Unraveling evolution through Next Generation Sequencing

Michael Westbury

Univ.-Diss.

zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Evolutionsgenetik."

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
Institut für Biochemie und Biologie
der Universität Potsdam

Hauptbetreuer: Prof. Michael Hofreiter (University of Potsdam)
weitere Gutachter: Associate Prof. Eline Lorenzen (Natural history museum of Denmark) and Prof. Jon Waters (University of Otago, New Zealand)

Published online at the
Institutional Repository of the University of Potsdam:
URN urn:nbn:de:kobv:517-opus4-409981
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-409981

Declaration of authorship

I, Michael Westbury, declare that the work contained within this thesis titled, "Unraveling evolution through Next Generation Sequencing", is purely my own unless stated otherwise.

Chapter 2:
"Complete mitochondrial genome of a bat eared fox (*Otocyon megalotis*), along with phylogenetic considerations"
This manuscript was published in "Mitochondrial DNA part B" in 2017 and can be found online at http://dx.doi.org/10.1080/23802359.2017.1331325. I carried out the extraction lab work, data analyses and wrote the manuscript.

Chapter 3:
"A mitogenomic timetree for Darwin's enigmatic South American mammal *Macrauchenia patachonica*"
This manuscript was published in "Nature communications" in 2017 under the doi 10.1038/ncomms15951. I carried out the majority of the lab work, including the sample that yielded DNA. Sina Baleka processed some samples that were unsuccessful in yielding DNA. I performed the majority of the data analyses assisted by Axel Barlow, Sina Baleka, Stefanie Hartmann, and Johanna L.A. Paijmans. I wrote the majority of the methods and results in the manuscript. I further supplied input in the introduction and discussion of the manuscript together with Ross D.E. MacPhee and Michael Hofreiter.

Chapter 4:
"Population and conservation genomics of the world's rarest hyena species, the brown hyena (*Parahyena brunnea*)"
This manuscript has been submitted to "Genome research" and is also available online in the preprint server "Biorxiv" under https://doi.org/10.1101/170621. I conceived the project idea together with Michael Hofreiter. I performed all lab work excluding the two captive samples (brown and striped hyena). Love Dalén and the National Genomics Infrastructure (Stockholm) processed the two captive samples. I performed all analyses apart from the sliding trees which Stefanie Hartmann performed. I wrote the majority of the manuscript.

Signed:

Michael Westbury

Acknowledgements

Firstly, I would like to acknowledge and thank my supervisor, Prof. Michael Hofreiter, for giving me the opportunity to work on such interesting and state of art projects. Without your guidance and support, I am sure I would have been lost. I really appreciated the chances you gave me to travel all around the world, meet new people and learn new skills on the work budget.

Furthermore, I would like to thank ….

All of the collaborators on the projects presented within this thesis.
The Evolutionary Adaptive Genomics group, present and past members. We have grown so significantly over the years that there are too many to name.
Stefanie Hartmann for being the speediest proof reader of time and giving some valuable feedback on the thesis.
The Potsdam Porcupines for giving me something to do apart from work.
The Fellas over fame crew for supplying good banter in times of need and boredom and all of the lads around the world.

A special thanks to my family, especially mum and dad who encouraged me to leave the homeland in the pursue of greater things. I am grateful that even when what I was doing seemed a bit abstract, you always were or at least pretended to be interested.

Last but not least, I would like to thank the lovely Binia De Cahsan. You helped me through the nightmare of the German bureaucratic system, supplied ample German language knowledge and put up with all of my rubbish. I greatly appreciate all the time and help you give me and wouldn't have wanted it any other way.

Table of contents

List of abbreviations

Bp - base pairs

Kya - thousand years ago

Ma - million years ago

NGS - Next generation sequencing

DNA - deoxyribonucleic acid

Mbp - Mega bases (1,000,000bp)

PCR - polymerase chain reaction

RFLP - restriction fragment length polymorphisms

SNP - single nucleotide polymorphism

aDNA - ancient DNA

Summary

The sequencing of the human genome in the early 2000s led to an increased interest in cheap and fast sequencing technologies. This interest culminated in the advent of next generation sequencing (NGS). A number of different NGS platforms have arisen since then all promising to do the same thing, i.e. produce large amounts of genetic information for relatively low costs compared to more traditional methods such as Sanger sequencing. The capabilities of NGS meant that researchers were no longer bound to species for which a lot of previous work had already been done (e.g. model organisms and humans) enabling a shift in research towards more novel and diverse species of interest. This capability has greatly benefitted many fields within the biological sciences, one of which being the field of evolutionary biology. Researchers have begun to move away from the study of laboratory model organisms to wild, natural populations and species which has greatly expanded our knowledge of evolution. NGS boasts a number of benefits over more traditional sequencing approaches. The main benefit comes from the capability to generate information for drastically more loci for a fraction of the cost. This is hugely beneficial to the study of wild animals as, even when large numbers of individuals are unobtainable, the amount of data produced still allows for accurate, reliable population and species level results from a small selection of individuals.

The use of NGS to study species for which little to no previous research has been carried out on and the production of novel evolutionary information and reference datasets for the greater scientific community were the focuses of this thesis. Two studies in this thesis focused on producing novel mitochondrial genomes from shotgun sequencing data through iterative mapping, bypassing the need for a close relative to serve as a reference sequence. These mitochondrial genomes were then used to infer species level relationships through phylogenetic analyses. The first of these studies involved reconstructing a complete mitochondrial genome of the bat eared fox (*Otocyon megalotis*). Phylogenetic analyses of the mitochondrial genome confidently placed the bat eared fox as sister to the clade consisting of the raccoon dog and true foxes within the canidae family. The next study also involved reconstructing a mitochondrial

genome but in this case from the extinct *Macrauchenia* of South America. As this study utilised ancient DNA, it involved a lot of parameter testing, quality controls and strict thresholds to obtain a near complete mitochondrial genome devoid of contamination known to plague ancient DNA studies. Phylogenetic analyses confidently placed *Macrauchenia* as sister to all living representatives of Perissodactyla with a divergence time of ~66 million years ago. The third and final study of this thesis involved *de novo* assemblies of both nuclear and mitochondrial genomes from brown and striped hyena and focussed on demographic, genetic diversity and population genomic analyses within the brown hyena. Previous studies of the brown hyena hinted at very low levels of genomic diversity and, perhaps due to this, were unable to find any notable population structure across its range. By incorporating a large number of genetic loci, in the form of complete nuclear genomes, population structure within the brown hyena was uncovered. On top of this, genomic diversity levels were compared to a number of other species. Results showed the brown hyena to have the lowest genomic diversity out of all species included in the study which was perhaps caused by a continuous and ongoing decline in effective population size that started about one million years ago and dramatically accelerated towards the end of the Pleistocene.

The studies within this thesis show the power NGS sequencing has and its utility within evolutionary biology. The most notable capabilities outlined in this thesis involve the study of species for which no reference data is available and in the production of large amounts of data, providing evolutionary answers at the species and population level that data produced using more traditional techniques simply could not.

Zusammenfassung

Die Sequenzierung des ersten menschlichen Genoms Anfang der 2000er Jahre förderte das Interesse an kostengünstigen und gleichzeitig schnelleren Sequenziertechniken. Dieses Interesse erreichte seinen derzeitigen Höhepunkt in der Einführung des sogenannten *Next Generation Sequencings* (NGS). Seitdem wurden zahlreiche NGS-Plattformen entwickelt, die alle dem gleichen Prinzip folgen, nämlich das Erzeugen großer Mengen genetischer Information zu relativ geringen Preisen verglichen mit herkömmlichen Methoden wie der Sanger-Sequenzierung. Die neue Leistungsfähigkeit von NGS bedeutete, dass Forscher nicht mehr länger an Organismen gebunden waren an denen bereits seit Jahren geforscht wurde (bspw. Modellorganismen oder der Mensch), sondern ermöglichte eine Verschiebung in Richtung neuerer und unterschiedlicher Arten von Interesse. Dieses Potential hat viele Wissenschaftsfelder positiv beeinflusst innerhalb der Biowissenschaften, u.a. das Feld der Evolutionsbiologie. Forscher haben angefangen sich zunehmend von Modellorganismen in Laboratorien wegzubewegen hinzu wildlebenden, natürlich vorkommenden Populationen und Arten, was unser Verständnis von Evolution maßgeblich erweitert hat. NGS hat mehrere Vorteile aufzuweisen gegenüber den herkömmlichen Sequenziermethoden. Der wohl größte Vorteil ist die Gewinnung genetischer Daten für mehrere Genorte (Loci) gleichzeitig zu einem Bruchteil der bisherigen Kosten. Das ist besonders nützlich für die Untersuchung wildlebender Tiere da, selbst wenn nicht ausreichend viele Individuen vorliegen, die gewonnene Menge an Daten genaue und verlässliche Ergebnisse auf Populations- sowie Artebene für eine kleine Auswahl an Individuen liefert.

Die Verwendung von NGS zur Untersuchung von Arten, für die bisher wenig oder gar keine vorherigen Forschungsergebnisse vorliegen sowie die Gewinnung neuartiger

Informationen im Bereich Evolution ebenso wie die Erstellung eines Referenzdatensatzes, der der wissenschaftlichen Gemeinschaft zur Verfügung gestellt werden kann, waren der Fokus dieser Arbeit. Zwei Studien in dieser Arbeit setzten ihren Fokus in der Gewinnung noch nicht publizierter, mitochondrialer Genome, die mittels *iterative mapping* erstellt wurden und so das Vorhandensein einer Referenzsequenz eines nahen Verwandten der untersuchten Art unnötig machten. In beiden Fällen wurden *Shotgun* Sequenzierungsdaten verwendet. Die so gewonnenen mitochondrialen Genome wurden dann genutzt, um innerartliche Verwandtschaftsverhältnisse mit hilfe von phylogenetischen Analysen zu klären. Die erste Studie befasste sich mit der Rekonstruktion des kompletten mitochondrialen Genoms des Löffelhundes (*Otocyon megalotis*). Die phylogenetische Analyse des mitochondrialen Genoms positionierten den Löffelhund sicher als Schwestergruppe der Klade bestehend aus Marderhund und echten Füchsen innerhalb der Familie Canidae. Die zweite Studie hat sich ebenfalls mit der Rekonstruktion eines mitochondrialen Genoms auseinandergesetzt, diesmal von einer bereits ausgestorbenen Art Südamerikas, dem *Macrauchenia.* Da diese Studie auf sehr alter DNA (*ancient DNA*) basiert, schließt sie viele Parametertests, Qualitätskontrollen sowie strenge Filterkriterien ein um ein fast vollständiges mitochondriales Genom erhalten zu können, frei von den für ancient DNA typischen Kontaminationen. Phylogenetische Analysen positionieren *Macrauchenia* als Schwestergruppe zu allen anderen lebenden Vertretern der Perissodactyla mit einer Abspaltung vor ~66 Millionen Jahren. Die dritte und letzte Studie dieser Arbeit beinhaltet die *de novo* Konstruktionen von nukleären und mitochondrialen Genomen der Schabracken- und Streifenhyäne mit Fokus auf demographische, genetische Diversität sowie Populationsgenomische Analysen innerhalb der Schabrackenhyänen. Vorausgehende Studien an der Schabrackenhyäne gaben Hinweise für einen geringen Grad an genomischer Diversität und, waren vielleicht deshalb, bisher nicht in der Lage eine nennenswerte Populationsstruktur der Schabrackenhyäne aufzudecken. Zusätzlich wurde die genomische Diversität mit der von einer

Reihe anderer Arten verglichen. Die Ergebnisse zeigen, dass die Schabrackenhyäne die niedrigste genomische Diversität aufweist im Vergleich zu den in dieser Studie verwendeten Arten, was vielleicht mit einem kontinuierlichen und fortschreitenden Rückgang der effektiven Populationsgröße dieser Spezies zu erklären ist, der vor ca. einer Million Jahre eingesetzt hat und dramatisch zugenommen hat zum Ende des Pleistozän.

Die Studien dieser Arbeit zeigen das Potential von NGS Sequenzierung und ihren Nutzen innerhalb der Evolutionsbiologie. Die nennenswertesten Anwendungen von NGS, die in dieser Arbeit hervorgehoben wurden, sind zum Einen der Nutzen für Organismen bzw. Arten für die es keine verfügbaren Referenzdaten gibt sowie zum Anderen die Gewinnung von großen Datenmengen, die die Grundlage bilden zur Beantwortung evolutionsbiologischer Fragestellungen auf Art- und Populationsebene, was vorhergegangene, traditionelle Methoden bisher nicht leisten konnten.

Chapter 1: Introduction

All heritable information within in an organism is encoded in its DNA. The inclusion of DNA into the field of evolutionary biology, therefore, introduced a valuable tool for the study of evolutionary relationships. DNA and other molecular markers (e.g. proteins) can be used to confidently infer evolutionary relationships by providing an extra form of evidence to be used alongside the more traditional forms of evidence, such as behaviour and morphology (Hillis 1987). Molecular markers are considered to be more robust against confounding factors that limit morphological analyses, despite cases of molecular convergence in specific genes between distantly related species (Parker et al. 2013; Swanson et al. 1991). Convergent evolution, the limited number of markers and the slower rate at which morphological differences accumulate between reproductively isolated populations serve as the main factors limiting the power of morphology based comparisons to infer population and species relationships (Maxson and Wilson 1974; Hillis 1987).

The study of DNA and evolution began, however, not with the analysis of linear nucleotide order as it is today. One of the earliest methods using DNA to study species relationships was DNA-DNA hybridisation. The discovery that complementary single strands of DNA would hybridise with one another was utilised, and the temperature at which single strands would denature from one another was used as an estimate for genome-wide divergence between individuals (Sibley and Ahlquist 1984; Britten and Kohne 1968). While this provided valuable new insights into species comparisons, it was still considered as a single marker and was therefore limited in its power. Restriction fragment length polymorphisms (RFLP) later added more power to the use of DNA in evolutionary analyses. Restriction enzymes were used to cut DNA at specific sites, and depending on nucleotide differences in the DNA strand at these cut sites, differently sized fragments would be the result. Polymorphisms in the resulting fragment lengths were then correlated with other information such as putative species or population (Botstein et al. 1980). One could use different enzymes, providing different cuts across the genome allowing for multiple markers to be analysed. RFLP analyses provided revolutionary

information at the time, for example the maternal inheritance of mitochondrial DNA (Hutchison et al. 1974), which is still important and taken for granted as common knowledge today. However, the ability to sequence the nucleotide order and compare homologous DNA sequences was when the use of DNA to infer evolutionary relationships became a lot more powerful. As each nucleotide sequenced represents a single character or marker, and the more markers one has, the more reliable a comparative analysis is deemed to be, comparisons using DNA sequences had great power over all other methods in determining evolutionary relationships (Harrison 1989).

1.1 Traditional DNA sequencing

The ability to directly sequence DNA nucleotides came with the invention of the Sanger sequencing method, which was developed in 1975 (Sanger and Coulson 1975; Sanger et al. 1977). Sanger was the gold standard for DNA sequencing over the next decades (Grada and Weinbrecht 2013). It requires the sequencing of clonally amplified DNA fragments. Multiple clonal copies of a single fragment of DNA were generally produced using one of two methods; cloning or polymerase chain reaction (PCR). In brief, cloning involves the insertion of randomly fragmented DNA into a plasmid vector which is then transformed into a bacterial host. The host then replicates the plasmid along with the inserted DNA. Once this replication has occurred, the inserted fragment can be sequenced (Alberts et al. 2002). In contrast to cloning, PCR allows for a more selective method of amplification. The use of PCR primers flanking the DNA region of interest allow for targeted sequencing. Fundamentally, a single cycle of PCR reaction involves a series of temperature changes, PCR primers complementary to the flanking area of interest, free nucleotides and DNA polymerase. Double stranded DNA is initially heated up and denatured into single stranded DNA, cooled slightly to allow the binding of the PCR primers, and finally heated up slightly to allow the polymerase to bind and elongate the complementary strand by adding the free nucleotides. This process is repeated for a number of cycles producing large amounts of clonally amplified DNA fragments (Bartlett and Stirling 2003). Sequencing of DNA via Sanger works through a chain termination process, utilising a principle similar to PCR.

Sanger sequencing, however, also includes special nucleotides called di-deoxynucleotidetriphosphates (ddNTPs) that terminate DNA strand elongation prematurely. While there have been a number of modifications over time, such as the inclusion of fluorescent labelling (Prober et al. 1987), automation and capillary electrophoresis, the original method worked by first separating the amplified DNA sample into four independent reactions, each one containing a different one of the four possible ddNTPs (A, C, G, T), then by separating the resulting fragments by gel electrophoresis (Sanger et al. 1977).

Since its introduction, Sanger sequencing has allowed for the decoding of a number of large scale sequencing projects, the largest of them being the publicly funded human genome project (International Human Genome Sequencing Consortium 2004). The human genome project took over a decade to complete and ended up costing billions of US dollars. Despite the successful implementation of Sanger sequencing in the human genome project, due to the limited ability for parallelisation and high costs, it is generally not feasible to be more widely implemented in the sequencing of other organisms.

1.2 Next generation sequencing

The completion of the human genome led to an increased interest in cheaper and faster sequencing technologies. This interest culminated in the development of "next generation sequencing" (NGS) technologies (Grada and Weinbrecht 2013) also known as high-throughput sequencing. Next generation sequencing allows for millions of independent DNA fragments to be sequenced in parallel, greatly increasing the speed at which large amounts of genetic information can be obtained for a relatively low price. On top of increased speed and decreased price, there are also many other significant benefits NGS has over more traditional methods. This thesis focusses on the following benefits: the ability of the technologies to sequence many independent fragments of DNA in parallel without prior information on nucleotide order; the ability to sequence short fragments of highly degraded DNA; the large volume of data that can

be produced within a short time period of time at relatively low costs, and the sensitivity increase one obtains through the inclusion of many independent loci during downstream analyses.

Currently there are many different types of next generation sequencing platforms on the market, all consisting of their own unique chemistry and lab methodologies to prepare DNA for sequencing in the form of "library preparation" (Buermans and den Dunnen 2014). However, as this thesis focuses on the production and analysis of Illumina sequencing data, I will focus on the methodologies for Illumina data generation and analyses.

1.2.1 Illumina technologies

Illumina is arguably the most widely used NGS platform and consists of a number of different sequencing machines. The most powerful to date is the new Illumina Novaseq 6000, which can output up to three terabytes of data per flow cell over a 40 hour period. In order for DNA to be sequenced, it first needs to be prepared into a sequencing library. For shotgun sequencing, DNA is randomly fragmented into many small fragments, Illumina specific adapter sequences are then ligated onto blunt ended fragments, and fragments with adapters attached are clonally amplified using PCR. The PCR step uses primers that are complementary to the adapter sequences. Library molecules are then denatured into single stranded DNA and bound to the Illumina flowcell by the adapter sequence. As adapter sequences not only function to ligate the library molecules to the sequencer flowcell but also as a template for primers to ligate to, DNA can be sequenced without prior knowledge. Illumina sequencing then utilises a bridge amplification method (Fig 1.1A) which is used to form clusters of a single DNA fragment consisting of many clonal copies on a flow cell. Each independent DNA fragment forms its own cluster on the flowcell. Illumina sequencing relies on a sequencing by synthesis method (Fig 1.1B). DNA library molecules act as the template strand on which fluorescently labeled nucleotides are incorporated. Incorporated nucleotides also have reversible 3' blockers so that the polymerase adds only a single nucleotide per cycle. The incorporated base is then digitally read and converted into a nucleotide sequence along with corresponding base quality scores

determined by the type of and intensity of light produced by the cluster. The order of bases from a single cluster is read by the machine and output as a "read" (Grada and Weinbrecht 2013).

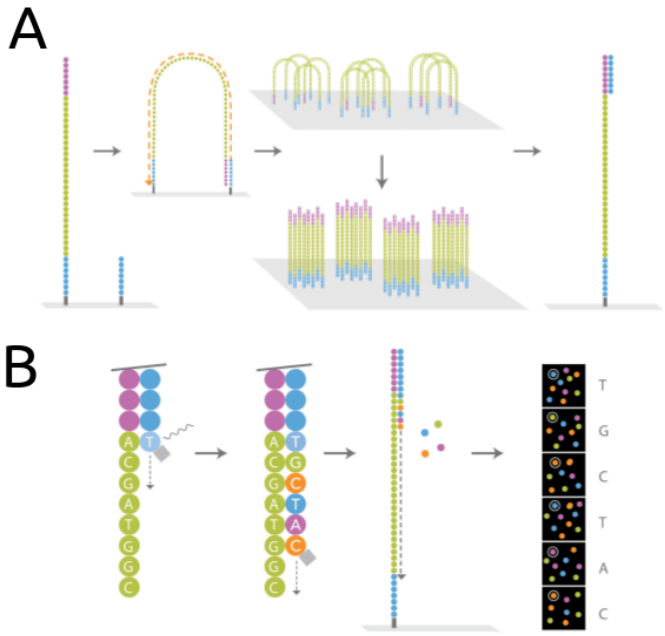


Figure 1.1: A) Bridge amplification. The library molecule is hybridised to the flow cell and a clonal cluster is generated from a single molecule through bridge amplification. **B) Sequencing by synthesis.** During each cycle a fluorescently tagged nucleotide is incorporated to the molecule, releasing a base specific emission which is registered by high-sensitivity cameras. The 3' blocker (seen in grey) is then removed before the next cycle can begin. (Figure modified from Illumina. Inc)

1.3 Applications of NGS

Pre-NGS technologies often limited researchers in their choice of study organism, especially those without vast amounts of expendable funding. Genetic resources from already studied species typically were required to provide comparative datasets or even simply for the designing of PCR primers. The field of evolutionary biology has greatly benefited from the introduction of NGS, as it provided a means to produce large comparative datasets for relatively low costs and allowed species to be studied for which no prior nucleotide information was available. These factors enabled researchers to move away from laboratory model organisms and

their close relatives towards research on wild, natural populations and species (Ekblom and Galindo 2011). The ability to produce previously unprecedented amounts of data has led to the development of a number of large consortiums, investigating whole genomic datasets within and between species. Two well known consortiums are the avian genomes consortium and the 1000 genomes project. The avian genome consortium sequenced complete nuclear genomes from 48 avian species representing all major extant clades within the avian lineage (Zhang et al. 2014). Through comparative genomic analyses this consortium was able to discover high levels of genomic conservation between species, while also discovering that there are certain selective constraints on different gene families depending on the which lineage within the avian phylogeny the species occupies. These results provided a large increase in the knowledge, not only about how the avian genome evolved, but also about avian evolution as a whole which would not have been possible without NGS. The avian genome consortium has now set out to become the "Bird 10k consortium" with the goal of sequencing 10,000 bird nuclear genomes. The 1000 genomes project was launched in 2008 and results were published in 2015 (1000 Genomes Project Consortium et al. 2015). This project originally set out to resequence 1000 human nuclear genomes to create a large dataset of human variation for future researchers to use but ended up sequencing nuclear genomes of 2,504 individuals from 26 populations around the world. Considering the first human genome took over a decade to sequence, these projects really show the efficiency and power NGS can have.

1.4 Benefits of NGS within evolutionary biology

The clarity of understanding of both species level relationships and population level relationships is greatly enhanced by the inclusion of more loci easily producible through NGS. Depending on the time since populations diverged from each other, genetic differentiation between populations may not be very high. In such a case, the higher sensitivity available through the inclusion of more loci to find is able the differentiate the finer differences between populations, allowing for more accurate estimates of population structure and the relationships between populations. The power of including more loci is highlighted in Chapter 4.

Similarly to the problems faced when inferring population relationships, inferring species relationships based on a single or just a few genes/loci also encounters difficulties. Phylogenetic trees are often used to infer species relationships. However, although results represent the evolutionary history of the data given (in this case a gene tree) it may not be representative of the species as a whole (species tree). Incongruences can arise due to external factors not considered such as selection and admixture. With the inclusion of many more loci, one can outweigh the noise introduced by these factors. By outweighing this noise, resultant topologies are more likely to present the true species tree, allowing species relationships to be more accurately inferred (Rokas et al. 2003). The polar and brown bears present an example of how external factors, in this case ancestral admixture, can complicate species inferences. When considering the exclusively maternally inherited mitochondrial genome, polar bears fall within brown bear diversity. Without looking at any other genetic data, this result would lead one to believe they are a single species despite large morphological and ecological differences. However, in contrast to the mitochondrial genome, when investigating nuclear genomes, the polar and brown bears were shown the be distinct lineages (Hailer et al. 2012). One possible explanation for this discrepancy is that an ancient hybridisation event introduced the brown bear mitochondrial genome into the polar bear gene pool which then became fixed within modern polar bears either due to genetic drift or some form of selection. Follow up studies, comparing nuclear genomes from polar and brown bears, have confirmed this admixture (Cahill et al. 2013).

1.5 Ancient DNA (aDNA)

Many fields involving the study of DNA have benefited greatly from the invention of NGS, the field of ancient DNA and palaeogenomics is arguably one of the fields that has benefitted most (Leonardi et al. 2017). Ancient DNA is a valuable tool within the field of evolutionary biology and allows one to observe changes in genetic diversity through real-time. It has been successfully implemented in studies to: infer relationships between evolutionary and environmental events in populations, resolve complex evolutionary relationships between species and even provide calibrations for molecular clock analyses (Shapiro and Hofreiter 2014).

Initially limited to short fragments of mitochondrial DNA (Shapiro and Hofreiter 2014), high-throughput NGS has allowed the sequencing of nuclear genomes from a number of ancient specimens (Miller et al. 2008; Palkopoulou et al. 2015; Meyer et al. 2012), making the above mentioned inferences much more powerful and reliable.

With the advent of NGS, many of the problems faced by PCR and Sanger sequencing based methods were overcome. These problems arise as ancient DNA presents a number of difficulties not commonly found within the study of modern DNA. Once a living cell is deceased, all DNA repair mechanisms cease to function, leading to the degradation of the DNA inside (Hofreiter et al. 2001). This degradation process leads to most fragments of ancient DNA being shorter than 100bp and damage to the nucleotides themselves, most commonly being deamination of cytosines to uracils (Dabney et al. 2013b). Furthermore, the specimen gets flooded with exogenous contaminant DNA in the form of environmental and microbial DNA (Willerslev and Cooper 2005). Authenticity of ancient DNA sequences is therefore very important, especially for ancient human specimens, requiring aDNA methods to have quality control measures. Multiple repetitions are often performed in order to detect whether the amplicon is authentic or contamination. This makes for a rather labour intensive process (Hofreiter et al. 2015). While DNA damage and contamination are hindrances in the study of aDNA with PCR and Sanger sequencing, there are ways around these problems. In contrast, the short fragment lengths of recovered DNA cannot be overcome and is the most limiting factor. The problem arises from the need for the primer binding sites to occur within the fragment itself. Primers are generally about 20bp each meaning that fragments with a length of 40bp or less are already unusable in downstream analyses. In order to overcome this and get meaningful information, most studies focus on the amplification of fragments of at least 100 bp. However, this results in the removal of the majority of DNA molecules in the extract from even being considered (Knapp and Hofreiter 2010).

The ability of NGS platforms to sequence very short fragments of DNA along with the sensitivity one gets from the number of sequences obtained are major factors that led to the

successful implementation of NGS into the field of ancient DNA. In order to take advantage of these, a number of new laboratory techniques have been developed. New extraction techniques allow for the isolation of the shortest fragments within an extract (Dabney et al. 2013a), and modified library building techniques can produce library molecules from these short fragments despite the presence of DNA damage (Gansauge and Meyer 2013). Furthermore, the sequencing of many independent fragments instead of a single clonally amplified fragment allows for investigation into the authenticity of the sequences. This comes in the form of aDNA damage patterns as C-T transitions are most commonly found at the single stranded overhangs at the ends of reads (Briggs et al. 2007; Brotherton et al. 2007). Moreover, the ability to sequence multiple fragments at once from the same extract means that a lot less DNA extract and therefore sample is required which is especially useful for precious samples with limited amounts of material.

1.6 Data analyses

With the advent of NGS came a necessity for new analysis techniques to deal with the new quantity and types of data being produced. The number of DNA molecules sequenced is orders of magnitude larger than what was achievable using Sanger sequencing. This meant that standard approaches could not be implemented. This is especially true for shotgun sequencing as all DNA molecules within a DNA extract are sequenced. To deal with this new influx of data, a large number of bioinformatic tools and software have been developed (Koboldt et al. 2013). The typical computational analysis for NGS data processing can generally be broken down into these major categories: removal of adapter sequences and low quality reads, mapping the data to a reference or a *de novo* assembly, and subsequent analysis of the processed data e.g. SNP calling, population genomic analyses, phylogenetic analyses, comparative genomic analyses (Grada and Weinbrecht 2013), and many others.

1.6.1 *De novo* assemblies

While lower costs and time requirements for data production were revolutionary, NGS was still limited by its maximal read length, preventing the use of NGS data alone to produce *de novo* assemblies of large, complex genomes. Assemblies using these short reads face difficulties as the reads do not span complete genes leading to incomplete gene assemblies and making genome annotations unreliable. Furthermore, repeats throughout the genome also cause trouble as many reads will align non-exclusively throughout the genome leading to very fragmented assemblies that prevent the analysis of structural variants. However, in 2010, the first *de novo* assembly using just Illumina short reads was produced for the Giant panda (Li et al. 2010a). This was achieved through the construction of specialised mate paired sequencing libraries with varying insert sizes of about 150 base pairs (bp), 500 bp, 2 kb, 5 kb and 10 kb. This project not only introduced a new laboratory method but also new software (Li et al. 2010b) to assemble the reads obtained from fragments of varying insert sizes produced using this method. This achievement opened up the use of short read NGS data alone for *de novo* assemblies at affordable prices. This same mate-paired library construction technique was used as a part of this thesis for the assembly of the high quality, striped hyena genome (Chapter 4).

1.6.2 Mapping assemblies

Once the initial *de novo* assembly has been completed, it is possible to perform whole genome re-sequencing (mapping assembly) of additional individuals from the same or closely related species for a fraction of the price of a *de novo* assembly. This can be useful for the identification of polymorphisms between individuals and population analyses. In contrast to *de novo* assembly, which assembles short reads into contiguous stretches of sequence information without a guide, mapping works by aligning reads against an existing reference sequence to generate an assembly that is similar but not necessarily identical to the original reference sequence. There are a lot of mapping assemblers but the most commonly used short read

mapping software are BWA (Li and Durbin 2009) and Bowtie (Langmead and Salzberg 2012; Langmead et al. 2009).

1.6.2.1 Low coverage mapping

By utilising a mapping approach and specialised downstream analyses, even low coverage re-sequencing genomes (<10x) are now amenable to a wide range analyses, further decreasing costs of large-scale genomic projects. One tool that allows for the analysis of low coverage genomes is ANGSD (Korneliussen et al. 2014). Instead of directly calling genotypes, which requires a high certainty and therefore high coverage, ANGSD can take into account the uncertainty of the data caused by coverage and uses genotype likelihood methods for downstream analyses. These techniques have been shown to be especially useful for population genomic analyses (Foote et al. 2016; Qiu et al. 2015; Laine et al. 2016). The biggest advantage of using these kinds of analyses is that one can sequence a greater number of individuals to lower coverage while still producing robust population level results.

1.6.2.2 Iterative mapping

Alternatively, when neither large insert library molecules are available for *de novo* assembly or a suitably close relative is available as a reference for a complete mapping assembly, iterative mapping is a good alternative and is becoming a more widely spread tool (Mitchell et al. 2014, 2016; Kehlmaier et al. 2017) and was used in all three case studies of this thesis. Iterative mapping functions essentially as a reference guided *de novo* assembly using an assembly of a close relative as an initial bait reference sequence. Reads are mapped to the initial bait reference sequence and mapped reads are then converted into a consensus sequence. This consensus sequence is then used as the reference for the next iteration. The iterative process is repeated until either the assembly is complete or no more new reads can be mapped. Iterative mapping is currently computationally unfeasible to run for large genomic datasets but is more than capable

of assembling smaller stretches of DNA such as complete mitochondrial genomes (Hahn et al. 2013).

1.7 Thesis outline

The aims of this thesis revolve around the production and analysis of next generation sequencing data through shotgun sequencing for use in both species level and population level comparisons to better understand the evolutionary history and current relationships of the three focus species. On top of the application of NGS, studies in the following case study chapters (Chapters 2, 3 and 4) all have a secondary common theme between them; the lack of a sequenced close relative on both the mitochondrial and nuclear genomic levels. This lack of closely related reference sequences required the utilisation of iterative mapping for the mitochondrial genomes and a *de novo* assembly in the case of the striped hyena nuclear genome. Following these assemblies, a number of analysis tools were implemented including phylogenetic analyses to infer species level relationships in the case of the bat eared fox (Chapter 2) and the *Macrauchenia* (Chapter 3) and population genomic, demographic and genetic diversity analyses in the case of the brown hyena (Chapter 4). Through these analyses, this thesis was able to gain a better understanding into the evolution of the species of interest by either resolving the closest living relatives of the species (Chapters 2 and 3) or by defining population structure within a species and the demographic history of the species (Chapter 4).

Chapter 2 involves the generation of the first bat eared fox (*Otocyon megalotis*) mitochondrial genome, which was used to determine its relationship to other species within the canidae family. The bat eared fox was previously estimated to have diverged from its closest living relative approximately eight million years ago. Even though this divergence date does not seem extremely deep, when coupled with the accelerated rate of evolution in the mitochondria it was deep enough that it was not suitable for a simple mapping assembly and required iterative mapping. Phylogenetic analyses using the reconstructed mitochondrial genome were able to confidently place the bat eared fox as sister to the clade consisting of the raccoon dog and true

foxes in good agreement with previous studies based on short fragments of mitochondrial and nuclear genes (Lindblad-Toh et al. 2005).

The *Macrauchenia* mitogenomic sequence described in Chapter 3 had a similar problem to the bat eared fox as no close relative was available but with the added difficulties of being comprised of ancient DNA and having a markedly deeper divergence time (almost 10 fold) from its closest living relative. Default iterative mapping parameters are optimised for the use of modern DNA which is devoid of contamination and has relatively long fragment lengths. This meant that iterative mapping protocols needed to be rigorously tested and optimised for use with ancient DNA when considering very deeply diverged reference sequences. Parameter testing resulted in the use of multiple mismatch values and initial bait reference sequences, in order to ensure high quality of the final assembly. Phylogenetic analyses of the resultant near complete mitochondrial genome placed the species as sister to all living Perissodactyla, with a divergence date of approximately 66 mya.

