Genomic implications of low effective population size in non-model mammalian species

A thesis submitted to the University of East Anglia
for the degree of Doctor of Philosophy

Jessica Peers

100347054

Earlham Institute, UK

University of East Anglia, UK

December 2025

Abstract

Small populations, particularly those that have experienced severe genetic bottlenecks, suffer from increased inbreeding and accumulations of deleterious mutations across the genome. Although many studies have examined the impact of these mutations in protein-coding genes, the effects of low effective population size (N_e) on the non-coding genome remain largely unknown, as most non-coding research to date focuses on model species. In this thesis, I investigate the effects of low N_e in a non-model mammalian species, the cheetah (*Acinonyx jubatus*), considering both the coding and non-coding genome. I identify premature termination codons in fertility-related genes that are shared across multiple unrelated cheetahs. I then evaluate the application of machine learning methods to annotate the non-coding genome of non-model species, negating the requirement for bespoke experimental data and expanding our ability to study species that lack such resources. Finally, I analyse inbreeding and mutation load in both captive and wild cheetahs and identify putative deleterious mutations across the genome in fertility-related genes. Overall, this thesis contributes to our understanding of the impacts of low N_e across the genome and provides a framework to study any non-model species, accelerating research into the non-coding genome and its applications in conservation.

Contents

List of Figures

List of Tables

List of Supplementary Figures

List of Supplementary Tables

List of Abbreviations

AUC-PR	Area under precision-recall curve
AUC-ROC	Area under receiver operator characteristic curve
ATAC-seq	Assay for transposase accessible chromatin sequencing
BERT	Bidirectional encoder representation from transformers
BLSTM	Bidirectional long short-term memory network
bp	Base pair
CADD	Combined Annotation Dependent Depletion
CDS	Coding DNA sequence
ChIP-seq	Chromatin immunoprecipitation sequencing
CITES	Convention on International Trade in Endangered Species of Wild Fauna and Flora
CNN	Convolutional neural network
ddRAD	Double-digest restriction site associated DNA
DNase-seq	DNase I hypersensitive sites sequencing
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements sequencing
FDR	False discovery rate
FN	False negative
FP	False positive
F_{IS}	Inbreeding coefficient
F_{ROH}	Fraction of genome in runs of homozygosity
GO	Gene ontology

GSTF	GeneSeqToFamily
GWAS	Genome-wide association study
IBD	Identical by descent
IUCN	International Union for Conservation of Nature
ML	Machine learning
MHC	Major histocompatibility complex
N_{ROH}	Number of runs of homozygosity
N_e	Effective population size
PCA	Principal component analysis
PTC	Premature termination codon
RADseq	Restriction site-associated DNA sequencing
RNN	Recurrent neural network
ROH	Runs of homozygosity
S_{ROH}	Size of runs of homozygosity
SNP	Single nucleotide polymorphism
SSP	Species Survival Plan
SVM	Support vector machine
TF	Transcription factor
TFBS	Transcription factor binding site
TLR	Toll-like receptor
TN	True negative
TP	True positive
UCE	Ultra-conserved element
US	United States of America
π	Nucleotide diversity

Acknowledgements

I've grown a lot as a scientist over the course of this PhD thanks to the amazing people I've had the privilege to work with. Firstly, I am immensely grateful to my supervisor, Dr Wilfried Haerty. Thank you for providing endless support and guidance, for always putting my wellbeing and interests (cats) first, and for giving me the opportunity to travel to some incredible places. I've learned so much from you and I appreciate all the time and energy you've given to supervise me.

Dr Will Nash, thank you for championing my wellbeing, especially when I was struggling with my health, and for always being willing to spend hours at a time helping me sketch out entire chapters on a whiteboard. Dr Dave Wright, thank you for providing both scientific guidance and emotional support (in the form of letting me cry to you when I'm stressed and letting me cat-sit). Past and present members of the Haerty group, thank you for creating a such welcoming and light-hearted environment and for always being happy to answer my stupid questions.

Thank you to my collaborators (Dr Ellie Armstrong, Dr Klaus-Peter Koepfli, Heather Sibley and African Parks) for giving me the opportunity to work on some exciting projects and to learn from scientists I greatly admire. I'm especially grateful to my colleagues at Fauna Bio for involving me in fascinating research and for making me feel like a part of the team despite the time difference (and the Atlantic Ocean).

I am also incredibly grateful to my fellow students at EI, particularly past and present members of the Earlham Student Body committee, for creating such a supportive and inspiring community. I especially enjoyed being part of the ESB in my second year and planning so many fun events (e.g. cow walks).

I've also grown a lot as a person in the last four years, and for that I must thank all the wonderful people in my life. Sofia, Becky, Mia, Insect George, Kate and Lu. We did it! None of us dropped out! I am so grateful to each of you and I'm so lucky to have had the last 4 years of lunchtimes, pub quizzes, crafts and laughter with you all. I hope when we're old and grey we still hang out and reminisce about our plans to start a cult in King's Lynn.

Megan, Tasha and Kiera. Thank you for always having my back, especially when I'm having Big Feelings, and for supporting me through another degree. I promise I won't do it again.

Mum and Dad. I am so lucky to have your unconditional and unwavering love and support. I'm so grateful for all the phone calls, cat pictures and last-minute visits to Norwich when I was stressing. Thank you for always being proud of me.

Ben. I am so lucky to have you in my life. Thank you for always believing in me and for helping me believe in myself. Sorry for crying at every opportunity and thanks for letting me cover the house with cheetah-related memorabilia. I love you.

All the cats I've met along the way (and their parents, of course): Fig, Poppy, Lychee, Hobbes, Nick, Mimi, Kotchik (who I haven't met in person but deserves a mention), Ron, Linneaus, Tiamat, Elminster, Peanut and Murphy.

During my PhD, I sadly lost both my grandmothers. Gran, who introduced me to cheese and jam sandwiches and Casualty, and Grandma, who loved to hear what I was up to and would tell everyone she knew. Thank you both for always being so proud of me.

I would like to dedicate this thesis to Holly and Pepper, who I love very dearly and who have inspired all my cat-themed research, and to Peanut, who kept me company while I wrote this thesis, even if he sometimes dribbled on my laptop. Meow meow meow.

Preface

Some of the material presented in this thesis has been published:

Chapter 2: **Jessica A Peers**, Will J Nash, Wilfried Haerty, Gene pseudogenization in fertility-associated genes in cheetah (*Acinonyx jubatus*), a species with long-term low effective population size, *Evolution*, Volume 79, Issue 4, 1 April 2025, Pages 574–585, https://doi.org/10.1093/evolut/qpaf005

Chapters 3 & 4: in preparation

Chapters have been adapted to fit the overall thesis structure and formatting requirements. All work described in the chapters was undertaken by the student unless stated otherwise.

Chapter 1

Introduction

1.1 Genetics of small populations

1.1.1 Genetic bottlenecks

Throughout their evolution, mammal species have been subject to population contractions and expansions. These have been caused by a variety of factors, such as hunting, climatic changes, migration or domestication. For example, the late-Pleistocene mass extinction event, where many megafauna species became extinct or experienced severe contractions, is hypothesised to have been caused by over-hunting by humans and sudden climatic changes (Barnosky et al., 2004), whilst humans experienced a population bottleneck as they expanded out of Africa (Lohmueller et al., 2008). Population bottlenecks can lead to genomic changes such as increased homozygosity, pseudogenization of genes, or accumulated deleterious mutations. These historic and recent demographic events have shaped the evolution of mammalian species. The recent availability of large-scale genomic resources across the whole Mammalia class offers a unique opportunity to study and compare the genomic impact of these population contractions. Here, I introduce some key considerations that are necessary for the successful interrogation and interpretation of such resources to identify signatures of low effective population size (N_e).

Genetic bottlenecks, rapid reductions in population size, may be due to environmental causes, such as floods, droughts, wildfires or diseases, or due to human interference, such as hunting, culling, or habitat destruction (Banks et al., 2013; Bijlsma & Loeschcke, 2012; Broquet et al., 2010; Galtier et al., 2000). These events can result in the fragmentation of populations and reduction of their genetic diversity as many alleles are lost in the process, increasing the risk of extinction (Figure 1.1) (Luikart et al., 1998). This makes a population less adaptable to environmental change, as remaining individuals contain much lower diversity and are therefore less likely to contain variation that may be beneficial in a changing environment. A reduction in population size also increases the likelihood of inbreeding, as the chance of mating with a related individual is greater, and can lead to an accumulation of deleterious mutations.

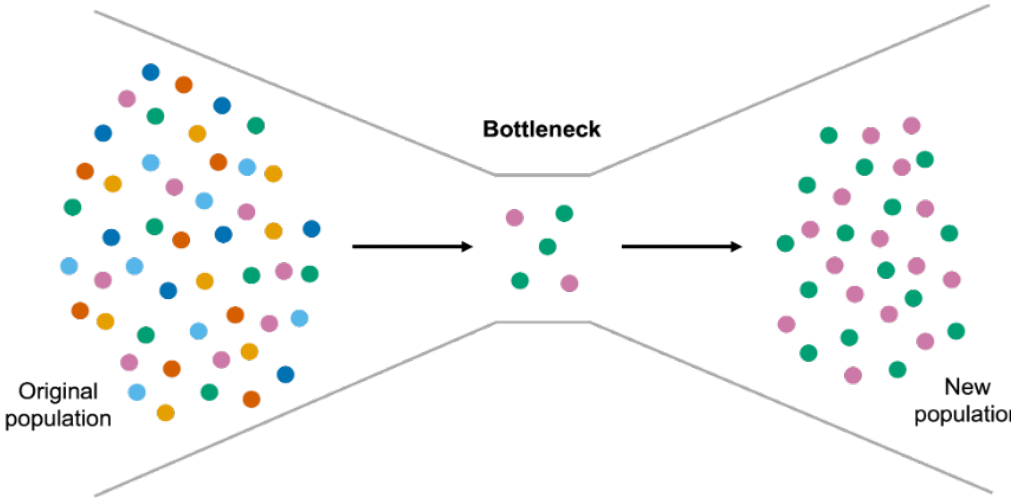


Figure 1.1: **Diagram illustrating a genetic bottleneck.** Alleles are represented by different coloured circles. After the bottleneck, the population expands but with limited alleles, and therefore diversity, compared to the original population.

Deleterious mutations are those which have a negative effect on an organism's fitness, defined as an organism's ability to reproduce (Lande, 1994; Lynch et al., 1993). In protein-coding sequences, these may occur in the form of frameshift, nonsense or missense mutations or premature termination codons (PTCs) (Hindorff et al., 2009; Kryukov et al., 2007; Raes & Van de Peer, 2005; Savas et al., 2006; Sonstegard et al., 2013). Deleterious mutations in the non-coding genome can alter the sequence of functional regions; for example, mutations in transcription

factor binding sites (TFBS) can affect the binding affinity of transcription factors, thereby affecting gene expression (Zheng et al., 2010).

1.1.2 Small populations

Deleterious mutations accumulate more quickly in populations with low N_e or populations with relaxed selective pressures (Björnerfeldt et al., 2006). The risk of extinction in such populations is more greatly impacted by weakly deleterious mutations than by demographic or environmental stochasticity, as the accumulation of weakly deleterious mutations can lead to loss of fitness and genetic inviability (Lande, 1994). Although selection typically acts to remove such mutations, some may persist or accumulate over time, particularly in populations with limited recombination or low N_e.

Muller's ratchet (Muller, 1964) describes this process, stating that with no recombination, irreversible deleterious mutations accumulate over generations (Felsenstein, 1974; Lynch et al., 1993; Muller, 1964; Olofsson et al., 2023). For example, on the Y chromosome, the absence of recombination results in an accumulation of deleterious mutations and subsequent degeneration of the chromosome (B. Charlesworth & Charlesworth, 2000). Mutations occur within a wide spectrum of deleteriousness, directly influencing their likelihood of remaining in the population. Severely deleterious mutations causing lethality or infertility will not be heritable, whereas weakly deleterious mutations with smaller effects on fitness may persist and accumulate over generations, particularly in small populations (Muller, 1964).

Small populations are also more prone to genetic drift, the random fixation or loss of alleles due to chance (Figure 1.2; Masel (2011)), which can lead to the fixation of weakly deleterious alleles in a population (Lande, 1994). The probability of fixation of an allele relates to the level of genetic drift; the likelihood that a deleterious allele will segregate at a high frequency is higher in populations with

a smaller N_e than in larger populations (Crow & Kimura, 1970). This can lead to a 'mutational meltdown': accumulated deleterious mutations and their fitness consequences lead to further population size reduction, increasing vulnerability to genetic drift that causes further fixation of deleterious mutations, eventually resulting in extinction (Lynch & Gabriel, 1990). Additionally, alleles can be lost through genetic erosion, where rare alleles are lost due to death of individuals. This process has a larger effect on smaller populations as it further decreases the population's genetic diversity (Bijlsma & Loeschcke, 2012).

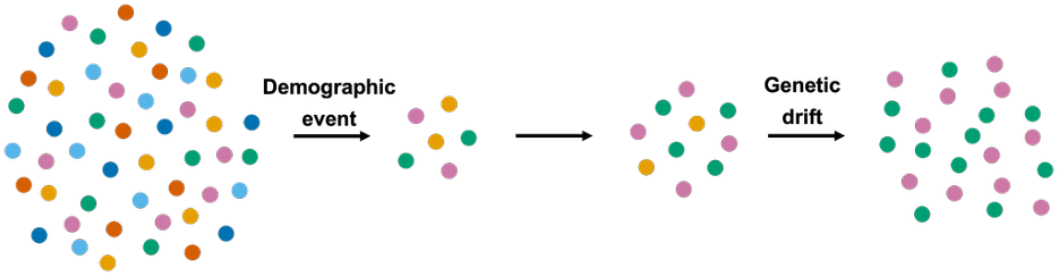


Figure 1.2: **Diagram illustrating genetic drift.** Alleles are shown in different colours. A demographic event (e.g., a bottleneck) results in a small population with reduced diversity. As the population expands, genetic drift can result in the loss (e.g., yellow) or fixation (e.g., pink/green) of alleles due to chance.

1.1.3 Nearly neutral theory of evolution

When investigating weakly deleterious mutations, it is important to consider the nearly neutral theory of evolution (Kimura, 1983; Ohta, 1992). This theory states that most mutations are selectively neutral, or nearly neutral, and segregate in a population due to genetic drift. In small populations, low N_e results in weaker selection, meaning such mutations can segregate at higher frequencies than expected under strong selection. As such, the effective selective value of a mutation S is estimated by the equation $S = 4N_e s$, where N_e is the effective population size and s is the selection coefficient (i.e., whether the mutation is selectively positive or negative, and to what extent) (Figure 1.3).

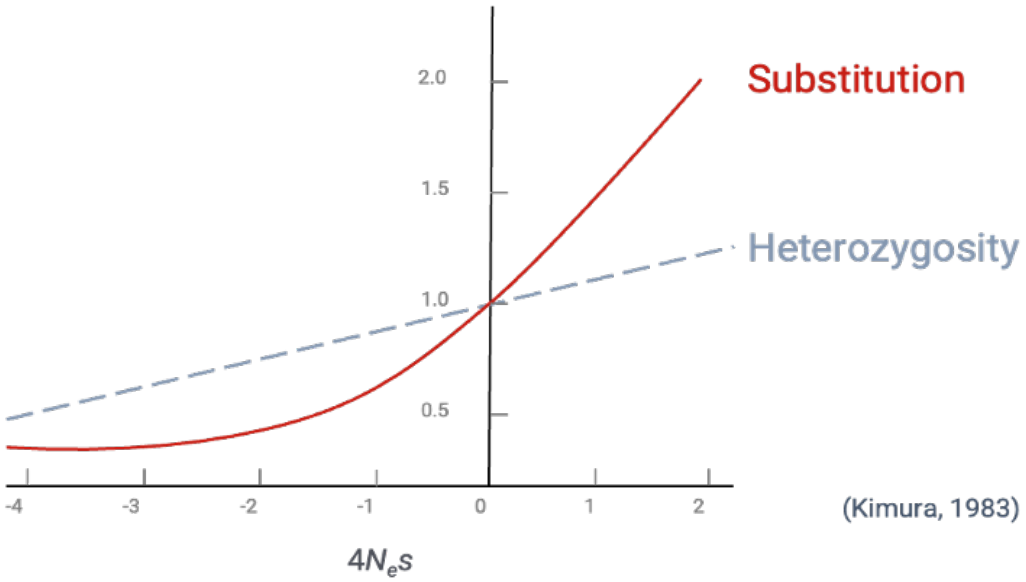


Figure 1.3: **The nearly neutral theory of evolution.** The substitution line shows the probability of a mutation reaching fixation whilst the heterozygosity line shows how mutations segregate over time. With a low N_e, $4N_es$ is close to zero and deleterious mutants behave nearly neutrally. Figure adapted from Kimura (1983).

In a population with large N_e, if the selection coefficient (s) of a mutation is negative (i.e., the mutation has a strong deleterious effect), the effective selective value of this mutation will be negative, resulting in a low probability of fixation and a high likelihood that the mutation will be purged from the population. For the same selective value in populations with a smaller N_e, the mutation will behave selectively nearly neutrally. Therefore, the chance of fixation is almost equal to that of a truly neutral mutation and the mutation may segregate in the population at high frequency. In a population with a low N_e, a mutation would need to be highly deleterious to be purged from the population, meaning weakly deleterious mutations can accumulate in these populations.

It is worth noting that this theory has some limitations, such as its reliance on the infinite-sites model, where each mutation occurs in a new site. However, this model is only accurate if the mutation rate is low enough that the chance of multiple mutations occurring at the same site is close to zero (D. Charlesworth, 2003). The model is therefore not accurate for mutation hotspots (D. Charlesworth, 2003), where the chance of multiple mutations occurring at the same site is increased.

Deleterious mutations may also segregate in a population through linkage, where a selectively neutral or deleterious mutant may be inherited alongside a selectively advantageous allele due to physical proximity on a chromosome (J. M. Smith & Haigh, 1974). As the advantageous allele increases in frequency due to selection, the linked deleterious mutation can "hitchhike" along with it, potentially becoming fixed in the population (Figure 1.4). This process of selection for advantageous alleles can result in a loss of genetic variation at linked loci and can contribute to an accumulation of deleterious alleles (Cruz et al., 2008; Lu et al., 2006; J. M. Smith & Haigh, 1974).

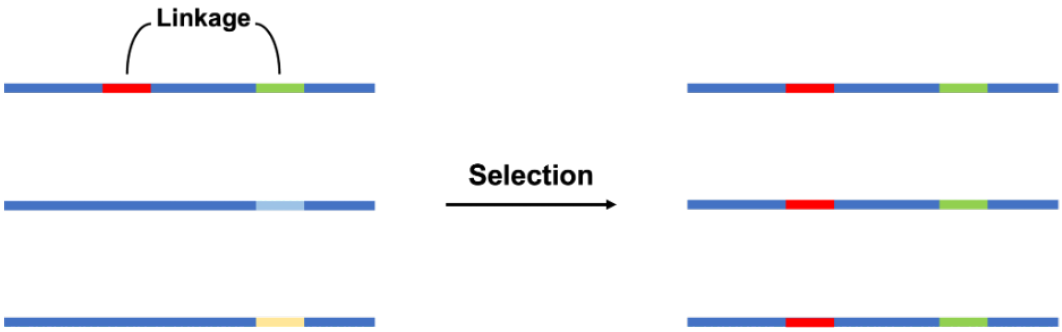


Figure 1.4: **Diagram illustrating linkage and hitchhiking.** A deleterious mutant (red) is linked to an advantageous variant (green). As the advantageous variant is selected for, the deleterious mutant is also inherited, resulting in both advantageous and deleterious mutants segregating in the population.

1.1.4 Inbreeding

Inbreeding, reproduction by related individuals, can occur in small populations due to demographic events (e.g., bottlenecks) or due to controlled breeding (e.g., domestication). In small populations, inbreeding increases the chance of homozygous expression of recessive deleterious alleles, many of which normally segregate at low frequency and would not otherwise be expressed, resulting in inbreeding depression (Bosse et al., 2019; Mukai et al., 1972). Inbreeding depression is the loss of fitness of a population due to increased homozygosity and expression of recessive deleterious mutations (D. H. Charlesworth & Charlesworth, 1987; D. Charlesworth & Willis, 2009). In *Drosophila melanogaster*, mutations with a large deleteri-

ous effect (e.g. lethal mutations) contribute significantly to inbreeding depression (B. Charlesworth & Charlesworth, 1999), but the cumulative impact of mutations with small effects can also significantly contribute to inbreeding depression (D. H. Charlesworth & Charlesworth, 1987).

Inbreeding can be estimated from molecular data using the inbreeding coefficient (F), with F statistics including excess homozygosity (F_{HOM}), homozygous-by-descent (F_{HBD}) and runs of homozygosity (F_{ROH}), although the efficacy of each of these measures to accurately estimate inbreeding depression has been debated (Alemu et al., 2021; Caballero et al., 2021). Runs of homozygosity (ROH) occur when identical haplotypes are inherited from both parents, resulting in long stretches of the genome with no heterozygosity (F. C. Ceballos et al., 2018). ROH is often used to detect recent and historical inbreeding, and recent studies support this use of ROH as a genomic marker of inbreeding depression (F. C. Ceballos et al., 2018; Curik et al., 2014; Kyriazis et al., 2025; Shafer & Kardos, 2025). For example, in grey wolves (*Canis lupus*), ROH were used to show that entire chromosome pairs were identical-by-descent (IBD) (Kardos et al., 2018), whilst in orca (*Orcinus orca*), variations in ROH lengths were used to support demographic reconstruction and identify populations at risk of inviability (Foote et al., 2021).

Another common estimator of inbreeding and extinction risk is mutation load, defined as the reduction in fitness due to repeated occurrence of deleterious mutations (Couvet & Ronfort, 1994; Haldane, 1937). This forms a component of overall genetic load, which considers the total reduction in population fitness relative to the genotype with maximum fitness (Agrawal & Whitlock, 2012; B. Charlesworth & Charlesworth, 1999; Dussex et al., 2023), however both metrics are commonly used in genomic studies of inbred species. Mutation load is a combination of the recessive deleterious alleles segregating in the population (segregation load) and mutations that have become fixed through genetic drift (drift load) (van Oosterhout, 2020). Whilst mutation load is not directly correlated with extinction risk, in small populations with increased homozygosity (such as inbred populations), the segregation load becomes exposed, leading to expression of deleterious mutations

(van Oosterhout, 2020).

Inbreeding can have severe and rapid effects on a population. For example, in a population of Scandinavian grey wolves (*Canis lupus*) founded by one female and two males, substantial inbreeding due to a severe founder effect has been described (Viluma et al., 2022). After approximately five generations, 10-24% of the diploid genome, containing 160,000 Single Nucleotide Polymorphisms (SNPs), had been lost (Viluma et al., 2022). Genetic drift in this population has been implicated in the loss of ancestral alleles, fixation of deleterious variants and an increase in homozygous deleterious alleles (realized load) (Smeds & Ellegren, 2023). This extreme inbreeding has resulted in a population with decreased diversity and adaptability, as well as increased expression of recessive deleterious alleles. The effects of extreme inbreeding can be long-lasting; in cheetahs (*Acinonyx jubatus*), inbreeding depression is observed in modern individuals as a result of a hypothesised bottleneck 10,000 years ago (O'Brien, 1994a, 1994b).

Inbreeding can result in severe phenotypic abnormalities, as has been observed in multiple wild mammals. The cheetah has incredibly low genetic variation, potentially due to a severe bottleneck at the end of the last ice age, with over 90% homozygosity observed in the genome (Dobrynin et al., 2015; Menotti-Raymond & O'Brien, 1993). Cheetahs have been observed with low quality and pleiomorphic sperm, asymmetrical skulls and high juvenile mortality in both wild and captive individuals (O'Brien et al., 1985; Wayne et al., 1986; Wildt et al., 1983), although more recent studies question the association between these symptoms and historical inbreeding (Crosier et al., 2018; Evermann et al., 1988; Heinrich et al., 2017; Terio et al., 2018).

Inbreeding depression symptoms have been observed across the tree of life, although the most well-known cases occur in charismatic mammal species. The Florida panther (*Puma concolor couguar*) experienced a severe bottleneck in the 1970s, leading to sperm abnormalities, congenital heart defects and increased susceptibility to deadly infectious diseases (Roelke et al., 1993). The black-footed

ferret (*Mustela nigripes*) also experienced a recent severe bottleneck, with high disease susceptibility, low fecundity and poor quality sperm observed as effects of inbreeding depression (Santymire et al., 2006, 2014). Similar symptoms have been observed in fish, where impaired reproductive success and growth measures have been linked to inbreeding depression (Ala-Honkola et al., 2009; Kincaid, 1983; Sheridan & Pomiankowski, 1997). Inbreeding depression has also been observed in wild and captive insects, although empirical data is lacking (Leung et al., 2025), as well as marine invertebrates and plants, where inbreeding predominantly occurs due to the difficulty of gamete dispersal (Lande et al., 1994; Olsen et al., 2021).

It is widely accepted that we are currently experiencing the sixth mass extinction event – the Holocene extinction – in which the current rate of extinction is 1,000 times the background rate (Pimm et al., 2014). Endangered species are particularly vulnerable to inbreeding as there are limited breeding populations, despite various conservation interventions applied to reduce inbreeding. In Scandinavian wolves, introducing immigrants to an inbred population temporarily shifted deleterious alleles from homozygous to heterozygous, however without permanent connectivity or continued introductions, inbreeding resulted in re-exposure of these deleterious mutations (Smeds & Ellegren, 2023), highlighting the importance of continued gene flow to rescue inbred populations. In captive populations of endangered species, breeding programmes aim to reduce inbreeding (Hedrick & Miller, 1992): without intensive management, genetic variation can decrease as the populations are often closed to novel diversity (Lande & Barrowclough, 1987; Woodworth et al., 2002).

1.1.5 Purging

Purging, the removal of strongly deleterious mutations by selection, occurs as recessive deleterious alleles are exposed to natural selection due to increased homozygosity caused by inbreeding (Urfer, 2009). Individuals with resultant expression of the deleterious allele are less likely to survive and reproduce, meaning the delete-

rious allele is purged from the population (Kalinowski et al., 2000). The success of purging depends on the characteristics of the deleterious loci; purging is most effective on mutations with large deleterious effects as these are most likely to prevent reproduction of affected individuals (D. Charlesworth & Willis, 2009). Rare or weakly deleterious mutations are less likely to be purged, despite the fact that such mutations may have a large cumulative contribution to inbreeding depression (D. Charlesworth & Willis, 2009).

Effective purging has been observed in several populations; for example, a substantial fraction of the inbreeding load of captive ungulate populations with low N_e was purged (López-Cortegano et al., 2021). An assessment of the relationship between purging and population size in this study suggested that larger populations would require more generations for purging to be detected compared to smaller populations (López-Cortegano et al., 2021). Evidence for purging was also observed in wild populations of isolated mountain gorillas (*Gorilla beringei beringei*) (Xue et al., 2015) and in Alpine ibex (*Capra ibex*) (Grossen et al., 2020), where highly deleterious mutations were purged from the population whilst mildly deleterious mutations continued to accumulate. Potential purging was observed in Indian tigers (*Panthera tigris tigris*), where a small isolated population had lower loss-of-function mutation load than two large connected populations (Khan et al., 2021). This empirical evidence suggests that purging in both captive and wild populations is possible, although care must be taken when studying purging in captive populations. An increase in fitness observed compared to the wild may instead be due to a fitness rebound caused by the captive environment and increased intensity of human husbandry (Clifford et al., 2007; Kalinowski et al., 2000). Indeed, adaptation to captivity can itself significantly alter the extent to which any given mutation is deleterious or beneficial and can change this relationship extremely quickly (Christie et al., 2016).

However, reviews of purging across a range of taxa, including mammals, birds, reptiles and amphibians, show that the resultant reduction in inbreeding depression is incredibly low ($<1\%$), suggesting very limited fitness benefits to purging are

likely to be observed (Boakes et al., 2007). Multiple reviews find little evidence of purging and emphasise the risks of fixation of deleterious alleles, suggesting inbreeding should still be avoided rather than the alternative practice of inducing purging (Boakes et al., 2007; Leberg & Firmin, 2008).

Deliberate inbreeding to induce purging has previously been attempted in the Speke's gazelle (*Gazella spekei*) (Templeton & Read, 1983). Whilst initial analyses suggested this reduced inbreeding depression in the population, subsequent reanalysis did not find any evidence of this (Boakes et al., 2007; Kalinowski et al., 2000). Deliberate inbreeding can be risky in captive populations; increased homozygosity and subsequent expression of deleterious alleles may pose an extinction risk if the mutations are not purged quickly enough, and will reduce genetic diversity (Hedrick, 1994; Hedrick & Garcia-Dorado, 2016). Additionally, as evidence suggests purging rarely counteracts the effects of inbreeding depression, it is unlikely that the risks posed by deliberate inbreeding will be outweighed by positive fitness effects (Boakes et al., 2007).

1.2 Genomics methods

To understand the impacts of demographic events on the genetic variation present within populations, it is essential to first identify this variation at the genomic level. Previous approaches, such as microsatellites or restriction site-associated DNA sequencing (RADseq), relied on the genotyping of specific markers in many individuals (Davey & Blaxter, 2010; Litt & Luty, 1989; Selkoe & Toonen, 2006; Weber & May, 1989). However, a major limitation of these approaches is that only a fraction of the genome is sampled. To accurately characterise the genomic consequences of population demographic events, it is necessary to investigate the whole genome.

1.2.1 Reference genomes

Since the first fully sequenced genome (the bacteria *Haemophilus influenzae*) in 1995 (Fleischmann et al., 1995), sequencing technology has advanced rapidly and thousands of more complex genomes have been sequenced (Hu et al., 2021; Shumate & Salzberg, 2021). Starting with the human (Lander et al., 2001; Venter et al., 2001) and mouse (Mouse Genome Sequencing Consortium et al., 2002), the genomes of over 200 mammals have been sequenced in the last few decades (Christmas et al., 2023; Zoonomia Consortium, 2020).

Current projects aim to produce reference quality genome assemblies for vast numbers of species; the Earth Biogenome Project aims to sequence the genomes of all eukaryotic species on the planet (Lewin et al., 2018), with projects focusing on specific geographic regions such as the Darwin Tree of Life project in the UK (Darwin Tree of Life Project Consortium, 2022). The Zoonomia Project, focusing on understanding the conservation landscape of the human genome through comparative genomics of mammals, published 131 genome assemblies and generated a whole-genome alignment of 240 eutherian mammals (Christmas et al., 2023; Zoonomia Consortium, 2020), which is an incredible resource for studying mammalian evolution.

The quality of reference genomes has advanced as new and improved genome assemblies are published (Shumate & Salzberg, 2021). In the mid 2000s, short-read next-generation sequencing was the cutting edge technology applied to most genome sequencing, however the short read length made it difficult to assemble repetitive or structurally complex genomic regions (S. A. Simon et al., 2009). Long-read sequencing technologies, such as PacBio (Rhoads & Au, 2015b) and Oxford Nanopore Technologies (Jain et al., 2016), can generate sequence reads up to megabases in length and can provide high accuracy, resolution and throughput, meaning these technologies are now becoming the standard (Logsdon et al., 2020; van Dijk et al., 2018). Further to this, chromosome conformation capture sequencing methods, such as Hi-C (Belton et al., 2012), can be used to order con-

tigs and scaffolds, enabling chromosome-level assembly. Using a combination of such technologies, the human genome has been refined over the last two decades to result in a telomere-to-telomere assembly, setting a precedent for future work (Church, 2022; Nurk et al., 2022). Whilst this is promising, at the time of writing, limitations in sample availability and financial constraints still limit our ability to apply long-read methods to endangered species and wild populations globally.

1.2.2 Genome annotation

To properly interpret the link between the phenotype and genotype of an organism, it is necessary to identify protein coding genes and associated functional non-coding regions. This process is known as genome annotation. Experimental evidence is required to determine the locations of coding regions; RNA-seq expression data provides such evidence as transcripts can be sequenced and mapped to the genome to identify their location (Salzberg, 2019). For a deeper understanding of the location of genes, tissue-specific expression patterns can be identified by using RNA-seq in different tissues (Zhu et al., 2016). Full-length isoforms can be sequenced with Iso-seq to understand splice variation and improve an annotation (Beiki et al., 2019). Another approach, *ab initio* gene prediction, uses sequence features such as start and stop codons and splice sites to identify coding sequences (Stein, 2001). These methods create a comprehensive dataset that can be used to generate an annotation.

The process of genome annotation presents a computational challenge (Salzberg, 2019). Initial genome annotation tools, such as Genscan (Burge & Karlin, 1997), used pattern recognition approaches to identify protein-coding regions. However, these patterns may occur by chance and this likelihood of false-positive results meant other approaches were required (Miller et al., 2004). Tools such as MAKER (Cantarel et al., 2008), BRAKER (Hoff et al., 2016), Funannotate (Palmer, 2016) and Augustus (Hoff & Stanke, 2019) use iterative unsupervised training and RNA-seq data to generate more accurate annotations

(Cantarel et al., 2008; Hoff & Stanke, 2019; Hoff et al., 2016). Automated annotation pipelines also exist; RefSeq and Ensembl both use an automated pipeline to annotate genomes as they are uploaded to the databases (Aken et al., 2016; W. Li et al., 2021).

Genome annotations are most accurate when experimental data is used to inform predictions (Miller et al., 2004). However, tissue-specific RNA-seq is not possible for all species, particularly rare wild species, as exhaustive sampling from multiple tissues and individuals is required. To generate annotations for species with a reference genome but no RNA-seq data, a "lift-over" approach can be used. Using homology and synteny information, orthologous genes are identified and annotations are lifted from a closely related species to provide gene information to the newly assembled genome (Shumate & Salzberg, 2021), such as in tools like UCSC liftOver (Kuhn et al., 2013) and halLiftover (Hickey et al., 2013). A more recent tool, LiftOff (Shumate & Salzberg, 2021), aligns gene sequences from the target genome to the reference genome of a closely related species to map genes with high accuracy. These methods avoid the necessity to generate a new annotation from scratch (and the associated requirement for experimental data) and enable annotations to be produced for species with limited genomic resources, which is particularly of interest when studying endangered or rare species.

One major issue with this method is that only the genes and genome features in the target species with homologs in the reference species will be annotated; features present in the target species but not reference species will not be annotated. Additionally, due to the fragmentation of some genome assemblies, genes may be split across different contigs, appearing to be absent. Therefore, genome annotations, particularly those lifted from one species to another, cannot be completely accurate.

1.2.3 Comparative genomics and non-model species

The generation of genome assemblies has enabled the comparison of genomes from different species to understand evolution, phylogeny and gene function (Miller et al., 2004). The first comparative study of multiple genomes investigated 12 *Drosophila* species, identifying signatures of positive selection in genes and regulatory regions (Drosophila 12 Genomes Consortium et al., 2007). Such comparisons rely on whole genome alignments, enabling the identification of evolutionarily conserved regions across species. Tools such as LastZ (Harris, 2007), MultiZ (Blanchette et al., 2004) and Cactus (J. Armstrong et al., 2020) can be used to generate whole genome alignments, which is a computationally intensive process as billions of bases must be aligned. Downstream analyses depend on accurate genome annotations, but care must be taken when comparing reference genomes that were annotated using different methods; orthologous genes may be annotated in one species but not another, therefore erroneously appearing to be lineage-specific (Weisman et al., 2022).

In 2011, the genomes of 29 eutherian mammals were published alongside a comparative study which identified regions of the human genome under positive or purifying selection (Lindblad-Toh et al., 2011). A subsequent study as part of the Zoonomia Project generated a reference-free whole genome alignment of 240 eutherian mammals, more than half of which were previously unsequenced, enabling novel investigation into mammalian evolutionary constraint at an unprecedented resolution and greater insight into genomic variants associated with increased disease risk in both mammals and humans (Christmas et al., 2023; Zoonomia Consortium, 2020). Whilst these large comparative projects have dramatically increased the availability of reference genomes for non-model species, corresponding experimental data required for bespoke genome annotations is still lacking. As previously described, genome annotations generated without experimental data can be inaccurate and miss species-specific patterns, preventing accurate analysis of the genome. This is particularly problematic when investigating species-specific dele-

terious mutations, as inaccurate annotations can result in both false positive and false negative results for deleterious alleles.

1.2.4 Genomic studies of deleterious mutations

Growing genomic resources from annotated genomes and population-level data can be used to identify high-priority variants for further study, such as disease-associated variants. For this, functionally important variants must be identified. Computational tools can functionally classify variants; for example, SIFT (Sorting Intolerant From Tolerant) predicts whether a SNP affects the function of the protein and distinguishes between functionally neutral and deleterious mutations (Ng & Henikoff, 2003). Identifying deleterious or pathogenic variants can give insight into diseases and associated healthcare, as well as evolution and gene expression (Frazer et al., 2021).

A pathogenic variant is one that increases the individual's susceptibility or predisposed risk to developing a disease or disorder. Methods to quantify the pathogenicity of protein-coding variants may rely on training machine learning (ML) models using known disease labels (Frazer et al., 2021). However, using this method relies on existing disease labels, which require an understanding of the phenotype. Due to the vast number of human genomes sequenced and the effort in identifying pathogenic variants in humans, there is a wealth of data on human deleterious mutations (Lappalainen et al., 2019), but this may not apply to other species. A recent study (Frazer et al., 2021) used deep generative models to predict pathogenic variants without using known disease labels, which removes the reliance on genomic experimental data. Another study utilised clustering, which is the non-random distribution of disease-associated variants (Quinodoz et al., 2022). The authors created a pathogenicity predictor score which accurately identified pathogenic mutations, particularly those associated with cancer and autosomal-dominant disease (Quinodoz et al., 2022).

Previous studies on the effects of inbreeding have mainly focused on protein-coding regions of the genome. Researchers have focused on identifying nonsense, missense or frameshift mutations in protein-coding regions, as these affect protein production and function (Z. Zhang et al., 2012). Pseudogenization is a key interest here; deleterious mutations occurring within genes, leading to PTCs, missing start codons or frameshift mutations, result in pseudogenisation (loss) of the gene (Khajavi et al., 2006; W. H. Li et al., 1981). However, there is growing evidence that regulation of gene expression is the key genetic mechanism involved in organisation, diversification and novel traits in organisms (Carroll, 2000; Wray et al., 2003; Lindblad-Toh et al., 2011). Therefore, the focus must shift to functional non-coding regions of the genome.

1.2.5 Non-coding regions of the genome

The majority of disease- and trait-associated variants identified in genome-wide association studies (GWAS) are located in functional non-coding regions (ENCODE Project Consortium, 2012; Hindorff et al., 2009; Ricaño-Ponce & Wijmenga, 2013). Based on a catalogue of SNP-trait associations from published data, it has been suggested that 88% of disease- and trait-associated SNPs are located in intergenic or intronic regions (Hindorff et al., 2009). Additionally, approximately 8-11% of non-coding sites in the human genome are suggested to be functional based on evidence of selective constraints (Christmas et al., 2023; Huang et al., 2017; Meader et al., 2010; Rands et al., 2014; Siepel et al., 2005).

Functional non-coding regions are often involved in regulating gene expression (Carroll, 2000; Lindblad-Toh et al., 2011; Wray et al., 2003), and include regulatory elements such as promoters, enhancers, silencers and insulators (Maston et al., 2006). TFBS are of particular interest in the non-coding genome as these play a key role in regulating gene expression (Boeva, 2016). A mutation in a TFBS can alter the binding affinity of the transcription factor to the site, resulting in a change to gene expression; a deleterious mutation in a TFBS can result in pseudogenisation

of a gene (Pinoli et al., 2019).

Despite the importance of investigating deleterious mutations in the non-coding genome, the difficulty in identifying functional non-coding regions has limited the scope of existing studies. The two methods predominantly used to identify such regions involve comparative genomics and experimental data. Based on the assumption that regulatory regions will be highly conserved over evolutionary time, highly-conserved regions can be identified through comparing genomes from multiple species. However, sequence conservation does not necessarily indicate functional conservation, and this method is not sufficient to accurately predict regions such as TFBS as regulatory activity can be maintained despite high sequence turnover (Cochran et al., 2022; Dermitzakis & Clark, 2002; Kellis et al., 2014; Rands et al., 2014). The use of experimental assays provides more accurate annotation of regulatory elements, however they require high-quality samples from multiple tissues and individuals, making them unfeasible for studies on non-model and wild species (Ernst et al., 2025; Ruiz Daniels et al., 2023; Z. Wang et al., 2015; Wiegleb et al., 2022).

Although GWAS and whole genome sequencing have enabled the identification of trait-associated variants and their abundance in non-coding regions (ENCODE Project Consortium, 2012), it can be difficult to determine whether the association between variant and trait is due to the specific variant itself or a linked variant (L. Chen et al., 2022). Recent studies have identified functional non-coding variants using high-throughput functional assays (Melnikov et al., 2014; Wen et al., 2020). To do this, the authors measured the functional effect of the variant by determining the molecular phenotypic change, such as change in gene expression, of the variant in different cell and tissue types. Whilst this does enable the identification and validation of functional non-coding variants, such laboratory protocols cannot feasibly be scaled up to characterise the millions of variants that exist in the non-coding genome (L. Chen et al., 2022; Koch, 2020). Therefore, to accurately identify functional non-coding regions in non-model species, and to predict the impact of variants within these regions, another method must be used.

ML offers a potential solution to this task. Using data from model species, models can be trained to predict functional non-coding regions, such as enhancers and promoters, with increasing accuracy as this technology develops (Ghandi et al., 2014; Kelley et al., 2018; Novakovsky et al., 2023; Quang & Xie, 2016). Models are typically trained on data from experimental assays such as ATAC-seq, ChIP-seq and DNAse-seq, which characterise regions of open chromatin or regulatory activity, and can then predict similar regions across the genome (Buenrostro et al., 2013; D. S. Johnson et al., 2007; L. Song & Crawford, 2010). Machine learning models can also accurately predict non-coding variant impacts, trained on human disease annotations or chromatin-profiling data (Caron et al., 2019; Schubach et al., 2017; Zhou & Troyanskaya, 2015). Preliminary evidence suggests these models can make accurate predictions across species, allowing the potential prediction of functional non-coding regions in non-model species (Kaplow et al., 2023; Kelley, 2020; Minnoye et al., 2020).

1.3 Cheetah: a unique case study

With a long-term low effective population size, a well-documented captive history and a range of publicly-available genomic resources, the cheetah provides an ideal case study to examine the genomic consequences of low N_e. Here, I outline the current and historic demography of the cheetah and synthesise progress to-date in cheetah conservation genetics.

The cheetah is the fastest living land mammal on the planet and a charismatic felid species, however its survival is currently under threat and just 7,100 individuals remain (Durant et al., 2017). Whilst the cheetah was previously found across Africa and southwest Asia, the species now inhabits just 9% of its original range, with the majority of cheetahs (67%) found outside protected areas (Durant et al., 2017). Cheetahs are exposed to a range of anthropogenic threats, such as habitat loss and fragmentation, human-wildlife conflict, prey scarcity, poaching and the illegal

wildlife trade (Belbachir et al., 2015; Durant et al., 2017; IUCN/SSC, 2015; L. Marker, 2019; Minja, 2025).

Four subspecies of cheetah are recognised by the IUCN: Southern and Eastern African (*A. j. jubatus*), Northeast African (*A. j. soemmeringii*), Northwest African (*A. j. hecki*) and Asiatic (*A. j. venaticus*). However, growing genomic evidence supports the previous classification of five distinct subspecies, with *A. j. jubatus* split into *jubatus* and *raineyi* (Krausman & Morales, 2005; Prost et al., 2022). Cheetahs are classified as Vulnerable by the International Union for Conservation of Nature (IUCN), with *A. j. venaticus* and *A. j. hecki* both classified as Critically Endangered (Belbachir et al., 2015; Durant et al., 2021; Jowkar et al., 2008). However, following a systematic review of global cheetah populations, Durant et al. (2017) called for a reclassification of cheetahs to Endangered. The Asiatic subspecies (*A. j. venaticus*) was once widespread across West, South and Central Asia, but now is only found in Iran, where as few as 24 individuals survive (Farhadinia et al., 2018; Taktehrani et al., 2025). Approximately 60% of the world's wild cheetahs occur in one contiguous population across southern Africa (Durant et al., 2017) and much of this population overlaps with commercial farmland, putting the population at risk of persecution and human-wildlife conflict (Prost et al., 2022; Weise et al., 2017).

The cheetah was one of the first species to be considered in the field of conservation genetics, with observations of phenotypic abnormalities being attributed to low genetic diversity since the early 1980s. O'Brien et al. (1983) observed monomorphism at 47 allozyme loci across 55 cheetahs, with significantly low average heterozygosity in fibroblast proteins. This observation was subsequently extended to major histocompatibility complex (MHC) genes, a vital part of the immune system; reciprocal skin grafts of 12 unrelated cheetahs showed no signs of rejection, suggesting little diversity in MHC loci (O'Brien et al., 1985). This has subsequently been revisited: the low number of MHC class I alleles previously observed was likely due to sample size rather than low genetic diversity, whilst low diversity in MHC class II alleles was confirmed (Castro-Prieto et al., 2011). In

comparison to other felids, cheetahs show lower MHC and toll-like receptor (TLR) diversity, however observations of free-ranging cheetahs suggest little phenotypic impact to their immunocompetence (Heinrich et al., 2017; Meißner et al., 2024). In fact, high disease susceptibility has only been observed in captive cheetahs, likely as a consequence of husbandry conditions rather than genetic diversity (Evermann et al., 1988; Heinrich et al., 2017; Terio et al., 2018).

Alongside high disease vulnerability, reproductive success (or lack thereof) has been highly studied in cheetahs as a suggested consequence of their demographic history (Wildt et al., 1983). Initial studies in wild cheetahs identified low sperm quality, with a lower concentration and percent motility and a higher proportion of morphologically abnormal spermatozoa compared to domestic cats (Wildt et al., 1983), although the cheetah had fewer abnormal sperm than leopards and pumas (Wildt et al., 1988). No evidence of reproductive impairment was observed in wild cheetahs, whilst breeding success varied significantly across captive collections (Lindburg et al., 1993; L. Marker & O'Brien, 1989). More recent research has questioned the link between heterozygosity and reproductive success, showing across captive and wild cheetahs that low heterozygosity did not correlate with low semen quality (Terrell et al., 2016). Studies of fertility of captive cheetahs suggest husbandry has more impact on reproductive success than sperm condition (Crosier et al., 2018; Koester et al., 2015, 2017; Lindburg et al., 1993; Woc Colburn et al., 2018).

Additional observations of phenotypic abnormalities in cheetah include skeletal deformities and high juvenile mortality, which have both been suggested as a consequence of low genetic diversity (Wielebnowski, 1996). Morphological asymmetry has been observed in the cheetah, interpreted as a result of developmental instability due to inbreeding (Wayne et al., 1986), with observations of kinked tails, focal palatine erosion and crowded incisors across a population of free-ranging Namibian cheetahs (Marker-Kraus, 1997). Over 70% infant mortality was observed in wild cheetahs, which is extremely high compared to other mammals (Laurenson, 1994), however captive-bred cheetahs have a higher average number of surviving cubs per

litter compared to other captive felids (Wielebnowski, 1996).

Whilst the cause of such abnormalities is currently still up for debate, the fact that cheetahs have low genetic diversity is widely supported by genetic data, with over 90% homozygosity observed in the genome (Dobrynin et al., 2015). The cause of this low genetic diversity has multiple hypotheses, all of which agree on a timescale predating the Anthropocene. Early genetic studies using a range of molecular markers suggested the cheetah experienced at least one severe bottleneck in the late Pleistocene, around 10-12 kya (Dobrynin et al., 2015; Menotti-Raymond & O'Brien, 1993; O'Brien et al., 1983, 1987), with additional evidence from microsatellites, MHC loci and whole genomes supporting this theory (Castro-Prieto et al., 2011; Dobrynin et al., 2015; Driscoll et al., 2002). Dobrynin et al. (2015) also suggested an ancient bottleneck or founder event over 100 kya when cheetahs migrated out of the Americas into Eurasia and Africa.

An alternative hypothesis suggests the low genetic diversity is predominantly caused by a long-term low effective population size, potentially reinforced by their polygynous mating system, where one male mates with multiple females (Pimm et al., 1989). Kim et al. (2016) utilised whole-genomes to construct demographic history of several felids, identifying a low N_e in the cheetah for at least the last three million years. The 'metapopulation dynamics' hypothesis suggests a continuous cycle of extinction and re-colonisation of habitats, which would also explain the observed low genetic diversity in cheetahs (Hedrick, 1996). Finally, a recent study applied Bayesian coalescent-based methods to 38 microsatellite loci to suggest a hypothesis of a gradual historical decline in population size rather than a sudden bottleneck (Fabiano et al., 2025), supported by Dobrynin et al. (2015)'s observation of a decreasing population size over the last 100,000 years. Fabiano et al. (2025) consistently estimated low effective population size across several coalescent models, with a present-day N_e of southern-African cheetahs ranging from 700 to 1,600.

Regardless of the hypothesis, it is widely agreed that the cheetah has suffered

some kind of historic demographic decline and evidence suggests the species has had a long-term low N_e. Whilst the extent to which phenotypic abnormalities in cheetahs reflect historic inbreeding remains debated, population genetics theory predicts that their prolonged low N_e is likely to have led to the accumulation of deleterious mutations. Several studies have investigated such mutations in the cheetah. Dobrynin et al. (2015) used whole-genome data to find an accelerated accumulation of non-synonymous mutations in the cheetah compared to the domestic cat, tiger, dog and human. Eighteen reproduction-related genes contained damaging mutations implicated in sperm function, including an excess of non-synonymous mutations in AKAP4, a gene involved in spermatozoal development (Dobrynin et al., 2015). Whilst these mutations were not observed in the tiger (*Panthera tigris*), domestic cat (*Felis catus*), wildcat or Asiatic Gir lion (*Panthera leo*), it was not possible to confirm these mutations are cheetah-specific as the sister species to the cheetah, the puma (*Puma concolor*), was not included in the study.

A subsequent study utilising whole genomes of Sumatran tigers (*P. tigris sumatrae*), cheetahs and snow leopards (*P. uncia*) also identified deleterious variants in 201 genes related to spermatogenesis, B cell mediated immunity and embryo development, although the distribution of these variants across cheetah populations is unknown (G. Samaha et al., 2021). Both of these studies focused solely on mutations within protein-coding genes, comparing a limited number of species with a limited amount of data from different cheetah populations. Therefore, whilst there is growing evidence for accumulations of deleterious mutations in the cheetah genome, there is still much work to be done to fully understand the impact of long-term low N_e on the cheetah genome and its link to phenotypes.

1.4 Thesis overview

This thesis broadly aims to contribute to our understanding of the impact of low effective population size at a genome-wide scale. I utilise the cheetah as a case study to investigate the accumulation of deleterious mutations due to long-term low N_e. The species' demographic history and resultant inbreeding depression, alongside its well documented conservation history of intensive management and captive breeding, make it an ideal taxon in which to understand the impacts of low effective population size on the genome. Specifically, I aim to identify deleterious mutations in the coding and non-coding genome, assess their distribution across populations of captive and wild cheetahs and evaluate the ability of current cutting-edge genomic tools to identify signatures and implications of low N_e.

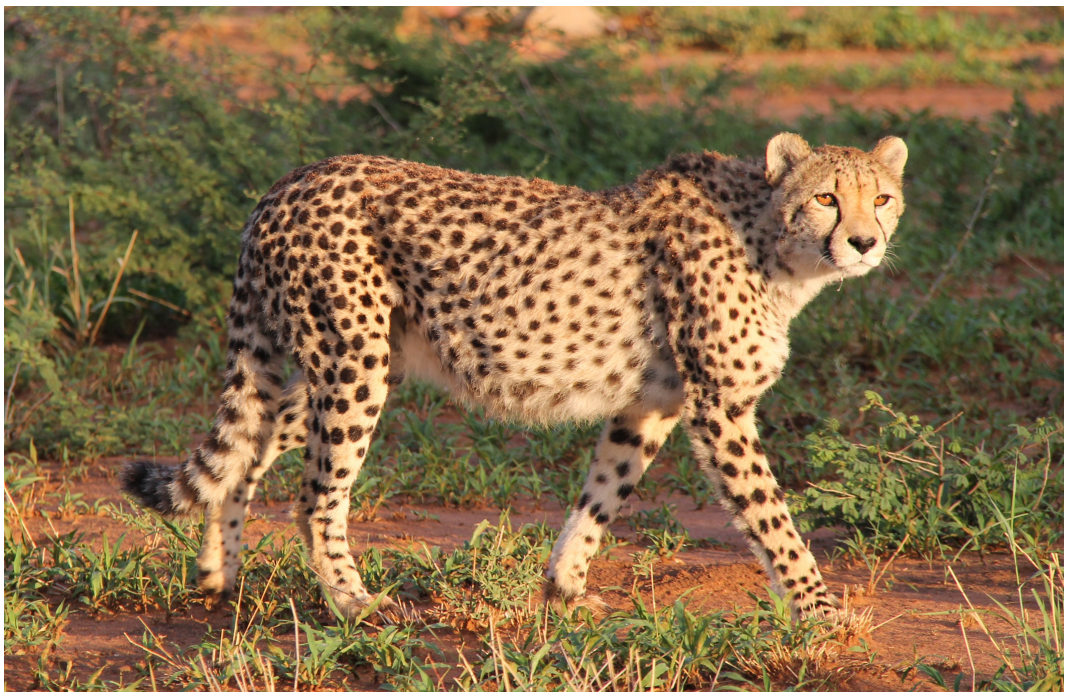
Chapter 2, "Gene pseudogenization in fertility-associated genes in cheetah (*Acinonyx jubatus*), a species with long-term low effective population size", provides an in-depth investigation into premature termination codons, a specific type of deleterious mutation, in cheetahs, linking novel mutations to previously identified symptoms of inbreeding depression.

Chapter 3, "Machine learning applications to predict functional non-coding regions in non-model species", investigates the application of ML to annotate functional non-coding regions of non-model genomes. By training a convolutional neural network on data from model species, I successfully predict regulatory regions across species, informed by sequence conservation scores, providing a framework for non-coding genomic research in non-model species.

Chapter 4, "Distribution of genome-wide deleterious variants in wild and captive cheetah populations", assesses the distribution of deleterious mutations in the loci identified in chapters 2 and 3 across multiple populations of wild and captive cheetahs. By sequencing the genomes of 32 cheetahs, I provide insight into population structure and signals of inbreeding in captive and wild cheetahs and identify segregating deleterious mutations in coding and non-coding regions.

Chapter 2

Gene pseudogenization in fertility-associated genes in cheetah *(Acinonyx jubatus)*, a species with long-term low effective population size



Photograph: Cheetah in Okonjima Nature Reserve, Namibia. Credit: Jessica Peers

The work in this chapter was completed by Jessica Peers with contribution from Dr Will Nash, who completed the variant calling. This chapter has been published in *Evolution* (DOI: https://doi.org/10.1093/evolut/qpaf005).

2.1 Abstract

We are witnessing an ongoing global biodiversity crisis, and an increasing number of mammalian populations are at risk of decline. Species that have survived severe historic bottlenecks, such as the cheetah (*Acinonyx jubatus*), exhibit symptoms of inbreeding depression including reproductive and developmental defects. Although it has long been suggested that such defects stem from an accumulation of weakly deleterious mutations, the implications of such mutations leading to gene pseudogenization has not been assessed. Here, I apply comparative analysis of eight felid genomes to better understand the impacts of deleterious mutations in the cheetah and identify novel pseudogenization events specific to the cheetah. Through careful curation, 65 genes with previously unreported premature termination codons that likely affect gene function are found. With the addition of population data (n=6), 22 of these premature termination codons are observed in at least one re-sequenced individual, four of which (DEFB116, ARL13A, CFAP119 and NT5DC4) are also observed in a more recent reference genome. Mutations within three of these genes are linked with sterility, including azoospermia, which is common in cheetahs. These results highlight the power of comparative genomic approaches for the discovery of novel causative variants in declining species.

2.2 Introduction

The ongoing biodiversity crisis is characterised by an increasing number of mammalian populations declining (G. Ceballos & Ehrlich, 2023). Lower effective population size (N_e) leads weakly deleterious alleles to behave 'nearly neutrally', seg-

regating at high frequencies and increasing genetic load (Björnerfeldt et al., 2006; Crow & Kimura, 1970; Kimura, 1983; Ohta, 1992; Wilder et al., 2023). Weakly deleterious mutations, deleterious mutations that are not affected by selection in normal conditions because their effects are too small (Loewe & Hill, 2010), can cumulatively reduce fitness and lead to genetic inviability (Pinto et al., 2024; Wilder et al., 2023), placing declining populations in a more precarious position (Lande, 1994; Lynch et al., 1993). Increased genetic load and a reduction in the effect of natural selection can lead to gene pseudogenization due to the accumulation of deleterious mutations (Casals et al., 2013; M. Kumar et al., 2023; Mathur & DeWoody, 2021; Ochman & Davalos, 2006).

