**MAIN PROGRESS REVIEW - SCIENTIFIC REPORT**

**Please consult online guidelines for further information (intranet.nbi.ac.uk/cms/3131)**

|  |  |
| --- | --- |
| Student Name | Jessica Peers |
| Institute | Earlham Institute |
| Student Number | 100347054 |
| Primary Supervisor | Wilfried Haerty |
| Degree | PhD |
| Year | 1 |

Please insert your completed Scientific Report below and submit as instructed.

Shining a light on the cost of domestication in the dark genome

Probation Review – Scientific Report

Jessica Peers

Student Number: 100347054



**Contents**

[1. Abstract 4](#_Toc101790675)

[2. Theory 4](#_Toc101790676)

[2.1. Bottlenecks 4](#_Toc101790677)

[2.2. Deleterious mutations 5](#_Toc101790678)

[2.3. Small populations 6](#_Toc101790679)

[2.4. Nearly neutral theory of evolution 6](#_Toc101790680)

[2.5. Hitchhiking 7](#_Toc101790681)

[2.6. Inbreeding 8](#_Toc101790682)

[2.7. Genetic purging 10](#_Toc101790683)

[3. Domesticated & wild mammals 10](#_Toc101790684)

[3.1. Wild mammals 10](#_Toc101790685)

[3.2. Domesticated mammals 11](#_Toc101790686)

[4. Genomics methods 12](#_Toc101790687)

[4.1. Reference genomes 12](#_Toc101790688)

[4.2. Genome annotation 13](#_Toc101790689)

[4.3. Comparative genomics 14](#_Toc101790690)

[4.4. Genomic studies of deleterious mutations 14](#_Toc101790691)

[5. Noncoding regions 15](#_Toc101790692)

[5.1. Functional noncoding regions 15](#_Toc101790693)

[5.2. Identification of noncoding regions 16](#_Toc101790694)

[5.3. Machine learning 16](#_Toc101790695)

[6. Project overview 17](#_Toc101790696)

[7. Work to date: Background 18](#_Toc101790697)

[7.1. Felidae 18](#_Toc101790698)

[7.2. Demographic histories 20](#_Toc101790699)

[7.3. Aims and objectives 23](#_Toc101790708)

[8. Progress and future work 23](#_Toc101790709)

[8.1. Deleterious mutations in functional noncoding regions 23](#_Toc101790710)

[8.2. Gene family analysis 24](#_Toc101790711)

[8.3. Research plan 28](#_Toc101790712)

[8.4. Training and conferences 28](#_Toc101790713)

[9. Acknowledgements 31](#_Toc101790714)

[10. References 32](#_Toc101790715)

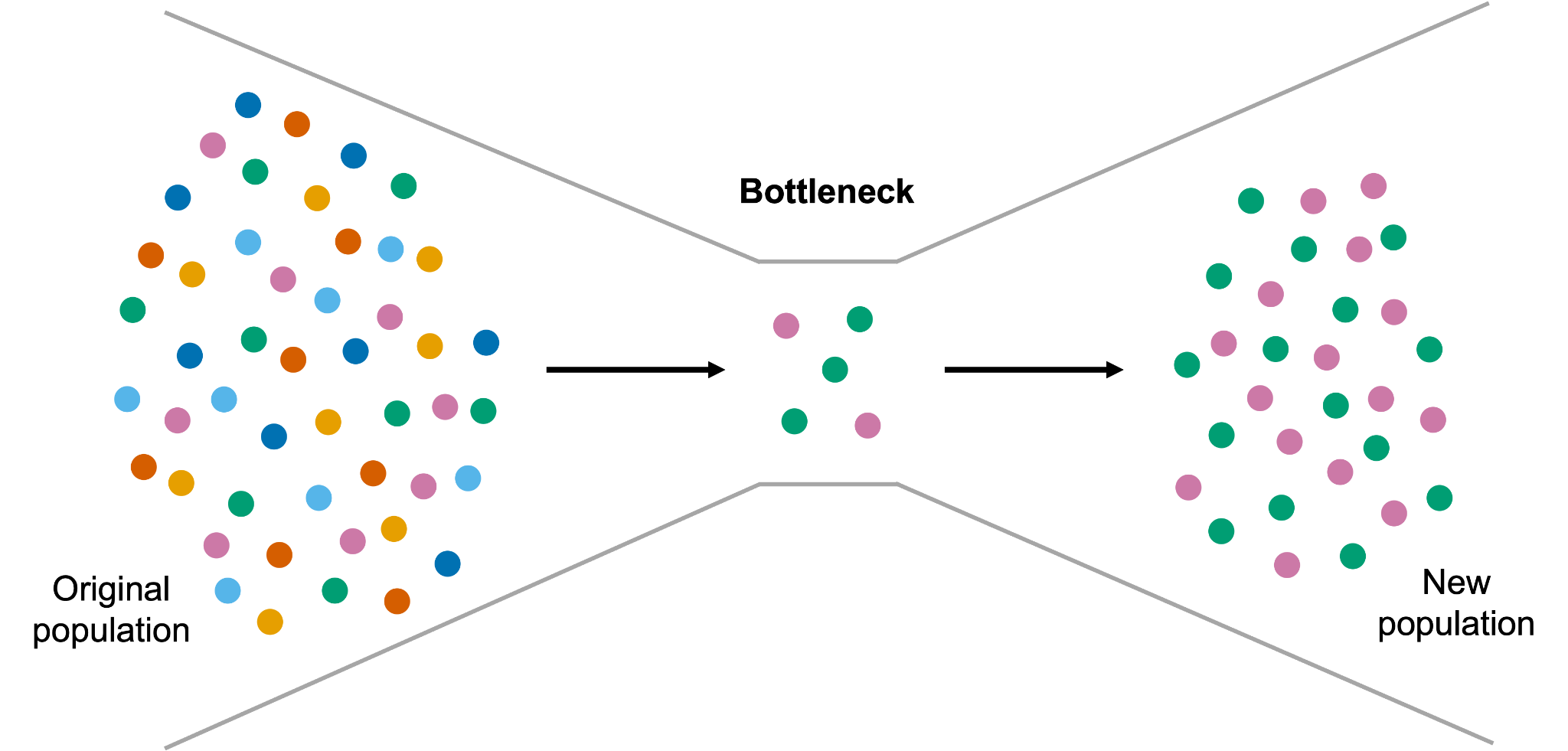
# Abstract

A decline in effective population size can result in a bottleneck, characterised by increased inbreeding and resulting in an accumulation of weakly deleterious mutations. This phenomenon occurs in declining wild mammal populations as well as inbred domestic mammals. The effects of inbreeding on protein coding sequences are well understood, but the majority of disease- and trait-associated variants are enriched in noncoding sequences. Therefore, investigating the accumulation of deleterious mutations in noncoding regions, particularly functional regions such as transcription factor binding sites, is crucial in understanding diseases and traits. The difficulty of detecting functional noncoding sequences has limited previous research, but recent advancements in machine learning methods enable the use of data from model species to predict functional noncoding regions in non-model species. Utilising the wealth of mammalian genome data now available, this project aims to investigate the accumulation of deleterious mutations in functional noncoding regions of genomes of domesticated and endangered wild mammals.

# Theory

## 2.1. Bottlenecks

Demographic events such as founder events or bottlenecks result in a reduction in population size [(Galtier *et al.*, 2000)](https://paperpile.com/c/HQttr3/VJr2). A bottleneck is a rapid reduction in population size which 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 [(Broquet *et al.*, 2010; Bijlsma and Loeschcke, 2012; Banks *et al.*, 2013)](https://paperpile.com/c/HQttr3/tBsW+IbOv+G24Q). 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) [(Luikart *et al.*, 1998)](https://paperpile.com/c/HQttr3/v0Vc). In turn, this makes a population less adaptable to environmental changes as the remaining alleles contain much less diversity and are therefore less likely to contain variation that is potentially beneficial in a changing environment. A reduction in population size also increases the chance of inbreeding as the chance of mating with a related individual is greater.



**Figure 1.** 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.

## 2.2. Deleterious mutations

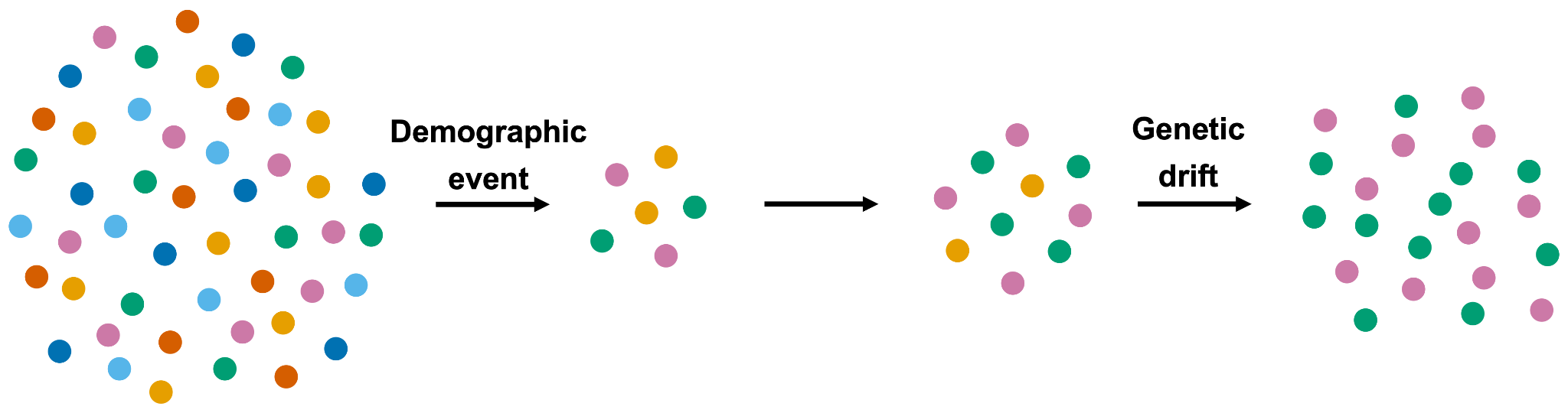
Deleterious mutations are those which have a negative effect on an organism's fitness [(Lynch *et al.*, 1993; Lande, 1994)](https://paperpile.com/c/HQttr3/C3HS+b5I0). They may occur within protein-coding sequences, in the form of frameshift, nonsense or missense mutations or premature termination codons (PTCs), or within noncoding regions, altering the sequence of functional regions [(Raes and Van de Peer, 2005; Savas *et al.*, 2006; Kryukov *et al.*, 2007; Hindorff *et al.*, 2009; Sonstegard *et al.*, 2013)](https://paperpile.com/c/HQttr3/HnZg+rhgV+uz1X+wpPj+SQeB). 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)](https://paperpile.com/c/HQttr3/yakP).

Muller’s ratchet [(Muller, 1964)](https://paperpile.com/c/HQttr3/GsGm) states that with no recombination, irreversible deleterious mutations accumulate. Offspring have at least as many deleterious mutations as their parents, so with more generations, deleterious mutations accumulate [(Muller, 1964)](https://paperpile.com/c/HQttr3/GsGm). Some deleterious mutations may have a strong effect on an organism, causing death or preventing reproduction, meaning the deleterious mutation is not heritable. Sexual reproduction enables the removal of such mutations, however weak mutations may still accumulate [(Muller, 1964)](https://paperpile.com/c/HQttr3/GsGm). Weakly deleterious mutations have a smaller effect on an organism, meaning reproduction may still occur, allowing the mutation to be heritable.

## 2.3. Small populations

Deleterious mutations accumulate more quickly in populations with small effective population sizes (*Ne*) or populations with relaxed selective pressures [(Björnerfeldt *et al.*, 2006)](https://paperpile.com/c/HQttr3/p9kK). 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)](https://paperpile.com/c/HQttr3/b5I0).

Small populations are also prone to genetic drift: the random fixation or loss of alleles due to chance (Figure 2) [(Masel, 2011)](https://paperpile.com/c/HQttr3/UAyi). Genetic erosion, where rare alleles are lost due to death of individuals, has a larger effect on smaller populations as it further decreases genetic diversity [(Bijlsma and Loeschcke, 2012)](https://paperpile.com/c/HQttr3/G24Q). Fixation of alleles in a population due to genetic drift may also be problematic for small populations as weakly deleterious mutations may be fixed by chance [(Lande, 1994)](https://paperpile.com/c/HQttr3/b5I0). 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 *Ne* than in larger populations [(Crow and Kimura, 1970)](https://paperpile.com/c/HQttr3/vK5t). This can lead to a ‘mutational meltdown’: accumulated deleterious mutations cause a reduced population size and subsequent vulnerability to genetic drift causes further fixation of deleterious mutations, eventually resulting in extinction [(Lynch and Gabriel, 1990)](https://paperpile.com/c/HQttr3/LI2Q).

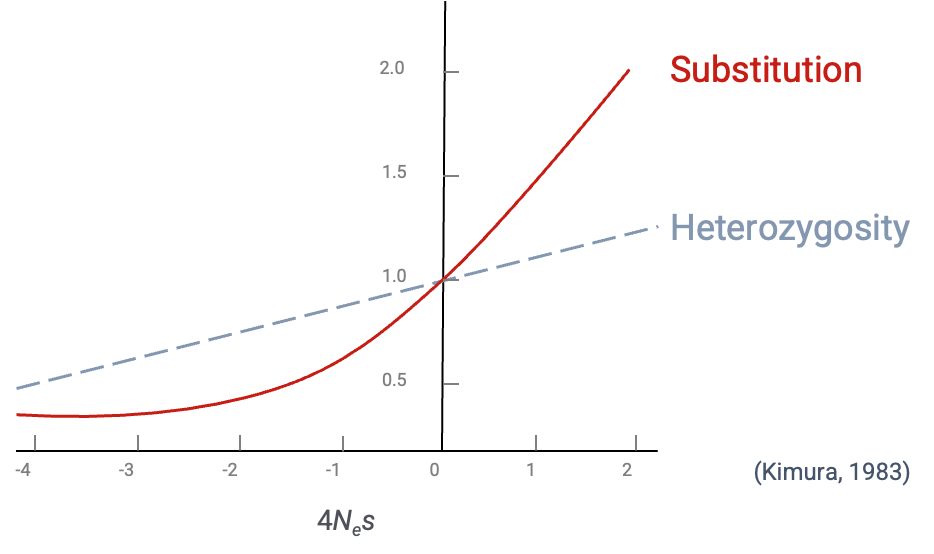


**Figure 2.** Genetic drift. Alleles are shown in different colours. A demographic event (e.g., a bottleneck) results in a small population with reduced diversity. As the population expands, genetic drift can result in the loss (e.g., yellow) or fixation (e.g., pink/green) of alleles due to chance.

## 2.4. 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)](https://paperpile.com/c/HQttr3/9Ze7+sgSi). This theory states that most mutations are selectively neutral, or nearly neutral, and segregate in a population due to genetic drift. Therefore, deleterious alleles with a small effect on fitness behave like neutral alleles. At high frequencies, these weakly deleterious alleles are selected against. However, in small populations, weakly deleterious alleles behaving nearly neutrally can segregate at high frequencies due to the low *Ne*.

The nearly neutral theory can be explained by the equation *S*=4*Nes*, where *S* is the probability of a mutant reaching fixation in a population, *Ne* 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 3).



**Figure 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 *Ne*, 4*Nes* is close to zero and deleterious mutants behave nearly neutrally [(Kimura, 1983)](https://paperpile.com/c/HQttr3/9Ze7).

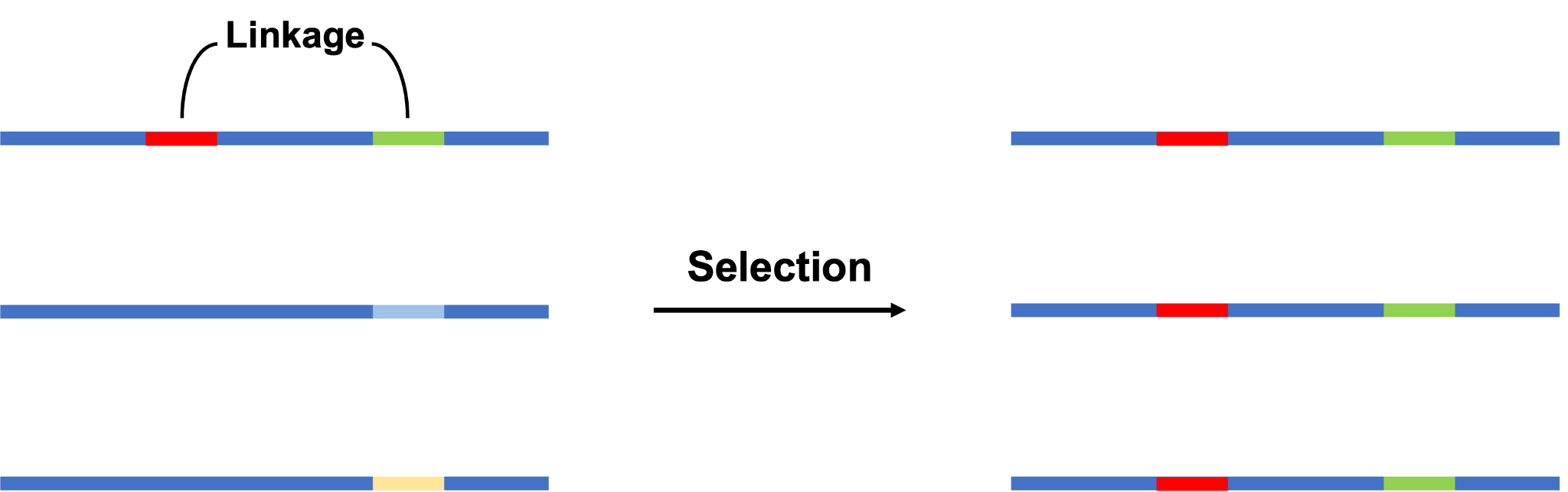
In a population with large *Ne*, if the selection coefficient (*s*)of a mutation is negative (i.e., the mutation has a strong deleterious effect), the probability of the mutation reaching fixation (4*Nes*) is highly negative and the mutation is likely to be purged from the population very quickly. For the same selective value in populations with a smaller *Ne*, the probability of deleterious mutations reaching fixation is very close to zero, meaning the mutation is behaving 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 *Ne*,a mutation would need to be highly deleterious to be purged from the population. Therefore, weakly deleterious mutations can accumulate in small populations.

