Author:
Howard-McCombe, Jo A

Title:
Hybridisation and introgression in the Scottish wildcat
implications for conservation

University of BRISTOL

Hybridisation and introgression in the Scottish wildcat: implications for conservation

Jo Howard-McCombe

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Life Sciences

School of Biological Sciences

October 2021

Word count: ~ 54,000

Abstract

Hybridisation is an important driver of evolutionary change. Anthropogenic hybridisation, however, (i.e., hybridisation mediated by human activity, for example, through habitat disturbance, introduction of non-native species, or climate change), is considered a threat to wild populations. An important example is the Scottish wildcat, *Felis silvestris*, which can hybridise with the domestic cat, *Felis catus*, to produce fertile offspring. Scottish wildcats are the most endangered carnivore species in the UK; this population is considered critically endangered and at serious risk of extinction, as a direct result of hybridisation, in the near future.

An overview of the current status of wildcats in Scotland was obtained using a representative sample of 108 individuals from wildcat and domestic cat populations. Hybridisation levels were assessed using 6,546 unlinked SNP markers, highlighting the 'hybrid swarm' observed in the wild, i.e., a genetic continuum between the two parent species.

Using this information, a subset of 45 individuals from across the hybrid swarm were selected for whole genome resequencing. Additional reference individuals were included in the dataset, specifically, seven samples from mainland populations of European wildcats and 17 domestic cats from a global distribution. Low-coverage data were obtained from historic specimens from early-20th century Scotland.

Multiple approaches were taken to model hybridisation and introgression in the Scottish wildcat population, using both the unlinked SNP dataset, and whole genome sequence data. Firstly, a demographic model for wildcats was developed within an approximate Bayesian computational framework. A second approach applied haplotype-based methods to identify local ancestry (wildcat or domestic) across the genome.

Results presented here support recent onset of wildcat hybridisation, probably within the last 70 years, and accelerating during the latter part of the 20th century, rapidly generating the 'hybrid swarm' structure observed in Scotland today. An improved understanding of past hybridisation dynamics is important for conservation management of this species in the face on continuing gene-flow from domestic cats, in Britain, but also across the species range.

Acknowledgements

I am very grateful to many people who have supported me throughout my PhD.

First and foremost, my supervisors, Mark Beaumont, Dan Lawson, Helen Senn, Mike Bruford and Andrew Kitchener. I feel incredibly lucky to have had such a wonderful supervisory team. Thank you especially to Mark and Dan, who are endlessly encouraging, enthusiastic, and generous with their time; your mentorship has shaped me as a scientist. Also, a special thank you to Helen, who is not only a fantastic supervisor, but who inspired me to pursue a PhD in the first place.

I would like to thank the organisations that have provided additional funding for this project, specifically the People's Trust for Endangered Species and the Royal Zoological Society of Scotland. I would also like to thank collaborators who have very generously provided additional samples: Violeta Muñoz-Fuentes, Carsten Nowak, Greger Larson, Laurent Frantz, Carlos Driscoll, and William Murphy. Thank you to everyone who has contributed to wildcat sample collection in Scotland over the last few decades.

Many thanks to everyone at the RZSS for enthusiastic conversations about wildcat genomics. In particular, a big thank you to Jennifer Kaden and Jal Ghazali for processing the DNA samples and coordinating their shipment. Also, to Jess Wise, for her work communicating my research to a wider audience, and David Barclay for helpful conversations about the wildcat studbook.

My work has been supported and enhanced through various collaborations. Many thanks to Isa-Rita Russo and Manon Pribille, at Cardiff University, for facing the daunting task of genome assembly with me. Also, thanks to Greger Larson, Laurent Frantz and Alex Jamieson for many interesting discussions about cat evolution and domestication. I am very grateful to Jon Bridle and Martin Genner, as my internal annual reviewers, for valuable feedback on my research.

There are many people who I need to thank for their invaluable friendship and support during my time in Bristol. Firstly, Helen, Nick, and Holly, thank you for putting up with all my ramblings about wildcats, and for being amazing friends over the last four years. Hannah and Claudia, my long-distance PhD support, thank you both for all the science (and non-science) chats (and cake). Also, to Anne and Sophia (and pretty much everyone who turned up for a Friday night dinner in High Kingsdown). As always, Evan, Hannah, Caitlin and Llinos, you never fail to put a smile on my face, even from so far away. Hatti, I love you more than Turk.

Matt, you have supported me through all the ups and downs over the last few years and had unwavering confidence in me which I sometimes didn't have myself. You have been my rock (sorry I can't be more specific about the type), and I am eternally grateful. Finally, I can't really put into words my gratitude to my parents, Sally and Tim, and my sister, Connie. You have been my cheerleaders from the very beginning and make me feel as if I can achieve anything I put my mind to! Your endless love and support mean the world to me.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed: ..Jo Howard-McCombe..... Date: 28th October 2021

Publications

Part of this work has been published in the following articles:

Howard-McCombe J., Ward D., Kitchener A.C., Lawson D., Senn H.V., Beaumont M. (2021) On the use of genome-wide data to model and date the time of anthropogenic hybridisation: An example from the Scottish wildcat. *Molecular Ecology*, 30, 3688-3702

Howard-McCombe, J., Senn, H., Lawson, D., Bruford, M., Beaumont, M. (2021) Hybridisation and introgression in the Scottish wildcat. In: German Wildlife Foundation (Ed.) *On the right track? The situation of the wildcat in Germany and Europe*. Proceedings of the European wildcat symposium 2019, Neuwied (pp. 124-127)

Additional samples

I am very grateful to collaborators who have provided additional samples for this study. This section details the samples provided, and the format data was provided in. Samples were provided by:

Violeta Muñoz-Fuentes (The European Bioinformatics Institute) and Carsten Nowak (Senckenberg Research Institute and Natural History Museum)

6 German wildcats, *whole genome resequencing data, FASTQ format*

Carlos Driscoll (National Institute on Alcohol Abuse and Alcoholism, NIH)

1 Portuguese wildcat, 1 *F. bieti*, 1 *F. margarita*, 1 *F. l. ornata, whole genome resequencing data, BAM format*

William Murphy (Texas A&M University)

1 *F. bieti*, 2 *F. l. ornata, whole genome sequence data, FASTQ format*

Greger Larson (University of Oxford) and Laurent Frantz (Ludwig Maximillian University)

4 historic samples from 20th century Scotland, *whole genome resequencing data, VCF*

23 historic samples from 20th century Scotland, *genetic screening data, VCF*

Alex Jamieson (University of Oxford)

2 ancient samples from Britain, *whole genome resequencing data, VCF*

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Hybridisation

1.1.1 Hybridisation as an important evolutionary process

Diverse organisms across the tree of life exchange genetic material, often driving important evolutionary change. For example, the horizontal gene transfer of antibacterial resistance between bacteria (de la Cruz & Davies, 2000), or tumour-inducing genes from *Agrobacterium* to plant hosts (Quispe-Huamanquispe, Gheysen, & Kreuze, 2017). Eukaryotic cells are themselves a product of symbiosis between prokaryotes and subsequent co-opting of aerobic/photosynthetic bacterial DNA in the mitochondria and chloroplasts (Cooper, 2000). Gene flow between populations is a fundamental part of evolutionary biology, interacting with natural selection and genetic drift to shape population trajectories. Here, I focus on hybridisation in sexually reproducing species as a mechanism for gene exchange. 'Hybridisation' refers to interbreeding between divergent populations or species, and 'introgressive hybridisation' the transfer of genetic material between parental groups via back-crossing with hybrid individuals (Mallet, 2005).

Since the beginnings of the field, hybridisation has been of significant interest to evolutionary biologists. Darwin referred to hybridisation in multiple publications (Darwin, 1875; 1876), including 'The Origin of Species' (1859), and conducted several breeding experiments exploring the fertility of plant hybrids. Zoologists and botanists have traditionally held contrasting views on the contribution of hybridisation to evolution. In plant species, hybridisation has long been considered a source of adaptive variation and driver of speciation (Anderson, 1949; Anderson & Stebbins, 1954; Stebbins, 1959; Grant, 1981). In animal species, however, hybridisation was thought to be an evolutionary dead end, occurring rarely in nature and only acting to reinforce reproductive isolation of the parental groups (Dobzhansky, 1937; Mayr, 1963).

It is likely that phenotypic and/or behavioural similarities between animal species and their hybrids have masked the extent of hybridisation (Mallet, 2005). Molecular methods to detect hybridisation and introgression have shown it to be far more common in animals than previously thought; Mallet (2005) estimated one in ten animal species hybridise with a related species. Hybridisation has been now been reported across a diverse range of taxa, with examples in insects (Mavárez et al., 2006), birds (Poelstra et al., 2014), mammals (Larsen, Marchán-Rivadeneira, & Baker, 2010), fish (Meier et al., 2017), amphibians (Novikova et al., 2020) and reptiles (Sovic, Fries, & Gibbs, 2016). Several studies have demonstrated ancient hybridisation between divergent groups, e.g., between cave bears and brown bears (Barlow et al., 2018) or between Neanderthals and modern humans (Kuhlwilm et al., 2016; Prüfer et al., 2014).

Hybridisation is difficult to reconcile with the biological species concept, which states that species exist in reproductive isolation, i.e., pre- or post-zygotic barriers prevent interbreeding between individuals of different species (Mayr, 1942). A strict interpretation of Mayr's definition is challenged by gene flow as a result of hybridisation. Though organisms can be grouped by genuine similarities in morphology, ecology, genetics, etc., designation of species is often contentious. Many definitions for 'species' have been proposed (Hausdorf, 2011), though a single workable definition remains controversial, and probably impossible, given species delineation is an attempt to discretise the continuum of biological diversity. Additionally, mechanisms for speciation are not fully understood. Mayr (1942) proposed that reproductive isolation develops through geographic isolation, with physical or ecological barriers to gene flow promoting divergence between populations (allopatric speciation). Other geographic mechanisms have since been described, including sympatric or parapatric speciation (involving no or only partial geographic separation) (Mallet, Meyer, Nosil, & Feder, 2009). Hybridisation itself provides a mechanism for new species to arise (hybrid speciation) (Mallet, 2007). Hybridisation rapidly generates genetic diversity, faster than by mutation or recombination within species (Schwenk, Brede, & Streit, 2008), and hybrid genomes contain novel combinations of genes and alleles, and unique epistatic interactions (Moran et al., 2021). Hybrid populations therefore provide important study systems to address fundamental aspects of evolutionary biology, including adaptive variation, the evolution of isolating mechanisms between species, and the concept of 'species' itself.

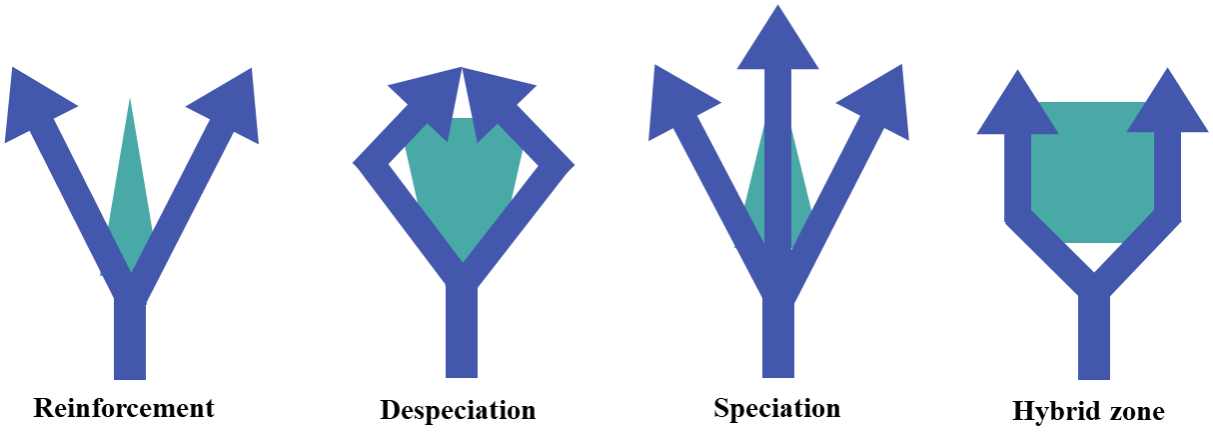

Figure 1.1. Hybridisation between diverged populations can lead to a variety of outcomes including (1) reinforcement, promoting further divergence of the parent groups, (2) despeciation, the merging of distinct lineages, (3) speciation, the formation of a third, hybrid, lineage or (4), the establishment of a stable hybrid zone.

Contact between diverged populations can lead to a range of outcomes (Fig. 1.1). In general, hybrids are expected to be less fit than the parental types (Barton, 2001). Hybridisation, and recombination between parental haplotypes in subsequent generations, generates novel genotypes at random. Unlike in the parent groups, natural selection has not acted on the recombinant genotypes, which are therefore expected to be less fit, on average. First generation (F_1) hybrids are a common

exception to this, through a phenomenon known as heterosis, or hybrid vigour. F_1 hybrids may be more fit than parent types due to (1) masking of deleterious recessive alleles from one parental population by dominant alleles of the second, (2) overdominance, the superiority of heterozygous sites in hybrids compared to homozygous parental sites, or (3) positive epistatic interactions across new combination of genes (Lamkey & Edwards, 1999). This phenomenon has been observed across diverse taxa and is commonly exploited in crop species (Fig. 1.2).

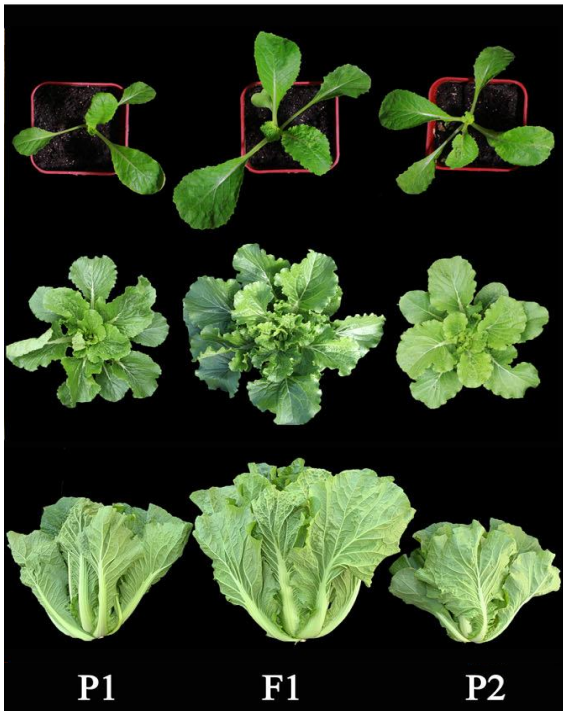

Figure 1.2. Example of heterosis in Chinese cabbage, where F_1 hybrids have a higher biomass than either parent (P_1, P_2).

From Li *et al.* (2021)

Later hybrids generally show a decline in fitness (Barton, 2001). Recombinant genotypes may be less fit in the parental environment, accumulate deleterious alleles from both parental groups, or harbour negative interactions between novel combinations of genes (Moran et al., 2021). Maladapted hybrid offspring can drive reinforcement, where a low rate of hybridisation promotes further divergence of the parental groups, potentially until reproductive isolation (Servedio & Noor, 2003). Hybridisation itself may drive the evolution of reproductive isolating mechanisms, such as hybrid incompatibility, or hybrid sterility, as has been observed between fruit fly species, *Drosophila simulans* and *D. mauritiana* (Wu & Ting, 2004).

Conversely, if poor reproductive isolation exists, or hybrids are as fit as the parent groups, high geneflow can homogenise the two mixing groups into a single population (Mallet, 2007), as has been observed in Darwin's finches (Grant & Grant, 1996). This is sometimes referred to as 'despeciation' (Mallet, 2007).

A third (hybrid) lineage can be established via hybridisation. Allopolyploid speciation, where hybrids carry double the parental number of chromosomes, instantly results in the reproductive isolation of hybrid individuals (Mallet, 2007). Allopolyploid speciation is common in plants, especially those capable of asexual reproduction or selfing, which mitigate the 'minority cytotype disadvantage' (where polyploid hybrids are initially rare and backcrossing with parent species is incompatible), allowing the hybrid population to become established. Homoploid hybrid speciation, (i.e., without a change in chromosome number), is rarer, given the potential for continued geneflow between hybrids and parent species. In general, homoploid hybrid speciation is driven by novel

variation, allowing the expansion of hybrids into a new ecological niche. For example, a hybrid of the butterfly species *Lycaeides melissa* and *L. idas* is adapted to alpine environments not occupied by either parent species (Gompert, Fordyce, Forister, Shapiro, & Nice, 2006). *Helianthus paradoxus*, a *Helianthus* sunflower hybrid, is tolerant of high saline environments as a result of the additive effects of loci inherited from both parental species (Lexer, Welch, Raymond, & Rieseberg, 2003). Hybridisation may generate variation to facilitate adaptive radiation and speciation (Seehausen, 2004), e.g., in African cichlids (Meier et al., 2017; Svardal et al., 2020). Although hybrids are expected to be less fit, on average, than the parent species (accounting for heterosis), some hybrids may outcompete parents in the parental environment (Barton, 2001). If recombinant genotypes reach a higher 'adaptive peak', parental genotypes will be displaced.

Hybrid zones are regions where divergent populations meet and hybridise (Barton & Hewitt, 1985). Hybrid zones are dynamic systems that can move in response to fluctuating population density, environmental conditions, or individual fitness, and can be transient or stable. A hybrid zone can represent a cline between the parent populations or have a mosaic-like structure. Hybrid zones are commonly maintained through a migration-selection equilibrium (also known as a tension zone); dispersal of hybrids works to widen the hybrid zone, whilst negative selection against hybrids narrows it. Selection for or against hybrids can be endogenous or exogenous, i.e., relating to innate hybrid

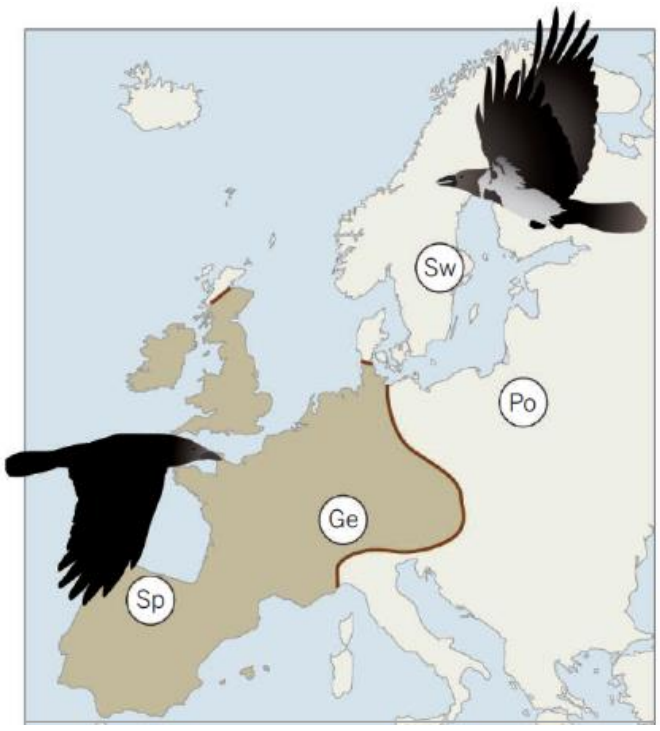

Figure 1.3. The hybrid zone (dark brown) between carrion and hooded crows runs north to south across central Europe and has been fairly stable over the last century. Assortative mating maintains phenotypic differences between all-black carrion crow and grey-coated hooded crows. There is substantial gene flow between the two groups, and German crows are closer, genetically, to Polish crows than to Spanish ones.

From De Knijff (2014)

fitness, or the relative fitness of hybrids in their environment. A fairly stable hybrid zone is maintained between carrion crows and hooded crows in Europe, for example, by assortative mating and sexual selection, despite substantial gene flow between the two species (Poelstra et al., 2014) (Fig. 1.3).

Hybridisation can contribute directly to the evolution of parent populations through introgression, the movement of genes and alleles between species via backcrossing with hybrids (Harrison & Larson, 2014). Differential rates of introgression across loci are observed at hybrid zones. Variation with a selective advantage can introgress quickly (adaptive introgression), whilst patterns of introgression at neutral loci are stochastic. There are several examples of adaptive introgression in wild animal species, including the introgression of genes linked to Müllerian mimicry in *Heliconius* butterflies (Pardo-Diaz et al., 2012) (Fig. 1.4) and rodenticide resistance in European house mice (Song et al., 2012). Conversely, regions of the genome, often linked to reproductive isolation, or local adaptation, have been shown to be resistant to introgression (Harrison & Larson, 2014).

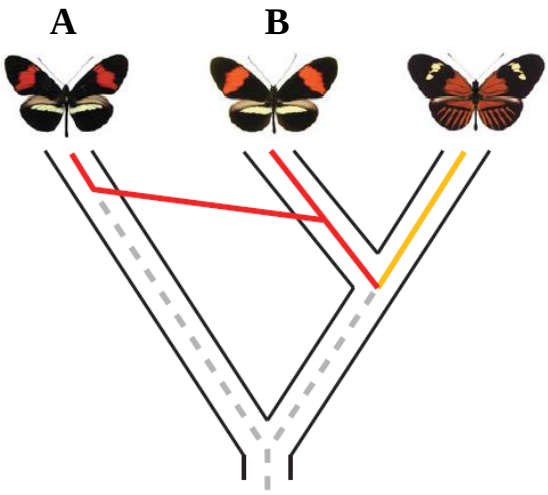

Figure 1.4. Adaptive introgression (as opposed to ancestral polymorphism) is the source of a shared mechanism of Müllerian mimicry in *Heliconius* butterflies. A gene tree view shows the relationship between three *Heliconius* species, where the gene for red wing colour (red line) has evolved in species B and introgressed into species A.

From Smith & Kronforst (2013)

The advent of molecular methods to detect hybridisation and introgression reconcile traditionally contrasting views of hybridisation. The role of hybridisation in evolution appears to be varied, acting at different times to reinforce speciation or homogenise divergent populations, create dynamic hybrid zones or promote genetic diversity and speciation.

1.1.2 Anthropogenic hybridisation

Natural hybridisation is widespread across diverse taxa (Barton, 2001), and likely to have contributed to the evolution of many species, including our own (Kuhlwilm et al., 2016; Prüfer et al., 2014). Human-mediated hybridisation, however, is of increasing concern in conservation biology (Allendorf, Leary, Spruell, & Wenburg, 2001; McFarlane & Pemberton, 2019; Quilodrán, Montoya-Burgos, & Currat, 2020; Todesco et al., 2016). Human activity alters species distributions directly, through the introduction of non-native and domesticated species, and indirectly, by habitat disturbance and climate change. Subsequent contact between species that would otherwise exist allopatrically, or the

breakdown of isolating mechanisms between sympatric species, leads to hybridisation often associated biodiversity loss, either through introgression and genetic erosion, or population or species extinction.

Translocation of non-native species can be deliberate, e.g., for fishing or hunting, or accidental, e.g., emptying of shipping ballast (Mack et al., 2000). Both are increasing as a result of global travel and trade. Invasive species are a major conservation issue for several reasons, including the spread of disease, habitat or ecosystem alteration, and competition or hybridisation with native taxa. Hybridisation with feral mallards (*Anas platyrhynchos*), for example, introduced in many regions as a game-bird, threatens the genetic distinction of an endemic duck species in Hawaii (*Anas wyvilliana*) (Wells et al., 2019) and New Zealand (Rhymer, Williams, & Braun, 1994). Allopolyploid hybrid speciation between British and North American cordgrasses (translocated in shipping ballast), produced the highly invasive *Spartina anglica* (Thompson, 1991). Initially established on the south coast of Britain in the 19th century, *S. anglica* has since spread as far as China (An et al., 2007) and Australia (Kriwoken & Hedge, 2000).

Hybridisation between wild and domestic species is an important aspect of this. Domesticated species are transported worldwide, where free-ranging or escaped individuals can hybridise with native species. European examples include dogs and wolves (Fan et al., 2016), pigs and wild boar (Scandura et al., 2008), ibex and goats (Grossen et al., 2014) and domestic cats and wildcats (Tiesmeyer et al., 2020). Domestic animals make up a significant proportion of the planet's total biomass (livestock species account for 60% of total mammal biomass, humans 36%, the remaining 4% are wild species), significantly outnumbering their wild progenitors (Bar-On, Phillips, & Milo, 2018). Hybridisation between groups with asymmetric population sizes can lead to the rapid extinction of rare species (Epifanio & Philipp, 2001). Escaped pollen and seeds from crop species, and subsequent hybridisation with native plants, is associated with the extinction of rare native forms, as well as the establishment of aggressive weed species (Ellstrand, Prentice, & Hancock, 1999). Additionally, hybridisation and introgression may result in the spread of maladaptive domestic variation; artificially selected variation is expected to be less fit in wild environments (Quilodrán, Montoya-Burgos, et al., 2020).

Human activity also alters species distributions indirectly. Habitat change or disturbance can modify the fitness landscape, promoting the establishment of hybrids (Anderson, 1948, Grabenstein & Taylor, 2018) and/or removing existing barriers between divergent populations (Grabenstein & Taylor, 2018). For example, human-induced eutrophication (algal or plant blooms in response to changing nutrient levels) limits light levels in freshwater and coastal ecosystems (Alexander, Vonlanthen, & Seehausen, 2017). The subsequent breakdown of assortative mating (based on mate colouration) as a mechanism of reproductive isolation is observed in many fish species. In Lake

Victoria, eutrophication has been proposed as a contributor to the extinction of at least 200 species of endemic cichlids over the last 30 years. Climate change impacts species distributions, habitat use, migration or breeding patterns, all of which may lead to contact between divergent populations or species (Hoffmann & Sgró, 2011).

Hybridisation can lead to extinction through two main mechanisms: demographic or genetic swamping (Todesco et al., 2016). Demographic swamping occurs when hybrid offspring are sterile or less fit than the parent populations (outbreeding depression). In this case, frequent hybridisation leads to a low or negative growth rate in one or both parent populations. Alternatively, frequent hybridisation producing relatively fit offspring may lead to homogenisation of the mixing groups, or complete replacement by invasive genotypes, known as genetic swamping. Selection against hybrids must be very strong to limit backcrossing with parental populations (Epifanio & Philipp, 2001); the literature review carried out by Todesco *et al.* (2016) suggests that genetic swamping is more common than demographic swamping (though, with the advent of molecular markers it may be easier to directly observe genetic swamping, unlike demographic swamping, where population decline may be attributed to other causes).

Introgression can impact parent populations through the loss of locally adaptive variation or spread of maladaptive variation (Todesco et al., 2016). The ecological impacts of the spread of modified genes through hybridisation is poorly understood and remains an important topic in the debate surrounding the use of genetically-modified organisms (Quilodrán, Montoya-Burgos, et al., 2020).

Anthropogenic hybridisation may also have a positive impact on wild populations through adaptive introgression, i.e., the transfer of adaptive variation between parental groups. For example, introduction of major histocompatibility complex diversity to Alpine ibex from goats (Grossen et al., 2014). Deliberate outbreeding (or 'genetic rescue') is used as a conservation management tool to boost genetic diversity of inbred populations, for example, in the Florida panther (Johnson et al., 2010).

In general, hybridisation poses a challenge for conservation management. Interbreeding between divergent populations complicates the delineation of units for conservation, with direct implications for conservation policy and legal protection (Allendorf et al., 2001; Quilodrán, Montoya-Burgos, et al., 2020; Wayne & Shaffer, 2016). Initial detection of hybrids can be challenging; relying on morphology alone can be misleading and does not provide an estimate of introgression level (Allendorf et al., 2001). Estimates of individual admixture proportions require appropriate genetic markers with the power to accurately discriminate between the mixing groups, and detect back-crosses spanning the number of generations of contact between populations (McFarlane & Pemberton, 2019). Systematic genetic sampling is needed to monitor the spatial and temporal patterns of

hybridisation and introgression, which may not be logistically or economically feasible for many conservation programmes.

Nevertheless, the last few decades have seen widespread application of molecular methods in conservation biology (Allendorf, Hohenlohe, & Luikart, 2010). This has been a double-edged sword in terms of practical management. Whilst providing an improved understanding of population history (important to disentangle the impacts of natural and anthropogenic hybridisation) and accurate quantification of hybridisation and introgression, fine-scale information about individual or population ancestry highlights both scale of hybridisation in natural systems and the complexity of populations' evolutionary histories (Wayne & Shaffer, 2016).

An improved understanding of hybridisation has encouraged a recent revisiting of the question of 'what to conserve' in relation to hybrid systems (Draper, Laguna, & Marques, 2021; Fitzpatrick, Ryan, Johnson, Corush, & Carter, 2015; Quilodrán, Montoya-Burgos, et al., 2020; Wayne & Shaffer, 2016). Two central goals of conservation biology are to protect biodiversity and conserve natural evolutionary processes (Frankham, Ballou & Briscoe, 2010). Anthropogenic hybridisation has historically been perceived as a threat to both, with hybridisation currently listed as a threat to the survival of species by the International Union for the Conservation of Nature (IUCN) (IUCN, 2021). The potential for rapid extinction of endangered populations following hybridisation promotes a 'precautionary principle' among conservationists. However, species-led conservation often disregards the potential value of hybrid individuals in terms of their ecological function, or as a source of genetic diversity or reservoir of native species genes (Chan, Hoffmann, & van Oppen, 2019).

Hybridisation remain a legislative blind spot, most hybrids are without any form of legal protection, even those in naturally-occurring hybrid zones. Neither the EU Habitats Directive (Council Directive 92/43/EEC, 1992) nor Canada's Species at Risk Act (2002) provide detailed guidance on managing hybrid populations. The US Endangered Species Act initially excluded hybrids. A hybrid policy, proposed in 1996, extended legal protection to admixed individuals which are more similar to the endangered parent species than F_1 hybrids, or hybrids that are the result of genetic rescue (US Fish and Wildlife Service 1996). However, this has yet to be officially accepted or rejected by the US Fish and Wildlife Service or the National Marine Fisheries Service (Wayne & Shaffer, 2016). Given the prevalence of hybridisation (both anthropogenic and natural), the number of mechanisms driving anthropogenic hybridisation, and its varied and unpredictable outcomes, many authors highlight the need for a more flexible approach moving forwards (Chan et al., 2019; Draper et al., 2021; Fitzpatrick et al., 2015; Quilodrán, Montoya-Burgos, et al., 2020; Wayne & Shaffer, 2016).

A distinction should be made between natural and anthropogenic hybridisation, where possible, with naturally-occurring hybrids made eligible for legal protection (Allendorf et al., 2001). Allendorf *et al.* (2001) describe three broad outcomes of anthropogenic hybridisation: (1)

hybridisation without introgression, i.e., few or sterile first-generation hybrids are created, potentially leading to demographic swamping, (2) widespread introgression (fertile F_1s backcrossing with parent populations), or (3) complete admixture, or genetic swamping, with few unadmixed individuals remaining and limited genetic distinction between parental populations. McFarlane & Pemberton (2019) propose (2) and (3) are the same outcome, only differing in the amount of time elapsed since secondary contact. Quilodrán *et al.* (2020) describe a fourth category, with fertile F_1 hybrids but no introgression. This is the outcome of hybridisation through genome exclusion, common in freshwater fish, where recombination between homologous chromosomes does not occur in F_1 hybrids, which transmit the complete genetic material of only one parent. Quilodrán *et al.* (2018) show that this mechanism of hybridisation can very quickly (within a few generations) result in the extinction of parental forms. Quilodrán *et al.* (2020) argue that these scenarios should be a priority for conservation.

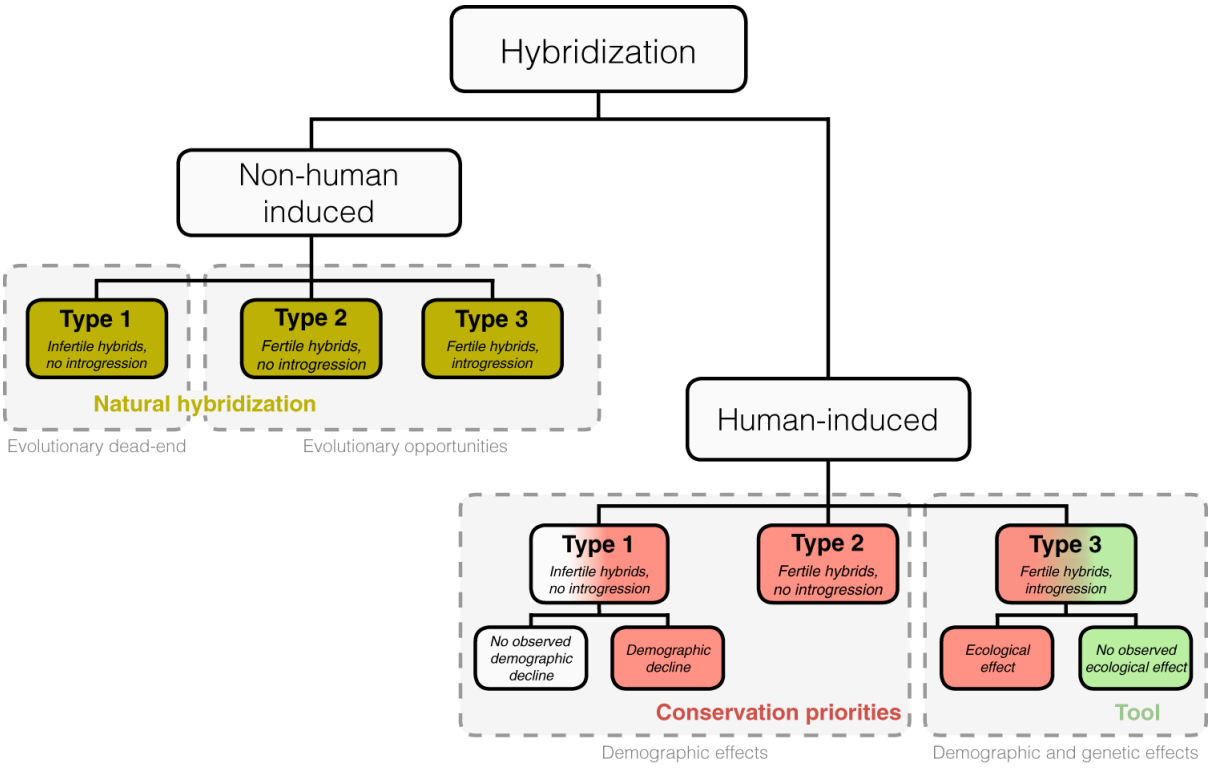

Figure 1.5. Priorities for conservation following anthropogenic hybridisation, as determined by Quilodrán *et al.* (2020). It is important to first distinguish between natural and anthropogenic hybridisation. The consequences of anthropogenic hybridisation have been classified into different groups, depending on, e.g., hybrid viability or introgression levels, to better support management decisions (see text). Different authors propose different scenarios where anthropogenic hybridisation might be acceptable, or even beneficial, for wild populations. Quilodrán *et al.* (2020) propose conservation intervention only if a direct demographic decline or negative ecological consequence is observed. Types of hybridisation with these potential outcomes are highlighted in red as priorities for conservation. Quilodrán *et al.* (2020) show scenarios where hybridisation may lead to increased genetic diversity in green.

From Quildorán *et al.* (2020)

Allendorf *et al.* (2001) argue for minimising anthropogenic hybridisation in general, with protection for hybrids only in exceptional circumstance (e.g., in cases of complete admixture). However, Quilodrán *et al.* (2020) and Wayne & Shaffer (2016) suggest hybrid fitness, demography and ecological function should be evaluated to inform hybrid management. Quilodrán *et al.* (2020) propose that hybridisation without an impact on ecological function or demographic decline of the parent species should not represent a priority for conservation intervention (Fig. 1.5). Wayne & Shaffer (2016) emphasise that an attempt at habitat restoration (to promote selection of native genes) should be carried out prior to direct management or elimination of the hybrid population.

It is clear that there is no 'one size fits all' approach to managing the impact of anthropogenic hybridisation on wild populations. In the midst of a biodiversity crisis (Ceballos, Ehrlich, & Dirzo, 2017), hybridisation and introgression pose a serious extinction risk to wild populations, with impacts on ecological interactions and adaptive capacity that remain poorly understood. Factors driving anthropogenic hybridisation, i.e., habitat disruption, globalisation, and climate change, show no signs of slowing into the 21st century. Scenarios must be considered on a case-by-case basis, and supported by accurate molecular methods, to make appropriate management decisions.

1.2 Study system

European wildcats, *Felis silvestris*, are an important example of the impact of anthropogenic hybridisation on wild populations. Widespread across Europe, in Britain the remaining wildcat population is distributed across the Scottish Highlands, where extensive hybridisation between wildcats and domestic cats is observed (Beaumont et al., 2001; Senn et al., 2019). This population is considered critically endangered and at serious risk of extinction via genetic swamping (Breitenmoser, Lanz, & Breitenmoser-Würsten, 2019).

1.2.1 European wildcats

Modern cat species (subfamily Felinae) are successful and widespread carnivores. Emerging in Eurasia during the late Miocene, modern Felinae constitute eight distinct evolutionary lineages found across all continents except Antarctica (Johnson et al., 2006). European wildcats, *Felis silvestris*, belong to the most recently diverged (~6.2mya) group, the domestic cat lineage, which includes small cat species such as sand cats (*Felis margarita*), jungle cats (*Felis chaus*) and black-footed cats (*Felis nigripes*), distributed across Africa and Eurasia.

Wildcats have a wide distribution across Africa, Asia, Europe, and the Middle East (Fig 1.6). Differences in morphology (Fig. 1.6), ecology (Nowell & Jackson, 1996) and phylogenetic clustering (Driscoll et al., 2007) are observed across the range, however, resolving the taxonomy of these

overlapping and interfertile populations is not straightforward. Phylogenetic analysis by Driscoll *et al.* (2007), using 36 microsatellites and ~2.6Kb of mitochondrial genome sequences, supported genetic clustering of four biogeographic groups, described as a single polytypic species, *Felis silvestris*, with four subspecies: *F. s. silvestris* in Europe, *F. s. lybica* in Africa, *F. s. ornata* in the Middle East and Asia, and *F. s. bieti* in China.

The subspecies referred to by Driscoll *et al.* (2007) are not used consistently across the available literature, with some disagreement as to whether these groups constitute separate species or subspecies. Scottish wildcats are variously referred to as a subpopulation of *F. silvestris* or *F. s. silvestris*, or even (historically) their own subspecies, *F. s. grampia* (Miller, 1907).

The most recent taxonomy (used here) splits wildcats into three full species, *Felis silvestris* (Europe, *F. s. silvestris*, and the Caucasus, *F. s. caucasica*), *Felis lybica* (south Africa, *F. l. cafra*, north Africa and the Middle East, *F. l. lybica*, and central and southwest Asia, *F. l. ornata*), and *Felis bieti* (China) (Kitchener et al., 2017) (Fig. 1.6). The delineation of these species is likely to be subject to further change in light of additional genomic data; recent whole-genome sequence based analyses, for example, disputes the status of Chinese desert cats (*F. bieti*) as a separate species (Yu et al., 2021).

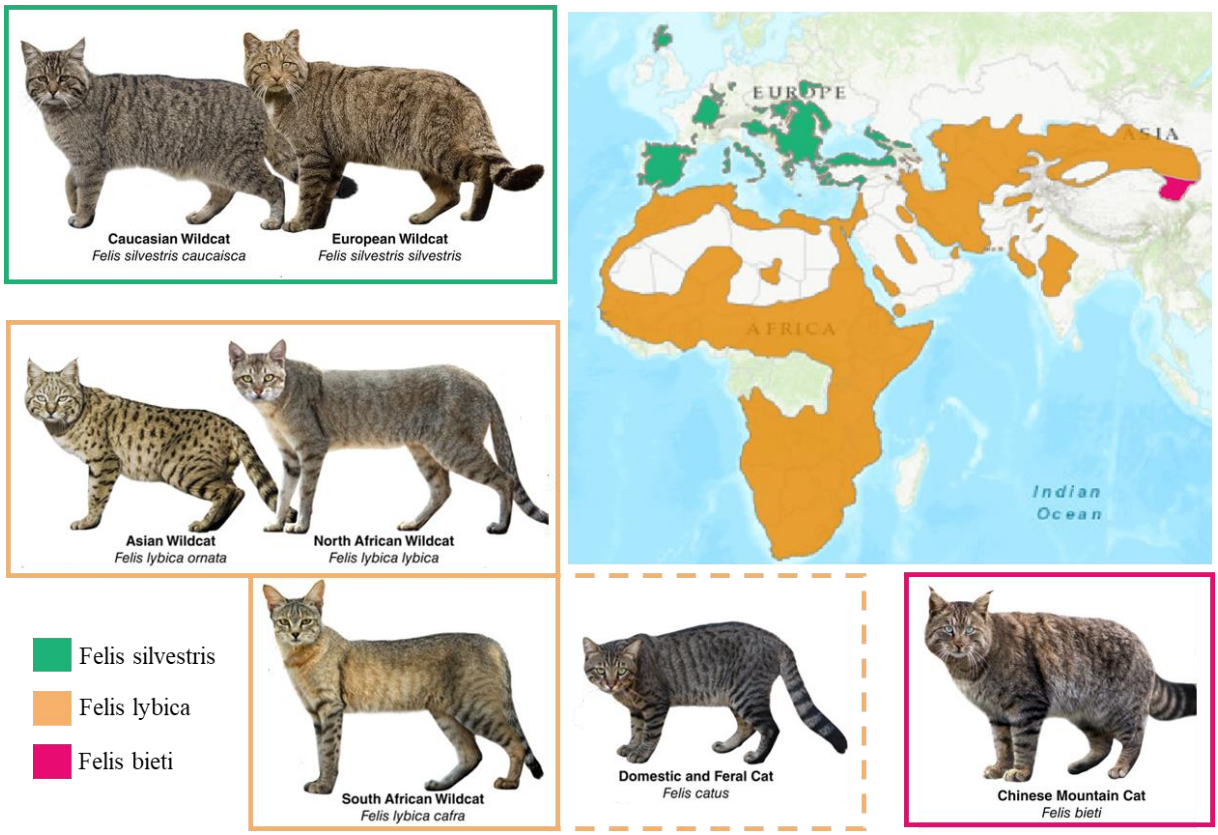

Figure 1.6. Distribution of wildcat species across Africa and Eurasia. Wildcats are currently classified into three species: *Felis silvestris* (green), *Felis lybica* (orange), and *Felis bieti* (pink) (Kitchener et al., 2017). Domestic cats are derived from the Near-Eastern wildcat species, *F. lybica* (Driscoll et al., 2007).

Map adapted from Yamaguchi *et al.*, (2015)
Photos from Castelló, (2020)

However, for now, we refer to wildcats using the taxonomy of Kitchener *et al.* (2017), and in concordance with studies by Beaumont *et al.* (2001) and Senn *et al.* (2019).

European wildcats are distributed from Turkey in the southeast to Scotland in the northwest (Kitchener et al., 2017). Historically a continuous range, since the 17th century the wildcat distribution in Europe has become increasingly fragmented as a result of hunting and habitat loss, with local extinctions in the Netherlands, Austria, England and Wales (Yamaguchi, Kitchener, Driscoll, & Nussberger, 2015). Recent recovery has been reported for some populations, for example, in Germany, where the population is expanding (Mueller et al., 2020), and is the likely source for recolonisation of the Netherlands (Canters, Thissen, Diepenbeek, Jansman, & Goutbeek, 2005). Other populations, such as those on the Iberian peninsula, continue to decline (Yamaguchi et al., 2015).

Mattucci *et al.* (2016) used genetic clustering at 31 microsatellite markers to classify European wildcats into five biogeographic groups: the Dinaric Alps, the Italian Peninsula, central Germany, central Europe and the Iberian Peninsula (Fig. 1.7). This population structure was shown to be the result of isolation (in glacial refugia) during the late Pleistocene, rather than recent

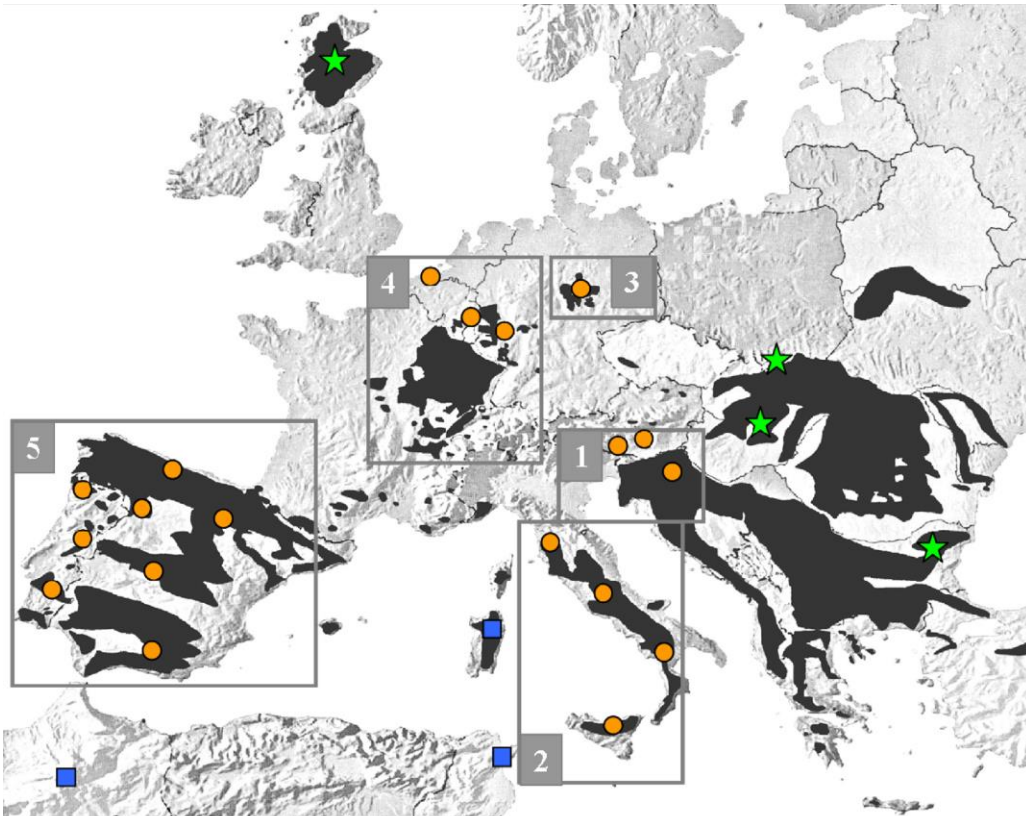

Figure 1.7. Biogeographic structure of European wildcats. Five main groups were described by Mattucci *et al.* (2016): (1) the Dinaric Alps, (2) the Italian Peninsula, (3) central Germany, (4) central Europe and (5) the Iberian Peninsula. Scottish and Hungarian populations (green stars) were excluded due to high levels of introgression.

From Mattucci *et al.* (2016)

fragmentation. Populations in Scotland and Hungary were excluded from this analysis due to high levels of introgression from domestic cats.

Wildcats occupy a mosaic habitat of woodland and open grassland, sheltering in woodland and hunting in the open (Yamaguchi et al., 2015). They are generally absent from urban areas, human settlements, or areas of intensive agriculture. Wildcats are solitary and largely nocturnal, hunting at night, with common prey species including rodents, rabbits, and small birds (Sunquist & Sunquist, 2002). Classified by the IUCN Red List as 'Least Concern' (in terms of species extinction risk), wildcat populations are threatened locally by habitat loss or fragmentation, hunting (as pests and indirectly as by-catch), road mortality and hybridisation with domestic cats (Yamaguchi et al., 2015).

1.2.2 Cat domestication

African wildcats, *Felis lybica*, are the wild progenitors of modern domestic cats (Driscoll et al., 2007) The process of cat domestication was initiated in the Near East, probably as a result of the attraction to rodents, who themselves were attracted to grain stores associated with settled agriculture ~9,500 years ago. The dispersal of domestic cats by humans appears to have been ongoing since that time, following routes (including by sea) of trade and travel (Ottoni et al., 2017). Mitochondrial lineages of modern domestic cats have predominantly been traced back to ancestors in Egypt and the Near East (Driscoll et al., 2007; Ottoni et al., 2017). Today, cats are one of the most popular companion animals, with an estimated 600 million individuals worldwide (excluding an unknown number of feral domestic cats) (Gehrt, Riley & Cypher, 2010).

Unlike most domesticates, the relationship between cats and humans has been largely commensal, with weaker artificial selection than in, for example, dogs (Montague et al., 2014). Modern cat breeds have all emerged within the last ~200 years, and have generally been selected for their appearance, rather than function. Nevertheless, domestication has altered the morphology, behaviour (most obviously tameness), and rate of reproduction in domestic cats (Driscoll, Macdonald, & O'Brien, 2009; Montague et al., 2014). As a result, they are sufficiently diverged from wildcats to be considered a separate species, *Felis catus* (International Commission on Zoological Nomencalture, 2003).

Domestic cats are widely considered to be an invasive species which pose a threat to local wildlife (Trouwborst, McCormack, & Martínez Camacho, 2020). They ranked third, behind rats and chytrid fungus, in a recent survey of invasive species that threaten the most vertebrate species worldwide. The largest, and best studied, impact of domestic cats is predation of wildlife. Cats are the highest source of human-mediated bird mortality in both the USA (Loss, Will, & Marra, 2015) and Canada (Blancher, 2013). In Australia, domestic cats are estimated to kill 377 million birds (Woinarski et al., 2017) and 649 million reptiles per year (Woinarski et al., 2018). Over a 5-month

period, domestic cats in the UK were estimated to predate 57 million mammals, 27 million birds and five million reptiles and amphibians (Woods, McDonald, & Harris, 2003). Their impact on endemic island species has been well-documented, for example, the extinction and extirpation of many endemic reptile and bird species on mainland Mauritius (Bell, 2002). Domestic cats can also negatively impact native species through competition, i.e., for habitat or prey, the spread of diseases, e.g., the spread of feline leukaemia virus to the endangered Florida panthers, and, importantly, hybridisation (Trouwborst et al., 2020).

1.2.3 Hybridisation as a threat to wildcats

Domestic cats and wildcats (*F. silvestris, F. lybica* and *F. bieti*) can hybridise to produce fertile offspring (Driscoll et al., 2007). Domestic cats are ubiquitous across the wildcat range, and hybridisation has been reported in several regions (Le Roux, Foxcroft, Herbst, & MacFadyen, 2015; Tiesmeyer et al., 2020; Yu et al., 2021). This is perhaps unsurprising, given the interfertility of the wild species, and limited divergence time between domestic cats and African wildcats (Driscoll et al., 2007). Several breeds of domestic cat are the product of deliberate hybridisation between *F. catus* and exotic cat species. For example, Bengal cats, which are a cross with Asian leopard cats (*Prionailurus bengalensis*), or Savannah cats, which are a serval (*Leptailurus serval*) hybrid.

Recent population declines and habitat fragmentation, alongside a continued increase in domestic cat ownership in Europe, mean domestic cats are considered a threat to wildcats (Yamaguchi et al., 2015). This is primarily due to hybridisation and introgression (i.e., the threat of genetic swamping, or spread of maladaptive variation), but also disease transmission and increased competition from feral domestic cats for prey and shelter. Contact between wildcats and domestic cats is therefore monitored across Europe. Systematic wildcat surveys have been carried out, e.g., in Scotland (Senn et al., 2019) and Switzerland (Nussberger, Wandeler, Weber, & Keller, 2014). Otherwise, sampling of wildcats is often opportunistic, e.g., from roadkill. Methods to detect and quantify hybridisation vary, using a combination of morphology (contentious due to the phenotypic similarity between wildcats and their hybrids) and genetics (using microsatellite markers or diagnostics SNPs) (Kitchener, Yamaguchi, Ward, & Macdonald, 2005; Nussberger, Greminger, Grossen, Keller, & Wandeler, 2013). Hybridisation rates may therefore not be directly comparable between studies, however, current estimates (excluding Scotland) find between 3% and 25% of wild-living cats to be hybrids (Tiesmeyer et al., 2020; Urzi et al., 2021). The estimated proportion of hybrids is highly variable across different regions.

Hunting and habitat loss have restricted the wildcat range in Britain to the highlands of Scotland, where hybridisation with domestic cats is now the most significant threat facing the remaining population. Following the extinction of the lynx in the 7th century, wildcats are the only extant felid species in Britain (Hetherington, Lord, & Jacobi, 2006), and most endangered carnivore

species (Mathews et al., 2018). Hybridisation in the wild is extensive, and much higher than the European-wide average, for reasons that remain poorly understood (Beaumont et al., 2001; Senn et al., 2019). The population is now at serious risk of genetic swamping, i.e., complete replacement by feral domestic cats, in the immediate future (Breitenmoser et al., 2019).

Domestic cats are thought to have arrived in southern Europe as early as 4400 BCE (Ottoni et al., 2017), but did not become widespread in Britain until the Roman occupation, ending 410 CE (Serpell, 2014). Domestic cats and wildcats have therefore been sympatric in Britain for at least 2000 years, however, the history of admixture between the two species remains unknown. Without a comprehensive understanding of hybridisation history or dynamics, or the impact of introgressive hybridisation on fitness, conservation of this species in Britain is not straightforward. Accurate population estimates are difficult to obtain due to the elusive nature of the species and limited ability to distinguish hybrids in the field based on morphology (Breitenmoser et al., 2019). This problem is compounded by the lack of a baseline reference for Scottish wildcats. The difficulties inherent in distinguishing wildcat and hybrid phenotypes results in haphazard protection, impedes accurate monitoring, and undermines the Scottish wildcat's legal status as a protected species.

1.3 General motivation and aims

Hybridisation and introgression are important evolutionary processes, generating genetic diversity and driving both speciation and extinction (Barton, 2001). Human-mediated hybridisation, however, is recognised as a threat to wild populations, the evolutionary and ecological impacts of which are still not fully understood (Allendorf et al., 2001). Genomic data provide an invaluable resource to understand demographic history and the impact of hybridisation, and the opportunity to develop methodologies to support conservation of at-risk species or populations.

Within this context, and using the critically endangered Scottish wildcat population as a case study, I aim to use genomic data to give a detailed picture of hybridisation in Scottish wildcats, specifically to:

1. Assess population structure and current levels of hybridisation in Scotland
2. Ascertain the timescale and mode of introgression in Scottish wildcats, testing the hypothesis that no significant introgression from domestic cats occurred prior to the last 200 years

1.4 References

Alexander, T. J., Vonlanthen, P., & Seehausen, O. (2017). Does eutrophication-driven evolution change aquatic ecosystems? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372, 20160041

Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, 11(10), 697–709

Allendorf, F. W., Leary, R. F., Spruell, P., & Wenburg, J. K. (2001). The problems with hybrids: Setting conservation guidelines. *Trends in Ecology and Evolution*, 16(11), 613–622

An, S. Q., Gu, B. H., Zhou, C. F., Wang, Z. S., Deng, Z. F., Zhi, Y. B., … Liu, Y. H. (2007). Spartina invasion in China: Implications for invasive species management and future research. *Weed Research*, 47, 183–191

Anderson, E. (1948). Hybridization of the habitat. *Evolution*, 2, 1–9.

Anderson, E., & Stebbins, G. L. (1954). Hybridization as an Evolutionary Stimulus. *Evolution*, 8, 378–388.

Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences of the United States of America*, 115(25), 6506–6511

Barlow, A., Cahill, J. A., Hartmann, S., Theunert, C., Xenikoudakis, G., Fortes, G. G., … Hofreiter, M. (2018). Partial genomic survival of cave bears in living brown bears Axel. *Nature Ecology and Evolution*, 2(10), 1563–1570

Barton, N. H. (2001). The role of hybridization in evolution. *Molecular Ecology*, 10, 551–568

Barton, N. H., & Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, 16, 113–148

Beaumont, M., Barratt, E. M., Gottelli, D., Kitchener, A. C., Daniels, M. J., Pritchard, J. K., & Bruford, M. W. (2001). Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology*, 10(2), 319–336

Bell, B. (2002). The eradication of alien mammals from five offshore islands, Mauritius, Indian Ocean. In Veitch, C. R. & Clout, M. N. (Eds.), *Turning the tide: the eradication of invasive species* (pp. 40-45). IUCN SSC Invasive Species Specialist Group. IUCN, Gland Switzerland and Cambridge, UK

Blancher, P. P. (2013). Estimated number of birds killed by house cats (*Felis catus*) in Canada. *Avian Conservation Ecology*, 8, 3

Breitenmoser, U., Lanz, T., & Breitenmoser-Würsten, C. (2019). *Conservation of the wildcat (*Felis silvestris*) in Scotland: Review of the conservation status and assessment of conservation activities*. IUCN SSC. http://www.scottishwildcataction.org/media/42633/wildcat-in-scotland-review-of-conservation-status-and-activities-final-14-february-2019.pdf

Canters, K., Thissen, J. B. M., Diepenbeek, M. a J., Jansman, H. a H., & Goutbeek, K. (2005). The wildcat (*Felix silvestris*) finally recorded in the Netherlands. *Lutra*, 48(2), 67–90

Castello, J. R. (2020) *Felids and Hyenas of the World: Wildcats, Panthers, Lynx, Pumas, Ocelots, Caracals, and Relatives.* Princeton, USA: Princeton University Press

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signalled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences of the United States of America*, 114(30), E6089–E6096

Chan, W. Y., Hoffmann, A. A., & van Oppen, M. J. H. (2019). Hybridization as a conservation management tool. *Conservation Letters*, 12, e12652

Cooper G. M. (2000) *The Cell: A Molecular Approach*. (2nd ed.) Washington (DC): ASM Press

Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31992L0043&from=EN

Darwin, C. R. (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* London: John Murray

Darwin, C. R. (1876) *The effects of cross and self-fertilisation in the vegetable kingdom.* London: John Murray

Darwin, C. R. (1875) *The variation of animals and plants under domestication* (2nd ed.). London: John Murray

De Knijff, P. (2014). How carrion and hooded crows defeat Linnaeus's curse. *Science*, 344(6190), 1345–1346

de la Cruz, F., & Davies, J. (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends In Microbiology*, 8(3), 128–133

Dobzhansky, T. (1937) *Genetics and the Origin of Species*. New York: Columbia Univ. Press

Draper, D., Laguna, E., & Marques, I. (2021). Demystifying Negative Connotations of Hybridization for Less Biased Conservation Policies. *Frontiers in Ecology and Evolution*, 9, 637100

Driscoll, C. A., Macdonald, D. W., & O'Brien, S. J. (2009). From Wild Animals to Domestic Pets, and Evolutionary View of Domestication. In J. C. Avise & F. J. Ayala (Eds.), *In the Light of Evolution III: Two Centuries of Darwin* (pp. 89–109). Washington (DC): National Academies Press

Driscoll, C. A., Menotti-Raymond, M., Roca, A. L., Hupe, K., Johnson, W. E., Geffen, E., … Macdonald, D. W. (2007). The Near Eastern Origin of Cat Domestication. *Science*, 317(5837), 519–523.

Ellstrand, N. C., Prentice, H. C., & Hancock, J. F. (1999). Gene flow and introgression from domesticated plants into their wild relatives. *Annual Review of Ecology and Systematics*, 30, 539–563.

Epifanio, J., & Philipp, D. (2001). Simulating the extinction of parental lineages from introgressive hybridization: the effects of fitness, initial proportions of parental taxa, and mate choice. *Reviews in Fish Biology and Fisheries*, 10, 339–354

Fan, Z., Silva, P., Gronau, I., Wang, S., Armero, A. S., Schweizer, R. M., … Wayne, R. K. (2016). Worldwide patterns of genomic variation and admixture in gray wolves. *Genome Research*, 26, 163–173

Fitzpatrick, B. M., Ryan, M. E., Johnson, J. R., Corush, J., & Carter, E. T. (2015). Hybridization and the species problem in conservation. *Current Zoology*, 61(1), 206–216

Frankham, R., Ballou, J., & Briscoe, D. (2010) *Introduction to Conservation Genetics* (2nd ed.). Cambridge: Cambridge University Press

Gehrt, S. D., Riley, S. P. D., Cypher, B. L. (2010) *Urban Carnivores: Ecology, Conflict, and Conservation*. Baltimore, USA: John Hopkins University Press.

Gompert, Z., Fordyce, J. A., Forister, M. L., Shapiro, A. M., & Nice, C. C. (2006). Homoploid hybrid speciation in an extreme habitat. *Science*, 314, 1923–1925

Grabenstein, K. C., & Taylor, S. A. (2018). Breaking Barriers: Causes, Consequences, and Experimental Utility of Human-Mediated Hybridization. *Trends in Ecology and Evolution*, 33(3), 198–212

Grant, V. (1981) *Plant Speciation* (2nd ed.). New York: Columbia University Press

Grant, B. R., & Grant, P. R. (1996). High Survival of Darwin's Finch Hybrids: Effects of Beak Morphology and Diets, *Ecology*, 77(2), 500–509

Grossen, C., Keller, L., Biebach, I., Zhang, W., Tosser-Klopp, G., Ajmone, P., … Croll, D. (2014). Introgression from Domestic Goat Generated Variation at the Major Histocompatibility Complex of Alpine Ibex. *PLoS Genetics*, 10(6), e1004438

Harrison, R. G., & Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity*, 105, 795–809

Hausdorf, B. (2011). Progress toward a general species concept. *Evolution*, 65(4), 923–931

Hetherington, D. A., Lord, T. C., & Jacobi, R. M. (2006). New evidence for the occurrence of Eurasian lynx (*Lynx lynx*) in medieval Britain. *Journal of Quaternary Science*, 21(1), 3–8

Hoffmann, A. A., & Sgró, C. M. (2011). Climate change and evolutionary adaptation. *Nature*, 470, 479–485

International Commission on Zoological Nomencalture. (2003). Opinion 2027 (Case 3010). *Bulletin of Zoological Nomenclature*, 60, 81–84

International Union for Conservation of Nature (2020). Threats Classification Scheme (Version 3.2). Retrieved August 10, 2020, from https://www.iucnredlist.org/resources/threat-classificationscheme

Johnson, W. E., Onorato, D. P., Roelke, M. E., Land, E. D., Cunningham, M., Belden, R. C., … O'Brien, S. J. (2010). Genetic Restoration of the Florida Panther. *Science*, 9(1), 76–99

Johnson, W. E., Eizirik, E., Pecon-Slattery, J., Murphy, W. J., Antunes, A., Teeling, E., & O'Brien, S. J. (2006). The late miocene radiation of modern felidae: A genetic assesstment. *Science*, 311(5757), 73–77

Kitchener, A. C., Breitenmoser-Würsten, C., Eizirik, E., Gentry, A., Werdelin, L., Wilting, A., … Tobe, S. (2017). A revised taxonomy of Felidae. The final report of the Cat Classification Task Force of the IUCN/SSC Cat Specialist Group. *Cat News Special Issue 11*

Kitchener, A. C., Yamaguchi, N., Ward, J. M., & Macdonald, D. W. (2005). A diagnosis for the Scottish wildcat (*Felis silvestris*): A tool for conservation action for a critically-endangered felid. *Animal Conservation*, 8(3), 223–237

Kriwoken, L. K., & Hedge, P. (2000). Exotic species and estuaries: Managing *Spartina anglica* in Tasmania, Australia. *Ocean and Coastal Management*, 43, 573–584

Kuhlwilm, M., Gronau, I., Hubisz, M. J., De Filippo, C., Prado-Martinez, J., Kircher, M., … Castellano, S. (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*, 530, 429–433

Lamkey, K. R., Edwardsm J. W. (1999) Quantitative Genetics of Heterosis. In Coors, J. G. & Pandey, S. (Eds.) *The Genetics and Exploitation of Heterosis in Crops* (pp. 31-48). Madison: American Society of Agronomy, Inc. and Crop Science Society of America Inc.