In addition to missense mutations that lead to amino acid replacement, deleterious mutations can occur in the form of nonsense or frameshift mutations in protein-coding genes, potentially resulting in the loss of coding potential (loss of function) due to premature termination codons (PTCs) forming. PTCs can lead to exon skipping and decreased mRNA stability followed by nonsense mediated decay (Khajavi et al., 2006). The resultant sequences are known as pseudogenes which, following W. H. Li et al. (1981), are defined here as vestigial coding sequences that remain in the genome but, due to premature termination codons, are not successfully translated. In some rare cases, PTCs can lead to partially or fully functional truncated proteins (Aartsma-Rus et al., 2006; Nicholson & Mühlemann, 2010). However, truncated proteins are more commonly associated with pathologies such as cystic fibrosis, for which stop codon or translational readthrough has potential as a therapeutic solution (De Boeck et al., 2014; Mort et al., 2008; Schilff et al., 2021; R. S. Williams et al., 2003).

Pseudogenization is predominantly observed in genes with little contribution to fitness, where deleterious mutations can accumulate in the absence of selection (Ochman & Davalos, 2006). However, in small, inbred populations, genes with greater contributions to fitness can become pseudogenized (Ochman & Davalos, 2006). Gene pseudogenization has previously been associated with increased mutational load in species with low N_e, such as dingoes (M. Kumar et al., 2023).

Many studies have considered the impacts of deleterious variants in species with low N_e (Daetwyler et al., 2014; Dobrynin et al., 2015; Marsden et al., 2016; G. Samaha et al., 2021), elevated rate of allele loss (Masel, 2011), and increased likelihood of homozygous expression of recessive deleterious alleles (Bosse et al., 2019; D. H. Charlesworth & Charlesworth, 1987; D. Charlesworth & Willis, 2009; Mukai et al., 1972). Until recently, a lack of high-quality reference genomes has meant that conservation genomics was limited to studies of single populations (Abascal et al., 2016; Colangelo et al., 2024; Dobrynin et al., 2015; G. Samaha et al., 2021). Whilst such approaches are incredibly valuable in identifying loss-of-function mutations, monomorphic mutations are not captured, meaning pseudogenization at a species-wide level due to population decline has often been overlooked.

To address this gap, I focus on the cheetah as it is a model system for populations with low effective population size due to its long-term low N_e (Kim et al., 2016), which allows us to understand the trajectory of pseudogenization events. Whilst the pseudogenization observed in the cheetah genome is likely due to ancient demographic events causing bottlenecks rather than modern population decline, it is nevertheless a useful model to understand the impact that current anthropogenic pressures, such as climate change and habitat fragmentation, will have on populations.

The cheetah has been suggested to have exhibited a much lower N_e over the past 3 million years compared to other felids (Kim et al., 2016). Cheetahs likely experienced a genetic bottleneck around 100,000 years ago, potentially caused by migration across Asia into Africa, although cheetah paleogeographic history is still debated (R. Barnett et al., 2005; O'Brien & Johnson, 2007; O'Regan & Steininger, 2017). Additionally, cheetah populations experienced a severe demographic decline beginning approximately 10,000 years ago, suggested to be the origin of the current lack of genetic diversity of the species, although multiple demographic hypotheses have been proposed, including a severe population bottleneck or a more prolonged, gradual decline (Dobrynin et al., 2015; Driscoll et al., 2002; Fabiano et al., 2025; Menotti-Raymond & O'Brien, 1993; O'Brien et al., 1987; Terrell et al., 2016).

Finally, poaching, habitat loss, prey loss, and human-wildlife conflict all impact current cheetah populations, leading them to be classified 'Vulnerable' by the IUCN (Durant et al., 2021). Contemporary population size estimates remain low, with an estimated census size of 7,100 (Durant et al., 2017; L. L. Marker et al., 2008).

Population genetic studies have shown the cheetah to have experienced prolonged inbreeding affecting the species as a whole (O'Brien & Johnson, 2005). This led to high homozygosity, accumulation of weakly deleterious mutations, and severe inbreeding depression (Dobrynin et al., 2015). Cheetahs exhibit increased susceptibility to infectious diseases (O'Brien et al., 1985; Terio et al., 2018), developmental instability (Wayne et al., 1986), a high frequency of spermatozoal abnormalities (Wildt et al., 1983) and high juvenile mortality in the wild and in captivity (Bell, 2005; O'Brien et al., 1985). High levels of homozygosity result in a lack of diversity at the major histocompatibility complex (MHC) loci enabling viable skin grafts to be made between unrelated cheetahs (O'Brien et al., 1985), although more recent studies with increased resolution observe higher MHC diversity than previously thought (Castro-Prieto et al., 2011; Prost et al., 2022; Schwensow et al., 2019). Additionally, deleterious alleles in Toll-like receptor (TLR) genes of the innate immune system have been identified in wild African cheetahs, which result in truncated and therefore potentially functionless proteins (Meißner et al., 2024). However, no reduction in immune response due to these mutations was observed in wild Namibian cheetahs and it has been suggested that high disease susceptibility in captive populations may be due to husbandry conditions (Evermann et al., 1988; Heinrich et al., 2017; Meißner et al., 2024; Terio et al., 2018).

Putatively deleterious mutations have been identified in genes with functions related to observed abnormalities in cheetah populations, including missense mutations and gained stop codons in reproductive genes (Dobrynin et al., 2015; G. Samaha et al., 2021). Whilst Dobrynin et al. (2015) investigate acquisition of PTCs in the cheetah, the authors focus on genes with 1:1 orthologs of human reproductive genes and identify gained stop codons in eight reproduction-related

genes. Therefore, there has not yet been any study of PTCs genome-wide in the cheetah in an unbiased manner.

Here, I investigate pseudogenization in the cheetah by taking advantage of the increased availability of genomic resources to compare the reference genomes of eight felid species: cheetah (*A. jubatus*), domestic cat (*Felis catus*), black-footed cat (*F. nigripes*), lion (*Panthera leo*), jaguar (*P. onca*), leopard (*P. pardus*), tiger (*P. tigris*) and puma (*Puma concolor*) (E. E. Armstrong et al., 2020; Christmas et al., 2023; Zoonomia Consortium, 2020). After careful curation, I identify 65 genes with novel cheetah-specific PTCs. In four of these genes (ARL13A, DEFB116, CFAP119, NT5DC4) potentially involved in reproduction and susceptibility to infectious diseases, issues previously reported in cheetah populations, I find support for these novel cheetah-specific PTCs in cheetahs from several distinct breeding populations. Pseudogenization of these genes could be of interest for conservation and this finding contributes to our understanding of the effect of long-term low N_e on the genome.

2.3 Materials and Methods

2.3.1 Genomic data

Publicly available reference genomes for twelve mammal species were used in this chapter (Table S1): eight felid species and four mammal outgroups (see Figure S2.1 for divergence times). The cheetah reference genome used for the core study was aciJub1 (GCA_001443585.1; Dobrynin et al., 2015), generated from a Namibian individual.

Additionally, publicly available low coverage whole genome resequencing data from six cheetahs (Dobrynin et al., 2015) were downloaded from Genbank (SRR2737543-SRR2737545). Three of these originated from Namibia, whilst the other three were from Tanzania. A long-read cheetah assembly

(VMU_Ajub_asm_v1.0, GCA_027475565.2) derived from an individual sampled at Lisbon Zoo was also downloaded (Winter et al., 2023). This individual originates from at least 4 managed generations in captivity, however the wild origin of this individual's ancestors is not listed in the cheetah studbook (L. Marker & Johnston, 2022).

2.3.2 Gene family identification

Coding sequences (CDS) were extracted from each of ten species using annotations generated in Christmas et al. (2023) (Table S1). CDS files for human (*Homo sapiens*, GCA_000001405.15) and mouse (*Mus musculus*, GCA_000001635.2) were downloaded from Ensembl release 99 (Cunningham et al., 2022) (Table S1). The longest transcript (nucleotide length) per gene was selected for analysis (https://github.com/EarlhamInst/GSTF_snakemake). A species tree was inferred using topologies from existing literature (E. E. Armstrong et al., 2020; Piras et al., 2018; Zoonomia Consortium, 2020) (see Figure 2.1). A gene tree for each gene family was generated using *GeneSeqToFamily (GSTF)* (Thanki et al., 2018) (https://github.com/EarlhamInst/GSTF_snakemake).

2.3.3 Gene tree reconciliation

To identify putative pseudogenes in the cheetah, tree reconciliation was conducted on gene trees generated by *GSTF* using *NOTUNG* v3.0.26 (K. Chen et al., 2000), resulting in quantification of gene gains and losses at each node in the species tree. As not every gene tree contained exactly one gene per species, it was necessary to interrogate each tree with a reported loss to identify the missing gene. Each gene with a reported loss in the cheetah was further investigated to determine presence or absence in the aciJub1 annotation (GCA_001443585.1).

2.3.4 Genes present in aciJub1 annotation

Genes in the cheetah annotation which did not follow coding logic and were therefore potential pseudogenes (length not a multiple of three, contained a premature termination codon) were filtered out by *GSTF* (Thanki et al. (2018); https://github.com/EarlhamInst/GSTF_snakemake). To determine whether such genes contain biologically feasible premature termination codons (as opposed to misannotation or sequencing errors), orthologous puma (*Puma concolor*, PumCon1.0, GCA_003327715.1) sequences were used as a closely related reference.

Stringent filtering was then undertaken to identify the most biologically feasible mutations (Figure S2.2). Genes with no annotated copy in the puma were filtered out. For each remaining gene, the difference in nucleotide length between the cheetah and puma sequences was calculated and genes with a difference in length of at least 10% were filtered out. Nucleotide sequences were translated to protein (*EMBOSS* v6.6.0 *Transeq*, Rice et al., 2000) and sequence identity between cheetah and puma sequences was calculated using the Smith-Waterman algorithm (*EMBOSS* v6.6.0 *Water*, Rice et al., 2000). Genes with a sequence identity of below 90% were removed due to potential misannotation and genes with a sequence identity of 100% were removed as a PTC is expected to be one of multiple mutations accumulating in a sequence; therefore, PTCs in an otherwise identical sequence are more likely to be sequencing errors. Following translation, amino acid sequences corresponding to each transcript were used to identify PTCs. Where no PTC was identified, the transcript was filtered out.

Of the remaining genes, highly-duplicated or fast-evolving gene families (Zinc-finger proteins (ZNF), olfactory receptors (OR), family with sequence similarity (FAM), transmembrane proteins (TMEM), cyclin-dependent kinases (CDK) and cytochrome P genes (CYP)) were filtered out, as pseudogenization is expected in these gene families (Nei et al., 2008; Albà, 2017). Transcripts retired from the current ENSEMBL annotation were also filtered out at this stage.

For the remaining genes, an amino acid sequence alignment was generated by adding the corresponding aciJub1 (GCA_001443585.1) sequence to the GSTF gene family alignment and realigning (*MAFFT v.7.271*, Katoh and Standley (2013), default settings). Through manual curation of alignments, the remaining genes were classified based on whether the PTC is unique to the cheetah (versus shared with the puma lineage) and whether the PTC is due to mis-annotation or a biologically feasible frameshift or point mutation. The sequence identity of regions upstream and downstream (excluding a buffer of 30 nucleotides either side to minimise the impact of frameshifts) of putative frameshifts and point mutations was calculated using the Smith-Waterman algorithm (*EMBOSS* v6.6.0 *Water*; Rice et al. (2000)) and genes with a sequence identity of over 97% were classified as less likely to be biologically feasible. This is because a pseudogene is likely to have accumulated multiple mutations alongside the PTC (Ochman & Davalos, 2006), so if the sequence identity is high, it is not possible to differentiate the PTC-causing variant from sequencing error.

During the course of this study, a novel long-read cheetah assembly from a captive individual (VMU_Ajub_asm_v1.0, GCA_027475565.2) was published (Winter et al., 2023). CDS sequences from this genome were added to PTC candidate alignments to identify if the biologically feasible PTCs are likely to be specific to the individual sequenced for the aciJub1 reference genome or if they are potentially present across the species.

2.3.5 Genes not present in aciJub1 annotation

Putative gene losses in the cheetah may also represent annotation or assembly errors. Genes with no annotated copy in the puma were filtered out as these gene losses may have occurred earlier in the cheetah-puma lineage. The cheetah genome was searched for evidence of unannotated genes, using puma CDS sequences (*blastn*, *BLAST* v2.10, Camacho et al., 2009). *BLAST* hits were filtered to retain only those with e-value $< 10^{-6}$, or e-value $< 10^{-3}$ if the length of the

hit was $<$ 50 nucleotides. *BLAST* hits that overlapped with annotated regions of the cheetah genome were filtered out. Successful hits were reciprocally *BLAST*ed to the puma genome to ensure that they represented all exons of the gene.

Exonic sequences of genes that had not been identified in the aciJub1 assembly were then searched for individually using *BLAT* v35 (Kent, 2002) to identify genes where at least one exon was missing, a PTC was present, or part of the gene sequence overlapped with another gene, resulting in misannotation in the cheetah. Highly conserved synteny between the cheetah and puma (E. E. Armstrong et al., 2020) was exploited to determine whether the *BLAST/BLAT* hits were located in the expected region of the genome. For genes with no significant *BLAST/BLAT* hit, this synteny was utilised to reject genes likely missing due to high fragmentation of the assembly.

2.3.6 Population analyses

Low coverage whole genome resequencing data for 6 cheetahs from Namibia and Tanzania (Dobrynin et al., 2015) was downloaded from NCBI GenBank (SRR2737543-SRR2737545). Reads were trimmed using *Trimmomatic* v039 (Bolger et al., 2014) and mapped to the aciJub1 (GCA_001443585.1) reference genome using *BWA-MEM* v0.7.17 (H. Li, 2013). Mapping was followed by *SAMtools* v1.15 *fixmate* and *sort* to ensure BAM files were correctly formatted and sorted prior to removing duplicates (Danecek et al., 2021). PCR duplicates were then removed using *Picard* v2.26.2 *RemoveDuplicates* ("Picard", n.d.). Finally, mappings were filtered for complete read pairs and those with a mapping quality (MAPQ) $>$25 using *SAMtools* v1.15 *view* (see Table S2 for full settings).

Joint genotyping was conducted using *BCFtools* v1.10.2 *mpileup* (Danecek et al., 2021). *BCFtools call* was then used to call multi-allelic variants. Variants were then filtered using *BCFtools filter* and single nucleotide polymorphisms (SNPs) were extracted using *BCFtools view*. I retained SNPs which were not within 3

bp of other variants, had a variant quality score >= 30, that were at a locus with sequencing depth greater than 12 and less than 106 (+/- 3 times average sequencing depth), had a minor allele count of 3 or more, and were represented by data at that locus in more than 50% of individuals.

SNPs were then intersected with the genomic coordinates of predicted PTCs using *bedtools* v2.30.0 *intersect* (Quinlan & Hall, 2010). *bedtools* v2.30.0 *coverage* (Quinlan & Hall, 2010) was used to identify candidates without sufficient depth (at least 12 total reads) to call SNPs. SNPs that intersected predicted PTC coordinates were manually assessed to determine the prevalence of each PTC in the population data. Predicted PTC sites that were monomorphic in the population data were interrogated to determine if the site was filtered out due to low coverage or if the site was not reported as all individuals matched the reference (and therefore the PTC was identified in all individuals). For sites that passed the filtering step, I then excluded individuals with less than 3 reads at the site, as these could not be confidently called.

2.4 Results

2.4.1 Gene family identification

From 230,740 coding sequences (Table S3), 9,673 CDS were filtered out by *GSTF* as they did not follow coding logic, resulting in 221,067 CDS for which the longest nucleotide transcript was extracted (Table S3). Following an internal filtering step (Table S3), *GSTF* sorted 207,314 transcripts into 13,979 clusters. Of these, 2,742 clusters with less than 3 genes were filtered out. The remaining clusters were split then into 19,349 gene trees. 6,671 1:1 orthologs were identified across all species, and 9,010 1:1 orthologs were present specifically in Felidae. Following gene tree reconciliation, gene gains and losses were identified across the dataset (Figure 2.1).

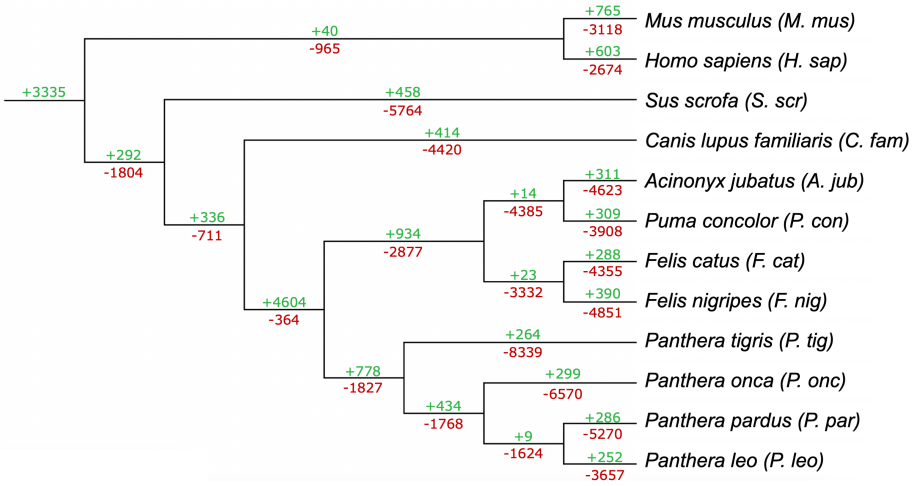


Figure 2.1: **Gene gains and losses in the Felidae.** *NOTUNG* (K. Chen et al., 2000), a tree reconciliation tool, was run on the gene trees generated by *GeneSeqToFamily* (Thanki et al., 2018). Gene gains (+) and losses (-) identified by *NOTUNG* for each branch of the tree are shown in green above the branch and red below the branch, respectively. The topology of the species tree was extracted from existing literature (E. E. Armstrong et al., 2020; Piras et al., 2018; Zoonomia Consortium, 2020).

2.4.2 Computational validation of results

4,623 gene trees with a putative gene loss in the cheetah were extracted (Table S4). Within these gene trees, 2,477 genes were absent from the expected orthogroup. 2,093 genes were subsequently characterised as putative gene losses as they were not present in any gene tree. 1,938 of these were identified as annotated aciJub1 transcripts. Following *BLAST* and *BLAT* searches, none of the remaining 155 unannotated genes were found to be pseudogenization candidates (Table S5).

2.4.3 Putative gene losses present in aciJub1 annotation

Of the 1,938 annotated transcripts not found in gene trees, following comparison to the puma genome and stringent filtering, 370 putative gene losses remained (Tables S6 & S7). Fifty-one of these genes were filtered out as they are part of highly duplicated or fast-evolving gene families (16 ZNF, 13 OR, 3 FAM, 2 TMEM,

2 CDK, 1 CYP) or did not have a known gene name annotated (11 genes), leaving 318 genes.

Of the 318 remaining pseudogenization candidates, 19 genes had a PTC in both the cheetah and puma (Table S8). Following manual assessment of nucleotide and amino acid alignments, one of these candidates (NT5DC4) was retained as a putative cheetah-specific pseudogenization event as multiple PTCs were identified. The remaining 299 genes did not have PTCs identified in the puma and were considered candidates for cheetah-specific pseudogenization (Table S9). Through careful manual curation of the coding sequence alignments, a further 88 genes were identified with a novel PTC unique to the cheetah (Table S9). Of the 89 total pseudogenization candidates, 62 were due to point mutations, 24 were due to frameshift mutations and 3 had occurrences of both (Table S10). Of these, 4 PTCs were shared between aciJub1 and VMU_Ajub_asm_v1.0 assemblies: DEFB116, ARL13A, NT5DC4 (Figure 2.2) and CFAP119.

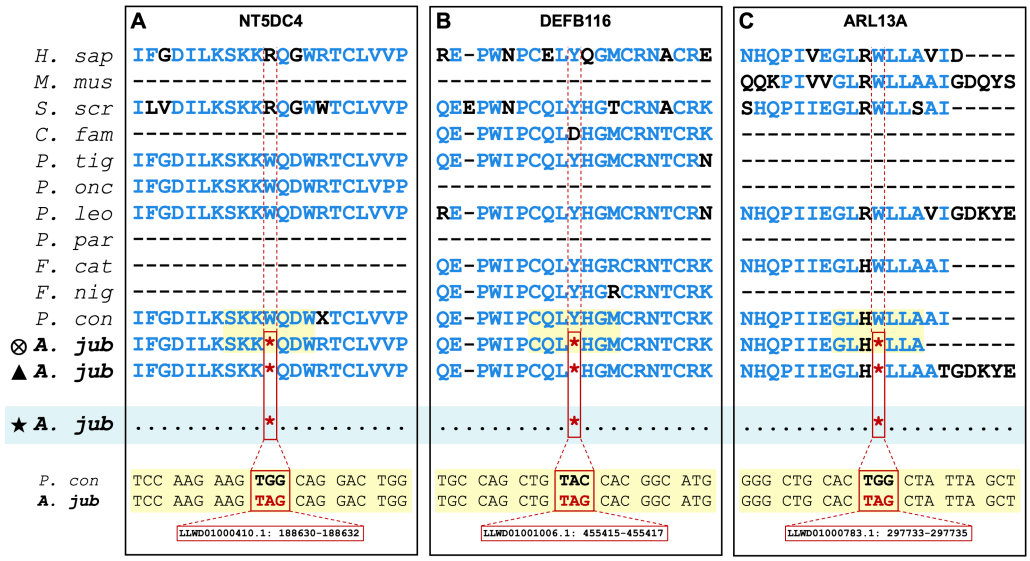


Figure 2.2: **Putative pseudogenization events in the aciJub1 cheetah genome.** Premature termination codons (PTCs) identified in *NT5DC4*, *DEFB116* and *ARL13A* in *aciJub1* (⊗), VMU_Ajub_asm_v1.0 (▲) and six African individuals (★) (Dobrynin et al., 2015). A protein alignment for each gene is shown for all species (expect those where the gene is unannotated, shown with dashes (-)). Species names correspond to those in (Figure 2.1). Conserved amino acids are shown in blue. The genomic position of each PTC is shown below the nucleotide sequence.

2.4.4 Population genomic analysis of PTCs

Population data for six individuals (Dobrynin et al., 2015) was mapped to the aciJub1 genome and variants were called. Following conservative filtering following *GATK* standard practices (Van der Auwera et al., 2013; Table S2), 3,894,283 variant sites were identified, from which 1,572,165 high confidence SNPs were retained.

None of the 24 genes with PTCs caused by frameshift mutations had sufficient population resequencing coverage to determine the prevalence of the frameshift mutation in the population (Table S10), suggesting these regions may be difficult to accurately map or assemble, so these were removed from the candidate list of novel PTCs. Of the 65 genes with PTCs caused by point mutations in aciJub1 (Table S10), 19 genes containing PTCs did not have sufficient population resequencing coverage to confidently determine the alleles in the population. A total of 22 PTCs were identified in all individuals with at least three reads (between 1 and 7 per PTC; Table S10), suggesting those PTCs are highly prevalent in those populations. These include the four previously identified pseudogenization candidates (DEFB116, ARL13A, CFAP119 and NT5DC4). The PTC identified in CFAP119 occurs upstream of one of two protein domains implicated in a previously identified pseudogenization event in Nordic Red dairy cattle (Figure 2.3, Iso-Touru et al., 2019).

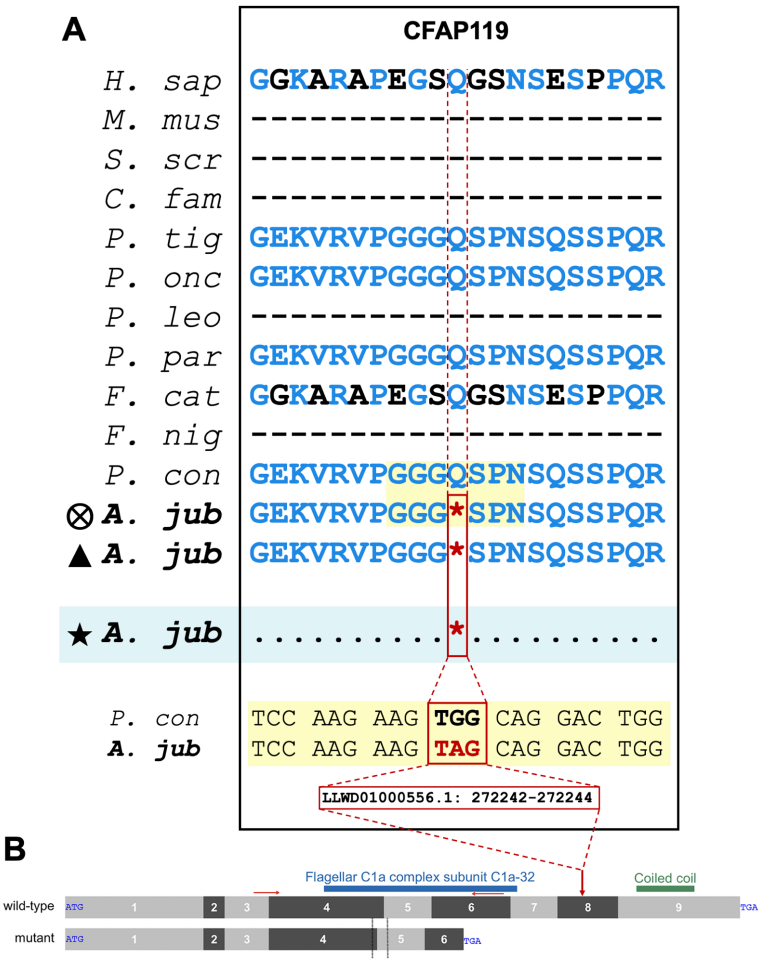


Figure 2.3: **A novel cheetah-specific premature termination codon in CFAP119 has direct similarity to a fertility relevant pseudogenization event in cattle.** (**A**) Novel cheetah-specific premature termination codon (PTCs) identified in CFAP119 in *aciJub1* (⊗), VMU_Ajub_asm_v1.0 (▲) and six African individuals (★) (Dobrynin et al., 2015). See Figure 2.2 for formatting. (**B**) A splice donor variant in CFAP119 is associated in asthenospermia in Nordic red dairy cattle (extracted from Figure 3B&C in Iso-Touru et al. (2019)). Approximate position of the cheetah PTC on the protein is shown with a red arrow.

2.5 Discussion

Historically bottlenecked populations, such as the cheetah, are characterised by low N_e and increased inbreeding. This has been shown to have led to an accumulation of deleterious mutations (Dobrynin et al., 2015), with the potential to cause gene pseudogenization. In this chapter, I compared felid genomes and identified

cheetah-specific gene pseudogenization caused by premature termination codons. Using coding sequences from eight felid species and four mammalian outgroups, I annotated gene families and applied tree reconciliation to identify cheetah-specific gene losses. Following validation of results, 65 genes with novel premature termination codons specific to the aciJub1 assembly were identified, including four genes (DEFB116, ARL13A, CFAP119, NT5DC4) with PTCs shared between eight cheetahs of wild and captive origin. These four genes have been experimentally linked to reproduction and infectious disease susceptibility in model species, which are reported issues in the cheetah.

2.5.1 Premature termination codons shared between multiple wild cheetahs

Of the 65 genes with biologically feasible premature termination codons in the original reference genome (aciJub1) caused by point mutations, 22 were fixed across resequenced samples from wild cheetahs generated by Dobrynin et al. (2015) (Table S10). Notable genes amongst these 22 include eight associated with male fertility (MOV10L1 (Fu et al., 2016); PHF7 (Cheng et al., 2023); ABHD10 (C. L. Smith & Eppig, 2009); CFAP119 (Iso-Touru et al., 2019); MARCHF6 (C. L. Smith & Eppig, 2009); MAGEB4 (Okutman et al., 2017); DEFB116 (Caballero-Campo et al., 2014; C. Zhang et al., 2018); ARL13A (Schürmann et al., 2002)). Additionally, I identify genes associated with immunity (DEFB116 (Dhople et al., 2006; Schneider et al., 2005; Schröder & Harder, 1999); ARL13A (P. Song & Perkins, 2018); IGBP1C (C. L. Smith & Eppig, 2009)); cancer (SSX5 (H. A. Smith & McNeel, 2010); TNFAIP1 (Tian et al., 2015); PRSS3 (Hockla et al., 2012); SLC38A7 (Haratake et al., 2021); HAS3 (N. Wang et al., 2022)); and developmental defects (NCDN (Fatima et al., 2021). There is no overlap between the acquired stop codons identified by Dobrynin et al. (2015) and those identified in this chapter. This is partially because Dobrynin et al. (2015) focused on reproduction-related genes, whilst I searched the full cheetah genome and applied more stringent filters.

Additionally, this work primarily focused on PTCs fixed in the species, whereas the sites reported by Dobrynin et al. (2015) showed variation within the population.

It has long been recognised that cheetah populations suffer defects in male fertility (Wildt et al., 1983), immunity (O'Brien et al., 1985; Terio et al., 2018) and development (Wayne et al., 1986), although high disease susceptibility has subsequently been linked with captive husbandry rather than genetic diversity (Terio et al., 2018). High cancer rates in wild and captive-managed felids have also been observed (Moresco et al., 2020), potentially due to the high fat and low fibre diet of the Carnivora (Chao et al., 2005; Vincze et al., 2022). Cheetahs have also been shown to have relatively high percentages of malignancy compared to other species with equivalent body mass, lifespan and litter size (Boddy et al., 2020). Although it is beyond the scope of this study to functionally validate them, the pseudogenization events I identify are in a range of genes associated with important defects such as these, and therefore may contribute to conservation-relevant traits associated with health and fitness.

All the pseudogenization events I report in this chapter are likely to be weakly deleterious as they are shared across multiple individuals, so each variant likely does not significantly impact fitness. Despite the low effective population size of this species, I still expect the most deleterious variants to have been removed through purifying selection and therefore not present in my analysis. However, I predict that although each mutation I identify may have a low fitness impact alone, they cumulatively contribute to developmental defects and disorders observed in cheetahs. It is also important to note that in rare cases, stop codon readthrough could cancel out the impact of the PTCs I identify, resulting in a functional protein; experimental validation, including proteomic analysis, is needed to determine the likelihood of this phenomenon (Loughran et al., 2014).

Gene pseudogenization has previously been linked with historic low effective population size and increased mutation load (Casals et al., 2013; M. Kumar et al., 2023; Mathur & DeWoody, 2021). Although cheetahs have been predicted to

have low N_e over a long period of time, it is beyond the scope of the data utilised in this study to date the variants described further than predicting that they either occurred in the last 3.86 – 6.92 million years since the cheetah diverged from the puma (W. E. Johnson et al., 2006) or in the 32,000 – 67,000 years since the Southeast African cheetah (*A. j. jubatus*) diverged from the Asiatic cheetah (*A. j. venaticus*) (Charruau et al., 2011).

2.5.2 PTCs shared between wild and captive cheetahs

Of the 22 genes with PTCs shared between multiple wild cheetahs, four were also identified in the chromosome-level cheetah assembly of a captive individual (VMU_Ajub_asm_v1.0, GCA_027475565.2): ARL13A, CFAP119, DEFB116 and NT5DC4. The function of NT5DC4 (5'-Nucleotidase Domain Containing 4) is not well known, although it is part of the family that catalyses the intracellular hydrolysis of nucleotides. However, the related gene NT5C2 has been associated with immunological and metabolic disorders in humans (Jordheim, 2018).

ARL13A (ADP Ribosylation Factor Like GTPase 13A) is predicted to be involved in ciliary structure and signalling (P. Song & Perkins, 2018). In mice and humans, there is evidence that defects in ciliary motility and structure may cause respiratory infections and poor fertility (Sironen et al., 2020; Tilley et al., 2015). Other genes in the ARL family have been associated with fertility; ARL4 is linked to significantly reduced sperm count in mice (Schürmann et al., 2002).

Poor fertility is a widely-reported issue in both captive and wild cheetahs; male cheetahs have low sperm concentrations and high proportions of malformed sperm (Crosier et al., 2007; Koester et al., 2015; Lindburg et al., 1993; Wildt et al., 1983). In addition to ARL13A, CFAP119 (Cilia and Flagella Associated Protein 119, also known as Coiled-Coil Domain-Containing Protein 189 (CCDC189)) is also potentially involved in male fertility. Many CCDC genes have previously been linked to male fertility in mice and humans (Tang et al., 2017; X. Zhang et al.,

2019), and a deleterious mutation in CCDC39 has previously been identified in the cheetah (G. Samaha et al., 2021).

Notably, a splice donor variant in CFAP119 was associated with asthenospermia (low sperm motility) in Nordic Red dairy cattle (Iso-Touru et al., 2019). The variant resulted in a frameshift mutation and premature translation termination, with the protein being truncated by over 40%, resulting in a lack of flagellar C1a complex subunit C1a-32, which modulates physiological movement of sperm flagella (Iso-Touru et al., 2019). A second coiled coil domain is found downstream of this, in exon 9 in cattle. The novel PTC I identify in the cheetah may therefore cause a similar asthenospermic effect, as the cheetah protein is also truncated by approximately 40% and the second coiled coil domain is likely lost. Further experimental validation is necessary to verify whether the PTC I observe in the cheetah impacts the protein in the same way.

Although the function of DEFB116 (Defensin Beta 116) is not known, other genes in the DEFB family are involved in antimicrobial defence in the skin and respiratory tract (Dhople et al., 2006; Schneider et al., 2005; Schröder & Harder, 1999) and regulation of sperm function (C. Zhang et al., 2018). Nearly all DEFB genes are preferentially expressed in the male reproductive tract and expression is enhanced during sexual maturation (Patil et al., 2005). In particular, DEFB29 is involved in sperm motility in mice and humans (Caballero-Campo et al., 2014) whilst DEFB23, DEFB26, and DEFB42 are linked with sperm motility and maturation in the epididymis of rats (C. Zhang et al., 2018).

The identification of the same PTC in more than one unrelated cheetah presents strong evidence that the mutations I identify are segregating across wild cheetah populations in Namibia and Tanzania. However, only 4 of the 22 PTCs supported by the population data are also observed in the long-read reference genome (VMU_Ajub_asm_v1.0, GCA_027475565.2). The population data I utilise (Dobrynin et al., 2015) and the original cheetah reference genome (aciJub1, GCA_001443585.1) are derived from individuals from Namibia and Tanzania, whilst

VMU_Ajub_asm_v1.0 is derived from a captive individual (Lisbon Zoo) with at least 4 generations of controlled breeding in captivity (L. Marker & Johnston, 2022). This may indicate a lower level of load in captive cheetahs than their wild counterparts, although this pattern may be due to reference genome bias and small sample size, highlighting the importance of further population resequencing. If important and functionally relevant pseudogenization events are being overlooked due to a lack of sequence data, this has the potential to limit effective design for captive breeding programmes and could potentially lead to wider fixation of deleterious mutations in this vulnerable species.

2.5.3 Limitations and considerations for future work

Because of genome assembly and annotation issues, I had to apply stringent filtering, resulting in a smaller number of high-confidence PTCs; more contiguous genomes would allow less stringent filtering and more identified PTCs. The study would have significantly benefitted from an estimation of the age of the mutations I report. However, the resources required for such analyses are not yet available for the cheetah. For instance, there are currently no high-quality genomic resources derived from museum samples or from three extant subspecies of cheetah (*A. j. venaticus, A. j. soemmeringii* and *A. j. hecki*). There is also currently no recombination map for the cheetah, which would enable an estimation of the age of these observed mutations through the estimation of shared ancestral segment lengths (Gandolfo et al., 2014). This limits inference of mutation age to the last common ancestor of the *A. j. jubatus* subspecies approximately 32,000 – 67,000 years ago (Charruau et al., 2011) or of the cheetah and the puma approximately 4.92 million years ago (W. E. Johnson et al., 2006). With exhaustive sampling of all extant subspecies and historic samples, it may be possible to derive a more accurate estimation of the age of these mutations. Additionally, experimental validation is necessary to confirm the deleterious effect of the pseudogenization events I identify. Therefore, my results highlight the importance of generating high-quality sequence data for non-model species of conservation concern.

2.6 Conclusion

The cheetah has experienced severe genetic bottlenecks and a prolonged low effective population size for millions of years. The resulting accumulation of weakly deleterious mutations across the genome is likely to have contributed to the prevalence of diseases and disorders in both captive and wild cheetahs today. The growing effort to generate high-quality reference genomes can contribute to comparative genomic investigations of species-specific gene family dynamics and pseudogenization. Here, I identified 65 genes with novel premature stop codons resulting in gene pseudogenization. Of these, at least 22 are shared in wild cheetahs and four are observed in both captive and wild cheetahs. The four genes with potentially fixed PTCs across the species are involved in fertility and immune response and may be contributing to the reproductive defects observed in cheetahs.

2.7 Supplementary material

2.7.1 Supplementary figures

Node	Divergence time (MYA)
Mammalia	91.5 – 97.4[1]
Carnivora	52.9 – 57.3[1] 55[2]
Felidae	11.3[2] 8.38 – 14.45[4] 11.6 – 14.7[1]
Panthera	3.80 – 4.31[6] 2.44 – 5.79[4]
Felinae	6.5 – 20.7[1] 5.27 – 9.20[4]
A	81.3 – 91.0[1]
B	72.8 – 78.3[1]
C	2.56 – 3.66[6]
D	3.72[5] 1.95 – 3.10[6]
E	4.85[5] 3.86 – 6.92[4]
F	3.07[3] 2.16 – 4.44[4]

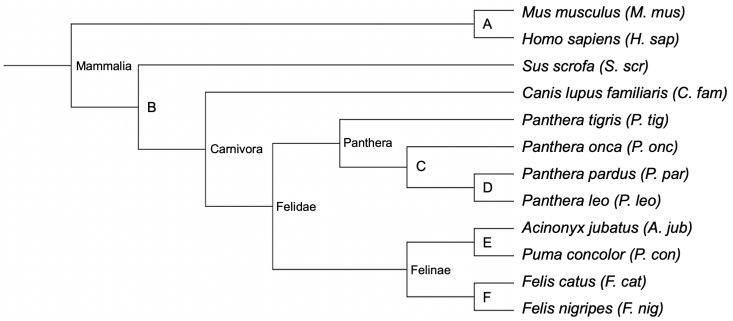


Figure S2.1. **Divergence times between the twelve mammalian species included in this study.** Values shown are median values and/or confidence intervals in millions of years as reported by the following studies and resources. 1: S. Kumar et al. (2022), 2: Yang and Yoder (2003), 3: Yuan et al. (2024), 4: W. E. Johnson et al. (2006), 5: W. Q. Zhang and Zhang (2013), 6: Davis et al. (2010). The topology of the species tree was extracted from existing literature (E. E. Armstrong et al., 2020; Piras et al., 2018; Zoonomia Consortium, 2020).

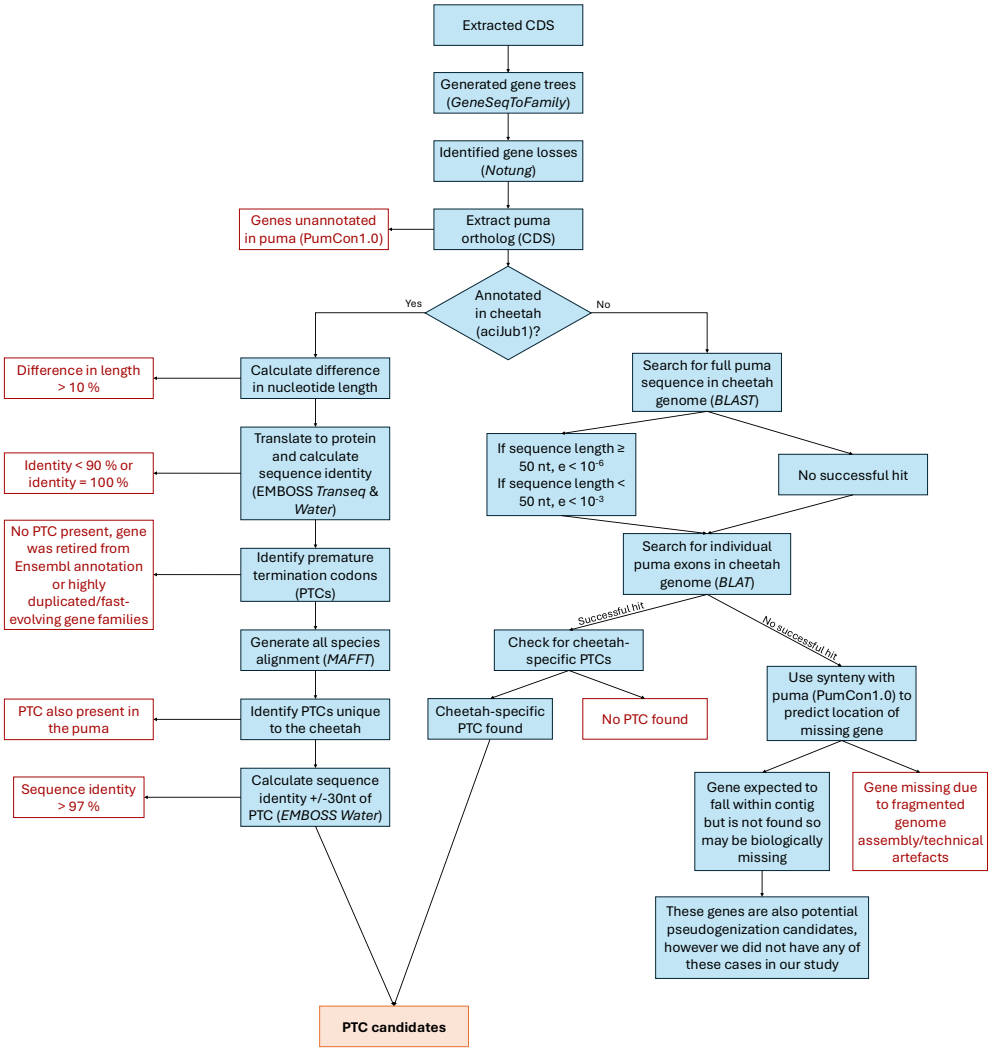


Figure S2.2. **Filtering pipeline applied to potential gene losses to identify high-confidence candidates.**

2.7.2 Supplementary tables

Supplementary tables can be found at github.com/EarlhamInst/JP_PhD.

Table S2.1. **Public genomic resources used in this study.** Genome assemblies used in this study and corresponding references. Data for all species except the lion (E. E. Armstrong et al., 2020) and VMU_Ajub_asm_v1.0 (Winter et al., 2023) were those used in the Zoonomia project (Christmas et al., 2023; Zoonomia Consortium, 2020). Note: In subsequent supplementary tables, VMU_Ajub_asm_v1.0 may be referred to as "VMU".

Table S2.2. **Filtering settings for variant calling.** Variants were called to identify prevalence of PTCs in a population of 6 cheetahs. This table describes the filtering settings that were used with SAMtools and BCFtools to filter the variants and extract high quality SNPs.

Table S2.3. **Primary GeneSeqToFamily (GSTF) results.** Coding sequences from twelve species were extracted, the longest nucleotide transcript per gene was identified and sequences that do not follow coding logic (i.e. the length is not a multiple of three, premature termination codon was identified) were filtered out in two stages, once whilst extracting the longest transcript and once during GSTF step one (Thanki et al., 2018). This table summarises the number of sequences for each species at each step of this process.

Table S2.4. **Putative gene losses.** A total of 4,623 gene trees with a loss in the cheetah were extracted, containing 5,541 genes. This table describes the status of each of these 5,541 genes. A key can be found below which explains each of the five categories: lost before GSTF, discarded by GSTF, in current file, in different file and not in annotation.

Table S2.5. **Unannotated putative losses.** Genes not in the annotation file were interrogated to determine if the gene is biologically missing or if it is a technical error (e.g. misannotation or genome fragmentation). This was done using *BLAST* and *BLAT*. Below, a key can be found which explains the three status options: present but unannotated, not present in genome and not cheetah specific.

Table S2.6. **Genes filtered out prior to GSTF.** 743 genes were filtered out during the process of extracting the longest transcript per gene (see Table S4). These genes were investigated to determine the cause for this. To determine if these genes are potential pseudogenes or are misannotations, the start codons, length of cheetah vs puma (or domestic cat) sequences, percent sequence identity and presence of premature termination codons were investigated. Genes with the same start codon, <10% difference in length, at least one stop codon and a sequence identity of 90-99.9% were carried through to further analysis. In the table below, "species" refers to the cheetah and "ref" refers to the puma or domestic cat.

Table S2.7. **Genes filtered out during GSTF.** 1196 genes were filtered out during the GeneSeqToFamily pipeline. These genes were investigated to determine the cause for this. To determine if these genes are potential pseudogenes or are misannotations, the start codons, length of cheetah vs puma (or domestic cat) sequences, percent sequence identity and presence of premature termination codons were investigated. Genes with the same start codon, <10% difference in length, at least one stop codon and a sequence identity of 90-99.9% were carried through to further analysis. In the table below, "species" refers to the cheetah and "ref" refers to the puma or domestic cat.

Table S2.8. **PTCs in multiple species.** 19 genes which had a PTC in the cheetah also had PTCs identified in other species included in the study. These genes were investigated to identify any cheetah-specific premature termination codons. A key can be found below explaining the potential statuses of each gene.

Table S2.9. **All putative novel PTCs.** Each of the genes with a potential novel cheetah-specific premature termination codon was investigated in more depth to determine if the pattern was consistent in the VMU reference genome. Percent sequence identity was also calculated upstream and downstream of the PTC to provide further information (if a PTC is the only mutation in a sequence, it is more likely to be erroneous annotation/sequencing than biologically feasible as genes with PTCs are expected to have accumulated multiple deleterious mutations).

Table S2.10. **89 genes with novel PTCs.** 89 genes were identified with premature termination codons that were considered biologically likely (see Tables S8 and S9). Population data of six African individuals (Dobrynin et al., 2015) was interrogated to determine whether each PTC was observed in the population data. Sites were either categorised as "supported in the population data" (all individuals with high quality data at the site matched the reference), "not supported in the population data" (none of the individuals with high quality data at the site matched the reference) or "filtered out" (there was not enough high quality data at the site to call a variant).

Chapter 3

Machine learning applications to predict functional non-coding regions in non-model species

Photograph: Cheetah siblings in Okonjima Nature Reserve, Namibia. Credit: Jessica Peers

The work in this chapter was completed by Jessica Peers.

3.1 Abstract

The non-coding genome harbours a wealth of variation associated with diseases and key adaptive traits, but is often overlooked when studying non-model species due to the difficulty in annotation. Experimental protocols to identify functional non-coding sequences, such as promoters and enhancers, are costly and difficult to apply to wild populations, whilst comparative genomics approaches rely on sequence conservation alone, requiring high-quality genomic resources and missing species-specific patterns. Machine learning (ML) techniques, such as convolutional neural networks, show potential for annotating the non-coding genome, but using such models to make predictions in non-model species has not been thoroughly tested. Here, I utilise publicly available data from model species to train and test multiple models to predict functional non-coding regions genome-wide and compare model performance within and across species. I find comparable precision–recall when transferring non-coding annotations between species and, when used in combination with sequence conservation information, I demonstrate accurate ML prediction of functional non-coding regions in non-model species.

3.2 Introduction

Deciphering the regulation of gene expression has potential for wide impact, spanning evolutionary biology, conservation, and health. Identifying functional regions of the non-coding genome is a crucial step towards this delivering this impact. Approximately 8-11% of mammalian genomes are evolutionarily constrained relative to neutrally evolving sequences, a signature suggestive of sequence function (Christmas et al., 2023; Rands et al., 2014). Functionally important sequences are unlikely to accumulate mutations, resulting in high levels of sequence con-

straint (hereafter termed conservation). Around 80% of significantly conserved nucleotides are located outside of protein-coding exons (Christmas et al., 2023). Despite this, our understanding of the function of the non-coding genome is currently far behind our understanding of protein coding genes. This is, in part, due to the relative ease with which coding sequences can be studied, through experimental approaches, such as cDNA screening and gene expression profiling (Adams et al., 1991; Schena et al., 1995), and validation through knockout experiments, in which gene disruption is used to hypothesis-test phenotypic changes (Hsu et al., 2014; Smithies, 1993).

The vast majority of trait- and disease-associated GWAS hits are found in the non-coding genome, such as functional variants implicated in many human cancers and autoimmune diseases (Edwards et al., 2013). Examples are also seen across model mammal species: polled intersex syndrome in goats (*Capra hircus*) is caused by a structural non-coding variant (R. Simon et al., 2020), and causal variants associated with heritable traits, such as brain size and vocal learning ability, across a range of mammals are non-coding (Kaplow et al., 2023; King & Wilson, 1975). Other examples of the functional importance of non-coding evolution include pelvic reduction in sticklebacks (*Gasterosteus aculeatus*), associated with a deletion of an enhancer of the Pitx1 gene (Chan et al., 2010), and milk fat percentage in dairy cattle (*Bos taurus*), affected by a tandem repeat variant in the DGAT1 promoter region (Kuehn et al., 2007).

The complexity of annotating the non-coding genome has led studies of genome evolution to focus on humans and other model species, as experimental assays involved in this process require high-quality samples (Ernst et al., 2025; Wiegleb et al., 2022). Whilst this focus has provided valuable insights, it also limits the understanding of non-coding genome evolution in non-model species. To understand the evolution of lineage-specific traits and adaptations across mammalian biodiversity, we must widen the investigation of non-coding genomes to non-model species. Comparative analyses of non-model species have also shown great impact in the field of human health and disease (Kaplow et al., 2023; Sullivan et al., 2023),

highlighting the additional importance of building a robust Mammalia-wide understanding of the functional non-coding genome (Gormley, 2023). This is especially prudent given the extinction crisis threatening many mammal species, as much of this biodiversity may be lost.

Our planet is currently experiencing its sixth mass extinction event, with 27% of mammal species being threatened with extinction (The IUCN Red List of Threatened Species, 2025). These taxa exhibit an overwhelming diversity of traits, with each species possessing unique adaptations that contribute to overall ecosystem diversity. As populations decline, genetic bottlenecks can occur, potentially giving rise to an accumulation of deleterious mutations, inbreeding depression, and the expression of phenotypic abnormalities (Dussex et al., 2023; Khan et al., 2021). To conserve species threatened by decreasing population sizes, contemporary conservation efforts can use population genomic methods. For example, it is possible to measure the impact of inbreeding through the calculation of runs of homozygosity (Shafer & Kardos, 2025), or to identify deleterious mutations and genetic load in protein coding genes (Hubisz et al., 2011; Schubach et al., 2024). However, without appropriate annotation of the non-coding genome, such methods lack the power to elucidate the genome-wide impacts of inbreeding on the non-coding genome. Due to their potential impact on gene expression, mutations in non-coding regions may be similarly deleterious to nonsense or missense mutations in protein-coding exons (Scacheri & Scacheri, 2015; Wells et al., 2019). Robust characterisation of non-coding variation requires accurate annotation of non-coding regulatory elements, such as promoters and enhancers (Joshi et al., 2021). In non-model species, such annotation has proven complex, resulting in a limited number of studies investigating deleterious variants in non-coding regions of non-model species, and very little information on the association between such variants and phenotypic changes (Bertorelle et al., 2022; L. M. Williams et al., 2010).

Whilst few previous studies have annotated functional non-coding regions in the genomes of non-model species, these studies have either used sequence-

conservation-based computational approaches, which are not always accurate (Huber et al., 2020), or experimental assays, which require high-quality tissue samples (Ernst et al., 2025; Wiegleb et al., 2022). Such samples are not readily available for most non-model species, especially those which are vulnerable or endangered (Zoonomia Consortium, 2020). Although both methods provide some basis for non-coding annotation, emerging developments in ML show potential for expediting the discovery of functional non-coding regions.

3.2.1 Approaches for annotating functional non-coding regions

Experimental approaches

Functional non-coding regions can be empirically identified using molecular signatures associated with regulatory activity such as chromatin state, methylation status, transcription factor (TF) binding, and histone modifications. Experimental approaches to quantify such signatures include ATAC-seq, DNAse-seq, ChIP-seq, Promoter Capture Hi-C, FAIRE-seq and CUT&Tag (Buenrostro et al., 2013; Giresi et al., 2007; D. S. Johnson et al., 2007; Kaya-Okur et al., 2019; Schoenfelder et al., 2018; L. Song & Crawford, 2010). These techniques can be used both individually and in combination to identify a range of regulatory regions (Elkon & Agami, 2017).

ATAC-seq (assay for transposase accessible chromatin sequencing) identifies regions of open chromatin, which can act as a proxy for functional non-coding elements because these accessible regions are more likely to be interacting with transcription factors and other transcription machinery. ATAC-seq requires sampling of a diverse set of tissues from the target species and the employment of complex experimental and validatory protocols (Buenrostro et al., 2013). ChIP-seq (chromatin immunoprecipitation sequencing), used to identify binding sites of regulatory proteins and histone modifications, also requires complex experimental

procedures (D. S. Johnson et al., 2007). As pre-existing antibodies against target transcription factors or histone modifications are required, the deployment of this approach in non-model species is often costly and time-consuming. DNase-seq (DNase I hypersensitive sites sequencing) uses the DNase I enzyme to identify nucleosome-depleted DNA, again as a proxy for functionality (L. Song & Crawford, 2010). FAIRE-seq (Formaldehyde-assisted isolation of regulatory elements sequencing) is also based on sequencing nucleosome-free DNA (Giresi et al., 2007) and finds regions similar to those found by DNase-seq, as well as transcription start sites and active promoters. However, FAIRE-seq is initiated *in vivo*, so requires considerable experimental resources. Hi-C, which crosslinks chromatin with formaldehyde, captures the three-dimensional conformation of DNA to detect chromatin interactions and provide spatial information (Belton et al., 2012). Promoter Capture Hi-C focuses specifically on promoter regions to identify long-range promoters (Schoenfelder et al., 2018), but also relies on large quantities of high-quality samples (Yamaguchi et al., 2021).

The data generated by such assays have the potential to be highly accurate but are intrinsically species-specific, and the assays are costly and difficult to apply to non-model species. The samples required for most of these protocols must be fresh or immediately frozen, as high quality nuclei are required, which are very sensitive to degradation (Ernst et al., 2025; Ruiz Daniels et al., 2023; Wiegleb et al., 2022). Additionally, as gene expression varies between tissues and developmental stages, representative sampling is essential to develop a comprehensive overview of gene regulation (Z. Wang et al., 2015; Wiegleb et al., 2022). For a non-model species, the process of sourcing fresh tissues, such as locating newly deceased individuals or via invasive procedures, can be prohibitively difficult and costly (Ruiz Daniels et al., 2023). The resources required to collect and process samples for non-model species, as well as lack of funding compared to model species and the justified increase in legislation, such as CITES (the Convention on International Trade in Endangered Species of Wild Fauna and Flora), may prevent use of experimental approaches in these systems. Therefore, the investigation of functional non-coding

elements in non-model species is vastly understudied.

Comparative Approaches

As sample access can be limited in non-model species, comparative genomics offers an alternative option for non-coding annotation. These methods are predominantly based on the theory that functional regions in the non-coding genome are highly conserved, showing little change in base identity over evolutionary time (Christmas et al., 2023; Sullivan et al., 2023). Methods such as GERP (Genomic Evolutionary Rate Profiling), PhyloP and PhastCons can be used to identify the genome-wide distribution of highly-conserved non-coding regions (Huber et al., 2020; Pollard et al., 2010).

GERP scores, which detect evolutionarily constrained elements by comparing the number of observed mutations to a hypothetical estimate of mutations under neutrality, have been used to identify putative functional non-coding elements across mammalian evolution (Cooper et al., 2005; Davydov et al., 2010; Huber et al., 2020). However, Huber et al. (2020) highlight the limitations of this method; high turnover of functional sequences and non-constant selection can limit the ability to identify loci under selection using GERP scores. Additionally, the authors find approximately 80 Mb of human functional non-coding sequence with low GERP scores that would be missed using standard detection thresholds, suggesting a potential lack of sensitivity in the GERP scoring method.

PhastCons, a part of the PHAST (Phylogenetic Analysis with Space/Time models) software package, identifies conserved elements using Hidden Markov Models, and is best suited for identifying long, consistently conserved regions (Hubisz et al., 2011; Siepel et al., 2005). Conserved element prediction requires a neutral model of evolution and a phylogenetic tree, against which observed substitution patterns can be compared. PhyloP, also included in the PHAST package, uses the GERP test, as well as a likelihood ratio test, a score test and a test based on distributions of substitution number, to detect site-specific conservation

or acceleration (Hubisz et al., 2011; Pollard et al., 2010). Resolution of PhyloP scores varies depending on the number of species compared; with 29 genomes, PhyloP scores can be generated at the level of several base pairs (Pollard et al., 2010). With enough high quality chromosome-level genomes (i.e. several hundred species), PhyloP scores can even be generated at single nucleotide resolution, allowing accurate identification of evolutionary constraint or rapidly evolving sites, as shown by the Zoonomia project (Christmas et al., 2023; Sullivan et al., 2023).

These methods rely on the assumption that sequence conservation directly correlates to functional conservation, but regulatory activity can be maintained despite high sequence turnover (Dermitzakis & Clark, 2002; Kellis et al., 2014; Rands et al., 2014). A further assumption is made that mutations in highly conserved sites are deleterious (Huber et al., 2020), but predicting the impact of mutations on conservation alone does not utilise functional information (Kircher et al., 2014). Relying on these signatures of purifying selection as a proxy for sequence functionality is not always accurate; identified elements may have lost or changed function in the species of interest and species-specific sequences may be missed. Another key caveat to the potential of comparative genomic prediction of non-coding regions is that tissue- or cell-type specific inference is also not possible, as these methods do not incorporate such data.