It is worth noting that this theory has some limitations. The theory relies 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 [(Charlesworth, 2003)](https://paperpile.com/c/HQttr3/pztn). The model is therefore not accurate for mutation hotspots [(Charlesworth, 2003)](https://paperpile.com/c/HQttr3/pztn), where the chance of multiple mutations occurring at the same site is increased.

## 

## 2.5. Hitchhiking

Due to linkage, a selectively neutral or deleterious mutant may be linked to a selectively advantageous gene [(Smith and Haigh, 1974)](https://paperpile.com/c/HQttr3/3bYT). This means that as the selectively advantageous gene is selected for and potentially becomes fixed in a population, so does the deleterious mutant (Figure 4).



**Figure 4.** 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 passed on, resulting in both advantageous and deleterious mutants segregating in the population.

Hitchhiking during artificial selection plays a large role in the accumulation of deleterious mutations in domesticated animals. Due to hitchhiking, strong selection can result in a loss of genetic variation at linked loci [(Smith and Haigh, 1974)](https://paperpile.com/c/HQttr3/3bYT). If weakly deleterious mutations are linked to desirable alleles, they can be passed through the population by hitchhiking, resulting in an accumulation [(Lu *et al.*, 2006; Cruz *et al.*, 2008)](https://paperpile.com/c/HQttr3/3XeP+tKxR). The domestication-associated Hill-Robertson (dHR) effect, which describes interference between natural and artificial selection due to the impact of domestication on recombination, increases the frequency of deleterious mutations due to hitchhiking [(Lu *et al.*, 2006)](https://paperpile.com/c/HQttr3/3XeP).

## 2.6. Inbreeding

Inbreeding is reproduction by related individuals. This can occur in small populations due to demographic events (e.g., bottlenecks) or due to controlled breeding. In small populations, inbreeding increases the chance of homozygous expression of recessive deleterious alleles (genetic load), many of which are at low frequency and would not otherwise be expressed, resulting in inbreeding depression [(Mukai *et al.*, 1972; Bosse *et al.*, 2019)](https://paperpile.com/c/HQttr3/H5eu+t7Ws). Inbreeding depression is the loss of fitness of a population due to increased homozygosity and expression of recessive deleterious mutations [(Charlesworth and Charlesworth, 1987; Charlesworth and Willis, 2009)](https://paperpile.com/c/HQttr3/eCj3+26Nq). In *Drosophila*, mutations with a large deleterious effect (e.g. lethal mutations) contribute significantly to inbreeding depression [(Charlesworth and Charlesworth, 1999)](https://paperpile.com/c/HQttr3/zYAK), but the cumulative impact of mutations with small effects can also significantly contribute to inbreeding depression [(Charlesworth and Charlesworth, 1987)](https://paperpile.com/c/HQttr3/eCj3).

A direct consequence of inbreeding is long homozygous regions: runs of homozygosity (ROH). ROH can be used to measure and detect recent and historical inbreeding in a population [(Curik *et al.*, 2014)](https://paperpile.com/c/HQttr3/f7A7). This method has been shown to give a resolution high enough to measure the identical-by-descent fraction of the genome (*FROH*), and in one study, runs of homozygosity were used to show that entire chromosome pairs of grey wolves (*Canis lupus*) were identical-by-descent [(Kardos *et al.*, 2018)](https://paperpile.com/c/HQttr3/C109).

Inbreeding can have severe and rapid effects on a population. The population of Scandinavian grey wolves was founded by one female and two males, leading to substantial inbreeding due to the founder effect [(Viluma *et al.*, 2022)](https://paperpile.com/c/HQttr3/DIBK). After approximately five generations, 10-24% of the diploid genomes were lost, including 160k SNPs for adaptive variants and wild-type alleles masking recessive deleterious alleles [(Viluma *et al.*, 2022)](https://paperpile.com/c/HQttr3/DIBK). Therefore, 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, inbreeding depression is observed in modern individuals as a result of a bottleneck 10 kya [(O’Brien, 1994b; O’Brien, 1994a)](https://paperpile.com/c/HQttr3/pRGG+bu0F).

Inbreeding can also result in severe phenotypic abnormalities, as has been observed in multiple wild mammals. The cheetah (*Acinonyx jubatus*)has incredibly low genetic variation due to a severe bottleneck at the end of the last ice age, with over 90% homozygosity observed in the genome [(Menotti-Raymond and O’Brien, 1993; Dobrynin *et al.*, 2015)](https://paperpile.com/c/HQttr3/9b8F+VeeI). As a result of this inbreeding, cheetahs have been observed with low quality and pleiomorphic sperm, asymmetrical skulls (a common sign of inbreeding) and high juvenile mortality in the wild and in captivity [(Wildt *et al.*, 1983; O’Brien *et al.*, 1985; Wayne *et al.*, 1986)](https://paperpile.com/c/HQttr3/fFS9+nAKi+aIKH). The Florida panther (*Puma concolor coryi*) experienced a severe bottleneck in the 1970s leading to sperm abnormalities, congenital heart defects and a high load of deadly infectious diseases [(Roelke *et al.*, 1993)](https://paperpile.com/c/HQttr3/h72c). Another species with a recent but severe bottleneck is the black-footed ferret (*Mustela nigripes*); disease susceptibility, low fecundity and poor quality sperm have been observed as effects of inbreeding depression [(Santymire *et al.*, 2006; Santymire *et al.*, 2014)](https://paperpile.com/c/HQttr3/SYxv+uW74).

Endangered species are particularly vulnerable to inbreeding as there are limited breeding populations. In captive populations of endangered species, breeding programmes aim to reduce inbreeding [(Hedrick and Miller, 1992)](https://paperpile.com/c/HQttr3/illR): without management, genetic variation would decrease as the populations are often closed to novel diversity [(Lande and Barrowclough, 1987)](https://paperpile.com/c/HQttr3/1ry4). Captive populations can be crucial to the survival of a species or population that is almost extinct in the wild through genetic rescue [(Sandler *et al.*, 2021)](https://paperpile.com/c/HQttr3/eRil). For example, in the black-footed ferret, novel diversity was introduced to the wild population through cloning from cryopreserved material [(Sandler *et al.*, 2021)](https://paperpile.com/c/HQttr3/eRil). Introductions of individuals from related populations can also be used for genetic rescue; in the Florida panther, this approach reduced phenotypic defects caused by inbreeding [(Johnson *et al.*, 2010)](https://paperpile.com/c/HQttr3/joOu).

Inbreeding has also been observed in domesticated species, like the dog (*Canis lupus familiaris*). Most dog breeds have closed breeding populations that started with a small founder population and many breeds have experienced additional bottlenecks since domestication [(Mellersh, 2008)](https://paperpile.com/c/HQttr3/fojQ), resulting in detrimental effects. For example, in golden retrievers, inbreeding has negatively impacted lifespan [(Yordy *et al.*, 2020)](https://paperpile.com/c/HQttr3/6WIy). In German shepherds and Labrador retrievers, inbreeding and increased homozygosity are significantly associated with hip dysplasia [(Mäki *et al.*, 2001)](https://paperpile.com/c/HQttr3/YOV7).

## 2.7. Genetic purging

Despite the fact that smaller populations are likely to have decreased fitness due to increased homozygosity and fixation of deleterious alleles, there is evidence that genetic purging can decrease inbreeding load and depression of these populations [(López-Cortegano *et al.*, 2021)](https://paperpile.com/c/HQttr3/HJ9c).

Purging occurs as recessive deleterious alleles are exposed to natural selection due to increased homozygosity caused by inbreeding [(Urfer, 2009)](https://paperpile.com/c/HQttr3/Cv6u). Individuals with expression of the deleterious allele do not survive or are unable to reproduce, meaning the deleterious allele is purged from the population [(Kalinowski *et al.*, 2000)](https://paperpile.com/c/HQttr3/TFOm). 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; 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 [(Charlesworth and Willis, 2009)](https://paperpile.com/c/HQttr3/26Nq).

Purging has been suggested to be a method by which inbreeding depression can be reduced in endangered captive populations through deliberate inbreeding, as this would expose deleterious mutations to selection [(Hedrick, 1994)](https://paperpile.com/c/HQttr3/ufIy). However, purging is risky in these populations as the increased homozygosity and therefore expression of deleterious alleles poses an extinction risk, which is dangerous for species which are already endangered [(Hedrick, 1994)](https://paperpile.com/c/HQttr3/ufIy). This deliberate inbreeding may also cause the loss of genetic diversity [(Hedrick, 1994)](https://paperpile.com/c/HQttr3/ufIy); in a population where genetic diversity is already limited, further loss of diversity may increase the risk of extinction.

Despite these risks, purging has been shown to effectively reduce mutational load in captive ungulate populations; a substantial fraction of the inbreeding load of these populations with low *Ne* was purged [(López-Cortegano *et al.*, 2021)](https://paperpile.com/c/HQttr3/HJ9c). It was suggested that larger populations would require more generations for purging to be detected compared to smaller populations [(López-Cortegano *et al.*, 2021)](https://paperpile.com/c/HQttr3/HJ9c). This evidence suggests that purging in captive populations is possible, but care must be taken when studying such populations as the increased fitness observed may be due to a fitness rebound caused by the captive environment and increased husbandry [(Kalinowski *et al.*, 2000; Clifford *et al.*, 2007)](https://paperpile.com/c/HQttr3/TFOm+ZBBb).

# Domesticated & wild mammals

## 3.1. Wild mammals

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 1000 times the background rate of extinction [(Pimm *et al.*, 2014)](https://paperpile.com/c/HQttr3/pGyj). Therefore, wild mammal populations are at risk. The process of population decline, bottlenecks and resultant inbreeding occurs in wild populations but can be exacerbated by direct or indirect human interference. For example, while some wild populations may decline due to hunting/poaching or human-wildlife conflict, populations may also decline due to habitat fragmentation or the effects of climate change. As previously described (section 2.6), the impact of genetic bottlenecks on wild mammals has been observed in multiple species, such as the cheetah, Florida panther and black-footed ferret [(Wildt *et al.*, 1983; O’Brien *et al.*, 1985; Wayne *et al.*, 1986; Roelke *et al.*, 1993; Santymire *et al.*, 2006; Santymire *et al.*, 2014)](https://paperpile.com/c/HQttr3/fFS9+nAKi+aIKH+h72c+SYxv+uW74).

## 3.2. Domesticated mammals

Inbreeding and decreased *Ne*s are not only observed in wild mammals, but in domesticated mammals the causation can differ. The process of domestication can be characterised by a “domestication bottleneck” followed by strong artificial selection for desired traits, leading to inbreeding [(Eyre-Walker *et al.*, 1998; Galtier *et al.*, 2000; Bosse *et al.*, 2019)](https://paperpile.com/c/HQttr3/PORF+VJr2+H5eu). The mutational load in domestic mammals increases [(Bosse *et al.*, 2019)](https://paperpile.com/c/HQttr3/H5eu) whilst genetic diversity and efficacy of purifying selection is reduced [(Cruz *et al.*, 2008)](https://paperpile.com/c/HQttr3/tKxR), resulting in an increased frequency of weakly deleterious mutations [(Cruz *et al.*, 2008; Schubert *et al.*, 2014)](https://paperpile.com/c/HQttr3/tKxR+Nwdx). As a consequence, low reproductive fitness is commonly observed [(Lu *et al.*, 2006)](https://paperpile.com/c/HQttr3/3XeP). This is often referred to as the “cost of domestication” [(Lu *et al.*, 2006)](https://paperpile.com/c/HQttr3/3XeP), which has been observed in animal and plant species (see section 2.6) [(Lu *et al.*, 2006; Björnerfeldt *et al.*, 2006; Cruz *et al.*, 2008; Wang *et al.*, 2011; Koenig *et al.*, 2013; Nabholz *et al.*, 2014; Schubert *et al.*, 2014)](https://paperpile.com/c/HQttr3/tKxR+Nwdx+3XeP+yvjg+EEMn+p9kK+Yk59).

The proportion of weakly deleterious alleles segregating in dogs is increased due to domestication, which could be associated with the high levels of diseases in domestic dogs [(Cruz *et al.*, 2008)](https://paperpile.com/c/HQttr3/tKxR). A similar phenomenon is observed in domestic horses: the deleterious mutation load of modern domestic horse genomes was significantly higher than in pre-domesticated horses [(Schubert *et al.*, 2014)](https://paperpile.com/c/HQttr3/Nwdx). Schubert *et al*. (2014) reported that their calculation of mutation load was robust to inbreeding, meaning their findings are specifically due to domestication itself, rather than resultant or historic inbreeding.

The process of domestication also causes the relaxation of selection: individuals with weakly deleterious mutations are still able to reproduce resulting in an increase in the rate of non-synonymous to synonymous mutations [(Björnerfeldt *et al.*, 2006)](https://paperpile.com/c/HQttr3/p9kK). This relaxation of selective constraint can be caused by a reduction in *Ne* due to inbreeding, as is seen in domesticated species [(Gu *et al.*, 2005; Lu *et al.*, 2006)](https://paperpile.com/c/HQttr3/3XeP+xQjl).

It is crucial to understand domestication-associated genetic disorders in mammals for both economic and research purposes. For example, bovine leucocyte adhesion deficiency led to an economic loss of $5 million annually in the US [(Shuster *et al.*, 1992)](https://paperpile.com/c/HQttr3/pbK6), whilst strong artificial selection in dogs has resulted in genetic disorders that resemble human disorders [(Ostrander and Kruglyak, 2000; Sutter and Ostrander, 2004)](https://paperpile.com/c/HQttr3/8keu+pwAz), highlighting the value of studying domesticated species to investigate human diseases and disorders.

# 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 relied on the genotyping of specific markers in many individuals (e.g. Genome-Wide Association Studies [(Visscher *et al.*, 2017)](https://paperpile.com/c/HQttr3/xvcU)). However, a major limitation of this approach 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.

## 4.1. Reference genomes

Since the first fully sequenced genome (the bacteria *Haemophilus influenzae*)in 1995 [(Fleischmann *et al.*, 1995)](https://paperpile.com/c/HQttr3/6QrD), sequencing technology has advanced rapidly [(Hu *et al.*, 2021)](https://paperpile.com/c/HQttr3/mN2p) and thousands of more complex genomes have been sequenced [(Shumate and Salzberg, 2021)](https://paperpile.com/c/HQttr3/JWF6). Starting with the human [(Lander *et al.*, 2001; Venter *et al.*, 2001)](https://paperpile.com/c/HQttr3/3pp5+s61q) and mouse [(Mouse Genome Sequencing Consortium *et al.*, 2002)](https://paperpile.com/c/HQttr3/CjpC), the genomes of over 200 mammals have been sequenced in the last few decades [(Zoonomia Consortium, 2020)](https://paperpile.com/c/HQttr3/QHTM).

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)](https://paperpile.com/c/HQttr3/qeHO), 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)](https://paperpile.com/c/HQttr3/52vP). 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 [(Zoonomia Consortium, 2020)](https://paperpile.com/c/HQttr3/QHTM), which is an incredible resource for researchers.

The quality of reference genomes has advanced as new and improved genome assemblies are published [(Shumate and Salzberg, 2021)](https://paperpile.com/c/HQttr3/JWF6). In the mid 2000s, short-read next-generation sequencing was the cutting edge technology used for most genome sequencing [(Simon *et al.*, 2009)](https://paperpile.com/c/HQttr3/VpbC). Long-read sequencing technologies, such as PacBio [(Rhoads and Au, 2015)](https://paperpile.com/c/HQttr3/opft) and Oxford Nanopore [(Jain *et al.*, 2016)](https://paperpile.com/c/HQttr3/ajNg), 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 [(van Dijk *et al.*, 2018; Logsdon *et al.*, 2020)](https://paperpile.com/c/HQttr3/xxaa+OXlH). Further to this, chromosome conformation capture sequencing methods, such as Hi-C [(Belton *et al.*, 2012)](https://paperpile.com/c/HQttr3/IVFB), can be used to order contigs 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 [(Nurk *et al.*, 2022; Church, 2022)](https://paperpile.com/c/HQttr3/MV13+ItP0).

