INTRODUCTION

Hybridisation and introgression are important drivers of evolutionary change (Barton, 2001). Evolutionary processes may be disrupted by human activity, particularly when species distributions are altered by, for example, climate change, landscape use, or introduction of non-native species, leading to contact between populations that were previously allopatric. Whilst it is recognised this can generate a range...
of outcomes, some of which may be positive (e.g., genetic rescue; Johnson et al., 2010 or adaptive introgression; Pardo-Díaz et al., 2012), hybridisation and introgression are often considered threats to wild populations (Rhymer & Simberloff, 1996; Todesco et al., 2016). Loss of locally adaptive variation, outbreeding depression or genetic swamping can all result in population or species extinction. Furthermore, introgressive hybridisation between domesticated species and wild populations increases the spread of potentially maladaptive, artificially selected variants in the wild (Randi, 2008).

The wildcat population in Scotland is an example of the threat of genetic extinction as a result of hybridisation (Mathews et al., 2018). The wildcat, Felis silvestris, is Britain’s most endangered carnivore and last remaining wild feline species. Wildcats have faced a long history of persecution and habitat loss and can hybridise with domestic cats to produce fertile offspring. Introgressive hybridisation is an increasingly serious threat to the dwindling population of this species in the Britain, which is now at risk of complete genetic replacement by hybrids in the wild (Breitenmoser et al., 2019).

Modern domestic cats are derived from the Near Eastern wildcat species Felis lybica (Driscoll et al., 2007). Both F. lybica and F. silvestris belong to the domestic cat lineage of the Felidae family, diverging from a common ancestor ~1.4 million years ago (Johnson et al., 2006). The process of cat domestication was probably initiated as a result of their attraction to rodents who themselves were attracted to grain stores associated with settled agriculture ~9500 years ago (Driscoll et al., 2007). Although Driscoll et al. (2007) described just one wildcat species, Felis silvestris, distributed across Europe, Asia, and Africa, a recently revised Felidae taxonomy recognises two species of wildcat, Felis silvestris present in Europe, Caucasus and Turkey, and Felis lybica distributed in Africa and Asia (Kitchener et al., 2017).

Artificial selection has altered the morphology, behaviour, and rate of reproduction of domestic cats (Driscoll et al., 2009). As a result, they are sufficiently diverged from wildcats to be considered a separate species, Felis catus (International Commission on Zoological Nomenclature, 2003). Domestic cats are widespread globally and found throughout the Felis silvestris range. Hybridisation between domestic cats and wildcats is variable across the wildcat range in Europe (Yamaguchi et al., 2015) and is particularly acute in Scotland where the ratio of un-neutered hybrids to wildcats was estimated at 6:1 (Breitenmoser et al., 2019).

The Scottish wildcat has served as a canonical example of domestic-wild hybridisation more generally. The aim of this study was, first, to clarify the population structure of wildcats in Scotland using a two-fold increase in the number of genetic markers compared to the most recent study (Senn et al., 2019). For this, we used ddRAD-seq data; ddRAD-seq is an efficient way to sample thousands of markers for genome-wide estimates of hybridisation (Peterson et al., 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 used an expanded set of markers to evaluate the effectiveness of current tests to identify hybrid individuals.

Importantly, we aimed to estimate the timescale of hybridisation using a model that predicts the observed pattern of population structure. A demographic model for Scottish wildcats has been developed using an approximate Bayesian computational (ABC) framework (Beaumont et al., 2002), a model-based approach to parameter inference rooted in Bayesian statistics. For this we have developed an efficient two-stage modelling approach, and a novel method for choosing summary statistics to improve model fit.

Finally, we show how this neutral demographic model is a useful tool to evaluate the accuracy of methods to identify genes that are subject to natural selection in structured populations, and hence calibrate them.

2 | MATERIALS AND METHODS

2.1 | Data processing

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 Scottish Wildcat Action www.scottishwildcataction.org trapped cats, 31 roadkill samples) and five Scottish domestic cats (for full sample details see Table S1).