3.2.2 Machine learning approaches for functional non-coding annotation

With increasing availability of taxonomically diverse genomic resources (Darwin Tree of Life Project Consortium, 2022; Lewin et al., 2018; Zoonomia Consortium, 2020), the potential for annotation of constrained genomic regions with increasing resolution and precision has greatly increased. Recent advances in ML methodologies may allow the utilisation of experimentally validated data from well-studied organisms in the accurate prediction of functional non-coding regions in non-model species. As there is a wealth of experimental data already publicly avail-

able for model species, such as humans (*Homo sapiens*) and mice (*Mus musculus*) (ENCODE Project Consortium, 2004), this alleviates aforementioned requirements associated with experimental methods. Additionally, ML algorithms can make predictions based on more than just sequence conservation, meaning that when provided with experimentally validated data, ML predictions can be more accurate than approaches like PhyloP (Huang et al., 2017).

ML methods for the prediction or lift-over of functional non-coding regions can be comprised of several different model architectures, including Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Bidirectional Long Short-Term Memory networks (BLSTMs) and Bidirectional Encoder Representations from Transformers (BERTs) (Table 3.1). RNNs are better suited to sequence data and are based on recurrent connections where data can move in multiple directions, whilst CNNs are ideal for image data and predicting the spatial relationship between data ("ANN vs CNN vs RNN: Neural Networks Guide", n.d.). CNNs have proven more popular than RNNs as they are more powerful, despite requiring more training data (Bai et al., 2018; Dhaka et al., 2021).

Table 3.1: **Summary of machine learning and deep learning architectures commonly used to predict functional non-coding regions of genomes.** Each architecture is briefly described, with typical applications and a key review article or architecture publication for further reading.

Architecture	Best for	Brief description	Citation
Support Vector Machine (SVM)	Classification, regression, outlier detection	Finds the optimal hyperplane (line, plane or generalisation) to separate different classes of data with maximum margin	Cervantes et al. (2020)
Convolutional Neural Network (CNN)	Classification (images, sequences)	A collection of neurons in interconnected layers, which learn hierarchical representations of input features	Yamashita et al. (2018)
Recurrent Neural Network (RNN)	Handling sequential data (e.g. natural language processing, time series)	Maintains a memory of previous inputs, learning contexts of data	Mienye et al. (2024)
Bidirectional Long Short-Term Memory (BLSTM)	Context-aware sequence modelling	A type of RNN, processes the sequence forwards and backwards to capture context in both directions	Mienye et al. (2024)
Bidirectional Encoder Representations from Transformers (BERT)	Sequence representation and classification	Transformer-based model that learns contextual embeddings by considering sequences in both directions	Devlin et al. (2018)

Multiple ML approaches to annotate non-coding sequences have been developed, using various input data and ML algorithms and yielding various output information (summarised in Table 3.2). Selecting the optimal tool to meet the aims of a project can be difficult, particularly for bioinformaticians with a background in biology rather than computer science, as tool documentation varies greatly.

Table 3.2: **Overview of machine learning tools for regulatory sequence prediction and analysis.** Summarises widely used tools designed to predict enhancers, promoters, transcription factor binding, and other regulatory features using various genomic input data types (e.g., ATAC-seq, CAGE, ChIP-seq). Output type, citation, and availability of documentation or code at time of writing are also listed. Associated machine learning architecture is provided, including CNN (Convolutional Neural Network), RNN (Recurrent Neural Network), BLSTM (Bidirectional Long Short-Term Memory network), SVM (Support Vector Machine) and BERT (Bidirectional Encoder Representations from Transformers). Tools vary in accessibility and complexity, from fully documented platforms with tutorials to models with limited public code.

Tool name	**Citation**	**Input data**	**Output**	**Documentation**	**ML method**
gmk-SVM	Ghandi et al. (2014)	ChIP-seq	Regulatory sequences	Code available online	SVM
Basset	Kelley et al. (2016)	DNase-seq	Functional activity of DNA sequences	Scripts and tutorials on GitHub	CNN
DanQ	Quang and Xie (2016)	ChIP-seq, DNase-seq peaks	Non-coding function	No scripts available	CNN, RNN, BLSTM
DeepEnhancer	Min et al. (2017)	DNA sequences	Enhancers	No scripts available	CNN
Basenji	Kelley et al. (2018)	DNA sequences	Promoters, distal regulatory elements	Scripts and tutorials on GitHub	CNN
AI-TAC	Maslova et al. (2020)	ATAC-seq	Enhancers, promoters	Scripts and tutorial on GitHub	CNN
DeepMEL	Minnoye et al. (2020)	Chromatin accessibility	TFBS, enhancers	Scripts on GitHub	BLSTM
Enformer	Avsec et al. (2021)	DNA sequences	Chromatin marks, TF binding	Scripts and tutorial on GitHub	Transformer (attention-based CNN)

Continued on next page

Table 3.2 (continued)

Tool name	Citation	Input data	Output	Documentation	ML method
DNABERT	Ji et al. (2021)	DNA sequences	TF binding, enhancers, promoters	Scripts and tutorial on GitHub	BERT
TACIT	Kaplow et al. (2023)	ATAC-seq	Enhancers, association with phenotype	CNN not published, phenotype association scripts on GitHub	CNN
ExplaiNN	Novakovsky et al. (2023)	ATAC-seq, TF binding (JASPAR)	TF binding, chromatin accessibility, de novo motifs	Scripts and tutorials on GitHub	CNN
Puffin	Dudnyk et al. (2024)	CAGE, RAMPAGE, GRO/PRO-cap	Promoters	Scripts on GitHub and user-friendly browser interface with tutorials	CNN

Origin of machine learning tools for functional non-coding sequence annotation

One of the first tools to predict functional non-coding regions was gmk-SVM (Ghandi et al., 2014), based on an SVM. Here, k-mers are used to predict regulatory elements and tissue-specific enhancers. Many of the successors of this tool, such as DanQ (Quang & Xie, 2016) and Basset (Kelley et al., 2016), use it as a benchmark or are based on its theoretical approach. Since the advent of these foundational tools, CNN-based models have been increasingly used to predict regulatory activity directly from DNA sequence. Basenji (Kelley et al., 2018), successor to Basset, predicts promoters and distal regulatory elements based on sequences, making gene expression predictions which align to existing knowledge of causal variants of

eQTLs in humans. Kelley et al. (2018) applied Basenji to cross-species prediction, simultaneously training CNNs on multiple genomes, and applying them to learn sequence predictors from human and mouse data. Their approach applies models trained on mouse data to analyse human genetic variants, as their aim is to further understand human disease-associated mutations. This provides a good example of the potential for cross-species predictions, which could be applied to non-model species. More recent CNN-based tools, such as Puffin (Dudnyk et al., 2024), DeepEnhancer (Min et al., 2017) and AI-TAC (Maslova et al., 2020), extend these methods to different regulatory prediction tasks. AI-TAC also utilises deep convolutional neural networks to infer chromatin accessibility from specific cell types, establishing a hierarchy of transcription factors and enabling identification of specific interactions in cell-types. AI-TAC is trained on mouse ATAC-seq data, however the tool was transferred to human DNA and a very similar ranking of TFs was generated, further highlighting the potential of CNNs for cross-species regulatory inference.

Using concepts from both CNNs and linear models, ExplaiNN (Novakovsky et al., 2023) utilises neural additive models, which combine the accuracy of deep learning models such as CNNs with the transparency and interpretability of linear models. Therefore, the benefit of ExplaiNN over other tools is its explainability: the features and properties that contribute to the predictions of the model are clearly provided. As the layers of the CNN get deeper, it becomes more difficult to interpret the model's predictions. However, ExplaiNN performs as well as state-of-the-art CNN models whilst allowing more intuitive and straightforward interpretation. The interpretability of ExplaiNN makes it a good candidate for application to non-model organisms, where interpretability and transparency of the model's predictions is particularly important. Additionally, ExplaiNN is provided as a plug-and-play platform, making it accessible to researchers less familiar with ML methods.

Cross-species predictions

More recently, several tools have begun to investigate transferring genomic annotations between species to enable prediction in non-model species. With a focus on melanoma, Minnoye et al. (2020) trained DeepMEL on chromatin accessibility of melanoma samples from six species, although the tool is reported to be applicable to non-cancerous cell types. Additionally, DeepMEL is exploited to identify transcription factor binding sites and identify orthologous enhancers in distantly related species. The cross-species analysis performed by its authors demonstrates the potential for DeepMEL to be used on non-model species, however at the time of writing, code documentation is sparse, making the tool difficult to apply.

TACIT (Kaplow et al., 2023) associates open chromatin regions identified by ATAC-seq and combines this with phenotype information to identify putative functional non-coding regions associated with specific phenotypes in a phylogenetically-aware way. From this, Hi-C data can be used to predict the genes that these regions may be associated with. TACIT is based on a convolutional neural network that learns tissue- and cell-type specific regulatory code based on candidate enhancers from ATAC-seq. From this, predictions can also be made in species without experimental data. However, although the scripts to associate predictions with phenotypes are available on GitHub, the convolutional neural networks have not yet been published, again, preventing the application of the tool.

3.2.3 Use of machine learning to annotate non-model genomes

Here, I test the use of an ML tool, ExplaiNN, to predict functional non-coding regions across species, with the aim of annotating the non-coding genome of a non-model species. ExplaiNN was selected based on the explainability of the tool, making it more accessible and appropriate for conservation applications. Using publicly available data from model species (human, mouse and dog (*Canis lupus*

familiaris)), multiple iterations of ExplaiNN are trained and tested to determine model performance when transferring between model species. I show that a model trained on ATAC-seq data from multiple species and transferred to a species outside the training set performs equally as well as a model trained only on the species of interest.

Subsequently, I train a model using input data from all three model species and use this to annotate the functional non-coding genome of the cheetah (*Acinonyx jubatus*), a non-model conservation-priority species with sustained low effective population size (Fabiano et al., 2025; Kim et al., 2016). I then compare model predictions to regions identified by traditional comparative genomics methods and discuss the accuracy of ML compared to existing methods to annotate functional non-coding regions in non-model species.

3.3 Methods

3.3.1 Genomic and ATAC-seq data

Publicly available ATAC-seq data for three model mammal species (human, mouse and dog) was used in this study. For the human and mouse, raw ATAC-seq read files were downloaded from ENCODE (https://www.encodeproject.org/) (Table S1) (ENCODE Project Consortium, 2012; Luo et al., 2020). Dog ATAC-seq read files were downloaded from Barkbase (https://www.barkbase.org/) (Table S2) (Megquier et al., 2019). To create a multi-species dataset with equal representation from all species, tissues with data available for all three species were selected: stomach, heart and liver. For the human, mouse and dog, 29, 18 and 10 samples were downloaded, respectively (Table S1,2). The corresponding reference genome for each species was downloaded from NCBI (Table 3.3).

Table 3.3: **Reference genomes used for ATAC-seq read processing.** Latin and common names are provided for each species alongside genome assembly and NCBI RefSeq accession

Species	Common name	Genome	NCBI RefSeq
Canis lupus familiaris	Dog	ROS_Cfam_1.0	GCF_014441545.1
Mus musculus	Mouse	GRCm39	GCF_000001635.27
Homo sapiens	Human	GRCh38.p14	GCF_000001405.40

3.3.2 Processing ATAC-seq data and calling peaks

ATAC-seq data was processed using the *nf-core atacseq* pipeline v2.1.2 (P. A. Ewels et al., 2020; Patel et al., 2023) to call narrow peaks, with read lengths corresponding to the fastq files specified at 25, 50 and 75 bp for dog, human and mouse, respectively. This pipeline, built in *Nextflow* v24.04.2, follows the ENCODE project ATAC-seq processing standards (Lee et al., 2016). Raw read files were provided as input to the pipeline, which first carried out QC (*FastQC* v0.11.9 (Andrews, 2010); *Cutadapt* v3.4 (Martin, 2011)), genome alignment (*BWA-MEM* 0.7.17 (H. Li, 2013)) and duplicate removal (*Picard* v3.0.0 https://github.com/broadinstitute/picard). Alignments from multiple libraries or replicates were merged where required. Alignments were filtered (*SAMtools* v1.15.1 (Danecek et al., 2021); *BEDTools* v2.30.0 (Quinlan & Hall, 2010); *BamTools* v2.5.2 (D. W. Barnett et al., 2011)) to remove duplicate reads, unmapped reads and mitochondrial reads, among others. Alignment-level QC was carried out (*Picard* v3.0.0, https://github.com/broadinstitute/picard) and normalized BigWig files scaled to 1 million reads were generated (*BEDTools* v2.30.0 (Quinlan & Hall, 2010); *UCSC BedGraphToBigWig* v445 (Kuhn et al., 2013)). *MACS2* v2.2.7.1 (Y. Zhang et al., 2008) was used to call narrow peaks, which were annotated relative to gene features (*HOMER* v4.11 (Heinz et al., 2010)) and *BEDTools* v2.30.0 (Quinlan & Hall, 2010) was used to create consensus peaks across all samples.

Reads in consensus peaks were counted (*featureCounts* v2.0.1 (Liao et al., 2014)) and differential accessibility analysis, PCA and clustering was conducted (*R* v4.0.3; *DESeq2* v1.28.0 (Love et al., 2014)). Finally, results were presented in the form of an IGV session (*IGV* v3.8.3 (Robinson et al., 2011)) and QC was provided for the whole pipeline (*ataqv* v1.3.1 (Orchard et al., 2020); *MultiQC* v4.7 (P. Ewels et al., 2016); *R* v4.0.3).

3.3.3 ExplaiNN: testing on mouse

The mouse was selected as the focal species for the initial iterations of ExplaiNN as it had the middle quantity of available data compared to the human and dog. For each run of ExplaiNN, labelled 'positive' and 'negative' sequences were required, which I defined as follows: 'positive' sequences associated with ATAC-seq peaks, which are likely to be functional non-coding regions, are provided to the model as examples of the pattern to learn. 'Negative' sequences do not bear any ATAC-seq evidence and are unlikely to be functional. They are provided to the model as examples of non-coding sequences which do not have a known function.

Following the ExplaiNN documentation provided on GitHub (https://github.com/wassermanlab/ExplaiNN), mouse peak summits generated by the *nf-core atacseq* pipeline from chromosomes 2-39 were extended by $\pm$ 100 bp, resulting in 201 bp sequences. This set of 'positive' sequences was subsampled to 10,000 sequences using *subsample-seqs-by-gc.py* (Novakovsky et al., 2023). An equivalent number of 'negative' sequences were generated by dinucleotide shuffling of the 'positive' sequences using *BiasAway*. Training, validation and test data sets were generated using *fasta2explainn.py* (Novakovsky et al., 2023), which separated the data into 80/10/10% respectively (Khan et al., 2021; Worsley Hunt et al., 2014). An ExplaiNN model (Model 1) was trained using *train.py* (Novakovsky et al., 2023) with the BCEWithLogits criterion and the reverse complement function, which gave as output the loss values for training and validation, the trained model and a parameter file. The trained model and parameter file were provided as input to

test.py (Novakovsky et al., 2023) along with the test dataset (10% of the original dataset) and performance metrics were calculated and provided as output. The performance metrics used were area-under-receiver-operator-characteristic curve (AUC-ROC) and area-under-precision-recall curve (AUC-PR).

Subsequently, a new model (Model 2) was trained to ensure the model was learning the difference between ATAC-seq peak sequences and non-ATAC-seq peaks, rather than the difference between real DNA and dinucleotide shuffled DNA. 'Positive' sequences from the prior run were used to train Model 2. 'Negative' sequences for Model 2 were generated by defining 201 bp windows across the genome, excluding windows that overlapped with annotated features or ATAC-seq peaks using *BEDTools* v2.17.0 *intersect* (Quinlan & Hall, 2010) and then hardmasking the remaining windows by replacing all repetitive sequences with Ns. Any window containing ≥ 1 N was then removed, leaving only sequences with no repetitive regions. As there were far more 'negative' sequences than 'positives', a set of 'negative' sequences was randomly extracted using *shuf* to approximately match the number of 'positive' sequences. Additionally, 'positives' and 'negatives' from chromosome 1 were separated from the main dataset to form an additional test set. The 'positives' and a subset of the 'negative' sequences from the remaining chromosomes were then analysed with *match-seqs-by-gc.py* (Novakovsky et al., 2023) to extract 100,000 'positive' and 'negative' sequences matched by GC content. From these, train/test/validation datasets were generated (80/10/10%) and Model 2 was trained. Model 2 was then tested on the set of 'positives' and 'negatives' from chromosome 1.

Equivalent tests were also run in the human (Model 3) and dog (Model 4), as well as a combined human and mouse model (Model 5). Again, for each species, chromosome 1 was separated from the rest of the genome to act as an independent test set and the remaining chromosomes were used to generate a training, test and validation data set. Training, test, and validation sets were generated using 201 bp peak-summit-flanking sequences as 'positives' and 201 bp unannotated non-ATAC-peak genomic windows as 'negatives'. The same was done for chromosome

1 to generate a labelled independent test set. Once trained, Models 3 and 4 were run on their corresponding chromosome 1 test set. For the combined human and mouse model (Model 5), the existing test sets, generated as described above, were concatenated to ensure equal representation from each species.

3.3.4 ExplaiNN: transferring between species

After calculating model performance in scenarios where both training and testing data were derived from the same species, models trained on one species were then tested on another species. Each of the single species models (mouse (Model 2), human (Model 3) and dog (Model 4)) were tested against the chromosome 1 test set of the other two species and model performance was computed. The human and mouse combined model (Model 5) was tested on the human and mouse combined chromosome 1 test set and the dog chromosome 1 test set.

Thus far, ExplaiNN models 1-5 made predictions on 201 bp sequences which were either 'positive' ATAC-seq peaks or 'negative' non-coding, non-repetitive, non-ATAC-seq windows. To test model performance in a non-model species genome, where 'positives' and 'negatives' are unknown, it was necessary to investigate model performance when predicting on genome-derived, unlabelled, genomic windows.

A new chromosome 1 test data set for each species was generated by calling 201 bp windows with 151 bp overlap across the whole chromosome using *SeqKit* v0.10.0 (Shen et al., 2016). Genomic windows were labelled as 'positives' if they had at least 75% overlap with one of the 201 bp ATAC-seq peak-summit-flanking sequences previously used as positives, to ensure each of these sequences was only labelled as a positive once. 'Negative' genomic windows were those with no overlap to any ATAC-seq peaks, no repetitive regions and no overlap with coding regions. To assess model performance, AUC-PR and AUC-ROC were calculated. Due to the severe imbalance in the test datasets (far more 'negatives' than 'positives'), AUC-

PR and AUC-ROC may not be the most informative metrics (Jeni et al., 2013; Movahedi et al., 2023). Therefore, precision, recall and F1 score were calculated (Jeni et al., 2013) and the number of true and false positives and negatives at different classification thresholds was quantified in the dog, as it had the least available data and thus acted as a 'test' non-model species.

3.3.5 ExplaiNN: relationship between ATAC-seq peaks and sequence conservation/distance to gene

Methods of differentiating between true- and false-positive predictions were explored with the aim of increasing confidence in the predicted functional regions in the cheetah. Using per-base PhyloP conservation scores in the dog genome, provided by Dr Michael Dong and Prof Kerstin Lindblad-Toh (Uppsala University, pers comm, 2025), the relationship between sequence conservation and ATAC-seq peaks was investigated. Using the UCSC chain file (https://hgdownload.soe.ucsc.edu/goldenPath/canFam4/liftOver/) and *liftOver* (*KentTools* v1) (Kuhn et al., 2013), PhyloP scores were converted to from canFam4 (GCA_011100685.1) to ROS_Cfam_1.0 (GCA_014441545.1) (the reference genome used in the rest of this study). Predictions from the dog-dog model (Model 4) were classified based on both ExplaiNN prediction score and positive/negative label (generated using aforementioned methods). The four classification categories were: true positives (TP; labelled as positives, ExplaiNN prediction $\geq$ 0.5), false positives (FP; labelled as negatives, ExplaiNN prediction $\geq$ 0.5), true negatives (TN; labelled as negatives, ExplaiNN prediction $<$0.5) and false negatives (FN; labelled as positives, ExplaiNN prediction $<$0.5). The median PhyloP score for each 201 bp window was calculated and medians plotted for each of the four classification categories. Additionally, the distance from the end of each window to the nearest downstream coding sequence (CDS) was calculated using *BEDTools* v2.31.0 *closest* (Quinlan & Hall, 2010) and plotted for each of the four classification categories. Pairwise Mann-Whitney U tests were run on each pair of

categories.

Additional investigation of the filters created by ExplaiNN was also conducted to identify those that were most important in making its predictions. A filter refers to a 19 bp DNA motif generated by ExplaiNN as a predictive feature. For each of the three species, *interpret.py* (Novakovsky et al., 2023) was deployed on the chromosome 2+ model (Models 2, 3 and 4) using the corresponding test dataset (a subset of the chromosome 2+ set), which outputs position weight matrices (PWM) representing the top 100 19 bp filters in MEME format. This was followed by *meme2logo.py* to turn each PWM into an image. The JASPAR 2024 Vertebrate Core Non-redundant position frequency matrices of transcription factor motifs were downloaded in MEME format from the JASPAR database (https://jaspar.elixir.no/downloads/). *tomtom.py* was then used to identify any of the ExplaiNN filters that significantly matched (q-value < 0.05) a JASPAR motif. For each model, the predictive importance of each filter that matched a TF-motif was extracted from ExplaiNN *interpret.py* output and a heatmap of these values was plotted using Python package *seaborn* (Waskom, 2021).

3.3.6 ExplaiNN: running on cheetah and comparing to sequence conservation

A final ExplaiNN model (Model 6) was trained on a total of 100,000 positive and negative sequences taken from the human, mouse and dog datasets, with equal representation from each species. The cheetah genome (VMU_Ajub_asm_v1.0, GCA_027475565.2, Winter et al. (2023)) was split into 201 bp sliding windows with a 151 bp overlap. The trained Model 6 was then deployed on these cheetah windows to predict functional non-coding regions.

To compare model predictions with ultra-conserved elements, conservation scores were calculated. A genome alignment of thirteen felids and four mammalian outgroups (Table S3) was generated using *Cactus* v2.9.8-gpu (J. Armstrong

et al., 2020). Genomes were selected based on chromosome-level of assembly, corresponding RefSeq annotation, and to ensure even taxonomic sampling across Felidae. From the HAL file generated by *Cactus*, the ancestral genome (Anc00) was extracted using *HAL* v2.2 *hal2fasta* (Hickey et al., 2013). Using *SeqKit* v2.1 (Shen et al., 2016), the ancestral genome was unmasked and re-masked using *RepeatMasker* v4.0.8 (Smit et al., 2013) to identify ancestral repeats. Ancestral repeats were extracted from the cheetah genome using *HAL* v2.2 *Liftover* (Hickey et al., 2013) and filtered to remove any complex repeat regions or coding regions. A MAF file was generated using *Cactus* v2.9.8-gpu *hal2maf* (J. Armstrong et al., 2020) with the cheetah as a reference. From this, a neutral model was generated using *phyloFit* v1.4 (Hubisz et al., 2011) with a species tree extracted from existing literature (Table S3; G. A. Samaha (2021) and Yu et al. (2025)). Using *KentUtils* v1.0 *mafSplit* (Kuhn et al., 2013), the MAF file was separated into chromosomes. *BEDTools* v2.30.0 *makewindows* (Quinlan & Hall, 2010) was used to bin the cheetah genome into 25 bp windows. The PhyloP score for each 25 bp window of the 19 chromosomal scaffolds of the cheetah genome was generated using the neutral model and *PhyloP* v1.4 (Hubisz et al., 2011) with the CONACC mode (tests for conservation and acceleration relative to neutral model) and LRT method (uses a likelihood ratio test to compute significance score). The majority of smaller contigs did not have sufficient alignment to calculate conservation scores so all non-chromosome contigs were discarded at this stage. The proportion of the genome in chromosomes was calculated to ensure only a small fraction was in these scaffolds. The mean conservation score for each predicted functional non-coding region in the chromosomes was calculated using the PhyloP scores. The neutral model generated by *phyloFit* was also provided to *PhastCons* v1.6 (Siepel et al., 2005) with the "–most-conserved" flag to identify ultra-conserved elements (UCEs).

3.3.7 ExplaiNN: filtering positive windows

Based on testing in the dog, several filtering steps were run on the cheetah genomic windows predicted as positives by ExplaiNN Model 6. Firstly, windows with any overlap to genes were filtered out using *BEDTools* v2.31.0 *intersect* (Quinlan & Hall, 2010) to retain only non-coding windows. Next, windows with an ExplaiNN prediction score ≥ 0.75 were extracted to retain only those windows with confident positive predictions. The distance from each window to the nearest downstream gene start site was calculated using *BEDtools* v2.31.0 (Quinlan & Hall, 2010) and windows were filtered to retain only those within 50,000 bp of a gene start site, as these regions are more likely to contain promoters or enhancers (Elango & Yi, 2011; Haberle & Stark, 2018; Symmons & Spitz, 2013; Vermunt et al., 2019; M. Q. Zhang, 1998; M. Q. Zhang, 2007). Finally, remaining windows were filtered based on mean PhyloP score to keep only windows with a mean score ≥ 2, suggesting significant evolutionary constraint compared to the neutral model (Christmas et al., 2023; Vy et al., 2021).

To compare the resultant set of predicted functional non-coding regions to alternative methods of annotating the non-coding genome, UCEs were extracted from the MAF file using *PhastCons* v1.6 (Siepel et al., 2005). A conservative set of high-confidence genomic windows with likelihood to be functional non-coding regions was generated by filtering based on ExplaiNN score (≥ 0.75), distance to nearest downstream gene ($\leq$ 50kb) and a measure of sequence conservation (either PhyloP score or overlap with a UCE). Additionally, *FIMO* (*MEME suite* v5.1.0) (Grant et al., 2011) was used with default settings to scan the resultant set of windows for transcription factor binding sites using the JASPAR 2024 Vertebrate Core Non-redundant motif database (https://jaspar.elixir.no/downloads/).

3.4 Results

3.4.1 Processing ATAC-seq data and calling peaks: dog

Input reads per sample ranged from 22,763,816 to 96,521,458, with 18.3 to 64.2% of these marked as duplicates (Figure S3.1). Per-sequence GC content showed a roughly normal distribution, suggesting a normal random library (Figure S3.2). All reads passed initial filtering steps. After adapter trimming, 20,624,461 to 70,507,822 reads remained. Following alignment, only 6.6 to 10.4% of reads were mapped. After alignment filtering, 1,805,110 to 5,346,126 mapped reads per sample remained. Between 623 and 5,084 narrow peaks were called per sample (Figure S3.3). FRiP (Fraction of Reads in Peaks) scores ranged from 0.05 to 0.24. Filtered narrow peaks were subsequently annotated relative to gene features and a periodicity plot was generated by Picard to show insert sizes (Figure S3.4).

3.4.2 Processing ATAC-seq data and calling peaks: human

Input reads ranged from 13,593,438 to 375,684,903, with 10.9 to 73.0% marked as duplicates (Figure S3.5). Per-sequence GC content showed a roughly normal distribution for about half of the samples, suggesting a normal random library (Figure S3.8). All reads passed initial filtering steps. 13,585,130 to 375,452,793 reads remained after adapter trimming. Following alignment, 0.2 to 6.7% of reads remained unmapped. Between 35,572,576 and 730,577,690 reads per sample remained after merging of libraries and replicates and alignment filtering. Between 37,992 and 340,687 narrow peaks per sample were called (Figure S3.7). FRiP scores ranged from 0.05 to 0.24, peaks were annotated relative to gene features, and a periodicity plot was generated (Figure S3.8).

3.4.3 Processing ATAC-seq data and calling peaks: mouse

Input reads ranged from 15,591,507 to 75,813,558, with 14.6 to 50% of these marked as duplicates (Figure S3.9). Per-sequence GC content showed a roughly normal distribution and all reads passed initial filtering steps (Figure S3.10). After adapter trimming, 15,523,342 to 75,552,933 reads remained. Following alignment, 0.9 to 4.0% of reads remained unmapped. After merging of libraries and replicates for each sample and alignment filtering, 50,713,172 to 173,331,152 mapped reads remained per sample. Between 17,603 and 70,592 narrow peaks were called per sample (Figure S3.11). FRiP scores ranged from 0.10 to 0.42. Peaks were subsequently annotated relative to gene features and a periodicity plot was generated (Figure S3.12).

3.4.4 ExplaiNN: testing on mouse

Performance metrics for Model 1 (trained on 201 bp sequences corresponding to extended ATAC-seq peak summits ('positives') and dinucleotide-shuffled positives ('negatives')) were calculated using the test dataset: AUC-ROC = 0.9160 and AUC-PR = 0.9209. Model 2 (trained on a similar set of positives and genomic windows with no overlap to ATAC-seq peaks (negatives)) was trained on all chromosomes except chromosome 1. When tested on the test dataset, AUC-ROC = 0.9002 and AUC-PR = 0.8993. Model 2 was then run on the test data derived from chromosome 1 and scored AUC-ROC = 0.9035 and AUC-PR = 0.9015.

3.4.5 ExplaiNN: predicting across species

Models 2, 3 and 4 were trained for each species on all chromosomes except chromosome 1. Each model was then tested on chromosome 1 of all three species. Additionally, a multi-species model (Model 5) was trained on combined human and mouse data and tested on combined human and mouse data and on dog data.

AUC-ROC and AUC-PR were calculated for each run (Figure 3.1).

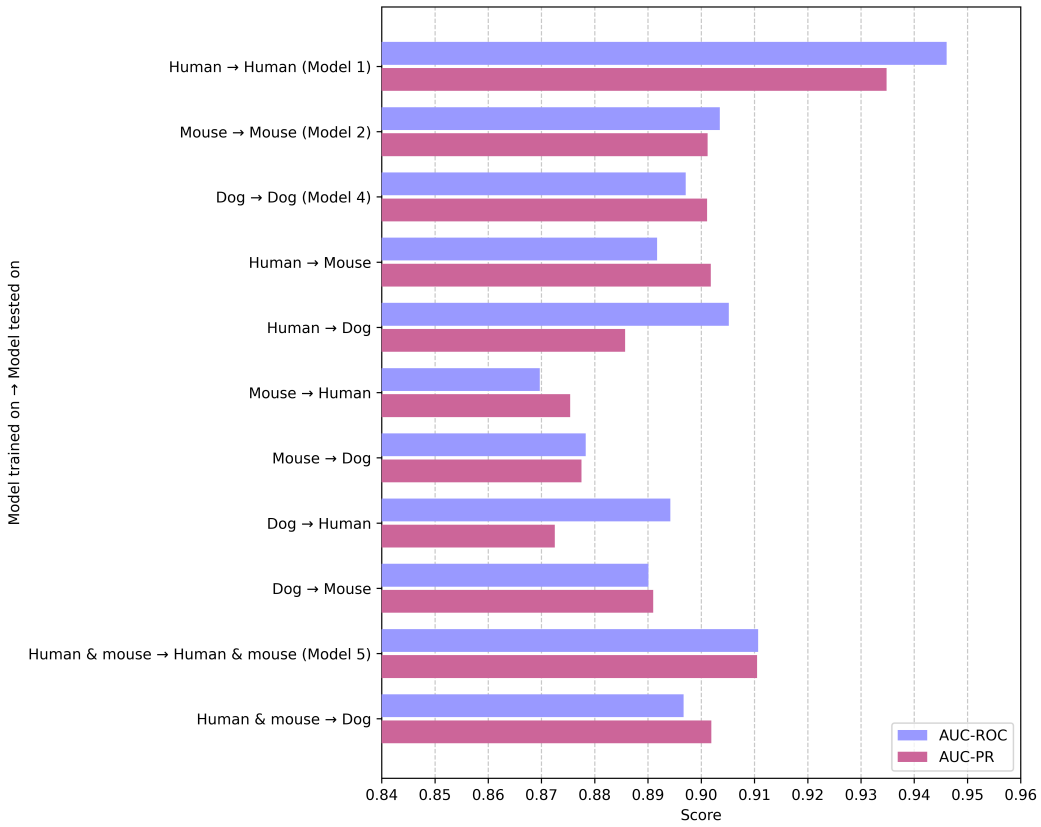


Figure 3.1: **ExplaiNN model performance for within and between species non-coding sequence prediction evaluated using AUC-ROC and AUC-PR.** Each bar represents the performance of a model trained on chromosomes 2+ from one species (or a combination of species) and tested on chromosome 1 of another species. Area under receiver operator curve (AUC-ROC, blue) and area under precision recall curve (AUC-PR, pink) are displayed for each model. Intra-species predictions (e.g. human → human) generally outperform cross-species predictions (e.g. human → mouse), but combined training (e.g. human & mouse → dog) improves generalization compared to single-species training.

When running ExplaiNN on genome-wide windows rather than balanced datasets, the resulting AUC values collapsed, indicating that the performance metrics were affected by class imbalance. For each of the three species (human, mouse, dog), AUC-ROC and AUC-PR scores were as follows: 0.8692 & 0.5088, 0.8410 & 0.1875, 0.8096 & 0.1368. As ATAC-seq peaks are relatively rare given the number of windows generated, far more negatives than positives were present (e.g. 14,738 positives versus 1,088,657 negatives in dog), meaning calculating AUC-PR and AUC-ROC was no longer sufficient to measure model performance. Even at very

high ExplaiNN prediction scores, predicted windows in the dog were dominated by negatives due to the severe imbalance of the dataset (Figure 3.2). A similar effect was seen in human and mouse. Precision, recall and F1 score (the harmonic mean of precision and recall), which are better suited to imbalanced data, were calculated and the number of true and false positives and negatives was quantified for each model (Models 2, 3 and 4; Table 3.4).

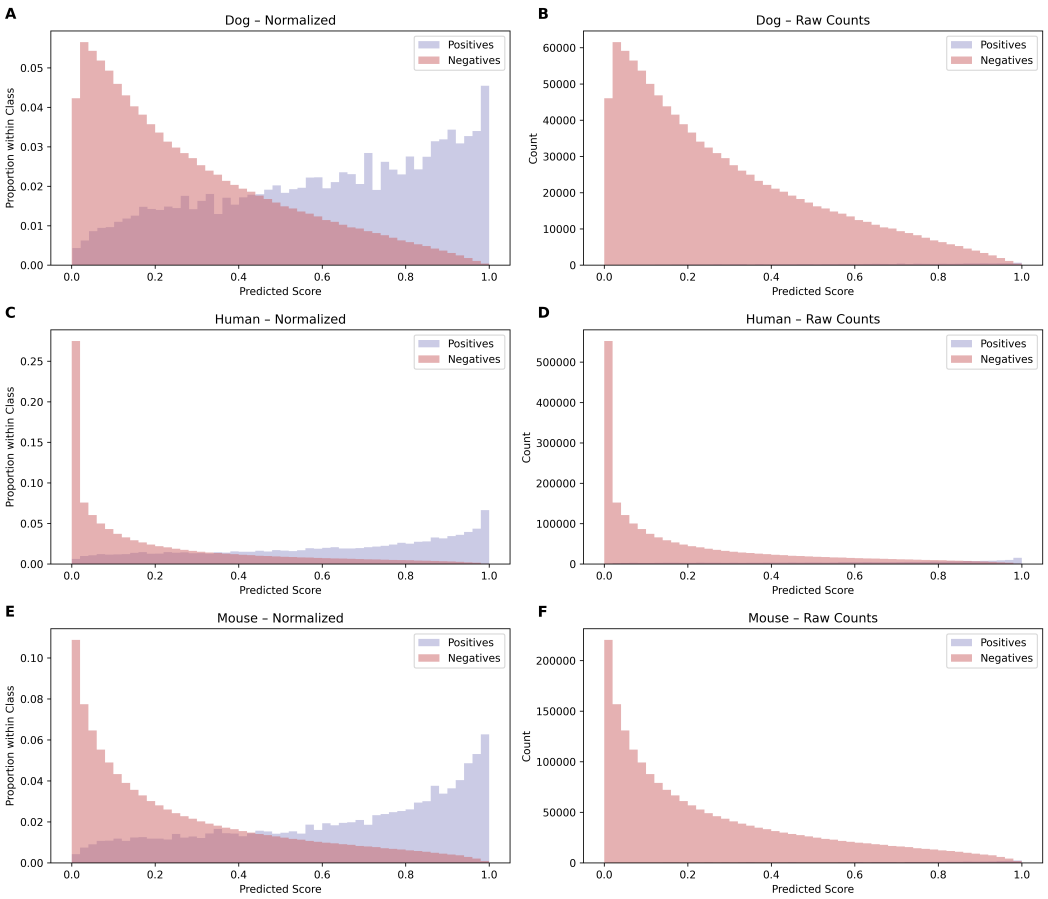


Figure 3.2: **Distribution of model-predicted scores for 201 bp windows across chromosome 1 of the dog (A & B), human (C & D) and mouse (E & F) genome. A, C, E:** Normalized distribution of predicted scores, showing the proportion of windows within each class (positive or negative). This view corrects for the class imbalance between positives and negatives, revealing clearer separation in predicted score profiles. **B, D, F:** Raw count distribution of prediction scores without normalization, highlighting the large number of negative windows relative to positives.

Table 3.4: **Precision, recall, F1 score, and confusion matrix values for ExplaiNN predictions on genome-wide 201 bp windows.** To better evaluate model performance under extreme class imbalance (see Figures 2 and 3), standard AUC metrics were supplemented with precision, recall, and F1 scores. Results are shown for models trained and tested on each species (dog, human, mouse). True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are also reported.

Species	Precision	Recall	F1	TP	FP	TN	FN
Dog	0.04	0.65	0.08	5910	206898	881746	4190
Human	0.37	0.67	0.47	153220	264240	1744763	76850
Mouse	0.07	0.69	0.13	26757	362221	1663723	11950

3.4.6 ExplaiNN: reducing number of false positives

Due to the difficulty of discerning true and false positives, ExplaiNN prediction score, conservation score and distance to nearest coding region were used to filter false-positives in dog. Prior to any filtering of predictions, ExplaiNN had labelled 5,910 true positives and 206,898 false positives – a ratio of 1:35. To account for the overrepresentation of false positives introduced by class imbalance, the distribution of ExplaiNN prediction scores across the four categories (TP, FP, TN, FN) was examined. Based on pairwise Mann-Whitney U testing, significant differences ($p < 0.05$) between the distributions of ExplaiNN score for TP and FP and for TN and FN were identified, as well as all other combinations of pairs of categories (Figure 3.3). When filtering the set of TP and FP for ExplaiNN scores ≥ 0.75, 3,162 TP and 54,038 FP were identified, a ratio of 1:17.

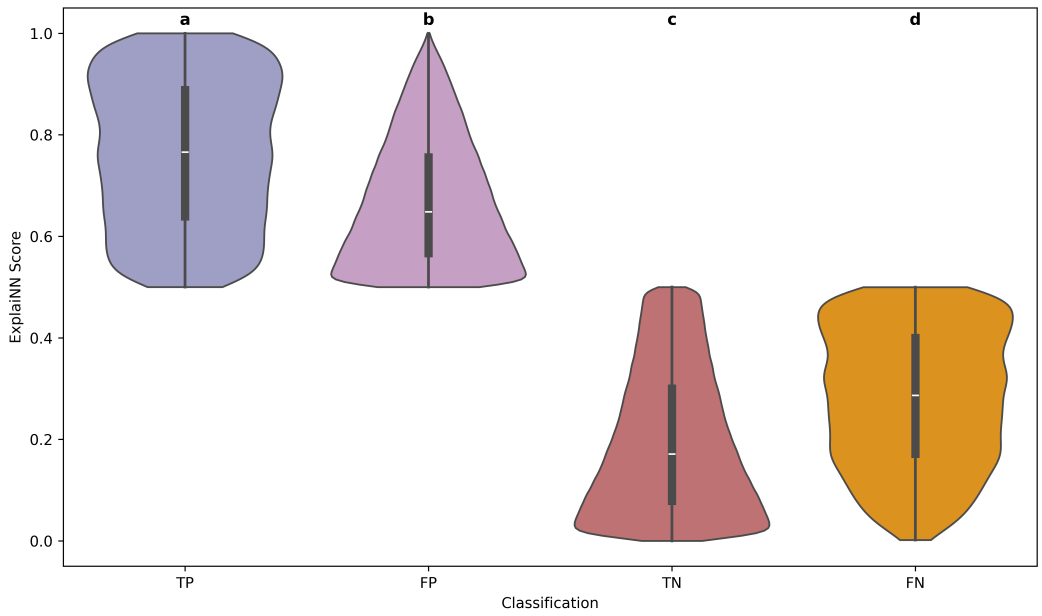


Figure 3.3: **Distribution of ExplaiNN prediction scores by classification category.** Violin plots show the distribution of prediction scores for each of the four classes: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). True positives and false positives had significantly higher scores than true and false negatives, respectively. Statistical significance was assessed using the Mann–Whitney U test, with all pairwise comparisons showing significance (indicated by a, b, c, d annotations). Mann-Whitney U scores and p-values were as follows: TP vs FP: U = 835,736,486.5, $p < 0.001$; TP vs TN: U = 5,211,118,860.0, $p < 0.001$; TP vs FN: U = 24,762,900.0, $p < 0.001$; FP vs TN: U = 182,431,483,908.0, $p < 0.001$; FP vs FN: U = 866,902,620.0, $p < 0.001$; TN vs FN: U = 1,207,370,575.5, $p < 0.001$.

When comparing distance to gene start site for each of the categories, a significant difference (Mann-Whitney U, $p < 0.05$) was found when comparing all categories (Figure 3.4). When filtering the raw set of TP and FP for a distance to nearest gene start site of $\leq$ 50,000 bp, 4,723 TP and 136,203 FP were observed, a ratio of 1:29.

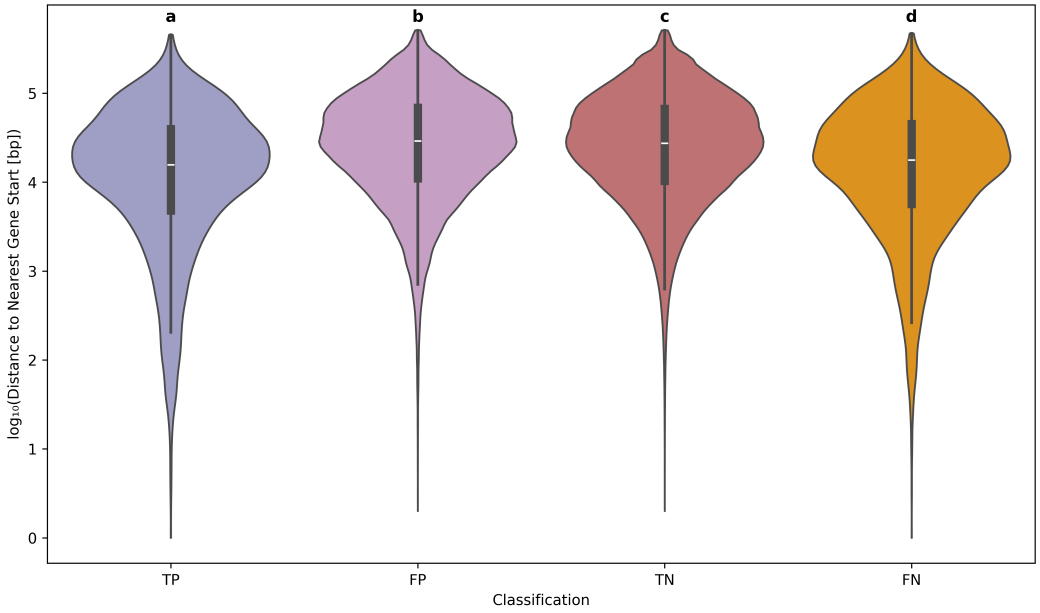


Figure 3.4: **Log-transformed distance to the nearest downstream gene for each classification category.** Violin plots show the distribution of distances (log_{10}-transformed) from the end of each 201 bp window to the nearest downstream gene, stratified by classification outcome: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Statistical significance was assessed using the Mann–Whitney U test, with all pairwise comparisons showing significance (indicated by a, b, c, d annotations). Mann-Whitney U scores and p-values were as follows: TP vs FP: U = 458445952.0, $p = 1.43 \times 10^{-236}$; TP vs TN: U = 2000798092.0, $p = 2.36 \times 10^{-208}$; TP vs FN: U = 11727248.0, $p = 5.86 \times 10^{-6}$; FP vs TN: U = 92956966276.5, $p = 1.10 \times 10^{-41}$; FP vs FN: U = 520019286.0, $p = 7.07 \times 10^{-109}$; TN vs FN: U = 2182495071.5, $p = 1.36 \times 10^{-91}$.

When comparing median PhyloP score for each category (Figure 3.5), a statistically significant difference was found between true and false positives (Mann-Whitney U, $p<0.05$) but not between true and false negatives. When filtering the set of true and false positives for a PhyloP score ≥ 2, 277 TP and 2,824 FP were observed, resulting in a ratio of 1:10.

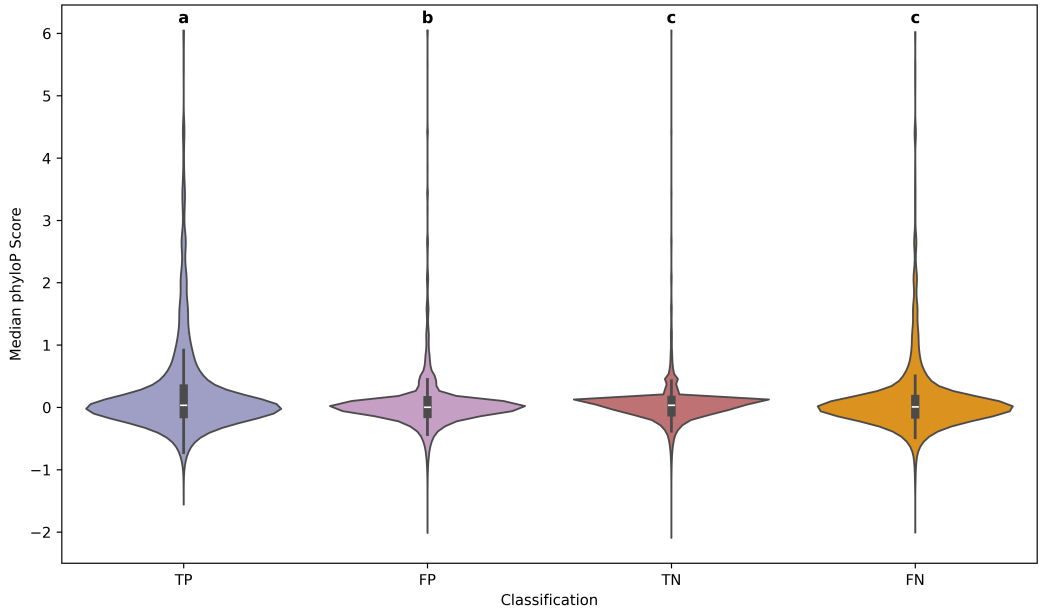


Figure 3.5: **Median PhyloP score per window across classification categories.** Violin plots show the distribution of median PhyloP conservation scores for each classification type: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Mann–Whitney U tests indicated significant differences between most category pairs, with the exception of TN vs FN (p = 0.193), as indicated by labels a, b, c, d. Mann-Whitney U and p values are as follows: TP vs FP: U = 677629912.0, $p = 1.23 \times 10^{-53}$; TP vs TN: U = 2772366354.5, $p = 4.17 \times 10^{-24}$; TP vs FN: U = 13208499.0, $p = 2.49 \times 10^{-11}$; FP vs TN: U = 83912000029.0, p < 0.001; FP vs FN: U = 410101715.5, $p = 1.24 \times 10^{-5}$; TN vs FN: U = 1834978705.5, $p = 1.93 \times 10^{-1}$.

Finally, transcription factor binding motifs were investigated. ExplaiNN filters that significantly matched a known TF motif in the JASPAR database were extracted. For the dog, of the 100 most informative filters, only 15 matched known transcription factor motifs (Table S4). For human and mouse, 17 and 19 filters matched known TF motifs, respectively. The importance of each of these TF-matching-filters was plotted per species (Figure 3.6, Table S4).

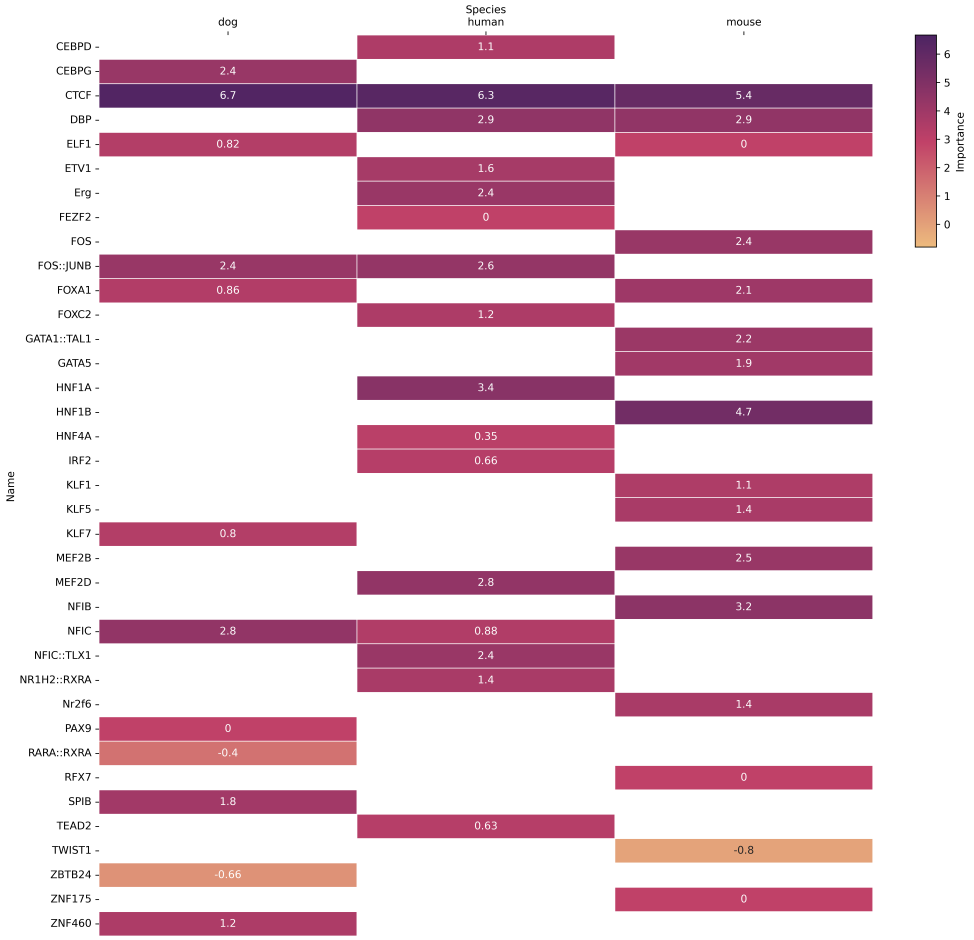


Figure 3.6: **Motif importance by species for ExplaiNN predictions.** The relative importance scores of transcription factor (TF) motif-like filters learned by ExplaiNN for dog, human, and mouse models is shown. Each row represents a distinct motif, and each column corresponds to a species. Where a motif filter appeared multiple times within a species, the highest importance score was retained. Higher scores indicate greater contribution to model predictions. White signifies no occurrences of the motif in that species.

3.4.7 ExplaiNN: running on cheetah and comparing to sequence conservation

When ExplaiNN Model 6 (trained on human, mouse and dog) was applied to the cheetah genome, 10,887,913 windows were predicted to be 'positive' (prediction score ≥ 0.5) and 36,659,747 windows were predicted to be 'negative' (prediction

score < 0.5). The total set of windows were filtered to retain those outside of protein-coding genes, leaving 44,330,717 windows. Windows were then filtered to retain those with ExplaiNN prediction scores ≥ 0.75, leaving 3,194,395 positive predictions. 6,263 windows found on non-chromosomal contigs were removed due to poor alignment quality, leaving only windows located in chromosomal scaffolds, which encompass 99.72% of the reference assembly. Remaining windows more than 50,000 bp from the nearest downstream gene start site were filtered out, leaving 1,767,739 windows. Of these, only 542 had a mean PhyloP score ≥ 2, all of which overlapped a UCE. Therefore, the intersect of the 1,767,739 positive ExplaiNN windows and UCEs was extracted, resulting in 292,912 remaining windows. As filtering for PhyloP score was too stringent for my data, I used UCE-intersect as a conservation filter instead, taking this set as the 'final' set of predicted functional non-coding regions.

A skewed distribution of distances to the nearest gene start site was observed (Figure 3.7), with the highest density occurring within approximately 2–5 kb upstream of the gene, and the majority of predicted functional windows located within 10–15 kb. When FIMO was run, 94,945 of the 292,912 windows had at least one significant hit for a TF binding motif. The top 20 most common TF binding motifs identified were extracted (Figure 3.8) and compared to the motifs learned by ExplaiNN in Figure 3.6: three were exact matches for ExplaiNN motifs and nine were within the same family as ExplaiNN motifs.

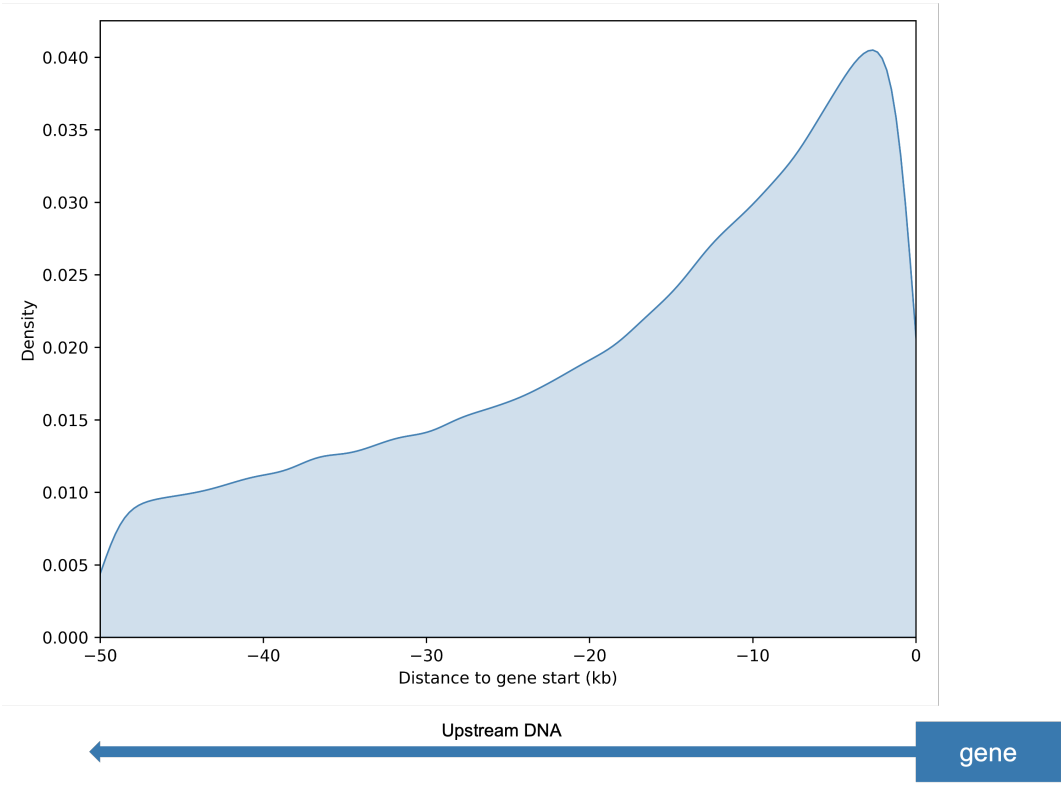


Figure 3.7: **Distribution of genomic windows relative to gene start sites.** Kernel density estimate (KDE) of distances from 292,912 genomic windows to the nearest gene start site. Distances are shown in kilobases (kb), with negative values corresponding to upstream positions relative to the TSS (0 kb). The density increases toward the gene start site, with the highest concentration of windows located within approximately 5 kb upstream. The schematic below the x-axis illustrates the orientation of the upstream region in relation to the gene.

