**Simulation Based Inference of the Evolutionary History of Wildcats (*Felis silvestris*): Machine Learning for Population Genomics**

Harry Gordon\*, Daniel Ward\*, Jo Howard-McCombe, Daniel Lawson, Mark Beaumont

\*Joint

**Abstract**

Software for simulating the evolution of genomes has become increasingly widespread in the field of population genomics due to its numerous applications. One such application is for simulation-based inference, where simulated data are compared with observed data to draw conclusions about a system. A popular implementation of this is approximate Bayesian computation (ABC), but recently machine learning approaches to ABC have been developed and offer a more flexible and efficient option. The Scottish wildcat is a sub-population of the European wildcat (*Felis silvestris*) under threat of eradication by genetic swamping due to hybridisation with domestic cats (*F. catus*). Previous studies have used ABC with population genomics simulators to explore the extent of this hybridisation for a three-population model of domestics, Scottish *F. silvestris*, and a captive population of *F. silvestris*. This study expands the previously studied model to include European *F. silvestris* and African wildcats (*F. lybica*), and to use Sequential Neural Posterior Estimation (SNPE) to infer parameter distributions based on whole-genome data. Initial rounds of inference indicated some model misspecification, so we implemented methods for pruning summary statistics to improve model fit. This improved our single-round estimates of wildcat demography considerably, particularly regarding parameters describing ancient demography, which were previously unresolved. We argue that NPE is very sensitive to model misspecification and poorly fitting summary statistics. Our approach provides a valuable and flexible template for inferring models of evolution and demography based on whole genome sequence data, and we describe novel tools for tackling model misspecification.

**Keywords**

Population Genomics, Machine Learning, Simulation, wildcats, Bayesian Inference.

**1. Introduction**

In genetic data, patterns of mutation and inheritance give us clues as to the demographic history of the species. However, in isolation, it is hard to piece together exactly what these patterns indicate, so if one wants to infer a detailed account of the species’ demographic history, a more comprehensive approach is often needed. A common method of inference in population genomics is Bayesian Inference, for which the main aim is to find the posterior *P(θ|D)* - the probability distribution of model parameter values given some observed data. According to Bayes’ theorem (below), this distribution can be obtained by evaluating the joint distribution between the prior probability distribution (the prior belief of the probability distribution of *θ*) and the likelihood, which describes how likely each particular set of parameters is to produce the observed sample. In this way, previous estimates for parameters can be incorporated into the analysis and ‘updated’ with observed empirical data (Beaumont et al., 2002; Pritchard et al., 1999).

However, models in population genomics often possess a complexity that renders the likelihood intractable, meaning the explicit evaluation of the likelihood function is impossible. Approximate Bayesian Computation, or ABC, presents a solution to this problem by drawing samples from a simulator that implicitly defines the likelihood, an approximation of this likelihood using model parameters sampled from the prior, thereby generating the joint distribution between the likelihood and the prior and side-stepping the mathematically intensive step of explicit likelihood evaluation (Beaumont & Rannala, 2004). As high dimensional data are hard to handle computationally, summary statistics describing the simulated data are calculated and compared with that of observed data, and an algorithm that accepts models that are consistent with the observed data is implemented to obtain the posterior distribution.

For models of population genomics, simulation tools that can generate this likelihood have been in constant development for many years and are becoming widespread within the field. The Wright-Fisher model and it’s continuous-time counterpart, the coalescent are mechanistic models describing the genealogy of a population of individuals, and underpin all simulators of this nature (Barroso et al., 2020; Hudson, 2002; Kingman, 1982). The main applications of these simulations are in evaluating population genomics methods, investigating scenarios of evolution, and simulation-based inference (Hoban, 2014). Since their arrival, these simulators have improved massively in their efficiency and scope of possible models, which has allowed for increasingly complex analyses (Peng et al., 2015).

In ABC, the posterior is typically estimated by use of a rejection algorithm, which involves rejecting data that is not within a tolerance of the observed data and then weighting or adjusting this approximate posterior to improve bias (Beaumont & Rannala, 2004). Sequential Neural Posterior Estimation (SNPE) harnesses the flexibility of neural networks to greatly improve the accuracy of conditional density estimation and makes inferences based on all the simulated data without the need for rejection (Durkan et al., 2018; Greenberg et al., 2019; Lueckmann et al., 2017; Papamakarios et al., 2019; Ward, 2024). In this study we use normalising flows, one of the multiple variants of SNPE, to handle density estimation and sampling. Normalising flows learn a bijective transformation between our target distribution and a simple, known distribution such as a multivariate normal distribution. This way they can sample from the complex target distribution by sampling from a normal distribution and then applying the transformation to those samples (Papamakarios et al., 2021). These learned transformations can be conditioned using observed data to obtain samples from the posterior distribution that are consistent with our observed data. If carried out sequentially, SNPE can make robust inferences using relatively little simulated data when compared to ABC, as sequential simulations are sampled from the previous rounds of the posterior approximation, providing a narrower posterior distribution with each round. In modelling scenarios where there are many parameters it is likely that the posterior distribution is very concentrated compared to the prior, and thus a more focussed proposal distribution is important for making accurate approximations. The principal aim of this project was to develop a versatile and flexible inference framework using SNPE with simulated data, that can be applied to a range of population genetics scenarios. SNPE is a relatively recently developed tool in the field of statistics and is not yet widely used in scientific disciplines such as population genomics. Most studies that have used this method so far have investigated topics within subjects such as neuroscience (Gonçalves et al., 2020; Groschner et al., 2022) and physics (Akhmetzhanova et al., 2023; Furia & Churchill, 2022). However, very few studies, if any, have tried to use this approach sequentially to infer a model of evolution in population genomics, so it remains an untested tool in this discipline.

