Faculty of Science, Engineering, and Agriculture

Department: Biological sciences

University of Venda

Thohoyandou Limpopo

South Africa

Genomic mechanisms underpinning phenotypic diversity in the spiral-horned antelopes (Genus: *Tragelaphus*)

by

Rambuda Thabelo

15018146

Research dissertation for Masters in Biological Science: Zoology

Supervisor: Professor Yoshan Moodley

Co-supervisor: Dr Andrinajoro Rakotoarivelo

Table of Contents

Declaration

I, Thabelo Rambuda (15018146), hereby declare that this work was conducted by me and has not been handed in for a degree at this university or any other university. This work was carried out for the Master of Science degree in Zoology.

Signature  …….. Date……01.07.2022……

Abstract

The world is old, full of diversity, and the history of all organisms that once lived on Earth is recorded in the DNA of its descendants. The genomes of living organisms contain ancestral information that, if analyzed, reveals the underlying mechanisms to explain an organism's evolutionary history. Therefore, it is crucial to study the whole genomes of highly diverse and specialized groups of organisms that could help our understanding of the speciation process. The continent of Africa is home to phenotypically diverse spiral-horned antelopes (genus *Tragelaphus*) which have gone through a recent adaptive radiation. Previous studies on *Tragelaphus* have argued that they comprise either nine or ten species based on mtDNA and nuclear DNA respectively/chromosomal number difference. With the same mentioned molecular data, there is discordance in their previously reconstructed species tree, placing species in different clades with different markers. At the mtDNA level, the nyala (*T. angasi*) is sister to the bushbuck (*T. scriptus*) making the mtDNA diversity polyphyletic within the bushbuck complex. The two bushbuck species and other phenotypically similar non-sister *Tragelaphus* lineages lead to the suggestion by scientists that some phenotypes evolved through convergent evolution. In this study, one whole genome of each *Tragelaphus* lineage was sampled with the aim to analyze the genome-wide relationship of these species by reconstructing their phylogenetic species tree. The study also aims to assess the genome-wide levels of diversity and to assess whether there has been gene flow between species which could have led to phylogenetic discordance among traditional markers. The relationship was analyzed with non-model based PCA and biological model-based IBS and maximum likelihood. All the methods used for structure analysis revealed the same genomic structure and confirmed other studies showing that morphologically similar *Tragelaphus* species were not most closely related at the genome level. The reconstructed genome-wide species tree was used for the assessment of introgression between species. Most of the observed gene flow was ancestral, the alleles of which are randomly kept in some lineages and passed from generation to generation but lost in others. Therefore, I propose that some phenotypic similarities between unrelated species could be due to high ancestral gene flow between these non-sister lineages. To confirm this would require further investigation using more samples for each species.

Keywords: *Tragelaphus,* Whole genome, evolution, speciation, convergent phenotype, Gene flow.

1.Introduction

1.1 Evolution and tracing lines of descent.

When one considers all the living things on the planet, the differences between them are most striking. This is because there are forces at work to bring about evolution (Stephens 2004), which is a continuous process of change. These forces of evolution are mutation, genetic drift, gene flow and selection. Non-random mating may also be seen as another force of evolution but unlike other forces mentioned above, it does not change the alleles frequency. Mutation alters the DNA of an individual and brings about genetic variation (Nei 1983). These mutations are mainly due to the response of individuals to the environment where natural selection will select the advantageous mutation in coding regions to be passed on to the next generation or be eliminated from the population (Nei 1983). Genetic drift was proposed by Fisher (1923) and Wright (1931) as a force of evolution, it is a random chance of eliminating an allele from population (Lande 1976), so it has neither high nor low fitness. All these changes in the genetic material are migrated in and out of the population by gene flow, together bringing about evolution.

The concept of evolution is so important that nothing in biology makes sense if not in the eyes of evolution (Dobzhansky 2013). Since every living organism originated through descent from a common ancestor (Ayala 2009), so their lines of descent can be traced backwards in time towards a common ancestor. Tracing lines of descent, in the presence of all four forces of evolution is the main task of evolutionary biologists (Garamszegi 2014). This helps in understanding the world, its living things, and how these interact with the environment in which they live. It is also important to know such information in order to make better-informed conservation management decisions.

The lines of descent can be defined as the genetic make-up that individuals inherit from their parents and which they in turn pass on to their offspring in an unbroken line. With each cycle of replication, the inherited genetic material undergoes mutations and segregation which creates variation (Gould 2002). Thus, offspring slightly differ from their parents. These mutations can be affected in different ways. The new mutations can be affected by genetic drift in a way of keeping or removing them from a population (Masel 2011). If the population is large enough new mutations may become common in a population and having many alleles in a population means many lines of descent in a population, making it difficult to trace back. As for small population, there is increased chance that new mutations may be lost and not passed to the next generations, increasing the rate at which lineages sort, thus makes it easier to trace lines of descent. Genetic

drift affects the whole genome, meaning that the altered DNA will be passed on to next generations.

The genetic changes are also affected by the environment the individuals live in. If the changes in genetic make-up of the offspring help them to adapt better in their environment, there is a greater probability that they survive to reproduce (increase fitness) thus passing on the essential genetic changes to their offspring. These changes that help in adaptation and are passed, are said to have been selected (Walsh 2000). This is the selection force of evolution which affects only the part of the genome that increases survival fitness of the individual. All the changes that may be adaptive and selected, may cause the formation of new species (speciation) in the next generations.

The selective conditions causing speciation may promote constant and gradual change over time, as suggested by Darwin (1859). It is generally accepted that in animals, the formation of new species occurs mainly by mechanisms that hinder or stop gene flow between populations that are diverging (Dobzhansky 1940) due to genetic drift and selection. Speciation can be divided into different modes depending on the way in which gene flow is stopped. Ernst Mayr proposed that new species arise when there is geographic barrier, that is the allopatric model, where new species emerge when a biological population is separated by geography, preventing gene flow between the separated populations (allopatric model) also further explained by Stanley (1975). Peripatric speciation also proposed by Ernst Mayr in (1954) also prevents gene flow between original and formed populations by a geographic barrier but in this case one population is smaller than the other, the small population will be highly affected by genetic drift and loses its variability (Barton 1996). When two diverging populations occupy part of (are parapatric) or all of (are sympatric, Smith 1966) the same geographic space, gene flow may still be reduced if individuals in a population develop different behaviors such as sexual selection (Darwin 1871) or any other kind of non-random mating.

Diverging species that are closely related and living in close proximity to one another geographically may still have gene flow (Wagner 1873). This can even happen when a portion of each population are in actual genetic contact, and the majority of individuals do not have frequent contact. In such cases new species may arise where there is contact. This was seen in salamanders species which was rapidly evolving in a short period of time (adaptive radiation), in response to environmental changes (Niemiller 2008). The Tennessee cave salamanders thus evolved from spring salamanders through divergence with gene flow (Niemiller 2008). In such adaptive radiations, the new species generally arise when there has first been allopatric

divergence that created phenotypic differences or adaptations, then followed by gene flow during secondary contact (Schield *et al.* 2019). Another classic example of adaptive radiation is of Darwin's finches in response to the El Niño phenomenon and glacial and interglacial cycles (Lamichhaney *et al.* 2015). The finches evolved the adaptive functional beak morphology and which throughout the radiation was being shared extensively (adaptive introgression) among the finches leading to the formation of new species (Lamichhaney *et al.* 2015). This adaptive introgression is witnessed also in the *Heliconius* butterflies (Pardo-Diaz *et al.* 2012) where mimicry evolved from introgression of adaptive alleles (gene regulating wing pattern) from *H. melpomene* to *H. cydno*, two different species known to have a strong barrier but exchange genes occasionally.

The most recent animal adaptive radiations occurred during Plio-Pleistocene, during the rising and falling of paleoclimate (Vrba 1995) and tectonic uplift (Pickford 1990). During this adaptive radiation, speciation was at its most spectacular and the bovids underwent radiation into many tribes, becoming more diverse and abundant amongst all herbivores (Du Toit and Cumming 1999). *Tragelaphus* forms part of the bovid radiation, wherein within this genus, there was great diversification amongst species which led to what is seen today.

Gene flow in the form of adaptive introgression, admixture, or hybrid speciation destroys the genomic structure created by genetic drift and selection, this makes tracing lines of descent more complicated when diverging populations have not fully developed reproductive isolation mechanisms and they come into secondary contact having gene flow, as described above. When estimating the lines of evolutionary descent (reconstructing phylogeny) it is difficult to distinguish between gene flow and conserved ancestral polymorphism, which is also known as incomplete lineage sorting (ILS) (Tajima 1983; Zhou 2017) since the two phenomena show the same pattern of shared diversity. When reconstructing a tree of admixed populations, they can be clustered in intermediate positions between parental clades since they share great proportions of alleles with both. Their branch lengths will be short and not very differentiated from either parental clade. This might be one way to distinguish whether allele sharing between species is due to ILS or gene flow.

1.2 Reconstructing lines of descent (phylogenies)

Studying phylogenies before the 1970's was mainly based on morphological traits of species, this was first practiced by Carl Linnaeus the taxonomist. To trace the lines of descent, traits were ordered in terms of similarities: However, similar traits or non-similar traits does not actually mean the sharing or not sharing of common ancestor respectively. Only after the birth of systematics and major advances in molecular genetics came in the 1970's, was it possible then to determine whether traits among species could be explained by homology or analogy. Homology refers to the acquisition of traits from a common ancestor (Boyden 1947; Abouhef 1997), whereas analogous traits are acquired through exposure to similar selective pressures, indicating that species have independently evolved adaptive similarities (Boyde 1947). Homology and analogy lead to the concept of divergent and convergent evolution respectively. Divergent evolution occurs when a population of species is exposed to different selective pressures (Rosenblum 2006), resulting in divergent traits, and convergent evolution occurs when populations of species are exposed to similar selective pressures (Rosenblum 2006), resulting in analogous traits. The above-mentioned evolution leads to the formation of new species from the existing which is known as speciation. The acquisition of new traits in species must show functionality that makes them adapt in their environment (Fisher 1985).

In 1904, Nuttall employed immunological test to deduce the relationship of human beings with other primates. That was when the use of molecular data was introduced. But during the time of Nuttall the molecular data used were proteins and they had their limitations, that is when in the 1980s DNA-based phylogenetics began to be used widely. Phylogeny could be reconstructed simply by measuring pairwise distance of morphological characters or genetic data (DNA sequence alignment). The use of genetic data is a modern and more reliable way of tracing the lines of decent since the forces that structure these data are taken into consideration when reconstructing phylogeny.

Harrison and Langdale (2006) explains how molecular data can be used to reconstruct phylogenies by computers, where first step is to align the DNA sequences, which can be done by different programs e.g Burrow-willer alignment (BWA) (Li and Durbin2009) and then employ the phylogenetic analysis to get the comparative data in a right format to reconstruct a tree. Phylogenetic analysis methods are different, with main methods being assessment of genetic distance, where some includes maximum likelihood (ML) (Holder and Lewis 2003). The distance method calculates the pairwise distances between sequences (identity by sequence) and the

sequences that are most similar will cluster (Harrison and Langdale 2006), however this is disadvantageous because reducing data into distances comes with loss of some information (Holder and Lewis 2003) like the ancestral information, since it is only comparison of aligned sequences. The maximum likelihood method will calculate the likelihood that the data set fits a tree generated from the data set. In this method there should be a biological model specified that best fits the DNA changes of sequences in data set (Harrison and Langdale 2006) to avoid bias phylogenetic inferences (Posada and Crandall 2001). These biological models describe the evolutionary changes of the DNA sequences from the simplest Jukes Cantor (JC) that assumes similar rate of change between all nucleotides (Jukes and Cantor 1969) to the most complex General time-reversible model (GTR) that assumes different rates of change. Confidence in the reconstructed tree can be obtained by performing bootstrap analysis (Hillis and Bull 1993). Bootstrap is statistically resampling the data with replacement, done a number of times to estimate significance of the model (Silakari and Singh 2020).

1.3 Spiral-horned antelopes (Genus: *Tragelaphus*)

The spiral-horned antelopes (*Tragelaphus*), which are the focus group of this project, consist of one of the more remarkable radiations within the Bovidae that occurred from the late Miocene-early Pliocene (Hassanin and Douzery 1999). *Tragelaphus* belong to the Tragelaphini tribe and are in the Bovidae family (Hassanin and Douzery 1999). The genus occurs solely in Africa, and several species appear to be adapted to the different and changing environments of that continent, leading to great phenotypic variation (Rakotoarivelo *et al.* 2019). Scientists suggest that this variation has resulted from a combination of divergent evolution, when closely related species or populations become phenotypically different from one another as they become adapted to different ecological niches, and convergent evolution, where non-sister species or populations evolve to become phenotypically similar to one another because they are exposed to similar environment conditions (Moodley and Bruford 2007; Wronski and Moodley 2009; Willows-Munro *et al.* 2005). This will be elaborated further in the next sections.

There is also a debate about how many (between 9 and 10) species the spiral-horned antelopes comprise. The nine *Tragelaphus* Blainville recognized extant species are, *Tragelaphus oryx* (eland), *Tragelaphus imberbis* (lesser kudu), *Tragelaphus angasi* (nyala), *Tragelaphus derbanius* (giant eland), *Tragelaphus strepciceros* (kudu), *Tragelaphus buxtoni* (mountain nyala),

Tragelaphus spekei (sitatunga), *Tragelaphus euryceros* (bongo) and *Tragelaphus scriptus* (bushbuck). However, Moodley *et al.* (2009) and Hassanin *et al.* (2018) suggested that the bushbuck may comprise two species, the *T. s. scriptus* (kewel) and *Tragelaphus s. sylvaticus* (imbabala). Moodley and Bruford (2007) revealed evidence of high levels of genetic diversity within both bushbuck lineages which were polyphyletic on the mitochondrial level, with *T. s. scriptus* occupying North-western part of Africa being sister with *T. angasi* (Moodley *et al.* 2009), and *T. s. sylvaticus* occupying Africa's South-eastern part being closely related to the clade comprising *T. euryceros–T. spekei* (Moodley *et al.* 2009). Furthermore Moodley *et al.* (2009) suggested that for the two lineages to be referred to using the two species names, their findings must be corroborated with nuclear data and not only mtDNA. The bushbucks were later said to comprise two species based on the different karyotype number (Hassanin *et al.* 2018), but these findings need further investigations with genome level data, and with respect to the forces of evolution acting upon them.

Despite not being very species rich, *Tragelaphus* species nevertheless possess large differences in habitat preference, coiling of horns and sexual dimorphism in presence/absence of horns, body size as well as coat colour (Kingdon 1982). Spiral-horned antelopes are found in almost every habitat throughout sub-Saharan Africa (Matthee and Robinson 1999; Willows-Munro *et al.* 2005; Moodley *et al.* 2009). Paragraphs below describes the distribution, adaptation as well as the morphology of each *Tragelaphus* species.

Lesser kudu

Photo: WhoZoo (http://whozoo.org), male and female lesser kudu

Redrawn from Institute of Applied ecology, 1998.

T. imberbis occupies African bushlands and thickets in the peninsula of northeast Africa (Bock *et al.* 2014). Their habitat is the dense thickets, and they are mostly active at night and dawn. Males reach about 95-105 cm at their shoulders and females reach 90-100 cm (Estes 2020). In this spiral-horned antelope only the males have horns, and they have two to two and a half twists. There is distinguishable sexual dimorphism (different characteristics

of sexes in the same species), one of which is male possession of horns and the males being larger than females with their bodies weighing 92-108 kg and females 56-70 kg, respectively.

Nyala

Photo: Steve Cornish, 2007, male nyala

T. angasi is found in Lowveld vegetation in south-eastern Africa (Malawi, Mozambique, South Africa, Swaziland, and Zimbabwe) (Tello and Van Gelder 1975). *T. angasi* shows great sexual dimorphism having only males with yellow tipped horns with females having no horns. The males have an apron of long hair, and their bodies can grow to be 104 -121 cm tall and weigh up to 58 kg (Tello and Van Gelder 1975). Young males have colour pattern like that of a female and darkens to greyish-black as they grow older (Tello and Van Gelder 1975) having 10-14 vertical, parallel white stripes on the flanks. These stripes are reduced on males as they grow older (Tello and Van Gelder 1975). Females are much smaller, weighing 54-68 kg and standing 82-106 cm tall at the shoulder. They have bright chestnut-brown colour, commonly confused with female bushbuck and they lack male's hairy coat (Furstenburg 2002).

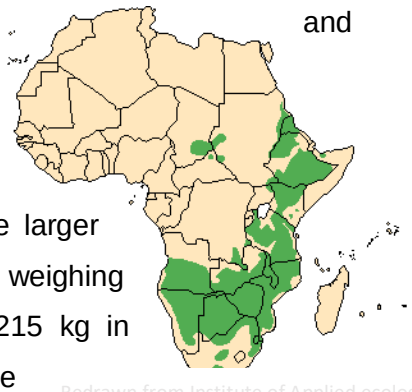
Redrawn from Institute of Applied ecology, 1998.

Photo: Bernard Dupont, 2011, female nyala

Greater kudu

Photo: Bernard Dupont, 2001, male greater kudu

T. strepciceros lives in thickets and dense woodlands throughout eastern and southern Africa. They show sexual dimorphism with body size where males are larger than females, with males weighing between 190-315 kg and 120-215 kg in females (Nersting and Arctande 2001).



Redrawn from Institute of Applied ecology, 1998.

Amongst the *Tragelaphus* genus, *T. strepciceros* males have the largest horns, known to reach length of up to 180 cm and the females do not possess horns. Mostert and Hoffman (2007) referred to this species as the most magnificent of Africa's antelopes and Nersting and Arctande (2001) revealed that they are great demand of trophy.

Photo: Bernard Dupont, 2019, female greater kudu

Common eland

Photo: Bernard Dupont, 2018, male common eland

T. oryx occurs throughout the savanna woodlands of eastern and southern Africa. It was suggested to be the second largest of the spiral-horned antelopes (Pappas 2002) having males larger than females. The male grow to a shoulder height of 163 cm and 142 cm in females, that is a 125-178 cm range with weight range of between 340–

Redrawn from Institute of Applied ecology, 1998.