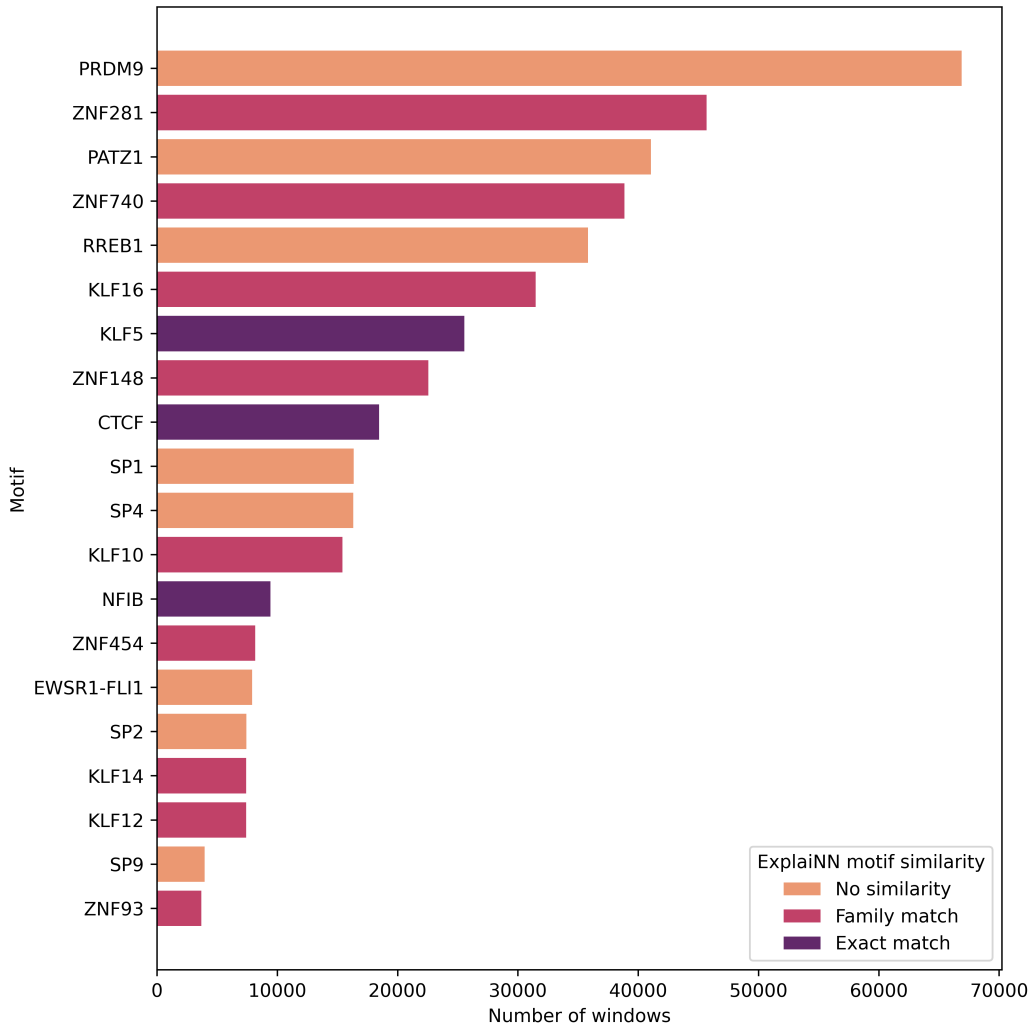


Figure 3.8: **Transcription factor binding motifs identified in predicted functional non-coding windows.** The twenty most frequently occurring transcription factor motifs and their number of significant hits (q-value < 0.05, FIMO) across the final set of predicted functional non-coding windows. Colours indicate similarity to motifs used by ExplaiNN for prediction (see Figure 3.6): orange = no similarity, pink = similarity at the family level, and purple = exact match.

3.5 Discussion

In this chapter, I present the novel application of an ML tool, ExplaiNN, to predict functional non-coding sequences in a non-model species. Through comparative analysis of a suite of model mammals and utilising a multiomic dataset, I apply this method to the cheetah. I demonstrate that it is possible to train a model that

accurately annotates the non-coding genome, with the caveat that data imbalance, caused by the relatively small amount of regulatory sequence compared to non-regulatory sequence in the non-coding genome, impacts the accuracy of this method. These findings provide support for the utility of this novel approach in studying the non-coding genome of non-model species, expanding the potential of research into crucial trait- and disease-associated variants in such species, as well as enabling the identification of deleterious mutations impacting gene expression.

3.5.1 Preparing ATAC-seq data for ExplaiNN

Prior to running ExplaiNN, the ATAC-seq data required preprocessing and formatting correctly. Whilst the majority of human and mouse data showed low duplication in reads and a high number of reads passing filtering steps, the processing steps highlighted several issues with the dog data. Read duplication was high, especially in the liver of Adult 3, which suggests a potential issue with the sequencing of this sample. Additionally, the percentage of mapped reads was very low for all samples, again suggesting a potential sequencing issue. Finally, FRiP scores were all below 0.25, which is considered low compared to ENCODE's recommended standards ("ATAC-seq data standards and processing pipeline", n.d.). As the dataset was an external published dataset and therefore there was not a possibility to resolve the potential sequencing errors, the subset of data that passed filtering was retained. Whilst a higher volume of data is generally considered better for model training (Banko & Brill, 2001), models trained on a subset of this data performed equally well as models trained on the full dataset, demonstrating that this filtering step did not impact model performance.

3.5.2 ExplaiNN: testing on mouse

The first run of ExplaiNN was trained and tested on real positive sequences and synthetic negative sequences, as the negatives were dinucleotide shuffled positive

sequences. Whilst the model performed well, it was not possible to know if the model was learning the difference between ATAC-seq peak and non-ATAC-seq peak or just the difference between real and synthetic DNA. When this model was tested on a dataset of real positives and negatives, it performed much worse, suggesting the model was indeed learning the difference between real and synthetic DNA.

A model trained on real positives and negatives performed less well than the first model, likely because the difference between positives and negatives was not as obvious. This is a phenomenon noted by the authors of ExplaiNN, so this pattern was expected. So far, the models had been trained on 201 bp sequences as ExplaiNN requires training, testing and predicting sets to be comprised of sequences of equal length. However, ATAC-seq peaks are used in this study as a proxy for open chromatin, which suggests some regulatory function in the region. Typical regulatory elements, such as enhancers and promoters, range from 100 bp to several kb in length, so limiting the model to only predict on 201 bp regions may not be as biologically informative and is a key caveat with the analysis presented here. Extending the model to encompass the diversity of lengths of functional non-coding sequences is an important step in any future work, however this involves significant edits to the model training and prediction scripts.

3.5.3 ExplaiNN: predicting across species

Whilst ExplaiNN is not necessarily designed to predict across species, the model performed well when trained on a different species to the species it was tested on. When transferring the ExplaiNN model across species, a decrease in model performance scores (AUC-ROC and AUC-PR) was observed. However, this decrease was not observed if more than one species was used to train the model, resulting in a model that performed as well as the model trained on the same species. This contrasts observations by Cochran et al. (2022), who suggest that cross-species neural networks perform worse than within-species models. This is likely due to my inclusion of multiple species as training data. In this chapter, model perfor-

mance was tested across an evolutionary distance of approximately 82.5 million years (Wu et al., 2017), showing that model performance did not decrease across this distance, provided that either a large amount of training data was used or that more than one species was used to train the model. Therefore, this method with the training data used here is expected to be applicable to any placental mammal with similar model performance. This study is a novel use of ExplaiNN, which has not been trialled across species before, and provides the opportunity to study a wide range of non-model species using only model species data.

3.5.4 ExplaiNN: relationship between ATAC-seq peaks and sequence conservation/distance to gene

One of the key issues when using ExplaiNN to predict functional non-coding regions in this study was the pronounced difference between the number of positives (ATAC-seq peaks) and negatives (non-coding regions with no overlap to ATAC-seq peaks) in the genome. Although artificially balancing the training and testing data was possible to confirm model performance, the mammalian genome is inherently unbalanced when broken into windows, with far fewer functional non-coding regions than non-functional windows. This means that even with very high model performance (AUC metrics ≈ 0.9), when run on an imbalanced dataset, a high number of false positives remain. As the aim of this work is to annotate functional non-coding regions in a non-model species with no available data or annotation, it is not possible to force the dataset to be balanced. Therefore, it was necessary to find additional metrics to separate true and false positives.

Existing methods to annotate the non-coding genome use sequence conservation as a proxy for functional elements (Hubisz et al., 2011). These assume that functional non-coding regions are highly conserved between species, as mutations within those regions will be under strong purifying selection (Christmas et al., 2023; Rands et al., 2014). Therefore, it is expected that true positives have higher sequence conservation than false positives. Whilst a significant difference

between sequence conservation of true and false positives was observed, there was still considerable overlap. However, selecting the putative functional non-coding regions with the highest conservation scores removed a higher proportion of false positives than true positives, giving higher confidence in the model predictions.

As sequence conservation was not enough to accurately distinguish between true and false positives, the location of each window relative to the nearest downstream coding regions was also considered. Although the distance between a functional non-coding region and a gene can vary, most promoters are usually within a few kilobases of coding sequences, with enhancers and distal regulatory elements found several megabases away (Elango & Yi, 2011; Haberle & Stark, 2018; Symmons & Spitz, 2013; Vermunt et al., 2019; M. Q. Zhang, 1998; M. Q. Zhang, 2007). Therefore, true positive windows are expected to be closer to coding regions than false positive windows. Again, whilst a statistically significant difference was observed, there was still considerable overlap between true and false positives, although the true-to-false-positive ratio improved when filtering for distance to coding regions. It is worth noting that the statistical significance observed may, in part, be due to the large number of values in the dataset.

Whilst distance to gene start did improve predictions by reducing the number of false positive results, it is important to note that this step is heavily impacted by the quality and method of genome annotation used. The cheetah genome annotation used here (VMU_Ajub_asm_v1.0; Winter et al., 2023) was generated using a homology-based approach, using nine mammalian genomes as references. This annotation will be missing cheetah-specific orphan genes (genes without homologous sequences across related species) as well as complex gene families such as immune genes, which are known to be poorly annotated using automated annotation approaches (Peel et al., 2022). Therefore, by restricting the set of positive predictions to those within 50 kb of annotated genes, regulatory elements upstream of unannotated complex gene families may be missed. However, these gene families typically require extensive manual curation to provide accurate annotation (Peel et al., 2022), which was beyond the scope of this study.

Another potential way to filter between true and false positives is to overlay transcription factor binding motifs onto the windows, as this can show where functional regions lie (ENCODE Project Consortium, 2012). When analysing the filters of ExplaiNN (the features generated by the model) that were most important in making predictions, I observed a relatively low proportion of filters that matched known TF motifs. Among these, most had positive importance scores, indicating that these motifs contributed primarily to predicting positive sequences (i.e. ATAC-seq peaks). The motif with the highest importance, CTCF, is critical for gene activation and repression, chromatin looping and enhancer blocking (Burcin et al., 1997; Filippova et al., 1996; Hark et al., 2000; Vostrov & Quitschke, 1997). Several motifs had negative importance scores, suggesting they were more important for predicting negative sequences. These were RARβ–RXRα (a heterodimer key in retinoid signalling (le Maire et al., 2019)), TWIST1 (involved in embryonic development (Qin et al., 2012)) and ZBTB24 (involved in DNA methylation at centromeres (Grillo et al., 2025)). Given RARβ–RXRα is a specialised motif involved in retinoid signalling and TWIST1 and ZBTB24 impact centromeres and embryos, it makes sense that these motifs are not likely to be associated with the ATAC-seq peak training data used in this study. This is further evidenced by the lack of expression of these motifs in the heart, liver and stomach in mice based on GTEx data ("GTEx Portal", n.d.).

As ExplaiNN's motif selection is consistent with our understanding of transcriptional regulation, this, combined with other filtering parameters, gave confidence in the final set of predicted 'positives'. However, as discussed previously, many false positives are expected due to the imbalanced number of functional non-coding regions compared to the rest of the genome. Whilst this is not an issue uniquely experienced by ML methods, the problem is more prevalent due to the nature of classification tasks. Therefore, I compared my predictions to existing methods of predicting functional non-coding regions: ultra-conserved elements and TF motif scanning. A vast majority of predicted windows contained at least one significant hit for a TF binding motif, which may suggest some functionality, however the

presence of a motif-like sequence does not guarantee a functional motif. With population-level data, it may be possible to use this information to identify putatively deleterious mutations occurring in TF motifs, but it is not especially valuable for the prediction of functional non-coding regions alone (GTEx Consortium, 2013).

Therefore, to increase confidence in the final set of predicted windows, I also compared to ultra-conserved elements (UCEs), one of the most common computational approaches to predict functional non-coding regions. The majority of predicted windows overlap a UCE, suggesting high correlation between ExplaiNN's predictions and sequence conservation. Windows with no overlap to a UCE may still be true functional regions, as function can be conserved even if sequence is not (Kircher et al., 2014; Rands et al., 2014). However, with the current lack of species-specific data and severely imbalanced training and testing data, it is not possible to discern whether these predictions are accurate or not.

3.5.5 Challenges using machine learning tools

Although ML tools can be applied to species with limited experimental data (beyond the requirement for a reference genome), the models themselves rely on substantial experimental datasets for training. As a result, study systems with little available data from closely related species or relevant tissues may lack sufficient training data. Also, many of these tools are designed to specifically work for humans and transferring to non-human species has not been tested; for example, Puffin, gmk-SVM and DanQ are all trained on human-specific promoter data and transferability is unknown (Dudnyk et al., 2024; Ghandi et al., 2014; Quang & Xie, 2016). Additionally, many tools are not regularly updated and well maintained which, when coupled with poor documentation, can make these tools more difficult to run, particularly by conservation biologists who may not have any training in ML.

Whilst there is a great deal of potential for the application of ML tools, work

must be done to close the knowledge gap between the computational scientists developing the tools and the biologists applying them. Some publications of deep learning methods provide clear explanations of the architecture of the tool and the jargon used in the field of ML, such as DeepMEL (Minnoye et al., 2020), whilst other tools provide clear documentation and tutorials on how to run different parts of the tool (e.g. Basenji (Kelley et al., 2018), ExplaiNN (Novakovsky et al., 2023)). However, this is not consistent across tools, as some do not provide clear explanations of the method or straightforward scripts, making the tools inaccessible for those without training in machine learning. To close this gap, we must prioritise accessibility of machine learning tools and facilitate collaboration between computer scientists and biologists.

Finally, the most critical issue with current machine learning tools for functional non-coding prediction is that genomic data is inherently severely imbalanced (Schubach et al., 2017). Even with the most accurate classifier trained on a large amount of data, due to the sparsity of functional non-coding regions throughout the genome, resultant predictions may suffer from a lack of sensitivity, resulting in false positives. This issue persisted here even when alternative performance metrics were used, and precision and recall were both impacted. Future machine learning tools must therefore be built with this imbalance in mind to enable robust predictions.

3.6 Conclusion

In this chapter, I show the potential for machine learning methods to predict functional non-coding regions of the genome in non-model species. Using this method, I have annotated the non-coding genome of a non-model species and demonstrate the use of ML to study species- and lineage-specific patterns in functional non-coding regions of any species, albeit with appropriate tentativeness. Whilst this CNN-based approach can accurately predict such regions, class imbalance in ge-

nomic datasets introduces substantial noise. Integrating sequence conservation information can mitigate this effect and as ML methods improve, so too will their applicability to non-model species. This framework provides a novel approach to annotate non-coding genomes of non-model species without the requirement for bespoke experimental data, providing valuable insights into the genetic basis of traits and diseases and enabling the study of deleterious mutations outside of genes.

3.7 Supplementary material

3.7.1 Supplementary figures

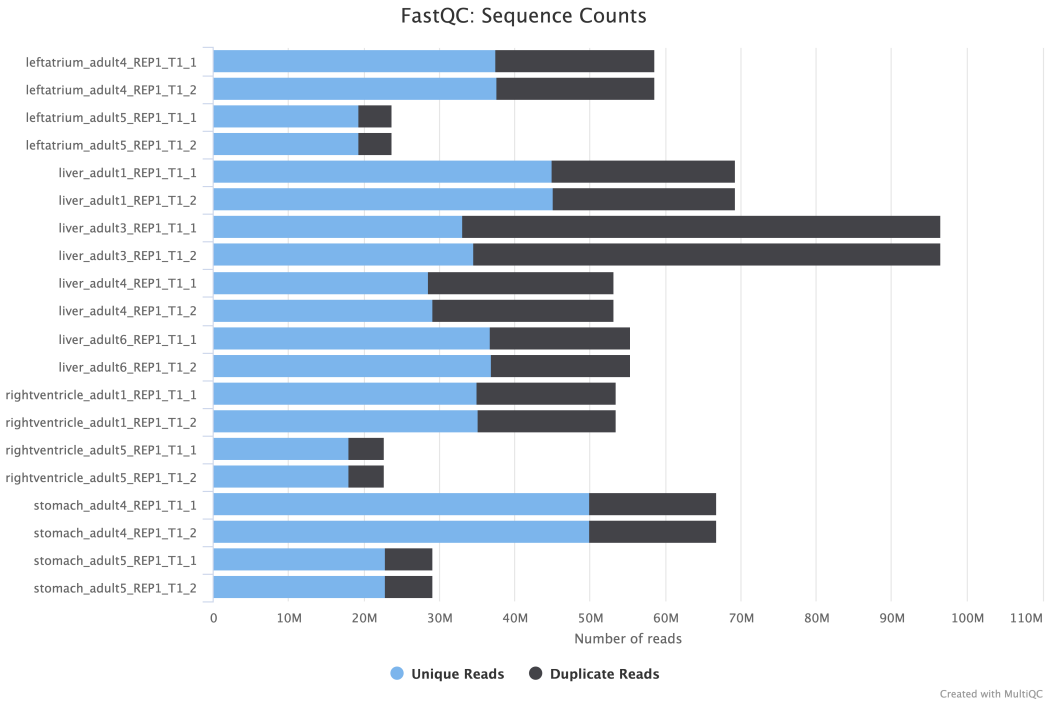


Figure S3.1. **Total number of ATAC-seq reads in dog.** Total number of reads provided as input for each dog sample (ranging from 22,763,816 to 96,521,458) and proportion of those marked as duplicates (ranging from 18.3 to 64.2%). Plot generated by MultiQC as part of the *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

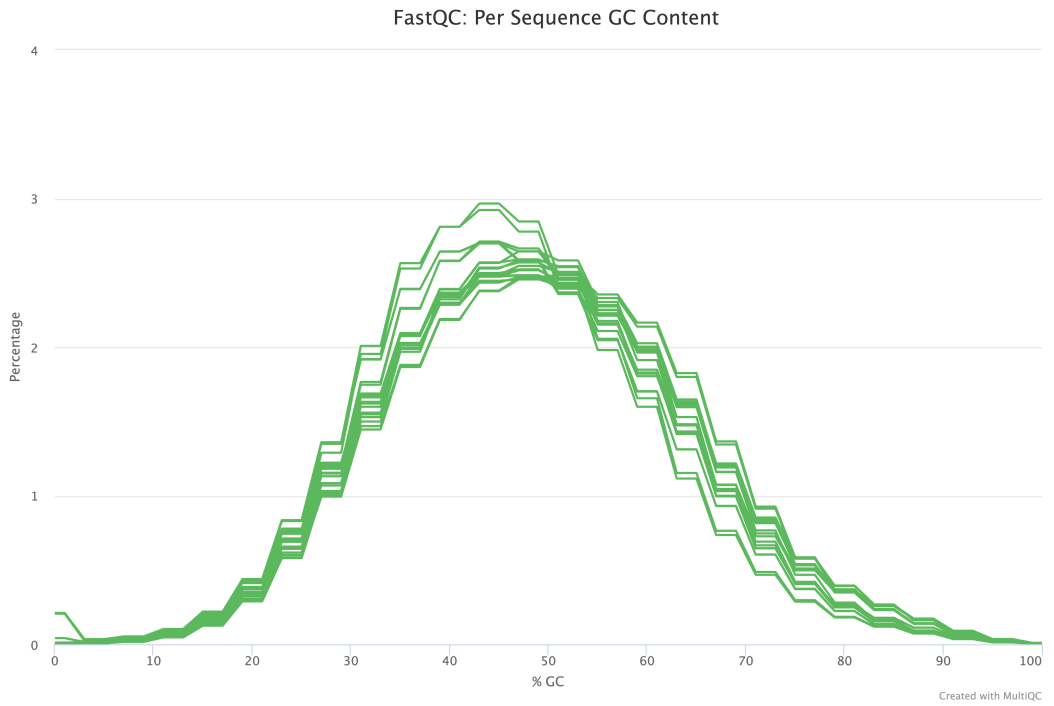


Figure S3.2. **GC content of ATAC-seq reads in dog.** Per-sequence GC content calculated by FastQC showing roughly normal distribution. Plot generated by MultiQC as part of *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

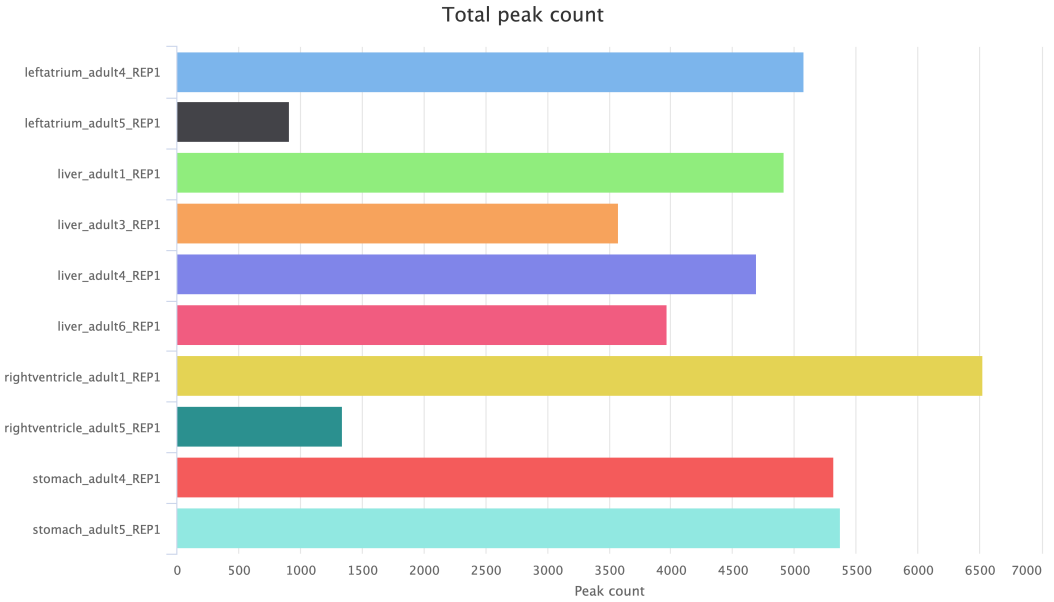


Figure S3.3. **Total peak count in dog.** Total narrow peak count for each sample as called by MACS2, ranging from 623 to 5,084 peaks per sample. Plot generated by MultiQC as part of *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

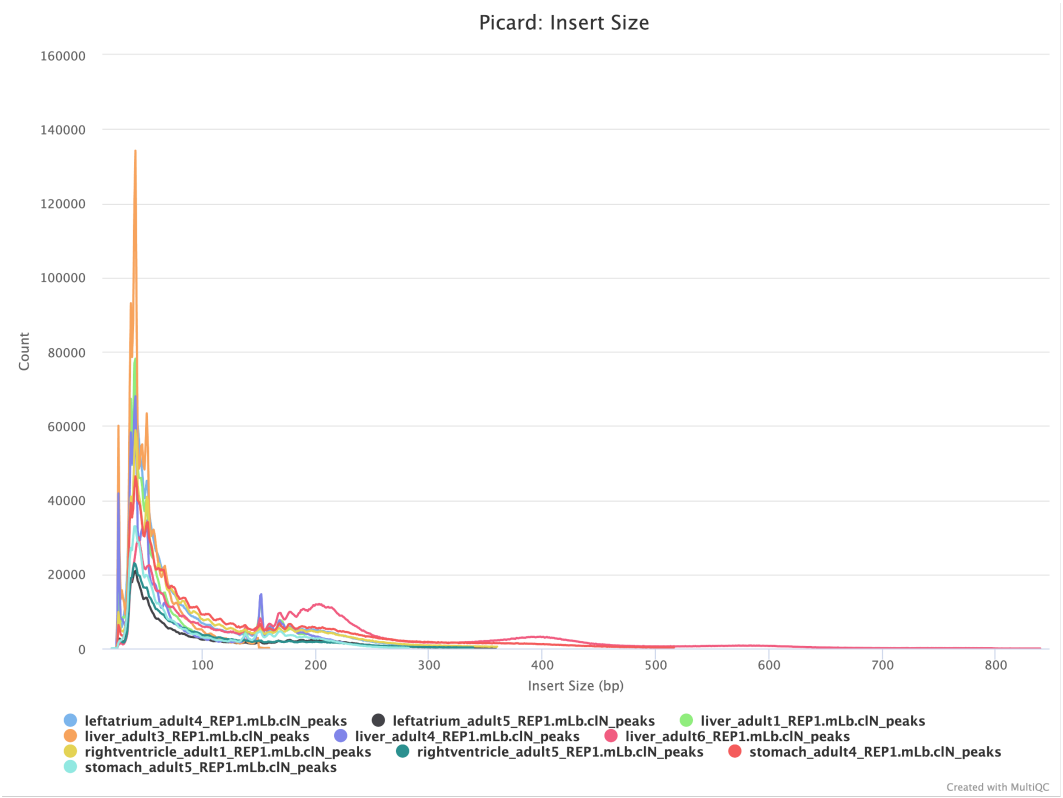


Figure S3.4. **Periodicity plot in dog.** Periodicity plot showing insert size of each peak. Plot generated by MultiQC as part of *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

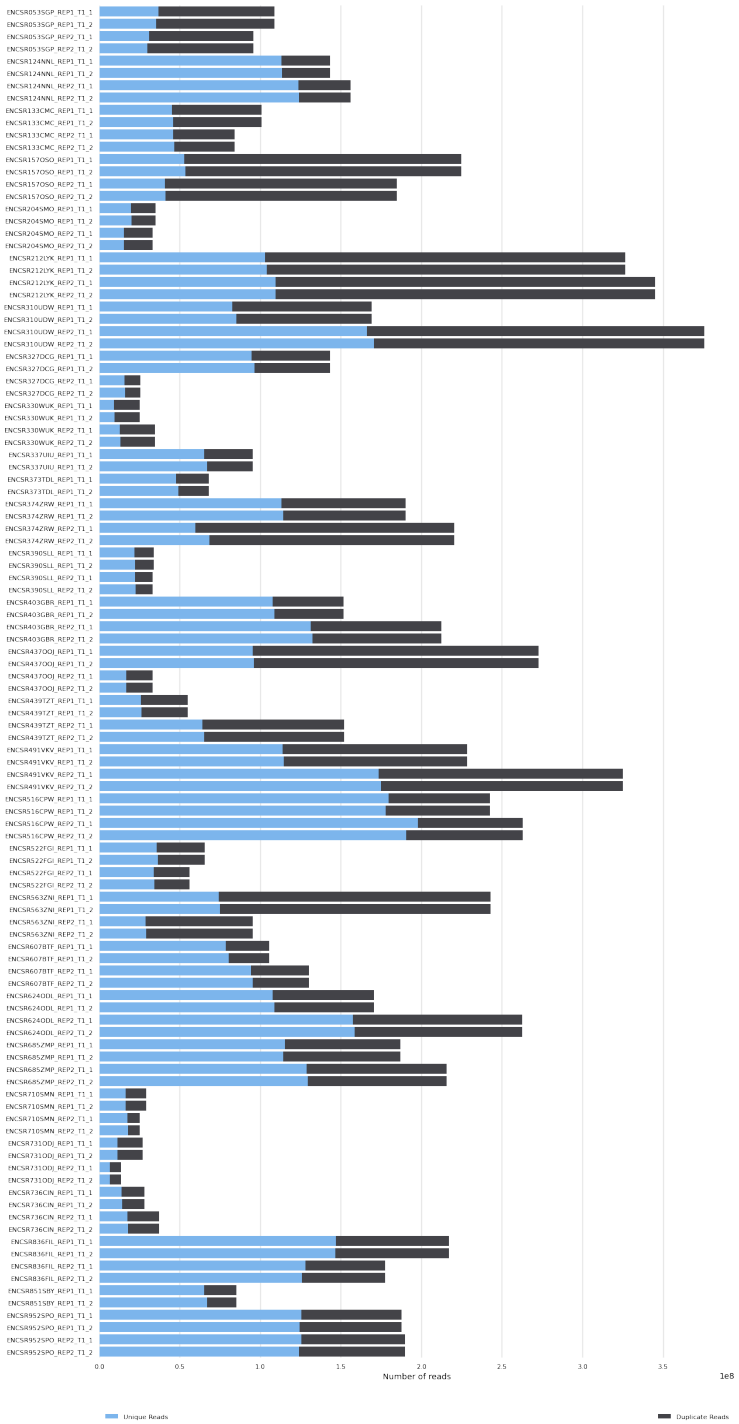


Figure S3.5. **Total number of ATAC-seq reads in human.** Total number of reads provided as input for each sample (ranging from 13,593,438 to 375,684,903) and proportion of those marked as duplicates. Plot generated by MultiQC as part of *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

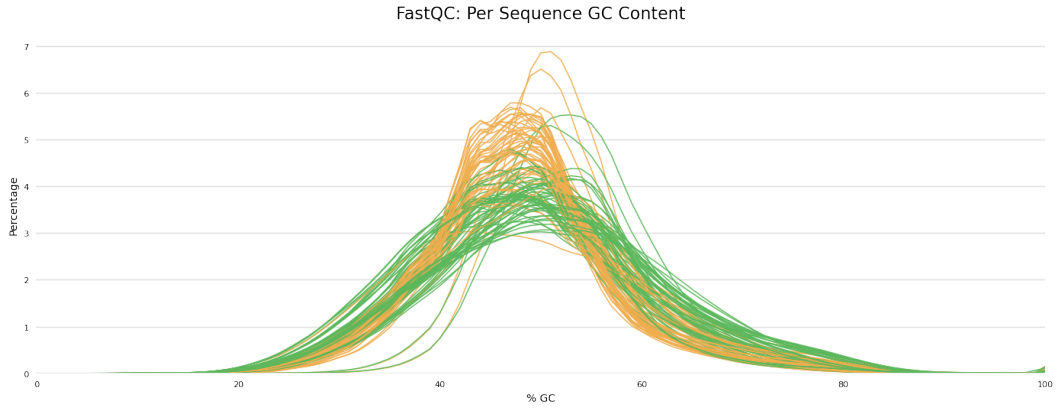


Figure S3.6. **GC content of ATAC-seq reads in human.** Per-sequence GC content calculated by FastQC showing roughly normal distribution. Plot generated by MultiQC as part of *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

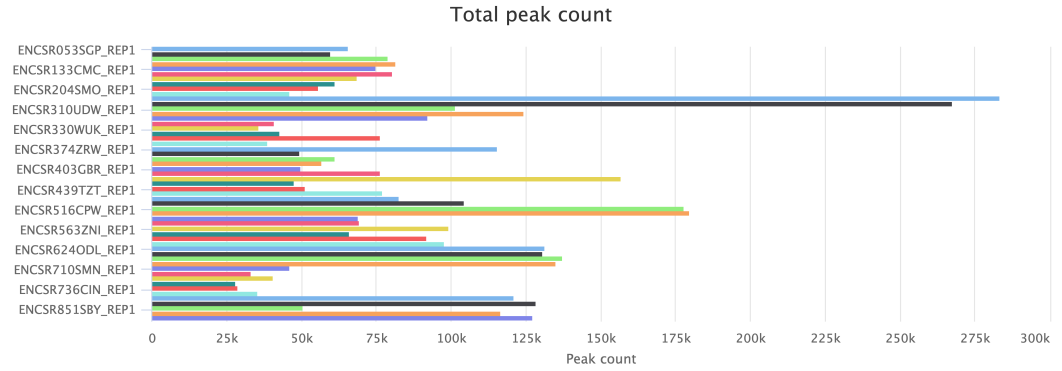


Figure S3.7. **Total peak count in human.** Total narrow peak count for each sample as called by MACS2, ranging from 37,992 to 340,687 peaks per sample. Plot generated by MultiQC as part of *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

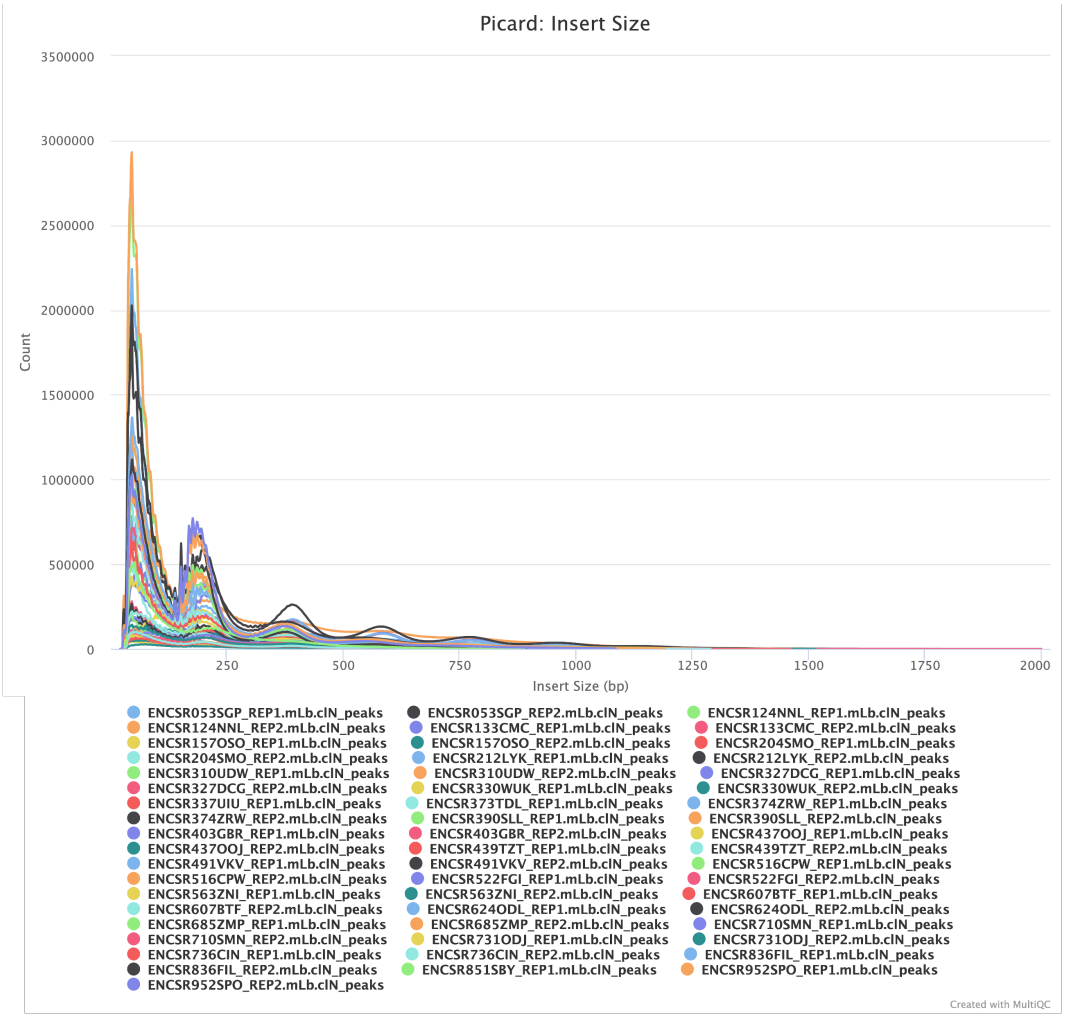


Figure S3.8. **Periodicity plot in human.** Periodicity plot showing insert size of each peak. Plot generated by MultiQC as part of *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

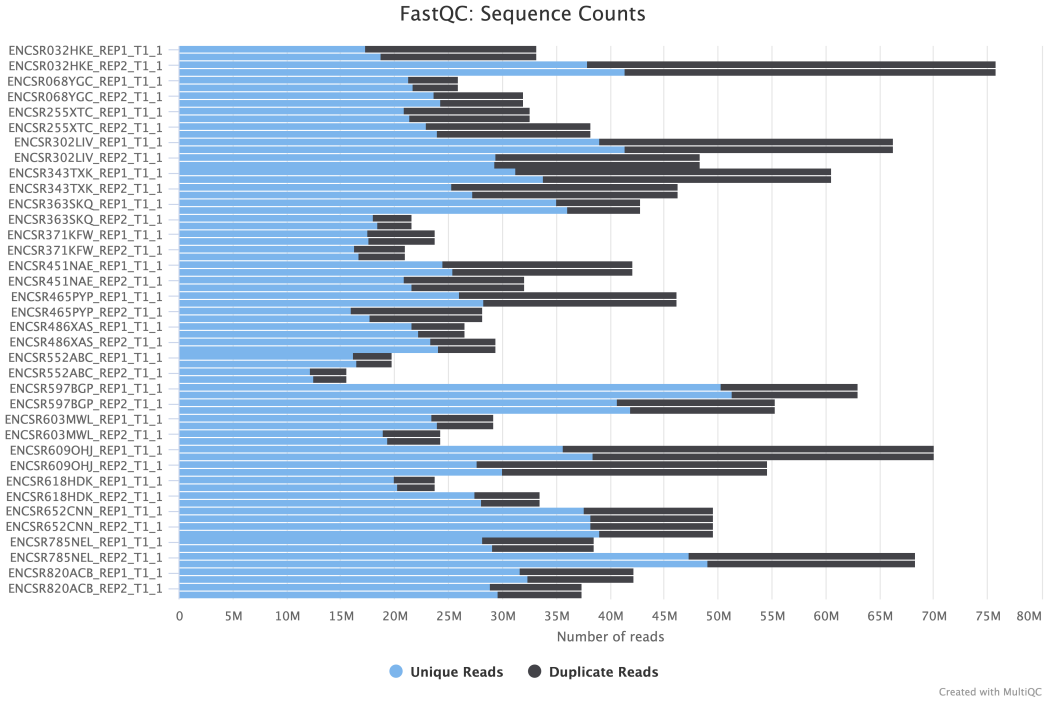


Figure S3.9. **Total number of ATAC-seq reads in mouse.** Total number of reads provided as input for each sample (ranging from 15,591,507 to 75,813,558) and proportion marked as duplicates (14.6–50.0%). Plot generated by MultiQC as part of *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

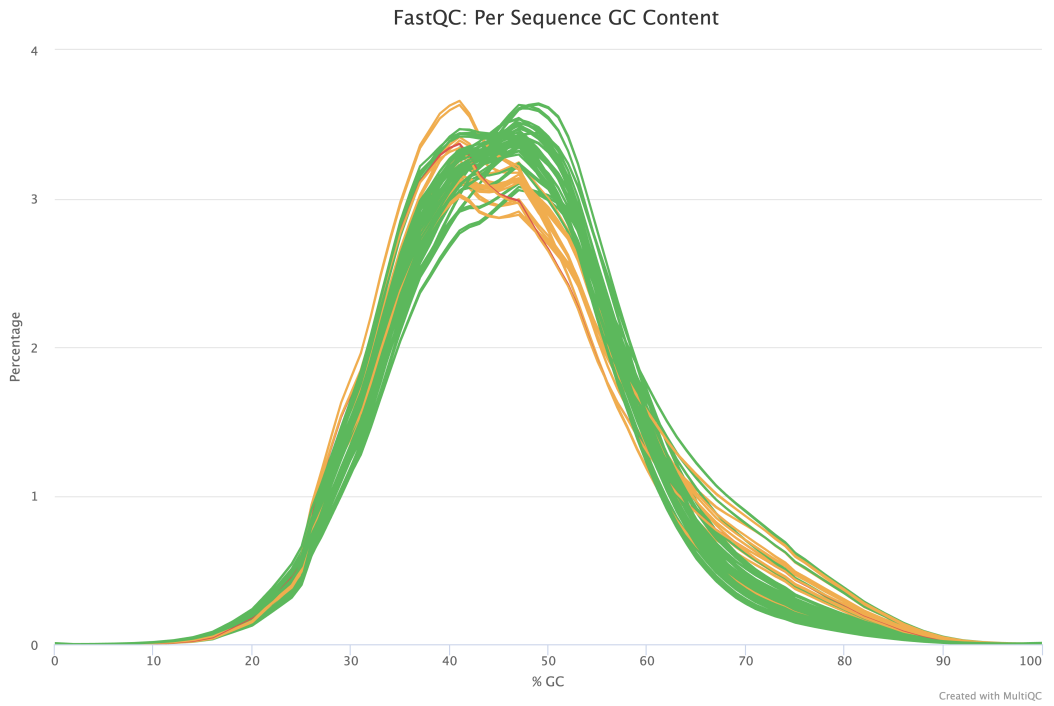


Figure S3.10. **GC content of ATAC-seq reads in mouse.** Per-sequence GC content calculated by FastQC showing roughly normal distribution. Plot generated by MultiQC as part of *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

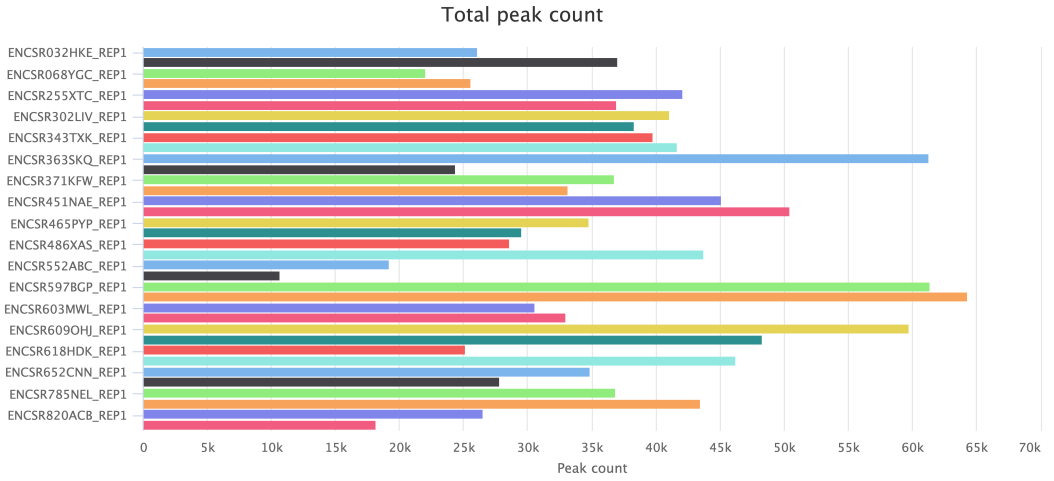


Figure S3.11. **Total peak count in mouse.** Total narrow peak count for each sample as called by MACS2, ranging from 17,603 to 70,592 peaks per sample. Plot generated by MultiQC as part of *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

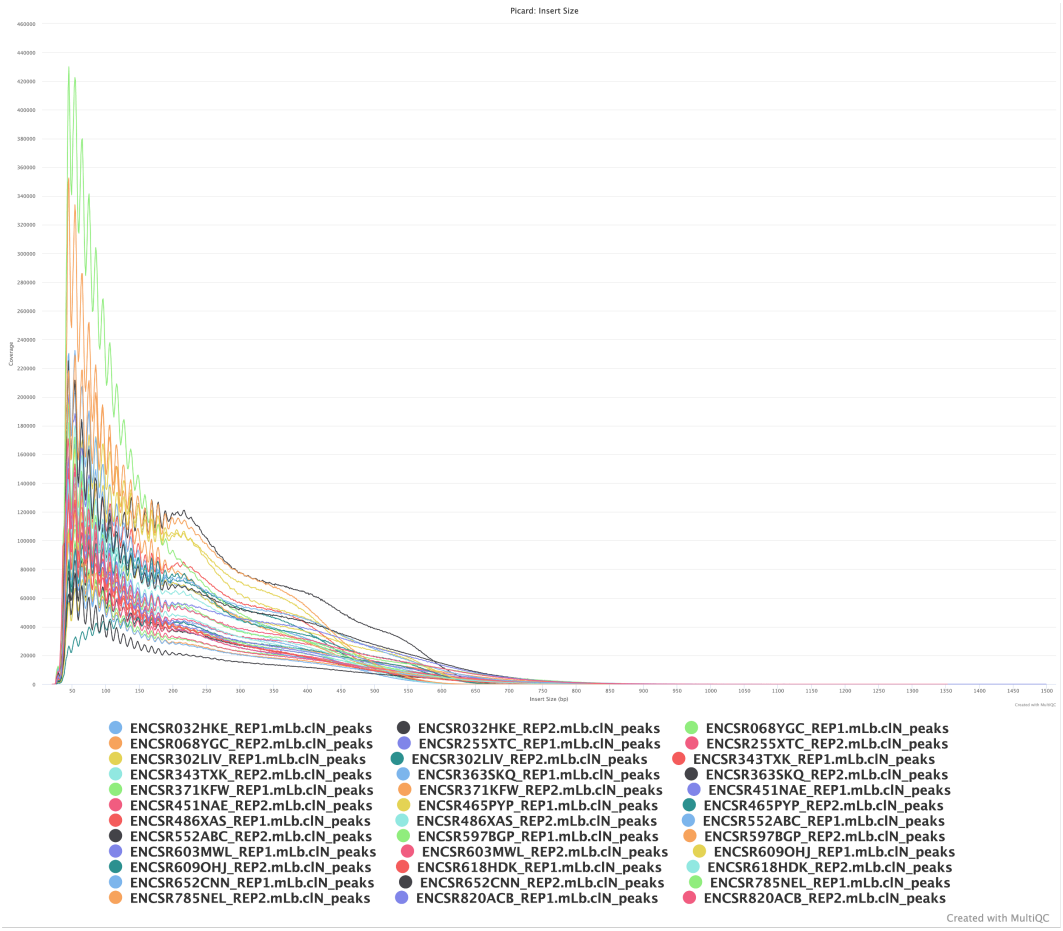


Figure S3.12. **Periodicity plot in mouse.** Periodicity plot showing insert size of each peak. Plot generated by MultiQC as part of *nf-core atacseq* pipeline v2.1.2 (P. Ewels et al., 2016; P. A. Ewels et al., 2020; Patel et al., 2023).

3.7.2 Supplementary tables

Supplementary tables can be found at github.com/EarlhamInst/JP_PhD.

Table S3.1. **Mouse & Human ATAC-seq samples.** Mouse and human ATAC-seq samples used in this study, downloaded from ENCODE (https://www.encodeproject.org/). For each sample, the species, tissue, age, sex, health status and ethnicity is provided (where relevant). Experiment, library and read accessions are provided alongside the relevant DOI. For this study, raw ATAC-seq read files (fastq) were used, but information for all files related to each sample is provided.

Table S3.2. **Dog ATAC-seq samples.** Dog ATAC-seq samples used in this study, downloaded from BarkBase (https://www.barkbase.org/). For each sample, the species, tissue, individual and corresponding accessions are provided.

Table S3.3. **Reference genomes for Felidae alignment.** Reference genomes used to generate an alignment of all chromosome-level Felid genomes, which was generated with Progressive Cactus and used to calculate conservation scores.

Table S3.4. **Transcription factor (TF) motifs learned by ExplaiNN for each species-specific model (Models 2, 3, 4).** The importance for each filter for predicting positives (ATAC-seq peaks) and negatives (non-ATAC-seq peaks) is shown alongisde the corresponding JASPAR motif and the significance of the JASPAR motif predictions.

Chapter 4

Distribution of genome-wide deleterious variants in wild and captive cheetah populations

Photograph: Cheetah siblings in Okonjima Nature Reserve, Namibia. Credit: Jessica Peers

The work in this chapter was completed by Jessica Peers with contribution from Dr Graham Etherington, who provided a script for variant calling with GATK, Dr Sam Speak, who produced cheetah-specific CADD scores, and Heather Sibley, who generated a pedigree figure.

4.1 Abstract

Small populations tend to harbour an increased load of genome-wide deleterious mutations as purifying selection becomes less effective and inbreeding increases. Reflecting this pattern, deleterious mutations in genes associated with fertility and immunity have previously been identified in the cheetah (*Acinonyx jubatus*), which has had a low effective population size for the last 10,000 years. However, the distribution of deleterious mutations across cheetah populations is currently unknown. Here, I present analysis of novel whole genome resequencing data sourced from 32 captive and wild cheetahs. I investigate variation in genetic diversity, genomic measures of inbreeding and the distribution of deleterious mutations across cheetah populations. I identify higher inbreeding in South Sudan and Tanzanian cheetahs, but a higher load of population-specific deleterious mutations in Namibian cheetahs. Protein-coding genes containing high- or moderate-impact deleterious mutations were significantly enriched for sperm-related functions, highlighting putative causative loci associated with poor sperm quality in cheetahs. Similar levels of genetic diversity and inbreeding were observed in captive cheetahs compared to their wild counterparts, providing empirical evidence of the efficacy of the captive breeding programme in maintaining genetic variation in *ex situ* populations.

4.2 Introduction

Small populations are more prone to increased inbreeding and the impacts of genetic drift, which can result in an accumulation of deleterious mutations across the

genome (Björnerfeldt et al., 2006; Lande, 1994; Masel, 2011). In wild populations, this can lead to 'extinction vortex' effects, where these factors continually interact in a feedback loop, eventually resulting in population extinction (Gilpin & Soulé, 1986). However, in captivity, these populations are typically carefully managed to ensure stable population sizes whilst minimising inbreeding.

Despite this potential for careful management and subsequent mitigation of inbreeding effects, captive populations often originate from few individuals, resulting in strong founder effects. Such effects decrease genetic diversity in the population and increase the likelihood of inbreeding, resulting in an accumulation of deleterious mutations (Björnerfeldt et al., 2006; Muller, 1964). Additionally, in captivity, selection may be relaxed, meaning such deleterious alleles can spread through the population (Lynch & O'Hely, 2001). Purging, where deleterious alleles are exposed to selection through increased homozygosity, can remove the most deleterious alleles from the genome. In some cases, purging has the potential to make the captive population healthier than the wild one, as has been observed in some captive ungulates (López-Cortegano et al., 2021). However, studies spanning mammals, birds, reptiles and amphibians have suggested that purging rarely has a significant effect on fitness and it is unlikely to counteract inbreeding depression (Boakes et al., 2007; Leberg & Firmin, 2008).

4.2.1 Cheetahs in captivity

The cheetah is an excellent study system for the impacts of captivity on a species which has experienced significant inbreeding and prolonged low effective population size. The cheetah is hypothesised to have experienced severe demographic decline beginning 10,000 years ago (Dobrynin et al., 2015; Fabiano et al., 2025; O'Brien & Johnson, 2007). This resulted in sustained low effective population size and a range of symptoms associated with inbreeding depression. Currently classed as 'Vulnerable' by the IUCN Red List, approximately 7,000 cheetahs remain in the wild (Durant et al., 2017). As a charismatic species that is easier to train than

other large cats, cheetahs have a long history in captivity. The earliest record of a captive cheetah comes from the Sumerians in approximately 3,000 BCE, and royals and emperors from the 5th to 16th centuries are thought to have kept hundreds of cheetahs at any one time (Marker-Kraus, 1997). Due to continued capture of cheetahs since that period, alongside persecution and habitat loss, wild populations declined; by the 1950s, the cheetah was declared extinct in India and Israel (Kraus & Marker-Kraus, 1991).

The first record of a cheetah in a zoological collection was an animal at the Zoological Society of London in 1829, although very few cheetahs were exhibited globally in the 1800s and no captive births were recorded (Marker-Kraus, 1988). The first authenticated captive birth of a cheetah litter occurred in 1956 at Philadelphia Zoo, with captive breeding steadily increasing from 1973 onwards (L. Marker & O'Brien, 1989; Marker-Kraus, 1997). Whilst the captive population in 1956 was comprised of 45 individuals, continual imports of wild animals reduced the potential impact of a founder effect. However, by 1986, issues with captive breeding of the population became apparent; despite a population size of almost 200 cheetahs, the effective breeding size was just 28, and high infant mortality compared to other captive species was reported (L. Marker & O'Brien, 1989). This sparked a sequence of genetic studies, identifying low genetic diversity, particularly in MHC genes, and poor sperm quality in cheetahs (O'Brien et al., 1985; Wayne et al., 1986; Wildt et al., 1983).

4.2.2 Captive breeding programmes

Conservation initiatives began in the 1980s, when a Species Survival Plan (SSP) was implemented and an international studbook, containing a record of all cheetah births and deaths in captivity, was published by the Cheetah Conservation Fund (Marker-Kraus, 1997). The SSP initiated a formal captive breeding programme for cheetahs, coordinated by the Association of Zoos and Aquariums (AZA); prior to this, captive breeding records were informal and wild-caught animals were still

being imported. This lack of coordinated management meant that the origins of founding individuals were poorly documented, increasing the risk of inbreeding and hybridisation between lineages.

In the latter half of the 20th century, there were at least 1,440 cheetahs in zoological collections worldwide (Marker-Kraus, 1997). Many of these animals originated in East Africa, namely Kenya and Somalia, but export from these countries was halted when local populations began to decline (Marker-Kraus, 1988). After this point, cheetahs were primarily imported from Namibia, where cheetah populations were more stable. In 1975, new CITES legislation restricted imports of wild-caught cheetahs (Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), 1992). Cheetah imports continued into the 1990s, with wild cheetahs being imported into captivity as recently as 2020 (L. Marker & Johnston, 2022). These individuals were either imported to Ashia Cheetah Conservation, a reserve in South Africa, or were confiscated by the Cheetah Conservation Fund in Hargeisa, Somaliland, and subsequently moved to the Somaliland Cheetah Rescue and Conservation Centre as part of a coordinated conservation action (L. Marker & Johnston, 2022). Today, there are almost 2,000 cheetahs in captivity globally, although this number does not include privately-owned animals.

In the late 1990s, almost all founders in the known global captive population originated from the Southern African subspecies (*A. j. jubatus*), except for seven individuals from East Africa (*A. j. jubatus*, previously *A. j. raineyi*), 2 of which did not breed (Marker-Kraus, 1997). Over 90% of all captive cheetahs at this point were thought to descend from Namibian wild-caught animals (Marker-Kraus & Kraus, 1997). The captive population was founded by approximately 424 individuals, although by 1994 most founders had not yet bred successfully (Marker-Kraus, 1997). However, among the breeding individuals in 1994, 13% were identified as hybrids between the Southern and Eastern populations (Marker-Kraus, 1997). As a result, *A. j. jubatus* is over-represented in the captive population, whereas lineages such as the Northern African, Asian, and Eastern African cheetahs are either poorly represented or not represented at all (Marker-Kraus, 1997).

4.2.3 Genetic diversity of captive cheetahs

Although founder effects in the captive cheetah population may have been reduced by consistent introduction of wild animals throughout the 20th century, such imports do not act as a guarantee of high genetic diversity in the captive population. Captive breeding programmes select breeding pairs based on relatedness, typically informed by a pedigree (Couvet & Ronfort, 1994; Fernández & Caballero, 2001). However, whilst this practice is likely to reduce breeding between closely related individuals, it does not necessarily reduce the spread and fixation of deleterious alleles. This is particularly problematic for captive animals, as selection in captivity can differ in both strength and directionality compared to the wild, meaning deleterious alleles can segregate at higher frequencies (Frankham, 2008; Lynch & O'Hely, 2001; Woodworth et al., 2002). Additionally, this method relies on the accuracy of the studbook, which is not well-maintained in all species (Giontella et al., 2020; Ivy & Lacy, 2010). Together, these factors can facilitate the accumulation and spread of deleterious variants within captive populations.

These processes are particularly concerning in a reproductive context. Cheetahs are known to exhibit signs of poor sperm quality and reproductive defects (O'Brien et al., 1985; Wildt et al., 1983), which may prevent reproduction in the wild. However, in captivity, individuals carrying such mutations may still breed, sustaining the mutations in the population. For example, in vitro fertilisation (IVF) was carried out on cheetahs in 2020, resulting in two offspring (Crosier et al., 2020). Whilst such actions maintain numbers of captive animals, deleterious mutations which may have otherwise prohibited successful reproduction could be passed to offspring and maintained in the population.

Since the first observation of low genetic diversity in captive cheetahs (O'Brien et al., 1985), researchers have continued to investigate this phenomenon (Terrell et al., 2016). To date, the majority of studies have sequenced microsatellite panels or specific loci, with just eight whole genomes published, meaning genome-wide genetic diversity across cheetah populations is unknown (Castro-Prieto et al., 2011;

Dobrynin et al., 2015; Driscoll et al., 2002; Winter et al., 2023). In particular, the genetic diversity and mutation load of the non-coding genome has not been studied, preventing the identification of mutations in regulatory regions, such as promoters and enhancers, which can impact gene expression (Joshi et al., 2021; Scacheri & Scacheri, 2015; Wells et al., 2019).

The severe genetic bottleneck and low effective population size of cheetahs predates captivity (R. Barnett et al., 2005; Kim et al., 2016; O'Brien & Johnson, 2007; O'Regan & Steininger, 2017), meaning it is likely that the majority of accumulated deleterious mutations are present in both wild and captive populations. However, these populations have been exposed to contrasting selection pressures and breeding histories. As captive breeding programmes often aim to act as reserve populations for potential reintroduction (Fraser, 2008; Witzenberger & Hochkirch, 2011), it is important to understand the relationship between the genetic diversity of wild populations and their captive equivalents. Studying the genome-wide genetic diversity of wild and captive cheetahs will allow us to consider the genetic load of the captive population compared to their wild counterparts.

Here, whole genomes of 30 captive and 2 wild cheetahs are sequenced and analysed alongside existing data from wild populations to investigate population structure, genetic diversity, and the genome-wide distribution of deleterious variants. A consistent separation of captive US and wild Namibian cheetahs from other wild populations is identified, consistent with historical records of cheetah imports. Deleterious mutations are found to be over-represented in genes associated with spermatozoal flagella in both captive and wild populations, suggesting potential causative loci of poor sperm function in cheetahs. Previously identified premature termination codons are revealed to be fixed across all cheetah populations studied. Notably, a higher proportion of unique deleterious SNPs is observed in Namibian cheetahs, highlighting important considerations for future translocation projects.

4.3 Methods

4.3.1 Samples

Whole genome re-sequencing data from 39 cheetahs was used in this study; 32 newly generated and seven sourced from Dobrynin et al. (2015) (Table 4.1). Of the newly sequenced individuals, 30 were sourced from the US captive population, collected by Drs Adrienne Crosier and Klaus-Peter Koepfli from individuals owned by several institutions (Table S1). Captive US samples were sequenced from whole blood (except AJU6540, which was skeletal muscle tissue). As part of a larger collaborative study on captive cheetahs in the US, several known family groups were included in the sample set (Figure 4.1). The remaining two newly sequenced genomes were blood samples collected in South Sudan by Kelsey Greene at African Parks South Sudan.

