Genomic analysis of wild and captive chimpanzee populations from non-invasive samples using target capture methods

Clàudia Fontserè Alemany

TESI DOCTORAL UPF / 2020

Directors de la tesi

Dr. Tomàs Marquès Bonet

Dra. Esther Lizano González

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Universitat Pompeu Fabra Barcelona

A les meves àvies,

Fina i Margarita

"If you know you are on the right track, if you have this inner knowledge,
then nobody can turn you off... no matter what they say."

Barbara McClintock

Agraïments

Sembla que era ahir quan començava aquesta tesi doctoral i ja han passat 5 anys. En aquest temps, no només he après sobre genòmica i ximpanzés, sinó que han sigut uns anys plens de creixement personal i aquí vull donar les gràcies a molta gent.

A en Tomàs, m'agradaria agrair-te la confiança que vas dipositar en mi per liderar un projecte tan ambiciós com el PanAf on he après i gaudit moltíssim, on he conegut a gent increïble. Gràcies, de veritat.

A l'Esther, el PanAf i la meva tesi no haurien arribat a tan bon port si no fos per tu. Vas aparèixer al grup quan més et necessitàvem. Vas engegar el motoret del projecte, i em vas ensenyar tot el que sé de treballar al laboratori, que et seré sincera, em feia pànic! A part d'això, has sigut el meu suport no només acadèmic sinó també emocional que és tan necessari, gràcies.

To Moose, you have been a mentor to me since the beginning and you have given me the self-confidence that I lacked so many times. Thank you.

To Mimi, I admire you very much. Thanks for your amazing work with the PanAf project and for always supporting me.

A tots els integrants del grup, perquè els dinars i els jocs s'han convertit en el millor moment del dia, i que tant he trobat més a faltar aquests mesos de confinament. Esperem que puguem tornar-hi ben aviat.

A la Jéssica, per ser tan bona, sempre disposada ajudar amb el que faci falta, per ser la meva guia quan vaig començar, per les nostres aventures per Brusel·les. A Irene porque he crecido mucho a tu lado, me encantan nuestra discusiones y que siempre haya espacio para aprender contigo. A en Marc, perquè t'admiro com a científic, per les teves idees, intuïció i interès, per tot el que he après de tu, pels moments divertits. Se us troba molt a faltar!

A l'Aitor, és impossible donar-te les gràcies per tot, sempre que t'he necessitat has estat allà per donar un cop de mà sense dubtar-ho ni un moment, especialment aquests últims mesos amb la tesi. Gràcies. A Manolo, aunque eres (demasiado?) simpático, has llenado el grupo de una energía y una alegría muy necesaria.

A l'"Hotelito team" perquè aquests mesos de quarantena han sigut més amens amb els nostres esmorzars a través de pantalla. Perquè el suport que ens hem donat les unes amb les altres és vital i ens ajuda a superar moments difícils. A la Marina, perquè aquesta tesi sense tu hagués sigut pitjor. Gràcies per ser la meva mà dreta fent llibreries i captures, pels nostres clean-ups conjunts a Leipzig quan el nostre cervell ja no donava més de sí. Gràcies per fer que les meves presentacions, gràfics i portada de la tesi siguin molt més bonics. Gràcies per dir les coses sense embuts i per donar-me ànims i suport en tot. A la Paula, perquè t'admiro moltíssim, jo de gran vull ser com tu, tenir la teva energia i seguretat. Gràcies per ser un puntal de grup, per ser tan valenta. A la Laura, perquè la teva vitalitat i energia positiva s'encomana per allà on passes, i això és molt necessari, en ciència i en la vida. Et trobem a faltar!

To Martin, for being a mentor in the PanAf project, for helping me focus and guide me on how to proceed. Thanks for your support on many aspects of this PhD, and for the long talks we have been having lately. You are going to be an amazing PI.

To Sojung, for your kindness on everything you do. For giving the group a different point of view, for being a good listener, for not hesitating on helping someone else.

To Harvi, for helping me with English spelling and grammar every time I asked without any complains. The group is in the best possible hands.

A Lukas, David, Joe y Luis, por todos estos años compartidos a vuestro lado, gracias.

M'agradaria donar un agraïment immens per tota la gent de Bioevo, als membres passats i als presents. Gràcies Marina BV, Raquel, Diego, Txema, Nino, Álex, Lara, Pere, Carla, Marco, Laia, Xavi, Alejandro, André, Toni, Simone, Pablo, Julen, Sandra, Meritxell, Ferriol, Neus, Borja, Fabio, Bárbara, Rocío, Ana i a tots els que m'hagi oblidat. A la Judit i a la Mònica, perquè vosaltres sou el puntal que aguanta l'IBE mentre els estudiants de doctorat anem passant. Us trobaré molt a faltar.

To Jonas and Fátima, I am very glad to have met both of you. You will always be welcome to Barcelona.

A Miquel Llorente, per cedir-me la fotografia per a la portada, on hi surt Àfrica, un ximpanzé de la Fundació Mona.

Al grup de Ioga, gràcies Marina BV per iniciar-m'hi i a totes les que ho heu fet possible al llarg dels anys. Espero que segueixi molts anys.

Un agraïment molt sentit a la música i tots els que han fet que em torni a enamorar de tocar en una orquestra. Primer de tot a l'Orquestra del PRBB i als seus membres. Però especialment a en Borja Mascaró, director de la Orquestra del PRBB, que em va encoratjar a tocar el Rèquiem de Fauré amb la Jove Orquestra de Cerdanyola (JOC). Gràcies a la JOC per deixar-me tocar amb vosaltres, sou fantàstics. Especialment a la secció de cellos, que tothom sap que som els millors: gràcies Marina, Víctor, Rosa, Paula, Mar...

A en Ricard, Sílvia i Maria perquè aquesta tesi no hauria passat sense els nostres inicis a Biologia Humana. Gràcies Ricard per encoratjar-me a triar la bioinformàtica, a perdre-li la por, i per ser un amic poc ortodox. Gràcies Sílvia, per ser-hi sempre, per poder comptar amb tu. Gràcies Maria, perquè encara que ens veiem poc sempre ens retrobem.

A la colla de Vic, Guillem, Alba, Josep, Mariona i Anna, perquè ens hem vist créixer i espero que continuem compartint moltes coses junts.

A les nenes, Cristina, Olga, Bego, Maria, Margalida, Marta, Anna, Leire… per què l'atzar ens va ajuntar i aquí seguim.

A la meva família, Amàlia, Toni i Ricard, perquè el vostre suport incondicional és la clau de l'èxit de tot el que he aconseguit. Gràcies.

A en Juan, el millor company de vida que podria tenir, per estimar-me tal i com sóc.

Clàudia

Abstract

Wild chimpanzee populations are considered to be under threat of extinction due to the damaging consequences of human impact into their natural habitat and illegal trade. Conservation genomics is an emerging field that has the potential to guide conservation efforts not only in the wild (*in situ*) but also outside their natural range (*ex situ*). In this thesis, we have explored to which extent target capture methods on specific genomic regions can provide insights into chimpanzee genetic diversity in captive and wild populations. Specifically, we have characterized the ancestry and inbreeding of 136 European captive chimpanzees to aid their management in captivity and inferred the origin of 31 confiscated individuals from illegal trade by sequencing ancestry informative SNPs. Also, we have examined molecular strategies to maximize the library complexity in target capture methods from fecal samples so they can be applied in large-scale genomic studies. Finally, we have captured the chromosome 21 from 828 fecal samples collected across the entire extant chimpanzee range. As a result of our high-density sampling scheme, we have found strong evidence of population stratification in chimpanzee populations and we have discovered new local genetic diversity that is linked to its geographic origin. Finally, with this newly generated dataset and fine-grained geogenetic map, we have implemented a strategy for the geolocalization of chimpanzees which has a direct conservation application.

Resum

Les poblacions salvatges de ximpanzés estan en perill d'extinció a causa de les dramàtiques conseqüències associades a l'impacte humà en el seu hàbitat natural i al tràfic il·legal. La genòmica de la conservació és un camp emergent que té el potencial de guiar esforços de conservació d'espècies en perill d'extinció no només en el seu hàbitat natural (*in situ*) sinó també en captivitat (*ex situ*). En aquesta tesi, hem analitzat fins a quin punt els mètodes de captura de regions específiques del genoma són una bona eina per explorar la diversitat genètica dels ximpanzés tant en poblacions captives com salvatges. Concretament, hem caracteritzat la subespècie i els nivells de consanguinitat de 136 ximpanzés de zoos europeus amb l'objectiu de guiar-ne la seva gestió en captivitat, i hem inferit l'origen de 31 individus confiscats del tràfic il·legal a través de la seqüenciació de SNPs informatius de llinatge. També hem posat en pràctica estratègies moleculars per maximitzat la complexitat de les llibreries en la captura de regions específiques a partir de mostres fecals i així poder ser aplicades en estudis genòmics a gran escala. Finalment, hem capturat el cromosoma 21 de 828 mostres fecals recollides per tota la distribució geogràfica dels ximpanzé. Arran de l'alta densitat de mostreig, hem trobat evidències que apunten a una alta estratificació poblacional en els ximpanzés i hem desxifrat nova diversitat genètica vinculada a l'origen geogràfic dels individus. Finalment, amb el conjunt de dades generat i el mapa geogenètic obtingut, hem implementat una estratègia per la geolocalització de ximpanzés amb aplicació directe per a la conservació.

Preface

Chimpanzees are our kin, we resemble each other in many ways: together with bonobos, they are our closest living relatives and we greatly relate to them in terms of social behavior and tool use. To quote Jane Goodall: "Perhaps the greatest difference between *Homo* and our ape relatives is the fact that we have developed a sophisticated language that enables us to plan far into the future and learn from the distant past [...]. Our highly evolved intellect gives us the ability to make decisions regarding the life and deaths of entire species. Only we can make the decision to preserve the apes."

However, ever since industrialization, humans have been transforming the natural habitat of many species, at an even larger scale. In tropical Africa, where chimpanzees live, we have converted what used to be an impenetrable jungle into agriculture fields, opened new roads as well as increased logging and mining leading to sometimes irreparable deforestation. It is our duty to start making decisions to protect biodiversity.

As evolutionary biologists we can make a tiny but valuable contribution in furthering our understanding of endangered species, in this case, the chimpanzee. Genetics and genomics can be a powerful tool to study chimpanzee populations, their genomic diversity and population history. Cataloguing diversity in wild and captive chimpanzee populations could help conservation managers, practitioners and policymakers to redesign established conservation programmes, as well as enforce laws and regulations against poaching and illegal trade.

In this PhD thesis I have explored mechanisms to increase the recovery of DNA from non-invasive samples and have used next-generation sequencing techniques to explore genetic variation and learn about captive and wild chimpanzee populations from invasive and non-invasive samples.

Table of Contents

1. INTRODUCTION

1.1. Chimpanzees in the great ape family

The *Hominidae* family, or the great apes, consists of four genera (*Pan*, *Homo*, *Gorilla* and *Pongo*) and eight living species: chimpanzee (*Pan troglodytes*), bonobo (*Pan paniscus*), eastern gorilla (*Gorilla beringei*), western gorilla (*Gorilla gorilla*), Sumatran orangutan (*Pongo abelii*), Bornean orangutan (*Pongo pygmaeus*), modern humans (*Homo sapiens*) (Groves, 2001), and the recently defined Tapanuli orangutan (*Pongo tapanuliensis*) (Nater *et al.*, 2017). All great ape species but humans live exclusively in the tropics of Africa or Asia (Figure 1) and are considered to be under threat of extinction by the International Union for Conservation of Nature (IUCN) Red List of Threatened Species. On the other hand, *Homo sapiens* is one of the most widely spread species and it is largely responsible for the critical situation of the other great apes (Caldecott and Miles, 2005).

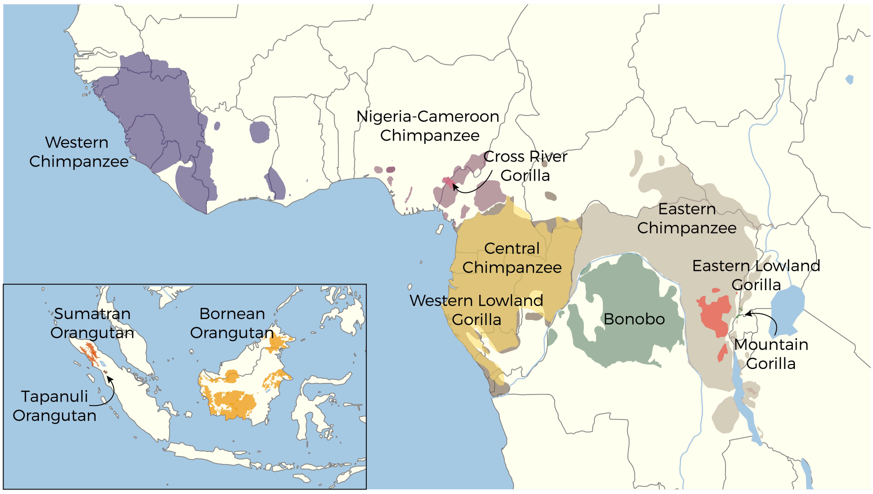


Figure 1. Distribution of the extant non-human great apes (from IUCN shapefiles, https://www.iucnredlist.org).

Of the great ape species occurring in Africa, chimpanzees have the largest and most widespread distribution, a total of 2.6 million km^2 (derived from the IUCN shapefiles). They have a discontinuous distribution ranging from southern Senegal to Uganda and Tanzania. The chimpanzee distribution overlaps with the gorilla distribution range in certain countries, such as in Nigeria, Cameroon, Equatorial Guinea, Congo, Central African Republic (CAR) and Gabon. The Congo river in Democratic Republic of Congo (DRC) is the natural barrier that separates chimpanzees from the other species in the *Pan* genus, the bonobos (Kormos *et al.*, 2003; Caldecott and Miles, 2005).

There are four recognized chimpanzee subspecies with their common name referring to their geographical location in equatorial Africa: central chimpanzee (*Pan troglodytes troglodytes*), eastern chimpanzee (*Pan troglodytes schweinfurthii*), western chimpanzee (*Pan troglodytes verus*) and, latest to be recognized as a subspecies, the Nigeria-Cameroon chimpanzee (*Pan troglodytes ellioti*) (Gonder *et al.*, 1997; Groves, 2001; Prado-Martinez *et al.*, 2013).

Their distribution encompasses different climates: from evergreen forest, through mosaic woodlands and deciduous forest, to dry savanna and from sea level to about 2,600m elevation (Caldecott and Miles, 2005). Most probably, this species spanned most of equatorial Africa, with a range over at least 25 countries in the beginning of the 20th Century (Humle *et al.*, 2016). Currently, chimpanzees can be found in 22 countries: Angola, Burundi, Cameroon, CAR, Congo, DRC, Côte d'Ivoire, Equatorial Guinea, Gabon, Ghana, Guinea, Guinea-Bissau, Liberia, Mali, Nigeria, Rwanda, Senegal, Sierra Leone, South Sudan, Tanzania and Uganda. Most possibly, they are extinct in Benin, Burkina Faso and Togo (Humle *et al.*, 2016).

Central chimpanzees occur in Cameroon, south of the Sanaga river which separates them from the Nigeria-Cameroon chimpanzees. They extend across seven countries to the Ugangi/Congo River at DRC, which delimitates the separation of this subspecies with eastern chimpanzees (IUCN, 2014). The eastern chimpanzee distribution ranges from the Ubangi/Congo river in southeast CAR and DRC to Burundi, Rwanda, western Uganda and western Tanzania, with some populations in South Sudan (Plumptre *et al.*, 2010). As their name indicates, the Nigeria-Cameroon chimpanzees occur in southern Nigeria in small highly fragmented populations and also along the border with Cameroon, north of the Sanaga River. This chimpanzee subspecies has the smallest geographical range of the four (Morgan *et al.*, 2011). Finally, the western chimpanzee range goes from southeast Senegal into southwest Mali and southern Guinea-Bissau (IUCN SSC Primate Specialist Group, 2020). Their geographical range is currently highly fragmented, but the western chimpanzee distribution may have been almost continuous until the middle of the last century (Jolly, Oates and Disotell, 1995).

1.2. Chimpanzee conservation

Chimpanzees are the most numerous non-human great ape species. Nonetheless, their populations are suffering dramatic declines as human impact into their habitat intensifies (Humle *et al.*, 2016). Current estimates of chimpanzee census population sizes suggest that there are fewer than 3,500-9,000 Nigeria-Cameroon chimpanzees (Morgan *et al.*, 2011), approximately 130,000 central chimpanzees (Strindberg *et al.*, 2018) and between 15,000 and 65,000 western chimpanzees (best estimates pointing to around 35,000 (Kühl *et al.*, 2017)). The most numerous chimpanzee subspecies is the eastern chimpanzee, with estimates of 173,000-248,000 individuals (Plumptre *et al.*, 2010; Hicks *et al.*, 2014).

Chimpanzees, as well as the other great apes, have low reproductive rate, long generation time (~25 years) and long interbirth intervals making them especially vulnerable to loss or modification of their habitat (Caldecott and Miles, 2005; Langergraber *et al.*, 2012; Kühl *et al.*, 2017). Although their flexible behavior and high degree of cultural variation can favor their adaptation to changing environments (Hockings *et al.*, 2015), chimpanzees have gone extinct in parts of their historical range in the last century (Funwi-Gabga *et al.*, 2014; Humle *et al.*, 2016).

Their sustained population decline in the last decades has been noted by the IUCN Red List of Threatened Species. Since 1996, chimpanzees have been classified as "Endangered" and in 2016 the western chimpanzee status was upgraded to "Critically Endangered" due to a population decline of about 80% in 24 years (Humle *et al.*, 2016; Kühl *et al.*, 2017).

1.2.1. Threats

Although there are national parks and protected areas, the majority of chimpanzees live outside these, and so are vulnerable to habitat disturbance and illegal trade (Figure 2).

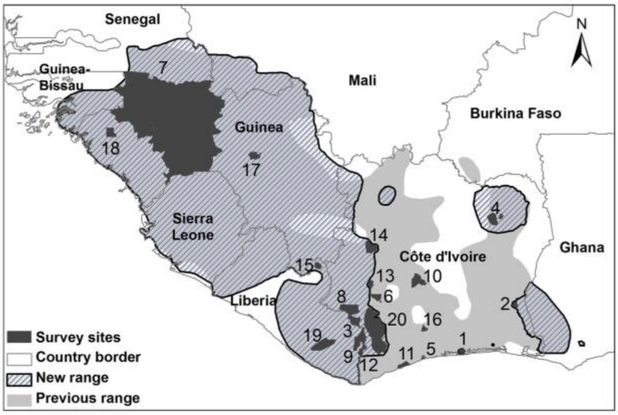


Figure 2. Reduction of habitat range in western chimpanzees. Adapted from (Kühl *et al.*, 2017).

Subsistence and industrial agriculture are converting forest to farmland, reducing the availability of chimpanzee habitat. In West Africa more than 80% of forest cover has been lost since the 19th Century (Norris *et al.*, 2010). Also, palm oil plantations, which have extensively altered Southeast Asian forests and negatively impacted biodiversity, in particular orangutans (Wich *et al.*, 2012), are recently expanding into tropical Africa (Rival and Levang, 2014). There are concerns that the implementation of palm oil agriculture in Africa will lead to similar biodiversity losses since 42% of great ape ranges overlap lands suitable for palm oil plantations (Wich *et al.*, 2014). Extractive industries such as logging, mining and oil are reducing the forest carrying capacity for chimpanzees, disrupting the forest ecosystem, degrading its integrity, fragmenting the continuity of the habitat and making it more prone to fire (Caldecott and Miles, 2005; Morgan and Sanz, 2007).

As a direct consequence of human activities, major transportation infrastructures are being constructed, fragmenting the species' habitat even more and increasing human accessibility to remote areas (Laurance *et al.*, 2014). This situation exacerbates hunting, especially for the commercial trade of bushmeat. Although killing or capturing great apes is illegal, poaching and live-animal trade still pose one of the greatest threats to chimpanzees. Sometimes, when chimpanzees are killed for meat, their infants are captured for the pet trade, entertainment industry and biomedical research (Hicks *et al.*, 2010), although this is explicitly forbidden by CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora). Between 2005 and 2011 it was reported that a minimum of 643 chimpanzees were trafficked, although the number of total losses is extrapolated to be 20 times bigger (Stiles *et al.*, 2013). This increase in accessibility into previously remote and isolated forests is favoring the frequent human-great apes interaction which make great apes more vulnerable to human infectious diseases, such as Ebola virus disease (EVD) (Leroy *et al.*, 2004; Bermejo *et al.*, 2006) and respiratory infections (Kaur *et al.*, 2008; Köndgen *et al.*, 2008), as well as increase the risk of spillover from great apes to humans (Gilardi *et al.*, 2015; Devaux *et al.*, 2019).

1.2.2. Conservation Measures

The biodiversity loss due to anthropogenic activities in the tropical, moist forest has motivated the development of conservation measures to maintain diversity of living organisms as well as their habitats. Traditionally, wild species conservation approaches have been divided into *in situ* and *ex situ*. According to Article 2 of the Convention on Biological Diversity (CBD, 1992), *ex situ* conservation is defined as the "conservation of components of biological diversity outside their natural habitat" whereas *in situ* conservation is defined as the "conservation of ecosystems and natural habitats and the maintenance and recovery of viable populations of species in their natural surroundings". In the next sections I will outline the most important features of both strategies.

a) *In situ*

Although chimpanzees are protected by national and international laws, law enforcement against poaching or illegal logging and mining is poor, even in protected areas. Specific plans and stricter enforcement of wildlife laws are urgently needed to counteract the otherwise likely extinction of chimpanzees (Humle *et al.*, 2016).

Awareness of this situation has driven the creation of major international initiatives, such as the Great Ape Survival Project (GRASP) and the Section on Great Apes of the IUCN Species Survival Commission (SSC) Primate Specialist Group. The SSC Primate Specialist Group has published regional conservation action plans for each subspecies of chimpanzee, where the current situation of the species is reviewed and the conservation priority actions to be carried out are reported (Tutin *et al.*, 2005; Plumptre *et al.*, 2010; Morgan *et al.*, 2011; IUCN, 2014; IUCN SSC Primate Specialist Group, 2020). The action plans identify priority populations in which immediate action must be taken, relying on information collected from many field sites over the past decades. Although every population may experience slightly different threats and require specific actions, all subspecies conservation plans converge to the need for support from local communities, as well as the creation of education programmes to raise awareness among local people. In the action plans, the importance of research as well as the precise monitoring and status surveillance of each chimpanzee population is reported. To fight against poaching in the short term, the usefulness of strict enforcement of wildlife protection laws through guard patrols is highlighted. In the long term, education campaigns to change the sometimes negative attitude towards chimpanzees, inform about conservation importance and rarity of this species and reduce the demand on bushmeat would greatly diminish trade. Finally, forest restoration and habitat protection to interconnect protected areas can largely reduce extinction risk (Edroma, Rosen and Miller, 1997).

None of these actions alone will prove sufficient to avoid chimpanzee population decline and eventual extinction, but the partnership of many non-governmental institutions as well as governments, international conventions and regional agreements is setting the groundwork for protecting these species in their natural habitat.

b) *Ex situ*

Captive management of the threatened species, outside their natural range, occurs in zoos and sanctuaries and might contribute to their conservation. Historically, zoos were designed as amusement parks, and this perception might still be common and justified in some cases. Nowadays, zoos have the moral obligation to play an active part in conservation, as they can maintain carefully managed populations of animals and minimize loss of genetic diversity. Moreover, captive populations are valuable in furthering our knowledge of the biology of a species and can play a role in education. Importantly, *ex situ* conservation actions should have a direct link to conservation in the wild, whether through breeding for reintroduction whenever possible or by rising awareness or funds for *in situ* conservation. All these principles are set out in the World Zoo and Aquarium Conservation

Strategy (WZACS) and the IUCN SSC Guidelines on the Use of *Ex situ* Management for Species Conservation (WAZA, 2005; IUCN/SSC, 2014).

The management of captive chimpanzee populations outside of Africa is done by regional breeding programmes such as the American Zoo and Aquarium Association (AZA), the European Association of Zoos and Aquaria (EAZA) or the Australiasian Regional Association of Zoological Parks and Aquaria (ARAZPA). The EAZA Ex situ Programme (EEP) is the largest and manages chimpanzee subspecies separately with specific breeding programmes for the western chimpanzee and the central chimpanzee (Carlsen and de Jongh, 2019) while the AZA manages all four subspecies as one population (AZA, Ape and TAG, 2010).

Given the current state of decline of wild chimpanzee populations and their clear extinction risk, it might become necessary to turn to future reintroductions of captive-born individuals. To do so, it is of high importance that breeding programmes preserve healthy (physically, behaviorally and genetically) self-sustaining populations that resemble their wild counterparts (Carlsen and de Jongh, 2019). Studbook pedigree information is usually used to minimize hybridization between subspecies, inbreeding and loss of genetic diversity. However, the completeness of studbooks is limited, mainly because of an unknown origin and genetic relationship of the chimpanzees, which has led to admixture of subspecies in the captive populations (Hvilsom *et al.*, 2013).

Wild-born apes placed at sanctuaries are currently the most viable option for reintroduction into the wild (Beck *et al.*, 2007). Sanctuaries have arisen due to the necessity to find a place for chimpanzees confiscated from illegal pet trade (IUCN/SSC, 2000). Many primate and wildlife sanctuaries have been established throughout Africa and in 2002 they formed the network Pan African Sanctuary Alliance (PASA, 2002). Usually animals that arrive at sanctuaries have poor health so these facilities need to fulfill a role in animal welfare. Besides, sanctuaries help the authorities to implement confiscation policies and serve as educational centers to local people and visitors. Some of the rescued chimpanzees require specialised lifetime care but others, after extensive rehabilitation and preparation, may be suitable for reintroduction into the wild (Beck *et al.*, 2007; IUCN/SSC, 2013). For such reintroduction, knowledge of their geographic origin is essential to preserve local adaptation and avoid a potential reduction of fitness (outbreeding depression) by the mixture of different populations. Such a situation has already been reported in orangutan reintroductions (Banes, Galdikas and Vigilant, 2016). Hence, genetic assessment should then be performed (Beck *et al.*, 2007; Banes, Galdikas and Vigilant, 2016). However, reintroduction will be ineffective unless the habitat is secure and without hunting pressure (Kormos *et al.*, 2003).

Recently, the IUCN SSC Conservation Breeding Specialist Group (CBSG) proposed to integrate *in situ* and *ex situ* conservation planning into a "One Plan Approach" (Byers *et al.*, 2013): one comprehensive conservation plan that joins management strategies from all conservation parties (Traylor-Holzer, Leus and Mcgowan, 2013; Traylor-Holzer, Leus and Bauman, 2019).

1.3. Studying the great apes' genomes

As discussed above (section 1.2.1.), it is clear that chimpanzees are under risk of extinction and actions must be taken immediately. As evolutionary biologists, we can contribute to conservation by exploring their genomes and studying genetic diversity, demographic history and other aspects that are reviewed in the next sections of this thesis. But first we need to recapitulate what we know in terms of the genetic relationships within the great apes and their genetic composition, and also how such knowledge was learnt.

1.3.1. History of great ape taxonomy

Nowadays there is no doubt that humans belong to the great ape family. However, this classification was not always assumed to be true or even considered possible. The first systematic classification of animal and plant species was done by the swedish botanist Carl Linnaeus, who devised the binomial system of nomenclature (*Genus species*). Linnaeus, in his book *Systema Naturae* (1758) placed chimpanzees and humans within the same genus, but without an evolutionary perspective at the time. After the publication of *On the Origin of Species* (1859) by Charles Darwin, Thomas Henry Huxley, who was a fervent opponent of Darwin's theory of evolution by natural selection, proposed that humans should be considered within their own order and separated from the rest of great apes and primates in his book *Man's Place in Nature* (1863). Later on, Darwin in *The Descent of Man* (1871) disagreed with Huxley's views and considered that humans and great apes should at least form a family or a sub-family.

During the 20th century, although it was clear that great apes were humans' closest relatives, the relationship between them was still a mystery. We know now that the results from morphological studies to resolve the great ape phylogeny were neither conclusive nor correct (Collard and Wood, 2000). In the 80s, chromosome karyotype reconstruction for human, chimpanzee, gorilla and orangutan species allowed the deduction of their phylogeny, with chimpanzees being closest related to humans and orangutans the outgroup (Yunis and Prakash, 1982). However, those studies were not capable of estimating the time of these speciation events.

When molecular approaches started to become available the possibility to correctly elucidate the phylogenetic relationship and estimate the timing of speciation of great apes became possible. Initial work in molecular anthropology used immunological assays to date the split between humans and the other great apes, although the gorilla-chimpanzee-human split was unresolved because the technology did not have enough resolution (Sarich and Wilson, 1967). Next, with DNA-DNA hybridization experiments (Sibley and Ahlquist, 1984) and the sequencing of genes and many neutral, single-copy orthologous loci the gorilla-chimpanzees-human trichotomy was resolved with chimpanzees and bonobos being the closest relatives to humans (Figure 3) (Miyamoto, Slightom and Goodman, 1987; Bailey *et al.*, 1992; Chen and Li, 2001).

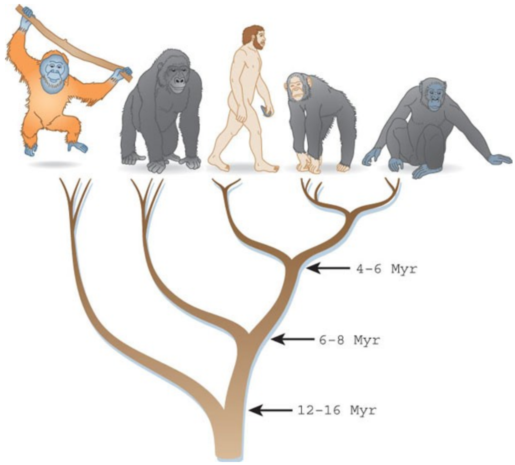


Figure 3. Great ape phylogeny with approximate dates of divergences. Adapted from (Pääbo, 2003).

Once the taxonomic classification of the great apes was resolved, attempts to reconstruct their demographic history followed. First, studies using a limited number of genetic autosomal markers already pointed to a complex demographic history (Fischer *et al.*, 2004, 2006; Steiper, 2006; Becquet and Przeworski, 2007). However, these studies lacked a full representation of the genome, and given the complexity of the demographic history of great apes, a more complete study including the whole genome was needed. The era of genome sequencing was about to start and would revolutionize the way we study and understand the biology of species.

1.3.2. Great ape reference genomes

The human genome was the first to be sequenced (Lander *et al.*, 2001). The Human Genome Project (HGP), an international project with the collaboration of many centers and countries, took over 13 years and around 3 billion dollars to complete (Hayden, 2014). The funding mainly came from the National Institutes of Health (NIH) as well as other groups around the world and the main technology used was Sanger sequencing (Sanger, Nicklen and Coulson, 1977), a laborious and low throughput technology. At the same time, a parallel project conducted by Celera Genomics Corporation applied whole-genome shotgun sequencing (Myers, 1999). With this new methodology, the sequence of the human genome was faster and cheaper than the HGP but at lower quality. In 2001 both the Celera and the HGP published the human genome (Lander *et al.*, 2001; Venter *et al.*, 2001). It was a massive effort with an optimistic forecast that the human genome would enable the mapping and identification of the genetic causes of traits and diseases. Now, with many genome-wide datasets, from many populations and a much clearer understanding of the human genome it is obvious that heritability and disease are far more complex than initially anticipated (Manolio *et al.*, 2009).

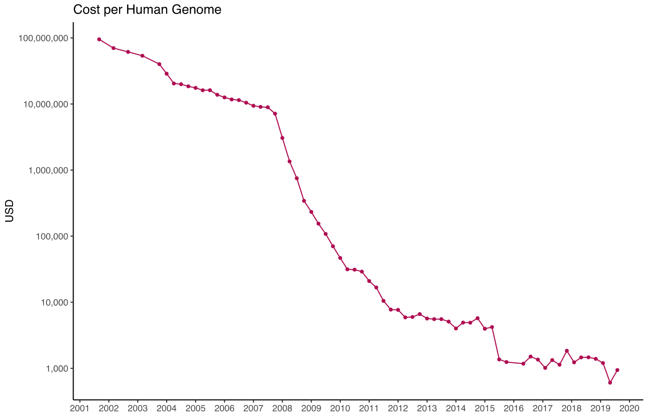


Figure 4. Cost of whole genome sequencing. Notice that the y axis is log10 transformed. Adapted from National Human Genome Research Institute (NHGRI), data: https://www.genome.gov/sequencingcostsdata .

After the human genome, sequencing of model organisms followed, such as the mouse genome (Chinwalla *et al.*, 2002) still using Sanger sequencing. At this point the cost of sequencing kept reducing, with the newly developed methods of next-generation sequencing (NGS) technologies by first 454 pyrosequencing (Margulies *et al.*, 2005) and then Solexa/Illumina (Bentley *et al.*, 2008) (Figure 4).

Therefore, at a more affordable price, the sequencing of other species started. Primate genomes became a high priority for whole genome sequencing (Marques-Bonet, Ryder and Eichler, 2009). Specifically, the NIH gave high priority to the chimpanzee genome (Check, 2002) with the hope to examine the genetic changes that are associated with the rapid evolution of new phenotypic characteristics in humans, as well as to identify genetic differences of medical interest (Olson and Varki, 2003). The genome of a male captive-born western chimpanzee (named Clint) was sequenced using whole genome shotgun sequencing and published in 2005 (Waterson, Lander and Wilson, 2005). After that, it was a matter of time until the rest of great apes would have their genomes sequenced. The orangutan reference genome was published in 2011 (Locke *et al.*, 2011) whereas the bonobo (Prüfer *et al.*, 2012) and the gorilla genomes (Scally *et al.*, 2012) were published one year later.

After the release of the first reference genomes and thanks to the reduction on sequencing costs of the high-throughput sequencing technologies, there have been many resequencing studies in humans. One example is the 1000 Genomes Project, an international collaboration to produce an extensive catalogue of human genetic diversity (Durbin *et al.*, 2010). Such an approach had not been fully explored in other great apes at that moment. Only the gorilla and orangutan genome projects included resequencing of a few individuals from different populations to provide insights into the demographic history of those species. Other studies used resequenced chimpanzee genomes to elucidate recombination, mutation rate and balancing selection (Auton *et al.*, 2012; Leffler *et al.*, 2013; Venn *et al.*, 2014), but not in a population genetic diversity approach as explained in the following section.

1.3.3. Great ape genetic diversity

Although single genomes already proved successful in describing some aspects of the biology of great apes, having more complete catalogues of great ape genetic variation was necessary to determine which characteristics were unique to each lineage, understand their past demographic events and particularly to interpret functional significance of the variation observed only in one lineage (Kuhlwilm, de Manuel, *et al.*, 2016).

Before going into great ape genetic diversity, I will briefly describe what we understand as genetic variation.

a) Genetic variation

Genetic variation is the difference in DNA sequences between individuals. The ultimate force that generates genetic variation is mutation, while recombination creates new allele combinations. Genetic variation can be shaped by various factors such as migration, isolation, admixture, genetic drift and selection. Therefore by studying how patterns of genetic variation change in different populations we have a powerful tool to reconstruct population demography and evolutionary events (Jobling *et al.*, 2013).

Genetic variation can range from single nucleotide variation (SNV) to translocations of chromosomal segments or even changes in chromosome number. In between both events of genetic variation, we find small insertions and deletions of a few bases (indels), tandem repeated DNA sequence expansions or contractions, insertion of transposable elements and structural variation involving millions of base pairs of DNA sequence (deletions, duplications or inversions) (Figure 5).

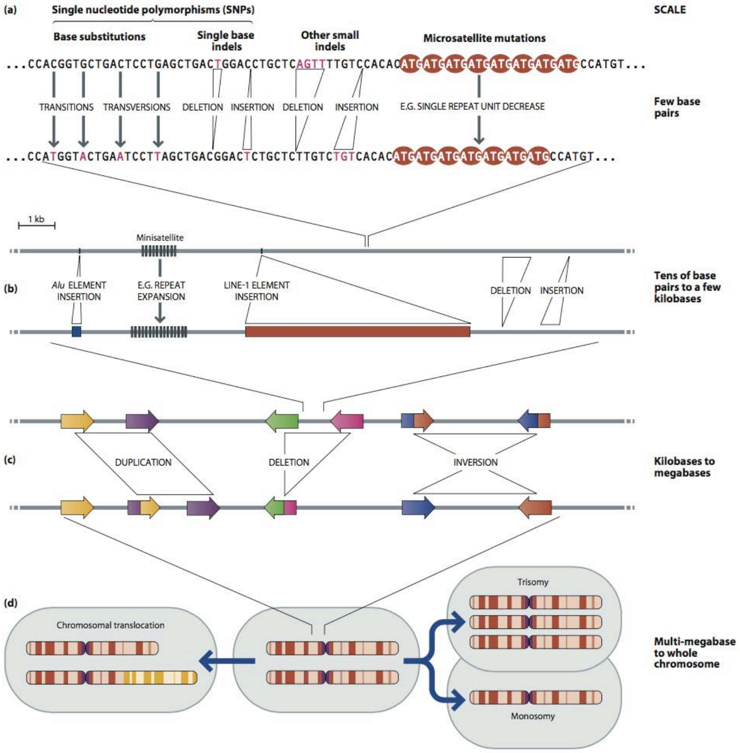

(a)
Single nucleotide polymorphisms (SNPs)
SCALE
Base substitutions
Single base indels
Other small indels
Microsatellite mutations
...CCACGGTGCTGACTCCTGAGCTGACTGGACCTGCTCAGTTTTGTCCACACATGATGATGATGATGATGATGATGCCATGT...
TRANSITIONS
TRANSVERSIONS
DELETION
INSERTION
DELETION
INSERTION
E.G. SINGLE REPEAT UNIT DECREASE
Few base pairs
...CCATGGTACTGAATCCTTAGCTGACGGACTCTGCTCTTGTCTGTCACACATGATGATGATGATGATGATGCCATGT...
1 kb
Minisatellite
(b)
Alu ELEMENT INSERTION
E.G. REPEAT EXPANSION
LINE-1 ELEMENT INSERTION
DELETION
INSERTION
Tens of base pairs to a few kilobases
(c)
DUPLICATION
DELETION
INVERSION
Kilobases to megabases
(d)
Chromosomal translocation
Trisomy
Monosomy
Multi-megabase to whole chromosome

Figure 5. Types of genetic variation according to their size: (a) single or a few base pairs, (b) from tens of base pairs to a few kilobases, (c) events involving kilobases or megabases, and (d) multi-megabase and whole chromosome rearrangements. Adapted from (Jobling *et al.*, 2013).

Apart from the variation affecting genome sequence and structure, there are other forms of variation such as methylation, histone modifications or even variation in the 3D structure and folding of genomic material (Jobling *et al.*, 2013).

In this thesis, I will focus on the SNVs (also known as point mutations or base substitutions) that can be observed in a sufficiently large fraction of the population. Those variants are called single nucleotide polymorphisms (SNPs).

b) Great ape population dynamics, demography and genome-wide variation

To describe genetic variation patterns, we first need to define and delimit a population and then we can measure the allele frequencies in it. Taxonomic classifications such as species and subspecies are usually the units in which population genetic parameters are estimated.

The biological species concept describes species as an interbreeding natural population that is reproductively isolated from others (Mayr, 1970). Historically, taxonomic classification (see section 1.3.1.) has defined species and subspecies by grouping individuals that share a morphology, ecology, behaviour and geographical distribution. More recently, molecular data has been used to delimitate species and subspecies. In great apes, recent changes in their classification included a new chimpanzee subspecies (Gonder *et al.*, 1997), separation of gorillas into two species, as well as the description of a new orangutan species (Xu and Arnason, 1996; Nater *et al.*, 2017).

Making such a classification is not straightforward. However, population classification has clear implications for conservation since it defines the conservation units which are taken into consideration by the policy makers on the management of that specific population (Supple and Shapiro, 2018) (see section 1.4.1.).

First analyses of genetic variation in great apes used a limited number of autosomal markers and mtDNA loci. These early studies already suggested that non-African humans have the lowest genetic diversity while the rest of the great apes bear high levels of diversity, pointing to very different population histories in each lineage (Gagneux *et al.*, 1999; Kaessmann *et al.*, 2001; Fischer *et al.*, 2006). As a result of the complex demographic history of great apes, sampling for genetic studies should be done with extreme caution to discover the whole genetic landscape of a population (Fischer *et al.*, 2006).

Still, these studies had a partial view of the genome by only looking at a few genetic markers, while whole genome resequencing studies could give an unbiased and much more complete view of the genome (Allendorf, Hohenlohe and Luikart, 2010).

	Species	Samples	Accession codes
Locke *et al* (2011)	**Orangutan**	**10**	**SRA** SRR032885-SRR032893 SRR032956-SRR032969 SRR032971-SRR032972 SRR032981-SRR032986 SRR033075-SRR033076 SRR033085- SRR033092 SRR033101-SRR033156 SRR033334-SRR033363 SRR033402-SRR033541
	Sumatran	5	
	Bornean	5	
Auton *et al* (2012)	**Chimpanzee**	**10**	http://panmap.uchicago.edu
	Western	10	

Scally *et al* (2012)	**Gorilla**	**4**	**SRA**
			ERS004138; ERS008713;
			ERS008712; SRX023771;
			SRX023772; SRX023773

	Western Lowland	3	
	Eastern Lowland	1	
Prado-Martinez *et al* (2013)	**Chimpanzee**	**25**	**SRA** PRJNA189439; SRP018689
	Western	5*	
	Central	4	
	Nigeria-Cameroon	10	
	Eastern	6	
	Bonobo	**13**	
	Gorilla	**31**	
	Western Lowland	27	
	Cross-River	1	
	Eastern Lowland	3	
	Orangutan	**10**	
	Sumatran	5	
	Bornean	5	
	Human	**9**	
Venn *et al* (2014)	**Chimpanzee**	**9**	**ENA** PRJEB5937
	Western	9	

Xue *et al* (2015)	**Gorilla**	**13**	**ENA**
			ERS168204;ERS525616;
			ERS525618;ERS525617;
			ERS168207;ERS168410;
			ERS168174;ERS525621;
			ERS168205;ERS525620;
			ERS525622;ERS525619;
			ERS168206;

	Mountain	7	
	Eastern Lowland	6	
de Manuel *et al* (2016)	**Chimpanzee**	**34**	**ENA** PRJEB15086
	Western	7	
	Central	14	
	Eastern	13	
Nater *et al* (2017)	**Orangutan**	**17**	**ENA** PRJEB19688
	Tapanuli	2	
	Sumatran	15	

Table 1. Great ape whole genomes sequenced in each study.
* One chimpanzee is a hybrid between western and central subspecies.
SRA: Sequence Read Archive; ENA: European Nucleotide Archive.

Before the first comprehensive catalogue of whole-genome genetic diversity and population history of great apes was published in 2013 (Prado-Martinez *et al.*, 2013), there were just a handful of great ape whole genomes sequenced, definitely not representative of the entire variation of the great ape family (Locke *et al.*, 2011; Auton *et al.*, 2012; Scally *et al.*, 2012; Venn *et al.*, 2014). In this effort by Prado et al. (2013), resequencing experiments by shotgun sequencing were conducted for all great ape species described at that time, excluding mountain gorillas. It included a total of 88 genomes at an average of 25x coverage: 9 humans, 25 chimpanzees, 13 bonobos, 31 gorillas and 10 orangutans. Two years later, 13 eastern gorillas, including 7 mountain gorilla genomes, were sequenced (Xue *et al.*, 2015) and 17 new orangutans were sequenced in 2017 (Nater *et al.*, 2017), where a new orangutan species *P. tapaluniensis* was proposed (Table 1).

This new available catalogue of variation allowed the characterization of genomic diversity in the great ape family. It also allowed scans of shared and private diversity in each lineage and increased the understanding of their population history. Prado-Martinez *et al.* (2013) found support for genetically distinct populations and subpopulations within each great ape species. They constructed a model of the great ape populations over the last 15 million years by estimating their divergence and ancestral and present effective population sizes (Figure 6).

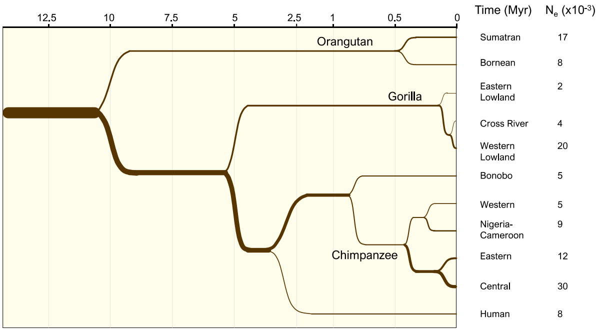
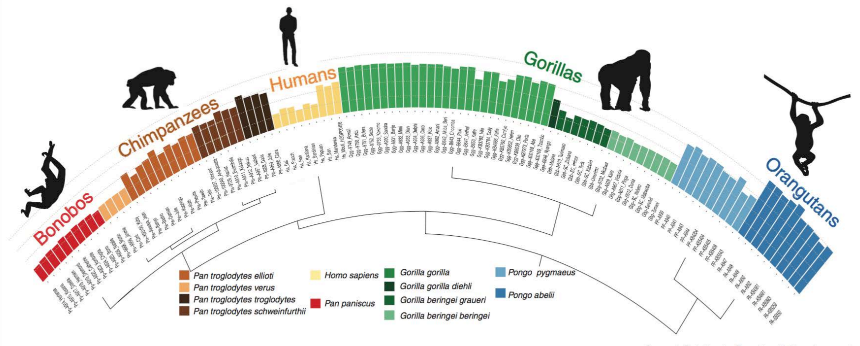


Figure 6. Population splits and effective population sizes (Ne) in the great apes. Adapted from Prado-Martinez (2013).

Moreover, this study also quantified the genetic diversity present at each lineage by estimating the genome-wide heterozygosity (the fraction of the loci that are heterozygous in each individual). The global heterozygosity patterns determined that non-African humans, eastern gorillas (lowland and mountain), bonobos and western chimpanzees show the lowest genetic diversity. In contrast, central chimpanzees, western lowland gorillas and orangutan species show the greatest genetic diversity (Figure 7).



Current Opinion in Genetics & Development

Figure 7. Genome-wide diversity and phylogenetic relationships among the great apes. Adapted from (Kuhlwilm, de Manuel, *et al.*, 2016).

The in-depth analysis of regions with depleted heterozygosity or Runs of Homozygosity (RoHs) to study inbreeding revealed that almost all wild populations presented some degree of inbreeding and eastern gorillas were the most extreme case with evidence of not only recent but also ancient inbreeding (Prado-Martinez *et al.*, 2013). This is precisely the case of mountain gorillas which suffered a prolonged population decline and extensive inbreeding that has led to purging of deleterious recessive mutations (Xue *et al.*, 2015).

c) Population history, gene flow and genetic diversity in chimpanzee populations

Being among the closest relatives to our own species, unraveling the demographic history of chimpanzees would provide an excellent opportunity for comparisons with our own history. This premise favored the focus of phylogenetic and demographic studies on chimpanzees, even before genome-wide data was widely available. At that time, the majority of our knowledge on chimpanzee diversity relied on population genetic data from mitochondrial genomes (Stone *et al.*, 2010), nuclear fragments (Fischer *et al.*, 2004, 2006; Caswell *et al.*, 2008), and microsatellites (Becquet *et al.*, 2007; Wegmann and Excoffier, 2010). These studies already hinted at a complex evolutionary history of the four taxonomically recognized chimpanzee subspecies, later confirmed with the analysis of 59 whole-genomes (Table 1) (Prado-Martinez *et al.*, 2013; de Manuel *et al.*, 2016).

Chimpanzee taxonomy has been under debate for many years. Today it is divided into two monophyletic clades, each including two subspecies (split time= ~500 Kya): a first clade composed by central and eastern chimpanzees (split time = ~150 Kya), and a second one composed by Nigeria-Cameroon and western chimpanzees (split time = ~250 Kya) (de Manuel *et al.*, 2016). Present-day population estimates show that central chimpanzees harbour the highest diversity, followed by eastern, Nigeria-Cameroon and western chimpanzees (Figure 7). All chimpanzee subspecies but the central chimpanzee show patterns of population bottlenecks, with stronger drift effects in western and Nigeria-Cameroon chimpanzees (Prado-Martinez *et al.*, 2013; de Manuel *et al.*, 2016).

Also, multiple events of recent gene flow between chimpanzee populations have influenced their genetic diversity. There is evidence of genetic exchange between central and eastern (Wegmann and Excoffier, 2010; Prado-Martinez *et al.*, 2013), central and Nigeria–Cameroon (Gonder *et al.*, 2011), and central and western chimpanzee (Wegmann and Excoffier, 2010; Gonder *et al.*, 2011) subspecies. Gene flow involving ancestral populations has been suggested between the western chimpanzee subspecies and the ancestor of the central and the eastern chimpanzees subspecies (Hey, 2010). Also, in captivity, chimpanzee subspecies are known to hybridize (Hvilsom *et al.*, 2013).

Interestingly, gene flow from bonobos into the ancestors of central and eastern chimpanzees has been described (Wegmann and Excoffier, 2010) and dated to happen between 200 and 550 thousand years ago (kya), to an extent of less than 1% (de Manuel *et al.*, 2016) as well as, ancient admixture from a ghost population into bonobos (Kuhlwilm *et al.*, 2019).

This complex picture of recurrent introgression events and gene flow from extinct populations has already been described in the human lineage, with neanderthal (Green *et al.*, 2010; Vernot and Akey, 2014; Fu *et al.*, 2015) and denisovan (Reich *et al.*, 2010; Meyer *et al.*, 2012; Prüfer *et al.*, 2014; Vernot *et al.*, 2016) introgression into modern humans, and from humans to neanderthals (Kuhlwilm, Gronau, *et al.*, 2016). Also, introgression from a ghost population into modern humans has been proposed (Hammer *et al.*, 2011; Mondal *et al.*, 2016; Durvasula and Sankararaman, 2020). Such events are less explored in *Gorilla* (Prado-Martinez *et al.*, 2013; McManus *et al.*, 2015) and *Pongo* (Nater *et al.*, 2017), although evidence of admixture has also been described. Actually, admixture appears to be abundant in primate and other mammalian species (Fontsere *et al.*, 2019).

d) Population structure in chimpanzees

The sequencing of new chimpanzee genomes from a known geographical origin has proven effective for the discovery of a substantial amount of new genetic diversity (de Manuel *et al.*, 2016), highlighting the importance of sampling strategies to target different geographical regions. Importantly, this approach allowed to explore to which extent genetic information correlates with geographical origin, without samples from known origin such analysis would not be possible. This idea had previously been explored with human populations, where it is seen that among Europeans there is a close correspondence between genetic composition and geographic distances (Figure 8) (Novembre *et al.*, 2008).

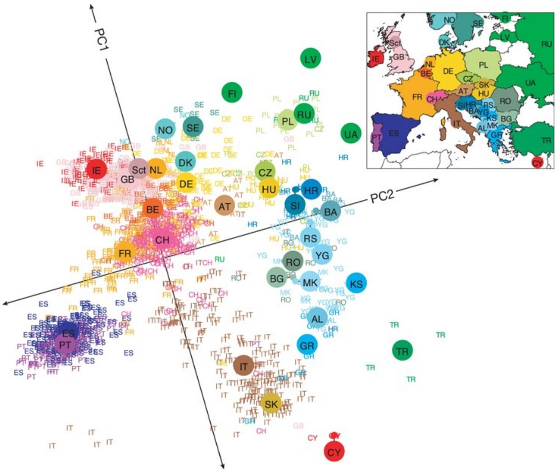


Figure 8. "Genes mirror geography". Principal component analysis of 1,387 Europeans based on genetic data, where there is high similarity to the geographic map of Europe. Adapted from (Novembre *et al.*, 2008).

Similar to what was found in Europeans, population clustering analysis revealed local stratification in central and eastern chimpanzees (de Manuel *et al.*, 2016) (Figure 9). It was not possible to assess population structure in western chimpanzees probably due to their low genetic diversity compared to central and eastern chimpanzees. Also, the limited number of geolocalized samples in both western and Nigeria-Cameroon chimpanzees precluded the determination of geographic and genetic differentiation patterns, although similar stratification would be expected with broader sampling. The described patterns found in eastern and central chimpanzees were then confirmed by chromosome 21 capture resequencing of chimpanzee fecal samples with known GPS coordinates (triangles in Figure 9).

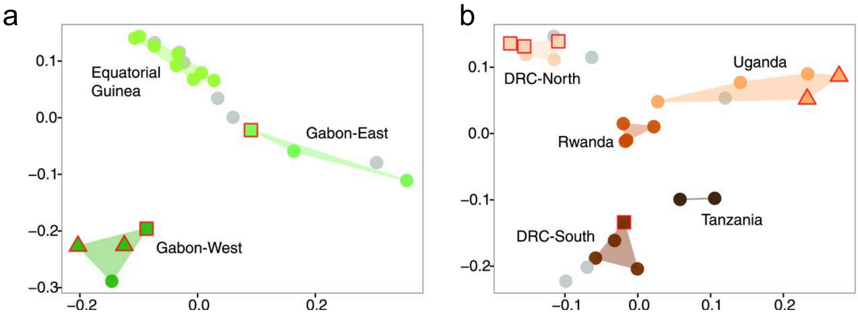


Figure 9. Population clustering of central (a) and eastern (b) chimpanzees correlates with their geographical origin. Fecal samples are shown in triangles. Adapted from de Manuel et al. (2016).

These results highlight the importance of knowing the provenance of the samples used for population studies to correctly describe genetic variation in chimpanzee populations and how it is distributed along their geographical range.

The ability to predict the geographical origin of a sample from its genetic data can be extremely valuable for the conservation of declining populations. In principle, it would be possible to determine the origin of confiscated animals and thus help to locate hotspots of poaching activity (as has been done in elephants (Wasser *et al.*, 2015)). However, for this to be applied into conservation action plans, we first need to have an extensive catalogue of genetic diversity from populations distributed along the extant geographical range of the chimpanzees. Sampling in previous studies was sparse, so a dense sampling effort, precisely in western (which have the lowest genetic diversity) and Nigeria-Cameroon chimpanzees for which this information was lacking, should be a priority. I will explore this approach in chapter 3.3.

To sum up, the catalogue of genetic diversity has not only shed light into past demographic events, current population diversity and structure, but also it can become a valuable tool to design conservation strategies.

1.4. Conservation genetics

The potential of genetics to be applied in conservation was initially suggested more than 40 years ago (Frankel, 1974). Since then, genetics have proved to be an important tool in conservation of threatened species with an entire field devoted to it. Conservation genetics is a discipline that involves the application of evolutionary and molecular genetics to biodiversity conservation (Frankham, Ballou and Briscoe, 2010; Allendorf, Luikart and Aitken, 2012). First, conservation genetics aims to understand the consequences of habitat loss and fragmentation into the population by analyzing their genetic diversity and fitness. And second, it aims to implement genetic tools to design and evaluate conservation plans (Angeloni *et al.*, 2012).

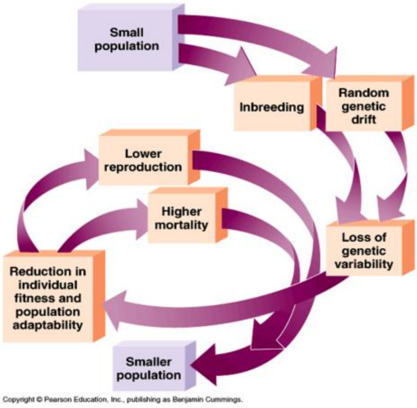

Small population
Inbreeding
Random genetic drift
Lower reproduction
Higher mortality
Loss of genetic variability
Reduction in individual fitness and population adaptability
Smaller population
Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Figure 10. Events in the extinction vortex. Adapted from (Campbell and Reece, 2008).

One of the focuses of conservation genetics is the study of the extinction risk in small and isolated populations, because genetic threats have a greater influence on them. Together with stochastic demographic and environmental events, an extinction vortex might be triggered (Figure 10): populations that are small in size tend to be more susceptible to random genetic drift and inbreeding, reducing their genetic diversity and fitness, which at the same time reduces even more their population size and eventually results in extinction (Blomqvist *et al.*, 2010).

