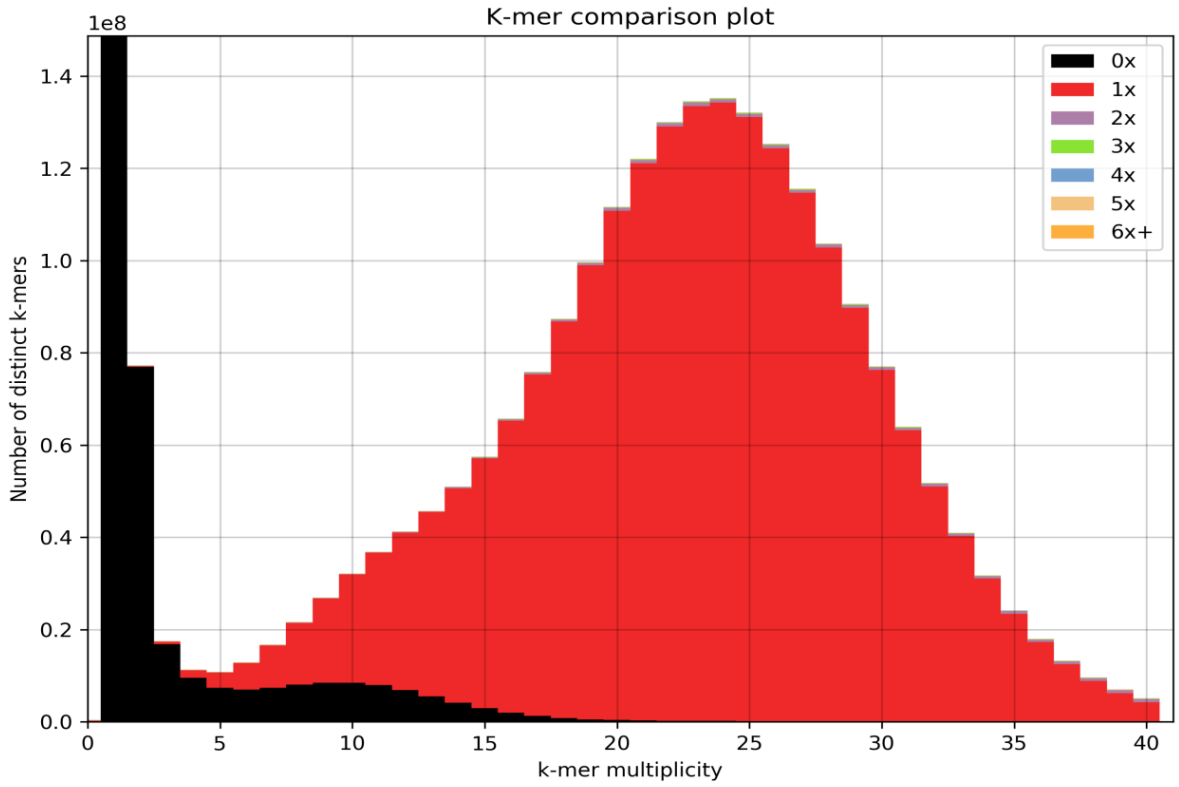


Figure 6: Kmer-spectra plot generated using Kmer analysis tool (KAT) showing motif and copy number representation in *C. nubiana* genome. The colored plot shows how many times fixed-length words (k-mers) from the sequence reads appeared in the assembly; frequency of occurrence (multiplicity; x-axis) and the number of distinct k-mers (y-axis). Black represents sequence reads absent in the assembly; red, sequence reads that appear once in the assembly; green, twice. The plot was generated using k = 31. Sequence reads representation (red bars, 1 copy in the assembly) shows the assembly is good. The level of heterozygosity was low (black peak at k-mer multiplicity 12); heterozygous content was collapsed by ABySS assembler. The long black bars at the y-axis represent k-mers at low frequencies (usually sequencing errors) which were not assembled hence a good indication of good assembly

The BUSCO.v3 mammalian gene dataset showed that 51.6% complete and 28.8% fragmented of the 4,104 BUSCO gene groups were present in *C. nubiana* genome. While, 19.6% of the 4,104 BUSCO were missing in *C. nubiana* genome. BUSCO analysis statistics is provided in Table 3.

Table 3: Completeness of *C. nubiana* genome assembly as assessed by BUSCO

Mammalian BUSCOs groups	Number of BUSCO (percentage)
Complete and single-copy BUSCOs	2 102 (51.2%)
Complete and duplicated BUSCOs	16 (0.4%)
Fragmented BUSCOs	1 182 (28.8%)
Missing BUSCOs	804 (19.6%)
Total BUSCO groups searched	4 104 (100%)

(v) Protein-coding genes predictions

A combination of *de novo* and similarity-based approaches were used to predict protein-coding sequences in *C. nubiana* genome. *De novo* gene predictions resulted in 39 190 protein-coding genes; 25 674 out of the 39 190 had significant similarity support. The largest predicted gene was 22 380 amino acids in length with the majority of the genes being less than 300 amino acids in length. A total of 19 065 protein-coding genes were assigned function by mapping it to the SwissProt database, while domain annotation using Interproscan assigned a total of 19 001 Pfam domains, 8 545 SMART domains, and 23 562 were assigned to GO terms in PANTHER database. The predicted protein-coding genes and their annotation have been deposited in (https://doi.org/10.6084/m9.figshare.11777595).

(vi) Phylogenetic analysis

The phylogenetic position of *C. nubiana* was determined using the maximum likelihood approach implemented in PhyML v.3.0 using GTR, and 1000 bootstrap replicates (Guindon *et al.,* 2010). A concatenated alignment of 802 orthologous genes shared among nine species; *C. nubiana*, *C. ibex*, Bezoar, domestic goat, cow, sheep, and horse were used to generate the phylogeny tree. The phylogenetic analysis showed that *C. nubiana* had a close relationship with *C. ibex,* and they diverged from each other around 4.75 mya. *Capra nubiana* diverged from domestic goat and *Capra aegagrus* 5.61 mya as shown in Fig 7a. The estimated divergence time from the present (million years ago; mya) is given at the nodes. The generated *C. nubiana* phylogeny shared a few similarities with the previous *Capra* species phylogeny tree obtained from the Timetree database (Kumar *et al.,* 2017) shown in Fig 7b; for instance, *C. ibex* and *C. nubiana* shared the same clade if *C. pyrenaica* is collapsed from the previous

phylogeny tree. The phylogeny trees also show that *C. hircus* and *C. aegagrus* share the same clade as expected.

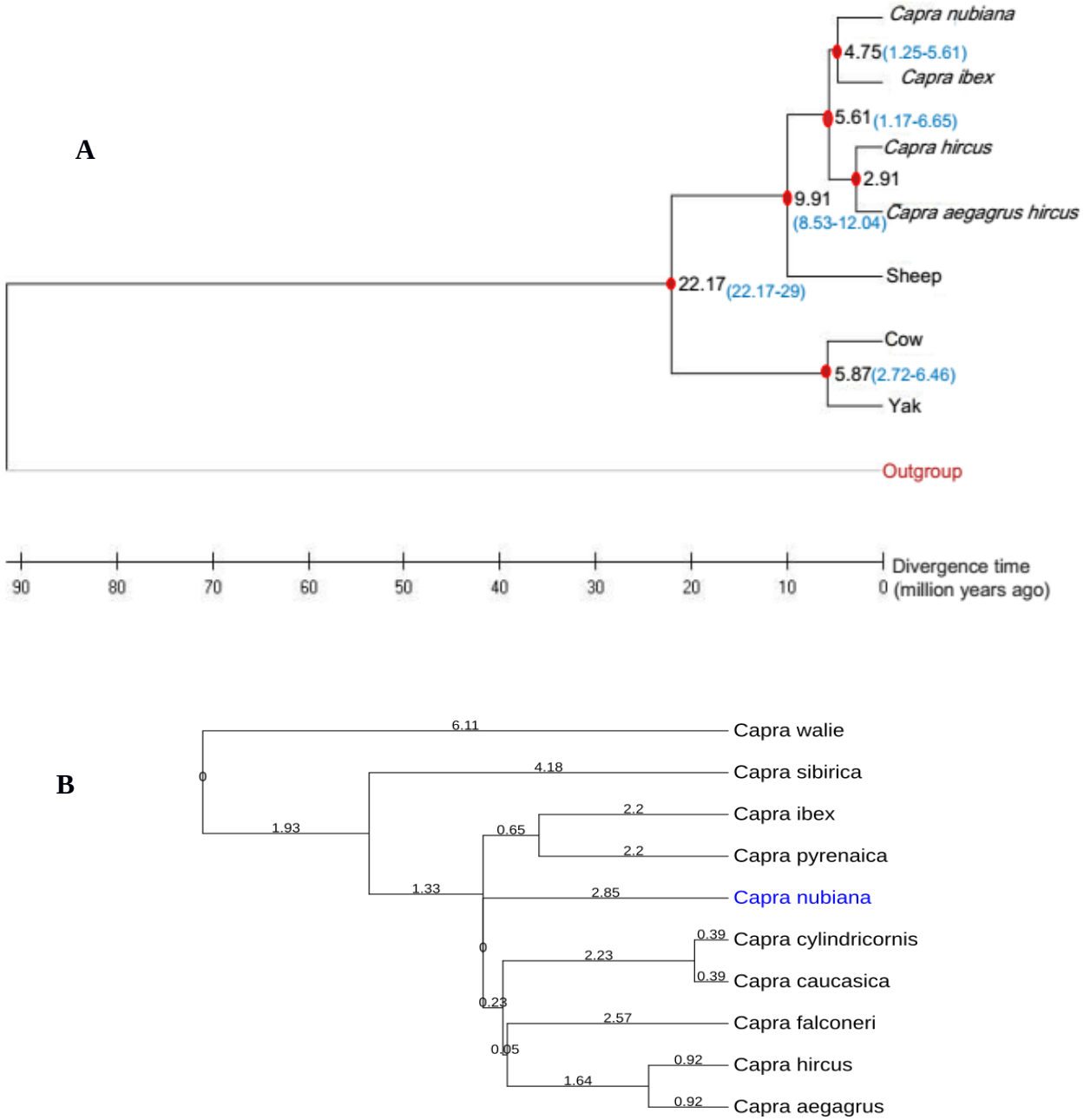


Figure 7: ***Capra* nubiana phylogenetic position. A. Phylogenetic tree of *C. nubiana* and its relatives. *Capra* species generated in this study. B. *Capra* species phylogeny tree was obtained from Timetree database (Kumar *et al.*, 2017). The figures on the branches shows the time in million years (mya) since divergence of Capra species from a common ancestor**

4.2.2 The SNVS and positively selected genes

(i) The SNVs and InDels calling in *C. nubiana*

Approximately 98% of the 781 955 700 *C. nubiana* genome sequence reads aligned onto unique sites in *C. hircus* genome. A total of 1 726 573 InDels and 19 468 467 SNV sites were detected from the sequence data of *C. nubiana* generated in this study. While 21 851 698 and 22 446 813 SNV sites were called from sequence data of the other two *C. nubiana* obtained from public database. Comparison of all the SNV sites identified in this study showed that 15 672 749 (81%) of it were shared across the three analyzed *C. nubiana* individuals. Further analysis of SNVs called from *C. nubiana* generated from this study showed that the 3 024 701 of the sites that were heterozygous while 16 443 766 that were homozygous. SNV sites were classified as transitions (13 720 855, Ts: G/A and C/T) or transversions (5 731 167, Tv: T/A, C/G, G/T and A/C); the Ts/Tv ratio was 2.39. The InDels comprised of 247 356 heterozygous sites and 1 479 217 homozygous sites; 903 070 were insertions, 820 893 were deletions and 2610 were both insertion and deletion at a given region. Deletions and insertions length ranged from 1-55 bp and 1-38 bp, respectively.

Variant annotations showed that a large percentage of the variants were in intergenic (69.2% SNVs and 69.1% InDels) and intronic (29.60% SNVs and 30.1%) genomic regions (Fig. 8). Only 0.7% of the SNVs and 0.1% InDels were found in exonic regions. The SNVs and InDels which affected: (a) splice acceptor or donor sites, (b) stop codons (gain or loss), start codons, or (iv) frameshift insertion or deletion were classified as 'high-impact' variants since they lead to protein truncation or loss of function (Rausell *et al.*, 2014). A sum of 1706 SNVs affecting 1373 gene sequences and 1350 InDels affecting 1141 genes were classified as high-impact variants.

A total of 57080 missense SNVs and 871 inframe InDels were categorized as having a moderate impact, and they overlapped with 13191 and 800 gene sequences respectively. Synonymous and splice region variants were classified as low-impact variants. In addition, 87977 SNVs overlapped with UTRs of 6723 protein-coding genes, while 9756 InDels overlapped with UTRs of 4021 protein-coding genes. Variants in the intergenic, intronic, UTRs, upstream and downstream regions of the genome were classified as modifiers. Summary statistics showing the distribution of SNV and InDels in the various genomic region and the type of impact for each variant type is provided in Table 4.

Additional genetic variants data in *C. nubiana* genome is available at Figshare (https://figshare.com/s/3041e34bc83934ba5797).

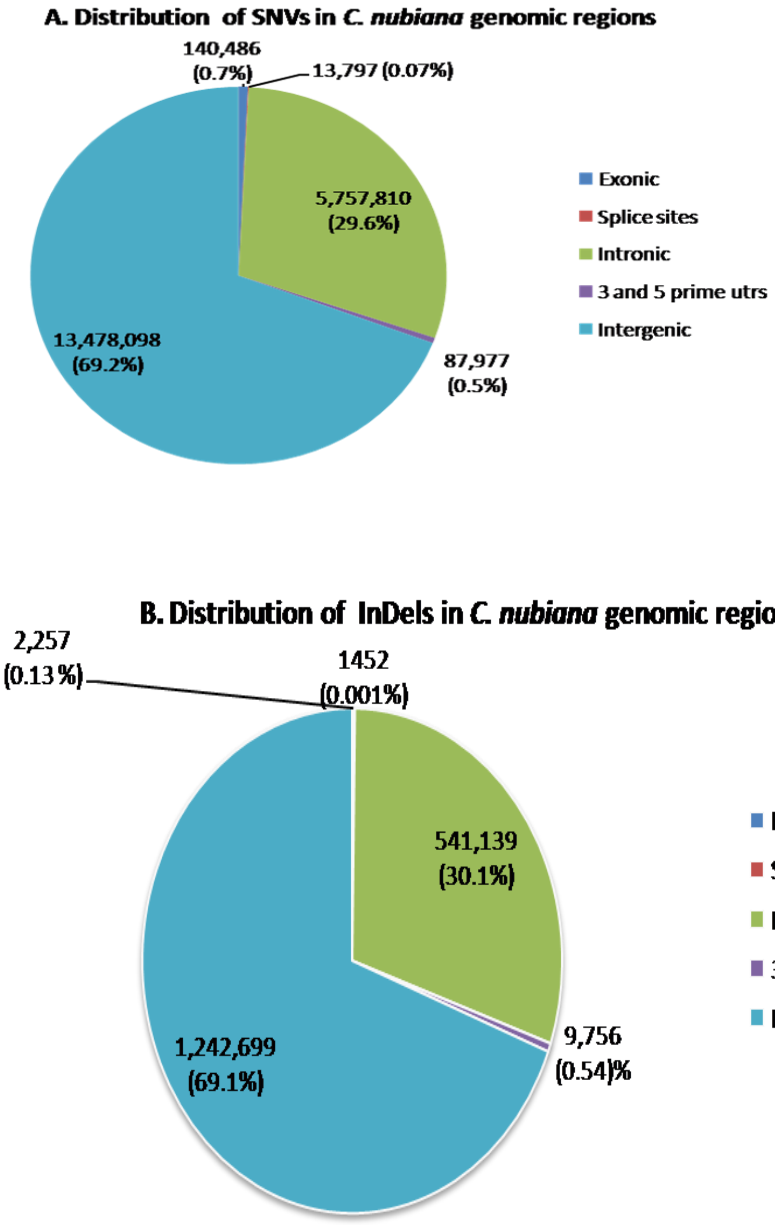


Figure 8: Distribution of SNVs and INDELS in *C. nubiana* genomic regions (introns, coding regions, intergenic and untranslated regions. The largest percentage (69%) of the SNVs and InDels are in the intergenic regions, followed by intronic regions (30%)

Table 4: Summary of SNVs and InDels sequence annotation detected in *C. nubiana* sequenced in this study

Variant category	Variant type	No. of Hom SNVs	No. of Het SNVs	No. of Hom InDels	No. of Het InDels	Impact
Exon	Stop gain	377	416	14	4	High
	Stop loss	76	37	4	1	High
	Start loss	150	0	16	3	High
	Frameshift	-	-	851	192	High
	Missense	40 206	16 874	-	-	Moderate
	Inframe deletion	-	-	381	147	Moderate
	Inframe insertion	-	-	269	74	Moderate
	Synonymous	65 427	15 979	-	-	Low
	Stop retained	55	17	-	-	Low
	Non coding transcript exon	462	341	32	10	Modifier
	Coding	33	3	137	33	Modifier
Splice site	Splice acceptor	197	66	128	15	High
	Splice donor	301	86	105	17	High
	Splice region	10 966	2 181	1026	161	Low
Intronic	Intron	4 895 896	861 914	467 063	74 076	Modifier
UTR	3 prime UTR	59 544	12 670	7121	1 135	Modifier
	5 prime UTR	12 995	2 768	1279	221	Modifier
Intergenic	Upstream gene	449 251	98 777	61 199	11 627	Modifier
	Downstream gene	464 156	103 621	64 810	10 691	Modifier
	Intergenic	10 445 066	1 917 227	907 683	186 689	Modifier

No. Het (Number of Heterozygous), No. Hom (Number of Homozygous)

(ii) Genes under adaptive evolution in *C. nubiana*

A total of 19 418 *C. nubiana* CDS were generated by projecting *C. nubiana* SNVs alleles to *C. hircus* CDS. The bash script used generating *C. nubiana* CDS, corresponding domestic goat CDS and SNVs annotation file is provided online at Figshare (https://doi.org/10.6084/m9.figshare.11954382.v1). Fifteen thousand five hundred and twenty-seven (15 527) gene orthologs shared between *C. nubiana*, and its related species were identified. Initial positive selection (dN/dS) analysis using 7-15 species in the background branches showed that 34 protein-coding are under adaptive evolution in C. nubiana. A re-run of dN/dS analysis using more background taxa data (10-19 species) confirmed that 28 out of the 34 candidate genes were under adaptive evolution. Bayes empirical Bayes (BEB) indicated that 43 amino acid sites in 22 evolving genes were positively selected. Validation of the SNV sites in 22 positively selected genes showed that 98% of the sites are shared across the three

analyzed *C. nubiana*. Functional impact analysis showed that 17 mutations were likely to alter the protein function and structure, while 13 were neutral. Positively selected genes and corresponding dN/dS ratios, Likelihood ratio test (LRT) and p values are provided in Table 5. A list and corresponding metadata of genes shown to be under adaptive evolution in *C. nubiana* are available in Appendix 4 and Appendix 5.