Table 4.1: **Metadata for cheetah individuals included in this study.** Information for each sample is included: sample identifier, origin (captive or wild), population (country of origin), sex, and house name (where available). Sample IDs marked with * are from Dobrynin et al. (2015).

Sample ID	Captive/Wild	Population	Sex	House name
AJU6540	Captive	US	M	Alberto
AJU7221	Captive	US	F	Sanurra
AJU7225	Captive	US	M	Sampson
AJU8426	Captive	US	F	Carmelita
AJU8957	Captive	US	F	Hope/Happy
AJU8965	Captive	US	M	Sukuri
AJU9216	Captive	US	F	Echo
AJU9459	Captive	US	M	Scott
AJU9460	Captive	US	M	Asante
AJU9461	Captive	US	F	Rosalie
AJU9481	Captive	US	F	Darlene
AJU9660	Captive	US	F	Teona

Table 4.1 (continued)

Sample ID	Captive/Wild	Population	Sex	House name
AJU9746	Captive	US	F	Rebel
AJU9840	Captive	US	M	Mozzie
AJU9889	Captive	US	M	Flash
AJU9890	Captive	US	M	Dougie
AJU9907	Captive	US	M	Roosevelt
AJU9910	Captive	US	F	Fossey
AJU9914	Captive	US	F	Zuri
AJU10188	Captive	US	M	Padfoot
AJU10270	Captive	US	M	Barafu
AJU10272	Captive	US	F	Amani
AJU10277	Captive	US	M	Upepo
AJU10682	Captive	US	M	Ellis
AJU10689	Captive	US	F	Kingsley
AJU10766	Captive	US	F	Bam Bam
AJU10899	Captive	US	F	Lane
AJU10902	Captive	US	M	Ragnar
AJU10949	Captive	US	M	Vogan
AJU10970	Captive	US	M	3D
18032FL-172-01-05_S5_L005	Wild	South Sudan	F	NA
18032FL-172-01-06_S6_L005	Wild	South Sudan	F	NA
SRR2737543*	Wild	Tanzania	M	NA
SRR2737544*	Wild	Tanzania	M	NA
SRR2737545*	Wild	Tanzania	F	NA
SRR2737540*	Wild	Namibia	F	NA
SRR2737541*	Wild	Namibia	M	NA
SRR2737542*	Wild	Namibia	M	NA
SRR27375chewbacca*	Wild	Namibia	M	Chewbacca

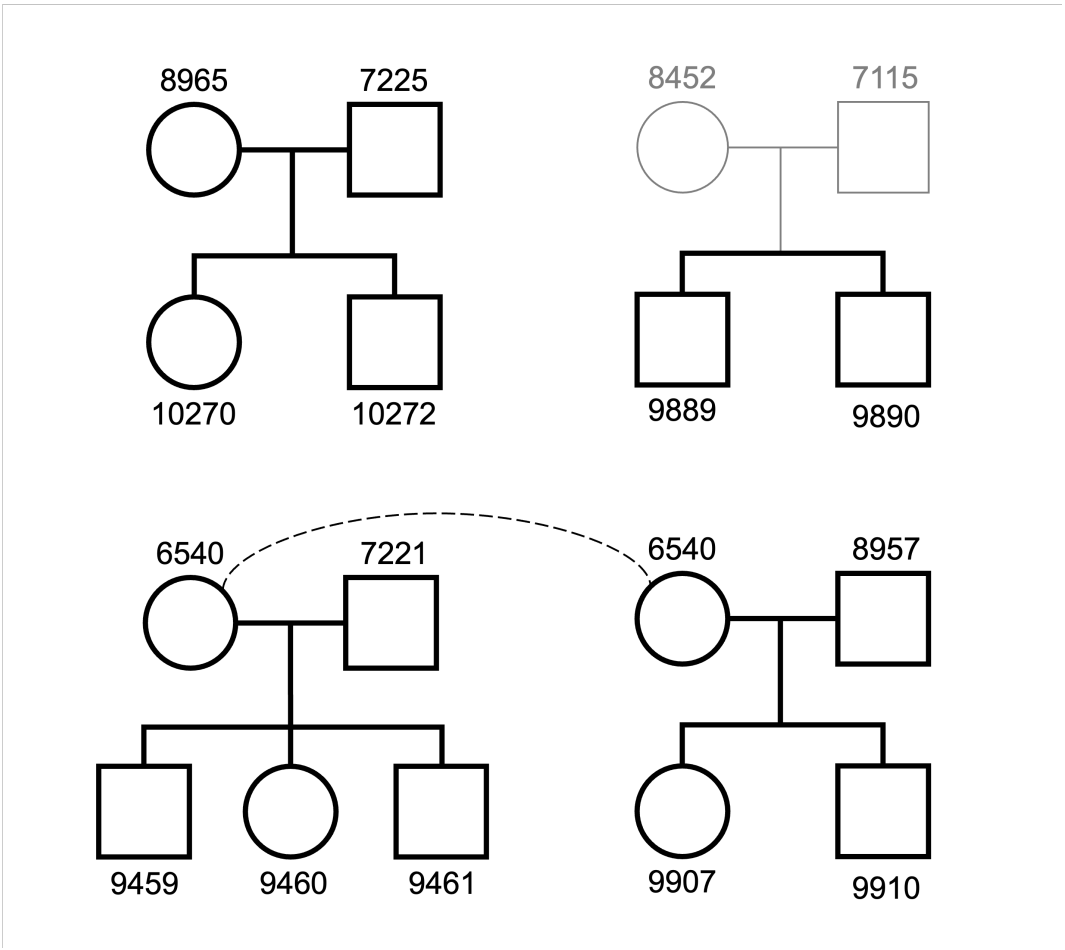


Figure 4.1: **Pedigrees of family groups included in this study.** Females are represented as circles and males as squares. Individuals shaded in grey (8452, 7115) were not sampled, but their offspring were included. The dashed line indicates that female 6540 contributed offspring to two different family groups.

4.3.2 DNA extraction and sequencing

For captive US samples, DNA extraction, library preparation and sequencing was conducted by Psomagen (Maryland, USA). DNA extraction was conducted using the QIAamp DNA Blood Mini Kit. Libraries were prepared using the TruSeq DNA PCR Free Library Preparation Kit, with an insert size of 350 bp. Sample AJU6540 failed initial library preparation due to low initial DNA quantity and so required a second DNA extraction and library preparation attempt. Libraries were sequenced by Psomagen on a single 10B flowcell on the Illumina NovaSeqX Plus platform.

For wild South Sudan samples, DNA was extracted by Inqaba Biotec (Pretoria, South

Africa) using the Quick-DNA Miniprep Plus Kit. Library preparation and sequencing were conducted by Admera Health (New Jersey, USA). Library preparation was conducted using the NEBNext Ultra II Library Preparation Kit, generating 150 bp insert libraries. Libraries were sequenced on a 25B flowcell on the Illumina NovaSeqX Plus platform.

4.3.3 QC, mapping and variant calling

Raw reads were initially assessed using *FastQC* v0.11.9 (Andrews, 2010) on each sample fastq file. Some samples had overrepresented strings of guanines (approximately 0.3% of sequences) however these were removed by subsequent filtering. To avoid a potential sequencing batch effect it is recommended to trim all sequencing reads to the same length (Leigh et al., 2018; Shaw et al., 2025). Therefore, reads from US and South Sudan samples were trimmed to 100 bp to match samples from Namibia and Tanzania using *Trimmomatic* v0.39 (Bolger et al., 2014). Adapters were trimmed using *Trim_galore* v0.6.10 with flags 'paired' and 'retain_unpaired' (Krueger, n.d.). Paired and unpaired reads were mapped to the most recent cheetah reference genome (VMU_Ajub_asm_v1.0, GCA_027475565.2, Winter et al., 2023) using *BWA-MEM* v0.7.17 (H. Li, 2013) and resultant BAM files were merged with *SAMtools* v1.18 (Danecek et al., 2021). Finally, read groups were added using *GATK* v4.6.0.0 (McKenna et al., 2010) and BAM files indexed with *SAMtools* v1.18 (Danecek et al., 2021).

Variant calling and initial filtering steps were performed using *GATK* v4.6.0.0 (McKenna et al., 2010). For each sample, the workflow included the following steps: *SortSam*, *MarkDuplicates*, and *HaplotypeCaller* (run in ploidy-aware mode for male samples), followed by execution of the gatk_params.py script (https://github.com/EarlhamInst/JP_PhD/blob/main/gatk_params.py; script written by Dr Graham Etherington). Variants were processed through *VariantFiltration* and *SelectVariants*, after which *BaseRecalibration* and *ApplyBQSR* were applied. *HaplotypeCaller* was then run again to produce the recalibrated variant calls. GVCF files were indexed using *IndexFeatureFile* and a database of variants was generated using *GenomicsDBImport*. Finally, a multi-sample VCF was created using this database with *GenotypeGVCFs*. Single Nucleotide Polymorphisms (SNPs) were extracted using *SelectVariants* and variants were filtered using default *GATK* hard filtering parameters with reduced FS (FisherStrand) requirement

corresponding to the data and a minimum and maximum depth of 3 (1st percentile) and 17 (99th percentile), respectively (GATK Team, 2025; Kryvokhyzha, 2016).

BCFtools v1.22 (Danecek et al., 2021) was used to extract biallelic sites, exclude the X chromosome and exclude sites missing in >15% of samples. Depending on the requirements of each analysis, additional filters were applied using *BCFtools* (Table 4.2). For some analyses (specified below), sites with minor allele frequency (MAF) < 0.025 and related individuals were excluded. Linkage-disequilibrium (LD) pruning was performed with r^2 of 0.125 and window size of 50,000 kb, again only applied for some analyses (stated below). Finally, invariant sites were removed.

Table 4.2: **Additional filtering steps applied for each VCF file.** All VCFs were filtered using GATK hard filters, biallelic sites extracted and sites missing in >15 % of samples removed. Following this, analysis-specific filters were applied.

VCF	Additional filters applied	Number of SNPs
1	N/A	4,142,757
2	Relatives removed	4,110,227
3	MAF	3,533,757
4	LD prune & MAF	279,045
5	Relatives removed, LD prune & MAF	273,831

4.3.4 Kinship

As many population genetic analyses require individuals to be unrelated, it was necessary to quantify relatedness between samples. *PLINK* v.2.0.0 (Chang et al., 2015; Purcell et al., 2007) was used to calculate kinship using the *KING* algorithm on VCF4 (see Table 4.2). Results were compared to information from the International Cheetah Studbook (L. Marker & Johnston, 2022) and confirmed expected familial relationships. Therefore, for subsequent analyses (unless stated otherwise), the following related individuals (offspring of parents included in the dataset and full siblings) were removed from the VCF using *PLINK* v1.9 (Chang et al., 2015; Purcell et al., 2007): AJU10270, AJU10272, AJU9459, AJU9460, AJU9461, AJU9907, AJU9910 and AJU9890. For all sibling sets except AJU9889 and AJU9890, both parents were included in the sample set so all siblings

were excluded from analyses. For AJU9889 and AJU9890, the individual with highest genomic coverage (AJU9889) was retained.

4.3.5 Estimates of N_e

PLINK v1.9 (Chang et al., 2015; Purcell et al., 2007) was used to remove relatives and generate .ped and .map files from VCF3 for each of the four populations, as well as all samples combined and all wild samples combined. Resultant .map files were manually adjusted using a linkage map from the domestic cat (G. Li et al., 2016). *GONE2* (Santiago et al., 2025) was then run on each of the six sets and effective population size was plotted using Python. The -x flag was used when running *GONE2* on all wild samples as the subpopulations were of roughly equal size.

4.3.6 Population structure and genetic diversity

To investigate population structure between US, Namibian, Tanzanian and South Sudanese cheetahs, a principal component analysis (PCA) was generated from VCF5 (see Table 4.2) using *PLINK* v1.9 (Chang et al., 2015; Purcell et al., 2007) and visualised using Python package *seaborn* (Waskom, 2021). PCAs were also run on all individuals (VCF4, see Table 4.2), on the US population only and on the US and Namibian individuals. *ADMIXTURE* v1.3.0 (Alexander et al., 2009) was run on the same set of SNPs from unrelated cheetahs (VCF5, see Table 4.2) to perform ancestry runs for K values 1-5. Pairwise F_{ST} values were computed between populations using the same set of SNPs with *VCFtools* v0.1.16 (Danecek et al., 2011).

To generate a population tree, VCF4 (see Table 4.2) was converted to PHYLIP format using *vcf2phylip* v2.0 (Ortiz, 2019). A population tree was generated from this PHYLIP file using *RAxML-NG* v0.8.0 (Kozlov et al., 2019). Model evaluation was performed to select the optimal model, GTR+G, and the tree was generated with 1,000 bootstraps. To estimate genetic diversity and inbreeding of each population, *VCFtools* v0.1.16 (Danecek et al., 2011) was used to calculate inbreeding coefficient (F_{IS}), SNP heterozygosity (observed heterozygosity at segregating sites), Tajima's D and nucleotide

diversity (π) on VCF2 (see Table 4.2). Population-specific SNPs were identified using *BCFtools* v1.22 (Danecek et al., 2021) and plotted using R package *ggVennDiagram* (Gao et al., 2024).

Runs of homozygosity (ROH) were identified from VCF3 (see Table 4.2) using *BCFtools* v1.22 (Danecek et al., 2021) with a recombination rate of 1.9×10^{-8} per bp per generation, based on estimations in the domestic cat (*Felis catus*) (G. Li et al., 2016). The total length of ROH ($\mathrm{S_{ROH}}$) was plotted against the total number of ROH ($\mathrm{N_{ROH}}$) per individual. The distribution of the fraction of the callable genome in ROH (F_{ROH}) was plotted per population for ROH $>$ 1 Mb in order to exclude short ancestral ROH and focus the analysis on more recent inbreeding that may differ between populations. All plots were generated using Python package *seaborn* (Waskom, 2021). IDrisk (Inbreeding Depression risk), a statistic combining long ROH with non-ROH heterozygosity to predict risk of inbreeding depression, was calculated following the method by Kyriazis et al. (2025).

4.3.7 Deleterious coding variants

SnpEff v5.3 (Cingolani et al., 2012a) was used to predict the functional impacts of variants in VCF1 (see Table 4.2) and to identify potentially deleterious mutations. Variants were categorised by predicted impact using *SnpSift* (Cingolani et al., 2012b). Variants classified as having a high or moderate predicted impact were extracted and corresponding gene names were identified. From this, *Ensembl BioMart* v115 (Cunningham et al., 2022) was used to extract corresponding human Ensembl gene IDs.

Orthofinder v2.5.4 (Emms et al., 2025) was run on the longest protein isoform per gene from the most recent cheetah (*A. jubatus*), human (*Homo sapiens*), dog (*Canis lupus familiaris*) and cat reference genomes (GCA_027475565.2, GCA_000001405.29, GCA_014441545.1, GCA_018350175.1, respectively). One-to-one (1:1) orthologs between the human and cheetah were identified and used as the background set for Gene Ontology (GO) enrichment analysis. The foreground set of genes were those with a corresponding 1:1 ortholog and a high or moderate impact SNP. GO enrichment analysis was run for the high and moderate impact SNPs separately, using *ShinyGO* v0.85 (Ge

et al., 2020) with an enrichment false discovery rate (FDR) corrected q-value cut-off of 0.05. Population-specific SNPs with high and moderate impact were identified using *BCFtools* v1.22 (Danecek et al., 2021) and plotted using R package *ggVennDiagram* (Gao et al., 2024).

To determine the prevalence of previously identified premature termination codons (see Chapter 2, Peers et al. (2025)), reads for all samples were mapped to the previous cheetah reference genome, aciJub1 (GCA_001443585.1; Dobrynin et al. (2015)) (except SRR27375chewbacca, which is the sample used for this assembly) using *BWA-MEM* v0.7.17 (H. Li, 2013). Resultant BAM files were merged with *SAMtools* v1.18 (Danecek et al., 2021). Read groups were added using *GATK* v4.6.0.0 (McKenna et al., 2010) and BAM files indexed with *SAMtools* v1.18 (Danecek et al., 2021). Joint genotyping was completed using *BCFtools* v1.10.2 *mpileup* (Danecek et al., 2021) and the resultant BCF file was converted to VCF format using *BCFtools view*. Filters were not applied as depth was calculated separately using *BCFtools query*. The 65 loci previously identified (see Chapter 2, Peers et al. (2025) were extracted alongside their genotype and per-sample depth using *BCFtools view* and manually examined to determine prevalence in the population data.

4.3.8 Deleterious non-coding variants

To predict the biological impact of variants occurring within functional non-coding regions of the cheetah genome, the 292,912 251 bp windows predicted by ExplaiNN in Chapter 3 were used. SNPs in VCF1 (see Table 4.2) which fell within these windows were extracted. As tools like *SnpEff* are not applicable for non-coding variants, several other methods were applied to predict functional impact of these SNPs.

Combined Annotation Dependent Depletion (CADD) scores quantify the potential deleteriousness of genetic variants using genomic annotations to predict their functional impact. Human CADD scores (hCADD) were downloaded from *CADD* v1.7 (Schubach et al., 2024) and lifted over from the human to the cheetah reference genome by Dr Sam Speak using *LoadLift* (Speak et al., 2024). As this method relies on the use of ultra-conserved elements (UCEs), it was not possible to generate a CADD score for every

SNP falling within a predicted functional non-coding region. Therefore, precomputed *NCBoost* scores were also applied (Caron et al., 2019). *NCBoost* (Caron et al., 2019) uses supervised learning with gradient tree boosting to assess pathogenic non-coding variants associated with monogenic Mendelian diseases. A chain file from the human to cheetah reference genomes was generated using *Cactus* v2.9.8-gpu *hal2chains* on the hal alignment file generated in Chapter 3 (J. Armstrong et al., 2020; Hickey et al., 2013). Using this, pre-computed NCBoost scores for over 850 million positions in the human genome were lifted over to the cheetah genome with *UCSC LiftOver* (Caron et al., 2019; Kuhn et al., 2013).

The nearest downstream gene for each SNP was identified using *BEDtools* v2.31.1 *closest* (Quinlan & Hall, 2010) in a strand-aware manner. *Ensembl BioMart* v115 (Cunningham et al., 2022) was then used to extract the gene IDs of corresponding orthologues in the human genome. Using both sets of score annotations, putative highly deleterious SNPs were extracted. These were classified as SNPs with either a CADD or NCBoost score greater than the 75th percentile for each score type. The closest gene to each putative highly deleterious SNP was extracted and those with a corresponding 1:1 ortholog in the human were retained and used as the foreground for GO enrichment analysis. GO enrichment analysis was run for each set of genes (CADD and NCBoost) separately and in combination, using *ShinyGO* v0.85 (Ge et al., 2020) with an enrichment false discovery rate (FDR) corrected q-value cutoff of 0.05. Unique putative high-impact SNPs were extracted for each population to determine the proportion of unique non-coding mutations. *FIMO* (MEME suite v5.1.0) (Grant et al., 2011) was used to scan the resultant set of windows for transcription factor binding sites using the JASPAR 2024 Vertebrate Core Non-redundant motif database (https://jaspar.elixir.no/downloads/). SNPs overlapping a significant FIMO hit for a TFBS were extracted.

4.4 Results

4.4.1 QC, mapping and variant calling

The total number of paired-end reads per individual ranged from 103,469,302 to 494,748,773 (Table S2). Quality trimming removed, on average, 0.8% of bases from Read 1 and 1.6% of bases from Read 2. Mapping efficiency to the reference genome was high, with between 98.9 and 99.9% of reads successfully aligned per individual.

6,896,511 variants were called by *GATK* across 39 individuals. From this, 4,986,770 SNPs were extracted. After *GATK* quality filtering, 4,269,772 SNPs remained, 4,210,600 of which were biallelic. Further filtering of biallelic SNPs using PLINK removed 873,535 sites due to missing genotype data and 1,032,070 sites due to the minor allele threshold, leaving 2,304,995 SNPs. LD pruning removed 1,862,015 sites, leaving 442,980 SNPs. Eight related individuals were removed, leaving 31 individuals.

4.4.2 Kinship

Pairwise kinship coefficients were calculated between all individuals and visualised as heatmaps (Figures 4.2,4.3). Individuals within the US population had relatively higher kinship compared to the other populations, although Namibian individuals also showed weakly positive kinship to the US population (Figure 4.2). In comparison, South Sudanese and Tanzanian cheetahs showed lower kinship. Within the captive US population (Figure 4.3), pairwise kinship values were generally low, but elevated kinship was observed within known family groups identified in the cheetah studbook. For example, kinship values of 0.25 (suggesting first-degree relatives) were calculated between AJU10270 and AJU10272, as well as between these individuals and AJU7225 and AJU8965. This is consistent with the studbook record listing AJU10270 and AJU10272 as offspring of AJU7225 and AJU8965.

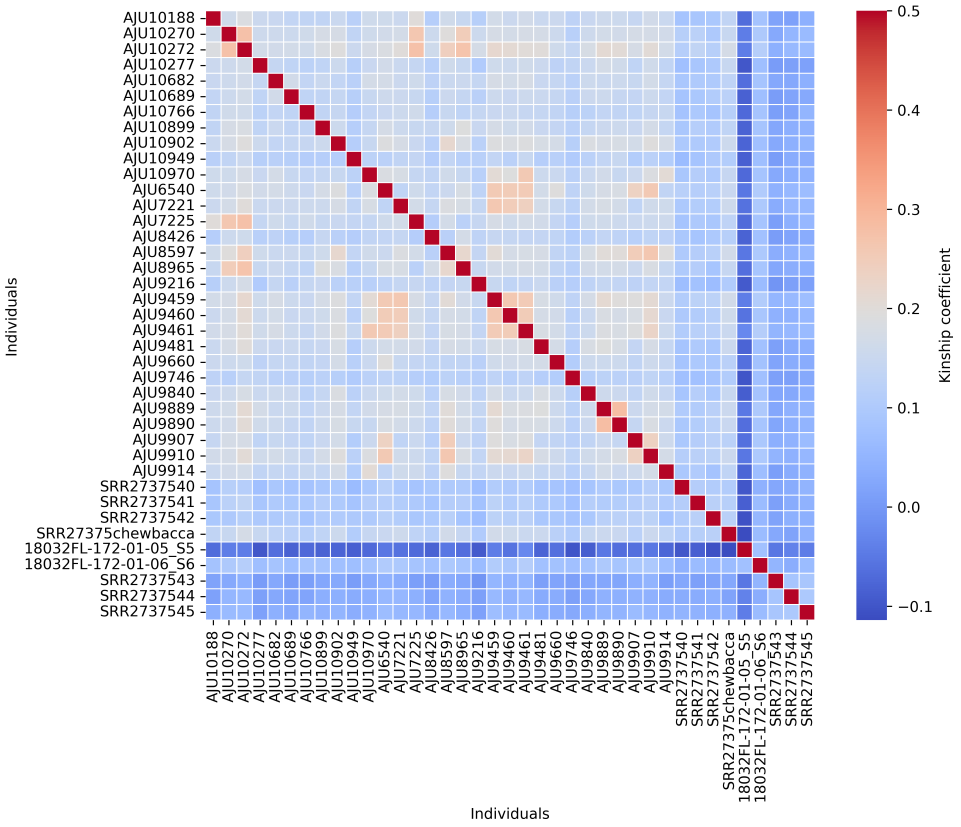


Figure 4.2: **Pairwise kinship values for all cheetahs in this study.** Each cell represents the estimated kinship coefficient between two individuals, with red indicating higher relatedness and blue indicating negative values, which suggest unrelatedness. The diagonal represents self-relatedness (kinship coefficient = 0.5). High kinship is observed within the US, Namibia and Tanzania populations and US and Namibian cheetahs share weakly positive kinship. Negative kinship is observed between South Sudan and Tanzanian populations compared to US and Namibian cheetahs.

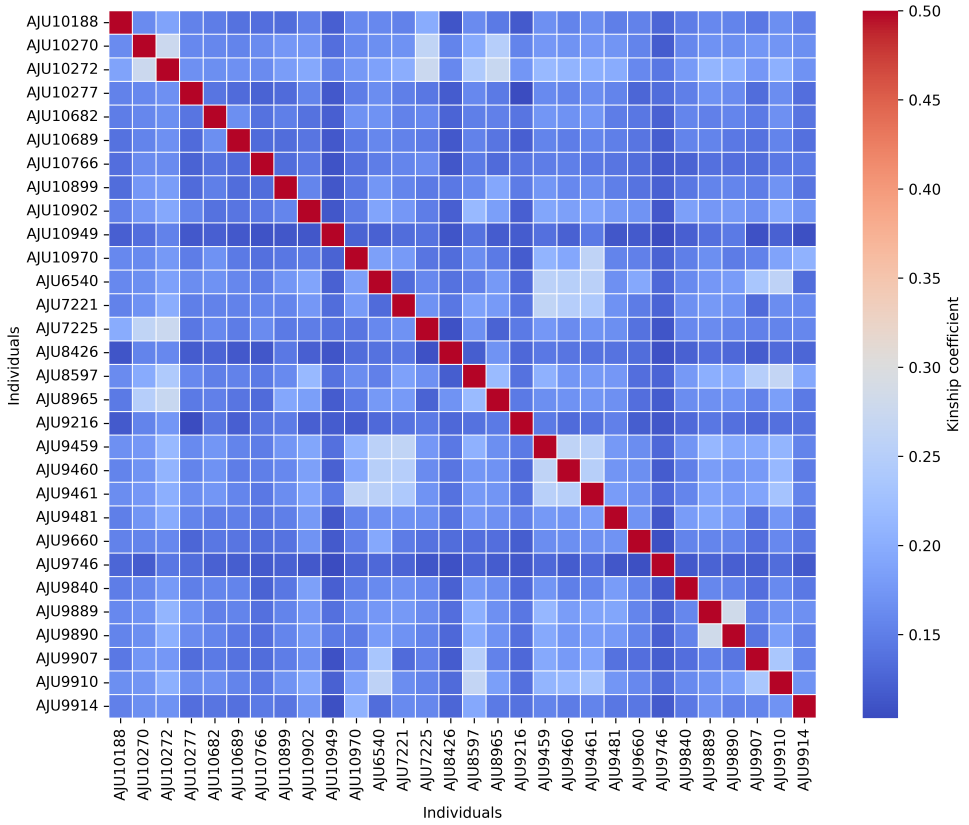


Figure 4.3: **Pairwise kinship values for US captive cheetahs.** Each cell represents the estimated kinship coefficient between two individuals, with red indicating higher relatedness and blue indicating negative values, suggesting unrelatedness. The diagonal represents self-relatedness (kinship coefficient = 0.5). Kinship between most pairs of individuals is low, but first degree family relationships can be observed between known family groups (see Figure 4.1).

4.4.3 Estimates of N_e

Effective population size over the last 150 generations was estimated and showed steep recent population contractions in each of the populations, ranging from 5 (all samples) to 88 (Namibia) generations ago (Figure 4.4). Estimates of N_e pre-contraction were as follows: 1,689 (all); 7,891 (wild); 1,276 (US); 1,295,270 (Namibia); 3,957,060 (South Sudan); and 1,643,780 (Tanzania). Inflated estimates of N_e in Namibia, South Sudan and Tanzania are likely due to the insufficient number of samples used as input to *GONE2*.

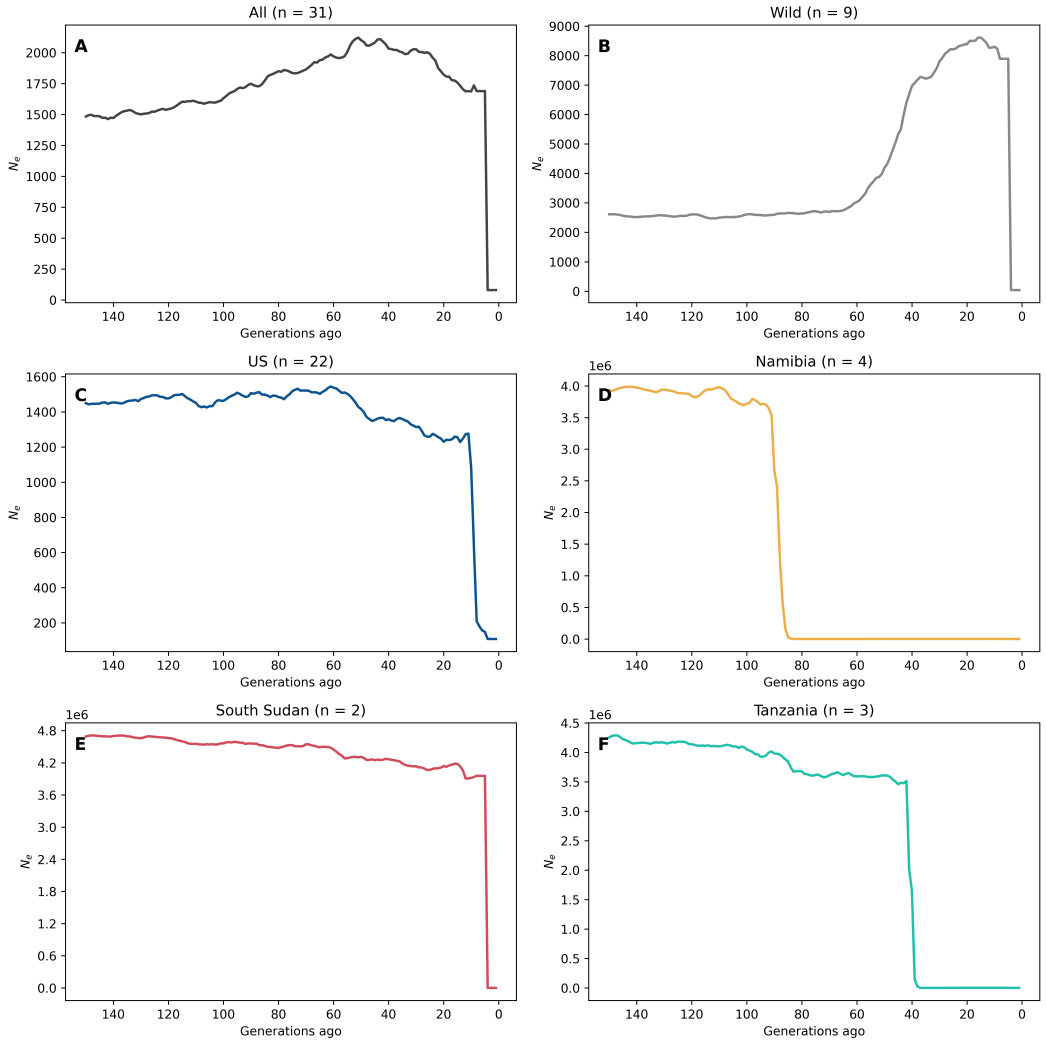


Figure 4.4: Estimates of effective population size (N_e) for each cheetah population for the last 150 generations calculated by *GONE2*. Estimated N_e for each population is shown: (A) 31 unrelated cheetahs, (B) nine wild cheetahs, (C) 22 unrelated US cheetahs, (D) four Namibian cheetahs, (E) two South Sudanese cheetahs, (F) three Tanzanian cheetahs. Every population (or combination of populations) shows a decrease in N_e between 5 and 88 generations ago.

4.4.4 Population structure and genetic diversity

Distinct separation was identified between South Sudanese, Tanzanian, and Namibian and US cheetahs (Figure 4.5). PC1, separating South Sudan and Tanzania from Namibia and the US, represents 12.91% of diversity. PC2, which shows variation between the African populations, represents 8.25% of diversity, with remaining PCs representing less than 6% each (Figure S1). A PCA with all individuals, including relatives, showed a

consistent pattern (Figure S2). PCAs of the US and Namibian populations (Figure S3,4) showed some separation of Namibian individuals from the US population, however the variation between US and Namibian individuals was less than that observed within the US population.

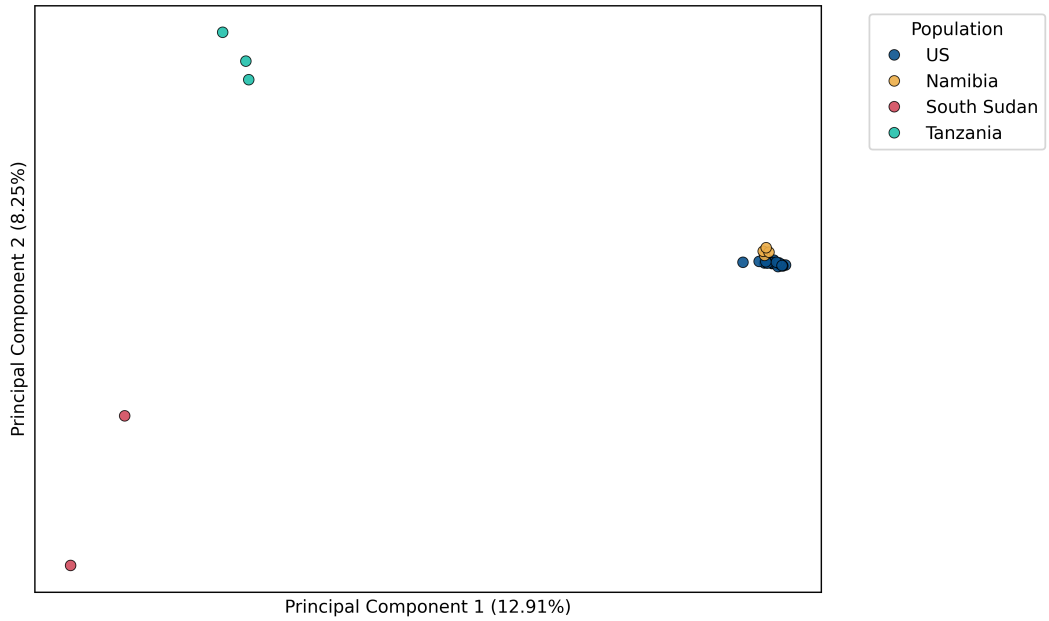


Figure 4.5: **Principal Component Analysis (PCA) of unrelated cheetahs.** PCA based on filtered genome-wide SNPs showing genetic clustering by population: US (blue), Namibia (yellow), South Sudan (red), and Tanzania (green). The percentage of variance explained by each principal component is shown in axes labels.

Admixture cross validation (CV) errors for K = 1 to 5 were 0.54, 0.60, 0.69, 0.89 and 0.89, respectively with the lowest error at K = 1, suggesting no population structure (Figure 4.6). For K > 1, the Tanzanian and South Sudan cheetahs form distinct clusters from the US and Namibian populations, consistent with PCA results (Figure 4.5). Across all analyses except K = 4, Tanzanian and South Sudanese individuals consistently cluster together. At K = 4, one Tanzanian individual clusters with the South Sudan cluster, although this pattern is not observed when relatives are included in the analysis (Figure S5).

Figure 4.6: **ADMIXTURE-derived ancestry clustering of unrelated cheetahs.** Model-based clustering of cheetah populations at K = 2–5 using only unrelated individuals. Each vertical bar represents an individual, and colours represent the proportion of ancestry assigned to each genetic cluster. Admixture cross-validation errors for K = 1 to 5 were 0.54, 0.60, 0.69, 0.89 and 0.89, respectively.

Genetic differentiation between populations, indicated by pairwise F_{ST} values (Figure 4.7), mirrored the pattern observed with PCA (Figure 4.5). The US and Namibian populations showed the lowest pairwise differentiation ($F_{\mathrm{ST}} = 0.042$), suggesting substantial shared ancestry. Differentiation between the US and African populations as a whole was moderate ($F_{\mathrm{ST}} = 0.046$), likely driven by the relationship between the US and Namibia. The highest level of pairwise differentiation was observed between South Sudan and Tanzania ($F_{\mathrm{ST}} = 0.175$), suggesting strong genetic divergence.

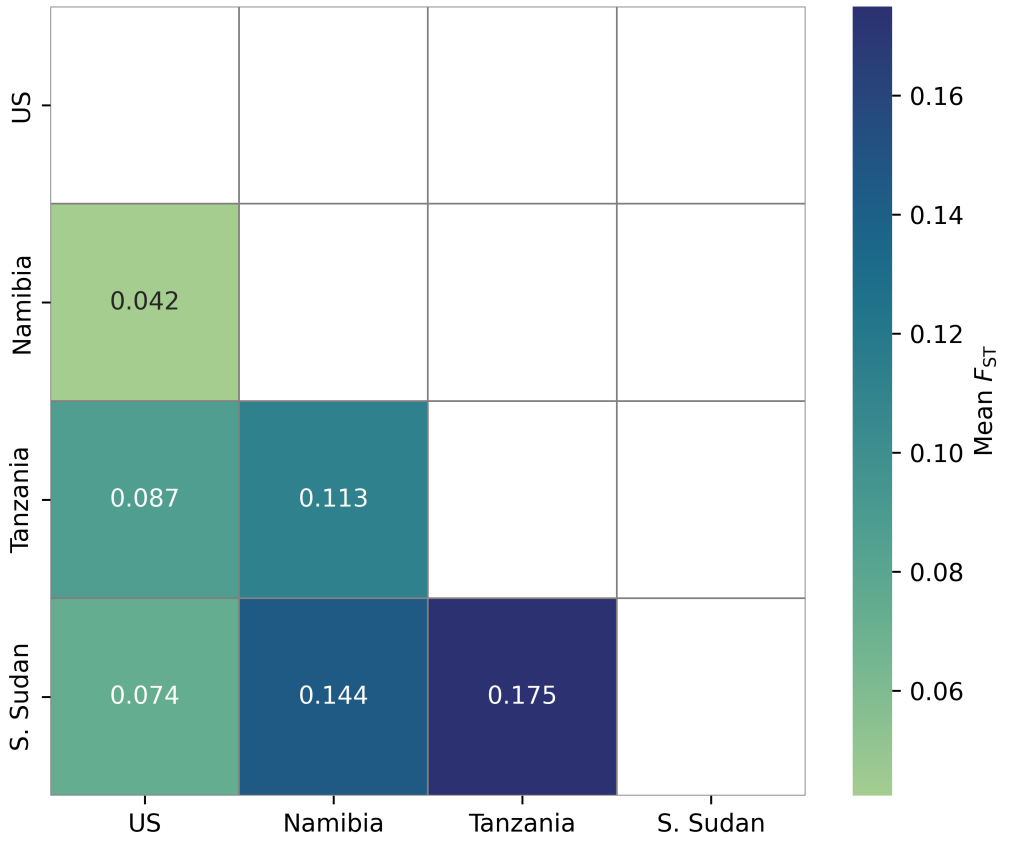


Figure 4.7: **Pairwise genetic differentiation between cheetah populations.** Heatmap showing mean pairwise Weir and Cockerham's F_{ST} values among cheetah populations. Darker colours represent greater genetic differentiation, with lighter colours indicating higher genetic similarity.

Phylogenetic analysis of genome-wide SNPs provided concurrent evidence of population structure, similar to that observed in PCA and ADMIXTURE analyses (Figure 4.8). Clear separation of South Sudanese and Tanzanian cheetahs from US and Namibian individuals was observed, with high bootstrap support for most branches, including those separating South Sudan from Tanzania and the Namibian individuals from the US.

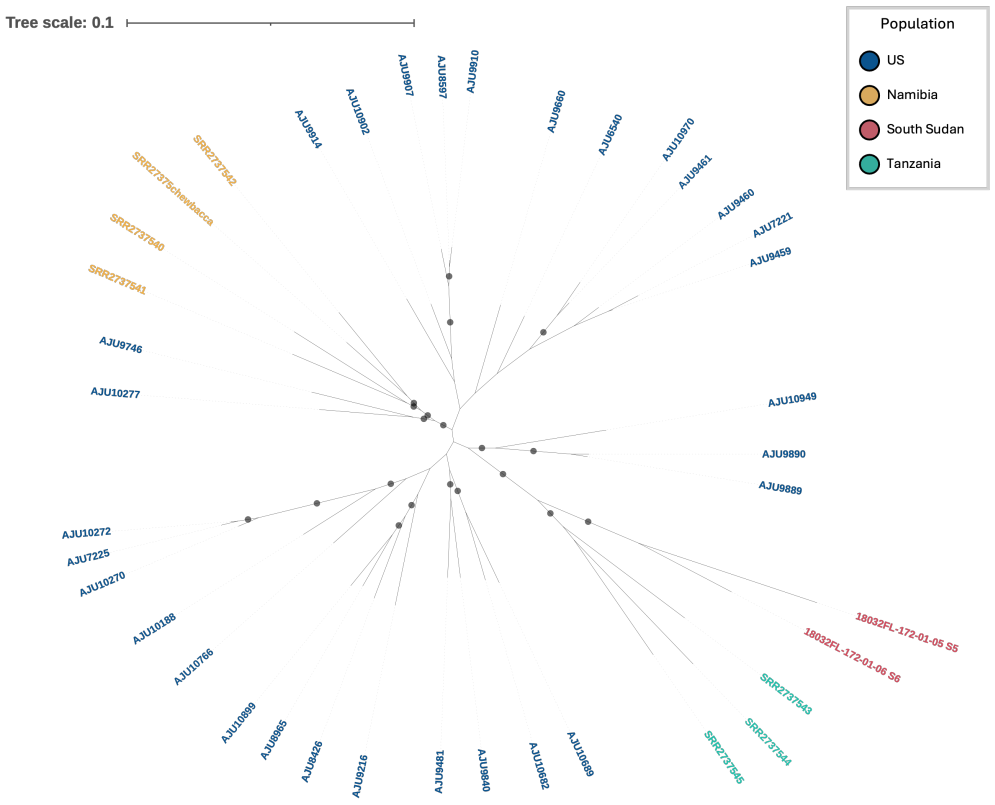


Figure 4.8: **Maximum-likelihood phylogeny of 39 cheetahs based on genome-wide SNPs.** Samples are coloured by population of origin: US (blue), Namibia (yellow), South Sudan (red), and Tanzania (green). The scale bar indicates substitutions per site. Branches with bootstrap values >90% are marked with a grey circle.

Population genetic diversity measures showed an overall pattern of higher inbreeding and lower genetic diversity in the South Sudan population (Figure 4.9). F_{IS}, a measure of inbreeding, was comparatively lower in the US and Namibian populations, however this result was not significant (Kruskal-Wallis: H=0.65, p=0.88). SNP heterozygosity was highest in South Sudan, although this is likely an artefact of the small sample size in this population. Across populations, SNP heterozygosity differed significantly (Kruskal-Wallis: H=19.20, p=2.49×10^{-4}), with the US population having significantly lower observed heterozygosity than each of the three African populations (Bonferroni-corrected Wilcoxon rank-sum tests: US vs Namibia: W = 88, p = 8.03×10^{-4}; US vs South Sudan: W = 44, p = 4.35×10^{-2}; US vs Tanzania: W = 66, p = 5.21×10^{-3}). Tajima's D was positive in all populations, suggesting a population contraction or bottleneck, and was highest in the US population, consistent with a founder effect and population contraction induced by captivity. A significant overall difference was

identified (Kruskal-Wallis: H=5919.46, p $<$ 0.001) and pairwise Wilcoxon rank-sum tests identified a significant difference between all pairs of populations (Table S3). π was also significantly different between populations (Kruskal-Wallis: H=4486.73, p $<$ 0.001), with significant differences between all pairs of populations except the US and Tanzania (Table S3).

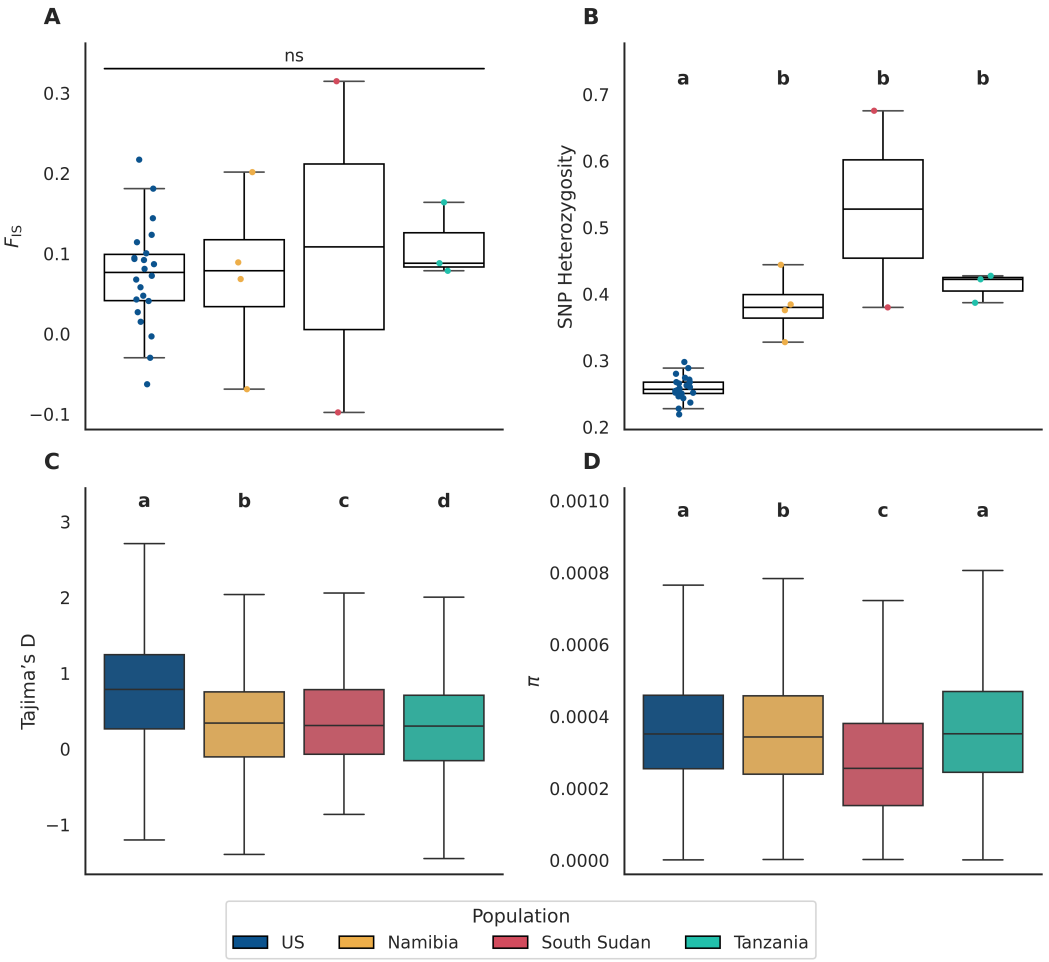


Figure 4.9: **Genome-wide estimates of genetic diversity across cheetah populations.** Boxplots showing population-level estimates of (**A**) inbreeding coefficient (F_{IS}), (**B**) SNP heterozygosity, (**C**) Tajima's D, and (**D**) nucleotide diversity (π). Colours indicate populations: US (blue), Namibia (yellow), South Sudan (red), and Tanzania (green). Significant results (Kruskal-Wallis, p $>$ 0.05) are indicated by labels (a, b, c, d) for each plot, with "ns" representing results with no signficant difference. See Table S3 for full significance values.

Population-specific SNPs were identified in all populations, with most SNPs unique to the US population (Figure 4.10). The Namibian population had the fewest unique SNPs, with almost 95% of Namibian SNPs also identified in the US population.

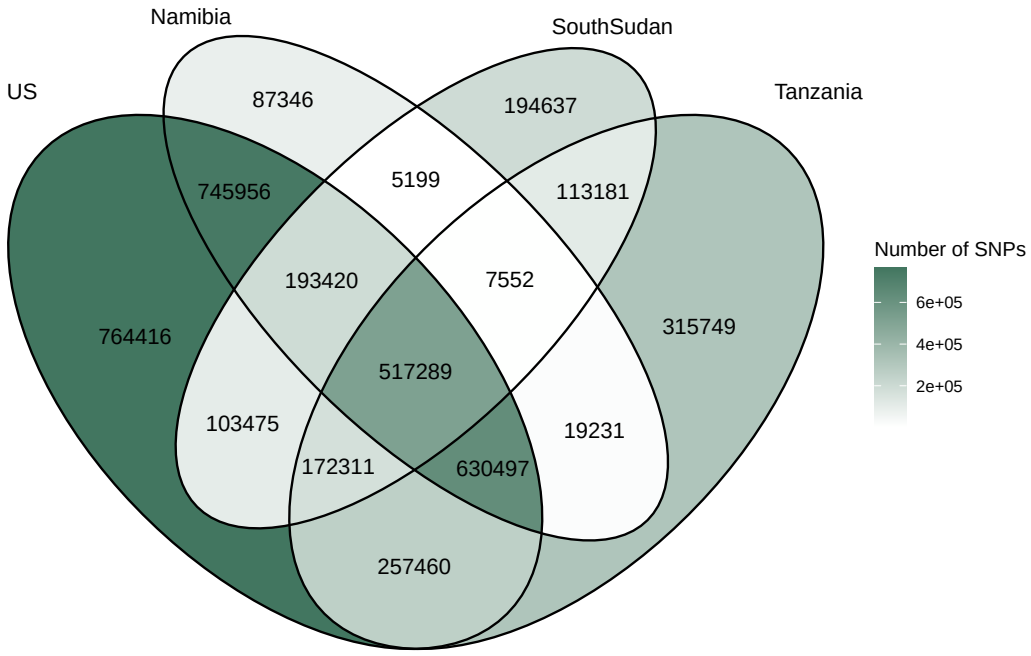


Figure 4.10: **Distribution of SNPs across the four cheetah populations in this study.** The colour of each segment indicates the number of SNPs shared between those populations, with darker green indicating a higher number. The analysis includes 4,142,757 unfiltered SNPs from VCF1; details of the VCF filtering steps are provided in Table 4.2.

Runs of homozygosity showed highest inbreeding in South Sudan (Figure 4.11). South Sudanese individuals, closely followed by Tanzanian individuals, had a large number of ROH ($\mathrm{N_{ROH}}$) and the longest total ROH. There were generally fewer, and shorter, ROH in the US and Namibian populations, with several outliers. Individual AJU9216 had a higher quantity of short and long ROH than most other samples, whilst individual AJU7225 had a notably higher quantity of ROH, particularly in runs of over 5 Mb. There was a significant overall difference in $\mathrm{F_{ROH}}$ among populations (Kruskal–Wallis test, H $= 11.70$, p $= 0.0085$), although no individual pairwise comparison was significant after Bonferroni correction (Mann–Whitney U tests: all adjusted p > 0.05). IDrisk scores showed low risk of population extinction in all populations, based on the score thresholds provided by Kyriazis et al. (2025) (Figure S6).

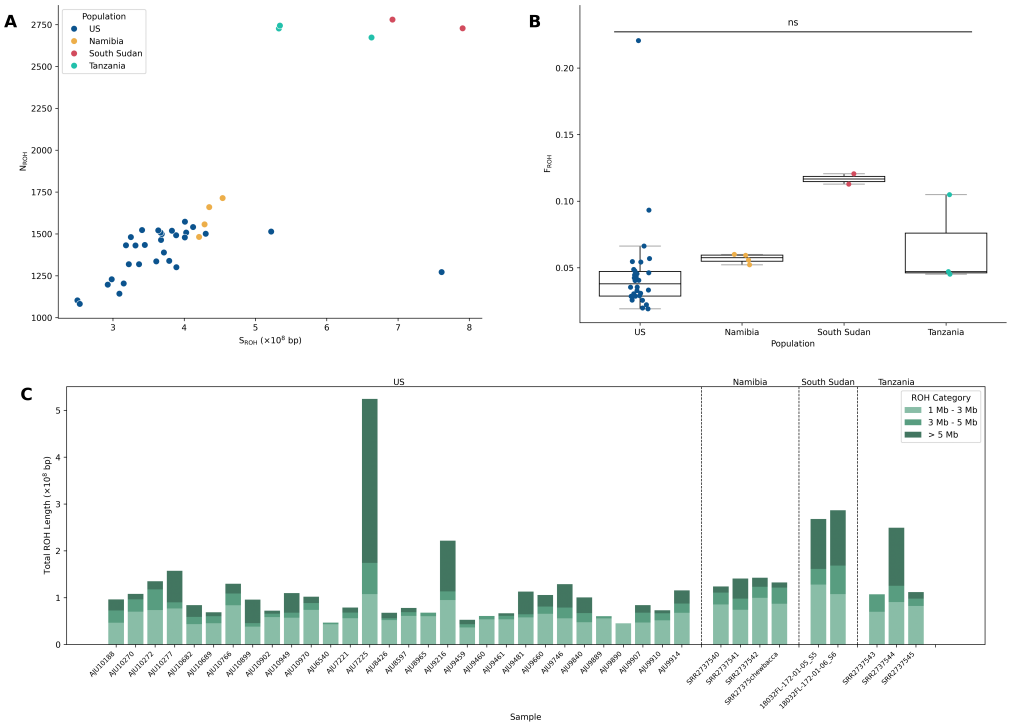


Figure 4.11: **Runs of homozygosity (ROH) across cheetah populations.** (**A**) The relationship between the number of ROH segments ($\mathrm{N_{ROH}}$) and the sum of ROH lengths ($\mathrm{S_{ROH}}$), where colours correspond to populations: US (blue), Namibia (yellow), South Sudan (red), and Tanzania (green). (**B**) Fraction of the genome in ROH ($\mathrm{F_{ROH}}$) per population, with a data point for each individual overlaid on the boxplot. Colours correspond to populations as shown in (A). Pairwise Mann-Whitney U tests found no significance (indicated by "ns" label), as all Bonferroni-corrected p-values > 0.05. (**C**) Total ROH length per individual, grouped by ROH length category 1-3 Mb, 3-5 Mb, >5 Mb.

4.4.5 Deleterious coding variants

SnpEff predicts the impact of each SNP and categorises it as either 'high', 'medium', 'low' or 'modifier', with the possibility of multiple categories per SNP as SnpEff considers the impact of a SNP in all transcripts it overlaps. High impact SNPs are likely to severely impact gene function and therefore be deleterious. SnpEff categorised the impact of 4,155,005 SNPs as follows: 684 high impact, 18,448 moderate impact, 27,506 low impact and 4,142,676 modifiers (variant is outside coding regions or its impact is unknown).

703 annotated genes were associated with the 684 high-impact SNPs. Of these,

562 had a 1:1 ortholog in the human and so were used as the foreground in GO enrichment analysis. Eleven Biological Process GO terms were significantly (FDR-corrected q $<$ 0.05) overrepresented in this gene list (Figure 4.12): sperm flagellum assembly (GO:0120316), axoneme assembly (GO:0035082), cilium movement involved in cell motility (GO:0060294), cilium or flagellum-dependent cell motility (GO:0001539), cilium-dependent cell motility (GO:0060285), cilium movement (GO:0003341), cilium assembly (GO:0060271), cilium organization (GO:0044782), microtubule-based movement (GO:0007018), plasma membrane bounded cell projection assembly (GO:0120031), and organelle assembly (GO:0070925).

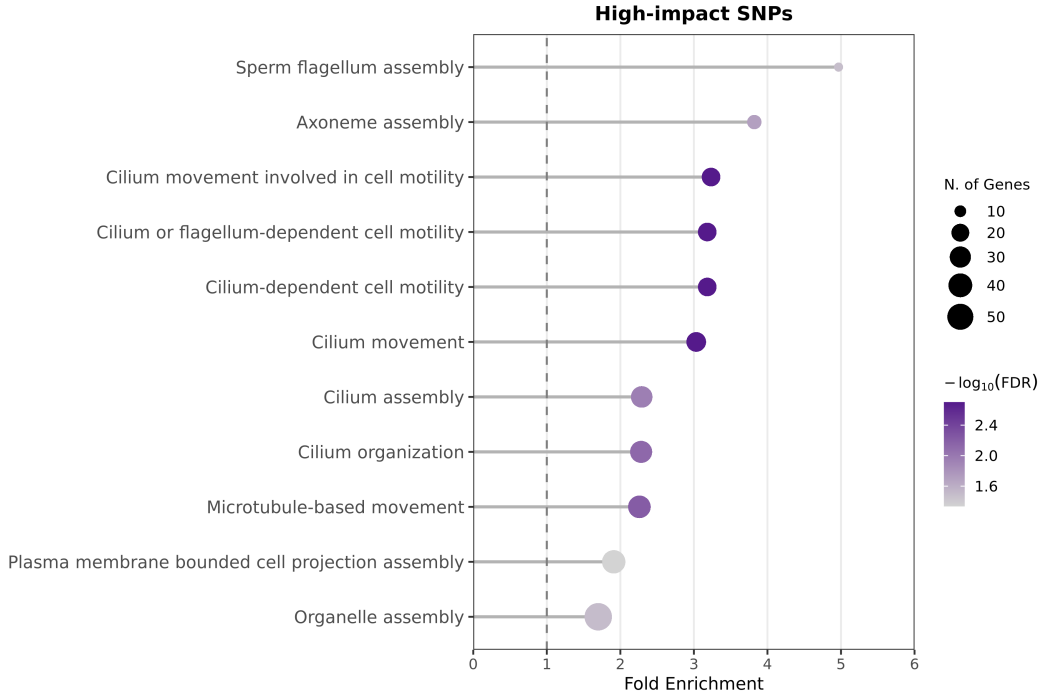


Figure 4.12: **Gene ontology (GO) enrichment of genes containing predicted high-impact deleterious SNPs.** GO Biological Process terms significantly enriched (FDR-corrected q-value $<$ 0.05) among genes containing high-impact SNPs across all the cheetah populations. Fold enrichment is shown on the x-axis, with the size of the point reflecting the number of genes annotated with each GO term. The colour of each point shows the statistical significance ($-log_{10}$ FDR) of the enrichment of each GO term, where darker points are more significantly enriched. A fold enrichment score of 1, shown by a dotted line, means no enrichment in the gene set. Enriched terms are largely associated with cilium and flagellum assembly and motility.