The consequence of both genetic drift and inbreeding is a reduction of genetic diversity. Genetic drift is the random fluctuation of allele frequencies over time with the eventual fixation or loss of alleles. This effect is higher when populations have small effective population size, since any fluctuation in allele frequency can easily get to fixation or loss in a shorter amount of time (Hartl and Clark, 2007).

Inbreeding is defined as the mating between closely related individuals, which leads to an increased frequency of homozygotes in the population, thus increasing the probabilities to carry identity by descent (IBD) loci (Wright, Tregenza and Hosken, 2008). The consequences of the increased homozygosity and the fixation of deleterious alleles (genetic load) are the reduction of fitness and short-term viability of the population, a phenomenon described as inbreeding depression (Charlesworth and Charlesworth, 1999; Frankham, 2005; Ouborg *et al.*, 2010).

Genetic data has the potential to provide insights on diverse areas of conservation biology, including the identification of potential extinction risks associated with demographic changes and inbreeding.

Besides, genetic data can also be useful for resolving taxonomic uncertainties, kinship and measuring population history, genetic connectivity between populations, population substructure, population size, disease risk and hybridization events (Shafer *et al.*, 2015). This knowledge can then be applied in the management of captive and wild populations to minimize inbreeding and loss of genetic diversity. Also, in the current context of climate change, human habitat destruction and biodiversity loss, the integration of genetics into a broader context taking into consideration demographic and environmental variables is a good strategy to monitor changes in genetic diversity, identify extinction risks and compare and apply suitable conservation strategies (Frankham, 2010).

1.4.1. From conservation genetics to conservation genomics

Traditionally, conservation genetics has used a few number of markers, including allozymes, mtDNA and microsatellites (Frankham, Ballou and Briscoe, 2010). The rise of genomic resources is offering new opportunities for conservation by sequencing whole genomes. With high-density markers across the entire genome we can have a much broader representation of the genetic variation within individuals and populations (Allendorf, Hohenlohe and Luikart, 2010). In this scenario is where the new field of conservation genomics thrives, with the idea that genome-wide data will provide better resources for the species protection and conservation (Supple and Shapiro, 2018).

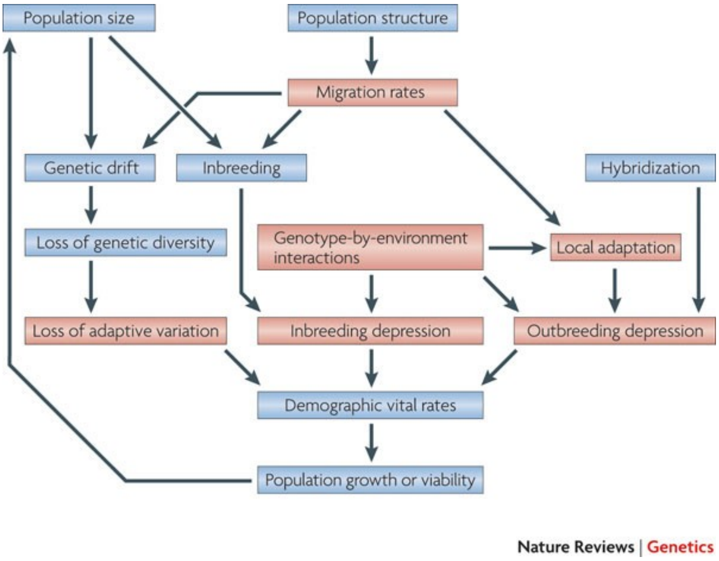


Figure 11. Comparison of factors which can be studied by traditional conservation genetics (blue) and conservation genomics (red). While conservation genetics can provide direct estimates of some factors, conservation genomics can address a wider range of factors and provide more precise estimates than traditional markers. Adapted from (Allendorf, Hohenlohe and Luikart, 2010).

Genomic approaches have stirred much expectation within the conservation community with the promise to address questions unanswered with traditional neutral markers, such as which loci are involved in adaptation, as well as to provide greater resolution and accuracy on demographic parameters (Figure 11) (Allendorf, Hohenlohe and Luikart, 2010; Ouborg *et al.*, 2010; Steiner *et al.*, 2013; McMahon, Teeling and Höglund, 2014; Shafer *et al.*, 2015; Garner *et al.*, 2016; Supple and Shapiro, 2018; Funk *et al.*, 2019). Using genome-wide data rather than a few markers, could result in different conservation recommendations (Supple and Shapiro, 2018).

Still, traditional genetic markers might continue to be the most economic and efficient solution to answer particular conservation biology questions (Kristensen *et al.*, 2010; Steiner *et al.*, 2013), especially in remote locations where using genomic technologies is not currently feasible (Flanagan *et al.*, 2018).

In the next sections, I will show specific scenarios where whole genome data can address fundamental evolutionary biology questions not fully resolved using traditional methods. However, the higher cost of sequencing compared to traditional methods, the need for quality reference genomes and the requirement of big data processing and storage resources still limits the advance of conservation genomics.

a) Conservation genomics opportunities

Whole genome data allows to increase the statistical power, accuracy and resolution in population genetics analyses (Allendorf, Hohenlohe and Luikart, 2010). It also can be used to resolve unknown phylogenetic relationships and delimitate conservation units, comprehend the genetic composition of present-day species and the events that lead to their present point. As well as this, it can determine loci responsible for speciation and local adaptation and understand the ultimate consequences of inbreeding depression in fitness

(Ouborg *et al.*, 2010; McMahon, Teeling and Höglund, 2014; Fuentes-Pardo and Ruzzante, 2017).

Defining conservation units (CU)[1] is the basic framework to support law enforcement and to allocate resources for conservation (Supple and Shapiro, 2018). The most common categorization is usually based on evolutionary significant units (ESU)[2] (Funk *et al.*, 2012). Although their correct identification is key for the successful implementation of conservation plans, complex evolutionary histories involving admixture, hybridization and introgression can make delineating ESUs with traditional markers difficult (Supple and Shapiro, 2018). For that, genomic data can be a good resource to robustly reconstruct evolutionary relationships among populations and detect hybridization events (Fuentes-Pardo and Ruzzante, 2017).

Genomics can also be a powerful tool to shed light into past population demographic changes and historical effective population sizes, with more resolution and accuracy than traditional markers (Fuentes-Pardo and Ruzzante, 2017; Supple and Shapiro, 2018). Events such as population bottlenecks, isolation, migrations and expansions are of high interest for conservation to fully comprehend the genetic composition of present-day populations (Fuentes-Pardo and Ruzzante, 2017). These events change the allele frequencies in the population and leave signatures in the genomes. In a bottleneck, population sizes reduce drastically in a short period of time with the subsequent decrease of standing genetic diversity, most importantly in those variants at low frequency (Hartl and Clark, 2007). The smaller the population, the more susceptible it is to the fixation of deleterious variants. However, if the population has been small for a long period of time, efficient

[1] CU is a population unit identified within species that is used to help guide management and conservation efforts (Fraser and Bernatchez, 2001)

[2] ESU is a population or group of populations that warrant separate management or priority for conservation because of high genetic and ecological distinctiveness (Funk *et al.*, 2012)

purging selection can reduce the otherwise expected high genetic load (van der Valk *et al.*, 2019).

In a continuous population without ecological or geographical barriers, genetic diversity follows an isolation-by-distance pattern. But when barriers are found, the reduction of gene flow leads to subdivision. Genomic data can enable high-resolution analyses to detect subtle population structure in nearby populations that can have big conservation implications (Steiner *et al.*, 2013). Also, with the increasing fragmentation of the habitat, populations are becoming more isolated. When migration is reduced, gene flow stops, and the allele frequencies among subpopulations become increasingly different. The resulting genetic differentiation could be due to selection or random genetic drift, the later having greater effect if the population is small.

The ultimate consequence of the aforementioned demographic events (population structure and bottlenecks) is a reduction in heterozygous genotypes and thus a reduction of genetic diversity (Hartl and Clark, 2007). In that regard, conservation genomics aims to maintain the genetic diversity levels and avoid the consequences of inbreeding depression in vulnerable populations. Therefore, understanding these two processes is highly relevant (Allendorf, 2017). In the case of inbreeding depression, genomics provide tools to time inbreeding events and also estimate the deleterious fitness consequences (Supple and Shapiro, 2018).

Finally, another main contribution of conservation genomics is the potential to identify genomic regions involved in adaptation to local environments (McMahon, Teeling and Höglund, 2014; Fuentes-Pardo and Ruzzante, 2017). Adaptive potential might be extremely relevant for the fitness of those populations and hence it should be fully considered in the evaluation of long-term extinction risks as well as restoration and reintroduction planning. If local adaptation is not taken into consideration, the reintroduction process may be less successful (Flanagan *et al.*, 2018; Funk *et al.*, 2019).

b) Limitations and challenges for conservation genomics

Conservation biology that involves threatened wild species is usually limited by the availability of samples (Steiner *et al.*, 2013). If captive populations are available, one solution is to use blood or tissue samples whenever those individuals have a veterinary check-up. These samples yield high DNA quality, although captive populations might not be a good representation of their wild counterpart (Prado-Martinez *et al.*, 2013). On the contrary, to avoid the disturbance of endangered wild animal species, sampling usually relies on non-invasive (NI) samples (feces, hair, urine…) which usually have lower DNA quality and quantity (see section 1.5.) (McMahon, Teeling and Höglund, 2014).

The advances achieved in NGS have greatly reduced the cost of whole genome sequencing (section 1.3.2.). However, this cost might still be too expensive for population studies that require the sequencing of numerous individuals. Also, most conservation projects have limited budgets and often prioritize other aspects, such as the creation of protected areas, awareness and education campaigns, community development and veterinary interventions (Caldecott and Miles, 2005). Sometimes, there is a huge tradeoff between number of samples and number of loci to be sequenced (Supple and Shapiro, 2018). To reduce sequencing costs, NGS approaches have been developed to perform SNP discovery in a reduced proportion of the genome. Those methods involve RAD-seq (sequence DNA adjacent to restriction enzyme cut sites) (Andrews *et al.*, 2016) and target DNA sequencing (capture target sequences with complementary probe hybridization) (see section 1.6.) (Grover, Salmon and Wendel, 2012).

Besides cost, whole-genome resequencing studies require the availability of reference genomes, which has limited its implementation to non-model species, as well as significant computational resources, storage and bioinformatic knowledge for the analysis (Shafer *et al.*, 2015; Fuentes-Pardo and Ruzzante, 2017).

All these factors challenge the integration of genomics into conservation, with academic research and policy-practitioner communities operating in largely separated groups (Shafer *et al.*, 2015; Taylor, Dussex and van Heezik, 2017). However, the fast pace of computation and sequencing development, the further reduction in costs and the development of new NGS techniques (i.e. target capture and RAD-seq) are making genomics an accessible tool for conservation managers, with already many examples of successful applications (Hoelzel, 2015; Garner *et al.*, 2016).

1.4.2. Translation of genetics and genomics into chimpanzee conservation

The usage of a few nuclear regions and microsatellites has been essential to elucidate the genetic diversity, date split times and estimate population size changes and structure in chimpanzee populations (Fischer *et al.*, 2004; Becquet *et al.*, 2007; Becquet and Przeworski, 2007; Gonder *et al.*, 2011). Later on, whole genome sequencing studies have complemented these findings (section 1.3.3.) (Prado-Martinez *et al.*, 2013; de Manuel *et al.*, 2016). These studies are the starting point to understand the species biology so they can be translated into conservation (Frankham, 2010). The application of genetics into the study of wild chimpanzee populations is usually based on non-invasive (NI) samples (feces, hair, urine, saliva…). Although they carry some complexity (explored in section 1.5. of this thesis), it is possible to extract DNA and genotype genetic markers (Vigilant and Guschanski, 2009). This approach has been applied for molecular censusing to assess population sizes and structure. Direct counts are usually impossible, so indirect methods such as nest counts or collecting NI samples are an effective way to obtain an accurate population estimate (Inoue *et al.*, 2007; Arandjelovic *et al.*, 2011; McCarthy *et al.*, 2015). Also, NI sampling allows to track individual chimpanzees after reintroduction efforts (Goossens *et al.*, 2003) and detect and monitor outbreaks of infectious diseases (Kaur *et al.*, 2008).

Molecular tracking of populations can be used to infer dispersal and migration events. Comparing patterns of variation from uniparental markers can inform about sex-biased dispersal, which for chimpanzees results in strong signal of male philopatry (Langergraber *et al.*, 2007). Furthermore, describing connectivity, isolation and barriers to dispersal are fundamental to the understanding of population structure. In spite of extensive habitat fragmentation, corridors of dispersal between chimpanzee communities have been described (McCarthy *et al.*, 2015). On the other hand, rivers are thought to be a strong barrier to gene-flow (Gonder, Disotell and Oates, 2006; Becquet *et al.*, 2007).

The description of population genetic subdivisions and the characterization of isolated or distinct wild populations has proven useful to be applied in direct conservation measures. In elephants, the availability of genomic resources and the testing of confiscated tusks has allowed the detection of hotspots of poaching (Wasser *et al.*, 2015). Another study has evaluated the consequences of reintroduction without proper genetic assessment in orangutans, resulting in outbreeding depression and admixture (Banes, Galdikas and Vigilant, 2016). Therefore, they strongly advise that future reintroduction should follow international guidelines that require genetic assessment (IUCN/SSC, 2013).

On the one hand, these direct conservation measures could be implemented in conservation plans of wild chimpanzee populations after compiling an extensive genome-wide dataset linked to geographical origin and performing a proper assessment and implementation of a methodology to be used on site (de Manuel *et al.*, 2016).

On the other hand, for the purpose of *ex situ* chimpanzee management, knowing the ancestry and inbreeding of chimpanzees is relevant for planning captive breeding (Hvilsom *et al.*, 2013). However, traditional markers might not have the necessary resolution to disentangle several generations of hybridizations in the EEP breeding population. Thus, a precise genome-wide characterization of captive populations would resolve this situation. Also, if the extant genetic diversity in the wild is catalogued, *ex situ* conservation programmes can attempt to preserve as much genetic diversity as possible in captive populations, which might serve as a reservoir for eventual reintroduction and supplementation to wild populations (Traylor-Holzer, 2011; Lacy, Traylor-Holzer and Ballou, 2013; Traylor-Holzer, Leus and Bauman, 2019).

1.5. Non-invasive samples to study wild populations

The majority of great ape population genomic studies (section 1.3.3.) use blood or tissue as a source of DNA (Prado-Martinez *et al.*, 2013; Xue *et al.*, 2015; de Manuel *et al.*, 2016), usually from captive populations (sanctuaries, zoos…). Obtaining invasive samples from the wild would require trapping or darting the animal causing physical distress and harm, elevating the risk of infection, altering their behaviour and even causing the animal's death (Morin *et al.*, 1993; Taberlet, Waits and Luikart, 1999). Moreover, the international transportation of invasive samples for endangered animals is regulated by CITES (Convention on International Trade in Endangered Species), which adds administrative complexity to genetic research (Perry *et al.*, 2010).

Therefore, the usage of NI samples as sources of DNA to study wild endangered populations is gaining importance since their collection is done without disturbance or harm to the animal. Moreover, NI samples from the wild provide information about geographical origin and are a good representation of the extant genetic diversity (Vigilant and Guschanski, 2009).

However, DNA isolated from NI samples is often highly degraded and in low quantities (Taberlet, Waits and Luikart, 1999; Perry *et al.*, 2010). Fecal samples are typically composed of low proportions of host or endogenous DNA (eDNA) (Perry *et al.*, 2010), genetic material from the host's microbiota and from species living in the environment where the sample was collected (i.e., exogenous DNA) (Hicks *et al.*, 2018). Moreover, NI samples can contain PCR inhibitors (Morin *et al.*, 2001) and are usually collected in warm and humid environments which accelerates DNA degradation (Hernandez-Rodriguez *et al.*, 2018).

This inherent complex nature of NI samples and the difficulties to extract good quality DNA has for years precluded the usage of these samples for genomic studies. This is why, studies using NI samples have traditionally been restricted to neutral markers or genetic loci such as autosomal and Y chromosome microsatellites (Arandjelovic *et al.*, 2011; Inoue *et al.*, 2013; Fünfstück *et al.*, 2014; Orkin *et al.*, 2016), autosomal regions (Thalmann *et al.*, 2007; Hans *et al.*, 2015) and the mitochondrial genome (Thalmann, Hebler, *et al.*, 2004; Thalmann, Serre, *et al.*, 2004).

The direct application of NGS methods to NI samples would not be economically feasible due to their low proportions of eDNA and the presence of PCR inhibitors (Morin *et al.*, 2001; Perry *et al.*, 2010). In chimpanzee fecal samples the proportion of eDNA can be highly variable (Hernandez-Rodriguez *et al.*, 2018) but is usually below 2% (White *et al.*, 2019).

The emergence of target enrichment methods (section 1.6.) to selectively recover genomic loci of interest has allowed for a more cost-effective use of NGS on complex and degraded samples, such as ancient DNA samples (Carpenter *et al.*, 2013) but also feces (Perry *et al.*, 2010; Snyder-Mackler *et al.*, 2016; van der Valk *et al.*, 2017). Still, the difficulties to work with fecal samples have motivated the appearance of technical studies describing the usage of NI samples for the genomic study of wild chimpanzees (Hernandez-Rodriguez *et al.*, 2018; White *et al.*, 2019). A precise scenario to improve the retrieval of genomic information from fecal samples through target capture is explored in chapter 3.2. of this thesis.

1.6. Target enrichment methods

Target enrichment methods allow for a reduction in sequencing cost by selecting and sequencing only genomic loci of interest (Mertes *et al.*, 2011). There are several target enrichment strategies according to their molecular reaction principle (Figure 12) (Mamanova *et al.*, 2010):

a. PCR (Polymerase Chain Reaction) amplification: PCR is directed towards the regions of interest with primers. It can be multiplexed using several primer pairs in a single reaction to generate multiple amplicons.
b. Selective circularization, also called molecular inversion probes (MIPs): probes are designed to be complementary to a target region at their edges and linked with an approximately 40-bp long connection sequence. When annealed to the DNA, the probe edges become ligated and thus circularized. Afterwards, all non-circular DNA is removed.
c. Hybridization capture: fragmented DNA prepared as a library is hybridized to probes complementary to the regions of interest. The hybridization reaction can be performed either on solid support (microarrays) or in-solution.

These methods have been applied in many studies from different disciplines such as evolutionary biology, ecology, population genetics and conservation targeting specific SNPs (Patterson *et al.*, 2012; Haak *et al.*, 2015; Olalde *et al.*, 2018), mitochondrial genome (Fu *et al.*, 2013; van der Valk *et al.*, 2018), exomes (Castellano *et al.*, 2014), whole chromosomes (Perry *et al.*, 2010) or entire genomes (Carpenter *et al.*, 2013; Snyder-Mackler *et al.*, 2016; Cruz-Dávalos *et al.*, 2018). Since in-solution hybridization capture is the method used in all three projects presented in this thesis I have included the protocols in the Annex of this thesis.

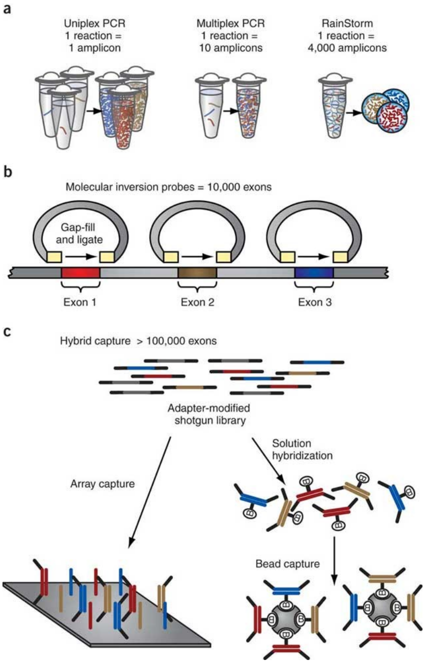


Figure 12. Target-enrichment strategies for next-generation sequencing: a) PCR based; b) MIP-based and c) hybrid capture–based. Adapted from (Mamanova *et al.*, 2010).

2. Objectives

The main objective of this thesis is to explore how genomics and non-invasive samples can provide insights into chimpanzee genetic diversity, either in captivity or in the wild.

For each result chapter there are specific objectives.

In the first project (chapter 3.1.) we explore conservation management outside chimpanzee natural range (*ex situ*) to:

- Characterize the captive chimpanzee population assessing their ancestry and inbreeding.
- Determine the geographical origin of rescued chimpanzees from illegal trade.
- Evaluate the captive chimpanzee population as a potential source to complement conservation initiatives in the wild.

In the second project (chapter 3.2.) we analyze technical aspects of target capture methods with fecal samples to:

- Provide a comprehensive exploration of target enrichment efficiency for very low endogenous DNA fecal samples.
- Explore how library complexity may be increased without repeating DNA extractions and generating new libraries.
- Provide guidelines to ensure the maintenance of the captured molecule diversity or library complexity.

In the third project (chapter 3.3.) we generated a genomic dataset from geolocalized non-invasive samples from the whole extant range of chimpanzees to:

- Describe chimpanzee genetic diversity and population structure as well as demography, migration and isolation patterns between different chimpanzee communities.

- Generate a fine-scale geographic and genetic variation map by increasing local geographic resolution and discovering new chimpanzee diversity.
- Explore the usage of the developed geogenetic map as a conservation tool to precisely infer the geographical origin of chimpanzees.

3. Results

3.1. Targeted conservation genetics of the endangered chimpanzee

Peter Frandsen*, **Claudia Fontsere***, Svend Vendelbo Nielsen, Kristian Hanghøj, Natalia Castejon-Fernandez, Esther Lizano, David Hughes, Jessica Hernandez-Rodriguez, Thorfinn Sand Korneliussen, Frands Carlsen, Hans Redlef Siegismund, Thomas Mailund, Tomas Marques-Bonet & Christina Hvilsom.

Targeted conservation genetics of the endangered chimpanzee.

Heredity 125, 15–27 (2020).

https://doi.org/10.1038/s41437-020-0313-0

Heredity
https://doi.org/10.1038/s41437-020-0313-0

the geneticssociety

ARTICLE

Targeted conservation genetics of the endangered chimpanzee

Peter Frandsen[1,2] · Claudia Fontsere[3] · Svend Vendelbo Nielsen[4] · Kristian Hanghøj[2] · Natalia Castejon-Fernandez[4] · Esther Lizano[3,5] · David Hughes[6,7] · Jessica Hernandez-Rodriguez[3] · Thorfinn Sand Korneliussen[8,9] · Frands Carlsen[1] · Hans Redlef Siegismund[2] · Thomas Mailund[4] · Tomas Marques-Bonet[3,5,10,11] · Christina Hvilsom[1]

Received: 19 February 2019 / Revised: 31 March 2020 / Accepted: 31 March 2020

Abstract

Populations of the common chimpanzee (*Pan troglodytes*) are in an impending risk of going extinct in the wild as a consequence of damaging anthropogenic impact on their natural habitat and illegal pet and bushmeat trade. Conservation management programmes for the chimpanzee have been established outside their natural range (ex situ), and chimpanzees from these programmes could potentially be used to supplement future conservation initiatives in the wild (in situ). However, these programmes have often suffered from inadequate information about the geographical origin and subspecies ancestry of the founders. Here, we present a newly designed capture array with ~60,000 ancestry informative markers used to infer ancestry of individual chimpanzees in ex situ populations and determine geographical origin of confiscated sanctuary individuals. From a test panel of 167 chimpanzees with unknown origins or subspecies labels, we identify 90 suitable non-admixed individuals in the European Association of Zoos and Aquaria (EAZA) Ex situ Programme (EEP). Equally important, another 46 individuals have been identified with admixed subspecies ancestries, which therefore over time, should be naturally phased out of the breeding populations. With potential for future re-introduction to the wild, we determine the geographical origin of 31 individuals that were confiscated from the illegal trade and demonstrate the promises of using non-invasive sampling in future conservation action plans. Collectively, our genomic approach provides an exemplar for ex situ management of endangered species and offers an efficient tool in future in situ efforts to combat the illegal wildlife trade.

Introduction

In an era of human-induced acceleration of species loss, often referred to as the sixth mass extinction era (Ceballos et al. 2015), conservation efforts to save endangered species are calling for novel approaches to mitigate the ongoing extinction crisis.

Since the discovery of the common chimpanzee (*Pan troglodytes*), humans have been drawn to this charismatic species. Despite our fascination, human activities have led to a drastic decline in the population size of the chimpanzee. In the last two decades, chimpanzees have been listed as 'Endangered' at the species level on the IUCN Red List, with one of the four recognised subspecies, the western chimpanzee (*P. t. verus*) being listed as 'Critically Endangered' in the latest assessment (Humle et al. 2016). Human encroachment on the natural range of the chimpanzee has further caused an intensified conflict between humans and chimpanzees (Hockings et al. 2015). One by-product of the human wildlife conflicts has been a rise in opportunistic trafficking of chimpanzees, which, in recent years has become more organised and systematic (Stiles et al. 2013). Besides wildlife trade, other continuous threats including habitat destruction, poaching for local consumption, and human linked disease outbreaks has led to a drastic decline in the

These authors contributed equally: Peter Frandsen, Claudia Fontsere

These authors jointly supervised this work: Tomas Marques-Bonet, Christina Hvilsom

Associate Editor: Xiangjiang Zhan

Supplementary information The online version of this article (https://doi.org/10.1038/s41437-020-0313-0) contains supplementary material, which is available to authorised users.

✉ Peter Frandsen
pef@zoo.dk

✉ Claudia Fontsere
claudia.fontsere@upf.edu

Extended author information available on the last page of the article

Published online: 28 April 2020

wild chimpanzee populations (Humle et al. 2016). Together, these threats emphasise the importance of a '*One Plan Approach*' conservation programme linking in situ and ex situ efforts (Traylor-Holzer et al. 2019) to prevent the predicted extinction of chimpanzees within the current century (Estrada et al. 2017).

Outside Africa, several regional chimpanzee conservation programmes exist, with the largest being the European Association of Zoos and Aquaria (EAZA) Ex situ Programme (henceforth EEP). The EEP targets the subspecies level and today, breeding programmes for two of the four recognised subspecies, the western chimpanzee (*P. t. verus*) and the central chimpanzee (*P. t. troglodytes*) have been established (Carlsen and de Jongh 2019). The primary aim of the EEP is to safeguard the survival of healthy self-sustaining populations targeting the taxonomical level of subspecies (Carlsen and de Jongh 2019). The extant EEP populations consist of wild founders and descendants thereof. However, in times before high resolution genetic technologies were available and even in its early development, knowledge of subspecies labels and relatedness between founders were inaccurate and has led to admixture of subspecies in the captive population (Hvilsom et al. 2013). Early attempts to add a genetic layer to the EEP management has confirmed that knowledge of subspecies ancestries, inbreeding and relatedness estimates are instrumental to preserve genetic diversity in captive populations (Hvilsom et al. 2013). Yet, most recent attempts based on microsatellite markers (Hvilsom et al. 2013), did not have the necessary resolution or predictive power to disentangle several generations of hybridisations in the EEP breeding population. Although we still do not know its full extent, hybridisation between neighbouring subspecies of chimpanzees has been shown to occur in the wild (Hvilsom et al. 2013; Prado-Martinez et al. 2013; de Manuel et al. 2016) and therefore, it is not unlikely that some founders in the EEP harbour shared ancestries from more than one subspecies. The current strategy in the EEP targets un-admixed breeding individuals and with the current methods, it is impossible to tell if small admixture proportions arose from an early ex situ hybridisation event followed by several generations of backcrossing or from a naturally admixed founder. Therefore, founders are potentially being wrongfully excluded from the breeding programme due to their admixed ancestry.

The scenario outlined above, is by no means exclusive to captive management of chimpanzees but extends to practically any ex situ management programme of populations based on wild born founders with a taxonomical subdivision. When morphology alone is insufficient in taxonomical delimitation between subspecies or the targeted conservation units, genetic resources becomes increasingly important. Yet, the choice of genetic resource is not always trivial. In response to a growing availability of different types of genetic resources with widely different applications, several studies have tried to develop guidelines based on the management requirements (see e.g. Grueber et al. 2019; Norman et al. 2019).

As described, the complexities in EEP management of chimpanzees requires a new rigorous solution as previous attempts using either mitochondrial DNA, or microsatellites have proven insufficient. With a genome-wide set of ancestry informative markers, we predict that it will be possible to obtain the desired depth of predictive power to infer ancestries in the present and previous generations and classify individuals with shared ancestries as either descendants of admixed founders or *ex situ* hybrids. This could provide the foundation of a possible reassessment of the current management strategies under the EEP and in turn, allow for inclusion of wild born hybrids in the breeding programme if these are found to resemble the diversity of the species in the wild.

In their natural range, chimpanzees have become a commodity and organised illegal trade poses a serious threat to the species. Over the period from 2005 to 2011 a reported minimum of 643 chimpanzees were harvested from the wild for illegal trade activities (Stiles et al. 2013). However, extrapolations suggest that 20 times as many individuals have become victims of the illegal wildlife trade in that relatively short time span (Stiles et al. 2013). While most of the captured individuals are sold as bushmeat, a considerable number of mostly juvenile chimpanzees end up in the illegal pet trade. When conservation authorities confiscate illegally kept chimpanzees, they are placed at wildlife sanctuaries, often arbitrarily based on availability of space and proximity to the confiscation site. Whilst some of the rescued chimpanzees require specialised lifetime care, others may be successfully reintroduced into their natural habitats after extensive preparation (Beck et al. 2007). For chimpanzees destined to lifetime care, proper management planning requires knowledge about relatedness among sanctuary chimpanzees in order to set up family groups. In cases, were chimpanzees are suitable for reintroduction, knowledge of geographical origin is essential as several studies have shown lineage-specific adaptations in all four subspecies in their respective geographical ranges (e.g. Nye et al. 2018). In the first complete geo-referenced genomic map of the chimpanzee, de Manuel et al. (2016) portrayed a strong correlation between geographical origin and genetic diversity, where the former can be inferred solely based on the latter. Employing genetic testing at the site of confiscation (e.g. airports and transport hubs) would enable conservation authorities to infer geographical origin of confiscated individuals and with time, strive to facilitate a return of these individuals to a protected area in the region where they were captured. Alternatively, confiscated chimpanzees can be sent to a neighbouring sanctuary with housing capacity, where specialised care and rehabilitation can be provided, and if possible, future reintroduction can be planned. Genetic testing at an early stage of confiscation also has the potential to understand and help break trafficking routes and enable CITES authorities to track and enforce law control in situations where chimpanzees are housed in disreputable zoos and entertainment facilities. However, to be a practical tool in conservation, the genetic test needs to maximise the inference accuracy, require very little investment, and pose as little risk to animal health as possible. These requirements limit our choice of applicable data types. With a novel SNP array design where the level of genetic information is only surpassed by costly whole genome sequencing, we argue that our approach constitutes the most cost-efficient option for conservation management in situations where funding is often scarce and demands for rigorous solutions are high.

Using a selected panel of 59,800 targeted ancestry informative markers, we demonstrate the ability to infer robust estimates of ancestry in several generations of the EEP chimpanzee breeding population. We further show how this set of ancestry informative markers can be used to determine geographical origin of confiscated individuals and demonstrate how these methodologies can readily be applied to using non-invasive sampling. In combination, these methods harbour great potential for future global management plans for the chimpanzee and provides an important exemplar for management of endangered species in general.

Materials and methods

Samples

A total of 179 chimpanzee samples were collected and analysed in the present study (Supplementary File S1 SequencingStatistics.xlsx). For the purpose of cross-validation between sequencing batches and to test our methodology on non-invasive hair sampling, a number of individuals were sequenced in duplicates and triplicates, which lead to 167 unique individuals. 136 from the EEP population housed in 47 different European zoos and primate rescue and rehabilitation centres (Table S2), and 31 from eight sanctuaries across Africa (Table S3). To form a reference panel, we complemented the genotypes of EEP and sanctuary chimpanzees with whole genome data from 58 geo-referenced wild-born chimpanzees, representing the four chimpanzee subspecies, and additionally, one known admixed individual (*Ptv-Donald*) and one known descendant of wild born individuals (*Ptv-Clint*) (Prado-Martinez et al. 2013; de Manuel et al. 2016).

DNA extraction and library preparation

DNA was extracted using a standard phenol-chloroform protocol. Samples were quantified with a Qubit 2.0 fluorometer, Qubit® dsDNA BR Assay Kit (Thermo Fisher Scientific). DNA library preparation was carried out in three batches. For the first batch (24 samples) and the second batch (63 samples), extracted DNA was sheared with a Covaris S2 ultrasonicator using the recommended fragmentation settings to obtain a 350 bp insert size. For the third batch (92 samples) DNA was sheared using the recommended settings of Covaris S2 to obtain 200 bp insert size. The first batch of 24 libraries (with 6 more samples not used in this study) were prepared using 1.5 µg of DNA and the TruSeq DNA HT Sample Prep Kit (Illumina), following manufacturer's instructions and 14 cycles of polymerase chain reaction (PCR) amplification. The second batch of 63 samples (with 17 more samples not used in this study) were processed using 500 ng of starting DNA and following the custom dual-indexed protocol described by Kircher et al. (2012) and 12 cycles of PCR were done for indexing and amplification. The remaining 92 samples (with two more samples not used in this study) were processed using 200 ng of starting DNA following the BEST protocol (Carøe et al. 2018) with minor modifications (initial reaction volume was incremented up to 50 µl to accommodate a larger amount of starting DNA and 10 cycles of PCR amplification). For this third batch, we used inline barcoded short adaptors with the same seven nucleotide barcodes at the P5 and P7 adaptors. Clean-ups were done using homemade SPRI beads (Rohland and Reich 2012). Libraries were eluted in 25 µl of ddH_2O and quantified with an Agilent 2100 Bioanalyzer using a DNA 1000 assay kit.

Target capture design

We performed a target capture enrichment experiment using baits synthesised by Agilent Technologies. We targeted 59 800 autosomal sites that were ancestry informative markers and designed using the panTro4 genome. Marker selection was done using published chimpanzee genomes (Prado-Martinez et al. 2013) and by applying a sparse PCA method on 10 Mbp bins of the genomes (Lee et al. 2012). Variant sites were then weighted to identify the most informative markers for the first two principal components (PCs) and 200 AIMs were extracted per segment. The genome was binned to have an unbiased and evenly distributed sampling of the genome and to have enough resolution to provide estimates of ancestry in highly admixed individuals.

For target enrichment hybridisation, libraries were pooled equimolarly based on a library prep method to obtain a total of 19 pools (see Supporting Information for a detailed description of the targeted enrichment hybridisation). PCR amplification product was cleaned up using our homemade SPRI beads (Rohland and Reich 2012). Each enriched sample was then quantified on a NanoDrop, BioAnalyzer and then sequenced.

Fastq filtering and mapping

Libraries were sequenced on five lanes of a HiSeq 2500 ultra-high-throughput sequencing system, one lane for 24 chimpanzee samples, 2 lanes for 63 chimpanzee samples and 2 lanes for the remaining 92 samples. Inline barcoded libraries captured in the same pool (92 from Batch 3) were de-multiplexed using Sabre software v. 1.0 (https://github.com/najoshi/sabre).

Prior to mapping, paired-end reads were filtered to remove PCR duplicates using FASTUNIQ v. 1.1 (Xu et al. 2012) and adaptors (*Illuminaclip*) and low quality first five bases in a read (*Slidingwindow:5:20*) were trimmed using TRIMMOMATIC v. 0.36 (Bolger et al. 2014). Overlapping reads were merged with a minimum overlap of 10 bp and minimum length of final read to 50 bp, using PEAR v. 0.9.6 (Zhang et al. 2014). Then, reads were mapped using BWA v. 0.7.12 (Li and Durbin 2009) to the Hg19 reference genome (GRCh37, Feb.2009 (GCA_000001405.1)). PCR duplicates were removed using PICARDTOOLS v. 1.95 (http://broadinstitute.github.io/picard/) with the *MarkDuplicates* option. Further filtering of the reads was done to discard secondary alignments and reads with mapping quality lower than 30 using SAMTOOLS v. 1.5 (Li et al. 2009). We then filtered for the targeted space (4 bp around the selected SNP) using BEDTOOLS intersect v. 2.16.2 (Quinlan and Hall 2010).

The total aligned reads were calculated by dividing the number of uniquely mapped reads (the remaining reads after removing duplicates) by the number of production reads. The on-target aligned reads were calculated by dividing the target filtered reads by the production reads. Then, the total coverage was calculated by dividing aligned bases by the length of the assembly (Hg19) and the target effective coverage dividing the on-target bases by the targeted genomic space. Finally, the enrichment factor of the capture performance was calculated by taking the ratio between the on-target reads by total mapped reads over the target size by genome size.

Variant calling

Variant discovery was performed using GATK '*Unified Genotyper*' (DePristo et al. 2011) for each sample independently with the following parameters *-out_mode EMIT_ALL_SITES -stand_call_conf* 5.0 *-stand_emit_conf* 5.0 *-A BaseCounts -A GCContent -A RMSMappingQuality -A BaseQualityRankSumTest*. Genotypes from each sample were combined in a single VCF using GATK '*CombineVariants*' (DePristo et al. 2011) with *-genotypeMergeOptions UNIQUIFY –excludeNonVariant* parameters. We also included the genotype information of available whole genome data of aforementioned 58 wild-born geo-referenced chimpanzees and *Ptv-Donald* and *Ptv-Clint* (Prado-Martinez et al. 2013; de Manuel et al. 2016). Unless differently stated in separate analysis, the variants with a depth of coverage less than 3, a quality score less than 30 (QUAL < 30), minor allele frequency (MAF) of 0.005 and a missingness rate of >60 % were removed using VCFTOOLS v. 0.1.12 (Danecek et al. 2011). We only kept the genotypes that were inside the target space by using the -bed option in VCFTOOLS v. 0.1.12 (Danecek et al. 2011).

Ancestry inference and inbreeding

We inferred proportions of shared ancestries in two approaches. First, to detect underlying genetic structure with a reduction of the dimensionality in the data, we performed a principle component analysis (PCA) using EIGENSOFT v. 6.1.3. (Price et al. 2006). All samples were included without pruning of sites in linkage disequilibrium or MAF, in order to avoid exclusion of fixed sites between populations. Analyses on shared ancestry in ex situ and sanctuary populations were done with reference to the genetic structure in the wild born individuals with ADMIXTURE v. 1.2 (Alexander et al. 2009). To avoid any bias introduced from a joint analysis with related individuals, each of the 167 unique individuals from the EEP and sanctuary populations were analysed separately one by one against a reference panel of all wild born individuals. After applying a MAF filter (--maf 0.05) in PLINK v. 1.07 (Purcell et al. 2007) to exclude sites polymorphic in only one individual, a set of 45,542 sites where kept for analysis. Each analysis of ADMIXTURE v. 1.2 (Alexander et al. 2009) was iterated 100 times under an EM optimisation algorithm and termination criteria of a log-likelihood increase of 10^{-5} between iterations. A value of $K = 4$ was chosen to obtain clusters in line with the four recognised subspecies of chimpanzees. To assess convergence, the 100 iterations were evaluated to ensure that iterations did not differ by more than 1 log-likelihood value.

For each of the individuals with admixture coefficients >0.99, we applied NGSRELATEv2 (Hanghøj et al. 2019) to estimate pairwise relatedness and individual inbreeding coefficients based on population allele frequencies from each of the inferred admixture clusters, after excluding MAF < 0.05 (see Supplementary Information for details along with per population and global estimates of F_{IS}).

Hybrid classification

To further explore the ancestry sharing in the EEP and sanctuary individuals and to be able to differentiate shared ancestry originating from the founding individuals and EEP hybrids, we developed a hidden Markov model (available on GitHub http://gihub.com/svendvn/ImmediateAncestry) to allow for an inference of the posterior proportion of ancestries in the three immediate previous generations. In addition, we estimate where these immediate ancestors belong in the pedigree. For full documentation of the model, see Supplementary Information.

Re-assignment of geographical origin

We applied the methodology of ORIGEN (CRAN R package https://cran.r-project.org/web/packages/OriGen/index.html) as described by Rañola et al. (2014), to re-assign the geographical origin of confiscated sanctuary individuals. We applied the *FitOriGenModelFindUnknowns* parameter to the 1690 highest ranked informative markers to assign individual geographical origin onto the allele frequency surface, inferred from the wild born reference panel.

Non-invasive sampling

To test our targeted capture approach on non-invasively collected hair samples, we sequenced three individuals where we had both blood samples, whole genome reference data and hair samples. Hair samples were capture sequenced using the same methodology as described above for blood samples, except we added a pre-treatment step in the DNA extraction of hair samples to enhance lysis of keratin. Shared ancestry and geo-graphical origin was analysed as described above.

Results

Capture sequencing and variant calling

First we quantified and assessed the performance of our capture methodology in the selected targeted space. We wanted to ensure sufficient representation of the targeted genomic regions to reliably call the selected variants. In a total of five lanes of HiSeq2500 we obtained ~1000 million production reads, and on average, each sample received five million reads. After removing PCR duplicates and considering only primary alignments with a mapping quality higher than 30, we obtained an average of 3.6 million mapped reads (74.31%) per sample (Supplementary File S1). The average effective target coverage on the 59,800 autosomal SNPs was 21.69X with 12.91% of on-target reads (four base pairs around the targeted SNP, Supplementary File S1) which fulfilled our theoretical prediction of 20X. In terms of capture performance, this last statistic is an underestimate since the full length of the capture bait is 120 base pairs and in this analysis, we only considered the four base pairs around the targeted SNP. Still, we considered it to be more accurate since it is the true space where the informative SNP falls. Lastly, to summarise the performance of the capture methodology, we computed the enrichment factor that relates the number of aligned reads on the target space divided by the production reads, with the size of the target space to the size of the whole genome. The resulting enrichment factor of 89.31X reasserts the advantages of capture to ensure enough coverage for genotyping purposes (Supplementary File S1).

Considering all samples without overlap, we obtained a total of ~150,000 genotypes. However the average number of SNPs called per sample was 30,337 sites passing the filtering steps (MAF 0.05 and max-missing 0.6, after we excluded samples ‘12103’ and ‘12349’ due to low coverage). The maximum number of SNPs called in one individual was 51,952 and the minimum was 10,783 (Fig. S1). Among the variation found in western chimpanzees, only a third of these were polymorphic in the western chimpanzee (Table S1), yet, of the 46,260 polymorphic sites, 15,738 were private in the western chimpanzee (Fig. S2). For fixed sites, the western chimpanzee also had the highest number of private sites (Fig. S2). Among the four subspecies, the eastern chimpanzee had the highest total number of polymorphic sites, followed by the central chimpanzee, Nigerian-Cameroon chimpanzee, and western chimpanzee, respectively (Table S1).

Population structure, ancestry, and inbreeding

The major axes of variance in EEP and sanctuary individuals were explored with a PC analysis with reference to the panel of geo-referenced individuals with known subspecies label from Prado-Martinez et al. (2013) and de Manuel et al. (2016). The first PC (PC1) explained 70.49% of the variance in our data, separating the western chimpanzees from the three other subspecies in the reference panel (Fig. 1b). With 16.53 % of explained variance, PC2 separated the Nigerian-Cameroon chimpanzee, central chimpanzee, and eastern chimpanzee.

The majority of the 167 tested individuals from the EEP and sanctuary populations, clustered with either of the four reference populations, while a minor part of the individuals scattered in between the defined populations (Fig. 1b). The inferred ancestries from the ADMIXTURE analysis conveyed the same patterns of genetic population structure separating the geo-referenced individuals into four distinct

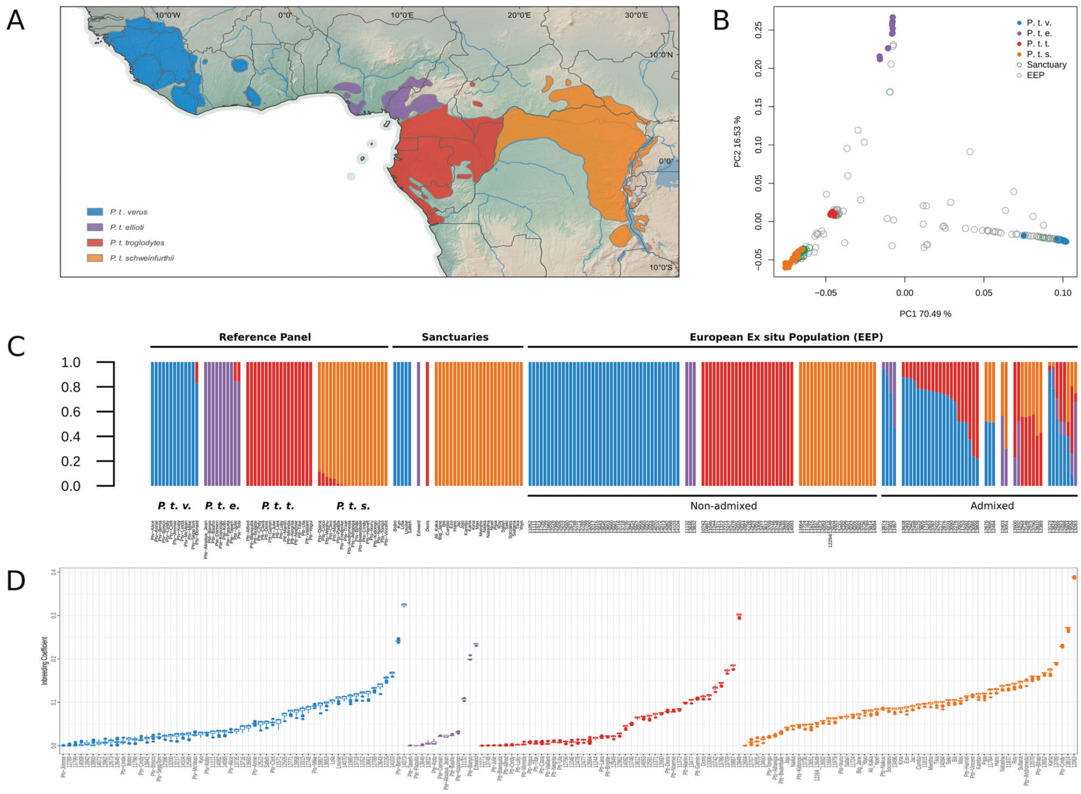


Fig. 1 Subspecies ancestry and inbreeding in wild and captive populations of chimpanzees. a Geographical distribution ranges of the four chimpanzee subspecies (IUCN 2015; QGIS 2018). **b** Population structure by principal component decomposition of sanctuary and the EAZA Ex situ Programme (EEP) populations with reference to wild born individuals. **c** Shared ancestry inferences of sanctuary and EEP individuals summarised from individual ADMIXTURE analysis against the reference panel of wild born individuals. Individuals from the reference panel are labelled with a subspecies ancestry prefix and known sample name in previous literature (Prado-Martinez et al. 2013; de Manuel et al. 2016), sanctuary individuals are labelled with common sample name identifiers, and individuals from the EEP are labelled by studbook number (Tables S2 and S3). **d** Individual inbreeding coefficients for all individuals with admixture proportions >0.99 in either of the four inferred clusters. Inbreeding estimates were estimated within each cluster independently. Clusters are colour labelled in accordance to (**a–c**).

clusters with varying degree of ancestry sharing between geographically neighbouring subspecies (Fig. 1c). With this as a reference, we assigned the EEP and sanctuary individuals into groupings in terms of their ancestry patterns of either non-admixed or hybrids with multiple components of ancestry. Of the 167 tested individuals, 121 could be confidently assigned as non-admixed (admixture proportion from one subspecies ≥ 0.99). All 31 sanctuary individuals were assigned to subspecies level without evidence of admixture, where five clustered with the western chimpanzee, one with the Nigerian-Cameroon chimpanzee, one with the central chimpanzee, and 24 with the eastern chimpanzee. In the EEP population, we inferred the majority of the 90 non-admixed individuals to belong to the western chimpanzee (41), three with the Nigerian-Cameroon chimpanzee, 25 with the central chimpanzee, and 21 with the eastern chimpanzee. Of the remaining 46 EEP individuals, 38 were inferred to be hybrids with two ancestry components while the last eight had three ancestry components.

Of all the individuals from the EEP, sanctuary, and the reference panel with admixture coefficients >0.99, relatedness estimates were low (Figs. S3–S6) while we identified eight individuals with inbreeding coefficients above 0.2 (Fig. 1d). Within these eight individuals, all four subspecies were represented, as were wild and captive born chimpanzees.

Hybrid classification

To explore ancestry patterns in the previous three generations, we ran our ancestry classification model going back

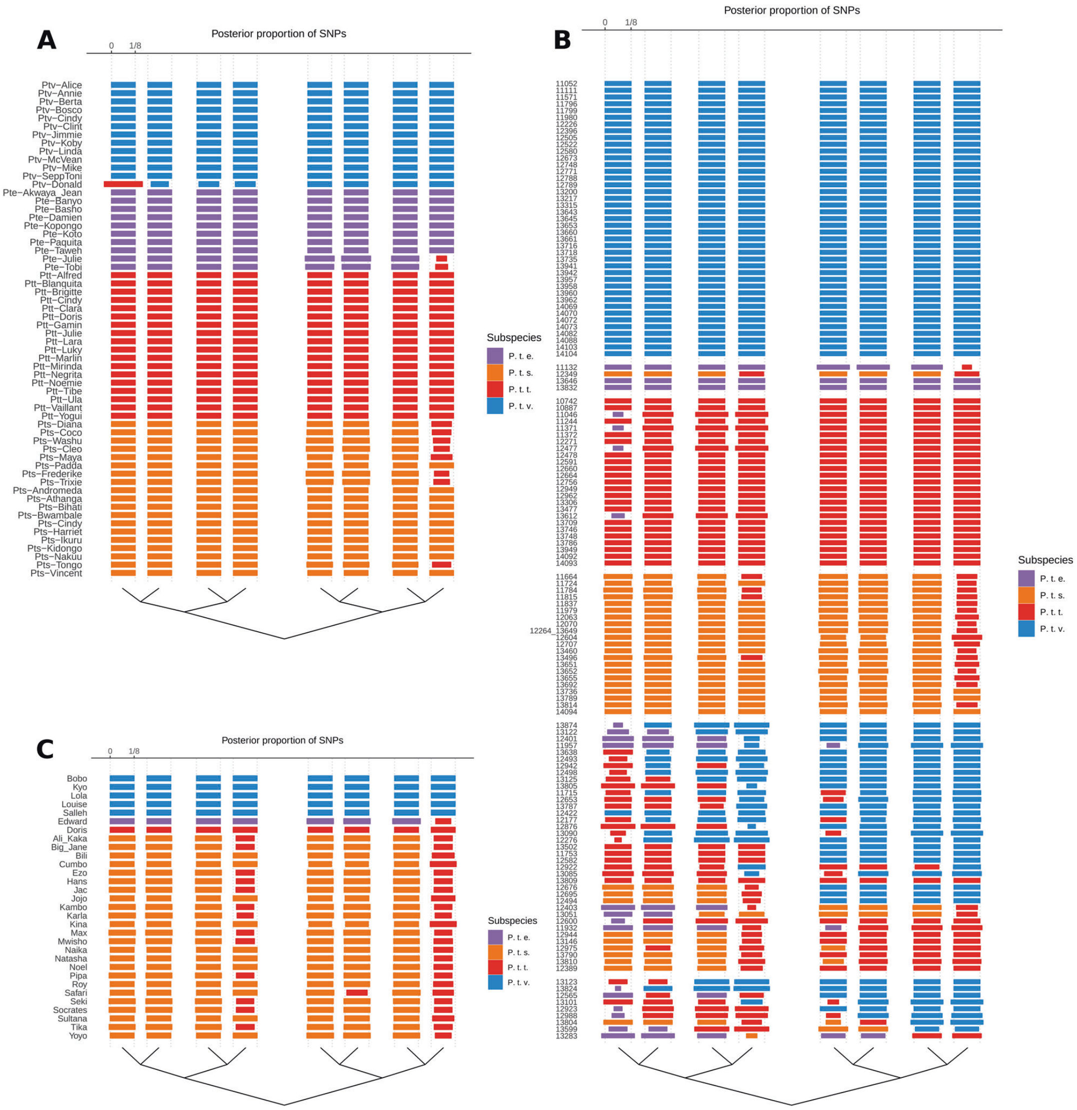


Fig. 2 Hybrid classification. Hybrid ancestry in **a** the reference panel, **b** the EEP population, and **c** the sanctuary population. The estimated posterior ancestries, θ is shown for the eight ancestors $k = 3$ generations back in time, for each individual in the three populations. The ancestors are orderedaccording to the "unphased" pedigree in the bottom of the plot. The width of each rectangle indicate the expected proportion of loci that are assigned to thatancestor (conditioned on the estimate of θ). Small widths suggest deviations from the model and features that could be improved by posterior correction.

$k = 3$ generations and visualised the number of loci each ancestor in generation k contributed to the ancestral informative part of the genome (see Supplementary Information). In general, our method correctly estimated the expected ancestries of our reference panel individuals (Fig. 2a). Several eastern and Nigerian-Cameroonian chimpanzee individuals were estimated to contain substantial ancestry components from the mutually neighbouring central subspecies. The known hybrid *Ptv-Donald* (Prado-Martinez et al. 2013) was estimated by the method to be at least one-eighth central chimpanzee, yet the large proportion of loci that were assigned to the central chimpanzee in the posterior distribution might suggest that *Ptv-Donald* could be as much as one-fourth central chimpanzee.

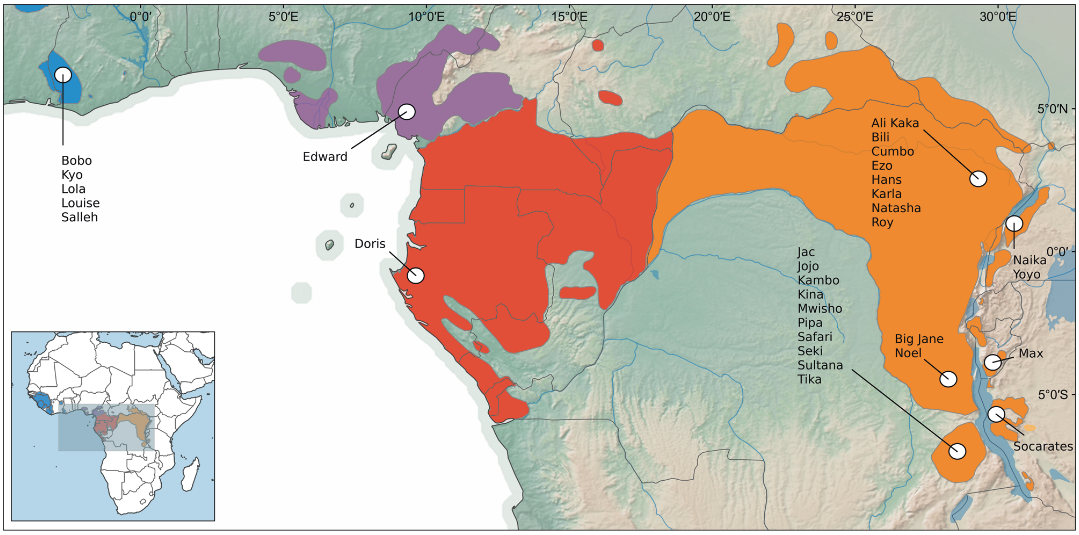


Fig. 3 Geographical origin estimates for sanctuary individuals. Based on the allele frequency surface map of the reference panel, sanctuary individuals are assigned probabilities of geographical origin, here summarised from individual estimates.

Similar to the ancestries inferred with ADMIXTURE, our method classified a large fraction of the EEP and sanctuary individuals to have ancestors from only one subspecies in the last three generations (Figs. 1c, 2b, c). In general, individuals inferred to belong to the eastern chimpanzee had third generation ancestors of central chimpanzee ancestry (Fig. 2b, c). Similarly, four inferred central chimpanzees in the EEP population, showed small proportions of ancestry from the Nigeria-Cameroon chimpanzee. Comparably, one sanctuary individual, *Edward*, was inferred here as a Nigeria-Cameroon chimpanzee with a small proportion of central chimpanzee ancestry. However, performing posterior correction by replacing the low central chimpanzee ancestor with another high posterior Nigeria-Cameroon ancestor, would likely make a more accurate estimate. Among the admixed EEP individuals, our model showed similar results to those obtained with ADMIXTURE but as ancestry patterns became increasingly complex (more than two ancestral subspecies) our inferred posterior proportions became increasingly uncertain (Figs. 2b, S14). We further observed that in some cases, small deviating (possibly deep coalescing) segments could have let the model to prefer configurations in the ancestry patterns to switch halves (Fig. 2c), while the correct configuration would probably be a simple case of hybridisation in the parent generation.