Larsen, P. A., Marchán-Rivadeneira, M. R., & Baker, R. J. (2010). Natural hybridization generates mammalian lineage with species characteristics. *Proceedings of the National Academy of Sciences of the United States of America*, 107(25), 11447–11452

Le Roux, J. J., Foxcroft, L. C., Herbst, M., & MacFadyen, S. (2015). Genetic analysis shows low levels of hybridization between African wildcats (*Felis silvestris lybica*) and domestic cats (*F. s. catus*) in South Africa. *Ecology and Evolution*, 5(2), 288–299

Lexer, C., Welch, M. E., Raymond, O., & Rieseberg, L. H. (2003). The Origin of Ecological Divergence in *Helianthus paradoxus* (Asteraceae): Selection on Transgressive Characters in a Novel Hybrid Habitat. *Evolution*, 57(9), 1989–2000

Li, P., Su, T., Zhang, D., Wang, W., Xin, X., Yu, Y., … Zhang, F. (2021). Genome-wide analysis of changes in miRNA and target gene expression reveals key roles in heterosis for Chinese cabbage biomass. *Horticulture Research*, 8, 1–15

Loss, S. R., Will, T., & Marra, P. P. (2015). Direct mortality of birds from anthropogenic causes. *Annual Review of Ecology, Evolution, and Systematics*, 46, 99–120

Mack, R. N., Simberloff, D., Lonsdale, M. W., Evans, H., Clout, M., & Bazzaz, F. A. (2000). Biotic invasions: causes, epidemiology, global consequences and control. *Ecological Applications*, 10(3), 689–710

Mallet, J., Meyer, A., Nosil, P., & Feder, J. L. (2009). Space, sympatry and speciation. *Journal of Evolutionary Biology*, 22, 2332–2341

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology and Evolution*, 20(5), 229–237

Mallet, J. (2007). Hybrid speciation. *Nature*, 446(7133), 279–283

Mathews, F., Kubasiewicz, L. M., Gurnell, J., Harrower, C. A., McDonald, R. A., & Shore, R. F. (2018). *A Review of the Population and Conservation Status of British Mammals: Technical Summary*. A report by the Mammal Society under contract to Natural England, Natural Resources Wales and Scottish Natural Heritage. Natural England, Peterborough

Mattucci, F., Oliveira, R., Lyons, L. A., Alves, P. C., & Randi, E. (2016). European wildcat populations are subdivided into five main biogeographic groups: Consequences of Pleistocene climate changes or recent anthropogenic fragmentation? *Ecology and Evolution*, 6(1), 3–22

Mavárez, J., Salazar, C. A., Bermingham, E., Salcedo, C., Jiggins, C. D., & Linares, M. (2006). Speciation by hybridization in Heliconius butterflies. *Nature*, 441, 868–871

Mayr, E. (1942) *Systematics and the Origin of Species.* New York: Columbia University Press

Mayr, E. (1963). *Animal Species and Evolution*. Cambridge, MA: Harvard University Press

McFarlane, S. E., & Pemberton, J. M. (2019). Detecting the True Extent of Introgression during Anthropogenic Hybridization. *Trends in Ecology and Evolution*, 34(4), 315–326

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, 8(14363), 1–11

Miller, G. S. (1907). Some new European Insectivora and Carnivora. *Annals and Magazine of Natural History* (7th series) 20, 389-398.

Montague, M. J., Li, G., Golfi, B., Khan, R., Aken, B. L., Searle, S. M. J., … Warren, W. C. (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), 17230–17235

Moran, B. M., Payne, C., Langdon, Q., Powell, D. L., Brandvain, Y., & Schumer, M. (2021). The genomic consequences of hybridization. *ELife*, 10, e69016

Mueller, S. A., Reiners, T. E., Steyer, K., von Thaden, A., Tiesmeyer, A., & Nowak, C. (2020). Revealing the origin of wildcat reappearance after presumed long-term absence. *European Journal of Wildlife Research*, 66, 94

Novikova, P. Y., Brennan, I. G., Booker, W., Mahony, M., Doughty, P., Lemmon, A. R., … Donnellan, S. C. (2020). Polyploidy breaks speciation barriers in Australian burrowing frogs Neobatrachus. *PLoS Genetics*, 16(5), e1008769

Nowell, K., & Jackson, P. (1996). *Wild cats. Status Survey and Conservation Action Plan*. Gland, Switzerland: IUCN

Nussberger, B., Greminger, M. P., Grossen, C., Keller, L. F., & Wandeler, P. (2013). Development of SNP markers identifying European wildcats, domestic cats, and their admixed progeny. *Molecular Ecology Resources*, 13(3), 447–460

Nussberger, B., Wandeler, P., Weber, D., & Keller, L. F. (2014). Monitoring introgression in European wildcats in the Swiss Jura. *Conservation Genetics*, 15, 1219–1230

Ottoni, C., Van Neer, W., De Cupere, B., Daligault, J., Guimaraes, S., Peters, J., … Geigl, E.-M. (2017). The paleogenetics of cat dispersal in the ancient world. *Nature Ecology & Evolution*, 1, 0139

Pardo-Diaz, C., Salazar, C., Baxter, S. W., Merot, C., Figueiredo-Ready, W., Joron, M., … Jiggins, C. D. (2012). Adaptive introgression across species boundaries in Heliconius butterflies. *PLoS Genetics*, 8(6)

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Müller, I., … Wolf, J. B. W. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, 344(6190), 1410–1414

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., … Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505, 43–49

Quilodrán, C. S., Currat, M., & Montoya-Burgos, J. I. (2018). Effect of hybridization with genome exclusion on extinction risk. *Conservation Biology*, 32(5), 1139–1149

Quilodrán, C. S., Montoya-Burgos, J. I., & Currat, M. (2020). Harmonizing hybridization dissonance in conservation. *Communications Biology*, 3, 1–10

Quispe-Huamanquispe, D. G., Gheysen, G., & Kreuze, J. F. (2017). Horizontal gene transfer contributes to plant evolution: The case of agrobacterium T-DNAs. *Frontiers in Plant Science*, 8, 1–6

Rhymer, J. M., Williams, M. J., & Braun, M. J. (1994). Mitochondrial analysis of gene flow between New Zealand mallards (*Anas platyrhynchos*) and Grey Ducks (*A. superciliosa*). *The Auk*, 111(4), 970–978

Scandura, M., Iacolina, L., Crestanello, B., Pecchioli, E., Di Benedetto, M. F., Russo, V., … Bertorelle, G. (2008). Ancient vs. recent processes as factors shaping the genetic variation of the European wild boar: Are the effects of the last glaciation still detectable? *Molecular Ecology*, 17, 1745–1762

Schwenk, K., Brede, N., & Streit, B. (2008). Extent, processes and evolutionary impact of interspecific hybridization in animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363, 2805–2811

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends in Ecology and Evolution*, 19(4), 198–207

Senn, H. V., Ghazali, M., Kaden, J., Barclay, D., Harrower, B., Campbell, R. D., … Kitchener, A. C. (2019). Distinguishing the victim from the threat: SNP-based methods reveal the extent of introgressive hybridization between wildcats and domestic cats in Scotland and inform future in situ and ex situ management options for species restoration. *Evolutionary Applications*, 12(3), 399–414

Serpell, J. A. (2014). Domestication and history of the cat. In Turner, D. C. & Bateson, P. (Eds.), *The Domestic Cat: The Biology of its Behaviour* (3rd ed., pp. 83–100). Cambridge, UK: Cambridge University Press

Servedio, M. R., & Noor, M. A. F. (2003). The Role of Reinforcement in Speciation: Theory and Data. *Annual Review of Ecology, Evolution, and Systematics*, 34, 339–364

Smith, J., & Kronforst, M. R. (2013). Do Heliconius butterfly species exchange mimicry alleles? *Biology Letters*, 9, 20130503

Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F., Shih, C., … Kohn, M. H. (2012). Adaptive introgression of anticoagulant rodent poison resistance by hybridization between Old World mice. *Current Biology*, 21(15), 1296–1301

Sovic, M. G., Fries, A. C., & Gibbs, H. L. (2016). Origin of a cryptic lineage in a threatened reptile through isolation and historical hybridization. *Heredity*, 117, 358–366

Species at Risk Act S. C. 2002, c. 29. Available at https://laws-lois.justice.gc.ca/PDF/S-15.3.pdf

Stebbins, G. L. (1959). The Role of Hybridization in Evolution. *Proceedings of the National American Philosophical Society*, 103(2), 231–251

Sunquist, M., Sunquist, F. (2002) *Wild Cats of the World.* Chicago: The University of Chicago Press

Svardal, H., Quah, F. X., Malinsky, M., Ngatunga, B. P., Miska, E. A., Salzburger, W., … Durbin, R. (2020). Ancestral hybridization facilitated species diversification in the lake malawi cichlid fish adaptive radiation. *Molecular Biology and Evolution*, 37(4), 1100–1113

Thompson, J. D. (1991). The biology of an invasive plant. *BioScience*, 41(6), 393–401

Tiesmeyer, A., Ramos, L., Manuel Lucas, J., Steyer, K., Alves, P. C., Astaras, C., … Nowak, C. (2020). Range-wide patterns of human-mediated hybridisation in European wildcats. *Conservation Genetics*, 21, 247–260

Todesco, M., Pascual, M. A., Owens, G. L., Ostevik, K. L., Moyers, B. T., Hübner, S., … Rieseberg, L. H. (2016). Hybridization and extinction. *Evolutionary Applications*, 9, 892–908

Trouwborst, A., McCormack, P. C., & Martínez Camacho, E. (2020). Domestic cats and their impacts on biodiversity: A blind spot in the application of nature conservation law. *People and Nature*, 2, 235–250

Urzi, F., Šprem, N., Potočnik, H., Sindičić, M., Konjević, D., Ćirović, D., … Buzan, E. (2021). Population genetic structure of European wildcats inhabiting the area between the Dinaric Alps and the Scardo-Pindic mountains. *Scientific Reports*, 11

US Fish and Wildlife Service (1996) Endangered and threatened wildlife and plants; proposed policy and proposed rule on the treatment of intercrosses and intercross progeny (the issue of 'hybridization'); Request for public comment. *Federal Register*, 61, 4710–4713

Wayne, R. K., & Shaffer, H. B. (2016). Hybridization and endangered species protection in the molecular era. *Molecular Ecology*, 25, 2680–2689

Wells, C. P., Lavretsky, P., Sorenson, M. D., Peters, J. L., DaCosta, J. M., Turnbull, S., … Engilis, A. (2019). Persistence of an endangered native duck, feral mallards, and multiple hybrid swarms across the main Hawaiian Islands. *Molecular Ecology*, 28, 5203–5216

Woinarski, J. C. Z., Murphy, B. P., Palmer, R., Legge, S. M., Dickman, C. R., Doherty, T. S., … Stokeld, D. (2018). How many reptiles are killed by cats in Australia? *Wildlife Research*, 45, 247-266

Woinarski, J., Murphy, B. P., Legge, S. M., Garnett, S. T., Lawes, M. J., Comer, S., … Woolley, L. A. (2017). How many birds are killed by cats in Australia? *Biological Conservation*, 214, 76–87

Wu, C. I., & Ting, C. T. (2004). Genes and speciation. *Nature Reviews Genetics*, 5(2), 114–122

Yamaguchi, N., Kitchener, A., Driscoll, C., & Nussberger, B. (2015). *Felis silvestris. The IUCN Red List of Threatened Species 2015*, e.T60354712A50652361

Yu, H., Xing, Y. T., Meng, H., He, B., Li, W. J., Qi, X. Z., … Luo, S. J. (2021). Genomic evidence for the Chinese mountain cat as a wildcat conspecific (*Felis silvestris bieti*) and its introgression to domestic cats. *Science Advances*, 7, eabg0221

Chapter 2 Current status of the Scottish wildcat

2.1 Introduction

2.1.1 Status of the wildcat in Britain

Historically, wildcats were widespread in Britain and persisted in England and Wales until the late 19th century (Langley & Yalden, 1977). The combined pressures of persecution and habitat loss have resulted in a dramatic decline and range constriction over the last few centuries (Fig. 2.1). Wildcats were hunted for sport during the Mediaeval period and are likely to have been extirpated from southern England as early as the 16th century. Hunting (for sport, fur or as pests) and habitat loss (particularly forest habitat) continued to drive decline during the 17th and 18th centuries. The establishment of many sporting estates during the 19th century, often for gamebirds such as grouse and pheasants, increased persecution from gamekeepers and accelerated decline. As late as 1984-1985 the Game Conservancy's Vermin Returns reported 274 wildcats killed across 40 estates in Scotland (Breitenmoser et al., 2019). The presence of wildcats on or near sporting estates creates conflict that continues today.

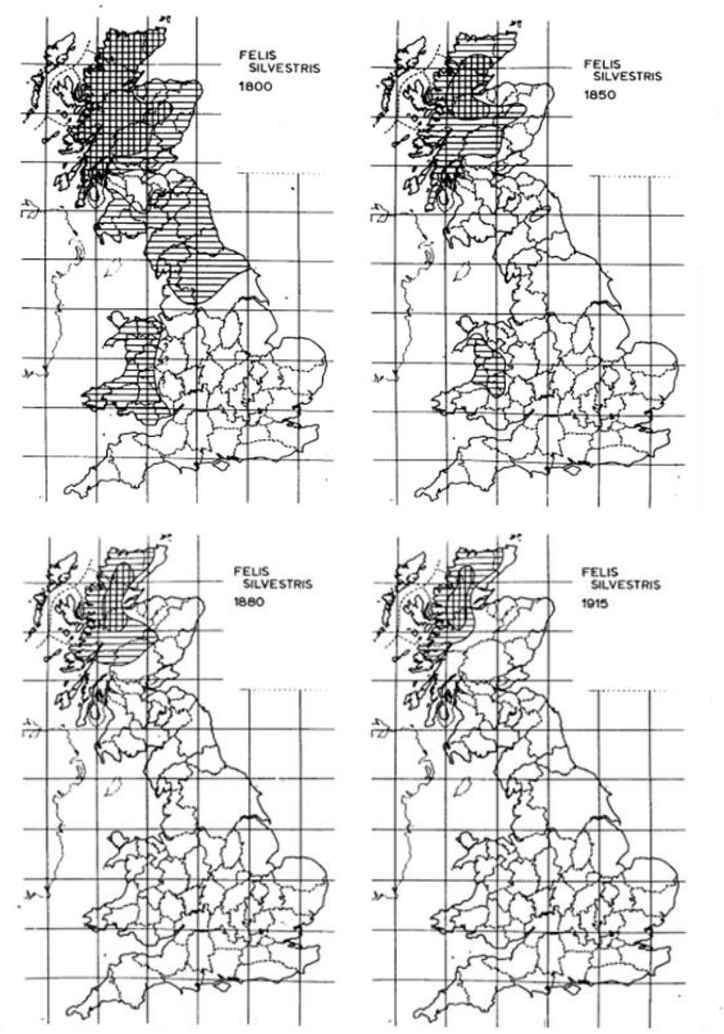

Figure 2.1. The distribution of wildcats in Britain 1800-1915. Wildcats were experiencing dramatic declines before the 19th century and were already extirpated from southern and central England. They were extinct in England and Wales by 1880. The wildcat range at its smallest was observed in 1915, limited to the northwest highlands of Scotland.

From Langley & Yalden (1977)

By the start of the 20th century, wildcat range in Britain was limited to the north-west highlands of Scotland (Langley & Yalden, 1977) (Fig. 2.2). Wildcats probably persisted in this area due to its remoteness, i.e., low human density, rather than high quality habitat (Easterbee et al., 1991). The wildcat population is thought to have reached its lowest level around

1915. Following the conscription of many gamekeepers during World War I and the establishment of the Forestry Commission in 1919, pressure from persecution was eased and areas of fast-growing, coniferous woodland were planted. The wildcat population appears to have recovered quickly, expanding into central and eastern Scotland. By the 1940s wildcats had re-established much of their current range (Fig. 2.2). Range expansion had slowed by the mid-20th century as all suitable habitat north of the central belt became occupied. The central belt is the most industrialised and densely populated area of Scotland, running between Glasgow in the west and Edinburgh in the east; it is considered a firm boundary to population expansion.

A major survey of wildcats was carried out by the Nature Conservancy Council (NCC) between 1983 and 1987 (Easterbee et al., 1991). Records were collected from 499 locations across Scotland, relying mostly on sightings, supplemented by data from road traffic accidents. A definition for 'wildcat' (versus feral or hybrid cat) was not used. A more recent survey carried out by NatureScot (formerly Scottish Natural Heritage, SNH) between 2006 and 2008, incorporated pelage information to verify wildcat sightings (Davis & Gray, 2010). Both surveys reported a wide distribution across Scotland (north of the central belt), with wildcats more prevalent in the east than the west. Davis & Gray (2010) proposed strongholds in the Cairngorms, the Black Isle and Aberdeenshire in the east, and Ardnamurchan in the west. Easterbee *et al.* (1991) reported a low density of wildcats, even in suitable habitat. In the more recent survey by Davis and Gray (2010), the population appeared fragmented, and declining in the west. Differences between the two surveys may be due to variation in survey methodology, including the use of pelage characteristics to identify wildcats.

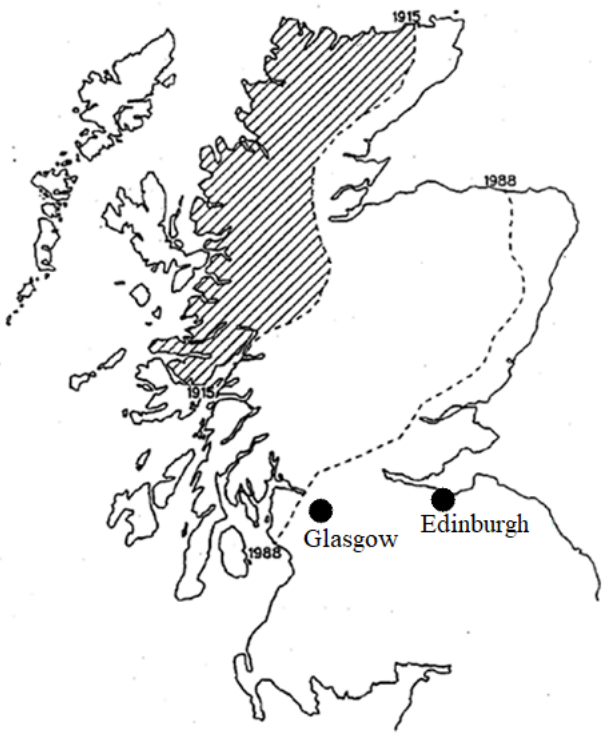

Figure 2.2. The shaded area shows the wildcat range at its smallest in 1915. The second dashed line shows the limit of 20th century range expansion, as estimated by the NCC 1983-1987 survey (Easterbee et al., 1991). This broadly corresponds to the current range in Scotland. The central belt between Glasgow and Edinburgh is a barrier to expansion further south.

Adapted from Balharry & Daniels (1993)

An extensive camera trapping survey carried out by Kilshaw *et al.* (2016) between 2010 and 2013 deployed 546 camera traps at 27 sites across the Scottish Highlands. Wildcats were distinguished from hybrids and feral domestic cats using a pelage score (Kitchener et. al., 2005) (see 2.1.3). An occupancy model was developed predicting wildcat presence across Scotland as a function

of habitat co-variates (such as land cover type, or distance to urban/suburban areas). In concurrence with Easterbee *et al.* (1991) and Davis & Gray (2010), higher occupancy was predicted in the central and eastern Highlands, the edges of the Cairngorms National Park, coastal regions in the western Highlands and some pockets in the far north.

Estimates of population size have fallen dramatically in recent decades (though see below). In 1995, Harris *et al.* estimated the wildcat population to be 3,500 breeding individuals. The same report estimated the feral cat population in Scotland to be 130,000 (Harris, Morris, Wray, & Yalden, 1995). In 2015, the IUCN Red List evaluation reported an estimated 400 individuals in Scotland that met genetic and morphological criteria to qualify as "the furthest from the domestic" (Yamaguchi et al., 2015). Recent estimates are around 200 individuals (Kilshaw, Johnson, Kitchener, & Macdonald, 2015; Mathews et al., 2018) and declining (Breitenmoser et al., 2019). The feline welfare charity, Cats Protection, estimate there are now 1.5 million feral cats in the UK.

Hybridisation limits our ability to obtain an accurate picture of population distribution or size. Changing approaches to identify wildcats may obscure trends in time series data. It is difficult to understand whether revised recent estimates reflect genuine decline, or an improved understanding of hybridisation and stricter criteria to identify wildcats. Either way, it is clear that the current population of putative wildcats in Britain is small, with a fragmented distribution across the Scottish Highlands.

As discussed in Chapter 1, hybridisation is now the biggest threat to the remaining Scottish wildcat population (Mathews et al., 2018; Sainsbury et al., 2019; Senn et al., 2019). Domestic and hybrid cats are prevalent across wildcat range (Kilshaw et al., 2016), and introgression appears to be extensive (Senn et al., 2019). Wildcats are at serious risk of genetic replacement by hybrids in the wild. In a 2017/18 survey of wildcat conservation "Priority Areas" (Littlewood et al., 2014) the ratio of un-neutered hybrids to wildcats was estimated at 6:1 (Breitenmoser et al., 2019). Hybridisation impedes effective conservation and results in haphazard legal protection. Domestic cats also pose a disease transmission risk to wildcats; many common feline infectious diseases have been detected in the wild-living population, including feline immunodeficiency virus, feline calicivirus and *Mycoplasma haemofelis* (Meredith et al., 2018). Feral domestic cats compete with wildcats for resources, e.g. shelter or prey, both of which are already limited or in decline (Breitenmoser et al., 2019). Between 1995 and 2002 there was 57.3% decline in the rabbit population in Scotland, the main prey species for wildcats (Battersby, 2005).

Persecution of wildcats, and accidental killing by snares or poisoned bait, continues in Scotland, especially as a pest species on sporting estates (Breitenmoser et al., 2019; Yamaguchi et al., 2015). Wildcats are a protected species in the UK, but feral domestic cats can be controlled legally. Inherent difficulties distinguishing wildcats, hybrids and feral domestics based on phenotype results in

ineffectual legal protection. The current impact of persecution or accidental killing on the wildcat population is unknown (Sainsbury et al., 2019).

Habitat loss and fragmentation remain a concern (Breitenmoser et al., 2019). The proportion of woodland in Scotland has gradually increased from ~5% in 1900 to 17% by the early 2000s, but this is still below the European-wide average of ~38% (Forest Research, 2021). Wildcats require a mosaic habitat of woodland and open areas for shelter and hunting (Breitenmoser et al., 2019). They are generally absent from urban areas, human settlements, or areas of intensive agriculture, all of which have also increased during the 20th century. Transport networks associated with urban development may further degrade or fragment habitat and traffic collisions are an additional threat to wildcats.

2.1.2 Wildcat conservation and management

The IUCN classifies wildcats as Least Concern on the Red List of Threatened Species (Yamaguchi et al., 2015). It notes that some populations are at risk of local extinction and, if accurate, recent estimates would classify the Scottish population as Critically Endangered.

The first concerted effort to monitor wildcats in Scotland was the 1983-1987 NCC survey (Easterbee et al., 1991). A conservation action plan for the species in Britain was not published until 2004 (Macdonald et al., 2004), where hybridisation and introgression were identified as threats to wildcats. Recommendations from this report were not implemented, but influenced subsequent conservation work (Breitenmoser et al., 2019). The Species Action Framework (2007-2012) targeted 32 species, including the Scottish wildcat, and funded work to develop and test field survey methods for wildcats. Under the Species Action Framework the Cairngorms Wildcat Project (2009-2012) was developed. The Cairngorms Wildcat Project was the first practical trial of wildcat conservation, operating within the Cairngorms National Park. Specifically, the main activities were a public awareness and engagement campaign, neutering and vaccination of domestic cats, close working with estates to improve wildcat identification and wildcat-friendly predator control, and research and monitoring (Hetherington & Campbell, 2012).

In 2013 the Scottish Wildcat Conservation Action Plan (SWCAP) was published (Scottish Natural Heritage, 2013). This was the first national plan for wildcat conservation, with the objective of halting decline within six years. The SWCAP set out to protect a group of cats that "look like wildcats, but may not all be genetically pure wildcats". For this, the SWCAP aimed to identify Priority Areas for wildcat conservation, carry out conservation in these areas (as trialled by the Cairngorms Wildcat Project), as well as develop a wider conservation programme.

The SWCAP was implemented by Scottish Wildcat Action (SWA, www.scottishwildcataction.org) between 2013 and 2019 (Breitenmoser et al., 2019). SWA worked in

five wildcat Priority Areas (PAs) identified by Littlewood *et al.* (2014): Morvern, Strathpeffer, Northern Strathspey, Strathbogie and the Angus Glens. PAs were putative strongholds of the remaining population, with sufficient habitat to support twenty adult females (~4,000 ha). *In situ* conservation work aimed to reduce the risk of hybridisation and disease transmission, implementing a Trap Neuter Vaccinate and Return (TNVR) programme for feral domestic and hybrid cats and promoting responsible pet ownership (Breitenmoser et al., 2019). Feral and hybrid cats were identified by pelage, as per Kitchener *et al.* (2005), with a cut-off of 17 (see 2.1.3). During the first season of TNVR in 2016/17 90 cats were treated across the PAs. During the same period one wildcat was trapped and released. SWA also worked to promote wildcat-friendly estate management, specifically, forestry practice and predator control. Monitoring was carried out, including camera trap surveys in all PAs.

SWA also worked *ex situ*, improving the management of the captive breeding programme (Barclay, 2019). The captive wildcat population in the UK was established in the 1960s and has historically remained small. In 2015 the Royal Zoological Society of Scotland (RZSS), an SWA partner organisation, took over the management of the captive population, which consisted of 64 individuals, 7% of which had a known pedigree. Genetic screening (using 2,230 SNPs) took place between 2015 and 2017 to reconstruct a molecular pedigree for the population, which is now 100% known. This is important to monitor genetic diversity and prevent inbreeding (Lacy 1994). Using pedigree information, the number of founders was estimated to be 30. Genetic screening also allowed hybridisation levels to be assessed, resulting in the removal of two individuals from the breeding programme. Pelage was scored independently, and all information incorporated into the studbook. Since 2015 the population has expanded to 107 individuals, including two additional wild founders from outside the SWA PAs. The number of captive holders has increased from 20 to 27. Detailed studbook records are available for all individuals (Barclay, 2019).

A 2019 IUCN review of wildcat conservation in Scotland suggested the SWCAP be revised with reintroduction or reinforcement in mind (Breitenmoser et al., 2019). Work carried out by the Cairngorms Wildcat Project and SWA has improved understanding of hybridisation, which seems far more extensive than previously thought. It now seems that the SWCAP aim to halt decline by 2020 was unachievable from the outset. The IUCN considered the wild-living population too small, fragmented and hybridised to be considered viable. They concluded it was now too late to conserve the Scottish wildcat as a phylogenetic unit, and strongly recommended translocation of wildcats from continental Europe (reinforcement or reintroduction will also require rigorous suppression of feral cats and hybrids).

Following on from SWA, an updated strategy for wildcat conservation is based around population reinforcement. The Saving Wildcats Project (www.savingwildcats.org.uk) started in 2020

with the aim of re-establishing "a genetically and demographically viable wildcat population in the highlands of Scotland through threat mitigation and reinforcement" (EU LIFE, 2021). This will include establishing a conservation breeding for release facility, and control of hybridisation in the release area through TNVR of feral domestic cats.

2.1.3 Monitoring hybridisation

The ability to accurately identify and monitor wildcats, and quantify levels of hybridisation, is vital to wildcat conservation. As methods to distinguish wildcats have developed, it has become increasingly clear that there is extensive hybridisation in the wild-living population (Beaumont et al., 2001; Kilshaw et al., 2015; Senn et al., 2019). It is now more important than ever to be able to accurately identify the remaining wildcats in Scotland (to conserve the remaining gene-pool) and also hybrids, in order to understand hybridisation dynamics.

Identification was initially based on morphology, specifically pelage characteristics. Several morphological characteristics are known to differentiate wildcats and domestics cats, e.g., intestinal length (Schauenberg, 1977), cranial index (Schauenberg, 1969) and skull characteristics (French, Corbett & Easterbee, 1988), though these are obviously not useful to survey living cats. A non-invasive methodology was developed by Kitchener *et al.* (2005) using 135 museum specimens of 20th century, wild-living cats. Twenty pelage characteristics were examined (Fig. 2.3), scoring each on an ordinal scale of 1, 2, or 3 for domestic, hybrid or wildcat features, respectively. The combined pelage score classified individuals into three groups, including a group of putative wildcats. Seven characteristics were identified as the most diagnostic for wildcats. Applying these to 187 wild-living cats sampled in the 1990s (Daniels et al., 1998), Kitchener *et al.* (2005) found 12% were classified as wildcats, and 50% closest to the domestic group. The twenty pelage characteristics described by Kitchener *et al.* (2005) were tested on a sample of German wildcats, of which three were found to be diagnostic: tail banding, stripes on the neck and stripes on the shoulder (Krüger, Hertwig, Jetschke, & Fischer, 2009).

The seven-point pelage scoring system (7PS) has been widely adopted by the conservation programme to identify wildcats in the field. Putative wildcats score 19 or higher on this test (maximum score 21), though a lower threshold of 17 can be used to overcome possible recorder error, e.g., from poor quality camera-trap photos. In the 2006-2008 survey pelage information was used to help classify sightings as 'probable' versus 'possible' wildcats (Davis & Gray, 2010). Kilshaw & Macdonald (2011) successfully trialled camera-trapping as a method to monitor wildcats, using a 7PS score of ≥14 (and no scores of one) to identify wildcats. The 7PS score was a key diagnostic tool for SWA *in situ* work, for both camera-trap surveys and TNVR.

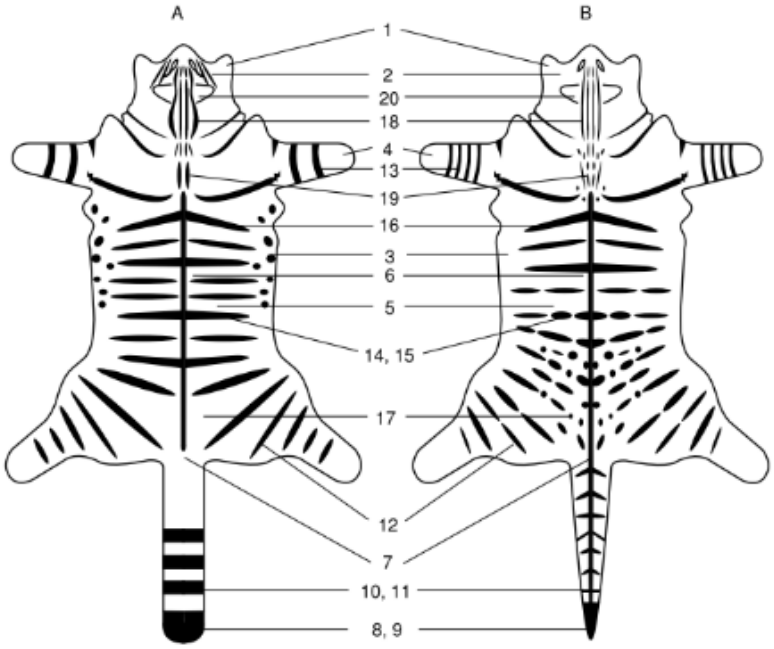

Figure 2.3. Pelage characteristics examined by Kitchener *et al.* (2005). A typical wildcat pelage is shown on the left (A), domestic tabby cat on the right (B). Seven traits were most diagnostic and are used to identify hybrids: the extent of the dorsal line (7), tail tip shape (8), tail banding (10), stripes/broken stripes (15) and spots (17) on hindquarters, stripes on nape (18) and shoulder (19).

From Kitchener *et al.* (2005)

The first major survey of hybridisation in Scotland using genetic markers was by Beaumont *et al.* (2001). Data at nine microsatellite loci were obtained from 230 wild-living Scottish cats (including 13 museum samples) and 74 British house cats, sampled between 1989 and 1994. All cats received a pelage score (assessed by Andrew Kitchener.) based on five characteristics. Beaumont *et al.* (2001) found evidence for three genetic groups of wild-living cats: domestics, putative wildcats and intermediates (hybrids). The putative wildcat group contained all of the individuals judged to be wildcats based on morphology, but also contained individuals classified as domestic. Overall, morphology and genetics were moderately correlated. In the first application of the program STRUCTURE (Pritchard, Seielstad, & Feldman, 1998), individuals were assigned a Q value between 0 and 1, higher values corresponding to individuals with more wildcat ancestry. Of the 230 individuals sampled from the wild in Scotland, 95 had a Q score greater than 0.9 (putative wildcats) and 96 had a Q score between 0.1 and 0.9 (hybrid) (Beaumont et al., 2001).

A genetic test for wildcats was developed in 2015 (Senn & Ogden, 2015). From a panel of 83 diagnostic SNPs developed to monitor hybridisation in the Swiss Jura wildcat population (Nussberger, Greminger, Grossen, Keller, & Wandeler, 2013), a subset of 35 SNPs was tested in Scotland. 82 individuals, including domestic, hybrid and wildcat representatives from the Swiss population, and wild-living and captive Scottish cats, were used to test the application of these markers in Scotland (Senn & Ogden, 2015). STRUCTURE Q values (Pritchard et al., 1998) and associated 90%

confidence interval (lower and upper bound, LBQ and UBQ, respectively), were generated for each individual (Senn & Ogden, 2015). The subset of 35 SNPs was able to detect up to a third-generation backcross. Based on Q score, a genetic continuum was observed between wildcats and domestic cats in Scotland, described as a 'hybrid swarm' (Mayr, 1963). For management purposes, a cut-off of LBQ≥0.75 was used to distinguish wildcats from hybrids. Senn & Ogden (2015) reported a weak correlation between 35 SNP and 7PS scores, and describe a decision matrix for selecting individuals to incorporate into the captive breeding programme using the two tests as separate lines of evidence.

The morphological (7PS) and genetic (35 SNP) tests are now routinely used to monitor both the wild-living and captive populations in Scotland. A recent study by Senn *et al.* (2019) screened 295 individuals. Using the 35 SNP test, 21 out of 144 wild-living samples were classified as wildcats and 106 as hybrids. 63 out of 72 captive individuals screened were classified as wildcats. For a subset of individuals with available morphological information, a weak correlation between genetic and pelage scores was reported; for 75 individuals classified as 35 SNP 'wildcats', 58 would also have been classified as wildcats based on pelage. ddRAD-seq data (3,097 SNPs) were generated for a subset of samples; Q scores from these data were generally found to be within the confidence intervals of the 35 SNP test.

Outside of Scotland, a of combination of morphology and genetics is also used to monitor wildcat hybridisation, though a lack of standardisation limits comparisons between studies. Data from microsatellite markers and mitochondrial DNA sequencing are often used (Hertwig et al., 2009; Steyer et al., 2016; Velli, Bologna, Silvia, Ragni, & Randi, 2015). Larger panels of SNP markers are increasingly common, e.g., Oliveira *et al.* (2015), including application of the Nussberger *et al.* (2013) panel, e.g. in Germany and Luxembourg (Steyer, Tiesmeyer, Muñoz-Fuentes, & Nowak, 2018).

2.1.4 Aims

The wildcat population has faced a long history of habitat loss and persecution in Britain (Langley & Yalden, 1977). Hybridisation in now the biggest threat facing this species, though it remains unclear to what extent this has historically affected the population. Urgent conservation intervention is now required if we wish to retain this species in the UK (Breitenmoser et al., 2019). Conservation management since the early 2000s has developed survey methods and diagnostic tests for wildcats, increased public awareness and engagement, and coordinated management of the *ex situ* population.

The aim of this chapter is to give an overview of the current status of the wildcat in Scotland, expanding on the work of Senn *et al.* (2019) with a larger number of both samples and genetic markers. Firstly, we clarify population structure using a two-fold increase in the number of genetic markers compared to Senn *et al.*, (2019). For this we use ddRAD-seq data; ddRAD-seq is an efficient way to sample thousands of markers for genome-wide estimates of hybridisation (Peterson, Weber,

Kay, Fisher, & Hoekstra, 2012). Increasing the number of markers increases power to accurately identify complex hybrids and backcrosses (McFarlane & Pemberton, 2019), giving the greatest resolution to date of the hybrid swarm in Scotland. We also use the expanded set of markers to evaluate the effectiveness of current tests to identify hybrid individuals.

2.2 Methods

2.2.1 ddRAD-seq dataset

ddRAD-seq data were generated for 129 individuals sampled between 1996 and 2017 (Senn et al., 2019). This included 71 individuals from the UK captive wildcat population, 53 individuals from the wild in Scotland (22 SWA trapped cats, 31 roadkill samples, see Fig. 2.4) and five Scottish domestic cats (domestic shorthairs). Blood samples were taken from captive and trapped cats, and tissue samples from roadkill specimens. For a full list of individuals see Table 2.3, Appendix 1.

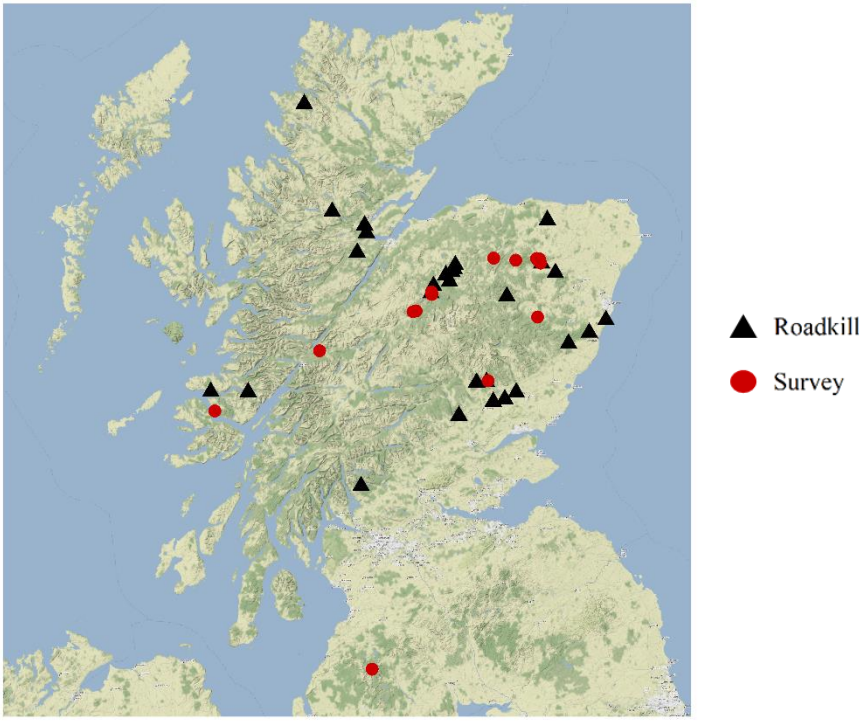

Figure 2.4. Sampling locations, where known, for wild-living individuals included in this study. Wild-living samples were collected from both SWA survey cats and from roadkill specimens.

This study represents a new bioinformatic analysis of the sequence reads produced by Senn *et al.* (2019), incorporating an additional 51 captive and two wild individuals, as well as the 76 original samples. Sequence reads were generated using the Illumina MiSeq Platform, as described in Senn *et al.* (2019). As per Senn *et al.* (2019), reads were demultiplexed by barcode and quality filtered using

the STACKS module, process_radtags (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013). Demultiplexed reads were trimmed to 135bp and concatenated into a single read file per individual. Note that historical wildcat samples (derived from museum specimens) reported in Senn *et al.* (2019) could not be used for this study due to poor DNA quality. Analysis of raw sequence reads diverges from that of Senn *et al.* (2019) from this point forward (described below), significant differences include alignment of reads to the domestic cat reference genome, a lower read depth threshold to identify loci using STACKs and stringent filtering of missing data.

2.2.2 Data processing

Sequence reads were aligned using BWA (Burrows-Wheeler Aligner, Li & Durbin, 2009) to the *Felis catus* reference genome v9.0 (GCF_000181335.3) (Buckley et al., 2020; Pontius et al., 2007). BWA is an efficient method to align short read sequences to longer reference sequences, allowing gaps or mismatches (Li & Durbin, 2009). The proportion of mapped reads appeared to be high, even for putative wildcat samples (Table 2.4, Appendix 2). Read alignment and linkage disequilibrium (LD) was checked visually using Haploview (Barrett, Fry, Maller, & Daly, 2005). A proportion of pairwise comparisons were affected by LD, but this was judged to be small and unlikely to affect downstream analysis. Mapped reads were processed using STACKS v2.1 (Catchen et al., 2013) . In STACKs a minimum of three reads were required to form a 'stack'. Multiple SNPs were allowed per read, the mean number of SNPs per read across the final dataset was 1.6. Variants were filtered using a minimum minor allele frequency of 0.05 and maximum proportion of heterozygotes of 0.7, treating the three sample sources (domestic, wild-living, and captive) as separate populations.

PLINK v1.9 (Chang et al., 2015) and VCFtools v1.15 (Danecek et al., 2011) were used to filter data from STACKs. Specifically, this led to the removal of individuals with >30% missing data and stringent subsequent filtering of loci to remove all sites with missing data. Closely related individuals (up to third degree relatives) were identified using IBD estimates calculated by PLINK, corrected to account for admixture using the method described by Morrison (2013). Corrected IBD estimates were used as input for PRIMUS (Staples et al., 2014) which uses genetic data to reconstruct pedigrees up to third degree relatives (e.g., Fig. 2.8, Appendix 2). Individuals were then removed from the dataset to limit relatedness (for the full list of excluded individuals see Table 2.3, Appendix 1).

Population genetic summary statistics (observed and expected heterozygosity, H_O and H_E, inbreeding coefficient, F_{IS}, and population pairwise F_{ST}; Weir & Cockerham, 1984) were generated for the final dataset using PLINK and VCFtools. H_O and H_E are common measures of genetic diversity, summarising the observed and expected number of heterozygous sites per individual (the population mean is reported here). Inbreeding coefficients were generated using PLINK's method of moments estimate. Wright's F-statistics (Wright, 1931) summarise the partitioning of genetic

variation within and between populations, and are among the most commonly used descriptive statistics in population genetics (Holsinger & Weir, 2009). F_{ST} is related to the variance of allele frequencies between populations and is an important summary of genetic differentiation as a result of population structure.

2.2.3 Population structure

Principal component analysis (PCA) and ADMIXTURE (Alexander, Novembre, & Lange, 2009) were used to examine population structure. PCA was completed in R using *prcomp*. ADMIXTURE analyses were performed for seven values of K, ranging from two to eight, and included a calculation of cross-validation error to estimate the optimal value of K. All SNPs were included, the data were not considered dense enough to require thinning (to minimise background LD) prior to the analysis (Alexander et al., 2009).

2.2.4 Evaluating hybrid tests

Hybrid individuals are currently identified using a combination of genetic and morphological diagnostic tests: the seven-point pelage scoring system (Kitchener, Yamaguchi, Ward, & Macdonald, 2005) and a 35 SNP genetic test (Senn & Ogden, 2015). The pelage test (7PS) scores seven key morphological characteristics, with putative wildcats scoring 19 or higher (maximum score 21). A lower threshold of 17 can be used to mitigate recorder error. The genetic test uses 35 SNPs that differentiate between wildcats and domestic cats (Nussberger et al., 2013; Senn & Ogden, 2015). A STRUCTURE LBQ score (i.e. the lower boundary of the Q35 value 90% CI) of 0.75 is proposed as the threshold to class individuals as putative wildcats, as distinct from hybrids (Senn & Ogden, 2015). Individuals with an LBQ≥0.75 are currently considered wildcats from a conservation management perspective.

We assessed the performance of these hybrid tests using ADMIXTURE Q values from the ddRAD-seq data (i.e., the Q value based on 6546 SNPs, Q6546) to determine hybrid status. Performance was assessed using receiver operating characteristic (ROC) curves, which evaluate the trade-off between sensitivity (ability to identify true positives) and specificity (ability to identify true negatives) of diagnostic tests (Fawcett, 2006). ROC curves were drawn in R using *pROC* (Robin, et al., 2011). None of the 35 SNPs from the genetic test were present in the ddRAD-seq data. Data were only included from individuals where both 35 SNP and pelage scores were available (n=59). The aim of this analysis was to compare the performance of these tests with diagnoses from a relatively dense marker set. Given the continuum of Q values observed in wild-living cats, a strict threshold (Q6546≥0.9) was used to select reference wildcat samples, but we recognise that this threshold is somewhat arbitrary and does not necessarily denote 'true wildcat' status. Individuals with Q6546≥0.9 were classified as wildcat reference samples, and those below 0.9 as hybrids. (Note the

threshold for the genetic test used by the conservation program, LBQ35≥0.75, is a management decision, and a higher threshold was used here to select reference samples for ROC analysis). Given the reference diagnosis, the true positive and false positive rates were calculated for both diagnostic tests at all possible threshold values. Plotting false positive rate against true positive rate (specificity vs sensitivity) for each classification threshold generated an ROC curve for each test (Robin et al., 2011). The area under the curve (AUC) is equivalent to the probability a test will rank a random positive instance higher than a random negative instance and is a useful metric to compare diagnostic tests. An AUC of 0.5 is essentially a random guess and an AUC of less than 0.5 is worse than random.

2.3 Results

2.3.1 Data processing

The final dataset included 108 individuals: four Scottish domestic cats and 104 putative wildcats (45 wild individuals and 59 from the captive population), genotyped at 6,546 SNPs. 21 samples were excluded from the analysis to minimise relatedness in the dataset and/or as a result of stringent filtering of missing data. Population summary statistics are given in Table 2.1.

Table 2.1. Summary statistics for the three source populations: captive wildcats, wild individuals, and domestic cats. This gives a basic summary of the dataset (e.g., number of sites) as well as genetic diversity estimates for each population, including observed and expected heterozygosity (H_o and H_E) and mean inbreeding coefficient (F_{IS}). Weir & Cockerham's (1984) estimates for population pairwise F_{ST} are shown on the right-hand side.

Summary	**Population**			**Pairwise F_{ST}**		
	Captive	**Wild**	**Domestic**		Captive	Wild
Number of individuals	59	45	4	Captive		
Number of loci	6546	6546	6546	Wild	0.130	
Number of alleles	12258	13075	11448	Domestic	0.446	0.128
% missing data	0	0	0			
H_O	0.178	0.307	0.270			
H_E	0.285	0.285	0.285			
F_{IS}	0.375	-0.077	0.055			

2.3.2 Population structure

Principal component analysis (Fig. 2.5) showed a large proportion of the genotypic variation (23.9%) was explained by the first principal component (PC1). PC1 supported strong differentiation between domestic cats and a group of almost exclusively captive individuals, only two wild-living individuals

were found at similarly extreme PC1 values. A large F_{ST} (0.446, Table 2.1) was observed between domestic cats and the captive wildcat population. The distinct PCA clustering and high F_{ST} value supports this as a cluster of putative wildcats. Most wild-living individuals were distributed across PC1, between these two groups, and therefore considered putative hybrids. A much smaller proportion of the variance is explained by PC2 (2.8%) and PC3 (2.7%).

An ADMIXTURE model with two ancestral populations (Fig. 2.6A, K=2) also supported distinct clustering of domestic cats and captive wildcats. The majority of wild individuals sampled had probable ancestry assigned to both groups, with varying amounts of 'domestic' ancestry. PC1 position was strongly correlated with ADMIXTURE Q value at K=2 (Spearman's r = 0.998, p<0.001; Fig. 2.9, Appendix 2). At K=3 further clustering within putative wildcats is observed, including within the captive population. The lowest cross-validation error was reported for K=5 (Fig. 2.6B), indicating this was the best fitting model for the data. Fig. 2.6C shows sampling locations for the wild individuals (where available), coloured by ADMIXTURE proportions at K=2. Individuals with domestic ancestry appear geographically widespread, with no clear single point of introgression.

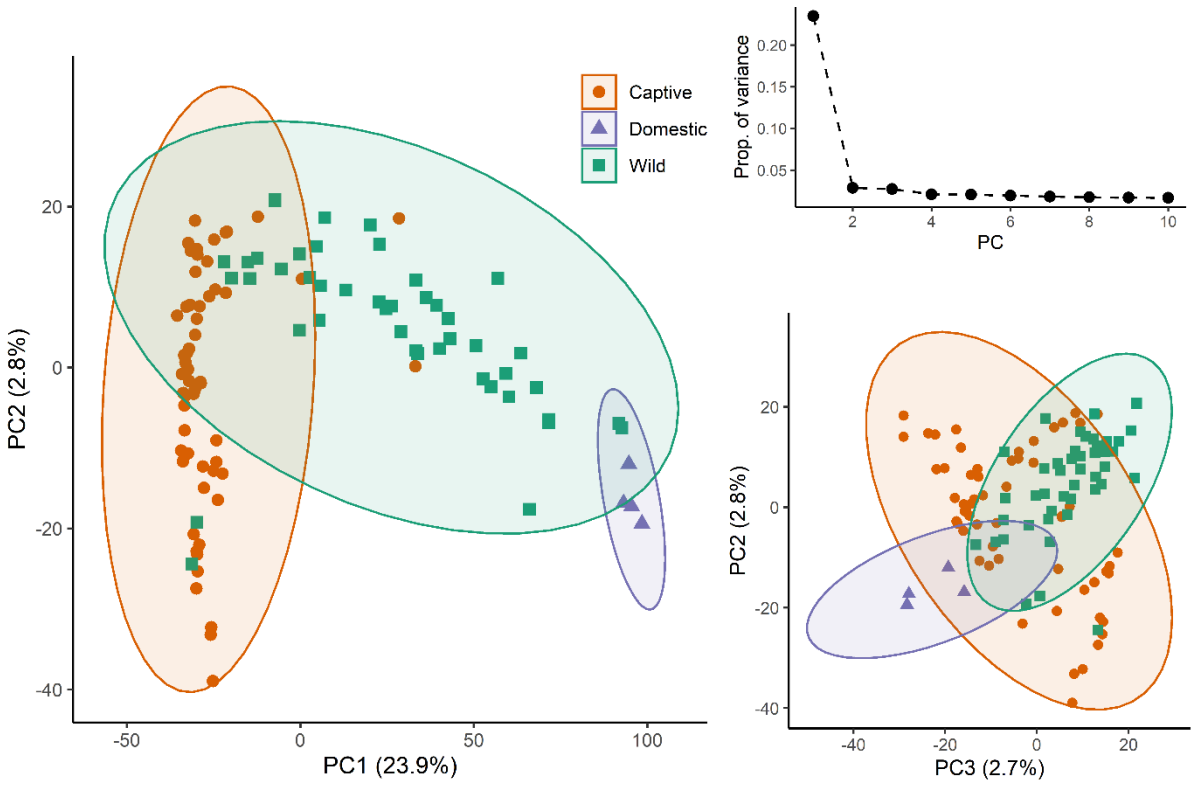

Figure 2.5. Principal component analysis (PCA) showed a strong genetic differentiation between domestic cats and a group of putative wildcats across PC1. In the wild-living population a 'hybrid swarm' is observed, with a continuum of genetic backgrounds between the putative parental groups. The captive population do not form a tight cluster on the PCA, but instead have a wide distribution across PC2 and PC3. A scree plot in the top right shows the proportion of variance explained by the first ten principal components.

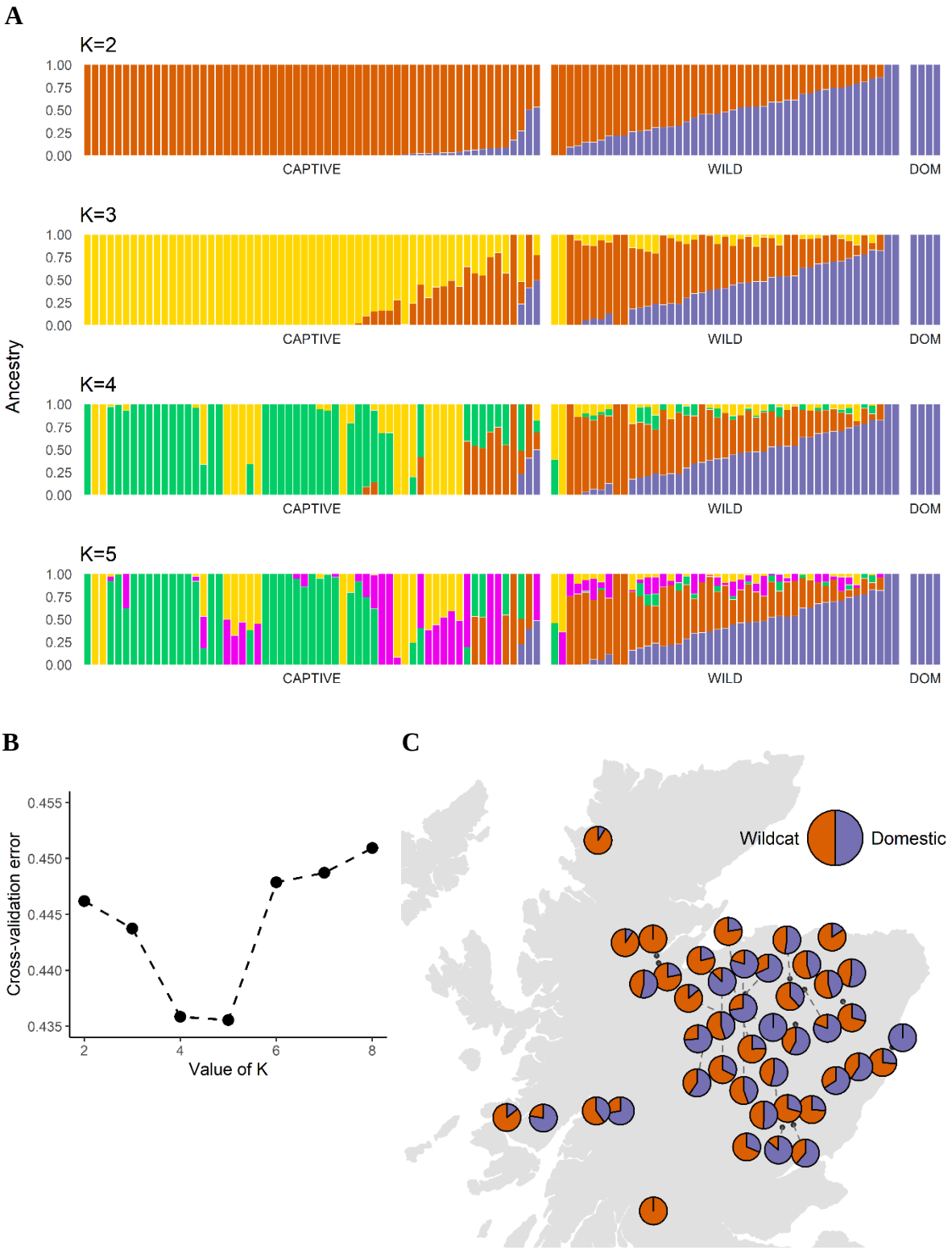

Figure 2.6. (A) ADMIXTURE clustering, K=2 to K=5. A small number of captive individuals and almost all wild individuals sampled appear to have a proportion of domestic ancestry. At increasing values of K further structure is observed within the putative wildcat population, the domestic cats are the only group to remain constant across all values of K. (B) The model with the lowest CV-error was K=5. (C) Sampling locations of wild individuals (where known), pie charts show probable ancestry for each individual at K=2, as modelled using ADMIXTURE

2.3.3 Evaluating hybrid tests

ROC curves showed that both diagnostic tests performed well, with AUC values of 0.984 (35 SNP) and 0.854 (7PS) (Fig. 2.7). The 35 SNP test (LBQ≥0.75) outperformed the morphology-based test, with a low rate of both false positives and false negatives (Table 2.2). Using a threshold of 17 the 7PS test showed nine false negatives and six false positives (i.e., individuals with few wildcat markings or features, but a high proportion of probable wildcat ancestry, and *vice versa*). At the higher threshold of 19 there was only one instance of a false positive, but 19 false negatives. The 35 SNP test showed two false negatives and four false positives.

Table 2.2. Calculating false positive rate (FPR) and true positive rate (TPR) for existing hybrid tests. The current threshold of each test is given in brackets.

		Q35 SNP (LBQ=0.75)		7PS (17)		7PS (19)	
		wildcat	hybrid	wildcat	hybrid	wildcat	hybrid
ddRAD Q value (Q=0.9)	wildcat	26	2	19	9	9	19
	hybrid	4	27	6	25	1	30
TPR		**0.929**		**0.679**		**0.321**	
FPR		**0.129**		**0.194**		**0.032**	

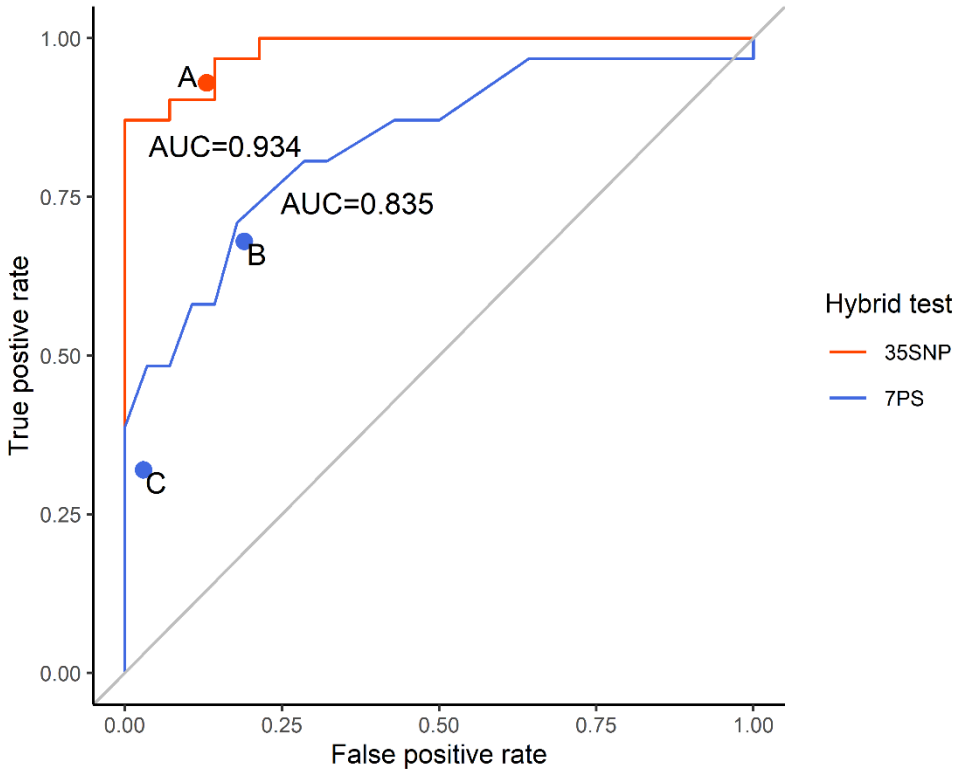

Figure 2.7. ROC curves for current tests to identify wildcat and hybrids: the 35 SNP genetic test (red) and seven-point pelage score (blue). The area under the curve (AUC) is shown for each test. True and false positive rates at the current thresholds for each test are shown using a circle at the corresponding coordinate, (A) LBQ≥0.75, (B) 7PS≥17, (C) 7PS≥19.

2.4 Discussion

2.4.1 Current status of the wildcat in Scotland

PCA and ADMIXTURE analysis (Figs. 2.5 & 2.6) demonstrated that a group of individuals genetically distinct from domestic cats (putative wildcats) persists in Scotland. Genetic differentiation between these groups was supported by a high F_{ST}, as would be anticipated between two species (Hartl & Clark, 2007), and comparable to that between dogs and wolves (Cronin et al., 2015) or red and sika deer (McFarlane et al., 2020). This supports the findings of previous microsatellite (Beaumont et al., 2001) and SNP studies (Senn et al., 2019) that were able to differentiate between domestic cats and a group of putative wildcats in Scotland. Here we reanalyse the 76 samples used by Senn *et al.* (2019) and an additional 53 individuals. We increase the resolution of the previous analysis by 3,449 SNPs, and the data show the same broad patterns. Putative wildcats reported in this study were sampled almost exclusively from the UK captive population. Hybridisation in the wild appeared extensive. A continuum of genetic backgrounds is observed, the result of repeated

hybridisation, backcrossing, and mating between hybrids referred to as a 'hybrid swarm' (Mayr, 1963); almost all wild-living individuals sampled showed some evidence of introgression from domestic cats. This supports the conclusion of Breitenmoser *et al.* (2019) that the remaining 'true wildcat' population is small, and at the current rate of introgression from domestic cats, at high-risk of extinction in the near-future.

Captive individuals have a wide distribution across PC2 and PC3 (Fig. 2.5), though this explains only a small proportion of the variation in the genetic data (2.8% and 2.7%, respectively). ADMIXTURE plots show clustering within the captive population (Fig. 2.6A), however, there does not appear to be a clear explanation for the clustering observed across the dataset at values of K greater than two. It is hard to disentangle the impacts of maintaining a (historically small) captive breeding population, e.g. inbreeding, genetic drift, or adaption to captivity (Frankham, 1995; Woodworth, Montgomery, Briscoe, & Frankham, 2002), from genuine population structure. The presence of family groups was limited following the identification of close relatives using PRIMUS. However, estimates of relatedness are complicated by potential admixture (Morrison, 2013). Our results (Fig. 2.10, Appendix 2) imply the distribution of individuals across PC2 or PC3 is not a gradient of inbreeding across the population.

