Insights into the genomics, evolution, and conservation of antelopes

Corey Kirkland

School of Biosciences

University of Kent

2025

A thesis submitted to the University of Kent for the degree of Doctor of Philosophy in Genetics

Declaration

No part of this thesis has been submitted in support of an application for any degree or other qualification of the University of Kent, or any other University or Institution of learning.

Corey Lee Kirkland

2025

Acknowledgements

I would like to start by thanking my supervisor, Dr Marta Farré Belmonte. I'm very honoured to have had her invaluable supervision over the past 6 years. This started with undergraduate research project in phylogenetics back in 2018, which sparked my interests in evolution and bioinformatics, to a MSc-R in conservation genetics and phylogenetics where I developed an interest in applying genetics to conservation. And now, through this PhD, I have developed a research interest in evolutionary genomics and the applications of this to conservation. Her support, knowledge, and encouragement have enabled me to continuously learn and grow as a researcher and this thesis would not be possible without her.

I would also like to thank the collaborators who have also contributed to this thesis. I had the pleasure of visiting Dr Rasmus Heller's group at the University of Copenhagen, where I learned population genomic methods and enjoyed exploring the city of Copenhagen. His ongoing support and sharing of genomic data have enhanced this study. Tony King from the Aspinall Foundation has been invaluable in the early creation of this project and for organising and sampling historical museum collections. Collaborators at the University of Barcelona, Professor Aurora Ruiz-Herrera and Dr Lucía Álvarez-González, and the DNA Zoo consortium, supported the Hi-C data and genome assembly work, and I really appreciate their contribution and support.

I'm also grateful to Dr Peter Ellis, members in the Ellis lab (Marie-Claire, Izzy, Liv, Sally, Richard) and Farré lab (Cristina, Carla, Frances, Dadu, Sarah), and the wider School of Biosciences who have been so supportive during my time at Kent. It's been an incredible place to study and work over the last 8 years.

Finally, I would like to thank my family and friends for their love and support. I would not have made it to this point without you.

Table of Contents

List of Figures

List of Tables

List of Supplementary Materials

Supplementary Figures

Supplementary Tables

Abstract

Genomic advancements have increased the feasibility of generating chromosome-level genomes for non-model animals, enabling evolutionary and conservation related research of understudied species. Here, we focus on antelopes, in the ruminant subfamily Bovidae, and explore their chromosome evolution, 3D genome organisation, and population genomics. Antelopes have diploid chromosome numbers ranging between 2*n* = 30 and 2*n* = 60, predominantly due to Robertsonian fusions of acrocentric chromosomes. Additionally, some species have chromosome fusions that have remained polymorphic within species. One example of this is the waterbuck (*Kobus ellipsiprymnus*), with chromosome polymorphisms within and between two recognised subspecies, causing karyotypes of between 2*n* = 50 and 2*n* = 54. To explore inter- and intra-species genomics of antelopes we firstly sequenced and assembled a chromosome-level genome for the waterbuck using PacBio HiFi long-reads and Hi-C. We then extracted DNA and whole genome sequenced (WGS) 24 historical waterbuck samples and combined this with WGS data for 119 modern samples to explore the population genomics and chromosome rearrangements of waterbuck. Lastly, we combined the newly assembled waterbuck genome and Hi-C with published chromosome-level genomes and Hi-C data to investigate chromosome evolution and 3D genome organisation across antelopes. This study provides a highly contiguous and complete genome assembly for the waterbuck. At the population level we support previous studies by showing genomic differences between and within the two subspecies, with strong barriers to gene flow, but also varying degrees of admixture. We identified multiple regions of high genomic differentiation between the two subspecies, several of them near chromosome fusions, suggesting putative signatures of these rearrangements, with these regions containing genes that may be involved in speciation. Between species, we reconstructed the chromosomes of two bovid ancestors and explored chromosome fusions. We found that commonly fused chromosomes were not found to be located closer within the 3D genome organisation of the nucleus in species without the fusion. This was also the case for the polymorphic chromosome fusion in waterbuck. This study provides an overview of the genomics and evolution of antelopes, raising important questions on how we conserve species with variable karyotypes and in early speciation, and with ongoing hybridisation.

Abbreviations

aDNA	Ancient DNA
CCS	Circular Consensus Sequencing
CDS	Coding Sequence
CENP-A	Centromere-Specific Protein A
CLR	Continuous Long Reads
CNVs	Copy Number Variants
CRs	Chromosome Rearrangements
CRT	Cyclic Reversible Termination
ddNTPs	Dideoxynucleotide Triphosphates
DIN	DNA Integrity Number
dNTPs	Deoxynucleotide Triphosphates
DRC	Democratic Republic of Congo
DSBs	Double-Strand Breaks
EBRs	evolutionary breakpoint regions
EEMS	Estimated Effective Migration Surface
EH	Evolutionary Highway
ESUs	Evolutionary Significant Units
FISH	Fluorescence in situ Hybridisation
GBS	Genotype-by-Sequencing
GO	Gene Ontology
hDNA	Historical DNA
HMW	High Molecular Weight
HSBs	Homologous Synteny Blocks
HWE	Hardy Weinberg Equilibrium
ILS	Incomplete Lineage Sorting
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
NAHR	Non-Allelic Homologous Recombination
NJ	Neighbour-Joining
NORs	Nucleolus Organiser Regions
ONT	Oxford Nanopore Technologies
PCA	Principle Component Analysis
PCM	Powell Cotton Museum
PHRs	Pseudo-Homologous Regions
PSMC	Pairwise Sequentially Markovian Coalescent
RACFs	Reconstructed Ancestral Chromosome Fragments
Rb	Robertsonian
RMCA	Royal Museum for Central Africa
ROH	Runs of Homozygosity
SBS	Sequencing by Synthesis
SDs	Segmental Duplications
SFS	Site Frequency Spectrum
SMRT	Single Molecule Real-Time
SNPs	Single-Nucleotide Polymorphisms
SVs	Structural Variants
T2T	Telomere to Telomere
TADs	Topological Associating Domains
TEs	Transposable Elements
TRs	Tandem Repeats
VGP	Vertebrate Genomes Project
WGS	Whole Genome Sequencing
ZMW	Zero-Mode Waveguide

1. General Introduction

1.1. The structure and function of genomes

The genome refers to the entire set of genetic information in an organism. In eukaryotic species this is in the form of DNA, frequently structured in the well-recognised right-handed double helix (B-DNA) and is packaged into chromosomes within the nucleus (**Figure 1**). DNA is compacted with histones (H2A, H2B, H3, and H4) that bind to DNA, forming DNA-histone complexes called chromatin. DNA is wrapped around eight histones, two of each type, in regular spacing across the genome to form nucleosomes, creating a structure that looks similar to beads (histones) on a string (DNA). These nucleosomes are further compacted together to form the 30 nm fibre, and these fibres can then form DNA loops.

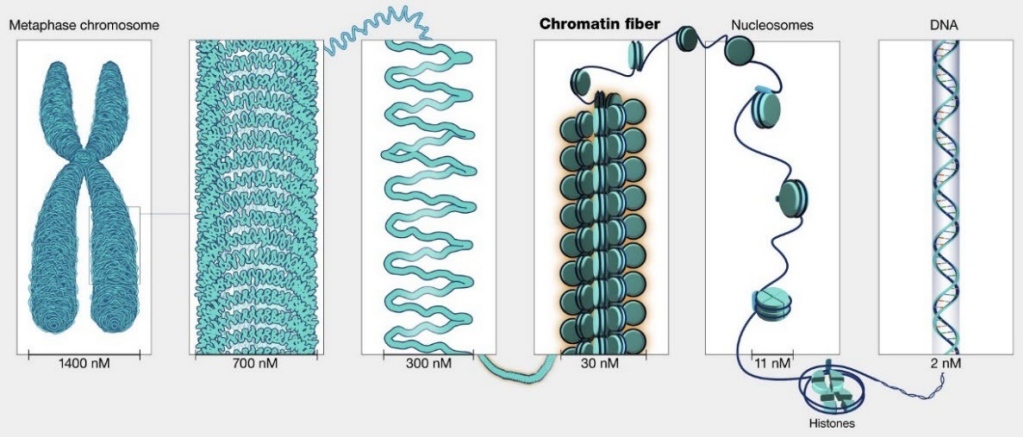


Figure 1: *The structure of DNA within the nucleus of eukaryotic cells. Adapted from https://www.genome.gov/genetics-glossary/Chromatin.*

The majority (~91%) of the genome is wrapped into the "beads on a string" structure, with these regions known as euchromatin. Here, the genome is more open, enriched in genes, and therefore more transcription can take place. The remaining regions of the genome are termed heterochromatin and are further condensed into the 30 nm fibre. These regions have less genes and are more often involved in gene expression. Two forms exist: constitutive heterochromatin and facultative heterochromatin. The former are found in mostly pericentric regions and telomeres, while the latter are regions that are transcriptionally silent where chromatin can decondense to allow transcription to take place (Trojer & Reinberg, 2007).

During mitosis, DNA is condensed even further into the characteristic structure of a chromosome, the metaphase chromosome. Three main types of chromosome structure exist in mammals: (i) metacentric, where the centromere is found at the centre of the

chromosome, (ii) submetacentric, where the centromere is positioned off-centre, or (iii) acrocentric, where the centromere is located at the end of the chromosome.

At interphase most mammalian chromosomes are organised into chromosome territories within the nucleus, where the position of genomic regions is correlated with transcriptional activity and where gene-rich regions are often near the borders of these territories (**Figure 2**; Zheng & Xie, 2019). Transcriptionally active regions often interact with each other and are termed compartment A and contain euchromatin. Whilst inactive regions are termed compartment B and contain heterochromatin. Within both types of compartments, the genome is further organised into topological associating domains (TADs), which are between 100 Kb and 1 Mb in length (Jerkovic & Cavalli, 2021). TAD boundaries are usually separated by CTCF-binding factors and cohesin-mediated loops to enable further chromatin folding within TADs.

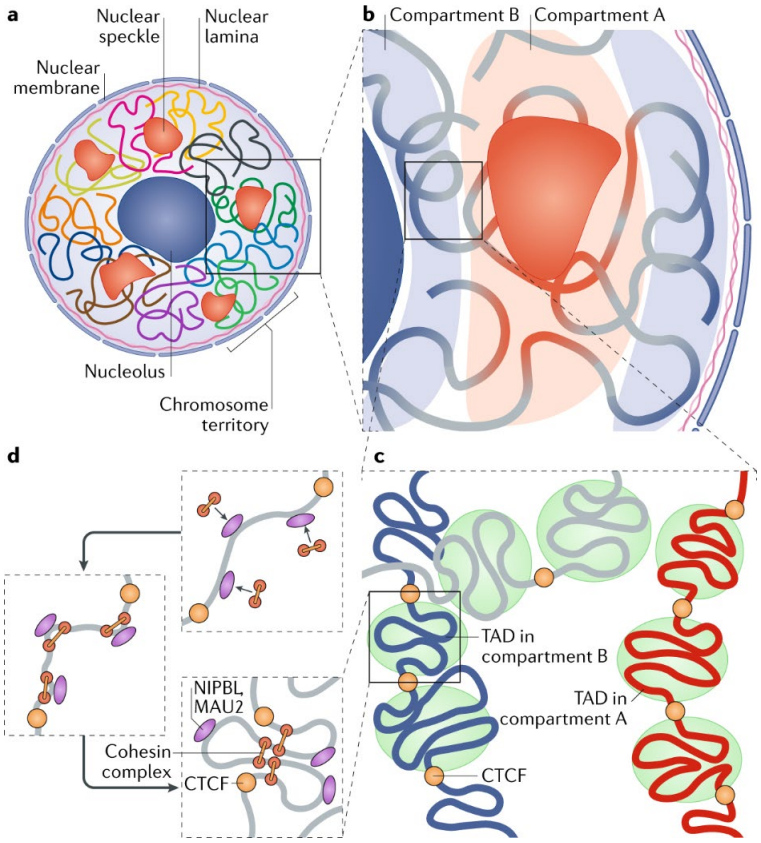


Figure 2: *3D genome organisation in the nucleus. From* Zheng & Xie, 2019.

1.2. Advances in DNA sequencing technologies

Genomes can be sequenced by a range of technologies to uncover the DNA sequence of organisms, whether that be for individual genes or whole genomes, revolutionising the fields of genetics and genomics. These DNA sequences can represent a reference for a

particular species (a reference genome assembly), several genomes can be sequenced for a species and compared to the reference (e.g., whole genome sequencing; WGS), or reference genomes can be assembled for several species and compared (e.g., comparative genomics). These approaches provide a wealth of information on the structure, function, and evolution of genomes.

1.2.1. Sanger sequencing

One of the first DNA sequencing technologies to be widely used in genetics (known as the 1st generation) was chain-termination sequencing or Sanger sequencing and can sequence DNA up to approximately 1,500 bp in length (Sanger et al., 1977). The technique makes use of a capillary gel, PCR products, and DNA polymerase to synthesise a new strand of DNA complementary to the starting PCR product. A solution is used containing the four deoxynucleotide triphosphates (dNTPs; dATP, dCTP, dGTP, and dTTP). The solution also contains dideoxynucleotide triphosphates (ddNTPs; ddATP, ddCTP, ddGTP, and ddTTP), each labelled with a fluorescent marker, which block further elongation of the DNA strand. As both types of nucleotides are present and are not preferentially used, some DNA strands elongate longer than others, resulting in DNA strands of varying lengths.

The identity of the “chain-terminated” molecule (i.e., the ddNTP) and the length of the DNA strand (and therefore its position) are resolved using capillary electrophoresis. DNA is separated by size and a fluorescence detector is used to determine the ddNTP. Sanger sequencing has provided a useful method to sequence PCR products of genes, mitochondrial markers, and microsatellites, and is still actively used in genetics.

1.2.2. Short-read sequencing

In order to sequence whole genomes more easily and cheaper than 1st generation technologies, next generation sequencing was developed. One of the most widely used platforms is Illumina, which utilises sequencing by synthesis (SBS), specifically cyclic reversible termination (CRT). This technique shares a similar approach to Sanger sequencing, where elongation is blocked by a ribose 3'-OH group (Guo et al., 2008; Ju et al., 2006).

Before SBS, a genomic library is created where DNA is fragmented, and adapters are ligated to each end. These adapters also include the sequencing binding site, indices, and regions complementary to the flow cell. The ligated fragments then attach to a dense

lawn of oligonucleotides on the solid flow cell slide using the complementary adapter region (**Figure 3**). Unlabelled dNTPs and DNA polymerase initiate solid-phase bridge amplification, where the unbound end binds to a nearby oligonucleotide and nucleotides are incorporated onto the DNA strand forming double-stranded DNA. The bridges are denatured, leaving two single-stranded DNA fragments bound to the two oligonucleotides. This step, known as clustering, is repeated and generates copies of the DNA fragment.

The reverse strands are then cleaved and four types of fluorescently labelled dNTPs that block elongation are added to begin SBS. For each DNA strand, a single dNTP hybridises onto the first base following the adapter, the remaining dNTPs are removed, and the slide is imaged to identify the nucleotide base. The fluorophores are then cleaved, and the reaction is repeated with new dNTPs to further elongate the DNA strand. After SBS, the index is read, and the sequencing products are washed away. For paired-end sequencing, the reverse strand is also sequenced. The forward DNA template folds over and binds to the second oligonucleotide. DNA is elongated by DNA polymerase and unlabelled dNTPs, forming the double-stranded DNA bridge. The template is cleaved, leaving only the reverse strands, and SBS then occurs.

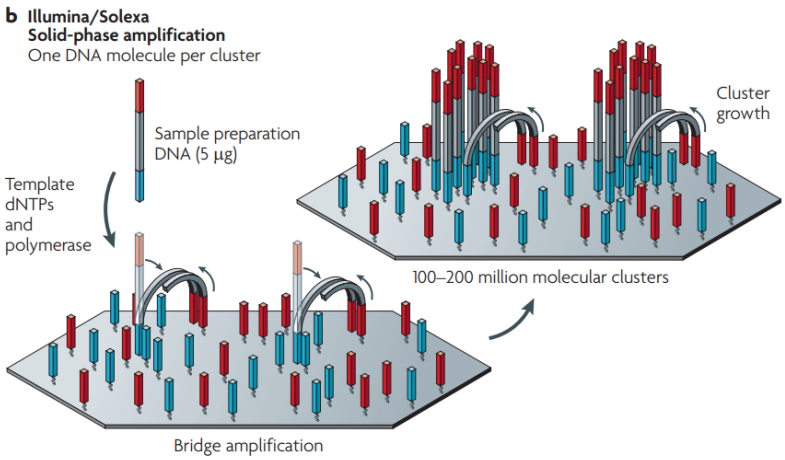


Figure 3: *Illumina sequencing. Clustering of oligonucleotides on the surface of the solid slide, showing the formation of DNA bridges through bride amplification. Adapted from* Metzker, 2010.

Illumina platforms typically sequence 150 bp DNA reads. For paired-end sequencing, this includes 150 bp at the 5' ends of the forward and reverse strands. DNA is often fragmented to a set size (the insert size) so that a region in the middle of the DNA fragment is not sequenced (known as the inner distance). The insert size is often 350 bp

in length, but additional techniques such as mate pair sequencing have enabled even longer insert sizes to be used, which can reach up to kilobases in length. Illumina sequencing technology additionally has a high level of accuracy (> 99%) and has been used to sequence whole genomes.

1.2.3. Long-read sequencing

Third generation, or long-read sequencing, was developed to enable even longer fragments of DNA to be sequenced, from kilobases to megabases in length. One technology is PacBio's Single Molecule Real-Time (SMRT) sequencing (Eid et al., 2009), which has been developed into two platforms, continuous long reads (CLR) and circular consensus sequencing (CCS; **Figure 4**). Both types require the preparation of a DNA library (a SMRTbell library) which includes the ligation of hairpin adapters to either end of the double-stranded DNA fragment, capping each end, and creating a circular template. Primers and a modified ϕ29 DNA polymerase are then included, which bind to the adapter. Both platforms also make use of a nanophotonic structure, called the zero-mode waveguide (ZMW), which detects the incorporation of fluorescently labelled dNTPs within the active site of the enzyme. The fluorescent labels are quickly cleaved and diffuse, before the next dNTP is incorporated. This results in real time sequencing of DNA fragments at kilobases in length.

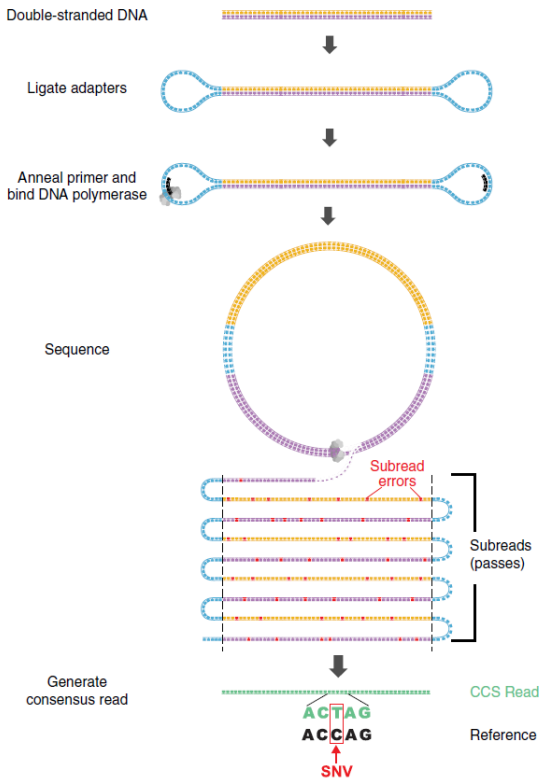


Figure 4: *PacBio HiFi sequencing. Adapted from* Wenger et al., 2019.

CLR sequencing results in one sequence for each template (with an insert size between 25 Kb and 175 Kb) but has lower accuracy than short-read technologies (~90%), requiring higher read coverage for more accurate genotyping. Whereas CCS sequencing (Travers et al., 2010), also known as HiFi sequencing, was developed to improve the accuracy of PacBio long-read sequencing (Wenger et al., 2019). Here, sequencing takes place similar to CLR but in a circular fashion, with several rounds of sequencing of the same DNA fragment. This creates raw HiFi reads which can be bioinformatically merged to call the consensus at each nucleotide base, improving the accuracy of the basecalled reads. HiFi reads have an accuracy of around 99.95%, but with a maximum insert size of 25 Kb (Hon et al., 2020).

The other commonly used long-read technology is Oxford Nanopore Technologies (ONT) sequencing (**Figure 5**). As the name suggests, this makes use of specially designed voltage-biased nanoscale membrane pores which create an ionic current, that draws single-stranded DNA from the *cis* chamber into the *trans* chamber (Deamer et al., 2016). Attached to the nanopore on the *cis* side is a motor protein that controls DNA translocation. The translocated nucleotide bases have different mass and electrical fields, and the change in the ionic current through the nanopore is measured which reveals each individual nucleotide base.

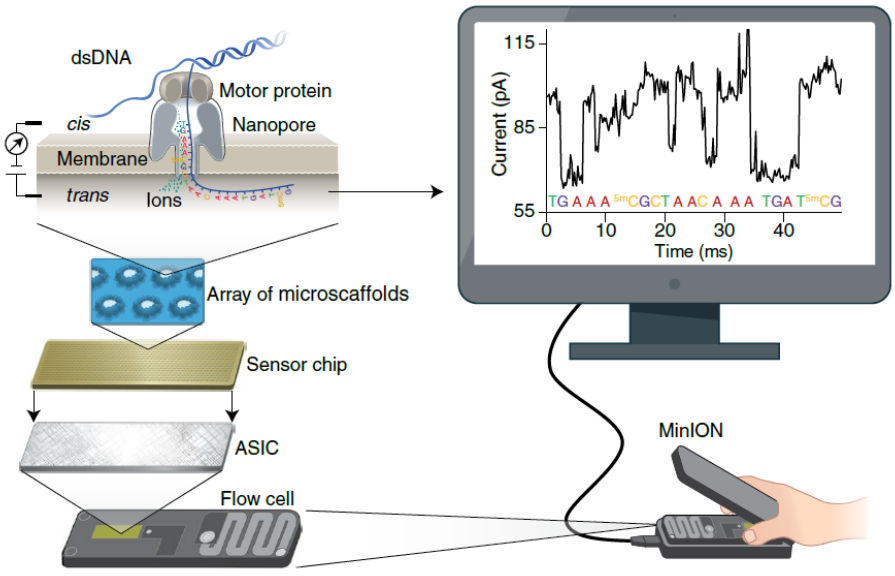


Figure 5: *Oxford Nanopore Technologies (ONT) sequencing using a MinION device. Adapted from* Y. Wang et al., 2021.

Genomic libraries are prepared by repairing DNA ends of DNA fragments and attaching de-A tails, and adapters are then ligated onto the repaired ends. The prepared libraries

are added to ONT flow cells, which contain the nanopores. ONT can sequence DNA reads up to several megabases but has lower accuracy than PacBio HiFi sequencing, with the latest products now reaching greater than 99% accuracy.

Short- and long-read technologies are enabling the assembly of highly accurate genome assemblies, the genotyping of single-nucleotide polymorphisms (SNPs) and structural variants (SVs), and the sequencing of hundreds of individuals or species (Lewin et al., 2018).

1.2.4. Chromatin conformation capture (Hi-C)

Short- and long-read sequencing have revolutionised the study of whole genomes. But studies requiring complete genome assemblies have also required techniques to establish the spatial organisation of chromosomes within the nucleus. One chromatin conformation capture technique (3C) termed Hi-C (Lieberman-Aiden et al., 2009) has enabled the assembly of large mammalian genomes (**Figure 6**). Moreover, the technique has been fundamental in uncovering chromosome territories, compartments, TADs, and DNA loops (Jerkovic & Cavalli, 2021; Zheng & Xie, 2019).

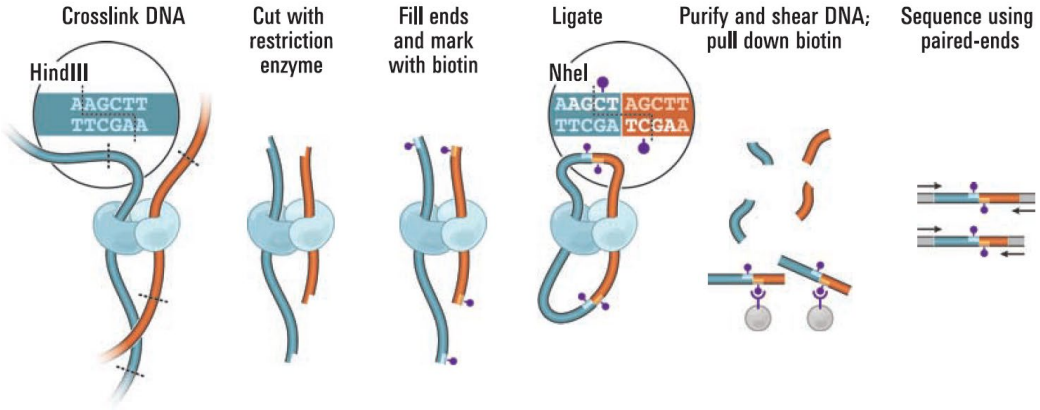


Figure 6: *Hi-C technique. Adapted from* (Lieberman-Aiden et al., 2009).

Hi-C is a method which makes use of paired-end short-read sequencing technologies. DNA within cells is firstly fixed by adding formaldehyde, causing cross-linking to occur between spatially adjacent chromatin fragments (Lieberman-Aiden et al., 2009). The chromatin is then digested with a restriction enzyme and nucleotides are added to the sticky ends, one of which contains biotin. DNA is ligated, purified, and sheared. Streptavidin beads are used to isolate the marked biotin, and the DNA fragment is sequenced using Illumina paired end sequencing, providing information on chromosome interactions in the nucleus.

1.3. Assembling a eukaryotic genome

Advances in sequencing technology, from sequencing small DNA fragments using the Sanger platform, to DNA sequences tens to hundreds of kilobases in length with long-reads, have enabled the improvement in the quality of genome assemblies. Quality here can be defined as the length of assembled sequences of DNA (contiguity) or the presence of orthologous genes (completeness). Metrics such as N50 are used to assess the contiguity of genome assemblies, where contigs are ordered and summed from largest to smallest, and the length of the smallest contig that makes up the top 50% of the assembly is the N50, while the L50 is the number contigs this equals to. The final aim of creating a contiguous and complete genome assembly is for sequences to be long enough to span entire chromosomes, without gaps or missing data, with some assemblies now spanning from telomere to telomere (T2T; e.g., Nurk et al., 2022).

The first step in genome assembly is to use bioinformatic programs to assemble sequencing reads into longer sequences called contigs, creating a contig-level genome assembly. Typically, this is done using a *de novo* genome assembly approach, without the use of a reference genome. Short-read sequencing was traditionally used for this due to higher accuracy than traditional long-reads. However, the improvement in long-read sequencing has made these longer and higher quality reads more useful in assembling more contiguous assemblies (Rhie et al., 2021). For example, for the 99 mammalian genomes in GenBank in 2015 using short-reads the average contig N50 was 41 Kb, whereas in 2020 out of 800 genomes using long-reads the average N50 was 5 Mb (Logsdon et al., 2020).

Bioinformatic programs use a range of methods to assemble sequencing reads. One of the most common types is graph-based assembly, using either a De Bruijn or string graph. Here, graphs are made up of vertices (nodes or points) and edges (links or lines between vertices). De Bruijn graphs are typically used for short-reads (e.g., the SOAPdenovo2 program; Luo et al., 2012), where reads are broken into smaller sequences of a specific size (K) called k-mers. K-mer prefixes and suffixes are represented as nodes on the graph, whilst the k-mers that have the particular prefix and suffix are the edges (**Figure 7**; Compeau et al., 2011). If repeats in the genome exist, then k-mers will have a multiplicity of *m*, and thereby connecting a prefix and suffix with *m* number of edges, denoting the number of times the k-mer is found in the genome.

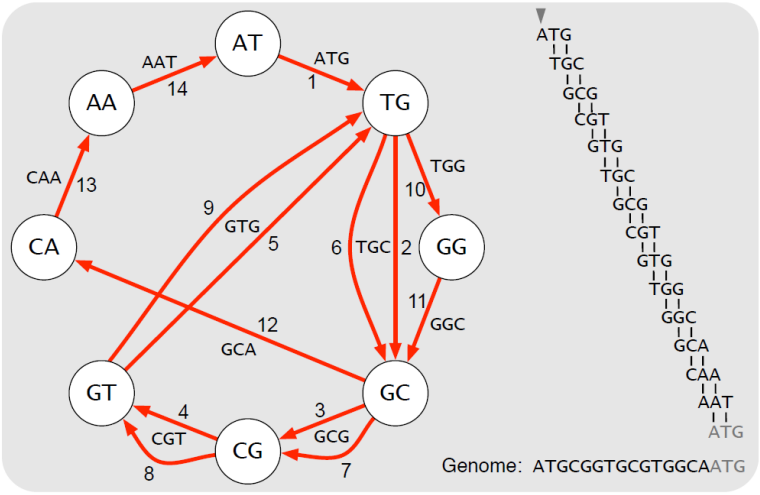


Figure 7: *An example of a De Bruijn graph of an assembled genome (ATGCGGTGCGTGGCA) with repeats denoted by more than one edge between vertices. Adapted from Compeau et al., 2011.*

This creates an Eulerian cycle, where a path is created through the graph that visits every edge once and returns back to the starting vertices, creating the genome sequence (Compeau et al., 2011). Whilst one path through the graph may not exist, multiple paths can be found throughout the graph to complete an Eulerian cycle and complete the genome sequence. These paths represent different contig sequences. In De Bruijn assembly programs, errors create bulges in the graph and require error-correcting approaches (**Figure 7**).

Several programs have been developed for the latest long-read sequencing reads which make use of the second type of graph, string graphs, and directly use long-reads rather than k-mers (Myers, 2005). A widely used program to assemble PacBio HiFi reads is Hifiasm (Cheng et al., 2021). An all-versus-all read overlap alignment is performed and is followed by haplotype-aware error correction (**Figure 8**). During correction, target reads containing informative genomic positions are defined as those with two types of bases (heterozygous alleles) and are supported by three or more reads. These represent inconsistent reads that are not identical across all overlap positions due to haplotype differences. After several rounds of error correction, Hifiasm then carries out another overlap alignment step and builds the string graph.

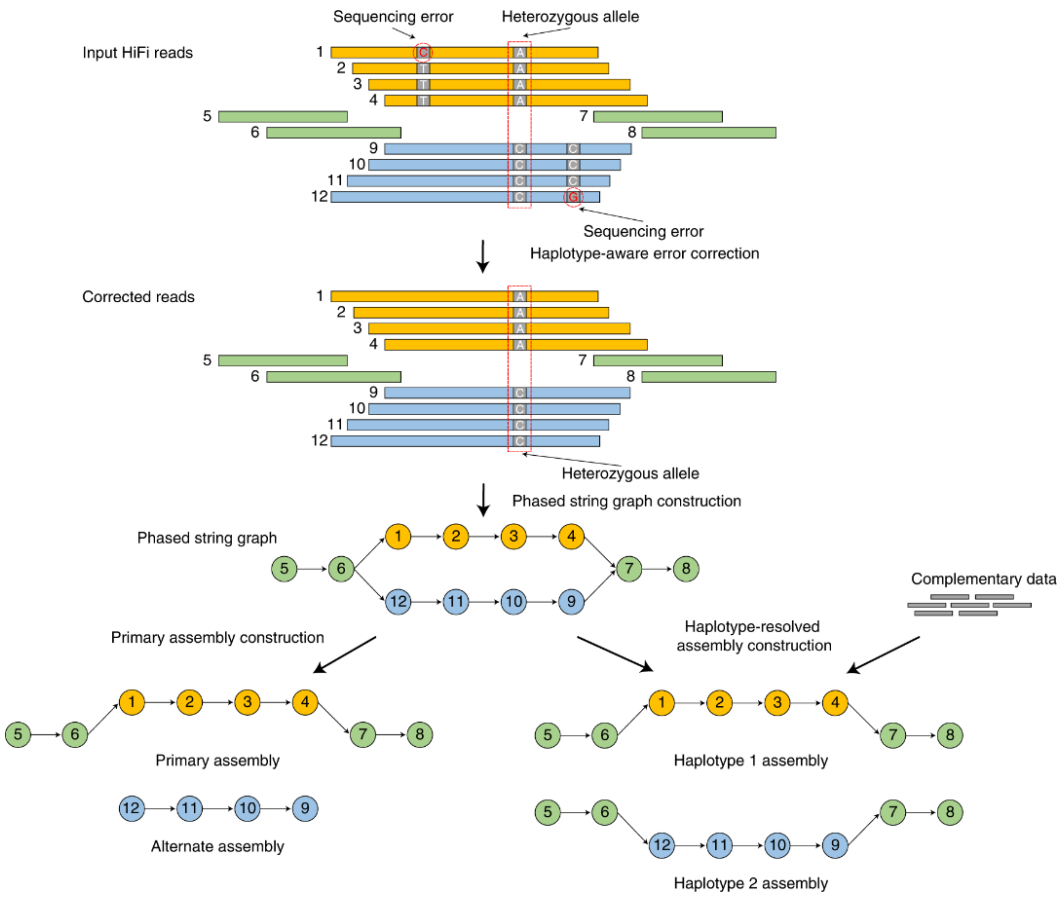


Figure 8: *Assembly of HiFi reads with Hifiasm showing alignment, haplotype-aware error correction, and the phased string graphs (from* Cheng et al., 2021*).*

Each vertex of the string graph is an orientated read, whilst the edge is a consistent overlap (**Figure 8**). Heterozygous alleles are represented as a "bubble" in the graph. The program randomly selects one side of each bubble if no additional data is present, and this results in a primary assembly. This primary assembly may contain more than one haplotig (a contig of one haplotype) and therefore purging is carried out to remove these duplicated haplotigs, and these are added to the alternative assembly. This results in a pseudohaplotype assembly. If parental data is present, then a fully phased genome can be completed, separately containing the maternal and paternal assemblies.

The next step in genome assembly is scaffolding, where contigs are joined into larger sequences with gaps called scaffolds to create a scaffold-level genome. If scaffolds span entire chromosomes, then this results in a chromosome-level genome assembly. Scaffolding programs often make use of Hi-C data, providing knowledge of the spatial proximity of loci within the nucleus and therefore which contigs should be scaffolded together. For example, in the human genome there are around eight times more contacts between loci at 10 Kb distance than those at 100 Kb (Dudchenko et al., 2017). One such

scaffolding program is 3D-DNA (Dudchenko et al., 2017) which firstly corrects misjoins through the identification of positions where there is a change in the pattern of long-range contacts. Contigs are anchored, ordered, and orientated, by estimating their proximity to one another, in an iterative manner. Contigs are then merged into scaffolds if they have strong sequence homology and strong similarity in their contact patterns. This produces the final scaffold- or chromosome-level genome assembly, and further improvements or corrections can be carried out through genome curation.

The advances in DNA sequencing and genome assembly technologies have enabled large numbers of species to be sequenced. Large consortia now exist with the aim of sequencing as many species as possible. These include the Earth BioGenome Project (Lewin et al., 2018), the Vertebrate Genomes Project (VGP; Rhie et al., 2021), the Darwin Tree of Life Project (Blaxter et al., 2022), and the Ruminant T2T Consortium (Kalbfleisch et al., 2024).

1.4. From population genetics to genomics

The technological innovations in DNA sequencing and assembly described in the previous sections have enabled the study of populations of animals from a few individuals to thousands, as well as the move from sequencing single markers such as the mitochondrial control region or repeats like microsatellites, to sequencing whole genomes several times over (e.g., WGS). This has resulted in the field of population genetics shifting towards larger datasets and is therefore now often termed as the field of population genomics. Population genomics can be formally defined as “the simultaneous study of numerous loci or genome regions to better understand the roles of evolutionary processes…that influence variation across genomes and populations” (Luikart et al., 2003).

1.4.1. Population Genomics Concepts

Variation in genotypes between individuals and populations can be due to a number of factors. DNA mutations are random in an evolutionary context but occur non-randomly across genomes. For example, transition mutations (A ↔ G or C ↔ T) are more likely to occur than transversion mutations (A/G ↔ C/T). Also, different regions of the genome have different mutation rates (Hodgkinson & Eyre-Walker, 2011). Mutations can have several consequences, either they are (i) advantageous and under positive selection, where a mutation is becoming fixed in a population, (ii) deleterious where negative (or

purifying) selection removes mutations that are functionally important such as in genes and the mutation is conserved, or (iii) the mutation is neutral. Additionally, balancing selection can maintain selected mutations in a population, which may fluctuate, and includes heterozygote advantage, where heterozygote genotypes have a higher fitness than the two homozygote genotypes. Several models have been proposed for molecular evolution. One of these is the neutral theory, which states that the majority of mutations are not because of selection but actually the random fixation of neutral or nearly neutral mutations (Kimura, 1983).

Mutations can become fixed in populations due to genetic drift, which is the stochastic change in allele frequencies from one generation to the next because of the finite nature of populations and the differing number of offspring in each generation. One model of genetic drift is the Wright-Fisher model, which assumes an idealised population with random mating, discrete generations, and where individuals are diploid. Another approach is coalescent theory, where alleles are considered instead of populations. These alleles are traced back to their ancestry and will eventually undergo coalescence or convergence, where they will derive from the same ancestral allele (Charlesworth, 2009). Genetic drift can be measured by the effective population size (N_e), which calculates an idealised constant population that has the same rate of genetic drift as the real population (Charlesworth, 2009). It can be calculated using several models of genetic drift, including the Wright-Fisher model or using coalescent theory, and several factors interact with N_e, including mutation, migration, recombination, and selection.

Population genomics often aims to separate species into subpopulations. Independent (or nearly independent) subpopulations will diverge genetically over time due to changes and differences in allele frequencies, due to mutation, selection, genetic drift, and migration, leading to population structure. All these processes can cause deviations from Hardy Weinberg Equilibrium (HWE) for the overall population due to the non-random mating among subpopulations. These subpopulations could be isolated due to barriers in the movement of alleles between subpopulations (gene flow), which may be geographic, climatic, temporal, or behavioural. The relative levels of genetic differentiation between subpopulations can be measured using the fixation index (F_{ST}) or absolute genetic differentiation can be calculated using Nei's *D* metric (D_{xy}). Genetic differentiation can be affected by different types of selection. Negative and balancing selection will reduce F_{ST}, whilst positive selection or local adaptation will increase F_{ST}. Lastly, the rate of migration (*m*) will also impact population structure, and several models

have been proposed (e.g., the infinite island, isolation, and isolation-with-migration models).

As well as population structure, genomic variation can also be established during meiosis in the form of recombination, where sister chromatids “crossover” and exchange chromatid segments with each other, which may break the association (linkage) between loci on the same chromosome. Linkage disequilibrium (LD) measures the non-random association of alleles in a population. For example, regions with lower recombination will have higher LD between alleles in these regions, because they are commonly inherited together. LD patterns across the genome can vary depending on the mating system, recombination and mutation rates, genetic drift, and population structure (Fox et al., 2019).

1.4.2. Genomic approaches and methods

Population genomic studies firstly utilised reduced-representation methods (e.g., RADseq) and genotype-by-sequencing methods (GBS), but now more commonly utilise the power of whole genome sequencing with short-read data, allowing the sequencing of potentially all sites across the genome. To enable the sampling of many individuals and populations with WGS, researchers sometimes opt to sequence samples at low coverage (< 5X). However, this may result in only one of the two chromosomes being sampled at a specific site, causing errors when genotyping and the inability to call single-nucleotide polymorphisms (Nielsen et al., 2011). In order to avoid these biases, genotype likelihood methods were developed to circumvent directly calling SNPs. They consider the sequencing quality of the nucleotide base at each position (the Phred score) and the number of reads that supports each base and use a probabilistic approach. These likelihoods are calculated on each possible genotype (e.g., there are 10 possible genotypes for diploids; Lou et al., 2021) and allow the calculation of several population genomic metrics.

As well as technological advances in sequencing and genotyping, studies are now beginning to utilise museum collections, carrying out extraction of historical DNA (hDNA) and WGS for use in population genomics. These collections, sometimes from as early as the 1600s, provide a wealth of genomic information on extant and extinct species (Raxworthy & Smith, 2021). This can be valuable when the collection of samples from the wild is difficult to obtain, to increase taxonomic coverage, and to sample extinct species (reviewed in Raxworthy & Smith, 2021). Moreover, genomic studies have

sequenced ancient DNA (aDNA) from thousands to millions of year old organisms (e.g., the woolly mammoth; van der Valk et al., 2021).

Once genomic data has been generated, several population genomic methods require the calculation of allele frequencies. Some of these make use of the allele frequency spectrum or the site frequency spectrum (SFS) for a population (1D-SFS) or for several populations (2D-SFS). An SFS that uses the minor allele frequency (MAF), the frequency of the less common allele, is known as a folded SFS, whereas if information is available on which alleles are derived and which are ancestral, then the derived alleles can be used to create an unfolded SFS.

A common approach in genomic studies is to firstly assess structure within the sampled subpopulations (reviewed in Lou et al., 2021). Principle component analysis (PCA) has been adopted and can be used with individual allele frequencies from genotype likelihoods (e.g., PCAngsd; Meisner & Albrechtsen, 2018) to assess population structure. Another approach is to use model-based clustering to estimate ancestry proportions using population allele frequencies and a given number of discrete ancestral populations (e.g., ADMIXTURE; Alexander et al., 2009). Besides population structure, this also provides information about recent admixture and possible introgression between populations.

To test whether populations are isolated, the program Estimated Effective Migration Surface (EEMS; Petkova et al., 2015) was developed and is based on the stepping stone migration model where migration between subpopulations (or demes) occurs locally and migration rates vary by location. The program creates a triangular grid on a given map, assigns sampled populations to the closest vertex, and then estimates migration rates along every edge, and diversity rates for each vertex, based on genetic dissimilarity.

Individual and population-level statistics can also be calculated from SFS data such as nucleotide diversity (π), heterozygosity, Tajima's D, and F_{ST} (Lou et al., 2021). Recombination can be explored using LD and this is often calculated from genotyping data with the r^2 *statistic* (e.g., ngsLD; Fox et al., 2019) and can be used to measure rates of recombination along the genome.

1.4.3. Empirical studies

The power of population genomics and short-read sequencing has enabled the assessment of a variety of wildlife taxa across the globe. These include delineation of

species such as giraffe, where WGS of 50 individuals confirmed four separately evolving species, with different levels of heterozygosity, limited gene flow, and different demographic histories (Coimbra et al., 2021). Or confirming the structure of populations or subspecies, such as in wildebeest, where 143 WGS samples were used to confirm the two subspecies, the blue and black wildebeest, with signs of historical introgression but no recent admixture between them (Liu et al., 2024). This was in contrast to previous mtDNA studies, therefore showing the power of genomics to resolve population structure.

Admixture and hybridisation have also been studied in several taxa, with many studies finding species have benefitted from adaptive introgression. For example, hybridisation between the snowshoe hare and black-tailed jackrabbit resulted in the introgression of a winter-brown camouflage allele into the snowshoe hare at the time of the last glacial maximum and this may have enabled the species to adapt to a changing climate and expand its range (Jones et al., 2018, 2020). More recently, a study using ancient and modern genomic data for the Iberian lynx found an increase in genomic diversity over time and proposed that this was due to historical introgression with the Eurasian lynx (Lucena-Perez et al., 2024). However, admixture and hybridisation can also have negative impacts, where invasive species have affected the population fitness and local adaptation of native species (Hohenlohe et al., 2021).

Genomic studies have also determined changes in the population sizes of species by calculating the effective population size (N_e). For example, the large ruminant sequencing project published in 2019 used short-read assembled genomes of 44 species and ran the N_e Pairwise Sequentially Markovian Coalescent (PSMC) program (H. Li & Durbin, 2011) to link declines in N_e in ruminant species to human population expansion around 100 to 50 Kya (Chen et al., 2019). The wide variety of open-access programs makes population genomics widely accessible for empirical studies.

1.4.4. Applications of population genomics to conservation

Population genomics can also be utilised to answer questions in conservation biology and aid in the management of species, and this has given rise to the field of conservation genomics. Conservation biology often aims to identify population units, assess population size and connectivity, detect hybridisation, and assess whether populations can adapt to changing environmental conditions (Hohenlohe et al., 2021). Each of these can be assessed and supported by the power of genomics. Understanding biodiversity

at the genomic level is critical and as such is included in the Convention on Biological Diversity, as well as part of the IUCN Red List.

Conservation genomics relies on several measures. Populations of conservation concern are predominantly declining or small in size, and these populations will often have higher levels of inbreeding. This increases the effects of genetic drift. Population bottlenecks result in a loss of genetic diversity, the fixation of alleles resulting in lower heterozygosity, and a change in allele frequencies. Genetic drift in small populations can cause inbreeding depression, a reduction in an individual's reproductive fitness, which further decreases population size. This can further increase the probability of the population becoming extinct and is known as the extinction vortex. The increase in homozygosity due to inbreeding also increases the likelihood of recessive deleterious alleles.

These negative impacts on small populations depend on N_e rather than the actual census size of a population. Additionally, if subpopulations become fragmented then this can further impact gene flow, reducing genetic diversity and increasing inbreeding depression. As genetic drift is often the dominant force changing allele frequencies in small populations, selection is less effective. These populations are therefore less able to adapt to changing environmental conditions, increasing their risk of extinction.

Applied studies in conservation genomics are able to delineate species and populations, for example organising populations that have substantial reproductive isolation and adaptive differences into groups known as evolutionary significant units (ESUs). In Cabrera voles on the Iberian Peninsula, genomic data improved the resolution in population structure, and using neutral and adaptive variants delineated four ESUs (Barbosa et al., 2018).

Studies have also genomically monitored diversity and levels of inbreeding in small populations of conservation concern. For the critically endangered Sumatran rhinoceros, a species with less than 100 individuals, a genomic study made use of historical and modern samples to assess the mammal (von Seth et al., 2021). Using historical DNA, they found that a population that had suffered local extinction had high levels of inbreeding. Whereas the extant populations had lower levels of inbreeding but high mutational loads and therefore may be subject to inbreeding depression in the future. However, interactions between population declines and deleterious impacts are complex, as discovered in a genomic study of kakapo in New Zealand (Dussex et al.,

2021). They found that a small island population had surprisingly reduced mutational load compared to the larger mainland population. This may have been due to genetic drift and the purging of deleterious mutations, through an increase in inbreeding and purifying selection.

Anthropogenic climate change is also starting to impact wildlife, with some species now shifting their range to maintain their particular niche. However, some species who are unable to track this change rely on genetic variation to be able to adapt to and persist in their changing environment (Brauer et al., 2023). A study on a species of rainbowfish, sampling both generalists and narrow range endemics, found that hybrid populations had a reduced genomic vulnerability to the projected change in climate compared with the narrow range endemics. This could suggest adaptive introgression in hybrids and supports the need to further understand the benefits of hybrids in conservation.

Lastly, ex situ management can be used to support the conservation of wild populations and can benefit from genomic approaches. This includes assessing captive breeding programs for genetic diversity and levels of inbreeding. The scimitar-horned oryx went extinct in the wild in 2000 and a genomic assessment found that the proportion of the genome in runs of homozygosity (ROH) was lower in the managed populations than the unmanaged populations, and the unmanaged populations additionally had higher inbreeding coefficients (Humble et al., 2023), demonstrating the importance of careful management.

Genomic rescue can also be utilised in conservation, where the genomes of individuals are sequenced, and selected individuals are introduced into a population to increase gene flow, improve the fitness of a population, and decrease its probability of extinction (Bell et al., 2019). This was notably carried out for the Florida panther in the USA where individuals were translocated from Texas using genetic information, resulting in an increase in heterozygosity and fitness (Johnson et al., 2010). But concerns around outbreeding depression due to genomic rescue have limited its use in conservation management (Bell et al., 2019) and more genomic studies are needed.

1.5. Chromosome evolution and its role in species divergence

Population genomics studies have predominantly focused on SNPs; however, new research has found that structural variants (SVs) and chromosome rearrangements (CRs) can also have an impact on population diversification and have been linked to speciation (Mérot et al., 2020). CRs have created the diversity of chromosome numbers

(karyotypes) seen across species, and in mammals this ranges from 2*n* = 6 to 2*n* = 102 (Ruiz-Herrera et al., 2012). This genomic reorganisation in mammals has often occurred rapidly and during or after cladogenesis, suggesting that it may play a role in evolution, including adaptation and speciation (reviewed in Dobigny et al., 2017). Moreover, these variants can also be useful for conservation genomics (Wold et al., 2021).

1.5.1. Types of CRs and SVs

While the terms CR and SV are used interchangeably in the literature, here we define CRs as large changes (several Mbs) in chromosome structure, such as fusions, fissions, large inversions, large translocations, changes in centromere position, and polyploidy (**Figure 9**). Whereas SVs are defined as larger than 50 bp but less than several Mbs in length. These can include small inversions, small translocations, and copy number variants (CNVs) like insertions, deletions, and duplications (**Figure 9**).

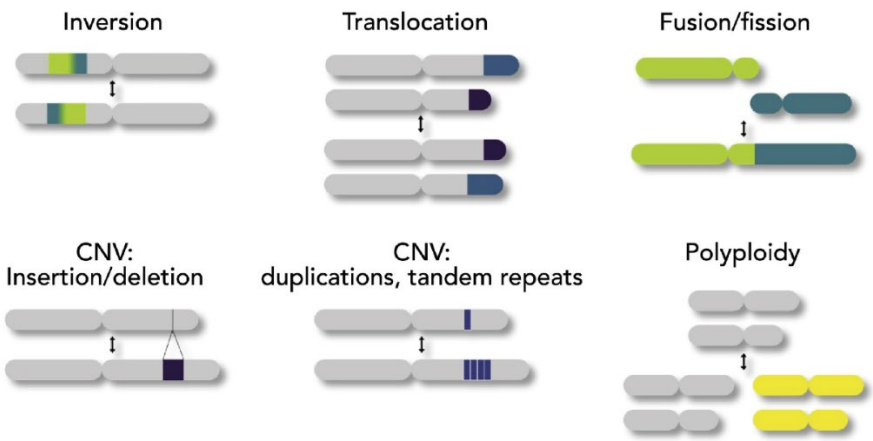


Figure 9: *Types of chromosome rearrangements (CRs) and structural variants (SVs). Adapted from* Mérot et al., 2020.

CRs have been naturally found across a variety of taxa (reviewed in Dobigny et al., 2017). These include interchromosomal rearrangements such as fusions, which can be either tandem or centric. Tandem fusions occur between the centromere of one chromosome and the telomere of another, or the fusion of two telomeres. These are highly deleterious and strongly underdominant. Whereas centric, or Robertsonian (Rb), fusions occur between the centromeres of two acrocentric chromosomes. These are likely the most common CR within mammals and can be polymorphic within species. They may be underdominant and suppress recombination in the heterozygous form. Rb fusions will eliminate or inactivate the telomeres of chromosomes, and several models have been

proposed. These include the inactivation of telomeres, breakage within centromeric satellite sequences, and telomere shortening (Sánchez-Guillén et al., 2015).

For example, the well-studied Rb system in western house mice in Barcelona is surrounded by mice with a standard karyotype of $2n = 40$, but populations in Barcelona show diploid numbers ranging from 27 to 39 due to polymorphisms of up to seven Rb chromosomes (Medarde et al., 2012). These mice were found to have shorter telomeres, with telomere length of the q-arm negatively correlated with diploid number and this supported the mechanism of telomere shortening in the origin of Rb fusions in this species (Sánchez-Guillén et al., 2015).

The opposite of a fusion is a fission, where a chromosome breaks into two functional chromosomes. This however requires the centromeres and telomeres in both of the new chromosomes to be functional and few studies have found polymorphic fissions. Another type of rearrangement is a translocation, the non-homologous exchange of chromosome segments or whole chromosome arms. These have been found to contribute to the differentiation of karyotypes in some mammals (for example in cattle and pigs) and there is some evidence of polymorphisms, for example in gibbons (Dobigny et al., 2017).

Intrachromosomal rearrangements include inversions which can rearrange the contents of chromosomes, and they can either be paracentric (not involving the centromere) or pericentric (occurring across the centromere). Some inversions are heterozygous, but this may cause unbalanced gamete production and lead to lower reproductive fitness, and so have low underdominance. However, some mammals have synaptic adjustments to prevent unbalance gametes and reproductive impairment, which may be neutral or close to neutral. Inversions usually result in no loss or gain of DNA and at the DNA-level are often marked at their breakpoints by identical inverted repeats, but they can disrupt genes and gene expression (Giner-Delgado et al., 2019). Larger inversions (from 100 Kb to several Mbs) are often associated with adaptive phenotypes and the maintenance of differentiation (Mérot et al., 2020).

The location of the centromere can also change, known as centromere repositioning. Here, the original centromere becomes non-functional and a neocentromere develops. This occurs epigenetically by centromere-specific protein A (CENP-A) nucleosomes, and these may “slide” along chromosomes (Purgato et al., 2015).

Lastly, smaller SVs such as CNVs (insertions, deletions, and duplications) cause variation in the number of copies of a particular region of the genome, and which modifies gene dosage and gene expression (Gamazon & Stranger, 2015). They may also contribute to local adaptation, for example in lobster where they have been proposed to contribute to the species adaptation to sea temperatures (Dorant et al., 2020).

1.5.2. Origins and models of chromosome evolution and speciation

CRs are generally caused by the incorrect repair of double-strand breaks (DSBs) which then become fixed in the germline. This occurs when two incorrect DSBs are joined together (Branco & Pombo, 2006) or through non-allelic homologous recombination (NAHR; I. Schubert & Lysak, 2011), where recombination occurs between two lengths of DNA that have high sequence similarity but are not alleles, and these are usually repetitive in nature. DSBs have been associated with segmental duplications (SDs; Carbone et al., 2014), tandem repeats (TRs; Farré et al., 2011), and transposable elements (TEs; Christmas et al., 2019).

Several models have been proposed in chromosomal evolution. One of these is the "Intergenic Breakage Model" which states that evolutionary rearrangement breakages occur uniformly across the genome but are deleterious in genes and regulatory regions (Peng et al., 2006). An advancement on this model is the "Integrative Breakage Model" which states that regions of the genome involved in rearrangements interact physically within the nucleus during the formation of the germ line, are found in open chromatin regions where DNA is more accessible, and rearrangements will become fixed in regions that do not contain essential genes or are involved in essential gene expression (Farré et al., 2015). A study in the mice germline showed that evolutionary breakpoint regions (EBRs), regions involved in structural evolutionary changes that break homologous synteny blocks (HSBs), are located in open chromatin states during post-meiotic stages of spermatogenesis (Álvarez-González, Burden, et al., 2022). Furthermore, DSBs were found to be associated with sites of DNA damage in post-meiotic cells. This supports the Integrative Breakage Model and the importance of chromatin in evolutionary reshuffling within the germline.

Chromosome evolution may play a major role in speciation and several models have been proposed. These include the hybrid dysfunction model where hybrids for a rearrangement have lower fertility and underdominance, and so populations of each rearrangement will become reproductively isolated (White, 1978). Another model

proposes that recombination is suppressed in rearranged regions, leading to reduced gene flow and divergence and ultimately reproductive isolation (Rieseberg, 2001). Finally, rearrangements may also be associated with different levels of gene expression through the disruption of genes, regulatory pathways, and gene dosage (Harewood & Fraser, 2014).

1.5.3. Detection and applications of CRs and SVs

Large CRs have been studied with cytogenetic techniques for decades in a wide range of taxa. Initially this was carried out by karyotyping, where metaphase chromosomes are stained and visualised by microscopy, showing the number and structure of chromosomes. Later, techniques such as fluorescence *in situ* hybridisation (FISH) were developed which use fluorescent probes that hybridise to specific locations on chromosomes, enabling the comparison of homology between species. Cytogenetics enabled the study of fusions, translocations, and large inversions.

But smaller SVs have been more difficult to resolve with cytogenetic techniques. DNA sequencing has enabled further SVs to be detected. Firstly, putative SVs can be detected by indirect methods. These include scanning chromosomes for large regions of high genomic differentiation (F_{ST}) and low recombination, sometimes termed haploblocks, but these can also be caused by selective sweeps or introgression (Mérot et al., 2020). For example, in seaweed flies several putative inversions were detected using a combination of local PCA, F_{ST}, and LD (Mérot et al., 2021).

Second, direct approaches are also used for the detection of SVs through the development of bioinformatic programs that inspect the alignment of short- and long-reads to a reference genome assembly. These programs look for the overlap, orientation, splitting, coverage, and insert sizes of reads (Mérot et al., 2020). Advancements in long-read sequencing technologies, such as PacBio and ONT, have improved the detection of CRs and SVs due to the increased read size and their ability to sequence across repetitive regions. This has enabled the identification of breakpoints in repetitive regions.

The availability of high-quality genome assemblies for several closely related species (for comparative genomics) or several genomes of the same species (for pangenomics) means that whole genome alignments can be carried out to detect CRs and SVs of all sizes. Overall, a combination of indirect, direct, and comparative approaches is useful in validating and understanding the effects of CRs and SVs.

Lastly, the study of CRs and SVs can be applied to conservation biology and management, expanding on the commonly studied SNPs which are now widely used in conservation genomics. SVs can influence hybridisation and introgression, population structure, local adaptation, and speciation (reviewed in Wold et al., 2021), and so have importance in species delineation, maintaining genetic diversity, and adaptation. Further studies related to CRs/SVs and conservation are needed to support their use.

1.6. Antelopes and the family Bovidae

Antelopes are large herbivorous mammals that belong in the family Bovidae, within the suborder Ruminantia and the order Artiodactyla. The term "antelope" refers to all species within Bovidae, except wild and domesticated cattle (in the genus *Bos*), and most individuals from the subfamily Caprinae (e.g., goats and sheep). The group is therefore not monophyletic, a term used to define a clade that consists of the last common ancestor and all its descendants; instead, antelopes form a polyphyletic group that is not based on phylogeny.

Bovidae is the largest mammalian family and consists of approximately 143 extant species, with around 91 of these antelopes (**Figure 10**). Bovids are distributed across Africa, Asia, Europe, and North America, whilst antelopes are naturally absent from North America. The majority of bovids and antelopes are found in Africa and here they often inhabit savannah ecosystems. The evolution of the rumen in ruminants and the third stomach compartment in most ruminant families, including bovids, has allowed for a herbivorous lifestyle and adaptation to savannah habitats (Clauss & Rössner, 2014).

Figure 10: *Antelopes within the family Bovidae. From left to right, top to bottom: waterbuck (subfamily Reduncinae), dik-dik (subfamily Antilopinae), oryx (subfamily Hippotraginae), blackbuck (subfamily Antilopinae), wildebeest (subfamily Alcelaphinae), and bongo (subfamily Bovinae). Waterbuck photo was taken by Dr Marta Farré Belmonte. All other antelope photos via Unsplash.*

1.6.1. Phylogeny

Ruminants emerged from a radiation of artiodactyls during the Eocene period (56 to 33.9 MYA), and then subsequently expanded into 6 extant families: Antilocapridae (pronghorn), Bovidae (cattle, sheep, goats, and antelopes), Cervidae (deer), Giraffidae (giraffe and okapi), Moschidae (musk deer), and Tragulidae (chevrotain; Hernández Fernández & Vrba, 2005). The family Bovidae first appeared in the fossil record around 20 MYA, and rapid diversification of the family occurred during the Miocene period (23.03 to 5.333 MYA), resulting in the diversity of extant species seen today.

The phylogenetic placement of Bovidae within suborder Ruminantia, and the phylogeny between bovid subfamilies and genera, has long been controversial. Studies have included fossil, morphological, and molecular datasets; however, many taxa are still unresolved. In morphological studies, the intermittent radiations during evolutionary time, as well as homoplasy in morphological characters due to convergent evolution, resulted in a lack of available synapomorphic characters (Calamari, 2021; Hernández Fernández & Vrba, 2005).

Molecular data has also provided further controversies in the placement of some clades due to incomplete lineage sorting (ILS) and short internal branches within the ruminant radiation (Hernández Fernández & Vrba, 2005). Recently, Bovidae has been confirmed to be a sister family with Moschidae using whole genome data (**Figure 11**; Chen et al., 2019), supporting some previous studies using molecular markers (e.g., Hassanin & Douzery, 2003).

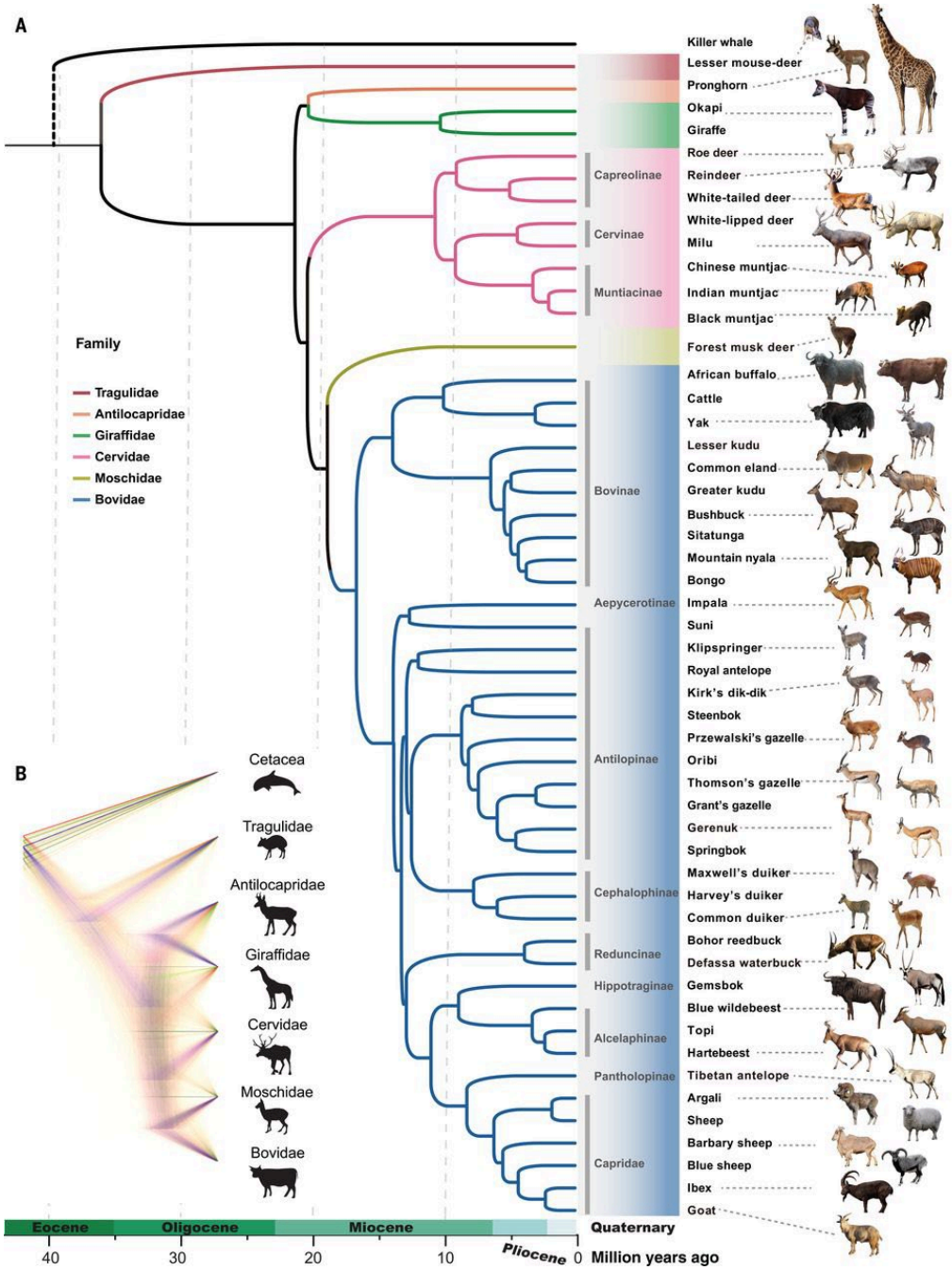


Figure 11: *Phylogeny of Ruminantia from Chen et al. 2019. (A) Maximum-likelihood phylogenetic tree based on whole genome data and fossil calibrations and (B) discordance among 10,000 random window-based gene trees.*

The phylogenetic relationship between subfamilies in Bovidae, or tribes in some phylogenies, paints an even more complicated picture, which is yet to be fully resolved. Genomic data is providing more clarity, resolving some of the relationships between subfamilies (**Figure 11**; Chen et al., 2019). In this genomic study, the subfamily Bovinae diverges from all extant subfamilies, containing genera such as *Bison* (bison), *Bos* (cattle), *Boselaphus* (nilgai), *Bubalus* (water buffalo), *Syncerus* (buffalo), *Taurotragus* (eland), and *Tragelaphus* (e.g., bongo, bushbuck, and sitatunga). The remaining subfamilies have sometimes been grouped into the clade Aegodontia or classified as the subfamily Antilopinae, with tribes instead of subfamilies (Calamari, 2021; Matthee & Robinson, 1999). For simplicity, in this thesis we will refer to the subfamilies, rather than tribes, as some tribes are not supported when using genomic data (e.g., tribe Neotragini as suggested in Matthee & Robinson, 1999).