Photo: Michael Seeley, 2018, female common eland

942 kg (Wronski *et al.* 2009) making it the largest of the *Tragelaphus*. The colour of their whole-body coat is almost uniform (Pappas 2002).

Giant eland

Photo: Charles J. Sharp, 2016, male giant eland

T. derbianius is distributed in West and northern sub-Saharan Africa and inhabits the Sudanian and Guinean savannas (Rakotoarivelo *et al.* 2019). This was said to be the largest of the *Tragelaphus* followed by *T.oryx* (Pappas 2002). *T. derbanius* grow to a shoulder height of 140–176 cm with the weight of 300–907 kg. They have visible white spot on each cheek of the *T.derbanius* (Pappas 2002).

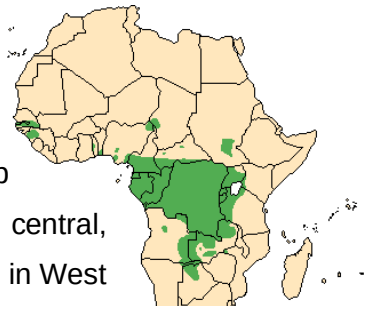
Redrawn from Institute of Applied ecology, 1998.

Photo: Dr. Alexey Yakovlev, 2015, female giant eland

Sitatunga

Photo: Fernando Losada Rodríguez, 2009, male sitatunga

T. spekei is adapted to wetlands, with its distribution in the Congo Basin's lowland forests, as well as near swamp areas within the savannas of central, eastern, and southern Africa, and in West Africa, where populations are primarily found in forests bordering river deltas (Furstenburg 1863). Their shoulder height and weight range are between 75-125 cm and 40-120 kg respectively (Furstenburg 1863).



Redrawn from Institute of Applied ecology, 1998.

Photo: H Zell, 2019, female sitatunga

Bongo

Photo: Joe Schneid, 2007, male bongo

Redrawn from Institute of Applied ecology, 1998.

The forest-savanna mosaics of West and Central African lowlands as well as the Kenya highlands are home to the ***T. euryceros*** (Kingdon 1982)***.*** They possess sexual dimorphism where both males (~80 cm) and female have large, dark, blackish-brown horns which can develop white or yellowish tip as they grow. These reddish-brown with white and black markings (Elkan 2003). *T. euryceros* are the largest of the antelopes that live in the forest (Ralls 1978) with males larger than the females. The number of stripes on each side of their bodies is usually between 10 -15 (Ralls 1978).

Photo: Linda, 2010, female bongo with her calf

Mountain nyala

Photo: KrisMaes, 2007, male mountain nyala

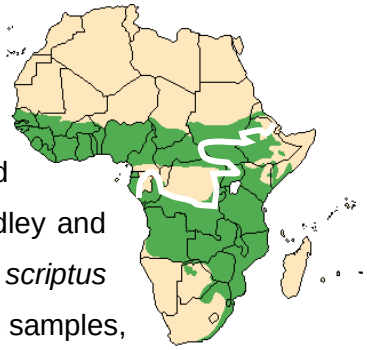
T. buxtoni is endemic to the highlands of Ethiopia (Evangelista *et al.* 2007; Tadesse and Kotler 2014). Tadesse and Kotler (2013) says grassland and woodland are the two major habitats of *T. buxtoni.* This species is listed as endangered by the IUCN (Tadesse and Kotler 2013; Atickem *et al.* 2022), it is prey to the predators such as leopards (Tadesse and Kotler 2013). The *T. buxtoni* coat color is tan grey and has distinctive white markings on their chest and face connecting the eyes (Evangelista *et al.* 2007). They may also have a row of six to ten white spots visible in the middle of the sides. They show sexual dimorphism in a way that adult males are significantly larger than

Photo: KrisMaes, 2007, female mountain nyala

adult females (Tadesse and Kotler 2013). The male weighs between 180-320 kg and females between 150-200 kg (Malcolm and Evangelista 2005) with only males possessing horns (Evangelista *et al.* 2007).

Bushbuck

Photo: Congo Forest Safari group, 2019, male kewel



Redrawn from Institute of Applied ecology, 1998.

T. scriptus is the most widespread of the spiral-horned antelopes (Moodley *et al.* 2009) and the smallest in size. Moodley and Bruford (2007) collected *T. scriptus* genetic data from museum samples, and their data accounted for all known phenotypic variation over the whole species range. The study revealed that of the collected data, the genetic variation divides the species to major groups: *Scriptus* (kewel) *and Sylvaticus* (imbabala). This division (white line on map) runs almost along the axis of Rift valley. In the Western-Northern side of the valley is *Scriptus* and the Eastern-Southern side is the *Sylvaticus.* The bushbuck *T. s. scriptus* show shoulder height of approximately 71 cm and can grow to be between the sizes of 24 kg and 80 kg in weight (Furstenburg 1766). With the bushbuck *T. s. sylvaticus* showing shoulder height of between 60-85 cm and weight of between 25-65 kg (Furstengurg 1766).

Photo: Bilal photography, 2019, female kewel

Photo: Bernard Dupont, 2016, male imbabala

Photo: Charles J. Sharp, 2018, female imbabala

1.4 The evolution of the spiral-horned antelopes

1.4.1The fossil record

Hassanin and Douzery (2003) suggest that *Tragelaphus* is one of the grazing and browsing bovid genera that went through several evolutionary radiations. The high divergence of *Tragelaphus* which inhabit different habitats of the African continent is suggested to have started to occur approximately 14 million years ago (Mya) when the global cooling peaked (Willows-Munro *et al.* 2005). As environments were changing the species had to adapt to new environments which led to isolation of some species (e.g *T.angasi* and *T. imberbis*) whilst new species of the Tragelaphini were being formed (speciation) co-inciding with the period of most recent adaptive radiations in many mammals (Willows-Munro *et al.* 2005).

Approximately 14 Mya was when the tribe Tragelaphine diverged from other tribes of the Bovidae (Willows-Munro *et al*. 2005). The fossils evidence for the genus *Tragelaphus* are found in a number of late Miocene sites, possibly as early as 6.5 Mya (Gentry 2010). Some researchers suggest that fossils species which are now extinct, may be ancestors to current species. Bibi 2011 presented the fossil of an extinct species (*Tragelaphus moroitu*) from the Mio-Pliocene (5-23 Mya) with horns similar to *T. angasi* and *T. spekei*. Its horn morphology is thought to represent the ancestral state from which all other *Tragelaphus* horn forms descended, these horns are distinguished by three-quarter-whorl spiraling and are weakly inclined, having three well-pronounced keels (anterior, posterolateral, posteromedial) (Bibi 2011).

During the Pliocene, another *Tragelaphus* species (*Tragelaphus rastafari*) existed (Blondel *et al.* 2018). This species existed until late Pliocene 2.8 Mya when it went extinct as a result of shifts in global and regional environment, as well as biotic turnover and was replaced by *Tragelaphus nakuae* (Bibi 2011). The *T. nakuae* fossil was found at Hadar and depicted that its females were horned as observed in extant *T. oryx* and *T. euryceros* (Reed and Bibi 2011). *T. nakuae* was the most regular occurring species at hominin sites during Plio-Pleistocene (1.8-5.3Mya) (Bibi 2011). Its fossils show significant morphological variation over the duration of its survival and this species has survived global and regional climatic and environmental changes. Knowing the morphological relationships of extinct species, we can appreciate that *Tragelaphus* contained more variation in its morphology than the current nine or ten species.

1.4.2 Morphological studies.

There have been morphological studies that attempted to reconstruct the evolutionary history of *Tragelaphus*, with varying degrees of success. Amongst the extant species, evolutionary relationship is controversial with different markers depicting different relationships. These are detailed below. The earliest attempts were based on morphology. Morphology (phenotype) is the result of both an individual's genetic makeup and the influence of their environment. It can therefore be unwise to base evolutionary history only on morphological data. Nevertheless, before the days of genetic data, morphology was the only means of reconstructing evolutionary relationships

Previously by the 1970s, morphological study generally accepted relationships among spiral-horned antelopes to be classified into three genera, based on their morphology (phenotype). The three genera were *Tragelaphus* (containing *T. strepciceros*, *T. imberbis*, *T. angasi*, *T. buxtoni*, *T. spekei*, and *T. scriptus*), *Taurotragus* (containing *T. oryx* and *T. derbianus*) and *Boocercus* (for *T. euryceros*) (Ansell 1971 in Willows-Munro *et al.* 2005). A further morphological study by Kingdon (1982) on the cranial similarities, placed *T. spekei* and *T. angasi* as sister taxa, as well as *T. imberbis* and *T. strepsiceros* were placed as sister taxa. Gentry (1990) latter suggested that the presence of horns in both sexes united the former genera *Boocerus* (*T. euryceros*) and *Taurotragus* (*T. oryx* and *T. derbanius*). Beside uniting the two genera, it was suggested that the genus *Taurotragus* is united with *Tragelaphus* due to the elands (*T. derbanius* and *T.* oryx) and kudus (*T. strepciceros* and *T. imberbis*) having similar body conformation and pelt coloration. It was also suggested to include the *Boocerus* genus within *Tragelaphus* due to the ability of *T. euryceros* and *T. spekei* to hybridize in captivity (Tijskens 1968; Wilson and Reeder 2005). These were later reconciled with molecular data to all belong to the genus *Tragelaphus* (Hassanin and Douzery 1999; Matthee and Robinson 1999).

1.4.3 Molecular studies

Molecular studies came after the morphological studies, and they have employed many different genetic markers. First, the molecular data was mitochondrial DNA (mtDNA) cytochrome *b* sequences. The cytochrome *b* gene had been useful in resolving phylogenetic patterns among various ruminant artiodactyls during the last 20–30 million years (Hassanin and Douzery 1999).

Cytochrome *b* is a protein-coding gene that has been used for phylogenetic inferences mostly because it is a mitochondrial gene, and it evolves at different rate among different lineages (Castresana 2001). This gene has definitive alignment because of the high conservation of sequence length across mammals which makes positional homology easy to discover (Irwin *et al.* 1991), and there is great amount of data available about its structure and function (Howell 1989; Degli Espoti *et al.* 1993) in a phylogenetic context (Griffiths 1997).

Hassanin and Douzerry (1999) sequenced mtDNA cytochrome *b* of the following six *Tragelaphus* species: *T. euryceros, T. spekei, T. scriptus, T. oryx, T. strepsiceros and T. imberbis. Tragelaphus imberbis* was found to be basal-most by Hassanin and Douzery (1999). However, when Matthee and Robinson (1999) conducted their study using the same mtDNA marker, they added data for *T. derbianus and T. angasi*, and showed that the former was sister to *T. oryx* and the latter a basal species together with *T. imberbis.*

Although neither study represented all *Tragelaphus* species, the mtDNA studies by Hassanin and Douzery (1999) and Matthee and Robinson (1999) differed greatly from what was expected due to morphology. Firstly, mtDNA strongly suggested that the tribe Tragelaphini has one genus: *Tragelaphus*. In both the studies distance, parsimony and maximum likelihood phylogenetic reconstructions supported a recent sister relationship between *T. euryceros* and *T. spekei* with *T. scriptus* next most closely related, supporting Tijskens (1968); and Wilson and Reeder (2005) suggestion to include the *Boocerus* genus within *Tragelaphus* due to the ability of *T. euryceros* and *T. spekei* to hybridize in captivity. The mtDNA divergence estimates revealed that the two species (*T. spekei* and *T. euryceros*) that differed most morphologically, were in fact the two most closely related *Tragelaphus* taxa (Matthee and Robison 1999).

Additionally, *T. spekei* and *T. angasi*, which were suggested sister taxa based on similar cranial features, were not monophyletic in the mtDNA phylogeny. Similarly, the two kudu-like phenotypes *T. imberbis* and *T. strepciceros* which show marked morphological and dietary similarities, were paraphyletic (Hassanin and Douzery 1999; Matthee and Robinson 1999). Finally, Hassanin and Douzery 1999 found that their sequence for the common eland "*Taurotragus oryx*" clustered within the diversity of other *Tragelaphus* species, suggesting that it should be better included within *Tragelaphus* following previous results by Georgisdis *et al.* (1990) and Gatesy *et al.* (1997). Based on the mtDNA, the arid environment adapted species (*T. imberbis, T. angasi, T. strepciceros, T. oryx, and T. derbianus*) were the most basal and the species adapted to moist forest habitats (*T. euryceros*, *T. spekei*, and *T. scriptus*) were derived (Matthee and Robinson, 1999).Later studies added nuclear DNA data to help resolve the mtDNA-morphological topological conflict. A larger

molecular data set was used with expectation to investigate the definite pattern and timing of when speciation occurred within the group, as well as give insights on what could be driving speciation in these African antelopes (Willows-Munro *et al.* 2005). Willows-Munro *et al.* (2005) suggested that the incongruency of previous morphological and mtDNA studies was caused by shared ancestral characters being fixed in only a portion of the descendants of polymorphic ancestor, comparing with the contradicting nuclear DNA results found in the study, which is portrayed below.

Willows-Munro *et al.* (2005) sequenced segments of four nuclear genes of all nine extant species, that is all species mentioned in the two studies above with the addition of *T. buxtoni*. Selected nuclear markers were previously used to illustrate evolutionary relationships among the cetartiodactyla (Matthee *et al.* 2001), Leporidae (Matthee *et al.* 2004) and Bovidae (Matthee and Davis 2001). The idea of using a greater number of genes for phylogenetic reconstruction was to increase its resolution (Willows-Munro *et al.* 2005) and nuclear DNA has lower levels of homoplasy and is possible to get information on rapid radiation events.

Nuclear data in this study was studied together with three mitochondrial genes (cytochrome *b*, 12S rRNA and 16S rRNA) of each species. The supermatrix topology (mtDNA and nuclear data) in the study of Willows-Munro *et al.* (2005) showed the monophyly of closed forest species (added species data *T. buxtoni* and *T. euryceros, T. spekei, T. scriptus*) which Matthee and Robinson (1999) refer to as the derived moist habitat adapted species. *Tragelaphus imberbis* and *T. angasi* were the most basal as in previous mitochondrial studies. These results also disagreed with the morphological data, thus Willows-Munro *et al.* (2005) saw it reasonable to suggest that morphological similarities i.e cranial similarities, presence of horns in both sexes, marked morphological and dietary similarities are due to convergence and not common ancestry. *Tragelaphus* genus is known to have gone through rapid evolutionary radiation (Willows-Munro *et al.* 2005), and it has been revealed that different groups of animals experiencing several evolutionary radiations can develop the same traits at the same time (Hassanin and Douzery 2003).

The most basal species (*T. imberbis* and *T. angasi*) both occur in dense dry woodlands and are likely to have been isolated because of fragmentation of these habitats. Between 14.08 and 10.89 Mya there was an expansion of savanna grassland which promoted the evolution of open savanna specialists like *T. strepciceros* as well as the two eland species (*T. derbanius* and *T. oryx*). Approximately 7.31 Mya there was diversification of tropical/wet habitat adapted species which are the closed forest group (*T. buxtoni, T. euryceros, T. spekei* and *T. scriptus*), this was due to

more fragmentation of forest habitats. This raised a concern with the placement of *T. strepciceros* which is always changing position on a tree based on markers used. Willows-Munro *et al.* (2005) proposed that paleoclimatic shifts, vicariance, and the sudden availability of open niches on African continent resulted in the rapid diversification of tragelaphids.

In a continent-wide phylogeographic study of the most widespread *Tragelaphus* species, the bushbuck (*T. scriptus*), Moodley and Bruford (2007) described the two highly divergent bushbuck mtDNA lineages which they called the *Scriptus* and *Sylvaticus* super lineages, each with very high genetic diversity. Moodley *et al.* (2009) used mtDNA sequences to deduce the evolutionary relationships between *Scriptus* and *Sylvaticus* by comparing them with those of other tragelaphines. They used *Tragelaphus* mitochondrial cytochrome *b* dataset from GenBank, this included all data from Willows-Munro *et al.* (2005), Hassanin and Douzerry (1999) and Matthee and Robinson (1999) and sequenced the same mtDNA loci for *Scriptus* and *Sylvaticus* super lineages identified by Moodley and Bruford (2007).

Moodley *et al.* (2009) showed that the two bushbuck super lineages were not monophyletic but instead polyphyletic. The *Scriptus* super lineage was closely related to *T. angasi* and *Sylvaticus* was sister to the *T. euryceros* and *T. spekei* clade. Willows-Munro *et al.* (2005) cytochrome *b* data did not show this previously because their bushbuck data set only included the S*ylvaticus* super lineage. Moodley *et al.* (2009) further suggested that bushbuck may comprise two species (*T. scriptus and T. sylvaticus*). Hassanin *et al.* (2012) used mtDNA genomes to show the same polyphyly of bushbuck super lineages, also suggesting they were separate species. The mtDNA polyphyly of the bushbuck was again demonstrated with full mitochondrial genomes when Bibi (2013) implemented fossil calibrations to account for changing rate of molecular evolution when reconstructing the evolutionary history of Ruminantia. For the *Tragelaphus* clade, Bibi (2013) used three fossil calibration points with the oldest being for crown *Tragelaphus* at 5.7 Ma.

The dates of Bibi (2013) differ from those of Willows-Munro (2005) described above. This is due to the fact that Bibi (2013) used many (16) fossil calibration points considering morphological and ecological characteristics of clades, assuming that crown *Tragelaphus* was 5.7 ma, whereas Willows-Munro (2005) used a specified mutation rate and a Bayesian molecular clock, that suggested the divergence of basal *T. angasi* and *T. imberbis* from other Tragelaphus to have occurred 10.8 Mya. Bibi 2013 suggests that the higher diverse the species are, the higher the possibility of their molecular rate of evolution, thus many fossil calibration points must be employed.

Furthermore, when Bibi (2013) also suggested that the bushbuck might be two separate species, he cautioned that confirmation from the nuclear genome was still lacking. The mtDNA polyphyly could have resulted from incomplete lineage sorting (Bibi 2013). He also suggested that mitochondrial lineage introgression between geographically close non-sister taxa could also be a possible explanation (Bibi 2013).

