UNIVERSITY OF LEICESTER

Genomic tools for great ape population dynamics and conservation

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

Ettore Fedele

Department of Genetics and Genome Biology
College of Life Sciences, University of Leicester

September 2022

Abstract

Genomic tools for great ape population dynamics and conservation

Ettore Fedele

Gorillas were once abundant in the forests of Central Africa, playing a fundamental role in the maintenance of ecosystems. Habitat loss, poaching and illegal wildlife trade have led populations to the brink of extinction. Genetic analysis of forensic DNA markers can be used to assess the effects of reduced population sizes and the effectiveness of conservation measures. This study aimed to assess whether cross-species genotyping of forensic DNA markers could provide both individual and sub-species identification. Firstly, massively parallel sequencing (MPS, via Illumina technology) analysis of 27 human orthologous aSTRs (autosomal short tandem repeats, via the Verogen ForenSeq™ DNA Signature Prep Kit) was applied to 52 individuals (14 chimpanzees; 4 bonobos; 16 western lowland; 6 eastern lowland, and 12 mountain gorillas). Thirteen loci could be genotyped across all individuals, and were individually and sub-species identifying. This showed that allelic diversity and the power to discriminate sub-species were greater when considering STR and flanking sequences rather than allele lengths alone, which are conventionally typed by capillary electrophoresis (CE). When comparing with humans, interruptions within long repeat arrays in African great apes did not appear to reduce allelic diversity, indicating a possible mutational difference to humans. Secondly, portable sequencing technology (via Oxford Nanopore Technologies Ltd sequencing) was exploited to investigate a panel of about 90 single nucleotide polymorphisms (SNPs) designed *ad hoc* for individual and sub-species identification in gorilla from non-invasive (*i.e.* poor quality) DNA samples (*e.g.* hair and faeces). A core set of 42 loci amplified across all individuals (34 western lowland; 6 eastern lowland, and 6 mountain gorillas) and provided both individual and sub-species identification. This approach helps overcome the current lack of well-established research facilities in habitat countries while facilitating genetic research and thus aid gorilla conservation.

Acknowledgements

First and foremost, I would like to thank my supervisors Prof Mark Jobling, Dr Jon Wetton, and Dr Celia May for the opportunity to embark on this incredible journey. When I started, as an economics and ecology graduate, my knowledge of the subject was (*euphemistically*) extremely limited. Their support, passion, and knowledge have inspired and motivated me all the way since the very the start. While it has been challenging at times, it has also been the most constructive experience and, for that, I will forever be grateful. Thank you for your immense patience and understanding.

I would like to thank all members of Mark Jobling's group, their support and friendship have helped me beyond words. Thank you Emily, Yahya, Jodie, Margherita, Tunde, Orie and Jordan. Many thanks to Rita Neumann, your expertise and knowledge can only be matched by your kindness, patience and great sense of humour. Thanks to all members of Ed's and Yuri's labs too. I would also like to take the chance to express my gratitude to my panel review committee (Dr Robert Hammond and Dr Edward Hollox) for taking the time to comment on my work. I would like to thank all collaborators. Wellcome Sanger Institute (Chris Tyler-Smith, Yali Xue and Pille Hallast) for sharing data and precious samples with us. I am immensely grateful to Lisa Gillespie (Twycross Zoo) and Tony King (Aspinall Foundation) for providing me with additional gorilla samples. Thanks to NERC – CENTA for funding this project and giving me the opportunity to learn a great deal about rocks and volcanos. I did enjoy it!

I would like to thank all my friends in the UK, Italy, Ireland, and farther afield for their support during these last years. Despite the distance and the pandemic, I have deeply felt your affection and support. You have all been my lighthouse and inspiration, and I am proud of you all. Thanks to all the people that have instilled in me a great passion for wildlife, Caz Schiess (WEI), Dr Alex Monro (RBG Kew), Dr Lynne MacTavish (Makwe Game Reserve), Kuki Gallmann (Ol Ari Nyiro Ranch), and Dr Patrick White (ENU). Your efforts and commitment are the best example.

Thanks to my family, for their patience, support, and love. Thanks to Giorgia, with whom I have shared a third of my life and with whom I wish to spend the rest of it. *Grazie!*

List of abbreviations

AIM	Ancestry Informative Marker
aSTR	Autosomal short tandem repeat
BAM	Binary Alignment Map
BIC	Bayesian information criterion
BINP	Bwindi Impenetrable National Park
BOLD	Barcode of Life Data System
bp	Base pairs
CE	Capillary electrophoresis
CITES	Convention on International Trade in Endangered Species of Wild Fauna and Flora
CNV	Copy number variation
COXI	Cytochrome c oxidase subunit 1
DAPC	Discriminant Analysis of Principal Components
df	degree of freedom
DNA	Deoxyribonucleic acid
EB	Elution buffer
FB	Flush buffer
F_{IS}	inbreeding coefficient
FLT	Flush tether
FS	Full siblings
Gb	Giga base
Gbb	*Gorilla beringei beringei*
Gbg	*Gorilla beringei graueri*

Ggg	*Gorilla gorilla gorilla*
H_E	Expected heterozygosity
H_O	Observed heterozygosity
HS	Half siblings
HWE	Hardy-Weinberg equilibrium
IBD	Identity-by-Descent
*ii*SNP	Individual identification SNP
IUCN	International Union for Conservation of Nature
LBII	Loading beads II
LD	Linkage disequilibrium
MCMC	Markov Chain Monte Carlo
mg	milligram
MPS	Massively parallel sequencing
MSY	Male-specific region of the Y chromosome
mtDNA	Mitochondrial DNA
N	Genotype count
Nall	Allele count
NCBI	National Centre for Biotechnology Information
ng	Nanogram
ONT	Oxford Nanopore Technologies Ltd
Pab	*Pongo abelii*
PCA	Principal Component Analysis
PCR	Polymerase chain reaction
PD	Power of discrimination

PE	Power of exclusion
PIC	Polymorphism information content
PM	Match probability
PO	Parent-offspring
Ppa	*Pan paniscus*
Ptr	*Pan troglodytes*
QC	Quality check
SBII	Sequencing buffer II
SD	Standard deviation
SFB	Short fragment buffer
*si*SNP	Species identification SNP
SNP	Single nucleotide polymorphism
STR	Short tandem repeat
TPI	Typical paternity index
UAS	Universal Analysis Software
UN	United Nations
VCF	Variant calling format
VM	Virunga mountain
µl	Microliter
µM	Micromolar

Table of contents

Table of Figures

Table of Tables

Chapter 1 Introduction

1.1 Background

1.1.1 Overview of the ecology and conservation of gorillas

It is now widely accepted that unprecedented rates of biodiversity loss worldwide represent a major threat to a sustainable future for ecosystems and local natural and human communities (Zimov *et al.*, 1995; Pimm *et al.*, 2014; Tong *et al.*, 2022). Owing to the tightly linked interactions that make up the living world around us, anticipating and understanding the consequences of species loss remain a difficult issue for researchers and conservationists; yet there is a growing concern that changes in species abundance can trigger important knock-on effects, disrupting the mechanisms that support inter-specific coexistence, resulting in a sequence of additional species extinctions (Sanders *et al.*, 2015; Donohue *et al.*, 2017). It is therefore of primary importance to equip conservationists with operational tools to assess risks for species survival in order to design and adopt effective countermeasures to control species loss, hence limiting the detrimental consequences that would follow (Evans and Rittenhouse, 2018).

In spite of species extinction being a global concern, it is local populations, not species, that are the focus of conservation management. This is because it is at the local scale that the effects of extinctions have the potential to affect the integrity of the ecosystem and the livelihood of the communities that depend on it (Zimov *et al.*, 1995; Hooper *et al.*, 2005; Ross *et al.*, 2021; Hallast and Jobling, 2017). Local changes in species richness and abundance are naturally more common than global extinctions and also particularly difficult to reverse (Razgour *et al.*, 2018) both in managed and unmanaged ecosystems (Hooper *et al.*, 2005). It is thus crucial for conservationists to act promptly to limit changes in local biodiversity, thereby preventing an escalation of detrimental effects on the entire ecosystem.

The rate at which natural populations go extinct is the result of the combination of the intensity of local threats and species biology. The aspects of species biology that should be considered in designing conservation plans include habitat, home-range area, ecological flexibility, diet, body size, sociality and population genetic variability (Isaac and Cowlishaw, 2004). For instance, a study conducted on primate species shows that

large-bodied populations occurring in small areas are particularly vulnerable to hunting, since their body-size makes them conspicuous and the restricted area of occupancy makes their movements predictable, becoming an easy target for hunters and poachers alike (Isaac and Cowlishaw, 2004). In this regard, the reduced number of gorillas is an emblematic example of the effects of the combination of specific biology characteristics and local threats (*e.g.* illegal hunting) on population survival and viability. There are currently two species of gorillas, namely western (*G. gorilla*) and eastern (*G. beringei*) gorilla, each subdivided further into two sub-species: western lowland (*G. g. gorilla*) and Cross River (*G. g. diehli*), and eastern lowland (*G. b. graueri*) and mountain gorilla (*G. b. beringei*). With the exception of mountain gorilla, which is currently classified as "Endangered" (Hickey *et al.*, 2018) by the International Union for Conservation of Nature (IUCN) Red List of Threatened Species, all other sub-species are currently "Critically Endangered" due to continuous decline in their populations (Plumptre *et al.*, 2016a; Maisels *et al.*, 2017; Bergl *et al.*, 2016). Different gorilla sub-species distribution is displayed in Figure 1.1.

Mountain gorillas occur in only two populations; one is found in the Virunga Massif ecoregion (VM, including the Southern sector of the Virunga National Park and the Volcano National Park, see Figure 1) on the border with the Democratic Republic of the Congo, Rwanda and Uganda, and the other is found in the Bwindi Impenetrable National

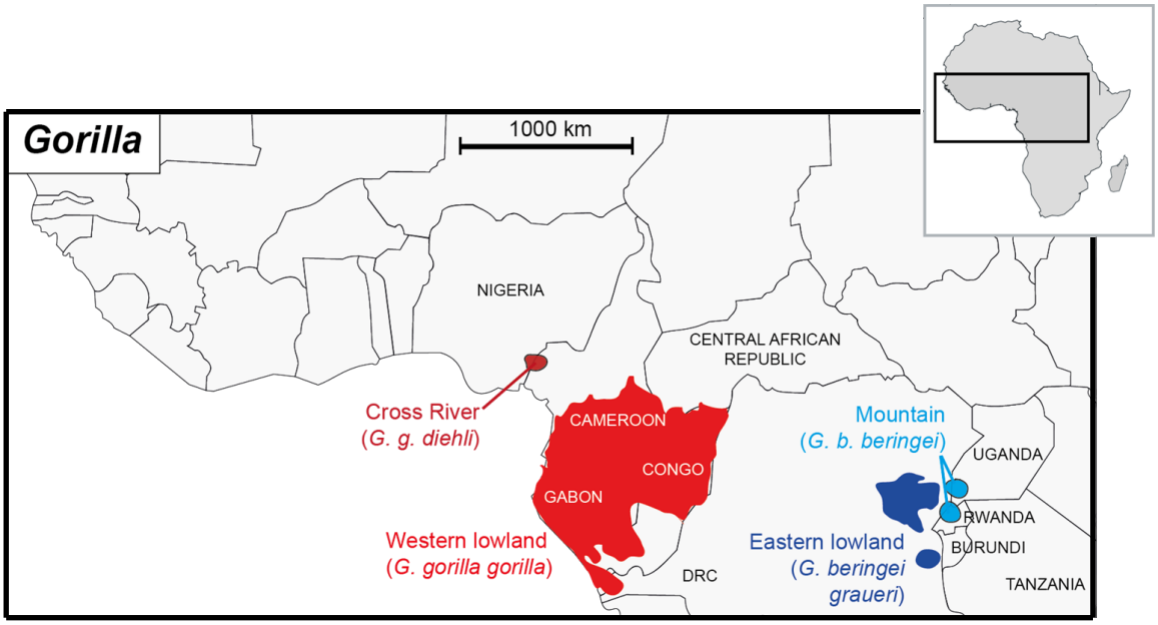


Figure 1.1: Gorilla sub-species distribution. Dark red (*G. g. diehli*), red (*G. g. gorilla*), dark blue (*G. b. graueri*) and blue (*G. b. beringei*). Map adapted from Africa just countries grayish.svg, published on Wikimedia under a Creative Commons Attribution-Share Alike 4.0 International license. DRC: Democratic Republic of the Congo.

Park (BINP) in Uganda (Figure 1.1). Estimates indicate that there are about 1000 mountain gorillas left in the wild, with just over 480 individuals left in the VM (Gray *et al.*, 2013) and some 340 in BINP (Guschanski *et al.*, 2009). The main threats to mountain gorillas are poaching (Yamagiwa *et al.*, 1993; Kühl, 2008; Estrada *et al.*, 2017), infectious diseases (*e.g.* Ebola) (Gilardi *et al.*, 2015; Rwego *et al.*, 2008; Estrada *et al.*, 2017), and habitat loss and degradation (Kühl, 2008; Xue *et al.*, 2015; Estrada *et al.*, 2017).

With agriculture being the major economic activity in the region, one of the most densely populated in Africa, mountain gorilla populations are increasingly isolated, and numerous conflicts with humans arise as they venture outside protected areas in search for food, mates and new areas to colonise (Kayitete *et al.*, 2019). Agriculture is also responsible for preventing the recruitment of important plant species (*i.e.* a crucial step for forest regeneration) upon which gorillas depend for food and shelter (Kayitete *et al.*, 2019), hence adding to the aforementioned issues. The ecological implications of the loss of this species are yet to be fully understood; this is because mountain gorillas are a keystone species, providing a vital role for the integrity and maintenance of the ecosystem and the processes therein. Gorillas are indeed fundamental seed dispersers moving long distances in search for food and maintaining habitat heterogeneity through the opening up of thick woodland and the creation of paths that allow other species to thrive in the same area (Nkurunungi *et al.*, 2004; Haurez *et al.*, 2018). In this regard, Effiom *et al.* (2013) found evidence that hunting of large-bodied seed dispersers (including Cross river gorillas) is likely to lead to drastic changes in tree community over few generations

Despite the close proximity of the two areas of occupancy of mountain gorillas, VM and BINP, there are significant ecological differences between the two sub-populations that were first observed by Sarmiento et al. (1996) and that should be considered while studying the species in the wild. In particular, it was found that the Bwindi population prefers feeding on fruit whereas the Virunga gorillas were found to feed predominantly on bamboo (Table 1); this is mirrored by different habitat preferences – the altitudinal range between the BINP and VM populations varies remarkably, occurring between 1,500 – 2,300 and between 2,200 – 4,000 metres above sea level respectively (Elliott, 1976; Butynski and Kalina, 1993; Sarmiento, 1996; Watts, 1994). Some behavioural differences were also noted: 54% of the Bwindi gorillas were seen nesting on or close to the ground and nests were rarely defecated in, whereas up to 97% of the Virunga gorillas would nest on the ground or in its proximity and nests were often defecated in and were built in

bamboo. Furthermore, in BINP gorillas would often use trees for feeding and nesting, while in Virunga gorillas would only occasionally visit trees for feeding and even more rarely to build nests (Sarmiento, 1996). This information is relevant to field work operations for the collection of non-invasive samples (the subject of this project) and for the later development of effective conservation plans aimed at protecting some of the most emblematic species of megafauna still roaming our planet.

Table 1.1: Dietary differences between BINP and VM mountain gorilla (*G. b. beringei*) populations (adapted from Sarmiento (1996)).

Diet items	Bwindi (BINP)	Virunga (VM)
Leaves	<10.0	67.7
Stems	<10.0	25.0
Pith	>30.0	2.4
Fruit	>20.0	0.2
Other*	–	4.7

** Other includes invertebrates such as termites, ants, and snails*

The eastern lowland gorilla (or Grauer's gorilla, *G. b. graueri*), the larger of the two sub-species of eastern gorilla, is endemic to the evergreen low elevation rainforests and high elevation mountain forests in the eastern Democratic Republic of the Congo (Plumptre *et al.*, 2016b; Van der Hoek *et al.*, 2021) (Figure 1.1). Estimates reveal a decline of 77% of the population over recent decades (Plumptre *et al.*, 2016b; Baas *et al.*, 2018), in addition, years of civil conflict have made research in the region extremely challenging so little is known about this subspecies (Van der Hoek *et al.*, 2021). The sparse information available on its ecology is the result of studies of only two recently-habituated groups of eastern lowland gorilla occurring at high elevation (2000 m) in Khuzi-Biega National Park and on Mt. Tshiaberimu (Yamagiwa and Basabose, 2006; Maldonado *et al.*, 2012; Yamagiwa *et al.*, 2012), which make up only 2% of the 3800 individuals estimated to exist in 2016 (Plumptre *et al.*, 2016b). Studies of individuals occurring at lower elevations are limited to census surveys and opportunistic counts (Van der Hoek *et al.*, 2021). As a result, there is still a dearth of knowledge regarding their behaviour, dispersal rate, population dynamics and structure. Similarly, the subspecies has been subject to limited genetic research; indeed, the first complete mitochondrial genome of eastern lowland gorilla was only published in 2016 (Baas *et al.*, 2018; Hu and Gao, 2016). In terms of genetic diversity, previous studies have revealed that eastern lowland gorillas are more diverse than mountain gorillas but less diverse than western lowland gorillas (Baas *et al.*, 2018), and that conservation attention should aim at preserving peripheral populations

and support habitat connectivity, so as to ensure the long-term viability of these animals (van der Valk *et al.*, 2018).

Of the two sub-species of western gorillas, Cross River gorilla (*G. g. diehli*) has the smallest population size. In fact, with only 250-300 mature individuals left in the wild, these are the rarest of all African great apes (Wade and Malone, 2021), occurring in only three genetically distinct populations along the border between Cameroon to the South and Nigeria to the North (Bergl *et al.*, 2008) (Figure 1.1). Analysis of genetic diversity has shown that Cross River gorilla populations compared well with mountain gorillas, registering a lower level of diversity than the other western sub-species (*i.e.* western lowland gorilla) (Bergl *et al.*, 2008). Continuous hunting and habitat fragmentation are the major drivers of population decline and hence of low genetic diversity. Despite various conservation efforts and, to a certain extent, evidence of gene flow across populations (Bergl and Vigilant, 2007), there have been increasing signs of forest fragmentation over recent years, which weakens the dispersal possibility, hence hindering inter-group gene flow (Fitz *et al.*, 2022), further reducing global population diversity.

With an estimated population of about 360,000 individuals, western lowland gorillas (*G. g. gorilla*) are the most numerous of all four sub-species (Strindberg *et al.*, 2018). However, owing to dramatic population declines following enormous die-offs caused by Ebolavirus outbreaks (Rizkalla *et al.*, 2007; Forcina *et al.*, 2019), and further predicted future declines (Rizkalla *et al.*, 2007), they have been listed as "Critically Endangered" according to the IUCN Red List (Maisels *et al.*, 2017). In addition, western lowland gorillas are particularly vulnerable to logging and poaching as 77% of their range falls outside protected areas (Strindberg *et al.*, 2018), where they occur in low elevation forests and swamps of central Africa (Figure 1.1). Groups are formed by one fully mature male and several adult females with their offspring, or by non-sexually mature individuals (Magliocca *et al.*, 1999; Gatti *et al.*, 2004; Robbins *et al.*, 2004; Forcina *et al.*, 2019). However, owing to the high mobility and lower observability, which impedes simultaneous observations of different groups, both structure and social group dynamics in western lowland gorillas remain poorly understood (Doran-Sheehy *et al.*, 2004; Douadi *et al.*, 2007). The information gathered on social interactions among western lowland gorillas is almost completely limited to studies conducted in "bais" (Magliocca *et al.*, 1999; Gatti *et al.*, 2004; Robbins *et al.*, 2004; Parnell, 2002; Levréro *et al.*, 2006; Forcina

et al., 2019), which are swampy clearings in the thick of the forest where different groups meet to feed on the salt-rich grasses (Metsio Sienne *et al.*, 2014). Nevertheless, comparative behavioural studies have revealed that, as opposed to the more aggressive mountain gorilla (Forcina *et al.*, 2019; Jeffery *et al.*, 2007a), inter and intra-group interactions in western lowland gorillas are often non-aggressive (Doran-Sheehy *et al.*, 2004). As a result infanticide has never been observed in this sub-species (Robbins *et al.*, 2004; Stokes *et al.*, 2003) nor have group takeovers by outside males (Robbins *et al.*, 2004; Stokes *et al.*, 2003; Gatti *et al.*, 2003). Furthermore, in juxtaposition to mountain gorillas where groups can contain several reproductive males (or silverbacks) and more than 15% of the offspring are not sired by the dominant male (Bradley *et al.*, 2005; Forcina *et al.*, 2019), western lowland gorilla social structure is characterised by the presence of a single silverback, who fathers the majority of the infants (Forcina *et al.*, 2019). Because of obligate dispersal by both sexes at maturity and a high degree of tolerance (Doran-Sheehy *et al.*, 2004), and a high rate of acceptance of members from other groups, western lowland gorilla groups are expected to be more genetically diverse than the more aggressive and territorial mountain gorillas (Xue *et al.*, 2015). This was indeed discussed in two important studies in which genetic diversity is compared across gorilla species (Prado-Martinez *et al.*, 2013; Xue *et al.*, 2015).

1.1.2 Genetic diversity and the effects of small population size

The factors limiting genetic variation in small fragmented wildlife populations are random genetic drift and inbreeding (Ouborg *et al.*, 2006; Ouborg *et al.*, 2010; Frankham *et al.*, 2017). Genetic drift – random fluctuations of allele frequencies over time and space – can lead to random loss of adaptive alleles and fixation of deleterious alleles in the population, whereas inbreeding, often referred to as biparental inbreeding or consanguinity (with closely related parents), can increase homozygosity in the population (Ouborg *et al.*, 2010; Jobling *et al.*, 2013).

Genetic drift and inbreeding events can have far-reaching consequences for the survival of threatened populations. High levels of homozygosity and increased frequencies of deleterious alleles often lead to inbreeding depression – an overall reduction of individual fitness – which further reduces the short-term viability of a population (Ouborg *et al.*, 2010; Frankham *et al.*, 2017). For instance, a recent study on the effects of inbreeding depression on a threated population of mountain lions in Southern California, has found

phenotypical evidence of both physical abnormalities (*i.e.* kinked tail) and reproductive problems (*e.g.* abnormal sperm production) in highly inbred individuals, raising further concern for the survival of these animals (Huffmeyer *et al.*, 2022). In addition, loss of genetic diversity can compromise the evolutionary adaptive potential of a population, thereby reducing its long-term viability, and exacerbating the impacts of stochastic events such as diseases and climate change (Ouborg *et al.*, 2010; Gilardi *et al.*, 2015). Finally, as isolation between populations increases, the genetic divergence between them increases as well as a result of independent genetic drift; this, in turn, increases the risk of outbreeding depression (*i.e.* reduction in fitness of offspring of an outcross between two distinct populations compared to the fitness of offspring of an outcross within a single population), whenever individuals belonging to different populations meet and interbred (Ouborg *et al.*, 2010; Frankham *et al.*, 2017), although this has rarely been demonstrated. For these reasons, it is key to investigate the genetic diversity and population history of species and identify those factors that underpinned their evolution.

The most comprehensive study to date into the diversity and relationships of great-ape populations was conducted by Prado-Martinez *et al.* (2013), who sequenced the whole genomes of examples of all species of great apes to high coverage, supporting the taxonomy of genetically distinct populations within species (Figure 1.2, 1.3) Results generated from SNPs (Single Nucleotide Polymorphisms) and AIMs (Ancestry Informative Markers) analyses show that, compared to other great-ape genera (Pan – chimpanzee and bonobo; and *Pongo* – orang-utan), gorillas have a simpler evolutionary history, consisting of only two species, the Western (*Gorilla gorilla*) and Eastern gorilla (*Gorilla beringei*), both occurring in equatorial Africa and separated by about 1000 km (Thalmann *et al.*, 2006; Groves, 2001; Grubb *et al.*, 2003; Prado-Martinez *et al.*, 2013; Xue *et al.*, 2015).

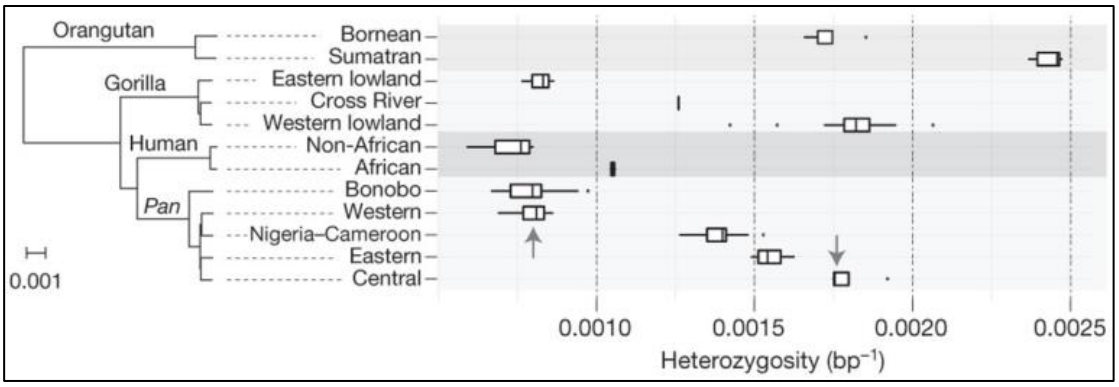


Figure 1.3: Heterozygosity estimates of each great-ape species (Prado-Martinez *et al.*, 2013). Heterozygosity estimates of each of the individual species and subspecies are superimposed onto a neighbour-joining tree from genome-wide genetic distance estimates (branch lengths in units of substitutions per bp). Arrows indicate heterozygosities previously reported for western and central chimpanzee populations. Non-African humans, eastern lowland gorillas, western chimpanzees, and bonobos show the lowest genetic diversity ($\sim 0.8 \times 10^{-3}$ heterozygotes per bp), whereas orangutans, western gorillas, and central chimpanzees the highest levels of genetic diversity ($1.6 \times 10^{-3} - 2.4 \times 10^{-3}$ heterozygotes per bp) (see Prado-Martinez *et al.*, 2013).

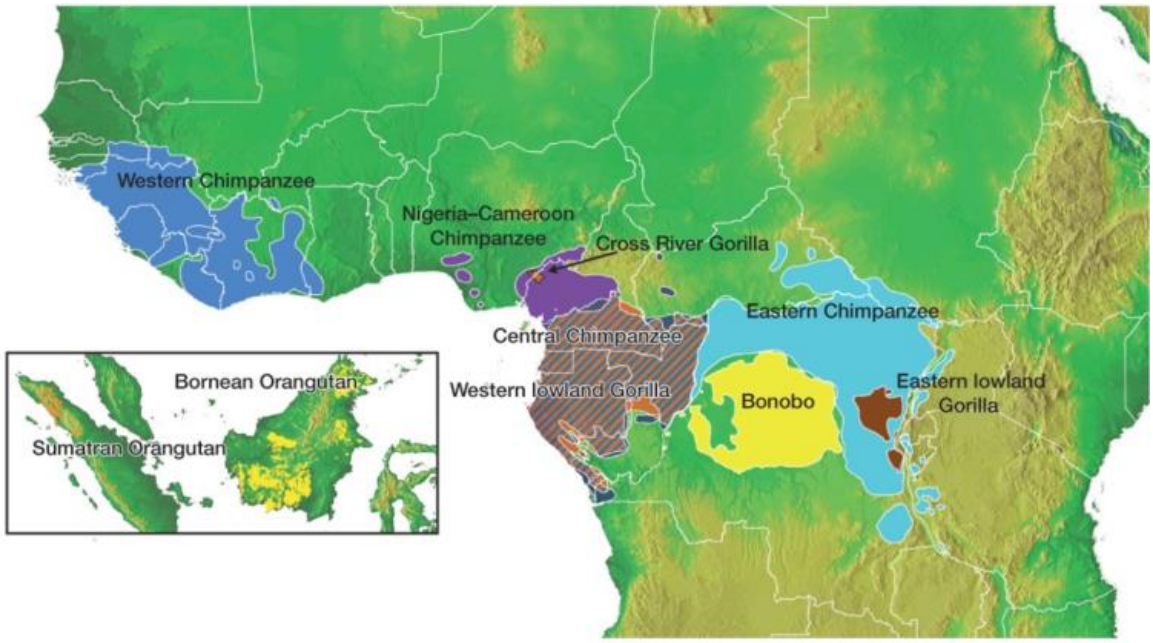


Figure 1.2: Geographic distribution of great-ape species (Prado-Martinez *et al.*, 2013).

Both Eastern lowland and Mountain gorilla subspecies have small effective population sizes (Ne = 290 ± 18 and Ne = 273 ± 54 respectively), and thus lower genomic variation, than other subspecies in their genus (Thalmann *et al.*, 2011; Prado-Martinez *et al.*, 2013; Xue *et al.*, 2015; Marques-Bonet and Hvilsom, 2018). A recent study has reported that the decline in genetic diversity of Eastern gorilla species was the result of the past extirpation of populations that were outside the current range of the species (van der Valk *et al.*, 2018), *i.e.* peripheral populations that once allowed enough gene flow for small populations to survive. This is the result of repeated ancient bottlenecks and founder effects that have hampered genetic diversity, which have also led to an increase in long

runs of homozygosity in their genomes (Marques-Bonet and Hvilsom, 2018). As a consequence of reduced population size, mountain gorillas show very low genetic diversity and an increased overall burden of deleterious variation (Xue *et al.*, 2015).

Recent whole genome sequencing analysis has showed that mountain gorilla populations have experienced continuous decline over the past 100,000 years, and that further recent decline has led to extensive inbreeding, such that individuals are homozygous at 34.5 – 38.4% of their sequence (Xue *et al.*, 2015; Marques-Bonet and Hvilsom, 2018). These levels of homozygosity are much higher than in western lowland gorillas (13.8%), and exceed that of even the most inbred human population (Pemberton *et al.*, 2012), although these have accumulated over long periods of time (Xue *et al.*, 2015). Furthermore, evidence suggests that a further recent decline in mountain gorillas has contributed extensively to the high rate of inbreeding among populations but it has also led to the purging of severely recessive deleterious mutations (Xue *et al.*, 2015). The effects of inbreeding are visible and cases of syndactyly (possibly due to homozygosity for recessive mutations) have already been reported (Xue *et al.*, 2015).

Recent findings confirmed that the divergence between the two gorilla species (*i.e.* Eastern gorilla – *Gorilla beringei spp*., and Western gorilla – *Gorilla gorilla spp*.) began some 150,000 years ago, although gene flow between the two may have lasted until 20,000 years ago (Xue *et al.*, 2015). Previous analyses, however, have indicated an initial population split at around 0.9 – 1.6 MYA, with simulations revealing that the majority of gene flow took place from eastern to western gorilla populations rather than vice versa (Thalmann *et al.*, 2006). Remarkably, this period of most recent gene flow between populations coincides with the Last Glacial Maximum (26,000 – 19,000 years ago) during which dry savannah replaced vast areas of primary rain forest, causing a collapse in the western population and hence completing the separation of the two species (Prado-Martinez *et al.*, 2013; Xue *et al.*, 2015). A comparison of the analyses of variation in mitochondrial DNA (mtDNA) and the Y chromosome (see Box 1), for reconstruction of species lineages, showed very little diversity in mountain gorillas, with only three and two haplotypes respectively (Xue *et al.*, 2015).

Box 1: Mitochondrial DNA (mtDNA) and the Y chromosome

Both mtDNA and Y-chromosome are inherited in a haploid manner: mtDNA through the female and the Y through the male lineages. The maternal inheritance of the mtDNA is ensured by a mechanism of proteolysis of the sperm midpiece during embryogenesis, which is species-specific and based on ubiquitination of the mitochondria in spermiogenesis (Cummins, 2001). In 2003, Hebert et al. introduced the analysis of a 648-bp region of the *cytochrome c oxidase subunit 1* (*COXI*) mitochondrial gene for evolutionary studies. Indeed, the region possesses several advantages: it is a short segment of DNA, therefore easy to isolate and sequence; it is ubiquitous in all animal species; it expresses significant species-specific variability, and it is flanked by highly conserved regions, facilitating primer design (Kress and Erickson, 2012). The resulting technique is widely known as species barcoding (Hebert *et al.*, 2003; Taberlet *et al.*, 1999). For the Y chromosome, the accumulation of genes that enhance male reproductive fitness has led to a rapid evolution across species (Bellott *et al.*, 2014; Hallast and Jobling, 2017). The haploid state and the lack of crossing over have led to degeneration of the male- specific region (MSY). Any intra-specific differences within the MSY can be attributed the accumulation of mutations, such as single nucleotide polymorphisms (SNPs) or short tandem repeats (STRs or microsatellites) (see Box 3), and these mutations can facilitate the differentiation of individuals into paternal lineages (Jobling and Tyler-Smith, 2003). Both MSY and mtDNA markers serve as universal genetic means to reclassify formally taxonomically defined mammalian species. The combination of both genetic facets allows to obtain a high-resolution picture of species dispersal times, sex- biased behaviours, and group dynamics.

By performing whole-genome sequencing and analysis for 13 eastern gorillas, Xue *et al.* (2015) have compiled a list of 25,628 autosomal and 89 mtDNA ancestry-informative SNP markers (AIMs) uniquely identifying eastern lowland and mountain gorillas, as well as 1127 lineage-specific CNV loss and gain events (Xue *et al.*, 2015). Careful consideration and detailed knowledge of the genetic diversity and structure of the focus species is central to an effective management of the small remnant populations of eastern gorillas (Marques-Bonet and Hvilsom, 2018). The available information regarding genetic variation by great-ape species and subspecies is summarised in Table 1.2.

Table 1.2: Genetic variation summary by species and subspecies (adapted from Prado-Martinez *et al.* (2013))

Genus	Scientific name species/subspecies	Common name	N	AIMs*
Homo	*Homo sapiens*	Non-African	6	12,316
		African	3	12,316
		Humans	9	NA
Pan	*Pan troglodytes ellioti*	Nigeria-Cameroon	10	2213
	Pan troglodytes schweinfurthii	Eastern	6	1265
	Pan troglodytes troglodytes	Central	4	619
	Pan troglodytes verus	Western	4	145,548
		Common Chimpanzees	24	149,645
Pan	*Pan paniscus*	Bonobos	13	NA
Gorilla	*Gorilla beringei beringei*	Mountain	7	25,717
	Gorilla beringei graueri	Eastern lowland	9	317,028
	Gorilla gorilla diehli	Cross river	1	35,693
	Gorilla gorilla gorilla	Western lowland	23	19,902
		Gorillas	40	398,340
Pongo	*Pongo abellii*	Sumatran	5	1,132,808
	Pongo pygmaeus	Bornean	5	1,132,808
		Orangutans	10	NA
		All	96	NA

* Ancestry informative markers: variants only found in a single group within each species

1.1.3 Past research and future developments for population surveys and monitoring

Following the advances in molecular genetics in the 1980s, and the availability of neutral molecular markers, analyses of the impact of genetic drift, inbreeding, and gene flow between and within populations were finally possible (Ouborg *et al.*, 2010). However, it was only with the introduction of non-invasive sampling and DNA amplification for paternity exclusion and the analysis of community structure, and the subsequent technological developments of the early '90s that genetic studies of free-ranging animal species became an important field in conservation biology (Morin *et al.*, 1993; da Silva *et al.*, 2012). These studies were impeded by obvious difficulties of sampling, storing and transportation of tissues, which were first overcome by Morin et al. (1993) who used non-invasive genetics to sequence hair samples for the analysis of patterns of gene flow in the Gombe chimpanzee community (Tanzania) (Morin *et al.*, 1993).

Specific to the mountain gorilla populations of Bwindi and Virunga, two independent studies were published in 2009 and 2013, where non-invasive faecal sample sequencing techniques (16 autosomal STR loci analysed using Capillary Electrophoresis [CE]) were exploited to estimate population densities in the two areas but often not with comprehensive sampling (Guschanski *et al.*, 2009; Gray *et al.*, 2013). Comparisons of the information obtained through this genetic census showed discrepancies with previously collected data where estimates were derived by means of traditional field methods (*e.g.* nest counting). On average, traditional methods yielded a higher rate of miscounting and were unable to accurately differentiate different groups within populations which, by contrast, was successfully done through individual profiling using genetic analysis (Guschanski *et al.*, 2009; Gray *et al.*, 2013). A virtually identical approach was adopted by Hagemann *et al.* (2018) to study long-term group membership and dynamics in a wild western lowland gorilla population in Loango National Park, Gabon.

Despite the importance of mtDNA and autosomal STR (aSTR) markers to the understanding of the evolution of species, there is a growing concern that these genomic sequences hold some significant drawbacks (Marques-Bonet and Hvilsom, 2018). In particular, mtDNA, which is inherited maternally (see Box 1), reflects only one lineage in a population's history, and thus just one realisation of the evolutionary process. On the other hand, for aSTRs, which are biparentally inherited, there are too few markers that

have been developed for the analysis of non-human great-ape species population histories. In addition, STR analysis has commonly been conducted via CE only, generating length-based information results, and thus failing to deal with such issues as homoplasy – alleles with different evolutionary histories, shared by different species or individuals, that share the same length but may be characterised by different structures.

For instance, past chimpanzee population genomic studies based on mtDNA (Stone *et al.*, 2010; Hvilsom *et al.*, 2014; Marques-Bonet and Hvilsom, 2018; Lobon *et al.*, 2016) or aSTR data (Becquet *et al.*, 2007) were not conclusive in the taxonomic identification of two subspecies of Central chimpanzee (*Pan* troglodytes troglodytes) and Eastern chimpanzee (Pan troglodytes schweinfurthii). Novel technologies have revolutionised the way in which population genetic studies are conducted; screening of whole genome sequences has disclosed the complete genome variation at the population scale, providing a more comprehensive view of the diversity, evolutionary relationships, population structure, admixture and even assessments of relatedness and inbreeding (Marques-Bonet and Hvilsom, 2018). In this regard, based on the knowledge from complete genomes (Prado-Martinez *et al.*, 2013), a panel of 60,000 SNPs has been compiled to capture the regions in the chimpanzee genome that have been identified as being highly informative about subspecies ancestry (AIMs) (Marques-Bonet and Hvilsom, 2018). In addition, new sequencing technologies allow the structure of STRs to be assessed, thereby overcoming the issues deriving from homoplasy, and enabling researchers to distinguish between isoalleles – alleles identical by length but different in sequence.

This approach has now been adopted to target other species' conservation-related questions. In fact, the conservation of great-ape species involves a significant amount of *ex situ* actions, the majority of which take place in zoos exploiting pedigree information for captive breeding programmes. This is often problematic due to gaps in pedigrees and the numerous assumptions around the relatedness among founders that may lead to more inbreeding or outbreeding in managed and captive populations. The application of AIMs can help assess the accuracy of pedigree records, identify hybrids between sub- species, and obtain a better overview of ancestry and genetic structure, hence allowing zoos to re-consider how management plans should be designed to best mimic nature (Marques-Bonet and Hvilsom, 2018). The very same information is a focal point for *in situ* conservation plans, including reintroduction activities and the designation of wildlife corridors for the passage of animals from one area to another (Frankham *et al.*, 2017).

1.1.4 Problems with CITES and invasive sampling

All gorilla species are listed in Appendix I of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES, accessible at: https://cites.org/eng) – which regulates the export of any specimen of a species (including non-invasive DNA samples) and requires prior permission and export permits. This process can be extremely long and laborious, causing significant logistical difficulties (Goossens and Bruford, 2009).

Due to the lack of well-established research facilities in habitat countries, biological samples have routinely been exported for analysis to sites where sequencing can be conducted. This has important effects on the cost and time of sequencing and it also deprives local scientists of the full benefits of research.

In order to reverse this trend, the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity (accessible at: https://www.cbd.int/abs/), enforced in 2014, requires scholars to seek the approval of local authorities and of all participants (*e.g.* field sample collectors) before publishing genetic data. The aim is to establish more predictable conditions for access to genetic resources, while helping to ensure equitable benefit-sharing when genetic resources leave the country that provides the resources. While the Nagoya Protocol represents an important step forwards to the implementation of the UN Sustainable Development Goals (accessible at: https://sdgs.un.org/goals) and to the decolonisation of research (Pettorelli *et al.*, 2021; Pomerantz *et al.*, 2022), it has long raised concerns about the possible negative impacts on academic research following the constrained exchange and use of genetic resource imposed by the Protocol (Jinnah and Jungcurt, 2009; Welch *et al.*, 2013). In fact, there is a potential risk that it hinders research by increasing the time that elapses between sample collection and publication of the results. Furthermore, growing concerns have raised about the effectiveness of the Access and Benefit-sharing (or ABS) procedures for the sharing of open-access sequence data imposed by the Nagoya Protocol (Scholz *et al.*, 2022). In fact, while the system is meant to ensure that at least a portion of the advantages and economic benefits of using open-access genetic data goes to the providing country, the strict and rigid imposition of transactions, compliance requirements, and checkpoints has also led to an increase in the

cost of obtaining the necessary permits to the detriment of research and biodiverse low- to mid-income countries, which have been calling for a more mutualistic and equitable system (Scholz *et al.*, 2022). In conclusion, researchers that use samples from CITES species should be aware of the possible complications that may arise and plan their research accordingly.

1.1.5 Challenges of using non-invasive samples

Over the last three decades, there has been a rapid increase in non-invasive DNA analysis, with great ethical and practical advantages for the study of endangered and elusive species (Bourgeois *et al.*, 2019; Höss *et al.*, 1992; Morin *et al.*, 1994; Pomerantz *et al.*, 2022). These advantages include lower disturbance and reduced risk of spreading diseases as well as a more accurate measure of population size (Gray *et al.*, 2013). Non- invasive samples, such as hair or faeces, contain either very little or highly degraded endogenous DNA (see Box 2).

Faecal samples contain comparatively large amounts of DNA, but most of it is from the microbiome and digested food products whereas the remaining target DNA (*i.e.* host organism's DNA) is often highly degraded; this is due to environmental variables such as ultraviolet light, rain and ground cover vegetation (Schultz *et al.*, 2018). The presence of polymerase chain reaction (PCR) inhibitors further reduces the usability of already degraded DNA and leads to the co-recovery of non-target DNA (Bourgeois *et al.*, 2019). In addition, volatile compounds released by decaying plant matter in herbivore faeces are also responsible for further DNA degradation (Schultz *et al.*, 2018). Therefore, there is a real difficulty in recovering good quality DNA from the target animal, which increases the risk of genotyping errors (Bourgeois *et al.*, 2019; Carlsson *et al.*, 2013) and the problem of allele dropout – stochastic amplification of only one of two alleles at a heterozygous locus (Morin *et al.*, 2001) – and the subsequent production of inaccurate genotypes (Taberlet *et al.*, 1996).

In this regard, a study conducted on the mountain gorilla population of BINP showed that a two-step storage protocol consisting of a short period of storage in ethanol followed by silica desiccation produced significantly higher amounts of target DNA than samples stored solely in ethanol or silica-dried samples (mean = 152.6 pg/µL, SD = 91.0 vs. mean

= 34.5 pg/µL, SD = 15.0 and mean = 48.3 pg/µL, SD = 37.0 respectively) (Nsubuga *et al.*, 2004). The extra manipulation of the field sample demanded by the two-step approach is offset by the substantially higher success rate in the laboratory (Nsubuga *et al.* (2004) reported that 95.2 vs. 54.6% of samples yielded concentrations of target DNA above the 10 pg/µL threshold required for reliable amplification for the two-step vs silica storage methods, respectively). As a result, a lower number of samples would be needed to obtain quality DNA from all individuals, thereby reducing the expense involved in laboratory work (Nsubuga *et al.*, 2004; White *et al.*, 2019).

Since the 1990's there has been considerable development of techniques to decrease error rates associated with low quality DNA during the amplification process, such as the multiple-tubes approach, whereby small quantities of DNA are amplified independently for each locus several times (Navidi *et al.*, 1992; Taberlet *et al.*, 1999), which became the gold standard for STR genotyping (Taberlet *et al.*, 1996; Bourgeois *et al.*, 2019). With the identification of aSTR loci, which are multiallelic and highly variable (for individual identification) and produce amplification products less than 300 bp long, it is possible to mitigate the aforementioned difficulties and yield a high rate of amplification efficiency and accuracy even from degraded DNA samples (Bradley *et al.*, 2000). STR typing is normally accomplished via PCR multiplexing for the analysis of multiple loci simultaneously recovered from small amounts of biological material. This is made possible by the use of fluorescently-labelled primers (normally labelled with a total of 4 to 5 dyes within one multiplex) to obtain different sized PCR products, of approximately 100-500 bp (Butler, 2007), that are compatible with degraded DNA (see Morin *et al.* (2001)). Multiple loci generate a relatively high power of discrimination in a single test without consuming much DNA. After PCR amplification, the length of each STR amplicon is measured to determine the number of repeats present in each allele via a sized-based separation of the PCR products involving CE (Butler, 2007). By recording the migration time of each fragment, relative to an internal size standard, the size for each STR allele, which can be recognised by the dye colour, may be determined following its separation from other STR alleles (Butler, 2007). The limited number of dyes that can be used together in a single multiplex, and the fact that the only information retrieved is the overall lengths of the STRs, and not their actual sequences, make this approach costly and not entirely accurate. Following the advances in molecular genetic techniques, researchers have started to use SNPs as novel markers in population and forensic genetic

analyses (Morin *et al.*, 2004; Butler, 2007; Bourgeois *et al.*, 2019).

SNP typing for individual and sub-species identification allows the use of particularly short amplicons thereby increasing the rate of amplification success with degraded DNA samples. One disadvantage of SNPs compared to STRs is that they are generally biallelic rather than multiallelic, so a larger number of markers is required to give sufficient resolution. In addition, SNPs in virtually all small samples suffer from ascertainment bias, which could lead to systematic deviations of population genetic statistics from theoretical expectations (Lachance and Tishkoff, 2013). This is because small samples are more likely to include more common alleles than rare ones (Lachance and Tishkoff, 2013). For example, in humans ~50 unlinked autosomal SNPs give the same power of discrimination as 15 unlinked aSTRs (Gill, 2001). The information retrieved from SNP typing is of great aid to conservationists who increasingly require reliable data on patterns of individual genetic diversity, dispersal, gene flow, population-level diversity, levels of inbreeding, and effective population size (Ne) (Schultz *et al.*, 2018). In addition, SNP genotyping can provide a higher resolution that can give less biased measures of genetic diversity than STRs (Munshi-South *et al.*, 2016) and is significantly more effective in species with low genetic diversity (Tokarska *et al.*, 2009; Schultz *et al.*, 2018), thereby reducing sequencing errors.

Box 2: Non-invasive sampling, hair vs faeces

Non-invasive genetic sampling refers to the situation in which sources of DNA, such as hair and faeces, are left behind by the animal and can be collected without causing any stress or disturbance to the animal. When it was introduced, non-invasive sampling raised many expectations among field biologists in terms of exploiting the full potential of DNA analysis; in other words, it was thought capable of providing the same information as DNA extracted from invasive, and hence good quality, samples such as blood and tissues (Taberlet *et al.*, 1999). To some extent, this proved to be true for studies requiring PCR amplification of mtDNA (present in a high number of copies per individual cell) such as species identification and intraspecific phylogeographic analyses (Morin *et al.*, 1994; Taberlet and Bouvet, 1994). However, the study of appropriate nuclear markers for individual identification using non-invasive samples proved more difficult mainly due to either low DNA quantity, low DNA quality, or poor extract quality (Taberlet *et al.*, 1999). DNA contamination and difficulties of amplifying long sequences are the obvious drawbacks, which can be avoided by following stringent guidelines and by choosing primers that amplify short DNA markers to deal with high levels of DNA degradation. When exploiting non-invasive samples for genetic analyses there are three possible outcomes: a) no PCR product is obtained, b) a PCR product but incorrect genotypes are obtained, or c) both PCR product and correct genotype are obtained (Taberlet *et al.*, 1999). When an incorrect genotype is generated, there is a higher risk of either detecting only one allele in a heterozygous individual (*i.e.* allelic dropout) or amplifying artefacts that can be misinterpreted as true alleles (*i.e.* false alleles) (Taberlet *et al.*, 1999). Several techniques were developed to deal with this type of error (see Taberlet *et al.* (1999) and references therein). Much of the success in exploiting non-invasive sampling for genetic studies depends on the attention paid to treating different kinds of samples in the appropriate manner. Indeed, the quantity and quality of DNA contained in non- invasive samples, and the issues that could follow, vary greatly depending on the nature of the sample (see Table 1.B).

Table 1.B Brief comparison of faecal and hair non-invasive samples for genetic analysis

Sample type	Characteristics	References
Faeces	Relatively high quantity of DNA but most of it from microorganisms and digested product remains Presence of PCR inhibitors DNA degrades fast in the environment	(Taberlet *et al.*, 1996; Taberlet *et al.*, 1999; Morin *et al.*, 2001; Schultz *et al.*, 2018)
Hair	Low quantity of DNA but relatively better quality Persists longer in the environment	(Morin *et al.*, 1993; Taberlet *et al.*, 1993; Taberlet *et al.*, 1999; Goossens and Bruford, 2009; Goossens *et al.*, 1998; Jeffery *et al.*, 2007b)

1.1.6 Genetic marker applications in gorilla monitoring

Autosomal STR-based length polymorphism analysis, associated with CE typing, has been the gold standard in forensic and population genetics for the three past decades (Tan *et al.*, 2018; Taberlet *et al.*, 1996; Avila *et al.*, 2019). Similarly, the control region sequence of mtDNA has also been the tool of choice in molecular studies in ecology and evolution for species and individual identification (Morin *et al.*, 2004). Both types of genetic marker are rapidly evolving DNA sequences that are informative for answering population-level questions, such as population and group dynamics based on individual identification, and as such have received great attention; large databases have been established worldwide both for animal species and humans (Morin *et al.*, 2004; Marques-Bonet and Hvilsom, 2018; Tan *et al.*, 2018; Dejean, 2018). Their application in conservation genomics increased significantly with recent technological advances allowing researchers to genotype non-invasive samples, such as faeces and hair, thereby reducing disturbance in the field (Höss *et al.*, 1992; Gulsby *et al.*, 2016; Morin *et al.*, 1994).