*Felis silvestris*, the European wildcat, is a species of wildcat native to many parts of continental Europe and Scotland. Due to decreased habitat range and population size, these wildcats can hybridize with feral domestic cats (*F. catus*) which has resulted in high levels of domestic cat genetic material in some populations (Yamaguchi et al., 2015). Hybridisation can be beneficial, as in the case of genetic rescue in small populations of highly inbred individuals. However, it is often a driver of species extinction due to genetic swamping, where rare genetic material is replaced by hybrid material, which is especially severe if the population size of the species at risk is far smaller than that of the invasive species. One population which suffers from extensive hybridization with domestic cats is the Scottish wildcat, which some have described as unviable and on the verge of eradication (Breitenmoser et al., 2019). This hybridisation has been driven by years of anthropogenic persecution and habitat destruction, and forces of selection such as disease have increased the rate of introgression (Beaumont et al., 2001; Howard-McCombe et al., 2023). To conserve the species and protect genetically pure individuals, a captive population of *F. silvestris* was founded with the aim of breeding and releasing pure individuals into the wild. However, even individuals in captivity have high levels of hybridisation and the rescue of this species remains a challenge.

Very little is known about the broader demographic history of European and African wildcats, despite recent advancements in this area (Jamieson et al., 2023). Presently, there are no estimates for the divergence time of African and European wildcats, and no estimates for the historic population sizes for these species. Current estimates for the time of cat domestication are based on archaeological evidence and therefore limited by preservation of materials (Hu et al., 2014). Furthermore, the divergence time between Scottish and continental European wildcats has only been inferred in the context of the geographic isolation of the British Isles (Mellett et al., 2012), without the use of genomic data. The secondary aim of this project is to use SNPE to infer a demographic model describing the flow of genetic material from *F. catus* to *F. silvestris* and wildcat evolution more broadly. Successfully inferring the timing and strength of this hybridisation would provide useful information to assist in conservation efforts and could be compared with existing estimates using other methods (Howard-McCombe et al., 2023). Inferring a broader demographic history for European and African wildcats would fill key gaps in current understanding and give a basis for a reintroduction programme for the Scottish population.

An earlier study used an ABC framework to model Scottish wildcat hybridisation (Howard-McCombe et al., 2021). The forward-time simulator SLiM (Haller & Messer, 2019) was used to simulate 500 generations of recent demography for domestic cats, Scottish wildcats and a captive population in a three-population model using reduced-representation sequencing data. We expand on this model to include continental European and African populations of wildcats, providing genetic reference points to more accurately inform a model of evolution, and providing posteriors for key demographic parameters such as timing of domestication, geographic isolation of the British isles, the divergence of European and African wildcats and effective population sizes for these ancestral populations. In this paper, we use a combination of SLiM and a coalescent simulator, msprime (Baumdicker et al., 2022), to model the demographic history of wildcats from the MRCA European and African wildcats to the present day. The motivation for this coupled simulation approach will be covered in more detail in the methods section. By using SNPE instead of traditional rejection ABC, we hope to make more efficient and robust inferences, using fewer simulations compared to the previous study (<30,000 compared to >500,000), and using all the simulated data to condition the posterior, rather than taking a percentage of the data within a tolerance of the observed data. Using the Flowjax package (Ward, 2024) for SNPE will have the added benefit of simplifying the process of inference from a relatively involved rejection and regression algorithm to one which is more hassle free. We also build on methods for mitigating model misspecification used in the previous study, which uses an iterative algorithm for dropping summary statistics affecting model fit. Finally, we use whole genome wildcat data, rather than reduced representation sequencing data, to improve the statistical power and accuracy of our analyses.

**2. Methods**

*2.1. Model of wildcat demography*

The model outline was derived from a previous study on Scottish wildcat introgression by Howard-McCombe *et al.* in 2021 which defined a three-population model. Our model (*Fig. 1*) is an expansion of this and features two more populations: A European population of *F. silvestris* and African wildcats (*F. lybica*) which are ancestral to domestic cats. For purposes of simulation, there are parameters for:

* 5 modern population sizes: *F. lybica* (N5), Scottish *F. catus* (N4), and European (N3), Scottish (N2), and captive Scottish (N1) populations of *F. silvestris*.
* 2 ancestral populations: *F. lybica* (N7) and *F. silvestris* (N6).
* 4 population divergence times representing the divergence of: *F. lybica* (T1), *F. catus* (T2), Scottish *F. silvestris* (T3), and captive Scottish *F. silvestris* (T4) populations.
* Migration between Scottish wildcats and domestic Scottish cats. Including the migration rate (M1) and duration (t) from domestic to wild-living *F. silvestris* and the rate from wild-living *F. silvestris* to the captive population (M2) to model the introduction of progressively more hybridized individuals to the captive population.
* Mutation rate (m) and recombination rate (r) for all individuals.

All population sizes, divergence times, and rates should be thought of as effective sizes, effective times, and effective rates as our model is a simplified version of wildcat demography and does not exhaustively parametrize all the forces acting on the wildcat genome.