Two recent studies, Hassanin *et al.* (2018) and Rakotoarivelo *et al.* (2019), finally presented a complete set data of *Tragelaphus* species (including both bushbuck lineages) at mitochondrial and nuclear levels. Both papers revealed that the two bushbuck were monophyletic at nuclear DNA. Since genetic drift must sort mtDNA into monophyly four times faster than nuclear DNA, it is impossible for nuclear DNA to show monophyly and mtDNA from the same samples to show ILS. Thus, for *Scriptus* bushbuck to be sister with *T. angasi* at the mitochondrial level, there must have been gene flow between an early *Scriptus* bushbuck and a now extinct species that was closely related to *T. angasi.*

Hassanin *et al.* (2018) showed a difference in karyotype number between *Scriptus* and *Sylvaticus* bushbuck and suggested them to nevertheless comprise two species, despite their nuclear monophyly. However, Rakotoarivelo *et al.* (2019) preferred to keep the two bushbuck as a single species noting that the difference in chromosome number can be explained by centric fissions or fusions which would not necessarily lead to reproductive isolation of the two bushbuck. This is evidenced by the fact that both come into regular secondary contact along the African rift (white line bushbuck distribution figure).

The complete set of *Tragelaphus* data raised another concern in the placement of *T. strepciceros* species on the phylogenetic tree. Rakotoarivelo *et al.* (2019) showed mtDNA tree placing *T. strepciceros* sister to *T.buxtoni-T. spekei-T. euryceros-T. s. sylvaticus* clade. On the nuclear gene level Rakotoarivelo *et al.* (2019) still revealed the sister relationship of *T.strepciceros* with monophyletic bushbucks and *T.buxtoni-T. spekei-T. euryceros* clade. Hassanin *et al.* (2018) on the nuclear gene tree placed *T. strepciceros* as sister with *T. derbanius-T. oryx* clade and stated that they are not certain with its placement. This has been an issue since the morphological classification of data, Matthee and Robinson (1999) showed it as well as Willows-Munro *et al.* (2005).

All the other studies so far point to a complex evolutionary history for *Tragelaphus* that involves at least one known instance of gene flow (bushbuck hybridization). It is plausible that there was more gene flow than we are aware of because the rapid radiation. This could explain why, aside

from bushbuck, the placements of other species on the tree, such as *T. strepciceros*, vary depending on the study and markers employed.

1.5 The benefit of whole genomes

As reviewed in the previous section, traditional markers provided interesting but limited information with mitochondrial DNA only portraying the maternal history and nuclear DNA gene fragments providing little genome-wide information.

There is development of high-throughput sequencing (Next-Generation Sequencing) producing whole genomes in a large quantity (Goodwin *et al.* 2016). Some of the examples of technologies being used now are short-read sequencing (e.g illumina) and long-read sequencing (e.g PacBio) (McCombie *et al.* 2019). This development is aiding the advancement of bioinformatics tools to handle large genomic data being generated for analyses (Purcell *et al.* 2007). Even when a single genome is sequenced to low coverage, there are bioinformatic tools designed to handle it like the software Analysis of Next Generation Sequence Data (ANGSD; Korneliussen *et al.* 2014). A single sequenced genome also represents a significant proportion of the ancestral genomes of the population. The evolutionary history of a species, both divergence due to genetic drift and introgression due to gene flow, is faithfully recorded in the genome (Frantz *et al*. 2013) and there is a growing number of analyses that can infer which parts of the genome have been affected by which force.

Assessment of whole genomes will reveal if the *Tragelaphus* radiation occurred with a lot of gene flow, which could be linked to the presence of convergent phenotypes. Gene flow could have aided convergent evolution by the process of adaptive introgression, which is where one species evolves an advantageous phenotype, but it is able to transmit this phenotype to another non-sister species through gene flow. So, the second species did not evolve the phenotype by itself but can enjoy the benefits of the phenotype because it obtained it through gene flow as seen in Darwin's finches (Lamichhaney *et al.* 2015) explained above.

Because of the possibility of extensive gene flow during the *Tragelaphus* radiation, the true evolutionary lines of descent cannot be determined using traditional markers because they do not have enough polymorphism to look far back in time with a high resolution. And they also do not contain enough polymorphism to detect the possible events of gene flow that may or may not

have occurred. It is thus important to study whole genomes which contains all genes, regulatory sequences, and noncoding information of an organism (genetic materials contained within and shared between organisms) to fully reconstruct the evolutionary history of *Tragelaphus*.

1.6 Problem statement and research questions

Despite mitochondrial and nuclear gene studies being published, there remain several unanswered questions about the evolutionary history of the genus *Tragelaphus*. In this dissertation, I will attempt to answer the following questions:

1. What are the true lines of evolutionary descent relating the species within the genus *Tragelaphus*?
2. How much gene flow has occurred throughout the *Tragelaphus* radiation?
3. Can the evolutionary history be recovered, even if gene flow was considerable?

1.7 Aims and objectives

- The main aim of this dissertation is to reconstruct the evolutionary history of the genus *Tragelaphus* for the first time using whole genomes.
- The first objective to accomplish this aim is the reconstruction of the *Tragelaphus* species tree, which is assumed to have occurred due to genetic drift.
- The second objective is to quantify levels of genomic diversity in each *Tragelaphus* species.
- The third objective is to use the topology of the species tree to investigate gene flow events during the *Tragelaphus* radiation.

2.Methods

2.1 Samples

The study was conducted on ten *Tragelaphus* genomes, whose samples were acquired from Zoos, museum collections and taxidermists (Table 1). *Bos taurus* (cow) which is available on the Bovine Genome Database (genome available http://bovinegenome.org/?q=node/61) (Shamimuzzaman 2020) and accessible on GenBank with accession number GCA_000003055.5, was used as a reference genome and as an outgroup for downstream analyses. The cow belongs to the bovines which are sister to the tragelaphines. This cow genome was used as a reference for mapping since there was no available reference genome of the *Tragelaphus* species. Galla *et al.* (2018) mentions that a reference genome from related species can be used for discovery of SNPs, in our case the cow is a distant related species to *Tragelaphus*. The cow's genome was sequenced to chromosome level and had a total sequence length of 2 670 123 310 bp, that is 30 chromosomes (2 660 906 405 bp) and 6336 bp of unassigned scaffolds.

Table 1: Table showing the samples collected for the study, their location and type of collected sample.

Common name	**Species**	**Type of sample collected**	**Total amount of sample extract (ng)**	**Source**	**Location**
Imbabala	*T. s. sylvaticus*	Dried skin	8131	Taxidermist, South Africa	Eastern Cape, South Africa
Kewel	*T. s. scriptus*	Skin DMSO	10944	Hans Siegismund, Copenhagen	Assin Manso, Central Region of Ghana

Common name	Species	Type of sample collected	Total amount of sample extract (ng)	Source	Location
Greater Kudu	*T. strepciceros*	Blood	1648	Gabrielle Stalder, FIWI, Austria	Zoo
Nyala	*T. angasi*	Liver	2561	Gabrielle Stalder, FIWI, Austria	Zoo
Common Eland	*T. oryx*	Liver	2301	Gabrielle Stalder, FIWI, Austria	Zoo
Giant Eland	*T. derbanius*	Skin EtOH	5939	Chinko Project	Central African Republic
Bongo	*T. euryceros*	Ear notches	7337	Paul O'Donoghue, United Kingdom	Port Lympne Zoo
Sitatunga	*T. spekei*	Skin (ear notch)	11178	Jan Robovsky, Czech Republic	Prague Zoo
Mountain Nyala	*T. buxtoni*	Dried skin	12240	Powell Cotton Museum, United Kingdom	Ethiopia
Lesser Kudu	*T. imberbis*	Blood	4080	Gabrielle Stalder, FIWI, Austria	Zoo

2.2 Genome sequencing

DNA extraction, library prep and sequencing platform was performed at the University of Potsdam as part of a PhD project by Ulrike Taron. However, I will describe the molecular steps here for clarity. First the DNA was extracted from collected samples (Table 1) using DNA extraction protocol known as batch method (Rohland *et al.* 2010) in a laboratory. The collected samples were digested in an extraction buffer and incubated overnight. The supernatant was then added to the binding buffer where DNA molecules bind to the silica surface then the mixture was incubated under agitation for 3hours. The DNA was eluted with Tri-EDTA buffer from silica surface which was dried first.

The extracts were then diluted (concentration: 15-27 ng/µl, Qubit) and sheared. Then the double stranded, double indexed NGS libraries were prepared. Double indexing was to prevent sequencing what is not the sample's DNA (Henneberger *et al.* 2019) and to optimize DNA yield of the collected samples. Double strand library preparation was according to Henneberger *et al.* (2019), where there was an extension of blunt-end adapters with fault on template DNA, these adapters were extended to complete adapters during indexing pylymerase chain reaction (PCR). TapeStation (DNA D1000 kit, Agilent) and Qubit were then employed to check the successful library preparation and re-amplified to improve their quality. All libraries were filtered for fragment size, and the size selected was of 541 bp average length, these are libraries with trimmed reads for mapping. Sequencing was executed on an illumina NextSeq 500 system performing paired-end sequencing with 150 cycles resulting in 300 analysed cycles.

2.3 Bioinformatics

2.3.1. Mapping

The cow reference genome was downloaded from the bovine genome database site (available http://bovinegenome.org/?q=node/61). Then mitochondrial sequences (which gives maternal known evolution and not needed in this study) were removed from the assembly using the grep command by getting the line numbers where the mitochondria start and ends. The modified reference genome without mitochondria was then indexed using SAMtools (Li *et al.* 2009). All

short reads of all samples had the illumina adapter sequences removed using Cutadapt v1.8.1 (Martin 2011). Trimmed reads were then mapped to the cow reference genome without the mitochondria using Burrow-willer alignment (BWA-MEM v0.7.15) with default parameters (Li and Durbin 2009). The mapped reads were sorted by coordinates and the duplicates were removed using SAMtools (Li *et al.* 2009). At that point the PhD student stopped her project for personal reasons. The mapped reads (samples) were then sent as zipped Binary Alignment Map (BAM) files to the University of Venda so that I may conduct the analyses for my Masters project.

2.3.2 Quality and removal of sex chromosomes

The first step I took was to download the BAM files then used Bcftools (Li 2011) to unzip the files. During mapping there were unassigned data which I removed with Bcftools and remained with only the assigned chromosomes which I needed to analyze. That was followed by the removal of sex chromosomes from the unzipped BAM files. Because this data contains short read sequences mapped to a reference genome, it makes it possible to filter the data (Benestan *et al*. 2016).

The sex chromosomes have smaller effective population size compared to the autosomes therefore this has an impact of lower diversity within the genome (Ellegren 2009). The Y chromosome is only paternal, with X chromosome passed by both male and female parents to the offspring. Sex chromosomes were therefore removed from each sample using the software suite Analysis of Next Generation Sequencing Data (ANGSD) by specifying a file with sex chromosome regions. Then eleven genomes (BAM files) were listed in one text file (BAMLIST) using ANGSD and with the same software the following flags were applied -doQsdist, -doDepth and -doCounts. These flags are for outputting the base qualities, calculating read depth and computing summery statistics respectively.

2.3.3 Converting BAM files to a multi-sample Variant Call Format (VCF) file

Converting BAM files was done with the aim to call variants in each sample (BAM) and do the downstream analyses on variants. VCF format represents mutations (SNPs), insertions and deletions (INDELS) against a reference genome (Garrison 2021). Bcftools (Li 2011) was used to convert BAM files to VCF. The first step was to use bcftools mpileup command on each BAM file

individually to extract genotype likelihoods for each sequenced base. The genotype likelihoods were indexed using Bcftools and then the variants among them were called with the bcftools call command, using the multiallelic caller which is more sensitive, and it was specified that the output should be variants only of the Autosomes. This was done for all eleven genomes and after this I had eleven VCF files, one for each genome. At the end, all eleven genomes were merged with bcftools merge command to create one multisample VCF file.

2.3.4 Filtering multisample VCF for downstream analyses

To make sure that there were no indels on the multisample VCF file, they were removed with the flag "—remove-indels" on bcftools (Li 2011) and the output was a multisample VCF without indels. Indels were removed as they are the insertion or deletion of small nucleotides which alters the length of genome sequence and may influence some analysis. The multisample VCF without indels was then filtered further using Plink (Purcell *et al.* 2007). Whole genomes VCF files are very big and require larger computer memory for analyses (Akgun and Demirci 2017), so they were filtered (or down sampled) to reduce the size of the data, with cautions against over filtering. The filter steps on plink were to filter for missing data, where the sites with missing data were filtered by setting the genotyping rate of 90%, where only 10% missing data was allowed for any given SNP. Then I filtered for minor allele frequency (maf ≥0.05) this meant that only single-nucleotide polymorphism (SNPs) with maf greater or equal to 0.05 were kept. Filtering for missing data and maf was crucial because there could have been errors during sequencing (Carson *et al.* 2014).

2.4 Evolutionary genomic analyses

In this study, different methods have been employed to determine the structure of the dataset, with the aim to reconstruct phylogenetic species tree to understand the relationship of *Tragelaphus*. A biological model free Principal component analysis (PCA) which is a statistical method used to determine structure in the pattern of dataset (McVean 2009) was employed. This method recognizes the main axes (principal components) of data variation and projects the samples onto these components graphically. The components are determined so that, among all available components, the projection of samples down the first principal component explains the

most variation in the data (McVean 2009). PCA analysis was done to look at the structure of the dataset without taking into consideration the biological processes that influence the sequences. Although PCA is model free, it is good in identifying structure created by a variety of processes, it is computationally fast, and it can visually arrange or separate samples.

Other methods employed in this study are Identity by sequence (IBS) and maximum likelihood (ML), the two methods employ biological models. These methods take biological parameters like mutation rate into consideration when inferring a tree (Felsenstein 1981) and that minimizes the amount of time to infer nonoptimal trees (Holder and Lewis 2003). In IBS, the genetic distances are calculated between given samples and compute a matrix out of them then a tree is computed out of the distance matrix. In ML we work with the likelihood of a biological process occurring. The two methods were employed because the analysis softwares that were going to be used incorporate several biological models and use the one that best fit the data. Of the two methods, ML employs bootstrapping and not IBS.

2.4.1 Principal Component Analysis (PCA)

The PCA is model free method for assessing the structure of data, meaning it determines the relationship of the data by considering them as variables. The filtered multisample VCF was used, and the cow genome was removed using vcftools (Danecel *et al.* 2011). It was removed because its relationship with the *Tragelaphus* is known and does not provide any information that is important for a PCA. If included, the first principal component will show the large differentiation of the cow from the rest of the *Tragelaphus* species, which is already known and would represent most variance of the dataset. After the cow was removed, plink was used to perform the PCA with the flag “-pca” that calculated a variance-standardized relationship matrix and extracted the top 10 principal components (Appendix 1), the flag "--allow-extra-chr” was also used so that plink can recognize the naming of chromosomes in the VCF file.

2.4.2 Phylogenomic reconstruction

2.4.2.1 Identity by sequence (IBS) neighbor-joining species tree

The BAM files which were listed in a text file named BAMLIST (mentioned in 2.3.2) were used for the first method to reconstruct a phylogeny that will draw closer to the species tree. This distance-based reconstruction method was implemented with the *Bos taurus* as an outgroup to polarize relationships among the *Tragelaphus* ingroup. This distance first requires an estimation of the site frequency spectrum using the reference genome information to infer two alleles for each position. This method was implemented in ANGSD, which then calculated the probability of each site of being IBS and then computed an IBS distance matrix relating all eleven genomes in the BAMLIST. The distance matrix was then converted into a nexus file and reconciled into a phylogenetic tree using the neighbor-joining algorithm (Saito and Nei 1987) and then visualized using Figtree v1.4.3 (Rambaut 2010).

2.4.2.2 Maximum likelihood species tree

I also used the maximum likelihood (Felsenstein 1981) approach to reconstruct the *Tragelaphus* species tree, since the IBS above does not provide bootstrap (Hoang 2018). This analysis was carried out in IQTREE (Nguyen *et al.* 2014; Minh *et al*. 2020), which was developed specifically to handle big high-throughput sequencing data for phylogenomic inference. IQTREE uses better computational methods than previously existing approaches, testing a number (26 in this study) of substitution models to determine the best fit model (Kalyaanamoorthy *et al.* 2017) before tree reconstruction. This software also employs the ultrafast bootstrap procedure (Hoang *et al.* 2018), which can quickly compute support values for each node on the tree.

Before, running the analysis, however, the genome alignment (VCF) file was further filtered for linked sites (Linkage disequilibrium; LD). LD is a nonrandom linkage of alleles at two or more loci and it tells us how forces of evolution (natural selection, genetic drift, recombination and mutation) have structured the genome (Slatkin 2008). Since it was not clear how much linkage could affect the phylogeny, I filtered the multisample VCF according to the following series of linkage thresholds (r^2): 0, 0.2, 0.5, 0.8 and 0.99 in sliding windows, and recoded the resulting VCF output so that the original multisample VCF file was unchanged.

During all these r^2 filter runs the window sizes were kept to 50 kilobases (kb) with a step size of 10 kb. The greater the linkage threshold, the fewer linked sites were allowed. The number of retained variants were recorded after each r^2 filter. Each of five r^2 filtered VCFs were then converted into Phylip format using a Python script vcf2phylip.py and used as an input in IQTREE to reconstruct five different species trees using maximum likelihood, specifying 1000 ultrafast bootstraps iterations for each run. The Modelfinder (Kalyaanamoorthy *et al.* 2017) automatically determined the best substitution model for each run, which was consistently the General time reversible (GTR+F) and models each base independently of each other. In this model, F represents the rate of a base changing to another, which is computed individually, where the state frequencies of each base are computed from the provided alignments.

.

2.4.3 Heterozygosity and Nucleotide diversity (P_i)

Genome wide heterozygosity, which is the state of a diploid organism containing two different alleles at a homologous locus, was calculated using ANGSD (Korneliussen *et al*. 2014). The software employs genotype likelihoods. Using ANGSD software, the distribution of alleles for each genome (with autosomes only) is called the site frequency spectrum (SFS), was calculated using a sliding window 2000 kb in size. The output results were homozygous and heterozygous positions for each window. To get an estimate of genome-wide heterozygosity-estimate, all window estimates were averaged, and the number of heterozygous positions was divided by the sum of homozygous and heterozygous positions.