The final case study chapter in this thesis, Chapter 4, involved the study of brown hyena on both mitochondrial and nuclear genomic levels. Similar to the above mentioned chapters, the lack of a published brown hyena mitochondrial genome necessitated the use of iterative mapping to reconstruct the genome. On top of this, a *de novo* assembly of a striped hyena nuclear genome was conducted using Illumina short reads and was used as a reference for both high and low coverage brown hyena nuclear genome mapping assemblies. Analyses of low coverage genomes originating from wild individuals across the brown hyena's range of southern Africa for signs of population structure showed that there are four separate populations within the dataset. This study also investigated a single high coverage brown hyena nuclear genome from a captive bred individual for genomic diversity. This analysis found very low levels of genomic diversity, in agreement with mitochondrial diversity estimates but at the same time, surprisingly, an absence of detectable signs of inbreeding.

Finally, Chapter 5 details the general conclusions of this thesis, highlighting the wider implications in evolutionary and conservation genetics and also the methods used to analyse it. Moreover, technical advances in both sequencing technologies and data analyses and the future impact NGS can have within the field of evolutionary biology are discussed.

Chapter 2: Article I

Complete mitochondrial genome of a bat eared fox (*Otocyon megalotis*), along with phylogenetic considerations*

Michael Westbury, Fredrik Dalerum, Karin Norén, Michael Hofreiter

*This manuscript was published in "Mitochondrial DNA part B" in 2017 and can be found online at http://dx.doi.org/10.1080/23802359.2017.1331325

2.1 Abstract

The bat eared fox, *Otocyon megalotis,* is the only member of its genus and is thought to occupy a basal position within the dog family. These factors can lead to challenges in complete mitochondrial reconstructions and accurate phylogenetic positioning. Here we present the first complete mitochondrial genome of the bat eared fox recovered using shotgun sequencing and iterative mapping to three distantly related species. Phylogenetic analyses placed the bat eared fox basal in the Canidae family within the clade including true foxes (*Vulpes*) and the raccoon dog (*Nyctereutes*) with high support values. This position is in good agreement with previously published results based on short fragments of mitochondrial and nuclear genes therefore adding more support to the basal positioning of the bat eared fox within Canidae.

2.2 Main text

The bat eared fox (*Otocyon megalotis*) is a small member of the Canidae family and the only species of the genus *Otocyon.* It occurs in two allopatric populations across Africa (Clark 2005) and is considered a basal canid species (Sillero-Zubiri and Macdonald 2004). Studies using short mitochondrial and nuclear genes support a basal placement of the bat eared fox within Canidae, as sister group to the clade including true foxes (*Vulpes)* and the Raccoon dog

(*Nyctereutes procyonoides*) (Lindblad-Toh et al. 2005). However, despite previous genetics studies, the complete mitochondrial genome of the bat eared fox has so far not been published.

Our bat eared fox sample was captured on Benfontein game farm outside of Kimberley, central South Africa (28°99′ S, 24°81′ E, e.g., (le Roux et al. 2014)) under permits from the animal care and use committee of the University of Pretoria (EC031-07) and from the provincial government in the Northern Cape (FAUNA 846/2009, FAUNA 847/2009). DNA was extracted using a Zymo genomic DNA clean and concentrator extraction kit, built into Nextera Illumina sequencing libraries and sequenced on an Illumina Nextseq 500 at the University of Potsdam, Germany. We trimmed raw reads using Cutadapt v1.4 (Martin 2011), merged overlapping fragments using FLASH v1.2.10 (Magoč and Salzberg 2011) and removed duplicate reads with Prinseq (Schmieder and Edwards 2011).

We undertook iterative mapping using MITObim v1.8 (Hahn et al. 2013) with default parameters apart from mismatch value, where we used 3%. Three independent runs were performed using different bait sequences from three species; *Canis lupus* (Genbank accession KC461238.1), *Vulpes vulpes* (Genbank accession JN711443.1) and *Urocyon littoralis* (Genbank accession KP128962.1). Consensus sequences were called using a minimum read coverage of 10x and a 75% base call threshold in Geneious v9.0.5 (Kearse et al. 2012). Automatic annotation was performed using MITOS (Bernt et al. 2013). We aligned our sequence with representatives of Carnivora using Mafftv7.271 (Katoh and Standley 2013). We constructed a phylogenetic tree using BEAST version 1.8.4 (Drummond et al. 2012) specifying GTR I+G as the substitution model, as determined using Jmodeltest v2.1.7 (Posada 2008), and a Yule speciation model (Yule 1925; Gernhard 2008).

Our bat eared fox mitochondrial sequence (Genbank accession KY776502) contains 16,431 bp. The genomes produced were identical regardless of starting bait reference. Our bat eared fox mitochondrial assembly contained all 13 protein-coding genes with expected open reading frames, 22 transfer RNA genes, and 2 ribosomal RNA genes found within a typical

vertebrate mitochondrial genome. Phylogenetic analyses found further support for a basal placement of the bat eared fox within the Canidae. *Otocyon* was placed with high confidence (posterior probabilities >0.99) as sister to the clade containing both the raccoon dog (*Nyctereutes*) and true foxes (*Vulpes*) (Fig 2.1). This result is consistent with previous conclusions based on short regions of mitochondrial and nuclear genes (Lindblad-Toh et al. 2005).

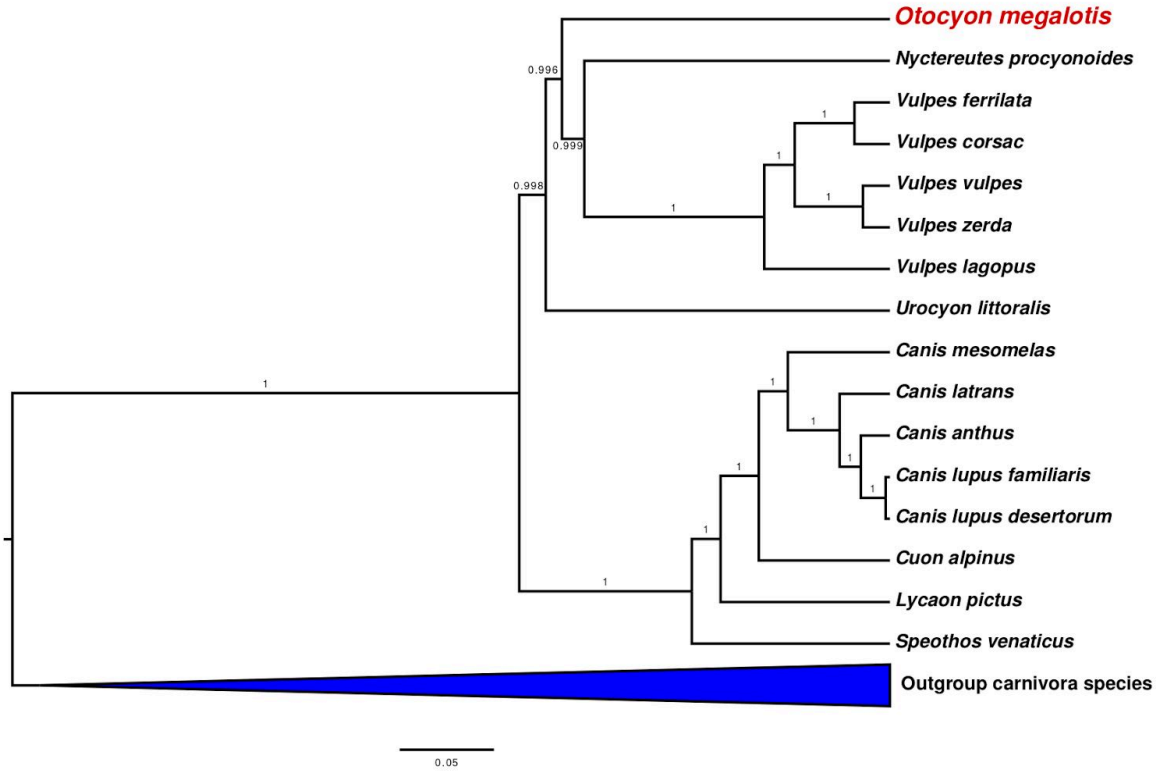


Figure 2.1: Bayesian tree showing the phylogenetic positioning of *Otocyon megalotis* within the Canidae and other clades within Carnivora. Numbers on branches represent posterior probabilities.

Acknowledgements:

We would like to thank Armanda Bastos for the use of her laboratory facilities at the University of Pretoria, South Africa.

Author contributions:

Lab work was carried out by M.W. Data analyses were carried out by M.W. Locating and sampling of our specimen was performed by F.D and K.N. Manuscript preparations and coordinations were performed by M.W and M.H. All contributing authors read and agreed to the final manuscript

Chapter 3: Article II

A mitogenomic timetree for Darwin‘s enigmatic South American mammal *Macrauchenia patachonica**

Michael Westbury, Sina Baleka, Axel Barlow, Stefanie Hartmann, Johanna L. A. Paijmans, Alejandro Kramarz, Analía M. Forasiepi, Mariano Bond, Javier N. Gelfo, Marcelo A. Reguero, Patricio López Mendoza, Matias Taglioretti, Fernando Scaglia, Andrés Rinderknecht, Washington Jones, Francisco Mena, Guillaume Billet, Christian de Muizon, José Luis Aguilar, Ross D. E. MacPhee, Michael Hofreiter

*This manuscript was published in “Nature communications” in 2017 under the doi 10.1038/ncomms15951

3.1 Abstract

The unusual mix of morphological traits displayed by extinct South American native ungulates (SANUs) confounded both Charles Darwin, who first discovered them, and Richard Owen, who tried to resolve their relationships. Here we report an almost complete mitochondrial genome for the litoptern *Macrauchenia*. Our dated phylogenetic tree places *Macrauchenia* as sister to Perissodactyla, but close to the radiation of major lineages within Laurasiatheria. This position is consistent with a divergence estimate of ~66 Ma (95% credibility interval, 56.64-77.83 Ma) obtained for the split between *Macrauchenia* and other Panperissodactyla. Combined with their morphological distinctiveness, this evidence supports the positioning of Litopterna (possibly in company with other SANU groups) as a separate order within Laurasiatheria. We also show that, when using strict criteria, extinct taxa marked by deep divergence times and a lack of close living relatives may still be amenable to palaeogenomic analysis through iterative mapping against more distant relatives.

3.2 Introduction

It is now well accepted that ancient DNA (aDNA) is a valuable tool for uncovering phylogenetic relationships of extinct animals (Hofreiter et al. 2001). However, in order to obtain correct DNA sequences from ancient remains, it is usual practice to utilise sequences from a close extant relative to produce primer sequences for PCR (Hofreiter et al. 2001; Pääbo et al. 1989), baits for hybridisation capture (Westbury et al. 2016; Carpenter et al. 2013), or reference frameworks for mapping shotgun data (Meyer et al. 2012). In theory, reconstructing an ancient genome *de novo* can be undertaken without relying on a close relative's DNA for guidance, but due to contaminant DNA and low average fragment lengths, *de novo* assembly is generally considered not computationally feasible (Dabney et al. 2013b; Hofreiter et al. 2015). These difficulties are compounded when targeted extinct species lived in tropical or subtropical regions, where aDNA preservation is characteristically poor (Hofreiter et al. 2015; Welker et al. 2015) —as in the case of the enigmatic South American mammal *Macrauchenia patachonica*.

Macrauchenia patachonica was among the last of the Litopterna, an endemic order whose fossil record extends from the Paleocene to the end of the Pleistocene and includes some 50 described genera. Over the past 180 years, remains of *Macrauchenia* and its close allies have been found in Quaternary deposits in various parts of the continent, first and most notably by Charles Darwin in 1834, near Puerto San Julián in southern Patagonia (Appendix A: Supplementary Note 1). However, neither Darwin nor Richard Owen, who described the species in 1838 (Owen and Darwin 1840), were able to place this taxon securely among placentals (Herbert 1980). Owen, who had only a few limb bones and vertebrae to work with, originally described *Macrauchenia* as a form 'transitional' between camelids and other ruminant artiodactyls and so-called Pachydermata, a polyphyletic miscellany that included elephants, horses, hippos, and hyraxes. Owen's analysis indicated that *Macrauchenia* was, at least in terms of grade, an ungulate of some sort, but it was otherwise inconclusive. Similar uncertainties have marked all subsequent morphology-based efforts to ascertain the affinities not only of Litopterna

but also other SANU orders (Simpson 1945; McKenna 1975; Soria 2001; O'Leary et al. 2013) (Appendix A: Supplementary Note 1).

The central problem in SANU systematics has long been how to evaluate the remarkable level of similarity individual orders display to various non-SANU taxa from other parts of the world. Unsurprisingly, different studies have reached very different conclusions. One such study (O'Leary et al. 2013), utilising a large set of morphological characters, found that Litopterna belonged within Pan-Euungulata, but another SANU order, Notoungulata, grouped with Afrotheria, indicating that SANUs were not monophyletic. By contrast, utilising protein (collagen) sequence information, two recently published molecular studies (Welker et al. 2015; Buckley 2015) found that litopterns as well as notoungulates formed a monophyletic unit that shared more recent common ancestry with Perissodactyla than with any other extant placental group (Beck and Lee 2014) (justifying recognition of the new unranked taxon Panperissodactyla (Welker et al. 2015)).

Although the collagen (I) evidence for the position of litopterns and notoungulates is informative, given the historical instability of SANU systematics, it is important to corroborate the proteomic results with additional, preferably molecular sources of evidence. However, to date, attempts to use standard aDNA methodologies to collect genetic material from specimens from low-latitude localities have been largely unsuccessful (Welker et al. 2015). A promising new approach is shotgun sequencing applied with iterative mapping, which functions in a way similar to reference-assisted *de novo* assembly (Hahn et al. 2013), bypassing the need for a close relative as reference. Initially, reads are mapped to a specified bait reference sequence. As analysis proceeds, a consensus of mapped reads becomes the new reference for each following iteration until no new reads are found to map. Using this approach and utilising a set of strict parameters, we report the successful collection of mitogenomic data from a South American native ungulate. Phylogenetic analyses place the *Macrauchenia* as a sister taxon to all living Perissodactyla, with the origin of Panperissodactlya at ~66Ma, successfully demonstrating that

even taxa marked by deep divergence times with no close living relatives are amenable to palaeogenomic analysis.

3.3 Results

3.3.1 Sample screening

We extracted DNA from 6 *Macrauchenia* and 11 *Toxodon* bone samples obtained from various sites across South America (Fig. 3.1, Appendix A: Supplementary Table 1), using a DNA extraction method specifically developed for recovering short fragments typical of aDNA (Dabney et al. 2013a). We converted the resulting extracts into Illumina libraries applying a single-strand library building approach, also specifically developed for aDNA (Gansauge and Meyer 2013) and carried out low-level sequencing (ranging from 2 to 20 million raw reads) to investigate endogenous DNA content. Only a 2nd phalanx, from Bano Nuevo-1 Cave (Coyhaique, Chile, Appendix A: Supplementary Fig. 1) and here coded MAC002, yielded a high number of reads mapping to the horse and rhinoceros nuclear genomes (1.9% and 2.8%, respectively, as opposed to <1% for all others; Appendix A: Supplementary Table 2). This result led us to conduct further shotgun sequencing of this individual for a total of approximately 69 million, paired-end 75 bp reads and 43 million after quality controls (see Methods).

3.3.2 Validation of the iterative mapping approach using MITObim

The lack of a suitable reference mitochondrial sequence necessitated an iterative mapping approach. We used the iterative mapping software package MITObim (Hahn et al. 2013), which has been used successfully for mitochondrial reconstructions using modern DNA. Reference sequences for the following four species (with GenBank accession numbers) were selected: guanaco (*Lama guanicoe*, NC_011822.1), rhinoceros (*Ceratotherium simum*, Y07726.1), horse (*Equus caballus*, HQ439492.1), and tapir (*Tapirus indicus*, KJ417810.1). MITObim reconstructions using the default MITObim mismatch value (15%) and consensus calling method

resulted in complete mitochondrial sequences being recovered for each bait reference used, but many discrepancies were evident among the consensus sequences (Appendix A: Supplementary Table 3). Upon visual inspection of the assembly, random read mapping was clearly visible, because large differences between reads mapping to the same region of the bait sequence could be seen. This led us to perform a software validation using Pleistocene cave hyena (*Crocuta crocuta spelaea*) DNA sequences. We compared consensus sequences produced using iterative mapping to distant references with one produced using direct mapping to a cave hyena mitochondrial genome. When using 80% of the average read coverage as the minimum coverage threshold for the consensus sequence base calling, regardless of the mismatch value or reference sequence we tried, the sequence produced with MITObim matched perfectly with the presumably correct consensus sequence produced by direct mapping to the reference cave hyena mitochondrial genome (NC_020670.1; Appendix A: Supplementary Table 4). This result led us to conclude that MITObim was a suitable platform for reconstructing the mitochondrial genome of *Macrauchenia* if appropriately stringent mismatch values and consensus calling parameters were implemented.

3.3.3 Mitochondrial genome reconstruction of MAC002

When implementing the 80% average coverage minimum cutoff value as predicted using the cave hyena (Appendix A: Supplementary Table 4), the initial strict value of 0% mismatch between reference and mapping reads produced four consensus sequences displaying exact identity to each other regardless of whether the guanaco, rhinoceros, horse or tapir bait reference was used. As using 0% mismatch value only recovered between ~10 and 25% of the complete mitochondrial genome (Appendix A: Supplementary Table 5), we relaxed the mismatch value in 1% increments in order to recover more of the genome. At a mismatch value of 7%, we noted that a number of sites could not be called unambiguously since discrepancies arose between consensus sequences obtained using different references. We therefore considered 6% as the upper threshold mismatch value as these disagreements may have arisen due to random read mappings.

Regardless of the implemented mismatch value, when using mismatch values from 0% to 6%, all mappings resulted in alignments of similar average depth (~ 40x) and identical sequences, the only difference being in the varying amounts of mitogenome coverage (Appendix A: Supplementary Table 5). The relationship between assembly completeness and mismatch value was, however, not entirely predictable, with some regions of the mitochondrion being covered in assemblies with low mismatch values that were not recovered at higher mismatch values (Appendix A: Supplementary Table 5). The guanaco (artiodactyl) reference produced sequences compatible with those produced using the perissodactyl references, ruling out the possibility of ascertainment biases based on phylogenetic relatedness. In total, we recovered 13,269 basepairs (79.1%) of the *Macrauchenia* mitogenome. Most remaining sites were also covered, but due to our strict consensus calling parameters (see Methods) involving a minimum coverage threshold of 80% of the average, many of these were considered as missing data. Stretches of missing data predominantly occur in the cytb, cox2 and nad6 genes (Appendix A: Supplementary Fig. 1).

3.3.4 *Macrauchenia* mitochondrial sequence validation

MITOS (Bernt et al. 2013) automated annotation confirmed the presence of most tRNA sequences, apart from glutamic acid and serine, all protein coding genes, and both ribosomal RNAs (Appendix A: Supplementary Fig. 1). Amino acid translations of manually predicted protein coding genes showed no indication of premature stop codons. We also conducted an analysis of pairwise sequence identity between our reconstructed *Macrauchenia* sequence and all reference sequences used for mapping as well as the human mitochondrial DNA sequence as outgroup. This analysis showed all regions of the reconstructed *Macrauchenia* mitogenome sequence as approximately equidistant to all reference sequences, with the human mitogenome having a consistently lower pairwise identity throughout (Fig. 3.2). No regions showed a large increase in pairwise identity to the human sequence, indicating that no regions were constructed from human contaminant DNA. Mapdamage (Jónsson et al. 2013) analysis of the mapped reads showed characteristic patterns of DNA damage and short read length distributions indicative of

aDNA (Appendix A: Supplementary Figs 2 and 3). Retrospective mapping of reads from the other analysed *Toxodon* and *Macrauchenia* samples to our *Macrauchenia* mitochondrial genome sequence produced either very few or no hits. This result suggests that none of the other samples contained detectable quantities of endogenous *Macrauchenia* or *Toxodon* mitochondrial DNA.

3.3.5 Phylogenetic reconstruction

We used both Maximum Likelihood and Bayesian inference approaches for phylogenetic tree reconstruction in order to determine the phylogenetic position of *Macrauchenia*. Both approaches recovered *Macrauchenia* as a sister taxon to the order Perissodactyla, represented by the genera *Hippidion*, *Equus*, *Dicerorhinus*, *Tapirus, Rhinoceros, Ceratotherium, Coelodonta* and *Diceros* (Appendix A: Supplementary Figs 4 and 5). We undertook a molecular dating analysis of the superorder Laurasiatheria with Eulipotyphla specified as outgroup, using four fossil calibrations. The calibrated nodes were: the basal divergence of extant lineages of Laurasiatheria (based on the fossil *Protictis haydenianus* (Benton et al. 2009)), the basal divergence of extant lineages of Bovidae (based on the fossil *Eotragus noyei* (Benton et al. 2009)), the basal divergence of extant lineages of Perissodactyla (based on the fossil *Sifrhippus sandrae* (Froehlich 2002; Secord et al. 2012)) and the basal divergence of extant lineages of Carnivora (based on the fossil *Hesperocyon gregarius* (Meredith et al. 2011)). This analysis produced an estimated Panperissodactyla divergence time of 66.15 Ma with a 95% CI of 56.64-77.83 Ma (Fig. 3.3, Appendix A: Supplementary Table 6). The resulting tree is in good agreement with divergence estimates of other major lineages within Laurasiatheria obtained using larger nuclear DNA data sets (Meredith et al. 2011) (Appendix A: Supplementary Table 7). We additionally investigated the potential variability in divergence time estimates among the individual calibrations used by reanalysing the dataset using each calibration independently. The mean estimated age for the basal divergence of Panperrissodactyla was broadly similar using the Carnivora, Laurasiatheria, or Bovidae calibrations. The mean age obtained using the Perissodactyla calibration was older than that produced using any of the other three calibrations, but the 95% credibility intervals generated using these different calibrations all overlapped

except in the case of the Perrissodactyla and Carnivora calibrations (Appendix A: Supplementary Table 6).

3.4 Discussion

We successfully recovered a nearly complete mitochondrial genome for the extinct South American native ungulate, *Macrauchenia.* This allowed us to confidently place *Macrauchenia*, and thus Litopterna, as the sister group of crown Perissodactyla, in agreement with collagen sequences obtained by proteomic analyses (Welker et al. 2015; Buckley 2015), but with a better-resolved divergence time (~66 Ma, 95% CI of 56.64-77.83 Ma). Our results confirm that Litopterna and Perissodactyla together form (non-exclusively) the unranked taxon Panperissodactyla (Welker et al. 2015). Didolodontidae, a 'condylarth' group with North American affinities and thought to be directly ancestral to Litopterna, were present in South America by the earliest part of the Palaeocene but (as far as we now know) not earlier (Cifelli 1993) (Appendix A: Supplementary Note 1). Since there are no accepted litopterms in the Paleogene record of North America, and no crown Perissodactyla in that of South America, stem panperissodactyls most likely dispersed into the latter continent before the divergence of Lipoterna. Thus, fossils, palaeobiography, and molecules are mutually concordant in suggesting that early divergences within South American Panperissodactyla probably took place very early in the Cenozoic (Welker et al. 2015), perhaps immediately subsequent to the K/Pg transition.

Macrauchenia samples used for this study came from the southern cone of South America (Fig. 3.1). Although this portion of the continent is more temperate than the equatorial region, we nonetheless expected DNA preservation to be poor (Hofreiter et al. 2015). This is vividly illustrated by the lack of detectable endogenous *Macrauchenia* and *Toxodon* DNA in all but the southernmost *Macrauchenia* sample tested in this study. This result underlines the importance of macro- and microenvironmental factors, with the latter being mostly unknown, affecting DNA survival. A second challenge in analysing the relationships of species such as *Macrauchenia patachonica* comes from the lack of closely related extant relatives which can be

used for sequence authentication and detecting potential contamination. This problem arises as the degree to which the genome of an extinct species can be mapped successfully against a corresponding reference genome correlates inversely with phylogenetic distance, which is especially true for the mitochondrial genome with its comparatively rapid substitution rate in vertebrates (Prüfer et al. 2010; Shapiro and Hofreiter 2014).

On the other hand, the small size of the mitochondrial genome simplifies the assembly of fossil sequences using *de novo* methods. Such an approach is likely unsuitable for the much larger nuclear genome. As we demonstrate here, iterative mapping strategies permit recovery of nearly complete mitochondrial genomes even from extinct species with only distant living relatives available for comparisons, thereby also permitting resolution of their phylogenetic relationships (Mitchell et al. 2014, 2016; Kehlmaier et al. 2017). Caution is, however, required when defining algorithmic parameters or consensus-building steps (Appendix A: Supplementary Table 4) because of the danger of inferring incorrect sequences. The use of a single bait reference sequence may also introduce some biases as final consensus sequences can differ depending on the starting reference bait sequence (Appendix A: Supplementary Table 3). Further, random read mapping of short contaminating DNA sequences is especially likely with aDNA datasets (Skoglund et al. 2014) when using less stringent parameters. To control for these and other issues, we adopted a highly conservative approach involving strict minimum coverage and threshold values for consensus construction, a range of mismatch values and the combination of four different bait references, including one phylogenetically more distant sequence. Our method may result in the loss of some correctly sequenced nucleotide sites, but we are confident that all reconstructed positions reported here for *Macrauchenia* are authentic. We suggest that similarly stringent approaches are implemented in future efforts to reconstruct the mitogenomes of extinct organisms without a close living relative, in order to avoid partially incorrect sequences that may occur with more relaxed approaches.

Although progress in sorting out the molecular palaeontology of Darwin's peculiar mammals is being made, uncertainties remain. The SANU orders Astrapotheria, Pyrotheria, and

Xenungulata still lack firm placement within Placentalia. Like Notoungulata, Xenungulata has recently been grouped with Afrotheria on morphological grounds (O'Leary et al. 2013), implying that this clade is not part of Panperissodactyla. Pyrotheres and xenungulates, never very diverse, had already disappeared by the end of the Palaeogene (Simpson 1980), far beyond the present empirical reach of any molecular method, including collagen proteomics (Welker et al. 2015). Prospects may be better for the more diverse astrapotheres, which persisted into the Middle Miocene (Goillot et al. 2011). However, while we were unable to successfully recover DNA from samples of *Toxodon*, the results from our study underscore the reliability of the collagen results (Welker et al. 2015; Buckley 2015) and its use in phylogenetic analyses despite the fact that these approaches are methodologically quite different.

3.5 Methods

3.5.1 Samples

We carried out genetic analyses on 6 *Macrauchenia* and 11 *Toxodon* subfossils originating from various locations in the southern portion of South America (Fig. 3.1, Appendix A: Supplementary Table 1). A schematic overview of the methods can be seen in Appendix A: Supplementary Fig. 6.

3.5.2 DNA preparation

For all samples, approximately 50mg of bone was ground to powder using a mortar and pestle, and DNA was extracted following the protocol described in Dabney *et al*. (Dabney et al. 2013a). Initially, bone powder was rotated at 37ºC in 1ml extraction buffer (0.45M EDTA, 0.25mg/ml proteinase K, pH 8.0). The remaining bone powder was then pelleted by centrifugation at maximum speed (16,000 x g). The supernatant was removed and added to 13ml of binding buffer (5M guanidine hydrochloride, 40% isopropanol, 0.05% Tween-20 and 90mM sodium acetate (pH 5.2)), then passed through a MinElute silica spin column (Qiagen) and

centrifuged at 450 x g until the solution had completely passed through the spin column. The silica membrane was then washed twice by adding 750µl PE buffer (Qiagen) to the column, centrifuging at 3,300 x g and discarding the flow through. Finally, DNA was eluted from the spin column by adding 25µl TET buffer followed by an incubation of fives minutes and centrifuged at maximum speed for one minute. An additional 50mg of bone powder from six of these samples, MAC001-004,TOX008 and TOX009 (details in Appendix A: Supplementary Table 8), was pretreated with 1ml of 0.5% bleach (sodium hypochlorite) for 15 minutes prior to DNA extraction, in an attempt to increase endogenous content (Korlević et al. 2015). All DNA extracts were converted to barcoded Illumina sequencing libraries using a method based on single stranded DNA specifically developed for highly degraded ancient samples (Meyer et al. 2012; Gansauge and Meyer 2013). Extracts initially underwent a uracil excision and DNA cleavage at abrasic sites step. 20µl of DNA extract was added to a solution containing 29µl water, 8µl Circligase buffer II (10x), 4µl $MnCl_2$ (50mM), 0.5µl Endonuclease VIII (10U/µl) and 0.5 Afu UDG (2U/µl). We then incubated this solution at 37ºC for 1 hour. Samples then underwent dephosphorylation and denaturation. 1µl of FastAP (1U) was added to the solution then incubated at 37ºC for 10 minutes followed by 95ºC for two minutes before being returned to room temperature. The first adapter was ligated by adding 32µl PEG-4000 (50%), 1µl adapter oligo CL78 (10µM) and 4µl Circligase II (100 U/µl) to the 43µl solution and incubated for one hour at 60ºC. 20µl of MyOne C1 beads were pelleted using a magnetic rack, the supernatant removed and the beads were washed twice using 500µl of bead binding buffer (1M NaCl, 10mM Tris-HCl (pH 8), 0.4mM EDTA (pH 8), 0.05% Tween and 0.5% SDS). The beads were then resuspended in 250µl bead binding buffer. Ligation products were then immobilized onto the beads. The adapter ligation solutions were incubated at 95ºC for one minute, before being cooled to room temperature and added to the bead buffer solution. The bead suspension was then rotated for 20 minutes at room temperature. The solution was then pelleted on a magnetic rack, the supernatant was removed and the beads were then washed once with wash buffer A (0.1M NaCl, 10mM Tris-HCl, 1mM EDTA, 0.05% Tween and 0.5% SDS) followed by a wash with wash buffer B (0.1M NaCl, 10mM Tris-HCl, 1mM EDTA and 0.05% Tween). The beads were pelleted and the wash buffer was discarded. Samples then underwent primer annealing and

extension. 5µl isothermal amplification buffer (10x), 0.5µl dNTP mix (25mM each), 1µl extension primer CL9 (100µM) and 40.5µl water was added to the pelleted beads followed by incubation at 65ºC for minutes before being cooled to room temperature. After incubation, 2µl Bst 2.0 polymerase (24 U) was added. The mixture was then incubated by increasing the temperature by 1ºC per minute from 15ºC to 37ºC with a final incubation of 5 minutes at 37ºC. Beads were then washed once in wash buffer A, once in stringency wash buffer (0.1% SDS and 0.1x SSC) with an incubation at 45ºC for four minutes and once in wash buffer B. Samples were then blunt end repaired. 10µl Buffer Tango (10x), 2.5µl Tween 20 (1%), 0.4µl dNTP (25mM each), 1µl T4 polymerase (5U) and 86.1µl water was added to the pelleted beads followed by a 15 minute incubation at 25ºC. Beads were again washed once in wash buffer A, once in stringency wash buffer with an incubation at 45ºC for four minutes and once in wash buffer B. Samples then had the second adapter ligated. 10µl T4 DNA ligase buffer (10x), 10µl PEG-4000 (50%), 2.5µl Tween (1%), 2µl double stranded adapter mixture (100µM), 2µl T4 DNA ligase (10U) and 73.5µl water was added to the beads, followed by a one hour incubation at room temperature. Beads were again washed once in wash buffer A, once in stringency wash buffer with an incubation at 45ºC for four minutes and once in wash buffer B. Sample bead pellets were then eluted by re-suspension in 25µl TET buffer followed by an incubation at 95ºC for one minute and pelleted using a magnetic rack. The supernatant contained the eluted library. Libraries were amplified and indexed by adding 10µl Accuprime Pfx reaction mix (10x), 4µl P7 indexing primer (10µM) 4µl P5 indexing primer (10µM), 24µl library, 1µl Accuprime Pfx polymerase (2.5U/µl) and 57µl water followed by a selected number of PCR cycles, involving denaturation for 15 seconds at 95 °C, annealing for 30 seconds at 60 °C and primer extension for one minute at 68 °C. Amplified libraries were cleaned up using a Minelute PCR purification kit following the manufacturer's protocol. Sequencing of library and extraction blanks were included to check for the presence of contamination.

3.5.3 Test sequencing and analysis

For each library, we sequenced approximately 2-20 million 75bp PE read pairs on an Illumina Nextseq 500 sequencing platform. Raw Illumina intensity data were demultiplexed and converted to nucleotide sequences using the Illumina software, bcl2fastq. Adapter sequences were trimmed from the raw reads, reads with lengths of 30bp or less were discarded using Cutadapt 1.4 (Martin 2011). Remaining trimmed reads were then merged using FLASH v1.2.10 (Magoč and Salzberg 2011). As *Macrauchenia* and *Toxodon* have been previously shown to be related to the order Perissodactyla (Welker et al. 2015), trimmed and merged reads were mapped to both the horse (GCA_000002305.1) and the rhinoceros (GCA_000283155.1) nuclear genomes using BWA 0.7.8 (Li and Durbin 2009) to evaluate the presence of endogenous DNA. Duplicate reads and reads of low mapping quality were then removed using SAMtools v0.1.19 (Li et al. 2009). Endogenous content was estimated by the fraction of merged reads that were successfully mapped to the reference nuclear genomes (Appendix A: Supplementary Table 2).

3.5.4 Further extractions of MAC002

As sample MAC002 showed potentially high endogenous DNA content (Appendix A: Supplementary Table 2), we further investigated the potential to increase endogenous DNA recovery, by extracting DNA from this sample a third time using the predigestion method proposed by Damgaard *et al*. (Damgaard et al. 2015), but utilising the extraction buffer and DNA purification method of Dabney *et al.* (Dabney et al. 2013a). Bone powder was mixed with 1ml of Dabney extraction buffer (0.45M EDTA, 0.25mg/ml proteinase K, pH 8.0) and incubated for one hour at 37°C. Samples were then centrifuged for one minute at maximum speed, the supernatant was removed, added to 13ml of binding buffer, and DNA was purified from it following the method described above. The remaining undigested bone powder was then re-extracted by suspension in 1ml of Dabney extraction buffer followed by an overnight incubation at 37°C. The bone powder was pelleted by 1 minute of at maximum speed and the extraction supernatant was

then subjected to the same extraction procedures as those described above. These extracts were converted into libraries, test sequenced, and analysed as described above.