Geo-localisation

Based on an allele frequency surface map, built from our reference panel of wild born individuals, we determined the geographical origin of all 31 sanctuary individuals. Generally, the inferred probabilities of geographical origin gave accurate estimates (i.e. high probabilities assigned to just one or a few adjacent grid cells) for all sanctuary individuals (Fig. 3). Also, all individuals assigned to the natural range of their inferred subspecies label. The majority of our tested sanctuary individuals belonged to the eastern chimpanzee where the geographical origins were inferred to six provinces along the eastern part of the natural range of the subspecies. Seven of the eastern individuals had low probability estimates divided over a cluster of adjacent grid cells, with the highest ranking cell assigned probability of less than 0.1. All five western chimpanzee individuals were assigned to the same grid cell in the eastern limits of their range. The single individual from the Nigeria-Cameroon chimpanzee was assigned to a locality in Cameroon while the one central chimpanzee was assigned to the coastal region of Gabon.

Non-invasive sampling

Expanding our targeted capture approach to non-invasively collected hair samples, corroborated the results obtained with blood samples. ADMIXTURE estimates converged to the same result in the two sample types for all tested individuals and geographical origin was assigned to the same locality between samples (Figs. 4, S15–S19). Compared to the reference, ancestry estimates in our capture array approach did not always reveal the minor components of shared ancestries found when including all variant sites in the genome (Fig. 4).

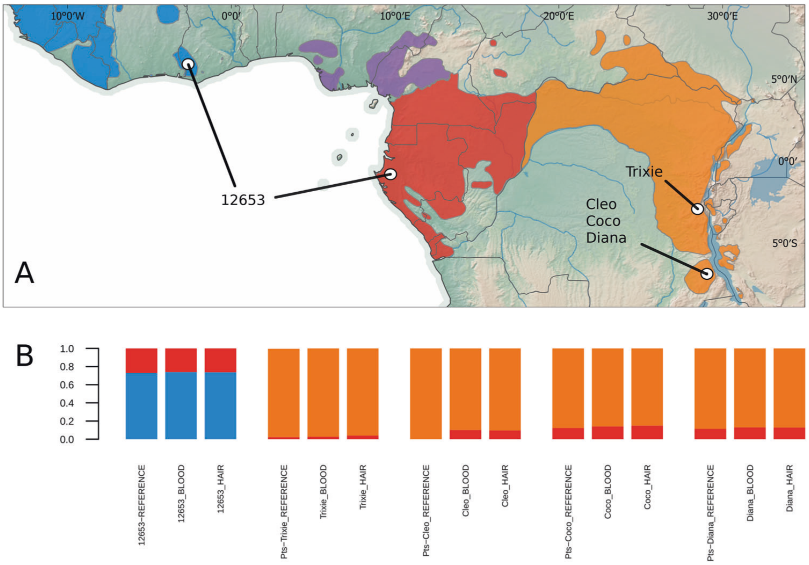


Fig. 4 Ancestry and geographical origin estimates from non-invasive samples. a Geographical origin estimates from hair samples based on the allele frequency surface map of the reference panel, tested individuals are assigned probabilities of geographical origin, here summarised from individual estimates with comparison to blood samples (Figs. S15–S19). **b** Shared ancestry estimates for hair samples compared to whole genome reference data and capture sequenced data from blood.

Discussion

As an exemplar for conservation genetics of endangered species, we have designed a novel capture array that targets identified ancestry informative markers across the genomes of 24 wild born chimpanzees (Prado et al. 2013) and the PanTro4 reference genome. Acknowledging that the selected ancestry markers were derived from a relatively limited set of genomes, which could potentially introduce an ascertainment bias towards specific subspecies, we confirmed that our design has the power to correctly identify the subspecies of an extended panel of newly sequenced chimpanzee genomes (de Manuel et al. 2016) (Fig. 1). Based on this proof of concept, we sequenced 167 chimpanzees from the EEP and sanctuary populations and analysed subspecies ancestries and geographical origin. We further show how this approach can be extended to non-invasive samples with robust results.

Ancestry of the ex situ population

In our test panel of 167 chimpanzees, 136 were from the EEP population housed at 47 European zoos and rehabilitation centres. Based on information on disembarkation or place of capture, we know that the majority of chimpanzees who founded the current EEP population came from West Africa. In accordance to this, a majority of the 90 non-admixed individuals could be assigned to the western chimpanzee (Fig. 1c). Our findings confirm that for the western chimpanzee, early efforts of the EEP that sought to identify a core group of non-admixed western chimpanzees using mitochondrial DNA (Jepsen and Carlsen *unpublished*) and microsatellites (Hvilsom et al. 2013), have been momentarily successful. Yet, using similar methodologies, previous attempts have only managed to identify a small group of central chimpanzees since the breeding effort for this subspecies was established (Carlsen and de Jongh 2019). Here, we identify 25 central chimpanzee individuals in the EEP population that show no evidence of shared ancestry with other subspecies (Fig. 1c), and hence from a genetic viewpoint, would qualify as a suitable bolster to the current breeding population. Similarly, the 21 inferred non-admixed eastern chimpanzee individuals could form the crucial starting point from where a separate breeding effort could be established under the EEP. In contrast to this, of our tested 136 EEP individuals, only three could be assigned to the Nigerian-Cameroonian subspecies (Fig. 1c) and in general, of the four subspecies, the Nigeria-Cameroon chimpanzee is by far the least represented in the EEP population (Carlsen and de Jongh 2019). Yet, with our targeted capture approach, it will now be feasible to scan the remaining EEP population (~1000 housed individuals) for additional non-admixed chimpanzee individuals in order to explore the possibilities of creating separate breeding populations for the two remaining subspecies.

Still, with a presumed small EEP population of eastern and Nigerian-Cameroonian chimpanzees, it might prove difficult to avoid inbreeding, although our estimates suggests, that high inbreeding coefficients are not exclusive to these particular subspecies. In fact, individuals with inbreeding coefficients in the range of 0.2–0.4 were found in each of the four subspecies and includes both wild and captive born individuals (Fig. 1d). It is therefore difficult to establish whether the amount of inbreeding in EEP individuals are a consequence of breeding among closely related individuals or whether it stems from inbred founders. In a few cases, like individual '*14073*', we know from reliable pedigree information, that this individual is the offspring of two full-siblings (Carlsen and de Jongh 2019). For the large majority of the EEP population, this knowledge is not available or is associated with high levels of uncertainties. Together with accurate ancestry inferences, genetically-based inbreeding estimates will be of high importance in management of the breeding population as will other factors such as age, fecundity, behaviour and housing capacities.

Of our 136 tested EEP individuals, 46 were inferred to be of hybrid origin (Fig. 1c). In terms of distinguishing founder individuals with shared ancestry components (wild born hybrids) from ex situ hybrids, our ancestry analyses show that the majority of our inferred hybrids are between non-neighbouring populations in the wild (e.g. between the western chimpanzee and either of the three other subspecies) and are therefore most likely the result of hybridisation in the EEP breeding population. From a management standpoint, these should eventually be phased out of the breeding programme. Yet, some known hybrids have been allowed to breed under the current management. This has been done with the purpose to maintain population numbers in an interim period while the populations reach their target size and also to allow experienced females to pass on up-bringing behaviour to young individuals in the housed groups. To explore the extent of wild born hybrids in the EEP and the possibility of including these in the breeding efforts, we developed a new method for hybrid classification that can trace ancestry patterns three generations back. This could possibly allow us to distinguish between hybrids bred in captivity and wild born hybrids, where the latter could be included in breeding programmes, as they represent natural processes in the wild. However, two key requirements to such an inclusion are a better understanding of the extent of hybridisation in the wild and an EEP management decision on what a suitable admixture threshold would be.

As validation for the hybrid classification model (see also Supplementary Information), our method infers the known hybrid background of *Ptv-Donald* to have received at least 12.5% of its ancestry from the central chimpanzee, which is in the range of what was previously estimated using whole-genome sequencing data (Prado-Martinez et al. 2013). Yet, in the EEP population, only a few of the inferred hybrids fit with the expectations of ancestry patterns in wild born hybrids. The majority of the inferred hybrids include a western chimpanzee ancestry component (Fig. 2b), which is highly unlikely to occur in the wild due to the vast geographical distance to any neighbouring subspecies (Fig. 1a). Of the eight inferred hybrids with adjacent distribution ranges, one central/Nigerian-Cameroonian and seven central/eastern hybrids (Fig. 2b), we know from studbook information that all eight individuals were captive born (Carlsen and de Jongh 2019) (Table S2). The only cases where our model might have picked up remnants from natural hybridisations are the ancestry components of central chimpanzee in what we inferred to be non-admixed eastern chimpanzees using ADMIXTURE (Fig. 1c, Fig. 2b). However, this could likely be due to a general limitation of our model to separate these two subspecies due to their evolutionary close relationship and history of allele sharing (Prado-Martinez et al. 2013; de Manuel et al. 2016). Although we did not identify any wild born hybrids in the tested set of individuals, our model predictions will be highly useful in terms of pinpointing the timing of admixture and help to illuminate blanks in the studbook regarding possible sires.

Sanctuary ancestry and geographical origin

In contrast to the predominance of western chimpanzee individuals in the EEP population, the majority of the tested sanctuary individuals are inferred to belong to the eastern chimpanzee. Of the 31 tested individuals, we only find four that can be assigned to the western chimpanzee and a single individual from each of the Nigeria-Cameroon chimpanzee and the central chimpanzee (Fig. 1c). When exploring ancestry patterns in the last three generations, we obtained similar results as in the EEP population, where small posterior proportions of central chimpanzee were found in individuals of the eastern chimpanzee (Fig. 2c). This is most likely due to the limitations of our model when it comes to distinguishing shared alleles between these two subspecies, and we do not infer any geographical origin close to possible contact zones between the two subspecies (Fig. 3).

For western and Nigerian-Cameroonian chimpanzees, we obtained high probabilities in the assigned origins but with little spatial resolution. Essentially, all five western chimpanzee individuals assign to the same grid cell. As de Manuel et al. (2016) have previously shown, population structure inferred in the western and Nigerian-Cameroonian populations, may not offer enough resolution to provide fine scale determination of geographical origin. To improve origin estimates in these populations, it is crucial to obtain a better representation of georeferenced samples across their distribution ranges. This has been achieved for most of the central and eastern chimpanzee ranges, but with only one central chimpanzee individual (*Doris*), we cannot fully evaluate the prediction power and resolution for this subspecies. Nevertheless, the estimated geographical origin of *Doris* is very close to the reported confiscation site (Table S3), which gives us some assurance that future efforts to determine origins in the central chimpanzee will be possible. With a larger set of individuals from the eastern chimpanzee, we can start to appreciate the full potential of the method. The 24 analysed individuals can be assigned to geographical origins in six localities along the eastern edge of the distribution range of the eastern chimpanzee, where the majority originates from two locations in the northern and southern regions of the Democratic Republic of Congo (DRC) (Fig. 3). First of all, this might tell us that these regions are heavily affected by poaching and illegal trafficking, although the abundance of confiscation sites might also be biased by the locality of contributing sanctuaries. Only further testing of individuals from sanctuaries across the species range will allow us to assess regional threat levels. However, with the inferred origins of the eastern chimpanzee individuals all along the eastern edge of the range, we can conclude that the threats are not confined to a few regions for this subspecies but are distributed across the eastern boarders of the DRC.

When comparing the inferred geographical origins with the reported confiscation sites for all our tested sanctuary individuals (Table S3), it becomes apparent that the trafficking routes generally operate within a relatively local scale. Overall, we see that most of the tested individuals originate from locations that are within close proximity to where they have been confiscated, though with two notable exceptions, *Louise* and *Edward*. *Louise* was confiscated in Moscow, Russia and inferred to have originated from West Africa, while *Edward* was confiscated in Nairobi Airport, Kenya with inferred origin in Cameroon. This confirms that the illegal trade of wild chimpanzees spans beyond country borders and the African continent as reported in Stiles et al. (2013). Both individuals are now housed in sanctuaries where specialised care can be provided, yet, in these cases, both individuals have been placed in sanctuaries far from their geographical origin and possibly within mixed subspecies groups (other individuals from these sanctuaries have been assigned to different subspecies). Without proper knowledge of their ancestry, sanctuaries might face the same challenges as we have seen in the EEP population, with admixture of subspecies as a result of (unintended) breeding. Genetic testing at an early stage could help to ameliorate these challenges and as we have shown, our genomic approach extents to non-invasive sampling (Fig. 4), making these methods both an accurate and practical tool in conservation efforts to help combat the illegal trade of chimpanzees.

We further predict that this approach will be self-empowering as sampling gaps in the distribution range of the chimpanzee are continuously covered and DNA extraction methods for non-invasive samples improve. This will significantly advance our predictive power of geographical origin and provide valuable insight to shared ancestries in natural populations with positive knock-on effects to hybrid assessment in the ex situ populations.

Our capture array approach of targeting ancestry informative markers offers a standardised and cost-effective method that accurately guides ex situ and in situ conservation management programmes. At the current rate of decline, chimpanzees are predicted to go extinct within the current century (Estrada et al. 2017). Conservation efforts might therefore, in a foreseeable future, be obligated to supplement wild populations with individuals from the ex situ populations as a last resort to prevent them from going extinct. Should it come to this, our approach facilitates the safeguarding of genetically self-sustainable populations that will have preserved a genetic profile that resembles their wild counterparts.

The current extinction crisis however, extends well beyond chimpanzees and the demand for molecular genetics to help guide future population management programmes is immense, ranging across the taxonomical scale of birds, reptiles, amphibians, and mammals. For the latter alone, more than ten EEP genetic projects are underway and globally, regional zoo associations are undertaking molecular genetic studies for which the present study serves as an important blueprint for linking in situ and ex situ conservation efforts.

Data archiving

The genetic data used in the present study is a publicly accessible through the Dryad Digital Repository, https://doi.org/10.5061/dryad.31zcrjdh7.

Acknowledgements The authors would like to thank all institutions who provided samples for this study: The EAZA Biobank, Stichting AAP Amersfoort Zoo, Antwerp Zoo, Zoo Delle Star ad Aprilia, Royal Burgers' Zoo Arnhem, Monde Sauvage Safari, Badoca Safari Park, Barcelona Zoo, Zoo Parc de Beauval, Bioparc Valencia, Borås Zoo, Parco Natura Viva, Edinburgh Zoo, Ölands Zoo, Plättli Zoo, Givskud Zoo/Zootopia, Safaripark Beekse Bergen, Hodonin Zoo, Xanthus Zoo, Kolmården Zoo, Krakow Zoo, Kristiansand Zoo, Lagos Zoo, Le Pal Zoo, Leipzig Zoo, Liberec Zoo, Lisbon Zoo, Madrid Zoo, Magdeburg Zoo, Olmense Zoo, Zoo Osnabrück, African Safari, Plaisance du Touch, Plzen Zoo, Centro de Rescate de Primates Rainfer, Zoological Center Tel Aviv—Ramat Gan, Safari Ravenna, Zoo di Roma, Leintal Zoo, Serengeti-park Hodenhagen, Reserve Africaine de Sigean, Wilhelma Zoo, Loro Parque Zoo, Touroparc, Twycross Zoo, AAP Primadomus, Warsaw Zoo, Schwabenpark, Zagreb Zoo, Cente International de Recherches Médicales de Franceville, Chimfunshi Wildlife Orphanage, Jeunes Animaux Confisques au Katanga, Ngamaba Island Chimpanzee Sanctuary, Tacugama Chimpanzee Sanctuary, The Chimpanzee Conservation Center. We further wish to thank Tom de Jongh and Lisbeth Borbye for valuable language editing and comments to the manuscript and Abigail Ramsøe for early developmental stages of the hybrid classification model. PF is supported by the Innovation Fund Denmark doctoral fellowship programme and the Candys Foundation. CF is supported by "la Caixa" doctoral fellowship programme. TSK is funded by Carlsberg grant CF19-0712 prepared within the framework of the HSE University Basic Research Program. TMB is supported by BFU2017-86471-P (MINECO/FEDER, UE), U01 MH106874 grant, Howard Hughes International Early Career, Obra Social "La Caixa" and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880). EL is supported by CGL2017-82654-P (MINECO/FEDER, UE).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664

Beck B, Walkup K, Rodriques M, Unwin S, Travis D, Stoinski T (2007) Best practice guidelines for the re-introduction of Great Apes. Gland, Switzerland

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Carlsen F, de Jongh T (2019) European studbook for the chimpanzee Pan troglodytes. Copenhagen

Carøe C, Gopalakrishnan S, Vinner L, Mak SST, Sinding MHS, Samaniego JA et al. (2018) Single-tube library preparation for degraded DNA. Methods Ecol Evol 9:410–419

Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, Palmer TM (2015) Accelerated modern human–induced species losses: entering the sixth mass extinction. Sci Adv 1:e1400253

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA et al. (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J et al. (2016) Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science 354:477–481

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498

Estrada A, Garber PA, Rylands AB, Roos C, Fernandez-duque E, Fiore A Di et al. (2017) Impending extinction crisis of the world's primates: why primates matter. Sci Adv 3:e1600946

Grueber CE, Fox S, McLennan EA, Gooley RM, Pemberton D, Hogg CJ et al. (2019) Complex problems need detailed solutions: harnessing multiple data types to inform genetic management in the wild. Evol Appl 12:280–291

Hanghøj K, Moltke I, Andersen PA, Manica A, Korneliussen TS (2019) Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. Gigascience 8:1–9

Hockings KJ, McLennan MR, Carvalho S, Ancrenaz M, Bobe R, Byrne RW et al. (2015) Apes in the anthropocene: flexibility and survival. Trends Ecol Evol 30:215–222

Humle T, Maisels F, Oates JF, Plumtre A, Williamson EA (2016) Pan troglodytes. IUCN Red List Threat Species

Hvilsom C, Frandsen P, Børsting C, Carlsen F, Sallé B, Simonsen BT et al. (2013) Understanding geographic origins and history of admixture among chimpanzees in European zoos, with implications for future breeding programmes. Heredity 110:586–593

IUCN (2015) IUCN Red List of Threatened Species. Version 20153: www.iucnredlist.org

Jepsen BI, Carlsen F Genetic identification of West African Chimpanzee, Pan troglodytes verus, based on mitochondrial DNA analysis. Unpublished

Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 40:e3. https://doi.org/10.1093/nar/gkr771

Lee S, Epstein MP, Duncan R, Lin X (2012) Sparse principal component analysis for identifying ancestry-informative markers in genome-wide association studies. Genet Epidemiol 36:293–302

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 et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079

Norman AJ, Putnam AS, Ivy JA (2019) Use of molecular data in zoo and aquarium collection management: benefits, challenges, and best practices. Zoo Biol 38:106–118

Nye J, Laayouni H, Kuhlwilm M, Mondal M, Marques-Bonet T, Bertranpetit J (2018) Selection in the introgressed regions of the chimpanzee genome. Genome Biol Evol 10:1132–1138

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B et al. (2013) Great ape genetic diversity and population history. Nature 499:471–475

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

QGIS (2018) QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842

Rañola JM, Novembre J, Lange K (2014) Fast spatial ancestry via flexible allele frequency surfaces. Bioinformatics 30:2915–2922

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res 22:939–946

Stiles D, Redmond I, Cress D, Nellemann C, Formo RK (2013) Stolen apes—the illicit trade in chimpanzees, Gorillas, Bonobos and Orangutans. Birkeland Trykkeri AS, Norway

Traylor-Holzer K, Leus K, Bauman K (2019) Integrated Collection Assessment and Planning (ICAP) workshop: helping zoos move toward the One Plan Approach. Zoo Biol 38:95–105

Xu H, Luo X, Qian J, Pang X, Song J, Qian G et al. (2012) FastUniq: a fast de novo duplicates removal tool for paired short reads. PLoS ONE 7:e52249. https://doi.org/10.1371/journal.pone.0052249

Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614–620

Affiliations

Peter Frandsen[1,2] · **Claudia Fontsere**[3] · **Svend Vendelbo Nielsen**[4] · **Kristian Hanghøj**[2] · **Natalia Castejon-Fernandez**[4] · **Esther Lizano**[3,5] · **David Hughes**[6,7] · **Jessica Hernandez-Rodriguez**[3] ·

Thorfinn Sand Korneliussen[8,9] · Frands Carlsen[1] · Hans Redlef Siegismund[2] · Thomas Mailund[4] · Tomas Marques-Bonet[3,5,10,11] · Christina Hvilsom[1]

[1] Research and Conservation, Copenhagen Zoo, Roskildevej 38, 2000 Frederiksberg, Denmark

[2] Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark

[3] Institute of Evolutionary Biology, (UPF-CSIC), PRBB, Dr. Aiguader 88, 08003 Barcelona, Spain

[4] Bioinformatics Research Center, Department of Mathematics, Aarhus University, C. F. Møllers Allé 8, 8000 Aarhus C, Denmark

[5] Institut Català de Paleontologia Miquel Crusafant, Universitat Autònoma de Barcelona, Edifici ICTA-ICP, c/ Columnes s/n, Cerdanyala del Vallès, 08193 Barcelona, Spain

[6] MRC Integrative Epidemiology Unit at University of Bristol, Bristol BS8 2BN, UK

[7] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol BS8 2BN, UK

[8] GLOBE, Section for Geogenetics, Øster Voldgade 5-7, 1350 Copenhagen, Denmark

[9] National Research University, Higher School of Economics, 20 Myasnitskaya Ulitsa, 101000 Moscow, Russia

[10] Catalan Institution of Research and Advanced Studies (ICREA), Passeig de Lluís Companys 23, 08010 Barcelona, Spain

[11] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri I Reixac, 408028 Barcelona, Spain

3.2. Maximizing the acquisition of unique reads in non-invasive capture sequencing experiments

Claudia Fontsere, Marina Alvarez-Estape, Jack Lester, Mimi Arandjelovic, Martin Kuhlwilm, Paula Dieguez, Anthony Agbor, Samuel Angedakin, Emmanuel Ayuk Ayimisin, Mattia Bessone, Gregory Brazzola, Tobias Deschner,Manasseh Eno-Nku, Anne-Céline Granjon, Josephine Head, Parag Kadam, Ammie K. Kalan, Mohamed Kambi, Kevin Langergraber, Juan Lapuente, Giovanna Maretti, Lucy Jayne Ormsby, Alex Piel, Martha Robbins, Fiona Stewart, Virginie Vergnes, Roman M. Wittig, Hjalmar S. Kühl, Tomas Marques-Bonet, David A. Hughes * & Esther Lizano*.

Maximizing the acquisition of unique reads in non-invasive capture sequencing experiments.

Accepted with minor revisions *Molecular Ecology Resources* (2020).

Maximizing the acquisition of unique reads in non-invasive capture sequencing experiments

Claudia Fontsere[1], Marina Alvarez-Estape[1], Jack Lester[2], Mimi Arandjelovic[2], Martin Kuhlwilm[1], Paula Dieguez[2], Anthony Agbor[2], Samuel Angedakin[2], Emmanuel Ayuk Ayimisin[2], Mattia Bessone[2], Gregory Brazzola[2], Tobias Deschner[2], Manasseh Eno-Nku[3], Anne-Céline Granjon[2], Josephine Head[2], Parag Kadam[4], Ammie K. Kalan[2], Mohamed Kambi[2], Kevin Langergraber[5,6], Juan Lapuente[2,7], Giovanna Maretti[2], Lucy Jayne Ormsby[2], Alex Piel[8], Martha M. Robbins[2], Fiona Stewart[4,8], Virginie Vergnes[9], Roman M. Wittig[2,10], Hjalmar S. Kühl[2,11], Tomas Marques-Bonet[1,12,13,14 †], David A. Hughes[15,16 *] and Esther Lizano[1,14 † *]

[1] Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Doctor Aiguader 88, Barcelona, 08003, Spain.
[2] Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.
[3] WWF Cameroon Country Programme Office, BP6776; Yaoundé, Cameroon.
[4] School of Biological and Environmental Sciences, Liverpool John Moores University, James Parsons Building, Byrom street, Liverpool, L3 3AF, UK.
[5] School of Human Evolution and Social Change, Arizona State University, 900 Cady Mall, Tempe, AZ 85287 Arizona State University, PO Box 872402, Tempe, AZ 85287-2402 USA.
[6] Institute of Human Origins, Arizona State University, 900 Cady Mall, Tempe, AZ 85287 Arizona State University, PO Box 872402, Tempe, AZ 85287-2402 USA.
[7] Comoé Chimpanzee Conservation Project, Kakpin, Comoé National Park, Ivory Coast.
[8] Department of Anthropology, University College London, 14 Taviton St, Bloomsbury London.
[9] Wild Chimpanzee Foundation (WCF) 23BP238 Abidjan, Côte d'Ivoire 23.
[10] Taï Chimpanzee Project, Centre Suisse de Recherches Scientifiques, BP 1301, Abidjan 01, CI, Côte d'Ivoire.
[11] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig.
[12] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain.
[13] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia 08010, Spain.
[14] Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Columnes s/n, 08193 Cerdanyola del Vallès, Spain.
[15] MRC Integrative Epidemiology Unit at University of Bristol, Bristol, BS8 2BN, UK.
[16] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, BS8 2BN, UK.

* Esther Lizano and David A. Hughes should be considered joint senior author.

[†] **Corresponding author:** Esther Lizano and Tomas Marques-Bonet

Abstract

Non-invasive samples as a source of DNA are gaining interest in genomic studies of endangered species. However, their complex nature and low endogenous DNA content hamper the recovery of good quality data. Target capture has become a productive method to enrich the endogenous fraction of non-invasive samples, such as feces, but its sensitivity has not yet been extensively studied. Coping with fecal samples with an endogenous DNA content below 1% is a common problem when prior selection of samples from a large collection is not possible. However, samples classified as unfavorable for target capture sequencing might be the only representatives of unique specific geographical locations or to answer the question of interest.

To explore how library complexity may be increased without repeating DNA extractions and generating new libraries, here we have captured the exome of 60 chimpanzees (*Pan troglodytes*) using fecal samples with very low proportions of endogenous content (< 1%).

Our results indicate that by performing additional hybridizations of the same libraries, the molecular complexity can be maintained to achieve higher coverage. Also, whenever possible, the starting DNA material for capture should be increased. Lastly, we have specifically calculated the sequencing effort needed to avoid exhausting the library complexity of enriched fecal samples with low endogenous DNA content.

This study provides guidelines, schemes and tools for laboratories facing the challenges of working with non-invasive samples containing extremely low amounts of endogenous DNA.

Keywords: Non-invasive samples, fecal samples, target capture, molecular complexity, conservation genomics, chimpanzees.

Introduction

Studies of wild animal populations that are unamenable to invasive sampling (eg: trapping or darting) often rely on the usage of low quality and/or quantity DNA samples (Schwartz, Luikart, & Waples, 2007; Vigilant & Guschanski, 2009), traditionally restricting the analysis to neutral markers or genetic loci such as microsatellites (Arandjelovic et al., 2011; Inoue et al., 2013; Mengüllüoğlu, Fickel, Hofer, & Förster, 2019; Orkin, Yang, Yang, Yu, & Jiang, 2016), autosomal regions (Fischer, Wiebe, Pääbo, & Przeworski, 2004) and the mitochondrial genome (Fickel, Lieckfeldt, Ratanakorn, & Pitra, 2007; Thalmann, Hebler, Poinar, Pääbo, & Vigilant, 2004). Depending on the researcher's question, these neutral genetic markers may continue to be the most economical and efficient method (Shafer et al., 2015). However, for other questions such as cataloging genetic diversity, assessing kinship, making fine inferences of demographic history, or evaluating disease susceptibility, it is increasingly relevant to acquire a more representative view of the genome (Ouborg, Pertoldi, Loeschcke, Bijlsma, & Hedrick, 2010; Primmer, 2009; Shafer et al., 2015; Städele & Vigilant, 2016; Steiner, Putnam, Hoeck, & Ryder, 2013).

Conservation genomics of ecologically-crucial, non-model organisms, and especially threatened species such as great apes, have largely benefited from the current advances in next-generation sequencing (NGS) technologies (Gordon et al., 2016; Locke et al., 2011; Mikkelsen et al., 2005; Scally et al., 2012). The ability to simultaneously interrogate hundreds of thousands of genetic markers across an entire genome allows greater resolution on inferences of demographic parameters, genetic variation, gene flow, inbreeding, natural selection, local adaptation and the evolutionary history of the studied species (De Manuel et al., 2016; Prado-Martinez et al., 2013; Xue et al., 2015).

The major impediment to the study of wild, threatened, natural populations continues to be the difficulties in acquiring samples of known location from a large number of individuals. To avoid disturbing and negatively influencing endangered species (alteration of social group dynamics, infections and stress) (Morin, Wallis, Moore, Chakraborty, & Woodruff, 1993; Taberlet, Luikart, & Waits, 1999), but also to track cryptic or monitor reintroduced species (De Barba et al., 2010; Ferreira et al., 2018; Reiners, Encarnação, & Wolters, 2011; Stenglein, Waits, Ausband, Zager, & Mack, 2010), sampling often relies on non-invasive (NI) sources of DNA such as feces and hair, rather than invasive samples such as blood or other tissues, which yield better DNA quality and quantity.

NI samples have a complex nature: they are typically composed of low proportions of host or endogenous DNA (eDNA), are highly degraded (Perry, Marioni, Melsted, & Gilad, 2010; Taberlet et al., 1999), and contain genetic material from the host's microbiota and from species living in the environment where the sample was collected (i.e., exogenous DNA) (Hicks et al., 2018). The proportion of endogenous versus exogenous DNA can be highly variable (Hernandez-Rodriguez et al., 2018) and as previous literature has proposed, may depend on the environmental conditions, with humidity and ambient temperature having the highest influence (Goossens, Chikhi, Utami, De Ruiter, & Bruford, 2000; Harestad & Bunnell, 1987; King, Schoenecker, Fike, & Oyler-McCance, 2018; Nsubuga et al., 2004). Because of this, the employment of techniques that generate sequences of the whole genomic content of the samples, such as NGS, has not been economically feasible until recently. Target enrichment technologies, also known as capture, have become a common and successful methodology in ancient DNA studies (Burbano et al., 2010; Carpenter et al., 2013; Maricic, Whitten, & Pääbo, 2010) and have allowed for a more cost-effective use of NGS on NI samples, as the endogenous to exogenous DNA ratio greatly improves, thus reducing the sequencing effort (Perry et al., 2010; Snyder-Mackler et al., 2016; van der Valk, Lona Durazo, Dalén, & Guschanski, 2017). Capture methods reduce the relative cost of sequencing and improve the quality of the data by building DNA libraries that are hybridized to complementary baits for selected target regions (partial genomic regions, a chromosome, the exome, or the whole genome) increasing the proportion of the targeted eDNA to be sequenced.

Despite the existence of technical studies describing the use of NI samples for the genomic study of wild chimpanzees (*Pan troglodytes*) (Hernandez-Rodriguez et al., 2018; White et al., 2019) many aspects remain to be investigated. For instance, in Hernandez-Rodriguez et al., samples were selected to cover the entire range of observed average fragmentation lengths and percentage of eDNA, in order to be as representative as possible. As a result, they observed a sequencing bias due to the different percentage of endogenous content in captured samples. To avoid that outcome, they proposed performing equi-endogenous pools instead of the standard pooling of libraries according to molarity. White et al. followed this recommendation and yielded a more balanced representation across samples. However, their experiments were limited to only those samples with a proportion of eDNA above 2% (White et al., 2019). As shown by Hernandez-Rodriguez et al. there is a positive association between endogenous content and the amount of data acquired from a sample, such that when possible, one should use those samples with higher endogenous content. However, the proportion of chimpanzee fecal samples with eDNA above 2% is often very low (<20%) (White et al., 2019).

The NI chimpanzee samples used in this study were collected from 15 different geographic sites across the whole species' ecological habitat in Africa and included all four subspecies, thus representing a wide variety of sampling and environmental conditions. With this screening approach we were able to examine how the proportion of eDNA content varies between each site, revealing that the majority of collected samples in some sites have low proportions of eDNA (<1%). Therefore, when prior selection of samples from a large collection is not possible, the only ones representing a specific location or that are relevant to the scientific question, might be those with extremely low proportions of endogenous content. Because of that, we have focused our efforts on developing approaches to retrieve the maximum data possible from challenging samples.

In that regard, we sought to capture the exome of 60 chimpanzee fecal samples as part of the Pan African Programme: The Cultured Chimpanzee (PanAf) (http://panafrican.eva.mpg.de/) (Kühl et al., 2019) with eDNA estimates below 1%. We used a commercial human exome to evaluate how the coverage of targeted genomic regions may be increased in a collection of samples that may be regarded as unfavorable for target capture sequencing. We confirmed the importance of the correct estimation of eDNA and the pooling of libraries accordingly to avoid sequencing bias across samples (Hernandez-Rodriguez et al., 2018). We also expanded on previously explored and unexplored guidelines to ensure the maintenance of the captured molecule diversity or library complexity such as the number of libraries in a pool, the performance of additional hybridizations and increasing the total DNA starting material for capture (Hernandez-Rodriguez et al., 2018; Perry et al., 2010; Snyder-Mackler et al., 2016; White et al., 2019).

Our results provide the most comprehensive exploration to date of target enrichment efficiency in very low eDNA fecal samples, and guidelines to improve the quality of the data without re-extracting DNA and preparing new libraries. These findings could greatly benefit the conservation effort on great apes, as well as any other species with similar DNA sampling limitations.

Material and Methods

Samples and Library Preparation

Chimpanzee fecal samples from 15 different sites in Africa were collected as part of the PanAf (Figure 1A). Approximately 5g (“hazelnut-size”) of feces were collected from each chimpanzee fecal sample and stored in the field using a two-step ethanol-silica preservation method (Nsubuga et al., 2004). Depending on the density of the sample, between 10 and 80 mg of dry fecal sample were extracted using a Qiagen robot with the QIAamp Fast DNA Stool Mini Kit (Qiagen) with modifications (Lester et al, in review, 2020). The extractions were screened using a microsatellite genotyping assay (Arandjelovic et al., 2009; Arandjelovic et al., 2011) and up to 20 samples from each PanAf field site were selected as follows: (1) those that amplified at the most loci of the 15 loci panel, (2) represented unique individuals, and (3) were ascertained to have a low probability of being first degree relatives (Csilléry et al., 2006) (302 samples) (Supporting Information Table S1). To ensure sufficient template DNA for library preparation, the 302 samples were re-extracted using the same QIAamp kit and between 100 and 200 mg of dry fecal sample. Total DNA concentration and fragmentation were measured on a Fragment Analyzer using a Genomic DNA 50Kb Analysis kit (Advanced Analytical) and the fragmentation level was calculated with PROSize Data Analysis Software (Agilent Technologies). Endogenous DNA content (fraction of mammalian DNA, relative to gut microbial and other environmental genetic material) was estimated by qPCR (Morin, Chambers, Boesch, & Vigilant, 2001). Finally, percentage of endogenous content for each sample was calculated by dividing the chimpanzee eDNA concentration by the total DNA concentration. We selected 60 samples with an intermediate percentage of eDNA (0.41-0.85%, average 0.61%) from the 302 screened samples (range of endogenous distribution: 0-47.57%, average 1.49%) (Supporting Information S1 and Table S2).

A single library was prepared for each of the 60 samples following the BEST protocol (Carøe et al., 2018) starting with 200 ng total DNA (from a sample) with minor modifications. Specifically, double in-line barcoded adapters were used, barcoding each sample at both ends of its library to allow for its unique identification within a pool (Rohland & Reich, 2012). Library concentration was calculated using Agilent 2100 BioAnalyzer and DNA7500 assay kit. A detailed protocol for library construction can be found in Supplementary Information.

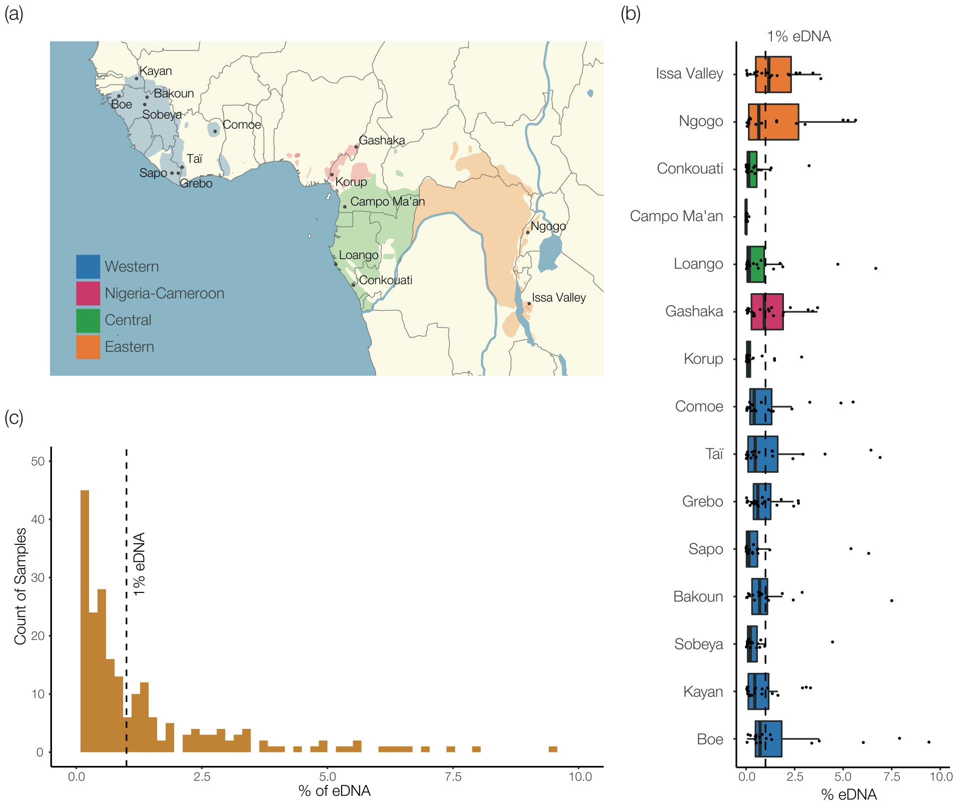


FIGURE 1. Sample description. (a) Geographical location of the 15 sites from the Pan African Programme: The Cultured Chimpanzee (PanAf). (b) Endogenous DNA (eDNA) content for all screened samples according to geographic origin. The maximum value of the x-axis has been set to 10% eDNA for visual purposes. (c) eDNA distribution for all screened samples. Samples with > 10% eDNA are excluded (N=5). In the boxplot, lower and upper hinges correspond to first and third quartiles and the lower and upper whiskers extend to the smallest or largest value no further than 1.5 times the interquartile range (distance between the 1st and 3rd quartile).

Pooling and Capture

Endogenous DNA content is a key factor in target-capture experiments directly influencing the yield of on-target reads and molecule diversity (Hernandez-Rodriguez et al., 2018). Our equi-endogenous sample pooling strategy follows two criteria. First, samples belonging to a pool have similar eDNA proportions according to a 1:2 ratio rule: the sample with highest proportion of eDNA cannot double the sample with the lowest. Second, each sample within a pool contributes the same total amount of eDNA (µg) to the final pool, creating an equi-endogenous pool. So, the sample with the lowest percentage of eDNA will contribute more total DNA to the final pool compared to the sample with the highest, but the amount of eDNA per sample will be equivalent.

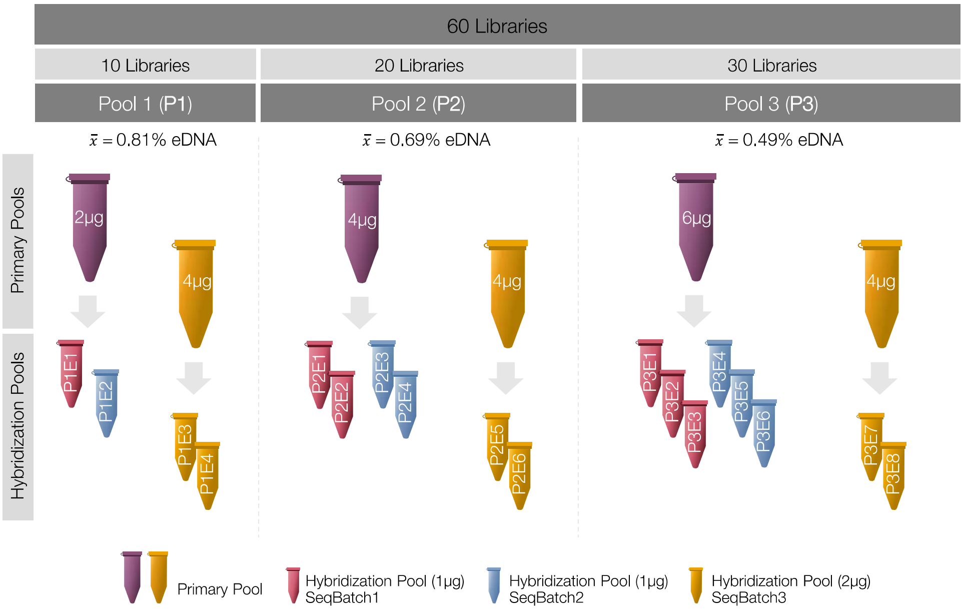


FIGURE 2. Pooling strategy illustration. P1 has 10 libraries with average endogenous of 0.81%. We performed two primary pools of 2 µg and 4 µg each that were further divided into four hybridization pools, two at 1 µg and two at 2 µg. P2 has 20 libraries with average endogenous of 0.69%. Two primary pools of 4 µg were divided into four hybridization pools of 1 µg each and two hybridizations pools of 2 µg. P3 has 30 libraries and an average endogenous of 0.49%. Two primary pools of 6 µg and 4 µg were distributed into six hybridization pools of 1µg and two hybridization pools of 2 µg each. Colors represent the sequencing batch.

According to the estimates of eDNA, we pooled the 60 libraries into three primary pools (see graphical representation in Figure 2). The first pool (P1) with 2 µg total DNA (in the pool) consisted of 10 samples with an average endogenous content of 0.81% (range 0.69-0.85%). The second pool (P2) had 4 µg total DNA and consisted of 20 samples and an average endogenous content of 0.69% (range 0.58-0.80%). The 30 remaining libraries were pooled into the third pool (P3) of 6 µg total DNA with an average endogenous content of 0.49% (range 0.41-0.66%) (Table 1 and Figure 3A, Supporting Information Table S2). Subsequently, each initial primary pool was subdivided into two (P1E1, P1E2), four (P2E1, P2E2, P2E3, P2E4) and six (P3E1, P3E2, P3E3, P3E4, P3E5, P3E6) exome capture (E) replicates each consisting of 1 µg of total DNA.

Independently, we repeated the construction of the primary pools (P1, P2 and P3), but with each having 4 µg total DNA. Each of these new primary pools was then divided into two replicates of 2 µg each (P1E3, P1E4, P2E5, P2E6, P3E7, P3E8). As a consequence of generating replicate primary pools, six of the 60 libraries were exhausted and are not present in these replicate primary pools. As a result, across all 60 samples and 18 hybridizations there are a total of 388 individual hybridization experiments (Figure 2). All details are provided in Table 1.

Each exome capture experiment consisted of two consecutive hybridizations, or dual-capture reactions as previously recommended (Hernandez-Rodriguez et al., 2018) using the SureSelect Human All Exon V6 RNA library baits from Agilent Technologies and was performed following the manufacturer's protocol with some modifications (full protocol is available in Supporting Information), and started with either 1 µg or 2 µg total DNA (Table 1 and Figure 2). After the first hybridization reaction and the subsequent PCR enrichment, we performed the second hybridization reaction with all available material. The final captured pool was amplified with indexed primers (Kircher, Sawyer, & Meyer, 2012), double-indexing each library within a pool, thereby tagging each library to a specific hybridization experiment. Double inline barcoded (sample specific) and double indexed (pool specific) libraries allow for multiplexing many libraries into a single pool and sequencing many pools into a single sequencing lane, even when the same sample library is present in multiple hybridization reactions. This permits the tracking of unique experiments.

For the reminder of the article when we use the word “capture” or “hybridization”, we will always be referring to the dual-capture or two consecutive rounds of capture hybridizations that are described above.

Pool	Average eDNA content (range)	Hybridization ID	Number of pooled libraries	Total DNA	Sequencing Batch
Pool 1 (**P1**)	0.81% (0.60% - 0.85%)	P1E1	10	1 µg	SeqBatch1
		P1E2	10	1 µg	SeqBatch2
		P1E3	9	2 µg	SeqBatch3
		P1E4	9	2 µg	SeqBatch3
Pool 2 (**P2**)	0.69% (0.58% - 0.80%)	P2E1	20	1 µg	SeqBatch1
		P2E2	20	1 µg	SeqBatch1
		P2E3	20	1 µg	SeqBatch2
		P2E4	20	1 µg	SeqBatch2
		P2E5	19	2 µg	SeqBatch3
		P2E6	19	2 µg	SeqBatch3
Pool 3 (**P3**)	0.49% (0.41% - 0.66%)	P3E1	30	1 µg	SeqBatch1
		P3E2	30	1 µg	SeqBatch1
		P3E3	30	1 µg	SeqBatch1
		P3E4	30	1 µg	SeqBatch2
		P3E5	30	1 µg	SeqBatch2
		P3E6	30	1 µg	SeqBatch2
		P3E7	26	2 µg	SeqBatch3
		P3E8	26	2 µg	SeqBatch3

TABLE 1. Pooling Strategy. Sixty libraries were divided into 3 pools for capture hybridization experiments in 4 replicates for P1, 6 replicates for P2 and 8 replicates for P3. Total DNA represents the starting material for each capture hybridization.

Sequencing and Mapping

Captured libraries were pooled into 3 sequencing batches and sequenced on a total of 3.75 lanes of a HiSeq 4000 with 2x100 paired-end reads: SeqBatch1 (P1E1, P2E1, P2E2, P3E1, P3E2, P3E3), SeqBatch2 (P1E2, P2E3, P2E4, P3E4, P3E5, P3E6) and SeqBatch3 (P1E3, P1E4, P2E5, P2E6, P3E6, P3E7, P3E8) (Table 1).

Demultiplexed FASTQ files were trimmed with Trimmomatic (version 0.36) (Bolger, Lohse, & Usadel, 2014) to remove the first 7 nucleotides corresponding to the in-line barcode (HEADCROP: 7), the Illumina adapters (ILLUMINACLIP:2:30:10), and bases with an average quality less than 20 (SLIDINGWINDOW:5:20). Paired-end reads were aligned to human genome Hg19 (GRCh37, Feb.2009 (GCA_000001405.1)) using BWA (version 0.7.12) (Li & Durbin, 2009). Duplicates were removed using PicardTools (version 1.95) (http://broadinstitute.github.io/picard/) with MarkDuplicates option. Further filtering of the reads was carried out to discard secondary alignments and reads with mapping quality lower than 30 using samtools (version 1.5) (Li et al., 2009). From now on, we will refer to those reads remaining after filtering as “reliable reads”. To retrieve the reliable reads on-target we used intersectBed from BEDTOOLS package (version 2.22.1) (Quinlan & Hall, 2010) using exome target regions provided by Agilent. In cases where we combined sequencing data, we merged filtered bam files from different hybridizations using MergeSamFiles option from PicardTools (version 1.95) (http://broadinstitute.github.io/picard/). Since the merged bam files can still contain duplicates generated during library preparation, we removed duplicates and then retrieved the reliable reads on-target using the same methodology as above.

For all previous steps, the total number of reads were counted using PicardTools (version 1.95) (http://broadinstitute.github.io/picard/) with

CollectAlignmentSummaryMetrics option. The percentage of human contamination was estimated by using positions where modern humans and chimpanzees consistently differ. We used previously published diversity data on high-coverage genomes from the *Pan* species (chimpanzee and bonobos) (De Manuel et al., 2016) and human diversity data from the 1000 Genomes Project (Auton et al., 2015), selecting positions where the human allele is observed at more than 98% frequency, and a different allele is observed in almost all *Pan* individuals (136 out of 138 chromosomes). Genome-wide, 5,646,707 chimpanzee-specific positions were identified. Using samtools mpileup (Li et al., 2009), we retrieved the number of observations of human-like alleles at these positions in the mapped reads, and estimated the human contamination as the fraction of observations for the human-like allele across all positions.

Capture performance

Capture performance was evaluated by calculating the enrichment factor (EF), capture specificity (CSp), library complexity (LC), and capture sensitivity (CS) as described in Hernandez-Rodriguez *et al* (2018). EF is calculated as the ratio of the number of reliable reads on-target to the total reads sequenced divided by the fraction of the target space (64Mb) to the genome size (~3Gb). CSp is defined as the ratio of reliable on-target reads to the total number of reliable reads. LC is defined as the number of reliable reads divided by the total number of mapped reads (containing duplicated reads). Capture sensitivity (CS) is defined as the number of target regions with an average coverage of at least one (DP1) - but also four (DP4), ten (DP10), twenty (DP20) or fifty (DP50) - divided by the total number of target regions provided by the manufacturer (n = 243,190). To calculate the average coverage of the target regions we used samtools (version 1.5) with the option bedcov (Li et al., 2009).

To generate molecular complexity or library complexity curves (MC), we used the subsampling without replacement strategy implemented in Preseq software (version 2.0.7) with c_curve option (http://smithlabresearch.org/software/preseq/) from the bam files without removing duplicates. MCs were sequentially estimated by adding the production reads, i.e. raw reads produced by sequencing, from additional hybridizations, one at a time until all hybridizations from the same library were merged (schematic representation in Figure S2).

Correlation coefficients among all pairs of study variables were estimated. Spearman's rho (cor.test(, method = "sp") from R stats package) was estimated when comparing two numeric variables. Among two categorical variables we estimated Cramér's V, derived from a chi-squared test (chisq.test() from R stats package). When comparing a numeric and categorical variable we took the square root of the R-squared statistic derived from a univariate linear model (lm() from R stats package) with a rank normal transformation (rntransform() modified from the GenABEL package to randomly split tied values) on the dependent, numerical values. In addition, univariate and multivariate type I hierarchical analysis of variances (ANOVA; anova() from R stats package) were performed to estimate the variance explained (or eta-squared) each experimental variable has on performance summary statistics (number of unique reads, reliable reads, EF, LC, CS and CSp). We down-sampled libraries to 1,500,000 reads (n=274) to remove production reads as a confounding factor. Each performance statistic was rank normal transformed with ties being randomly split to ensure normality of the dependent variable. Univariate analysis focused on the effect that subspecies, geographic sampling site, total DNA concentration, endogenous DNA concentration, percent endogenous DNA, average fragment length, pool, amount of DNA in a hybridization, hybridization and sequencing batch had on each performance statistic. A multivariate model was built to conform with experimental (hierarchical) order, such that each dependent variable (performance summary statistic, CS at DP1) was explained by ~ subspecies + site + % eDNA + average fragment size + pool + amount of DNA + hybridization + sequencing batch + error. Again, the variance explained by each independent variable was summarized by computing the eta-square statistic derived from the sums of squares for each variable using a type I hierarchical ANOVA. All statistical analyses were performed in R (version 3.5.2) (R Core Team, 2018).

Results

Sample Description

Samples were collected from 15 different PanAf sites distributed across the entire range of chimpanzees in Africa (Figure 1A and Supporting Information Table S1). The 302 screened samples had an average eDNA of 1.49%, ranging from 0 to 47.75% (Figure 1B, Supporting Information Figure S1A and Table S1) with 70.2% of the samples below 1% eDNA, according to qPCR estimates (Figure 1C). The average fragment length for screened samples was 3,479.94 bp (ranging from 72 to 17,966 bp) (Supporting Information Figure S1B and Table S1).

We observe variation on the average endogenous content among geographical sites (Figure 1B), and also variation on fragment length among geographical sites (Supporting Information Figure S1B). For instance, samples collected in a specific location such as Campo Ma'an (Cameroon) have an average eDNA of 0.02%, an extremely low value compared to the average of all sites of 1.49%. On the other hand, some sites such as Ngogo (Uganda) have samples with higher than average eDNA (6.95%) (Supporting Information Table S3). This might be explained by the influence of weather, humidity and temperature on DNA preservation and bacterial growth in the fecal sample before collection as well as a product of sample age and quality of sampling conditions (Brinkman, Schwartz, Person, Pilgrim, & Hundertmark, 2010; Goossens et al., 2000; Harestad & Bunnell, 1987; King et al., 2018; Nsubuga et al., 2004; Wedrowicz, Karsa, Mosse, & Hogan, 2013).

A total of 60 samples with a mean percent endogenous content of 0.58% (range from 0.41% to 0.85%), and with a median human contamination of 0.0875% (range from 0.04% to 7.50%) from all four chimpanzee subspecies and 14 geographic sites were carried forward into target capture enrichment experiments (Table S2). After double-inline-barcoded library production, the 60 samples were placed into 3 pools with 10, 20 and 30 samples each. Samples were divided into pools based on their percent endogenous content, such that those samples with higher levels of percent endogenous content were in P1 with 10 samples (mean = 0.81) and those with the smallest were in P3 with 30 samples (mean = 0.49; P2 mean = 0.69) (Figure 3A). As such the percent endogenous DNA is highly structured among the three pools, explaining 81% of the variation in eDNA (univariate linear model using rank normal transformed % eDNA; p-value = $2.05\text{x}10_{-91}$) (Supporting Information Figure S4A).