For example, Gray *et al.* (2013) and Guschanski *et al.* (2009) analysed 15 autosomal STRs and one sex-linked marker from the XY-homologous amelogenin gene to genotype faecal samples from the mountain gorillas of Virunga and Bwindi to distinguish among individuals in an attempt to estimate population size. Their findings, albeit useful and innovative, were flawed by the limited number of markers they could analyse via CE at every run. When a large proportion of these markers fails to amplify or cannot be genotyped correctly, the amount of information being lost can significantly affect the outcome of the research. For the samples collected in the BINP, genotypes were on average 84.9% complete (Guschanski *et al.*, 2009), whereas in the case of VM samples, this number increased to 85.9% (Gray *et al.*, 2013), with the majority of individuals being fully genotyped at eight or more loci.

With the development of Massively Parallel Sequencing (MPS), it is now possible to obtain the sequences of multiple markers amplified simultaneously. Researchers can therefore sequence SNPs, which were shown to be more informative in terms of species and sub-species identification than STR and mtDNA markers (Marques-Bonet and Hvilsom, 2018) (Table 1.3). Not only that, SNP typing can succeed based on very short

amplicons thus greatly improving the success of analysing non-invasive samples that contain highly degraded and fragmented DNA. The development of MPS and the introduction of commercially available kits for humans comprising autosomal STRs, Y-STRs, X-STRs, Identity SNPs and Ancestry SNPs (*e.g.* Verogen - ForenSeq™ DNA Signature Prep Kit) have alleviated the limitations associated with CE-based fragment analysis. Another advantage deriving from SNP analysis is that SNPs can also be analysed on portable devices such as Oxford Nanopore Technology's MinION for in field identification (Cornelis *et al.*, 2017).

Table 1.3: A simple summary comparing the characteristics of mtDNA, STRs, and SNPs.

Characteristics	**mtDNA**	**STRs**	**SNPs**
Maternal inheritance	+	+	+
Paternal inheritance	-	+	+
Autosomal inheritance	-	+	+
Individual identification	-	+	+
Species identification	+	-	+
Performance on degraded DNA	+	-	+

Box 3: Short Tandem Repeats (STRs) and Single Nucleotide Polymorphisms (SNPs)

Short tandem repeats are tandem arrays of repeat units 1-7 bp in length distributed throughout the genome characterised by a number of copies of 10-30 units upon which they are categorised (Jobling *et al.*, 2013). The STR structures are further categorised into either "simple", with only one repeat type present, "compound", containing more than one type of repeat, and "complex", consisting of varying composition of repeats with additional internal sequences (Fan and Chu, 2007). STRs have been observed to mutate mostly via a single-step (±1 repeat) process. Under the stepwise mutation model (Ohta and Kimura, 1973), expansions and contractions of STR lengths, observable between individuals at specific loci, are the result of replication slippage. Slippage is due to DNA polymerase experiencing difficulties in replicating repeated sequence motifs during replication (Figure 1.B). STRs can be inherited autosomally or paternally (Y-STRs), and are normally typed in PCR multiplexes, with fluorescently-labelled primers, and separated and detected via CE (see section 1.5).

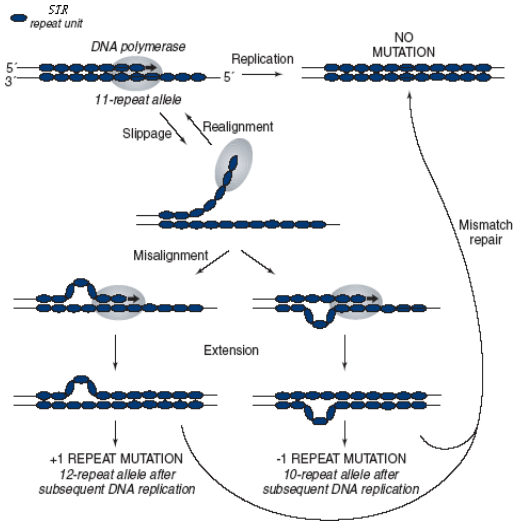


Figure 1.B Stepwise mutation model and formation of STR polymorphisms

Single nucleotide polymorphisms represent the simplest smallest-scale difference between two homologous DNA sequences. SNPs are the most common form of genetic variation and consists of base substitutions, accounting for up to 85% of human genetic variation (Budowle and Van Daal, 2008), in which one base is mistakenly exchanged for another, but single base insertions/deletions are also possible during DNA replication and repair (Jobling *et al.*, 2013). SNPs may generate via direct mutagenesis either through internal or external mutagens. SNP mutation rate (*i.e.* $\sim 10^{-8}$ per SNP per generation) is much lower than that of STRs (*i.e.* $\sim 10^{-3}$ per STR per generation), meaning that they are inherited from one generation to another, making them particularly suitable for evolutionary studies and species identification. SNPs can be inherited maternally (in mtDNA), paternally (Y-chromosome) or autosomally, and therefore can be used for a wide variety of population analyses.

There is a consensus among the scientific community that 40-60 highly heterozygous SNPs are capable of providing the same accuracy as 13-15 STR loci (Butler, 2007; Amorim and Pereira, 2005), and thus SNP loci need to be carefully chosen so as to maximise heterozygosity for individual identification. One, however, must be aware of the limitation of ascertainment bias arising from, for example, the non-random sampling of individuals for the identification of suitable forensic genetic markers so as to minimise

the risk of systematic distortions of true allele frequency measures in populations, which could influence the results of genomic analyses (McTavish and Hillis, 2015; Lachance and Tishkoff, 2013). In this regard, the power of MPS resides in the fact that it allows for a greater variety of genetic analysis based on non-invasive samples, thereby enabling researchers to analyse a higher number of markers than with CE, without the restrictions of non-overlapping amplicons and limited numbers of fluorescent dye channels. MPS data simulation by Fumagalli (2013) shows that genotyping small sample sizes at high sequencing depth can yield highly precise detection of polymorphic sites (n = 20, depth = 50 ×, precision = 1). However, accurate detection can also be achieved by genotyping larger sample sizes at lower sequencing depth; Schultz *et al.* (2018) suggested that comparable results can indeed be achieved with a sample size of 100 individuals sequenced at 10 × depth, yielding a precision of 0.779 ± 0.0441 with a confidence interval of 95%. In conclusion, the development of genomic analyses based on SNP typing could represent the way forward for future ecological studies that aim at producing accurate estimates of threatened species' population sizes.

1.1.7 The evolution of genomic science: Next Generation Sequencing

After the advent of the Sanger chain termination method (Sanger *et al.*, 1977), which, for the first time, allowed scientists to sequence DNA in a reliable and reproducible manner, DNA sequencing has undergone remarkable improvement. In the '80-'90s, Applied Biosystems introduced the first automated CE sequencer, the AB370 and the AB370xl, which soon became the primary workhorses for numerous DNA sequencing projects (Collins *et al.*, 2003). However, it was only with the introduction of the Genome Analyzer in 2005 that high throughput sequencing became a reality, taking sequencing runs from 84 kb per run to 1 Gb per run (Illumina, 2015b). This revolution arrived with a short-read-based massively parallel sequencing technique that enabled scientists to more than double the data output each year, finally launching the "next generation" in genomic science (Goodwin et al., 2016).

The technological development of next generation sequencing (or massively parallel sequencing – MPS) allowed a dramatic cut in costs, remarkably reducing the cost of

sequencing a whole human genome from over three billion dollars in 2001 to approximately $1000 in less than two decades (Illumina, 2015b). This reduction allows researchers to analyse thousands of samples in a relatively short period of time, thereby enabling population-scale studies, and more recently found applications in other fields such as forensics and conservation biology. A major player in the development of the MPS technology is Illumina® and its technology has become the most widely established.

The great advance of MPS resides in its biochemistry. Automated Sanger sequencing proceeds via the incorporation of deoxynucleotides into the growing daughter strand during the polymerase chain reaction (PCR) process, which is terminated by incorporation of a fluorescently-labelled 3´- dideoxynucleotide lacking of the hydroxyl (-OH) group necessary to create phosphodiester bonds (Jobling *et al.*, 2013). The result is a set of fragments of different lengths, each terminating with one of the four bases labelled with a different dye, allowing the sequence to be read from an electropherogram after fluorescence base detection (Jobling *et al.*, 2013) (Figure 1.4). It is possible to process more than 1000 samples in one day using a single capillary sequencer, yielding 850-bp reads each, with an accuracy of nearly 99% (Jobling *et al.*, 2013).

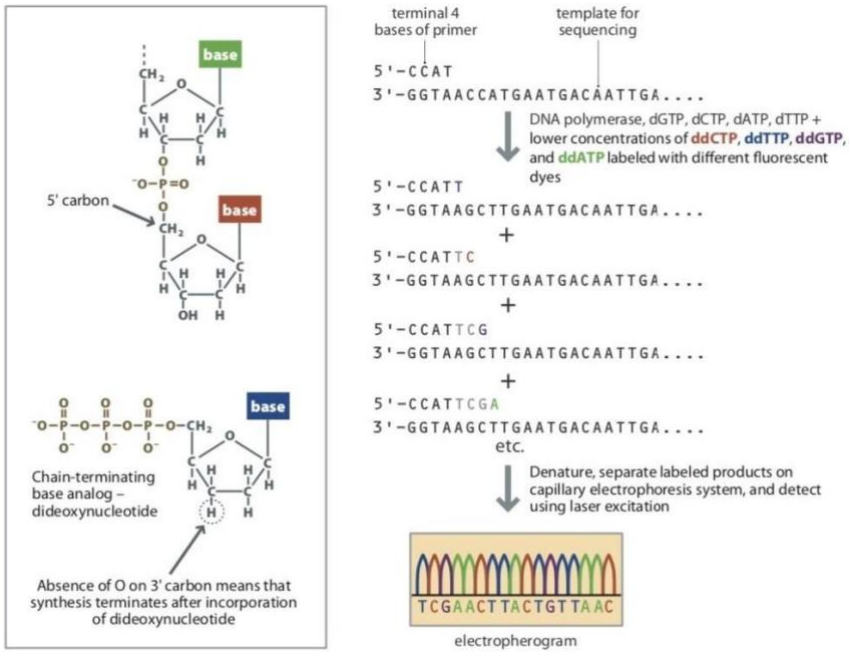


Figure 1.4: Principle of Sanger Sequencing of DNA (Jobling *et al.*, 2013).

Illumina® uses a different workflow for next-generation sequencing (Figure 1.5) that

comprises four main steps (Illumina, 2015b):

A. Library Preparation: the DNA is randomly fragmented and 5´ and 3´ adapters are ligated to ends of each fragment. Through reduced cycle amplification, additional motifs are introduced, such as the sequencing binding sites, indices, and sequences complementary to the flow cell oligos.
B. Cluster Amplification: each fragment is then loaded into a flow cell and isothermally amplified. The flow cell consists of glass lanes covered in surface-bound oligos complementary to the library adapter regions. Each fragment attaches to the complementary oligos and amplifies into distinct clonal clusters through bridge amplification. When clustering is complete, the templates are ready for sequencing.
C. Sequencing: sequencing begins with the extension of the first sequencing primer to produce the first read. With each cycle fluorescently tagged nucleotides compete for addition to the growing chain, minimising incorporation bias thereby reducing raw error rates compared to other technologies. Only one is incorporated based on the sequence of the template. After the addition of each chain-terminating nucleotide the clusters are excited by a light source and a characteristic fluorescent signal is emitted. An important feature is that chain-terminating nucleotides are reversible, meaning that after imaging the terminating group can be removed (deblocking) and the next nucleotide inserted until the entire fragment is completely read. This proprietary process is called Sequencing-by-Synthesis (SBS). For a given cluster all identical fragments are read simultaneously. After the completion of the first read, the second read is read the same way. Hundreds of millions of clusters are sequenced in a massively parallel process.
D. Data Analysis: newly identified sequence reads are aligned to a reference genome. Following alignment, it is possible to analyse the variations, such as SNPs and STRs. Dedicated software are now available to aid the analysis of the output data.

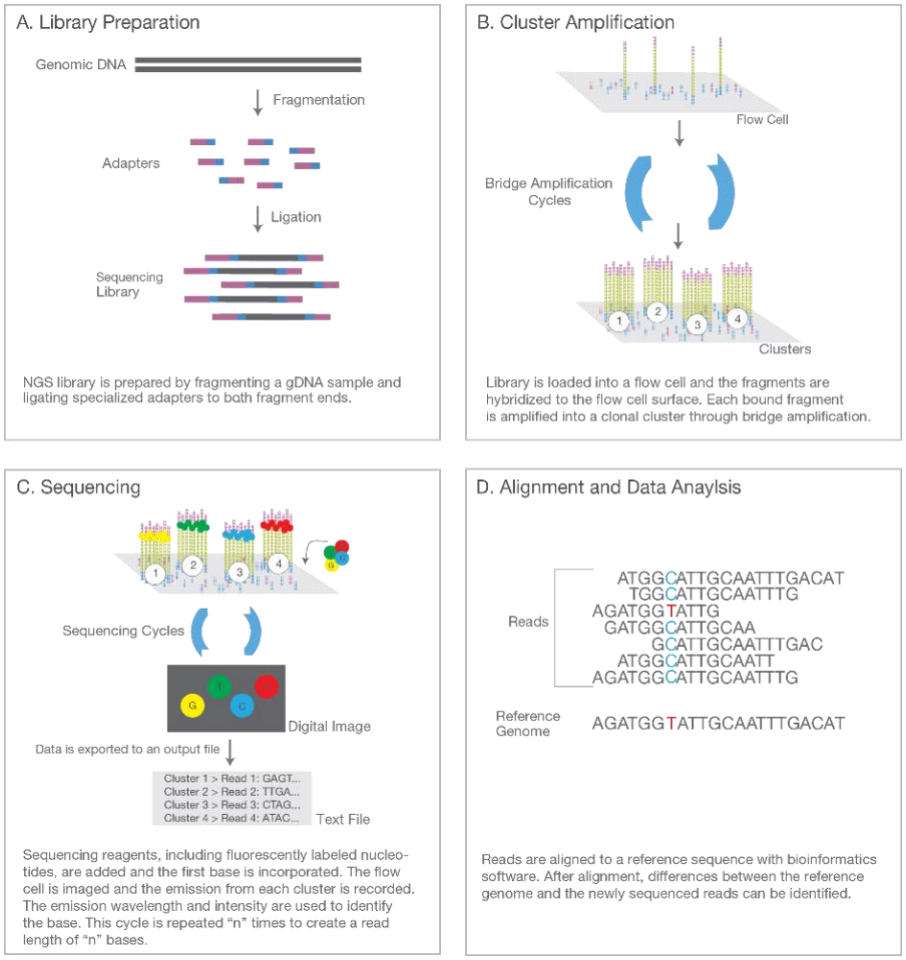


Figure 1.5: Principles of Illumina Massively Parallel Sequencing (Illumina, 2015a)

Another advance in MPS technology was the introduction of paired-end (PE) sequencing. PE allows sequencing from both ends of the DNA fragments and aligning the forward and reverse reads as read pairs (Illumina, 2015b). Different platforms exploit different sequencing principles and also yield reads of different lengths, for example Illumina® HiSeq™ has a read length of 2 x 250 bp whereas Illumina® MiSeq™ 2500 reads are 2 x 300 bp in length.

The PE process generates twice the number of reads and enables more accurate alignment with the ability to detect indels which is not possible otherwise (Illumina, 2015b). In addition to a significant increase in data output per run, the sample throughput per run in MPS has risen over time. This is enabled by multiplexing, which allows the preparation and pooling of multiple libraries simultaneously (Figure 1.6) (Illumina, 2015b). With short, specific indices attached to both ends of the fragment during library preparation, each read can be identified and sorted bioinformatically before the final analysis; as a

result, the time to data for the analysis of multiple samples has been drastically reduced. Owing to the reduction in costs and time for the generation of sequencing data, following the advent of MPS, the amount of data being generated has increased enormously. This requires careful computational storage and processing of these data in order to produce meaningful information (Nielsen *et al.*, 2011). Several steps need to be taken in order to deal with the various errors that might be generated from MPS: while mapping (see Section 1.3) some reads may be wrongly aligned to the reference genome, owing to a combination of sequencing mistakes and true variation including SNPs and indels. The number of MPS sequencing errors is high if compared with Sanger sequencing (Jobling *et al.*, 2013). For instance, individual reads in Illumina® have a mean error rate of >0.1%, which increases towards the end of the read. This means that base calling should be seen as probabilistic rather than deterministic, as it is the case for Sanger sequencing. The number of reads per individual base (*i.e.* coverage) generated through MPS methods is therefore a powerful statistical tool to assess the degree of confidence about the nature of that base; the higher the coverage the higher the probability that both alleles are sampled equally (Jobling *et al.*, 2013).

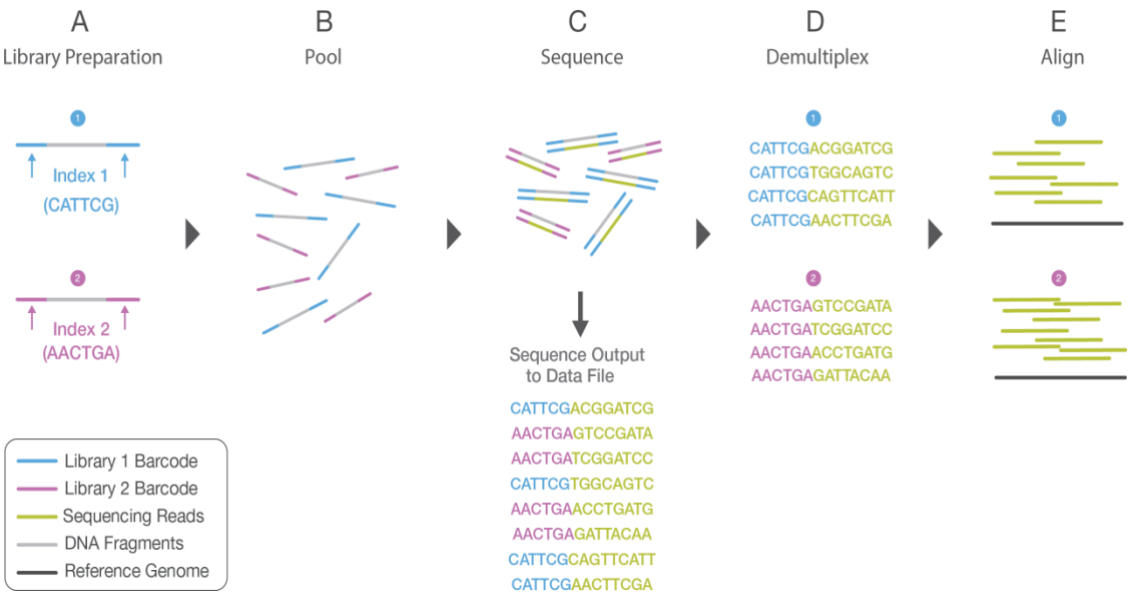


Figure 1.6: MPS library multiplexing overview (Illumina, 2015a)

Because of these advances MPS, and Illumina® technology in particular, is becoming increasingly important in a wide variety of fields, including conservation biology where it is increasingly used to infer kinship to inform meta-population management of

endangered species by identifying candidates for translocation, reintroduction, etc., and to easily identify species and individuals for the fight against illegal hunting and wildlife trafficking (Purisotayo *et al.*, 2019; Gueuning *et al.*, 2019). Also, due to the reduced length of the DNA fragments, MPS is particularly suitable for the sequencing of degraded DNA samples such as faeces and hair, thereby overcoming the issues described in Section 1.5.

A step further in the development of sequencing technology is represented by the introduction of the MinION™, a single-molecule nanopore real-time sequencing device from Oxford Nanopore Technologies Ltd. (hereafter, ONT), that offers portability, lower overall instrument cost, and ease of use (Plesivkova *et al.*, 2019; Pomerantz *et al.*, 2022). The MinION™ is potentially suitable for the study of free roaming wildlife in remote areas and has already been tested to work even in extreme climates and remote areas where it has been successfully used for *in situ* species barcoding (Pomerantz *et al.*, 2017; Pomerantz *et al.*, 2022). This technology exploits nanopores (*i.e.* engineered proteins) embedded in an electrically resistant membrane of synthetic polymers immersed in an electrically charged ionic solution (Slatko *et al.*, 2018). When a DNA molecule traverses the nanopore it produces a characteristic disruption of the electrical field, producing alterations in the current, that are used to identify individual bases (Plesivkova *et al.*, 2019) (Figure 1.7). This approach allows the MinION™ to generate real-time sequencing information but it can also be a potential source of sequencing errors, especially when dealing with homopolymers and repetitive regions such as STRs (Lu et al., 2016), which represent nearly half of all sequencing errors (Delahaye and Nicolas, 2021). In addition, specifically to nanopore, recent findings suggest that low-GC reads have fewer errors than high-GC reads (Delahaye and Nicolas, 2021). Indeed, despite continuous development, the base calling accuracy of MinION™ has increased from 60% to 85%, though recent studies show an accuracy of nearly 95% on some datasets (Boža *et al.*, 2021), which is still well below other commonly used platforms such as those manufactured by Illumina® (Plesivkova *et al.*, 2019; Napieralski and Nowak, 2022). Illumina® sequencing in fact can reach an accuracy of 99.9%, which is still considerably higher than that of nanopore-based sequencing. A plethora of basecalling software have been developed through the years (Jäger *et al.*, 2017; Boža *et al.*, 2017; Stoiber and Brown, 2017; Teng *et al.*, 2018) with the aim of increasing the accuracy and therefore reducing the error rate. The current official ONT basecaller is Guppy (Napieralski and Nowak, 2022), a proprietary software

accessible to all ONT users. Using SNPs instead of STRs for individual identification can thus be beneficial, as it reduces the issues related to the sequencing of repetitive regions (*e.g*. STRs) and works best with degraded DNA, requiring shorter length amplicons (Plesivkova *et al.*, 2019). ONT has produced two different kinds of flow cells that can be used with the MinION device, the MinION flow cell with 512 channels and four nanopores per channel and the Flongle flow cell with 126 channels (Grädel *et al.*, 2019). Flongle flow cells are single-use, relatively cheap flow cells that can generate up to 2.8 Gb for sequencing DNA in real-time, whereas MinION flow cells can generate up to 50 Gb of can be used for in real-time sequencing of DNA, cDNA and RNA (Sessegolo *et al.*, 2019). While Flongle flow cells are disposable, MinION flow cells can be washed, stored and reused, until all pores are inactive.

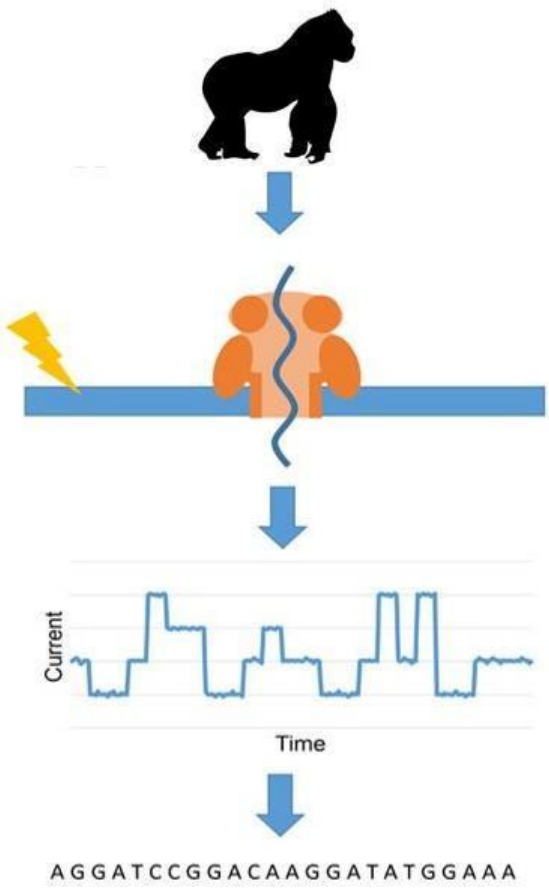


Figure 1.7: Nanopore Sequencing technology adapted from Plesivkova *et al.* (2019)

1.2 Aims & Objectives

This thesis seeks to develop genomic tools for the identification of gorilla individuals and species exploiting DNA analysis in non-invasive samples. The main aims and objectives of the project are:

- To use MPS to analyse the orthologs of human aSTRs in a diverse set of non-human African great ape samples, and to compared these to human data, in order to assess the potential for individual and (sub)species identification
  - To test the currently available human-based ForenSeq DNA Signature Kit to generate data for 27 a-STRs, and related flanking regions, in a panel of non-human African great apes
  - To ask if this kit can perform well in matched blood and faecal/hair DNA samples from animals at Twycross Zoo (this includes a breeding group of six individuals)
  - To compare the derived orthologous STR sequences across great ape (sub)species and illuminate the evolution and species-specificity of STRs
  - To ask if MPS provides better resolution for both individual and (sub)species identification than conventional CE-based methods

- To develop SNP panels for gorilla sub-species and individual identification, and design PCR multiplexes to amplify these
  - To identify individual identification SNPs (*ii*SNPs) and sub-species identification SNPs (*si*SNPs) from published whole genome sequencing data
  - To test *ii*- and *si*SNP panel amplification on a good quality gorilla DNA sample of known genome sequence

- To use ONT sequencing technology (i.e the MinION sequencing device) to develop a portable and low-cost laboratory system for the analysis of *ii*- and *si*SNPs in gorilla DNA samples
  - To assess the performance of ONT sequencing with short DNA fragments
  - To compare the SNP genotypes across gorilla sub-species and individuals with published whole genome sequencing data
  - To analyse the performance of *ii*- and *si*SNP panels in samples of gorilla DNAs, comparing good quality and non-invasively collected samples samples (*e.g.* hair and faeces)

- To implement a straightforward analysis of SNP data to facilitate the adoption of genomic techniques in gorilla conservation and the fight against wildlife crimes
- To use population genetic methods to compare genetic diversity among gorilla sub-species and perform parentage analysis.

The findings will improve current efforts for the conservation of the sub-species both in their home-range and *ex situ* (*e.g.* zoos and zoological parks).

Chapter 2 Massively parallel sequencing of the orthologs of human autosomal STRs in great ape species

This Chapter is planned for submission as a research paper to BMC Ecology & Evolution, and is currently available on BioRxiv as:
doi: https://doi.org/10.1101/2022.08.03.502616

2.1 Introduction

Habitat loss, disease, climate change and hunting are among the main drivers of localised and global extinctions (Maxwell *et al.*, 2016). As species become increasingly restricted to fragmented habitats it is necessary to assess their genetic viability to support effective management decisions. Increasing global awareness has drawn attention towards the preservation of charismatic flagship species (Williams *et al.*, 2000), among which the African great apes have been a focal interest: most of these species remain critically endangered throughout their home ranges (Kuhlwilm *et al.*, 2016) (Figure 2.1). However, when threat status is measured merely on the basis of species decline and habitat degradation (Rivers *et al.*, 2014), it can neglect the biological and ecological impacts of shifts in population size and distribution (Ouborg *et al.*, 2010). As populations decline and inbreeding intensifies, high rates of homozygosity spread among groups of individuals (Keller and Waller, 2002). In turn, reduced allelic diversity can affect the adaptive ability of the species and potentially lead to the emergence of genetic defects underpinned by recessive alleles (Xue *et al.*, 2015).

As a response, conservation efforts have seen an upsurge in the use of DNA testing to assess animal population parameters playing an important role in the implementation of effective wildlife management and preservation policies (Guschanski *et al.*, 2009). Measuring polymorphism at sets of autosomal short tandem repeats (aSTRs) via capillary electrophoresis (CE) has been an important tool in population genetic analysis (Allendorf, 2017). Because they assort independently at meiosis, sets of unlinked aSTRs can also yield genotypes that provide individual identification within a species: this forms the

basis of human forensic identification technologies (Jobling and Gill, 2004), and can be applied in forensic casework involving animals, for example in poaching or illegal trade cases (Linacre, 2021). Such genotypes also have the potential to distinguish between species and subspecies when allelic spectra are suitably differentiated and characteristic.

Because of the high levels of sequence similarity among great-ape genomes (Wall, 2013), PCR primers for aSTR markers developed in humans are expected often to amplify their orthologs. Indeed, some STR multiplexes designed for human forensic analysis have been shown to have cross-species application for the analysis of orthologous loci in other great apes and that these are indeed polymorphic (*e.g.* Thakur *et al.* (2018) and Singh *et al.* (2021)). The underlying assumption is that amplicons generated at orthologous loci are generally commensurable across species (Kwong and Pemberton, 2014). In other words, it is assumed that the overall structure of STRs is maintained across species and that these can be used to conduct comparative studies. However, this assumption is often incorrect; indeed, the presence of species-specific indels in flanking sequence together with different organisation and variability of STRs present difficulties with great-ape cross-species comparisons (Kwong and Pemberton, 2014). In translating multiplexes designed in humans to other species, there is also a practical problem of interpretation, since allele size ranges for different loci (labelled with the same fluorescent dye) were designed to be non-overlapping in humans, but may well overlap in non-human primates.

These issues arise because of the nature of CE, which assesses polymorphism by measuring the length of PCR fragments and converting this to an assumed number of repeat units within each allele. An alternative approach is multiplex massively parallel sequencing (MPS), in which the sequences of STRs, rather than their lengths, are analysed. This obviates the problem of size-range overlap, since it is the sequence itself that identifies the locus, and also permits larger numbers of STRs to be simultaneously analysed than is possible with length-based CE genotyping. MPS-based analysis is now becoming established in human forensic genetics. For example, the ForenSeq DNA Signature Prep Kit (Verogen) (Churchill *et al.*, 2016; Just *et al.*, 2017) includes multiple autosomal, X- and Y-chromosomal STRs, as well as autosomal SNPs for individual identification.

The aim of this Chapter is to assess how the ForenSeq multiplex system designed for

humans performs in amplifying and sequencing autosomal STRs in a set of chimpanzees, bonobos and gorillas, and to ask if the orthologous loci are both individually identifying and can robustly distinguish groups at the species and subspecies levels. Sequencing across subspecies and species may also reveal aspects of the mutation processes of these widely used STRs across several million years of primate evolution.

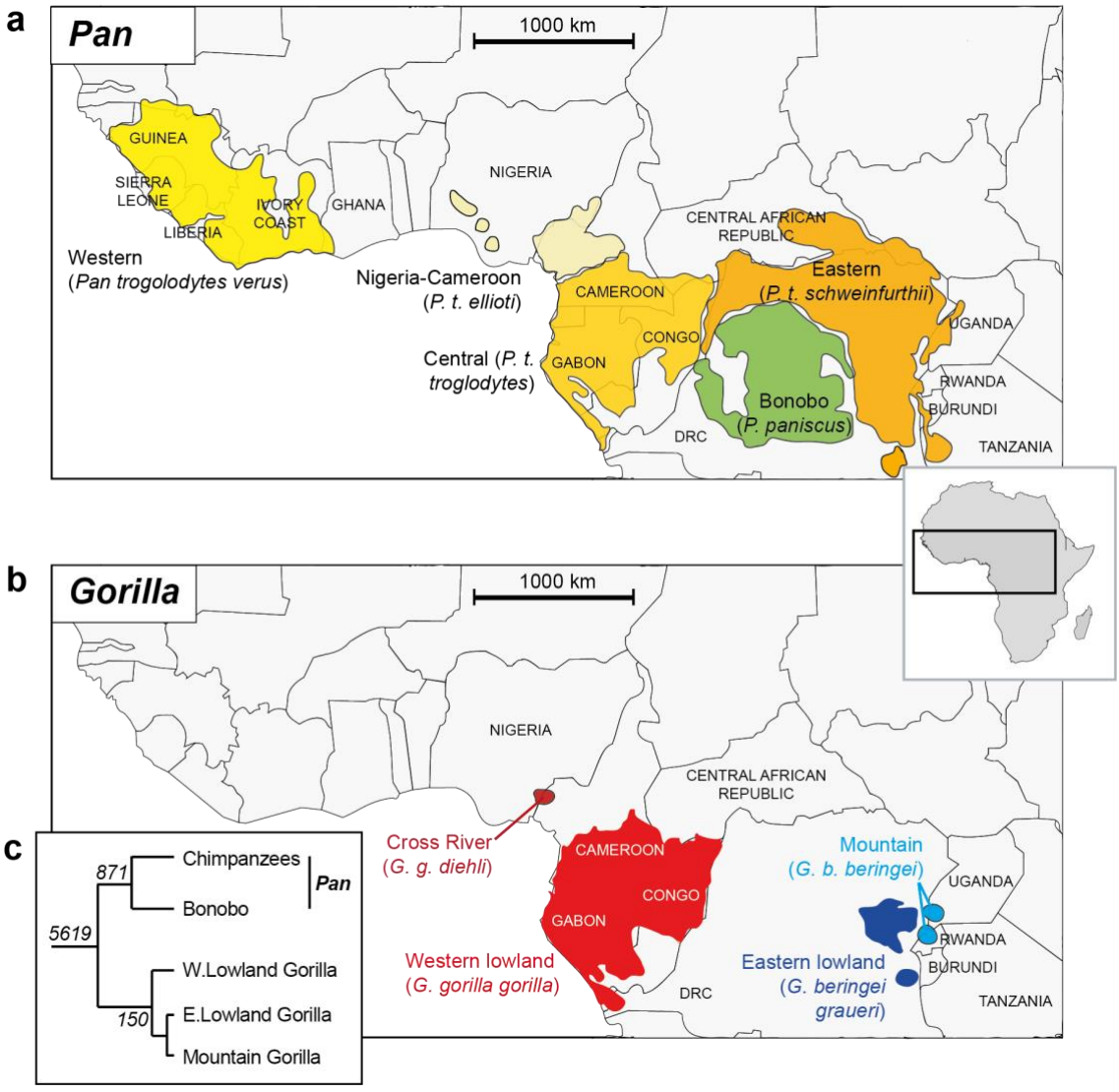


Figure 2.1 Pan and Gorilla species and sub-species distributions, and phylogenetic relationships. Distributions of a) *Pan*, and b) *Gorilla*, adapted from (Jobling *et al.*, 2019). c) Phylogeny showing relationships between (sub)species (branch lengths not to scale), with classifications reflecting those used in this study. Italic numbers at nodes are split times in thousands of years, based on a mutation rate of 1 x 10-9 per bp per year (Prado-Martinez *et al.*, 2013). Map adapted from Africa just countries grayish.svg, published on Wikimedia under a Creative Commons Attribution-Share Alike 4.0 International license. DRC: Democratic Republic of the Congo

2.2 Methods

2.2.1 DNA samples and data

DNA samples were collated from a variety of sources including laboratory collections, as detailed in Table 2.1. For chimpanzees, subspecies definition was sometimes unclear, and where it was defined, sample sizes for individual subspecies were small: chimpanzees were therefore considered at the species level. By contrast, gorilla samples were better defined, with at least six individuals in each of three of the four known subspecies, and therefore gorillas were considered at this level. As a result, comparison groups were five in number: chimpanzee - Pan troglodytes (n=14), bonobo - P. paniscus (n=4), western lowland gorilla - Gorilla gorilla gorilla (n=16), eastern lowland gorilla - G. beringei graueri (n=6) and mountain gorilla - G. b. beringei (n=12). To provide comparative information on the same set of loci in humans a published dataset based on analysis of the ForenSeq™ DNA Signature Prep Kit in 89 unrelated Saudi Arabian human males (Khubrani *et al.*, 2019) was used, as well as information from STRBase.nist.gov and (Gettings *et al.*, 2016).

2.2.2 Library preparation and sequencing

DNA samples from DNA extracts collections were quantified using the Qubit™ 2.0 Fluorometer with the Qubit™ dsDNA HS (High Sensitivity) Assay Kit for double-stranded DNA (dsDNA) (Table 2.1). Sequencing libraries were prepared with the human-based ForenSeq™ DNA Signature Prep Kit according to the manufacturer's recommendations (Verogen®, San Diego, CA, USA). Primer mix A was used to target 58 STRs (27 autosomal STRs, 7 X-STRs and 24 Y-STRs) and 94 identity-informative SNPs (iiSNPs). As per recommended protocol, purified DNA template was diluted to an input concentration of 0.2 ng/µl; also, as the kit has been specifically designed to work with traces of degraded DNA no additional DNA integrity assessment was conducted prior to library preparation. Extracted DNA purity ratios were measured utilizing the OD260/280 ratios and all samples were >1.8 on the NanoDrop 2K (data not shown). Steps for library preparation include amplifying, indexing, purifying, normalising and pooling, prior to sequencing on an Illumina MiSeq FGx, all of which were performed in accordance with the manufacturer's recommended protocols.

2.2.3 Sequence data analysis

Quality-checked FASTQ files were generated using Trimmomatic v.0.36 (Bolger *et al.*, 2014) for adapter sequence and poor-quality base trimming using the Linux terminal. The threshold for minimum read length was set at 50 bp.

Analysis of human data using the DNA Signature Kit is usually undertaken using the ForenSeq™ Universal Analysis Software (UAS), but for the non-human analysis done here the software FDSTools (Hoogenboom *et al.*, 2017) was employed. This is laborious, but has the advantage that tailored anchor, flanking and repeat-array sequences can be designed, hence obviating the need for a reliable reference genome, which is still lacking for *Gorilla* beringei and most *Pan* sub-species. In order to develop library files for variant calling for Pan and Gorilla, trimmed BAM files were visualised and aligned with the human reference (GRCh38/hg38) using the Integrative Genomic Viewer (IGV) (Robinson *et al.*, 2011) allowing the identification of suitable flanking sequences as anchors (Hoogenboom *et al.*, 2017). Considering the kit chemistry, which produces short and unreliable second reads, the 5´ anchor was set close to the 5´ end of the repeat array of each locus, so as to maximise the coverage for each marker. Flanking sequences to be added to the final version of the library file were obtained through repeated runs of FDSTools.

If there are null (non-amplifying) alleles at a given STR locus, these may exist in a heterozygous state, and it then becomes necessary to distinguish between such heterozygotes and true non-null homozygotes, in which two identical alleles are amplified. This was done using a sequence read-depth approach, normalised against known heterozygote calls, since a true homozygote's read-depth should be equal to the sum of two heterozygous alleles (following the approach for duplicated Y-STR alleles set out in Huszar *et al.* (2018)). In particular, the contribution of each individual to the total number of reads across loci was computed in Excel whereby putative homozygous alleles were treated as truly homozygous (*i.e.* alleles were counted only once). The cumulative number of reads (*i.e.* sum of reads for each individual and locus) was then divided by the sum number of reads per individual to compute the proportion of individual cumulative contribution to the number of reads.

In order to assess the expected individual contribution to the total number of reads for each locus, the individual total number of reads per locus was multiplied by the proportion of individual contribution to the cumulative number of reads. This was compared with the observed individual contribution to each locus's total number of reads, which was computed by dividing the total number of reads per locus by the individual sum number of reads per locus.

Information from the comparison between observed and expected individual contribution to total number of reads for each locus was used to assign individuals to different ploidy categories. In particular, focusing on putatively homozygous individuals, the observed and expected contribution of individuals to the total number of reads of any given locus was compared. If the former was half (or close to half) the latter, the individual would classified as "null hemizygote" for that particular locus. Similarly, if the expected contribution of an individual to the total number of reads was relatively high (*i.e.* suggesting good quality) but the observed contribution to a given locus was zero, it would then be considered to be "null homozygous" for that particular locus. If the number of reads was low (*i.e.* < 20 reads per locus) the individual would be classified as "sub-threshold" for that particular marker. Indeed, it would be inaccurate to classify that individual as either "homozygote" or "heterozygote" for the locus, given the minimum threshold of 10 reads per allele (*i.e.* equivalent to 20 for homozygous and heterozygous loci). "No amplification" was assigned when no amplification (*i.e.* 0 reads) was observed and the contribution to the total number of reads was either zero or close to zero.

2.2.4 Population, forensic and statistical analysis

STRAF (Gouy and Zieger, 2017) was used to calculate forensic statistics, including genotype count (N), allele count based on sequence (Nall), observed and expected heterozygosity (Hobs and Hexp), polymorphism information content (PIC), match probability (PM), power of discrimination (PD), power of exclusion (PE), and typical paternity index (TPI).

Clustering of genetically similar individuals was investigated using both STRUCTURE (Pritchard *et al.*, 2000), and discriminant analysis of principal components (DAPC). DAPC was conducted using the package *adegenet* (version 2.1-3) (Jombart and Ahmed,

2011) implemented in R version 3.6.3 (Team, 2018). For DAPC, the function *find.clusters()* was used to determine the optimal cluster number without prior information, and the Bayesian information criterion (BIC) was used to identify abrupt changes in fit models for successive runs of increasing k-means clustering with $K = 1 - 8$. The number of PCs to retain was cross-validated using the function *xvalDapc()* with 50 repetitions in order to avoid overfitting.

ML-Relate (Kalinowski *et al.*, 2006) was used to screen the sample set for closely related individuals within (sub)species. Based on this, together with some prior information (Table 2.1) some individuals were removed for some analyses, as described in the first paragraph of the Results section.

In considering STR repeat arrays across species, consider four basic types were considered: perfect (an uninterrupted array of a single repeat type, *e.g.* [GATA]n); interrupted (two or more arrays of the same repeat type interrupted by non-repeat material, *e.g*. [GATA]nNNNNN[GATA]m); imperfect (two or more arrays of the same repeat type interrupted by repeat-derived material, *e.g*. [GATA]nGACA[GATA]m or [GATA]nGAT[GATA]m); compound (two or more variable arrays of different repeat types of the same length, *e.g*. [GATA]n[GACA]m). Two hybrid categories were also included, compound interrupted (two or more variable arrays of different repeat types of the same length, interrupted by non-repeat material, *e.g.* [GATA]nNNNNN[GACA]m), and imperfect compound (two or more arrays of different repeat types of the same length, interrupted by repeat-derived material, *e.g*. [GATA]nAATA[GACA]m).

The initial hypothesis is that the amplification of orthologous markers will be successful in closely related non-human great apes and that the analysis of STRs structure will provide some level of individual, and *genus* and/or species identification.

Table 2.1: Sample information. Each colour indicates a different (sub)species.

Sample ID	Sample name	Common name	Scientific name	Sex	Geographic origin/birth origin	Relatives in dataset? (MLRelate)a	ng/µl	Notes
Ptr_3	Tommy	Hybrid	*Pan troglodytes troglodytes / Pan troglodytes verus*	M	Wild born (?)		19.11	
Ptr_4	PTR7 - Johnny	Western chimpanzee	*Pan troglodytes verus*	M	Captive born		12.00	
Ptr_5	Buttons	Western chimpanzee	*Pan troglodytes verus*	M	-		1.07	
Ptr_6	NA03450	Western chimpanzee	*Pan troglodytes verus*	M	Coriell Institute for Medical Research	FS Ptr_18 (GM3450)	15.00	Excluded from population analyses
Ptr_7	EB176JC	Hybrid	*Pan troglodytes verus / Pan troglodytes ellioti*	M	ECACC cell line no. 89072704		3.17	
Ptr_14	Zurich	Chimpanzee	*Pan troglodytes sp.*	M	-		7.22	
Ptr_15	GM3452	Western chimpanzee	*Pan troglodytes verus*	M	Coriell Institute for Medical Research		1.05	
Ptr_16	GM3448	Western chimpanzee	*Pan troglodytes verus*	M	Coriell Institute for Medical Research		20.07	
Ptr_17	J19	Western chimpanzee	*Pan troglodytes verus*	M	Unknown		11.02	
Ptr_18	GM3450	Western chimpanzee	*Pan troglodytes verus*	M	Coriell Institute for Medical Research	FS Ptr_6 (GM3450)	14.37	
Ptr_21	Flynn	Hybrid	*Pan troglodytes verus / unknown*	M	Twycross Zoo	FS Ptr_22 (Jomar)	1.56	Excluded from population analyses
Ptr_22	Jomar	Hybrid	*Pan troglodytes verus / Pan troglodytes schweinfurthii*	M	Twycross Zoo	FS Ptr_21 (Flynn)	7.86	
Ptr_23	Peter	Eastern chimpanzee	*Pan troglodytes schweinfurthii*	M	Twycross Zoo		4.08	
Ptr_29	Tina	Chimpanzee	*Pan troglodytes sp.*	F	Unknown		19.11	
Ppa_1	Bono	Bonobo	*Pan paniscus*	M	Wild born		17.78	
Ppa_2	PPA2	Bonobo	*Pan paniscus*	M	Captive born		15.77	
Ppa_19	Moko	Bonobo	*Pan paniscus*	M	Twycross Zoo		5.98	
Ppa_20	Keke	Bonobo	*Pan paniscus*	M	Twycross Zoo		5.95	
Ggg_8	Fritz	Western lowland gorilla	*Gorilla gorilla gorilla*	M	Cameroon, wild born		23.99	
Ggg_9	Tomoka	Western lowland gorilla	*Gorilla gorilla gorilla*	M	Captive born	PO Ggg_11 (Tomoka)	13.48	

Ggg_10	Guy	Western lowland gorilla	*Gorilla gorilla gorilla*	M	Cameroon, wild born		22.37	Repeated sample Ggg_24 Ggg_25 Ggg_27
Ggg_11	Nikumba	Western lowland gorilla	*Gorilla gorilla gorilla*	M	Republic of the Congo (?), wild born	PO Ggg_9 (Tomoka)	60.24	Excluded from population analyses
Ggg_26	Magawi985	Western lowland gorilla	*Gorilla gorilla gorilla*	F	-		31.44	
Ggg_28	GoM7	Western lowland gorilla	*Gorilla gorilla gorilla*	M	-	HS Ggg_36 (Matadi)	36.11	
Ggg_31	Mjukuu	Western lowland gorilla	*Gorilla gorilla gorilla*	F	-		85	
Ggg_33	EB (JC)	Western lowland gorilla	*Gorilla gorilla gorilla*	F	ECACC		21.33	
Ggg_34	Kaja	Western lowland gorilla	*Gorilla gorilla gorilla*	F	Chessington Zoo	FS Ggg_39 (Kanghu); HS Ggg_41 (Fubu)	41.43	
Ggg_35	Effie	Western lowland gorilla	*Gorilla gorilla gorilla*	F	London Zoo RP		10.00	
Ggg_36	Matadi	Western lowland gorilla	*Gorilla gorilla gorilla*	M	Paignton Zoo	HS Ggg_28	110	Excluded from population analyses
Ggg_37	Kesho	Western lowland gorilla	*Gorilla gorilla gorilla*	M	ZSL Zoo		57	
Ggg_38	Bikira	Western lowland gorilla	*Gorilla gorilla gorilla*	F	Belfast Zoological Gardens		49.57	
Ggg_39	Kanghu	Western lowland gorilla	*Gorilla gorilla gorilla*	M	Howletts Zoo Park	FS Ggg_24 (Kaja); HS Ggg_41 (Fubu)	25.16	Excluded from population analyses
Ggg_41	Fubu	Western lowland gorilla	*Gorilla gorilla gorilla*	M	Howletts Zoo Park	HS Ggg_34 (Kaja) Ggg_39 (Kanghu) Ggg_42 (Kuoyou)	10.02	Excluded from population analyses
Ggg_42	Kuoyou	Western lowland gorilla	*Gorilla gorilla gorilla*	M	Howletts Zoo Park	HS Ggg_41 (Fubu)	9.80	
Gbg_43	Itebero	Eastern lowland gorilla	*Gorilla gorilla gorilla*	F	Confiscated		110	
Gbg_44	Tumani	Eastern lowland gorilla	*Gorilla beringei graueri*	F	Confiscated		110	
Gbg_45	Pinga	Eastern lowland gorilla	*Gorilla beringei graueri*	F	Confiscated		120	
Gbg_46	Ntabwoba	Eastern lowland gorilla	*Gorilla beringei graueri*	M	Confiscated		110	
Gbg_47	Dunia	Eastern lowland gorilla	*Gorilla beringei graueri*	F	Confiscated		115	
Gbg_48	Serufull	Eastern lowland gorilla	*Gorilla beringei graueri*	F	Confiscated		95	

Gbb_49	Bwiruka 1	Mountain gorilla	*Gorilla beringei beringei*	F	Bwindi (Uganda)	HS Gbb_50 (Kahungye)	112	
Gbb_50	Kahungye	Mountain gorilla	*Gorilla beringei beringei*	F	Bwindi (Uganda)	HS Gbb_49 (Bwiruka 1) Gbb_51 (Katungi)	110	
Gbb_51	Katungi	Mountain gorilla	*Gorilla beringei beringei*	F	Bwindi (Uganda)	HS Gbb_49 (Bwiruka 1) Gbb_50 (Kahungye)	20.87	
Gbb_53	Nyamunwa	Mountain gorilla	*Gorilla beringei beringei*	F	Bwindi (Uganda)		60	
Gbb_55	Ishema	Mountain gorilla	*Gorilla beringei beringei*	F	Rwanda	PO Gbb_56 (Imfura) HS Gbb_61 (Turimaso)	90	
Gbb_56	Imfura	Mountain gorilla	*Gorilla beringei beringei*	M	Rwanda	PO Gbb_55 (Ishema); HS Gbb_60 (Zirikana)	1010	Excluded from population analyses
Gbb_57	Umurimo	Mountain gorilla	*Gorilla beringei beringei*	F	Rwanda	PO Gbb_60 (Zirikana)	1000	Excluded from population analyses
Gbb_58	Kaboko	Mountain gorilla	*Gorilla beringei beringei*	M	Democratic Republic of the Congo	FS Gbb_59 (Maisha)	980	
Gbb_59	Maisha	Mountain gorilla	*Gorilla beringei beringei*	F	Democratic Republic of the Congo	FS Gbb_58 (Kaboko)	110	Excluded from population analyses
Gbb_60	Zirikana	Mountain gorilla	*Gorilla beringei beringei*	M	Rwanda	HS Gbb_56 (Imfura) Gbb_61 (Turimaso); PO Gbb_57 (Umurimo)	1145	
Gbb_61	Turimaso	Mountain gorilla	*Gorilla beringei beringei*	F	Rwanda	HS Gbb_55 (Ishema) Gbb_60 (Zirikana) Gbb_62 (Tuck)	1010	
Gbb_62	Tuck	Mountain gorilla	*Gorilla beringei beringei*	F	Rwanda	HS Gbb_61 (Turimaso)	1200	

2.3 Results

A set of DNA samples was assembled from 52 non-human great-ape individuals (14 chimpanzees, 4 bonobos, 16 western lowland gorillas, 6 eastern lowland gorillas, and 12 mountain gorillas) for sequencing. Both prior information on some sampled individuals and later deductions from data using the software ML-Relate (Kalinowski *et al.*, 2006) (Table 2.1) indicated that the sample set included close relatives within (sub)species, including some parent-offspring, full-sib and half-sib pairs, though no mother-father-child trios. In describing the diversity of STR sequences and in considering identification at the individual and (sub)species levels, all individuals were retained as they contributed new alleles to the dataset. When considering population structure, heterozygosity, inbreeding (F_{IS}), and forensic parameters close relatives were removed from the dataset, apart from in mountain gorillas. In fact, for this sub-species whole-genome sequencing (Xue *et al.*, 2015) has shown chromosomes to be homozygous over >38% of their lengths; given this very high general relatedness, it seems reasonable to retain pairs suggested as half-sibs by ML-Relate.