*2.2. Prior distributions*

To generate the genomic data to be used in training the neural network, we simulated genomes according to parameters sampled from prior probability distributions, i.e. a previous belief of the parameter values. Sampling from these distributions generates a range of genomic data that describes all the possible demographic models supported by our prior beliefs. Table 1 specifies the prior probability distributions for each parameter and the source or justification for each prior. Many of the priors are based upon the posterior distributions obtained from a previous study of simulation based inference using ABC on RADSeq wildcat data (Howard-McCombe et al., 2021). In general, priors were chosen to be reasonably wide, i.e. conservative estimates of current belief, causing some initial simulations, modelling a combination of large parameters, to take a long time to compute. Regression of runtimes was conducted on a sample set of 5000 simulations to determine the causes of simulation time, and it was found that most population sizes and the divergence time of *F. lybica* significantly affected simulation time; In general terms, simulations took longer to compute the more individuals are modelled and the further back in time modelled. After simulating the first round of data points (~10,000) the narrower posterior from this round will be used as the prior for the next, meaning fewer models with large populations and early divergences would be simulated. Thus, long simulation times caused by wide priors may only be present for the first round.

*2.3. Genomic data*

We used the whole genome SNP data for the E3 wildcat chromosome of 112 individuals from 5 different populations. These individuals included: 65 captive Scottish *F. silvestris*, 22 Scottish *F. silvestris*, 15 European *F. Silvestris*, 6 Scottish domestics, and 4 *F. lybica*. The E3 chromosome was used as it is the smallest of the cat chromosomes (~45Mb) and chromosomes larger than this would drastically increase simulation times and therefore reduce the efficiency of our inference framework. Using VCFtools (v0.1.16), the genomic data was filtered to remove sites with missing data and sites with a minor allele count (MAC) of two or less (removing singletons and doubletons) to remove potential sequencing errors or sites with low statistical power (Danecek et al., 2011). As our simulation software only simulates mutations with one alternate allele (variants 0 and 1), we also removed sites in the genomic data with more than one alternate allele. This left approximately 370k SNPs. The WGS data was generously provided by authors of recent studies on wildcats who used BGISEQ and Illumina methods to sequence samples from a variety of sources (Howard-McCombe et al., 2023; Jamieson et al., 2023).

*2.4. Coupling forward-time and coalescent simulators*

Simulations were carried out using a coupled framework of SLiM (v4.0.1) (Haller & Messer, 2023), in forward-time and msprime (v1.2.0) (Baumdicker et al., 2022) in the coalescent. The motivation for this double framework is a concern that the coalescent is not a good model for recent pedigree structure (Wakeley et al., 2012). Concomitantly, forward Wright-Fisher modelling requires a suitable prior distribution for the gene frequency. Therefore, our approach simulates the complex recent history of wildcat hybridisation in forward-time, using the coalescent as a prior for initial gene frequency. We note that recent changes in msprime now allow for this within a single simulation. (find citation)

Four starting populations were established and modelled by SLiM forwards in time from 100 generations in the past, with the captive population diverging after this. SLiM simulated a 45 Mb genome (length of the E3 wildcat genome) under our demographic model to the present-day generation. This ‘decapitated’ tree was then passed to msprime, which started at 100 generations in the past and simulated the genome in the coalescent according to our model backwards in time to the *F. lybica* divergence time, ‘recapitating’ the tree. Mutations were then generated and ‘overlaid’ onto the tree by msprime. The simulations were carried out in terms of generations, which are approximately 3 years for wildcats (Beaumont et al., 2001). During simulations, SLiM and msprime recorded ancestry using ‘tree sequences’. This is a data model created by the authors of msprime which records local ancestry at SNPs along the genome, which reduces the memory requirement to handle and store data, and provides efficient calculation of descriptive statistics (Kelleher et al., 2018). After simulation, the tree sequence was simplified to include only the ancestry of a sample set matching our observed dataset and a MAC filter was applied to remove singletons and doubletons to match our observed data. To obtain simulated data and observed data in the same format for inference, tsinfer (v0.3.1) was used to infer the succinct tree sequence for the observed genomic data (Kelleher et al., 2019).

The simulations were carried out using the University of Bristol’s High-Performance Computing cluster, which allowed as many as ~400 simulations to run in parallel. (For the first round) 54Gb of memory was allocated for each simulation and a single processor was used. Simulations taking longer than 8 hours were aborted, as simulations exceeding this had combinations of very large populations and very early divergence times which were unlikely to be consistent with our observed data. Approximately 1% of simulations in the first round were discarded due to the time limit.

*2.5. Summary statistics*

To reduce the dimensionality of the data so it can be used by the neural network, descriptive summary stats were calculated for the tree sequences obtained from simulations and observed data. Overall, 135 summary statistics were calculated. These included:

* Diversity (Nei & Li, 1979), number of segregating sites, Tajima’s D (Tajima, 1989), divergence (Nei & Li, 1979), genetic relatedness (Speed & Balding, 2015), Patterson’s f statistics (Reich et al., 2009), Y statistics (Jaime et al., 2018), and Fst (Holsinger & Weir, 2009), all efficiently computed directly from the tree sequence by tskit (v0.5.5) (Kelleher et al., 2018).
* PCA median, inter-quartile range, and pairwise distance summary stats using scikit-learn (v1.2.2) (Pedregosa et al., 2011). For the calculation of PCA statistics, the 012 genotype matrix was generated from the tree sequence, which required a large amount of memory. Inference procedures like this, which are affected by simulation times, would be greatly benefitted if tskit possessed built-in features for calculating PCA statistics directly and efficiently from the tree sequence.

For each simulation, summary stats were calculated for all populations and collated along with the corresponding parameters into a reference table and used as input for SNPE.