The Aegodontia clade contains approximately 10 subfamilies, depending on the phylogeny. These include the subfamilies Aepycerotinae (impala), Alcelaphinae (e.g., hartebeest and wildebeest), Antilopinae (e.g., springbok, blackbuck, gazelle, and gerenuk), Caprinae (e.g., goat, sheep, takin, and goral), Cephalophinae (duiker), Hippotraginae (e.g., addax, gemsbok, and oryx), Nesotraginae (suni and bates's antelope), Oreotraginae (klipspringer), Peleinae (rhebok), and Reduncinae (e.g., waterbuck, lechwe, and reedbuck).

1.6.2. Karyotype variability

As well as variation in morphology, bovids also show diversity in the number of chromosomes they possess. Whilst the total amount of genomic DNA and the number of chromosome arms are similar in most species of Bovidae (~3 Gb and FN = 58, respectively), the diploid number of chromosomes ranges from 2*n* = 30 in blackbuck, sitatunga, and steenbok, to 2*n* = 60 in at least 29 taxa (**Figure 12**; Arias-Sardá et al., 2023; Gallagher & Womack, 1992). Within Bovidae, subfamilies range from having small differences in diploid number (e.g., Cephalophinae and Hippotraginae) to being highly variable (e.g., Alcelaphinae, Antilopinae, and Bovinae; **Figure 12**).

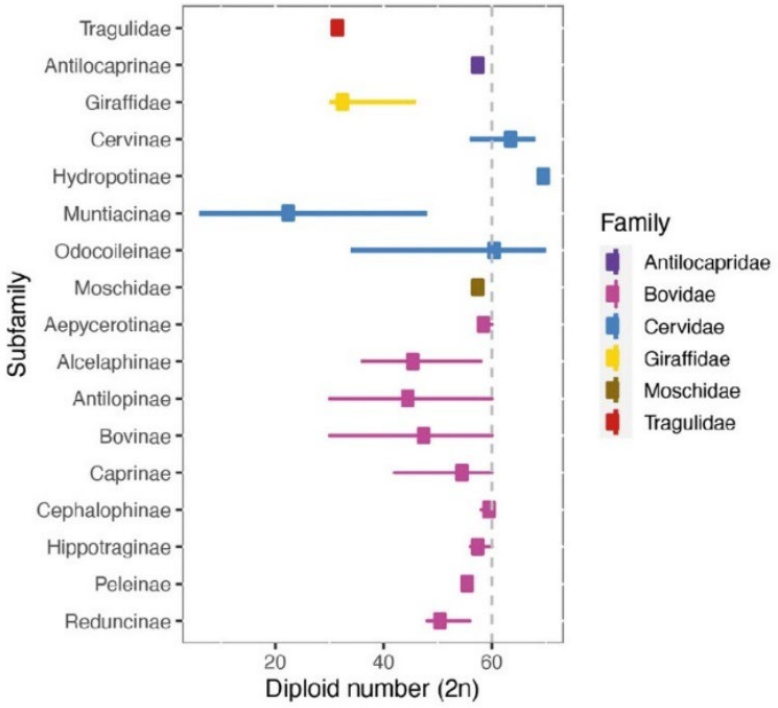


Figure 12: *Variation in the diploid number of chromosomes within and between selected ruminant subfamilies. Adapted from* Arias-Sardá et al., 2023.

Variation in chromosome number is predominantly due to the Rb fusions of acrocentric chromosomes (Gallagher & Womack, 1992), which are a common type of chromosome rearrangement in this clade. Several species also have polymorphic fusions resulting in variable karyotypes and these include the two subspecies of African buffalo (2*n* = 52/54; Buckland & Evans, 1978), impala (2*n* = 58/60; Wallace, 1979), and waterbuck (2*n* = 50-54; S. Kingswood et al., 1998; S. C. Kingswood et al., 2000).

1.6.3. Conservation status

Many antelope species are declining in numbers or have become lost from their historical range, with some now a conservation concern. Out of the 139 bovids listed on the IUCN Red List of Threatened Species (IUCN 2024), 19 of these are categorised as endangered and six critically endangered, and three antelopes have recently gone extinct (the Yemen gazelle, Saudi gazelle, and bluebuck). Threats listed for antelopes on the IUCN Red List include hunting, habitat loss through urban expansion and agriculture, climate change, and war.

Due to successful conservation management some species have recently been downgraded on the IUCN Red List. One example is the scimitar-horned oryx which was

once extinct in the wild and was only found in captive populations (Woodfine & Gilbert, 2016). Ex situ conservation using captive breeding and reintroduction programs have been a success, and populations can now be found in Tunisia, Morocco, and Senegal. The species is now listed in the category Endangered, showing the impact conservation management can have on threatened species.

1.7. Waterbuck (*Kobus ellipsiprymnus*), a model species to study genome and chromosome evolution

In the next three results chapters we focus on one particular antelope, the waterbuck, as it is a perfect model species to study population genomics and the origins and effects of chromosome rearrangements on evolution. Waterbuck are large antelopes found across central and southern Africa and are currently classified into two subspecies; the common waterbuck (*Kobus ellipsiprymnus ellipsiprymnus*) and the defassa waterbuck (*Kobus ellipsiprymnus defassa*). The species inhabits scrubland and savannah and requires a close locality to water sources such as rivers and lakes. Waterbucks are also listed on the IUCN Red List, with the overall species listed as Least Concern, whilst the defassa subspecies has been upgraded to Near Threatened ("Kobus Ellipsiprymnus: IUCN SSC Antelope Specialist Group," 2016). The species faces a number of threats, as do most antelopes, and has been lost from some of its former range.

1.7.1. Cytogenetics

The first studies exploring the genomes of waterbucks used cytogenetics to study their karyotypes. The species was found to have a variable karyotype ranging from 2*n* = 50 to 2*n* = 52 in the common subspecies, and 2*n* = 53 or 2*n* = 54 in the defassa subspecies, due to polymorphic Robertsonian fusions of acrocentric chromosomes (**Figure 13**; S. Kingswood et al., 1998; S. C. Kingswood et al., 2000). Specifically, these studies found that Rb fusions occurred between chromosomes syntenic to cattle chromosomes BTA6;18 and BTA7;11. In the common subspecies, the BTA6;18 fusion is fixed and the BTA7;11 is polymorphic, with individuals having either no fusion, or being heterozygous or homozygous for the fusion. Whilst in the defassa, the BTA6;18 is polymorphic (although only the wild type and heterozygous fusions were found) and the BTA7;11 fusion was not present.

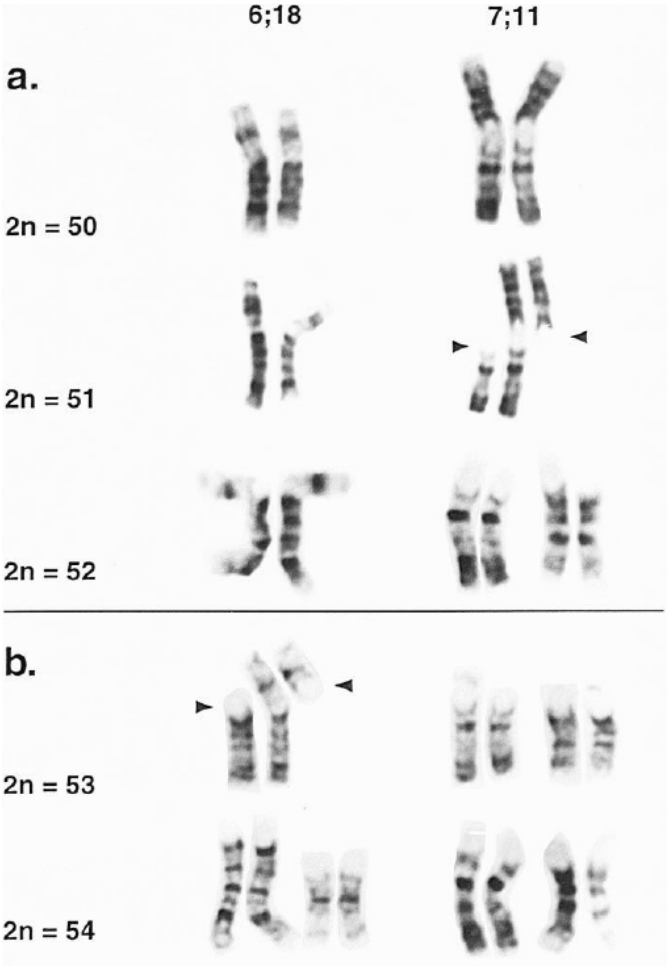


Figure 13: *Polymorphic Robertsonian fusions in (a) common and (b) defassa waterbuck identified using karyotyping. Adapted from* S. Kingswood et al., 1998.

As well as these polymorphic Rb fusions forming submetacentric chromosomes, waterbucks also have three submetacentric chromosomes due to ancestral Rb fusions which are now fixed in the species and are syntenic to cattle chromosomes BTA1;19, BTA2;25, and BTA5;17 (S. Kingswood et al., 1998; S. C. Kingswood et al., 2000). These studies also found a large submetacentric X chromosome in waterbuck, whilst the small Y chromosome is acrocentric in the common and submetacentric in the defassa subspecies. This was speculated to be due to a pericentric inversion, but this was inconclusive due to the resolution of karyotyping. Also, when using the staining technique of C-banding, the pericentric regions of autosomes were much larger than those of the fused metacentric chromosomes.

1.7.2. Population genetics

Studies then began to perform population genetic analysis of waterbuck using individual genetic markers. Mitochondrial and microsatellite markers were sequenced in 186

waterbuck across some of the species' distribution in Africa, mostly in eastern and southern Africa, along the contact zone between the two subspecies (**Figure 14**; Eline D. Lorenzen et al., 2006).

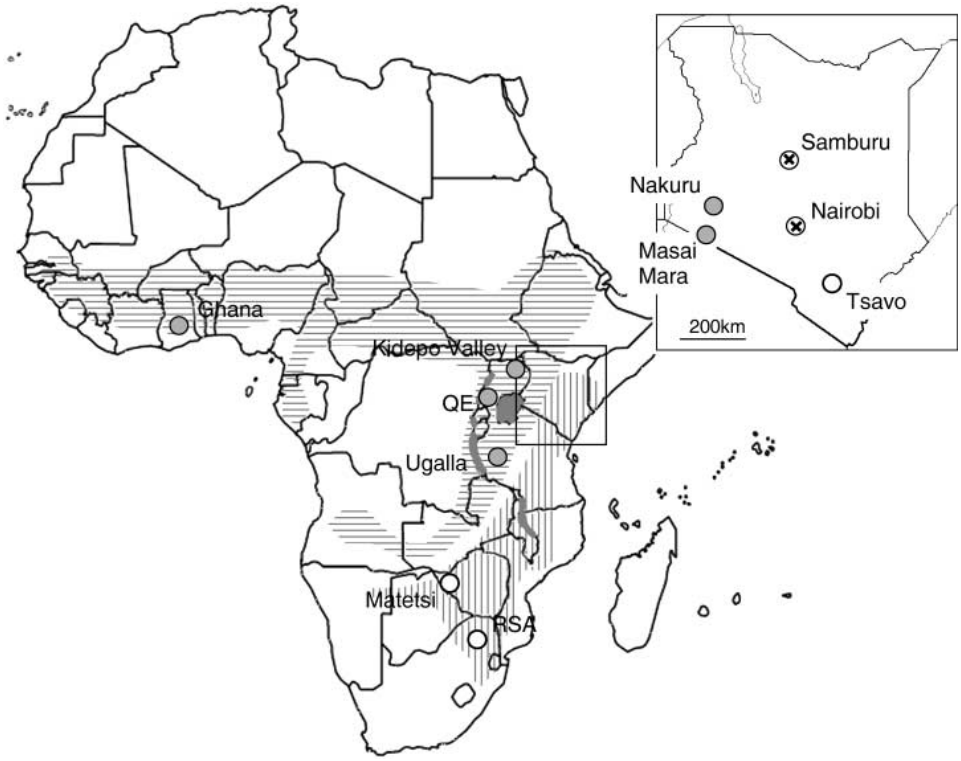


Figure 14: *Sampling map of the 11 waterbuck populations studied in* Eline D. Lorenzen et al., 2006. *The defassa subspecies distribution is denoted by horizontal shading and the common by the vertical shading.*

Using a 478-bp fragment of the mitochondrial control region they found 88 haplotypes across the 11 populations, which were further grouped into four haplogroups, with high genetic differentiation between subspecies. Two of these were defassa, one was common, and one haplogroup had individuals from both subspecies present. However, the mitochondrial control region was unable to individually differentiate the 11 populations. The 14 microsatellite loci were more useful in resolving population structure, grouping subspecies and populations into separate clades, whilst populations in Nairobi (Kenya), where the subspecies distributions overlap, were intermediate on the phylogenetic tree, and therefore suggested to be admixed due to recent hybridisation. This was supported by intermediate rump patterns.

The results of this study suggested that nuclear data may be more useful in resolving population structure and this was later supported by another population genetic study

using the mitochondrial control region that found that it could not accurately be used to determine the origin of captive samples (Ogden et al., 2018). Therefore, an assessment of waterbuck populations using further nuclear markers or at the genomic level was needed to resolve population structure, as well as support conservation efforts.

1.7.3. Population genomics

A recent genomic study sequenced 145 individuals from 10 waterbuck populations using WGS (X. Wang et al., 2024), from most of the population studied in Eline D. Lorenzen et al., 2006 (**Figure 14**). Here, they found population structure both within and between the two subspecies, based on their geographic locality. They also found a restriction of gene flow along the Great Rift Valley that separates the two subspecies. However, they also uncovered evidence of recent and historical admixture along several regions of this barrier, suggesting that the barrier is permeable in places and the species is not yet completely reproductively isolated. Regions of high genomic differentiation were also found across the genome between the two subspecies, with one containing a gene that may be related to differences in fur colouration. These regions did not have a reduction in gene flow or selection against hybrids and so were not deemed to be speciation islands.

However, this study used a scaffold-level genome assembly for the waterbuck (Chen et al., 2019), and so they were unable to place this work in the context of the waterbuck chromosomes. This is essential in a species with polymorphic karyotypes, which may impact speciation. Therefore, a high-quality chromosome-level genome assembly is needed to enable a further assessment of population genomics, but also to further understand chromosome evolution in the waterbuck. As waterbuck have a wide distribution across central and southern Africa, a wider sampling of genomic data is also needed.

1.8. Thesis aims

The main aim of this thesis is to understand the genomics and evolution of antelopes, and to apply this to their conservation. To do so we focus on the waterbuck (*Kobus ellipsiprymnus*), an antelope with polymorphic chromosome fusions within and between two subspecies. Specifically, the thesis is split into three main aims as follows:

1. To sequence and assemble a high-quality chromosome-level genome assembly for the waterbuck and perform genome annotation.

2. To assess population structure, gene flow, and genomic differentiation within and between waterbuck populations using whole genome sequencing data and a chromosome-level reference genome.
3. To detect chromosome evolution and structural variation between and within antelopes in the family Bovidae using chromosome-level reference genomes, and to further understand the role of 3D genome organisation within the nucleus in chromosome evolution.

2. Chromosome-Level Genome Assembly for the Waterbuck

2.1. Introduction

In order to further understand genomics and evolution within and between antelopes at the chromosome-level, highly contiguous and complete genome assemblies are required. The genome of a waterbuck was previously sequenced to scaffold-level using Illumina short-read sequencing in a large ruminant genome sequencing project (Chen et al., 2019). However, due to the technology used, the assembly was highly fragmented, and scaffolds could not be placed accurately into chromosomes. This waterbuck genome assembly totalled 2.895 Gb in length and was made up of 88,848 scaffolds with an N50 of 782.102 Kb. This was one of the least contiguous assemblies in the study, where the scaffold N50s of the ruminant genomes ranged from 11,258 Kb to 15,190,720 Kb. The waterbuck genome is much lower than the VGP project standards (Rhie et al., 2021) and therefore is not very useful for studying population genomics in a chromosome context or for further understanding intra- and inter-species chromosome evolution.

The length of the genome was lower than that estimated using the *C*-values at 3.374 Gb. This may be due to the repetitive nature of bovid genomes and the inability of short reads to span repeats. These repeats are important in bovids with Robertsonian fusions, like the waterbuck, where centromeric satellite DNA (satDNA) repeats are reorganised after the chromosome rearrangement (Escudeiro et al., 2021). Thus, long-read sequencing is required in order to sequence across these repetitive regions. Additionally, whilst the subspecies of the sample was known (defassa) and the location (Nakuru, Kenya), the karyotype had not been established for this sample and could not accurately be identified from the scaffold-level genome assembly. The sample was also from the hybrid zone and may be admixed with the common subspecies.

We therefore aimed to sequence and assemble a high-quality chromosome-level genome assembly for the waterbuck. To do this, we firstly established a cell culture of a captive defassa waterbuck sample, karyotyped the sample to clarify the number of chromosomes and the presence or absence of polymorphic Rb fusions. We then extracted high molecular weight (HMW) DNA from this sample and sequenced the DNA using PacBio HiFi long-reads. Hi-C chromatin conformation capture data was then utilised to scaffold the genome into chromosomes. By using a combination of the gold standard PacBio HiFi long-read sequencing and Hi-C, we aimed to produce a high-quality chromosome-level genome assembly, similar to those produced by the likes of

the VGP (Rhie et al., 2021). Additionally, homology-based genome annotation was used to annotate genes and repeats across the genome. This will not only provide a framework for our study but also will be invaluable for research in genomics and evolution in antelopes.

2.2. Materials and Methods

2.2.1. Primary mammalian cell culture and karyotyping

A fibroblast mammalian cell line was previously established in the laboratory using a tissue sample from a female captive defassa waterbuck (*Kobus ellipsiprymnus defassa*) donated by the Aspinall Foundation (Kent, UK). The cell culture sample was stored long-term in liquid nitrogen before being thawed and re-established. The culture was maintained in a medium containing DMEM, 10% FBS, and 1% Pen-Strep, and was incubated at 37°C with 5% CO_2. Media was refreshed every two or three days. Cells were passaged once they reached approximately 80% confluency. This firstly involved removing the culture media and washing cells with 1% PBS. Approximately 1-2 ml of Trypsin-EDTA was then added to cover cells and incubated at 37°C for several minutes until cells detached from the flask. Culture medium was then immediately added to inactivate the Trypsin and pipetted into additional flasks, with the volume of media and number of flasks depending on the splitting ratio required and passage number of the flasks. The passage number was then increased by one each time the cells were split. Cells were limited to a passage number of ten.

Cells were harvested and karyotyped following the protocol of Howe et al., 2014, which is described below. Colcemid (10 µl/ml) was added to confluent cell culture flasks and incubated at 37°C with 5% CO_2 for 45 min, and cells were then washed with 1x PBS. Approximately 1-2 ml of Trypsin was added to the flasks for 2 min before transferring the contents to a sterile tube. Tubes were centrifuged at 200 x g for 5 min, followed by removal of the supernatant and resuspension in 5 ml of Carnoy's Fixative (3:1 methanol and glacial acetic acid) with vortexing. A further 5 ml of fixative was added without vortexing, centrifuged at 200 x g for 5 min, and the supernatant discarded. This step was repeated once, and we then stored the chromosome preps at 4°C.

For karyotyping, chromosome preps were firstly concentrated by centrifugating for 5 min at 1500 rpm, the supernatant was then decanted until approximately 2 ml remained. A

clean slide was placed over a container containing warm water to cover the slide in steam, before two 10 µl drops of the concentrated chromosome preps were added at each end of the slide. Once the slides were dried, they were immersed in 70% glacial acetic acid for 5 sec and then vertically airdried. A drop of DAPI stain was added to each end of the slide, a coverslip was placed on top, and the slides were placed in the dark for 30 min before being viewed with fluorescent microscopy. Several photographs were taken and the karyotyping software, SmartTypeDemo v3.3.2, was used to visualise, crop, and arrange the chromosomes. Moreover, cells were also harvested during later stages of the cell culture to verify the stability of the karyotype, which was maintained.

2.2.2. DNA extraction of waterbuck cell culture

HMW DNA was extracted from the defassa waterbuck cell line using the QIAGEN Blood and Cell DNA extraction kit following the cell culture protocol. A haemocytometer was firstly used to estimate the number of cells and to avoid blocking the “Genomic Tips”, which required a maximum of 10 M cells/ml. The DNA extraction also included an RNase A treatment before the lysis step to remove any RNA within the sample. DNA was eluted in 10 mM Tris-HCl (pH 8.5) and quantified using both NanoDrop and Qubit.

2.2.3. PacBio long-read HiFi sequencing

Extracted HMW DNA from the cell culture sample was sent to Edinburgh Genomics (Edinburgh, UK), where quality control for DNA quantity and quality using Nanopore and Qubit machines, as well as the measurement of DNA fragment lengths with an Agilent Technologies Genomic DNA ScreenTape TapeStation, was performed. One PacBio SMRTbell library was then prepared and sequenced on two PacBio Sequel IIe SMRT 8M Cells with HiFi mode enabled to approximately 20X coverage of the waterbuck genome.

2.2.4. Contig-level genome assembly

PacBio HiFi data was received from Edinburgh Genomics and a bioinformatic pipeline for genome assembly was created based on the pipeline used in the Vertebrate Genome Project (VGP). The pipeline is described as follows. We firstly converted HiFi read data from BAM into FASTQ format using Samtools v1.6 ‘fastq’ (Danecek et al., 2021), and quality controlled the reads with FASTQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc), MultiQC v1.0.dev0 (Ewels et al., 2016), and Nanoplot v1.32.1 (De Coster et al., 2018). Any reads that contained the PacBio adapters

(ATCTCTCTCAACAACAACAACGGAGGAGGAGGAAAAGAGAGAGAT and ATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGAT) were discarded from the FASTQ files with Cutadapt v1.18 (Martin, 2011), with parameters minimum error rate 0.1%, minimum overlap length 35, minimum length 5000 bp, and maximum length 30000. Genome size, maximum read depth, and the transition coverage between haploid and diploid peaks was estimated using Meryl v1.3 (Rhie et al., 2020), with a k-mer size of 32. This was then profiled using GenomeScope2 v2.0 (Ranallo-Benavidez et al., 2020), a program that uses a nonlinear least-squares optimization to fit a mixture of negative binomial distributions to the data. The contig-level genome was assembled with Hifiasm v0.16.1-r375 (Cheng et al., 2021), with the final genome assembly (Assembly 2) parameters including the purging of all types of haplotigs, a purge maximum of 45, and the removal of 20 bp from both ends of the HiFi reads. We also tested purging only contained haplotigs (Assembly 1) and no purging (Assembly 3).

This outputted a contig-level primary genome assembly which was quality checked with QUAST v5.0.2 (Mikheenko et al., 2018) and Merqury v1.3 (Rhie et al., 2020). The genome was additionally assessed for completeness and contamination with the BlobToolKit v2 pipeline (Challis et al., 2020) using BUSCO v5 (Manni, Berkeley, Seppey, & Zdobnov, 2021; Manni, Berkeley, Seppey, Simão, et al., 2021), BLAST v2.10.0, and Uniprot 2023_01 (Consortium, 2023) databases. We also visualised the assembly graph using BandageNG v2022.09 (Wick et al., 2015; https://www.github.com/asl/BandageNG).

2.2.5. Mitochondrial genome assembly

The mitochondrial genome was assembled from the trimmed HiFi reads with the MitoHiFi v2.2 pipeline (Uliano-Silva et al., 2023), followed by gene annotation and visualisation in Geneius Prime (https://www.geneious.com).

2.2.6. Chromatin conformation capture (Hi-C), Illumina short-read sequencing, and additional Hi-C data

The female waterbuck cell line described above was also prepared for chromatin conformation capture (Hi-C) as follows. Culture medium was discarded, and cells were washed twice with 1x PBS. A total of 11 ml of freshly made 1% formaldehyde in 1x PBS was added to the cell culture flasks, gently mixed, and incubated for 10 min at 4°C with gentle rocking every 2 min. After 10 min, 687 µl of glycine (final concentration of 0.125 M) was added and incubated at RT for 5 min, then incubated at 4°C for 15 min. After incubation, cells were washed with 1x PBS. A total of 2 ml of Trypsin was then added to

detach the cells from the flask and incubated for 8 min at RT. Additionally, cells were manually scraped and collected into sterile tubes. Additional washes of 1x PBS were used to remove all cells from the flask. Cells were pelleted by centrifugation at 3,000 rpm for 5 min, and the supernatant was removed. This step was repeated, and the pellet of fixed cells was frozen at -80°C until the protocol was continued.

The remainder of the protocol was undertaken at the University of Barcelona (Lucía Álvarez González and Aurora Ruiz-Herrera) and described in Álvarez-González et al., 2022 and summarised here. Lysis buffer was added to cells, incubated on ice for 30 min, and centrifuged for 5 min at 1,800 x g. The cell pellets were washed twice with 1 x NEB2 buffer and resuspended in fresh NEB buffer with 10% SDS at RT, followed by incubation at 65°C for 10 min with mixing. A NEB2 buffer containing 10% Triton X-100 was added to the cells and incubated for 30 min at 37°C. Cells were pelleted at 1,800 x g for 5 min at 4°C, then washed twice with 1 x NEB2 buffer. A total of 400 U of Mbol was added and incubated at 37°C overnight with mixing to digest the chromatin, followed by the addition of Proteinase K (10 mg/ml) the next day and a second incubation for 45-60 min at 65°C. A standard phenol:chloroform extraction was then undertaken.

The sample was then centrifuged for 5 min at 1,800 x g and washed twice with 1 x NEB2, followed by resuspension in the reparation mix which contained 1x NEB2 buffer, 0.05 mM dCTP, 0.05 mM dTTP, 0.05 mM, 0.05mM biotin-dATP, and 50U Klenow. This was incubated for 45 min at 37°C, then 10 min at 65°C. The sample was then pelleted at 1,800 x g and resuspended in a ligation buffer containing 1x NEB T4 ligase buffer, 0.83% Triton X-100, 0.1 mg/mL BSA, 5 mL ligase (2000 U/ml), and 963 mL H2O, before incubating for at least 4 hours at 16°C with mixing, centrifuged for 5 min at 1,800 x g, and resuspended in 1 x NEB2 buffer. RNase A (10 mg/ml) was and incubated for 15 min at 37°C. Proteinase K (10 mg/ml) was added to reverse the cross-linking and incubated overnight at 65°C. After the incubation, the sample was left at RT to cool, then purified with phenol:chloroform.

The sample was then sonicated (20 sec time ON, 60 sec OFF, for eight cycles), before checking the DNA fragment sizes with gel electrophoresis. Dynabeads MyOne Streptavidin T1 beads and 2x Binding Buffer (10 mM TrisHCl, 1mM EDTA, 2M NaCl) were added to the sample, and then incubated at RT for 30 min with mixing. The beads were washed twice using a magnetic rack, and the sample resuspended in the end repair mix, containing 1x NEB T4 DNA ligase buffer with 10 mM ATP, 25 mM dNTP mix, 10U/ml NEB T4 PNK, 3U/ml T4 DNA polymerase I, and 5U/ml NEB DNA polymerase I (Klenow).

Following a 30 min incubation at RT, the beads were washed again with 1x Binding Buffer. The beads were resuspended in a dATP attachment master mix containing 1x NEBuffer 2, 0.5mM dATP, and 5U/ml NEB Klenow exo minus, incubated at 37°C for 30 min, and washed with 1 x Binding Buffer. DNA was resuspended in 1x NEB Quick ligation buffer. The DNA library was sent for Illumina sequencing using pair-end reads (2 x 150 bp).

Moreover, we collaborated with DNA Zoo, a consortium that produces and releases genomic assemblies, to obtain Hi-C data from additional waterbuck samples. This included one high coverage Hi-C dataset for a female captive common waterbuck (*Kobus ellipsiprymnus ellipsiprymnus*) with a karyotype of 2*n* = 52, and one low coverage Hi-C dataset from a female captive common waterbuck sample with a karyotype of 2*n* = 51, which was not used in this study. Both samples were prepared from blood samples donated to DNA Zoo.

2.2.7. Genome scaffolding

Hi-C sequencing data from the fibroblast cell line was used to scaffold the contig-level genome assembly. FASTQ files were aligned to the genome assembly with Juicer (Durand et al., 2016) and Hi-C interaction signals were viewed with Juicebox Assembly Tools (JBAT; Dudchenko et al., 2018). However, the Hi-C interactions signals were noisy and unable to be used for genome scaffolding. In order to complete the genome assembly, we collaborated with DNA Zoo who scaffolded the contig-level genome with their common waterbuck Hi-C data (2*n* = 52 high-coverage data) using the program 3D-DNA (Dudchenko et al., 2017). Manual curation was then completed with JBAT. The scaffolded genome was made publicly available, and we manually curated it further to reorder and reorientate chromosomes.

2.2.8. Genome synteny to cattle

The chromosome-level genome assembly was verified using synteny to the cattle reference genome and known cytogenetic studies (i.e., Kingswood et al., 1998). The genome was aligned to cattle (*Bos taurus*; ARS-UCD2.0) using MashMap (Jain et al., 2018). Customised scripts were used to construct syntenic blocks, which were visualised with the R package syntenyPlotteR v1.0.0 (Quigley et al., 2023).

2.2.9. Genome annotation

The chromosome-level genome was annotated for genes and repeats using homology-based methods. Genes were annotated with GeMoMa v1.9 (Keilwagen et al., 2016, 2018)

using both goat (*Capra hircus*; ARS1.2) and cattle (*Bos taurus*; ARS-UCD2.0) as reference. Repeats were annotated with RepeatMasker v2.6.0+ (http://www.repeatmasker.org) using cattle as a reference, and the Dfam_Consensus-20181026 and RepBase-20181026 gene databases.

2.2.10. Circos plot

The final chromosome-level genome assembly, GC content, gene annotation, and repeat annotation were visualised with Circos 0.69.8 (Krzywinski et al., 2009). The percentage of GC content was calculated with SeqKit v2.6.1 fx2tab (Shen et al., 2016) in 100 Kb windows. Repeat content was calculated as the total number of masked bases in each window divided by 100 Kb. Gene density was calculated as the number of genes in each 100 Kb window (with annotations that had several transcripts, or identical coordinates, excluded from the calculation).

2.3. Results

2.3.1. Cell culture and karyotyping

We first established a cell line for the defassa waterbuck from a captive sample donated by the Aspinall Foundation (Kent, UK). Once the cell line was established, and after several passages, we assessed the karyotype at passage P7 to check chromosome morphology and diploid number. The sample had the “standard” or wild-type (wt) karyotype for the defassa waterbuck, with no polymorphic fusions and a diploid number of 2*n* = 54. (**Figure 15**). Three autosomal chromosomes were submetacentric, having a smaller p-arm to q-arm, and 23 chromosomes were acrocentric, with the centromere displayed close to the end of the chromosome. The X chromosome was metacentric, with the p- and q-arms a similar size, and the centromere in the middle of the chromosome. Two homologous metacentric X chromosomes were found, and therefore the sample was confirmed as female (XX).

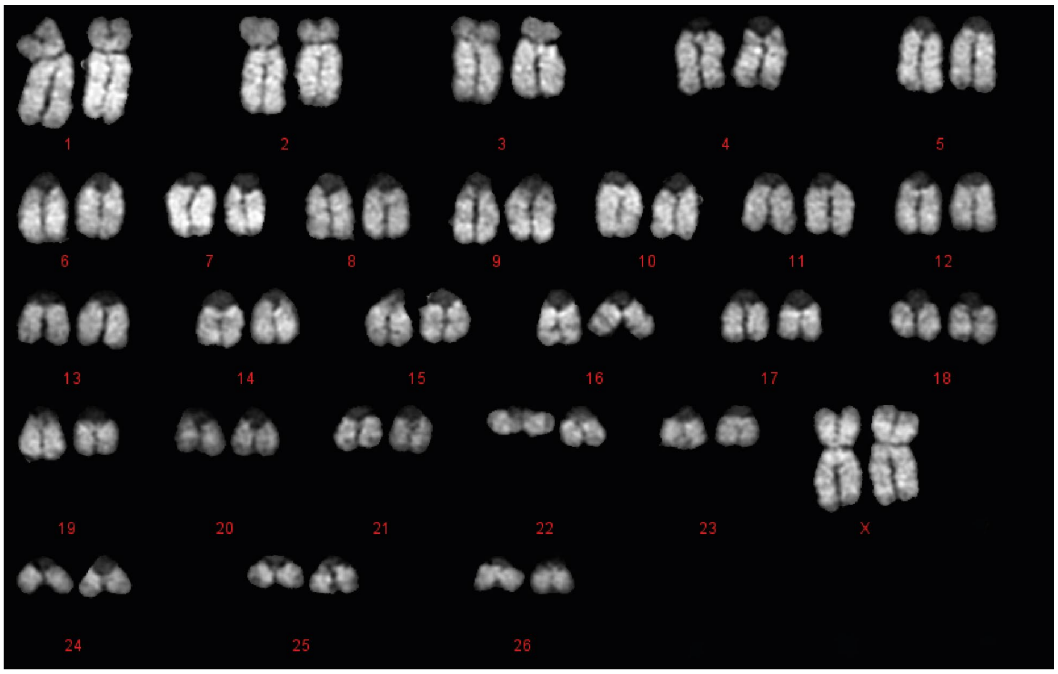


Figure 15: *Karyotype of the defassa waterbuck cell line. Autosomal chromosomes were labelled in size order from 1 to 26, and the X chromosome was labelled separately.*

2.3.2. DNA extraction, PacBio HiFi sequencing, and adapter trimming

After establishing the cell line and karyotyping, we extracted DNA from four confluent T75 flasks at passage P8. This was carried out using the QIAGEN Blood and Cell Culture DNA extraction kit. For the first DNA extraction (DW1) RNase A was not added during the lysis stage and DNA was eluted in 500 µl of TE buffer. This led to a discrepancy between the DNA concentration measured on the NanoDrop (32.500 ng/µl) and the measurement on the Qubit (21.700 ng/µl), due to contamination of the DNA sample with RNA (**Table 1**). In the second DNA extraction (DW2), RNase A was added during the lysis step and DNA was eluted into 380 µl of TE buffer. DNA concentration was estimated as 54.700 ng/µl with NanoDrop and 54.000 ng/µl with Qubit, with similar results indicating a lack of RNA contamination in this DNA sample. The Qubit reading therefore gave a final DNA yield of 20.520 µg in 380 µl. DNA quality was also measured using NanoDrop with the DW2 DNA sample having a 260/280 ratio of 1.810 and a 260/230 ratio of 2.140.

Table 1: *Estimation of DNA quality and quantity of the two extracted DNA samples (DW1 and DW2) from the defassa waterbuck cell line (2n = 54).*

		NanoDrop			Qubit	
DNA Sample	**Volume (µl)**	**260/ 380**	**260/ 230**	**Conc. (ng/µl)**	**Conc. (ng/µl)**	**Yield (µg)**
DW1	500	1.910	2.160	32.500	21.700	10.850
DW2	380	1.810	2.140	54.700	54.00	20.520

The DNA was then sent to Edinburgh Genomics (Edinburgh, UK) where they undertook further QC of the DNA sample (DW2). This included running the DNA sample on an Agilent Technologies TapeStation to assess DNA fragment length, which resulted in an average DNA fragment length of 48.500 Kb and a DNA Integrity Number (DIN) of 9.500 (**Figure 16**). RNA contamination was also quantified in the sample with Qubit and was estimated as being low at 2.900 ng/µl (5.600% of the DNA concentration). A PacBio SMRTbell library was then prepared and sequenced on two PacBio Sequel IIe SMRT 8M Cells in HiFi mode.

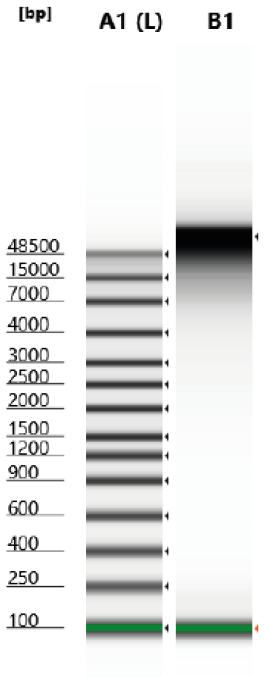


Figure 16: *DNA fragment lengths assessed by Agilent Technologies TapeStation from the defassa waterbuck cell line DNA sample (DW2). DNA Ladder with sizes in bp (A1) and DW2 DNA sample (B1). Green band represents the lower marker (100 bp) used to align the DNA ladder and DNA sample.*

The PacBio sequencing of the defassa waterbuck sample yielded 3.504 million HiFi reads totalling 64.107 Gb of genomic data (~21X coverage of a 3 Gb genome), with a mean read length of 18.294 Kb and read length N50 of 18.890 (**Table 2**). Adapter trimming was used to remove any HiFi sequences containing adapters, and this reduced the total number of bases by 9 Mb to 64.098 Gb. Read lengths were unaffected by the trimming (**Figure 17**).

Table 2: *Quality control of the waterbuck PacBio HiFi reads before and after trimming.*

	Mean Read Length (Kb)	Mean Read Quality (Phred)	Number of Reads	Read Length N50 (Kb)	Total Bases (Gb)
Before	18.294	32.600	3,504,337	18.890	64.107
After	18.294	32.600	3,503,872	18.890	64.098

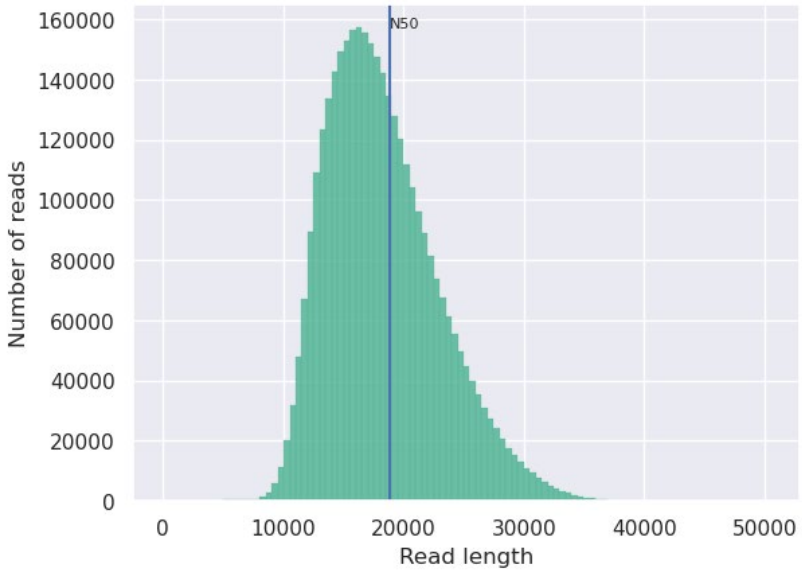


Figure 17: *Number of trimmed PacBio HiFi reads by length.*

2.3.3. Contig-level genome assembly

The trimmed HiFi reads were then used to estimate genome properties using a k-mer approach. The dataset was split into 32-mers and a k-mer spectrum was calculated by creating a histogram of the k-mer distribution. The k-mer spectrum was then used to estimate genome properties. This resulted in an estimated haploid genome size of 3.059-3.063 Gb, percentage of homozygosity between 99.74-99.76% and heterozygosity between 0.240-0.260%, and a read error rate of 0.140% (**Table 3**).

Table 3: *Genome estimation statistics based on the 32-mer histogram of the trimmed waterbuck PacBio HiFi reads.*

	Min	Max
Homozygous (%)	99.740	99.760
Heterozygous (%)	0.240	0.260
Genome Haploid Length (Gb)	3.059	3.063
Genome Repeat Length (Gb)	0.910	0.911
Genome Unique Length (Gb)	2,149	2.151
Model Fit (%)	71.660	99.530
Read Error Rate (%)	0.140	0.140

The program additionally plots the k-mer spectrum and the fitted models (**Figure 18**). Looking at the k-mer spectrum plot, we see the two characteristic peaks of a diploid organism. One main peak, at ~18X coverage, indicating the homozygous regions of the genome, and a smaller, less pronounced peak at ~10X coverage, showing the heterozygous k-mers. This small 'heterozygous peak' agrees with the results in **Table 3**, showing a heterozygosity between 0.240-0.260%.

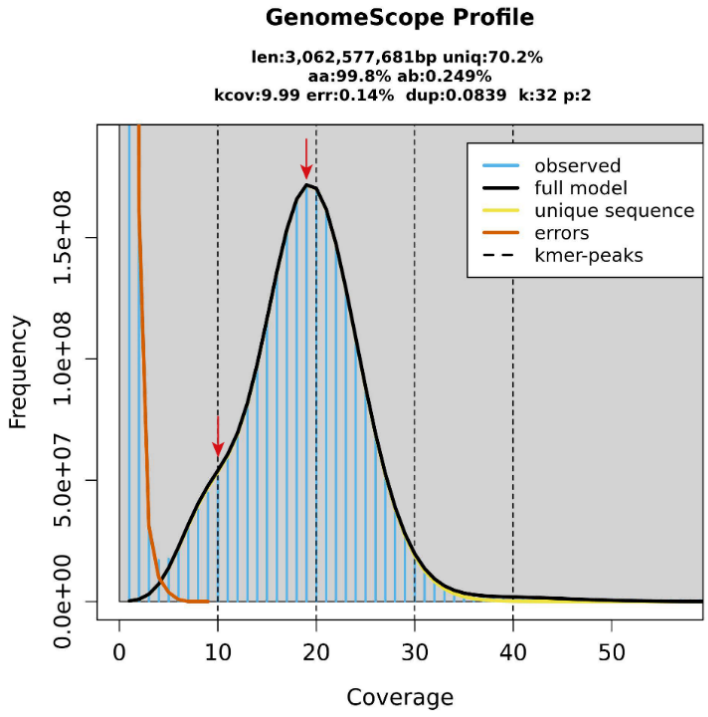


Figure 18: *Estimation of genome properties. Observed k-mer spectrum is represented by blue vertical lines, the full model by the black line, unique sequences by the yellow line, and errors the orange line. Arrows indicate heterozygous and homozygous peaks.*

The HiFi reads were then assembled into a contig-level genome with the program Hifiasm (Cheng et al., 2021). During assembly, programs can sometimes duplicate highly heterozygous regions in diploid genomes, creating a contig for each allele, rather than separate contigs in the primary and secondary assemblies, and therefore increasing the genome assembly size. Purging is done on the primary assembly to remove these duplicate contigs, creating a clean or "purged" assembly. To test the level of purging needed in the assembly process, we ran Hifiasm three times with different purging parameters, which produced three primary assemblies and their respective secondary assemblies. All assemblies were purged to a maximum coverage of 45 (based on 2.500 times coverage of the homozygous peak). The purging level of "Assembly 1" was set to "-l 1" to only purge contained haplotigs and produced a primary assembly of 2,381 contigs totalling 3.600 Gb, a contig N50 of 14.365 Mb, and a contig L50 of 64 (**Table 4**). Primary "Assembly 2" was set to a purge level of "-l 3", which purges all haplotigs, and performed the best, with the lowest number of contigs at 1,072 totalling 3.154 Gb (91 Mb longer than the estimated genome size), the highest contig N50 of 71.171 Mb, and lowest L50 of 18. "Assembly 3" was set to no purging ("-l 0") and performed the poorest, with a primary assembly of 5,373 contigs totalling 5.298 Gb, an N50 of 4.276 Mb, and L50 of 243. The secondary assemblies were also quality controlled but were not taken forward.

Table 4: *Genome assembly statistics for the primary (P) and alternative assemblies (A) tested with different levels of purging in Hifiasm.*

	Assembly 1		**Assembly 2**		**Assembly 3**	
Assembly Type	P	A	P	A	P	A
Number of Contigs	2,381	5,090	1,072	7,525	5,373	2,357
Largest Contig (Mb)	71.769	9.496	137.412	9.161	6.573	0.316
Total Length (Gb)	3.600	1.771	3.154	0.224	5.298	0.085
GC (%)	45.850	44.460	45.250	45.750	45.340	51.130
Contig N50 (Mb)	14.365	1.263	71.171	1.127	4.276	0.039
Contig L50	64	399	18	536	243	696

All three primary assemblies were then assessed for completeness with BUSCO (**Figure 19**). Primary "Assembly 2" had the highest percentage of complete and single-copy genes from the "mammalia_odb10" database (96.400%), 1.800% were duplicated, 0.900% were fragmented, and 2.700% genes were missing. Primary assemblies 1 and 3 had lower percentages of complete genes and higher percentages of duplicated genes.

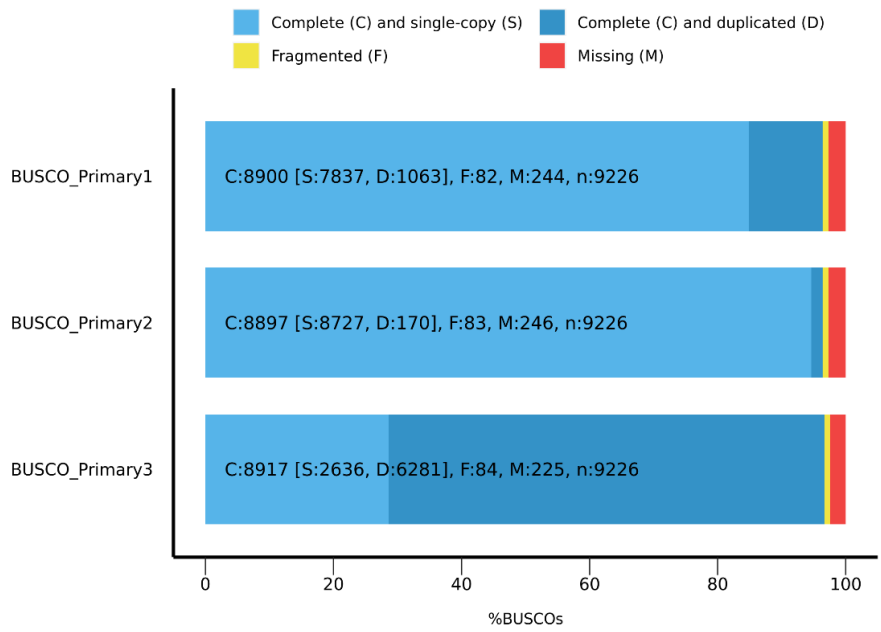


Figure 19: *BUSCO assessment of genome completeness of the three primary assemblies (Assembly 1, Assembly 2, and Assembly 3) using the mammalia_odb10 database. The total number of genes in the database is given by n:9226. The number of complete genes (C), complete and single copy (S), complete and duplicated (D), fragmented (F), and missing (M).*

Overall, primary "Assembly 2" was the most contiguous and complete of the three assemblies and therefore was taken forward for further quality control and genome scaffolding. Primary "Assembly 2", now denoted as the "waterbuck contig-level genome assembly", was then quality controlled with the k-mer-based program Merqury (Rhie et al., 2020). The program compares k-mers in the trimmed HiFi reads with k-mers in the genome assembly, and the copy number of each k-mer. There were a high number of k-mers only found in the trimmed HiFi reads and not in the assembly (black line) which may be indicative of sequencing errors or missing sequencing data from the assembly (**Figure 20**). However, the red line (1) representing one-copy k-mers had a maximum k-mer multiplicity of approximately 18, similar to the k-mer analysis undertaken before genome assembly with a coverage of ~18X (**Figure 18**). K-mers with copy numbers of two and above had a very small count in the genome assembly. The assembly graph of the contig-level genome assembly was visualised (**Figure 21**), which showed that the largest contigs had very few tangles and assembly problems, whereas the smaller contigs were more tangled, had more haplotigs, and more assembly problems. We found no contigs that were joined together.

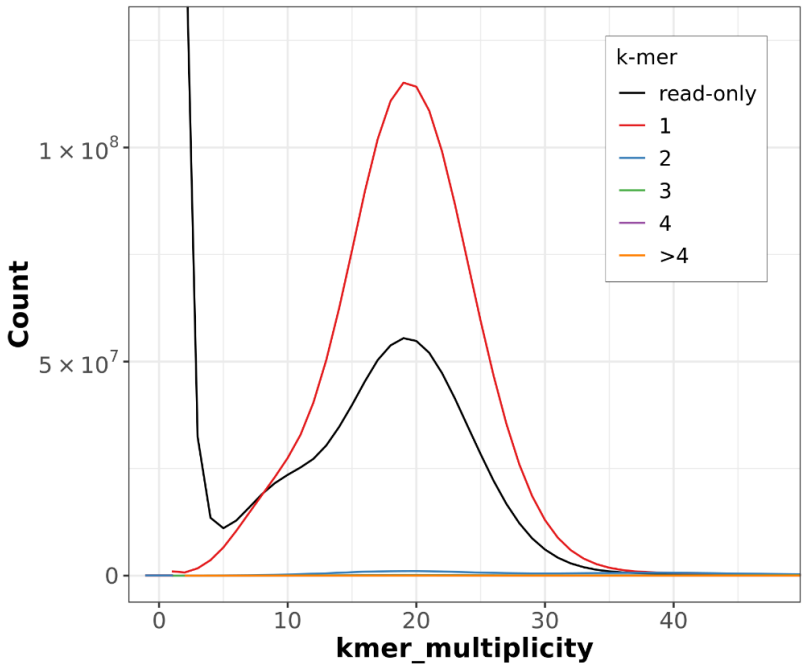


Figure 20: *Copy number (k-mer multiplicity) and count in the trimmed HiFi reads, with the colour denoting the number of times the k-mer is found within the waterbuck contig-level genome assembly.*

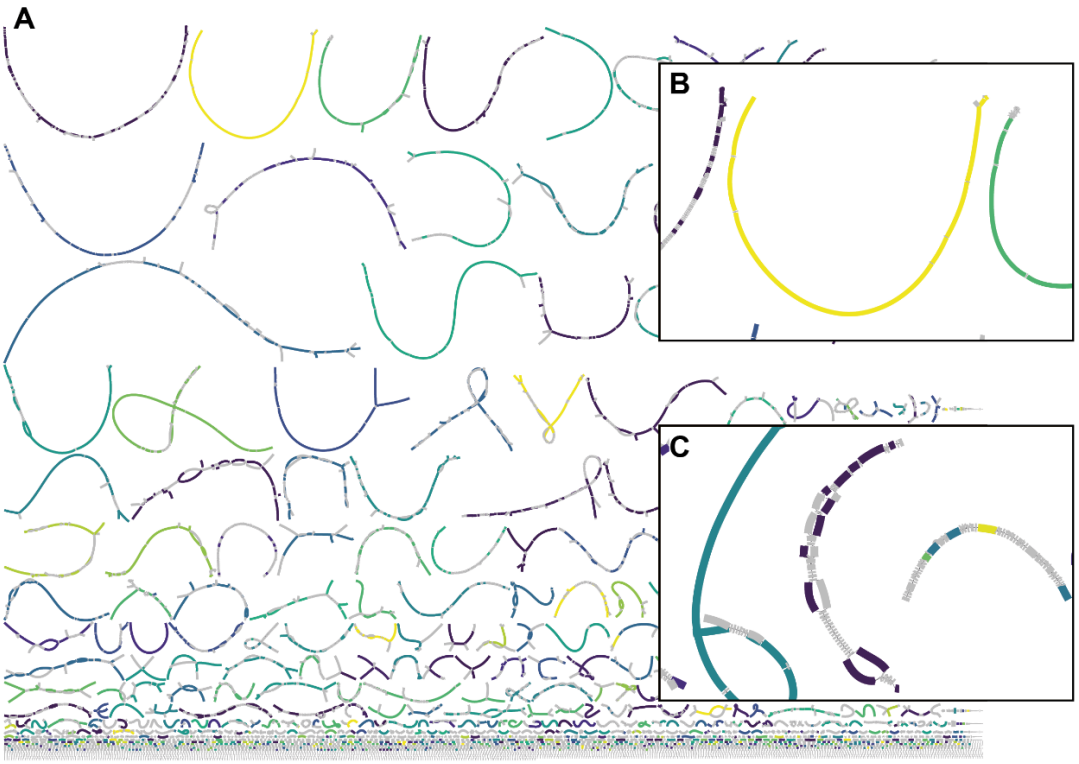


Figure 21: *Bubble plot of the contig-level waterbuck genome assembly graph. (A) All contigs, (B) example of a contig with little variation, and (C) example of a contig with several haplotigs.*

The waterbuck contig-level genome was aligned to the cattle chromosome-level genome and syntenic blocks were constructed (**Figure 22)**. Some contigs spanned entire cattle chromosomes, such as BTA22, and some almost entire chromosomes but with a small contigs at the ends of chromosomes (e.g. BTA2) or near centromeres. Several of the contigs that mapped to the ends of chromosomes or near centromeres mapped to two different chromosomes, such as ptg000105l at the start of BTA1 which also mapped to BTA10. This was potentially due to the repetitive nature of these regions. Other contigs were split, such as ptg000020l on BTA1, which was broken by a small inversion. The X chromosome had several contigs, with some split, however this chromosome is known to be difficult to assemble in mammals due to its highly repetitive nature. Overall, the synteny analysis verified that the contig-level genome was contiguous. Based on the high contiguity and completeness, and k-mer and synteny analysis, the contig-level waterbuck genome passed all our QC tests, summarised with the program BlobToolKit (**Figure 23**).

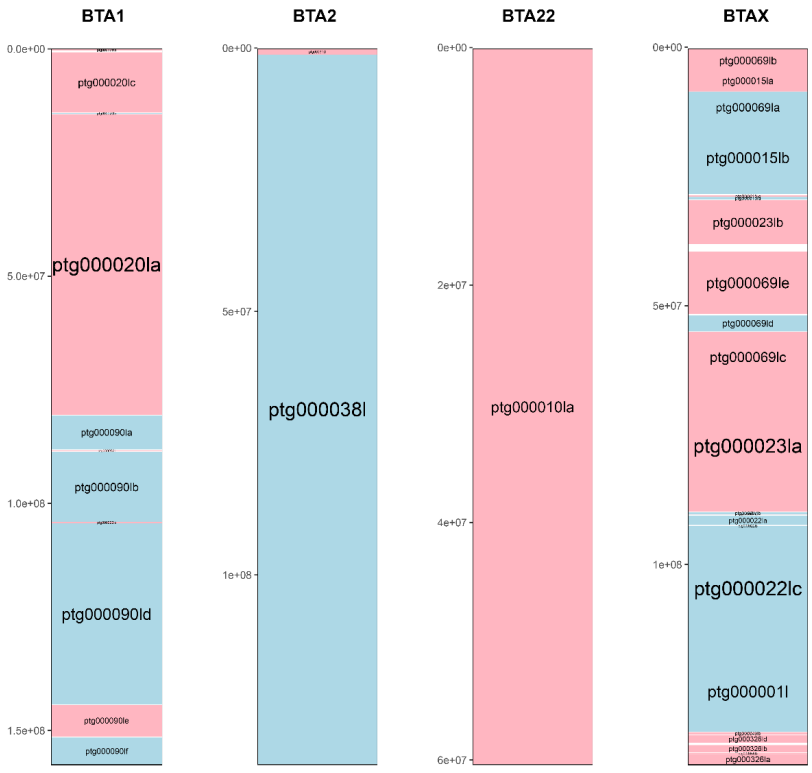


Figure 22: *Synteny between the contig-level waterbuck and cattle chromosome-level genome. Blue synteny blocks represent the same orientation between cattle and waterbuck, whilst red synteny blocks represent the reverse orientation. IDs inside syntenic blocks refer to the contig name and the final letter denotes a split contig. Ideograms shown for only cattle chromosomes BTA1, BTA2, BTA22 and BTAX.*

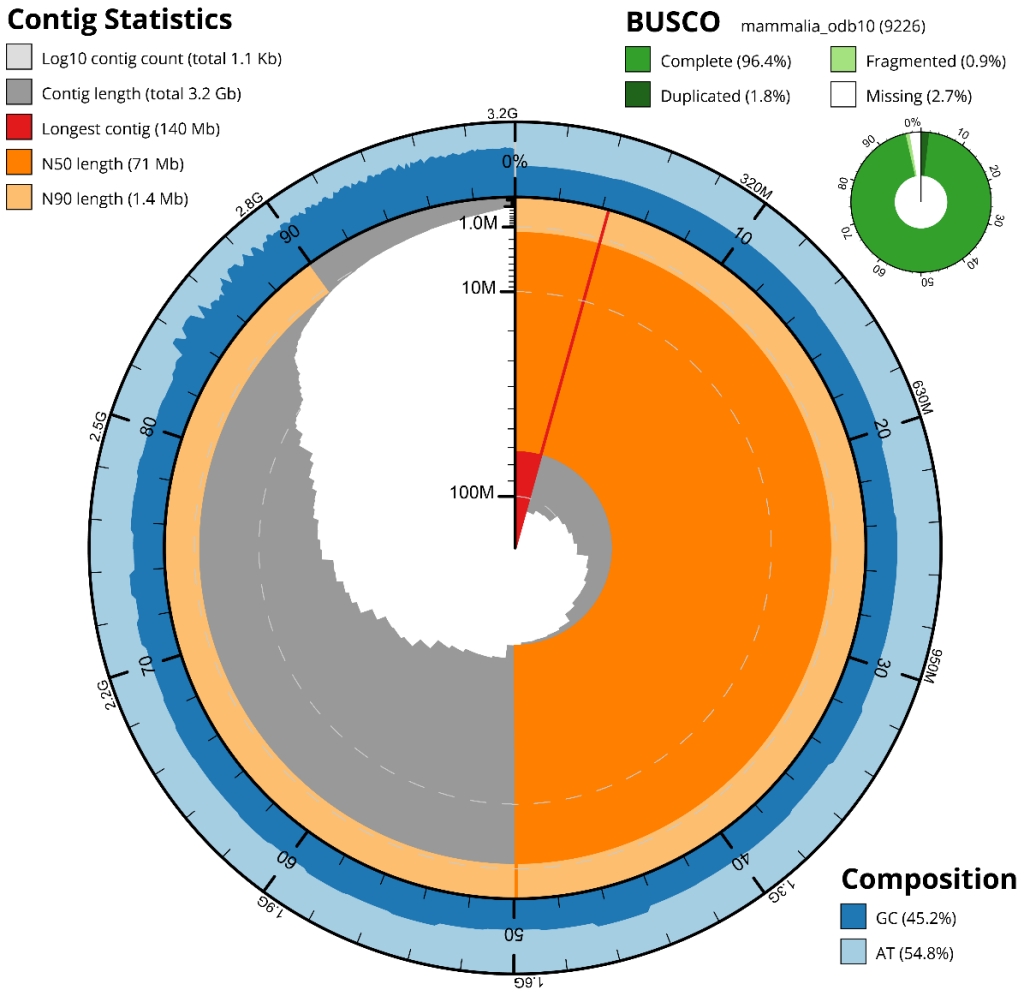


Figure 23: *BlobToolKit plot of the final waterbuck contig-level genome assembly. The figure contains the contig statistics, BUSCO completeness, and "snail plot". The latter plot summarises the percentage of GC bases in dark blue and AT bases in light blue. The contig lengths in grey are ordered in size from largest to smallest, with red representing the longest contig. Dark orange represents the contig N50 and light orange represents the contig N90.*

2.3.4. Mitochondrial genome assembly

The mitochondrial genome was also assembled from the trimmed PacBio HiFi reads from the defassa waterbuck cell line using MitoHiFi. This resulted in a genome of 16,427 bp in length with 37 annotated genes (**Figure 24**). Of these, 13 genes were coding sequence (CDS), 22 were tRNA, and two were rRNA.

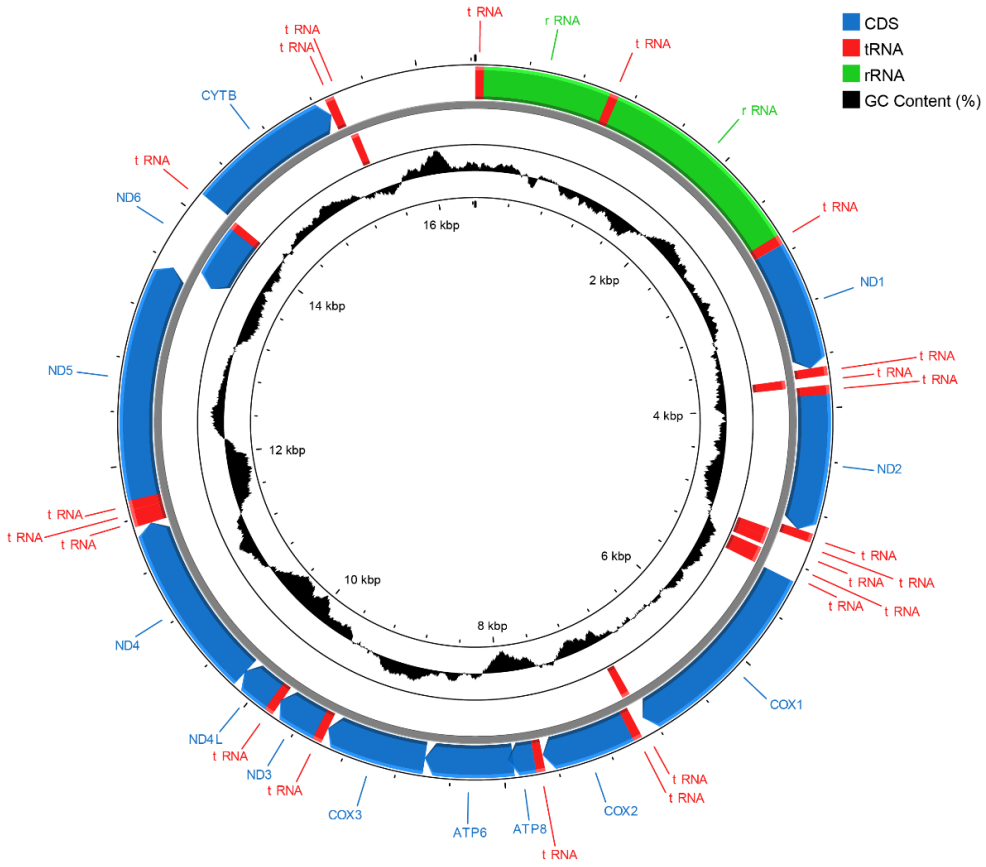


Figure 24: *Mitochondrial genome assembly and annotation. GC content (%; black) centred at 50%. Annotation is classified into coding sequence (CDS; blue), tRNA (red), and rRNA (green).*

2.3.5. Chromosome-level genome assembly

To scaffold the contig-level waterbuck genome into chromosomes, we used chromatin conformation capture and specifically Hi-C. The defassa waterbuck cell line (2*n* = 54), used in the PacBio long-read sequencing, was fixed and sequenced with Illumina short-reads. This resulted in 199.569 M paired end reads (2 x 150 bp). QC of the raw reads (**Table 5**) revealed a very high level of duplication (51-53%) and high mean GC content (50-51%), suggesting an issue with the sequencing data.