Hybridisation appeared extensive across the wildcat range in Scotland, though a limited number of samples were collected from the north and west. In terms of introgression it seems clear there have been multiple admixture events, possibly due to continuing high levels of persecution maintaining wildcat populations at low levels and pervasiveness of domestic cats in wildcat habitat (Kitchener & O'Connor 2010). Patterns of genetic clustering corresponding to the geographic origin of the wild samples were unclear due to the high levels of introgression (Fig. 2.6C). The evidence presented here does not rule out that the observed clustering in the captive population reflects biogeographic structure in the Scottish wildcat population. The Great Glen, for example, has been suggested as a barrier to gene flow in the Scottish red deer population (Pérez-Espona et al., 2008). The Great Glen is a ~100km long valley, running along part of the Great Glen fault that bisects the Scottish Highlands. In red deer, strong population differentiation is observed between the eastern and western sides. Wild-living individuals belonging to a single cluster at K=3, however, were sampled from both sides of the Great Glen (potentially unsurprising given the recent range expansion). Other geographical barriers may need to be considered and tested with additional sampling.

A second possibility is that ADMIXTURE clustering at values of K greater than two reflect temporal patterns of hybridisation, i.e., snapshots of the genetic composition of the wild-living population at various points since the mid-20th century (a number of wild founders have been incorporated into the captive population since it was founded in 1960). The value of K with the lowest cross-validation error was five, this may be an effect of trying to break a continuum of

hybridisation levels into discrete units. It is interesting to note that captive individuals with probable domestic ancestry at K=2 all belong to the same cluster at K=3.

2.4.2 Existing tests for hybrids

Accurately identifying hybrids in the field is crucial to effective conservation of the wildcat in Scotland. In the absence of uncontroversial reference samples, we have used a score based on 6,546 ddRAD SNPs and investigated the relative effectiveness of current hybrid tests in recovering this. An ROC analysis (Fig. 2.7) showed both diagnostic tests to be informative in identifying hybrid individuals as judged by scores from the ddRAD SNPs. The pelage score was a less reliable indicator of wildcat ancestry; this is unsurprising given the characteristics scored by this test are likely to be controlled by a limited number of genes (Cieslak, Reissmann, Hofreiter, & Ludwig, 2011; Eizirik et al., 2010), the transmission of which is still poorly understood. Devillard *et al.* (2014) and Kitchener *et al.* (2005) reported a greater degree of accuracy when using anatomical characteristics (skull size and shape and intestinal length) as opposed to pelage in order to identify hybrids. Mattucci *et al.* (2019) found genomic regions in hybrid individuals with a high frequency of wildcat-type alleles contained (amongst others) genes relating to morphology. If selection is acting on key morphological features, as this result suggests, pelage may not give an accurate picture of hybridisation across the genome. Using a more lenient threshold (7PS ≥17 for putative wildcats) appeared to give a number of false negatives and false positives, i.e., individuals with probable wildcat ancestry that did not necessarily score highly for wildcat features and vice versa. A more conservative threshold of 7PS≥19 reduced the number of false positives but increased the false negative rate - a number of individuals with high proportions of putative wildcat ancestry are not classified as wildcats at this threshold.

We found the 35 SNP test to be a highly accurate predictor of the ddRAD SNP score; hybrids could be identified almost as well using 35 SNPs as with a dense marker set of over 6,000 SNPs. Four false positives and two false negatives were identified, though similar Q values were recovered using both marker sets for these individuals, so this may partly reflect the stringent threshold used to select reference wildcats from the ddRAD data. Interestingly, in a separate panel of 158 SNPs, Oliveria *et al.* (2015) found the 35 most differentiated SNPs also correctly identified hybrids in 99% of cases.

Without accurate information on the history of hybridisation in Britain there is no uncontroversial baseline for Scottish wildcats with which to calibrate either diagnostic test. Therefore, we recommend the continued use of the pelage score and 35 SNP test in conjunction to identify hybrids, especially when considering individuals to be incorporated into the captive breeding programme.

2.5 Conclusion

The results presented here give a detailed picture of Scottish wildcat hybridisation. As reported in previous studies, the wild-living population now resembles a hybrid swarm. The captive population appears to cluster at one end of the observed genetic continuum, and a high F_{ST} is reported between this group and domestic cats. These individuals now represent the last putative wildcats in Scotland and are an important resource for conservation management. Management of Scottish wildcats is supported by accurate tests for hybrids.

2.6 References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Research*, 19(9), 1655–1664

Barclay, D. (2019) *Scottish Wildcat Studbook*, Species360 Zoological Information Management System (ZIMS). zims.Species360.org

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2), 263–265

Battersby, J. (Ed) & Tracking Mammals Partnership. 2005. *UK Mammals: Species Status and Population Trends. First Report by the Tracking Mammals Partnership*. JNCC/Tracking Mammals Partnership, Peterborough

Beaumont, M., Barratt, E. M., Gottelli, D., Kitchener, A. C., Daniels, M. J., Pritchard, J. K., & Bruford, M. W. (2001). Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology*, 10(2), 319–336

Breitenmoser, U., Lanz, T., & Breitenmoser-Würsten, C. (2019). *Conservation of the wildcat (*Felis silvestris*) in Scotland: Review of the conservation status and assessment of conservation activities*. IUCN SSC. http://www.scottishwildcataction.org/media/42633/wildcat-in-scotland-review-of-conservation-status-and-activities-final-14-february-2019.pdf

Buckley, R. M., Davis, B. W., Brashear, W. A., Farias, F. H. G., Kuroki, K., Graves, T., … Warren, W. C. (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), 1–28

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, 22(11), 3124–3140

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 1–16

Cieslak, M., Reissmann, M., Hofreiter, M., & Ludwig, A. (2011). Colours of domestication. *Biological Reviews*, 86(4), 885–899

Cronin, M. A., Cánovas, A., Bannasch, D. L., Oberbauer, A. M., MeDrano, J. F., & Ostrander, E. (2015). Single nucleotide polymorphism (SNP) variation of wolves (*Canis lupus*) in Southeast Alaska and comparison with wolves, dogs, and Coyotes in North America. *Journal of Heredity*, 106(1), 26–36

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158

Daniels, M. J., Balharry, D., Hirst, D., Kitchener, A. C. &Aspinall, R. J. (1998). Morphological and pelage characteristics of wild living cats in Scotland: implications for defining the 'wildcat'. *Journal of Zoology*, 244, 231–247

Davis, A. R., & Gray, D. (2010). *Scottish Wildcat Survey 2006-2008*. Scottish Natural Heritage Commissioned Report No. 360, 60

Devillard, S., Jombart, T., Léger, F., Pontier, D., Say, L., & Ruette, S. (2014). How reliable are morphological and anatomical characters to distinguish European wildcats, domestic cats and their hybrids in France? *Journal of Zoological Systematics and Evolutionary Research*, 52(2), 154–162

Easterbee. N., Hepburn. L. V., Jeffries, D. J. (1991) *Survey of the status and distribution of the wildcat in Scotland, 1983–1987*. Nature Conservancy Council for Scotland, Edinburgh

Eizirik, E., David, V. A., Buckley-Beason, V., Roelke, M. E., Schäffer, A. A., Hannah, S. S., … Menotti-Raymond, M. (2010). Defining and mapping mammalian coat pattern genes: Multiple genomic regions implicated in domestic cat stripes and spots. *Genetics*, 184(1), 267–275

EU LIFE (2021), SWAforLIFE: Scottish Wildcat Action Phase 2 Wildcat recovery through threat mitigation and translocation. Retrieved June 2 2021, from https://webgate.ec.europa.eu/life/publicWebsite/index.cfm?fuseaction=search.dspPage&n_proj_id=7323

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861–874

Frankham, R. (1995). Conservation genetics.*Annual Review of Genetics*, 29, 382–390

French, D., D., Corbett, L., K. & Easterbee, N. (1988). Morphological discriminant functions of Scottish wildcats (*Felis silvestris*), domestic cats (*F. catus*) and their hybrids. *Journal of Zoology*, 214, 235–259

Forest Research (2021) Forest cover - international comparisons. Retrieved May 29, 2021, from https://www.forestresearch.gov.uk/tools-and-resources/statistics/forestry-statistics/forestry-statistics-2018/international-forestry/forest-cover-international-comparisons/

Harris, S., Morris, P., Wray, S., & Yalden, D. (1995). *A review of British mammals: population estimates and conservation status of British mammals other than cetaceans*. Peterboroough: The Joint Nature Conservation Committee

Hartl, D. L,, & Clark, A. G. (2007) *Principles of population genetics* (4th ed.). Oxford: Oxford University Press

Hertwig, S. T., Schweizer, M., Stepanow, S., Jungnickel, A., Böhle, U. R., & Fischer, M. S. (2009). Regionally high rates of hybridization and introgression in German wildcat populations (*Felis silvestris*, Carnivora, Felidae). *Journal of Zoological Systematics and Evolutionary Research*, 47(3), 283–297

Hetherington D., & Campbell, R (2012) *The Cairngorms Wildcat Project Final Report*. Report to Cairngorms National Park Authority, Scottish Natural Heritage, Royal Zoological Society of Scotland, Scottish Gamekeepers Association and Forestry Commission Scotland

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting Fst. *Nature Reviews Genetics*, 10(9), 639–650

Kilshaw, K. & Macdonald, D., W., (2011). *The use of camera trapping as a method to survey for the Scottish wildcat*. Scottish Natural Heritage Commissioned Report 479

Kilshaw, K., Johnson, P. J., Kitchener, A. C., & Macdonald, D. W. (2015). Detecting the elusive Scottish wildcat *Felis silvestris silvestris* using camera trapping. *Oryx*, 49(2), 207–215

Kilshaw, K., Montgomery, R. A., Campbell, R. D., Hetherington, D. A., Johnson, P. J., Kitchener, A. C., … Millspaugh, J. J. (2016). Mapping the spatial configuration of hybridization risk for an endangered population of the European wildcat (*Felis silvestris silvestris*) in Scotland. *Mammal Research*, 61(1), 1–11

Kitchener, A. C., Yamaguchi, N., Ward, J. M., & Macdonald, D. W. (2005). A diagnosis for the Scottish wildcat (*Felis silvestris*): A tool for conservation action for a critically-endangered felid. *Animal Conservation*, 8(3), 223–237

Krüger, M., Hertwig, S. T., Jetschke, G., & Fischer, M. S. (2009). Evaluation of anatomical characters and the question of hybridization with domestic cats in the wildcat population of thuringia, Germany. *Journal of Zoological Systematics and Evolutionary Research*, 47(3), 268–282

Lacy, R.C., 1994. Managing genetic diversity in captive populations of animals, in: Bowles, M.L., Whelan, C.J. (Eds.), *Restoration of Endangered Species* (pp. 63-89). Cambridge: Cambridge University Press

Langley, P. J. W., & Yalden, D. W. (1977). The decline of the rarer carnivores in Great Britain during the nineteenth century. *Mammal Review*, 7(3–4), 95–116

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

Littlewood, N. A., Campbell, R. D., Dinnie, L., Gilbert, L., Hooper, R., Iason, G., … Ross, A. (2014*). Survey and scoping of wildcat priority areas.* Scottish Natural Heritage Commissioned Report No. 768.

Mathews, F., Kubasiewicz, L. M., Gurnell, J., Harrower, C.A., McDonald, R.A., Shore, R.F. (2018) *A Review of the Population and Conservation Status of British Mammals: Technical Summary*. A report by the Mammal Society under contract to Natural England, Natural Resources Wales and Scottish Natural Heritage. Natural England, Peterborough

Mayr, E. (1963) *Animal Species and Evolution*. Cambridge, MA: Harvard University Press

McFarlane, S. E., Hunter, D. C., Senn, H. V., Smith, S. L., Holland, R., Huisman, J., & Pemberton, J. M. (2020). Increased genetic marker density reveals high levels of admixture between red deer and introduced Japanese sika in Kintyre, Scotland. *Evolutionary Applications*, 13(2), 432–441

McFarlane, S. E., & Pemberton, J. M. (2019). Detecting the True Extent of Introgression during Anthropogenic Hybridization. *Trends in Ecology and Evolution*, 34(4), 315–326

Meredith, A., Bacon, A., Allan, B., Kitchener, A., Senn, H., Brooks, S., … Davies, S. (2018). Domestic cat neutering to preserve the Scottish wildcat. *Veterinary Record*, 183(1), 27–28

Morrison, J. (2013). Characterization and correction of error in genome-wide IBD estimation for samples with population structure. *Genetic Epidemiology* 37(6), 635–641

Nussberger, B., Greminger, M. P., Grossen, C., Keller, L. F., & Wandeler, P. (2013). Development of SNP markers identifying European wildcats, domestic cats, and their admixed progeny. *Molecular Ecology Resources*, 13(3), 447–460

Oliveira, R., Randi, E., Mattucci, F., Kurushima, J. D., Lyons, L. A., & Alves, P. C. (2015). Toward a genome-wide approach for detecting hybrids: Informative SNPs to detect introgression between domestic cats and European wildcats (*Felis silvestris*). *Heredity*, 115(3), 195–205

Pérez-Espona, S., Pérez-Barbería, F. J., Mcleod, J. E., Jiggins, C. D., Gordon, I. J., & Pemberton, J. M. (2008). Landscape features affect gene flow of Scottish Highland red deer (*Cervus elaphus*). *Molecular Ecology*, 17(4), 981–996

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, 7(5), e37135

Pontius, J. U., Mullikin, J. C., Smith, D. R., Lindblad-Toh, K., Gnerre, S., Clamp, M., … McKernan, K. (2007). Initial sequence and comparative analysis of the cat genome. *Genome Research*, 17(11), 1675–1689

Pritchard, J. K., Seielstad, M. T., & Feldman, M. W. (1998). Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites. *Molecular Biology and Evolution*, 16(12), 1791–1798

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Miller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 8, 12–77

Sainsbury, K. A., Shore, R. F., Schofield, H., Croose, E., Campbell, R. D., & McDonald, R. A. (2019). Recent history, current status, conservation and management of native mammalian carnivore species in Great Britain. *Mammal Review*, 49(2), 171–188

Schauenberg, P. (1969). L'identification du chat forestier d'Europe *Felis s. silvestris* Schreber 1777 par une méthode ostéométrique. *Rev. Suisse Zool*, 76, 433–441

Schauenberg, P. (1977). Longeur de l'intestin du chat forestier *Felis silvestris* Schreber. *Mammalia*, 41, 357–360

Scottish Natural Heritage. (2013). *Scottish Wildcat Conservation Action Plan*

Senn, H., & Ogden, R. (2015). *Wildcat hybrid scoring for conservation breeding under the Scottish Wildcat Conservation Action Plan*. Royal Zoological Society of Scotland

Senn, H. V., Ghazali, M., Kaden, J., Barclay, D., Harrower, B., Campbell, R. D., … Kitchener, A. C. (2019). Distinguishing the victim from the threat: SNP-based methods reveal the extent of introgressive hybridization between wildcats and domestic cats in Scotland and inform future in situ and ex situ management options for species restoration. *Evolutionary Applications*, 12(3), 399–414

Staples, J., Qiao, D., Cho, M. H., Silverman, E. K., Nickerson, D. A., & Below, J. E. (2014). PRIMUS: Rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *American Journal of Human Genetics*, 95(5), 553–564

Steyer, K., Kraus, R. H. S., Mölich, T., Anders, O., Cocchiararo, B., Frosch, C., … Nowak, C. (2016). Large-scale genetic census of an elusive carnivore, the European wildcat *(Felis s. silvestris*). *Conservation Genetics*, 17(5), 1183–1199

Steyer, K., Tiesmeyer, A., Muñoz-Fuentes, V., & Nowak, C. (2018). Low rates of hybridization between European wildcats and domestic cats in a human-dominated landscape. *Ecology and Evolution*, 8(4), 2290–2304

Velli, E., Bologna, M. A., Silvia, C., Ragni, B., & Randi, E. (2015). Non-invasive monitoring of the European wildcat (*Felis silvestris silvestris* Schreber, 1777): comparative analysis of three different monitoring techniques and evaluation of their integration. *European Journal of Wildlife Research*, 61(5), 657–668

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6), 1358–1370

Woodworth, L. M., Montgomery, M. E., Briscoe, D. A., & Frankham, R. (2002). Rapid genetic deterioration in captive populations: Causes and conservation implications. *Conservation Genetics*, 3(3), 277–288

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16, 97–159

Yamaguchi, N., Kitchener, A., Driscoll, C., & Nussberger, B. (2015). *Felis silvestris*. *The IUCN Red List of Threatened Species 2015*, e.T60354712A50652361

2.7 Appendix 1. Sample information

Table 2.3. Sample information. 'Q6546' indicates the ADMIXTURE Q values generated using the ddRAD-seq data. Q values are also given for the 35 SNP test, with the lower and upper bounds (LBQ/UBQ, 90% CI) also shown. Pelage scores (7PS) were available for a subset of samples. The final column indicates whether a sample was included in the final dataset, 21 samples were excluded due to high levels of missing data, or to limit relatedness in the data.

Sample ID	Source pop	Year sampled	Q6546	Q35	LBQ35	UBQ35	7PS	Included in the final dataset? (Yes/No)
WCQ0047	Wild	1996	0.892	0.931	0.854	0.987	18	Y
WCQ0052	Wild	1999	0.853	0.91	0.838	0.966	18	Y
WCQ0099	Wild		0	0.012	0	0.049	16	Y
WCQ0100	Wild		0.315	0.289	0.184	0.402	10	Y
WCQ0107	Wild	1996	0.522	0.599	0.476	0.718		Y
WCQ0110	Wild	2008	0.78	0.851	0.754	0.932		Y
WCQ0132	Captive		NA	0.822	0.709	0.922	14	N
WCQ0158	Wild	1997	0.855	0.876	0.787	0.949	11	Y
WCQ0165	Wild		0	0.035	0	0.102	9	Y
WCQ0168	Wild	2007	0.412	0.402	0.284	0.523	10	Y
WCQ0172	Wild		0.253	0.056	0	0.147	13	Y
WCQ0208	Wild	2013	0.693	0.724	0.612	0.826	14	Y
WCQ0209	Wild	2013	NA	0.666	0.552	0.773	16	N
WCQ0210	Wild	2005	0.78	0.851	0.754	0.932		Y
WCQ0211	Wild	1999	0.908	0.884	0.8	0.952	14.5	Y
WCQ0212	Wild	2011	0.722	0.794	0.692	0.884	16	Y
WCQ0213	Wild	2002	0.737	0.732	0.618	0.835	15	Y
WCQ0214	Wild	2003	0.832	0.909	0.834	0.966	13	Y
WCQ0216	Wild	2014	0.581	0.599	0.478	0.714	15	Y
WCQ0217	Captive	2014	1	0.986	0.941	1	17	Y
WCQ0218	Wild	2011	0.674	0.777	0.67	0.873	13	Y
WCQ0222	Wild	2014	NA	0.734	0.625	0.833	12	N
WCQ0223	Wild	2014	NA	0.556	0.436	0.673	15	N
WCQ0224	Wild	2014	NA	0.796	0.691	0.89	8	N
WCQ0227	Wild	2012	0.206	0.288	0.182	0.401	12.5	Y
WCQ0229	Wild	2014	0.272	0.282	0.171	0.401	10	Y
WCQ0230	Wild	2011	0.321	0.367	0.251	0.488	15	Y
WCQ0231	Wild	2011	0.411	0.466	0.348	0.585	17	Y
WCQ0234	Wild	2014	0.463	0.626	0.507	0.74	7	Y
WCQ0236	Wild	2011	0.499	0.489	0.371	0.607	9	Y
WCQ0243	Captive	2014	1	0.937	0.858	0.992	19	Y
WCQ0245	Captive	2014	NA	0.959	0.877	1	19	N
WCQ0246	Wild	2014	0.627	0.433	0.315	0.553	14	Y
WCQ0247	Wild	2014	0.682	0.721	0.608	0.825		Y
WCQ0248	Wild	2014	1	0.974	0.905	1	17	Y
WCQ0249	Wild	2013	0.462	0.632	0.513	0.745	10	Y
WCQ0252	Wild	2014	NA	0.528	0.406	0.648	15	N
WCQ0255	Wild	2012	0.391	0.327	0.216	0.445	18	Y
WCQ0340	Captive	2014	1	0.938	0.846	1	18.5	Y
WCQ0343	Captive	2014	1	0.991	0.961	1	20	Y
WCQ0344	Captive	2014	1	0.99	0.959	1	20	Y

WCQ0358	Captive		NA	0.98	0.919	1	17	N
WCQ0383	Wild	2014	0.541	0.431	0.315	0.55	8	Y
WCQ0387	Wild	2014	0.688	0.548	0.427	0.666	15.5	Y
WCQ0390	Wild	2014	NA	0.641	0.523	0.753	13	N
WCQ0402	Captive	2017	1	0.967	0.881	1	16.5	Y
WCQ0404	Captive	2017	1	0.974	0.901	1	15	Y
WCQ0408	Captive	2017	0.495	0.537	0.418	0.654	21	Y
WCQ0419	Captive	2006	NA	0.984	0.933	1		N
WCQ0420	Captive	2013	0.992	0.934	0.856	0.99	16	Y
WCQ0421	Captive	2013	0.978	0.962	0.901	0.999		Y
WCQ0422	Captive	2013	1	0.993	0.97	1		Y
WCQ0427	Captive	2012	0.829	0.865	0.773	0.942		Y
WCQ0428	Captive	2015	1	0.967	0.881	1	14	Y
WCQ0429	Captive	2015	1	0.989	0.952	1		Y
WCQ0430	Captive	2015	NA	0.826	0.72	0.918		N
WCQ0431	Captive	2015	NA	0.828	0.724	0.918		N
WCQ0432	Captive	2015	0.92	0.856	0.752	0.944		Y
WCQ0433	Captive	2015	0.919	0.867	0.77	0.948		Y
WCQ0434	Captive	2015	NA	0.85	0.744	0.941		N
WCQ0435	Captive	2015	1	0.993	0.968	1		Y
WCQ0436	Captive	2015	1	0.991	0.962	1		Y
WCQ0437	Captive	2015	0.73	0.745	0.633	0.847		Y
WCQ0439	Captive	2015	1	0.976	0.906	1		Y
WCQ0443	Domestic	2015	NA	0.01	0	0.041		N
WCQ0485	Wild	2014	0.138	0.146	0.061	0.246		Y
WCQ0486	Wild	2014	0.152	0.128	0.042	0.228		Y
WCQ0487	Captive		NA	0.979	0.922	1	15	N
WCQ0488	Captive	?	NA	0.803	0.7	0.893	19	N
WCQ0489	Captive	2017	1	0.985	0.939	1		Y
WCQ0490	Captive	2015	1	0.989	0.954	1		Y
WCQ0491	Captive	2015	1	0.983	0.931	1		Y
WCQ0492	Domestic	2007	0	0.023	0	0.079		Y
WCQ0494	Domestic	2010	0	0.012	0	0.051		Y
WCQ0501	Domestic	2008	0	0.023	0	0.086		Y
WCQ0504	Domestic	2011	0	0.043	0.001	0.112		Y
WCQ0515	Wild	2014	1	0.992	0.966	1	14-16	Y
WCQ0519	Wild	2010	0.283	0.271	0.164	0.387		Y
WCQ0527	Wild	2015	0.229	0.294	0.183	0.412	11-14	Y
WCQ0528	Wild	2009	0.732	0.727	0.613	0.831		Y
WCQ0529	Wild	2015	0.783	0.733	0.621	0.834		Y
WCQ0531	Captive	2015	1	0.979	0.926	1	16	Y
WCQ0540	Captive	2015	1	0.988	0.95	1		Y
WCQ0541	Captive	2015	1	0.986	0.942	1		Y
WCQ0544	Captive	2015	1	0.987	0.944	1	17	Y
WCQ0545	Captive	2015	1	0.984	0.934	1	18	Y
WCQ0546	Captive	2015	1	0.985	0.941	1	19.5	Y
WCQ0547	Captive	2015	1	0.977	0.917	1	18	Y
WCQ0549	Captive	2015	1	0.985	0.936	1		Y
WCQ0550	Captive	2015	1	0.977	0.923	1	19.5	Y
WCQ0551	Captive	2015	1	0.975	0.918	1	20	Y
WCQ0552	Captive	2015	1	0.968	0.89	1	19.5	Y

WCQ0553	Captive	2015	1	0.981	0.935	1	17.5	Y
WCQ0554	Captive	2016	1	0.955	0.861	1		Y
WCQ0555	Captive	2016	1	0.943	0.84	1	18	Y
WCQ0556	Captive	2016	1	0.975	0.908	1		Y
WCQ0557	Captive	2016	0.938	0.843	0.74	0.931	13.5	Y
WCQ0558	Captive	2016	NA	0.895	0.798	0.975		N
WCQ0559	Captive	2016	0.912	0.825	0.721	0.915	13	Y
WCQ0560	Captive	2016	1	0.968	0.891	1	18	Y
WCQ0564	Captive	2016	1	0.931	0.822	1	17.5	Y
WCQ0567	Captive	2016	1	0.99	0.957	1		Y
WCQ0578	Wild	2016	0.189	0.144	0.065	0.236	15	Y
WCQ0586	Captive	2016	0.93	0.93	0.849	0.989		Y
WCQ0588	Captive	2016	1	0.982	0.931	1		Y
WCQ0589	Captive	2016	0.467	0.328	0.213	0.448		Y
WCQ0603	Wild	2015	NA	0.646	0.528	0.757		N
WCQ0604	Wild	2015	0.546	0.641	0.523	0.751	10.5	Y
WCQ0606	Wild	2015	0.458	0.424	0.306	0.544		Y
WCQ0612	Captive	2016	NA	0.927	0.835	0.998		N
WCQ0613	Wild	2016	0.47	0.445	0.326	0.565	<14	Y
WCQ0614	Captive	2016	0.947	0.89	0.796	0.967		Y
WCQ0615	Captive	2016	1	0.985	0.938	1		Y
WCQ0616	Captive	2016	1	0.986	0.941	1		Y
WCQ0617	Captive	2016	1	0.979	0.916	1		Y
WCQ0618	Captive	2016	0.971	0.88	0.774	0.967		Y
WCQ0619	Captive	2016	0.961	0.854	0.754	0.938		Y
WCQ0620	Captive	2016	0.969	0.882	0.786	0.96		Y
WCQ0621	Captive	2016	0.974	0.938	0.859	0.996		Y
WCQ0622	Captive	2016	0.977	0.886	0.789	0.963		Y
WCQ0624	Captive	2016	NA	0.965	0.884	1	16	N
WCQ0626	Captive	2016	1	0.956	0.87	1		Y
WCQ0627	Captive	2016	1	0.956	0.884	1	20	Y
WCQ0628	Captive	2016	0.981	0.974	0.907	1	20.5	Y
WCQ0629	Captive	2016	1	0.988	0.949	1	15.5	Y
WCQ0901	Wild	2015	0.389	0.418	0.3	0.539	12	Y
WCQ0902	Wild	2015	0.255	0.249	0.141	0.365		Y
WCQ0903	Wild	2013	0.545	0.644	0.523	0.759	13 (2 markings not scored)	Y
WCQ0904	Wild	2015	NA	0.344	0.233	0.461	13	N

2.8 Appendix 2. Supplementary material

Table 2.4. For a subset of samples (three captive, three domestic and five wild) alignment rate of raw sequence reads to the *Felis catus* reference genome was assessed. These metrics are reported as the mean across the sampled individuals. The proportion of aligned reads was approximately even across the three samples sources, despite captive and wild populations containing putative *Felis silvestris* individuals. Coverage was slightly lower in the captive and wild populations.

	No. of reads	**No. of aligned reads**	**Proportion of aligned reads (%)**	**Mean coverage**
Captive	798394	629063	79	35.4
Wild	1109107	863287	78	43.4
Domestic	1270779	1011224	80	53

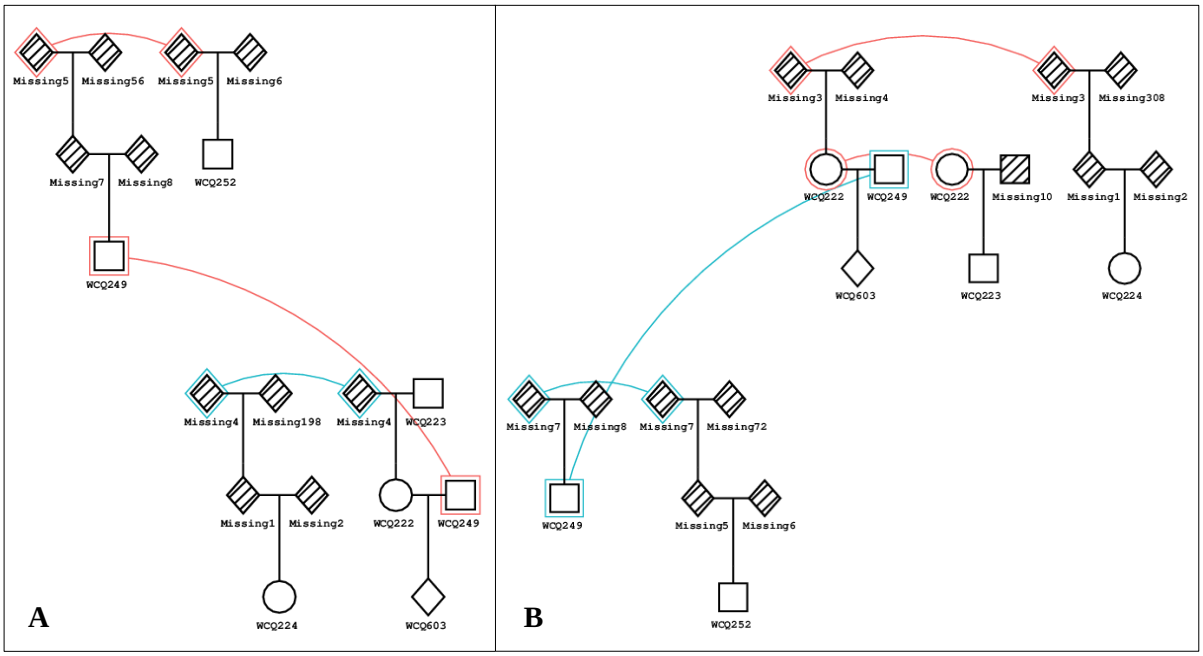

Figure 2.8. Two possible pedigrees (A and B) for a family group of individuals. Squares represent males, circles females and diamonds individuals of unknown sex. Coloured lines indicate the same individual in multiple networks, 'missing' shaded diamonds are dummies, i.e., individuals that are needed to construct the network, but whose data were not collected as part of this study. Constructing pedigrees using genetic data is a useful way to minimise relatedness in the dataset; one trio (WCQ222, WCQ249 and WCQ603) was successfully identified using this method.

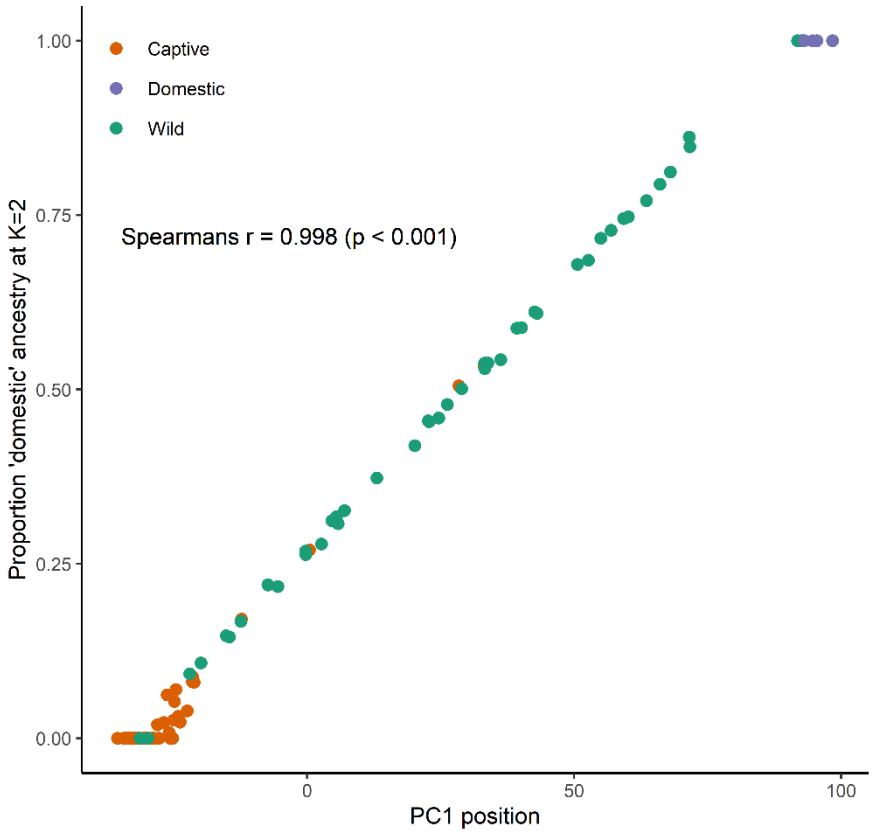

Figure 2.9. Relationship between PC1 position and putative 'domestic' ancestry at K=2. The higher the PC1 coordinate the more domestic cat ancestry an individual is likely to have. This supports a strong genetic differentiation between domestic cats and a group of putative wildcats.

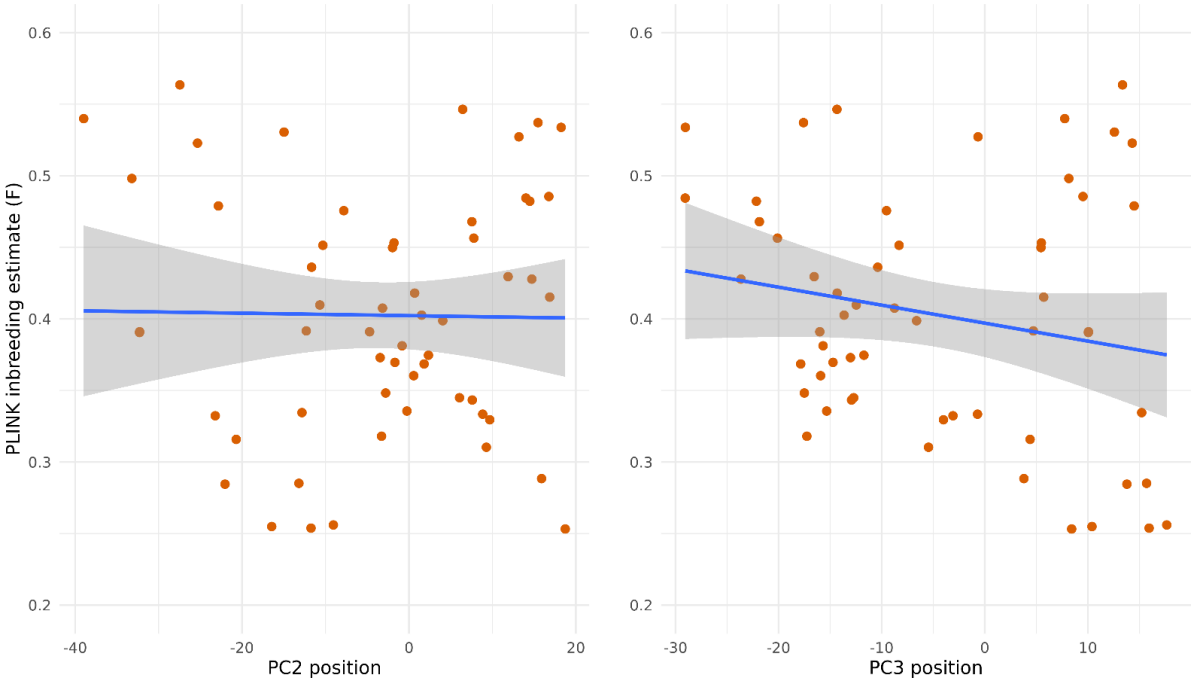

Figure 2.10. For captive animals only, the relationship between PC2 or PC3 position and inbreeding coefficient (F), excluding three hybrid individuals with negative F estimates (i.e., higher than expected heterozygosity). Inbreeding does not seem to explain the distribution of captive individuals across PC2 and PC3.

Chapter 3 Modelling hybridisation dynamics

3.1 Introduction

3.1.1 Dating admixture in the Scottish wildcat population

Uncertainty surrounds the temporal patterns of hybridisation in Scotland. Domestic cats are thought to have become widespread during the Roman occupation of Britain ~2,000 years ago (Serpell, 2014), though cat remains have also been found at Iron Age sites, including on the Orkney islands off the north coast of Scotland (Macdonald et al., 2010; Smith, 1994). Wildcats and domestic cats have therefore been sympatric, and potentially hybridising, for over 2000 years. Results from Chapter 2 and previous studies (Beaumont et al., 2001; Kilshaw, Drake, Macdonald, & Kitchener, 2010; Kitchener et al., 2005; Senn et al., 2019), however, show a group of putative wildcats persists in Scotland. Significant introgression is thought to have occurred within the last hundred years, following wildcat population expansion, when recolonisation of central and eastern Scotland by wildcats is thought to have increased contact between the small remaining population of wildcats and domestic cats (Breitenmoser et al., 2019).

Historic samples, collected over the last c. 100 years, support an acceleration of hybridisation in Scotland over this period (Beaumont et al., 2001; Senn et al., 2019). The lowest STRUCTURE Q value reported by Beaumont *et al.* (2001) (using nine microsatellites) for samples pre-dating 1970 was 0.95, only six out of twenty individuals sampled post-1970 scored 0.95 or more. A similar pattern of high-scoring (35 SNP test) historic samples was recently reported by Senn *et al.* (2019), based on a sample of 60 individuals collected between 1895 and 1985. TNVR and camera-trap surveys have reported large numbers of hybrids since conservation began in earnest in the early 2000s (Breitenmoser et al., 2019; Hetherington & Campbell, 2012; Kilshaw et al., 2015; Littlewood et al., 2014), but a lack of standardisation across historic surveys limits our ability to distinguish temporal trends. Current tests to identify hybrids (35 SNP test and seven-point pelage score) were both developed in the 21st century (Kitchener et al., 2005; Senn & Ogden, 2015). There has been an increasing awareness of hybridisation as a major threat to wildcats since the first large-scale wildcat survey in the 1980s (Easterbee et al., 1991), and only in the last few decades has introgression been monitored at all. The patterns of both recent and historic introgression are therefore poorly understood, and it is difficult to know the extent to which historic or ancient introgression has impacted the modern wildcat population in Scotland.

An improved understanding of hybridisation history and dynamics would support conservation of wildcats in Scotland. Confidence in reference wildcat samples would improve diagnostic tests for wildcats and hybrids, and therefore efficiency of conservation management,

monitoring, and legal protection. Understanding how hybridisation has proceeded in the past may allow us to manage it better in the future, in Scotland and across the *Felis silvestris* range.

3.1.2 Approximate Bayesian computation

Demographic and adaptive processes give rise to natural populations with complex histories, in evolutionary biology we are often interested in understanding these processes over long timescales (Servedio et al., 2014). Mathematical modelling is a powerful and flexible tool to test hypotheses explaining the existing genetic complexity of populations and make predictions about the future. DNA sequencing technology and computing power continue to advance rapidly, and increasingly large amounts of genetic data are available for non-model species (Csilléry, Blum, Gaggiotti, & François, 2010; Ellegren, 2014). This has led to the development of sophisticated statistical methods to model complex scenarios and datasets. One such group of methods is approximate Bayesian computation, or ABC (Beaumont, 2010).

In Bayesian inference the probability of a hypothesis is updated based on the available evidence using Bayes' theorem,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

which estimates the conditional probability of a parameter (or parameters, θ) given the data (x), i.e., the posterior probability, $p(\theta|x)$. To estimate the posterior probability, we must define the prior probability, $p(\theta)$, using our knowledge of θ before x was observed, and the likelihood, $p(x|\theta)$, the probability of observing the data given the parameter(s). The marginal likelihood, the probability of the observed data, $p(x)$, is a normalising constant in the equation (Beaumont, 2010) and can be ignored when comparing relative posterior probabilities for different values of θ (Sunnåker et al., 2013).

The likelihood function can be difficult to define or compute due to the large number of unobservable (latent) variables in complex models, such as those needed to understand the demographic histories of natural populations (Beaumont, 2010). Markov chain Monte Carlo (MCMC) or sequential Monte Carlo (SMC) methods can be used to approximate the likelihood (Gelman et al. 2003, Robert & Casella 2004), but again, for complex models, this is often computationally intensive, inefficient or in some cases, completely unfeasible (Beaumont, 2010; Green, Łatuszyński, Pereyra, & Robert, 2015). The problem is compounded by increasingly large DNA datasets available from high-throughput sequencing (Csilléry et al., 2010). Likelihood-based inference is therefore often limited to simple evolutionary or molecular models.

Approximate Bayesian computation (ABC, Beaumont, Zhang & Balding, 2002) is an approach to model-based inference that bypasses the likelihood calculation. Instead, ABC methods approximate the likelihood using simulated data, which can be compared to the observed data using summary statistics. Summary statistics capture information in both the simulated and observed data. This method was first developed in a population genetics context (Beaumont et al., 2002; Pritchard, Seielstad, Perez-Lezaun, & Feldman, 1999; Tavaré, Balding, Griffiths, & Donnelly, 1997) but its flexibility has led to its application in many subject areas, including ecology (Jabot & Chave, 2009), epidemiology (Tanaka, Francis, Luciani, & Sisson, 2006) and cell biology (Vo, Drovandi, Pettitt, & Pettet, 2015).

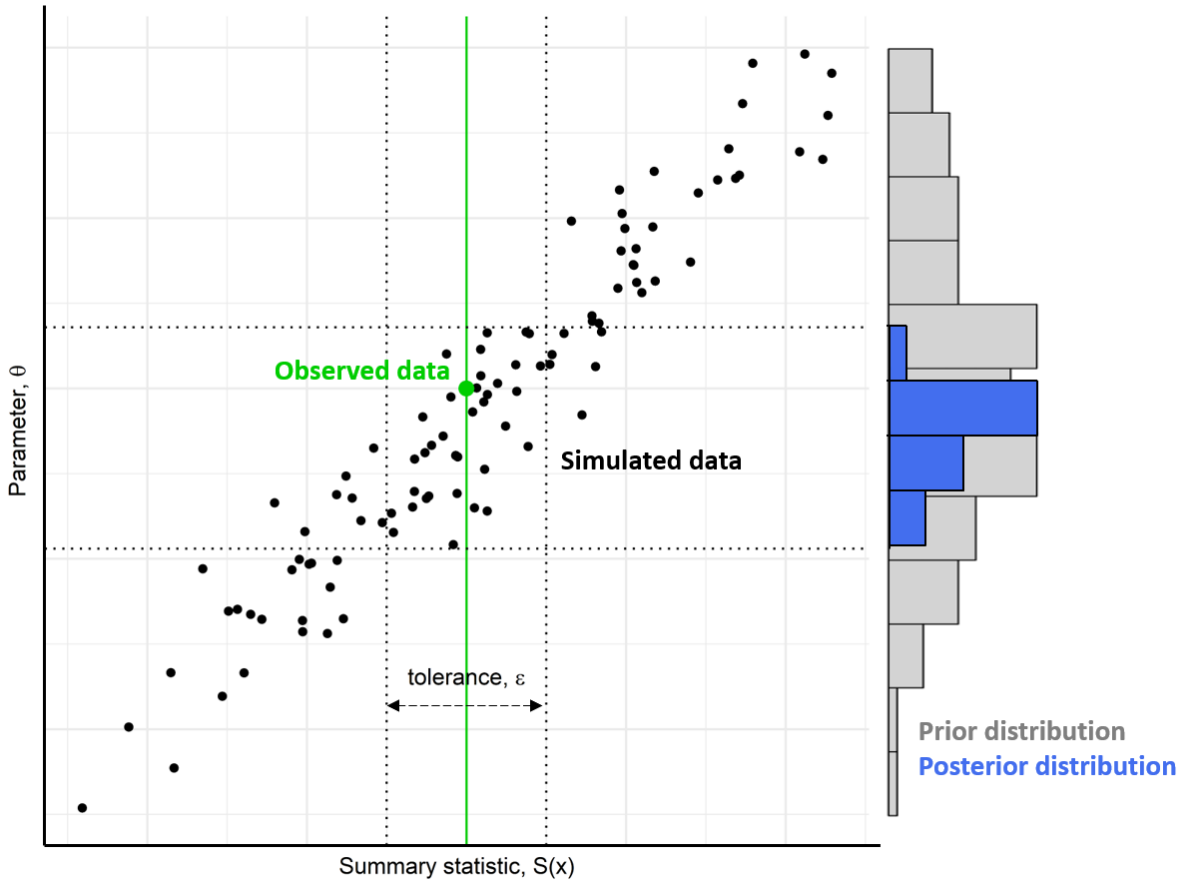

Figure 3.1. The ABC rejection algorithm. In a simple example using one model parameter, θ, and one summary statistic, S(x), data is simulated under a hypothesised model, with values of θ sampled from the prior distribution. S(x) is used to summarise information in both the observed and simulated data. Simulated data with values of S(x) closest to the observed S(x) (within a given tolerance, ε) are treated as samples from the posterior distribution of θ.

ABC follows a rejection algorithm, where data are simulated under a hypothesised model with parameters sampled from a prior distribution (Fig. 3.1) (Beaumont et al., 2002). Summary statistics are taken from both the simulated and observed data. Simulations with summary statistics closest to the observed data (within a given distance tolerance, ε) are used to generate posterior

distributions of model parameters. Posterior distributions from the rejection algorithm can then be improved, post-sampling, using local linear (Beaumont et al., 2002) or non-linear regression (Blum & François 2010). Alternatively, ABC-MCMC (Marjoram *et al.* 2003) or SMC-ABC (Sisson, Fan, & Tanaka, 2007) methods can be used to sample the parameter space more efficiently.

The probability of simulating data that exactly resembles the observed data ($\varepsilon = 0$) is impossible for continuous datasets, however, a non-zero tolerance introduces bias (Beaumont et al., 2002). A tolerance that is too large will essentially recover the prior distribution. The standard rejection algorithm is therefore sensitive to the choice of tolerance threshold. Local linear and non-linear regression (Blum & François, 2010) can be used to adjust accepted parameter values closer to the posterior, in principle correcting for this bias.

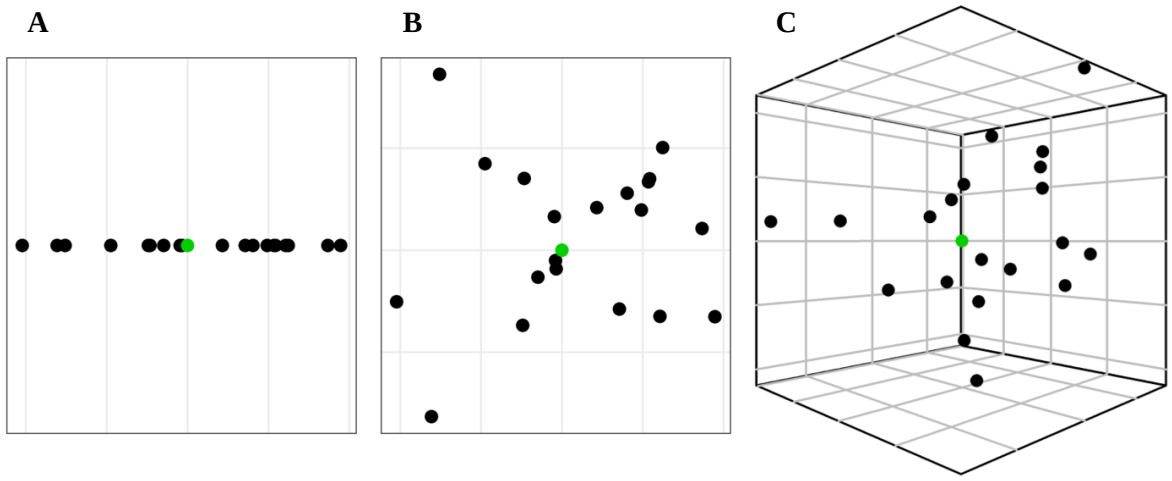

Figure 3.2. The 'curse of dimensionality'. For an example dataset with twenty simulations (black points) in (A) a single summary statistic is used (one dimension); many simulations generate points close to the observed value (green). Increasing the number of summary statistics to two (B) and then three (C), the number of simulated points close to the observed data rapidly decreases as the size of the search space increases exponentially.

Summary statistics are used in ABC to reduce the dimensionality of complex datasets, capturing useful information about the observed data (Beaumont et al., 2002). It is generally an unrealistic aim to capture all available information (i.e., sufficient summary statistics) in a low-dimensional set of summaries (Beaumont, 2010; Sunnåker et al., 2013). In practice, a set of insufficient summary statistics must be chosen that are informative about the parameter, or parameters, of interest. Increasing the number of summary statistics captures more information, but at the cost of increasing dimensionality. The dimensions of the parameter search space increase with every summary statistic (or parameter) added to the model. The larger the search space, the greater the amount of simulated data needed to accurately infer posterior densities (i.e., to simulate data that closely resembles the observed data); the number of simulations required increases exponentially with every additional dimension. This is referred to as the 'curse of dimensionality' (Fig. 3.2); for high-

dimensional data simulation becomes computationally intensive, or completely intractable. Choosing summary statistics is therefore a trade-off between informativeness and model fit. Linear and non-linear adjustment methods can help mitigate this problem by accepting and adjusting a larger sample of simulated points (using a wider tolerance interval) for posterior estimates. Other methods select useful subsets of summary statistics from a large number of candidates (Joyce & Marjoram, 2008; Nunes & Balding, 2010). Another set of methods again can handle large numbers of summary statistics, reducing dimensionality by, for example, projection (Blum, Nunes, Prangle, & Sisson, 2013).

ABC performance is highly dependent on the choice of model, summary statistics and tolerance threshold used (Beaumont, 2010; Sunnåker et al., 2013). Like other methods of statistical inference, it is sensitive to the choice of parameters and their prior distributions, which are limited by the investigator's knowledge of the model system. Model-based methods also draw criticism for not exhaustively exploring the hypothesis space (Templeton, 2009). An iterative approach to model development is therefore generally taken (Fig. 3.3), and in this way the model is refined and confidence in posterior estimates improved (Csilléry et al., 2010).

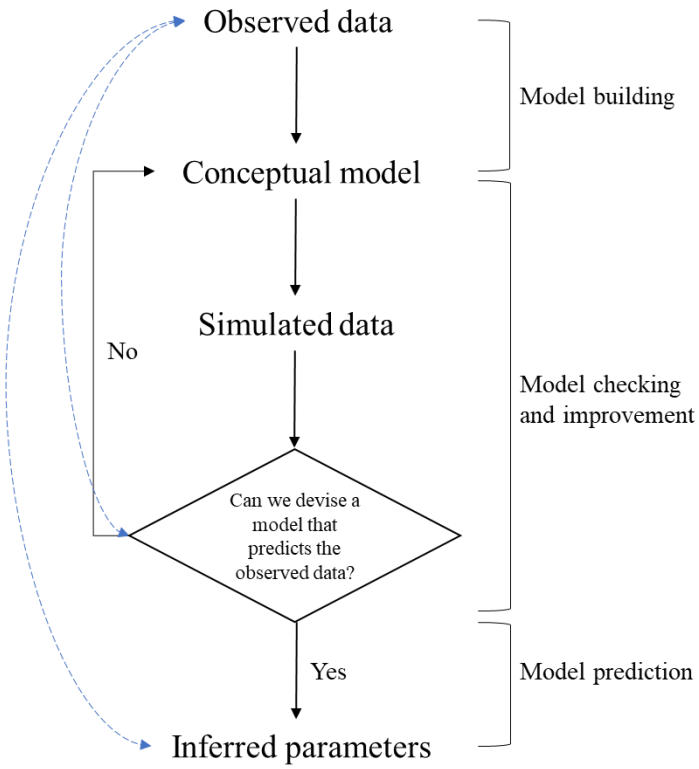

Figure 3.3. The approach to ABC inference. A conceptual model is initially designed based on knowledge of the real system, from which you have a set of observed data. Model development is an iterative process where model checking (comparing the fitted model to the empirical data) is crucial to refine the model. Often multiple rounds of model improvement, data simulation and model checking are needed for reliable parameter inference.

Firstly, a model, or models, are formulated to test a given hypothesis. There is no set procedure for devising a conceptual model, and this step relies heavily on knowledge of model system (Csilléry et al., 2010). It is sensible to test multiple scenarios, with the aim of finding one that best explains the data as simply as possible. Accurate representation of the model system must also be balanced with mathematical and computational tractability (Gelman & Shalizi, 2013). Secondly, data is simulated from the model to assess model fit, importantly, to evaluate parameterisation and the choice of summary statistics, and avoid model misspecification. This step may also involve model choice; ABC can be applied here to generate posterior probabilities of candidate models (Csilléry et al., 2010). Posterior predictive checks should be used to evaluate the candidate model. Posterior predictive checks involve further simulation of data from the posterior distribution, and the use of test statistics (not used for model fit) for comparison to the observed data. Model checking is crucial to reject poorly fitting models outright, or to refine the existing model. Several rounds of model improvement, simulation and comparison with empirical data are needed to find a well-fitting model that is useful for testing the original hypothesis. Model checking is important to understand the ways in which the model does not fit the data. All models are a simplification of reality, it is important to identify the key scientific question being addressed, and to critically evaluate the model and its limitations. Once the devised model is considered to fit well, and is useful for making predictions about model system, the posterior distribution can be used to infer parameter values.

3.1.3 Aims

In this chapter I describe a demographic model for wildcats developed within an approximate Bayesian computational framework. Using this model, we aimed to understand the demographic history of both the wild-living and captive wildcat populations in Scotland, including historical patterns of introgression from domestic cats. Specifically, we wished to test the hypothesis that no significant hybridisation took place before population expansion in Scotland in the early 20th century. Additionally, we demonstrate the value of the model as a tool to understand other evolutionary processes in the wildcat population by applying it to calibrate tests for selection.

3.2 Methods

3.2.1 SNP dataset

The ddRAD-seq dataset set described in the previous chapter (2.3.1) constituted the observed data for the ABC model. This consisted of four Scottish domestic cats, 59 individuals from the captive wildcat population and 45 from the wild in Scotland, genotyped at 6,546 SNPs. For details of bioinformatic processing see 2.2.2.

3.2.2 A demographic model for wildcats

A simple model for wildcats was first considered using the two parent populations (domestic and wildcat) and an admixed (hybrid) population. However, it is clear from recent studies (Senn et al., 2019) and previous results (see Chapter 2), that almost all putative Scottish wildcat individuals are found in the captive population. Few, if any, wild-living individuals exist in Scotland without introgression from domestic cats. It was therefore more informative to model the history of the wild and captive populations separately, treating the three sample sources (domestic, wild-living and captive) as separate populations. The captive population was an important element of the model as these individuals represent the population most genetically distant to domestic cats in Scotland. It is important to understand their relationship to the wild-living population, as well as the demographic history of the Scottish population (captive and wild) as a whole.

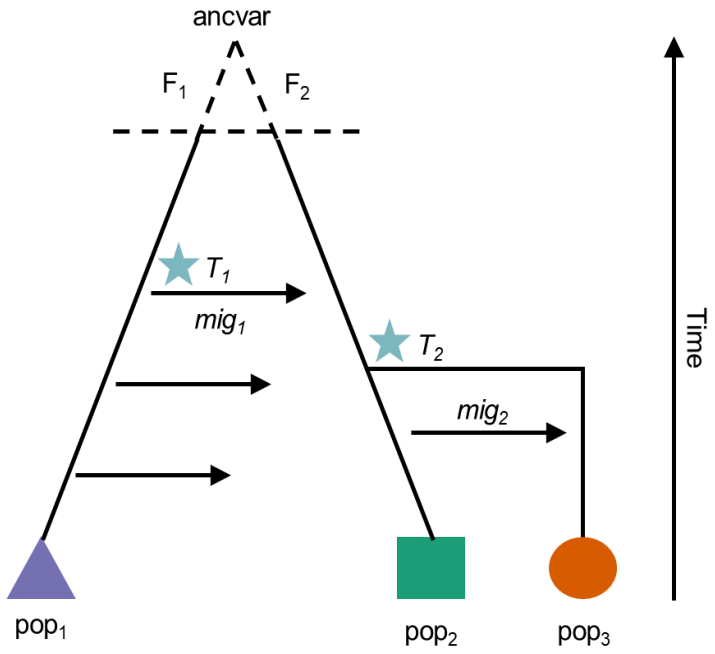

Figure 3.4. Demographic model for wildcats. *Ancvar*, F_1 and F_2 are used to initialise the starting gene frequencies in each parent population (modelled as drift, F_1/F_2 from an ancestral baseline, *ancvar* [see text]). The two parent populations (pop_1, *F. catus*, and pop_2, *F. silvestris*) diverge under a neutral model of evolution. Starting SNP frequencies are generated using three nuisance parameters, F_1, F_2, and *ancvar*. Gene-flow (introgression) from domestic cats begins at time T_1, at a rate of mig_1 for every subsequent generation. At time T_2 the captive population (pop_3) is formed from a random sample of wild-living cats. Limited gene-flow from the wild population into the captive population occurs at a rate of mig_2. Note T_1 is shown here occurring before T_2, but the prior on T_1 allowed hybridisation to begin at any point during the simulation, before or after T_2.

A demographic model for wildcats was created within an ABC framework (Beaumont et al., 2002). Model development is described below (3.2.3). Under the final model (Fig. 3.4), wildcat and domestic cat populations diverge under a neutral model of evolution for 500 generations. Generation time for a wildcat was estimated to be three years (Beaumont et al., 2001; Nussberger, Currat, Quilodran, Ponta, & Keller, 2018), 500 generations (or ~1,500 years) therefore approximately spans

the time period domestic cats and wildcats are thought to have been sympatric in Britain (Serpell, 2014). Given the focus on recent demography, and in view of the low mutation rate of SNPs, a two-stage 'mutation free' approach (Beaumont, 2004) was used. We firstly model the divergence of the two populations from a common ancestor, using a computationally efficient method in which the starting SNP frequencies for each population were simulated from a beta-binomial distribution, parameterised by F_{ST} (Balding & Nichols, 1995). We achieve this by simulating from three beta distributions, the parameters for which we treat as nuisance parameters in the statistical model. The metapopulation SNP frequency, X, is simulated from beta(1, *ancvar*), which assumes that the non-reference allele is typically rarer (empirically confirmed). Parameters F_1 and F_2, are population-specific F_{ST}s (Balding, 2003) modelling drift from the ancestral baseline for domestic and wildcat, giving frequencies beta($X(1-F_1)/F_1$,$(1-X)(1-F_1)/F_1$) and beta($X(1-F_2)/F_2$,$(1-X)(1-F_2)/F_2$), respectively. The finite population frequency is then a binomial sample of size $2pop_1$ and $2pop_2$.

This step initialises an individual-based model of genetic inheritance in which at time T_1 gene-flow from domestics to wildcats begins at a rate of mig_1 per generation. Gene-flow occurs at the same rate in every subsequent generation. At time T_2 the captive wildcat population is established from a random sample (of size pop_3) of wildcat individuals (referred to as the wild-living population from this point forward). There is (limited) gene-flow (mig_2) from the wild-living population to the captive wildcats (reflecting a number of wild-caught founders that have been incorporated into the captive population over its history). Population sizes remain constant throughout the simulation; we do not model any fluctuations in wildcat population size (e.g., recent population expansion), or a decline in the wildcat population as a direct result of hybridisation. Furthermore, unlike Quilodrán *et al*. (2020), we did not consider a spatial model for hybridisation. Previous analysis indicates a complex and patchy pattern of hybridisation, difficult to model on a large scale (Kilshaw et al., 2016; Senn et al., 2019).