## 4.2. 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 noncoding 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)](https://paperpile.com/c/HQttr3/6cTn). 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)](https://paperpile.com/c/HQttr3/erbk). Iso-seq enables the sequencing of full-length isoforms to understand splice variation and improve an annotation [(Beiki *et al.*, 2019)](https://paperpile.com/c/HQttr3/GK7h). Another approach uses sequence features such as start codons, stop codons, and splice sites to identify coding sequences [(Stein, 2001)](https://paperpile.com/c/HQttr3/Q3fU). These methods create a comprehensive set of data that can be used to generate an annotation.

The process of genome annotation presents a computational challenge [(Salzberg, 2019)](https://paperpile.com/c/HQttr3/6cTn). Initial genome annotation tools, such as Genscan [(Burge and Karlin, 1997)](https://paperpile.com/c/HQttr3/5v2K), used pattern recognition approaches to identify protein-coding regions. However, these patterns may occur by chance so there was a likelihood of false-positive results, meaning other approaches were required [(Miller *et al.*, 2004)](https://paperpile.com/c/HQttr3/aMsB). More recent tools such as MAKER [(Cantarel *et al.*, 2008)](https://paperpile.com/c/HQttr3/BT72), BRAKER [(Hoff *et al.*, 2016)](https://paperpile.com/c/HQttr3/YyKR), Funannotate [(Palmer, 2016)](https://paperpile.com/c/HQttr3/HXwx) and Augustus [(Hoff and Stanke, 2019)](https://paperpile.com/c/HQttr3/CEuX) use iterative unsupervised training and RNA-seq information to generate more accurate annotations [(Cantarel *et al.*, 2008; Hoff *et al.*, 2016; Hoff and Stanke, 2019)](https://paperpile.com/c/HQttr3/YyKR+CEuX+BT72). 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; Li *et al.*, 2021)](https://paperpile.com/c/HQttr3/S96K+Tl37).

Genome annotations are most accurate when experimental data is used to inform predictions [(Miller *et al.*, 2004)](https://paperpile.com/c/HQttr3/aMsB). 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 and Salzberg, 2021)](https://paperpile.com/c/HQttr3/JWF6), such as in tools like UCSC liftOver [(Kuhn *et al.*, 2013)](https://paperpile.com/c/HQttr3/7dfz) and halLiftover [(Hickey *et al.*, 2013)](https://paperpile.com/c/HQttr3/RvVV). A more recent tool, LiftOff [(Shumate and Salzberg, 2021)](https://paperpile.com/c/HQttr3/JWF6), 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 in the target species with homologs in the reference species will be annotated; genes 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.

## 4.3. Comparative genomics

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)](https://paperpile.com/c/HQttr3/aMsB). 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)](https://paperpile.com/c/HQttr3/kJg6).

Such comparisons rely on whole genome alignments, enabling the identification of

conserved or non-conserved regions across species. Tools such as LastZ [(Harris, 2007)](https://paperpile.com/c/HQttr3/j0ox), MultiZ [(Blanchette *et al.*, 2004)](https://paperpile.com/c/HQttr3/U8bG) and Cactus [(J. Armstrong *et al.*, 2020)](https://paperpile.com/c/HQttr3/Pkqv) can be used to generate whole genome alignments, which is a computationally intensive process as billions of bases must be aligned. However, care must be taken when comparing reference genomes that were annotated using different methods as orthologous genes may be annotated in one species but not another, therefore erroneously appearing to be lineage-specific [(Weisman *et al.*, 2022)](https://paperpile.com/c/HQttr3/XoTv).

In 2011, the genomes of 29 eutherian mammals were published alongside a comparative study [(Lindblad-Toh *et al.*, 2011)](https://paperpile.com/c/HQttr3/L4Qy) which identified regions of the human genome under positive or purifying selection. A recent breakthrough 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 uncharacterized, enabling novel investigation into mammalian evolutionary constraint at a higher resolution than before and greater insight into genomic variants associated with increased disease risk in both mammals and humans [(Zoonomia Consortium, 2020)](https://paperpile.com/c/HQttr3/QHTM).

## 4.4. 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 predicts whether a single nucleotide polymorphism (SNP) affects the function of the protein and distinguishes between functionally neutral and deleterious mutations [(Ng and Henikoff, 2003)](https://paperpile.com/c/HQttr3/2io9). Identifying deleterious or pathogenic variants can give insight into diseases and associated healthcare, as well as evolution and gene expression [(Frazer *et al.*, 2021)](https://paperpile.com/c/HQttr3/fYdu).

A pathogenic variant is one that increases the individual’s susceptibility/predisposed risk to developing a disease or disorder. Methods to quantify the pathogenicity of protein variants may rely on training machine learning models using known disease labels [(Frazer *et al.*, 2021)](https://paperpile.com/c/HQttr3/fYdu). 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)](https://paperpile.com/c/HQttr3/Ctx9), but this may not apply to other species.

A recent study [(Frazer *et al.*, 2021)](https://paperpile.com/c/HQttr3/fYdu) 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)](https://paperpile.com/c/HQttr3/fquX). 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)](https://paperpile.com/c/HQttr3/fquX).

Previous studies on the effects of inbreeding have mainly focused on gene 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. Pseudogenization is a key interest here; deleterious mutations occurring within genes, leading to PTCs, missing start codons or frameshift mutations, results in pseudogenisation (loss) of the gene. 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)](https://paperpile.com/c/HQttr3/pcPN+L4Qy+lZRZ). Therefore, the focus must shift to functional noncoding regions of the genome.

# Noncoding regions

## 5.1. Functional noncoding regions

The majority of disease- and trait-associated variants identified in GWAS studies are located in functional noncoding regions [(Hindorff *et al.*, 2009; ENCODE Project Consortium, 2012; Ricaño-Ponce and Wijmenga, 2013)](https://paperpile.com/c/HQttr3/wJ8u+SQeB+En9K). 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)](https://paperpile.com/c/HQttr3/SQeB). Approximately 8% of noncoding sites in the human genome are suggested to be functional based on evidence of selective constraints [(Siepel *et al.*, 2005; Meader *et al.*, 2010; Rands *et al.*, 2014; Huang *et al.*, 2017)](https://paperpile.com/c/HQttr3/5oMi+b8If+XWXs+WdHb).

Functional noncoding regions are often involved in regulating gene expression [(Carroll, 2000; Wray *et al.*, 2003; Lindblad-Toh *et al.*, 2011)](https://paperpile.com/c/HQttr3/L4Qy+pcPN+lZRZ), and include regulatory elements such as promoters, enhancers, silencers and insulators [(Maston *et al.*, 2006)](https://paperpile.com/c/HQttr3/qxCG). Transcription factor binding sites (TFBS) are of particular interest in the noncoding genome as these play a key role in regulating gene expression [(Boeva, 2016)](https://paperpile.com/c/HQttr3/Dhgj). 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)](https://paperpile.com/c/HQttr3/JcU4).

## 5.2. Identification of noncoding regions

Despite the importance of investigating deleterious mutations in noncoding regions, the difficulty in identifying functional noncoding regions has limited the scope of existing studies. Despite the fact that functional noncoding regions are usually highly conserved, this is not sufficient to accurately predict regions such as TFBS [(Cochran *et al.*, 2022)](https://paperpile.com/c/HQttr3/1cI3). Additionally, although GWAS and whole genome sequencing have enabled the identification of trait-associated variants and their abundance in noncoding regions [(ENCODE Project Consortium, 2012)](https://paperpile.com/c/HQttr3/wJ8u), it can be difficult to determine whether the association between variant and trait is due to the specific variant itself or a linked variant [(Chen and Wang, 2022)](https://paperpile.com/c/HQttr3/7pI2).

There are multiple methods to identify putative TFBS [(Boeva, 2016)](https://paperpile.com/c/HQttr3/Dhgj). The Position Weight Matrix (PWM) is a widely-used mathematical model using a probability of each nucleotide at each position in the binding motif [(Stormo, 2000)](https://paperpile.com/c/HQttr3/SqyF). Supervised classification methods, such as Bayesian networks, approaches based on Markov models and searching for k-mers via a support vector machine (SVM), can be used to scan a whole genome to identify potential TFBS [(Boeva, 2016)](https://paperpile.com/c/HQttr3/Dhgj). These methods rely on recognised binding motifs or a set of DNA sequences containing binding sites [(Boeva, 2016)](https://paperpile.com/c/HQttr3/Dhgj).

Recent studies have identified functional noncoding variants using high-throughput functional assays [(Melnikov *et al.*, 2014; Wen *et al.*, 2020)](https://paperpile.com/c/HQttr3/2hr2+fB9M). 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 noncoding variants, such laboratory protocols cannot feasibly be scaled up to identify the millions of variants that exist in the noncoding genome [(Koch, 2020; Chen and Wang, 2022)](https://paperpile.com/c/HQttr3/ldWs+7pI2). Therefore, to fully understand these variants, another method must be used.

## 5.3. Machine learning

Machine learning can be used to identify noncoding variants. Deep learning algorithms can be used for large feature-rich datasets, as this method is scalable and able to identify complex patterns, but large training data sets are required to train deep neural networks [(LeCun *et al.*, 2015)](https://paperpile.com/c/HQttr3/scQl). CADD (Combined Annotation-Dependent Depletion) is an annotation method which uses an SVM to annotate coding and noncoding variants in a genome [(Kircher *et al.*, 2014)](https://paperpile.com/c/HQttr3/jZKY). The SVM separates simulated and observed genetic variants, but cannot identify non-linear relationships among the features [(Kircher *et al.*, 2014)](https://paperpile.com/c/HQttr3/jZKY). A subsequent method, DANN (Deleterious Annotation of genetic variants using Neural Networks), uses a deep neural network which can capture non-linear relationships among features (unlike CADD) and performs better than SVMs with large numbers of samples and features [(Quang *et al.*, 2015)](https://paperpile.com/c/HQttr3/GvUB). DANN is more accurate when using a testing set which contains mostly noncoding variants. However, a sufficient testing set is not necessarily available for many species.

A recent approach, TLVar, predicts context-specific functional noncoding variants using deep transfer learning with convolutional neural networks [(Chen and Wang, 2022)](https://paperpile.com/c/HQttr3/7pI2). This method has been shown to outperform deep learning approaches, but it has not been tested on cross-species analyses – its strength is in identifying tissue/cell specific variants or causal variants of diseases [(Chen and Wang, 2022)](https://paperpile.com/c/HQttr3/7pI2).

Machine learning can be used to identify functional noncoding regions in species with limited resources. Data from highly studied species can be used to train a machine learning model, which is then applied to the genomes of less-well studied species to identify functional noncoding regions in these genomes. However, cross-species predictions with neural networks perform worse than predictions within the same species due to species-specific sequence repeats [(Cochran *et al.*, 2022)](https://paperpile.com/c/HQttr3/1cI3). For cross-species predictions of functional noncoding regions such as TFBS to be accurate, the neural network model must be discouraged from learning the species-specific features – this can be done automatically using an augmented network architecture [(Cochran *et al.*, 2022)](https://paperpile.com/c/HQttr3/1cI3).

# Project overview

Whilst previous research has investigated the relationship between decreased population size and an accumulation of deleterious mutations, little is known about deleterious mutations in functional noncoding regions, particularly in domesticated and wild mammal species. However, a wealth of genomic data is now available for these species, and combined with machine learning techniques, it is possible to investigate these mutations. Therefore, this project will investigate the accumulation of deleterious mutations in noncoding regions of genomes of species which have experienced a bottleneck in population size, either due to domestication-related inbreeding or threats in the wild.

There are three main aims of this project:.

1. Quantify population demographic history:

The initial aim of the project is to assess the demographic history of domesticated and wild endangered species since their most recent common ancestor. Sequentially Markovian Coalescent models (PSMC/MSMC) will be used to infer past effective population sizes, and population data will be used where possible to further refine the demographic models.

1. Identify noncoding regulatory elements and impact of mutations:

Using the wealth of genomic data available for model organisms (e.g., mice and humans), deep learning models (specifically convolutional neural networks) will be trained to predict functional elements such as TFBS in species with less genomic resources. Genome alignments and population data where available will be used to assess the evolution of TFBS; the demographic inferences from Aim 1 will be applied here to refine predictions. This data will also be used to identify sites with potentially deleterious mutations segregating at a high frequency.

1. Characterise the impact of noncoding variants:

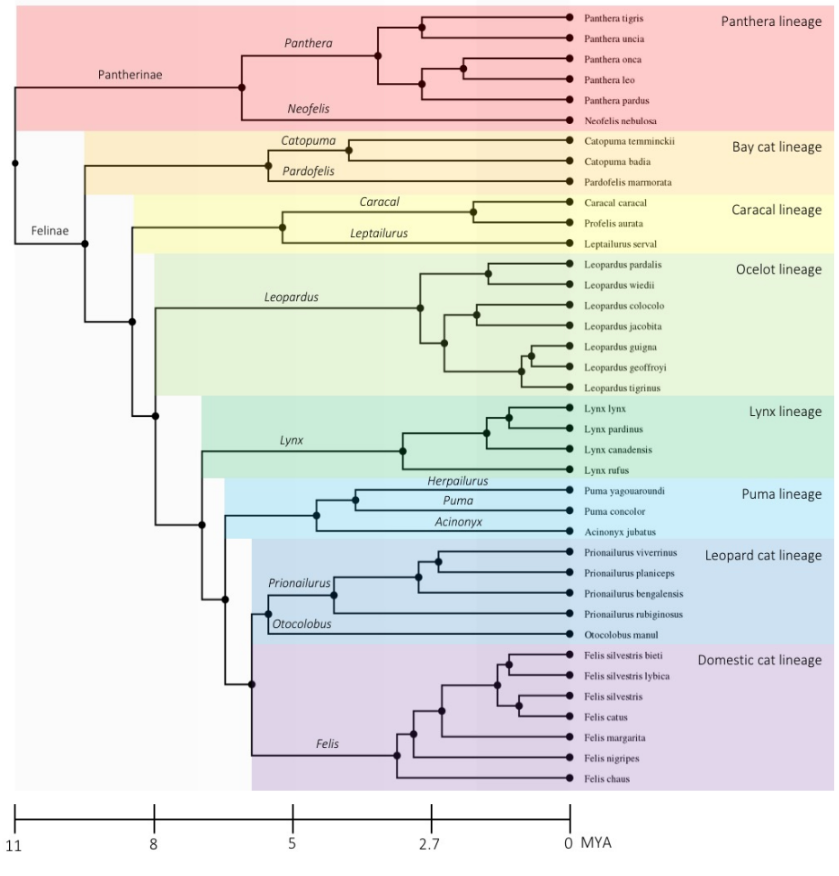
The final aim of the project is to determine the impact of the variants identified in Aim 2. A low-cost plate-based luminescence assay will be used to quantify transcription factor binding affinity to synthesised DNA fragments. As transcription factors are highly conserved, these will be used from a model organism [check?].

# Work to date: Background

To gain the skills and experience needed to undertake the project, an initial study was conducted using a subset of genomes from the Zoonomia project: the Felidae.

## 7.1. Felidae

Felidae is a family of mammals in the Carnivora order containing 38 extant species across the globe [(Samaha, 2021)](https://paperpile.com/c/HQttr3/KJnw). The species are divided into two subfamilies, Pantherinae and Felinae, split into 14 genera [(Samaha, 2021)](https://paperpile.com/c/HQttr3/KJnw) (Figure 5). Pantherinae includes the Panthera and Neofelis genera, whilst Felinae spans the rest of the family, including the puma, cheetah and domestic cat.



**Figure 5.** Phylogeny of all extant felid species [(Samaha, 2021)](https://paperpile.com/c/HQttr3/KJnw).

This study utilises the reference genomes of eight felid species: *Acinonyx jubatus* (cheetah) [(Dobrynin *et al.*, 2015)](https://paperpile.com/c/HQttr3/VeeI), *Felis catus* (domestic cat) [(Buckley *et al.*, 2020)](https://paperpile.com/c/HQttr3/JSpl), *Felis nigripes* (black-footed cat) [(Zoonomia Consortium, 2020)](https://paperpile.com/c/HQttr3/QHTM), *Panthera leo* (lion) [(E. E. Armstrong *et al.*, 2020)](https://paperpile.com/c/HQttr3/lLXE), *Panthera onca* (jaguar) [(Zoonomia Consortium, 2020)](https://paperpile.com/c/HQttr3/QHTM), *Panthera pardus* (leopard) [(Kim *et al.*, 2016)](https://paperpile.com/c/HQttr3/L7br), *Panthera tigris* (tiger) [(Cho *et al.*, 2013)](https://paperpile.com/c/HQttr3/dkXg) and *Puma concolor* (puma) [(Saremi *et al.*, 2019)](https://paperpile.com/c/HQttr3/yCWN).