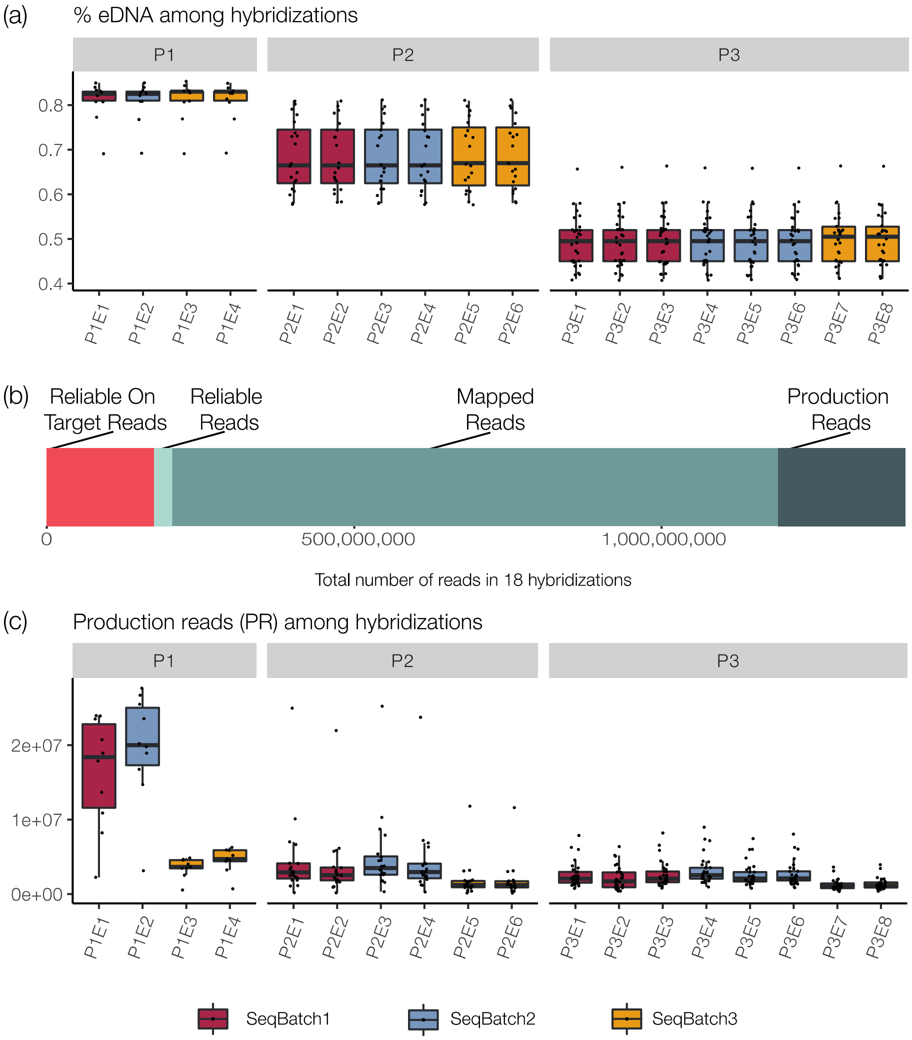


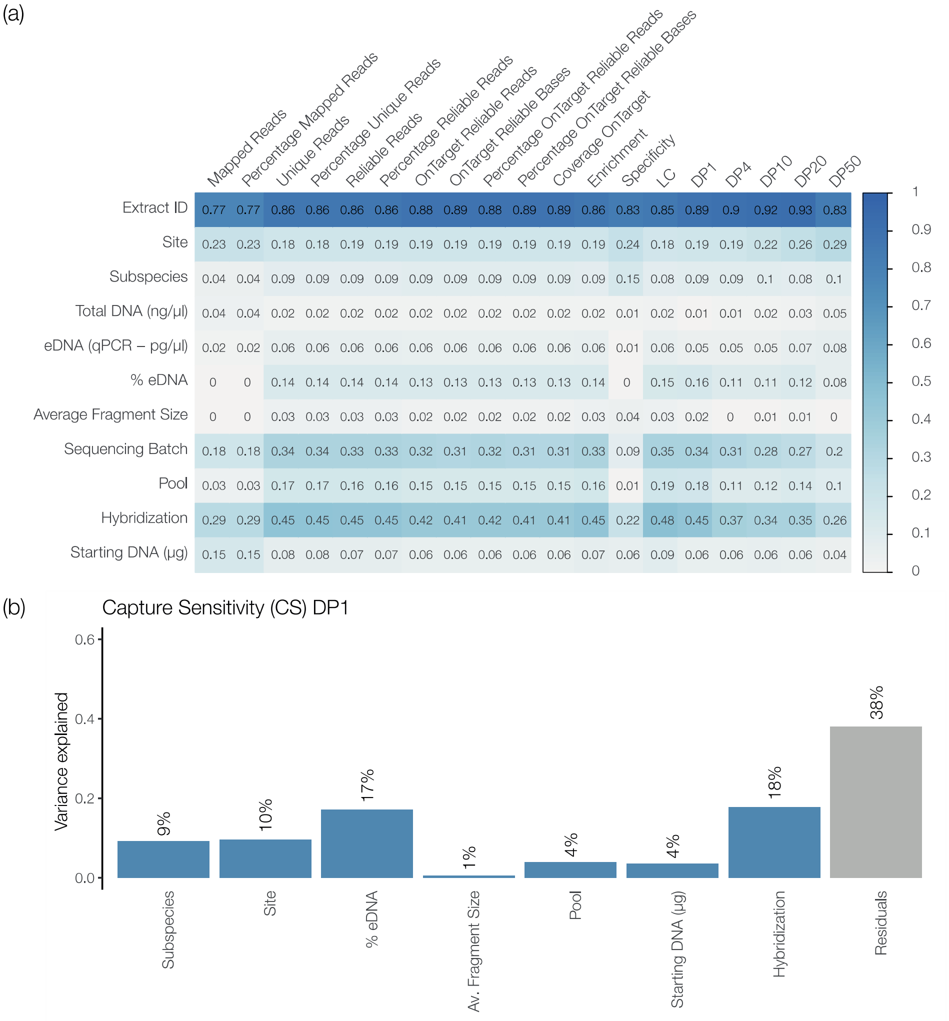
FIGURE 3. Capture performance and sequencing. (a) Percentage of eDNA among hybridizations, structured by pools (P1, P2 and P3). (b) Sequencing stats across all samples for the 18 hybridizations in 3,75 HiSeq 4000 lanes. (c) Distribution of production reads across 18 hybridizations. The colors red, blue and yellow found in the box plots for figure (a) and (c) denote the sequencing batch to which each hybridization was assigned. In the boxplots, lower and upper hinges correspond to first and third quartiles and the lower and upper whiskers extend to the smallest or largest value no further than 1.5 times the interquartile range (distance between the 1st and 3rd quartile).

Read Summary Statistics and Capture Performance

As illustrated in Figure 3B across a total of 18 hybridization experiments sequenced we obtained ~1.40 billion reads distributed among 3 pools. Of those, ~1.19 billion were mapped reads (85.19%), with ~203 million reads being considered duplicate-free, reliable reads (14.6%). After removing off-target reads, we obtained a total of ~174 million on-target-reliable reads (12.48%) (Supporting Information Table S4, Supporting Information Figure S3A). However, on average each hybridization experiment yielded an average of 17.35% on-target-reliable reads, with a range of 4.15% in our earliest experiments to 34.85% in our later experiments (Supporting information Table S5). The observed high levels of duplicates are a consequence of the low endogenous content of the samples and the exhaustion of library complexity during sequencing; we will elaborate on outcome and improvements below.

The ~1.40 billion reads were not equally distributed among the 3 pools (production reads explained by pools; r_2 = 0.41, p-value = 3.24x10_{-16}) or 18 hybridizations (r_2 = 0.62, p-value = 2.59x10_{-30}). In fact, two hybridizations of P1 (P1E1, P1E2) were sequenced to an average depth of 18 million reads, while all other hybridizations had an average depth of 3 million reads (Figure 3C). This very deep sequencing, in P1E1 and P1E2, led to a point where the library complexity was exhausted, leading to the sequencing of a high number of PCR duplicates (Supporting Information Figure S3A, S3B and Supporting Information Table S5). We therefore reduced subsequent sequencing efforts, as discussed in section “Optimization of required production reads”, for the remaining replicate hybridizations.

All capture performance summary statistics (Supporting Information Table S4), to the exception of capture specificity (CSp), are strongly correlated with the number of production reads acquired (median correlation coefficient = 0.422, CI = 0.03 to 0.93; Supporting information Figure S4A, Table S6). Given this, and also because of the distinct difference in the number of production reads between P1E1 and P1E2 and all other hybridizations we down-sampled all experiments to 1.5 million production reads, retaining only those 274 sample/hybridization experiments with 1.5 million production reads, and re-estimated all capture performance summary statistics (Supporting Information Figure S4B, Table S7 and S8). The effect each experimental variable has on performance was estimated in a univariate linear model after rank normal transforming each summary statistic (Figure 4A). We observed a near uniformity in the variance explained by each experimental variable across each performance statistics. In short, the average, ranked order of variance explained by each explanatory variable are sample (86.50%), hybridization (38.72%), sequencing batch (28.78%), site (20.5%), pool (13%), % endogenous DNA (11%), subspecies (8.85%), starting DNA amount (7.35%), endogenous DNA concentration (5.14%), average fragmentation size (2.12%,), and total DNA concentration (2.07%). Given these observations we may conclude that variation in hybridization and sequencing are crucial to performance. However, sample quality and starting material varies among our hybridizations and sequencing batches. These tendencies can be observed in Figure 5A-C. We account for this in a multivariate linear model followed by a decomposition of the variance in a type I hierarchical analysis of variance (ANOVA). To do so we fit a linear model ordered by experimental choices, as described in materials and methods, to explain Capture Sensitivity (CS) at DP1 which is being used here as an example of capture performance. This model indicates that hybridization explains, on average, an attenuated 17.80% of the variation in performance, followed by percent endogenous content (17.11%), site (9.62%), subspecies (9.26%), pool (3.92%) and then the amount of DNA in the hybridization (3.58 %) (Figure 4B). Results for all other performance summary statistics mirror those for CS at DP1 and can be seen in Figure S5.



	Mapped Reads	Percentage Mapped Reads	Unique Reads	Percentage Unique Reads	Reliable Reads	Percentage Reliable Reads	OnTarget Reliable Reads	OnTarget Reliable Bases	Percentage OnTarget Reliable Reads	Percentage OnTarget Reliable Bases	Coverage OnTarget	Enrichment	Specificity	LC	DP1	DP4	DP10	DP20	DP50
Extract ID	0.77	0.77	0.86	0.86	0.86	0.86	0.88	0.89	0.88	0.89	0.89	0.86	0.83	0.85	0.89	0.9	0.92	0.93	0.83
Site	0.23	0.23	0.18	0.18	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.24	0.18	0.19	0.19	0.22	0.26	0.29
Subspecies	0.04	0.04	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.15	0.08	0.09	0.09	0.1	0.08	0.1
Total DNA (ng/µl)	0.04	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.01	0.02	0.03	0.05
eDNA (qPCR – pg/µl)	0.02	0.02	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.01	0.06	0.05	0.05	0.05	0.07	0.08
% eDNA	0	0	0.14	0.14	0.14	0.14	0.13	0.13	0.13	0.13	0.13	0.14	0	0.15	0.16	0.11	0.11	0.12	0.08
Average Fragment Size	0	0	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.03	0.04	0.03	0.02	0	0.01	0.01	0
Sequencing Batch	0.18	0.18	0.34	0.34	0.33	0.33	0.32	0.31	0.32	0.31	0.31	0.33	0.09	0.35	0.34	0.31	0.28	0.27	0.2
Pool	0.03	0.03	0.17	0.17	0.16	0.16	0.15	0.15	0.15	0.15	0.15	0.16	0.01	0.19	0.18	0.11	0.12	0.14	0.1
Hybridization	0.29	0.29	0.45	0.45	0.45	0.45	0.42	0.41	0.42	0.41	0.41	0.45	0.22	0.48	0.45	0.37	0.34	0.35	0.26
Starting DNA (µg)	0.15	0.15	0.08	0.08	0.07	0.07	0.06	0.06	0.06	0.06	0.06	0.07	0.06	0.09	0.06	0.06	0.06	0.06	0.04

FIGURE 4. Analysis of variance. (a) Estimated variance explained from univariate linear models after rank normal transforming each performance summary statistic (columns). LC stands for library complexity and DP describes read depth at different cutoffs (1, 4, 10, 20 and 50 reads) (b) Multivariate type I ANOVA of the experimental variables affecting Capture Sensitivity (CS) at depth 1. Both models are built down-sampling libraries to 1,500,000 reads.

Relevance of Equi-Endogenous Pools

The observations of Hernandez-Rodriguez et al. and White et al. suggest that pooling libraries by eDNA concentration (in equi-endogenous pools) prior to hybridization capture should reduce or remove the effect of variation in eDNA across samples on targeted capture sequencing performance. Indeed, eDNA did not have a major influence on production reads or on-target reads, although a slightly positive trend can be observed in some hybridizations of P2 (Supporting Information Figure S6). Without equi-endogenous pooling, it is expected that samples with higher eDNA would accumulate more on-target reads than other samples with lower eDNA as observed by Hernandez-Rodriguez et al. The reason why in P2 we find some outliers might be traced to both pipetting variations and inaccurate endogenous measurements from qPCR values due to the presence of inhibitors (Morin et al., 2001). Avoiding outliers is extremely important in limiting variability within a pool. For example, sample N183-5 accumulated 29.4% of total raw reads in P2, when a value 5% (1/20 of 100%) was expected (Supporting Information Figure S7).

Impact of Amount of Starting DNA for Capture on Library Complexity

One major decision when performing capture experiments is the amount of starting DNA in the pool. In twelve hybridizations we used the manufacturer's suggested amount of starting material, 1 µg for each pool. For the last two hybridizations of each pool (a total of six hybridizations) we doubled the starting material, up to 2 µg of pooled libraries (Table 1). With this approach we aimed to test the effect on the final LC when doubling the amount of DNA and to determine how much DNA should be used for fecal capture experiments.

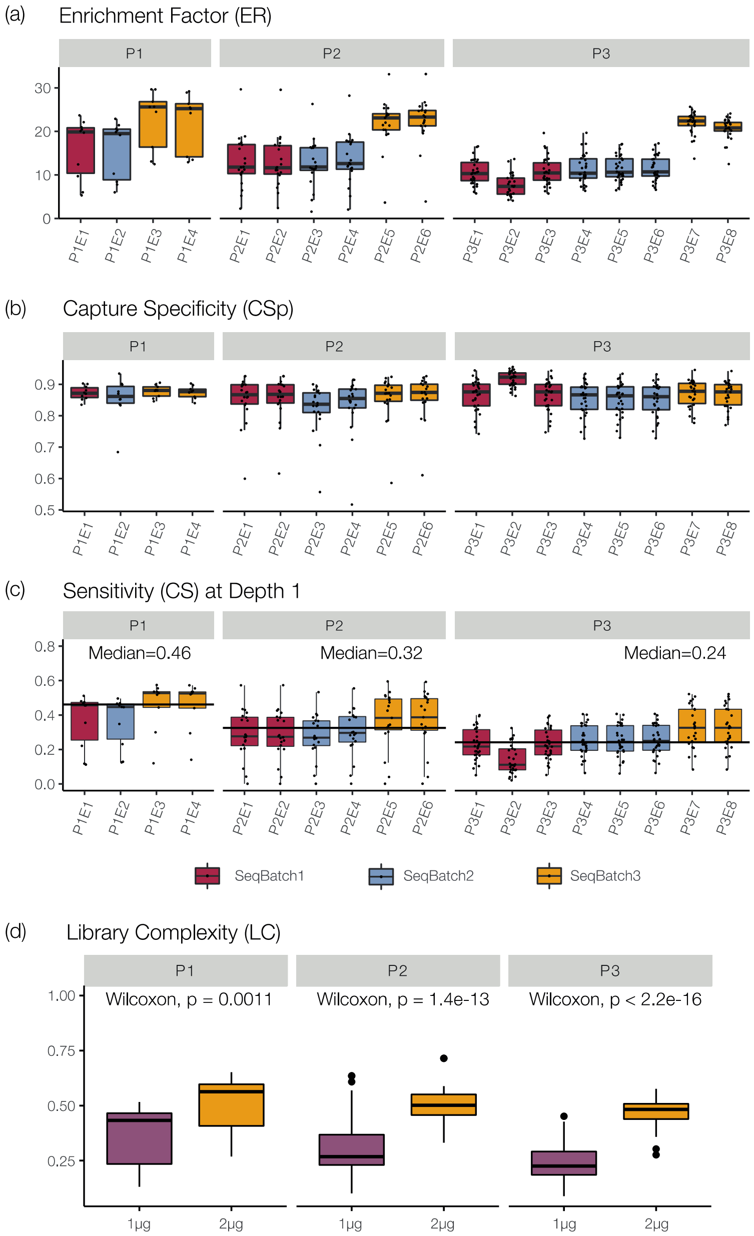


FIGURE 5. Summary stats after down-sampling to 1,500,000 reads: (a) Enrichment factor and (d) Capture Specificity (c) Capture Sensitivity at depth 1 for the 18 hybridizations in P1, P2 and P3; colors illustrate sequencing batch. (d) Library complexity contrasting the amount of starting DNA (1 μg or 2 μg) in down-sampled data and structured by pools (P1=Pool1, P2=Pool2, P3=Pool3). See Figure 2 for more details on pools. In the boxplots, lower and upper hinges correspond to first and third quartiles and the lower and upper whiskers extend to the smallest or largest value no further than 1.5 times the interquartile range (distance between the 1st and 3rd quartile).

We observed an average increase of 2.8-fold in LC for experiments using 2 µg of total DNA in the hybridization relative to those using 1 µg (Supporting Information Figure S3B). However, given that production reads also vary between these two conditions, we down-sampled the data to 1,500,000 reads per library. After this correction we still observed 2-fold higher LC when starting the experiments with 2 µg of total DNA in all pools (Figure 5D).

Molecular complexity, as influenced by the amount total DNA in a hybridization, was further investigated by evaluating the relationship between MC and production reads in a MC curve analysis. The MC curve for each hybridization was obtained by subsampling without replacement their reads. The results supported the conclusion above: increasing the amount of total DNA in the hybridization increased the MC (Supporting Information Figure S8). Therefore, whenever there is sufficient library available, it is advisable to start with 2 µg rather than 1 µg.

Molecular Complexity and Capture Sensitivity

One of the critical aspects to increase coverage is to acquire as many unique on-target reads as possible without exhausting the library's molecular complexity. We applied a subsampling without replacement method to assess how many mapped reads are unique after incrementally adding production reads from replicate hybridizations. In principle, molecular complexity curves that plateau quickly are derived from low complexity libraries, and conversely high complexity libraries may not reach plateau. Thereby the plateau indicates when there are no new unique reads to be sampled or sequenced (see Supporting Information Figure S2 for a schematic representation).

We performed the analysis of molecular complexity in libraries belonging to P3 since more hybridization replicates were available (8 in total) for 30 libraries. We found that for the majority of the libraries, performing additional hybridizations increased the number of unique reads retrieved (Supporting Information Figure S9, example library N259-5). However, there were libraries that quickly hit exhaustion where performing additional hybridizations would add little extra information (Supporting Information Figure S9, example library Kay2-32). Overall, by performing additional hybridizations, it was possible to retrieve new unique reads and thus increase the final coverage (Figure 6A), because libraries themselves were not exhausted but merely their hybridization-captured molecules reached exhaustion.

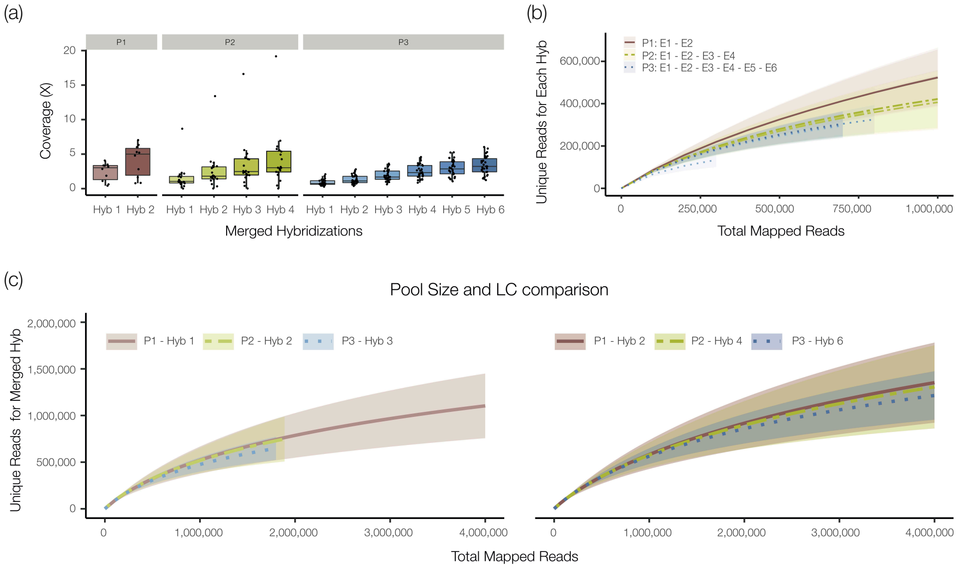


FIGURE 6. Analysis of coverage and LC with hybridizations done with 1 µg. (a) Coverage after merging data from additional hybridizations with up to 2, 4 and 6 for P1, P2 and P3. (b) Comparison of average LC curves of individual hybridizations belonging to pools with different size. Each line is the average of libraries within each hybridization and the surrounding area is the standard deviation. (c) Two examples comparing the effect of pool size on the average LC curves from merged hybridization: P1 (10 samples) - 1 hybridization, P2 (20 samples) – 2 hybridizations and P3 (30 samples) – 3 hybridizations; and P1 (10 samples) - 2 hybridizations,

P2 (20 samples) – 4 hybridizations and P3 (30 samples) – 6 hybridizations. Sample Lib1-6D in P2 was removed from the analysis due to low coverage.

Following the same strategy, we calculated the sensitivity in P1, P2 and P3 (4, 6 and 8 replicates respectively). After cumulatively adding data from replicate hybridizations we covered 85.57% in P1 (95% CI: 74.78-96.36%), 76.23% in P2 (95% CI: 64.55-87.91%) and 79.83% in P3 (95% CI: 74.44-85.22%) on average of the target space, with at least 1 read (Supporting Information Figure S10). Interestingly, no sample covered 100% of target space. Looking carefully into this, we observed that precisely the same 3,804 regions (1.54%) were never covered in any replicate hybridizations, suggesting that some regions are either difficult to capture (Kong, Lee, Liu, Hirschhorn, & Mandl, 2018) or are too divergent between *Homo* and *Pan* to either capture or map these particular sequences (Supporting Information Figure S11).

For deeper coverage of at least 4 or 10 reads, we still observed a positive progression, with each additional hybridization increasing coverage, indicating that additional hybridizations would result in an increase of the proportion of the genome covered at these depths as well (Supporting Information Figure S10).

Optimization of Required Production Reads

Assessing the amount of sequencing needed is one of the major decisions when planning an experiment. As a result of the low eDNA content of most fecal samples, derived libraries can easily reach saturation (i.e., high levels of duplicated reads). Therefore, sequencing depth should be carefully calculated. Without previous knowledge, we sequenced the first 2 hybridizations for P1, the first 4 hybridizations for P2, and the first 6 hybridizations for P3 in three lanes of a HiSeq 4000. For P1 only

~6% and for P2 and P3 only ~13% of production reads were unique reads (Supporting Information Table S5), indicative of high levels of PCR duplicates due to library exhaustion. To avoid over-sequencing in our next experiments, we set an arbitrary threshold to recover approximately 20% of the “informative” data (unique reads) available in a hybridization experiment. Using the data from SeqBatch 1 and 2, we estimated that on average, for samples with less than 1% eDNA, we would sequence at most 2 million mapped reads per library (Figure S12). Given that 80% of reads mapped to the genome in these experiments, we estimated that we would need to sequence at most 2.5 million production reads per library (Supporting Information Table S5).

To test these estimates, we sequenced the remaining hybridizations (P1E3, P1E4, P2E5, P2E6, P3E7, P3E8) in three-fourths of a HiSeq 4000 lane. The number of average production reads obtained were 3.5, 2.0 and 1.5 million for libraries in hybridizations from P1, P2, and P3, respectively. On average ~38% (range: 8.09-50.81%) of reads were unique reads in all pools (Supporting Information Figure S13). We note that these values exceeded what we observed in the previous hybridization experiments. An outcome we attribute to the increase in starting material (2 μg), also used in these experiments, as noted above.

Pooling Strategy

Choosing how many samples to pool is a difficult decision, since little is known on how the pool size will affect the final molecular complexity. Taking advantage of our pooling strategy (Figure 2), we assessed the effect of size on the average library complexity for all samples within each hybridization with a subsampling without replacement strategy.

When only a single hybridization was performed, a single library within a pool of 10, 20 or 30 would, on average, result in a similar number of unique molecules (Figure 6B, Supporting Information Figure S14). However, there is a tendency for samples in smaller pools (P1) to perform better than those in larger pools. This could be explained by our experimental design, where samples with higher eDNA content are in smaller pools. However, let us address this possibility here. Using CS as an example summary statistic, we observed that CS is higher for pools with smaller numbers of samples in them (Figure 5C). Given median estimates, a pool of 10 libraries (median CS = 0.46) had 1.44-fold higher CS than a pool of 20 libraries (median CS = 0.32), and 1.92-fold higher than a pool of 30 libraries (median CS = 0.24). Between a pool of 20 and a pool of 30, the ratio was 1.33-fold (Figure 5C and Supporting Information Figure S15). If we remove the effect of having a variable number of production reads across experiments by down-sampling, this observation still remains (Supporting Information Figure S16). That is, smaller pools do have higher CS estimates, and pools linearly account for 18% of the variation in CS (univariate ANOVA, p-value=$3.47x10_{-12}$ (Figure 4A)). Finally, if we correct for all experimental variables with a multivariate analysis, as done above, we show that 'Pool' only accounts for 4% of the variation in CS (Figure 4B), but the effect of pool size remains significant (multivariate ANOVA, p-value = $2.7x10_{-4}$; Supporting Information Figure S16). However, this effect on CS attenuates with additional hybridizations (4, 6 and 8, for P1, P2 and P3 respectively) for the same pool (Supporting Information Figure S17). Moreover, a similar outcome can be observed when comparing the effect of pool size on LC. After sequentially adding data from replicate hybridizations in each pool (see Supporting Information Figure S2 for a schematic representation), we can acquire the same number of unique reliable reads (Figure 6C, Supporting Information S16).

Discussion

Capturing host DNA from fecal samples is a challenging endeavor. Previous work has shown that the retrieval of genomic data from fecal samples by target enrichment methodologies is a feasible and powerful tool for conservation and evolutionary studies (Perry, 2014; Snyder-Mackler et al., 2016). However, obtaining good quality and quantity DNA from fecal samples is not always possible. Because of that, many studies have characterized the technical difficulties of capturing DNA from non-invasive samples and proposed different strategies (Hernandez-Rodriguez et al., 2018; van der Valk et al., 2017; White et al., 2019). Van der Valk et al. (2017) captured the whole mitochondrial genome but no autosomal regions, and describe the biases introduced during capture such as DNA fragment size, jumping PCR and divergence between bait and target species. The study performed by Hernandez-Rodriguez et al. (2018) systematically analyzed the capture performance and library complexity. While they described that pooling different libraries into the same hybridization is feasible, they did not discuss how many of them should be pooled. Also, they concluded that performing multiple libraries from the same extract or even from different extracts from the same sample can increase the final complexity. Finally, they recommended performing two capture rounds for the same library. On the other hand, White et al. (2019) suggested to do only one capture round, at least when eDNA is higher than 2-

3%, stressing the importance of pooling libraries as well as taking into consideration the eDNA content, as first proposed by Hernandez-Rodriguez et al.

The present study addresses these gaps left unexplored by the previous studies. We focused our analysis on a representative set of samples with very low proportions of endogenous content (< 1%) as are often found in the field. After screening 302 samples, we found that up to 70% of samples are below this threshold, similar to what was already described (White et al., 2019). Hence, if time and economic reasons hinder the ability to collect and select the best samples, the only available one(s) might have low eDNA. This may be a common situation when using historical samples, aiming for a large sample size, or if an interesting sampling location is particularly challenging in terms of low eDNA (such as Campo Ma'an, Figure 1B).

For these reasons, it is of utmost importance to characterize ways to maximize the amount of data to be recovered from these types of samples. In this regard, we have extensively evaluated how to increase library complexity without doing more extractions or library preparations from the same sample, how many libraries to pool together, and how much starting amount of DNA should be used in a capture, as well as the impact of endogenous content for pooling.

Consistent with previous findings (Hernandez-Rodriguez et al., 2018; White et al., 2019), we determined that assessing the endogenous content of fecal samples and pooling them equi-endogenously is a practical way to equally distribute raw reads between samples. Importantly, the correct estimation of the proportion of eDNA is key for the success of this method. Thus, we recommend the usage of shotgun sequencing (Hernandez-Rodriguez et al., 2018) rather than qPCR estimates, since the later can easily fluctuate due to the presence of inhibitors (Morin et al., 2001).

In regard to the performance of target capture sequencing experiments, gaining new unique reads is crucial to reach higher sensitivity, which is a good predictor of capture success. Here, we have established an approach to obtain new unique reads using the same prepared libraries. Since it is mainly during capture experiments when the molecular diversity is reduced, we propose to perform additional hybridizations from the same library so the final coverage can reach higher values. If the library complexity is already very low, the only solution is to re-extract DNA or prepare a new library from the same sample (Hernandez-Rodriguez et al., 2018).

We observed a better performance (MC and CS) in small pools, when evaluating initial results derived from the entire dataset. However, after correcting for other variables that differ among pools, the effect is attenuated and can only explain ~4% of the variance, an effect that may be largely negligible for most studies. Moreover, performing additional hybridizations can also compensate for this effect. Therefore, we do not conclude, based on this data, that pool size is a major contributor to performance. However, in cases where libraries have small proportions of eDNA, we would advocate for the reduction of the number of samples per pool so that pipetting volumes may remain larger, and as a consequence variability due to pipetting error may be reduced. Otherwise when the eDNA proportion is not a limiting factor, pooling more libraries together and performing additional hybridizations can be a good strategy.

It is worth noting that without taking into consideration individual sample quality and the amount of starting material used, one of the most influential variables on the performance of target capture enrichment experiments is the hybridization experiment itself. After accounting for all other variables, it still explains 18% of the variation. This is due to the technical complexity and variability inherent to these experiments. Careful equipment optimization, material selection, preparation and experience will aid in minimizing this variation, although it is likely to remain a sensitive experiment that requires diligence.

Finally, we have illustrated that a sequencing effort of exome-captured fecal samples with low eDNA (< 1%) should be set at ~3 million reads per library in a pool to avoid exhausting the molecular complexity. We have benefited from the usage of double-barcoded and double-indexed libraries to multiplex many samples in a single sequencing lane. This becomes a great advantage because we can utilize high throughput sequencing technologies at a lower price per read.

To summarize, when starting a project involving fecal samples, we recommend screening your set of samples based on quantity and quality of the DNA extracted. If having related or identical individuals in the study should be avoided, microsatellite genotyping could be an option, helping as well to discard samples with high amount of PCR inhibitors. Further selection of samples should be based on the proportion of eDNA; we recommend using shotgun sequencing from the prepared libraries. Performing re-extractions of the most valuable samples and preparing replicate libraries from each extract can help increase the final molecular complexity. As we have shown here, another approach to achieve higher molecular complexity is based on conducting additional hybridizations of the captured libraries, always pooling libraries in an equi-endogenous manner, and starting with more library material than the standard protocol suggests. Finally, we suggest not sequencing the captured libraries very deeply, since their molecular complexity is already very low and over-sequencing can result in rapidly depleting the economic feasibility of the experiment.

In the study presented here we have thoroughly explored approaches to increase the molecular diversity and capture sensitivity and hence the final coverage of exome captured fecal samples with extremely low endogenous content in an attempt to help laboratories facing the challenges of working with non-invasive samples.

Acknowledgments

We would like to thank Linda Vigilant, Christophe Boesch and Marco Telford for helpful discussion and Roland Schroeder, Alan Riedel and Katharina Madl, for guidance and assistance in the laboratory. We thank Emmanuel Dilambaka, Devla Dowd, Annemarie Goedmakers, Vincent Lapeyre, Vera Leinert, Mizuki Murai, Emmanuelle Normand, Robinson Orume, Alexander Tickle, Els Ton, Joost van Schijndel, Sergio Marrocoli, Amelia Meier, Volker Sommer, Martijn Ter Heegde, Nadege Wangue Njomen, Joshua M Linder, Hilde Vanleeuwe, Jean Claude Dengui, Paul Telfer and Yasmin Moebius for assistance in field site coordination and sample collection. C.F. is supported by the "La Caixa" doctoral fellowship program. M.A.E. is supported by an FPI (Formación de Personal Investigador) PRE2018-083966 from Ministerio de Ciencia, Universidades e Investigación. The Pan African Programme: The Cultured Chimpanzee (PanAf) is generously funded by the Max Planck Society, the Max Planck Society Innovation Fund and the Heinz L. Krekeler Foundation. E.L is supported by CGL2017-82654-P (MINECO/FEDER,UE). M.K. is supported by "la Caixa" Foundation (ID 100010434), fellowship code LCF/BQ/PR19/11700002. T.M.-B is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 864203), BFU2017-86471-P (MINECO/FEDER, UE), "Unidad de Excelencia María de Maeztu", funded by the AEI

(CEX2018-000792-M), Howard Hughes International Early Career, Obra Social "La Caixa" and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880). We thank the following government agencies for their support in conducting field research in their countries: Ministère de la Recherche Scientifique et de l'Innovation, Cameroon, Ministère des Forêts et de la Faune, Cameroon, Ministère des Eaux et Forêts, Cote d'Ivoire, Ministère de l'Enseignement Supérieur et de la Recherche Scientifique, Cote d'Ivoire, Agence Nationale des Parcs Nationaux, Gabon, Centre National de la Recherche Scientifique (CENAREST), Gabon, Société Equatoriale d'Exploitation Forestière (SEEF), Gabon, Ministère de l'Agriculture de l'Elevage et des Eaux et Forets, Guinea, Instituto da Biodiversidade e das Áreas Protegidas (IBAP), Guinea Bissau, Ministro da Agricultura e Desenvolvimento Rural, Guinea-Bissau, Forestry Development Authority, Liberia, National Park Service, Nigeria, Ministère de l'Economie Forestière, R-Congo, Ministère de le Recherche Scientifique et Technologique, R-Congo, Direction des Eaux, Forêts et Chasses, Senegal,Tanzania Commission for Science and Technology, Tanzania, Tanzania Wildlife Research Institute, Tanzania, Uganda National Council for Science and Technology (UNCST), Uganda, Uganda Wildlife Authority, Uganda, National Forestry Authority, Uganda.

References

Arandjelovic, M., Guschanski, K., Schubert, G., Harris, T. R., Thalmann, O., Siedel, H., & Vigilant, L. (2009). Two-step multiplex polymerase chain reaction improves the speed and accuracy of genotyping using DNA from noninvasive and museum samples. *Molecular Ecology Resources*, *9*(1), 28–36. doi: 10.1111/j.1755-0998.2008.02387.x

Arandjelovic, M., Head, J., Rabanal, L. I., Schubert, G., Mettke, E., Boesch, C., … Vigilant, L. (2011). Non-invasive genetic monitoring of wild central chimpanzees. *PLoS ONE*, *6*(3), e14761. doi: 10.1371/journal.pone.0014761

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, Vol. 526, pp. 68–74. doi: 10.1038/nature15393

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Brinkman, T. J., Schwartz, M. K., Person, D. K., Pilgrim, K. L., & Hundertmark, K. J. (2010). Effects of time and rainfall on PCR success using DNA extracted from deer fecal pellets. *Conservation Genetics*, *11*(4), 1547–1552. doi: 10.1007/s10592-009-9928-7

Burbano, H. A., Hodges, E., Green, R. E., Briggs, A. W., Krause, J., Meyer, M., … Pääbo, S. (2010). Targeted investigation of the neandertal genome by array-based sequence capture. *Science*, *328*(5979), 723–725. doi: 10.1126/science.1188046

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. H. S., Samaniego, J. A., … Gilbert, M. T. P. (2018). Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution*, *9*(2), 410–419. doi: 10.1111/2041-210X.12871

Carpenter, M. L., Buenrostro, J. D., Valdiosera, C., Schroeder, H., Allentoft, M. E., Sikora, M., … Bustamante, C. D. (2013). Pulling out the 1%: Whole-Genome capture for the targeted enrichment of ancient dna sequencing libraries. *American Journal of Human Genetics*, *93*(5), 852–864. doi: 10.1016/j.ajhg.2013.10.002

Csilléry, K., Johnson, T., Beraldi, D., Clutton-Brock, T., Coltman, D., Hansson, B., … Pemberton, J. M. (2006). Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. *Genetics*, *173*(4), 2091–2101. doi: 10.1534/genetics.106.057331

De Barba, M., Waits, L. P., Genovesi, P., Randi, E., Chirichella, R., & Cetto, E. (2010). Comparing opportunistic and systematic sampling methods for non-invasive genetic monitoring of a small translocated brown bear population. *Journal of Applied Ecology*, *47*(1), 172–181. doi: 10.1111/j.1365-2664.2009.01752.x

De Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., … Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, *354*(6311), 477–481. doi: 10.1126/science.aag2602

Ferreira, C. M., Sabino-Marques, H., Barbosa, S., Costa, P., Encarnação, C., Alpizar-Jara, R., … Alves, P. C. (2018). Genetic non-invasive sampling (gNIS) as a cost-effective tool for monitoring elusive small mammals. *European Journal of Wildlife Research*, *64*(4). doi: 10.1007/s10344-018-1188-8

Fickel, J., Lieckfeldt, D., Ratanakorn, P., & Pitra, C. (2007). Distribution of haplotypes and microsatellite alleles among Asian elephants (Elephas maximus) in Thailand. *European Journal of Wildlife Research*, *53*(4), 298–303. doi: 10.1007/s10344-007-0099-x

Fischer, A., Wiebe, V., Pääbo, S., & Przeworski, M. (2004). Evidence for a Complex

Demographic History of Chimpanzees. *Molecular Biology and Evolution*, *21*(5), 799–808. doi: 10.1093/molbev/msh083

Goossens, B., Chikhi, L., Utami, S. S., De Ruiter, J., & Bruford, M. W. (2000). A multi-samples, multi-extracts approach for microsatellite analysis of faecal samples in an arboreal ape. *Conservation Genetics*, *1*(2), 157–162. doi: 10.1023/A:1026535006318

Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., Kronenberg, Z. N., Munson, K. M., … Eichler, E. E. (2016). Long-read sequence assembly of the gorilla genome. *Science*, *352*(6281), aae0344. doi: 10.1126/science.aae0344

Harestad, A. S., & Bunnell, F. L. (1987). Persistence of Black-Tailed Deer Fecal Pellets in Coastal Habitats. *The Journal of Wildlife Management*, *51*(1), 33. doi: 10.2307/3801624

Hernandez-Rodriguez, J., Arandjelovic, M., Lester, J., de Filippo, C., Weihmann, A., Meyer, M., … Marques-Bonet, T. (2018). The impact of endogenous content, replicates and pooling on genome capture from faecal samples. *Molecular Ecology Resources*, *18*(2), 319–333. doi: 10.1111/1755-0998.12728

Hicks, A. L., Lee, K. J., Couto-Rodriguez, M., Patel, J., Sinha, R., Guo, C., … Williams, B. L. (2018). Gut microbiomes of wild great apes fluctuate seasonally in response to diet. *Nature Communications*, *9*(1), 1786. doi: 10.1038/s41467-018-04204-w

Inoue, E., Akomo-Okoue, E. F., Ando, C., Iwata, Y., Judai, M., Fujita, S., … Yamagiwa, J. (2013). Male genetic structure and paternity in western lowland gorillas (Gorilla gorilla gorilla). *American Journal of Physical Anthropology*, *151*(4), 583–588. doi: 10.1002/ajpa.22312

King, S. R. B., Schoenecker, K. A., Fike, J. A., & Oyler-McCance, S. J. (2018). Long-term persistence of horse fecal DNA in the environment makes equids particularly good candidates for noninvasive sampling. *Ecology and Evolution*, *8*(8), 4053–4064. doi: 10.1002/ece3.3956

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, *40*(1), 1–8. doi: 10.1093/nar/gkr771

Kong, S. W., Lee, I. H., Liu, X., Hirschhorn, J. N., & Mandl, K. D. (2018). Measuring coverage and accuracy of whole-exome sequencing in clinical context. *Genetics in Medicine*, *20*(12), 1617–1626. doi: 10.1038/gim.2018.51

Kühl, H. S., Boesch, C., Kulik, L., Haas, F., Arandjelovic, M., Dieguez, P., … Kalan, A. K. (2019). Human impact erodes chimpanzee behavioral diversity. *Science (New York, N.Y.)*, *363*(6434), 1453–1455. doi: 10.1126/science.aau4532

Lester, J.D., Vigilant, L., Gratton, P., McCarthy, M.S., Barratt, C.D., Dieguez, P., ... Arandjelovic, M., (2020). Recent genetic connectivity and clinal variation in chimpanzees. In review.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V., Muzny, D. M., … Wilson, R. K. (2011). Comparative and demographic analysis of orangutan genomes. *Nature*, *469*(7331), 529–533. doi: 10.1038/nature09687

Maricic, T., Whitten, M., & Pääbo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE*, *5*(11), e14004. doi: 10.1371/journal.pone.0014004

Mengüllüoğlu, D., Fickel, J., Hofer, H., & Förster, D. W. (2019). Non-invasive faecal sampling reveals spatial organization and improves measures of genetic diversity for the conservation assessment of territorial species: Caucasian lynx as a case species. *PLoS ONE*, *14*(5). doi: 10.1371/journal.pone.0216549

Mikkelsen, T. S., Hillier, L. W., Eichler, E. E., Zody, M. C., Jaffe, D. B., Yang, S. P., … Waterston, R. H. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87. doi: 10.1038/nature04072

Morin, P. A., Chambers, K. E., Boesch, C., & Vigilant, L. (2001). Quantitative PCR analysis of DNA from noninvasive samples fro accurate microsatellite genotyping of wild chimpanzees. *Molecular Ecology*, 1835–1844.

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*(3), 347–356. doi: 10.1007/BF02382630

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*(7), 2089–2094. doi: 10.1111/j.1365-294X.2004.02207.x

Orkin, J. D., Yang, Y., Yang, C., Yu, D. W., & Jiang, X. (2016). Cost-effective scat-detection dogs: Unleashing a powerful new tool for international mammalian conservation biology. *Scientific Reports*, *6*(1), 34758. doi: 10.1038/srep34758

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*(4), 177–187. doi: 10.1016/j.tig.2010.01.001

Perry, G. H. (2014). The Promise and Practicality of Population Genomics Research with Endangered Species. *International Journal of Primatology*, *35*(1), 55–70. doi: 10.1007/s10764-013-9702-z

Perry, G. H., Marioni, J. C., Melsted, P., & Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular Ecology*, *19*(24), 5332–5344. doi: 10.1111/j.1365-294X.2010.04888.x

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., … Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471–475. doi: 10.1038/nature12228

Primmer, C. R. (2009). From conservation genetics to conservation genomics. *Annals of the New York Academy of Sciences*, Vol. 1162, pp. 357–368. doi: 10.1111/j.1749-6632.2009.04444.x

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

Reiners, T. E., Encarnação, J. A., & Wolters, V. (2011). An optimized hair trap for non-invasive genetic studies of small cryptic mammals. *European Journal of Wildlife Research*, *57*(4), 991–995. doi: 10.1007/s10344-011-0543-9

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, *22*(5), 939–946. doi: 10.1101/gr.128124.111

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., … Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, *483*(7388), 169–175. doi: 10.1038/nature10842

Schwartz, M. K., Luikart, G., & Waples, R. S. (2007). Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology and Evolution*, Vol. 22, pp. 25–33. doi: 10.1016/j.tree.2006.08.009

Shafer, A. B., Wolf, J. B., Alves, P. C., Bergströ, L., Bruford, M. W., Brä nnströ, I., … Zielin, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, *30*(2), 78–87. doi: 10.1016/j.tree.2014.11.009

Snyder-Mackler, N., Majoros, W. H., Yuan, M. L., Shaver, A. O., Gordon, J. B., Kopp, G. H., … Tung, J. (2016). Efficient genome-wide sequencing and low-coverage pedigree analysis from noninvasively collected samples. *Genetics*, *203*(2), 699–714. doi: 10.1534/genetics.116.187492

Städele, V., & Vigilant, L. (2016). Strategies for determining kinship in wild populations using genetic data. *Ecology and Evolution*, *6*(17), 6107–6120. doi: 10.1002/ece3.2346

Steiner, C. C., Putnam, A. S., Hoeck, P. E. A., & Ryder, O. A. (2013). Conservation Genomics of Threatened Animal Species. *Annual Review of Animal Biosciences*, *1*(1), 261–281. doi: 10.1146/annurev-animal-031412-103636

Stenglein, J. L., Waits, L. P., Ausband, D. E., Zager, P., & Mack, C. M. (2010). Efficient, Noninvasive Genetic Sampling for Monitoring Reintroduced Wolves. *Journal of Wildlife Management*, *74*(5), 1050–1058. doi: 10.2193/2009-305

Taberlet, P., Luikart, G., & Waits, L. P. (1999). Noninvasive genetic sampling: Look before you leap. *Trends in Ecology and Evolution*, *14*(8), 323–327. doi: 10.1016/S0169-5347(99)01637-7

Thalmann, O., Hebler, J., Poinar, H. N., Pääbo, S., & Vigilant, L. (2004). Unreliable mtDNA data due to nuclear insertions: A cautionary tale from analysis of humans and other great apes. *Molecular Ecology*, *13*(2), 321–335. doi: 10.1046/j.1365-294X.2003.02070.x

van der Valk, T., Lona Durazo, F., Dalén, L., & Guschanski, K. (2017). Whole mitochondrial genome capture from faecal samples and museum-preserved specimens. *Molecular Ecology Resources*, *17*(6), e111–e121. doi: 10.1111/1755-0998.12699

Vigilant, L., & Guschanski, K. (2009). Using genetics to understand the dynamics of wild primate populations. *Primates*, *50*(2), 105–120. doi: 10.1007/s10329-008-0124-z

Wedrowicz, F., Karsa, M., Mosse, J., & Hogan, F. E. (2013). Reliable genotyping of the koala (Phascolarctos cinereus) using DNA isolated from a single faecal pellet. *Molecular Ecology Resources*, *13*(4), 634–641. doi: 10.1111/1755-0998.12101

White, L. C., Fontsere, C., Lizano, E., Hughes, D. A., Angedakin, S., Arandjelovic, M., … Vigilant, L. (2019). A roadmap for high-throughput sequencing studies of wild animal populations using noninvasive samples and hybridization capture. *Molecular Ecology Resources*, *19*(3), 609–622. doi: 10.1111/1755-0998.12993

Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., …

Scally, A. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, *348*(6231), 242–245. doi: 10.1126/science.aaa3952

Data Accessibility

All raw sequencing data have been deposited at ENA and are available under the accession code PRJEB37173 (http://www.ebi.ac.uk/ena/data/view/PRJEB37173).

Author Contributions

CF, TMB, DAH and EL designed the study. MA and HSK direct the Pan African Programme: The Cultured Chimpanzee. MA and HSK obtained funding for the project. MA, PD, AA, SA, EAA, MB, GB, TD, MEN, ACG, JH, PK, AKK, MK, KL, JL, GM, LJO, AP, MMR, FS, VV and RMW supervised, conducted field work and collected samples. CF, MAE, EL, JL, MA performed experiments. CF and DAH performed the analysis. MAE, MK, DAH, TMB, EL provided analytical support. CF wrote the manuscript with input from all authors.

Supporting Information

Additional supporting information with extended methods and supplementary figures and tables can be found at the end of the manuscript.

Conflict of Interest

Authors declare no conflict of interest.

Supporting Information

Extended methods

Library Preparation

A single library was prepared for each sample following the BEST protocol published by Caroe *et al*. with minor modifications. A total of 200 ng of DNA in 35 µl of lowTE was sheared using a Covaris S2 ultrasonicator with the following settings to obtain 200 bp fragments: duty cycle: 10%, intensity: 5, cycles per burst: 200, time: 120 s.

Next, DNA was end-repaired using 0.5 µl T4 polymerase (5U/µl, Thermo Scientific) 1.5 µl T4 PNK (10 U/µl, Thermo Scientific), 0.4 µl dNTPs (25mM, GE Healthcare), 10 µl T4 DNA ligase buffer (5x, Invitrogen) and 2.5 µl Reaction Enhancer (20% PEG-4000 (Thermo Scientific), 2 mg/µL BSA (New England BioLabs), 400 mM NaCl (Sigma-Aldrich). The mix was incubated 30 min at 20ºC and 30 min at 65ºC (lid at 80ºC).

For adapter ligation reaction we used 2.5 µl T4 DNA ligase buffer (5x, Invitrogen), 1.25 µl T4 DNA ligase (5 U/µl, Invitrogen) and 6.25 µl ddH_2O. At each well we added unique inline barcoded short adapters (1.25 µl each at 100uM; F_P5_7nt_XX Indexed Adapter 5'-CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNN-3'; F_P7_7nt_XX Indexed Adapter 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNN-3'; R_P5/P7_7nt_XX Indexed Common Adapter 5'-NNNNNNNAGATCGGAA-3') with the same 7 nucleotide barcode for the P5 and P7 adapters (Figure S18). Previous studies have shown a better capture efficiency when the library size is small (Rohland & Reich, 2012). Moreover, an early barcoding of the library (in the adapter ligation step rather than in the final amplification PCR) lowers the probability of indiscernible contamination from close wells. Ligation reaction was incubated 45 min at 20ºC and 10 min at 64ºC (lid at 80ºC). Fill-in reaction was done using 2 µl of Bst 2.0 WarmStart Polymerase (8 U/µl, New England BioLabs), 2.5 µl of Isothermal amp. buffer (10x, New England BioLabs)), 0.5 µl of dNTPs (25 mM, GE Healthcare) and 7.5 µl ddH_2O. Reaction was incubated for 20 min at 65ºC (lid 80ºC) and 20 min at 80ºC (lid 110ºC).

The product was purified using homemade SPRI beads (Rohland & Reich, 2012) and eluting in a final volume of 25 µl of lowTE. Finally, each library was amplified using 25 µl of Kapa HIFI HS RM (2x, Roche), and 2.5 µl of each PreHyb primers (P5: 5'-CTTTCCCTACACGACGCTCTTC-3' and P7: 5'-GTGACTGGAGTTCAGACGTGTG-3', 10 µM) and incubated 2 min at 95ºC (lid at 110ºC), followed by 8 to 12 cycles of 15 s at 98ºC, 30 s at 55ºC and 30 s at 72ºC, with a final elongation of 1 min at 72ºC.

The final library was purified using homemade SPRI beads (Rohland & Reich, 2012) and eluting in a final volume of 30 µl of ddH_2O. Libraries were quantified with an Agilent 2100 Bioanalyzer using a DNA 7500 assay kit.

Hybridization Capture

Each hybridization reaction was performed with 1 or 2 µg of pooled library (7 µl) a blocking mix containing 2.5 µg of Human cot-1 (1 µg/µl, Invitrogen), 2.5 µg of salmon sperm (10 µg/µl, Invitrogen), 2 µM of P5 and P7 blocking oligos (Rohland & Reich, 2012), heated 5 min at 95ºC (lid 105ºC) and held at 65ºC for at least 5 minutes.

Then, the prewarmed 22 µl of hybridization buffer (10x SSPE (20x, Invitrogen), 10x Denhardt's Solution (50x, Invitrogen), 10mM EDTA (0.5M, Sigma-Aldrich), 0.2% SDS (20%, Invitrogen)) was added to the previously warmed to 65 ºC for 2 min bait mix: 3 µl of SureSelect Human All Exon V6 RNA library baits (Agilent Technologies), 1 µl of SUPERase-In and 1 µl of ddH$_2$O. The capture mix was added to the pools and incubated overnight at 65ºC. After the incubation we performed several washes with homemade wash buffers (Wash Buffer #1: 1x SSC (20x, Invitrogen) and 0.1% SDS (20%, Invitrogen); Wash Buffer #2: 0.1% SSC (20x, Invitrogen) and 0.1% SDS (20x, Invitrogen)) and Streptavidin-coated beads (Dynabeads MyOne Streptavidin T1 beads, Invitrogen). Beads were washed following the manufacturer's protocol and resuspended in 200 µl of binding buffer (1M NaCl (5M, Sigma-Aldrich), 10mM Tris-HCl pH 7.5 (1M, Invitrogen), 1mM EDTA (0.5M, Sigma-Aldrich)). The captured library was transferred to the beads and incubated at room temperature on a thermomixer at 700 RPM for 30 min. Using a magnetic rack, we removed the supernatant and washed the beads with Wash Buffer #1 for 15 min at room temperature on the thermomixer at 700 RPM. Then, the beads were placed in the magnetic rack again and washed with Wash Buffer #3 three times for 10 min at 68°C and 700 RPM. Finally, the beads were resuspended in 20 µl of H$_2$O followed by an enrichment PCR with PreHyb primers (P5-F: 5'-CTTTCCCTACACGACGCTCTTC-3' and P7-R: 3'-GTGTGCAGACTTGAGGTCAGTG-5'), with the same incubation protocol as in library preparation amplification but with 10-12 cycles. After cleaning the PCR product with homemade SPRI beads (Rohland & Reich, 2012) a second capture experiment was performed as recommended by Hernandez-Rodriguez et al. PCR amplification (9-12 cycles) of the final captured pool was done using the same protocol as before but with indexed primers (P5-F: 5'-AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTT-3' and P7-R: 3'-TGTGCAGACTTGAGGTCAGTGNNNNNNNTAGAGCATACGGCAGAAGACGAAC-5') (Kircher, Sawyer, & Meyer, 2012) to double-index each pool of libraries with a unique pair of indices (Figure S18).

As previously described, the use of inline barcodes and P5 and P7 indexing primers allows the multiplexing of numerous libraries in a single pool. Thus, for the experiments presented here, the usage of such adapters was of high utility, since after the libraries were build, we pooled them together for capture, and subsequently pools were indexed using P5 and P7 (Rohland & Reich, 2012).

Since the captured pools were indexed, it was possible to sequence many libraries in one sequencing lane. Also, these short adapters do not interfere with hybridization experiments as complete adapters did. As suggested in Rohland et al., we increased by one nucleotide the barcode sequence in the adapters, from 6nt to 7nt, thus increasing the multiplexing power.

Supplementary Table Legends *(tables are not provided here, access upon request)*

Supplementary T1. Sample description of screened samples.

Sample description for all screened samples in this study; provided in the additional excel file.

Supplementary T2. Sample description for capture samples.

Sample description for the selected samples for capture; provided in the additional excel file.

Supplementary T3. Endogenous content by site.

Average endogenous content of samples according to site; provided in the additional excel file.

Supplementary T4. Sequencing summary statistics.

Summary of sequencing stats for each sample in each hybridization; provided in the additional excel file.

Supplementary T5. Sequencing summary statistics for independent hybridizations.

Summary of sequencing stats for independent hybridizations, each row contains the sum of all samples belonging to each hybridization; provided in the additional excel file.

Supplementary T6. Correlation matrix among all study variables.

Correlation matrix of all variables analyzed in this study. Spearman's rho was estimated when comparing two numeric variables. Cramér's V was estimated among two categorical variables. When comparing a numeric and categorical variable we took the square root of the R-squared statistic derived from a univariate linear model with no transformation on the dependent, numerical values; provided in the additional excel file.

Supplementary T7. Sequencing summary statistics for down-sampled data.

Summary of sequencing stats for each down-sampled library at 1,500,000 in each hybridization; provided in the additional excel file.

Supplementary T8. Correlation matrix among all study variables for down-sampled data.

Correlation matrix of all variables analyzed in this study after each library has been down-sampled to 1,500,000 reads. Spearman's rho was estimated when comparing two numeric variables. Cramér's V was estimated among two categorical variables. When comparing a numeric and categorical variable we took the square root of the R-squared statistic derived from a univariate linear model with no transformation on the dependent, numerical values; provided in the additional excel file.

Supplementary Figures

Figure S1: Endogenous content and fragment size across sampling sites.

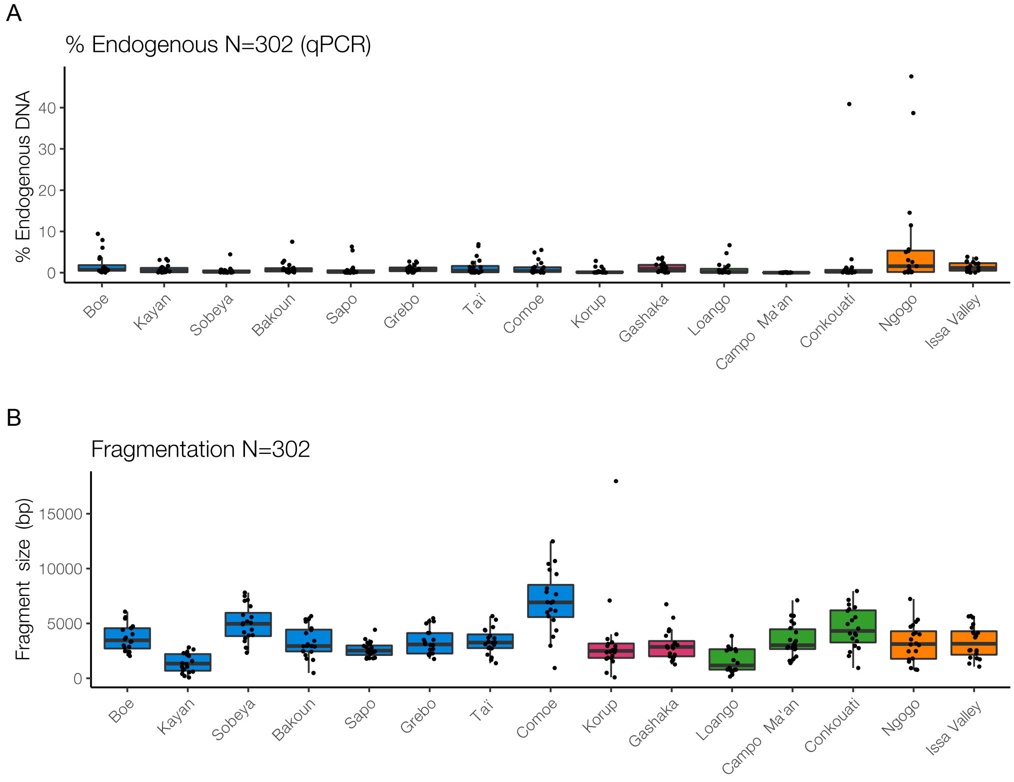


Figure S1 Legend: Distribution of (A) % endogenous content and (B) fragment size for the 302 screened samples from the 15 screened African sites in the PanAfrican programme. The boxplot colors indicate the subspecies membership as seen in Figure 1: blue (western chimpanzee), pink (Nigeria-Cameroon chimpanzee), green (central chimpanzee) and orange (eastern chimpanzee).