2.3.1 Amplification of orthologs of human loci in the multiplex

The ForenSeq™ DNA Signature Prep Kit, with a human-based design, was used to amplify autosomal, X- and Y-STRs and autosomal SNP-containing loci in the set of 52 African great ape samples. Appendix A summarises amplification results across the entire set of 152 amplicons in the multiplex. Here, despite the focus on the 27 autosomal STRs (sequences given in Appendix B), the sequences of amplified X-STRs are also reported in Appendix C. Not all samples were males, so Y-STR data were less extensive, and there was also a relatively high failure rate for amplifying orthologs of human loci (ranging from nearly 50% to 95% of the total number of markers). The amelogenin sex test loci (Sullivan *et al.*, 1993) amplified in all individuals and gave results consistent with previously known sex (data not shown). Sequence information for the human identity-informative SNP amplicons is not reported here.

Of the 27 autosomal STRs targeted in the multiplex, two (D7S820 and D9S1122) failed to amplify in any individuals. D5S818 amplifies only in gorillas, but contains a low-diversity STR array with the structure (AGAT)1-2(AG)9-13, unlike the human ortholog

which is a highly variable tetranucleotide repeat, (AGAT)6-18; therefore it was not considered further here. Of the remaining 24 STRs, six (Figure 2.2) could be analysed only in particular species, likely due to inter-specific sequence differences affecting primer sites. A set of 18 STRs amplifies in all species, but with some missing data in particular individuals. Missingness could be due to null alleles arising from sequence variants affecting primer sites, or to poor sequence quality. Neglecting all STRs that show missing data leaves a 'core' set of thirteen STRs that were sequenced across all individuals; this set allows cross-species comparisons to be done.

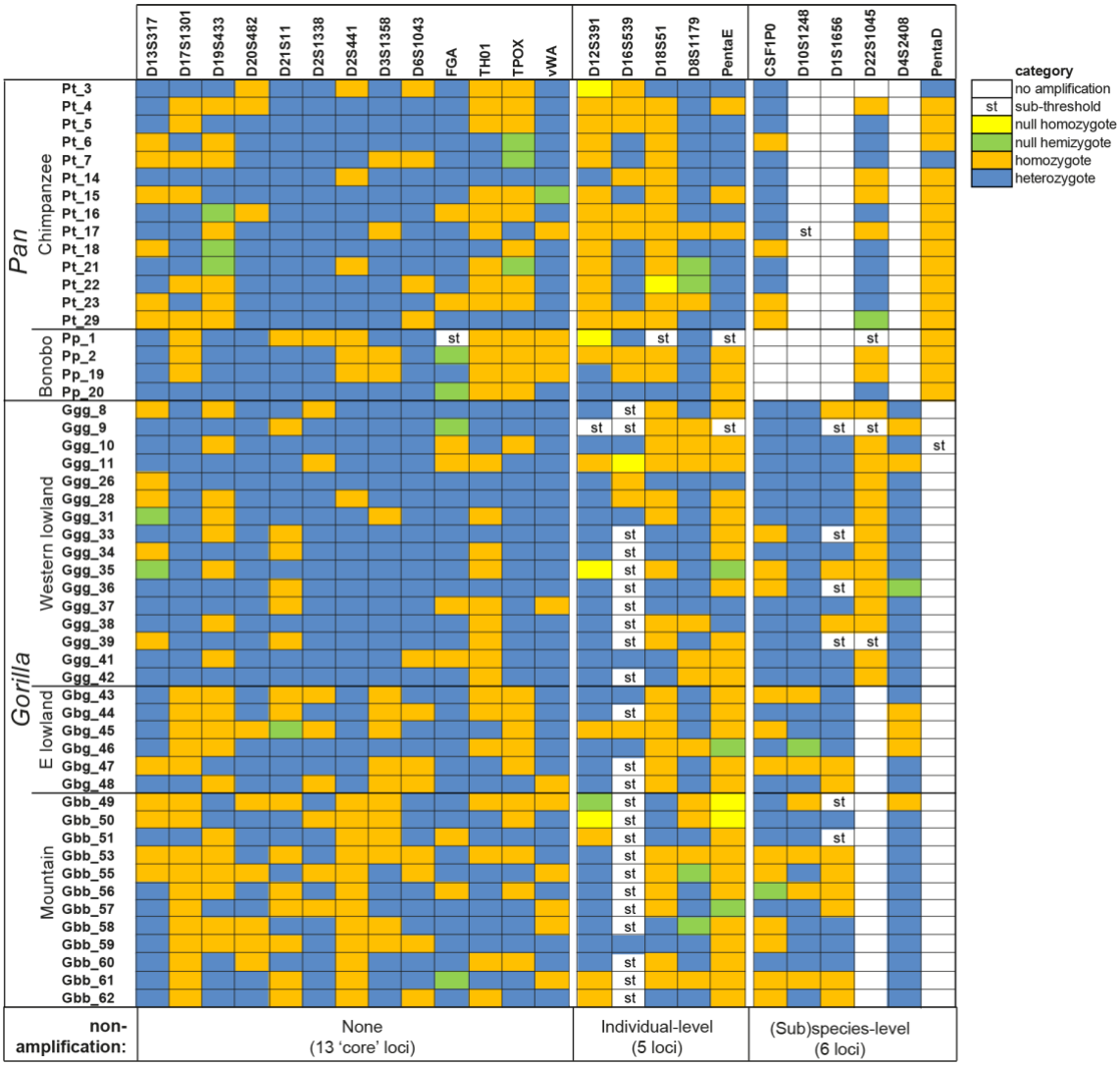


Figure 2.2: Summary of amplification behaviours of autosomal STRs across individuals. For each STR and each great-ape individual, amplification behaviour is summarised, as indicated in the key to the right. Distinction between categories is based on sequence read-depth analysis. STRs are organised into three groups reflecting the amplifiability and degree of data completeness as indicated below the figure.

2.3.2 Sequence diversity in autosomal STRs

In humans, sequencing autosomal STRs increases allelic diversity (Gettings *et al.*, 2016) (Gettings *et al.*, 2016; Khubrani *et al.*, 2019) by allowing variation within both the repeat array and flanking DNA to be observed (Figure 2.3a). This was also the case in the great apes studied here (Figure 2.3b-f; Table 2.2). Focusing on variation within the repeat array (since the lengths of flanking regions are not completely comparable between species), STRs showing sequence variants are not well conserved across species. In humans, D12S391 showed by far the greatest increase in diversity due to repeat array sequence variation (Gettings *et al.*, 2016; Khubrani *et al.*, 2019), but this feature was not observed in the great apes studied here. D2S1338 showed the greatest degree of repeat array sequence variants across (sub)species.

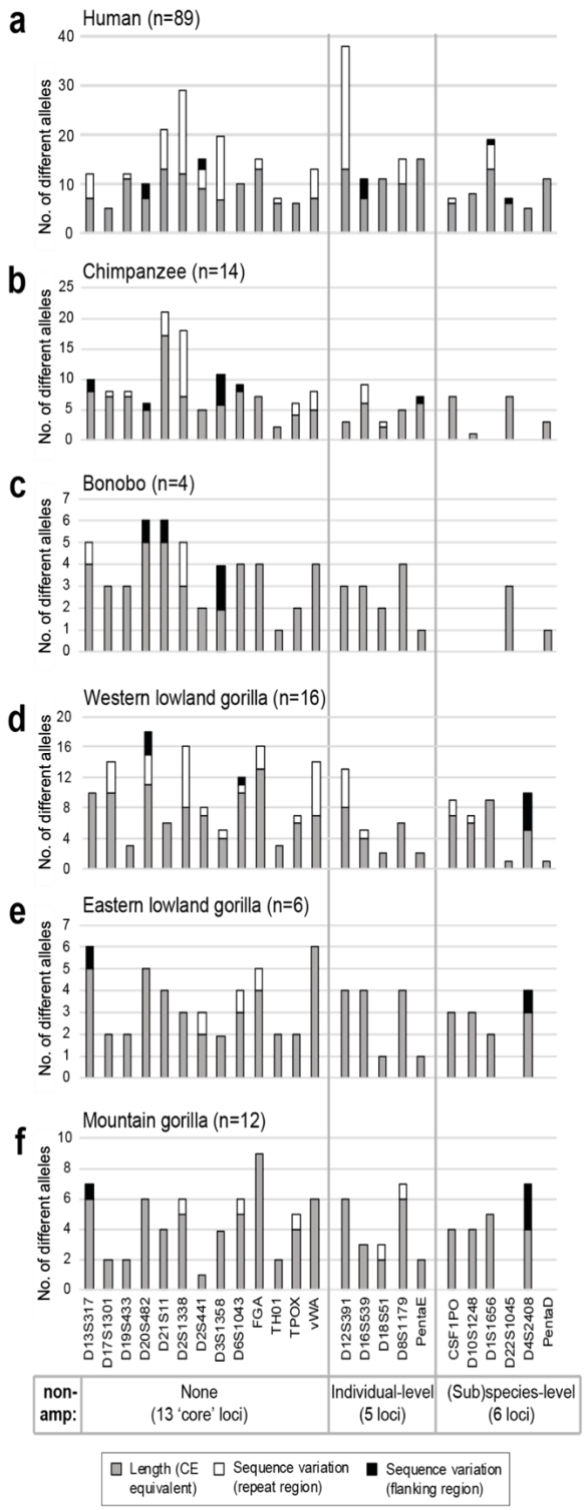


Figure 2.3: Counts of distinguishable alleles in each (sub)species by STR locus, and per-locus increment due to sequence variants. The observed numbers of length variants among individuals are shown as grey bars, and the number of additional alleles resulting from sequence variation within and flanking the repeat array are shown in white and black respectively. STRs are organised into three groups as in Figure 2, and shown below the figure. a) Human (Khubrani *et al.*, 2019); b) Chimpanzee; c) Bonobo; d) Western lowland gorilla; e) Eastern lowland gorilla; f) Mountain gorilla. Note that, although repeat array sequence variation is comparable across species, flanking sequence variation is not strictly comparable because the amount of sequence considered in different species varies somewhat.

Table 2.2: Summary of number of autosomal STR alleles per locus per (sub)species.

Sequence-based alleles	Ppa (n=4)	Ptr (n=14)	Ggg (n=16)	Gbg (n=6)	Gbb (n=12)	*CE-equivalent alleles*	Ppa (n=4)	Ptr (n=14)	Ggg (n=16)	Gbg (n=6)	Gbb (n=12)
CSF1P0	NA	7	9	3	4	**CSF1P0**	NA	7	7	3	4
D10S1248	NA	1	7	3	4	**D10S1248**	NA	1	6	3	4
D12S391	3	3	13	4	6	**D12S391**	3	3	8	4	6
D13S317	5	10	10	6	7	**D13S317**	4	8	10	5	6
D16S539	3	9	5	4	3	**D16S539**	3	6	4	4	3
D17S1301	3	8	14	2	2	**D17S1301**	3	7	10	2	2
D18S51	2	3	2	1	3	**D18S51**	2	2	2	1	2
D19S433	3	8	3	2	2	**D19S433**	3	7	3	2	2
D1S1656	NA	NA	9	2	5	**D1S1656**	NA	NA	9	2	5
D20S482	6	6	18	5	6	**D20S482**	5	5	11	5	6
D21S11	6	21	6	4	4	**D21S11**	5	17	6	4	4
D22S1045	3	7	1	NA	NA	**D22S1045**	3	7	1	NA	NA
D2S1338	5	18	16	3	6	**D2S1338**	3	7	8	3	5
D2S441	2	5	8	3	1	**D2S441**	2	5	7	2	1
D3S1358	4	11	5	2	4	**D3S1358**	2	6	4	2	4
D4S2408	NA	NA	10	4	7	**D4S2408**	NA	NA	5	3	4
D6S1043	4	9	12	4	6	**D6S1043**	4	8	10	3	5
D8S1179	4	5	6	4	7	**D8S1179**	4	5	6	4	6
FGA	4	7	16	5	9	**FGA**	4	7	13	4	9
PentaD	1	3	1	NA	NA	**PentaD**	1	3	1	NA	NA
PentaE	1	7	2	1	2	**PentaE**	1	6	2	1	2
TH01	1	2	3	2	2	**TH01**	1	2	3	2	2
TPOX	2	6	7	2	5	**TPOX**	2	4	6	2	4
vWA	4	8	14	6	6	**vWA**	4	5	7	6	6
TOTAL	66	164	197	72	101	**TOTAL**	59	128	149	67	92

NA: no amplification.

2.3.3 STR variant classes within and between (sub) species

To consider the sequence variation in the 18 cross-species amplifiable STRs in an evolutionary framework, the *Pan* and *Gorilla* data was compared to the predominant sequence structures of human orthologs (retrieved from STRBase.nist.gov and Gettings *et al.* (2016)), as well as to a single orangutan orthologous allele for all loci (where this could be identified), extracted from the orangutan (*Pongo abelii*) reference sequence (ponAbe3 assembly). Figure 2.4a summarises the STR structural categories observed; the range of allele structures for each locus is shown in a phylogenetic context in Figure 2.4 b-h.

Several loci (including D13S317, D19S433, TH01, TPOX and D16S539) show evolutionary conservation across the great apes, with perfect repeat arrays of the same repeat unit across all (sub)species examined, and similar repeat ranges (Figure 2.4). There are no examples in which the major variable repeat unit differs in sequence between (sub)species, but among the remaining loci there is variation in structural types and little obvious relationship with the phylogeny, suggesting stochastic origins of mutations giving rise to diverse non-perfect repeat arrays. Repeat array length distributions are particularly well understood in humans because of very large sample sizes, whereas here great-ape sample sizes are small and may be highly unrepresentative. However, given this caveat, the number of repeats observed in all species fall within the range of human variation, with the exception of D13S317 (based on the lists given by STRBase.nist.gov and Gettings *et al.* (2016)).

Here, some features of structures for the 18 STRs that were amplifiable and sequenced across Pan and Gorilla are summarised. Many of these evolutionary comparisons appear to confirm anthropocentric ascertainment bias: for several STRs (in particular D6S1043, D18S51, D19S433, PentaE and TH01), recorded human allele repeat number ranges are much wider than those seen in sample of great apes analysed in this Chapter. In fact, across all 18 STRs, there is only one case, D13S317 in western lowland gorilla, where the observed non-human primate allele size range exceeds that seen in humans. While this may be influenced by the relatively large surveyed human sample sizes, it seems surprising given that, for example, western lowland gorillas have an effective population size more than twice that of humans (Prado-Martinez *et al.*, 2013), and it may suggest

differences in mutational processes.

a) Some loci behave simply across the phylogeny with a lack of variant structures, and straightforward patterns of variation in perfect arrays. An example is TH01 (Figure 2.4b), which is a simple, perfect array of AATG repeats in humans, and the same across Pan and Gorilla, albeit with narrower repeat number ranges (and invariant in bonobos). The orangutan allele is very short and interrupted, and unlikely to be variable. Similarly simple structures are seen across species at PentaE (Figure 2.4c), D18S51, and D19S433 (Figure 2.4).

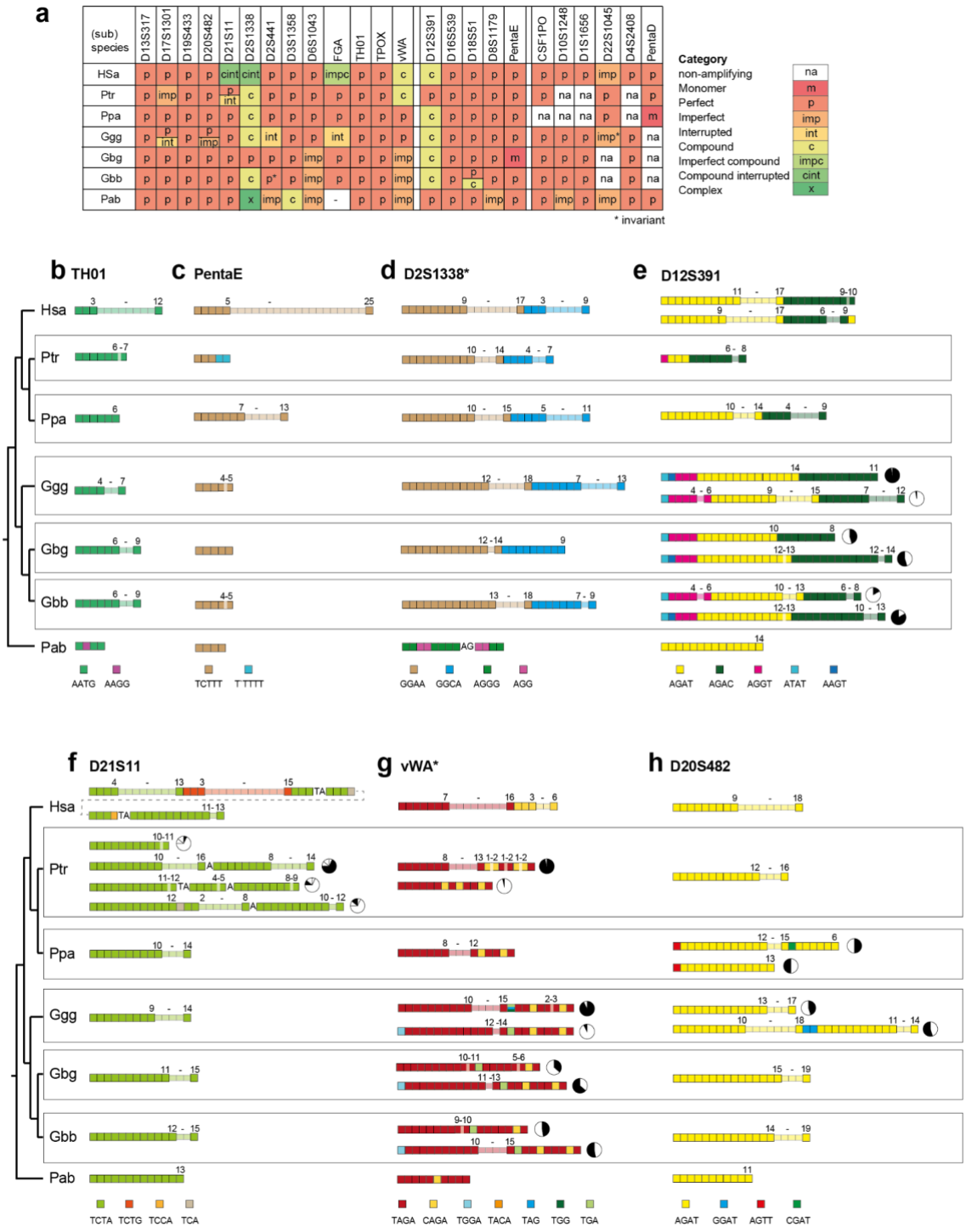
b) Two of the human loci, D2S1338 and D12S391, are compound in humans with two variable blocks of different repeat types. These features are conserved: D2S1338 (Figure 4d) shows similar structure and approximate array length ranges in humans, Pan and Gorilla, as a compound and polymorphic [GGAA]n[GGCA]m STR. Surprisingly, the orangutan allele here comprises short arrays of different repeat units (AGGG and AGG). D12S391 (Figure 2.4e) shows variable arrays of AGAT and AGAC repeats, and in orangutan is a simple perfect array of just one of these repeat types, AGAT.

c) There is little evidence of novel repeat arrays arising and expanding in particular species. One exception is D21S11 (Figure 2.4f), which in all species shows one or more arrays of TCTA repeats, but in humans also includes a highly variable array of TCTG repeats that is not seen in any other species. The other example is at D12S391 (Figure 2.4e), where (as well as the AGAT and AGAC arrays mentioned above) an array of AGGT repeats is specific to Gorilla, and polymorphic in western lowland and mountain gorillas.

d) STR mutation processes are generally thought of as rapid compared to single-nucleotide changes in non-repetitive material, and (unless there has been recent gene flow) little identity-by-descent might therefore be expected in the features of repeat arrays over the several million years of primate evolution. However, this is not so, and the distribution of structures identical by descent appears to be non-uniform across the great apes. There are no examples of distinctive *Pan*-specific derived features in any of the 18 STRs analysed. However, the picture is different in *Gorilla*. For D12S391 (Figure 2.4e), vWA (Figure 4g), D2S441, D16S539, FGA, and TPOX (Figure 2.4), all gorilla

(sub)species studied carry more than one allele structure, and these are shared among western lowland, eastern lowland and mountain gorillas (which have an estimated divergence time of ~150 KYA; Figure 2.1c). Only one locus, D8S1179 (Figure 2.4), shows distinctive structural features restricted to the two eastern subspecies.

e) Considerations of STR array evolution based on human diversity and pedigree data have shown that interrupting a long perfect repeat array with a variant repeat or indel leads to a marked reduction of mutation rate (Sun *et al.*, 2012) and consequent lower allelic diversity. However, in both *Pan* and *Gorilla* there are several allele structures featuring polymorphic arrays separated by interruptions (variant repeats, or insertions). In most of these cases, other variant structures in the same (sub)species are short and perfect, and these are shared across species suggesting they may be ancestral. This suggests that the long interrupted alleles might arise via a non-slippage-like process. Chimpanzee shows this phenomenon at D21S11 (Figure 2.4f) and D17S1301, while it is seen in *Gorilla* at D20S482 (Figure 2.4h), D13S317, D17S1301, and FGA (Figure 2.4).


a
(sub) species
D13S317
D17S1301
D19S433
D20S482
D21S11
D2S1338
D2S441
D3S1358
D6S1043
FGA
TH01
TPOX
vWA
D12S391
D16S539
D18S51
D8S1179
PentaE
CSF1PO
D10S1248
D1S1656
D22S1045
D4S2408
PentaD
HSa
Ptr
Ppa
Ggg
Gbg
Gbb
Pab
* invariant
Category
non-amplifying
Monomer
Perfect
Imperfect
Interrupted
Compound
Imperfect compound
Compound interrupted
Complex
b TH01
c PentaE
d D2S1338*
e D12S391
Hsa
Ptr
Ppa
Ggg
Gbg
Gbb
Pab
AATG AAGG
TCTTT T TTTT
GGAA GGCA AGGG AGG
AGAT AGAC AGGT ATAT AAGT
f D21S11
g vWA*
h D20S482
TCTA TCTG TCCA TCA
TAGA CAGA TGGA TACA TAG TGG TGA
AGAT GGAT AGTT CGAT


i D2S441
j D3S1358
k D6S1043*
l D8S1179
m D16S539
Hsa
Ptr
Ppa
Ggg
Gbg
Gbb
Pab
TCTA TCTG TCCA TTTA CCTA TCA
TCTA TCTG TTTA TCTA
ATCT ATGT GTCT ATC TCCT
TCTA TCTG TCCA TTTA
GATA CATA TAGA TAAA GAGA
n D17S1301
o D18S51
p D19S433
q FGA*
r TPOX
No orangutan ortholog found
AGAT AGAA AAAT GAT AGGT
AGAA TGAA AGAG AGGA
CCTT CTTT CCCT CCTA
GGAA GGAG AAAG AGAA GAAA AAAA GAAA ACAG AGA GAG AAG
AATG GATG ACTG AATA AACG

Figure 2.4: Summary of STR structures across (sub)species, and examples of inter- and intra-specific structural variation. a) For each (sub)species and each locus, the structural class of the STR is summarised as indicated in the key to the right. In cases where two classes are both present at high frequencies, the two classes are given as a split cell in the table. Human structures are taken from the predominant observed class listed at STRBase.nist.gov. Orthologous orangutan (Pab: *Pongo abelii*) alleles are based on the reference sequence. Hsa: *Homo sapiens*; Ptr: *Pan troglodytes*; Ppa: *P. paniscus*; Ggg: *Gorilla gorilla gorilla*; Gbg: *G. beringei graueri*; Gbb: *G. b. beringei*. b - r) Examples of variation across (sub)species, phylogenetically arranged, for seventeen STRs. Human structures are from STRBase.nist.gov and (Gettings *et al.*, 2016). In each case, tetra- or tri-nucleotide repeat motifs are indicated by boxes coloured according to the keys below. Ranges of repeat numbers within variable arrays are indicated. Where more than one structural class is observed within a *Pan* or *Gorilla* (sub)species, pie-charts indicate their proportions.

2.3.4 Within-(sub)species variability of multilocus STR genotypes

Within (sub)-species, all individuals (including related individuals; Table 2.1) are distinguishable by their STR genotypes, and this is true for both CE-equivalent and sequence-based allele designations.

After removing related individuals (Table 2.1) observed vs expected heterozygosity was assessed for the tested loci (Table 2.3); following Bonferroni correction, only one locus in one species (D16S539 in chimpanzee), shows a significant deviation from expectation. To ask if the tested loci reveal any evidence of inbreeding, F_{IS} was estimated (Table 2.4); following Bonferroni correction, significant positive F_{IS} values are seen for three loci (D16S539, D19S433, TPOX) in chimpanzee, two (D13S317, D16S539) in western lowland gorilla, and one (D8S1179) in eastern lowland gorilla. As shown in Figure 2, all except one of these (TPOX) show evidence of null alleles or low read-depth in the relevant (sub)species, suggesting that the F_{IS} results reflect technical issues rather than evidence of inbreeding. Forensic statistics derived from the data are given in Appendix D, and Table 2.5 presents the combined random match probabilities (RMPs) in each (sub)species. The values obtained strongly reflect the sample sizes, which in turn influence the mean number of alleles observed per locus. RMPs are in all cases lower for MPS than CE allele designations, and in the range 10^{-8} to 10^{-18}. Any comparison with human RMPs, where sample sizes and numbers of observed alleles are much larger, is not very meaningful. For example, the 24 loci analysable in western lowland gorillas give respective RMPs for CE- and MPS-based designations of 1.49×10^{-27} and 1.98×10^{-30} in a sample of 89 Saudi Arabian humans (Khubrani *et al.*, 2019).

Table 2.3: Observed vs expected heterozygosity.

Locus	*Pan paniscus*			*Pan troglodytes*			*Gorilla gorilla gorilla*			*Gorilla beringei graueri*			*Gorilla beringei beringei*		
	H_O	H_E	*p*	H_O	H_E	*p*	H_O	H_E	*p*	H_O	H_E	*p*	H_O	H_E	*p*
CSF1P0	n/a	n/a	n/a	0.75	0.78	0.4239	0.83	0.8	0.8800	0.5	0.57	0.0687	0.44	0.44	0.4747
D10S1248	n/a	n/a	n/a	0	0	1.0000	1	0.82	0.5213	0.5	0.57	0.4530	0.67	0.5	0.8953
D12S391	0.67	0.67	0.1116	0.09	0.24	0.0116	0.91	0.87	0.6882	0.83	0.65	0.8866	0.5	0.56	0.9572
D13S317	1	0.72	0.3528	0.58	0.78	0.0754	0.5	0.82	0.3128	0.83	0.69	0.9365	0.56	0.77	0.0992
D16S539	0.5	0.41	0.9309	0.42	0.81	**0.0010**	0.22	0.74	0.0062	0.5	0.69	0.2073	0	0.5	0.1573
D17S1301	0.25	0.53	0.2447	0.42	0.49	0.0098	1	0.85	0.1333	0.17	0.38	0.1736	0.11	0.28	0.0719
D18S51	0.25	0.47	0.3506	0.09	0.24	0.0116	0.42	0.41	0.9768	0	0	1.0000	0.56	0.51	0.8784
D19S433	1	0.59	0.2615	0.33	0.8	0.0027	0.42	0.35	0.8420	0.17	0.15	0.8238	0.56	0.4	0.2486
D1S1656	n/a	n/a	n/a	n/a	n/a	n/a	0.73	0.84	0.3635	0.67	0.44	0.2207	0.44	0.73	0.4138
D20S482	1	0.78	0.6790	0.75	0.65	0.9291	1	0.93	0.6288	0.83	0.76	0.6288	0.56	0.78	0.2498
D21S11	0.75	0.81	0.3821	1	0.95	0.6037	0.67	0.73	0.5868	0.67	0.65	0.2911	0.56	0.57	0.8407
D22S1045	0.25	0.53	0.2447	0.5	0.8	0.2657	0	0	1.0000	n/a	n/a	n/a	n/a	n/a	n/a
D2S1338	0.75	0.75	0.5855	1	0.93	0.7664	0.92	0.92	0.6196	0.5	0.4	0.8810	0.78	0.72	0.8118
D2S441	0.25	0.22	0.7751	0.83	0.7	0.9040	0.92	0.83	0.3180	1	0.62	0.2530	0	0	1.0000
D3S1358	0.5	0.66	0.5438	0.83	0.82	0.9972	0.92	0.72	0.8268	0.17	0.15	0.8238	0.44	0.61	0.1999
D4S2408	n/a	n/a	n/a	n/a	n/a	n/a	0.92	0.85	0.4547	0.5	0.51	0.4665	0.89	0.77	0.6635
D6S1043	1	0.75	0.6767	0.67	0.78	0.0363	1	0.9	0.7579	0.5	0.65	0.3208	0.67	0.74	0.0628
D8S1179	1	0.66	0.6767	0.75	0.75	0.2237	0.58	0.72	0.5717	0.83	0.69	0.3147	0.33	0.81	0.0804
FGA	0.25	0.72	0.2073	0.83	0.82	0.4388	0.83	0.89	0.1867	1	0.78	0.2429	0.78	0.85	0.5041
PentaD	0	0	1.0000	0.17	0.16	0.9919	0	0	1.0000	n/a	n/a	n/a	n/a	n/a	n/a
PentaE	0	0	1.0000	0.75	0.75	0.3062	0.25	0.22	0.6207	0	0	1.0000	0.14	0.13	0.8387
TH01	0	0	1.0000	0.33	0.28	0.4884	0.5	0.54	0.7212	0.5	0.49	0.9442	0.56	0.4	0.2486
TPOX	0	0.38	0.0455	0.33	0.8	0.0335	0.92	0.77	0.6198	0.17	0.15	0.8238	0.56	0.62	0.9603
vWA	0.25	0.72	0.2073	0.83	0.82	0.2554	0.92	0.87	0.3347	0.83	0.75	0.4688	0.56	0.78	0.3237
no. of loci	20			22			24			22			22		
Bonferroni-corrected *p*-value	0.0025			0.0023			0.0021			0.0023			0.0023		

p-values in bold are significant following Bonferroni correction: H_O: observed heterozygosity; H_E: expected heterozygosity; n/a: data not available.

Table 2.4: F_{IS} values.

locus	*Pan paniscus*		*Pan troglodytes*		*Gorilla gorilla gorilla*		*Gorilla beringei graueri*		*Gorilla beringei beringei*	
	F_{IS}	*p*-value	F_{IS}	*p*-value	F_{IS}	*p*-value	F_{IS}	*p*-value	F_{IS}	*p*-value
CSF1P0	n/a		0.0791	0.3088	-0.0046	0.6908	0.2105	0.2134	0.0588	0.5189
D10S1248	n/a		n/a		-0.1733	0.6082	0.2105	0.3894	-0.28	1
D12S391	0.2	0.4712	0.6552	0.0418	0	0.4875	-0.1905	1	0.1765	0.769
D13S317	-0.2632	0.7708	0.2936	0.0268	0.4286	**0.0019**	-0.1111	1	0.3333	0.0541
D16S539	-0.0909	1	0.5154	**0.0007**	0.7288	**0.0009**	0.4	0.3168	1	0.3307
D17S1301	0.625	0.1457	0.1852	0.1878	-0.1282	0.3901	0.6154	0.2706	0.6364	0.1738
D18S51	0.5714	0.4248	0.6552	0.0489	0.0351	1	n/a		-0.0256	1
D19S433	-0.6	0.3072	0.6089	< **0.0001**	-0.1458	1	n/a		-0.3333	1
D1S1656	n/a		n/a		0.1837	0.5383	-0.4286	1	0.4435	0.0576
D20S482	-0.1429	1	-0.1061	0.8088	-0.0353	1	0	0.9442	0.3388	0.0436
D21S11	0.2174	0.3099	-0.0115	1	0.133	0.4911	0.0698	0.3482	0.0909	0.8042
D22S1045	0.625	0.1413	0.4133	0.0174	n/a		n/a		n/a	
D2S1338	0.1429	0.6633	-0.0312	1	0.0472	0.4904	-0.1538	1	-0.0182	0.7428
D2S441	n/a		-0.1399	0.9652	-0.0661	0.8085	-0.5385	0.6539	n/a	
D3S1358	0.3684	0.3088	0.0308	0.9678	-0.241	0.849	n/a		0.3263	0.0979
D4S2408	n/a		n/a		-0.0298	0.8684	0.1176	0.5261	-0.1034	0.727
D6S1043	-0.2	1	0.1852	0.3494	-0.0732	0.7444	0.3182	0.3296	0.1579	0.0267
D8S1179	-0.4118	1	0.0388	0.3049	0.2261	0.189	-0.1111	0.3225	0.6279	**0.0007**
FGA	0.7273	0.0301	0.0265	0.5796	0.1093	0.3763	-0.2	0.5571	0.1385	0.5395
PentaD	n/a		-0.0233	1	n/a		n/a		n/a	
PentaE	n/a		0.0388	0.1553	-0.1	1	n/a		n/a	
TH01	n/a		-0.1579	1	0.1141	0.786	0.0625	1	-0.3333	1
TPOX	1	0.1432	0.6106	**0.0006**	-0.1524	0.4299	n/a		0.1667	0.8178
vWA	0.7273	0.0277	0.0222	0.0278	-0.0126	0.2645	-0.0204	0.9554	0.3388	0.1528
no. of loci		16		21		22		17		20
Bonferroni-corrected *p*-value		0.0031		0.0024		0.0023		0.0029		0.0025

p-values in bold are significant following Bonferroni correction.

Table 2.5: Observed per genotype combined RMPs for different great ape (sub)species.

(sub)species	Mean N Chr	N loci	Mean N alleles/locus		Combined RMP	
			CE	MPS	CE	MPS
P. paniscus	6.6	20	3.0	3.1	6.91^{-8}	5.15^{-8}
P. troglodytes	20.9	22	5.8	7.0	7.78^{-15}	2.17^{-15}
G. g. gorilla	22.6	24	6.0	7.3	1.46^{-17}	1.61^{-18}
G. b. graueri	10.8	22	3.0	3.2	6.21^{-10}	5.70^{-10}
G. b. beringei	15.7	22	4.1	4.3	3.96^{-12}	8.53^{-13}

2.3.5 Between-(sub)species variability of STR genotypes

To compare multilocus STR genotypes for the 13 'core' loci across (sub)species, cluster analysis was conducted using STRUCTURE and DAPC, both for data at the full sequence level and for CE-equivalent (length-based) data. In STRUCTURE analysis of CE-equivalent data (Figure 2.5a), the best-supported value of *K* is 2, in which *Pan* and *Gorilla* form two clusters. DAPC analysis reveals three clusters, with *Gorilla* divided into clusters corresponding to western and eastern species (Figure 2.5b), reflecting the behaviour of this method in minimising differences within, while maximising differences between, populations. In STRUCTURE analysis of sequence-level data, $K = 4$ is best supported (data not shown), differentiating clusters corresponding to bonobo, chimpanzee, western gorilla and eastern gorilla (Figure 2.6a). DAPC analysis gives five clusters, separating out the two eastern gorilla subspecies (Figure 2.6b). Sequence-based analysis therefore performs better in distinguishing between (sub)species. Given the sharing of repeat motif variation across *Gorilla* (sub)species (Figure 2.4), it seems likely that the differences contributing to differentiation here reflect variation in the flanking sequences.

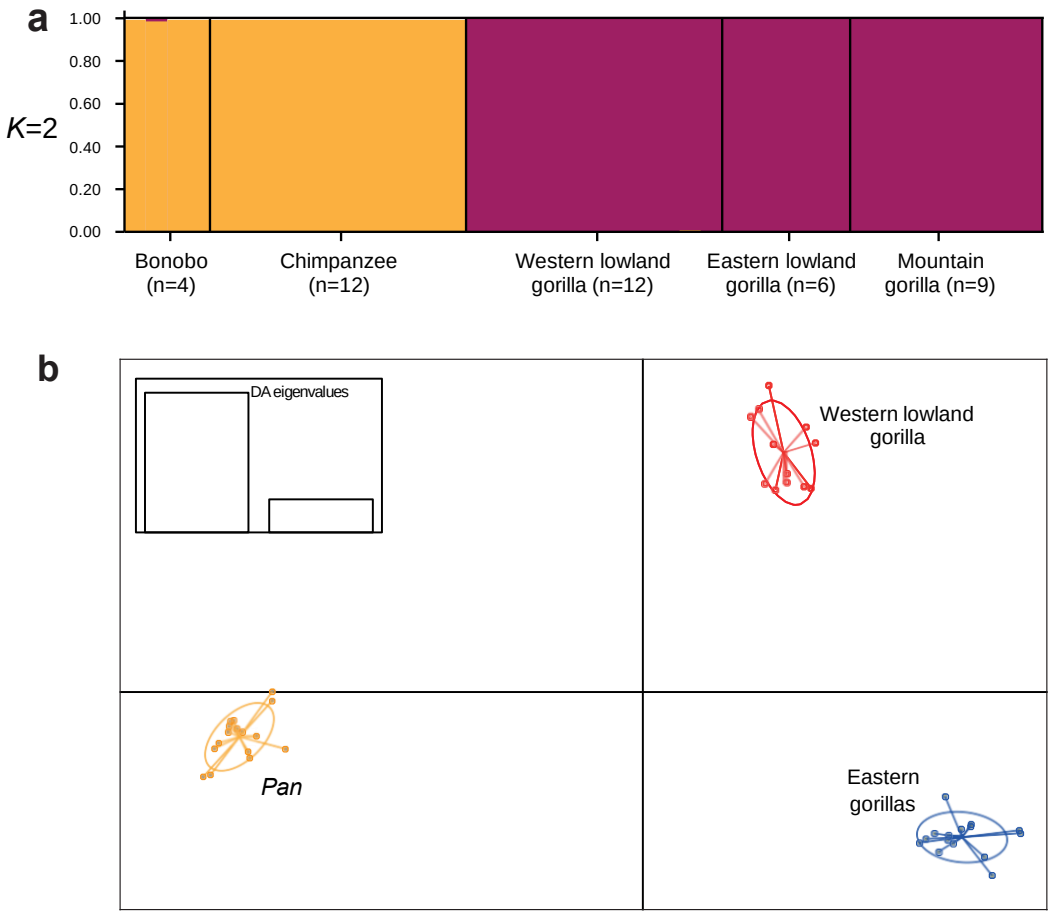


Figure 2.5: Cluster analysis based on CE-equivalent autosomal STR genotypes. a) Results based on STRUCTURE, for *K = 2*; b) Results based on DAPC analysis. Related individuals are removed for this analysis (see Table 2.1).

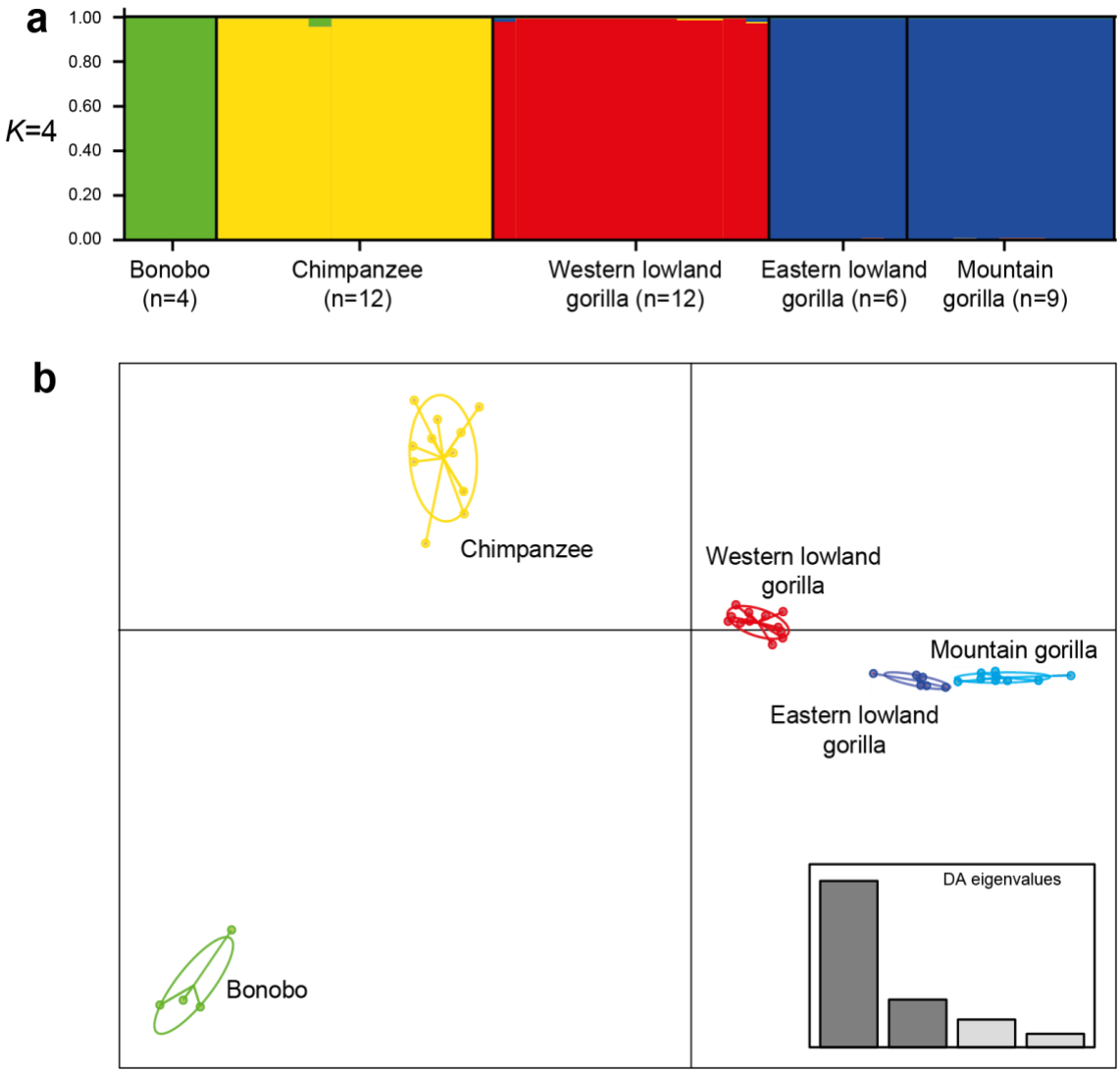


Figure 2.6: Cluster analysis based on sequence-based autosomal STR genotypes. a) Results based on STRUCTURE, for K=4; b) Results based on DAPC analysis. Full sequence information was used here (both array and flanking sequence data). Related individuals are removed for this analysis (see Table 2.1).

2.4 Discussion

Recent conservation initiatives have witnessed a considerable increase in the use of DNA testing for the implementation of effective wildlife conservation and management plans throughout the world. The current rate of biodiversity loss has prompted researchers to utilise markers that can be readily transferred between species to facilitate the study of taxa in which allelic diversity is poorly characterised (Barbara *et al.*, 2007; Maduna *et al.*, 2014). In this context, aSTRs have been a dominant source of neutral genetic markers for a variety of applications, including individual identification, assessment of population diversity and structure, and evolutionary studies (FitzSimmons *et al.*, 1995). Cross-species amplification depends on the presence of flanking sequences that, despite sometimes long divergence times, are conserved across organisms, and is directly related to the phylogenetic distance between the source and the target species (Miles *et al.*, 2009; Primmer *et al.*, 2005). This has enabled the exploitation of common sets of PCR primers to type orthologous aSTR loci via capillary electrophoresis (CE) for the study of non-model organisms (Blanquer-Maumont and Crouau-Roy, 1995; Brohede and Ellegren, 1999; Clisson *et al.*, 2000; FitzSimmons *et al.*, 1995; Gugerli *et al.*, 2008; Kwong and Pemberton, 2014). Following CE, PCR fragment lengths are converted into numbers of repeats at STR regions to produce individual genotypes. Recent studies, however, have identified several caveats to this approach, especially when it is used in cross-species analyses. Firstly, owing to convergent mutations, repetitive regions that are identical by state (*i.e.* have the same length) may not be identical by descent (Estoup *et al.*, 2002), therefore estimates of differentiation across species can be inaccurate. Secondly, CE fails to distinguish indels occurring within STR flanking sequences from changes in the structure of the repetitive regions, compromising the assessment of the organisation and variability of STRs (Kwong and Pemberton, 2014). As a result, the underlying assumption, under which orthologous STRs are commensurable across species, is often incorrect.

In recent years, the advent of MPS has obviated these problems by allowing researchers to investigate the structures of STR alleles, virtually in unlimited numbers. Consequently, MPS tolerates size homoplasy and the occurrence of overlapping ranges between loci that arise when homologous primers are used to genotype different species, as both STR and flanking sequences may not be invariant across species (Blanquer-Maumont and Crouau-

Roy, 1995; Brohede and Ellegren, 1999; Clisson *et al.*, 2000; FitzSimmons *et al.*, 1995; Gugerli *et al.*, 2008; Kwong and Pemberton, 2014). Because MPS does not rely on length discrimination, primer pairs can be strategically designed to target shorter fragments and increase multiplexing capability, thus making this technology particularly suitable for the analysis of highly degraded DNA found in non-invasive samples (*e.g.* faeces and hair (Tytgat *et al.*, 2020)).

Here, the human-designed ForenSeq kit was applied to amplify and sequence human loci of forensic interest in 52 DNA samples from chimpanzees, bonobos, and gorillas, focusing on the results obtained for 27 autosomal STRs (aSTRs). As expected, given the low average sequence divergence between African great ape genomes (~1.3% between human and chimpanzee/bonobo (Prüfer *et al.*, 2012); ~1.75% between human and western lowland gorilla (Scally *et al.*, 2012), most of the aSTRs amplified successfully in most cases. Thirteen STRs could be genotyped in all individuals, and a further five showed only individual-level dropouts or sub-threshold amplification. The remaining nine either failed amplification altogether or failed in a particular species or genus. Failure to amplify is likely due to sequence divergence in primer-binding sites, though this has not been investigated this since the ForenSeq kit's primer sequences are proprietary and therefore not known exactly.

Results showed that MPS analysis of STR alleles can provide accurate individual and sub-species identification. In addition, STR structures show evidence of allele stability over long evolutionary times and reveal unexpectedly high levels of IBD across shared gorilla alleles (the only exception being D8S1179 in eastern gorilla subspecies, which reflects the short divergence time). Contrary to what has been reported in human pedigrees (Sun *et al.*, 2012), long interrupted alleles were found to share a high degree of polymorphism across species, suggesting possible differences in mutation processes between species. In the future, increasing whole-genome sequence data at the population level and the application of genome-wide STR calling tools (*e.g.* HipSTR (Willems *et al.*, 2017), LobSTR (Gymrek *et al.*, 2012)) should illuminate these questions further.

Despite the advantages of MPS, the widespread adoption of high throughput sequence-based STR typing for wildlife conservation purposes is still hindered by high start-up costs (*e.g.* for equipment and reagents), labour-intensive sample preparation, and steep

learning curves associated with MPS data analysis (Hall *et al.*, 2020; Hall *et al.*, 2022; Zascavage *et al.*, 2013). Additionally, the lack of well-established research facilities in biodiverse countries means that biological samples must be shipped to sites where sequencing can be performed (Pomerantz *et al.*, 2018). Stringent international restrictions on the export of endangered species biological samples further contribute to increasing the cost and time of sequencing, *de facto* limiting the feasibility of DNA testing for wildlife conservation purposes (Fernandez, 2011; Gilbert, 2010).