3.5.5 Deep sequencing of MAC002

We resequenced all MAC002 libraries produced from the three different extraction methods on an Illumina Nextseq 500. Sample pre-treatment using either bleach or predigestion did not result in libraries with increased endogenous DNA content (Appendix A: Supplementary Table 2). To ensure that only high quality sequences were used for subsequent mitochondrial reconstruction, we processed the raw reads using stringent criteria. First, potential PCR duplicates were removed using fastuniq (Xu et al. 2012). Next, we trimmed adapter sequences and low quality bases from read ends, and discarded any reads less than 31bp, using Cutadapt 1.4 (Martin 2011). We then merged the trimmed PE reads using FLASH v1.2.10 (Magoč and Salzberg 2011) with a maximum difference allowed for merging of 0.1. We undertook all further analyses with these trimmed, merged reads (see Appendix A: Supplementary Table 9)

3.5.6 Mitochondrial reconstruction

We reconstructed the mitochondrial sequence of MAC002 using MITObim v1.8, a wrapper script for the Mira v4.0.2 (Chevreux et al. 1999) assembler. Direct reconstruction without prior mapping assembly with default parameters was implemented. We initially tested MITObim with the default mismatch value and using four different reference bait mitochondrial sequences, three from the order Perissodactyla and one from the order Artiodactyla. Mira output maf files were then converted to sam files and visualised using Geneious v9.0.5 (Kearse et al. 2012). Visual inspection revealed that MITObim assemblies generated using default parameters contained very large numbers of spuriously mapped reads or misassemblies, and that these regions were typically associated with either very high or very low coverage.

3.5.7 MITObim validation

In order to better evaluate and optimise the MITObim assembly, we tested the ability of MITObim to reconstruct the correct mitochondrial sequence of a Late Pleistocene cave hyena (*Crocuta crocuta spelaea*), an extinct taxon for which published mitogenome sequences are available that can be used to validate the MITObim assembly. These validation tests used cave hyena shotgun sequencing data with a similar predicted endogenous content and similar number of reads as MAC002. We iteratively mapped these reads to both the dog (*Canis lupus familiaris* GenBank accession: NC002008) and the brown bear (*Ursus arctos*, GenBank HQ685964) mitochondrial genome, as they are thought to have diverged at a similar phylogenetic time (~ 50 million years) from *Crocuta* as *Macrauchenia* from the Perissodactyla reference sequences used. We implemented mismatch values ranging from 0-12%. Output consensus sequences were then compared to one generated from the same raw reads mapped to a cave hyena mitochondrial genome (Genbank, NC_020670.1) using BWA 0.7.8 (Appendix A: Supplementary Table 4).

3.5.8 MAC002 mitochondrial genome reconstruction

We generated assemblies of the MAC002 mitochondrial genome using mismatch values of 0-6% in steps of 1% for each of the four reference sequences described above, resulting in a total of 28 assemblies. Visual examination of these assemblies revealed that both mismatch values and the reference used affected both the regional coverage and regional accumulation of spurious alignments for the assembly. We therefore processed the 28 assemblies into a single consensus sequence, with the aim of maximising total coverage information, while removing any incongruent sites among assemblies generated using different reference sequences or mismatch values, which may potentially result from analytical bias or contamination. This processing involved three stages of analysis: initial consolidation of each sequence read assembly into a consensus sequence, then consolidation of consensus sequences generated using different references and the same mismatch value into a mismatch value consensus sequence, and finally consolidation of mismatch value consensus sequences into a final consensus sequence. Initial

consensus sequences were generated using strict coverage filters, to ensure that any regions containing incorrect assemblies did not contribute to the final consensus sequence. We estimated the average read depth of coverage for the mitochondrion provided by the MAC002 data to be ~43x. Based on our previous tests using the cave hyena (Appendix A: Supplementary Table 4), we generated consensus sequences by applying a minimum of 34x coverage, representing around 80% of the mean coverage, and a maximum coverage of two times the mean read depth (86x). Nucleotide positions within this range were only included in the consensus sequence if a minimum of 95% of reads supported the same nucleotide, and otherwise entered as missing data.

Mismatch value consensus sequences were generated by aligning all consensus sequences (one for each reference used) corresponding to each mismatch value using Mafftv7.123b (Katoh and Standley 2013). A majority rule base call was applied to these alignments to produce a preliminary mismatch value consensus sequence, which was then aligned back to the original consensus sequences and visualised in MEGA6 (Tamura et al. 2013). Any nucleotide positions that were variable among consensus sequences were scored as missing data (N) in the mismatch value consensus.

The final consensus sequence was generated by aligning all seven mismatch value consensus sequences (one for each mismatch value). The final consensus sequence was then produced as described above, scoring any nucleotide positions that were variable among mismatch values consensus sequences as missing data.

3.5.9 Final sequence validation

The online automated mitochondrial annotation program MITOS (Bernt et al. 2013) was used to evaluate the orientation and positions of tRNAs and protein coding genes within our final consensus sequence. Protein coding genes were manually identified based on the horse mitochondrial sequence, extracted, and translated into their respective amino acid sequences using MEGA6 to check for premature stop codons.

We then calculated pairwise distances for the MAC002 sequence from the four references used for assembly, as well as the human mitochondrial sequence (Genbank accession J01415.2), along a sliding window. Following alignment, sites with missing data were removed manually in MEGA6 and pairwise distances calculated for windows of 500bp at 50bp intervals using a custom perl script. Finally, the trimmed and merged MAC002 reads were mapped back to our final consensus sequence using BWA (Li and Durbin 2009) and parsed using Samtools (Li et al. 2009), allowing investigation of aDNA damage patterns and read length distributions using mapdamage2.0.2-8 (Jónsson et al. 2013) (Appendix A: Supplementary Figs 2 and 3).

3.5.10 Retrospective mapping of other *Macrauchenia* and *Toxodon* samples

We retrospectively mapped reads from the other analyzed samples to our final *Macrauchenia* consensus sequence using BWA0.7.8 (Li and Durbin 2009), to better assess the presence of endogenous mitochondrial fragments in these samples.

3.5.11 Phylogenetic reconstruction

Our final *Macrauchenia* consensus sequence was aligned with 94 other mitochondrial sequences, including representatives from all major clades of the Laurasiatheria superorder (Appendix A: Supplementary Table 10). All sites containing missing data (N) were removed from the alignment manually, along with the control region, resulting in 12,997bp of aligned sequence. tRNA and gene positions within this alignment were determined manually for later partitioning analyses.

A maximum likelihood phylogenetic analysis was carried out using Raxml-HPC2 (Stamatakis 2014) on XSEDE. The appropriate partitioning scheme for all possible combinations of genes and tRNAs, and appropriate substitution models for each partition (GTR considering all possible combinations of invariant sites and gamma distributed rate heterogeneity parameters)

was selected under the Bayesian Information Criterion (BIC) using PartitionFinder (Lanfear et al. 2012) (Appendix A: Supplementary Table 11). We did not partition by codon position as the removal of columns with missing data from our alignment would have led to some individual codon partitions being extremely small, confounding the ability to optimise both partitioning scheme and substitution models across all possible combinations of partitions. Furthermore, by including substitution models which accommodate substitution rate heterogeneity among nucleotide positions (+I, +G), our model selection approach is able to accommodate such heterogeneity resulting from codon positioning within individual data partitions. We then carried out five-hundred bootstrap replicates using unlinked GTR+CAT (which approximates the GTR+G model) substitution models for each partition, with a final maximum likelihood tree calculated using GTR+G models. The Eulipotyphla clade was specified as outgroup (comprising *Uropsilus* sp.*, Crocidura attenuata, Episoriculus caudatus, Neomys fodiens, Nectogale elegans, Uropsilus soricipes, Crocidura shantungensis, Blarinella quadraticauda, Talpa europaea, Mogera wogura, Galemys pyrenaicus, Sorex araneus,* and *Anourosorex squamipes*).

Phylogeny and divergence times were then jointly estimated using a Bayesian approach in BEAST 1.8.3 (Drummond et al. 2012). For the Bayesian approach, we found the appropriate partition scheme and substitutions models through a second run of PartitionFinder (Lanfear et al. 2012) (this time including GTR, TrNef, TrN, HKY, K80, HKY or SYM, and again considering all possible combinations of invariant sites and gamma distributed rate heterogeneity parameters) (Appendix A: Supplementary Table 12). Time calibration was achieved by applying informative exponentially distributed priors on the ages of four internal nodes of the tree, based on information from the fossil record (Appendix A: Supplementary Table 13). These were: 'Bovidae', incorporating the basal diverge of the bovid clade, which must have occurred prior to 18 Ma based on the age of the fossil bovid *Eotragus noyei* (Benton et al. 2009); 'Carnivora', describing the basal diverge of the carnivore clade, which must have occurred prior to 37.1 Ma based on the age of the fossil *Hesperocyon gregarius* (Meredith et al. 2011); 'Perissodactyla', incorporating the basal divergence of the perissodactyl clade, which must have occurred prior to 47.8 Ma based on the age of the fossil *(Sifrhippus sandrae* (Froehlich 2002; Secord et al. 2012))

and ‘Laurasiathera’ incorporating the basal divergence of the Laurasiatheria clade, which must have occurred prior to 62.5 Ma based on the age of the fossil *Protictis haydenianus* (Benton et al. 2009). We implemented hard minimum ages of 18.0, 37.1, 47.8 and 62.5 Ma, respectively, and mean ages of 3.357, 6.015, 2.131 and 21.95 Ma, respectively. These parameters gave,providing soft maximum 95% bounds of 28.06, 55.12, 54.18 and 128.3 Ma, respectively. Fossil age distribution values for crown Laurasiatheria, Bovidae and Carnivora were based on those described in Welker *et al.* (Welker et al. 2015) , while the crown Perissodactlya fossil age distribution was based on an early Eocene distribution as fossils are most abundant during this time period (Froehlich 2002). Divergence dates on the tree were estimated using each of the four calibrations in combination, and also independently. For each of these analyses, we specified Eulipotyphla as outgroup. An uncorrelated lognormal relaxed clock model was utilised to accommodate variation in substitution rates along individual branches of the tree and a birth-death speciation process (Gernhard 2008) was specified as the tree prior. Preliminary runs showed that for some partitions, individual parameters of the substitution model suggested by PartitionFinder (Lanfear et al. 2012) failed to converge, indicating over-parameterisation. We therefore implemented the simpler HKY+I+G substitution model for these partitions in order to achieve convergence. The Markov Chain Monte Carlo (MCMC) chain was sampled every 20,000 generations, and ran for a sufficient number of generations to reach convergence and provide sufficient sampling of the posterior distributions of all parameters (ESS > 200), as determined using the program Tracerv1.6 (Rambaut et al.). The maximum clade credibility tree was extracted, with node heights scaled to the median of the posterior sample, and visualised using Figtree v1.4.2 (Rambaut 2014).

Data availability

The mitochondrial sequence for MAC002 can be found under the accession code KY611394 on Genbank.

Acknowledgements

We thank members of the following institutions for permission to sample specimens in their care: in Argentina, the Museo Argentino de Ciencias Naturales ‘‘Bernardino Rivadavia’’ (Buenos Aires), Museo de La Plata, Museo Municipal de Ciencias Naturales “Lorenzo Scaglia” (Mar del Plata), and Museo Paleontológico de San Pedro “Fray Manuel de Torres” (San Pedro); in Chile, the Departamento de Antropología (Universidad de Chile) and the Facultad de Patrimonio Cultural y Educación (Universidad SEK Chile); in Uruguay, Museo Nacional de Historia Natural (Montevideo); and in France, Muséum national d’Histoire naturelle (Paris). All fossil samples were collected before 2010. We also thank Rheon Slade for assisting in the modifications of some figures. This work was partly supported by National Science Foundation DEB 1547414 (R.D.E.M.). This work was also supported by the European Research Council (consolidator grant GeneFlow # 310763 to M.H.). The NVIDIA TITAN X GPU used for BEAST analyses was kindly donated by the NVIDIA Corporation.

Author contributions:

R.D.E.M. and M.H. conceived the project, M.W. and S.B performed lab work, M.W., S.B., A.B and J.L.A.P performed DNA analyses and interpretation of results, M.W. and A.B conducted the phylogenetic analyses and constructed trees, A.K., A.M.F, M.B., J.N.G., M.A.R., P.L.M., M.T., F.S., A.R., W.J., G.B., C.M., F.M. and J.L.A. assisted with locating and sampling specimens or provided logistical help, A.K., A.M.F., M.B., J.N.G., M.A.R., P.L.M. and R.D.E.M. provided the palaeontological and systematic framework for this study and wrote the palaeontological portion of the supplementary file. Final editing and manuscript preparation was coordinated by M.W., A.B., R.D.E.M., and M.H. All contributing authors read and agreed to the final manuscript

Figure 3.1. Map of sites yielding specimens of *Toxodon* and *Macrauchenia*. MAC002 *Macrauchenia*, the sample from which mitogenomic data was successfully collected, came from a metapodial recovered at the locality Baño Nuevo-1 Cave (in red). For locality context, see Appendix A: Supplementary Note 1 and Appendix A: Supplementary Figure 8. Map generated using QGIS 2.8 (QGIS Development Team, 2016. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://www.qgis.org/).

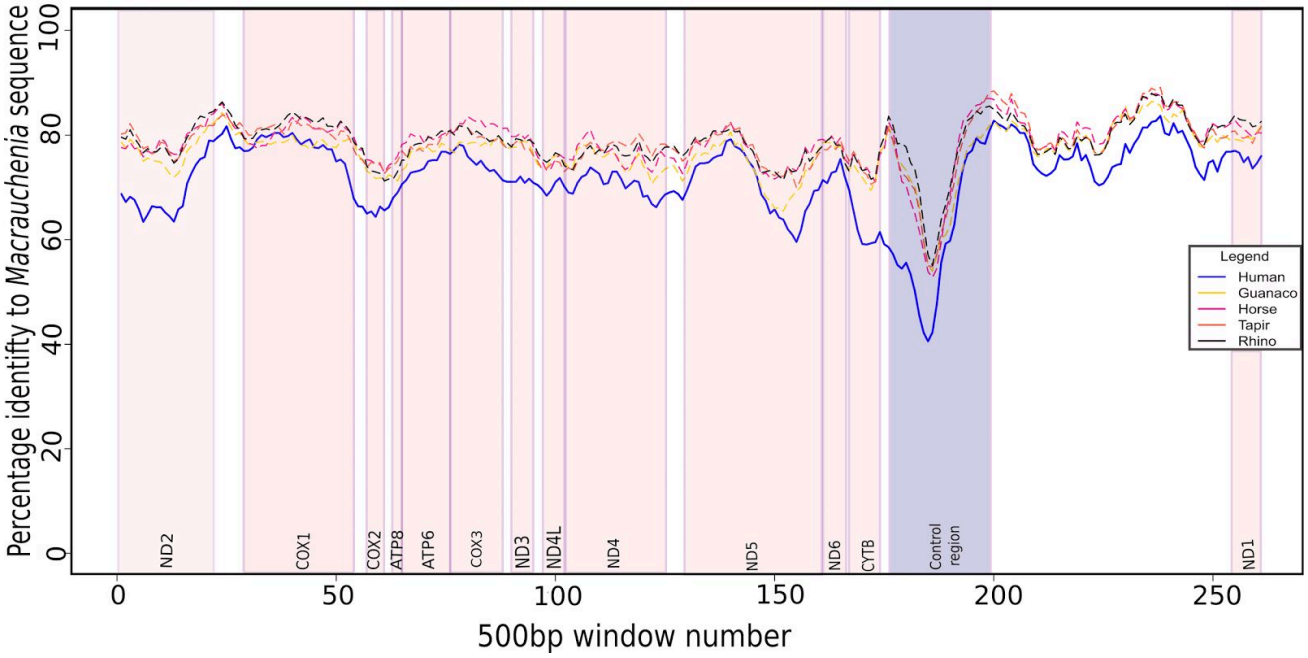


Figure 3.2. Contamination check using pairwise sliding-window comparisons. Comparisons were undertaken in 500bp windows with 50bp overlaps. X axis represents the sliding window number. Approximate gene locations within sliding windows are indicated by pink coloured boxes, and the control region indicated by a blue box. Five sliding window pairwise comparisons are shown: MAC002-human (blue), MAC002-rhino (black), MAC002-guanaco (yellow), MAC002-tapir (orange), and MAC002-horse (red).

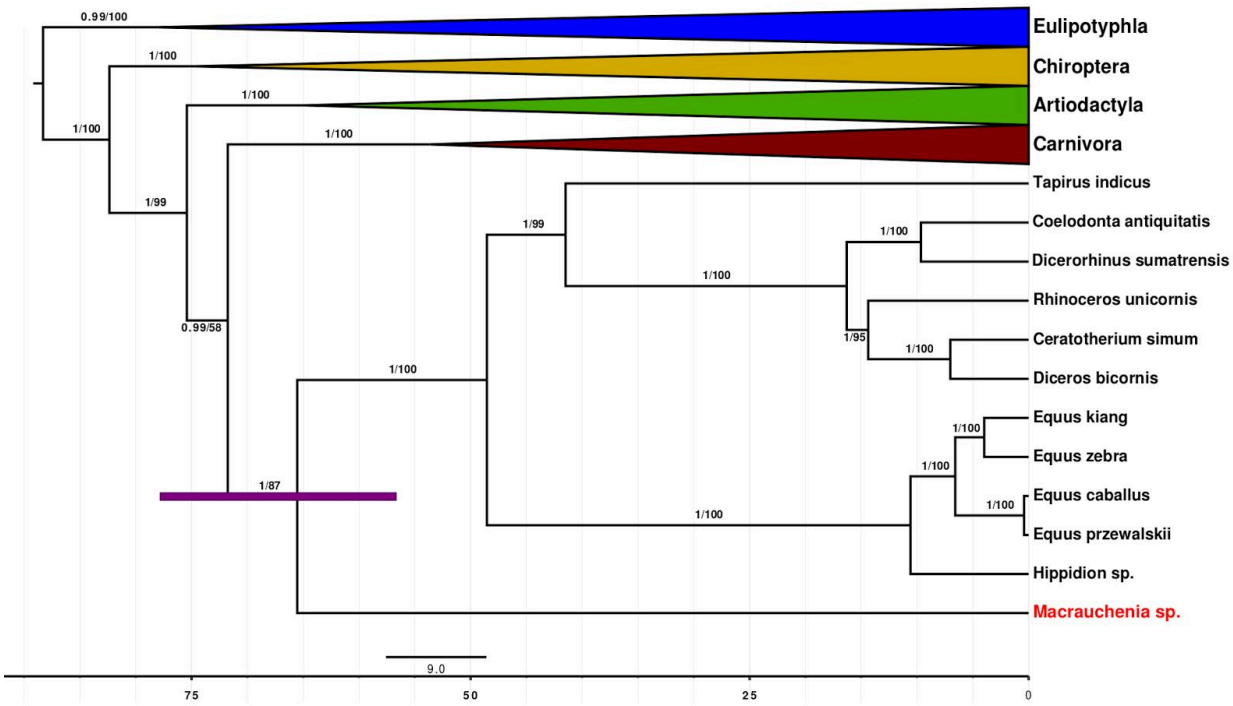


Figure 3.3. Dated mitogenomic phylogenetic tree. Posterior probabilities and bootstrap values are indicated on the tree branches. The purple node bars represents the 95% CI for the Panperissodactyla clade divergence date based on the combination of all four calibrations used in this study. Scale bar represents time in millions of years.

Chapter 4: Article III

Population and conservation genomics of the world's rarest hyena species, the brown hyena (*Parahyena brunnea*)*

Michael Westbury, Stefanie Hartmann, Axel Barlow, Ingrid Wiesel, Viyanna Leo, Rebecca Welch, Daniel M Parker, Florian Sicks, Arne Ludwig, Love Dalén, Michael Hofreiter

* This manuscript has been submitted to "Genome research" and is also available online in the preprint server "Biorxiv" under https://doi.org/10.1101/170621

4.1 Abstract

With an estimated population size of less than 10,000 individuals worldwide, the brown hyena (*Parahyaena brunnea*) has been listed as 'near threatened' by the IUCN. Despite this rank, studies involving DNA analyses of the brown hyena are limited. Little consideration has been focussed towards population structure within the brown hyena, which could provide valuable insights about its evolutionary history and aid in conservation efforts of the species.

Here we report both mitochondrial and nuclear genomes of wild-caught brown hyena individuals from across southern Africa. Mitochondrial DNA shows little to no phylogeographic structure, whereas low-coverage nuclear genomes reveal several potential sub-populations. Moreover, we find that brown hyenas harbour the lowest genetic diversity for a species on both the mitochondrial and nuclear level when compared to a number of mammalian species for which such information is currently available. Our data also reveal that at least on the nuclear DNA level, this low diversity could be the result of a continuous and ongoing decline in effective population size that started about one million years ago and dramatically accelerated towards the end of the Pleistocene. Moreover, our findings also show that the correlation between genetic diversity and the perceived risk of extinction is not particularly strong, since many species with higher genetic diversity than the brown hyena are considered to be at greater risk of extinction.

Taken together, our results have important implications for the conservation status and conservation approaches of the brown hyena.

4.2 Introduction

With the ever-decreasing price of next generation sequencing (NGS), population genomics has become a rapidly evolving field within evolutionary biology (Buerkle and Gompert 2013). First appearing in the late 1990s with the analysis of large-scale datasets on human single-nucleotide-polymorphisms (Hartl and Clark 1997), population genomics analyses have now been implemented in the study of a range of species and has greatly aided in understanding their evolution (Ellegren 2014; Liti et al. 2009; Nielsen et al. 2009). A major argument in favour of the analysis of whole genomic data is that the high number of independent loci available provides both power and accuracy and allows for the separation of real evolutionary effects from noise. This remains true even when considering just a few individuals, making genomic approaches much more versatile than more traditional methods utilising a few loci and large numbers of individuals (McMahon et al. 2014).

Population genomic analyses have a wide variety of applications and can be used to accurately gauge population structure and connectivity, taxonomic relationships, genetic diversity, and the demographic history of a species (Shafer et al. 2015; Allendorf et al. 2010; Steiner et al. 2013). Despite the recent rise in the number of genomic sequences obtained, many of these studies are still restricted to already well-studied species such as humans, model organisms and domesticates. However, the use of population genomic analyses is often extremely useful in solving questions about the evolutionary history of less studied, yet evolutionarily important lineages.

One little studied but ecologically important lineage is the family Hyaenidae (hyenas and aardwolf). Hyaenidae occupy a species-poor (four extant species) branch within the Feliformia suborder in a sister position to the Felidae family. However, despite its close relationship to the

well-studied Felidae family, of which a number of genomes have already been sequenced (Abascal et al. 2016; Dobrynin et al. 2015; Cho et al. 2013), very few genetic studies have been carried out on the Hyaenidae family. Members of Hyaenidae occupy a variety of different niches. The most notable and arguably most important niche being that of the scavenger (Gusset and Burgener 2005; Watts and Holekamp 2007). Scavengers are known to be important for maintaining healthy ecosystem function with profound roles in nutrient cycling and influencing disease dynamics (Benbow et al. 2015).

The brown hyena (*Parahyaena brunnea*) is predominantly a scavenger, mainly feeding on large vertebrate carrion (Watts and Holekamp 2007). It is generally found in arid areas across southern Africa and is listed as 'Near Threatened' by the International Union for Conservation of Nature (IUCN). The brown hyena is the rarest of all extant hyena species with estimates of population size being less than 10,000 individuals worldwide (Wiesel 2015). Despite its listing as Near Threatened, brown hyenas continue to be persecuted, often considered as problem animals by farmers or killed for trophy hunting. Incidental and often deliberate poisoning, shooting and trapping of these animals all hamper the survival of this ecologically important species (Kent and Hill 2013).

Genetic studies of the brown hyena are very limited but have hinted towards very low genetic diversity within the species (Rohland et al. 2005; Knowles et al. 2009). Species-wide genetic comparisons using a short fragment of the mitochondrial cytochrome b gene found no variability in this region regardless of sample origin (Rohland et al. 2005). Moreover, a study investigating population structure within Namibian brown hyena, utilising microsatellites, also found no detectable population structure (Knowles et al. 2009). The inability of both of these studies to find population structure perhaps stems from very low levels of genetic diversity within the brown hyena. These early indications of low diversity are important to follow up on and investigate. Even though genetic diversity does not necessarily correlate with current day population sizes (Bazin et al. 2006; Leffler et al. 2012), it is still an important factor in understanding past evolutionary events. Importantly, knowledge of the evolutionary processes

affecting a species is critical to inform conservation plans aimed at the long term management of its evolutionary potential (Romiguier et al. 2014).

Here we present complete mitochondrial and nuclear genomic analyses of brown hyena from populations across its range in Southern Africa (Fig. 3.1). We find both mitochondrial and nuclear genomic diversity to be extremely low, lower than in any other mammalian species published to date. Our data suggest that this low diversity results from a continuous decrease in effective population size over the last million years. Moreover, our low-coverage genomes reveal a number of potential sub-populations of brown hyena across its range.

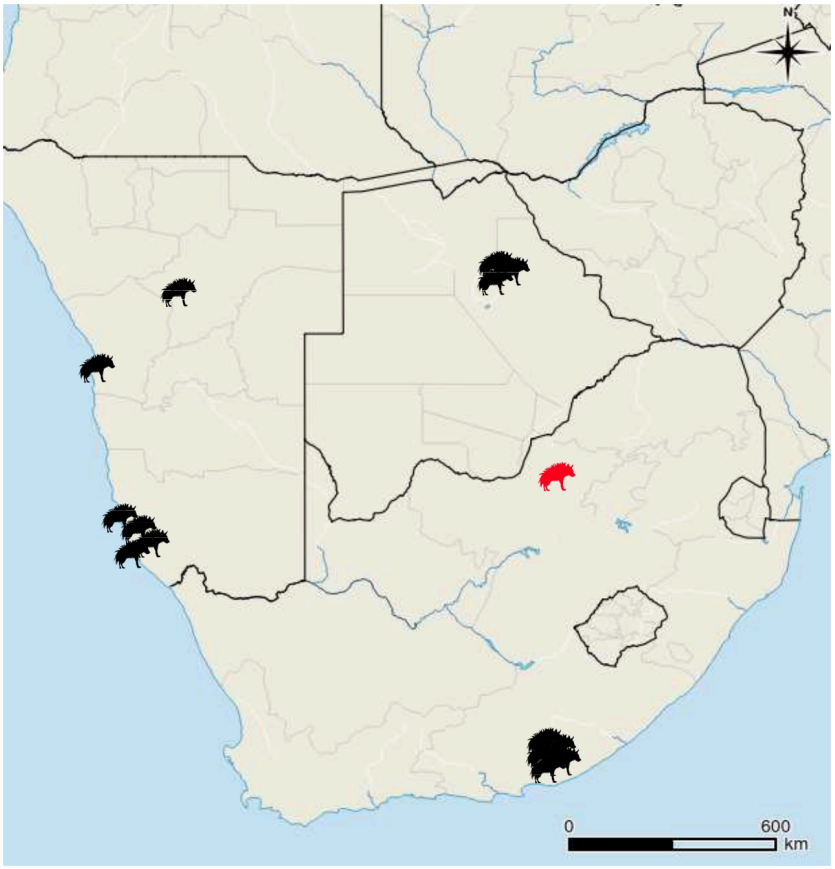


Figure 4.1 : Map of the sampling locations of the wild-caught brown hyena included in this study. The red hyena indicates the original area of the South African population prior to translocations in 2003.

4.3 Results

4.3.1 Genome reconstructions

A *de novo* assembly of a striped hyena nuclear genome with a scaffold N50 of ~2Mbp was assembled using Allpaths LG (Butler et al. 2008), default parameters and an additional gap closing step using Gapcloser (Luo et al. 2012) (Appendix B: Supplemental Table S1). BUSCO analyses (Simão et al. 2015) using both the eukaryotic and mammalian databases show high levels of complete BUSCOs (Appendix B: Supplemental Table S2) giving us confidence that our assembly is of good quality and completeness. With this as reference, we successfully mapped low coverage nuclear genomes (2.1 - 3.7x) from 14 wild-caught brown hyena individuals originating from Namibia, South Africa and Botswana and a high coverage nuclear genome from a captive individual (Appendix B: Supplemental Tables S3 and S4).

Due to the lack of a published brown hyena mitochondrial genome, we assembled the complete *de novo* mitochondrial genome from a captive individual using MITObim (Hahn et al. 2013). We used default parameters apart from mismatch value, where we used zero, and three different bait reference sequences (domestic cat (U20753.1), spotted hyena (JF894377.1) and striped hyena (NC_020669.1)). All three independent MITObim runs produced identical brown hyena mitochondrial sequences, suggesting that our reconstructed mitochondrial genome is correct. We then mapped the remaining wild samples to this sequence (Appendix B: Supplemental Table S4).

4.3.2 Genetic diversity

Mitochondrial DNA diversity estimates (Fig. 4.2) using the 14 wild caught mitogenomes from our data set showed that the brown hyena has the lowest diversity when compared to a number of other mammalian mitochondrial genomes as analysed in previous studies (Dobrynin et al. 2015; Miller et al. 2011). All brown hyena individuals shared an average of only three

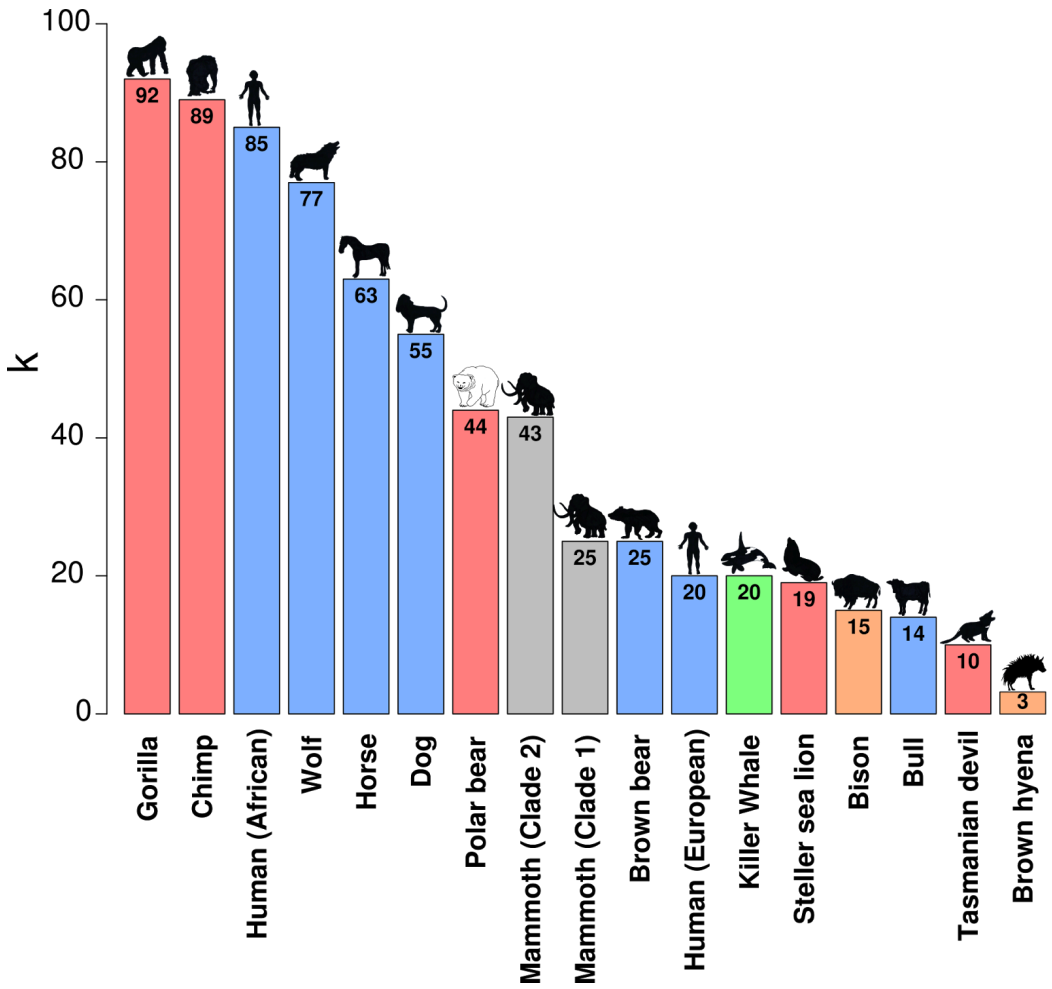


Figure 4.2: Mitochondrial diversity comparisons of the brown hyena to a number of other species/populations for which such data are available (Dobrynin et al. 2015; Miller et al. 2011). k represents the average number of substitutions expected between two randomly selected individuals of the same population, bar colours represent conservation status (red - endangered, grey - extinct, orange - near threatened, blue - least concern, green - unavailable) according to the IUCN.

mutational mismatches (k) among them. This is about three times lower than the species with the next lowest level of mitochondrial diversity, the tasmanian devil (*Sarcophilus harrisii*), with a k value of 10, and 30 times lower than the gorilla (consisting of both *Gorilla gorilla* and *Gorilla beringei*), which has a k value of 92. We also calculated genome-wide nuclear heterozygosity estimates of our high coverage captive individual as these are considered a good

proxy for nuclear genomic diversity. Results show the brown hyena to have the lowest of any species included in this study, most of which are considered "endangered" (Fig. 4.3A). These results are consistent with the very low levels of diversity we found within the mitochondrial genome.

The brown hyena individual sequenced to high coverage is known to have been born from wild-caught parents, albeit in a captive environment, and should not display large amounts of inbreeding that can be found in captive bred populations. However, to investigate signs of inbreeding, we measured levels of heterozygosity in various window sizes across the brown hyena nuclear genome. The results show no considerable stretches of homozygosity (Appendix B: Supplemental Fig. S1), suggesting that there are no significant signs of inbreeding.