Figure S2. Schematic of library complexity analysis

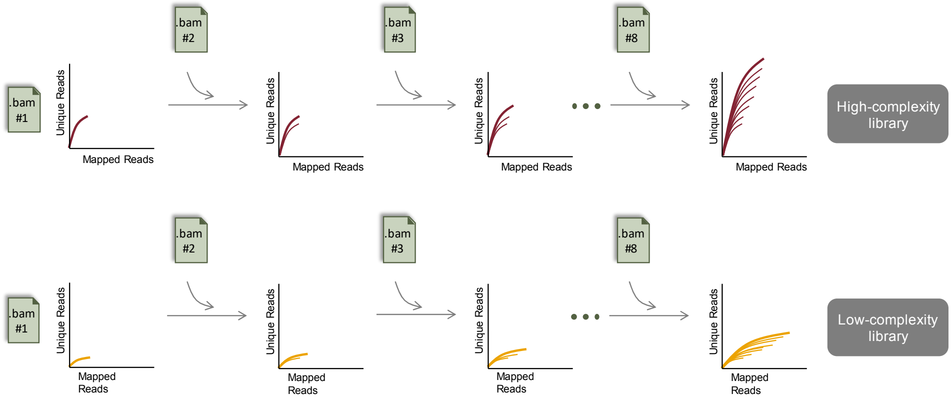
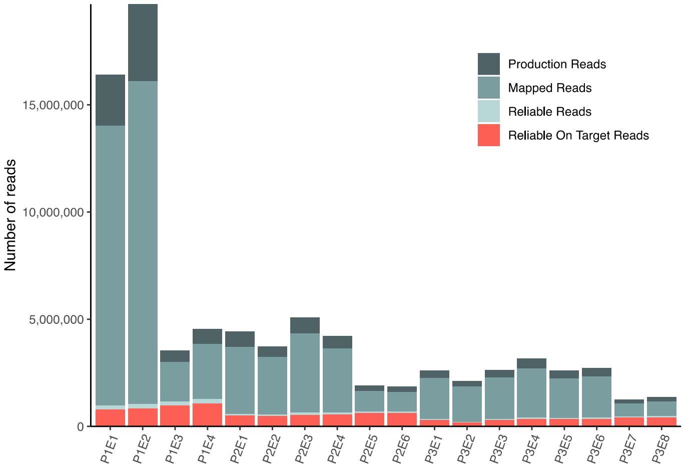


Figure S2 Legend. Schematic representation of library complexity analysis. We add data sequentially, coming from replicate hybridizations through merging BAM files. For each step we subsample without replacement each merged bam file. If the library has high molecular complexity (in red) we see a feathered distribution, where the more data we add, the more unique reads are retrieved. On the other hand, if the library has low molecular complexity, performing additional replicate hybridization does not improve the recovery of new unique reads.

Figure S3. Capture performance

A



B

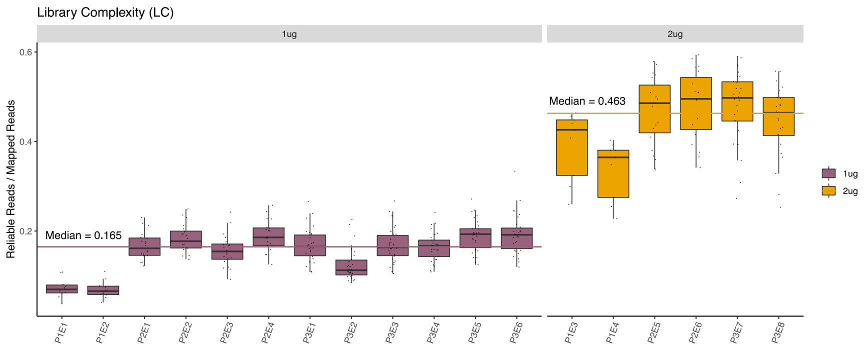
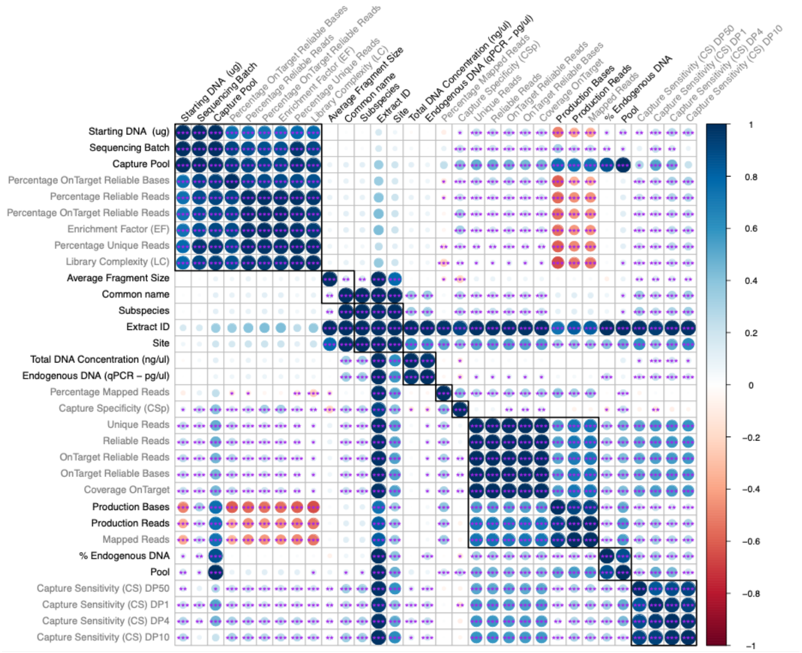


Figure S3 Legend. Capture performance analysis for each 18 capture experiments in 3,75 HiSeq 4000 lanes. (A) Sequencing stats and (B) Library complexity separated by experiments using 1 µg and 2 µg of pooled library, solid lines represent the median LC.

Figure S4. Correlation matrixes of all variables

A



B

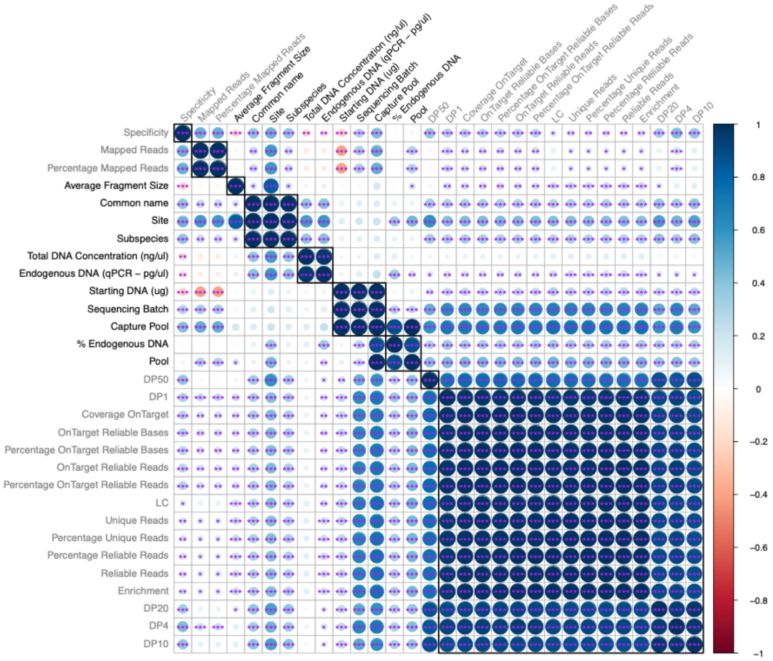
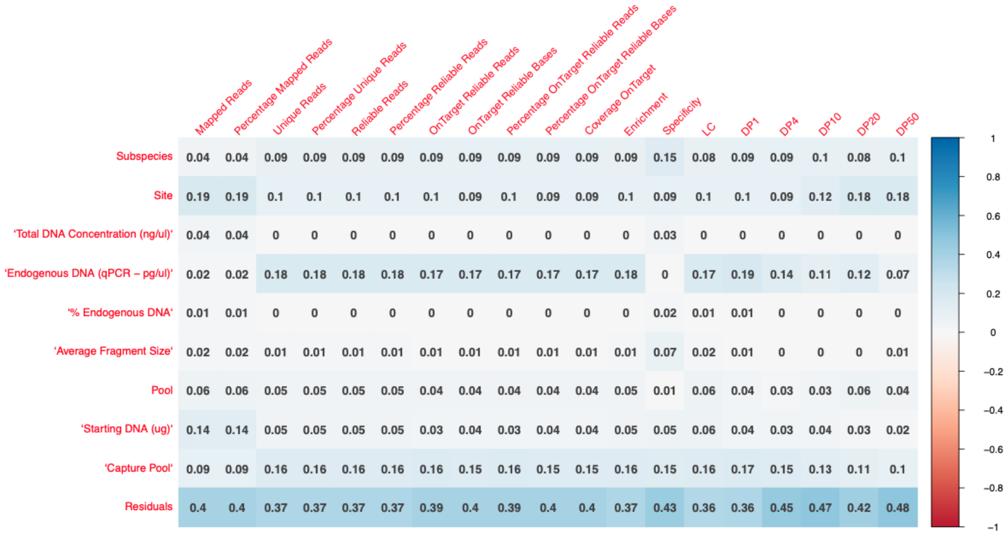


FIGURE S4 Legend. Correlation matrix of all variables included in this study in the (a) full dataset and (b) after having down-sampled each library to 1,500,000 reads. Spearman's rho was estimated when comparing two numeric variables. Cramér's V was estimated among two categorical variables. When comparing a numeric and categorical variable we took the square root of the R-squared statistic derived from a univariate linear model with no transformation on the dependent, numerical values. Experimental variables are illustrated in black text. Performance variables are illustrated in grey text. Clusters of strongly correlated variables where identified, and illustrated by the black squares, using the function cutree() on a hierarchical clustering dendrogram of the same data transformed to distances (1-abs(data)). A cut height of 0.5 was used to identify clusters where intra-cluster distances among variables are greater than or equal to 0.5, and inter-cluster correlations are smaller than 0.5.

Figure S5. Multivariate type I ANOVA

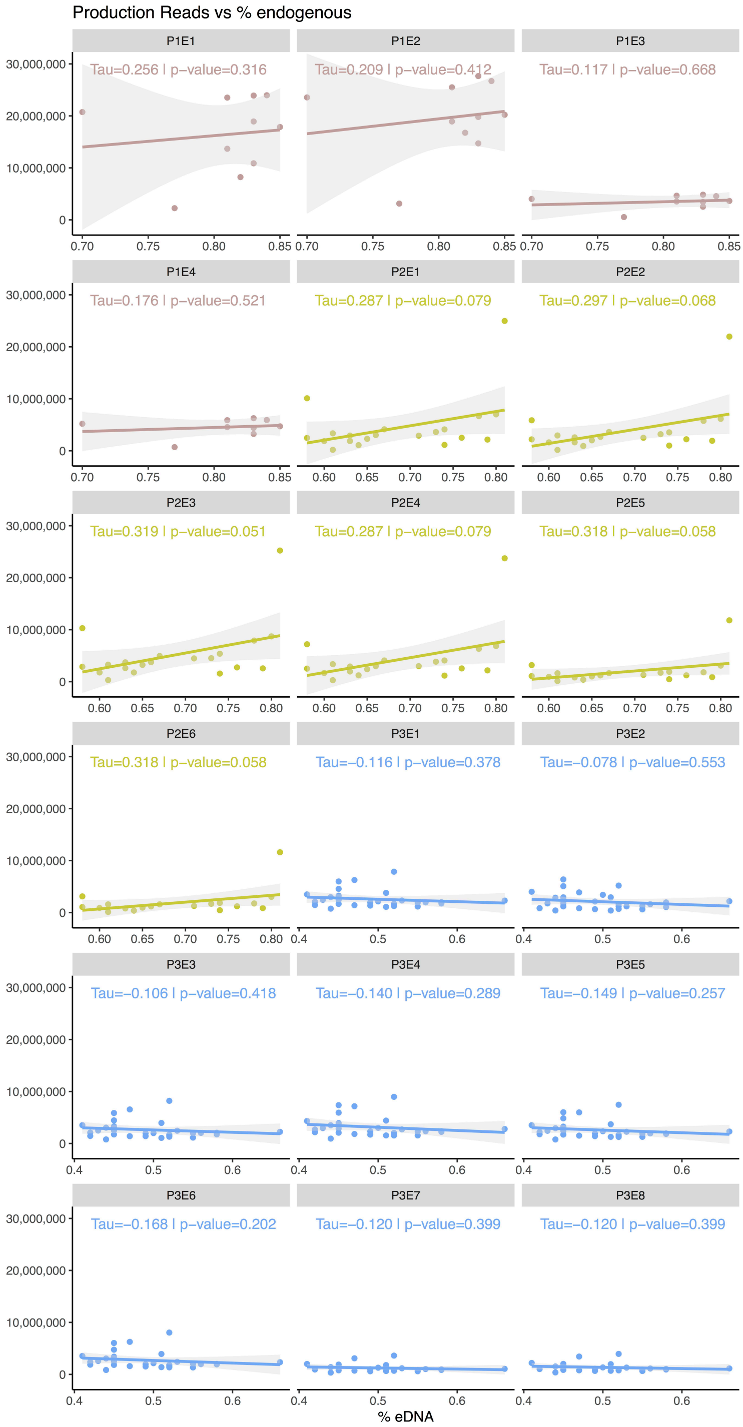


	Mapped Reads	Percentage Mapped Reads	Unique Reads	Percentage Unique Reads	Reliable Reads	Percentage Reliable Reads	OnTarget Reliable Reads	OnTarget Reliable Bases	Percentage OnTarget Reliable Reads	Percentage OnTarget Reliable Bases	Coverage OnTarget	Enrichment	Specificity	LC	DP1	DP4	DP10	DP20	DP50
Subspecies	0.04	0.04	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.15	0.08	0.09	0.09	0.1	0.08	0.1
Site	0.19	0.19	0.1	0.1	0.1	0.1	0.1	0.09	0.1	0.09	0.09	0.1	0.09	0.1	0.1	0.09	0.12	0.18	0.18
‘Total DNA Concentration (ng/ul)‘	0.04	0.04	0	0	0	0	0	0	0	0	0	0	0.03	0	0	0	0	0	0
‘Endogenous DNA (qPCR – pg/ul)‘	0.02	0.02	0.18	0.18	0.18	0.18	0.17	0.17	0.17	0.17	0.17	0.18	0	0.17	0.19	0.14	0.11	0.12	0.07
‘% Endogenous DNA‘	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0.02	0.01	0.01	0	0	0	0
‘Average Fragment Size‘	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.07	0.02	0.01	0	0	0	0.01
Pool	0.06	0.06	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.05	0.01	0.06	0.04	0.03	0.03	0.06	0.04
‘Starting DNA (ug)‘	0.14	0.14	0.05	0.05	0.05	0.05	0.03	0.04	0.03	0.04	0.04	0.05	0.05	0.06	0.04	0.03	0.04	0.03	0.02
‘Capture Pool‘	0.09	0.09	0.16	0.16	0.16	0.16	0.16	0.15	0.16	0.15	0.15	0.16	0.15	0.16	0.17	0.15	0.13	0.11	0.1
Residuals	0.4	0.4	0.37	0.37	0.37	0.37	0.39	0.4	0.39	0.4	0.4	0.37	0.43	0.36	0.36	0.45	0.47	0.42	0.48

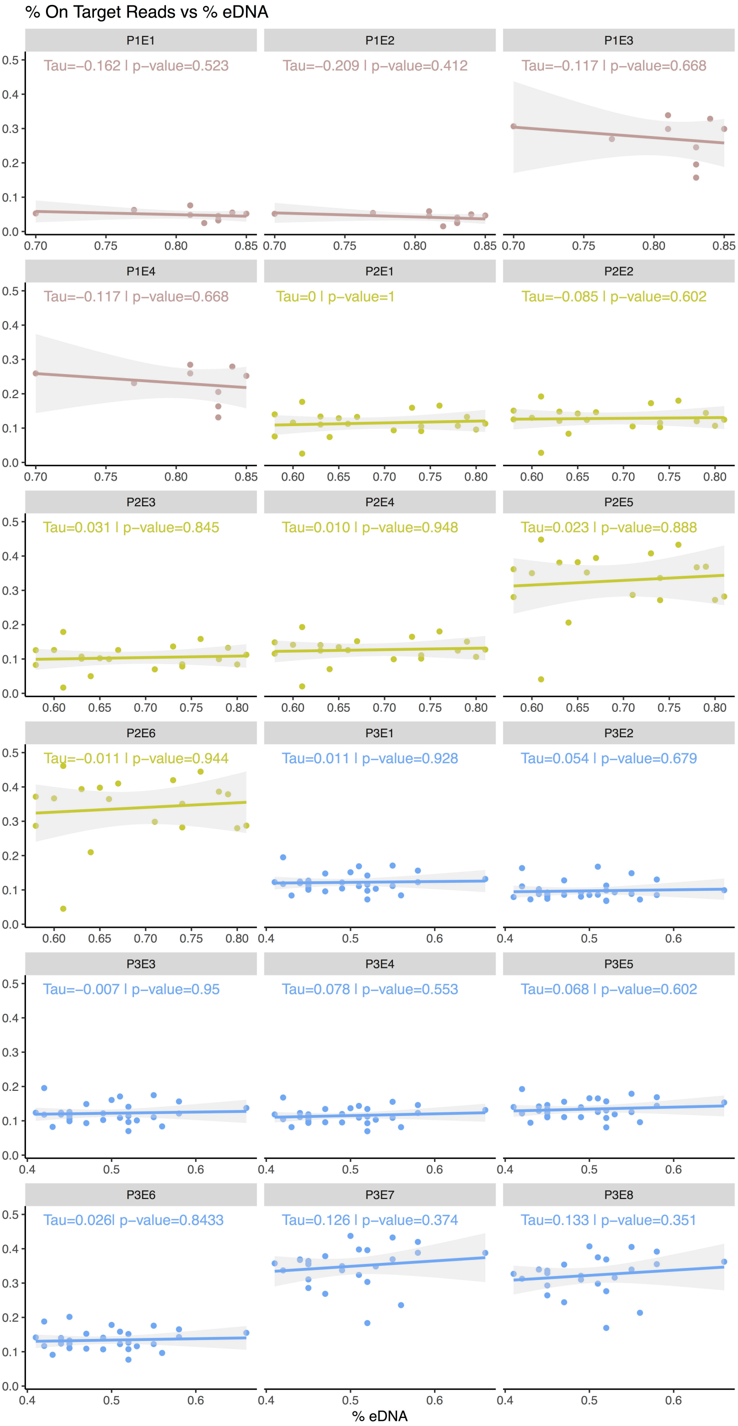
Figure S5 Legend. Multivariate type I analysis of variance. Estimated variance explained from multivariate type I ANOVA of the experimental variables affecting performance summary statistics. Figure is an extension of Figure 4. Estimates are derived from 1,500,000 read down-sampled libraries.

Figure S6. Correlation dot plots

(A)



(B)


% On Target Reads vs % eDNA
P1E1
Tau=−0.162 | p−value=0.523
P1E2
Tau=−0.209 | p−value=0.412
P1E3
Tau=−0.117 | p−value=0.668
P1E4
Tau=−0.117 | p−value=0.668
P2E1
Tau=0 | p−value=1
P2E2
Tau=−0.085 | p−value=0.602
P2E3
Tau=0.031 | p−value=0.845
P2E4
Tau=0.010 | p−value=0.948
P2E5
Tau=0.023 | p−value=0.888
P2E6
Tau=−0.011 | p−value=0.944
P3E1
Tau=0.011 | p−value=0.928
P3E2
Tau=0.054 | p−value=0.679
P3E3
Tau=−0.007 | p−value=0.95
P3E4
Tau=0.078 | p−value=0.553
P3E5
Tau=0.068 | p−value=0.602
P3E6
Tau=0.026l p−value=0.8433
P3E7
Tau=0.126 | p−value=0.374
P3E8
Tau=0.133 | p−value=0.351
0.0
0.1
0.2
0.3
0.4
0.5
0.70
0.75
0.80
0.85
0.60
0.65
0.70
0.75
0.80
0.4
0.5
0.6
% eDNA

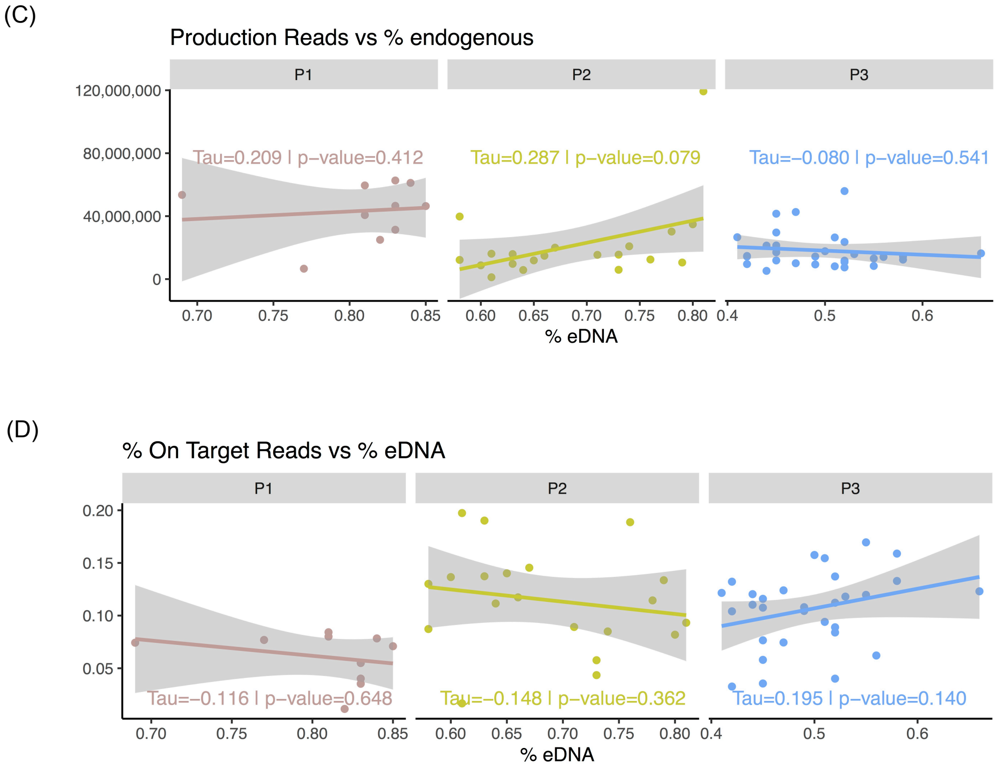


Figure S6 Legend. Kendall's correlation between (A) Production Reads and (B) % On Target Reads versus % eDNA in each Hybridization experiment. No statistically significant correlation of eDNA content with both summary statistics although some hybridizations in P2 exhibit a slight positive correlation, possibly due to one outlier. In (C) Production Reads and (D) % On Target Reads we show the same correlation plots with % eDNA but now with data coming from merged hybridizations.

Figure S7. Distribution of raw reads across pools

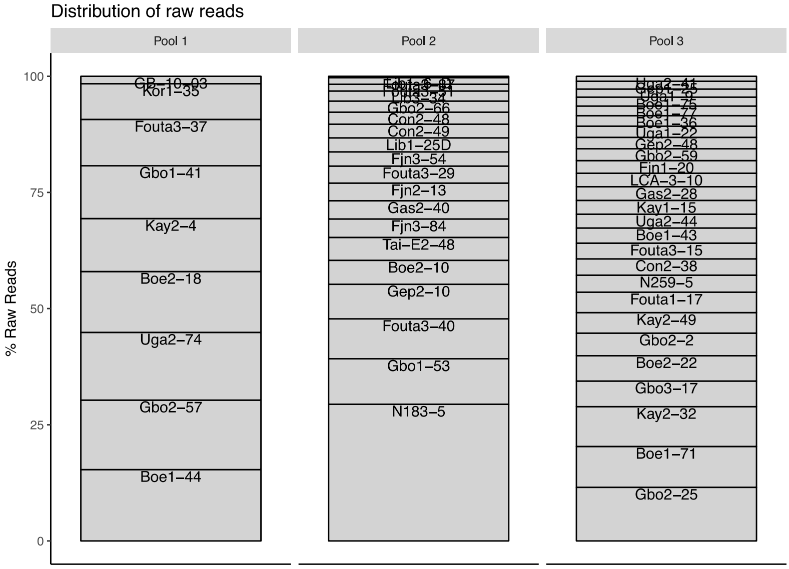


Figure S7 Legend. Percentage of raw reads (production reads) sequenced for each library in each pool to detect which samples are taking a greater proportion of the total production reads.

Figure S8. Impact of total DNA in pooled libraries on average unique read count.

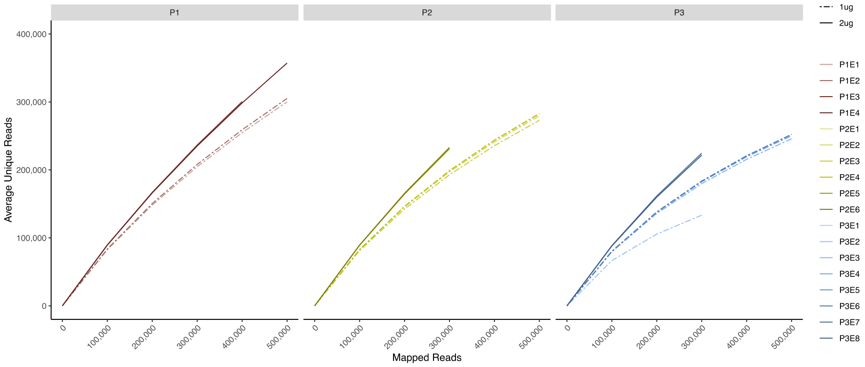


Figure S8 Legend. Comparison of pooling 1 μg or 2 μg DNA for capture. We subsampled without replacement reads in each hybridization (average of all samples within a pool) and obtained the corresponding average unique reads. The averages are done if all samples in the pool have data in any given point (for that reason sample Lib1-6D from P2 is excluded). Dashed lines indicate 1 μg of starting DNA for capture while solid lines are the hybridizations with 2 μg of starting DNA. Colors indicate each hybridization.

Figure S9. Library complexity by replicate hybridizations.

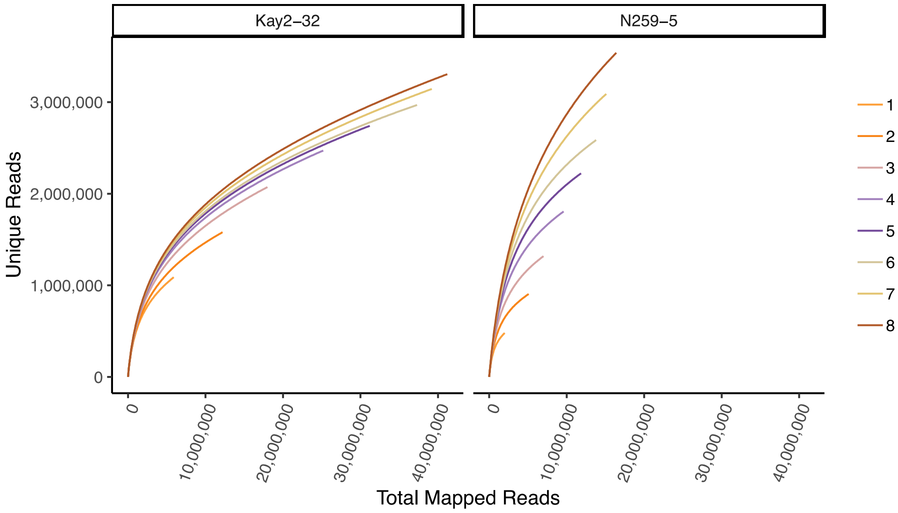
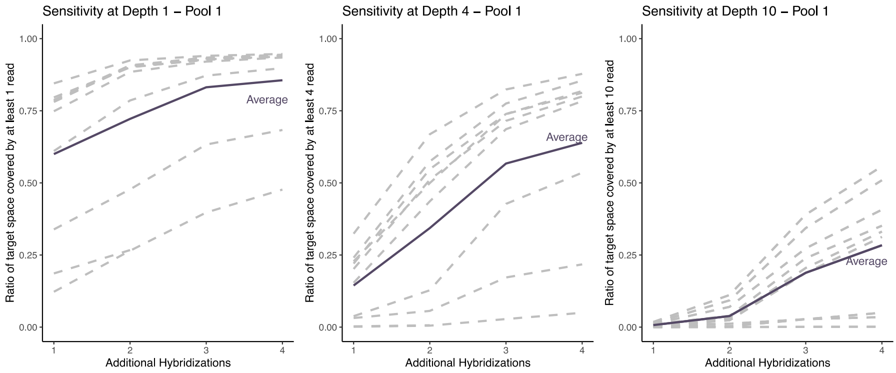


Figure S9 Legend. Library complexity plots of two samples belonging to P3. Each line represents data coming from cumulative replicate hybridizations. *Line 1* indicates data coming for only one hybridization, *line 2* indicates combined data from 2 hybridization, until *line 8* that indicates combined data from all 8 hybridization replicates. Library Kay2-32 has low library complexity and cannot be increased by additional hybridizations. However, the majority of samples behave similar to the example sample N259-5. By performing additional hybridizations, it is possible to retrieve new unique reads.

Figure S10. Capture sensitivity by depth and pool.

A



B

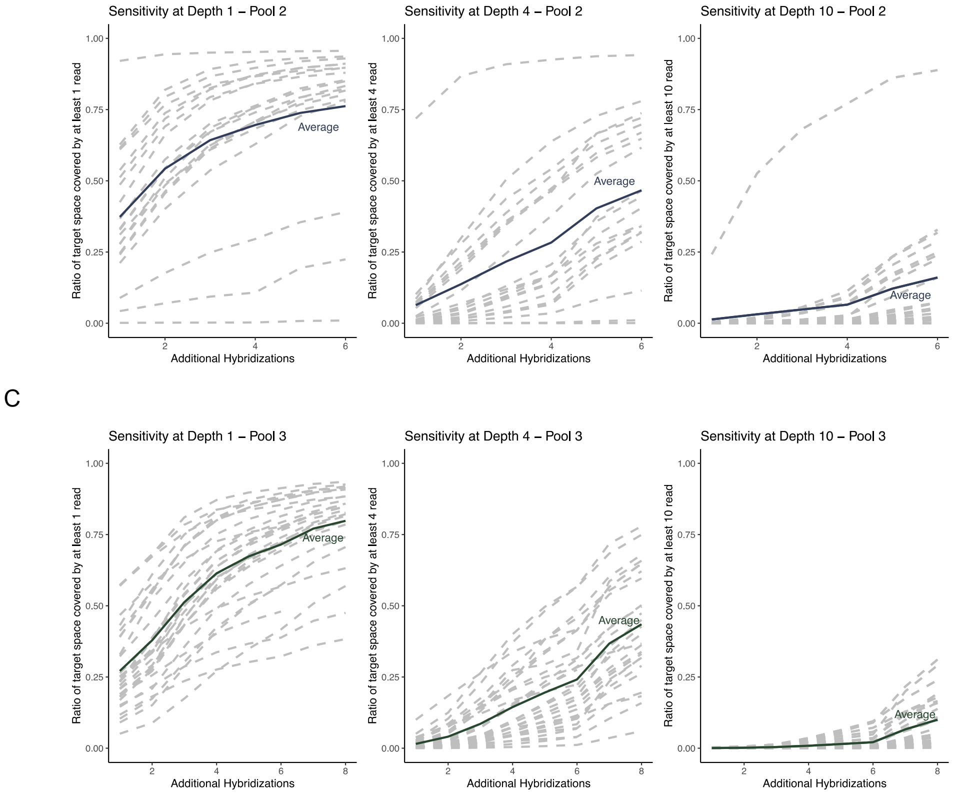


Figure S10 Legend. Sensitivity (ratio of target space covered by at least a certain number of reads) at depth 1, 4 and 10 for samples in (A) P1, (B) P2 and (C) P3. Each grey dashed line represents a sample from each pool and the colored solid line is the average of all samples within the pool.

Figure S11. Venn diagram never covered regions.

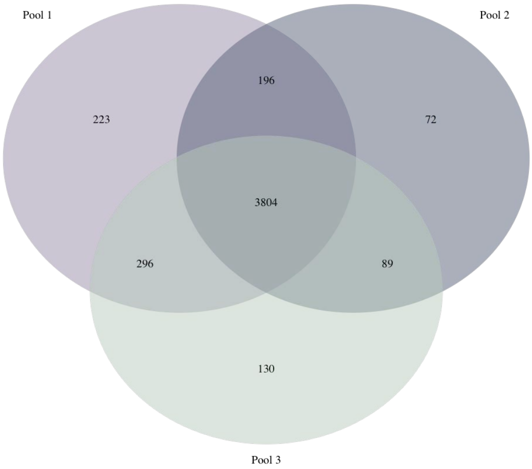


Figure S11 Legend. Intersection of regions never covered after 4, 6 and 8 additional hybridizations for Pool1, Pool2 and Pool3, respectively. In Pool1, out of the total 243,190 regions, 4,519 are never covered (1.85%); in Pool 2, it is 4161 out of 243,190 total regions (1.71%); and for Pool 3 it is 4319 out of 243,190 total regions (1.77%). From those, the same 3804 regions are never covered in all experiments (1,564%).

Figure S12. Sequencing effort data saturation.

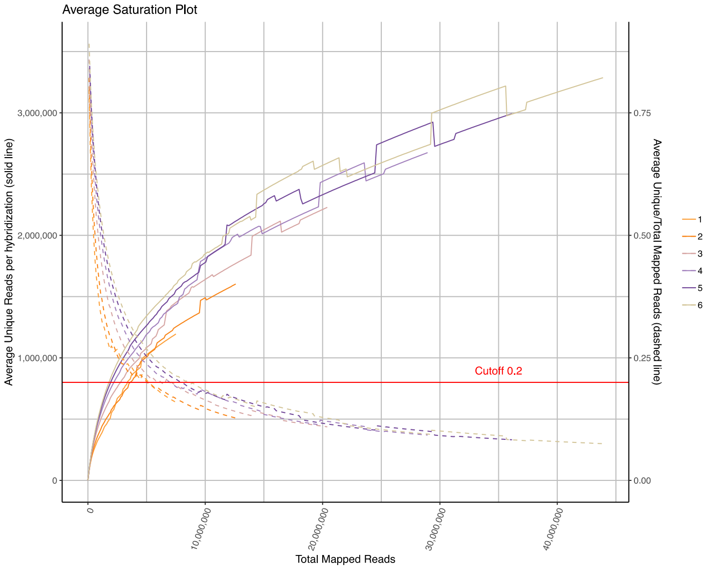


Figure S12 Legend. Sequencing Effort. Solid lines represent the sample average number of unique reads after merging data from additional hybridizations (numeric key). Dashed lines represent the average number of unique reads normalized by the number of mapped reads. The cutoff is set at 20% (right Y axis). We estimated for each additional hybridization a sample average and plotted the number of unique reads averaged across samples (left Y axis) and also the proportion of unique reads by total mapped reads averaged across samples (right Y axis), with the total mapped reads (X axis).

Figure S13. Sequencing summary statistics by SeqBatch.

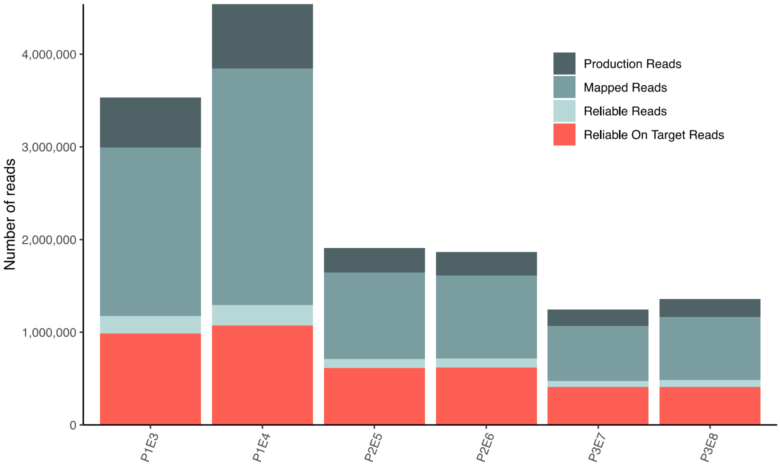
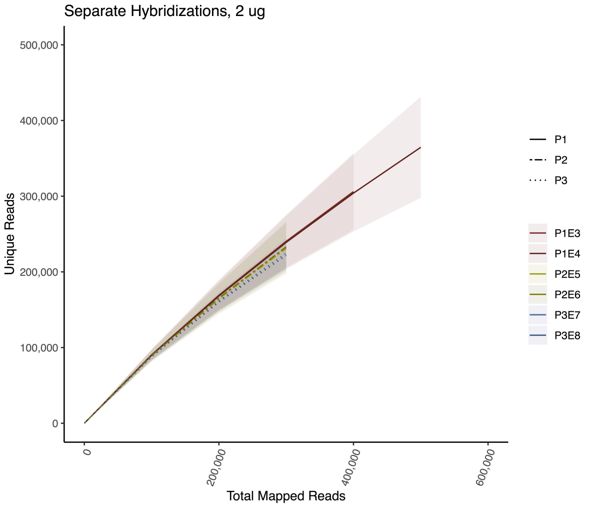


Figure S13 Legend. Sequencing stats for the SeqBatch 3 (P1E3, P1E4, P2E5, P2E6, P3E7, P3E8). Y axis represents the average number of reads per library belonging to each pool. On average we obtain 3.5 million reads per library in hybridizations from P1, around 2 million reads per library in hybridizations from P2 and around 1.5 million reads per library for hybridizations from P3. The percentage of reliable reads is 27.87% in P1E3 and 23.58% in P1E4; 32.12% in P2E5 and 33.06% in P2E6; 32.71% in P3E7 and 30.17% in P3E8.

Figure S14. Average library complexity curves

A



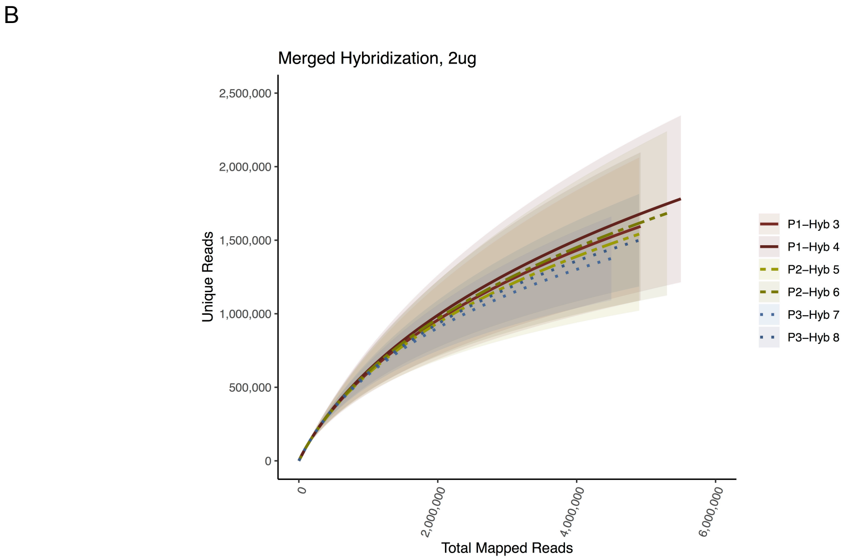
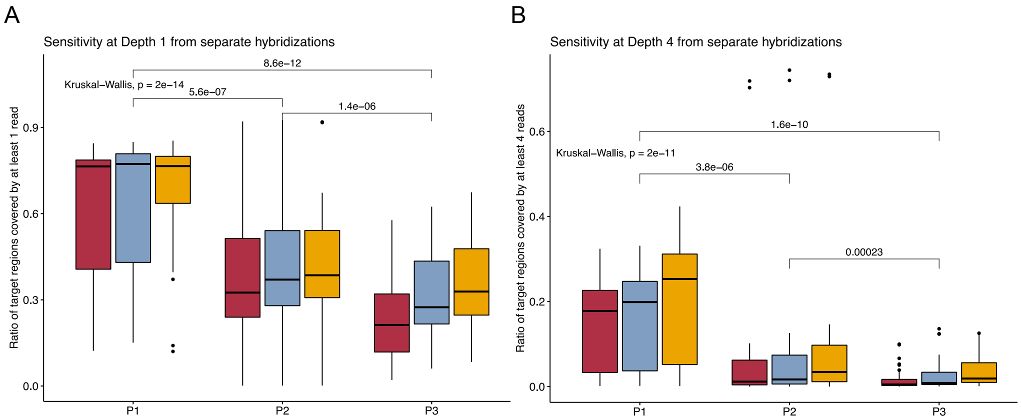


Figure S14 Legend. A) Average library complexity curve for each individual hybridization (starting with 2µg). B) Average library complexity curve for merged hybridizations (only hybridizations with starting DNA of 2 µg). Solid line is P1, two-dashed line is P2 and dotted line is P3. Sample Lib1-6D in P2 was removed from the analysis due to low coverage.

Figure S15. Sensitivity by pool at various depth.



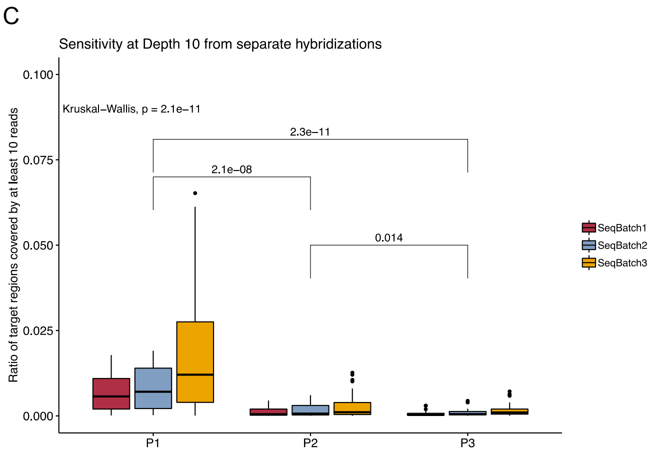


Figure S15 Legend. Capture performance analysis of sensitivity from separate hybridizations and plotting together the data coming from the same Sequencing Batch (color). Small pools have higher sensitivity than larger pools. (A) Capture sensitivity at depth 1, (B) capture sensitivity at depth 4 and (C) capture sensitivity at depth 10.

Figure S16. Variance explained by pool on capture sensitivity.

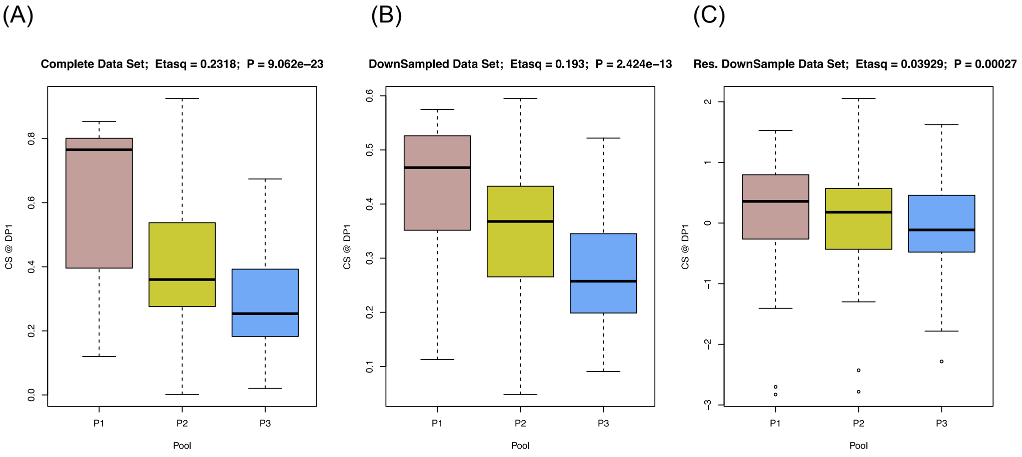


Figure S16 Legend. Multivariate Type I ANOVA of the variance explained of ‘Pool’ on capture sensitivity (CS) at Depth 1. (A) Whole data set. (B) Libraries down-sampled at 1,500,000 reads. (C) Residuals.

Figure S17. Variation in capture sensitivity across pools.

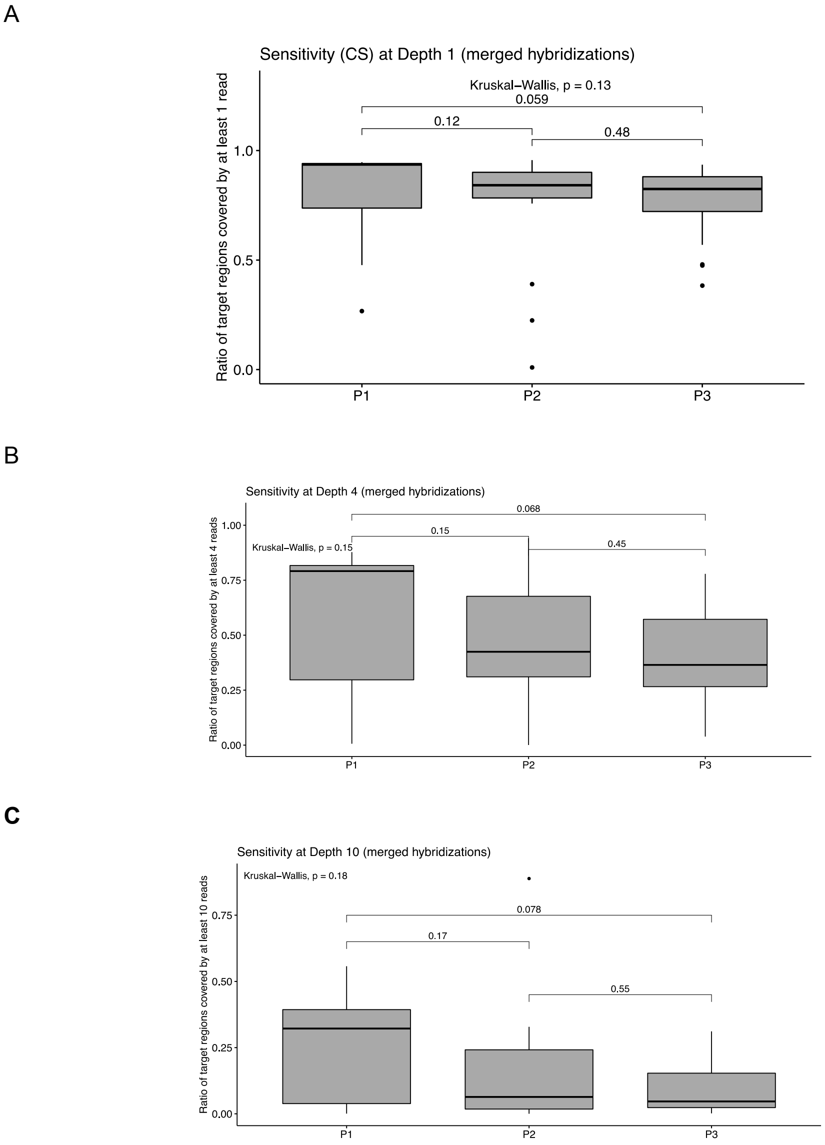


Figure S17 Legend. Capture performance analysis of sensitivity after merging data from additional hybridizations. (A) Capture sensitivity at depth 1, (B) Capture sensitivity at depth 4 and (B) capture sensitivity at depth 10.

Figure S18. Illustration of library construction

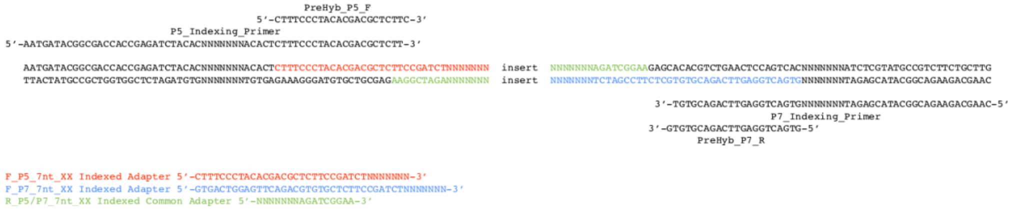

PreHyb_P5_F
5'-CTTTCCCTACACGACGCTCTTC-3'
P5_Indexing_Primer
5'-AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTT-3'
AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNN insert NNNNNNNAGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTAGATGTGNNNNNNNTGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGANNNNNNN insert NNNNNNNTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNNTAGAGCATACGGCAGAAGACGAAC
3'-TGTGCAGACTTGAGGTCAGTGNNNNNNNTAGAGCATACGGCAGAAGACGAAC-5'
P7_Indexing_Primer
3'-GTGTGCAGACTTGAGGTCAGTG-5'
PreHyb_P7_R
F_P5_7nt_XX Indexed Adapter 5'-CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNN-3'
F_P7_7nt_XX Indexed Adapter 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNN-3'
R_P5/P7_7nt_XX Indexed Common Adapter 5'-NNNNNNNAGATCGGAA-3'

Figure S18. Final library structure showing the sequences of the indexed adapters and primers used as well as the primers used for amplification of the partial library before and after the first round of hybridization.

References

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. H. S., Samaniego, J. A., … Gilbert, M. T. P. (2018). Single-tube library preparation for degraded DNA. Methods in Ecology and Evolution, 9(2), 410–419. doi: 10.1111/2041-210X.12871

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, *40*(1), 1–8. doi: 10.1093/nar/gkr771

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, *22*(5), 939–946. doi: 10.1101/gr.128124.111

3.3. A fine-scale genomic survey of wild chimpanzee populations reveals past connectivity and provides a tool for geolocalization

Claudia Fontsere, Martin Kuhlwilm, Marina Alvarez-Estape, Jack Lester, Paula Dieguez, Thierry Aebischer, Anthony Agbor, Paula Alvarez Varona, Samuel Angedakin, Alfred K. Assumang, Floris Aubert, Emmanuel A. Ayimisin, Emma Bailey, Donatienne Barubiyo, Mattia Bessone, Matthieu Bonnet, Gregory Brazzola, Andrea Carretero-Alonso, Rebecca Chancellor, Heather Cohen, Katherine Corogenes, Charlotte Coupland, Bryan Curran, Emmanuel Danquah, Tobias Deschner, Emmanuel Dilambaka, Dervla Dowd, Andrew Dunn, Jef Dupain, Villard E. Egbe, Henk Eshuis, Olga Feliu, Annemarie Goedmakers, Anne-Celine Granjon, Josephine Head, Daniela Hedwig, Veerle Hermans, Inaoyom Imong, Kathryn J. Jeffery, Sorrel Jones, Parag Kadam, Mike Kaiser, Mbangi Kambere, Mohamed Kambi, Magloire V. Kambale, Ivonne Kienast, Deo Kujirakwinja, Kevin Langergraber, Vincent Lapeyre, Juan Lapuente, Bradley Larson, Anne Laudisoit, Kevin Lee, Vera Leinert, Miquel Llorente, Giovanna Maretti[2], Sergio Marrocoli, Rumen Martin, Amelia Meier, Dave Morgan, Felix Mulindahabi, Mizuki Murai, Emily Neil, Stuart Nixon, Protais Niyigabae, Emma Normand, Chris Orbell, Lucy J.Ormsby, Liliana Pacheco, Alex Piel, Jodie Preece, Sebastien Regnaut, Laura Riera, Martha Robbins, Aaron Rundus, Crickette Sanz, Lilah Sciaky, Volker Sommer, Fiona A. Stewart, Nikki Tagg, Emilien Terrade, Alexander Tickle, Els Ton, Joost van Schijndel, Hilde Vanleeuwe, Virginie Vergnes, Jacob Willie, Roman M. Wittig, Yisa G. Yuh, Kyle Yurkiw, Klaus Zuberbuehler, Richard McElreath, Linda Vigilant, Christophe Boesch, Aida M. Andrés, David A. Hughes, Hjalmar Kühl, Esther Lizano, Mimi Arandjelovic & Tomas Marques-Bonet

A fine-scale genomic survey of wild chimpanzee populations reveals past connectivity and provides a tool for geolocalization.

Manuscript under preparation

A fine-scale genomic survey of wild chimpanzee populations reveals past connectivity and provides a tool for geolocalization

Claudia Fontsere[1], Martin Kuhlwilm[1], Marina Alvarez-Estape[1], Jack Lester[2], Paula Dieguez[2], Thierry Aebischer[3], Anthony Agbor[2], Paula Alvarez Varona[4], Samuel Angedakin[2], Alfred K. Assumang[5], Floris Aubert[6], Emmanuel A. Ayimisin[2], Emma Bailey[2], Donatienne Barubiyo[2], Mattia Bessone[2], Matthieu Bonnet[7], Gregory Brazzola[2], Andrea Carretero-Alonso[8], Rebecca Chancellor[9], Heather Cohen[2], Katherine Corogenes[2], Charlotte Coupland[2], Bryan Curran[7], Emmanuel Danquah[5], Tobias Deschner[2], Emmanuel Dilambaka[10], Dervla Dowd[6], Andrew Dunn[10], Jef Dupain[11], Villard E. Egbe[2], Henk Eshuis[2], Olga Feliu[12], Annemarie Goedmakers[13], Anne-Celine Granjon[2], Josephine Head[2], Daniela Hedwig[7], Veerle Hermans[14], Inaoyom Imong[10], Kathryn J. Jeffery[15,16], Sorrel Jones[2], Parag Kadam[17], Mike Kaiser[2], Mbangi Kambere[2], Mohamed Kambi[2], Magloire V. Kambale[2], Ivonne Kienast[2], Deo Kujirakwinja[10], Kevin Langergraber[18,19], Vincent Lapeyre[6], Juan Lapuente[2], Bradley Larson[2], Anne Laudisoit[20], Kevin Lee[2], Vera Leinert[6], Miquel Llorente[12,21], Giovanna Maretti[2], Sergio Marrocoli[2], Rumen Martin[2], Amelia Meier[2], Dave Morgan[22], Felix Mulindahabi[10], Mizuki Murai[2], Emily Neil[2], Stuart Nixon[23], Protais Niyigabae[10], Emma Normand[6], Chris Orbell[24], Lucy J.Ormsby[2], Liliana Pacheco[4], Alex Piel[25], Jodie Preece[2], Sebastien Regnaut[6], Laura Riera[12], Martha Robbins[2], Aaron Rundus[9], Crickette Sanz[26], Lilah Sciaky[2], Volker Sommer[25], Fiona A. Stewart[25,27], Nikki Tagg[14], Emilien Terrade[2], Alexander Tickle[2], Els Ton[13], Joost van Schijndel[2], Hilde Vanleeuwe[10], Virginie Vergnes[6], Jacob Willie[14], Roman M. Wittig[2], Yisa G. Yuh[2], Kyle Yurkiw[2], Klaus Zuberbuehler[28], Richard McElreath[29], Linda Vigilant[2], Christophe Boesch[2,6], Aida M. Andrés[30], David A. Hughes[31,32], Hjalmar Kühl[2,33], Esther Lizano[1], Mimi Arandjelovic[2] & Tomas Marques-Bonet[1,34,35,36]

1 Institut de Biologia Evolutiva (UPF-CSIC), Barcelona, Spain.
2 Max Planck Institute for Evolutionary Anthropology (MPI EVA), Leipzig, Germany.
3 University of Fribourg, Fribourg, Switzerland.
4 Instituto Jane Goodall España, Barcelona, Spain.
5 Department of Wildlife and Range Management, Faculty of Renewable Natural Resources, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana.
6 Wild Chimpanzee Foundation (WCF), Leipzig, Germany.
7 The Aspinall Foundation, Kent, United Kingdom,
8 Centro de Rescate de Primates "RAINFER", Fuente el Saz de Jarama, Madrid, Spain.
9 Depts of Anthropology & Sociology and Psychology, West Chester University, West Chester,
10 Wildlife Conservation Society (WCS), New York, United States.
11 Africa Wildlife Foundation, Washington, DC, USA.
12 Unitat de Recerca i Laboratori d'Etologia, Fundació Mona, Riudellots de la Selva, Girona, Spain.
13 Chimbo Foundation, Amsterdam, Netherlands.
14 KMDA, Centre for Research and Conservation, Royal Zoological Society of Antwerp, Antwerp, Belgium.
15 School of Natural Sciences, University of Stirling, UK.
16 Agence National des Parcs Nationaux (ANPN) Batterie 4, BP20379, Libreville, Gabon.
17 University of Cambridge, Cambridge, United Kingdom.
18 School of Human Evolution and Social Change, Arizona State University, Tempe, USA.
19 Institute of Human Origins, Arizona State University, Tempe, USA.
20 EcoHealth Alliance, New York.
21 Institut Català de Paleoecologia Humana i Evolució Social (IPHES), Universitat Rovira i Virgili (URV), Spain.
22 Lester E. Fisher Center for the Study and Conservation of Apes, Chicago, United States.
23 Chester Zoo, Chester.
24 STICS - Stirling Conservation Science, Stirling.