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 original 76 samples. Historical wildcat samples derived from museum specimens reported in Senn et al. (2019) could not be used in this study due to poor DNA quality. 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 et al., 2013). Demultiplexed reads were trimmed to 135 bp and concatenated into a single read file per individual. Analysis of raw sequence reads diverged from that of Senn et al. (2019) from this point forward (described below), significant differences included the 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.

Sequence reads were aligned using BWA (Li & Durbin, 2009) to the Felis catus reference genome v9.0 (GCF_000181335.3) (Pontius et al., 2007). The proportion of mapped reads appeared to be high, even for putative wildcat samples (Table S2). We visually checked read alignment and linkage disequilibrium (LD) using Haploview (Barrett et al., 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". We allowed multiple SNPs per read, and the mean number of SNPs per read across the final data set was 1.6. Variants were filtered using a minimum allele frequency of 0.05 and maximum proportion of heterozygous individuals of 0.7, and the three sample sources (domestic, wild-living, and captive) were treated as separate populations.

PLINK v1.9 (Chang et al., 2015) and VCFtools v1.15 (Danecék 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 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. Individuals were then removed from the data set to limit relatedness (for the full list of excluded individuals see Table S1). Population genetic summary statistics (observed and expected heterozygosity, inbreeding coefficient and pairwise $F_{st}$; Weir & Cockerham, 1984) were generated using PLINK and VCFtools.

### 2.2 Population structure

Principal component analysis (PCA) and ADMIXTURE (Alexander et al., 2009) were used to examine population structure. PCA was completed in R using pcreg. ADMIXTURE analyses were performed for seven values of $K$, ranging from 2–8, and included a calculation of cross-validation error to estimate the optimal value of $K$. All SNPs were included, and the data were not considered dense enough to require thinning (to minimise background linkage disequilibrium) prior to the analysis (Alexander et al., 2009).

### 2.3 Existing hybrid tests

Hybrid individuals are currently identified using a combination of genetic and morphological diagnostic tests: a seven-point pelage scoring system (Kitchener et al., 2005) and a 35 SNP genetic test (Senn & Ogden, 2015). The pelage test (7PS) scores seven key morphological characteristics on an ordinal scale of 1,2,3 for domestic, hybrid or wildcat features, respectively. Putative wildcats score 19 or higher on this test (maximum score 21), although a lower threshold of 17 can be used to overcome possible recorder error, for example, from poor quality camera-trap photos. The genetic test uses 35 SNPs that differentiate between wildcats and domestic cats (Nussberger et al., 2013; Senn & Ogden, 2015). A "hybrid score" is generated using STRUCTURE Q values between 0 and 1 (Pritchard et al., 2000); higher values correspond to individuals with more wildcat ancestry. An LBQ score (i.e., the lower boundary of the Q 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 compared the performance of these hybrid tests using ADMIXTURE Q values (Q6546) from the ddRAD-seq data to determine hybrid status. 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 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 to select reference samples for ROC analysis). Receiver operating characteristic (ROC) curves were then constructed to assess performance (Robin et al., 2011). 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. 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.
2.4 | Modelling wildcat demography

A demographic model for wildcats was developed within an ABC framework (Beaumont et al., 2002). ABC was developed as a rejection algorithm (Pritchard et al., 1999), in which simulated data are generated under a hypothesised model of evolution, with model parameters sampled from a known prior distribution. Summary statistics are taken from both the simulated data and observed data. An accepted sample of simulations (those with summaries closest to the observed data) are then used to estimate posterior distributions of the model parameters. Posterior estimates from the basic rejection algorithm can be improved with local linear (Beaumont et al., 2002) or nonlinear regression (Blum & François, 2010).