The Zoonomia project generated two new reference genomes (*P. onca* and *F. nigripes*), meaning limited work has been done on these genomes so far. Additionally, several genomes are chromosome-level, namely the domestic cat [(Buckley *et al.*, 2020)](https://paperpile.com/c/HQttr3/JSpl) and the lion [(E. E. Armstrong *et al.*, 2020)](https://paperpile.com/c/HQttr3/lLXE). Therefore, the species chosen for this work will enable new insight into the genomics of Felidae.

Felidae was chosen as it is a good representative family with examples of the demographic features that will be seen in the domesticated and wild endangered species later in my project. For example, some species have experienced historic bottlenecks and historically low *Ne*s whereas other species have had more recent declines. It is also possible to compare the domestic cat to wild felids. The Felidae are also a good family to use due to the high levels of conserved synteny between felids, as has been seen between the lion, Panthera genus and the domestic cat [(E. E. Armstrong *et al.*, 2020)](https://paperpile.com/c/HQttr3/lLXE) and between the domestic cat, cheetah, snow leopard and Sumatran tiger [(Samaha *et al.*, 2021)](https://paperpile.com/c/HQttr3/gg7F).

## 7.2. Demographic histories

The demographic histories of these species are varied, making them an interesting family to study. Here, I will summarise current literature on the demographic histories of the eight felid species I am investigating.

### Cheetah (Acinonyx jubatus)

It has been suggested that the cheetah has experienced multiple severe genetic bottlenecks, with a low effective population size for the last 3 million years [(Kim *et al.*, 2016)](https://paperpile.com/c/HQttr3/L7br). The cheetah experienced a severe bottleneck at the end of the last ice age in the late Pleistocene: a near extinction event where all but a handful went extinct [(Menotti-Raymond and O’Brien, 1993; Dobrynin *et al.*, 2015)](https://paperpile.com/c/HQttr3/9b8F+VeeI). An additional bottleneck over 100 kya has also been identified, thought to be when the cheetah migrated out of the Americas and into Eurasia/Africa [(Dobrynin *et al.*, 2015)](https://paperpile.com/c/HQttr3/VeeI).

Due to these severe bottlenecks, cheetahs suffer from inbreeding depression leading to increased susceptibility to infectious diseases, developmental deformities, a high frequency of spermatozoal abnormalities and high juvenile mortality in the wild and in captivity [(Wildt *et al.*, 1983; O’Brien *et al.*, 1985; Wayne *et al.*, 1986)](https://paperpile.com/c/HQttr3/fFS9+aIKH+nAKi). Surgical skin grafts between 12 unrelated individuals were accepted as the MHC alleles were all identical [(O’Brien *et al.*, 1985)](https://paperpile.com/c/HQttr3/aIKH). Cheetahs are the only felid species where the whole species is inbred: there is no non-inbred population [(O’Brien and Johnson, 2005)](https://paperpile.com/c/HQttr3/wdtK).

### Domestic cat (Felis catus)

The cat is hypothesised to have been domesticated in the Near East around 9.5 kya based on archaeological evidence from Cyprus [(Vigne *et al.*, 2004; Driscoll *et al.*, 2007)](https://paperpile.com/c/HQttr3/vKgI+LD4r). Cat domestication is thought to have begun as a symbiotic relationship between cats and humans as cats had value to humans as vermin control [(Driscoll *et al.*, 2007)](https://paperpile.com/c/HQttr3/vKgI). Unlike other domesticated species, cats were unlikely to have experienced strong selective forces during the domestication process as they were already capable rodent hunters, although coat colour variability was selected for [(Montague *et al.*, 2014)](https://paperpile.com/c/HQttr3/ojjx). Since domestication, cats have experienced artificial selection mainly for aesthetic traits and most cat breeds originated in the last 150 years [(Alhaddad *et al.*, 2013; Montague *et al.*, 2014)](https://paperpile.com/c/HQttr3/ojjx+53vG). Several cat breeds show evidence of recent bottlenecks, which potentially occurred due to artificial selection as the breeds were formed [(Menotti-Raymond *et al.*, 2008; Alhaddad *et al.*, 2013)](https://paperpile.com/c/HQttr3/53vG+yADr). However, no studies have identified major historic bottlenecks in the domestic cat.

### Black-footed cat (Felis nigripes)

Very few studies have been undertaken on the black-footed cat [(Mattucci *et al.*, 2019)](https://paperpile.com/c/HQttr3/KBkV), and there has been little research on their demographic history. Mattucci *et al.* (2019) determined that the black-footed cat did not experience a bottleneck when it split from the domestic cat and wildcat and high levels of genome-wide heterozygosity were identified. A contrasting study suggests that there is evidence of a possible bottleneck but gives no citation [(Wilson, 2015)](https://paperpile.com/c/HQttr3/jgAI).

### Lion (Panthera leo)

The lion is divided into two key populations: the Asiatic lion in India and the African lion in West and Central Africa. The Asiatic lion has evidence of multiple historic bottlenecks; it is now only found in the Gir Forest and surrounding areas in Gujarat, India [(Singh, 2017; Singh and Nala, 2018)](https://paperpile.com/c/HQttr3/QwoN+tFLs). One bottleneck occurred 1.1-4.3 kya, with another approximately 100 years ago due to hunting and habitat loss, reducing the population to less than 20 individuals [(Gilbert *et al.*, 1991; Driscoll *et al.*, 2002)](https://paperpile.com/c/HQttr3/dn2s+RJVK). The current population is estimated to be approximately 500 individuals [(Singh, 2017; Singh and Nala, 2018)](https://paperpile.com/c/HQttr3/QwoN+tFLs). Spermatozoa abnormalities have been identified in Asiatic lions [(Wildt *et al.*, 1987)](https://paperpile.com/c/HQttr3/Krwm), as has been seen in the cheetah [(Wildt *et al.*, 1983; O’Brien *et al.*, 1985)](https://paperpile.com/c/HQttr3/fFS9+aIKH).

Antunes *et al.* (2008) suggest there is evidence of several lion expansions across Asia and Africa (169-324 kya, 100 kya and 7-14 kya). Based on nuclear and mitochondrial DNA, it is suggested that lions have also experienced severe bottlenecks, although estimated dates are not stated [(Antunes *et al.*, 2008)](https://paperpile.com/c/HQttr3/5PPq).

### Jaguar (Panthera onca)

Although there is limited research on the demographic history of the jaguar, multiple studies suggest potential bottlenecks of Mexican, Brazilian and some Amazonian populations [(Roques *et al.*, 2016; Lorenzana *et al.*, 2020)](https://paperpile.com/c/HQttr3/9k0X+BuJc). Based on mitochondrial DNA haplotypes, one study identified a potential recent population bottleneck and expansion in the jaguar, although there is no estimate of the timescale [(Eizirik *et al.*, 2001)](https://paperpile.com/c/HQttr3/4MnI).

### Leopard (Panthera pardus)

There are eight subspecies of leopard and the species has a range across Africa, Asia and Southern Russia [(Jacobson *et al.*, 2016)](https://paperpile.com/c/HQttr3/Rm8I), the widest range of any cat species [(Miquelle *et al.*, 1996)](https://paperpile.com/c/HQttr3/YcFS). The Amur or Far Eastern leopard (*Panthera pardus orientalis*)is the most threatened subspecies [(Pečnerová *et al.*, 2021)](https://paperpile.com/c/HQttr3/VXM7) but also the subspecies from which the leopard reference genome was generated [(Kim *et al.*, 2016)](https://paperpile.com/c/HQttr3/L7br). Based on a PSMC comparison with African leopards, the Amur leopard was shown to have experienced a recent bottleneck [(Pečnerová *et al.*, 2021)](https://paperpile.com/c/HQttr3/VXM7): the range of the leopard reduced dramatically in the 20th century, with surveys in the 1990s estimating 25-40 individuals remained in just one population [(Miquelle *et al.*, 1996; Miquelle *et al.*, 2010)](https://paperpile.com/c/HQttr3/Wqy1+YcFS). It has been predicted that the leopard experienced another severe genetic bottleneck 900 thousand to 2 million years ago [(Kim *et al.*, 2016)](https://paperpile.com/c/HQttr3/L7br).

Additionally, a recent (last 30 ky) explosion in effective population size was identified in the leopard reference genome but not in wild leopard genomes. As the reference genome is from an individual with 30% admixture with the North-Chinese leopard, it was suggested that this explosion is actually an artefact of this admixture [(Kim *et al.*, 2016)](https://paperpile.com/c/HQttr3/L7br).

Amur leopards also carry a signature of an out-of-Africa expansion event in the Pleistocene (300 to 400 kya) and potentially additional founder effects since [(Pečnerová *et al.*, 2021)](https://paperpile.com/c/HQttr3/VXM7). The out-of-Africa expansion event, in which leopards first colonised Asia, is estimated to have occurred 500 to 600 kya based on PSMC analyses, resulting in a strong founder effect and decreased heterozygosity in Asian leopards compared to their African counterparts [(Paijmans *et al.*, 2021)](https://paperpile.com/c/HQttr3/pJBu).

### Tiger (Panthera tigris)

The tiger has been split into multiple subspecies (exact number debated [(Armstrong *et al.*, 2021)](https://paperpile.com/c/HQttr3/T9Lh)) including the Amur tiger, from which samples were used for the reference genome [(Cho *et al.*, 2013)](https://paperpile.com/c/HQttr3/dkXg). The Amur tiger is mainly found in the Sikhote-Alin Nature Reserve in the Russian Far East [(Kaplanov, 2005)](https://paperpile.com/c/HQttr3/tg8D) and multiple studies on its demographic history have been undertaken. It is widely accepted that the Amur tiger experienced a demographic collapse in the early 1940s, where the population was reduced to 20-30 individuals, resulting in a severe genetic bottleneck [(Kaplanov, 2005; Alasaad *et al.*, 2011)](https://paperpile.com/c/HQttr3/tg8D+9151). Very low levels of mitochondrial DNA variation were identified in the Amur tiger, likely due to this recent bottleneck [(Russello *et al.*, 2004)](https://paperpile.com/c/HQttr3/lMc7).

The existence of this bottleneck in tiger DNA has been debated; one microsatellite study did not find any evidence of a recent bottleneck [(Henry *et al.*, 2009)](https://paperpile.com/c/HQttr3/94Qh), however a more recent study with higher quality data and more microsatellite loci did find evidence [(Alasaad *et al.*, 2011)](https://paperpile.com/c/HQttr3/9151). Tigers were found to have long runs of homozygosity, suggesting recent inbreeding, and demographic models such as PSMC showed strong recent bottlenecks [(Armstrong *et al.*, 2021)](https://paperpile.com/c/HQttr3/T9Lh). This evidence further supports Alasaad *et al.*’s (2011) identification of a recent bottleneck. A very strong bottleneck has also been suggested approximately 234 kya [(Armstrong *et al.*, 2021)](https://paperpile.com/c/HQttr3/T9Lh). Using PSMC and mitochondrial DNA, bottlenecks were also estimated at 7 to 70 kya and 72 to 108 kya respectively [(Cho *et al.*, 2013)](https://paperpile.com/c/HQttr3/dkXg). It is important to note that it is unlikely that signatures of the recent bottleneck will be identified in the reference genome.

### Puma (Puma concolor)

The puma (also called panther or cougar) is one of the most famous examples of a genetic bottleneck. The Florida panther (*P. c. coryi*) experienced a severe bottleneck in the 20th century [(Culver *et al.*, 2008)](https://paperpile.com/c/HQttr3/aIHd), to the extent that the genetic diversity in modern individuals is much lower than is observed in 1890s museum samples [(Culver *et al.*, 2008)](https://paperpile.com/c/HQttr3/aIHd). Resultant symptoms of inbreeding depression include sperm abnormalities, congenital heart defects and a high load of deadly infectious diseases [(Roelke *et al.*, 1993)](https://paperpile.com/c/HQttr3/h72c).

The puma reference genome used samples from a Californian individual, so the bottleneck in the Florida panther population will not be observed. However, the species as a whole experienced a severe demographic decline in the late Pleistocene and went extinct in North America [(Culver *et al.*, 2000; Matte *et al.*, 2013)](https://paperpile.com/c/HQttr3/8KCa+JxtB). South American individuals were shown to experience a demographic expansion approximately 8 kya [(Matte *et al.*, 2013)](https://paperpile.com/c/HQttr3/JxtB) and were then hypothesised to have repopulated North America [(Culver *et al.*, 2000)](https://paperpile.com/c/HQttr3/8KCa). All North American pumas investigated in a recent study had short ROH dating to the early 20th century, suggesting small isolated populations and a potential bottleneck corresponding to extensive hunting [(Saremi *et al.*, 2019)](https://paperpile.com/c/HQttr3/yCWN).

## 7.3. Aims and objectives

Using genomic data from the species described in Section 7.2, the two main aims will be investigated: to identify deleterious functional noncoding mutations and to investigate gene family dynamics and evolution. For the first aim, a multiple genome alignment will be generated, highly conserved noncoding regions identified (with the assumption that highly conserved regions are likely to be functional) and then putatively deleterious variants within these regions identified. For the second aim, gene families will be identified that have expanded, contracted or been lost entirely from particular species or lineages. It will then be possible to consider whether such dynamics can be related to demographic events within those species or lineages, such as historic bottlenecks.

# 8. Progress and future work

# 8.1. Deleterious mutations in functional noncoding regions

Methods

To identify deleterious mutations in functional noncoding regions of felid genomes, the first key step was to generate a multiple genome alignment. Progressive Cactus [(J. Armstrong *et al.*, 2020)](https://paperpile.com/c/HQttr3/Pkqv) was chosen for this as it generates a reference-free alignment, meaning it is not biased to one reference species.

Alongside the eight felid species (Section 7.2), four well-annotated model species were chosen as outgroups: human (*Homo sapiens*), mouse (*Mus musculus*), pig (*Sus scrofa*) and dog (*Canis lupus familiaris*). Prior to running Cactus, it was necessary to soft-mask the genomes. RepeatMasker was run on 10 of the genomes (human and lion soft-masked genomes were downloaded from Ensembl instead).

Cactus was run on the genomes of the 12 mammal species with guidance from Will Nash. Issues were first encountered with allocated resources, as Cactus is computationally intensive and required increased disk space from our initial predictions. Further issues were encountered, so the version of Cactus was updated to the most recent version. One more issue was identified, this time with file paths, which was solved by altering the flags used when running Cactus. Cactus is currently still running.

Future work

Once the multiple genome alignment has been generated, I will investigate functional noncoding evolution by generating conservation scores to identify ultra-conserved (and therefore likely functional) noncoding regions using PhyloP [(Pollard *et al.*, 2010)](https://paperpile.com/c/HQttr3/yQTz). I will then identify deleterious variants within these regions using machine learning approaches. Tools such as FunSeq [(Bahcall, 2013)](https://paperpile.com/c/HQttr3/5AsT) will be used to assess the potential impact of variants (i.e., no impact, deleterious effect or novel binding). Machine learning tools to identify functional noncoding regions and deleterious variants mostly rely on training sets with identified features: for species with available information, these will be used. Neural network approaches will be used to identify functional noncoding sequences first in the Felidae species and then a wider set of species.