The same analysis was then completed for SNPs with moderate impact. 7,818 genes were associated with the 18,448 moderate impact SNPs. Of these, 6365 had a 1:1 ortholog in the human; this set of genes was used as the foreground for this

GO enrichment analysis. 143 Biological Process GO terms were significantly enriched (FDR-corrected $q < 0.05$) in this gene set. The top 20 most significant (Figure 4.13) were: axoneme assembly (GO:0035082), microtubule bundle formation (GO:0001578), cilium movement (GO:0003341), cilium or flagellum-dependent cell motility (GO:0001539), cilium-dependent cell motility (GO:0060285), cilium movement involved in cell motility (GO:0060294), cilium organization (GO:0044782), flagellated sperm motility (GO:0030317), sperm motility (GO:0097722), cilium assembly (GO:0060271), microtubule-based movement (GO:0007018), microtubule cytoskeleton organization (GO:0000226), microtubule-based process (GO:0007017), cell projection assembly (GO:0030031), cellular process involved in reproduction in multicellular organism (GO:0022412), plasma membrane bounded cell projection assembly (GO:0120031), organelle assembly (GO:0070925), cytoskeleton organization (GO:0007010), cell projection organization (GO:0030030), and organelle organization (GO:0006996).

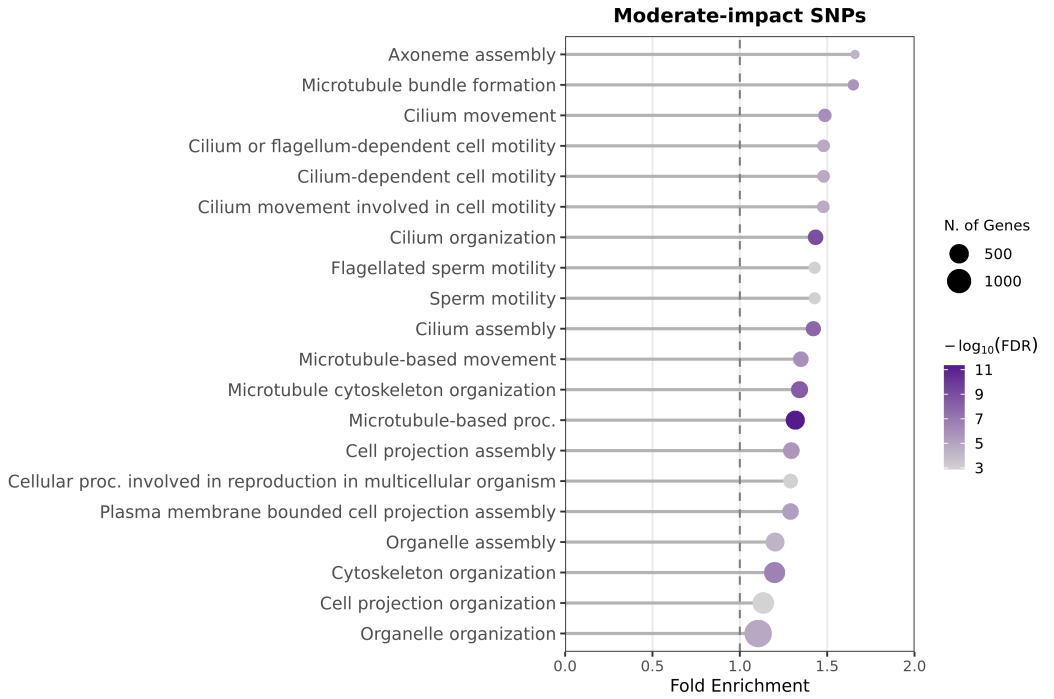


Figure 4.13: **Gene ontology (GO) enrichment of genes containing predicted moderate impact deleterious SNPs.** GO Biological Process terms significantly enriched (FDR-corrected q-value < 0.05) among genes containing moderate-impact SNPs across all the cheetah populations. Fold enrichment is shown on the x-axis, point size corresponds to the number of genes annotated with each GO term and the colour of each point shows the statistical significance ($-log_{10}$ FDR) of the enrichment, where darker points are more significantly enriched. A fold enrichment score of 1, shown by a dotted line, means no enrichment in the gene set. Enriched terms are largely associated with cilium and flagellum assembly and motility, particularly of sperm, and microtubule organization.

High and moderate impact SNPs showed a similar population distribution as the total set of SNPs, with each population containing at least 41 unique SNPs with predicted high impact and at least 528 unique SNPs with predicted moderate impact (Figure 4.14A,C). When comparing these values to overall numbers of population-specific SNPs, the Namibian population contained a higher proportion of unique high-impact SNPs than other populations, whilst Tanzania contained a lower proportion of unique high- and medium-impact SNPs (Figure 4.14B,D).

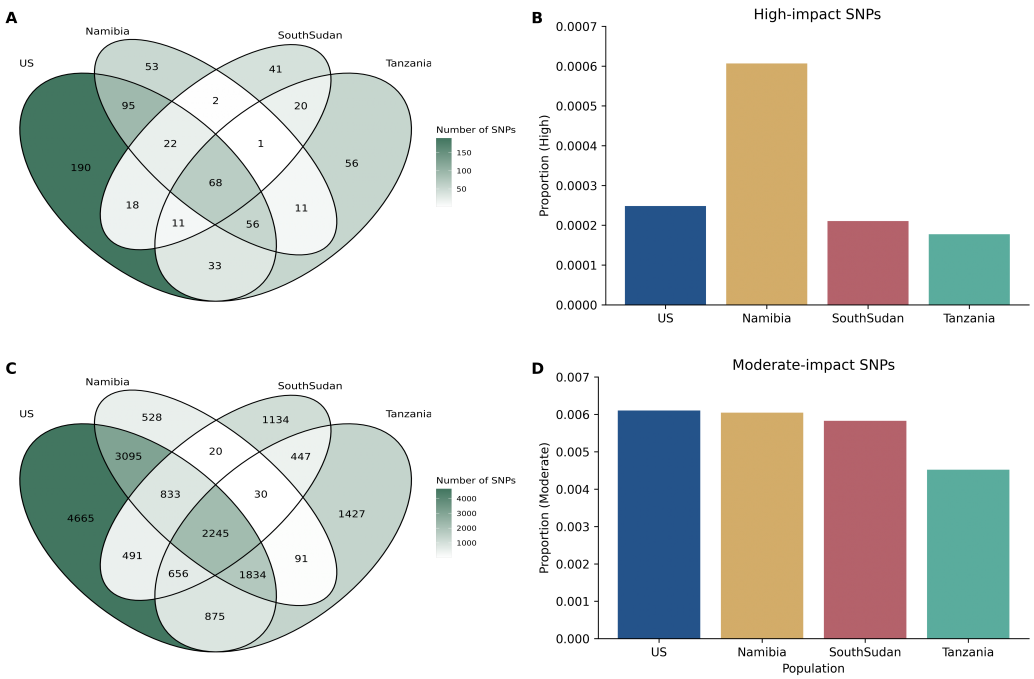


Figure 4.14: **Distribution of predicted high- and moderate-impact SNPs between populations.** (**A**) Distribution of high-impact SNPs predicted by SnpEff between the four populations. Number of SNPs specific to each combination of populations is shown in each segment, with a darker colour indicating a higher number of SNPs. (**B**) Unique high-impact SNPs as a proportion of the overall number of unique SNPs per population (see Figure 4.10). Panels **C** and **D** show the same information as panels A and B, respectively, for moderate-impact SNPs.

Allele frequency was calculated for the 684 high-impact coding SNPs and pairwise Mann-Whitney U tests found no significant difference between populations (Figure 4.15A; all Bonferroni-corrected p-values > 0.05). The majority of high-impact SNPs had a low allele frequency (< 0.2), with similar patterns observed in moderate- and low-impact SNPs (Figures S4.7,S4.8). The 64 SNPs with an allele frequency over 0.9 were extracted and their distribution across populations was plotted (Figure 4.15B). A large proportion of high-frequency SNPs unique to South Sudan was observed, likely due to the small sample size. 40 unique genes were associated with these high-frequency, high-impact SNPs, and GO enrichment analysis found no significant functional enrichment (FDR-corrected q > 0.05) of these genes.

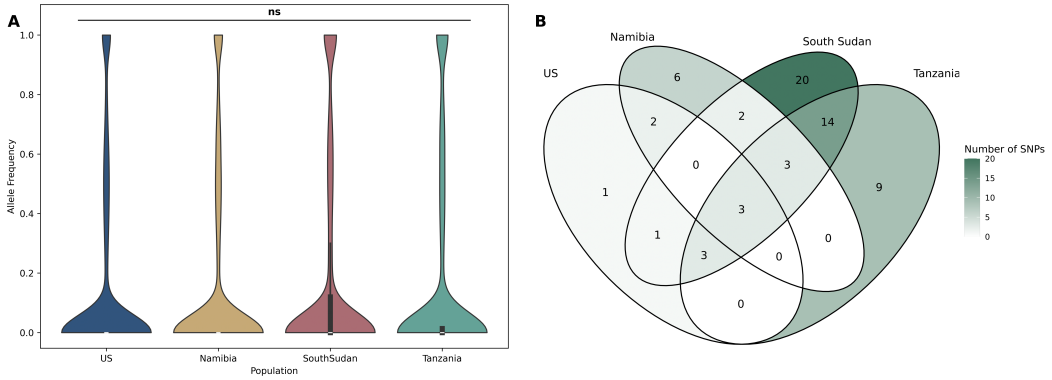


Figure 4.15: **Allele frequency of high-impact SNPs in coding regions.** (**A**) Violin plot showing the allele frequency of all high-impact SNPs per population. "ns" indicates that no significant result was identified by pairwise Mann-Whitney U tests (Bonferroni-corrected p-values > 0.05). A frequency of 0 indicates that the variant is not present in the given population but is present elsewhere in the dataset. (**B**) The distribution of 64 high-impact SNPs with an allele frequency > 0.9 across the populations. A darker colour indicates a higher number of SNPs.

Of the 65 previously identified premature termination codons (PTCs) (see Chapter 2, Peers et al. (2025)), 36 were not found in any population data and were therefore considered unique to the reference genome used in that analysis (aciJub1, GCA_001443585.1, Dobrynin et al. (2015)). Two loci did not have sufficient coverage in the present analysis to determine population distribution. 23 loci were homozygous for the reference allele in all individuals, meaning the PTC-causing mutation is fixed across all samples. The remaining four loci had some variation in the population data. Two loci were identified as indels in the present analysis. One mutation, in the gene FHL5, was only found in eight captive samples and was otherwise not present in the population data. The final mutation, in gene DEFB116, was fixed in all but two captive samples, a mother and daughter (AJU8965 and AJU10270).

4.4.6 Deleterious non-coding variants

Across all individuals, 79,025 SNPs fell within predicted functional non-coding windows. Of these, 7,103 SNPs had a CADD score and 25,817 SNPs had an NCBoost score,

resulting in a total of 29,357 SNPs with at least one score and 3,563 SNPs with both scores. The distribution of these scores showed a skew towards low CADD scores and a skew towards high NCBoost scores (Figure 4.16). The 25th, 50th and 75th percentiles for CADD scores were 1.15, 3.58 and 7.66 and for NCBoost scores were 0.49, 0.71 and 0.86. Predicted highly deleterious variants, which had a CADD score >7.66 or an NCBoost score >0.86, were extracted. This resulted in 597 SNPs with a high CADD score and 6,882 SNPs with a high NCBoost score, with 105 SNPs falling into both categories.

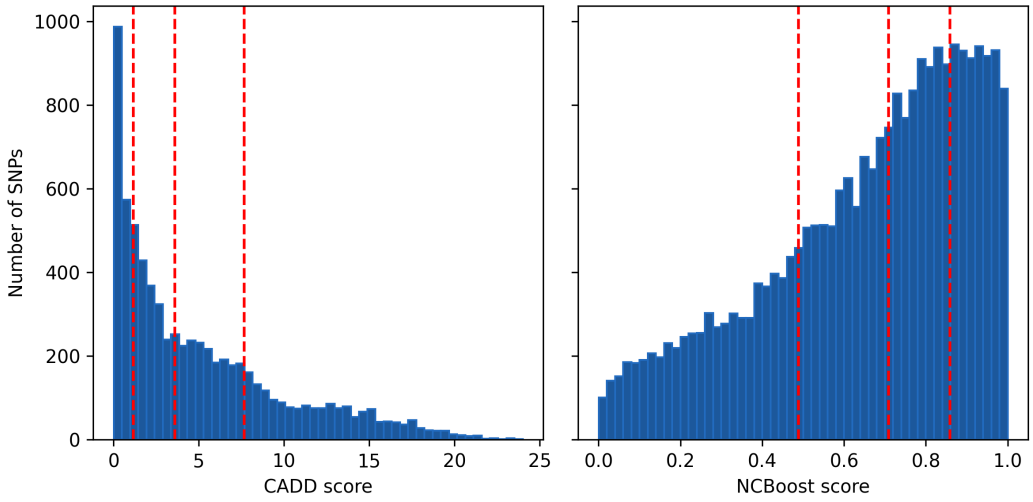


Figure 4.16: **Distribution of CADD (left) and NCBoost (right) scores of SNPs occurring within predicted functional non-coding windows.** Red dashed lines show the 25th, 50th and 75th percentile (from left to right) of the data. The majority of SNPs had a low CADD score, with few highly pathogenic variants, whereas the distribution of NCBoost scores is skewed towards a higher score.

275 of the 597 genes closest to a SNP annotated with a high CADD score and 2,147 of the 6,882 genes closest to a SNP with a high NCBoost score mapped to a human Ensembl ID. GO enrichment analysis identified no significant functional enrichment (FDR-corrected q-value $>$ 0.05) in either set of genes. Within the 7,352 total deleterious SNPs, population distribution was investigated, and a lower proportion of deleterious SNPs was observed in Tanzania (Figure 4.17).

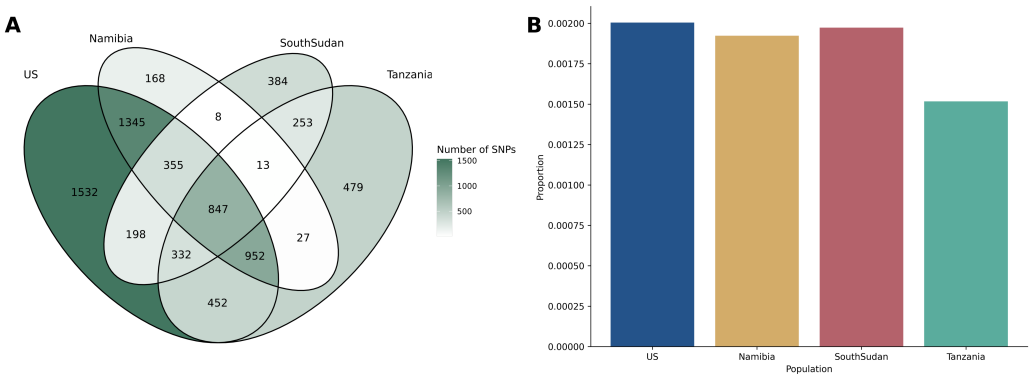


Figure 4.17: **Population distribution of predicted high-impact SNPs within predicted functional non-coding windows.** (**A**) Distribution of high-impact noncoding SNPs, based on NCBoost and CADD scores, between the four populations. Number of SNPs specific to each combination of populations is shown in each segment, with a darker colour indicating a higher number of SNPs. (**B**) Unique high-impact noncoding SNPs as a proportion of the overall number of unique SNPs per population (see Figure 4.10).

Of the 79,025 SNPs within predicted functional non-coding regions, 50,426 overlapped a significant FIMO hit. Of these, 4794 were predicted high-impact SNPs, meaning 65.2% of the predicted high-impact SNPs overlapped a transcription factor binding motif. These SNPs were found in predicted regulatory regions closest to 3024 genes and overlapping 728 different motifs. Of the motifs overlapped by a high-impact SNP, the most impacted were CTCF (CCCTC-binding factor), followed by several TFAP (Transcription Factor Activating enhancer binding Protein) and ZNF (Zinc Finger Protein) motifs (Table S6).

Allele frequency was calculated for the 4794 high-impact non-coding SNPs overlapping a TF motif (Figure 4.18). As with the coding SNPs, the majority of high-impact SNPs had a low allele frequency. 752 SNPs with an allele frequency over 0.9 were extracted and their distribution across populations was plotted. A large proportion of high-frequency SNPs unique to South Sudan was observed, likely due to the small sample size. 347 unique genes were associated with these SNPs. GO enrichment analysis found no significant enrichment (FDR-corrected $q > 0.05$) of these genes associated with a high-frequency, high-impact SNP.

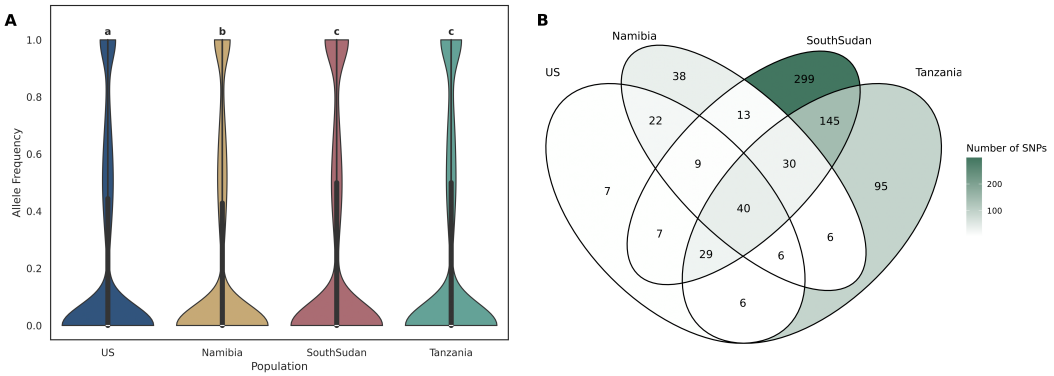


Figure 4.18: **Allele frequency of high-impact non-coding SNPs.** **A** Violin plot showing the allele frequency of all high-impact SNPs per population. A frequency of 0 indicates that the variant is not present in the given population but is present elsewhere in the dataset. All pairwise tests were significant (indicated by labels a, b, c). Mann–Whitney U and Bonferroni-corrected p-values were as follows: US vs Namibia: U = 1,374,855,595.0, $p = 9.17 \times 10^{-4}$; US vs South Sudan: U = 644,232,683.5, $p = 2.65 \times 10^{-11}$; US vs Tanzania: U = 988,335,587.0, $p = 9.38 \times 10^{-7}$; Namibia vs South Sudan: U = 83,456,171.5, $p = 4.79 \times 10^{-15}$; Namibia vs Tanzania: U = 128,047,881.0, $p = 1.50 \times 10^{-10}$; South Sudan vs Tanzania: U = 66,218,724.0, $p = 2.30 \times 10^{-1}$. **B** Distribution of high-impact SNPs with an allele frequency > 0.9 across the populations. A darker colour indicates a higher number of SNPs.

4.5 Discussion

Understanding the distribution of deleterious mutations in and across cheetah genomes and populations can provide us with invaluable information about both the past (by inferring impacts of population contractions) and the future (by considering the long-term impacts of such contractions) of cheetah demography. Here, I identify a significant over-representation of deleterious mutations in sperm-associated genes, particularly in Namibian cheetahs, whilst Tanzanian cheetahs contained the lowest proportion of deleterious SNPs across both coding and non-coding genomic regions. I show low genetic differentiation between captive and Namibian cheetahs, consistent with the proposed origin of the majority of captive cheetahs.

4.5.1 Population structure and relatedness

Firstly, the genomic evidence presented herein corroborates and supports the International Cheetah Studbook, as genomic kinship values reflect the familial relationships recorded in the studbook. This is an important result, as captive breeding pairs are usually selected based on kinship estimates from the studbook. Therefore, ensuring that the studbook is accurate is crucial to prevent accidental inbreeding in captivity.

Population structure analyses consistently showed US and Namibian cheetahs clustering together, with greater differentiation between those and the other African populations. This is consistent with the history of captive cheetahs in the US; the majority of individuals brought into captivity from the wild were sourced from populations in Namibia and South Africa (Marker-Kraus, 1988). There is some evidence that few individuals from Eastern African populations were brought into captivity (Marker-Kraus, 1988), which may explain the lower genetic differentiation between South Sudan and Tanzanian populations with the US. I observe lower F_{ST} values across all populations compared to a previous study based on double-digest restriction site associated DNA (ddRAD) sequencing, despite similar numbers of samples used (Prost et al., 2022). This difference is likely due to the sequence data used; whole-genome data can produce more robust estimates of F_{ST} than ddRAD sequencing (Lou et al., 2021).

Namibian and Tanzanian cheetahs are currently recognised as the same subspecies (*A. j. jubatus*) by the IUCN, although the Tanzanian population was previously classified as *A. j. raineyi* (Krausman & Morales, 2005). Based on population structure analyses, the Namibian and Tanzanian samples used in this study show proportionally high genetic differentiation, supporting Prost et al. (2022)'s call for the IUCN to return to the previous subspecies classification. However, as only four Namibian and three Tanzanian cheetahs were included here, a larger sample set is needed to confirm this observation.

Sequencing for each population was carried out with different sequencing chemistries and at different times, resulting in the potential for a technical batch effect inflating population differentiation (Lou & Therkildsen, 2022). Previous studies have highlighted the importance of sequence read trimming and base recalibration to remove such batch effects in population genetics studies (Leigh et al., 2018; Lou & Therkildsen, 2022;

Shaw et al., 2025). Whilst some data may be lost by trimming longer read lengths to match older data, not trimming these reads would have resulted in lower quality scores for variants in Namibia and Tanzania, biasing the data. Therefore, by applying these techniques, I am confident that the population structure I observe is biological and not an artefact of a sequencing batch effect.

4.5.2 Estimates of N_e

Historic effective population size was estimated to confirm the widely accepted hypothesis that cheetahs have experienced long-term low N_e (Castro-Prieto et al., 2011; Dobrynin et al., 2015; Driscoll et al., 2002; Fabiano et al., 2025; Kim et al., 2016; Menotti-Raymond & O'Brien, 1993). Estimates for wild cheetah populations suggested an N_e of several million per population as recently as five generations ago, corresponding to 25 years based on a generation time of 5 years (Durant et al., 2021; Durant et al., 2017). This is biologically implausible, as large carnivores typically occur at low densities (Carbone & Gittleman, 2002) and the global cheetah population is estimated at 7,100 individuals (Durant et al., 2017). LD-based methods for estimating effective population size, such as *GONE2*, are highly sensitive to sample size (England et al., 2006; Santiago et al., 2020). In this chapter, datasets ranged from two to 31 samples. The inflated N_e estimates observed for the Namibian, South Sudanese, and Tanzanian populations are therefore likely artefacts of small sample size.

Combining samples from multiple populations increases sample size, but introduces additional biases, as differences in sub-population size can distort LD patterns (Santiago et al., 2025). The combination of wild samples took these sub-populations into account, but the data quantity was still low. However, this illustrates the effect of increased sample size, as the estimated N_e was 7,891 approximately five generations ago, which more closely aligns with contemporary estimates of global wild cheetah populations (Durant et al., 2017).

Whilst the larger sample size of the US population allows more stable estimates, these individuals have been bred in captivity for at least 8 generations, resulting in high relatedness which can bias LD-based estimates of N_e (England et al., 2006; Santiago

et al., 2025). This pattern of captive breeding is reflected in the *GONE2* analysis, which shows a severe population decline around 8 generations ago, when the ancestors of these samples were first brought into captivity. As the majority of captive cheetahs were sourced from Namibia (Marker-Kraus, 1997), the estimated N_e prior to this decline ($\sim$1,200) may reflect the ancestral Namibian population. This is consistent with recent estimates of a southern African N_e ranging from 700 to 1,600 (Fabiano et al., 2025). However, care must be taken when interpreting this estimate as intensive captive breeding may introduce bias in the predictions. Finally, although the estimates made here may be indicative of current N_e, the historic demographic decline in cheetahs that has resulted in severe inbreeding has been estimated to occur around 10 kya (Dobrynin et al., 2015; Fabiano et al., 2025; Menotti-Raymond & O'Brien, 1993), meaning this decline lies beyond the temporal resolution of LD-based methods such as *GONE2* (Santiago et al., 2020).

4.5.3 Genetic diversity and measures of inbreeding

Nucleotide diversity (π) for all populations was comparable to previous studies (Dobrynin et al., 2015), with median values between 0.0003-0.0004, supporting the idea first proposed by O'Brien et al. (1983) that cheetahs contain some of the lowest genetic diversity of any mammal. π was lowest in South Sudan with a median value of $\approx$0.00025, although unexpectedly, SNP heterozygosity was highest in this population. However, this is likely due to the low sample size skewing the output, as each individual in this population showed vastly different values, rather than a real biological observation. This emphasises the need for a suite of statistics to understand the genetic diversity of a population rather than relying on any one measure alone. The lower genetic diversity observed in South Sudan supports a previous observation of continued decline in genetic diversity in wild cheetahs (Terrell et al., 2016). This pattern is more subtly observed in Namibian individuals and is not observed at all in Tanzanian individuals, who hold comparative genetic diversity to US cheetahs. However, this is likely due to a combination of low sample size used in this study and the fact that previous observations were made on males, whereas this study contains mixed-sex samples (Terrell et al., 2016).

F_{IS}, indicating the level of inbreeding compared to random mating expectations

(Wright, 1965), was greater than zero in all populations, with few outliers in all populations except for Tanzania, suggesting all populations have experienced inbreeding. Previous research comparing cheetah subspecies did not find a difference in inbreeding between wild and captive cheetahs (Prost et al., 2022); although F_{IS} values vary between populations, the result observed in this study was not significant, supporting this previous observation and suggesting inbreeding is comparable between wild and captive cheetahs. However, as with the previous study, sample sizes for wild cheetah populations remain low, meaning more samples are required to confirm this finding. Tajima's D, which was also positive for all populations, supports the theory that all populations have experienced a genetic bottleneck or population contraction. This supports previous hypotheses suggesting a severe population contraction or bottleneck in the cheetah (Dobrynin et al., 2015; Fabiano et al., 2025; Menotti-Raymond & O'Brien, 1993).

Inbreeding can also be estimated using runs of homozygosity (ROH), where the length of the ROH indicates the age of the inbreeding. Longer ROH suggest more recent inbreeding, as the ROH have been exposed to less recombination, whilst shorter ROH indicate historic bottlenecks or inbreeding. F_{ROH} (the fraction of the genome in ROH above a specified length) has often been used to compare inbreeding between populations (McQuillan et al., 2008). Here, I demonstrate that F_{ROH} was highest for South Sudan and lowest in the captive US population, mirrored by the calculation of $\mathrm{N_{ROH}}$ (number of ROH) against $\mathrm{S_{ROH}}$ (total size of ROH). South Sudan and Tanzania had much higher $\mathrm{N_{ROH}}$ than all US and Namibian individuals, with a higher $\mathrm{S_{ROH}}$ than the majority of US and all Namibian individuals. This suggests more recent inbreeding in these populations compared to Namibia and US cheetahs, consistent with the observation of a higher F_{IS} in these populations.

One key finding comes from the comparison of long ROH ($>$ 1 Mb) across the populations. Here, I observe a larger proportion of long ROH ($>$ 5 Mb) in South Sudan and one Tanzanian individual compared to the other populations. However, two individuals in the captive US population have significantly higher proportions of long ROH compared to the rest of the population, with AJU7225 containing over double the total length of long ROH than the South Sudanese and Tanzanian individuals. This individual contains a consistent number of ROH compared to other US samples, but a much greater total length of ROH. This suggests significant recent inbreeding in

this individual, which was confirmed by extracting the pedigree of AJU7225 from the studbook (Figure 4.19). Although captive breeding programmes should select mating pairs based on relatedness (Couvet & Ronfort, 1994; Fernández & Caballero, 2001), it is clear that this individual has experienced multiple occurrences of severe inbreeding in captivity. However, as two of AJU7225's offspring were included in this study, it is possible to show that by mating AJU7225 with an individual with lower $\mathrm{F_{ROH}}$, their offspring can be prevented from inheriting the long ROH observed in their father. It is not known why close relatives were bred together to result in such severe inbreeding in this individual, but this is a reassuring finding highlighting that such inbreeding is not irreversible provided effective action is taken.

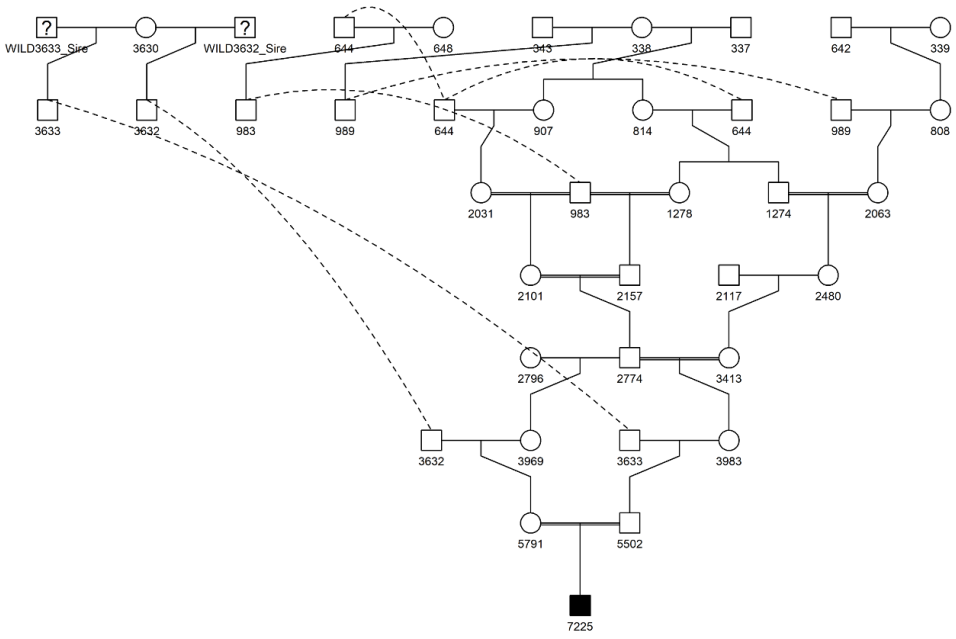


Figure 4.19: **Pedigree of AJU7225.** Males are shown with squares and females with circles. Number IDs for each individual correspond to the International Cheetah Studbook (L. Marker & Johnston, 2022). Individuals with wild origin are denoted with "WILD"; all other individuals were born in captivity. Severe inbreeding can be observed between multiple sets of AJU7225's ancestors. This figure was generated by Heather Sibley (George Mason University, USA) using the International Cheetah Studbook.

Overall, these data show the captive cheetah population exhibits comparable measures of inbreeding and genetic diversity to the wild populations, suggesting captive

breeding programmes are genetically effective, as has been previously suggested (S. E. Williams & Hoffman, 2009; Witzenberger & Hochkirch, 2011). This is strong support for the cheetah captive breeding programme, as signatures of a captivity-induced founder effect, which would reduce genetic diversity and increase mutational load, are not observed. Captive populations are important reservoirs for genetic diversity for the future. That the data presented here shows that the captive population has not lost diversity is testament to the effective management of the captive population to date.

4.5.4 Deleterious mutations

SNPs with a predicted 'high' or 'moderate' deleterious impact within protein coding genes were significantly enriched for functions associated with male reproductive fitness. Although the relationship between inbreeding and male fertility in cheetahs has been widely debated (Crosier et al., 2018; Terrell et al., 2016; Wildt et al., 1983), this finding provides undeniable evidence that cheetahs have experienced an accumulation of deleterious mutations in sperm-associated genes. This corroborates previous research focusing on premature termination codons and human-cheetah orthologs of fertility-related genes, where highly deleterious mutations were identified in several sperm-associated genes (Dobrynin et al., 2015; Peers et al., 2025). Whilst any link between this accumulation of mutations and the cheetah's demographic history remains unresolved, these mutations are key candidates which could be contributing to the poor sperm quality observed in cheetahs.

Despite observing high levels of ROH and a relatively high F_{IS} in Tanzania compared to the other populations, this population contains the lowest proportion of unique high- and moderate-impact SNPs, suggesting that higher inbreeding does not directly correlate with mutation load. However, this observation may be skewed by low sample size, so future work should aim to sample more individuals from this subspecies.

When comparing the distribution of these SNPs across populations, I observe a significantly higher proportion of unique high-impact SNPs in Namibian individuals, with a lower proportion of unique high- and moderate-impact SNPs in the Tanzanian population. This finding suggests that Namibian cheetahs may suffer decreased sperm

quality compared to other wild or captive populations; comparison of semen samples between different wild populations are crucial to confirm this, as previous studies of wild cheetah semen have focused on the southern African subspecies (Crosier et al., 2007; Lindburg et al., 1993; Wildt et al., 1983, 1988). As the majority of captive cheetahs are sourced from the Namibian subspecies, this could suggest that these mutations have occurred more recently in the Namibian population, or that these mutations have been removed from the captive population through purging or controlled breeding. However, as only four Namibian genomes were used in this study, future work to sequence more genomes from this population is necessary to confirm this finding.

This observation could have major impacts for conservation programmes. Augmentation, a form of translocation whereby individuals from one population are moved to another population typically to maintain or bolster genetic diversity and reduce inbreeding, is a conservation strategy currently being applied to the cheetah (L. L. Marker et al., 2008). This strategy is primarily being implemented in southern Africa, where individuals are translocated to maintain fragmented populations and increase gene flow (Buk et al., 2018). However, if this population does contain such a significantly higher proportion of unique highly deleterious SNPs, with the majority occurring at low allele frequency, moving these individuals between populations could have a detrimental effect on the augmented populations. Thorough profiling of the distribution of highly deleterious SNPs across populations in Namibia and southern Africa is necessary to determine the level of this threat. Additionally, cheetahs from southern Africa are being used to reintroduce the cheetah to India (Tordiffe et al., 2023; Venugopal, 2025), which has already been questioned due to poor choices of suitable protected habitats (Gopalaswamy et al., 2022). Taking individuals from a population with a high mutation load and exposing them to the founder effect caused by creating a new population could result in these deleterious SNPs becoming exposed through increased homozygosity.

Whilst no significant enrichment for gene function was observed in genes associated with high impact SNPs in functional non-coding regions, we find strong evidence for the deleteriousness of these variants across the genome. A combination of high CADD score, NCBoost score and an overlap of these SNPs with TFBS suggests these variants could have widespread deleterious impact on gene expression. Future work to confirm the associations of such mutations with genes using Hi-C data is necessary to confirm

the predicted impacts identified here. The distribution of deleterious non-coding SNPs mirrors that of the coding SNPs, with a lower proportion of unique SNPs identified in Tanzania. Again, this shows that the strong evidence of inbreeding in the Tanzanian population is not reflected in the volume of population-specific deleterious SNPs.

Previous research has shown that genomics-informed captive breeding can reduce inbreeding and mutational load, such as in the pink pigeon (*Nesoenas mayeri*) where genome-wide genetic load was used to identify optimal mate pairs with minimal load in offspring (Speak et al., 2024). In the pink pigeon, this method is more feasible due to the low number of individuals in captivity, making it possible to sample and sequence whole genomes of a larger proportion of the captive population. In species with larger captive populations, like the cheetah, the time and cost of collecting and sequencing every breeding individual in captivity makes this method infeasible. However, identification of deleterious mutations, particularly those impacting key functions like male fertility, segregating at high frequency in the captive population could enable the generation of a SNP panel to sequence these loci and prevent fixation of such deleterious mutations (Bertola et al., 2022; Kleinman-Ruiz et al., 2017; Wehrenberg et al., 2024).

It is important to note here that the SNP counts of population-specific mutations are biased to the reference genome used to call SNPs, raising the important issue of reference bias in population genetic studies, which can underestimate genetic diversity and differentiation (Akopyan et al., 2025). As the reference genome used in this study is derived from a captive cheetah, I would expect this bias to falsely reduce the number of unique captive US SNPs and inflate the number of unique SNPs found in South Sudan and Tanzania. However despite this, the US cheetahs still have a higher proportion of unique high and moderate impact SNPs than South Sudan and Tanzania, suggesting that reference bias has not impacted the overall patterns I observe.

4.6 Conclusion

In this chapter, I interrogate population resequencing data to examine the distribution of deleterious mutations across previously-identified genomic loci, such as premature termination codons and predicted functional non-coding regions. I confirm the fixation

of PTCs in multiple genes associated with male fertility across captive and wild cheetahs and identify an over-representation of deleterious SNPs across sperm-associated genes. I also observe a high load of these SNPs unique to Namibia, with high inbreeding metrics in Tanzania and South Sudan, whilst the captive US population holds similar levels of diversity and inbreeding. Finally, I discuss several conservation impacts of this work, highlighting the importance of genomic studies to inform conservation management of threatened species.

4.7 Supplementary material

4.7.1 Supplementary figures

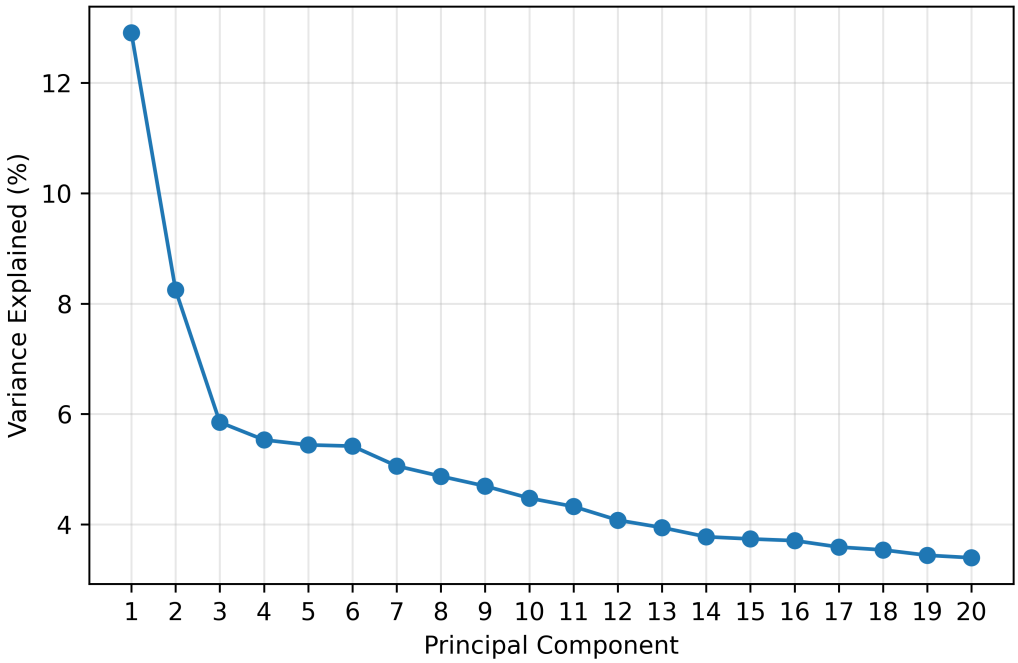


Figure S4.1. **Scree plot showing variance explained by each principal component of a principal component analysis (PCA) run on unrelated cheetahs (Figure 4.5).** Principal components 1 and 2 explain at least 8% of variance each, with subsequent PCs each contributing to $< 6\%$.

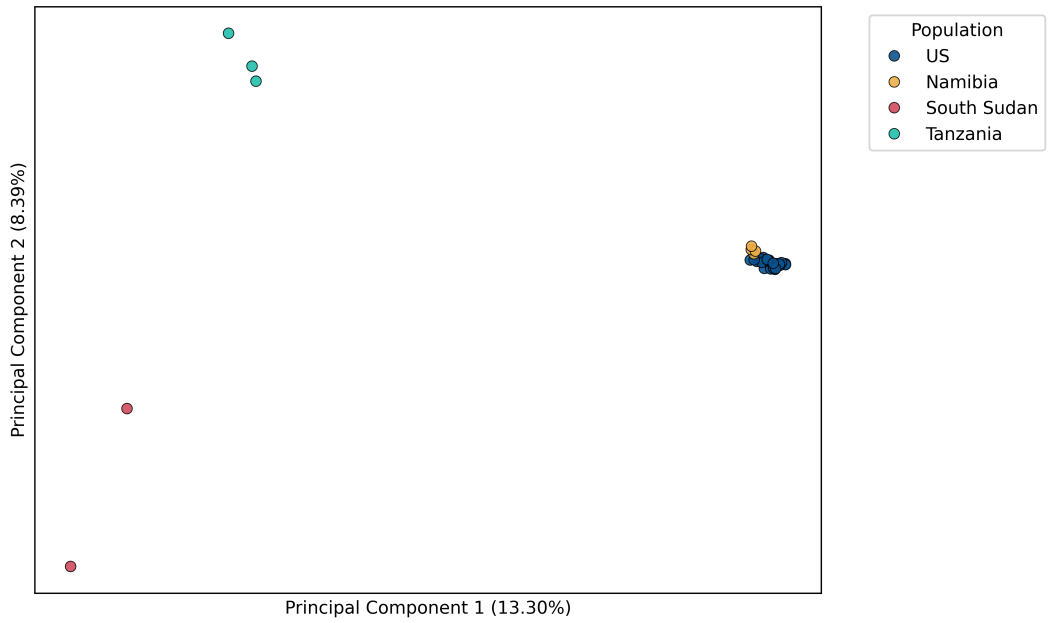


Figure S4.2. **Principal Component Analysis (PCA) of all cheetahs in this study.** PCA based on filtered genome-wide SNPs showing genetic clustering by population: US (blue), Namibia (yellow), South Sudan (red), and Tanzania (green). The percentage of variance explained by each principal component is shown in axes labels.

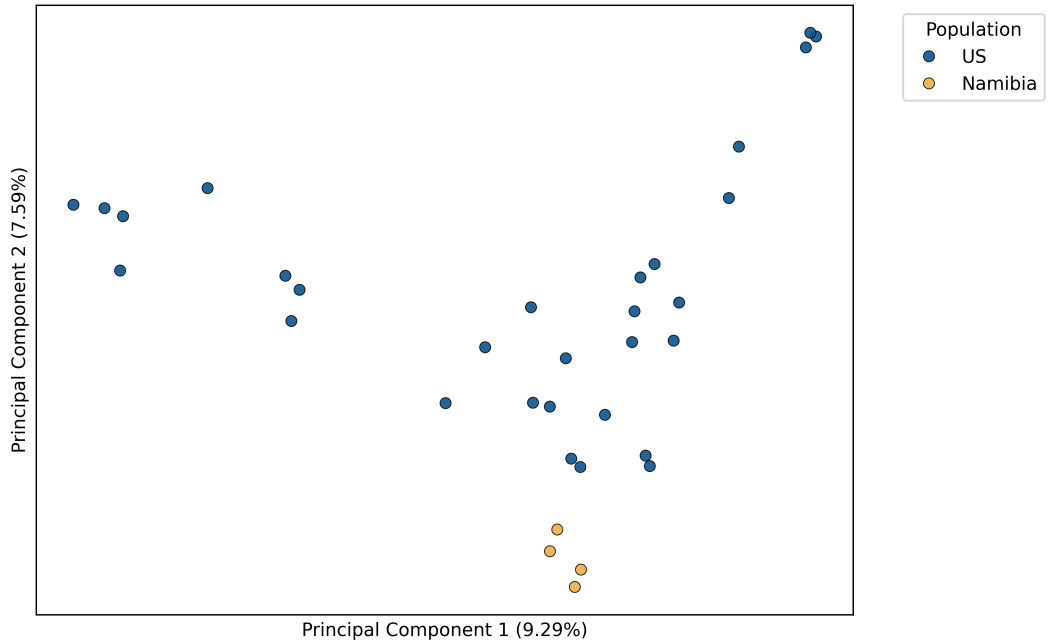


Figure S4.3. **Principal Component Analysis (PCA) of US and Namibian cheetahs.** PCA based on filtered genome-wide SNPs of captive US (blue) and wild Namibian (yellow) cheetahs. The percentage of variance explained by each principal component is shown in axes labels.

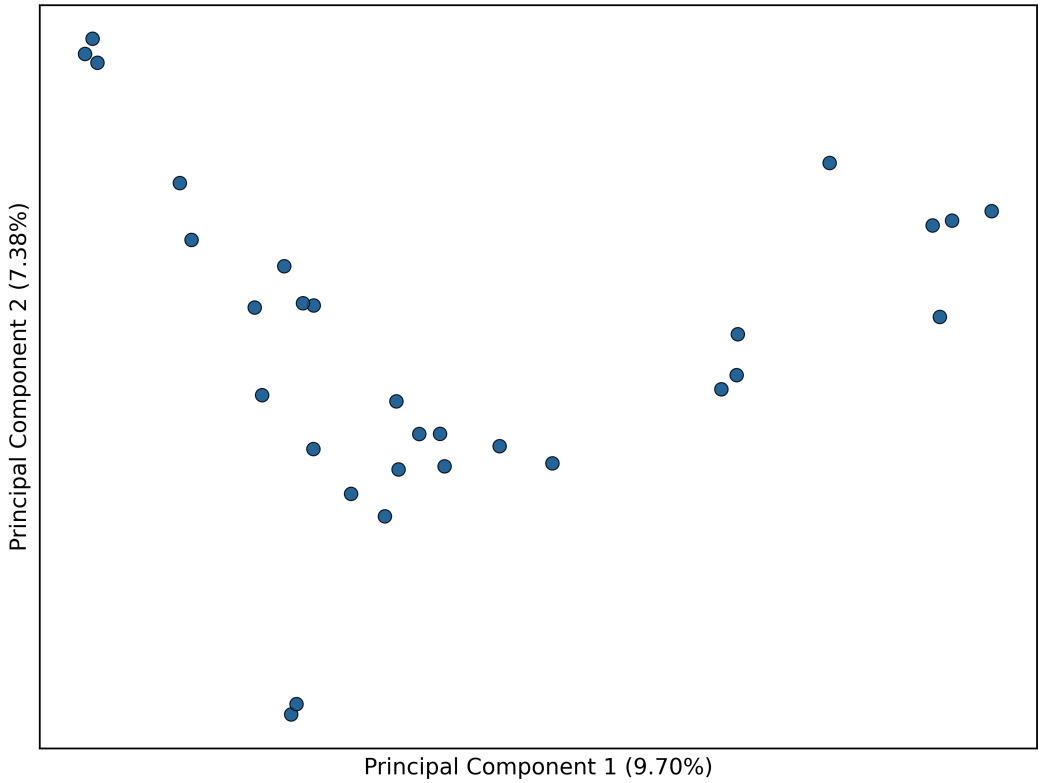


Figure S4.4. **Principal Component Analysis (PCA) of captive US cheetahs.** PCA based on filtered genome-wide SNPs of US cheetahs. The percentage of variance explained by each principal component is shown in axes labels.

Figure S4.5. **ADMIXTURE analysis of all cheetahs in this study.** Model-based clustering of cheetah populations at K = 2–5 using only unrelated individuals. Each vertical bar represents an individual, and colours represent the proportion of ancestry assigned to each genetic cluster. Admixture cross-validation errors for K = 1–5 were 0.49, 0.52, 0.55, 0.57 and 0.62.

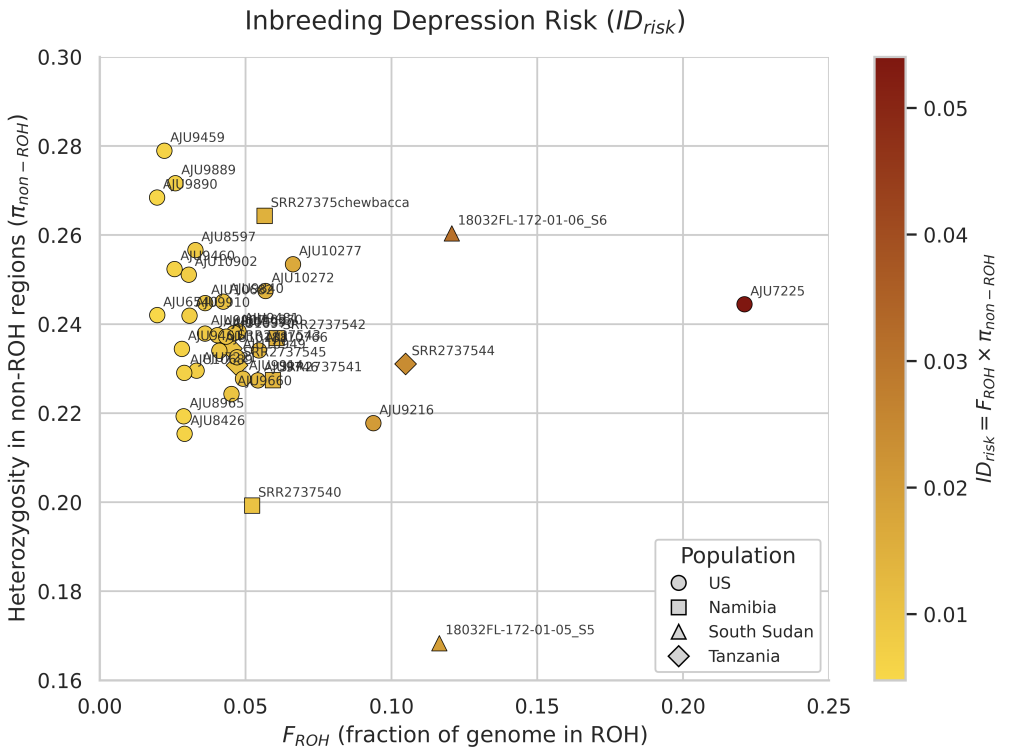


Figure S4.6. **Inbreeding depression risk (IDrisk) for each individual.** IDrisk scores combine $\mathrm{F_{ROH}}$ with average heterozygosity in non-ROH genomic regions. Individuals are coloured by severity (dark red = highest IDrisk). Shapes correspond to populations: US (circle), Namibia (square), South Sudan (triangle), Tanzania (diamond). As per thresholds of Kyriazis et al. (2025), all individuals show low risk (IDrisk $\leq$ 0.05).

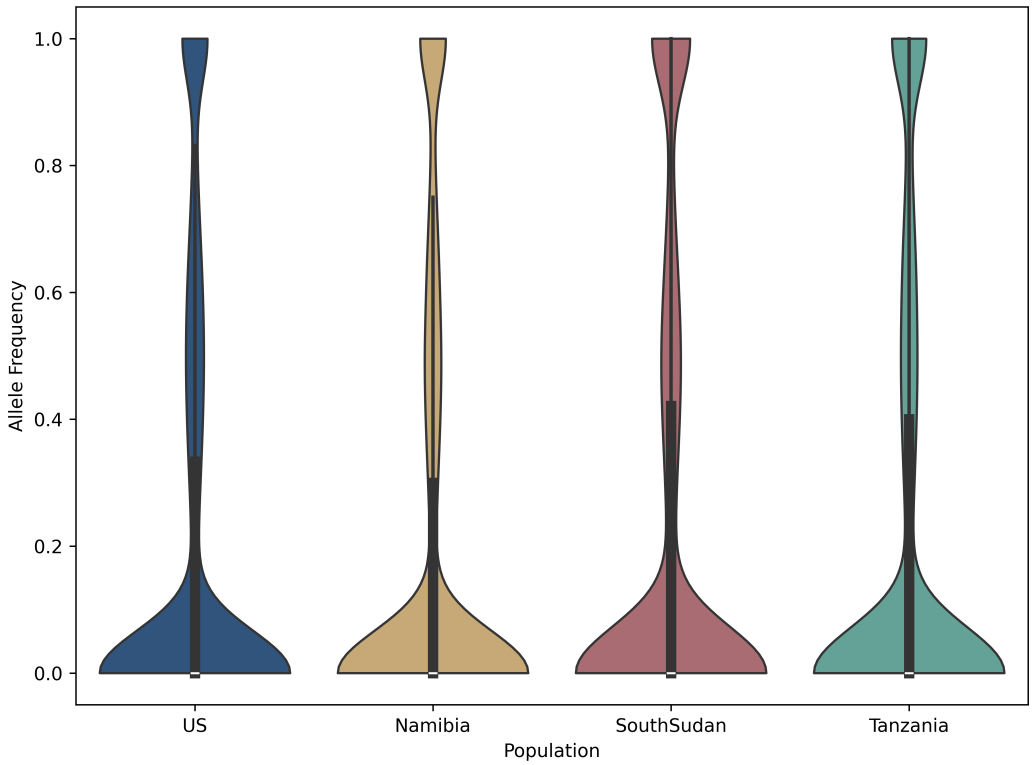


Figure S4.7. **Allele frequencies of moderate-impact coding SNPs per population.** Allele frequencies were calculated for each SNP classed as moderate impact by SnpEff: US (blue), Namibia (yellow), South Sudan (red), Tanzania (green).

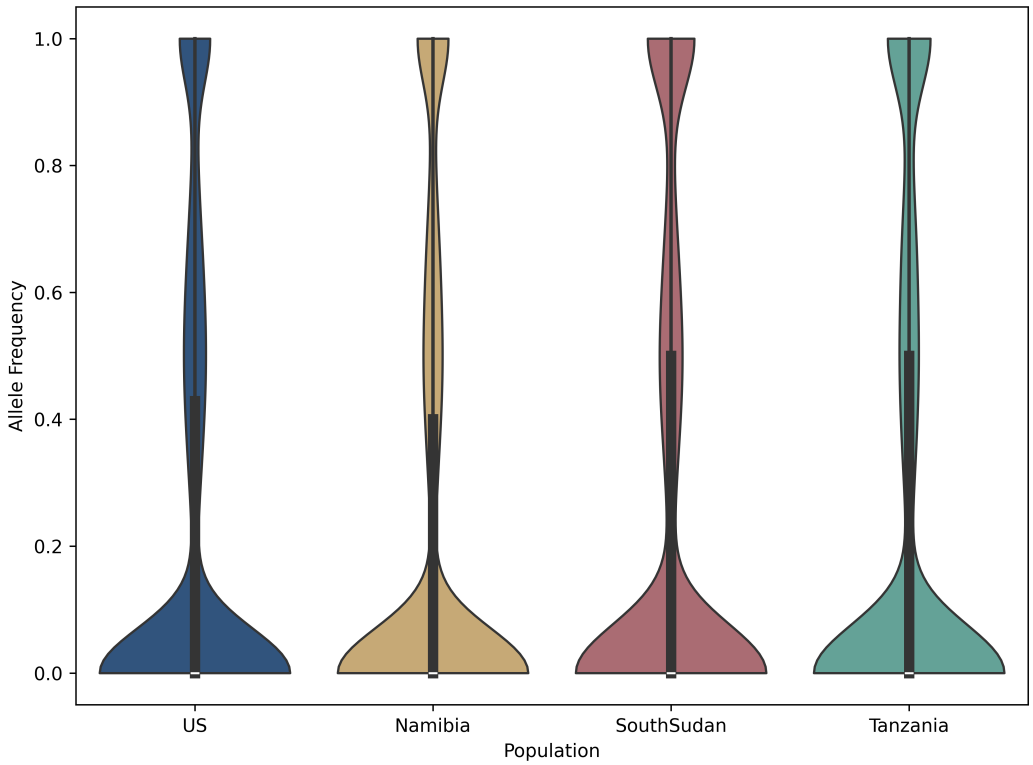


Figure S4.8. **Allele frequencies of low-impact coding SNPs per population.** Allele frequencies were calculated for SNPs classed as low impact by SnpEff: US (blue), Namibia (yellow), South Sudan (red), Tanzania (green).

4.7.2 Supplementary tables

Supplementary tables can be found at github.com/EarlhamInst/JP_PhD.

Table S4.1. **Sample metadata.** Metadata for all cheetah samples used in this study. For all samples, their sex, source population, and owner/provider is provided. For relevant samples, their house name, collection date, studbook ID, and accession is also provided. Captive samples were collected and provided by Dr Klaus-Peter Koepfli and Dr Adrienne Crosier (Smithsonian Conservation Biology Institute/National Zoo). South Sudan samples were collected by Kelsey Greene (African Parks South Sudan) and sequenced by Dr Ellie Armstrong at the University of California, Riverside.

Table S4.2. **Read quality and mapping.** For each sample, the total number of raw sequence reads is provided, followed by the number and percentage of base pairs removed through quality trimming (Trim_Galore) and the remaining read numbers. The number and percentage of reads that mapped successfully (BWA-MEM) is shown alongside the average sequencing depth for each individual.

Table S4.3. **Significance values for genetic diversity statistics.** For each population statistic, a Kruskal–Wallis test was carried out to determine overall significance. H and p-values are provided. For each pair of populations, pairwise Wilcoxon rank-sum tests were carried out, followed by Bonferroni correction of p-values.

Table S4.4. **Gene ontology (GO) enrichment analysis for high-impact coding SNPs.** GO enrichment analysis was run using ShinyGO v0.85 (Ge et al., 2020). The foreground gene set consisted of genes with a predicted high-impact SNP and a corresponding 1:1 human ortholog. The background set was all 1:1 orthologs between cheetah and human. Results shown meet an enrichment FDR q-value < 0.05.

Table S4.5. **Gene ontology (GO) enrichment analysis for moderate-impact coding SNPs.** GO enrichment analysis was run using ShinyGO v0.85 (Ge et al., 2020). The foreground gene set consisted of genes with predicted moderate-impact SNPs and a 1:1 human ortholog. The background set was all 1:1 orthologs between cheetah and human. Reported results meet an enrichment FDR q-value < 0.05.