Nevertheless, recent technological advances have circumvented these issues by greatly reducing the cost for the acquisition of sequencing and laboratory equipment, with positive repercussions for the implementation of wildlife conservation genomics initiatives (Hall *et al.*, 2022; Hall *et al.*, 2020). In this regard, the commercialisation of portable nanopore sequencing devices by Oxford Nanopore Technologies Ltd. promises to revolutionise the field of molecular ecology by permitting *in situ* analysis of DNA samples (Blanco *et al.*, 2020; Chang *et al.*, 2020; Menegon *et al.*, 2017; Pomerantz *et al.*, 2018; Pomerantz *et al.*, 2022). The shift from a laboratory-centralised workflow to on-site DNA analysis overcomes the fundamental challenge of transporting biological material to a site where sequencing can be performed (Pomerantz *et al.*, 2018). While few studies to date have assessed the applicability of the ONT MinION™ device for sequencing forensic STRs (Asogawa *et al.*, 2020; Cornelis *et al.*, 2017; Ren *et al.*, 2021; Tytgat *et al.*, 2020), recent findings suggest that STR panels can be compatible with ONT sequencing platforms (Hall *et al.*, 2022), which opens up new opportunities in the field of wildlife forensics and conservation genetics.

Chapter 3 Development of a SNP panel for individual and sub-species identification in gorillas

3.1 Introduction

As discussed in the previous chapter, the advent of massively parallel sequencing (MPS) has played a central role in the progress of conservation genetics and forensic disciplines (Butler, 2007; Shokralla *et al.*, 2015; Pertoldi and Randi, 2018). The ability to simultaneously analyse large numbers of samples, together with the development of universal primers for cross-taxa amplification of short fragments of mtDNA (*e.g.* CytB and COXI), has indeed facilitated the implementation of large-scale monitoring assessment of species diversity and wildlife conservation-related projects (Hebert *et al.*, 2003). Recent improvements in sequencing technology have also favoured the rapid expansion of key reference databases including the Barcode of Life Data System (BOLD - www.boldsystems.org) and the National Center for Biotechnology Information (NCBI - www.ncbi.nlm.nih.gov) GenBank database, further reinforcing the potential of DNA sequencing as a valuable tool for wildlife conservation and monitoring programmes. Following the advances in MPS, researchers have also started to sequence single nucleotide polymorphisms (SNPs), which are reportedly more informative in terms of species and sub-species identification than either STR or mtDNA markers (Marques-Bonet and Hvilsom, 2018; Bourgeois *et al.*, 2019). Not only that, being single-base changes in the genome, SNPs can be amplified successfully even when DNA is scarce or highly fragmented (Andrews *et al.*, 2018; von Thaden *et al.*, 2020). Confiscated samples sent for forensic DNA examination, for example, are often degraded or low in quantity, which makes amplification of standard DNA markers (mtDNA and STRs) and downstream analysis challenging, especially when targeting a few relatively long amplicons (Kumar *et al.*, 2014; Natesh *et al.*, 2019). A similar issue arises when targeting non-invasive samples (*e.g.* hair and faeces) in order to reduce animal disturbance to a minimum while conducting population-level studies (Hohenlohe *et al.*, 2021). In fact, hair contains little good quality DNA (Jeffery *et al.*, 2007b) and the presence of PCR inhibitors and microorganisms in faeces hinders targeted DNA analysis (Taberlet *et al.*, 1999; Morin *et al.*, 1993; Morin *et al.*, 2001; Schultz *et al.*, 2018). SNP-typing via MPS technologies,

paired with ever-more efficient methods for DNA recovery from non-invasive samples (Bourgeois *et al.*, 2019), has allowed researchers to investigate the conservation status of several threatened species (de Flamingh *et al.*, 2023; Khan *et al.*, 2020; Latorre-Cardenas *et al.*, 2020).

However, two main limitations, namely the high initial capital investment for the acquisition of laboratory equipment and the lack of well-established research facilities in many biodiverse countries, have been hampering the application of MPS to wildlife conservation around the globe. Additionally, the widespread adoption of strict international regulations on the export of biological material – such as CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora; https://www.cites.org) – has restricted the transport of samples for analysis, further limiting the applicability of DNA testing.

In this context, miniaturized laboratory instruments, such as portable thermocyclers (BentoLab produced by Bento Bioworks Ltd. and MiniPCR produced by miniPCR bio™) and nanopore sequencing devices (Oxford Nanopore Technologies plc (ONT)), offer the possibility to overcome the limitations posed by CITES on the transport of biological samples and promise to democratise sequencing-based studies by reducing the costs and time, and simplifying the steps for DNA processing and sequencing (Edwards *et al.*, 2016; Quick *et al.*, 2016; Carroll *et al.*, 2018; Krehenwinkel *et al.*, 2019). As a consequence, sequencing is now possible virtually anywhere, including developing countries and remote biodiversity hotspot areas. In addition, the ability to sequence biological samples on site will not only reduce costs, with important repercussions for species monitoring around the globe, but it will also create new opportunities for improving local scientific capacity; the latter is a key aspect of the Nagoya Protocol (www.cbd.int/abs/about/) and the United Nations Sustainable Development Goals (https://sdgd.un.org/goals) (Pomerantz *et al.*, 2022).

For the second part of my PhD, I have focused on the development of a portable laboratory system that can provide sub-species and individual identification in gorillas, based on ONT sequencing of autosomal single nucleotide polymorphism (SNP) amplicons from a variety of biological sources (*e.g.* blood, hair and faeces). The idea was to deliver an innovative approach to the identification of gorillas in a non-laboratory

environment to support *in situ* efforts for the conservation of endangered populations and aid the fight against illegal wildlife trafficking and poaching. SNPs have increased in popularity because they hold clear advantages over more commonly used forensic markers such as STRs (Kidd *et al.*, 2006; Sobrino *et al.*, 2005; Tytgat *et al.*, 2022; Morin *et al.*, 2004; Stronen *et al.*, 2022). In fact, the estimated low mutation rates of 10-8 per base per generation (Reich *et al.*, 2002), compared with rates of 10-3 to 10-5 per locus per generation for STRs (Huang *et al.*, 2002; Dupuy *et al.*, 2004), means that SNPs show virtually no recurrent mutation (Kidd *et al.*, 2006). Additionally, the diallelic nature of SNP markers makes variant calling a qualitative rather than quantitative process, and hence typing is more amenable to automated high-throughput genotyping (Kidd *et al.*, 2006; Gavriliuc *et al.*, 2022; Stronen *et al.*, 2022). Lastly, as SNPs are single base changes, the distance between primers can be reduced considerably; decreasing the amplicon size to 100 – 200 bp can significantly improve PCR success rates especially in degraded DNA samples (Kidd *et al.*, 2006; Tytgat *et al.*, 2022). This has important repercussions for the study of endangered and elusive species in the wild, as it allows the acquisition of information from non-invasively collected samples, such as faeces and hair that often contain either low amounts or highly fragmented DNA (Morin *et al.*, 2004; Ekblom *et al.*, 2021; Stronen *et al.*, 2022; Carroll *et al.*, 2018).

The use of selectively neutral DNA markers has played a central role in the development of molecular ecology over recent decades (Gavriliuc *et al.*, 2022). Several studies have indeed relied on the analysis of neutral markers to infer such parameters as genetic diversity, local adaptation, evolutionary potential, effective population size and taxonomic designations (Allendorf *et al.*, 2010; Frankham, 2010; von Thaden *et al.*, 2020). Among these, STRs have long represented the markers of choice (Taberlet *et al.*, 1996; Tan *et al.*, 2018; Avila *et al.*, 2019). With the advent of MPS however, SNPs have been rapidly replacing conventionally used genetic markers (Marques-Bonet and Hvilsom, 2018; Morin *et al.*, 2004; Amorim and Pereira, 2005; Tytgat *et al.*, 2022).

For the purpose of this project, I utilised published whole genome sequencing (WGS) data to identify a candidate set of SNPs for both gorilla individual and sub-species identification, designed primer pairs to amplify these, and then proceeded to test them using the ONT MinION™ sequencing device.

3.2 MinION™ Sequencing Summary

Here, two sets of SNPs for individual and sub-species identification were identified by exploiting existing whole genome sequencing data from 50 individuals (Table 3.1) published by Prado-Martinez et al. 2013 and Xue et al. 2005. The first set was designed to include individual informative SNP sites (ii-SNPs) – highly heterozygous SNPs across individuals (*i.e.* these are SNPs that are found to be polymorphic in most individuals) – while the second was selected to include species informative SNPs (si-SNPs) for species identification, where one allele is fixed in each species but absent, or at least very rare, in the others. This preliminary selection was conducted by Javi Prado-Martinez at the Wellcome Sanger Institute, utilising the western gorilla Susie3_GSMRT3/gorGor5 assembly (March 2016) deposited in the NCBI under the accession number GCA_900006654.1 (Gordon *et al.*, 2016) resulting in two separate VCF files, one for each set of SNPs, the input was Repeatmasked to exclude repetitious loci. Overall, the aim was to select around 60 ii-SNPs, a number that was considered to be in line with similar studies of forensic SNPs for individual identification in humans (Wei *et al.*, 2012; Pakstis *et al.*, 2010; Børsting *et al.*, 2009), and 30 si-SNPs (*i.e.* 10 per sub-species, including *G. g. gorilla, G. b. beringei*, and *G. b. graueri*).

3.2.1 SNP identification and primer design

For the identification of a subsample of 60 candidate *ii*-SNPs, the VCF file containing 2,098,376 potential individually identifying markers was filtered using VCFtools (Danecek *et al.*, 2011). In order to minimise linkage disequilibrium (LD) among sites, a minimum distance of 1 Mb between SNPs was selected; only diallelic markers were allowed, and finally the lowest minor allele frequency (within the sequenced dataset) was set at 0.5 and maximum minor allele count at 2, in order to avoid both rare alleles and sequencing errors. The individual named Nyango, the only member belonging to the sub-species *G. g. diehli*, was removed from the list of samples as it reportedly showed contamination issues (Javi Prado-Martinez, personal communication). An exact test was used to identify and exclude sites that were found to be out of Hardy-Weinberg Equilibrium (HWE), since such sites might indicate sequencing problems or the influence of factors such as natural selection. Finally, only sites with a minimum read depth equal to 10 were retained, to maximise data quality.

```
vcftools --gzvcf iiSNPs.vcf.gz --thin 1000000 --max-missing
1 --maf 0.5 --min-alleles 2 –max-mac 2 --minDP 10 --minQ 30
--hwe 0.001 --remove-indv Ggd_F_Nyango -recode
```

where:

`--thin 1000000`: 1 Mb of interspacing between sites to avoid LD

`--max-missing 1`: to exclude sites with missing data (1 means no missing data allowed)

`--min-alleles 2`: includes sites with a number of alleles greater or equal to 2

`--max-alleles 2`: to include sites with a number of alleles less or equal to 2

`--maf 5`: minor allele frequency equal to 0.5 (to avoid rare alleles that would bias het)

`--max-mac 2`: maximum minor allele count equal to 2 (to avoid sequencing errors)

`--minDP 10`: to include only sites with read depth greater or equal to 10

`--hwe 0.001`: to assess sites for HWE using an exact test. Sites with a p-value below 0.001 threshold are taken to be out of HWE, and therefore excluded

`--remove-indv Ggd_F_Nyango`: to remove individual G.d.g. Nyango. This showed contamination issues

`--recode`: to generate a new VCF file

After filtering, 49 out of 50 individuals (Gdg_F_Nyango being removed) and 375 out of 2,098,376 sites were retained.

3.2.2 Individual *ii*-SNP selection in R

In order to maximise discriminatory power for individual identification, selected ii-SNPs should ideally display a high degree of heterozygosity and no obvious structure amongst sub-species. Therefore, the derived VCF file was visually inspected using a suite of dedicated R (Team, 2013) packages: Adegenet (Jombart, 2008), vcfR (Knaus and Grünwald, 2017) and dartR (Gruber *et al.*, 2018). A convenient way to assess the behaviour of the 375 candidate SNPs was to plot a heat-map for the individual genotypes (Figure 3.1), where diallelic loci are recorded as 0 (homozygous for the reference allele), 1 (heterozygous) or 2 (homozygous for the alternative allele).

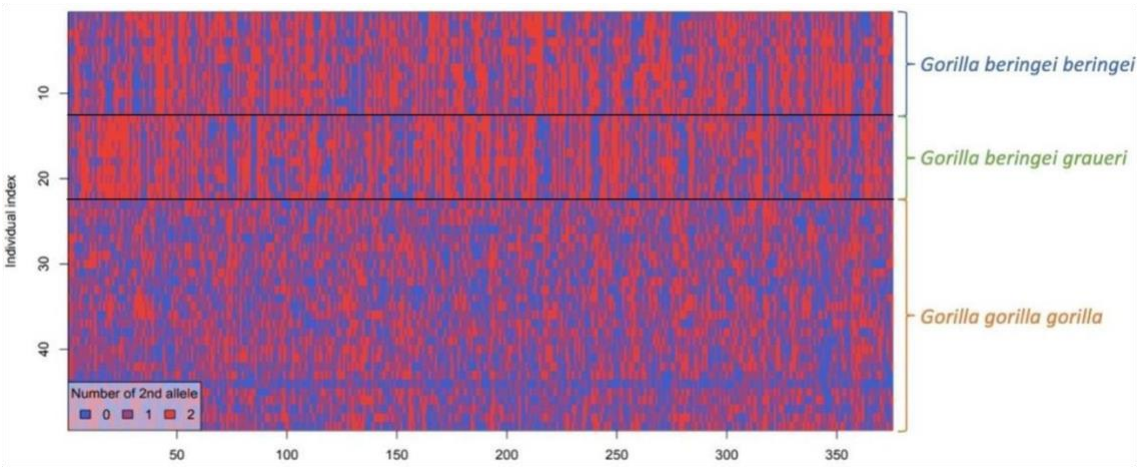


Figure 3.1 Filtered individual genotypes for individual identification SNPs. Individuals are ordered according to sub-species belonging along the y-axis. Consistent with Figure 2, sub-species patterns are still visible, suggesting that SNPs may not be equally heterozygous in each sub-species.

Three different patterns, corresponding to the three gorilla sub-species, are clearly visible in Figure 3.1. This was confirmed via PCA analysis of the SNP genotypes (Figure 3.2), which showed clear clustering, and hence structure, in gorilla samples.

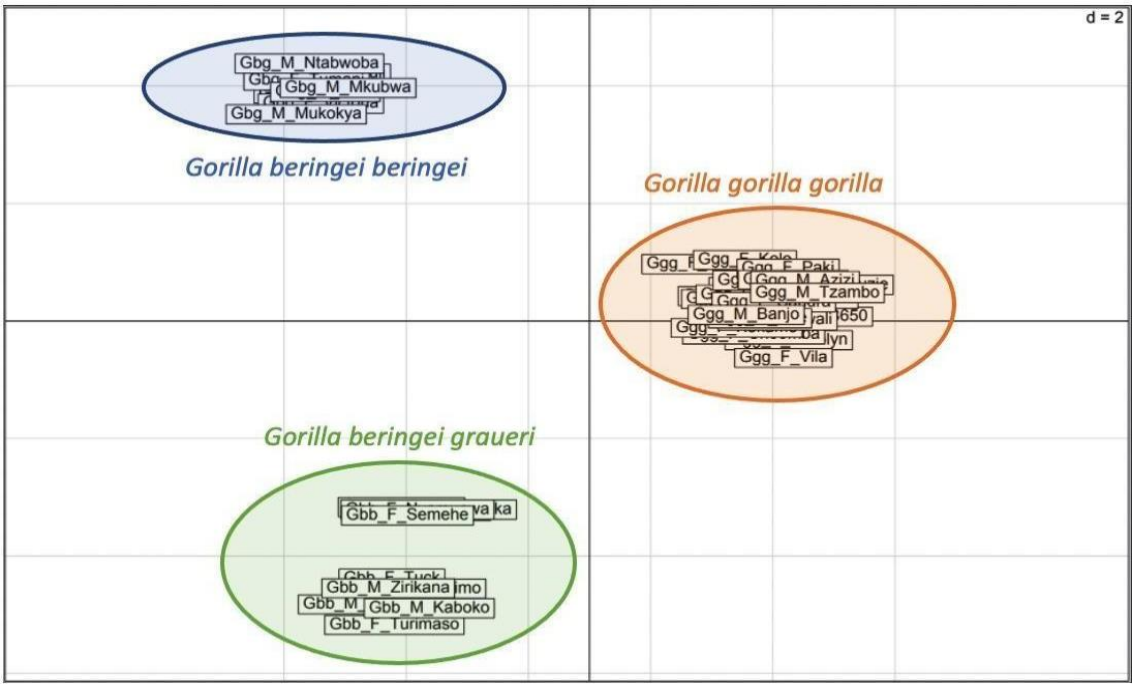


Figure 3.2: PCA scatter plot showing sub-species clustering based on data from the *ii*SNPs.

Following these findings, the number of heterozygous individuals per marker was calculated.

Two to three markers with the highest score (*i.e.* highest number of heterozygotes), for a total of 60 SNPs, were then selected from each chromosome. Once again, data were visually inspected in R (Figure 3.3) and confirmed via PCA analysis (Figure 3.4).

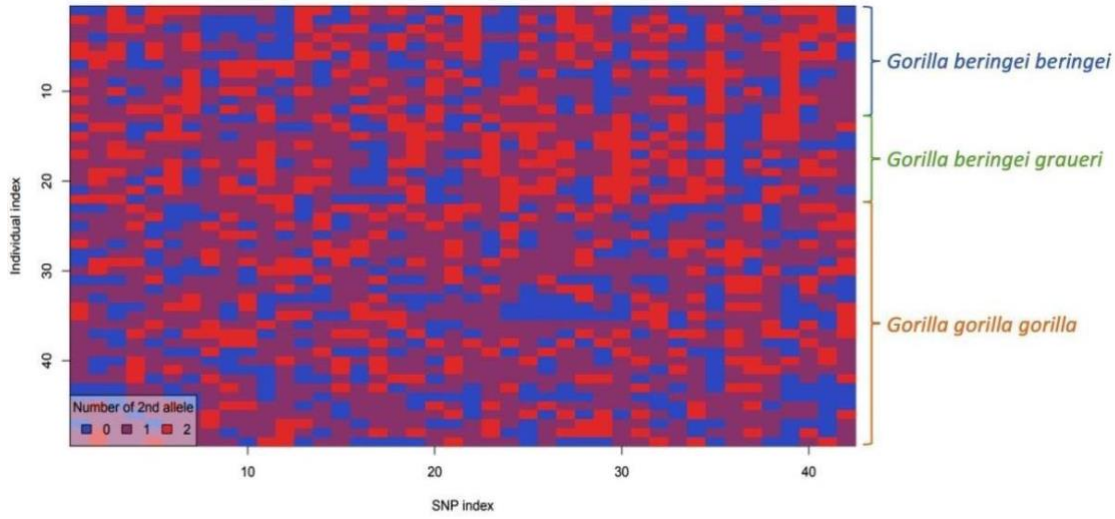


Figure 3.3: **Target iiSNP genotypes for each individual**. Individuals are ordered as in Figure 1, however, in contrast with Figure 3.1, no pattern is now visible (*i.e.* it is not possible to distinguish sub-species), this is consistent with the PCA displayed in Figure 3.4.

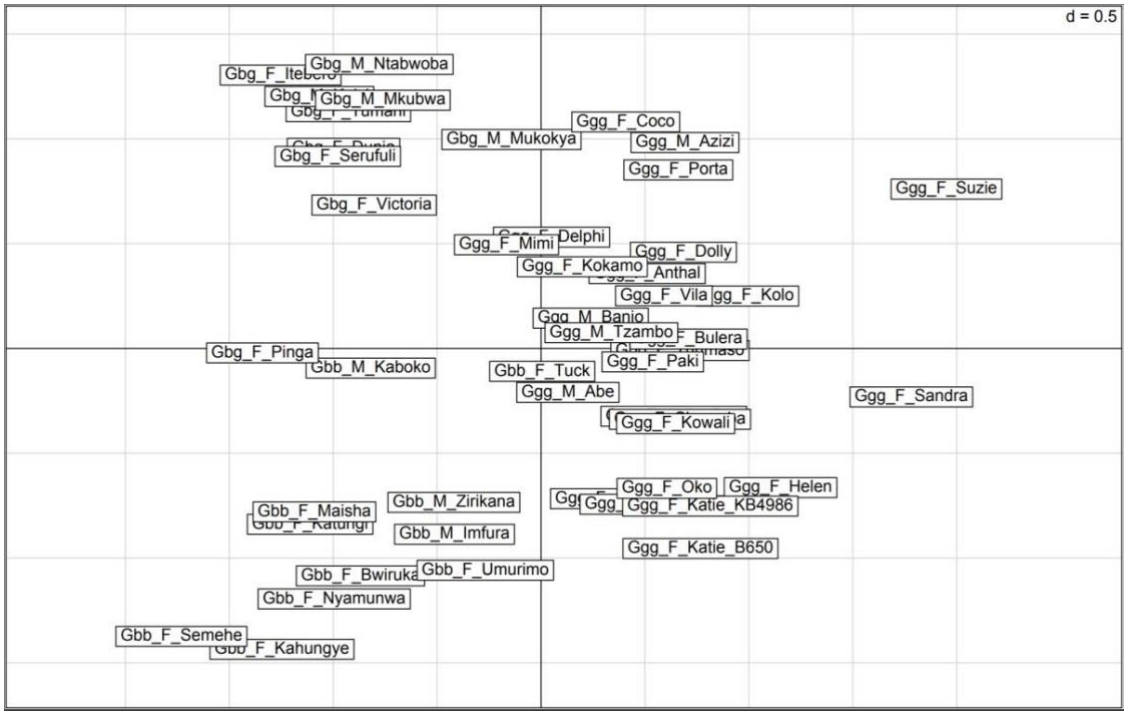


Figure 3.4: PCA scatterplot of iiSNPs showing no clear clustering of sub-species.

Despite a low degree of visible structure, individuals appeared to be more scattered – therefore, the decision was made to retain these SNP markers for individual identification.

As for si-SNPs, ten candidate markers per sub-species were selected randomly from the

provided VCF files, while choosing different chromosomes to avoid LD.

Once the target SNPs were identified, 500 to 650 bp of flanking sequence were extracted from the UCSC Genome Browser using the Kamilah_GGO_v0/gorGor6 assembly (August 2019) deposited in the NCBI under the accession number GCF_008122165.1 (Coordinators, 2015) and added to each side of the target SNP.
The retrieved sequences were then saved in txt files and later stored in FASTA format.

3.2.3 Primer design

Primers flanking ii- and si-SNPs were designed using the proprietary PrimerPlex v2.50 software (www.premierbiosoft.com). The software allows the simultaneous design of primer pairs for a maximum of 30 target sequences at once. Specifically for SNP markers, the user has to specify mutations within DNA sequences (i.e target SNPs) by enclosing both the "wild type" (reference) and the "mutant" (alternative) in square brackets [] separated by a forward slash "/". Sequence information, including SNP mutations, must be provided in FASTA format. FASTA files were generated by appending 500 to 650 bases flanking target ii- and si-SNPs from the UCSC Genome Browser database (Kent *et al.*, 2002) using the Kamilah (G. g. gorilla) assembly ("Kamilah_GGO_v0/gorGor6").

PrimerPlex v2.50 allows the user to set different parameters to avoid cross hybridisation and interactions among primers, and designs primers suited to uniform reaction conditions to enable multiplex amplification under standard experimental conditions. By default, primers for standard multiplex PCR are designed for different amplicon lengths, so as to yield distinguishable bands when visualised on agarose gel. Specifically for this project, primers were designed to amplify segments of DNA 100 - 230 bp in length. Short fragments were preferred so that the panel of SNPs could be used for the analysis of degraded DNA.

The following parameters were selected:

```
Primer Tm: 58.0 +/- 4.0
Primer length: 18 to 25 bp
Amplicon length 100 to 230 bp
Exclude primer binding up and downstream of SNP 30 Amplicon
size difference Min: 3 Max: 130 Alternate Results. 5
```

Advanced primer parameters

```
Hairpin Maximum ΔG (3′ End): 3.0 -kcal/mol Hairpin Maximum
ΔG (Internal): 5.0 -kcal/mol 3′ End Maximum ΔG: 9.0 -
kcal/mol
Self Dimer Maximum ΔG (3′ End): 5.0 -kcal/mol Self Dimer
Maximum ΔG (Internal): 6.0 -kcal/mol Run/Repeat
(dinucleotide) Maximum Length: 3 bp G/C clamp – Target
Consecutive G/Cs at 3′ End: 1 GC% 40.0 to 60.0
```

The software failed to design primer pairs for 28 SNP sequences, which were later designed manually using Primer3 (https://primer3.ut.ee) under the same parameters. Finally, three loci were excluded from the downstream analysis as even Primer3 failed to design primers that could satisfy the parameters described above. Primer sequences and fragment lengths (in bp) are reported in Table 3.2. Allele frequencies for the SNPs of interest within the WGS dataset are displayed in Table 3.3.

In addition to the ii-SNPs and the si-SNPs, a single pair of primers was designed to target the Amelogenin X- and Y-linked genes for sex confirmation of individuals (Pfeiffer and Brenig, 2005; Ensminger and Hoffman, 2002; Sullivan *et al.*, 1993). The same marker, which is also included in the ForenSeq™ DNA Signature Prep kit, already proved successful in determining the sex of great apes. This information is indeed of crucial importance to both conservationists, park managers and authorities to study gorilla population dynamics and fight the illegal wildlife trade (Ensminger and Hoffman, 2002).

Table 3.1: Individuals included in the VCFfiles provided by the Wellcome Trust Sanger Institute. The last two columns indicate whether the individual was from Prado-Martinez *et al.* (2013) or from Xue *et al.* (2015), or if WGS data was provided by Sanger Wellcome Institute (SWI).

Sample ID	Binomial name	Common name	Sex	Prado-Martinez et al.	Xue et al.	SWI
Bwiruka	*Gorilla beringei beringei*	Mountain gorilla	F			✓
Kahungye	*Gorilla beringei beringei*	Mountain gorilla	F			✓
Katungi	*Gorilla beringei beringei*	Mountain gorilla	F			✓
Maisha	*Gorilla beringei beringei*	Mountain gorilla	F		✓	
Nyamunwa	*Gorilla beringei beringei*	Mountain gorilla	F			✓
Semehe	*Gorilla beringei beringei*	Mountain gorilla	F			✓
Tuck	*Gorilla beringei beringei*	Mountain gorilla	F		✓	
Turimaso	*Gorilla beringei beringei*	Mountain gorilla	F		✓	
Umurimo	*Gorilla beringei beringei*	Mountain gorilla	F		✓	
Imfura	*Gorilla beringei beringei*	Mountain gorilla	M		✓	
Kaboko	*Gorilla beringei beringei*	Mountain gorilla	M		✓	
Zirikana	*Gorilla beringei beringei*	Mountain gorilla	M		✓	
Dunia	*Gorilla beringei graueri*	Eastern lowland gorilla	F		✓	
Itebero	*Gorilla beringei graueri*	Eastern lowland gorilla	F		✓	
Pinga	*Gorilla beringei graueri*	Eastern lowland gorilla	F		✓	
Serufuli	*Gorilla beringei graueri*	Eastern lowland gorilla	F		✓	
Tumani	*Gorilla beringei graueri*	Eastern lowland gorilla	F		✓	
Victoria	*Gorilla beringei graueri*	Eastern lowland gorilla	F	✓		
Kaisi	*Gorilla beringei graueri*	Eastern lowland gorilla	M	✓		
Mkubwa	*Gorilla beringei graueri*	Eastern lowland gorilla	M	✓		
Mukokya	*Gorilla beringei graueri*	Eastern lowland gorilla	M			✓
Ntabwoba	*Gorilla beringei graueri*	Eastern lowland gorilla	M		✓	
Nyango	*Gorilla gorilla diehli*	Cross river gorilla	F	✓		
Akiba-beri	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Amani	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Anthal	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Bulera	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Carolyn	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		

Choomba	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Coco	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Delphi	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Dian	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Dolly	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Helen	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Katie_B650	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Katie_KB4986	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Kokamo	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Kolo	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Kowali	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Mimi	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Oko	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Paki	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Porta	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Sandra	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Suzie	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Villa	*Gorilla gorilla gorilla*	Western lowland gorilla	F	✓		
Abe	*Gorilla gorilla gorilla*	Western lowland gorilla	M	✓		
Azizi	*Gorilla gorilla gorilla*	Western lowland gorilla	M	✓		
Banjo	*Gorilla gorilla gorilla*	Western lowland gorilla	M	✓		
Tzambo	*Gorilla gorilla gorilla*	Western lowland gorilla	M	✓		

Table 3.2: **ii- and si-SNP loci, locus IDs, product lengths (bp), and respective primer sequences**

Susie3	Kamilah	Locus ID	Ind/Spp	Length	Forward	Reverse
chr15: 75404625	NC_044617.1: 74600328	GggLoc1	*Ggg*	144	ACCATTTGCTGAATAGGGTGTCTTT	GGTACTGGCATAAATACAGGCACA
chr7: 26430492	NC_044609.1: 27063360	GggLoc2	*Ggg*	169	GATCTTAGTCCAATCTGTACTCCCA	ATCATCACCACCATCATCATTGTCT
chr6: 75612424	NC_044608.1: 76391190	GggLoc3	*Ggg*	188	AGGGCGGGTGTCTTCTTGATAT	ACAACAGATATTGGTAAGGAGGGAA
chr3: 62842266	NC_044605.1: 62489626	GggLoc4	*Ggg*	193	GGACCTGGCCACTATGTTGA	CTTCCCTGCCAAAGACACCT
chr9: 23334056	NC_044611.1: 21246985	GggLoc5	*Ggg*	197	GCCTCCCTCTCATTGTTTTCC	TCCAGAAGCCAACCAAGGAC
chr4: 65145778	NC_044606.1: 65662558	GggLoc6	*Ggg*	198	ACTTGTCCCAGCTCAGATGC	CAAAAGGCAATGAGGGGCAC
chr7: 146303540	NC_044609.1: 144698546	GggLoc7	*Ggg*	199	GGTGGGACAGGGCATCACAT	GCAGTAGTGTTGAGAGGTGGGA
chr3: 69798696	NC_044605.1: 69444435	GggLoc8	*Ggg*	206	GCTAACTGTGGCTGCACTTG	CACGATGGAGCTGGACTGTG
chr3: 73863565	NC_044605.1: 73508748	GggLoc9	*Ggg*	207	GCCTCAAAGCATCTACCTACAG	CCAGGTCATACGTTTTGGGC
chr6: 11537807	NC_044608.1: 11728215	GggLoc10	*Ggg*	207	CGATATTTCCACCTGTCACATGCT	CCAGGGTAGTGCTCAGGGTAAA
chr4: 65173146	NC_044606.1: 65689911	GggLoc11	*Ggg*	220	GGAGTATTGGTGTGCTGGGA	CACAAGCCAACCCACCATTG
chr6: 126820530	NC_044608.1: 129726948	GbgLoc1	*Gbg*	132	ACACAGTTGTCGCTTGGTAT	ACGGGGAGGATGAACATAGGT
chr7: 14127617	NC_044609.1: 14768484	GbgLoc2	*Gbg*	141	TGGATGACATAGAAGGATGGAACTG	AGTCTATCAGGAGACCTCAGAATCT
chr4: 21221995	NC_044606.1: 21687583	GbgLoc3	*Gbg*	148	AAGCCATGCCATTACAACAGACT	TTCCAACCAGTAGAAGAACTGCTAA
chr9: 53904382	NC_044611.1: 51645682	GbgLoc4	*Gbg*	161	GACTGAGGCTGAGATGGCTGAA	ATTCCTCTCTAAATGGGCTGTTCTG
chr5: 89550295	NC_044607.1: 88455911	GbgLoc5	*Gbg*	165	ATGTATGCACAACGGAGGACC	AAAGAGGTGAGGTAGAAAGGTAAC
chr3: 1567661	NC_044605.1: 1531022	GbgLoc6	*Gbg*	178	ACATACTAGGTAGCTTACGGATCAG	CTATTCTCTGAGGCTTGGTACACTT
chr16: 39950317	NC_044618.1: 38039430	GbgLoc7	*Gbg*	196	CTGTCATTACGATGCTAGGTGGTT	TCATCAGAGATATTGGCCTGAAACT
chr15: 14795214	NC_044617.1: 14118600	GbgLoc8	*Gbg*	214	GAGGAATGGTTTGATGCCTTGTGTA	CCGTGACGCAAGTTTACCTATGT
chr11: 28238775	NC_044613.1: 22074032	GbgLoc9	*Gbg*	217	CCTCTTCTCCTCCAAAGGAAAGC	AGGTCTTCTCTGTGCTGTCTATTCT
chr1: 156088577	NC_044602.1: 153402393	GbgLoc10	*Gbg*	226	AGAGCAATCAGGCAAGAGAAAGAAA	CTGGGCAACAGAGCAAGACTC
chr11: 114861666	NC_044613.1: 106702934	GbbLoc1	*Gbb*	133	TTGTAGCTCAAGTGAAGCGC	GCCTGCTACCCTCAAAATGG
chr4: 25789657	NC_044606.1: 26266581	GbbLoc2	*Gbb*	172	TGTGGACCCTAAAGGTACTTCAAAG	GCAGGATGGAGCAAAGATGAGAA
chr7: 14919870	NC_044609.1: 15560133	GbbLoc3	*Gbb*	202	GTACAGAAAGCACTTCTCTGTTCCT	CCAGCACCACCTTTCTTGATGA
chr6: 127004586	NC_044608.1: 129910681	GbbLoc4	*Gbb*	210	CCAACGATCCAATTCTTCCAATGAA	TGTCTGTCTTGATATTCTTCTCCCT
chr16: 13739443	NC_044618.1: 13097759	GbbLoc5	*Gbb*	214	TGTAAGACATCCCCAGCCATG	CTGCACTGTGACCTTGGGAA

chr1: 56707045	NC_044602.1: 52658998	GbbLoc6	*Gbb*	220	GCTGTGTTAAGCATCACCTCGTA	TGACCATACTCTTCCCTGCTACTC
chr8: 135291188	NC_044610.1: 135068928	GbbLoc7	*Gbb*	223	ATGGAGTCTGACCCTGAACTATGG	TGCAGCCTGATCAAGAGTGGAA
chr10: 54022094	NC_044612.1: 53187893	GbbLoc8	*Gbb*	229	AAAGGAGACTATGGCAGAGTGTC	ATGTTTGTCAGTGATTGAGGATTGC
chr15: 13377591	NC_044617.1: 12703779	IndLoc1	Ind	111	ACTTGTGAAGATACAGCCCTAAGAC	AACCGTATTGATGTGGTGCCAAA
chr16: 6051853	NC_044618.1: 4448609	IndLoc2	Ind	122	TCGAGACTCTAGGCAGTAGGAC	GTGACAATGTTGAGATTTCCATCCA
chr1: 90239302	NC_044602.1: 86318636	IndLoc3	Ind	130	GCTCCACCTGAAGTAGTAATGTGTT	CCTGAGTGACAGAGTGAGAACCT
chr14: 52255083	NC_044616.1: 52513763	IndLoc4	Ind	141	GGGCATTTAGGTTGATTCCATGTCT	AACTACCATTTGACTCAGCAATCCT
chr18: 16878850	NC_044620.1: 17267399	IndLoc5	Ind	144	AGTGAGATGGCTAGATTGTATGGTA	CAAATGCTGGAAAGGATGTGAAGAA
chr4: 81099232	NC_044606.1: 81239185	IndLoc6	Ind	149	CCTTGGCTTCTGGGAGAGTATCA	ATATGGTTGGTCTTTCTGAAATGGC
chr1: 106837670	NC_044602.1: 102643419	IndLoc7	Ind	151	GTTGCCAAGGCTAAGGACTGTT	CCTGAAGGAAGAGTGTCAATTAGGT
chr6: 121355018	NC_044608.1: 124273564	IndLoc8	Ind	154	GACAGACTAGGAAGCTCAAGAGGA	CACATGCAGATTCATCTGACACCT
chr4: 113446649	NC_044606.1: 113648078	IndLoc9	Ind	157	TGGCACCACTCACGTTACC	TGGCCACTTCCAGTTCAATGA
chr16: 34023265	NC_044618.1: 32067006	IndLoc10	Ind	159	TAGAGCCACTAAGGTAAGAAGGAAG	ACAACAGGATAATGTCTCACAACTC
chr18: 12783648	NC_044620.1: 13169469	IndLoc11	Ind	161	TGGATGCCCTTTACTTCCTTCTTG	TGACAAACCCACAGCCAACATT
chr13: 86378203	NC_044615.1: 84934104	IndLoc12	Ind	162	AGTGCATAGCTTCAGCATTCTTCT	CTCTACTTTAGATCATGGCGTGCTA
chr5: 66963817	NC_044607.1: 65998574	IndLoc13	Ind	165	ACTAATCAACACAGTTGTCGGAGTA	CCTGAATAGTGGGAACTTGGTTTG
chr14: 2976167	NC_044616.1: 2991602	IndLoc14	Ind	166	GCAGCATACTCATCACTATTGTTAC	TAGCTCACTACTGAAAGCACAGGT
chr12: 72507169	NC_044614.1: 71512743	IndLoc15	Ind	169	AGAAGCTCAAAGAACATCTGGGAAA	TCTGGGAAATCTGCTGTTAATCTGA
chr17: 80853952	NC_044619.1: 79198718	IndLoc16	Ind	170	TCCCCGACAATCAAATGCCT	TGCTCATATCTGCTTGGGGA
chr19: 25933443	NC_044621.1: 21480063	IndLoc17	Ind	170	ACCACAGACTGGGTATTAAACAACA	GTGAGGACACAGCAAGATGGC
chr10: 86116187	NC_044612.1: 85470022	IndLoc18	Ind	172	GGCTGAGAAGAAGATAGGAAGATGT	GCTAACGCATTACAAGAGTAACCAT
chr11: 96511288	NC_044613.1: 88188051	IndLoc19	Ind	175	ATTCCACAGAAACAGAAACCAAAGC	TCTATGTGCTCCAGTGTTAGGTGT
chr5: 101567183	NC_044607.1: 100523769	IndLoc20	Ind	177	TGACAGTGTAAGAAAGATTGCCTTC	CCACCACCAGTATTCTGCCATT
chr20: 33609773	NC_044622.1: 34942601	IndLoc21	Ind	177	AGGTCTGACTCATGAGGGTC	TCAACCCCGAGAAAGCCAAT
chr10: 24112515	NC_044612.1: 23402210	IndLoc22	Ind	179	CCATTGTCTGAATGTATCACACTTG	TGATCCAGTAATTCCACTTCTAGGT
chr4: 8671005	NC_044606.1: 9147002	IndLoc23	Ind	181	CATCTACGCTGTACATGAAACTCCT	GCTTGGTGTAATTGAATCCATCACT
chr12: 52619777	NC_044614.1: 51663608	IndLoc24	Ind	182	AAACCAAGACACAAACCCACACA	GCAGGTAGGTCTTTCAGGAGATGT
chr9: 91636372	NC_044611.1: 89090674	IndLoc25	Ind	182	ACATGCTGCATTGGTTCCAC	GCTTAAAGACCAGGCAGCCT
chr15: 55698155	NC_044617.1: 54887639	IndLoc26	Ind	187	GAGCAGCTACTATATGCCAGGTG	AGTTCAAGGTGAGTTTATGGAGGAG

chr21: 1108782	NC_044623.1: 769212	IndLoc27	Ind	188	TCCAATTAGTACAAGCTCACAGACT	TAAGGACACCAGTAGGAACGGATTA
chr15: 9848855	NC_044617.1: 9202319	IndLoc28	Ind	190	GCCTGTTTGTGTGCCCTGATAA	GCCCAGGTTGTGCAGTAGGTA
chr3: 187034328	NC_044605.1: 186026040	IndLoc29	Ind	191	ATTGAGAGGGCTTGGGTGGA	GCATTCATTCCCTTGGCTCC
chr3: 98988769	NC_044605.1: 97935501	IndLoc30	Ind	191	TCCGTCTAGCATTCTGTGTGAGG	ATGGGTTAATGAGCCTTTGTCCTTG
chr13: 33183132	NC_044615.1: 32288038	IndLoc31	Ind	194	TTATGATCTACCCACAGAAACAGTG	GCTGGATGCTGGAACCCTTC
chr9: 37716631	NC_044611.1: 35551558	IndLoc32	Ind	197	TTCTCACCGCTCTCCCATCT	AGGGGAACACACAGCTTTGT
chr1: 62084611	NC_044602.1: 58070426	IndLoc33	Ind	198	CTCGCACAATTCCTGGGACAAA	TGTTGGGATTACAGTTCATCTGACA
chr8: 96937369	NC_044610.1: 96745909	IndLoc34	Ind	199	GCTCCTGACATATGCCAGTGT	AGCCACCTCACCTAACCTGA
chr16: 63239765	NC_044618.1: 62104656	IndLoc35	Ind	201	CAGTTAGCCAGTTACTGCTGGTATA	GGCACAGAGAAGAGACAGGTACA
chr22: 31942809	NC_044624.1: 30873402	IndLoc36	Ind	201	CACACCTGTAATCCTAGCACTTGG	ACCTCCGCCTCCTGAGTTCA
chr7: 148757113	NC_044609.1: 148024724	IndLoc37	Ind	203	GGGGAAGAGAGCACTTAAGTATCC	GCACAAATGAGATGGGGAAGG
chr14: 36647016	NC_044616.1: 36916075	IndLoc38	Ind	204	GCAGTACAAACTGGGACATTGAGA	ACCTTCTCCTACCATCAGACTGTAT
chr3: 27060183	NC_044605.1: 26996865	IndLoc39	Ind	205	TGCAGCATAACTGAGCTTATCCTG	CAATCTCCTGACCTCGTGATCCA
chr13: 75011706	NC_044615.1: 73564862	IndLoc40	Ind	207	GGATCTCAACGGTCCCATCTCA	AGAAGACAATGCTGGTGCTAGTG
chr18: 15428264	NC_044620.1: 15817105	IndLoc41	Ind	208	TTTGAGCAGCCCTAAACAACTTG	TTCCCACCTTACACATTCTATACCA
chr3: 146742439	NC_044605.1: 145828527	IndLoc42	Ind	209	ATGAGAGTCCAGCTGCAGTG	ACCCTCCCTCATACCCCTTC
chr1: 173509211	NC_044602.1: 170798991	IndLoc43	Ind	210	ACGGCAGGACATAGTGTACCAA	TCAATTTAGTGAGTCGTGACCAGAA
chr9: 23588884	NC_044611.1: 21501651	IndLoc44	Ind	210	AGAAGCTCGATTTGCCTGGA	TGAAGGCAAGGAAATCATGGA
chr6: 54790900	NC_044608.1: 54724583	IndLoc45	Ind	211	ACCTAGAAGGAGAAGAGAAGCAAAG	CAAGCACCGTATCAGGCACTAA
chr12: 7209327	NC_044614.1: 6454099	IndLoc46	Ind	213	TGGTTGGGCAACTTAGGTTTCATT	AGCCAAAGGAAGTTCTTAAAGCAGA
chr11: 70306499	NC_044613.1: 60399321	IndLoc47	Ind	216	ATCTGCACTCCTATGTTCAATGAAG	TGTCTTTCTACGCCTTGCTTATTTC
chr6: 91286337	NC_044608.1: 92052234	IndLoc48	Ind	217	GGATGGACATGCGCTTCACTG	AGACCTCATCAGACACCCAATCTT
chr7: 132581742	NC_044609.1: 131328013	IndLoc49	Ind	218	TGGGGAAGTGCATAGAAGGG	TGGCAGGAGACATCAGAACG
chr17: 74503189	NC_044619.1: 72708222	IndLoc50	Ind	220	AGGGGTGAAAGAGGACAGGT	TGCCTTTCACTTTCTTCCTGC
chr1: 3589914	NC_044602.1: 1240358	IndLoc51	Ind	220	ATGAGAAGAGCATGGAGGTAACTG	AACCACACTGAGATACAGCTATGTC
chr7: 24543502	NC_044609.1: 25177014	IndLoc52	Ind	221	GGAAAGACAACTAAACTAGCCTCAG	TCTGTAACTGCATGGATATAGGAGT
chr22: 2781730	NC_044624.1: 2094293	IndLoc53	Ind	225	CAGCTTGGGTAACAAGAGTGAAACT	TACAAGTGCGAGCCACCATCC
chr11: 80255082	NC_044613.1: 70326181	IndLoc54	Ind	225	AGGAGTAGGCTGGCAGTCATG	ATCTGGTGGCTGGTTAAAGAAGAAA
chr10: 6791289	NC_044612.1: 6626313	IndLoc55	Ind	228	CCATATGAAGACGTGCCTGCTT	AAAGAGGGTAGGGAGCTGAGG

chr21: 3609237	NC_044623.1: 3268087	IndLoc56	Ind	229	ACAACGTATCTGACCAGCACTTC	GCTAACATCGTACATTAGCACAGTT
chrXY: AmelX	NC_044625.1: 9595126	AmelLocXY	Ind	variable	CCCTGGGCTCTGTAAAGAAA	GCCAACCATCAGAGCTTAAAC
chr17: 76725259	NC_044619.1: 75065665	IndLoc57	Ind	FAILED	FAILED	FAILED
chr8: 33277278	NC_044610.1: 33389566	IndLoc58	Ind	FAILED	FAILED	FAILED
chr8: 65820950	NC_044610.1: 65865729	IndLoc59	Ind	FAILED	FAILED	FAILED

Table 3.3: **WGS dataset allele frequencies for the three gorilla sub-species Western (Gbb), Eastern (Gbg), and Mountain (Gbb)**

Locus	Locus ID	Ind/Spp	*Ggg* allele	Frequency	*Gbg* allele	Frequency	*Gbb* allele	Frequency
NC_44617.1: 12703779	IndLoc1	Ind	A	0.54	A	0.2	A	0.67
			G	0.46	G	0.8	G	0.33
NC_44618.1: 4448609	IndLoc2	Ind	T	0.50	T	0.7	T	0.33
			G	0.50	G	0.3	G	0.67
NC_44602.1: 86318636	IndLoc3	Ind	C	0.57	C	0.3	C	0.50
			T	0.43	T	0.7	T	0.50
NC_44616.1: 52513763	IndLoc4	Ind	T	0.35	T	0.65	T	0.71
			G	0.65	G	0.35	G	0.29
NC_44620.1: 17267399	IndLoc5	Ind	G	0.50	G	0.35	G	0.63
			C	0.50	C	0.65	C	0.38
NC_44606.1: 81239185	IndLoc6	Ind	C	0.37	C	0.85	C	0.50
			G	0.63	G	0.15	G	0.50
NC_44602.1: 102643419	IndLoc7	Ind	T	0.67	T	0.45	T	0.17
			C	0.33	C	0.55	C	0.83
NC_44608.1: 124273564	IndLoc8	Ind	T	0.39	T	0.35	T	0.88
			C	0.61	C	0.65	C	0.13
NC_44606.1: 113648078	IndLoc9	Ind	C	0.35	C	0.9	C	0.50
			T	0.65	T	0.1	T	0.50
NC_44618.1: 32067006	IndLoc10	Ind	G	0.59	G	0	G	0.71
			A	0.41	A	1	A	0.29
NC_44620.1: 13169469	IndLoc11	Ind	C	0.56	C	0.2	C	0.63
			T	0.44	T	0.8	T	0.38
NC_44615.1: 84934104	IndLoc12	Ind	G	0.48	G	0.35	G	0.67
			A	0.52	A	0.65	A	0.33
NC_44607.1: 65998574	IndLoc13	Ind	C	0.41	C	0.35	C	0.83
			T	0.59	T	0.65	T	0.17

NC_44616.1: 2991602	IndLoc14	Ind	C	0.67	C	0.25	C	0.33
			A	0.33	A	0.75	A	0.67
NC_44614.1: 71512743	IndLoc15	Ind	C	0.39	C	0.5	C	0.75
			T	0.61	T	0.5	T	0.25
NC_44619.1: 79198718	IndLoc16	Ind	C	0.61	C	0.45	C	0.29
			T	0.39	T	0.55	T	0.71
NC_44621.1: 21480063	IndLoc17	Ind	G	0.76	G	0.25	G	0.13
			A	0.24	A	0.75	A	0.88
NC_44612.1: 85470022	IndLoc18	Ind	G	0.44	G	0.7	G	0.46
			T	0.56	T	0.3	T	0.54
NC_44613.1: 88188051	IndLoc19	Ind	C	0.69	C	0.4	C	0.17
			G	0.31	G	0.6	G	0.83
NC_44607.1: 100523769	IndLoc20	Ind	G	0.39	G	0.45	G	0.79
			C	0.61	C	0.55	C	0.21
NC_44622.1: 34942601	IndLoc21	Ind	G	0.56	G	0.25	G	0.58
			C	0.44	C	0.75	C	0.42
NC_44612.1: 23402210	IndLoc22	Ind	T	0.37	T	0.85	T	0.50
			C	0.63	C	0.15	C	0.50
NC_44606.1: 9147002	IndLoc23	Ind	T	0.39	T	0.65	T	0.63
			C	0.61	C	0.35	C	0.38
NC_44614.1: 51663608	IndLoc24	Ind	G	0.41	G	0.75	G	0.50
			A	0.59	A	0.25	A	0.50
NC_44611.1: 89090674	IndLoc25	Ind	A	0.50	A	0.25	A	0.71
			G	0.50	G	0.75	G	0.29
NC_44617.1: 54887639	IndLoc26	Ind	G	0.48	G	0.45	G	0.58
			C	0.52	C	0.55	C	0.42
NC_44623.1: 769212	IndLoc27	Ind	A	0.26	A	0.65	A	0.92
			C	0.74	C	0.35	C	0.08
NC_44617.1: 9202319	IndLoc28	Ind	G	0.48	G	0.7	G	0.38