## 8.2. Gene family analysis

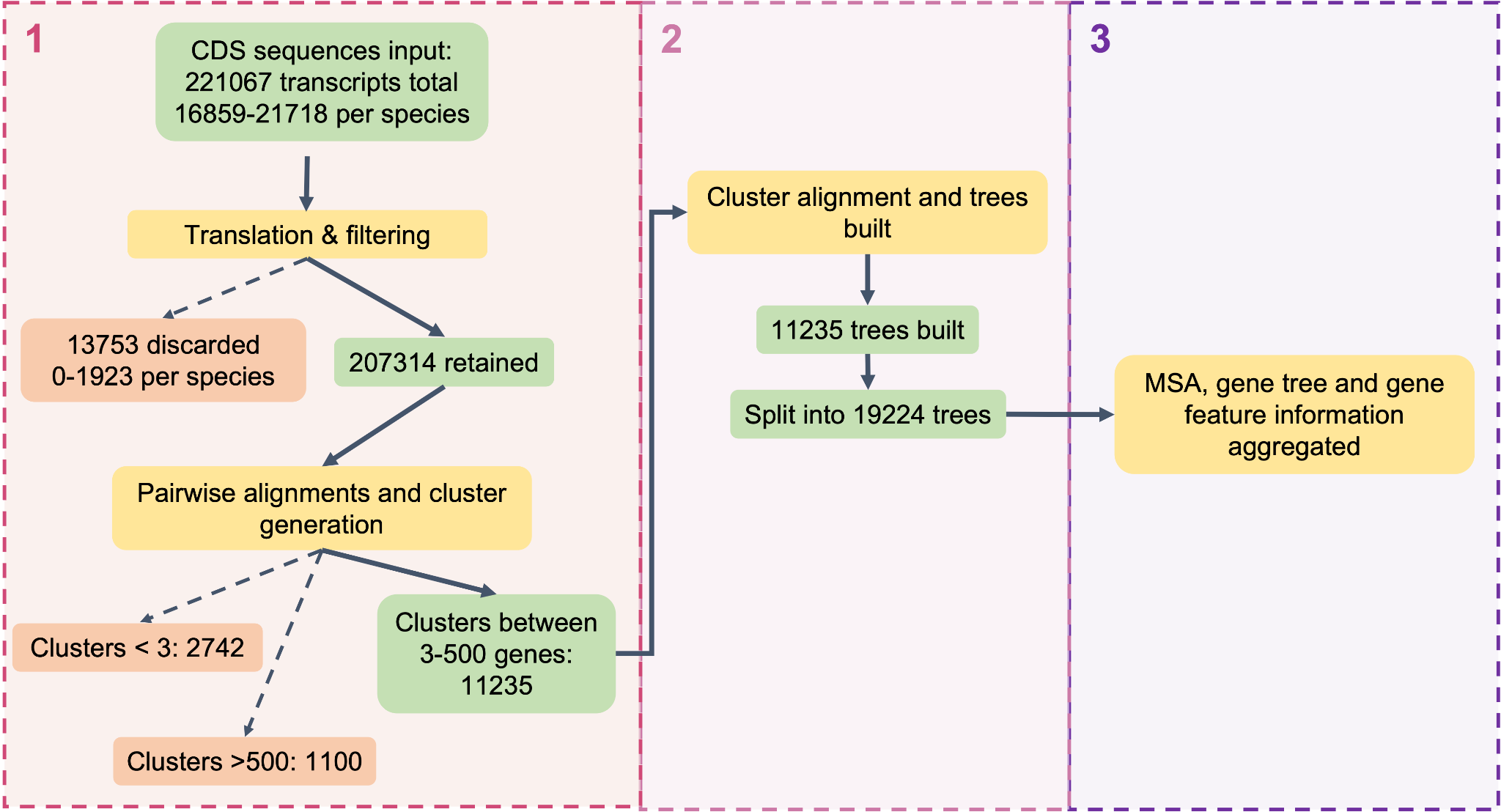
Methods

To investigate gene family expansions and contractions potentially related to historic demographic events, GeneSeqToFamily (GSTF) [(Thanki *et al.*, 2018)](https://paperpile.com/c/HQttr3/XEIT) was used on the 12 mammal species (see above). GSTF required coding sequences (CDS) from each species, as well as a Newick species tree. A custom Python script was written to extract CDS from each of the species (except human and mouse, for which the files were downloaded from Ensembl) using the annotations and reference genomes provided by the Zoonomia project. The number of coding sequences are shown in Table 1. Between 3-8% of the Zoonomia CDS were subsequently identified as noncoding (non-multiples of three). A script written by Anil Thanki was run to extract the longest transcript per gene. This required GFF files which were generated from GTF files using Gffread [(Pertea and Pertea, 2020)](https://paperpile.com/c/HQttr3/pxC8). Outputs are shown in Table 1. 9,673 transcripts were not included as these appear to be pseudogenes or were not protein coding genes. In total, 221,067 coding sequences were used as input for GeneSeqToFamily.

**Table 1.** Input and output of GeneSeqToFamily step 1. Columns show number of CDS extracted from each species, percentage of these sequences that are noncoding (non-multiple of 3), number of genes remaining after longest transcripts were extracted, number of transcripts filtered out by GSTF step 1 and number of transcripts that were subsequently put into clusters by GSTF.

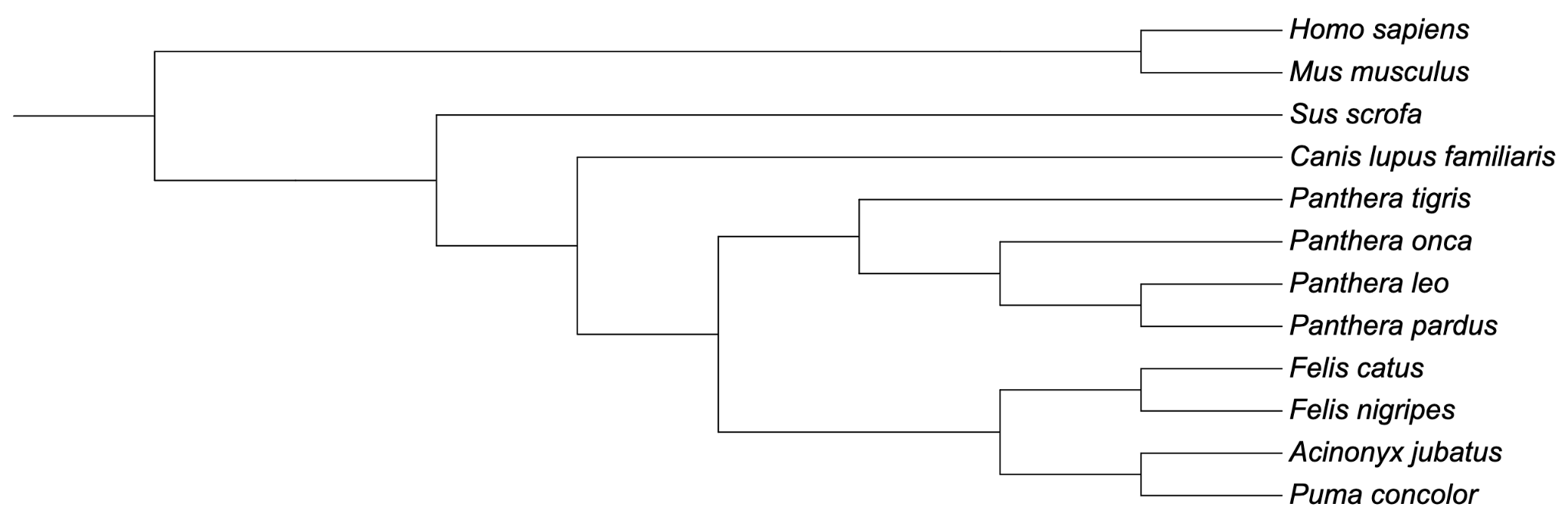
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Species | No. of CDS | % Noncoding | No. of genes (longest transcripts) | No. of transcripts removed by GSTF in step 1 | No. of transcripts remaining after step 1 |
| A. jubatus | 18632 | 7 | 17402 | 1923 | 15479 |
| F. catus | 18892 | 5 | 17885 | 1471 | 16414 |
| F. nigripes | 18772 | 3 | 18265 | 1280 | 16985 |
| P. leo | 19550 | <1 | 19546 | 0 | 19546 |
| P. onca | 18895 | 3 | 18251 | 1319 | 16932 |
| P. pardus | 18926 | 4 | 18216 | 1387 | 16829 |
| P. tigris | 18632 | 4 | 17804 | 1674 | 16130 |
| P. concolor | 18514 | 5 | 17665 | 1551 | 16114 |
| C. l. familiaris | 18855 | 5 | 17832 | 1582 | 16250 |
| H. sapiens | 20402 | 35 | 19624 | 42 | 19582 |
| M. musculus | 22438 | 25 | 21718 | 51 | 21667 |
| S. scrofa | 18232 | 8 | 16859 | 1473 | 15386 |
| Total | 230740 |  | 221067 | 13753 | 207314 |

GSTF was run on the HPC using Snakemake [(Koster and Rahmann, 2012)](https://paperpile.com/c/HQttr3/VYrV) scripts by Anil Thanki. The first step translates coding sequences to protein sequences, uses a BLAST parser tool [(Thanki *et al.*, 2018)](https://paperpile.com/c/HQttr3/XEIT) to find pairwise alignments of the sequences and then uses hcluster\_sg [(Li, 2006)](https://paperpile.com/c/HQttr3/uQ02) to generate clusters from the alignments. 13,753 CDS transcripts were filtered out, leaving 207,314 transcripts remaining (Table 1; Figure 6). From these, 13,350 clusters were generated. Clusters with less than 3 genes (of which there were 2742) were not included as a gene tree cannot be built from these, and clusters with over 500 genes (of which there were 1100) need to be analysed manually. 11,235 clusters with between 3 and 500 genes were carried through to the second step of GSTF.



**Figure 6.** The GeneSeqToFamily workflow. Yellow boxes describe GSTF’s method in its three key steps. Inputs/retained outputs are shown in green and discarded outputs are shown in red.

The second step, which uses T-Coffee [(Notredame *et al.*, 2000)](https://paperpile.com/c/HQttr3/7Zfg) to generate a multiple sequence alignment for each cluster and TreeBest [(Ruan et al. 2008)](https://paperpile.com/c/HQttr3/5hq6) to generate a gene tree, required a Newick species tree. A tree was generated using the Zoonomia tree [(Zoonomia Consortium, 2020)](https://paperpile.com/c/HQttr3/QHTM) as well as relevant literature [(Piras *et al.*, 2018; E. E. Armstrong *et al.*, 2020)](https://paperpile.com/c/HQttr3/lLXE+eIS2) (Figure 7). Gene trees were split if there was a duplication event prior to the most recent common ancestor of the species. In total, 19,224 split gene trees were generated. The third step of GSTF was then run, which merged the multiple sequence alignment, gene tree and gene feature information into a database.



**Figure 7.** Species tree generated using the Zoonomia tree [(Zoonomia Consortium, 2020)](https://paperpile.com/c/HQttr3/QHTM) and literature [(Piras *et al.*, 2018; E. E. Armstrong *et al.*, 2020)](https://paperpile.com/c/HQttr3/lLXE+eIS2).

Upon analysing the results of the first run, issues were identified with the tree-splitting algorithm. These were identified by investigating the gene trees for highly conserved gene families such as *HOX* genes; in initial runs of GSTF, these families were over-split and the gene trees did not look as expected. This algorithm was amended multiple times by Anil Thanki, using my data set to test and refine each iteration, resulting in an accurate tree-splitting algorithm which splits large gene trees into smaller trees based on the number of in- and out-group species in each clade.

Other minor issues were encountered during the process. For example, three species were missing a *HOXB7* gene in the penultimate run of the tool. However, on further investigation, it was identified that the genes for two of these species had been erroneously filtered out as duplicates, whilst the third (*P. tigris*) had a PTC. The duplicate filtering issue was easily resolved, and the PTC is likely a sequencing error. It is noteworthy that the Ensembl annotation for the tiger *HOXB7* gene annotates the PTC as a three-nucleotide intron. Further investigation into Ensembl vs. Zoonomia annotations will be undertaken.

Results

The first step in analysing the results of GSTF was to identify 1:1 orthologs. 6,671 orthologs were identified across all twelve species. 9,010 1:1 orthologs present specifically in cats (may or may not be present in outgroups) were identified, whilst 10,558 1:1 orthogroups present specifically in the outgroups were identified.

As a preliminary investigation to confirm the results appeared accurate, the *HOX* genes were investigated. Of the 39 *HOX* genes identified by GSTF, 26 had at least one species missing. The cheetah, puma and tiger were missing from more than ten of these genes, suggesting that the genome annotations for these species are of poorer quality. Therefore, the decision was made to focus on the remaining five felid species for further analyses, and to include the cheetah, puma and tiger once particular genes of interest had been identified.

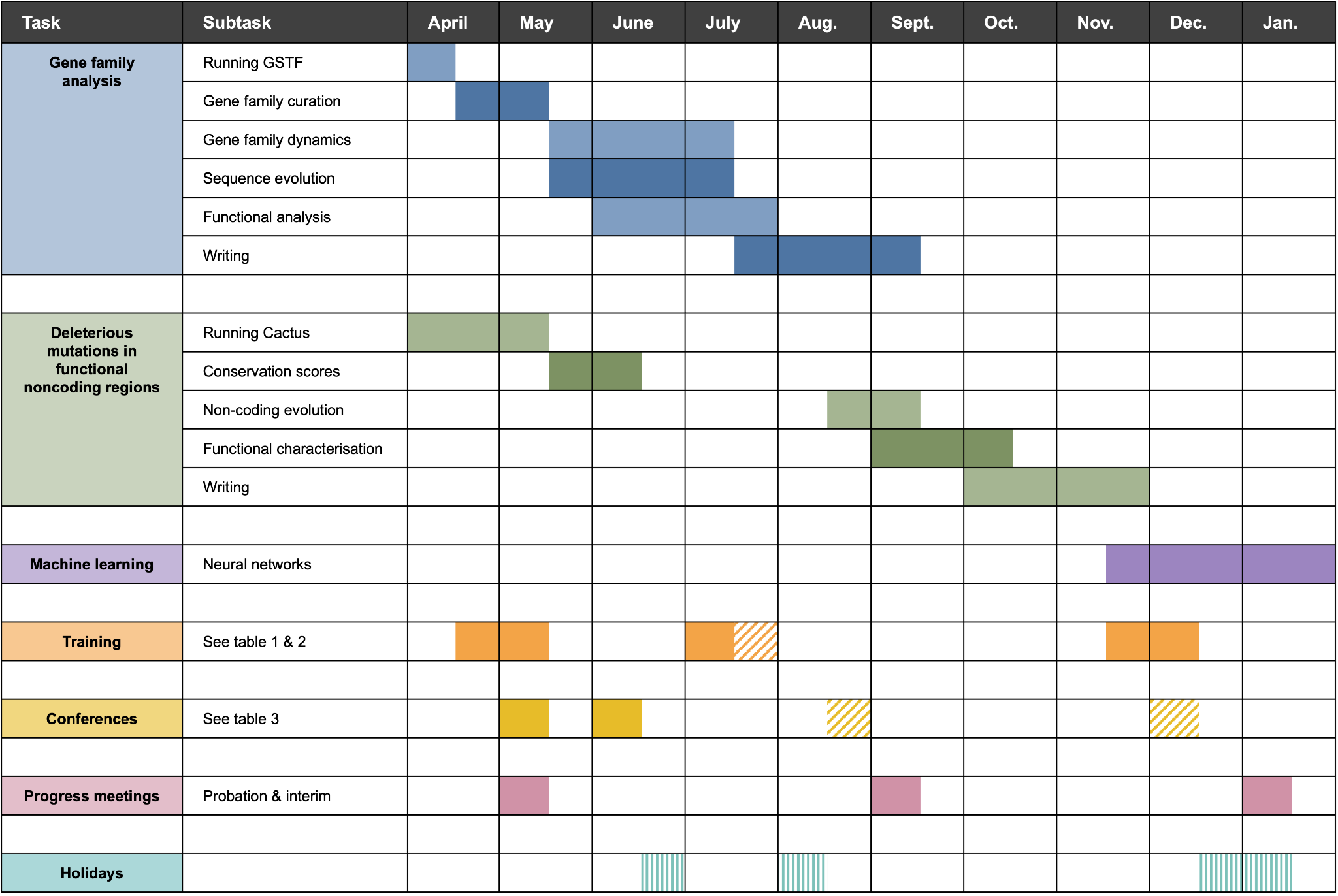
The genes with the greatest number of copies across the species were identified. Many of these contained multiple copies of the gene from the lion but no copies in any other species. When the transcript IDs were searched on Ensembl, they were listed as protein coding “novel lion transcripts”, but the transcripts appear to be transposable elements rather than protein coding genes, suggesting erroneous annotation. Therefore, the lion was also removed from initial analyses.

In addition to missing genes, some *HOX* genes had been split into more than one tree by GSTF. By aligning the sequences of the genes together alongside Ensembl annotations of the genes where available, it was identified that exons present in Ensembl annotations were missing from the Zoonomia annotations, resulting in low sequence identity matches. This suggests that the Zoonomia annotations are not the best quality, so investigation \*\*\*is being undertaken\*\*\* [check?] to solve this issue.

Future work

I will investigate gene family dynamics by identifying gene families with expansions or contractions in specific lineages. Felidae-specific expansions and contractions will be identified first. I will then focus my search on the *Panthera* and *Felis* genera to identify lineage-specific expansions and contractions and broaden the search to investigate *Carnivora*-specific changes. I will also investigate species-specific expansions and contractions. Sequence evolution will subsequently be investigated. I will determine whether expansions or contractions may be linked to demographic events such as bottlenecks or domestication, which have been identified from the literature. Functional analysis will determine the implication of expansions and contractions on the species.

## 8.3. Research plan

****

**Figure 8.** Research plan. Diagonal striped boxes show training and conferences that are not yet confirmed. Holidays are shown in vertical stripes as they do not take up the full time.

## 8.4. Training and conferences

I have completed a range of mandatory and non-mandatory training since beginning my PhD. Mandatory training is listed in Table 2. I have attended non-mandatory training covering both technical and non-scientific topics (Table 3). To develop my personal skills, I have attended sessions on wellbeing, productivity and mental health. To develop my scientific skills, I have attended training on poster-making and figure-making using R, which will be useful throughout my PhD. To gain specific skills required for my project, I attended a course on machine learning using R. I have also registered for training courses in Python and machine learning, which will enable me to further develop the skills required for my project, as well as more broad sessions on networking and project planning. I believe the training I have already completed, as well as the training I plan to complete in the coming months, will give me the skills required for future work. As I progress, I will ensure I identify any further training needs. As well as training, I have also attended multiple virtual conferences (Table 4), which have broadened my understanding of fields related to my project.