Table 5: Genes displaying strong positive selection signals in *C. nubiana*

Gene name	Ensembl goat gene id	ω0	ω*f*	LRT	p-value
Zinc finger and SCAN domain containing 23	ENSCHIT00000026283	0.078	999	13.004	0.0003
Olfactory receptor 2G2-like	ENSCHIT00000004434	0.11	999	11.661	0.0006
Atpase H+ transporting V1 subunit E2	ENSCHIT00000004084	0.07	999	11.12	0.0009
F-box protein 21	ENSCHIT00000018881	0.006	999	9.379	0.0022
Achaete-scute family bhlh transcription factor 4	ENSCHIT00000040177	0.113	999	9.094	0.0026
Toll like receptor adaptor molecule 2	ENSCHIT00000015914	0.042	615.743	9.027	0.0027
Olfactory receptor 1P1	ENSCHIT00000040379	0.067	471.664	8.818	0.003
Tripartite motif containing 16	ENSCHIT00000034768	0.093	223.251	8.301	0.004
Centrosomal protein 112	ENSCHIT00000041152	0.074	129.156	7.917	0.0049
Storkhead box 2	ENSCHIT00000003090	0.03	335.66	7.544	0.006
Matrix AAA peptidase interacting protein 1	ENSCHIT00000010253	0.028	580.881	6.893	0.0087
F-box and WD repeat domain containing 2	ENSCHIT00000030384	0.025	83.055	6.817	0.009
ATP binding cassette subfamily A member 12	ENSCHIT00000028741	0.053	255.305	6.775	0.0092
PATJ crumbs cell polarity complex component	ENSCHIT00000035903	0.084	322.085	6.519	0.0107
Rho gtpase activating protein 42	ENSCHIT00000036547	0.045	106.727	6.432	0.0112
Multimerin 2	ENSCHIT00000000612	0.11	142.837	5.562	0.0184
Serine protease 56	ENSCHIT00000008957	0.068	999	5.552	0.0185
LY6/PLAUR domain containing 6B	ENSCHIT00000020934	0.073	45.53	5.29	0.0214
Eukaryotic translation initiation factor 2 subunit beta	ENSCHIT00000016318	0.021	183.165	5.14	0.0234
UV stimulated scaffold protein A	ENSCHIT00000028977	0.091	156.533	5.017	0.0251
Prostaglandin I2 synthase	ENSCHIT00000015750	0.061	105.035	4.46	0.0347

ω0 =background branches, ω*f*= foreground and LRT= Likelihood ratio test

Table 6: Positively selected sites that likely to change protein structure and function based on Polyphen-2 analysis

Gene Name	Positively Selected sites (Posterior probabilities > 0.8)		
	Ancestral residue	**Positively selected sites**	***C. nubiana* residue**
Atpase H+ transporting V1 subunit E2	M	72	N
Olfactory receptor 2G2-like	F	73	T
Serine protease 56	R	425	G
Putative olfactory receptor 52P1	M	67	L
Prostaglandin I2 synthase	R	320	H
F-box protein 21	K	616	R
Zinc finger and SCAN domain containing 23	P	213	N
UV stimulated scaffold protein A	D	361	G
F-box and WD repeat domain containing 2	L	82	C
Multimerin 2	S	214	H
Eukaryotic translation initiation factor 2 subunit beta	K	83	I
ATP binding cassette subfamily A member 12	M	570	T
	I	1738	F
Rho gtpase activating protein 42	W	773	R
Achaete-scute family bhlh transcription factor 4	L	30	S
Olfactory receptor 1P1	H	159	C

The table shows specific sites of genes that are targets of selections (posterior probability>0.5). Sites with high posterior pobabilities>0.8 are interpreted as a sign of strog positive selection at that given site. The mutations are functionally consequential (Polyphen-2 score>0.7).

Positively selected genes discovered are associated with several molecular functions, including signal transduction, protein binding, transcription, translation, serine-type endopeptidase activity, transmembrane transport, G protein-coupled receptor signalling pathway, and ATP binding. The evolving genes participate in diverse biological processes, including prostaglandin metabolic processes, protein ubiquitination, camera-type eye development, transmembrane transport, positive regulation of the Notch signalling pathway, blood pressure regulation, and spermatogenesis. The gene ontology terms are provided in Appendix 6. Notably, genes involved in developing the skin barrier and DNA repair, which may have an adaptive role, were reported to be under adaptive evolution in *C. nubiana.*

Adaptive evolution in skin development and DNA repair genes

Skin and hair follicle development genes (*ASCL4* and *ABCA12*) were under adaptive evolution in *C. nubiana*. The *ABCA12* gene had an amino acid change at position 570 with posterior

probability >0.9, classified as functionally consequential (Polyphen-2 score >0.7). The mutation (M570T) is outside the *ABCA12* functional domains. *ABCA12* is associated with lipid transport activity, ceramide transport, keratinocyte differentiation, and skin barrier establishment. The *ABCA12* gene tree showing Nubian ibex (*C. nubiana*) branch (foreground) and other related species (background branches), and multiple sequence alignment data showing positively selected site (M570T) is depicted in Fig. 9.

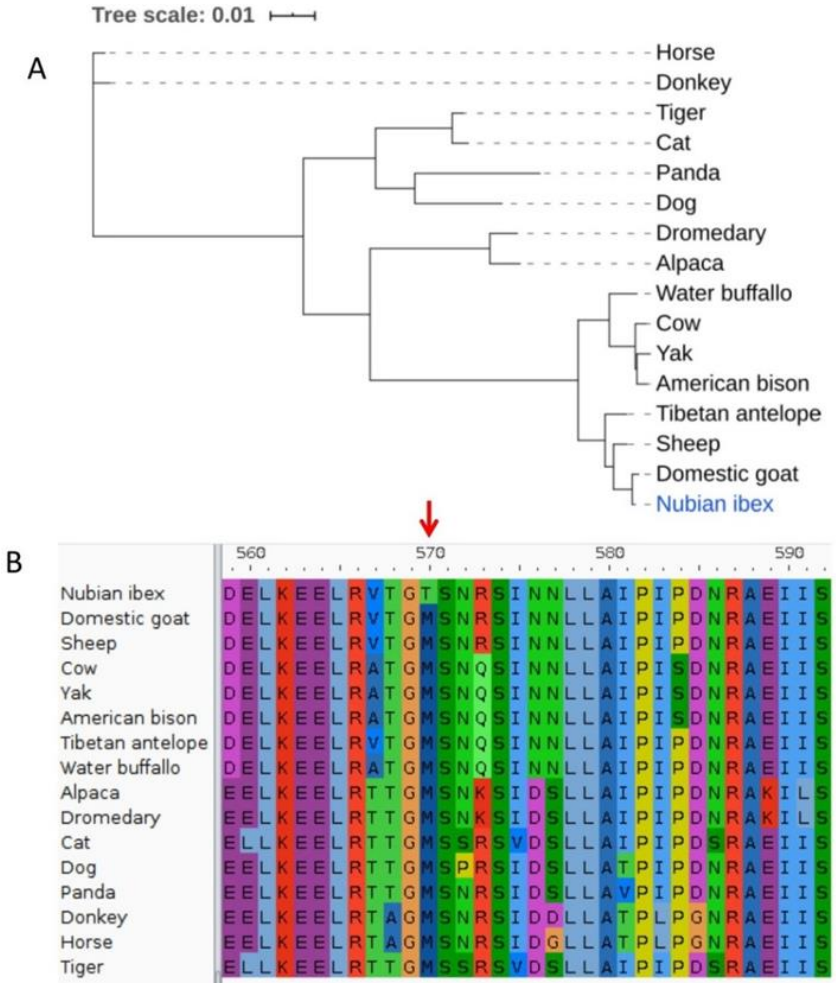


Figure 9: ABCA12 phylogenetic tree and multiple sequence alignment data used for dN/dS analysis. The multiple sequence aligment shows a mutation at position 570 of ABCA12 gene in C. nubiana classified as functionally important (Polyphen-2 score> 0.7)

Achaete-scute family (basic helix-loop-helix) bHLH transcription factor 4 (*ASCL4*) were shown to be rapidly evolving in *C. nubiana*. The *ASCL4* gene had one mutation (S30L) located outside of the gene functional domain, predicted as functionally important (Polyphen-2 score > 0.99). The *ASCL4* gene is Achaete-Scute basic helix-loop-helix (bHLH) transcriptional regulatory proteins together with *ASCL1, ASCL2, ASCL3* and *ASCL5* (Wang *et al.*, 2017). The *ASCL1* and *ASCL2* genes are involved in neural crest cells differentiation, while *ASCL3 and*

ASCL5 play key roles in generation of salivary gland duct cells and brain, respectively (Ball *et al.,* 1993; Wang *et al.,* 2017). The *ASCL4* is associated with gene ontology terms such as regulation of transcription and hair follicle development. The *ASCL4* gene tree showing Nubian ibex (*C. nubiana*) branch (foreground) and other related species (background branches), and multiple sequence alignment data showing positivel selected site provided in Appendix 7.

Additionally, UV-stimulated scaffold protein A (*UVSS*A) was reported to be under positively selected. Two positively selected sites with posterior probability >0.9 were reported at positions 361 and 517. The mutation at position 361 was classified as functionally important (Polyphen-2 score>0.99), while the amino acid change at position 517 is neutral. The *UVSSA* gene is key in some biological processes such as protein ubiquitination, response to ultraviolet radiations (UV), and transcription-coupled nucleotide excision. The *UVSSA* gene tree showing Nubian ibex (*C. nubiana*) branch (foreground) and other related species (background branches), and multiple sequence alignment data showing positively selected site provided in Appendix 7.

4.2.3 Copy number variable regions

(i) Evaluation of CNVnator sensitivity

Tandem duplications at 58 sites representing 2 sites per chromosome were simulated. The CNVnator successfully called 38 to 43 artificial CNVs across the three analyzed *C. nubiana* genomes, indicating that CNVnator has a 66-75% sensitivity. As expected, the normalized read depths for the artificial CNVs regions were 0.5 or less. Notably, the normalized read depths for simulated CNVs with 2 copies of tandem repeats were approximately 0.5, while those with 4 copies of tandem repeats were 0.25. All the artificial CNVs had a mean zero mapping quality (q0) between 0.7 and 0.99, implying that any CNV site with a q0 score greater than 0.7 is a true CNV reflecting copy number gain events in the reference genome. The CNVnator evaluation showed that it is not precise in determining CNVstart and end positions, with CNV calls being on average 1654 base pairs off from the actual positions.

(ii) The CNVs in *C. nubiana* and domestic goat genomes

The CNVnator identified 13472, 9064 and 7724, raw CNV sites in *C. nubiana* individuals from South Africa, Saudi Arabia, and Egypt, respectively (available at Figshare (https://doi.org/10.6084/m9.figshare.13633943)). A total of 4504 CNVs detected in *C. nubiana* genomes were retained after stringent filtering; CNV retained for each *C. nubiana* is summarized in Table 6. Detailed data for the CNVs are available at Figshare (https://doi.org/10.6084/m9.figshare.13633943). The CNVs (27) discovered earlier on in the *C. hircus* were excluded. Three hundred and sixty-seven (367) CNV loci common to the three *C. nubiana* analyzed in this study (Table 7) were subjected to further analysis. The 367 CNV covered < 1% (5.6Mbp) of the *C. nubiana* genome. The sizes of the 367 CNVs were between 1100bp and 214000bp (see Fig. 10), and the number of copies was between 0 and 463.

Table 7: A summary of the total CNVs detected in *C. nubiana* and *C. hircus* genomes

CNV type	***C. nubiana* sample origin**			**Number of common CNVs in the three *C. nubiana* individuals**
	South Africa (SA)	Egypt (E)	Saudi Arabia (A)	SA_E_A (intersect)
Loss of copy CNVs in *C. nubiana*	971	750	682	97
Gain of copy CNVs in *C. nubiana*	483	543	424	206
Gain of copy CNVs in *C. hircus*	251	234	125	62
Gain of copy CNVs in *C. hircus* with more copies in *C. nubiana* genome	21	17	3	2
Total CNVs	1726	1544	1234	367

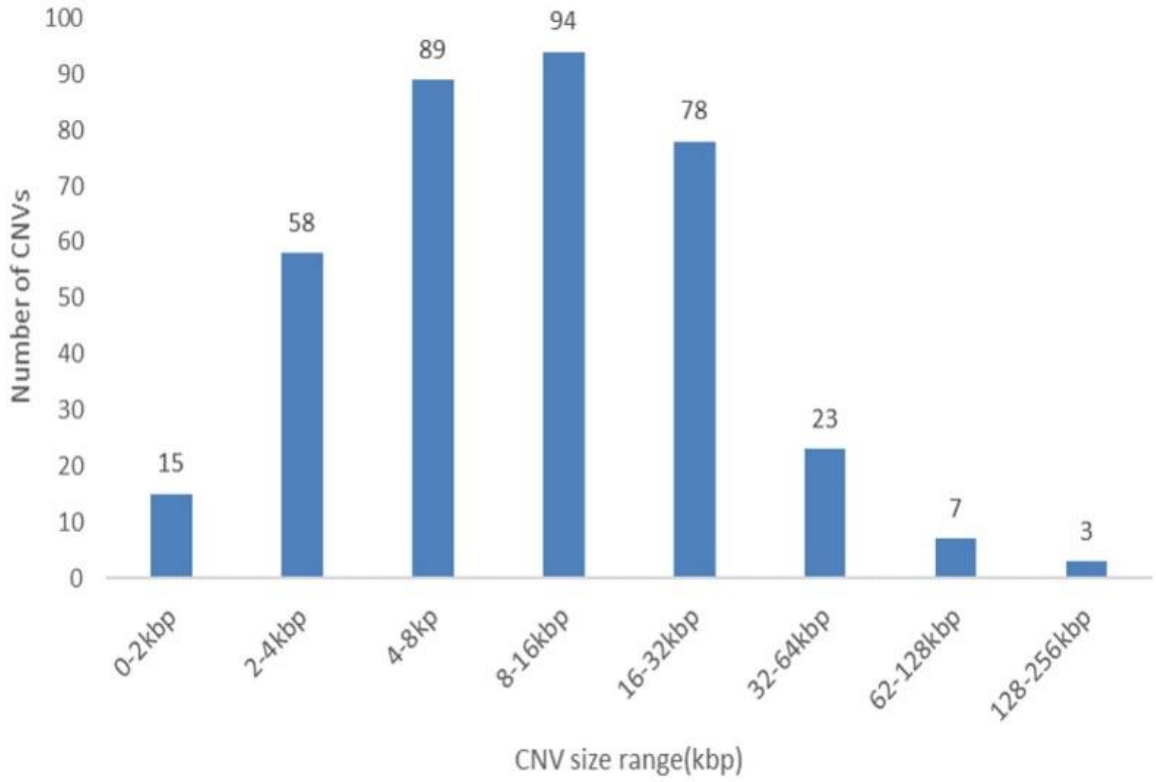


Figure 10: The CNV size distribution in *C. nubiana* genome

(iii) The CNVs distribution across *C. nubiana* genome

The CNVs were distributed across various genomic regions with 36% (161) overlapping with intergenic regions of the genome. Approximately 14% (62) CNVs were found in gene exons, while 13% (60) were located in the intronic regions (Fig. 11). A detailed description of the locations of CNVs in various genomic regions (exons, introns, intergenic, upstream gene regions, downstream regions and untranslated genomic regions) is provided in the figshare link provided previously.

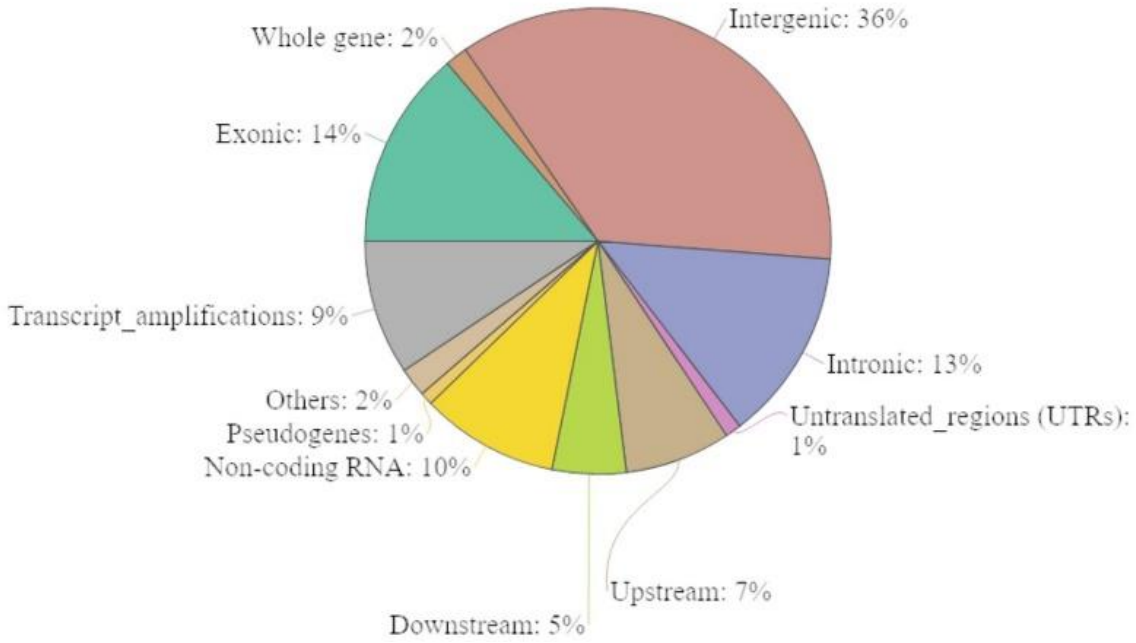


Figure 11: The distribution of the CNVs across various genomic regions in *C. nubiana.* The pie chat illustrates the genomic locations (coding sequence region, non-coding sequence regions and intergenic regions) of the 367 CNVs detected in *C. nubiana*

(iv) The CNV-associated genes

A total of 191 CNV-associated protein-coding genes and 20 non-coding genes were reported. The CNVs spanning genes were found mainly in introns, untranslated gene regions, and upstream gene regions, while CNVs in the coding sequence region overlapped with exons or entire genes. Ninety-six (96) CNVs were found in exons and upstream regions of genes. Five gain of copy number CNVs overlapped with five full genes in *the C. nubiana* genome. Figure 12 depicts examples of CNV-associated genes in *C. nubiana*. In addition, 83 CNVs were within intronic and downstream regions of the protein-coding genes. Cumulatively, 126 CNV associated-genes were in duplicated regions, and 47 were found in deleted regions in *C. nubiana* genome. The CNVs overlapping with protein-coding genes is provided in Appendix 8.

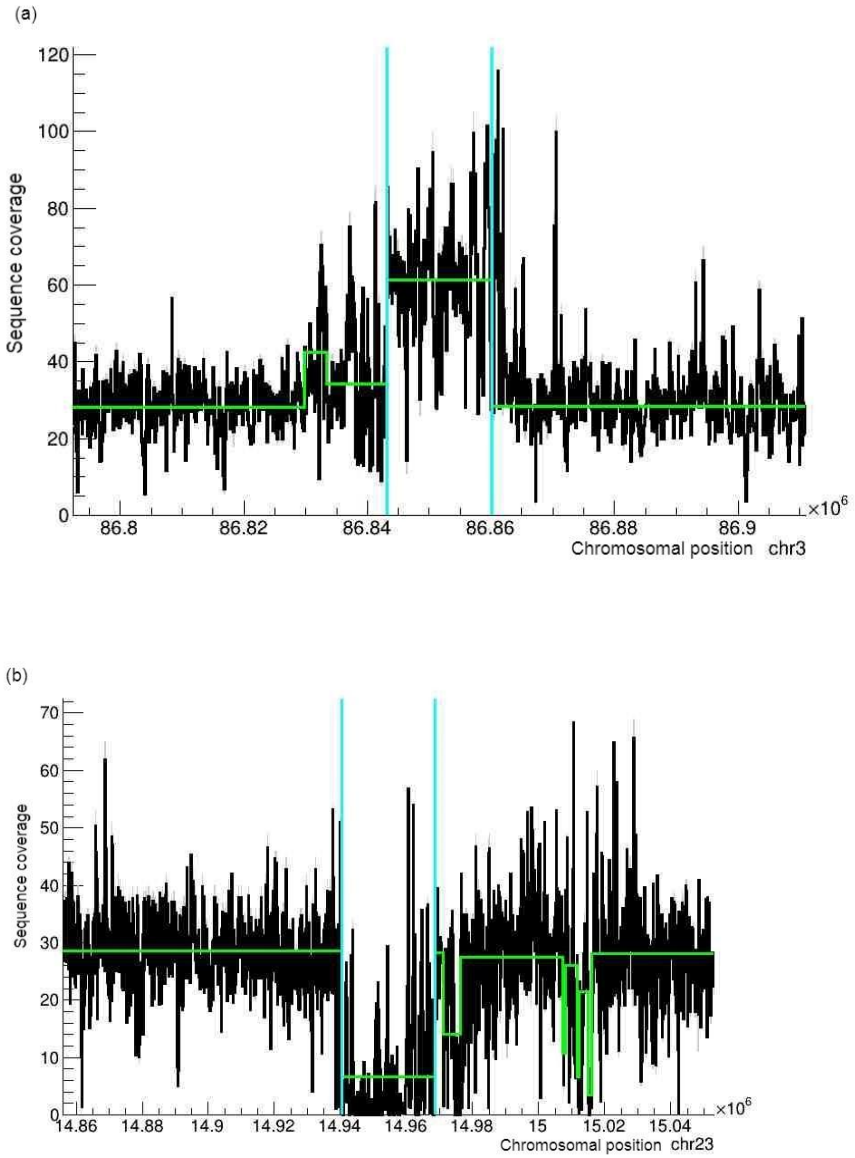


Figure 12: Read depth plots against chromosomes 3 and 24 illustrating gain and loss of copy number events. The read depth plots were generated using CNVnator –view program (Abyzov *et al.*, 2011); from sequence data of *C. nubiana* sampled from South Africa. (a) Gain of copy number event. The green lines indicate normalized read depth, while the section enclosed in blue vertical lines depicts gain of copy number region (chr3:86843200-86860100) in C. nubiana. The gain of copy number region overlaps the first two exons of GSTM4 gene, which is found in chr3:86846458-86878863. (b) Loss of copy number event. The green lines indicate normalized read depth, while regions enclosed in blue vertical lines depict a loss of copy number region (chr23:14940500-14968600) in *C. nubiana.* The loss of copy number region overlaps with SERPIN B6L gene, which is found at chr23:14938589-14955225: -1

Table 8: CNV-associated protein-coding genes in *C. nubiana* and *C. hircus*

CNV type	CNVs overlapping with coding sequence regions		CNV in non-coding sequence regions			
	Exonic	Whole gene	Introns	Upstream	Downstream	UTRs
Gain of copy CNVs in *C. nubiana*	3	2	30	8	4	0
Loss of copy CNVs in *C. nubiana*	54	5	22	22	19	4
Gain of copy CNVs in *C. hircus*	6	0	8	3	0	1
Total CNV genes	63	7	60	33	23	5

Six protein-coding genes were shown to have more copies of exons in the domestic goat, while eight others had more copies in their introns when compared to *C. nubiana*. In addition, three protein-coding genes were reported in the upstream CNVs regions in the *C. hircus* genome. Figure 13 depicts a dot plot for tandem repeats in the *C. hircus* genome.