			A	0.52	A	0.3	A	0.63
NC_44605.1: 186026040	IndLoc29	Ind	A	0.41	A	0.6	A	0.63
			G	0.59	G	0.4	G	0.38
NC_44605.1: 97935501	IndLoc30	Ind	T	0.44	T	0.3	T	0.79
			C	0.56	C	0.7	C	0.21
NC_44615.1: 32288038	IndLoc31	Ind	C	0.48	C	0.45	C	0.58
			A	0.52	A	0.55	A	0.42
NC_44611.1: 35551558	IndLoc32	Ind	G	0.37	G	0.5	G	0.79
			A	0.63	A	0.5	A	0.21
NC_44602.1: 58070426	IndLoc33	Ind	G	0.59	G	0.4	G	0.38
			T	0.41	T	0.6	T	0.63
NC_44610.1: 96745909	IndLoc34	Ind	T	0.46	T	0.65	T	0.46
			A	0.54	A	0.35	A	0.54
NC_44618.1: 62104656	IndLoc35	Ind	A	0.33	A	0.85	A	0.58
			T	0.67	T	0.15	T	0.42
NC_44624.1: 30873402	IndLoc36	Ind	C	0.41	C	0.6	C	0.63
			T	0.59	T	0.4	T	0.38
NC_44609.1: 148024724	IndLoc37	Ind	C	0.35	C	0.55	C	0.79
			T	0.65	T	0.45	T	0.21
NC_44616.1: 36916075	IndLoc38	Ind	A	0.59	A	0.25	A	0.50
			G	0.41	G	0.75	G	0.50
NC_44605.1: 26996865	IndLoc39	Ind	T	0.33	T	0.75	T	0.67
			C	0.67	C	0.25	C	0.33
NC_44615.1: 73564862	IndLoc40	Ind	T	0.46	T	0.25	T	0.79
			C	0.54	C	0.75	C	0.21
NC_44620.1: 15817105	IndLoc41	Ind	A	0.59	A	0.6	A	0.21
			G	0.41	G	0.4	G	0.79
NC_44605.1: 145828527	IndLoc42	Ind	G	0.61	G	0.35	G	0.38
			A	0.39	A	0.65	A	0.63

NC_44602.1: 170798991	IndLoc43	Ind	G	0.54	G	0.3	G	0.58
			C	0.46	C	0.7	C	0.42
NC_44611.1: 21501651	IndLoc44	Ind	A	0.37	A	0.75	A	0.58
			G	0.63	G	0.25	G	0.42
NC_44608.1: 54724583	IndLoc45	Ind	C	0.50	C	0.5	C	0.50
			G	0.50	G	0.5	G	0.50
NC_44614.1: 6454099	IndLoc46	Ind	C	0.61	C	0.2	C	0.50
			T	0.39	T	0.8	T	0.50
NC_44613.1: 60399321	IndLoc47	Ind	A	0.54	A	0.2	A	0.67
			G	0.46	G	0.8	G	0.33
NC_44608.1: 92052234	IndLoc48	Ind	C	0.50	C	0.5	C	0.50
			T	0.50	T	0.5	T	0.50
NC_44609.1: 131328013	IndLoc49	Ind	A	0.59	A	0.15	A	0.58
			G	0.41	G	0.85	G	0.42
NC_44619.1: 72708222	IndLoc50	Ind	A	0.57	A	0.2	A	0.58
			T	0.43	T	0.8	T	0.42
NC_44602.1: 1240358	IndLoc51	Ind	G	0.59	G	0.15	G	0.58
			A	0.41	A	0.85	A	0.42
NC_44609.1: 25177014	IndLoc52	Ind	A	0.33	A	0.55	A	0.83
			C	0.67	C	0.45	C	0.17
NC_44624.1: 2094293	IndLoc53	Ind	G	0.41	G	0.55	G	0.67
			A	0.59	A	0.45	A	0.33
NC_44613.1: 70326181	IndLoc54	Ind	A	0.57	A	0.55	A	0.29
			G	0.43	G	0.45	G	0.71
NC_44612.1: 6626313	IndLoc55	Ind	T	0.33	T	0.7	T	0.71
			C	0.67	C	0.3	C	0.29
NC_44623.1: 3268087	IndLoc56	Ind	A	0.50	A	0.45	A	0.54
			T	0.50	T	0.55	T	0.46
NC_44619.1: 75065665	IndLoc57	Ind	G	0.48	G	0.55	G	0.50

			A	0.52	A	0.45	A	0.50
NC_44617.1: 74600328	GggLoc1	*Ggg*	G	1.00	C	1	C	1.00
NC_44609.1: 27063360	GggLoc2	*Ggg*	C	1.00	T	1	T	1.00
NC_44608.1: 76391190	GggLoc3	*Ggg*	T	1.00	C	1	C	1.00
NC_44605.1: 62489626	GggLoc4	*Ggg*	T	1.00	A	1	A	1.00
NC_44611.1: 21246985	GggLoc5	*Ggg*	C	1.00	G	1	G	1.00
NC_44606.1: 65662558	GggLoc6	*Ggg*	A	1.00	G	1	G	1.00
NC_44609.1: 144698546	GggLoc7	*Ggg*	T	1.00	A	1	A	1.00
NC_44605.1: 69444435	GggLoc8	*Ggg*	C	1.00	A	1	A	1.00
NC_44605.1: 73508748	GggLoc9	*Ggg*	C	1.00	T	1	T	1.00
NC_44608.1: 11728215	GggLoc10	*Ggg*	A	1.00	G	1	G	1.00
NC_44606.1: 65689911	GggLoc11	*Ggg*	A	1.00	G	1	G	1.00
NC_44608.1: 129726948	GbgLoc1	*Gbg*	A	1.00	C	1	A	1.00
NC_44609.1: 14768484	GbgLoc2	*Gbg*	A	1.00	C	1	A	1.00
NC_44606.1: 21687583	GbgLoc3	*Gbg*	G	1.00	T	1	G	1.00
NC_44611.1: 51645682	GbgLoc4	*Gbg*	A	1.00	T	1	A	1.00
NC_44607.1: 88455911	GbgLoc5	*Gbg*	A	1.00	G	1	A	1.00
NC_44605.1: 1531022	GbgLoc6	*Gbg*	T	1.00	C	1	T	1.00
NC_44618.1: 38039430	GbgLoc7	*Gbg*	C	1.00	A	1	C	1.00
NC_44617.1: 14118600	GbgLoc8	*Gbg*	T	1.00	C	1	T	1.00
NC_44613.1: 22074032	GbgLoc9	*Gbg*	C	1.00	T	1	C	1.00
NC_44602.1: 153402393	GbgLoc10	*Gbg*	T	1.00	C	1	T	1.00
NC_44613.1: 106702934	GbbLoc1	*Gbb*	G	1.00	G	1	A	1.00
NC_44606.1: 26266581	GbbLoc2	*Gbb*	G	1.00	G	1	A	1.00
NC_44609.1: 15560133	GbbLoc3	*Gbb*	C	1.00	C	1	T	1.00
NC_44608.1: 129910681	GbbLoc4	*Gbb*	G	1.00	G	1	A	1.00
NC_44618.1: 13097759	GbbLoc5	*Gbb*	G	1.00	G	1	A	1.00
NC_44602.1: 52658998	GbbLoc6	*Gbb*	T	1.00	T	1	C	1.00
NC_44610.1: 135068928	GbbLoc7	*Gbb*	C	1.00	C	1	T	1.00

NC_44612.1: 53187893	GbbLoc8	*Gbb*	C	1.00	C	1	T	1.00

3.3 Discussion

This Chapter summarised the work done to find suitable autosomal *ii*- and *si*-SNPs for the identification of individuals and sub-species in gorillas. For this aim, published whole genome sequencing data from 11 male and 38 female individuals of the three sub-species western lowland (*G. g. gorilla* – 27 individuals), eastern lowland (*G. b. graueri* – 10 individuals), and mountain gorilla (*G. b. beringei* – 12 individuals), was filtered by Pille Hallast at the Wellcome Sanger Institute to select two sets of SNPs, one containing highly heterozygous variants across individuals (*i.e. ii*-SNPs) and another comprising SNPs that were fixed in one sub-species but absent in the others (*i.e. si*-SNPs). The set of genome sequences also included one individual of Cross River gorilla (*G. g. dielhi*), however, as this sample suffered from contamination issues (Javi Prado-Martinez, personal communication), it was excluded from any further consideration and analysis. VCF files containing the two sets of SNPs were filtered further to exclude sites that were not in HWE (and therefore possibly under selection or subject to sequencing error, which would have led to biases in the dataset) and of poor quality (see Materials and Methods Section 3.2.1). Visual inspection of genotype data based on filtered *ii*-SNPs was conducted to check for any obvious residual structure. Finally, two sites with the highest degree of variability were selected from each chromosome. For the *si*-SNPs, an average of ten loci per sub-species (11 for western lowland, 10 for eastern lowland, and 8 for mountain gorillas, depending on the number of chromosomes in the VCF files provided by Pille Hallast) were selected randomly from different chromosomes.

Non-complementary target-specific primer pairs were designed utilising the proprietary PrimerPlex v2.50, and the freely available Primer3 software to simultaneously amplify the 90 identified loci in short fragments (*i.e.* no longer than 230 bp) so as to increase the chances of successful amplification in non-invasive and degraded samples, containing either scarce or highly degraded DNA. The Kamilah (*G. g. gorilla*) assembly ("Kamilah_GGO_v0/gorGor6") was used as reference. In addition to the autosomal SNPs for individual and sub-species identification, a primer pair for the amplification of the amelogenin sex test locus (Sullivan *et al.*, 1993) was also designed using the aforementioned software.

The software failed to design primers for three out of the 90 identified loci, under the

imposed conditions described in Section 3.3.3. Given the surplus of identified markers and the difficulties that may be encountered when designing a large number of primer pairs for multiplex PCR (Henegariu *et al.*, 1997), it was decided to continue with only 86 markers, hence leaving one locus for sex determination, 29 si- and 56 ii-SNPs. In this regard, previous studies of forensic SNPs for human identification have obtained accurate results with an even lower number of markers (Wei *et al.*, 2012; Pakstis *et al.*, 2010; Børsting *et al.*, 2009), even though humans, together with eastern gorillas, show a general dearth of genetic diversity when compared to other great ape species (Pemberton *et al.*, 2012; Prado-Martinez *et al.*, 2013). This suggests that even only 56 (or fewer) ii-SNPs could have sufficient discriminative power to provide effective individual identification in gorillas.

Specifically for the si-SNPs, the utilised software was successful in designing primer pairs for all the identified candidate loci. The use of autosomal SNPs, which are biparentally inherited, for species identification holds the advantage of overcoming issues of sex-biased effects at the species level that characterise more standard analytical approaches. In fact, while the analysis of selected mtDNA fragments – highly abundant in animal cells, and therefore allowing identification to be based on trace quantities of material, including processed and degraded samples – has now become the basis of molecular barcoding for species identification based on sequence polymorphisms analysis (such as polymorphisms in the gene fragments involved in the respiratory chain - cytochrome b and cytochrome oxidase), autosomal markers are better suited for the detection of hybridisation events between closely related species (Kowalczyk *et al.*, 2021; Matsudaira *et al.*, 2022). The latter represents an important aspect for the study and management of endangered animal populations both in situ and *ex situ*.

An important caveat of this Chapter is that the population allele frequencies reported in Table 3.3 may not be representative of the real population. This is because the individuals that compose the sample set were either confiscated or belonged to zoos and local orphanages (Prado-Martinez *et al.*, 2013; Xue *et al.*, 2015). Therefore, SNPs that are either heterozygous or homozygous in these individuals may well not commonly be in different individuals. To this regard, Chapter 4 will focus on the analysis of the SNP panel on a previously sequenced individual of mountain gorilla, so as to compare results between nanopore sequencing and published whole genome sequencing data; while

Chapter 5 will expand the sample size to include individuals that are, allegedly, unrelated to other individuals in the sample set.

Chapter 4 Nanopore sequencing of individual and sub-species identification SNPs in Mountain gorillas (*G. b. beringei)*

4.1 Introduction

This Chapter will describe the experimental and analytical methodology used to assess the performance of the primer plex designed in the previous Chapter for the amplification and sequencing of ii- and si-SNP loci using portable ONT sequencing. The validation was conducted on DNA from a previously sequenced gorilla (Nyamunwa – *G. b. beringei*) for which WGS data was already available (Prado-Martinez *et al.*, 2013). Experiments were conducted following good laboratory practice to prevent contamination. Positive (Kamilah – *G. g. gorilla*, whose assembly was used for primer design and therefore more likely to amplify) and a non-template control were used to test for contamination. Protective clothing, including laboratory coats, gloves, face masks and hair coverings were worn where appropriate. Disposal of laboratory waste was conducted in compliance with University of Leicester guidelines.

4.2 Materials and methods

4.2.1 DNA sample information and quantification

The Nyamunwa (G. b. beringei) DNA sample was a kind donation of Chris Tyler-Smith and Yali Xue from Wellcome Sanger Institute (Cambridge, UK). WGS data for Nyamunwa can be found and downloaded from Prado-Martinez *et al.* (2013). The sample was provided dissolved in nuclease-free water in a 1.5 ml Eppendorf tube and stored at -20°C.

Quantification was conducted using a NanoDrop 2000 (Thermo Scientific NanoDrop Products) in accordance to manufacturer's instructions (Desjardins and Conklin, 2010), utilising 1 µl of DNA in solution. To test for DNA purity, the ratio of absorbance at 260 nm to 280 nm was used, whereby a ratio of 1.8 indicates good sample purity.

4.2.2 DNA oligos ordering, resuspension and storage

Lyophilized desalted standard DNA oligos were ordered from MERCK (www.sigmaaldrich.com). Oligos were resuspended in nuclease-free water to a final concentration of 100 µM and stored at -20°C in labelled 100-well freezer boxes. Aliquots from each oligo stock solution were mixed in a multiplex working solution with a final concentration of 2 µM per oligo as per Type-it® Microsatellite PCR Kit (QIAGEN) Handbook recommendations (QIAGEN, 2009).

4.2.3 DNA amplification, purification and library preparation

Target loci were amplified using the Type-it® Microsatellite PCR Kit (QIAGEN, www.qiagen.com) in 25 µl reactions according to manufacturer's protocol, with the only variation regarding the concentration of primers added to the PCR, which was reduced to 0.05 µM per primer to accommodate all primers and avoid inhibition (QIAGEN, 2009). The PCR contained 12.5 µl Type-it® MasterMix, 2.5 µl 10x Primer Mix (2 µM), 8 µl RNase-free water, and 2 µl DNA template (Nyamunwa, 50 ng/µl). The PCR thermal cycling protocol consisted of a 95°C activation step for 5 min, followed by 35 cycles of 95°C for 30 sec, 60°C for 90 sec and 72°C for 30 sec, and a final extension step at 68°C for 10 min. For PCR validation, standard benchtop laboratory equipment was used,

including a micro-centrifuge (Eppendorf® 5415 D) and Veriti™ 96-well thermal cycler (Applied Biosystems™). PCR fragments were then separated by electrophoresis in a 4% (w/v) agarose gel in 0.5x TBE buffer at 120 volts for 2.5-3 h, using HyperLadder™ 50bp (Bioline) as a molecular size marker. PCR products were then purified using the Monarch® PCR & DNA Cleanup Kit (5 µg, Catalogue #T1030L) and eluted in 12 µl Nuclease-free water. 1 µl of PCR product was then quantified using the Qubit fluorometer dsDNA HS Assay Kit 500 (Catalogue #Q32854, ThermoFisher Scientific).

Library preparation was carried out utilising the ONT Ligation Sequencing Kit (SQK-LSK109) following the Genomic DNA by Ligation protocol (Version: GDE_9063_v109_revX_14Aug2019) designed for sequencing via Flongle flow cell (R9.4.1). However, as the aforementioned protocol is designed for use with 500 ng (or 50 to 100 fmol) of genomic DNA, the amount of input material in fmol was computed using the NEBioCalculator (New England Biolabs® Inc., www.nebiocalculator.com) and setting an estimated molecule length of 189 bp (median length of fragments). Considering possible product loss due to repeated purification steps, as outlined in the manufacturer's protocol, 70 ng of PCR product were used as input material. The protocol for library preparation consists of two main steps, namely "DNA repair and end-prep" and "Adapter ligation and clean-up". After each step, 1 µl of eluted sample was quantified using a Qubit™ fluorometer in conjunction with the dsDNA HS Assay Kit 500. Purification was conducted using AppMag PCR Clean Up Beads (AMB001, Appleton) in 1.8x PCR volume on DynaMag™-2 Magnet (Catalogue #12321D, Invitrogen™) and eluted either in water or ONT Elution Buffer (EB). Beads were washed with 80% (v/v) ethanol and with ONT Short Fragment Buffer (SFB) for the final purification step as per manufacturer's recommendations. The purified library (40 ng) was then stored on wet ice until loading.

4.2.4 Flongle flow cell loading

Following the installation of MinKNOW™ v.5.3.1, a first hardware check was conducted both on the MinION™ hardware and on the Flongle flow cell – which had been stored at 4°C upon arrival – following the dedicated wizards located in the "Start" interface of MinKNOW™. Upon meeting QC satisfying conditions – *i.e.* more than 50 pores available for sequencing – the Flongle flow cell was primed by injecting a mix of 117 µl Flush

Buffer (FB) and 3 µl Flush Tether (FLT) through the sample port using a P200 pipette tip. The Sequencing Mix was then prepared by mixing 15 µl Sequencing Buffer II (SBII), 10 µl Loading Beads II (LBII), and 5 µl DNA library in a 1.5 ml Eppendorf DNA LoBind® tube, and injected with great care through the sample port to prevent introducing air bubbles that would damage the pores.

MinKNOW™ guides the user through the appropriate kit selection process; in this case SQK-LSK109 was selected for sequencing. Run options were set to 24 hours for the run length and -180 mV for the bias voltage. Basecalling was disabled on MinKNOW™ and sequencing output was stored in FAST5 extension container file in Hierarchical Data Format 5 (HDF5 - https://support.hdfgroup.org/products/hdf5_tools/), and was performed using the ONT basecalling algorithm implemented in the Guppy v.5.0.7 (www.nanoporetech.com) software that runs on Graphics Processing Units (GPUs) on Linux in High Accuracy mode.

4.2.5 Basecalling, alignment and variant calling pipeline

Due to the lack of a standard pipeline (but see Clair (Luo *et al.*, 2020) and ARTIC NETWORK (https://artic.network/ncov-2019), developed after this project) for the analysis of short fragments sequenced using nanopore technology, a tailored-made approach to the analysis of FAST5 files, including basecalling, mapping and variant calling, was developed. This research was conducted using the Linux terminal of the SPECTRE High Performance Computing Facility at the University of Leicester.

The first step of the analysis involved basecalling – the process of converting FAST5 files into FASTQ files that can then be manipulated using standard bioinformatics tools (Cock *et al.*, 2010). Basecalling was conducted following the standard Guppy v.5.0.7 approach, available for ONT customers, using the High Accuracy setting. Resultant FASTQ files were then merged into one, which was later used for the alignment. Contrary to what is reported in similar studies (*e.g.* Cornelis *et al.* (2017)) no base quality filter was applied at this step, but rather the aligned sequences (in .BAM format) were filtered on the basis of mapping quality. The alignment was performed using Minimap2 (Li, 2018), a dedicated software developed by ONT for mapping sequences contained in FASTQ files, to the latest version of the western gorilla Kamilah_GGO_v0/gorGor6 assembly (August

2019) deposited in the NCBI under the accession number GCF_008122165.1 (Coordinators, 2016), which had been previously indexed using SAMtools (Li *et al.*, 2009) and BWA (Li and Durbin, 2009). Only regions containing the SNPs of interest were retained from the aligned files and these were then indexed and sorted using SAMtools. The BAM file was then visually inspected in Integrative Genomics Viewer – IGV v1.9 (Thorvaldsdóttir *et al.*, 2013). Finally variant calling was performed following the SAMtools | BCFtools v1.15 (Li, 2011) approach described in (Cornelis *et al.*, 2017), using a mapping quality filter value of 20. Finally, using VCFtools (Danecek *et al.*, 2011), the resultant VCF file containing the target variants was filtered to retain only the positions corresponding to the *ii*- and *si*-SNPs. The scripts used for the pipeline are summarised in Appendix E.

The resultant VCF file was later explored and manipulated in R v.4.2.0 (Team, 2018) using the package *vcfR* (Knaus and Grünwald, 2017) to extract genotype information. A read depth approach was adopted to reassess the heterozygosity of SNP loci; in particular, those SNPs for which the balance ratio in allele depth (*i.e.* AD1 / (AD1 + AD2) and AD2 / (AD1 + AD2)) fell between 0.2 and 0.8 were called heterozygous, while all the others were categorised as homozygous either for the reference (REF) or the alternative (ALT) allele. The newly called genotypes were then exported as a matrix in a .txt file and compared with the known genotypes from Prado-Martinez *et al.* (2013).

4.3 Results

DNA from Nyamunwa and controls was subject to PCR to amplify 86 loci (see Chapter 3), as described in Material and Methods – Section 4.2.3, and the products assessed using agarose gel electrophoresis. This provided promising results (Fig. 4.1). In fact, even though it was practically impossible to assess whether all loci amplified due to the limited size difference (*i.e.* 3 bp apart only), segments appeared to be of the right size range (*i.e.* between 111 and 230 bp in length). Any off-target amplification in Figure 4.1 (*i.e.* bands < 111 bp and >230 bp) will be filtered out during sequencing quality check.

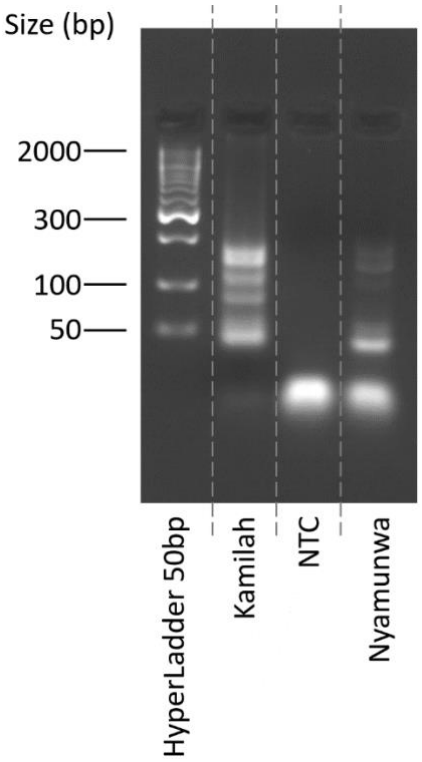


Figure 4.1: Gel electrophoresis of PCR products for the 86 identified loci. From the left, HyperLadder 50bp, Kamilah (*G. g. gorilla*), non-template control (NTC), Nyamunwa (*G. b. beringei*).

Based on these findings, the purification of PCR products and library preparation was commenced and conducted at room temperature according to manufacturer's instructions. The library was then sequenced using a Flongle flow cell.

MinKNOW™ produces valuable information that can be explored by the user to assess the quality of the sequencing run and make comparisons with other runs. The process of ONT sequencing technology is indeed prone to a degree of variability, depending

primarily on the number of active pores available for sequencing prior to Flongle flow cell priming and after loading the library. In addition, room temperature, library purity and concentration, and even storage conditions are all factors that are known to influence the quality of the sequencing run. Because of this, Flongles were QC'd immediately before use to assess (and minimize) pore loss during the storage period (McDonald *et al.*, 2021). The aforementioned information is summarised in a "Run report" that is generated by MinKNOW at the end of each run. For the experimental sequencing run of the *ii*- and *si*-SNPs of the genotyped mountain gorilla sample Nyamunwa, used for the validation of the panel of SNPs, MinKNOW registered a total of 86 pores before loading and 72 pores after loading. The sequencing run lasted for 17.5 hours and produced an estimated total of 1.2 million reads and 501.02 megabases of sequence.

Processed data – *i.e.* the matrix with extracted genotypes – were then compared with Nyamunwa's genotypes known from Prado-Martinez *et al.* (2013) to assess the efficacy of ONT sequencing and, in particular, of the Flongle flow cell – which reportedly has lower accuracy and sequencing output than standard ONT flow cell (Grädel *et al.*, 2019). Unexpectedly, sequencing of short amplicons yielded promising results. Indeed, while previous studies have highlighted issues related to sequencing short fragments using ONT technology (Wei and Williams, 2016; Wei *et al.*, 2018), all but four amplicons were sequenced to an average coverage of 2990 reads per amplicon ($min = 25; max = 32{,}342;\ SD = 5080$), for a total of 248,208 reads (Table 4.1). Comparison with published sequencing data from Prado-Martinez *et al.* (2013) demonstrated that the sequencing and analysis of short target *ii*- and *si*-SNP amplicons provided accurate information; in fact, no mismatches were found. The absence of four amplicons was attributed to issues with amplification rather than problems with the sequencing.

Having successfully tested the two SNP-plexes using nanopore sequencing, analysis of additional gorilla samples was undertaken including both good quality (*i.e.* blood) and poor quality (*i.e.* hair and faeces) DNA; this will be described in the next chapter.

Table 4.1: Summary of genotypes, allele read depth and overall coverage in Nyamunwa (*G. b. beringei*)

Locus ID	REF	ALT	Called genotypes	Genotypes	REF depth	ALT depth	Coverage

GggLoc1	G	C	1/1	C/C	5	5445	5450
GggLoc2	C	.	na	na	0	0	0
GggLoc3	T	C	1/1	C/C	13	2863	2876
GggLoc4	T	.	na	na	0	0	0
GggLoc5	C	G	1/1	G/G	1	637	638
GggLoc6	A	G	1/1	G/G	0	315	315
GggLoc7	T	A	1/1	A/A	299	3518	3817
GggLoc8	C	A	1/1	A/A	0	199	199
GggLoc9	C	T	1/1	T/T	1	278	279
GggLoc10	A	G	1/1	G/G	14	738	752
GggLoc11	A	G	1/1	G/G	0	25	25
GbgLoc1	A	.	0/0	A/A	235	0	235
GbgLoc2	A	.	0/0	A/A	865	0	865
GbgLoc3	G	.	0/0	G/G	4902	0	4902
GbgLoc4	A	.	0/0	A/A	20617	0	20617
GbgLoc5	A	.	0/0	A/A	12404	0	12404
GbgLoc6	T	.	0/0	T/T	1988	0	1988
GbgLoc7	C	.	0/0	C/C	262	0	262
GbgLoc8	T	.	0/0	T/T	1448	0	1448
GbgLoc9	C	.	0/0	C/C	8177	0	8177
GbgLoc10	T	.	0/0	T/T	135	0	135
GbbLoc1	G	A	1/1	A/A	0	70	70
GbbLoc2	G	A	1/1	A/A	7	6563	6570
GbbLoc3	C	T	1/1	T/T	8	1445	1453
GbbLoc4	G	A	1/1	A/A	24	4624	4648
GbbLoc5	G	A	1/1	A/A	0	457	457
GbbLoc6	T	C	1/1	C/C	17	1791	1808
GbbLoc7	C	T	1/1	T/T	24	4757	4781
GbbLoc8	C	T	1/1	T/T	1	456	457
IndLoc1	G	A	1/1	A/A	2	511	513
IndLoc2	T	.	0/0	T/T	3202	0	3202
IndLoc3	T	C	0/1	T/C	228	146	374
IndLoc4	T	G	0/1	T/G	24	11	35
IndLoc5	G	.	0/0	G/G	2709	0	2709
IndLoc6	G	C	0/1	G/C	16551	15791	32342
IndLoc7	T	C	1/1	C/C	19	1085	1104
IndLoc8	C	T	0/1	C/T	2674	2437	5111
IndLoc9	T	.	0/0	T/T	2619	0	2619
IndLoc10	A	G	1/1	G/G	0	50	50
IndLoc11	T	C	1/1	C/C	88	7527	7615
IndLoc12	A	G	1/1	G/G	26	2194	2220
IndLoc13	T	C	1/1	C/C	20	3712	3732
IndLoc14	C	A	0/1	C/A	311	380	691

IndLoc15	C	.	0/0	C/C	104	0	104
IndLoc16	T	C	0/1	T/C	158	122	280
IndLoc17	A	G	0/1	A/G	2589	2137	4726
IndLoc18	G	.	0/0	G/G	2708	0	2708
IndLoc19	C	G	0/1	C/G	2226	950	3176
IndLoc20	C	G	0/1	C/G	3376	4544	7920
IndLoc21	C	.	0/0	C/C	406	0	406
IndLoc22	T	.	0/0	T/T	1789	0	1789
IndLoc23	T	C	0/1	T/C	256	184	440
IndLoc24	A	G	1/1	G/G	75	18470	18545
IndLoc25	A	G	0/1	A/G	213	228	441
IndLoc26	C	G	0/1	C/G	7004	7138	14142
IndLoc27	A	.	0/0	A/A	3243	0	3243
IndLoc28	G	A	na	na	0	0	0
IndLoc29	G	A	1/1	A/A	16	2963	2979
IndLoc30	C	T	1/1	T/T	15	10142	10157
IndLoc31	C	.	0/0	C/C	5143	0	5143
IndLoc32	G	.	0/0	G/G	1267	0	1267
IndLoc33	G	T	1/1	T/T	23	1285	1308
IndLoc34	T	A	0/1	T/A	101	110	211
IndLoc35	T	.	0/0	T/T	947	0	947
IndLoc36	C	.	0/0	C/C	38	0	38
IndLoc37	C	.	0/0	C/C	325	0	325
IndLoc38	A	G	0/1	A/G	852	844	1696
IndLoc39	T	.	0/0	T/T	167	0	167
IndLoc40	C	T	1/1	T/T	16	3372	3388
IndLoc41	A	G	1/1	G/G	14	2413	2427
IndLoc42	A	G	0/1	A/G	611	526	1137
IndLoc43	C	G	1/1	G/G	1	218	219
IndLoc44	G	A	1/1	A/A	0	95	95
IndLoc45	G	C	0/1	G/C	1539	1329	2868
IndLoc46	C	.	0/0	C/C	395	0	395
IndLoc47	A	.	0/0	A/A	1746	0	1746
IndLoc48	C	T	0/1	C/T	407	196	603
IndLoc49	A	.	0/0	A/A	581	0	581
IndLoc50	A	.	0/0	A/A	159	0	159
IndLoc51	A	G	1/1	G/G	6	1439	1445
IndLoc52	A	C	na	na	0	0	0
IndLoc53	A	G	1/1	G/G	2	33	35
IndLoc54	G	A	0/1	G/A	165	139	304
IndLoc55	C	T	1/1	T/T	6	1081	1087
IndLoc56	T	.	0/0	T/T	405	0	405
AmelLocXY	T	.	0/0	T/T	1319	0	1319

4.4 Discussion

In this Chapter, the panel of 86 SNPs designed in Chapter 3, including 56 SNPs for individual identification, 29 for sub-species identification, and an additional sex-test locus, was tested on a good quality sample of previously sequenced mountain gorilla (Nyamunwa – *G. b. beringei*) using the portable ONT MinION™ in conjunction with the Ligation Sequencing Kit (SQK-SLK109). Drawing from the pipeline developed by Cornelis *et al.* (2017), a dedicated workflow was developed for the analysis of short reads.

Initially, and based on previous studies (Cornelis *et al.*, 2017), a more complex approach was adopted, whereby library preparation was preceded by amplicon ligation so as to overcome the (widely accepted) lower length limitations that were thought to characterise ONT sequencing. This was based on the catalytic activity of the T4 Ligase enzyme, which is commonly used for the ligation of blunt-ended DNA fragments (Lohman, 2018). However, as amplification via the Type-it kit is mediated by the HotStarTaq *Plus* polymerase, the addition of an A-base to the 3′ end of each fragment was thought to be a likely cause of inhibition of ligation. To circumvent this issue, two independent tests were run to assess the best approach to ligation.

Firstly, the PCR product of a single phosphorylated primer pair (*i.e.* GbbLoc5, length 214 bp) from the designed panel was purified using the Monarch® PCR & DNA Cleanup Kit (5 µg), incubated with 10 units of T4 Polynucleotide Kinase (T4 PNK; New England BioLabs®) enzyme for 3′-end A-tail removal and then ligated with 400 units of T4 Ligase (Catalogue #M0202, New England BioLabs®). The purification step prior to the incubation with T4 PNK was necessary to remove the salts in the Type-it Kit buffer which could inhibit the activity of the enzyme. All incubation steps were carried out in accordance with the manufacturer's instructions. As T4 Ligase is active in T4 PNK buffer, no additional purification step was needed.
Secondly, primers for the GbbLoc5 locus were re-designed to include the XbaI restriction enzyme's cut site (*i.e.* T/CTAGA) at the 5′-end of each primer with the addition of a short flanking sequence (6 bases) to facilitate the activity of the enzyme itself. PCR was conducted following the same Type-it protocol described above.

PCR products from the two different approaches were then compared via gel electrophoresis (Fig 4.2), and gave similar results.

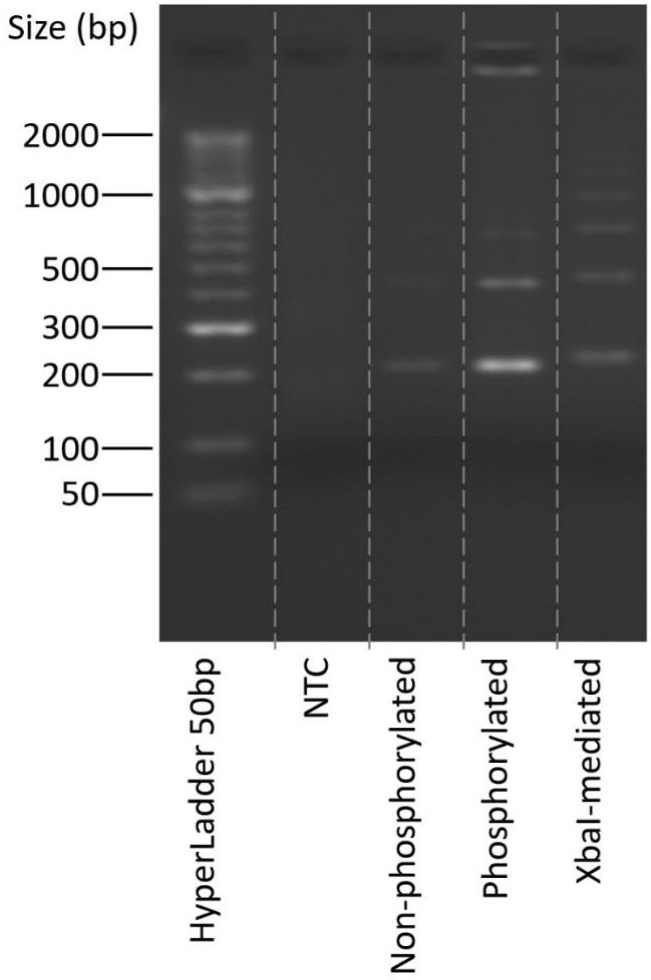


Figure 4.2: Agarose gel comparing the results of the different ligation approaches for the GbbLoc5 locus. From the left, HyperLadder 50bp, negative control (NTC), non-phosphorylated product, phosphorylated product, restriction enzyme (XbaI) product

While robust and effective, both ligation protocols proved to be time consuming, costly, and more prone to errors due to the numerous steps required; the combination of these factors hampers the portability and practicality of MinION™ sequencing, originally meant for in situ identification of sub-species and individuals. For this reason, it was finally decided to attempt sequencing short amplicons directly – *i.e.* with no prior ligation. As this approach – described in this Chapter – gave promising results, it was adopted for all subsequent experiments.

A brief investigation of the causes of the sequencing failure of four loci, conducted using IGV v1.9 on the unfiltered BAM file, revealed that although these amplified they failed

to meet the selected minimum quality criteria (Appendix E). Again, for the reasons outlined in Chapter 3 regarding the surplus of loci when compared with similar studies in humans (Wei *et al.*, 2012; Pakstis *et al.*, 2010; Børsting *et al.*, 2009), it was decided to proceed with the application of the developed SNP panel – regardless of the four failed loci – on a larger sample so that include all three available sub-species. This will be described in Chapter 5.

Chapter 5 Nanopore sequencing of *ii*- and *si*-SNPs in gorilla samples (good quality *vs* non-invasive samples)

5.1 Introduction

The previous Chapter assessed the efficacy of nanopore sequencing in analysing short amplicons using a good quality DNA sample (*i.e.* Nyamunwa – *G. b. beringei*; see Chapter 4). In this Chapter, those methods are applied to analyse a larger number of samples per species and from different sources (*i.e.* blood, hair, and faeces) to assess the feasibility of the system to study gorillas in the field.

Whilst ecological and conservation research have benefited significantly from technological advances in remote sensing (Rhodes *et al.*, 2022), camera trap surveys (Steenweg *et al.*, 2017), GPS (Schieltz *et al.*, 2017) and acoustics (Pijanowski *et al.*, 2011) in recent decades, to date the adoption of genomic analysis in wildlife management and conservation plans remains marginal (von Thaden *et al.*, 2020; Taylor *et al.*, 2017). In particular, the high capital investments and complex bioinformatics skills required for the analysis have delayed the adoption of genomics on a broad scale (Pires and Olah, 2022). Nonetheless, despite these issues, genomic analysis holds some clear advantages over the aforementioned approaches to the study of wildlife. In fact, while camera trapping is an excellent method to study both the occurrence and behaviour of animal species (Dheer *et al.*, 2022; Mori *et al.*, 2022) and GPS tracking is now commonly used to monitor individual movements over large distances (Mandl *et al.*, 2022; Duporge *et al.*, 2022), both methods fail to provide essential information on population structure and dynamics. For that, direct in field observations remain the only valuable alternative (Tella *et al.*, 2013). Recent improvements in molecular and sequencing technologies now enable researchers to utilise non-invasive samples such as hair and faeces and even just soil or water samples, to address crucial questions for the conservation of threatened and elusive species (Schwartz *et al.*, 2007). With the advent of massively parallel sequencing and, in particular, of portable sequencing technology, it is now possible to analyse a virtually unlimited number of genetic markers almost anywhere in the world, including biallelic

single nucleotide polymorphisms (SNPs), the study of which is more amenable to automation and hence more accessible to people with no previous experience and background (von Thaden *et al.*, 2020; Menegon *et al.*, 2017). In recent years, the field of forensic genomics has indeed been acquiring an ever more prominent role in the fight against illegal wildlife trafficking and poaching, thereby contributing considerably to the implementation of wildlife export and trade protocols (*e.g.* CITES) (Ogden *et al.*, 2009; Williams *et al.*, 2021; Pires and Olah, 2022).

In this context, nanopore sequencing has emerged as a valuable approach to DNA testing for species monitoring programmes in remote areas (Pomerantz *et al.*, 2017; Pomerantz *et al.*, 2018; Pomerantz *et al.*, 2022), and the portable ONT MinION™ device has now been used to conduct genetic analysis in remote areas around the globe (Edwards *et al.*, 2016; Quick *et al.*, 2016; Carroll *et al.*, 2018; Chang *et al.*, 2020). The portability, relatively low acquisition costs, and simplified library preparation protocols for MinION™ sequencing, allow researchers to operate in areas that lack well-established research facilities, such as biodiversity rich regions and developing countries. This, in turn, facilitates the study of elusive species and brings support to the fight against wildlife crime by providing an *in situ* analytical system for the identification of species, thereby overcoming the issues related to the export of biological samples, which has traditionally been considerably expensive and time consuming (Pomerantz *et al.*, 2022).

For the reasons outlined above, the final part of this PhD project aimed at assessing the performance of SNP sequencing through nanopore portable technologies on non-invasive samples containing highly degraded DNA. The aim was to test the practicality and efficacy of portable nanopore sequencing and thus to assess how it can support field studies of elusive gorillas from non-invasively collective samples. This Chapter describes the full analytical process from DNA extraction to data analysis using a group of gorilla samples belonging to the three subspecies; western lowland, eastern lowland and mountain. Particular attention was paid to the development of a relatively straightforward, inexpensive and streamlined approach to DNA testing of both good quality and poor quality samples, in order to offer a tool that can support *in situ* studies of gorilla populations while aiding the fight against illegal wildlife trafficking and poaching. Towards this aim, it is demonstrated here that the designed SNP-plex was indeed effective for both sub-species and individual identification and could be used to conduct tests for

parentage exclusion and reconstruction of family pedigrees.

5.2 Materials and methods

5.2.1 Collation of samples

Samples came from a variety of sources (Table 5.1). These included DNA extracts donated by Chris Tyler-Smith and Yali Xue from samples that had been previously analysed, matched hair and blood on FTA cards from individuals hosted at Aspinall Foundation, and matched blood, hair and faecal samples of individuals from Twycross Zoo. DNA extracts from other places (*e.g.* Belfast Zoo, Paignton Zoo Park, ECACC, etc.) were part of the collection of Prof Mark Jobling's Lab (UoL). Reported parentage information of individuals hosted at the respective zoos was also provided together with the aforementioned samples. Biological samples from zoos (including Twycross Zoo and the Aspinall Foundation) were collected by qualified members of staff either during veterinary procedures, within safe and acceptable limits for the animals concerned, or during routine zoo keeping activities (*e.g.* enclosure maintenance). The University of Leicester is an approved CITES-registered institution and the use of great ape samples (and this research in general) was approved by the University of Leicester Animal Welfare and Ethical Review Body (ref.: AWERB/2021/159).

Table 5.1: Sample information. This includes name and species of belonging, DNA source, sex, origin, flow cell type used for sequencing and known relationships.

Sample name	Common name	Scientific name	Source	Sex	Ng/uL	Origin	Flow cell type	Relatives
Nyamunwa	Mountain gorilla	*Gorilla beringei beringei*	DNA	F	60	Bwindi (Uganda)	Flongle	
Bwiruka	Mountain gorilla	*Gorilla beringei beringei*	DNA	F	112	Bwindi (Uganda)	Flongle	
Kahungye	Mountain gorilla	*Gorilla beringei beringei*	DNA	F	110	Bwindi (Uganda)	Flongle	
Katungi	Mountain gorilla	*Gorilla beringei beringei*	DNA	F	20.9	Bwindi (Uganda)	Flongle	
Zirikana	Mountain gorilla	*Gorilla beringei beringei*	DNA	M	1145	Rwanda	Flongle	
Imfura	Mountain gorilla	*Gorilla beringei beringei*	DNA	M	1010	Rwanda	Flongle	
Tumani	Eastern lowland gorilla	*Gorilla beringei graueri*	DNA	F	110	Confiscated	Flongle	
Pinga	Eastern lowland gorilla	*Gorilla beringei graueri*	DNA	F	120	Confiscated	Flongle	
Itebero	Eastern lowland gorilla	*Gorilla beringei graueri*	DNA	F	110	Confiscated	Flongle	
Ntabwoba	Eastern lowland gorilla	*Gorilla beringei graueri*	DNA	M	110	Confiscated	Flongle	
Dunia	Eastern lowland gorilla	*Gorilla beringei graueri*	DNA	F	115	Confiscated	Flongle	
Serufuli	Eastern lowland gorilla	*Gorilla beringei graueri*	DNA	F	95	Confiscated	Flongle	
Oumbi_B	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	M	57	Twycross Zoo	Flongle	PO Shufai
Oumbi_H	Western lowland gorilla	*Gorilla gorilla gorilla*	Hair	M	-	Twycross Zoo	Flongle	PO Shufai
Joshi_B	Western lowland gorilla	*Gorilla gorilla gorilla*	FTA	M	13	Aspinall Foundation	MinION	
Joshi_H	Western lowland gorilla	*Gorilla gorilla gorilla*	Hair	M	-	Aspinall Foundation	MinION	
Koujou	Western lowland gorilla	*Gorilla gorilla gorilla*	DNA	M	105	Howletts Zoo Park	Flongle	
Matadi	Western lowland gorilla	*Gorilla gorilla gorilla*	DNA	M	110	Paignton Zoo	Flongle	
Mjukuu	Western lowland gorilla	*Gorilla gorilla gorilla*	DNA	F	85	-	Flongle	
Kibi	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	7.6	Aspinall Foundation	Flongle	
FouFou	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	8.4	Aspinall Foundation	Flongle	
Mbwambe	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	7.9	Aspinall Foundation	Flongle	
Akou	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	27.7	Aspinall Foundation	Flongle	
Louna	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	M	7.6	Aspinall Foundation	Flongle	
Kishi	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	6.1	Aspinall Foundation	MinION	
Boula	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	3.6	Aspinall Foundation	Flongle	

Fubu	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	M	6.8	Aspinall Foundation	Flongle	
Kebu	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	M	4.5	Aspinall Foundation	Flongle	
Masindi	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	10	Aspinall Foundation	Flongle	
Emmie	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	11	Aspinall Foundation	MinION	
MahMah	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	9	Aspinall Foundation	MinION	
Tamba	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	7.5	Aspinall Foundation	MinION	
EB(JC)	Western lowland gorilla	*Gorilla gorilla gorilla*	DNA	F	21	ECACC	Flongle	
Effie	Western lowland gorilla	*Gorilla gorilla gorilla*	DNA	F	10	London Zoo RP	Flongle	
Kesho	Western lowland gorilla	*Gorilla gorilla gorilla*	DNA	M	57	ZSL	Flongle	
Floquet	Western lowland gorilla	*Gorilla gorilla gorilla*	DNA	M	25	-	Flongle	
Kanghu	Western lowland gorilla	*Gorilla gorilla gorilla*	DNA	M	25	Howletts Zoo Park	Flongle	
GoM6	Western lowland gorilla	*Gorilla gorilla gorilla*	DNA	M	23	-	Flongle	
GoM4	Western lowland gorilla	*Gorilla gorilla gorilla*	DNA	M	60	-	Flongle	
Ggo5	Western lowland gorilla	*Gorilla gorilla gorilla*	DNA	M	20	-	Flongle	
Biddy	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	8	Twycross Zoo	MinION	PO Ozala
Ozala	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	F	7	Twycross Zoo	MinION	PO Biddy PO Shufai
Oumbi	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	M	64	Twycross Zoo	MinION	PO Shufai
Shufai_B	Western lowland gorilla	*Gorilla gorilla gorilla*	Blood	M	12	Twycross Zoo	MinION	PO Ozala PO Oumbi
Shufai_F	Western lowland gorilla	*Gorilla gorilla gorilla*	Faeces	M	50	Twycross Zoo	MinION	PO Ozala PO Oumbi
Shufai_H	Western lowland gorilla	*Gorilla gorilla gorilla*	Hair	M	-	Twycross Zoo	MinION	PO Ozala PO Oumbi

5.2.2 DNA extraction and quantification

Depending on the biological source (blood, hair or faeces) different protocols were followed for the extraction of DNA for downstream analysis.

5.2.2.1 DNA extraction from blood samples on FTA cards

Extraction of DNA from blood samples stored on FTA (Flinders Technology Associates) cards followed a modified version of the Chelex® 100 protocol described in (Hailemariam *et al.*, 2017). For each FTA sample, a 4 x 8 mm rectangle was cut and added to a 1.5 mL Eppendorf tube. The sample was then washed twice in 500 µl FTA purification reagents for 15 minutes. It was then washed twice in 500 µl Nuclease-Free water (Sigma-Aldrich, catalogue #W4502) for 15 minutes, and allowed to air dry for ~1 h. The sample was then suspended in 100 µl 5% (w/v) Chelex® 100 solution and incubated at 90°C for 30 minutes on a heating block. The sample was then spun down at maximum speed (13,200 rpm) for 3 minutes in a micro-centrifuge (Eppendorf® 5415 D). The supernatant was then transferred into a new 1.5 mL Eppendorf tube and the pellet discarded.

5.2.2.2 DNA extraction from whole blood samples

DNA extraction from whole blood samples utilised the QIAamp® DNA Investigator kit (QIAamp, 2010a) in accordance with the manufacturer's indications. An aliquot of 45 µl whole blood was incubated at 56°C for 10 minutes in a reaction containing 55 µl Buffer ATL, 10 µl proteinase K, and 100 µl Buffer AL. After incubation, the sample was transferred into a mini column and washed using ethanol (96 – 100% v/v) and proprietary buffers (AW1 and AW2), spun down at maximum speed for 1 minute in a micro-centrifuge (Eppendorf® 5415 D), and finally eluted in 50 µl Nuclease-Free water (Sigma-Aldrich, catalogue #W4502). Pipette tips and Eppendorf tubes used for the extraction of DNA from whole blood (and FTA card) gorilla samples were disinfected in virucidal solution overnight (Rely^{+}On™ Virkon® tablets) prior to final disposal.

5.2.2.3 DNA extraction from faecal samples

For DNA extraction from faecal samples, the QIAamp DNA Stool Mini Kit (QIAamp, 2010b) was utilised in accordance with the manufacturer's instructions. Circa 220 mg of

frozen faeces were dissected from the sample surface using a scalpel and added to a 1.5 mL Eppendorf™ Polypropylene DNA LoBind microcentrifuge tube. DNA was eluted in Nuclease-Free water and stored at -20°C for downstream analysis. The tools used for dissecting and handling faecal samples (*e.g.* scalpel, cutting mat, and forceps) were washed in a 80% (v/v) bleach solution for 5-10 minutes, rinsed with water, and finally sterilised with ethanol over flames twice after each use.

5.2.2.4 Direct PCR from hair samples

Five hairs per individual were cut to retain pieces of 1 cm in length, including the follicle. This is because gorilla hair is rich in melanin, which is known to interfere with PCR by forming a reversible complex with the DNA polymerase (Eckhart *et al.*, 2000). Hair samples were then washed twice in 500 µl 70% (v/v) ethanol solution and twice in 700 µl Nuclease-Free water, and allowed to dry for 2-3 hours at room temperature. The 50 µl direct PCR consisted of 25 µl Phire Tissue Direct PCR Master Mix (Thermo Scientific™, catalogue #F170S), 9 µl Nuclease-free water and 16 µl Primer mix (2 µM). Washed hair was added directly to the PCR in a 200 µl PCR tube. Thermal cycling was conducted at 98°C for 5 minutes, followed by 35 cycles of denaturation at 98°C for 5 seconds, annealing at 60°C for 5 seconds, and extension at 72°C for 20 seconds, and a final extension at 72°C for 1 minute.