**Table 2.** Descriptions provided where necessary

|  |  |  |
| --- | --- | --- |
| **Mandatory training** | | |
| **Training course** | **Date** | **Description** |
| Information Security Training | 01/10/21 | NBI |
| Fire Safety Awareness Training | 04/10/21 | NBI |
| Equality and Diversity | 04/10/21 | NBI |
| Understanding Bullying and Harassment | 04/10/21 | NBI |
| Reviewing Progress and Probation | 04/10/21 | NRPDTP |
| L&D Induction | 05/10/21 | NBI - Introduction to learning and development |
| Computing Orientation Seminar | 05/10/21 | NBI |
| Health and Safety Induction (Generic) | 06/10/21 | NBI |
| Quality Assurance and Intellectual Property | 07/10/21 | NRPDTP - Introduction to IP protocol on research park |
| Research Ethics, Integrity and Regulations | 07/10/21 | NRPDTP |
| Health and Safety (Local & Site Inductions) | 08/10/21 | EI - H&S on site at EI |
| Introduction to IP Management at EI | 12/10/21 | EI - Introduction to EI’s IP protocols |
| Endnote training | 26/10/21 | NBI |
| Chemical Safety for Laboratory Workers | 28/10/21 | NBI - H&S in the laboratory |
| Introduction to HPC | 28/10/21 | NBI - Introduction to HPC and command line |
| HPC Induction | 28/10/21 | EI - EI introduction to HPC |
| Electronic Lab Notebook Training | 29/10/21 | NBI - Introduction to LabArchives |
| Statistical Analysis & Experimental Design | 10/11/21 | NRPDTP - How to remove bias and design a study |
| Introduction to Global Citizenship | 11/11/21 | NRPDTP - How to make research ethical and accessible |
| Mental Health Awareness | 29/11/21 | Norfolk MIND - Mental wellbeing |
| Research & Publication Ethics & Integrity | 1-2/12/21 | NRPDTP |
| Probation Briefing | 13/12/21 | NRPDTP - Preparing for the probationary review |
| Computational Biology & Big Data Science | 20/1/22 | NRPDTP - Intro to bioinformatics and machine learning, useful tools and potential career paths |
| CV & Networking workshop | 10/2/22 | NRPDTP - How to design a CV for different contexts |
| Introduction to Ethics in Publishing | 8/3/22 | NRPDTP |
| Masterclass - Generating & Protecting Knowledge | 10/3/22 | NRPDTP - IP and patenting processes |
| Masterclass - Innovation and Entrepreneurial Skills | 24/3/22 | NRPDTP - Entrepreneurship career paths and recommendations/advice |

**Table 3.**

|  |  |  |
| --- | --- | --- |
| **Non-mandatory training** | | |
| **Training course** | **Date** | **Description** |
| Active bystander training | 19/10/21 | NBI - How to be an active bystander |
| Getting started with R | 3/11/21 | UEA - Basics of R including data and variable types |
| Making figures in R | 19/11/21 | UEA - Make figures using base R and ggplot |
| The productive researcher | 19/1/22 | NBI - Efficiency and productivity in academia |
| Mental health and wellbeing: a journey through postgraduate experiences | 1/2/22 | UEA - How to look after your mental health during a PhD |
| Surviving your PhD - five ways to wellbeing | 7/2/22 | UEA - Key features to take care of for your wellbeing |
| Machine learning: a hands-on introduction | 21 - 25/2/22 | Physalia - different machine learning methods and usage in R (30 contact hours) |
| How to design an excellent scientific poster | 15/3/22 | NBI - Making eye-catching and effective posters for poster sessions/conferences |
| **Future non-mandatory training:** | | |
| LinkedIn | 25/4/22 | UEA Career Central |
| Making figures in Python | 27/4/22 | UEA |
| How to plan your PhD | 28/4/22 | NBI |
| Version control with Git and Github | 3/5/22 | NBI |
| Machine learning for beginners | 6/5/22 | NBI |
| Advanced Python for biologists | 4-15/7/22 | EI (Martin Jones) |
| Computational Molecular Evolution | 18-29/7/22 | Wellcome course - Waitlisted |
| Data exploration with Python | 21/11-2/12/22 | EI (Martin Jones) |

**Table 4.**

|  |  |  |
| --- | --- | --- |
| **Conferences:** | | |
| **Name** | **Dates** | **Organisers/location** |
| UK Conference of Bioinformatics and Computational Biology | 28-30/09/21 | EI (Virtual) |
| Biodiversity Genomics | 27/09-01/10/21 | Darwin Tree of Life (Virtual) |
| PopGroup 55 | 5-7/1/22 | NRP (Virtual) |
| PAGBio Day | 15/2/22 | PacBio (Virtual) |
| **Future conferences:** | | |
| Biology of Genomes | 10-14/5/22 | Cold Spring Harbor (Virtual) |
| Population, Evolutionary and Quantitative Genetics Conference | 7-10/6/22 | Genetics Society of America (Virtual) |
| European Society for Evolutionary Biology | 14-19/8/22 | ESEB Prague (Not registered) |
| European Conservation Genetics Meeting | 31/8-2/9/22 | Edinburgh (Not registered) |
| Conservation Genomics at the Population Level | 30/11-2/12/22 | Wellcome (Not registered) |

# 9. Acknowledgements

I would like to thank my supervisor, Wilfried Haerty, for his guidance and feedback during my project. I would also like to thank Anil Thanki for helping me run GeneSeqToFamily and answering my endless questions and to Will Nash for helping me to get Cactus running despite all its many challenges. Thank you to the whole Haerty group for their advice and ideas for my project.

BBSRC/DTP?

# 10. References

[Aken, B.L. *et al.* (2016) ‘The Ensembl Gene Annotation System’. *Database: The Journal of Biological Databases and Curation*, 2016. DOI:](http://paperpile.com/b/HQttr3/Tl37) [10.1093/database/baw093](http://dx.doi.org/10.1093/database/baw093)[.](http://paperpile.com/b/HQttr3/Tl37)

[Alasaad, S. *et al.* (2011) ‘Siberian Tiger’s Recent Population Bottleneck in the Russian Far East Revealed by Microsatellite Markers’. *Mammalian Biology = Zeitschrift Fur Saugetierkunde*, 76(6), pp. 722–726.](http://paperpile.com/b/HQttr3/9151)

[Alhaddad, H. *et al.* (2013) ‘Extent of Linkage Disequilibrium in the Domestic Cat, Felis Silvestris Catus, and Its Breeds’. *PloS One*, 8(1), p. e53537.](http://paperpile.com/b/HQttr3/53vG)

[Antunes, A. *et al.* (2008) ‘The Evolutionary Dynamics of the Lion Panthera Leo Revealed by Host and Viral Population Genomics’. *PLoS Genetics*, 4(11), p. e1000251.](http://paperpile.com/b/HQttr3/5PPq)

[Armstrong, E.E. *et al.* (2020) ‘Long Live the King: Chromosome-Level Assembly of the Lion (Panthera Leo) Using Linked-Read, Hi-C, and Long-Read Data’. *BMC Biology*, 18(1), p. 3.](http://paperpile.com/b/HQttr3/lLXE)

[Armstrong, E.E. *et al.* (2021) ‘Recent Evolutionary History of Tigers Highlights Contrasting Roles of Genetic Drift and Selection’. *Molecular Biology and Evolution*, 38(6), pp. 2366–2379.](http://paperpile.com/b/HQttr3/T9Lh)

[Armstrong, J. *et al.* (2020) ‘Progressive Cactus Is a Multiple-Genome Aligner for the Thousand-Genome Era’. *Nature*, 587(7833), pp. 246–251.](http://paperpile.com/b/HQttr3/Pkqv)

[Bahcall, O. (2013) *FunSeq for Cancer Genomics*. *Nature Publishing Group US*. DOI:](http://paperpile.com/b/HQttr3/5AsT) [10.1038/ng.2819](http://dx.doi.org/10.1038/ng.2819)[.](http://paperpile.com/b/HQttr3/5AsT)

[Banks, S.C. *et al.* (2013) ‘How Does Ecological Disturbance Influence Genetic Diversity?’ *Trends in Ecology & Evolution*, 28(11), pp. 670–679.](http://paperpile.com/b/HQttr3/tBsW)

[Beiki, H. *et al.* (2019) ‘Improved Annotation of the Domestic Pig Genome through Integration of Iso-Seq and RNA-Seq Data’. *BMC Genomics*, 20(1), p. 344.](http://paperpile.com/b/HQttr3/GK7h)

[Belton, J.-M. *et al.* (2012) ‘Hi-C: A Comprehensive Technique to Capture the Conformation of Genomes’. *Methods* , 58(3), pp. 268–276.](http://paperpile.com/b/HQttr3/IVFB)

[Bijlsma, R. and Loeschcke, V. (2012) ‘Genetic Erosion Impedes Adaptive Responses to Stressful Environments’. *Evolutionary Applications*, 5(2), pp. 117–129.](http://paperpile.com/b/HQttr3/G24Q)

[Björnerfeldt, S., Webster, M.T. and Vilà, C. (2006) ‘Relaxation of Selective Constraint on Dog Mitochondrial DNA Following Domestication’. *Genome Research*, 16(8), pp. 990–994.](http://paperpile.com/b/HQttr3/p9kK)

[Blanchette, M. *et al.* (2004) ‘Aligning Multiple Genomic Sequences with the Threaded Blockset Aligner’. *Genome Research*, 14(4), pp. 708–715.](http://paperpile.com/b/HQttr3/U8bG)

[Boeva, V. (2016) ‘Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells’. *Frontiers in Genetics*, 7, p. 24.](http://paperpile.com/b/HQttr3/Dhgj)

[Bosse, M. *et al.* (2019) ‘Deleterious Alleles in the Context of Domestication, Inbreeding, and Selection’. *Evolutionary Applications*, 12(1), pp. 6–17.](http://paperpile.com/b/HQttr3/H5eu)

[Broquet, T. *et al.* (2010) ‘Genetic Bottlenecks Driven by Population Disconnection’. *Conservation Biology: The Journal of the Society for Conservation Biology*, 24(6), pp. 1596–1605.](http://paperpile.com/b/HQttr3/IbOv)

[Buckley, R.M. *et al.* (2020) ‘A New Domestic Cat Genome Assembly Based on Long Sequence Reads Empowers Feline Genomic Medicine and Identifies a Novel Gene for Dwarfism’. *PLoS Genetics*, 16(10), p. e1008926.](http://paperpile.com/b/HQttr3/JSpl)

[Burge, C. and Karlin, S. (1997) ‘Prediction of Complete Gene Structures in Human Genomic DNA’. *Journal of Molecular Biology*, 268(1), pp. 78–94.](http://paperpile.com/b/HQttr3/5v2K)

[Cantarel, B.L. *et al.* (2008) ‘MAKER: An Easy-to-Use Annotation Pipeline Designed for Emerging Model Organism Genomes’. *Genome Research*, 18(1), pp. 188–196.](http://paperpile.com/b/HQttr3/BT72)

[Carroll, S.B. (2000) ‘Endless Forms: The Evolution Minireview of Gene Regulation and Morphological Diversity’. *Cell*, 101, pp. 577–580.](http://paperpile.com/b/HQttr3/lZRZ)

[Charlesworth, B. and Charlesworth, D. (1999) ‘The Genetic Basis of Inbreeding Depression’. *Genetical Research*, 74(3), pp. 329–340.](http://paperpile.com/b/HQttr3/zYAK)

[Charlesworth, D. (2003) ‘Effects of Inbreeding on the Genetic Diversity of Populations’. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 358(1434), pp. 1051–1070.](http://paperpile.com/b/HQttr3/pztn)

[Charlesworth, D.H. and Charlesworth, B. (1987) ‘Inbreeding Depression and Its Evolutionary Consequences’. *Annual Review of Ecology and Systematics*, 18(1), pp. 237–268.](http://paperpile.com/b/HQttr3/eCj3)

[Charlesworth, D. and Willis, J.H. (2009) ‘The Genetics of Inbreeding Depression’. *Nature Reviews. Genetics*, 10(11), pp. 783–796.](http://paperpile.com/b/HQttr3/26Nq)

[Chen, L. and Wang, Y. (2022) ‘Exploiting Deep Transfer Learning for the Prediction of Functional Noncoding Variants Using Genomic Sequence’. *bioRxiv*, p. 2022.03.19.484983. DOI:](http://paperpile.com/b/HQttr3/7pI2) [10.1101/2022.03.19.484983](http://dx.doi.org/10.1101/2022.03.19.484983)[.](http://paperpile.com/b/HQttr3/7pI2)

[Cho, Y.S. *et al.* (2013) ‘The Tiger Genome and Comparative Analysis with Lion and Snow Leopard Genomes’. *Nature Communications*, 4, p. 2433.](http://paperpile.com/b/HQttr3/dkXg)

[Church, D.M. (2022) ‘A Next-Generation Human Genome Sequence’. *Science*, 376(6588), pp. 34–35.](http://paperpile.com/b/HQttr3/ItP0)

[Clifford, D.L. *et al.* (2007) ‘Using Pregnancy Rates and Perinatal Mortality to Evaluate the Success of Recovery Strategies for Endangered Island Foxes’. *Animal Conservation*, 10(4), pp. 442–451.](http://paperpile.com/b/HQttr3/ZBBb)

[Cochran, K. *et al.* (2022) ‘Domain Adaptive Neural Networks Improve Cross-Species Prediction of Transcription Factor Binding’. *Genome Research*. DOI:](http://paperpile.com/b/HQttr3/1cI3) [10.1101/gr.275394.121](http://dx.doi.org/10.1101/gr.275394.121)[.](http://paperpile.com/b/HQttr3/1cI3)

[Crow, J.F. and Kimura, M. (1970) *An Introduction to Population Genetics Theory*. Burgess Publishing Company.](http://paperpile.com/b/HQttr3/vK5t)

[Cruz, F., Vilà, C. and Webster, M.T. (2008) ‘The Legacy of Domestication: Accumulation of Deleterious Mutations in the Dog Genome’. *Molecular Biology and Evolution*, 25(11), pp. 2331–2336.](http://paperpile.com/b/HQttr3/tKxR)

[Culver, M. *et al.* (2008) ‘Estimation of the Bottleneck Size in Florida Panthers’. *Animal Conservation*, 11(2), pp. 104–110.](http://paperpile.com/b/HQttr3/aIHd)

[Culver, M. *et al.* (2000) ‘Genomic Ancestry of the American Puma (Puma Concolor)’. *The Journal of Heredity*, 91(3), pp. 186–197.](http://paperpile.com/b/HQttr3/8KCa)

[Curik, I., Ferenčaković, M. and Sölkner, J. (2014) ‘Inbreeding and Runs of Homozygosity: A Possible Solution to an Old Problem’. *Livestock Science*, 166, pp. 26–34.](http://paperpile.com/b/HQttr3/f7A7)

[Darwin Tree of Life Project Consortium. (2022) ‘Sequence Locally, Think Globally: The Darwin Tree of Life Project’. *Proceedings of the National Academy of Sciences of the United States of America*, 119(4). DOI:](http://paperpile.com/b/HQttr3/52vP) [10.1073/pnas.2115642118](http://dx.doi.org/10.1073/pnas.2115642118)[.](http://paperpile.com/b/HQttr3/52vP)

[van Dijk, E.L. *et al.* (2018) ‘The Third Revolution in Sequencing Technology’. *Trends in Genetics: TIG*, 34(9), pp. 666–681.](http://paperpile.com/b/HQttr3/xxaa)

[Dobrynin, P. *et al.* (2015) ‘Genomic Legacy of the African Cheetah, Acinonyx Jubatus’. *Genome Biology*, 16, p. 277.](http://paperpile.com/b/HQttr3/VeeI)

[Driscoll, C.A. *et al.* (2002) ‘Genomic Microsatellites as Evolutionary Chronometers: A Test in Wild Cats’. *Genome Research*, 12(3), pp. 414–423.](http://paperpile.com/b/HQttr3/dn2s)

[Driscoll, C.A. *et al.* (2007) ‘The Near Eastern Origin of Cat Domestication’. *Science*, 317(5837), pp. 519–523.](http://paperpile.com/b/HQttr3/vKgI)

[Drosophila 12 Genomes Consortium. *et al.* (2007) ‘Evolution of Genes and Genomes on the Drosophila Phylogeny’. *Nature*, 450(7167), pp. 203–218.](http://paperpile.com/b/HQttr3/kJg6)

[Eizirik, E. *et al.* (2001) ‘Phylogeography, Population History and Conservation Genetics of Jaguars (Panthera Onca, Mammalia, Felidae)’. *Molecular Ecology*, 10(1), pp. 65–79.](http://paperpile.com/b/HQttr3/4MnI)