3.2.3 Model development

Initially four different models were tested, each with a different approach to incorporating the captive population (Fig. 3.5). Acknowledging the presence of hybrid individuals in captivity, an additional migration parameter, mig_2, was devised, allowing limited gene-flow from the wild-living (hybridising) population. Also, given selection criteria are used to identify individuals included in the captive breeding program (see Chapter 2), a filtered set of observed (target) data was tested, removing probable hybrids in the captive population. Probable hybrids (n=13) were identified using a Q35 threshold of 0.9. Filtering of the target data was used as a proxy for the selection of putative wildcats, in the model the captive population is formed from a random subset of wild-living individuals. All combinations of these two components, an additional migration parameter and filtered target data, were tested under four different models, ~30,000 simulations were

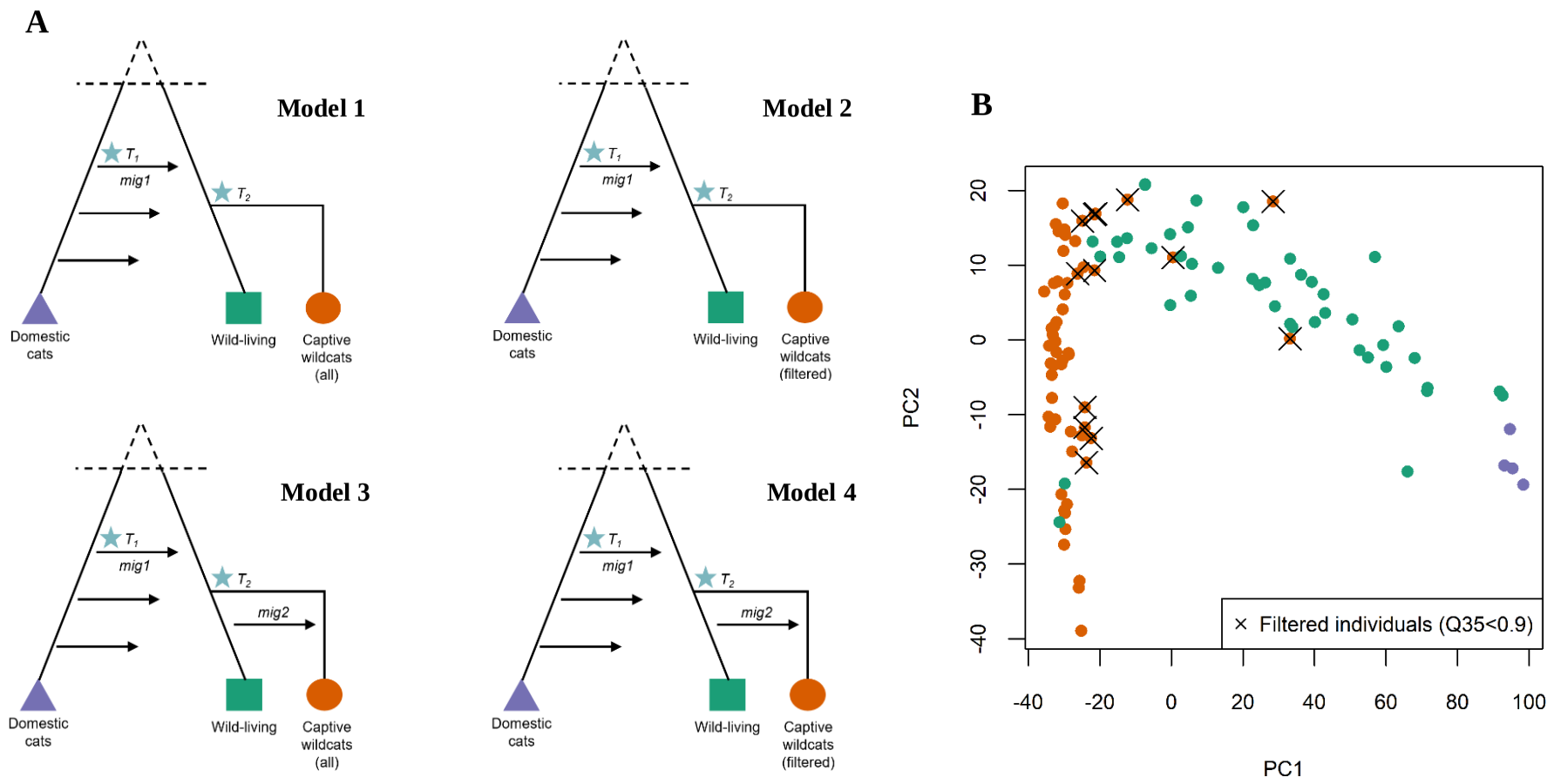

Figure 3.5. Data were initially simulated from four different models (A), with 22 summary statistics (Table 3.1). These were used to identify the best combination of including/excluding an additional migration parameter (mig_2), allowing gene flow from the wild-living population into the captive population, and including/excluding introgressed captive individuals in the observed data. Approximately 30,000 simulations were generated under each model. In the filtered dataset individuals with a Q35 score less than 0.9 were excluded from the target data, these individuals are marked with a cross on the PCA (B).

generated from each model. 22 summary statistics were devised with which to perform ABC (Table 3.1, see also 3.2.6).

Performance of each model was evaluated using the R package *abc* (Csilléry, François, & Blum, 2012). Different target data were used by different models (i.e., with different numbers of captive individuals in the filtered versus unfiltered data), preventing direct comparison using the relative posterior probabilities of each model. Instead, model performance was evaluated using the goodness-of-fit test, *gfit*. *Gfit* generates a null distribution for the distance between the observed data and accepted summary statistics, using simulated data as pseudo-observed data. 100 samples from the simulated data closest to the observed data (tol=0.01) were used to generate the null distribution. The actual distance between the observed and simulated data was calculated, and a p-value generated to indicate whether this significantly differed from the null distribution. This appeared to be the case for all four models (Fig. 3.6), indicating poor model fit. The smallest distances were reported for models 3 and 4 (distances of 11.29, p-val=0.02, and 12.08, p-val=0.02, respectively). These were the models using the additional migration parameter, mig_2.

To improve fit a novel method for dropping summary statistics was devised. This method used the observed summary statistics (target data) and simulated summary statistics (with parameters drawn from the prior) to compute for each point the Mahalanobis distance to its nearest neighbour. The target and simulated summary statistics were scaled to have unit variance prior to PCA rotation. The nearest neighbour distance (nnd) is an estimate of a quantity proportional to density (Silverman, 1998), in this case the prior predictive density. The idea was to compare the nnd of the target to the nnd of all the simulated points. We can then define a highest prior predictive density (HPPD) band, e.g., $HPPD_{0.95}$, such that 95% of all simulated points have nnd < $HPPD_{0.95}$. Nnds were computed, each time leaving out one summary statistic, allowing summaries resulting in the largest distance between the target and simulated data to be identified. The process was iterated (permanently dropping the worst performing summary statistic from the previous round) until nnd of the target < $HPPD_{0.95}$.

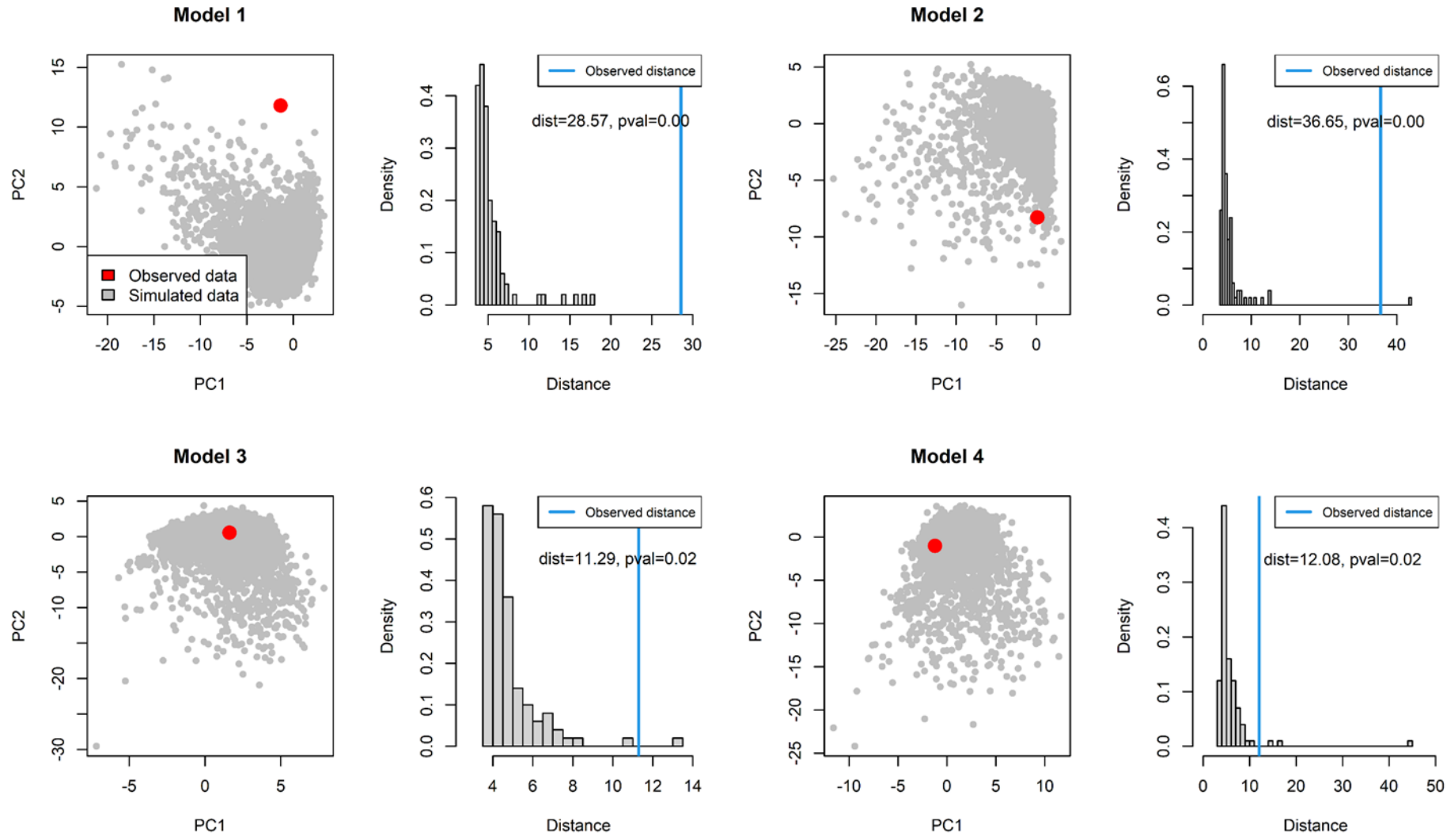

Figure 3.6. Goodness-of-fit test, *gfit* (Csilléry, Francois & Blum 2012). For each model (1-4) the left-hand plot shows the first two principal components following PCA of the projected summary statistics, with target data shown in red. The right-hand plot shows the results of the goodness-of-fit test. The null distribution for the mean distance between the observed data and accepted summary statistics (tol=0.01) was generated using 100 replicates of pseudo-observed data. The blue line shows the actual mean distance between the observed and simulated data (distance and associated p-value shown). For each model the observed data was outside the null distribution (p-value < 0.05). Models 3 and 4, using parameter mig_2, appeared to fit better.

Using this method to drop the five worst performing summary statistics from models 3 and 4 significantly improved model fit (as assessed using *gfit*, Fig. 3.7). Model 4 (using mig_2 and filtered target data) appeared to be the best fitting, reporting the smallest distance between the simulated and target data (5.38, p-val=0.24). This model was used for large scale simulation and parameter inference. Among the worst performing summary statistics were population-specific PCA measures, for consistency all of these were removed from the final analysis. For the full list of dropped summary statistics see Table 3.1.

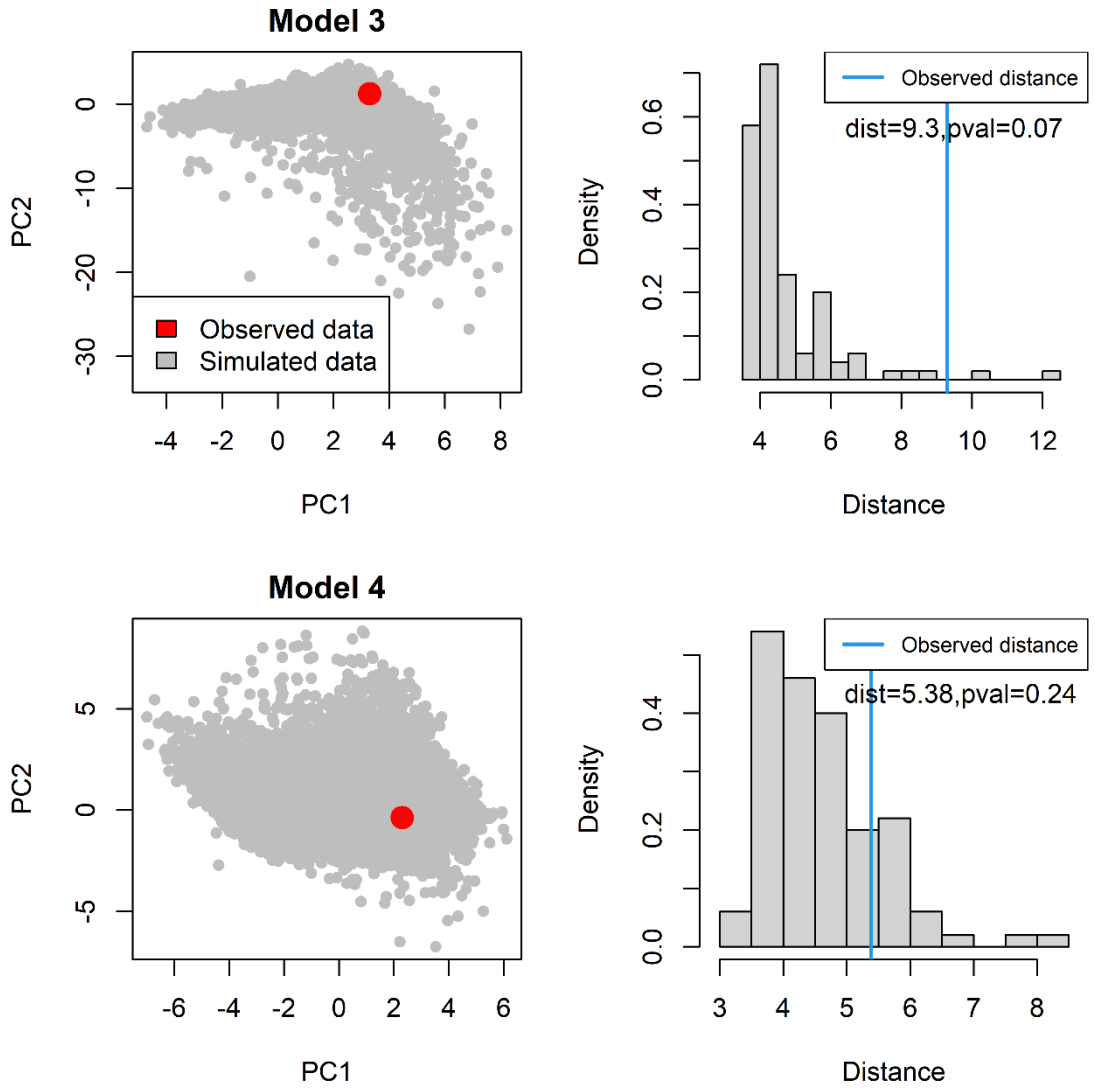

Figure 3.7. Dropping poorly performing summary statistics improved fit for both models; mean distances were smaller than reported for the full set of summary statistics (Fig, 3.6), and these distances were not significantly different from the null distribution (p-value > 0.05). Model 4, using the second migration parameter (mig_2) and filtering introgressed captive individuals from the target data, appeared to be the best fitting and was used for larger-scale simulations.

Table 3.1. Summary statistics used for ABC. Following a novel approach dropping summary statistics to improve model fit, 14 of the 22 summary statistics initially devised were used in the final analysis.

No.	Name	Description	Included in final analysis?
1	ss1	Pairwise Euclidean distance between individuals was computed using the SNP genotype (0,1,2) matrix (scaled to have unit variance). A hierarchical clustering algorithm (*hclust*, in R, with the complete linkage option) was then applied to the distance matrix. ss1 was the mean node height divided by the square root of the total number of SNPs.	No
2	ss2	Using the results from *hclust* (described for ss1) we recorded the number of cluster groupings at the height corresponding to each node of the cluster tree (using the function *cutree* in R). For each height, we then computed an effective number of groups as $1/\sum x_i^2$ where x_i is the relative frequency of the *i*th group. For each height we then computed the ratio of the actual number of groups to the effective number of groups. This was regarded as a measure of structuring in the cluster tree, and the natural logarithm of the mean of this across all heights was reported as ss2.	Yes
3	ss3.1	Following PCA, eight evenly spaced quantiles for PC1 values were computed for all samples. To create a statistic robust to PCA rotation we record the difference between the two outermost quantiles (ss3.1), working inwards to the innermost two quantiles (ss3.4). In this way we capture information about the distribution of individuals across PC1	Yes
4	ss3.2		Yes
5	ss3.3		Yes
6	ss3.4		Yes
7	ss4.1	As above, for PC2	Yes
8	ss4.2		Yes
9	ss4.3		Yes
10	ss4.4		Yes
11	ss5	Pairwise genetic distance between populations, using *hierfstat* in R (Goudet, 2005)	Yes
12	ss6		Yes
13	ss7		Yes
14	ss8	A statistic related to the effective population size corresponding to the expected LD, using the formula $\frac{1}{3\overline{R^2}-1/n}$ where n is the sample size.	Yes
15	ss9	The squared pairwise Neyman-Pearson correlation coefficients were computed from the SNP genotype (0,1,2) matrix (corresponding to Burrow's R^2 for unphased SNPs). The standard deviation of these was then reported as ss9	Yes
16	ss10.1	As ss3, but using domestic individuals only	No
17	ss10.2	As ss4, but using domestic individuals only	No
18	ss11.1	As ss3, but using wild individuals only	No
19	ss11.2	As ss4, but using wild individuals only	No
20	ss12.1	As ss3, but using captive individuals only	No
21	ss12.2	As ss4, but using captive individuals only	No
22	ss13	Total number of simulated SNPs	No

3.2.4 Simulating data with SLiM

Data were simulated using SLiM (Haller & Messer 2017), a toolkit for evolutionary modelling. SLiM is individual-based, forward-simulating, and implements a Wright-Fisher model of evolution (amongst others) in which generations are non-overlapping, individuals are diploid, and offspring are generated through recombination and mutation of parental genotypes. 12,000 independent, unlinked, sites were modelled per individual. A large number of variable sites needed to be initialised in order to replicate the observed SNP data as a proportion of sites reached fixation over the course of a simulation. After 500 generations the genotypes of 46 captive wildcats, 45 wild-living and four domestic cats were sampled at random, and summary statistics were calculated in R. The total number of simulations used for ABC was 509,070.

3.2.5 Prior distributions

Prior distributions for demographic parameters were chosen based on existing knowledge of the model system. The prior for *ancvar* followed an exponential distribution. The F_{ST} between captive wildcats and domestic cats reported by this study was 0.446 (Table 1.1, Chapter 2), therefore priors for F_1 and F_2 followed a beta distribution sampling values around 0.2. Priors for effective population sizes followed a log normal distribution, with a fixed lower bound for the captive population of 60 individuals (preventing simulations with a fewer number of individuals then the observed data). Fairly wide priors were used for wild-living and domestic cat population sizes; accurate estimates of census population size, both historic and current, are difficult to obtain, especially considering difficulties distinguishing wildcats from hybrids (Macdonald et al., 2010). Mig_1 was a parameter of particular interest as it corresponded to the rate of introgression from domestic cats. The prior for this parameter followed a beta distribution allowing a migration rate of up to 0.6 per generation. The UK studbook for wildcats informed priors relating to the captive population (Barclay, 2019). Gene-flow between the wild-living and captive populations (mig_2) was constrained to be relatively small (around 0.01); we know from studbook records that only a small number of additional wild founders (between one and six) have been incorporated at any one time over the population's history. A more informative prior was given to T_2 as we know the captive population was established in 1960. Importantly, a wide prior was chosen for T_1, allowing hybridisation to begin at any point in the simulation, before or after T_2. The priors for T_1 and T_2 were completely independent. For a summary of prior distributions see Table 3.2.

Table 3.2. Full set of model parameters, the prior distribution and posterior mean are given for each parameter.

Parameter	Description	Prior distribution	Posterior mean (95% HPD)
ancvar	Generates baseline ancestral variation	Exponential† λ=0.1	4.155 (2.441 -5.801)
F_1	Drift from baseline (pop_1)	Beta α=2, β=10	0.211 (0.047 – 0.391)
F_2	Drift from baseline (pop_2)		0.183 (0.036 – 0.336)
log(pop_1)	Log population size	Normal μ=6.5, σ=0.5	6.429 (5.813 – 7.167)
log(pop_2)			6.580 (5.924 – 7.426)
log(pop_3)		Normal‡ μ=4.6, σ=0.5	4.469 (3.986 – 5.099)
T_1	Onset of gene flow from pop_1 to pop_2 (number of generations)	Exponential λ=0.02	3.326 (1.209 – 5.602)
T_2	Time pop_3 is established from a sample of pop_2 (number of generations)	Gamma α=9, θ=0.5	19.272 (9.430 – 30)
mig_1	Migration (per generation) pop_1 to pop_2	Beta α=5, β=20	0.128 (0.067 – 0.192)
mig_2	Migration (every three generations) pop_2 to pop_3	Gamma§ α=1, θ=1	0.012 (0 – 0.037)

† (exponential distribution with rate parameter λ=0.1)+1 to avoid values of ancvar less than 1
‡The lower bound of this distribution was limited to 60 to avoid simulating a population of captive individuals smaller than the target data
§ (gamma distribution with shape parameter α=1 and scale parameter θ=1)/size of captive population

3.2.6 Summary statistics

Given the strong separation of domestic cats and wildcats across the first principal component (2.3.2), a set of PCA-based summaries were devised (measures of the distribution of points across PC1 and PC2). Additional summaries included pairwise genetic distance (F_{ST}) and linkage disequilibrium measures, for a detailed list see Table 3.1. The final number of summary statistics was 14, permanently dropping eight with a detrimental impact on model fit (3.2.3). Owing to the correlation within and between parameters and the final set of summary statistics (Fig. 3.10, Appendix 3), projection was used to improve posterior estimates, following the approach of Fearnhead and Prangle (2012). Projection involves fitting a regression model between each parameter and the summary statistics. The regression model gives an estimate of the posterior mean for a given set of summary statistics. This prediction for each parameter can be viewed as a projection of the 14-dimensional summary statistics onto a 10-dimensional set of new summary statistics (Blum, Nunes, Prangle, & Sisson, 2013). To fit the regression model a subset of simulated points closest to the observed data (tol=0.2) were used.

3.2.7 Parameter inference

Parameter inference was carried out in R using the package *abc* (Csilléry et al., 2012). The closest 5,091 points (1%) were used to generate the posterior distributions, correcting for an imperfect match between the projected summary statistics and observed data using non-linear regression (neural network) (Blum et al., 2013; Raynal et al., 2019).

3.2.8 Using the wildcat model to calibrate tests for selection

A model for wildcats is important to understand the demographic history of this species in Britain. It is also a valuable tool for understanding other processes in wildcat or hybrid populations. For example, we applied simulated data, generated under the best-fitting neutral model, to calibrate methods for detecting selection in admixed populations. Currently, the consequences of introgression of domestic cat genes into wildcat populations, or the fitness of hybrid offspring, are poorly understood. It is unknown whether introgressed domestic cat genes confer any selective advantage or disadvantage.

The data were screened for selection using two methods, the R program *pcadapt* (Luu, Bazin, & Blum, 2017) and bgc (Gompert & Buerkle, 2011). *Pcadapt* and bgc are complementary methods to examine patterns of variable introgression in admixed populations, generating a null model from the data in order to detect outlying regions potentially associated with adaptive variation. *Pcadapt* uses a PCA-based method to identify candidate loci that are outliers with respect to population structure. Bgc implements a Bayesian genomic cline model to quantify locus-specific patterns as a function of genome-wide admixture. These methods were also applied to ten simulated datasets selected at random from the posterior distribution of the model.

For *pcadapt* the first three principal components were used in the analysis, following Cattell's Rule that smaller eigenvalues, relating to random variation, lie on a straight line and those relating to population structure depart from the line (Cattell, 1966). We focused on outliers correlated with PC1, i.e., SNPs with large variation in allele frequency between the parent populations (included in the analysis) and which therefore represent 'wildcat' or 'domestic' loci under selection in the hybrid population. P-values $< 1 \times 10^{-6}$ were investigated as outliers (equivalent to 0.01 Bonferroni corrected).

Unlike *pcadapt*, bgc requires parental populations to be defined *a priori*. For this ADMIXTURE Q scores (Q6546) were used to classify individuals as wildcat, $Q6546 \geq 0.9$, hybrid, $0.9 > Q6546 \geq 0.1$, or domestic, $Q6546 < 0.1$. Using these thresholds 58 individuals were classified as wildcat, 44 as hybrids and six as domestic cats. For simulated datasets, the hybrid population corresponded to the simulated wild-living population and the captive population was used as a proxy for wildcats. Following the approach described by McFarlane *et al.* (2021) bgc was run independently five times for the observed data, using the run with the widest reported confidence interval per loci to identify those deviating from the genome wide expectation. Bgc was run once per set of simulated data. Each bgc run consisted of 50,000 iterations, with a burnin of 25,000 and recording MCMC samples every 200th iteration. Loci in 'excess' were defined as loci with confidence intervals that did not span zero, 'outlying' loci were defined as loci with α and/or β estimates outside the 95% distribution across all SNPs. α and β are genomic cline parameters that

describe the cline centre and rate (Gompert & Buerkle, 2011). Positive values of α indicate an increased probability of ancestry from one parental population (in this case wildcat), negative values a decrease in probability (i.e., probable domestic ancestry), given the genome-wide expectation. β values indicate an increase (positive estimates) or decrease (negative) in the rate of transition from one parent population to the other.

3.3 Results

3.3.1 Demographic modelling

The demographic model was capable of simulating data within the range of the observed data and appeared to fit well (Fig. 3.8; Fig. 3.11, Appendix 3). The first two axes of the posterior predictive PCA plots (Fig. 3.9) show broadly the same patterns as the observed data, particularly with respect to the distribution of wild-living individuals across PC1.

Prior and posterior distributions for the three parameters of interest (T_1, T_2, and mig_1) are shown in Fig. 3.9. The posterior mean for T_1, the onset of gene flow from domestic cats to wildcats, was 3.3 generations (95% HPD: 1.21– 5.6). For T_2, the time the captive population was established, the mean was 19.3 generations (95% HPD: 9.4 – 30). Note that the estimate for T_1 is not constrained by the prior to any marked degree, whereas the historically informed prior for T_2 has a stronger effect. The migration rate of domestic cats into the wild-living population was estimated to be 0.13 (95% HPD: 0.076 – 0.19). For posterior means and distributions for all parameters see Table 3.2 and Fig. 3.12, Appendix 3.

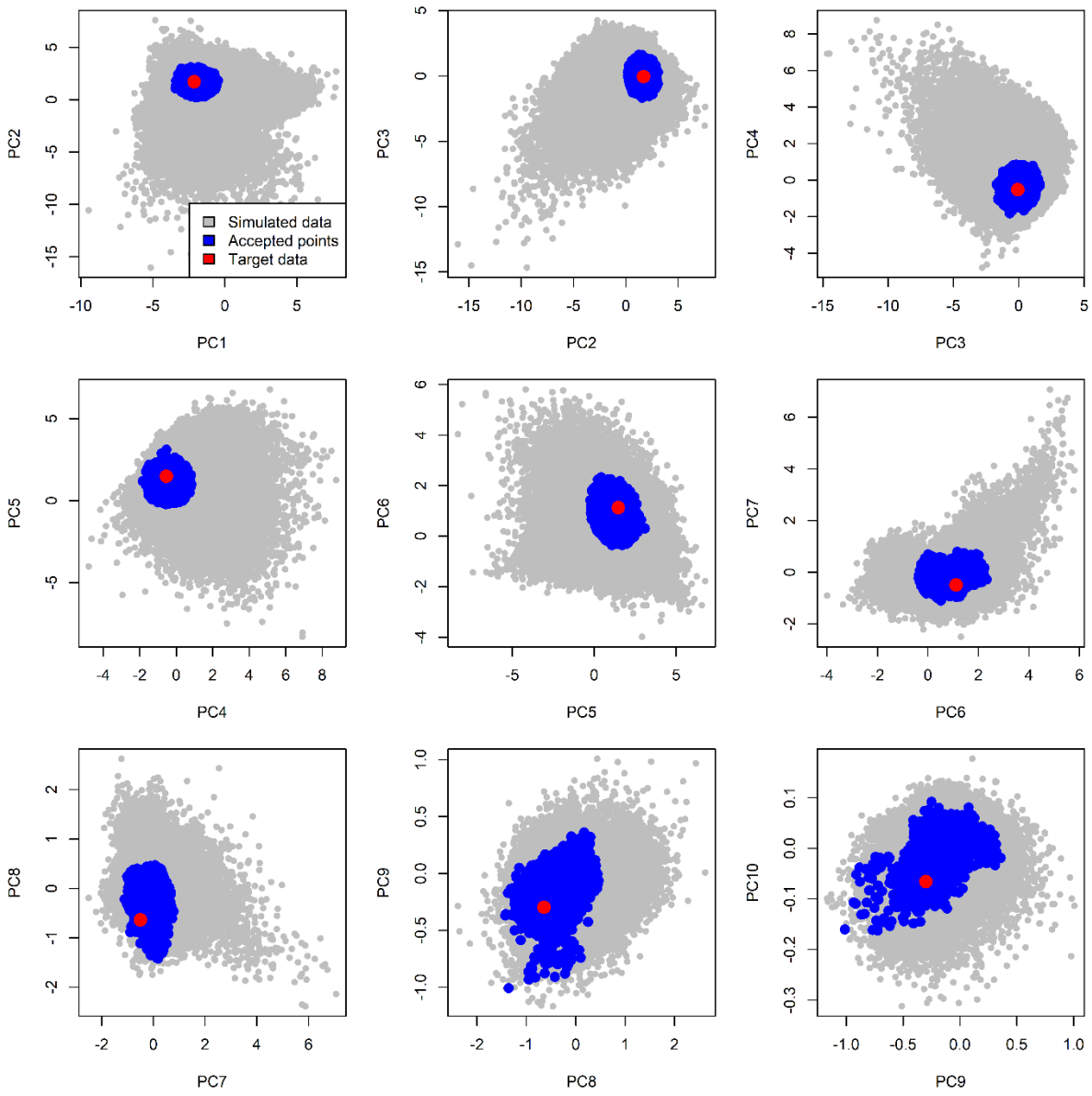

Figure 3.8. Model fit can be assessed visually using PCA of the summary statistics. Simulated data are shown in grey, with accepted points (tol=0.01) highlighted in blue and target data shown in red. If the target data (i.e., observed summary statistics) lie outside of the cloud of accepted points this indicates poor model fit.

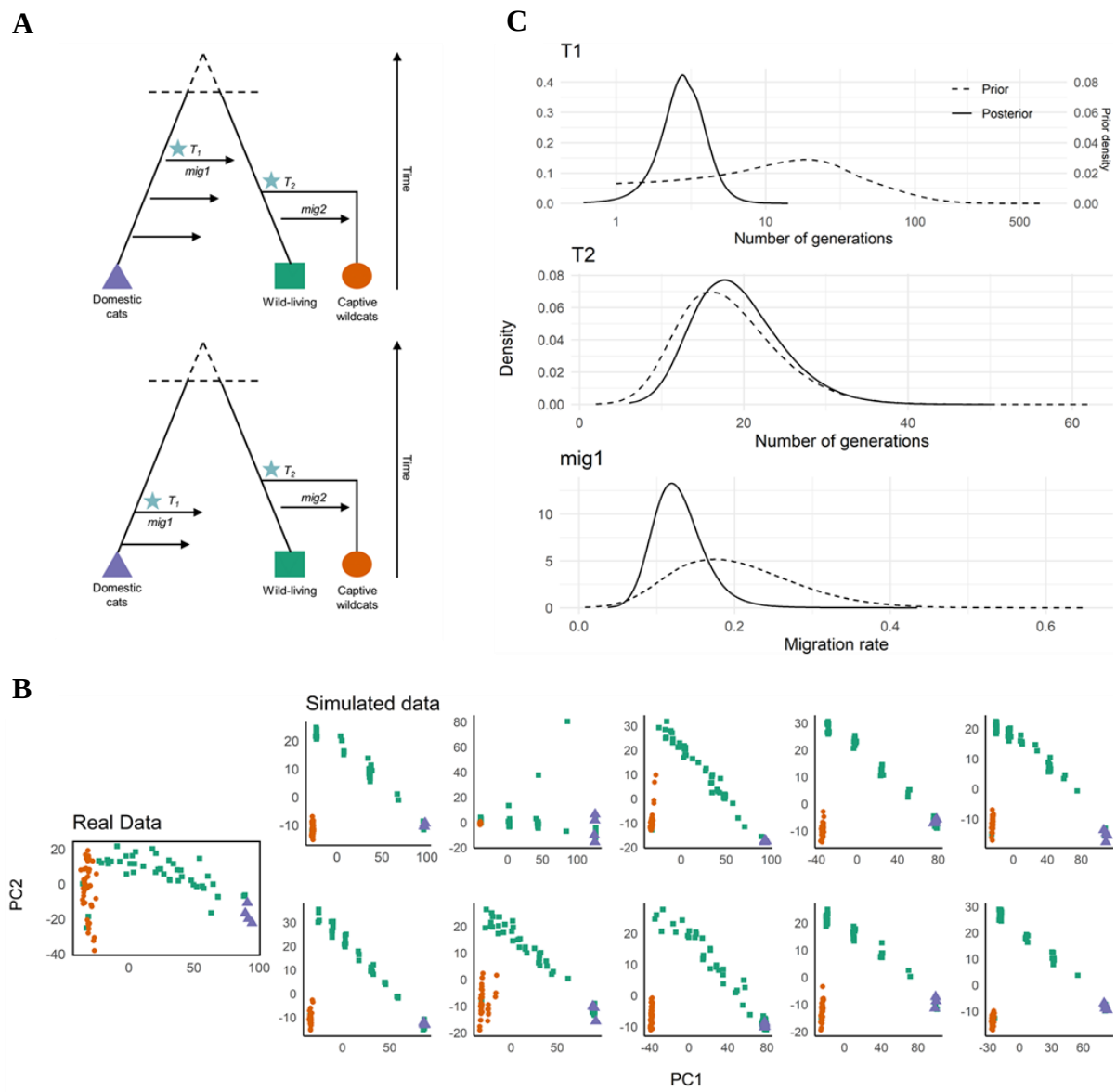

Figure 3.9. Modelling wildcat demography. (A) The model under which data were simulated; two possible scenarios are shown, one where gene-flow from domestics precedes the founding of the captive population i.e., T_1 precedes T_2, (top), and an alternative scenario where T_1 occurs after T_2 (bottom). On both models an important subset of parameters is shown, for the full set of parameters see Fig. 3.4. (B) PCA plots for the real data (left) and for a random sample from the posterior distribution (right). The model is broadly able to simulate the same patterns as we observe in the real data. (C) Prior and posterior distributions following ABC, dashed lines indicate the prior. Curves were fitted in R using *locfit* (Loader, 2013). Modelling supported recent introgression in the Scottish wildcat population following high gene-flow from domestics.

3.3.2 Evidence for natural selection

Using the observed data *pcadapt* reported three outlying SNPs most correlated with PC1 (Table 3.4, Appendix 4; Fig. 3.13, Appendix 4). Bgc found the majority of SNPs (90.1%) to be in excess for α and/or β estimates; 5901 were in excess for α estimates, 4318 for β estimates, 3935 were in excess for both α and β. A total of 280 (4.3%) SNPs had outlying values of α, 243 (3.7%) were β outliers, 6 (0.09%) of these had outlying values of both α and β (five with negative values of α and β, one with

positive α and negative β) (Table 3.4, Appendix 4). Two SNPs were found to be outliers by both *pcadapt* and bgc. These results are summarised in Table 3.3.

Table 3.3. Summary of scans for selection using *pcadapt* and bgc. Results are shown for the observed and simulated data. Reported for the simulated data is the mean across ten datasets, sampled at random from the posterior distribution of the demographic model. The proportion of total SNPs (%) is given in brackets. A similar proportion of outlying SNPs was reported across the two methods for both the observed data and data simulated under a neutral model of evolution.

pcadapt			
	Outlying SNPs correlated with PC1		
Observed	3 (0.05)		
Simulated	8.1 (0.1)		
bgc			
	In excess for α and/or β estimates	**Outlying α and/or β values**	**Outlying α and β values**
Observed	5901 (90.1)	517 (7.9)	6 (0.09)
Simulated	6336.9 (85.4)	586.4 (7.9)	17.6 (0.2)

Simulated data generated a comparable number of outliers with both methods. Using *pcadapt*, nine out of the ten simulated datasets contained at least one SNP correlated with PC1 found to be outlying with respect to population structure (Table 3.5, Appendix 4). The mean proportion of outlying SNPs across the ten simulations was 0.1%. Similarly, using bgc, the mean proportion of SNPs with α and β estimates in excess was 85.4%. 4.3% and 3.9% of SNPs had outlying values of α and β, respectively. 0.2% were outlying for both α and β estimates (Table 3.3; Table 3.6, Appendix 4).

Outlier SNPs are candidates for loci under selection though extreme outliers can also be generated via neutral processes, a result of pre-existing population structure, emphasised by genetic drift. Using a neutral model of wildcat demography to calibrate these tests, we show that the number of outlying SNPs detected in the observed data does not deviate from expectations under neutrality.

3.4 Discussion

3.4.1 The recent history of wildcat hybridisation in Scotland

Demographic modelling supported a rapid emergence of the hybrid swarm in the Scottish wildcat population as a result of high gene-flow from domestic cats. We take generation time for wild-living cats to be ~3 years (Beaumont et al., 2001; Nussberger et al., 2018). The T_1 posterior mean (3.3

generations, or ~10 years) was implausibly recent, yet extensive model-checking (Figs. 3.5-3.8; Fig. 3.10-3.11, Appendix 3) suggests the model generally fits the observed data well. The model appears to consistently underestimate admixture times older than ~20 generations (see below), however, the posterior mean falls within the range that can be well predicted (Fig. 3.11, Appendix 3). The exact history of hybridisation in Britain remains poorly understood (and is likely to show geographic variation), but hybridisation has been of increasing conservation concern since the 1980s (Hubbard et al., 1992, Kitchener et al. 1992, Easterbee et al. 1991) and is generally thought to be a consequence of wildcat range expansion in Scotland during the early 20th century, coupled with continuing high levels of persecution, especially in eastern Scotland. This does not exclude the onset of significant introgression within the last few decades. Senn *et al.* (2019) generated Q35 scores for 60 samples collected in Scotland between 1895 and 1985. These were predominantly cats shot by gamekeepers and subsequently incorporated into museum collections, so there is potential bias towards individuals with wildcat features, nonetheless, only five samples were classified as hybrids (using LBQ<0.75), and one as a domestic cat.

Mattucci *et al.* (2019) used SNP array data to date admixture in continental European wildcat populations. Individuals were sampled across the five main biogeographic groups (from Mattucci et al., 2016): Iberia, Central Europe, Central Germany, Italy and the Dinaric Alps v. The study found hybridisation across all populations, occurring between three and 22 generations before present. The most recent admixture time reported by this study was 3.15 generations. Mattucci *et al.* (2019) reported admixture times for individuals previously classified as wildcats using microsatellite data, highlighting the power of a SNP-based approach to detect historic and/or complex patterns of admixture (Gärke et al., 2012; Haasl & Payseur, 2011). In an example of another hybridising species, Galaverni *et al.* (2017) date recent admixture between wolves and dogs in Italy to the 1940s, but peaking in the 1990s.

A recent hybridisation time for Scottish wildcats only seems likely in the face of high gene-flow from domestic cats. Our model estimated gene flow to be 13% (95% HPD: 7-19%). This estimate implies 13% of gene copies in wild-living cats come from the domestic population per generation. Quilodrán *et al.* (2020), using a forward simulating approach to model introgression in the Swiss Jura wildcat population, estimated the rate of introgression to be 6%. At this lower rate of introgression, it took 26 generations for the wildcat population to become 50% introgressed.

The demographic model for Scottish wildcats has limited power to detect ancient or complex patterns of admixture. Results presented here suggest our model is unable to detect signals of admixture beyond ~20 generations or c. 60 years (Fig. 3.11, Appendix 3). Haplotype and linkage disequilibrium information (from sequence data) are needed for accurate dating of admixture events,

especially to separate historical admixture from the very recent (Hellenthal et al., 2014; Loh et al., 2013).

Tentative evidence is presented here that the 'hybrid swarm' effect can develop rapidly following the breakdown of isolating mechanisms between two species, as has been observed in other hybridising species, such as deer (Smith, Carden, Coad, Birkitt, & Pemberton, 2014), loaches (Kwan, Ko, & Won, 2014) and honey-bees (Pinto, Rubink, Patton, Coulson, & Johnston, 2005). Our results may also support a recent acceleration of hybridisation in Britain. Though it is difficult to conclude using the current model whether historical admixture has occurred (and to what extent), it is clear there has been significant recent introgression within the last few decades.

An important feature of the model was the captive wildcat population. There is significant interest surrounding this population, which comprises individuals that are among the last putative wildcats in Britain, especially regarding its value to continuing conservation efforts. It is therefore important to understand the extent to which hybridisation has impacted this population. It is clear from the previous chapter that hybrids are present, though the number appears to be low (2.3.2). From the ABC posterior distribution, T_2 (the time the captive population is established) occurs consistently before gene-flow from domestic cats begins (T_1). This suggests the formation of the captive population in the 1960s and 1970s may have occurred prior to significant recent admixture, and that this population is an important reservoir of wildcat genes in Britain (probably aided in recent years by accurate tests for hybrids, see Chapter 2). How closely modern captive animals resemble the British post-glacial population of wildcats, especially considering sympatry with domestic cats over the last 2000 years, remains to be determined.

3.4.2 Modelling approach

Overall, the demographic model developed for wildcats appeared to fit well (Figs. 3.5-3.8; Fig. 3.10-3.11, Appendix 3). The modelling approach we have taken has been to assume that our data does not have sufficient information from mutations occurring over the period of hybridisation to warrant a detailed evolutionary model (Beaumont, 2004). Although linkage has been assumed absent, our model allows for linkage disequilibrium due to finite population size and migration (Waples & England, 2011), which is why we have favoured an individual-based simulation using SLiM, rather than using the coalescent to simulate independent SNPs. The posterior predictive checks (Fig. 3.9) show the broad patterns, in the terms of introgression, were recovered when simulating from the posterior distribution, as judged by the distribution of individuals across PC1.

The model shown in Fig. 3.4 performed best compared to those shown in Fig. 3.5, making use of an additional migration parameter, *mig2*, and a set of Q35 filtered captive individuals. This appeared to be an effective way to model the incorporation of additional wild founders to the captive

population. The goodness-of-fit test (Csilléry et al., 2012) was valuable to make comparisons between models using different target data (Fig. 3.6-3.7).

The distance-based method to drop poorly performing summary statistics was shown to improve model fit (Fig. 3.7). ABC methods are highly sensitive to the choice of summary statistics, which is a somewhat arbitrary step of model development. Similar to methods developed by Joyce & Marjoram (2008) and Nunes & Balding (2010), we aimed to identify a useful subset of summary statistics from a larger set of trial statistics. Nunes & Balding (2010) identify subsets that minimise the entropy of the posterior distribution, using a (computationally expensive) two stage method testing all possible subsets. Like Joyce & Marjoram (2008), we employ a stepwise approach. Joyce & Marjoram (2008) propose incorporating summary statistics until approximate sufficiency is reached, but this means the order in which summaries are tested determines the final subset. In the method described here all summary statistics are initially considered, and the worst performing removed first. Using the approximate sufficiency approach, different subsets are selected when tested with different simulated datasets. It is unlikely that the subset used here represent a single optimal set, and likely that other combinations of summary statistics would also have improved model fit. The main drawback of methods to subset summary statistics is that they ultimately result in the loss of information, hence the motivation for methods to project a larger number of summary statistics onto a lower-dimensional space (Blum et al., 2013).

Many of the worst-performing summary statistics in this analysis were population specific measures of the distribution of individuals across PC1 and PC2 (Table 3.1). In particular, the distribution of the captive population across PC2 was a difficult feature to replicate in the model. It is likely that this is due to genuine, but as yet uncharacterised, structure within the captive wildcats (see Chapter 2) not captured in the model. Our understanding of demographic and/or adaptive processes in the captive wildcat population would benefit from additional sampling and modelling. Here, we have tried to focus on a simple model to explore wildcat demography and patterns of recent hybridisation in the wild. Broad patterns of hybridisation in the wild population are best captured by the distribution of individuals across PC1, which our model is able to replicate well (Fig. 5.9B). Further improvements may be possible, at the cost of increased parameterisation, by considering, for example, variable population size or migration rates.

Quilodrán *et al.* (2020) used a spatial model to quantify introgression. Although this would be challenging at the scale of the model presented here, especially considering the complex patterns of introgression observed in the wild (see Chapter 2), it may be helpful to apply the approach of Quilodrán *et al.* (2020), in conjunction with parameter estimates from the current model, to focus on a geographical area of interest to better understand hybridisation dynamics in a priority area for conservation management.

3.4.3 Calibrating tests for selection

Simulated data were applied to understanding methods for detecting selection in admixed populations, specifically *pcadapt* (Luu et al., 2017) and bgc (Gompert & Buerkle, 2011). These methods are designed to be robust to demographic biases and handle genetically continuous, admixed populations. However, simulation results, based on our best-fitting demographic model for wildcats, show evidence of a high number of false-positives in this setting (Table 3.3), even using a conservative approach to controlling false discovery rate. For these analyses the wildcat model was useful for deriving a null distribution specific to Scottish wildcats.

Even at neutral loci the demographic history of a population can cause allele frequency to vary hugely in space due to genetic drift and/or migration (Hoban et al., 2016; Lotterhos & Whitlock, 2014), as demonstrated by the variability in outcomes from the simulated data (each using a different set of demographic parameters sampled from the posterior distribution). Differences in allele frequencies between domestic cats and wildcats are not surprising considering the genetic differentiation between the two populations, and do not necessarily correspond to deviations from neutrality. Previous simulation studies (Gompert & Buerkle, 2011; McFarlane, Senn, Smith, & Pemberton, 2021) have demonstrated patterns of introgression are highly stochastic and subsequently exaggerated by genetic drift, and this is especially true in cases of recent admixture. Based on our current results we do not have the power to make conclusive statements about natural selection in Scottish wildcats, or fitness consequences for hybrid populations.

3.5 Conclusion

Demographic modelling supported an acceleration of hybridisation in Scotland in recent decades. Using unlinked SNP data, we do not have the power to rule out ancient hybridisation with domestic cats, which would require haplotype and linkage information from sequence data (Chapter 5). A wildcat-specific model of admixture is nonetheless a useful tool to evaluate specific statistical approaches in genomic analysis and provides a baseline with which to develop scenarios of increasing complexity, e.g., incorporating selection, fluctuations in populations size or spatial models. In this regard, our study support the conclusions of recent studies of hybridisation (McFarlane et al., 2021; Quilodrán, Nussberger, et al., 2020). Furthermore, it will be straightforward to extend the approach to incorporate whole-genome sequence data in the future.

3.6 References

Balding, D. J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology*, 63, 221–230

Balding, D. J., & Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96, 3–12

Barclay, D. (2019) *Scottish Wildcat Studbook*, Species360 Zoological Information Management System (ZIMS). zims.Species360.org

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41, 379–406

Beaumont, M. A. (2004). Recent developments in genetic data analysis: What can they tell us about human demographic history? *Heredity*, 92, 365–379

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162, 2025–2035

Beaumont, M., Barratt, E. M., Gottelli, D., Kitchener, A. C., Daniels, M. J., Pritchard, J. K., & Bruford, M. W. (2001). Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology*, 10, 319–336

Blum, M. G. B., & François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20, 63–73

Blum, M. G. B., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28, 189–208

Breitenmoser, U., Lanz, T., & Breitenmoser-Würsten, C. (2019). *Conservation of the wildcat (*Felis silvestris*) in Scotland: Review of the conservation status and assessment of conservation activities*. IUCN SSC. http://www.scottishwildcataction.org/media/42633/wildcat-in-scotland-review-of-conservation-status-and-activities-final-14-february-2019.pdf

Cattell, R. B. (1966) The scree test for the number of factors. *Multivariate Behavioural Research*, 1(2), 245-276

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, 25, 410–418

Csilléry, K., François, O., & Blum, M. G. B. (2012). Abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3, 475–479

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution*, 29, 51–63

Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Royal Statistical Society*, 74(3), 419–474

Galaverni, M., Caniglia, R., Pagani, L., Fabbri, E., Boattini, A., & Randi, E. (2017). Disentangling timing of admixture, patterns of introgression, and phenotypic indicators in a hybridizing Wolf population. *Molecular Biology and Evolution*, 34(9), 2324–2339

Gärke, C., Ytournel, F., Bed'Hom, B., Gut, I., Lathrop, M., Weigend, S., & Simianer, H. (2012). Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations. *Animal Genetics*, 43, 419–428

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2003). *Bayesian Data Analysis*. London: Chapman & Hall/CRC Press

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66, 8–38

Gompert, Z., & Buerkle, C. A. (2011). Bayesian estimation of genomic clines. *Molecular Ecology*, 20, 2111–2127

Goudet, J., & Jombart, T. (2020). hierfstat: Estimation and tests of hierarchical F-statistics. R package version 0.5-7. https://CRAN.R-project.org/package=hierfstat

Green, P. J., Łatuszyński, K., Pereyra, M., & Robert, C. P. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25, 835–862

Haasl, R. J., & Payseur, B. A. (2011). Multi-locus inference of population structure: A comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, 106, 158–171

Haller, B. C., & Messer, P. W. (2017). SLiM 2: Flexible, interactive forward genetic simulations. *Molecular Biology and Evolution*, 34(1), 230–240

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, 343, 747–751

Hetherington D., & Campbell, R (2012) *The Cairngorms Wildcat Project Final Report*. Report to Cairngorms National Park Authority, Scottish Natural Heritage, Royal Zoological Society of Scotland, Scottish Gamekeepers Association and Forestry Commission Scotland

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., … Whitlock, M. C. (2016). Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist*, 188(4), 379–397

Hubbard, A. L., McOris, S., Jones, T. W., Boid, R., Scott, R., & Easterbee, N. (1992). Is survival of European wildcats *Felis silvestris* in Britain threatened by interbreeding with domestic cats? Biological Conservation, 61(3), 203–208

Jabot, F., & Chave, J. (2009). Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. *Ecology Letters*, 12, 239–248

Joyce, P., & Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7(1), Article 26

Kilshaw, K., Drake, A., Macdonald, D. W., & Kitchener, A. C. (2010). *The Scottish wildcat: A comparison of genetic and pelage characteristics*. Scottish Natural Heritage Commisioned Report No. 356, 70

Kilshaw, Kerry, Johnson, P. J., Kitchener, A. C., & Macdonald, D. W. (2015). Detecting the elusive Scottish wildcat *Felis silvestris silvestris* using camera trapping. *Oryx*, 49(2), 207–215

Kilshaw, Kerry, Montgomery, R. A., Campbell, R. D., Hetherington, D. A., Johnson, P. J., Kitchener, A. C., … Millspaugh, J. J. (2016). Mapping the spatial configuration of hybridization risk for an endangered population of the European wildcat (*Felis silvestris silvestris*) in Scotland. *Mammal Research*, 61(1), 1–11

Kitchener, A. C., Yamaguchi, N., Ward, J. M., & Macdonald, D. W. (2005). A diagnosis for the Scottish wildcat (*Felis silvestris*): A tool for conservation action for a critically-endangered felid. *Animal Conservation*, 8(3), 223–237

Kwan, Y. S., Ko, M. H., & Won, Y. J. (2014). Genomic replacement of native *Cobitis lutheri* with introduced *C. tetralineata* through a hybrid swarm following the artificial connection of river systems. *Ecology and Evolution*, 4(8), 1451–1465

Littlewood, N. A., Campbell, R. D., Dinnie, L., Gilbert, L., Hooper, R., Iason, G., … Ross, A. (2014). *Survey and scoping of wildcat priority areas.* Scottish Natural Heritage Commissioned Report No. 768

Loader, C. (2020). R package "locfit": *Local Regression, Likelihood and Density Estimation* v 1.5-9.4. R. Available from https://CRAN.R-project.org/package=locfit

Loh, P. R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., & Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193(4), 1233–1254

Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, 23(9), 2178–2192

Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, 17(1), 67–77

Macdonald, D. W., Yamaguchi, N., Kitchener, A. C., Daniels, M., Kilshaw, K., & Driscoll, C. (2010). Reversing cryptic extinction: the history, present, and future of the Scottish wildcat. In Macdonald, D. W. & Loveridge, A. J. (Eds.) *Biology and Conservation of Wild Felids* (pp. 471–491). Oxford: Oxford University Pres

Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America,* 100(26), 15324-15328

Mattucci, F., Galaverni, M., Lyons, L. A., Alves, P. C., Randi, E., Velli, E., … Caniglia, R. (2019). Genomic approaches to identify hybrids and estimate admixture times in European wildcat populations. *Scientific Reports*, 9, 11612

Mattucci, F., Oliveira, R., Lyons, L. A., Alves, P. C., & Randi, E. (2016). European wildcat populations are subdivided into five main biogeographic groups: Consequences of Pleistocene climate changes or recent anthropogenic fragmentation? *Ecology and Evolution*, 6(1), 3–22

McFarlane, S. E., Senn, H. V., Smith, S. L., & Pemberton, J. M. (2021). Locus-specific introgression in young hybrid swarms: Drift may dominate selection. *Molecular Ecology*, 30, 2104-2115

Nunes, M. A., & Balding, D. J. (2010). On optimal selection of summary statistics for approximate Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 9(1), Article 34

Nussberger, B., Currat, M., Quilodran, C. S., Ponta, N., & Keller, L. F. (2018). Range expansion as an explanation for introgression in European wildcats. *Biological Conservation*, 218(2018), 49–56

Pinto, M. A., Rubink, W. L., Patton, J. C., Coulson, R. N., & Johnston, J. S. (2005). Africanization in the United States: Replacement of feral European honeybees (*Apis mellifera L.*) by an African hybrid swarm. *Genetics*, 170(4), 1653–1665

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12), 1791–1798

Quilodrán, C. S., Nussberger, B., Macdonald, D. W., Montoya-Burgos, J. I., & Currat, M. (2020). Projecting introgression from domestic cats into European wildcats in the Swiss Jura. *Evolutionary Applications*, 13, 1–12

Raynal, L., Marin, J. M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10), 1720–1728

Robert, C. P., Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer

Senn, H., & Ogden, R. (2015). *Wildcat hybrid scoring for conservation breeding under the Scottish Wildcat Conservation Action Plan*. Royal Zoological Society of Scotland

Senn, H. V., Ghazali, M., Kaden, J., Barclay, D., Harrower, B., Campbell, R. D., … Kitchener, A. C. (2019). Distinguishing the victim from the threat: SNP-based methods reveal the extent of introgressive hybridization between wildcats and domestic cats in Scotland and inform future in situ and ex situ management options for species restoration. *Evolutionary Applications*, 12(3), 399–414

Serpell, J. A. (2014). Domestication and history of the cat. In D. C. Turner & P. Bateson (Eds.), *The Domestic Cat: The Biology of its Behaviour* (3rd ed., pp. 83–100). Cambridge: Cambridge University Press

Servedio, M. R., Brandvain, Y., Dhole, S., Fitzpatrick, C. L., Goldberg, E. E., Stern, C. A., … Yeh, D. J. (2014). Not Just a Theory—The Utility of Mathematical Models in Evolutionary Biology. *PLoS Biology*, 12(12), 1–5

Sisson, S. A., Fan, Y., & Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6), 1760-1765

Smith, B. B. (1994). *Howe: four millennia of Orkney prehistory excavations 1978-1982*.Edinburgh: Society of Antiquaries Scotland

Smith, S. L., Carden, R. F., Coad, B., Birkitt, T., & Pemberton, J. M. (2014). A survey of the hybridisation status of Cervus deer species on the island of Ireland. *Conservation Genetics*, 15(4), 823–835

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian Computation. PLoS Computational Biology, 9(1). https://doi.org/10.1371/journal.pcbi.1002803

Tanaka, M. M., Francis, A. R., Luciani, F., & Sisson, S. A. (2006). Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, 173(3), 1511–1520

Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2), 505–518

Templeton, A. R. (2009). Statistical hypothesis testing in intraspecific phylogeography: Nested clade phylogeographical analysis vs. approximate Bayesian computation. *Molecular Ecology*, 18(2), 319–331

Vo, B. N., Drovandi, C. C., Pettitt, A. N., & Pettet, G. J. (2015). Melanoma Cell Colony Expansion Parameters Revealed by Approximate Bayesian Computation. *PLoS Computational Biology*, 11(12), 1–22

Waples, R. S., & England, P. R. (2011). Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics*, 189(2), 633–644

Wei, T. & Simko, V. (2021). R package 'corrplot': *Visualization of a Correlation Matrix (Version 0.90).* Available from https://github.com/taiyun/corrplot

3.7 Appendix 3. Demographic modelling

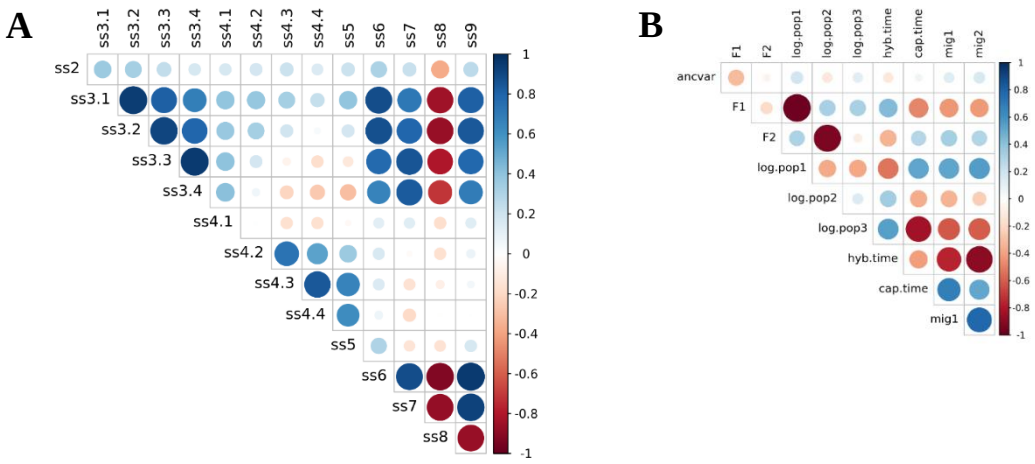

Figure 3.10. Correlation between summary statistics (see Table 3.1), with (B) showing the correlation between projected summary statistics (i.e., the new summaries generated by fitting a regression model between each parameter and the summary statistics shown in [A]). Ideally, summary statistics should be highly correlated with the parameters, and minimally correlated with each other. Overall, the projection appears to have performed well here to reduce dimensionality and limit correlation between summary statistics. Plotted using *corrplot* in R (Wei & Simko, 2021).

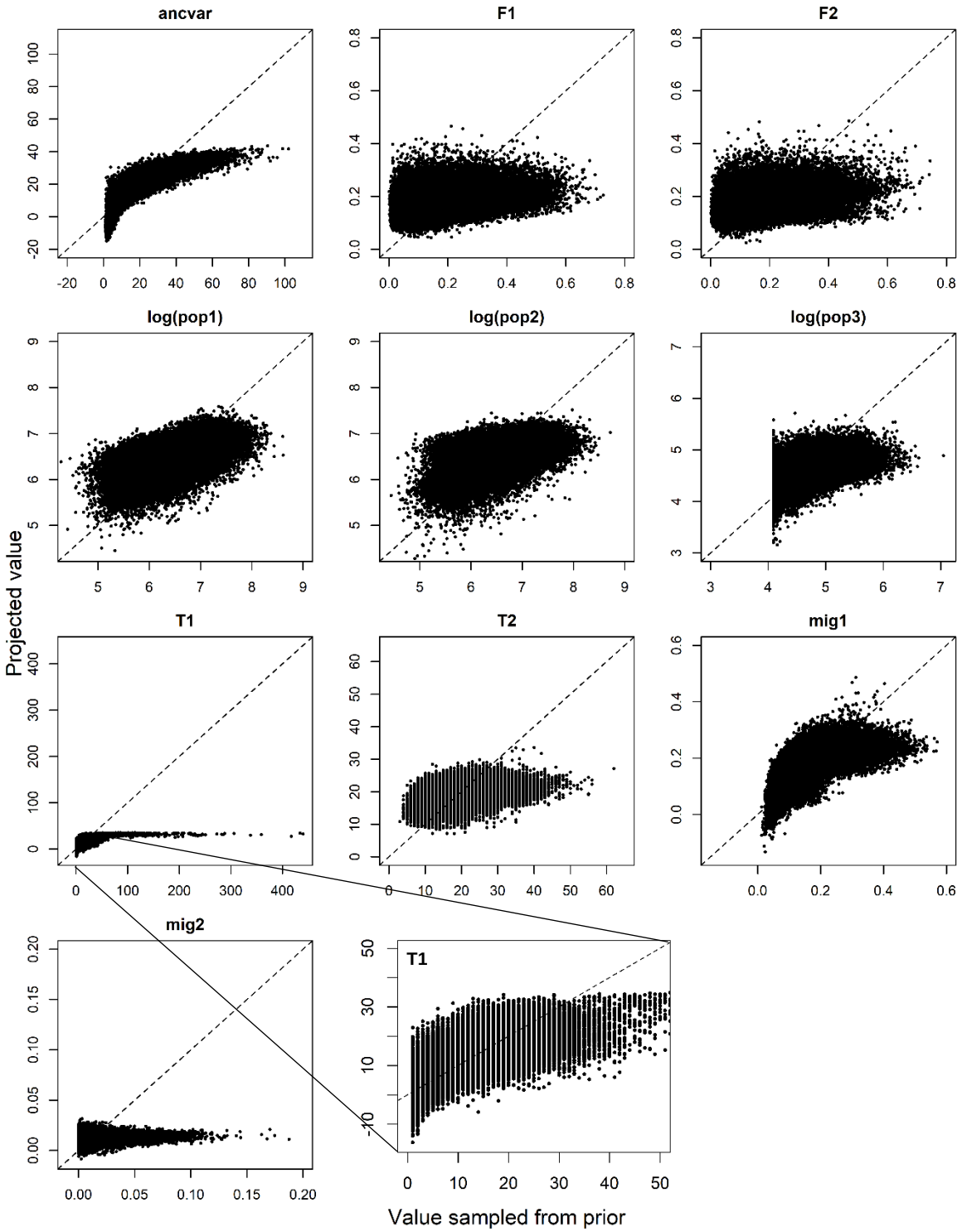

Figure 3.11. For each model parameter, values drawn from the prior distribution are plotted against the parameter value recovered using projection of the simulated data. Points shown are the 20% of simulated points closest to the target data. Generally, the model is able to recover the prior values, indicating a good fit. For T_1, the model does not seem able to recover values greater than ~20 generations, i.e., it performs poorly for older admixture events.