Table 5: *Defassa waterbuck (2n = 54) Hi-C sequencing quality control. Sample name refers to sequencing type (Hi-C), the karyotype (2n = 54), and the read pair (R1 or R2).*

Sample	Duplicated (%)	GC Content (%)	Mean Length (bp)	Total Reads
Hi-C_54_R1	51.442	50	151	199,568,812
Hi-C_54_R2	52.818	51	151	199,568,812

Mapping the Hi-C sequencing data to the contig-level genome and visualising the chromosome interactions (Hi-C matrix; **Figure 25**) revealed high levels of inter-chromosomal and intra-chromosomal interactions. The highest signal was on the diagonal line, representing strong close proximity intra-chromosomal interactions, as expected. However, the high levels of inter-chromosomal interactions were unexpected and represented an error in the Hi-C preparation. Thus, we were unable to use this Hi-C sequencing data to scaffold the contig-level genome assembly.

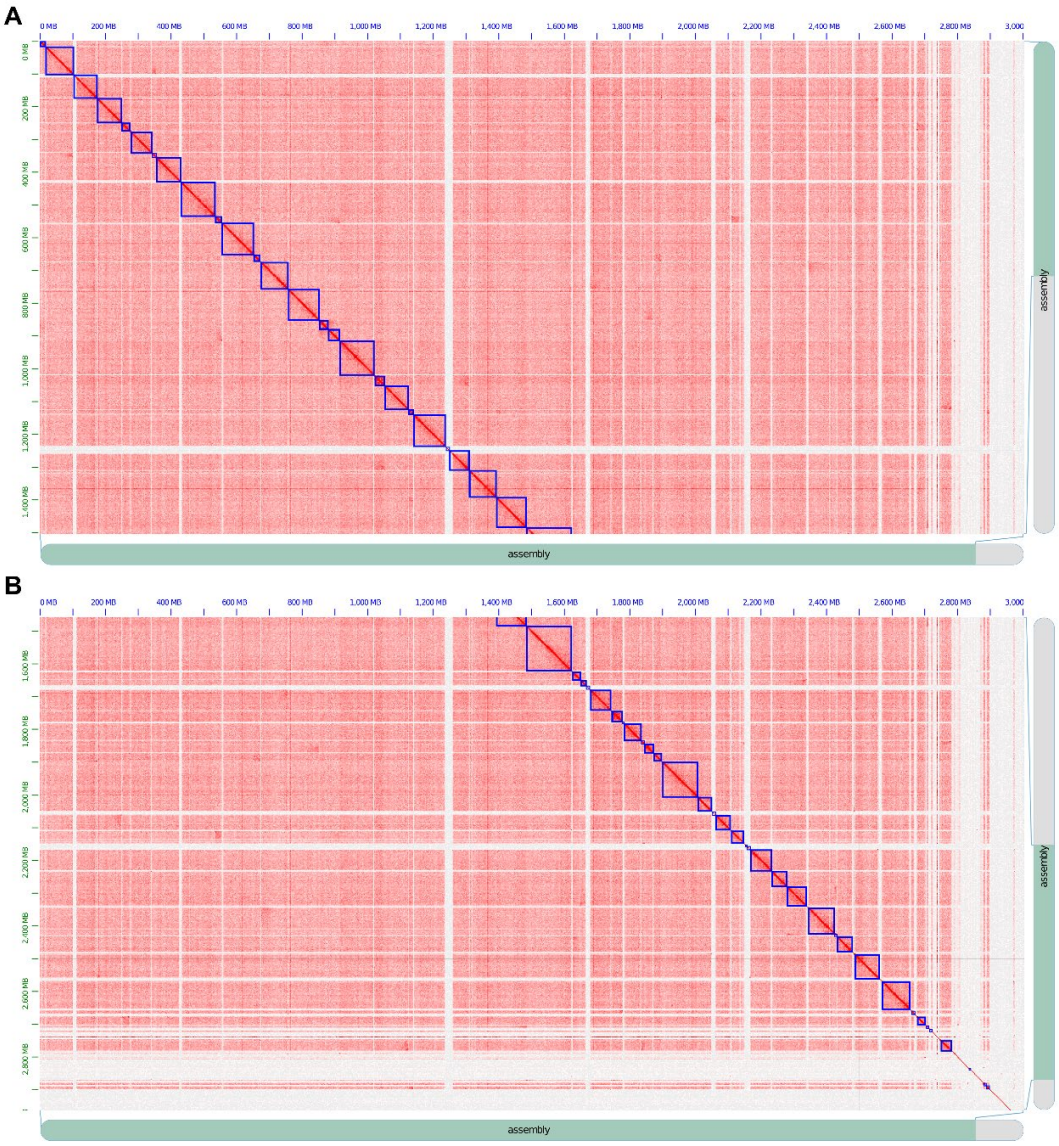


Figure 25: *Hi-C matrix for the defassa waterbuck sample (2n = 54) mapped to the contig-level waterbuck genome assembly. The figure is split between the start (A) and end (B) of the genome. Blue boxes represent contigs and are ordered by contig number.*

In order to scaffold the contig-level genome assembly, we collaborated with the DNA Zoo consortium who provided Hi-C data and performed the scaffolding step. The Hi-C data consisted of 375 million paired-end reads, with a GC content of 45% and a level of duplication (between 22.396% and 25.695%; **Table 6**) lower than the 2n = 54 waterbuck Hi-C dataset, suggesting less error.

Table 6: *Defassa waterbuck (2n = 52) Hi-C sequencing quality control. Sample name refers to sequencing type (Hi-C), the karyotype (2n = 52), and the read pair (R1 or R2).*

Sample	Duplicated (%)	GC Content (%)	Mean Length (bp)	Total Reads
Hi-C_52_R1	25.695	45	151	375,493,105
Hi-C_52_R2	22.396	45	151	375,493,105

The Hi-C dataset was from a captive common waterbuck blood sample with a karyotype of 2*n* = 52, containing the BTA6;18 homozygous fusion. However, our PacBio sequencing data was from a defassa waterbuck sample without any fusions (2*n* = 54). In order for the final chromosome-level genome to have the wt karyotype of the defassa waterbuck (2*n* = 54), we firstly scaffolded the genome into chromosomes with the common waterbuck (2*n* = 52) Hi-C dataset and then broke the two scaffolds of the fused chromosomes. To confirm this, we looked at the synteny between the contig-level genome assembly and cattle chromosomes BTA6 and BTA18 in the chromosome-level genome assembly (**Figure 26**). Cattle BTA6 was syntenic to contigs ptg000021l and ptg000056l, whilst BTA18 was syntenic to contigs ptg000003l and ptg000060l. We used this information to locate the scaffolded contigs in the Hi-C matrix, and to manually break them into two scaffolds, representing the two unfused chromosomes in the 2*n* = 54 waterbuck karyotype (**Figure 27**).

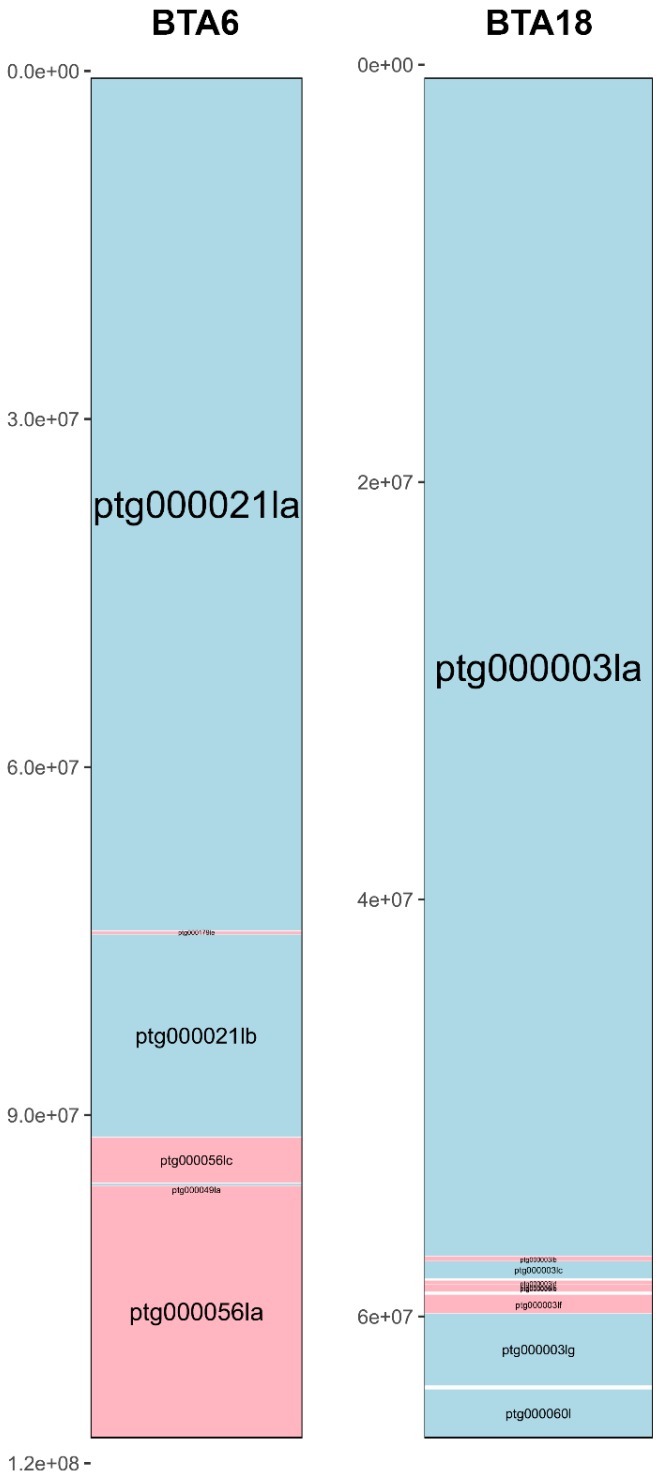


Figure 26: *Synteny of the contig-level genome assembly to cattle chromosomes BTA6 and BTA18. Blue synteny blocks represent the same orientation between cattle and waterbuck, whilst red synteny blocks represent the reverse orientation. IDs inside syntenic blocks refer to the contig name and the final letter denotes a split contig.*

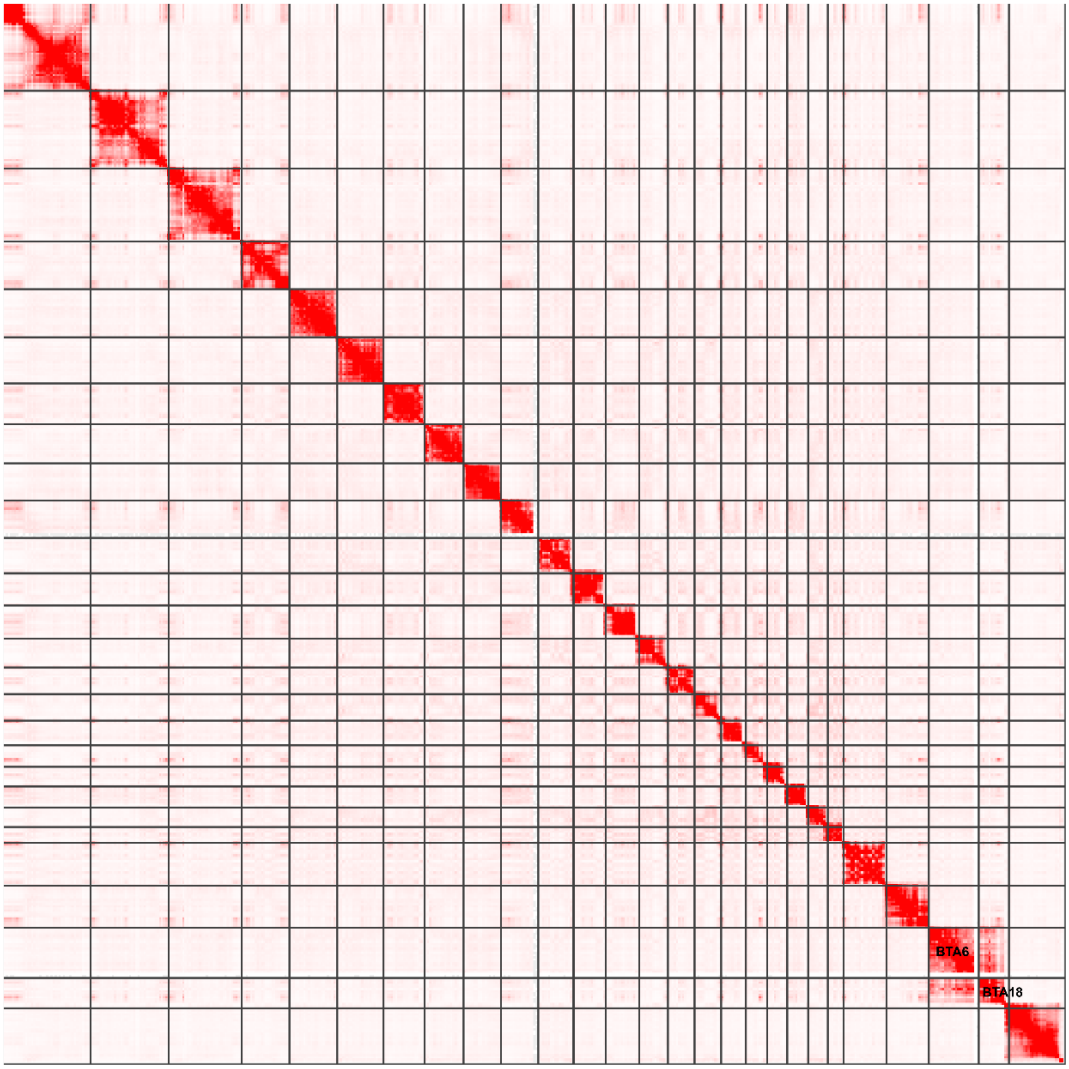


Figure 27: *Interaction matrix of the waterbuck Hi-C sample 2n = 52, mapped to the waterbuck chromosome-level assembly prior to genome curation.*

The chromosome-level genome assembly was then curated using synteny to the cattle chromosome-level genome (**Figure 28**), where acrocentric chromosomes had been orientated with centromere (or the region located closest to the centromere) at the start of the sequence and submetacentric chromosomes with the centromere (or the region surrounding the centromere) closer to the start of the sequence than the end. This resulted in the reorientation of waterbuck chromosomes HiC_scaffold_2, HiC_scaffold_10, HiC_scaffold_12, HiC_scaffold_15, HiC_scaffold_16, HiC_scaffold_17, HiC_scaffold_19, HiC_scaffold_23, HiC_scaffold_24, and HiC_scaffold_25 (**Table 7**). Additionally, waterbuck chromosomes were reordered by length and renamed (i.e., chr1 to chr26, chrX, followed by the unplaced scaffolds).

Table 7: *Chromosome names before and after curation. Chromosomes were reordered by size and renamed. Chromosomes that were reorientated are denoted by 'Y'.*

Before Curation	Size (bp)	After Curation	Reorientated?
HiC_scaffold_1	224,392,242	chr1	N
HiC_scaffold_2	197,800,822	chr2	Y
HiC_scaffold_3	183,868,402	chr3	N
HiC_scaffold_4	124,412,615	chr4	N
HiC_scaffold_5	122,906,609	chr5	N
HiC_scaffold_25	120,449,952	chr6	Y
HiC_scaffold_6	115,378,235	chr7	N
HiC_scaffold_24	108,576,873	chr8	Y
HiC_scaffold_23	107,344,814	chr9	Y
HiC_scaffold_7	105,395,809	chr10	N
HiC_scaffold_8	98,446,286	chr11	N
HiC_scaffold_9	96,839,935	chr12	N
HiC_scaffold_10	93,141,722	chr13	Y
HiC_scaffold_11	90,291,058	chr14	N
HiC_scaffold_12	83,106,378	chr15	Y
HiC_scaffold_13	81,989,244	chr16	N
HiC_scaffold_26	75,501,084	chr17	N
HiC_scaffold_14	73,691,448	chr18	N
HiC_scaffold_15	70,443,932	chr19	Y
HiC_scaffold_16	64,013,170	chr20	Y
HiC_scaffold_17	63,006,212	chr21	Y
HiC_scaffold_18	53,227,197	chr22	N
HiC_scaffold_19	53,015,818	chr23	Y
HiC_scaffold_20	52,569,133	chr24	N
HiC_scaffold_21	48,126,938	chr25	N
HiC_scaffold_22	46,130,729	chr26	N
HiC_scaffold_27	148,333,903	chrX	N

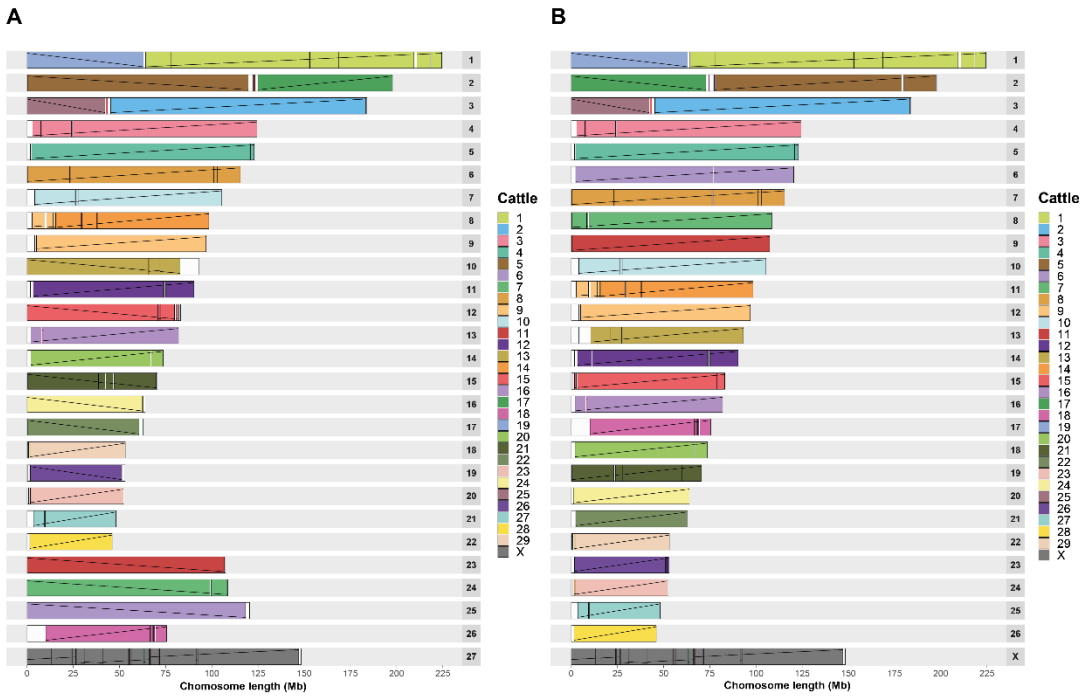


Figure 28: *Synteny of the waterbuck chromosome-level genome to the cattle genome, before (A) and after (B) curation. Rows indicate waterbuck chromosomes (1-26 and X). Syntenic cattle chromosomes painted onto waterbuck chromosomes. Horizontal lines represent an evolutionary breakpoint region (EBR). Diagonal lines represent the orientation of syntenic blocks (reverse orientation represented by a diagonal line from top to bottom).*

Following genome curation, the Hi-C dataset (2*n* = 52) was remapped to the curated chromosome-level genome assembly (**Figure 29**) and quality controlled (**Table 8** and **Figure 30**). In summary, the final waterbuck genome was 3.154 Gb in length with a total of 27 chromosomes (26 autosomes and the X chromosome), a scaffold N50 of 98.446 Mb, and L50 of 12, with 0.452 Gb in unplaced scaffolds. The unplaced scaffolds had higher GC content than the chromosomes, suggesting they contained repetitive sequences that were unable to be assembled correctly. BUSCO analysis resulted in 96.300% of the genes in the mammalia_odb10 database being complete in the final assembly. Consequently, the final waterbuck chromosome-level genome assembly was both highly contiguous and highly complete.

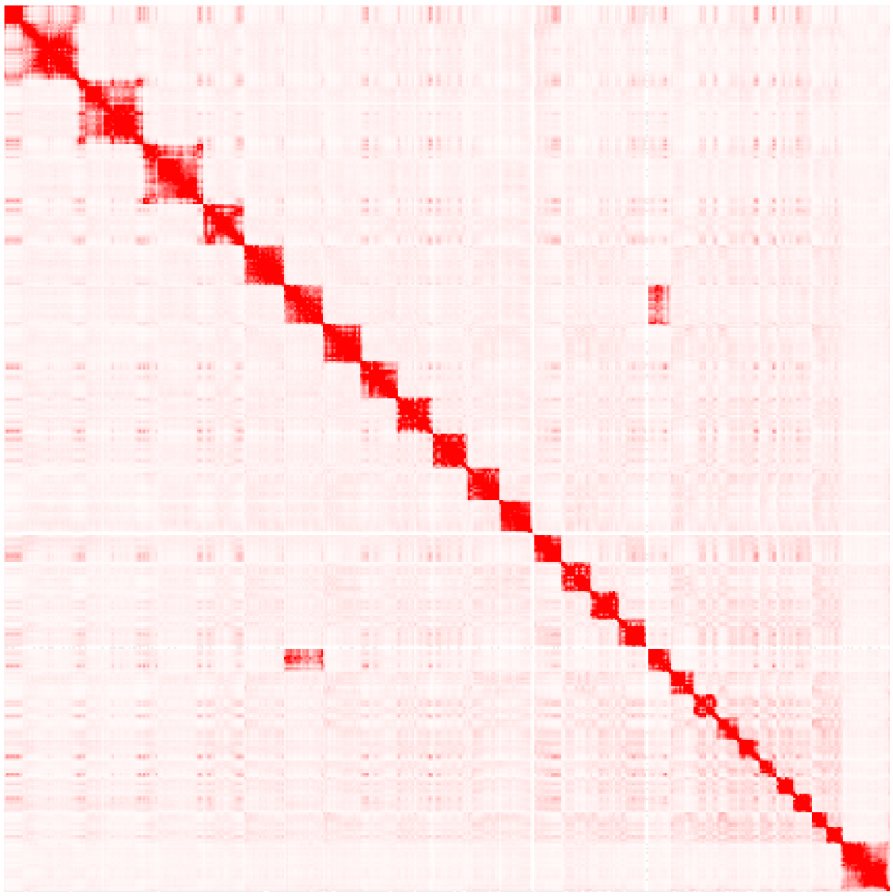


Figure 29: *Interaction matrix of the Hi-C 2n = 52 waterbuck sample mapped to the curated chromosome-level genome assembly. Chromosomes were ordered by size and in the correct orientation. The X chromosome was placed at the end.*

Table 8: *Summary of the waterbuck genome assembly statistics.*

Contig Length (Gb)	3.154
No. Contigs	1,071 7
Contig N50 (Kb)	71.171
Contig L50	18
Scaffold Length (Gb)	3.154
No. Scaffolds	1,014
No. Chromosomes	27
Scaffold N50 (Kb)	98.446
Scaffold L50	12
GC Content (%)	45.200

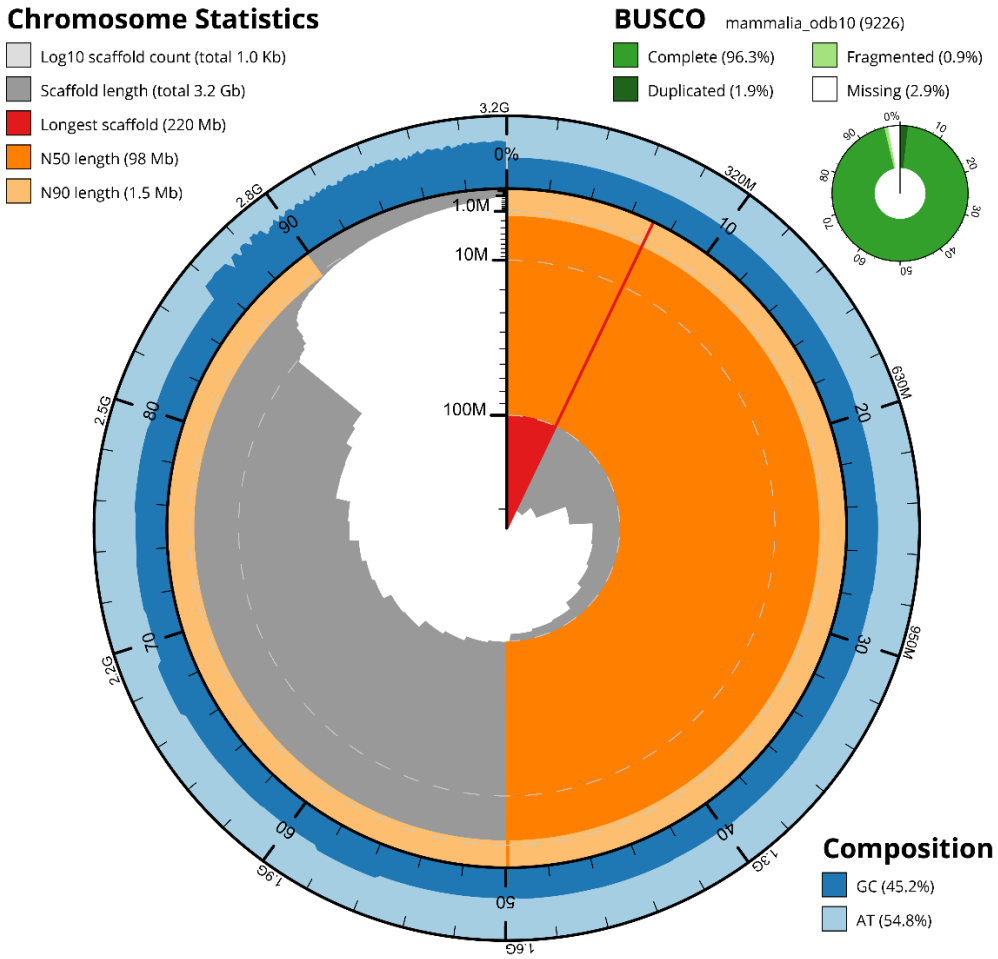


Figure 30: *Snail plot of the waterbuck chromosome-level genome assembly. The figure contains the scaffold statistics, BUSCO completeness, and "snail plot". The latter plot summarises the percentage of GC bases in dark blue and AT bases in light blue. The scaffold lengths in grey are ordered in size from largest to smallest, with red representing the longest scaffold. Dark orange represents the scaffold N50, and light orange represents the scaffold N90.*

2.3.6. Genome annotation

Homology-based gene annotation was undertaken using the cattle (*Bos taurus*; ARS-UCD2.0) and goat (*Capra hircus*; ARS1.2) annotations as references with the program GeMoMa, resulting in 33,077 annotated protein-coding genes, which were then filtered to 24,645 genes, removing duplicates. For each annotated position, the reference gene was used, and any alternative genes were omitted.

Moreover, repeats were annotated with the tool RepeatMasker using cattle as reference. A total of 1,728,574,685 bp (54.801%) of the waterbuck genome was masked, with LINEs making up 24.259%, satellite and centromeric repeats 14.242%, SINEs 9.697%, LTR

elements 3.937%, and DNA elements 1.908% (**Table 9**). RTE-BovB elements which are characteristic of ruminant genomes were the majority of the LINE repeats (11.022%).

Table 9: *Homology-based repeat annotation of the waterbuck genome assembly from RepeatMasker. Only selected classes and families were included.*

Class/Family	bp	%
SINEs	**305,882,625**	**9.697**
SINE/tRNA	200,532,122	6.357
SINE/MIR	55,849,322	1.771
SINE/Core-RTE	49,351,161	1.565
LINEs	**765,195,990**	**24.259**
LINE/L1	346,414,515	10.982
LINE/L2	62,582,782	1.984
LINE/RTE-BovB	347,653,427	11.022
LINE/RTE-X	1,678,818	0.053
LINE/CR1	6,740,296	0.214
LTR Elements	**124,171,252**	**3.937**
LTR/ERV1	34,934,861	1.108
LTR/ERVK	17,841,991	0.566
LTR/ERVL	67,582,681	2.143
DNA Elements	**60,187,108**	**1.908**
DNA/hAT	44,505,333	1.411
DNA/TcMar	14,371,705	0.456
Satellites/Centromeric	**449,248,494**	**14.242**
Simple Repeats	**24,560,056**	**0.779**
Low Complexity	**4,273,955**	**0.135**
Total Masked	**1,728,574,685**	**54.801**
Total Annotated	**1,735,188,903**	**55.010**

We compared repeats found in the 27 chromosomes with those found in the unplaced scaffolds in the chromosome-level assembly. We found that the majority of repeats in the unplaced scaffolds were annotated as satellite/centromeric repeats (397,174,998 bp) compared to the assembled chromosomes which had a total of 52,073,496 bp of these repeats (**Figure 31**).

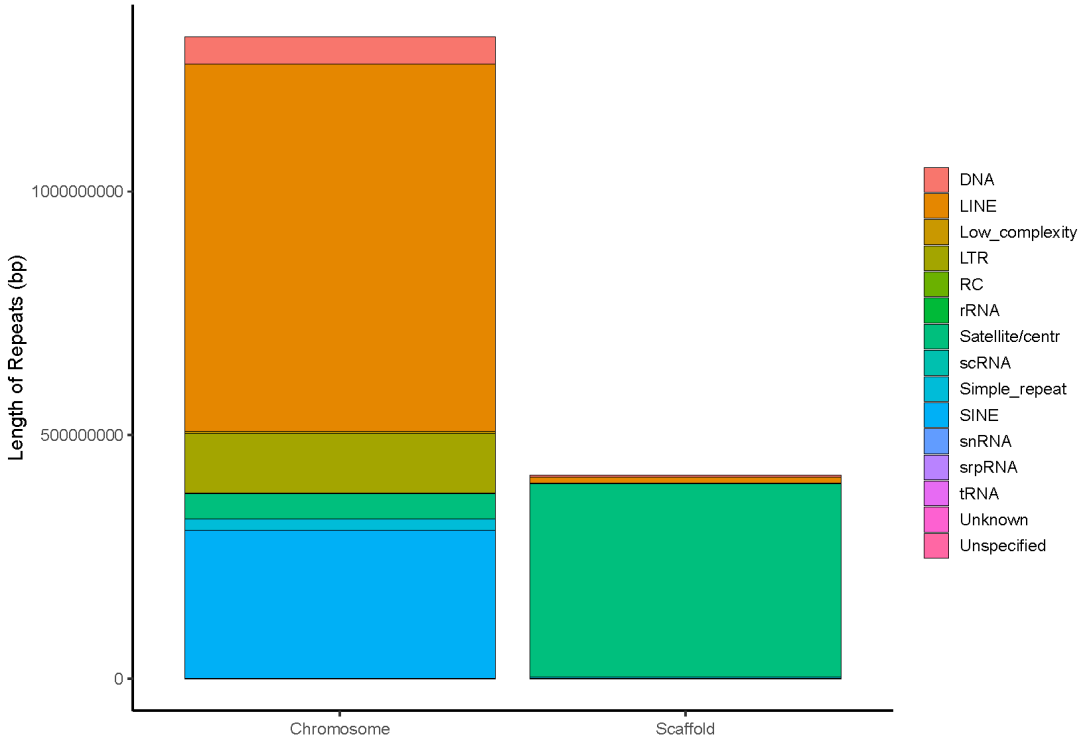


Figure 31: *Length of repeats (bp) in the chromosomes and scaffolds in the waterbuck chromosome-level genome, grouped by type of repeat.*

GC content, repeat density, and gene density were also calculated in 100 Kb windows and visualised across the genome (**Figure 32**). GC content was highest at the starts and ends of chromosomes, and near the likely locations of centromeres (either off centre in submetacentric chromosomes or near the beginning of acrocentric chromosomes). Whilst repeat density was only highest near the centromeres. Gene density was variable on each chromosome.

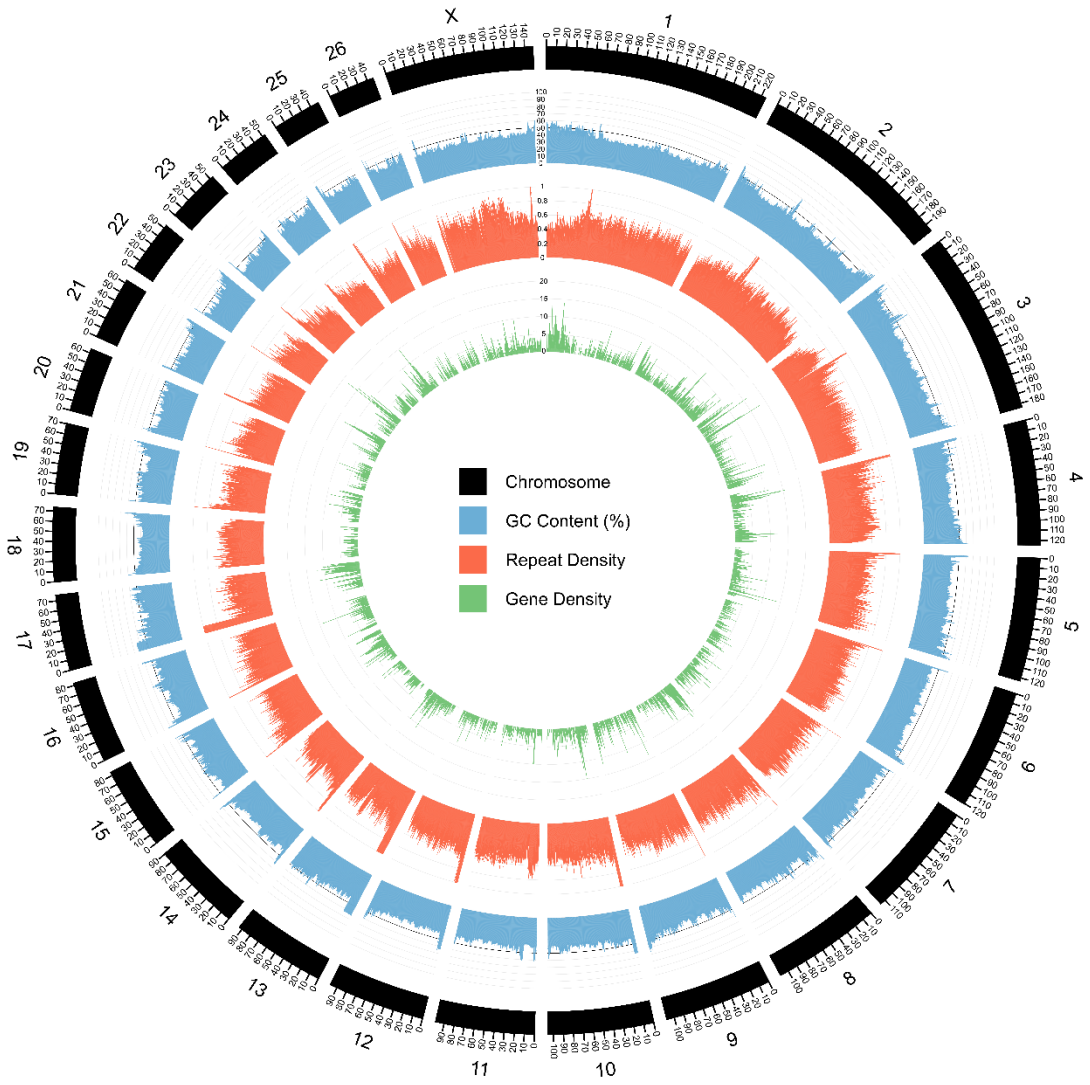


Figure 32: *Genome annotation summary of the waterbuck chromosome-level genome – GC content (%), repeat density, and gene density in 100 Kb windows.*

2.4. Discussion

In this chapter we sequenced and assembled a genome for the waterbuck at chromosome-level using a combination of PacBio HiFi long-reads and Hi-C sequencing. This resulted in a genome of 3.154 Gb in length (**Table 8**), which was 0.258 Gb longer than the scaffold-level assembly produced in the ruminant genome project (Chen et al., 2019) and closer to their estimated genome size of 3.374 Gb with k-mers. We also estimated the genome size with k-mers using the HiFi reads, and this was predicted to be slightly shorter than our assembly, between 3.059 and 3.063 Gb (**Table 3**). This suggests that we have produced a relatively complete genome assembly for the waterbuck. This was

supported by the BUSCO analysis where 96.300% of the orthologous genes across mammals were present in our genome, with only 2.95% missing (**Figure 30**).

The waterbuck assembly was also highly continuous, composed of 1,071 contigs and 1,014 scaffolds, with a contig N50 of 71.171 Mb and a scaffold N50 of 98.446 Mb (**Table 8**). Whereas the previous assembly had 279,061 contigs with an N50 of 0.207 Mb and 88,848 scaffolds with an N50 of 0.782 Mb (Chen et al., 2019). This demonstrates the power of long-read sequencing for assembling a genome to contig-level, where the N50 was 343 times greater than using short reads. Also, the use of Hi-C to scaffold genomes, rather than the previous short-read scaffolding programs, increased the scaffold N50 by 125 times. The final curated genome assembly had 26 autosomal chromosomes and one sex chromosome (**Figure 32**).

Our genome meets the VGP-2020 standards, including a contig NG50 greater than 10 Mb, a scaffold NG50 that equals the chromosome NG50, and gene completeness greater than 95% (Rhie et al., 2021). The waterbuck genome also shares similar statistics to other bovid genomes assembled with PacBio HiFi sequencing. For example, the takin genome had a contig N50 of 68.054 Mb , scaffold N50 of 101.266 Mb, and a BUSCO completeness of 94.2% (A. Li et al., 2023). However, most currently assembled bovid genomes have used either short-reads or PacBio CLR sequencing, with the former resulting in fragmented genome assemblies. For example, the chromosome-level roan antelope assembly used short-read sequencing and had a contig N50 of 0.237 Mb, but a scaffold N50 of 99.088 Mb. This assembly may have some inaccuracies in the correct number of copies of repetitive regions due to the sequencing technology. Therefore, a greater number of assemblies using the latest long-read sequencing technologies is needed to accurately achieve high-quality antelope genomes.

Some antelopes have polymorphic Robertsonian fusions and therefore it is important to know the karyotype of the animal being sequenced. We did this by karyotyping the sample before extracting DNA and sequencing. Our waterbuck sample had the standard karyotype for the defassa subspecies, without any polymorphic fusion (**Figure 15**). We were then cautious during the assembly process to maintain this karyotype. This was done through the use of a synteny analysis to the cattle genome (**Figure 26**), which had been used in previous cytogenetic studies of waterbuck (S. Kingswood et al., 1998; S. C. Kingswood et al., 2000), and therefore we could identify which contigs or scaffolds were syntenic to these polymorphic chromosomes. This was important as our Hi-C dataset

produced from the cell culture sample had spurious interaction signals and thus we had to rely on a common waterbuck sample donated to us by DNA Zoo. This sample had a karyotype of 2*n* = 52, so we manually broke the scaffold that was syntenic to the cattle chromosomes BTA6 and BT18, in order to maintain the 2*n* = 54 karyotype (**Figure 29**).

Synteny and karyotyping data was also useful when curating the genome, as we had knowledge of the synteny between cattle and waterbuck chromosomes, including the fixed Rb fusions. This enabled us to be confident that our genome was assembled into relatively complete and accurate chromosomes, as well as in the correct orientation. For example, the genome confirmed results from the previous cytogenetic studies (S. Kingswood et al., 1998; S. C. Kingswood et al., 2000) that waterbuck have three submetacentric chromosomes due to ancestral Rb fusions, a submetacentric X chromosome, and the remaining 23 chromosomes are acrocentric. Our sample was from a female waterbuck (XX), and therefore we were unable to assemble the Y chromosome. This may be worth sequencing in the future as it has been proposed to be variable between the two subspecies (S. Kingswood et al., 1998; S. C. Kingswood et al., 2000). However, the Y chromosome has been difficult to sequence in mammals due to its highly-repetitive nature (Rhie et al., 2023).

We additionally made use of homology-based programs to annotate genes and repeats across the genome. For genes, this was carried out using cattle and goat as reference. A total of 24,645 protein-coding genes were annotated in the waterbuck genome. Compared to the cattle genome (ARS-UCD2.0), our annotation had slightly more protein-coding genes, as the total number of genes in cattle is 37,073, with 21,667 coding and 10,686 non-coding genes. To achieve a full and accurate set of annotated genes, RNA sequencing of multiple tissues is needed. However, we were unable to obtain samples from several tissues to extract RNA in this study.

For repeats, cattle was used as reference and this resulted in 54.801% of the waterbuck genome being identified as repetitive (**Table 9**), with the majority annotated as LINEs (24.259%), followed by satellite and centromere repeats (14.242%), SINEs (9.697%), and then LTR elements (3.937%), similar to previous studies on bovid genomes that found TEs were higher in this clade compared to others (Chen et al., 2019). Our assembly had a higher percentage of repeats annotated as satellite or centromeric (14.242%), compared to other short-read assemblies such as the sable antelope (1.58%; Koepfli et al., 2019) and the scimitar-horned oryx (1.73%; Humble et al., 2020), and long-read

assemblies such as the takin (6.97%; A. Li et al., 2023). This may be due to PacBio HiFi sequencing being able to sequence across these highly repetitive regions, as KEL13 and KEL17 are highly repetitive near the beginning of these chromosomes (**Figure 32**).

RTE-BovB LINEs were the most common LINEs in our waterbuck genome assembly (11.022%; **Table 9**). These TEs have been shown to have occurred through horizontal transfer in the ancestor of ruminants and then subsequently expanded (Adelson et al., 2009; Ivancevic et al., 2018). BovB elements encode the machinery needed to transpose SINE elements such as BovA2, BOV-tA, and ART2A (Adelson et al., 2009). Bov-A2 elements may have played a role in immune gene regulation during bovid evolution, as well as creating genomic variation through polymorphisms (Kelly et al., 2022). With the increasing availability of long-read genome assemblies that are able to accurately sequence these repeats, the effect of BovB elements and their associated SINE elements on genome evolution and variation will be further uncovered.

Advances in genome assembly are ongoing, with new genome assemblies beginning to be published that are sequencing entire chromosomes without gaps from telomere to telomere (e.g., the human T2T genome; Nurk et al., 2022). These new T2T human genome assemblies now include satellite repeat arrays and the short p-arms of the five acrocentric chromosomes, which has not been sequenced correctly in previous versions (Nurk et al., 2022). To achieve this, researchers used PacBio HiFi long-reads for sequencing accuracy and supplemented this with Oxford Nanopore Technologies ultra-long reads (> 100 Kb) to span the entirety of repeats, establishing a new path forward in achieving complete genomes.

One key finding using these T2T human genomes has been the identification of recombination between heterologous human acrocentric chromosomes (Guarracino et al., 2023). During the detangling phase of genome assembly, researchers found that the five human acrocentric chromosomes created tangles among the short arms of these chromosomes, particularly in regions of rRNA arrays. A population-level analysis of these regions revealed Mb-sized homology blocks shared between acrocentric chromosomes, termed pseudo-homologous regions (PHRs). The existence of PHRs implies frequent interhomolog recombination and might explain how Robertsonian fusions might form between these acrocentric chromosomes (Guarracino et al., 2023). However, rRNA arrays in bovids are not located in the short arms of acrocentric chromosomes (Gallagher et al., 1999), instead, satellite DNA repeats are in these areas

(Escudeiro et al., 2021). We did not find any tangles between the homologous short arms in waterbuck (**Figure 21**), potentially due to the length of PacBio sequencing reads which cannot stretch across these long repetitive regions. Satellite repeats were mostly found in the short contigs that were unable to be scaffolded with Hi-C alone (**Figure 31**). A new consortium is being established to sequence T2T genomes for most ruminant species (Kalbfleisch et al., 2024) which will help to sequence the p-arms and centromeres, and investigate the role of satellite repeats in the formation of Rb fusions.

Chromosome-level genome assemblies are a useful resource for genomic studies, particularly those studying chromosome evolution. For example, in Lepidoptera, they identified 32 ancestral chromosomes and found whilst most lineages remained conserved to the ancestral state, some lineages had extensive rearrangements due to fusions and fissions, with fusions mostly occurring in smaller chromosomes (Wright et al., 2024). This was also the case in frogs, where synteny of the 13 chromosomes was relatively conserved since the ancestor, but with a limited number of Rb fusions in smaller chromosomes (Bredeson et al., 2024).

Whilst scaffold-level genomes can be useful in population genomics, studies that have used chromosome-level genome assemblies as a reference have been able to study genomic differentiation, genome-wide association, and selection across chromosomes, identifying important genes. In the white-tailed jackrabbit, the genetic basis of winter hair colour was determined using genome-wide association studies across chromosomes and identified populations that were able to adapt to a warming climate due to loss of snow cover (Ferreira et al., 2023). Whilst in oysters, a selective sweep analysis was used to detect genes that were associated with heat response and salinity adaptation (Wu et al., 2022).

Our genome will provide a useful resource for studying chromosome evolution in ruminants and bovids, where extensive Rb fusions have been found both between and within species. It will also be useful in population genomics studies, improving the current reference genome, and enabling studies to explore population genomic metrics across chromosomes. Lastly, the genome could also be utilised with Hi-C data to study the 3D genome organisation of chromosomes within the nucleus.

3. Population Genomics of the Waterbuck

3.1. Introduction

With a new high-quality chromosome-level genome assembled for the waterbuck, as well as the annotation of genes and repeats, a further assessment of the population genomics of this species can be carried out. Previous studies have utilised the mitochondrial control region (Eline D. Lorenzen et al., 2006; Ogden et al., 2018), microsatellites (Eline D. Lorenzen et al., 2006), and WGS data (X. Wang et al., 2024) to study waterbuck genetics and genomics. However, the latter genomic study made use of the scaffold-level reference genome assembly for the waterbuck (Chen et al., 2019). Whilst this study provided several insights into the evolution of the waterbuck using low coverage WGS data and a scaffold-level reference, further work is needed to fully understand evolutionary processes affecting the species across its chromosomes. Especially as this species has polymorphic Rb fusions which cannot be studied using a scaffold-level reference.

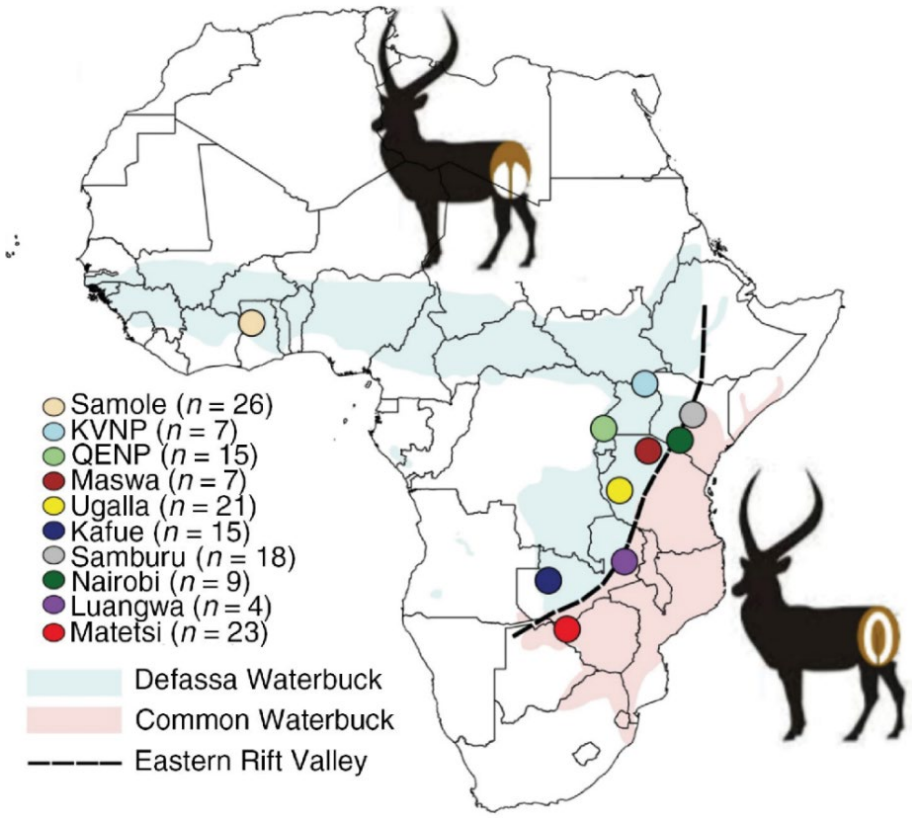


Figure 33: *Sampling map of the whole genome sequencing (WGS) data from 10 waterbuck populations adapted from* X. Wang et al., 2024. *The Great Rift Valley is shown by the dotted line, with the distributions of defassa and common waterbuck given by the blue and red shading, respectively. Morphological differences in rump fur colouration displayed.*

Moreover, a wider sampling of the species distribution is needed to ensure that all populations are considered, which will also benefit the conservation of the species across its entire range. For example, there is missing genomic data from defassa populations in central Africa between Samole and KVNP, populations in the north (i.e., Ethiopia) and in the southwest (Angola, Democratic Republic of Congo; DRC, the Republic of Congo, and Gabon; **Figure 33**). Sampling from these regions will improve the accuracy of population structure, which may be useful for determining conservation priorities. In the common subspecies, studying populations away from the Rift Valley is needed to further understand how far historical gene flow has occurred, as well as population structure in this subspecies.

We therefore set out to use our chromosome-level genome assembly for the waterbuck, utilise the 119 WGS individuals previously published (from X. Wang et al., 2024), and enhance the sampling of the species with an additional 48 historical samples. To do this, we extracted DNA from historical dried skin samples from two museum collections and sequenced samples using low coverage WGS. We then mapped the historical samples, along with the modern samples, to our chromosome-level genome for waterbuck. This allowed us to estimate population structure, admixture, gene flow, and heterozygosity across a wider sampling range for the species. We then explored genomic differentiation across chromosomes, both between subspecies and different population groups, utilising the chromosome-level reference and genome annotation. This chapter aims to further uncover the evolutionary process within waterbuck, as well as apply this to the conservation of the species.

3.2. Materials and Methods

3.2.1. Historical museum sampling

In collaboration with the Aspinall Foundation, we sampled historical waterbuck skin samples from two museums, the Powell Cotton Museum (PCM; Kent, UK) and the Royal Museum for Central Africa (RMCA; Tervuren, Belgium). This totalled 48 waterbuck samples, 24 from each museum, of which 42 were classified as the defassa subspecies and 6 were common (**Table 10**). These samples were chosen to provide a wide representation of both subspecies and geographic localities, however common

waterbuck samples were rare in both museums, limiting their availability and numbers in this study. Several samples from each locality were also prioritised where possible.

Initially, samples were grouped by currently recognised country names to avoid splitting populations too frequently. This naming however did provide a limitation when geographically close samples were grouped into separate populations, or larger countries where individuals are more distantly located. Samples for the defassa waterbuck included the populations of the DRC (n = 19), Kenya (n = 5), Cameroon (n = 4), Angola (n = 3), Chad (n = 3), Guinea-Bissau (n = 2), South Sudan (n = 2), Ethiopia (n = 2), and Rwanda (n = 2). Whilst samples from the common waterbuck included the populations of Somalia (n = 2), Tanzania (n = 2), and Kenya (n = 2).

Table 10: *Waterbuck historical museum samples collected from the Powell Cotton Museum (PCM) and the Royal Museum for Central Africa (RCMA). Sample name used throughout the thesis (ID) and sample name given by the museum (Museum ID). Democratic Republic of Congo is abbreviated to DRC, Guinea-Bissau to G-Bissau, and South Sudan to S. Sudan.*

ID	Museum ID	Museum	Subspecies	Sex	Date	Country
WB_1a	NH.UG.288	PCM	Defassa	M	21/10/1902	Kenya
WB_1b	NH.ANG1.82	PCM	Defassa	M	14/12/1921	Angola
WB_1c	NH.ANG1.93	PCM	Defassa	M	22/12/1921	Angola
WB_1d	NH.CAM1.79	PCM	Defassa	F	28/04/1929	Cameroon
WB_1e	NH.CAM2.140	PCM	Defassa	M	30/07/1931	Cameroon
WB_1f	NH.MER35.8	PCM	Defassa	F	04/03/1935	Cameroon
WB_1g	NH.MER36.218	PCM	Defassa	M	09/03/1936	Cameroon
WB_1h	NH.NN.103	PCM	Defassa	M	15/04/1925	Chad
WB_1i	NH.NN.131	PCM	Defassa	M	08/05/1925	Chad
WB_1j	NH.NN.134	PCM	Defassa	M	08/05/1925	Chad
WB_1k	NH.CON.454	PCM	Defassa	F	04/12/1906	DRC
WB_1l	NH.ABYS1.147	PCM	Defassa	M	22/05/1900	Ethiopia
WB_2a	NH.ANG1.94	PCM	Defassa	M	22/12/1921	Angola
WB_2b	NH.ABYS1.148	PCM	Defassa	M	22/05/1900	Ethiopia
WB_2c	NH.GUI.28	PCM	Defassa	M	20/06/1911	G-Bissau
WB_2d	NH.GUI.29	PCM	Defassa	M	20/06/1911	G-Bissau

WB_2e	NH.UG.147	PCM	Defassa	M	30/05/1902	Kenya
WB_2f	NH.UG.38	PCM	Defassa	M	26/03/1902	Kenya
WB_2g	NH.CON.44	PCM	Defassa	M	16/02/1905	S. Sudan
WB_2h	NH.SUD1.15	PCM	Defassa	M	08/02/1933	S. Sudan
WB_2i	NH.SOM2.7	PCM	Common	M	31/01/1935	Somalia
WB_2j	NH.SOM2.8 A	PCM	Common	F	31/01/1935	Somalia
WB_2k	NH.TAN.8	PCM	Common	M	14/12/1938	Tanzania
WB_2l	NH.TAN.9	PCM	Common	F	18/12/1938	Tanzania
WB_3a	3596	RMCA	Defassa	M	27/07/1914	Kenya
WB_3b	6281	RMCA	Defassa	M	27/04/1923	Kenya
WB_3c	9	RMCA	Defassa	M	NA	DRC
WB_3d	98	RMCA	Defassa	NA	1898	DRC
WB_3e	107	RMCA	Defassa	M	1898	DRC
WB_3f	293	RMCA	Defassa	F	NA	DRC
WB_3g	8603	RMCA	Defassa	M	NA	DRC
WB_3h	13661	RMCA	Defassa	M	03/11/1936	DRC
WB_3i	25425	RMCA	Defassa	M	12/08/1956	DRC
WB_3j	36520	RMCA	Defassa	M	17/02/1950	DRC
WB_3k	294	RMCA	Defassa	F	NA	DRC
WB_3l	520	RMCA	Defassa	M	07/05/1903	DRC
WB_4a	1237	RMCA	Defassa	F	07/08/1912	DRC
WB_4b	2458	RMCA	Defassa	NA	19/12/1913	DRC
WB_4c	13662	RMCA	Defassa	M	03/11/1936	DRC
WB_4d	36523	RMCA	Defassa	F	11/07/1952	DRC
WB_4e	5516	RMCA	Defassa	M	NA	DRC
WB_4f	5517	RMCA	Defassa	M	NA	DRC
WB_4g	16852	RMCA	Defassa	M	01/03/1938	DRC
WB_4h	16853	RMCA	Defassa	M	01/03/1938	DRC
WB_4i	22472	RMCA	Defassa	NA	01/09/1954	Rwanda
WB_4j	R6 78-30-M	RMCA	Defassa	NA	30/08/1978	Rwanda
WB_4k	2177	RMCA	Common	F	05/05/1913	Kenya
WB_4l	2190	RMCA	Common	M	11/05/1913	Kenya

3.2.2. Historical DNA extraction

DNA extraction of the historical skin samples was firstly optimised to improve final DNA yield and quality, and to reduce laboratory contamination. This involved setting up a pre-PCR laboratory, where no PCR equipment or work was completed, and a dedicated sterile workspace for undertaking historical DNA extractions. To do this, benches were thoroughly and frequently cleaned with a bleach-based product to denature DNA, and separate consumables were used. Plasticware was sterilised in an autoclave and placed under UV light before being used. Solutions made in the laboratory were autoclaved, filter-sterilised, and placed under UV light before use. This minimised the likelihood that DNA from modern samples, or PCR products, contaminated the historical DNA extractions.

The optimised DNA extraction protocol for historical skin samples is described below and adapted from McDonough et al., 2018; Roycroft et al., 2021 and Molecular Cloning Vol I. Skin samples were placed on a sterile cutting surface and small sections (5-10 mm) were cut with sterile scalpels and forceps. The sample was cleaned briefly in dilute bleach solution (approximately 1% dilution) and then washed immediately with sterile MilliQ water, before being placed into a sterile Eppendorf tube containing UltraPure water and incubated at RT. To limit any DNA contamination of the next sample, gloves and the cutting surface were replaced after each sample, and the scalpel and forceps were washed in a diluted bleach solution before rinsing with distilled water. The surface was also cleaned with a bleach-based product. Once all samples were prepared into Eppendorf tubes, the incubation timer was set, and the water was replaced at least three times the first day to remove any residual bleach and then subsequently incubated overnight to rehydrate the skin.

The following day samples were cut into smaller pieces and any hair removed using sterile scalpels and forceps. We followed a similar set up as before, with the same sterile precautions taken. Fragmented skin was placed into sterile Eppendorf tubes and incubated in UltraPure water to rehydrate during the remained of the day and then overnight.

On the third day of the protocol, the water was removed from each Eppendorf and replaced with 320 µl of lysis buffer containing 100 mM Tris-HCl (pH 8.000), 5 mM EDTA (pH 8.000), 200 mM NaCl, and 0.200% SDS (Z. Wang & Storm, 2006), 40 µl of proteinase K (20 mg/ml), and 40 µl of 1M DTT. Samples were then vortexed for 10 sec and briefly

centrifuged. Parafilm was placed on tube lids to limit evaporation, and the tubes were incubated for 48 hours at 56°C, with occasional vortexing to lyse samples and redistribute the lysis buffer.

After 48 hours, the samples were briefly centrifuged to pellet any remaining debris (e.g. 3 min at 5000 x g). Sterile QIAGEN MaXtract tubes were centrifuged for 20 sec at 16,000 x g immediately before use. The contents of the sample were pipetted into a QIAGEN MaXtract tube (maximum of 500 µl) and 1 volume of phenol:chloroform:isoamyl alcohol (24:25:1) was added to the tube and manually mixed for around 30 sec to form an homogenous solution. The tube was then centrifuged for 5 min at 16,000 x g. The upper aqueous phase was pipetted into a sterile Eppendorf. DNA was precipitated by adding an equal volume of isopropanol, thoroughly mixed, and centrifuged for 15 min at 13,000 x g. Isopropanol was carefully removed and DNA was washed with 70% ethanol, then centrifuged for 10 min at 13,000 x g. Ethanol was removed and the pellet was partially dried in a fume cupboard. Between 50 and 100 µl of TE buffer was added to the DNA and vortexed briefly or incubated overnight at 4°C to completely dissolve the pellet.

DNA was measured for quantity and quality with NanoDrop and Qubit, and gel electrophoresis (1% agarose) was used to measure fragment lengths. Some DNA samples required further precipitation to improve the purity. This was carried out by adding 1/10 volume of 3 M Na-Acetate (pH 5.200) and 2 volumes of 100% ethanol to the DNA sample, incubating at -20°C for at least one hour, centrifuging at max speed for 30 min, and then two washes with 70% ethanol. DNA was resuspended in TE buffer, and quality controlled as described above.

3.2.3. Historical DNA sequencing

Quality controlled DNA was sent for sequencing at Novogene. DNA was further quality controlled and only samples that passed these checks underwent Illumina library preparation and whole genome sequencing (WGS) on an Illumina NovaSeq6000, with a 2 x 150 bp read length and 350 bp insert size. Samples were either sequenced at 1X coverage of the waterbuck genome (the trial run with sample 1b) or 5X (the final run with 24 samples). Sequencing data was returned in FASTQ format.

3.2.4. Quantifying the percentage of waterbuck DNA with qPCR

Because DNA extracted from museum samples can contain varying amounts of endogenous DNA, we estimated the amount of endogenous nuclear DNA of the target

species in each sample before sending them for sequencing. To do so, we compared the concentration of DNA in each sample to the concentration in the WGS trial run sample.

Primers were designed to specifically target waterbuck nuclear DNA. Firstly, waterbuck gene sequences were searched using NCBI Nucleotide and the gene MGF (stem cell factor MGF gene) was selected. Primer-BLAST (NCBI) was then used, with the parameter for PCR product size set between 50 bp and 100 bp, and the best primer pair was selected (F- GGAAGCAGGCCTGGAAAGTA and R- GTCAGTGTTGCATGGCGATT) based on the PCR product size, similar annealing temperatures (F- 57.200°C and R- 56.300°C), and specificity to waterbuck. The primer pair showed some specificity to other bovid species; however, this was acceptable for the requirement of this qPCR study.

Historical DNA samples were then standardised to approximately 10 ng for qPCR quantification. Two reactions were prepared for each sample containing 1x Thermo Fisher Scientific PowerTrack SYBR Green, 500 nM of each primer, 10 ng DNA, and UltraPure water in a 96-well plate. The reaction was then run on a QuantStudio 3 qPCR machine with an enzyme activation stage of 10 min at 95°C, followed by 40 cycles of 15 sec at 95°C and 1 min at 60°C. The percentage of endogenous DNA was approximately estimated by comparing the plots of each sample to those of sample 1b, which was sequenced to 1X coverage. Two negative controls were also used and showed no amplification signal.

3.2.5. Quality control and mapping of WGS data

FASTQ files of WGS data were quality controlled with FASTQC v0.11.9 and MultiQC v1.0.dev0 (Ewels et al., 2016). Illumina adapters (AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT and GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGATGACTATCTCGTATGCCGTCTTCTG CTTG) were trimmed from paired-end reads using AdapterRemoval v2.3.3 (M. Schubert et al., 2016) with the parameters --mm 3, --collapse, --collapse-conservatively, --trimns, and --trimqualities. This created FASTQ files for both pair-end reads and reads that were collapsed due to DNA fragments sizes smaller than the specified insert size. Both types of reads were aligned separately with BWA-MEM v0.7.17 (H. Li, 2013), using the chromosome-level waterbuck genome as reference. Mate information was fixed in the BAM files with Picard v3.0.0 FixMateInformation (https://broadinstitute.github.io/picard/) and Samtools v1.6 calmd (Danecek et al.,

2021) was used to add further BAM file tags. These files were then validated with Picard v3.0.0 ValidateSAMFile to check for errors in formatting.