5.2.2.5 Quantification on NanoDrop

DNA extracts were quantified on a NanoDrop 2000 (Thermo Scientific NanoDrop Products) in accordance with manufacturer's instructions (Desjardins and Conklin, 2010), utilising 1 µl of DNA in solution and blanking with 1 µl Nuclease-Free water (Sigma-Aldrich, catalogue #W4502). As different kits were used for the extraction, DNA quality was assessed by running PCR products on agarose gels to check whether SNPs amplification had been successful (see Figure 4.1 – Chapter 4) and no cutoffs were set for DNA quality prior to amplification. Extracted DNA concentrations (excluding DNA extracted from hair for which a direct PCR approach was followed, see session 5.2.2.4) are reported in Table 5.1.

5.2.3 DNA amplification, purification and quantification

DNA amplification from blood and faecal samples (hair samples were amplified following the protocol described above), and PCR purification (with the Monarch PCR and DNA CleanUp Kit) were performed following the protocols described in Chapter 4. Purified PCR products were then quantified using the Qubit HS Assay and 1 µl per sample.

5.2.3.1 Library Prep and Sequencing

Sequencing was conducted using the ONT ligation sequencing kit (SQK-LSK109) in conjunction with the native barcoding expansion 1-12 barcodes (EXP-NBD104) for sample multiplexing. Differences in sequencing accuracy and throughput between Flongle and MinION flow cells were assessed by running side-by-side comparisons (Table 5.1). In fact, there is a widespread consensus among Nanopore users that MinION flow cells provide higher accuracy (nanopore community and ONT events). For this, a maximum of eight samples (*i.e.* eight ONT barcodes) per run were sequenced using both flow cell types, following virtually identical library preparation protocols. The major difference between the Flongle and the MinION protocols is the input concentration of purified PCR products (70 ng/µl for the Flongle flow cell and 180 ng/µl for the MinION flow cell per sample) and the reaction volume, which doubles when preparing a library for sequencing with the MinION flow cell. Library loading also differs and requires the user to prime the MinION flow cell with proprietary priming buffer before adding the sample for sequencing, whereas no priming is required for loading a Flongle flow cell. In terms of hardware, MinION flow cells can sequence for a longer period of time due to the greater number of selectable pores which, in turn, increases the number of reads. In both cases, the purification steps described in the protocol were modified slightly for use with AppMag PCR Clean Up Beads (catalogue #AMB001, appleton®); for that, the manufacturer's recommended protocol (Appleton, 2020) was followed. Purified samples were quantified after each library preparation step using the Qubit HS assay as per ONT recommendations. The purified library was then stored either on wet ice for immediate loading, or at -20°C for delayed sequencing. Only half the library was loaded for each sequencing run, while the remainder was stored at -20°C in DNA LoBind Eppendorf tubes as a backup in case the flow cell failed during loading.

5.2.4 Data Analysis

Base calling, variant calling, and genotype filtering were performed following the procedures described in Chapter 4. Once again the latest version of Kamilah_GGO_v0/gorGor6 assembly (August 2019) deposited with NCBI under the accession number GCF_008122165.1 (Coordinators, 2016) was used for the alignment and mapping of SNPs, with the exception of the amelogenin marker for sex determination (Sullivan *et al.*, 1993) which was analysed using the Susie3_GSMRT3/gorGor5 assembly (March 2016) deposited in the NCBI database under the accession number GCA_900006654.1 (Gordon *et al.*, 2016), which includes both X and Y sequences. Consistent with Sullivan *et al.* (1993), individuals were categorised as females when only X-chromosomal products were observed, or as males when both X- and Y-chromosomal products were observed. Clustering analysis was conducted to assess the ability of *si*-SNPs to assign individuals to the appropriate species, utilising the DAPC function contained in the packages "*adegenet*" (version2.1-3) (Jombart, 2008) implemented in the statistical program R version 3.6.3 (Team, 2013) using coded identifiers for unique allele genotypes (Appendix F). This approach produces uncorrelated variables that serve as input for discriminant analysis (DA), via principal component analysis (PCA) transformation of multi-locus genotypes (Viengkone *et al.*, 2016). The *find.clusters* function was used to detect the number of clusters in the dataset, running successive K-means clustering with increasing numbers of clusters (*K*) for a maximum of 10 clusters (de Oliveira *et al.*, 2022). The Bayesian information criterion (BIC) was used to identify sharp changes in fit models for the successive runs of increasing *K*-means (Vallejo-Marin and Lye, 2013; Buchalski *et al.*, 2016; Viengkone *et al.*, 2016). In order to avoid overfitting (when too many PCs are retained) and losing important information (when too few PCs are retained), the number of PCs to retain was cross-validated using the function *xvalDapc* with 100 repetitions (Viengkone *et al.*, 2016). The advantage of using DAPC resides in the maximisation of between-group (or among-group) variation and the minimisation of within-group variation (Jombart *et al.*, 2010), making this method ideal for identifying species via SNP analysis. Conveniently, the *glPlot* function included in the *adegenet* package produces a relatively straightforward visual representation of individual genotypes, using different colours for homozygous and heterozygous loci (either for the REF or the ALT). This function was exploited both for the graphical representation of both *ii*- and *si*-SNPs. In addition, the *si*-SNPs efficiency in discriminating species was assessed using the software STRUCTURE v.2.3.4 (Pritchard *et al.*, 2000). As different species were present in the dataset, the admixture option was

disabled for the STRUCTURE analysis, which consisted of 150,000 Markov Chain Monte Carlo (MCMC) repetitions and 50,000 burn-ins for $K = 1 - 10$. The Evanno method (Evanno *et al.*, 2005) implemented in the web-based program STRUCTURE HARVESTER (available at http://taylor0.biology.ucla.ed/structureHarvester) was used to assess and visualise likelihood values across multiple K values generated in STRUCTURE (Earl and VonHoldt, 2012).

Relationship between individuals were estimated using the software ML-Relate (Montana State University, 2006) (Kalinowski *et al.*, 2006), after separating individuals into sub-species and removing repeated samples from the same individual but different biological sources (*e.g.* hair and faecal samples). When available, genotypes from previously sequenced individuals were added to the dataset to increase the sample size and allow the most realistic estimate of population allele frequencies, which is then used by ML-Relate to estimate relatedness between individuals. Relationship sets were constructed using a level of confidence of 95% with 1000 randomisations. However, because the presence of siblings can lead to errors in parentage identification, especially when levels of polymorphism are low (Ling *et al.*, 2020), parentage was then inferred using a maximum likelihood-based approach implemented in CERVUS version 3.0.7 (Field Genetics Ltd, www.fieldgenetics.com) (Kalinowski *et al.*, 2007) using both *ii*- and *si*-SNPs. For each candidate parent, CERVUS tests two alternative hypotheses: a) the candidate parent is indeed the true parent, and b) the candidate parent is not the true parent; and calculates the probability of obtaining the observed genotypes under these two hypotheses. It then provides the likelihood ratio – the likelihood that the candidate parent is the true parent divided by the likelihood that the candidate parent is not the true parent – expressed as the Likelihood of Odds (or LOD) score, whereby a positive value indicates that candidate parents might be the actual parents of the progeny (Marshall *et al.*, 1998; Wikberg *et al.*, 2017). For the parentage analysis in CERVUS, all possible combinations between related individuals (according to the ML-Relate output) were investigated, where all male and female samples were considered as candidate parents. Firstly, based on simulations of parentage analysis, CERVUS exploits allele frequencies to generate pairs of parental genotypes as well as a series of random genotypes representing unrelated candidate parents. It then produces an offspring via Mendelian sampling of true parental alleles, and uses this information to calculate the likelihood of parentage among the true parent (measured by its LOD score) and unrelated candidate parents for a large number (here set at 10,000) of simulated offspring (Kalinowski *et al.*, 2007). Relaxed confidence levels

were set at $LOD = 80\%$ whereas strict levels were set at $LOD = 95\%$ (Wikberg *et al.*, 2017). This approach was also used to reconstruct the family pedigree for a group of western gorillas hosted at Twycross Zoo (Table 5.1). The proportion of genotyped parents was set to 0.90 to account for the possibility of extra-group paternity (Wikberg *et al.*, 2017). Finally, CERVUS was used to perform identity analysis to assess whether genotypes of individuals from previous studies (Prado-Martinez *et al.*, 2013; Xue *et al.*, 2015) matched the results of MinION sequencing for the same individuals. When exact genotype matches are found (*i.e.* complete genotype sharing between two individuals), CERVUS calculates the probability that two individuals share the same genotype, excluding typing errors, in two forms. The basic formula assumes that the two individuals are unrelated (hereafter, p_{ID}), whereas a more conservative formula assumes that the two individuals are full siblings (hereafter, p_{IDsib}) (Waits *et al.*, 2001; Kalinowski *et al.*, 2007). A maximum of two mismatches per individual – equal to ~5% of the available genotyped markers – were allowed.

Furthermore, the same analyses were repeated using microhaplotype data obtained by retaining 25-bp sequences on both sides of each target SNP using the function *extract.haps* included in the R package *vcfR* (Knaus and Grünwald, 2017), and surveying this additional sequence for variants. While extracting longer haplotypes from sequencing data was indeed possible, the characteristic drop in quality of raw sequences at both ends could have led to a relatively high sequencing noise with the risk of detecting false variants (Delahaye and Nicolas, 2021). The aim was to assess whether microhaplotype analysis could increase the per-locus discriminative power for both individual and species identification as well as for parentage analysis, following the discovery of multiple alleles as phase-known haplotypes (Kidd *et al.*, 2013). Previous studies have indeed shown the utility of microhaplotypes in providing accurate individual and species identification as well as improving relationship testing (Kidd *et al.*, 2013; Oldoni *et al.*, 2019). For simplicity, non-invasive samples were excluded from this part of the analysis.

Related individuals were removed from population genetic analysis performed in STRAF version 1.0.5 online application (Univertät Bern, 2018) (Gouy and Zieger, 2017) to measure observed heterozygosity (H_{obs}) and derive expected heterozygosity (H_{exp}), Random Match Probability (PM), Power of Discrimination (PD), and Power of Exclusion (PE). Owing to the high rate of inbreeding among eastern gorilla species (Xue *et al.*, 2015; McGrath *et al.*, 2022), which would likely result in a higher rate of relatedness among individuals, half siblings (HS) detected through the ML-Relate analysis belonging to *G.*

b. beringei and *G. b. graueri* were retained for the analysis of population parameters while all related individuals (including half siblings) of *G. g. gorilla* were excluded.

Finally in R, the function *test_HW* implemented in the package *Genepop* was used to derive estimates of inbreeding coefficient (F_{IS}) as the proportion of the variance in a subpopulation contained in an individual (Weir and Cockerham, 1984). Specifically for this set of individuals and excluding *si*-SNP loci, which are *de facto* homozygous (and would bias the analysis), the expectation was that no evidence of inbreeding should be found among *ii*-SNPs given the high level of heterozygosity at these loci, which were selected for their ability to readily discriminate individuals. This is because samples used in this analysis were collected from individuals hosted in a variety of venues, including different zoos and orphanages, which may or may not reflect natural population structures and dynamics. Indeed, while zoo populations would in general be small and inbred (Che-Castaldo *et al.*, 2021), the "artificial" populations considered in this study – *i.e.* populations created *ad hoc* with individuals from different zoos – may appear to be less inbred than natural populations, where females would generally disperse and join bordering groups (Robbins and Robbins, 2015; Manguette *et al.*, 2020). In support of this, previous studies have indeed argued that considerations of sampling design are key to avoid biases in population genetic studies (Tajima, 1995; Gorospe *et al.*, 2015).

5.3 Results

Of the 84 *ii*- and *si*-SNP markers 42 survived filtering across all sequencing runs and samples (Table 6.2); these included 27 *ii*-SNPs, 14 *si*-SNPs, and one locus for sex determination. Microhaplotypes (51-bp long) were also derived for each of the 42 sequenced loci using the *bed* option contained in VCFtools (Danecek *et al.*, 2011). When only "good" quality samples (*i.e.* blood and DNA extracts) were considered a total of 49 SNPs survived filtering. The seven failing loci included both *si*- and *ii*-SNPs and had amplicons that varied in length between 160 bp to 226 bp. The analysis of variance (*i.e.* Kruskal-Wallis test) gave no evidence of length difference between the two sets of 42 and the 49 loci ($\chi^2 = 0.049,\ df = 1,\ P\ value = 0.824$). This suggests that the amplicon length is not a likely cause of failure in "poor" quality samples. This corroborates the idea that the designed panel is indeed suitable for the analysis of degraded DNA samples. For simplicity, and in for comparison purposes, the analysis was reduced down to the core set of 42 loci that worked across all samples and runs.

The number of reads varied considerably across Flongle and MinION sequencing runs, and to a lesser extent between Flongle runs: loci were sequenced to an average depth of 2,334 reads (Table 5.2). Figure 5.1 shows cross-species relative number of reads to the total number of reads per locus in relation to fragment size (bp). Kruskal-Wallis tests for the analysis of variance provided evidence for significant cross-species differences in the relative number of reads in relation to amplicon length (bp) ($\chi^2 = 50.724,\ df = 2,\ P\ value = 9.62^{-12}$). In particular, negative correlations between length and reads were found in *G. b. beringei* ($r = -0.441,\ P\ value = 0.003$) and in *G. g. gorilla* ($r = -0.374,\ P\ value = 0.016$). However, no significant difference was found between good quality (*i.e.* blood and DNA extracts) and poor quality (*i.e.* hair and faeces) in *G. g. gorilla* ($\chi^2 = 0.614,\ df = 1,\ P\ value = 0.433$), for which both types of sources were available.

Table 5.2: **Total number of reads per locus across gorilla sub-species.**

locus	Length (bp)	Western lowland	Eastern lowland	Mountain
IndLoc1	111	9060	1780	425
IndLoc2	122	4881	1916	625
IndLoc6	149	14314	3579	4331

IndLoc8	154	7665	1163	636
IndLoc9	157	5059	2729	566
IndLoc11	161	7269	1130	802
IndLoc13	165	700	805	622
IndLoc17	170	4544	2758	830
IndLoc19	175	1487	1314	495
IndLoc24	182	9307	2822	2733
IndLoc25	182	920	636	92
IndLoc26	187	2670	2604	1833
IndLoc27	188	678	406	494
IndLoc29	191	3708	2574	502
IndLoc30	191	8446	2484	1422
IndLoc31	194	2927	2179	853
IndLoc32	197	4327	2871	275
IndLoc35	201	1005	816	192
IndLoc37	203	2702	968	170
IndLoc38	204	794	552	349
IndLoc40	207	901	1218	582
IndLoc42	209	4227	3018	317
IndLoc45	211	1799	602	441
IndLoc48	217	3271	2673	184
IndLoc49	218	3043	1394	197
IndLoc51	220	943	1257	298
IndLoc55	228	3521	3122	399
GbbLoc2	172	4140	1376	1190
GbbLoc3	202	1809	838	275
GbbLoc6	220	3443	1823	310
GbbLoc7	223	2786	1447	957
GbgLoc3	148	4938	1112	839
GbgLoc4	161	7098	4696	3307
GbgLoc5	165	5720	2264	1949
GbgLoc9	217	7579	5718	1672
GggLoc1	144	12608	3387	984
GggLoc3	188	3384	944	526
GggLoc5	197	904	539	103
GggLoc7	199	9148	3118	875
GggLoc8	206	267	254	58
GggLoc9	207	1779	573	133
Total average number of reads per locus				**2334**

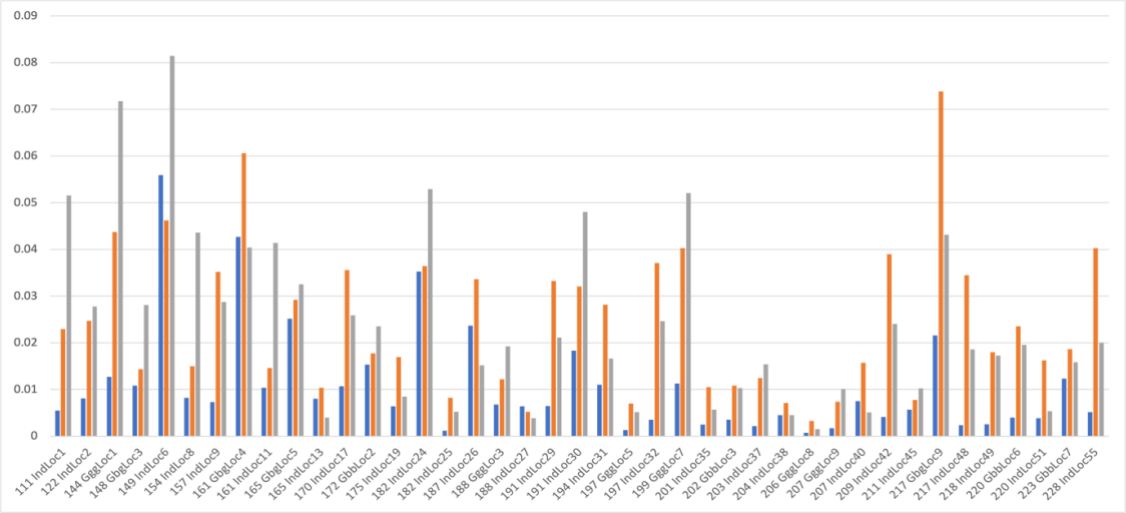


Figure 5.1: Cross-species relative number of reads per locus (y axis) in relation to fragment size (bp) for mountain gorilla (blue), eastern lowland (orange), and western lowland (grey).

Additionally, comparing ONT sequencing data – for the loci that survived filtering – with available SNP genotypes (Prado-Martinez *et al.*, 2013; Xue *et al.*, 2015) provided complete matching regardless of the flow cell type (*i.e.* either Flongle or MinION flow cells) used for sequencing.

ONT individual genotypes for *ii*- and *si*-SNP loci and respective microhaplotypes are summarised in Appendix G and Appendix H respectively.

5.3.1 Individual and species identification, and sex determination

Visual inspection of target SNP markers through the *glPlot* function in R produced easily interpretable results on the efficacy of selected markers for the identification of both individuals and sub-species using both good and poor quality DNA samples. Specifically for the *ii*-SNPs (Fig 5.2), it was observed that each individual (represented by rows in the plot) was different from all others. Additionally, individuals that were sequenced from multiple sources (*i.e.* blood, hair, and faeces) gave almost identical results across loci. Discrepancies were detected at IndLoc17 for Joshi (*G. g. gorilla* sequenced from blood and hair samples) and IndLoc48 for Shufai_F (*G. g. gorilla* sequenced from blood, hair, and faecal samples) (Fig 5.2). As "good" (*i.e.* blood) and "poor" (*i.e.* hair) samples were sequenced using the same flow cells, it is reasonable to assume that, all things being equal, genotyping errors are more likely due to issues with amplification, especially for "poor" samples, rather than variability in pores.

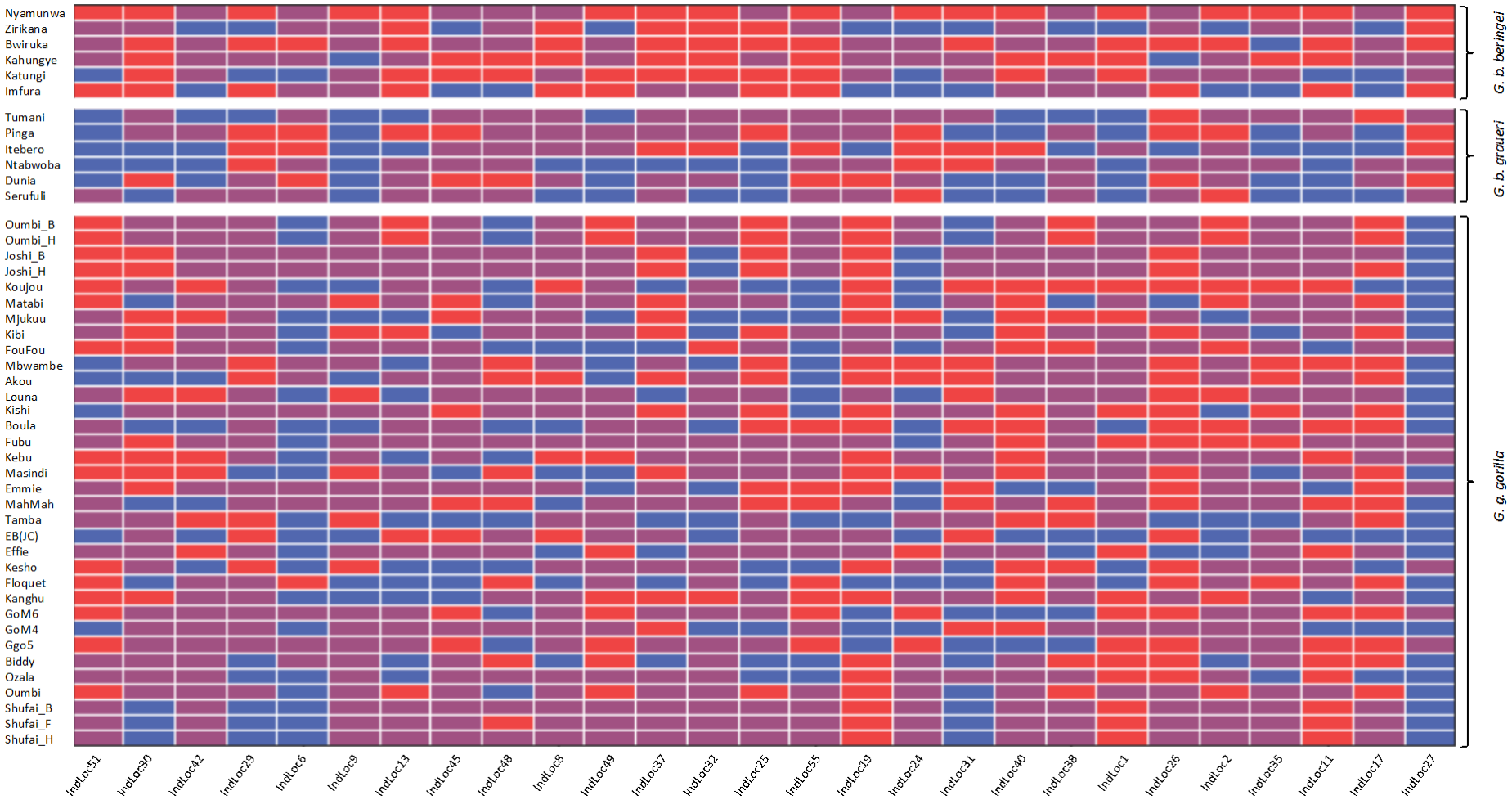


Figure 5.2: Colour coded cross-species Individual identification SNPs (ii-SNPs); REF homozygous (blue), ALT homozygous (red), and heterozygous (purple)

5.3.1.1 Identity analysis for repeated samples

The identity analysis performed in CERVUS (Appendix I for *G. b. beringei*, Appendix J for *G. b. graueri*, and Appendix K for *G. g. gorilla*) validated the deductions made by inspecting Fig 5.2 and confirmed the presence of recurring profiles, namely Joshi, Shufai, and Oumbi (*G. g. gorilla* sequenced from blood and hair samples). In addition, it also confirmed a complete correspondence between previously sequenced individuals belonging to the sub-species *G. b. beringei* and *G. b. graueri* and the *ii*-SNP genotypes obtained via nanopore sequencing on the same individuals, suggesting a good level of sequencing accuracy. The analysis also revealed the presence of two additional repeated samples belonging to the western lowland gorilla sub-species, namely Ggo5 and GoM6 ($p_{ID} = 1.82^{-13}$ and $p_{IDsib} = 7.19^{-7}$), that were sequenced from DNA extracts of the same individual belonging to the lab collection that had previously been labelled as two different individuals. A later inspection of laboratory records confirmed this finding.

5.3.1.2 Species-identification SNPs (*si*-SNPs) analysis

Regarding *si*-SNPs, on the other hand, the visual inspection of markers yielded promising results, whereby three clearly different patterns, containing only homozygous SNPs either for the reference or for the alternative, allowing the user to readily identify the sub-species of gorilla (Fig 5.3). The exception here relates to the presence of two heterozygous SNPs in four individuals belonging to the species *G. g. gorilla* (Fig 5.3). These SNP loci are GbbLoc2 in Mbwambe and Akou and GggLoc9 in Boula, MahMah and Biddy. It is important to note that these individuals were not part of the dataset that was used during the process of SNP selection described in Chapter 4, so direct comparison with available sequencing data was not possible at that stage. These suppositions were supported by the DAPC analysis conducted in R, which resulted in $K = 3$ ($PCs = 3; n_{DA} = 2; proportion\ of\ conserved\ variance = 0.996$), separating out the three subspecies as shown in Fig 5.4. Nevertheless, in STRUCTURE analysis of *si*-SNPs, the best supported value of *K* is 2 ($\Delta K = 503.388$ - Table 5.3), in which eastern (*Gorilla beringei*) and western gorillas (*Gorilla gorilla*) form two separate clusters (Fig 5.5). This result reflects the ability of DAPC to minimise differences within and maximise differences between populations.

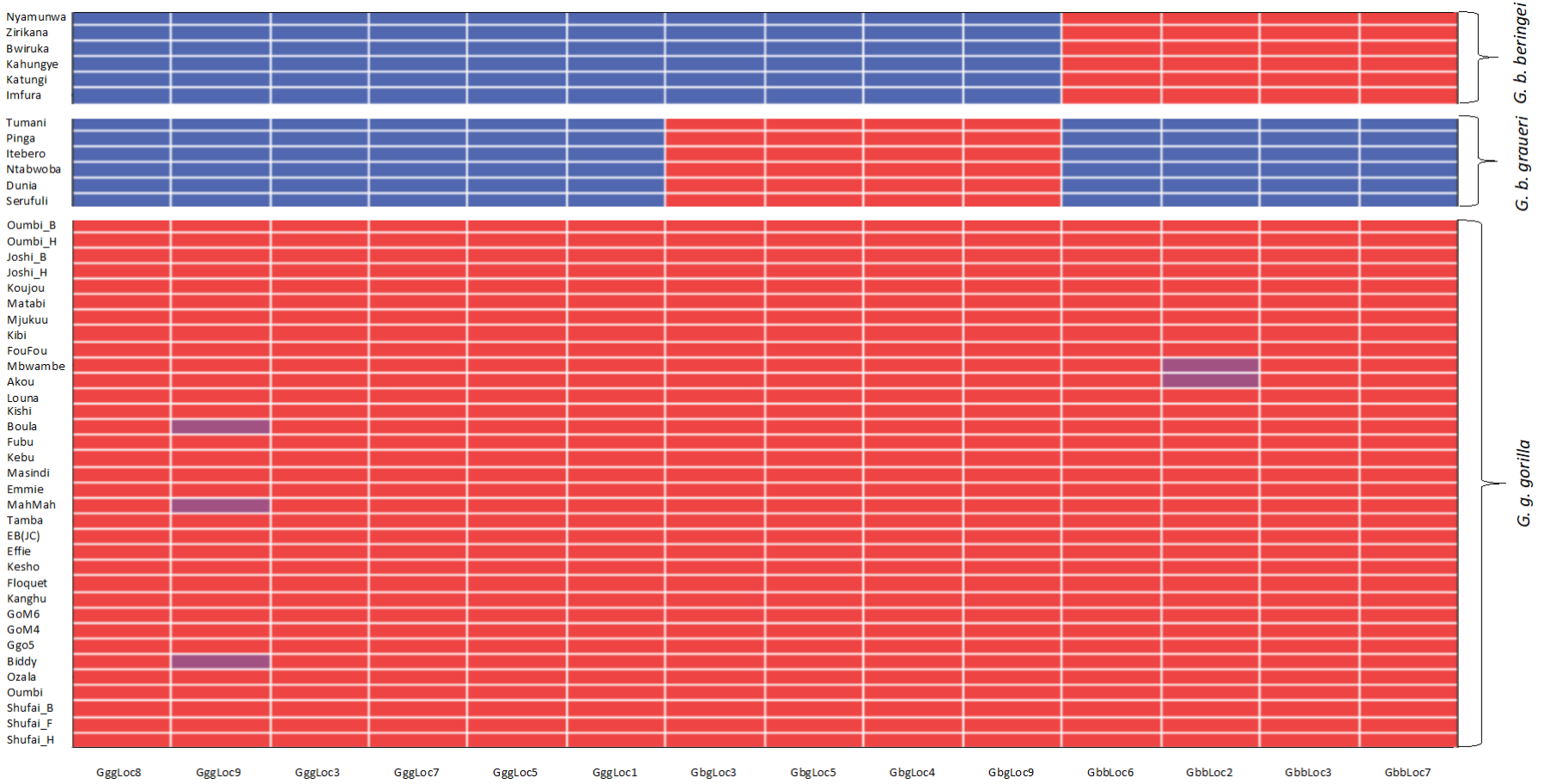


Figure 5.3: Colour-coded cross species species identification SNPs (si-SNPs); REF homozygous (red), ALT homozygous (blue), heterozygous (purple).

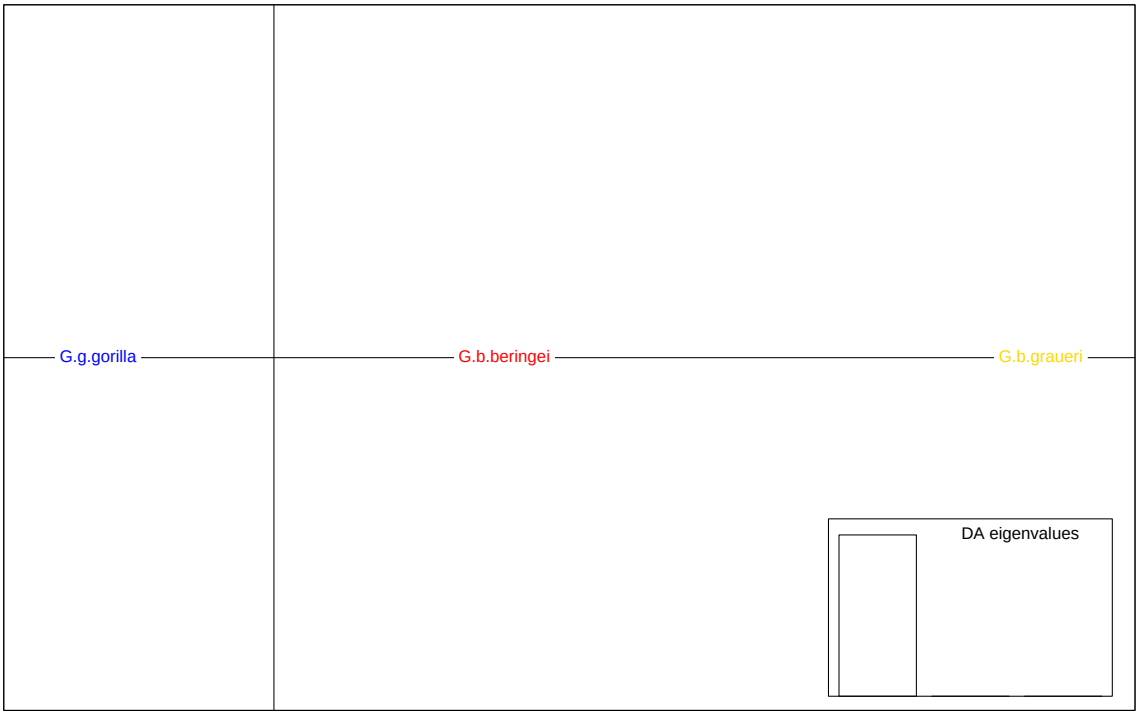


Figure 5.4: DAPC plot based on species identification SNPs (si-SNPs).

Table 5.3: Evanno values for K cluster analysis in STRUCTURE.

K	Reps	Mean LnP(*K*)	Stdev LnP(*K*)	LnP'(*K*)	|LnP''(*K*)|	Delta *K*
1	8	-586.763	0.635	NA	NA	NA
2	8	-170.013	0.656	416.75	330.025	503.388
3	8	-83.287	71.831	86.725	34.175	0.476
4	8	-30.737	1.297	52.550	53.987	41.619
5	8	-32.175	0.381	-1.437	0.279	0.733
6	3	-33.333	1.102	-1.158	0.508	0.461
7	3	-35.000	0.755	-1.667	0.167	0.221
8	3	-36.833	0.569	-1.833	103.967	182.487
9	3	-142.433	89.146	-105.6	97.967	1.099
10	3	-150.067	94.022	-7.633	NA	NA

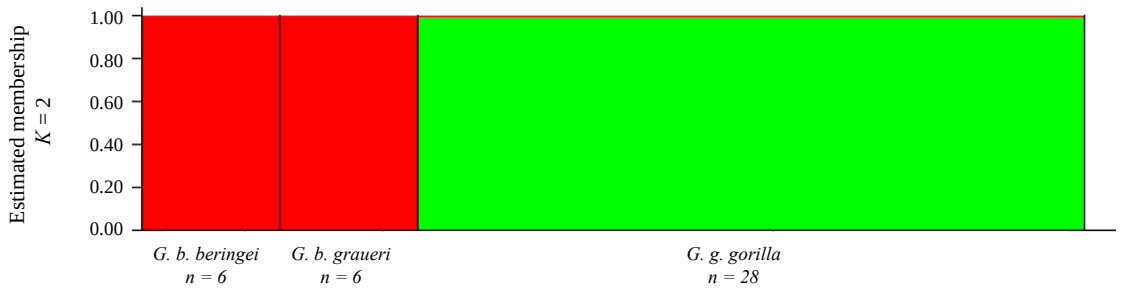


Figure 5.5: Cluster analysis based on si-SNPs in STRUCTURE

5.3.1.3 Analysis of microhaplotypes

Loci were surveyed for additional variants that would define microhaplotypes of 50 bp long. When comparing microhaplotypes with the core set of 41 SNP loci, hence excluding the amelogenin locus, 19 showed additional variants (Fig. 5.6).

Nevertheless, the analysis of microhaplotypes confirmed the results presented above for the analysis of SNPs. In particular, microhaplotypes for GoM6 and Ggo5 genotypes matched (Appendix H) and, in STRUCTURE, $K = 2$ (with $\Delta K_{Evanno} = 232.515$) was once again the best supported value of K number of clusters (Appendix L), suggesting that microhaplotype data added no discriminative power to the informative target *si*-SNP markers. However, microhaplotypes may represent a resource should the number of detectable loci fall even further, as variants in SNP flanking regions may well add discriminative power for the identification of individuals.

The amelogenin-related sex test locus amplified in all individuals and gave results consistent with previously described sex, whereby only X-chromosomal products were observed in females and both X- and Y-chromosomal products were observed in males (data not shown).

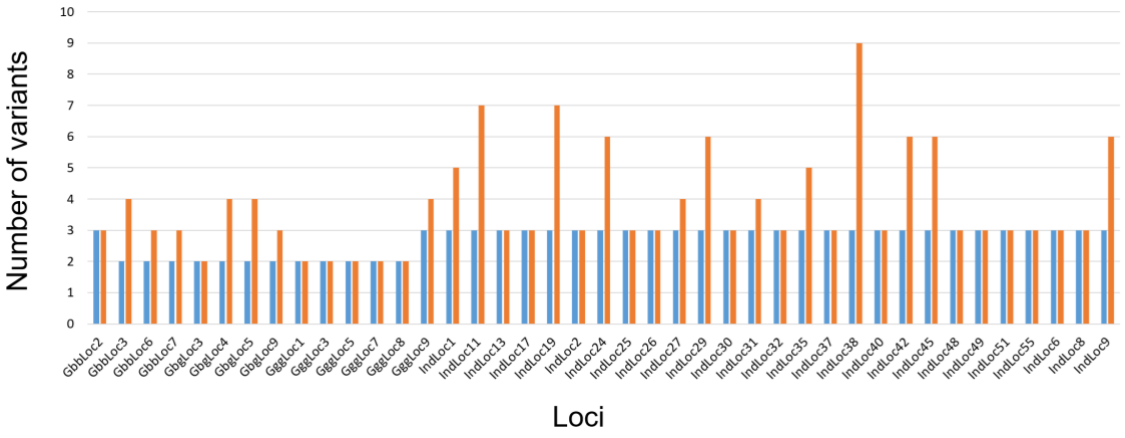


Figure 5.6: Comparison of number of variants for SNPs (blue) and microhaplotypes (orange).

5.3.2 Parentage analysis and family trio pedigree

The results for the maximum likelihood estimation analysis of relatedness between individuals are displayed in Appendix M (*G. b. beringei*), Appendix N (*G. b. graueri*), and Appendix O (*G. g. gorilla*). Because ML-Relate is designed for polymorphic codominant markers (*e.g.* microsatellite data), the analysis of biallelic SNPs can impede the correct estimation of parental relationships between individuals (Ling *et al.*, 2020). Therefore, all ML-Relate-estimated relationships between individuals were re-tested in CERVUS to identify the most likely parent-offspring relationships (Table 5.4).

CERVUS confirmed parent-offspring relationships detected in ML-Relate and, more importantly, it validated the family trio for samples from Ozala (mother), Oumbi (father), and Shufai (offspring) from Twycross Zoo, as shown in Table 5.4 and Figure 5.7.

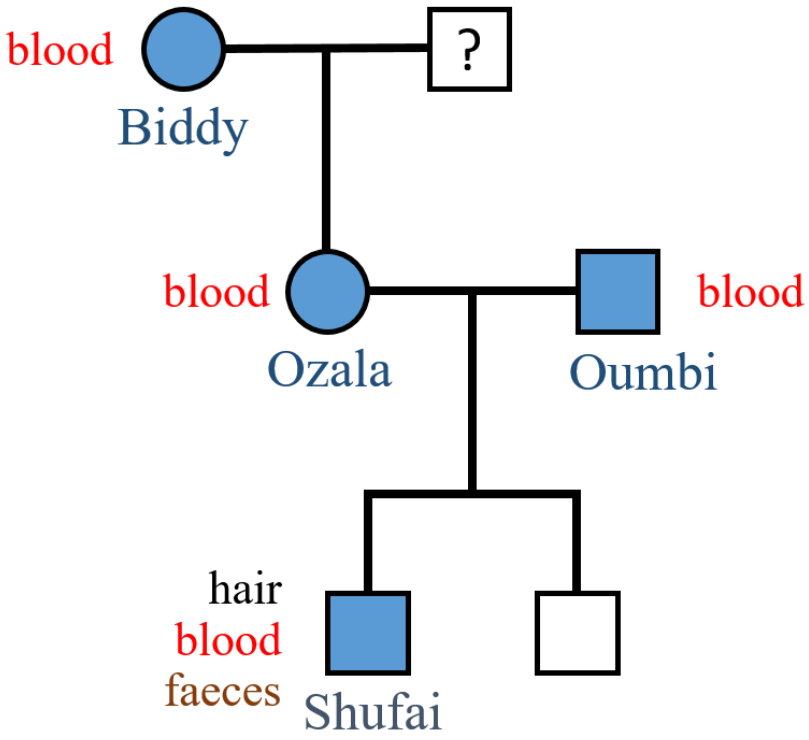


Figure 5.7: Family pedigree structure for samples from Twycross Zoo. Circles indicate female samples whereas square are used to indicate males.

Table 5.4: Estimates of parental analysis in CERVUS. Parentage analysis was performed on all possible combinations between related individuals (according to the ML-Relate output) only. Each male and female individuals was considered a candidate parent.

Species	Offspring ID	Loci Typed	Candidate mother ID	Loci typed	Pair loci compared	Pair loci mismatching	Pair LOD score	Pair top LOD	Pair confidence	Candidate father ID	Loci typed	Pair loci compared	Pair loci mismatching	Pair LOD score	Pair top LOD	Pair confidence	Trio loci compared	Trio loci mismatching	Trio LOD score	Trio top LOD	Trio confidence
Gbb	Nyamunwa	27	Bwiruka	27	27	1	1.78	1.78	+												
Gbb	Bwiruka	27	Nyamunwa	27	27	1	1.78	1.78	+												
Gbg	Dunia	27	Tumani	27	27	0	1.10	1.10	+	Ntabwoba	27	27	2	-7.86	-7.86	+	27	2	-5.89	-5.89	+
Gbg	Serufuli	27	Dunia	27	27	1	-4.41	0.00		Gbg_M_Kaisi	27	27	0	-1.55	-1.55	+	27	2	-2.42	-2.42	+
Gbg	Tumani	27	Dunia	27	27	0	1.10	1.10	+	Ntabwoba	27	27	1	-2.94	-2.94	+	27	3	-4.67	-4.67	+
Gbg	Pinga	27	Dunia	27	27	1	-1.75	-1.75	+	Ggg_M_kaisi	27	27	3	-1.27	0.00		27	5	-1.62	-1.62	-
Gbg	Ntabwoba	27	Tumani	27	27	1	-2.94	0.00		Gbg_M_Mukokya	27	27	2	-8.95	0.00		27	3	-8.21	-8.21	+
Ggg	**Shufai**	**27**	**Ozala**	27	27	**0**	**3.06**	**3.06**	**+**	**Oumbi**	27	27	**0**	**0.97**	**0.97**	**+**	27	**0**	**6.61**	**6.61**	**+**
Ggg	Akou	27	Mbwambe	27	27	0	10.73	10.73	*	Fubu	27	27	2	-8.30	0.00		27	2	3.88	3.88	+
Ggg	Emmie	27	Ggg_F_Carolyn	27	27	1	0.21	0.00		Joshi	27	27	0	3.51	3.51	+	27	1	4.81	4.81	+
Ggg	Fubu	27	Ggg_F_Carolyn	27	27	0	1.81	0.00		Koujou	27	27	0	2.82	2.82	+	27	1	6.15	6.15	+
Ggg	GoM4	27	EB(JC)	27	27	1	2.97	0.00	-	Fubu	27	27	0	1.96	1.96	+	27	2	2.11	2.11	+
Ggg	Kanghu	27	Ggg_F_Porta	27	27	1	-2.11	-2.11		Kebu	27	27	1	0.09	0.00		27	2	-0.9	-0.9	-
Ggg	Kebu	27	Ggg_F_Sandra	27	27	0	1.11	0.00	+	Fubu	27	27	0	0.72	0.00		27	0	3.59	3.59	+
Ggg	Kesho	27	FouFou	27	27	0	3.80	3.80	+	Kebu	27	27	1	-0.40	-0.4	+	27	2	1.41	1.41	+
Ggg	Kibi	27	Masindi	27	27	0	5.14	5.14	+	Fubu	27	27	1	-2.70	0.00		27	1	3.53	3.53	+
Ggg	Kishi	27	Akou	27	27	0	5.06	5.06	+	Joshi	27	27	1	-2.50	0.00		27	1	2.81	2.81	+
Ggg	Louna	27	Ggg_F_Delphi	27	27	0	3.80	3.80	+	Ggg_M_Banjo	27	27	0	2.23	2.23	+	27	0	6.95	6.95	+
Ggg	MahMah	27	Boula	27	27	1	4.03	4.03	+	Ggg_M_Banjo	27	27	1	-3.90	0.00		27	3	-1.1	-1.1	-
Ggg	Masindi	27	Kibi	27	27	0	5.14	5.14	+	Kanghu	27	27	2	-3.80	0.00		27	2	2.20	2.20	+
Ggg	Oumbi	27	Ggg_F_Helen	27	27	0	3.35	3.35	+	Shufai	27	27	0	0.97	0.97	+	27	0	5.73	5.73	+
Ggg	Ozala	27	Mjukuu	27	27	0	3.46	3.46	+	Shufai	27	27	0	3.06	3.06	+	27	1	3.67	3.67	+
Ggg	Joshi	27	Emmie	27	27	0	3.51	3.51	+	Kebu	27	27	0	-0.10	0.00		27	0	5.90	5.90	+
Ggg	Koujou	27	Ozala	27	27	1	0.15	0.15	-	Fubu	27	27	0	2.82	0.00		27	1	2.96	2.96	+
Ggg	Mjukuu	27	Kishi	27	27	1	-0.66	0.00		Kebu	27	27	1	-1.60	-1.6	+	27	2	-.08	-.08	-
Ggg	FouFou	27	Ggg_F_Anthal	27	27	1	-0.51	-0.51	-	Louna	27	27	0	2.47	0.00		27	1	4.92	4.92	+
Ggg	Mbwambe	27	Akou	27	27	0	10.73	10.73	*	Ggg_Tzambo	27	27	1	-0.40	-0.4	+	27	1	13.3	13.3	*
Ggg	Boula	27	MahMah	27	27	1	4.03	4.03	+	Joshi	27	27	1	-2.30	0.00		27	2	2.19	2.19	+
Ggg	EB(JC)	27	Emmie	27	27	1	1.23	1.23	-	GoM4	27	27	1	2.97	2.97	+	27	2	4.27	4.27	+
Ggg	Effie	27	Ggg_F_Dian	27	27	0	2.34	2.34	-	GoM6	27	27	1	-1.20	0.00		27	1	3.87	3.87	+
Ggg	Biddy	27	Ggg_F_Dolly	27	27	0	0.38	0.38	-	Shufai	27	27	0	0.89	0.89	+	27	1	-0.9	-0.9	-

5.3.3 Summary of *ii*- and *si*-SNP forensic statistics

Combined *ii*- and *si*-SNP statistics for *G. b. beringei* (Table 5.5), *G. b. graueri* (Table 5.6), and *G. g. gorilla* (Table 5.7) were generated using the browser-based application STRAF 1.0.5 (Gouy and Zieger, 2017). Related individuals (*i.e.* HS, FS and PO in *G. g. gorilla* and FS and PO in *G. b. beringei* and *G. b. graueri*) were removed for the analysis of population statistics.