Table S4.6. **TF motifs containing high-impact SNPs.** The number of high-impact SNPs overlapping each transcription factor binding motif. Motifs were identified using the JASPAR 2024 Vertebrate Core non-redundant database (https://jaspar.elixir.no/downloads/) with FIMO (MEME Suite v5.1.0) (Grant et al., 2011).

Table S4.7. **Population distribution of previously identified PTCs.** Sixty-five previously identified premature termination codon (PTC) loci (Peers et al., 2025) were extracted from the full population VCF using aciJub1 as the reference. For each locus, the PTC position, gene name, VCF status, reference and alternate alleles, site quality, and genotype counts are provided. The final outcome is annotated as reference-unique, polymorphic, or fixed. Two loci had insufficient coverage to call.

Chapter 5

Discussion

5.1 Thesis summary

In this thesis, I use the cheetah as a model to consider the impacts of long-term low effective population size on the genome, considering both coding and non-coding regions. I utilise both publicly available genomic resources and novel sequencing data and apply a convolutional neural network approach to annotate the non-coding genome. My results have implications in our understanding of accumulations of deleterious mutations, particularly those associated with inbreeding depression symptoms, and the analysis of deleterious non-coding mutations in non-model species.

5.1.1 Chapter summaries

Chapter 2 uses a comparative genomics approach to identify 65 genes with novel premature termination codons (PTCs) resulting in potential gene pseudogenization in cheetahs. Of these PTC-causing mutations, at least 22 were shared in wild cheetahs and four were observed in both captive and wild cheetahs. The four genes with PTCs shared across all seven samples are thought to be involved in male fertility and immune response, so may be contributing to the reproductive and immune defects observed in cheetahs. This chapter demonstrates the utility of growing public genomic datasets in a comparative framework to uncover previously uncharacterised signatures of genetic bottlenecks.

In **Chapter 3**, I apply machine learning (ML) in a non-model species. Whilst the

field of ML is rapidly growing, the majority of biological applications are in model species, but here I demonstrate that there is great potential for these tools to be used in non-model species. I apply an ML model not designed for non-model species and find that it can accurately annotate the non-coding genome through transfer learning. This provides a framework for the annotation of any non-model species, which is crucial for fully understanding the evolution of traits and diseases across species, characterising the mutation load of the whole genome and informing conservation programmes, such as *in-situ* captive breeding or *ex-situ* population management.

Chapter 4 leverages population data across multiple distinct populations of wild and captive cheetahs to assess deleterious mutation load across the genome. High-impact deleterious SNPs were identified within protein-coding genes associated with sperm function and in functional non-coding regions, with each population containing unique SNPs segregating at high frequency. Whilst higher measures of inbreeding were observed in South Sudan and Tanzanian populations, a higher proportion of unique deleterious mutations were identified in Namibia. The captive population showed comparable or higher genetic diversity to the sampled wild populations with fewer unique deleterious mutations. This highlights the potential for captive breeding populations to act as a reservoir for diversity.

Here, I discuss the implications of my results within a wider scientific and conservation context, including key caveats to my findings that must be considered, as well as suggestions for future work and routes to inform and positively impact conservation management.

5.2 Assessment of results and impact

5.2.1 Accumulation of mutations in coding regions

In Chapter 2, I identify novel PTCs unique to the cheetah in genes associated with male fertility, which appeared to be fixed in the species based on the limited population data available at the time of study. By examining these loci in a broader set of population data

in Chapter 4, it was possible to confirm that these mutations are present in all captive and wild cheetahs that were sequenced. This supports Chapter 2's results and shows the accuracy of my pipeline to identify real deleterious mutations rather than technical artefacts. However, as the reference genomes used in Chapter 2 were highly fragmented, my pipeline involved several conservative steps to remove any potential false positive results. Although these genomes enabled me to identify cheetah-specific mutations, I hypothesise that repeating these analyses with higher-quality genome assemblies would yield more identified mutations, as filtering steps could be relaxed.

The PTC-causing mutations I identified are likely to be weakly to moderately deleterious, as strongly deleterious alleles would result in infertility or death and would therefore not be inherited. Due to the low N_e of cheetahs and resultant decreased efficiency of purifying selection, deleterious mutations can segregate at higher frequency than in larger populations (Crow & Kimura, 1970). Therefore, these mutations, which contribute to the mutational load, have persisted despite purifying selection and can accumulate to the point that entire genes become pseudogenized. Supporting this observation, in Chapter 4, I observe deleterious SNPs occurring in protein-coding genes associated with sperm function. As with the PTC-causing mutations, these mutations are not likely to result in infertility or death, but may still contribute significantly to reduced fitness, including effects such as severe spermatozoal abnormalities. However, experimental evidence is needed to validate the impact of these mutations on fertility (Hsu et al., 2014; Smithies, 1993).

One key assumption throughout this thesis is that long term low N_e is directly linked with the accumulation of deleterious mutations observed. Although spermatozoal defects are observed in wild and captive cheetahs (Wildt et al., 1983, 1988), it is currently unknown whether this is a direct result of their demographic history. This could be confirmed by sequencing ancient samples, ideally spanning the timescale of cheetah population decline. The oldest sample which has been sequenced at the time of writing is an *A. j. hecki* museum specimen collected in the 1830s (Prost et al., 2022). As population decline and subsequent inbreeding in cheetahs is widely agreed to have begun over 10,000 years ago, much older samples are required to confirm that the high mutation load observed in cheetahs is a result of their severe population decline over the last 10,000 years. The fossil record of cheetahs (or their ancestors) spans millions of years of

evolutionary history, however controversy exists over exact species classifications meaning no suitable ancient cheetah samples are currently available (Gimranov et al., 2024; Hemmer et al., 2011; Orlando et al., 2021; Van Valkenburgh et al., 2018; Werdelin & Lewis, 2005; Werdelin et al., 2010). More recent fossils, dated to around the time of cheetahs' population decline, have not yet been identified and may not exist (Van Valkenburgh et al., 2018; Werdelin & Lewis, 2005). If such samples were identified in the future, this introduces issues of DNA degradation commonly observed when sequencing ancient DNA (aDNA); even the 1830s museum sample had such low coverage that accurate estimates of genetic diversity could not be made (Prost et al., 2022).

In both Chapter 2 and 4, I observe deleterious mutations in genes associated with male fertility. Whilst reduced fertility is a phenotype often associated with inbreeding depression in many species (Kincaid, 1983; Roelke et al., 1993; Santymire et al., 2006, 2014; Sheridan & Pomiankowski, 1997; Tsheten et al., 2023; Vasudeva et al., 2025), it is important to consider why male reproductive traits are particularly affected. Cheetahs exhibit a polygynous mating system in which a small number of males contribute disproportionately to the next generation, resulting in high variance in male reproductive success (Gottelli et al., 2007). Such variance further reduces N_e, increasing the influence of genetic drift and reducing the efficacy of purifying selection. Genes involved in male fertility are highly polygenic and often rapidly evolving, meaning that weakly deleterious mutations occurring in multiple loci may have cumulative effects on fertility (Hirawatari et al., 2015; Torgerson et al., 2002). Under prolonged periods of low N_e, mildly deleterious mutations affecting male fertility may accumulate, potentially explaining the enrichment of deleterious variants observed in sperm-associated genes.

5.2.2 Accumulation of mutations in non-coding regions

The non-coding genome contains often-overlooked but functionally-important genetic variation that influences gene regulation and organismal traits. To assess the mutation load of cheetahs outside of protein-coding genes, I annotated the functional non-coding landscape of the cheetah genome and identified deleterious mutations within these regions. Unlike the deleterious SNPs I identify in protein-coding genes, the mutations I observe in the non-coding genome do not significantly associate with a particular gene

function and are distributed throughout the genome. SNPs were associated with genes based solely on their distance, as the nearest downstream gene for each SNP was extracted (Levine & Tjian, 2003; Urtecho et al., 2023; C. Wang et al., 2020). An important caveat is that these analyses do not take into account the potential for overlapping transcripts or three-dimensional chromatin structure. It is becoming more widely accepted that 3D chromatin structure is associated with regulation of gene expression and the link between regulatory elements and the genes they regulate is not straightforward (Fulco et al., 2016; Pennacchio et al., 2013; Sanyal et al., 2012). Generation and incorporation of Hi-C data into such analyses in the future is necessary to more accurately associate genes with deleterious non-coding SNPs. This would allow the use of Topologically Associating Domains (TADs), three-dimensional structures that constrain gene regulatory interactions, to associate predicted regulatory elements with likely target genes and infer the phenotypic impact of SNPs within these regions (Tena & Santos-Pereira, 2021). Additionally, experimental validation, such as transcription factor binding assays, are necessary to validate the impact of mutations on binding affinity and gene expression regulation (Elnitski et al., 2006; Whitfield et al., 2012). For example, by introducing the mutations I identify upstream of a reporter gene, luciferase reporter assays can be used to measure the impact of a SNP on regulatory activity (Whitfield et al., 2012; H. Zhang et al., 2010).

The deleteriousness of the mutations I identify was predicted using a combination of models (CADD and NCBoost). Both sets of scores were originally calculated in the human genome and I transferred them onto the cheetah (Caron et al., 2019; Schubach et al., 2024). This could have resulted in some inaccuracies; both methods rely on alignment of the cheetah and human genome, so errors within the alignment, particularly around species-specific regions, will result in inaccurate deleteriousness scores. Additionally, mutations with deleterious effects in humans may not have the same selection pressures or functional impact in cheetahs. However, the mutations I identify fall within genomic windows with high sequence conservation and are within 50 kb of a gene, which are both indications of regulatory elements (Christmas et al., 2023; ENCODE Project Consortium, 2012; Levine & Tjian, 2003; Rands et al., 2014). Therefore, even if the predictions of functional non-coding regions or deleteriousness of mutations are inaccurate, it is likely that mutations occurring within these conserved regions have

deleterious impact, as sequence conservation has repeatedly been identified as an indicator of functional importance (Caron et al., 2019; Christmas et al., 2023; Rands et al., 2014).

5.2.3 Population distribution of deleterious mutations

When making comparisons between *in-situ* and *ex-situ* populations, and subsequent inferences about evolution in the wild and captivity, it is important to consider that wild cheetah populations are not entirely unmanaged. Less than half of wild cheetah populations exist in actively protected areas, and those that do not are restricted by anthropogenic impacts restricting movement (Durant et al., 2017; Marnewick & Somers, 2015). The majority of remaining cheetahs exist in small fragmented populations with limited gene flow between them, although translocations may have introduced novel gene flow (Buk et al., 2018; L. L. Marker et al., 2008).

Within this context, a major conservation impact of the present work is understanding the population-specific distribution of deleterious alleles. Prior to examining SNP distribution within populations, I investigated population structure. The three wild African populations separated into distinct clusters, with the captive US population clustering closely with the Namibian cheetahs. This likely reflects the historical origin of captive cheetahs, as the majority were sourced from the southern African population (Marker-Kraus, 1988). However, this result, in combination with the comparable levels of inbreeding and genetic diversity observed between wild and captive cheetahs, also indicates that there has not been a significant founder effect in the captive population. This is particularly important, as captive populations can be crucial to the survival of a species or population that is almost extinct in the wild through genetic rescue (W. E. Johnson et al., 2010; Sandler et al., 2021); maintaining genetic diversity in captive populations is imperative to enable future genetic rescue projects.

I observe a lower proportion of both coding and non-coding SNPs unique to Tanzania, with a higher proportion of highly-deleterious coding SNPs unique to Namibia. This suggests that the high mutation load in Namibian cheetahs has either been purged in the captive population or has arisen more recently in the wild population. However,

I have a small sample size ($n \leq 4$) in the wild populations, meaning the patterns I observe may be not be representative of the real population structure and mutation load in wild cheetahs. More comprehensive sampling of the captive and wild populations, ideally temporally resolved, would confirm the patterns I observe and inform whether the Namibian mutations are recent or have been purged in captivity.

The unprecedented spatial resolution afforded by the data in this thesis provides much needed evidence for optimising population management. However, it is important to note that not all cheetah populations are represented. Whole-genome samples for Northwest Africa and Asiatic cheetah subspecies are not currently available, meaning it is not possible to understand their genetic diversity or mutation load. This is particularly important as these populations are more threatened than the Southern African subspecies, with fewer individuals distributed in highly-fragmented sub-populations (Durant et al., 2017). Therefore, understanding their genetic diversity and load of deleterious mutations is crucial to inform conservation management and prevent the loss of private genetic variation unique to these populations. Additionally, the current subspecies classification in cheetahs has recently been debated, with a call for the separation of *A. j. jubatus* into two distinct subspecies (Prost et al., 2022). My results, albeit based on small sample sizes, support this separation, as I observe clear population differentiation between Tanzanian and Namibian cheetahs. Whilst this subspecies delineation is important from a taxonomic standpoint, it also has key conservation impacts: conservation management programmes typically operate at a subspecies level, so subspecies classifications have both conservation and financial implications (Garner et al., 2005; Morrison et al., 2009; Zink & Klicka, 2022). Accurate taxonomic descriptions are crucial to ensure appropriate funding of conservation projects and to conserve genetic diversity of all subspecies (Morrison et al., 2009; S. Wang et al., 2025; Zink & Klicka, 2022). In cheetahs, as the southern African population is the most numerate and least threatened, the south-eastern population (which includes Tanzania) may be overlooked unless they are classified as a separate subspecies. Again, more whole-genome samples from these populations are necessary to confirm their genetic differentiation and support any future subspecies reclassification.

Although the genomic data used in this thesis is not comprehensive enough to support taxonomic inferences, my results still have several major conservation implications.

A key finding of my research is the identification of a high proportion of deleterious SNPs in sperm-related genes, particularly in the Namibian population, supporting previous reports of poor sperm quality in wild cheetahs (Koester et al., 2015; Wildt et al., 1983). As stated previously, cheetah conservation programmes are utilising translocations to support wild cheetah populations, with Namibian cheetahs being relocated within southern Africa and reintroduced into India (Buk et al., 2018; L. L. Marker et al., 2008; Tordiffe et al., 2023; Venugopal, 2025). However, assessing these individuals for their load of deleterious mutations, particularly in sperm-related genes, is crucial to prevent the further spread of such mutations. Future work needs to sample more exhaustively, encompassing the entire species distribution, to confirm whether these mutations are indeed unique to the Namibian population.

Although it has long been agreed that increased inbreeding leads to the expression of deleterious alleles and subsequent decreased fitness, empirical evidence linking such deleterious mutations with direct fitness effects remains scarce and was unfortunately beyond the scope of this thesis. However, a recent study combined long-term health data with genomics to link a deleterious allele directly to reduced fitness in a reintroduced Eurasian lynx (*Lynx lynx*) population (Niehaus et al., 2025). Additional research in wild Soay sheep (*Ovis aries*) and arctic fox (*Vulpes lagopus*) populations also integrate fitness observations with genomic data, identifying key loci linked to reduced fitness (Hasselgren et al., 2024; Stoffel et al., 2021). These studies show the potential for directly linking deleterious mutations with fitness effects in inbred populations. Conservation organisations, like the Cheetah Conservation Fund (Namibia) and the Smithsonian Zoo (USA), have been collecting sperm samples from wild and captive cheetahs over many years (Dr Laurie Marker & Dr Adrienne Crosier, pers. comm., 2025). Following the lynx study framework (Niehaus et al., 2025), by integrating such phenotypic data with my genomic results, it may be feasible to associate the mutations I identify with fitness effects. Future work should utilise such phenotypic data to confirm the predicted impacts of the deleterious mutations I identify, particularly those in sperm-associated genes.

Notwithstanding the need for additional work to confirm the phenotypic impact of the mutations I observe, the results from my thesis inform conservation management. The mutations I identified could be included in a panel to inform captive breeding pair selection, monitor mutation load and select individuals for translocations. Genome-

wide SNP panels have been used to inform conservation decision-making in a variety of species, including the Iberian lynx (*Lynx pardinus*; Kleinman-Ruiz et al. (2017)), Pacific lamprey (*Entosphenus tridentatus*; Hess et al. (2015)), European bison (*Bos bonasus*; Wehrenberg et al. (2024)), cichlid fish (*Oreochromis* sp.; Ciezarek et al. (2022)), sun parakeet (*Aratinga solstitialis*; Spitzer et al. (2020)) and lion (*Panthera leo*; Bertola et al. (2022)). These panels can enable non-invasive monitoring of populations without requiring whole-genome sequencing, making the method more accessible for financially-constrained conservation projects.

Monitoring the prevalence of deleterious mutations is particularly important when considering *in vitro* fertilisation (IVF), an intervention used in *ex-situ* cheetah reproduction (Crosier et al., 2020). Whilst this method successfully circumvents reproduction-related issues, it risks removing any selection that would act on the potential causative mutations, resulting in persistence of deleterious variants that might otherwise be purged. In recent applications of IVF in the cheetah, genetic testing was carried out, however Crosier et al. (2020) sequenced only nine loci, meaning the overall genetic load of the selected mating pairs was not considered. My finding of deleterious mutations in sperm-associated genes segregating at low frequency across captive cheetah genomes provides a potential panel of fertility loci to consider in subsequent IVF attempts. This could prevent the spread and potential fixation of these low-frequency but highly deleterious mutations, which could otherwise pose a threat to the survival of both captive and future reintroduced populations.

5.2.4 Use of comparative genomics and machine learning

The use of comparative genomics has recently increased significantly thanks to large-scale projects publishing increasing numbers of high-quality reference genomes (Darwin Tree of Life Project Consortium, 2022; Zoonomia Consortium, 2020). Here, I applied comparative genomics to identify mutations in the cheetah that are not present in related species, enabling me to confirm the specificity of these mutations. As stated previously (see section 5.2.1), the quality of reference genomes used can have a significant impact on the accuracy of the results. Despite this, due to the conservative filtering steps I applied in Chapter 2, the mutations I observed were valid (as evidenced in Chapter

4) and have potential conservation impacts, showing that even fragmented reference genomes can yield biologically-relevant results. However, even in the time between the start of this analysis and its publication (Peers et al., 2025), higher quality chromosome-level assemblies for many of the included species have been published (Figueiró et al., 2017; Plasil et al., 2025; Winter et al., 2023). As the cost of sequencing continues to decrease, we can expect increasingly high quality reference genomes and annotations to become available, further reducing the requirement for such stringent filtering.

A significant contribution of my thesis lies in demonstrating the application of a convolutional neural network ML approach alongside publicly-available genomic data to obtain useful conservation-orientated data. A major barrier to the wider application of machine learning in conservation is the limited technical expertise within the field. Most conservation researchers lack formal training in ML, and although online resources are available, they are often time-consuming to engage with and rarely tailored to the specific needs of conservation biologists (Christin et al., 2019; Miao et al., 2025; Pichler & Hartig, 2023). However, my results offer a simple, easy to use application of ML that is relatively accessible and straightforward to implement, opening up this approach to a wider range of projects. Although Chapter 3 focuses on one tool, ExplaiNN, this method of transferring between species could be applied to other ML tools, depending on what specific predictions are required.

One major caveat of the application of ML to the task of annotating the non-coding genome is that of imbalanced data classes. My results show that when tested on a balanced dataset (i.e. an equivalent number of 'positive' and 'negative' sequences), models can make predictions with high accuracy. However, when applied to an imbalanced dataset, model performance decreases considerably due to the quantity of false positives. This is because the non-coding genome contains significantly more non-functional than functional sequence (Rands et al., 2014). Although I was able to reduce the number of false positives in my final predictions by applying additional filters, it is likely that a substantial number of false positive predictions remain. This means that a number of the predicted regulatory elements I report may not be functional. However, by applying sequence conservation information and identifying mutations that overlap a known transcription factor binding motif, I was able to generate a subset of predicted deleterious mutations which are likely to impact gene expression.

There are multiple approaches that could be used to confirm this result. Firstly, experimental chromatin accessibility data (such as ATAC-seq data) could be used to validate the functional regions I identify; those which overlap a peak in ATAC-seq data are more likely to be functional, whilst those which are located in closed chromatin are more likely to be false positive results. Additionally, TF binding assays or knock-out experiments of the deleterious mutations predicted could be used to functionally validate their impact on gene expression (Hsu et al., 2014; Smithies, 1993; Tewhey et al., 2016). Expanding this work to test the model on a wider taxonomic range of mammalian species could both increase our understanding of the model's predictions and, by training on more species, potentially increase model accuracy. Finally, future developments in ML methods could result in a model that is better suited to imbalanced data. It is important to note that the issue with data imbalance is not unique to this type of research question; it is a widely documented phenomenon across the field of ML, particularly in biological applications (Branco et al., 2017; Chawla et al., 2002; Ghosh et al., 2024; Haque et al., 2014; Schubach et al., 2017).

5.3 Summary and considerations for future work

Whilst I have made several suggestions for future work throughout this chapter, some of these steps are more pressing than others. Firstly, a major caveat to my work is the low sample size for wild cheetah populations. Therefore, any future work should aim to sample more thoroughly across wild populations, ensuring to include individuals from all subspecies, to fully characterise genetic diversity and mutation load across the species. Additionally, as emphasised by Chapter 2, it is crucial to include a sister species to confirm mutations are cheetah-specific. Full demographic reconstruction of the cheetah is also necessary. Multiple hypotheses to explain low genetic diversity in cheetah populations have been proposed, however these hypotheses have mostly been tested based on microsatellite data or a small number of whole genomes (Castro-Prieto et al., 2011; Dobrynin et al., 2015; Driscoll et al., 2002).

Future work should prioritise sequencing more whole genomes of wild cheetahs to enable demographic reconstruction. Tools based on coalescent rates (PSMC, MSMC,

SMC++) or the site frequency spectrum (dadi, fastsimcoal2) typically require multiple high-coverage genomes to accurately reconstruct demographic history (Excoffier et al., 2013; Gutenkunst et al., 2009; H. Li & Durbin, 2011; Schiffels & Wang, 2020; Terhorst et al., 2017). Approaches based on genomic linkage patterns (GONE2) or identity-by-descent (IBDNe) predict more recent demographic history but require larger samples sizes (typically at least 20 unrelated individuals) (Browning & Browning, 2015; Santiago et al., 2025). Increased whole genome sampling across the cheetah's range would allow a combination of these tools to test current demographic hypotheses and confirm the severity and timing of any population bottlenecks.

Although I identified deleterious mutations across the genome, including often-overlooked regulatory elements, my results focus solely on single nucleotide mutations. However structural variants (SV), such as deletions, insertions, translocations, inversions, and duplications, are increasingly thought to contribute to adaptive evolution, resulting in unique phenotypes and diseases (Radke & Lee, 2015; Weischenfeldt et al., 2013; Wellenreuther et al., 2019, 2025). Studies of structural variation in non-human species are lacking, but there is growing evidence in model species that structural variants are associated with metabolic processes (Axelsson et al., 2013; Radke & Lee, 2015), fertility (Hamilton et al., 2012) and immune function (Rodriguez et al., 2023). Therefore, such variation could be highly relevant to inbred species like the cheetah, who experience decreased fertility and a potential increased susceptibility to disease. Uncovering structural variation relies on long-read sequencing, which is becoming increasingly affordable and therefore applicable to conservation studies, although high-quality samples are required (Rhoads & Au, 2015a; Zhao et al., 2021). Future work to sample across the range of the cheetah should take this into consideration and therefore aim to generate long-read whole genomes where possible. This would enable a more thorough understanding of the mutation load in the cheetah; the pattern of SV diversity across populations could be compared to the pattern of SNP diversity I report in this thesis to confirm or expand on our understanding of population differentiation. Additionally, deleterious SVs could be identified, which may also have a significant impact on the decreased fitness observed in cheetahs.

The work presented in this thesis demonstrates how comparative genomics and machine learning can be integrated to address key questions in conservation genetics. By

combining existing genomic resources with novel sequencing data and developing an approach to annotate and interpret deleterious mutations across coding and non-coding regions, this thesis contributes to our understanding of the genomic consequences of small effective population size in the cheetah. More broadly, the analytical framework introduced here provides a generalisable strategy for assessing mutation load in non-model species and highlights the potential of computational approaches to inform conservation management.

Bibliography

Aartsma-Rus, A., Van Deutekom, J. C. T., Fokkema, I. F., Van Ommen, G.-J. B., & Den Dunnen, J. T. (2006). Entries in the leiden duchenne muscular dystrophy mutation database: An overview of mutation types and paradoxical cases that confirm the reading-frame rule. *Muscle Nerve*, *34*, 135–144.

Abascal, F., Corvelo, A., Cruz, F., Villanueva-Cañas, J. L., Vlasova, A., Marcet-Houben, M., Martínez-Cruz, B., Cheng, J. Y., Prieto, P., Quesada, V., Quilez, J., Li, G., García, F., Rubio-Camarillo, M., Frias, L., Ribeca, P., Capella-Gutiérrez, S., Rodríguez, J. M., Câmara, F., ... Godoy, J. A. (2016). Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered iberian lynx. *Genome Biol.*, *17*, 251.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., & Moreno, R. F. (1991). Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science*, *252*, 1651–1656.

Agrawal, A. F., & Whitlock, M. C. (2012). Mutation load: The fitness of individuals in populations where deleterious alleles are abundant. *Annu. Rev. Ecol. Evol. Syst.*, *43*, 115–135.

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., Howe, K., Kähäri, A., Kokocinski, F., Martin, F. J., Murphy, D. N., Nag, R., Ruffier, M., Schuster, M., Tang, Y. A., ... Searle, S. M. J. (2016). The ensembl gene annotation system. *Database*, *2016*.

Akopyan, M., Genchev, M., Armstrong, E. E., & Mooney, J. A. (2025). Reference genome choice compromises population genetic analyses. *Cell*.

Ala-Honkola, O., Uddström, A., Pauli, B. D., & Lindström, K. (2009). Strong inbreeding depression in male mating behaviour in a poeciliid fish. *J. Evol. Biol.*, *22*, 1396–1406.

Alemu, S. W., Kadri, N. K., Harland, C., Faux, P., Charlier, C., Caballero, A., & Druet, T. (2021). An evaluation of inbreeding measures using a whole-genome sequenced cattle pedigree. *Heredity (Edinb.)*, *126*, 410–423.

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

Andrews, S. (2010). *Babraham bioinformatics - FastQC a quality control tool for high throughput sequence data*.

ANN vs CNN vs RNN: Neural networks guide. (n.d.).

Armstrong, E. E., Taylor, R. W., Miller, D. E., Kaelin, C. B., Barsh, G. S., Hadly, E. A., & Petrov, D. (2020). Long live the king: Chromosome-level assembly of the lion (panthera leo) using linked-read, hi-C, and long-read data [Chromosome level assembly from captive lion. Synteny is highly conserved between lion, Panthera and domestic cat. Variability in ROH across lion genomes - inbreeding/bottleneck? Choice of reference genome important in inferring demographic history and comparing heterozygosity across species.]. *BMC Biol.*, *18*, 3.

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E. D., . . . Paten, B. (2020). Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, *587*, 246–251.

ATAC-seq data standards and processing pipeline. (n.d.).

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, *18*, 1196–1203.

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M. T., Perloski, M., Liberg, O., Arnemo, J. M., Hedhammar, A., & Lindblad-Toh, K.

(2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, *495*, 360–364.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv [cs.LG]*.

Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*, 26–33.

Banks, S. C., Cary, G. J., Smith, A. L., Davies, I. D., Driscoll, D. A., Gill, A. M., Lindenmayer, D. B., & Peakall, R. (2013). How does ecological disturbance influence genetic diversity? *Trends Ecol. Evol.*, *28*, 670–679.

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., & Marth, G. T. (2011). BamTools: A c++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, *27*, 1691–1692.

Barnett, R., Barnes, I., Phillips, M. J., Martin, L. D., Harington, C. R., Leonard, J. A., & Cooper, A. (2005). Evolution of the extinct sabretooths and the american cheetah-like cat. *Curr. Biol.*, *15*, R589–90.

Barnosky, A. D., Koch, P. L., Feranec, R. S., Wing, S. L., & Shabel, A. B. (2004). Assessing the causes of late pleistocene extinctions on the continents. *Science*, *306*, 70–75.

Beiki, H., Liu, H., Huang, J., Manchanda, N., Nonneman, D., Smith, T. P. L., Reecy, J. M., & Tuggle, C. K. (2019). Improved annotation of the domestic pig genome through integration of iso-seq and RNA-seq data. *BMC Genomics*, *20*, 344.

Belbachir, F., Pettorelli, N., Wacher, T., Belbachir-Bazi, A., & Durant, S. M. (2015). Monitoring rarity: The critically endangered saharan cheetah as a flagship species for a threatened ecosystem. *PLoS One*, *10*, e0115136.

Bell, K. (2005). Mortality in hand reared cheetah cubs, 306–314.

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*, *58*, 268–276.

Bertola, L. D., Vermaat, M., Lesilau, F., Chege, M., Tumenta, P. N., Sogbohossou, E. A., Schaap, O. D., Bauer, H., Patterson, B. D., White, P. A., de Iongh, H. H., Laros, J. F. J., & Vrieling, K. (2022). Whole genome sequencing and the application of a SNP panel reveal primary evolutionary lineages and genomic variation in the lion (panthera leo). *BMC Genomics*, *23*, 321.

Bertorelle, G., Raffini, F., Bosse, M., Bortoluzzi, C., Iannucci, A., Trucchi, E., Morales, H. E., & van Oosterhout, C. (2022). Genetic load: Genomic estimates and applications in non-model animals. *Nat. Rev. Genet.*

Bijlsma, R., & Loeschcke, V. (2012). Genetic erosion impedes adaptive responses to stressful environments. *Evol. Appl.*, *5*, 117–129.

Björnerfeldt, S., Webster, M. T., & Vilà, C. (2006). Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome Res.*, *16*, 990–994.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., & Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, *14*, 708–715.

Boakes, E. H., Wang, J., & Amos, W. (2007). An investigation of inbreeding depression and purging in captive pedigreed populations. *Heredity (Edinb.)*, *98*, 172–182.

Boddy, A. M., Abegglen, L. M., Pessier, A. P., Aktipis, A., Schiffman, J. D., Maley, C. C., & Witte, C. (2020). Lifetime cancer prevalence and life history traits in mammals. *Evol. Med. Public Health*, *2020*, 187–195.

Boeva, V. (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Front. Genet.*, *7*, 24.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, *30*, 2114–2120.

Bosse, M., Megens, H.-J., Derks, M. F. L., de Cara, Á. M. R., & Groenen, M. A. M. (2019). Deleterious alleles in the context of domestication, inbreeding,

and selection [Good intro. Reviews inbreeding/mutational load in domestic species]. *Evol. Appl.*, *12*, 6–17.

Branco, P., Torgo, L., & Ribeiro, R. P. (2017). A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, *49*, 1–50.

Broquet, T., Angelone, S., Jaquiery, J., Joly, P., Lena, J.-P., Lengagne, T., Plenet, S., Luquet, E., & Perrin, N. (2010). Genetic bottlenecks driven by population disconnection. *Conserv. Biol.*, *24*, 1596–1605.

Browning, S. R., & Browning, B. L. (2015). Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet.*, *97*, 404–418.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, *10*, 1213–1218.

Buk, K. G., van der Merwe, V. C., Marnewick, K., & Funston, P. J. (2018). Conservation of severely fragmented populations: Lessons from the transformation of uncoordinated reintroductions of cheetahs (acinonyx jubatus) into a managed metapopulation with self-sustained growth. *Biodivers. Conserv.*, *27*, 3393–3423.

Burcin, M., Arnold, R., Lutz, M., Kaiser, B., Runge, D., Lottspeich, F., Filippova, G. N., Lobanenkov, V. V., & Renkawitz, R. (1997). Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. *Mol. Cell. Biol.*, *17*, 1281–1288.

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, *268*, 78–94.

Caballero, A., Villanueva, B., & Druet, T. (2021). On the estimation of inbreeding depression using different measures of inbreeding from molecular markers. *Evol. Appl.*, *14*, 416–428.

Caballero-Campo, P., Buffone, M. G., Benencia, F., Conejo-García, J. R., Rinaudo, P. F., & Gerton, G. L. (2014). A role for the chemokine receptor CCR6 in mammalian sperm motility and chemotaxis. *J. Cell. Physiol.*, *229*, 68–78.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., & Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, *18*, 188–196.

Carbone, C., & Gittleman, J. L. (2002). A common rule for the scaling of carnivore density. *Science*, *295*, 2273–2276.

Caron, B., Luo, Y., & Rausell, A. (2019). NCBoost classifies pathogenic non-coding variants in mendelian diseases through supervised learning on purifying selection signals in humans. *Genome Biol.*, *20*, 32.

Carroll, S. B. (2000). Endless forms: The evolution minireview of gene regulation and morphological diversity. *Cell*, *101*, 577–580.

Casals, F., Hodgkinson, A., Hussin, J., Idaghdour, Y., Bruat, V., de Maillard, T., Grenier, J.-C., Gbeha, E., Hamdan, F. F., Girard, S., Spinella, J.-F., Larivière, M., Saillour, V., Healy, J., Fernández, I., Sinnett, D., Michaud, J. L., Rouleau, G. A., Haddad, E., ... Awadalla, P. (2013). Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet.*, *9*, e1003815.

Castro-Prieto, A., Wachter, B., & Sommer, S. (2011). Cheetah paradigm revisited: MHC diversity in the world's largest free-ranging population. *Mol. Biol. Evol.*, *28*, 1455–1468.

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., & Wilson, J. F. (2018). Runs of homozygosity: Windows into population history and trait architecture. *Nat. Rev. Genet.*, *19*, 220–234.

Ceballos, G., & Ehrlich, P. R. (2023). Mutilation of the tree of life via mass extinction of animal genera. *Proc. Natl. Acad. Sci. U. S. A.*, *120*, e2306987120.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215.

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Jr, Shapiro, M. D., Brady, S. D., Southwick, A. M., Absher, D. M., Grimwood, J., Schmutz, J., Myers, R. M., Petrov, D., Jónsson, B., Schluter, D., Bell, M. A., & Kingsley, D. M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science*, *327*, 302–305.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*, 7.

Chao, A., Thun, M. J., Connell, C. J., McCullough, M. L., Jacobs, E. J., Flanders, W. D., Rodriguez, C., Sinha, R., & Calle, E. E. (2005). Meat consumption and risk of colorectal cancer. *JAMA*, *293*, 172–182.

Charlesworth, B., & Charlesworth, D. (1999). The genetic basis of inbreeding depression. *Genet. Res.*, *74*, 329–340.

Charlesworth, B., & Charlesworth, D. (2000). The degeneration of Y chromosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *355*, 1563–1572.

Charlesworth, D. H., & Charlesworth, B. (1987). Inbreeding depression and its evolutionary consequences. *Annu. Rev. Ecol. Syst.*, *18*, 237–268.

Charlesworth, D. (2003). Effects of inbreeding on the genetic diversity of populations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *358*, 1051–1070.

Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. *Nat. Rev. Genet.*, *10*, 783–796.

Charruau, P., Fernandes, C., Orozco-Terwengel, P., Peters, J., Hunter, L., Ziaie, H., Jourabchian, A., Jowkar, H., Schaller, G., Ostrowski, S., Vercammen, P., Grange, T., Schlötterer, C., Kotze, A., Geigl, E.-M., Walzer, C., & Burger, P. A. (2011). Phylogeography, genetic structure and population divergence time of cheetahs in africa and asia: Evidence for long-term geographic isolates. *Mol. Ecol.*, *20*, 706–724.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, *16*, 321–357.

Chen, K., Durand, D., & Farach-Colton, M. (2000). NOTUNG: A program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.*, *7*, 429–447.

Chen, L., Wang, Y., & Zhao, F. (2022). Exploiting deep transfer learning for the prediction of functional non-coding variants using genomic sequence. *Bioinformatics*, *38*, 3164–3172.

Cheng, J., Li, T., Zheng, Z., Zhang, X., Cao, M., Tang, W., Hong, K., Zheng, R., Shao, J., Zhao, X., Jiang, H., Xu, W., & Lin, H. (2023). Loss of histone reader Phf7 leads to immune pathways activation via endogenous retroviruses during spermiogenesis. *iScience*, *26*, 108030.

Christie, M. R., Marine, M. L., Fox, S. E., French, R. A., & Blouin, M. S. (2016). A single generation of domestication heritably alters the expression of hundreds of genes. *Nat. Commun.*, *7*, 10676.

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods Ecol. Evol.*, *10*, 1632–1644.

Christmas, M. J., Kaplow, I. M., Genereux, D. P., Dong, M. X., Hughes, G. M., Li, X., Sullivan, P. F., Hindle, A. G., Andrews, G., Armstrong, J. C., Bianchi, M., Breit, A. M., Diekhans, M., Fanter, C., Foley, N. M., Goodman, D. B., Goodman, L., Keough, K. C., Kirilenko, B., ... Karlsson, E. K. (2023). Evolutionary constraint and innovation across hundreds of placental mammals. *Science*, *380*, eabn3943.

Church, D. M. (2022). A next-generation human genome sequence. *Science*, *376*, 34–35.

Ciezarek, A., Ford, A. G. P., Etherington, G. J., Kasozi, N., Malinsky, M., Mehta, T. K., Penso-Dolfin, L., Ngatunga, B. P., Shechonge, A., Tamatamah, R., Haerty, W., Di Palma, F., Genner, M. J., & Turner, G. F. (2022). Whole genome resequencing data enables a targeted SNP panel for conservation and aquaculture of oreochromis cichlid fishes. *Aquaculture*, *548*, 737637.

Cingolani, P., Platts, A., Wang Le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012a). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, *6*, 80–92.

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012b). Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.*, *3*, 35.

Clifford, D. L., Woodroffe, R., Garcelon, D. K., Timm, S. F., & Mazet, J. A. K. (2007). Using pregnancy rates and perinatal mortality to evaluate the success of recovery strategies for endangered island foxes. *Anim. Conserv.*, *10*, 442–451.

Cochran, K., Srivastava, D., Shrikumar, A., Balsubramani, A., Hardison, R. C., Kundaje, A., & Mahony, S. (2022). Domain adaptive neural networks improve cross-species prediction of transcription factor binding. *Genome Res.*

Colangelo, P., Di Civita, M., Bento, C. M., Franchini, P., Meyer, A., Orel, N., das Neves, L. C. B. G., Mulandane, F. C., Almeida, J. S., Senczuk, G., Pilla, F., & Sabatelli, S. (2024). Genome-wide diversity, population structure and signatures of inbreeding in the african buffalo in mozambique. *BMC Ecol Evol*, *24*, 29.

Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). (1992). Quotas for trade in specimens of cheetah.

Cooper, G. M., Stone, E. A., Asimenos, G., NISC Comparative Sequencing Program, Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, *15*, 901–913.

Couvet, D., & Ronfort, J. (1994). Mutation load depending on variance in reproductive success and mating system. In *Conservation genetics* (pp. 55–68). Birkhäuser Basel.

Crosier, A. E., Lamy, J., Bapodra, P., Rapp, S., Maly, M., Junge, R., Haefele, H., Ahistus, J., Santiestevan, J., & Comizzoli, P. (2020). First birth of cheetah cubs from in vitro fertilization and embryo transfer. *Animals (Basel)*, *10*, 1811.

Crosier, A. E., Marker, L., Howard, J., Pukazhenthi, B. S., Henghali, J. N., & Wildt, D. E. (2007). Ejaculate traits in the namibian cheetah (acinonyx jubatus): Influence of age, season and captivity. *Reprod. Fertil. Dev.*, *19*, 370–382.

Crosier, A. E., Wachter, B., Schulman, M., Lüders, I., Koester, D. C., Wielebnowski, N., Comizzoli, P., & Marker, L. (2018). Reproductive physiology of the cheetah and assisted reproductive techniques. In *Cheetahs: Biology and conservation* (pp. 385–402). Elsevier.

Crow, J. F., & Kimura, M. (1970). *An introduction to population genetics theory*. Burgess Publishing Company.

Cruz, F., Vilà, C., & Webster, M. T. (2008). The legacy of domestication: Accumulation of deleterious mutations in the dog genome. *Mol. Biol. Evol.*, *25*, 2331–2336.

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., ... Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Res.*, *50*, D988–D995.

Curik, I., Ferenčaković, M., & Sölkner, J. (2014). Inbreeding and runs of homozygosity: A possible solution to an old problem. *Livest. Sci.*, *166*, 26–34.

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R. F., Liao, X., Djari, A., Rodriguez, S. C., Grohs, C., Esquerré, D., Bouchez, O., Rossignol, M.-N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P. J., Coote, D., ... Hayes, B. J. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.*, *46*, 858–865.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*.

Darwin Tree of Life Project Consortium. (2022). Sequence locally, think globally: The darwin tree of life project. *Proc. Natl. Acad. Sci. U. S. A.*, *119*.

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: Next-generation population genetics. *Brief. Funct. Genomics*, *9*, 416–423.

Davis, B. W., Li, G., & Murphy, W. J. (2010). Supermatrix and species tree methods resolve phylogenetic relationships within the big cats, panthera (carnivora: Felidae). *Mol. Phylogenet. Evol.*, *56*, 64–76.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, *6*, e1001025.

De Boeck, K., Zolin, A., Cuppens, H., Olesen, H. V., & Viviani, L. (2014). The relative frequency of CFTR mutation classes in european patients with cystic fibrosis. *J. Cyst. Fibros.*, *13*, 403–409.

Dermitzakis, E. T., & Clark, A. G. (2002). Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. *Mol. Biol. Evol.*, *19*, 1114–1121.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv [cs.CL]*.

Dhaka, V. S., Meena, S. V., Rani, G., Sinwar, D., Kavita, Ijaz, M. F., & Woźniak, M. (2021). A survey of deep convolutional neural networks applied for prediction of plant leaf diseases. *Sensors (Basel)*, *21*, 4749.

Dhople, V., Krukemeyer, A., & Ramamoorthy, A. (2006). The human beta-defensin-3, an antibacterial peptide with multiple biological functions. *Biochim. Biophys. Acta*, *1758*, 1499–1512.

Dobrynin, P., Liu, S., Tamazian, G., Xiong, Z., Yurchenko, A. A., Krasheninnikova, K., Kliver, S., Schmidt-Küntzel, A., Koepfli, K.-P., Johnson, W., Kuderna, L. F. K., García-Pérez, R., Manuel, M. d., Godinez, R., Komissarov, A., Makunin, A., Brukhin, V., Qiu, W., Zhou, L., ... O'Brien, S. J. (2015). Genomic legacy of the african cheetah, acinonyx jubatus. *Genome Biol.*, *16*, 277.

Driscoll, C. A., Menotti-Raymond, M., Nelson, G., Goldstein, D., & O'Brien, S. J. (2002). Genomic microsatellites as evolutionary chronometers: A test in wild cats. *Genome Res.*, *12*, 414–423.

Drosophila 12 Genomes Consortium, Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., Pollard, D. A., Sackton, T. B., Larracuente, A. M., Singh, N. D., Abad, J. P., Abt, D. N., Adryan, B., Aguade, M., ... MacCallum, I. (2007). Evolution of genes and genomes on the drosophila phylogeny. *Nature*, *450*, 203–218.

Dudnyk, K., Cai, D., Shi, C., Xu, J., & Zhou, J. (2024). Sequence basis of transcription initiation in the human genome. *Science*, *384*, eadj0116.

Durant, S. M., Groom, R., Ipavec, A., Mitchell, N., & Khalatbari, L. (2021, May 17). *Acinonyx jubatus (IUCN)*. IUCN.

Durant, S. M., Mitchell, N., Groom, R., Pettorelli, N., Ipavec, A., Jacobson, A. P., Woodroffe, R., Böhm, M., Hunter, L. T. B., Becker, M. S., Broekhuis, F., Bashir, S., Andresen, L., Aschenborn, O., Beddiaf, M., Belbachir, F., Belbachir-Bazi, A., Berbash, A., Brandao de Matos Machado, I., ... Young-Overton, K. (2017). The global decline of cheetah acinonyx jubatus and what it means for conservation. *Proc. Natl. Acad. Sci. U. S. A.*, *114*, 528–533.

Dussex, N., Morales, H. E., Grossen, C., Dalén, L., & van Oosterhout, C. (2023). Purging and accumulation of genetic load in conservation. *Trends Ecol. Evol.*, *38*, 961–969.

Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: Illuminating the dark road from association to function. *Am. J. Hum. Genet.*, *93*, 779–797.

Elango, N., & Yi, S. V. (2011). Functional relevance of CpG island length for regulation of gene expression. *Genetics*, *187*, 1077–1083.

Elkon, R., & Agami, R. (2017). Characterization of noncoding regulatory DNA in the human genome. *Nat. Biotechnol.*, *35*, 732–746.

Elnitski, L., Jin, V. X., Farnham, P. J., & Jones, S. J. M. (2006). Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. *Genome Res.*, *16*, 1455–1464.

Emms, D. M., Liu, Y., Belcher, L. J., Holmes, J., & Kelly, S. (2025). OrthoFinder: Scalable phylogenetic orthology inference for comparative genomics. *bioRxiv*.

ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science*, *306*, 636–640.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*, 57–74.

England, P. R., Cornuet, J.-M., Berthier, P., Tallmon, D. A., & Luikart, G. (2006). Estimating effective population size from linkage disequilibrium: Severe bias in small samples. *Conserv. Genet.*, *7*, 303–308.

Ernst, K. J., Okonechnikov, K., Bageritz, J., Perera, A. A., Mallm, J.-P., Wittmann, A., Maaß, K. K., Leible, S., Boutros, M., Pfister, S. M., Zuckermann, M., & Jones, D. T. W. (2025). A simplified preparation method for single-nucleus RNA-sequencing using long-term frozen brain tumor tissues. *Sci. Rep.*, *15*, 12849.

Evermann, J. F., Heeney, J. L., Roelke, M. E., McKeirnan, A. J., & O'Brien, S. J. (1988). Biological and pathological consequences of feline infectious peritonitis virus infection in the cheetah. *Arch. Virol.*, *102*, 155–171.

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

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.*, *38*, 276–278.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet.*, *9*, e1003905.

Fabiano, E. C., Bonatto, S. L., Schmidt-Küntzel, A., O'Brien, S. J., Marker, L., & Eizirik, E. (2025). Inferring the historical demography of southern african cheetahs (acinonyx jubatus) using bayesian analyses of molecular genetic data. *Genet. Mol. Biol.*, *48*, e20240253.

Farhadinia, M. S., Hunter, L. T. B., Jowkar, H., Schaller, G. B., & Ostrowski, S. (2018). Asiatic cheetahs in iran: Decline, current status and threats. In *Cheetahs: Biology and conservation* (pp. 55–69). Elsevier.

Fatima, A., Hoeber, J., Schuster, J., Koshimizu, E., Maya-Gonzalez, C., Keren, B., Mignot, C., Akram, T., Ali, Z., Miyatake, S., Tanigawa, J., Koike, T., Kato, M., Murakami, Y., Abdullah, U., Ali, M. A., Fadoul, R., Laan, L., Castillejo-López, C., ... Dahl, N. (2021). Monoallelic and bi-allelic variants in NCDN cause neurodevelopmental delay, intellectual disability, and epilepsy. *Am. J. Hum. Genet.*, *108*, 739–748.

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, *78*, 737–756.

Fernández, J., & Caballero, A. (2001). Accumulation of deleterious mutations and equalization of parental contributions in the conservation of genetic resources. *Heredity (Edinb.)*, *86*, 480–488.

Figueiró, H. V., Li, G., Trindade, F. J., Assis, J., Pais, F., Fernandes, G., Santos, S. H. D., Hughes, G. M., Komissarov, A., Antunes, A., Trinca, C. S., Rodrigues, M. R., Linderoth, T., Bi, K., Silveira, L., Azevedo, F. C. C., Kantek, D., Ramalho, E., Brassaloti, R. A., ... Eizirik, E. (2017). Genome-wide signatures of complex introgression and adaptive evolution in the big cats. *Sci Adv*, *3*, e1700299.

Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P. E., Collins, S. J., & Lobanenkov, V. V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol. Cell. Biol.*, *16*, 2802–2813.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., & Merrick, J. M. (1995). Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, *269*, 496–512.

Foote, A. D., Hooper, R., Alexander, A., Baird, R. W., Baker, C. S., Ballance, L., Barlow, J., Brownlow, A., Collins, T., Constantine, R., Dalla Rosa, L., Davison, N. J., Durban, J. W., Esteban, R., Excoffier, L., Martin, S. L. F., Forney, K. A., Gerrodette, T., Gilbert, M. T. P., ... Morin, P. A. (2021). Runs of homozygosity in killer whale genomes provide a global record of demographic histories. *Mol. Ecol.*, *30*, 6162–6177.

Frankham, R. (2008). Genetic adaptation to captivity in species conservation programs. *Mol. Ecol.*, *17*, 325–333.

Fraser, D. J. (2008). How well can captive breeding programs conserve biodiversity? a review of salmonids: Genetic diversity and fitness in captive breeding. *Evol. Appl.*, *1*, 535–586.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., & Marks, D. S. (2021). Disease variant prediction with deep generative models of evolutionary data [new approach using deep generative models to predict variant pathogenicity]. *Nature*.

Fu, Q., Pandey, R. R., Leu, N. A., Pillai, R. S., & Wang, P. J. (2016). Mutations in the MOV10L1 ATP hydrolysis motif cause piRNA biogenesis failure and male sterility in mice. *Biol. Reprod.*, *95*, 103.

Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S., & Engreitz, J. M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*, *354*, 769–773.

Galtier, N., Depaulis, F., & Barton, N. H. (2000). Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics*, *155*, 981–987.

Gandolfo, L. C., Bahlo, M., & Speed, T. P. (2014). Dating rare mutations from small samples with dense marker data. *Genetics*, *197*, 1315–1327.

Gao, C.-H., Chen, C., Akyol, T., Dusa, A., Yu, G., Cao, B., & Cai, P. (2024). ggVennDiagram: Intuitive venn diagram software extended. *Imeta*, *3*, e177.

Garner, A., Rachlow, J. L., & Hicks, J. F. (2005). Patterns of genetic diversity and its loss in mammalian populations: Mammalian genetic diversity. *Conserv. Biol.*, *19*, 1215–1221.

GATK Team. (2025, January 22). *(how to) filter variants either with VQSR or by hard-filtering*.

Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, *36*, 2628–2629.

Ghandi, M., Lee, D., Mohammad-Noori, M., & Beer, M. A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.*, *10*, e1003711.

Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., & Japkowicz, N. (2024). The class imbalance problem in deep learning. *Mach. Learn.*, *113*, 4845–4901.

Gilpin, M. E., & Soulé, M. E. (1986). Minimum viable populations: Processes of extinction. In M. E. Soulé (Ed.), *Conservation biology: The science of scarcity and diversity* (pp. 19–34).

Gimranov, D. O., Madurell-Malapeira, J., Jiangzuo, Q., Lavrov, A. V., & Lopatin, A. V. (2024). Cheetah acinonyx pardinensis (felidae, carnivora) from the early pleistocene of crimea (taurida cave). *Dokl. Biol. Sci.*, *518*, 234–238.

Giontella, A., Cardinali, I., Lancioni, H., Giovannini, S., Pieramati, C., Silvestrelli, M., & Sarti, F. M. (2020). Mitochondrial DNA survey reveals the lack of accuracy in maremmano horse studbook records. *Animals (Basel)*, *10*, 839.

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.*, *17*, 877–885.

Gopalaswamy, A. M., Khalatbari, L., Chellam, R., Mills, M. G. L., Vanak, A. T., Thuo, D., Karanth, K. U., & Broekhuis, F. (2022). Introducing african cheetahs to india is an ill-advised conservation attempt. *Nat. Ecol. Evol.*, *6*, 1794–1795.

Gormley, B. (2023, December 21). Startup fauna bio studies animal genomes for clues to human diseases.

Gottelli, D., Wang, J., Bashir, S., & Durant, S. M. (2007). Genetic analysis reveals promiscuity among female cheetahs. *Proc. Biol. Sci.*, *274*, 1993–2001.

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, *27*, 1017–1018.

Grillo, G., Boyarchuk, E., Mihic, S., Ivkovic, I., Bertrand, M., Jouneau, A., Dahlet, T., Dumas, M., Weber, M., Velasco, G., & Francastel, C. (2025). ZBTB24 is a conserved multifaceted transcription factor at genes and centromeres that governs the DNA methylation state and expression of satellite repeats. *Hum. Mol. Genet.*, *34*, 161–177.

Grossen, C., Guillaume, F., Keller, L. F., & Croll, D. (2020). Purging of highly deleterious mutations through severe bottlenecks in alpine ibex. *Nat. Commun.*, *11*, 1001.

GTEx Consortium. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.*, *45*, 580–585.

GTEx portal. (n.d.).

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.*, *5*, e1000695.

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.*, *19*, 621–637.

Haldane, J. B. S. (1937). The effect of variation of fitness. *Am. Nat.*, *71*, 337–349.

Hamilton, C. K., Verduzco-Gómez, A. R., Favetta, L. A., Blondin, P., & King, W. A. (2012). Testis-specific protein Y-encoded copy number is correlated to its expression and the field fertility of canadian holstein bulls. *Sex Dev.*, *6*, 231–239.

Haque, M. M., Skinner, M. K., & Holder, L. B. (2014). Imbalanced class learning in epigenetics. *J. Comput. Biol.*, *21*, 492–507.

Haratake, N., Hu, Q., Okamoto, T., Jogo, T., Toyokawa, G., Kinoshita, F., Takenaka, T., Tagawa, T., Iseda, N., Itoh, S., Yamada, Y., Oda, Y., Shimokawa, M., Kikutake, C., Suyama, M., Unoki, M., Sasaki, H., & Mori, M. (2021). Identification of SLC38A7 as a prognostic marker and potential therapeutic target of lung squamous cell carcinoma. *Ann. Surg.*, *274*, 500–507.

Hark, A. T., Schoenherr, C. J., Katz, D. J., Ingram, R. S., Levorse, J. M., & Tilghman, S. M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, *405*, 486–489.

Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*. Pennsylvania State University.

Hasselgren, M., Dussex, N., von Seth, J., Angerbjörn, A., Dalén, L., & Norén, K. (2024). Strongly deleterious mutations influence reproductive output and longevity in an endangered population. *Nat. Commun.*, *15*, 8378.

Hedrick, P. W. (1994). Purging inbreeding depression and the probability of extinction: Full-sib mating. *Heredity*, *73 (Pt 4)*, 363–372.

Hedrick, P. W. (1996). Bottleneck(s) or metapopulation in cheetahs. *Conserv. Biol.*, *10*, 897–899.

Hedrick, P. W., & Garcia-Dorado, A. (2016). Understanding inbreeding depression, purging, and genetic rescue. *Trends Ecol. Evol.*, *31*, 940–952.

Hedrick, P. W., & Miller, P. S. (1992). Conservation genetics: Techniques and fundamentals. *Ecol. Appl.*, *2*, 30–46.

Heinrich, S. K., Hofer, H., Courtiol, A., Melzheimer, J., Dehnhard, M., Czirják, G. Á., & Wachter, B. (2017). Cheetahs have a stronger constitutive innate immunity than leopards. *Sci. Rep.*, *7*, 44837.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, *38*, 576–589.

Hemmer, H., Kahlke, R.-D., & Vekua, A. K. (2011). The cheetah acinonyx pardinensis (croizet et jobert, 1828) s.l. at the hominin site of dmanisi (georgia) – a potential prime meat supplier in early pleistocene ecosystems. *Quat. Sci. Rev.*, *30*, 2703–2714.

Hess, J. E., Campbell, N. R., Docker, M. F., Baker, C., Jackson, A., Lampman, R., McIlraith, B., Moser, M. L., Statler, D. P., Young, W. P., Wildbill, A. J., & Narum, S. R. (2015). Use of genotyping by sequencing data to develop a high-throughput and multifunctional SNP panel for conservation applications in pacific lamprey. *Mol. Ecol. Resour.*, *15*, 187–202.

Hickey, G., Paten, B., Earl, D., Zerbino, D., & Haussler, D. (2013). HAL: A hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, *29*, 1341–1342.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits [talks about SNPs in noncoding sequences being associated with disease]. *Proc. Natl. Acad. Sci. U. S. A.*, *106*, 9362–9367.

Hirawatari, K., Hanzawa, N., Kuwahara, M., Aoyama, H., Miura, I., Wakana, S., & Gotoh, H. (2015). Polygenic expression of teratozoospermia and normal fertility in B10.MOL-TEN1 mouse strain: Teratozoospermia and normal fertility. *Congenit. Anom. (Kyoto)*, *55*, 92–98.

Hockla, A., Miller, E., Salameh, M. A., Copland, J. A., Radisky, D. C., & Radisky, E. S. (2012). PRSS3/mesotrypsin is a therapeutic target for metastatic prostate cancer. *Mol. Cancer Res.*, *10*, 1555–1566.

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1: Unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, *32*, 767–769.

Hoff, K. J., & Stanke, M. (2019). Predicting genes in single genomes with AUGUSTUS. *Curr. Protoc. Bioinformatics*, *65*, e57.

Hsu, P. D., Lander, E. S., & Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, *157*, 1262–1278.

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Hum. Immunol.*, *82*, 801–811.

Huang, Y.-F., Gulko, B., & Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.*, *49*, 618–624.

Huber, C. D., Kim, B. Y., & Lohmueller, K. E. (2020). Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. *PLoS Genet.*, *16*, e1008827.

Hubisz, M. J., Pollard, K. S., & Siepel, A. (2011). PHAST and RPHAST: Phylogenetic analysis with space/time models. *Brief. Bioinform.*, *12*, 41–51.