We further analysed the distribution of heterozygosity across the genome by randomly selecting 1000 non-overlapping windows of 1Mbp in size and comparing the results to the four other species with the lowest heterozygosity included in this study, the orca (*Orcinus orca*), cheetah (*Acinonyx jubatus*), Channel Island fox (*Urocyon littoralis*) and Iberian lynx (*Lynx pardinus*) (Fig. 4.3B). The brown hyena genome has consistently lower levels of heterozygosity across the genome when compared to the Iberian lynx, orca and cheetah, excluding the windows with the lowest levels of heterozygosity which may indicate higher levels of inbreeding in the other species. However, it can be seen that, while the brown hyena has a lower level of average genome-wide diversity, for the majority of the windows, the Channel Island fox has lower heterozygosity than the brown hyena. Regions of high levels of heterozygosity amongst a sea of low heterozygosity have been previously reported for the Channel Island fox (Robinson et al. 2016). Moreover, it should also be noted that as each estimate only shows the genomic diversity of a single individual and levels are expected to vary within natural populations, some caution must be taken when interpreting these results.

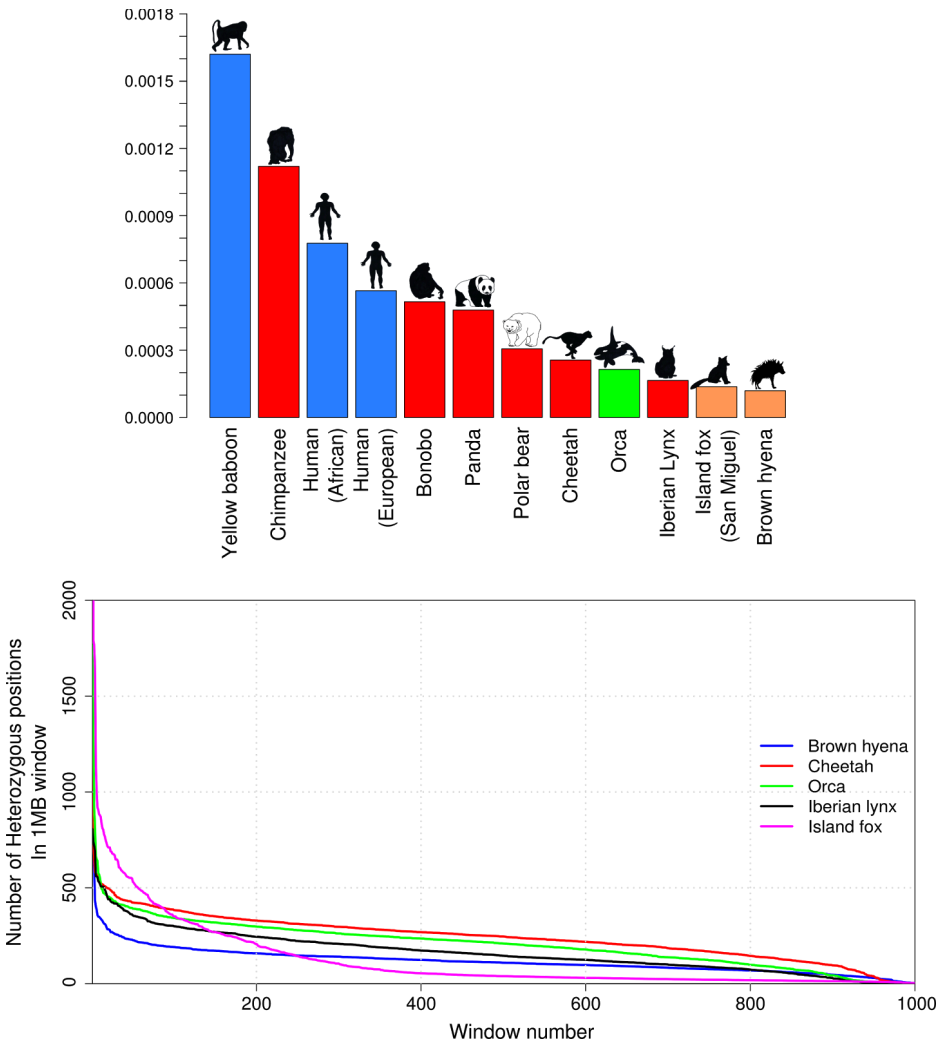


Figure 4.3: Estimated heterozygosity levels in the brown hyena and a comparison to other mammalian species. **A)** Average genome-wide heterozygosity comparisons. Y axis represents the average proportion of sites within the nuclear genome to be heterozygous. Bar colours represent conservation status (red - endangered, orange - near threatened, blue - least concern, green - insufficient data) according to the IUCN. **B)** Heterozygosity density comparisons between the four species with the lowest estimated average genome-wide heterozygosity levels in this study. Y axis represents the number of heterozygous sites within the 1Mbp window. X axis represents the window. Colours represent species (blue - brown hyena, black - Iberian lynx, green - orca, red - cheetah, magenta - Island fox).

4.3.3 Demographic history

As sex chromosomes are known to have a different demographic history than autosomes, we found and removed scaffolds in our assembly related to the X chromosome before running demographic analyses. We aligned our striped hyena assembly by synteny using Satsuma (Grabherr et al. 2010) to the cat X chromosome (126,427,096bp) and found 195 scaffolds (Appendix B: Supplemental Table S5), totaling 117,479,157bp. The alignment was visualised using Circos (Krzywinski et al. 2009) (Appendix B: Supplemental Fig. S2). The Circos alignment shows that the scaffolds found using Satsuma cover the complete cat X chromosome. These scaffolds were then removed before running a Pairwise Sequentially Markovian Coalescent (PSMC) model (Li and Durbin 2011). PSMC analyses using the brown hyena autosomes were consistent with its low genomic diversity (Fig. 4.4; Appendix B: Supplemental Fig. S3) and revealed a continuous gradual decrease in effective population size over the last one million years. This was then followed by a more rapid recent decrease in effective population size at the end of the Late Pleistocene.

4.3.4 Population structure

In order to infer potential population structure, we reconstructed a haplotype network using the mitochondrial genomes of the wild-caught individuals. This network revealed some phylogeographic structure with all but one haplotype being geographically restricted. This one shared haplotype was shared among individuals from all three sampled countries (Fig. 4.5A).
We then used the mapped genomes of the wild-caught individuals to infer nuclear population structure by carrying out principal component analyses (PCA) and maximum likelihood (ML) phylogenetic analyses. The results showed that clusterings or clades of individuals were in all cases consistent with the geographical origin of the individuals (Fig. 4.5B; Appendix B: Supplemental Fig. S4). PCA analyses using both single base identity by state (IBS) and genotype likelihoods (GL) produced similar results (Fig. 4.5B; Appendix B: Supplemental Fig. S5).

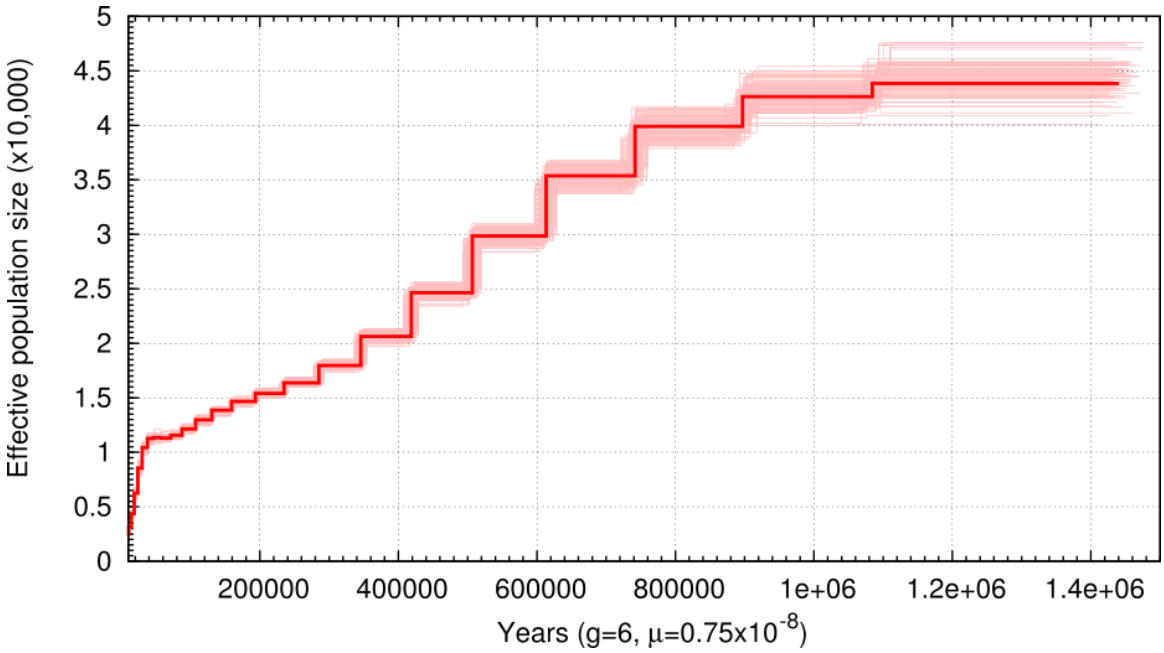


Figure 4.4: Pairwise Sequentially Markovian Coalescent (PSMC) model plot of the autosomes of one high coverage brown hyena. The Y axis represents effective population size and the X axis represents time in years. Light red bars show bootstrap support values. Calibrated using a generation time (g) of 6 years and a mutation rate (μ) of 7.5x10-9 per generation.

In contrast to whole nuclear genomic data, both PCA and phylogenetic analyses using single scaffolds did not produce unanimous phylogeographic results. Independent PCA results for the nine longest scaffolds can be seen in Appendix B: Supplemental Fig. S6. Although they do generally support some phylogeographic structure, individuals from different regions partially intermingle in these plots. When running independent per-scaffold ML phylogenetic analyses with scaffolds over 2MB, we find that in 49 out of 333 trees, the South African samples form a monophyletic clade, in 72 out of 333 trees the Botswana samples form a monophyletic clade and in 17 out of 333 trees the six Namibian samples form a monophyletic clade. This non-unanimous pattern shows up as clouds surrounding all nodes within the brown hyena lineage when visualising all independent ML trees simultaneously using Densitree (Bouckaert 2010) (Fig. 4.5C). Although few individual trees support monophyly of either of the three geographical groups, the root canal (consensus tree with highest clade support), defined by Densitree, nevertheless shows a similar topology as the tree constructed using the complete nuclear genome

(Fig. 4.5C; Appendix B: Supplemental Fig. S4). Importantly, it is again fully consistent with the geographical origin of the individuals.

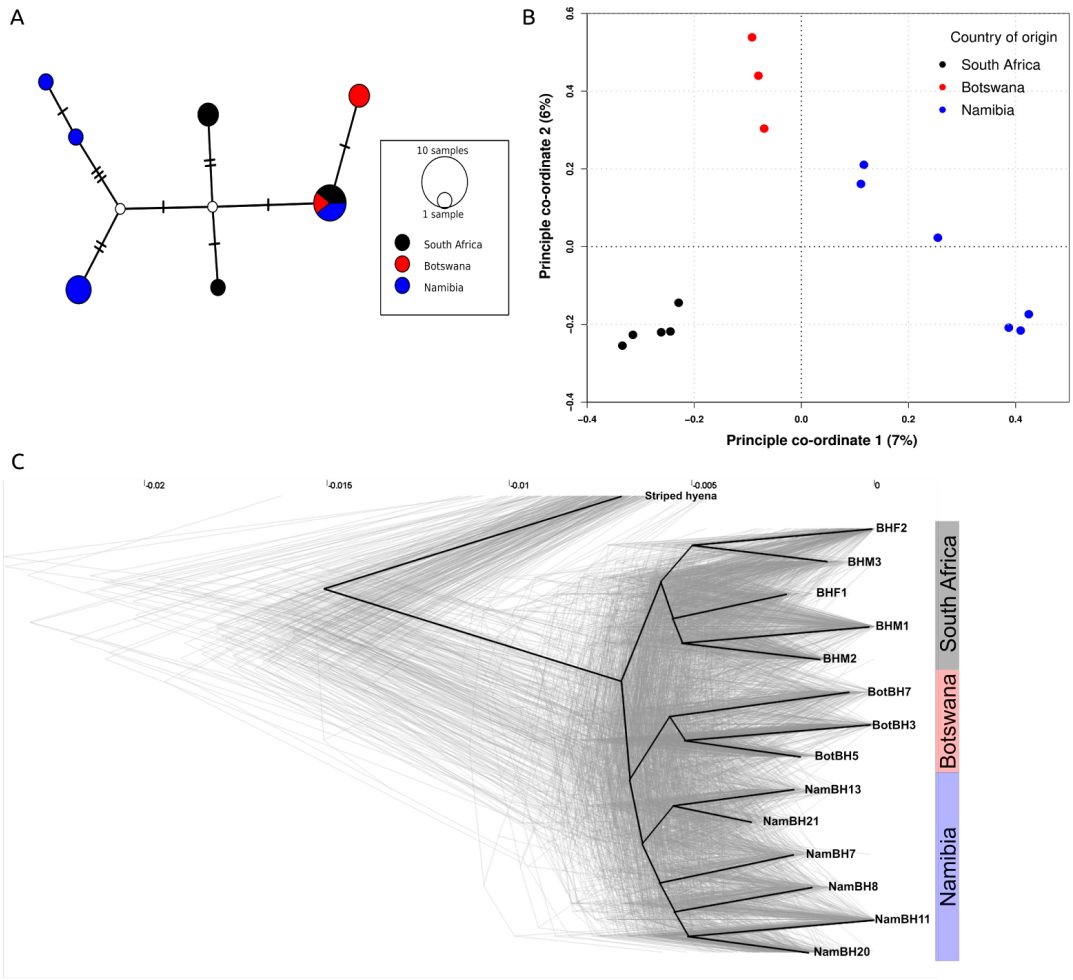


Figure 4.5: Genomic diversity analyses A) Median joining haplotype network of the 14 wild-caught brown hyena individuals included in this study. Lines on the connecting branches represent single base differences, size of the circle represents number of individuals belonging to a single haplotype and colours represent sampling country (black - South Africa, blue - Namibia, red - Botswana). B) Principal components analysis produced using genotype likelihoods for the low coverage genomes of the 14 wild-caught brown hyena individuals in this study. Colours represent country of origin (black - South Africa, blue - Namibia, red - Botswana). Percentages on the X and Y axis represent the percentage of variance explained by each respective component. C) Densitree phylogenetic tree. Light grey lines represent phylogenetic trees produced from single scaffolds. The dark black line represents the root canal as defined by Densitree.

Admixture and population structure analyses using NGSadmix (Skotte et al. 2013) reached convergence for K values of two to five, indicating that there may be anywhere from two to five different populations inhabiting southern Africa (Fig. 4.6). When considering a K value most congruent with the number of countries of origin, K3 (Fig. 4.6B), individuals NamBH13 and NamBH21 stand out as being "admixed" between all three populations. This could be a byproduct of the analysis trying to find a population for them. A K value of four however, shows them as an independent population (Fig. 4.6C). The latter is the most parsimonious result as it is consistent with their geographic origins (northern Namibia while all other Namibian samples are from the south) and previous findings investigating the distribution patterns of individuals across Namibia (Appendix B: Supplemental Fig. S7).

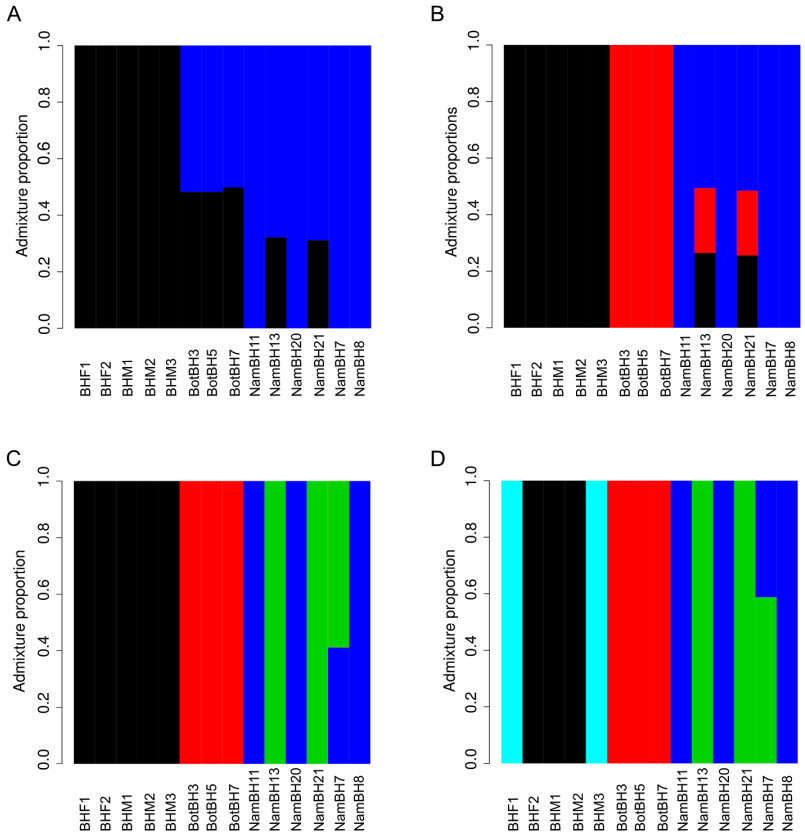


Figure 4.6: Admixture plots produced using different K values in NGSAdmix. **A)** K2 **B)** K3 **C)** K4 **D)** K5. Y axis represents the admixture proportion and x axis represents the individual.

We carried out D-statistics comparisons testing for population structure (Fig. 4.7). A high D-value could either represent differential levels of admixture or more recent common ancestry and therefore an incorrect predefined topology. Taking the latter into account, we placed individuals into one of three (defined as: Namibia, Botswana, or South Africa) or one of four (defined as: northern Namibia, southern Namibia, Botswana, or South Africa) predefined populations and compared the D values produced from "correct" topologies, i.e branches H1 and H2 belong within the same population (e.g. Fig. 4.7A), and "incorrect" topologies, i.e branches H2 and H3 belong within the same population (e.g. Fig. 4.7B). When considering four populations (Fig. 4.7C) it can consistently be seen that when we break the predefined population structure and therefore topology, we find a higher D value than when individuals in the same predefined population are in the H1 and H2 positions. Assuming that the correct topology is that with the lowest D value and as there is clear separation between D values recovered when breaking and not breaking the predetermined population structure when using four predefined populations, we conclude that there are indeed four populations within our dataset. This pattern is not seen when only three predefined populations (Fig. 4.7D) are considered as there are many overlapping D values within the Namibian population when "correct" and "incorrect" topologies are tested. This led us to reject a possibility of one single population within Namibia. These results suggest four populations, with a split between northern and southern Namibia. This analysis revealed the same population structure as the NGSadmix analyses, thus corroborating the observation that this dataset consists of four populations rather than three.

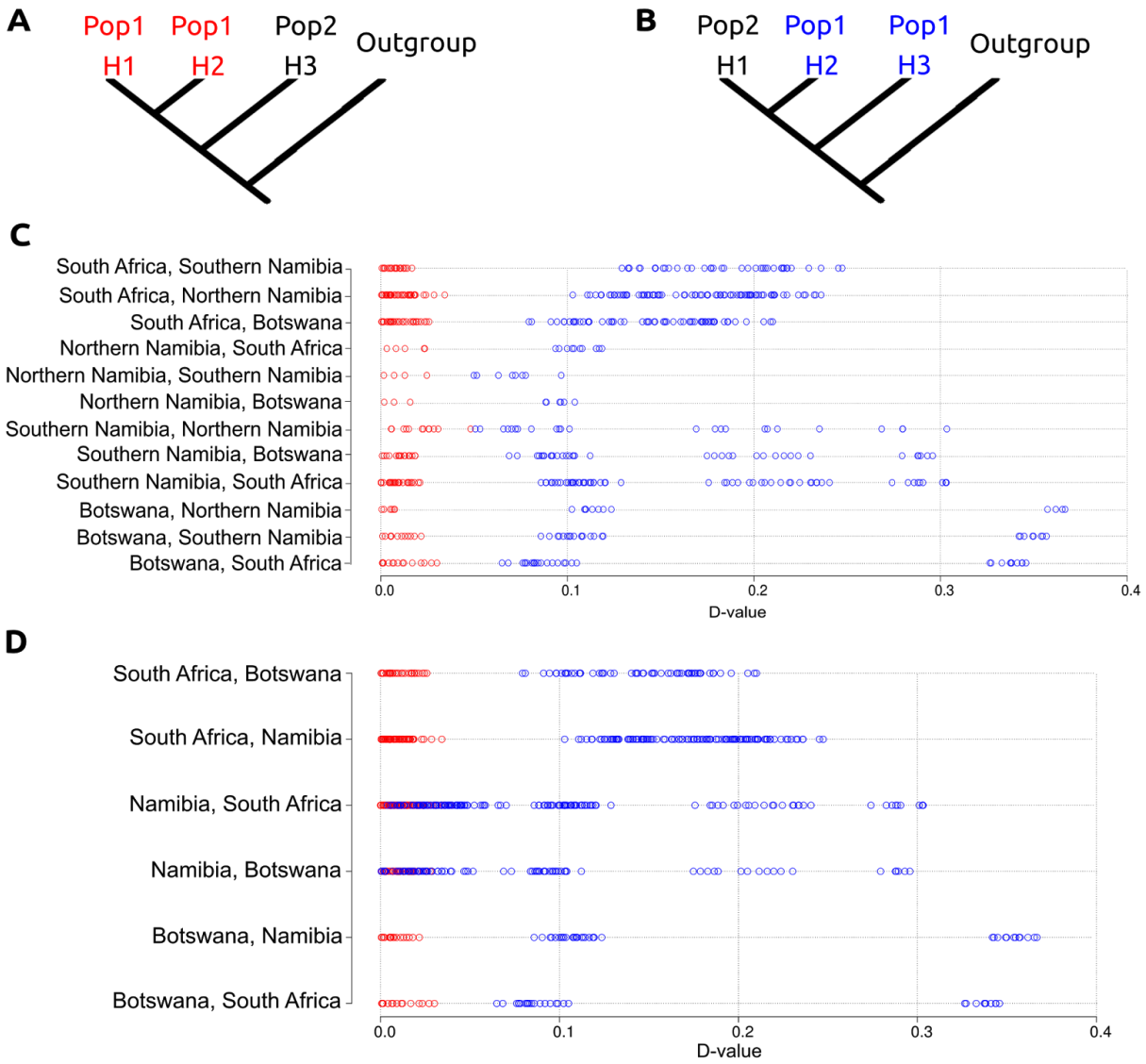


Figure 4.7: Topology test using D-statistics. **A)** D-statistic analysis demonstrating the correct predetermined population structure. **B)** D-statistic analysis demonstrating the incorrect predetermined population structure. **C)** D-statistics comparisons when four "populations" are determined a priori. **D)** D-statistics comparisons when three "populations" are determined a priori. Red coloured circles represent comparisons when the predefined population structure is not broken within the comparison. Blue coloured circles represent comparisons when the predefined population structure is broken within the comparison. X axis represents the D-value.

We carried out IBS analyses on a number of other mammalian species for which species-wide low coverage data was available in order to compare these with the population structure within the brown hyena. Each species, including the brown hyena, had 10 representatives randomly selected and pairwise distance comparisons were carried out within each species (Appendix B: Supplemental Figure. S8). For the brown hyena, most pairwise distance values clustered together, representing a relatively high level of shared diversity between different putative populations and suggesting extensive gene flow among these populations. This pattern was not as strong, however, as in some species, such as domesticated sheep which are known to be fairly panmictic (Peter et al. 2007) and still suggests some level of differentiation within the brown hyena. We also found no abnormalities, i.e. potential subspecies, large differences between populations or indications that a number of the individuals were extraordinarily related which could be driving the "population" signal we find in the brown hyena.

4.4 Discussion

Hyaenidae occupy a major, albeit species poor branch within Feliformia. The family has a rich fossil history but is now restricted to only four extant species (Koepfli et al. 2006). The abundance of fossil specimens has led to a number of studies focused on the taxonomic relationships within fossil Hyaenidae (Werdelin and Solounias 1991; Turner et al. 2008). Little attention, however, has been focused towards the evolution and population status of the extant species, especially on a molecular level. Despite the brown hyenas rarity, molecular analyses of the brown hyena are currently limited to two studies on microsatellites and short regions of mitochondrial DNA (Rohland et al. 2005; Knowles et al. 2009). Both hinted at low genetic diversity but had very limited results with regard to population structure.

4.4.1 Genomic diversity

Using complete nuclear and mitochondrial genomes of 14 wild-caught brown hyena originating from across southern Africa, brown hyenas displayed the lowest level of both mitochondrial and nuclear genomic diversity when compared against other mammalian species for which comparable data were available, including many endangered species (Fig. 4.2; Fig. 4.3A). Genomic diversity was even lower than in species famously known to have extremely low levels of genomic diversity, such as the cheetah (Dobrynin et al. 2015) and the Iberian lynx (Abascal et al. 2016), both of which have gained considerable research attention because of this characteristic. In addition to finding very low levels of genomic diversity, demographic analyses of the brown hyena show a gradual, yet steady decline in effective population size over the last million years, with a more rapid decline towards the end of the Late Pleistocene (Fig. 4.4). The brown hyena is known to have once had a more extensive range, with Middle Pleistocene fossils having been found in Kenya (Werdelin and Barthelme 1997). The continuous decrease in effective population size and low levels of genomic diversity seen today may have occurred with the shrinking of suitable habitats during the Pleistocene (deMenocal 2004) and potentially coincided with the migration of new competitors, such as jackals, into Africa (Koepfli et al. 2015).

Interestingly, despite very low levels of genetic diversity within both the nuclear and mitochondrial genomes, we found no strong signs of inbreeding in the nuclear genome (Fig. 4.3B; Appendix B: Supplemental Fig. S1). This was an unexpected result as low genomic diversity is generally expected to arise from high levels of inbreeding (Willoughby et al. 2015). Low genomic diversity has commonly been associated with decreased fitness and higher extinction rates (Spielman et al. 2004; Reed and Frankham 2003). However, the lack of detectable inbreeding in our high coverage individual could mean that this species has evolved into a state of genetic stasis, allowing low genetic diversity to persist across the genome, not strongly influencing the survivability of the species. This genetic stasis could have evolved as a result of the slow but continuous decrease in effective population size seen over the last million

years and lack of detectable bottlenecks during that time (Fig. 4.4). This slow continuous decrease could hamper the influence of genetic drift therefore allowing purifying selection to maintain variability in a number of loci potentially important for the survival of the species. The retention of relatively high levels of potentially important adaptive variation allowing for the retention of evolutionary potential has recently been found in Channel Island foxes (Robinson et al. 2016). Heterozygosity hotspots were shown to be present in a number of genes with high levels of ancestral variation, despite low levels of genome-wide heterozygosity in the Channel Island fox. However, as we see heterozygosity to be fairly evenly distributed across the genome, this doesn't seem to be the case for the brown hyena and it has perhaps been able to persist despite the lack of adaptive variation due to some other factor. Further research would however be required to definitively know whether the brown hyena goes against the commonly perceived notion that high levels of adaptive variation are required to ensure a species survival.

4.4.2 Brown hyena population structure

By using a combination of different population structure analyses on the entire nuclear genome we were also able to, for the first time, define population structure within the brown hyena (Fig. 4.5; Fig. 4.6 and Fig. 4.7) despite its very low levels of genetic diversity. Results were in concordance with geographic structuring. Furthermore, by comparing these results to those produced using mitochondrial genomes and single scaffolds, we show that even when using millions of loci (i.e. single scaffolds >2Mbp (Fig. 4.5C; Appendix B: Supplemental Fig S6)) difficulties with accurately defining population structure can arise. This maybe due to the very low genomic diversity within the brown hyena. This finding reinforces the value of using whole nuclear genomes in population analyses over other approaches such as microsatellites and reduced genomic representation techniques, especially in species suspected to have low genomic diversity.

4.4.3 Conservation implications

Our findings provide a greater understanding into the evolution and population structure of the brown hyena, which has wider implications in aiding conservation of this species. Due to the increase in agriculture, many habitable areas for the brown hyena have now been fragmented, and the connectivity between populations has therefore been reduced. The wide geographical range of the brown hyena, coupled with the extremely low levels of genomic diversity, indicates that the brown hyena persists at low population densities, which is in concordance with previous studies analysing natural population densities (Kent and Hill 2013). This demonstrates the importance of connectivity within the species. Given the overall evidence for substantial gene flow between the different populations, well implemented translocations may certainly become important for the survivability of this species in the future. With better definitions of population boundaries, translocations can be implemented in such a way as to avoid outbreeding depression and increase the chances of an animal successfully adapting to its new environment (Allendorf et al. 2010). Moreover, previous studies have shown that translocated brown hyena are able to settle into their new environments and do not return to their original locations (Weise et al. 2015), and that human mediated translocations can lead to the rise of highly dense and successful populations (Welch and Parker 2016). These studies further demonstrate the ability of this species to adapt and accept translocations, and the utility these techniques could have in the future.

4.5 Methods

4.5.1 Samples

In total, 15 brown hyena samples were used for this study; five from South Africa, six from Namibia, three from Botswana (Fig. 4.1) and one from Tierpark Berlin (Appendix B: Supplemental Table S3). One female striped hyena (*Hyaena hyaena*) from Tierpark Berlin was also included to be used for reference based mapping.

4.5.2 Striped hyena *de novo* assembly

We extracted DNA from Hyena2069, the Tierpark Berlin striped hyena sample, on a KingFisher Duo robot using the blood DNA extraction kit according to the manufacturer's instructions. The extract was then built into two PCR free Truseq Illumina sequencing libraries, one with 180bp and one with 670bp short inserts. Two Nextera mate-pair libraries were also constructed with sizes of 3kbp and 6kbp. All of the above-mentioned libraries were constructed at the National Genomics Infrastructure (NGI) in Stockholm. All libraries were then sequenced on an Illumina HighseqX using 2x150bp paired-end sequencing at the NGI in Stockholm. The 180bp and 670bp insert libraries were sequenced on one lane each, whereas the mate-pair libraries were multiplexed and sequenced together on a single lane.

The NGI trimmed Illumina adapter sequences from the raw illumina reads using Trimmomatic (Bolger et al. 2014) and performed a *de novo* assembly using Allpaths LG (Butler et al. 2008) with default parameters. We then performed an additional gap closing step using Gapcloser (Luo et al. 2012). Assembly quality and completeness was assessed using BUSCOv2 (Simão et al. 2015) using both the eukaryote and mammalian BUSCO databases (Appendix B: Supplemental Table S2).

4.5.3 Captive brown hyena sample

We extracted DNA from a single brown hyena sample from Tierpark Berlin on a KingFisher Duo robot using the cell and tissue DNA extraction kit. The extract was then built into a PCR free Truseq Illumina sequencing library using a 350bp insert size by the NGI in Stockholm. This library was then sequenced on an Illumina Highseq X using 2x150bp paired-end sequencing at the NGI in Stockholm.

4.5.4 Wild caught brown hyena samples

We extracted DNA from the six blood and three tissue samples using a DNeasy blood and tissue extraction kit, following the manufacturer's protocol. We extracted the five hair samples using the DY04 user modified version of the DNeasy kit and protocol. We fragmented DNA extracts using a Covaris sonicator into ~500bp fragments. Fragmented extracts were then constructed into Illumina sequencing libraries using a modified version of the protocol set out by Meyer and Kircher (Meyer and Kircher 2010; Gonzales-Fortes and Paijmans 2015). Library molecules from 400bp to 900bp were selected using a Pippin Prep Instrument (Sage Science) and sequenced on an Illumina Nextseq 500 at Potsdam University, Germany.

4.5.5 Raw data treatment

We trimmed Illumina adapter sequences and removed reads shorter than 30bp from the raw reads of the 15 brown hyena samples using Cutadapt v1.8.1 (Martin 2011) and merged overlapping reads using FLASH v1.2.1 (Magoč and Salzberg 2011).

4.5.6 Mitochondrial genome reconstruction

As no brown hyena mitochondrial sequence was available, we reconstructed one using the shotgun data from our single high coverage individual. We assembled the mitochondrial genome through iterative mapping using MITObimv1.8 (Hahn et al. 2013) on 40 million trimmed and merged reads, subsampled using seqtk (Li 2012). We removed duplicate reads using prinseq (Schmieder and Edwards 2011). MITObim was performed in three independent runs using three different starting bait reference sequences. The references included the domestic cat (U20753.1), spotted hyena (JF894377.1) and striped hyena (NC_020669.1). We implemented MITObim using default parameters apart from mismatch value where we used zero. Output maf files were converted to sam files and visualized using Geneious v9.0.5 (Kearse et al. 2012).

Consensus sequences were constructed in Geneious using a 75% base call consensus threshold, and only sites with over 20x coverage were considered.

The reconstructed mitochondrion served as a reference sequence for subsequent mitochondrial mapping analyses. We mapped the trimmed and merged reads from our 14 wild brown hyenas to the reconstructed reference sequence using BWAv0.7.15 (Li and Durbin 2009), using the mem algorithm and default parameters and parsed the mapped files using Samtools v1.3.1 (Li et al. 2009). The consensus sequences were constructed using ANGSDv0.913 (Korneliussen et al. 2014), only considering reads and bases of quality scores greater than 25.

4.5.7 Mitochondrial analyses

The mitochondrial genomes from the 14 wild-caught brown hyena were aligned using Mafftv7.271 (Katoh and Standley 2013). We constructed a median joining haplotype network of the alignment using Popart (Leigh and Bryant 2015). Nucleotide diversity (pi = 0.000140667) was estimated using Popart. The k value was calculated by multiplying the pi value by the total length of the brown hyena mitochondria. This number was then compared against those from previous studies (Dobrynin et al. 2015; Miller et al. 2011) (Fig. 3.1B).

4.5.8 Low coverage nuclear genome analyses

Trimmed and merged data were mapped to the *de novo* striped hyena assembly using BWA v0.7.15 (Li and Durbin 2009) and parsed using Samtools v1.3.1 (Li and Durbin 2009).
We applied the following filtering options for all analyses involving ANGSD (Korneliussen et al. 2014): we only considered sites where at least 10 individuals had coverage (-minInd 10), only included sites for which the per-site coverage across all individuals was less than 75. We implemented quality filtering by setting a minimum base quality score of 25 (-minQ 25), minimum mapping quality score of 25 (-minMapQ 25) and only allowed reads that mapped

uniquely to one location (-unique_only 1). We also adjusted quality scores around indels (-baq 1) (Li 2011).