To calculate the nucleotide diversity (P_i) a multisample VCF file was used on Python scripts created by Simon Martin (http://github/simonmartin/genomics_general). First the multisample VCF was processed with the script parseVCF.py into geno format, that is a file of the genotypes, then the popgenWindows.py (http://github/simonmartin/genomics_general) script was used to calculate the P_i in window sizes of 50 kb and step size 25 kb. The script computed Pi for each sample on a multisample VCF. The nucleotide diversity is the pairwise average number of nucleotide differences between two DNA sequences in all possible pairs in the sample population (multiple VCF).

2.4.4 Introgression (gene flow)

With the previous uncertainty placement of species on the species tree, it was important that introgression be assessed to identify potential non-sister secondary contact. Introgression was calculated using four-taxon method known as ABBA/BABA (Durand *et al.* 2011) implemented in the software Dsuite (Malinsky *et al.* 2022). The fourth position in this method is always the outgroup since introgression is assessed within the ingroup taxa (positions 1-3). In this method Introgression detection is based on a statistically significant imbalance in the number of sampled discordant biallelic site patterns "ABBA" and "BABA." (Pease and Hahn 2015). Under a model of genetic drift there should be an equal number of ABBA and BABA sites. More ABBA than BABA means introgression between taxon in 2nd and 3rd position, more BABA than ABBA means introgression between taxon in 1st and 3rd position.

This software will calculate D-statistics between every combination of taxa in positions 1-3 of the standard ABBA-BABA (Durand *et al.* 2011) test. However, many of these combinations will reflect common ancestry, and not gene flow and to control for this, I specified the maximum likelihood species tree topology so that Dsuite only returned those tests that tested for gene flow, contingent on the topology of the species tree. To do this, the maximum likelihood species tree was converted into Newick format using Figtree and then used as basis for the Dsuite geneflow analysis on the multisample VCF file, also having specified the cow as the outgroup (position 4) for all ABBA-BABA tests. One hundred and twenty (120) different trios were thus calculated, and the results were recorded. To further identify excess sharing of derived alleles, correlated f4-ratio values produced by Dsuite 'Dtrios' were interpreted by an F-branch (Malinsky *et al.* 2021) method implemented in Dsuite. The F-branch method was performed having provided the *Tragelaphus* species tree with the option '–tree' to produce a gene flow heatmap. Dsuite outputs a lot of difficult-to-interpret strongly correlated f values when f4-admixture ratio statistics are generated for multiple subsets of samples within the same phylogeny (Malinsky *et al.* 2018), thus f-branch creates an f-branch metric given a specific tree with hypothesized relationships to disentangle correlated f4-ratio values and assign gene flow evidence to specific, presumably internal, branches on phylogeny (Malinsky *et al.* 2021).

2.4.5 Directional introgression

DFOIL (Pease and Hahn 2015) which detects introgression in a five-taxon phylogeny was used to detect the direction of introgressions detected by *DSUITE* and whether introgression occurred between ancestral lineages. First the fasta consensus files were created for each individual genome using ANGSD. The five fasta files of interest according to the species tree were merged with “cat” command into five taxon file, with the outgroup always fixed at fifth position (P5). When merging fasta files according to the species tree, P1 and P2 must be monophyletic pair of taxa, P3 and P4 must be monophyletic pair of taxa with the divergence of P3 and P4 having occurred before the divergence of P1 and P2. Using Mvftools (Pease and Rosenzweig 2015) the merged fasta files were then converted into mvf file with the command “ConvertFasta2MVF”. The mvf file was then converted into three Dfoil’s input files with the command “CalcPatternCount” on Mvftools. Dfoil python script (dfoil.py) was then run on the correctly merged mvf file. Dfoil then calculated four independent D statistics in sliding window (500kb window size) fashion which were then combined before inferences were made.

3.Results

3.1 Data filtering

Table 2 below shows the mapped reads and depth coverage of the autosomes BAM files assessed by ANGSD for each *Tragelaphus* species ordered according to degree of relatedness from results of structure (PCA, IBS and Maximum likelihood) from the most basal *T. imberbis.* The coverage of the genomes range between 7 X and 13 X.

Table 2. A table showing mapped reads and depth coverage of autosomes BAM file.

Autosomal BAM Quality check		
Sample	Mapped reads	BAM depth (X)
B. taurus	167722177	10
T. imberbis	213871511	13
T. angasi	147639863	9
T. strepciceros	145016702	9
T. oryx	127848578	7
T. derbanius	153114521	9
T. spekei	173910489	10
T. euryceros	140088040	8
T. buxtoni	190353150	12
T. s. scriptus	135516814	8
T. s. sylvaticus	127731885	8

Below is table 3 which depicts the number of variants remaining after filtering the multisample VCF from indels removal, minor allele frequency filtering to filtering for missing data. All these filtering reduce the data to a more manageable size which decreases the computer memory needed for downstream analyses.

Table 3: A table showing number of variants remaining after filtering out indels, minor alleles and missing data.

Removal of indels, minor alleles and missing data by filtering	
Filter	#SNP Variants
Raw multisample VCF (no filter)	218057242
Indels removed	208044306
Maf filtering	181789590
Missing data filtering	141315256

3.2 Genomic structure

Principal component analysis plot (Figure 1) below is depicting the first five components which accounts for about 70% (Appendix 1) of total data variation. The analysis depicts four major clusters. The 1st component (PC1) defines the separation of the two basal-most species *T. imberbis* and *T. angasi* from the other *Tragelaphus.* The 2nd component (PC2) separates the *T. angasi* from the rest of the *Tragelaphus*. In the 3rd component (PC3) there is a separation of *Tragelaphus*, grouping them into a dry savanna cluster (*T. strepciceros, T. oryx* and *T. derbanius*) with *T. strepciceros* closer to *T. oryx* and two closed forest groups comprising firstly *T. spekei, T. buxtoni* and *T. euryceros* and secondly the two bushbuck (*T. s. sylvaticus* and *T. s. scriptus*). Interestingly in 4th component (PC4) *T. strepciceros* is seen being separated from the rest of the *Tragelaphus*. 5th component depicts what 3rd component depicted, the grouping of dry savanna species and closed forest species with two bushbuck clustered together but, unlike in PC3 *T. strepciceros* is placed away from *T. oryx* in PC5.

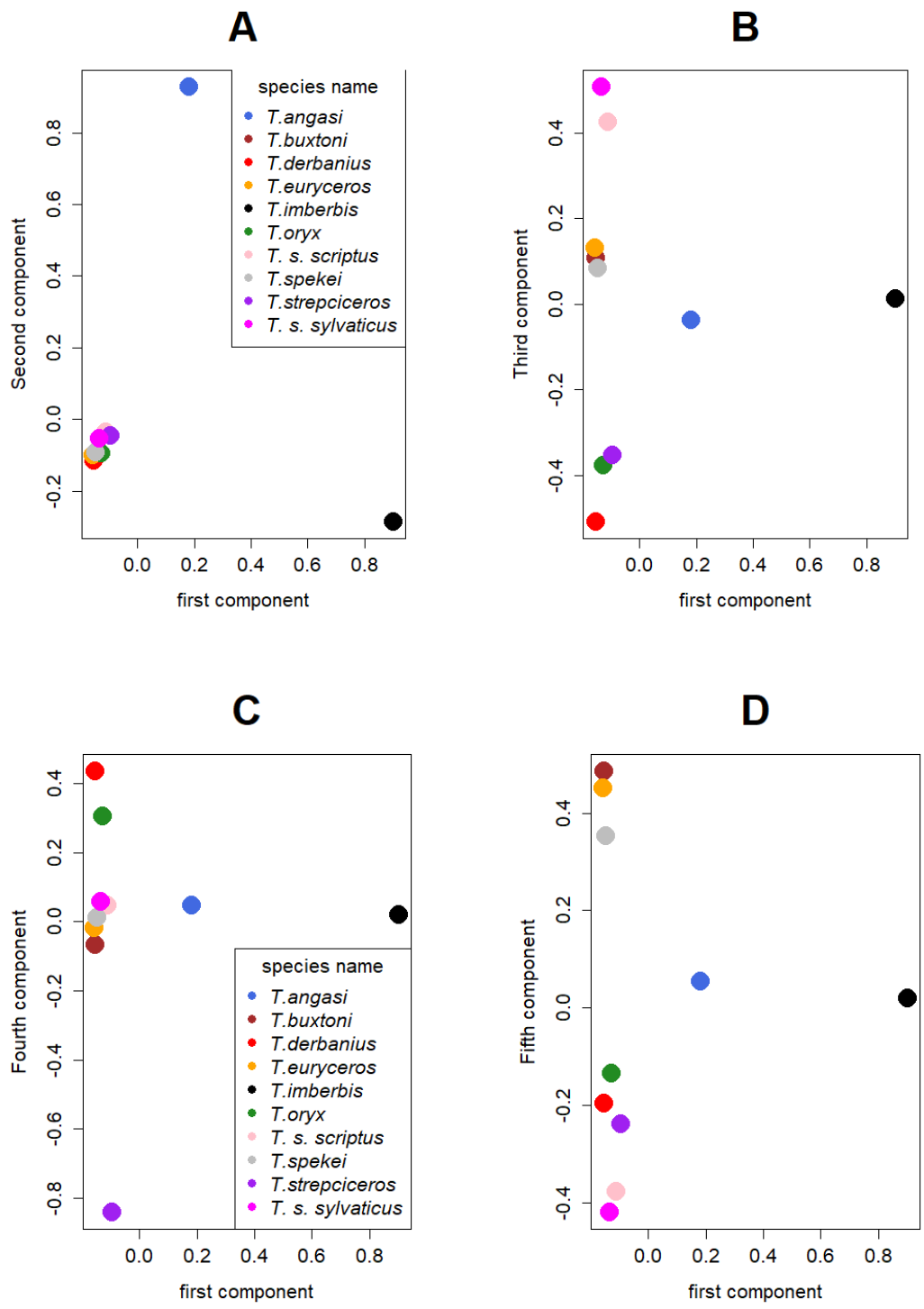


Figure 1. PCA plot of *Tragelaphus* genomic data. (A) 1st and 2nd components. (B) 1st and 3rd components. (C) 1st and 4th components. (D) 1st and 5th components. PCA which takes no biological model into consideration was conducted on the genomic dataset of 141315256 SNPs in PLINK having corrected for the missing data.

The 1st model-based method results of reconstructing a species tree is depicted below (Figure 2). The genetic distance-based method revealed what is depicted on 1st and 2nd components of the PCA, these are the different clades of species. On the Neighbor-joining tree below, the most significant clades being: (1) The basal *T. angasi* and *T. imberbis* clade and (2) the clade of the rest of the *Tragelaphus.* Where on the rest of the *Tragelaphus* clade, they were further placed into two distinct clades: *T. strepciceros-T. derbanius-T. oryx* clade and *T. spekei-T. euryceros-T. buxtoni-T. s. scriptus-T. s. sylvaticus* clade.

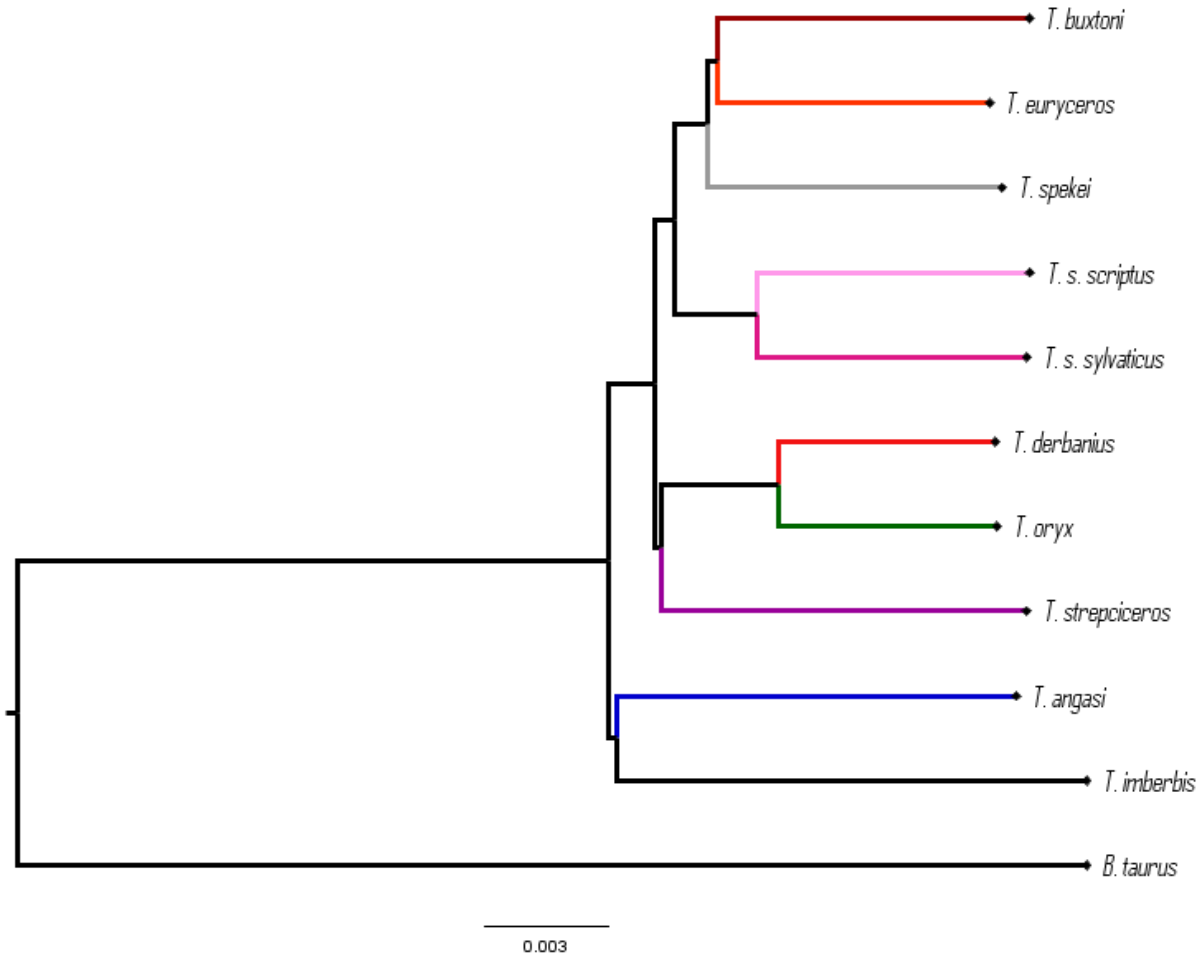


Figure 2. A neighbor Joining, identity by sequence (IBS) species tree generated by ANGSD. Bootstrap resampling was not an option for this analysis. The tree is based on whole genomes BAM files (218057242 SNPs) for all *Tragelaphus* and the cow (11 individuals total). The tree was plotted in Figtree.v1.4.3 and the colors correspond with species colours in PCA as It gives the relationship among species as depicted in PCA plot Figure 1.

The maximum likelihood species tree (Figure 3) below was reconstructed with ultrafast bootstrap and nodes had strong support of 100%. The ML tree depicted the similar relationship of species as the IBS tree. On ML tree, it was showed that *T. angasi* and *T. imberbis* are the most basal and the rest form a monophyletic clade. Furthermore, *T. strepciceros-T. derbanius-T. oryx* of the dry savanna form their own clade, with *T. strepciceros* being a sister taxa to *T. derbanius-T. oryx* clade. The other clade separated the bushbuck into their own monophyletic clade and the *T. spekei-T. euryceros-T. buxtoni* clade with *T. spekei* as a sister to *T. euryceros-T. buxtoni* clade.

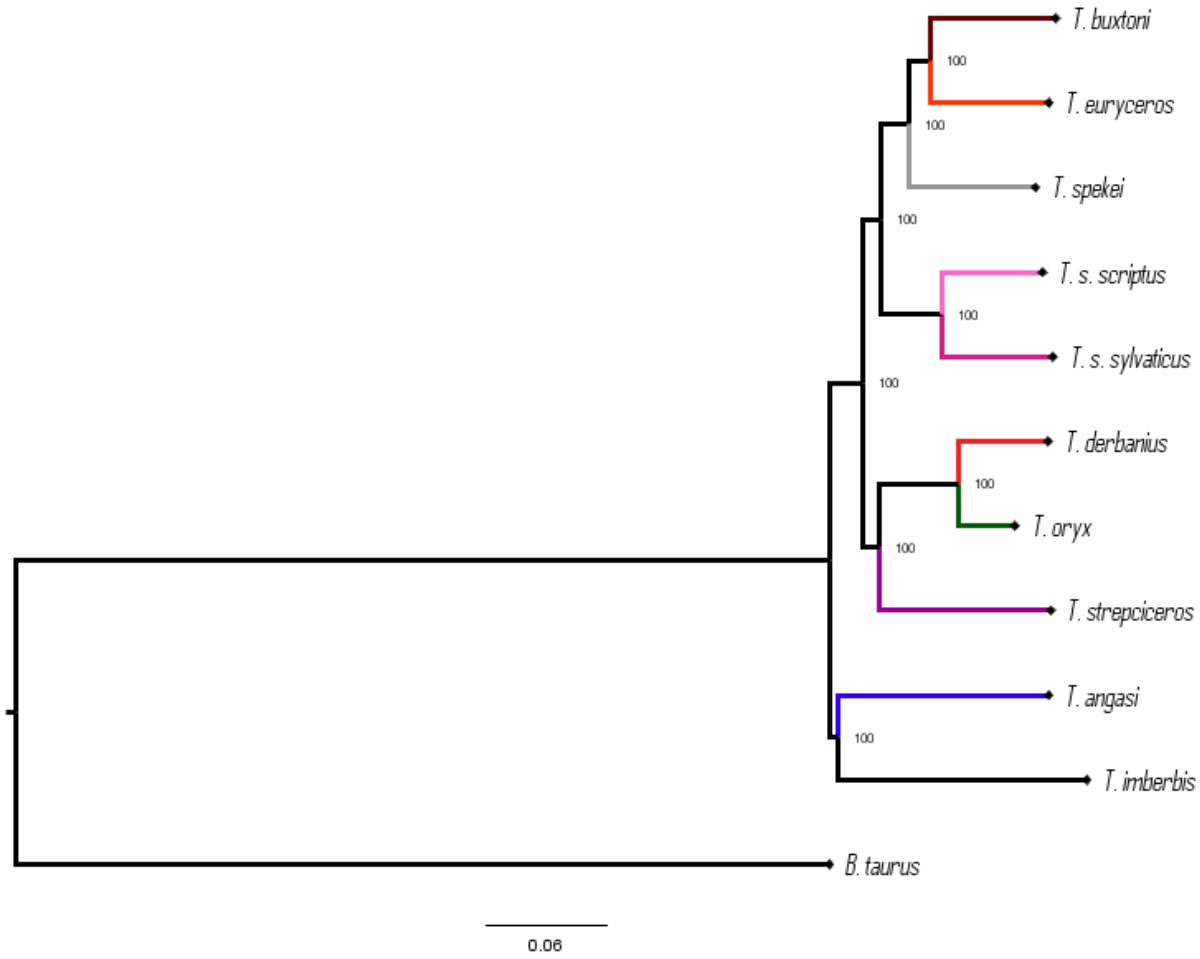


Figure 3. A maximum likelihood species tree generated from filtered multisample VCF (11 sequences) of 141315256 SNPs by IQtree which implemented the GTR+F (general time reversible) model for unequal base frequencies and plotted on Figtreev1.4.3 showing the nodes bootstrap support. The colors correspond with PCA (Figure 1) and IBS tree (Figure 2).