*2.6. SNPE procedure*

For inferring the posterior distribution, the Flowjax package (v12.0.1) was used to carry out SNPE (Ward, 2024). For stability during training, parameters were first log transformed and then normalised using affine transformations as outlined in the Flowjax documentation. Summary stats were also normalised using the same affine transformation method as carrying out inference without normalisation of the simulated data caused the neural network to fail to converge. The same affine transformation was used in subsequent rounds. The prior distribution provided to SNPE was a multivariate normal PyTorch (v2.0.1) distribution with mean 0 and standard deviation 1 (Paszke et al., 2019). The neural network was then trained with the first dataset of simulated data (10,000 simulations) to learn relationships between summary stats and parameters. The validation proportion was 0.1 (1000 out of 10,000 each round), the batch size was 100 and a learning rate of 1e-4 was used. After training, the estimated distribution of parameters was conditioned using the observed data, and then sampled from to obtain parameters consistent with our observed data. 10,000 sets of model parameters were sampled from this posterior, and inversely transformed and exponentiated to obtain the posterior distributions and the parameters for the next round of simulation. For the inference to be carried out sequentially, it was necessary to serialise the previous round’s normalising flow and carrying this forward to be used in following rounds (Kidger & Garcia, 2021) retaining information from previous rounds and preventing SNPE from overconfidently estimating a new posterior based on resampled data. The basic outline of this approach is illustrated in Figure 2. Posterior predictive checks were carried out to assess whether sequential inference was feasible.

*2.7. Addressing model misspecification*

Initial rounds of inference using the full set of summary statistics yielded posteriors that had strong indications of model misspecification. These were characterized by overconfident first round posteriors followed by ‘exploded’ posteriors in subsequent rounds that were far wider than the prior. Prior and posterior predictive checks revealed that the observed data was far outside of the prior and posterior support for many summary statistics. For example, the observed Tajima’s D statistic for the European population of *F. silvestris* was negative (~-0.3), and well outside of the support of the simulated data, which ranged from 0.5 to 2.0. The negative value of Tajima’s D indicates a disproportionate level of rare alleles which can be a sign of population growth (Tajima, 1989), admixture (Stajich & Hahn, 2005), or a structured population (Wakeley, 1999). However, this demography is not consistent with current understanding (Langley & Yalden, 1977; Pierpaoli et al., 2003) and it would be difficult to replicate this statistic in the simulated data. To provide flexibility in the model, mutation and recombination rates, which were originally point values, were given prior distributions and included in the parameters to be inferred. This also has the added benefit of potentially providing a posterior estimate for these values, which do not currently have good estimates in the literature.

To address the clear misspecification in the initial set of 134 summary statistics, we implemented three methods to remove the statistics for which there was the largest discrepancy between the simulated and observed data.

1. **Removal of summary statistics with >0.99 correlation.** Pairwise correlations were calculated and one of the correlated stats were removed to limit the maximum correlation in the set of summary statistics to 0.99. 24 summary statistics were removed using this method
2. **Noise detection using normalising flows.** If we assume, in the misspecified case, that the observed data we are using has been generated according to our model plus some noise in the sample i.e. p(xobs) = p(x) + noise, we can attempt to infer a posterior over the noise to identify misspecified summary statistics. A flow was trained on 10,000 simulations from the prior to approximate the density function of the summary data, and MCMC was used to fit a noise model over the data with a Laplace distribution (μ=0, b=1). 10 summary statistics were removed using this method, with an absolute noise threshold of 0.8.
3. **Iterative improvement of HDR (highest density region) of observed data.** Previous studies have iteratively dropped summary statistics based on nearest neighbour distances, however, this method yielded inconsistent results with our data. Instead, we trained normalising flows on 10,000 simulations to approximate the joint distribution between all the summary statistics, and then evaluated the log probabilities of 5000 held-out simulations and the observed data. Thus, taking the rank of the observed data within these held-out simulations yielded the HDR of the observed data (between 0 and 1). To achieve consistent results in training the flows, we used a batch size of 25 and a learning rate of 5e-5. Using this method, we estimated observed percentiles for 100 sets of summary statistics (each with one stat dropped) and accepted the set with the lowest percentile i.e. the most specified set. This was repeated 7 times until an observed percentile of 0.78 was achieved (the best percentile achieved using this method). 7 summary statistics were removed in total. This method for evaluating the HDR of the observed data was also used as a posterior predictive check to evaluate the viability of SNPE (Gabry et al., 2019).

This multi-faceted approach reduced our initial set of 134 summary statistics down to 93 statistics, minimising the effects of model misspecification. The statistics removed can be found in the supplementary materials (Table S1).

**3. Results**

*3.1. Posterior distributions*

After one round of inference, the marginal posteriors estimated were narrower than priors for all but one parameter (*Fig. 3, 3i*). However, evaluation of the percentile of the observed data within the simulations sampled from the first round posterior revealed a percentile of 1.0, meaning the observed data had the lowest posterior probability out of 5000 posterior samples. This indicates some model misspecification still remains which is inhibiting robust sequential inference. Despite this, our first round marginal posteriors showed no sign of obvious model misspecification, and were a huge improvement on the first round posteriors using the full set of summary statistics (*Fig*. *3, S1*).

Our posterior for the establishment of the captive population of Scottish wildcats was slightly narrower than our prior, and shifted to an earlier establishment time centred around a posterior mean of 25.3 generations (75.9 years) (*Fig. 3a)*. The posterior for *F. lybica* divergence time (MRCA of European and African wildcats) was narrower than our prior, and the posterior mean was shifted to an earlier divergence, distributed around a posterior mean of 104k generations ago (~316 kya) (*Fig. 3b*). The posterior for domestic divergence time (domestication) was narrower than our prior and shifted marginally to an earlier divergence, with a posterior mean of 4170 generations (12.5 kya) (*Fig. 3c)*. Our posterior for the divergence of the Scottish *F. silvestris* population was narrower than our prior, and shifted to a more recent divergence around a posterior mean of 2900 generations (8.7 kya) (*Fig. 3d*).