4.5.9 Brown hyena population structure

Principle component analyses (PCA) were carried out using both single read IBS analyses and GL analyses in ANGSDv0.913 (Korneliussen et al. 2014). IBS analyses were restricted to SNPs occurring in at least two individuals. This was done to remove singletons which could represent sequencing errors. We computed genotype likelihood in ANGSD and converted outputs to a covariance matrix using ngsTools (Fumagalli et al. 2014). Covariance matrices were converted into PCA outputs and visualized using R (R Development Core Team 2008). For the phylogenetic analyses, we performed Maximum likelihood analyses with Raxml v8.2.10 (Stamatakis 2014), specifying the striped hyena as outgroup and using the GTR+GAMMA substitution model. We prepared the infile for this by computing consensus sequences using ANSGD with the above-mentioned filters. We then performed genome-wide alignments, removed sites with missing data in three or more individuals, sites where singletons occurred within the brown hyena ingroup and invariant site positions using a custom Perl script.
We then repeated the phylogenetic and IBS PCA analyses using single scaffolds. PCA analyses were carried out using nine independent analyses (scaffolds 0-8) and Maximum likelihood analyses were carried out independently for single scaffolds with a length larger than 2MB. We calculated admixture proportions using NGSadmix (Skotte et al. 2013) setting K values from 2-7. We used ANGSD genotype likelihood values as input, only including SNPs with a p value of less than 1e-6. NGSadmix analyses were repeated a maximum of 100 times per K. Only those that converged (produced a consistently identical likelihood score) within these 100 analyses were considered as meaningful. D-statistic analyses were implemented in ANGSDv0.913, sampling a single base per site while specifying the striped hyena as outgroup with default parameters.

4.5.10 Comparative population structures

In order to compare the population structure within the brown hyena to those of other species, 10 individuals per species were randomly selected for a number of different mammalian species for which such data was publically available (Appendix B: Supplemental Table S6). Comparisons between individuals were performed using single base IBS, only considering sites where at least 7 individuals had coverage and SNPs that occurred in at least 2 individuals. Other filtering options included; a minimum base quality score of 25 (-minQ 25), minimum mapping quality score of 25 (-minMapQ 25) and reads that mapped uniquely to one location (-unique_only 1). Quality scores around indels were also adjusted for (-baq 1).

4.5.11 Species heterozygosity estimates

High coverage, single individual representatives for a number of species were assessed for heterozygosity levels. Raw data were selected from a range of different species (Appendix B: Supplemental Table S7). Raw reads were all treated comparably, using Cutadapt v1.8.1 (Martin 2011) to trim Illumina adapter sequences and FLASH v1.2.1 (Magoč and Salzberg 2011) to merge overlapping reads. We mapped each species to its respective reference sequence using BWAv0.7.15 (Li and Durbin 2009) and processed the mapped reads further using Samtools v1.3.1 (Li et al. 2009). To adjust for biases introduced by unequal levels of coverage, the resulting bam files were all subsampled to 20x using Samtools (Li et al. 2009). The folded site frequency spectrum and therefore heterozygosity was then calculated from sample allele frequencies, taking genotype likelihoods into account for each species representative calculated using ANGSDv0.913 (Korneliussen et al. 2014). For the 1Mbp window analysis, we used the same analysis to estimate heterozygosity as before but performed the analysis in non-overlapping windows of 1MB. We also removed any scaffolds less than 1Mbp in length.

4.5.12 Demographic inference

The demographic history of the brown hyena was calculated using PSMC (Li and Durbin 2011), considering only the autosomal chromosomes. Scaffolds representing the X chromosome of the striped hyena were determined through a synteny analysis to the cat X chromosome (CM001396.2) using Satsuma synteny (Grabherr et al. 2010). These scaffolds were then removed along with any scaffold shorter than 1Mbp. A consensus diploid sequence was constructed using Samtools (Li et al. 2009) to be used as input for PSMC. PSMC was implemented using parameters previously shown to be meaningful when considering human data. 100 bootstrap analyses were undertaken. When plotting, we used a generation time of 6 years and a mutation rate of 7.5x10-9 per generation for autosomes.

In order to estimate the mutation rate, we carried out a pairwise distance analysis on the striped and brown hyena's autosomes using a consensus base IBS approach in ANGSDv0.913. We then calculated the average per generation mutation rate assuming a divergence date of the two species to be 4.2mya (Koepfli et al. 2006), a genome-wide strict molecular clock and a generation time of 6 years. Additional analyses utilising different mutation rates based on the 95% confidence interval of the brown and striped hyena divergence (2.6mya and 6.4mya) can be seen in Appendix B: Supplemental Fig. S3.

Data access:

Raw sequencing reads can be found under the accession codes XXXXXX. The striped hyena nuclear genome assembly can be found at XXXXXX and the brown hyena mitochondrial genomes can be found at XXXXXX.

Acknowledgements and funding:

This work was supported by the European Research Council (consolidator grant GeneFlow # 310763 to M.H.). The authors also acknowledge support from Science for Life Laboratory, the Knut and Alice Wallenberg Foundation, the National Genomics Infrastructure funded by the

Swedish Research Council, and Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing, as well as *de novo* assembly of the striped hyena and access to the UPPMAX computational infrastructure. L.D. acknowledges support from the Swedish Research Council and FORMAS. We would like to thank Prof. Yoshan Moodley for his suggestions on the manuscript. We would finally like to thank Binia De Cahsan for producing the animal icons found in Figures 4.1, 4.2 and 4.3.

Author Contributions:

The project was conceived by M.W. and M.H. M.W. and L.D. performed lab work; M.W., S.H., and A.B. performed DNA analyses and interpretation of results. I.W., V.L., R.W, D.M.P., F.S., and A.L. assisted with locating and sampling of specimens. Final editing and manuscript preparation was coordinated by M.W. All contributing authors read and agreed to the final manuscript.

Chapter 5: Discussion

A wide variety of species currently occupy the earth, all of which have undergone their own, unique evolutionary paths. Therefore, in order to understand evolution as a whole, the evolutionary history and relationships of a phylogenetically diverse range of species need to be investigated. Unfortunately, not all species were created equal in a scientific sense and not all species can be model organisms with a lot of funding and research devoted to them. The increase in low cost, high throughput sequencing, led by the invention of NGS, has allowed for a shift in the species that are studied within evolutionary genetics as massive sums of money and big international consortiums necessary are no longer necessary to produce large amounts of nuclear genomic sequence information. On top of the low price and high output, NGS also allows obtaining relatively easily high volumes of genetic information from species on which little to no previous work has been performed as it does not require prior sequence knowledge for designing primers or baits for capture. Thanks to NGS, large nuclear genomic datasets can now be relatively easily and cost effectively produced on a smaller scale, allowing for more research to be done on little studied, wild species and populations. The three case studies presented within this thesis highlight the ability NGS has in allowing us to study such understudied species.

5.1 Aims and importance of the thesis

This thesis set out to use NGS approaches, more specifically shotgun sequencing of Illumina data, to better understand the evolutionary history and current relationships of three focus species. All three species within this thesis, the bat eared fox (*Otocyon megalotis*), *Macrauchenia patachonica* and the brown hyena *(Parahyaena brunnea)* are hardly studied when it comes to genetic information. The first two case study chapters (Chapters 2 and 3) set out to infer species level relationships through phylogenetic analyses of the mitochondrial genome. The third case study (Chapter 4) set out to infer relationships on a population level and demographic and genetic diversity estimates on a species level. The results presented in this thesis not only produced novel findings about the evolutionary history and relationships of the species involved,

but also produced valuable scientific resources in the form of new mitochondrial and nuclear genomes that can be used by the wider research community to further the research on these species.

5.2 Evolutionary insights

The three case studies presented within this thesis highlight, in their own unique ways, the power that NGS has over more traditional techniques to not only better understand the evolution of a species but also for assisting in the generation of novel ideas, data and methods for future research. For the case of the bat eared fox seen in Chapter 2, a shotgun sequencing approach followed by an iterative mapping assembly of the mitochondrial genome produced a novel complete mitochondrial genome. Phylogenetic analyses using this novel mitochondrial genome provided additional evidence for the phylogenetic placement of the bat eared fox as sister to true foxes and the raccoon dog, confirming previous studies investigating the phylogenetic placement of the species with nuclear and mitochondrial genes (Lindblad-Toh et al. 2005). Due to the basal positioning of this species within the canidae family tree, making phylogenetic inferences more difficult, it is important to include additional data in order to provide further evidence to previous findings. Furthermore, not only does this mitochondrial genome provide insight into the phylogenetic placement of the bat eared fox within canidae but it will also be a valuable resource for future studies as it can be used to design primers for PCR or produce baits for hybridisation capture approaches.

For the case of the *Macrauchenia*, seen in Chapter 3, the work done in this thesis showed that even when a close reference is unavailable for a direct mapping assembly, the use of iterative mapping of shotgun data complemented with strict parameters and multiple bait reference sequences consisting of very distantly related species can be used to successfully reconstruct long stretches of mitochondrial DNA from ancient specimens. Phylogenetic analyses of this near complete mitochondrial genome placed the *Macrauchenia* as sister to all living Perissodactyla with a divergence time of approximately 66Ma, very close to the radiation of the

major orders within Laurasiatheria. This deep divergence coupled with the fact that the *Macrauchenia* were only found in South America, where DNA preservation is known to be poor, were most likely what prevented previous attempts at recovering DNA from this species using more traditional methods such as PCR and Sanger sequencing. Results showed that reference sequences from species over 66 million years diverged from the target species can produce consistent results if other parameters were implemented correctly. This result will widely open up the field of ancient DNA to a number of species that were previously deemed impossible to study because of their deep divergence times from their closest relatives. It also provides initial evidence for the power these types of methods can have, especially for the use of aDNA, giving hope that perhaps in the future, as computer power and sequencing costs further decrease, nuclear genomes from ancient species could be assembled in a similar manner.

The case study presented in Chapter 4 supplies a major argument in favour of the analysis of whole genomic data for the study of populations, especially species suspected of having low levels of genetic diversity. The high number of independent loci found within complete nuclear genomes provides both power and accuracy to bioinformatic analyses. Complete nuclear genomes allow studying population differences on a much finer scale, uncovering structure that may not be found with fewer markers. Low genetic diversity within the brown hyena hindered the ability of population structure analyses when considering only a few loci. This hindrance could be seen in the case of microsatellites and short mitochondrial fragments investigated in previous studies using Sanger sequencing (Rohland et al. 2005; Knowles et al. 2009) and even when considering whole scaffolds greater than 2Mbp is size (Chapter 4). Analyses using complete nuclear genomes, however, were able to find population structure within the brown hyena suggesting the presence of an isolating mechanism between populations which will have important implications for how future conservation approaches need to be implemented. Furthermore, if one does not look into the complete nuclear genome, signs of inbreeding may be harder to detect. When considering long contiguous stretches of genetic information such as those present in whole genome assemblies, levels of homozygosity across the genome can be analysed. Long stretches of homozygosity have been shown to be a good proxy for evaluating

levels of inbreeding (Kirin et al. 2010). This kind of analysis would not be possible to analyse if only SNP data or short contiguous regions of the genome were available. Without a high quality nuclear genome, the interesting result of low heterozygosity despite no signs of inbreeding would not have been discovered in the brown hyena. As this result goes against the normally perceived concept that low levels of heterozygosity are caused by inbreeding, without this additional information, one may have just assumed it followed the expected pattern without further investigations, therefore missing an important result.

5.3 Conservation

While NGS has a lot of benefits for the scientific community, it also has a number of benefits to the wider community. Conservation is a topic that many people in the wider community understand as important. The results of this thesis, most specifically those in Chapter 4, have important implications for the wider community and especially for the implementation of conservation approaches.

Since the inclusion of genetic approaches into the field, the ability of conservation biologists to accurately define important factors for conservation has increased. The use of genetic techniques can aid with the definition of conservation units, population structure and connectivity, taxonomic relationships, genetic diversity and demographic history of a species (Shafer et al. 2015; Allendorf et al. 2010; Steiner et al. 2013), among others. This information can greatly help to accurately determine conservation approaches specifically tuned to the species and populations of interest. The implementation of traditional conservation genetics approaches, involving the use of just a few genetic loci from a large number of individuals, into the conservation of many species has been successful (McCormack et al. 2013). However, although using just a few genetic loci can provide valuable information, difficulties in obtaining large enough quantities of individuals can often hamper results. Due to the limited number of loci, the differentiation between real relationships and noise can be difficult, making this kind of analysis unfeasible for a number of species especially when only a few individuals remain and

are available to study (McMahon et al. 2014). This problem is very clearly highlighted within the brown hyena as even when whole scaffolds, consisting of many Mbp of genetic data, were analysed, one could see inconsistencies within the results, perhaps stemming from the very low diversity within the species. An obvious solution to this problem would be to include more loci from the few individuals available in the form of whole genomic information.

With the ever decreasing price of next generation sequencing, conservation genetics has already begun to move towards the inclusion of new genomic techniques to solve conservation related questions. The use of new genomic techniques for conservation purposes has given rise to the term "conservation genomics" (Allendorf et al. 2010). Conservation genomics allows the inclusion of drastically more loci than traditional methods, producing more accurate, reliable population level results even when using a small number of individuals (Shafer et al. 2015; Allendorf et al. 2010). Some analyses even allow the use of a single individual for demographic inferences (Li and Durbin 2011). In my opinion, this is the natural next step for conservation genetics as it also limits the potential impacts that sampling of endangered species can cause while not decreasing the accuracy.

5.4 Bioinformatic advances

Black-box tools available for the analysis of NGS data have become a valuable resource for many analyses but in order to get the best results, one must also understand the type of data produced and adjust the analyses to suit their specific data. This thesis highlighted the importance of bioinformatic parameters and the correct implementation of the software specifically tailored to one's data. For example, when singletons were not removed from population analyses involving the brown hyena (Chapter 4), the low coverage nature of the data, coupled with the error rates produced by Illumina sequencing (1/1000 for >Q30) led to difficulties differentiating relationships from noise. This difficulty arose as all individuals had large numbers of unique "SNPs" many of which were probably originating from the sequencing errors within the data. Furthermore, the standardisation of the data when performing

heterozygosity analyses was important in order to reduce inconsistencies in the results which may arise from the use of different analysis approaches, software and coverage.

Publically available online databases, such as Genbank sequence read archive (SRA), containing the raw sequencing reads from other projects are very important resources as they allow independently produced data to be processed in comparable ways, thus enabling the production of the most consistent results possible. In Chapter 4, reads were downloaded from this database and equally subsampled for all species undergoing heterozygosity comparisons against the brown hyena as, depending on the coverage of the individual, it was noted that the levels of genome-wide heterozygosity changed. This stems from the fact that the higher the coverage of a genome, the more likely one is to find heterozygous positions simply because there are more reads mapping to a single locus. This highlights the importance of maintaining consistency between analyses as opposed to simply relying on previously reported numbers which may have been produced following different methods and contain biases not accounted for.

On top of the analyses implemented post-assembly, this thesis shows that the selection of reference can also have a dramatic influence on the final product, especially with the use of ancient DNA and species with only very distantly related modern reference sequences. This is very clearly seen in the *Macrauchenia* of Chapter 3. Depending on which starting reference one selects, variabilities arise between the final products. In order to achieve the most accurate results, it is important to use multiple references, strict mismatch values and strict consensus calling thresholds.

5.5 NGS in the future

While there has been a more recent induction of new sequencing technologies termed "third generation sequencing" in the form of, for example, PacBIO (Biosciences) and Oxford Nanopore, next generation sequencing technologies will still be a valuable resource for many

years to come. These new technologies utilise long read sequencing technologies making them very useful for *de novo* assemblies of complex genomes (Bleidorn 2016). The long read lengths however, have their own downfalls. Reads obtained with these technologies are currently plagued with high sequencing error rates and relatively high costs when compared to NGS technologies such as Illumina. Due to this, in many cases, Illumina sequencing reads are used as a necessary companion to the long reads. Short reads can be used to correct for the high error rate of long reads for much lower costs than if purely third generation sequencing reads were used (Goodwin et al. 2015). Just like with NGS, the prices and output of these third generation technologies are also constantly changing and improving perhaps eventually leading to a decrease in the need for NGS reads. However, despite these constant improvements, it does not seem like NGS will become superfluous in the near future. New, more powerful NGS platforms will allow this technology to stay relevant. The most recent of these is the Illumina Novaseq. The Novaseq promises to produce extremely large amounts of data for very low costs and will be extremely useful for use in resequencing projects and mapping assemblies. On top of the increase in power of the platforms, new machines complementing Illumina short read technologies have been developed. One example is the 10x genomics machine. The 10x allows for many separate long fragments to be built into independently barcoded sequencing libraries, which can be sequenced on an Illumina platform. The output from these libraries then undergo standalone assemblies to reconstruct the initial long fragments. These assemblies are further assembled together into the complete genome. These independent libraries are designed to overcome the problem repeats dispersed throughout the genome can cause to *de novo* assemblies while also being useful for the phasing of genomes (Mostovoy et al. 2016). In theory, this works in a similar fashion to the reads produced using third generation sequencing technologies in their ability to overcome repeats while still utilising short read sequencing and therefore are only a fraction of the price of the new technologies. Furthermore, new NGS platforms and technologies are still being developed, providing competition for current technologies. One recently developed machine is the BGIseq-500. This machine can currently only produce short reads with a maximum size of 100bp but has shown promise, especially for its use in ancient DNA studies as only short reads are needed for the highly fragmented aDNA (Mak et al. 2017) and has been

shown to produce results comparable to those produced through Illumina sequencing. Current NGS technologies also began only being able to produce short fragments so this is an important first step for this technology. This new machine provides competitive prices when compared to Illumina, which currently more or less has the monopoly over the NGS world. This should in turn force the pricing of this well established sequencing technology to further decrease.

The decrease in price possible from upgrading existing NGS platforms, the production of new technologies complementing and taking full advantage of existing NGS platforms and new competitive NGS platforms will allow NGS to persist at competitive prices playing a valuable role in evolutionary biology for years to come.

5.6 General conclusions

The inclusion of DNA into the field of evolutionary biology was a huge milestone and led to a great increase in the understanding and knowledge on the process of evolution as a whole. The next major milestone for the field of evolutionary biology was the invention of NGS. By opening up the use of large amounts of DNA markers to the study of a wide range of phylogenetically diverse taxa, our understanding of the natural world has only just begun. The massive amounts of data can at times be overwhelming but as this thesis shows, if one has an understanding of the nature of the data and adjusts the analyses to suit their data, one can expand our understanding of evolution through novel and cost-effective analyses.

Bibliography

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Abascal F, Corvelo A, Cruz F, Villanueva-Cañas JL, Vlasova A, Marcet-Houben M, Martínez-Cruz B, Cheng JY, Prieto P, Quesada V, et al. 2016. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. *Genome Biol* **17**: 251.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. 2002. *Isolating, Cloning, and Sequencing DNA*. Garland Science.

Allendorf FW, Hohenlohe PA, Luikart G. 2010. Genomics and the future of conservation genetics. *Nat Rev Genet* **11**: 697–709.

Bartlett JMS, Stirling D. 2003. A short history of the polymerase chain reaction. *Methods Mol Biol* **226**: 3–6.

Bazin E, Glémin S, Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. *Science* **312**: 570–572.

Beck RMD, Lee MSY. 2014. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. *Proc Biol Sci* **281**: 20141278.

Benbow ME, Tomberlin JK, Tarone AM. 2015. *Carrion ecology, evolution, and their applications*. CRC Press.

Benton MJ, Donoghue P, Asher RJ. 2009. The timetree of life. *Calibrating and constraining molecular clocks* 35–86.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M, Stadler PF. 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol* **69**: 313–319.

Bleidorn C. 2016. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *System Biodivers* **14**: 1–8.

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

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**: 314–331.

Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* **26**:

1372–1373.

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, et al. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A* **104**: 14616–14621.

Britten RJ, Kohne DE. 1968. Repeated Sequences in DNA. *Science* **161**: 529–540.

Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, Austin J, Cooper A. 2007. Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res* **35**: 5717–5728.

Buckley M. 2015. Ancient collagen reveals evolutionary history of the endemic South American "ungulates." *Proc Biol Sci* **282**: 20142671.

Buerkle AC, Gompert Z. 2013. Population genomics based on low coverage sequencing: how low should we go? *Mol Ecol* **22**: 3028–3035.

Buermans HPJ, den Dunnen JT. 2014. Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta* **1842**: 1932–1941.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* **18**: 810–820.

Cahill JA, Green RE, Fulton TL, Stiller M, Jay F, Ovsyanikov N, Salamzade R, St John J, Stirling I, Slatkin M, et al. 2013. Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. *PLoS Genet* **9**: e1003345.

Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, Rasmussen M, Gravel S, Guillén S, Nekhrizov G, et al. 2013. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am J Hum Genet* **93**: 852–864.

Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. In *German conference on bioinformatics*, Vol. 99 of, pp. 45–56, Heidelberg.

Cho YS, Hu L, Hou H, Lee H, Xu J, Kwon S, Oh S, Kim H-M, Jho S, Kim S, et al. 2013. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nat Commun* **4**: 2433.

Cifelli RL. 1993. The phylogeny of the native South American ungulates. *Mammal phylogeny* **2**: 195–216.

Clark HO Jr. 2005. Otocyon megalotis. *Mammalian Species* **776**: 1–5.

Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, Valdiosera C, García N, Pääbo S, Arsuaga J-L, et al. 2013a. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc Natl Acad Sci U S A* **110**: 15758–15763.

Dabney J, Meyer M, Pääbo S. 2013b. Ancient DNA damage. *Cold Spring Harb Perspect Biol* **5**: a012567.

Damgaard PB, Margaryan A, Schroeder H, Orlando L, Willerslev E, Allentoft ME. 2015. Improving access to endogenous DNA in ancient bones and teeth. *Sci Rep* **5**: 11184.

deMenocal PB. 2004. African climate change and faunal evolution during the Pliocene–Pleistocene. *Earth Planet Sci Lett* **220**: 3–24.

Dobrynin P, Liu S, Tamazian G, Xiong Z, Yurchenko AA, Krasheninnikova K, Kliver S, Schmidt-Küntzel A, Koepfli K-P, Johnson W, et al. 2015. Genomic legacy of the African cheetah, Acinonyx jubatus. *Genome Biol* **16**: 277.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol* **29**: 1969–1973.

Ekblom R, Galindo J. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**: 1–15.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends Ecol Evol* **29**: 51–63.

Foote AD, Vijay N, Ávila-Arcos MC, Baird RW, Durban JW, Fumagalli M, Gibbs RA, Bradley Hanson M, Korneliussen TS, Martin MD, et al. 2016. Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat Commun* **7**: 11693.

Froehlich DJ. 2002. Quo vadis eohippus? The systematics and taxonomy of the early Eocene equids (Perissodactyla). *Zool J Linn Soc* **134**: 141–256.

Fumagalli M, Vieira FG, Linderoth T, Nielsen R. 2014. ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics* **30**: 1486–1487.

Gansauge M-T, Meyer M. 2013. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat Protoc* **8**: 737–748.

Gernhard T. 2008. The conditioned reconstructed process. *J Theor Biol* **253**: 769–778.

Goillot C, Antoine P-O, Tejada J, Pujos F, Gismondi RS. 2011. Middle Miocene Uruguaytheriinae (Mammalia, Astrapotheria) from Peruvian Amazonia and a review of the

astrapotheriid fossil record in northern South America. *Geodiversitas* **33**: 331–345.

Gonzales-Fortes G, Paijmans JLA. 2015. Analysis of Whole Mitogenomes from Ancient Samples. *Methods Mol Biol* **1347**: 179–195.

Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res* **25**: 1750–1756.

Grabherr MG, Russell P, Meyer M, Mauceli E, Alfoldi J, Di Palma F, Lindblad-Toh K. 2010. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26**: 1145–1151.

Grada A, Weinbrecht K. 2013. Next-generation sequencing: methodology and application. *J Invest Dermatol* **133**: e11.

Gusset M, Burgener N. 2005. Estimating larger carnivore numbers from track counts and measurements. *Afr J Ecol* **43**: 320–324.

Hahn C, Bachmann L, Chevreux B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res* **41**: e129.

Hailer F, Kutschera VE, Hallström BM, Klassert D, Fain SR, Leonard JA, Arnason U, Janke A. 2012. Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. *Science* **336**: 344–347.

Harrison RG. 1989. Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. *Trends Ecol Evol* **4**: 6–11.

Hartl DL, Clark AG. 1997. *Principles of population genetics*. Sinauer associates Sunderland.

Herbert S. 1980. *The red notebook of Charles Darwin*. Cornell University Press.

Hillis DM. 1987. Molecular Versus Morphological Approaches to Systematics. *Annu Rev Ecol Syst* **18**: 23–42.

Hofreiter M, Paijmans JLA, Goodchild H, Speller CF, Barlow A, Fortes GG, Thomas JA, Ludwig A, Collins MJ. 2015. The future of ancient DNA: Technical advances and conceptual shifts. *Bioessays* **37**: 284–293.

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S. 2001. Ancient DNA. *Nat Rev Genet* **2**: 353–359.

Hutchison CA 3rd, Newbold JE, Potter SS, Edgell MH. 1974. Maternal inheritance of mammalian mitochondrial DNA. *Nature* **251**: 536–538.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**: 1682–1684.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Kehlmaier C, Barlow A, Hastings AK, Vamberger M, Paijmans JLA, Steadman DW, Albury NA, Franz R, Hofreiter M, Fritz U. 2017. Tropical ancient DNA reveals relationships of the extinct Bahamian giant tortoise Chelonoidis alburyorum. *Proc Biol Sci* **284**: 20162235.

Kent VT, Hill RA. 2013. The importance of farmland for the conservation of the brown hyaena Parahyaena brunnea. *Oryx* **47**: 431–440.

Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. 2010. Genomic Runs of Homozygosity Record Population History and Consanguinity. *PLoS One* **5**: e13996.

Knapp M, Hofreiter M. 2010. Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives. *Genes* **1**: 227–243.

Knowles JC, Van Coeverden de Groot PJ, Wiesel I, Boag PT. 2009. Microsatellite Variation in Namibian Brown Hyenas (Hyaena brunnea): Population Structure and Mating System Implications. *J Mammal* **90**: 1381–1391.

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. 2013. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**: 27–38.

Koepfli K-P, Jenks SM, Eizirik E, Zahirpour T, Van Valkenburgh B, Wayne RK. 2006. Molecular systematics of the Hyaenidae: relationships of a relictual lineage resolved by a molecular supermatrix. *Mol Phylogenet Evol* **38**: 603–620.

Koepfli K-P, Pollinger J, Godinho R, Robinson J, Lea A, Hendricks S, Schweizer RM, Thalmann O, Silva P, Fan Z, et al. 2015. Genome-wide Evidence Reveals that African and Eurasian Golden Jackals Are Distinct Species. *Curr Biol* **25**: 2158–2165.

Korlević P, Gerber T, Gansauge M-T, Hajdinjak M, Nagel S, Aximu-Petri A, Meyer M. 2015. Reducing microbial and human contamination in DNA extractions from ancient bones and

teeth. *Biotechniques* **59**: 87–93.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**: 356.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**: 1639–1645.

Laine VN, Gossmann TI, Schachtschneider KM, Garroway CJ, Madsen O, Verhoeven KJF, de Jager V, Megens H-J, Warren WC, Minx P, et al. 2016. Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nat Commun* **7**: 10474.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol* **29**: 1695–1701.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol* **10**: e1001388.

Leigh JW, Bryant D. 2015. popart: full-feature software for haplotype network construction. *Methods Ecol Evol* **6**: 1110–1116.

Leonardi M, Librado P, Der Sarkissian C, Schubert M, Alfarhan AH, Alquraishi SA, Al-Rasheid KAS, Gamba C, Willerslev E, Orlando L. 2017. Evolutionary Patterns and Processes: Lessons from Ancient DNA. *Syst Biol* **66**: e1–e29.

le Roux A, Beishuizen R, Brekelmans W, Ganswindt A, Paris M, Dalerum F. 2014. Innovative parental care in a myrmecophageous mammal. *Acta Ethol* **17**: 63–66.

Li H. 2011. Improving SNP discovery by base alignment quality. *Bioinformatics* **27**: 1157–1158.

Li H. 2012. seqtk Toolkit for processing sequences in FASTA/Q formats. https://github.com/lh3/seqtk.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome

sequences. *Nature* **475**: 493–496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ 3rd, Zody MC, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803–819.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al. 2010a. The sequence and de novo assembly of the giant panda genome. *Nature* **463**: 311–317.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al. 2010b. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**: 265–272.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**: 18.

Magoč T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957–2963.

Mak SST, Gopalakrishnan S, Carøe C, Geng C, Liu S, Sinding M-HS, Kuderna LFK, Zhang W, Fu S, Vieira FG, et al. 2017. Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *Gigascience* **6**: 1–13.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Maxson LR, Wilson AC. 1974. Convergent morphological evolution detected by studying proteins of tree frogs in the Hyla eximia species group. *Science* **185**: 66–68.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol* **66**: 526–538.

McKenna MC. 1975. Toward a Phylogenetic Classification of the Mammalia. In *Phylogeny of the Primates* (eds. W. Patrick Luckett and F.S. Szalay), pp. 21–46, Springer US.

McMahon BJ, Teeling EC, Höglund J. 2014. How and why should we implement genomics into

conservation? *Evol Appl* **7**: 999–1007.

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TLL, Stadler T, et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334**: 521–524.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010**: db.prot5448.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**: 222–226.

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, Packard MD, Zhao F, Sher A, et al. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456**: 387–390.

Miller W, Hayes VM, Ratan A, Petersen DC, Wittekindt NE, Miller J, Walenz B, Knight J, Qi J, Zhao F, et al. 2011. Genetic diversity and population structure of the endangered marsupial Sarcophilus harrisii (Tasmanian devil). *Proc Natl Acad Sci U S A* **108**: 12348–12353.

Mitchell KJ, Llamas B, Soubrier J, Rawlence NJ, Worthy TH, Wood J, Lee MSY, Cooper A. 2014. Ancient DNA reveals elephant birds and kiwi are sister taxa and clarifies ratite bird evolution. *Science* **344**: 898–900.

Mitchell KJ, Scanferla A, Soibelzon E, Bonini R, Ochoa J, Cooper A. 2016. Ancient DNA from the extinct South American giant glyptodont Doedicurus sp.(Xenarthra: Glyptodontidae) reveals that glyptodonts evolved from Eocene armadillos. *Mol Ecol* **25**: 3499–3508.

Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Džakula Ž, et al. 2016. A hybrid approach for de novo human genome sequence assembly and phasing. *Nat Methods* **13**: 587–590.

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D. 2009. Population genomics of marine fishes: identifying adaptive variation in space and time. *Mol Ecol* **18**: 3128–3150.

O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo Z-X, Meng J, et al. 2013. The Placental Mammal Ancestor and the Post–K-Pg Radiation of Placentals. *Science* **339**: 662–667.

Owen R, Darwin C. 1840. The Zoology of the voyage of HMS Beagle. *Part I Mammalia Smith Elder & Co , London*.

Pääbo S, Higuchi RG, Wilson AC. 1989. Ancient DNA and the polymerase chain reaction: The emerging field of molecular archaeology (Minireview). *J Biol Chem* **264**: 9709–9712.

Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, Omrak A, Vartanyan S, Poinar

H, Götherström A, et al. 2015. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr Biol* **25**: 1395–1400.

Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**: 228–231.

Peter C, Bruford M, Perez T, Dalamitra S, Hewitt G, Erhardt G. 2007. Genetic diversity and subdivision of 57 European and Middle-Eastern sheep breeds. *Anim Genet* **38**: 37–44.

Posada D. 2008. jModelTest: phylogenetic model averaging. *Mol Biol Evol* **25**: 1253–1256.

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**: 336–341.

Prüfer K, Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE. 2010. Computational challenges in the analysis of ancient DNA. *Genome Biol* **11**: R47.

Qiu Q, Wang L, Wang K, Yang Y, Ma T, Wang Z, Zhang X, Ni Z, Hou F, Long R, et al. 2015. Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. *Nat Commun* **6**: 10283.

Rambaut A. 2014. FigTree v1. 4.2. *Univ Edinb J*.

Rambaut A, Suchard MA, Xie D, Drummond AJ. Tracer v1. 6, 2014 http://beast. bio. ed. ac. uk. *Tracer*.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. http://www.R-project.org.

Reed DH, Frankham R. 2003. Correlation between fitness and genetic diversity. *Conserv Biol* **17**: 230–237.

Robinson JA, Ortega-Del Vecchyo D, Fan Z, Kim BY, vonHoldt BM, Marsden CD, Lohmueller KE, Wayne RK. 2016. Genomic Flatlining in the Endangered Island Fox. *Curr Biol* **26**: 1183–1189.

Rohland N, Pollack JL, Nagel D, Beauval C, Airvaux J, Pääbo S, Hofreiter M. 2005. The population history of extant and extinct hyenas. *Mol Biol Evol* **22**: 2435–2443.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**: 798–804.

Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, et al. 2014. Comparative population genomics in animals uncovers the

determinants of genetic diversity. *Nature* **515**: 261.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**: 441–448.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463–5467.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863–864.

Secord R, Bloch JI, Chester SGB, Boyer DM, Wood AR, Wing SL, Kraus MJ, McInerney FA, Krigbaum J. 2012. Evolution of the earliest horses driven by climate change in the Paleocene-Eocene Thermal Maximum. *Science* **335**: 959–962.

Shafer ABA, Wolf JBW, Alves PC, Bergström L, Bruford MW, Brännström I, Colling G, Dalén L, De Meester L, Ekblom R, et al. 2015. Genomics and the challenging translation into conservation practice. *Trends Ecol Evol* **30**: 78–87.

Shapiro B, Hofreiter M. 2014. A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science* **343**: 1236573.

Sibley CG, Ahlquist JE. 1984. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *J Mol Evol* **20**: 2–15.

Sillero-Zubiri C, Macdonald DW. 2004. *The Biology and Conservation of Wild Canids*. Oxford University Press.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Simpson GG. 1980. *Splendid isolation: the curious history of South American mammals*. Yale University Press.

Simpson GG. 1945. The principles of classification and a classification of mammals. *Bull Am Mus Nat Hist* **85**: xvi+350.

Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, Jakobsson M. 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc Natl Acad Sci U S A* **111**: 2229–2234.

Skotte L, Korneliussen TS, Albrechtsen A. 2013. Estimating individual admixture proportions from next generation sequencing data. *Genetics* **195**: 693–702.

Soria MF. 2001. Los Proterotheriidae (Mammalia, Litopterna), sistemática, origen y filogenia.

Monogr Mus Argent Cienc Nat **1**: 1–167.