Table 5.5: Standard forensics parameters for ii- and si-SNPs in *G. b. beringei*.

locus	N	Nall	GD	PIC	PM	PD	Hobs	PE	TPI	pHW
IndLoc1	12	2	0.409	0.305	0.500	0.500	0.167	0.021	0.600	0.263
IndLoc11	12	2	0.409	0.305	0.500	0.500	0.167	0.021	0.600	0.286
IndLoc13	12	2	0.167	0.141	0.722	0.278	0.167	0.021	0.600	1.000
IndLoc17	12	2	0.409	0.305	0.500	0.500	0.500	0.188	1.000	1.000
IndLoc19	12	2	0.485	0.346	0.556	0.444	0.667	0.379	1.500	0.524
IndLoc2	12	2	0.545	0.375	0.333	0.667	0.333	0.078	0.750	0.487
IndLoc24	12	2	0.485	0.346	0.389	0.611	0.333	0.078	0.750	1.000
IndLoc25	12	2	0.303	0.239	0.556	0.444	0.333	0.078	0.750	1.000
IndLoc26	12	2	0.530	0.368	0.389	0.611	0.500	0.188	1.000	1.000
IndLoc27	12	2	0.303	0.239	0.556	0.444	0.333	0.078	0.750	1.000
IndLoc29	12	2	0.530	0.368	0.389	0.611	0.167	0.021	0.600	0.148
IndLoc30	12	2	0.167	0.141	0.722	0.278	0.167	0.021	0.600	1.000
IndLoc31	12	2	0.545	0.375	0.333	0.667	0.333	0.078	0.750	0.481
IndLoc32	12	2	0.167	0.141	0.722	0.278	0.167	0.021	0.600	1.000
IndLoc35	12	2	0.545	0.375	0.333	0.667	0.333	0.078	0.750	0.512
IndLoc37	12	2	0.167	0.141	0.722	0.278	0.167	0.021	0.600	1.000
IndLoc38	12	2	0.545	0.375	0.500	0.500	0.667	0.379	1.500	1.000
IndLoc40	12	2	0.409	0.305	0.500	0.500	0.500	0.188	1.000	1.000
IndLoc42	12	2	0.485	0.346	0.556	0.444	0.667	0.379	1.500	0.499
IndLoc45	12	2	0.545	0.375	0.333	0.667	0.333	0.078	0.750	0.437
IndLoc48	12	2	0.530	0.368	0.389	0.611	0.500	0.188	1.000	1.000
IndLoc49	12	2	0.485	0.346	0.389	0.611	0.333	0.078	0.750	1.000
IndLoc51	12	2	0.530	0.368	0.389	0.611	0.500	0.188	1.000	1.000
IndLoc55	12	2	0.167	0.141	0.722	0.278	0.167	0.021	0.600	1.000
IndLoc6	12	2	0.545	0.375	0.500	0.500	0.667	0.379	1.500	1.000
IndLoc8	12	2	0.303	0.239	0.556	0.444	0.333	0.078	0.750	1.000
IndLoc9	12	2	0.545	0.375	0.500	0.500	0.667	0.379	1.500	1.000
GbbLoc2	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbbLoc3	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbbLoc6	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbbLoc7	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc3	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc4	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc5	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc9	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc1	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc3	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc5	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc7	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc8	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc9	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000

Table 5.6: Standard forensics parameters for ii- and si-SNPs in *G. b. graueri*.

locus	N	Nall	GD	PIC	PM	PD	Hobs	PE	TPI	pHW
IndLoc1	12	2	0.167	0.141	0.722	0.278	0.167	0.021	0.600	1.000
IndLoc11	12	2	0.303	0.239	0.556	0.444	0.333	0.078	0.750	1.000
IndLoc13	12	2	0.530	0.368	0.389	0.611	0.500	0.188	1.000	1.000
IndLoc17	12	2	0.485	0.346	0.389	0.611	0.333	0.078	0.750	1.000
IndLoc19	12	2	0.545	0.375	0.500	0.500	0.667	0.379	1.500	1.000
IndLoc2	12	2	0.485	0.346	0.556	0.444	0.667	0.379	1.500	0.497
IndLoc24	12	2	0.303	0.239	0.556	0.444	0.333	0.078	0.750	1.000
IndLoc25	12	2	0.409	0.305	0.500	0.500	0.167	0.021	0.600	0.261
IndLoc26	12	2	0.485	0.346	0.389	0.611	0.333	0.078	0.750	1.000
IndLoc27	12	2	0.409	0.305	0.500	0.500	0.500	0.188	1.000	1.000
IndLoc29	12	2	0.485	0.346	0.389	0.611	0.333	0.078	0.750	1.000
IndLoc30	12	2	0.485	0.346	0.389	0.611	0.333	0.078	0.750	1.000
IndLoc31	12	2	0.530	0.368	0.389	0.611	0.167	0.021	0.600	0.166
IndLoc32	12	2	0.530	0.368	0.389	0.611	0.500	0.188	1.000	1.000
IndLoc35	12	2	0.303	0.239	0.556	0.444	0.333	0.078	0.750	1.000
IndLoc37	12	2	0.545	0.375	0.500	0.500	0.667	0.379	1.500	1.000
IndLoc38	12	2	0.485	0.346	0.556	0.444	0.667	0.379	1.500	0.533
IndLoc40	12	2	0.409	0.305	0.500	0.500	0.167	0.021	0.600	0.262
IndLoc42	12	2	0.303	0.239	0.556	0.444	0.333	0.078	0.750	1.000
IndLoc45	12	2	0.485	0.346	0.556	0.444	0.667	0.379	1.500	0.522
IndLoc48	12	2	0.530	0.368	0.722	0.278	0.833	0.662	3.000	0.377
IndLoc49	12	2	0.303	0.239	0.556	0.444	0.333	0.078	0.750	1.000
IndLoc51	12	2	0.167	0.141	0.722	0.278	0.167	0.021	0.600	1.000
IndLoc55	12	2	0.485	0.346	0.556	0.444	0.667	0.379	1.500	0.504
IndLoc6	12	2	0.409	0.305	0.500	0.500	0.500	0.188	1.000	1.000
IndLoc8	12	2	0.485	0.346	0.556	0.444	0.667	0.379	1.500	0.520
IndLoc9	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbbLoc2	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbbLoc3	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbbLoc6	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbbLoc7	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc3	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc4	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc5	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc9	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc1	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc3	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc5	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc7	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc8	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc9	12	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000

Table 5.7: Standard forensics parameters for ii- and si-SNPs in *G. g. gorilla*.

locus	N	Nall	GD	PIC	PM	PD	Hobs	PE	TPI	pHW
IndLoc1	58	2	0.479	0.360	0.396	0.604	0.483	0.173	0.967	1.000
IndLoc11	58	2	0.487	0.364	0.375	0.625	0.448	0.146	0.906	0.696
IndLoc13	58	2	0.479	0.360	0.434	0.566	0.552	0.237	1.115	0.455
IndLoc17	58	2	0.448	0.344	0.394	0.606	0.310	0.068	0.725	0.095
IndLoc19	58	2	0.422	0.329	0.432	0.568	0.241	0.042	0.659	0.024
IndLoc2	58	2	0.506	0.374	0.386	0.614	0.517	0.203	1.036	1.000
IndLoc24	58	2	0.494	0.368	0.348	0.652	0.345	0.084	0.763	0.139
IndLoc25	58	2	0.506	0.374	0.356	0.645	0.448	0.146	0.906	0.695
IndLoc26	58	2	0.354	0.287	0.498	0.502	0.241	0.042	0.659	0.096
IndLoc27	58	2	0.216	0.190	0.634	0.366	0.241	0.042	0.659	1.000
IndLoc29	58	2	0.508	0.375	0.524	0.476	0.690	0.412	1.611	0.054
IndLoc30	58	2	0.499	0.370	0.363	0.637	0.448	0.146	0.906	0.699
IndLoc31	58	2	0.509	0.375	0.337	0.664	0.379	0.102	0.806	0.262
IndLoc32	58	2	0.479	0.360	0.486	0.514	0.621	0.316	1.318	0.130
IndLoc35	58	2	0.506	0.374	0.491	0.509	0.655	0.362	1.450	0.175
IndLoc37	58	2	0.506	0.374	0.356	0.645	0.448	0.146	0.906	0.693
IndLoc38	58	2	0.506	0.374	0.386	0.614	0.517	0.203	1.036	1.000
IndLoc40	58	2	0.436	0.336	0.406	0.595	0.345	0.084	0.763	0.402
IndLoc42	58	2	0.506	0.374	0.432	0.568	0.586	0.275	1.208	0.440
IndLoc45	58	2	0.508	0.375	0.406	0.595	0.552	0.237	1.115	0.707
IndLoc48	58	2	0.499	0.370	0.363	0.637	0.448	0.146	0.906	0.735
IndLoc49	58	2	0.508	0.375	0.406	0.595	0.552	0.237	1.115	0.730
IndLoc51	58	2	0.479	0.360	0.372	0.628	0.414	0.123	0.853	0.685
IndLoc55	58	2	0.494	0.368	0.348	0.652	0.345	0.084	0.763	0.122
IndLoc6	58	2	0.334	0.274	0.505	0.495	0.345	0.084	0.763	1.000
IndLoc8	58	2	0.503	0.372	0.463	0.538	0.621	0.316	1.318	0.262
IndLoc9	58	2	0.508	0.375	0.406	0.595	0.552	0.237	1.115	0.715
GbbLoc2	58	2	0.068	0.064	0.872	0.128	0.069	0.004	0.537	1.000
GbbLoc3	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbbLoc6	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbbLoc7	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc3	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc4	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc5	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GbgLoc9	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc1	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc3	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc5	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc7	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc8	58	1	0.000	0.000	1.000	0.000	0.000	0.000	0.500	1.000
GggLoc9	58	2	0.100	0.093	0.815	0.186	0.103	0.009	0.558	1.000

N number of samples

N_{all} number of alleles

GD	Genetic diversity/expected heterozygosity
PIC	Polymorphism Information Content
PM	match probability
PD	Power of discrimination
H_{obs}	Observed heterozygosity
PE	power of exclusion
TPI	typical paternity index
pHW	Hardy-Weinberg p-value

5.3.4 Analysis of inbreeding coefficient (F_{IS}) in *ii*-SNPs and microhaplotypes

The analysis of F_{IS} produced, as expected, no significant values. Both the high level of heterozygosity of *ii*-SNPs and the reduced number of individuals being analysed are likely causes of this result. The same analysis conducted on microhaplotype data gave similar results, hence corroborating the previous observation that microhaplotypes add little information to the analysis.

5.4 Discussion

This Chapter extended the work presented in Chapter 4 by focusing on a larger number of samples belonging to the three available sub-species of gorilla, namely mountain (*G. b. beringei*), eastern lowland (*G. b. graueri*), and western lowland (*G. g. gorilla*). It also described the work conducted for the extraction, preparation, and analysis of DNA from non-invasively collected samples with the aim to test the utility of nanopore sequencing in conducting DNA testing on site, where non-invasive samples represent the most abundant and easily accessible source of biological material.

As expected, only a portion (circa 46% or 41 out of 89) of the identified markers gave comparable results across sequencing runs, species, and biological sources, most likely due both to variability in the number of pores, which affects sequencing throughput, and inconsistent sample quality, which influenced locus amplification across samples (Taberlet *et al.*, 1999). In fact, DNA from non-invasive samples is often of poor quality due to environmental degrading factors (*e.g.* ultraviolet radiation, heat, moisture), and can lead to lower loci-called rates, increased allelic dropout, fewer informative loci, and an overall reduction in genotyping accuracy (Valiere *et al.*, 2007; Schultz *et al.*, 2022). Nevertheless, the approach discussed in this Chapter was effective in providing both species and individual identification through a variety of DNA sources, suggesting that the method could indeed find important applications for the study of elusive gorillas in the wild and support the fight against wildlife crimes. Indeed, the relative ease of use, paired with the reduced costs and portability of the DNA testing system described in this Chapter, open up the opportunity to support the use of genetic testing for conservation-related purposes.

To date, the complexity of genetic data analysis and results interpretation represents one of the main impediments to the widespread adoption of DNA testing in ecology and conservation research (Ogden, 2011; Sahlin *et al.*, 2021; Pomerantz *et al.*, 2022). For this reason, the analysis of sequencing data discussed in this Chapter was intentionally performed by exploiting widely used analytical software for which a plethora of information can be freely retrieved from the internet. In addition, the qualitative approach to the identification of different individuals and species of belonging based on the visual inspection of Fig 5.2 and Fig 5.3 can provide immediate access to information that allows

the estimation of fundamental measures of biological diversity such as species richness and abundance (Gotelli and Colwell, 2001). Alternatively, the provision of sequence-based data for species and individual identification allows the user to create records of individual profiles, finding important applications in the field of wildlife forensics studies (Alacs *et al.*, 2010; Johnson *et al.*, 2014; Gondhali and Mishra, 2022).

As expected, the analysis of genetic diversity among species revealed localised (but not globally, $\chi^2 = 3.890,\ df = 2,\ P\ value = 0.143$) higher levels of observed heterozygosity in western gorillas than in the two sub-species of eastern gorilla. This corroborates previous findings (McGrath *et al.*, 2022; Xue *et al.*, 2015) and supports the idea that conservation efforts should focus on facilitating gene flow between different populations to prevent genetic defect issues (*e.g.* syndactyly, see Xue et al., 2015), as has already been observed in highly inbred mountain gorilla populations. In this regard, landscape-based management activities, such as the creation of green corridors to provide connectivity between isolated populations, as well as animal translocations, can play a crucial role in enhancing genetic diversity, which is key to the long-term survival of endangered species (Frankham, 2003; Frankham *et al.*, 2017).

Pedigree analysis for the family trio from Twycross Zoo confirmed parent-offspring transmission of alleles, hence suggesting that detected variants were indeed real. This is an important step for ensuring the reliability of sequencing data. In fact, while genotype accuracy is key to downstream analysis, it is often limited by the introduction of sequencing errors intrinsic to high-throughput sequencing techniques (Mastretta-Yanes *et al.*, 2015). Usually, the frequency with which these occur depends on the platform being used, and appropriate filtering protocols (like the one applied in this study) are crucial to ensure reliability of the data (Roques *et al.*, 2019), despite the unavoidable consequential loss of loci available for analysis; here only a core set of 42 loci survived. While it is difficult to predict what consequences further reductions in the number of loci might have on the accuracy of relatedness testing, this study confirmed the importance of including replicates (including technical replicates) and family groups to facilitate the detection of sequencing errors (Roques *et al.*, 2019). Moreover, specifically for ONT technology which, despite several recent improvements to the hardware and basecalling software, still suffers from a degree of throughput and accuracy variability in response to the pore viability on individual flow cells, it was found that sequencing of matched good

and poor-quality DNA samples added reliability to the final results. Finally, the additional variants discovered through the analysis of SNP flanking regions (*i.e.* microhaplotype sequences) might help increase power of discrimination and the reliability of parentage analysis should the number of loci decrease below what was reported in this Chapter.

Crucially, the lack of complete correspondence between parental and offspring genotypes due to sequencing errors can impede correct assignment and bias results (Congiu *et al.*, 2011). Accuracy of relatedness testing can further be hampered by the combination of small sample size and non-random sampling of individuals. In fact, marker-based relatedness estimators have traditionally been developed on the assumption that allele frequencies are computed without errors from large random-mating reference populations, at Hardy-Weinberg equilibrium (Ritland, 1996; Lynch and Ritland, 1999; Wang, 2017; Wang, 2014). These conditions however are rarely met, particularly in the study of non-model organisms, like non-human great apes, where population dynamics and structure may not be fully known (Wang, 2017). This issue is made worse in the analysis of captive-bred populations that are inherently small and inbred, and where long homologous stretches between sampled individuals and the reference population may be identical by descent (IBD). As a result, estimators of relatedness are often biased, undermining the accuracy of parent-offspring and sibling detection. Furthermore, most commonly used estimators were developed mainly for application to microsatellite (and hence multiallelic) data, whereby bias due to small sample size is still deemed unimportant (Ritland, 1996; Wang, 2017). Following the recent improvements in sequencing technology, it is now possible to analyse an ever growing number of loci simultaneously which, however, is seldom followed by an increase in sample size, leading to a small number of genotypes (due to the paucity of sampled individuals) and a consequent high rate of missing data. Under these circumstances, commonly used relatedness estimators can become biased and the accuracy is influenced by bias rather than sampling errors, particularly when the number of loci is higher than the number of individuals being analysed (Wang, 2017; Wang, 2014). Efforts should be made to increase sample sizes and with that the accuracy of population allele frequencies to be used in relatedness testing. Importantly, the development of accessible and straightforward methods of analysis will play a crucial role in facilitating the acquisition of additional population genotype data.

Recent technological advances, such as the commercialisation of portable thermal cyclers like the miniPCR and the BentoLab, which comes with a fitted centrifuge for 12 samples, a thermal cycler block and a small gel tank for gel electrophoresis, have now opened up the possibility of operating in field laboratory conditions (Edwards *et al.*, 2016; Quick *et al.*, 2016; Pomerantz *et al.*, 2017; Pomerantz *et al.*, 2018; Pomerantz *et al.*, 2022; Carroll *et al.*, 2018; Chang *et al.*, 2020). In addition, the reduced initial capital investment makes this equipment affordable by a much larger group of people that could now operate in virtually any context. While the current SARS-CoV-2 pandemic, and the consequent ban on international travel, have prevented me from conducting tests *in situ*, for which I would have originally travelled to the Volcano National Park (Rwanda), previous studies, based on side-by-side comparisons between standard benchtop and portable thermal cyclers, have provided supporting evidence that portable technology represents a valid and effective alternative for DNA extraction and amplification (Chang *et al.*, 2020).

Chapter 6 General discussion and future directions

6.1 MPS technology in wildlife forensics

The introduction of massively parallel sequencing (MPS) has rapidly changed the way genetic research is conducted (Børsting and Morling, 2015). Forensic genetics has benefited from the potential of MPS to define variants in STRs – both in their internal arrays and flanking regions – which allows the discrimination of isoalleles (Xavier and Parson, 2017; Huszar *et al.*, 2018; Beasley *et al.*, 2021) and hence reduces random match probability, while also effectively detecting associated SNPs.

The improved resolution offered by MPS over conventional sequencing methods (*e.g.* CE-based approaches), together with greater multiplexing capability (*e.g.* the ForenSeq kit in human forensics has been validated to interrogate 27 autosomal STRs, 24 Y-chromosomal and 7 X-chromosomal STRs, along with 172 SNPs (Jäger *et al.*, 2017)), enables the analysis of traces of degraded DNA. This has significant repercussions for forensic analyses, which have now found important applications in the implementation of effective wildlife conservation measures. The ability to exploit non-invasively collected samples, often containing highly fragmented DNA, allows researchers to study elusive and cryptic animal species while reducing any form of disturbance and lowering the risk of spreading infectious diseases. In addition, MPS has the potential to increase the accuracy of the analysis of mixed DNA samples (Novroski, 2021) and hence to identify different contributors, which is a crucial aspect in the fight against the illegal wildlife trade.

However, despite the clear advantages over conventional methods, MPS requires considerable investments in new and expensive technology platforms, laboratory procedures, analysis, and expertise for the interpretation of the results (Huszar *et al.*, 2018; Watsa *et al.*, 2020), which make it largely inaccessible to research and conservation groups in developing countries. As a result, biological samples need to be exported

outside the habitat country to sites where sequencing can be conducted, thereby increasing the cost and time for sequencing (Krehenwinkel *et al.*, 2019; Watsa *et al.*, 2020), and limiting the role of MPS in the implementation of conservation projects on a global scale.

Recent technological improvements, such as the introduction of portable thermal cyclers (*e.g.* Bento Lab and miniPCR) and miniaturised sequencing devices (*e.g.* ONT MinION™), now enable DNA testing virtually anywhere. This technology has successfully been used to conduct species barcoding analysis in remote areas (Edwards *et al.*, 2016; Quick *et al.*, 2016; Carroll *et al.*, 2018; Pomerantz *et al.*, 2017; Pomerantz *et al.*, 2022), contributing to the adoption of genetic techniques to address fundamental ecological questions and conservation issues.

In this context, the current study applied a human-based forensic kit to interrogate the orthologs of 27 autosomal STRs for individual and (sub)species identification in non-human great apes, following a lab-centred MPS approach. It then explored the opportunity to exploit novel portable sequencing technology for the analysis of a novel panel of SNPs – designed for in situ applications – to provide individual and sub-species identification in Gorilla, using non-invasive samples (*e.g.* hair and faeces).

6.2 Summary of the results

In Chapter 2, the Verogen ForenSeq DNA Signature Prep Kit was used to investigate the orthologs of 27 human autosomal forensic STRs in 52 African great apes (14 chimpanzees; 4 bonobos; 16 western lowland, 6 eastern lowland; and 12 mountain gorillas). The orthologs of 24 out 27 aSTRs amplified across species, while a core set of thirteen could be genotyped in all individuals, providing individual and (sub)species identification. Allelic diversity and power of discrimination were greater when considering STR sequences rather than allele lengths. Comparisons with human and orangutan (*Pongo spp*.) showed general conservation of repeat types and allele size ranges across loci. However, variation in repeat array structures and little relationship with the known phylogeny suggest stochastic origins of mutations giving rise to diverse imperfect repeat arrays. In addition, interruptions within long repeat arrays in African great apes do not appear to reduce allelic diversity, which might be indicative of a possible mutational difference compared to humans. This Chapter concluded that despite some variability in amplification success, orthologs of most aSTRs in the ForenSeq DNA Sequencing Prep Kit can be analysed in African great apes and that MPS provides better resolution for both individual and (sub)species identification than standard CE-based approaches (Fedele *et al.*, 2022). There followed a brief discussion on the limitations of MPS, which focused on the high start-up costs (*e.g.* for equipment and reagents), labour intensive sample preparation and steep learning curves associated with data analysis. In addition, the lack of well-established research facilities in remote areas and in developing countries was indicated as one of the main reasons that hamper the applicability of MPS for wildlife conservation purposes.

Focusing on the potential applications of recent developments in portable sequencing technologies, in Chapter 3 a novel panel of autosomal SNPs for in situ gorilla individual and sub-species identification was designed. Compared to conventionally used markers for individual (*e.g.* microsatellite) and species (*e.g.* mitochondrial genome) identification, autosomal SNPs have the advantage that they work well on degraded DNA – ideal for the analysis of non-invasively collected samples – and are biparentally inherited – providing introgression information about hybrids of crosses between closely-related species (Vilà *et al.*, 2003). Drawing from available whole genome sequencing data (Xue *et al.*, 2015; Prado-Martinez *et al.*, 2013) for three different sub-species of gorilla, namely western

lowland (*G. g. gorilla*), eastern lowland (*G. b. graueri*) and mountain gorilla (*G. b. beringei*), two sets of SNPs were identified. A first set consisting of highly variable SNPs across individuals, here named ii-SNPs, was selected to provide individual identification; and a second, consisting of SNPs fixed in one sub-species but absent in the others, named si-SNPs, was developed to provide sub-species identification. This resulted in a panel of 85 autosomal SNPs (*i.e.* 56 ii-SNPs, 29 si-SNPs) and one sex-test locus.

In Chapter 4, MinION™ sequencing was applied to investigate variation in the 85 autosomal SNPs contained in the panel developed in Chapter 3 on a single individual of previously sequenced (Prado-Martinez *et al.*, 2013) mountain gorilla (Nyamunwa). Comparing nanopore-generated genotypes with previously available WGS data (for the same individual) gave complete correspondence for 82 out of 86 loci; the remaining four failed to meet the quality criteria and were therefore excluded from the analysis. In addition, it was shown that, following the improvements in basecalling software (Guppy), SNP amplicon ligation protocols such as the one applied by Cornelis et al. (2017), which was developed to overcome minimum length requirements, are no longer necessary for effective sequencing of short DNA fragments with the MinION™. This saves time and resources for sequencing, hence making this approach accessible for the study of gorillas in remote areas and, more generally, in sites that lack well-established research facilities.

In Chapter 5, ONT sequencing of the ii- and si-SNPs was applied to a sample of 46 gorillas, belonging to the three sub-species western lowland (*G. g. gorilla*), eastern lowland (*G. b. graueri*), and mountain gorilla (*G. b. beringei*), for which DNA was extracted from a variety of sources, including blood, hair and faeces. A core set of 41 SNPs, consisting of 27 *ii* and 14 *si*-SNPs, and one sex-locus marker, produced comparable results across individuals and sequencing runs. Identity analysis revealed the presence of recurrent individuals, while pedigree analysis confirmed Mendelian transmission of variants, suggesting that these were indeed real.

Additional polymorphic sites were observed within the amplicons of 12 of the 27 ii-SNPs and 7 of the 14 si-SNPs, forming up to 10 microhaplotypes per locus. The analysis of microhaplotypes corroborated the results of SNP data analyses. Furthermore, and in accordance with previous studies, observed heterozygosity at a number of loci was found to be higher in western lowland gorillas than in individuals belonging to the two eastern

sub-species. This reflects the reduced effective population size in eastern gorillas, which has been reported to be the cause of the insurgence of genetic defects among individuals of this species.

6.3 Limitations and caveats of the study

Appropriate sampling is central to drawing robust and significant conclusions in population studies. However, the outbreak of SARS-Cov-2, and the following ban on international travel, meant that this study had to rely on samples collected from UK zoos and donations from local research institutions (*e.g.* Wellcome Sanger Institute) for the analysis. This raises concerns around true population allele frequencies and, consequently, the reliability of estimated population parameters. In addition, due to the small size and the characteristics (*e.g.* non-random sampling from non-natural populations) of the sample being analysed here, which in the case of eastern gorilla species here coincided to the quasi totality of the reference population, interpolation of genetic diversity measures was deemed not worthwhile.

While the focal point of this study was to apply MPS for individual and (sub)species identification in non-human great apes, and in particular in gorilla species, it also showed the potential to conduct population studies as genetic data from wild reference populations becomes available. In addition, pedigree analysis of a family trio, based on MinION™ sequencing data, was still successful in detecting Mendelian variant transmission, suggesting that the statistical methods used for the analysis of population parameters may well be effective for the identification of closely-related individuals even in small and inbred populations.

Specifically for ONT sequencing, the variability in the number and viability of flow cell pores still represents an obstacle to the comparison of data across runs. As a result, this hinders the ability to readily identify the causes of genotyping failures. In fact, while DNA sample quality is crucial for successful loci amplification, it was found that loci for which genotyping failed were not statistically different in length to those that were successfully sequenced and analysed, suggesting that inconsistent flow cell quality, rather than DNA quality, is the main reason for such failures. Similarly, the issue of variability among different flow cells makes it practically impossible to investigate whether genotyping failure is due to variants in primer binding sites that could potentially inhibit the amplification of markers. Nevertheless, Chapter 5 demonstrated that sequencing sample replicates represents an effective way to address this issue, especially when DNA quality represents an additional source of uncertainty. Notwithstanding the consistency

issue, this study provided evidence that (at least currently), sequencing of up to nine samples on one Flongle flow cell can provide sufficient accuracy, enabling both individual and species identification.

6.4 Final considerations and future directions

Here, it was shown that MPS analysis using the commercially available human-based forensic kit provided greater detail and resolution than conventional approaches for the study of STRs in non-human great ape species. In Chapter 2, the analysis of the orthologs of 27 human STRs provided reliable individual identification for a core set of 13 orthologous loci that amplified across all analysed (sub)species. Additionally, cluster analysis (using DAPC) of sequence-based data was successful in assigning individuals to the right (sub)species, resolving five different clusters, whereas length-based (or CE equivalent) data analysis gave only three clusters, separating the Pan genus from the two species *G. gorilla* and *G. beringei*. This is because the additional information extracted with the use of MPS, such as detection of isometric heterozygotes and variants in STR-flanking regions, provides better resolution and hence increases the discriminative power of the analysis.

MPS also enables the detection of SNPs which allow researchers to exploit non-invasively collected samples, an accessible source of biological material in the wild, to conduct DNA testing on a large number of samples simultaneously. This reduces the time required for sequencing, reduces any form of disturbance to the animals, and also provides an effective tool for the investigation of degraded DNA samples, hence facilitating the fight against illegal wildlife trade.

One of the major difficulties with MPS applications in wildlife conservation is that it is costly and it requires sound bioinformatics skills for the analysis of sequencing data. As laboratory infrastructure is not spread evenly across the globe, and it does not match areas of high biodiversity, samples have routinely been exported outside the habitat countries for the analysis. This involves time-consuming applications for export permits, additional financial investments and the risk of losing and even further degrading DNA samples. In addition, exporting samples takes away the opportunity for local scientists to actively participate in research studies beyond on-site sample collection activities (Watsa, 2022). As discussed in Chapters 3, 4 and 5 the ONT MinION™ portable sequencing device, effectively circumvents these issues. Here, it was demonstrated that MinION™ sequencing is indeed effective for the analysis of degraded samples and protocols can be developed to make its use field friendly. Initiatives like ORG.one by ONT to support

rapid DNA sequencing of critically endangered species (accessible at: http://org.one/oo), are likely to play a fundamental role in facilitating the introduction of this technology to a wider range of activities and environments. In this regard, financially and operationally self-sufficient field-based genomics labs (*e.g.* Sumak Kawsay In Situ, accessible at : https://www.sumakkawsayinsitu.org/) have been created to realise genetic research and capacitate local scientists and conservationists (Watsa *et al.*, 2020). In addition, by reducing the time for sequencing, this technology may well represent an effective tool in the fight against wildlife crimes. Yet, the lack of a well-established and widespread system of distribution for the hardware and reagents, can still represent a significant impediment for research groups to access the technology in very remote areas (Zane Libke pers. comm.).

Specifically to the study of gorillas, as for many other wildlife species, additional efforts are needed to expand sample size beyond captive bred individuals, in order to increase the reliability of population genetic analyses. This will in turn facilitate the introduction of genetic analysis in field-based conservation programmes around the globe. In this context, once again, the portability of MinION™, alongside that of DNA preparation equipment (*e.g.* miniPCR and Bento Lab), opens the possibility to generate data faster and with (comparatively) limited resources.

As the human population continues to grow (Roser, 2013), wildlife faces continuous and increasing pressures (Daszak *et al.*, 2000; Beebee, 2022; van der Wal *et al.*, 2022). The extreme events that have characterised the last decades are a stark reminder that we are already seeing the effects of nature decline on our lives (Maxwell *et al.*, 2019; Keesing and Ostfeld, 2021). Climate warming, loss of natural spaces, species extinctions, and pollution are some of the biggest challenges in our history and, while technology can help, no single approach can be applied to solve the diverse array of issues we have come to face (Berger-Tal and Lahoz-Monfort, 2018). There follows the need to develop interdisciplinary expertise so as to make effective use of the large amounts of data that are being generated to inform, monitor, and implement wildlife conservation and management programmes around the globe.

Electronic Appendices

Appendix A	Amplification_behaviour_of_loci_included_in_the_multiplex
Appendix B	aSTR_Sequences
Appendix C	XSTR_Sequences
Appendix D	ForensicStats_ForenSeq_data
Appendix E	SNP_&_MicroHap_Calling_Scripts
Appendix F	SNP_&_MicroHap_codes
Appendix G	ii_&_si_SNP_genotypes
Appendix H	Microhaplotype_sequences
Appendix I	IdentityAnalysis_Gbb
Appendix J	IdentityAnalysis_Gbg
Appendix K	IdentityAnalysis_Gbb
Appendix L	Microhaplotype_STURCTURE_K_2
Appendix M	MLRelate_Gbb_SNPs_&_MicroHaps
Appendix N	MLRelate_Gbg_SNPs_&_MicroHaps
Appendix O	MLRelate_Ggg_SNPs_&_MicroHaps

References

ALACS, E. A., GEORGES, A., FITZSIMMONS, N. N. & ROBERTSON, J. 2010. DNA detective: a review of molecular approaches to wildlife forensics. *Forensic science, medicine, and pathology,* 6**,** 180-194.

ALLENDORF, F. W. 2017. Genetics and the conservation of natural populations: allozymes to genomes. Wiley Online Library.

ALLENDORF, F. W., HOHENLOHE, P. A. & LUIKART, G. 2010. Genomics and the future of conservation genetics. *Nature reviews genetics,* 11**,** 697-709.

AMORIM, A. & PEREIRA, L. 2005. Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic science international,* 150**,** 17-21.

ANDREWS, K. R., DE BARBA, M., RUSSELLO, M. A. & WAITS, L. P. 2018. Advances in using non-invasive, archival, and environmental samples for population genomic studies. *Population genomics: wildlife.* Springer.

APPLETON, W. L. 2020. AppMag PCR Clean Up Beads. Clean Up of Post PCR and NGS Library Construction.

ASOGAWA, M., OHNO, A., NAKAGAWA, S., OCHIAI, E., KATAHIRA, Y., SUDO, M., OSAWA, M., SUGISAWA, M. & IMANISHI, T. 2020. Human short tandem repeat identification using a nanopore-based DNA sequencer: a pilot study. *Journal of Human Genetics,* 65**,** 21-24.

AVILA, E., FELKL, A. B., GRAEBIN, P., NUNES, C. P. & ALHO, C. S. 2019. Forensic characterization of Brazilian regional populations through massive parallel sequencing of 124 SNPs included in HID ion Ampliseq Identity Panel. *Forensic Science International: Genetics,* 40**,** 74-84.

BAAS, P., VAN DER VALK, T., VIGILANT, L., NGOBOBO, U., BINYINYI, E., NISHULI, R., CAILLAUD, D. & GUSCHANSKI, K. 2018. Population-level assessment of genetic diversity and habitat fragmentation in critically endangered G rauer's gorillas. *American Journal of Physical Anthropology,* 165**,** 565-575.

BARBARA, T., PALMA-SILVA, C., PAGGI, G. M., BERED, F., FAY, M. F. & LEXER, C. 2007. Cross-species transfer of nuclear microsatellite markers: potential and limitations. *Molecular ecology,* 16**,** 3759-3767.

BEASLEY, J., SHORROCK, G., NEUMANN, R., MAY, C. A. & WETTON, J. H. 2021. Massively parallel sequencing and capillary electrophoresis of a novel panel of falcon STRs: Concordance with minisatellite DNA profiles from historical wildlife crime. *Forensic Science International: Genetics,* 54**,** 102550.

BECQUET, C., PATTERSON, N., STONE, A. C., PRZEWORSKI, M. & REICH, D. 2007. Genetic structure of chimpanzee populations. *PLoS genetics,* 3**,** e66.

BEEBEE, T. J. 2022. *Impacts of Human Population on Wildlife*, Cambridge University Press.

BELLOTT, D. W., HUGHES, J. F., SKALETSKY, H., BROWN, L. G., PYNTIKOVA, T., CHO, T.-J., KOUTSEVA, N., ZAGHLUL, S., GRAVES, T. & ROCK, S. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature,* 508**,** 494.

BERGER-TAL, O. & LAHOZ-MONFORT, J. J. 2018. Conservation technology: The next generation. *Conservation Letters,* 11**,** e12458.

BERGL, R., DUNN, A., FOWLER, A., IMONG, I., NDELOH, D., NICHOLAS, A. & OATES, J. 2016. Gorilla gorilla ssp. diehli (errata version published in 2016). *The IUCN Red List of Threatened Species 2016: e. T39998A102326240*.

BERGL, R. A., BRADLEY, B. J., NSUBUGA, A. & VIGILANT, L. 2008. Effects of habitat fragmentation, population size and demographic history on genetic diversity: the Cross River gorilla in a comparative context. *American Journal of Primatology: Official Journal*

of the American Society of Primatologists, 70**,** 848-859.
BERGL, R. A. & VIGILANT, L. 2007. Genetic analysis reveals population structure and recent migration within the highly fragmented range of the Cross River gorilla (Gorilla gorilla diehli). *Molecular Ecology,* 16**,** 501-516.
BLANCO, M. B., GREENE, L. K., RASAMBAINARIVO, F., TOOMEY, E., WILLIAMS, R. C., ANDRIANANDRASANA, L., LARSEN, P. A. & YODER, A. D. 2020. Next-generation technologies applied to age-old challenges in Madagascar. *Conservation genetics,* 21**,** 785-793.
BLANQUER-MAUMONT, A. & CROUAU-ROY, B. 1995. Polymorphism, monomorphism, and sequences in conserved microsatellites in primate species. *Journal of molecular evolution,* 41**,** 492-497.
BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.
BØRSTING, C. & MORLING, N. 2015. Next generation sequencing and its applications in forensic genetics. *Forensic Science International: Genetics,* 18**,** 78-89.
BØRSTING, C., ROCKENBAUER, E. & MORLING, N. 2009. Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard. *Forensic Science International: Genetics,* 4**,** 34-42.
BOURGEOIS, S., KADEN, J., SENN, H., BUNNEFELD, N., JEFFERY, K. J., AKOMO-OKOUE, E. F., OGDEN, R. & MCEWING, R. 2019. Improving cost-efficiency of faecal genotyping: New tools for elephant species. *PloS one,* 14**,** e0210811.
BOŽA, V., BREJOVÁ, B. & VINAŘ, T. 2017. DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PloS one,* 12**,** e0178751.
BOŽA, V., PEREŠÍNI, P., BREJOVÁ, B. & VINAŘ, T. 2021. Dynamic Pooling Improves Nanopore Base Calling Accuracy. *arXiv preprint arXiv:2105.07520*.
BRADLEY, B. J., BOESCH, C. & VIGILANT, L. 2000. Identification and redesign of human microsatellite markers for genotyping wild chimpanzee (Pan troglodytes verus) and gorilla (Gorilla gorilla gorilla) DNA from faeces. *Conservation Genetics,* 1**,** 289-292.
BRADLEY, B. J., ROBBINS, M. M., WILLIAMSON, E. A., STEKLIS, H. D., STEKLIS, N. G., ECKHARDT, N., BOESCH, C. & VIGILANT, L. 2005. Mountain gorilla tug-of-war: silverbacks have limited control over reproduction in multimale groups. *Proceedings of the National Academy of Sciences,* 102**,** 9418-9423.
BROHEDE, J. & ELLEGREN, H. 1999. Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 266**,** 825-833.
BUCHALSKI, M. R., SACKS, B. N., GILLE, D. A., PENEDO, M. C. T., ERNEST, H. B., MORRISON, S. A. & BOYCE, W. M. 2016. Phylogeographic and population genetic structure of bighorn sheep (Ovis canadensis) in North American deserts. *Journal of Mammalogy,* 97**,** 823-838.
BUDOWLE, B. & VAN DAAL, A. 2008. Forensically relevant SNP classes. *Biotechniques,* 44**,** 603-610.
BUTLER, J. M. 2007. Short tandem repeat typing technologies used in human identity testing. *Biotechniques,* 43**,** Sii-Sv.
BUTYNSKI, T. M. & KALINA, J. 1993. Three new mountain national parks for Uganda. *Oryx,* 27**,** 214-224.
CARLSSON, J., GAUTHIER, D. T., CARLSSON, J. E. L., COUGHLAN, J. P., DILLANE, E., FITZGERALD, R. D., KEATING, U., MCGINNITY, P., MIRIMIN, L. & T.F., C. 2013. Rapid, economical single-nucleotide polymorphism and microsatellite discovery based on de novo assembly of a reduced representation genomein a non-model organism: a case study of Atlantic cod Gadus morhua. . *Journal of Fish Biology,* 82**,** 944-958.
CARROLL, E. L., BRUFORD, M. W., DEWOODY, J. A., LEROY, G., STRAND, A., WAITS, L. & WANG, J.

2018. Genetic and genomic monitoring with minimally invasive sampling methods. *Evolutionary applications,* 11**,** 1094-1119.
CHANG, J. J. M., IP, Y. C. A., NG, C. S. L. & HUANG, D. 2020. Takeaways from mobile DNA barcoding with BentoLab and MinION. *Genes,* 11**,** 1121.
CHE-CASTALDO, J., GRAY, S. M., RODRIGUEZ-CLARK, K. M., SCHAD EEBES, K. & FAUST, L. J. 2021. Expected demographic and genetic declines not found in most zoo and aquarium populations. *Frontiers in Ecology and the Environment,* 19**,** 435-442.
CHURCHILL, J. D., SCHMEDES, S. E., KING, J. L. & BUDOWLE, B. 2016. Evaluation of the Illumina® beta version ForenSeq™ DNA signature prep kit for use in genetic profiling. *Forensic Science International: Genetics,* 20**,** 20-29.
CLISSON, I., LATHUILLIERE, M. & CROUAU-ROY, B. 2000. Conservation and evolution of microsatellite loci in primate taxa. *American Journal of Primatology: Official Journal of the American Society of Primatologists,* 50**,** 205-214.
COCK, P. J., FIELDS, C. J., GOTO, N., HEUER, M. L. & RICE, P. M. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research,* 38**,** 1767-1771.
COLLINS, F. S., MORGAN, M. & PATRINOS, A. 2003. The Human Genome Project: lessons from large-scale biology. *Science,* 300**,** 286-290.
CONGIU, L., PUJOLAR, J. M., FORLANI, A., CENADELLI, S., DUPANLOUP, I., BARBISAN, F., GALLI, A. & FONTANA, F. 2011. Managing polyploidy in ex situ conservation genetics: the case of the critically endangered Adriatic sturgeon (Acipenser naccarii). *PloS one,* 6**,** e18249.
COORDINATORS, N. R. 2015. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research,* 44**,** D7-D19.
COORDINATORS, N. R. 2016. Database resources of the national center for biotechnology information. *Nucleic acids research,* 44**,** D7.
CORNELIS, S., GANSEMANS, Y., DELEYE, L., DEFORCE, D. & VAN NIEUWERBURGH, F. 2017. Forensic SNP genotyping using nanopore MinION sequencing. *Scientific reports,* 7**,** 41759.
CUMMINS, J. 2001. Mitochondrial DNA and the Y chromosome: parallels and paradoxes. *Reproduction, Fertility and Development,* 13**,** 533-542.
DA SILVA, F., MINHOS, M., SA, R. & BRUFORD, M. 2012. Using genetics as a tool in primate conservation. *Nature Education Knowledge,* 3**,** 10.
DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T. & SHERRY, S. T. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158.
DASZAK, P., CUNNINGHAM, A. A. & HYATT, A. D. 2000. Emerging infectious diseases of wildlife--threats to biodiversity and human health. *science,* 287**,** 443-449.
DE FLAMINGH, A., ISHIDA, Y., PEČNEROVÁ, P., VILCHIS, S., SIEGISMUND, H. R., MALHI, R. S. & ROCA, A. L. 2023. Combining methods for non-invasive fecal DNA enables whole genome and metagenomic analyses in wildlife biology. *Frontiers in Genetics,* 13.
DE OLIVEIRA, G. L., NIEDERAUER, G. F., DE OLIVEIRA, F. A., RODRIGUES, C. S., HERNANDES, J. L., DE SOUZA, A. P. & MOURA, M. F. 2022. Genetic Diversity, Population Structure and Parentage Analysis in Brazilian Grapevine Hybrids after Half a Century of Genetic Breeding. *bioRxiv*.
DEJEAN, C. B. 2018. Genetic markers (SNPs/Satellites/STRs/RFLPs). *The International Encyclopedia of Biological Anthropology***,** 1-4.
DELAHAYE, C. & NICOLAS, J. 2021. Sequencing DNA with nanopores: Troubles and biases. *PLoS One,* 16**,** e0257521.
DESJARDINS, P. & CONKLIN, D. 2010. NanoDrop microvolume quantitation of nucleic acids. *JoVE (Journal of Visualized Experiments)***,** e2565.
DHEER, A., SAMARASINGHE, D., DLONIAK, S. M. & BRACZKOWSKI, A. 2022. Using camera traps to study hyenas: challenges, opportunities, and outlook. *Mammalian Biology***,** 1-8.

DONOHUE, I., PETCHEY, O. L., KÉFI, S., GÉNIN, A., JACKSON, A. L., YANG, Q. & O'CONNOR, N. E. 2017. Loss of predator species, not intermediate consumers, triggers rapid dramatic extinction cascades. *Global Change Biology,* 23**,** 2962-2972.

DORAN-SHEEHY, D. M., GREER, D., MONGO, P. & SCHWINDT, D. 2004. Impact of ecological and social factors on ranging in western gorillas. *American Journal of Primatology: Official Journal of the American Society of Primatologists,* 64**,** 207-222.

DOUADI, M. I., GATTI, S., LEVRÉRO, F., DUHAMEL, G., BERMEJO, M., VALLET, D., MÉNARD, N. & PETIT, E. J. 2007. Sex-biased dispersal in western lowland gorillas (Gorilla gorilla gorilla). *Molecular Ecology,* 16**,** 2247-2259.

DUPORGE, I., FINERTY, G. E., IHWAGI, F., LEE, S., WATHIKA, J., WU, Z., MACDONALD, D. W. & WANG, T. 2022. A satellite perspective on the movement decisions of African elephants in relation to nomadic pastoralists. *Remote Sensing in Ecology and Conservation*.

DUPUY, B. M., STENERSEN, M., EGELAND, T. & OLAISEN, B. 2004. Y-chromosomal microsatellite mutation rates: Differences in mutation rate between and within loci. *Human mutation,* 23**,** 117-124.

EARL, D. A. & VONHOLDT, B. M. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation genetics resources,* 4**,** 359-361.

ECKHART, L., BACH, J., BAN, J. & TSCHACHLER, E. 2000. Melanin binds reversibly to thermostable DNA polymerase and inhibits its activity. *Biochemical and biophysical research communications,* 271**,** 726-730.

EDWARDS, A., DEBBONAIRE, A. R., SATTLER, B., MUR, L. & HODSON, A. J. 2016. Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N. *BioRxiv,* 10**,** 073965.

EFFIOM, E. O., NUÑEZ-ITURRI, G., SMITH, H. G., OTTOSSON, U. & OLSSON, O. 2013. Bushmeat hunting changes regeneration of African rainforests. *Proceedings of the Royal Society B: Biological Sciences,* 280**,** 20130246.

EKBLOM, R., ARONSSON, M., ELSNER-GEARING, F., JOHANSSON, M., FOUNTAIN, T. & PERSSON, J. 2021. Sample identification and pedigree reconstruction in Wolverine (Gulo gulo) using SNP genotyping of non-invasive samples. *Conservation Genetics Resources,* 13**,** 261-274.

ELLIOTT, R. 1976. Observations on a small group of mountain gorillas (Gorilla gorilla beringei). *Folia Primatologica,* 25**,** 12-24.

ENSMINGER, A. L. & HOFFMAN, S. M. 2002. Sex identification assay useful in great apes is not diagnostic in a range of other primate species. *American Journal of Primatology: Official Journal of the American Society of Primatologists,* 56**,** 129-134.

ESTOUP, A., JARNE, P. & CORNUET, J. M. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular ecology,* 11**,** 1591-1604.

ESTRADA, A., GARBER, P. A., RYLANDS, A. B., ROOS, C., FERNANDEZ-DUQUE, E., DI FIORE, A., NEKARIS, K. A.-I., NIJMAN, V., HEYMANN, E. W. & LAMBERT, J. E. 2017. Impending extinction crisis of the world's primates: Why primates matter. *Science advances,* 3**,** e1600946.

EVANNO, G., REGNAUT, S. & GOUDET, J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology,* 14**,** 2611-2620.

EVANS, M. J. & RITTENHOUSE, T. A. 2018. Evaluating spatially explicit density estimates of unmarked wildlife detected by remote cameras. *Journal of applied ecology,* 55**,** 2565-2574.

FAN, H. & CHU, J.-Y. 2007. A brief review of short tandem repeat mutation. *Genomics, Proteomics & Bioinformatics,* 5**,** 7-14.

FEDELE, E., WETTON, J. H. & JOBLING, M. A. 2022. Sequencing the orthologs of human autosomal forensic short tandem repeats provides individual-and species-level

identification in African great apes. *bioRxiv*.
FERNANDEZ, F. 2011. The greatest impediment to the study of biodiversity in Colombia. *Caldasia*, iii-v.
FITZ, J., ADENLE, A. A. & SPERANZA, C. I. 2022. Increasing signs of forest fragmentation in the Cross River National Park in Nigeria: Underlying drivers and need for sustainable responses. *Ecological indicators,* 139, 108943.
FITZSIMMONS, N. N., MORITZ, C. & MOORE, S. S. 1995. Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. *Molecular biology and evolution,* 12, 432-440.
FORCINA, G., VALLET, D., LE GOUAR, P. J., BERNARDO-MADRID, R., ILLERA, G., MOLINA-VACAS, G., DRÉANO, S., REVILLA, E., RODRÍGUEZ-TEIJEIRO, J. D. & MÉNARD, N. 2019. From groups to communities in western lowland gorillas. *Proceedings of the Royal Society B,* 286, 20182019.
FRANKHAM, R. 2003. Genetics and conservation biology. *Comptes Rendus Biologies,* 326, 22-29.
FRANKHAM, R. 2010. Where are we in conservation genetics and where do we need to go? *Conservation genetics,* 11, 661-663.
FRANKHAM, R., BALLOU, J. D., RALLS, K., ELDRIDGE, M., DUDASH, M. R., FENSTER, C. B., LACY, R. C. & SUNNUCKS, P. 2017. *Genetic management of fragmented animal and plant populations*, Oxford University Press.
FUMAGALLI, M. 2013. Assessing the effect of sequencing depth and sample size in population genetics inferences. *PloS one,* 8, e79667.
GATTI, S., LEVRÉRO, F., MÉNARD, N. & GAUTIER-HION, A. 2004. Population and group structure of western lowland gorillas (Gorilla gorilla gorilla) at Lokoue, Republic of Congo. *American Journal of Primatology: Official Journal of the American Society of Primatologists,* 63, 111-123.
GATTI, S., LEVRÉRO, F., MÉNARD, N., PETIT, E. & GAUTIER-HION, A. 2003. Bachelor groups of western lowland gorillas (Gorilla gorilla gorilla) at Lokoué Clearing, Odzala National Park, Republic of Congo. *Folia Primatol,* 74, 195-196.
GAVRILIUC, S., REZA, S., JEONG, C., GETACHEW, F., MCLOUGHLIN, P. D. & POISSANT, J. 2022. Targeted genome-wide SNP genotyping in feral horses using non-invasive fecal swabs. *Conservation Genetics Resources*, 1-11.
GETTINGS, K. B., KIESLER, K. M., FAITH, S. A., MONTANO, E., BAKER, C. H., YOUNG, B. A., GUERRIERI, R. A. & VALLONE, P. M. 2016. Sequence variation of 22 autosomal STR loci detected by next generation sequencing. *Forensic Science International: Genetics,* 21, 15-21.
GILARDI, K. V., GILLESPIE, T. R., LEENDERTZ, F. H., MACFIE, E. J., TRAVIS, D. A., WHITTIER, C. A. & WILLIAMSON, E. A. 2015. Best practice guidelines for health monitoring and disease control in great ape populations. *Occasional Papers of the IUCN Species Survival Commission,* 56.
GILBERT, N. 2010. Biodiversity law could stymie research. *Nature,* 463, 598.
GILL, P. 2001. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *International journal of legal medicine,* 114, 204-210.
GONDHALI, U. & MISHRA, A. 2022. Wildlife DNA Profiling and Its Forensic Relevance. *Handbook of DNA Profiling*, 823.
GOOSSENS, B. & BRUFORD, M. W. 2009. Non-invasive genetic analysis in conservation. *Population genetics for animal conservation. Cambridge University Press, Cambridge*, 167-201.
GOOSSENS, B., WAITS, L. P. & TABERLET, P. 1998. Plucked hair samples as a source of DNA: reliability of dinucleotide microsatellite genotyping. *Molecular Ecology,* 7, 1237-1241.
GORDON, D., HUDDLESTON, J., CHAISSON, M. J., HILL, C. M., KRONENBERG, Z. N., MUNSON, K. M., MALIG, M., RAJA, A., FIDDES, I. & HILLIER, L. W. 2016. Long-read sequence assembly of the gorilla genome. *Science,* 352, aae0344.

GOROSPE, K. D., DONAHUE, M. J. & KARL, S. A. 2015. The importance of sampling design: spatial patterns and clonality in estimating the genetic diversity of coral reefs. *Marine Biology,* 162**,** 917-928.
GOTELLI, N. J. & COLWELL, R. K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology letters,* 4**,** 379-391.
GOUY, A. & ZIEGER, M. 2017. STRAF—a convenient online tool for STR data evaluation in forensic genetics. *Forensic Science International: Genetics,* 30**,** 148-151.
GRÄDEL, C., TERRAZOS MIANI, M. A., BARBANI, M. T., LEIB, S. L., SUTER-RINIKER, F. & RAMETTE, A. 2019. Rapid and cost-efficient enterovirus genotyping from clinical samples using flongle flow cells. *Genes,* 10**,** 659.
GRAY, M., ROY, J., VIGILANT, L., FAWCETT, K., BASABOSE, A., CRANFIELD, M., UWINGELI, P., MBURANUMWE, I., KAGODA, E. & ROBBINS, M. M. 2013. Genetic census reveals increased but uneven growth of a critically endangered mountain gorilla population. *Biological Conservation,* 158**,** 230-238.
GROVES, C. P. 2001. Primate taxonomy.
GRUBB, P., BUTYNSKI, T. M., OATES, J. F., BEARDER, S. K., DISOTELL, T. R., GROVES, C. P. & STRUHSAKER, T. T. 2003. Assessment of the diversity of African primates. *International Journal of Primatology,* 24**,** 1301-1357.
GRUBER, B., UNMACK, P. J., BERRY, O. F. & GEORGES, A. 2018. dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular ecology resources,* 18**,** 691-699.
GUEUNING, M., GANSER, D., BLASER, S., ALBRECHT, M., KNOP, E., PRAZ, C. & FREY, J. E. 2019. Evaluating next-generation sequencing (NGS) methods for routine monitoring of wild bees: Metabarcoding, mitogenomics or NGS barcoding. *Molecular ecology resources*.
GUGERLI, F., BRODBECK, S. & HOLDEREGGER, R. 2008. Insertions–deletions in a microsatellite flanking region may be resolved by variation in stuttering patterns. *Plant Molecular Biology Reporter,* 26**,** 255-262.
GULSBY, W. D., KILLMASTER, C. H., BOWERS, J. W., LAUFENBERG, J. S., SACKS, B. N., STATHAM, M. J. & MILLER, K. V. 2016. Efficacy and precision of fecal genotyping to estimate coyote abundance. *Wildlife Society Bulletin,* 40**,** 792-799.
GUSCHANSKI, K., VIGILANT, L., MCNEILAGE, A., GRAY, M., KAGODA, E. & ROBBINS, M. M. 2009. Counting elusive animals: comparing field and genetic census of the entire mountain gorilla population of Bwindi Impenetrable National Park, Uganda. *Biological Conservation,* 142**,** 290-300.
GYMREK, M., GOLAN, D., ROSSET, S. & ERLICH, Y. 2012. lobSTR: a short tandem repeat profiler for personal genomes. *Genome research,* 22**,** 1154-1162.
HAGEMANN, L., BOESCH, C., ROBBINS, M. M., ARANDJELOVIC, M., DESCHNER, T., LEWIS, M., FROESE, G. & VIGILANT, L. 2018. Long-term group membership and dynamics in a wild western lowland gorilla population (Gorilla gorilla gorilla) inferred using non-invasive genetics. *American Journal of Primatology,* 80**,** e22898.
HAILEMARIAM, Z., AHMED, J. S., CLAUSEN, P.-H. & NIJHOF, A. M. 2017. A comparison of DNA extraction protocols from blood spotted on FTA cards for the detection of tick-borne pathogens by Reverse Line Blot hybridization. *Ticks and tick-borne diseases,* 8**,** 185-189.
HALL, C., ZASCAVAGE, R., SEDLAZECK, F. & PLANZ, J. 2020. Potential applications of nanopore sequencing for forensic analysis. *Forensic science review,* 32**,** 23-54.
HALL, C. L., KESHARWANI, R. K., PHILLIPS, N. R., PLANZ, J. V., SEDLAZECK, F. J. & ZASCAVAGE, R. R. 2022. Accurate profiling of forensic autosomal STRs using the Oxford Nanopore Technologies MinION device. *Forensic Science International: Genetics,* 56**,** 102629.
HALLAST, P. & JOBLING, M. A. 2017. The Y chromosomes of the great apes. *Human genetics,* 136**,** 511-528.
HAUREZ, B., TAGG, N., PETRE, C. A., BROSTAUX, Y., BOUBADY, A. & DOUCET, J. L. 2018. Seed dispersal effectiveness of the western lowland gorilla (Gorilla gorilla gorilla) in Gabon.