All our parameters describing migration of genetic material from domestic cats to wild living Scottish wildcats, and subsequently from wild-living to captive Scottish wildcats, were narrower than our priors and were centred around similar values to our priors (*Fig. 3e, 3f, 3g*). The posteriors for captive and wild migration rates were centred around means of 0.0373 and 0.0464 migrants per generation respectively, and the length of migration from domestics to wild-living Scottish wildcats was centred around a mean of 14.2 generations (42.6 years).

The posterior for the effective population size of the captive population of Scottish wildcats was only slightly narrower than our prior, and was centred around 113 individuals (*Fig. 3h*). The posterior for the wild-living Scottish wildcat population was narrower than our prior and shifted slightly to a larger population size around a posterior mean of 3390 individuals (*Fig. 3l)*. The posterior for domestic cat population size was the only parameter with a posterior that was wider than our prior (*Fig. 3i*). It was shifted considerably to a larger population size, around a posterior mean of 185k individuals. Posteriors for effective population sizes of contemporary and ancestral African wildcat populations were both narrower than our priors (*Fig 3j, 3k*). The contemporary African wildcat population was shifted to a smaller population size around a posterior mean of 4150 individuals, and the ancestral population was shifted to a larger population size of 107K individuals. The posteriors for the population size of contemporary and ancestral European wildcat populations were both narrower than our prior, however the contemporary population size was only slightly narrower (*Fig. 3m, 3n*). Both contemporary and ancient population size posteriors were shifted to slightly larger sizes, centred around 27.3k and 47.7k individuals respectively.

Finally, our posteriors for mutation and recombination rates were both narrower than our priors, and had posterior means of 1.03e-8 mutations per site per generation and 1.39e-8 centimorgans respectively (*Fig. 3o, 3p*). There was correlation in the posterior density between some pairs of parameters, indicated by linear relationships in joint distributions; These parameters have been plotted as a pairplot (*Fig 4*).

*3.2. Posterior mean model of evolution*

The means were evaluated for all marginal parameter posteriors, and this posterior mean model of evolution was simulated. PCA was carried out on the resulting genotype matrix and that of the observed data, and principal components 1 and 2 were plotted for comparison as part of a posterior predictive check . Procrustes rotation was applied to the simulation to provide a better comparison between the simulated and observed populations. (<- should this bit be in methods?) Overall, there is very good concordance between the simulated and observed populations. All populations show a very similar pattern across PC1 and PC2, with the possible exception of the European *F. silvestris* population, which is clustered slightly further on both axes from Scottish *F. silvestris* in the MAP simulation than in the observed data PCA plot. Both plots show a ‘smear’ of increasingly hybridised individuals from the Captive population cluster towards the domestic cluster.

**4. Discussion**

Plan

* NPE unable to make estimates outside of prior support
* Talk about posteriors and implications for evolution/conservation etc, widen prior for domestic pop size maybe
* Talk about model misspecification and sensitivity to summary statistics, how to improve choice of summary statistics
* Talk about possible sources, e.g. pca noise, low sample size, unrepresentative samples, missingness, population structure,
* Removing summary stats – all models are wrong but some are useful – George Box
* Talk about methods to remove summary stats, how to improve
* Talk about how posteriors improved with removal of summary stats
* Correlation of parameters

As the HDR of the observed data was evaluated as 1.00 after the first round of inference, further rounds were not feasible given the apparent misspecification remaining and the poor ability of neural density estimators to condition the posterior towards an observation outside of prior support (citation?) (maybe put this after main discussion of posteriors). However, after one round of inference, informative posteriors were recovered for all parameters in our model of wildcat evolution.

*4.1. Conclusions for demography*

Firstly, the distributions for migration rate and the duration of migration of genetic material from domestic cats to wild *F. silvestris* in Scotland were considerably narrower than our prior distributions, and are ecologically significant (*Fig. 3f, g*). The posterior mean for rate of migration estimated at 0.0464 migrants per generation and the date for onset of this migration estimated at 14.2 generations (~43 years) is a considerably lower rate and earlier onset than a previous estimate by a paper using a similar method of 13% migrants for 3.3 generations (Howard-McCombe et al., 2021). It should be noted that these two different findings represent a similar effective level of migration, although over different timeframes. The date for onset of migration estimated here is consistent with a recent paper using a different approach, which estimated the onset of hybridisation to have occurred approximately between 1970 and 2000 (Howard-McCombe et al., 2023). Both the migration rate parameter posteriors were estimated with higher density than our priors, which suggests this method is viable for quantifying hybridisation. Another potentially meaningful insight into the recent demography of *F. silvestris* is that the posterior mean for the effective date for onset of hybridisation from domestic cats (11.6 generations) was inferred as simultaneous with that of the effective date for the establishment of the captive population (11.7 generations), which is broadly in line with findings of a previous study which dates the onset after the establishment of the captive population (*Fig. 3d, l*) (Howard-McCombe et al., 2021). This finding suggests that the captive population was established before much of the hybridisation between domestics and wildcats occurred, emphasizing the importance of this captive population as a resource for repopulation.