PCR duplicates from DNA library preparation were removed in BAM files using Picard v3.0.0 MarkDuplicates (--REMOVE_DUPLICATES) in paired-end reads and from collapsed reads using the PALEOMIX v1.3.7 rmdup_collapsed script with parameter '--remove-duplicates' (M. Schubert et al., 2014). DNA damage, which is higher in historical and ancient samples, was checked with mapDamage v2.2.0 (Jónsson et al., 2013). This program also rescales mapping quality scores in BAM files, to take into account DNA damage. BAM files were corrected to ensure paired ends were properly aligned with the parameter -f 0x2 in Samtools v1.6 view (Danecek et al., 2021). The paired-end and collapsed mapped reads for a given sample were then merged into one BAM file using Samtools v1.6 merge. This merged mapping file was then quality controlled with Samtools v1.6 depth and stats (Danecek et al., 2021), as well as Qualimap v2.2.2-dev (Okonechnikov et al., 2016).

3.2.6. Modern sequencing data

Additional WGS FASTQ files from 119 modern waterbuck samples (X. Wang et al., 2024) were also mapped to the waterbuck chromosome-level genome assembly by collaborators at the University of Copenhagen (Dr Xi Wang and Dr Rasmus Heller). This was carried out using a modified version of the PALEOMIX bam pipeline (M. Schubert et al., 2014). The modern WGS data included 10 populations, of which six populations were defassa and included 15 QENP (Uganda) samples, five KVNP (Uganda), 12 Kafue (Zambia), seven Maswa (Tanzania), 19 Ugalla (Tanzania), and 13 Samole (Ghana; **Figure 33**). The remaining four waterbuck populations were common and included 18 Matetsi (Zimbabwe) samples, four Luangwa (Zambia), 17 Samburu (Kenya) and nine Nairobi (Kenya). Combined with the historical data, this totalled 143 individuals across the waterbuck's distribution (**Figure 34**).

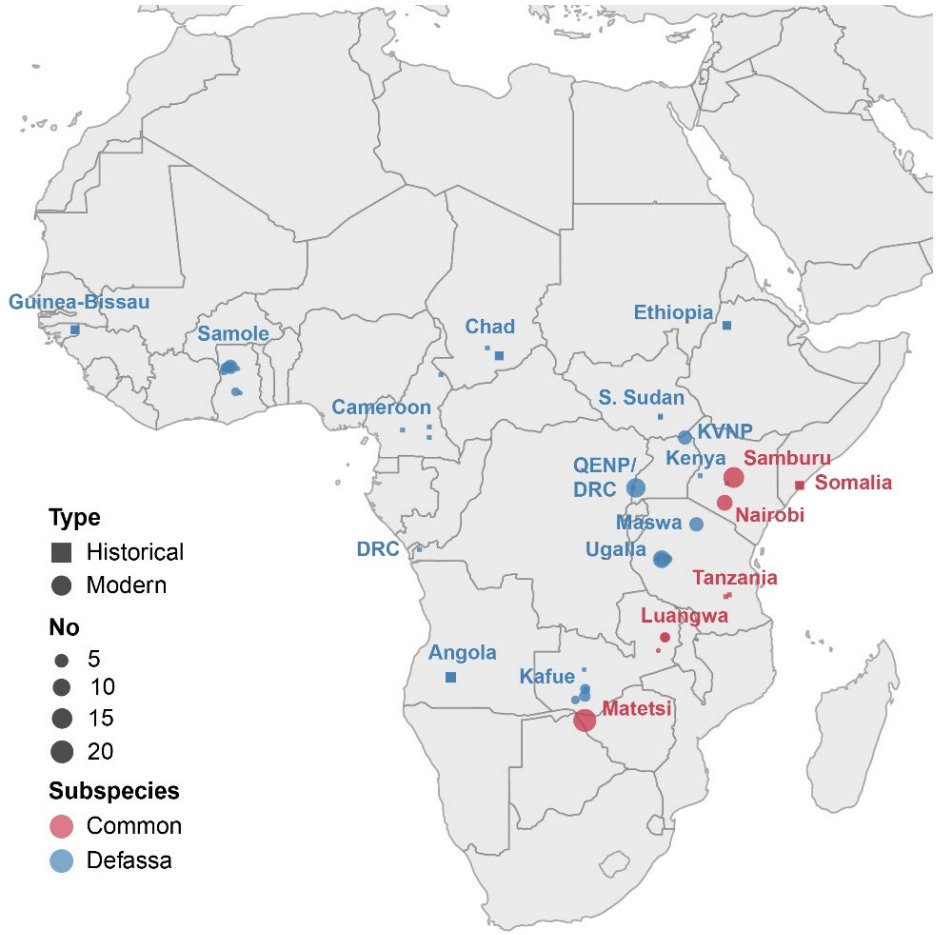


Figure 34: *Sampling map of the historical (n=24) and modern (n=119) whole genome sequencing (WGS) data.*

3.2.7. Filtering mapping data

To obtain reliable genotype likelihoods and perform further analysis, we filtered all alignment files and the reference genome to obtain high confidence mapping data. First, all BAM files were filtered using a customised python script looking at several SAM flags. The script removed unmapped reads (0x4), reads with an unmapped mate (0x8), secondary (0x100) and supplementary alignments (0x800), PCR duplicates (0x400), reads that failed QC (0x200), reads with insert sizes less than 50 bp in length or greater than 1000 bp, reads less than 50 bp in length or where less than 50% of the read aligned, and reads that mapped to different chromosomes or had unexpected orientations. Then, both the reference genome and BAM files were filtered to remove challenging regions that may lead to biases when genotyping (as described in Pečnerová et al., 2021). This used scripts from https://github.com/popgenDK/sites_filters in a stepwise manner as follows:

1. Unplaced scaffolds and sex chromosomes were removed from the BED file containing all genomic sites, leaving only sites found on autosomal chromosomes.
2. Repetitive sites were then documented using a python script that identified lowercase bases from the soft-masked waterbuck FASTA file. These repetitive sites were either removed from the BED file (leaving only non-repetitive sites) or included (leaving only repetitive sites), generating two BED files.
3. Both non-repetitive and repetitive BED files were then filtered further. Regions with high excess heterozygosity were identified in the historical samples and genotype likelihoods were calculated for common polymorphic sites (MAF ≥ 0.050 and SNP p value < 10−6) with a base quality ≥30. Per-site inbreeding coefficients (*F*) were calculated with PCAngsd v0.99 (Meisner & Albrechtsen, 2018), and a Hardy-Weinberg Equilibrium (HWE) likelihood ratio was performed to account for population structure. Individual allele frequencies were then computed using three principle components with PCAngsd v0.99 (Meisner & Albrechtsen, 2018). We excluded sites, and 10 Kb windows around sites, with significant heterozygosity estimates ($F < -0.900$ and p value $< 10^{-6}$).
4. The historical samples were also used to detect regions that had low depth of coverage by calculating the global depth, across samples, for each site in the genome using ANGSD v0.940 (Korneliussen et al., 2014). Sites below the lower 1% percentile and above the upper 1% percentile were removed from the BED files.
5. Mappability scores for each site in the reference genome were estimated using GenMap v1.3.0 (Pockrandt et al., 2020) in 100 bp k-mers, with a total of 2 mismatches (-K 100 -E 2) allowed, and sites with a score < 1 were removed from both BED files.

These steps resulted in two BED files; one containing filtered non-repetitive sites and the other containing filtered repetive sites. These BED files were used when calculating genotype likelihoods for all samples and in all population genomic analyses.

3.2.8. Error rates

Error rates were estimated for the historical samples using the “perfect individual” approach (described in Orlando et al., 2013) to filter samples before further population genomic analyses. This method required samples to be mapped to an outgroup species

and therefore the goat reference genome (*Capra hircus*; ARS1.2) was chosen. The 24 individuals were mapped as above to this reference genome. Sample WB_1b was selected as the "perfect individual", having the highest genomic coverage.

A consensus FASTA sequence was created for the most common base at each position with ANGSD v0.940 (Korneliussen et al., 2014) using the goat reference genome and the parameters -doCounts 1, -doFasta 2, -minMapQ 30, and -minQ 20. Error rates were then calculated in ANGSD v0.940 (Korneliussen et al., 2014) with -doAncError, -minMapQ 30, and -minQ 20, using the goat reference genome as the ancestral state (-anc) and the FASTA file of the "perfect individual" as the reference (-ref). Specifically, the error rates were calculated as an excess or deficit of derived alleles from the outgroup compared to the "perfect individual".

Additionally, overall heterozygosity per sample for both mapping datasets (waterbuck and goat references) was calculated on a reduced number of genomic sites. ANGSD v0.940 (Korneliussen et al., 2014) was used to calculate genotype likelihoods with parameters -gl 2, -dosaf 1, -minMapQ 30, and -minQ 20. Site allele frequency (SAF) files were converted into site frequency spectrum (SFS) files with winsfs (Rasmussen et al., 2022) using parameter -s 1, and heterozygosity was calculated from SFS files with parameter -s heterozygosity.

3.2.9. Population structure

To explore the population structure of waterbuck samples, Principal Component Analyses (PCAs) were computed. Genotype likelihoods were called with ANGSD v0.940 (Korneliussen et al., 2014) on all filtered sites for each chromosome with the parameters -GL 2, -doMajorMinor 1, -doMaf 1, -SNP_pval 1e-6, -minMaf 0.05, -minMapQ 30, and -minQ 20. Genotypes were also called for only filtered transversion sites (-rmtrans 1). This produced Beagle files which were combined for all chromosomes and samples, and inputted into PCAngsd v0.99 (Meisner & Albrechtsen, 2018) with -minMaf 0.05. Admixture proportions were also calculated from these Beagle files using ANGSD v0.940 NGSadmix (Korneliussen et al., 2014) between K=2 and K=20, however only K=2, K=3, K=4, and K=12 populations were included in this study.

Neighbour-Joining (NJ) phylogenetic trees were constructed with ANGSD v0.940 (Korneliussen et al., 2014) and the parameters -GL 2, -minMapQ 30, -minQ 20, -doMajorMinor 1, -doMaf 1, -SNP_pval 1e-6, -doIBS 1, -doCounts 1, -doCov 1, -

makeMatrix 1, and -minMaf 0.05. The pairwise IBS matrix was visualised with a custom R script.

Estimated Effective Migration Surfaces (EEMS) were computed by firstly computing genotype likelihoods using ANGSD v0.940 (Korneliussen et al., 2014) with -GL 2, -minMapQ 30, -minQ 20, -doMajorMinor 1, -doMaf 1, -SNP_pval 1e-6, -doIBS 1, -doCounts 1, -doCov 1, -makeMatrix 1, and -minMaf 0.05. This output was then used in the EEMS runeems_snps pipeline (Petkova et al., 2015) which was run three times with the options nIndiv 143, nSites 143, nDemes 600, diploid, numMCMCIter 20000000, numBurnIter 10000000, and numThinIter 9999. The output was then visualised with the R package rEEMSplots.

3.2.10. Heterozygosity

Overall genome-wide heterozygosity per sample was computed. ANGSD v0.940 (Korneliussen et al., 2014) was used to call genotype likelihoods -doSaf 1, -GL 2, -minMapQ 30, and -minQ 20, which produced SAF files. The waterbuck reference genome was used as the ancestral state. We also estimated heterozygosity without transversions using the parameter -noTrans 1. SAF files were converted into SFS files using ANGSD v0.940 realSFS (Korneliussen et al., 2014) with a folded spectrum (-folded). This was visualised in R with ggplot2.

3.2.11. Genomic differentiation and diversity

We estimated genomic differentiation between the two subspecies of waterbuck (*K. e. defassa* and *K. e. ellipsiprymnus*) and between population groups using the fixation index, F_{ST}. Genotype likelihoods were computed with ANGSD v0.940 (Korneliussen et al., 2014) for each subspecies on all filtered sites (-GL 2, -doSaf 1, -minMapQ 30, and -minQ 20) with the waterbuck genome as the ancestral, the folded SFS files were generated with ANGSD v0.940 realSFS (-fold 1), and fst index (-whichFst 1 and -fold 1) and fst stats2 was run with a window size and step size of 10 Kb. We then visualised the window F_{ST} values, with the addition of synteny data to cattle (as described in chapter 2), using a custom R script with ggplot2 and qqman. Lastly, gene ontology statistical overrepresentation tests were computed on lists of waterbuck genes using Panther (Mi et al., 2019; Thomas et al., 2022) and cattle (*Bos taurus*) as reference.

3.3. Results

3.3.1. DNA extraction of museum samples

DNA was extracted from a total of 48 waterbuck museum skin samples, with 10 samples extracted at a time. A negative control was used throughout each extraction to check for contamination during the protocol. The DNA quality measured with NanoDrop varied between samples (**Table 11**). DNA extracted from samples from the PCM (WB_1a to WB_2l) had overall better 260/280 and 260/230 ratios than samples from the RCMA (3a to 4l). DNA yield measured with NanoDrop, in a total volume of 40 µl, also varied between samples, ranging from 1.600 ng to 240,940.400 ng. The negative controls had DNA concentrations between 0.900 ng/µl and 78.490 ng/µl.

However, when measuring with the more sensitive Qubit DNA yield decreased in each sample, suggesting the presence of chemical contaminants. DNA yield was between 11 ng and 4,640 ng. As Qubit requires the use of two DNA standards to calculate DNA concentration, samples with a DNA concentration below 0.1 ng/µl and above 120 ng/µl (for the High Sensitivity Assay Kit) cannot be accurately measured. Therefore, these samples were denoted as being below ('<') or above ('>') this DNA concentration range. All negative controls were measured as having DNA concentrations below 0.1 ng/µl when using Qubit, again suggesting that the DNA measured with NanoDrop was susceptible to false readings due to chemical contaminants and that there was no contamination of DNA during the DNA extraction steps.

Table 11: *DNA quality control of waterbuck museum samples. DNA concentrations that were too low or too high to be measured accurately with Qubit are denoted as '<' and '>', respectively. Negative controls (NC) were used for each of the four extractions.*

		NanoDrop				Qubit	
ID	**Vol. (µl)**	**260/280**	**260/230**	**ng/µl**	**Yield (ng)**	**ng/µl**	**Yield (ng)**
WB_1a	40	1.890	1.490	377.500	15,100.000	83.400	3,336
WB_1b	40	1.890	2.110	3,411.600	136,464.000	116.000	4,640
WB_1c	40	1.890	1.920	2,970.100	118,804.000	112.000	4,480
WB_1d	40	1.880	1.890	907.700	36,308.000	102.000	4,080
WB_1e	40	1.910	2.010	304.900	12,196.000	102.000	4,080
WB_1f	40	1.900	2.140	2,183.400	87,336.000	96.000	3,840

WB_1g	40	1.880	1.990	671.600	26,864.000	100.000	4,000
WB_1h	40	1.900	1.850	1,295.800	51,832.000	108.000	4,320
WB_1i	40	1.880	1.970	886.800	35,472.000	97.400	3,896
WB_1j	40	1.910	1.930	471.000	18,840.000	91.400	3,656
WB_1k	40	1.910	1.880	470.600	18,824.000	98.800	3,952
WB_1l	40	1.860	1.130	47.200	1,888.000	22.000	880
NC 1	40	1.870	0.400	0.900	36.000	<	<
WB_2a	40	1.800	2.04	4,570.900	182,836.000	>	>
WB_2b	40	1.750	1.370	154.900	6,196.000	59.600	2,384
WB_2c	40	1.810	2.000	4,252.070	170,082.800	>	>
WB_2d	40	1.870	1.980	2,525.990	101,039.600	>	>
WB_2e	40	1.550	0.820	19.500	780.000	7.560	302
WB_2f	40	1.850	1.850	882.730	35,309.200	102.000	4,080
WB_2g	40	1.850	1.940	957.150	38,286.000	>	>
WB_2h	40	1.850	1.970	1,846.040	73,841.600	>	>
WB_2i	40	1.840	1.940	2,835.240	113,409.600	>	>
WB_2j	40	1.860	2.140	1799.580	71,983.200	>	>
WB_2k	40	1.850	2.090	1626.990	65,079.600	>	>
WB_2l	40	1.840	2.030	2947.330	117,893.200	>	>
NC 2	40	1.510	2.000	78.490	3,139.600	<	<
WB_3a	40	1.350	0.160	5.960	238.400	<	<
WB_3b	40	1.640	0.530	28.230	1,129.200	1.840	74
WB_3c	40	-1.550	0.200	1.630	65.200	0.284	11
WB_3d	40	1.750	0.360	16.490	659.600	2.200	88
WB_3e	40	1.850	1.230	35.200	1,408.000	7.300	292
WB_3f	40	1.710	1.170	119.080	4,763.200	24.800	992
WB_3g	40	-4.420	0.080	0.810	32.400	<	<
WB_3h	40	1.560	0.950	20.240	809.600	3.000	120
WB_3i	40	-0.080	0.010	0.040	1.600	<	<
WB_3j	40	1.830	1.490	221.040	8,841.600	45.400	1,816
WB_3k	40	1.800	1.880	2,342.560	93,702.400	>	>
WB_3l	40	1.240	0.420	50.940	2,037.600	<	<
NC 3	40	1.450	0.700	17.970	718.800	<	<

WB_4a	40	1.390	1.050	8.180	327.200	1.420	57
WB_4b	40	1.170	0.420	7.060	282.400	<	<
WB_4c	40	1.810	1.780	367.810	14,712.400	92.400	3,696
WB_4d	40	1.820	1.860	4,140.080	165,603.200	>	>
WB_4e	40	1.200	0.350	17.840	713.600	<	<
WB_4f	40	1.260	0.360	9.820	392.800	26.800	1072
WB_4g	40	1.780	1.680	65.540	2,621.600	<	<
WB_4h	40	1.400	0.420	6.930	277.200	3.240	130
WB_4i	40	1.350	0.340	47.760	1,910.400	<	<
WB_4j	40	1.530	1.260	55.040	2,201.600	7.840	314
WB_4k	40	1.260	0.530	3.830	153.200	<	<
WB_4l	40	1.460	6.010	42.100	1,684.000	1.080	43
NC 4	40	1.140	1.020	5.900	236.000	<	<

A total of 26 DNA samples that passed the NanoDrop and Qubit quality control were sent to Novogene (UK) for WGS. Here, the samples were additionally quality controlled for DNA quantity with Qubit (**Table 12**), as well as the measurement of DNA fragment lengths with gel electrophoresis (**Figure 35**). All samples had smeared DNA bands, with some samples having DNA fragment lengths up to 10,000 bp (e.g., 3/WB_1b, 7/WB_1g, and 9/WB_1i), whilst samples 24/WB_3k, 25/WB_4c, and 26/WB_4d had low DNA fragment lengths of less than 500 bp. From the gel, it was apparent that none of the DNA samples required fragmentation during WGS library preparation.

Table 12: *Quantification of the passed DNA samples sent for whole genome sequencing (WGS) with Qubit.*

			Qubit	
No.	**ID**	**Vol. (µl)**	**ng/µl**	**DNA Yield (ng)**
1	WB_1a	20	117.000	2,340.000
2	WB_1b	19	365.000	6,935.000
3	WB_1c	18	674.000	12,132.000
4	WB_1d	20	268.000	5,360.000
5	WB_1e	20	102.000	2,040.000
6	WB_1f	19	708.000	13,452.000
7	WB_1g	20	286.000	5,720.000
8	WB_1h	20	282.000	5,640.000
9	WB_1i	22	339.000	7,458.000
10	WB_1j	19	162.000	3,078.000
11	WB_1k	20	159.000	3,180.000
12	WB_1l	17	9.500	161.500
13	WB_2a	19	742.000	14,098.000
14	WB_2b	19	22.500	427.500
15	WB_2c	19	696.000	13,224.000
16	WB_2d	19	580.000	11,020.000
17	WB_2f	21	262.000	5,502.000
18	WB_2g	20	297.000	5,940.000
19	WB_2h	22	399.000	8,778.000
20	WB_2i	19	387.000	7,353.000
21	WB_2j	23	312.000	7,176.000
22	WB_2k	22	420.000	9,240.000
23	WB_2l	18	658.000	11,844.000
24	WB_3k	20	172.000	3,440.000
25	WB_4c	18	15.800	284.400
26	WB_4d	19	306.000	5,814.000

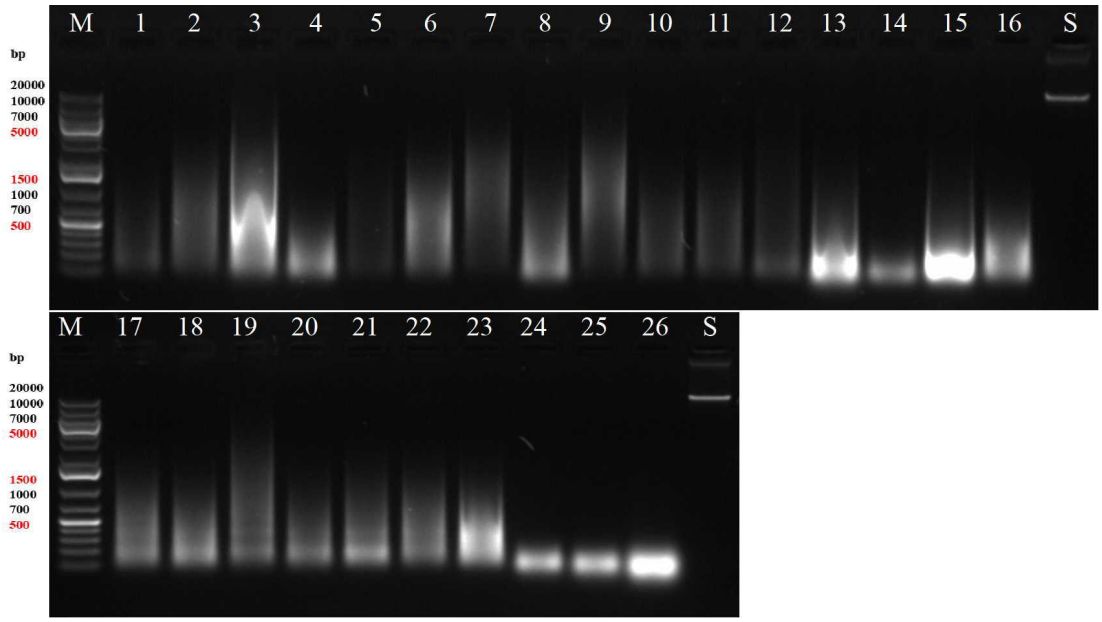


Figure 35: *Gel electrophoresis of historical DNA samples that passed quality control to assess DNA fragment lengths. DNA was run on a 1% agarose gel at 100 V for 40 min. M refers to the DNA ladder, with sizes of key DNA bands given in bp, and the numbers above lanes refer to the sample number (No. in* ***Table 12****).*

3.3.2. Testing the whole-genome resequencing and bioinformatics pipeline

Before sequencing all samples, we resequenced one sample (WB_1b) at approximately 1X coverage (of a 3 Gb genome) as a trial to see if the DNA from an historical museum sample that passed QC was able to be sequenced successfully. A total of 12,657,961 paired-end reads were successfully sequenced with a mean length of 150 bp, and 8,343,746 paired-end reads left after trimming adapters, stretches of low-quality bases and/or N's, and collapsing overlapping mates (**Table 13**).

Table 13: *Museum sample WB_1b_1X resequencing quality control before and after trimming. Sample name refers to the sample (WB_1b), requested sequencing coverage (1X), and the read pair (R1 or R2).*

	Before Trimming		After Trimming	
Sample	**Duplicates (%)**	**Total Reads**	**Duplicates (%)**	**Total Reads**
WB_1b_1X_R1	28.195	12,657,961	28.119	8,343,746
WB_1b_1X_R2	27.166	12,657,961	26.853	8,343,746

After mapping the reads to the waterbuck reference genome, 20.491% of reads were identified as being duplicates, and a total of 79.291% of sequences successfully mapped, equating to an average coverage of 0.625X. Mean mapping quality was 36, the error rate was 0.007, and mean insert size was 354.300 bp. Additionally, DNA damage was estimated bioinformatically using the program mapDamage (Jónsson et al., 2013). The sample had low levels of DNA damage, with the highest rate of C to T substitutions at 0.020, at the start of reads (**Figure 36**). Whereas G to A substitutions were below 0.050 and remained more constant at the start and ends of reads.

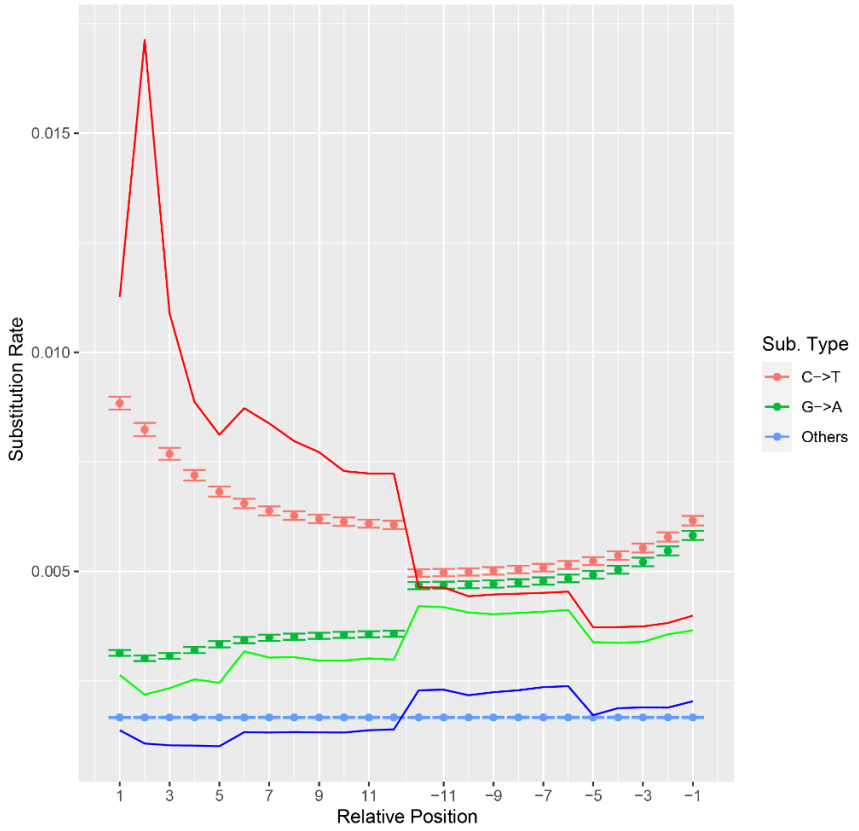


Figure 36: *Estimation of DNA damage in sequencing sample WB_1b_1X. Empirical misincorporation frequencies (solid line) and 95% simulated posterior predictive intervals from the fitted model (confidence intervals), grouped by substitution type (Sub. Type). Relative position is the start (1-12 bp) and end (-11--1) of each read.*

3.3.3. Endogenous DNA quantification with qPCR:

It is well known that DNA extractions from historical and museum samples might contain both endogenous and exogenous DNA (e.g., Molbert et al., 2023). To quantify the amount of endogenous DNA in each of the museum samples, we used qPCR and firstly required a pair of primers with specificity to waterbuck. We chose the nuclear gene MGF and

designed primers to amplify an 80 bp region. We performed two qPCRs per sample in the 24 historical DNA samples, and then their melt and amplification curves were compared. As sample WB_1b was resequenced to 1X coverage and had a mean coverage of 0.625X we could compare the other waterbuck samples to this, and approximately estimate the coverage we would expect from WGS these samples. For the amplification curve, most samples had a log phase at a lower number of qPCR cycles than WB_1b (**Figure 37**), suggesting that they contained more endogenous DNA and would sequence to a higher coverage. Whereas samples WB_1a, WB_1e, WB_1h, WB_1j, WB_2j, WB_3k, and WB_4d required a greater number of qPCR cycles to reach the log phase, and therefore may have less endogenous DNA than WB_1b and this could lead to less coverage when sequenced. Both reactions for WB_2b did not amplify, which suggests there may have been an issue with the preparation of the sample rather than a lack of endogenous DNA.

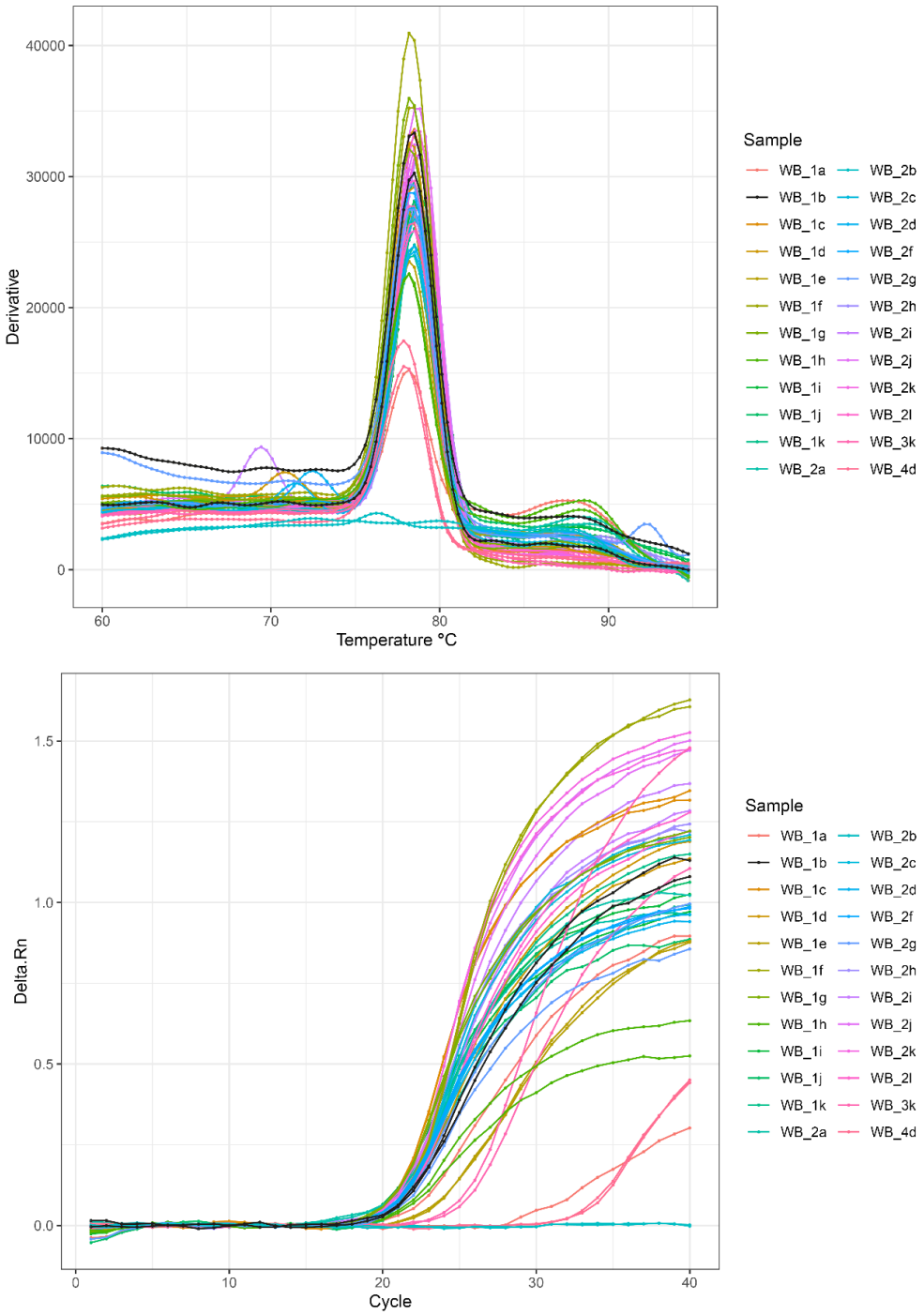


Figure 37: *qPCR melt curve (top) and amplification curve (bottom) for each waterbuck museum sample. The black line is sample WB_1b which was resequenced to 1X coverage and used as a comparison to all other samples.*

3.3.4. Whole-genome resequencing and mapping of museum samples

Of the 26 samples sent for WGS library preparation, 24 samples were resequenced to 5X coverage. Two samples performed poorly during the library preparation and were not able to be sequenced (WB_4c and WB_4d). Both these samples had low DNA fragment

lengths (**Figure 35**) and were from RCMA, where the majority of samples failed the DNA extraction QC. Each samples paired end reads (R1 and R2) were quality controlled (**Table 14**). The percentage of duplicated reads ranged from 26.832% to 45.236%, whilst the number of paired-end reads was between 49,987,740 and 72,458,130. During adapter trimming the reads were split into either paired end reads containing the forward (R1) and reverse mates (R2), or the collapsed reads (COL) where the two mates overlapped. WB_3k_5X had the highest percentage of collapsed reads at 90.781%, whereas sample WB_2f_5X had the lowest at 25.269%.

Samples were then mapped to the chromosome-level waterbuck genome, duplicates were removed, and the mapping data was quality controlled (**Table 15**). The percentage of duplicate reads ranged from 14.196% in WB_1c_5X to 27.559% in WB_3k_5X, whilst the final percentage of mapped reads compared with the total reads after trimming ranged from 35.780% in WB_2b_5X to 85.473% in WB_2d_5X. Whole genome coverage ranged between 1.587X in WB_2b_5X to 4.863X in WB_1b_5X. Mapping quality was between 36 and 37 for all samples and error rates were below 0.010. The mean insert size for the majority of samples was between 308.800 and 397.900, whilst sample WB_3k_5X had a mean insert size of 252.600, due to high levels of DNA degradation (**Figure 35**) and the collapsing of overlapping reads (**Table 14**).

Table 14: *Museum sample resequencing (5X) quality control before and after adapter trimming. Sample name refers to the sample (e.g., WB_1a) and the sequencing coverage (5X). R1 is the forward mate, R2 the reverse mate, and COL the collapsed reads. Dup. is the percentage of duplicated reads.*

	Before Trimming				After Trimming						
	R1		R2		R1		R2		COL		All
Sample	**Dup. (%)**	**Total Reads**	**Dup. (%)**	**Total Reads**	**Dup. (%)**	**Total Reads**	**Dup. (%)**	**Total Reads**	**Dup. (%)**	**Total Reads**	**Total Reads**
WB_1a_5X	33.405	61,233,926	32.331	61,233,926	30.912	18,484,611	29.419	18,484,611	32.320	42,699,336	79,668,558
WB_1b_5X	33.058	72,458,130	31.660	72,458,130	31.881	42,289,868	30.281	42,289,868	31.834	30,131,999	114,711,735
WB_1c_5X	29.554	55,837,218	28.565	55,837,218	28.763	22,857,924	27.243	22,857,924	27.811	32,939,036	78,654,884
WB_1d_5X	31.094	68,904,390	29.672	68,904,390	26.941	13,074,088	25.266	13,074,088	30.635	55,788,057	81,936,233
WB_1e_5X	30.964	62,972,744	29.429	62,972,744	29.553	30,340,222	27.683	30,340,222	29.710	32,607,423	93,287,867
WB_1f_5X	32.887	62,147,971	31.753	62,147,971	30.817	27,605,650	29.266	27,605,650	31.883	34,509,068	89,720,368
WB_1g_5X	27.792	49,987,740	26.832	49,987,740	27.130	23,433,867	25.825	23,433,867	26.103	26,541,495	73,409,229
WB_1h_5X	36.157	62,670,073	34.272	62,670,073	33.258	26,117,602	31.212	26,117,602	35.332	36,516,056	88,751,260
WB_1i_5X	32.710	68,658,154	31.981	68,658,154	31.557	26,208,190	30.264	26,208,190	31.466	42,417,823	94,834,203
WB_1j_5X	32.296	70,636,472	30.928	70,636,472	30.890	29,145,908	29.218	29,145,908	30.916	41,458,987	99,750,803
WB_1k_5X	29.893	56,126,487	28.479	56,126,487	27.982	26,682,355	26.143	26,682,355	29.226	29,423,197	82,787,907
WB_1l_5X	33.597	57,476,215	32.341	57,476,215	31.907	28,998,761	30.345	28,998,761	30.925	28,461,480	86,459,002
WB_2a_5X	35.543	63,579,419	34.826	63,579,419	31.600	16,043,824	30.269	16,043,824	34.873	47,505,380	79,593,028

WB_2b_5X	45.236	64,263,689	43.639	64,263,689	56.344	22,818,857	53.623	22,818,857	31.772	41,396,749	87,034,463
WB_2c_5X	36.623	50,120,036	35.758	50,120,036	35.090	10,886,594	33.488	10,886,594	35.033	39,206,997	60,980,185
WB_2d_5X	35.147	56,092,309	33.744	56,092,309	28.953	9,508,545	27.153	9,508,545	34.541	46,536,229	65,553,319
WB_2f_5X	35.010	69,426,285	33.636	69,426,285	34.998	41,406,959	33.386	41,406,959	32.010	28,001,760	110,815,678
WB_2g_5X	33.591	52,068,949	32.117	52,068,949	28.731	15,050,348	27.010	15,050,348	33.447	36,985,356	67,086,052
WB_2h_5X	34.634	61,829,195	33.188	61,829,195	33.536	32,077,773	31.216	32,077,773	32.766	29,722,543	93,878,089
WB_2i_5X	35.859	62,887,220	34.335	62,887,220	35.750	32,250,194	33.847	32,250,194	32.832	30,604,367	95,104,755
WB_2j_5X	37.739	55,338,597	36.163	55,338,597	36.123	24,045,565	34.120	24,045,565	35.667	31,258,844	79,349,974
WB_2k_5X	35.493	60,581,904	34.337	60,581,904	34.196	27,099,373	32.602	27,099,373	33.781	33,456,786	87,655,532
WB_2l_5X	36.188	60,598,317	34.993	60,598,317	32.993	17,274,567	31.374	17,274,567	35.259	43,284,966	77,834,100
WB_3k_5X	31.532	53,832,048	30.553	53,832,048	33.716	2,597,745	32.389	2,597,745	30.586	51,162,888	56,358,378

Table 15: *Mapping stats for the museum samples sequenced to 5X coverage. Percentage of duplicated reads (Dup. %), percentage of mapped reads after trimming out of the total reads (Map. %), mean mapping quality (Mean Map. Qual), and mean coverage (Mean Cov.).*

Sample	Dup. (%)	No. Reads Mapped	Map. (%)	Error Rates	Mean Map. Qual.	Mean Insert Size	Mean Cov.
WB_1a_5X	17.287	65,903,684	82.722	0.008	36.300	362.800	3.485
WB_1b_5X	19.861	90,865,404	79.212	0.009	36.200	381.200	4.863
WB_1c_5X	14.196	66,713,167	84.818	0.008	36.300	384.800	3.566
WB_1d_5X	17.908	68,013,977	83.008	0.008	36.500	342.400	3.591
WB_1e_5X	17.193	74,896,252	80.285	0.008	36.200	373.800	4.060
WB_1f_5X	16.578	74,502,607	83.039	0.008	36.300	370.800	4.019
WB_1g_5X	16.492	60,175,307	81.972	0.008	36.200	357.800	3.404
WB_1h_5X	17.579	71,895,902	81.008	0.008	36.300	373.200	3.818
WB_1i_5X	17.960	77,175,367	81.379	0.007	36.400	361.600	4.273
WB_1j_5X	19.639	77,956,143	78.151	0.008	36.300	359.500	4.293
WB_1k_5X	17.532	67,497,794	81.531	0.008	36.200	366.000	3.701
WB_1l_5X	17.972	61,339,405	70.946	0.008	36.300	376.900	3.295
WB_2a_5X	17.673	65,733,476	82.587	0.007	36.500	358.700	3.432
WB_2b_5X	18.722	31,140,538	35.780	0.008	36.500	308.800	1.587
WB_2c_5X	14.537	51,517,659	84.483	0.007	36.500	354.500	2.660
WB_2d_5X	16.348	56,030,220	85.473	0.008	36.500	348.600	2.822
WB_2f_5X	22.304	85,402,812	77.067	0.008	36.200	376.500	4.596
WB_2g_5X	16.033	56,263,004	83.867	0.008	36.400	359.900	2.987
WB_2h_5X	17.685	76,278,977	81.253	0.008	36.100	397.900	3.986
WB_2i_5X	20.289	75,279,994	79.155	0.010	36.300	379.300	3.988
WB_2j_5X	19.027	64,328,475	81.069	0.009	36.300	376.300	3.391
WB_2k_5X	19.643	69,735,481	79.556	0.010	36.300	365.200	3.802
WB_2l_5X	20.020	62,307,901	80.052	0.009	36.400	358.400	3.289
WB_3k_5X	27.559	42,693,426	75.753	0.010	36.600	252.600	1.952

3.3.5. Bioinformatic estimation of DNA damage in historical samples

The 24 historical samples were then assessed bioinformatically for DNA damage with mapDamage (Jónsson et al., 2013). DNA damage was measured by the rates of C to T, G to A, or other types of substitutions. C to T substitutions were the highest at the start of paired end reads, with WB_2h_5X having the lowest rate at 0.006 and WB_1d_5X having the highest rate at 0.026 (**Figure 38**). Whereas G to A substitutions were the highest at the end of collapsed reads, ranging from 0.09 to 0.044 in sample WB_2h_5X and WB_1d_5X, respectively. Other substitution types showed low levels at both the start and end of reads.

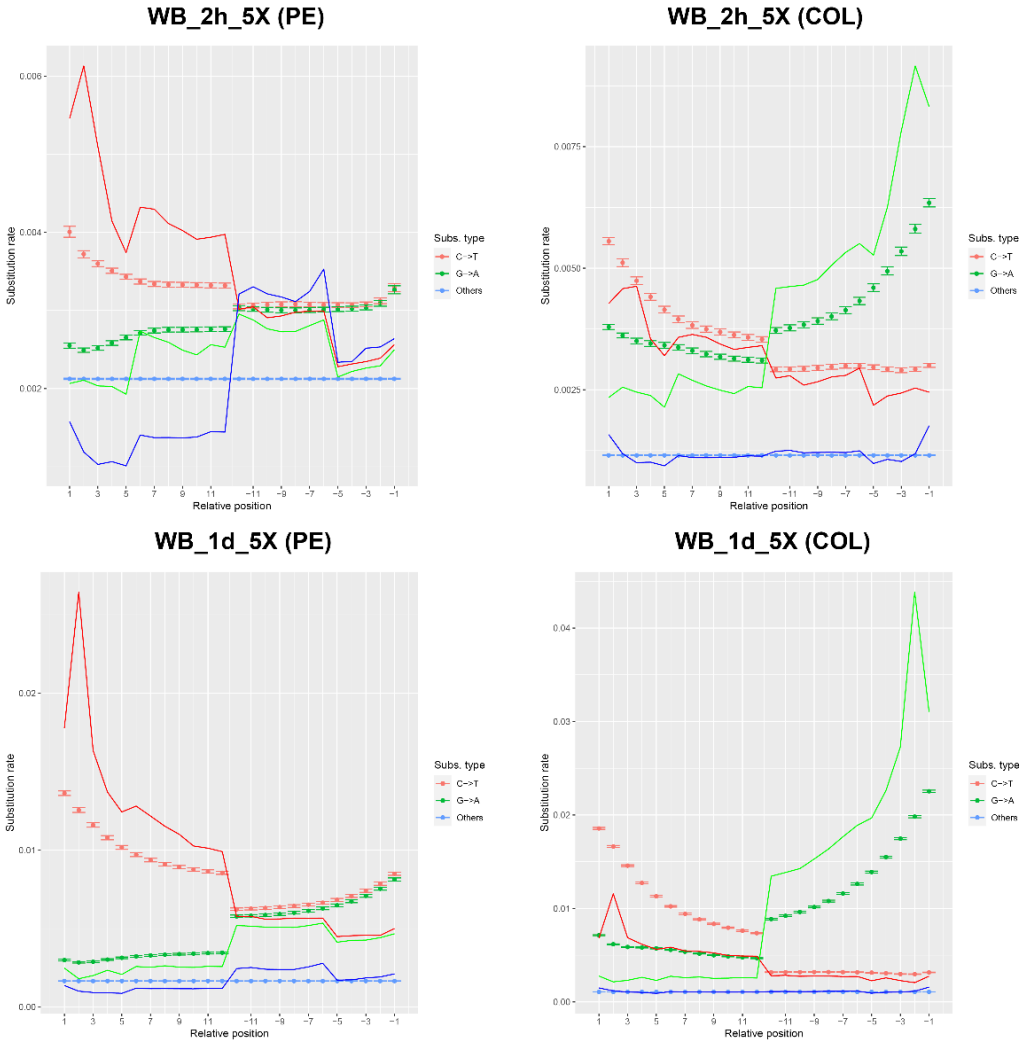


Figure 38: *Estimation of DNA damage in two historical samples with the lowest (WB_2h) and highest DNA damage (WB_1d), in the paired end (PE) and collapsed mate (COL) alignments. Empirical misincorporation frequencies (solid line) and 95% simulated posterior predictive intervals from the fitted model (confidence intervals), grouped by substitution type. Relative position is the start and end of each read.*

3.3.6. Error rates in historical samples

Before running any population genomic analyses, we firstly estimated error rates by calculating the excess or deficit of derived alleles from the outgroup species compared to the "perfect individual" (WB_1b), which had the highest whole genome coverage. Error rates were lowest in samples WB_1c_5X, WB_1g_5X, WB_1i_5X, WB_1l_5X, and WB_2h_5X (< 0.0005; **Figure 39**). Whereas samples WB_1d_5X and WB_2b_5X had higher error rates (> 0.0010). Sample WB_3k_5X had a considerably higher error rate to all other samples (> 0.0025), suggesting an issue with the sequencing of this waterbuck sample from the RMCA collection.

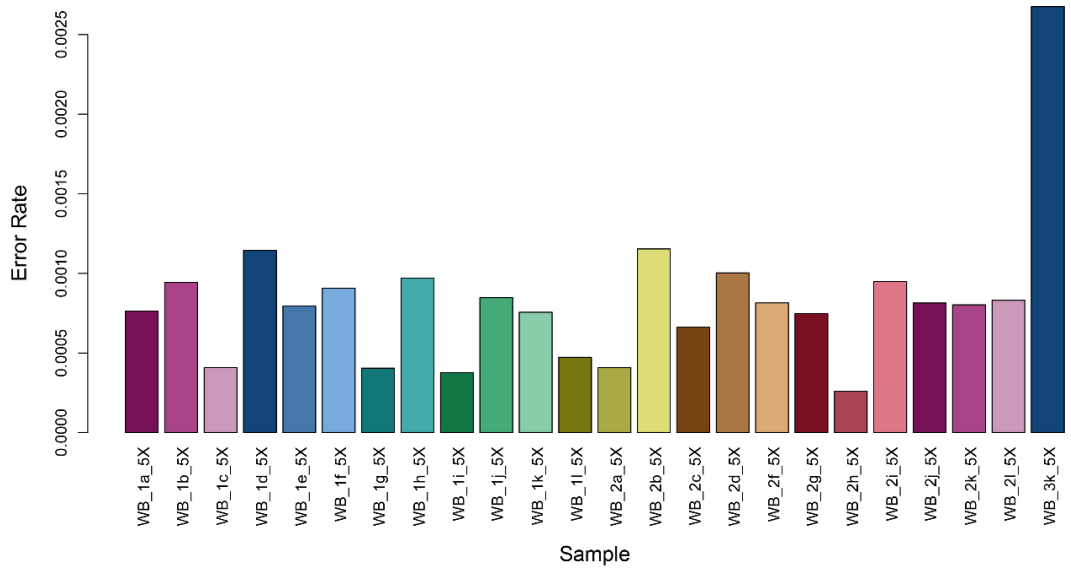


Figure 39: *Error rates in each of the historical samples using the "perfect individual" approach. Sample WB_1b_5X was selected as the "perfect individual".*

Error rates also correlated positively with heterozygosity in each of the historical samples (**Figure 40**), with sample WB_3k_5X also having very high levels of heterozygosity compared to all other samples. These results suggested that sample WB_3k_5X should be removed from further population genomic analyses. However, this sample was the only waterbuck sample from southern DRC, and so we decided to use this sample to assess population structure but remain cautious with the results.

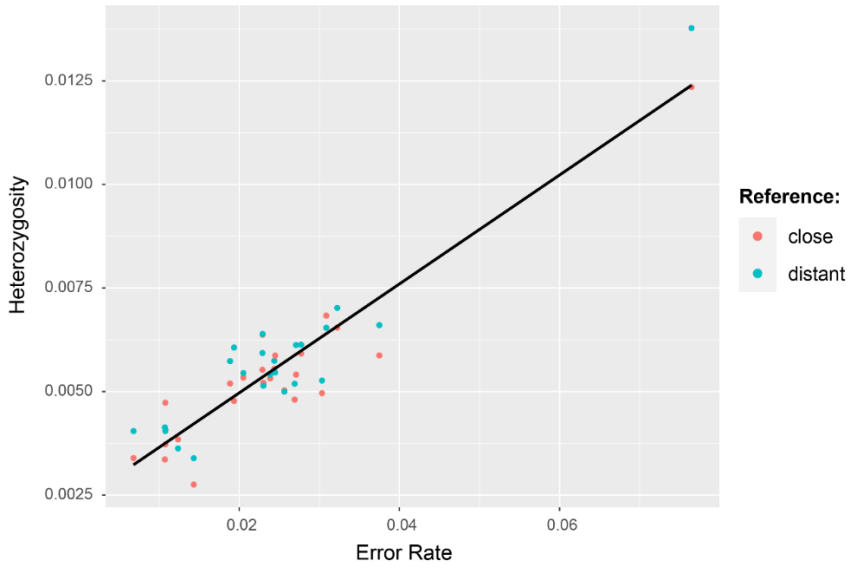


Figure 40: *Heterozygosity and error rates in the historical samples, using the "close" reference genome (waterbuck) or "distant" reference genome (goat).*

3.3.7. Genomic sites filtering

To avoid biases when genotyping low coverage WGS data, we conservatively removed genomic sites with repeats, low mappability, excess heterozygosity, low or high sequencing depth, and those that were found on the sex chromosome X or on unplaced scaffolds. Repeats were previously annotated and masked in Chapter 2, with a total of 1,728,574,685 bp or 54.801% of the genome masked as repetitive. As repeats make up the majority of the sites within the genome and would normally be filtered out (as in Pečnerová et al., 2021), we tested whether these repetitive sites could be informative for population genomic analyses if further filtered following the same steps as non-repetitive sites. Therefore, we ran the population genomics analyses using filtered non-repetitive genomic sites and also using filtered repetitive sites.

In a stepwise manner, we firstly retained only autosomes, resulting in 80.971% of the genome remaining (**Table 16**). In our first analysis, we kept only single-copy non-repetitive regions of the genome, and as such repeats were removed from the autosomes, resulting in 42.079% of sites remaining. Sites with excess heterozygosity (0.059%) were then filtered out, leaving 42.055% of filtered genomic sites. Genomic sites with low or high depth of coverage were also identified (20.652%; **Figure 41**) and were filtered out, resulting in 40.331% remaining. Lastly, sites in the reference genome with low mappability were identified totalling 19.834% (625,623,511 bp) of the genome, and these were filtered out, resulting in a final filtered genomic sites list consisting of only 40.185 % of the genome.

Because using single-copy, non-repetitive regions identified above excludes more than half of the genome, we performed the same filtering steps in the repetitive areas of the genome. For the repetitive sites, starting with only autosomes, 38.892% of the genome was filtered after removing non-repetitive sites, 38.870% filtered sites remained after removing those with excess heterozygosity, 34.692% when low or high depth sites were removed, and 32.933% when low mappability sites were removed (**Table 16**).

Table 16: *Genomic sites filtering steps (1-5) for the non-repetitive and repetitive sites. Order of filtering in a stepwise manner: 1. autosomes, 2. repeats, 3. heterozygosity, 4. depth, and 5. mappability. The number (bp) and the percentage of filtered genomic sites remaining are displayed after each filtering step (out of 3,154,302,869 bp).*

	Non-Repetitive		Repetitive	
Filtering Step	**Remaining Sites (bp)**	**Remaining Sites (%)**	**Remaining Sites (bp)**	**Remaining Sites (%)**
1 - Autosomes	2,554,066,657	80.971	2,554,066,657	80.971
2 – Rep/non-rep	1,327,283,576	42.079	1,226,783,081	38.892
3 – Heterozygosity	1,326,534,284	42.055	1,226,062,865	38.870
4 – Coverage	1,272,151,388	40.331	1,094,276,795	34.692
5 – Mappability	1,267,547,030	40.185	1,038,816,161	32.933

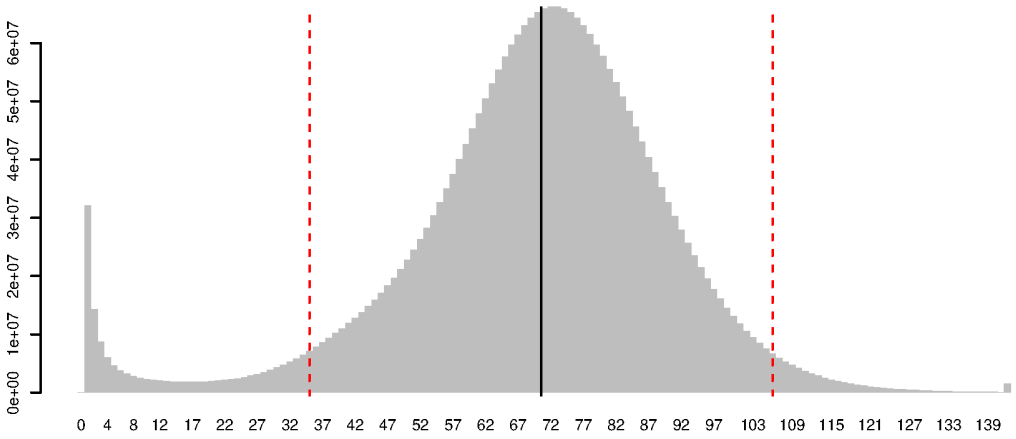


Figure 41: *Genomic sites filtering by sequencing depth. Histogram of the total sequencing depth for all historical samples. Median (black line), lower threshold (red line; 0.5 x the median), and the upper threshold (red line; 1.5 x the median).*

After these conservative filtering steps, we assessed where the filtered repetitive sites were located. Overall, the filtered sites file contained a smaller number of repetitive sites for each type of class than the original unfiltered repetitive sites. The largest difference between the two files was for satellite and centromeric repeats, with 446,103,674 (85.850%) fewer sites in the filtered file (**Figure 42**). The final filtered repetitive sites file contained mostly LINE/L1 (25.990%), LINE/RTE-BovB (20.112%), and SINE/tRNA-Core-RTE (20.039%) repeat classes. The filtered non-repetitive sites and filtered repetitive sites were then separately used to estimate genotype likelihoods for each population genomics analyses.

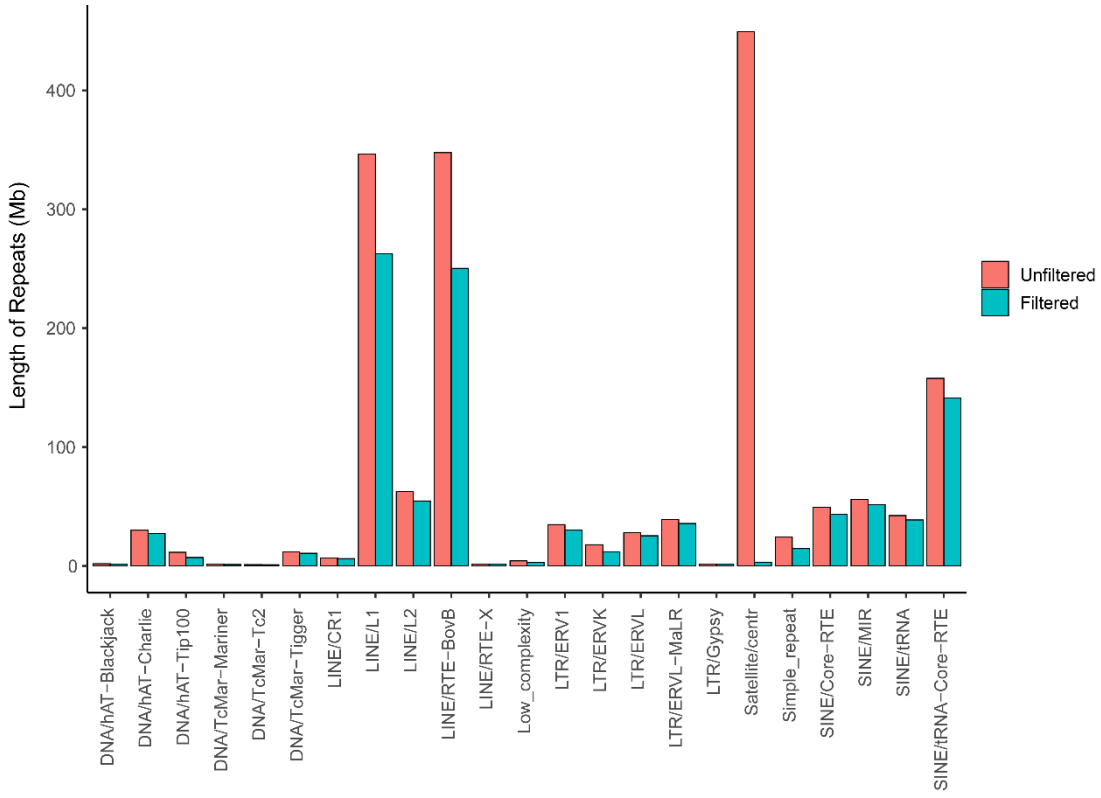


Figure 42: *Length of repeats by family for the unfiltered and filtered repetitive sites. Only repeat families totalling > 1 Mb were visualised.*

3.3.8. Modern resequencing data

We additionally obtained published WGS data (X. Wang et al., 2024) from 119 modern waterbuck samples from 10 populations across Africa, which were mapped to the chromosome-level waterbuck genome. The mapped BAM files were quality controlled and had a genome-wide coverage between 2.083X and 9.009X, with a mean of 4.127X. The number of mapped reads ranged from 31,577,524 to 190,632,490, the minimum mapping quality was 34.700, and the maximum mapping quality was 39.800.

3.3.9. Population structure

To explore the population structure of the waterbuck across the species distribution, we firstly genotyped the 24 historical samples and constructed a PCA, with either filtered non-repetitive or repetitive genomic sites. Both PCAs showed similar results, where PC1 (25.07%-26.03%) showed a split between the common and defassa subspecies, and PC2 (7.36%) split populations in the north and in the south separately (**Figure 43**). There was a larger percentage of variance between the two defassa populations, than the two common populations.

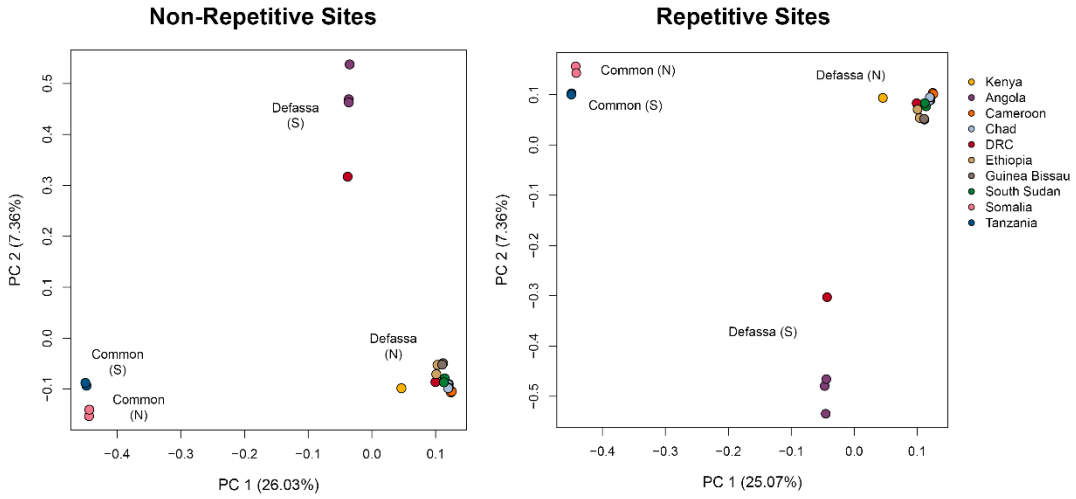


Figure 43: *Principal component analyses (PCAs) of the 24 historical waterbuck samples using all filtered genomic sites (non-repetitive and repetitive).*

We then combined the 24 historical and 119 modern waterbuck samples, genotyped, and constructed a PCA, for both the filtered non-repetitive and repetitive genomic sites. Because historical samples have DNA damage resulting in an increase in transition mutations in the sequences (C to T or G to A), this might incorrectly increase the genotype differences between samples, and as such we assessed this impact by genotyping only transversion sites. The PCAs again showed similar results between non-repetitive and repetitive sites, but also between all sites and transversion sites, confirming that DNA damage was low and not causing biases in the PCAs (**Figure 44**). The two subspecies group separately in PC1 (19.53%-21.48%), but with the defassa (N) population (Samole, Angola, and the southern DRC individual) centred in between. On PC2 (3.05%-3.26%) the north and south populations grouped separately. The Nairobi sample (Common N) grouped closer to the northern defassa population, suggesting recent admixture.

In the north, Maswa, Ugalla, and a Kenyan sample clustered together and were positioned further towards the common subspecies in PC1 than the remaining northern defassa populations, reflecting their closer locality to the common waterbuck's range. Samole and Guinea-Bissau individuals had the lowest values on PC2 of the northern defassa grouping and were closer to southern defassa populations in PC2, and therefore were not organised by latitude or geographic distance. Whereas, for the common subspecies, populations within the northern and southern groupings on PC2 were organised by latitude.

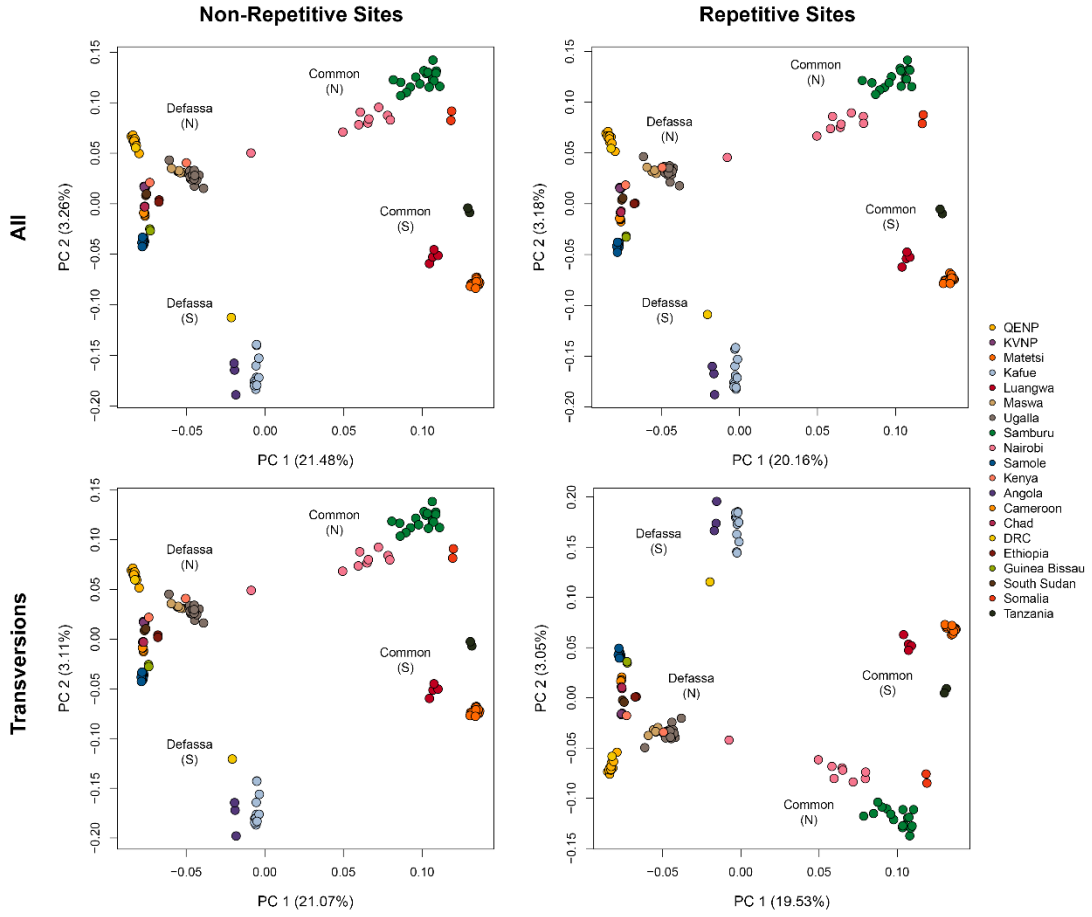


Figure 44: *Principal component analyses (PCAs) of historical and modern waterbuck samples using all filtered non-repetitive sites (top left), all filtered repetitive filtered sites (top right), filtered non-repetitive transversion sites (bottom left), or filtered repetitive transversion sites (bottom right).*

A NJ tree of the 143 samples showed similar results to the PCAs, with four clades representing the four populations, and the Nairobi (Common N) individual grouping in the defassa clade (**Supplementary Figure 1**).