The filtered multisample VCF (from Table 3) was further filtered with different linkage disequilibrium (LD) pruning number (r^2) in PLINK and the SNP variants remaining after filtering with each r^2 were recorded and depicted on table 4 below. It is evident that the higher the LD number, the more variants are kept. For each r^2 pruned VCF output, an ML tree was generated from that data (Appendix 2) and the best model as well as Akaike information criterion (AIC) number of the models are depicted on the table below. ALL topologies for each r^2 however were too identical, depicting the same topology and more or less equal branch lengths. (Figure 3 and Appendix 2).

Table 4: A table showing the number of SNP variants before and after VCF LD filtering for reconstructing maximum likelihood tree.

Linkage disequilibrium threshold (r^2) filtering			
r^2	#SNP Variants	ML Model	#AIC
0	2084117	GTR+F	1646715589.3269
0.2	13404302		1646715589.3171
0.5	33546125		1646715589.2433
0.8	44982334		1646715589.2405
0.99	48169119		1646715589.2754

3.3 Genome-wide diversity

The figure (Figure 4) below shows genomic diversity of the *Tragelaphus.* Heterozygosity is the proportion of a diploid organism that has two distinct alleles at a homologous locus, whereas nucleotide diversity is the pairwise average number of nucleotide differences between two DNA sequences in all possible pairs in the sample population. This was done in a sliding window and averaged since variation may vary along the genome. Both heterozygosity and nucleotide diversity of the sampled lineages correlates with the following order from highest to lowest: *T.*

oryx, T. buxtoni, T. spekei, T.s. scriptus, T. s. sylvaticus, T. strepciceros, T. imberbis, T. euryceros, T. derbanius to T. angasi

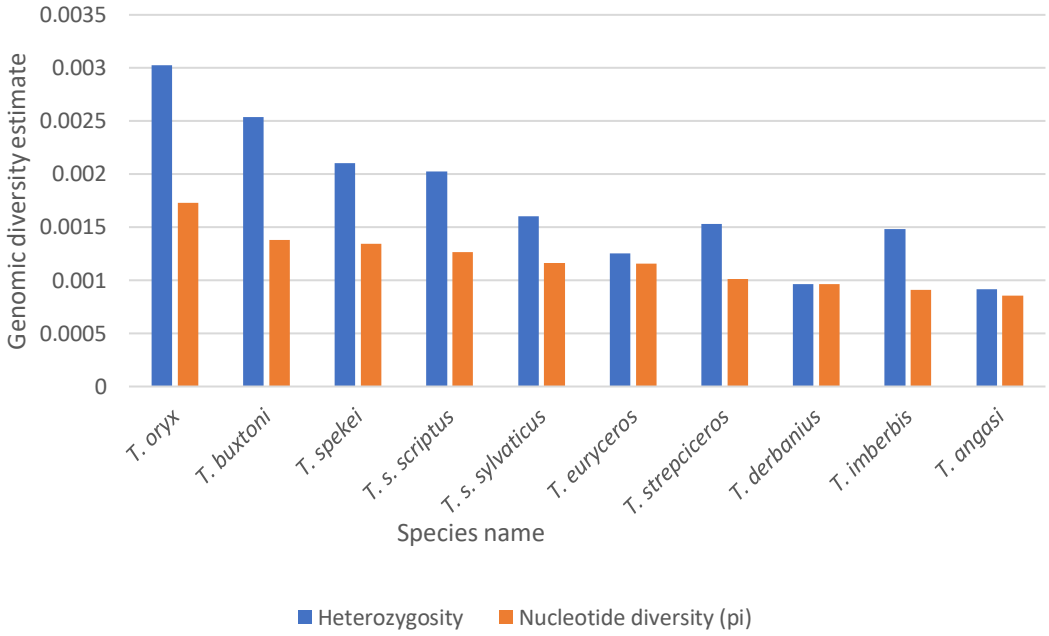


Figure 4. Genomic diversity investigated by two independent measures of genetic diversity, Heterozygosity (Blue) calculated by ANGSD, and Nucleotide diversity (Orange), independent estimations were calculated on BAM files (Table 2) and VCF (141315256 SNPs) respectively.

3.4 Gene flow

The F-branch heatmap (Figure 5) below depict f4-ratio results generated by ABBA-BABA Dsuite 'Dtrios' showing gene flow signals to specific, possibly internal branches on a tree, since the test was done relying on reconstructed species tree. The analysis showed that gene flow has occurred between non-sister lineages. *T. angasi* showed stronger gene flow signals with *T. scriptus* and it also showed strong signals with other *Tragelaphus* species. Gene flow signals with *T. imberbis* was detected in very low levels. There are also signals of gene flow between the *T. buxtoni-T. euryceros-T. spekei* clade and the *T. derbanius-T. oryx-T. strepciceros* clade visible which is depicted to be ancestral geneflow. Geneflow was also seen depicted between the ancestor of *T. derbanius-T. oryx* and *T. spekei,* and furthermore, within the *T. buxtoni-T. euryceros-T. spekei* clade, that is between *T. euryceros* and *T. spekei.* There are also slight gene flow signals shown

between the ancestor of bushbuck and *T. angasi* as well as between: *T. strepciceros* with *T. angasi and T. imberbis, T. spekei* with *T. oryx and T. derbanius.*

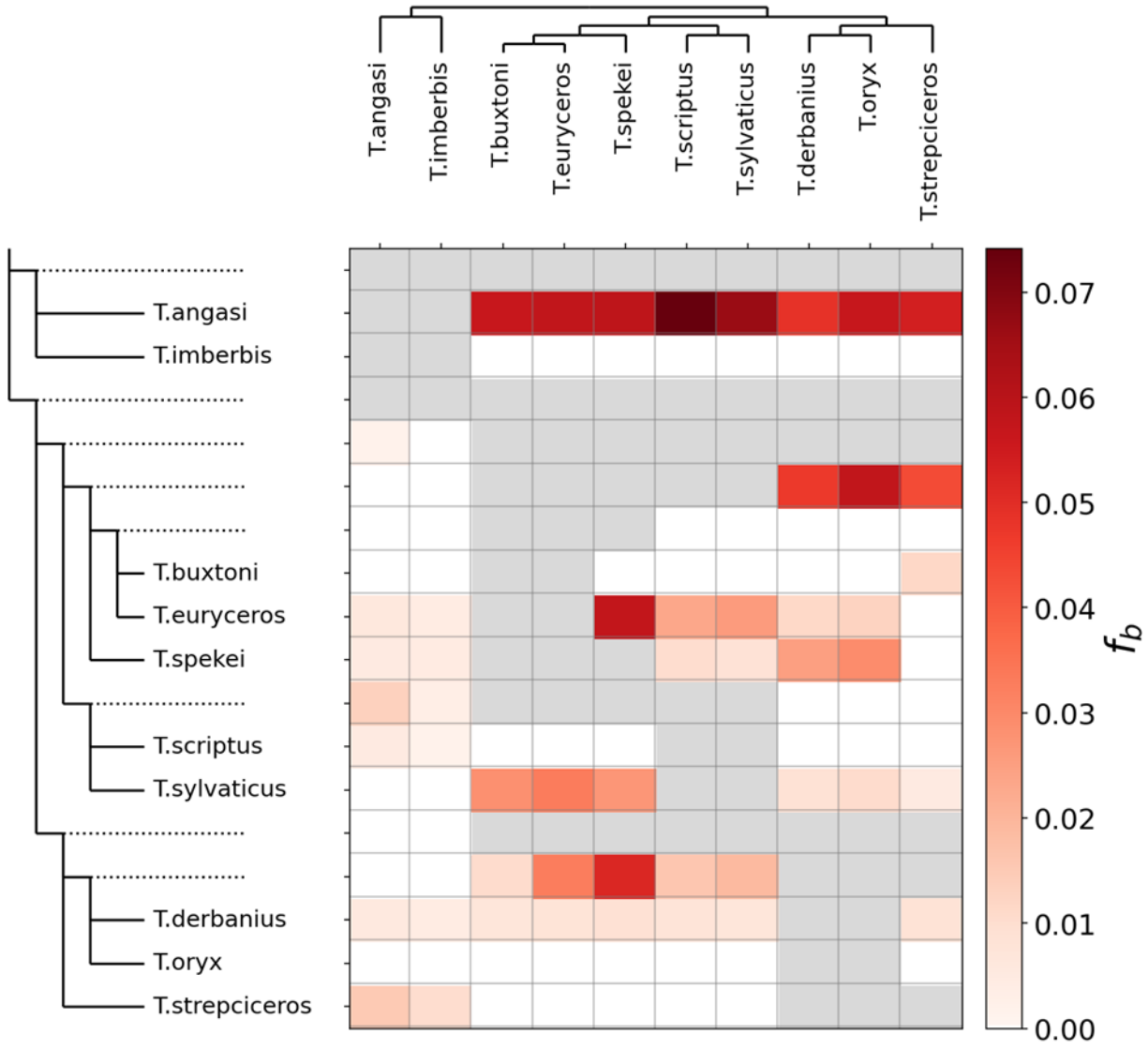


Figure 5: Dsuite Fbranch heatmap based on 120 different possible species trios assessed for introgression tests, identifying tree violating branches (internal branches of the tree, which could be an indicator of gene flow with ancestral branch) and possible gene-flow events. The branch-specific statistic *fb* identifies excess sharing of derived alleles between the branch of the tree on the *y*-axis and the species on the *x*-axis. The *Tragelaphus* species tree was provided as a basis for the branch statistic. Grey data points on the map correspond to tests that are not consistent with the phylogeny. The intensity of the colour corresponds to the excess sharing of alleles between species and descendants of branches.

Below are the introgression results of Dfoil analysis (Table 5, Figure 6). The five-taxon phylogenies with four in-group taxa (P1-P4) and an outgroup (O) that gave the highest proportion of introgressed windows, depicting significant introgression. P12 represents an ancestral branch of P1 and P2. From the presented table below, the tests correspond with the introgression arrows on the schematic summary (Figure 6) results of Dfoil analysis. T. *angasi* showed significant introgression with non-sister taxa. Some introgressions could not have occurred directly between species because the species exist in different habitats but may be remnants of ancestral gene flow. Thus, it was detected by Dfoil analyses, and the results are depicted below. In Table 5 and Figure 6 it is shown that *T. angasi* had ancestral geneflow with the ancestor of bushbuck (Test 1, Table 5), common ancestor of *T. buxtoni-T. euryceros-T. spekei* clade (Test 2), common ancestor of *T. derbanius-T. oryx-T. strepciceros* clade (Test 3), ancestor of *T. derbanius-T. oryx* (Test 4), and with ancestor of *T. buxtoni-T. euryceros* clade (Test 5). Ancestral introgression was also found between the ancestor of *T. derbanius-T. oryx* and *T. spekei* lineage (Test 6), as well as between the ancestor of *T. derbanius-T. oryx* and *T. euryceros* (Test 7). Directional introgression between lineages (Test 8 and 9) was seen into *T. angasi* from *T. scriptus* and *T. strepciceros.*

Table 5: Five-taxon phylogenies with the highest proportion of introgressed windows which were inferred using Dfoil analysis. The cow (*Bos taurus*) was used as the outgroup taxon (P5).

#Test	P1	P2	P3	P4	P5	Introgression
1	*T. s. sylvaticus*	*T. s. scriptus*	*T. angasi*	*T. imberbis*	*B. taurus*	P12⇔P3
2	*T. buxtoni*	*T. spekei*	*T. angasi*	*T. imberbis*	*B. taurus*	P12⇔P3
3	*T. derbanius*	*T. strepciceros*	*T. angasi*	*T. imberbis*	*B. taurus*	P12⇔P3
4	*T. derbanius*	*T. oryx*	*T. angasi*	*T. imberbis*	*B. taurus*	P12⇔P3
5	*T. buxtoni*	*T. euryceros*	*T. angasi*	*T. imberbis*	*B. taurus*	P12⇔P3
6	*T. derbanius*	*T. oryx*	*T. spekei*	*T. buxtoni*	*B. taurus*	P12⇔P3
7	*T. derbanius*	*T. oryx*	*T. buxtoni*	*T. euryceros*	*B. taurus*	P12⇔P4

#Test	P1	P2	P3	P4	P5	Introgression
8	*T. s. scriptus*	*T. s. sylvaticus*	*T. angasi*	*T. imberbis*	*B. taurus*	P1 ⇒ P3
9	*T. derbanius*	*T. strepciceros*	*T. angasi*	*T. imberbis*	*B. taurus*	P2 ⇒ P3

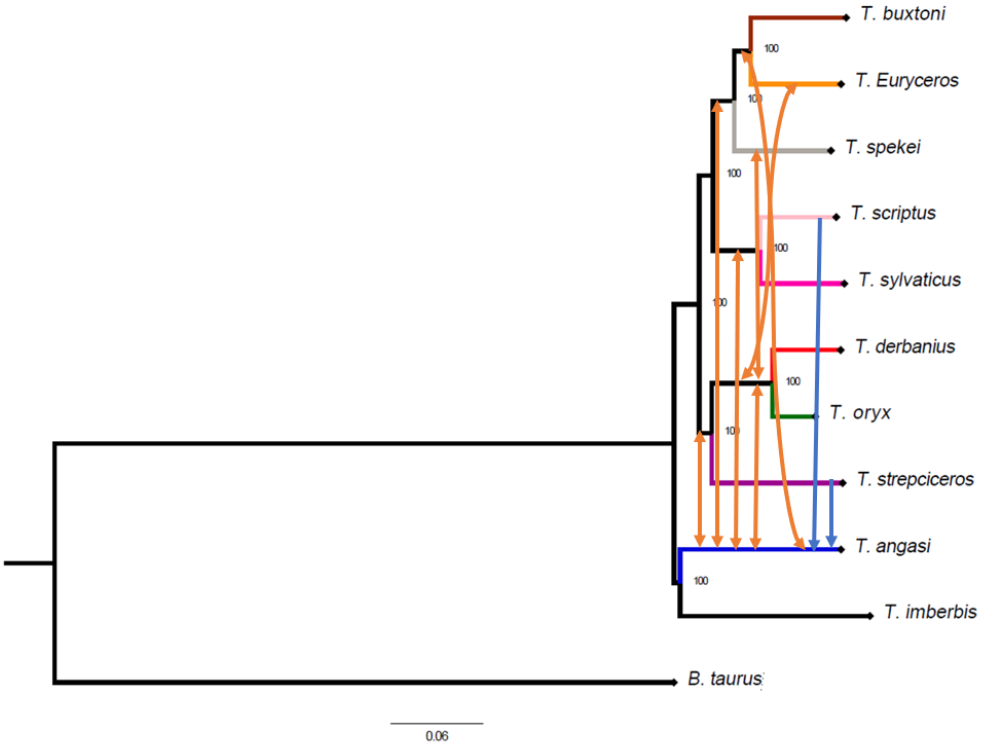


Figure 6: Graphical summary of directional gene flow analyses using *DFOIL* onto the maximum likelihood species tree. The arrows indicate gene flow occurring between non-sister lineages. Orange bi-directional arrows show ancestral gene flow, and the blue arrows indicate the directional gene flow between species lineages.

4. Discussion

4.1 Consistency of genomic structure

All the genomic structure analysis done in this project were seen giving the same overall evolutionary pattern of the genus *Tragelaphus*. From the model-independent PCA to model based neighbor-joining and maximum likelihood. The whole genome data used reveals *T. imberbis* on the 1st component of PCA method to be separated from the rest of the *Tragelaphus,* explaining approximately 19% of the data (Appendix1) and the species trees depicts it being the most ancient of the *Tragelaphus* looking at its longest branch length in ML tree (Figure 4; Appendix 2) and that it diverged earlier than the rest.

I can say that the similar structures given by different methods depict the true genome-wide species tree reconstructed for *Tragelephus.* Although it is known that evolution is driven by different forces (Ackermann and Cheverud 2004) this revelation gives confidence about the evolutionary history of genetic drift on the group. Genetic drift causes structure by random chance of keeping or removing an allele in a population. The spiral-horned antelopes are diverse and populate most habitat in sub-Saharan Africa (Rakotoarivelo *et al.* 2019) but their structure is seen over and above gene flow that has been detected, meaning that genetic drift was strong. It is possible that the ancestral effective population sizes of *Tragelaphus* populations might have been small, leading to strong structuring. However, Pairwise Sequential Markovian Coalescent (PSMC) (Li and Durbin 2011) could be employed to check the effective population size of *Tragelaphus* using high coverage data. The PSMC method relies on the detection of regions of accumulated heterozygous sites. Unfortunately, the data generated in this project was of low coverage thus PSMC was not run to avoid inconclusive results.