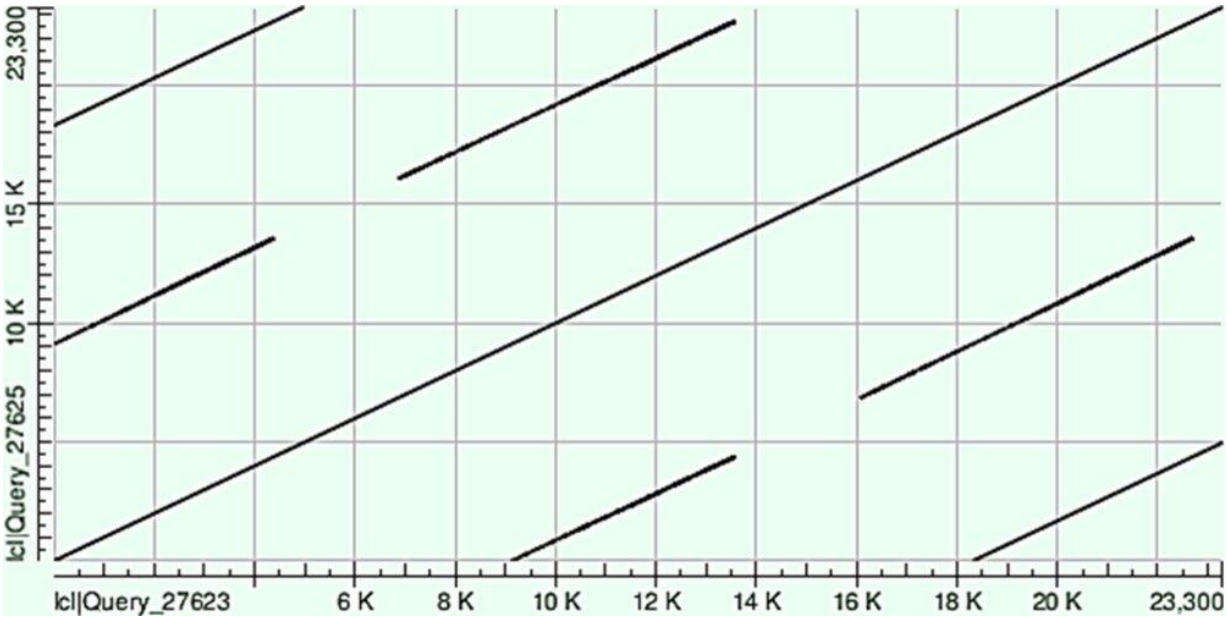


Figure 13: An illustration of four tandem repeats in C. hircus genome. The read depth of the CNV event is 0.21, reflecting that C. hircus has 4 copies in this region (chr1: 67890500-67913800). The dot plot was created from a DNA sequence extracted from the CNV (chr1: 67890500-67913800) region using NCBI BLAST website (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

(v) Biological roles of copy number variable genes

Copy number variable genes discovered in the present study play various biological processes, including complement activation, lipid, and energy metabolism, cell growth, inflammatory response, and reproduction. There were no significantly enriched gene ontology (GO) terms (available at figshare: https://doi.org/10.6084/m9.figshare.13633943).

Immune response genes mainly complement activation genes such as Complement Factor H Related 4 (*CFHR4*)), complement C3 (*C3*) and complement C4B (*C4B*) were duplicated in *C. nubiana.* Additionally, Cluster of Differentiation 54 (*CD54*), Cluster of Differentiation 48 (*CD48*) and Bactericidal/permeability-increasing fold containing family A, member 1 (*BPIFA1*) were shown to be expanded in *C. nubiana*. Notably, Natural Killer Group 2D receptor ligands (UL16 binding protein 3 (*ULBP3*), NKG2D ligand 1-like (*NKG2D LIGAND 1-L*) and Natural Killer Group 2D ligand 4-like (*NKG2D LIGAND 4L*)) were also shown to be in more copies in *C. nubiana* compared to the domestic goat.

Similarly, oxidizing enzymes such as cytochrome P450 family 2 subfamily B member 6 (*CYP2B6*), Cytochrome P450 2D6 (*CYP2D6*) and carboxylesterase 1 (*CES1*)) that play a crucial role in the biotransformation of xenobiotics compounds were shown to be in more copies in *C. nubiana*. Additionally, conjugative enzymes such as UDP-glucuronosyltransferase 2B31 (*UGT2B31*) and Glutathione S-transferase Mu 4 (*GSTM4*) that are involved in phase II of xenobiotic compounds metabolisms were also found to be expanded in *C. nubiana.* Furthermore, genes involved in xenobiotic compounds transport out of the cell, such as Multidrug resistance protein 4 (*MRP4 and MRP4L*), were reported to be expanded in *C. nubiana*.

4.2 Discussion

4.2.1 Genome sequence data

(i) Genome data and assembly

Capra nubiana is an endangered animal species; hence it is protected in its native countries. Therefore, it is essential to conserve its genetic resources. The C. *nubiana* genome sequence data and a draft genome assembly were generated since it was unavailable at the time of this study. Approximately 781 million high-quality paired-end sequence reads with coverage of 37x generated were *denovo* assembled using two de Bruijn graph-based assemblers;

Soapdenovo2 (Luo *et al.,* 2012) and ABySS (Jackman *et al.,* 2017). The two de Bruijn-based assemblers were used to determine the best assembler which can generate an optimal assembly for downstream analysis purposes. Genome assembly assessments showed that ABySS assembled genome was better than Soapdenovo2 assembled genome. Reference-assisted *de novo* assembly was carried out to improve the genome assembly contiguity; however, there was no improvement in the assembly. For example, the largest contig produced by the Alignngraph algorithm was 150 046 bp, and N50 was 10 597 bp, slightly lower than that for the ABySS assembler. Reference-assisted denovo assembly approaches are expected to improve the contiguity of a given fragmented assembly; however, it is unclear why it did not solve the fragmentation issue in *C. nubiana* genome. As such, the ABySS denovo assembled genome was considered the best for subsequent analysis.

(ii) Genome size estimation

GenomeScope (Vurture *et al.,* 2017) estimated *C. nubiana* genome size to be 2.6 Gbps. The *de novo* assembled genome size was 2.8 Gbps, nearly the same size as the estimated genome size. *C. nubiana* genome size is close to that of the *C. hircus* reference genome (2.9 Gbps) (Bickhart *et al.,* 2017) and the Bezoar (2.9 Gbps) (Dong *et al.,* 2015), and *Capra ibex* (2.7 Gbps) (Chen *et al.,* 2019).

(iii) Contiguity and completeness of *C. nubiana* denovo assembled genome

Contiguity assessment (number of contigs and N50 length) showed that *C. nubiana* assembly was fragmented when compared to that of the domestic goat (Bickhart *et al.,* 2017) and bezoar (Dong *et al.,* 2015). The contigs and scaffolds were many and short when compared to the ones' for the reference genome. This fragmentation is attributed to the use of short reads 125 bp and one insert size library (450 bp (Bao *et al.,* 2014; Schatz *et al.,* 2010). Additional sequence reads preferable mate-pair, and long-read sequences need to be generated to improve the assembly. Kmer analysis showed that most of the sequence contents were present in the assembly. Gene content analysis showed that 80.4% of the BUSCOs mapped to the draft genome, with 28.8% being fragmented, indicating a moderately complete genome. The presence of 28.8% of truncated genes is attributed to the fragmented nature of the assembly (Schatz *et al.,* 2010). The draft genome was suitable for gene predictions; however, it may not resolve entire gene sequences in the genome due to its fragmented nature (Parra *et al.,* 2009).

(iv) Gene predictions and annotation

A total of 25 674 protein-coding genes comprising matches to 80.4% of the single-copy gene orthologs in mammalian species were predicted in the *C. nubiana* genome, suggesting that the assembled genome contained a significant proportion of genes despite its fragmented nature. The number of predicted genes is slightly higher compared to what has been reported in other related species: domestic goat (21 361) (Bickhart *et al.*, 2017), Bezoar (23 217) (Dong *et al.*, 2015), and *C. ibex* (21 204) (Chen *et al.*, 2019). The differences in the number of genes identified may be due to the annotation approach used in this study; a combination of a single *de novo* approach and a single homology approach was used. The other studies used a combination of several *de novo* and homology approaches (Chen *et al.*, 2019; Dong *et al.*, 2015); therefore, the predicted genes will require further evaluation.

(v) Phylogenetic analysis

The phylogenetic tree placed *C. nubiana* and Alphine ibex (*C. ibex*) in the same clade (Fig. 8a), while *C. hircus* and *C. aegegrus* were in same clade. Since the genome sequences for few *Capra* species is available, it was not possible to make the comparison of the generated phylogeny tree with previous ones generated from mitochondrial and Y chromosome data. However, there are some similarities; for instance, *C. hircus* and *C. aegagrus* share the same clade. If *C. pyrenica* is collapsed in the previous phylogeny tree then *C. nubiana* and *C. ibex* will cluster together, as illustrated in Fig. 8b.

4.2.2 Positive selection signatures in *C. nubiana* genome

(i) Genetic variants and their annotations

Approximately 98% of *C. nubiana* sequence reads were aligned to unique sites in *C. hircus* (domestic goat), an indicator of reliable data for variant calling (Wu *et al.*, 2017). Approximately 1.7 million InDels and 19.5 million SNVs were discovered in the genome of *C. nubiana* generated in the present study. The SNVs in *C. nubiana* genome are close to those reported in other ruminants such as donkeys (18 million) (Bertolini *et al.*, 2018). The SNVs Ts/Tv ratio was 2.39, almost similar to human Ts/Tv ratio (>2.1), suggesting that potential sequencing errors are relatively low (Danecek *et al.*, 2011). Approximately 21 million and 22 million SNVs were detected from *C. nubiana* sequence data downloaded from a public database, a number close to detected in *C. nubiana* sequence data generated in this study. The SNVs and InDels present important genomic resources for further studies on *Capra* species genomic evolution.

(ii) Genes displaying strong selection signals in *C. nubiana*

Twenty-eight candidate genes were reported to be under adaptive evolution in the genome of the *C. nubiana*. However, 98% of the positively selected sites in 22 out of the 28 genes shown to be evolving were common to the three investigated *C. nubiana* genomes, implying that selection signatures reported are species-specific variations. The positively selected genes play vital roles, including visual development (Serine protease 56), blood pressure regulation (Rho GTPase activating protein 42) and reproduction (Storkhead box 2). Other evolving genes such as Olfactory receptor 1P1 are involved in signal transduction, while F-box protein 21 plays a crucial role in protein ubiquitination. On the other hand, nucleolus and neural progenitor protein is a crucial gene in Notch signalling pathway regulation. The biological functions of several other genes reported to be evolving in *C. nubiana* are unclear. Further, functional characterization studies would be necessary to explore the possible biological roles of each of the discovered positively selected genes. Notably, two genes (*ABCA12* and *ASCL4*), crucial in establishing a skin barrier and a DNA repair (*UVSSA*) were evolving in *C. nubiana* and may have an adaptive role.

Adaptive signatures of evolution in Capra nubiana

Capra nubiana is subjected to temperatures and solar radiation extremes in their environment, which has some implications such as excessive water loss through the skin or skin damages. For this reason, *C. nubiana* has an exceptional skin barrier characterized by a shiny waterproof coat (Castello, 2016). The skin barrier protects the inner body from environmental stressors such as extreme solar radiation and temperature as well as pathogens (Jensen & Proksch, 2009). Genes involved in the establishment of skin barriers such as *ABCA12* and *ASCL4* were found to be under adaptive evolution in *C. nubiana*. The ABCA12 gene's primary function is to transport lipids and ceramides to the top layer of the skin, which forms a skin-lipid barrier (Akiyama, 2014). Mutations in the *ABCA12* conserved domains in humans lead to a skin disorder known as ichthyosis (Kelsell *et al.*, 2005; Scott *et al.*, 2013). The amino acid change reported in *the ABCA12* gene in this study was functionally important, suggesting that it might have a crucial role in *C. nubiana* adaptations to the hot desert. Similarly, *the ASCL4* gene involved in epidermal development (Quan & Hassan, 2005), was reported to be evolving in *C. nubiana*. Studies have shown that the *ASCL4* gene is only expressed in skin and is involved in the development of hair follicles (Jonsson *et al.*, 2004; Rezza *et al.*, 2016). Positive selection

of *ABCA12* and *ASCL4* genes suggests that *C. nubiana* has acquired adaptive mechanisms to cope with the scorching sun and temperature extremes in its environment.

Furthermore, the findings showed that *C. nubiana* had evolved adaptive strategies to cope with UV-induced DNA damages. The *UVSSA*, a DNA repair gene found to be positively selected in *C. nubiana* is possibly involved in protecting it from harmful desert solar radiation. The *UVSSA* gene primary role is to remove damaged DNA from actively transcribed genes caused by UV (Sarasin, 2012); its mutations lead to UV-sensitive syndrome in humans (Nakazawa *et al.*, 2012).

4.2.3 Copy number variations

(i) Copy number variable regions in *C. nubiana* genome

The CNVs were called from *C. nubiana* genome sequence data using the depth of coverage approach (Abyzov *et al.*, 2011). Depth of coverage is a reliable sequence-based method for CNV calling, which was validated experimentally in several studies (Paudel *et al.*, 2015; Pezer *et al.*, 2015). For example, experimental validation of CNVs detected by CNVnator using droplet digital PCR showed a strong correlation between the CNVs detected by the two approaches (Pezer *et al.*, 2015). Therefore, the CNVs detected in this study were not confirmed experimentally; hence caution should be taken when interpreting the results. However, the simulation experiment showed that CNVnator was able to call 71% of the artificial CNVs, confirming that most CNVs reported here are potential true positives.

The CNV sites detected in three *C. nubiana* genomes were 1234, 1544, and 1726 respectively. In total, 367 CNVs were discovered across the three *C. nubiana* genomes. CNVs that were common between *C. nubiana* and *C. hircus* (Di Gerlando *et al.*, 2020; Guan *et al.*, 2020) were discarded because they indicate common genomic variations in the wild and domesticated goats genomes. The sequence data used in the present study were obtained from *C. nubiana* originating from different geographical regions (Pretoria in South Africa, Sinai in Egypt and Howtat in Saudi Arabia), suggesting that the CNVs reported here may reflect *C. nubiana*-specific CNVs.

(ii) The CNV-associated protein-coding genes

Copy number variable genes were discovered, providing a precious resource for future research into the relationship between CNVs phenotypes and adaptations in livestock species. Diverse biological processes were associated with copy number variable genes, but none were significantly enriched. Nonetheless, the CNV genes are involved in several biological processes such as energy metabolism, reproduction, cell growth lipid metabolism. Similar to other ruminants (Bickhart *et al.,* 2012; Zhang *et al.,* 2016), xenobiotic metabolism and immune- related genes were found to be copy number variable in *C. nubiana.*

Copy number variable genes associated with adaptations

Although several copy number variable genes were reported in this study, clusters of immune response and xenobiotic compounds metabolism genes reported in *C. nubiana* warranted further investigation due to their well-known functions. For instance, several immune response genes, including *BPIFA1, CD48, ULBP3, NKG2D ligand 1-like*, and *NKG2D ligand 4-like*, were duplicated in *C. nubiana*. *BPIFA1*is particularly expressed in the upper airways, and it offers crucial anti-bacterial and anti-viral roles (Akram *et al.,* 2018; Zhou *et al.,* 2008). The *CD48,* a signalling lymphocyte activation molecular family, was reported in more copies in *C. nubiana*. The *CD48* is involved in diverse immune response functions such as defense against viral and microbial infections (McArdel *et al.,* 2016). Although *CD48* serves a variety of immune response functions, it is a target of viral immune evasion (McArdel *et al.,* 2016). An increase in copy numbers of *CD48* might be an adaptation that allows species such *C. nubiana* to produce a diverse set of functional proteins to provide robust immunosurveillance. Ligands for NKG2D, an activation receptor including *NKG2D ligand 1-like, NKG2D ligand 4-like*, and *ULBP3* genes were duplicated in *C. nubiana*. NKG2D ligand regulates innate and adaptive immune responses (Sutherland *et al.,* 2006). Expressions of *NKG2D ligand 1-like, NKG2D ligand 4-like and ULBP3* are usually induced by viral stressors, tumorigenesis, or DNA damages (Lanier, 2015). The NKG2D ligands' expressions and binding to NKG2D receptors lead to immune response, enhanced immune surveillance, and antimicrobial immune response (Zingoni *et al.,* 2018). The *C. nubiana* is vulnerable to viral infections such as the Malignant catarrhal fever virus (Gasper *et al.,* 2012; Okeson *et al.,* 2007). Expansion of viral response genes in *C. nubiana,* suggests that it has acquired defense mechanisms to cope with viral stressors in its environment.

Other duplicated immune response genes identified in *C. nubiana* genome that play critical roles in the complement system included *C3, C4A*, and *C4B. The complement system* is key in the innate and acquired immune response against pathogens (Miyagawa *et al.,* 2008; Yang *et al.,* 2007). For instance, *C3* gene deficiency is linked to high susceptibility to systemic lupus erythematosus (SLE), while many copy numbers of *C4A* are known to alleviate susceptibility to SLE (Miyagawa *et al.,* 2008; Juptner *et al.,* 2018). Additionally, increased *C4A* copy numbers protect against macular degeneration associated with ageing (Grassmann *et al.,* 2016). Further studies into the specific roles of complement component genes in *C. nubiana* will be necessary to understand their immunosurveillance roles.

Oxidizing enzymes such as *CYP2D6, CYP2B6, CYP2D14, CYP2C31,* and *CES1* are key in bioconversion of lipid-soluble xenobiotic compounds to water-soluble form were shown to be expanded in *C. nubiana* (Marechal *et al.,* 2008). Similarly, conjugative enzymes such as *UGT2B7, UGT2B1,* and *GSTM4* key in the glucuronidation of biotransformed toxic compounds were also reported to be duplicated in *C. nubiana* (Iyanagi, 2007). Furthermore, xenobiotic efflux genes, including *MRP4-Like* and *MRP4*, that transport xenobiotics compounds that have undergone conjugation were found to be in more copies *C. nubiana* (Russel *et al.,* 2008) were reported to be expanded in *C. nubiana.* Desert plants such as cactus are affluent in toxic compounds such as oxalates and alkaloids (Robertson *et al.,* 2018). It is known that *C. nubiana* consumes a lot of alkaloid-producing plants (Habibi, 1997; Hakham & Ritte, 1993). Expansion of xenobiotic compounds metabolisms and transport genes in *C. nubiana*, suggests that it has evolved an excellent detoxification system to deal with harmful substances in their diet.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The *C. nubiana,* a wild African goat endemic to Sahara deserts, and Arabia thrive well in hot desert environments. The economic value of the domestic goat (*C. hircus* breeds) and current global warming trends highlight the need to know the genetic basis of well-adapted species such as *C. nubiana*. The adaptive signatures identified in the present study could then be used as selection makers for developing relevant goat breeding programs to enhance their adaptations to less favourable habitats in response to the climate emergency. The main objective of the present study was to investigate the adaptive signatures of evolution in *C. nubiana* genome. We achieved the goal using three specific activities.