Spielman D, Brook BW, Frankham R. 2004. Most species are not driven to extinction before genetic factors impact them. *Proc Natl Acad Sci U S A* **101**: 15261–15264.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Steiner CC, Putnam AS, Hoeck PEA, Ryder OA. 2013. Conservation genomics of threatened animal species. *Annu Rev Anim Biosci* **1**: 261–281.

Swanson KW, Irwin DM, Wilson AC. 1991. Stomach lysozyme gene of the langur monkey: tests for convergence and positive selection. *J Mol Evol* **33**: 418–425.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**: 2725–2729.

Turner A, Antón M, Werdelin L. 2008. Taxonomy and evolutionary patterns in the fossil Hyaenidae of Europe. *Geobios Mem Spec* **41**: 677–687.

Watts HE, Holekamp KE. 2007. Hyena societies. *Curr Biol* **17**: R657–60.

Weise FJ, Wiesel I, Jr JL, van Vuuren RJ. 2015. Evaluation of a Conflict-Related Brown Hyaena Translocation in Central Namibia. *S Afr J Wildl Res* **45**: 178–186.

Welch RJ, Parker DM. 2016. Brown hyaena population explosion: rapid population growth in a small, fenced system. *Wildl Res* **43**: 178–187.

Welker F, Collins MJ, Thomas JA, Wadsley M, Brace S, Cappellini E, Turvey ST, Reguero M, Gelfo JN, Kramarz A, et al. 2015. Ancient proteins resolve the evolutionary history of Darwin/'s South American ungulates. *Nature* **522**: 81–84.

Werdelin L, Barthelme J. 1997. Brown hyena (Parahyaena brunnea) from the Pleistocene of Kenya. *J Vert Paleontol* **17**: 758–761.

Werdelin L, Solounias N. 1991. The Hyaenidae: taxonomy, systematics and evolution. *Fossils and Strata* **30**: 1–104.

Westbury M, Prost S, Seelenfreund A, Ramírez J-M, Matisoo-Smith EA, Knapp M. 2016. First complete mitochondrial genome data from ancient South American camelids - The mystery of the chilihueques from Isla Mocha (Chile). *Sci Rep* **6**: 38708.

Wiesel I. 2015. Parahyaena brunnea. *The IUCN Red List of Threatened Species 2015: eT10276A82344448*. http://dx.doi.org/10.2305/IUCN.UK.2015-4.RLTS.T10276A82344448.en.

Willerslev E, Cooper A. 2005. Ancient DNA. *Proc Biol Sci* **272**: 3–16.

Willoughby JR, Fernandez NB, Lamb MC, Ivy JA, Lacy RC, DeWoody JA. 2015. The impacts of inbreeding, drift and selection on genetic diversity in captive breeding populations. *Mol Ecol* **24**: 98–110.

Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S. 2012. FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One* **7**: e52249.

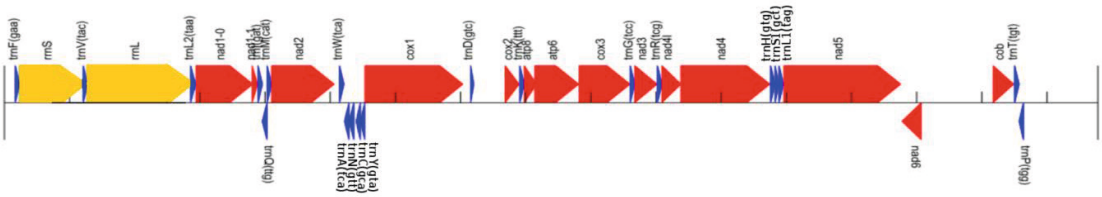
Yule GU. 1925. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical transactions of the Royal Society of London Series B, containing papers of a biological character* **213**: 21–87.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**: 1311–1320.

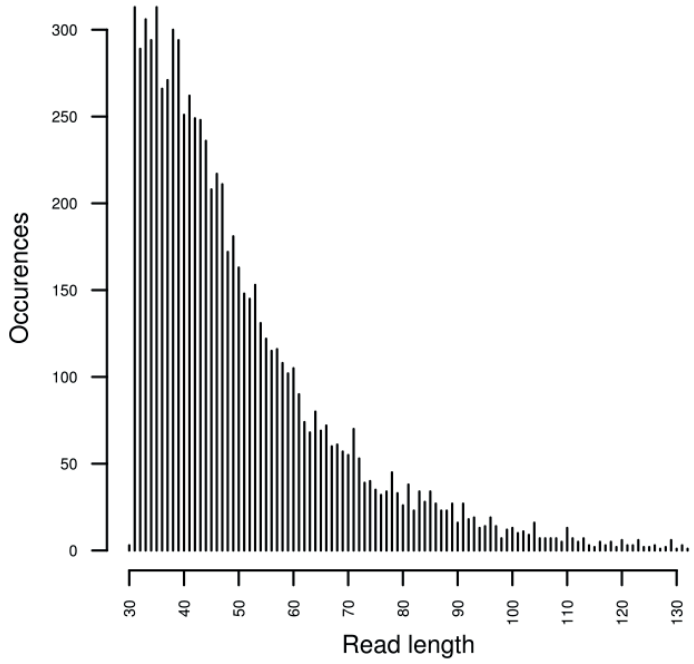
Appendix A.

Supporting information for Chapter 4: Article II, Westbury et al 2017 **A mitogenomic timetree for Darwin‘s enigmatic South American mammal *Macrauchenia patachonica***

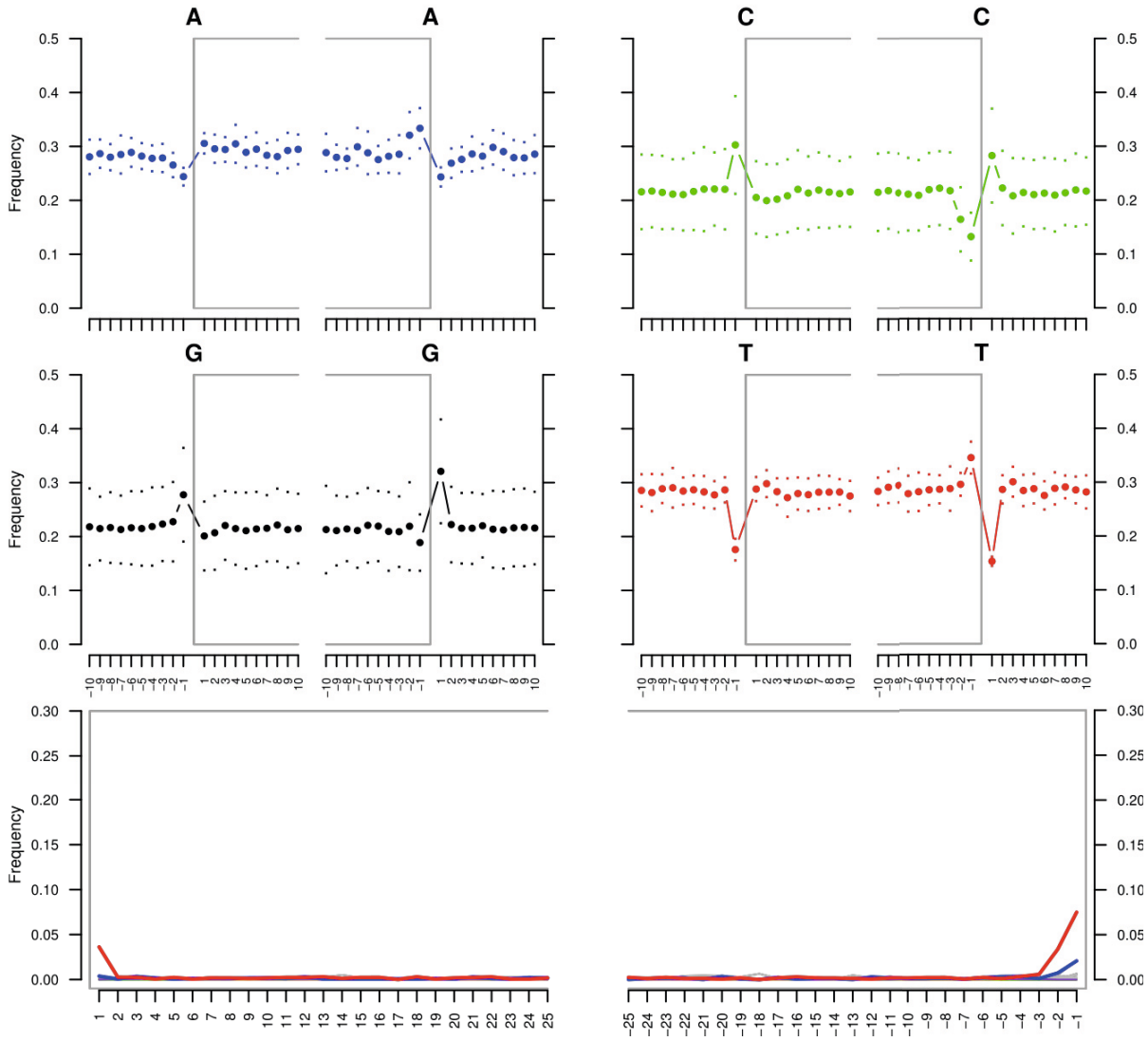
Supplementary Figures



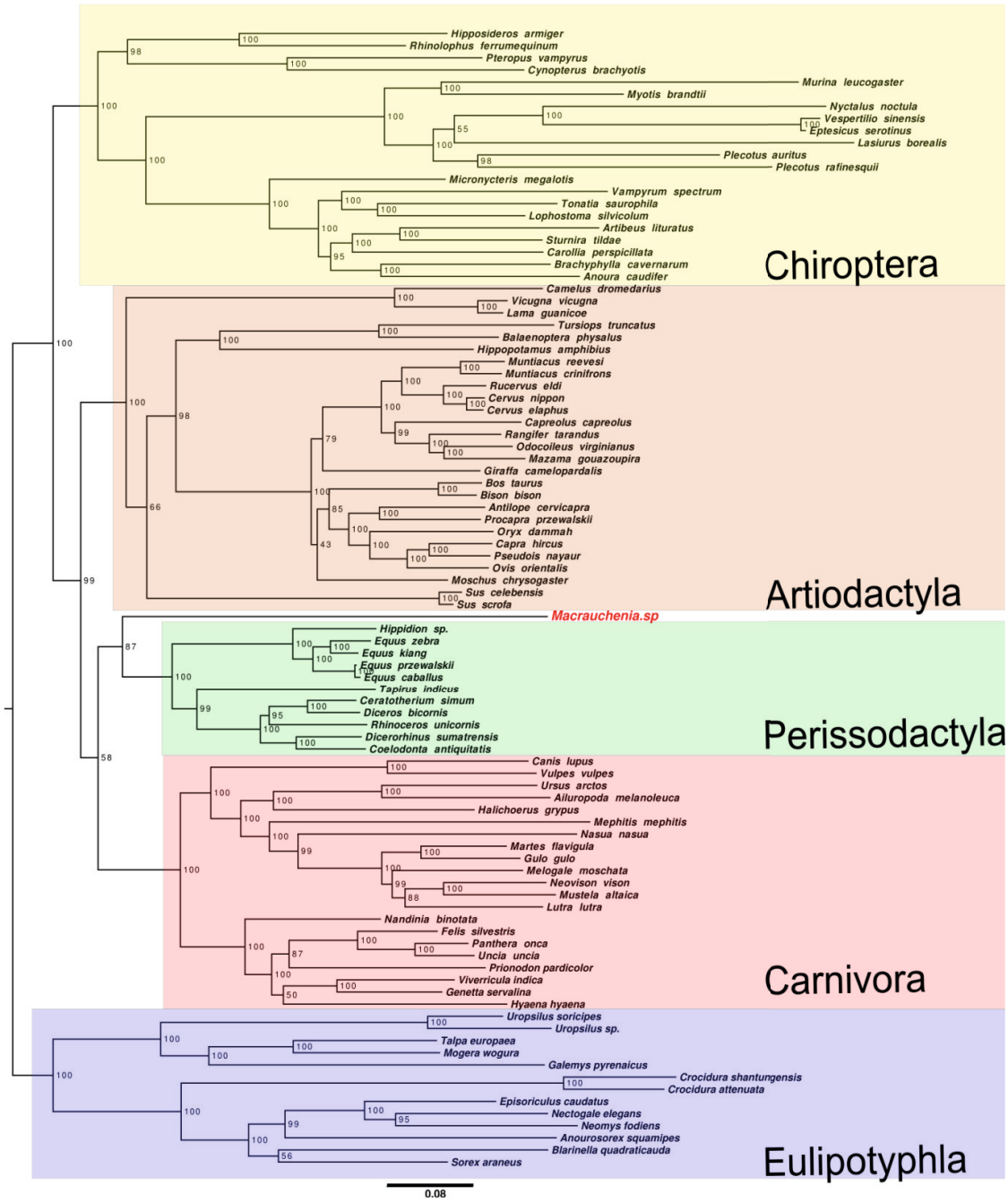
Supplementary Figure 1. MITOS output. Protein coding gene (red), tRNA (blue) and rRNA (yellow) presence and approximate location along the reconstructed MAC002 mitochondrion.



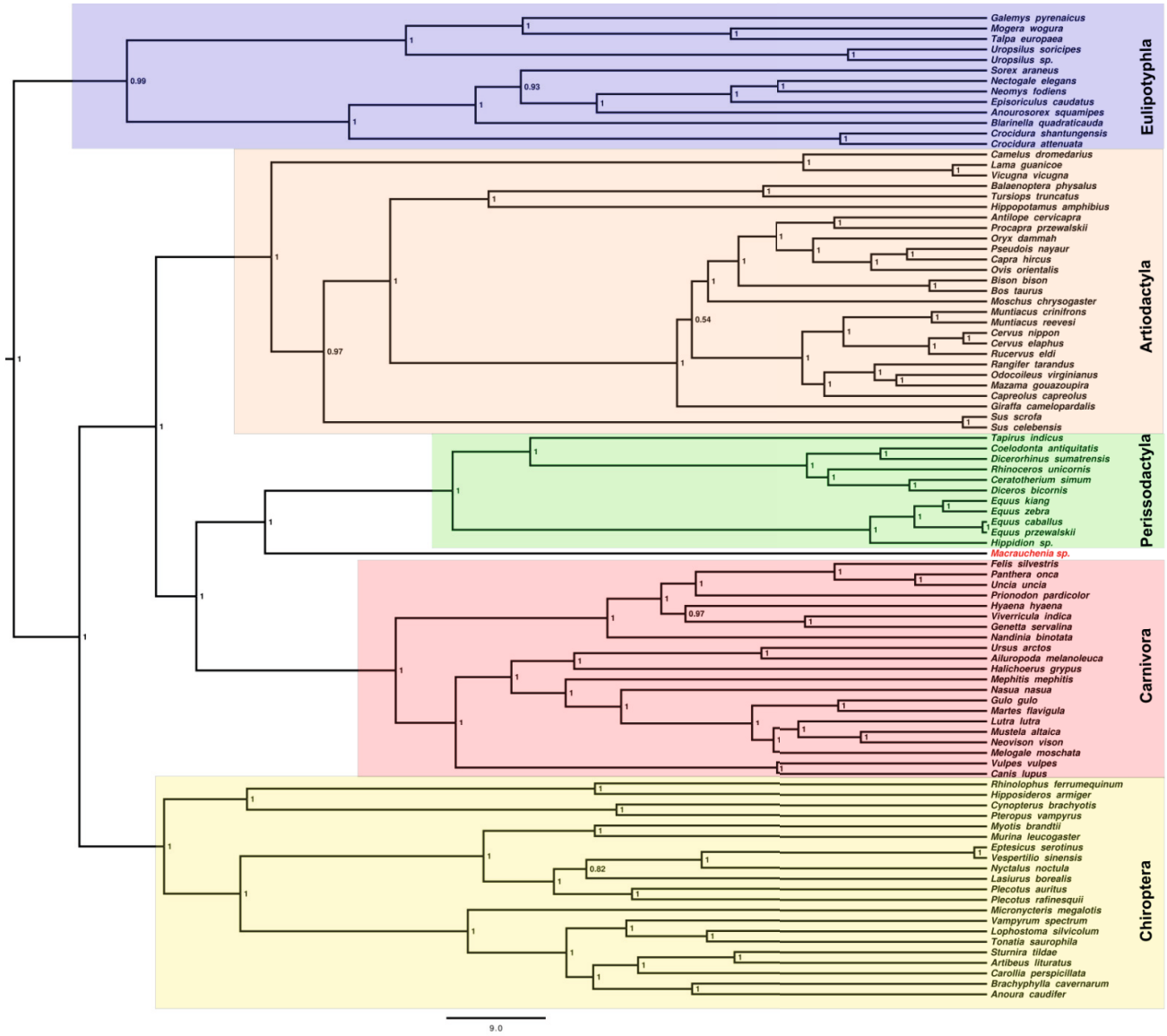
Supplementary Figure 2. Read length Mapdamage output. Read length distribution of MAC002 reads mapped to our *Macrauchenia* mitochondrial sequence.



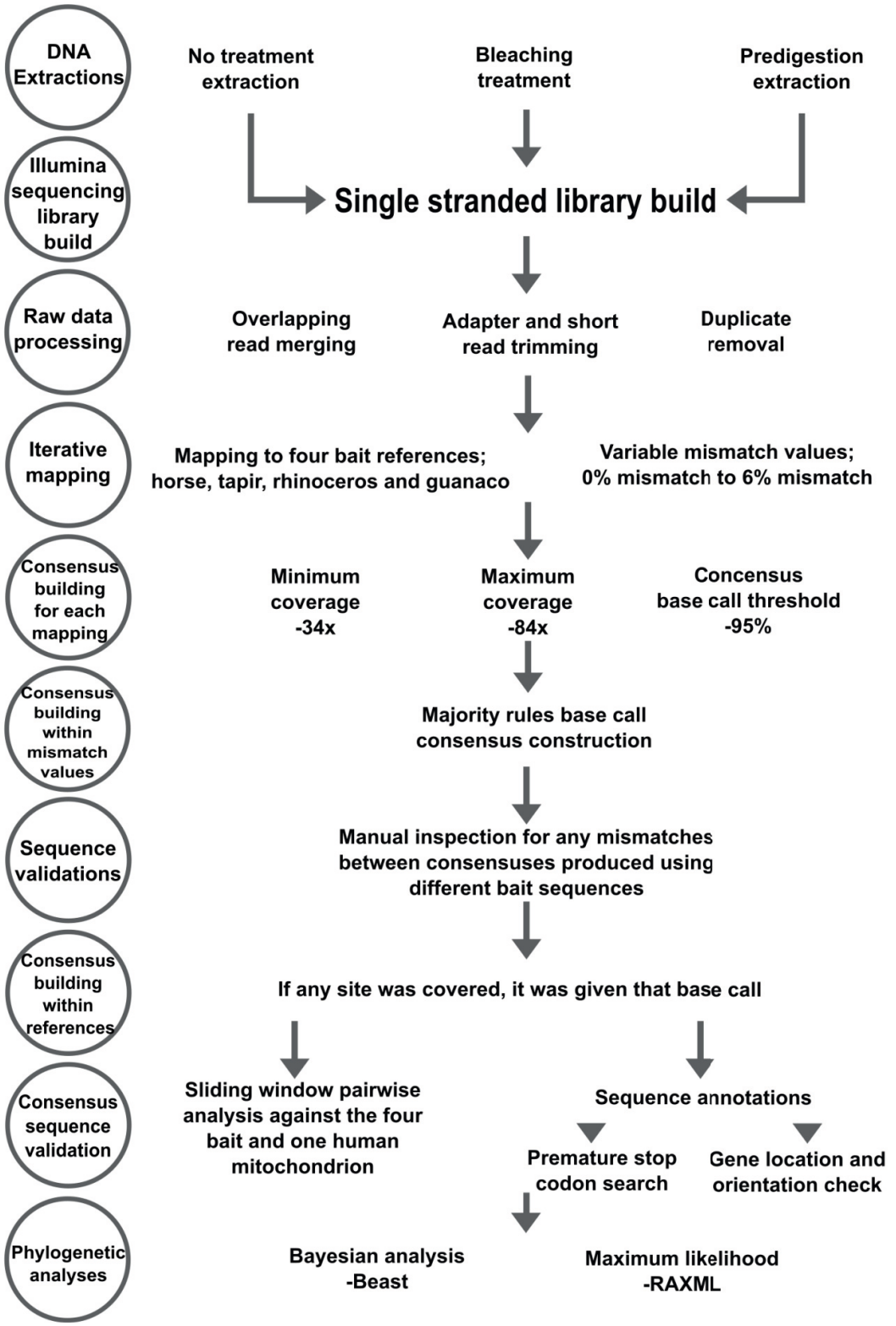
Supplementary Figure 3: Damage pattern Mapdamage results. There is an increased number of T bases (indicated in red) at the ends of reads when reads were mapped back to our MAC002 *Macrauchenia* mitochondrial sequence. X axis represents position from 5' (left) and 3' (right) read end.



Supplementary Figure 4. Maximum likelihood mitochondrial genome tree for our complete dataset consisting of the *Macrauchenia* and selected representatives of the Laurasiatheria superorder. Numbers at nodes represent bootstrap values. *Macrauchenia* is sister to the Perissodactyla clade. Scale bar indicates substitutions per site.



Supplementary Figure 5. Bayesian mitochondrial genome tree for our complete dataset consisting of the *Macrauchenia* and selected representatives of the Laurasiatheria superorder. Posterior clade probabilities are indicated on nodes. Scale bar represents millions of years.



Supplementary Figure 6. Diagram presenting the work-flow of methods used to construct and analyse the mitochondrial genome of MAC002.

Supplementary Tables

Supplementary Table 1. Details of *Macrauchenia* and *Toxodon* samples used in this report.

Collection No.	Sample ID	Species	Locality	Latitude (S); longitude (W)	Material
GCF-1	TOX2	*Toxodon platensis*	Campo Spósito, San Pedro, Argentina	33°44'; 59°36'	petrosal
MACN-PV 5712	TOX005	*Toxodon platensis*	Tapalqué, Buenos Aires, Argentina	36°21'; 60°01' *	Phalanx & carpal bone
MACN-PV 5718	TOX008	*Toxodon platensis*	Tapalqué, Buenos Aires, Argentina	36°21'; 60°01' *	petrosal
MACN-PV 11382	TOX007	*Toxodon platensis*	Carcarañá, Santa Fe, Argentina	32°51'; 61°09' *	tibia
MACN-PV 11527	TOX1	*Toxodon platensis*	Carcarañá, Santa Fe, Argentina	32°51'; 61°09' *	petrosal
MACN-PV 14110	TOX009	*Toxodon platensis*	Argentina, old collections	NA	humerus
MACN-PV 17710	TOX004	*Toxodon platensis*	Tapalqué, Buenos Aires, Argentina	36°21'; 60°01' *	tibia
MLP 12-1174	TOX001	*Toxodon platensis*	Arrecifes, Buenos Aires, Argentina	34°04'; 60°07' *	petrosal
MNHN(U) 150	TOX002	*Toxodon platensis*	South bank of Río Negro River, 5ta Sección, Dto Durazno, Uruguay	33°22'; 56°31' *	petrosal
MNHN(U) 379	TOX003	*Toxodon platensis*	Departomento Colonia, Uruguay	34°28'; 57°50' *	humerus
MNHN(U) no number	TOX006	*Toxodon platensis*	Uruguay, old collections	NA	humerus
MMP 5019-M	RAU2	*Macrauchenia patachonica*	Camet Norte, Buenos Aires, Argentina	37°49'; 57°29'	mandible
MNHN-F-TAR 817	RAU1	*Macrauchenia patachonica*	Tarija, Bolivia	21°32'; 64°44'	Metatarsal
UISEK/KM/No number	MAC001	*Macrauchenia patachonica*	Kamac Mayu, Calama, Chile	22°28'; 68°56'	Molar root
UISEK/KM/B2/3	MAC003	*Macrauchenia patachonica*	Kamac Mayu, Calama, Chile	22°28'; 68°56'	Metapodial
UISEK/KM/No number	MAC004	*Macrauchenia patachonica*	Kamac Mayu, Calama, Chile	22°28'; 68°56'	Petrosal
FACSO/BN-1/2A/5	MAC002	*Macrauchenia patachonica*	Baño Nuevo-1 Cave, Coyhaique, Chile	45°17'; 71°32'	Middle phalanx

*As specific localities are rarely indicated in older collections, coordinates are of nearest town

Institutional abbreviations:
FACSO/BN, Facultad de Ciencias Sociales, Universidad de Chile, Baño Nuevo-1 collection, Santiago, Chile
UISEK/KM, Universidad Internacional SEK-Chile, Kamac Mayu collection, Santiago, Chile
MACN-PV, Museo Argentino de Ciencias Naturales, vertebrate paleontology collection, Buenos Aires, Argentina
MNHN(U), Museo Nacional de Historia Natural, Montevideo, Uruguay
MLP, Museo de La Plata, vertebrate paleontology collection, Buenos Aires, Argentina
MNHN-F-TAR, Musee National de Histoire Natural, Tarija collection, Paris, France

Supplementary Table 2. ***Macrauchenia*** **and** ***Toxodon*** **test sequencing and mapping results.**

Given sample code name	Number of raw reads	Total number of read pairs after trimming and adapter removal	Number of combined read pairs	Number of reads uniquely mapped to horse	% of merged, trimmed reads mapping to horse	% of raw read pairs mapping to horse	Number of reads uniquely mapped to rhino	% of merged, trimmed reads mapping to rhino	% of raw read pairs mapping to rhino
MAC001	3105006	1699322	1646515	5896	0.3580896621	0.1898869117	7733	0.4696586426	0.249049438
MAC002	1934987	1560104	1512035	36822	2.44	1.9	41992	2.78	2.17
MAC002 (bleach) *	2303144	1668567		26311	1.576862062	1.142394918	30904	1.852128203	1.341817967
MAC002 (predigest) *	3470116	2609499		37138	1.423185063	1.070223589	43466	1.665683719	1.252580605
MAC003	3002543	1197249	1166376	6235	0.5345617537	0.2076573092	7992	0.6851992839	0.266174372
MAC004	3018104	1362016	1326751	7652	0.576747257	0.2535366575	10067	0.7587708621	0.333553780
TOX008*	2599640	1422642		1961	0.1378421275	0.0754335216	2382	0.1674349555	0.091628071
TOX009*	2609664	1529771		816	0.0533413171	0.0312683932	996	0.0651077841	0.038165832
RAU1	19932093	18360464	7373163	986	0.0053702347	0.0049467961	1066	0.0058059535	0.005348158
RAU2	11191261	10305156	4997463	680	0.0065986386	0.0060761696	717	0.0069576822	0.006406784
TOX1	7023066	6112488	4448756	68	0.0011124766	0.0009682381	92	0.0015051154	0.001309969
TOX2	8058659	7833419	2275358	427	0.0054510042	0.0052986483	428	0.00546377	0.005311057
TOX001	6493167	6106757	2047118	2347	0.0384328376	0.0361456898	2629	0.0430506732	0.040488716
TOX002	6717421	6055758	3905117	179	0.0029558645	0.0026647131	197	0.0032531023	0.002932673
TOX003	6506929	5812021	3959820	27476	0.4727443345	0.4222575657	13804	0.2375077447	0.212143086
TOX004	10062032	9272949	4390415	176	0.0018979938	0.0017491497	187	0.0020166184	0.001858471
TOX005	2910385	2692588	1463995	247	0.0091733306	0.0084868497	255	0.0094704426	0.008761727
TOX006	7729600	7274714	2919959	177	0.0024330853	0.0022898986	85	0.0011684308	0.001099668
TOX007	7900849	7419317	3819959	112	0.0015095729	0.0014175692	112	0.0015095729	0.001417569

* Note: sequenced using single ended reads

Supplementary Table 3. Pairwise distance comparisons between consensus sequences from different bait sequences when using MITObim default parameters.

Bait reference 1	Bait reference 2	Pairwise distance
Horse	Rhino	0.16
Horse	Guanaco	0.43
Rhino	Guanaco	0.42

Supplementary Table 4. Comparisons of number of mismatches between MITObim produced cave hyena mitochondrial sequence and the sequence produced using BWA when using different minimum coverage cutoffs for consensus calling.

Mismatch %	MITObim bait reference	Number of sites of mitogenome covered	Number of mismatches at 1x coverage consensus calling compared to bwa sequence	Number of mismatches 80% of the average coverage compared to bwa sequence
0	Brown bear	1393	25	0
1	Brown bear	6121	51	0
2	Brown bear	6361	53	0
6	Brown bear	7016	24	0
10	Brown bear	7611	24	0
12	Brown bear	6931	24	0
0	Dog	3985	27	0
1	Dog	7223	17	0
2	Dog	7586	18	0
6	Dog	7430	42	0
10	Dog	7571	5	0
12	Dog	7567	9	0

Supplementary Table 5. Percentages of mitochondrial genome covered when using different mismatch values and bait reference sequences.

Bait reference	mismatch value	% of the mitogenome covered
Horse	0	23.2
Horse	1	52.8
Horse	2	49.0
Horse	3	50.5
Horse	4	54.7
Horse	5	58.0
Horse	6	60.4
Tapir	0	23.6
Tapir	1	58.0
Tapir	2	62.4
Tapir	3	61.1
Tapir	4	69.6
Tapir	5	66.6
Tapir	6	67.9
Rhinoceros	0	14.1
Rhinoceros	1	32.6
Rhinoceros	2	36.2
Rhinoceros	3	40.1
Rhinoceros	4	41.2
Rhinoceros	5	39.9
Rhinoceros	6	66.9
Guanaco	0	10.1
Guanaco	1	10.1
Guanaco	2	10.1
Guanaco	3	50.6
Guanaco	4	50.7
Guanaco	5	67.2
Guanaco	6	74.6

Supplementary Table 6. Estimated *Macrauchenia* (Panperissodactyla) divergence dates based on different fossil calibrations.

Calibration node	Mean divergence time (MYA)	95% CI upper limit	95% CI lower limit
Crown Laurasitheria	54.82	75.82	40.22
Crown Carnivore	48.91	61.61	38.99
Crown Bovidae	55.37	72.48	41.76
Crown Perissodactyla	78.82	94.51	63.09
Combination of the above four	66.15	77.83	56.64

Supplementary Table 7. Estimated divergence times for each major clade using the combination of the four calibration points described in Supplementary Table 6.

Clade	Mean (MYA)	Lower 95% CI (MYA)	Upper 95% CI (MYA)
Perissodactyla	48.88	47.8	51.00
Carnivora	54.09	45.16	64.92
Artiodactyla	65.51	54.66	78.08
Panperissodactyla	66.15	56.64	77.83
Chiroptera	75.48	63.46	89.02
Eulipotyphla	78.82	63.06	95.92
Laurasiatheria	89.19	73.88	104.62

Supplementary Table 8. *Macrauchenia* and *Toxodon* samples that received pretreatment with bleach.

Given code name	Sample details
MAC001	Molar root Kamac mayu(Calama) Grid B2/Layer:3 Fondecyt Integracion Nueva Calama
MAC002	2nd phalanx Bano nuevo-1 (Coyhaique) Grid 2A/Layer:5 Fondecyt 1030560
MAC003	Metapodial Kamac mayu(Calama) Grid B2/Layer:3 Fondecyt Integracion Nueva Calama
MAC004	Petrosal Kamac mayu (Calama) Grid: 0 Layer 0 Fondecyt Integracion Nueva Calama
TOX008	MACN 14110
TOX009	MACN 5718

Supplementary Table 9. Total number of MAC002 reads remaining after various control stages.

	Number of reads
Raw paired end reads	68891960
PE reads post-duplicate removal	65268335
Post adapter, low quality and short read trimming	43917870
Post PE merging	42928963

Supplementary Table 10. Species names and Genbank accession numbers of mitochondrial sequences used in the multiple sequence alignment.