4.2 Linkage correction

When correcting for linked sites, with application of different linkage thresholds (r^2) for maximum likelihood trees, this demonstrated that a high r^2 number allowed for more SNPs to be counted, resulting in reduced linked sites. The remaining variants in each r^2 gave out consensus species trees that were too identical (Appendix 2). LD filtering had very slight differences of branch lengths

on the topologies (Appendix 2). LD was corrected for since whole genomes are large markers that may give biased results (Kling 2015) due to strong linked loci. The *Tragelaphus* analyses gave the same topology even when LD was accounted for. Thus, it builds further confidence that the reconstructed tree is a phylogeny drawing closer to the true species tree. When generating a tree in IQtree, since LD was accounted for, the AIC number was recorded to see how much information was lost. I can say that LD has very little effect on this whole genome phylogenetic reconstruction.

4.3 Comparison with previous studies

The genome-wide phylogenetic trees reconstructed in this study depicts that in the *Tragelaphus* genus, *T. imberbis* and *T. angasi* are the most basal with all the other *Tragelaphus* forming two clades, one of dry savanna and the other of closed forest species with the two bushbucks monophyly in the closed forest species clade as was shown by the previous nuclear gene studies (Willows-Munro *et al.* 2005; Hassanin *et al.* 2018; Rakotoarivelo *et al.* 2019). The species trees in this study, supported by (Willows-Munro *et al.* 2005; Hassanin *et al.* 2018; Rakotoarivelo *et al.* 2019), does not only dispute (1) the mtDNA sister relationship of bushbuck *T. s. scriptus* to *T. angasi* and (2) mtDNA paraphyly of bushbucks, but also differ from the previously reconstructed trees using traditional markers with the placement of species within each clade mentioned above. *T. strepciceros* was shown to belong to the closed forest species clade (Willows-Munro *et al.* 2005; Rakotoarivelo *et al.* 2019) whereas genome-wide data shows *T. strepciceros* belonging to the dry savanna clade as sister to *T. derbanius*-*T. oryx* clade, this was shown also by mtDNA (Willows-Munro *et al.* 2005; Bibi 2013; Hassanin *et al.* 2018). Furthermore, within the closed forest clade previous traditional markers placed *T. euryceros* in close sister relationship with *T. spekei* (Willows-Munro *et al.* 2005; Bibi 2013; Hassanin *et al.* 2018; Rakotoarivelo *et al.* 2019) and in the genome-wide level, *T. euryceros* is sister to *T. buxtoni* with *T. spekei* sister to *T. euryceros*-*T. buxtoni* clade.

4.3.1 The placement of *T. strepciceros*

As mentioned above that genome-wide placement of species on a tree, was depicted also by the non-model based PCA (Figure 1; components 1st-4th), on that note in 3rd component of PCA, *T. oryx* was close to *T. strepciceros* and later reconciled with *T. derbanius* in 4th component. The 3rd and 4th component had approximately 13% and 12% (Appendix 1) variation respectively, which it could be said that the difference is so small that a biological model could give a more definite placement of these species. On the same note, it is also interesting that *T. strepciceros* was placed further away from the same cluster and from all the *Tragelaphus* on the 4th component (Figure 1), this could be the evidence of significant ancestral introgression of *T. strepciceros* with other *Tragelaphus* species detected (Dfoil; Figure 5). The significant gene flow signal involving *T. strepciceros* detected was with *T.angasi* (Figure 5; Table 5 Test 9), and non-significant with other lineages, depicted in Dsuite (Figure 5).

The longer branch of the *T. strepciceros* on the ML tree (Appendix 2) suggests that *T. strepciceros* is the older branch than all other *Tragelaphus* except of the most basal longer branched *T. angasi* and *T. imberbis*, with *T.derbanius* and *T. oryx* as the youngest thus shorter branches. The long branch of *T. strepciceros* and its separation in fourth component of PCA (Figure 1) may be suggesting sharing of allele with ancient *T. angasi* (Figure 5; Table 5 Test 9) or slight sharing of allele with *T. imberbis* (Figure 5), thus the previous placement of *T. strepciceros* in an intermediate position between basal most species and the *T. oryx* (Willows-Munro *et al.* 2005).

4.4 Gene flow and convergent phenotypes

The level of introgression found in the *Tragelaphus* genus was great between non-sister lineages. This could probably explain the uncertainty of previous studies to reconstruct evolutionary relationships using traditional markers (as mentioned in the previous paragraphs) but could also explain phenotypic convergence among these non-sister taxa within the genus. A lot of detected gene flow was ancestral and between non-sister lineages. It could be suspected that during speciation, these lineages must have not been reproductively isolated as they are now, and they had secondary contact exchanging beneficial alleles. Interestingly, *T. angasi* showed more ancestral gene flow with non-sister species but not its closely related sister taxa *T. imberbis*.

Although this can be explained by the long branch lengths of these basal most species (distant sister taxa).

There was strong allele sharing of non-sister lineages detected between *T. angasi* and *T. s. scriptus* (Figure 5) which was shown to be the directional introgression from *T. s. scriptus* lineage into the *T. angasi* lineage (Table 5; Figure 6). This raises concern of whether mtDNA of *T. angasi* was introgressed into the bushbuck as previously suggested. Evidence of an ancestral gene flow between the ancestor of bushbuck and *T. angasi* could have given the *T. angasi*-like mtDNA to both bushbuck. This secondary contact must have occurred then before bushbucks diverged but *T. angasi*-like mtDNA was kept in the *Scriptus* lineage and lost in *Sylvaticus* due to genetic drift that happened after the ancestral gene flow event.

From the most basal *T. imberbis* F-branch (Figure 5) showed slight introgression with *T. strepciceros* which could serve as an explanation that the two species *T. imberbis* and *T. strepciceros* due to cranial similarities were morphologically placed as sister taxa (Hassanin and Douzery 1999; Matthee and Robinson 1999) but are non-sister at the genome-wide level. But this introgression could not be concluded as great evidence, as in F-branch (Figure 5) *T. imberbis* also portrayed to have slight introgression with other lineages. *T. strepciceros* and *T. angasi* also depicted signal of introgression. *T. strepciceros* was placed away from other species on a PCA (Figure 1), and its placement on species tree was previously not certain, this could be because it slightly shares alleles with other species in the *Tragelaphus* genus. It was evident in ABBA-BABA (Figure 5) results, showing slight (non-significant) allele sharing between *T. strepciceros* and *T.buxtoni* and non-significant ancestral alleles sharing of *T. strepciceros* with the ancestor of *T. buxtoni-T. euryceros-T. spekei* was also depicted (Figure 5).

The phenotypes of *T. spekei*, *T. buxtoni* and bushbucks (closed forest group) generally resembles the *T. angasi* (nyala-like) phenotype, and in this genome-wide study, the directional gene flow results (Table 5 and Figure 6) showed that of the mentioned closed forest species, the ancestor of *T. spekei* - *T. buxtoni* and ancestor of bushbucks had gene flow with non-sister *T. angasi.* Thus, these multiple gene flow events with the *T. angasi* could explain the observed nyala-like phenotype of the closed forest group. On the same group of the closed forest, *T. s. sylvaticus, T. euryceros* and basal *T.angasi* show great levels of sexual dimorphism in coat color, this coat color trait amongst these non-sister lineages could also be result of gene flow between *T. angasi* and ancestor of the bushbuck (Table 5, Test 1), and *T. angasi* with ancestor of *T. euryceros-T. buxtoni* (Table 5, Test 5) may explain the high levels of the same sexual dimorphism.

Another convergent phenotype of non-sister taxa is the presence of horns in females and males of *T. euryceros,* and monophyletic *T. oryx*-*T. derbanius*. Ancestral gene flow between these species was slightly shown in f-branch analysis (Figure 5) and more clearly by Dfoil (Table 5: Test 7 and Figure 6). The ancestor of *T.oryx*-*T.derbanius* had gene flow with *T. euryceros* (Figure 5; Table 5: Test 7 and Figure 6) this gene flow could also explain the large size trait of the *T. euryceros* larger than other closed forest species but similar to that of the dry savanna *T.oryx*-*T.derbanius.*

However, it is not known whether these convergent phenotypes are truly transmitted by gene flow. To assess this, it would require looking at each segment introgressed between non-sister species and investigate whether they contain genes responsible for these convergent phenotypes. Although to do this analysis, it will require more genomes per species to accurately estimate the site frequency spectrum of the species that will account for the species populations.

4.5 Genomic diversity

It must be noted that, gene flow destroys structure of a population, whereas genetic drift increase diversity between isolated populations. The genome-wide diversity in this study showed that *T. angasi,* that had gene flow with most of the *Tragelaphus* species has lowest genomic diversity (Figure 4), this is expected since gene flow decreases structure. Most of the dry savanna group including the basal *T. imberbis* had lower genomic diversity than some closed forest group species. On average the closed forest group depict greater genome-wide diversity (with pi of 0.00126 and heterozygosity of 0.00190), than dry savanna (with pi of 0.00121 and heterozygosity of 0.00183) species. Although, the overall genome-wide diversity of *Tragelaphus,* noting that there were great signals of introgression between non-sister lineages cannot account for the whole population. So, the level of diversity in this study could not conclusively explain the survival of the *Tragelaphus* populations but more the quality of the data. More population whole genome data for further and more clearer analyses.

It is evident in Table 2 that the BAM files had low coverage (Table 2), and although the ANGSD software used to calculate heterozygosity was made to handle the low coverage data (Korneliussen *et al.* 2014), it could be accepted that the trend of genome-wide diversity results given, depicts the diversity trend within *Tragelaphus* populations. The ANGSD software only calculates site frequency spectrum for each species (provided samples). Since in this project only one sample of each species was sampled, the true genomic diversity of each species cannot be

accurately estimated, thus cannot be concluded. Having highest heterozygosity simply mean that the individual has more heterozygous sites. Furthermore, for nucleotide diversity which also gave the same diversity trend as heterozygosity, was calculated along the genome (sliding window), this was calculated with the knowledge and expectation of it varying along the genome, but average diversity was taken to compare with heterozygosity. There were regions (windows) along the genome with high diversity than others, which could be investigated and assessed whether they are involved in adaptive introgression and selected amongst species. That (selection) could be the next study.

5. Conclusion

Whole genomes provide a greater insight on the relationships of species. In this study *Tragelaphus* whole genome species tree was reconstructed and more analyses (gene flow) to get an insight on the relationship of unclear placement of some species in previous studies was done. High ancestral gene flow between non-sister lineages was revealed, explaining the incongruency of species trees previously reconstructed using traditional markers. The main example of this is placing *T. angasi* with *T. s. scriptus* as sister taxa at mtDNA level, which I show here was because of gene flow between the ancestor of both bushbuck and *T. angasi.* The nyala-like mtDNA must have been lost on the *Sylvaticus* lineage due to genetic drift. The observed non-sister gene flows may also explain the similar phenotypes of different lineages (convergent phenotypes). It can be said that these lineages must not have been reproductively isolated during speciation, and that enhanced greater levels of introgression between non-sister lineages. However, a greater number of sequenced genomes in each species would give much greater insight on evolution of this genus and would also allow looking at gene flow together with selection, that is, which genes are being introgressed and/or selected for.

6. Acknowledgements

I am grateful to have worked on this project under the supervision of Professor Yoshan Moodley. Appreciation goes to my co-supervisor Dr Andrinajoro Rakotoarivelo for his outstanding guidance

and assistance in the Bioinformatics laboratory. My fellow postgraduates' students in our lab have also played a huge role in keeping me sane when it was tough, I am grateful to them.

The data in this project was generated at the University of Potsdam as part of a PhD project by Ulrike Taron. I'm thankful to her for doing the molecular laboratory work. Further thanks extended to Professor Moodley and Professor Michael Hofreiter of Potsdam for allowing me to work with their data.

I acknowledge the financial support of the National Research Foundation (NRF) that I received and the University of Venda for allowing me to work on this project registered in their institution. Words fail to express my gratitude to my parents; I direct many appreciations to them.

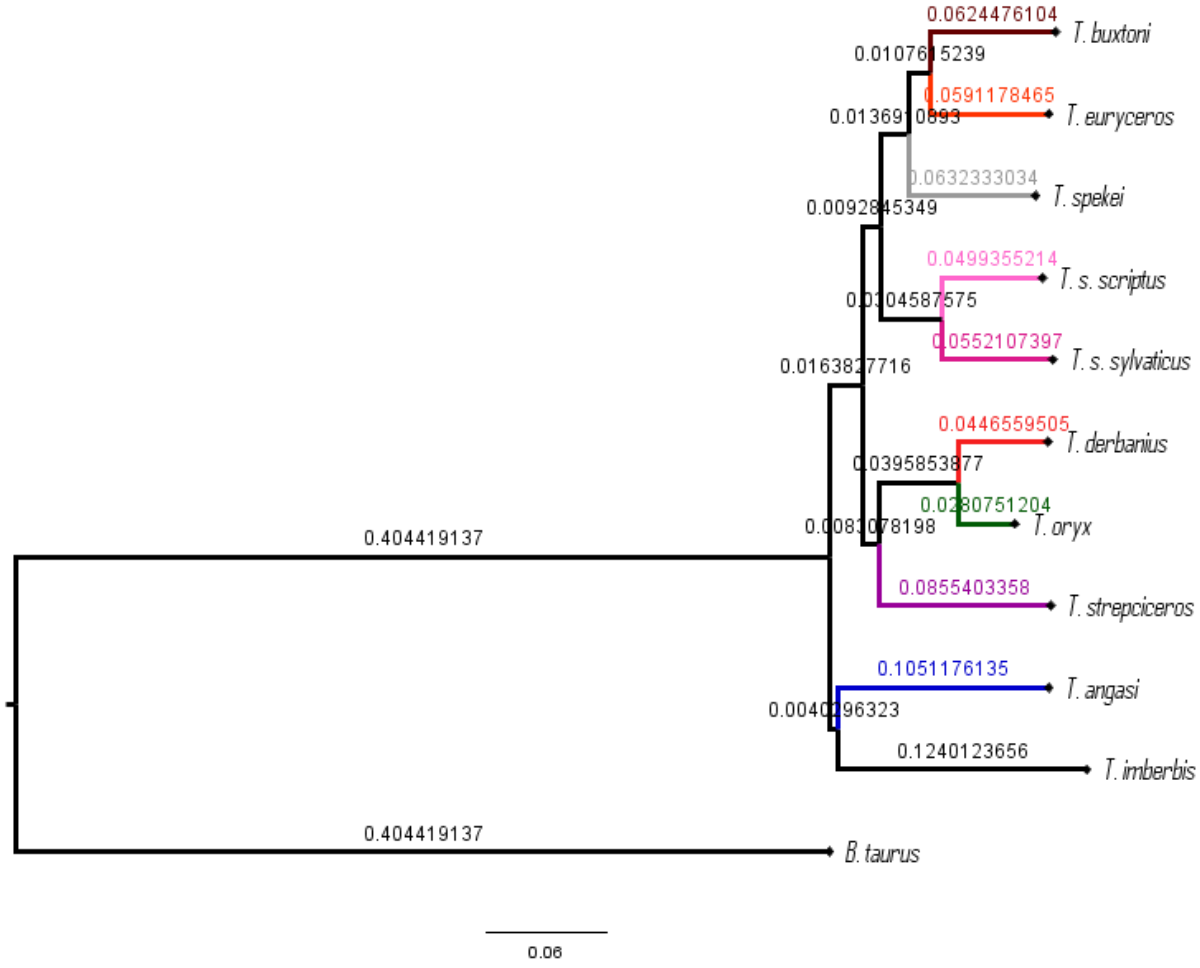
7. Appendix

Appendix 1: A table showing the amount of variation in each principal components (PC) of the graphs shown in figure 2.

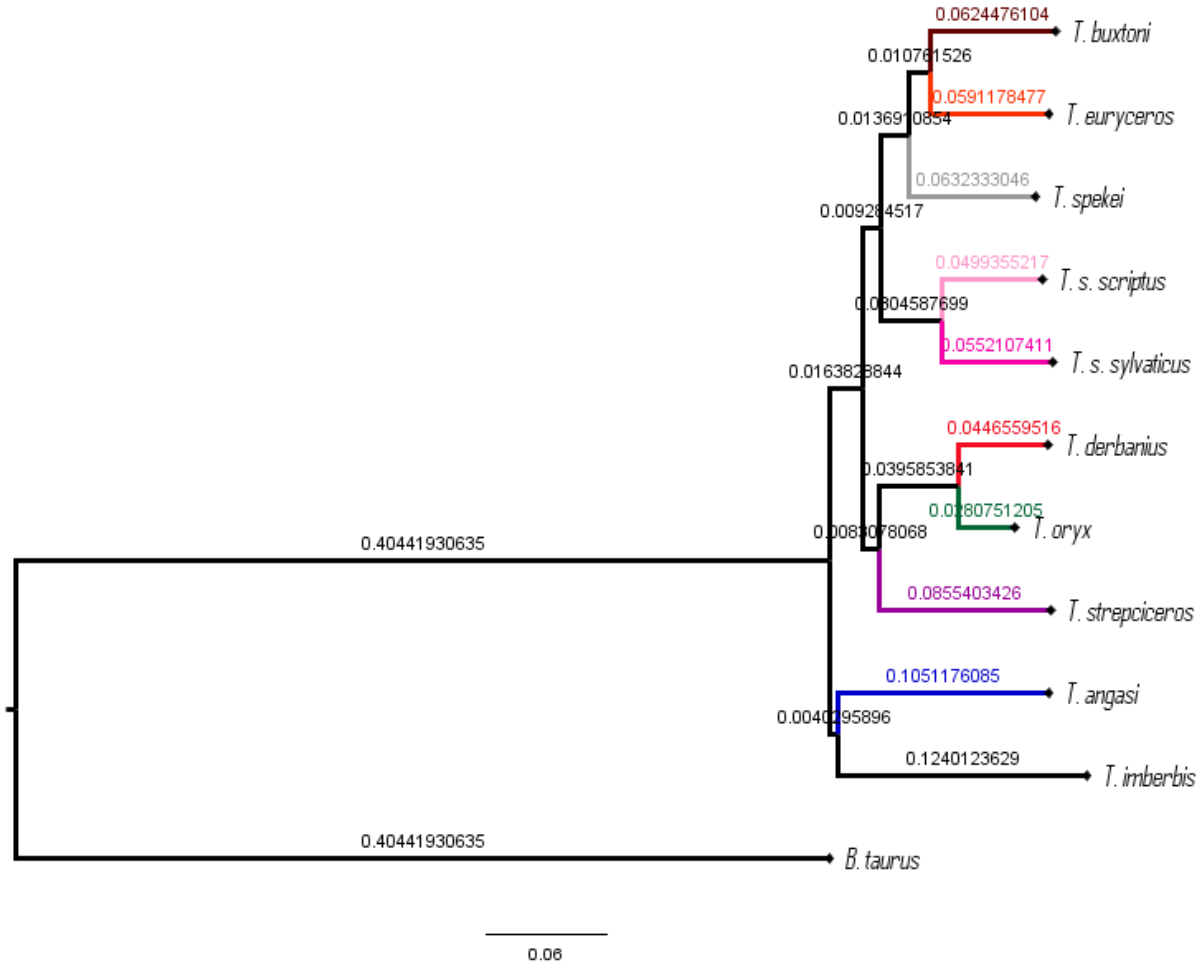
Principal Components	Variation (%)
First	18.88
Second	15.56
Third	12.73
Fourth	11.6
Fifth	11.31
Sixth	8.62
Seventh	8.16
Eighth	7.66
Ninth	5.45
Tenth	0.03
Total	100