African Journal of Ecology, 56**,** 185-193.
HEBERT, P. D., RATNASINGHAM, S. & DE WAARD, J. R. 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 270**,** S96-S99.
HENEGARIU, O., HEEREMA, N., DLOUHY, S., VANCE, G. & VOGT, P. 1997. Multiplex PCR: critical parameters and step-by-step protocol. *Biotechniques,* 23**,** 504-511.
HICKEY, J., BASABOSE, A., GILARDI, K., GREER, D., NAMPINDO, S., ROBBINS, M. & STOINSKI, T. 2018. Gorilla beringei ssp. beringei. *The IUCN Red List of Threatened Species*.
HOHENLOHE, P. A., FUNK, W. C. & RAJORA, O. P. 2021. Population genomics for wildlife conservation and management. *Molecular Ecology,* 30**,** 62-82.
HOOGENBOOM, J., VAN DER GAAG, K. J., DE LEEUW, R. H., SIJEN, T., DE KNIJFF, P. & LAROS, J. F. 2017. FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise. *Forensic Science International: Genetics,* 27**,** 27-40.
HOOPER, D. U., CHAPIN, F. S., EWEL, J. J., HECTOR, A., INCHAUSTI, P., LAVOREL, S., LAWTON, J. H., LODGE, D., LOREAU, M. & NAEEM, S. 2005. Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecological monographs,* 75**,** 3-35.
HÖSS, M., KOHN, M., PÄÄBO, S., KNAUER, F. & SCHRÖDER, W. 1992. Excrement analysis by PCR. *Nature,* 359**,** 199.
HU, X.-D. & GAO, L.-Z. 2016. The complete mitochondrial genome of eastern lowland gorilla, Gorilla beringei graueri, and comparative mitochondrial genomics of Gorilla species. *Mitochondrial DNA Part A,* 27**,** 1484-1485.
HUANG, Q.-Y., XU, F.-H., SHEN, H., DENG, H.-Y., LIU, Y.-J., LIU, Y.-Z., LI, J.-L., RECKER, R. R. & DENG, H.-W. 2002. Mutation patterns at dinucleotide microsatellite loci in humans. *The American Journal of Human Genetics,* 70**,** 625-634.
HUFFMEYER, A. A., SIKICH, J. A., VICKERS, T. W., RILEY, S. P. & WAYNE, R. K. 2022. First reproductive signs of inbreeding depression in Southern California male mountain lions (Puma concolor). *Theriogenology,* 177**,** 157-164.
HUSZAR, T. I., JOBLING, M. A. & WETTON, J. H. 2018. A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing. *Forensic Science International: Genetics,* 35**,** 97-106.
HVILSOM, C., CARLSEN, F., HELLER, R., JAFFRÉ, N. & SIEGISMUND, H. R. 2014. Contrasting demographic histories of the neighboring bonobo and chimpanzee. *Primates,* 55**,** 101-112.
ILLUMINA 2015a. DNA Signature Prep Reference Guide. Document.
ILLUMINA, I. 2015b. An introduction to next-generation sequencing technology.
ISAAC, N. J. & COWLISHAW, G. 2004. How species respond to multiple extinction threats. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 271**,** 1135-1141.
JÄGER, A. C., ALVAREZ, M. L., DAVIS, C. P., GUZMÁN, E., HAN, Y., WAY, L., WALICHIEWICZ, P., SILVA, D., PHAM, N. & CAVES, G. 2017. Developmental validation of the MiSeq FGx forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories. *Forensic Science International: Genetics,* 28**,** 52-70.
JEFFERY, K. J., ABERNETHY, K. A., TUTIN, C. E., ANTHONY, N. A. & BRUFORD, M. W. 2007a. Who killed Porthos? Genetic tracking of a gorilla death. *Integrative Zoology,* 2**,** 111-119.
JEFFERY, K. J., ABERNETHY, K. A., TUTIN, C. E. & BRUFORD, M. W. 2007b. Biological and environmental degradation of gorilla hair and microsatellite amplification success. *Biological Journal of the Linnean Society,* 91**,** 281-294.
JINNAH, S. & JUNGCURT, S. 2009. Could access requirements stifle your research? *Science,* 323**,** 464-465.
JOBLING, M., HURLES, M. & TYLER-SMITH, C. 2013. *Human evolutionary genetics: origins, peoples & disease*, Garland Science.

JOBLING, M. A. & GILL, P. 2004. Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics,* 5**,** 739-751.
JOBLING, M. A., HURLES, M. & TYLER-SMITH, C. 2019. *Human evolutionary genetics: origins, peoples and disease*, Garland Science.
JOBLING, M. A. & TYLER-SMITH, C. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews Genetics,* 4**,** 598.
JOHNSON, R. N., WILSON-WILDE, L. & LINACRE, A. 2014. Current and future directions of DNA in wildlife forensic science. *Forensic Science International: Genetics,* 10**,** 1-11.
JOMBART, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics,* 24**,** 1403-1405.
JOMBART, T. & AHMED, I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics,* 27**,** 3070-3071.
JOMBART, T., DEVILLARD, S. & BALLOUX, F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics,* 11**,** 1-15.
JUST, R. S., MORENO, L. I., SMERICK, J. B. & IRWIN, J. A. 2017. Performance and concordance of the ForenSeq™ system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens. *Forensic Science International: Genetics,* 28**,** 1-9.
KALINOWSKI, S. T., TAPER, M. L. & MARSHALL, T. C. 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular ecology,* 16**,** 1099-1106.
KALINOWSKI, S. T., WAGNER, A. P. & TAPER, M. L. 2006. ML-Relate: a computer program for maximum likelihood estimation of relatedness and relationship. *Molecular Ecology Notes,* 6**,** 576-579.
KAYITETE, L., VAN DER HOEK, Y., NYIRAMBANGUTSE, B. & DERHÉ, M. A. 2019. Observations on regeneration of the keystone plant species Hagenia abyssinica in Volcanoes National Park, Rwanda. *African Journal of Ecology*.
KEESING, F. & OSTFELD, R. S. 2021. Impacts of biodiversity and biodiversity loss on zoonotic diseases. *Proceedings of the National Academy of Sciences,* 118**,** e2023540118.
KELLER, L. F. & WALLER, D. M. 2002. Inbreeding effects in wild populations. *Trends in ecology & evolution,* 17**,** 230-241.
KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome research,* 12**,** 996-1006.
KHAN, A., PATEL, K., BHATTACHARJEE, S., SHARMA, S., CHUGANI, A. N., SIVARAMAN, K., HOSAWAD, V., SAHU, Y. K., REDDY, G. V. & RAMAKRISHNAN, U. 2020. Are shed hair genomes the most effective noninvasive resource for estimating relationships in the wild? *Ecology and Evolution,* 10**,** 4583-4594.
KHUBRANI, Y. M., HALLAST, P., JOBLING, M. A. & WETTON, J. H. 2019. Massively parallel sequencing of autosomal STRs and identity-informative SNPs highlights consanguinity in Saudi Arabia. *Forensic Science International: Genetics,* 43**,** 102164.
KIDD, K., PAKSTIS, A., SPEED, W., LAGACE, R., CHANG, J., WOOTTON, S. & IHUEGBU, N. 2013. Microhaplotype loci are a powerful new type of forensic marker. *Forensic Science International: Genetics Supplement Series,* 4**,** e123-e124.
KIDD, K. K., PAKSTIS, A. J., SPEED, W. C., GRIGORENKO, E. L., KAJUNA, S. L., KAROMA, N. J., KUNGULILO, S., KIM, J.-J., LU, R.-B. & ODUNSI, A. 2006. Developing a SNP panel for forensic identification of individuals. *Forensic science international,* 164**,** 20-32.
KNAUS, B. J. & GRÜNWALD, N. J. 2017. vcfr: a package to manipulate and visualize variant call format data in R. *Molecular ecology resources,* 17**,** 44-53.
KOWALCZYK, M., STANISZEWSKI, A., KAMIŃSKA, K., DOMARADZKI, P. & HORECKA, B. 2021. Advantages, Possibilities, and Limitations of Mitochondrial DNA Analysis in Molecular

Identification. *Folia Biologica (Kraków),* 69**,** 101-111.
KREHENWINKEL, H., POMERANTZ, A. & PROST, S. 2019. Genetic biomonitoring and biodiversity assessment using portable sequencing technologies: current uses and future directions. *Genes,* 10**,** 858.
KRESS, W. J. & ERICKSON, D. L. 2012. DNA barcodes: methods and protocols. *DNA Barcodes.* Springer.
KÜHL, H. 2008. *Best practice guidelines for the surveys and monitoring of great ape populations*, IUCN.
KUHLWILM, M., DE MANUEL, M., NATER, A., GREMINGER, M. P., KRÜTZEN, M. & MARQUES-BONET, T. 2016. Evolution and demography of the great apes. *Current opinion in genetics & development,* 41**,** 124-129.
KUMAR, V. P., KUMAR, D. & GOYAL, S. P. 2014. Wildlife DNA forensic in curbing illegal wildlife trade: species identification from seizures. *Int J Forensic Sci Pathol,* 2**,** 38-42.
KWONG, M. & PEMBERTON, T. J. 2014. Sequence differences at orthologous microsatellites inflate estimates of human-chimpanzee differentiation. *BMC genomics,* 15**,** 990.
LACHANCE, J. & TISHKOFF, S. A. 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays,* 35**,** 780-786.
LATORRE-CARDENAS, M. C., GUTIÉRREZ-RODRÍGUEZ, C. & LANCE, S. L. 2020. Isolation and characterization of 13 microsatellite loci for the Neotropical otter, Lontra longicaudis, by next generation sequencing. *Molecular Biology Reports,* 47**,** 731-736.
LEVRÉRO, F., GATTI, S., MÉNARD, N., PETIT, E., CAILLAUD, D. & GAUTIER-HION, A. 2006. Living in nonbreeding groups: an alternative strategy for maturing gorillas. *American Journal of Primatology: Official Journal of the American Society of Primatologists,* 68**,** 275-291.
LI, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics,* 27**,** 2987-2993.
LI, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics,* 34**,** 3094-3100.
LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics,* 25**,** 1754-1760.
LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.
LINACRE, A. 2021. Animal forensic genetics. *Genes,* 12**,** 515.
LING, C., LIXIA, W., RONG, H., FUJUN, S., WENPING, Z., YAO, T., YAOHUA, Y., BO, Z. & LIANG, Z. 2020. Comparative analysis of microsatellite and SNP markers for parentage testing in the golden snub-nosed monkey (Rhinopithecus roxellana). *Conservation Genetics Resources,* 12**,** 611-620.
LOBON, I., TUCCI, S., DE MANUEL, M., GHIROTTO, S., BENAZZO, A., PRADO-MARTINEZ, J., LORENTE-GALDOS, B., NAM, K., DABAD, M. & HERNANDEZ-RODRIGUEZ, J. 2016. Demographic history of the genus Pan inferred from whole mitochondrial genome reconstructions. *Genome biology and evolution,* 8**,** 2020-2030.
LOHMAN, G. 2018. Substrate specificity and mismatch discrimination in DNA ligases.
LUO, R., WONG, C.-L., WONG, Y.-S., TANG, C.-I., LIU, C.-M., LEUNG, C.-M. & LAM, T.-W. 2020. Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nature Machine Intelligence,* 2**,** 220-227.
LYNCH, M. & RITLAND, K. 1999. Estimation of pairwise relatedness with molecular markers. *Genetics,* 152**,** 1753-1766.
MADUNA, S. N., ROSSOUW, C., ROODT-WILDING, R. & BESTER-VAN DER MERWE, A. E. 2014. Microsatellite cross-species amplification and utility in southern African elasmobranchs: a valuable resource for fisheries management and conservation. *BMC research notes,* 7**,** 1-13.

MAGLIOCCA, F., QUEROUIL, S. & GAUTIER-HION, A. 1999. Population structure and group composition of western lowland gorillas in North-Western Republic of Congos. *American Journal of Primatology: Official Journal of the American Society of Primatologists,* 48**,** 1-14.
MAISELS, F., STRINDBERG, S., BREUER, T., GREER, D., JEFFERY, K. & STOKES, E. 2017. Gorilla gorilla ssp. gorilla (amended version of 2016 assessment).
MALDONADO, O., AVELING, C., COX, D., NIXON, S., NISHULI, R., MERLO, D., PINTEA, L. & WILLIAMSON, E. A. 2012. *Grauer's Gorillas and Chimpanzees in Eastern Democratic Republic of Congo (Kahuzi-Biega, Maiko, Tayna and Itombwe Landscape): Conservation Action Plan 2012–2022*, IUCN.
MANDL, I., HOUMADI, A., SAID, I., ABDOU, B. B. A., FARDANE, A.-K., EGGER-PEITLER, K., OLEKSY, R., DOULTON, H. & CHAIHANE, S. S. A. 2022. Using GPS tracking for fruit bat conservation. *Oryx,* 56**,** 50-53.
MANGUETTE, M. L., ROBBINS, A. M., BREUER, T., STOKES, E. J., PARNELL, R. J. & ROBBINS, M. M. 2020. Female dispersal patterns influenced by male tenure duration and group size in western lowland gorillas. *Behavioral Ecology and Sociobiology,* 74**,** 1-15.
MARQUES-BONET, T. & HVILSOM, C. 2018. Genomic variation of the great apes and the application to conservation. *International Zoo Yearbook,* 52**,** 25-33.
MARSHALL, T., SLATE, J., KRUUK, L. & PEMBERTON, J. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular ecology,* 7**,** 639-655.
MASTRETTA-YANES, A., ARRIGO, N., ALVAREZ, N., JORGENSEN, T. H., PIÑERO, D. & EMERSON, B. C. 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular ecology resources,* 15**,** 28-41.
MATSUDAIRA, K., REICHARD, U. H., ISHIDA, T. & MALAIVIJITNOND, S. 2022. Introgression and mating patterns between white-handed gibbons (Hylobates lar) and pileated gibbons (Hylobates pileatus) in a natural hybrid zone. *PloS one,* 17**,** e0264519.
MAXWELL, S. L., BUTT, N., MARON, M., MCALPINE, C. A., CHAPMAN, S., ULLMANN, A., SEGAN, D. B. & WATSON, J. E. 2019. Conservation implications of ecological responses to extreme weather and climate events. *Diversity and Distributions,* 25**,** 613-625.
MAXWELL, S. L., FULLER, R. A., BROOKS, T. M. & WATSON, J. E. 2016. Biodiversity: The ravages of guns, nets and bulldozers. *Nature News,* 536**,** 143.
MCDONALD, T. L., ZHOU, W., CASTRO, C. P., MUMM, C., SWITZENBERG, J. A., MILLS, R. E. & BOYLE, A. P. 2021. Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nature communications,* 12**,** 1-13.
MCGRATH, K., ERIKSEN, A. B., GARCÍA-MARTÍNEZ, D., GALBANY, J., GÓMEZ-ROBLES, A., MASSEY, J. S., FATICA, L. M., GLOWACKA, H., ARBENZ-SMITH, K. & MUVUNYI, R. 2022. Facial asymmetry tracks genetic diversity among Gorilla subspecies. *Proceedings of the Royal Society B,* 289**,** 20212564.
MCTAVISH, E. J. & HILLIS, D. M. 2015. How do SNP ascertainment schemes and population demographics affect inferences about population history? *BMC genomics,* 16**,** 1-13.
MENEGON, M., CANTALONI, C., RODRIGUEZ-PRIETO, A., CENTOMO, C., ABDELFATTAH, A., ROSSATO, M., BERNARDI, M., XUMERLE, L., LOADER, S. & DELLEDONNE, M. 2017. On site DNA barcoding by nanopore sequencing. *PLoS One,* 12**,** e0184741.
METSIO SIENNE, J., BUCHWALD, R. & WITTEMYER, G. 2014. Plant mineral concentrations related to foraging preferences of western lowland gorilla in central African forest clearings. *American Journal of Primatology,* 76**,** 1115-1126.
MILES, L. G., ISBERG, S. R., GLENN, T. C., LANCE, S. L., DALZELL, P., THOMSON, P. C. & MORAN, C. 2009. A genetic linkage map for the saltwater crocodile (Crocodylus porosus). *Bmc Genomics,* 10**,** 1-11.
MORI, E., FEDELE, E., GRECO, I., GIAMPAOLI RUSTICHELLI, M., MASSOLO, A., MINIATI, S., PUPPO, F., SANTINI, G. & ZACCARONI, M. 2022. Spatiotemporal activity of the pine marten

Martes martes: Insights from an island population. *Ecological Research,* 37**,** 102-114.
MORIN, P. A., CHAMBERS, K. E., BOESCH, C. & VIGILANT, L. 2001. Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (Pan troglodytes verus). *Molecular ecology,* 10**,** 1835-1844.
MORIN, P. A., LUIKART, G. & WAYNE, R. K. 2004. SNPs in ecology, evolution and conservation. *Trends in ecology & evolution,* 19**,** 208-216.
MORIN, P. A., MOORE, J. J., CHAKRABORTY, R., JIN, L., GOODALL, J. & WOODRUFF, D. S. 1994. Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science,* 265**,** 1193-1201.
MORIN, P. A., WALLIS, J., MOORE, J. J., CHAKRABORTY, R. & WOODRUFF, D. S. 1993. Non-invasive sampling and DNA amplification for paternity exclusion, community structure, and phylogeography in wild chimpanzees. *Primates,* 34**,** 347-356.
MUNSHI-SOUTH, J., ZOLNIK, C. P. & HARRIS, S. E. 2016. Population genomics of the Anthropocene: Urbanization is negatively associated with genome-wide variation in white-footed mouse populations. *Evolutionary applications,* 9**,** 546-564.
NAPIERALSKI, A. & NOWAK, R. 2022. Basecalling Using Joint Raw and Event Nanopore Data Sequence-to-Sequence Processing. *Sensors,* 22**,** 2275.
NATESH, M., TAYLOR, R. W., TRUELOVE, N. K., HADLY, E. A., PALUMBI, S. R., PETROV, D. A. & RAMAKRISHNAN, U. 2019. Empowering conservation practice with efficient and economical genotyping from poor quality samples. *Methods in Ecology and Evolution,* 10**,** 853-859.
NAVIDI, W., ARNHEIM, N. & WATERMAN, M. 1992. A multiple-tubes approach for accurate genotyping of very small DNA samples by using PCR: statistical considerations. *American journal of human genetics,* 50**,** 347.
NIELSEN, R., PAUL, J. S., ALBRECHTSEN, A. & SONG, Y. S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics,* 12**,** 443.
NKURUNUNGI, J. B., GANAS, J., ROBBINS, M. M. & STANFORD, C. B. 2004. A comparison of two mountain gorilla habitats in Bwindi Impenetrable National Park, Uganda. *African Journal of Ecology,* 42**,** 289-297.
NOVROSKI, N. M. 2021. Exploring new short tandem repeat markers for DNA mixture deconvolution. *Wiley Interdisciplinary Reviews: Forensic Science,* 3**,** e1390.
NSUBUGA, A. M., ROBBINS, M. M., ROEDER, A. D., MORIN, P. A., BOESCH, C. & VIGILANT, L. 2004. Factors affecting the amount of genomic DNA extracted from ape faeces and the identification of an improved sample storage method. *Molecular ecology,* 13**,** 2089-2094.
OGDEN, R. 2011. Unlocking the potential of genomic technologies for wildlife forensics. *Molecular ecology resources,* 11**,** 109-116.
OGDEN, R., DAWNAY, N. & MCEWING, R. 2009. Wildlife DNA forensics—bridging the gap between conservation genetics and law enforcement. *Endangered Species Research,* 9**,** 179-195.
OHTA, T. & KIMURA, M. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetics Research,* 22**,** 201-204.
OLDONI, F., KIDD, K. K. & PODINI, D. 2019. Microhaplotypes in forensic genetics. *Forensic Science International: Genetics,* 38**,** 54-69.
OUBORG, N., VERGEER, P. & MIX, C. 2006. The rough edges of the conservation genetics paradigm for plants. *Journal of Ecology,* 94**,** 1233-1248.
OUBORG, N. J., PERTOLDI, C., LOESCHCKE, V., BIJLSMA, R. K. & HEDRICK, P. W. 2010. Conservation genetics in transition to conservation genomics. *Trends in genetics,* 26**,** 177-187.
PAKSTIS, A. J., SPEED, W. C., FANG, R., HYLAND, F. C., FURTADO, M. R., KIDD, J. R. & KIDD, K. K.

2010. SNPs for a universal individual identification panel. *Human genetics,* 127**,** 315-324.
PARNELL, R. J. 2002. Group size and structure in western lowland gorillas (Gorilla gorilla gorilla) at Mbeli Bai, Republic of Congo. *American Journal of Primatology: Official Journal of the American Society of Primatologists,* 56**,** 193-206.
PEMBERTON, T. J., ABSHER, D., FELDMAN, M. W., MYERS, R. M., ROSENBERG, N. A. & LI, J. Z. 2012. Genomic patterns of homozygosity in worldwide human populations. *The American Journal of Human Genetics,* 91**,** 275-292.
PERTOLDI, C. & RANDI, E. 2018. The ongoing transition at an exponential speed from Conservation genetics to Conservation genomics. *Genetics & Biodiversity Journal,* 2**,** 47-54.
PETTORELLI, N., BARLOW, J., NUÑEZ, M. A., RADER, R., STEPHENS, P. A., PINFIELD, T. & NEWTON, E. 2021. How international journals can support ecology from the Global South.
PFEIFFER, I. & BRENIG, B. 2005. X-and Y-chromosome specific variants of the amelogenin gene allow sex determination in sheep (Ovis aries) and European red deer (Cervus elaphus). *BMC genetics,* 6**,** 1-4.
PIJANOWSKI, B. C., FARINA, A., GAGE, S. H., DUMYAHN, S. L. & KRAUSE, B. L. 2011. What is soundscape ecology? An introduction and overview of an emerging new science. *Landscape ecology,* 26**,** 1213-1232.
PIMM, S. L., JENKINS, C. N., ABELL, R., BROOKS, T. M., GITTLEMAN, J. L., JOPPA, L. N., RAVEN, P. H., ROBERTS, C. M. & SEXTON, J. O. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. *Science,* 344**,** 1246752.
PIRES, S. F. & OLAH, G. 2022. Wildlife Crime: Issues and Promising Solutions. MDPI.
PLESIVKOVA, D., RICHARDS, R. & HARBISON, S. 2019. A review of the potential of the MinION™ single-molecule sequencing system for forensic applications. *Wiley Interdisciplinary Reviews: Forensic Science,* 1**,** e1323.
PLUMPTRE, A., NIXON, S., CAILLAUD, D., HALL, J., HART, J., NISHULI, R. & WILLIAMSON, E. 2016a. Gorilla beringei ssp. graueri (errata version published in 2016). *The IUCN Red List of Threatened Species,* 2016**,** e. T39995A102328430.
PLUMPTRE, A. J., NIXON, S., KUJIRAKWINJA, D. K., VIEILLEDENT, G., CRITCHLOW, R., WILLIAMSON, E. A., NISHULI, R., KIRKBY, A. E. & HALL, J. S. 2016b. Catastrophic decline of world's largest primate: 80% loss of Grauer's Gorilla (Gorilla beringei graueri) population justifies critically endangered status. *PloS one,* 11**,** e0162697.
POMERANTZ, A., PENAFIEL, N., ARTEAGA, A., BUSTAMANTE, L., PICHARDO, F., COLOMA, L. A., BARRIO-AMOROS, C. L., SALAZAR-VALENZUELA, D. & PROST, S. 2017. Real-time DNA barcoding in a remote rainforest using nanopore sequencing. *bioRxiv***,** 189159.
POMERANTZ, A., PEÑAFIEL, N., ARTEAGA, A., BUSTAMANTE, L., PICHARDO, F., COLOMA, L. A., BARRIO-AMORÓS, C. L., SALAZAR-VALENZUELA, D. & PROST, S. 2018. Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *GigaScience,* 7**,** giy033.
POMERANTZ, A., SAHLIN, K., VASILJEVIC, N., SEAH, A., LIM, M., HUMBLE, E., KENNEDY, S., KREHENWINKEL, H., WINTER, S. & OGDEN, R. 2022. Rapid in situ identification of biological specimens via DNA amplicon sequencing using miniaturized laboratory equipment. *Nature Protocols***,** 1-29.
PRADO-MARTINEZ, J., SUDMANT, P. H., KIDD, J. M., LI, H., KELLEY, J. L., LORENTE-GALDOS, B., VEERAMAH, K. R., WOERNER, A. E., O'CONNOR, T. D. & SANTPERE, G. 2013. Great ape genetic diversity and population history. *Nature,* 499**,** 471-475.
PRIMMER, C. R., N. PAINTER, J., T. KOSKINEN, M., U. PALO, J. & MERILÄ, J. 2005. Factors affecting avian cross-species microsatellite amplification. *Journal of Avian Biology,* 36**,** 348-360.
PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. 2000. Inference of population structure using

multilocus genotype data. *Genetics,* 155**,** 945-959.
PRÜFER, K., MUNCH, K., HELLMANN, I., AKAGI, K., MILLER, J. R., WALENZ, B., KOREN, S., SUTTON, G., KODIRA, C. & WINER, R. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature,* 486**,** 527-531.
PURISOTAYO, T., JONSSON, N. N., MABLE, B. K. & VERREYNNE, F. J. 2019. Combining molecular and incomplete observational data to inform management of southern white rhinoceros (Ceratotherium simum simum). *Conservation Genetics,* 20**,** 639-652.
QIAAMP. 2010a. *Investigator Hanbook*. Hilden, Germany patent application.
QIAAMP. 2010b. *QIAamp DNA Stool Mini Kit Handbook*. Hilden, Germany patent application.
QIAGEN 2009. Type-it® Microsatellite PCR Handbook.
QUICK, J., LOMAN, N. J., DURAFFOUR, S., SIMPSON, J. T., SEVERI, E., COWLEY, L., BORE, J. A., KOUNDOUNO, R., DUDAS, G. & MIKHAIL, A. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature,* 530**,** 228-232.
RAZGOUR, O., TAGGART, J. B., MANEL, S., JUSTE, J., IBANEZ, C., REBELO, H., ALBERDI, A., JONES, G. & PARK, K. 2018. An integrated framework to identify wildlife populations under threat from climate change. *Molecular ecology resources,* 18**,** 18-31.
REICH, D. E., SCHAFFNER, S. F., DALY, M. J., MCVEAN, G., MULLIKIN, J. C., HIGGINS, J. M., RICHTER, D. J., LANDER, E. S. & ALTSHULER, D. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature genetics,* 32**,** 135-142.
REN, Z.-L., ZHANG, J.-R., ZHANG, X.-M., LIU, X., LIN, Y.-F., BAI, H., WANG, M.-C., CHENG, F., LIU, J.-D. & LI, P. 2021. Forensic nanopore sequencing of STRs and SNPs using Verogen's ForenSeq DNA signature prep kit and MinION. *International journal of legal medicine,* 135**,** 1685-1693.
RHODES, M. W., BENNIE, J. J., SPALDING, A., FFRENCH-CONSTANT, R. H. & MACLEAN, I. M. 2022. Recent advances in the remote sensing of insects. *Biological Reviews,* 97**,** 343-360.
RITLAND, K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetics Research,* 67**,** 175-185.
RIVERS, M. C., BRUMMITT, N. A., LUGHADHA, E. N. & MEAGHER, T. R. 2014. Do species conservation assessments capture genetic diversity? *Global ecology and conservation,* 2**,** 81-87.
RIZKALLA, C., BLANCO-SILVA, F. & GRUVER, S. 2007. Modeling the impact of Ebola and bushmeat hunting on Western Lowland Gorillas. *EcoHealth,* 4**,** 151-155.
ROBBINS, A. M. & ROBBINS, M. M. 2015. Dispersal patterns of females in the genus Gorilla. *Dispersing primate females.* Springer.
ROBBINS, M. M., BERMEJO, M., CIPOLLETTA, C., MAGLIOCCA, F., PARNELL, R. J. & STOKES, E. 2004. Social structure and life-history patterns in western gorillas (Gorilla gorilla gorilla). *American Journal of Primatology: Official Journal of the American Society of Primatologists,* 64**,** 145-159.
ROBINSON, J. T., THORVALDSDÓTTIR, H., WINCKLER, W., GUTTMAN, M., LANDER, E. S., GETZ, G. & MESIROV, J. P. 2011. Integrative genomics viewer. *Nature biotechnology,* 29**,** 24-26.
ROQUES, S., CHANCEREL, E., BOURY, C., PIERRE, M. & ACOLAS, M. L. 2019. From microsatellites to single nucleotide polymorphisms for the genetic monitoring of a critically endangered sturgeon. *Ecology and evolution,* 9**,** 7017-7029.
ROSER, M. 2013. Future population growth. *Our world in data*.
ROSS, S. R.-J., ARNOLDI, J.-F., LOREAU, M., WHITE, C. D., STOUT, J. C., JACKSON, A. L. & DONOHUE, I. 2021. Universal scaling of robustness of ecosystem services to species loss. *Nature communications,* 12**,** 1-7.
RWEGO, I. B., ISABIRYE-BASUTA, G., GILLESPIE, T. R. & GOLDBERG, T. L. 2008. Gastrointestinal bacterial transmission among humans, mountain gorillas, and livestock in Bwindi Impenetrable National Park, Uganda. *Conservation Biology,* 22**,** 1600-1607.

SAHLIN, K., LIM, M. C. & PROST, S. 2021. NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data. *Ecology and evolution,* 11**,** 1392-1398.
SANDERS, D., KEHOE, R. & VAN VEEN, F. F. 2015. Experimental evidence for the population-dynamic mechanisms underlying extinction cascades of carnivores. *Current Biology,* 25**,** 3106-3109.
SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences,* 74**,** 5463-5467.
SARMIENTO, E. B., TM; KALINA, J; 1996. Gorillas of Bwindi-Impenetrable Forest and the Virunga Volcanoes: Taxonomic Implications of Morphological and Ecological Differences. *American journal of primatology,* 40**,** 1-21.
SCALLY, A., DUTHEIL, J. Y., HILLIER, L. W., JORDAN, G. E., GOODHEAD, I., HERRERO, J., HOBOLTH, A., LAPPALAINEN, T., MAILUND, T. & MARQUES-BONET, T. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature,* 483**,** 169-175.
SCHIELTZ, J. M., OKANGA, S., ALLAN, B. F. & RUBENSTEIN, D. I. 2017. GPS tracking cattle as a monitoring tool for conservation and management. *African journal of range & forage science,* 34**,** 173-177.
SCHOLZ, A. H., FREITAG, J., LYAL, C. H., SARA, R., CEPEDA, M. L., CANCIO, I., SETT, S., HUFTON, A. L., ABEBAW, Y. & BANSAL, K. 2022. Multilateral benefit-sharing from digital sequence information will support both science and biodiversity conservation. *Nature Communications,* 13**,** 1086.
SCHULTZ, A., CRISTESCU, R. H., LITTLEFORD-COLQUHOUN, B. L., JACCOUD, D. & FRÈRE, C. H. 2018. Fresh is best: Accurate SNP genotyping from koala scats. *Ecology and Evolution,* 8**,** 3139-3151.
SCHULTZ, A. J., STRICKLAND, K., CRISTESCU, R. H., HANGER, J., DE VILLIERS, D. & FRÈRE, C. H. 2022. Testing the effectiveness of genetic monitoring using genetic non-invasive sampling. *Ecology and evolution,* 12**,** e8459.
SCHWARTZ, M. K., LUIKART, G. & WAPLES, R. S. 2007. Genetic monitoring as a promising tool for conservation and management. *Trends in ecology & evolution,* 22**,** 25-33.
SESSEGOLO, C., CRUAUD, C., DA SILVA, C., COLOGNE, A., DUBARRY, M., DERRIEN, T., LACROIX, V. & AURY, J.-M. 2019. Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Scientific reports,* 9**,** 1-12.
SHOKRALLA, S., PORTER, T. M., GIBSON, J. F., DOBOSZ, R., JANZEN, D. H., HALLWACHS, W., GOLDING, G. B. & HAJIBABAEI, M. 2015. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific reports,* 5**,** 1-7.
SINGH, A., SAHAJPAL, V., THAKUR, M., SHARMA, L. K., CHANDRA, K., BHANDARI, D. & SHARMA, A. 2021. Applicability of human-specific STR systems, GlobalFiler™ PCR Amplification Kit, Investigator 24plex QS Kit, and PowerPlex® Fusion 6C in chimpanzee (Pan troglodytes). *BMC research notes,* 14**,** 1-5.
SLATKO, B. E., GARDNER, A. F. & AUSUBEL, F. M. 2018. Overview of next-generation sequencing technologies. *Current protocols in molecular biology,* 122**,** e59.
SOBRINO, B., BRIÓN, M. & CARRACEDO, A. 2005. SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic science international,* 154**,** 181-194.
STEENWEG, R., HEBBLEWHITE, M., KAYS, R., AHUMADA, J., FISHER, J. T., BURTON, C., TOWNSEND, S. E., CARBONE, C., ROWCLIFFE, J. M. & WHITTINGTON, J. 2017. Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. *Frontiers in Ecology and the Environment,* 15**,** 26-34.
STOIBER, M. & BROWN, J. 2017. BasecRAWller: Streaming nanopore basecalling directly from raw signal. *BioRxiv***,** 133058.
STOKES, E. J., PARNELL, R. J. & OLEJNICZAK, C. 2003. Female dispersal and reproductive success in wild western lowland gorillas (Gorilla gorilla gorilla). *Behavioral Ecology and Sociobiology,* 54**,** 329-339.
STONE, A. C., BATTISTUZZI, F. U., KUBATKO, L. S., PERRY JR, G. H., TRUDEAU, E., LIN, H. &

KUMAR, S. 2010. More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 365**,** 3277-3288.

STRINDBERG, S., MAISELS, F., WILLIAMSON, E. A., BLAKE, S., STOKES, E. J., ABA'A, R., ABITSI, G., AGBOR, A., AMBAHE, R. D. & BAKABANA, P. C. 2018. Guns, germs, and trees determine density and distribution of gorillas and chimpanzees in Western Equatorial Africa. *Science advances,* 4**,** eaar2964.

STRONEN, A. V., MATTUCCI, F., FABBRI, E., GALAVERNI, M., COCCHIARARO, B., NOWAK, C., GODINHO, R., RUIZ-GONZÁLEZ, A., KUSAK, J. & SKRBINŠEK, T. 2022. A reduced SNP panel to trace gene flow across southern European wolf populations and detect hybridization with other Canis taxa. *Scientific reports,* 12**,** 1-14.

SULLIVAN, K. M., MANNUCCI, A., KIMPTON, C. P. & GILL, P. 1993. A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of XY homologous gene amelogenin. *Biotechniques,* 15**,** 636-8, 640.

SUN, J. X., HELGASON, A., MASSON, G., EBENESERSDÓTTIR, S. S., LI, H., MALLICK, S., GNERRE, S., PATTERSON, N., KONG, A. & REICH, D. 2012. A direct characterization of human mutation based on microsatellites. *Nature genetics,* 44**,** 1161-1165.

TABERLET, P. & BOUVET, J. 1994. Mitochondrial DNA polymorphism, phylogeography, and conservation genetics of the brown bear Ursus arctos in Europe. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 255**,** 195-200.

TABERLET, P., GRIFFIN, S., GOOSSENS, B., QUESTIAU, S., MANCEAU, V., ESCARAVAGE, N., WAITS, L. P. & BOUVET, J. 1996. Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acid Research,* 24.

TABERLET, P., MATTOCK, H., DUBOIS-PAGANON, C. & BOUVET, J. 1993. Sexing free-ranging brown bears Ursus arctos using hairs found in the field. *Molecular Ecology,* 2**,** 399-403.

TABERLET, P., WAITS, L. P. & LUIKART, G. 1999. Noninvasive genetic sampling: look before you leap. *Trends in ecology & evolution,* 14**,** 323-327.

TAJIMA, F. 1995. Effect of non-random sampling on the estimation of parameters in population genetics. *Genetics Research,* 66**,** 267-276.

TAN, Y., BAI, P., WANG, L., WANG, H., TIAN, H., JIAN, H., ZHANG, R., LIU, Y., LIANG, W. & ZHANG, L. 2018. Two-person DNA mixture interpretation based on a novel set of SNP-STR markers. . *Forensic Science International: Genetics,* 37**,** 37-45.

TAYLOR, H. R., DUSSEX, N. & VAN HEEZIK, Y. 2017. Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners. *Global Ecology and Conservation,* 10**,** 231-242.

TEAM, R. C. 2013. R: A language and environment for statistical computing.

TEAM, R. C. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria: Vienna.

TELLA, J. L., ROJAS, A., CARRETE, M. & HIRALDO, F. 2013. Simple assessments of age and spatial population structure can aid conservation of poorly known species. *Biological Conservation,* 167**,** 425-434.

TENG, H., CAO, M. D., HALL, M. B., DUARTE, T., WANG, S. & COIN, L. J. 2018. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience,* 7**,** giy037.

THAKUR, M., CHANDRA, K., SAHAJPAL, V., SAMANTA, A., SHARMA, A. & MITRA, A. 2018. Functional validation of human-specific PowerPlex® 21 System (Promega, USA) in chimpanzee (Pan Troglodytes). *BMC research notes,* 11**,** 1-4.

THALMANN , O., FISCHER, A., LANKESTER, F., PÄÄBO, S. & VIGILANT, L. 2006. The Complex Evolutionary History of Gorillas: Insights from Genomic Data. *Molecular Biology and Evolution,* 24**,** 146-158.

THALMANN, O. H., WEGMANN, D., SPITZNER, M., ARANDJELOVIC, M., GUSCHANSKI, K., LEUENBERGER, C., BERGL, R. A. & VIGILANT, L. 2011. Historical sampling reveals

dramatic demographic changes in western gorilla populations. *BMC Evolutionary Biology,* 11**,** 85.
THORVALDSDÓTTIR, H., ROBINSON, J. T. & MESIROV, J. P. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics,* 14**,** 178-192.
TOKARSKA, M., MARSHALL, T., KOWALCZYK, R., J.M., W., PERTOLDI, C., KRISTENSEN, T. N., LOESCHCKE, V., GREGERSEN, V. R. & BENDIXEN, C. 2009. Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: the case of European bison. *Heredity,* 103**,** 326-332.
TONG, S., BAMBRICK, H., BEGGS, P. J., CHEN, L., HU, Y., MA, W., STEFFEN, W. & TAN, J. 2022. Current and future threats to human health in the Anthropocene. *Environment international,* 158**,** 106892.
TYTGAT, O., GANSEMANS, Y., WEYMAERE, J., RUBBEN, K., DEFORCE, D. & VAN NIEUWERBURGH, F. 2020. Nanopore sequencing of a forensic STR multiplex reveals loci suitable for single-contributor STR profiling. *Genes,* 11**,** 381.
TYTGAT, O., ŠKEVIN, S., DEFORCE, D. & VAN NIEUWERBURGH, F. 2022. Nanopore sequencing of a forensic combined STR and SNP multiplex. *Forensic Science International: Genetics,* 56**,** 102621.
VALIERE, N., BONENFANT, C., TOÏGO, C., LUIKART, G., GAILLARD, J.-M. & KLEIN, F. 2007. Importance of a pilot study for non-invasive genetic sampling: genotyping errors and population size estimation in red deer. *Conservation Genetics,* 8**,** 69-78.
VALLEJO-MARIN, M. & LYE, G. C. 2013. Hybridisation and genetic diversity in introduced Mimulus (Phrymaceae). *Heredity,* 110**,** 111-122.
VAN DER HOEK, Y., BINYINYI, E., NGOBOBO, U., STOINSKI, T. S. & CAILLAUD, D. 2021. Daily travel distances of unhabituated Grauer's gorillas (Gorilla beringei graueri) in a low elevation forest. *Folia Primatologica,* 92**,** 112-125.
VAN DER VALK, T., SANDOVAL-CASTELLANOS, E., CAILLAUD, D., NGOBOBO, U., BINYINYI, E., STOINSKI, T., GILISSEN, E., SONET, G., SEMAL, P., KALTHOFF, D. C. & DALÉN, L. 2018. Significant loss of mitochondrial diversity within the last century due to extinction of peripheral populations in eastern gorillas. *Scientific Reports,* 8**,** 6551.
VAN DER WAL, J. E., SPOTTISWOODE, C. N., UOMINI, N. T., CANTOR, M., DAURA-JORGE, F. G., AFAN, A. I., ATTWOOD, M. C., AMPHAERIS, J., BALASANI, F. & BEGG, C. M. 2022. Safeguarding human–wildlife cooperation. *Conservation Letters***,** e12886.
VIENGKONE, M., DEROCHER, A. E., RICHARDSON, E. S., MALENFANT, R. M., MILLER, J. M., OBBARD, M. E., DYCK, M. G., LUNN, N. J., SAHANATIEN, V. & DAVIS, C. S. 2016. Assessing polar bear (Ursus maritimus) population structure in the Hudson Bay region using SNP s. *Ecology and Evolution,* 6**,** 8474-8484.
VILÀ, C., WALKER, C., SUNDQVIST, A.-K., FLAGSTAD, Ø., ANDERSONE, Z., CASULLI, A., KOJOLA, I., VALDMANN, H., HALVERSON, J. & ELLEGREN, H. 2003. Combined use of maternal, paternal and bi-parental genetic markers for the identification of wolf–dog hybrids. *Heredity,* 90**,** 17-24.
VON THADEN, A., NOWAK, C., TIESMEYER, A., REINERS, T. E., ALVES, P. C., LYONS, L. A., MATTUCCI, F., RANDI, E., CRAGNOLINI, M. & GALIÁN, J. 2020. Applying genomic data in wildlife monitoring: Development guidelines for genotyping degraded samples with reduced single nucleotide polymorphism panels. *Molecular ecology resources,* 20**,** 662-680.
WADE, A. H. & MALONE, N. 2021. Ecological, Historical, Economic, and Political Factors Shaping the Human–Gorilla Interface in the Mone-Oku Forest, Cameroon. *Diversity,* 13**,** 175.
WAITS, L. P., LUIKART, G. & TABERLET, P. 2001. Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular ecology,* 10**,** 249-256.
WALL, J. D. 2013. Great ape genomics. *ILAR journal,* 54**,** 82-90.

WANG, J. 2014. Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods. *Journal of Evolutionary Biology,* 27**,** 518-530.
WANG, J. 2017. Estimating pairwise relatedness in a small sample of individuals. *Heredity,* 119**,** 302-313.
WATSA, M. London Calling 2022 Oxford Nanopore Technologies Ltd, 2022 London.
WATSA, M., ERKENSWICK, G. A., POMERANTZ, A. & PROST, S. 2020. Portable sequencing as a teaching tool in conservation and biodiversity research. *PLoS biology,* 18**,** e3000667.
WATTS, D. P. 1994. Composition and variability of mountain gorilla diets in the central Virungas. *American journal of primatology,* 7**,** 323-256.
WEI, S., WEISS, Z. R. & WILLIAMS, Z. 2018. Rapid multiplex small DNA sequencing on the MinION nanopore sequencing platform. *G3: Genes, Genomes, Genetics,* 8**,** 1649-1657.
WEI, S. & WILLIAMS, Z. 2016. Rapid short-read sequencing and aneuploidy detection using MinION nanopore technology. *Genetics,* 202**,** 37-44.
WEI, Y. L., LI, C. X., JIA, J., HU, L. & LIU, Y. 2012. Forensic identification using a multiplex assay of 47 SNPs. *Journal of Forensic Sciences,* 57**,** 1448-1456.
WEIR, B. S. & COCKERHAM, C. C. 1984. Estimating F-statistics for the analysis of population structure. *evolution***,** 1358-1370.
WELCH, E. W., SHIN, E. & LONG, J. 2013. Potential effects of the Nagoya Protocol on the exchange of non-plant genetic resources for scientific research: Actors, paths, and consequences. *Ecological Economics,* 86**,** 136-147.
WHITE, L. C., FONTSERE, C., LIZANO, E., HUGHES, D. A., ANGEDAKIN, S., ARANDJELOVIC, M., GRANJON, A. C., HANS, J. B., LESTER, J. D., RABANUS-WALLACE, M. T., ROWNEY, C., STÄDELE, V., MARQUES-BONET, T., LANGERGRABER, K. E. & VIGILANT, L. 2019. A roadmap for high-throughput sequencing studies of wild animal populations using noninvasive samples and hybridization capture. . *Molecular Ecology Resources* 19**,** 609-622.
WIKBERG, E. C., JACK, K. M., FEDIGAN, L. M., CAMPOS, F. A., YASHIMA, A. S., BERGSTROM, M. L., HIWATASHI, T. & KAWAMURA, S. 2017. Inbreeding avoidance and female mate choice shape reproductive skew in capuchin monkeys (Cebus capucinus imitator). *Molecular ecology,* 26**,** 653-667.
WILLEMS, T., ZIELINSKI, D., YUAN, J., GORDON, A., GYMREK, M. & ERLICH, Y. 2017. Genome-wide profiling of heritable and de novo STR variations. *Nature methods,* 14**,** 590-592.
WILLIAMS, P. H., BURGESS, N. D. & RAHBEK, C. 2000. Flagship species, ecological complementarity and conserving the diversity of mammals and birds in sub-Saharan Africa. *Animal Conservation,* 3**,** 249-260.
WILLIAMS, V. L., COALS, P. G., DE BRUYN, M., NAUDE, V. N., DALTON, D. L. & KOTZÉ, A. 2021. Monitoring compliance of CITES lion bone exports from South Africa. *Plos one,* 16**,** e0249306.
XAVIER, C. & PARSON, W. 2017. Evaluation of the Illumina ForenSeq™ DNA Signature Prep Kit–MPS forensic application for the MiSeq FGx™ benchtop sequencer. *Forensic Science International: Genetics,* 28**,** 188-194.
XUE, Y., PRADO-MARTINEZ, J., SUDMANT, P. H., NARASIMHAN, V., AYUB, Q., SZPAK, M., FRANDSEN, P., CHEN, Y., YNGVADOTTIR, B. & COOPER, D. N. 2015. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science,* 348**,** 242-245.
YAMAGIWA, J. & BASABOSE, A. K. 2006. Diet and seasonal changes in sympatric gorillas and chimpanzees at Kahuzi–Biega National Park. *Primates,* 47**,** 74-90.
YAMAGIWA, J., BASABOSE, A. K., KAHEKWA, J., BIKABA, D., ANDO, C., MATSUBARA, M., IWASAKI, N. & SPRAGUE, D. S. 2012. Long-term research on Grauer's gorillas in Kahuzi-Biega National Park, DRC: life history, foraging strategies, and ecological differentiation from sympatric chimpanzees. *Long-term field studies of primates.* Springer.
YAMAGIWA, J., MWANZA, N., SPANGENBERG, A., MARUHASHI, T., YUMOTO, T., FISCHER, A. &

STEINHAUER-BURKART, B. 1993. A census of the eastern lowland gorillas Gorilla gorilla graueri in Kahuzi-Biega National Park with reference to mountain gorillas G. g. beringei in the Virunga region, Zaire. *Biological Conservation,* 64**,** 83-89.

ZASCAVAGE, R. R., SHEWALE, S. J. & PLANZ, J. V. 2013. Deep-sequencing technologies and potential applications in forensic DNA testing. *Forensic Sci Rev,* 25**,** 79-105.

ZIMOV, S. A., CHUPRYNIN, V., ORESHKO, A. P., CHAPIN III, F. S., REYNOLDS, J. F. & CHAPIN, M. C. 1995. Steppe-tundra transitions: a herbivore-driven biome shift at the end of the Pleistocene. . *American Naturalist,* 146**,** 765-794.