Firstly, we generated *C. nubiana* genome sequence data and draft denovo assembly to serve as the reference for future studies and act as a starting point for evolutionary genomics studies to answer pertinent questions touching on its adaptations. The experiment yielded approximately 780 million clean paired-end sequence reads with coverage of 37x. The sequence reads were de novo assembled, resulting in a 2.6 Gbps genome with a scaffold N50 length of 13 812 bp. Genome annotation identified 25 674 protein-coding genes, including an estimated 80% of core mammalian genes (complete and partial). However, 28.8% of the detected core mammalian genes were fragmented; this is attributed to short reads (125 bp) and one library size (Schatz *et al.,* 2010) used in this study. The *C. nubiana* draft genome and predicted protein-coding genes serve as a starting point for further genome refinements. Being the first *C. nubiana* draft genome, the assembly is incomplete, and more work needs to be done to fill the gaps. There is a need to generate more sequence data, preferably long sequence reads, to develop a contiguous genome. Notably, the genome sequence data were of excellent quality (coverage>37x) to detect adaptive signatures of evolution in *C. nubiana*.

Secondly, a catalogue of single nucleotide variants and InDels in *C. nubiana* was generated using an alignment-based approach relative to the domestic goat genome. Approximately 19 million SNV and 1.7 million InDels were detected, which presents a rich resource for future *C. nubiana* genomic variation studies. Through comparative genomic analysis, twenty-two positively selected genes were identified in *C. nubiana*. Genes found to be evolving in C. nubiana play key biological functions such as visual development, immune response, blood

pressure regulation, and protein ubiquitination. Strong selection signals were reported in genes involved in skin barrier development (*ABCA12*), hair follicle developments (*ASCL4*) and DNA repair (*UVSSA*), implying that *C. nubiana* have acquired adaptive strategies to cope with the prevailing harsh conditions in the desert.

Thirdly, copy number variable genes were discovered in *C. nubiana* genome. Copy number variants (CNV) are differences in the number of copies of DNA segments between genomes and are the source of genetic variation in mammalian species (Redon *et al.,* 2006). Many copy number variable regions were reported; however, immune-related genes featured prominently with genes such as *BPIFA1, CD48, ULBP3, NKG2D LIGAND 1-L,* and *NKG2D LIGAND 4-L* being found in duplicated regions in *C. nubiana*. *C. nubiana* is vulnerable to viral infections such as the Malignant catarrhal fever virus (Gasper *et al.,* 2012; Okeson *et al.,* 2007). It is known that a healthy *C. nubiana* transmits the Ibex-MCF virus to antelopes, implying that it has a robust immune system hence the reason they get infected but not diseased (Gasper *et al.,* 2012; Okeson *et al.,* 2007). Expansion of viral response genes is possibly an adaptive trait that confers *C. nubiana* with a robust immunosurvillance system to cope with viral stressors in its environment. Xenobiotic compounds metabolism genes involved in various phases of toxic compound elimination were expanded in *C. nubiana*. For example, genes involved in biotransformation (*CES1, CYP2D14, CYP2B6, CYP2D6, CYP2C31*), conjugation (*UGT2B31, UGT2B7, GSTM4*) and transport (*MRP4* and *MRP4-L*) of xenobiotic compounds were reported to be expanded in *C. nubiana*. Xenobiotics metabolism enzymes and genes play crucial roles conversion of toxic secondary plant metabolites to less harmful compounds and subsequently transport them out of the cells (Marechal *et al.,* 2008). It is has been observed that *C. nubiana mainly* consumes alkaloid-rich plants (Habibi, 1997; Hakham & Ritte, 1993). These findings demonstrate that *C. nubiana* has an excellent detoxification system to handle xenobiotic compounds in its diet.

Genome sequence analysis provided insight into the adaptive sequences of evolution in *C. nubiana* adaptation in relation to its environment. The study showed that skin barrier development and function, xenobiotic compounds metabolisms, and viral response genes had undergone adaptive evolution in *C. nubiana*. The results suggest that *C. nubiana* have evolved adaptive attributes to cope with stressors in its environments, such as extreme temperatures, intense solar radiation, toxic diet, and viral infections. The adaptive signatures detected could be used as selection markers for designing goat breeding programs. *C. nubiana* could, for instance, be crossbred with local domestic goats. Furthermore, the adaptive signatures could

be used as selection markers to make specific changes to local goat genomes to improve their disease resistance or adapt to changing climates using genome editing technologies such as clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9). In conclusion, comparative genomics is a viable tool for studying the adaptation of species to their environment at the genome level; it shed light on the possible signatures of selection in *C. nubiana*.

5.2 Future recommendations

This study was pilot, and it provides a starting point for further genomics studies in *C. nubiana*. Sequence data from three *C. nubiana* was used to detect candidate adaptive genes; hence, further studies involving a large population would be necessary. The candidate genes detected were not confirmed experimentally; thus, it will be interesting to carry out gene expression analysis to verify if genes under strong selection signals involved in skin barrier development functions are expressed in the skin. Furthermore, candidate copy number variable genes involved viral response and xenobiotic compounds metabolisms need to be validated using PCR assays**.**

REFERENCES

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, *21*(6), 974–984.

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen- 2. *Current Protocols in Human Genetics*, *76*(1), 7–20.

Agaba, M., Ishengoma, E., Miller, W. C., McGrath, B. C., Hudson, C. N., Reina, O. C. B., Ratan, A., Burhans, R., Chikhi, R., & Medvedev, P. (2016). Giraffe genome sequence reveals clues to its unique morphology and physiology. *Nature Communications*, *7*(1), 1–8.

Akiyama, M. (2014). The roles of ABCA12 in epidermal lipid barrier formation and keratinocyte differentiation. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, *1841*(3), 435–440.

Akram, K. M., Moyo, N. A., Leeming, G. H., Bingle, L., Jasim, S., Hussain, S., Schorlemmer, A., Kipar, A., Digard, P., Tripp, R. A., Shohet, R. V., Bingle, C. D., & Stewart, J. P. (2018). An innate defense peptide BPIFA1/SPLUNC1 restricts influenza A virus infection. *Mucosal Immunology*, *11*(1), 71–81. https:// doi. org/10. 1038/mi. 2017.45

Alasaad, S., Fickel, J., Rossi, L., Sarasa, M., BenÃ-tez-Camacho, B., Granados, J. E., & Soriguer, R. C. (2012). Applicability of major histocompatibility complex DRB1 alleles as markers to detect vertebrate hybridization: A case study from Iberian ibex× domestic goat in southern Spain. *Acta Veterinaria Scandinavica*, *54*(1), 1–6.

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, *12*(5), 363–376.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andersson, L., & Georges, M. (2004). Domestic-animal genomics: Deciphering the genetics of complex traits. *Nature Reviews Genetics*, *5*(3), 202–212.

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Anisimova, M., Bielawski, J. P., & Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution*, *18*(8), 1585–1592.

Anisimova, M., Bielawski, J. P., & Yang, Z. (2002). Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution*, *19*(6), 950–958.

Anisimova, M., Nielsen, R., & Yang, Z. (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, *164*(3), 1229–1236.

Arlt, M. F., Wilson, T. E., & Glover, T. W. (2012). Replication stress and mechanisms of CNV formation. *Current Opinion in Genetics & Development*, *22*(3), 204–210.

Bactrian Camels Genome Sequencing and Analysis Consortium, Jirimutu, Wang, Z., Ding, G., Chen, G., Sun, Y., Sun, Z., Zhang, H., Wang, L., Hasi, S., Zhang, Y., Li, J., Shi, Y., Xu, Z., He, C., Yu, S., Li, S., Zhang, W., Batmunkh, M., … Meng, H. (2012). Genome sequences of wild and domestic bactrian camels. *Nature Communications*, *3*, 1202–1202. PubMed. https://doi.org/10.1038/ncomms2192

Baharav, D., & Meiboom, U. (1981). The status of the Nubian ibex Capra ibex nubiana in the Sinai Desert. *Biological Conservation*, *20*(2), 91–97.

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, *28*(1), 45–48. PubMed. https://doi.org/10.1093/nar/28.1.45

Ball, D. W., Azzoli, C. G., Baylin, S. B., Chi, D., Dou, S., Donis-Keller, H., Cumaraswamy, A., Borges, M., & Nelkin, B. D. (1993). Identification of a human achaete-scute homolog highly expressed in neuroendocrine tumors. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(12), 5648–5652. PubMed. https://doi.org/10.1073/pnas.90.12.5648

Bamshad, M., & Wooding, S. P. (2003). Signatures of natural selection in the human genome. *Nature Reviews Genetics*, *4*(2), 99–110.

Bao, E., Jiang, T., & Girke, T. (2014). AlignGraph: Algorithm for secondary de novo genome assembly guided by closely related references. *Bioinformatics*, *30*(12), i319–i328.

Benjelloun, B., Alberto, F. J., Streeter, I., Boyer, F., Coissac, E., Stucki, S., BenBati, M., Ibnelbachyr, M., Chentouf, M., & Bechchari, A. (2015). Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data. *Frontiers in Genetics*, *6*, 107.

Bentley, R. W., Pearson, J., Gearry, R. B., Barclay, M. L., McKinney, C., Merriman, T. R., & Roberts, R. L. (2010). Association of HigherDEFB4Genomic Copy Number With Crohn's Disease. *American Journal of Gastroenterology*, *105*(2), 354–359.

Berihulay, H., Abied, A., He, X., Jiang, L., & Ma, Y. (2019). Adaptation Mechanisms of Small Ruminants to Environmental Heat Stress. *Animals*, *9*(3). https:// doi. org/10. 3390/ ani9030075

Bertolini, F., Servin, B., Talenti, A., Rochat, E., Kim, E. S., Oget, C., Palhière, I., Crisà, A., Catillo, G., Steri, R., Amills, M., Colli, L., Marras, G., Milanesi, M., Nicolazzi, E., Rosen, B. D., Van Tassell, C. P., Guldbrandtsen, B., Sonstegard, T. S., … the AdaptMap consortium. (2018). Signatures of selection and environmental adaptation across the goat genome post-domestication. *Genetics Selection Evolution*, *50*(1), 57. https://doi.org/10.1186/s12711-018-0421-y

Bickhart, D. M., Hou, Y., Schroeder, S. G., Alkan, C., Cardone, M. F., Matukumalli, L. K., Song, J., Schnabel, R. D., Ventura, M., & Taylor, J. F. (2012). Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Research*, *22*(4), 778–790.

Bickhart, D. M., & Liu, G. E. (2014). The challenges and importance of structural variation detection in livestock. *Frontiers in Genetics*, *5*, 37.

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., Lee, J., Lam, E. T., Liachko, I., Sullivan, S. T., Burton, J. N., Huson, H. J., Nystrom, J. C., Kelley, C. M., Hutchison, J. L., Zhou, Y., Sun, J., Crisà, A., Ponce de León, F. A., … Smith, T. P.

L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, *49*(4), 643–650. https://doi.org/10.1038/ng.3802

Bielawski, J. P., & Yang, Z. (2003). Maximum likelihood methods for detecting adaptive evolution after gene duplication. *Genome Evolution*, 3, 201–212.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Boschiero, C., Gheyas, A. A., Ralph, H. K., Eory, L., Paton, B., Kuo, R., Fulton, J., Preisinger, R., Kaiser, P., & Burt, D. W. (2015). Detection and characterization of small insertion and deletion genetic variants in modern layer chicken genomes. *BMC Genomics*, *16*(1), 562-567. https://doi.org/10.1186/s12864-015-1711-1

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

Buchanan, J. A., & Scherer, S. W. (2008). Contemplating effects of genomic structural variation. *Genetics in Medicine, 10*(9), 639–647.

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, *268*(1), 78–94.

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*, *39*(7), S16–S21.

Castello, J. R. (2016). *Bovids of the world: Antelopes, gazelles, cattle, goats, sheep, and relatives* (Vol. 104). Princeton University Press.

Chavez, D. E., Gronau, I., Hains, T., Kliver, S., Koepfli, K. P., & Wayne, R. K. (2019). Comparative genomics provides new insights into the remarkable adaptations of the African wild dog (Lycaon pictus). *Scientific Reports*, *9*(1), 1–14.

Chebii, V. J., Oyola, S. O., Kotze, A., Domelevo Entfellner, J.-B., Musembi Mutuku, J., & Agaba, M. (2020). Genome-Wide Analysis of Nubian Ibex Reveals Candidate Positively Selected Genes That Contribute to Its Adaptation to the Desert Environment. *Animals*, *10*(11), 1-17. https://doi.org/10.3390/ani10112181

Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., Bibi, F., Yang, Y., Wang, J., & Nie, W. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science*, *364*(6446), 1-18

Chikhi, R., & Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, *30*(1), 31–37. https:// doi. org/10. 1093/bioinformatics/btt310

Chiruvella, K. K., Liang, Z., & Wilson, T. E. (2013). Repair of double-strand breaks by end joining. *Cold Spring Harbor Perspectives in Biology*, *5*(5), a012757–a012757. PubMed. https://doi.org/10.1101/cshperspect.a012757

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, *44*(D1), D67–D72. PubMed. https:// doi. org/10. 1093/ nar/ gkv1276

Cook, D. E., Lee, T. G., Guo, X., Melito, S., Wang, K., Bayless, A. M., Wang, J., Hughes, T. J., Willis, D. K., & Clemente, T. E. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science*, *338*(6111), 1206–1209.

Criscuolo, A., & Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*, *10*(1), 1–21.

Dakal, T. C., Kala, D., Dhiman, G., Yadav, V., Krokhotin, A., & Dokholyan, N. V. (2017). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms in IL8 gene. *Scientific Reports*, *7*(1), 1–18.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., & Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

De Smith, A., Walters, R., Froguel, P., & Blakemore, A. (2008). Human genes involved in copy number variation: Mechanisms of origin, functional effects and implications for disease. *Cytogenetic and Genome Research*, *123*(1–4), 17–26.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology*, *4*(9), 1–11.

Di Gerlando, R., Mastrangelo, S., Moscarelli, A., Tolone, M., Sutera, A. M., Portolano, B., & Sardina, M. T. (2020). Genomic Structural Diversity in Local Goats: Analysis of Copy-Number Variations. *Animals : An Open Access Journal from MDPI*, *10*(6). https://doi.org/10.3390/ani10061040

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*(1), 1–11.

Dorshorst, B., Harun-Or-Rashid, M., Bagherpoor, A. J., Rubin, C. J., Ashwell, C., Gourichon, D., Tixier-Boichard, M., Hallböök, F., & Andersson, L. (2015). A genomic duplication is associated with ectopic eomesodermin expression in the embryonic chicken comb and two duplex-comb phenotypes. *PLoS Genet*, *11*(3), e1004947.

Du, X., Servin, B., Womack, J. E., Cao, J., Yu, M., Dong, Y., Wang, W., & Zhao, S. (2014). An update of the goat genome assembly using dense radiation hybrid maps allows detailed analysis of evolutionary rearrangements in Bovidae. *BMC Genomics*, *15*(1), 1–16.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 1–19.

Ekblom, R., & Wolf, J. B. (2014). A field guide to whole- genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042.

Fontanesi, L., Beretti, F., Martelli, P., Colombo, M., Dall'Olio, S., Occidente, M., Portolano, B., Casadio, R., Matassino, D., & Russo, V. (2011). A first comparative map of copy number variations in the sheep genome. *Genomics*, *97*(3), 158–165.

Fontanesi, L., Martelli, P. L., Beretti, F., Riggio, V., Dall'Olio, S., Colombo, M., Casadio, R., Russo, V., & Portolano, B. (2010). An initial comparative map of copy number variations in the goat (Capra hircus) genome. *BMC Genomics*, *11*(1), 639. https://doi.org/10.1186/1471-2164-11-639

Gao, Y., Jiang, J., Yang, S., Hou, Y., Liu, G. E., Zhang, S., Zhang, Q., & Sun, D. (2017). CNV discovery for milk composition traits in dairy cattle using whole genome resequencing. *Bmc Genomics*, *18*(1), 1–12.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., & Conesa, A. (2012). Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics*, *28*(20), 2678–2679. https:// doi. org/10. 1093/bioinformatics/bts503

Gasper, D., Barr, B., Li, H., Taus, N., Peterson, R., Benjamin, G., Hunt, T., & Pesavento, P. A. (2012). Ibex-associated malignant catarrhal fever–like disease in a group of bongo antelope (Tragelaphus eurycerus). *Veterinary Pathology*, *49*(3), 492–497.

Gebreyohanes, M., & Assen, A. (2017). Adaptation mechanisms of camels (Camelus dromedarius) for desert environment: A review. *Journal of Veterinary Science Technology*, *8*(6), 1–5.

Gharib, W. H., & Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Molecular Biology and Evolution*, *30*(7), 1675–1686.

Giacometti, M., Roganti, R., De Tann, D., Stahlberger-Saitbekova, N., & Obexer-Ruff, G. (2004). Alpine ibex Capra ibex ibex x domestic goat C. aegagrus domestica hybrids in a restricted area of southern Switzerland. *Wildlife Biology*, *10*(1), 137–143.

Giuffra, E., Evans, G., Törnsten, A., Wales, R., Day, A., Looft, H., Plastow, G., & Andersson, L. (1999). The Belt mutation in pigs is an allele at the Dominant white (I/KIT) locus. *Mammalian Genome*, *10*(12), 1132–1136.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333.

Grassmann, F., Cantsilieris, S., Schulz-Kuhnt, A.-S., White, S. J., Richardson, A. J., Hewitt, A. W., Vote, B. J., Schmied, D., Guymer, R. H., Weber, B. H. F., & Baird, P. N. (2016). Multiallelic copy number variation in the complement component 4A (C4A) gene is associated with late-stage age-related macular degeneration (AMD). *Journal of Neuroinflammation*, *13*(1), 81–81. PubMed. https://doi.org/10.1186/s12974-016-0548-0

Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., Rogel-Gaillard, C., Park, C., Milan, D., & Megens, H.-J. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, *491*(7424), 393–398.

Gross, J. E., Alkon, P., & Demment, M. (1995). Grouping patterns and spatial segregation by Nubian ibex. *Journal of Arid Environments*, *30*(4), 423–439.

Grossen, C., Guillaume, F., Keller, L. F., & Croll, D. (2020). Purging of highly deleterious mutations through severe bottlenecks in Alpine ibex. *Nature Communications*, *11*(1), 1–12.

Guan, D., Martínez, A., Castello, A., Landi, V., Luigi-Sierra, M. G., Fernández-Álvarez, J., Cabrera, B., Delgado, J. V., Such, X., Jordana, J., & Amills, M. (2020). A genome-wide analysis of copy number variation in Murciano-Granadina goats. *Genetics Selection Evolution*, *52*(1), 44. https://doi.org/10.1186/s12711-020-00564-4

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Guo, F., Si, R., He, J., Yuan, L., Hai, L., Ming, L., Yi, L., & Ji, R. (2019). Comprehensive transcriptome analysis of adipose tissue in the Bactrian camel reveals fore hump has more specific physiological functions in immune and endocrine systems. *Livestock Science*, *228*, 195–200.

Guo, J., Tao, H., Li, P., Li, L., Zhong, T., Wang, L., Ma, J., Chen, X., Song, T., & Zhang, H. (2018). Whole-genome sequencing reveals selection signatures associated with important traits in six goat breeds. *Scientific Reports*, *8*(1), 10405. PubMed. https://doi.org/10.1038/s41598-018-28719-w

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075.

Gurgul, A., Jasielczuk, I., Semik-Gurgul, E., Szmatoła, T., Majewska, A., Sosin-Bzducha, E., & Bugno-Poniewierska, M. (2019). Diversifying selection signatures among divergently selected subpopulations of Polish Red cattle. *Journal of Applied Genetics*, *60*(1), 87–95.

Habibi, K. (1997). Group dynamics of the Nubian ibex (Capra ibex nubiana) in the Tuwayiq Canyons, Saudi Arabia. *Journal of Zoology*, *241*(4), 791–801.

Hakham, E., & Ritte, U. (1993). Foraging pressure of the Nubian ibex Capra ibex nubiana and its effect on the indigenous vegetation of the En Gedi Nature Reserve, Israel. *Biological Conservation*, *63*(1), 9–21.

Hammer, S. E., Schwammer, H. M., & Suchentrunk, F. (2008). Evidence for introgressive hybridization of captive markhor (Capra falconeri) with domestic goat: Cautions for reintroduction. *Biochemical Genetics*, *46*(3–4), 216–226.

Haraksingh, R. R., & Snyder, M. P. (2013). Impacts of variation in the human genome on gene regulation. *Journal of Molecular Biology*, *425*(21), 3970–3977.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, *10*(8), 551–564.