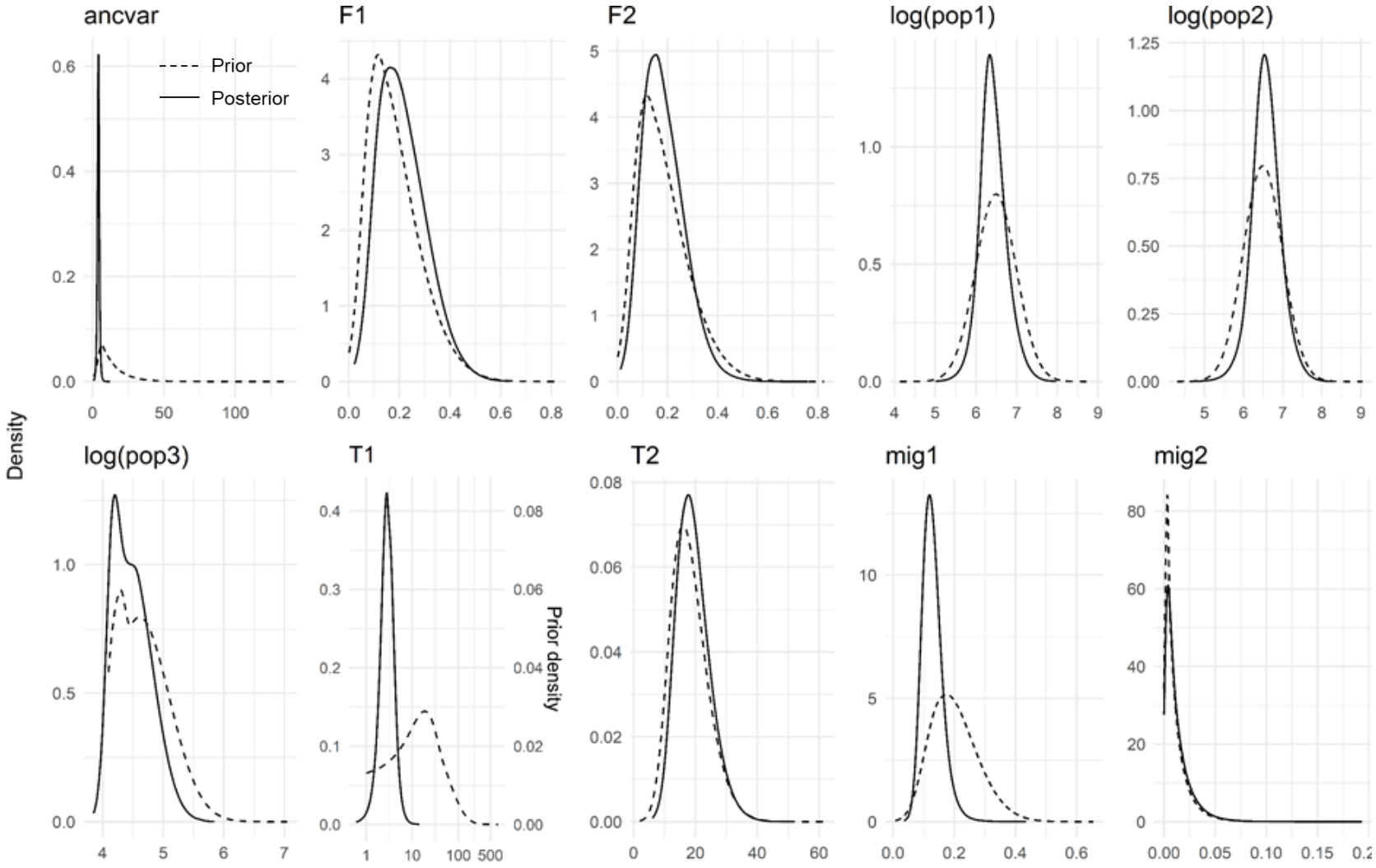

Figure 3.12. Prior (dashed line) and posterior distributions for all model parameters, fitted with *locfit* (Loader, 2020)

3.8 Appendix 4. Calibrating tests for selection

Table 3.4. SNPs of interest from scans for selection (observed data). For *pcadapt* these were SNPs associated with PC1 that were outliers with respect to population structure (p-value < $1x10^{-6}$), and for bgc, SNPs with outlying values (outside the 95% distribution) of both α and β estimates

pcadapt							
No.	**SNP No.**	**Chromosome**	**Position (bp)**	**Allele 1**	**Allele 2**	**PC**	**P-value**
1	2022	B2	52989424	G	T	1	1.403×10^{-11}
2	5147	D4	75300817	G	A	1	1.991×10^{-7}
3	5885	E3	20260711	A	G	1	1.794×10^{-7}
bgc							
No.	**SNP No.**	**Chromosome**	**Position (bp)**	**Allele 1**	**Allele 2**	**α**	**β**
1	1930	B2	16622716	C	T	-1.13	-0.45
2	2022	B2	52989424	G	T	-1.30	-0.56
3	2898	B4	132128702	C	T	0.71	-0.56
4	3133	C1	24488332	G	C	-0.83	-0.48
5	5076	D4	34582615	G	A	-0.77	-0.47
6	5885	E3	20260711	A	G	-1.16	-0.48

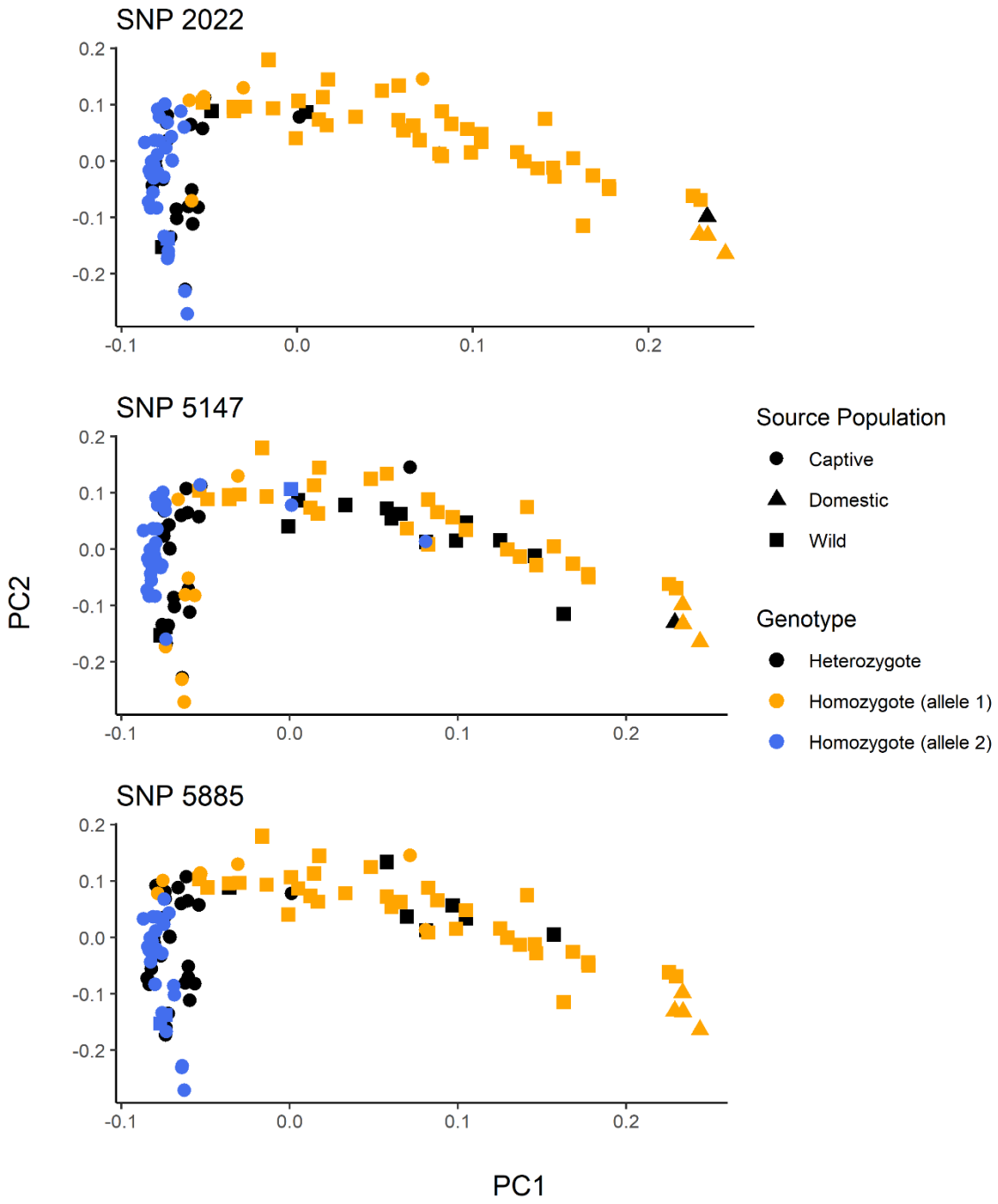

Figure 3.13. PCA plots coloured by individual genotype for each outlying SNP found by *pcadapt* to be correlated with PC1. Parent populations (wildcat and domestic) were included in this analysis and appear on the PCA at either extreme of PC1. There appears to be a high frequency of each of the 'domestic type' alleles in the hybrid population, though results from simulated data indicate this is not unexpected under neutrality.

Table 3.5. *Pcadapt* results from simulated data. Simulated data contained a number of outlying SNPs correlated with PC1. For each of the 10 simulated datasets the total number of SNPs is given, followed by the number of outlying SNPs (associated with PC1) with p-values at least as small as the largest and smallest p-values reported in the observed data (unadjusted p-values). Following a Bonferroni correction (adjusted p-values), the number of outlying SNPs below a threshold of 0.01 is also reported.

Simulation No.	**Total number of SNPs**	**Number of outlying SNPs associated with PC1**		
		Unadjusted p-val ≤1.991 x 10^{-7}	**Unadjusted p-val ≤1.403 x 10^{-11}**	**Adjusted p-val <0.01**
1	7492	8	0	14
2	6858	3	0	15
3	7542	0	0	2
4	7358	5	0	5
5	7101	17	1	24
6	8208	1	0	1
7	7286	0	0	1
8	7570	4	0	3
9	7296	0	0	0
10	7502	14	4	16
Total	74213	52	5	81

Table 3.6. Summary of results from bgc, observed and simulated data. Negative α estimates indicate a SNP with excess domestic ancestry, negative β estimates indicate faster introgression than expected. The opposite is true for positive estimates (i.e., excess wildcat ancestry, slow introgression). A locus was considered in 'excess' if the 95% confidence intervals did not span zero and outlying if the point estimate (mean across 125 MCMC samples) was outside the 95% distribution. For simulated data, the mean number of loci is reported from ten simulated datasets (random samples from the posterior distribution, see Fig. 3.12, Appendix 3). For each result, the proportion of the total number of SNPs is given in brackets.

alpha	**beta**	**observed data**				**simulated**			
		excess CI	**α outlier**	**β outlier**	**α and β outlier**	**excess CI**	**α outlier**	**β outlier**	**α and β outlier**
Negative	Negative	1147 (0.175)	97 (0.015)	43 (0.007)	5 (0.001)	1174.5 (0.157)	109.9 (0.015)	66.7 (0.009)	13.9 (0.002)
Negative	Not significant	837 (0.128)	14 (0.002)	0	0	700.8 (0.096)	17 (0.002)	0.3 (0.000)	0
Negative	Positive	863 (0.132)	5 (0.001)	58 (0.009)	0	1207.4 (0.161)	3.3 (0.000)	101.7 (0.014)	0
Not significant	Negative	167 (0.026)	0	4 (0.001)	0	364.7 (0.049)	0	8.2 (0.001)	0
Not significant	Not significant	262 (0.040)	0	0	0	408 (0.057)	0	0	0
Not significant	Positive	216 (0.033)	0	14 (0.002)	0	311.7 (0.042)	0	25.9 (0.004)	0
Positive	Negative	836 (0.128)	34 (0.005)	32 (0.005)	1 (0.000)	1142.1 (0.152)	93.6 (0.012)	26.7 (0.004)	3.2 (0.000)
Positive	Not significant	1129 (0.172)	89 (0.014)	0	0	1011.9 (0.138)	61.3 (0.008)	0.1 (0.000)	0
Positive	Positive	1089 (0.166)	41 (0.006)	92 (0.014)	0	1100.2 (0.148)	31.2 (0.004)	58 (0.008)	0.5 (0.000)

Chapter 4 Whole genome resequencing: bioinformatics pipeline

4.1 Introduction

Early work to identify genetic variation typically relied on a small number of molecular markers, including allozymes, microsatellites and mitochondrial DNA (Allendorf, 2017). Today, most studies apply single nucleotide polymorphisms (SNPs) as markers that can be sampled at high densities, genome-wide. Furthermore, the advent of high-throughput sequencing has resulted in a step-change in the number of markers and sequences available, of which there can be in the order of thousands to millions. High-throughput methods sampling markers or sequences across the genome ('genomic methods') generally fall into three categories: (1) marker-based genotyping (e.g., SNP arrays), (2) reduced representation sequencing (such as ddRAD-seq, see Chapter 2) or (3) whole-genome sequencing (Allendorf et al., 2010).

Marker-based genotyping and reduced representation sequencing (RRS) are considered cost-effective methods to sample across the genome, mitigating the historically high cost of whole-genome sequencing (Fuentes-Pardo & Ruzzante, 2017). Several SNP arrays have been designed for model or domestic species which can also be applied in related wild species. For example, the application of a 50K goat SNP chip to examine genetic diversity of Nubian ibex (Hassan et al., 2018) or a 600K chicken SNP chip used for gene ontology analysis across North American prairie grouse species (Minias, Dunn, Whittingham, Johnson, & Oyler-McCance, 2019). RRS methods are advantageous as they can be used even in the absence of other genomic resources (e.g., a reference genome) (Peterson et al., 2012), and have been applied to a wide range of non-model species, including, for example, snow leopards (Janjua et al., 2021), or Berthelot's pipits (Martin et al., 2021).

SNP arrays must be carefully designed to limit ascertainment bias, i.e., systematic bias in the allele frequency distribution (Helyar et al., 2011). Often, development using a limited number of individuals and/or individuals from a diverged species or population can lead to the omission of rare SNPs, biasing downstream analyses, including estimates of genetic diversity, population structure and demographic parameters. Ascertainment bias is significantly reduced in RRS. RRS methods generally use restriction enzymes to fragment genomic DNA, selecting a subset of fragments (e.g., by size) for sequencing (Scheben, Batley, & Edwards, 2017). However, these methods are not without limitations, and can be affected by allelic drop-out (i.e., polymorphism altering a restriction enzyme cutting site), PCR duplication bias, or variation in site coverage. Both methods are vulnerable to criticism of the density of markers, particularly in studies designed to detect loci under selection (Lowry et al., 2017).

Whole genome sequence data is a powerful resource with wide-ranging applications, including in medicine, evolutionary, synthetic and cell biology (Goodwin, McPherson, & McCombie,

2016; Luikart, England, Tallmon, Jordan, & Taberlet, 2003; Mardis, 2008; Purnick & Weiss, 2009). It has the potential to give the most comprehensive picture of genetic variation to date, including SNPs, insertions and deletions (indels), copy number variants (CNVs) and larger structural rearrangements (Ekblom & Wolf, 2014). Since the completion of the Human Genome Project (National Human Genome Research Institute, NIH, www.genome.gov/human-genome-project), the pioneering work to generate the first complete sequence of the human genome, in 2003, and the introduction of the first commercially available sequencer, in 2005, there has been rapid expansion in the available whole genome sequence data and tools (Reinert, Langmead, Weese, & Evers, 2015).

Traditional Sanger sequencing, primarily used by the Human Genome Project, has largely been replaced by massively high throughput 'next generation' sequencers, with 'third generation' long-read technologies becoming increasingly competitive (Goodwin et al., 2016). Sequencing cost has dropped 50,000 fold since the 2000s (Goodwin et al., 2016), outpacing Moore's law, that computing power doubles, and cost halves, every two years (Moore, 1965). As both timescale and cost of sequencing projects decrease, this approach has become feasible for non-model organisms, including species or populations of conservation concern (Fuentes-Pardo & Ruzzante, 2017).

Whole genome resequencing (WGR) refers to the sequencing of multiple individuals or populations to identify variation. WGR data was generated for the Scottish wildcat population with the aim of accurately dating the onset of admixture with domestic cats (Chapter 5). This chapter aims to give an overview of the application of genomic data in conservation (4.1.1), as well as the general workflow of a WGR project, including genome assembly and variant calling, and some common pitfalls and limitations (4.1.2). I then describe the pipeline established to process the Scottish wildcat WGR data and evaluate its effectiveness (4.2-4.4).

4.1.1 Conservation in the genomics era

There are several advantages of applying genomic data to conservation problems (Allendorf et al., 2010). Most obviously, the expansion of small sets of neutral markers traditionally used for population genetic analyses. Genomic methods can sample thousands or millions of markers, increasing power and reliability of parameter estimates, such as effective population size or relatedness (Allendorf et al., 2010). An increased number of genetic markers better resolves population structure, tree topologies and estimates of population or species divergence (Ellegren, 2014). Accounting for any ascertainment bias issues (see above), it can provide reliable inference of past demographic events, e.g., population growth, decline or migration (Allendorf et al., 2010; Luikart et al., 2003). An accurate understanding of population history, genetic diversity and demography supports conservation management decisions regarding, for example, conservation status, population viability or designation of management units (Allendorf et al., 2010). A recent landscape genomics study of the foothill yellow-legged frog (*Rana boylii*), for example, was used to identify putative

management units for this species on west coast of the USA (McCartney-Melstad, Gidiş, & Shaffer, 2018). Using genomic data, this study highlighted populations with low genetic diversity in need of more intensive conservation management, including genetic rescue via assisted migration.

There are also novel applications of genomic data in conservation biology (Allendorf et al., 2010). Whole genome sequences can identify structural variation and genomic incompatibilities between species for the first time. Genomic data can be screened for adaptive variation; identifying local adaptation supports designation of conservation management units and expands our understanding of the interaction between individuals (or populations) and their environment (Flanagan, Forester, Latch, Aitken, & Hoban, 2018). This feeds into prediction of population viability and adaptive potential in the face of environmental change. The genome of the Tasmanian devil, for example, has been sequenced and annotated for the first time in order to understand the evolution of transmissible cancer in this species, and the adaptive response to this selective pressure (Murchison et al., 2012).

Haplotype inference, using sequence data, can improve estimates of population structure and demography (Leitwein, Duranton, Rougemont, Gagnaire, & Bernatchez, 2020). Importantly, haplotype information is a valuable tool to detect introgression and understand hybridisation dynamics. This includes identification of anthropogenic hybridisation and prediction of the impacts of hybridisation on fitness.

A third important application is to understanding the mechanism of inbreeding depression (Kardos, Taylor, Ellegren, Luikart, & Allendorf, 2016). Inbreeding depression is often a primary concern for the management of small, fragmented or isolated populations (wild and captive). An understanding of the molecular drivers, e.g., the number and/or effect of loci involved, would help prevent or reverse the impacts. It would also better support management or reintroduction programs to avoid founder-specific inbreeding depression, allowing screening of potential founders for deleterious recessive alleles.

The field of metagenomics, sequencing genetic material from environmental samples, has potentially interesting applications in conservation biology (Trevelline, Fontaine, Hartup, & Kohl, 2019). For example, sequencing microbial communities as a means to assess ecosystem function (DeLong, 2009) or disease status of individuals (Vega Thurber et al., 2008). Many other novel aspects, e.g., exploring epigenetics or environmental DNA sampling, are still at an exploratory stage (Shafer et al., 2015).

In practice, these applications, though promising, may not be easy to translate into everyday conservation management. High performance computing resources and bioinformatic expertise are needed for most analyses (Shafer et al., 2015). Genomic resources are currently available for a limited number of species, predominantly model species or domesticates, and existing tools generally

require high quality reference data (da Fonseca et al., 2016). The cost-benefit of whole genome resequencing should be considered on a case-by-case basis; large amounts of sequence data are unnecessary to answer most conservation questions (Shafer et al., 2015). It important to consider the biological questions that need addressing, and the available resources to answer them, before designing a whole genome resequencing project.

4.1.2 Genome assembly and variant calling

Most sequencing projects aim to identify variation within or between genomes. The general pipeline for variant discovery is shown in Fig. 4.1, which illustrates the WGR workflow from DNA extraction to analysis-ready variants.

The first step, once DNA of sufficiently high molecular weight has been sampled and extracted, is to determine its nucleic acid sequence (Fuentes-Pardo & Ruzzante, 2017). Even the smallest genomes cannot be sequenced in a single next generation sequencing (NGS) run, so DNA is first broken up into smaller fragments, either mechanically, by sonication, or using enzymes (Goodwin et al., 2016). Multiple sequencing technologies have been developed, but the general principle is to use a single stranded DNA template to build the complementary sequence using known bases, the order of which can be recorded. This was initially achieved by Sanger sequencing, where DNA polymerase synthesises the complementary strand, incorporating normal deoxynucleotides (dNTPs) and modified, fluorophore or radioactively labelled nucleotides (dideoxynucleotide triphosphates, ddNTPs), which terminate chain elongation (Sanger, Nicklen, & Coulson, 1977). Labelled ddNTPs are incorporated at random, at multiple positions across different copies of the template sequence. Synthesised sequences are separated by size using gel electrophoresis, and in this way, the order of bases in the sequence is determined.

Sanger sequencing is highly accurate (~99.99%) and capable of generating relatively long reads (up to 1kb), but is also slow, low throughput, and therefore expensive (Fuentes-Pardo & Ruzzante, 2017). The main advantage of NGS platforms, such as Roche 454, Illumina or BGISEQ, is high throughput. NGS uses multiple reaction centres, each with thousands of copies of template DNA, to parallelise sequencing (Goodwin et al., 2016). Some platforms, e.g., Illumina, employ a sequencing by synthesis approach similar to that of Sanger sequencing, where DNA polymerase incorporates chain-terminating nucleotides. Unlike Sanger sequencing, the process is cyclical: signal from the incorporated nucleotide is recorded and the 3' blocking group removed for the next cycle. This removes the requirement for gel electrophoresis, allowing sequencers to become increasingly compact (Reinert et al., 2015). Other platforms, e.g., BGISEQ, use sequencing by ligation, where labelled probes hybridise to the template and the molecule is imaged (Goodwin et al., 2016).

Cleavage of the fluorophore label and/or probes containing additional degenerative or universal nucleotides can be used to interrogate different template positions in a cyclical fashion.

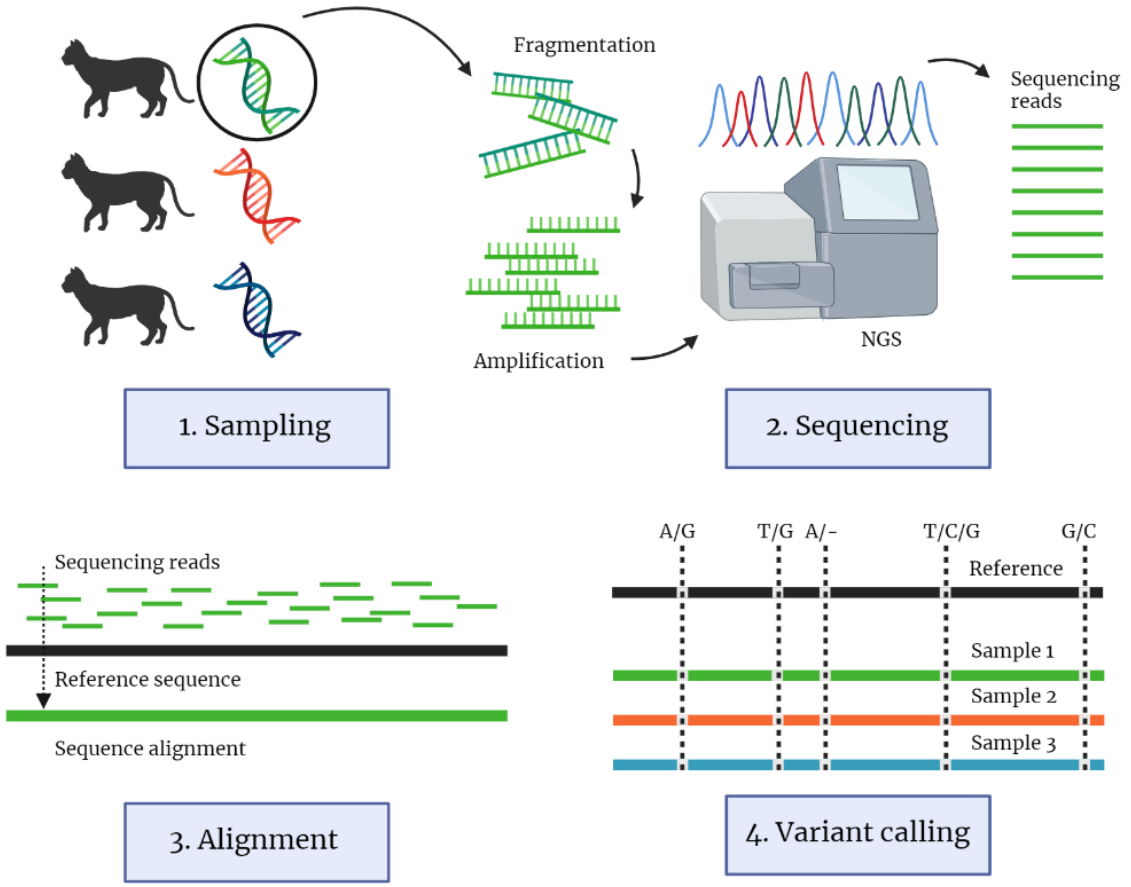

Figure 4.1. Overview of the variant calling pipeline for population genomics. (1) DNA is sampled from a population (or populations) of interest. (2) Samples are sequenced, generally using high-throughput NGS platforms generating short read data. (3) Read data can be aligned to a reference genome, reconstructing the genome sequence of each individual in the sample. (4) Variation within the population can then be identified.

The major drawback of NGS platforms is short read length (usually between 50 and 700bp) (Goodwin et al., 2016). Each platform also has a specific error model. Mistakes in DNA polymerase replication, for example, result in high rates of substitution error during sequencing by synthesis (Illumina accuracy is ~99.5%). Sequencing by ligation is more accurate as each position is probed multiple times, but AT rich regions may be under-represented using these methods. For platforms capable of incorporating homopolymers (runs of more than one identical base) intensity of signal is used to infer homopolymer length (Reinert et al., 2015). Poor signal resolution or background noise can lead to error, especially as homopolymer length increases. 'De-phasing' is another common source of error, where the sequencing of identical templates within a reaction centre gets out of sync. This effect is exaggerated after multiple cycles and can therefore lead to a high error rate at read ends. To account for miscalls and sequencing errors, base quality scores are provided to estimate confidence

of each base call (see Box 1) and this information is taken into account by downstream processes, such as alignment.

Sequencing reads are then assembled into the genome sequence of an individual. Reconstructing this sequence from short reads is complex, especially for large or repetitive genomes, and several different algorithms and tools have been developed to improve the speed and accuracy of genome alignment (Reinert et al., 2015). *De novo* assembly assumes no prior knowledge of the genome under investigation (Ng & Kirkness, 2010). Overlapping reads are aligned to form contigs (longer contiguous sequences), and the subsequent joining of contigs forms scaffolds (where contigs are ordered and orientated correctly). *De novo* assembly is computationally intensive and requires high sequencing depth (Fuentes-Pardo & Ruzzante, 2017). Alternatively, reads can be mapped to an existing reference sequence (Reinert et al., 2015). Instead of aligning reads to each other, the aim is to find the origin of each read with respect to the reference. The reference sequence must be selected carefully; ideally, there should be minimal divergence between the reference and target populations (Fuentes-Pardo & Ruzzante, 2017). Otherwise, there is a bias towards reference alleles (Reinert et al., 2015) and any novel sequences, not present in the reference, will be excluded from the alignment (leaving a proportion of reads unmapped) (Fuentes-Pardo & Ruzzante, 2017).

Alignment algorithms have been designed to solve the approximate matching problem, i.e., to find the best alignment for a read, allowing gaps or mismatches. They must search the reference sequence efficiently for the best alignment. For this, most methods use either filtering, where regions with no match to a short sequence (or seed) within the read are excluded, or indexing, where a sequence index (of the reads, reference, or both) is generated prior to alignment, avoiding scanning the entire dataset or reference sequence for each query.

It is important to note that there is a distinction between the sequence that can be obtained for an individual, through sequencing and assembly, and its actual genome sequence (Ellegren, 2014). Many regions of the genome are hard to reconstruct due to their repetitive nature, e.g., heterochromatin at the centromeres and telomeres, which have only recently been characterised in well studied species, such as humans (Miga et al., 2020). Approximately 50% of the human genome consists of repeat sequences and, though poorly understood, many have functional importance, e.g., in epigenetics, or during genome expansion or speciation (Reinert et al., 2015).

Repetitive regions lead to errors during alignment. For exact repeats it is impossible to identify the true origin of a read. Even for inexact repeats, high coverage and stringent alignment criteria are needed to resolve these regions (Miller, Koren, & Sutton, 2010). Alignment is generally improved using longer reads (Reinert et al., 2015), the solution of NGS (short-read) platforms is paired-end data. Paired-end reads are generated by sequencing template DNA from both the 5’ and 3’

ends, producing forward and reverse reads. Forward/reverse reads are found within a known distance apart and can overlap, improving alignment.

Third generation platforms have been developed with aim of generating long-read data, including *in silico* synthetic methods, PacBio's SMRT-seq (single molecule real-time sequencing) (read lengths of 8-20Kb) or Oxford Nanopore MinION (200Kb) (Goodwin et al., 2016). Though advantageous for resolving structural variation or repetitive regions, error rates are much higher than Sanger sequencing or NGS platforms, and cost remains high (Fuentes-Pardo & Ruzzante, 2017). It is advisable to combine data from multiple platforms, using long and short reads to resolve alignment issues and avoid systemic sequencing biases (see Rhie et al., 2021), though, in practice, the associated cost may not be feasible for many projects (Goodwin et al., 2016).

Alignment score and mapping quality are used to evaluate alignment success (Reinert et al., 2015). Alignment score is essentially a measure of how closely the read sequence matches the reference. Mismatches or gaps can be the result of genuine polymorphism, or alignment or sequencing error; there is an important trade-off between tolerating mismatches to find true variation and tolerating errors, leading to a high false positive rate.

Mapping quality (MAPQ) quantifies mapping confidence, similarly to base calling quality score (Box 1):

$$\mathrm{MAPQ} = -10 \log_{10} P$$

where P is the probability of incorrect mapping. E.g., for an alignment with a MAPQ score of 10, there is a one in ten chance the read originated from a different place in the genome. Alignment score and mapping quality can be related through 'uniqueness'. If a read aligns to multiple places in the reference sequence equally well, the highest scoring alignment is not unique, and should therefore have a lower MAPQ score.

Once aligned, sequences are ready for genotype and SNP calling (indels, copy number and structural variation will not be discussed further here). Early genotyping methods used a single sample, allele counting approach, retaining only high-quality variants with a non-reference allele frequency within an acceptable range (Nielsen, Paul, Albrechtsen, & Song, 2011). This works well for high coverage data (>20X), but otherwise leads to under-calling of heterozygotes. Importantly, it does not give a measure of confidence in the called genotypes.

A probabilistic framework was introduced using Bayes' theorem to compute genotype likelihoods given the read data at a site (Nielsen et al., 2011). The genotype with the highest posterior probability is called and the posterior probability used as a measure of confidence in the call. Other information can be incorporated into the prior, such as expected allele frequencies, or linkage disequilibrium information, to improve posterior estimates. Genotype likelihoods can also be improved with a larger sample size or by recalibrating quality scores using empirical data.

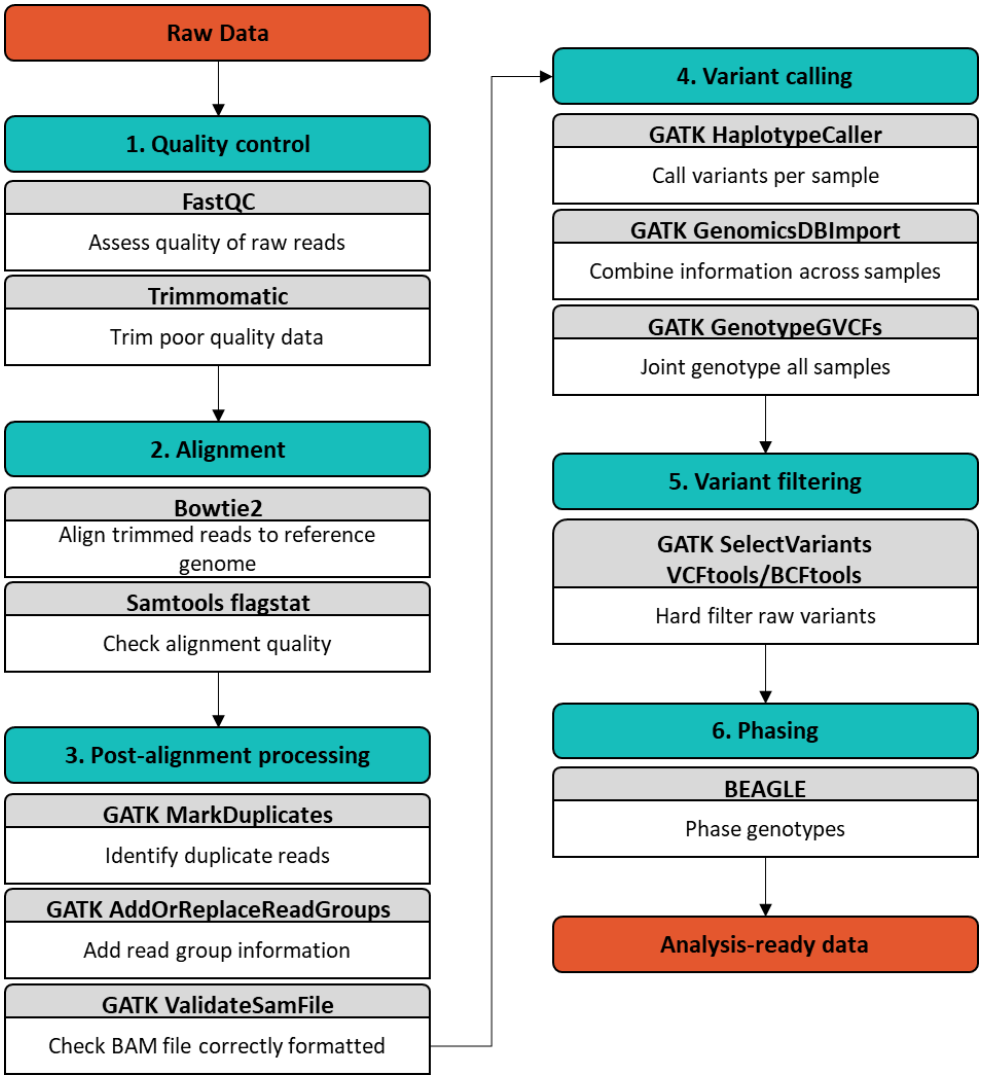

Figure 4.2. Tools used at each step of the variant-calling pipeline developed for Scottish wildcats.

Accumulated errors from sequencing and alignment can result in variant calling error (Nielsen et al., 2011). Genotype-likelihood methods assume reads are independent from each other, an assumption which can be violated by alignment errors or PCR artefacts (generated during template amplification). Variant calls are therefore improved by subsequent filtering. Common filters include low confidence calls, extreme reads depths, and strand or positional bias. Filtering is another important trade-off between retaining large numbers of false positives and losing too many true

positives. It is important to carefully consider potential sources of error and tailor the filtering approach to the analysis.

The workflow described in the following sections (and shown in Fig. 4.2), was guided by the Genome Analysis Toolkit (GATK) best practice for variant discovery (https://gatk.broadinstitute.org/hc/en-us) (DePristo et al., 2011, Van der Auwera & O'Connor, 2020). GATK provides a suite of tools to handle genomic data, originally designed to standardise analysis of human genomes, but now expanded for all organisms. Tools used at each stage are shown in Fig. 4.2. GATK best practice results in a set of high-quality variants ready for downstream processing, such as phasing, and population genetic analyses.

4.1.3 Aims

For questions relating to adaptive variation and genome or haplotype structure, whole genome sequence data are a powerful resource previously unavailable for conservation research. Whole genome resequencing data provides essential information to understand hybridisation dynamics in the Scottish wildcat population, where haplotype information is needed to fully resolve the history of hybridisation in Scotland. This chapter describes the sampling approach, sequencing methodology and bioinformatic pipeline developed to process wildcat genomic data, with the aim of generating high-confidence variants ready for downstream analyses.

4.2 Methods

4.2.1 Sampling

Forty-five individuals from Scotland were selected for WGR: five Scottish domestic cats, ten captive wildcats and 30 wild-living individuals. This included a subset of 35 from the ddRAD-seq dataset (Chapter 2). Wild individuals were selected with the aim of sampling a wide geographic distribution (Fig. 4.3A) and range of genetic backgrounds from across the hybrid swarm (assessed using Q35 score and PC1 position, Fig. 4.3B). Samples were collected between 1997 and 2018, and processed and stored by the WildGenes laboratory, RZSS.

Wildcat and domestic cat data from outside of Scotland were available from several other sources. This included raw read data for six German wildcat samples, provided by Carsten Nowak and Violeta Muñoz-Fuentes, and mapped read data from one Portuguese wildcat, provided by Carlos Driscoll. Public databases contained read data for a further 17 domestic cats with a global distribution (for detailed sample information see Table 4.6, Appendix 5). The final dataset is summarised in Table 4.1.

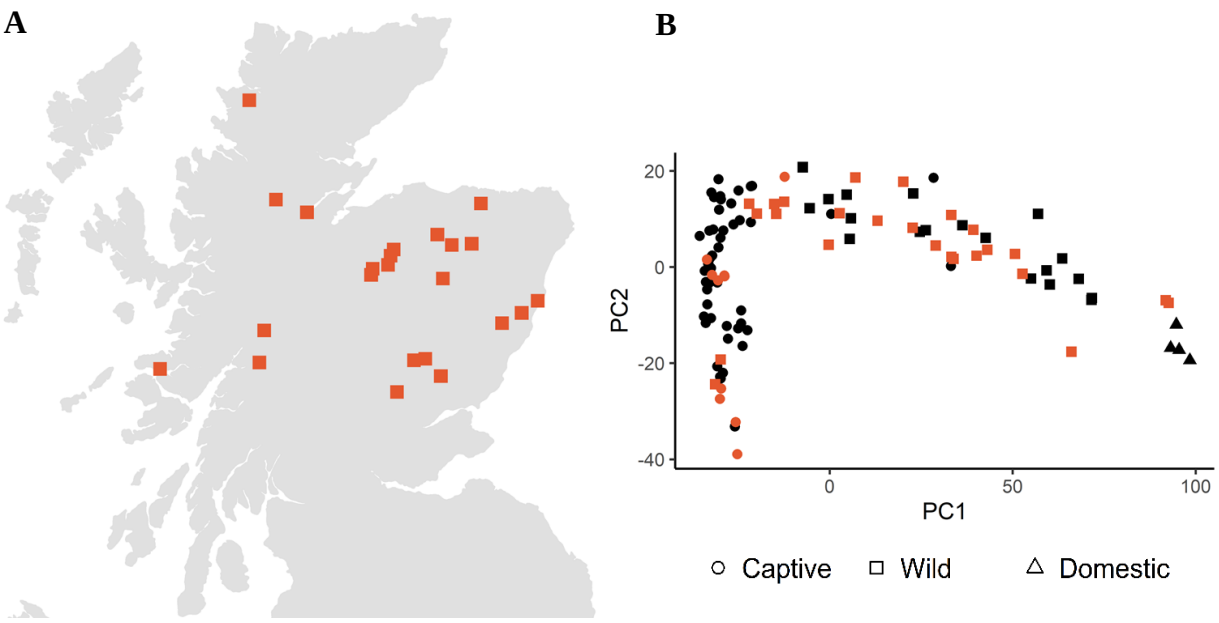

Figure 4.3. Sampling approach for whole genome resequencing. (A) Sampling locations, where known, for wild-living individuals. (B) PCA, as per Chapter 2, using the ddRAD-seq dataset (108 individuals genotyped at 6,546 SNPs). A subset of 35 individuals were selected for WGR (highlighted in red), including 10 captive and 25 wild-living cats. Wild individuals were sampled from across the hybrid swarm.

Table 4.1. Summary of samples and source populations included in whole-genome sequence analysis.

Population	**Number of samples**	**Sampling location(s)**
Scottish captive	10	UK
Scottish wild	30	Scotland
Scottish domestic	5	Scotland
Continental wildcat	7	Germany, Portugal
Other domestic	17	Portugal, Denmark, Italy, USA, Oman, Iraq, Jordan, China, South Korea, Thailand, Madagascar
Total	69	

4.2.2 Sequencing

DNA samples from Scotland were sequenced by the Beijing Genomics Institute (BGI) using the BGISEQ-500 platform. BGISEQ uses a sequencing by ligation approach (Goodwin et al., 2016). Adaptors are ligated and double stranded template DNA is circularised and amplified via rolling circle amplification. This generates long, single-stranded DNA molecules containing multiple copies of the template sequence, which are compacted into nanoballs and distributed onto a flow cell; each nanoball occupies a discrete location. Sequencing by ligation can then take place, using probes containing a known base in the n+1 position for each subsequent round of imaging. Sequencing can occur up or downstream of the bound probe.

BGISEQ generated paired-end read data in FASTQ format (Box 1). Medium sequence coverage was obtained, ~15X (i.e., on average, 15 reads per sample covered each base position). Adaptor sequences and low-quality reads were removed as part of BGI's quality control.

Data collected from outside Scotland was sequenced at variable depths (up to 30X), predominantly using Illumina platforms.

4.2.3 Quality control

FastQC (Andrews, 2010) was used to perform quality checks on all raw read data (68 samples). FastQC was run twice per sample, processing forward and reverse reads separately. For a description of FastQC analyses see Table 4.2.

Table 4.2. FastQC modules. FastQC employs a traffic light system to indicate a pass, warning or fail for each module as a means to assess raw read quality.

Analysis Module	**Description**	**Warning/fail**
Basic Statistics	Basic summary of the data, including the number of reads, read length, etc.	NA
Per Base Sequence Quality	Range of quality scores at each base positions across all reads	Lower quartile for any position Warning: <10 Fail: <5 Or median at any position Warning: <25 Fail: <20
Per Sequence Quality Scores	Mean quality per read	Most frequent score across dataset Warning: <27 Fail: <20
Per Base Sequence Content	Proportion of each of the four bases at every position	Difference between A and T or G and C at any position Warning: >10% Fail: >20%
Per Sequence GC Content	Distribution of GC content per read across the dataset (should be approximately normal)	Deviation from normal distribution Warning: >15% of reads Fail: > 30%
Per Base N Content	Proportion of missing information at each position	Warning: >5% Fail: >20%
Sequence Length Distribution	Distribution of sequence lengths across the dataset	Warning: not all sequences are the same length Fail: any sequences have length 0
Duplicate Sequences	Amount of duplication of each sequence examined. Gives an approximation of duplication levels; examines the first 50bp of 100,000 sequences	Proportion of sequences that are duplicated Warning: >20% Fail: >50 %
Overrepresented Sequences	Lists all sequences that make up more than 0.1% of the total (same approximation as Duplicate Sequences)	Warning: >0.1% of total Fail: >1% of total
Adaptor Content	Fraction of reads where Illumina adapter sequences are identified	Warning: >5% Fail: >10%

Trimmomatic (Bolger, Lohse, & Usadel, 2014) was used to filter poor quality read data. A sliding window approach removed bases from the 3' end of any sequence of four positions with a mean quality score less than 20. Bases at the leading or trailing ends of reads with a quality score less than three were also trimmed. Any reads shorter than 36 bases were discarded. Removal of poor quality read data is an important step to minimise sequencing error before alignment.

4.2.4 Alignment and processing

Bowtie 2 (Langmead & Salzberg, 2012) was used to align trimmed read data to the domestic cat reference genome (felCat9, GCF_000181335.3) (Buckley et al., 2020); there is currently no reference sequence for *F. silvestris.* A reference genome for domestic cats has been available since 2006 (Pontius et al., 2007), and has since been improved by long-read sequencing and optical mapping (Buckley et al., 2020). The current assembly contains 4,909 contigs with an N50 of 42Mb, comparable to that of the human genome (Zhang & Schoenebeck, 2020). The current (ungapped) length is 2.48Gb, including 18 autosomes and the X chromosome (Buckley et al., 2020).

Bowtie 2 is optimised to align short read data to relatively large, mammalian, genomes. It uses an indexing approach which is fast and memory efficient (Langmead & Salzberg, 2012). During alignment, mismatches, gaps, or missing data are penalised, especially at sites with high quality base calls. A perfect alignment (where the read and reference sequences match exactly) receives a score of zero, alignments with gaps, mismatches, etc., receive a negative score. In this way multiple alignments can be tested per read, and the most likely origin identified. End-to-end alignment was used here, i.e., involving all base positions, as opposed to local alignment, where the ends of a read can be ignored to maximise the alignment score. For paired data, Bowtie 2 also takes the relative orientation and distance between mate pairs into account. A pair aligns 'concordantly' if reads are in the expected orientation and within the expected distance apart. Discordant alignment can occur where both reads have unique alignments (see 4.1.2), but do not meet paired-end expectations. This can be an indicator of structural variation. If a paired-end alignment could not be found, pairs were aligned separately (Bowtie 2 'mixed mode').

Alignment information was outputted in SAM format (Box 1). SAMtools (Li et al., 2009) was used to sort and compress SAM files to BAM format, and index the resulting BAM files. Further processing was required to be compatible with the GATK toolkit (DePristo et al., 2011; McKenna et al., 2010), this included:

1. *Addition of read groups.* Read groups give technical information about the sequencing process, e.g., sequencing lane and platform. This information was added using GATK AddOrReplaceReadGroups

2. *Identification of duplicates.* Marking duplicates is the process of identifying reads that originate from duplicated DNA templates. Duplication inflates coverage at a specific region, which is especially problematic if propagating sequencing or copying error. Reads are tagged using GATK MarkDuplicates, to take duplication into account during variant calling. There are two types of duplication:

   - PCR duplicates: PCR duplicates occur when copies of a single template are found at more than one location on the sequencing flowcell.
   - Optical duplicates: Optical duplicates are copies of one template found at one flowcell location falsely called as two locations.

GATK ValidateSamFile was used to check the BAM files for errors, e.g., incorrect formatting or faulty alignments, before proceeding to variant calling.

Box 1. Common file formats

FASTQ

Text files containing nucleotide sequence(s) and quality score per base. There are generally four lines:

- Sequence ID
- Nucleotide sequence
- '+' and sequence ID/description (optional)
- Phred-scaled quality scores

Quality scores are coded by ASCII characters. Sequence identifiers often contain useful information about the sequencing platform, e.g., instrument name, lane, etc.

Phred-scaled quality score (Q) is calculated by:

$$\mathrm{Q} = -10 \log_{10} P$$

Where P is the estimated probability an incorrect base call, e.g.,

Q	P	Accuracy
10	0.1	90%
20	0.01	99%
30	0.001	99.9%

SAM/BAM

SAM and BAM files store sequence alignment data, BAM files are compressed binary versions of SAM files. There are two sections:

- Header (optional) containing general information about the file
- Alignment information per sequence

Box 1. (cont.)

The alignment section has 11 mandatory fields:

Field	Description
QNAME	Sequence ID
FLAG	Sequence/alignment information coded using bitwise flags (see below)
RNAME	Reference ID
POS	Start position
MAPQ	Mapping quality score
CIGAR	CIGAR (Concise Idiosyncratic Gapped Alignment Report) string. Codes for the positions of mismatches, insertions or deletions with respect to the reference
RNEXT	Reference name of mate/next read
PNEXT	Position of mate/next read
TLEN	Sequence length
SEQ	Nucleotide sequence
QUAL	Base quality (ASCII encoded Q score)

The FLAG field is reported as the sum of any applicable flags, each giving different information about a sequence alignment.

	FLAG (as interpreted for paired data)
1	Read is paired
2	Read is properly paired (i.e., both reads in a pair have been mapped onto the same chromosome, within the expected distance apart)
4	Read is unmapped
8	Mate is unmapped
16	Reverse complement
32	Mate is the reverse complement
64	First in pair
128	Second in pair
256	Secondary alignment (i.e., for reads that could align well in multiple places)
512	Failed quality checks
1024	PCR or optical duplicate
2048	Supplementary alignment (i.e., part of a chimeric alignment, where parts of the read align in different places)

VCF

This is the standard format for storing SNP and indel information. Again, there are two main sections:

- Header containing information about the file/dataset, including definitions of any annotations used to quantify variant calls
- Records with one line per variant

The records section contains the following information per variant:

Field	Description
CHROM	Reference sequence name (usually chromosome name)
POS	Variant position on reference
ID	Variant ID
REF	Reference base
ALT	Alternate allele(s)
QUAL	Quality score
FILTER	Pass/fail for given set of filters
INFO	Site level annotations, e.g., frequency of alternate allele or depth across all samples
FORMAT	Sample level annotations, e.g., genotype, genotype likelihood, read depth
SAMPLE(s)	Information per sample for all fields given in FORMAT

4.2.5 Genotyping

Genotypes and variable sites were called using the GATK toolkit. First, HaplotypeCaller identified variants per individual, outputting a gVCF (or genomic VCF, for information about VCF files see Box 1) containing genotype information at all sites. At variable sites HaplotypeCaller calculates genotype likelihoods, and the most likely genotype is reported. HaplotypeCaller was run for all individuals (n=69), including the Portuguese wildcat sample for which alignment data was available (in BAM format).

GenomicsDBImport aggregated gVCFs for joint genotyping. This is computationally intensive, so was performed separately across shorter genomic regions (or intervals), in this case per chromosome. Processing a subset (or batch) of gVCFs at a time reduces memory consumption (but increases running time); a batch size of 15 samples was used for this analysis. Joint genotyping of all samples was then performed per interval using GenotypeGVCFs.

4.2.6 Filtering

Variant filtering improves calls where genotype likelihoods have not been calculated accurately and/or not all error information considered (Nielsen et al., 2011). Two rounds of filtering were carried out on the wildcat dataset. Firstly, using GATK SelectVariants and the filters and thresholds shown in Table 4.3. Read depth, variant quality and SNP density were then assessed per chromosome (Fig. 4.9, Appendix 6) to inform a second round of filtering using VCFtools (Danecek et al., 2011) and BCFtools (Li et al., 2009), to remove sites with low-quality calls (QUAL < 50) or excessive read depth (DP > 2000). Closely related individuals in the captive population were removed from the dataset (as identified using the molecular studbook described in 2.1.2). Missingness per source population (domestic, putative wildcat or hybrid) was checked (Fig. 4.10, Appendix 6) and sites with missing data discarded. The putative wildcat group included samples from continental European wildcat populations and the captive Scottish wildcats (as the most genetically distant to domestic cats in Scotland, see Chapter 2). Finally, a minor allele count of three was imposed, ensuring all variant sites were called in at least two individuals.

4.2.7 Phasing

Beagle v.5.2 (Browning & Browning, 2007) was used to phase the data, i.e., determine parental haplotypes for all samples. Beagle uses a Hidden Markov Model (HMM) to infer haplotypes using linkage disequilibrium information. A genetic linkage map has been generated for domestic cats (Li et al., 2016). This was modified to phase the wildcat dataset. The original marker set was pruned to remove non-contiguous SNPs and a minimum recombination rate of 5×10^{-7} was imposed. The final

map used for phasing included 5,860 markers. Beagle was run using 100 iterations and a burnin of 50.

Table 4.3. Filters and thresholds used for first round of variant filtering with GATK SelectVariants

Filter	Description	Discard
QD	Variant quality accounting for the read depth of the alternate allele	< 2
FS	Phred-scaled probability of strand bias	> 60
SOR	A second estimate of strand bias taking into account the ratio of reads covering both alleles	> 3
MQ	Root mean square of mapping quality for all reads covering a variant site	< 40
MQRankSum	Rank sum test of mapping qualities, comparing forward and reverse reads supporting the reference versus alternate allele	< -12.5
ReadPosRankSumTest	Rank sum test for variant position within reads, comparing position within forward and reverse reads supporting the reference versus alternate allele	< -8

4.3 Results

4.3.1 Data quality and alignment

The FastQC results can be seen in Fig. 4.4. For two modules, Per Base Sequence Content and Per Sequence GC content, most samples received a warning, and some failed completely. Warnings from both these modules can indicate the presence of overrepresented sequences, for example adaptor sequences, or contamination. Visual inspection of the GC distribution per sample (Fig. 4.11, Appendix 6), showed skewed distributions, but no sharp peaks, which demonstrate the presence of a specific contaminant.

Following the removal of poor-quality base calls and truncated sequences by Trimmomatic, the mean proportion of read pairs retained per sample was 89.3%. For 8.5% of read pairs, on average, only one mate was retained for downstream analyses (Fig. 4.5). FastQC checks highlighted some samples with a decrease in base quality at read ends (Fig. 4.11, Appendix 6), which was improved by trimming.

All samples aligned well to the domestic cat reference sequence (Fig. 4.6), with an overall mean alignment rate of 97.4% (Table 4.4). Alignment rate did not seem to vary between source populations, though a larger variation in alignment rate was observed in domestic cats and putative wildcats. This is likely to reflect the mix of sample sources and sequencing approaches within these groups.

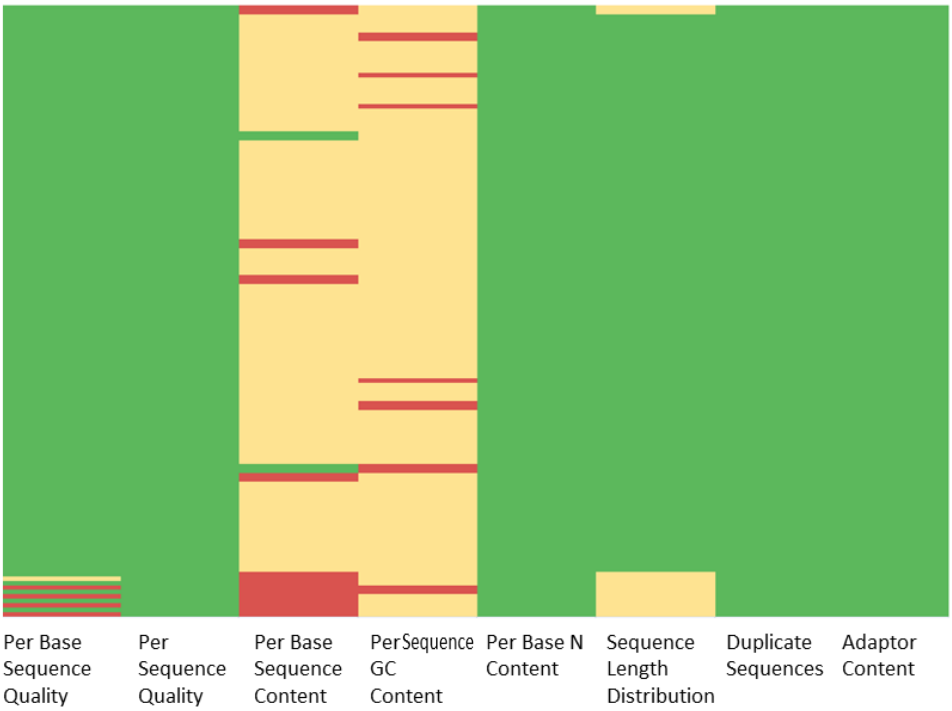

Figure 4.4. FastQC traffic light system to flag warnings (yellow) or failures (red) of quality checks on raw data (for description see Table 4.2). Forward and reverse reads were processed separately for each sample. Overall, the quality of the raw data appeared high, with a low rate of missing data, duplication and no strong signals of contamination.

Table 4.4. Mean alignment rate across all individuals and for each source population. 'Wildcat' samples included the captive Scottish wildcats (n=10) and German wildcat samples (n=6), 'hybrid' samples included all individuals sampled from the wild in Scotland. The Portuguese wildcat sample was excluded from this as the data were not aligned by the author of this study (data supplied as BAM file).

Population	**Number of samples**	**% Alignment**			
		Mean	**Var**	**Max**	**Min**
Domestic cat	22	96.11	33.38	99	79.3
Putative wildcat	16	97.57	18.27	99	81.6
Hybrid	30	98.32	1.05	99	95.6
All	68	97.43	15.94	99	79.3

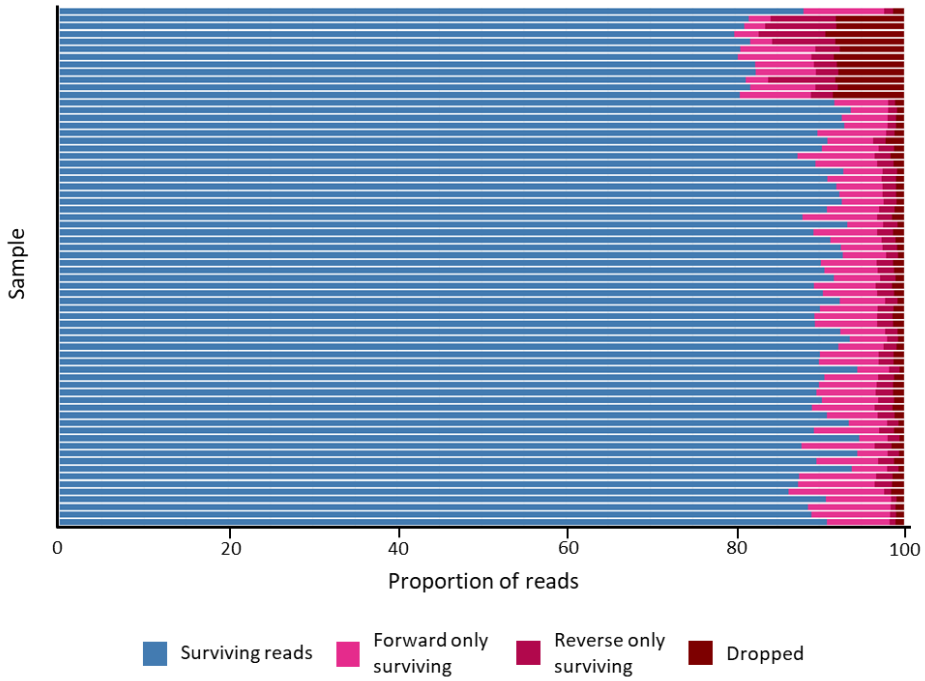

Figure 4.5. Each horizontal bar shows the proportion of reads retained or dropped by Trimmomatic per sample. Overall, raw read data was of a high quality and only a small proportion of reads were dropped per sample. Many of the German wildcat samples failed the base quality module of FastQC (see Fig. 4.11. Appendix 6), and a higher proportion of reads were dropped for these samples.

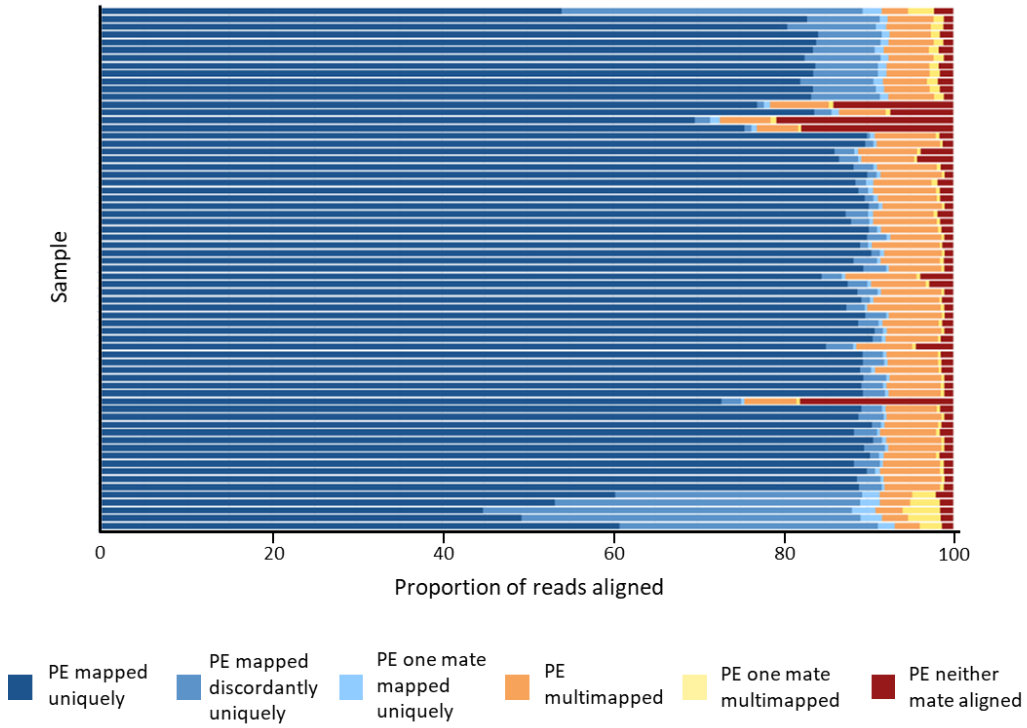

Figure 4.6. Each horizontal bar shows the proportion of reads aligned to the domestic cat genome per sample. Colour-coding is based on whether alignment was unique or multimapping (i.e., aligned to one or multiple places in the reference sequence), whether it met paired-end expectations (concordant/discordant paired-end alignment), and whether both mates in a pair were mapped together.

4.3.2 Genotyping and filtering

34,471,462 SNPs were initially called following joint genotyping of 69 samples. During the first round of filtering 8,869,640 SNPs were discarded, and a further 13,737,929 SNPs were discarded during the second round (Fig. 4.7). Variant filtering successfully retained only high confidence variants, at sites genotyped across all samples (Fig. 4.8). The final number of SNPs for analyses was 11,863,892.

Four captive wildcat samples were removed from the analysis to limit relatedness in the dataset (see Table 4.6, Appendix 5).