Iso-Touru, T., Wurmser, C., Venhoranta, H., Hiltpold, M., Savolainen, T., Sironen, A., Fischer, K., Flisikowski, K., Fries, R., Vicente-Carrillo, A., Alvarez-Rodriguez, M., Nagy, S., Mutikainen, M., Peippo, J., Taponen, J., Sahana, G., Guldbrandtsen, B., Simonen, H., Rodriguez-Martinez, H., ... Pausch, H. (2019). A splice donor variant in CCDC189 is associated with asthenospermia in nordic red dairy cattle. *BMC Genomics*, *20*, 286.

IUCN/SSC. (2015). *Review of the regional conservation strategy for the cheetah and african wild dogs in southern africa* (research rep.). IUCN/SSC Gland, Switzerland, Range Wide Conservation Program for Cheetah, and African Wild Dogs.

Ivy, J. A., & Lacy, R. C. (2010, June 14). Using molecular methods to improve the genetic management of captive breeding programs for threatened species. In J. A. DeWoody, J. W. Bickham, C. H. Michler, K. M. Nichols, G. E. Rhodes, & K. E. Woeste (Eds.), *Molecular approaches in natural resource conservation and management* (pp. 267–295). Cambridge University Press.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The oxford nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol.*, *17*, 239.

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data recommendations for the use of performance metrics. *Int. Conf. Affect. Comput. Intell. Interact. Workshops*, *2013*, 245–251.

Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, *37*, 2112–2120.

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, *316*, 1497–1502.

Johnson, W. E., Eizirik, E., Pecon-Slattery, J., Murphy, W. J., Antunes, A., Teeling, E., & O'Brien, S. J. (2006). The late miocene radiation of modern felidae: A genetic assessment. *Science*, *311*, 73–77.

Johnson, W. E., Onorato, D. P., Roelke, M. E., Land, E. D., Cunningham, M., Belden, R. C., McBride, R., Jansen, D., Lotz, M., Shindle, D., Howard, J., Wildt, D. E., Penfold, L. M., Hostetler, J. A., Oli, M. K., & O'Brien, S. J. (2010). Genetic restoration of the florida panther. *Science*, *329*, 1641–1645.

Jordheim, L. P. (2018). Expanding the clinical relevance of the 5'-nucleotidase cN-II/NT5C2. *Purinergic Signal.*, *14*, 321–329.

Joshi, M., Kapopoulou, A., & Laurent, S. (2021). Impact of genetic variation in gene regulatory sequences: A population genomics perspective. *Front. Genet.*, *12*, 660899.

Jowkar, H., Hunter, L., Ziaie, H., Marker, L., Breitenmoser-Wursten, C., & Durant, S. (2008, June 30). *Acinonyx jubatus ssp. venaticus*. IUCN.

Kalinowski, S. T., Hedrick, P. W., & Miller, P. S. (2000). Inbreeding depression in the speke's gazelle captive breeding program. *Conserv. Biol.*, *14*, 1375–1384.

Kaplow, I. M., Lawler, A. J., Schäffer, D. E., Srinivasan, C., Sestili, H. H., Wirthlin, M. E., Phan, B. N., Prasad, K., Brown, A. R., Zhang, X., Foley, K., Genereux, D. P., Zoonomia Consortium**, Karlsson, E. K., Lindblad-Toh, K., Meyer, W. K., Pfenning, A. R., Andrews, G., Armstrong, J. C., ... Zhang, X. (2023). Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. *Science*, *380*, eabm7993.

Kardos, M., Åkesson, M., Fountain, T., Flagstad, Ø., Liberg, O., Olason, P., Sand, H., Wabakken, P., Wikenros, C., & Ellegren, H. (2018). Genomic conse-

quences of intensive inbreeding in an isolated wolf population. *Nat Ecol Evol*, *2*, 124–131.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, *30*, 772–780.

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., & Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.*, *10*, 1930.

Kelley, D. R. (2020). Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.*, *16*, e1008050.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., & Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.*, *28*, 739–750.

Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, *26*, 990–999.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., Dunham, I., Elnitski, L. L., Farnham, P. J., Feingold, E. A., Gerstein, M., Giddings, M. C., Gilbert, D. M., Gingeras, T. R., Green, E. D., ... Hardison, R. C. (2014). Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U. S. A.*, *111*, 6131–6138.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.*, *12*, 656–664.

Khajavi, M., Inoue, K., & Lupski, J. R. (2006). Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease. *Eur. J. Hum. Genet.*, *14*, 1074–1081.

Khan, A., Patel, K., Shukla, H., Viswanathan, A., van der Valk, T., Borthakur, U., Nigam, P., Zachariah, A., Jhala, Y. V., Kardos, M., & Ramakrishnan, U. (2021). Genomic evidence for inbreeding depression and purging of dele-

terious genetic variation in indian tigers. *Proc. Natl. Acad. Sci. U. S. A.*, *118*.

Kim, S., Cho, Y. S., Kim, H.-M., Chung, O., Kim, H., Jho, S., Seomun, H., Kim, J., Bang, W. Y., Kim, C., An, J., Bae, C. H., Bhak, Y., Jeon, S., Yoon, H., Kim, Y., Jun, J., Lee, H., Cho, S., ... Yeo, J.-H. (2016). Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly [Evolution of carnivory using mammal genomes incl. Felidae. All felidae are obligate carnivores. Present high-quality leopard genome assembly. Found contractions in gene families for starch/sucrose metabolism in carnivores. Carnivores under strong selective pressure related to diet. Felids showed recent reductions in gen. div. associated w/ decreased pop. sizes - could be due to inflexibility of diet. High level of genomic similarity within felids compared to hominids and bovids. Felidae - adaptations for muscle power, agility and specialised diet - make them good predators but vulnerable to extinction.]. *Genome Biol.*, *17*, 211.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Kincaid, H. L. (1983). Inbreeding in fish populations used for aquaculture. *Aquaculture*, *33*, 215–227.

King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, *188*, 107–116.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, *46*, 310–315.

Kleinman-Ruiz, D., Martínez-Cruz, B., Soriano, L., Lucena-Perez, M., Cruz, F., Villanueva, B., Fernández, J., & Godoy, J. A. (2017). Novel efficient genome-wide SNP panels for the conservation of the highly endangered iberian lynx. *BMC Genomics*, *18*, 556.

Koch, L. (2020). Exploring human genomic diversity with gnomAD. *Nat. Rev. Genet.*, *21*, 448.

Koester, D. C., Freeman, E. W., Brown, J. L., Wildt, D. E., Terrell, K. A., Franklin, A. D., & Crosier, A. E. (2015). Motile sperm output by male cheetahs (acinonyx jubatus) managed ex situ is influenced by public exposure and number of care-givers. *PLoS One*, *10*, e0135847.

Koester, D. C., Freeman, E. W., Wildt, D. E., Terrell, K. A., Franklin, A. D., Meeks, K., & Crosier, A. E. (2017). Group management influences reproductive function of the male cheetah (acinonyx jubatus). *Reprod. Fertil. Dev.*, *29*, 496–508.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*, 4453–4455.

Kraus, D., & Marker-Kraus, L. (1991). *Current status of the cheetah (acinonyx jubatus)*.

Krausman, P. R., & Morales, S. M. (2005). Acinonyx jubatus. *Mammalian Species*, *771*, 1.

Krueger, F. (n.d.). *TrimGalore: A wrapper around cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data* (comp. software).

Kryukov, G. V., Pennacchio, L. A., & Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *Am. J. Hum. Genet.*, *80*, 727–739.

Kryvokhyzha, D. (2016, September 22). *GATK: The best practice for genotype calling in a non-model organism*.

Kuehn, C., Edel, C., Weikard, R., & Thaller, G. (2007). Dominance and parent-of-origin effects of coding and non-coding alleles at the acylCoA-diacylglycerol-acyltransferase (DGAT1) gene on milk production traits in german holstein cows. *BMC Genet.*, *8*, 62.

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Brief. Bioinform.*, *14*, 144–161.

Kumar, M., Conroy, G., Ogbourne, S., Cairns, K., Borburgh, L., & Subramanian, S. (2023). Genomic signatures of bottleneck and founder effects in dingoes. *Ecol. Evol.*, *13*, e10525.

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., & Hedges, S. B. (2022). TimeTree 5: An expanded resource for species divergence times. *Mol. Biol. Evol.*, *39*.

Kyriazis, C. C., Robinson, J. A., & Lohmueller, K. E. (2025). Long runs of homozygosity are reliable genomic markers of inbreeding depression. *Trends Ecol. Evol.*, *40*, 874–884.

Lande, R. (1994). Risk of population extinction from fixation of new deleterious mutations. *Evolution*, *48*, 1460–1469.

Lande, R., & Barrowclough, G. F. (1987, August). Effective population size, genetic variation, and their use in population management. In *Viable populations for conservation* (pp. 87–124). Cambridge University Press.

Lande, R., Schemske, D. W., & Schultz, S. T. (1994). High inbreeding depression, selective interference among loci, and the threshold selfing rate for purging recessive lethal mutations. *Evolution*, *48*, 965–978.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., ... International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

Lappalainen, T., Scott, A. J., Brandt, M., & Hall, I. M. (2019). Genomic analysis in the age of human genome sequencing. *Cell*, *177*, 70–84.

Laurenson, M. K. (1994). High juvenile mortality in cheetahs (*Acinonyx jubatus*) and its consequences for maternal care. *J. Zool. (1987)*, *234*, 387–408.

le Maire, A., Teyssier, C., Balaguer, P., Bourguet, W., & Germain, P. (2019). Regulation of RXR-RAR heterodimers by RXR- and RAR-specific ligands and their combinations. *Cells*, *8*, 1392.

Leberg, P. L., & Firmin, B. D. (2008). Role of inbreeding depression and purging in captive breeding and restoration programmes. *Mol. Ecol.*, *17*, 334–343.

Lee, J., Christoforo, G., Christoforo, G., Foo, C. S., Probert, C., Kundaje, A., Boley, N., kohpangwei, Kim, D., & Dacre, M. (2016). *Kundaje-lab/atac_dnase_pipelines: 0.3.0* (comp. software). Zenodo.

Leigh, D. M., Lischer, H. E. L., Grossen, C., & Keller, L. F. (2018). Batch effects in a multiyear sequencing study: False biological trends due to changes in read lengths. *Mol. Ecol. Resour.*, *18*, 778–788.

Leung, K., Beukeboom, L. W., & Zwaan, B. J. (2025). Inbreeding and outbreeding depression in wild and captive insect populations. *Annu. Rev. Entomol.*, *70*, 271–292.

Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, *424*, 147–151.

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., ... Zhang, G. (2018). Earth BioGenome project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.*, *115*, 4325–4333.

Li, G., Hillier, L. W., Grahn, R. A., Zimin, A. V., David, V. A., Menotti-Raymond, M., Middleton, R., Hannah, S., Hendrickson, S., Makunin, A., O'Brien, S. J., Minx, P., Wilson, R. K., Lyons, L. A., Warren, W. C., & Murphy, W. J. (2016). A high-resolution SNP array-based linkage map anchors a new domestic cat draft genome assembly and provides detailed patterns of recombination. *G3*, *6*, 1607–1616.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*.

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

Li, W. H., Gojobori, T., & Nei, M. (1981). Pseudogenes as a paradigm of neutral evolution. *Nature*, *292*, 237–239.

Li, W., O'Neill, K. R., Haft, D. H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M. K., Durkin, A. S., Gonzales, N. R., Gwadz, M., Lanczycki, C. J., Song, J. S., Thanki, N., Wang, J., Yamashita, R. A., Yang, M., Zheng, C., ... Thibaud-Nissen, F. (2021). RefSeq: Expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res.*, *49*, D1020–D1028.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*, 923–930.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., ... Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals [Comparative analysis of 29 eutherian mammals]. *Nature*, *478*, 476–482.

Lindburg, D. G., Durrant, B. S., Millard, S. E., & Oosterhuis, J. E. (1993). Fertility assessment of cheetah males with poor quality semen. *Zoo Biol.*, *12*, 97–103.

Litt, M., & Luty, J. A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.*, *44*, 397–401.

Loewe, L., & Hill, W. G. (2010). The population genetics of mutations: Good, bad and indifferent. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *365*, 1153–1167.

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nat. Rev. Genet.*, *21*, 597–614.

Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., Clark, A. G., & Bustamante, C. D. (2008). Proportionally more deleterious genetic variation in european than in african populations. *Nature*, *451*, 994–997.

López-Cortegano, E., Moreno, E., & García-Dorado, A. (2021). Genetic purging in captive endangered ungulates with extremely low effective population sizes. *Heredity*, *127*, 433–442.

Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol. Ecol.*, *30*, 5966–5993.

Lou, R. N., & Therkildsen, N. O. (2022). Batch effects in population genomic studies with low-coverage whole genome sequencing data: Causes, detection and mitigation. *Mol. Ecol. Resour.*, *22*, 1678–1692.

Loughran, G., Chou, M.-Y., Ivanov, I. P., Jungreis, I., Kellis, M., Kiran, A. M., Baranov, P. V., & Atkins, J. F. (2014). Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res.*, *42*, 8928–8938.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, *15*, 550.

Lu, J., Tang, T., Tang, H., Huang, J., Shi, S., & Wu, C.-I. (2006). The accumulation of deleterious mutations in rice genomes: A hypothesis on the cost of domestication. *Trends Genet.*, *22*, 126–131.

Luikart, G., Sherwin, W. B., Steele, B. M., & Allendorf, F. W. (1998). Usefulness of molecular markers for detecting population bottlenecks via monitoring genetic change. *Mol. Ecol.*, *7*, 963–974.

Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U. K., Graham, K., Litton, C., Miyasato, S. R., Strattan, J. S., Jolanki, O., Lee, J.-W., Tanaka, F. Y., Adenekan, P., ... Cherry, J. M. (2020). New developments on the encyclopedia of DNA elements (ENCODE) data portal. *Nucleic Acids Res.*, *48*, D882–D889.

Lynch, M., Bürger, R., Butcher, D., & Gabriel, W. (1993). The mutational meltdown in asexual populations. *J. Hered.*, *84*, 339–344.

Lynch, M., & Gabriel, W. (1990). MUTATION LOAD AND THE SURVIVAL OF SMALL POPULATIONS. *Evolution*, *44*, 1725–1737.

Lynch, M., & O'Hely, M. (2001). Captive breeding and the genetic fitness of natural populations. *Conserv. Genet.*, *2*, 363–378.

Marker, L., & O'Brien, S. J. (1989). Captive breeding of the cheetah (acinonyx jubatus) in north american zoos (1871-1986). *Zoo Biology*, *8*, 3–16.

Marker, L. (2019, November 20). Cheetahs race for survival: Ecology and conservation. In *Wildlife population monitoring*. IntechOpen.

Marker, L., & Johnston, B. (2022). 2020 international cheetah studbook. *Cheetah Conservation Fund, Namibia*.

Marker, L. L., Pearks Wilkerson, A. J., Sarno, R. J., Martenson, J., Breitenmoser-Würsten, C., O'Brien, S. J., & Johnson, W. E. (2008). Molecular genetic insights on cheetah (acinonyx jubatus) ecology and conservation in namibia. *J. Hered.*, *99*, 2–13.

Marker-Kraus, L. (1988). *Smithsonian institution's national zoological park's NOAHS center*.

Marker-Kraus, L. (1997). History of the cheetah: Acinonyx jubatus in zoos 1829-1994. *Int. Zoo Yearb.*, *35*, 27–43.

Marker-Kraus, L., & Kraus, D. (1997). Conservation strategies for the long-term survival of the cheetah acinonyx jubatus by the cheetah conservation fund, windhoek. *Int. Zoo Yearb.*, *35*, 59–66.

Marnewick, K., & Somers, M. J. (2015). *Home ranges of cheetahs (acinonyx jubatus) outside protected areas in south africa*.

Marsden, C. D., Vecchyo, D. O.-D., O'Brien, D. P., Taylor, J. F., Ramirez, O., Vilà, C., Marques-Bonet, T., Schnabel, R. D., Wayne, R. K., & Lohmueller, K. E. (2016). Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences*, *113*, 152–157.

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

Masel, J. (2011). Genetic drift. *Curr. Biol.*, *21*, R837–8.

Maslova, A., Ramirez, R. N., Ma, K., Schmutz, H., Wang, C., Fox, C., Ng, B., Benoist, C., Mostafavi, S., & Project, I. G. (2020). Deep learning of immune

cell differentiation. *Proceedings of the National Academy of Sciences*, *117*, 25655–25666.

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, *7*, 29–59.

Mathur, S., & DeWoody, J. A. (2021). Genetic load has potential in large populations but is realized in small inbred populations. *Evol. Appl.*, *14*, 1540–1557.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, *20*, 1297–1303.

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., Macleod, A. K., Farrington, S. M., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S. H., Dunlop, M. G., Wright, A. F., ... Wilson, J. F. (2008). Runs of homozygosity in european populations. *Am. J. Hum. Genet.*, *83*, 359–372.

Meader, S., Ponting, C. P., & Lunter, G. (2010). Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res.*, *20*, 1335–1343.

Megquier, K., Genereux, D. P., Hekman, J., Swofford, R., Turner-Maier, J., Johnson, J., Alonso, J., Li, X., Morrill, K., Anguish, L. J., Koltookian, M., Logan, B., Sharp, C. R., Ferrer, L., Lindblad-Toh, K., Meyers-Wallen, V. N., Hoffman, A., & Karlsson, E. K. (2019). BarkBase: Epigenomic annotation of canine genomes. *Genes*, *10*.

Meißner, R., Mokgokong, P., Pretorius, C., Winter, S., Labuschagne, K., Kotze, A., Prost, S., Horin, P., Dalton, D., & Burger, P. A. (2024). Diversity of selected toll-like receptor genes in cheetahs (acinonyx jubatus) and african leopards (panthera pardus pardus). *Sci. Rep.*, *14*, 3756.

Melnikov, A., Zhang, X., Rogov, P., Wang, L., & Mikkelsen, T. S. (2014). Massively parallel reporter assays in cultured mammalian cells. *J. Vis. Exp.*

Menotti-Raymond, M., & O'Brien, S. J. (1993). Dating the genetic bottleneck of the african cheetah. *Proc. Natl. Acad. Sci. U. S. A.*, *90*, 3172–3176.

Miao, Z., Zhang, Y., Fabian, Z., Hernandez Celis, A., Beery, S., Li, C., Liu, Z., Gupta, A., Nasir, M., Li, W., Holmberg, J., Palmer, M., Gaynor, K., Arbelaez, P., Wang, P., Dodhia, R., & Ferres, J. L. (2025). New frontiers in artificial intelligence for biodiversity research and conservation with multimodal language models. *Methods Ecol. Evol.*

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information (Basel)*, *15*, 517.

Miller, W., Makova, K. D., Nekrutenko, A., & Hardison, R. C. (2004). Comparative genomics. *Annu. Rev. Genomics Hum. Genet.*, *5*, 15–56.

Min, X., Zeng, W., Chen, S., Chen, N., Chen, T., & Jiang, R. (2017). Predicting enhancers with deep convolutional neural networks. *BMC Bioinformatics*, *18*, 478.

Minja, D. (2025). *The influence of anthropogenic activities on cheetah space use and hunting success across a landscape of coexistence*.

Minnoye, L., Taskiran, I. I., Mauduit, D., Fazio, M., Van Aerschot, L., Hulselmans, G., Christiaens, V., Makhzami, S., Seltenhammer, M., Karras, P., Primot, A., Cadieu, E., van Rooijen, E., Marine, J.-C., Egidy, G., Ghanem, G.-E., Zon, L., Wouters, J., & Aerts, S. (2020). Cross-species analysis of enhancer logic using deep learning. *Genome Res.*, *30*, 1815–1834.

Moresco, A., Muñoz, K. E., Gutiérrez, F., Arias-Bernal, L., Yarto-Jaramillo, E., Teixeira, R. H. F., Peña-Stadlin, J., & Troan, B. V. (2020). Taxonomic distribution of neoplasia among non-domestic felid species under managed care. *Animals (Basel)*, *10*, 2376.

Morrison, W. R., III, Lohr, J. L., Duchen, P., Wilches, R., Trujillo, D., Mair, M., & Renner, S. S. (2009). The impact of taxonomic change on conservation:

Does it kill, can it save, or is it just irrelevant? *Biol. Conserv.*, *142*, 3201–3206.

Mort, M., Ivanov, D., Cooper, D. N., & Chuzhanova, N. A. (2008). A meta-analysis of nonsense mutations causing human genetic disease. *Hum. Mutat.*, *29*, 1037–1047.

Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., ... Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*, 520–562.

Movahedi, F., Padman, R., & Antaki, J. F. (2023). Limitations of receiver operating characteristic curve on imbalanced data: Assist device mortality risk scores. *J. Thorac. Cardiovasc. Surg.*, *165*, 1433–1442.e2.

Mukai, T., Chigusa, S. I., Mettler, L. E., & Crow, J. F. (1972). Mutation rate and dominance of genes affecting viability in drosophila melanogaster. *Genetics*, *72*, 335–355.

Muller, H. J. (1964). THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE. *Mutat. Res.*, *106*, 2–9.

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, *31*, 3812–3814.

Nicholson, P., & Mühlemann, O. (2010). Cutting the nonsense: The degradation of PTC-containing mRNAs. *Biochem. Soc. Trans.*, *38*, 1615–1620.

Niehaus, J., Imlau, M., Keller, S., Marti, I., Breitenmoser, C., Letko, A., Jagannathan, V., Grossen, C., Leeb, T., & Pewsner, M. (2025). Evolutionary dynamics of a lethal recessive allele in reintroduced fragmented lynx populations. *bioRxiv*, 2025.11.06.686959.

Novakovsky, G., Fornes, O., Saraswat, M., Mostafavi, S., & Wasserman, W. W. (2023). ExplaiNN: Interpretable and transparent neural networks for genomics. *Genome Biol.*, *24*, 154.

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

O'Brien, S. J. (1994a). Genetic and phylogenetic analyses of endangered species. *Annual Review of Genetics*, *28*, 467–489.

O'Brien, S. J. (1994b). A role for molecular genetics in biological conservation. *Proc. Natl. Acad. Sci. U. S. A.*, *91*, 5748–5755.

O'Brien, S. J., Joslin, P., Smith, G. L., Wolfe, R., Schaffer, N., Heath, E., Ott-Joslin, J., Rawal, P. P., Bhattacharjee, K. K., & Martenson, J. S. (1987). Evidence for african origins of founders of the asiatic lion species survival plan. *Zoo Biol.*, *6*, 99–116.

O'Brien, S. J., Roelke, M. E., Marker, L., Newman, A., Winkler, C. A., Meltzer, D., Colly, L., Evermann, J. F., Bush, M., & Wildt, D. E. (1985). Genetic basis for species vulnerability in the cheetah. *Science*, *227*, 1428–1434.

O'Brien, S. J., Wildt, D. E., Goldman, D., Merril, C. R., & Bush, M. (1983). The cheetah is depauperate in genetic variation. *Science*, *221*, 459–462.

O'Brien, S. J., & Johnson, W. E. (2005). Big cat genomics. *Annu. Rev. Genomics Hum. Genet.*, *6*, 407–429.

O'Brien, S. J., & Johnson, W. E. (2007). The evolution cats. *Sci. Am.*, *297*, 68–75.

Ochman, H., & Davalos, L. M. (2006). The nature and dynamics of bacterial genomes. *Science*, *311*, 1730–1733.

Ohta, T. (1992). The nearly neutral theory of molecular evolution [doi: 10.1146/annurev.es.23.110192.001403]. *Annu. Rev. Ecol. Syst.*, *23*, 263–286.

Okutman, O., Muller, J., Skory, V., Garnier, J. M., Gaucherot, A., Baert, Y., Lamour, V., Serdarogullari, M., Gultomruk, M., Röpke, A., Kliesch, S., Herbepin, V., Aknin, I., Benkhalifa, M., Teletin, M., Bakircioglu, E., Goossens, E., Charlet-Berguerand, N., Bahceci, M., ... Viville, S. (2017). A no-stop mutation in MAGEB4 is a possible cause of rare X-linked azoospermia

and oligozoospermia in a consanguineous turkish family. *J. Assist. Reprod. Genet.*, *34*, 683–694.

Olofsson, P., Chipkin, L., Daileda, R. C., & Azevedo, R. B. R. (2023). Mutational meltdown in asexual populations doomed to extinction. *J. Math. Biol.*, *87*, 88.

Olsen, K. C., Ryan, W. H., Kosman, E. T., Moscoso, J. A., Levitan, D. R., & Winn, A. A. (2021). Lessons from the study of plant mating systems for exploring the causes and consequences of inbreeding in marine invertebrates. *Mar. Biol.*, *168*.

Orchard, P., Kyono, Y., Hensley, J., Kitzman, J. O., & Parker, S. C. J. (2020). Quantification, dynamic visualization, and validation of bias in ATAC-seq data with ataqv. *Cell Syst.*, *10*, 298–306.e4.

O'Regan, H. J., & Steininger, C. (2017). Felidae from cooper's cave, south africa (mammalia: Carnivora). *Geodiversitas*, *39*, 315–332.

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., & Warinner, C. (2021). Ancient DNA analysis. *Nat. Rev. Methods Primers*, *1*, 14.

Ortiz, E. M. (2019). *vcf2phylip v2.0: Convert a VCF matrix into several matrix formats for phylogenetic analysis* (comp. software). Zenodo.

Palmer, J. (2016). *Funannotate: Eukaryotic genome annotation pipeline* (comp. software).

Patel, H., Espinosa-Carrasco, J., Langer, B., Ewels, P., Bot, N.-C., Garcia, M. U., Syme, R., Peltzer, A., Talbot, A., Behrens, D., Gabernet, G., Jin, M., Hörtenhuber, M., Rodriguez, J. G., Menden, K., & An, Ö. (2023). *Nf-core/atacseq: [2.1.2] - 2022-08-07* (comp. software). Zenodo.

Patil, A. A., Cai, Y., Sang, Y., Blecha, F., & Zhang, G. (2005). Cross-species analysis of the mammalian β-defensin gene family: Presence of syntenic gene clusters and preferential expression in the male reproductive tract [doi: 10.1152/physiolgenomics.00104.2005]. *Physiol. Genomics*, *23*, 5–17.

Peel, E., Silver, L., Brandies, P., Zhu, Y., Cheng, Y., Hogg, C. J., & Belov, K. (2022). Best genome sequencing strategies for annotation of complex immune gene families in wildlife. *Gigascience*, *11*, giac100.

Peers, J. A., Nash, W. J., & Haerty, W. (2025). Gene pseudogenization in fertility-associated genes in cheetah (acinonyx jubatus), a species with long-term low effective population size. *Evolution*, *79*, 574–585.

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: Five essential questions. *Nat. Rev. Genet.*, *14*, 288–295.

Picard. (n.d.).

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—a review for ecologists. *Methods Ecol. Evol.*, *14*, 994–1016.

Pimm, S. L., Gittleman, J. L., McCracken, G. F., & Gilpin, M. (1989). Plausible alternatives to bottlenecks to explain reduced genetic diversity. *Trends Ecol. Evol.*, *4*, 176–178.

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., Raven, P. H., Roberts, C. M., & Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, *344*, 1246752.

Pinoli, P., Stamoulakatou, E., Ceri, S., & Piro, R. (2019). Deleterious impact of mutational processes on transcription factor binding sites in human cancer. *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, 188–193.

Pinto, A. V., Hansson, B., Patramanis, I., Morales, H. E., & van Oosterhout, C. (2024). The impact of habitat loss and population fragmentation on genomic erosion. *Conserv. Genet.*, *25*, 49–57.

Piras, P., Silvestro, D., Carotenuto, F., Castiglione, S., Kotsakis, A., Maiorino, L., Melchionna, M., Mondanaro, A., Sansalone, G., Serio, C., Vero, V. A., & Raia, P. (2018). Evolution of the sabertooth mandible: A deadly ecomorphological specialization [Used Figure 1 to write Newick tree]. *Palaeogeogr. Palaeoclimatol. Palaeoecol.*, *496*, 166–174.

Plasil, M., Winter, S., Stejskalova, K., Vychodilova, L., Jelinek, A., Futas, J., Burger A, P., & Horin, P. (2025). A chromosome-level genome assembly of the snow leopard, panthera uncia. *J. Hered.*, esaf046.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, *20*, 110–121.

Prost, S., Machado, A. P., Zumbroich, J., Preier, L., Mahtani-Williams, S., Meissner, R., Guschanski, K., Brealey, J. C., Fernandes, C. R., Vercammen, P., Hunter, L. T. B., Abramov, A. V., Plasil, M., Horin, P., Godsall-Bottriell, L., Bottriell, P., Dalton, D. L., Kotze, A., & Burger, P. A. (2022). Genomic analyses show extremely perilous conservation status of african and asiatic cheetahs (acinonyx jubatus). *Mol. Ecol.*, *31*, 4208–4223.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, *81*, 559–575.

Qin, Q., Xu, Y., He, T., Qin, C., & Xu, J. (2012). Normal and disease-related biological functions of Twist1 and underlying molecular mechanisms. *Cell Res.*, *22*, 90–106.

Quang, D., & Xie, X. (2016). DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, *44*, e107.

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

Quinodoz, M., Peter, V. G., Cisarova, K., Royer-Bertrand, B., Stenson, P. D., Cooper, D. N., Unger, S., Superti-Furga, A., & Rivolta, C. (2022). Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity. *Am. J. Hum. Genet.*, *0*.

Radke, D. W., & Lee, C. (2015). Adaptive potential of genomic structural variation in human and mammalian evolution. *Brief. Funct. Genomics*, *14*, 358–368.

Raes, J., & Van de Peer, Y. (2005). Functional divergence of proteins through frameshift mutations. *Trends Genet.*, *21*, 428–431.

Rands, C. M., Meader, S., Ponting, C. P., & Lunter, G. (2014). 8.2% of the human genome is constrained: Variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet.*, *10*, e1004525.

Rhoads, A., & Au, K. F. (2015a). PacBio sequencing and its applications genomics proteomics bioinformatics. *PacBio Sequencing and Its Applications Genomics Proteomics Bioinformatics*, *13*, 278–289.

Rhoads, A., & Au, K. F. (2015b). PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*, *13*, 278–289.

Ricaño-Ponce, I., & Wijmenga, C. (2013). Mapping of immune-mediated disease genes. *Annu. Rev. Genomics Hum. Genet.*, *14*, 325–353.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The european molecular biology open software suite. *Trends Genet.*, *16*, 276–277.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat. Biotechnol.*, *29*, 24–26.

Rodriguez, O. L., Safonova, Y., Silver, C. A., Shields, K., Gibson, W. S., Kos, J. T., Tieri, D., Ke, H., Jackson, K. J. L., Boyd, S. D., Smith, M. L., Marasco, W. A., & Watson, C. T. (2023). Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nat. Commun.*, *14*, 4419.

Roelke, M. E., Martenson, J. S., & O'Brien, S. J. (1993). The consequences of demographic reduction and genetic depletion in the endangered florida panther. *Curr. Biol.*, *3*, 340–350.

Ruiz Daniels, R., Taylor, R. S., Dobie, R., Salisbury, S., Furniss, J. J., Clark, E., Macqueen, D. J., & Robledo, D. (2023). A versatile nuclei extraction protocol for single nucleus sequencing in non-model species-optimization in various atlantic salmon tissues. *PLoS One*, *18*, e0285020.

Salzberg, S. L. (2019). Next-generation genome annotation: We still struggle to get it right. *Genome Biol.*, *20*, 92.

Samaha, G., Wade, C. M., Mazrier, H., Grueber, C. E., & Haase, B. (2021). Exploiting genomic synteny in felidae: Cross-species genome alignments and SNV discovery can aid conservation management. *BMC Genomics*, *22*, 601.

Samaha, G. A. (2021). *Advances in felid genetics and genomics*.

Sandler, R. L., Moses, L., & Wisely, S. M. (2021). An ethical analysis of cloning for genetic rescue: Case study of the black-footed ferret. *Biol. Conserv.*, *257*, 109118.

Santiago, E., Köpke, C., & Caballero, A. (2025). Accounting for population structure and data quality in demographic inference with linkage disequilibrium methods. *Nat. Commun.*, *16*, 6054.

Santiago, E., Novo, I., Pardiñas, A. F., Saura, M., Wang, J., & Caballero, A. (2020). Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. *Mol. Biol. Evol.*, *37*, 3642–3653.

Santymire, R. M., Livieri, T. M., Branvold-Faber, H., & Marinari, P. E. (2014). The black-footed ferret: On the brink of recovery? In W. V. Holt, J. L. Brown, & P. Comizzoli (Eds.), *Reproductive sciences in animal conservation: Progress and prospects* (pp. 119–134). Springer New York.

Santymire, R. M., Marinari, P. E., Kreeger, J. S., Wildt, D. E., & Howard, J. (2006). Sperm viability in the black-footed ferret (mustela nigripes) is influenced by seminal and medium osmolality. *Cryobiology*, *53*, 37–50.

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, *489*, 109–113.

Savas, S., Tuzmen, S., & Ozcelik, H. (2006). Human SNPs resulting in premature stop codons and protein truncation. *Hum. Genomics*, *2*, 274–286.

Scacheri, C. A., & Scacheri, P. C. (2015). Mutations in the noncoding genome. *Curr. Opin. Pediatr.*, *27*, 659–664.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*, 467–470.

Schiffels, S., & Wang, K. (2020). MSMC and MSMC2: The multiple sequentially markovian coalescent. In J. Y. Dutheil (Ed.), *Statistical population genomics* (pp. 147–166). Springer US.

Schilff, M., Sargsyan, Y., Hofhuis, J., & Thoms, S. (2021). Stop codon context-specific induction of translational readthrough. *Biomolecules*, *11*, 1006.

Schneider, J. J., Unholzer, A., Schaller, M., Schäfer-Korting, M., & Korting, H. C. (2005). Human defensins. *J. Mol. Med.*, *83*, 587–595.

Schoenfelder, S., Javierre, B.-M., Furlan-Magaril, M., Wingett, S. W., & Fraser, P. (2018). Promoter capture hi-C: High-resolution, genome-wide profiling of promoter interactions. *J. Vis. Exp.*

Schröder, J. M., & Harder, J. (1999). Human beta-defensin-2. *Int. J. Biochem. Cell Biol.*, *31*, 645–651.

Schubach, M., Maass, T., Nazaretyan, L., Röner, S., & Kircher, M. (2024). CADD v1.7: Using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res.*, *52*, D1143–D1154.

Schubach, M., Re, M., Robinson, P. N., & Valentini, G. (2017). Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. *Sci. Rep.*, *7*, 2959.

Schürmann, A., Koling, S., Jacobs, S., Saftig, P., Krauss, S., Wennemuth, G., Kluge, R., & Joost, H.-G. (2002). Reduced sperm count and normal fertility in male mice with targeted disruption of the ADP-ribosylation factor-like 4 (Arl4) gene. *Mol. Cell. Biol.*, *22*, 2761–2768.

Schwensow, N., Castro-Prieto, A., Wachter, B., & Sommer, S. (2019). Immunological MHC supertypes and allelic expression: How low is the functional MHC diversity in free-ranging namibian cheetahs? *Conserv. Genet.*, *20*, 65–80.

Selkoe, K. A., & Toonen, R. J. (2006). Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. *Ecol. Lett.*, *9*, 615–629.

Shafer, A. B. A., & Kardos, M. (2025). Runs of homozygosity and inferences in wild populations. *Mol. Ecol.*, *34*, e17641.

Shaw, R., MacPherson, J., Kitchener, A. C., Etherington, G. J., & Haerty, W. (2025). Characterisation of the historic demographic decline of the british european polecat population. *Mol. Ecol.*, *34*, e70091.

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One*, *11*, e0163962.

Sheridan, L., & Pomiankowski, A. (1997). Fluctuating asymmetry, spot asymmetry and inbreeding depression in the sexual coloration of male guppy fish. *Heredity (Edinb.)*, *79*, 515–523.

Shumate, A., & Salzberg, S. L. (2021). Liftoff: Accurate mapping of gene annotations. *Bioinformatics*, *37*, 1639–1643.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, *15*, 1034–1050.

Simon, R., Lischer, H. E. L., Pieńkowska-Schelling, A., Keller, I., Häfliger, I. M., Letko, A., Schelling, C., Lühken, G., & Drögemüller, C. (2020). New genomic features of the polled intersex syndrome variant in goats unraveled by long-read whole-genome sequencing. *Anim. Genet.*, *51*, 439–448.

Simon, S. A., Zhai, J., Nandety, R. S., McCormick, K. P., Zeng, J., Mejia, D., & Meyers, B. C. (2009). Short-read sequencing technologies for transcriptional analyses. *Annu. Rev. Plant Biol.*, *60*, 305–333.

Sironen, A., Shoemark, A., Patel, M., Loebinger, M. R., & Mitchison, H. M. (2020). Sperm defects in primary ciliary dyskinesia and related causes of male infertility. *Cell. Mol. Life Sci.*, *77*, 2029–2048.

Smeds, L., & Ellegren, H. (2023). From high masked to high realized genetic load in inbred scandinavian wolves. *Mol. Ecol.*, *32*, 1567–1580.

Smit, A. F. A., Hubley, R., & Green, P. (2013). *RepeatMasker open-4.0*.

Smith, C. L., & Eppig, J. T. (2009). The mammalian phenotype ontology: Enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, *1*, 390–399.

Smith, H. A., & McNeel, D. G. (2010). The SSX family of cancer-testis antigens as target proteins for tumor therapy. *Clin. Dev. Immunol.*, *2010*, 1–18.

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.*, *23*, 23–35.

Smithies, O. (1993). Animal models of human genetic diseases. *Trends Genet.*, *9*, 112–116.

Song, L., & Crawford, G. E. (2010). DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, *2010*, db.prot5384.

Song, P., & Perkins, B. D. (2018). Developmental expression of the zebrafish arf-like small GTPase paralogs arl13a and arl13b. *Gene Expr. Patterns*, *29*, 82–87.

Sonstegard, T. S., Cole, J. B., VanRaden, P. M., Van Tassell, C. P., Null, D. J., Schroeder, S. G., Bickhart, D., & McClure, M. C. (2013). Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in jersey cattle. *PLoS One*, *8*, e54872.

Speak, S. A., Birley, T., Bortoluzzi, C., Clark, M. D., Percival-Alwyn, L., Morales, H. E., & van Oosterhout, C. (2024). Genomics-informed captive breeding can reduce inbreeding depression and the genetic load in zoo populations. *Mol. Ecol. Resour.*, *24*, e13967.

Spitzer, R., Norman, A. J., Königsson, H., Schiffthaler, B., & Spong, G. (2020). De novo discovery of SNPs for genotyping endangered sun parakeets (aratinga solstitialis) in guyana. *Conserv. Genet. Resour.*, *12*, 631–641.

Stein, L. (2001). Genome annotation: From sequence to biology. *Nat. Rev. Genet.*, *2*, 493–503.

Stoffel, M. A., Johnston, S. E., Pilkington, J. G., & Pemberton, J. M. (2021). Genetic architecture and lifetime dynamics of inbreeding depression in a wild mammal. *Nat. Commun.*, *12*, 2972.

Sullivan, P. F., Meadows, J. R. S., Gazal, S., Phan, B. N., Li, X., Genereux, D. P., Dong, M. X., Bianchi, M., Andrews, G., Sakthikumar, S., Nordin, J., Roy, A., Christmas, M. J., Marinescu, V. D., Wang, C., Wallerman, O., Xue, J., Yao, S., Sun, Q., ... Lindblad-Toh, K. (2023). Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science*, *380*, eabn2937.

Symmons, O., & Spitz, F. (2013). From remote enhancers to gene regulation: Charting the genome's regulatory landscapes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *368*, 20120358.

Taktehrani, A., Shah Hosseini, M., Gholikhani, N., Hobeali, K., Karimi, M. H., Samadzadeh, N., Abolghasemi, H., Ranjbaran, A., Radman, A., Safarzadeh, A., Pourmirzai, M., & Farhadinia, M. S. (2025). Will they survive? alarming circumstances of asiatic cheetah (*Acinonyx jubatus venaticus*) in iran's drylands. *bioRxiv*.

Tang, S., Wang, X., Li, W., Yang, X., Li, Z., Liu, W., Li, C., Zhu, Z., Wang, L., Wang, J., Zhang, L., Sun, X., Zhi, E., Wang, H., Li, H., Jin, L., Luo, Y., Wang, J., Yang, S., & Zhang, F. (2017). Biallelic mutations in CFAP43 and CFAP44 cause male infertility with multiple morphological abnormalities of the sperm flagella. *Am. J. Hum. Genet.*, *100*, 854–864.

Templeton, A. R., & Read, B. (1983). The elimination of inbreeding depression in a captive herd of speke's gazelle. In C. M. Schonewald-Cox, S. M. Chambers, B. MacBryde, & W. L. Thomas (Eds.), *Genetics and conservation: A reference for managing wild animal and plant populations* (pp. 241–261).

Tena, J. J., & Santos-Pereira, J. M. (2021). Topologically associating domains and regulatory landscapes in development, evolution and disease. *Front. Cell Dev. Biol.*, *9*, 702787.

Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.*, *49*, 303–309.

Terio, K. A., Mitchell, E., Walzer, C., Schmidt-Küntzel, A., Marker, L., & Citino, S. (2018). Diseases impacting captive and free-ranging cheetahs. *Cheetahs: Biology and Conservation*, 349.

Terrell, K. A., Crosier, A. E., Wildt, D. E., O'Brien, S. J., Anthony, N. M., Marker, L., & Johnson, W. E. (2016). Continued decline in genetic diversity among wild cheetahs (acinonyx jubatus) without further loss of semen quality. *Biol. Conserv.*, *200*, 192–199.

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, *165*, 1519–1529.

Thanki, A. S., Soranzo, N., Haerty, W., & Davey, R. P. (2018). GeneSeqToFamily: A galaxy workflow to find gene families based on the ensembl compara GeneTrees pipeline [Anil's workflow. Used to look for gene families, ancestral gene duplication events and genes diverged from a common ancestor under positive selection]. *Gigascience*, *7*, 1–10.

Tian, X., Zhang, J., Yan, L., Dong, J.-M., & Guo, Q. (2015). MiRNA-15a inhibits proliferation, migration and invasion by targeting TNFAIP1 in human osteosarcoma cells. *Int. J. Clin. Exp. Pathol.*, *8*, 6442–6449.

Tilley, A. E., Walters, M. S., Shaykhiev, R., & Crystal, R. G. (2015). Cilia dysfunction in lung disease. *Annu. Rev. Physiol.*, *77*, 379–406.

Tordiffe, A. S. W., Jhala, Y. V., Boitani, L., Cristescu, B., Kock, R. A., Meyer, L. R. C., Naylor, S., O'Brien, S. J., Schmidt-Küntzel, A., Stanley Price, M. R., van der Merwe, V., & Marker, L. (2023). The case for the reintroduction of cheetahs to india. *Nat. Ecol. Evol.*, *7*, 480–481.

Torgerson, D. G., Kulathinal, R. J., & Singh, R. S. (2002). Mammalian sperm proteins are rapidly evolving: Evidence of positive selection in functionally diverse genes. *Mol. Biol. Evol.*, *19*, 1973–1980.

Tsheten, G., Fuerst-Waltl, B., Pfeiffer, C., Sölkner, J., Bovenhuis, H., & Mészáros, G. (2023). Inbreeding depression and its effect on sperm quality traits in pietrain pigs. *J. Anim. Breed. Genet.*, *140*, 653–662.

Urfer, S. R. (2009). Inbreeding and fertility in irish wolfhounds in sweden: 1976 to 2007. *Acta Vet. Scand.*, *51*, 21.

Urtecho, G., Insigne, K. D., Tripp, A. D., Brinck, M. S., Lubock, N. B., Acree, C., Kim, H., Chan, T., & Kosuri, S. (2023). Genome-wide functional characterization of escherichia coli promoters and sequence elements encoding their regulation. *eLife*.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The third revolution in sequencing technology. *Trends Genet.*, *34*, 666–681.

van Oosterhout, C. (2020). Mutation load is the spectre of species conservation. *Nat Ecol Evol*, *4*, 1004–1006.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, *43*, 11.10.1–11.10.33.

Van Valkenburgh, B., Pang, B., Cherin, M., & Rook, L. (2018). The cheetah: Evolutionary history and paleoecology. In *Cheetahs: Biology and conservation* (pp. 25–32). Elsevier.

Vasudeva, R., Sales, K., Gage, M. J. G., & Hosken, D. J. (2025). Inbreeding depression in male reproductive traits. *J. Evol. Biol.*, *38*, 504–515.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., ... Zhu, X. (2001). The sequence of the human genome. *Science*, *291*, 1304–1351.

Venugopal, V. (2025, September 24). *India to receive new batch of cheetahs as survival rates outpace global average*.

Vermunt, M. W., Zhang, D., & Blobel, G. A. (2019). The interdependence of gene-regulatory elements and the 3D genome. *J. Cell Biol.*, *218*, 12–26.

Viluma, A., Flagstad, Ø., Åkesson, M., Wikenros, C., Sand, H., Wabakken, P., & Ellegren, H. (2022). Whole-genome resequencing of temporally stratified

samples reveals substantial loss of haplotype diversity in the highly inbred scandinavian wolf population. *Genome Res.*

Vincze, O., Colchero, F., Lemaître, J.-F., Conde, D. A., Pavard, S., Bieuville, M., Urrutia, A. O., Ujvari, B., Boddy, A. M., Maley, C. C., Thomas, F., & Giraudeau, M. (2022). Cancer risk across mammals. *Nature*, *601*, 263–267.

Vostrov, A. A., & Quitschke, W. W. (1997). The zinc finger protein CTCF binds to the APBβ domain of the amyloid β-protein precursor promoter. *J. Biol. Chem.*, *272*, 33353–33359.

Vy, H. M. T., Jordan, D. M., Balick, D. J., & Do, R. (2021). Probing the aggregated effects of purifying selection per individual on 1,380 medical phenotypes in the UK biobank. *PLoS Genet.*, *17*, e1009337.

Wang, C., Gibbons, J., Adapa, S. R., Oberstaller, J., Liao, X., Zhang, M., Adams, J. H., & Jiang, R. H. Y. (2020). The human malaria parasite genome is configured into thousands of coexpressed linear regulatory units. *J. Genet. Genomics*, *47*, 513–521.

Wang, N., Lysenkov, V., Orte, K., Kairisto, V., Aakko, J., Khan, S., & Elo, L. L. (2022). Tool evaluation for the detection of variably sized indels from next generation whole genome and targeted sequencing data. *PLoS Comput. Biol.*, *18*, e1009269.

Wang, S., Chen, Z., Luo, A., You, X., Kitchener, A. C., Tu, X., Thakur, M., Umapathy, G., Hu, S., Zhang, T., Zhang, Y., Liu, S., Ding, Y., Liu, F., Dai, Q., Feng, X., Li, L., Pan, Y., Zhang, M., ... Wu, D.-D. (2025). Genome sequences of extant and extinct gibbons reveal their phylogeny, demographic history, and conservation status. *Cell*, *0*.

Wang, Z., Yuan, W., & Montana, G. (2015). Sparse multi-view matrix factorisation: A multivariate approach to multiple tissue comparisons. *arXiv [stat.ML]*.

Waskom, M. (2021). Seaborn: Statistical data visualization. *J. Open Source Softw.*, *6*, 3021.

Wayne, R. K., Modi, W. S., & O'Brien, S. J. (1986). Morphological variability and asymmetry in the cheetah (acinonyx jubatus), a genetically uniform species. *Evolution*, *40*, 78–85.

Weber, J. L., & May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.*, *44*, 388–396.

Wehrenberg, G., Tokarska, M., Cocchiararo, B., & Nowak, C. (2024). A reduced SNP panel optimised for non-invasive genetic assessment of a genetically impoverished conservation icon, the european bison. *Sci. Rep.*, *14*, 1875.

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nat. Rev. Genet.*, *14*, 125–138.

Weise, F. J., Vijay, V., Jacobson, A. P., Schoonover, R. F., Groom, R. J., Horgan, J., Keeping, D., Klein, R., Marnewick, K., Maude, G., Melzheimer, J., Mills, G., van der Merwe, V., van der Meer, E., van Vuuren, R. J., Wachter, B., & Pimm, S. L. (2017). The distribution and numbers of cheetah (acinonyx jubatus) in southern africa. *PeerJ*, *5*, e4096.

Weisman, C. M., Murray, A. W., & Eddy, S. R. (2022). Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Curr. Biol.*

Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Mol. Ecol.*, *28*, 1203–1209.

Wellenreuther, M., Oomen, R. A., Han, K. Y., Krohman, R., & Reusch, T. B. H. (2025). Beyond supergenes: The diverse roles of inversions in trait evolution. *Trends Ecol. Evol.*

Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., Telenti, A., & di Iulio, J. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat. Commun.*, *10*, 5241.

Wen, L., Zhao, C., Song, J., Ma, L., Ruan, J., Xia, X., Chen, Y. E., Zhang, J., Ma, P. X., & Xu, J. (2020). CRISPR/Cas9-mediated TERT disruption in cancer cells. *Int. J. Mol. Sci.*, *21*.

Werdelin, L., & Lewis, M. E. (2005). Plio-pleistocene carnivora of eastern africa: Species richness and turnover patterns. *Zoological Journal of the Linnean Society*, *144*, 121–144.

Werdelin, L., Yamaguchi, N., Johnson, W. E., & O'Brien, S. J. (2010). Phylogeny and evolution of cats (felidae). *Biology and conservation of wild felids*, 59–82.

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., Myers, R. M., & Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.*, *13*, R50.

Wiegleb, G., Reinhardt, S., Dahl, A., & Posnien, N. (2022). Tissue dissociation for single-cell and single-nuclei RNA sequencing for low amounts of input material. *Front. Zool.*, *19*, 27.

Wielebnowski, N. (1996). Reassessing the relationship between juvenile mortality and genetic monomorphism in captive cheetahs. *Zoo Biol.*, *15*, 353–369.

Wilder, A. P., Supple, M. A., Subramanian, A., Mudide, A., Swofford, R., Serres-Armero, A., Steiner, C., Koepfli, K.-P., Genereux, D. P., Karlsson, E. K., Lindblad-Toh, K., Marques-Bonet, T., Munoz Fuentes, V., Foley, K., Meyer, W. K., Zoonomia Consortium‡, Ryder, O. A., & Shapiro, B. (2023). The contribution of historical processes to contemporary extinction risk in placental mammals. *Science*, *380*, eabn5856.

Wildt, D. E., Bush, M., Howard, J. G., O'Brien, S. J., Meltzer, D., Van Dyk, A., Ebedes, H., & Brand, D. J. (1983). Unique seminal quality in the south african cheetah and a comparative evaluation in the domestic cat. *Biol. Reprod.*, *29*, 1019–1025.

Wildt, D. E., Phillips, L. G., Simmons, L. G., Chakraborty, P. K., Brown, J. L., Howard, J. G., Teare, A., & Bush, M. (1988). A comparative analysis of

ejaculate and hormonal characteristics of the captive male cheetah, tiger, leopard, and puma. *Biol. Reprod.*, *38*, 245–255.

Williams, L. M., Ma, X., Boyko, A. R., Bustamante, C. D., & Oleksiak, M. F. (2010). SNP identification, verification, and utility for population genetics in a non-model genus. *BMC Genet.*, *11*, 32.

Williams, R. S., Chasman, D. I., Hau, D. D., Hui, B., Lau, A. Y., & Glover, J. N. M. (2003). Detection of protein folding defects caused by BRCA1-BRCT truncation and missense mutations. *J. Biol. Chem.*, *278*, 53007–53016.

Williams, S. E., & Hoffman, E. A. (2009). Minimizing genetic adaptation in captive breeding programs: A review. *Biol. Conserv.*, *142*, 2388–2400.

Winter, S., Meißner, R., Greve, C., Ben Hamadou, A., Horin, P., Prost, S., & Burger, P. A. (2023). A chromosome-scale high-contiguity genome assembly of the cheetah (acinonyx jubatus). *J. Hered.*

Witzenberger, K. A., & Hochkirch, A. (2011). Ex situ conservation genetics: A review of molecular studies on the genetic consequences of captive breeding programmes for endangered animal species. *Biodivers. Conserv.*, *20*, 1843–1861.

Woc Colburn, A. M., Sanchez, C. R., Citino, S., Crosier, A. E., Murray, S., Kaandorp, J., Kaandorp, C., & Marker, L. (2018). Clinical management of captive cheetahs. In *Cheetahs: Biology and conservation* (pp. 335–347). Elsevier.

Woodworth, L. M., Montgomery, M. E., Briscoe, D. A., & Frankham, R. (2002). Rapid genetic deterioration in captive populations: Causes and conservation implications. *Conserv. Genet.*, *3*, 277–288.

Worsley Hunt, R., Mathelier, A., Del Peso, L., & Wasserman, W. W. (2014). Improving analysis of transcription factor binding sites within ChIP-seq data based on topological motif enrichment. *BMC Genomics*, *15*, 472.

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., & Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, *20*, 1377–1419.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, *19*, 395–420.

Wu, J., Yonezawa, T., & Kishino, H. (2017). Rates of molecular evolution suggest natural history of life history traits and a post-K-pg nocturnal bottleneck of placentals. *Curr. Biol.*, *27*, 3025–3033.e5.

Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., Frandsen, P., Chen, Y., Yngvadottir, B., Cooper, D. N., de Manuel, M., Hernandez-Rodriguez, J., Lobon, I., Siegismund, H. R., Pagani, L., Quail, M. A., Hvilsom, C., Mudakikwa, A., Eichler, E. E., ... Scally, A. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, *348*, 242–245.

Yamaguchi, K., Kadota, M., Nishimura, O., Ohishi, Y., Naito, Y., & Kuraku, S. (2021). Technical considerations in hi-C scaffolding and evaluation of chromosome-scale genome assemblies. *Mol. Ecol.*, *30*, 5923–5934.

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights Imaging*, *9*, 611–629.

Yang, Z., & Yoder, A. D. (2003). Comparison of likelihood and bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Syst. Biol.*, *52*, 705–716.

Yu, J., Yu, X., Bi, W., Li, Z., Zhou, Y., Ma, R., Feng, F., Huang, C., Gu, J., Wu, W., Lan, G., Zhang, L., Chen, C., Xue, F., & Liu, J. (2025). Mitogenome diversity and phylogeny of felidae species. *Diversity (Basel)*, *17*, 634.

Yuan, J., Kitchener, A. C., Lackey, L. B., Sun, T., Jiangzuo, Q., Tuohetahong, Y., Zhao, L., Yang, P., Wang, G., Huang, C., Wang, J., Hou, W., Liu, Y., Chen, W., Mi, D., Murphy, W. J., & Li, G. (2024). The genome of the black-footed cat: Revealing a rich natural history and urgent conservation priorities for small felids. *Proc. Natl. Acad. Sci. U. S. A.*, *121*, e2310763120.

Zhang, C., Zhou, Y., Xie, S., Yin, Q., Tang, C., Ni, Z., Fei, J., & Zhang, Y. (2018). CRISPR/Cas9-mediated genome editing reveals the synergistic effects of β-

defensin family members on sperm maturation in rat epididymis. *FASEB J.*, *32*, 1354–1363.

Zhang, H., Gelernter, J., Gruen, J. R., Kranzler, H. R., Herman, A. I., & Simen, A. A. (2010). Functional impact of a single-nucleotide polymorphism in the OPRD1 promoter region. *J. Hum. Genet.*, *55*, 278–284.

Zhang, M. Q. (1998). Identification of human gene core promoters in silico. *Genome Res.*, *8*, 319–326.

Zhang, M. Q. (2007). Computational analyses of eukaryotic promoters. *BMC Bioinformatics*, *8 Suppl 6*, S3.

Zhang, W. Q., & Zhang, M. H. (2013). Complete mitochondrial genomes reveal phylogeny relationship and evolutionary history of the family felidae. *Genet. Mol. Res.*, *12*, 3256–3262.

Zhang, X., Shen, Y., Wang, X., Yuan, G., Zhang, C., & Yang, Y. (2019). A novel homozygous CFAP65 mutation in humans causes male infertility with multiple morphological abnormalities of the sperm flagella. *Clin. Genet.*, *96*, 541–548.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-seq (MACS). *Genome Biol.*, *9*, R137.

Zhang, Z., Miteva, M. A., Wang, L., & Alexov, E. (2012). Analyzing effects of naturally occurring missense mutations. *Comput. Math. Methods Med.*, *2012*, 805827.

Zhao, X., Collins, R. L., Lee, W.-P., Weber, A. M., Jun, Y., Zhu, Q., Weisburd, B., Huang, Y., Audano, P. A., Wang, H., Walker, M., Lowther, C., Fu, J., Human Genome Structural Variation Consortium, Gerstein, M. B., Devine, S. E., Marschall, T., Korbel, J. O., Eichler, E. E., ... Talkowski, M. E. (2021). Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Hum. Genet.*, *108*, 919–928.

Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M., & Snyder, M. (2010). Genetic analysis of variation in transcription factor binding in yeast. *Nature*, *464*, 1187–1191.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, *12*, 931–934.

Zhu, J., Chen, G., Zhu, S., Li, S., Wen, Z., Bin Li, Zheng, Y., & Shi, L. (2016). Identification of tissue-specific protein-coding and noncoding transcripts across 14 human tissues using RNA-seq. *Sci. Rep.*, *6*, 28400.

Zink, R. M., & Klicka, L. B. (2022). The taxonomic basis of subspecies listed as threatened and endangered under the endangered species act. *Front. Conserv. Sci.*, *3*, 971280.

Zoonomia Consortium. (2020). A comparative genomics multitool for scientific discovery and conservation [Describes zoonomia project and various key findings]. *Nature*, *587*, 240–245.