25 Department of Anthropology, University College London, 14 Taviton St, Bloomsbury London.
26 Department of Anthropology, Washington University in Saint Louis, St. Louise, United States
27 School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool, United Kingdom.
28 Université de Neuchâtel, Institut de Biologie, Neuchâtel, Switzerland.
29 Max Planck Institute for Human Behavior, Ecology & Culture, Leipzig, Germany.
30 UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, London, UK
31 MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom.
32 Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, BS8 2BN, UK.
33 German Centre for Integrative Biodiversity Research (iDiv), Leipzig, Germany.
34 Catalan Institution of Research and Advanced Studies (ICREA).
35 CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona.
36 Institut Catala de Paleontologia Miquel Crusafont, Universitat Autonoma de Barcelona, Cerdanyola del Vallès, Spain.

Abstract

Chimpanzees are under an enormous threat of extinction due to human impact on their natural habitat, poaching and illegal trade. The genomic study of wild endangered populations has always been hampered by the availability of samples. Non-invasive samples, such as feces, can be obtained without harm to the animal and from their natural habitat but low quality and endogenous DNA content are obstacles for the generation of genome-wide data. Here, we have captured the complete chromosome 21 from 828 fecal samples collected at 51 field sites in Africa and produced the first non-invasive geo-localized catalogue of genomic diversity from the whole extant range of chimpanzees discovering more than 50% of variation not described before. We find strong genetic evidence in support of the four known population clusters or subspecies with distinct patterns of demographic history and barriers impeding genetic exchange. In particular, our results show a clear genetic differentiation even between geographically close groups of central and eastern chimpanzees with some particular locations showing an additional pulse of ancient admixture from bonobos.

The nature of the data allows to determine intra-subspecies population structure, in line with barriers to gene flow between areas that overlap with known geographical barriers. However, we find evidence for connectivity between subspecies in central Africa between 1,500-5,000

years ago. Remarkably, the most endangered populations, western chimpanzees, have been highly interconnected during the last ~800 years, mainly in their north-east distribution.

With our extensive sampling approach, we discovered substantial amounts of new variation, representative of local geographic sites. This fine-grained geo-genetic map allows now for the geo-localization of individual samples as close as 100km from their true location, even for samples at a coverage of less than 1-fold on a single chromosome. These findings provide a novel tool for conservation purposes, such as the determination of the geographical origin for poached and confiscated chimpanzees.

Main

Wild chimpanzee populations were suffering dramatic declines in the last few years (Humle *et al.*, 2016; Kühl *et al.*, 2017), mainly due to anthropogenic factors such as illegal pet and bushmeat trade (Hicks et al. 2010) and habitat destruction linked to extractive industries and intensive agriculture (Funwi-Gabga *et al.*, 2014). This situation led to the classification of chimpanzees as "Endangered" and western chimpanzees as "Critically Endangered" on the IUCN Red List (Humle *et al.*, 2016). The critical status of this species demands urgent actions to avoid the potential extinction of this species. From the point of view of conservation genomics, it is crucial to gain a comprehensive knowledge of the genomic landscape of a threatened species (Supple and Shapiro, 2018). This information can then guide conservation plans both *in situ* and *ex situ* management (Frandsen *et al.*, 2020) as well as become a tool to infer the origin of confiscated individuals from illegal trade, detect hotspots of poaching (Wasser *et al.*, 2015) and inform putative reintroductions based on genetic criteria (Banes, Galdikas and Vigilant, 2016).

As of today, the available genetic data on chimpanzee populations has described their genome-wide diversity and population structure as well as characterize their past demographic history and patterns of admixture (Fischer *et al.*, 2006; Becquet *et al.*, 2007; Prado-Martinez *et al.*, 2013; de Manuel *et al.*, 2016). Since the whole-genome resequencing studies relied on blood samples, they were restricted to wild-born individuals currently placed in sanctuaries or zoos, after being rescued from illegal trade. The knowledge on their origin is important to draw meaningful conclusions and explain the patterns of diversity linked to geography (de Manuel *et al.*, 2016). In genomic studies on great apes, the confiscation site is usually assigned as the place of origin of each individual, and such assumptions carry inherent uncertainties (Frandsen *et al.*, 2020). Specifically for chimpanzees, the available knowledge on geographical origin is sparse, with substantial sampling gaps, particularly for the western and Nigeria-Cameroon chimpanzees (de Manuel *et al.*, 2016) due to lack of local genome diversity and limited sample sizes.

An alternative to overcome these limitations is the usage of non-invasive (NI) samples, such as feces (Vigilant and Guschanski, 2009). They can be obtained without physical disturbance and harm to the animal and have known GPS coordinates (Morin *et al.*, 1993; Taberlet, Waits and Luikart, 1999). However, low quality of DNA due to degradation, and usually with low levels of endogenous DNA (eDNA) (Perry *et al.*, 2010), has hindered the use of NI samples in large-scale genomic studies, essentially limiting it to the analysis on small numbers of autosomal or uniparental genetic markers (Thalmann *et al.*, 2004; Arandjelovic *et al.*, 2011; Inoue *et al.*, 2013).

Here, we overcome the lack of fine-grained spatial geographic resolution across the entire range of extant chimpanzees and provide the first geo-localized catalogue of genomic diversity from wild chimpanzee populations from non-invasive samples. A total of 828 fecal samples were collected from 51 field sites across the chimpanzee range (Fig. 1a) as part of the PanAfrican Programme: ”The Cultured Chimpanzee” (http://panafrican.eva.mpg.de). Using previously developed strategies for the capture of genomic DNA from such samples (Hernandez-Rodriguez *et al.*, 2018; White *et al.*, 2019; Fontsere et al. 2020, Under Review) (Fig. 1b), we generated sequencing data covering the chromosome 21 from these samples at a median target coverage of 1.61-fold (0 - 90.14-fold) (Fig. 1c), with variability according to site of origin (Extended Data Fig. 1b) and a wide range of endogenous DNA (eDNA) content (Extended Data Fig. 1a, Supporting Information Fig. S1). A substantial proportion of samples (N=100) showed high levels of contamination from other primates (Extended Data Fig. 1c and

Supplementary Information Fig. S2, S4), most probably due to diet or sample misidentification, while others yielded less than 0.5-fold coverage in the target space (N=173) (Fig. 1d and Extended Data Fig. 1d). We analyzed the remaining samples (N=555) for a variety of other features and quality measures (Supplementary Information), to confirm their utility for genomic analyses.

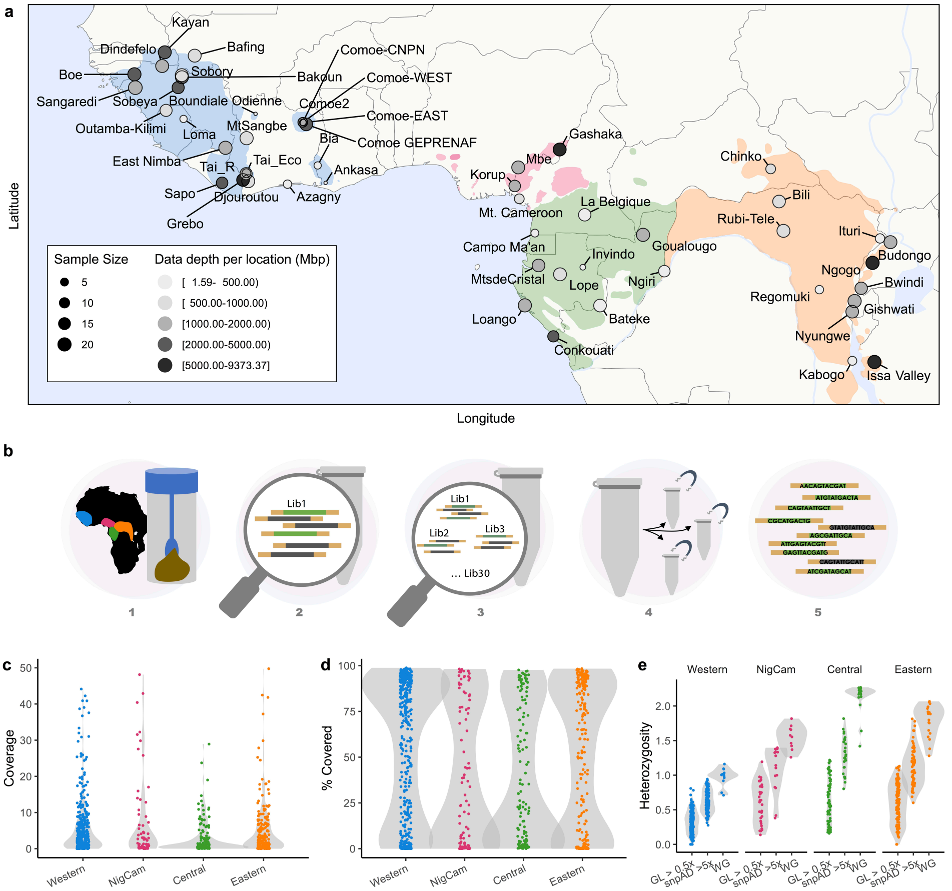


Figure 1 | Chimpanzee geography, sampling scheme, genetic diversity and capture performance.

a, Geographic distribution of chimpanzee populations and PanAf sampling locations. Western chimpanzees are in blue, Nigeria-Cameroon in pink, central in green and eastern in orange. Size of the dots represent the number of sampled feces and the color intensity represents the amount of data generated (Mbp of mapped sequence) at each field site. **b**, Experimental pipeline followed: 1. 828 samples were collected from 51 field sites representing the four chimpanzee subspecies; 2. We prepared one library per sample (Carøe *et al.*, 2018); 3. Between 10 and 30 libraries were pooled equi-endogenously (Fontsere et al., 2020, Under Review); 4. Libraries were enriched for chromosome 21 using target capture methods, each library was captured between three and five times (Fontsere et al., 2020, Under Review); 5.

We generated sequencing data for each library. **c**, Average coverage on the target region of chromosome 21. **d**, Percentage of the target space covered by at least 1 read. **e**, Heterozygosity estimates per subspecies derived from ANGSD genotype likelihood on PanAf samples with more than 0.5-fold coverage (GL >0.5x), from snpAD genotype calls on PanAf samples with more than 5-fold coverage and from GATK genotype calls on previously published whole-genome (WG) chimpanzee samples (de Manuel et al, 2016).

Geographic stratification of chimpanzee populations

We deemed the samples with more than 0.5-fold coverage to be of sufficient quality (N=555) (Supplementary Information), and recovered the clustering according to the four known subspecies (Extended Data Fig. 1e) which most likely formed during the Middle Pleistocene (de Manuel *et al.*, 2016). We find that heterozygosity was higher in central chimpanzees, followed by eastern, Nigeria-Cameroon and western chimpanzee subspecies (Fig. 1e), coherent with previously known patterns from high-quality samples (Prado-Martinez *et al.*, 2013; de Manuel *et al.*, 2016).

Local population stratification was already described for the eastern and central chimpanzee subspecies (de Manuel *et al.*, 2016). However, with our increased sampling density and reliability of the true origin of samples, we can now explore fine-scale population structure across the whole geographic range (Extended Data Fig. 2a,b). With this new approach we also determined structure in Nigeria-Cameroon and western chimpanzees (Extended Data Fig. 2c,d). In the latter, we find a variation cline from east to west (Fig. 2a), with three groups: a Northwestern group, a Southern group and the Ivory Coast east group (Comoé). In eastern chimpanzees, a north-south cline can be observed from the northern Democratic Republic of Congo (DRC) and Central African Republic (CAR) to samples from Tanzania (Fig. 2b), and in central chimpanzees we detect two clusters, a northern group and a southern group. These patterns of stratification are also supported by F3 statistics and F_{ST} (Extended Data Fig. 2e,f,g). Sites from the same subspecies cluster together and geographically close field sites tend to share more drift and exhibit higher genetic similarity.

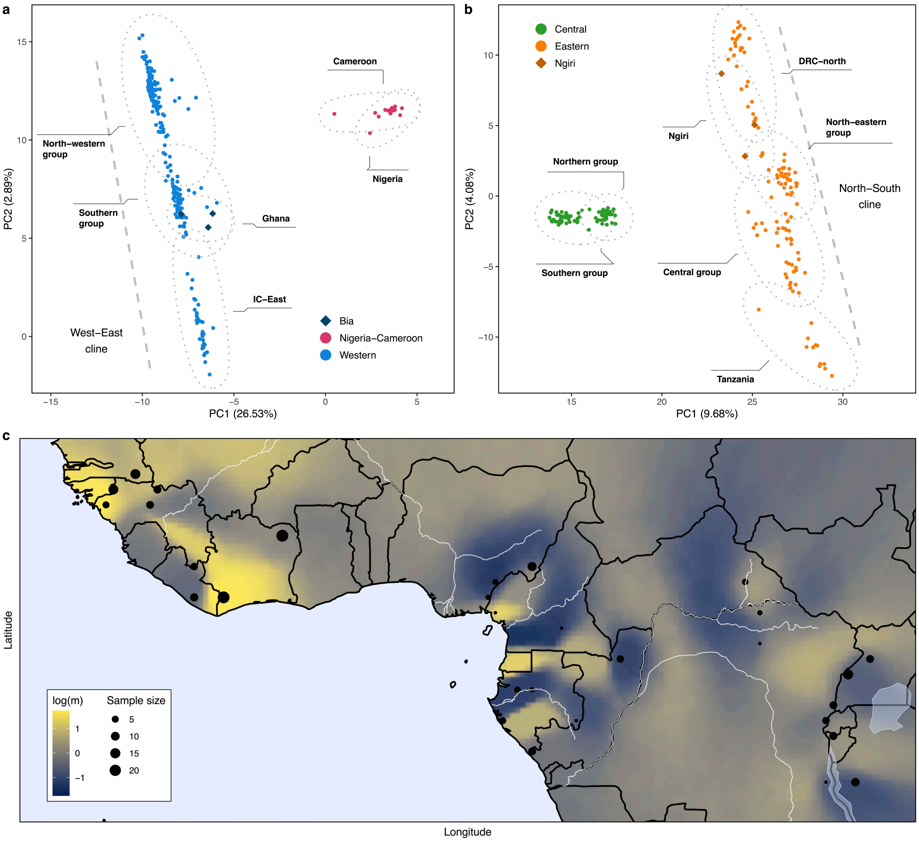


Figure 2 | Chimpanzee genetic structure and effective migration.

a, Procrustes transformed PCA of western (blue) and Nigeria-Cameroon (pink) chimpanzee subspecies. Dark blue diamonds mark Bia field site in Ghana representing the easternmost site from the extant western chimpanzee range. In western chimpanzees, the northwestern group includes Kayan, Dindefelo, Bafing, Sobory, Bakoun, Boe, Sangaredi, Sobeya, Bakoun and Outamba-Kilimi field sites. The southern group is formed by Tai_Eco, Tai_R, Sapo, Grebo, East Nimba, MtSangbe and Djouroutou. The Ivory Coast (IC) east group includes all Comoe field sites. In Nigeria-Cameroon chimpanzees, Mbe and Gashaka are located in Nigeria while Mt. Cameroon and Korup are located in Cameroon. **b**, Procrustes transformed PCA of central (green) and eastern (orange) chimpanzee subspecies. Dark orange diamonds represent the Ngiri field site located on the westernmost part of the eastern chimpanzee distribution, geographically close to central chimpanzee field sites. In central chimpanzees, the northern group includes Campo Ma'an, Mts de Cristal, Goualougo, La Belgique and Invindo, and the southern group includes Loango, Conkouati, Lope and Bateke. In eastern chimpanzees, the DRC-north group is formed by Bili, Rubi-Tele and Chinko (which is located in CAR), the north-east group includes Ngogo, Budongo and Regomuki, and the central group includes Bwindi, Gishwati, Nyungwe and Kabogo. Finally, Issa Valley is the sole representative of Tanzania. **c**, Effective migration rates obtained with EEMS, darker blue represents lower than expected migration under isolation by distance and thus points to a barrier of gene flow while light yellow describes more than expected migration under isolation-by-distance, and therefore marks putative gene-flow corridors. The size of the dots indicates the number of samples used. Major rivers are marked in white.

The question whether the genetic diversity in central and eastern chimpanzee populations reflects two distinctly separated subspecies, or rather a cline of variation under isolation-by-distance, has not been fully resolved previously due to sampling gaps (Fünfstück *et al.*, 2015; de Manuel *et al.*, 2016). In the PanAf dataset, we collected feces from Ngiri, an eastern chimpanzee field site from western DRC, located at the border separating the two subspecies. Ngiri is geographically much closer to a central chimpanzee field site (Goualougo, 281.5 km) than to any other eastern chimpanzee site (Rubi-Tele, 843.6 km and Bili, 898.9 km) in the sampled dataset. However, chimpanzees from Ngiri clearly fall within the genetic diversity of eastern chimpanzees (Fig. 2b, and Extended Data Fig. 4a), pointing to a clear separation of both subspecies for an extended period of time. The same strategy was explored with samples from Ghana (field site Bia), located at the easternmost edge of the current western chimpanzee distribution. These samples fall within the southern group of western chimpanzees (Fig. 2a).

The genetic stratification of chimpanzee subspecies can be interpreted in the context of geographical barriers impeding gene flow, such as major rivers present in tropical Africa (Mitchell *et al.*, 2015). We applied the EEMS method (Petkova, Novembre and Stephens, 2016) to analyze long-term migration landscapes during the Late Pleistocene and early Holocene (Al-Asadi *et al.*, 2019). We find evidence for regions of reduced effective migration that overlap with the Sanaga River (separating Nigeria-Cameroon and central chimpanzees) and the Ubangi river (separating central and eastern chimpanzees) (Fig. 2c and Supplementary Information Fig. S7). We also detected a clear reduction of migration between central chimpanzee populations, overlapping the Ogooué river in Gabon, which separates the northern group from the southern group (Fig. 2c, Supplementary Information Fig. S8). In contrast, western chimpanzees show a rather high historical connectivity across their range in comparison to non-western chimpanzees (Fig. 2c and Supplementary Information Fig. S8). A few field sites, such as Mbe (Nigeria-Cameroon chimpanzee) and Issa Valley (eastern chimpanzee) (Supplementary Information Fig. S8) appear to be relatively more isolated than the rest.

Ancient admixture from bonobos into the non-western chimpanzee subspecies to a small extent (less than 1%) has been described before (de Manuel *et al.*, 2016), most likely as a part of a complex population history of the *Pan* clade (Kuhlwilm *et al.*, 2019) and probably with differential consequences in terms of selection in the different subspecies (Nye *et al.*, 2018). Unfortunately, tests for a significant enrichment of allele sharing (D-statistic, F-statistics) (Peter, 2016) on the target space of chromosome 21 from fecal samples did not provide statistically significant results due to low coverage, capture bias on chimpanzee alleles and high levels of missing data, although we could replicate previous findings with WG data only on the chromosome 21. We inferred introgressed fragments (Peter, 2020) from bonobos into chimpanzees from all field sites, and we find that the southern group of central chimpanzees carries significantly more bonobo ancestry than the northern group (Wilcoxon rank sum test, p-value = 3.441e-08) or any other chimpanzee population (p-value < 0.01; Extended Data Fig. 3e). This provides further evidence for multiple phases of genetic exchange between chimpanzees and bonobos, with differential effects in specific geographical areas.

Connectivity between chimpanzee populations

We investigated the relationships between chimpanzee field sites in a recent timeframe beyond family relationships by analyzing shared DNA segments also termed ‘identical by descent’ (IBD) tracts. This approach can reveal recent coalescent events during the Holocene, up to ~5,000 years (Thompson, 2013). The timing of such genetic exchange is related to the length of the shared segments, with more recent migration resulting in longer IBDs (Supplementary Information). We observed an exponential decay of IBD lengths with geographical distance within eastern and western chimpanzees (Extended Data Fig. 3a), as expected for a scenario where isolation by distance took place.

We quantified the pairwise average lengths and numbers of IBDs present per field site. Western chimpanzee communities appear to have high levels of connectivity between them, represented by more and longer IBDs segments than any other subspecies, especially within their northern group (Kayan, Dindefelo, Boe, Sangaredi, Bakoun-Sobory and Sobeya) (Fig. 3c). Such connectivity could relate to recent migration within the past ~800 years (range of 155-3,300 years), but also population expansion. We also detected high effective migration (Fig. 2c and Supplementary Information Fig. S7), which suggests a long-term connectivity between western chimpanzee communities. Interestingly, we find MtSangbe and at lesser extent Comoe, to have been isolated rather recently, with relatively fewer shared IBD segments with nearby sites, while the longer-term historical connectivity appears to have been higher.

Southern central chimpanzees show recent connectivity until ~1,500 years ago (Loango, Lope, Bateke and Conkouati), with more and slightly longer IBD segments compared to the connectivity within their northern clade (MtsdeCristal, La Belgique and Goaulougo), and between both groups. This is consistent with a rather strong barrier between them that has been maintained until recent times (Fig. 3a, 2c), as also supported by the genetic differentiation represented by F_{ST} (Extended Data Fig. 2). All sites of Nigeria-Cameroon chimpanzees seem to have been connected within the past 2,500 years (Fig. 3d), suggesting that Mbe was isolated only further back in time (Fig. 2c, Supporting Information Fig. S8). The connectivity between Mt.Cameroon and Korup has been very high until ~600 years ago, as they share a large number of long IBDs segments. Within eastern chimpanzee sites, we observed three clusters of connectivity (Fig. 3b): Bili-Chinko, Budongo-Ngogo, and Gishwati-Nyungwe. Past migration (Fig. 2c) between Bili and Chinko was estimated to be high, which has apparently continued until as recently as 350 years ago. A corridor of dispersal between chimpanzee communities between Budongo and nearby forests had already been proposed (McCarthy et al., 2015). Interestingly, although Bili and Rubi-Tele are located in close proximity (~198 km), Rubi-Tele shares no IBDs with Bili, but is connected with Ngogo, Bwindi and Budongo towards the east. Kabogo is the most isolated site of eastern chimpanzees; while geographically close to Issa Valley, it is connected with other sites to the north, demonstrating that Lake Tanganyika seems to have been a barrier to gene flow.

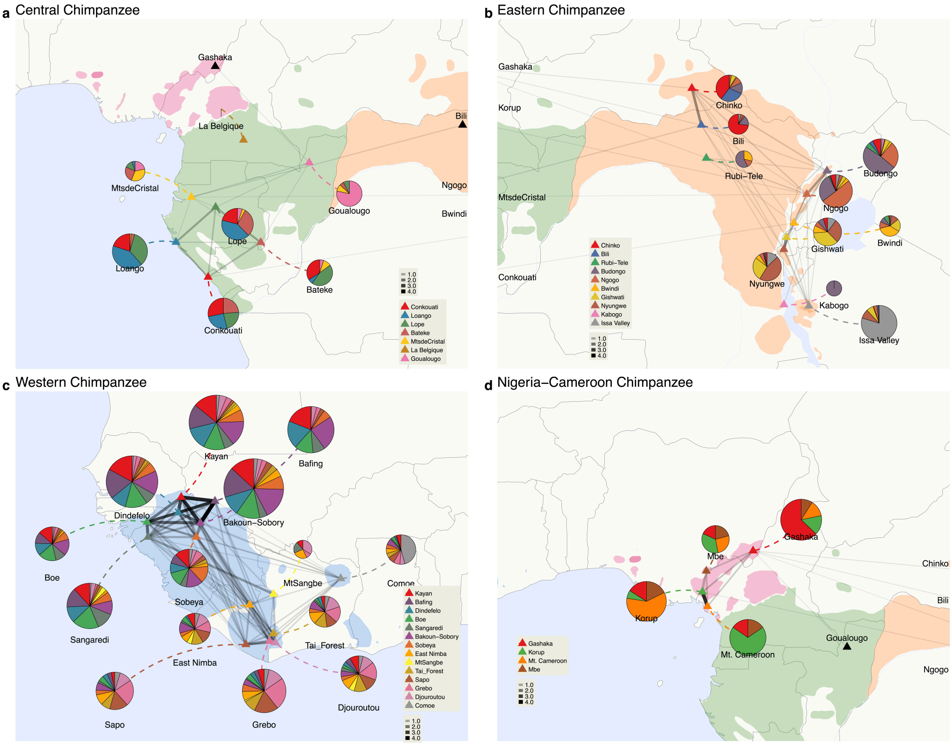


Figure 3 | Recent connectivity between chimpanzee populations.
Size of the pie charts represents the pairwise number of shared fragments, normalized by the number of pairs. The thickness of the lines indicates the average length of the IBDs (in Mbps). Triangles show the location of sites. **a**, Central chimpanzees **b**, eastern chimpanzees **c**, western chimpanzees and **d**, Nigeria-Cameroon chimpanzees.

Overlapping the timeframes of both IBD and EEMS, we explored the sharedness of rare alleles, which likely represents connectivity between 1.5kya and 15kya (Schiffels et al., 2016). We calculated the proportion of variants which are almost uniquely derived in a given population but shared with any other population (Supplementary Information). The observations largely agree with the conclusions derived from both IBD and EEMS methods. Interestingly, we find an area of extensive allele sharing in the central chimpanzee southern region compared to the northern group. In eastern chimpanzees, we propose a recent expansion from populations in the central-eastern part (Budongo, Bwindi, Gishwati, Ngogo, Nyungwe) to the south (Issa Valley, Kabogo), west (Regomuki) and northwest (Rubi-Tele, Bili, Chinko). In western chimpanzees, we propose that in the northwestern range may have been subject to recent expansions into the fringes (Bafing and Sangaredi), while the intermediate region (East Nimba and Outamba-Kilimi) might have been a corridor of genetic exchange between the northwestern and southern areas.

We also detected short IBD segments (less than ~0.5 Mbp) shared between individuals from different subspecies, suggesting that this observation may be the consequence of connectivity further back in time (around 5000 years ago). Genetic exchange between chimpanzee subspecies and possible corridors of migration in the past have been suggested before (de Manuel *et al.*, 2016). We find evidence of a past connection between the northern group in central chimpanzee subspecies, especially Goualougo, and Gashaka (Nigeria-Cameroon chimpanzees), which is supported by rare allele sharing (Extended Data Fig. 4b,c) as well as introgressed fragments (Extended Data Fig. 3b,d). Gashaka also has more eastern ancestry compared to the rest of Nigeria-Cameroon chimpanzees, possibly due to shared ancestry between central and eastern chimpanzee subspecies, since Goualougo has the highest eastern ancestry in central chimpanzee populations (Extended Data Fig. 3c, 4c). Surprisingly, Issa Valley in Tanzania exhibits the highest levels of central chimpanzee ancestry (Extended Data Fig. 3b).

Geolocalization of rescued chimpanzees

Our dense sampling of chimpanzee populations distributed across their extant range allowed for the discovery of ~50% new variation previously not reported only on chromosome 21 (Extended data Fig. 5). Importantly such variation is linked to specific geographical locations, which has direct implications for conservation biology. We developed a strategy to use rare variation almost private to each chimpanzee population to infer their origin (Supplementary Information). Rare alleles have been a useful resource for studying human populations (Gravel *et al.*, 2011), since these variants were arising during the past few 100s or 1,000s of years. Therefore, they should represent the recent variation emerging locally in chimpanzee groups which were only loosely connected after their split.

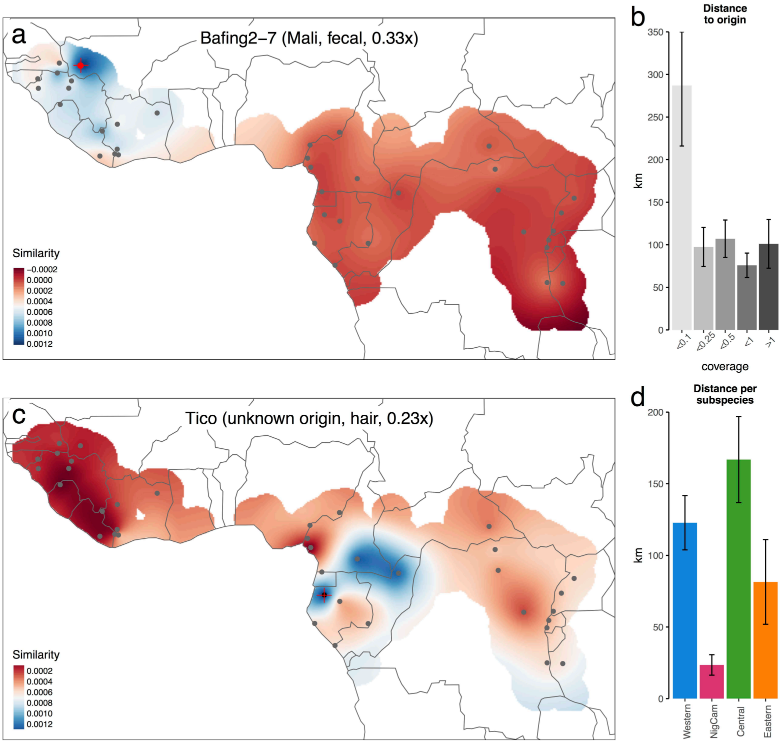


Figure 4 | Chimpanzee origin inferences based on rare variation.
a, Spatial representation of sharedness of rare alleles. Red color indicates lower amounts of shared rare variation, while blue indicates larger amounts of sharedness of rare alleles. Black dots are the field sites used as reference for the geolocalization, the red dot indicates the known place of origin, in this case the sample Baf2-7 is from Bafing and the red cross represents the inferred origin. Both the red cross and the red dot overlap since the sample is correctly assigned to its true origin. **b**, Average distance of best matching to true location in bins of coverage. Samples with >1x coverage included here have more than 0.5% human contamination. **c**, Geolocation of a chimpanzee from a rescue center, based on rare variation. Sample Tico is assigned to come from the northern area of central chimpanzees. **d**, Average distance of best matching to true location per subspecies.

We tested the method on a panel of 196 samples sequenced to less than 1-fold coverage and with known locations within the reference panel (Fig. 4a), as well as 28 samples with a coverage of more than 1-fold which were excluded from the reference panel due to human contamination. We find that 62% (139) were correctly assigned to their reference population, and for 86%, the correct population was among the top three ranked populations. We determined the most likely region of origin, rather than relying on the best match only by using a spatial representation of matching to all reference sites (Fig. 4a). At a coverage below 0.1-

fold, the average distance to the true origin is 287km, while already at a coverage of more than 0.1-fold and beyond, we determined the samples to be within 100 km of their true origin, even when human contamination was present (Fig. 4b and d).

Finally, we used this strategy to estimate the most probable origin of 20 chimpanzees from two Spanish rescue centers (Fundació Mona and Centro de Rescate de Primates Rainfer), which were sequenced at low coverage from hair and blood samples (median 0.35-fold coverage, ranging from 0.15-fold to 4.3-fold). We find that these samples show very similar patterns compared to the fecal samples included in the PanAf dataset (Fig 4c, Fig. S10, S11).

Discussion

Geographic origin is a good predictor of genetic structure not only in chimpanzee subspecies but also in local-scale population stratification. In some instances, population stratification in chimpanzee populations can be explained by isolation-by-distance. But known ecological or geographical barriers cause a reduced gene flow between populations. We find strong evidence supporting the four known clusters of chimpanzee populations, and we link genetic barriers of gene flow between them with geographical barriers: the Sanaga river and the Ubangi river. The Ogooué river also acts as a barrier for central chimpanzee populations and the Lake Tanganyika induces isolation in eastern chimpanzees. In more recent times, populations belonging to the same subspecies have shared genetic material. We detect possible recent expansions and increased connectivity in the northwest distribution of western chimpanzees, and increased isolation in some areas such as MtSangbe, Mbe and Kabogo and between the northern group of central chimpanzees. Further back in time, corridors of gene-flow between non-western chimpanzee subspecies have also been detected, mainly between Goualougo (northern central chimpanzees) and Gashaka (Nigeria-Cameroon chimpanzees). Finally, we suggest an additional event of ancient introgression from Bonobos particularly to the southern group of central chimpanzees.

Our high-density sampling scheme has proven effective to discover new genetic variation linked to specific locations, even though capture of a single chromosome in non-invasive samples. We demonstrate the need to have a comprehensive catalogue of genetic diversity avoiding sampling gaps which could hamper proper conclusions.

With our sampling approach and after determining population structure linked to geographical origin, we devised a strategy for the geo-localization of samples of unknown origin. This strategy has a precision of ~100km, even when using low-coverage or contaminated samples. This newly developed resource for the geolocalization of chimpanzees can have direct conservation applications. On the one hand, it can inform the captivity management of rescued chimpanzees, so they are placed into a suitable sanctuary as well as to guide future reintroduction into the wild whenever possible. On the other hand, it will allow the detection of poaching hotspots, so the competent authorities can enforce the law in the countries where these animals come from.

In conclusion, through the capture of chromosome 21 on hundreds of chimpanzee fecal samples we presented the first non-invasive catalogue of genomic diversity in extant wild chimpanzee populations. This new resource has allowed not only to describe fine-scale population structure, past and recent gene-flow and migration events but also to build a geogenetic map for the geolocalization of rescued chimpanzees.

Methods

Fecal DNA was extracted and screened with a microsatellite genotyping assay (Arandjelovic *et al.*, 2009, 2011). A unique double-inline barcoded library was prepared for each sample following the BEST protocol with minor modifications (Carøe *et al.*, 2018; Fontsere et al. 2020 Under Review). Pooling for capture was devised based on the endogenous DNA content (fraction of chimpanzee DNA, relative to gut microbial and exogenous DNA) (Supplementary Information) (Hernandez-Rodriguez *et al.*, 2018; Fontsere et al. 2020 Under Review). Each pool was divided into several aliquots to perform multiple hybridizations. Afterwards, with predesigned RNA baits (SureSelect Agilent), we captured the chromosome 21 following the protocol provided by Agilent Sureselect Custom Array, adding two consecutive hybridization rounds for pools containing samples with <5% eDNA.

Captured libraries were sequenced on the HiSeq 4000 Illumina platform with 2x100 paired-end reads. We processed the data to demultiplex libraries belonging to the same hybridization pool using Sabre (https://github.com/najoshi/sabre) and reads were trimmed with Trimmomatic (version 0.36) (Bolger, Lohse, & Usadel, 2014). Paired-end reads were then aligned to the human genome Hg19 (GRCh37, Feb.2009 (GCA_000001405.1)) using BWA (version 0.7.12) (Li and Durbin, 2009). Duplicates were removed using PicardTools (version 1.95) (http://broadinstitute.github.io/picard/) and further filtering of the reads was done using samtools (version 1.5) (Li *et al.*, 2009). To retrieve the on-target reads we used intersectBed from the BEDTOOLS package (version 2.22.1) (Quinlan and Hall, 2010). Average coverage of the target space was calculated as the number of bases in the target region divided by the size of the target space.

We obtained genotype likelihoods using ANGSD version 0.916 (Korneliussen, Albrechtsen and Nielsen, 2014) and genotype calls using snpAD v0.3.2 (Prüfer, 2018), a software that takes DNA damage into account for genotype calling.

PCAs were obtained using PCAangsd (Meisner and Albrechtsen, 2018) (Supplementary Information Fig. S2). The sources of primate contamination on fecal samples were devised using BBsplit software (https://sourceforge.net/projects/bbmap/), mapping to 11 different primate genomes (Supplementary Information Fig. S4, S5). Human contamination was estimated as the fraction of the number of observations of human-like alleles across all positions where chimpanzees and humans consistently differ, as described previously (Fontsere et al., 2020 Under Review) (Supplementary Information). Although samples had been screened prior to library preparation with a microsatellite assay (Arandjelovic *et al.*, 2009, 2011) to identify and discard identical or first order relative individuals, we used NgsRelate (Korneliussen and Moltke, 2015) (Supplementary Information Fig. S6).

Due to the high variability of sample qualities and specific requirements for the application of different methods, a variety of filtering procedures was applied. In most analyses, samples with evidence of contamination from either human (>1% or >0.5%) or other primate species were filtered out, as well as 1st degree and identical samples, and those that were found to not belong to the expected subspecies or site (Supplementary Information). Finally, we used samples with different coverage cutoffs for different analysis (0.5-fold, 1-fold or 5-fold). To obtain pairwise F_{ST} estimates between field sites, we computed the 2-dimensional SFS (2d SFS) between each pair of geographical sites with ANGSD -doSaf 1 and realSFS (Korneliussen, Albrechtsen and Nielsen, 2014). The genetic relationships between populations were used to build a matrix, from which we constructed a neighbour-joining tree using the ape package (Paradis, Claude and Strimmer, 2004) in R (version 3.5.2). F3 outgroup statistics were calculated between field sites using qp3Pop (Patterson *et al.*, 2012) and taking an orangutan as the outgroup (pygmaeus_ERS1986511) (Nater *et al.*, 2017).

Long-term effective migration rates were calculated using EEMS (Petkova, Novembre and Stephens, 2016) with samples of more than 5-fold coverage (Supplementary Information). The same dataset was used to obtain IBDs with IBDseq software (Browning and Browning, 2013) (Supplementary Information). Sharedness of IBD segments is likely the consequence of recent migration events between geographic sites or areas. The length of the shared segments would be correlated to the time of such genetic exchange, with more recent migration causing longer IBDs. To time the events, we followed the rule of G=100/(2*Mbp) (Thompson, 2013). Mbp stands for the length of the fragments in IBD, and G is the number of generations. Then, we assumed a generation time of 25 years to calculate the time (Langergraber *et al.*, 2012). We took the maximum IBD length per pair of individuals between sites to estimate the timeframe of connectivity. Subspecies ancestry introgressed fragments and bonobo introgression were calculated with Admixfrog (Peter, 2020). The reference panel on the chromosome 21 (*source*) was built using an equal number of individual genomes of each chimpanzee subspecies (6 genomes), 10 bonobo genomes (de Manuel *et al.*, 2016), two human genomes (Mallick *et al.*, 2016) and 1 orangutan (Nater *et al.*, 2017). Rare allele variation nearly-private to each field site was used to estimate the most probable origin of chimpanzee fecal samples (Supplementary Information). We tested our approach by performing shallow sequencing at median 0.25-fold coverage (ranging from 0.15-fold to 4.3-fold) on blood and hair samples from 20 rescued chimpanzees from two Spanish rescue centers (Supplementary Information).

References

Al-Asadi, H. *et al.* (2019) “Estimating recent migration and population-size surfaces,” *PLoS genetics*, 15(1), p. e1007908.

Arandjelovic, M. *et al.* (2009) “Two-step multiplex polymerase chain reaction improves the speed and accuracy of genotyping using DNA from noninvasive and museum samples,” *Molecular ecology resources*, 9(1), pp. 28–36.

Arandjelovic, M. *et al.* (2011) “Non-invasive genetic monitoring of wild central chimpanzees,” *PloS one*, 6(3), p. e14761.

Banes, G. L., Galdikas, B. M. F. and Vigilant, L. (2016) “Reintroduction of confiscated and displaced mammals risks outbreeding and introgression in natural populations, as evidenced by orang-utans of divergent subspecies,” *Scientific Reports*, 6(1). doi: 10.1038/srep22026.

Becquet, C. *et al.* (2007) “Genetic structure of chimpanzee populations,” *PLoS genetics*, 3(4), p. e66.

Browning, B. L. and Browning, S. R. (2013) “Detecting identity by descent and estimating genotype error rates in sequence data,” *American journal of human genetics*, 93(5), pp. 840–851.

Carøe, C. *et al.* (2018) “Single-tube library preparation for degraded DNA,” *Methods in Ecology and Evolution*, 9(2), pp. 410–419.

Fischer, A. *et al.* (2006) “Demographic history and genetic differentiation in apes,” *Current biology: CB*. Elsevier BV, 16(11), pp. 1133–1138.

Fontsere, C *et al.* (2020 Under Review) "Maximizing the acquisition of unique reads in non-invasive capture sequencing experiments," *Molecular Ecology Resources*

Frandsen, P. *et al.* (2020) “Targeted conservation genetics of the endangered chimpanzee,” *Heredity*, 125(1–2), pp. 15–27.

Fünfstück, T. *et al.* (2015) “The sampling scheme matters: Pan troglodytes troglodytes and P. t. schweinfurthii are characterized by clinal genetic variation rather than a strong subspecies break,” *American journal of physical anthropology*. Wiley, 156(2), pp. 181–191.

Funwi-Gabga, N. *et al.* (2014) “State of the Apes 2013: Extractive Industries and Ape Conservation,” in Arcus Foundation (ed.) *The status of apes across Africa and Asia*. Cambridge, UK: Cambridge University Press., pp. 252–277.

Gravel, S. *et al.* (2011) “Demographic history and rare allele sharing among human populations,” *Proceedings of the National Academy of Sciences of the United States of America*, 108(29), pp. 11983–11988.

Hernandez-Rodriguez, J. *et al.* (2018) “The impact of endogenous content, replicates and pooling on genome capture from faecal samples,” *Molecular ecology resources*, 18(2), pp. 319–333.

Humle, T. *et al.* (2016) *Pan troglodytes (errata version published in 2018)*, *The IUCN Red List of Threatened Species 2016: e.T15933A129038584*. Available at: https://dx.doi.org/10.2305/IUCN.UK.2016-2.RLTS.T15933A17964454.en (Accessed: August 31, 2020).

Inoue, E. *et al.* (2013) “Male genetic structure and paternity in western lowland gorillas (Gorilla gorilla gorilla),” *American journal of physical anthropology*, 151(4), pp. 583–588.

Korneliussen, T. S., Albrechtsen, A. and Nielsen, R. (2014) “ANGSD: Analysis of Next Generation Sequencing Data,” *BMC bioinformatics*, 15, p. 356.

Korneliussen, T. S. and Moltke, I. (2015) “NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data,” *Bioinformatics*, p. btv509.

Kühl, H. S. *et al.* (2017) “The Critically Endangered western chimpanzee declines by 80,” *American journal of primatology*, 79(9). doi: 10.1002/ajp.22681.

Kuhlwilm, M. *et al.* (2019) “Ancient admixture from an extinct ape lineage into bonobos,” *Nature ecology & evolution*, 3(6), pp. 957–965.

Langergraber, K. E. *et al.* (2012) “Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution,” *Proceedings of the National Academy of Sciences of the United States of America*. Proceedings of the National Academy of Sciences, 109(39), pp. 15716–15721.

Li, H. *et al.* (2009) “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, 25(16), pp. 2078–2079.

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

Mallick, S. *et al.* (2016) “The Simons Genome Diversity Project: 300 genomes from 142 diverse populations,” *Nature*, 538(7624), pp. 201–206.

de Manuel, M. *et al.* (2016) “Chimpanzee genomic diversity reveals ancient admixture with bonobos,” *Science*, 354(6311), pp. 477–481.

McCarthy, M. S. *et al.* (2015) “Genetic censusing identifies an unexpectedly sizeable population of an endangered large mammal in a fragmented forest landscape,” *BMC Ecology*, 15(1). doi: 10.1186/s12898-015-0052-x.

Meisner, J. and Albrechtsen, A. (2018) “Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data,” *Genetics*, 210(2), pp. 719–731.

Mitchell, M. W. *et al.* (2015) “Environmental variation and rivers govern the structure of chimpanzee genetic diversity in a biodiversity hotspot,” *BMC evolutionary biology*. Springer Nature, 15(1), p. 1.

Morin, P. A. *et al.* (1993) “Non-invasive sampling and DNA amplification for paternity exclusion, community structure, and phylogeography in wild chimpanzees,” *Primates*, 34(3), pp. 347–356.

Nater, A. *et al.* (2017) “Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species,” *Current biology: CB*, 27(22), pp. 3576–3577.

Nye, J. *et al.* (2018) “Selection in the Introgressed Regions of the Chimpanzee Genome,” *Genome biology and evolution*, 10(4), pp. 1132–1138.

Paradis, E., Claude, J. and Strimmer, K. (2004) “APE: Analyses of Phylogenetics and Evolution in R language,” *Bioinformatics* , 20(2), pp. 289–290.

Patterson, N. *et al.* (2012) “Ancient Admixture in Human History,” *Genetics*, 192(3), pp. 1065–1093.

Perry, G. H. *et al.* (2010) “Genomic-scale capture and sequencing of endogenous DNA from feces,” *Molecular ecology*, 19(24), pp. 5332–5344.

Peter, B. M. (2016) “Admixture, Population Structure, and F-Statistics,” *Genetics*, 202(4), pp. 1485–1501.

Peter, B. M. (2020) “100,000 years of gene flow between Neandertals and Denisovans in the Altai mountains,” *bioRxiv*. doi: 10.1101/2020.03.13.990523.

Petkova, D., Novembre, J. and Stephens, M. (2016) “Visualizing spatial population structure with estimated effective migration surfaces,” *Nature genetics*, 48(1), pp. 94–100.

Prado-Martinez, J. *et al.* (2013) “Great ape genetic diversity and population history,” *Nature*, 499(7459), pp. 471–475.

Prüfer, K. (2018) “snpAD: an ancient DNA genotype caller,” *Bioinformatics* , 34(24), pp. 4165–4171.

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

Schiffels, S. *et al.* (2016) “Iron Age and Anglo-Saxon genomes from East England reveal British migration history,” *Nature communications*, 7, p. 10408.

Supple, M. A. and Shapiro, B. (2018) “Conservation of biodiversity in the genomics era,” *Genome biology*, 19(1), p. 131.

Taberlet, P., Waits, L. P. and Luikart, G. (1999) “Noninvasive genetic sampling: look before you leap,” *Trends in ecology & evolution*, 14(8), pp. 323–327.

Thalmann, O. *et al.* (2004) “Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes,” *Molecular ecology*, 13(2), pp. 321–335.

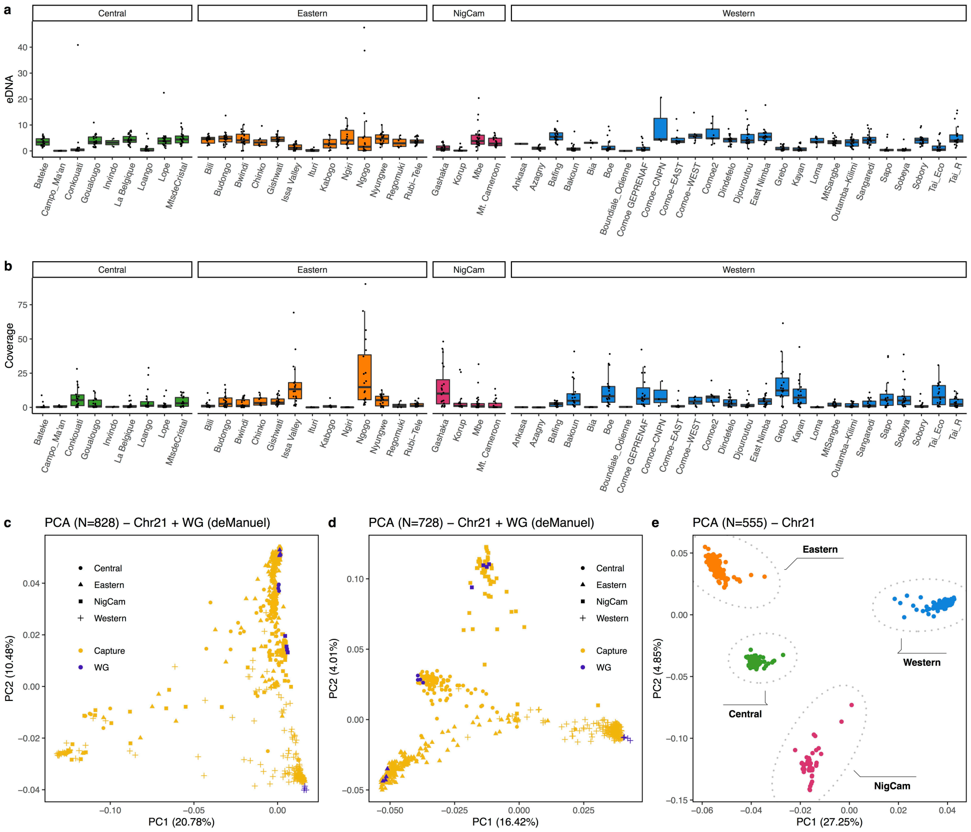
Thompson, E. A. (2013) “Identity by descent: variation in meiosis, across genomes, and in populations,” *Genetics*, 194(2), pp. 301–326.

Vigilant, L. and Guschanski, K. (2009) “Using genetics to understand the dynamics of wild primate populations,” *Primates; journal of primatology*, 50(2), pp. 105–120.

Wasser, S. K. *et al.* (2015) “Genetic assignment of large seizures of elephant ivory reveals Africa’s major poaching hotspots,” *Science*, 349(6243), pp. 84–87.

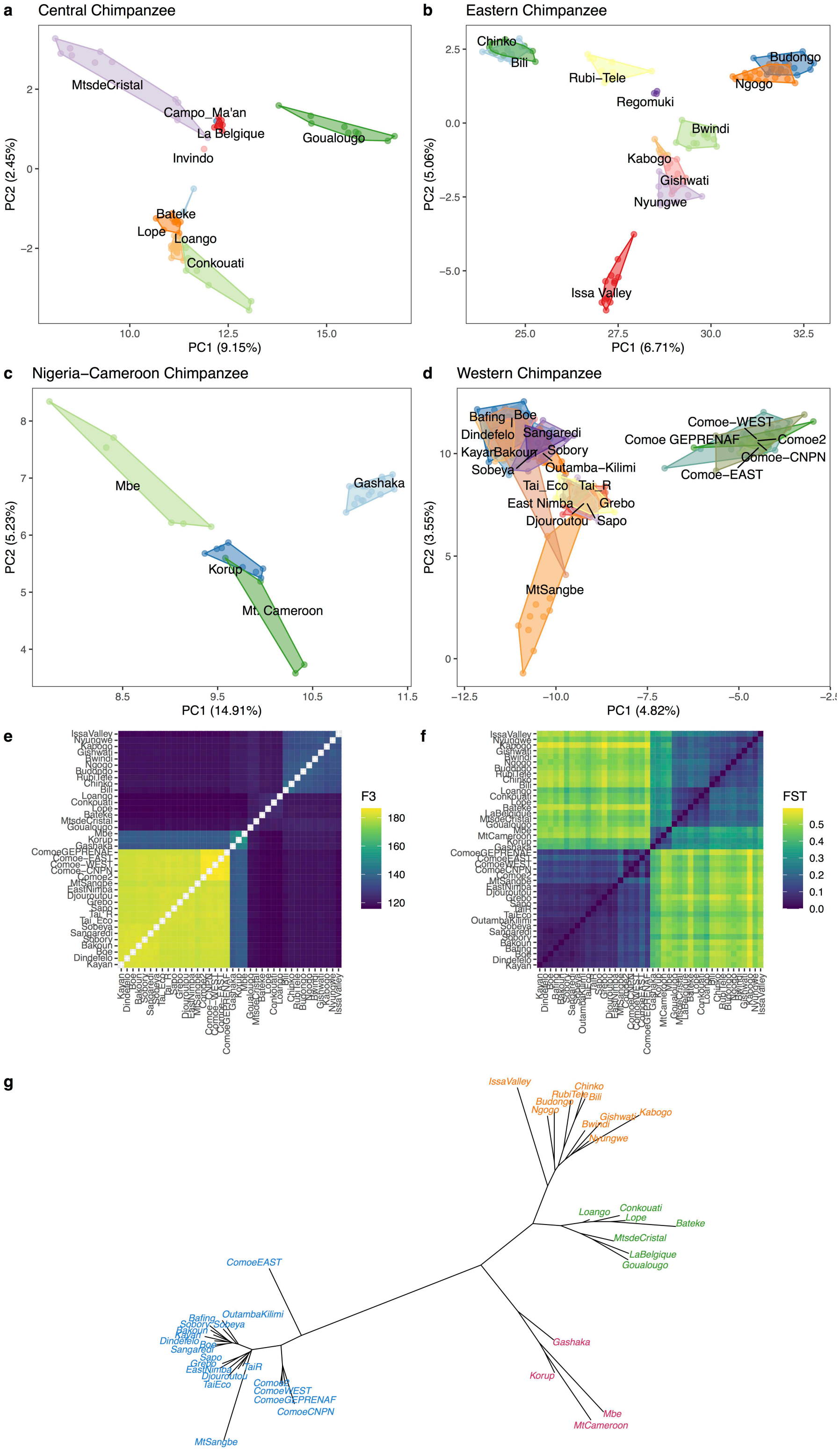
White, L. C. *et al.* (2019) “A roadmap for high-throughput sequencing studies of wild animal populations using noninvasive samples and hybridization capture,” *Molecular Ecology Resources*, 19(3), pp. 609–622.