The model developed for wildcat demography is outlined below. 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 et al., 2018), 500 generations (~1500 years) therefore approximately spans the time 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 first 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 (Figure S1A), and the priors for which are given in Table 1. The metapopulation SNP frequency, $X$, is simulated from beta(1, ancvar), which assumes that the nonreference allele is typically rarer (empirically confirmed). Parameters $F_1$ and $F_2$, model 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 $2p_0$, and $2p_2$. 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 $p_3$) 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 since it was established). Population sizes remain constant throughout the simulation; we did not model any fluctuations in wildcat population size (e.g., recent population expansion), and we did not model 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.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Prior distribution</th>
<th>Posterior mean (95% HPD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ancvar</td>
<td>Generates baseline ancestral variation</td>
<td>Exponential $^a$ $\lambda = 0.1$</td>
<td>4.155 (2.441–5.801)</td>
</tr>
<tr>
<td>$F_1$</td>
<td>Drift from baseline (pop$_1$)</td>
<td>Beta $\alpha = 2, \beta = 10$</td>
<td>0.211 (0.047–0.391)</td>
</tr>
<tr>
<td>$F_2$</td>
<td>Drift from baseline (pop$_2$)</td>
<td>Beta $\alpha = 2, \beta = 10$</td>
<td>0.183 (0.036–0.336)</td>
</tr>
<tr>
<td>$\log(p_1)$</td>
<td>Log population size</td>
<td>Normal $\mu = 6.5, \sigma = 0.5$</td>
<td>6.429 (5.813–7.167)</td>
</tr>
<tr>
<td>$\log(p_2)$</td>
<td></td>
<td>Normal $\mu = 6.5, \sigma = 0.5$</td>
<td>6.580 (5.924–7.426)</td>
</tr>
<tr>
<td>$\log(p_3)$</td>
<td></td>
<td>Normal $\mu = 6.5, \sigma = 0.5$</td>
<td>4.469 (3.986–5.099)</td>
</tr>
<tr>
<td>$T_1$</td>
<td>Onset of gene flow from pop$_3$ to pop$_2$ (number of generations)</td>
<td>Exponential $\lambda = 0.02$</td>
<td>3.326 (1.209–5.602)</td>
</tr>
<tr>
<td>$T_2$</td>
<td>Time pop$_3$ is established from a sample of pop$_2$ (number of generations)</td>
<td>Gamma $\alpha = 9, \theta = 0.5$</td>
<td>19.272 (9.430–30)</td>
</tr>
<tr>
<td>$mig_1$</td>
<td>Migration (per generation) pop$_3$ to pop$_2$</td>
<td>Beta $\alpha = 5, \beta = 20$</td>
<td>0.128 (0.067–0.192)</td>
</tr>
<tr>
<td>$mig_2$</td>
<td>Migration (every three generations) pop$_2$ to pop$_3$</td>
<td>Gamma $^a$ $\alpha = 1, \theta = 1$</td>
<td>0.012 (0–0.037)</td>
</tr>
</tbody>
</table>

$^a$(exponential distribution with rate parameter $\lambda = 0.1$) + 1 to avoid values of ancvar less than 1.

$^b$The lower bound of this distribution was limited to 60 to avoid simulating a population of captive individuals smaller than the target data.

$^c$(gamma distribution with shape parameter $\alpha = 1$ and scale parameter $\theta = 1$)/size of captive population.
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).

Prior distributions for these 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 is 0.446, 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 than the target 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). \( M_{ig} \) was a parameter of particular interest as it corresponds to the rate of introgression from domestic cats. This prior 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 \( (m_{ig}) \) 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 populations 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 1.

Data were simulated under this model 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 nonoverlapping, 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 \( n \). Captive individuals with a Q35 score (35 SNP test STRUCTUCE Q value) of <0.9 \( (n = 13) \) were filtered from the observed data. This functioned as a proxy for the selection of putative wildcats for incorporation into the captive breeding programme, in the model migrants were selected at random. The total number of simulations used for ABC was 509,070.

Given the strong separation of domestic cats and wildcats across the first principal component (Figure 1a), 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 S3. The final number of summary statistics was 14 (dropping eight with a detrimental impact on model fit, see below). Owing to the correlation within and between parameters and summary statistics (Figure S2), 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 et al., 2013). To fit the regression model for the projection we chose 20% of simulated points that were closest to the observed set of summary statistics.