Henry, B. K., Eckard, R. J., & Beauchemin, K. A. (2018). Review: Adaptation of ruminant livestock production systems to climate changes. *Animal*, *12*(s2), s445–s456. Cambridge Core. https://doi.org/10/gk6jpx

Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-genome annotation with BRAKER. In *Gene prediction* (pp. 65–95). Springer.

Huang, Y., Chen, S.-Y., & Deng, F. (2016). Well-characterized sequence features of eukaryote genomes and implications for ab initio gene prediction. *Computational and Structural Biotechnology Journal*, *14*, 298–303.

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, *33*(6), 1635–1638.

Iyanagi, T. (2007). Molecular mechanism of phase I and phase II drug- metabolizing enzymes: Implications for detoxification. *International Review of Cytology*, *260*, 35–112.

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., Jahesh, G., Khan, H., Coombe, L., & Warren, R. L. (2017). ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, *27*(5), 768–777.

Jenkins, G. M., Goddard, M. E., Black, M. A., Brauning, R., Auvray, B., Dodds, K. G., Kijas, J. W., Cockett, N., & McEwan, J. C. (2016). Copy number variants in the sheep genome detected using multiple approaches. *BMC Genomics*, *17*(1), 441.

Jensen, J., & Proksch, E. (2009). The skin's barrier. *Giornale Italiano Di Dermatologia e Venereologia: Organo Ufficiale, Societa Italiana Di Dermatologia e Sifilografia*, *144*(6), 689–700.

Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J. F., Faraut, T., Wu, C., Muzny, D. M., Li, Y., & Zhang, W. (2014). The sheep genome illuminates biology of the rumen and lipid metabolism. *Science*, *344*(6188), 1168–1173.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics (Oxford, England)*, *30*(9), 1236–1240. PubMed. https://doi.org/10.1093/bioinformatics/btu031

Jonsson, M., Björntorp Mark, E., Brantsing, C., Brandner, J. M., Lindahl, A., & Asp, J. (2004). Hash4, a novel human achaete-scute homologue found in fetal skin. *Genomics*, *84*(5), 859–866. https://doi.org/10.1016/j.ygeno.2004.07.004

Juptner, M., Flachsbart, F., Caliebe, A., Lieb, W., Schreiber, S., Zeuner, R., Franke, A., & Schröder, J. O. (2018). Low copy numbers of complement C4 and homozygous deficiency of C4A may predispose to severe disease and earlier disease onset in patients with systemic lupus erythematosus. *Lupus*, *27*(4), 600–609. PubMed. https://doi.org/10.1177/0961203317735187

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361. PubMed. https://doi.org/10.1093/nar/gkw1092

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T., Lua, R. C., Wilkins, A. D., & Lichtarge, O. (2014). Single nucleotide variations: Biological impact and theoretical interpretation. *Protein Science*, *23*(12), 1650–1666.

Kchouk, M., Gibrat, J. F., & Elloumi, M. (2017). Generations of sequencing technologies: From first to next generation. *Biology and Medicine*, *9*(3).

Kelsell, D. P., Norgett, E. E., Unsworth, H., Teh, M. T., Cullup, T., Mein, C. A., Dopping-Hepenstal, P. J., Dale, B. A., Tadini, G., Fleckman, P., Stephens, K. G., Sybert, V. P., Mallory, S. B., North, B. V., Witt, D. R., Sprecher, E., Taylor, A. E. M., Ilchyshyn, A., Kennedy, C. T., … O'Toole, E. A. (2005). Mutations in ABCA12 underlie the severe congenital skin disease harlequin ichthyosis. *American Journal of Human Genetics*, *76*(5), 794–803. PubMed. https://doi.org/10.1086/429844

Kijas, J., Serrano, M., McCulloch, R., Li, Y., Salces Ortiz, J., Calvo, J., Pérez- Guzmán, M., & International Sheep Genomics Consortium. (2013). Genomewide association for a dominant pigmentation gene in sheep. *Journal of Animal Breeding and Genetics*, *130*(6), 468–475.

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., & Kerhornou, A. (2011). *Ensembl BioMarts: A hub for data retrieval across taxonomic space. Database, 2011*. www. google.com

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., & Du, L. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science*, *318*(5849), 420–426.

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547.

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, *34*(7), 1812–1819.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357.

Lanier, L. L. (2015). NKG2D receptor and its ligands in host defense. *Cancer Immunology Research*, *3*(6), 575–582.

Lee, W., Ahn, S., Taye, M., Sung, S., Lee, H. J., Cho, S., & Kim, H. (2016). Detecting positive selection of korean native goat populations using next-generation sequencing. *Molecules and Cells*, *39*(12), 862.

Li, H. (2012). Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*, *28*(14), 1838–1844.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858. PubMed. https://doi.org/10.1101/gr.078212.108

Li, W., Bickhart, D. M., Ramunno, L., Iamartino, D., Williams, J. L., & Liu, G. E. (2019). Comparative sequence alignment reveals River Buffalo genomic structural differences compared with cattle. *Genomics*, *111*(3), 418–425. https:// doi. org/ 10. 1016/j. ygeno.2018.02.018

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., & Liu, B. (2012). Comparison of the two major classes of assembly algorithms: Overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, *11*(1), 25–37.

Lightner, J. K. (2006). Identification of species within the sheep-goat kind (Tsoan monobaramin). *Journal of Creation*, *20*(3), 61–65.

Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., & Guo, J. (2017). Effects of short indels on protein structure and function in human genomes. *Scientific Reports*, *7*(1), 9313. https://doi.org/10.1038/s41598-017-09287-x

Lin, Z., Chen, L., Chen, X., Zhong, Y., Yang, Y., Xia, W., Liu, C., Zhu, W., Wang, H., & Yan, B. (2019). Biological adaptations in the Arctic cervid, the reindeer (Rangifer tarandus). *Science*, *364*(6446).

Lischer, H. E., & Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, *18*(1), 1–12.

Liu, M., Zhou, Y., Rosen, B. D., Van Tassell, C. P., Stella, A., Tosser-Klopp, G., Rupp, R., Palhière, I., Colli, L., Sayre, B., Crepaldi, P., Fang, L., Mészáros, G., Chen, H., Liu, G. E., & the ADAPTmap Consortium. (2019). Diversity of copy number variation in the worldwide goat population. *Heredity*, *122*(5), 636–646. https://doi.org/10.1038/s41437-018-0150-6

Loytynoja, A. (2014). Phylogeny-aware alignment with PRANK. In *Multiple sequence alignment methods* (pp. 155–170). Springer.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., & Liu, Y. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience*, *1*(1), 2047-217X.

Magi, A., Tattini, L., Pippucci, T., Torricelli, F., & Benelli, M. (2012). Read count approach for DNA copy number variants detection. *Bioinformatics*, *28*(4), 470–478. https://doi.org/10.1093/bioinformatics/btr707

Manceau, V., Després, L., Bouvet, J., & Taberlet, P. (1999). Systematics of the genus Capra inferred from mitochondrial DNA sequence data. *Molecular Phylogenetics and Evolution*, *13*(3), 504–510.

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., & Clavijo, B. J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, *33*(4), 574–576.

Marechal, J. D., Kemp, C. A., Roberts, G. C. K., Paine, M. J. I., Wolf, C. R., & Sutcliffe, M. J. (2008). Insights into drug metabolism by cytochromes P450 from modelling studies of CYP2D6-drug interactions. *British Journal of Pharmacology*, *153 Suppl 1*(Suppl 1), S82–S89. PubMed. https://doi.org/10.1038/sj.bjp.0707570

McArdel, S. L., Terhorst, C., & Sharpe, A. H. (2016). Roles of CD48 in regulating immunity and tolerance. *Clinical Immunology*, *164*, 10–20. https:// doi. org/10. 1016/j. clim. 2016. 01. 008

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., & Daly, M. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Medeiros de Oliveira Silva, R., Bonvino Stafuzza, N., De O. F. B., Miguel, F. C. G., Matos, C. T., Noely, S. G. C. J., Baldi, F., Augusti, B. A., Zerlotti, M. M. E., & Lino, L. D. (2017). Genome-wide association study for carcass traits in an experimental Nelore cattle population. *PLoS One*, *12*(1), e0169860.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315–327.

Miyagawa, H., Yamai, M., Sakaguchi, D., Kiyohara, C., Tsukamoto, H., Kimoto, Y., Nakamura, T., Lee, J.-H., Tsai, C. Y., Chiang, B.-L., Shimoda, T., Harada, M., Tahira, T., Hayashi, K., & Horiuchi, T. (2008). Association of polymorphisms in complement component C3 gene with susceptibility to systemic lupus erythematosus. *Rheumatology*, *47*(2), 158–164. https://doi.org/10.1093/rheumatology/kem321

Miyata, N., & Roman, R. J. (2005). Role of 20-hydroxyeicosatetraenoic acid (20-HETE) in vascular system. *Journal of Smooth Muscle Research*, *41*(4), 175–193.

Montgomery, S. B., Goode, D. L., Kvikstad, E., Albers, C. A., Zhang, Z. D., Mu, X. J., Ananda, G., Howie, B., Karczewski, K. J., & Smith, K. S. (2013). The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Research*, *23*(5), 749–761.

Nakazawa, Y., Sasaki, K., Mitsutake, N., Matsuse, M., Shimada, M., Nardo, T., Takahashi, Y., Ohyama, K., Ito, K., Mishima, H., Nomura, M., Kinoshita, A., Ono, S., Takenaka, K.,

Masuyama, R., Kudo, T., Slor, H., Utani, A., Tateishi, S., … Ogi, T. (2012). Mutations in UVSSA cause UV-sensitive syndrome and impair RNA polymerase IIo processing in transcription-coupled nucleotide-excision repair. *Nature Genetics*, *44*(5), 586–592. https://doi.org/10.1038/ng.2229

Naval-Sanchez, M., Nguyen, Q., McWilliam, S., Porto-Neto, L. R., Tellam, R., Vuocolo, T., Reverter, A., Perez-Enciso, M., Brauning, R., & Clarke, S. (2018). Sheep genome functional annotation reveals proximal regulatory elements contributed to the evolution of modern breeds. *Nature Communications*, *9*(1), 1–13.

NCBI Resource Coordinators. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *44*(D1), D7–D19. PubMed. https://doi.org/10.1093/nar/gkv1290

Ng, P. C., & Henikoff, S. (2006). Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annual Review of Genomics and Human Genetics*, *7*(1), 61–80. https://doi.org/10.1146/annurev.genom.7.080505.115630

Nicoloso, L., Bomba, L., Colli, L., Negrini, R., Milanesi, M., Mazza, R., Sechi, T., Frattini, S., Talenti, A., & Coizet, B. (2015). Genetic diversity of Italian goat breeds assessed with a medium-density SNP chip. *Genetics Selection Evolution*, *47*(1), 1–10.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annual Review Genetinetics*, *39*, 197–218.

Nielsen, R., & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, *148*(3), 929–936.

Okeson, D. M., Garner, M. M., Taus, N. S., Li, H., & Coke, R. L. (2007). Ibex-associated malignant catarrhal fever in a bongo antelope (Tragelaphus euryceros). *Journal of Zoo and Wildlife Medicine*, *38*(3), 460–464.

Olson, N. D., Lund, S. P., Colman, R. E., Foster, J. T., Sahl, J. W., Schupp, J. M., Keim, P., Morrow, J. B., Salit, M. L., & Zook, J. M. (2015). Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in Genetics*, *6*, 235-243.

Onzima, R. B., Upadhyay, M. R., Doekes, H. P., Brito, L., Bosse, M., Kanis, E., Groenen, M. A., & Crooijmans, R. P. (2018). Genome-wide characterization of selection signatures and runs of homozygosity in Ugandan goat breeds. *Frontiers in Genetics*, *9*, 318-324.

Oyola, S. O., Otto, T. D., Gu, Y., Maslen, G., Manske, M., Campino, S., Turner, D. J., Macinnis, B., Kwiatkowski, D. P., Swerdlow, H. P., & Quail, M. A. (2012). Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics*, *13*, 1–1. PubMed. https://doi.org/10.1186/1471-2164-13-1

Parra, G., Bradnam, K., Ning, Z., Keane, T., & Korf, I. (2009). Assessing the gene space in draft genomes. *Nucleic Acids Research*, *37*(1), 289–297.

Parrini, F., Cain, J. W., & Krausman, P. R. (2009). Capra ibex (Artiodactyla: Bovidae). *Mammalian Species*, *830*, 1–12.

Paszkiewicz, K., & Studholme, D. J. (2010). De novo assembly of short sequence reads. *Briefings in Bioinformatics*, *11*(5), 457–472.

Paudel, Y., Madsen, O., Megens, H. J., Frantz, L. A., Bosse, M., Bastiaansen, J. W., Crooijmans, R. P., & Groenen, M. A. (2013). Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics*, *14*(1), 449-460. https://doi.org/10.1186/1471-2164-14-449

Paudel, Y., Madsen, O., Megens, H. J., Frantz, L. A., Bosse, M., Crooijmans, R. P., & Groenen, M. A. (2015). Copy number variation in the speciation of pigs: A possible prominent role for olfactory receptors. *BMC Genomics*, *16*(1), 330-340.

Pezer, Z., Harr, B., Teschke, M., Babiker, H., & Tautz, D. (2015). Divergence patterns of genic copy number variation in natural populations of the house mouse (Mus musculus domesticus) reveal three conserved genes with major population-specific expansions. *Genome Research*, *25*(8), 1114–1124.

Pidancier, N., Jordan, S., Luikart, G., & Taberlet, P. (2006). Evolutionary history of the genus Capra (Mammalia, Artiodactyla): Discordance between mitochondrial DNA and Y-chromosome phylogenies. *Molecular Phylogenetics and Evolution*, *40*(3), 739–749.

Pirooznia, M., Goes, F. S., & Zandi, P. P. (2015). Whole-genome CNV analysis: Advances in computational approaches. *Frontiers in Genetics*, *6*, 138-145.

Qiu, Q., Zhang, G., Ma, T., Qian, W., Wang, J., Ye, Z., Cao, C., Hu, Q., Kim, J., & Larkin, D. M. (2012). The yak genome and adaptation to life at high altitude. *Nature Genetics*, *44*(8), 946–949.

Quan, X. J., & Hassan, B. A. (2005). From skin to nerve: Flies, vertebrates and the first helix. *Cellular and Molecular Life Sciences CMLS*, *62*(18), 2036–2049. https:// doi. org/10. 1007/ s00018-005-5124-1

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.

Rahmatalla, S. A., Arends, D., Reissmann, M., Ahmed, A. S., Wimmers, K., Reyer, H., & Brockmann, G. A. (2017). Whole genome population genetics analysis of Sudanese goats identifies regions harboring genes associated with major traits. *BMC Genetics*, *18*(1), 1–10.

Rausell, A., Mohammadi, P., McLaren, P. J., Bartha, I., Xenarios, I., Fellay, J., & Telenti, A. (2014). Analysis of stop-gain and frameshift variants in human innate immunity genes. *PLoS Computation Biology*, *10*(7), e1003757.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., & Chen, W. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454.

Reis, M. dos, & Yang, Z. (2011). Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Molecular Biology and Evolution*, *28*(7), 2161–2172.

Rezza, A., Wang, Z., Sennett, R., Qiao, W., Wang, D., Heitman, N., Mok, K. W., Clavel, C., Yi, R., Zandstra, P., Ma'ayan, A., & Rendl, M. (2016). Signaling Networks among Stem Cell Precursors, Transit-Amplifying Progenitors, and their Niche in Developing Hair Follicles. *Cell Reports*, *14*(12), 3001–3018. PubMed. https:// doi. org/10. 1016/j. celrep. 2016. 02.078

Robertson, L. P., Hall, C. R., Forster, P. I., & Carroll, A. R. (2018). Alkaloid diversity in the leaves of Australian Flindersia (Rutaceae) species driven by adaptation to aridity. *Phytochemistry*, *152*, 71–81. https://doi.org/10.1016/j.phytochem.2018.04.011

Rosenberg, M. S., Subramanian, S., & Kumar, S. (2003). Patterns of transitional mutation biases within and among mammalian genomes. *Molecular Biology and Evolution*, *20*(6), 988–993.

Ross, S., Elalqamy, H., Al Said, T., & Saltz, D. (2020). *Capra nubiana. The IUCN Red List of Threatened Species 2020: E. T3796A22143385*.

Russel, F. G. M., Koenderink, J. B., & Masereeuw, R. (2008). Multidrug resistance protein 4 (MRP4/ABCC4): A versatile efflux transporter for drugs and signalling molecules. *Trends in Pharmacological Sciences*, *29*(4), 200–207. https:// doi. org/10. 1016/j. tips.2008.01.006

Sarasin, A. (2012). UVSSA and USP7: New players regulating transcription-coupled nucleotide excision repair in human cells. *Genome Medicine*, *4*(5), 44–44. PubMed. https://doi.org/10.1186/gm343

Schatz, M. C., Delcher, A. L., & Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Research*, *20*(9), 1165–1173.

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H., & Graur, D. (2009). Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biology and Evolution*, *1*, 114–118.

Schrider, D. R., & Hahn, M. W. (2010). Gene copy-number polymorphism in nature. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1698), 3213–3221.

Scott, C. A., Rajpopat, S., & Di, W. L. (2013). Harlequin ichthyosis: ABCA12 mutations underlie defective lipid transport, reduced protease regulation and skin-barrier dysfunction. *Cell and Tissue Research*, *351*(2), 281–288. https://doi. org/10.1 007/ s00441-012-1474-9

Sehn, J. K. (2015). Chapter 9—Insertions and Deletions (Indels). In S. Kulkarni & J. Pfeifer (Eds.), *Clinical Genomics* (pp. 129–150). Academic Press. https:// doi. rg/10. 1016/ B978-0-12-404748-8.00009-5

Shackleton, D. M. (1997). *Wild sheep and goats and their relatives-status survey and conservation action plan for caprinae* (Issue 333.959 W668w). IUCN, Gland (Suiza). Species Survival Commission.

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

Smit, A., Hubley, R., & Green, P. (1996). *RepeatMasker Open-3.0*. www.google.com

Spencer, D. H., Zhang, B., & Pfeifer, J. (2015). *Single nucleotide variant detection using next generation sequencing. In Clinical Genomics (pp. 109–127)*. Elsevier. www.google.com

Stafuzza, N. B., Zerlotini, A., Lobo, F. P., Yamagishi, M. E. B., Chud, T. C. S., Caetano, A. R., Munari, D. P., Garrick, D. J., Machado, M. A., & Martins, M. F. (2017). Single nucleotide variants and InDels identified from whole-genome re-sequencing of Guzerat, Gyr, Girolando and Holstein cattle breeds. *PLoS One*, *12*(3), e0173954.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(suppl_2), W435–W439.

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, *61*, 437–455.

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

Stuwe, M., & Grodinsky, C. (1987). Reproductive biology of captive Alpine ibex (Capra i. Ibex). *Zoo Biology*, *6*(4), 331–339.

Sutherland, C. L., Rabinovich, B., Chalupny, N. J., Brawand, P., Miller, R., & Cosman, D. (2006). ULBPs, human ligands of the NKG2D receptor, stimulate tumor immunity with enhancement by IL-15. *Blood*, *108*(4), 1313–1319.

Tadesse, S. A., & Kotler, B. P. (2012). Impact of tourism on Nubian Ibex (Capra nubiana) revealed through assessment of behavioral indicators. *Behavioral Ecology*, *23*(6), 1257–1262.

Tamura, K., Battistuzzi, F. U., Billing-Ross, P., Murillo, O., Filipski, A., & Kumar, S. (2012). Estimating divergence times in large molecular phylogenies. *Proceedings of the National Academy of Sciences*, *109*(47), 19333–19338.

Tang, H., & Thomas, P. D. (2016). Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics*, *203*(2), 635–647.

Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in Bioengineering and Biotechnology*, *3*, 92.

Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S., & Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, *28*(21), 2711–2718.

Tollis, M., DeNardo, D. F., Cornelius, J. A., Dolby, G. A., Edwards, T., Henen, B. T., Karl, A. E., Murphy, R. W., & Kusumi, K. (2017). The Agassiz's desert tortoise genome provides a resource for the conservation of a threatened species. *PLoS One*, *12*(5), e0177708.

Tosser-Klopp, G., Bardou, P., Bouchez, O., Cabau, C., Crooijmans, R., Dong, Y., Donnadieu-Tonon, C., Eggen, A., Heuven, H. C., & Jamli, S. (2014). Design and characterization of a 52K SNP chip for goats. *PloS One*, *9*(1), e86227.