Acession code	**Genus**	**species**
GU946995	*Bison*	*bison*
NC004577	*Muntiacus*	*crinifrons*
HQ832482	*Cervus*	*nippon*
JQ608470	*Moschus*	*chrysogaster*
KJ772514	*Mazama*	*gouazoupira*
KM612279	*Odocoileus*	*virginianus*
KF312238	*Ovis*	*orientalis*
KJ681486	*Capreolus*	*capreolus*
KP172593	*Cervus*	*elaphus*
KP662715	*Capra*	*hircus*
KM506758	*Rangifer*	*tarandus*
NC024860	*Sus*	*celebensis*
JX101652	*Pseudois*	*nayaur*
JN869311	*Oryx*	*dammah*
NC014701	*Rucervus*	*eldi*
NC014875	*Procapra*	*przewalskii*
NC012100	*Giraffa*	*camelopardalis*
NC012098	*Antilope*	*cervicapra*
NC009849	*Camelus*	*dromedarius*
EF035447	*Muntiacus*	*reevesi*
KC572860	*Balaenoptera*	*physalus*
NC027237	*Nyctalus*	*noctula*
NC026465	*Cynopterus*	*brachyotis*
NC025949	*Murina*	*leucogaster*
NC024558	*Vespertilio*	*sinensis*
NC018540	*Hipposideros*	*armiger*
NC016872	*Plecotus*	*rafinesquii*
NC016871	*Artibeus*	*lituratus*
JN209842	*Lasiurus*	*borealis*
NC022474	*Eptesicus*	*serotinus*
NC022429	*Vampyrum*	*spectrum*
NC022428	*Tonatia*	*saurophila*
NC022427	*Sturnira*	*tildae*
NC022424	*Lophostoma*	*silvicolum*

NC022422	*Carollia*	*perspicillata*
NC022421	*Brachyphylla*	*cavernarum*
NC022420	*Anoura*	*caudifer*
NC022419	*Micronycteris*	*megalotis*
NC015484	*Plecotus*	*auritus*
HQ685964	*Ursus*	*arctos*
KM488625	*Neovison*	*vison*
KM347744	*Martes*	*flavigula*
NC025296	*Viverricula*	*indica*
NC021751	*Mustela*	*altaica*
NC024568	*Genetta*	*servalina*
NC024567	*Nandinia*	*binotata*
KJ636050	*Prionodon*	*pardicolor*
KF387633	*Vulpes*	*vulpes*
EF672696	*Lutra*	*lutra*
KM236783	*Panthera*	*onca*
KR611313	*Gulo*	*gulo*
NC020648	*Mephitis*	*mephitis*
NC001602	*Halichoerus*	*grypus*
EF551004	*Uncia*	*uncia*
HM106331	*Nasua*	*nasua*
HM106328	*Melogale*	*moschata*
EF196663	*Ailuropoda*	*melanoleuca*
KP202275	*Felis*	*silvestris*
KF926377	*Bos*	*taurus*
FJ905816	*Dicerorhinus*	*sumatrensis*
NC020433	*Equus*	*kiang*
NC020476	*Equus*	*zebra*
NC002008	*Canis*	*lupus*
NC012059	*Tursiops*	*truncatus*
NC011822	*Lama*	*guanicoe*
KM881677	*Hippidion*	*sp.*
NC000889	*Hippopotamus*	*amphibius*
EU939445	*Equus*	*caballus*
KT368758	*Equus*	*przewalskii*
NC020669	*Hyaena*	*hyaena*
JX034737	*Uropsilus*	*sp.*
NC026204	*Crocidura*	*attenuata*
NC026131	*Episoriculus*	*caudatus*
NC025559	*Neomys*	*fodiens*
KC503902	*Nectogale*	*elegans*
NC023244	*Uropsilus*	*soricipes*
NC021398	*Crocidura*	*shantungensis*
NC023950	*Blarinella*	*quadraticauda*
NC002391	*Talpa*	*europaea*
NC005035	*Mogera*	*wogura*
AY833419	*Galemys*	*pyrenaicus*
NC024563	*Anourosorex*	*squamipes*

NC025308	*Myotis*	*brandtii*
KP126954	*Sus*	*scrofa*
Y07726	*Ceratotherium*	*simum*
NC016191	*Rhinolophus*	*ferrumequinum*
NC012682	*Diceros*	*bicornis*
NC001779	*Rhinoceros*	*unicornis*
NC012681	*Coelodonta*	*antiquitatis*
NC027963	*Sorex*	*araneus*
KJ417810	*Tapirus*	*indicus*
KP214033	*Pteropus*	*vampyrus*
FJ456892	*Vicugna*	*vicugna*

Supplementary Table 11. Genes and RNA sequences associated with each partition used in the Raxml analysis.

partition number	tRNA and gene in partition
1	tRNA-Ile, tRNA-Leu, tRNA-Met2, tRNA-Pro, tRNA-Ser2
2	ATP8, ND2
3	tRNA-Asp, tRNA-Gly, tRNA-His, tRNA-Leu2, tRNA-Lys, tRNA-Ser, tRNA-Trp, tRNA-Tyr, tRNA-Val
4	12S, tRNA-Ala, tRNA-Arg, tRNA-Phe, tRNA-Thr
5	16S, tRNA-Asn, tRNA-Met1
6	TRNA-Cys
7	COX1
8	ATP6, COX2, COX3, ND3, ND4L, tRNA-Gln
9	ND4, ND5, ND6, tRNA-Glu
10	CYTB, ND1

Supplementary Table 12. Genes and RNA sequences associated with each partition along with the substitution model associated with each partition for the BEAST analysis.

partition number	tRNA and gene in partition	Substitution models
1	COX1, tRNA-Met2	GTR I+G
2	ND2, tRNA-Glu	GTR I+G
3	12S, tRNA-Ala, tRNA-Gly, tRNA-His, tRNA-Leu2, tRNA-Lys, tRNA-Phe, tRNA-Ser, tRNA-Thr, tRNA-Trp, tRNA-Tyr, tRNA-Val	GTR I+G
4	16S, tRNA-Asn, tRNA-Met1	GTR I+G
5	tRNA-Cys	SYM+G
6	tRNA-Arg, tRNA-Asp, tRNA-Ile, tRNA-Leu, tRNA-Ser2	GTR I+G
7	ATP6, COX2, COX3, ND3, ND4L, tRNA-Gln, tRNA-Pro	GTR I+G
8	ATP8, ND4, ND5, ND6	GTR I+G
9	CYTB, ND1	GTR I+G

Supplementary Table 13. Species considered as ingroup for each fossil calibration analysis.

Ingroup species when using crown Bovidae fossil calibration	Ingroup species when using crown Laurasiatheria fossil calibration	Ingroup species when using crown Carnivora fossil calibration	Ingroup species when using crown Perissodactyla fossil calibration
Antilope cervicapra	All species in the alignment	*Ailuropoda melanoleuca*	*Equus kiang*
Bison bison		*Canis lupus*	*Equus zebra*
Bos taurus		*Felis silvestris*	*Equus cabullus*
Capra hircus		*Genetta servalina*	*Equus przewalskii*
Oryx dammah		*Gulo gulo*	*Tapirus indicus*
Ovis orientalis		*Halichoerus grypus*	*Coelodonta antiquitatis*
Procapra przewalskii		*Hyaena hyaena*	*Dicerorhinus sumatrensis*
Pseudois nayaur		*Lutra lutra*	*Rhinoceros unicornis*
		Martes flavigula	*Ceratotherium simum*
		Melogale moschata	*Diceros bicornis*
		Mephitis mephitis	*Hippidion sp.*
		Mustela altaica	
		Nandinia binotata	
		Nasua nasua	
		Neovison vison	
		Panthera onca	
		Prionodon pardicolor	
		Uncia uncia	
		Ursus arctos	
		Viverricula indica	
		Vulpes vulpes	

Supplementary Note 1: Systematic Context of *Macrauchenia patachonica*

Overview of litoptern systematics. Litopterns were cursorial herbivores with mesaxonic limbs and several unusual cranial and dental features. In distinction to most other SANUs, litopterns primitively retained rooted teeth and developed varied cheektooth morphologies, including bunodonty (*e.g.,* Megadolodinae), bunolophodonty (*e.g.,* Proterotheriinae) and lophoselenodonty (*e.g.,* Macraucheniidae)[1,2,3]. As the result of continuing discovery over many decades, we now know that litopterns were present in South America for most of the Cenozoic. One group (Sparnotheriodontidae) even reached West Antarctica when this was connected to southernmost South America prior to the appearance of the Drake Passage[4,5.6].

Unquestionable litopterns are traditionally classified in three families, Adianthidae, Proterotheriidae, and Macraucheniidae. The latter two were taxically dominant, with macraucheniids persisting from the Late Eocene (Mustersan South American Land Mammal Age or SALMA) through to the end of the Pleistocene (Lujanian, local Stage/Age)[7,8]. Other, mostly Paleogene groups have a less certain connection with these core litoptern families. Anisolambdidae, Notonychopidae, Indalecidae, Protolipternidae, Amilnedwardsiidae, Sparnotheriodontidae and even archaic ungulates such as Didolodontidae have traditionally been included within the order (*e.g.*, refs.[1,9-11]), but their positions have been repeatedly questioned (*e.g.*, refs.[12-15]). Among these putative litopterns the oldest known is *Requisia vidmari*, from the Early Paleocene locality of Punta Peligro in central Patagonia[10,16], currently dated to 63.2 – 63.8 Ma[17]. Isolated astragali from the Early Eocene locality of Itaboraí (eastern Brazil) resemble those of definite proterotheriids, leading some authors to speculate that this clade was already in existence by this time (*e.g.*, refs.[14,18-20]). However, proterotheriid dentitions are first recorded in the Late Oligocene, during the Deseadan SALMA[3].

Late Eocene (Mustersan SALMA) *Polymorphis*, the oldest known definite macraucheniid[14], serves to date the origin of the family paleontologically. Differing from the horse-like proterotheriids, macraucheniids had a robust body structure more like that of modern camels, featuring three-toed autopodia, a long neck, and reduced nasals. In Late Miocene–Pleistocene macraucheniines, some of these features evolved in bizarre directions, as in the case of the retraction of the nasal aperture[21]. From a position at the end of the rostrum, as in typical mammals, this opening progressively moved dorsally. Pleistocene *Macrauchenia*, in which the nasal aperture is perched between the orbits, near the summit of the skull, represents a morphological extreme. It has long been speculated whether so drastic a modification implies that the snout was elaborated into a proboscis[8,22-25].

Macraucheniidae includes either two or three subfamilies, depending on whether Theosodontinae is separately recognized. Macraucheniinae and Cramaucheniinae contain most of the approximately 18 genus-level taxa currently recognized for the family (ignoring possible synonyms). Progressive increases in body size, cheektooth crown height, and snout length strongly marked the evolution of Macraucheniinae[1,21, 23,24,26-28]. From the standpoint of taxonomic richness,

macraucheniine peak diversity was achieved in the Late Miocene. They became extinct early in the Holocene, *Macrauchenia* itself being the last surviving taxon (last appearance date: 8390 ± 140 14C yr BP[8,29]). *Macrauchenia* (and its close relatives *Macraucheniopsis* and *Xenorhinotherium*) had a broad distribution across the continent, although as in the case of other South American native ungulates known fossils overwhelmingly come from the southern cone (Supplementary fig. 7, 8). Unlike toxodontid notoungulates, which managed to penetrate Central and southern North America after the completion of the Isthmus of Panama, *Macrauchenia* and other litopterns failed to participate in the Great American Biotic Interchange.

History of classification and phylogenetic relationships. The relationships of litopterns and other SANUs have been controversial since the first specimens were described 180 years ago[30,31]. Lydekker[32] included litopterns within "Ungulata", a wastebasket comprising fossil and extant ungulates generally, and in this he was followed by Osborn[33], Scott[1], and Schlosser[34]. Although Scott[1] presciently remarked on numerous resemblances between litopterns and perissodactyls, in the end he concluded that they were more likely the result of parallelism than actual close relationship. Ameghino[35], by contrast, claimed that it was "absolutely certain" that macraucheniids and other litopterns had a common origin with perissodactyls, distinct from the ancestry of other SANUs.

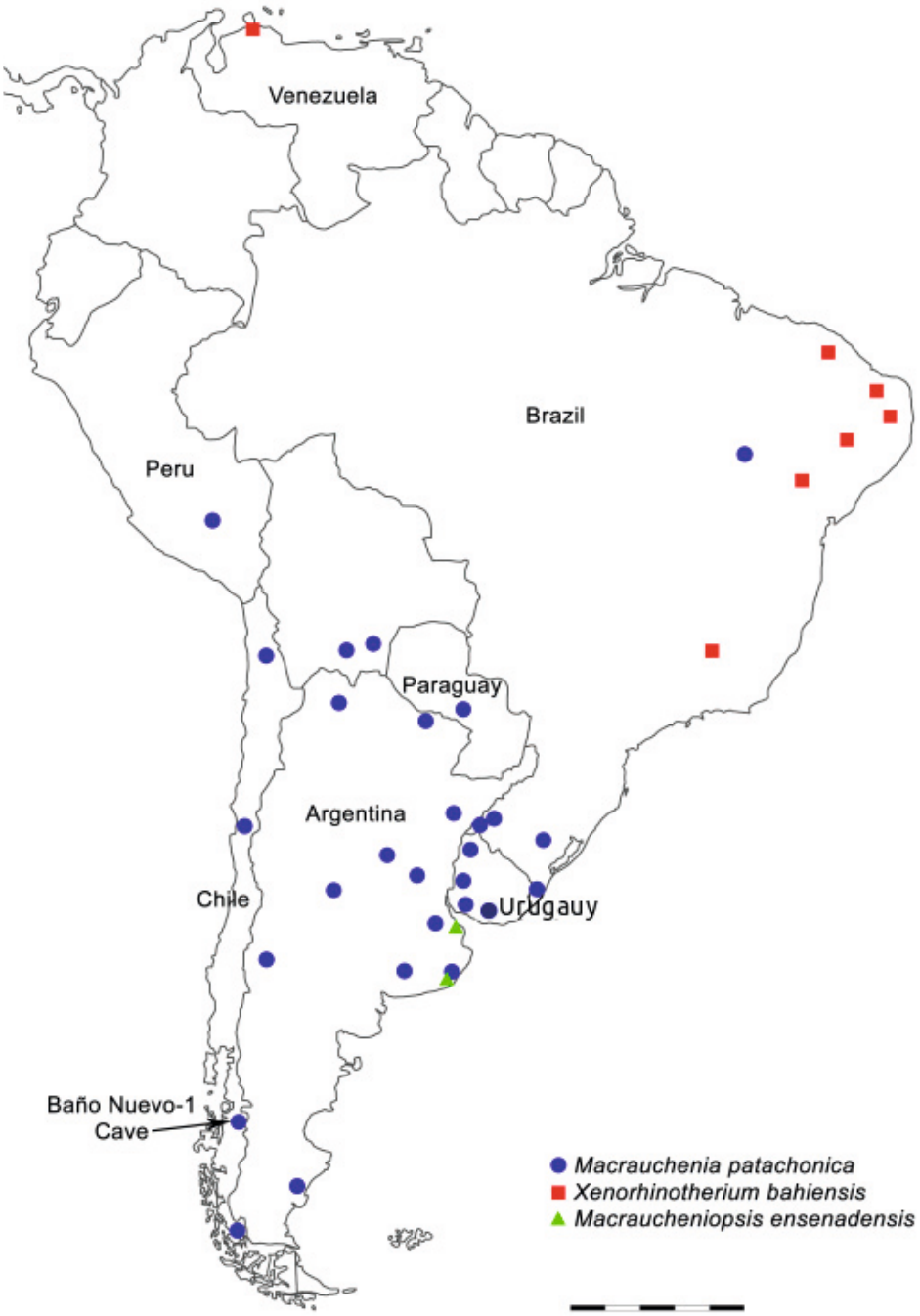
Simpson[36] viewed litopterns as directly derived from Condylarthra, a heterogeneous group of early Cenozoic mammals structurally and presumably phyletically ancestral to later, more advanced ungulates, and grouped them with Notoungulata, Astrapotheria, and Tubulidentata in a paraphyletic supraordinal entity, Protungulata. Subsequently, Simpson[2,37] suggested that all SANUs (including Pyrotheria and Xenungulata) evolved in South America from a North American condylarth ancestral stock, which arrived in South America during the Late Cretaceous or Early Paleocene. This scenario has been accepted by many later authors (*e.g.*, refs.[14, 18,19, 38, 39]).

McKenna[40,41] refined Simpson's proposal, contending that all SANUs (but not Tubulidentata) could be derived from a single North American condylarthran common ancestor. He proposed the name Meridiungulata for this putative monophyletic ensemble, but did not present an explicit cladistic analysis in support of it (see also ref. [42]).
Soria[3,43] also accepted a condylarthran origin for SANUs, but differed from McKenna in proposing a diphyletic origin for the latter. He argued that litopterns were explicitly related to certain North and South American condylarthrans and could thus be included within Protungulata (with content differing from Simpson's[36] concept). The remaining SANU orders were excluded from this grouping. A similar evolutionary scenario was proposed by Muizon and Cifelli[11], who coined the term Panameriungulata for a group including litopterns, South American didolodontids, and Kollpaniinae, an endemic subfamily of North American mioclaenids. The monophyly of Meridiungulata sensu McKenna[40,41] was also rejected by Tong and Lucas[44], Lucas[45], and Kondrashov and Lucas[46]. Horovitz[47] likewise found that Meridiungulata was paraphyletic, based on a cladistic study of postcranial elements. She concluded that litopterns and notoungulates were sister taxa, allied with meniscotheriid and phenacodontid condylarthrans, but separate from astrapotheres. In a similar vein, the preferred tree of placental cladistic relationships published by O´Leary *et al.*[48] positioned Litopterna (represented by *Protolipterna*) within stem Pan-Euungulata

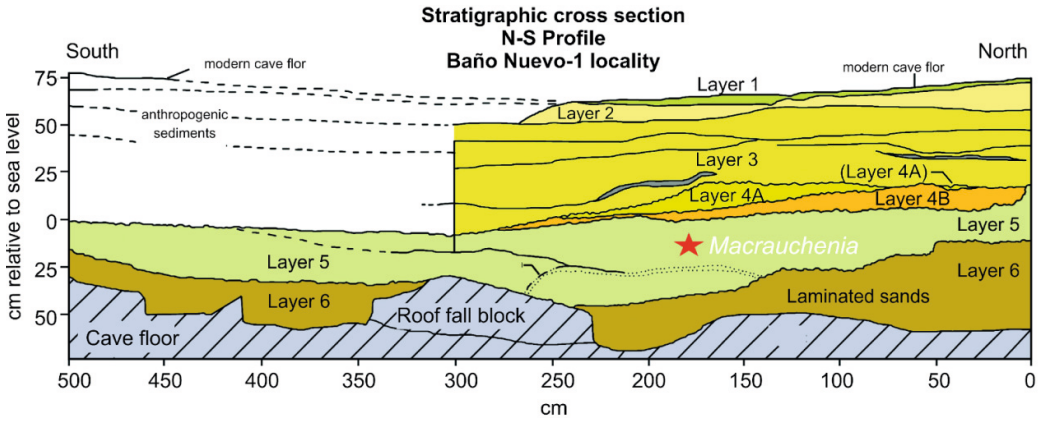
while Notoungulata (represented by *Thomashuxleya*) grouped with Afrotheria, thereby denying once again the monophyly of Meridiungulata.

Muizon *et al*.[49] recently took up the question again, ruling in favor of meridiungulate monophyly but favoring close relationship with Artiodactyla. However, inasmuch as their data matrix did not include any perissodactyls, their analysis does not contradict the results of recent proteomic studies[50,51], which conclude that both litopterns (represented by *Macrauchenia*) and notoungulates (represented by *Toxodon*) are more closely related to Perissodactyla than to any other extant placental group (see also ref. [52]). In the absence of consensus in phenomic analyses, and lack of molecular information for other SANU orders, it remains unsettled whether all South American native ungulates are related monophyletically.

Supplementary Figure 7. Geographic distribution of Pleistocene *Macrauchenia patachonica* and closely related taxa *Xenorhinotherium bahiensis* and *Macraucheniopsis ensenadensis* (based on ref. 53; see also refs 54-61). The arrow indicates the locality of Cueva Baño Nuevo-1, the source of the sample of *Macrauchenia patachonica* (FACSO/BN-1/2A/5) utilized for this study (see Supplementary Figure 8).



Supplementary Figure 8. Stratigraphic cross-section of Baño Nuevo-1 Cave, situated *ca.* 80 km NE of Coyhaique (45° 17′ S; 71° 32′ W), Región XI, Chile. The site is located within the Cerro Grande del Campo Seis, an Aptian (E. Cretaceous) volcanic complex. The cave has a depth of 20 m and an average width of 4 m. A middle phalanx of *Macrauchenia* (red star), designated in this study as MAC002 (and originally catalogued as fondecyt 1030560) was recovered from Layer 5 (clay and organic sands). This specimen yielded a date of 11,115 ± 30 ^{14}C yr BP (UCIAMS 166314), which is consistent with its stratigraphic position and association with faunal elements typical of late Pleistocene Patagonian faunas, including representatives of Ursidae, Equidae, Felidae, Camelidae, and Mylodontidae. *Macrauchenia* specimens were also recovered from overlaying Layer 4B[62].



Supplementary References

1. Scott, W.B. Mammalia of the Santa Cruz beds. Part I. Litopterna. *Reports of the Princeton University Expedition to Patagonia* **7**, 1–156 (1910).
2. Simpson, G.G. *Splendid Isolation* (Yale Univ. Press, New Haven, 1980).
3. Soria, M.F. Los Proterotheriidae (Mammalia, Litopterna): sistemática, origen y filogenia. *Monogr. Mus. Argent. Cienc. Nat. "Bernardino Rivadavia"* **1**, 1–167 (2001).
4. Bond, M., Reguero, M.A., Vizcaíno, S.F. & Marenssi, S.A. in *Cretaceous-Tertiary high-latitude palaeoenvironments, James Ross Basin, Antarctica* (eds Francis J.E., Pirrie D., & Crame J.A.) 163–176. (Geological Society of London, 2006).
5. Reguero, M.A., Gelfo, J.N., López, G.M., Bond, M., Abello, A., Santillana, S.N. & Marenssi, S.A. Final Gondwana breakup: the Paleogene South American native ungulates and the demise of the South America–Antarctica land connection. *Global Planet. Change* **123**, 400–413 (2014).
6. Gelfo, J.N., Mörs T., Lorente, M., López, G.M. & Reguero, M. The oldest mammals from Antarctica, Early Eocene of the La Meseta Formation, Seymour Island. *Palaeontology* **58**, 101–110 (2015).
7. Cifelli, R.L. & Soria, M.F. Notes on Deseadan Macraucheniidae. *Ameghiniana* **20**, 141–153 (1983).
8. Bond, M. Quaternary native ungulates of Southern South America. A synthesis. *Quatern. South Amer. Antarc. Pen.* **12**, 177–205 (1999).
9. Scott, W.B. *A History of Land Mammals in the Western Hemisphere* (Macmillan, 1913).
10. Bonaparte, J.F. & Morales, J. Un primitivo Notonychopidae (Litopterna) del Paleoceno Inferior de Punta Peligro, Chubut Argentina. *Estud. Geol.* **53**, 263–274 (1997).
11. De Muizon, C. & Cifelli, R.L. The "condylarths" (archaic Ungulata, Mammalia) from the Early Palaeocene of Tiupampa (Bolivia): implications on the origin of the South American ungulates. *Geodiversitas* **22**, 47–150 (2000).
12. Soria, M.F. Notopterna: un nuevo orden de mamíferos ungulados eógenos de América del Sur. Parte I. Los Amilnedwardsidae. *Ameghiniana* **25**, 245–258 (1989).
13. Soria, M.F. Notopterna: un nuevo orden de mamíferos ungulados eógenos de América del Sur. Parte II. *Notonychops powelli* gen. et sp. nov. (Notonychopidae nov.) de la Formación Río Loro (Paleoceno medio, Provincia de Tucumán, Argentina). *Ameghiniana* **25**, 259–272 (1989).
14. Cifelli, R.L. in *Mammal phylogeny* (eds F.S. Szalay F.S., M.J. Novacek, N.J. & McKenna, M.C.) **2**, 195–216 (Springer, 1993).
15. Billet, G., Muizon, C., Schellhorn, R., Ruf, I., Ladevèze, S. & Bergqvist, L. Petrosal and inner ear anatomy and allometry amongst specimens referred to Litopterna (Placentalia). *Zool. Jour. Linn. Soc.* **173**, 956–987 (2015).
16. Gelfo, J.N., Ortiz-Jaureguizar, E. & Rougier, G.W. New remains and species of the 'condylarth' genus *Escribania* (Mammalia: Didolodontidae) from the Palaeocene of Patagonia, Argentina. *Trans. R. Soc. Edinburgh* **98**, 127–138 (2007).

17. Woodburne, M.O., Goin, F.J., Raigemborn, M.S., Heizler, M., Gelfo, J. N., Oliveira, E. V. Revised timing of the South American Early Paleogene land mammal ages. *J. South Am. Earth Sci.* **54**, 109–119 (2014).

18. Paula Couto, C. Fossil mammals from the beginning of the Cenozoic in Brazil. Condylarthra, Litopterna, Xenungulata and Astrapotheria. *Bull. Amer. Mus. Nat. Hist.* **99**, 359–394 (1952).

19. Cifelli, R.L. The origin and affinities of the South American Condylarthra and Early Tertiary Litopterna (Mammalia). *Amer. Mus. Novitates* **2772**, 1–49 (1983).

20. Bergqvist, L.P. Deciduous premolars of Paleocene litopterns of Saõ José de Itaboraí Basin, Rio de Janeiro, Brazil. *Jour. Paleont.* **84**, 858–867 (2010).

21. Forasiepi, A.M., MacPhee, R.D.E., Hernández Del Pino, S., Schmidt, G.I., Amson, E. & Grohé, C. Exceptional skull of *Huayqueriana* (Mammalia, Litopterna, Macraucheniidae) from the Late Miocene of Argentina: anatomy, systematics, and paleobiological implications. *Bull. Amer. Mus. Nat. Hist.* **404**, 1–76 (2016).

22. Burmeister, G. Beschreibung der *Macrauchenia patachonica* Owen (*Opisthorhinus falkoneri* Brav.) nach A. Bravard´s Zeichnungen und den im Museo zu Buenos Aires vorhandenen Resten entworfen. *Abh. Naturforsch. Ges. Halle* **1**, 75–112 (1864).

23. Scott, W.B. *A History of Land Mammals in the Western Hemisphere* (MacMillan, 1937).

24. Rusconi, C. Evolución de la trompa en las macrauquenias. *Rev. Mus. His. Nat. Mendoza* **10**, 111–118 (1957).

25. Soria, M.F. Los Litopterna del Colhuehuapense (Oligoceno Tardío) de la Argentina. *Rev. Mus. Argent. Cienc. Nat. "Bernardino Rivadavia"* **3**, 1–54 (1981).

26. Ameghino, F. Apuntes preliminares sobre el género *Theosodon*. *Rev. Jard. Zool. Buenos Aires* **1**, 20–29 (1893).

27. Soria, M.F. in *Actas de IV Congreso Argentino de Paleontología y Bioestratigrafía*, 157–164 (1986).

28. Schmidt, G.I. & Ferrero, B.S. 2014. Taxonomic reinterpretation of *Theosodon hystatus* Cabrera and Kraglievich, 1931 (Litopterna, Macraucheniidae) and phylogenetic relationships of the family. *Jour. Vert. Paleon.* **34**, 1231–1238 (2014).

29. Tonni E.P. 1990. Mamíferos del Holoceno en la Provincia de Buenos Aires. *Paulacoutiana* **4**, 3–21.

30. Owen, R. A description of the cranium of *Toxodon platensis*, a gigantic extinct mammiferous species, referable by its dentition to the Rodentia, but with affinities to the Pachydermata and the herbivorous Cetacea. *Proc. Geol. Soc. London* **2**, 541–542 (1837).

31. Owen, R. Fossil Mammalia. In *The Zoology of the voyage of H.M.S. Beagle, Under the Command of Captain Fitzroy, during the Years 1832 to 1836* (ed. Darwin, C), **1**(1), 1–40 (Smith Elder, 1838).

32. Lydekker, R. Contributions to knowledge of the fossil vertebrates of Argentina, 3. A study of the extinct ungulates of Argentina. *An. Mus. La Plata* **2**, 1–91 (1893).

32. Osborn, H.F. *The Age of Mammals in Europe, Asia, and North America* (Macmillan, 1910).

34. Schlosser, M. in *Grundzüge der Paläontologie (Paläozoologie). II. Abteilung: Vertebrata* (ed. Zittel, K.A.) 402–689 (Oldenbourg, 1923).

35. Ameghino, F. Les formations sédimentaires du Crétacé Supérieur et du Tertiaire de Patagonie avec un parallèle entre leurs faunes mammalogiques et celles de l'ancien continent. *An. Mus. Nac. Buenos Aires* **8**, 1–568 (1906).

36. Simpson, G.G. The principles of classification and a classification of mammals. *Bull. Amer. Mus. Nat. Hist.* **85**, 1–350 (1945).

37. Simpson, G.G. The beginning of the age of mammals in South America. Part 1. Introduction. Systematics: Marsupialia, Edentata, Condylarthra, Litopterna and Notioprogonia. *Bull. Amer. Mus. Nat. Hist.* **91**, 1–232 (1948).

38. Reig, O.A. Teoría del origen y desarrollo de la fauna de mamíferos de América del Sur. *Publ. Mus. Mun. Cien. Nat. "Lorenzo Scaglia"* 1-162 (1981).

39. Cifelli, R.L. in *The Great American Biotic Interchange* (eds Stehli, F.G. & Webb, S.D.) 249–266 (Plenum, 1985).

40. McKenna M. C. in *Phylogeny of the Primates* (eds Luckett, W.P. & Szalay, F.S.) 21–46 (Plenum, 1975). Press, New York.

41. McKenna M.C. in *Evolutionary Biology of the New World Monkeys and Continental Drift* (eds Ciochon, R.L. & Chiarelli, A.B.) 43–77 (Plenum, 1981).

42. McKenna, M.C. & Bell, S.K. *Classification of Mammals above the Species Level* (Columbia Univ. Press, 1997).

43. Soria, M.F. Estudios sobre los *Astrapotheria* (Mammalia) del Paleoceno y Eoceno. Parte II: Filogenia, origen y relaciones. *Ameghiniana* **25**, 47-59 (1988).

44. Tong, Y. & Lucas, S.G. in *Proc. Third North Amer. Paleon. Conv.* **2**, 551–556 (1982).

45. Lucas, S..in *Mammal phylogeny* (eds F.S. Szalay F.S., M.J. Novacek, N.J. & McKenna, M.C.) **2**, 182–194 (Springer, 1993).

46. Kondrashov, P.E. & Lucas, S.G. *Palaeostylops iturus* from the Upper Paleocene of Mongolia and the status of Arctostylopida (Mammalia, Eutheria). *Bull. New Mexico Mus. Nat. Hist. Sci.* **26**, 195–203 (2004).

47. Horovitz, I. Eutherian mammal systematics and the origins of South American ungulates as based on postcranial osteology. *Bull. Carnegie Mus. Nat. Hist.* **36**, 63 –79 (2004).

48. O'Leary, M.A. *et al.* The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* **33**, 662–667 (2013).

49. De Muizon, C., Billet, G. Argot, C. Ladevèze, S. & Goussard, F. *Alcidedorbignya inopinata*, a basal pantodont (Placentalia, Mammalia) from the early Palaeocene of Bolivia: anatomy, phylogeny and palaeobiology. *Geodiversitas* **37**, 397–634 (2015).

50. Welker, F. *et al.* Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature* **522**, 81–84 (2015).

51. Buckley, M. Ancient collagen reveals evolutionary history of the endemic South American "ungulates". *Proc. R. Soc.* **B282**, 20142671 (2015).

52. Beck, R.M.D. & Lee, M.S.Y. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. *Proc. R. Soc.* **281**, 20141278 (2014).

53. Scherer, C.S., Pitana V.G. & Ribeiro A.M. Proterotheriidae and Macraucheniidae (Litopterna, Mammalia) from the Pleistocene of Rio Grande Do Sul State, Brazil. *Rev. Brasil. Paleon.* **12**, 231–246 (2009).

54. Politis, G., Prado J.L., & Beukens, R. in *Ancient peoples and landscapes* (ed. Johnson, E.) 187–205 (Texas Tech Univ., 1995).

55. Panarello, H.O. & Fernández, J. Palaeoenvironmental changes in Leuto Caballo (Neuquén, Argentina) during Late Pleistocene - Holocene, evidenced by stable isotopes on marl and *Lymnaea*: first results. *An. Direc. Nac. Serv. Geol.* **34**, 418–421 (1999).

56. Ubilla, M. & Perea, D. Quaternary vertebrates of Uruguay: A biostratigraphic, biogeographic and climatic overview. *Quat. South Amer. Antarc. Pen.* **12**, 75–90 (1999).

57. Velásquez, H. & Mena, F. Distribuciones óseas de ungulados en la cueva Baño Nuevo 1 (XI Región, Chile): un primer acercamiento. *Magallania* **34**, 91–105 (2006).

58. López, P. & Labarca, R.O. *Macrauchenia* (Litopterna), *Hippidion* (Perissodactyla), Camelidae y Edentata en Calama (II Región): comentarios taxonómicos y tafonómicos. *Not. Mens. Mus. Nac. Hist. Nat.* **355**, 7–10 (2005).

59. Labarca, R.O. El Yacimiento paleontológico "Kamac Mayu": tafonomía y procesos de formación en el Cuaternario kárstico de la Cuenca de Calama (Región de Antofagasta-Chile). *Ameghiniana* **46**, 3–16 (2009).

60. Borrero, L.A. in *American Megafaunal Extinctions at the End of the Pleistocene* (ed. Haynes, G.)145–168 (Springer, 2009).

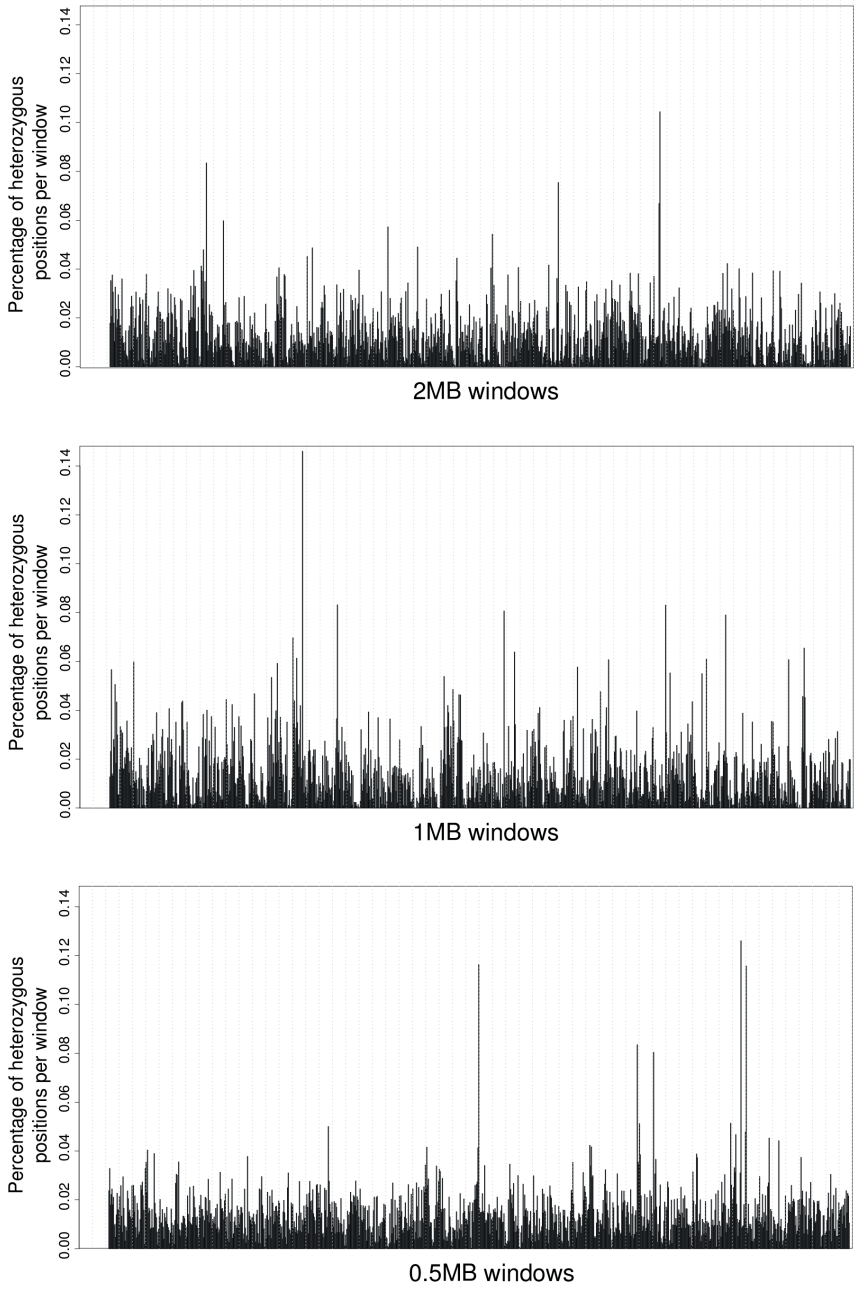
61. Tassara, D.A. and Cenizo, M.M. El patrimonio paleontológico en el sector costero al NE de Mar del Plata (Provincia de Buenos Aires, Argentina): Estado del conocimiento, vulnerabilidad y propuestas para su conservación. *Rev. Mus. Argent. Cienc. Nat. "Bernardino Rivadavia"* **16**, 165–183 (2014).