The posterior for the domestic cat effective population size was also narrower than the prior, with a posterior mean of ~110k individuals (*Fig. 3h*). This is far lower than the current census estimate of population size for domestic cats in Scotland of 685k, however, it has been shown that the ratio of effective population size to census population size (Ne/N) for domestics can be as low as 20% in feral populations (CATS Report Scotland, 2021; Kaeuffer et al., 2004). The posterior for effective population size of Scottish wild *F. silvestris* was the only parameter that was estimated with higher density after two rounds of inference than both the prior and first-round posteriors, indicating some ability of this method to sequentially provide better estimates for parameters (*Fig. 3f*). Although only slightly narrower than the prior, the European *F. silvestris* effective population size posterior mean of 15.3k individuals suggests, despite their decline, this European wildcats are still relatively numerate (*Fig. 3g*) (Eckert et al., 2010).

Our second round posteriors for the effective divergence times of domestic cats (*Fig. 3b*) and Scottish *F. silvestris* (*Fig. 3c*) were both distributed almost identically to our priors for these parameters indicating a limited ability of our model to estimate these with any more precision, although both of the posterior means were closer than our priors to the divergence times estimated in the literature (Driscoll et al., 2007; Mellett et al., 2012). Similarly, posteriors for *F.* lybica divergence (*Fig. 3a*), captive population size (*Fig. 3e*), mutation rate (*Fig. 3o*), and recombination rate (*Fig. 3p*) were not estimated with any more density than our priors, highlighting the inability of our current method to provide good posteriors for all parameters. It was clear that our method was entirely unable to recover any reasonable estimate for ancestral *F. silvestris* and *F. lybica* population sizes (*Fig. 3j, k*), as the posteriors were especially wide when compared to our prior, suggesting our approach struggles particularly with inferring ancient demography. Despite the shortfalls of this method in inferring narrow posterior distributions for many parameters, some posterior distributions were estimated somewhat precisely, demonstrating the potential benefits of this approach. The following discussion will focus primarily on potential reasons for the failure of our method to provide good posterior estimates for all parameters.

*4.2. Choice of summary statistics*

In traditional ABC approaches, much importance was placed on the selection and optimization of summary statistics. This was mainly due to the unnecessary increase in dimensionality caused by uninformative summary statistics, which resulted in less precise posterior estimates. As a result, much time and effort was taken to develop techniques to optimize the summary statistics used (Aeschbacher et al., 2012). In our approach, however, unoptimized summary statistics were not such a barrier to inference. MAF posterior estimation methods have been shown to be able to learn relevant features of the data and place lower importance on features that are uninformative (Greenberg et al., 2019). Consequently, one advantage of our approach is that the only drawback of using a very large number of summary statistics is the extra time and memory it takes to calculate and handle these statistics. The presence of non-identifiable parameters in our first round posteriors (*Fig. 4*) indicates that the features included in our range of summary statistics may not have been informative enough to infer narrow posterior distributions in these parameters. The inclusion of more summary statistics to attempt to differentiate between these parameters could be a solution worth exploring for inferring better posteriors because if they turn out to be uninformative, it would have little to no effect on posterior estimation.

* Rewrite this paragraph

*4.3. Possible model improvements*

The correlation between the mutation rate (m), ancestral *F. silvestris* (N6) and *F. lybica* (N7) population size, and *F. lybica* divergence (T1)first-round posteriors could be driven by the lack of differentiating summary statistics (*Fig. 4*). Summary statistics dependent on these parameters would include the number of segregating sites and diversity, as large ancestral populations, early divergences, large mutation rate or any combination of these could all cause higher diversity and numbers of segregating sites observed in current populations. Inversely, small ancestral populations, recent divergences, and a small mutation rate could each result in low diversity and number of segregating sites in current populations. A simple solution to this problem would be to provide point values for parameters, e.g. mutation rate, that can be estimated from related studies, which would reduce the uncertainty in the model and allow the estimator to infer better posteriors, however, as discussed previously, this is not always possible if flexibility in the model for such parameters is necessary to obtain an adequate fit for the observed data. As previously mentioned, another solution to explore would be the addition of summary statistics that improve the separate identifiability of these parameters, although this may not be achievable, as signals in genetic data tend to be far weaker and more cryptic for ancient demography compared to recent demography. Migration rate and duration of migration from domestics to Scottish Wild *F. silvestris* (M1, t) were also very slightly correlated and, thus, estimates may be improved by the implementation of the same changes (*Fig. 4)*.

As Tajima’s D is most likely an informative feature of the genetic data, it is possible that posterior estimates of European *F. silvestris* population size could be significantly improved with the inclusion of the Tajima’s D statistic for this population. The statistic was removed as the observed value (-0.3) fell well outside of the support of the simulations, indicating this statistic could negatively affect posterior estimates given our simplistic model. To obtain the most accurate posterior estimates possible, future studies of wildcats using this approach could investigate the population structure of European *F. silvestris* to find the cause of this value, with the aim of adding events to the demography of European *F. silvestris* to allow the model to fit this summary statistic. Negative values of Tajima’s D are sometimes observed in populations that have recently experienced a bottleneck (Tajima, 1989), so a possibility future studies could explore is including this European bottleneck in the model for simulation. However, it remains likely that this statistic is due to some cryptic demography that may be hard to replicate in simulations. If this is the case, fitting a model using a spatial simulator may provide insight into this European demography, as the populations are highly fragmented in some areas and spatial simulations could provide a better fit for the observed data for these populations. In particular, it has been shown that Scottish wildcats are most genetically similar to wildcats in north-western regions of continental Europe, so a possible spatial model that could describe this association well would be one that differentiates between subpopulations of European *F. silvestris* and implements a partial reproductive barrier separating the north-western population and having the Scottish population diverge from this (Neaves & Hollingsworth, 2013). An approach like this would require observed samples representing each subpopulation of European *F. silvestris*. Using spatial simulators could also provide a better fit for describing extent of hybridisation in Scotland, as these populations are also highly fragmented, which is associated with increased levels of hybridisation (Beugin et al., 2020).