Tull, J. C., Krausman, P. R., & Steidl, R. J. (2001). Bed-site selection by desert mule deer in southern Arizona. *The Southwestern Naturalist*, 354–357.

Utsunomiya, Y. T., Do Carmo, A. S., Carvalheiro, R., Neves, H. H., Matos, M. C., Zavarez, L. B., O'Brien, A. M. P., Sölkner, J., McEwan, J. C., & Cole, J. B. (2013). Genome-wide association study for birth weight in Nellore cattle points to previously described orthologous genes affecting human and bovine height. *BMC Genetics*, *14*(1), 1–12.

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, *34*(3), 275–305.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204.

Wall, D., Fraser, H., & Hirsh, A. (2003). Detecting putative orthologs. *Bioinformatics*, *19*(13), 1710–1711.

Wang, C. Y., Shahi, P., Huang, J. T. W., Phan, N. N., Sun, Z., Lin, Y. C., Lai, M. D., & Werb, Z. (2017). Systematic analysis of the achaete-scute complex-like gene signature in clinical cancer patients. *Molecular and Clinical Oncology*, *6*(1), 7–18. PubMed. https://doi.org/10.3892/mco.2016.1094

Wang, G.-D., Shao, X. J., Bai, B., Wang, J., Wang, X., Cao, X., Liu, Y. H., Wang, X., Yin, T.-T., Zhang, S. J., Lu, Y., Wang, Z., Wang, L., Zhao, W., Zhang, B., Ruan, J., & Zhang, Y. P. (2019). Structural variation during dog domestication: Insights from gray wolf and dhole genomes. *National Science Review*, *6*(1), 110–122. https:// doi. org/10. 1093/nsr/nwy076

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164.

Wang, X., Liu, J., Zhou, G., Guo, J., Yan, H., Niu, Y., Li, Y., Yuan, C., Geng, R., & Lan, X. (2016). Whole-genome sequencing of eight goat populations for the detection of selection signatures underlying production and adaptive traits. *Scientific Reports*, *6*(1), 1–10.

Wang, Z., Chen, Y., & Li, Y. (2004). A Brief Review of Computational Gene Prediction Methods. *Genomics, Proteomics & Bioinformatics*, *2*(4), 216–221. https:// doi. org/ 10.1016/ S1672-0229(04)02028-5

Ward, N., & Moreno-Hagelsieb, G. (2014). Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? *PloS One*, *9*(7), e101850.

Waterhouse, R. M., Zdobnov, E. M., Tegenfeldt, F., Li, J., & Kriventseva, E. V. (2011). OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Research*, *39*(suppl_1), D283–D288.

Wernersson, R., & Pedersen, A. G. (2003). RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Research*, *31*(13), 3537–3539.

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., Zhang, L., Abutarboush, M. H., Xing, Y., & Xie, Z. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nature Communications*, *5*(1), 1–10. https://doi.org/10/f6n79n

Wu, L., Yavas, G., Hong, H., Tong, W., & Xiao, W. (2017). Direct comparison of performance of single nucleotide variant calling in human genome with alignment-based and assembly-based approaches. *Scientific Reports*, *7*(1), 1–9.

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C. Y., & Wei, L. (2011). KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Research*, *39*(Web Server issue), W316–W322. PubMed. https://doi.org/10.1093/nar/gkr483

Xie, C., & Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, *10*(1), 1–9.

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329–342.

Yang, J., Li, W. R., Lv, F.-H., He, S.-G., Tian, S. L., Peng, W. F., Sun, Y. W., Zhao, Y. X., Tu, X. L., & Zhang, M. (2016). Whole-genome sequencing of native sheep provides insights into rapid adaptations to extreme environments. *Molecular Biology and Evolution*, *33*(10), 2576–2592.

Yang, L., Xu, L., Zhou, Y., Liu, M., Wang, L., Kijas, J. W., Zhang, H., Li, L., & Liu, G. E. (2018). Diversity of copy number variation in a worldwide population of sheep. *Genomics*, *110*(3), 143–148.

Yang, Y., Chung, E. K., Wu, Y. L., Savelli, S. L., Nagaraja, H. N., Zhou, B., Hebert, M., Jones, K. N., Shu, Y., Kitzmiller, K., Blanchong, C. A., McBride, K. L., Higgins, G. C., Rennebohm, R. M., Rice, R. R., Hackshaw, K. V., Roubey, R. A. S., Grossman, J. M., Tsao, B. P., … Yu, C. Y. (2007). Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): Low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *American Journal of Human Genetics*, *80*(6), 1037–1054. PubMed. https://doi.org/10.1086/518257

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591.

Yang, Z., & Dos Reis, M. (2010). Statistical properties of the branch-site test of positive selection. *Molecular Biology and Evolution*, *28*(3), 1217–1228.

Yang, Z., & Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, *19*(6), 908–917.

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A. M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, *155*(1), 431–449.

Yang, Z., Wong, W. S., & Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, *22*(4), 1107–1118.

Yates, C. M., & Sternberg, M. J. (2013). Proteins and domains vary in their tolerance of non-synonymous single nucleotide polymorphisms (nsSNPs). *Journal of Molecular Biology*, *425*(8), 1274–1286.

Zeng, L., Cao, Y., Wu, Z., Huang, M., Zhang, G., Lei, C., & Zhao, Y. (2019). A missense mutation of the HSPB7 gene associated with heat tolerance in Chinese Indicine cattle. *Animals*, *9*(8), 554.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., & Girón, C. G. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761.

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, *10*, 451–481.

Zhang, J. (2004). Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular Biology and Evolution*, *21*(7), 1332–1339.

Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*, *22*(12), 2472–2479.

Zhang, X., Wang, K., Wang, L., Yang, Y., Ni, Z., Xie, X., Shao, X., Han, J., Wan, D., & Qiu, Q. (2016). Genome-wide patterns of copy number variation in the Chinese yak genome. *BMC Genomics*, *17*(1), 1–12.

Zhang, Z. D., Du, J., Lam, H., Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). Identification of genomic indels and structural variations using split reads. *BMC Genomics*, *12*(1), 1–12.

Zhang, Z., Miteva, M. A., Wang, L., & Alexov, E. (2012). Analyzing effects of naturally occurring missense mutations. *Computational and Mathematical Methods in Medicine*, *2012*.

Zhao, M., Wang, Q., Wang, Q., Jia, P., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. *BMC Bioinformatics*, *14*(11), 1–16.

Zhou, H. D., Li, X. L., Li, G. Y., Zhou, M., Liu, H. Y., Yang, Y. X., Deng, T., Ma, J., & Sheng, S.-R. (2008). Effect of SPLUNC1 protein on the Pseudomonas aeruginosa and Epstein-Barr virus. *Molecular and Cellular Biochemistry*, *309*(1), 191–197. https://doi.org/10.1007/s11010-007-9659-3

Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C. P., & Sonstegard, T. S. (2009). A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biology*, *10*(4), 1–10.

Zingoni, A., Molfetta, R., Fionda, C., Soriani, A., Paolini, R., Cippitelli, M., Cerboni, C., & Santoni, A. (2018). NKG2D and Its Ligands: "One for All, All for One". *Frontiers in Immunology*, *9*, 476-480. https://doi.org/10.3389/fimmu.2018.00476

APPENDICES

Appendix 1: Extraction of C. nubiana DNA using the Phenol Chloroform Protocol

DNA was extracted from 21 liver tissues which been harvested from *C. nubiana* which had died of natural causes and kept at the National Zoological Gardens, Pretoria, South Africa. The tissues had been tranoosted from South Africa in absolute ethanol and stored in ethanol at -80 o C in a freezer at the Biosciences eastern and central Africa laboratories in Kenya.

1. Tissue samples stored in ethanol (25mg) were cut and put in a sterile microcentrifuge tube. Phosphate Buffered Saline (PBS) was then added to aid the tissue to regain its physiological and structural integrity; the tissue was left for 1 hour.
2. The tissue was removed from the Phosphate Buffered Saline (PBS) and cut into small pieces and put in a sterile microcentrifuge tube. 700µl of digestion buffer was added plus 10µl of proteinase K and was incubated at 50^{o}C overnight.
3. 20µl of RNASe A was then added and incubated at 37^{o}C for 1 hour
4. 700 µl of phenol/chloroform/isoamyalcohol (25:24:1) was added and mixed well for 15 minutes then centrifuged at 13000 rpm for 10 minutes; the aqueous phase was then transferred into new tube
5. Step 3 once was then repeated once.
6. 700 µl of chloroform was then added and centrifuged at 13000 rpm for 10 minutes and the aqueous phase was transferred into a new tube.
7. Step 5 was repeated once
8. One tenth volume of 3m Sodium acetate (NaOAc) was then added, then two volumes of 100% ethyl alcohol were added.
9. The DNA started precipitating after mixing it gently
10. The DNA precipitate was then transferred using a pipette tip into a new Eppendorf tube
11. 70% ethanol was then added to the precipitated and left on the bench for 5 minutes
12. The 70% ethanol was then removed and the DNA pellets were air-dried for 15 minutes and then it was re-suspended in 100 µl of Tris EDTA buffer.
13. The DNA was then quantified using UV spectroscopy using Thermo Scientific Nanodrop 2000c.
14. DNA qualitative analysis was finally carried using gel electrophoresis in 1.5% agarose gel

Appendix 2: Data sources for the species used as background data in positive selection analysis

Species name	Data source
Bos taurus (Cow)	ftp://ftp.ensembl.org/pub/release-97/fasta/bos_taurus/
Ovis aries (Sheep)	ftp://ftp.ensembl.org/pub/release-97/fasta/ovis_aries/
Equus caballus (Horse)	ftp://ftp.ensembl.org/pub/release-97/fasta/equus_caballus/
Equus asinus asinus (Donkey)	ftp://ftp.ensembl.org/pub/release-97/fasta/equus_asinus_asinus/
Sus scrofa (Pig)	ftp://ftp.ensembl.org/pub/release-97/fasta/sus_scrofa/
Pantera tigris altaica (Tiger)	ftp://ftp.ensembl.org/pub/release-97/fasta/panthera_tigris_altaica/
Felis catus (Cat)	ftp://ftp.ensembl.org/pub/release-97/fasta/felis_catus/
Canis familiaris (Dog)	ftp://ftp.ensembl.org/pub/release-97/fasta/canis_familiaris/
Capra hircus (Domestic goat)	ftp://ftp.ensembl.org/pub/release-97/fasta/capra_hircus/
Bos mutus (Wild Yak)	ftp://ftp.ensembl.org/pub/release-97/fasta/bos_mutus/
Bison bison bison (American Bison)	ftp://ftp.ensembl.org/pub/release-97/fasta/bison_bison_bison/
Ailuropoda melanoleuca (Krishnan & Panda)	ftp://ftp.ensembl.org/pub/release-97/fasta/ailuropoda_melanoleuca/
Bubalus bubalis (Water Buffallo)	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/121/395/GCF_003121395.1_UOA_WB_1/
Pantholops hodgsonii (Tibetian Antelope)	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/400/835/GCF_000400835.1_PHO1.0/
Capra aegagrus (Bezoar)	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/978/405/GCA_000978405.1_CapAeg_1.0/
C. nubiana (*C. nubiana*)	Generated and described in chapter 3

Appendix 3: CodeML control file

Alternate model (H1)

Input data and parameters	**Description of parameters**
seqfile = Ibex-aligments.phy	**sequence data file name
treefile =Ibex-trees.nwk	**result file name
outfile =CodeML-output.txt	**tree structure file name
CodonFreq = 2	** 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table
cleandata = 1	** remove sites with ambiguity data (1:yes, 0:no)?
NSsites = 2	** 0:one w; 1:NearlyNeutral; 2:PositiveSelection; 3:discrete;
	** 4:freqs; 5:gamma;6:2gamma;7:beta;8:beta&w;9:betaγ10:3normal
model = 2	** models for codons:
	* 0:one, 1:b, 2:2 or more dN/dS ratios for branches
fix omega = 0	** 1: omega or omega 1 fixed, 0: estimate
omega = 1	** initial or fixed omega, for codons or codon-based AAs
fix kappa = 0	** 1: kappa fixed, 0: kappa to be estimated
kappa = 2	** initial or fixed kappa
fix alpha = 1	** 0: estimate gamma shape parameter; 1: fix it at alpha
alpha = 0	** initial or fixed alpha, 0:infinity (constant rate)
clock = 0	** 0: no clock, unrooted tree, 1: clock, rooted tree
runmode = 0	**0: user tree; 1: semi-automatic; 2: automatic
	* 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise
Small Diff = .45e-6	** Default value.
method = 1	** 0: simultaneous; 1: one branch at a time
aaDist = 0	** 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a, 7:AAClasses
RateAncestor = 1	** (0,1,2): rates (alpha>0) or ancestral states (1 or 2)
icode = 0	** 0:standard genetic code; 1:mammalian mt; 2-10:see below
seqtype = 1	** 1:codons; 2:AAs; 3:codons-->AAs
getSE = 0	** 0: don't want them, 1: want S.E.s of estimates
noisy = 0	** 0,1,2,3,9: how much rubbish on the screen
ndata = 1	** specifies the number of separate data sets in the file
verbose = 1	** 1: detailed output, 0: concise output
fix blength =0	** 0: ignore, -1: random, 1: initial, 2: fixed

Null model (H0)

Input data and parameters	**Description of parameters**
seqfile = Ibex-aligments.phy	**sequence data file name
treefile =Ibex-trees.nwk	**result file name
outfile= CodeML- output.txt	**tree structure file name
CodonFreq = 2	** 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table
cleandata = 1	** remove sites with ambiguity data (1:yes, 0:no)?
NSsites = 2	** 0:one w; 1:NearlyNeutral; 2:PositiveSelection; 3:discrete;
	** 4:freqs; 5:gamma;6:2gamma;7:beta;8:beta&w;9:betaγ10:3normal
model = 2	** models for codons:
	* 0:one, 1:b, 2:2 or more dN/dS ratios for branches
fix omega = 1	** 1: omega or omega 1 fixed, 0: estimate
omega = 1	** initial or fixed omega, for codons or codon-based AAs
fix kappa = 0	** 1: kappa fixed, 0: kappa to be estimated
kappa = 2	** initial or fixed kappa
fix alpha = 1	** 0: estimate gamma shape parameter; 1: fix it at alpha
alpha = 0	** initial or fixed alpha, 0:infinity (constant rate)
clock = 0	** 0: no clock, unrooted tree, 1: clock, rooted tree
runmode = 0	**0: user tree; 1: semi-automatic; 2: automatic
	* 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise
Small Diff = .45e-6	** Default value.
method = 1	** 0: simultaneous; 1: one branch at a time
aaDist = 0	** 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a, 7:AAClasses
RateAncestor = 1	** (0,1,2): rates (alpha>0) or ancestral states (1 or 2)
icode = 0	** 0:standard genetic code; 1:mammalian mt; 2-10:see below
seqtype = 1	** 1:codons; 2:AAs; 3:codons-->AAs
getSE = 0	** 0: don't want them, 1: want S.E.s of estimates
noisy = 0	** 0,1,2,3,9: how much rubbish on the screen
ndata = 1	** specifies the number of separate data sets in the file
verbose = 1	** 1: detailed output, 0: concise output
fix blength =0	** 0: ignore, -1: random, 1: initial, 2: fixed

Appendix 4: CodeML output for the positively selected genes. lnL1 is the log likelihood for the alternate model, while lnL0 is the log likelihood for the null model. LRT is the Likelihood ratio test

Gene id	Omega(background)	Omega(foreground)	lnL1	lnL0	LRT=2 (lnL1-lnL0)	p-value
ENSCHIT00000003090	0.03	335.66	-8354.306	-8358.077	7.544	0.006
ENSCHIT00000004084	0.07	999	-2136.448	-2142.008	11.12	0.001
ENSCHIT00000004434	0.11	999	-4942.614	-4948.444	11.661	0.001
ENSCHIT00000008957	0.068	999	-5009.075	-5011.851	5.552	0.018
ENSCHIT00000010253	0.028	580.881	-2260.117	-2263.564	6.893	0.009
ENSCHIT00000012782	0.104	993.763	-4248.15	-4251.554	6.808	0.009
ENSCHIT00000015750	0.061	105.035	-4255.621	-4257.852	4.46	0.035
ENSCHIT00000018881	0.006	999	-4344.798	-4349.488	9.379	0.002
ENSCHIT00000026283	0.078	999	-3790.464	-3796.966	13.004	0
ENSCHIT00000028977	0.091	156.533	-8989.648	-8992.156	5.017	0.025
ENSCHIT00000029782	0.079	439.708	-5237.997	-5241.376	6.758	0.009
ENSCHIT00000030384	0.025	83.055	-3160.089	-3796.966	6.817	0.009
ENSCHIT00000000612	0.11	142.837	-8159.858	-8162.639	5.562	0.018
ENSCHIT00000015914	0.042	615.743	-2926.199	-2930.712	9.027	0.003
ENSCHIT00000016318	0.021	183.165	-2214.032	-2216.602	5.14	0.023
ENSCHIT00000020934	0.073	45.53	-1823.493	-1826.138	5.29	0.021
ENSCHIT00000028741	0.053	255.305	-23013.418	-23016.806	6.775	0.009
ENSCHIT00000035903	0.084	322.085	-14727.189	-14730.449	6.519	0.011
ENSCHIT00000036547	0.045	106.727	-7406.648	-7409.864	6.432	0.011
ENSCHIT00000040177	0.113	999	-1375.419	-1379.966	9.094	0.003
ENSCHIT00000040379	0.067	471.664	-3022.521	-3026.93	8.818	0.003
ENSCHIT00000034768	0.093	223.251	-5256.636	-5260.786	8.301	0.004
ENSCHIT00000041152	0.074	129.156	-9716.72	-9720.679	7.917	0.005

Appendix 5: Positively selected amino acid sites and impact on gene function predicted by Bayes empirical Bayes and Polyphen-2 Analysis. *Sites with posterior probabilities > 0.95 probably positively selected sites

		Bayes empirical Bayes (BEB) analysis		**Polyphen-2 analysis**	
Ensemble gene id	Gene name	Amino acid changes	Posterior probabilities	Score	Impact
ENSCHIT00000003090	Storkhead box 2	T734V	0.893	0.049	benign
		N835T	0.814	0.065	benign
ENSCHIT00000004084	ATPase H+ transporting V1 subunit E2	M72N	0.998**	0.711	Possibly damaging
ENSCHIT00000004434	olfactory receptor 2G2-like	F73T	0.997**	0.996	probably damaging
ENSCHIT00000008957	Serine protease 56	Q424L	0.917	0.001	benign
		R425G	0.846	0.992	probably damaging
		R436W	0.845	0.002	benign
		A548G	0.739	0	benign
ENSCHIT00000010253	Matrix AAA peptidase interacting protein 1	T76A	0.835	0	benign
		Q93P	0.943	0	benign
ENSCHIT00000012782	Putative olfactory receptor 52P1	M67L	0.924	0.889	Possibly damaging
ENSCHIT00000015750	Prostaglandin I2 synthase	A79M	0.747	0.832	possibly damaging
		R320H	0.831	0.816	possibly damaging
		D411E	0.898	0	benign
ENSCHIT00000018881	F-box protein 21	S603A	0.864	0	benign
		E606G	0.972*	0.002	benign
		K615E	0.974*	0.11	benign
		K616R	0.971*	0.884	possibly damaging

		E620G	0.999**	0	benign
ENSCHIT00000026283	Zinc finger and SCAN domain containing 23	P213N	0.969*	0.997	probably damaging
ENSCHIT00000028977	UV stimulated scaffold protein A	D361G	0.917	0.992	probably damaging
		A517T	0.897	0.001	benign
ENSCHIT00000029782	Leucine rich repeats and WD repeat domain containing 1	T61M	0.767	0.172	benign
		E99Q	0.789	1	probably damaging
		A588T	0.782	0.003	benign
ENSCHIT00000030384	F-box and WD repeat domain containing 2	L82C	0.954*	0.998	probably damaging
ENSCHIT00000000612	Multimerin 2	S214H	0.821	0.997	probably damaging
		A559T	0.541	0.009	benign
ENSCHIT00000015914	Toll like receptor adaptor molecule 2	R43H	0.56	0	benign
		I213N	0.965*	0.297	benign
ENSCHIT00000016318	eukaryotic translation initiation factor 2 subunit beta	K83I	0.982*	0.947	possibly damaging
		K205E	0.907	0.528	possibly damaging
ENSCHIT00000020934	LY6/PLAUR domain containing 6B	A7T	0.957*	0	benign
		F16L	0.911	0	benign
ENSCHIT00000028741	ATP binding cassette subfamily A member 12	M570T	0.904	0.74	possibly damaging
ENSCHIT00000035903	PATJ crumbs cell polarity complex component	V249I	0.891	0.376	benign
		I1738F	0.896	1	probably damaging
		I1739V	0.864	0.012	benign
ENSCHIT00000036547	Rho GTPase activating protein 42	I502L	0.92	0.004	benign
		M770T	0.986*	0.001	benign
		W773R	0.919	0.999	probably damaging