The final model parameters and summary statistics were decided via the process described in Supporting Information Box 1 and Figures S3–S5, which used the goodness-of-fit test included in the R package abc (Csilléry et al., 2012) and a novel method for dropping summary statistics to improve model fit. 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, for example, \( HPDD_{0.95} \) such that 95% of all simulated points have \( nnd > HPDD_{0.95} \). The nearest neighbour distances 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 repeated (permanently dropping the worst performing summary statistic from the previous round, PCA rotated and scaled, and Mahalanobis \( nnd \) recomputed) until \( nnd \) of the target > \( HPDD_{0.95} \).

Parameter inference was carried out in R using the package abc (Csilléry et al., 2012). The closest 5091 points (1%) were used to generate the posterior distributions, correcting for an imperfect match between the summary statistics and observed data using nonlinear regression (neural network) (Blum et al., 2013; Raynal et al., 2019).

2.5 | Using the wildcat model to calibrate tests for selection

A model for wildcats, such as the one described above, is important for understanding the demographic history of this species in Britain. It is also a valuable tool to understand other processes in wildcat or hybrid populations. For example, here we applied simulated data, generated under the best-fitting wildcat 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 et al., 2017) and bgc (Gompert & Buerkle, 2011). 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 of introgression as a function of genome-wide admixture. These methods were also applied to 10 simulated data sets selected at random from the posterior distribution of the wildcat model.

For pcdapt the first three principal components were used in the analysis, following Cattell’s Rule that 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, these relate to SNPs with large variation in allele frequency between the parent populations (included in the analysis) and may therefore represent “wildcat” or “domestic” loci under selection in the hybrid population. p-values <1 × 10⁻⁶ were investigated as outliers (equivalent to 0.01 Bonferroni corrected).

Unlike pcdapt, bgc requires parental populations to be defined a priori. For this admixture Q scores (Q6546) were used to classify individuals as wildcat, Q6546 > 0.9, hybrid, 0.9 > Q6546 ≥ 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 data sets, 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 | RESULTS

The final data set included 108 individuals: four Scottish domestic cats and 104 putative wildcats (45 wild individuals and 59 from the UK captive population), genotyped at 6546 SNPs. Twenty-one
samples were excluded from the analysis to minimise relatedness in the data set and/or as a result of stringent filtering of missing data. Population summary statistics are given in Table 2.

### 3.1 | Population structure

Principal component analysis (PCA) (Figure 1a) 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) was observed between domestic cats and the captive wildcat population. The distinct PCA clustering and high F_{ST} values supports this as a cluster of putative wildcats. Most wild-living individuals were distributed across PC1, between these two groups, and are therefore considered putative hybrids. A much smaller proportion of the variance is explained by PC2 (2.8%) and PC3 (2.7%, Figure S6).

An ADMIXTURE model with two ancestral populations (Figure 1b, 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 values at K = 2 (Spearman’s r = .998, p < .001; Figure S7). Figure 1c 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. At K = 3 further clustering within the putative wildcats was observed, including within the captive population. Cross-validation error indicated the most likely value of K for the whole data set is 5 (Figure S8).

### 3.2 | Existing hybrid tests

ROC curves showed that both diagnostic tests performed well, with AUC values of 0.984 and 0.854 (Figure 2). The 35 SNP test (LBQ ≥ 0.75) outperformed the morphology-based test, with a low rate of both false positives and false negatives. 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.

### 3.3 | Demographic modelling

Our demographic model (Figure 3a) is capable of simulating data within the range of the observed data and the model fits these data well (Figures S9 and S10). The first two axes of the posterior predictive PCA plots (Figure 3b) 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) are shown in Figure 3c. The posterior mean for T_1, onset of gene flow from domestics to wildcats, was 3.3 generations (95% HPD: 1.21–5). For T_2, time the captive population was established, the mean was 19.3 generations (95% HPD: 9.4–30). Note that the estimate for T_1 was not constrained by the prior to any marked degree, whereas the historically informed prior for T_2 had 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) that is, for an individual selected at random from the wild-living population there is a 13% chance it is a domestic cat.

### 3.4 | Evidence for natural selection

In the observed data padapt reported three outlying SNPs most correlated with PC1 (Figure S11). bgc found the majority of SNPs (90.1%) to be in excess for α and/or β estimates; S901 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, six (0.09%) of these had outlying values of both α and β.
and $\beta$ (five with negative values of $\alpha$ and $\beta$, one with positive $\alpha$ and negative $\beta$). Two SNPs were found to be outliers by both pcadapt and bgc. These results are summarised in Table 3 and Table S4.