When inferring a model for scenarios of evolution like this, there is often a trade-off between more simplistic models that have fewer variable parameters, and more complex models with more variable parameters that allow for more flexibility. Simpler models have less uncertainty so can potentially provide better posterior estimates if the model adequately fits the observed data. However, if the model is too simple to provide an adequate fit, the model is considered to be misspecified, which can often completely inhibit the performance of simulation-based inference methods using neural density estimators (Cannon et al., 2022). Although the PCA for our first round MAP model implies to some extent a good fit of our model to the observed data (*Fig. 5*), the second round posteriors clearly indicate that there is some level of misspecification, so future studies investigating this model should consider improving the model fit by adding population changes, bottlenecks, further migration, and parameters to spatially model potentially fragmented populations.

Correlation of these parameters indicates that the data may lack features that uniquely describe these parameters, and features determining one parameter will also determine the other (Li & Vu, 2013).

Use Procrustes rotation on PCA for pca based stats to reduce noise.

*4.4. Genomic data*

In the present study, we used whole genome data from multiple sources, meaning there were slight differences in the sequencing of some individuals, specifically a higher level of missing data in a few samples. Removing sites with missing data decreased the number of SNPs in our dataset by about ~26%, which could represent a significant loss of statistical power and present issues of model misspecification with regard to the ‘segregating sites’ summary statistic. Filtering multiallelic sites also removed a small number of sites (~0.7%). Future studies aiming to fit a model of demography to genomic data should use whole genome data with as close as possible to the full number of SNPs in the genome. This would involve using genomic data with a very low level of missingness and recovering multiallelic sites to multiple biallelic sites that match the simulated data and implementing a multiallelic alleles into the simulations to best match the observed data. Additionally, the possibility of introducing artificial genotyping errors to simulated data could be explored, as some studies have found data simulated this way can provide a better fit to observed genetic data when compared to data that has been trimmed to match filtered genetic data (Jay et al., 2019). Finally, as for any study using genetic data, obtaining greater numbers of samples for each population would give a more representative data set and improve the accuracy of summary statistics calculated, therefore providing better inferences.

Interestingly, the posteriors for the divergence and population size of the captive Scottish wildcats were both very similar to our priors and only marginally narrower. Given that this was the population for which we had the most samples (65) ….

*4.5. Simulation approach limitations*

It should be highlighted that a huge advantage of a multi-round approach like this one is that of preventing needlessly simulating lots of models that are well outside the eventual posterior density, so are not informative in inferring this posterior. This advantage is especially important if the model under investigation, like ours, involves many individuals that are traced all the way back to ancient divergence times, as the wall times for these simulations can be prohibitively long. In our study, persistent teething problems and these long simulation times meant that the eventual number of simulations was limited to ~20,000 (two rounds of SNPE). This is clearly a limitation of this study, and future studies should aim to iron out simulation difficulties and model misspecification as early as possible to allow for the simulation of a sufficient number of data points for robust inference. Nevertheless, this study demonstrates the possible advantages of resampling from the posterior for simulating data, and future studies may be able to make good inferences based on far fewer than the hundreds of thousands of simulations often needed for traditional ABC approaches with models of this type (Howard-McCombe et al., 2021). A resource that was invaluable during this study was the high-performance computing cluster of the University of Bristol, which allowed for up to ~400 simulations running simultaneously. Any future effort to develop this approach should aim to maximise the parallelisation offered by computing clusters and minimise the runtimes of simulations. As they are developed, tools for simulation will become more efficient and therefore increase the speed of analyses like this. However, future studies should take into account that analyses of demography over shorter time frames and with smaller population sizes will be faster to simulate data for, and therefore be less affected by common issues such as initial model misspecification problems.

*4.6. Final Conclusions*

The first-round distributions demonstrate that this method has the potential to estimate informative posterior distributions for parameters in models of this type and complexity. However, the power of posterior estimation can be limited by issues of model misspecification. Studies hoping to infer a model of evolution for a species like the wildcat should carefully consider the specification of their model and focus initial efforts on finding a demographic model that describes more completely the patterns of genetic data in their observed samples while being careful not to over-parametrise. We have shown that there can be difficulty in recovering reasonable distributions for mutation rate and recombination rate, so future studies should treat these as point parameters, and aim to obtain a good fit for their data through parametrisation of demographic events. Future efforts to investigate hybridisation in Scottish wildcats should also explore the incorporation of spatial aspects to models of demography to investigate and model the effect of fragmentation in the Scottish population.

As the efficiency and flexibility of simulation software improve, and high throughput facilities become more advanced, inference procedures like ours will become more efficient, informative and widespread. However, the scarcity in the literature of simulation-based inferences that use SNPE, especially within population genomics, makes the common pitfalls of such approaches difficult to avoid. Therefore, transparency in matters of model misspecification in such studies, especially with regard to the steps taken to mitigate it, would be beneficial in the wider understanding of both the true effect and potential solutions of the issue. We argue that the success of studies like this hinge greatly on the ability of the model to adequately describe the observed data.

* Future studies should try identifying samples that might be contributing to model misspecification.