[ENCODE Project Consortium. (2012) ‘An Integrated Encyclopedia of DNA Elements in the Human Genome’. *Nature*, 489(7414), pp. 57–74.](http://paperpile.com/b/HQttr3/wJ8u)

[Eyre-Walker, A. *et al.* (1998) ‘Investigation of the Bottleneck Leading to the Domestication of Maize’. *Proceedings of the National Academy of Sciences of the United States of America*, 95(8), pp. 4441–4446.](http://paperpile.com/b/HQttr3/PORF)

[Fleischmann, R.D. *et al.* (1995) ‘Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd’. *Science*, 269(5223), pp. 496–512.](http://paperpile.com/b/HQttr3/6QrD)

[Frazer, J. *et al.* (2021) ‘Disease Variant Prediction with Deep Generative Models of Evolutionary Data’. *Nature*. DOI:](http://paperpile.com/b/HQttr3/fYdu) [10.1038/s41586-021-04043-8](http://dx.doi.org/10.1038/s41586-021-04043-8)[.](http://paperpile.com/b/HQttr3/fYdu)

[Galtier, N., Depaulis, F. and Barton, N.H. (2000) ‘Detecting Bottlenecks and Selective Sweeps from DNA Sequence Polymorphism’. *Genetics*, 155(2), pp. 981–987.](http://paperpile.com/b/HQttr3/VJr2)

[Gilbert, D.A. *et al.* (1991) ‘Analytical DNA Fingerprinting in Lions: Parentage, Genetic Diversity, and Kinship’. *The Journal of Heredity*, 82(5), pp. 378–386.](http://paperpile.com/b/HQttr3/RJVK)

[Gu, Z. *et al.* (2005) ‘Elevated Evolutionary Rates in the Laboratory Strain of Saccharomyces Cerevisiae’. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4), pp. 1092–1097.](http://paperpile.com/b/HQttr3/xQjl)

[Harris, R.S. (2007) *Improved Pairwise Alignment of Genomic DNA*. Pennsylvania State University.](http://paperpile.com/b/HQttr3/j0ox)

[Hedrick, P.W. (1994) ‘Purging Inbreeding Depression and the Probability of Extinction: Full-Sib Mating’. *Heredity*, 73 (Pt 4), pp. 363–372.](http://paperpile.com/b/HQttr3/ufIy)

[Hedrick, P.W. and Miller, P.S. (1992) ‘Conservation Genetics: Techniques and Fundamentals’. *Ecological Applications: A Publication of the Ecological Society of America*, 2(1), pp. 30–46.](http://paperpile.com/b/HQttr3/illR)

[Henry, P. *et al.* (2009) ‘In Situ Population Structure and Ex Situ Representation of the Endangered Amur Tiger’. *Molecular Ecology*, 18(15), pp. 3173–3184.](http://paperpile.com/b/HQttr3/94Qh)

[Hickey, G. *et al.* (2013) ‘HAL: A Hierarchical Format for Storing and Analyzing Multiple Genome Alignments’. *Bioinformatics* , 29(10), pp. 1341–1342.](http://paperpile.com/b/HQttr3/RvVV)

[Hindorff, L.A. *et al.* (2009) ‘Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits’. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), pp. 9362–9367.](http://paperpile.com/b/HQttr3/SQeB)

[Hoff, K.J. *et al.* (2016) ‘BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS’. *Bioinformatics* , 32(5), pp. 767–769.](http://paperpile.com/b/HQttr3/YyKR)

[Hoff, K.J. and Stanke, M. (2019) ‘Predicting Genes in Single Genomes with AUGUSTUS’. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, 65(1), p. e57.](http://paperpile.com/b/HQttr3/CEuX)

[Huang, Y.-F., Gulko, B. and Siepel, A. (2017) ‘Fast, Scalable Prediction of Deleterious Noncoding Variants from Functional and Population Genomic Data’. *Nature Genetics*, 49(4), pp. 618–624.](http://paperpile.com/b/HQttr3/5oMi)

[Hu, T. *et al.* (2021) ‘Next-Generation Sequencing Technologies: An Overview’. *Human Immunology*, 82(11), pp. 801–811.](http://paperpile.com/b/HQttr3/mN2p)

[Jacobson, A.P. *et al.* (2016) ‘Leopard (Panthera Pardus) Status, Distribution, and the Research Efforts across Its Range’. *PeerJ*, 4, p. e1974.](http://paperpile.com/b/HQttr3/Rm8I)

[Jain, M. *et al.* (2016) ‘Erratum to: The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community’. *Genome Biology*, 17(1), p. 256.](http://paperpile.com/b/HQttr3/ajNg)

[Johnson, W.E. *et al.* (2010) ‘Genetic Restoration of the Florida Panther’. *Science*, 329(5999), pp. 1641–1645.](http://paperpile.com/b/HQttr3/joOu)

[Kalinowski, S.T., Hedrick, P.W. and Miller, P.S. (2000) ‘Inbreeding Depression in the Speke’s Gazelle Captive Breeding Program’. *Conservation Biology: The Journal of the Society for Conservation Biology*, 14(5), pp. 1375–1384.](http://paperpile.com/b/HQttr3/TFOm)

[Kaplanov, L.G. (2005) ‘The Tiger in the Sikhote-Alin’. In Kaplanov, L.G. (ed.) *Tiger. Deer. Moose*. pp. 1–71.](http://paperpile.com/b/HQttr3/tg8D)

[Kardos, M. *et al.* (2018) ‘Genomic Consequences of Intensive Inbreeding in an Isolated Wolf Population’. *Nature Ecology & Evolution*, 2(1), pp. 124–131.](http://paperpile.com/b/HQttr3/C109)

[Kim, S. *et al.* (2016) ‘Comparison of Carnivore, Omnivore, and Herbivore Mammalian Genomes with a New Leopard Assembly’. *Genome Biology*, 17(1), p. 211.](http://paperpile.com/b/HQttr3/L7br)

[Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press.](http://paperpile.com/b/HQttr3/9Ze7)

[Kircher, M. *et al.* (2014) ‘A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants’. *Nature Genetics*, 46(3), pp. 310–315.](http://paperpile.com/b/HQttr3/jZKY)

[Koch, L. (2020) ‘Exploring Human Genomic Diversity with gnomAD’. *Nature Reviews. Genetics*, 21(8), p. 448.](http://paperpile.com/b/HQttr3/ldWs)

[Koenig, D. *et al.* (2013) ‘Comparative Transcriptomics Reveals Patterns of Selection in Domesticated and Wild Tomato’. *Proceedings of the National Academy of Sciences of the United States of America*, 110(28), pp. E2655–62.](http://paperpile.com/b/HQttr3/EEMn)

[Koster, J. and Rahmann, S. (2012) DOI:](http://paperpile.com/b/HQttr3/VYrV) [10.1093/bioinformatics/bts480](http://dx.doi.org/10.1093/bioinformatics/bts480)[.](http://paperpile.com/b/HQttr3/VYrV)

[Kryukov, G.V., Pennacchio, L.A. and Sunyaev, S.R. (2007) ‘Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies’. *American Journal of Human Genetics*, 80(4), pp. 727–739.](http://paperpile.com/b/HQttr3/wpPj)

[Kuhn, R.M., Haussler, D. and Kent, W.J. (2013) ‘The UCSC Genome Browser and Associated Tools’. *Briefings in Bioinformatics*, 14(2), pp. 144–161.](http://paperpile.com/b/HQttr3/7dfz)

[Lande, R. (1994) ‘RISK OF POPULATION EXTINCTION FROM FIXATION OF NEW DELETERIOUS MUTATIONS’. *Evolution; International Journal of Organic Evolution*, 48(5), pp. 1460–1469.](http://paperpile.com/b/HQttr3/b5I0)

[Lande, R. and Barrowclough, G.F. (1987) ‘Effective Population Size, Genetic Variation, and Their Use in Population Management’. In *Viable Populations for Conservation*. Cambridge University Press, pp. 87–124.](http://paperpile.com/b/HQttr3/1ry4)

[Lander, E.S. *et al.* (2001) ‘Initial Sequencing and Analysis of the Human Genome’. *Nature*, 409(6822), pp. 860–921.](http://paperpile.com/b/HQttr3/s61q)

[Lappalainen, T. *et al.* (2019) ‘Genomic Analysis in the Age of Human Genome Sequencing’. *Cell*, 177(1), pp. 70–84.](http://paperpile.com/b/HQttr3/Ctx9)

[LeCun, Y., Bengio, Y. and Hinton, G. (2015) ‘Deep Learning’. *Nature*, 521(7553), pp. 436–444.](http://paperpile.com/b/HQttr3/scQl)

[Lewin, H.A. *et al.* (2018) ‘Earth BioGenome Project: Sequencing Life for the Future of Life’. *Proceedings of the National Academy of Sciences of the United States of America*, 115(17), pp. 4325–4333.](http://paperpile.com/b/HQttr3/qeHO)

[Li, H. (2006) *Hcluster\_sg*. Github Available at:](http://paperpile.com/b/HQttr3/uQ02) <https://github.com/douglasgscofield/hcluster> [(Accessed: 22 April 2022).](http://paperpile.com/b/HQttr3/uQ02)

[Lindblad-Toh, K. *et al.* (2011) ‘A High-Resolution Map of Human Evolutionary Constraint Using 29 Mammals’. *Nature*, 478(7370), pp. 476–482.](http://paperpile.com/b/HQttr3/L4Qy)

[Li, W. *et al.* (2021) ‘RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline Reach with Protein Family Model Curation’. *Nucleic Acids Research*, 49(D1), pp. D1020–D1028.](http://paperpile.com/b/HQttr3/S96K)

[Logsdon, G.A., Vollger, M.R. and Eichler, E.E. (2020) ‘Long-Read Human Genome Sequencing and Its Applications’. *Nature Reviews. Genetics*, 21(10), pp. 597–614.](http://paperpile.com/b/HQttr3/OXlH)

[López-Cortegano, E., Moreno, E. and García-Dorado, A. (2021) ‘Genetic Purging in Captive Endangered Ungulates with Extremely Low Effective Population Sizes’. *Heredity*, 127(5), pp. 433–442.](http://paperpile.com/b/HQttr3/HJ9c)

[Lorenzana, G. *et al.* (2020) ‘Large-Scale Assessment of Genetic Diversity and Population Connectivity of Amazonian Jaguars (Panthera Onca) Provides a Baseline for Their Conservation and Monitoring in Fragmented Landscapes’. *Biological Conservation*, 242, p. 108417.](http://paperpile.com/b/HQttr3/BuJc)

[Luikart, G. *et al.* (1998) ‘Usefulness of Molecular Markers for Detecting Population Bottlenecks via Monitoring Genetic Change’. *Molecular Ecology*, 7(8), pp. 963–974.](http://paperpile.com/b/HQttr3/v0Vc)

[Lu, J. *et al.* (2006) ‘The Accumulation of Deleterious Mutations in Rice Genomes: A Hypothesis on the Cost of Domestication’. *Trends in Genetics: TIG*, 22(3), pp. 126–131.](http://paperpile.com/b/HQttr3/3XeP)

[Lynch, M. *et al.* (1993) ‘The Mutational Meltdown in Asexual Populations’. *The Journal of Heredity*, 84(5), pp. 339–344.](http://paperpile.com/b/HQttr3/C3HS)

[Lynch, M. and Gabriel, W. (1990) ‘MUTATION LOAD AND THE SURVIVAL OF SMALL POPULATIONS’. *Evolution; International Journal of Organic Evolution*, 44(7), pp. 1725–1737.](http://paperpile.com/b/HQttr3/LI2Q)

[Mäki, K. *et al.* (2001) ‘Population Structure, Inbreeding Trend and Their Association with Hip and Elbow Dysplasia in Dogs’. *Animal Science* , 73(2), pp. 217–228.](http://paperpile.com/b/HQttr3/YOV7)

[Masel, J. (2011) ‘Genetic Drift’. *Current Biology: CB*, 21(20), pp. R837–8.](http://paperpile.com/b/HQttr3/UAyi)

[Maston, G.A., Evans, S.K. and Green, M.R. (2006) ‘Transcriptional Regulatory Elements in the Human Genome’. *Annual Review of Genomics and Human Genetics*, 7, pp. 29–59.](http://paperpile.com/b/HQttr3/qxCG)

[Matte, E.M. *et al.* (2013) ‘Molecular Evidence for a Recent Demographic Expansion in the Puma (Puma Concolor) (Mammalia, Felidae)’. *Genetics and Molecular Biology*, 36(4), pp. 586–597.](http://paperpile.com/b/HQttr3/JxtB)

[Mattucci, F. *et al.* (2019) ‘How to Spot a Black-Footed Cat? Successful Application of Cross-Species Markers to Identify Captive-Bred Individuals from Non-Invasive Genetic Sampling’. *Mammal Research*, 64(1), pp. 133–145.](http://paperpile.com/b/HQttr3/KBkV)

[Meader, S., Ponting, C.P. and Lunter, G. (2010) ‘Massive Turnover of Functional Sequence in Human and Other Mammalian Genomes’. *Genome Research*, 20(10), pp. 1335–1343.](http://paperpile.com/b/HQttr3/WdHb)

[Mellersh, C. (2008) ‘Give a Dog a Genome’. *Veterinary Journal* , 178(1), pp. 46–52.](http://paperpile.com/b/HQttr3/fojQ)

[Melnikov, A. *et al.* (2014) ‘Massively Parallel Reporter Assays in Cultured Mammalian Cells’. *Journal of Visualized Experiments: JoVE*, (90). DOI:](http://paperpile.com/b/HQttr3/2hr2) [10.3791/51719](http://dx.doi.org/10.3791/51719)[.](http://paperpile.com/b/HQttr3/2hr2)

[Menotti-Raymond, M. *et al.* (2008) ‘Patterns of Molecular Genetic Variation among Cat Breeds’. *Genomics*, 91(1), pp. 1–11.](http://paperpile.com/b/HQttr3/yADr)

[Menotti-Raymond, M. and O’Brien, S.J. (1993) ‘Dating the Genetic Bottleneck of the African Cheetah’. *Proceedings of the National Academy of Sciences of the United States of America*, 90(8), pp. 3172–3176.](http://paperpile.com/b/HQttr3/9b8F)

[Miller, W. *et al.* (2004) ‘Comparative Genomics’. *Annual Review of Genomics and Human Genetics*, 5, pp. 15–56.](http://paperpile.com/b/HQttr3/aMsB)

[Miquelle, D.G. *et al.* (2010) *A Program for Reintroduction of the Far Eastern Leopard into Southern Sikhote-Alin, Primorsky Krai, Russian Far East*.](http://paperpile.com/b/HQttr3/Wqy1)

[Miquelle, D.G., Arzhanova, T.D. and Solkin, V.A. (1996) ‘Recovery Plan for Conservation of the Far Eastern Leopard’. In *Results of an International Conference Held in Vladivostok*. USAID/WWF/EPT. Available at:](http://paperpile.com/b/HQttr3/YcFS) <http://www.catsg.org/fileadmin/filesharing/3.Conservation_Center/3.4._Strategies___Action_Plans/leopard_Asia/Miquelle_et_al_1996_Recovery_plan_for_conservation_of_the_far_eastern_leopard.pdf>[.](http://paperpile.com/b/HQttr3/YcFS)

[Montague, M.J. *et al.* (2014) ‘Comparative Analysis of the Domestic Cat Genome Reveals Genetic Signatures Underlying Feline Biology and Domestication’. *Proceedings of the National Academy of Sciences of the United States of America*, 111(48), pp. 17230–17235.](http://paperpile.com/b/HQttr3/ojjx)

[Mouse Genome Sequencing Consortium. *et al.* (2002) ‘Initial Sequencing and Comparative Analysis of the Mouse Genome’. *Nature*, 420(6915), pp. 520–562.](http://paperpile.com/b/HQttr3/CjpC)

[Mukai, T. *et al.* (1972) ‘Mutation Rate and Dominance of Genes Affecting Viability in Drosophila Melanogaster’. *Genetics*, 72(2), pp. 335–355.](http://paperpile.com/b/HQttr3/t7Ws)

[Muller, H.J. (1964) ‘THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE’. *Mutation Research*, 106, pp. 2–9.](http://paperpile.com/b/HQttr3/GsGm)

[Nabholz, B. *et al.* (2014) ‘Transcriptome Population Genomics Reveals Severe Bottleneck and Domestication Cost in the African Rice (Oryza Glaberrima)’. *Molecular Ecology*, 23(9), pp. 2210–2227.](http://paperpile.com/b/HQttr3/yvjg)

[Ng, P.C. and Henikoff, S. (2003) ‘SIFT: Predicting Amino Acid Changes That Affect Protein Function’. *Nucleic Acids Research*, 31(13), pp. 3812–3814.](http://paperpile.com/b/HQttr3/2io9)

[Notredame, C., Higgins, D.G. and Heringa, J. (2000) ‘T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment’. *Journal of Molecular Biology*, 302(1), pp. 205–217.](http://paperpile.com/b/HQttr3/7Zfg)