Simulated data produced a comparable number of outliers using both methods. With pcadapt, nine out of the 10 simulated data sets contained at least one SNP correlated with PC1 found to be outlying with respect to population structure (Table S5). The mean proportion of outlying SNPs across the 10 simulations was 0.1%. Similarly, using bgc, the mean proportion of SNPs with $\alpha$ and $\beta$ estimates in excess was 85.4%. A total of 4.3% and 3.9% of SNPs had outlying $\alpha$ and $\beta$ estimates, respectively and 0.2% were outlying for both $\alpha$ and $\beta$ estimates (Tables 3 and S6).

Outlier SNPs are candidates for loci under selection although extant 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 the LBQ <0.75 threshold, 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 from five main biogeographic groups: Iberia, Central Europe, Central Germany, Italy and the Dinaric Alps (Mattucci et al., 2016). 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 introgression.

4 | DISCUSSION

4.1 | Current status of the wildcat in Scotland

PCA and ADMIXTURE analysis (Figure 1) 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 reanalysed the 76 samples used by Senn et al. (2019) and an additional 53 individuals. We increased the resolution of the previous analysis by 3449 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 was 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 (Figure 1). This supports the conclusion of Breitenmoser et al. (2019) that at the current rate of introgression from domestic cats, the wildcat population in Scotland is at high-risk of extinction in the near-future.

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 took generation time for wild-living cats to be ~3 years (Beaumont et al., 2001; Nussberger et al., 2018). The $T_2$ posterior mean (3.3 generations, or ~10 years) is implausibly recent, yet extensive model-checking (Figures 3b, S3–S5, S9 and S10) suggests the model generally fits well. 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 (Easterbee et al., 1991; Hubbard et al., 1992; Kitchener, 1992) 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. Although no historical samples were included in this study, 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 the LBQ <0.75 threshold, and one as a domestic cat.
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.

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 30 generations or c. 100 years (Figure S10). Haplotype admixture (Gärke et al., 2012; Haasl & Payseur, 2011).
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); this work in whole genome sequenced individuals is now underway.

A recent hybridisation time for Scottish wildcats only seems likely in the face of high gene-flow from domestic cats. Our model estimates 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. This appears to contrast with the estimate of one wildcat to six unneutered hybrids reported by Breitenmoser et al. (2019); however, it should be noted that these hybrids will not be exclusively F₁s and will have lines of descent from the domestic population extending over many generations. 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.

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 et al., 2014), loaches (Kwan et al., 2014) and honey-bees (Pinto et al., 2005). Our results may also support a recent acceleration of hybridisation in Britain. Although 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 is the captive wildcat population. There is significant interest surrounding this population, which comprises individuals that are among the last putative wildcats in Britain, and 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 Figure 1 that hybrids are present, although the number appears to be low. 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 below). 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.

Captive individuals had a wide distribution across PC2 and PC3 (although this explained only a small proportion of the variation in the genetic data, 2.8% and 2.7%, respectively), and ADMIXTURE plots show clustering within the captive population (Figure 1b, \( K = 3 \)). The distribution of captive individuals across PC2 was a difficult feature to replicate in the model (Figure 3b). It is hard to disentangle the impacts of maintaining a (historically small) captive breeding population, for example, inbreeding, genetic drift, or adaption to captivity (Frankham, 2018; Woodworth et al., 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 (Figure S12) imply the distribution of individuals across PC2 or PC3 was not a gradient of inbreeding across the population.

Patterns relating to geographical origin in the wild samples were unclear due to the high levels of introgression (Figure 1c). In terms of introgression it seems clear there have been multiple admixture events, possibly due to the pervasiveness of domestic cats in wildcat habitat in Scotland and continuing high levels of persecution that maintained wildcat populations at low levels (Kitchener & O’Connor, 2010). 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 ~100 km 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 \( K = 3 \) were sampled from both sides of the Great Glen, so other geographical barriers may need to be considered and tested with additional sampling and modelling.