Appendix 2:

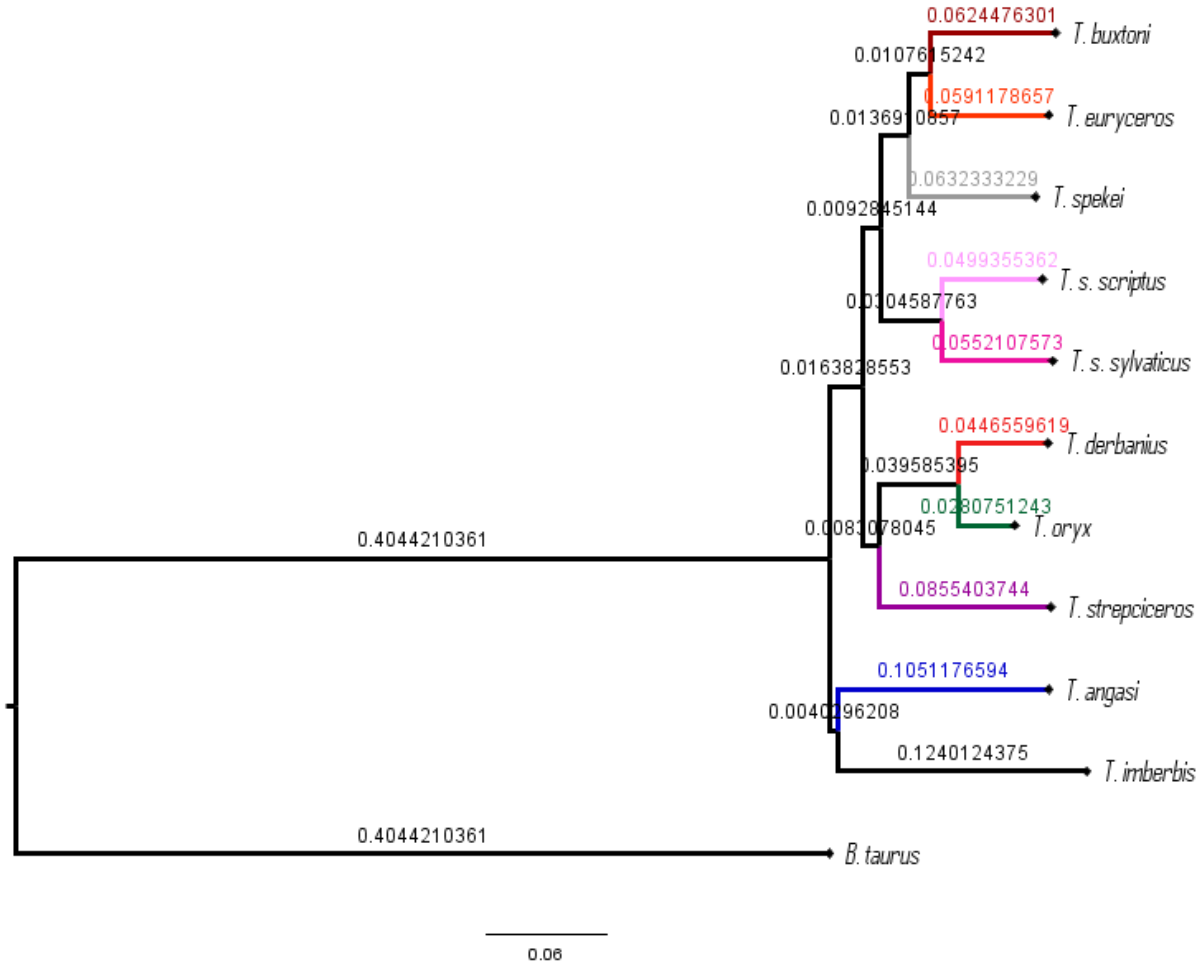
Consensus trees generated with different linkage disequilibrium thresholds (r^2) are displayed below. They were analysed independently, and their branches were given different colours which corresponds to other structure analyses (PCA Figure 1; IBS Figure 2) and their lengths (numbers on tree). For each tree different numbers of SNPs according to r^2 filtering were used (see figure legends).



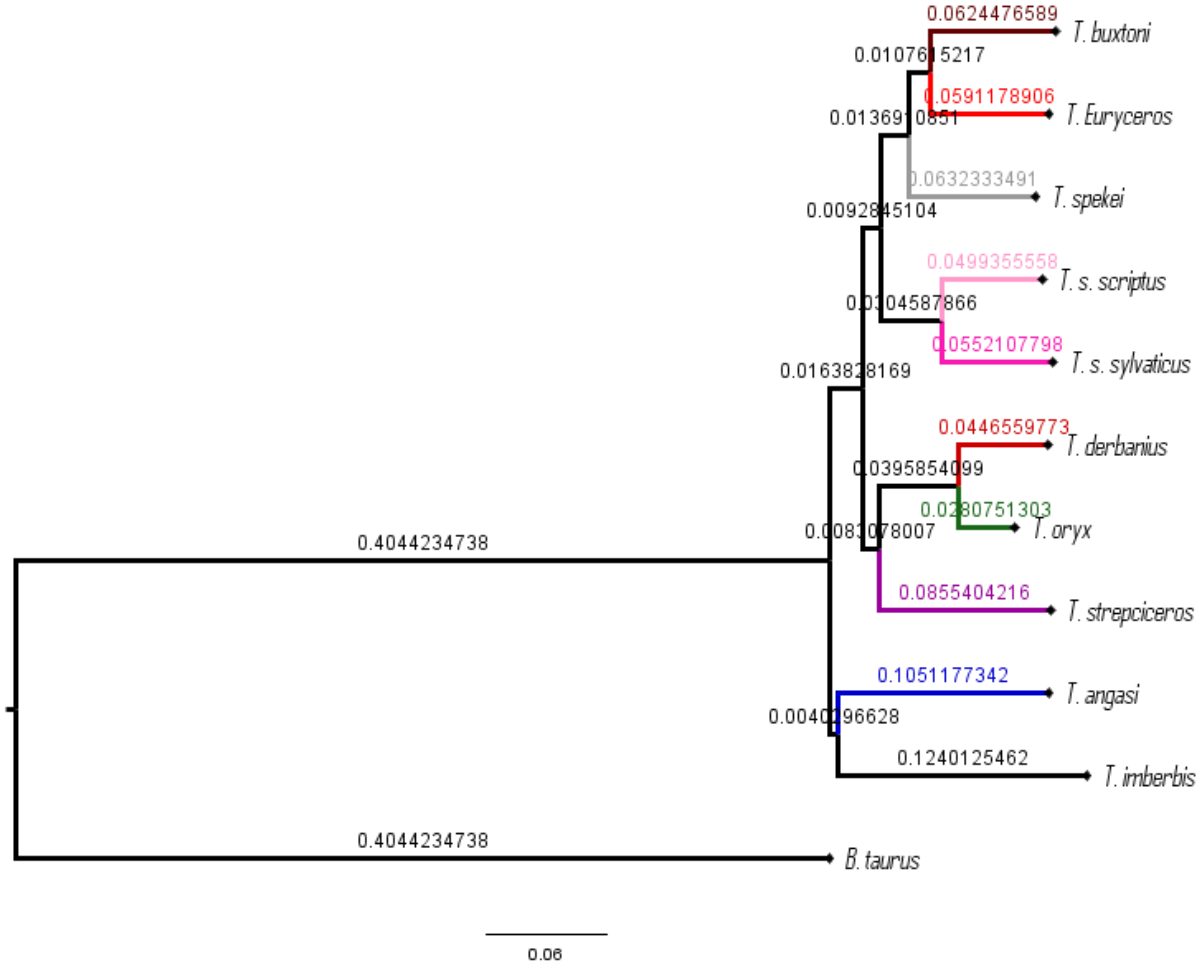
Appendix 2a: ML tree with r^2 of 0. The 11 sequences VCF input had 2084117 SNPs and from it, a tree was generated by IQtree which implemented the GTR+F (general time reversible) model for unequal base frequencies and plotted on Figtreev1.4.3. The numbers on tree depict branch lengths which are slightly different from other trees of different r^2. The nodes bootstrap support is 100% and not shown on tree for branch lengths visibility purpose. The branch colours correspond with the branch length, PCA (Figure 1) and IBS tree (Figure 2).



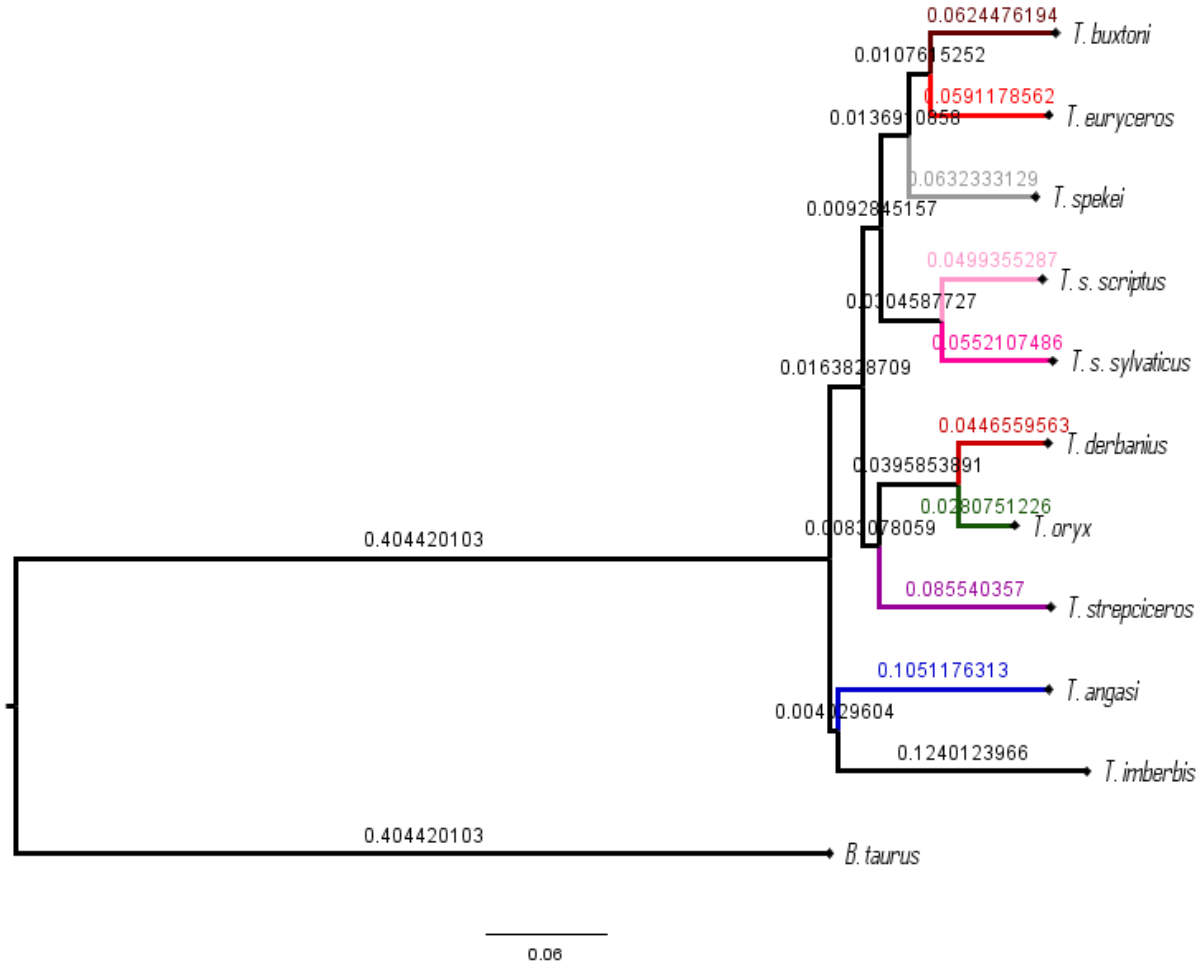
Appendix 2b: ML tree with r^2 of 0.2. The 11 sequences VCF input had 13404302 SNPs and from it, a tree was generated by IQtree which implemented the GTR+F (general time reversible) model for unequal base frequencies and plotted on Figtreev1.4.3. The numbers on tree depict branch lengths which are slightly different from other trees of different r^2. The nodes bootstrap support is 100% and not shown on tree for branch lengths visibility purpose. The branch colours correspond with the branch length, PCA (Figure 1) and IBS tree (Figure 2).



Appendix 2c: ML tree with r^2 of 0.5. The 11 sequences VCF input had 33546125 SNPs and from it, a tree was generated by IQtree which implemented the GTR+F (general time reversible) model for unequal base frequencies and plotted on Figtreev1.4.3. The numbers on tree depict branch lengths which are slightly different from other trees of different r^2. The nodes bootstrap support is 100% and not shown on tree for branch lengths visibility purpose. The branch colours correspond with the branch length, PCA (Figure 1) and IBS tree (Figure 2).



Appendix 2d: ML tree with r^2 of 0.8. The 11 sequences VCF input had 44982334 SNPs and from it, a tree was generated by IQtree which implemented the GTR+F (general time reversible) model for unequal base frequencies and plotted on Figtreev1.4.3. The numbers on tree depict branch lengths which are slightly different from other trees of different r^2. The nodes bootstrap support is 100% and not shown on tree for branch lengths visibility purpose. The branch colours correspond with the branch length, PCA (Figure 1) and IBS tree (Figure 2).



Appendix 2e: ML tree with r^2 of 0.99. The 11 sequences VCF input had 48169119 SNPs and from it, a tree was generated by IQtree which implemented the GTR+F (general time reversible) model for unequal base frequencies and plotted on Figtreev1.4.3. The numbers on tree depict branch lengths which are slightly different from other trees of different r^2. The nodes bootstrap support is 100% and not shown on tree for branch lengths visibility purpose. The branch colours correspond with the branch length, PCA (Figure 1) and IBS tree (Figure 2).

8. References

Abouheif, E., 1997. Developmental genetics and homology: a hierarchical approach. *Trends in Ecology & Evolution*, *12*(10), pp.405-408.

Ackermann, R.R. and Cheverud, J.M., 2004. Detecting genetic drift versus selection in human evolution. *Proceedings of the National Academy of Sciences*, *101*(52), pp.17946-17951.

Akgün, M. and Demirci, H., 2017. VCF-Explorer: filtering and analysing whole genome VCF files. *Bioinformatics*, *33*(21), pp.3468-3470.

Atickem, A., Klapproth, M., Fischer, M., Zinner, D. and Loe, L.E., 2022. Home range and habitat selection of female mountain nyalas (*Tragelaphus buxtoni*) in the human-dominated landscape of the Ethiopian Highlands. *Mammalian Biology*, pp.1-8.

Ayala, F.J., 2009. Molecular evolution. *Ruse & Travis*, *2009*, pp.132-151.

Barbato, M., Hailer, F., Upadhyay, M., Del Corvo, M., Colli, L., Negrini, R., Kim, E.S., Crooijmans, R.P., Sonstegard, T. and Ajmone-Marsan, P., 2020. Adaptive introgression from indicine cattle into white cattle breeds from Central Italy. *Scientific reports*, *10*(1), pp.1-11.

Barton, N.H., 1996. Natural selection and random genetic drift as causes of evolution on islands. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*(1341), pp.785-795.

Benestan, L.M., Ferchaud, A.L., Hohenlohe, P.A., Garner, B.A., Naylor, G.J., Baums, I.B., Schwartz, M.K., Kelley, J.L. and Luikart, G., 2016. Conservation genomics of natural and managed populations: building a conceptual and practical framework.

Bibi, F., 2011. *Tragelaphus nakuae*: evolutionary change, biochronology, and turnover in the African Plio-Pleistocene. *Zoological Journal of the Linnean Society*, *162*(3), pp.699-711.

Bibi, F., 2013. A multi-calibrated mitochondrial phylogeny of extant Bovidae (Artiodactyla, Ruminantia) and the importance of the fossil record to systematics. *BMC evolutionary biology*, *13*(1), pp.1-15.

Blondel, C., Rowan, J., Merceron, G., Bibi, F., Negash, E., Barr, W.A. and Boisserie, J.R., 2018. Feeding ecology of Tragelaphini (Bovidae) from the Shungura Formation, Omo Valley, Ethiopia:

contribution of dental wear analyses. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *496*, pp.103-120.

Bock, F., Gallus, S., Janke, A., Hailer, F., Steck, B.L., Kumar, V. and Nilsson, M.A., 2014. Genomic resources and genetic diversity of captive lesser kudu (*Tragelaphus imberbis*). *Zoo biology*, *33*(5), pp.440-445.

Boyden, A., 1947. Homology and analogy. A critical review of the meanings and implications of these concepts in biology. *American Midland Naturalist*, pp.648-669.

Carson, A.R., Smith, E.N., Matsui, H., Brækkan, S.K., Jepsen, K., Hansen, J.B. and Frazer, K.A., 2014. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC bioinformatics*, *15*(1), pp.1-15.

Castresana, J., 2001. Cytochrome *b* phylogeny and the taxonomy of great apes and mammals. *Molecular Biology and Evolution*, 18(4), pp.465-471.