[Nurk, S. *et al.* (2022) ‘The Complete Sequence of a Human Genome’. *Science*, 376(6588), pp. 44–53.](http://paperpile.com/b/HQttr3/MV13)

[O’Brien, S.J. (1994a) ‘A Role for Molecular Genetics in Biological Conservation’. *Proceedings of the National Academy of Sciences of the United States of America*, 91(13), pp. 5748–5755.](http://paperpile.com/b/HQttr3/bu0F)

[O’Brien, S.J. (1994b) ‘Genetic and Phylogenetic Analyses of Endangered Species’. *Annual Review of Genetics*, 28, pp. 467–489.](http://paperpile.com/b/HQttr3/pRGG)

[O’Brien, S.J. *et al.* (1985) ‘Genetic Basis for Species Vulnerability in the Cheetah’. *Science*, 227(4693), pp. 1428–1434.](http://paperpile.com/b/HQttr3/aIKH)

[O’Brien, S.J. and Johnson, W.E. (2005) ‘Big Cat Genomics’. *Annual Review of Genomics and Human Genetics*, 6, pp. 407–429.](http://paperpile.com/b/HQttr3/wdtK)

[Ohta, T. (1992) ‘The Nearly Neutral Theory of Molecular Evolution’. *Annual Review of Ecology and Systematics*, 23(1), pp. 263–286.](http://paperpile.com/b/HQttr3/sgSi)

[Ostrander, E.A. and Kruglyak, L. (2000) ‘Unleashing the Canine Genome’. *Genome Research*, 10(9), pp. 1271–1274.](http://paperpile.com/b/HQttr3/8keu)

[Paijmans, J.L.A. *et al.* (2021) ‘African and Asian Leopards Are Highly Differentiated at the Genomic Level’. *Current Biology: CB*, 31(9), pp. 1872–1882.e5.](http://paperpile.com/b/HQttr3/pJBu)

[Palmer, J. (2016) *Funannotate: Eukaryotic Genome Annotation Pipeline*. Github Available at:](http://paperpile.com/b/HQttr3/HXwx) <https://github.com/nextgenusfs/funannotate> [(Accessed: 19 April 2022).](http://paperpile.com/b/HQttr3/HXwx)

[Pečnerová, P. *et al.* (2021) ‘High Genetic Diversity and Low Differentiation Reflect the Ecological Versatility of the African Leopard’. *Current Biology: CB*, 31(9), pp. 1862–1871.e5.](http://paperpile.com/b/HQttr3/VXM7)

[Pertea, G. and Pertea, M. (2020) ‘GFF Utilities: GffRead and GffCompare’. *F1000Research*, 9. DOI:](http://paperpile.com/b/HQttr3/pxC8) [10.12688/f1000research.23297.2](http://dx.doi.org/10.12688/f1000research.23297.2)[.](http://paperpile.com/b/HQttr3/pxC8)

[Pimm, S.L. *et al.* (2014) ‘The Biodiversity of Species and Their Rates of Extinction, Distribution, and Protection’. *Science*, 344(6187), p. 1246752.](http://paperpile.com/b/HQttr3/pGyj)

[Pinoli, P. *et al.* (2019) ‘Deleterious Impact of Mutational Processes on Transcription Factor Binding Sites in Human Cancer’. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. pp. 188–193.](http://paperpile.com/b/HQttr3/JcU4)

[Piras, P. *et al.* (2018) ‘Evolution of the Sabertooth Mandible: A Deadly Ecomorphological Specialization’. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 496, pp. 166–174.](http://paperpile.com/b/HQttr3/eIS2)

[Pollard, K.S. *et al.* (2010) ‘Detection of Nonneutral Substitution Rates on Mammalian Phylogenies’. *Genome Research*, 20(1), pp. 110–121.](http://paperpile.com/b/HQttr3/yQTz)

[Quang, D., Chen, Y. and Xie, X. (2015) ‘DANN: A Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants’. *Bioinformatics* , 31(5), pp. 761–763.](http://paperpile.com/b/HQttr3/GvUB)

[Quinodoz, M. *et al.* (2022) ‘Analysis of Missense Variants in the Human Genome Reveals Widespread Gene-Specific Clustering and Improves Prediction of Pathogenicity’. *American Journal of Human Genetics*, 0(0). DOI:](http://paperpile.com/b/HQttr3/fquX) [10.1016/j.ajhg.2022.01.006](http://dx.doi.org/10.1016/j.ajhg.2022.01.006)[.](http://paperpile.com/b/HQttr3/fquX)

[Raes, J. and Van de Peer, Y. (2005) ‘Functional Divergence of Proteins through Frameshift Mutations’. *Trends in Genetics: TIG*, 21(8), pp. 428–431.](http://paperpile.com/b/HQttr3/HnZg)

[Rands, C.M. *et al.* (2014) ‘8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage’. *PLoS Genetics*, 10(7), p. e1004525.](http://paperpile.com/b/HQttr3/b8If)

[Rhoads, A. and Au, K.F. (2015) ‘PacBio Sequencing and Its Applications’. *Genomics, Proteomics & Bioinformatics*, 13(5), pp. 278–289.](http://paperpile.com/b/HQttr3/opft)

[Ricaño-Ponce, I. and Wijmenga, C. (2013) ‘Mapping of Immune-Mediated Disease Genes’. *Annual Review of Genomics and Human Genetics*, 14, pp. 325–353.](http://paperpile.com/b/HQttr3/En9K)

[Roelke, M.E., Martenson, J.S. and O’Brien, S.J. (1993) ‘The Consequences of Demographic Reduction and Genetic Depletion in the Endangered Florida Panther’. *Current Biology: CB*, 3(6), pp. 340–350.](http://paperpile.com/b/HQttr3/h72c)

[Roques, S. *et al.* (2016) ‘Effects of Habitat Deterioration on the Population Genetics and Conservation of the Jaguar’. *Conservation Genetics* , 17(1), pp. 125–139.](http://paperpile.com/b/HQttr3/9k0X)

[Russello, M.A. *et al.* (2004) ‘Potential Genetic Consequences of a Recent Bottleneck in the Amur Tiger of’. *Conservation Genetics* , 5(5), pp. 707–713.](http://paperpile.com/b/HQttr3/lMc7)

[Salzberg, S.L. (2019) ‘Next-Generation Genome Annotation: We Still Struggle to Get It Right’. *Genome Biology*, 20(1), p. 92.](http://paperpile.com/b/HQttr3/6cTn)

[Samaha, G. *et al.* (2021) ‘Exploiting Genomic Synteny in Felidae: Cross-Species Genome Alignments and SNV Discovery Can Aid Conservation Management’. *BMC Genomics*, 22(1), p. 601.](http://paperpile.com/b/HQttr3/gg7F)

[Samaha, G.A. (2021) *Advances in Felid Genetics and Genomics*. [Doctor of Philosophy Thesis]. The University of Sydney. Available at:](http://paperpile.com/b/HQttr3/KJnw) <https://ses.library.usyd.edu.au/handle/2123/25725> [(Accessed: 26 January 2022).](http://paperpile.com/b/HQttr3/KJnw)

[Sandler, R.L., Moses, L. and Wisely, S.M. (2021) ‘An Ethical Analysis of Cloning for Genetic Rescue: Case Study of the Black-Footed Ferret’. *Biological Conservation*, 257, p. 109118.](http://paperpile.com/b/HQttr3/eRil)

[Santymire, R.M. *et al.* (2006) ‘Sperm Viability in the Black-Footed Ferret (Mustela Nigripes) Is Influenced by Seminal and Medium Osmolality’. *Cryobiology*, 53(1), pp. 37–50.](http://paperpile.com/b/HQttr3/SYxv)

[Santymire, R.M. *et al.* (2014) ‘The Black-Footed Ferret: On the Brink of Recovery?’ In Holt, W.V.Brown, J.L.and Comizzoli, P. (eds.) *Reproductive Sciences in Animal Conservation: Progress and Prospects*. New York, NY: Springer New York, pp. 119–134.](http://paperpile.com/b/HQttr3/uW74)

[Saremi, N.F. *et al.* (2019) ‘Puma Genomes from North and South America Provide Insights into the Genomic Consequences of Inbreeding’. *Nature Communications*, 10(1), p. 4769.](http://paperpile.com/b/HQttr3/yCWN)

[Savas, S., Tuzmen, S. and Ozcelik, H. (2006) ‘Human SNPs Resulting in Premature Stop Codons and Protein Truncation’. *Human Genomics*, 2(5), pp. 274–286.](http://paperpile.com/b/HQttr3/rhgV)

[Schubert, M. *et al.* (2014) ‘Prehistoric Genomes Reveal the Genetic Foundation and Cost of Horse Domestication’. *Proceedings of the National Academy of Sciences of the United States of America*, 111(52), pp. E5661–9.](http://paperpile.com/b/HQttr3/Nwdx)

[Shumate, A. and Salzberg, S.L. (2021) ‘Liftoff: Accurate Mapping of Gene Annotations’. *Bioinformatics* , 37(12), pp. 1639–1643.](http://paperpile.com/b/HQttr3/JWF6)

[Shuster, D.E. *et al.* (1992) ‘Identification and Prevalence of a Genetic Defect That Causes Leukocyte Adhesion Deficiency in Holstein Cattle’. *Proceedings of the National Academy of Sciences of the United States of America*, 89(19), pp. 9225–9229.](http://paperpile.com/b/HQttr3/pbK6)

[Siepel, A. *et al.* (2005) ‘Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes’. *Genome Research*, 15(8), pp. 1034–1050.](http://paperpile.com/b/HQttr3/XWXs)

[Simon, S.A. *et al.* (2009) ‘Short-Read Sequencing Technologies for Transcriptional Analyses’. *Annual Review of Plant Biology*, 60, pp. 305–333.](http://paperpile.com/b/HQttr3/VpbC)

[Singh, A.P. and Nala, R.R. (2018) ‘Estimation of the Status of Asiatic Lion (Panthera Leo Persica) Population in Gir Lion Landscape, Gujarat, India’. *Indian Forester*, 144(10), pp. 887–892.](http://paperpile.com/b/HQttr3/QwoN)

[Singh, H.S. (2017) ‘Dispersion of the Asiatic Lion Panthera Leo Persica and Its Survival in Human-Dominated Landscape Outside the Gir Forest, Gujarat, India’. *Current Science*, 112(5), pp. 933–940.](http://paperpile.com/b/HQttr3/tFLs)

[Smith, J.M. and Haigh, J. (1974) ‘The Hitch-Hiking Effect of a Favourable Gene’. *Genetical Research*, 23(1), pp. 23–35.](http://paperpile.com/b/HQttr3/3bYT)

[Sonstegard, T.S. *et al.* (2013) ‘Identification of a Nonsense Mutation in CWC15 Associated with Decreased Reproductive Efficiency in Jersey Cattle’. *PloS One*, 8(1), p. e54872.](http://paperpile.com/b/HQttr3/uz1X)

[Stein, L. (2001) ‘Genome Annotation: From Sequence to Biology’. *Nature Reviews. Genetics*, 2(7), pp. 493–503.](http://paperpile.com/b/HQttr3/Q3fU)

[Stormo, G.D. (2000) ‘DNA Binding Sites: Representation and Discovery’. *Bioinformatics* , 16(1), pp. 16–23.](http://paperpile.com/b/HQttr3/SqyF)

[Sutter, N.B. and Ostrander, E.A. (2004) ‘Dog Star Rising: The Canine Genetic System’. *Nature Reviews. Genetics*, 5(12), pp. 900–910.](http://paperpile.com/b/HQttr3/pwAz)

[Thanki, A.S. *et al.* (2018) ‘GeneSeqToFamily: A Galaxy Workflow to Find Gene Families Based on the Ensembl Compara GeneTrees Pipeline’. *GigaScience*, 7(3), pp. 1–10.](http://paperpile.com/b/HQttr3/XEIT)

[Urfer, S.R. (2009) ‘Inbreeding and Fertility in Irish Wolfhounds in Sweden: 1976 to 2007’. *Acta Veterinaria Scandinavica*, 51, p. 21.](http://paperpile.com/b/HQttr3/Cv6u)

[Venter, J.C. *et al.* (2001) ‘The Sequence of the Human Genome’. *Science*, 291(5507), pp. 1304–1351.](http://paperpile.com/b/HQttr3/3pp5)

[Vigne, J.-D. *et al.* (2004) ‘Early Taming of the Cat in Cyprus’. *Science*, 304(5668), p. 259.](http://paperpile.com/b/HQttr3/LD4r)

[Viluma, A. *et al.* (2022) ‘Whole-Genome Resequencing of Temporally Stratified Samples Reveals Substantial Loss of Haplotype Diversity in the Highly Inbred Scandinavian Wolf Population’. *Genome Research*. DOI:](http://paperpile.com/b/HQttr3/DIBK) [10.1101/gr.276070.121](http://dx.doi.org/10.1101/gr.276070.121)[.](http://paperpile.com/b/HQttr3/DIBK)

[Visscher, P.M. *et al.* (2017) ‘10 Years of GWAS Discovery: Biology, Function, and Translation’. *American Journal of Human Genetics*, 101(1), pp. 5–22.](http://paperpile.com/b/HQttr3/xvcU)

[Wang, Z. *et al.* (2011) ‘Domestication Relaxed Selective Constraints on the Yak Mitochondrial Genome’. *Molecular Biology and Evolution*, 28(5), pp. 1553–1556.](http://paperpile.com/b/HQttr3/Yk59)

[Wayne, R.K., Modi, W.S. and O’Brien, S.J. (1986) ‘MORPHOLOGICAL VARIABILITY AND ASYMMETRY IN THE CHEETAH (ACINONYX JUBATUS), A GENETICALLY UNIFORM SPECIES’. *Evolution; International Journal of Organic Evolution*, 40(1), pp. 78–85.](http://paperpile.com/b/HQttr3/nAKi)

[Weisman, C.M., Murray, A.M. and Eddy, S.R. (2022) ‘Mixing Genome Annotation Methods in a Comparative Analysis Inflates the Apparent Number of Lineage-Specific Genes’. *bioRxiv*, p. 2022.01.13.476251. DOI:](http://paperpile.com/b/HQttr3/XoTv) [10.1101/2022.01.13.476251](http://dx.doi.org/10.1101/2022.01.13.476251)[.](http://paperpile.com/b/HQttr3/XoTv)

[Wen, L. *et al.* (2020) ‘CRISPR/Cas9-Mediated TERT Disruption in Cancer Cells’. *International Journal of Molecular Sciences*, 21(2). DOI:](http://paperpile.com/b/HQttr3/fB9M) [10.3390/ijms21020653](http://dx.doi.org/10.3390/ijms21020653)[.](http://paperpile.com/b/HQttr3/fB9M)

[Wildt, D.E. *et al.* (1987) ‘Reproductive and Genetic Consequences of Founding Isolated Lion Populations’. *Nature*, 329(6137), pp. 328–331.](http://paperpile.com/b/HQttr3/Krwm)

[Wildt, D.E. *et al.* (1983) ‘Unique Seminal Quality in the South African Cheetah and a Comparative Evaluation in the Domestic Cat’. *Biology of Reproduction*, 29(4), pp. 1019–1025.](http://paperpile.com/b/HQttr3/fFS9)

[Wilson, B. (2015) ‘THE BLACK-FOOTED CAT FELIS NIGRIPES (BURCHELL, 1824): A REVIEW OF THE GEOGRAPHICAL DISTRIBUTION AND CONSERVATION STATUS’. DOI:](http://paperpile.com/b/HQttr3/jgAI) [10.13140/RG.2.2.18294.34889](http://dx.doi.org/10.13140/RG.2.2.18294.34889)[.](http://paperpile.com/b/HQttr3/jgAI)

[Wray, G.A. *et al.* (2003) ‘The Evolution of Transcriptional Regulation in Eukaryotes’. *Molecular Biology and Evolution*, 20(9), pp. 1377–1419.](http://paperpile.com/b/HQttr3/pcPN)

[Yordy, J. *et al.* (2020) ‘Body Size, Inbreeding, and Lifespan in Domestic Dogs’. *Conservation Genetics* , 21(1), pp. 137–148.](http://paperpile.com/b/HQttr3/6WIy)

[Zheng, W. *et al.* (2010) ‘Genetic Analysis of Variation in Transcription Factor Binding in Yeast’. *Nature*, 464(7292), pp. 1187–1191.](http://paperpile.com/b/HQttr3/yakP)

[Zhu, J. *et al.* (2016) ‘Identification of Tissue-Specific Protein-Coding and Noncoding Transcripts across 14 Human Tissues Using RNA-Seq’. *Scientific Reports*, 6, p. 28400.](http://paperpile.com/b/HQttr3/erbk)

[Zoonomia Consortium. (2020) ‘A Comparative Genomics Multitool for Scientific Discovery and Conservation’. *Nature*, 587(7833), pp. 240–245.](http://paperpile.com/b/HQttr3/QHTM)