Moreover, overall F_{ST} was calculated between the two subspecies and the four populations (Defassa N, Defassa S, Common N, and Common S), using the filtered non-repetitive sites. Between the two subspecies, the overall F_{ST} was 0.214. F_{ST} between the four major populations reflected geography, with Defassa (N) and Common (S) having the highest value (0.303), followed by the two southern populations (0.245), and the two northern populations at 0.218 (**Figure 45**). Whereas F_{ST} within subspecies (north and south) was much lower at 0.139 for defassa and 0.115 for common. This additionally confirms results seen in the PCAs, where the variance was greater on PC1 (subspecies) compared to PC2 (north/south). The overall F_{ST} values were similar for repetitive sites (**Supplementary Figure 2**).

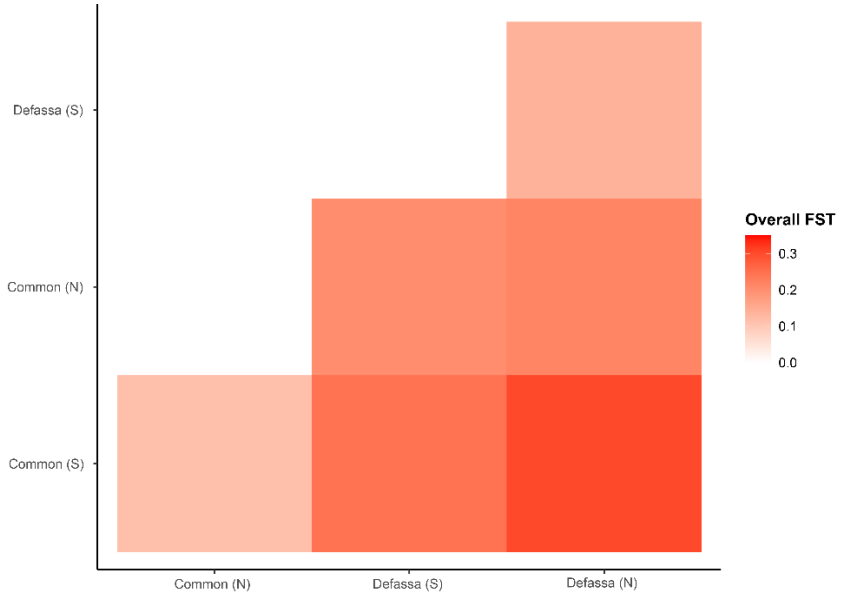


Figure 45: *Pairwise genomic differentiation (F_{ST}) between populations for the filtered non-repetitive sites.*

We then used the program Admixture to further explore the population structure of the waterbuck, by assessing levels of admixture within individuals. We firstly set the number of populations to two (k=2), which grouped the two subspecies separately, but with varying levels of admixture proportions (**Figure 46**). Non-repetitive and repetitive sites showed similar levels of admixture in each individual at k=2. For the defassa subspecies, some populations in the north (Maswa, Ugalla, and Kenya) had varying proportions of admixture with common waterbuck. Populations in the south for the defassa subspecies (Samole, Angola, and the southern DRC individual) also showed admixture with the common waterbuck. Moreover, all common waterbuck in the north had some proportion of admixture with defassa at k=2, whilst only Luangwa had some admixture with defassa

in the southern common population. The Nairobi sample, identified in the PCAs as being admixed, had the highest proportion of admixture out of all samples.

At three populations (k=3), the admixture proportions differed between individuals genotyped using the non-repetitive and the repetitive genomic sites (**Figure 47**). The non-repetitive sites showed a similar result to k=2 for common populations in the north, with varying degrees of admixture with Defassa N, and Luangwa (Common S) had admixture with defassa. For the defassa, populations in the south had admixture from both the common and defassa (populations in western and central Africa), confirming their intermediate placement in the PCAs. In the north, Maswa, Ugalla, and a Kenyan defassa waterbuck had admixture with common. Only Samole and QENP had no admixture (but were grouped into two different populations, red and blue, respectively). The admixture plot using repetitive genomic sites showed a different picture for the defassa subspecies, where individuals in the north had low levels of admixture, and individuals in the south were grouped separately into their own population (red).

At four populations (k=4), the non-repetitive sites grouped the common waterbuck with varying proportions of admixture with defassa (**Figure 48**). Most common populations in the north had admixture with defassa in the north, whilst Luangwa in the south had some admixture with defassa in the south. For the Defassa (N), two populations were identified, whereas defassa in the south grouped separately. Maswa, Ugalla, and the Kenyan individual all had admixture with common, as seen at the lower k values. Along with Ethiopia, they also had small proportions of admixture with Defassa (S). The repetitive sites showed some differences, with one notable difference being the splitting of defassa in the south into two populations, with a larger proportion being defassa and a smaller proportion from common in the south.

Furthermore, we ran Admixture up to k=20, however most k values above four were not informative. Twelve populations (k=12) had the lowest amount of admixture within individuals and most closely grouped the populations geographically (with some of the original populations shared, e.g. Tanzania and Luangwa, or Cameroon and Chad; **Figure 49**). Using the non-repetitive sites, Defassa (N) was grouped as one population, but when using the repetitive sites, Kafue groups separately to Angola and the DRC individual.

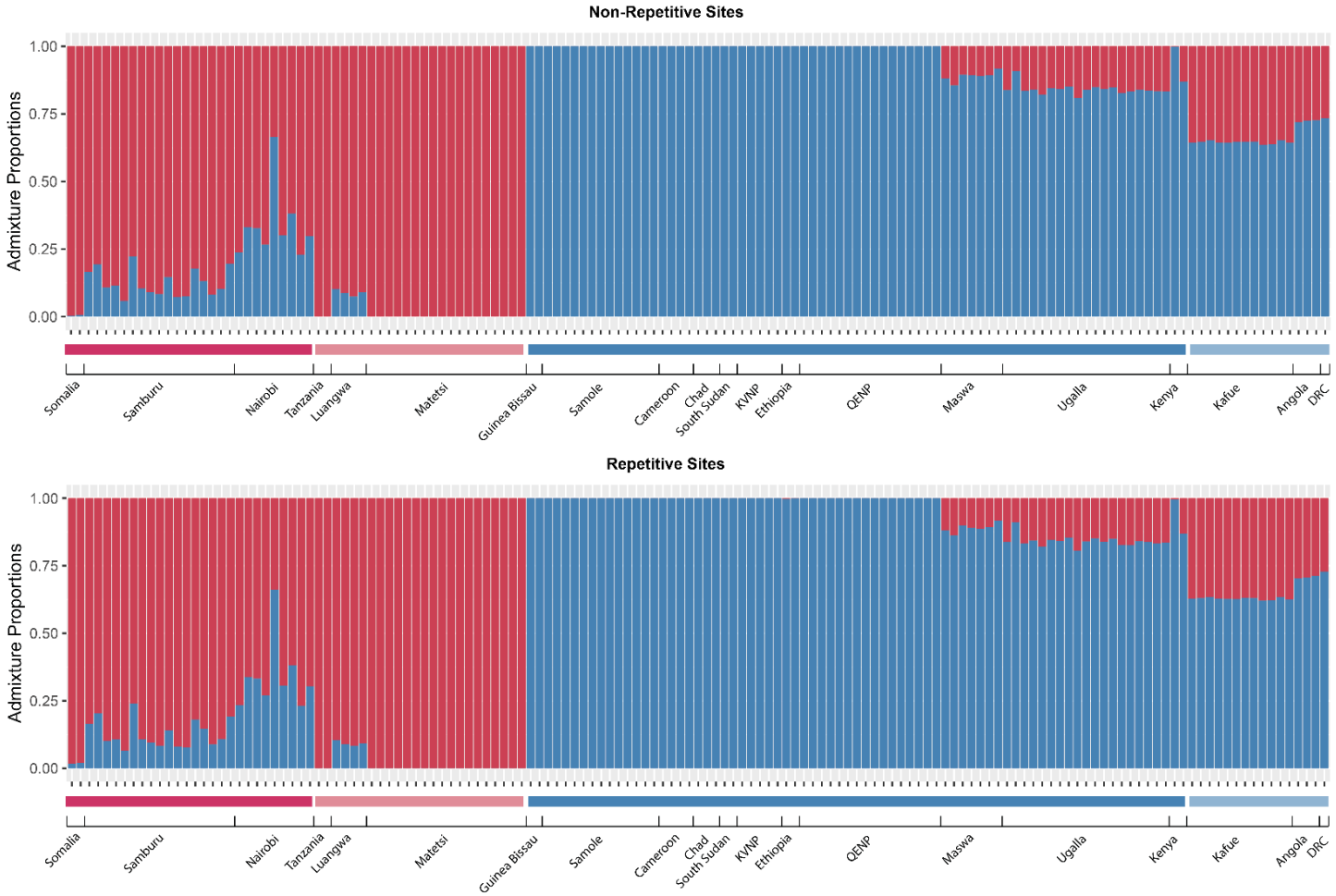


Figure 46: *Admixture proportions (k=2) for historical and modern samples using non-repetitive (top) and repetitive (bottom) filtered genomic sites. Horizontal bars represent the four populations: Common N (dark red), Common S (light red), Defassa N (dark blue), and Defassa S (light blue).*

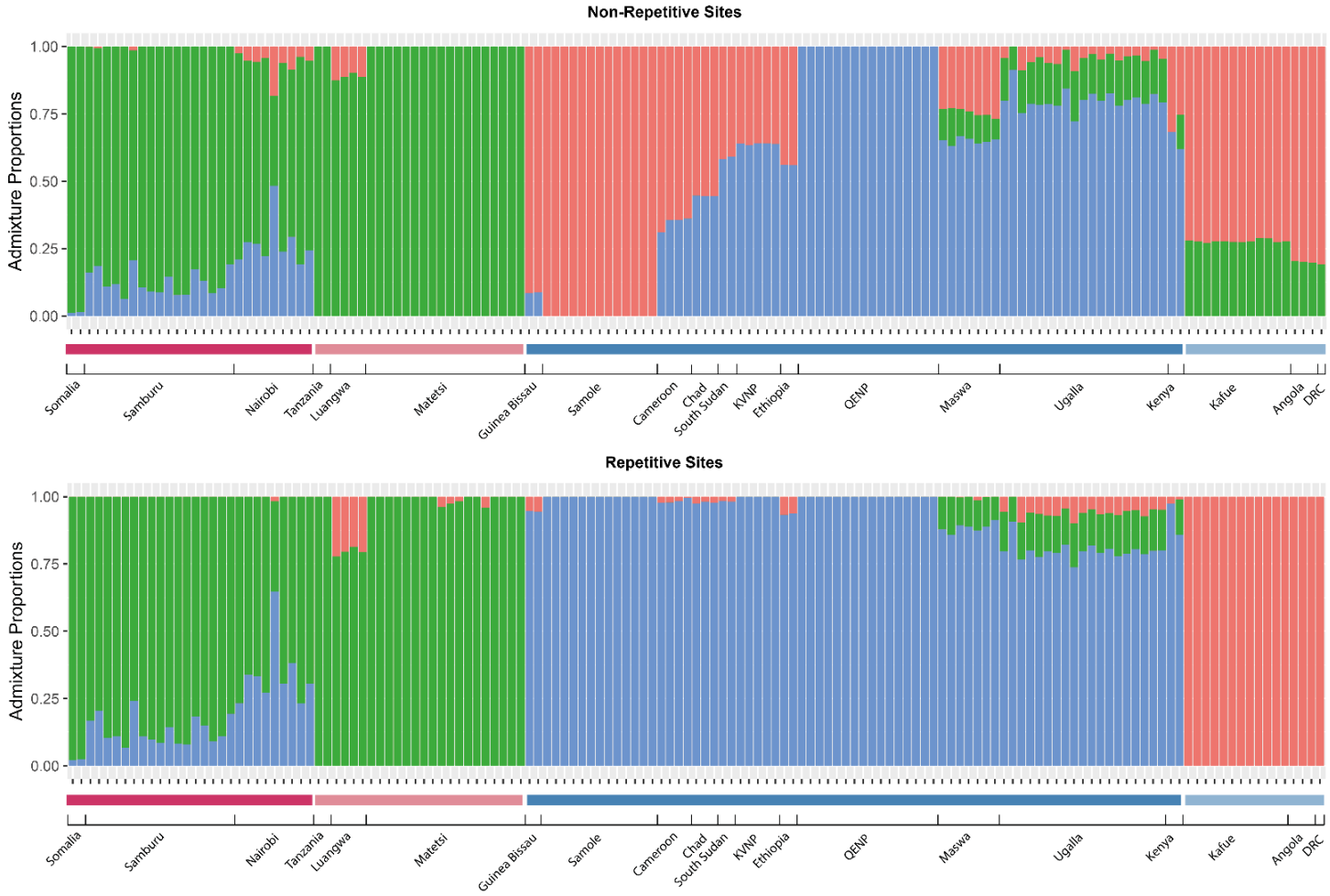


Figure 47: *Admixture proportions (k=3) for historical and modern samples using non-repetitive (top) and repetitive (bottom) filtered genomic sites. Horizontal bars represent the four populations: Common N (dark red), Common S (light red), Defassa N (dark blue), and Defassa S (light blue).*

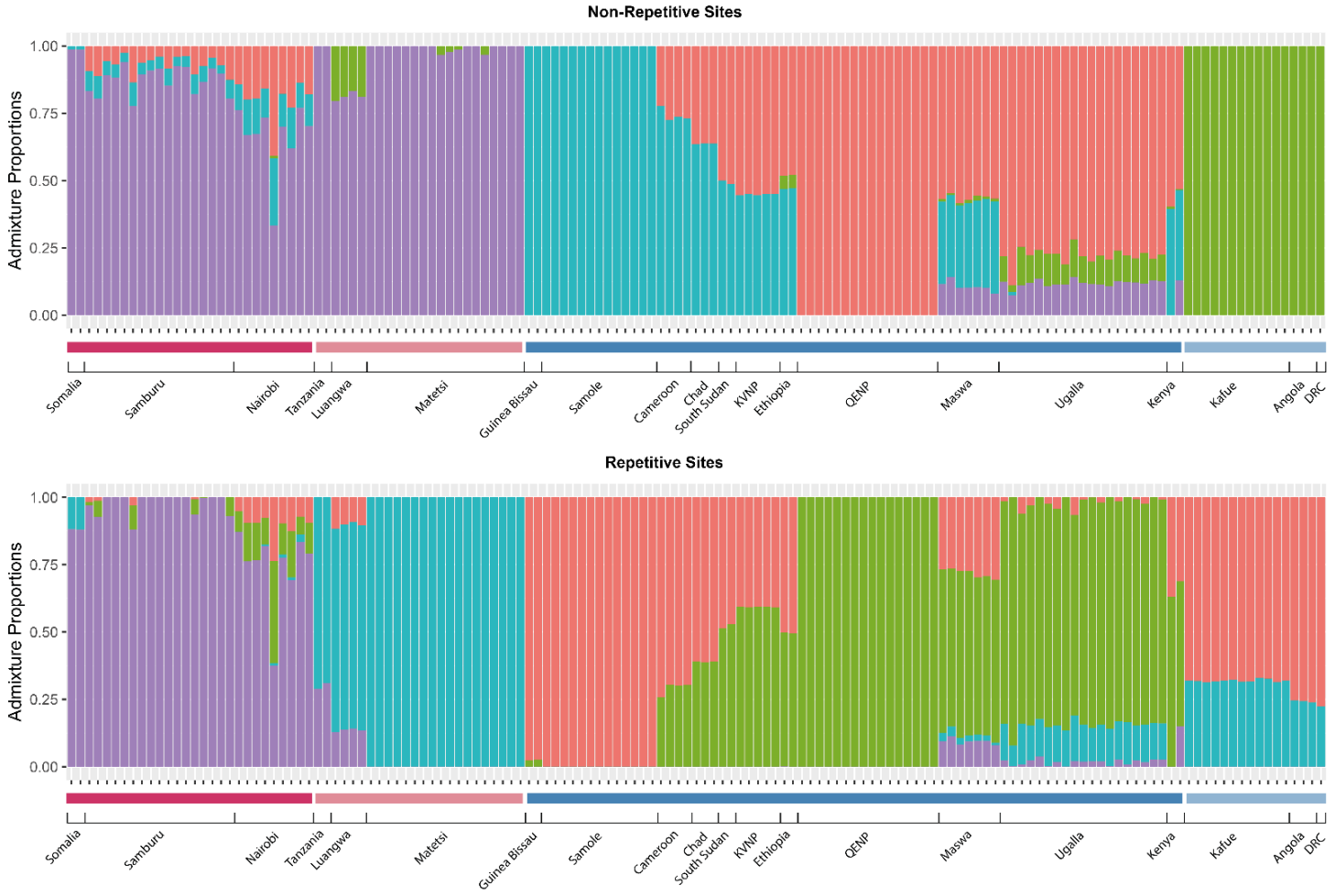


Figure 48: *Admixture proportions (k=4) for historical and modern samples using non-repetitive (top) and repetitive (bottom) filtered genomic sites. Horizontal bars represent the four populations: Common N (dark red), Common S (light red), Defassa N (dark blue), and Defassa S (light blue).*

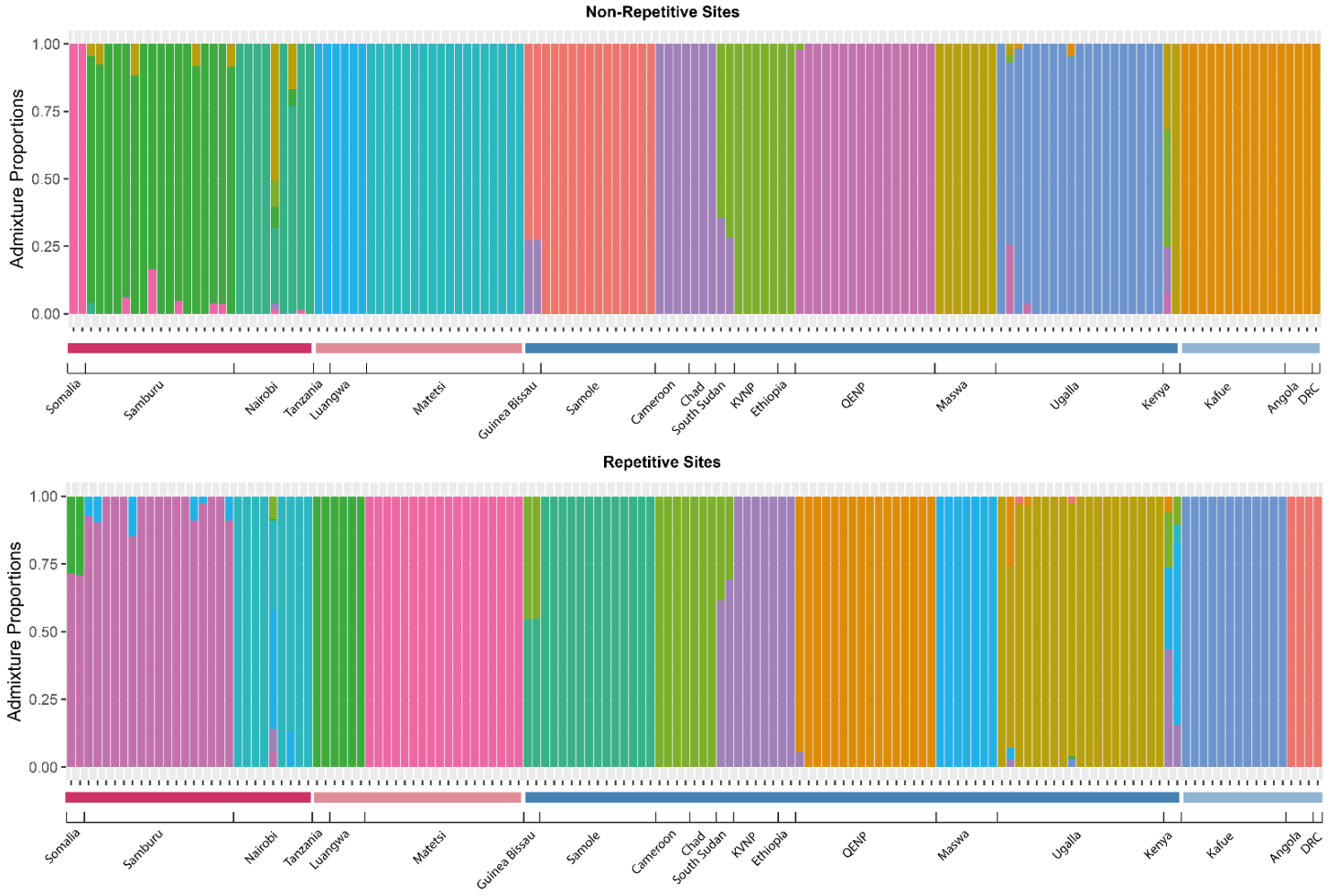


Figure 49: *Admixture proportions (k=12) for historical and modern samples using non-repetitive (top) and repetitive (bottom) filtered genomic sites. Horizontal bars represent the four populations: Common N (dark red), Common S (light red), Defassa N (dark blue), and Defassa S (light blue).*

3.3.10. Heterozygosity

Genome-wide heterozygosity was calculated both on filtered non-repetitive sites and filtered non-repetitive transversion sites, for historical and modern samples (**Figure 50**). Overall, heterozygosity was lower when using only transversion sites, than when using all sites, whilst historical samples had a higher heterozygosity than modern samples when using all sites, but similar when only using transversion sites. This suggests some differences due to historical DNA damage. Sample WB_3k_5X had very high levels of heterozygosity (0.013 for all sites and 0.005 for transversion sites), and so was removed from further analyses. Luangwa had the highest average heterozygosity for transversion sites, whilst an individual from Ethiopia had the lowest and the Cameroon population had the lowest average.

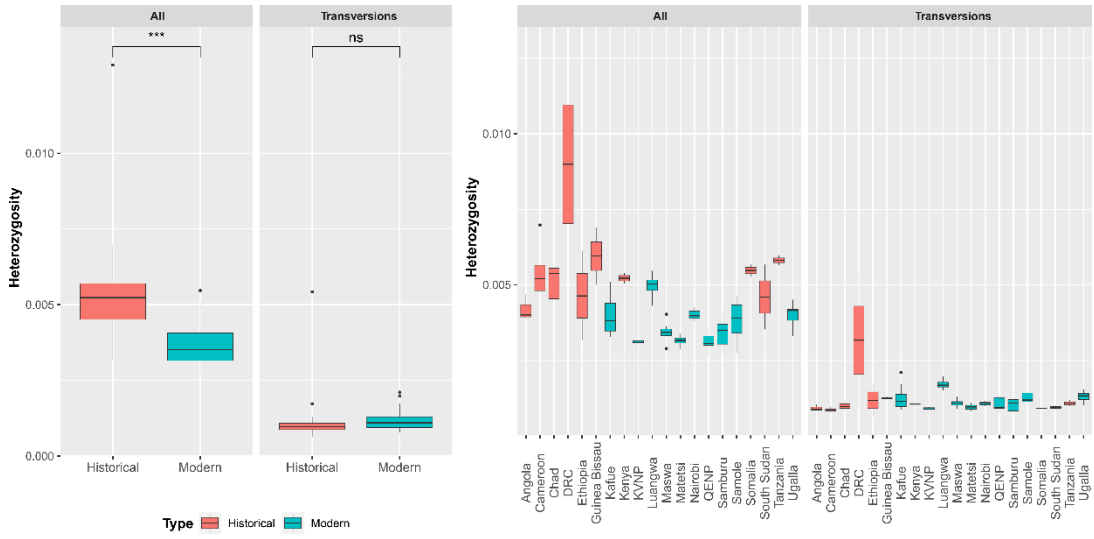


Figure 50: *Heterozygosity of all non-repetitive filtered sites and non-repetitive transversion sites, separated by age (left) and by population (right).*

Heterozygosity was also calculated for repetitive sites and showed similar results to non-repetitive sites (**Supplementary Figure 3**). Notable differences were the estimation of heterozygosity per population for transversion sites, where Kafue had two samples with the highest individual heterozygosity, and on average KVNP and Cameroon had the lowest levels.

3.3.11. Gene flow

EEMS was used to estimate gene flow between historical and modern populations, using the calculated migration rates (*m*), as well as to estimate genetic diversity between populations using the diversity rates (*q*). We calculated these with and without sample

WB_3k_5X, using the filtered non-repetitive genomic sites (**Figure 51**). We found a strong barrier to gene flow between the two subspecies, along the East African Rift Valley. This extends northwards and southwards, showing lower levels of gene flow between the subspecies. This suggests that whilst there is some admixture between common and defassa populations, there has been an historical barrier to the gene flow, and this could be caused by the climate and geography of the East African Rift Valley. There is also a lack of gene flow in the Congolian Rainforest, creating a barrier between defassa in the north and in the south, confirming the separation seen on the PCA and admixture analyses. This area of reduced gene flow, due to rainforest and lack of savannah habitat, extends towards the Rift Valley, creating a divide between defassa populations in Tanzania and southern defassa. Northern and southern common populations also have some reduction in gene flow, again confirming their differences seen in the PCA and admixture.

Mean diversity rates were highest between populations in central Africa around the Rift Valley (Ugalla, Maswa, Nairobi, Samburu, Somalia, and Luangwa), due to the genomic differences between the two subspecies, whilst lowest in the northwest of the defassa range and in Ethiopia (**Figure 51**). Repetitive genomic sites were also used to run EEMS but showed similar results to the non-repetitive sites (**Supplementary Figure 4**).

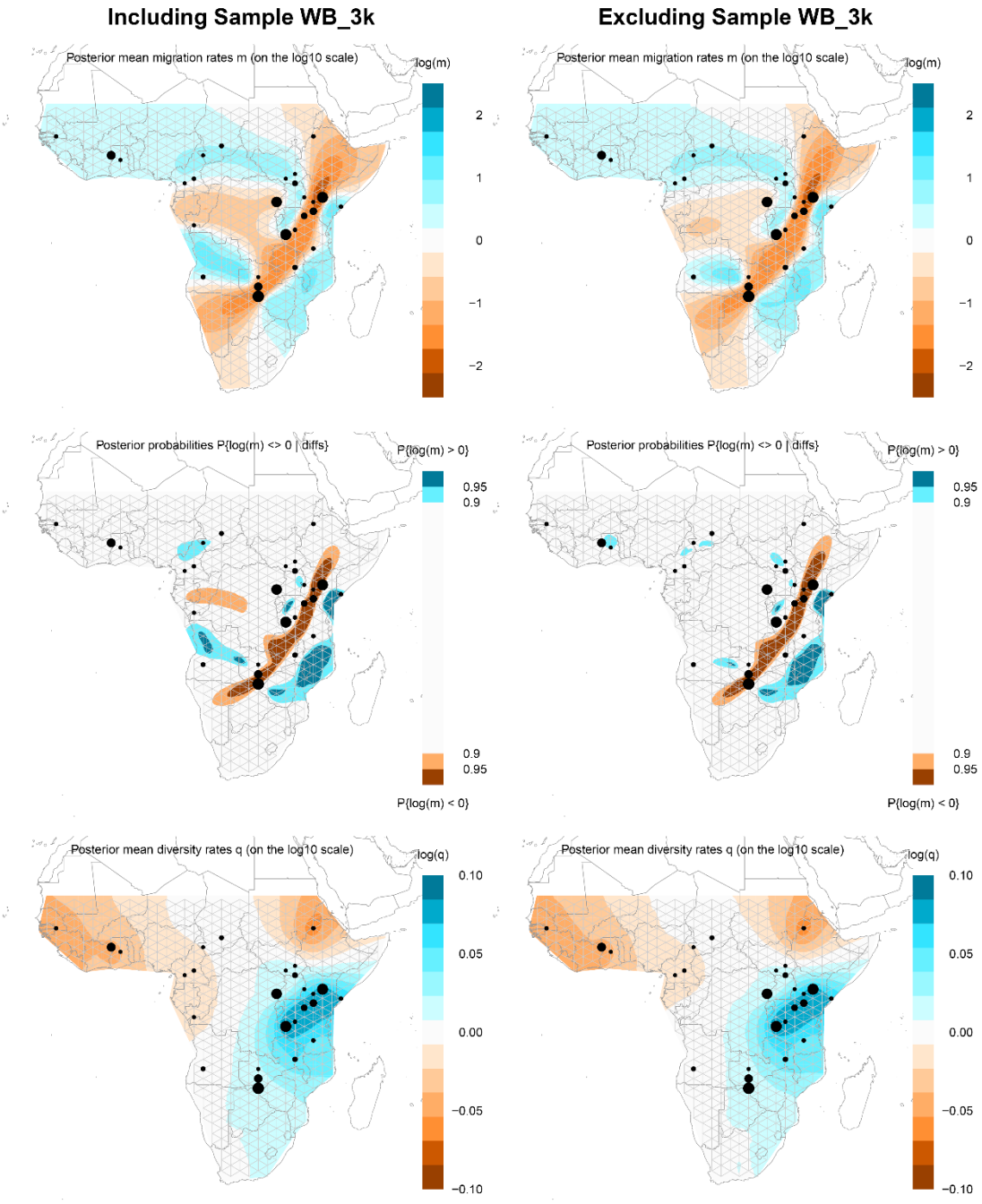


Figure 51: *Estimating Effective Migration Surfaces (EEMS) analysis of the filtered non-repetitive sites, with both migration (m) and diversity (q) rates shown. The analysis was run with all historical and modern samples (143 individuals) and additionally with sample WB_3k removed (142 individuals).*

3.3.12. Genomic differentiation

We next compared the two subspecies to see if there were any regions in the genome showing differentiation. To do this, F_{ST} was calculated in 10 Kb windows using either the non-repetitive sites or the repetitive sites. Both analyses showed similar regions of high

F_{ST} and so we focused on non-repetitive sites. A total of 242 windows were in the top 0.1% of F_{ST} values with greater than or equal to 1000 genomic sites (**Figure 52**). In these windows there were 104 annotated genes (**Supplementary Table 1**). The F_{ST} in these windows ranged from 0.671 to 0.770, on KEL3 and KEL21, respectively. The window with highest F_{ST} (0.770) on KEL21 contained the gene MTMR14, encoding for a myotubularin related protein. In KEL13, we found the second highest F_{ST} peak (0.760), containing genes RTEL1, involved in telomere-length regulation, DNA repair, and maintaining genomic stability; and STMN2, which regulates microtubule stability.

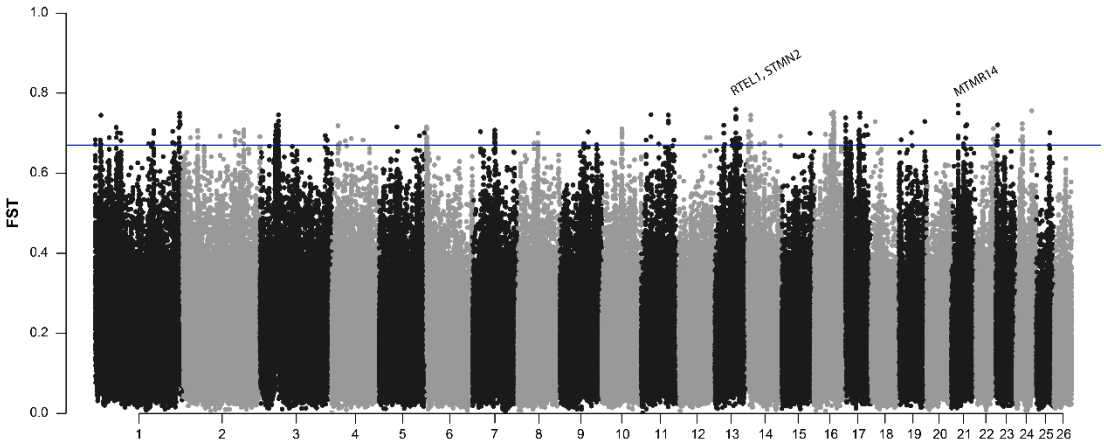


Figure 52: *Genomic differentiation (F_{ST}) calculated in 10 Kb windows between common and defassa waterbuck across the 26 autosomal chromosomes, using non-repetitive sites. Blue line represents the top 0.1% windows (with greater than or equal to 1000 genomic sites).*

Gene function was assessed using the gene ontology (GO) statistical overrepresentation test in Panther, with annotated waterbuck genes against the *Bos taurus* reference list. GO terms for Biological Process included negative regulation of t-circle formation, linoleic acid metabolic process, hepoxilin biosynthetic process, and lipoxygenase pathway (**Figure 53**). Whilst for Molecular Function, terms included RNA polymerase II binding, phosphatidylinositol activity, and linoleate activity. No statistically overrepresented terms were found for Cellular Component.

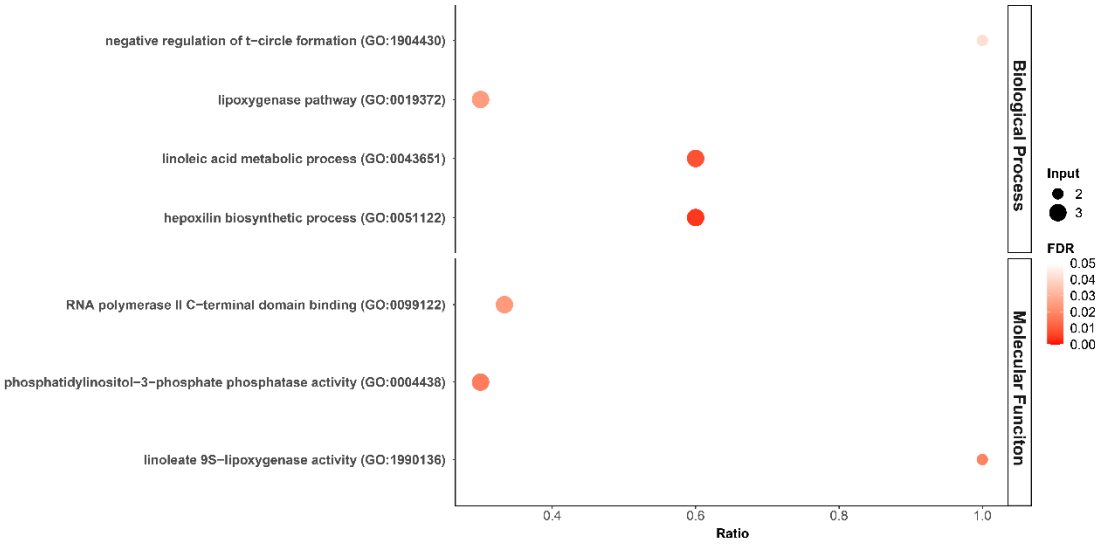


Figure 53: *Gene Ontology (GO) statistical overrepresentation of genes located in the top 0.1% of genomic differentiation (F_{ST}) windows between common and defassa waterbuck. Ratio is the number of input genes (input) out of the number of genes for a particular GO term.*

We then looked at each gene individually. IFT140 located on KEL3, with a F_{ST} of 0.720, is required during the development and maintenance of rod and cone photoreceptor cells and plays a key role in the function of ciliated cells, such as sperm. Two genes on KEL17 (F_{ST} window = 0.705), SLC7A11 and SLC7A6, control the production of pheomelanin pigment or may be involved in mediating the transport of ornithine in retinal pigment epithelial cells, respectively. The latter gene may also transport glycine betaine into the oocyte. The gene ALOX12B (on KEL1), found in a window with an F_{ST} of 0.689, synthesises the corneocyte lipid envelope and the skin barrier, to reduce water loss. It may also regulate the expression of airway mucins.

Some genes were found to be involved with chromatin, such as KMT2C on KEL5 (F_{ST} = 0.701) which forms H3K4me1 methylation marks at active chromatin sites to enable transcription or DNA repair, PPP1CA/PPP1CB (KEL22 with an F_{ST} of 0.686) which controls the structure of chromatin and the progression of the cell cycle from mitosis into interphase, and BPTF on KEL1 (F_{ST} = 0.672) that regulates chromatin remodelling complexes, orders nucleosome arrays, and facilitates access to DNA during DNA replication, transcription, or repair.

Several windows contained genes associated with microtubules. These included STMN2 (0.741 on KEL13) which regulates microtubules and may be involved, by similarity, with learned and innate fear. KATNAL1 (KEL14) with an F_{ST} of 0.673 which regulates

microtubules in Sertoli cells during spermiogenesis. Another gene in the same window, KATNA1, is involved in the reorganization of microtubules and during nucleation, the release of microtubules from the centrosome. The gene RABGAP1 (KEL9, F_{ST} = 0.671) may also have a similar function. While other windows contained genes involved in embryogenesis. For example, KEL17 contained a window (F_{ST} = 0.696) with the gene LEUTX, a homeobox transcription factor involved in embryogenesis and may help regulate the activation of the genome in the embryo. A gene on KEL1, HELZ (F_{ST} = 0.684), may metabolise RNA in the developing embryo.

Considering that both subspecies have a clearly defined population structure between north and south individuals (**Figure 44**), and that the overall F_{ST} between defassa populations and common populations showed some differentiation (**Figure 45**), we explored genomic differences between the two major population groups in each subspecies. F_{ST} was calculated in 10 Kb windows as above, using both non-repetitive (**Figure 54**). A total of 242 windows had the highest 0.1% F_{ST} values when using non-repetitive sites and contained 65 and 68 genes in the common and defassa comparisons, respectively.

In the top F_{ST} windows identified between the common populations containing genes, F_{ST} values ranged between 0.468 on KEL2 and 0.632 also on KEL1 (**Supplementary Table 2**). The highest contained the genes ASGR1 and ASGR2, which mediate endocytosis of plasma glycoproteins, and one gene which was not functionally annotated (LOC132346939). The second highest window (F_{ST} = 0.631) was on KEL9 which contained the gene ZC3H8 that encodes for a zinc finger protein. The third highest window was also on KEL9 (F_{ST} = 0.608), containing the gene GREB1 which may play a role in cell proliferation stimulated by oestrogen.

Whereas for the F_{ST} between defassa populations, values were between 0.555 on KEL19 and 0.750 on KEL1 for the top 0.1% of windows containing genes (**Supplementary Table 3**). The highest F_{ST} window contained the genes ALDH3A1, ALDH3A2, and ALDH3B1, which are aldehyde dehydrogenases. The second highest window was on KEL9 (F_{ST} = 0.716) and contained LIMS1 and LIMS2, involved with the cytoskeleton. KEL7 had the third highest F_{ST} window (0.701) and this contained the gene SMU1 which plays a role in pre-mRNA splicing in the spliceosome.

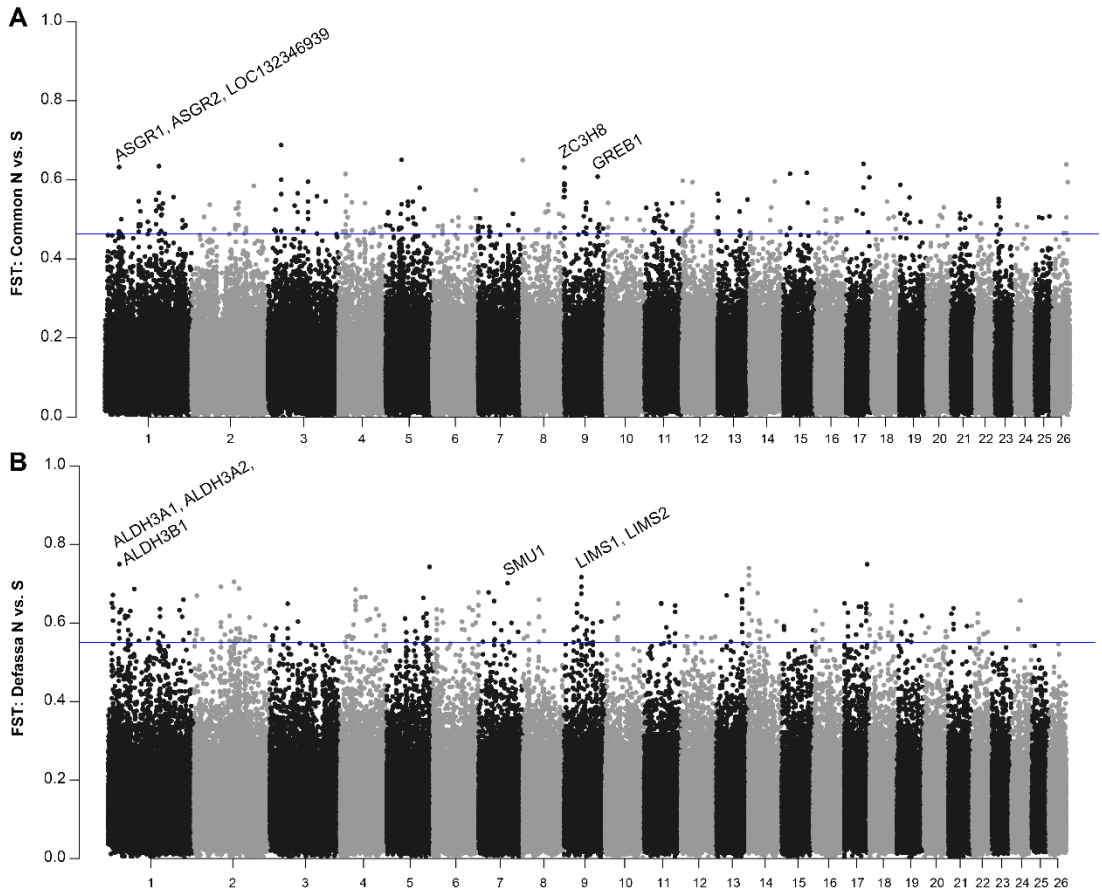


Figure 54: *Genomic differentiation (F_{ST}) in 10 Kb windows between waterbuck in the north and south for the common (A) and defassa subspecies (B). Analyses included all 26 autosomal chromosomes and filtered non-repetitive genomic sites. Blue lines represent the top 0.1% windows (with greater than or equal to 1000 genomic sites).*

Gene ontology (GO) statistical overrepresentation test was performed on each subspecies, with the genes found in the top 0.1% of F_{ST} windows. For the common waterbuck, 16 GO terms were identified as being overrepresented (**Figure 55**). These included sulphate assimilation, propionate catabolic processes, neuronal and membrane depolarisation during action potential, glycerol biosynthetic processes, glucose-6-phosphate transport, and 3'-phosphoadenosine 5'-phosphosulfate biosynthetic processes for Biological Process. GO terms for Cellular Component included voltage-gated sodium channel complex and heterotrimeric G-protein complex. For Molecular Function, voltage-gated sodium channel activity, phosphoenolpyruvate carboxykinase (GTP) activity, histone H3K4 demethylase activity, glucose 6-phosphate:phosphate antiporter activity, G-protein beta/gamma-subunit complex binding, fucose binding, and adenylylsulphate kinase activity were found.

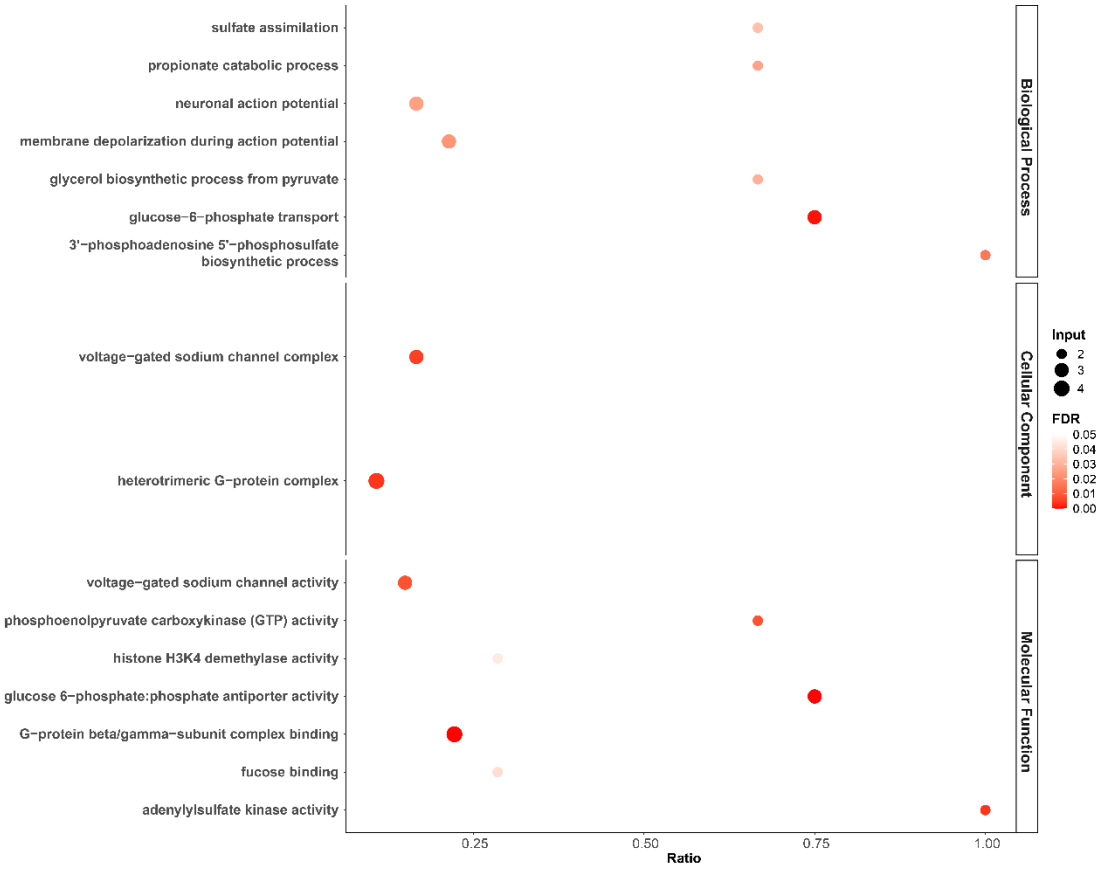


Figure 55: *Gene Ontology (GO) statistical overrepresentation of genes located in the top 0.1% of F_{ST} windows between common waterbuck populations in the north and south. Ratio is the number of input genes (input) out of the number of genes for a particular GO term.*

When comparing north and south defassa populations, four GO terms showed statistical overrepresentation for Molecular Function, including vitamin D response element binding, cell-cell adhesion mediator activity, benzaldehyde dehydrogenase (NAD+) activity, and 3-chloroallyl aldehyde dehydrogenase activity (**Figure 56**). No terms were statistically significant for Biological Process or Cellular Component.

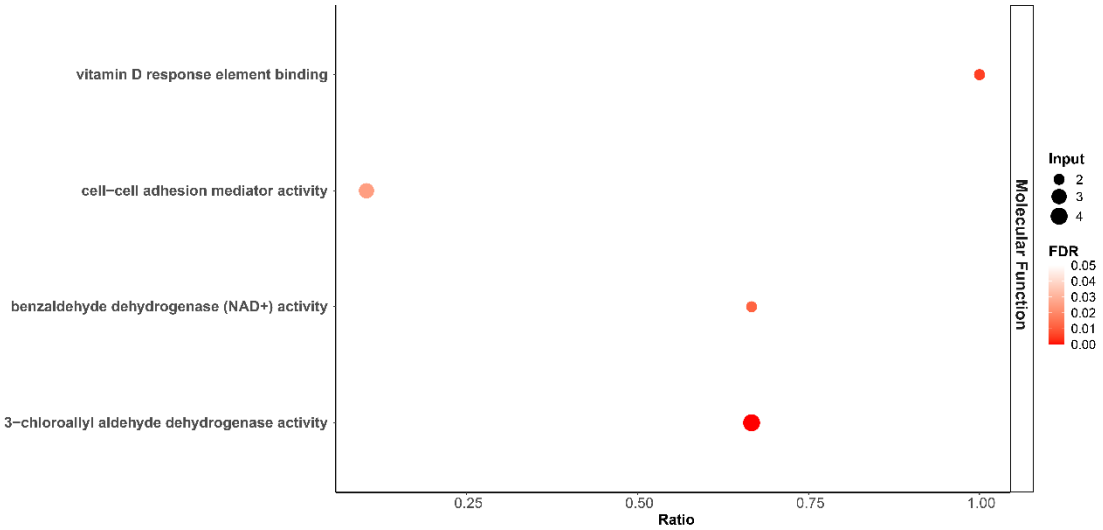


Figure 56: *Gene Ontology (GO) statistical enrichment of genes located in the top 0.1% of F_{ST} windows between defassa waterbuck populations in the north and south. Ratio is the number of input genes (input) out of the number of genes for a particular GO term.*

3.4. Discussion

Combining a chromosome-level genome with an increased sampling of the species distribution using historical museum samples, we were able to expand on the previous genomic study conducted in waterbuck (X. Wang et al., 2024). We firstly confirmed population structure across the species and within each subspecies using PCA, Admixture, and F_{ST} analyses. The increase in samples across central and southern Africa allowed us to study population structure on a finer scale. Overall, we confirmed that populations reflected geographic locations, with a separation of the two subspecies, and a north-south separation within each subspecies (**Figure 44**, **Figure 45**, **Figure 46**), as shown previously (X. Wang et al., 2024). We expand on the previous genomic study by discussing the finer scale population structure within the waterbuck.

Focusing on populations in the north, in defassa we find that populations group into four smaller clusters. These include populations in western Africa (Guinea-Bissau and Samole in Ghana), populations in central Africa (Cameroon, Chad, Ethiopia, KVNP in Uganda, and an historical sample in Kenya), waterbuck in Uganda (QENP), and waterbuck in Tanzania (Maswa and Ugalla). PC1 and Admixture (**Figure 44**, **Figure 46**) show that Maswa, Ugalla, and Kenya populations are clustered more closely to common waterbuck and have varying proportions of admixture. This reflects their proximity to the common subspecies ranges. Whilst in the common subspecies, Nairobi and Samburu

(Kenya) were positioned closer to the northern defassa group and had some admixture with defassa, again reflecting their geographic locality near the hybrid zone of the two subspecies. The two Somalia individuals were grouped separately and had only a very small amount of admixture with defassa. These results support the previous genomic study using PCA and admixture, as well as their TreeMix analysis which found historical admixture from Maswa to Nairobi (X. Wang et al., 2024).

In the south, defassa populations (Kafue, Angola, and DRC) were clustered between the two subspecies on PC1 (**Figure 44**) suggesting an admixed origin, which was supported by the Admixture analysis (**Figure 46**). Surprisingly, this admixture was found as far north as the historical Angola and DRC populations. Whereas in the southern common group only Luangwa had admixture with defassa, whereas Matetsi although geographically closer to Kafue did not present admixture with defassa populations. This could suggest that admixture occurred between Defassa (S) and Luangwa in both directions. Whilst it also suggests that there must be a barrier to gene flow between Defassa (N) and Matetsi. Together, these results suggest that many populations along the hybrid line are admixed, confirming that admixture has occurred along this entire region, in both directions, both recently and historically (X. Wang et al., 2024).

However, the two subspecies remain genomically differentiated in non-admixed populations. The EEMS analysis (**Figure 51**) shows a strong barrier to gene flow across the Great Rift Valley, but also the Congolian Rainforest, restricting gene flow between subspecies as well as between the northern and southern populations in defassa. However, on PC1 of the PCA (**Figure 44**) the Defassa (N) population of Samole (Ghana) clusters closer to the Defassa (S) group, whilst populations in Uganda and Tanzania are genomically more different. This is also evident in the Admixture analysis at k=3. This raises the question of whether historically there was movement of waterbuck around the west of the Congolian Rainforest, rather than the east. Sequencing samples from Gabon and the Republic of Congo may provide greater insight into this, as the historical sample we analysed from this region (DRC) had elevated DNA damage. Moreover, the reduced gene flow between defassa and Matetsi in the south may reflect the lower extensions of the Rift Valley and the Zambezi River. However, admixture has been found between Defassa (S) and Luangwa, therefore there must have been gene flow between these populations, and this could be the reason for the admixture found in all the Defassa (S) individuals.

There is also some restriction of gene flow between the two common groups (**Figure 51**), which might have been caused by extensions of the Rift Valley in this area. Within the larger population groups gene flow is high, supported by the PCA which shows similar clustering (**Figure 44**), suggesting any small genomic differences are due to distance. Compared to the previous study, we do not find regions along the Rift Valley that are more permeable to gene flow in the EEMS analysis. Our data indicates that whilst there is high genomic differentiation between the two subspecies and a restriction of gene flow along the Rift Valley, historical and recent admixture has occurred.

It has been suggested that the Great Rift Valley has periodically acted as a climatic barrier between the two subspecies, as during the Pleistocene there were warmer periods with very little available surface water (X. Wang et al., 2024). Climate, combined with geographic barriers such as mountains, rivers, and lakes, may have created the population structure and admixture seen today. However, while waterbuck have high genomic differentiation between subspecies and variable karyotypes, reproductive isolation has not yet set in for this species.

Zooming in at the chromosome-level, we identified several regions showing high genomic differentiation (F_{ST}) between the two subspecies, with some containing annotated genes (**Figure 52**). The window with the highest F_{ST} contained the gene MTMR14, encoding a myotubularin-related protein that affects autophagy (Gibbs et al., 2010), the removal and recycling of cellular components. It may also play a role in sperm cells and affect male fertility (Wen et al., 2018).

The second highest window contained the gene RTEL1, encoding a DNA helicase which controls telomere length, recombination, and replication (Vannier et al., 2014). When knocked-out in mice embryonic stem cells it causes reduced proliferation, chromosomal fusions, and telomere length heterogeneity (Ding et al., 2004). The protein also limits the excessive crossing-over of chromosomes during meiosis, reducing homologous recombination. Moreover, it also plays a role in T-loop disassembly at telomeres which protects chromosomes from degradation and DNA repair. The gene may therefore play a role in the polymorphic chromosome fusions in waterbuck, as shown in *C. elegans* where the absence of the gene causes complex rearrangements (León-Ortiz et al., 2018). RTEL1 reduces homologous recombination during meiosis. It is well known that recombination is a major barrier to the formation of new species, primarily because it opposes the establishment of linkage disequilibrium (Trickett &

Butlin, 1994). As such, a recombination suppressor causing even a slightly change on the recombination profile between two subspecies might enhance the processes leading to speciation.

Other genes within the highest F_{ST} windows include ALOX12B, encoding for a epidermal lipoxygenase; mutations in this gene have been associated with skin disease in humans called ichthyosis (Jobard et al., 2002). The gene KATNAL1 is expressed in Sertoli cells and is essential for the retention of spermatids during spermiogenesis, with a loss of function of this gene resulting in male infertility due to the disruption of microtubule dynamics and the premature release of spermatids (Smith et al., 2012). Finally, the homeobox gene LEUTX was differentiated between the subspecies and is expressed in human embryos during preimplantation development (Jouhilahti et al., 2016). These genes, especially involved in fertility and development, may be evidence of early speciation between the subspecies.

In order to sample across a broader distribution of the waterbuck's range, we made use of historical museum collections (**Figure 34**). Museum collections are beginning to be employed in genomic studies (Raxworthy & Smith, 2021). Caution is required when working with these samples, as they are prone to low endogenous DNA and DNA damage, including degradation of DNA fragment sizes, cross-linking, and DNA base modifications (described in Pääbo et al., 2004). We found DNA degradation was high (as shown by the small DNA fragment sizes measured using gel electrophoresis; **Figure 35**) but DNA damage was estimated to be relatively low in the historical WGS data (using bioinformatic tools; **Figure 38**). This resulted in shorter insertion sizes in the historical WGS data, with the majority of reads overlapping, and higher estimations of heterozygosity compared with the modern samples (**Table 15**, **Figure 50**). The difference in heterozygosity could reflect a higher genetic diversity within the species a hundred years ago, and a loss of diversity over time. But when using only transversion sites we find that heterozygosity is similar to the modern samples (**Figure 50**). This suggests that the increase in heterozygosity is more likely caused by minor DNA damage, as transitions mutations are the most common cause of DNA modification by deamination (Pääbo et al., 2004). This increase in heterozygosity did not appear to impact the population structure analyses, and therefore museum samples can play a vital role in increasing sampling of species that are widely distributed or difficult to sample.

Population genomic studies sometimes conservatively filter genomic sites in order to avoid incorrect mapping or regions where genotypes cannot be accurately called, especially when using low-coverage WGS data. This usually includes removing all repetitive sites from the genome (Pečnerová et al., 2021). In the waterbuck, repetitive sites make up the majority of the genome (54.801%; **Table 9**). Applying the same filtering criteria for non-repetitive areas, 32.933% of the repetitive genome was rescued for analysis (**Table 16**). Because TE insertions rarely provide an immediate fitness advantage to their host, those reaching fixation in the population do so largely by genetic drift, and largely decaying neutrally once inserted (Bourque et al., 2018). As such, these remaining repeats with high mappability could be TEs that had enough time to mutate over time and became more distinct to each other, behaving like non-repetitive areas of the genome. We show that these repeats can also be useful in population genomic studies and could be combined with non-repetitive sites to increase the proportion of the genome studied.

This study shows the importance of having a chromosome-level reference genome and genome annotation for the species being studied, especially when studying genomic differentiation across chromosomes. Waterbuck population genomics could be further studied by using higher coverage WGS data. For example, this could be used to study runs of homozygosity across the genome to further understand the genetic diversity of populations across the species distribution. It also provides a framework for further studies to utilise museum collections to increase genomic sampling.

4. Inter- and Intra-Species Chromosome Evolution in Antelopes

4.1. Introduction

The family Bovidae, containing antelopes, have variable karyotypes ranging from 2*n* = 30 to 2*n* = 60 (Arias-Sardá et al., 2023). Karyotype variation in the family is dominated by Robertsonian fusions and has been widely studied using cytogenetics (Gallagher et al., 1999; Gallagher & Womack, 1992; S. C. Kingswood et al., 2000; Vassart et al., 1995). Several species have also been identified as having polymorphic Rb fusions, including in the genus *Aepyceros*, *Kobus*, *Redunca*, and *Syncerus* (Pagacova et al., 2011). Several ancestral nodes on the ruminant phylogenetic tree have been proposed, including the Bovidae ancestor, which has been shown to have either *n* = 30 (Farré et al., 2019) or *n* = 29 chromosomes when using genomic methods (Arias-Sardá et al., 2023).

To the best of our knowledge, a complete and focused review of chromosome evolution within Bovidae has not been carried out recently, and previous studies have focused on either cytogenetics or genomics. Additionally, further studies are needed that combine chromosome evolution with 3D genome organisation, as chromosome rearrangements interact physically within the nucleus and are found in heterochromatic regions, as stated by the "Integrative Breakage Model" (Farré et al., 2015). Therefore, studying the impact of 3D genome organisation on chromosomes rearrangements in Bovidae is also needed to further understand the potential reasons behind the high number of Rb fusions within this group.

We therefore employed the availability of chromosome-level genomes and reviewed the literature, to further understand chromosome evolution within this group with extensive interchromosomal rearrangements, both between species, but also in the case of the waterbuck, within species. We firstly aligned bovid genome assemblies to our waterbuck chromosome-level genome and reconstructed ancestral chromosomes to review chromosome evolution across the family. We also utilised the extensive cytogenetic literature for bovids to understand the frequency of chromosome rearrangements in as many taxa as possible. Additionally, we used Hi-C data to understand if chromosome rearrangements are impacted by the 3D genome organisation of the nucleus. Finally, we used the waterbuck as a model for intraspecies chromosomal evolution, to uncover signatures of chromosome rearrangement, and novel structural variants, using indirect approaches.

4.2. Materials and Methods

4.2.1. Bovidae genome assemblies

In order to successfully align and reconstruct the ancestral chromosomes of the ruminant subfamily Bovidae, to study interspecies chromosome evolution, high-quality genomes were needed. We prioritised genomes at chromosome-level, as scaffold-level assemblies can result in fragmented ancestral genome reconstructions. Of the 151 Bovidae species, at least 50 were known to be sequenced to scaffold-level (23 with a scaffold N50 above 1 Mb), and 11 at chromosome-level (Arias-Sardá et al., 2023). Additional species have been sequenced since this publication, including the chromosome-level waterbuck genome assembled in this project. Therefore, after filtering the Bovidae genomes at chromosome-level that were duplicates, or from very closely related species with identical karyotypes, 13 chromosome-level genomes were taken forward (**Table 17**). The scaffold genome for the red lechwe (*Kobus leche*) was also included as it is a sister species to the waterbuck.

We then filtered the chromosome-level genomes by removing any unplaced scaffolds with the UCSC Kent Utilities tool faSomeRecord (Kuhn et al., 2013; Rhead et al., 2009), or removing scaffolds smaller than 10 Kb from the scaffold-level assemblies with faFilter. Chromosome or scaffold sizes were calculated with faSize '-detailed' (Kuhn et al., 2013; Rhead et al., 2009) and FASTA files were converted into 2bit format with faToTwoBit. For genomes without repeats annotated and masked, RepeatMasker v2.6.0+ (http://repeatmasker.org) was run before 2bit formatting of the FASTA file with cattle as reference and -frag set to 60,000.

Table 17: *Genome assemblies used in the pairwise alignments and the two ancestral chromosome reconstructions.*

Subfamily	Common name	Species Name	2*n*	Assembly	Size (Gb)	No. Scaffolds	Scaffold N50 (bp)
Alcelaphinae	Blue Wildebeest	*Connochaetes taurinus*	58	Chromosome	2.650	68,063	98,098,359
Antilopinae	Dama Gazelle	*Nanger dama*	38	Chromosome	3.010	38,032	156,272,796
Antilopinae	Gerenuk	*Litocranius walleri*	60	Chromosome	2.980	80,032	100,165,937
Bovinae	Cattle	*Bos taurus*	60	Chromosome	2.710	2,211	103,308,737
Bovinae	Eastern Bongo	*Tragelaphus eurycerus isaaci*	34	Chromosome	2.970	35,382	192,009,155
Caprinae	Goat	*Capra hircus*	60	Chromosome	3.040	3,972	102,339,471
Caprinae	Sheep	*Ovis aries*	54	Chromosome	2.770	1,015	103,700,880
Caprinae	Takin	*Budorcas taxicolor*	52	Chromosome	2.850	1,362	109,747,698
Hippotraginae	Sable Antelope	*Hippotragus niger*	60	Chromosome	2.590	15,586	100,235,793
Hippotraginae	Addax	*Addax nasomaculatus*	58	Chromosome	3.290	2,442	87,344,204
Hippotraginae	Roan Antelope	*Hippotragus equinus*	60	Chromosome	2.600	15,646	99,087,534
Hippotraginae	Scimitar-Horned Oryx	*Oryx dammah*	58	Chromosome	2.720	11,238	100,398,400
Reduncinae	Defassa Waterbuck	*Kobus ellipsiprymnus defassa*	54	Chromosome	3.150	1,014	98,450,000
Reduncinae	Red Lechwe	*Kobus leche leche*	48	Scaffold	2.770	57,712	3,233,651

4.2.2. Genome alignments

After filtering and masking repeats, a total of 14 genomes were used in the following alignment steps. Pairwise alignments were carried out for each genome against the reference, the chromosome-level waterbuck genome, with LASTZ version 1.04.22 (Harris, 2007) using default parameters (minScore = 1000, -linearGap = medium, C = 0, E = 30, H = 2000, K = 3000, L = 3000, O = 400 and the Q matrix). The outputted alignment files were converted to chain and net files with the following UCSC Kent Utilities (Kuhn et al., 2013; Rhead et al., 2009): axtChain (-psl, -verbose = 0, -minScore = 1000, -linearGap = medium), chainAntiRepeat, chainSort, chainPreNet, chainNet, and netSyntenic. The coverage of the net files were calculated by converting each chromosome net file into a BED file with netToBed UCSC Kent Utilities (Kuhn et al., 2013; Rhead et al., 2009). The BED files were then concatenated and inputted into BEDTools coverage (Quinlan & Hall, 2010), along with a BED file containing chromosome sizes, to calculate the mean coverage of each chromosome for each species. The chain and net files were used to construct HSBs with maf2Synteny (Kolmogorov et al., 2018) at a 300 Kb resolution. The HSBs were then viewed using the R package syntenyPlotteR (Quigley et al., 2023).