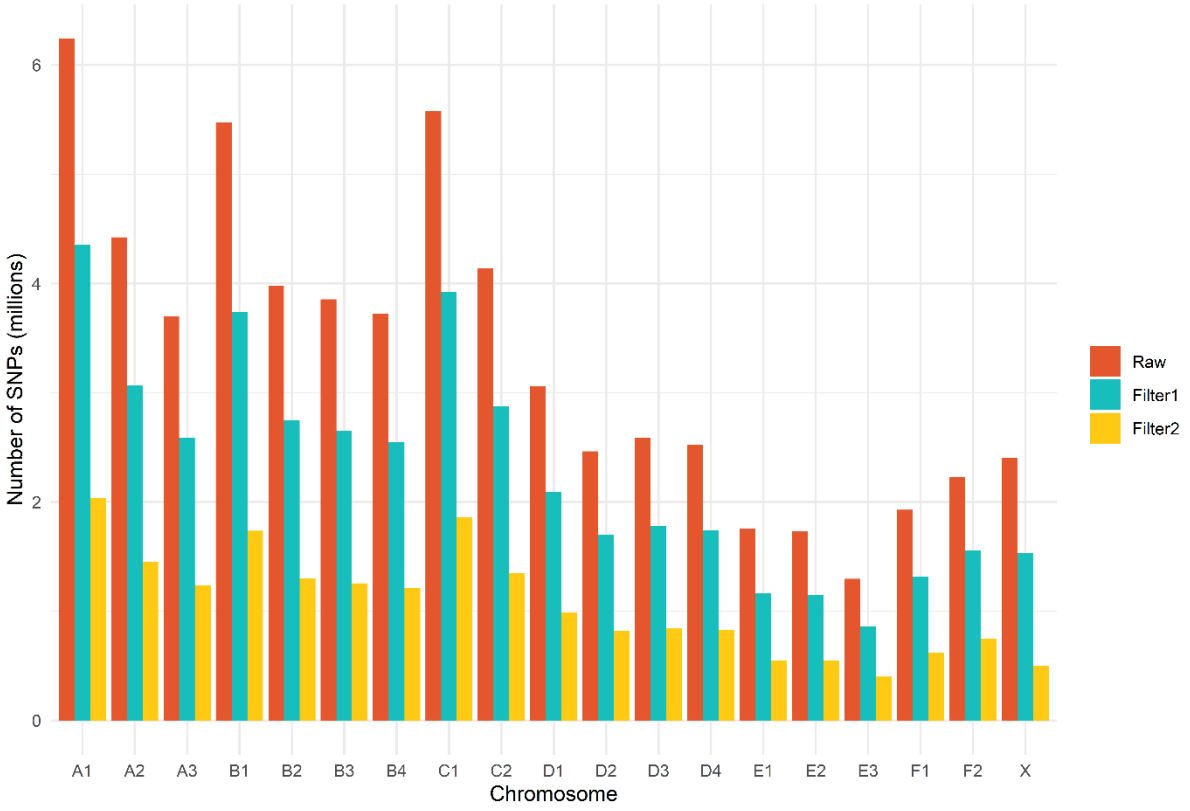

Figure 4.7. Number of SNPs per chromosome following joint-genotyping (red), initial filtering with GATK SelectVariants (blue) and a final round of filtering (yellow), based on call quality, read depth and site missingness.

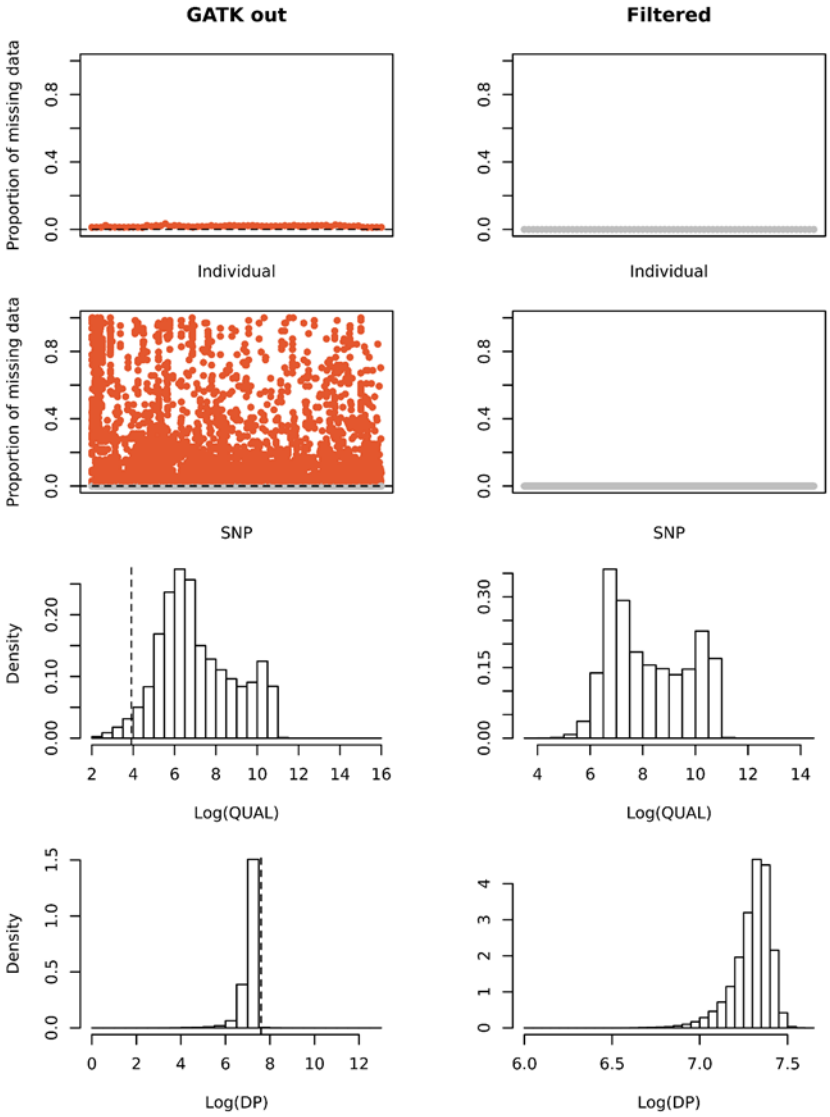

Figure 4.8. During the second round of variant filtering missing-ness was evaluated per sample and per site, as well as variant quality (QUAL) and read depth (DP). Above, these metrics are compared between the 'raw variants' following GATK joint-genotyping ('GATK out', left) and the final set of filtered variants ('Filtered', right). Dashed lines indicate the chosen thresholds.

4.3.3 Phasing

The final dataset used for phasing consisted of 65 individuals, genotyped at 11,863,892 SNPs. Without phased reference wildcat data, it was difficult to evaluate phasing accuracy. The VCFtools module '--diff-switch-error' compares two VCF files containing the same individuals and reports the number of phasing switches (i.e., where maternal and paternal haplotypes are reversed). This was used to compare phasing using Beagle's default recombination rate and/or fewer phasing iterations to the final dataset, phased using the modified domestic cat recombination map (Fig. 4.12, Appendix 6) and 100 iterations.

The number of iterations did not appear to improve the switch error rate (Table 4.5). Using the default recombination rate, a higher switch error rate (~4%) was reported, indicating information from the domestic cat recombination map was successfully used to phase the wildcat data, however, this still does not guarantee phasing accuracy.

Table 4.5. Switch error rate for four test scenarios when compared to the final dataset phased with the modified domestic cat recombination map and 100 iterations. One scenario used Beagle's default recombination rate, three used the domestic cat recombination map and increasing numbers of iterations (default=12). Switch error rate was greatest using the default recombination rate. Increasing the number of iterations does not appear to improve phasing error.

Recombination rate	**Number of iterations**		
	12	24	50
Default	0.038	-	-
Domestic cat map	0.019	0.021	0.020

4.4 Discussion

Overall, an effective pipeline was established to generate phased SNP data from WGR reads. Despite variability in sequencing platform, read depth, and quality between samples, the pipeline appeared to be robust, generating consistent output across the sample set.

The history of conservation management and monitoring of Scottish wildcats, both in captivity, and through TNVR and roadkill sampling in the wild, provided an indispensable databank of high-quality DNA samples. Obtaining DNA of sufficient quality for whole genome sequencing is often challenging for conservation genomics research, especially for wild populations of elusive species, like the Scottish wildcat, that are often monitored through non-invasive sampling (Piggott & Taylor, 2003). Samples were selected with the aim of representing a wide range of genetic backgrounds and geographic locations, from individuals sampled over the last 25 years.

BGISEG-500 was used for whole-genome sequencing. BGISEQ is a relatively new platform, available since 2016. It is comparable to Illumina platforms in terms of sequencing accuracy and read quality (Huang et al., 2017; Mak et al., 2017). Unique to BGISEQ is the application of nanoball technology (Drmanac et al., 2010). Nanoballs are a compact molecule containing multiple copies of the template DNA fragment. This provides a strong signal for sequencing, and the naturally negatively charged nanoballs repel each other, maintaining a discrete location on the flow cell and eliminating optical duplicates. Sequencing by synthesis is highly accurate, calling few false variants, but results in a loss of sensitivity (some true positives are missed) (Goodwin et al., 2016). Other disadvantages are short read length (150bp) and potential for PCR bias, but these are common to all

NGS platforms. Generally, the quality of the data was high, as can be seen in the FASTQC results (Fig. 4.4). A small proportion of reads were removed by Trimmomatic due to poor quality (Fig. 4.5).

Use of the domestic cat reference genome was a potential source of bias in assembly and variant calling. In general, reference-based assembly can lead to the propagation of errors in the reference sequence. The domestic cat assembly, however, is of high quality (Zhang & Schoenebeck, 2020), and the current version is the latest in a series of upgrades and improvements (Buckley et al., 2020; Pontius et al., 2007). Mapping wildcat data to a domestic reference, however, potentially favours domestic-like alleles (Reinert et al., 2015). Highly polymorphic regions may have been difficult to align at all. Use of the domestic cat reference may have resulted in the loss of information about novel wildcat sequences, copy number variation or larger structural rearrangements, which can occur even between closely related species (Fuentes-Pardo & Ruzzante, 2017).

For many species, including the European wildcat, high-quality reference data is not available, and it is common to use resources from a related species, primarily to avoid the expensive and time-consuming process of *de novo* assembly (Fuentes-Pardo & Ruzzante, 2017). Sequences from related species have been used to create draft genome assemblies for several species, for example the Egyptian water buffalo, using the cattle genome (El-Khishin et al., 2020), or to improve *de novo* assembly via reference-assisted assembly (Gnerre, Lander, Lindblad-Toh, & Jaffe, 2009), for example in African elephants (Lindblad-Toh et al., 2011) or Tibetan antelope (Kim et al., 2013). Alternate reference sequences are routinely used in resequencing projects where data from a related agricultural or model species are available, including bighorn sheep (Kardos et al., 2015) or Alpine ibex (Kessler et al., 2021).

Alignment to the domestic cat reference genome did not appear to impact the alignment rate of putative European wildcats in the dataset. Mean alignment rate across the three samples sources (domestic, putative wildcat and hybrid, Table 4.4) was highest in the hybrid group, and no group appeared to deviate far from the overall mean (~97%). The lowest alignment rate reported for an individual was from a domestic cat (79.3%), potentially indicating the presence of contamination or adaptor sequences within in this sample. Additional steps were taken to compensate for any potential bias towards 'domestic' alleles, specifically the removal of sites with missing data. Despite no obvious systemic bias in missingness across the three source populations (Fig. 4.10, Appendix 6) a stringent approach was taken, retaining only variant sites genotyped across all samples. This had the additional benefit that imputation of missing data was not required, which would necessitate a validated set of wildcat reference data (Browning, Zhou, & Browning, 2018). A cautious approach to mitigating any domestic bias in the dataset was important given the nature of the investigation in subsequent chapter, modelling hybridisation and introgression between the two species.

It is clear from Fig. 4.9 (Appendix 6) that misalignment of reads was an issue in some regions, as indicated by inflated coverage. This usually occurs in repetitive regions, where the true read origin could not be identified (Ekblom & Wolf, 2014). These sites were removed during the second round of filtering, which included a threshold for maximum read depth (Fig. 4.8).

Data generated by this study does not provide the read depth needed for *de novo* genome assembly (sample number was prioritised here over sequencing depth). Future work to generate a reference sequence for wildcats would be beneficial, using greater sequencing depth and a mix of long and short reads to improve alignment, resolve repetitive regions, and identify copy number or structural variation.

Domestic cat reference data was also used during phasing. Recombination rates can vary within and between species, and it has been hypothesised that the process of domestication elevates recombination rate (Coop & Przeworski, 2007; Li et al., 2016). For this analysis the existing domestic cat recombination map (Li et al., 2016) was modified to impose a minimum recombination rate of 5×10^{-7} cM/bp (removing any extremely low recombination rates reported in domestic cats). Development of a recombination map for wildcats and evaluation of recombination rates in hybrids would be valuable.

A joint genotyping approach to variant calling appeared to be successful. There are several advantages of joint genotyping (versus a single sample approach) (DePristo et al., 2011; Poplin et al., 2017). Firstly, calls are made for all individuals at sites where any individual shows variation. This allows a clear distinction between individuals with missing data and those that are homozygous for the reference allele, making it straightforward to evaluate missingness across the dataset. Importantly, calls can be made at low coverage sites if present in another sample. A large sample size improves the prior for genotype likelihood estimation and improves error modelling, e.g., variant quality score recalibration (though see below). Here, data from a range of sequencing platforms (with different error models) may have helped to avoid systemic sequencing bias in the dataset.

Processing from raw read data to SNP variants followed a modified GATK best practice protocol (DePristo et al., 2011; Van der Auwera et al., 2013). Important deviations from GATK best practice included the omission of recalibration steps: base quality score recalibration (BQSR) and variant quality score recalibration (VQSR). BQSR uses empirical data to adjust base quality scores, accounting for sequencing error. Similarly, VQSR uses a reference set of high confidence calls to train a machine learning approach to variant filtering (more nuanced than hard filtering). Both require a validated set of reference data, currently not available for domestic cats or wildcats. This is often the case for non-model species, where a bootstrapping approach offers an alternative (Kardos et al., 2018). Bootstrapping uses the highest scoring base or variant calls in the existing data as a training set for calibration, repeating variant calling and improvement of the training set until quality scores

converge. For future studies using this dataset, especially those investigating rare variants, or functional variation, it may be valuable to validate the current calls using a bootstrapping approach to recalibration.

Instead of VQSR, stringent hard filtering was used. First, using filters and thresholds proposed by GATK (Table 4.3, Van der Auwera et al., 2013) including important filters to identify patterns of mapping error, strand bias, and positional bias. Variant quality, read depth, missingness and SNP density were evaluated before a second round of filtering (Fig. 4.8; Fig. 4.9-4.10, Appendix 6). Hard filtering is not as flexible as VQSR, and only considers a single piece of information about a site at a time. It requires a trade-off between retaining a small number of high confidence variants, potentially removing true positives, or keeping a large number of variants, potentially containing a proportion of false positives. A stringent approach to filtering was taken here, with high thresholds for quality and missingness. This was primarily to ensure subsequent analysis of hybridisation and introgression was not biased by alignment/variant calling. In a recent analysis of 74 domestic cat samples (aligned to the domestic reference) ~40,000,000 SNPs were described (Buckley et al., 2020), approximately equal to the number of SNPs initially called here (~34 million), but much higher than the final number reported (~12 million). An important difference between the two approaches to variant calling was the application of GATK's BQSR by Buckley *et al.* (2002), using a set of strict filters to build a reference dataset, and much lower filtering thresholds to generate the final set of SNPs, post-BQSR. The difference between the two studies suggests that a stringent, hard-filtering approach has potentially led to the omission of a number of true positives here. However, strict filters were important to eliminate bias towards domestic or domestic-like individuals versus individuals with higher proportions of putative wildcat ancestry.

4.5 Conclusion

A set of high-confidence, analysis-ready SNPs were generated from a representative sample of modern Scottish wildcat and domestic cat populations, as well as additional reference samples from outside the UK. Data were successfully phased for the application of haplotype-based methods. The pipeline provides a workflow to generate reproducible data from future sampling of the wildcat population, though would benefit from additional genomic resources specific to wildcats.

4.6 References

Allendorf, F. W. (2017). Genetics and the conservation of natural populations: allozymes to genomes. *Molecular Ecology*, 26, 420–430

Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, 11(10), 697–709

Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data. Available online at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

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

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, 103(3), 338–348

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81(5), 1084–1097

Buckley, R. M., Davis, B. W., Brashear, W. A., Farias, F. H. G., Kuroki, K., Graves, T., … Warren, W. C. (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), 1–28

Coop, G., & Przeworski, M. (2007). An evolutionary view of human recombination. *Nature Reviews Genetics*, 8(1), 23–34

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., … Pereira, R. J. (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, 30, 3–13

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158

DeLong, E. F. (2009). The microbial ocean from genomes to biomes. *Nature*, 459(7244), 200–206

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491–501

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., … Reid, C. A. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 327(5961), 78–81

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, 7(9), 1026–1042

El-Khishin, D. A., Ageez, A., Saad, M. E., Ibrahim, A., Shokrof, M., Hassan, L. R., & Abouelhoda, M. I. (2020). Sequencing and assembly of the Egyptian buffalo genome. *PLoS ONE*, 15(8), e0237087

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution*, 29(1), 51–63

Flanagan, S. P., Forester, B. R., Latch, E. K., Aitken, S. N., & Hoban, S. (2018). Guidelines for planning genomic assessment and monitoring of locally adaptive variation to inform species conservation. *Evolutionary Applications*, 11(7), 1035–1052

Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology*, 26(20), 5369–5406

Gnerre, S., Lander, E. S., Lindblad-Toh, K., & Jaffe, D. B. (2009). Assisted assembly: How to improve a de novo genome assembly by using related species. *Genome Biology*, 10, R88

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351

Hassan, L. M. A., Arends, D., Rahmatalla, S. A., Reissmann, M., Reyer, H., Wimmers, K., … Brockmann, G. A. (2018). Genetic diversity of Nubian ibex in comparison to other ibex and domesticated goat species. *European Journal of Wildlife Research,* 64, 52

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., … Nielsen, E. E. (2011). Application of SNPs for population genetics of nonmodel organisms: New opportunities and challenges. *Molecular Ecology Resources*, 11(Suppl. 1), 123–136

Huang, J., Liang, X., Xuan, Y., Geng, C., Li, Y., Lu, H., … Gao, S. (2017). A reference human genome dataset of the BGISEQ-500 sequencer. *GigaScience*, 6(5), 1–9

Janjua, S., Peters, J. L., Weckworth, B., Abbas, F. I., Bahn, V., Johansson, O., Rooney, T. P. (2020) Improving our conservation genetic toolkit: ddRAD-seq for SNPs in snow leopards. *Conservation Genetics Resources,* 12, 257-261

Kardos, M., Åkesson, M., Fountain, T., Flagstad, Ø., Liberg, O., Olason, P., … Ellegren, H. (2018). Genomic consequences of intensive inbreeding in an isolated wolf population. *Nature Ecology and Evolution,* 2(1), 124–131

Kardos, M., Luikart, G., Bunch, R., Dewey, S., Edwards, W., McWilliam, S., … Kijas, J. (2015). Whole-genome resequencing uncovers molecular signatures of natural and sexual selection in wild bighorn sheep. *Molecular Ecology*, 24, 5616–5632

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., & Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, 9(10), 1205–1218

Kessler, C., Brambilla, A., Waldvogel, D., Camenisch, G., Biebach, I., Leigh, D. M., … Croll, D. (2021). A robust sequencing assay of a thousand amplicons for the high-throughput population monitoring of Alpine ibex immunogenetics. *Molecular Ecology Resources*, 00, 1–20

Kim, J., Larkin, D. M., Cai, Q., Asan, Zhang, Y., Ge, R. L., … Ma, J. (2013). Reference-assisted chromosome assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 110(5), 1785–1790

Knaus, B. J., & Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, 17(1), 44–53

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359

Leitwein, M., Duranton, M., Rougemont, Q., Gagnaire, P. A., & Bernatchez, L. (2020). Using Haplotype Information for Conservation Genomics. *Trends in Ecology and Evolution*, 35(3), 245–258

Li, G., Hillier, L. D. W., Grahn, R. A., Zimin, A. V., David, V. A., Menotti-Raymond, M., … Murphy, W. J. (2016). A high-resolution SNP array-based linkage map anchors a new domestic cat draft genome assembly and provides detailed patterns of recombination. *G3: Genes, Genomes, Genetics*, 6(6), 1607–1616

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

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., … Sodergren, E. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478, 476–482

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, 17(2), 142–152

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics*, 4(12), 981–994

Mak, S. S. T., Gopalakrishnan, S., Carøe, C., Geng, C., Liu, S., Sinding, M. H. S., … Gilbert, M. T. P. (2017). Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *GigaScience*, 6(8), 1–13

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9, 387–402

Martin, C. A., Armstrong, C., Illera, J. C., Emerson, B. C., Richardson, D. S., & Spurgin, L. G. (2021). Genomic variation, population history and within-archipelago adaptation between island bird populations. *Royal Society Open Science*, 8, 201146

McCartney-Melstad, E., Gidiş, M., & Shaffer, H. B. (2018). Population genomic data reveal extreme geographic subdivision and novel conservation actions for the declining foothill yellow-legged frog. *Heredity*, 121(2), 112–125

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20, 1297–1303

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., … Phillippy, A., M. (2020) Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585**,** 79–84

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6), 315–327

Minias, P., Dunn, P. O., Whittingham, L. A., Johnson, J. A., & Oyler-McCance, S. J. (2019). Evaluation of a Chicken 600K SNP genotyping array in non-model species of grouse. *Scientific Reports*, 9, 6407

Moore, G. M. (1965). Cramming more components onto integrated circuits with unit cost. *Electronics*, 38(8), 114

Murchison, E. P., Schulz-Trieglaff, O. B., Ning, Z., Alexandrov, L. B., Bauer, M. J., Fu, B., … Stratton, M. R. (2012). Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell*, 148(4), 780–791

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6), 443–451

Ng, P. C. & Kirkness, E. F. (2010) Whole Genome Sequencing. In Barnes, M. & Breen, G. (Eds.) *Genetic Variation. Methods in Molecular Biology (Methods and Protocols)*. Totowa, USA: Humana Press

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, 7(5), e37135

Piggott, M. P., & Taylor, A. C. (2003). Remote collection of animal DNA and its applications in conservation management and understanding the population biology of rare and cryptic species. *Wildlife Research*, 30(1), 1–13

Pontius, J. U., Mullikin, J. C., Smith, D. R., Lindblad-Toh, K., Gnerre, S., Clamp, M., … McKernan, K. (2007). Initial sequence and comparative analysis of the cat genome. *Genome Research*, 17(11), 1675–1689

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., … Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 1–22

Purnick, P. E. M., & Weiss, R. (2009). The second wave of synthetic biology: From modules to systems. *Nature Reviews Molecular Cell Biology*, 10(6), 410–422

Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of Next-Generation Sequencing Reads. Annual Review of Genomics and Human Genetics, 16, 133–151

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856), 737–746

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-termination inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5463–5467

Scheben, A., Batley, J., & Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnology Journal*, 15, 149–161

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., … Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, 30(2), 78–87

Trevelline, B. K., Fontaine, S. S., Hartup, B. K., & Kohl, K. D. (2019). Conservation biology needs a microbial renaissance: A call for the consideration of host-associated microbiota in wildlife management practices. *Proceedings of the Royal Society B: Biological Sciences*, 286(1895)

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43, 11.10.1-11.10.33

Van der Auwera, G. A. & O'Connor, B. D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition).* Sebastopol, USA: O'Reilly Media.

Vega Thurber, R. L., Barott, K. L., Hall, D., Liu, H., Rodriguez-Mueller, B., Desnues, C., … Rohwer, F. L. (2008). Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. *Proceedings of the National Academy of Sciences of the United States of America*, 105(47), 18413–18418.

Zhang, W., & Schoenebeck, J. J. (2020). The ninth life of the cat reference genome, felis_catus. *PLoS Genetics*, 16(10), 8–11

4.7 Appendix 5. WGR sample information

Table 4.6. Whole genome resequencing data. Sample ID, source population (wild/captive/domestic) and country of origin are provided per individual. Q35 scores were available for putative Scottish wildcats. Individuals or organisation that provided samples for this study are named in the source column, for public data this column contains the sample accession number.

Sample ID	Population	Country	Q35	Source	Included in final dataset?
WCQ0047	Wild	Scotland	0.931	RZSS	Yes
WCQ0052	Wild	Scotland	0.91	RZSS	Yes
WCQ0099	Wild	Scotland	0.012	RZSS	Yes
WCQ0100	Wild	Scotland	0.289	RZSS	Yes
WCQ0158	Wild	Scotland	0.876	RZSS	Yes
WCQ0165	Wild	Scotland	0.035	RZSS	Yes
WCQ0168	Wild	Scotland	0.402	RZSS	Yes
WCQ0209	Wild	Scotland	0.666	RZSS	Yes
WCQ0211	Wild	Scotland	0.884	RZSS	Yes
WCQ0212	Wild	Scotland	0.794	RZSS	Yes
WCQ0213	Wild	Scotland	0.732	RZSS	Yes
WCQ0214	Wild	Scotland	0.909	RZSS	Yes
WCQ0216	Wild	Scotland	0.599	RZSS	Yes
WCQ0218	Wild	Scotland	0.777	RZSS	Yes
WCQ0224	Wild	Scotland	0.796	RZSS	Yes
WCQ0227	Wild	Scotland	0.288	RZSS	Yes
WCQ0230	Wild	Scotland	0.367	RZSS	Yes
WCQ0231	Wild	Scotland	0.466	RZSS	Yes
WCQ0234	Wild	Scotland	0.626	RZSS	Yes
WCQ0236	Wild	Scotland	0.489	RZSS	Yes
WCQ0243	Captive	Scotland	0.937	RZSS	Yes
WCQ0246	Wild	Scotland	0.433	RZSS	Yes
WCQ0248	Wild	Scotland	0.974	RZSS	Yes
WCQ0249	Wild	Scotland	0.632	RZSS	Yes
WCQ0252	Wild	Scotland	0.528	RZSS	Yes
WCQ0255	Wild	Scotland	0.327	RZSS	Yes
WCQ0340	Captive	Scotland	0.938	RZSS	Yes
WCQ0343	Captive	Scotland	0.991	RZSS	Yes
WCQ0344	Captive	Scotland	0.99	RZSS	No
WCQ0402	Captive	Scotland	0.967	RZSS	No
WCQ0427	Captive	Scotland	0.865	RZSS	Yes
WCQ0428	Captive	Scotland	0.967	RZSS	Yes
WCQ0443	Domestic	Scotland	0.01	RZSS	Yes
WCQ0515	Wild	Scotland	0.992	RZSS	Yes
WCQ0550	Captive	Scotland	0.977	RZSS	No
WCQ0553	Captive	Scotland	0.981	RZSS	Yes
WCQ0564	Captive	Scotland	0.931	RZSS	No

WCQ0613	Wild	Scotland	0.445	RZSS	Yes
WCQ0903	Wild	Scotland	0.644	RZSS	Yes
WCQ0904	Wild	Scotland	0.344	RZSS	Yes
WCQ0915	Wild	Scotland	0.799	RZSS	Yes
WCQ1135	Domestic	Scotland		RZSS	Yes
WCQ1136	Domestic	Scotland		RZSS	Yes
WCQ1137	Domestic	Scotland		RZSS	Yes
WCQ1138	Domestic	Scotland		RZSS	Yes
FSX360	UNK	Portugal		Carlos Driscoll	Yes
ex19	Wild	Germany (east)		Carsten Nowak/Violeta Muñoz-Fuentes	Yes
ex21	Wild	Germany (west)		Carsten Nowak/Violeta Muñoz-Fuentes	Yes
ex38	Wild	Germany (east)		Carsten Nowak/Violeta Muñoz-Fuentes	Yes
ex40	Wild	Germany (west)		Carsten Nowak/Violeta Muñoz-Fuentes	Yes
ex9	Wild	Germany (east)		Carsten Nowak/Violeta Muñoz-Fuentes	Yes
FA661	Wild	Germany (east)		Carsten Nowak/Violeta Muñoz-Fuentes	Yes
SRR5040107	Domestic	Oman		SAMN05980322	Yes
SRR5040110	Domestic	USA		SAMN05980329	Yes
SRR5040111	Domestic	Madagascar		SAMN05980324	Yes
SRR5040113	Domestic	Iraq		SAMN05980320	Yes
SRR5040114	Domestic	Portugal		SAMN05980325	Yes
SRR5040116	Domestic	Denmark		SAMN05980323	Yes
SRR5040117	Domestic	Italy		SAMN05980326	Yes
SRR5040118	Domestic	Jordan		SAMN05980327	Yes
SRR5040120	Domestic	Thailand		SAMN05980319	Yes
SRR5040125	Domestic	USA		SAMN05980328	Yes
SRR5040126	Domestic	South Korea		SAMN05980321	Yes
SRR7621212	Domestic	China		SAMN09509226	Yes
SRR7621222	Domestic	China		SAMN09509239	Yes
SRR7621225	Domestic	China		SAMN09509242	Yes
SRR7621252	Domestic	China		SAMN09509250	Yes
SRR7621254	Domestic	China		SAMN09509256	Yes
SRR8377759	Domestic	USA		SAMN10661124	Yes

4.8 Appendix 6. Supplementary material

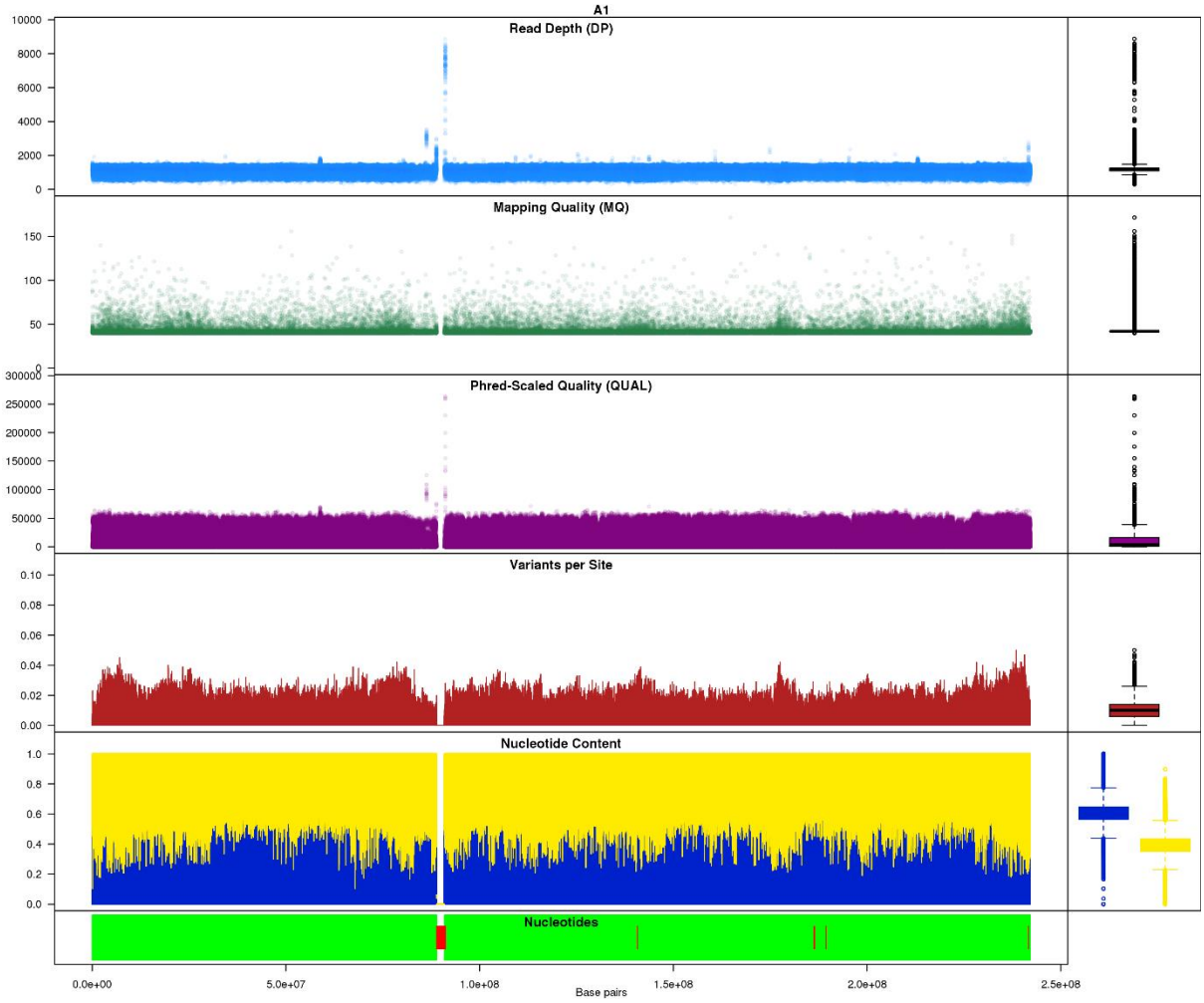

Figure 4.9. Example plot using vcfR (Knaus & Grünwald, 2017) to visual read depth, mapping quality, variant quality (QUAL), SNP density, nucleotide content and missing data across each chromosome. These plots were used to evaluate the output from the initial round of filtering and set thresholds for the second. For example, peaks of high read depth indicated repetitive regions, with poor alignment, leading to the threshold for maximum read depth, DP=2000.

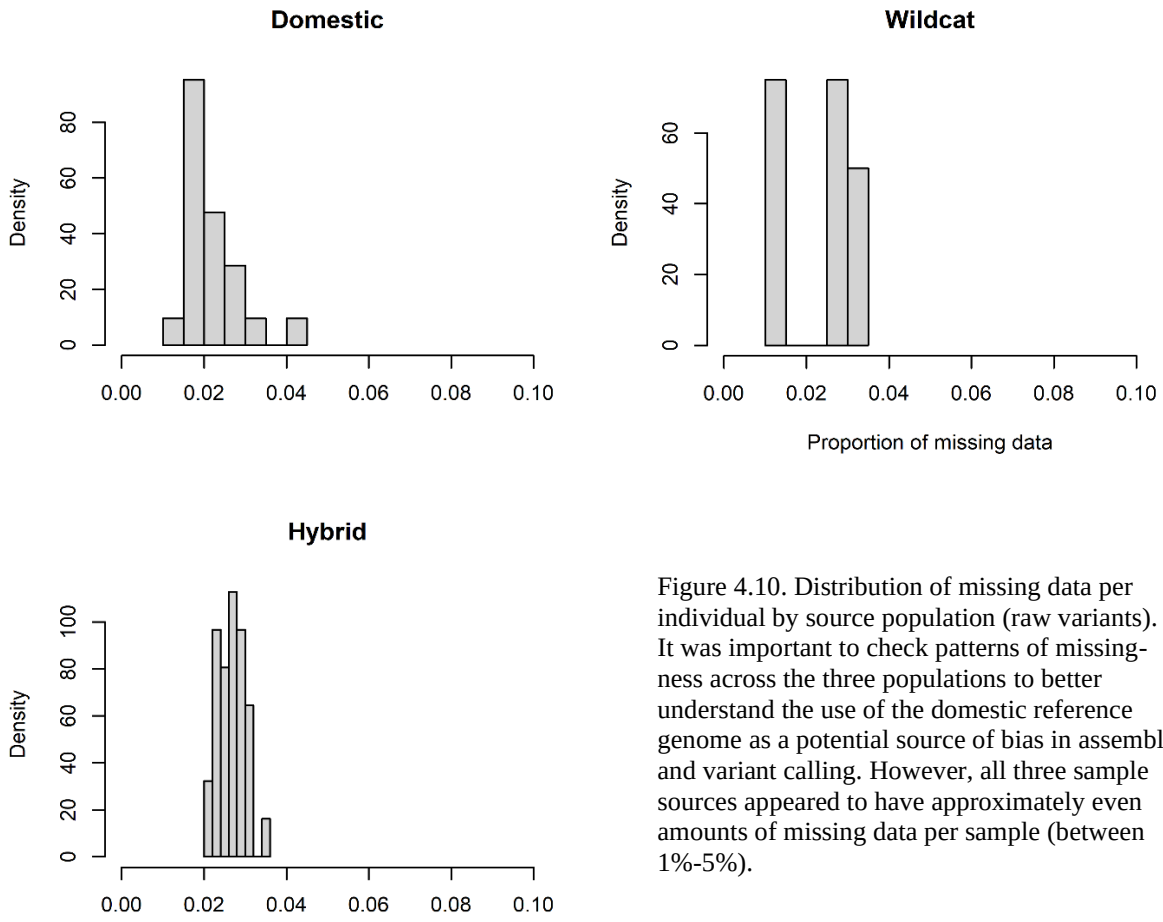

Figure 4.10. Distribution of missing data per individual by source population (raw variants). It was important to check patterns of missingness across the three populations to better understand the use of the domestic reference genome as a potential source of bias in assembly and variant calling. However, all three sample sources appeared to have approximately even amounts of missing data per sample (between 1%-5%).

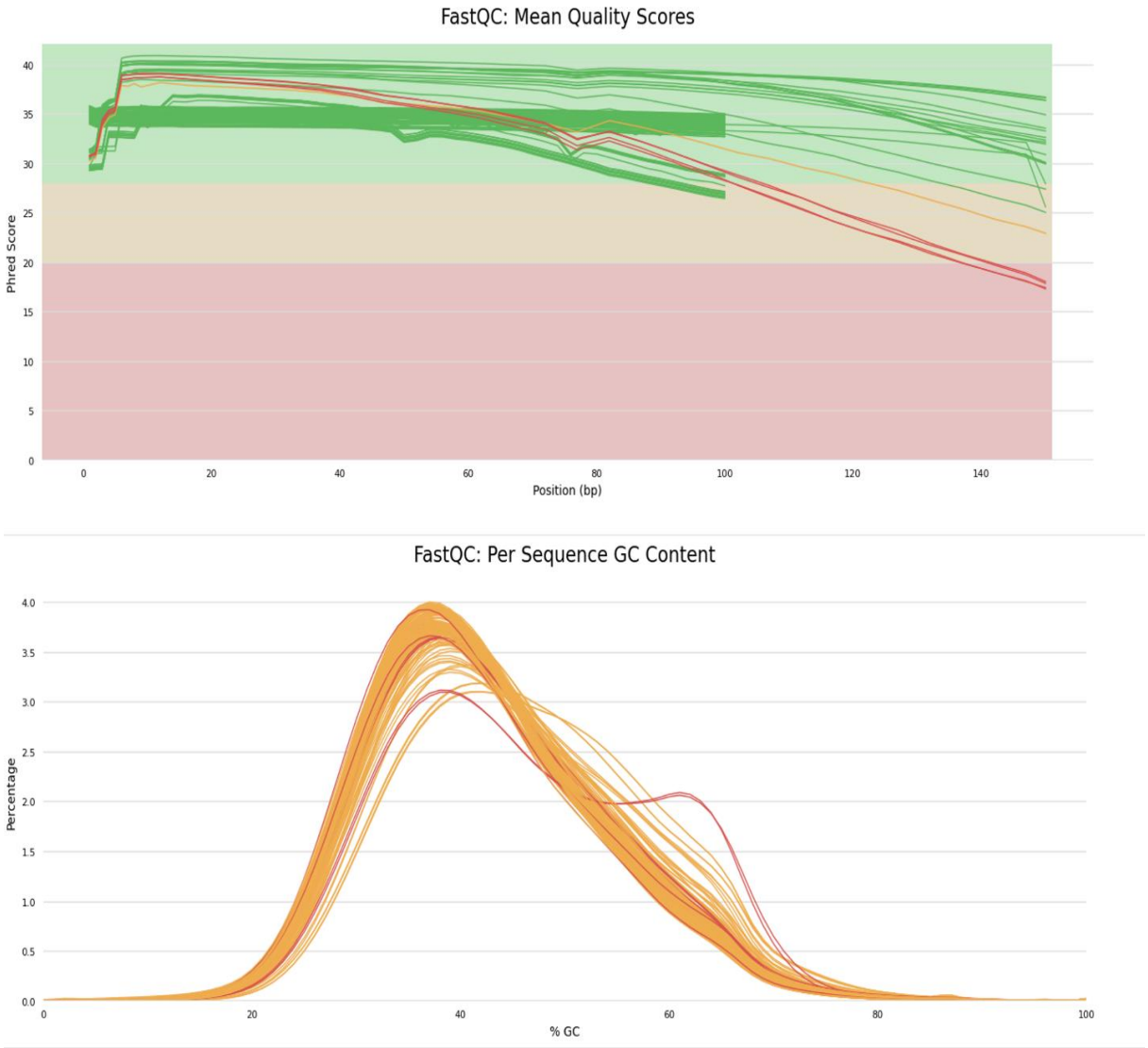

Figure 4.11. Two FastQC modules where a proportion of samples failed: Mean Quality Score Per Base (top) and Per Sequence GC Content (bottom). The top plot highlights some samples where base quality decreases towards the ends of reads, which was improved by trimming. The bottom plot shows skewed distributions of GC content for most samples, but without sharp peaks that can indicate the presence of contamination.

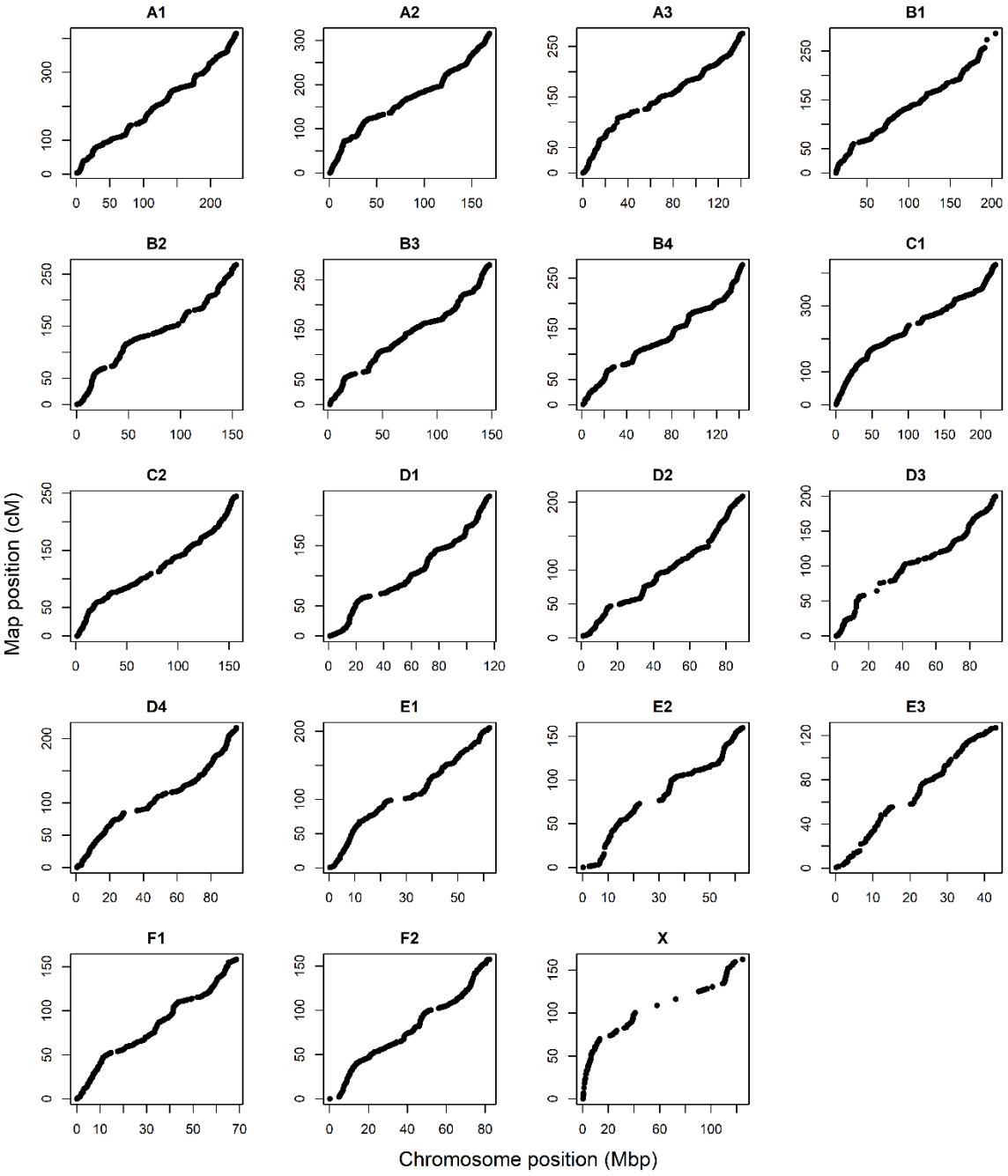

Figure 4.12. Recombination maps (per chromosome) used for phasing. Maps were modified from those generated for domestic cats (Li *et al.* 2016).

Chapter 5 Haplotype-based methods to date admixture

5.1 Introduction

Previous results (Chapter 3) demonstrate a strong signal of recent admixture in the Scottish wildcat population. However, unlinked SNP data has limited power to detect older admixture events, and thus far we have been unable to rule out early-20th century hybridisation as a result of population expansion, or more ancient signatures of admixture (important given the sympatry of domestic cats and wildcats over the last few thousand years). In this chapter I apply whole genome resequencing data (described in Chapter 4), to haplotype-based methods for accurate dating of admixture in Scottish wildcats.

5.1.1 Haplotype-based methods for characterising admixture

Principal component analysis (PCA) and clustering algorithms, such as STRUCTURE (Pritchard et al., 1999) or ADMIXTURE (Alexander et al., 2009), capture useful information regarding patterns of genetic variation between individuals, and are widely used to identify admixture, characterise admixing 'source' populations and estimate admixture proportions (Wangkumhang & Hellenthal, 2018). However, these methods are not informative about the timing of admixture events. For this, haplotype information, to examine patterns of admixture along individual chromosomes, is the primary tool.

A haplotype is the combination of alleles found together on a chromosome, inherited from a single parent (Delaneau & Zagury, 2012). Haplotypes are formed through mutation and recombination, and haplotype diversity within a population is determined by neutral and adaptive evolutionary processes, such as selection and drift. Linkage disequilibrium (LD) refers to the non-random association of alleles belonging to the same haplotype. The probability of a crossing over event separating two loci increases with distance; loci separated by larger distances are therefore less tightly linked (i.e., LD decays along the chromosome). Within a homogenous population, drift and selection act to maintain a background rate of LD. In humans, this typically becomes negligible over more than a few hundred kilobases (Reich et al., 2001).

Admixed individuals inherit intact haplotypes from two distinct populations. In the second generation after admixture, chromosomes from the mixing groups begin to recombine, breaking down ancestral haplotypes into successively smaller 'chunks' (Fig. 5.1) (Gravel, 2012; Pool & Nielsen, 2009). This provides an important signal corresponding to the length of time since admixture; longer contiguous 'chunks' are the result of recent admixture, where recombination has acted over fewer generations. The opposite is true for short 'chunks', which have been broken up by several generations of recombination. Following a single (pulse) admixture event, chunk length (in Morgans)

decays exponentially, with rate parameter, *r*, proportional to the number of generations since admixture.

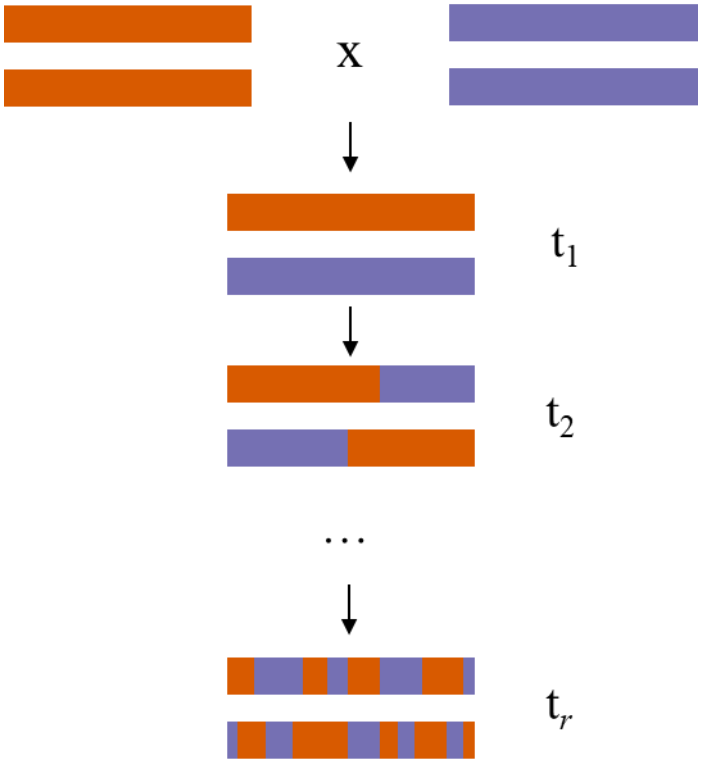

Figure 5.1. Admixture between two populations, 'orange' and 'purple', and the subsequent breakdown of ancestral haplotypes by recombination. First generation hybrids receive a single chromosome copy from each parent. In the second generation following admixture (t_2), recombination between the two chromosome copies results in the breakdown of ancestral haplotypes into smaller 'chunks'. At each successive generation (t_r), recombination continues to act on the ancestral haplotypes. The distribution of chunk lengths is therefore informative about the number of generations since admixture.

Local ancestry along the chromosome is not directly observable; understanding the history of admixed populations therefore relies on our ability to infer haplotype origin. Several statistical methods have been developed to this end. An important step was the extension of the STRUCTURE model to account for 'admixture LD', the non-random association between markers that are inherited together, in admixed populations, as part of the same ancestral 'chunk' (Falush, Stephens, & Pritchard, 2003). STRUCTURE approaches markers as unlinked, and therefore providing independent information about ancestry (Pritchard et al., 1999). In admixed populations, however, individuals often inherit linked alleles together on the same DNA 'chunk' (Fig. 5.1). The STRUCTURE 'linked model' uses a Hidden Markov Model (HMM) to identify correlations in ancestry along the chromosome and model inheritance of discrete 'chunks', rather than independent alleles (Falush et al., 2003). The linked model does not account for background LD, however, giving limited power to detect ancestry tracts, possibly leading to overestimation of admixture timings.

Other statistical methods have been developed to infer local ancestry tracts, accounting for background and admixture LD. These methods generally rely on reference samples (as surrogates for the mixing groups) to identify patterns of shared variation across the genome. The originally admixing groups are often hard to sample due to drift, extinction, or later admixture events.

A common approach of tract-based methods is to portion the genome into non-overlapping windows and determine local ancestry per window, an approach taken by, e.g., PCAdmix (Brisbin et al., 2012) or LAMP-LD (Baran et al., 2012). HAPMIX (Price et al., 2009), allows ancestry switches at any point across the genome. It extends Li and Stephens' (2003) approach to LD modelling, which constructs 'offspring' haplotypes as a mosaic of reference 'parental' haplotype blocks. HAPMIX applies the same principal to build admixed genomes from partial haplotypes of two reference populations representing the mixing groups (Price et al., 2009). This allows ancestry transitions at two scales: (1) small-scale, switching between haplotypes within a reference population (representing older, pre-admixture, recombination), and (2) large-scale, between reference populations (i.e., post-admixture recombination) (Fig. 5.2). Based on patterns of shared variation, the likelihood of inheritance from reference population A vs. population B can be estimated for a given point in the genome. This information is combined with that of neighbouring loci to give a probabilistic estimate of the ancestry at each locus.

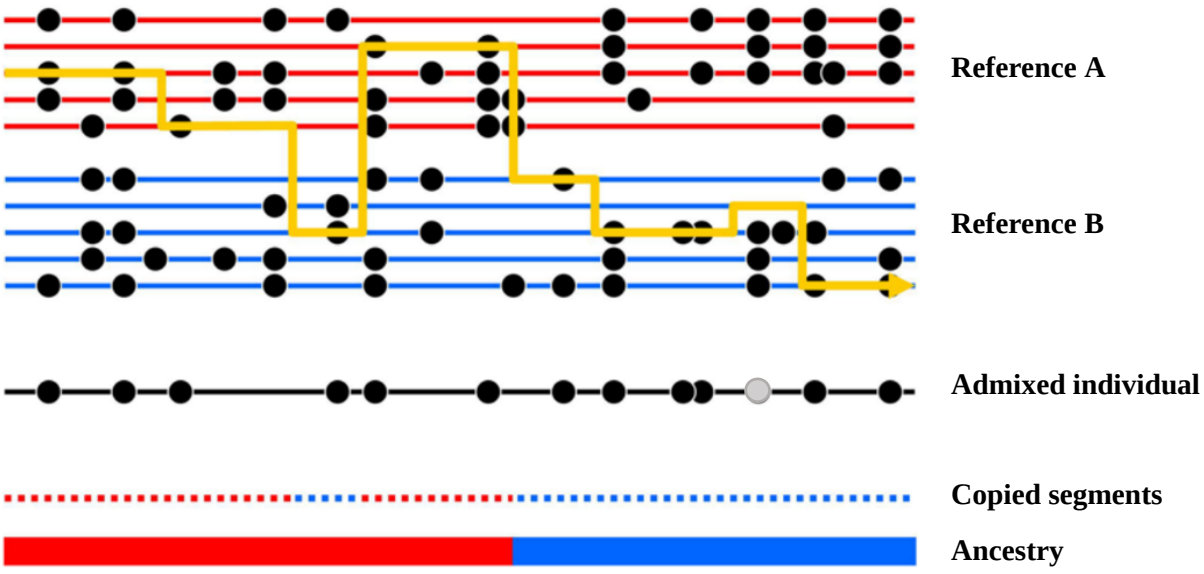

Figure 5.2. HAPMIX model to identify local ancestry along the genome. Haplotypes from two (unadmixed) populations (references A [red] and B [blue]) are sampled as surrogates for the two mixing groups. Black circles represent polymorphism in the sampled individuals. A haplotype from an admixed individual is shown below. The admixed chromosome is reconstructed as a mosaic of the sampled reference individuals, the yellow line indicates which individuals are copied from for any one segment. Local ancestry inference, based on the copied segments, is shown at the bottom, coloured by putative ancestry. Inference is improved by allowing a rate of miscopying (e.g., copying a 'B segment' under 'A ancestry', note the blue segment copied within a block of red ancestry) and mutation (the grey circle in the admixed haplotype, not present in the reference samples).

From Price *et al.* (2009)

Haplotype-copying methods rely on accurate phasing of admixed haplotypes (Price et al., 2009). Incorrect phasing may result in artefactual ancestry switches, with a knock-on effect on tract length and accurate dating. To account for this, HAPMIX includes a built-in phasing algorithm to average ancestry inference over all possible phasing combinations. Phasing of reference populations is treated as accurate, mostly because phasing errors within reference populations should not impact local ancestry switches. However, even for the admixed population alone this approach is computationally intensive (Salter-Townshend & Myers, 2019).

Identifying patterns of shared variation is complicated by mutation (or genotyping error) and recombination, and necessitates sufficient sampling of reference population diversity (Price et al., 2009). This is often complicated by drift of the sampled modern populations from the ancestral admixing groups. HAPMIX therefore parameterises miscopying and mutation rates, allowing for inconsistencies between admixed and reference individuals. The miscopying parameter is especially important to account for incomplete lineage sorting, either as a result of poor sampling of haplotype diversity or drift in the reference populations. This parameterisation improves HAPMIX local ancestry inference, particularly to accurately identify short tracts associated with ancient admixture events (up to 400 generations in human data). Nevertheless, close representatives of the mixing groups are needed as reference populations. HAPMIX only models two-way admixture scenarios.

A range of methods have been developed to estimate admixture timing without directly inferring local ancestry. Instead, these approaches measure the decay of admixture LD across the genome, i.e., the association of loci inherited on the same ancestral 'chunk', and its subsequent breakdown by recombination (Wangkumhang & Hellenthal, 2018). Tools such as ROLLOFF (Moorjani et al., 2011) and ALDER (Loh et al., 2013) measure LD decay between pairs of loci, weighted by the power of each locus to discriminate between reference populations. A curve can be fitted to the correlation between loci over genetic distance, which can be used to estimate time since admixture. This has the advantage over local ancestry methods in terms of the reference data required. For example, ALDER can be run using reference data that is diverged from the ancestral mixing groups, and in some cases, using reference data from a single population only (Loh et al., 2013).

GLOBETROTTER extends this approach to measure correlation between haplotype segments, increasing analysis power by combining information across successive markers (Hellenthal et al., 2014). GLOBETROTTER first employs a haplotype-copying model (related to that of Li and Stephens, 2003), CHROMOPAINTER (Lawson, Hellenthal, Myers, & Falush, 2012), to construct admixed ('target') haplotypes as a mosaic of multiple reference ('donor') populations (Hellenthal et al., 2014). The (unsampled) ancestral ('source') groups can also be described as a weighted mixture of the donor populations. Donors, therefore, do not need to be sampled from the source groups; the

haplotype-copying approach will use the most closely related population. Inference is improved, however, by using extant groups similar to the ancestral mixing groups. Model-fitting narrows down the donors to produce a 'clean' signal reflecting best-fit true ancestry, which closely represents the 'true' ancestry if the assumptions hold, of relatively simple recent admixture with no strong genetic bottleneck or multi-generational processes involving different populations. The estimate of ancestral groups is still informative about more complex scenarios.

GLOBETROTTER constructs coancestry curves to estimate admixture timing (Hellenthal et al., 2014). Coancestry curves capture information about tract length. Specifically, for an admixed genome, they show the probability of two loci being inherited under a given local ancestry, as a function of genetic distance. In admixed populations, coancestry curves will have the same exponential decay as the expected size distribution, with a rate parameter equal to the number of generations since admixture. Older admixture events generate steeper curves. A mixture of exponential curves indicates multiple admixture times.

LD decay methods are considered more accurate for characterising subtle admixture than approaches that directly infer local ancestry tracts (Wangkumhang & Hellenthal, 2018). GLOBETROTTER is informative about a range of admixture scenarios, including those with multiple admixture events and/or involving more than two populations. However, accurate estimates of admixture proportions can be difficult to obtain (depending on the donor populations used), and in cases of multiple admixture events, timing estimates may be biased towards the most recent event. It can be challenging to distinguish multiple admixture events from continuous admixture (Hellenthal et al., 2014). All of the methods described so far are limited by their reliance on reference populations to represent the original admixing groups, and, ultimately, a poor understanding of the relationship between modern reference populations and ancestral groups.

Here, we apply MOSAIC to estimate the timing of admixture in the Scottish wildcat population. Uniquely, MOSAIC does not require reference individuals to be close representatives of the admixing groups, instead, it infers relationships between populations from the data (Salter-Townshend & Myers, 2019). This is important to understand admixture in Scottish wildcats, where representatives of the originally admixing wildcat population are unlikely to exist. MOSAIC combines the approaches of GLOBETROTTER and HAPMIX to identify local ancestry tracts across the genome (like HAPMIX), combining information from multiple reference panels to better understand the unsampled ancestral populations (like GLOBETROTTER). Unlike GLOBETROTTER, this information is incorporated into an ancestry aware HMM, improving accuracy of local ancestry estimates.

Bi-allelic SNP data are first reduced to a grid of 60 points per centi-Morgan (cM), restricting recombination to between successive grid points (to make the algorithm computationally tractable).

SNPs are associated with their nearest grid point. Haplotype and ancestry are jointly inferred at each grid point along the chromosome, using a HMM similar to that of Li and Stephens (2003) and Price *et al* (2009). The main parameters estimated by the model are:

- Recombination rates pre- and post-admixture. These are parameterised separately, controlling copying switches within and between reference panels.
- A miscopying rate, allowing for genotyping error or mutation in target individuals
- Importantly, the relationship between each reference population and the admixed target group is parameterised by a copying matrix (the probability of copying from each reference population given the underlying ancestry). Reference populations that are poor surrogates for the originally admixing groups will have copying probabilities close to zero, and vice versa. Admixed reference populations will be copied under multiple ancestries.

Parameters are initialised using all possible donor individuals as part of an ancestry unaware, single layer HMM, similar to that of CHROMOPAINTER (Lawson et al., 2012). The MOSAIC algorithm then proceeds through several rounds of the following steps until convergence:

1. Thinning. The top one hundred donor individuals are selected from the full set of reference individuals, based on an ancestry unaware copying model (similar to CHROMOPAINTER [Lawson et al., 2012])
2. Rephasing. MOSAIC accounts for possible phasing errors (which impede accurate local ancestry estimates) by finding phase flips that lead to an increased likelihood of the data under the model
3. EM (expectation-maximisation) updates (10 iterations [Salter-Townshend & Myers, 2019]). Estimation of the hidden states (i.e., ancestry, in terms of the individual and population copied from at each grid point), and maximum-likelihood estimation of the parameters given the hidden states.

Following local ancestry inference, segments of the admixed ('target') genomes can be used to construct partial haplotypes considered to be drifted samples of the originally admixing populations (Salter-Townshend & Myers, 2019). Pairwise F_{ST} estimates between each reference panel and the partial 'ancestral' genomes provides important information about the relationships between the sampled data and inferred source populations. Coancestry curves can be fitted, as per GLOBETROTTER, to show the exponential decay of local ancestry tract length and estimate admixture timing. MOSAICs approach has been demonstrated to work well in human populations for complex, multi-way admixture scenarios, including older admixture events (shown up to 100 generations) (Salter-Townshend & Myers, 2019).

5.1.2 Aims

In this chapter I apply haplotype-based methods to the Scottish wildcat dataset, with the aim of obtaining a more accurate picture of the temporal patterns of hybridisation in Scotland. The chapter expands on the data analysed in previous chapters (Chapters 2 & 3) by incorporating additional reference samples from continental European wildcat populations, domestic cats from outside of the UK, as well as a sample of early 20th century putative Scottish wildcats. Using this data, we aim to examine population structure within the sampled individuals and define reference panels for local ancestry estimation. We then aim to use MOSAIC to model admixture in the Scottish wildcat population, and compare this to a window-based approach, PCAdmix, using local ancestry estimates to accurately date hybridisation in the Scottish wildcat population.

5.2 Methods

5.2.1 Final dataset

65 individuals were genotyped at 11,863,892 SNPs (see Chapter 4), including 30 wild-living and six captive putative Scottish wildcats and five Scottish domestic cats. An additional seven European wildcats and 17 domestic cats were sampled from outside Scotland (Fig. 5.3).

Additionally, low-coverage whole-genome sequence data were available from four historic samples, provided by Greger Larson and Laurent Frantz, and two archaeological samples (Jamieson et al., in prep.) from Britain, information about these samples is summarised in Table 5.1.