ENSCHIT00000040177	Achaete-scute family bHLH transcription factor 4	L30S	0.999**	0.999	probably damaging
ENSCHIT00000040379	olfactory receptor 1P1	A133T	0.835	0.001	benigh
		V135D	0.928	0.795	possibly damaging
		H159C	0.998**	0.999	probably damaging
ENSCHIT00000034768	tripartite motif containing 16	A155T	0.694	0.95	possibly damaging
		D159L	0.969*	0.418	benign
		S515L	0.872	0.151	benign
ENSCHIT00000041152	centrosomal protein 112	K338G	0.984*		unknown

Appendix 6: Gene ontology terms associated with positively selected genes

Ensemble id	Molecular functions	Cellular component	Biological process
ENSCHIT00000000612	GO:0005515~protein binding	GO:0005604~basement membrane,GO:0005615~extracellular space,GO:0031012~extracellular matrix,GO:0070062~extracellular exosome	GO:0001525~angiogenesis,GO:0030948~negative regulation of vascular endothelial growth factor receptor signaling pathway,GO:0090051~negative regulation of cell migration involved in sprouting angiogenesis
ENSCHIT00000003090			GO:0001893~maternal placenta development,GO:0009790~embryo development
ENSCHIT00000004084	GO:0005515~protein binding,GO:0008553~hydrogen-exporting ATPase activity, phosphorylative mechanism,GO:0046961~proton-transporting ATPase activity, rotational mechanism	GO:0001669~acrosomal vesicle, GO:0005829~cytosol, GO:0033178~proton-transporting two-sector ATPase complex, catalytic domain	GO:0008286~insulin receptor signaling pathway, GO:0015991~ATP hydrolysis coupled proton transport, GO:0016241~regulation of macroautophagy, GO:0033572~transferrin transport,GO:0034220~ion transmembrane transport, GO:0090383~phagosome acidification
ENSCHIT00000008957	GO:0004252~serine-type endopeptidase activity	GO:0005783~endoplasmic reticulum	GO:0006508~proteolysis, GO:0043010~camera-type eye development
ENSCHIT00000010253	GO:0005515~protein binding, GO:0043022~ribosome binding	GO:0005739~mitochondrion, GO:0005743~mitochondrial inner membrane, GO:0005759~mitochondrial matrix	GO:0007007~inner mitochondrial membrane organization, GO:0032979~protein insertion into mitochondrial membrane from inner side, GO:0036444~calcium ion transmembrane import into mitochondrion,GO:0051204~protein insertion into mitochondrial membrane,GO:0051560~mitochondrial calcium ion homeostasis,GO:0097033~mitochondrial

			respiratory chain complex III biogenesis,GO:0097034~mitochondrial respiratory chain complex IV biogenesis
ENSCHIT00000012782	GO:0004984~olfactory receptor activity, GO:0004930~G protein-coupled receptor activity	GO:0016021~ integral component of membrane	GO:0007186~G protein-coupled receptor signaling pathway
ENSCHIT00000015750	GO:0004497~monooxygenase activity, GO:0005506~iron ion binding, GO:0005515~protein binding, GO:0008116~prostaglandin-I synthase activity, GO:0016705~oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen,GO:0020037~heme binding	GO:0005615~extracellular space,GO:0005634~nucleus,GO:0005783~endoplasmic reticulum,GO:0005789~endoplasmic reticulum membrane,GO:0005901~caveola,GO:0016021~integral component of membrane	GO:0001516~prostaglandin biosynthetic process,GO:0006690~icosanoid metabolic process,GO:0006769~nicotinamide metabolic process,GO:0007566~embryo implantation,GO:0019371~cyclooxygenase pathway,GO:0032088~negative regulation of NF-kappaB transcription factor activity,GO:0035360~positive regulation of peroxisome proliferator activated receptor signaling pathway,GO:0045019~negative regulation of nitric oxide biosynthetic process,GO:0045766~positive regulation of angiogenesis,GO:0046697~decidualization,GO:0050728~negative regulation of inflammatory response,GO:0055114~oxidation-reduction process,GO:0071347~cellular response to interleukin-1,GO:0071354~cellular response to interleukin-6,GO:0071456~cellular response to hypoxia,GO:0097190~apoptotic signaling

			pathway,GO:1900119~positive regulation of execution phase of apoptosis
ENSCHIT00000016318	GO:0003723~RNA binding, GO:0003743~translation initiation factor activity, GO:0005515~protein binding, GO:0008135~translation factor activity, RNA binding, GO:0044822~poly(A) RNA binding, GO:0046872~metal ion binding	GO:0005634~nucleus, GO:0005737~cytoplasm, GO:0005829~cytosol, GO:0005850~eukaryotic translation initiation factor 2 complex	GO:0001701~in utero embryonic development, GO:0002176~male germ cell proliferation, GO:0006413~translational initiation, GO:0008584~male gonad development, GO:0055085~transmembrane transport
ENSCHIT00000018881	GO:0003677~DNA binding, GO:0004842~ubiquitin-protein transferase activity	GO:0000151~ubiquitin ligase complex	GO:0006511~ubiquitin-dependent protein catabolic process, GO:0016567~protein ubiquitination
ENSCHIT00000020934		GO:0005886~plasma membrane,GO:0031225~anchored component of membrane	
ENSCHIT00000023110	GO:0005515~protein binding,	GO:0000922~spindle pole,GO:0005737~cytoplasm,GO:0005813~centrosome,GO:0016021~integral component of membrane	
ENSCHIT00000026283	GO:0003700~transcription factor activity, sequence-specific DNA	GO:0005634~nucleus, GO:0070062~extracellular exosome	GO:0006351~transcription, DNA-templated, GO:0006355~regulation of transcription, DNA-templated

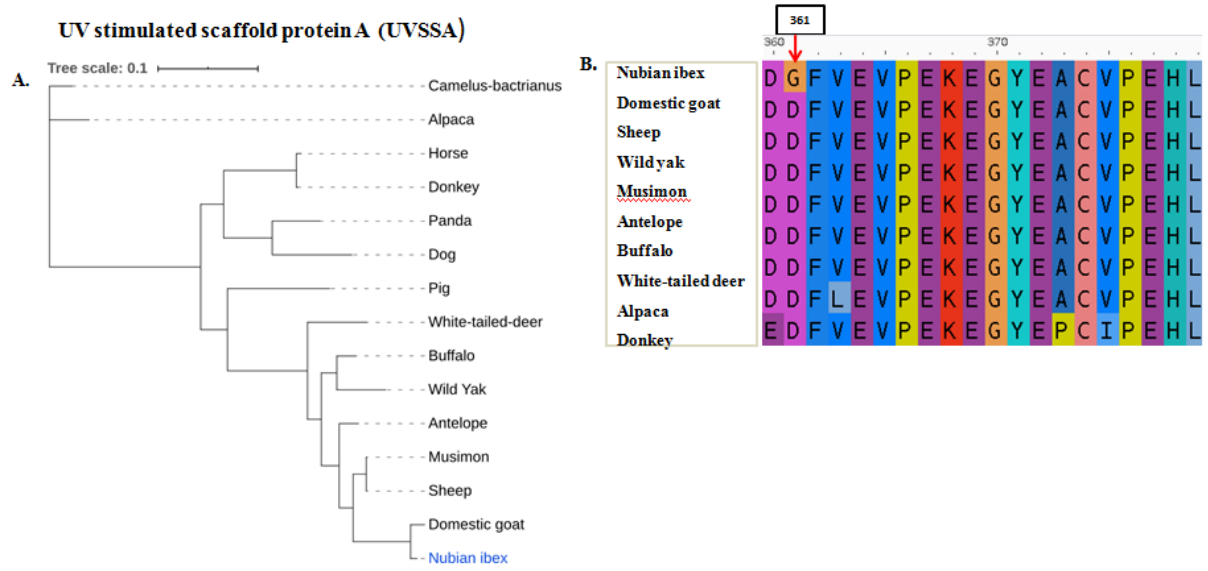
	binding,GO:0043565~sequence-specific DNA binding,GO:0046872~metal ion binding		
ENSCHIT00000028741	GO:0005102~receptor binding,GO:0005215~transporter activity,GO:0005319~lipid transporter activity,GO:0005515~protein binding,GO:0005524~ATP binding,GO:0016887~ATPase activity,GO:0034040~lipid-transporting ATPase activity,GO:0034191~apolipoprotein A-I receptor binding,GO:0042626~ATPase activity, coupled to transmembrane movement of substances	GO:0005737~cytoplasm,GO:0005743~mitochondrial inner membrane,GO:0005829~cytosol,GO:0005886~plasma membrane,GO:0016021~integral component of membrane,GO:0097209~epidermal lamellar body	GO:0006869~lipid transport,GO:0010875~positive regulation of cholesterol efflux,GO:0019725~cellular homeostasis,GO:0031424~keratinization,GO:0032940~secretion by cell,GO:0033700~phospholipid efflux,GO:0035627~ceramide transport,GO:0043129~surfactant homeostasis,GO:0045055~regulated exocytosis,GO:0048286~lung alveolus development,GO:0055085~transmembrane transport,GO:0055088~lipid homeostasis,GO:0061436~establishment of skin barrier,GO:0072659~protein localization to plasma membrane,GO:2000010~positive regulation of protein localization to cell surface
ENSCHIT00000028977	GO:0000993~RNA polymerase II core binding,GO:0005515~protein binding	GO:0005654~nucleoplasm,GO:0005694~chromosome	GO:0006283~transcription-coupled nucleotide-excision repair,GO:0009411~response to UV,GO:0016567~protein ubiquitination
ENSCHIT00000029782	GO:0005515~protein binding, GO:0003682~chromatin binding,	GO:0005634~nucleus, GO:0005664~nuclear origin of replication recognition complex, GO:0031933~telomeric	GO:0006260~DNA replication, GO:0071169~ establishment of protein localization to chromatin, GO:0006325~chromatin organization

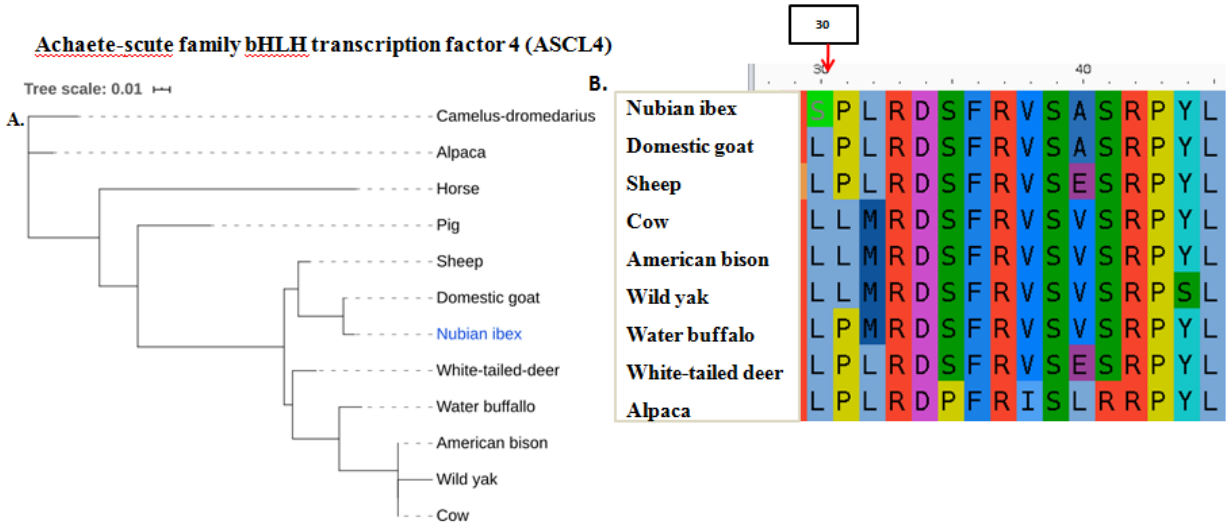
	GO:0035064~methylated histone binding	heterochromatin, GO:0005721~pericentric heterochromatin	
ENSCHIT00000030384	GO:0004842~ubiquitin-protein transferase activity,GO:0005515~protein binding		GO:0006464~cellular protein modification process,GO:0006508~proteolysis,GO:0016567~protein ubiquitination,
ENSCHIT00000034768			GO:0006355~regulation of transcription, DNA-templated,GO:0032526~response to retinoic acid,GO:0043966~histone H3 acetylation,GO:0043967~histone H4 acetylation,GO:0045618~positive regulation of keratinocyte differentiation,GO:0045893~positive regulation of transcription, DNA-templated,GO:0046683~response to organophosphorus,GO:0048386~positive regulation of retinoic acid receptor signaling pathway,GO:0050718~positive regulation of interleukin-1 beta secretion,GO:0060416~response to growth hormone
ENSCHIT00000035903	GO:0005515~protein binding,	GO:0005886~plasma membrane,GO:0005923~bicellular tight junction,GO:0016324~apical plasma membrane,GO:0043234~protein complex,GO:0048471~perinuclear region of	GO:0035556~intracellular signal transduction, GO:0070830~bicellular tight junction assembly,

		cytoplasm,GO:0070062~extracellular exosome	
ENSCHIT00000036547	GO:0005096~GTPase activator activity		GO:0003085~negative regulation of systemic arterial blood pressure,GO:0007165~signal transduction,GO:0035024~negative regulation of Rho protein signal transduction,GO:0043547~positive regulation of GTPase activity,GO:1904694~negative regulation of vascular smooth muscle contraction
ENSCHIT00000038507	GO:0015078~hydrogen ion transmembrane transporter activity, GO:0022857~transmembrane transporter activity	GO:0005739~mitochondrion,GO:0005743~mitochondrial inner membrane,GO:0005753~mitochondrial proton-transporting ATP synthase complex,GO:0016021~integral component of membrane,GO:0045263~proton-transporting ATP synthase complex, coupling factor F(o),GO:0070062~extracellular exosome	GO:0006754~ATP biosynthetic process, GO:0007568~aging, GO:0015986~ATP synthesis coupled proton transport, GO:0042776~mitochondrial ATP synthesis coupled proton transport, GO:0055093~response to hyperoxia
ENSCHIT00000040177	GO:0000977~RNA polymerase II regulatory region sequence-specific DNA binding,GO:0003700~transcription factor activity, sequence-specific DNA binding,GO:0005515~protein	GO:0090575~RNA polymerase II transcription factor complex	GO:0006351~transcription, DNA-templated, GO:0006357~regulation of transcription from RNA polymerase II promoter, GO:0043588~skin development

	binding,GO:0046983~protein dimerization activity		
ENSCHIT00000040379	GO:0004984~olfactory receptor activity, GO:0004930~G protein-coupled receptor activity	GO:0016021~integral component of membrane	GO:0007186~G protein-coupled receptor signaling pathway
ENSCHIT00000041152		GO:0005737~cytoplasm,GO:0005813~centrosome,GO:0005886~plasma membrane,GO:0060077~inhibitory synapse	GO:0097120~receptor localization to synapse,

Appendix 7: Gene trees and multiple sequence alignements used for positive selection analysis





Appendix 8: CNV associated protein-coding genes detected in three analyzed *C. nubiana*

Gene	Chr	CNV_start	Size	Read depth	SYMBOL	EXON	INTRON	CNV_typr
ENSCHIG00000027245	chr1	65862600	2500	0.206432	-	-	11/22	del_in_ibex
ENSCHIG00000015809	chr1	67890500	23300	0.211003	*MYLK*	1-2/31	1-2/30	dup_in_hircus
ENSCHIG00000020817	chr1	74335100	2300	0.00291621	*MB21D2*	-	1/1	del_in_ibex
ENSCHIG00000023257	chr1	143938000	5000	3.79449	*PIK3R4*	-	-	dup_in_ibex
ENSCHIG00000025923	chr1	151406400	9100	0.0521813	-	-	8/22	del_in_ibex
ENSCHIG00000026346	chr1	157369600	34000	2.10002	*PRDM9*	1-11/11	1-10/10	dup_in_ibex
ENSCHIG00000016852	chr2	117207700	17800	0.229459	-	-	9/19	dup_in_hircus
ENSCHIG00000017200	chr2	135392700	2100	0.00986789	*CYFIP1*	-	1/30	del_in_ibex
ENSCHIG00000010191	chr3	20800	26500	4.42965	*NBPF4*	1-8/13	1-8/12	dup_in_ibex
ENSCHIG00000019983	chr3	325700	10900	2.97466	-	-	-	dup_in_ibex
ENSCHIG00000022081	chr3	21602000	14700	1.91448	*CYP4A21*	1-14/15	1-14/14	dup_in_ibex
ENSCHIG00000026769	chr3	51665100	8800	1.92846	*SLC44A5*	-	1/22	dup_in_ibex
ENSCHIG00000026769	chr3	51674900	7900	1.92743	*SLC44A5*	-	1/22	dup_in_ibex
ENSCHIG00000024688	chr3	86843200	16900	2.16326	*GSTM4*	1-2/8	1-2/7	dup_in_ibex
ENSCHIG00000023592	chr3	86843200	16900	2.16326	-	-	-	dup_in_ibex
ENSCHIG00000005925	chr3	93562700	10700	0.209682	*DDX25*	1-4/5	1-4/4	dup_in_hircus
ENSCHIG00000010241	chr3	97722800	12900	3.29646	*PDE4DIP*	46/46	-	dup_in_ibex
ENSCHIG00000018067	chr3	111010400	15400	2.09274	*CD48*	2/4	1-2/3	dup_in_ibex
ENSCHIG00000018296	chr4	297600	4800	0.0209503	-	-	23/24	del_in_ibex
ENSCHIG00000024423	chr4	7003000	16400	1.80589	*GIMAP7*	2/2	-	dup_in_ibex
ENSCHIG00000009906	chr5	97803500	3400	0.199336	-	-	-	del_in_ibex
ENSCHIG00000009906	chr5	97808600	1600	0.117426	-	-	-	del_in_ibex
ENSCHIG00000015796	chr5	99356100	4600	1.82596	*Antigen*	3-4/19	2-4/18	dup_in_ibex
ENSCHIG00000013285	chr5	101608200	21100	2.48411	*CD163L1*	5-18/18	5-17/17	dup_in_ibex
ENSCHIG00000012612	chr5	101608200	21100	2.48411	-	-	-	dup_in_ibex
ENSCHIG00000012612	chr5	101656400	11700	0.288531	-	-	4/6	dup_in_hircus