A second possibility is that admixture clustering at values of \( K \) greater than two reflect temporal patterns of hybridisation, that is, 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 $K = 3$.

### 4.2 A demographic model for wildcats

The demographic model developed for wildcats appeared to fit well (Figures S3–S5, S9 and S10), and the method to improve fit by dropping poorly performing summary statistics was effective (Figure S5). The two-step approach was computationally efficient, and overall, the model is a useful tool to understand the recent demographic history of the wildcat population in Britain.

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. We have shown that the model fits the key features of the data reasonably well. We have tried to focus on a simple model to achieve this. Further improvements may be possible, at the cost of increased parameterisation, by considering, for example, variable population size, which is possible even with unlinked SNPs (Hollenbeck et al., 2016); however, this was not an aim of the current study.

Quilodrán et al. (2020) use 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 (Figure 1c), it may be helpful in a future study 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.

Simulated data was applied to understanding methods for detecting selection in admixed populations, specifically pcmdapt (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), 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 therefore 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 et al., 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.

This analysis demonstrates that a wildcat-specific model of admixture is a useful tool to evaluate specific statistical approaches in genomic analysis and provides a useful baseline with which to develop scenarios of increasing complexity, for example, 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 et al., 2020). Furthermore, it will be straightforward to extend the approach to incorporate whole-genome data in the future.

### 4.3 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 6546 ddRAD SNPs and investigated the relative effectiveness of field-based tests in recovering this. An ROC analysis (Figure 2) 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 as the characteristics scored by this test are likely to be controlled by a limited number of genes (Cieslak et al., 2011; Ezirik 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) pelage scoring appeared to give a number of false negatives and false positives, that is, individuals with probable wildcat ancestry that did not necessarily score highly for wildcat features and vice versa. A more conservative threshold of 7PS ≥ 19 reduces the number of false positives but increases 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 6000 SNPs. Four false positives and two false negatives were identified, although 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.

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.

4.4  Conclusion

We found a population of putative wildcats persists in Scotland. These individuals were almost exclusively found in the UK captive population, which appears to have been established prior to significant recent admixture and is supported by accurate tests for hybrids. The captive population is an important resource for wildcat conservation in Britain. We found the wild-living population to be a hybrid swarm and almost all wild individuals sampled showed evidence of introgression from domestic cats.

ACKNOWLEDGEMENTS

This study represents further analysis of ddRAD data first published in Senn et al. (2019); the authors of that study are thanked. The authors would like to thank again the many people who have provided samples to this data set over the years. Thanks to those who have handed wildcat samples to the National Museum of Scotland; without these repeated individual efforts, these types of study are not possible. We are also extremely grateful for the assistance of the wildcat captive holding community in the UK. We thank Danielle Gunn-Moore at the University of Edinburgh for access to domestic cat reference samples. We also thank staff, volunteers and collaborators of Scottish Wildcat Action for their participation in sampling during this project and David Barclay for discussions relating to the studybook. Storage and archiving of the study samples at https://www.cryoarks.org/ is supported via BBSRC grant BB/R015260/1. Generation of ddRAD data was funded by RZSS and the Heritage Lottery Fund via grant to Scottish Natural Heritage. We would like to thank Greger Larson for valuable feedback on the manuscript. J. Howard-McCombe is supported by the NERC Doctoral Training Partnership, with additional funding from the People’s Trust for Endangered Species.

AUTHOR CONTRIBUTIONS

J. Howard-McCombe designed the research, analysed the data, and wrote the paper. H. V. Senn provided data for analysis. M. Beaumont, D. Lawson and H. V. Senn conceived the study and designed the research. D. Ward and A. C. Kitchener analysed the data. All authors critically reviewed the manuscript.

DATA AVAILABILITY STATEMENT

SNP data have been made available from the Dryad Digital Repository (Howard-McCombe et al., 2021; https://doi.org/10.5061/dryad.z34TMPGdj). Materials for demographic modelling available have been made available github.com/johowardmcc/wildcat_abc.

ORCID

Jo Howard-McCombe https://orcid.org/0000-0002-4009-3323

REFERENCES


**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.