4.2.3. Ancestral chromosome reconstructions

To construct the ancestral chromosomes of two nodes on the Bovidae tree, we firstly constructed a phylogenetic tree of all 14 species used in this study. We used TimeTree (Kumar et al., 2022), with species names as input, which outputted a phylogenetic tree in Newick format that was then visualised with iTOL (Letunic & Bork, 2021). The program DESCHRAMBLER (Kim et al., 2017) was then used to reconstruct the ancestral genomes, with the resolution set to 300 Kb, the minimum adjacency score 0.0001, and the Newick file modified to specify the particular node to reconstruct. This resulted in an output file for each species aligned to the reconstructed ancestral chromosome fragments (RACFs). This was visualised with syntenyPlotteR (Quigley et al., 2023) and RACFs were reorientated and reordered based on the orientation of the waterbuck chromosomes.

4.2.4. 3D genome organisation with Hi-C

In order to explore the 3D genome organisation of the waterbuck and compare this to other bovids, we used Hi-C paired-end sequencing data from the common waterbuck (*K. e. ellipsiprymnus*; 2*n* = 52) sample (described in Chapter 2) and the addax (*Addax*

nasomaculatus; SRX16279305), blue wildebeest (*Connochaetes taurinus*; SRX7041756), dama gazelle (*Nanger dama*; SRX5415940), and sable antelope (*Hippotragus niger*; SRX7041758) from NCBI. FASTQ reads were trimmed to remove Illumina adapters and poor quality reads with Adapter Removal v2.3.3 (M. Schubert et al., 2016) with parameters –mm 3, --trimns, --trimqualities, --minlength 25, and --minquality 20. Trimmed reads were mapped to their corresponding reference genomes using BWA-MEM v0.7.17 (H. Li, 2013) with parameters -A1, -B4, -E50, -L0, and paired-end mates were each mapped separately. Samtools view v1.17 (Danecek et al., 2021) was used to convert SAM files into BAM files (parameters -S, -h, and -b).

HiCExplorer v3.7.3 (Ramírez et al., 2018) was then used to process BAM files and create Hi-C matrices for each species. Firstly, restriction enzyme sites (GATC) were located across each reference genome using HiCExplorer hicFindRestSite (Ramírez et al., 2018), creating a BED file. This restriction cut BED file, along with each mapped BAM file, were then input into HiCExplorer hicBuildMatrix along with the parameters –binSize 100000, --restrictionSequence GATC, and –danglingSequence GATC, producing a matrix for each species at 500 Kb resolution. The Hi-C matrices were then normalised to the smallest read number using HiCExplorer hicNormalize (Ramírez et al., 2018). Each normalised matrix was corrected using HiCExplorer hicCorrectMatrix diagnostic_plot and correct, removing bins with very low and very high Z-scores. Only assembled chromosomes were used, due to unassembled scaffolds creating problems with the correction. The Hi-C matrices were visualised with HiCExplorer hicPlotMatrix (Ramírez et al., 2018) and matrix files (.h5) were converted into interactions files (.ginteractions) using hicConvertFormat. Mean interactions per chromosome were calculated between and within each chromosome in R and plotted with ggplot2.

4.2.5. Chromosome rearrangements within waterbuck

The F_{ST} analysis undertaken in Chapter 3 was further analysed to look specifically for regions of the genome that showed "blocks" of high genomic differentiation (F_{ST}), as putative chromosome rearrangements or structural variation between the two species. As previously, F_{ST} was calculated in windows of 10 Kb. Synteny to the Anc1 ancestral chromosomes and synteny to the cattle genome (*Bos taurus*; ARS-UCD2.0) were visualised on the F_{ST} plots using ggplot2 in R. We also carried out PCA analysis of each putative region by firstly filtering the sites file between the start and end coordinates of the region and then specifying this sites file when genotyping with ANGSD v0.940

(Korneliussen et al., 2014). The PCAs were then calculated with PCAngsd v0.99 (Meisner & Albrechtsen, 2018). LD was also calculated in 100 Kb windows across the genome with ngsLD v1.2.1 (Fox et al., 2019).

4.3. Results

4.3.1. Bovidae chromosome evolution

The pairwise alignment of the 13 Bovidae species to the waterbuck chromosome-level genome assembly resulted in a high mean alignment coverage across chromosomes for each species (97.604-97.895%; **Table 18**). Some chromosomes had lower coverage, such as waterbuck chromosomes KEL13 (mean across species was 88.847%) and KEL17 (mean of 84.962%), but this may be due to larger regions of repeats at the start of these chromosomes that may not accurately map (see percentage repeats in Chapter 2; **Figure 32**).

To further assess the alignment quality, we constructed HSBs at 300Kb resolution. These HSBs were visualised using evolutionary highway (EH) plots. Some chromosomes were mostly syntenic, shown by large regions covered by HSBs across all species (e.g., KEL9; **Figure 57**), whilst others presented some chromosome rearrangements (e.g., KEL1), shown by breaks in synteny (breakpoints depicted by the white regions). In KEL1, we can see a fusion of two chromosomes and an inversion in all species. Red lechwe was highly fragmented due to it being a scaffold-level assembly.

Table 18: *Coverage of nets (%) per chromosome for the pairwise alignment of 13 Bovidae species to the waterbuck reference genome.*

	Coverage of Nets (%)													
Chr.	**Roan Antelope**	**Eastern Bongo**	**Blue Wilde-beest**	**Dama Gazelle**	**Red Lechwe**	**Goat**	**Cattle**	**Sheep**	**Takin**	**Gerenuk**	**Scimitar-Horned Oryx**	**Sable Antelope**	**Addax**	**Mean**
KEL1	99.591	99.576	99.588	99.588	99.256	99.595	99.578	99.603	99.631	99.594	99.627	99.601	99.637	**99.574**
KEL2	98.561	99.118	98.735	98.968	98.910	97.959	99.988	98.715	98.767	98.518	98.525	98.792	99.755	**98.870**
KEL3	99.326	99.041	99.859	99.402	99.290	99.240	99.179	99.301	99.983	99.327	99.218	99.309	99.278	**99.366**
KEL4	97.669	97.604	97.647	97.600	97.447	97.547	97.662	97.673	97.674	97.596	97.684	97.670	97.698	**97.629**
KEL5	98.727	98.584	98.698	98.900	98.689	98.721	98.657	98.817	98.818	98.717	98.684	98.728	98.726	**98.728**
KEL6	98.980	99.237	99.229	99.236	99.133	99.279	99.155	99.276	99.277	98.867	99.210	99.257	99.283	**99.186**
KEL7	99.853	99.851	99.841	99.850	99.822	99.840	99.850	99.848	99.844	99.848	99.853	99.854	99.854	**99.847**
KEL8	99.686	99.660	99.659	99.687	99.603	99.693	99.653	99.694	99.660	99.582	99.689	99.686	99.695	**99.665**
KEL9	99.574	99.576	99.634	99.605	99.530	99.577	99.593	99.633	99.613	99.611	99.594	99.588	99.594	**99.594**
KEL10	96.066	96.115	96.035	96.115	96.072	96.063	96.010	96.070	96.060	96.097	96.115	96.115	96.110	**96.080**
KEL11	99.505	99.969	99.807	99.833	99.527	99.980	99.730	99.988	99.973	98.925	99.978	99.948	99.985	**99.781**
KEL12	95.576	95.765	96.091	95.759	96.019	95.908	95.852	95.930	96.094	95.444	96.124	96.237	95.939	**95.903**
KEL13	88.850	88.850	88.848	88.850	88.787	88.851	88.850	88.860	88.857	88.850	88.847	88.850	88.859	**88.847**
KEL14	97.408	97.239	97.444	97.401	97.475	96.368	97.514	97.416	97.415	97.138	97.483	97.472	97.624	**97.338**
KEL15	98.890	98.829	98.897	98.881	98.936	98.888	98.842	98.901	98.891	98.841	98.871	98.869	99.099	**98.895**
KEL16	98.752	98.608	98.616	98.650	98.550	98.752	98.586	98.660	98.778	98.646	98.681	98.677	98.675	**98.664**
KEL17	86.324	86.272	86.321	86.323	86.131	86.304	86.287	86.324	86.323	86.323	68.931	86.322	86.324	**84.962**
KEL18	99.736	99.942	99.874	99.956	99.501	99.961	99.864	99.960	99.959	99.583	99.542	99.959	99.987	**99.833**

KEL19	99.257	99.284	99.367	99.242	98.899	99.426	99.275	99.429	99.433	99.292	99.421	99.425	99.435	**99.322**
KEL20	99.727	99.875	99.935	99.915	99.809	99.721	99.862	99.941	99.941	99.214	99.942	99.937	99.946	**99.828**
KEL21	97.050	96.894	97.034	96.788	96.949	96.973	96.979	97.006	97.024	97.295	97.055	97.040	97.064	**97.012**
KEL22	99.039	98.862	98.627	96.189	98.924	98.905	98.769	99.066	99.046	98.746	98.673	98.950	99.060	**98.681**
KEL23	98.433	98.315	98.340	98.327	98.494	98.649	98.556	98.390	98.415	98.256	98.452	98.336	98.927	**98.453**
KEL24	99.975	99.866	99.857	99.868	99.850	99.933	99.901	99.922	99.979	99.299	99.976	99.981	99.995	**99.877**
KEL25	95.553	95.089	95.418	95.215	95.452	95.351	95.556	95.605	95.612	94.614	95.370	95.672	95.391	**95.377**
KEL26	98.524	98.210	99.406	98.603	98.774	98.193	98.318	98.860	98.505	98.292	98.777	98.806	98.825	**98.623**
KELX	99.937	99.934	99.978	99.940	96.965	95.620	99.980	99.297	99.277	99.866	99.890	99.946	98.411	**99.157**
Mean	**97.799**	**97.784**	**97.881**	**97.729**	**97.659**	**97.604**	**97.853**	**97.859**	**97.883**	**97.644**	**97.193**	**97.890**	**97.895**	

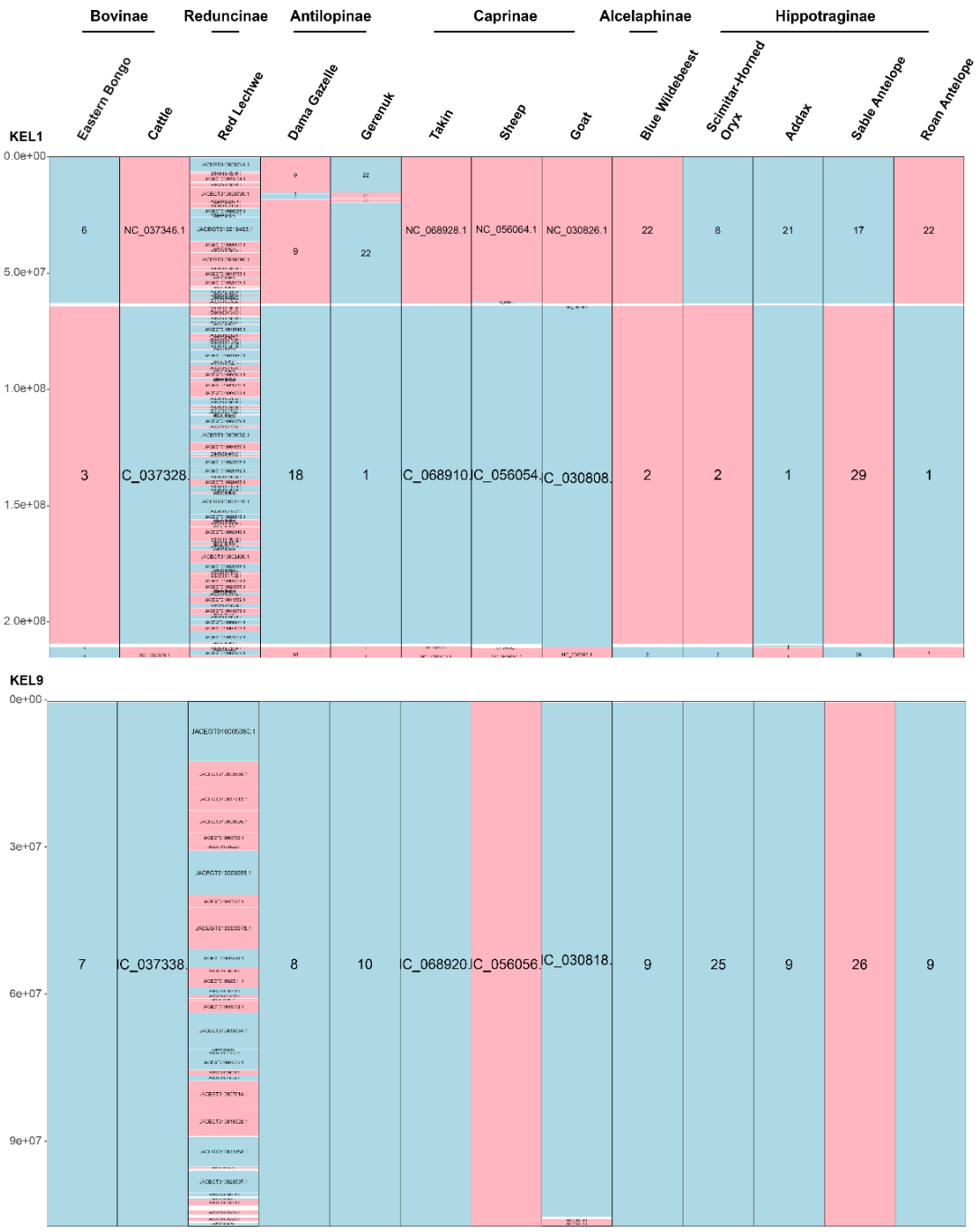


Figure 57: *Evolutionary highway (EH) plot of the synteny of the 13 Bovidae genomes to waterbuck chromosomes KEL1 (top) and KEL9 (bottom). Blue represents homologous synteny blocks (HSBs) with the same orientation as the reference, whilst red the opposite orientation. HSBs are labelled with chromosome numbers/names, and the y axis is the position on the chromosome (bp).*

To study chromosome evolution in antelopes, we then used the tool DESCHRAMBLER (Kim et al., 2017) to reconstruct two nodes on the Bovidae TimeTree phylogenetic tree (Anc1 and Anc2; **Figure 59**). Anc1 is the ancestor of the clade containing all sampled

subfamilies excluding Bovinae (i.e., Reduncinae, Antilopinae, Caprinae, Alcelaphinae, and Hippotraginae) known sometimes as Aegodontia, whilst Anc2 is the ancestor of Reduncinae and Antilopinae. We reconstructed these two nodes to explore the chromosome evolution across the ruminant family Bovidae, but also to allow us to trace the evolution to Reduncinae and the genus *Kobus*.

The two ancestral karyotypes, termed Anc1 and Anc2, had 31 and 32 RACFs, respectively (**Table 19**). However, for Anc1, RACF 31 was very small compared to all other RACFs (3,438,787 bp in length) and always shared synteny to current species with RACF 17, and therefore was merged (denoted as RACF 17a and RACF 17b for clarity), reducing the haploid number to 30 (Anc1 Curated), and which totalled 2,576,902,472 bp in length. The two RACFs as in Anc1 were also merged in Anc2. RACF 32 (471,112 bp) was removed due to its small size and difficulty in merging with another RACF, resulting in a haploid number of 30 and a total ancestral genome length of 2,576,431,360 Kb. The two curated ancestral chromosomes had conserved synteny across all RACFs (**Figure 58**).

Table 19: *Reconstructed ancestral chromosome fragment (RACF) statistics for the two Bovidae ancestors (Anc1 and Anc2), before and after manual curation.*

Ancestor Name	No. RACFs	Min. RACF Size (bp)	Max. RACF Size (bp)	Total Genome Size (Kb)
Anc1	31	3,438,787	158,137,021	2,576,902,472
Anc1 (Curated)	30	42,169,846	158,137,021	2,576,902,472
Anc2	32	471,112	158,512,868	2,612,757,919
Anc2 (Curated)	30	44,088,175	158,512,868	2,576,431,360

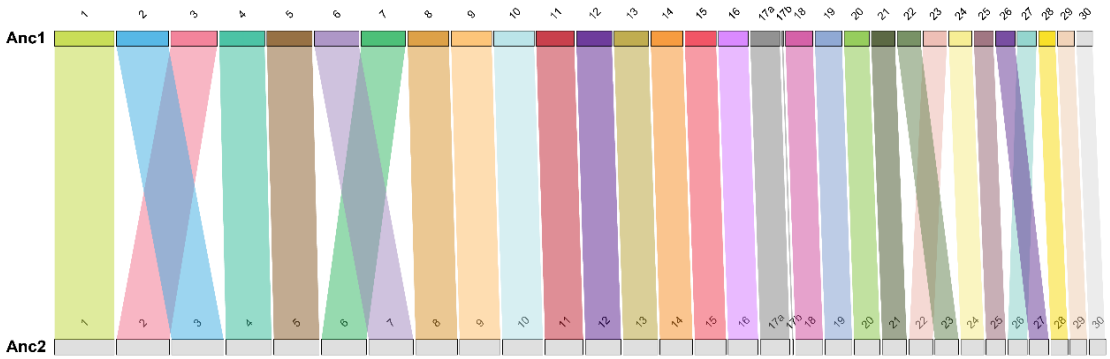


Figure 58: *Synteny between reconstructed ancestral chromosome fragments (RACFs) of the two curated ancestors (Anc1 and Anc2).*

No rearrangements were observed between Anc1 and Anc2. As the synteny was conserved between the two ancestral chromosome reconstructions, we focused on Anc1 and constructed HSBs between this ancestor and the 12 Bovidae species with chromosome-level genomes (**Figure 59**). We showed that chromosome fusions, specifically Rb fusions, were the major contributor to the diversity of chromosome structure and karyotype number in Bovidae, confirming previous results. Taxa within the same subfamily sometimes had very different chromosome configurations, for example in Bovinae, where the eastern bongo had 12 Rb fusions and 1 tandem fusion, whilst cattle only had a small translocation, but overall was conserved to Anc1. Other subfamilies had very few rearrangements, such as Hippotraginae, where two of the sampled species (*Addax* and *Oryx*) had one Rb fusion, whilst in the other two species (the Sable and Roan antelopes) synteny to Anc1 was completely conserved.

Two Rb fusions were shared between species (**Figure 59** and **Figure 60**). Anc1-1;30 was found in two sister species of Hippotraginae (addax and oryx), whilst Anc1-2;30 was found across three species (waterbuck, takin, and wildebeest) and three subfamilies (Reduncinae, Caprinae, and Alecelaphinae, respectively). Six species had Rb fusions involving chromosomes syntenic to Anc1-1 (bongo, waterbuck, takin, sheep, oryx, and addax), six to Anc1-2 (bongo, waterbuck, dama gazelle, takin, sheep, and wildebeest), and six to Anc1-30 (waterbuck, dama gazelle, takin, wildebeest, oryx, and addax). This was followed by four species for Anc1-7 (bongo, waterbuck, takin, and sheep). All Anc1 chromosomes were involved in at least one fusion in the species sampled.

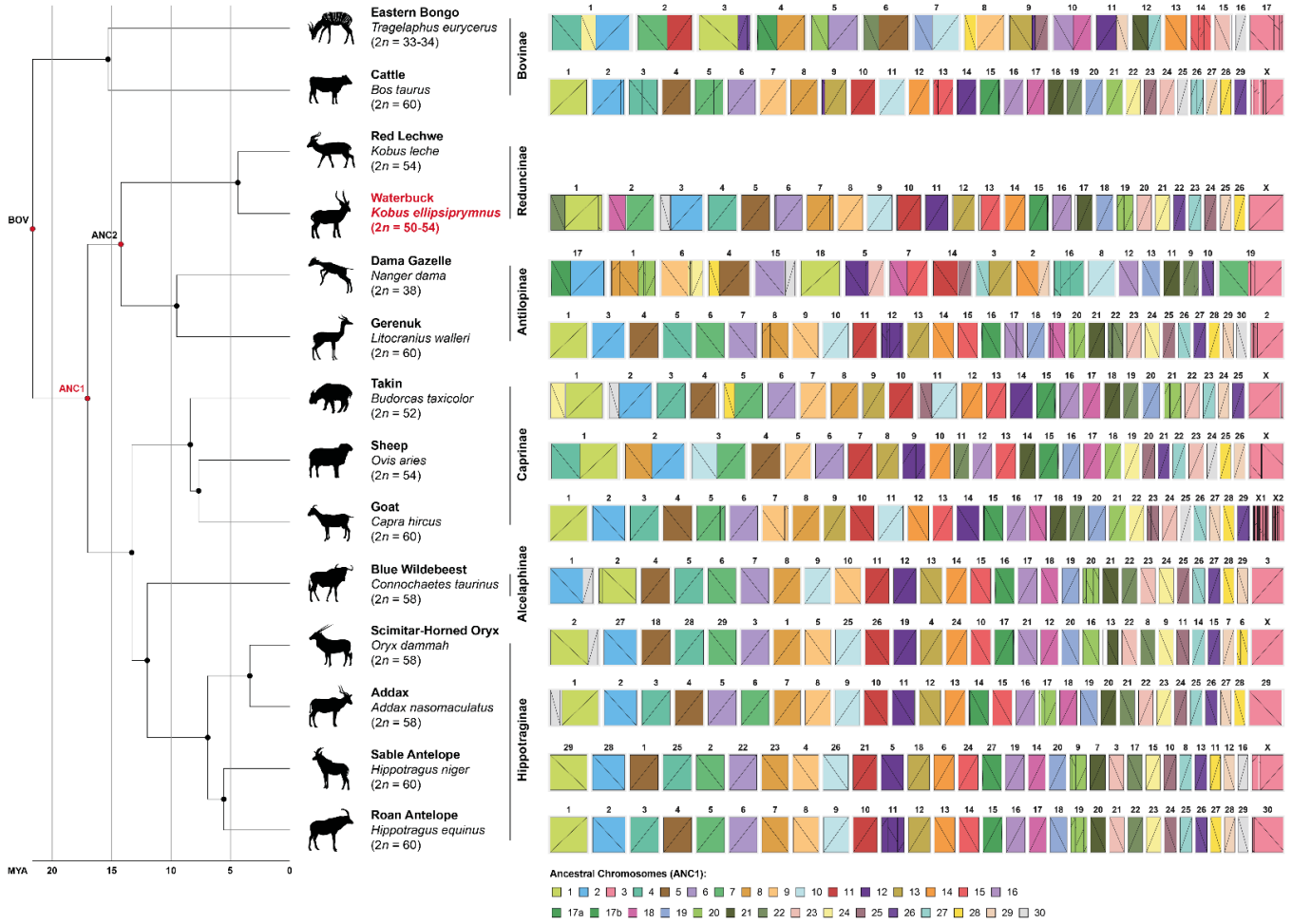


Figure 59: *Phylogenetic tree and synteny between Anc1 and the 13 chromosome-level Bovidae genomes. In the synteny plots, vertical lines represent an evolutionary breakpoint region (EBR), and diagonal lines represent the orientation of syntenic blocks.*

To assess whether the same chromosomes were involved in other Rb fusions in species without their genome available, we looked at previously published research. To do this, we searched the literature for cytogenetic publications which compared a species karyotypes with cattle chromosomes and found 55 species across seven subfamilies (Alcelaphinae, Antilopinae, Bovinae, Caprinae, Hippotraginae, Nesotraginae, and Reduncinae; **Supplementary Table 5**). We then converted the cattle chromosome numbers into the Anc1 chromosome number based on synteny between Anc1 and cattle (**Figure 59**). Using a phylogenetic approach, synapomorphic Rb fusions, shared by multiple species within a subfamily, were only counted once, to avoid overcounting fusions in subfamilies with many sampled species, and to identify fusions that were symplesiomorphic, shared across some of the subfamilies, or homoplasic, convergently evolved in two independent species.

The most common fusions in species from the literature review were Anc1-5;9 and Anc1-2;30 which occurred in four subfamilies, whilst the chromosomes Anc1-4;10, Anc1-2;4, Anc1-14;16, Anc1-10;25, and Anc1-1;11 were found in three subfamilies (**Figure 60**). It is worth noting that not all combinations of chromosomes were involved in the sampled chromosomes. The Anc1 chromosomes involved in the highest number of Rb fusions were Anc1-10 (n = 19), Anc1-6 (n = 18), Anc1-1 and Anc1-5 (n = 17), and Anc1-2 (n = 15). However, all ancestral chromosomes were involved in at least one Rb fusion. All chromosome fusions found in the selected genomes studied were also supported by the literature, but with varying frequencies.

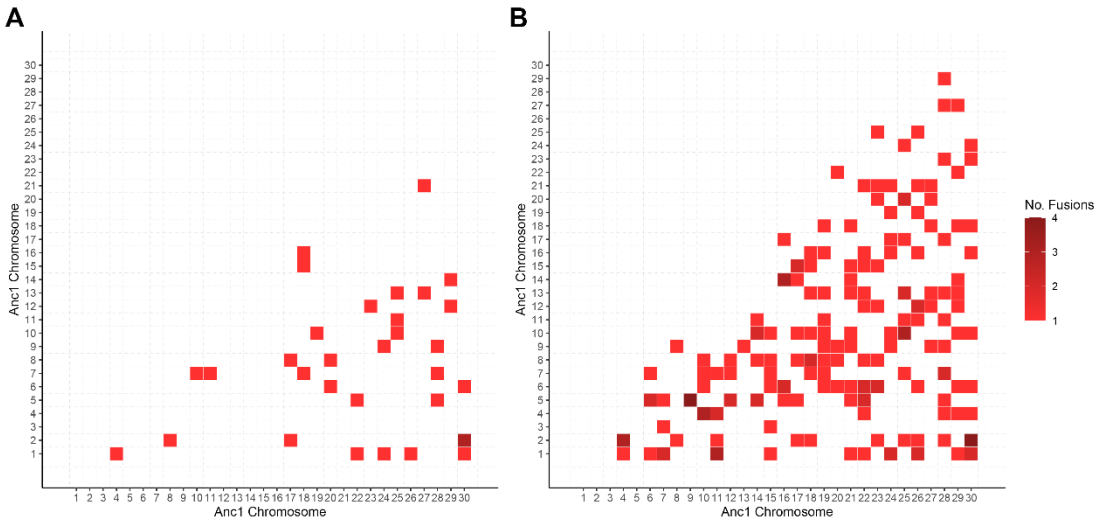


Figure 60: *Number of Robertsonian fusions between two Anc1 chromosomes in the 13 selected extant species analysed (A) and 55 extant species from the literature (B).*

Moreover, it should be noted that chromosome inversions were also discovered between Anc1 and the extant species (**Figure 59**). This includes the ancestral chromosome Anc1-20 which had inversions in several species. In gerenuk, takin, wildebeest, and the sable and roan antelopes, an inversion was shared across species. Whereas dama gazelle, waterbuck, and addax had separate inversions in chromosomes homologous to Anc1-20. There were no inversions found syntenic to Anc1-20 in the studied subfamily Bovinae species. The ancestral chromosomes Anc1-8, Anc1-16, and Anc1-22 all shared inversions between the two studied Antilopinae species (dama gazelle and gerenuk). Dama gazelle had by far the highest number of inversions out of the studied extant species (Anc1-4, Anc1-8, Anc1-12, Anc1-16, Anc1-18, Anc1-20, Anc1-22, Anc1-24, Anc1-26, Anc1-28, Anc1-30). Therefore, whilst Rb fusions may have been the most common rearrangement in Bovidae chromosome evolution, inversions may have also played a part and should be explored further.

4.3.2. 3D genome organisation in waterbuck and other antelopes

We next explored whether the 3D genome organisation in Bovidae is contributing to chromosome fusions being the dominant type of rearrangement in the group's evolution. Five bovids were selected from the 14 sampled in the previous section that had Hi-C data publicly available. This included the addax, wildebeest, dama gazelle, sable antelope, and the common waterbuck (2*n* = 52).

Hi-C sequencing data was quality controlled before and after adapter trimming (**Table 20**). Prior to trimming, the number of sequenced reads ranged from 143,880,368 to 375,493,105, with a duplication level between 12.314% in wildebeest to 48.932% in the sable antelope. After adapter trimming, the number of reads reduced slightly, but the percentage of duplication remained similar.

Paired-end sequencing reads were mapped to their respective genome assembly and HiCExplorer (Ramírez et al., 2018) was used to construct an Hi-C matrix for each species. The percentage of Hi-C contacts ranged from 26.220% (sable antelope) to 52.790% (addax; **Table 21**). The remaining Hi-C contacts were marked as having low mapping quality (13.180-16.640%), one mate not unique (12.890-32.160%), and one mate unmapped (1.830-5.580%). The majority of Hi-C contacts were intrachromosomal, with most representing long-range interactions (≥ 20 Kb distance) between 47.540% to 56.270%, and a small quantity short range interactions (< 20 Kb distance) from 4.980%

to 16.910%. The remaining interchromosomal Hi-C contacts ranged between 26.820% in dama gazelle and 47.480% in wildebeest, showing that the Hi-C data is of good quality.

Table 20: *Quality control of Hi-C sequencing reads before and after adapter trimming. R1 is the forward and R2 the reverse mate. Dup. is the percentage of duplicated reads.*

	Before Trimming			After Trimming		
Sample	**R1 Dup. (%)**	**R2 Dup. (%)**	**Total Reads**	**R1 Dup. (%)**	**R2 Dup. (%)**	**Total Reads**
Addax	22.968	20.429	219,176,075	22.884	20.256	219,108,271
Wildebeest	16.447	12.314	143,880,368	15.861	11.735	143,328,824
D. Gazelle	25.185	20.303	303,717,784	24.642	20.157	303,659,033
S. Antelope	48.932	42.540	353,749,438	43.601	38.840	324,842,689
Waterbuck	25.695	22.396	375,493,105	25.738	23.465	375,306,977

Table 21: *Quality control of Hi-C interactions. Percentage of Hi-C contacts, low mapping quality, one mate not unique, one mate unmapped, interchromosomal interactions, and intrachromosomal interactions (short and long range).*

Sample	**Hi-C Cont. (%)**	**Low Map. Q (%)**	**One Mate Not Unique (%)**	**One Mate Unmap-ped (%)**	**Inter (%)**	**Intra Short Range <20 Kb (%)**	**Intra Long Range ≥20 Kb (%)**
Addax	52.790	13.180	21.980	1.830	39.100	7.350	53.540
Wildebeest	50.220	14.840	12.890	5.580	47.480	4.980	47.540
D. Gazelle	52.530	16.640	17.460	2.110	26.820	16.910	56.270
S. Antelope	26.220	13.830	32.160	2.620	44.560	7.400	48.040
Waterbuck	47.820	13.330	24.600	2.350	42.620	6.440	50.940

A normalised Hi-C matrix was constructed for each species (**Figure 61**) and the mean number of interactions between chromosomes was calculated. Overall, mean intrachromosomal interactions were higher than interchromosomal interactions (> 10; **Figure 62**). Interchromosomal mean interactions between small chromosomes were statistically higher than those between small and large, and those between large chromosomes (P-value < 0.0001; **Figure 63**). This may be unclear in some species on the heatmaps (i.e., dama gazelle and sable antelope) where chromosomes have not been ordered and numerically labelled by size (**Figure 62**).

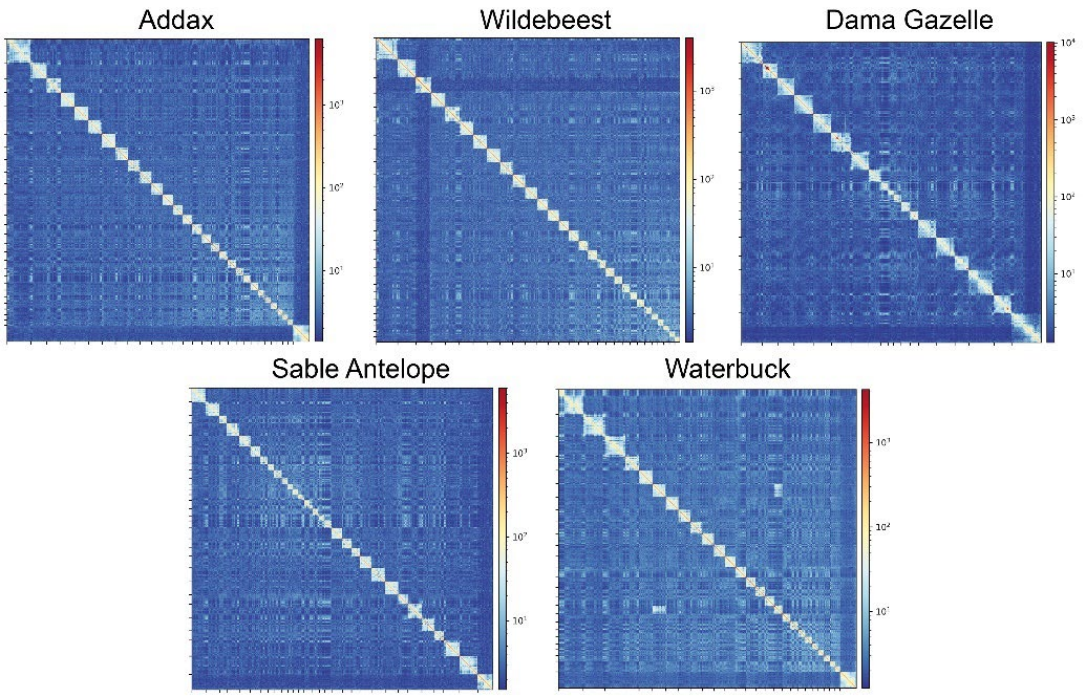


Figure 61: *Hi-C interaction matrices for the five Bovidae species sampled.*

In waterbuck, the fusion of chromosomes KEL6;17 (syntenic to cattle chromosomes BTA6;18) shows a high mean interaction compared with all other chromosomes, due to the Hi-C dataset used being the common subspecies karyotype of 2*n* = 52 and the reference the defassa subspecies karyotype of 2*n* = 54 (**Figure 62**). This is also visible in this Hi-C matrix, with interchromosomal interactions between the two chromosomes (**Figure 61**). However, the other polymorphic fusion in common waterbuck, KEL8;9 (syntenic to cattle BTA7;11), did not have high interchromosomal interactions (**Figure 61** and **Figure 62**), providing the first line of evidence that chromosomes involved in Rb fusions were not necessarily in close proximity in the 3D organisation of the nucleus.

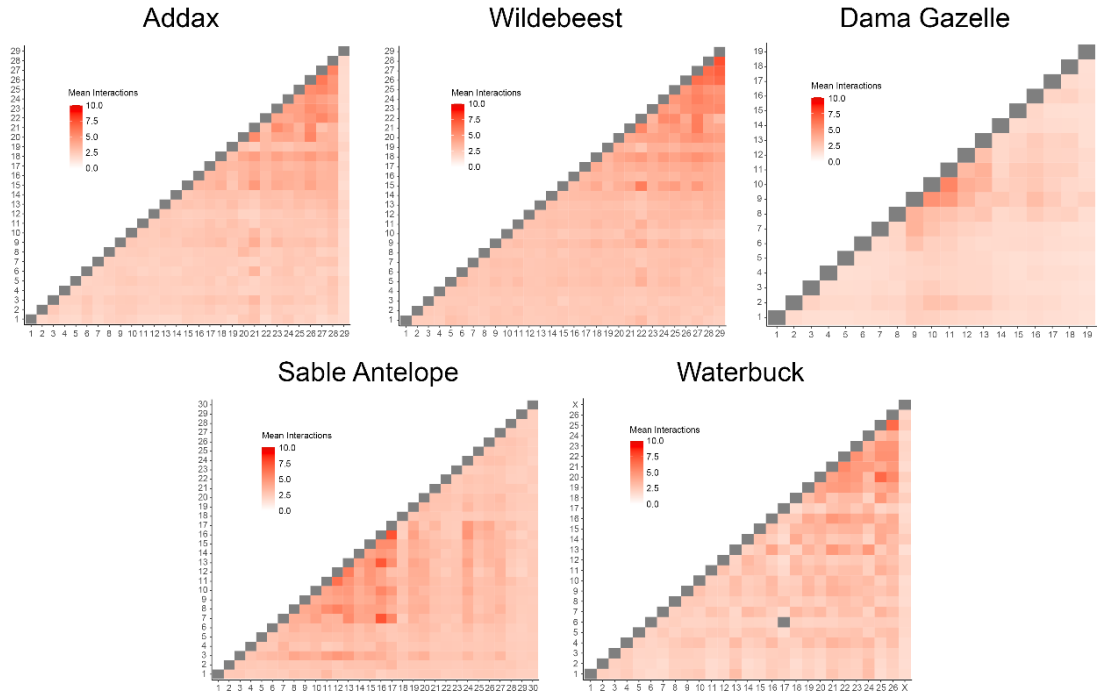


Figure 62: *Mean interactions from the Hi-C matrices for the five species sampled.*

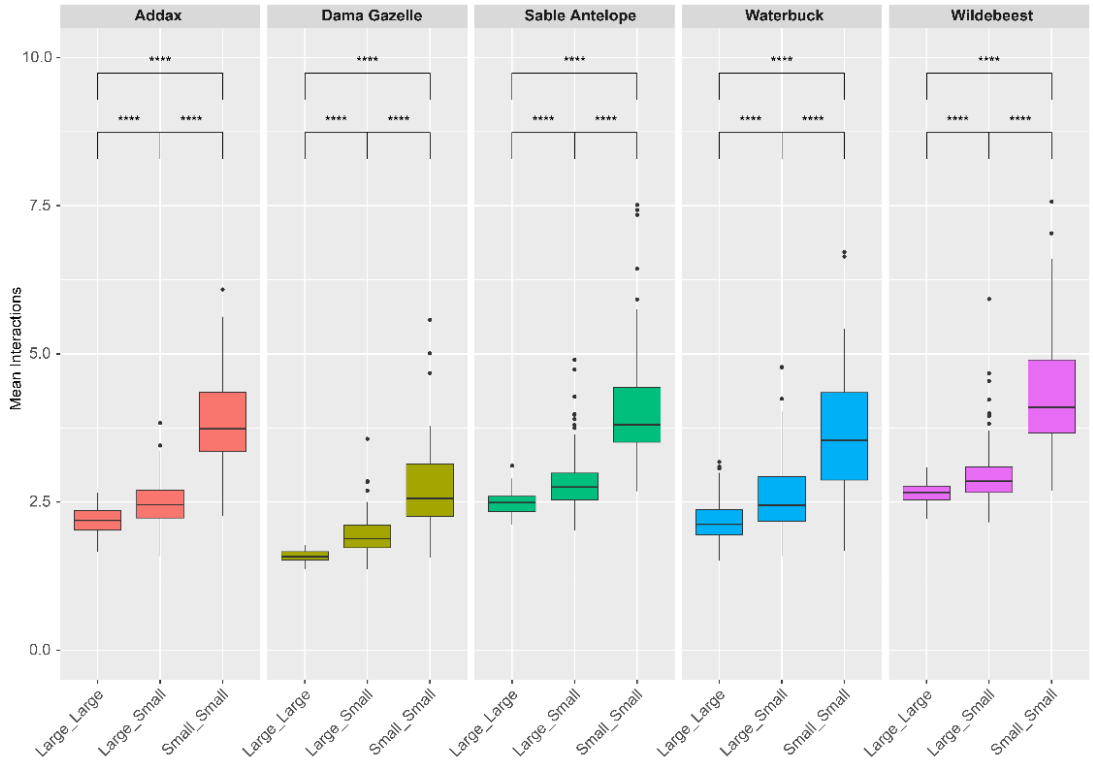


Figure 63: *Mean interchromosomal interactions between large, large and small, and small chromosomes for each of the five Bovidae species sampled. Large chromosomes were defined as those greater than or equal to the median chromosome size, whilst small chromosomes less than the median size.*

To further assess whether the 3D genome configuration of the nucleus is related to an increase in Rb fusions in antelopes, we tested if Anc1 chromosomes often involved in Rb fusions in the extant species (**Figure 60**) were also in close 3D proximity in species lacking the Rb fusion. To do so, we looked at the most common Rb fusions (Anc1-2;30 and Anc1-5;9) and the average interchromosomal interactions between the two chromosomes in species lacking the Rb fusion (**Figure 62**), and compared this to the average interchromosomal interactions between chromosomes of similar length (i.e., small or large chromosomes; **Figure 63**), as a way of correcting for differences in interactions between small and large chromosomes.

When we looked at the chromosomes syntenic to Anc1-2;30 in extant species that lack the fusion (NDA15 and NDA17 in dama gazelle, ANA1 and ANA2 in addax, and HNI16 and HNI28 in sable antelope) or in Anc1-5;9 (KEL5 and KEL8 in waterbuck, CTA4 and CTA10 in wildebeest, ANA4 and ANA8 in addax, and HNI4 and HNI8 in sable antelope) we found lower mean interactions between these chromosomes in each species than the median interchromosomal interaction between all chromosomes of a similar size in the particular species (**Table 22**). This further suggests that the proximity of chromosomes in the nucleus of somatic cells does not increase the likelihood of chromosomes fusing.

Table 22: *Mean interactions between chromosomes involved in the most common Robertsonian fusions (Anc1-2;30 and Anc1-5;9) in Hi-C samples lacking the fusion (from **Figure 62**) and the median interchromosomal interactions between chromosomes by sizes in the given species (from **Figure 63**).*

Anc1	Extant Species	Mean Interactions Between Chromosomes	Chromosome Sizes	Median Interactions by Size
Anc1-2;30	NDA15;17	1.514	Large_Large	1.577
Anc1-2;30	ANA1;2	1.852	Large_Large	2.191
Anc1-2;30	HNI16;28	2.522	Large_Small	2.751
Anc1-5;9	KEL5;8	2.050	Large_Large	2.125
Anc1-5;9	CTA4;10	2.596	Large_Large	2.664
Anc1-5;9	ANA4;8	2.042	Large_Large	2.191
Anc1-5;9	HNI4;8	2.714	Large_Small	2.751

4.3.3. Understanding the relationship between chromosome rearrangements and genetic differentiation within a species

As well as interspecies chromosome rearrangements, some antelopes also have intraspecies chromosome variation between subspecies or populations, which may impact on speciation. We used the waterbuck as a model antelope with known polymorphic Rb fusions between and within the two subspecies (common and defassa) to investigate the impact of these on genetic differentiation and speciation. We extended this to also look for putative signatures of structural variation between the two subspecies.

To do so, we utilised the WGS data and F_{ST} analyses from Chapter 3, as well as the chromosome-level genome assembled in Chapter 2. Focusing on each waterbuck chromosome independently, we identified blocks with elevated F_{ST}, pinpointing signatures of putative chromosome rearrangements. To support that these regions with elevated F_{ST} might be caused by chromosome rearrangements or structural variants, we calculated linkage disequilibrium (LD) in 100 Kb windows across the genome, and further investigated population structure in these regions using PCA.

First, we focused on chromosomes that have been involved in fusions in this antelope. We found two genomic regions with elevated blocks of F_{ST} on waterbuck chromosome KEL3, surrounding the centromere of the submetacentric chromosome (**Figure 64**, **Supplementary Table 6**). This chromosome is the result of an ancestral fusion between Anc1-2;30 (BTA2;25), which is now fixed in waterbuck, and other *Kobus* species (S. C. Kingswood et al., 2000). The chromosome showed two blocks of high F_{ST} either side of the centromere, ranging from 1,850,000 bp to 2,980,000 in size. Higher LD was found on the left-hand side of the centromere in the common subspecies (as well as one high LD window in the defassa), and higher LD was also found in the common subspecies on the right-hand side, indicating that recombination was lower in these regions.

We additionally computed a PCA of each of the two regions surrounding the centromere and found that PC1 explained between 61.440% and 65.620% of the variation. Three main groupings containing the same populations were present on PC1 of both PCAs. PCAs of regions with elevated F_{ST} most commonly show three different groups, potentially indicating the three genotypes: homozygous dominant, heterozygous and homozygous recessive for the CR or SV. The group with the lowest PC1 values contained defassa waterbuck from Angola, Cameroon, Chad, DRC, Ethiopia, Guinea-Bissau, Kafue, Kenya, KVNP, Maswa, QENP, Samole, South Sudan, and Ugalla. In the centre

group, both subspecies were present and included the populations of Nairobi (common), Samburu (common), and Ugalla (defassa). The third group, with the highest values on PC1, contained mostly common waterbuck from the populations of Luangwa, Matetsi, Nairobi, Samburu, Somalia, Tanzania, as well as two defassa individuals from Ugalla. Our data points to either a direct effect of the fixed chromosome fusion, or novel chromosome rearrangements near the centromere of this chromosome, on the genetic differentiation, recombination, and population structure between the two subspecies.

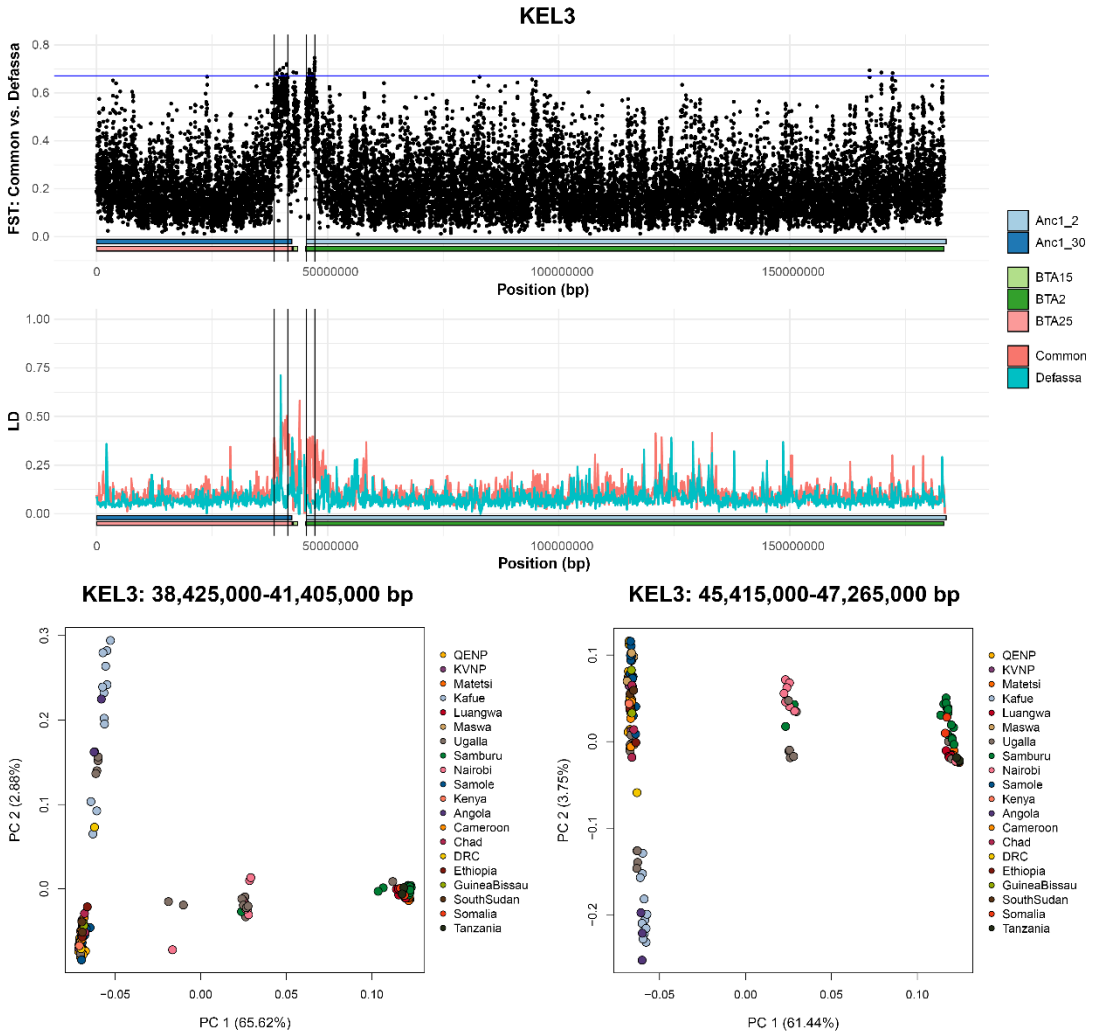


Figure 64: *Waterbuck chromosome KEL3. Genomic differentiation (F_{ST}) was calculated in 10 Kb windows between the two subspecies, linkage disequilibrium (LD) was calculated in 100 Kb windows for each subspecies, and a principal component analysis (PCA) was computed for each block of high F_{ST}. Horizontal bars show the synteny to Anc1 and cattle (BTA) chromosomes. Vertical lines represent the regions of interest.*

Within these two regions we found 128 annotated genes (**Supplementary Table 7**). The first block contained 113 genes, including UBN1 (F_{ST} = 0.627) which binds to

proliferation-promoting genes, the gene PRSS21 (F_{ST} = 0.643) that may regulate proteolytic events during the maturation of testicular germ cells, SEP12 (F_{ST} = 0.621) which is involved in the morphogenesis of sperm heads and the elongation of sperm tails, IFT140 (F_{ST} = 0.720) which develops and maintains outer segments of rod and cone photoreceptor cells, and UNK (F_{ST} = 0.630) involved in cortical neurons during development of the embryo. Genes in the first block were only statistically overrepresented for one GO term, presynapse (9/365 genes, ratio = 0.025, FDR = 0.023).

The second F_{ST} block on KEL3 contained 15 genes and included OCA2 (F_{ST} = 0.626) which had previously been reported to be differentiated between the two subspecies (X. Wang et al., 2024). This gene is part of the mammalian pigmentary system and has been suggested to be a reason for differences in coat fur colour between the two subspecies. The region also contained the gene TUBGCP5 (F_{ST} = 0.628) important for the nucleation of microtubules at the centrosome. There were no GO terms showing statistical overrepresentation in this block.

We also found two blocks of high F_{ST} at the starts of chromosomes KEL6 (3,110,000 bp) and KEL17 (1,830,000 bp), the two chromosomes involved in the polymorphic Rb fusion within waterbuck (syntenic to Anc1-6;21 or BTA6;18; **Figure 65**, **Supplementary Table 6**). Defassa waterbuck have been reported to be either heterozygous for the fusion or have the wt karyotype, whilst the fusion is homozygous and fixed within the common subspecies. High LD was found in both regions, with higher levels in the common subspecies. As these regions occur near the centromeres of the two chromosomes, with potentially low recombination in common waterbuck with the fixed homozygous fusion, these results may indicate two putative signatures of the Rb fusion. However, the results could also indicate separate SVs.

A PCA of each region also shows different population structuring than the PCA of all sites (**Figure 44**), with three main groups on PC1 but with a few individuals positioned between groups. The group with the lowest PC1 values contained only defassa waterbuck populations (Angola, Cameroon, Chad, DRC, Ethiopia, Guinea-Bissau, Kafue, KVNP, Maswa, Nairobi, QENP, Samole, South Sudan, and Ugalla). The second group with intermediate values contained mostly common waterbuck individuals from Luangwa, Matetsi, Nairobi, Samburu, and defassa waterbuck from Ugalla. The third group with the highest PC1 values contained only common waterbuck (Luangwa, Matetsi, Nairobi, Samburu, Somalia, Tanzania).

Our PCA for KEL6 and KEL7 (**Figure 65**) may show groupings based on the karyotype of the fusion. For example, a group containing defassa without the fusion (the wildtype), a group containing defassa heterozygous for the fusion (the intermediate group), and a group with common waterbuck that are homozygous for the fusion. Our PCAs agree with the wildtype (containing only defassa) and the homozygous (containing only common) groupings, however the intermediate group (the heterozygous individuals) contained both of the subspecies. The heterozygous fusion has not been identified in the common waterbuck, but this may be due to lack of sampling in the wild. However, this could also suggest that the two F_{ST} blocks seen are instead signatures of a novel structural variation between the subspecies, and direct methods are needed to confirm this.

The region on chromosome KEL6 contained 16 genes (**Supplementary Table 8**), including APELA (F_{ST} = 0.565) which is involved in cardiovascular formation and differentiation during early development and cardiovascular homeostasis in adults. NPY5R (F_{ST} = 0.305) functions as a receptor for neuropeptide Y and peptide YY and may be implicated with food intake. The highest F_{ST} window (0.630) in this region contained the gene EXOSC9 which is a component of the exosome. A total of 12 GO terms were statistically overrepresented in the F_{ST} block on KEL6 (**Figure 66**). These included snRNA 3'-end processing, the TRAMP-dependent tRNA surveillance pathway, nuclear mRNA surveillance, mitochondrial respiratory chain complex III assembly, and exonucleolytic trimming for Biological Process. For Molecular Function, the GO term pre-ribosome binding was statistically overrepresented, whilst the cytoplasmic and nuclear exosome (RNase complexes) and the cyclin A2-CDK2 complex were statistically significant for Cellular Component.

In the F_{ST} block on chromosome KEL17 there were 42 genes (**Supplementary Table 9**). Several genes were involved with embryonic development, including CFDP1 (window F_{ST} of 0.603), BCNT (0.579) and the ruminant specific gene duplication of the ancestral BCNT gene (P97BCNT; F_{ST} = 0.232), and GABARAPL2 (0.384). The latter gene is required for the sonic hedgehog pathway which plays a role in cell differentiation in the embryo. The gene also plays a role in ciliogenesis. The region also contained a gene which regulates telomere recombination (TERF2IP; F_{ST} = 0.532) by repressing homology-directed repair and influencing telomere length. Additionally, the block was statistically overrepresented for genes involved in the PAS complex (2/2 genes, ratio = 1, FDR = 0.003).

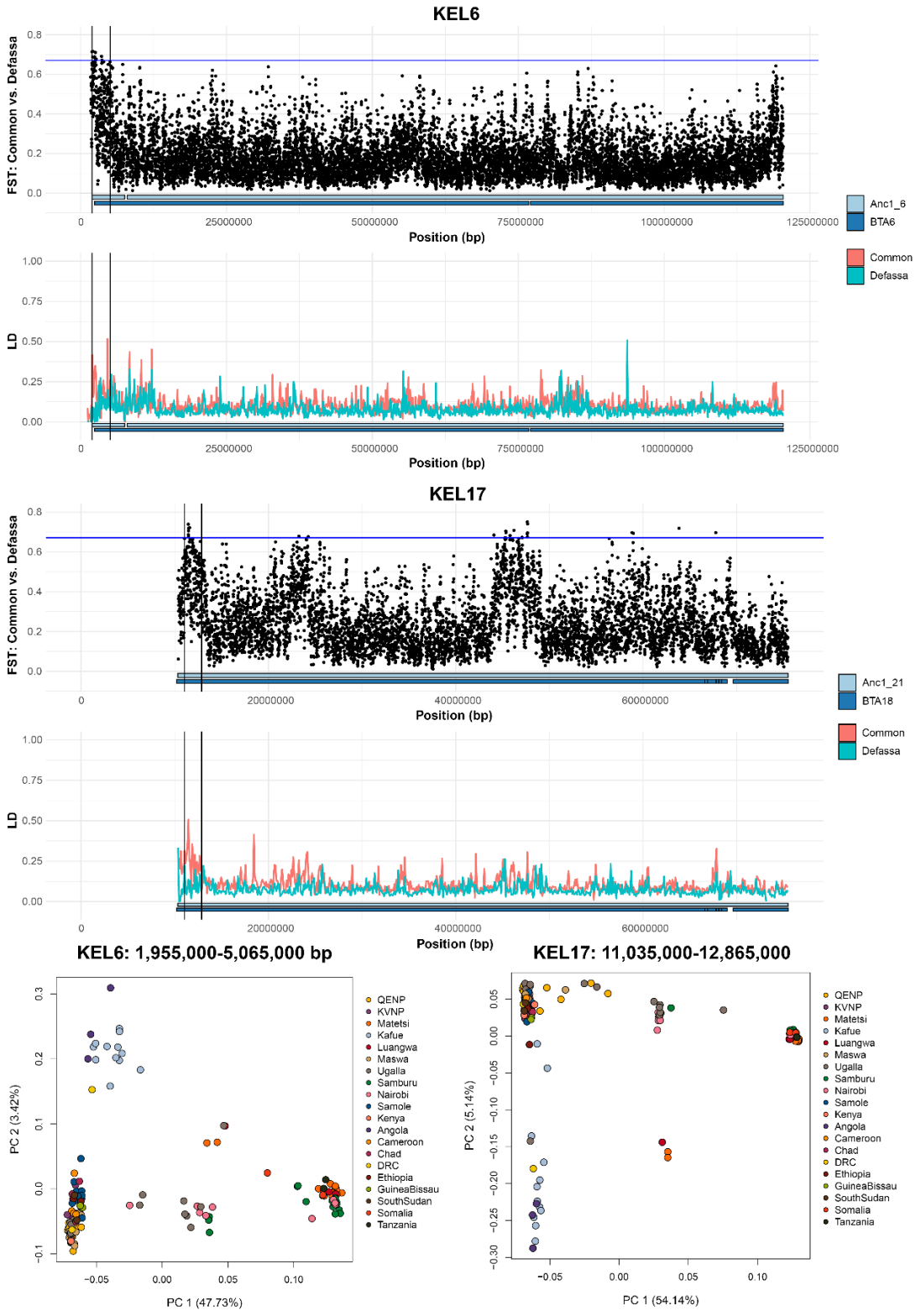


Figure 65: *Waterbuck chromosomes KEL6 and KEL17. Genomic differentiation (F_{ST}) was calculated in 10 Kb windows between the two subspecies, linkage disequilibrium (LD) was calculated in 100 Kb windows for each subspecies, and a principal component analysis (PCA) was computed for each block of high F_{ST}. Horizontal bars show the synteny to Anc1 and BTA. Vertical lines represent regions of interest.*

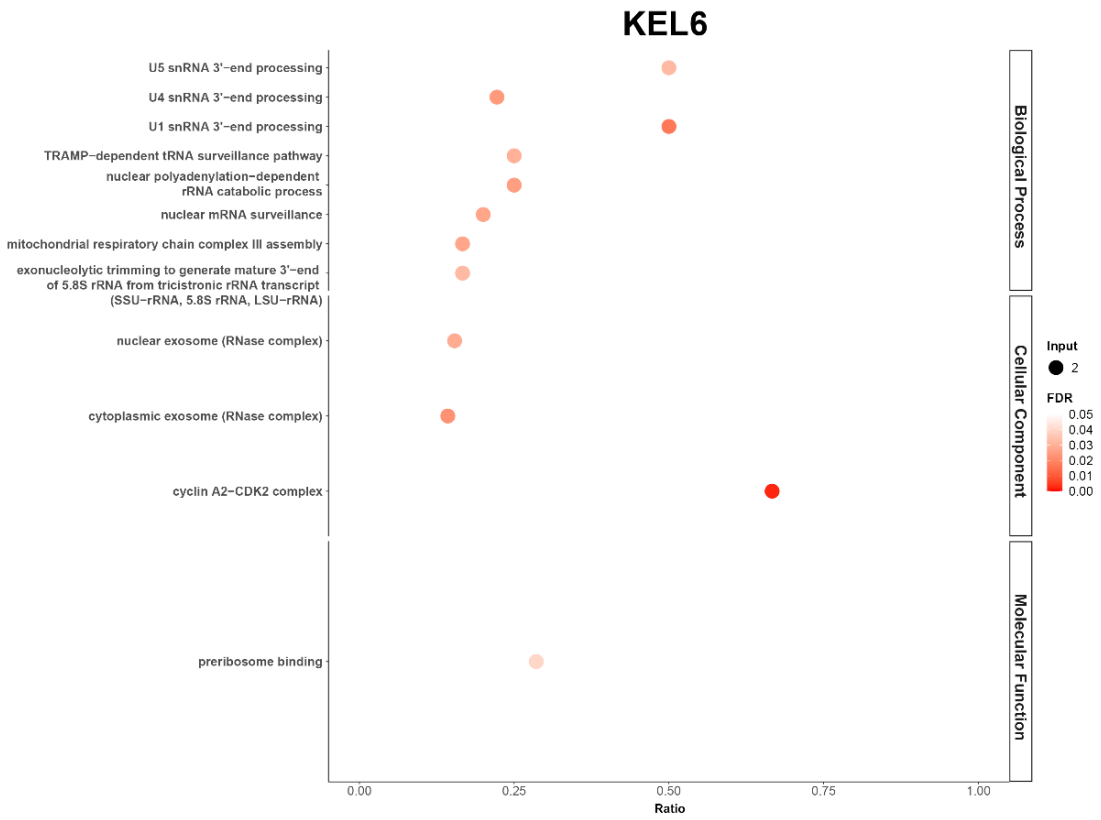


Figure 66: *Gene Ontology (GO) statistical overrepresentation of genes located in the blocks of high genomic differentiation (F_{ST}) on KEL6. The ratio is the number of input genes (input) out of the number of genes for a particular GO term.*

It should be noted that we did not find the same signatures for the other polymorphic Rb fusion in waterbuck, between KEL8 and KEL9 (BTA7;11 or Anc1-9;10; **Figure 67**). We do not find particularly high F_{ST} regions near the centromeres of these chromosomes, and only a small region on KEL8 with high LD in the common subspecies. However, the fusion is not present in the defassa or in common waterbuck with the 2*n* = 52 karyotype, and it is only present in individuals with the heterozygous (2*n* = 51) or homozygous (2*n* = 50) fused karyotype. These fused karyotypes could be less sampled in our dataset, and therefore the majority of individuals in the two subspecies would have the wt chromosomes, resulting in low F_{ST}. However, this was not the case in the previous cytogenetic study (S. Kingswood et al., 1998), where the fused karyotypes in the common waterbuck (2*n* = 50 and 2*n* = 51) were more common than the wild type (2*n* = 52). But this could be due to sampling mostly captive samples. As the coordinates of these chromosome regions could not be accurately identified, a PCA was not constructed.