62. López, P., Mena, F. & Bostelmann, E. Presence of extinct bear in a pre-cultural level of Baño Nuevo-1 cave (Central Patagonia, Chile). *Estud. Geol.* **71**, e041 (2015).

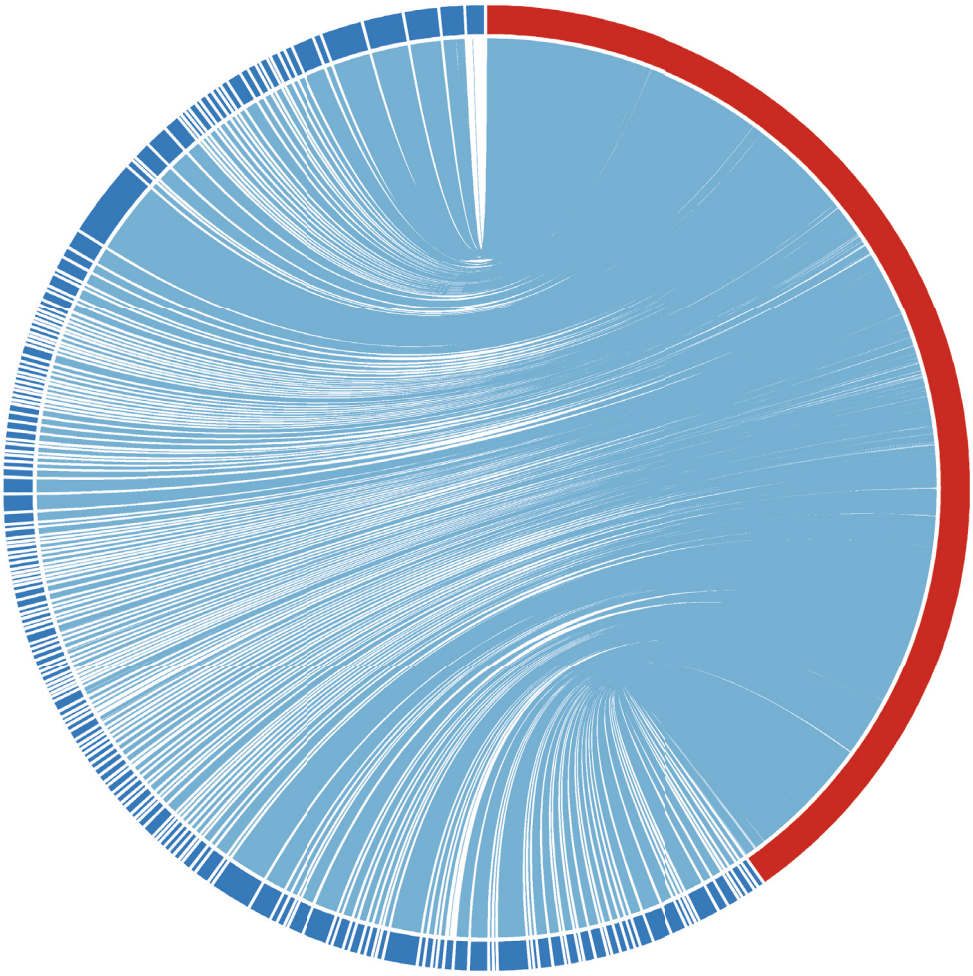
Appendix B.

Supporting information for Chapter 5: Article III, Westbury et al., 2017 **Population and conservation genomics of the world's rarest hyena species, the brown hyena (*Parahyena brunnea*)**

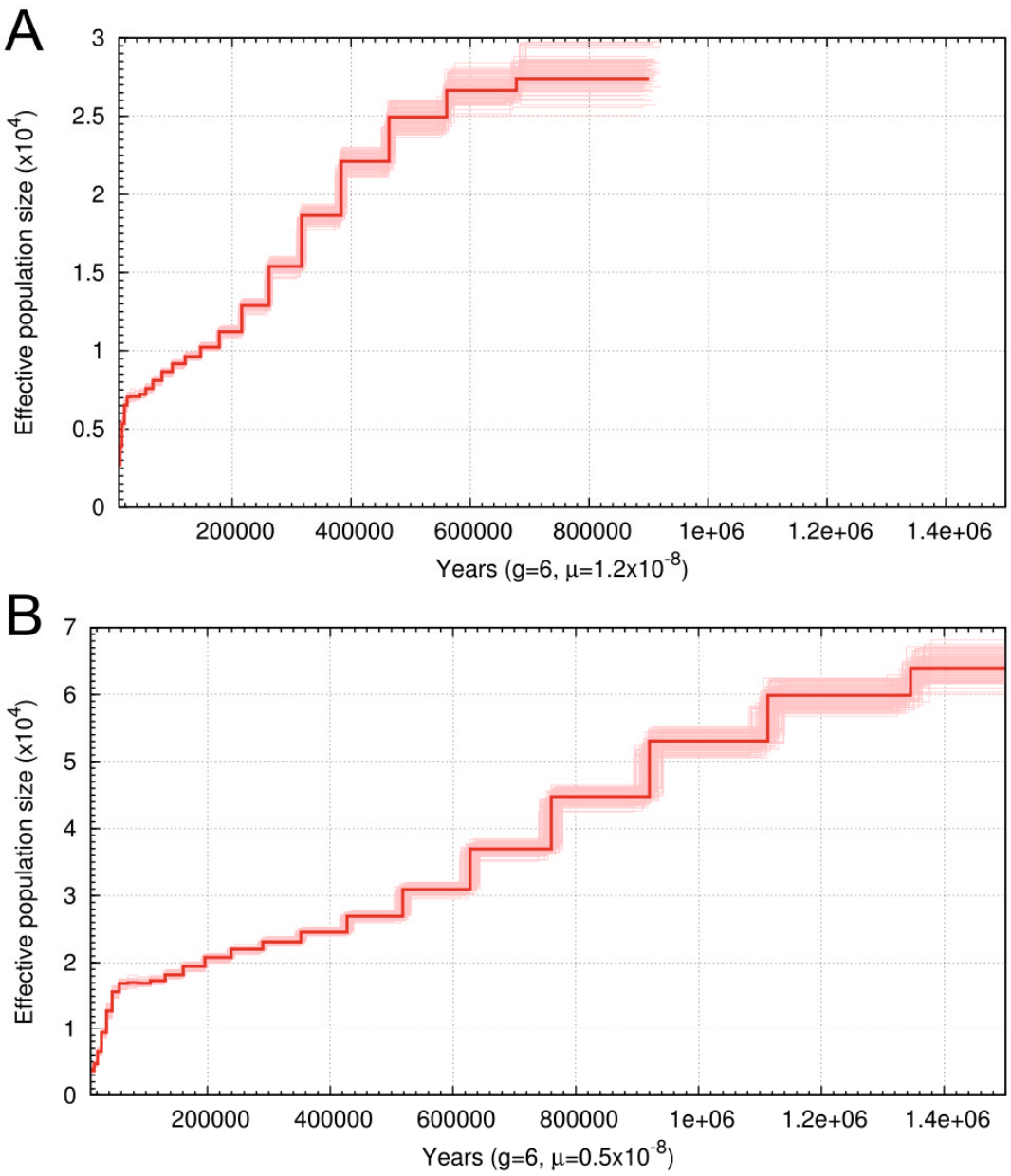
Supplemental figures



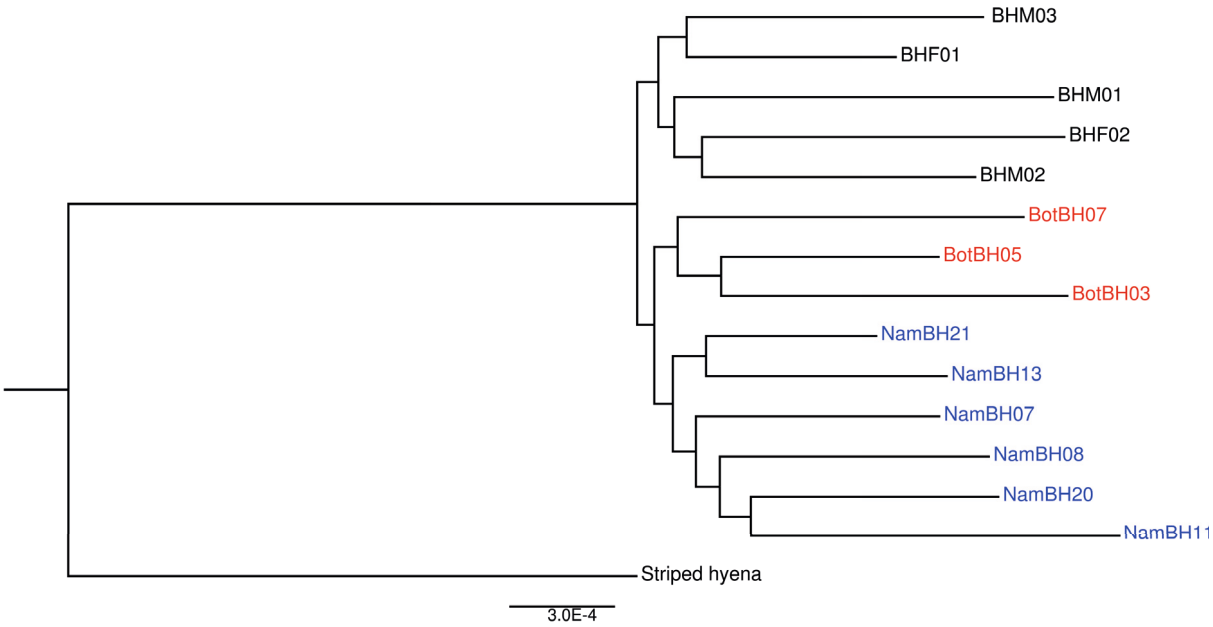
Supplemental figure S1. Non-overlapping sliding window analysis of heterozygosity across the captive brown hyena nuclear genome using window sizes of 2Mbp, 1Mbp and 0.5Mbp. Each graph consists of 2000 windows. The Y axis indicates the percentage of the window made up of heterozygous positions. The X axis indicates the window. No considerable stretches of homozygosity can be seen suggesting a lack of inbreeding in this individual.



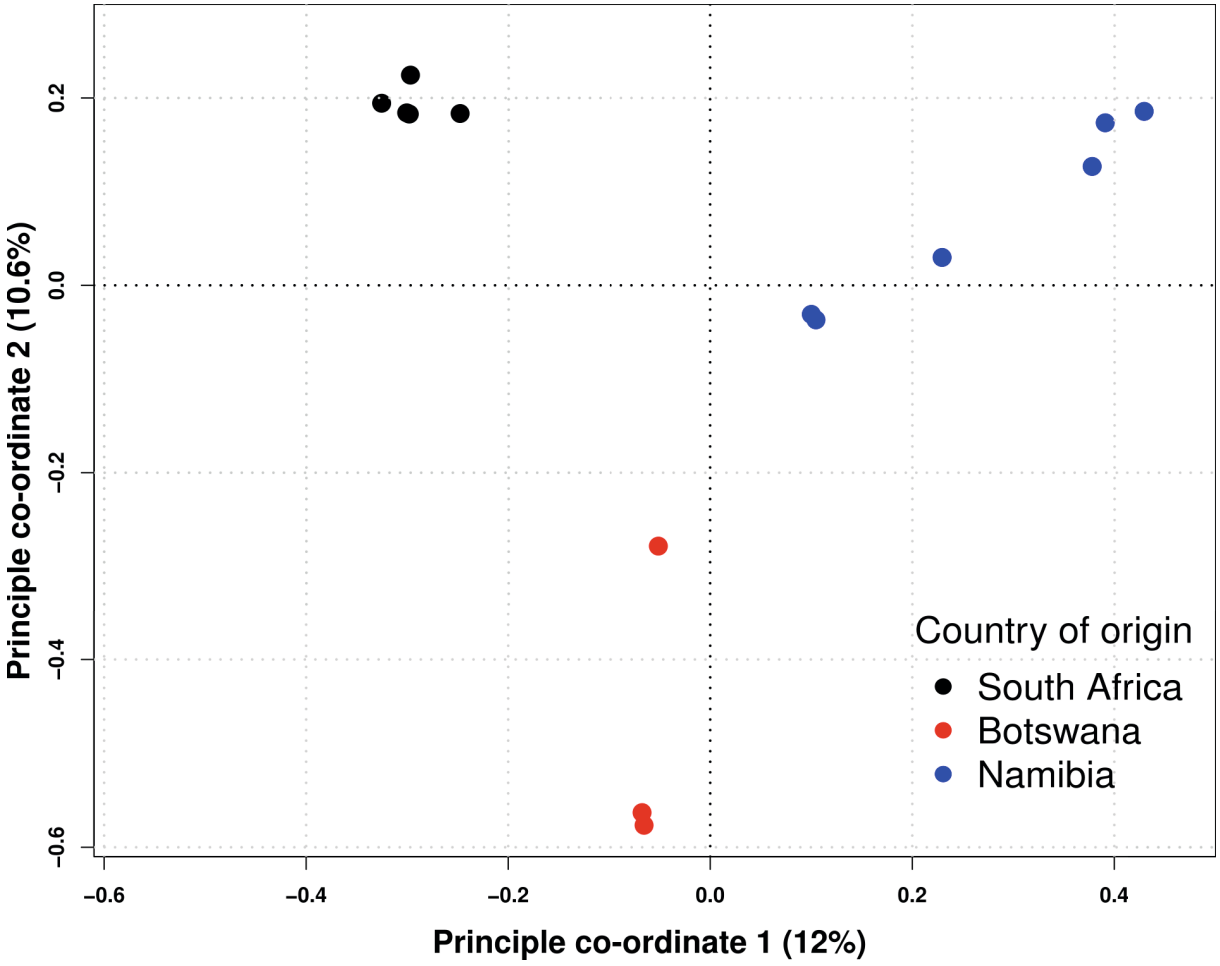
Supplemental figure S2. Circos plot of the cat X chromosome (red) and the corresponding scaffolds in the brown hyena (blue). Most of the cat X chromosome is covered by the brown hyena.



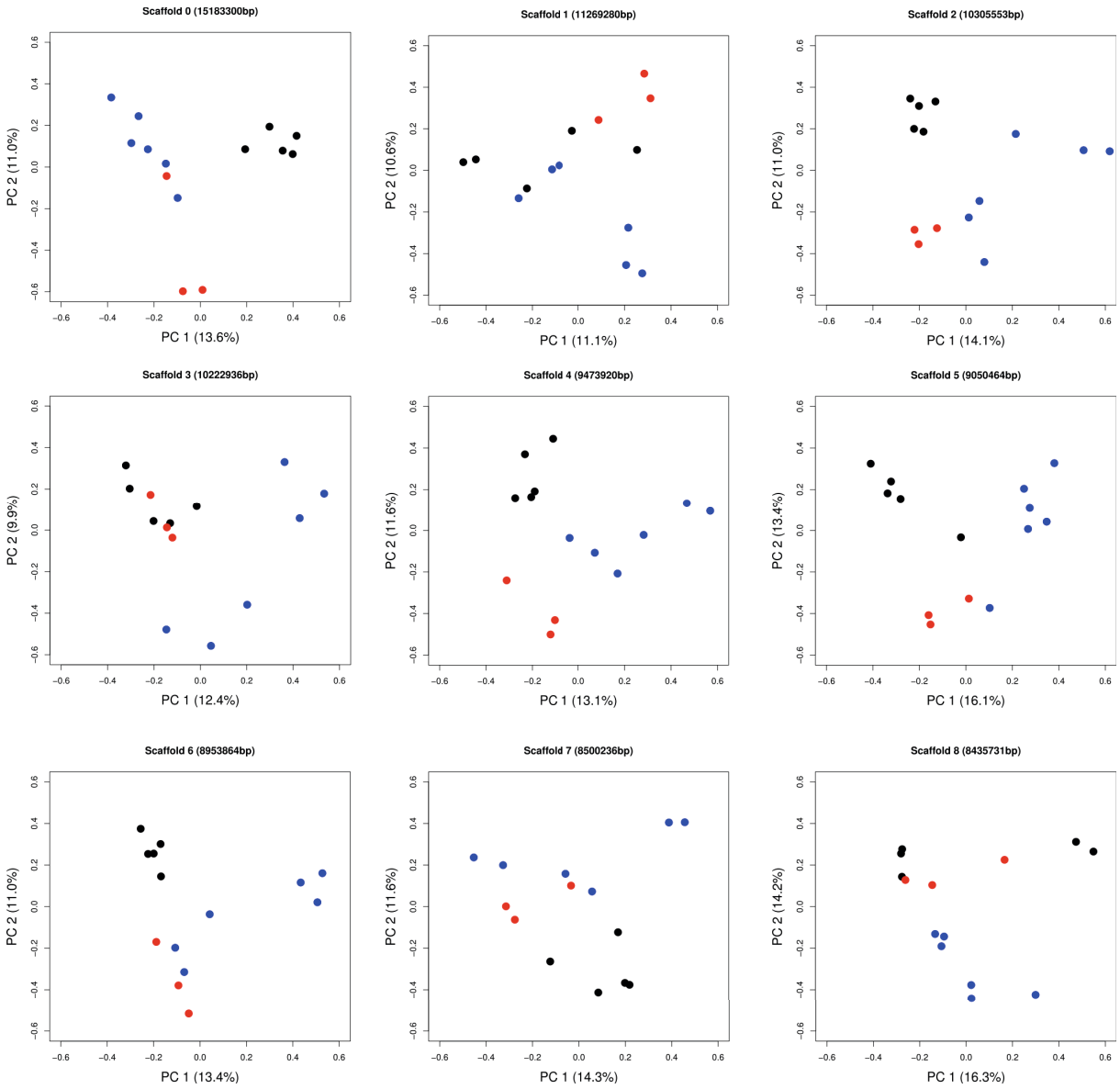
Supplemental figure S3. PSMC analyses utilising different per generation mutation rates calculated from the 95% confidence interval of the brown and striped hyena divergence date. A) PSMC plot using a per generation mutation rate calculated assuming a brown and striped hyena divergence date of 2.6mya. B) PSMC plot using a per generation mutation rate calculated assuming a brown and striped hyena divergence date of 6.4mya. g shows generation time and μ shows mutation rate per generation.



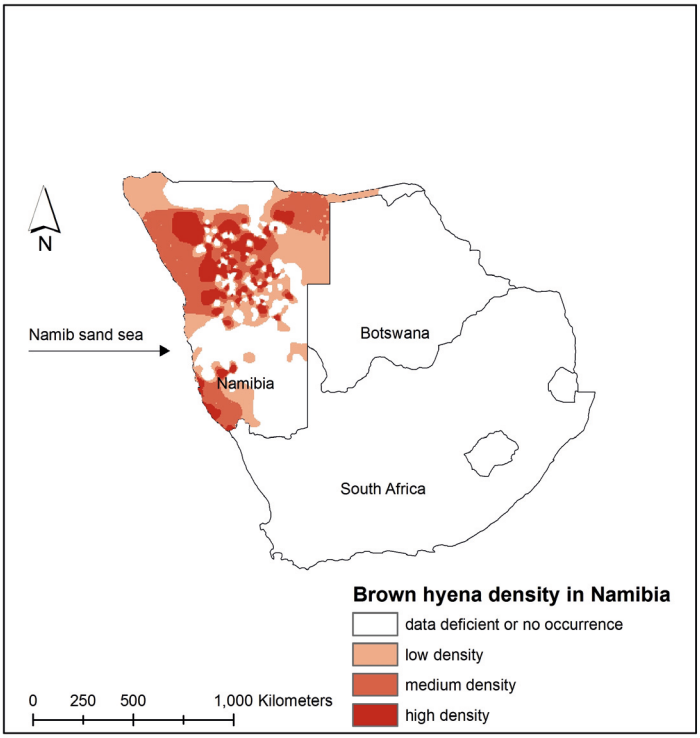
Supplemental figure S4. Wild caught brown hyena maximum likelihood tree rooted using the Striped hyena. Different colours represent the country of origin of the samples (blue - Namibia, red - Botswana, black - South Africa). Scale bar indicates substitutions per site.



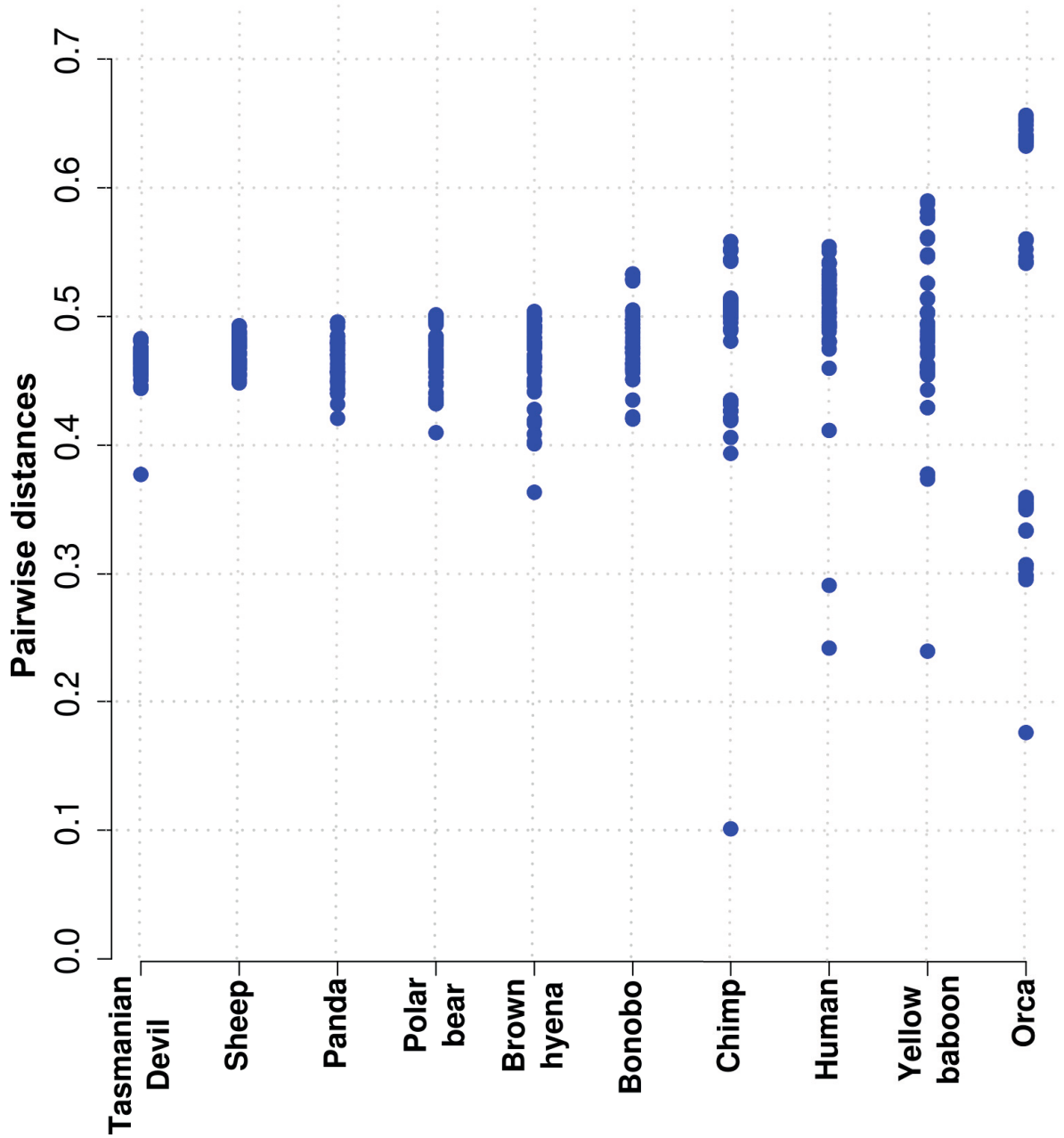
Supplemental figure S5. Principal components analysis produced using single site identity by state comparisons of the 14 wild caught brown hyena individuals in this study. Colours represent country of origin (black - South Africa, blue - Namibia, red - Botswana). Percentages on the X and Y axis represent the percentage of variance explained by each respective component. Results found here are consistant with those calculated using genotype likelihoods.



Supplemental figure S6. Single scaffold principal components analysis produced using single site identity by state comparisons of the 14 wild caught brown hyena individuals. Colours represent country of origin (black - South Africa, blue - Namibia, red - Botswana). Percentages on the X and Y axis represent the percentage of variance explained by each respective component. Size of scaffolds in base pairs can be seen in each respective main label. These plots do generally support some phylogeographic structure but individuals from different regions partially intermingle in a number of plots suggesting that there is not enough power in these single scaffolds to resolve phylogeographic structure in these cases.



Supplemental figure S7. Observed brown hyena density found across Namibia. Differing intensities of red correspond to the brown hyena density in that location. Distribution patterns within Namibia show a lack of overlap between individuals from Southern Namibia and Northern Namibia.



Supplemental figure S8. Comparative population structure analyses performed using single base identity by state comparisons while removing singletons. The Y axis indicates pairwise differences while the X axis shows the species. Each point on the plot represents a single pairwise comparison.

Supplemental tables

Supplemental table S1. Striped hyena *de novo* assembly contig/scaffold statistics.

Contig N50	82579
Scaffold N50	2001328
GC %	41.56
Longest scaffold (bp)	15183300
Assembly length (bp)	2374721933
Number of contigs	54939
Number of scaffolds	5760

Supplemental table S2. Striped hyena *de novo* assembly BUSCO scores calculated using the Eukaryote and Mammalian BUSCO databases.

	Eukaryote BUSCO scores	**Mammalian BUSCO scores**
Complete BUSCOs	271	3835
Complete Single-Copy BUSCOs	262	3807
Complete Duplicated BUSCOs	9	28
Fragmented BUSCOs	10	165
Missing BUSCOs	22	104
Total BUSCO groups searched	303	4104

Supplemental table S3. Brown hyena sample location and sample type details. * Indicates these individuals have approximate sample locations.

Codename	**Country of Origin**	**Sample type**	**latitude**	**longitude**
BHF1*	South Africa	Hair	-33.124584	26.537708
BHF2*	South Africa	Hair	-33.124584	26.537708
BHM1*	South Africa	Hair	-33.124584	26.537708
BHM2*	South Africa	Hair	-33.124584	26.537708
BHM3*	South Africa	Hair	-33.124584	26.537708
NamBH11	Namibia	Blood	-27.608632	15.497852
NamBH13*	Namibia	Blood	-22.79169	14.549477
NamBH20	Namibia	Blood	-26.680351	15.163507
NamBH21*	Namibia	Blood	-20.85327	16.647978
NamBH7	Namibia	Blood	-27.349512	15.912735
NamBH8	Namibia	Blood	-26.99631	15.650119
BotBH3*	Botswana	Tissue	-20.46989	25.12184
BotBH5*	Botswana	Tissue	-20.46989	25.12184
BotBH7*	Botswana	Tissue	-20.46989	25.12184
BH_Love	Tierpark Berlin	Blood		

Supplemental table S4. Brown hyena mapping statistics when mapping to the striped hyena nuclear and our reconstructed brown hyena mitochondrial genome.

Codename	Total SE reads after merging and adapter trimming	Unique reads mapped reads to nuclear genome	bp mapped to nuclear genome	Average coverage	Unique reads mapped reads to mitochondrial genome	bp mapped to mitochondrial genome	Average coverage
BHF1	64422612	47254713	7363724523	3.100	35110	5776562	335.905
BHF2	38641444	33331195	4975565124	2.095	18976	2850634	165.763
BHM1	49209478	43764811	6384610509	2.688	27345	4048045	235.393
BHM2	52780882	45508271	6744127558	2.839	33391	5084205	295.645
BHM3	53172160	46297247	6833441946	2.877	37079	5684861	330.573
NamBH11	52544400	42170064	6641423313	2.796	23041	3784082	220.043
NamBH13	68581118	56379724	8728844324	3.675	19598	3042953	177.132
NamBH20	55366276	43635935	6874142712	2.894	26319	4230429	245.998
NamBH21	90758418	48688930	7181669430	3.024	20916	3194780	185.948
NamBH7	65777674	50924181	8248476080	3.473	24683	4154046	241.556
NamBH8	63671350	47951458	7818176024	3.292	23623	4002347	232.735
BotBH3	38007734	31490650	4912299215	2.068	38406	6549645	380.860
BotBH5	63924040	52012046	8162041861	3.437	49010	8838756	513.971
BotBH7	54158322	43563771	6920001119	2.914	17771	2878601	167.390
BH_Love	881114120	632326566	97757690887	41.165			

Supplemental table S5. List of the numerical IDs of the scaffolds that successfully aligned to the cat X chromosome (CM001396.2) via synteny.

Striped hyena scaffolds aligning to the cat X chromosome
12, 77, 80, 85, 142, 147, 162, 243, 259, 295, 297, 318, 348, 351, 376, 383, 402, 451, 470, 524, 536, 555, 556, 576, 610, 611, 658, 669, 701, 744, 754, 758, 765, 808, 840, 844, 853, 865, 917, 924, 929, 932, 960, 982, 986, 1017, 1019, 1025, 1051, 1070, 1091, 1114, 1115, 1112, 1157, 1130, 1158, 1163, 1165, 1171, 1178, 1181, 1182, 1188, 1217, 1219, 1211, 1230, 1257, 1259, 1254, 1272, 1269, 1292, 1295, 1299, 1318, 1340, 1333, 1349, 1358, 1386, 1388, 1420, 1414, 1421, 1450, 1443, 1461, 1463, 1468, 1471, 1488, 1512, 1508, 1517, 1531, 1547, 1539, 1549, 1567, 1578, 1588, 1615, 1609, 1620, 1635, 1640, 1629, 1630, 1660, 1667, 1677, 1701, 1688, 1705, 1712, 1713, 1765, 1749, 1776, 1789, 1792, 1814, 1825, 1815, 1837, 1849, 1843, 1844, 1845, 1860, 1884, 1881, 1876, 1877, 1879, 1891, 1897, 1898, 1903, 1929, 1955, 1964, 1961, 1974, 2000, 1984, 1995, 1994, 2044, 2022, 2047, 2061, 2054, 2062, 2093, 2098, 2100, 2112, 2111, 2116, 2120, 2125, 2164, 2193, 2183, 2187, 2203, 2221, 2209, 2217, 2226, 2236, 2248, 2261, 2271, 2306, 2335, 2326, 2333, 2350, 2375, 2376, 2388, 2402, 2404, 2418, 2464, 2542, 2557, 2605, 2672, 2719, 3002

Supplemental table S6. Accession numbers for the raw reads from the low coverage genomes used in the population structure comparison and the accession number for the reference these were mapped against.

Species	Accession numbers	Reference sequence
Human (Homo sapian)	ERR010982, ERR010985, ERR010989, ERR010992, ERR011001, ERR011008, ERR019683, ERR019684, ERR019688, ERR033733	PRJNA31257
Sheep (Ovis aries)	SRR501839, SRR501846, SRR501849, SRR501856, SRR501860, SRR501864, SRR501870, SRR501877, SRR501896, SRR501910	GCA_000005525
Yellow baboon (Papio cynocephalus)	SRR3151894, SRR3151901, SRR3151905, SRR3151907, SRR3151909, SRR3151922, SRR3151925, SRR3151931, SRR3151932, SRR3151936	AHZZ00000000.2
Polar bear (Ursus maritimus)	SRR827537, SRR827574, SRR827584, SRR827585, SRR827587, SRR827600, SRR942195, SRR942202, SRR942223, SRR942231	PRJNA210951
Panda (Ailuropoda melanoleuca)	SRR504865, SRR504868, SRR504877, SRR504884, SRR504885, SRR504888, SRR504893, SRR504895, SRR504900, SRR504904	GCA_000004335
Bonobo (Pan paniscus)	ERR032963, SRR740773, SRR740792, SRR740800, SRR740807, SRR740821, SRR740823, SRR740833, SRR740941, SRR741276	GCA_000258655.2
Orca (Orcinus orca)	ERR637310, ERR637314, ERR637317, ERR637322, ERR637329, ERR637332, ERR637336, ERR637344, ERR637347, ERR637352	ANOL00000000
Tasmanian Devil (Sarcophilus harrisii)	ERR1474982, ERR1474983, ERR1474984, ERR1474985, ERR1474986, ERR789027, ERR789028, ERR789029, ERR789030, ERR789031, ERR789032	GCA_000189315
Chimpanzee (Pan troglodytes)	ERR032935, ERR032936, ERR032939, ERR032940, ERR032943-ERR032948, ERR032947-ERR032952, ERR032958, ERR032959, ERR032960, ERR032961	GCA_000001515.5

Supplemental table S7: Accession numbers for the raw reads used in the nuclear genome heterozygosity estimates and the accession code for the reference these were mapped against.

20x coverage genome species	Accesion number	Reference
Bonobo	SRR740794-SRR740801	GCA_000258655.2
Chimpanzee	ERR1709871,ERR1709872	GCA_000001515.5
Cheetah	SRR2737512-SRR2737518	GCA_001443585.1
Iberian lynx	ERR1255587-ERR1255590	FIZN00000000.1
Human (African)	SRR1295432	PRJNA31257
Human (European)	SRR1291026	PRJNA31257
Panda	SRR019734-SRR019754	GCA_000004335
Orca	SRR574970-SRR574972,SRR574975	ANOL00000000
Polar bear	SRR933670-SRR933696	PRJNA210951
Yellow baboon	SRR1513473,SRR1513475-SRR1513478	AHZZ00000000.2
Island fox	SRR5198007,SRR5198009	GCF_000002285.3