Extended Data Figures



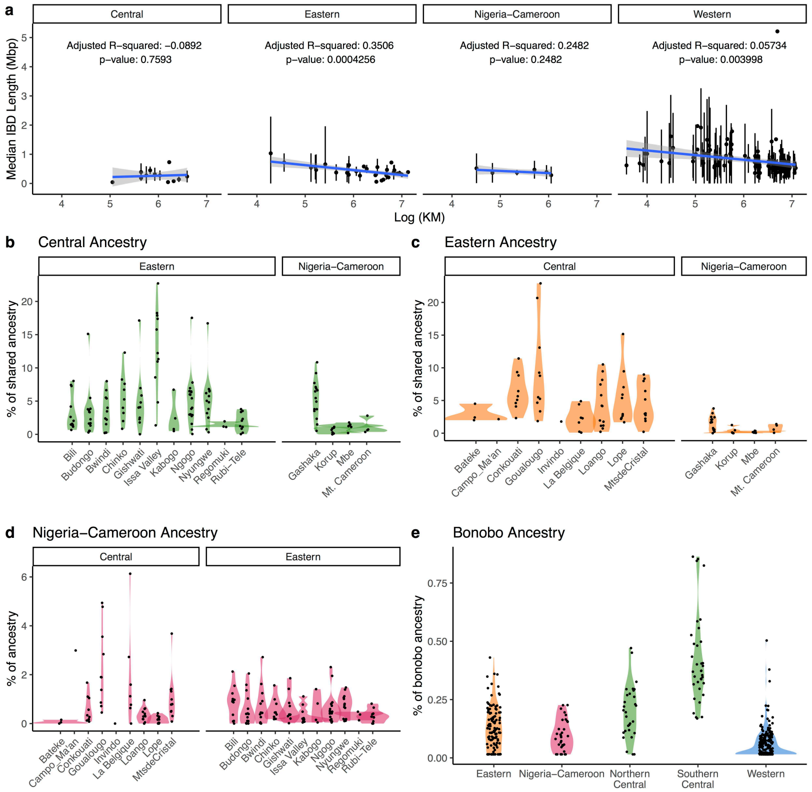
Extended Data Figure 1 | Sample quality assessment

a, Endogenous DNA (eDNA) distribution of fecal samples from 51 PanAfrican field sites. **b**, Average coverage on the target region of chromosome 21 from 51 PanAfrican field sites. PCA of chromosome 21 single-nucleotide polymorphisms on **c,** all captured fecal samples (N=828) and with 4 whole-genomes of each subspecies from de Manuel et al., (2016) on 1,533,092 variable positions in the target space. **d**, on chimpanzee fecal samples (N=728), excluding samples with high levels of primate contamination, and with 4 whole-genomes of each subspecies from de Manuel et al., (2016) (513,406 positions). **e**, and on chimpanzee samples with more than 0.5-fold coverage (N=555).


a
Central Chimpanzee
MtsdeCristal
Campo_Ma'an
La Belgique
Goualougo
Invindo
Bateke
Lope
Loango
Conkouati
PC2 (2.45%)
PC1 (9.15%)
2
0
−2
10.0
12.5
15.0
b
Eastern Chimpanzee
Chinko
Bili
Rubi−Tele
Budongo
Ngogo
Regomuki
Bwindi
Kabogo
Gishwati
Nyungwe
Issa Valley
PC2 (5.06%)
PC1 (6.71%)
2.5
0.0
−2.5
−5.0
25.0
27.5
30.0
32.5
c
Nigeria−Cameroon Chimpanzee
Mbe
Gashaka
Korup
Mt. Cameroon
PC2 (5.23%)
PC1 (14.91%)
8
7
6
5
4
8.5
9.5
10.5
11.5
d
Western Chimpanzee
Bafing
Boe
Dindefelo
Sangaredi
KayanBakoun
Sobory
Sobeya
Outamba−Kilimi
Tai_Eco
Tai_R
East Nimba
Grebo
Djouroutou
Sapo
MtSangbe
Comoe−WEST
Comoe GEPRENAF
Comoe2
Comoe−CNPN
Comoe−EAST
PC2 (3.55%)
PC1 (4.82%)
10
5
0
−12.5
−10.0
−7.5
−5.0
−2.5
e
IssaValley Nyungwe Kabogo Gishwati Bwindi Ngogo Budongo RubiTele Chinko Bili Loango Conkouati Lope Bateke MtsdeCristal Goualougo Mbe Korup Gashaka ComoeGEPRENAF Comoe−EAST Comoe−WEST Comoe−CNPN Comoe2 MtSangbe EastNimba Djouroutou Grebo Sapo Tai_R Tai_Eco Sobeya Sangaredi Sobory Bakoun Boe Dindefelo Kayan
Kayan Dindefelo Boe Bakoun Sobory Sangaredi Sobeya Tai_Eco Tai_R Sapo Grebo Djouroutou EastNimba MtSangbe Comoe2 Comoe−CNPN Comoe−WEST Comoe−EAST ComoeGEPRENAF Gashaka Korup Mbe Goualougo MtsdeCristal Bateke Lope Conkouati Loango Bili Chinko RubiTele Budongo Ngogo Bwindi Gishwati Kabogo Nyungwe IssaValley
F3
180
160
140
120
f
IssaValley Nyungwe Kabogo Gishwati Bwindi Ngogo Budongo RubiTele Chinko Bili Loango Conkouati Lope Bateke LaBelgique MtsdeCristal Goualougo Mbe MtCameroon Korup Gashaka ComoeGEPRENAF ComoeEAST ComoeWEST ComoeCNPN Comoe2 MtSangbe EastNimba Djouroutou Grebo Sapo TaiR TaiEco OutambaKilimi Sobeya Sangaredi Sobory Bakoun Bafing Boe Dindefelo Kayan
Kayan Dindefelo Boe Bafing Bakoun Sobory Sangaredi Sobeya OutambaKilimi TaiEco TaiR Sapo Grebo Djouroutou EastNimba MtSangbe Comoe2 ComoeCNPN ComoeWEST ComoeEAST ComoeGEPRENAF Gashaka Korup MtCameroon Mbe Goualougo MtsdeCristal LaBelgique Bateke Lope Conkouati Loango Bili Chinko RubiTele Budongo Ngogo Bwindi Gishwati Kabogo Nyungwe IssaValley
FST
0.5
0.4
0.3
0.2
0.1
0.0
g
IssaValley
Chinko
RubiTele
Bili
Budongo
Ngogo
Gishwati
Kabogo
Bwindi
Nyungwe
Loango
Conkouati
Lope
Bateke
MtsdeCristal
LaBelgique
Goualougo
ComoeEAST
OutambaKilimi
Bafing
Sobory
Sobeya
Bakoun
Kayan
Dindefelo
Boe
Sangaredi
Sapo
Grebo
EastNimba
TaiR
Djouroutou
TaiEco
Comoe2
ComoeWEST
ComoeGEPRENAF
ComoeCNPN
MtSangbe
Gashaka
Korup
Mbe
MtCameroon

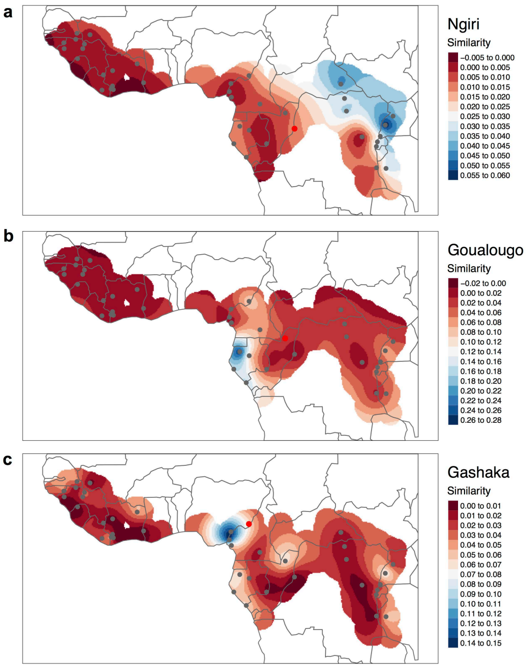
Extended Data Figure 2 | Chimpanzee genetic stratification.

Principal Component Analysis (PCA) of chromosome 21 in each chimpanzee subspecies **a**, Central chimpanzees (N=69). **b**, Eastern chimpanzees (N=119) **c,** Nigeria-Cameroon Chimpanzees (N=34). **d**, Western chimpanzees (N=227). **e**, F3 outgroup statistic, test for shared drift between chimpanzee field sites with orangutan as outgroup. **f**, Pairwise F_{ST} estimates between chimpanzee field sites. **g**, Neighbour-joining tree constructed from the pairwise F_{ST} estimates.

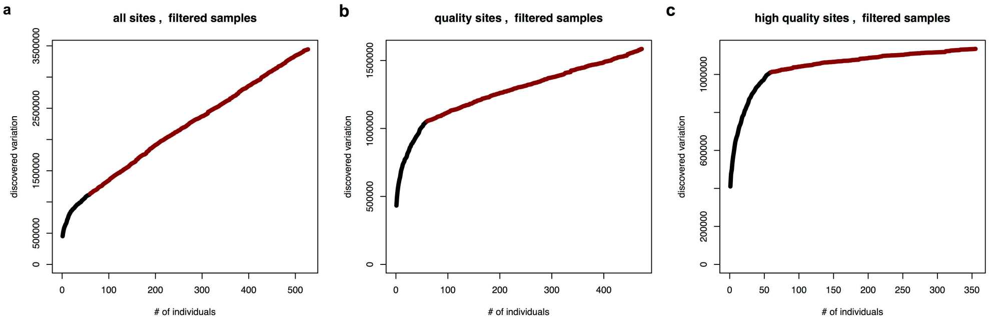


Extended Data Figure 3 | Recent connectivity and shared ancestry between chimpanzee populations.

a, Median length (Mbps) and standard deviation of IBD shared fragments between chimpanzee field sites. Western and eastern chimpanzee subspecies show an exponential decay of length with geographical distance. **b,** Cumulative percentage of central ancestry in eastern and Nigeria-Cameroon chimpanzees. **c**, Cumulative percentage of eastern ancestry in central and Nigeria-Cameroon chimpanzees. **d**, Cumulative percentage of Nigeria-Cameroon ancestry in central and eastern chimpanzees. **e**, Percentage of bonobo ancestry present in different chimpanzee subspecies, separating central chimpanzees in two groups: northern (Campo_Ma'an, MtsdeCristal, Goualougo, Invindo, La Belgique) and southern group (Loango, Conkouati, Lope, Bateke).



Extended Data Figure 4 | Connectivity based on rare variation.
Blue color represents higher affinity and red color represents lesser affinity. Black dots are the field sites used as reference and the red dots, the tested population. **a**, Ngiri share rare variants with northwestern and central areas of the eastern chimpanzees, but not with nearby central chimpanzee populations. **b**, Rare allele connectivity in Goualougo shows affinity towards other northern central chimpanzee populations and Gashaka (Nigeria-Cameroon). **c**, Rare alleles in Gashaka show high connectivity toward other Nigeria-Cameroon sites, but also with Goualougo and Ngogo.



Extended Data Figure 5 | Novel variant discovery.

Black line indicates the variants discovered with 59 high coverage genomes from de Manuel et al., (2016) and the red line indicated the novel variant discovery with PanAf dataset. **a,** When looking at all genotyped positions we discover a total of 2,328,613 variants (207% more) on 469 individuals, with an almost linear increase in discovered positions, which does not continue a flattening trend of the high-coverage samples. **b**, After applying quality filtering on the discovered variation (Supplementary Information), we increase the variant discovery from 1,050,120 variants (on the 59 high coverage individuals) to 1,585,194 variants (with 415 PanAf individuals). This novel discovery rate of 50.9% is likely a reasonable representation of the extent of new variants found. **c,** When considering only high quality sites (Supplementary Information), we find an increase of 10.6% from 1,001,898 to 1,108,367 high quality sites for the 228 PanAf individuals, which is most likely an underestimate due to strict filtering.

Supplementary Information

1. Quality assessment

1.1 Endogenous DNA

To quantify the proportion of endogenous DNA (eDNA) in each sample we used two different methods: qPCR and shallow shotgun sequencing. Both methods return different results, with qPCR under-estimating the eDNA, due to the possible presence of PCR inhibitors. Therefore, we did not pool samples together from which the eDNA estimates were derived from different methods. The eDNA estimates range from 0% to 47.57% with a median of 3.05% (Fig. S1).

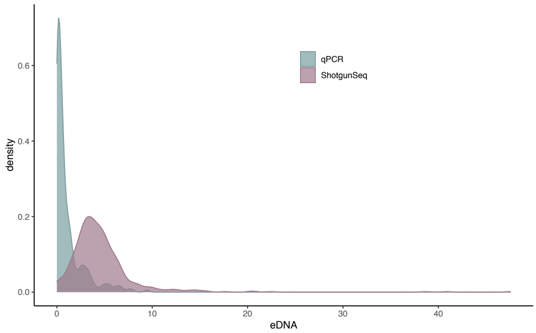


Figure S1. Distribution of eDNA using two methods: qPCR and Shotgun sequencing.

1.2 PCA

Principal component analysis (PCA) of all samples (N=828) was done with PCAngsd [Citation error] after obtaining genotype likelihoods with ANGSD (Extended Data Fig. 1c). We included 4 representatives of each known chimpanzee subspecies from de Manuel et al. (2016) to ensure the fecal samples analyzed in this study recapitulate the known genomic diversity of chimpanzees. The first component (PC1) is dominated by a small number of samples which are separated from the known variation from whole-genome sequencing (de Manuel et al., 2016) (Extended Data Fig. 1c). Most likely, these samples have extremely high levels of contamination, possibly through diet, or fecal samples from other primate species mistakenly collected for this study.

We performed the PCA only with fecal samples and defined a threshold to keep only those samples representative of chimpanzee diversity (Extended Data Fig. 1d). This threshold was set at -0.01 of the initial PCA, in order to retain the main two clusters of samples, while being permissive at this point (Fig. S2).

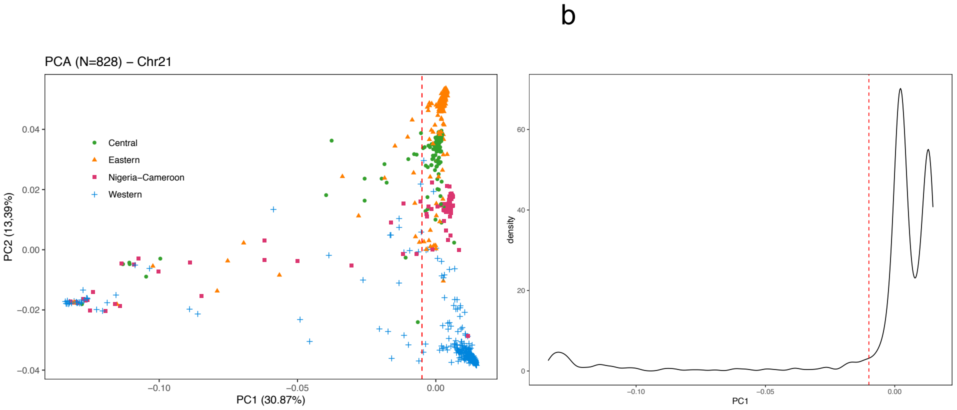


Figure S2. PCA obtained from only fecal samples with 1,430,461 markers, threshold is set at -0.01 PC1 to exclude the potentially contaminated samples. PCA of all samples (A), distribution of samples at PC1 (B).

The resulting PCA without 100 outlier samples results in the clustering of the four chimpanzee subspecies, also confirmed by adding 4 samples of each know subspecies from de Manuel et al. (2016) (Extended Data Fig. 1d). We do not observe clusters separating from the four known subspecies, while a number of samples is tending towards the center (0,0). These samples have significantly lower coverage than those falling close to the four major clusters (Fig. S3), but do not belong to specific sites, suggesting that this is due to missing data, rather than a genetic gradient.

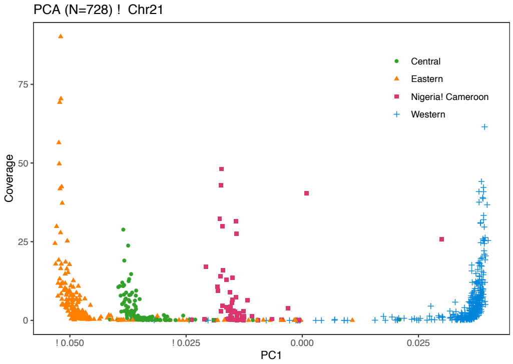


Figure S3. Coverage at target space with PC1. Colors represent each subspecies.

Then we kept only samples with coverage > 0.5-fold (N=556) and removed one Nigeria-Cameroon chimpanzee sample from Korup (Kor1-35) that clustered within western chimpanzee diversity. Samples (N=555) belonging to the four known subspecies clearly cluster together, supporting previous evidence for four distinct chimpanzee populations in the wild (de Manuel et al., 2016, Prado et al., 2013) (Extended Data Fig. 1e).

1.4 Sources of contamination

a) Primate contamination

According to Extended Data Fig. 1c, it is likely that a minority of samples (N=100) had significant amounts of primate contamination, either being samples from a different primate, or chimpanzee samples with some degree of contamination. Since other primates are closely related to chimpanzees and humans, we used the BBsplit (https://github.com/BioInfoTools/BBMap) software to competitively map sequencing reads obtained from each the sample to a range of other primate genomes, and thus obtain a summary of unambiguous mappings to each genome. We used the BAM files without duplicates mapped to the human genome (Hg19), converted them into FASTQ and then mapped them to these primate genomes. We restricted the analysis to only primates since we are only looking at the endogenous portion of the DNA that was already mapped to the human genome. The primate assemblies used were: chimpanzee (panTro6), human (hg19), olive baboon (Panu_3.0), green monkey (Chlorocebus_sabeus_1.1), Angolan colobus (Cang.pa_1.0), sooty mangabey (Caty_1.0), gorilla (gorGor4), drill (Mleu.le_1.0), Patas monkey (EryPat_v1_BIUU), Da Brazza monkey (CertNeg_v1_BIUU) and mandril (mandrill_1.0). We used the proportion of unambiguously mapping reads as a proxy to determine whether other primates than chimpanzees are more likely to be the source of the endogenous DNA, not to definitively determine these primate species, since we limited this test to only 11 primate assemblies.

We determined that the majority of samples removed following PCA criteria (Fig. S4) contained a higher proportion of reads mapping to other primate genomes compared to samples retained with that filtering (Fig. S5).

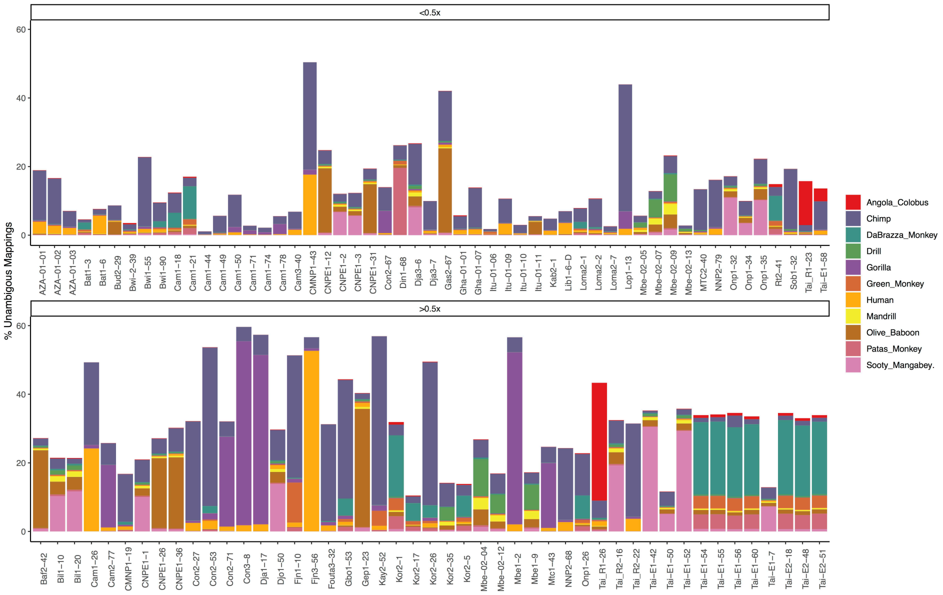


Figure S4. Proportion of unambiguous mappings to many primate species for samples that have been excluded by PCA filtering (N=100) at different coverage (higher or lower than than 0.5x coverage).

Some samples have the majority of their reads mapping to the gorilla genome. Most probably, these samples are misidentifications at the moment of collection, mistakenly taking a gorilla sample for a chimpanzee sample. We also find samples with a high proportion of olive baboon reads from sites within the geographic range of this species. Two samples from Tai-Eco may belong to sooty mangabey, a primate species that inhabits western Africa and is specifically present at Tai Forest. Many other samples from Tai-Eco seem to be from an unidentified primate species. Surprisingly, one sample, Fjn3-56, is probably a human feces. Other samples had some degree of primate contamination with the majority of reads still mapping to the chimpanzee genome. This might be explained by diet, considering that the chimpanzee diet includes other primate species, and thus DNA of the diet can survive the intestinal tract and be found in the feces (Hernandez-Rodriguez *et al.*, 2018).

On the other hand, for samples kept after initial PCA filtering, the majority of reads were mapping to chimpanzee genome (Fig. S5), although some low levels of reads mapping to the human genome were shared across all samples.

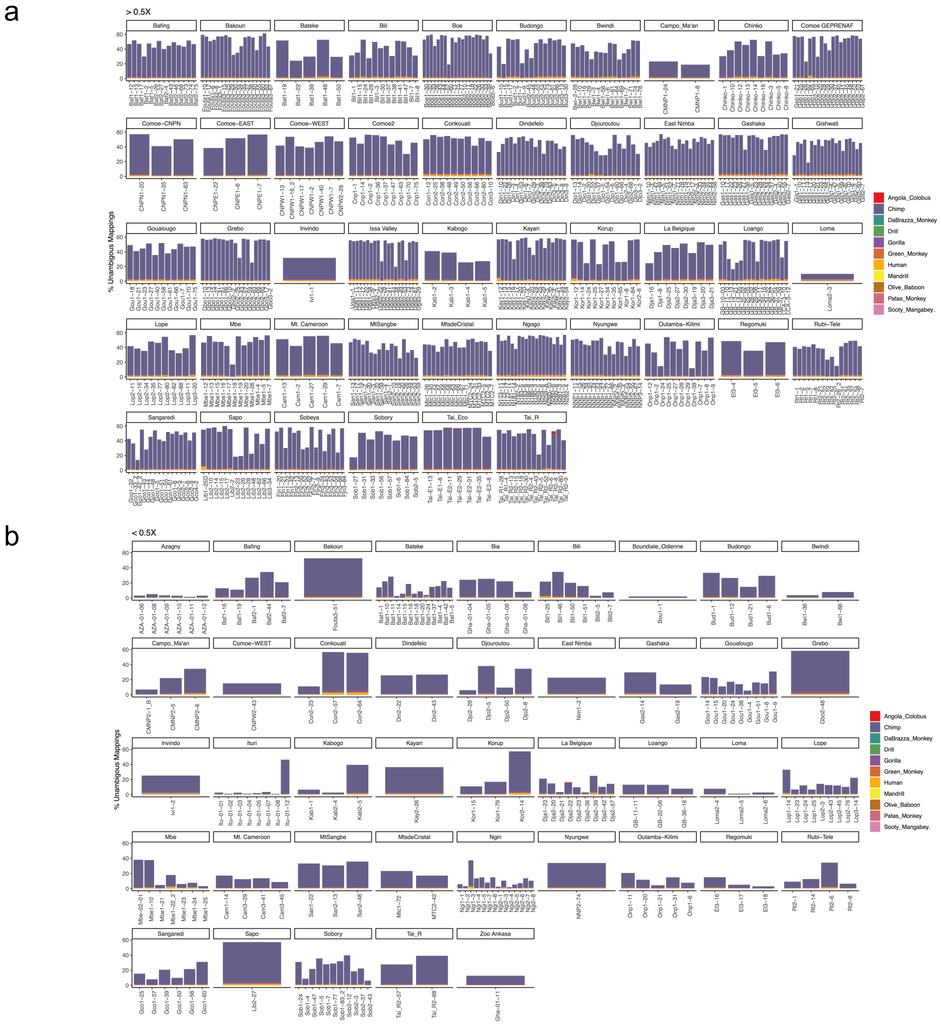


Figure S5. Proportion of unambiguous mappings to many primate species for samples that have been kept by PCA filtering (N=728) at different coverage cutoff (more (a) or less (b) than 0.5x coverage).

Since it is difficult to disentangle human contamination from mapping bias, we decided to apply another method to be more rigorous and remove from the dataset those samples with higher levels of human contamination.

b) Human contamination

Human contamination can occur at different stages, during sample collection, laboratory procedures and sequencing. We devised a human contamination test by using positions where modern humans and chimpanzees consistently differ, as described previously (Fontsere et al., 2020, under review). Using samtools mpileup (Li *et al.*, 2009), we retrieved the number of observations of human-like and chimpanzee-like alleles at these positions, considering the fraction of observations for the human-like allele across all positions as an estimate for possible human contamination. Within the working dataset that passed previous filters of primate contamination and coverage >0.5x (N=555) we found that 17 samples had more than 1% human contamination, which were removed from further analysis, unless stated otherwise. For more refined analyses, we kept a smaller dataset with more than 5-fold coverage. In this case we also reduced the threshold of allowed human contamination to less than 0.5%. In that case, 10 samples had more than 0.5% human contamination and were excluded from this dataset.

1.5 Relatedness

Fecal samples used in this study had been previously genotyped with microsatellites to discard those samples that belong to the same individual or were 1st order relatives. Still, we tested the dataset to remove any related pair at 1st degree or more that may have remained. We obtained genotype likelihoods for each geographical site independently using ANGSD (Korneliussen, Albrechtsen and Nielsen, 2014), extracted allele frequencies at each site and ran NgsRelate (Korneliussen and Moltke, 2015). We decided to calculate relatedness at each site separately to avoid population structure bias within each subspecies, that would result in an overestimation of related pairs (Fig. S6).

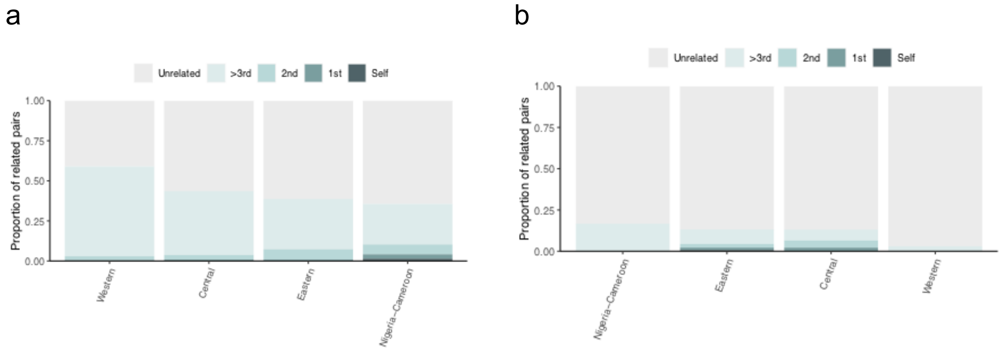


Figure S6. Proportion of related pairs of each subspecies, by calculating relatedness estimates with (A) all samples of each subspecies together or (b) by restricting the analysis by geographical site independently.

We consider unrelated individuals when their kinship coefficient has a value of 0, third degree or higher when it has a value between 0 and 0.0625, 2nd degree relatives when it fluctuates between 0.0625 and 0.1875 and 1st degree relatives when this coefficient has a value between 0.1875 and 0.375. Samples with a kinship coefficient higher than 0.375 are considered to be identical.

Whenever we encountered a pair with kinship coefficient > 0.1875, we kept the sample with the highest coverage. Out of 3581 total pairs analyzed at each site, only a small fraction (96 pairs) were first order or identical samples, with the majority being of unrelated pairs.

2. Novel variant discovery

In order to estimate the amount of variation discovered with new samples, we determined segregating sites (not fixed for the reference or alternative state) within the target space across individuals. Previous studies provided an overall picture of variation across the four chimpanzee subspecies (Prado-Martinez *et al.*, 2013; de Manuel *et al.*, 2016). We restrict the analysis to the 469 reliable individuals that do not show an excess of homozygous alternative alleles (ratio of alternative to reference calls < 0.013), that are not outliers in the PCA (Extended Data Fig. 1c), have less than 1% human contamination, and for which at least 10 million positions carry information. This results in a discovery of a total of 2,328,613 variants (207% more), with an almost linear increase in discovered positions, which does not continue a flattening trend of the high-coverage samples (Extended Data Fig. 5a). We conclude that errors and biases contribute largely to this trend, and apply further filtering to the data, as described below:

- "quality sites": Subset of these sites with a sequencing depth of at least 3 reads, less than 100 reads, and a genotype quality of more than 20.

- "high quality sites": Subset of these sites with a sequencing depth of at least 8 reads, less than 100 reads, a genotype quality of more than 40.

When using sites that fulfill "quality" criteria, we find 1,050,120 variants across the previously studied 59 individuals, and 1,585,194 variants when considering the 415 reliable individuals with at least 7.5 million observed sites additional to the conditions stated above, considering that less data is available per individual. This novel discovery rate of 50.9% is likely a reasonable representation of the extent of new variants found by sequencing this number of individuals (Extended Data Fig. 5b). The novel discovery rate per individual is flattening (0-0.16% for the last 20 individuals). We conclude that when sequencing this number of individuals, a saturation in the discovery of new variation is approached, although not fully reaching a plateau phase. However, only chromosome 21 is considered here, and due to the patchy distribution of the capture data many sites are not covered in all individuals.

Finally, when considering only "high quality" sites, we find an increase of 10.6% from 1,001,898 to 1,108,367 high quality sites for the 228 reliable individuals with at least

5 million sites (Extended Data Fig. 5c), which is most likely an underestimate due to strict filtering.

3. Effective Migration Surfaces (EEMS)

We applied EEMS (Petkova, Novembre and Stephens, 2016) to infer past patterns of migration between chimpanzee populations. This program calculates effective migration surfaces and provides a visualization of potential regions of higher-than-average and lower-than-average historical migration between those sites. Here, we used only samples with more than 5-fold coverage. The VCF file was filtered to keep only biallelic sites, at a minimum depth of 3, minimum genotype quality of 20, and allowing for 20% missing sites. We ran the program with all samples to obtain the overall measurement taking into consideration all subspecies. We obtained ten replicate runs of EEMS with the following parameters: nIndiv = 212, nSites = 1113936, nDemes = 1000, diploid = TRUE, numMCMCIter = 2000000, numBurnIter = 1000000, numThinIter = 9999.

When plotting the whole range of the chimpanzee distribution, we observed that there is more effective migration than average between western chimpanzee sites when compared to the other subspecies (Fig. 2c). We also observe significant barriers of gene flow between subspecies (Fig. 2c and S7).

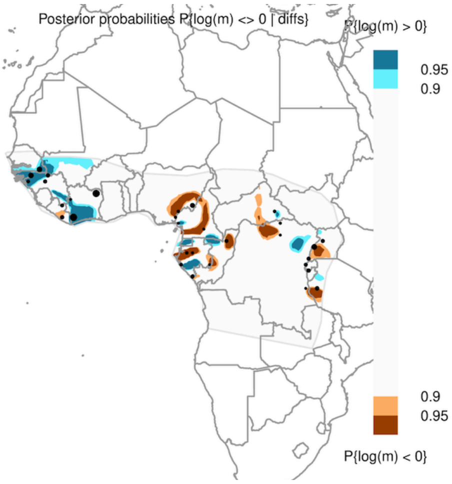
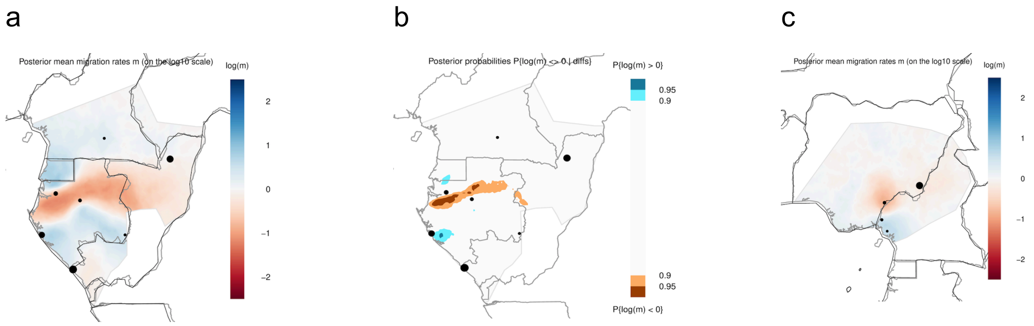


Figure S7. Posterior probabilities from EEMS between all samples and subspecies in the PanAf dataset.

There appears to be a clear barrier separating Nigeria-Cameroon and central chimpanzees, which is overlapping with the Sanaga River (Fig. 2a and S7). Between central and eastern chimpanzees we do not observe such a clear barrier, although the Goualougo site, close to Ubangi river, shows less migration towards eastern chimpanzees. We also observe another barrier at the west of Chinko, Bili and Rubi-Tele eastern chimpanzee sites that may represent low levels of migration of the Nigeria-Cameroon and central chimpanzees with eastern chimpanzee subspecies.

Some areas appear more isolated, while others seem more connected within each subspecies: In central chimpanzees (nIndiv = 25, nSites = 1067355, nDemes = 500), there is a significant barrier which overlaps with the Ogooué River crossing Gabon, separating MtdeCristal and La Belgique from Lope, Loango, Conkouati and Bateke, with both groups having more connectivity within their respective communities (Fig. S8a,b). Mbe in Nigeria-Cameroon chimpanzees (nIndiv = 17, nSites = 1258968, nDemes = 500) appears to be relatively isolated from the rest (Fig. S8c,d). For eastern chimpanzee communities (nIndiv = 60, nSites = 1098162, nDemes = 600), Bili and Chinko are more connected between them, while Issa Valley seems to have been more strongly isolated from the rest within the timeframe considered by EEMS (Fig. S8e,f).

In western chimpanzee sites (nIndiv = 109, nSites = 1121734, nDemes = 600), we observe a rather high connectivity across their range in comparison to non-western chimpanzees, with no site being significantly more isolated than others. Our data suggests that corridors of genetic connectivity existed between Boe and Sangaredi, but also between Kayan, Dindefelo and Sobory, and finally between MtSangbe and the southern clade of Tai Eco, Tai R and Grebo sites (Fig. S8g,h).


a
Posterior mean migration rates m (on the log10 scale)
log(m)
2
1
0
−1
−2
b
Posterior probabilities P{log(m) <> 0 | diffs}
P{log(m) > 0}
0.95
0.9
0.9
0.95
P{log(m) < 0}
c
Posterior mean migration rates m (on the log10 scale)
log(m)
2
1
0
−1
−2

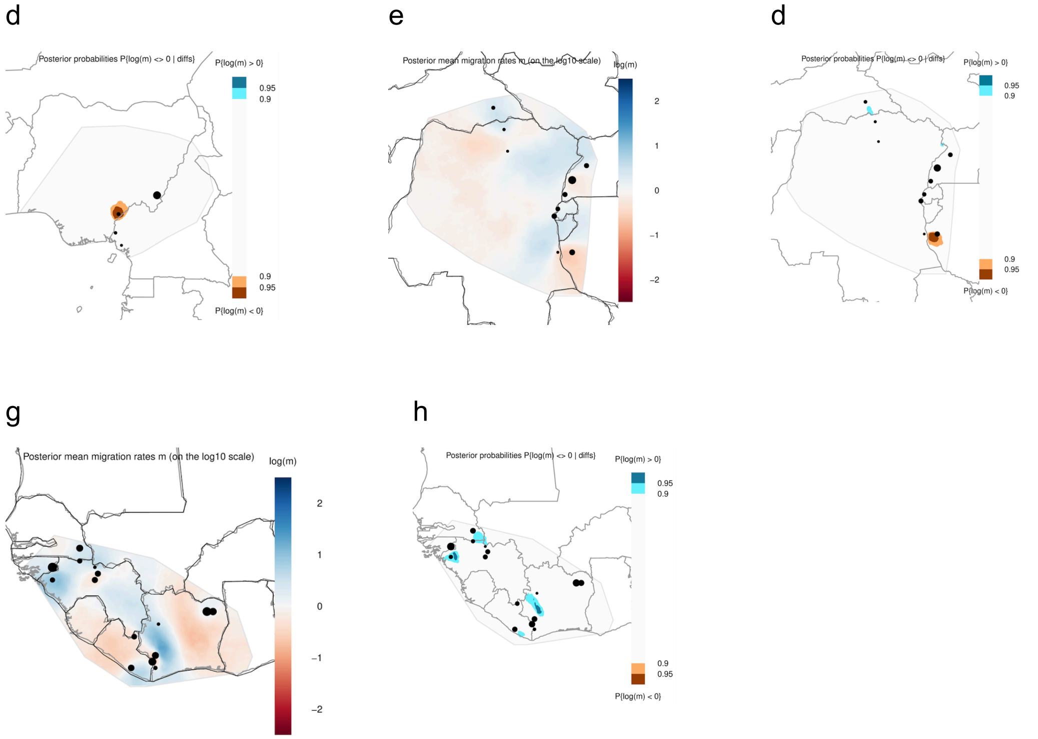

d
Posterior probabilities P{log(m) <> 0 | diffs}
P{log(m) > 0}
0.95
0.9
0.9
0.95
P{log(m) < 0}
e
Posterior mean migration rates m (on the log10 scale)
log(m)
2
1
0
−1
−2
d
Posterior probabilities P{log(m) <> 0 | diffs}
P{log(m) > 0}
0.95
0.9
0.9
0.95
P{log(m) < 0}
g
Posterior mean migration rates m (on the log10 scale)
log(m)
2
1
0
−1
−2
h
Posterior probabilities P{log(m) <> 0 | diffs}
P{log(m) > 0}
0.95
0.9
0.9
0.95
P{log(m) < 0}

Figure S8. EEMS per subspecies: **a & b,** central chimpanzee; **c & d,** Nigeria-Cameroon chimpanzee; **e & f**, eastern chimpanzee; and **g & h**, western chimpanzee. a, c, e and g are posterior mean migration rate; and b, d, f and h are posterior probabilities.

4. Fragments of shared ancestry

4.1 Detection of Identical-by-descent (IBD)

In order to identify Identical-By-Descent (IBD) segments between individuals within and between sites, we used IBDseq (Browning and Browning, 2013), since this program does not require phasing of the data. We applied this method to the dataset of samples with more than 5-fold coverage but including pairs of related samples. To increase the sample since and thus the power for IBD detection, we merged this present dataset with data on chromosome 21 from 59 whole genomes (de Manuel *et al.*, 2016). We removed indels and kept only biallelic positions with minimum genotype quality of 20, minor allele frequency of 0.01 and excluded variants with a missingness of more than 0.6.

We explored the effect of using genotypes at two different thresholds of minimum depth: 3 or 8 reads, since the accuracy in detecting true heterozygous positions likely influences the power to detect IBDs. We assessed the impact of the depth of coverage on mean heterozygosity, considering global missingness and coverage. Mean heterozygosity and missingness were computed using plink (Purcell *et al.*, 2007) parameters -missing and -het, respectively. At a minimum depth of 3 reads at each genotype, the mean heterozygosity expected by subspecies (taking whole genomes as benchmark) is hardly maintained in samples when the mean coverage is low and the genotype missingness high. This implies that at lower coverage, we lose power to detect heterozygous calls, which at the same time would decrease our sensitivity to detect IBDs. On the other hand, to increase the power to detect heterozygous calls, we increased the threshold of minimum depth of 8 reads at each genotype. As a consequence, while the mean heterozygosity resembles the values of the whole-genome dataset, the missingness is increased (Fig. S9).

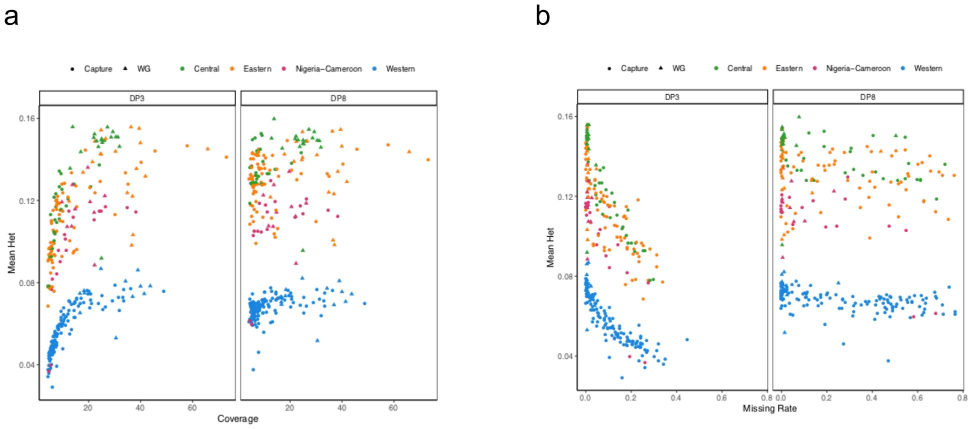


Figure S9. Impact of depth of coverage at each genotype (3 or 8 reads) on mean heterozygosity in the PanAf dataset (240 samples) and whole genome dataset from de Manuel 2016 (59 samples). **a,** Mean heterozygosity vs. mean coverage and **b,** Mean heterozygosity vs. missing rate.

The IBDseq software was applied to all samples (PanAf and Whole-genome samples) with standard parameters (ibdlod=3.0, ibdtrim=0.3, r2window=500, r2max=0.15), but exploring a range of errorProp and errorMax values. Since the Panaf dataset is of rather low coverage, and although we limited the discovery of SNPs with a minimum of 8 reads, any misassigned genotype may cause a break of any hypothetical IBD between two samples. Instead of using the standard parameters (errorProp=0.25 and errorMax=0.001), we tested 5 different errorProp values (0.1, 0.25, 0.5, 0.6 and 0.8) and 12 errorMax values (0.00025, 0.001, 0.004, 0.008, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 and 0.8) on samples previously determined to be from 14 identical individuals among the PanAf samples, processed and sequenced independently, and with of sufficient coverage (more than 5-fold). Theoretically, the whole chromosome 21 should be detected as one perfect IBD segment. We find that the empirical thresholds yielding a maximized length for all identical pairs were 0.1 for errorMax and 0.5 for errorProp.

Since we have different numbers of samples per site, we have summarized the IBD detected by computing the number of IBD segments per pair of samples and also the average length of pairs of samples between sites. Also, sites that are in close proximity (within 15 km) were merged into a single site, as described above (Tai_Eco and Tai_R; Sobory and Bakoun; Comoe Geprenaf, Comoe-WEST, Comoe-East, Comoe2 and Comoe-CNPN).

5. Rare alleles

5.1 Patterns of rare allele sharing

Rare alleles have been a useful resource for studying human populations (Gravel et al., 2011), since these variants were arising during the past few 100 or 1,000 years. Although this study provides the largest dataset of variation in chimpanzees so far, there are limitations to the sample size compared to human studies for defining rareness, and to the data quality (low coverage, DNA degradation, single chromosome) that prevent the use of existing methods exploiting rare variants for ancestry estimation, like rarecoal (Schiffels et al., 2016). However, here we approached a noise-tolerant strategy for detecting variants which are almost location-specific. These should represent the recent variation emerging locally in chimpanzee groups which were only loosely connected after their split, as suggested by the patterns of differentiation in the PCA analysis and other measures of connectivity.

We defined a reference panel of sufficient quality, using only individuals with more than 1-fold coverage in the target space, less than 0.5% human contamination, which were not detected as non-chimpanzee samples, as outliers in the PCA (see above), determined as belonging to another population in the PCA (Kor1-35, CMNP1-24, Uga2-81), or as carrying an excess of heterozygous alleles (Gas1-10). This yields a geographically distributed reference set of 449 individuals across 38 sites (on average 12 individuals per site), with the Comoe, Tai and Bakoun-Sobory sites each being merged. We used only bi-allelic on-target quality-filtered sites for these individuals and calculated the site-specific alternative allele frequency for each variant. We then defined near-private variants per site (instead of per individual), where a given variant had to be observed a) at a frequency larger than zero at one site, b) at a frequency of less than 0.5 when calculating the sum of frequencies across all other sites (for example, the variant may be observed at a frequency of 0.2 at two different sites, or 0.4 at only one site), c) with data for at least two sites (no NA values). We find a total of 985,906 of such variants, with on average 27,653 variants per site.

This panel of near-private variants can be used to estimate the matching of new individuals, even when sequenced at shallow depth. In order to do this, we use the quality genotype calls for each individual, and overlap these with the near-private variants. Alternative alleles in heterozygous and homozygous states were counted as derived alleles. For each comparison, we counted the number of positions that were near-private in a given reference population and carried information in the test sample, and the subset of these positions that carried the alternative allele in the test sample. We then calculated the proportion of shared near-private sites of all observed near-private sites for each reference population as the summary statistic of shared rare alleles. This strategy has the advantage that a large fraction of false alternative alleles in the test sample would not cause an enrichment of shared alternative alleles with any reference population.

Based on the observation that near-private allele sharing is high between the different western chimpanzee populations, it seems possible that regional substructure within the range of this subspecies could be used to determine a region of likely origin, rather than relying on the best match only. The dense sampling scheme in this project allows the use of the known geographic coordinates of the sites for an explicit spatial model of these matching scores using the R packages sf (Pebesma, 2018), sp (Bivand, Pebesma and Gómez-Rubio, 2013), and maptools (Bivand and Lewin-Koh, 2013). We used the known current distribution of chimpanzees according to the IUCN, and expanded this range by 1.5° in all directions to create a spatial grid in which to estimate the geographic areas. We then used the geographic coordinates of the known locations to create a simple variogram of the formula (match score) ~ X + Y, using the R package gstat (Gräler, Pebesma and Heuvelink, 2016), and fit the variogram using a spherical model and standard parameters, but without fitting ranges. Finally, we performed kriging to the expanded geographic range using the krige function in gstat. The focal point was estimated as the highest point within the surface, in order to calculate the distance to the correct origin of samples. Examples of this match score surface are shown in Fig. 4a and Extended Data Figure 4.

5.2 Connectivity based on rare variants

Since the rare variants used in our approach are not necessarily fixed at a given location, but can be present at other locations, the pattern of shared rare variants is informative for past connectivity. We calculated the proportion of derived variants in a given population shared with all other populations. We then used these data points to infer a landscape of sharing with other populations, and applied the kriging procedure described above, where we left out the test population from the landscape (see examples in Extended Data Fig. 4).

5.3 Inference of origin of confiscated chimpanzees

We tested the procedure described above on other chimpanzee samples. We sequenced at low coverage 20 chimpanzees from two rescue centers in Spain: Centro de Rescate de Primates Rainfer (http://rainfer.org) and Fundació Mona (https://fundacionmona.org/). We obtained the overlap of each individual with the near-private alleles from each location, calculated the proportion of matching variants, and applied the spatial model fitting.

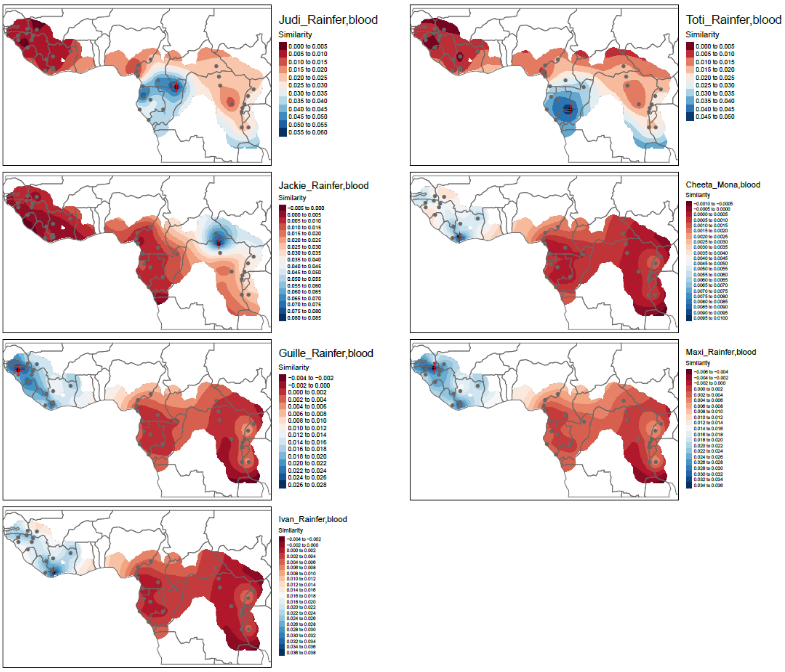


Figure 10. Spatial matching of 7 chimpanzee blood samples from sanctuaries.

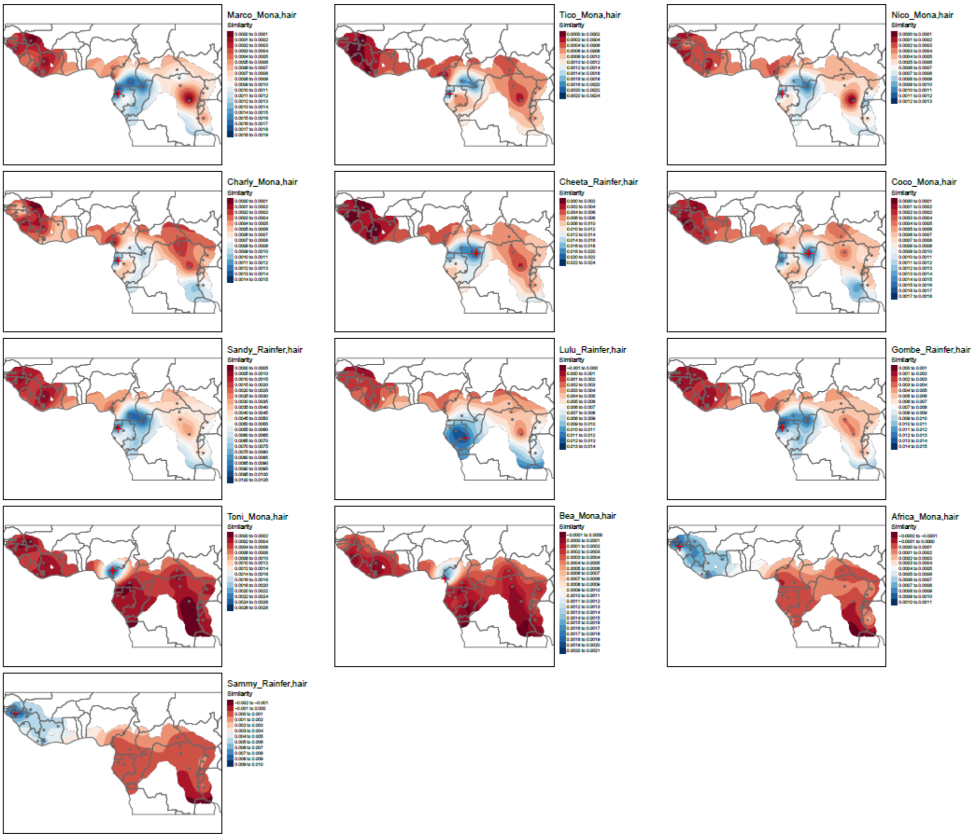


Figure S11. Spatial matching of 13 chimpanzee hair samples from sanctuaries.

References

Bivand, R. S., Pebesma, E. and Gómez-Rubio, V. (2013) “Classes for Spatial Data in R,” in *Applied Spatial Data Analysis with R*. New York, NY: Springer New York, pp. 21–57. doi: 10.1007/978-1-4614-7618-4_2.

Bivand, R and Lewin-Koh, N (2013) "maptools: Tools for reading and handling spatial objects," R package version 0.8.

Browning, B. L. and Browning, S. R. (2013) “Detecting identity by descent and estimating genotype error rates in sequence data,” *American journal of human genetics*, 93(5), pp. 840–851. doi: 10.1016/j.ajhg.2013.09.014.

Fontsere, C *et al.* (2020 Under Review) "Maximizing the acquisition of unique reads in non-invasive capture sequencing experiments," *Molecular Ecology Resources*

Gräler, B (2016) "Spatio-Temporal Interpolation using gstat," The R Journal (2016) 8:1, pages 204-218

Hernandez-Rodriguez, J. *et al.* (2018) “The impact of endogenous content, replicates and pooling on genome capture from faecal samples,” *Molecular ecology resources*, 18(2), pp. 319–333. doi: 10.1111/1755-0998.12728.

Korneliussen, T. S., Albrechtsen, A. and Nielsen, R. (2014) “ANGSD: Analysis of Next Generation Sequencing Data,” *BMC bioinformatics*, 15, p. 356. doi: 10.1186/s12859-014-0356-4.

Korneliussen, T. S. and Moltke, I. (2015) “NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data,” *Bioinformatics*, p. btv509. doi: 10.1093/bioinformatics/btv509.

Li, H. *et al.* (2009) “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

de Manuel, M. *et al.* (2016) “Chimpanzee genomic diversity reveals ancient admixture with bonobos,” *Science*, 354(6311), pp. 477–481. doi: 10.1126/science.aag2602.

Pebesma, E. (2018) “Simple features for R: Standardized support for spatial vector data,” *The R journal*. The R Foundation, 10(1), p. 439. doi: 10.32614/rj-2018-009.

Petkova, D., Novembre, J. and Stephens, M. (2016) “Visualizing spatial population structure with estimated effective migration surfaces,” *Nature Genetics*, 48(1), pp. 94–100. doi: 10.1038/ng.3464.

Prado-Martinez, J. *et al.* (2013) “Great ape genetic diversity and population history,” *Nature*, 499(7459), pp. 471–475. doi: 10.1038/nature12228.

Purcell, S. *et al.* (2007) “PLINK: a tool set for whole-genome association and population-based linkage analyses,” *American journal of human genetics*, 81(3), pp. 559–575. doi: 10.1086/519795.

4. Discussion

Ever since the sequencing of the first human genome, genomic resources for non-model organisms have steadily risen. Such is the case for great apes. As reviewed in section 1.3., the availability of multiple great ape whole genomes has been key to elucidate their population history and demography, which can be relevant for their conservation. Clearly, Prado-Martinez *et al.* (2013) and de Manuel *et al.* (2016) studies on great ape and chimpanzee genomics have set the groundwork for the development of this thesis. Here, we have gone a step further in the study of chimpanzee populations. We have taken advantage of the available genomic resources to design target capture arrays and avoid performing further whole-genome sequencing. First, this strategy has allowed us to reduce sequencing costs while extracting genome-wide data for the captive chimpanzee population. Second, the application of target capture methods to enrich a small proportion of the genome has been essential to retrieve genomic information out of fecal samples. In that regard, we have thoroughly explored strategies to improve data recovery from fecal samples and have designed an array that targets the chromosome 21.

In the next sections I will comment on the findings, recommendations and contributions that resulted from the work of this thesis, and also, their implications for biodiversity conservation.

4.1. Invasive or non-invasive samples

Through this thesis there has been a constant, the mention of invasive and non-invasive samples as a source of DNA for genomic studies.

The first project of this thesis (chapter 3.1) was mainly based on the use of blood samples. After being very compelling on how non-invasive samples can be successfully used to extract meaningful genomic data, why did we use invasive samples? Mainly because we were analyzing captive chimpanzee individuals from zoos and sanctuaries. In this context it is feasible to extract blood or tissue samples, for instance during a veterinarian check-up of the animals. If blood or tissue samples are available, they should always be the first choice. In fact, the accessibility to blood samples has been fundamental for the generation of genomic resources such as reference assemblies and whole-genome resequencing studies. But, as mentioned in section 1.4 and 1.5. of this thesis, blood samples cannot be obtained from wild endangered populations. Hence, the focus turns to the collection of NI samples. They have been successfully used as a source of DNA in many conservation genetics initiatives and are slowly gaining relevance in genomic studies, ever since the implementation of target enrichment methods (section 1.6). Otherwise, retrieving enough sequencing data from NI samples with very low levels of endogenous content, would require a substantial amount of sequencing. But most importantly, even if money were not a limiting factor, these libraries have low molecular complexity, which would hamper recovery of unique reads even with deep sequencing.

Extracting genomic data out of fecal samples with very low endogenous DNA content is not an easy task. Careful assessment of the quality of fecal samples and the technical approaches to follow should be done before moving into large-scale genomic studies. In that regard, the existence of technical studies on target capture of chimpanzee fecal samples has set the basis for the application of these types of samples in genomic research (Hernandez-Rodriguez *et al.*, 2018; White *et al.*, 2019). However, there were some aspects not fully resolved. To better comprehend target capture performance on fecal samples we devised the study presented in chapter 3.2. We aimed to examine the variation of eDNA content between different field sites and how the data recovery of samples could be increased in samples with very low proportions of eDNA (<1%). First of all, one critical aspect for the success of a fecal sample capture experiment is the eDNA content present in the feces (Hernandez-Rodriguez *et al.*, 2018). Therefore, the proportion of eDNA should guide the library pooling for capture experiments. Our recommended pooling strategy follows two criteria: first, samples belonging to the same pool should have similar eDNA proportions, and second, each sample within a pool should contribute the same total amount of eDNA, creating an equi-endogenous pool. Hence, a proper assessment of endogenous DNA content should be done prior to pooling and capture. Although in chapter 3.2 we estimated eDNA with qPCR (Morin *et al.*, 2001), we recommend using shallow shotgun sequencing in the case of large-scale genomic projects. qPCR estimates can easily fluctuate due to the presence of PCR inhibitors and represent a laborious task. Since sequencing costs have reduced significantly, when having enough samples, shotgun sequencing for eDNA estimation would be faster, more accurate and cheaper. In this project, we captured the exomes of 60 chimpanzee fecal samples from 14 different field sites in Africa and characterized a way to increase molecular complexity without the need to perform multiple extractions and libraries. Our results suggest that performing multiple additional hybridizations from the same library is a good method to maximize library complexity. We also propose to lower the number of samples pooled when the eDNA content is very low. Finally, we recommend not sequencing the captured libraries very deeply to avoid exhausting the already very low library complexity. All these findings, together with the recommendations provided by Hernandez-Rodriguez *et al.*, (2018) and White *et al.*, (2019), were applied in chapter 3.3. and will serve as guidelines to the community for the genomic studies from NI samples.

4.2. Captive populations

Captive management of endangered populations aims to maintain a self-sustaining population representative of the wild counterpart. In the critical situation of biodiversity loss, captive populations may represent the last reservoirs of genetic diversity for some species. Hence, genetic assessment of the existing populations should be undertaken to guide captive management of the species. This has been the aim of chapter 3.1., where we explored a direct application of the available genomic datasets for the management of the zoo chimpanzee population. We implemented a target capture strategy to sequence a panel of ancestry informative SNPs. We wanted to devise a strategy that would retrieve meaningful information about subspecies ancestry and inbreeding coefficients without the need to perform whole-genome sequencing, and that could work in invasive samples (blood) and in non-invasive samples (hair). We assigned the ancestry of 167 chimpanzees from European zoos and determined their inbreeding coefficients. The majority of tested individuals belonged to the western chimpanzee subspecies, but we also identified admixed individuals, which will be eventually excluded from captive breeding plans. Our findings reinforce the idea that sampling for genomic studies should be done from wild populations or at least from wild-born individuals, since captive-born individuals can be a result of admixture that does not represent the wild counterpart.