ENSCHIG00000024286	chr5	108495000	10200	1.95877	-	-	-	dup_in_ibex
ENSCHIG00000017340	chr5	112038800	9500	1.55899	*CYP2D14*	5-9/9	4-8/8	dup_in_ibex
ENSCHIG00000012914	chr5	112058600	9800	1.78983	*CYP2D6*	2-9/9	1-8/8	dup_in_ibex
ENSCHIG00000018238	chr5	112058600	9800	1.78983	*TCF20*	-	-	dup_in_ibex
ENSCHIG00000018021	chr5	115617400	3300	0.1129	*CELSR1*	-	26/35	del_in_ibex
ENSCHIG00000018021	chr5	115752600	3100	0.0159134	*CELSR1*	1-2/36	1-2/35	del_in_ibex
ENSCHIG00000013097	chr6	5876900	12800	2.45276	*ARHGAP20*	20-24/26	19-24/25	dup_in_ibex
ENSCHIG00000017623	chr6	11318300	9000	0.227487	*UGT8*	-	1/4	dup_in_hircus
ENSCHIG00000015853	chr6	85040500	19500	2.55871	*UGT2B7*	1-2/6	1-2/5	dup_in_ibex
ENSCHIG00000011403	chr6	113107600	35800	2.00817	*MAN2B2*	1-14/22	1-14/21	dup_in_ibex
ENSCHIG00000016572	chr6	113107600	35800	2.00817	-	-	-	dup_in_ibex
ENSCHIG00000025479	chr6	113244100	4500	1.85053	*MAN2B2-like*	13-16/19	12-16/18	dup_in_ibex
ENSCHIG00000019204	chr6	114905600	2800	0	*TRMT44*	-	8/10	del_in_ibex
ENSCHIG00000016417	chr7	2591400	5200	0.00322182	*FER*	-	12/18	del_in_ibex
ENSCHIG00000017448	chr7	61279200	11300	252.225	*MT-ATP6*	1/4	1/3	dup_in_ibex
ENSCHIG00000008320	chr7	92685000	48800	1.89	*C3*	3-42/42	2-41/41	dup_in_ibex
ENSCHIG00000022235	chr7	93096500	12900	1.80499	*ADGRE3*	11-16/17	10-16/16	dup_in_ibex
ENSCHIG00000015838	chr7	93096500	12900	1.80499	-	-	-	dup_in_ibex
ENSCHIG00000000516	chr7	94689600	4800	1.92372	*MRPL4*	-	-	dup_in_ibex
ENSCHIG00000015941	chr7	94689600	4800	1.92372	*ICAM1*	-	-	dup_in_ibex
ENSCHIG00000020226	chr7	96343900	7000	0.150812	*ADGRE2*	1-5/18	1-4/17	dup_in_hircus
ENSCHIG00000025309	chr7	96343900	7000	0.150812	*Pcp2*	-	-	del_in_ibex
ENSCHIG00000011512	chr7	96343900	7000	0.150812	*XAB2*	-	-	del_in_ibex
ENSCHIG00000025046	chr8	22709200	6400	3.08422	*WC1*	-	-	dup_in_ibex
ENSCHIG00000010419	chr8	23058100	58500	3.37647	*IFNB1*	-	-	dup_in_ibex
ENSCHIG00000010419	chr8	23121200	9200	2.42644	-	-	-	dup_in_ibex
ENSCHIG00000015822	chr8	23133700	4100	2.09312	*IFNA2*	-	-	dup_in_ibex
ENSCHIG00000018064	chr8	58914800	74200	1.877898501	*FAM205A*	2/2	1/1	exonic
ENSCHIG00000018064	chr8	58937400	4500	2.01999	-	-	1/1	dup_in_ibex
ENSCHIG00000018064	chr8	58948500	5000	1.98085	-	-	1/1	dup_in_ibex

ENSCHIG00000007111	chr8	74405500	2300	0.0498154	*EPHX2*	-	5/18	del_in_ibex
ENSCHIG00000010514	chr8	75739900	19200	2.33315	*CNTFR*	-	3/9	dup_in_ibex
ENSCHIG00000004128	chr9	74011600	28900	4.26357	*NKG2D ligand 1-like*	2-5/5	1-4/4	dup_in_ibex
ENSCHIG00000023938	chr9	74047100	213800	3.16891	*NKG2D ligand 4-like*	1-3/4	1-3/3	exonic
ENSCHIG00000023938	chr9	74080800	17100	2.13832	-	-	3/3	dup_in_ibex
ENSCHIG00000023938	chr9	74108700	38100	2.17336	-	-	3/3	dup_in_ibex
ENSCHIG00000023938	chr9	74165000	14400	2.02566	-	-	3/3	dup_in_ibex
ENSCHIG00000023938	chr9	74182900	39100	2.44854	-	-	3/3	dup_in_ibex
ENSCHIG00000023938	chr9	74226100	34800	2.56723	-	-	3/3	dup_in_ibex
ENSCHIG00000008160	chr9	74316600	139000	2.0597	*ULBP3*	2-5/5	1-4/4	exonic
ENSCHIG00000008160	chr9	74409600	5800	2.07603	-	-	2/4	dup_in_ibex
ENSCHIG00000019400	chr9	74409600	5800	2.07603	*LRP11*	-	-	dup_in_ibex
ENSCHIG00000009062	chr9	74448200	7400	2.01217	-	-	-	dup_in_ibex
ENSCHIG00000008160	chr9	74448200	7400	2.01217	-	-	1/4	dup_in_ibex
ENSCHIG00000008160	chr9	74459600	10000	2.19994	-	1/5	1/4	dup_in_ibex
ENSCHIG00000008160	chr9	74459600	10000	2.19994	-	1/5	1/4	dup_in_ibex
ENSCHIG00000009062	chr9	74459600	10000	2.19994	*LRP11-like*	1-5/6	1-5/5	exonic
ENSCHIG00000021177	chr9	88088200	6200	1.65078	-	1/5	-	dup_in_ibex
ENSCHIG00000016479	chr9	88868200	3700	0.0848431	*RPS6KA2*	-	16/22	del_in_ibex
ENSCHIG00000024776	chr9	90205300	2900	0	*SMOC2*	-	9/12	del_in_ibex
ENSCHIG00000013268	chr10	78472100	5000	1.72914	*TRAV22*	-	-	dup_in_ibex
ENSCHIG00000015167	chr11	47892700	2500	0	*RNF103*	-	2/3	del_in_ibex
ENSCHIG00000002954	chr11	92974800	5100	1.98928	*OR1L8-like*	1/1	-	dup_in_ibex
ENSCHIG00000005457	chr11	105471600	4400	0.0191717	*CACNA1B*	-	15/46	del_in_ibex
ENSCHIG00000024860	chr11	105822100	2000	0.00226006	*PNPLA7*	-	14/34	del_in_ibex
ENSCHIG00000005530	chr12	14187699	19100	2.836326545	*MRP4-like*	1-3/31	1/30	dup-ibex
ENSCHIG00000022427	chr12	14673200	11100	0.468941	-	-	3/20	del_in_ibex
ENSCHIG00000016126	chr12	14851899	139200	2.004654842	*MRP4*	3-29/30	28-29/29	dup-ibex
ENSCHIG00000013067	chr12	34980100	21400	3.44002	*MRP4-like*	1-7/7	1-6/6	dup_in_ibex
ENSCHIG00000022569	chr13	27293500	5500	1.80671	*PHYH*	7-9/9	6-8/8	dup_in_ibex

ENSCHIG00000023950	chr13	37872200	2600	1.83061	*DZANK1*	-	-	dup_in_ibex
ENSCHIG00000015086	chr13	37872200	2600	1.83061	-	-	-	dup_in_ibex
ENSCHIG00000023950	chr13	37877500	3400	2.04604	*DZANK1*	-	17/18	dup_in_ibex
ENSCHIG00000023950	chr13	37881500	6000	1.91498	*DZANK1*	17/19	16-17/18	dup_in_ibex
ENSCHIG00000025369	chr13	61119100	28500	0.274563	*ASXL1*	-	3/11	dup_in_hircus
ENSCHIG00000021482	chr13	62201600	24200	2.00343	*BPIFA2*	1-4/6	1-4/5	dup_in_ibex
ENSCHIG00000025440	chr13	75313200	17400	4.15153	*ZMYND8*	13-16/23	12-16/22	dup_in_ibex
ENSCHIG00000020529	chr14	4908500	7000	0.396292	*CA1-like*	-	-	del_in_ibex
ENSCHIG00000024121	chr14	4957300	18200	2.80078	*CA1*	1-6/7	1-6/6	dup_in_ibex
ENSCHIG00000018727	chr14	38832100	1500	0.0113663	*MRPS28*	-	1/2	del_in_ibex
ENSCHIG00000013539	chr15	3170000	5100	3.17816	*KRTAP1-1*	1/11	1/10	dup_in_ibex
ENSCHIG00000008349	chr15	3917900	10200	1.90576	-	-	-	dup_in_ibex
ENSCHIG00000026736	chr15	3960600	83000	3.556837597	*FADS2-like*	1-5/11	1-5/10	
ENSCHIG00000025967	chr15	27449400	3400	0	*UVRAG*	-	13/14	del_in_ibex
ENSCHIG00000007741	chr15	32435400	14900	0.280498	*SSU72P3*	-	-	dup_in_hircus
ENSCHIG00000013289	chr15	35993300	19800	2.20337	*GVINP1-like*	-	-	dup_in_ibex
ENSCHIG00000001685	chr15	36015400	28100	2.37436	*GVINP1-like*	1/1	-	dup_in_ibex
ENSCHIG00000019212	chr16	924300	10000	1.62388	*ATP2B4-like*	2-6/6	1-5/5	dup_in_ibex
ENSCHIG00000026966	chr16	3867600	2800	0.102629	*MAPKAPK2*	-	2/10	del_in_ibex
ENSCHIG00000020774	chr16	5113300	4500	2.03464	-	-	21/21	dup_in_ibex
ENSCHIG00000020774	chr16	5120600	6800	2.16978	-	-	21/21	dup_in_ibex
ENSCHIG00000020774	chr16	5133300	3400	1.75236	-	-	21/21	dup_in_ibex
ENSCHIG00000020774	chr16	5137400	5300	1.80752	-	-	21/21	dup_in_ibex
ENSCHIG00000000467	chr16	5137400	5300	1.80752	*CFHR4*	-	-	dup_in_ibex
ENSCHIG00000021774	chr16	5911800	23200	1.975404857	*ATP2B4*	2-15/22	1-15/21	exonic
ENSCHIG00000003216	chr16	14991400	1100	0.027472	-	-	-	dup_in_ibex
ENSCHIG00000016126	chr16	14991400	1100	0.027472	-	-	2/29	dup_in_ibex
ENSCHIG00000019930	chr16	42908700	23100	1.94538	*GPR157*	3-4/4	2-3/3	dup_in_ibex
ENSCHIG00000018753	chr16	42933800	12800	1.79468	*SLC2A5*	1/12	1/11	dup_in_ibex
ENSCHIG00000020091	chr16	49628400	1900	0.00352751	*ANKRD65*	-	-	del_in_ibex

ENSCHIG00000010545	chr16	49628400	1900	0.00352751	*MRPL20*	-	-	del_in_ibex
ENSCHIG00000014992	chr16	56282500	25000	0.298457	*PAPPA2*	-	1/21	dup_in_hircus
ENSCHIG00000016826	chr16	58269700	3200	0.0223062	*RASAL2*	-	1/17	del_in_ibex
ENSCHIG00000003453	chr17	263600	38200	0.115182	*IGLV2-11*	1-2/2	1/1	del_in_ibex
ENSCHIG00000025371	chr17	645700	2200	1.94319	*IGLV1-40*	-	-	dup_in_ibex
ENSCHIG00000016084	chr17	645700	2200	1.94319	*IGLV5-45*	-	-	dup_in_ibex
ENSCHIG00000019259	chr17	68517000	14000	463.908	*GUCY1B1*	-	-	dup_in_ibex
ENSCHIG00000016409	chr18	1793200	15000	1.89604	*CSH2*	1/5	1/4	exonic
ENSCHIG00000025848	chr18	26110200	18500	2.29226	*CES1*	3-12/15	2-12/14	dup_in_ibex
ENSCHIG00000021270	chr18	51520800	20700	1.86789	*CYP2B6*	1/9	1/8	dup_in_ibex
ENSCHIG00000021270	chr18	51520800	20700	1.86789	*CYP2B6*	1/9	1/8	exonic
ENSCHIG00000018895	chr18	52827500	4400	0.284695	-	-	4/4	del_in_ibex
ENSCHIG00000024358	chr18	52827500	4400	0.284695	*PINLYP*	-	-	del_in_ibex
ENSCHIG00000010931	chr18	57722800	6400	2.9253	*JOSD2*	-	-	dup_in_ibex
ENSCHIG00000023228	chr18	58611900	10700	2.49271	*SIGLEC5*	1/2	1/1	dup_in_ibex
ENSCHIG00000023228	chr18	58623100	20800	2.37875	-	1/2	-	dup_in_ibex
ENSCHIG00000012247	chr18	59974600	6100	1.61126	*GNB1*	-	-	dup_in_ibex
ENSCHIG00000019435	chr18	59974600	6100	1.61126	*ZNF345*	-	-	dup_in_ibex
ENSCHIG00000021602	chr18	60135700	8300	1.57576	*ZNF665-like*	1/2	1/1	dup_in_ibex
ENSCHIG00000021602	chr18	60145500	8200	1.62231	*ZNF665-like*	-	-	dup_in_ibex
ENSCHIG00000023653	chr18	60212000	21600	2.21944	*ZNF501*	2/2	1/1	dup_in_ibex
ENSCHIG00000021602	chr18	60212000	21600	2.21944	-	-	3/3	dup_in_ibex
ENSCHIG00000017914	chr18	60783100	11800	3.76127	-	-	1/1	dup_in_ibex
ENSCHIG00000011266	chr18	64361500	8700	2.56187	*KIR2DL1*	1-3/4	1-3/3	dup_in_ibex
ENSCHIG00000022734	chr18	64361500	8700	2.56187	-	-	-	dup_in_ibex
ENSCHIG00000001659	chr19	40856200	14800	0.16045	*KRTAP3-1*	-	-	dup_in_hircus
ENSCHIG00000014439	chr19	40893400	18700	0.439653	*KRTAP1-1*	3/3	2/2	dup_in_hircus
ENSCHIG00000024904	chr19	56300600	3900	0.0067174	*DNAI2*	-	-	del_in_ibex
ENSCHIG00000022827	chr19	56300600	3900	0.0067174	*KIF19*	-	-	del_in_ibex
ENSCHIG00000024318	chr19	61931400	23600	0.313423	*PRKCA*	1/13	1/12	dup_in_hircus

ENSCHIG00000022006	chr19	62256400	3500	0.000962845	*CACNG4*	-	2/2	del_in_ibex
ENSCHIG00000014564	chr21	3594400	22300	0.203476	-	-	3/4	dup_in_hircus
ENSCHIG00000014430	chr21	15569200	1900	0.0160302	*AKAP13*	-	1/36	del_in_ibex
ENSCHIG00000015347	chr21	24518500	58600	6.74714	-	-	-	dup_in_ibex
ENSCHIG00000005122	chr21	24577500	45300	5.12384	*TNFRSF10B-like*	1-5/8	1-5/7	dup_in_ibex
ENSCHIG00000017644	chr21	39705600	2400	0.0378049	*PRKD1*	-	2/18	del_in_ibex
ENSCHIG00000008107	chr21	50804800	3800	0.0966095	*LRFN5*	-	1/5	del_in_ibex
ENSCHIG00000012212	chr21	54972000	3600	0.535566	*WDR76*	-	8/12	del_in_ibex
ENSCHIG00000014586	chr21	58140600	26100	1.99648	*IFI27L2*	2-5/5	1-4/4	dup_in_ibex
ENSCHIG00000017277	chr21	58140600	26100	1.99648	-	-	5/5	dup_in_ibex
ENSCHIG00000014562	chr21	58721700	7500	1.86488	*SerpinA3-6*	3-4/4	2-3/3	dup_in_ibex
ENSCHIG00000013116	chr23	14940500	28100	0.23343	*SerpinB6-like*	1-7/7	1-6/6	del_in_ibex
ENSCHIG00000014090	chr23	14940500	28100	0.23343	*SerpinB6*	4-7/7	3-6/6	del_in_ibex
ENSCHIG00000026493	chr23	16889100	9000	2.00543	*ACOT13*	1-2/3	1-2/2	dup_in_ibex
ENSCHIG00000019632	chr23	16889100	9000	2.00543	*C6orf62*	-	-	dup_in_ibex
ENSCHIG00000019632	chr23	16901300	11400	2.02299	*C6orf62*	1-5/5	1-4/4	dup_in_ibex
ENSCHIG00000026493	chr23	16901300	11400	2.02299	-	-	-	dup_in_ibex
ENSCHIG00000021899	chr23	20958900	13100	1.85886	*BoLA*	-	-	dup_in_ibex
ENSCHIG00000018017	chr23	20958900	13100	1.85886	*IFITM3*	-	-	dup_in_ibex
ENSCHIG00000012639	chr23	20994300	9500	1.56011	*OR2H1D*	-	-	dup_in_ibex
ENSCHIG00000009184	chr23	20994300	9500	1.56011	*UBD*	-	-	dup_in_ibex
ENSCHIG00000004385	chr23	21030300	18200	1.9336	*C19orf12*	-	-	dup_in_ibex
ENSCHIG00000006985	chr23	22585900	5800	2.24464	*STK19*	6-7/7	6/6	dup_in_ibex
ENSCHIG00000021630	chr23	22585900	18500	2.049109508	*C4A*	1-17/44	1-16/43	dup_in_ibex
ENSCHIG00000023354	chr23	22585900	18500	2.049109508	*C4B*	45/45	-	dup_in_ibex
ENSCHIG00000022212	chr23	22598000	6400	1.87191	*CYP21A2*	-	-	dup_in_ibex
ENSCHIG00000004476	chr23	37615100	9800	0.184295	*TMEM217*	1/3	-	dup_in_hircus
ENSCHIG00000025055	chr23	37615100	9800	0.184295	*TBC1D22B*	-	-	dup_in_hircus
ENSCHIG00000018673	chr24	33359600	10500	1.87453	*ANKRD29*	-	-	dup_in_ibex
ENSCHIG00000024653	chr24	33378100	64200	1.902857697	*RMC1*	5-19/19	4-18/18	dup_in_ibex

ENSCHIG00000021287	chr24	33378100	64200	1.902857697	*NPC1*	1-16/25	1-16/24	dup_in_ibex
ENSCHIG00000016115	chr24	54392900	7800	0.441613	*RAB27B*	-	1/5	dup_in_hircus
ENSCHIG00000023320	chr24	60768400	2700	0.0468324	*PIGN*	-	1/28	del_in_ibex
ENSCHIG00000018449	chr25	19015900	19100	0.268511	*CRYM-AS1*	3/3	2/2	dup_in_hircus
ENSCHIG00000026607	chr25	34497300	3600	1.7447	*OR4C15-like*	2/2	-	dup_in_ibex
ENSCHIG00000014409	chr26	369700	1300	0.0210928	*TUBGCP2*	-	1/17	del_in_ibex
ENSCHIG00000014162	chr26	369700	1300	0.0210928	*ZNF511*	-	-	del_in_ibex
ENSCHIG00000024348	chr26	35121300	5100	1.63559	*CYP2C31*	2-3/9	1-3/8	dup_in_ibex
ENSCHIG00000022550	chr26	48384700	2800	0.0585765	-	-	1/2	del_in_ibex
ENSCHIG00000019119	chr27	1358000	15600	0.274547	*ZNF385D*	-	2/7	dup_in_hircus
ENSCHIG00000002303	chr27	8383100	14900	1.97891	*POLB*	13-14/14	12-13/13	dup_in_ibex
ENSCHIG00000021808	chr27	10956200	15200	2.0154	*ADAMTS18-like*	18-20/20	17-19/19	dup_in_ibex
ENSCHIG00000000321	chr28	15738200	4100	0.0175047	*VCL*	-	1/21	del_in_ibex
ENSCHIG00000011162	chr28	42851600	2300	0.00311939	*PGBD5*	-	3/6	del_in_ibex
ENSCHIG00000011568	chr29	46015400	13200	0.00405503	*PKP3*	-	-	del_in_ibex

RESEARCH OUTPUTS

Output one: Publications

Chebii, V. J., Mpolya, E. A., Muchadeyi, F. C., & Domelevo, E. J. B. (2021). Genomics of Adaptations in Ungulates. *Animals*, 11(6). 1-20. https://doi.org/10.3390/ani11061617

Chebii, V. J., Mpolya, E. A., Oyola, S. O., Kotze, A., Entfellner, J. B. D., & Mutuku, J. M. (2021). Genome Scan for Variable Genes Involved in Environmental Adaptations of Nubian Ibex. *Journal of Molecular Evolution*, 2020, 1-16. https://doi.org/10.1007/s00239-021-10015-3

Chebii, V. J., Oyola, S. O., Kotze, A., Domelevo Entfellner, J. B., Musembi Mutuku, J., & Agaba, M. (2020). Genome-Wide Analysis of Nubian Ibex Reveals Candidate Positively Selected Genes That Contribute to Its Adaptation to the Desert Environment. *Animals*, 10(11), 1-20 https://doi.org/10.3390/ani10112181

Output two: Conferences

Scitalk and poster presentation: Exploring association of copy number variations in Wild African goats (*C. nubiana*) adaptations (https://virtual.keystonesymposia.org/ks/sessions/997/view). Presented at KEYSTONE SYMPOSIA on Molecular and Cellular Biology Leveraging Genomic Diversity to Promote Animal and Human Health (S5) held in November 25-29, 2018 in Kampala, Uganda.

Output three: Bioinformatics training

A trainer at Eastern African Bioinformatics Networks Training (EANBIT) held in 2018 and 2019 in Kenya Medical Research Institute, Kilifi, Kenya

Output four: *Capra nubiana* genomic resources

(i) *Capra nubiana* genome sequence data available at National Center for Biotechnology Information (Accession: SRR12990712).

(ii) *Capra nubiana* single nucleotide variants data

(iii) *Capra nubiana* copy number variants data

(iv) *Capra nubiana* coding DNA sequences