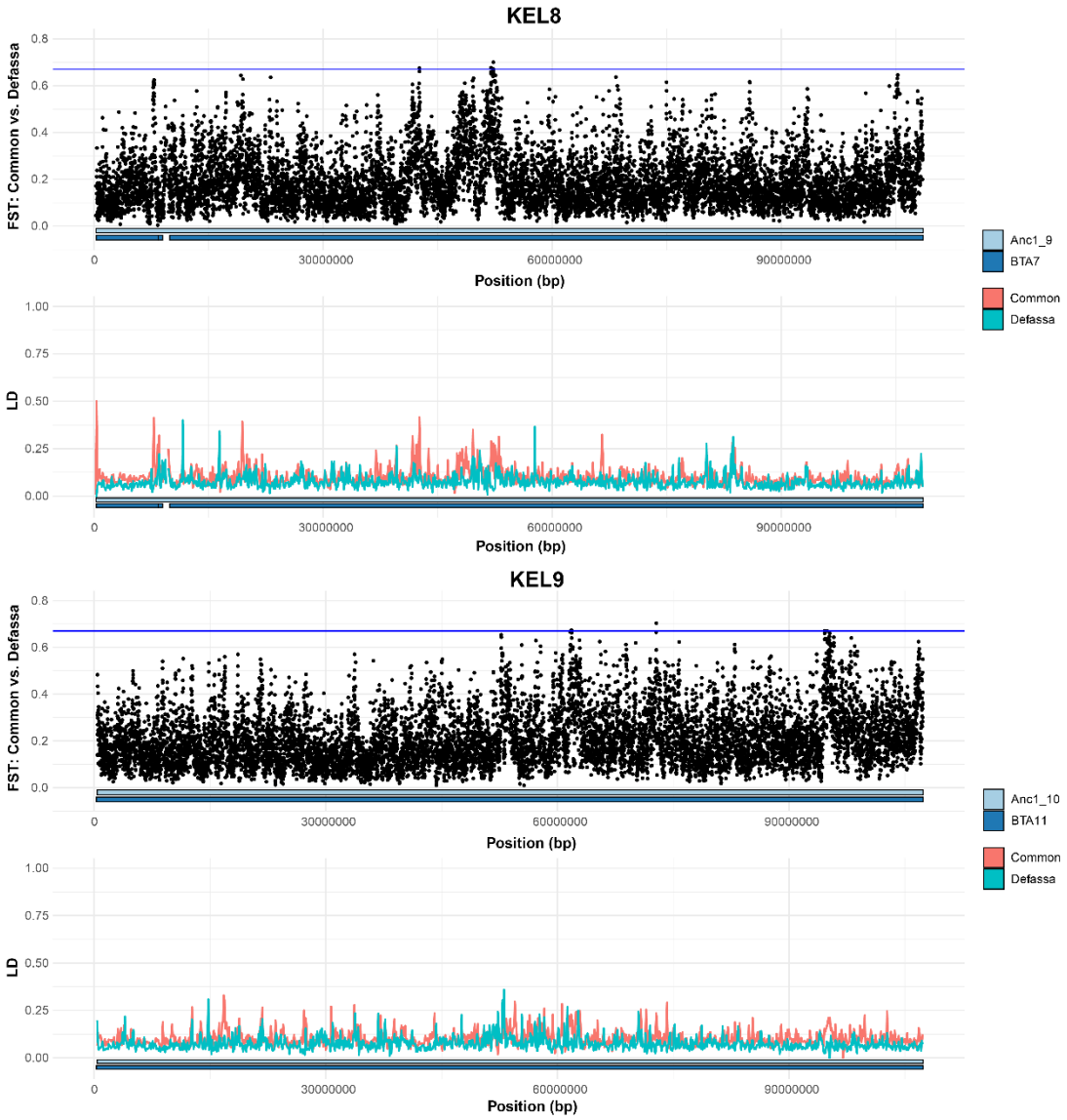


Figure 67: *Waterbuck chromosome KEL8 and KEL9. Genomic differentiation (F_{ST}) was calculated in 10 Kb windows between the two subspecies, linkage disequilibrium (LD) was calculated in 100 Kb windows for each subspecies, and a principal component analysis (PCA) was computed for each block of high F_{ST}. Horizontal bars show the synteny to Anc1 and BTA chromosomes.*

Furthermore, we found more regions in the genome showing blocks of high F_{ST} that were not in close proximity to the centromeres and could be signatures of structural variation between the two subspecies. We also investigated these regions by calculating LD in 100 Kb windows and PCAs of the genomic sites within the selected regions (**Supplementary Figure 6**, **Supplementary Table 6**). This resulted in at least 13 putative SVs with high F_{ST}, different population structure to the whole genome PCA (often grouping into three groups on PC1 with a high percentage of variance), and some blocks also having high LD.

Sizes of these blocks ranged from 630,000 bp on KEL21 to 5,540,000 bp on KEL8. This provides regions in the genome to further investigate and confirm the number and type of structural variation, and the impact of this on speciation between the two subspecies.

4.4. Discussion

In this chapter, we expanded on previous cytogenetic and genomic studies to closely examine chromosome evolution within Bovidae. We firstly studied interspecies chromosome rearrangements using available chromosome-level genome assemblies and reconstructed the ancestral chromosomes of two nodes on the Bovidae phylogenetic tree. The ancestral chromosome reconstructions both had *n* = 30 chromosomes after curation, which may suggest that the Bovidae ancestor also had *n* = 30 chromosomes, as found in Farré et al., 2019.

This allowed us to study the synteny between the oldest ancestor and the 13 extant chromosome-level genomes. We were not able to use the red lechwe scaffold-level genome because it had a low scaffold N50 size. As previously found, the most common rearrangement in Bovidae was Robertsonian fusions. We show that some taxa have extensive Rb fusions, whilst others, even in the same subfamily, have a conserved ancestral karyotype (**Figure 59**). Interestingly, the Rb fusion of Anc1-2;30 was shared across three subfamilies in the studied species, and Anc1-1, Anc1-2, and Anc1-30 were the chromosomes most commonly involved in fusions across the clade.

This led us to hypothesise that some chromosomes, and chromosome pairs, were more likely to be involved in Rb fusions than others. We checked the literature for previous cytogenetic publications to further increase the sampling of taxa and identified the homologous chromosomes in the ancestor (Anc1). The most common fusions were Anc1-2;30 and Anc1-5;9, and the most common chromosomes involved in fusions were Anc1-10 (n = 19), Anc1-6 (n = 18), Anc1-1 and Anc1-5 (n = 17), and Anc1-2 (n = 15; **Figure 60**). Whilst all chromosomes were involved in at least one Rb fusion in Bovidae evolution, this does suggest that some chromosomes were more likely to fuse than others. This could be due to similarities in the p-arms of these chromosomes, or as suggested by the Integrative Breakage Model, their 3D genome organisation (Farré et al., 2015; Guarracino et al., 2023).

It has been proposed that a combination of sequence homology between acrocentric chromosomes, recombination initiation during meiosis, and physical proximity in the nucleus are needed for Rb fusions to form (Gerton, 2024). In human genomes the short p-arms of the acrocentric chromosomes contain the nucleolus organiser regions (NORs), made up of rDNA genes, that are pseudo-homologous between acrocentrics and have been suggested to be associated with Rb fusions (Guarracino et al., 2023). Further studies in the p-arms of human acrocentric chromosomes have found that a combination of homologous SST1 repeats, an inversion, close proximity in 3D space due to NORs, and meiotic recombination hotspots have enabled the formation of Rb fusions (Gomes de Lima et al., 2024). However, in bovids NORs are mostly found in the telomeres (H. Cernohorska et al., 2012), except in Thomson's gazelle (Halina Cernohorska et al., 2015), and therefore the p-arms of bovids must contain other homologous repetitive regions.

To assess whether the location within the nucleus could influence the propensity of forming Rb fusions, we studied the locations of the most commonly fused chromosomes in species that do not harbour these fusions. Using Hi-C data available for several bovids, we explored whether the most common Rb fusions had higher interchromosomal interactions in species without the rearrangement, and therefore were in closer proximity within the nucleus. However, we found that the most common Rb fusions in Bovidae were not in closer proximity in species without the fusions (**Table 22**). This suggests that the 3D genome organisation in somatic cells does not influence chromosome rearrangements. This was also the case for the polymorphic KEL8;9 (BTA7;11) chromosome fusion in waterbuck, which did not have higher interchromosomal interactions in the 2*n* = 52 Hi-C dataset (**Figure 62**). Our data supports a previous study using genomic and Hi-C data from several muntjac deer, which have fixed chromosome fusions between species, that found there was no increase in chromosome interactions in individuals without the fusion (Yin et al., 2021).

These results could suggest that the 3D genome organisation within the nucleus of germline cells is different to somatic cells. Previous studies have found that Rb fusions affect heterologous interactions, chromosomal synapsis, and recombination in the germline in mice (Vara et al., 2021). The study found a reduced crossing-over frequency in homozygous Rb mice and higher F_{ST} estimates than standard mice. There was also different interchromosome interactions in the fibroblasts compared to the spermatids.

Using our approach in the germline could uncover the role of chromosome proximity in Rb fusions.

We then used the waterbuck as a model to investigate intraspecies chromosome rearrangements. We used indirect methods (F_{ST}, LD, and PCA) to detect putative chromosome rearrangements and structural variants between the two subspecies. Firstly, we found blocks of high F_{ST} around centromeres of chromosomes with fixed and polymorphic chromosome fusions. On waterbuck chromosome KEL3 there were two blocks of high F_{ST} either side of the centromere, with higher LD and a PCA clustering into three groups on PC1, a signature of a CR or SV (**Figure 64**). The Anc1-2;30 (BTA2;25) chromosome fusion is fixed in all Kobus species (S. C. Kingswood et al., 2000), and therefore it suggests that it could either be a signature of this historical fusion or novel SVs. One of the regions contained the gene OCA2, which had previously been found to be differentiated in waterbuck (X. Wang et al., 2024). The gene IFT140 was also found, which may play a role in sperm flagella formation and therefore impact male fertility (Zhang et al., 2018), with sperm morphology and motility an important trait during speciation.

We found the same signatures near the centromeres of the two chromosomes involved in Rb fusions between the two subspecies, KEL6 and KEL17 (syntenic to Anc1-6;21 or BTA6;18; **Figure 65**). In the defassa, individuals either have the standard karyotype or are heterozygous for the fusion, whilst in common the homozygous fusion is fixed (S. Kingswood et al., 1998). This could therefore be a signature of the fusion. In the region on KEL6 we found the genes APELA, involved in early embryogenesis (Chng et al., 2013; Norris et al., 2017; Pauli et al., 2014) and NPY5R which plays a role in food intake and appetite (Raposinho et al., 2004), whilst on KEL17 we found several genes involved in embryonic development. These included CFDP1 within the BCNT gene superfamily that is involved in chromatin organisation, condensin recruitment, cell cycle progression, and craniofacial development (Messina et al., 2015, 2017). We also found the gene TERF2IP/RAP1 which is associated with the shelterin complex of telomeres, and that when removed results in an increase in telomere recombination by homologous directed repair and increased fragility (Martinez et al., 2010). This may play a role in the Rb fusions in this species, as shelterin protects telomeres from DNA damage and fusion events (De Lange, 2005), with telomere loss or inactivation required for the formation of Rb fusions (Slijepcevic, 1998). Moreover, we also found 13 additonal blocks of high F_{ST} across the genome, with some having signatures of putative CRs or SVs (**Supplementary Figure 6**).

Further work is needed in order to confirm these putative rearrangements and could include ultra-long read ONT sequencing for each subspecies or karyotype.

5. General Discussion

5.1. Genomics and evolution of antelopes

In this thesis, we aimed to use the latest genomic techniques to further uncover evolution within antelopes, with a focus on the waterbuck (*Kobus ellipsiprymnus*). We firstly sequenced and assembled a high-quality chromosome-level genome for the waterbuck, which meets the VGP standards (Rhie et al., 2021). A high-quality reference genome is essential to accurately study evolution within and between species.

We utilised this reference to study the population genomics of the waterbuck, where combined with WGS data from historical and modern samples, we were able to uncover finer-scale population structure, admixture, and gene flow between populations across the entire species distribution in Africa, than previously studied (X. Wang et al., 2024). Our increase in sampling enabled the evolutionary dynamics between novel populations to be studied and the new reference genome enabled us to find regions across chromosomes that were differentiated between the two subspecies. Using homology-based gene annotation we found regions of high genomic differentiation containing genes that may be related to early speciation.

Placing this genomic differentiation across the genome into a chromosome context, with the support of karyotyping and synteny, enabled us to show that some of the chromosomes involved in fixed and polymorphic Rb fusions in waterbuck had high genomic differentiation near their centromeres. This suggests that Rb fusions may have impacted differentiation between the two subspecies, and with a potential impact on LD and recombination. These regions also contained genes that may influence speciation between the two subspecies, including those involved in spermatogenesis and embryonic development. We also found additional blocks of high F_{ST} across chromosomes, which could be putative SVs between the subspecies.

Finally, we explored chromosome evolution across antelopes, supporting previous cytogenetic studies, and found that the family Bovidae has been dominated by interchromosomal Robertsonian fusions, shaping the variation in karyotype of species from $2n = 30$ to $2n = 60$ (Arias-Sardá et al., 2023). By estimating two ancestral karyotypes within Bovidae, both of which had 30 acrocentric chromosomes, our study suggests that Rb fusions have occurred relatively recently in the family's evolution, with some subfamilies and genera having variable karyotypes and rearrangements. This suggests that Rb fusions may a have played a role in the evolutionary radiation within this clade. Whilst some species have more extensive Rb fusions than others, we also find that some

fusions and chromosomes have occurred more often. However, we did not find a correlation with these common Rb fusions and the 3D genome organisation of chromosomes within the nucleus of somatic cells, suggesting that the 3D genome organisation in the germline may be different in bovids.

Combining these genomic results suggest that Robertsonian fusions may be currently shaping speciation within the waterbuck, as well as having historically shaped evolution across the family Bovidae. Centromeres have been proposed to be involved in meiotic drive, the non-Mendelian inheritance of genetic material, by the modification of satellite DNA causing enhanced interactions with the meiotic spindle in females, favouring segregation to the egg rather than the polar body in heterozygotes (Searle & Pardo-Manuel de Villena, 2024). This may promote the fixation of Rb fusions leading to reproductive isolation. This, on top of the high genomic differentiation and low recombination around fused centromeres containing genes involved in fertility, may play a role in speciation. It has recently been proposed that "speciation genes", the suppression of recombination, and ancient genetic variation could be important in the speciation process (Schluter & Rieseberg, 2022). These results will have significance across antelopes, where several species have polymorphic Rb fusions.

5.2. Applications of genomics to the conservation of antelopes

Developing tools and approaches to support the conservation of wildlife has never been more important, as species are now facing a sixth mass extinction event, with species loss in the last century a hundred times higher than the background state (Ceballos et al., 2015). This loss of biodiversity, caused by human activity, includes destruction of habitat through land use change for agriculture, forestry, and urban development, pollution, the introduction of invasive species, and anthropogenic climate change (described in Des Roches et al., 2021). Intraspecific diversity is important for both ecological function and also nature's contributions to people, through habitat creation, pollination, climate control, water quality, pest and pathogen control, energy, food, materials, and medicine (Des Roches et al., 2021). This diversity includes variation in life history, chemistry, and morphology, as well as genetics.

Often conservationists aim to preserve species and populations, however a greater number of studies are showing the importance of conserving wildlife at the genetic and genomic level too. Species and population declines can reduce genetic diversity and increase inbreeding, which impacts fitness and the ability for adaptation to changing

environmental conditions that may be caused by humans, such as invasive species or climate change (Ceballos et al., 2017). Genomics can help to resolve species and population delineation, population structure, gene flow and connectivity, hybridisation, and historical population size (Hohenlohe et al., 2021; Kardos et al., 2021).

Our study provides an example of a species that does not fit the traditional species concept, composed of two subspecies undergoing early speciation but with ongoing admixture. Conservation of a species such as this one requires careful management. In order to maintain the genomic diversity between and within subspecies, supported by clustering on the PCA and F_{ST}, the four main populations groups (common N, common S, Defassa N, Defassa S) should be managed separately. Care should also be taken with waterbuck surrounding the hybrid zone between the subspecies, as we found several populations with high proportions of admixture. Whilst our study was unable to karyotype individuals, the polymorphic chromosome fusions should also be considered, as previously proposed (S. Kingswood et al., 1998), as we find regions around these centromeres with high genomic differentiation, that may be involved in speciation. Chromosome rearrangements and structural variants are rarely considered in conservation, but can affect genomic variation, adaptation, and population fitness (Wold et al., 2021) and should also be adopted more widely. The use of Hi-C data and chromosome interactions, as shown in our study, could be a useful method to detect polymorphic Rb fusions in antelopes. Reductions in sequencing costs and improved preservation would improve its use in wild and captive samples. Long-read sequencing, such as ONT, will help with other SVs.

Further work is also needed to understand the impact of hybridisation on the conservation of antelopes, as we found several populations near the hybrid zone in waterbuck with varying degrees of admixture. A study in rainbow fish found that natural hybrid population and adaptive introgression had reduced the vulnerability of the population to projected future climate change, and stated that hybrid population should be considered in conservation management (Brauer et al., 2023). With changes in climate and habitat, suitable habitats may change. Studies are now beginning to predict species responses to these changes and the levels of adaptive potential (Hohenlohe et al., 2021). Using a combination ecological, environmental, and genomic datasets, populations can be assessed for genomic vulnerability to changing environmental and climatic conditions (e.g., the life on the edge toolbox; Barratt et al., 2023).

Antelopes have undergone changes in their distribution during the Pleistocene due to changes in savannah and forest cover across Africa, caused by periods of warming and cooling (E. D. Lorenzen et al., 2012). However, threatened antelopes with currently small geographic ranges due to anthropogenic factors will be disproportionately affected by future climate change (Payne & Bro-Jørgensen, 2016). This study found that 59/72 antelopes studied would decline by 2080, and 19/59 species would decline by over 50%, based on projected climate change scenarios. Protected areas could be impacted if species shift their range in response to the changing climate (Rannow et al., 2014). Increasing protected areas that will be most beneficial in future climates and improving connectivity between them should be a high priority for antelope conservation (Payne & Bro-Jørgensen, 2016). Waterbuck could also be impacted by climate change, as they have a high dependency on water (Taylor et al., 1969), with the potential for a reduction in standing water in a warmer climate. This may impact the species future distribution and population structure.

Several conservation programs are translocating captive antelopes back into the wild with the use of genomics. This includes the oryx, which was once extinct in the wild, but now has several wild populations in North Africa, and is supported by a chromosome-level genome and WGS data (Humble et al., 2020, 2023). Our study worked with conservationists at the Aspinall Foundation (Kent, UK), who have several antelopes in captivity, with the potential for reintroductions in the future. Genomic studies of these captive individuals are needed to accurately reintroduce the animals into the most suitable regions. Our study on waterbuck provides a detailed plan of population structure in this species which could be used for waterbuck translocations, and this should be expanded to other species before translocations.

Conservation genomics requires financial and human resources that may be limited in low- and middle-income countries, and therefore, there is need to build capacity in these countries for laboratory, sequencing, and computing infrastructure, as well as training (Bertola et al., 2024). Genomic data and publications should also be made accessible for researchers across the globe for conservation. Collaborative groups have also created standardised metrics for assessing genetic variation, such as the Essential Biodiversity Variables (EBVs; Hoban et al., 2022), which include genetic diversity, genetic differentiation, inbreeding, and effective population size. These measures may increase the uptake of genomics by conservation managers.

5.3. Future work

With advances in genomics, further work could include sequencing the waterbuck genome from telomere to telomere, as proposed by the Ruminant T2T Consortium (Kalbfleisch et al., 2024). Sequencing several karyotypes with long-reads (ONT or PacBio HiFi) would allow the study of the short p-arms of acrocentric chromosomes and the centromeres involved in Rb fusions, uncovering the mechanisms behind these Rb fusions in bovids. High coverage Hi-C data of all karyotypes would further support this effort.

Our study prioritised higher sampling over sequencing coverage, to enable the assessment of population structure. Further studies could utilise our chromosome-level reference and sequence several individuals at high coverage. This would allow the study of ROH across chromosomes, another indicator of genomic diversity, and demographic histories using effective population size, enabling both historic and recent population histories to be studied. It would also provide a reference for the mapping of long-read data, such as ONT or PacBio, to confirm the putative SVs we found between the two subspecies of waterbuck. The impact of these SVs on genes could then be investigated.

Lastly, with the sequencing of more chromosome-level genomes, further work could explore chromosome evolution across a greater number of species in the family Bovidae and reconstruct further nodes within this clade. This genomic data would also support efforts in resolving taxonomic uncertainties between antelopes. Moreover, using Hi-C data from the germline may find a correlation between Rb fusions and 3D genome organisation, supporting the Integrative Breakage Model of chromosome evolution.

References

Adelson, D. L., Raison, J. M., & Edgar, R. C. (2009). Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proceedings of the National Academy of Sciences*, *106*(31), 12855–12860. https://doi.org/10.1073/pnas.0901282106

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Álvarez-González, L., Arias-Sardá, C., Montes-Espuña, L., Marín-Gual, L., Vara, C., Lister, N. C., Cuartero, Y., Garcia, F., Deakin, J., Renfree, M. B., Robinson, T. J., Martí-Renom, M. A., Waters, P. D., Farré, M., & Ruiz-Herrera, A. (2022). Principles of 3D chromosome folding and evolutionary genome reshuffling in mammals. *Cell Reports*, *41*(12). https://doi.org/10.1016/j.celrep.2022.111839

Álvarez-González, L., Burden, F., Doddamani, D., Malinverni, R., Leach, E., Marín-García, C., Marín-Gual, L., Gubern, A., Vara, C., Paytuví-Gallart, A., Buschbeck, M., Ellis, P. J. I., Farré, M., & Ruiz-Herrera, A. (2022). 3D chromatin remodelling in the germ line modulates genome evolutionary plasticity. *Nature Communications*, *13*(1), 1–15. https://doi.org/10.1038/s41467-022-30296-6

Arias-Sardá, C., Quigley, S., & Farré, M. (2023). Patterns of chromosome evolution in ruminants. *Molecular Ecology*, *June*, 1–11. https://doi.org/10.1111/mec.17197

Barbosa, S., Mestre, F., White, T. A., Paupério, J., Alves, P. C., & Searle, J. B. (2018). Integrative approaches to guide conservation decisions: Using genomics to define conservation units and functional corridors. *Molecular Ecology*, *27*(17), 3452–3465. https://doi.org/10.1111/mec.14806

Barratt, C. D., Onstein, R. E., Pinsky, M. L., Steinfartz, S., Kühl, H. S., Forester, B. R., & Razgour, O. (2023). Life on the edge: a new toolbox for population-level climate change vulnerability assessments. *BioRxiv*, *2024*(June 2023), 2023.06.23.543988. https://doi.org/10.1111/2041-210X.14429

Bell, D. A., Robinson, Z. L., Funk, W. C., Fitzpatrick, S. W., Allendorf, F. W., Tallmon, D.

A., & Whiteley, A. R. (2019). The Exciting Potential and Remaining Uncertainties of Genetic Rescue. *Trends in Ecology and Evolution*, *34*(12), 1070–1079. https://doi.org/10.1016/j.tree.2019.06.006

Bertola, L. D., Brüniche-Olsen, A., Kershaw, F., Russo, I. R. M., MacDonald, A. J., Sunnucks, P., Bruford, M. W., Cadena, C. D., Ewart, K. M., de Bruyn, M., Eldridge, M. D. B., Frankham, R., Guayasamin, J. M., Grueber, C. E., Hoareau, T. B., Hoban, S., Hohenlohe, P. A., Hunter, M. E., Kotze, A., … Segelbacher, G. (2024). A pragmatic approach for integrating molecular tools into biodiversity conservation. *Conservation Science and Practice*, *6*(1), 1–15. https://doi.org/10.1111/csp2.13053

Blaxter, M., Mieszkowska, N., Di Palma, F., Holland, P., Durbin, R., Richards, T., Berriman, M., Kersey, P., Hollingsworth, P., Wilson, W., Twyford, A., Gaya, E., Lawniczak, M., Lewis, O., Broad, G., Howe, K., Hart, M., Flicek, P., & Barnes, I. (2022). Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences*, *119*(4), 1–7. https://doi.org/10.1073/pnas.2115642118

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements 06 Biological Sciences 0604 Genetics. *Genome Biology*, *19*(1), 1–12. https://doi.org/10.1186/s13059-018-1577-z

Branco, M. R., & Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biology*, *4*(5), 780–788. https://doi.org/10.1371/journal.pbio.0040138

Brauer, C. J., Sandoval-Castillo, J., Gates, K., Hammer, M. P., Unmack, P. J., Bernatchez, L., & Beheregaray, L. B. (2023). Natural hybridization reduces vulnerability to climate change. *Nature Climate Change*, *13*(3), 282–289. https://doi.org/10.1038/s41558-022-01585-1

Bredeson, J. V., Mudd, A. B., Medina-Ruiz, S., Mitros, T., Smith, O. K., Miller, K. E., Lyons, J. B., Batra, S. S., Park, J., Berkoff, K. C., Plott, C., Grimwood, J., Schmutz, J., Aguirre-Figueroa, G., Khokha, M. K., Lane, M., Philipp, I., Laslo, M., Hanken, J., … Rokhsar, D. S. (2024). Conserved chromatin and repetitive patterns reveal slow genome

evolution in frogs. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-023-43012-9

Buckland, R. A., & Evans, H. J. (1978). Cytogenetic aspects of phylogeny in the Bovidae. *Cytogenetic and Genome Research*, *21*(1–2), 42–63. https://doi.org/10.1159/000130877

Calamari, Z. T. (2021). Total Evidence Phylogenetic Analysis Supports New Morphological Synapomorphies for Bovidae (Mammalia, Artiodactyla). *American Museum Novitates*, *2021*(3970), 1–40. https://doi.org/10.1206/3970.1

Carbone, L., Alan Harris, R., Gnerre, S., Veeramah, K. R., Lorente-Galdos, B., Huddleston, J., Meyer, T. J., Herrero, J., Roos, C., Aken, B., Anaclerio, F., Archidiacono, N., Baker, C., Barrell, D., Batzer, M. A., Beal, K., Blancher, A., Bohrson, C. L., Brameier, M., … Gibbs, R. A. (2014). Gibbon genome and the fast karyotype evolution of small apes. *Nature*, *513*(7517), 195–201. https://doi.org/10.1038/nature13679

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances*, *1*(5), 9–13. https://doi.org/10.1126/sciadv.1400253

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(30), E6089–E6096. https://doi.org/10.1073/pnas.1704949114

Cernohorska, H., Kubickova, S., Vahala, J., & Rubes, J. (2012). Molecular insights into X;BTA5 chromosome rearrangements in the tribe antilopini (Bovidae). *Cytogenetic and Genome Research*, *136*(3), 188–198. https://doi.org/10.1159/000336248

Cernohorska, Halina, Kubickova, S., Kopecna, O., Vozdova, M., Matthee, C. A., Robinson, T. J., & Rubes, J. (2015). Nanger, Eudorcas, Gazella, and Antilope form a well-supported chromosomal clade within Antilopini (Bovidae, Cetartiodactyla). *Chromosoma*, *124*(2), 235–247. https://doi.org/10.1007/s00412-014-0494-5

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit - interactive quality assessment of genome assemblies. *G3: Genes, Genomes,*

Genetics, *10*(4), 1361–1374. https://doi.org/10.1534/g3.119.400908

Charlesworth, B. (2009). Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195–205. https://doi.org/10.1038/nrg2526

Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., Bibi, F., Yang, Y., Wang, J., Nie, W., Su, W., Liu, G., Li, Q., Fu, W., Pan, X., Liu, C., Yang, J., Zhang, C., Yin, Y., … Wang, W. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science*, *364*(6446). https://doi.org/10.1126/science.aav6202

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Chng, S. C., Ho, L., Tian, J., & Reversade, B. (2013). ELABELA: A Hormone Essential for Heart Development Signals via the Apelin Receptor. *Developmental Cell*, *27*(6), 672–680. https://doi.org/10.1016/j.devcel.2013.11.002

Christmas, M. J., Wallberg, A., Bunikis, I., Olsson, A., Wallerman, O., & Webster, M. T. (2019). Chromosomal inversions associated with environmental adaptation in honeybees. *Molecular Ecology*, *28*(6), 1358–1374. https://doi.org/10.1111/mec.14944

Clauss, M., & Rössner, G. E. (2014). Old world ruminant morphophysiology, life history, and fossil record: Exploring key innovations of a diversification sequence. *Annales Zoologici Fennici*, *51*(1–2), 80–94. https://doi.org/10.5735/086.051.0210

Coimbra, R. T. F., Winter, S., Kumar, V., Koepfli, K. P., Gooley, R. M., Dobrynin, P., Fennessy, J., & Janke, A. (2021). Whole-genome analysis of giraffe supports four distinct species. *Current Biology*, *31*(13), 2929-2938.e5. https://doi.org/10.1016/j.cub.2021.04.033

Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987–991. https://doi.org/10.1038/nbt.2023

Consortium, T. U. (2023). *UniProt: the Universal Protein Knowledgebase in 2023 - Google Scholar*. *51*(November 2022), 523–531.

https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=UniProt%3A+the+Universal+Protein+Knowledgebase+in+2023&btnG=

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/gigascience/giab008

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics*, *34*(15), 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

De Lange, T. (2005). Shelterin: The protein complex that shapes and safeguards human telomeres. *Genes and Development*, *19*(18), 2100–2110. https://doi.org/10.1101/gad.1346005

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature Biotechnology*, *34*(5), 518–524. https://doi.org/10.1038/nbt.3423

Des Roches, S., Pendleton, L. H., Shapiro, B., & Palkovacs, E. P. (2021). Conserving intraspecific variation for nature's contributions to people. *Nature Ecology and Evolution*, *5*(5), 574–582. https://doi.org/10.1038/s41559-021-01403-5

Ding, H., Schertzer, M., Wu, X., Gertsenstein, M., Selig, S., Kammori, M., Pourvali, R., Poon, S., Vulto, I., Chavez, E., Tam, P. P. ., Nagy, A., & Lansdorp, P. M. (2004). Regulation of Murine Telomere Length by Rtel. *Cell*, *117*(7), 873–886. https://doi.org/10.1016/j.cell.2004.05.026

Dobigny, G., Britton-Davidian, J., & Robinson, T. J. (2017). Chromosomal polymorphism in mammals: an evolutionary perspective. *Biological Reviews*, *92*(1), 1–21. https://doi.org/10.1111/brv.12213

Dorant, Y., Cayuela, H., Wellband, K., Laporte, M., Rougemont, Q., Mérot, C., Normandeau, E., Rochette, R., & Bernatchez, L. (2020). Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species. *Molecular Ecology*, *29*(24), 4765–4782. https://doi.org/10.1111/mec.15565

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., & Aiden, E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95. https://doi.org/10.1126/science.aal3327

Dudchenko, O., Shamim, M. S., Batra, S. S., Durand, N. C., Musial, N. T., Mostofa, R., Pham, M., Hilaire, B. G. S., Yao, W., Stamenova, E., Hoeger, M., Nyquist, S. K., Korchina, V., Pletch, K., Flanagan, J. P., Tomaszewicz, A., McAloose, D., Estrada, C. P., Novak, B. J., … Nathaniel, T. (2018). The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *BioRxiv*, 254797. https://www.biorxiv.org/content/early/2018/01/28/254797%0Ahttps://www.biorxiv.org/content/early/2018/01/28/254797?utm_content=buffer421d5&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*, *3*(1), 95–98. https://doi.org/10.1016/j.cels.2016.07.002

Dussex, N., van der Valk, T., Morales, H. E., Wheat, C. W., Díez-del-Molino, D., von Seth, J., Foster, Y., Kutschera, V. E., Guschanski, K., Rhie, A., Phillippy, A. M., Korlach, J., Howe, K., Chow, W., Pelan, S., Mendes Damas, J. D., Lewin, H. A., Hastie, A. R., Formenti, G., … Dalén, L. (2021). Population genomics of the critically endangered kākāpō. *Cell Genomics*, *1*(1), 100002. https://doi.org/10.1016/j.xgen.2021.100002

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., DeWinter, A., Dixon, J., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138. https://doi.org/10.1126/science.1162986

Escudeiro, A., Adega, F., Robinson, T. J., Heslop-Harrison, J. S., & Chaves, R. (2021). Analysis of the Robertsonian (1;29) fusion in Bovinae reveals a common mechanism: insights into its clinical occurrence and chromosomal evolution. *Chromosome Research*, *29*(3–4), 301–312. https://doi.org/10.1007/s10577-021-09667-0

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Farré, M., Bosch, M., López-Giráldez, F., Ponsà, M., & Ruiz-Herrera, A. (2011). Assessing

the role of tandem repeats in shaping the genomic architecture of great apes. *PLoS ONE*, *6*(11). https://doi.org/10.1371/journal.pone.0027239

Farré, M., Kim, J., Proskuryakova, A. A., Zhang, Y., Kulemzina, A. I., Li, Q., Zhou, Y., Xiong, Y., Johnson, J. L., Perelman, P. L., Johnson, W. E., Warren, W. C., Kukekova, A. V., Zhang, G., O'Brien, S. J., Ryder, O. A., Graphodatsky, A. S., Ma, J., Lewin, H. A., & Larkin, D. M. (2019). Evolution of gene regulation in ruminants differs between evolutionary breakpoint regions and homologous synteny blocks. *Genome Research*, *29*(4), 576–589. https://doi.org/10.1101/gr.239863.118

Farré, M., Robinson, T. J., & Ruiz-Herrera, A. (2015). An Integrative Breakage Model of genome architecture, reshuffling and evolution: The Integrative Breakage Model of genome evolution, a novel multidisciplinary hypothesis for the study of genome plasticity. *BioEssays*, *37*(5), 479–488. https://doi.org/10.1002/bies.201400174

Ferreira, M. S., Thurman, T. J., Jones, M. R., Farelo, L., Kumar, A. V., Mortimer, S. M. E., Demboski, J. R., Mills, L. S., Alves, P. C., Melo-Ferreira, J., & Good, J. M. (2023). The evolution of white-tailed jackrabbit camouflage in response to past and future seasonal climates. *Science*, *379*(6638), 1238–1242. https://doi.org/10.1126/science.ade3984

Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). NgsLD: Evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*, *35*(19), 3855–3856. https://doi.org/10.1093/bioinformatics/btz200

Gallagher, D. S., Davis, S. K., De Donato, M., Burzlaff, J. D., Womack, J. E., Taylor, J. F., & Kumamoto, A. T. (1999). A molecular cytogenetic analysis of the tribe Bovini (Artiodactyla: Bovidae: Bovinae) with an emphasis on sex chromosome morphology and NOR distribution. *Chromosome Research*, *7*(6), 481–492. https://doi.org/10.1023/A:1009254014526

Gallagher, D. S., & Womack, J. E. (1992). Chromosome Conservation in the Bovidae. *Journal of Heredity*, *83*(4), 287–298. https://doi.org/10.1093/oxfordjournals.jhered.a111215

Gamazon, E. R., & Stranger, B. E. (2015). The impact of human copy number variation on gene expression. *Briefings in Functional Genomics*, *14*(5), 352–357. https://doi.org/10.1093/bfgp/elv017

Gerton, J. L. (2024). A working model for the formation of Robertsonian chromosomes. *Journal of Cell Science*, *137*(7). https://doi.org/10.1242/jcs.261912

Gibbs, E. M., Feldman, E. L., & Dowling, J. J. (2010). The role of MTMR14 in autophagy and in muscle disease. *Autophagy*, *6*(6), 819–820. https://doi.org/10.4161/auto.6.6.12624

Giner-Delgado, C., Villatoro, S., Lerga-Jaso, J., Gayà-Vidal, M., Oliva, M., Castellano, D., Pantano, L., Bitarello, B. D., Izquierdo, D., Noguera, I., Olalde, I., Delprat, A., Blancher, A., Lalueza-Fox, C., Esko, T., O'Reilly, P. F., Andrés, A. M., Ferretti, L., Puig, M., & Cáceres, M. (2019). Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-12173-x

Gomes de Lima, L., Guarracino, A., Koren, S., Potapova, T., McKinney, S., Rhie, A., Solar, S. J., Seidel, C., Fagen, B., Walenz, B. P., Bouffard, G. G., Brooks, S. Y., Peterson, M., Hall, K., Crawford, J., Young, A. C., Pickett, B. D., Garrison, E., Phillippy, A. M., & Gerton, J. L. (2024). The formation and propagation of human Robertsonian chromosomes. In *Tjyybjb.Ac.Cn* (Vol. 27, Issue 2, pp. 635–637). https://doi.org/10.1101/2024.09.24.614821

Guarracino, A., Buonaiuto, S., de Lima, L. G., Potapova, T., Rhie, A., Koren, S., Rubinstein, B., Fischer, C., Abel, H. J., Antonacci-Fulton, L. L., Asri, M., Baid, G., Baker, C. A., Belyaeva, A., Billis, K., Bourque, G., Carroll, A., Chaisson, M. J. P., Chang, P.-C., … Garrison, E. (2023). Recombination between heterologous human acrocentric chromosomes. *Nature*, *617*(7960), 335–343. https://doi.org/10.1038/s41586-023-05976-y

Guo, J., Xu, N., Li, Z., Zhang, S., Wu, J., Dae, H. K., Mong, S. M., Meng, Q., Cao, H., Li, X., Shi, S., Yu, L., Kalachikov, S., Russo, J. J., Turro, N. J., & Ju, J. (2008). Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(27), 9145–9150. https://doi.org/10.1073/pnas.0804023105

Harewood, L., & Fraser, P. (2014). The impact of chromosomal rearrangements on regulation of gene expression. *Human Molecular Genetics*, *23*(R1), 76–82. https://doi.org/10.1093/hmg/ddu278

Harris, R. S. (The P. S. U. (2007). *Improved pairwise alignment of genomic DNA*.

Hassanin, A., & Douzery, E. J. P. (2003). Molecular and morphological phylogenies of Ruminantia and the alternative position of the Moschidae. *Systematic Biology*, *52*(2), 206–228. https://doi.org/10.1080/10635150390192726

Hernández Fernández, M., & Vrba, E. S. (2005). A complete estimate of the phylogenetic relationships in Ruminantia: A dated species-level supertree of the extant ruminants. *Biological Reviews of the Cambridge Philosophical Society*, *80*(2), 269–302. https://doi.org/10.1017/S1464793104006670

Hoban, S., Archer, F. I., Bertola, L. D., Bragg, J. G., Breed, M. F., Bruford, M. W., Coleman, M. A., Ekblom, R., Funk, W. C., Grueber, C. E., Hand, B. K., Jaffé, R., Jensen, E., Johnson, J. S., Kershaw, F., Liggins, L., MacDonald, A. J., Mergeay, J., Miller, J. M., … Hunter, M. E. (2022). Global genetic diversity status and trends: towards a suite of Essential Biodiversity Variables (EBVs) for genetic composition. *Biological Reviews*, *97*(4), 1511–1538. https://doi.org/10.1111/brv.12852

Hodgkinson, A., & Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, *12*(11), 756–766. https://doi.org/10.1038/nrg3098

Hohenlohe, P. A., Funk, W. C., & Rajora, O. P. (2021). Population genomics for wildlife conservation and management. *Molecular Ecology*, *30*(1), 62–82. https://doi.org/10.1111/mec.15720

Hon, T., Mars, K., Young, G., Tsai, Y. C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P., & Rank, D. R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, *7*(1), 1–11. https://doi.org/10.1038/s41597-020-00743-4

Howe, B., Umrigar, A., & Tsien, F. (2014). Chromosome preparation from cultured cells. *Journal of Visualized Experiments*, *83*, 3–7. https://doi.org/10.3791/50203

Humble, E., Dobrynin, P., Senn, H., Chuven, J., Scott, A. F., Mohr, D. W., Dudchenko, O., Omer, A. D., Colaric, Z., Lieberman Aiden, E., Al Dhaheri, S. S., Wildt, D., Oliaji, S., Tamazian, G., Pukazhenthi, B., Ogden, R., & Koepfli, K. (2020). Chromosomal-level genome assembly of the scimitar-horned oryx: Insights into diversity and

demography of a species extinct in the wild. *Molecular Ecology Resources*, *20*(6), 1668–1681. https://doi.org/10.1111/1755-0998.13181

Humble, E., Stoffel, M. A., Dicks, K., Ball, A. D., Gooley, R. M., Chuven, J., Pusey, R., Remeithi, M. Al, Koepfli, K.-P., Pukazhenthi, B., Senn, H., & Ogden, R. (2023). Conservation management strategy impacts inbreeding and mutation load in scimitar-horned oryx. *Proceedings of the National Academy of Sciences*, *120*(18), 2017. https://doi.org/10.1073/pnas.2210756120

Ivancevic, A. M., Kortschak, R. D., Bertozzi, T., & Adelson, D. L. (2018). Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome Biology*, *19*(1), 1–13. https://doi.org/10.1186/s13059-018-1456-7

Jain, C., Koren, S., Dilthey, A., Phillippy, A. M., & Aluru, S. (2018). A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics*, *34*(17), i748–i756. https://doi.org/10.1093/bioinformatics/bty597

Jerkovic, I., & Cavalli, G. (2021). Understanding 3D genome organization by multidisciplinary methods. *Nature Reviews Molecular Cell Biology*, *22*(8), 511–528. https://doi.org/10.1038/s41580-021-00362-w

Jobard, F., Lefèvre, C., Karaduman, A., Blanchet-Bardon, C., Emre, S., Weissenbach, J., Özgüc, M., Lathrop, M., Prud'Homme, J. F., & Fischer, J. (2002). Lipoxygenase-3 (ALOXE3) and 12(R)-lipoxygenase (ALOX12B) are mutated in non-bullous congenital ichthyosiform erythroderma (NCIE) linked to chromosome 17p13.1. *Human Molecular Genetics*, *11*(1), 107–113. https://doi.org/10.1093/hmg/11.1.107

Johnson, W. E., Onorato, D. P., Roelke, M. E., Land, E. D., Cunningham, M., Belden, R. C., McBride, R., Jansen, D., Lotz, M., Shindle, D., Howard, J., Wildt, D. E., Penfold, L. M., Hostetler, J. A., Oli, M. K., & O'Brien, S. J. (2010). Genetic Restoration of the Florida Panther. *Science*, *329*(5999), 1641–1645. https://doi.org/10.1126/science.1192891

Jones, M. R., Mills, L. S., Alves, P. C., Callahan, C. M., Alves, J. M., Lafferty, D. J. R., Jiggins, F. M., Jensen, J. D., Melo-Ferreira, J., & Good, J. M. (2018). Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science*, *360*(6395), 1355–1358. https://doi.org/10.1126/science.aar5273

Jones, M. R., Mills, L. S., Jensen, J. D., & Good, J. M. (2020). The Origin and Spread of Locally Adaptive Seasonal Camouflage in Snowshoe Hares. *The American Naturalist*, *196*(3), 316–332. https://doi.org/10.1086/710022

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., & Orlando, L. (2013). MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, *29*(13), 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Jouhilahti, E. M., Madissoon, E., Vesterlund, L., Töhönen, V., Krjutškov, K., Reyes, A. P., Petropoulos, S., Månsson, R., Linnarsson, S., Buörglin, T., Lanner, F., Hovatta, O., Katayama, S., & Kere, U. (2016). The human PRD-like homeobox gene LEUTX has a central role in embryo genome activation. *Development (Cambridge)*, *143*(19), 3459–3469. https://doi.org/10.1242/dev.134510

Ju, J., Kim, D. H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M. S., Shi, S., Wu, J., Edwards, J. R., Romu, A., & Turro, N. J. (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(52), 19635–19640. https://doi.org/10.1073/pnas.0609513103

Kalbfleisch, T. S., McKay, S. D., Murdoch, B. M., Adelson, D. L., Almansa-Villa, D., Becker, G., Beckett, L. M., Benítez-Galeano, M. J., Biase, F., Casey, T., Chuong, E., Clark, E., Clarke, S., Cockett, N., Couldrey, C., Davis, B. W., Elsik, C. G., Faraut, T., Gao, Y., … Rosen, B. D. (2024). The Ruminant Telomere-to-Telomere (RT2T) Consortium. *Nature Genetics*, *56*(8), 1566–1573. https://doi.org/10.1038/s41588-024-01835-2

Kardos, M., Armstrong, E. E., Fitzpatrick, S. W., Hauser, S., Hedrick, P. W., Miller, J. M., Tallmon, D. A., & Chris Funk, W. (2021). The crucial role of genome-wide genetic variation in conservation. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(48). https://doi.org/10.1073/pnas.2104642118

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O., & Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2203-5

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic*

Acids Research, *44*(9). https://doi.org/10.1093/nar/gkw092

Kelly, C. J., Chitko-McKown, C. G., & Chuong, E. B. (2022). Ruminant-specific retrotransposons shape regulatory evolution of bovine immunity. *Genome Research*, *32*(8), 1474–1486. https://doi.org/10.1101/gr.276241.121

Kim, J., Farré, M., Auvil, L., Capitanu, B., Larkin, D. M., Ma, J., & Lewin, H. A. (2017). Reconstruction and evolutionary history of eutherian chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(27), E5379–E5388. https://doi.org/10.1073/pnas.1702012114

Kimura, M. (1983). Rare variant alleles in the light of the neutral theory. *Molecular Biology and Evolution*, *1*(1), 84–93. https://doi.org/10.1093/oxfordjournals.molbev.a040305

Kingswood, S. C., Kumamoto, A. T., Charter, S. J., Houck, M. L., & Benirschke, A. (2000). Chromosomes of the antelope genus *Kobus* (Artiodactyla, Bovidae): karyotypic divergence by centric fusion rearrangements. *Cytogenetic and Genome Research*, *91*(1–4), 128–133. https://doi.org/10.1159/000056832

Kingswood, S., Kumamoto, A., Charter, S., Aman, R., & Ryder, O. (1998). Brief communication. Centric fusion polymorphisms in waterbuck (Kobus ellipsiprymnus). *Journal of Heredity*, *89*(1), 96–100. https://doi.org/10.1093/jhered/89.1.96

Kobus ellipsiprymnus: IUCN SSC Antelope Specialist Group. (2016). In *IUCN Red List of Threatened Species*. https://doi.org/10.2305/IUCN.UK.2016-2.RLTS.T11035A50189324.en

Koepfli, K. P., Tamazian, G., Wildt, D., Dobrynin, P., Kim, C., Frandsen, P. B., Godinho, R., Yurchenko, A. A., Komissarov, A., Krasheninnikova, K., Kliver, S., Kolchanova, S., Gonçalves, M., Carneiro, M., Pinto, P. V., Ferrand, N., Maldonado, J. E., Ferrie, G. M., Chemnick, L., … Pukazhenthi, B. S. (2019). Whole Genome Sequencing and Re-sequencing of the Sable Antelope (Hippotragus niger): A resource for monitoring diversity in ex situ and in situ populations. *G3: Genes, Genomes, Genetics*, *9*(6), 1785–1793. https://doi.org/10.1534/g3.119.400084

Kolmogorov, M., Armstrong, J., Raney, B. J., Streeter, I., Dunn, M., Yang, F., Odom, D., Flicek, P., Keane, T. M., Thybert, D., Paten, B., & Pham, S. (2018). Chromosome

assembly of large and complex genomes using multiple references. *Genome Research*, *28*(11), 1720–1732. https://doi.org/10.1101/gr.236273.118

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 1–13. https://doi.org/10.1186/s12859-014-0356-4

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

Kuhn, R. M., Haussler, D., & James Kent, W. (2013). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, *14*(2), 144–161. https://doi.org/10.1093/bib/bbs038

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., & Hedges, S. B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, *39*(8), 1–6. https://doi.org/10.1093/molbev/msac174

León-Ortiz, A. M., Panier, S., Sarek, G., Vannier, J.-B., Patel, H., Campbell, P. J., & Boulton, S. J. (2018). A Distinct Class of Genome Rearrangements Driven by Heterologous Recombination. *Molecular Cell*, *69*(2), 292-305.e6. https://doi.org/10.1016/j.molcel.2017.12.014

Letunic, I., & Bork, P. (2021). Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, *115*(17), 4325–4333. https://doi.org/10.1073/pnas.1720115115

Li, A., Yang, Q., Li, R., Dai, X., Cai, K., Lei, Y., Jia, K., Jiang, Y., & Zan, L. (2023). Chromosome-level genome assembly for takin (Budorcas taxicolor) provides insights into its taxonomic status and genetic diversity. *Molecular Ecology*, *32*(6),

1323–1334. https://doi.org/10.1111/mec.16483

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. *00*(00), 1–3. http://arxiv.org/abs/1303.3997

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369

Liu, X., Lin, L., Sinding, M.-H. S., Bertola, L. D., Hanghøj, K., Quinn, L., Garcia-Erill, G., Rasmussen, M. S., Schubert, M., Pečnerová, P., Balboa, R. F., Li, Z., Heaton, M. P., Smith, T. P. L., Pinto, R. R., Wang, X., Kuja, J., Brüniche-Olsen, A., Meisner, J., … Heller, R. (2024). Introgression and disruption of migration routes have shaped the genetic integrity of wildebeest populations. *Nature Communications*, *15*(1), 2921. https://doi.org/10.1038/s41467-024-47015-y

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Lorenzen, E. D., Heller, R., & Siegismund, H. R. (2012). Comparative phylogeography of African savannah ungulates. *Molecular Ecology*, *21*(15), 3656–3670. https://doi.org/10.1111/j.1365-294X.2012.05650.x

Lorenzen, Eline D., Simonsen, B. T., Kat, P. W., Arctander, P., & Siegismund, H. R. (2006). Hybridization between subspecies of waterbuck (Kobus ellipsiprymnus) in zones of overlap with limited introgression. *Molecular Ecology*, *15*(12), 3787–3799. https://doi.org/10.1111/j.1365-294X.2006.03059.x

Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, *30*(23), 5966–5993. https://doi.org/10.1111/mec.16077

Lucena-Perez, M., Paijmans, J. L. A., Nocete, F., Nadal, J., Detry, C., Dalén, L., Hofreiter, M., Barlow, A., & Godoy, J. A. (2024). Recent increase in species-wide diversity after interspecies introgression in the highly endangered Iberian lynx. *Nature Ecology and Evolution*, *8*(2), 282–292. https://doi.org/10.1038/s41559-023-02267-7

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics*, *4*(12), 981–994. https://doi.org/10.1038/nrg1226

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. https://doi.org/10.1186/2047-217X-1-18

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols*, *1*(12), 1–41. https://doi.org/10.1002/cpz1.323

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Martinez, P., Thanasoula, M., Carlos, A. R., Gómez-López, G., Tejera, A. M., Schoeftner, S., Dominguez, O., Pisano, D. G., Tarsounas, M., & Blasco, M. A. (2010). Mammalian Rap1 controls telomere function and gene expression through binding to telomeric and extratelomeric sites. *Nature Cell Biology*, *12*(8), 768–780. https://doi.org/10.1038/ncb2081

Matthee, C. A., & Robinson, T. J. (1999). CytochromebPhylogeny of the Family Bovidae: Resolution within the Alcelaphini, Antilopini, Neotragini, and Tragelaphini. *Molecular Phylogenetics and Evolution*, *12*(1), 31–46. https://doi.org/10.1006/mpev.1998.0573

McDonough, M. M., Parker, L. D., McInerney, N. R., Campana, M. G., & Maldonado, J. E.

(2018). Performance of commonly requested destructive museum samples for mammalian genomic studies. *Journal of Mammalogy*, *99*(4), 789–802. https://doi.org/10.1093/jmammal/gyy080

Medarde, N., López-Fuster, M. J., Mũoz-Mũoz, F., & Ventura, J. (2012). Spatio-temporal variation in the structure of a chromosomal polymorphism zone in the house mouse. *Heredity*, *109*(2), 78–89. https://doi.org/10.1038/hdy.2012.16

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, *210*(2), 719–731. https://doi.org/10.1534/genetics.118.301336

Mérot, C., Berdan, E. L., Cayuela, H., Djambazian, H., Ferchaud, A. L., Laporte, M., Normandeau, E., Ragoussis, J., Wellenreuther, M., & Bernatchez, L. (2021). Locally Adaptive Inversions Modulate Genetic Variation at Different Geographic Scales in a Seaweed Fly. *Molecular Biology and Evolution*, *38*(9), 3953–3971. https://doi.org/10.1093/molbev/msab143

Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends in Ecology and Evolution*, *35*(7), 561–572. https://doi.org/10.1016/j.tree.2020.03.002

Messina, G., Atterrato, M. T., Prozzillo, Y., Piacentini, L., Losada, A., & Dimitri, P. (2017). The human Cranio Facial Development Protein 1 (Cfdp1) gene encodes a protein required for the maintenance of higher-order chromatin organization. *Scientific Reports*, *7*(September 2016), 1–10. https://doi.org/10.1038/srep45022

Messina, G., Celauro, E., Atterrato, M. T., Giordano, E., Iwashita, S., & Dimitri, P. (2015). The Bucentaur (BCNT) protein family: a long-neglected class of essential proteins required for chromatin/chromosome organization and function. *Chromosoma*, *124*(2), 153–162. https://doi.org/10.1007/s00412-014-0503-8

Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., & Thomas, P. D. (2019). Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols*, *14*(3), 703–721.

https://doi.org/10.1038/s41596-019-0128-8

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, *34*(13), i142–i150. https://doi.org/10.1093/bioinformatics/bty266

Molbert, N., Ghanavi, H. R., Johansson, T., Mostadius, M., & Hansson, M. C. (2023). An evaluation of DNA extraction methods on historical and roadkill mammalian specimen. *Scientific Reports*, *13*(1), 1–9. https://doi.org/10.1038/s41598-023-39465-z

Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*, *21*(SUPPL. 2), 79–85. https://doi.org/10.1093/bioinformatics/bti1114

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443–451. https://doi.org/10.1038/nrg2986

Norris, M. L., Pauli, A., Gagnon, J. A., Lord, N. D., Rogers, K. W., Mosimann, C., Zon, L. I., & Schier, A. F. (2017). Toddler signaling regulates mesodermal cell migration downstream of Nodal signaling. *ELife*, *6*, 1–18. https://doi.org/10.7554/eLife.22626

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Ogden, R., Ghazali, M., Hopper, J., Čulík, L., & King, T. (2018). Genetic assessments for antelope reintroduction planning in four European breeding programmes. *Journal of Zoo and Aquarium Research*, *6*(3), 79–84.

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), 292–294. https://doi.org/10.1093/bioinformatics/btv566

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P. L. F., Fumagalli, M., Vilstrup, J. T., Raghavan, M., Korneliussen, T., Malaspinas, A. S., Vogt, J., Szklarczyk, D., Kelstrup, C. D., … Willerslev, E. (2013). Recalibrating equus evolution using the

genome sequence of an early Middle Pleistocene horse. *Nature*, *499*(7456), 74–78. https://doi.org/10.1038/nature12323

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., & Hofreiter, M. (2004). Genetic Analyses from Ancient DNA. *Annual Review of Genetics*, *38*(1), 645–679. https://doi.org/10.1146/annurev.genet.37.110801.143214

Pagacova, E., Cernohorska, H., Kubickova, S., Vahala, J., & Rubes, J. (2011). Centric fusion polymorphism in captive animals of family Bovidae. *Conservation Genetics*, *12*(1), 71–77. https://doi.org/10.1007/s10592-009-9991-0

Pauli, A., Norris, M. L., Valen, E., Chew, G.-L., Gagnon, J. A., Zimmerman, S., Mitchell, A., Ma, J., Dubrulle, J., Reyon, D., Tsai, S. Q., Joung, J. K., Saghatelian, A., & Schier, A. F. (2014). Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science*, *343*(6172). https://doi.org/10.1126/science.1248636

Payne, B. L., & Bro-Jørgensen, J. (2016). Disproportionate Climate-Induced Range Loss Forecast for the Most Threatened African Antelopes. *Current Biology*, *26*(9), 1200–1205. https://doi.org/10.1016/j.cub.2016.02.067

Pečnerová, P., Garcia-Erill, G., Liu, X., Nursyifa, C., Waples, R. K., Santander, C. G., Quinn, L., Frandsen, P., Meisner, J., Stæger, F. F., Rasmussen, M. S., Brüniche-Olsen, A., Hviid Friis Jørgensen, C., da Fonseca, R. R., Siegismund, H. R., Albrechtsen, A., Heller, R., Moltke, I., & Hanghøj, K. (2021). High genetic diversity and low differentiation reflect the ecological versatility of the African leopard. *Current Biology*, *31*(9), 1862-1871.e5. https://doi.org/10.1016/j.cub.2021.01.064

Peng, Q., Pevzner, P. A., & Tesler, G. (2006). The fragile breakage versus random breakage models of chromosome evolution. *PLoS Computational Biology*, *2*(2), 100–111. https://doi.org/10.1371/journal.pcbi.0020014

Petkova, D., Novembre, J., & Stephens, M. (2015). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, *48*(1), 94–100. https://doi.org/10.1038/ng.3464

Pockrandt, C., Alzamel, M., Iliopoulos, C. S., & Reinert, K. (2020). GenMap: Ultra-fast computation of genome mappability. *Bioinformatics*, *36*(12), 3687–3692. https://doi.org/10.1093/bioinformatics/btaa222

Purgato, S., Belloni, E., Piras, F. M., Zoli, M., Badiale, C., Cerutti, F., Mazzagatti, A., Perini, G., Della Valle, G., Nergadze, S. G., Sullivan, K. F., Raimondi, E., Rocchi, M., & Giulotto, E. (2015). Centromere sliding on a mammalian chromosome. *Chromosoma*, *124*(2), 277–287. https://doi.org/10.1007/s00412-014-0493-6

Quigley, S., Damas, J., Larkin, D. M., & Farré, M. (2023). syntenyPlotteR: a user-friendly R package to visualize genome synteny, ideal for both experienced and novice bioinformaticians. *Bioinformatics Advances*, *November*, 2–4. https://doi.org/10.1093/bioadv/vbad161

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., Habermann, B., Akhtar, A., & Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications*, *9*(1), 189. https://doi.org/10.1038/s41467-017-02525-w

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-14998-3

Rannow, S., Macgregor, N. A., Albrecht, J., Crick, H. Q. P., Förster, M., Heiland, S., Janauer, G., Morecroft, M. D., Neubert, M., Sarbu, A., & Sienkiewicz, J. (2014). Managing Protected Areas Under Climate Change: Challenges and Priorities. *Environmental Management*, *54*(4), 732–743. https://doi.org/10.1007/s00267-014-0271-5

Raposinho, P. D., Pedrazzini, T., White, R. B., Palmiter, R. D., & Aubert, M. L. (2004). Chronic Neuropeptide Y Infusion into the Lateral Ventricle Induces Sustained Feeding and Obesity in Mice Lacking Either Npy1r or Npy5r Expression. *Endocrinology*, *145*(1), 304–310. https://doi.org/10.1210/en.2003-0914

Rasmussen, M. S., Garcia-Erill, G., Korneliussen, T. S., Wiuf, C., & Albrechtsen, A. (2022). Estimation of site frequency spectra from low-coverage sequencing data using stochastic EM reduces overfitting, runtime, and memory usage. *Genetics*, *222*(4). https://doi.org/10.1093/genetics/iyac148

Raxworthy, C. J., & Smith, B. T. (2021). Mining museums for historical DNA: advances and challenges in museomics. *Trends in Ecology and Evolution*, *36*(11), 1049–1060. https://doi.org/10.1016/j.tree.2021.07.009

Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J., Pohl, A., Pheasant, M., Meyer, L. R., Learned, K., Hsu, F., Hillman-Jackson, J., Harte, R. A., Giardine, B., Dreszer, T. R., … Kent, W. J. (2009). The UCSC genome browser database: Update 2010. *Nucleic Acids Research*, *38*(SUPPL.1), 613–619. https://doi.org/10.1093/nar/gkp939

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., Hook, P. W., Koren, S., Rautiainen, M., Alexandrov, I. A., Allen, J., Asri, M., Bzikadze, A. V., Chen, N. C., Chin, C. S., Diekhans, M., Flicek, P., Formenti, G., Fungtammasan, A., … Phillippy, A. M. (2023). The complete sequence of a human Y chromosome. *Nature*, *621*(7978), 344–354. https://doi.org/10.1038/s41586-023-06457-y

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, *21*(1), 1–27. https://doi.org/10.1186/s13059-020-02134-9

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology and Evolution*, *16*(7), 351–358. https://doi.org/10.1016/S0169-5347(01)02187-5

Roycroft, E., MacDonald, A. J., Moritz, C., Moussalli, A., Miguez, R. P., & Rowe, K. C. (2021). Museum genomics reveals the rapid decline and extinction of Australian rodents since European settlement. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(27). https://doi.org/10.1073/pnas.2021390118

Ruiz-Herrera, A., Farré, M., & Robinson, T. J. (2012). Molecular cytogenetic and genomic

insights into chromosomal evolution. *Heredity*, *108*(1), 28–36. https://doi.org/10.1038/hdy.2011.102

Sánchez-Guillén, R. A., Capilla, L., Reig-Viader, R., Martínez-Plana, M., Pardo-Camacho, C., Andrés-Nieto, M., Ventura, J., & Ruiz-Herrera, A. (2015). On the origin of Robertsonian fusions in nature: Evidence of telomere shortening in wild house mice. *Journal of Evolutionary Biology*, *28*(1), 241–249. https://doi.org/10.1111/jeb.12568

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Schluter, D., & Rieseberg, L. H. (2022). Three problems in the genetics of speciation by selection. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(30). https://doi.org/10.1073/pnas.2122153119

Schubert, I., & Lysak, M. A. (2011). Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends in Genetics*, *27*(6), 207–216. https://doi.org/10.1016/j.tig.2011.03.004

Schubert, M., Ermini, L., Sarkissian, C. Der, Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M. D., Fernández, R., Kircher, M., McCue, M., Willerslev, E., & Orlando, L. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols*, *9*(5), 1056–1082. https://doi.org/10.1038/nprot.2014.063

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*(1), 1–7. https://doi.org/10.1186/s13104-016-1900-2

Searle, J. B., & Pardo-Manuel de Villena, F. (2024). Meiotic Drive and Speciation. *Annual Review of Genetics*, *58*(1), 341–363. https://doi.org/10.1146/annurev-genet-111523-102603

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE*, *11*(10), 1–10. https://doi.org/10.1371/journal.pone.0163962

Slijepcevic, P. (1998). Telomeres and mechanisms of Robertsonian fusion.