In this project, we also used the available chimpanzee geogenetic map (de Manuel *et al.*, 2016) to infer the origin of 31 rescued chimpanzees, most of them located in the eastern chimpanzee range. At that point and given the lack of resolution of the available geogenetic map, we could not determine the location of the Nigeria-Cameroon and western chimpanzee samples. Certainly, there was the need for a much more fine-grained map. We explored this idea in the final project of this thesis (chapter 3.3.).

4.3. Wild populations

As mentioned before, there was a lack of fine-scale resolution to develop conservation tools and identify the origin of chimpanzees. Therefore, in chapter 3.3., we aimed to overcome this issue by constructing the largest to date NI geolocalized catalogue of genomic diversity from the whole extant range of chimpanzees. For this effort, we collected a total of 828 fecal samples from 51 field sites in Africa. It is worth noting that such an extensive sampling has only been possible thanks to the collaboration of many institutions under the *PanAfrican Programme: The Cultured Chimpanzee*.

We decided to capture chromosome 21 of all fecal samples to reduce the proportion of the genome to be sequenced. This strategy was devised to be economically viable and at the same time retrieve continuous genomic sequences avoiding the ascertainment bias caused by a panel of SNPs in an array (such as the one from chapter 3.1.). With this approach and by sequencing a single chromosome, we were able to discover ~50% of new genetic diversity and, most importantly, link it to specific geographical locations. In that regard, we validated that geographic origin is a good predictor of genetic structure. Population stratification in chimpanzee populations was determined to be largely consistent with isolation-by-distance, but known geographical barriers such as the Sanaga, Ogooué and Ubangi rivers and the Lake Tanganyika caused a reduction of genetic exchange. We found strong evidence supporting the four known clusters of chimpanzee populations, in particular between geographically close groups of central and eastern chimpanzees. A big sampling gap between both populations had formerly hindered the determination of whether they were two distinctly separated subspecies or rather a cline of variation under isolation-by-distance (de Manuel *et al.*, 2016). Furthermore, we could identify past corridors of gene-flow between non-western chimpanzees and link them to specific locations.

With the high-density sampling scheme in the PanAf dataset, we were able to build a connectivity map between populations. We detected possible recent expansions and increased connectivity in the western chimpanzees and uncovered recently isolated populations. This evidence could contribute to understanding recent demographic events in particular chimpanzee populations.

Finally, in this project we also developed a resource for chimpanzee geolocalization based on the presence of rare allele variation particular to each population. With our approach, we were able to locate samples to around 100 km from their true origin and provided a proof of concept of how genomics can be applied to fighting illegal trade of chimpanzees.

4.4. Perspectives

In this thesis, I have reviewed the critical situation of chimpanzees and other great apes facing extinction. But these charismatic species are not alone in this fate. We, humans, have modified the natural habitat to such an extent (Garcia *et al.*, 2020) that we have induced the Sixth Mass Extinction (Barnosky *et al.*, 2011; Ceballos *et al.*, 2015). Monitoring of biodiversity losses in vertebrates shows that between 1970 and 2016 vertebrate population sizes have dropped by 68% according to the Living Planet Index Report of 2020. It is clear that actions need to be taken urgently to bend that curve and restore as much biodiversity as possible. Limiting global warming should be the starting point (Warren *et al.*, 2018) and will imply a global change towards a more sustainable human economy growth. Also, land protection and restoration should become a priority to stop biodiversity loss and switch the current trend (Leclère *et al.*, 2020). It is clear that we are facing a massive challenge. A challenge that not only involves conservation practitioners but that should also draw attention to the whole society.

If we step into conservation specific actions, genetics and recently genomics can be extremely useful to delineate populations and better understand the species we want to preserve. Also, genomics can provide estimates of population changes through time, assess levels of genetic diversity and understand the consequences of population reduction. Then, this knowledge can be applied to guide management plans either in captivity or in the wild, and aid enforcement of conservation policies (Allendorf, Hohenlohe and

Luikart, 2010; Frankham, 2010; Shafer *et al.*, 2015). In recent years, genomic studies in endangered species have increased substantially. For instance, there are population genomic studies uncovering the genetic signatures of population reduction and isolation and their potential fitness consequences. Some examples are the Iberian (Abascal *et al.*, 2016) and Eurasian lynx (Lucena-Perez *et al.*, 2020), the grey wolf (Gómez-Sánchez *et al.*, 2018), the crested ibis (Feng *et al.*, 2019), the white rhinoceros (Sánchez-Barreiro *et al.*, 2020) and the mountain gorilla (Xue *et al.*, 2015). This is a far from complete list of all genomic resources being generated on endangered populations. I am sure that in the next few years, we will have many more studies elucidating the population trajectories and genetic diversity of endangered species.

In the three projects of this thesis, we have applied genomics to characterize chimpanzee populations and provided insights into how this could be translated into conservation actions, either in captivity or in the wild.

We have described the genetic ancestry and inbreeding levels of the European captive population to better guide conservation breeding programmes with the aim to maintain a self-sustainable population. That would be important in case it comes to a point where *in situ* conservation plans need supplementation from captive populations. However, this option is not currently feasible for chimpanzees that were born and lived in captivity for their whole life. Then, if reintroduction is not possible in the short term, should captive breeding programs be encouraged? The answer to this question has opened an intense debate, loaded with a great moral burden related to animal welfare. Zoos have provided us a vast range of resources to study wildlife, and they are nowadays involved in educational programs and *in situ* conservation actions. Also, there are initiatives to cryopreserve biological material from endangered species. The most famous one is the Frozen Zoo in San Diego (USA). Besides providing genetic material for research, they have stored the genetic material in an attempt to de-extinct species and recover lost genetic diversity for species that face clear risk of extinction. Although this seems like science fiction, they have already started the implementation of this strategy on the northern white rhinoceros (Hildebrandt *et al.*, 2018) and successfully cloned a przewalski horse, reviving genetic diversity from a 40-year-old cryopreserved sample. As exciting as this is, we don't need to forget that forest preservation is essential, otherwise there is going to be no habitat for these specimens to be reintroduced to. In my opinion, there are many actions that should be taken before reaching the point of extinction, and zoos should be the last resource. For instance, conservation action plans to fight against poaching and illegal pet trade are needed. In this thesis, we have devised a strategy to infer the origin of confiscated chimpanzees with a precision of around 100 km. These are the first steps to bridge the gap between the academic world and policy-practitioners. Still, the amount of resources invested to complete the project presented here are not affordable by many research groups and conservation agencies. To complete this project, we have needed state-of-the-art laboratories and massive storage, analytical, computational and financial resources. Therefore, for our geolocalization strategy to be implemented in real-world conservation action plans, it would require a simplification of the methodology. At the present time, our approach relies on performing shallow shotgun sequencing on each sample. Hence, a genomic laboratory facility where samples will be processed for library preparation and sequencing is required. Also, although next-generation sequencing costs have been significantly reduced, they still might not be affordable for many countries and institutions in a large-scale effort to fight poaching. I believe that this methodology should be available and affordable for the countries where chimpanzees naturally live, so that sample processing and analysis could be conducted avoiding dependency on European or North American laboratories.

In recent years, plenty of resources and research have been made available to describe many aspects of the chimpanzee genome and their evolutionary trajectory. However, the possibilities are not over yet. Besides quantifying genetic diversity and performing descriptive analysis of wild populations, genomics is also a powerful tool to understand functional aspects such as local adaptation and fitness, which should definitely be explored in detail following our fine-scale sampling scheme. Also, in this thesis we have *only* explored the extant genetic diversity in chimpanzees. It would be fascinating to recover DNA from samples that represent extinct populations. Unfortunately, there is a lack of chimpanzee fossils, but there are many historical samples placed in museums as a result of the European colonial times. Sampling these museum collections would entail a fantastic way to explore the recent past of chimpanzees, before the dramatic population declines in the past years. We may be able to recover lost genetic diversity and compare current estimates with the ones from around 100 years ago.

As I mentioned before, future research should focus on translating genomics findings into clear and practical conservation strategies, and for that new, cheaper and easier methodologies must be developed.

Finally, this thesis has demonstrated the utility of non-invasive samples for large-scale genomic studies. Our achievements are of great value for the scientific community, and we have set the grounds for its implementation in other great ape species.

Contributions to other publications

Orkin JD, Montague MJ, Tejada-Martinez D, de Manuel M, del Campo J, Di Fiore A, **Fontsere C** et al. Selection and local adaptation in capuchin monkeys revealed through fluorescence-activated cell sorting of feces (fecalFACS). bioRxiv. 2020. https://doi.org/10.1101/366112

Fontsere C, de Manuel M, Marques-Bonet T and Kuhlwilm M. Admixture in Mammals and How to Understand Its Functional Implications. BioEssays. 2019; https://doi.org/10.1002/bies.201900123

White LC, **Fontsere C**, Lizano E, Hughes DA, Angedakin S, Arandjelovic M, et al. A roadmap for high-throughput sequencing studies of wild animal populations using non-invasive samples and hybridization capture. Molecular Ecology Resources. 2019; https://doi.org/10.1111/1755-0998.12993

Solis-Moruno M, de Manuel M, Hernandez-Rodriguez J, **Fontsere C**, Gomara- Castaño A, Valsera-Naranjo C, et al. Potential damaging mutation in LRP5 from genome sequencing of the first reported chimpanzee with the Chiari malformation.Scientific Reports. 2017;7: 15224. https://doi.org/10.1038/s41598-017-15544-w

Bibliography

Abascal, F. *et al.* (2016) "Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx," *Genome biology*, 17(1), p. 251. doi: 10.1186/s13059-016-1090-1.

Allendorf, F. W. (2017) "Genetics and the conservation of natural populations: allozymes to genomes," *Molecular Ecology*, 26(2), pp. 420–430. doi: 10.1111/mec.13948.

Allendorf, F. W., Hohenlohe, P. A. and Luikart, G. (2010) "Genomics and the future of conservation genetics," *Nature Reviews Genetics*, 11(10), pp. 697–709. doi: 10.1038/nrg2844.

Allendorf, F. W., Luikart, G. H. and Aitken, S. N. (2012) *Conservation and the Genetics of Populations*. John Wiley & Sons.

Andrews, K. R. *et al.* (2016) "Harnessing the power of RADseq for ecological and evolutionary genomics," *Nature reviews Genetics*, 17(2), pp. 81–92. doi: 10.1038/nrg.2015.28.

Angeloni, F. *et al.* (2012) "Genomic toolboxes for conservation biologists," *Evolutionary Applications*, 5(2), pp. 130–143. doi: 10.1111/j.1752-4571.2011.00217.x.

Arandjelovic, M. *et al.* (2011) "Non-invasive genetic monitoring of wild central chimpanzees," *PloS one*, 6(3), p. e14761. doi: 10.1371/journal.pone.0014761.

Auton, A. *et al.* (2012) "A fine-scale chimpanzee genetic map from population sequencing," *Science*, 336(6078), pp. 193–198. doi: 10.1126/science.1216872.

AZA, Ape and TAG (eds.) (2010) *Chimpanzee (Pan troglodytes) Care Manual*. Association of Zoos and Aquariums, Silver Spring, MD.

Bailey, W. J. *et al.* (1992) "Reexamination of the African hominoid trichotomy with additional sequences from the primate beta-globin gene cluster," *Molecular phylogenetics and evolution*, 1(2), pp. 97–135. doi: 10.1016/1055-7903(92)90024-b.

Banes, G. L., Galdikas, B. M. F. and Vigilant, L. (2016) "Reintroduction of confiscated and displaced mammals risks outbreeding and introgression in natural populations, as evidenced by orang-utans of divergent subspecies," *Scientific reports*, 6, p. 22026. doi: 10.1038/srep22026.

Barnosky, A. D. *et al.* (2011) "Has the Earth's sixth mass extinction already arrived?," *Nature*, 471(7336), pp. 51–57. doi: 10.1038/nature09678.

Beck, B. *et al.* (eds.) (2007) *Best Practice Guidelines for the Re-introduction of Great Apes*. SSC Primate Specialist Group of the World Conservation Union, Gland, Switzerland.

Becquet, C. *et al.* (2007) "Genetic structure of chimpanzee populations," *PLoS genetics*, p. e66. doi: 10.1371/journal.pgen.0030066.eor.

Becquet, C. and Przeworski, M. (2007) "A new approach to estimate parameters of speciation models with application to apes," *Genome research*, 17(10), pp. 1505–1519. doi: 10.1101/gr.6409707.

Bentley, D. R. *et al.* (2008) "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, 456(7218), pp. 53–59. doi: 10.1038/nature07517.

Bermejo, M. *et al.* (2006) "Ebola outbreak killed 5000 gorillas," *Science*, 314(5805), p. 1564. doi: 10.1126/science.1133105.

Blomqvist, D. *et al.* (2010) "Trapped in the extinction vortex? Strong genetic effects in a declining vertebrate population," *BMC Evolutionary Biology*, 10(1), p. 33. doi: 10.1186/1471-2148-10-33.

Byers, O. *et al.* (2013) "The 'One Plan' Approach: The Philosophy and Implementation of CBSG's Approach to Integrated Species Conservation Planning," *WAZA Magazine*, 14, pp. 2–5.

Caldecott, J. and Miles, L. (eds.) (2005) *World Atlas of Great Apes and Their Conservation*. University of California Press, Berkeley, USA (Prepared at the UNEP World Conservation Monitoring Centre).

Campbell, N. A. and Reece, J. B. (2008) *Biology*. Pearson Benjamin Cummings.

Carlsen, F. and de Jongh, T. (2019) *European studbook for the chimpanzee Pan troglodytes*. Copenhagen Zoo.

Carpenter, M. L. *et al.* (2013) "Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries," *American journal of human genetics*, 93(5), pp. 852–864. doi: 10.1016/j.ajhg.2013.10.002.

Castellano, S. *et al.* (2014) "Patterns of coding variation in the complete exomes of three Neandertals," *Proceedings of the National Academy of Sciences of the United States of America*, 111(18), pp. 6666–6671. doi: 10.1073/pnas.1405138111.

Caswell, J. L. *et al.* (2008) "Analysis of chimpanzee history based on genome sequence alignments," *PLoS genetics*, 4(4), p. e1000057. doi: 10.1371/journal.pgen.1000057.

Ceballos, G. *et al.* (2015) "Accelerated modern human–induced species losses: Entering the sixth mass extinction," *Science Advances*, 1(5), p. e1400253. doi: 10.1126/sciadv.1400253.

Charlesworth, B. and Charlesworth, D. (1999) "The genetic basis of inbreeding depression," *Genetical Research*, 74(3), pp. 329–340. doi: 10.1017/s0016672399004152.

Check, E. (2002) "Priorities for genome sequencing leave macaques out in the cold," *Nature*, 417(6888), pp. 473–474. doi: 10.1038/417473a.

Chen, F. C. and Li, W. H. (2001) "Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees," *American journal of human genetics*, 68(2), pp. 444–456. doi: 10.1086/318206.

Chinwalla, A. *et al.* (2002) "Initial sequencing and comparative analysis of the mouse genome," *Nature*, 420(6915), pp. 520–562. doi: 10.1038/nature01262.

Collard, M. and Wood, B. (2000) "How reliable are human phylogenetic hypotheses?," *Proceedings of the National Academy of Sciences*, 97(9), pp. 5003–5006. doi: 10.1073/pnas.97.9.5003.

Cruz-Dávalos, D. I. *et al.* (2018) "In-solution Y-chromosome capture-enrichment on ancient DNA libraries," *BMC genomics*, 19(1), p. 608. doi: 10.1186/s12864-018-4945-x.

Devaux, C. A. *et al.* (2019) "Infectious Disease Risk Across the Growing Human-Non Human Primate Interface: A Review of the Evidence," *Frontiers in public health*, 7, p. 305. doi: 10.3389/fpubh.2019.00305.

Durbin, R. *et al.* (2010) "A map of human genome variation from population-scale sequencing," *Nature*, 467(7319), pp. 1061–1073. doi: 10.1038/nature09534.

Durvasula, A. and Sankararaman, S. (2020) "Recovering signals of ghost archaic introgression in African populations," *Science advances*, 6(7), p. eaax5097. doi: 10.1126/sciadv.aax5097.

Edroma, E., Rosen, N. and Miller, P. (eds.) (1997) *Conserving the Chimpanzees of Uganda. Population and Habitat Viability Assessment for Pan troglodytes schweinfurthii*. IUCN/SSC Conservation Breeding Specialist Group, Apple Valley, Minnesota.

Feng, S. *et al.* (2019) "The Genomic Footprints of the Fall and Recovery of the Crested Ibis," *Current biology*, 29(2), pp. 340-349.e7. doi: 10.1016/j.cub.2018.12.008.

Fischer, A. *et al.* (2004) "Evidence for a complex demographic history of chimpanzees," *Molecular biology and evolution*, 21(5), pp. 799–808. doi: 10.1093/molbev/msh083.

Fischer, A. *et al.* (2006) "Demographic History and Genetic Differentiation in Apes," *Current Biology*, 16(11), pp. 1133–1138. doi: 10.1016/j.cub.2006.04.033.

Flanagan, S. P. *et al.* (2018) "Guidelines for planning genomic assessment and monitoring of locally adaptive variation to inform species conservation," *Evolutionary Applications*, 11(7), pp. 1035–1052. doi: 10.1111/eva.12569.

Fontsere, C. *et al.* (2019) "Admixture in Mammals and How to Understand Its Functional Implications," *BioEssays*, 41(12), p. e1900123. doi: 10.1002/bies.201900123.

Frankel, O. H. (1974) "Genetic conservation: our evolutionary responsibility," *Genetics*, 78(1), pp. 53–65. Available at: https://www.ncbi.nlm.nih.gov/pubmed/17248668.

Frankham, R. (2005) "Genetics and extinction," *Biological Conservation*, 126(2), pp. 131–140. doi: 10.1016/j.biocon.2005.05.002.

Frankham, R. (2010) "Challenges and opportunities of genetic approaches to biological conservation," *Biological conservation*, 143(9), pp. 1919–1927. doi: 10.1016/j.biocon.2010.05.011.

Frankham, R., Ballou, J. D. and Briscoe, D. A. (2010) *Introduction to Conservation Genetics*. Cambridge University Press.

Fraser, D. J. and Bernatchez, L. (2001) "Adaptive evolutionary conservation: towards a unified concept for defining conservation units," *Molecular ecology*, 10(12), pp. 2741–2752. Available at: https://www.ncbi.nlm.nih.gov/pubmed/11903888.

Fu, Q. *et al.* (2013) "A revised timescale for human evolution based on ancient mitochondrial genomes," *Current biology: CB*, 23(7), pp. 553–559. doi: 10.1016/j.cub.2013.02.044.

Fu, Q. *et al.* (2015) "An early modern human from Romania with a recent Neanderthal ancestor," *Nature*, 524(7564), pp. 216–219. doi: 10.1038/nature14558.

Fuentes-Pardo, A. P. and Ruzzante, D. E. (2017) "Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations," *Molecular ecology*, 26(20), pp. 5369–5406. doi: 10.1111/mec.14264.

Fünfstück, T. *et al.* (2014) "The genetic population structure of wild western lowland gorillas (Gorilla gorilla gorilla) living in continuous rain forest," *American journal of primatology*, 76(9), pp. 868–878. doi: 10.1002/ajp.22274.

Funk, W. C. *et al.* (2012) "Harnessing genomics for delineating conservation units," *Trends in ecology & evolution*, 27(9), pp. 489–496. doi: 10.1016/j.tree.2012.05.012.

Funk, W. C. *et al.* (2019) "Improving conservation policy with genomics: a guide to integrating adaptive potential into U.S. Endangered Species Act decisions for conservation practitioners and geneticists," *Conservation Genetics*, 20(1), pp. 115–134. doi: 10.1007/s10592-018-1096-1.

Funwi-Gabga, N. *et al.* (2014) "State of the Apes 2013: Extractive Industries and Ape Conservation," in Arcus Foundation (ed.) *The status of apes across Africa and Asia*. Cambridge, UK: Cambridge University Press., pp. 252–277. doi: 10.1017/cbo9781107590274.013.

Gagneux, P. *et al.* (1999) "Mitochondrial sequences show diverse evolutionary histories of African hominoids," *Proceedings of the National Academy of Sciences of the United States of America*, 96(9), pp. 5077–5082. doi: 10.1073/pnas.96.9.5077.

Garcia, C. A. *et al.* (2020) "The Global Forest Transition as a Human Affair," *One Earth*, 2(5), pp. 417–428. doi: 10.1016/j.oneear.2020.05.002.

Garner, B. A. *et al.* (2016) "Genomics in Conservation: Case Studies and Bridging the Gap between Data and Application," *Trends in ecology & evolution*, pp. 81–83. doi: 10.1016/j.tree.2015.10.009.

Gilardi, K. V. *et al.* (2015) *Best practice guidelines for health monitoring and disease control in great ape populations*. Gland, Switzerland: IUCN SSC Primate Specialist Group, Gland, Switzerland. doi: 10.2305/iucn.ch.2015.ssc-op.56.en.

Gómez-Sánchez, D. *et al.* (2018) "On the path to extinction: Inbreeding and admixture in a declining grey wolf population," *Molecular Ecology*, 27(18), pp. 3599–3612. doi: 10.1111/mec.14824.

Gonder, M. K. *et al.* (1997) "A new west African chimpanzee subspecies?," *Nature*, 388(6640), p. 337. doi: 10.1038/41005.

Gonder, M. K. *et al.* (2011) "Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations," *Proceedings of the National Academy of Sciences*, 108(12), pp. 4766–4771. doi: 10.1073/pnas.1015422108.

Gonder, M. K., Disotell, T. R. and Oates, J. F. (2006) "New Genetic Evidence on the Evolution of Chimpanzee Populations and Implications for Taxonomy," *International Journal of Primatology*, 27(4), pp. 1103–1127. doi: 10.1007/s10764-006-9063-y.

Goossens, B. *et al.* (2003) "Successful reproduction in wild-released orphan chimpanzees (Pan troglodytes troglodytes)," *Primates; journal of primatology*, 44(1), pp. 67–69. doi: 10.1007/s10329-002-0003-y.

Green, R. E. *et al.* (2010) "A draft sequence of the Neandertal genome," *Science*, 328(5979), pp. 710–722. doi: 10.1126/science.1188021.

Grover, C. E., Salmon, A. and Wendel, J. F. (2012) "Targeted sequence capture as a powerful tool for evolutionary analysis," *American Journal of Botany*, 99(2), pp. 312–319. doi: 10.3732/ajb.1100323.

Groves, C. P. (2001) *Primate Taxonomy*. Smithsonian University Press, Washington, pp. 81–81. doi: 10.1086/343645.

Haak, W. *et al.* (2015) "Massive migration from the steppe was a source for Indo-European languages in Europe," *Nature*, 522(7555), pp. 207–211. doi: 10.1038/nature14317.

Hammer, M. F. *et al.* (2011) "Genetic evidence for archaic admixture in Africa," *Proceedings of the National Academy of Sciences of the United States of America*, 108(37), pp. 15123–15128. doi: 10.1073/pnas.1109300108.

Hans, J. B. *et al.* (2015) "Characterization of MHC class II B polymorphism in multiple populations of wild gorillas using non-invasive samples and next-generation sequencing," *American journal of primatology*, 77(11), pp. 1193–1206. doi: 10.1002/ajp.22458.

Hartl, D. L. and Clark, A. G. (2007) *Principles of Population Genetics*. Sinauer Associates Incorporated.

Hayden, E. C. (2014) "Technology: The $1,000 genome," *Nature*, 507(7492), pp. 294–295. doi: 10.1038/507294a.

Hernandez-Rodriguez, J. *et al.* (2018) "The impact of endogenous content, replicates and pooling on genome capture from faecal samples," *Molecular Ecology Resources*, 18(2), pp. 319–333. doi: 10.1111/1755-0998.12728.

Hey, J. (2010) "The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses," *Molecular biology and evolution*, 27(4), pp. 921–933. doi: 10.1093/molbev/msp298.

Hicks, A. L. *et al.* (2018) "Gut microbiomes of wild great apes fluctuate seasonally in response to diet," *Nature communications*, 9(1), p. 1786. doi: 10.1038/s41467-018-04204-w.

Hicks, T. C. *et al.* (2010) "Trade in orphans and bushmeat threatens one of the Democratic Republic of the Congo's most important populations of eastern chimpanzees (Pan troglodytes schweinfurthii)," *African primates*, 7(1), pp. 1–18.

Hicks, T. C. *et al.* (2014) "Absence of evidence is not evidence of absence: Discovery of a large, continuous population of Pan troglodytes schweinfurthii in the Central Uele region of northern DRC," *Biological Conservation*, 171, pp. 107–113. doi: 10.1016/j.biocon.2014.01.002.

Hildebrandt, T. B. *et al.* (2018) "Embryos and embryonic stem cells from the white rhinoceros," *Nature communications*, 9(1), p. 2589. doi: 10.1038/s41467-018-04959-2.

Hockings, K. J. *et al.* (2015) "Apes in the Anthropocene: flexibility and survival," *Trends in ecology & evolution*, 30(4), pp. 215–222. doi: 10.1016/j.tree.2015.02.002.

Hoelzel, A. R. (2015) "Can DNA foil the poachers?," *Science*, pp. 34–35. doi: 10.1126/science.aac6301.

Humle, T. *et al.* (2016) *Pan troglodytes (errata version published in 2018)*, *The IUCN Red List of Threatened Species.* e.T15933A129038584. Available at: https://dx.doi.org/10.2305/IUCN.UK.2016-2.RLTS.T15933A17964454.en. (Accessed: June 10, 2020).

Hvilsom, C. *et al.* (2013) "Understanding geographic origins and history of admixture among chimpanzees in European zoos, with implications for future breeding programmes," *Heredity*, 110(6), pp. 586–593. doi: 10.1038/hdy.2013.9.

Inoue, E. *et al.* (2007) "Wild chimpanzee infant urine and saliva sampled noninvasively usable for DNA analyses," *Primates; journal of primatology*, 48(2), pp. 156–159. doi: 10.1007/s10329-006-0017-y.

Inoue, E. *et al.* (2013) "Male genetic structure and paternity in western lowland gorillas (Gorilla gorilla gorilla)," *American journal of physical anthropology*, 151(4), pp. 583–588. doi: 10.1002/ajpa.22312.

IUCN (2014) *Regional action plan for the conservation of western lowland gorillas and central chimpanzees 2015-2025*. Edited by F. Maisels et al. Gland, Switzerland: IUCN SSC Primate Specialist Group.

IUCN SSC Primate Specialist Group (2020) *Regional action plan for the conservation of western chimpanzees (Pan troglodytes verus) 2020–2030*. Edited by E. G. Wessling et al. Gland, Switzerland: IUCN. doi: 10.2305/iucn.ch.2020.ssc-rap.2.en.

IUCN/SSC (ed.) (2000) *IUCN Guidelines for the Placement of Confiscated Animals*. IUCN Species Survival Commission, Gland, Switzerland.

IUCN/SSC (ed.) (2013) *Guidelines for Reintroductions and other Conservation Translocations*. IUCN Species Survival Commission, Gland, Switzerland, 2013.

IUCN/SSC (2014) *Guidelines on the Use of Ex Situ Management for Species Conservation*. IUCN Species Survival Commission; Gland, Switzerland:

Jobling, M. *et al.* (2013) *Human Evolutionary Genetics*. Garland Science, Taylor & Francis Group, LLC. doi: 10.1201/9781317952268.

Jolly, C. J., Oates, J. F. and Disotell, T. R. (1995) "Chimpanzee kinship," *Science*, pp. 185–188. doi: 10.1126/science.7716503.

Kaessmann, H. *et al.* (2001) "Great ape DNA sequences reveal a reduced diversity and an expansion in humans," *Nature Genetics*, 27(2), pp. 155–156. doi: 10.1038/84773.

Kaur, T. *et al.* (2008) "Descriptive epidemiology of fatal respiratory outbreaks and detection of a human-related metapneumovirus in wild chimpanzees (Pan troglodytes) at Mahale Mountains National Park, Western Tanzania," *American journal of primatology*, 70(8), pp. 755–765. doi: 10.1002/ajp.20565.

Köndgen, S. *et al.* (2008) "Pandemic human viruses cause decline of endangered great apes," *Current biology*, 18(4), pp. 260–264. doi: 10.1016/j.cub.2008.01.012.

Kormos, R. *et al.* (eds.) (2003) *West African Chimpanzees: Status Survey and Conservation Action Plan*. IUCN/SSC Primate Specialist Group. IUCN, Gland, Switzerland., pp. 103–104. doi: 10.1017/s0030605305250184.

Kristensen, T. N. *et al.* (2010) "Research on inbreeding in the 'omic' era," *Trends in Ecology & Evolution*, 25(1), pp. 44–52. doi: 10.1016/j.tree.2009.06.014.

Kühl, H. S. *et al.* (2017) "The Critically Endangered western chimpanzee declines by 80%," *American journal of primatology*, 79(9). doi: 10.1002/ajp.22681.

Kuhlwilm, M., Gronau, I., *et al.* (2016) "Ancient gene flow from early modern humans into Eastern Neanderthals," *Nature*, 530(7591), pp. 429–433. doi: 10.1038/nature16544.

Kuhlwilm, M., de Manuel, M., *et al.* (2016) "Evolution and demography of the great apes," *Current Opinion in Genetics & Development*, 41, pp. 124–129. doi: 10.1016/j.gde.2016.09.005.

Kuhlwilm, M. *et al.* (2019) "Ancient admixture from an extinct ape lineage into bonobos," *Nature ecology & evolution*, 3(6), pp. 957–965. doi: 10.1038/s41559-019-0881-7.

Lacy, R. C., Traylor-Holzer, K. and Ballou, J. D. (2013) "Managing for true sustainability of species," *WAZA magazine*, 14, pp. 10–14.

Lander, E. *et al.* (2001) "Initial sequencing and analysis of the human genome," *Nature*, 409(6822), pp. 860–921. doi: 10.1038/35057062.

Langergraber, K. E. *et al.* (2007) "The genetic signature of sex-biased migration in patrilocal chimpanzees and humans," *PloS one*, 2(10), p. e973. doi: 10.1371/journal.pone.0000973.

Langergraber, K. E. *et al.* (2012) "Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution," *Proceedings of the National Academy of Sciences*, 109(39), pp. 15716–15721. doi: 10.1073/pnas.1211740109.

Laurance, W. F. *et al.* (2014) "A global strategy for road building," *Nature*, 513(7517), pp. 229–232. doi: 10.1038/nature13717.

Leclère, D. *et al.* (2020) "Bending the curve of terrestrial biodiversity needs an integrated strategy," *Nature*, 366, p. eaax3100. doi: 10.1038/s41586-020-2705-y.

Leffler, E. M. *et al.* (2013) "Multiple instances of ancient balancing selection shared between humans and chimpanzees," *Science*, 339(6127), pp. 1578–1582. doi: 10.1126/science.1234070.

Leroy, E. M. *et al.* (2004) "Multiple Ebola virus transmission events and rapid decline of central African wildlife," *Science*, 303(5656), pp. 387–390. doi: 10.1126/science.1092528.

Locke, D. P. *et al.* (2011) "Comparative and demographic analysis of orang-utan genomes," *Nature*, 469(7331), pp. 529–533. doi: 10.1038/nature09687.

Lucena-Perez, M. *et al.* (2020) "Genomic patterns in the widespread Eurasian lynx shaped by Late Quaternary climatic fluctuations and anthropogenic impacts," *Molecular ecology*, 29(4), pp. 812–828. doi: 10.1111/mec.15366.

Mamanova, L. *et al.* (2010) "Target-enrichment strategies for next-generation sequencing," *Nature methods*, 7(2), pp. 111–118. doi: 10.1038/nmeth.1419.

Manolio, T. A. *et al.* (2009) "Finding the missing heritability of complex diseases," *Nature*, 461(7265), pp. 747–753. doi: 10.1038/nature08494.

de Manuel, M. *et al.* (2016) "Chimpanzee genomic diversity reveals ancient admixture with bonobos," *Science*, 354(6311), pp. 477–481. doi: 10.1126/science.aag2602.

Margulies, M. *et al.* (2005) "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, 437(7057), pp. 376–380. doi: 10.1038/nature03959.

Marques-Bonet, T., Ryder, O. A. and Eichler, E. E. (2009) "Sequencing Primate Genomes: What Have We Learned?," *Annual Review of Genomics and Human Genetics*, 10(1), pp. 355–386. doi: 10.1146/annurev.genom.9.081307.164420.

Mayr, E. (1970) *Populations, Species, and Evolution: An Abridgment of Animal Species and Evolution*. Harvard University Press.

McCarthy, M. S. *et al.* (2015) "Genetic censusing identifies an unexpectedly sizeable population of an endangered large mammal in a fragmented forest landscape," *BMC Ecology*, 15(1). doi: 10.1186/s12898-015-0052-x.

McMahon, B. J., Teeling, E. C. and Höglund, J. (2014) "How and why should we implement genomics into conservation?," *Evolutionary applications*, 7(9), pp. 999–1007. doi: 10.1111/eva.12193.

McManus, K. F. *et al.* (2015) "Inference of gorilla demographic and selective history from whole-genome sequence data," *Molecular biology and evolution*, 32(3), pp. 600–612. doi: 10.1093/molbev/msu394.

Mertes, F. *et al.* (2011) "Targeted enrichment of genomic DNA regions for next-generation sequencing," *Briefings in Functional Genomics*, 10(6), pp. 374–386. doi: 10.1093/bfgp/elr033.

Meyer, M. *et al.* (2012) "A High-Coverage Genome Sequence from an Archaic Denisovan Individual," *Science*, 338(6104), pp. 222–226. doi: 10.1126/science.1224344.

Miyamoto, M., Slightom, J. and Goodman, M. (1987) "Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region," *Science*, 238(4825), pp. 369–373. doi: 10.1126/science.3116671.

Mondal, M. *et al.* (2016) "Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation," *Nature genetics*, 48(9), pp. 1066–1070. doi: 10.1038/ng.3621.

Morgan, B. *et al.* (2011) "Regional Action Plan for the Conservation of the Nigeria-Cameroon Chimpanzee (Pan troglodytes ellioti)," *IUCN/SSC Primate Specialist Group and Zoological Society of San Diego, CA, USA.*

Morgan, D. and Sanz, C. (2007) *Best practice guidelines for reducing the impact of commercial logging on great apes in Western Equatorial Africa*. Edited by IUCN Species Survival Commission (SSC), Primate Specialist GroupConservation International, and Center for Applied Biodiversity Science (CABS). IUCN. doi: 10.2305/iucn.ch.2007.ssc-op.34.en.

Morin, P. A. *et al.* (1993) "Non-invasive sampling and DNA amplification for paternity exclusion, community structure, and phylogeography in wild chimpanzees," *Primates*, 34(3), pp. 347–356. doi: 10.1007/bf02382630.

Morin, P. A. *et al.* (2001) "Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (Pan troglodytes verus)," *Molecular Ecology*, 10(7), pp. 1835–1844. doi: 10.1046/j.0962-1083.2001.01308.x.

Myers, G. (1999) "Whole-genome DNA sequencing," *Computing in Science & Engineering*, 1(3), pp. 33–43. doi: 10.1109/5992.764214.

Nater, A. *et al.* (2017) "Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species," *Current biology: CB*, 27(22), pp. 3576–3577. doi: 10.1016/j.cub.2017.11.020.

Norris, K. *et al.* (2010) "Biodiversity in a forest-agriculture mosaic – The changing face of West African rainforests," *Biological Conservation*, 143(10), pp. 2341–2350. doi: 10.1016/j.biocon.2009.12.032.

Novembre, J. *et al.* (2008) "Genes mirror geography within Europe," *Nature*, 456(7218), pp. 98–101. doi: 10.1038/nature07331.

Olalde, I. *et al.* (2018) "The Beaker phenomenon and the genomic transformation of northwest Europe," *Nature*, 555(7695), pp. 190–196. doi: 10.1038/nature25738.

Olson, M. V. and Varki, A. (2003) "Sequencing the chimpanzee genome: insights into human evolution and disease," *Nature reviews. Genetics*, 4(1), pp. 20–28. doi: 10.1038/nrg981.

Orkin, J. D. *et al.* (2016) "Cost-effective scat-detection dogs: unleashing a powerful new tool for international mammalian conservation biology," *Scientific reports*, 6, p. 34758. doi: 10.1038/srep34758.

Ouborg, N. J. *et al.* (2010) "Conservation genetics in transition to conservation genomics," *Trends in Genetics*, 26(4), pp. 177–187. doi: 10.1016/j.tig.2010.01.001.

Pääbo, S. (2003) "The mosaic that is our genome," *Nature*, 421(6921), pp. 409–412. doi: 10.1038/nature01400.

PASA (ed.) (2002) *PanAfrican Sanctuary Alliance (PASA) Workshop Report.* Mount Kenya Safari Club, Kenya. http://www.panafricanprimates.org/.

Patterson, N. *et al.* (2012) "Ancient Admixture in Human History," *Genetics*, 192(3), pp. 1065–1093. doi: 10.1534/genetics.112.145037.

Perry, G. H. *et al.* (2010) "Genomic-scale capture and sequencing of endogenous DNA from feces," *Molecular ecology*, 19(24), pp. 5332–5344. doi: 10.1111/j.1365-294X.2010.04888.x.

Plumptre, A. J. *et al.* (2010) "Eastern Chimpanzee (Pan troglodytes schweinfurthii): Status Survey and Conservation Action Plan 2010–2020," *IUCN/SSC Primate Specialist Group, Gland, Switzerland.*

Prado-Martinez, J. *et al.* (2013) "Great ape genetic diversity and population history," *Nature*, 499(7459), pp. 471–475. doi: 10.1038/nature12228.

Prüfer, K. *et al.* (2012) "The bonobo genome compared with the chimpanzee and human genomes," *Nature*, 486(7404), pp. 527–531. doi: 10.1038/nature11128.

Prüfer, K. *et al.* (2014) "The complete genome sequence of a Neanderthal from the Altai Mountains," *Nature*, 505(7481), pp. 43–49. doi: 10.1038/nature12886.

Reich, D. *et al.* (2010) "Genetic history of an archaic hominin group from Denisova Cave in Siberia," *Nature*, 468(7327), pp. 1053–1060. doi: 10.1038/nature09710.

Rival, A. and Levang, P. (2014) *Palms of controversies: Oil palm and development challenges*. Bogor, Indonesia: CIFOR. doi: 10.17528/cifor/004860.

Sánchez-Barreiro, F. *et al.* (2020) "Historical population declines prompted significant genomic erosion in the northern and southern white rhinoceros (Ceratotherium simum)," *BioRxiv*. doi: 10.1101/2020.05.10.086686.

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

Sarich, V. M. and Wilson, A. C. (1967) "Immunological time scale for hominid evolution," *Science*, 158(3805), pp. 1200–1203. doi: 10.1126/science.158.3805.1200.

Scally, A. *et al.* (2012) "Insights into hominid evolution from the gorilla genome sequence," *Nature*, 483(7388), pp. 169–175. doi: 10.1038/nature10842.

Shafer, A. B. A. *et al.* (2015) "Genomics and the challenging translation into conservation practice," *Trends in ecology & evolution*, 30(2), pp. 78–87. doi: 10.1016/j.tree.2014.11.009.

Sibley, C. G. and Ahlquist, J. E. (1984) "The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization," *Journal of Molecular Evolution*, 20(1), pp. 2–15. doi: 10.1007/bf02101980.

Snyder-Mackler, N. *et al.* (2016) "Efficient genome-wide sequencing and low coverage pedigree analysis from non-invasively collected samples," *Genetics*, 203(2), pp. 699–714. doi: 10.1101/029520.

Steiner, C. C. *et al.* (2013) "Conservation genomics of threatened animal species," *Annual review of animal biosciences*, 1, pp. 261–281. doi: 10.1146/annurev-animal-031412-103636.

Steiper, M. E. (2006) "Population history, biogeography, and taxonomy of orangutans (Genus: Pongo) based on a population genetic meta-analysis of multiple loci," *Journal of human evolution*, 50(5), pp. 509–522. doi: 10.1016/j.jhevol.2005.12.005.

Stiles, D. *et al.* (2013) *Stolen Apes: The Illicit Trade in Chimpanzees, Gorillas, Bonobos, and Orangutans*. Birkeland Trykkeri AS, Norway.

Stone, A. C. *et al.* (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(1556), pp. 3277–3288. doi: 10.1098/rstb.2010.0096.

Strindberg, S. *et al.* (2018) "Guns, germs, and trees determine density and distribution of gorillas and chimpanzees in Western Equatorial Africa," *Science advances*, 4(4), p. eaar2964. doi: 10.1126/sciadv.aar2964.

Supple, M. A. and Shapiro, B. (2018) "Conservation of biodiversity in the genomics era," *Genome biology*, 19(1), p. 131. doi: 10.1186/s13059-018-1520-3.

Taberlet, P., Waits, L. P. and Luikart, G. (1999) "Noninvasive genetic sampling: look before you leap," *Trends in ecology & evolution*, 14(8), pp. 323–327. doi: 10.1016/s0169-5347(99)01637-7.

Taylor, H. R., Dussex, N. and van Heezik, Y. (2017) "Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners," *Global Ecology and Conservation*. Elsevier, 10, pp. 231–242. doi: 10.1016/j.gecco.2017.04.001.

Thalmann, O., Serre, D., *et al.* (2004) "Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA," *Molecular Ecology*, 14(1), pp. 179–188. doi: 10.1111/j.1365-294x.2004.02382.x.

Thalmann, O., Hebler, J., *et al.* (2004) "Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes," *Molecular Ecology*, 13(2), pp. 321–335. doi: 10.1046/j.1365-294x.2003.02070.x.

Thalmann, O. *et al.* (2007) "The complex evolutionary history of gorillas: insights from genomic data," *Molecular biology and evolution*, 24(1), pp. 146–158. doi: 10.1093/molbev/msl160.

Traylor-Holzer, K. (2011) "Identifying gaps and opportunities for inter-regional ex situ species management," *WAZA Magazine*, 12, pp. 30–33.

Traylor-Holzer, K., Leus, K. and Bauman, K. (2019) "Integrated Collection Assessment and Planning (ICAP) workshop: Helping zoos move toward the One Plan Approach," *Zoo biology*, 38(1), pp. 95–105. doi: 10.1002/zoo.21478.

Traylor-Holzer, K., Leus, K. and Mcgowan, P. (2013) "Integrating Assessment of Ex Situ Management Options into Species Conservation Planning," *WAZA Magazine*, 14, pp. 6–9.

Tutin, C. *et al.* (2005) *Regional Action Plan for the Conservation of Chimpanzees and Gorillas in Western Equatorial Africa*. IUCN/SSC Primate Specialist Group Conservation International. Washington, DC. doi: 10.2305/iucn.ch.2005.ssc-rap.1.en.

van der Valk, T. *et al.* (2017) "Whole mitochondrial genome capture from faecal samples and museum-preserved specimens," *Molecular ecology resources*, 17(6), pp. e111–e121. doi: 10.1111/1755-0998.12699.

van der Valk, T. *et al.* (2018) "Significant loss of mitochondrial diversity within the last century due to extinction of peripheral populations in eastern gorillas," *Scientific reports*, 8(1), p. 6551. doi: 10.1038/s41598-018-24497-7.

van der Valk, T. *et al.* (2019) "Estimates of genetic load in small populations suggest extensive purging of deleterious alleles," *BioRxiv*. doi: 10.1101/696831.

Venn, O. *et al.* (2014) "Nonhuman genetics. Strong male bias drives germline mutation in chimpanzees," *Science*, 344(6189), pp. 1272–1275. doi: 10.1126/science.344.6189.1272.

Venter, J. C. *et al.* (2001) "The sequence of the human genome," *Science*, 291(5507), pp. 1304–1351. doi: 10.1126/science.1058040.

Vernot, B. *et al.* (2016) "Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals," *Science*, 352(6282), pp. 235–239. doi: 10.1126/science.aad9416.

Vernot, B. and Akey, J. M. (2014) "Resurrecting surviving Neandertal lineages from modern human genomes," *Science*, 343(6174), pp. 1017–1021. doi: 10.1126/science.1245938.

Vigilant, L. and Guschanski, K. (2009) "Using genetics to understand the dynamics of wild primate populations," *Primates; journal of primatology*, 50(2), pp. 105–120. doi: 10.1007/s10329-008-0124-z.

Warren, R. *et al.* (2018) "The projected effect on insects, vertebrates, and plants of limiting global warming to 1.5°C rather than 2°C," *Science*, 360(6390), pp. 791–795. doi: 10.1126/science.aar3646.

Wasser, S. K. *et al.* (2015) "Genetic assignment of large seizures of elephant ivory reveals Africa's major poaching hotspots," *Science*, 349(6243), pp. 84–87. doi: 10.1126/science.aaa2457.

Waterson, R., Lander, E. and Wilson, R. (2005) "Initial sequence of the chimpanzee genome and comparison with the human genome," *Nature*, 437(7055), pp. 69–87. doi: 10.1038/nature04072.

WAZA (ed.) (2005) *Building a Future for Wildlife - The World Zoo and Aquarium Conservation Strategy*. WAZA Executive Office 3012 Bern, Switzerland.

Wegmann, D. and Excoffier, L. (2010) "Bayesian inference of the demographic history of chimpanzees," *Molecular biology and evolution*, 27(6), pp. 1425–1435. doi: 10.1093/molbev/msq028.

White, L. C. *et al.* (2019) "A roadmap for high-throughput sequencing studies of wild animal populations using noninvasive samples and hybridization capture," *Molecular ecology resources*, 19(3), pp. 609–622. doi: 10.1111/1755-0998.12993.

Wich, S. A. *et al.* (2012) "Understanding the Impacts of Land-Use Policies on a Threatened Species: Is There a Future for the Bornean Orang-utan?," *PLoS one*, 7(11), p. e49142. doi: 10.1371/journal.pone.0049142.

Wich, S. A. *et al.* (2014) "Will Oil Palm's Homecoming Spell Doom for Africa's Great Apes?," *Current Biology*, 24(14), pp. 1659–1663. doi: 10.1016/j.cub.2014.05.077.

Wright, L. I., Tregenza, T. and Hosken, D. J. (2008) "Inbreeding, inbreeding depression and extinction," *Conservation Genetics*, 9(4), pp. 833–843. doi: 10.1007/s10592-007-9405-0.

Xu, X. and Arnason, U. (1996) "The Mitochondrial DNA Molecule of Sumatran Orangutan and a Molecular Proposal for Two (Bornean and Sumatran) Species of Orangutan," *Journal of Molecular Evolution*, 43(5), pp. 431–437. doi: 10.1007/pl00006103.

Xue, Y. *et al.* (2015) “Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding,” *Science*, 348(6231), pp. 242–245. doi: 10.1126/science.aaa3952.

Yunis, J. and Prakash, O. (1982) “The origin of man: a chromosomal pictorial legacy,” *Science*, 215(4539), pp. 1525–1530. doi: 10.1126/science.7063861.

Annex

1. Experimental protocols

1.1. Library preparation

For any DNA to be sequenced, the DNA molecules need to be first converted into DNA libraries. In turn, these libraries can be the substrate for the subsequent in-solution hybridization capture.

In library preparation synthetic adapters are attached to both DNA fragment ends, thus rendering a molecular structure that enables their amplification and the recognition by the sequencing platform. There are two different methods to build libraries for the gold-standard sequencing platform (Illumina): double-stranded (Meyer and Kircher, 2010) and single-stranded library preparation (Gansauge and Meyer, 2013). On the one hand, double-stranded library preparation is the most widely adopted method and the one used in the three works presented in this thesis. On the other hand, single stranded library preparation is mainly limited to ancient DNA (aDNA) samples that have very low quantities of DNA. Therefore, when using the term "library" I will be referring to double-stranded libraries.

The first step of library preparation is DNA fragmentation. Blood, tissue, hair or fecal extracted DNA molecules have a long size, and although fecal DNA is more degraded (Figure S1), it still needs to be fragmented.

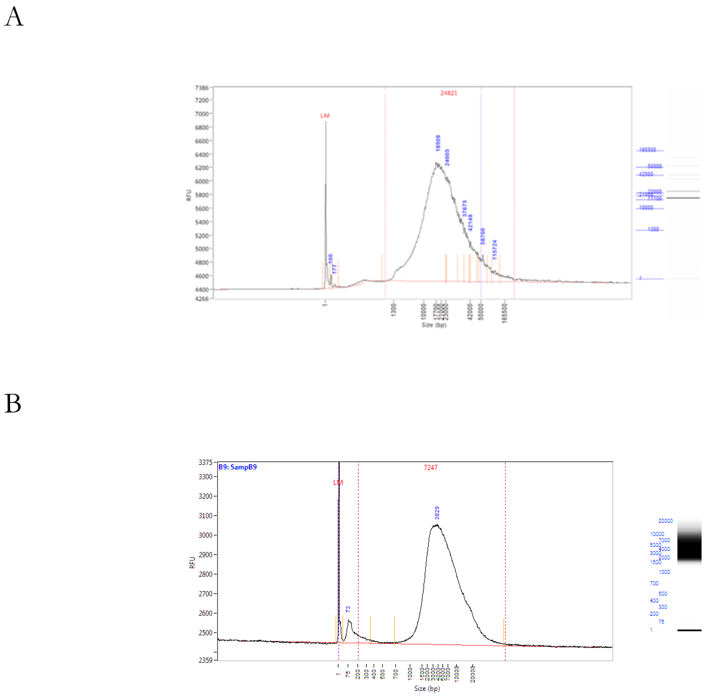


Figure S1. Fragment analyzer profile for an example of A) Genomic DNA and B) fecal DNA.

DNA is then sheared setting the parameters to obtain the desired insert size for the DNA fragments. Once the DNA is fragmented, the next step is end-repair. 3' and 5' fragment ends have overhangs as a result of fragmentation that need to be filled to create blunt ends. P5 and P7 adapters can then be ligated to the double stranded fragments (blunt-end ligation). Ligated adapters are not complete so before PCR amplification they need to be filled in. Adapters can have inline barcodes that are directly attached to the DNA fragment. Early barcoding of the library (in the adapter ligation step rather than in the final PCR amplification) lowers the probability of indiscernible contamination from close wells. Moreover, previous studies have shown a better capture efficiency when the library size is small (Rohland and Reich, 2012). Therefore, small-sized libraries with inline barcodes can already be pooled for capture without the need to fully extend the library at its final length.

Finally, the prepared library is amplified using PCR. At this step, depending on the primers used, the library can be maintained at a small size for capture (no indexing primers) or double-indexed and extended at full length for sequencing (indexing primers) (Kircher, Sawyer and Meyer, 2012) (Figure S2).

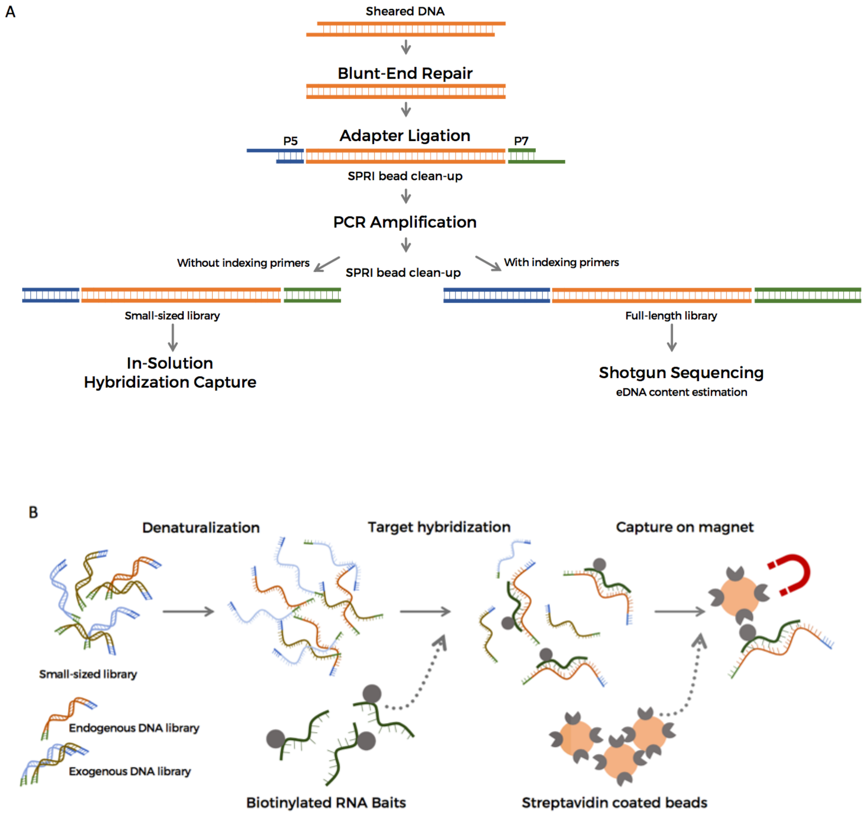


Figure S2. Library preparation (A) and in-solution hybridization capture (B) representation.

The method for library preparation that I have followed is a modification of the initial double-stranded illumina library protocol (Meyer and Kircher, 2010) that reduces the number of clean-ups by using SPRI beads (Rohland and Reich, 2012), and subsequently it is less time consuming with increased library yield (Carøe *et al.*, 2018).

1.2. Target capture

Once the libraries have been prepared, double-index, amplified libraries can be directly sequenced (Figure S2), in an approach called shotgun sequencing. In the case of good quality samples with sufficient depth of sequencing, complete genomes would be recovered. However, with complex samples such as NI samples or aDNA, where the percentage of eDNA is usually very low, the obtention of significant information would be economically unfeasible. Still, shotgun sequencing can provide valuable insight such as the percentage of eDNA, which will guide pool preparation for target capture. qPCR done directly from the extracted DNA is another method to obtain the percentage of eDNA (Morin *et al.*, 2001).

With pre-designed biotinylated RNA (or DNA) baits complementary to the regions of interest (SNPs in chapter 3.1., exome in chapter 3.2. and chromosome 21 in chapter 3.3.), it is possible to enrich and sequence only some parts of the genome and thus reach higher coverage in those regions. In this thesis I have followed Agilent protocols (Figure S2). First, library pools are mixed with blocking oligos (Rohland and Reich, 2012), Human cot-1 and salmon sperm at 95ºC to block repetitive regions and avoid the incorrect annealing between single-stranded DNA molecules after denaturalization. Next, at 65ºC prewarmed RNA baits with the hybridization buffer are mixed with the library pool and incubated for 24 h at 65ºC. After incubation, several washes with streptavidin-coated beads on a magnet are used to separate the captured library from the uncaptured fragments. In low endogenous DNA fecal samples, it is usually recommended to perform two rounds of capture to increase efficiency in the recovery of fragments (Hernandez-Rodriguez *et al.*, 2018). In such cases, after the first capture, PCR is done without indexing the primers and maintaining small adapter length. Otherwise, a final PCR amplification with indexed primers (tagging the capture pool) is used to obtain full-length adapters, which are now ready to be sequenced.

2. Annex bibliography

Carøe, C. et al. (2018) 'Single‐tube library preparation for degraded DNA', *Methods in Ecology and Evolution*, 9(2), pp. 410–419.

Gansauge, M.-T. and Meyer, M. (2013) 'Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA', *Nature Protocols*, 8(4), pp. 737–748.

Hernandez-Rodriguez, J. et al. (2018) 'The impact of endogenous content, replicates and pooling on genome capture from faecal samples', *Molecular Ecology Resources*, 18(2), pp. 319–333.

Kircher, M., Sawyer, S. and Meyer, M. (2012) 'Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform', *Nucleic acids research*, 40(1), p. e3.

Meyer, M. and Kircher, M. (2010) 'Illumina sequencing library preparation for highly multiplexed target capture and sequencing', *Cold Spring Harbor protocols*, 2010(6), p. db.prot5448.

Morin, P. A. et al. (2001) 'Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (Pan troglodytes verus)', *Molecular Ecology*, 10(7), pp. 1835–1844.

Rohland, N. and Reich, D. (2012) 'Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture', *Genome research*, 22(5), pp. 939–946.