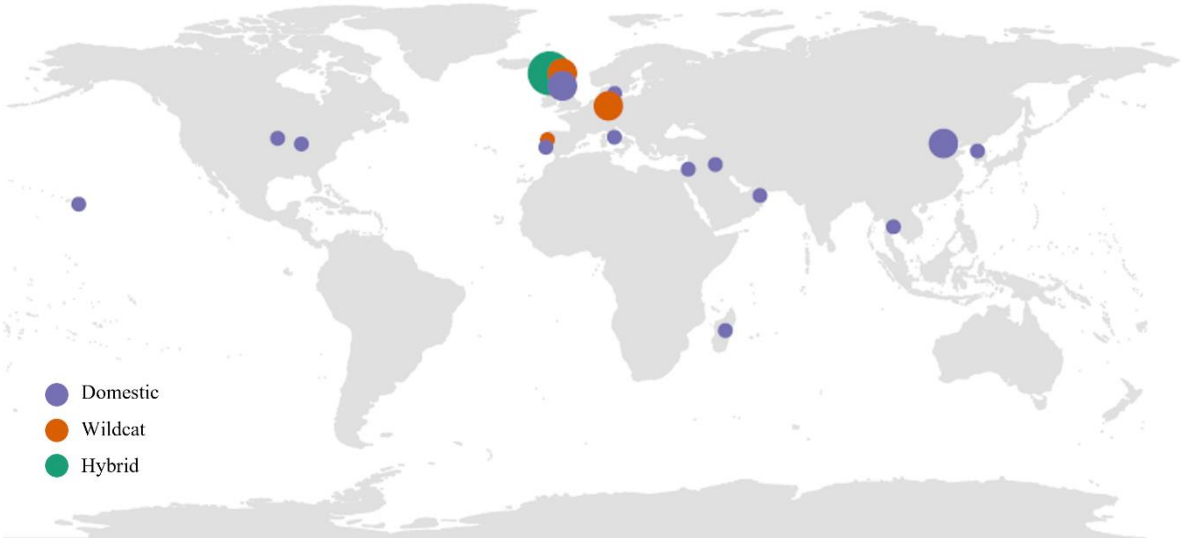

Figure 5.3. Sampling locations for whole-genome resequencing data. Locations are coloured by putative source population and with marker size corresponding to sample size. Domestic cat samples are shown in purple, putative wildcats in orange and putative hybrids (Scottish wild population) in green.

Table 5.1 Historic and archaeological sample information

Sample name	Location	Date	Sequence coverage	Genotyped SNPs
AJ324	England	~16th c.	0.9x	2,473,792
AJ419	Scotland	~6300 BCE	0.2x	616,385
WCQ0965	Scotland	1928	0.6x	7,731,714
WCQ0986	Scotland	1939	0.8x	9,530,472
WCQ1008	Scotland	1920	4.7x	19,598,613
WCQ1021	Scotland	1906	0.3x	4,684,265

A further 23 historic putative wildcat samples (primarily from Scotland) were genotyped at between 2,500 and 22,000 SNPs (data provided by Greger Larson and Laurent Frantz). These samples are summarised in Table 5.2 and referred to as the 'historic screening data'.

Table 5.2. Sampling date and location for the historic screening data

Sample name	Year sampled	Sampling location			
WCQ0948	1903	Scotland			
WCQ0949	1904	Scotland	WCQ1029	1938	Scotland
WCQ0956	1934	Scotland	WCQ1031	1938	Scotland
WCQ0966	1938	Scotland	WCQ1033	1938	Scotland
WCQ0967	1938	Scotland	WCQ1042	1907	Spain
WCQ1010	1929	Scotland	WCQ1057	1904	Scotland
WCQ1011	1946	Scotland	WCQ1058	1912	Romania
WCQ1016	1935	Scotland	WCQ1059	UNK	Spain
WCQ1023	1914	Scotland	WCQ1060	1906	Turkey
WCQ1026	1938	Scotland	WCQ1061	1956	Scotland
WCQ1027	1922	Scotland	WCQ1062	1956	Scotland
WCQ1028	1914	Scotland	WCQ1063	1956	Scotland

5.2.2 Principal component analysis

Principal component analysis (PCA) was completed for all 71 whole-genome sequencing samples (i.e., modern, historic and archaeological samples) using Eigensoft's smartpca (Patterson, Price, & Reich, 2006; Price et al., 2006). Outlier removal (of samples that appeared outlying with respect to principal component mean and standard deviation) was disabled; all samples were retained for the

analysis. The modern data were used to compute the first ten principal components (PCs), projecting the low coverage (historic/archaeological) samples onto these axes. Prior to PCA, low-coverage samples were filtered to include only bi-allelic sites with a genotyping quality score (QUAL) of at least 20. To minimise the amount of missing data for this analysis, PCs were computed based on SNPs that were also genotyped in at least one of the low-coverage samples, and then thinned to one SNP per 1kb. The final number of SNPs used for PCA was 862,730, the proportion of missing data for each low coverage sample is shown in Fig. 5.14 (Appendix 7).

Projection is a useful approach to PCA for samples with a high proportion of missing data. Eigenvectors capturing the most variation can be established using a set of high-quality reference data, and the coordinates of the low-coverage samples inferred (in the case of smartpca using a least squares solution). Computing PCs using samples with large amounts of missing data can otherwise lead to bias (Yi & Latch, 2021).

5.2.3 Defining reference populations

Most methods to detect or model admixture require reference samples from labelled populations as surrogates for the mixing groups. To confirm appropriate sampling from domestic and wildcat reference populations had taken place, population structure was first evaluated using ADMIXTURE and fineSTRUCTURE. These tools were run using a thinned set of markers (n=1,011,786); a minor allele count of at least three was imposed and markers were thinned at random to one SNP per 2kb.

As per Chapter 2, ADMIXTURE was run for values of K ranging from two to seven, including a calculation of cross-validation error.

FineSTRUCTURE uses CHROMOPAINTER to reconstruct the genome of each individual as a mosaic of haplotypes from the other samples in the analysis (Lawson et al., 2012). The expected number, or total length, of shared haplotype 'chunks' provides information about genetic relatedness between all pairs of individuals in the sample, referred to as the 'coancestry matrix'. FineSTRUCTURE uses this information about haplotype similarity to determine fine-scale population structure. Recombination information was provided using modified domestic cat recombination maps (4.2.7).

ADMIXTURE clustering showed the broad patterns in terms of domestic, wildcat and hybrid groups (the best fitting model used K=2). However, fineSTRUCTURE highlighted genetic differentiation between mainland European wildcats (sampled from three biogeographic groups, as per Mattucci *et al.* [2016]) and captive Scottish wildcats. On the PCA, ancient and historic British samples appeared to cluster with modern wildcats from mainland Europe, rather than the Scottish captive cats.

To formally test for domestic introgression in the captive Scottish wildcat population, F_4 statistics were calculated. F-statistics summarise allele-sharing between populations, and are used to determine whether relationships between populations conform to tree-based models, or are better explained by more complex models, e.g., involving admixture (Patterson et al., 2012; Reich, Thangaraj, Patterson, Price, & Singh, 2009). The F_4 statistic measures the correlation of allele frequencies between four populations, A, B, C and O (Fig. 5.4), and is calculated as

$$F_4(A, B; C, O) = E[(a' - b')(c' - o')]$$

where a', b', c' and o' represent the allele frequency in each population at a given site. In the absence of geneflow we would expect the allele frequency difference between A and B (across all genotyped sites) to be independent from that between C and O, and F_4 to be approximately equal to zero. Deviation from zero can be interpreted as evidence of geneflow.

For an admixed target population, X, the proportion of ancestry from the parental populations can be estimated using the F_4-ratio

$$\frac{F4(A, O; X, C)}{F4(A, O; B, C)}$$

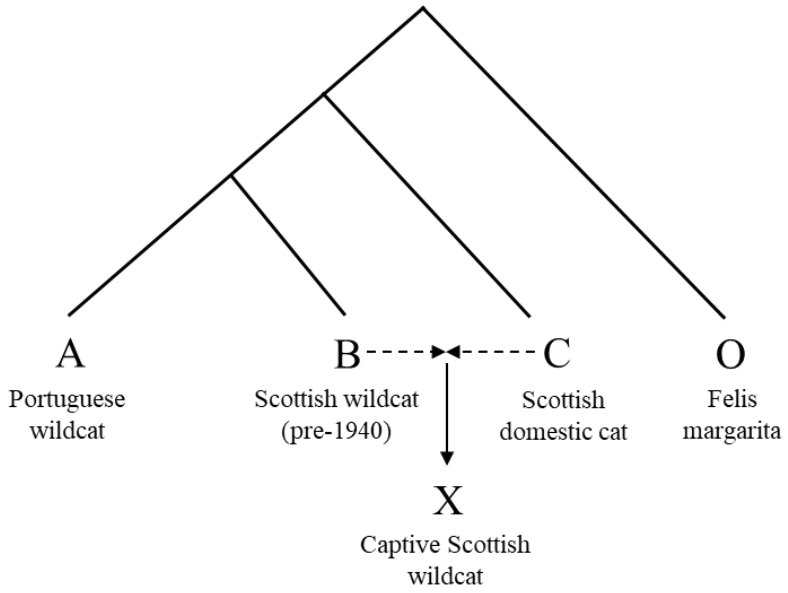

Figure 5.4. An example phylogeny used to calculate F_4 statistics. The target population (X) is a putative admixture of populations B and C (X is represented here by the captive Scottish wildcat population, a putative mix of early 20th-century Scottish wildcat and modern Scottish domestic ancestry).

Adapted from Patterson *et al.* (2012)

F_4 statistics were calculated using AdmixTools (Patterson et al., 2012), using the R interface, *admixr* (Petr, 2020). Ancestry proportions were estimated for all putative Scottish wildcats (historic, modern wild-living and modern captive) using both the F_4 ratio test and qpAdm (862,730 SNPs).

These tests were initially developed to model admixture in ancient human populations and are considered robust to missing data (Harney, Patterson, Reich, & Wakeley, 2021). Admixture proportions could therefore also be estimated for the historical screening data (79,382 SNPs).

QpAdm is considered a more powerful approach to estimating admixture proportions, requiring a set of source and reference populations to generate a matrix of F_4 statistics (Harney et al., 2021). QpAdm tests whether variance in the target population can be explained by the source populations with respect to the reference populations and estimates the admixture proportions of the target group. Outgroup species for this analysis included *Felis margarita* (n=2), *Felis bieti* (n=5) and *Felis lybica ornata* (n=4) from the domestic cat lineage (Johnson, Eizirik, Pecon-Slattery, Murphy, Antunes, Teeling, 2006) (Fig. 5.5). Genotype information for these samples was generated using the pipeline described in Chapter 4. Samples were provided by Carlos Driscoll (one each of *F. bieti, F. margarita* and *F. l. ornata*) and William Murphy (1 *F. bieti,* 2 *F. l. ornata*), the rest were sourced from public databases. The scenarios tested were informed by PCA and fineSTRUCTURE clustering (in terms of reference populations) and are summarised in Tables 5.3 and 5.4. An example is shown in Fig. 5.4.

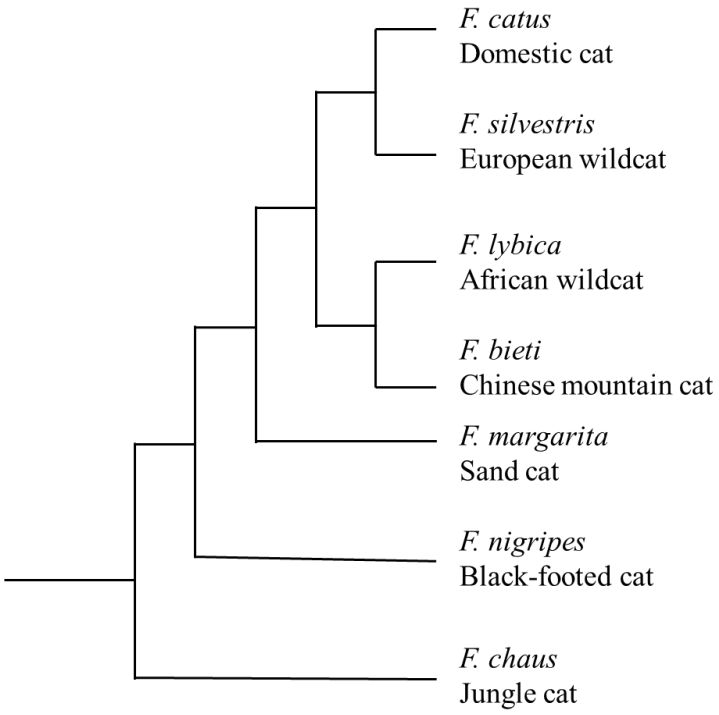

Figure 5.5. The domestic cat lineage, including domestic cats, *Felis catus*, and European wildcats, *Felis silvestris.* This phylogeny was used to select outgroup populations to calculate F_4 statistics.

Adapted from Johnson *et al.* (2006)

Table 5.3. F_4 ratio tests. The first test used modern populations of continental European wildcats (from western Germany and Portugal) to test for introgression in the historic samples from Scotland (whole genome sequence data only, collected pre-1940, n=4). Subsequent tests used the historic samples as a reference population to estimate the admixture proportions of the modern captive and wild-living individuals, as well as a larger sample of 20th century Scottish cats (historic screening data).

Scenario	X	A	B	C	O	Number of SNPs
1	20th c. Scotland (whole-genome data)	German wildcat (west)	Portuguese wildcat	Scottish domestic	*Felis margarita*	862,730
2	Scottish captive	Portuguese wildcat	20th c. Scotland (whole-genome data)			
3	Wild-living Scotland					
4	20th c. Scotland (screening data)					79,382

Table 5.4. Populations supplied to qpAdm. As per Table 5.3, the first test examined the historic samples (whole genome sequence data), which were then used as a reference population to test for admixture in the remaining Scottish individuals

Scenario	Target population	Source populations	Reference populations	Number of SNPs
1	20th c. Scotland (whole-genome data)	Scottish domestic, German wildcat (west)	*Felis margarita*, *Felis bieti*, *Felis lybica ornata*	862,730
2	Scottish captive	Scottish domestic, 20th c. Scotland (whole-genome data)		
3	Wild-living Scotland			
4	20th c. Scotland (screening data)			79,382

5.2.4 Effective population size

IBDNe was used to estimate recent effective population size (N_e) (Browning & Browning, 2015). Effective population size is the number of individuals in a theoretically idealised, randomly mating, population needed to explain genetic drift in the actual population (Wright, 1931). IBDNe estimates N_e using identity by descent (IBD) segments, i.e., regions of the genome shared by two individuals due to inheritance from a recent common ancestor (Browning & Browning, 2015). IBD can be informative about effective population size, for example, populations with a small N_e have higher IBD sharing as they are more closely related (on average) than populations with a large N_e. As with introgressed tracts (see 5.1.1), IBD segments are broken up by recombination: long IBD segments

indicate a short coalescence time and are informative about recent effective population size, short IBD segments are informative about older effective population size.

IBD segments were identified using Hap-IBD (Zhou, Browning, & Browning, 2020). Hap-IBD first estimates short segments that share identity by state (IBS), which are then extended into IBD segments. All modern individuals of putative Scottish wildcat ancestry were included in the analysis (i.e., all captive and wild-living individuals [n=36]), which used the thinned set of markers (n=1,011,786) described in 5.2.3. A genetic map with centi-Morgan (cM) units was generated using fineSTRUCTURE's convertrecfile.pl, which can extrapolate recombination rates, in this case from the modified domestic cat recombination maps (4.2.7), to estimate cM positions for all SNPs (Lawson et al., 2012).

223,718 IBD segments (identified by Hap-IBD) were used to infer recent effective population size. IBDNe (Browning & Browning, 2015) was run using a minimum IBD segment length of 2cM, for 1000 iterations, estimating N_e over the previous 300 generations. Confidence intervals were generated using 80 bootstrap samples.

5.2.5 Methods to date admixture

Two methods for dating admixture were applied to the wildcat dataset: PCAdmix (Brisbin et al., 2012) and MOSAIC (Salter-Townshend & Myers, 2019). Both aim to determine local ancestry across the genome and provide tract length information which can be used to infer admixture timing. The thinned set of SNPs described above (5.2.3) was applied to both methods (nSNPs=1,011,786, modern data only). Given the results from 5.2.3, all modern Scottish wildcats (wild and captive) were treated as admixed and included in the target population.

As described in 5.1.1, MOSAIC uses a number of reference panels to estimate local ancestry in the target individuals and construct a copying matrix relating the reference panels to the (unobserved) admixing groups (Salter-Townshend & Myers, 2019). MOSAIC was used to model two-way admixture, using three reference panels: Scottish domestic cats (n=5), non-Scottish domestic cats (n=17) and mainland European wildcats (n=7). All putative Scottish wildcat samples (captive and wild, n=36) constituted the admixed target population. MOSAIC uses R_{ST} as a measure of how useful the reference panels are to differentiate the ancestral mixing groups. R_{ST} varies between 0 and 2, a low R_{ST} indicates reference panels with a similar F_{ST} to both ancestral groups. Confidence in local ancestry estimates is summarised by r^2, the expected correlation between the local ancestry and (unobserved) true ancestry.

MOSAIC is designed to be robust to admixed panels (Salter-Townshend & Myers, 2019). To test this assumption a second MOSAIC analysis was run, including the captive Scottish wildcats in the 'wildcat' reference panel. The captive population contains valuable information about Scottish

wildcat haplotypes pertinent to the analysis. Despite evidence of introgression in these individuals (5.2.3), their inclusion in the reference panel did not appear to bias estimates of admixture timing (Fig. 5.15, Appendix 7). Reported F_{ST}, R_{ST} and r^2 values were 0.57, 0.62 and 0.97, respectively, comparable to those reported for analysis using only the continental wildcats in the reference panel (see 5.3.3). Only the MOSAIC results including the captive individuals in the target population are therefore presented in 5.3.3, as these have potentially important implications for conservation policy.

Coancestry curves were fitted for the target population as a whole, and for each sample individually (MOSAIC allows for the fact that individuals within a population may experience an admixture event at different points in their history) (Salter-Townshend & Myers, 2019). MOSAIC coancestry curves show the ratio of local ancestry probabilities for two positions separated by genetic distance, *d*, given the genome-wide average. Sampling date could then be accounted for, where possible, to give an estimated date of admixture per sample. Confidence intervals for the inferred population mean were generated using 100 bootstrap samples.

We compared MOSAIC inference to a window-based approach to identifying local ancestry, PCAdmix (Brisbin et al., 2012), using a method described by Johnson *et al.* (2011) to infer admixture timing using the number of ancestry chunks per individual as a summary statistic. PCAdmix is PCA-based method to infer the probable ancestry of pre-defined windows across the genome. It requires representative samples from the admixing groups to compute PCs, which hybrid individuals are projected onto. PC loadings are used to calculate an allele score per window, weighting SNPs which are informative about ancestry. The posterior probability of each putative ancestry is calculated, and a HMM used to smooth window calls and output the posterior probability of each ancestry per window, given data from the rest of the chromosome.

PCAdmix (Brisbin et al., 2012) was run per autosome using 22 domestic cats and seven mainland European wildcats as reference populations for PCA. All putative Scottish wildcat individuals (captive and wild) were projected onto these axes. PCAdmix employs several quality-control steps before PCA, filtering sites with low minor allele frequency (MAF), high missingness and high LD values. The wildcat dataset was previously filtered based on MAF and genotyping rate (MAF>2, lmiss=0); the default threshold for LD was used (r^2 >0.8). PCAdmix was run per chromosome, testing several window sizes of 5, 10, 20 (default), 40, 80 and 160 SNPs.

PCAdmix estimates overall ancestry proportion (wildcat and domestic), per haplotype, for each chromosome. Using this, the mean proportion of genome-wide wildcat cat ancestry was calculated per individual. Using only the windows with a high confidence in the inferred local ancestry (posterior probability > 95%) the total number of ancestry switches was summed across the genome for each individual. Number of generations since admixture (T) was estimated per individual

using the equation from Johnson *et al.* (2011), which relates admixture proportion (z), genome length (L, in cM) and number of local ancestry switches (B) in diploid genomes using

$$T = \frac{B}{(2 * 2 * 0.01) * L * z(1 - z)}$$

The domestic cat genome length (autosomes only) is estimated to be 4446.73cM (Li et al., 2016). This is an approximation based on a large Ne, but is expected to hold for recent admixture events. Admixture dates (in years) were estimated assuming a wildcat generation time of three years (Beaumont et al., 2001; Nussberger et al., 2018).

5.3 Results

5.3.1 Population structure

The results of PCA are shown in Fig. 5.6. As previously observed (Chapter 2), a large proportion of variance (24.7%) was captured by PC1, which separates wildcats and domestic cats. Wild-living (putative hybrid) cats from Scotland are found at intermediate positions across PC1, with captive Scottish wildcats clustered at one end of this continuum. Unlike previous analysis, this PCA includes wider geographic sampling, which is reflected in the clustering within parental groups. In domestic cats especially, a more global sample was available and PC2 appears to capture biogeographic structure in domestic cats across Eurasia from east to west (disregarding the individual from Madagascar, which represents the only sample from Africa). European and North American domestic cats cluster together, likely a result of the geopolitical history between these regions (and spread of domestic cats from Europe to the USA). Within European wildcats, biogeographic clustering is observed between Scotland, Germany, and Portugal (though this population has a sample size of one), which is better resolved by PC3 (3.1%). This axis appears to separate wildcat populations west to east, with the Scottish population appearing genetically closer to Portuguese wildcats than German wildcats. The German wildcats separate into two groups based on biogeographic clustering between eastern and western Germany.

Interestingly, the archaeological and historic samples from Britain cluster with wildcat populations from continental Europe across PC1. Both the archaeological and 20th century samples (collected between 1906 and 1938) form a tight cluster, distinct from modern captive individuals, and at a lower PC1 position.

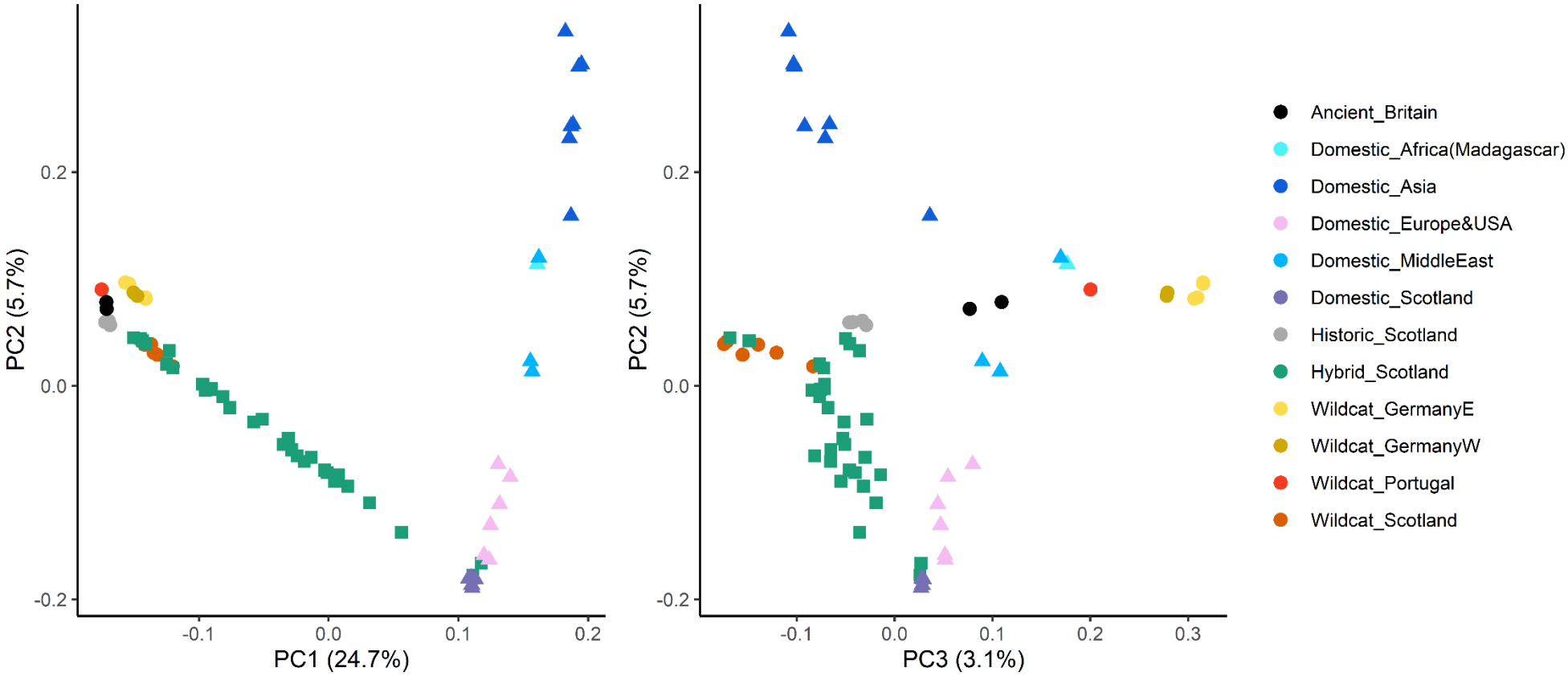

Figure 5.6. The first three principal components following PCA of the wildcat and domestic cat dataset. PC1 captures genetic differentiation between wildcats and domestic cats, with hybrids at intermediate positions between the parental groups. Ancient and historic samples of putative wildcats cluster with modern wildcats at the lower extreme of PC1. PC2 appears to capture geographic variation within domestic cats, and PC3 biogeographic clustering within wildcats.

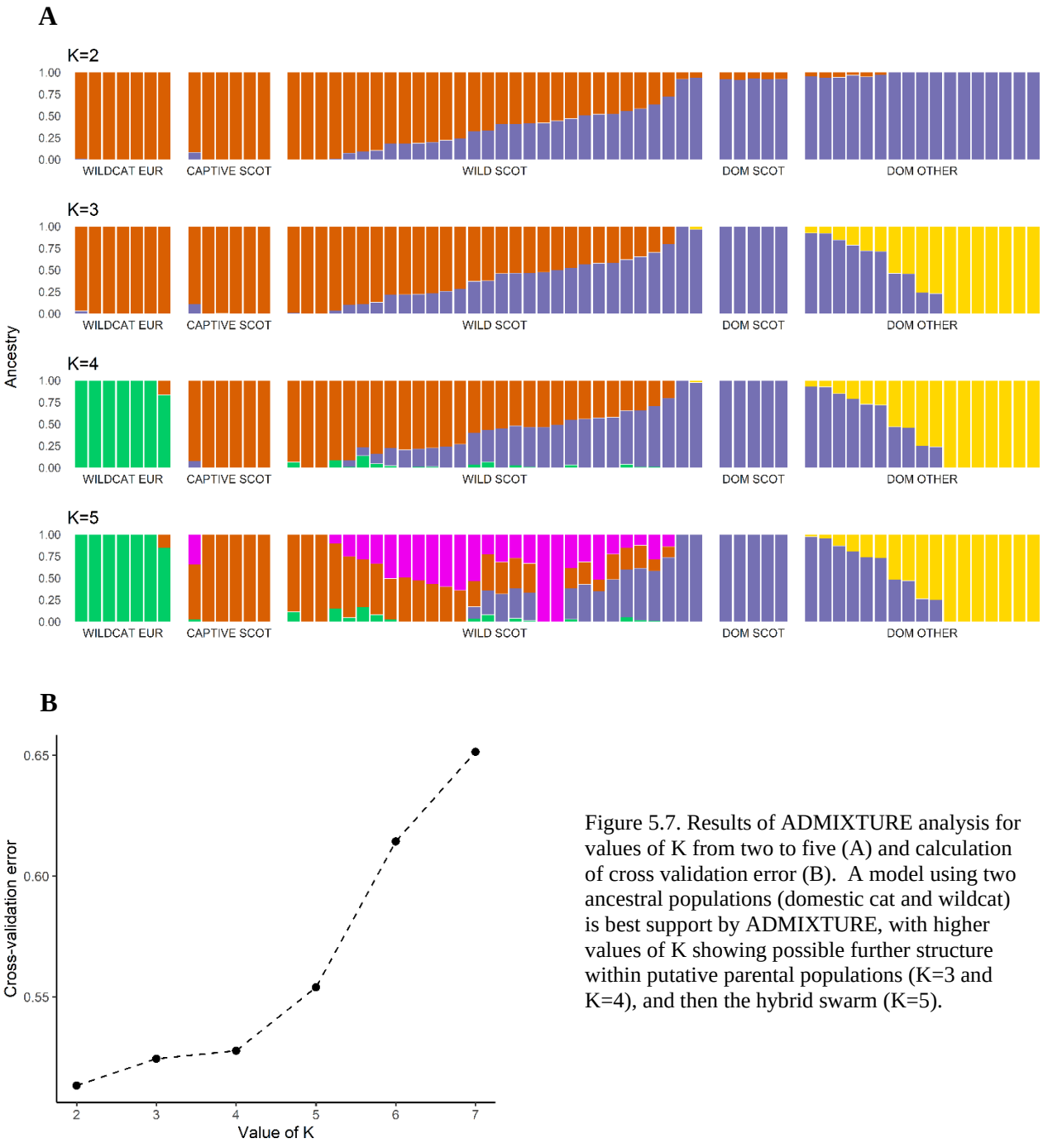

Figure 5.7. Results of ADMIXTURE analysis for values of K from two to five (A) and calculation of cross validation error (B). A model using two ancestral populations (domestic cat and wildcat) is best support by ADMIXTURE, with higher values of K showing possible further structure within putative parental populations (K=3 and K=4), and then the hybrid swarm (K=5).

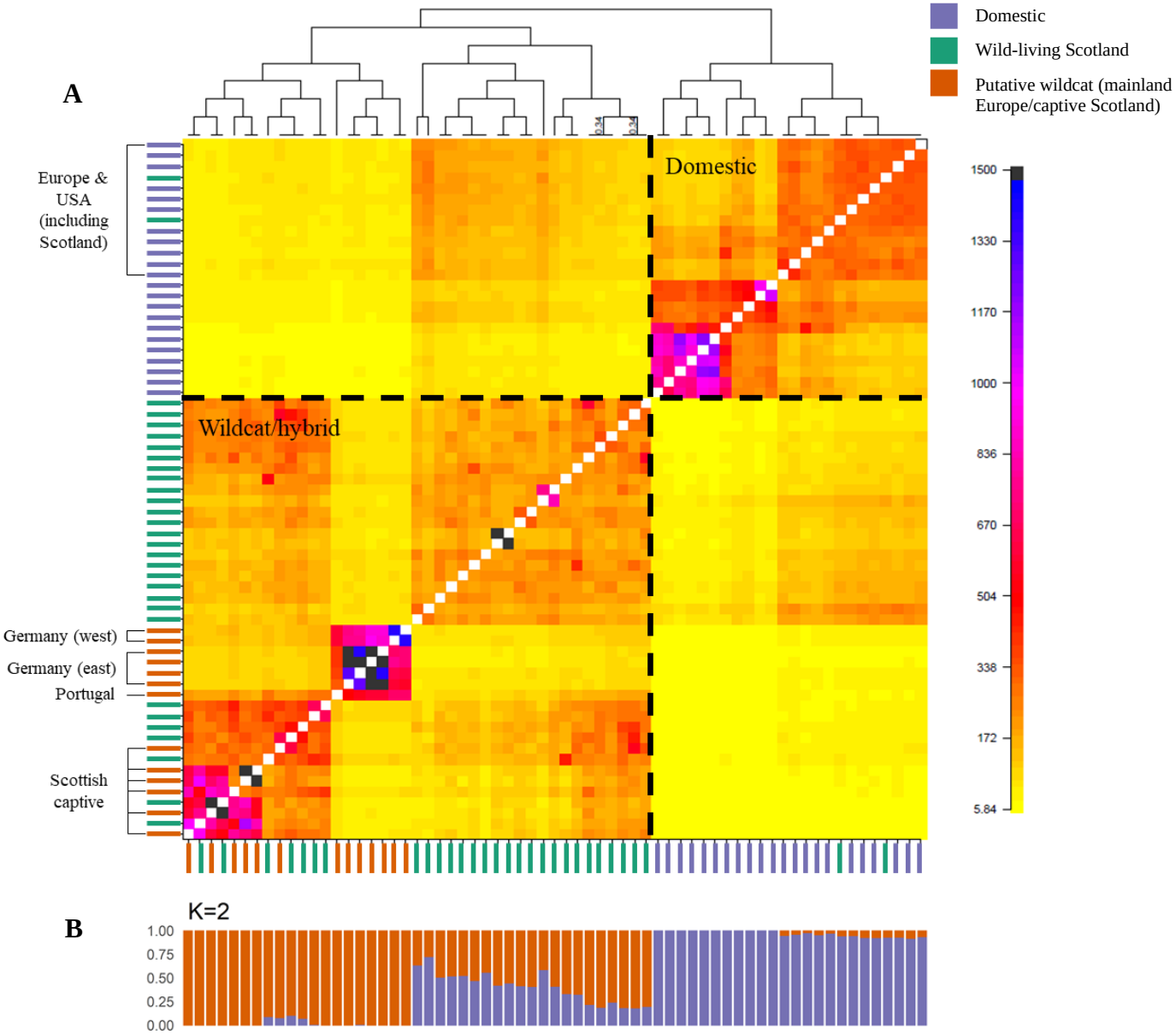

Figure 5.8. FineSTRUCTURE analysis (A) gives fine-scale detail about population structure, showing pairwise relationships between individuals. The heatmap shows the total length per donor used to reconstruct the genome of each individual. The source population for all individuals (domestic cat, Scottish wild-living, or putative wildcat [European wildcat/captive Scottish wildcat]) is shown by the coloured bar beside each row/column. 'Wild-living Scotland' comprises individuals sampled across the hybrid swarm, including some which cluster with the captive Scottish wildcats, as well as a number of feral domestic cats. Strong clustering is observed between the continental putative wildcats. An ADMIXTURE plot for K=2 is shown in (B), reordering individuals to match the fineSTRUCTURE heatmap, highlighting the broad clustering identified by both methods. Clustering methods (5.2.3) were used to confirm the sampling from distinct reference populations for admixture analyses.

ADMIXTURE (Fig. 5.7) and FineSTRUCTURE (Fig. 5.8) supported the general clustering shown on the PCA. The best supported value of K for ADMIXTURE analysis was 2 (Fig. 5.7B), separating putative representatives of the two parental species. Higher values of K provide greater resolution within these groups, first within domestic cats (K=3) and then wildcats (K=4). K=5 resulted in structure within the Scottish wild-living population. At all values of K shown in Fig. 5.7, only one captive Scottish wildcat appeared to have any domestic ancestry.

fineSTRUCTURE provided greater resolution, showing pairwise relationships between individuals (Fig. 5.8). It highlighted the distinction between continental European wildcats and individuals from Scotland. As observed on the PCA (Fig. 5.6), the captive population appeared to represent one extreme of the genetic continuum observed in the wild in Scotland. ADMIXTURE and fineSTRUCTURE were primarily used here to confirm reasonable sampling of the parental populations had taken place. Based on the observed clustering, reference panels for admixture analyses were chosen to consist of Scottish domestics/non-Scottish domestic cats and continental European wildcat.

5.3.2 Testing for introgression

Results from qpAdm are shown in Table 5.5. The historic samples appeared to show very little evidence of introgression from domestic cats. The captive population was estimated to be 18% introgressed. Estimated admixture proportions per individual are shown in full in Table 5.6 (Appendix 7). Overall, the historic samples (all collected pre-1960) appeared to have very little (to no) introgression from domestic cats (mean domestic admixture proportion, α=0.0). Modern samples (collected 1997-2018) show generally high levels of introgression (mean α=0.4, max α=0.8), with captive cats the least introgressed (mean α=0.2).

Table 5.5. qpAdm results. Both samples of historic individuals (whole-genome and screening data) showed minimal introgression from domestic cats, whilst the modern captive and wild-living population were shown to be introgressed.

Scenario	**Population**	**Proportion of ancestry (stderr)**	
		Domestic	**Wildcat**
1	20th c. Scotland (whole-genome data)	-0.006 (0.007)	1.01 (0.007)
2	Scottish captive	0.182 (0.009)	0.818 (0.009)
3	Wild-living Scotland	0.47 (0.004)	0.53 (0.004)
4	20th c. Scotland (screening data)	-0.009 (0.004)	1.01 (0.004)

5.3.3 Haplotype methods to date admixture

An estimate of Scottish wildcat effective population size is shown in Fig. 5.9. Effective population size appeared to increase by an order of magnitude ten generations before present and increased a further three orders of magnitude over those ten generations. The sudden increase in effective population size (and widening of confidence intervals) could be interpreted as a signal of putative admixture, a result of violating IBDNe's assumption that samples are from a single, unadmixed population. Taking 'present-day' to be the most recent sampling date, and wildcat generation time to be three years, this dates admixture in Scottish wildcats to the mid-1980s. (Upper 95% CI truncated on Fig. 5.9 for better visualisation, for unmodified plot see Fig. 5.16, Appendix 7).

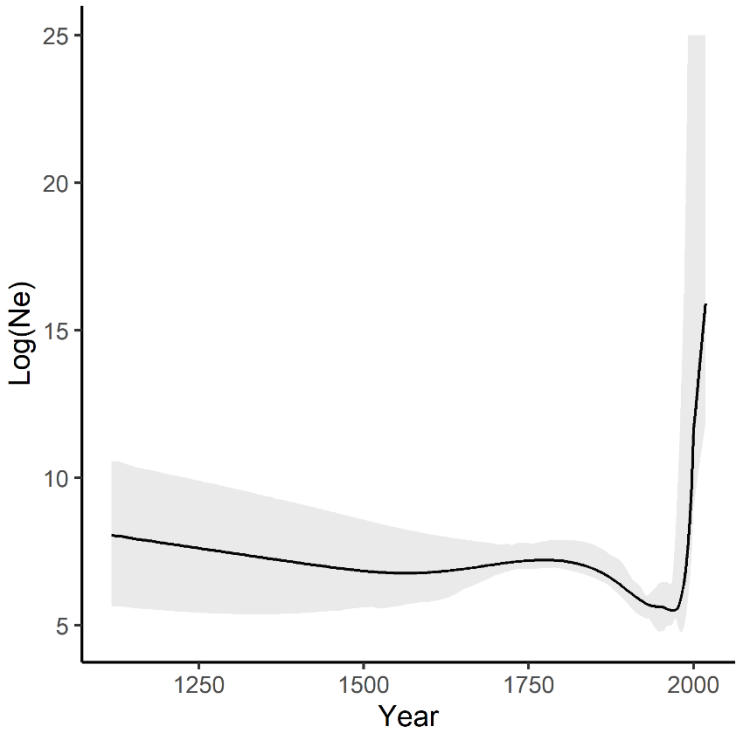

Figure 5.9. Estimate of effective population size (N_e) over the past 300 generations (~900 years). The dramatic increase in N_e in the late 20th century may be an artefact of admixture, violating the assumptions of IBDNe that individuals are sampled from a single, homogenous population. 95% confidence intervals are shown in grey, with the upper log(95% CI) capped at 25 to improve visualisation.

355,327 SNPs were used as input for PCAdmix, following LD thinning and removal of SNPs monomorphic in all ancestral individuals. Estimated admixture time appeared to be highly dependent on window size (Fig. 5.10 and Fig. 5.11), and reliable estimates could therefore not be obtained using this method. Number of generations since admixture varied between 0 and 67 across all window sizes shown in Fig. 5.10, with a large amount of individual variation. Intuitively, larger window sizes give less detailed information about local ancestry patterns across the genome; introgressed tracts appear artificially long and fewer ancestry switches reported. This may explain the especially poor fit

observed in Fig. 5.10 at larger window sizes, where individuals with variable proportions of putative wildcat ancestry have a similar number of reported local ancestry switches. Instead of decreasing window size converging on well-defined blocks of local ancestry, however, switch number continued to increase (Fig. 5.11).

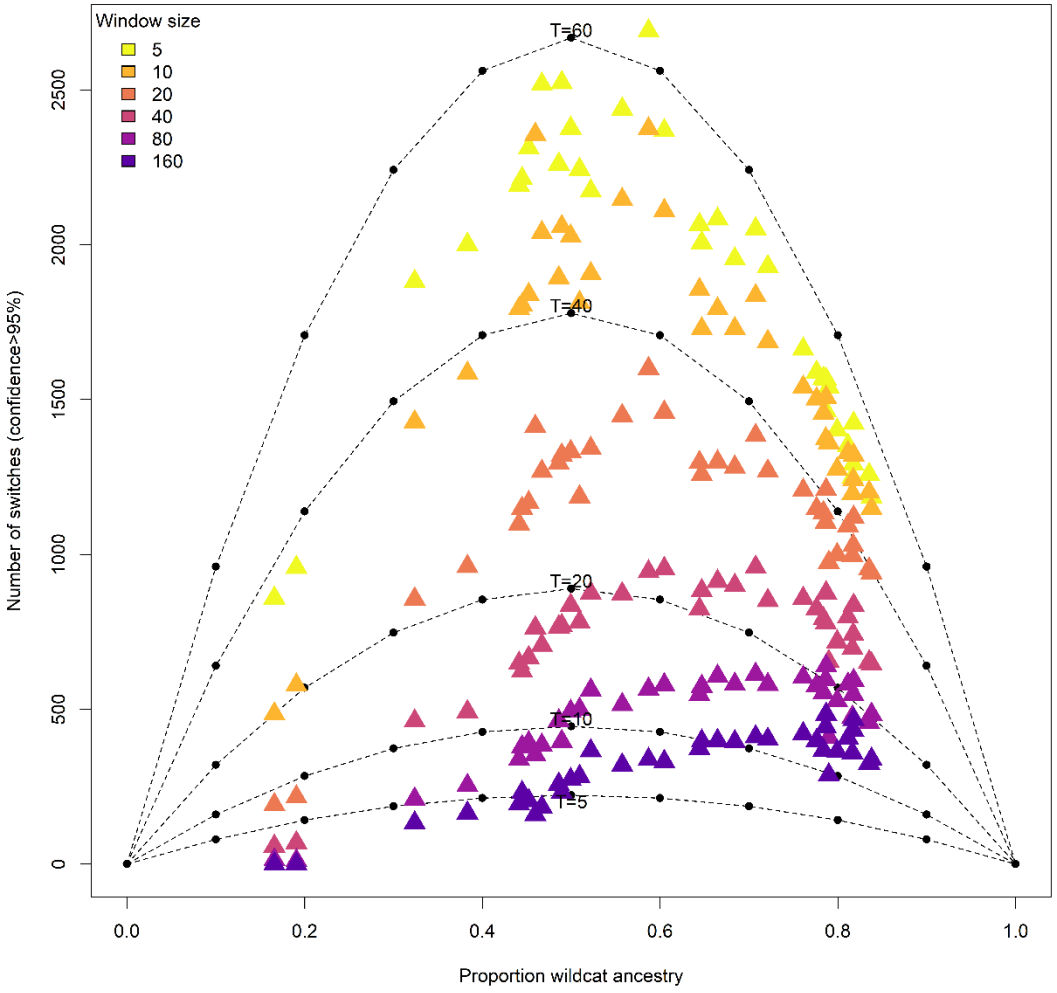

Figure 5.10. Inferred admixture time (in generations) per individual, using the proportion of genome-wide wildcat ancestry and number of local ancestry switches estimated by PCAdmix (Johnson et al. 2011). Dashed lines indicate the expected distributions for admixture times between five and 60 generations ago. Results are shown per individual for each run of PCAdmix, coloured by window size (in number of SNPs). Number of ancestry switches, and therefore admixture time, seemed to depend on the window size selected.

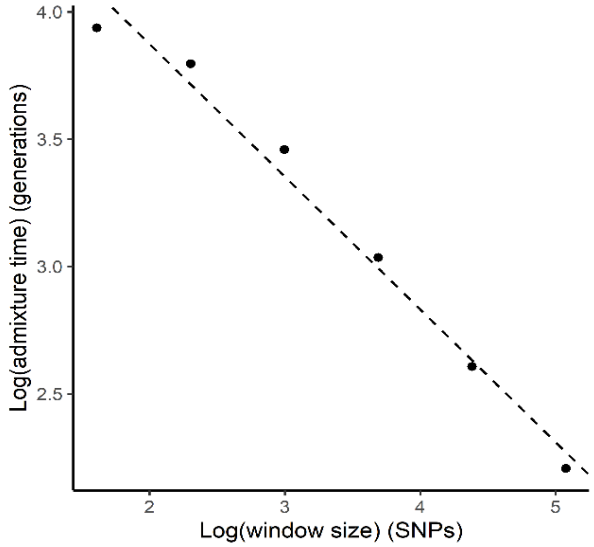

Figure 5.11. Relationship between window size (number of SNPs) and mean admixture time (in generations). PCAdmix does not converge on a mean admixture time for the hybrid population, instead, as window size decreases, admixture appears increasingly older as local ancestry blocks become progressively smaller.

MOSAIC modelling resulted in a high expected r^2 (0.96) and moderately high R_{ST} between the mixing groups (0.74). These measures indicated accurate local ancestry estimation based on informative reference panels. One ancestry appeared to copy preferentially from the domestic cat populations, and the other from the European wildcats. F_{ST} between the ancestral mixing groups was estimated to be 0.66. Pairwise F_{ST} estimates between the reference panels and inferred ancestral populations, as well as the copying matrix are shown in Table 5.7 and Fig. 5.17 of Appendix 7.

Coancestry curves fitted by MOSAIC are shown for all possible pairs of local ancestries (i.e., wildcat:wildcat, wildcat:domestic, domestic:domestic) (Fig. 5.12). Observed patterns of exponential decay appeared to be consistent with a single admixture event (Hellenthal et al., 2014). Using the mean estimate across the three curves, the admixture event in Scottish wildcats was dated to was 8.6 (95% CI 8.3-9.8) generations before present. Estimates per individual are given in Table 5.8 (Appendix 7), and the distribution of corresponding times (in years, accounting for sampling date) are shown in Fig. 5.13. Excluding two probable feral domestic cats, the oldest admixture event reported by MOSAIC occurred 17.9 generations before present (~53.8 years ago). No admixture was detected in the current sample prior to 1957, with the majority of admixture events occurring between 1970 and 1995.

Karyograms allow clear visualisation of local ancestry across the genome for each target individual. Some example plots are given in Fig. 5.18, Appendix 7. Estimated proportions of genome-wide ancestry are given in Table 5.6, Appendix 7.

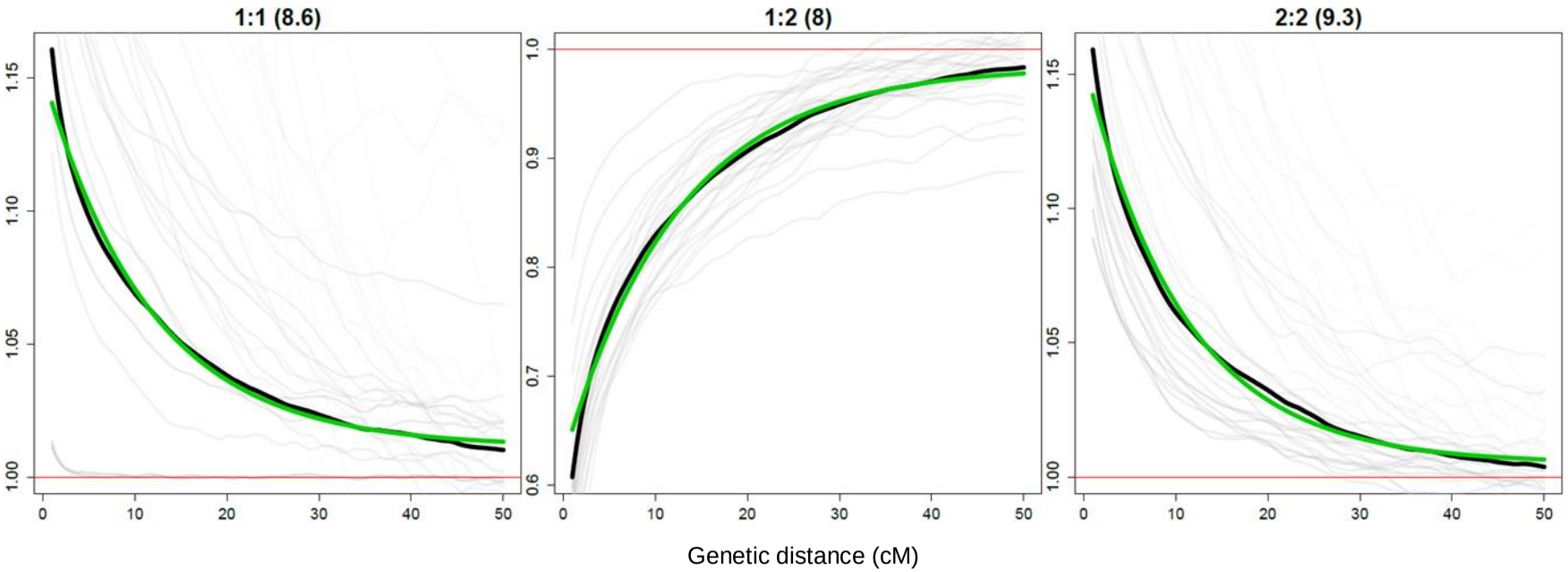

Figure 5.12. Coancestry curves generated by MOSAIC. The three plots show relative probability of pairs of local ancestries, wildcat:wildcat (1:1), wildcat:domestic (1:2) and domestic:domestic (2:2) over genetic distance. The grey lines show the curve per individual, the black line represents the population mean and the green line the fitted decay curve, the rate parameter of which is equal to the number of generations since admixture (estimated for the population as a whole). This number is given in brackets at the top of each plot, with a mean estimate of 8.6 generations for the sample of Scottish wildcats.

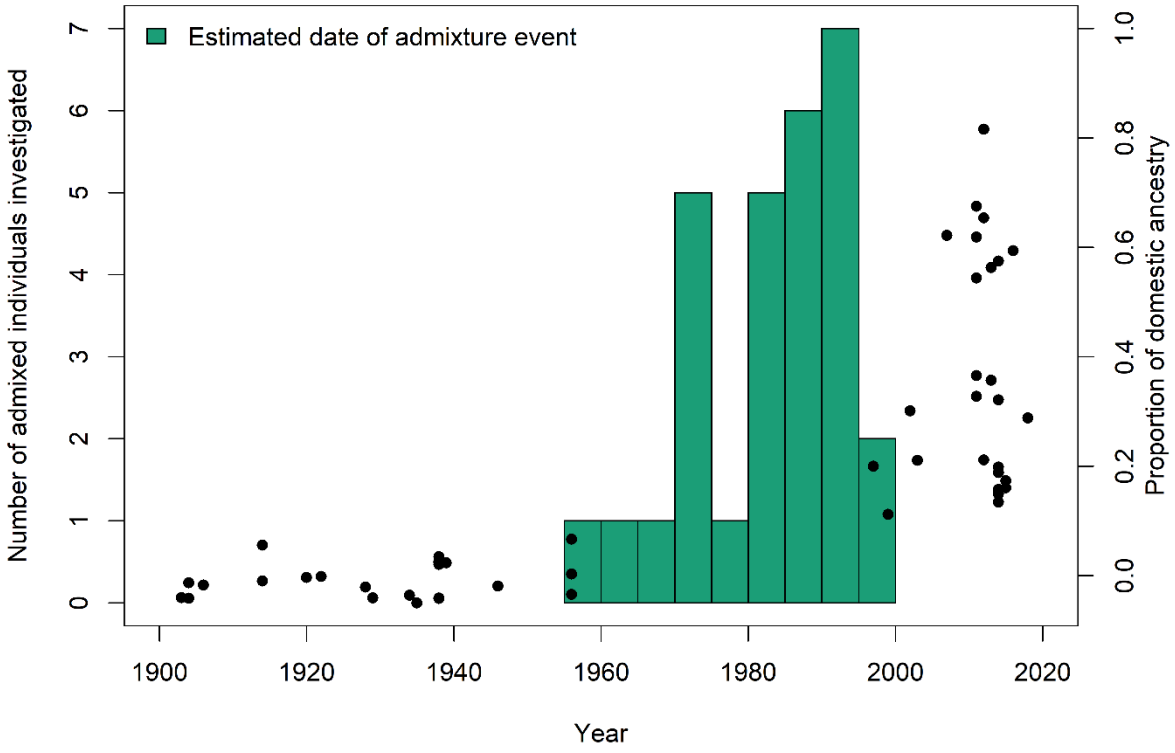

Figure 5.13. Predicted pattern of 20th century admixture in Scottish wildcats. Shown here is the distribution of estimated admixture times (per individual) from MOSAIC (green). Superimposed are the predicted proportions of wildcat ancestry in the historic screening data and modern Scottish samples (black points). Mosaic did not recover signals of admixture pre-dating 1957, historic screening data (sampled 1903-1956) showed little introgression from domestic cats.

5.4 Discussion

5.4.1 Inferring admixture history in the Scottish wildcat population

Results from this study support a recent date for the onset of hybridisation in Scotland. Reference samples from additional populations of European wildcats, as well as historic and ancient samples from Britain, highlight introgression in modern Scottish wildcats. Ancient and historic samples cluster with modern wildcats from continental Europe across PC1 and PC2 (Fig. 5.6), despite broad sampling of individuals in space and time. Importantly, this cluster includes all of the historic wildcat samples from early 20th-century Scotland (1906-1939). Using AdmixTools to quantify introgression (Patterson et al., 2012), low hybridisation rates were reported in the historic samples, including the historic screening data, all sampled pre-1956 (Table 5.5; Table 5.6, Appendix 7). All of the modern individuals sampled (1997-2018) showed evidence of introgression from domestic cats.

These results suggest hybridisation events in the early part of the 20th century were rare, supporting previous analysis of museum samples by Beaumont *et al.* (2001) (13 samples dating from 1945) and Senn *et al.* (2019) (60 samples dated 1895-1985) that reported high levels of putative

wildcat ancestry in historic samples. There is an obvious sampling gap in the present study spanning four decades between 1956 and 1997. Senn *et al.* (2018) find individuals with a high proportion of putative wildcat ancestry (evaluated using the 35 SNP test) into the 1980s, Beaumont *et al.* (2001) note a decline in putative wildcat ancestry (based on nine microsatellite markers) post-1970.

Here, ancient samples were also available to examine historic patterns of hybridisation in Scotland. Archaeological samples from mediaeval and Mesolithic sites provided a useful baseline to evaluate hybridisation in historic samples, as well as to examine less recent patterns of admixture. Genetic clustering of ancient and historic samples (Fig. 5.6) indicated 20th century Scottish wildcats were genetically similar to archaeological samples from Britain. It seems unlikely that historic and ancient populations, as well as modern populations of European wildcats (which cluster similarly), would have all experienced the same pattern of introgression from domestic cats. We therefore conclude that these samples represent (mostly) unadmixed wildcat individuals. Distinct clustering of 20th century and archaeological samples across PC3 (Fig. 5.6) may reflect the population decline and range contraction observed in Britain between 1600 and 1900. As noted by Senn *et al.* (2018) there is a bias towards wildcat phenotypes in the 20th century specimens, as these are samples selected to be incorporated into museum collections. These results, therefore, may not be inconsistent with post-WWI expansion of Scottish wildcats and subsequent introgression.

IBDNe uses haplotype information to estimate effective population size through time. However, here we interpret loss of signal as an indicator of putative admixture. IBDNe assumes individuals are sampled from a homogenous, non-admixed population; admixed individuals share IBD segments with recent ancestors from two distinct, diverged, populations (Browning & Browning, 2015). IBDNe appears to be unable to accurately infer effective population size in Scottish wildcats from approximately ten generations before present, corresponding to a putative pulse admixture event, or continuous admixture, from the mid-1980s (Fig. 5.9).

Results from MOSAIC, a haplotype-based approach developed to accurately infer admixture times (Salter-Townshend & Myers, 2019), appeared to corroborate these results. Based on the analysis of 36 modern individuals from Scotland, the mean estimate for the onset of hybridisation was 8.6 generations (or 25.8 years) before present. The distribution of dates from individual coancestry curves ranged between 17.9 and 4.7 generations. Accounting for sampling date, this corresponds to a period between the late 1950s and mid-1990s (Fig. 5.13). A more detailed evaluation of MOSAIC, and its application to studying admixture in the Scottish wildcat population, is given below (5.4.3). This method represents the most powerful analysis of wildcat hybridisation in Scotland to date, and the first application of haplotype-based methods to infer the population's admixture history.

All evidence presented here is consistent with recent onset of significant hybridisation in Scotland and the rapid formation of the 'hybrid swarm'. These analyses strongly support Scottish

wildcat hybridisation as a modern phenomenon, occurring from the mid-20th century onwards and increasing into 1980s and 1990s. As discussed briefly in Chapter 3, extensive hybridisation, and rapid emergence of hybrid swarms, has been observed between other wild populations and domesticated or invasive species. This includes red deer and introduced Japanese sika deer in Scotland (Senn & Pemberton, 2009) and Ireland (Smith et al., 2014), both believed to have occurred during the 20th century, and within the last 50 years in the Scottish Loch Awe population. Hybridisation between native and introduced fish species is common. This has rapidly (within decades) led to hybrid swarms in, for example, cut-throat trout and rainbow trout (Muhlfeld et al., 2014), and blacktail shiners and red shiners (Walters et al., 2008). A well-studied example of hybridisation between wild and domestic species is between wolves and dogs. Hybridisation rate is variable across the range of grey wolves in Eurasia, with evidence of ancient hybridisation events (Pilot et al., 2018). As observed for European wildcats and domestic cats, however, modern populations of the two species remain distinct, suggesting ancient admixture was likely to be rare. In the Italian population of grey wolves, a similar pattern of 20th century admixture has been reported, with the earliest admixture dated to the 1940s, and the onset of significant admixture in the 1980s, peaking in the 1990s (Galaverni et al., 2017). Galaverni *et al.* (2017) attribute this to wolf population expansion during this period, with populations on the leading edge at lower density and therefore a reduced frequency of available wolf mates. This is reflected in the spatial pattern of hybridisation, expanding from refugia in central and southern Italy. A decreasing rate of hybridisation post-2000 is speculated to be the result of subsequent establishment of wolf populations in the new range.

Given the sympatry of domestic cats and wildcats in Britain for over two thousand years (Jamieson et al, in prep.), and comparatively low levels of hybridisation across most of continental Europe (Nussberger et al., 2018; Tiesmeyer et al., 2020), it is evident that isolating mechanisms between the two species exist which have completely broken down in Scotland. The drivers of hybridisation remain poorly understood, but will be key to restoring this species in Scotland and managing the impacts of introgression across the rest of the species range.

5.4.2 Implications for the captive breeding population

Previous analyses (Chapter 2), using individuals sampled exclusively from Scotland, could not rule out historic introgression in the captive population, which could only be described as the 'most genetically distant' to domestic cats. Based on the results of this study, captive wildcats in Scotland appeared to be ~18% introgressed (Table 5.5). This highlights the importance of additional reference samples to accurately quantify hybridisation, and the value of a baseline for future conservation management. It has been valuable to compare hybridisation rates in Scotland to those from modern populations representing three of the biogeographic groups in Europe, as well as representatives of the historic Scottish wildcat population. In agreement with ddRAD-seq analysis

(Chapter 2), the captive population appears to be the most genetically distant from domestic cats in modern Scotland (interpreting PC1, Fig. 5.6, as an axis separating the two species). However, clustering of reference samples show captive wildcats are found at one end of the 'hybrid swarm' continuum. Given that the onset of significant hybridisation in Scotland (late 1950s) appears to have coincided with the establishment of the captive population (1960), it is perhaps unsurprising that hybrids have been incorporated over the population's history, especially as accurate tests for hybrids were only developed in the 21st century (Kitchener et al., 2005; Nussberger et al., 2013; Senn & Ogden, 2015).

Given the introgression level in captivity, and inbreeding risk associated with maintaining captive populations (Frankham et al. 2002), it seems reasonable to consider genetic augmentation from populations of continental wildcats as a conservation management strategy. IUCN guidelines for species reintroduction/reinforcement state that source populations should be genetically close to the native population, where possible (IUCN/SSC, 2013). Mattucci *et al.* (2016) found biogeographic structure in European wildcats corresponded to expansion from glacial refugia on the Iberian, Italian and Balkan peninsulas. Unfortunately, high rates of hybridisation in Scotland precluded its involvement in this study due to the high proportion of shared introgressed variation in Scottish individuals. The relationship between Scotland and genetic clusters observed within continental Europe is yet to be resolved. Patterns of expansion from Mediterranean refugia have been described for many European species, including brown bears (*Ursus arctos*), hedgehogs (*Erinaceus* spp.), and oaks (*Quercus* spp.) (Hewitt, 2001). However, multiple routes of colonisation into Britain have been reported, e.g., in pygmy shrews (*Sorex minutus*) (Vega et al., 2010), from north-eastern Europe, the common frog (*Rana temporaria*) (Teacher, Garner, & Nichols, 2009), from western Europe, and water voles (*Arvicola terrestris*), where English and Welsh mitochondrial haplotypes have been linked to refugia in eastern Europe and Scottish haplotypes from Iberian refugia (Piertney et al., 2005). Results presented here tentatively support the association between Scotland and Iberian refugia in wildcats; PCA clustering indicated historic populations in Scotland most closely resemble the Portuguese sample (Fig. 5.6). Sampling of biogeographic groups in this study, however, has not been exhaustive, and sample sizes from represented groups are small. Further sampling will be needed to confidently suggest source populations for Scottish wildcat reinforcement. This work would be supported by local ancestry assignment, e.g., from MOSAIC, to remove tracts of domestic cat ancestry (Fig. 5.18, Appendix 7) for improved analysis of population structure. Also, given the period of separation between British and continental wildcat populations, a full assessment of the risk of outbreeding depression should be carried out (Frankham et al., 2011).

5.4.3 Methods to date admixture

MOSAIC (Salter-Townshend & Myers, 2019) appeared to be able to infer useful information about the ancestral Scottish wildcat population from the reference and target panels supplied. Haplotype copying under each putative ancestry preferentially used either the domestic or the wildcat reference panels (Fig. 5.17, Appendix 7). A strong negative correlation between the copying matrix and pairwise F_{ST} (with the reconstructed ancestral haplotypes) was observed for both domestic panels, indicating these were good surrogates for the admixing domestic group. Modern European wildcat haplotypes were copied almost exclusively under putative Scottish wildcat local ancestry, and only 5% of the time under putative domestic ancestry, supporting these individuals as unadmixed. Moderately high F_{ST} values were reported between the modern European wildcats and both putative ancestral groups (0.506 and 0.340 for ancestral domestic and wildcat populations, respectively, Table 5.7, Appendix 7). This may highlight potential drift of the ancestral Scottish population from continental wildcats (unsurprising, given the isolation of British wildcats for ~10,000 years [Yalden, 1982]), and/or indicates this panel was a poor surrogate for the ancestral Scottish wildcats.

MOSAIC analysis performs well for complex admixture scenarios, in human populations, up to 100 generations before present (Salter-Townshend & Myers, 2019). Applying this tool to non-human populations for the first time generated confident estimates of local ancestry (Fig 5.18, Appendix 7) and admixture dating (bootstrapping approach), despite a limited number of panels with small sample sizes. This may be a result of the high divergence between wildcats and domestic cats (which constitute two separate species), and relatively simple two-way admixture scenario. Nonetheless, this analysis could potentially be improved by enlarging panel sizes for all reference populations, sampling more of the domestic cat and wildcat haplotype diversity. This could include the addition of historic samples from Scotland (sequenced at higher coverage), and the modern captive population. MOSAIC appeared to be robust to admixed panels; the captive population, representing the least introgressed individuals in modern-day Scotland, is a useful reservoir of Scottish wildcat haplotypes for MOSAIC analysis.