Chromosoma, *107*(2), 136–140. https://doi.org/10.1007/s004120050289

Smith, L. B., Milne, L., Nelson, N., Eddie, S., Brown, P., Atanassova, N., O'Bryan, M. K., O'Donnell, L., Rhodes, D., Wells, S., Napper, D., Nolan, P., Lalanne, Z., Cheeseman, M., & Peters, J. (2012). KATNAL1 regulation of sertoli cell microtubule dynamics is essential for spermiogenesis and male fertility. *PLoS Genetics*, *8*(5). https://doi.org/10.1371/journal.pgen.1002697

Taylor, C., Spinage, C., & Lyman, C. (1969). Water relations of the waterbuck, an East African antelope. *American Journal of Physiology-Legacy Content*, *217*(2), 630–634. https://doi.org/10.1152/ajplegacy.1969.217.2.630

Thomas, P. D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L. P., & Mi, H. (2022). PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science*, *31*(1), 8–22. https://doi.org/10.1002/pro.4218

Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S., & Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, *38*(15). https://doi.org/10.1093/nar/gkq543

Trickett, A. J., & Butlin, R. K. (1994). Recombination suppressors and the evolution of new species. *Heredity*, *73*(4), 339–345. https://doi.org/10.1038/hdy.1994.180

Trojer, P., & Reinberg, D. (2007). Facultative Heterochromatin: Is There a Distinctive Molecular Signature? *Molecular Cell*, *28*(1), 1–13. https://doi.org/10.1016/j.molcel.2007.09.011

Uliano-Silva, M., Ferreira, J. G. R. N., Krasheninnikova, K., Blaxter, M., Mieszkowska, N., Hall, N., Holland, P., Durbin, R., Richards, T., Kersey, P., Hollingsworth, P., Wilson, W., Twyford, A., Gaya, E., Lawniczak, M., Lewis, O., Broad, G., Martin, F., Hart, M., … McCarthy, S. A. (2023). MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics*, *24*(1), 1–13. https://doi.org/10.1186/s12859-023-05385-y

van der Valk, T., Pečnerová, P., Díez-del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E., Fidan, F. R., Barnes, I., Liu, S., Somel, M., Heintzman, P. D., Nikolskiy, P., Shapiro, B., Skoglund, P., Hofreiter, M., … Dalén, L. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, *591*(7849), 265–269.

https://doi.org/10.1038/s41586-021-03224-9

Vannier, J. B., Sarek, G., & Boulton, S. J. (2014). RTEL1: Functions of a disease-associated helicase. *Trends in Cell Biology*, *24*(7), 416–425. https://doi.org/10.1016/j.tcb.2014.01.004

Vara, C., Paytuví-Gallart, A., Cuartero, Y., Álvarez-González, L., Marín-Gual, L., Garcia, F., Florit-Sabater, B., Capilla, L., Sanchéz-Guillén, R. A., Sarrate, Z., Aiese Cigliano, R., Sanseverino, W., Searle, J. B., Ventura, J., Marti-Renom, M. A., Le Dily, F., & Ruiz-Herrera, A. (2021). The impact of chromosomal fusions on 3D genome folding and recombination in the germ line. *Nature Communications*, *12*(1), 2981. https://doi.org/10.1038/s41467-021-23270-1

Vassart, M., Séguéla, A., & Hayes, H. (1995). Chromosomal Evolution in Gazelles. *Journal of Heredity*, *86*(3), 216–227. https://doi.org/10.1093/oxfordjournals.jhered.a111565

von Seth, J., Dussex, N., Díez-del-Molino, D., van der Valk, T., Kutschera, V. E., Kierczak, M., Steiner, C. C., Liu, S., Gilbert, M. T. P., Sinding, M. H. S., Prost, S., Guschanski, K., Nathan, S. K. S. S., Brace, S., Chan, Y. L., Wheat, C. W., Skoglund, P., Ryder, O. A., Goossens, B., … Dalén, L. (2021). Genomic insights into the conservation status of the world's last remaining Sumatran rhinoceros populations. *Nature Communications*, *12*(1), 1–11. https://doi.org/10.1038/s41467-021-22386-8

Wallace, C. (1979). Some trends in the evolution of the chromosomes in the Bovidae. *Koedoe*, *22*(1). https://doi.org/10.4102/koedoe.v22i1.654

Wang, X., Pedersen, C. T., Athanasiadis, G., Garcia-Erill, G., Hanghøj, K., Bertola, L. D., Rasmussen, M. S., Schubert, M., Liu, X., Li, Z., Lin, L., Balboa, R. F., Jørsboe, E., Nursyifa, C., Liu, S., Muwanika, V., Masembe, C., Chen, L., Wang, W., … Heller, R. (2024). Persistent Gene Flow Suggests an Absence of Reproductive Isolation in an African Antelope Speciation Model. *Systematic Biology*, *Xx*, 1–23. https://doi.org/10.1093/sysbio/syae037

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, *39*(11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Wang, Z., & Storm, D. R. (2006). Extraction of DNA from mouse tails. *BioTechniques*,

41(4), 410–412. https://doi.org/10.2144/000112255

Wen, N., Yu, M. F., Liu, J., Cai, C., Liu, Q. H., & Shen, J. (2018). Deficiency of MTMR14 impairs male fertility in Mus musculus. *PLoS ONE*, *13*(11). https://doi.org/10.1371/journal.pone.0206224

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

White, M. J. D. (1978). Chain processes in chromosomal speciation. *Systematic Zoology*, *27*(3), 285–298. https://doi.org/10.2307/2412880

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*, *31*(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Wold, J., Koepfli, K. P., Galla, S. J., Eccles, D., Hogg, C. J., Le Lec, M. F., Guhlin, J., Santure, A. W., & Steeves, T. E. (2021). Expanding the conservation genomics toolbox: Incorporating structural variants to enhance genomic studies for species of conservation concern. *Molecular Ecology*, *30*(23), 5949–5965. https://doi.org/10.1111/mec.16141

Woodfine, T., & Gilbert, T. (2016). The Fall and Rise of the Scimitar-Horned Oryx. In *Antelope Conservation* (pp. 280–296). Wiley. https://doi.org/10.1002/9781118409572.ch14

Wright, C. J., Stevens, L., Mackintosh, A., Lawniczak, M., & Blaxter, M. (2024). Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera. *Nature Ecology and Evolution*, *8*(4), 777–790. https://doi.org/10.1038/s41559-024-02329-4

Wu, B., Chen, X., Yu, M., Ren, J., Hu, J., Shao, C., Zhou, L., Sun, X., Yu, T., Zheng, Y., Wang, Y., Wang, Z., Zhang, H., Fan, G., & Liu, Z. (2022). Chromosome-level genome and population genomic analysis provide insights into the evolution and environmental

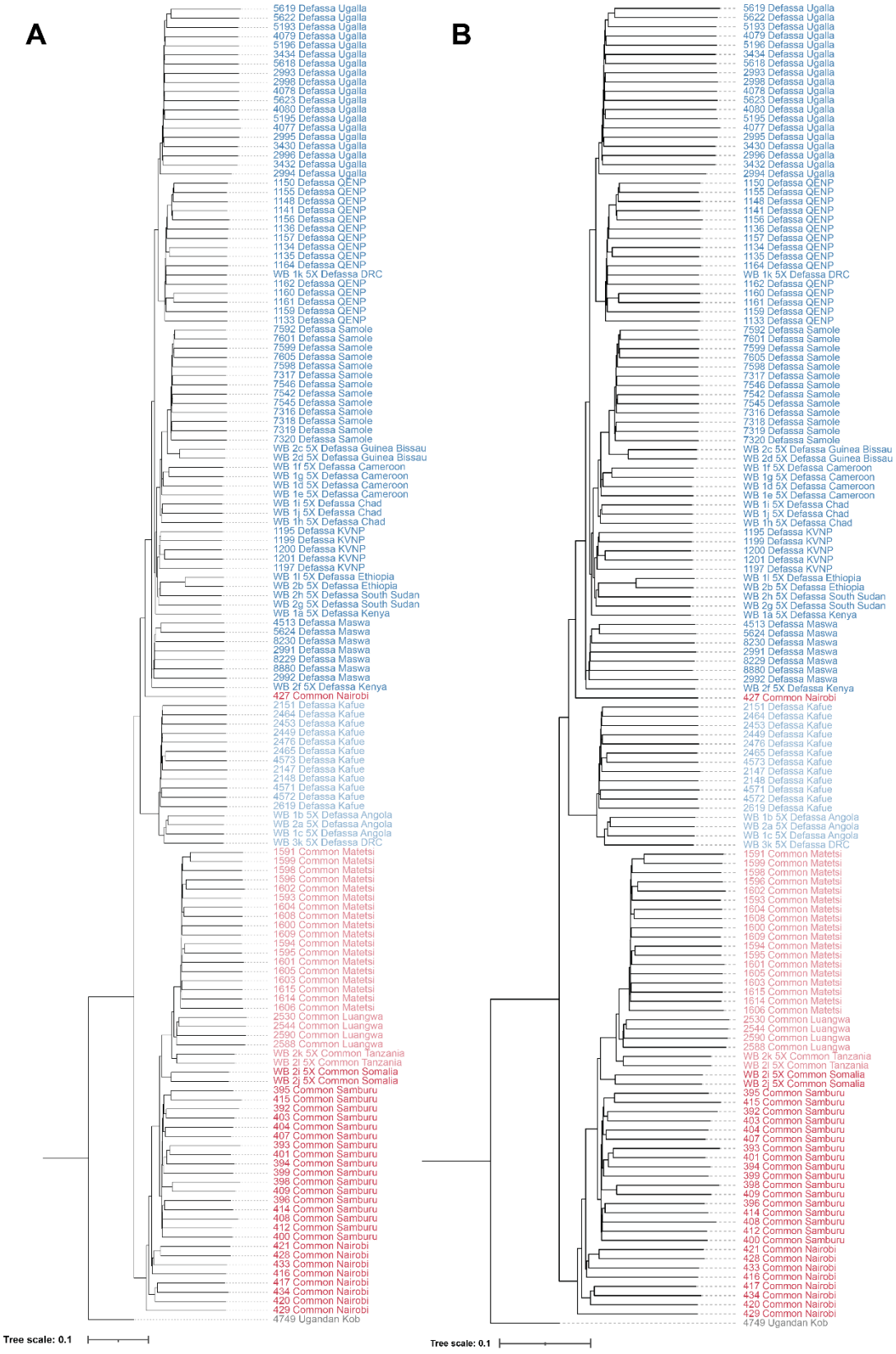
adaptation of Jinjiang oyster Crassostrea ariakensis. *Molecular Ecology Resources*, *22*(4), 1529–1544. https://doi.org/10.1111/1755-0998.13556

Yin, Y., Fan, H., Zhou, B., Hu, Y., Fan, G., Wang, J., Zhou, F., Nie, W., Zhang, C., Liu, L., Zhong, Z., Zhu, W., Liu, G., Lin, Z., Liu, C., Zhou, J., Huang, G., Li, Z., Yu, J., … Wei, F. (2021). Molecular mechanisms and topological consequences of drastic chromosomal rearrangements of muntjac deer. *Nature Communications*, *12*(1), 6858. https://doi.org/10.1038/s41467-021-27091-0

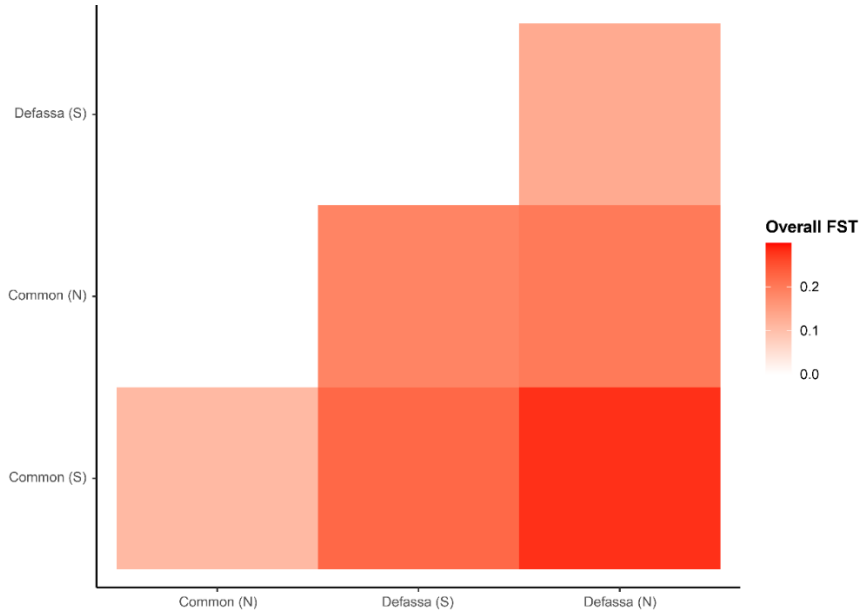
Zhang, Y., Liu, H., Li, W., Zhang, Z., Zhang, S., Teves, M. E., Stevens, C., Foster, J. A., Campbell, G. E., Windle, J. J., Hess, R. A., Pazour, G. J., & Zhang, Z. (2018). Intraflagellar transporter protein 140 (IFT140), a component of IFT-A complex, is essential for male fertility and spermiogenesis in mice. *Cytoskeleton*, *75*(2), 70–84. https://doi.org/10.1002/cm.21427

Zheng, H., & Xie, W. (2019). The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology*, *20*(9), 535–550. https://doi.org/10.1038/s41580-019-0132-4

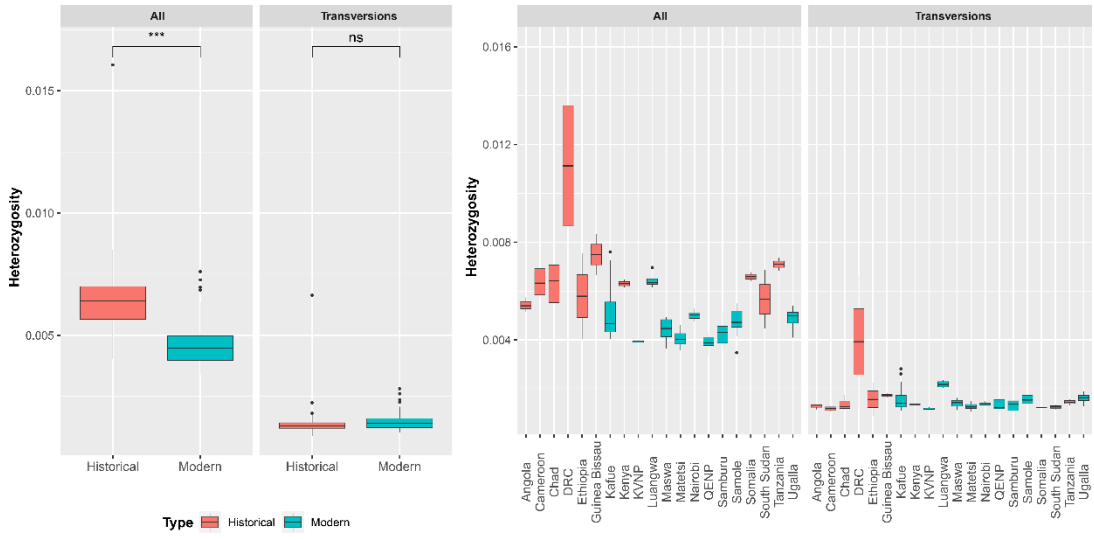
Supplementary Materials



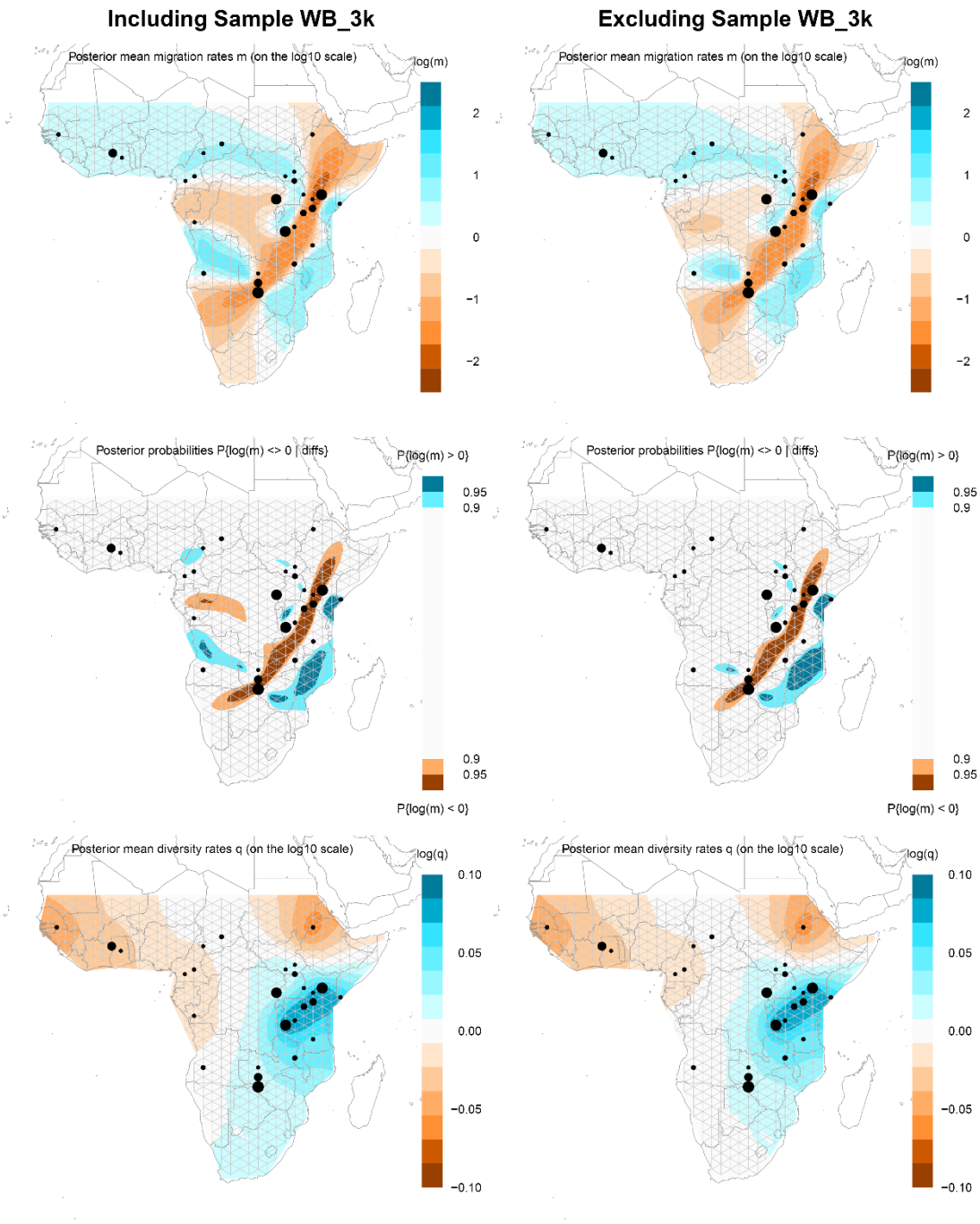
Supplementary Figure 1: *Neighbour joining tree of historical and modern samples using non-repetitive (A) and repetitive (B) filtered genomic sites. Defassa subspecies labelled in blue and common in red, with north in a darker shade and south lighter.*



Supplementary Figure 2: *Pairwise genomic differentiation (F_{ST}) between waterbuck groups for the filtered repetitive sites.*



Supplementary Figure 3: *Heterozygosity of all repetitive filtered sites and repetitive transversion sites, separated by historical or modern (left) and by population (right).*

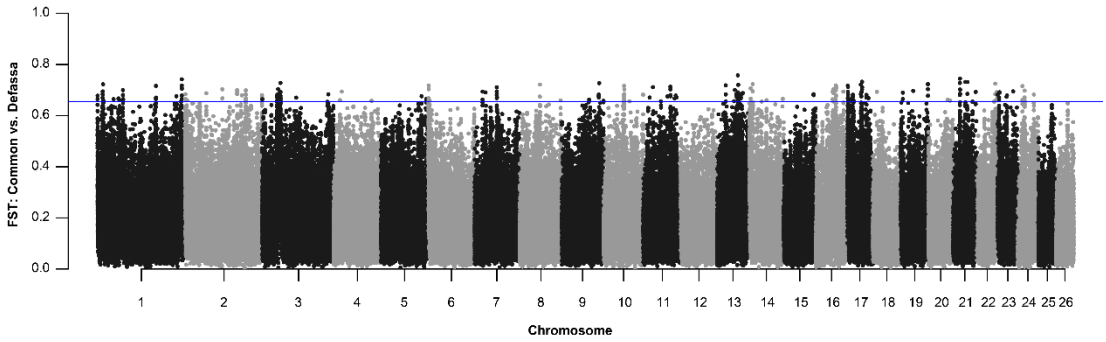


Supplementary Figure 4: *Estimating Effective Migration Surfaces (EEMS) analysis of filtered repetitive sites. The analysis was run with all historical and modern samples (143 individuals) and additionally with sample WB_3k removed (142 individuals).*

Supplementary Table 1: *List of genes found in the top 0.1% of genomic differentiation (F_{ST}) windows between common and defassa waterbuck for non-repetitive sites. Waterbuck chromosome (KEL), gene, and the highest F_{ST} window containing the gene.*

KEL	Gene	F_{ST}
21	MTMR14	0.770346
13	RTEL1	0.7599
13	STMN2	0.7599
13	STMN3	0.7599
24	DTNBP1	0.756076
16	SSU72	0.752172
16	TMEM240	0.752172
13	STMN1	0.741315
1	DIP2A	0.741171
17	DDX19A	0.738583
16	FNDC10	0.729349
18	CCNB1	0.728888
18	LOC102187875	0.728888
16	GNB1	0.727515
16	ATAD3A	0.726425
16	VWA1	0.72153
17	PDPR	0.721367
21	CADPS	0.721108
3	IFT140	0.72047
3	TMEM204	0.72047
4	ADAR	0.718694
17	DDX19B	0.718567
17	SLC8A2	0.718545
17	SLC8A3	0.718545
1	MTMR3	0.714602
1	MTMR4	0.714602
1	MORC3	0.714439
22	KDM2B	0.711907
16	NADK	0.711469
2	IPO8	0.708347
2	CAPRIN2	0.7078
2	LOC101906914	0.7078
3	LOC102183842	0.704764
17	SLC7A11	0.704524
17	SLC7A6	0.704524
14	DIAPH3	0.70448
19	FBXO22	0.701635
25	LOC100847151	0.701474
5	KMT2C	0.701
13	DNAJC5	0.70096
1	ITSN1	0.699898
3	ARHGEF4	0.699222
16	CDC42BPB	0.697316
3	LOC102186637	0.696921
17	PRMT7	0.696802
17	LEUTX	0.696428
16	SPEN	0.695365
17	AAGAB	0.693203
21	TADA3	0.692632
10	LOC108634660	0.689308
10	ZNF280D	0.689308
1	HSF5	0.689134
12	SCAF8	0.688961
1	ALOX12B	0.688919
1	ALOX15B	0.688919
1	ALOXE3	0.688919
13	FYN	0.687884
13	LCK	0.687884
13	YES1	0.687884

13	GMEB2	0.68767
16	DDI1	0.685997
16	DDI2	0.685997
22	PPP1CA	0.685716
22	PPP1CB	0.685716
22	RAD9A	0.685716
22	TBC1D10C	0.685716
13	ZBTB46	0.685524
3	PUM2	0.684865
1	HELZ	0.683641
1	CD7	0.683112
11	ESRP1	0.68282
11	ESRP2	0.68282
24	CMTR1	0.682617
16	LDHAL6B	0.682312
17	TANGO6	0.682212
2	BTBD11	0.681833
1	IFNAR1	0.681729
24	XPO5	0.681637
4	FNDC7	0.681245
16	PEX14	0.680757
3	LOC108633841	0.679589
3	IMP4	0.678545
3	KCTD17	0.677004
3	KCTD5	0.677004
3	CLCN7	0.676495
17	SPG7	0.676201
4	OTUD7B	0.675842
1	WDR45B	0.675407
16	TNFRSF18	0.675263
16	TTLL10	0.675263
8	ZCCHC10	0.675209
13	PSMF1	0.674795
9	EHBP1	0.674347
13	DIP2C	0.674319
16	TNFRSF4	0.674043
17	LOC510185	0.673183
11	COPS5	0.67318
14	KATNA1	0.672783
14	KATNAL1	0.672783
1	BPTF	0.672305
13	RPRD1A	0.671501
13	RPRD1B	0.671501
4	UBAP2L	0.671084
9	RABGAP1	0.670991



Supplementary Figure 5: *Genomic differentiation (F_{ST}) calculated in 10 Kb windows between common and defassa waterbuck across the 26 autosomal chromosomes, using repetitive sites. Blue line represents the top 0.1% windows (with greater than or equal to 1000 genomic sites).*

Supplementary Table 2: *List of genes found in the top 0.1% of genomic differentiation (F_{ST}) windows between the common waterbuck populations in the north and south. given. Waterbuck chromosome (KEL), gene, and the highest F_{ST} window containing the gene.*

KEL	Gene	F_{ST}
1	ASGR1	0.63211
1	ASGR2	0.63211
1	LOC132346939	0.63211
9	ZC3H8	0.630625
9	GREB1	0.607997
12	RPF2	0.594099
9	LOC102169298	0.590545
17	LOC102168852	0.580579
5	CADPS	0.57967
9	ZC3H8	0.573766
3	SCN1A	0.56592
3	SCN2A	0.56592
3	SCN4A	0.56592
13	ANKRD26	0.56424
4	GABPB2	0.560084
1	MME	0.556354
8	FBXL17	0.555701
23	PAPSS1	0.551506
23	PAPSS2	0.551506
1	LOC132346640	0.545706
4	GNAT1	0.543276
4	GNAT2	0.543276
4	GNAZ	0.543276
5	SEMA3E	0.537937
8	LOC100298064	0.537167
1	FGF12	0.532819
23	PAPSS1	0.532796
23	PAPSS2	0.532796
1	VPS8	0.526515
1	SEC22A	0.5216
8	HMGXB3	0.517469
17	WDR88	0.514229
5	PCLO	0.513536
19	LOC102180540	0.511077
12	RPF2	0.510194
21	PBRM1	0.508012
25	THAP1	0.507769
2	RIOX2	0.506146
2	RXFP1	0.506146
13	ANKRD26	0.504352
6	SLAIN2	0.504285
25	LOC616190	0.503244
6	TAPT1	0.503178
11	WASHC5	0.502043
1	EFCAB5	0.500197
12	RPF2	0.494528
1	SEC22A	0.487632
5	ZMIZ1	0.48718
5	ZMIZ2	0.48718
24	TINAG	0.486307
1	PDE9A	0.485685
6	SLAIN2	0.483759
1	SLC37A1	0.482824
1	SLC37A2	0.482824
4	LOC102169353	0.480593
9	FNBP1	0.476948
9	FNBP1L	0.476948

9	TRIP10	0.476948
3	GALNT3	0.475764
18	ANKRD33B	0.475529
21	CNTN3	0.475037
8	OR2AV10	0.474656
5	ABCB1	0.473897
3	KATNIP	0.473521
4	GNAT1	0.472606
4	GNAT2	0.472606
4	GNAZ	0.472606
4	GNAI3	0.472252
4	GNAT1	0.472252
13	PCK1	0.471374

3	HAPSTR1	0.470626
4	LOC112445868	0.470236
14	BIVM	0.46967
7	TTC39A	0.469626
7	TTC39B	0.469626
1	POLR2A	0.469454
2	DCAF13	0.468321
26	PLA2G12A	0.464623
26	PLA2G12B	0.464623
2	OR10U5	0.464041
1	AURKA	0.46399
1	AURKB	0.46399
1	DNAH2	0.463239

Supplementary Table 3: *List of genes found in the top 0.1% of genomic differentiation (F_{ST}) windows between defassa waterbuck populations in the north and south. Waterbuck chromosome (KEL), gene, and the highest F_{ST} window containing the gene.*

KEL	Gene	F_{ST}
1	ALDH3A1	0.749557
1	ALDH3A2	0.749557
1	ALDH3B1	0.749557
9	LIMS1	0.716455
9	LIMS2	0.716455
7	SMU1	0.70143
9	CCDC138	0.691862
2	CNTN1	0.687419
2	CNTN6	0.687419
6	LOC102187580	0.67799
6	RNF212	0.67799
9	CCDC138	0.674153
1	MYADML2	0.671094
2	MYO18B	0.66923
4	LOC507055	0.665315
5	CPA1	0.663775
5	LOC102182474	0.663775
8	HSPA4	0.659223
8	HSPH1	0.659223
13	LOC514978	0.658928
11	MCMDC2	0.650175
17	SF3B3	0.649572
17	BOSTAUV1R416	0.64907
3	TUBGCP5	0.649047
11	MCMDC2	0.648948
6	LOC102187580	0.64482
17	PMFBP1	0.641222
1	NOTUM	0.638944
13	LOC514978	0.63683
13	LOC104975936	0.636745
4	LOC102182655	0.635342
1	ALDH3A1	0.627266
1	ALDH3A2	0.627266
1	ALDH3B1	0.627266
17	BOSTAUV1R416	0.625449
21	TFCP2L1	0.623221
4	IPP	0.619257
6	RNF212	0.618559
19	EVL	0.618284
9	CCDC138	0.616066
8	HSPA4	0.61587
8	HSPH1	0.61587
8	HSPA4	0.614868
8	HSPH1	0.614868
5	CPA1	0.613198
5	LOC102182474	0.613198
10	LOC102176127	0.611418
1	LOC782391	0.605904
17	BOSTAUV1R416	0.605063
6	LOC102181374	0.600693
6	C1QTNF7	0.600563
7	WNK2	0.600014
1	SLC47A2	0.598547
8	GPR108	0.597948
21	ENTPD3	0.597628
16	MARK1	0.597204
16	MARK2	0.597204
16	MARK3	0.597204
13	LOC514978	0.596722

4	EEIG1	0.595576
4	FAM102B	0.595576
6	DMP1	0.594684
4	MMADHC	0.594584
5	SVOPL	0.593015
22	LOC108635411	0.592376
21	DNAH1	0.591776
17	OSGIN1	0.591293
2	CNTN1	0.585583
2	CNTN6	0.585583
3	TUBGCP5	0.585465
2	VDR	0.584795
18	PANK1	0.583881
4	LOC132344952	0.581884
15	LOC101908134	0.581325
13	LOC102171118	0.576609
13	LOC104975936	0.576609
19	MCTP2	0.571437
3	CUX1	0.567984
4	LOC102182655	0.567446
4	ZYG11B	0.567446
8	MUC16	0.56628
2	LOC132345313	0.565374
3	CUX1	0.565126
4	LOC108634873	0.564764

4	EEIG1	0.563312
4	FAM102B	0.563312
2	PPFIBP1	0.563215
5	SVOPL	0.562857
2	TMBIM6	0.561196
5	PSMA2	0.56069
1	SLC47A2	0.560467
8	OBSCN	0.560292
4	SLC16A1	0.559135
16	MARK1	0.558358
16	MARK2	0.558358
16	MARK3	0.558358
2	CNTN1	0.557613
2	CNTN6	0.557613
6	UGT8	0.557178
2	CABIN1	0.556511
3	CUX1	0.555702
19	ARNT2	0.554927
9	EML5	0.55478
2	CABIN1	0.554669
7	DDX58	0.552054
8	HSPA4	0.551969
8	HSPH1	0.551969
16	CFH	0.551051

Supplementary Table 4: *Synteny between Anc1 and cattle chromosomes.*

Anc1 Chromosome	**Cattle Chromosome**
Anc1-1	BTA1
Anc1-2	BTA2
Anc1-3	BTAX
Anc1-4	BTA3
Anc1-5	BTA4
Anc1-6	BTA6
Anc1-7	BTA5
Anc1-8	BTA8
Anc1-9	BTA7
Anc1-10	BTA11
Anc1-11	BTA10
Anc1-12	BTA9;14
Anc1-13	BTA9
Anc1-14	BTA12
Anc1-15	BTA13
Anc1-16	BTA16
Anc1-17	BTA15
Anc1-18	BTA17
Anc1-19	BTA20
Anc1-20	BTA21
Anc1-21	BTA18
Anc1-22	BTA19
Anc1-23	BTA24
Anc1-24	BTA22
Anc1-25	BTA23
Anc1-26	BTA29
Anc1-27	BTA26
Anc1-28	BTA28
Anc1-29	BTA27
Anc1-30	BTA25

Supplementary Table 5: *Cytogenetic publications of species of the family Bovidae with synteny to cattle chromosomes. Chromosomes denoted with an * are polymorphic within the species and were not included in the analysis.*

Reference	Subfamily	Species	Fusions (Cattle Chromosomes)
Pagacova et al. 2011	Aepycerotinae	*Aepyceros melampus*	14;20*
Steiner et al. 2014	Alcelaphinae	*Alcelaphus buselaphus*	1;10, 2;25, 3;19, 4;6, 5;14, 7;9, 8;17, 11;15, 12;16, 22;23
Steiner et al. 2014	Alcelaphinae	*Beatragus hunter*	1;18, 2;25, 3;19, 4;7, 5;14, 6;16, 8;11, 10;12
Steiner et al. 2014	Alcelaphinae	*Connochaetes gnou*	2;25
Steiner et al. 2014	Alcelaphinae	*Connochaetes taurinus*	2;25
Steiner et al. 2014	Alcelaphinae	*Damaliscus lunatus*	1;10, 2;25, 3;19, 4;14, 5;6, 7;9, 8;17, 11;23, 12;16, 13;15, 18;24, 20;22
Steiner et al. 2014	Alcelaphinae	*Damaliscus pygargus*	1;10, 2;25, 3;19, 4;14, 5;6, 7;9, 8;17, 11;23, 12;16, 13;15, 20;22
Cernohorska et al. 2012	Antilopinae	*Antilope cervicapra*	1;25, 2;29, 3;27, 4;19, 6;24, 7;20, 8;14, 9;17, 10;28, 11;22, 12;16, 13;18, 15;23, 21;26, X;5
Cernohorska et al. 2012	Antilopinae	*Gazella leptoceros*	2;3, 4;6, 7;21, 8;14, 9;17, 10;20, 11;27, 12;16, 13;19, 18;29, 22;25, 23;24, 26;28, X;5
Cernohorska et al. 2012	Antilopinae	*Nanger dama*	2;25, 4;28, 6;25, 7;22, 8;21, 9;26, 10;23, 12;27, 13;17, 14;24, X;5, Y;16
Cernohorska et al. 2015	Antilopinae	*Eudorcas thomsoni*	X;5, Y;16
Cernohorska et al. 2015	Antilopinae	*Eudorcas rufifrons*	X;5, Y;16
Vassart et al. 1995	Antilopinae	*Antidorcas marsupialis*	1;25, 2;29

Cernohorska et al. 2015	Antilopinae	*Nanger soemmerringii*	1;29, 2;15, 8;21, 7;22, 13;17, 12;27, 3;28, 4;28, 10;23, 14;24, 9;26, X;5, Y;16, 6;25
Cernohorska et al. 2015	Antilopinae	*Gazella leptoceros*	2;3, 4;6, 7;21, 8;14, 9;17, 10;20, 11;27, 12;16, 13;19, 18;29, 22;25, 23;24, 26;28, X;5
Cernohorska et al. 2015	Antilopinae	*Gazella cuvieri*	2;3, 4;6, 7;21, 8;14, 9;17, 10;20, 11;27, 12;16, 13;19, 18;29, 22;25, 23;24, 26;28, X;5
Cernohorska et al. 2015	Antilopinae	*Gazella dorcas*	1;10, 2;24, 3;25, 4;7, 6;19, 8;14, 9;18, 11;12, 13;15, 16;22, 17;26, 20;29, 21;23, 27;28, X;5
Cernohorska et al. 2015	Antilopinae	*Gazella marica*	1;6, 2;3, 4;15, 7;21, 8;14, 9;17, 10;20, 11;27, 12;16, 13;19, 18;29, 22;25, 23;24, 26;28, X;5
Cernohorska et al. 2015	Antilopinae	*Gazella erlangeri*	1;10, 2;24, 3;25, 4;7, 6;19, 8;14, 9;18, 11;12, 13;15, 16;22, 17;26, 23;29, X;5
Cernohorska et al. 2015	Antilopinae	*Nanger granti*	1;22, 2;15, 3;11, 4;28, 6;25, 7;20, 8;21, 9;26, 10;23, 12;27, 13;24, 14;29, 16;19, 17;18, X;5
Cernohorska et al. 2015	Antilopinae	*Gazella bennettii*	13;19, 24;28, 26;27, X;5
Cernohorska et al. 2015	Antilopinae	*Gazella saudiya*	4;12, 8;14, 9;23, 11;17, 13;19, 24;28, 26;27, X;5
Robinson et al. 2021	Antilopinae	*Raphicerus spp.*	1;25;14, 2;3, 7;8, 4;16, 11;13, 5;23, 9;19, 12;15, 6;27, 10;29, 17;20, 18;22, 21;24, 26;28
Nguyen et al. 2008	Bovinae	*Bos frontalis*	2;28
Gallagher et al. 1999	Bovinae	*Bos gaurus*	2;28

Ropiquet et al. 2008	Bovinae	*Bos javanicus javanicus*	1;29, 2;28
Gallagher et al. 1999	Bovinae	*Boselaphus tragocamelus*	1;5, 2;3, 6;13, 8;12, 19;27, 24;25
Gallagher et al. 1999	Bovinae	*Bubalus bubalis*	1;27, 2;23, 8;19, 5;28, 16;25
Gallagher et al. 1999	Bovinae	*Bubalus depressicornis*	1;27, 2;23, 8;19, 5;28, 11;20, 17;25
Nguyen et al. 2008	Bovinae	*Bubalus mindorensis*	5;28;11, 2;23, 8;19, 4;14, 16;29
Nguyen et al. 2008	Bovinae	*Pseudoryx nghetinhensis*	1;10, 8;13, 6;19, 4;18, 11;12
Gallagher et al. 1999	Bovinae	*Syncerus caffer*	1;13, 2;3, 5;20, 11;25
Rubes et al. 2008	Bovinae	*Taurotragus derbianus*	1;29, 3;22;2, 4;12, 5;10, 6;11, 7;28, 8;24, 9;20, 14;26, 15;16, 17;27, 18;19, 21;23, Y;13
Gallagher et al. 1999	Bovinae	*Taurotragus oryx*	1;29, 3;22;2, 4;12, 5;10, 6;11, 7;28, 8;24, 9;20, 14;26, 15;16, 17;27, 18;19, 21;23, Y;13
Rubes et al. 2008	Bovinae	*Tragelaphus angasii*	11;22;2, Y;13
Rubes et al. 2008	Bovinae	*Tragelaphus eurycerus*	3;22;2, 5;10, 1;29, 4;19, 6;21, 7;28, 8;15, 9;23, 11;20, 14;27, 16;17, 18;26, Y;13
Rubes et al. 2008	Bovinae	*Tragelaphus imberbis*	1;5, 2;10, 4;7, 3;11, 6;16, 12;18, 8;20, 9;27, 15;22, 14;29, X;13, Y;13
Hassanin et al. 2018	Bovinae	*Tragelaphus scriptus*	Y;13
Rubes et al. 2008	Bovinae	*Tragelaphus spekii*	3;22;2, 5;10, 4;12, 1;29, 8;15, 6;24, 11;18, 7;26, 16;20, 14;19, 9;28, 17;27, 21;23, X;13, Y;13
Rubes et al. 2008	Bovinae	*Tragelaphus strepsiceros*	24;22;2, 4;5, 3;10, 1;29, 6;20, 7;18, 8;17, 12;16, 11;23, 9;27, 19;21, 14;26, 15;28, Y;13

Gallagher et al. 1999	Caprinae	*Ovis aries*	1;3, 2;8, 5;11
Gallagher et al. 1992	Caprinae	*Rupicapra rupicapra*	1;3
Li et al. 2022	Caprinae	*Budorcas taxicolor*	1;22, 2;25, 5;28, 11;23
Claro et al. 1996	Hippotraginae	*Addux nusornuculutus*	1;25
Kumamoto et al. 1999	Hippotraginae	*Oryx beisa*	1;25
Kumamoto et al. 1999	Hippotraginae	*Oryx callotis*	1;25
Kumamoto et al. 1999	Hippotraginae	*Oryx dammah*	1;25, 2;15*
Kumamoto et al. 1999	Hippotraginae	*Oryx gazella*	1;25, 2;17
Kumamoto et al. 1999	Hippotraginae	*Oryx leucoryx*	1;25, 18;19*
Kingswood et al. 1998	Nesotraginae	*Neotragus moschatus*	1;5, 2;3, 4;19*, 18;22*
Kingswood et al. 2001 / Pagacova et al. 2011	Reduncinae	*Kobus ellipsiprymnus*	1;19, 2;25, 5;17, 6;18*, 7;11*, 7;29*
Kingswood et al. 2001	Reduncinae	*Kobus kob*	1;19, 2;25, 3;11, 5;13, 6;29
Kingswood et al. 2001	Reduncinae	*Kobus leche*	1;19, 2;25, 3;10, 4;7, 5;13, 6;18
Kingswood et al. 2001	Reduncinae	*Kobus megaceros*	1;19, 2;25, 4;7, 6;18
Pagacova et al. 2011	Reduncinae	*Redunca fulvorufula*	2;25, 6;10*

Supplementary Table 6: *Statistics of genomic differentiation (F_{ST}) blocks. Linkage disequilibrium (LD) calculated separately for the common (Co.) and defassa (De.) subspecies. 99.9 percentile given for the FST and LD of each region.*

Chr.	Start	End	Length	F_{ST} (99.9%)	LD - C (99.9%)	LD – D (99.9%)
KEL1	11,915,000	15,625,000	3,710,000	0.724	0.249	0.216
KEL2	37,385,000	38,445,000	1,060,000	0.707	0.154	0.256
KEL3	38,425,000	41,405,000	2,980,000	0.719	0.502	0.697
KEL3	45,415,000	47,265,000	1,850,000	0.744	0.398	0.097
KEL4	97,475,000	100,365,000	2,890,000	0.659	0.256	0.205
KEL5	96,035,000	98,055,000	2,020,000	0.654	0.231	0.163
KEL6	1,955,000	5,065,000	3,110,000	0.716	0.513	0.273
KEL8	47,875,000	53,415,000	5,540,000	0.687	0.349	0.237
KEL9	94,595,000	95,805,000	1,210,000	0.671	0.212	0.115
KEL10	55,875,000	57,005,000	1,130,000	0.710	0.267	0.144
KEL11	27,115,000	29,125,000	2,010,000	0.735	0.381	0.096
KEL13	32,565,000	33,375,000	810,000	0.718	0.398	0.109
KEL14	43,975,000	46,915,000	2,940,000	0.669	0.477	0.442
KEL16	50,715,000	52,965,000	2,250,000	0.751	0.178	0.252
KEL17	11,035,000	12,865,000	1,830,000	0.735	0.504	0.222
KEL17	44,085,000	48,945,000	4,860,000	0.746	0.304	0.261
KEL21	19,285,000	19,915,000	630,000	0.769	0.152	0.070

Supplementary Table 7: *List of genes found on the two blocks of high genomic differentiation (F_{ST}) surrounding the centromere of chromosome KEL3 and the highest F_{ST} between the common and defassa subspecies of the window containing the gene.*

Gene	**F_{ST}**
IFT140	0.720
LOC102183842	0.705
ARHGEF4	0.699
LOC102186637	0.697
LOC108633841	0.680
IMP4	0.679
KCTD17	0.677
CLCN7	0.676
AMER3	0.669
CRAMP1	0.658
TBC1D24	0.658
ZNF500	0.657
LOC108633241	0.654
ADCY9	0.651
HN1L	0.646
E4F1	0.645
ECI1	0.644
PRSS21	0.643
AMDHD2	0.642
LOC101903064	0.639
TELO2	0.635
ABCA3	0.635
TRAF7	0.633
ROGDI	0.632
PRSS22	0.631
UNK	0.630
CYFIP1	0.630
CACNA1G	0.629
RNPS1	0.629
TSC2	0.628
TUBGCP5	0.628
UBN1	0.627
PTPN18	0.627
OCA2	0.626
NTPCR	0.624
LOC108633325	0.624
ANKS3	0.622
MAPK8IP3	0.622
Sep-12	0.621
DNASE1	0.620
MEIOB	0.620
TFAP4	0.620
LOC102189536	0.619
FLYWCH2	0.618
RAB26	0.617
NTN3	0.615
PKD1	0.615
CASKIN1	0.614
CLUAP1	0.609
MGRN1	0.609
GLYR1	0.607
EME2	0.607
CCNF	0.606
LOC132346924	0.606
FLYWCH1	0.603
BRICD5	0.598
NIPA1	0.598
HAGH	0.597
NAA60	0.596
ATP6V0C	0.595
GNPTG	0.591
FAHD1	0.589
KREMEN2	0.588
CCDC115	0.587
TIGD7	0.584
LOC112442547	0.583
C25H16orf59	0.582
LOC102185624	0.580
RNF151	0.580
PPL	0.577
IGFALS	0.577
PRSS27	0.572
MEFV	0.571
GREP1	0.571
LOC102178686	0.571
PAQR4	0.570
HS3ST6	0.569
C25H16orf96	0.567
LOC102188914	0.567
ZNF205	0.562
LOC108633912	0.560
SRL	0.559
OR2C1	0.557
LOC107131805	0.547
ZNF75A	0.547
RPL3L	0.545
NTHL1	0.544
BAIAP3	0.544
GFER	0.539

UBE2I	0.534	PRSS33	0.505	DNAJA3	0.423
CDIP1	0.529	ELOB	0.500	NMRAL1	0.413
ZNF598	0.527	ZNF200	0.496	SLC9A3R2	0.412
C25H16orf71	0.525	SRRM2	0.482	ZNF263	0.390
LOC108633877	0.523	C25H16orf90	0.468	SNX17	0.388
GLIS2	0.520	OR1F1E	0.467	IL32	0.383
LGSN	0.519	NLRC3	0.461	TRAP1	0.381
CCDC154	0.517	BICDL2	0.460	MMP25	0.376
CASP16	0.516	SLX4	0.457	CLDN6	0.323
LOC108633901	0.514	ZNF174	0.454	ZG16B	0.323
HMOX2	0.510	LOC617663	0.453	ZSCAN10	0.298
TPSB2	0.510	LOC132346709	0.447	LOC108633902	0.217
NDUFB10	0.508	SYNGR1	0.438	GTSE1	0.210

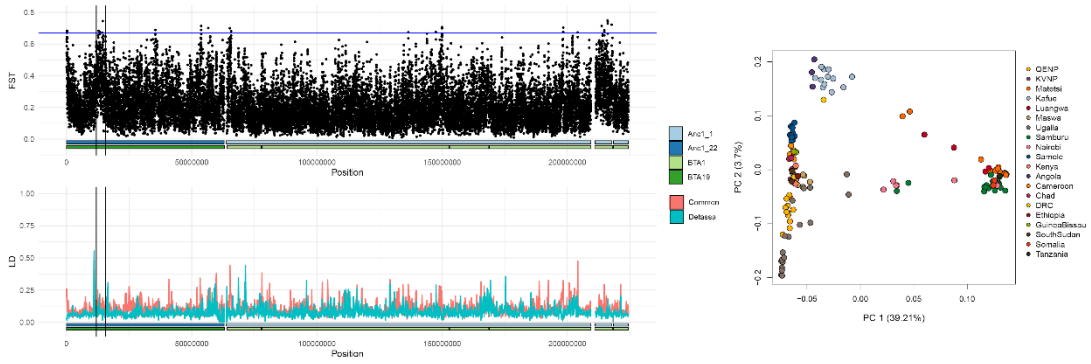
Supplementary Table 8: *List of genes found on the genomic differentiation (F_{ST}) block on KEL6 and the highest F_{ST} between the common and defassa subspecies of the window containing the gene.*

Gene	**F_{ST}**		
EXOSC9	0.630	NAF1	0.411
TMA16	0.587	LYRM7	0.388
APELA	0.565	LOC108638059	0.337
TKTL2	0.551	Mar-01	0.332
MARCHF1	0.546	RPS4Y1	0.314
CCNA2	0.510	NPY5R	0.305
EIF4B	0.478	LOC132344125	0.304
LOC102177135	0.411	CHCHD3	0.200

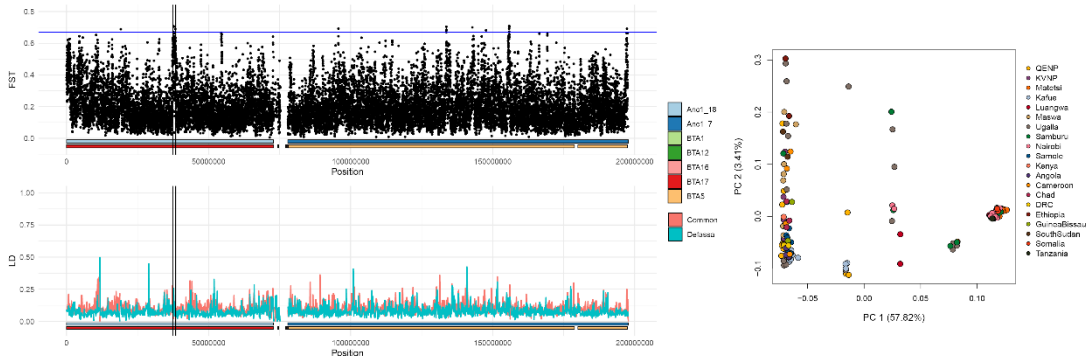
Supplementary Table 9: *List of genes found on the genomic differentiation (F_{ST}) block on KEL17 and the highest F_{ST} between the common and defassa subspecies of the window containing the gene.*

Gene	F_{ST}
DDX19A	0.739
PDPR	0.721
DDX19B	0.719
FA2H	0.659
WDR59	0.632
FUK	0.627
LOC102171180	0.625
SF3B3	0.620
VAC14	0.607
CFDP1	0.603
RBM4B	0.596
ADAT1	0.592
KARS	0.592
TMEM170A	0.581
BCNT	0.579
HNRNPA1	0.579
LOC102177976	0.579
LOC102172995	0.557
COG4	0.549
TERF2IP	0.532
ZFP1	0.519
IL34	0.493
AARS	0.488
EXOSC6	0.488
LOC108637950	0.488
LDHD	0.467
GABARAPL2	0.448
ZNRF1	0.443
TMEM231	0.434
ST3GAL2	0.429
MTSS1	0.395
MTSS1L	0.395
BCAR1	0.391
LOC102176156	0.391
P97BCNT	0.385
MLKL	0.383
LOC102174148	0.377
LOC132346553	0.372
RFWD3	0.372
CHST6	0.352
ZFAND3	0.329
LOC102173875	0.305

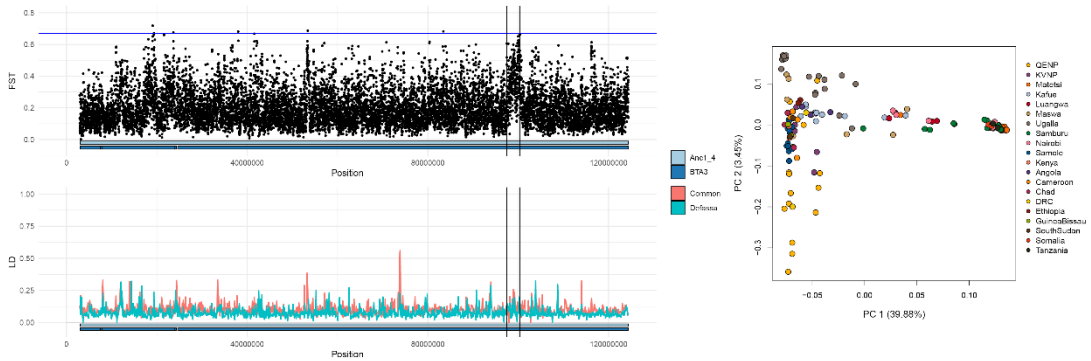
KEL1



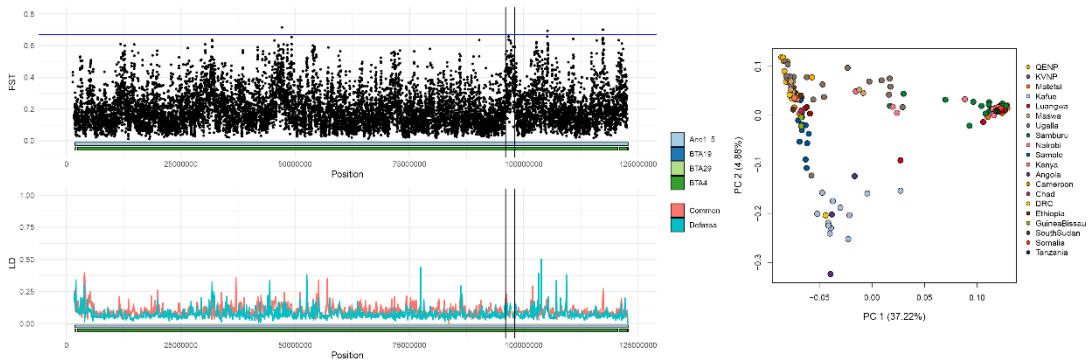
KEL2



KEL4

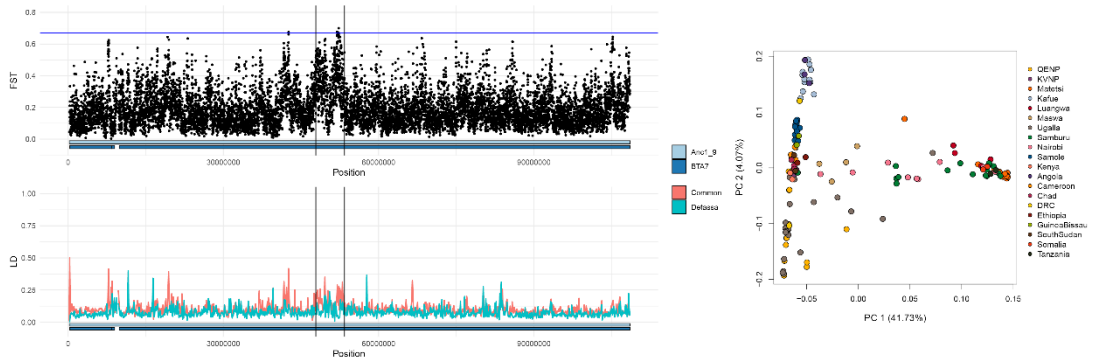


KEL5

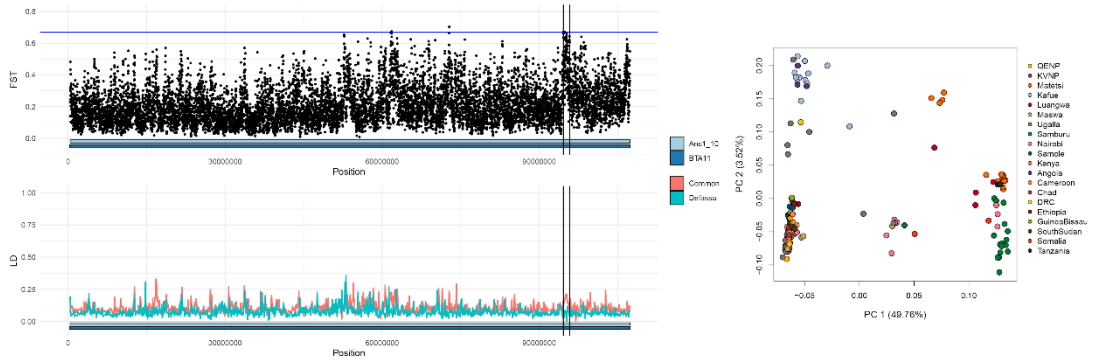


(Supplementary Figure 6 cont.)

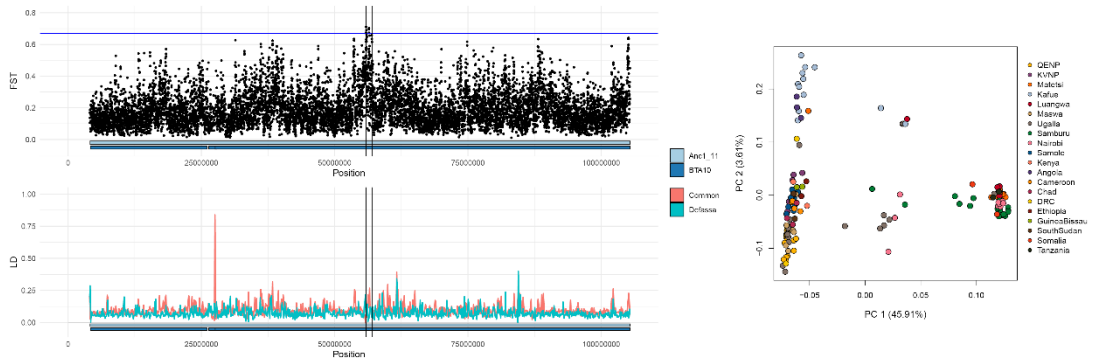
KEL8



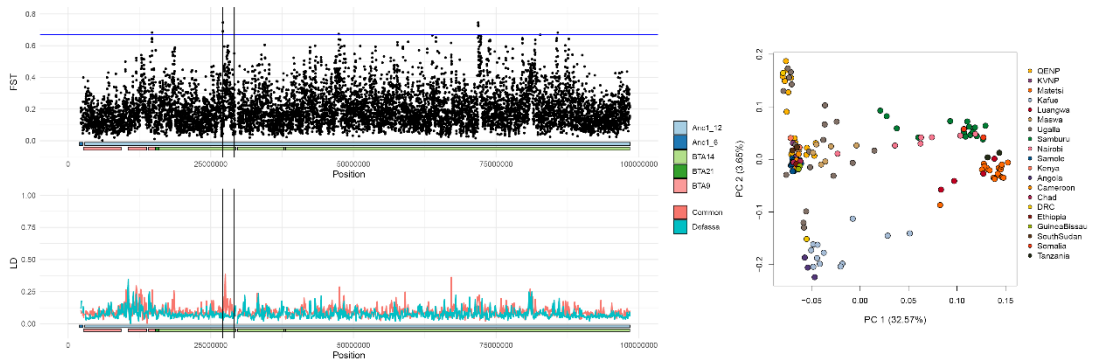
KEL9



KEL10

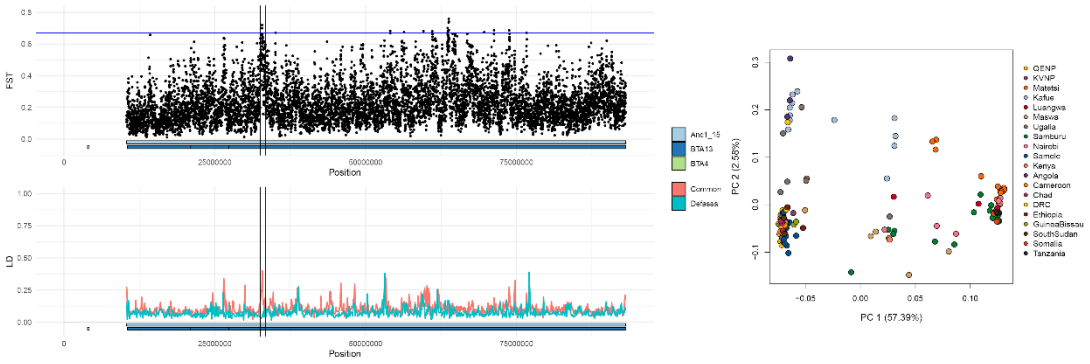


KEL11

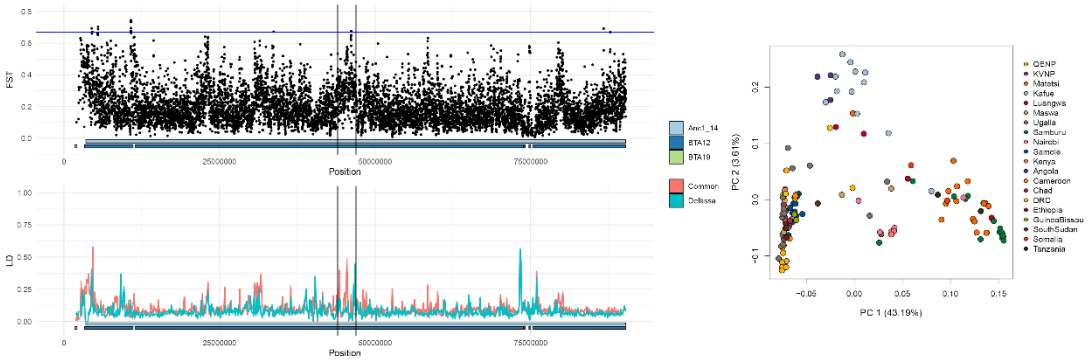


(Supplementary Figure 6 cont.)

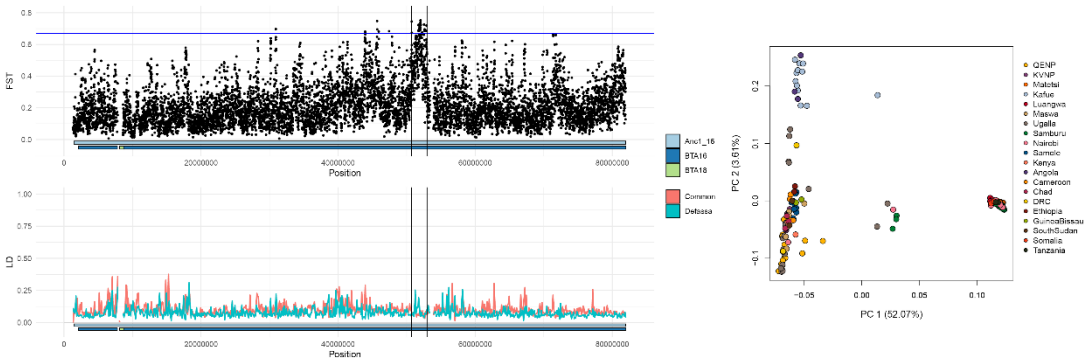
KEL13



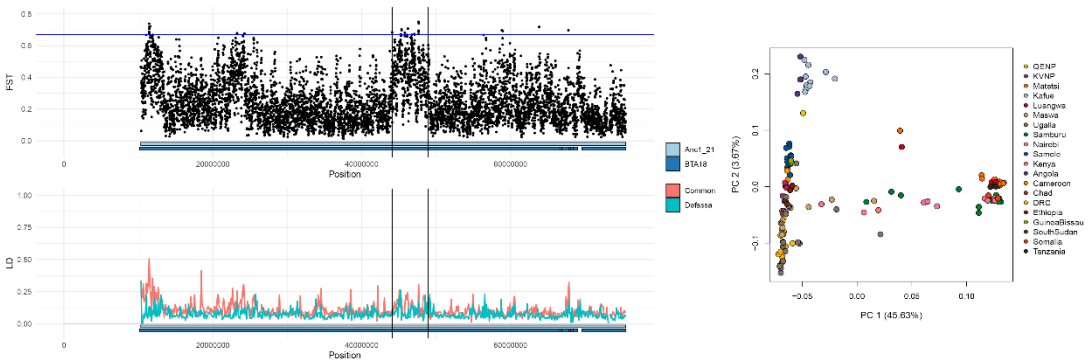
KEL14



KEL16

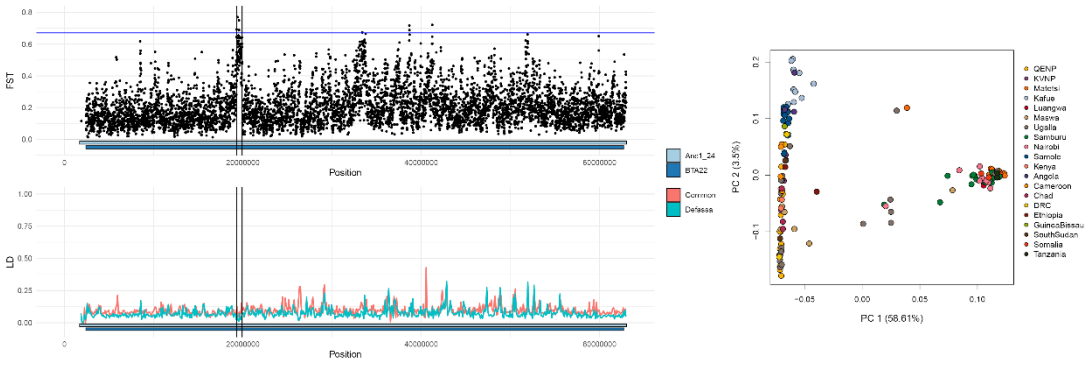


KEL17



(Supplementary Figure 6 cont.)

KEL21



Supplementary Figure 6: *Blocks of high genomic differentiation (F_{ST}) between common and defassa waterbuck calculated in 10 Kb windows. Linkage disequilibrium (LD) calculated in 100 Kb windows for each subspecies. Synteny of each chromosome to Anc1 and BTA (cattle) denoted by the two horizontal bars on each plot, respectively. Selected region of the putative SV shown by the two vertical lines. PCA computed on the genomic sites of the selected region.*