Dalton, D.L., Tordiffe, A., Luther, I., Duran, A., Van Wyk, A.M., Brettschneider, H., Oosthuizen, A., Modiba, C. and Kotzé, A., 2014. Interspecific hybridization between greater kudu and nyala. *Genetica*, *142*(3), pp.265-271.

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

Darwin, C., 2008. The descent of man, and selection in relation to sex. *Princeton University Press*.

Darwin, C.,1859. The Origin of Species by Means of Natural Selection. *John Murray.*

Degli Esposti, M., De Vries, S., Crimi, M., Ghelli, A., Patarnello, T. and Meyer, A., 1993. Mitochondrial cytochrome b: evolution and structure of the protein. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, *1143*(3), pp.243-271.

Dobzhansky, T., 1940. Speciation as a stage in evolutionary divergence. *The American Naturalist*, *74*(753), pp.312-321.

Dobzhansky, T., 2013. Nothing in biology makes sense except in the light of evolution. *The american biology teacher*, *75*(2), pp.87-91.

Du Toit, J.T. and Cumming, D.H., 1999. Functional significance of ungulate diversity in African savannas and the ecological implications of the spread of pastoralism. *Biodiversity & Conservation*, *8*(12), pp.1643-1661.

Durand, E.Y., Patterson, N., Reich, D. and Slatkin, M., 2011. Testing for ancient admixture between closely related populations. *Molecular biology and evolution*, *28*(8), pp.2239-2252.

Elkan Jr, P.W., 2003. Ecology and conservation of bongo antelope *(Tragelaphus eurycerus*) in lowland forest, northern Republic of Congo. *University of Minnesota*.

Ellegren, H., 2009. The different levels of genetic diversity in sex chromosomes and autosomes. *Trends in Genetics*, *25*(6), pp.278-284.

Estes, R.. "kudu." *Encyclopedia Britannica*, March 11, 2020. https://www.britannica.com/animal/kudu.

Evangelista, P., Swartzinski, P. and Waltermire, R., 2007. A profile of the mountain nyala (*Tragelaphus buxtoni*). *African Indaba*, *5*(2), pp.1-47.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, *17*(6), pp.368-376.

Fernández, M.H. and Vrba, E.S., 2005. A complete estimate of the phylogenetic relationships in Ruminantia: a dated species-level supertree of the extant ruminants. *Biological reviews*, *80*(2), pp.269-302.

Fisher, D.C., 1985. Evolutionary morphology: beyond the analogous, the anecdotal, and the ad hoc. *Paleobiology*, *11*(1), pp.120-138.

Fisher, R.A., 1923. XXI.—on the dominance ratio. *Proceedings of the royal society of Edinburgh*, *42*, pp.321-341.

Frantz, L.A., Schraiber, J.G., Madsen, O., Megens, H.J., Bosse, M., Paudel, Y., Semiadi, G., Meijaard, E., Li, N., Crooijmans, R.P. and Archibald, A.L., 2013. Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. *Genome biology*, *14*(9), pp.1-12.

Furstenburg, D., 2002. Nyala (*Tragelaphus angasii*). *Game and Hunt*, *8*, pp.8-9.

Furstenburg, D., *Tragelaphus scriptus* (Pallas, 1766).

Furstenburg, D., *Tragelaphus spekii* (Speke, 1863).

Galla, S.J., Forsdick, N.J., Brown, L., Hoeppner, M.P., Knapp, M., Maloney, R.F., Moraga, R., Santure, A.W. and Steeves, T.E., 2019. Reference genomes from distantly related species can be used for discovery of single nucleotide polymorphisms to inform conservation management. *Genes*, *10*(1), p.9.

Garamszegi, L.Z. ed., 2014. Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice. *Springer*.

Garrison, E., Kronenberg, Z.N., Dawson, E.T., Pedersen, B.S. and Prins, P., 2021. Vcflib and tools for processing the VCF variant call format. *BioRxiv*.

Gatesy, J., Amato, G., Vrba, E., Schaller, G. and DeSalle, R., 1997. A cladistic analysis of mitochondrial ribosomal DNA from the Bovidae. *Molecular phylogenetics and evolution*, *7*(3), pp.303-319.

Gentry, A.W., 1990. Evolution and dispersal of African Bovidae. In Horns, pronghorns, and antlers (pp. 195-227). *Springer, New York, NY*.

Gentry, A.W., Bovidae, L.W. and Sanders, W.J., 2010. Cenozoic mammals of Africa. (L. Werdelin, W. J. Sanders, eds.). *University of California Press, Berkley, Carlifornia*

Georgiadis, N.J., Kat, P.W., Oketch, H. and Patton, J., 1990. Allozyme divergence within the Bovidae. *Evolution*, *44*(8), pp.2135-2149.

Goodwin, S., McPherson, J.D. and McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), pp.333-351.

Gould, S.J., 2002. The structure of evolutionary theory. *Harvard University Press*.

Griffiths, C.S., 1997. Correlation of functional domains and rates of nucleotide substitution in cytochrome *b*. *Molecular Phylogenetics and Evolution*, *7*(3), pp.352-365.

Hall, T., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In *Nucleic Acids Symp. Ser.* (Vol. 41, pp. 95-98).

Hassanin, A. and Douzery, E.J., 1999. Evolutionary affinities of the enigmatic saola (Pseudoryx nghetinhensis) in the context of the molecular phylogeny of Bovidae. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *266*(1422), pp.893-900.

Hassanin, A. and Douzery, E.J., 1999. The tribal radiation of the family Bovidae (Artiodactyla) and the evolution of the mitochondrial cytochrome b gene. *Molecular phylogenetics and evolution*, *13*(2), pp.227-243.

Hassanin, A. and Douzery, E.J., 2003. Molecular and morphological phylogenies of Ruminantia and the alternative position of the Moschidae. *Systematic biology*, *52*(2), pp.206-228.

Hassanin, A., Delsuc, F., Ropiquet, A., Hammer, C., Van Vuuren, B.J., Matthee, C., Ruiz-Garcia, M., Catzeflis, F., Areskoug, V., Nguyen, T.T. and Couloux, A., 2012. Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. *Comptes rendus biologies*, *335*(1), pp.32-50.

Hassanin, A., Houck, M.L., Tshikung, D., Kadjo, B., Davis, H. and Ropiquet, A., 2018. Multi-locus phylogeny of the tribe Tragelaphini (Mammalia, Bovidae) and species delimitation in bushbuck: evidence for chromosomal speciation mediated by interspecific hybridization. *Molecular Phylogenetics and Evolution*, *129*, pp.96-105.

Hedrick, P.W., 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular ecology*, *22*(18), pp.4606-4618.

Hillis, D.M. and Bull, J.J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic biology*, *42*(2), pp.182-192.

Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q. and Vinh, L.S., 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular biology and evolution*, *35*(2), pp.518-522.

Holder, M. and Lewis, P.O., 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews genetics*, *4*(4), pp.275-284.

Howell, N., 1989. Evolutionary conservation of protein regions in the protonmotive cytochromeb and their possible roles in redox catalysis. *Journal of molecular evolution*, *29*(2), pp.157-169.

Irwin, D.M., Kocher, T.D. and Wilson, A.C., 1991. Evolution of the cytochrome b gene of mammals. *Journal of molecular evolution*, *32*(2), pp.128-144.

Jill Harrison, C. and Langdale, J.A., 2006. A step by step guide to phylogeny reconstruction. *The Plant Journal*, *45*(4), pp.561-572.

Jukes, T.H. and Cantor, C.R., 1969. Evolution of protein molecules. *Mammalian protein metabolism*, *3*, pp.21-132.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., Von Haeseler, A. and Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, *14*(6), pp.587-589.

Kingdon, J., 1982. East African mammals: an atlas of evolution in Africa, volume 3. *Academic Press, London.*

Kling, D., Tillmar, A., Egeland, T. and Mostad, P., 2015. A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations. *International journal of legal medicine*, *129*(5), pp.943-954.

Korneliussen, T.S., Albrechtsen, A. and Nielsen, R., 2014. ANGSD: analysis of next generation sequencing data. *BMC bioinformatics*, *15*(1), pp.1-13.

Lamichhaney, S., Berglund, J., Almén, M.S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerová, M., Rubin, C.J., Wang, C., Zamani, N. and Grant, B.R., 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*(7539), pp.371-375.

Lande, R., 1976. Natural selection and random genetic drift in phenotypic evolution. *Evolution*, pp.314-334.

Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, *25*(14), pp.1754-1760.

Li, H. and Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), pp.493-496.

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), pp.2987-2993.

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

Malcolm, J. and Evangelista, P., 2005. The range and status of the Mountain Nyala. *Redlands and Fort Collins, USA.*

Malinsky, M., Matschiner, M. and Svardal, H., 2021. Dsuite-fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, *21*(2), pp.584-595.

Malinsky, M., Svardal, H., Tyers, A.M., Miska, E.A., Genner, M.J., Turner, G.F. and Durbin, R., 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature ecology & evolution*, *2*(12), pp.1940-1955.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), pp.10-12.

Martin, S., 2018. Genomics_general. Available at: http://github/simonmartin/genomics_general. Accessed on: 14 08 2020.

Masel, J., 2011. Genetic drift. *Current Biology*, *21*(20), pp.R837-R838.

Matthee, C.A. and Robinson, T.J., 1999. Cytochrome *b* Phylogeny of the family Bovidae: resolution within the alcelaphini, antilopini, neotragini, and tragelaphini. *Molecular phylogenetics and evolution*, *12*(1), pp.31-46.

Matthee, C.A., Burzlaff, J.D., Taylor, J.F. and Davis, S.K., 2001. Mining the mammalian genome for artiodactyl systematics. *Systematic biology*, *50*(3), pp.367-390.

Matthee, C.A., Van Vuuren, B.J., Bell, D. and Robinson, T.J., 2004. A molecular supermatrix of the rabbits and hares (Leporidae) allows for the identification of five intercontinental exchanges during the Miocene. *Systematic biology*, *53*(3), pp.433-447.

Mayr, E., 1954. Change of genetic environment and evolution.

McCombie, W.R., McPherson, J.D. and Mardis, E.R., 2019. Next-generation sequencing technologies. *Cold Spring Harbor perspectives in medicine*, *9*(11), p.a036798.

McVean, G., 2009. A genealogical interpretation of principal components analysis. *PLoS genetics*, *5*(10), p.e1000686.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A. and Lanfear, R., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, *37*(5), pp.1530-1534.

Moodley, Y. and Bruford, M.W., 2007. Molecular biogeography: towards an integrated framework for conserving pan-African biodiversity. *PloS one*, *2*(5), p.e454.

Moodley, Y., Bruford, M.W., Bleidorn, C., Wronski, T., Apio, A. and Plath, M., 2009. Analysis of mitochondrial DNA data reveals non-monophyly in the bushbuck (Tragelaphus scriptus) complex. *Mammalian Biology*, *74*(5), pp.418-422.

Mostert, R. and Hoffman, L.C., 2007. Effect of gender on the meat quality characteristics and chemical composition of kudu (Tragelaphus strepsiceros), an African antelope species. *Food Chemistry*, *104*(2), pp.565-570.

Nei, M., 1983. Genetic polymorphism and the role of mutation in evolution. *Evolution of genes and proteins*, *71*, pp.165-190.

Nersting, L.G. and Arctander, P., 2001. Phylogeography and conservation of impala and greater kudu. *Molecular Ecology*, *10*(3), pp.711-719.

Nersting, L.G. and Arctander, P., 2001. Phylogeography and conservation of impala and greater kudu. *Molecular Ecology*, *10*(3), pp.711-719.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. and Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, *32*(1), pp.268-274.

Niemiller, M.L., Fitzpatrick, B.M. and Miller, B.T., 2008. Recent divergence with gene flow in Tennessee cave salamanders (Plethodontidae: Gyrinophilus) inferred from gene genealogies. *Molecular ecology*, *17*(9), pp.2258-2275.

Nuttall, G.H.F., Graham-Smith, G.S. and Pigg-Strangeways, T.S., 1904. Blood immunity and blood relationship: a demonstration of certain blood-relationships amongst animals by means of the precipitin test for blood.

Pappas, L.A., 2002. *Taurotragus oryx*. *Mammalian species*, *2002*(689), pp.1-5.

Pardo-Diaz, C., Salazar, C., Baxter, S.W., Merot, C., Figueiredo-Ready, W., Joron, M., McMillan, W.O. and Jiggins, C.D., 2012. Adaptive introgression across species boundaries in Heliconius butterflies. *PLoS genetics*, *8*(6), p.e1002752.

Pease, J.B. and Hahn, M.W., 2015. Detection and polarization of introgression in a five-taxon phylogeny. *Systematic biology*, *64*(4), pp.651-662.

Pease, J.B. and Rosenzweig, B.K., 2015. Encoding data using biological principles: the Multisample Variant Format for phylogenomics and population genomics. *IEEE/ACM transactions on computational biology and bioinformatics*, *15*(4), pp.1231-1238.

Pickford, M., 1990. Uplift of the roof of Africa and its bearing on the evolution of mankind. *Human Evolution*, *5*(1), pp.1-20.

Posada, D. and Crandall, K.A., 2001. Selecting the best-fit model of nucleotide substitution. *Systematic biology*, *50*(4), pp.580-601.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. and Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), pp.559-575.

Rakotoarivelo, A.R., O'Donoghue, P., Bruford, M.W. and Moodley, Y., 2019. An ancient hybridization event reconciles mito-nuclear discordance among spiral-horned antelopes. *Journal of Mammalogy*, *100*(4), pp.1144-1155.

Ralls, K., 1978. Tragelaphus eurycerus. *Mammalian Species*.

Rambaut, A., 2010. FigTree v1. 3.1 (http://tree. bio. ed. ac. uk/software/figtree/) Institute of Evolutionary Biology. *University of Edinburgh, Edinburgh, United Kingdom*.

Reed, K.E. and Bibi, F., 2011. Fossil Tragelaphini (Artiodactyla: Bovidae) from the Late Pliocene Hadar Formation, Afar Regional State, Ethiopia. *Journal of Mammalian Evolution*, *18*(1), pp.57-69.

Rohland, N., Siedel, H. and Hofreiter, M., 2010. A rapid column-based ancient DNA extraction method for increased sample throughput. *Molecular ecology resources*, *10*(4), pp.677-683.

Rosenblum, E.B., 2006. Convergent evolution and divergent selection: lizards at the White Sands ecotone. *The American Naturalist*, *167*(1), pp.1-15.

Saitou, N. and Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), pp.406-425.

Schield, D.R., Perry, B.W., Adams, R.H., Card, D.C., Jezkova, T., Pasquesi, G.I., Nikolakis, Z.L., Row, K., Meik, J.M., Smith, C.F. and Mackessy, S.P., 2019. Allopatric divergence and secondary

contact with gene flow: a recurring theme in rattlesnake speciation. *Biological Journal of the Linnean Society*, *128*(1), pp.149-169.

Shamimuzzaman, M., Le Tourneau, J.J., Unni, D.R., Diesh, C.M., Triant, D.A., Walsh, A.T., Tayal, A., Conant, G.C., Hagen, D.E. and Elsik, C.G., 2020. Bovine Genome Database: new annotation tools for a new reference genome. *Nucleic acids research*, *48*(D1), pp.D676-D681.

Silakari, O. and Singh, P.K., 2020. Concepts and Experimental Protocols of Modelling and Informatics in Drug Design. *Academic Press*.

Slatkin, M., 2008. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), pp.477-485.

Smith, J.M., 1966. Sympatric speciation. *The American Naturalist*, *100*(916), pp.637-650.

Stanley, S.M., 1975. A theory of evolution above the species level. *Proceedings of the National Academy of Sciences*, *72*(2), pp.646-650.

Stephens, C., 2004. Selection, drift, and the “forces” of evolution. *Philosophy of Science*, *71*(4), pp.550-570.

Tadesse, S.A. and Kotler, B.P., 2013. The impacts of humans and livestock encroachments on the habitats of mountain nyala (*Tragelaphus buxtoni)* in Munessa, Ethiopia. *International Journal of Biodiversity and Conservation*, *5*(9), pp.572-583.

Tadesse, S.A. and Kotler, B.P., 2014. Effects of habitat, group-size, sex-age class and seasonal variation on the behavioural responses of the mountain nyala (*Tragelaphus buxtoni*) in Munessa, Ethiopia. *Journal of tropical ecology*, *30*(1), pp.33-43.

Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, *105*(2), pp.437-460.

Tello, J.L. and Van Gelder, R.G., 1975. The natural history of nyala, *Tragelaphus angasi* (Mammalia, Bovidae) in Mozambique. Bulletin of the AMNH; v. 155, article 4.

Tijskens, J., 1968. Preliminary notes on the F1 Bongo antelope x sitatunga hybrids *Taurotragus eurycerus* x *Tragelaphus spekei* at Antwerp Zoo. *International Zoo Yearbook*, *8*(1), pp.137-139.

Vrba, E.S., 1995. On the connections between paleoclimate and evolution. In Paleoclimate and evolution, with emphasis on human origins (Vrba, E.S., Denton, G.H., Partridge, T.C. and Burckle, L.H. eds.) *Yale University Press*.

Wagner, M., 1873. The Darwinian theory and the law of the migration of organisms. *E. Stanford*.

Walsh, D.M., 2000. Chasing shadows: natural selection and adaptation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *31*(1), pp.135-153.

Willows-Munro, S., Robinson, T.J. and Matthee, C.A., 2005. Utility of nuclear DNA intron markers at lower taxonomic levels: Phylogenetic resolution among nine Tragelaphus spp. *Molecular Phylogenetics and Evolution*, *35*(3), pp.624-636.

Wilson, D.E. and Reeder, D.M. eds., 2005. Mammal species of the world: a taxonomic and geographic reference (Vol. 1). *JHU press*.

Wright, S., 1931. Evolution in Mendelian populations. *Genetics*, *16*(2), p.97.

Wronski, T. and Moodley, Y., 2009. Bushbuck, harnessed antelope or both. *Gnusletter*, *28*, pp.18-19.

Wronski, T., Apio, A., Plath, M. and Averbeck, C., 2009. Do ecotypes of bushbuck differ in grouping patterns?. *acta ethologica*, *12*(2), pp.71-78.

Zhou, Y., Duvaux, L., Ren, G., Zhang, L., Savolainen, O. and Liu, J., 2017. Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. *Heredity*, *118*(3), pp.211-220.