There was inconsistency between the dates obtained using MOSAIC (8.6 [8.3-9.8] generations before present), and ABC modelling of unlinked loci (3.3 [1.2-5.6], see Chapter 3). However, both support a recent onset of admixture, despite the application of very different datasets and methodologies. Future work to extend the ABC model to exploit sequence data would be valuable.

PCAdmix appeared to be unable to accurately infer local ancestry in the Scottish wildcat population. Unlike MOSAIC, PCAdmix assigns local ancestry to pre-defined windows spanning *x* number of SNPs. However, despite testing a range of window-sizes, PCAdmix did not converge on local ancestry tracts, and therefore expected admixture date (Figs. 5.10, 5.11). PCAdmix has been

applied to other admixing population, including dogs and wolves (Galaverni et al., 2017) and wildcats and domestic cats (Mattucci et al., 2019). Galaverni *et al.* (2017) found concordance in admixture date estimates between PCAdmix and ALDER (PCAdmix was run using the default window size, with no additional window sizes tested). Both studies used a larger number of reference samples for PCA (in the order of 100s). PCA projection is also sensitive to uneven sample samples (McVean 2009). It seems likely that the small (and uneven) reference sample sizes used here may have influenced projection of the hybrid samples, and therefore local ancestry estimates. In a study of dog-wolf hybridisation by Smeds *et al.* (2021), both the number of reference individuals and the number of markers used for analysis was shown to impact PCAdmix inference. Estimated proportion of dog ancestry appeared to converge using 20 reference individuals per population or more. Further work, including sampling a larger number of reference individuals, is needed to calibrate PCAdmix for use in the Scottish wildcat population.

Haplotype information has been applied to date admixture in wild populations of diverse taxa, including mammals (Galaverni et al., 2017; Mattucci et al., 2019), insects (Nelson, Wallberg, Simões, Lawson, & Webster, 2017), plants (Duranton, Bonhomme, & Gagnaire, 2019) and fish (Leitwein, Gagnaire, Desmarais, Berrebi, & Guinand, 2018). Admixture analysis of non-human populations has previously been limited by the availability of reference data, and technical aspects such as generating chromosome-level assemblies, recombination maps and accurate phasing. Increased marker density and accurate phasing are especially important to identify short tracts (and therefore signatures of ancient admixture). Smeds *et al.* (2021) identified ancestry switch errors using a known F_1 hybrid to calibrate local ancestry estimates, and consequently were not able to estimate admixture timing with any accuracy. Early methods to infer local ancestry required prior knowledge of admixture parameters, such as admixture time or proportion. The development of methods, such as MOSAIC, which simultaneously infer local ancestry and estimate admixture parameters are an important contribution to the study of admixture in wild populations.

5.5 Conclusion

Haplotype information is a powerful resource to characterise hybridisation dynamics in the Scottish wildcat population. Using whole-genome sequence data from 65 modern individuals we were able to identify admixture dating back to the mid-1950s. Hybridisation between wildcats and domestic cats appeared to accelerate during the end of the 20th century. Individuals sampled prior to 1956, including two low-coverage samples from mediaeval and Mesolithic archaeological sites, appeared to show no introgression from domestic cats, suggesting historic admixture events were rare. The 'hybrid swarm' appears to be a modern phenomenon following rapid demographic collapse of wildcats in Scotland.

5.6 References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Research*, 19(9), 1655–1664

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., … Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, 28(10), 1359–1367

Beaumont, M., Barratt, E. M., Gottelli, D., Kitchener, A. C., Daniels, M. J., Pritchard, J. K., & Bruford, M. W. (2001). Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology*, 10(2), 319–336

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., … Bustamante, C. D. (2012). Pcadmix: Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human Biology*, 84(4), 343–364

Browning, S. R., & Browning, B. L. (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *American Journal of Human Genetics*, *97*(3), 404–418

Delaneau, O., Zagury, J. (2012) Haplotype Inference. In: Pompanon, F., Bonin, A., (Eds.) *Data Production and Analysis in Population Genomics* (pp. 177-196). Totowa, USA: Humana Press

Duranton, M., Bonhomme, F., & Gagnaire, P. A. (2019). The spatial scale of dispersal revealed by admixture tracts. *Evolutionary Applications*, 12(9), 1743–1756

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics*, 164, 1567–1587

Frankham, R., Ballou, S. E. J. D., Briscoe, D. A., & Ballou, J. D. (2002). *Introduction to conservation genetics*. Cambridge: Cambridge University Press

Frankham, R., Ballou, J. D., Eldridge, M. D. B., Lacy, R. C., Ralls, K., Dudash, M. R., & Fenster, C. B. (2011). Predicting the Probability of Outbreeding Depression. *Conservation Biology*, 25(3), 465–475

Galaverni, M., Caniglia, R., Pagani, L., Fabbri, E., Boattini, A., & Randi, E. (2017). Disentangling timing of admixture, patterns of introgression, and phenotypic indicators in a hybridizing Wolf population. *Molecular Biology and Evolution*, 34(9), 2324–2339

Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, 191(2), 607–619

Harney, É., Patterson, N., Reich, D., & Wakeley, J. (2021). Assessing the performance of qpAdm: A statistical tool for studying population admixture. *Genetics*, 217(4), iyaa045

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, 343, 747–751

Hewitt, G. M. (2001). Speciation, hybrid zones and phylogeography - Or seeing genes in space and time. *Molecular Ecology*, 10(3), 537–549

IUCN/SSC. (2013). *Guidelines for Reintroductions and Other Conservation Translocations IUCN. Gland, Switzerland: IUCN Species Survival Commission* (Vol. viiii).

Jamieson, A., (2021) The evolutionary history of domestic cats in Europe. Manuscript in preparation.

Johnson, N. A., Coram, M. A., Shriver, M. D., Romieu, I., Barsh, G. S., London, S. J., & Tang, H. (2011). Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genetics*, 7(12), e1002410

Kitchener, A. C., Yamaguchi, N., Ward, J. M., & Macdonald, D. W. (2005). A diagnosis for the Scottish wildcat (*Felis silvestris*): A tool for conservation action for a critically-endangered felid. *Animal Conservation*, 8(3), 223–237

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1), 11–17

Leitwein, M., Gagnaire, P. A., Desmarais, E., Berrebi, P., & Guinand, B. (2018). Genomic consequences of a recent three-way admixture in supplemented wild brown trout populations revealed by local ancestry

tracts. *Molecular Ecology*, 27(17), 3466–3483

Li, G., Hillier, L. D. W., Grahn, R. A., Zimin, A. V., David, V. A., Menotti-Raymond, M., … Murphy, W. J. (2016). A high-resolution SNP array-based linkage map anchors a new domestic cat draft genome assembly and provides detailed patterns of recombination. *G3: Genes, Genomes, Genetics*, 6(6), 1607–1616

Li, N., & Stephens, M. (2003). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165(4), 2213–2233

Loh, P. R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., & Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193(4), 1233–1254

Mattucci, F., Galaverni, M., Lyons, L. A., Alves, P. C., Randi, E., Velli, E., … Caniglia, R. (2019). Genomic approaches to identify hybrids and estimate admixture times in European wildcat populations. *Scientific Reports*, 9, 11612

Moorjani, P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L., Atzmon, G., … Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genetics*, 7(4)

Muhlfeld, C. C., Kovach, R. P., Jones, L. A., Al-Chokhachy, R., Boyer, M. C., Leary, R. F., … Allendorf, F. W. (2014). Invasive hybridization in a threatened species is accelerated by climate change. *Nature Climate Change*, 4(7), 620–624

Nelson, R. M., Wallberg, A., Simões, Z. L. P., Lawson, D. J., & Webster, M. T. (2017). Genomewide analysis of admixture and adaptation in the Africanized honeybee. *Molecular Ecology*, 26(14), 3603–3617

Nussberger, B., Currat, M., Quilodran, C. S., Ponta, N., & Keller, L. F. (2018). Range expansion as an explanation for introgression in European wildcats. *Biological Conservation*, 218(2018), 49–56

Nussberger, B., Greminger, M. P., Grossen, C., Keller, L. F., & Wandeler, P. (2013). Development of SNP markers identifying European wildcats, domestic cats, and their admixed progeny. *Molecular Ecology Resources*, 13(3), 447–460

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., … Reich, D. (2012). Ancient admixture in human history. *Genetics*, 192(3), 1065–1093

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12), 2074–2093

Petr, M. (2020). admixr: An Interface for Running 'ADMIXTOOLS' Analyses. R package version 0.9.1. Available at https://CRAN.R-project.org/package=admixr

Piertney, S. B., Stewart, W. A., Lambin, X., Telfer, S., Aars, J., & Dallas, J. F. (2005). Phylogeographic structure and postglacial evolutionary history of water voles (*Arvicola terrestris*) in the United Kingdom. *Molecular Ecology*, 14(5), 1435–1444

Pilot, M., Greco, C., VonHoldt, B. M., Randi, E., Jędrzejewski, W., Sidorovich, V. E., … Wayne, R. K. (2018). Widespread, long-term admixture between grey wolves and domestic dogs across Eurasia and its implications for the conservation status of hybrids. *Evolutionary Applications*, 11, 662–680

Pool, J. E., & Nielsen, R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181(2), 711–719

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 904–909

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., … Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6), e1000519

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12), 1791–1798

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., … Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, 411, 199–204

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian Population History. *Nature*, 461(7263), 489–494

Salter-Townshend, M., & Myers, S. (2019). Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics*, 212, 869–889

Senn, H., & Ogden, R. (2015). *Wildcat hybrid scoring for conservation breeding under the Scottish Wildcat Conservation Action Plan*. Royal Zoological Society of Scotland.

Senn, H. V., Ghazali, M., Kaden, J., Barclay, D., Harrower, B., Campbell, R. D., … Kitchener, A. C. (2019). Distinguishing the victim from the threat: SNP-based methods reveal the extent of introgressive hybridization between wildcats and domestic cats in Scotland and inform future in situ and ex situ management options for species restoration. *Evolutionary Applications*, 12(3), 399–414

Senn, H. V., & Pemberton, J. M. (2009). Variable extent of hybridization between invasive sika (*Cervus nippon*) and native red deer (*C. elaphus*) in a small geographical area. *Molecular Ecology*, 18(5), 862–876

Smeds, L., Aspi, J., Berglund, J., Kojola, I., Tirronen, K., & Ellegren, H. (2021). Whole-genome analyses provide no evidence for dog introgression in Fennoscandian wolf populations. *Evolutionary Applications*, 14(3), 721–734

Smith, S. L., Carden, R. F., Coad, B., Birkitt, T., & Pemberton, J. M. (2014). A survey of the hybridisation status of Cervus deer species on the island of Ireland. *Conservation Genetics*, 15(4), 823–835

Teacher, A. G. F., Garner, T. W. J., & Nichols, R. A. (2009). European phylogeography of the common frog (*Rana temporaria*): Routes of postglacial colonization into the British Isles, and evidence for an Irish glacial refugium. *Heredity*, 102(5), 490–496

Tiesmeyer, A., Ramos, L., Manuel Lucas, J., Steyer, K., Alves, P. C., Astaras, C., … Nowak, C. (2020). Range-wide patterns of human-mediated hybridisation in European wildcats. *Conservation Genetics*, 21, 247–260

Vega, R., Fløjgaard, C., Lira-Noriega, A., Nakazawa, Y., Svenning, J. C., & Searle, J. B. (2010). Northern glacial refugia for the pygmy shrew Sorex minutus in Europe revealed by phylogeographic analyses and species distribution modelling. *Ecography*, 33(2), 260–271

Walters, D. M., Blum, M. J., Rashleigh, B., Freeman, B. J., Porter, B. A., & Burkhead, N. M. (2008). Red shiner invasion and hybridization with blacktail shiner in the upper Coosa River, USA. *Biological Invasions*, 10(8), 1229–1242

Wangkumhang, P., & Hellenthal, G. (2018). Statistical methods for detecting admixture. *Current Opinion in Genetics and Development*, 53, 121–127

Warren E. Johnson, Eduardo Eizirik, Jill Pecon-Slattery, William J. Murphy, Agostinho Antunes, Emma Teeling, S. J. O. (2006). The late Miocene Radiation of Modrn Felidae: A Genetic Assessment. *Science*, 311(2006), 73–76

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16, 97–159

Yalden, D. W. (1982). When did the mammal fauna of the British Isles arrive? *Mammal Review*, 12(1), 1–57

Yi, X., & Latch, E. K. (2021). Nonrandom missing data can bias Principal Component Analysis inference of population genetic structure. *Molecular Ecology Resources*, Epub ahead of print

Zhou, Y., Browning, S. R., & Browning, B. L. (2020). A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *American Journal of Human Genetics*, 106(4), 426–437

5.7 Appendix 7. Supplementary material

Table 5.6. Proportion of wildcat ancestry estimated per individual by ADMIXTURE (K=2), AdmixTools (F_4 ratio test), PCAdmix and MOSAIC. Historic whole-genome (Historic WGD) and screening data (Historic SD) were evaluated using AdmixTools only.

Indv	**Population**	**ADMIX-TURE (K=2)**	**AdmixTools F_4 ratio test**			**PCAdmix**	**MOSAIC**
			alpha	**stderr**	**Z**		
WCQ243	Captive	1.00	0.817	0.015	55.5	0.79	0.78
WCQ340	Captive	1.00	0.805	0.014	59.1	0.79	0.76
WCQ343	Captive	1.00	0.848	0.012	70.6	0.81	0.82
WCQ427	Captive	0.92	0.787	0.013	58.6	0.76	0.75
WCQ428	Captive	1.00	0.826	0.013	64.1	0.80	0.81
WCQ553	Captive	1.00	0.843	0.013	65.1	0.82	0.82
WCQ047	Wild	0.99	0.875	0.011	76.9	0.84	0.83
WCQ052	Wild	0.91	0.813	0.013	62.6	0.79	0.77
WCQ099	Wild	0.06	-0.022	0.005	-4.2	0.17	0.07
WCQ100	Wild	0.37	0.263	0.015	18.0	0.38	0.31
WCQ158	Wild	0.93	0.806	0.013	64.5	0.78	0.78
WCQ165	Wild	0.08	-0.005	0.005	-0.9	0.19	0.08
WCQ168	Wild	0.48	0.373	0.014	27.5	0.47	0.41
WCQ209	Wild	0.78	0.649	0.016	41.5	0.64	0.62
WCQ211	Wild	1.00	0.889	0.012	77.2	0.84	0.84
WCQ212	Wild	0.81	0.676	0.015	44.7	0.66	0.64
WCQ213	Wild	0.82	0.698	0.014	49.5	0.72	0.68
WCQ214	Wild	0.90	0.788	0.013	60.5	0.78	0.77
WCQ216	Wild	0.67	0.551	0.012	44.3	0.59	0.55
WCQ218	Wild	0.76	0.635	0.016	38.9	0.65	0.64
WCQ224	Wild	0.80	0.680	0.016	43.7	0.68	0.67
WCQ227	Wild	0.28	0.183	0.014	12.9	0.32	0.24
WCQ230	Wild	0.42	0.324	0.016	20.9	0.44	0.37
WCQ231	Wild	0.50	0.379	0.014	26.4	0.45	0.40
WCQ234	Wild	0.59	0.478	0.015	31.3	0.52	0.49
WCQ236	Wild	0.59	0.463	0.016	29.5	0.51	0.48
WCQ246	Wild	0.68	0.572	0.015	37.2	0.60	0.58
WCQ248	Wild	1.00	0.848	0.012	71.4	0.82	0.82
WCQ249	Wild	0.58	0.444	0.017	26.3	0.49	0.44
WCQ252	Wild	0.56	0.428	0.016	26.3	0.50	0.43
WCQ255	Wild	0.48	0.348	0.016	21.3	0.44	0.38
WCQ515	Wild	1.00	0.868	0.012	74.4	0.82	0.82
WCQ613	Wild	0.53	0.417	0.014	30.4	0.49	0.44
WCQ903	Wild	0.59	0.494	0.015	32.7	0.56	0.51
WCQ904	Wild	0.45	0.349	0.011	33.2	0.46	0.39
WCQ915	Wild	0.82	0.706	0.014	51.7	0.71	0.68
WCQ0965	Historic WGD	NA	0.948	0.006	171.4	NA	NA
WCQ0986	Historic WGD						

WCQ1008	Historic WGD						
WCQ1021	Historic WGD						
WCQ0948	Historic SD		1.041	0.007	152.8		
WCQ0949	Historic SD		1.042	0.009	119.0		
WCQ0956	Historic SD		1.036	0.007	159.1		
WCQ0966	Historic SD		1.041	0.007	158.5		
WCQ0967	Historic SD		1.042	0.006	184.3		
WCQ1010	Historic SD		1.041	0.006	172.8		
WCQ1011	Historic SD		1.019	0.008	131.3		
WCQ1016	Historic SD		1.051	0.007	155.2		
WCQ1023	Historic SD		1.010	0.017	59.4		
WCQ1026	Historic SD		0.966	0.013	72.6		
WCQ1027	Historic SD		1.002	0.013	78.1		
WCQ1028	Historic SD		0.945	0.014	66.3		
WCQ1029	Historic SD		0.975	0.009	110.9		
WCQ1031	Historic SD		0.977	0.010	93.2		
WCQ1033	Historic SD		0.980	0.012	82.1		
WCQ1042	Historic SD		1.048	0.011	97.8		
WCQ1057	Historic SD		1.014	0.008	130.0		
WCQ1058	Historic SD		1.003	0.008	118.7		
WCQ1059	Historic SD		1.060	0.009	117.0		
WCQ1060	Historic SD		0.860	0.025	34.7		
WCQ1061	Historic SD		1.035	0.007	151.6		
WCQ1062	Historic SD		0.998	0.010	97.7		
WCQ1063	Historic SD		0.934	0.011	86.5		

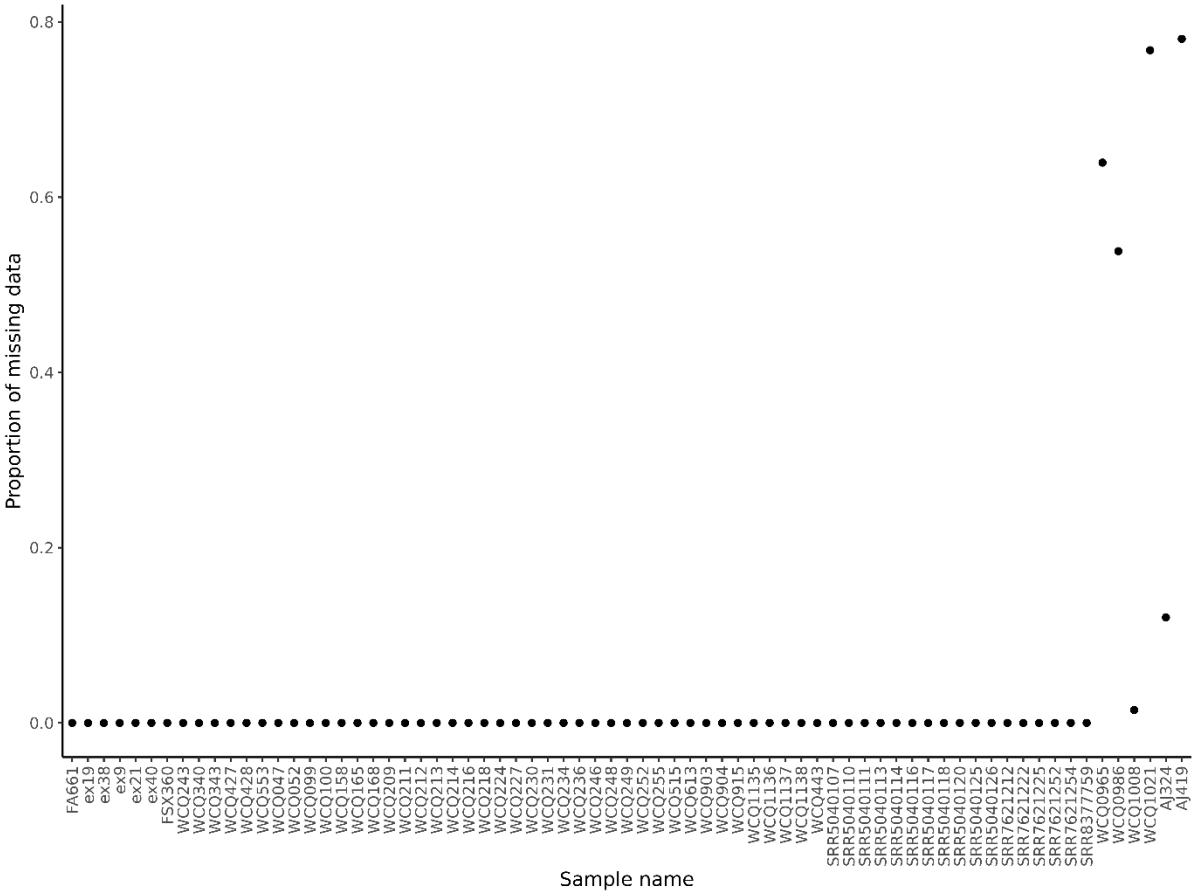

Figure 5.14. Proportion of missing data per individual across the 862,730 SNPs used for PCA. Modern samples were genotyped at all sites. These samples were used to build the principal components. Historic data were projected onto the PCA, missing data in these samples varied between 1.5% (WCQ1008) and 78.1% (AJ419).

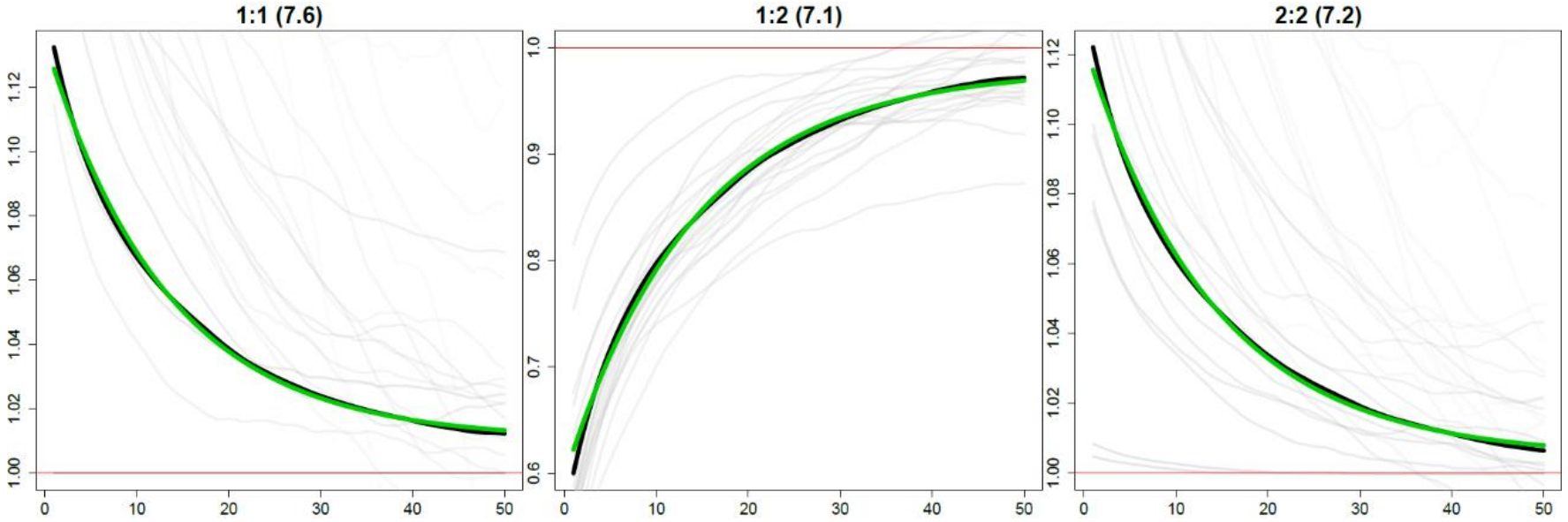

Figure 5.15. MOSAIC coancestry curves constructed using the captive population as part of the wildcat reference panel. Including introgressed individuals in the reference panel did not appear to bias individual coancestry curves (grey lines) (compared to Fig. 5.12), and the overall population mean estimate (7.3 generations) was not dissimilar to that reported for the expanded target panel (8.6, see 5.3.3).

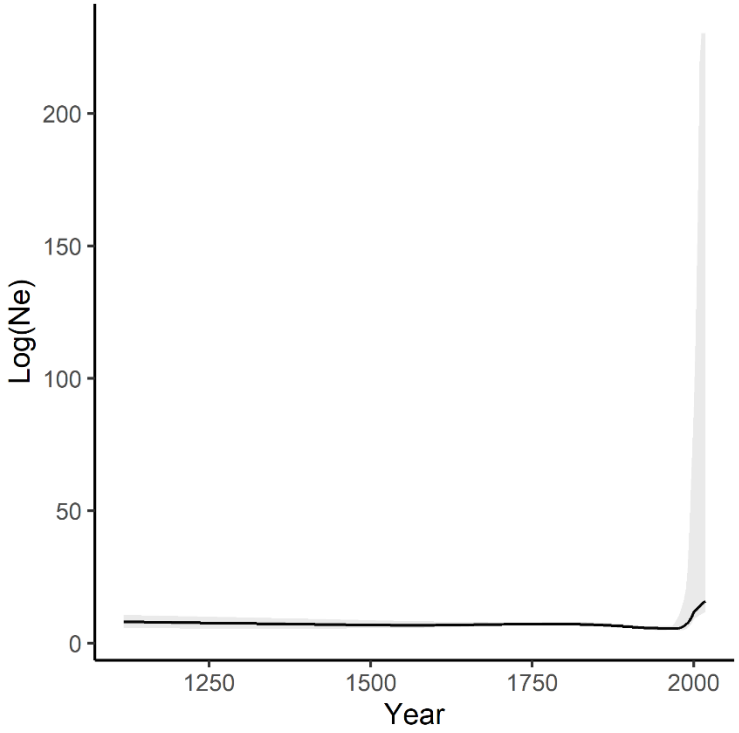

Figure 5.16. Results from IBDNe, plotted to show the extent of 95% CI widening once the assumption that individuals are from a non-admixed population had been violated. From the end of the 20th century IBDNe cannot give an accurate estimate of effective population size.

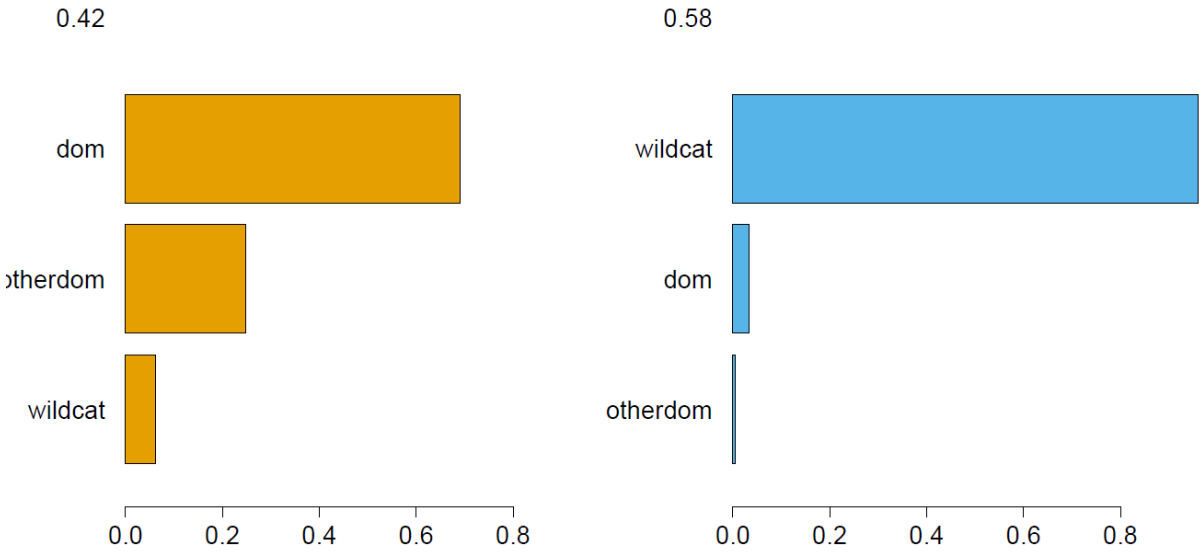

Figure 5.17. MOSAIC copying matrix. This shows the probability of haplotype copying from each of the reference populations, given the underlying ancestry (domestic, LHS, or wildcat, RHS). The three reference panels here are 'dom', Scottish domestics, 'otherdom', non-Scottish domestics and 'wildcat', mainland European wildcats. The inferred ancestral domestic group copies predominantly modern domestic cat haplotypes, and vice versa for the ancestral wildcat group.

Table 5.7. Pairwise F_{ST} values between each reference panel and the reconstructed ancestral partial genomes ('ancestral group 1', domestic, and 'ancestral group 2',wildcat). The three reference panels here are 'dom', Scottish domestics, 'otherdom', non-Scottish domestics and 'wildcat', mainland European wildcats.

Reference panel	**Ancestral group 1**	**Ancestral group 2**
dom	0.028	0.737
otherdom	0.089	0.640
wildcat	0.506	0.340

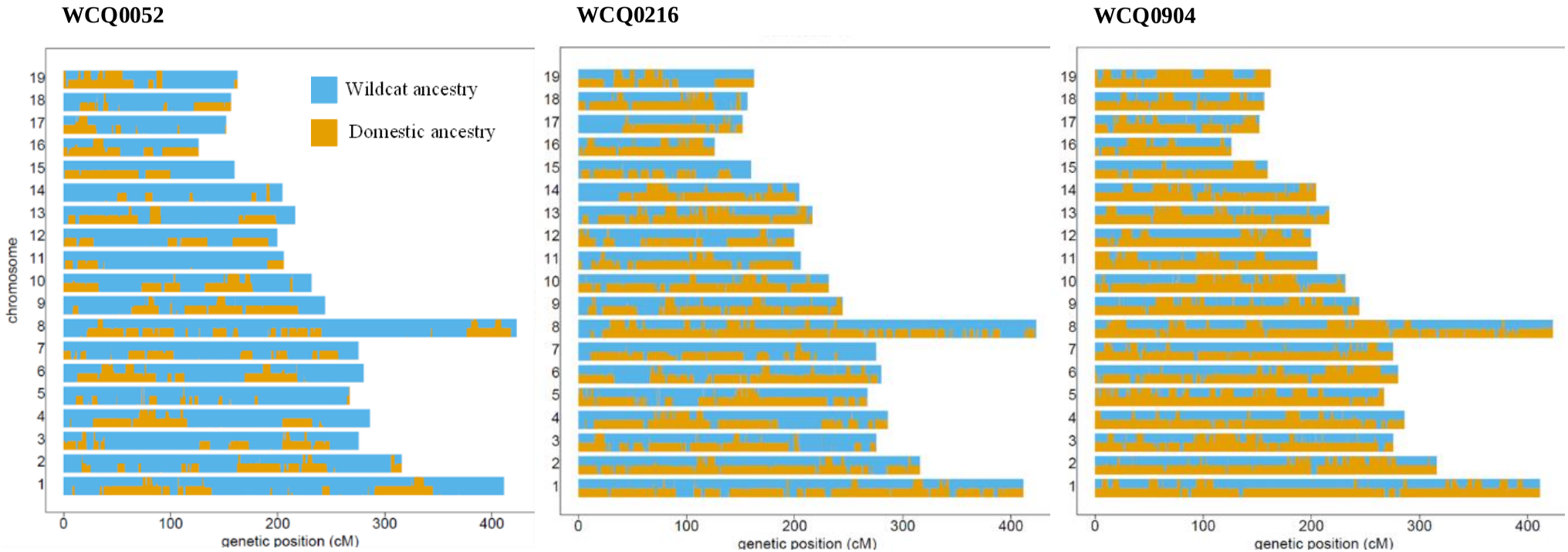

Figure 5.18. Example karyograms for three individuals: WCQ0052, WCQ0216 and WCQ0904. For each chromosome, 1-19, the two haplotypes are stacked on top of each other and coloured by inferred local ancestry across the chromosome (blue indicates wildcat and orange domestic local ancestry). Tracts of introgressed domestic cat DNA are clear for each individual

Table 5.8. Inferred date of admixture for each individual in the target population, taking sampling date into account and assuming a generation time for wildcats of three years. Probable feral domestic cats are indicated with a *.

Indv	Population	Est. time since admixture (generations)			Mean time since admixture (generations)	Mean time since admixture (years)	Sampling year	Est. date of admixture
		1:1	1:2	2:2				
WCQ047	Wild	8.6	8.6	8.5	8.6	25.7	UNK	NA
WCQ052	Wild	4.7	4.7	4.7	4.7	14.1	UNK	NA
WCQ099*	Wild	50.1	89	89.2	76.1	228.3	UNK	NA
WCQ100	Wild	10.2	10.2	10.2	10.2	30.6	UNK	NA
WCQ158	Wild	13.1	13	13	13.0	39.1	1997	1958
WCQ165*	Wild	43.4	62.1	62	55.8	167.5	UNK	NA
WCQ168	Wild	8.5	8.5	8.4	8.5	25.4	2007	1982
WCQ209	Wild	5.4	5.5	5.6	5.5	16.5	2013	1997
WCQ211	Wild	10.7	10.7	10.7	10.7	32.1	1999	1967
WCQ212	Wild	5.6	5.8	6	5.8	17.4	2011	1994
WCQ213	Wild	9.3	9.3	9.3	9.3	27.9	2002	1974
WCQ214	Wild	6.4	6.4	6.3	6.4	19.1	2003	1984
WCQ216	Wild	8.4	8.3	8.1	8.3	24.8	2014	1989
WCQ218	Wild	12.2	12.1	12	12.1	36.3	2011	1975
WCQ224	Wild	7.6	7.8	8	7.8	23.4	2014	1991
WCQ227	Wild	14.1	13.9	13.8	13.9	41.8	2012	1970
WCQ230	Wild	7.7	7.5	7.4	7.5	22.6	2011	1988
WCQ231	Wild	8.4	8.5	8.5	8.5	25.4	2011	1986
WCQ234	Wild	10.2	10.3	10.4	10.3	30.9	2014	1983
WCQ236	Wild	6.1	6.1	6.1	6.1	18.3	2011	1993
WCQ243	Captive	7.9	8	8.1	8	24	2014	1990
WCQ246	Wild	7.5	7.4	7.4	7.4	22.3	2014	1992
WCQ248	Wild	11.9	11.9	11.8	11.9	35.6	2014	1978
WCQ249	Wild	5.6	5.8	5.9	5.8	17.3	2013	1996
WCQ252	Wild	6.9	7.1	7.2	7.1	21.2	2014	1993
WCQ255	Wild	7.5	7.5	7.5	7.5	22.5	2012	1990
WCQ340	Captive	9.3	9.3	9.4	9.3	28	2014	1986
WCQ343	Captive	11.3	11.3	11.3	11.3	33.9	2014	1980
WCQ427	Captive	10.5	10.6	10.7	10.6	31.8	2012	1980
WCQ428	Captive	14.6	14.7	14.8	14.7	44.1	2015	1971
WCQ515	Wild	18.1	17.9	17.8	17.9	53.8	2014	1960
WCQ553	Captive	14.5	14.3	14.1	14.3	42.9	2015	1972
WCQ613	Wild	8.6	8.6	8.6	8.6	25.8	2016	1990
WCQ903	Wild	7.3	7.3	7.3	7.3	21.9	UNK	NA
WCQ904	Wild	15.6	15.6	15.5	15.6	46.7	UNK	NA
WCQ915	Wild	7.8	7.9	8.1	7.9	23.8	2018	1994

Chapter 6 General Discussion

6.1 A history of hybridisation in the Scottish wildcat population

An exact date for the arrival of modern domestic cats in Britain is disputed, however, it is clear from the available archaeological and genetic evidence that wildcats and domestic cats have been sympatric for thousands of years (Jamieson et al, in prep). Despite this, strong genetic differentiation is observed between the two species, the isolating mechanisms between which remain poorly understood. Strong genetic differentiation between domestic cats and (predominantly captive) wildcats in the UK (Chapter 2) supports the hypothesis that extensive hybridisation in Scotland is a modern phenomenon, i.e., not enough time has elapsed for high geneflow between the two species to result in homogenisation or 'despeciation'. 'Wildcat ancestry' is still identifiable in the modern Scottish population, and some wild-living cats appear to have a high proportion of putative wildcat ancestry. These individuals are clearly in the minority, however, and form one end of a continuum between domestic cats and wildcats, referred to as a hybrid swarm.

Evidence presented here (summarised in Table 6.1) supports the recent onset of hybridisation in Scotland, however, there are some discrepancies in the estimates obtained using different methods. The posterior distribution of the ABC model was implausibly recent, with a posterior mean of 3.3 generations (see 3.4.1). The application of unlinked SNP data seemed to identify signals of very recent admixture only (in this case from the mid-2000s, around five generations, or ~15 years later than the mean estimate from MOSAIC). As discussed in 5.1.1, haplotype information is an invaluable resource to infer local ancestry and therefore accurately date admixture. This information was leveraged for the analyses in Chapter 5. Of these analyses, MOSAIC (Salter-Townshend & Myers, 2019) appeared to give the most accurate estimate (8.6 generations before present), robust to introgressed reference panels (5.2.5), but potentially benefitting from additional sampling of domestic cat and wildcat haplotype diversity. This estimate was supported by the results from IBDNe, where loss of accurate information about effective population size was interpreted as a signal of admixture ~10 generations before present, just outside the upper bounds of the confidence intervals around the MOSAIC estimate (8.3-9.8 generations). However, IBDNe (Browning & Browning, 2015) is not specifically designed to infer admixture timing; the point estimate was inferred from the sudden population size increase and widening of confidence intervals (Fig. 5.9), and may therefore not reflect a precise date. The distinct clustering of ancient, historic, and modern Scottish samples following PCA (Fig. 5.6) also supports recent admixture, though the sampling gap (1939-1997) does not allow for a detailed picture of hybridisation during the 20th century or accurate estimation of the onset of admixture.

PCAdmix (Brisbin et al., 2012) did not converge on well-defined ancestry blocks, and consequently admixture date was highly dependent on the window size selected for analysis (an advantage of MOSAIC is that is does not require a pre-defined window size, allowing ancestry switches at any point along the chromosome [Salter-Townshend & Myers, 2019]).

Table 6.1. Summary of the methods used to estimate the onset of admixture in the Scottish wildcat population. For detailed results see the relevant section signposted below.

Method	**Section**	**Est. onset of admixture (generations before present)**	**Limitations**
ABC modelling	3.3.1	3.3 (1.21-5.6)	Unlinked SNP data has limited power to detect older admixture events. The current model did not seem to be able to infer admixture beyond 20 generations before present.
PCA clustering	5.3.1	8-27	Lack of samples collected between 1939-1997
IBDNe	5.3.3	10	IBDNe was not developed to give accurate estimates of admixture timings, and as such the point estimate is somewhat arbitrary.
PCAdmix	5.3.3	0.1-62.4	Estimates of admixture date were highly dependent on window size. PCAdmix was not able to converge on well-defined ancestry blocks and therefore unable to accurately date hybridisation in Scottish wildcats.
MOSAIC	5.3.3	8.6 (8.3-9.8)	Small reference panels sampled a limited amount of domestic cat/wildcat haplotype diversity.

Prior to this study, significant hybridisation was thought to have been the result of population expansion following reduced pressure from hunting and habitat loss at the start of the 20th century (Easterbee et al., 1991). Range expansion is a proposed driver of introgression in other species, including those that hybridise with domesticates, such as grey wolves (Galaverni et al., 2017), or polecats (Costa et al., 2013), and is thought to be due to the low density of conspecific mates on the leading edge of an expansion. Relative population densities are especially important in cases of wild/domestic hybridisation, where domestic species often significantly outnumber the native wild species. Importantly, hybridisation has been observed on the expanding edge of multiple wildcat populations (Nussberger et al., 2018; Randi, Pierpaoli, Beaumont, Ragni, & Sforzi, 2001), and is therefore a plausible explanation for the pattern of introgression in Scotland. However, the work presented here (Chapters 3 and 5, Table 6.1) supports a more recent onset of hybridisation, beginning in the mid-20th century and peaking in the 1980s and 1990s. This post-dates wildcat recovery in Scotland, where the population was believed to have been expanding during the early part of the 20th century, re-establishing the current range by the 1940s (Easterbee et al., 1991).

The captive population in Scotland provides an important resource for conservation of this species in the UK. Established during the 1960s, it has not been able to escape introgression (Chapter 5), though at a significantly lower rate than currently observed in the wild. This population is vital to

restoring this species in Britain, comprising of the individuals most closely resembling the historic wildcat population.

Further work is needed to understand the drivers of hybridisation in Scotland. As discussed in Chapter 5, a similar pattern of 20th century hybridisation has been observed in grey wolves, proposed to have occurred during a population expansion, and decreasing into the 21st century as the wolf population became established in the new range (Galaverni et al., 2017). Speculatively, it seems likely that the 20th century re-establishment of Scottish wildcats was limited, both in terms of geographic spread and population density, increasing vulnerability to hybridisation. This may have been driven by multiple factors. The current wildcat population is most obviously restricted from expanding southwards by the central belt between Edinburgh and Glasgow (Davis & Gray, 2010; Easterbee et al., 1991). The distribution within Scotland appears to be patchy, likely a result of slow habitat recovery (specifically afforestation), increasing urbanisation and intensive agriculture. A drive for increased food productivity post-World War II continued in Britain until the 1970s, supported by, for example, the 1946 Hill Farm Act and 1947 Agricultural Act, promoting the use of marginal land for farming (van der Wal et al., 2011). Even in areas of suitable habitat, wildcat numbers may have been limited by ongoing persecution. Despite a general decline in grouse moor management in the UK since the 1940s (Grant, Mallord, Stephen, & Thompson, 2012), grouse moorland still covers an estimated 8% of land in England and Scotland (van der Wal et al., 2011), with the southern uplands and north-eastern Highlands of Scotland among the most intensively managed areas (Grant et al., 2012). Grouse moor management is a source of wildlife conflict. The distribution and abundance of many birds of prey, for example, golden eagles, hen harriers or peregrines, are limited by the presence of grouse moors, the result of both legal management practices and illegal persecution (Amar et al., 2012; Fielding et al., 2011; Whitfield, Fielding, Mcleod, & Haworth, 2004). Mammal species such as wildcats, foxes, stoats, weasels, pine martins and polecats are also subject to heavy predator control, accidental or otherwise (Sainsbury et al., 2019). Note that wildcats did not receive legal protection from persecution in the UK until 1988. The late 20th century also corresponded to a period of technological advancement in terms of predator control (Kerr, Hodges, & Sandrini, 2017). A trend of increasing domestic cat ownership in the UK (PFMA, n.d.), including a large number of free-ranging or feral domestic cats, is in stark contrast to declines in the wildcat population observed over the last few hundred years.

A small, fragmented wildcat population and expanding domestic cat population seems likely to have promoted hybridisation and introgression. Similar to the dynamics at the leading edge of a population expansion, if the wildcat population in Scotland struggled to re-establish a critical density, or a sufficiently large, interconnected metapopulation, hybridisation events may have occurred more frequently. Poor habitat quality, and/or the extent of human-mediated environments, may have accelerated contact with domestic cats (Kilshaw et al., 2016). As shown by the modelling work of

Quilodrán *et al.* (2020), even seemingly low rates of hybridisation can lead to high introgression levels within a short time frame. In Scotland, hybridisation has led to the breakdown of isolating mechanisms, rapidly leading to a complete demographic collapse. This genetic analysis supports wildcat hybridisation and introgression in Scotland as symptoms of long-term threats, specifically heavy persecution and habitat loss and fragmentation, which have decimated the wildcat population in Britain, and which must be addressed to re-establish this species in the UK.

6.2 Implications for conservation

The rapid development of the hybrid swarm in Scotland may be cause for concern in other species at risk of genetic swamping, not least wildcat populations across continental Europe. Many threats to wildcats are common across the species range, including a pattern of heavy persecution and habitat loss since the Middle Ages, and a recent history of urbanisation, habitat degradation and predator control (Lozano & Malo, 2012; Yamaguchi et al., 2015). Domestic cat ownership continues to increase, there are now an estimated 65 million domestic cats in European countries with wildcat populations (EPFI, 2017). In Germany the number of pet cats almost doubled between 2010 and 2019, from 8.2 to 14.8 million (Koptyug, 2020).

Currently, wildcat hybrids are most commonly found at the edges of expanding wildcat populations, such as those in Germany, France, and Switzerland (Tiesmeyer et al., 2020). Habitat fragmentation is also believed to promote hybridisation. In France, for example, a higher rate of introgression is observed in the north-east, where forest habitat is interspersed with agricultural land, compared to populations in continuous, undisturbed, forest habitat in the French Pyrenees (Beugin et al., 2020). A low rate of hybridisation on the Iberian Peninsula (Oliveira, Godinho, Randi, & Alves, 2008; Oliveira, Godinho, Randi, Ferrand, & Alves, 2008) is thought to be the result of ecological separation between wildcats and domestic cats (Gil-Sánchez et al., 2020), specifically, feral domestic cats are excluded from wildcat habitat due to their dependence on humans for food, and subsequent proximity to human settlements, interspecific competition and aggression from wildcats and other carnivores, such as foxes, and predation by raptors (e.g., golden eagles or eagle owls). Predicted habitat occupancy in Scotland also showed limited overlap between domestic cats and wildcats, with the hybrid population providing a 'bridge for gene flow' (Kilshaw et al., 2016).

The question of 'what to do with hybrids' remains to be answered for wildcats. Clearly, this will be context dependent. Populations with more extensive hybridisation, such as in Scotland, may confer a higher value to hybrids (as a reservoir of 'wildcat genes'), compared to regions with relatively stable wildcat populations. Additionally, legislation to control hybrid and feral domestic cats, as well as the priorities of stakeholders (e.g., hunters), varies regionally. To better support

decisions about the potential conservation value of hybrids, it is important to establish their relative fitness, behaviour, and ecosystem function in both wildcat and human-mediated environments. The absence of feral domestic or hybrid individuals in some regions (discussed above) suggests hybrids may be less fit in wildcat environments, though analyses by Mattucci *et al.* (2019) showed regions of both domestic and wildcat ancestry potentially under selection in hybrids. Further research is needed in this area to support evidence-based conservation of the hybrid population.

As suggested by Wayne & Shaffer (2016), protecting or restoring habitat may therefore be an important first step for managing wildcat hybridisation, allowing natural selection to 'find' the best genes for the environment. Habitat restoration has been shown to reverse anthropogenic hybridisation in some cases, for example, Heiser (1979) describes hybridisation between *Helianthus* species in areas disturbed by grazing and road construction. Over a 22-year period, restoration of the habitat appeared to allow a return of the parental types. Continuous areas of high-quality habitat may also support higher densities of wildcats, limiting 'edge effect' hybridisation and promoting intraspecific mating. Quilodrán *et al.* (2020) show that relative population densities are key to limiting hybridisation, and propose improving the quality, quantity, and connectivity of wildcat habitats, and well as controlling the number of domestic cats and hybrids, as conservation priorities.

Many hybridising species, including wolves (Pilot et al., 2018), wild boar (Iacolina et al., 2018), and red deer (Senn & Pemberton, 2009), demonstrate variable geographic and temporal patterns of admixture, which have not necessarily resulted in local extinctions. In Scotland, however, wildcat hybridisation appears to have reached a 'tipping point' leading to rapid genetic swamping and near extirpation of the species. Analyses presented here suggest this has occurred within a 50-year period, illustrating how hybridisation can quickly overwhelm vulnerable populations. Effective conservation of this species into the future will require a better understanding of the factors driving localised hybridisation, and the ability to predict any potential 'tipping point' of genetic swamping. For this, comparative analyses across European wildcat populations would be valuable. A limited number of Europe-wide studies have been published to date, principally aimed at detecting introgression (Mattucci et al., 2019; Oliveira et al., 2015; Tiesmeyer et al., 2020). Expanding these datasets to explore spatial patterns and patterns of hybridisation over time, as well as including metainformation, e.g., about habitat quality or domestic cat presence, would support work to identify potential drivers of hybridisation. Standardisation of genetic and morphology-based methods to quantify hybridisation would better support co-ordinated analyses of hybridisation dynamics, including a systematic approach to monitoring. Detailed studies of wildcat hybridisation may also contribute to a better understanding of localised introgression and genetic swamping in other species.

6.3 Future directions

Fundamental research questions, in terms of wildcat conservation, have been outlined above, specifically, the need for comparative analyses across European wildcat populations in order to understand the drivers of hybridisation and fitness and ecological function of hybrid individuals. Whole-genome sequence data provides a valuable resource to support this, and other analyses, and some possible future directions are described below.

As discussed in Chapter 4, limited genomic resources are available for the wildcat. This study represents, as far as we are aware, the largest whole genome resequencing dataset described for wildcats to date. It would be beneficial to generate a high-quality wildcat reference sequence, especially to examine copy number or structural variation between wildcats and domestic cats, and to improve imputation of low-quality sequencing data, for example, from historic samples (Fuentes-Pardo & Ruzzante, 2017). Further sequencing of parent-offspring trios would improve phasing for wildcats (Marchini et al., 2006), with potential downstream improvements for local ancestry estimation (Price et al., 2009).

Once identified, accurate local ancestry estimates allow the introgressed domestic tracts of hybrid individuals to be masked, with the aim of addressing questions relating to wildcat population structure within Scotland, divergence between Scotland and continental Europe, and ancestry-specific population size and genetic diversity, which shared patterns of introgressed variation have previously precluded (e.g., Mattucci et al., 2016). This would have significant value for the conservation programme to give a picture of the historic population size, structure, and diversity, and could also inform selective breeding of the captive population. Selective breeding could be applied with the aim of reducing the overall proportion of introgressed domestic DNA in the captive population, using molecular methods to monitor introgression in subsequent generations (Amador, Fernández, & Meuwissen, 2013; Amador, Hayes, & Daetwyler, 2014; Amador, Toro, & Fernández, 2012).

Genomic data represent a powerful resource with which to investigate natural selection (Vitti, Grossman, & Sabeti, 2013). Genome-wide scans for selection or admixture mapping can be applied to evaluate the fitness of the hybrid population and contribute more widely to our understanding of wildcats and their interaction with the environment, and the process of cat domestication. The hybrid swarm in Scotland is also a valuable study system to investigate fundamental questions about evolutionary biology. As outlined in Chapter 1, recognition of the role of hybridisation, especially in animal species, is relatively recent; studies of hybridisation in wild populations are therefore valuable opportunities to examine processes such as recombination, adaptation, and genetic drift (Barton, 2001).

The demographic model developed for wildcats (Chapter 3) is a useful and flexible tool to simulate wildcat history, and potentially to predict patterns of introgression into the future. The current model could easily be extended for whole-genome sequence data, and to incorporate information from historic samples (Haller & Messer, 2017). It could also be adapted to examine spatial patterns of introgression over smaller scales. As demonstrated in Chapter 3, it provides a useful tool to calibrate various analyses, for example, it could be used here to assess the application of MOSAIC to wildcat populations (and its appropriateness for young hybrid swarms).

6.4 Conclusion

The main aim of this thesis has been to understand the history of admixture between wildcats and domestic cats in Scotland. Hybridisation drives evolutionary change, with varied and often unpredictable outcomes. It highlights the dynamic nature of evolutionary processes, such as natural selection and genetic drift, and can be challenging to identify or monitor. In the face of increasing globalisation, habitat disturbance and climate change, anthropogenic hybridisation is considered a threat to wild populations. The potential for rapid extinctions following genetic or demographic swamping promotes a 'precautionary principle' among many conservationists, especially for rare or endangered species. The negative impacts of hybridisation continue to be debated, however, stimulating important discussions about the nature of species and priorities for species conservation. Certainly, geneflow between populations can also increase genetic diversity, conferring adaptive variation and increasing the resilience of populations in the face of environmental change. Parental species are often the sole target of conservation programmes or policy, largely ignoring the hybrid population, which, in the face of massive biodiversity declines, may be valuable for maintaining a diverse, functioning ecosystem.

The Scottish wildcat is an important example of hybridisation between a rare native species and widespread domesticate which, without conservation intervention, is likely to result in genetic swamping and the permanent loss of wildcat ancestry from Britain. An accurate timeline of hybridisation supports practical management decisions concerning both the captive breeding programme and the potential value of the hybrid population. It helps to establish a baseline for wildcats in Scotland, which is especially important given that systematic monitoring of this species did not begin until after the onset of widespread hybridisation and introgression. A clear understanding of hybridisation dynamics in the past will support conservation of this species into the future, not just in Scotland, but across the European wildcat range.

6.5 References

Amador, C., Fernández, J., & Meuwissen, T. H. E. (2013). Advantages of using molecular coancestry in the removal of introgressed genetic material. *Genetics Selection Evolution*, 45(1), 1–10

Amador, C., Hayes, B. J., & Daetwyler, H. D. (2014). Genomic selection for recovery of original genetic background from hybrids of endangered and common breeds. *Evolutionary Applications*, 7(2), 227–237

Amador, C., Toro, M. Á., & Fernández, J. (2012). Molecular Markers Allow to Remove Introgressed Genetic Background: A Simulation Study. *PLoS ONE*, 7(11)

Amar, A., Court, I.R., Davison, M., Grimshaw, T., Pickford, T. & Raw, D. 2012. Linking nest histories, remotely sensed land use data and wildlife crime records to explore the impact of grouse moor management on peregrine falcon populations. *Biological Conservation*, **145**, 86-94

Barton, N. H. (2001). The role of hybridization in evolution. *Molecular Ecology*, 10, 551–568

Beugin, M. P., Salvador, O., Leblanc, G., Queney, G., Natoli, E., & Pontier, D. (2020). Hybridization between *Felis silvestris silvestris* and *Felis silvestris* catus in two contrasted environments in France. *Ecology and Evolution*, 10, 263–276

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., … Bustamante, C. D. (2012). Pcadmix: Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human Biology*, 84(4), 343–364

Browning, S. R., & Browning, B. L. (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *American Journal of Human Genetics*, *97*(3), 404–418

Costa, M., Fernandes, C., Birks, J. D. S., Kitchener, A. C., Santos-Reis, M., & Bruford, M. W. (2013). The genetic legacy of the 19th-century decline of the British polecat: Evidence for extensive introgression from feral ferrets. *Molecular Ecology*, 22(20), 5130–5147

Davis, A. R. & Gray, D. (2010) *The distribution of Scottish wildcats (Felis silvestris) in Scotland (2006-2008*). Scottish Natural Heritage Commissioned Report No. 360

Easterbee. N., Hepburn. L. V., Jeffries, D. J. (1991) *Survey of the status and distribution of the wildcat in Scotland, 1983–1987*. Nature Conservancy Council for Scotland, Edinburgh

European Pet Food Industry (2017) Facts and figures 2017. Retrieved July 17, 2018, from https://www.fedia f.org/who-we-are/european-statistics.html

Fielding, A., Haworth, P., Whitfield, P., McLeod, D. & Riley, H. 2011. *A Conservation Framework for Hen Harriers in the United Kingdom*. JNCC Report No: 441. JNCC, Peterborough

Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology*, 26(20), 5369–5406

Galaverni, M., Caniglia, R., Pagani, L., Fabbri, E., Boattini, A., & Randi, E. (2017). Disentangling timing of admixture, patterns of introgression, and phenotypic indicators in a hybridizing wolf population. *Molecular Biology and Evolution*, 34(9), 2324–2339

Gil-Sánchez, J. M., Barea-Azcón, J. M., Jaramillo, J., Herrera-Sánchez, F. J., Jiménez, J., & Virgós, E. (2020). Fragmentation and low density as major conservation challenges for the southernmost populations of the European wildcat. *PLoS ONE*, 15(1), 1–21

Grant, M., Mallord, J., Stephen, L., & Thompson, P. S. (2012). *The costs and benefits of grouse moor management to biodiversity and aspects of the wider environment: a review* (RSPB Research Report Number 43). RSPB, Sandy, Bedfordshire. http://www.rspb.org.uk/Images/grant_mallord_stephen_thompson_2012_tcm9-318973.pdf

Haller, B. C., & Messer, P. W. (2017). SLiM 2: Flexible, interactive forward genetic simulations. *Molecular Biology and Evolution*, 34(1), 230–240

Heiser, C. (1979). Hybrid populations of *Helianthus divaricatus* and *H. microcephalus* after 22 years. *Taxon*, 28(1), 71–75.

Iacolina, L., Pertoldi, C., Amills, M., Kusza, S., Megens, H. J., Bâlteanu, V. A., … Stronen, A. V. (2018). Hotspots of recent hybridization between pigs and wild boars in Europe. *Scientific Reports*, 8, 17372

Jamieson, A., (2021) The evolutionary history of domestic cats in Europe. Manuscript in preparation.

Kerr, A., Hodges, T., & Sandrini, J. (2017). *Exploring new technologies for hunting, review and recommendations.* Wyoming Game and Fish Department.

Kilshaw, K., Montgomery, R. A., Campbell, R. D., Hetherington, D. A., Johnson, P. J., Kitchener, A. C., … Millspaugh, J. J. (2016). Mapping the spatial configuration of hybridization risk for an endangered population of the European wildcat (*Felis silvestris silvestris*) in Scotland. *Mammal Research*, 61(1), 1–11

Koptyug, E. (2020) Number of pets in Germany 2000-2019, by type of animal. Retrieved October 1, 2021, from https://www.statista.com/statistics/552971/pets-number-by-type-germany/

Lozano, J., & Malo, A. F. (2012). Conservation of the European Wildcat (*Felis silvestris*) in Mediterranean environments: A reassessment of current threats. In G. S. Williams (Ed.), *Mediterranean Ecosystems: Dynamics, Management and Conservation* (pp. 1–31). Nova Science Publishers, Inc.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., … Donnelly, P. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics*, 78, 437–450

Mattucci, F., Galaverni, M., Lyons, L. A., Alves, P. C., Randi, E., Velli, E., … Caniglia, R. (2019). Genomic approaches to identify hybrids and estimate admixture times in European wildcat populations. *Scientific Reports*, 9(1), 1–15

Mattucci, F., Oliveira, R., Lyons, L. A., Alves, P. C., & Randi, E. (2016). European wildcat populations are subdivided into five main biogeographic groups: Consequences of Pleistocene climate changes or recent anthropogenic fragmentation? *Ecology and Evolution*, 6(1), 3–22

Nussberger, B., Currat, M., Quilodran, C. S., Ponta, N., & Keller, L. F. (2018). Range expansion as an explanation for introgression in European wildcats. *Biological Conservation*, 218(2018), 49–56

Oliveira, R., Randi, E., Mattucci, F., Kurushima, J. D., Lyons, L. A., & Alves, P. C. (2015). Toward a genome-wide approach for detecting hybrids: Informative SNPs to detect introgression between domestic cats and European wildcats (*Felis silvestris*). *Heredity*, 115(3), 195–205

Oliveira, Rita, Godinho, R., Randi, E., & Alves, P. C. (2008). Hybridization versus conservation: are domestic cats threatening the genetic integrity of wildcats (*Felis silvestris silvestris*) in Iberian Peninsula? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363, 2953–2961

Oliveira, Rita, Godinho, R., Randi, E., Ferrand, N., & Alves, P. C. (2008). Molecular analysis of hybridisation between wild and domestic cats (*Felis silvestris*) in Portugal: Implications for conservation. *Conservation Genetics*, 9, 1–11

Pet Food Manufacturers Association, (n.d.) Historical Pet Ownership 1965-2004. Retrieved October 1, 2021, from https://www.pfma.org.uk/historical-pet-ownership-statistics

Pilot, M., Greco, C., VonHoldt, B. M., Randi, E., Jędrzejewski, W., Sidorovich, V. E., … Wayne, R. K. (2018). Widespread, long-term admixture between grey wolves and domestic dogs across Eurasia and its implications for the conservation status of hybrids. *Evolutionary Applications*, 11, 662–680

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., … Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6), e1000519

Quilodrán, C. S., Nussberger, B., Macdonald, D. W., Montoya-Burgos, J. I., & Currat, M. (2020). Projecting introgression from domestic cats into European wildcats in the Swiss Jura. *Evolutionary Applications*, 13, 1–12

Randi, E., Pierpaoli, M., Beaumont, M., Ragni, B., & Sforzi, A. (2001). Genetic identification of wild and domestic cats (*Felis silvestris*) and their hybrids using Bayesian clustering methods. *Molecular Biology and Evolution*, 18(9), 1679–1693

Salter-Townshend, M., & Myers, S. (2019). Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics*, 212, 869–889

Sainsbury, K. A., Shore, R. F., Schofield, H., Croose, E., Campbell, R. D., & McDonald, R. A. (2019). Recent history, current status, conservation and management of native mammalian carnivore species in Great Britain. *Mammal Review*, 49(2), 171–188

Senn, H. V., & Pemberton, J. M. (2009). Variable extent of hybridization between invasive sika (*Cervus nippon*) and native red deer (*C. elaphus*) in a small geographical area. *Molecular Ecology*, 18(5), 862–876

Tiesmeyer, A., Ramos, L., Manuel Lucas, J., Steyer, K., Alves, P. C., Astaras, C., … Nowak, C. (2020). Range-wide patterns of human-mediated hybridisation in European wildcats. *Conservation Genetics*, 21, 247–260

van der Wal, R., Bonn, A., Monteith, D., Reed, M., Blackstock, K., Hanley, N., … Tinch, D. (2011). Chapter 5: Mountains, Moorlands and Heaths. In *The UK National Ecosystem Assessment Technical Report* (pp. 1–105). UNEP-WCMC, Cambridge

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, 47, 97–120

Wayne, R. K., & Shaffer, H. B. (2016). Hybridization and endangered species protection in the molecular era. *Molecular Ecology*, 25, 2680–2689

Whitfield, D. P., Fielding, A. H., Mcleod, D. R. A., Haworth, P. F. (2004) The effects of persecution of age of breeding and territory occupation in golden eagles in Scotland. *Biological Conservation*, 118, 249-259

Yamaguchi, N., Kitchener, A., Driscoll, C., & Nussberger, B. (2015). *Felis silvestris. The IUCN Red List of Threatened Species 2015*, e.T60354712A50652361