The code written for this project will be available at: <https://github.com/HGord2022/wildcats>.

**References**

Aeschbacher, S., Beaumont, M. A., & Futschik, A. (2012). A novel approach for choosing summary statistics in approximate Bayesian computation. In *Genetics* (Vol. 192, Issue 3, pp. 1027–1047).

Akhmetzhanova, A., Mishra-Sharma, S., & Dvorkin, C. (2023). Data Compression and Inference in Cosmology with Self-Supervised Machine Learning. In *arXiv preprint arXiv:2308.09751*.

Barroso, G. V., Moutinho, A. F., & Dutheil, J. Y. (2020). A Population Genomics Lexicon. In J. Y. Dutheil (Ed.), *Statistical Population Genomics* (pp. 3–17). Springer US. https://doi.org/10.1007/978-1-0716-0199-0\_1

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., … Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. In *GENETICS* (Vol. 220, Issue 3). https://doi.org/10.1093/genetics/iyab229

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. https://doi.org/10.1046/j.1365-294x.2001.01196.x

Beaumont, M., & Rannala, B. (2004). The Bayesian revolution in genetics. In *Nature Reviews Genetics* (Vol. 5, Issue 4, pp. 251–261). https://doi.org/10.1038/nrg1318

Beaumont, M., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. In *Genetics* (Vol. 162, Issue 4, pp. 2025–2035).

Beugin, M., 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. In *Ecology and Evolution* (Vol. 10, Issue 1, pp. 263–276).

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. In *IUCN SSC Cat Specialist Group*.

Cannon, P., Ward, D., & Schmon, S. M. (2022). Investigating the impact of model misspecification in neural simulation-based inference. In *arXiv preprint arXiv:2209.01845*.

*CATS Report Scotland*. (2021). [Report]. Cats Protection.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & Genomes Project Analysis, G. (2011). The variant call format and VCFtools. In *Bioinformatics* (Vol. 27, Issue 15, pp. 2156–2158). https://doi.org/10.1093/bioinformatics/btr330

Driscoll, C. A., Menotti-Raymond, M., Roca, A. L., Hupe, K., Johnson, W. E., Geffen, E., Harley, E. H., Delibes, M., Pontier, D., & Kitchener, A. C. (2007). The Near Eastern origin of cat domestication. In *Science* (Vol. 317, Issue 5837, pp. 519–523).

Durkan, C., Papamakarios, G., & Murray, I. (2018). Sequential neural methods for likelihood-free inference. *arXiv Preprint arXiv:1811.08723*.

Eckert, I., Suchentrunk, F., Markov, G., & Hartl, G. B. (2010). Genetic diversity and integrity of German wildcat (Felis silvestris) populations as revealed by microsatellites, allozymes, and mitochondrial DNA sequences. In *Mammalian Biology* (Vol. 75, pp. 160–174).

Furia, C. S., & Churchill, R. M. (2022). Normalizing flows for likelihood-free inference with fusion simulations. In *Plasma Physics and Controlled Fusion* (Vol. 64, Issue 10, p. 104003).

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian Workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Gonçalves, P. J., Lueckmann, J.-M., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., Chintaluri, C., Podlaski, W. F., Haddad, S. A., Vogels, T. P., Greenberg, D. S., & Macke, J. H. (2020). Training deep neural density estimators to identify mechanistic models of neural dynamics. In *eLife* (Vol. 9, p. e56261). https://doi.org/10.7554/eLife.56261

Greenberg, D., Nonnenmacher, M., & Macke, J. (2019). *Automatic posterior transformation for likelihood-free inference* (pp. 2404–2414). PMLR.

Groschner, L. N., Malis, J. G., Zuidinga, B., & Borst, A. (2022). A biophysical account of multiplication by a single neuron. In *Nature* (Vol. 603, Issue 7899, pp. 119–123).

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. In *Molecular Biology and Evolution* (Vol. 36, Issue 3, pp. 632–637). https://doi.org/10.1093/molbev/msy228

Haller, B. C., & Messer, P. W. (2023). SLiM 4: Multispecies eco-evolutionary modeling. In *The American Naturalist* (Vol. 201, Issue 5, pp. E127–E139).

Hoban, S. (2014). An overview of the utility of population simulation software in molecular ecology. In *Molecular Ecology* (Vol. 23, Issue 10, pp. 2383–2401). https://doi.org/10.1111/mec.12741

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: Defining, estimating and interpreting F ST. In *Nature Reviews Genetics* (Vol. 10, Issue 9, pp. 639–650).

Howard-McCombe, J., Jamieson, A., Carmagnini, A., Russo, I.-R., Ghazali, M., Campbell, R., Driscoll, C., Murphy, W., Nowak, C., & O’Connor, T. (2023). *Genetic swamping of the critically endangered Scottish wildcat was recent and accelerated by disease*.

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. In *Molecular Ecology* (Vol. 30, Issue 15, pp. 3688–3702). https://doi.org/10.1111/mec.16000

Hu, Y., Hu, S., Wang, W., Wu, X., Marshall, F. B., Chen, X., Hou, L., & Wang, C. (2014). Earliest evidence for commensal processes of cat domestication. In *Proceedings of the National Academy of Sciences* (Vol. 111, Issue 1, pp. 116–120). https://doi.org/10.1073/pnas.1311439110

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. In *Bioinformatics* (Vol. 18, Issue 2, pp. 337–338). https://doi.org/10.1093/bioinformatics/18.2.337

Jaime, A., Peter, R., Evan, M.-M., & Shaffer, H. B. (2018). Demographic inference in a spatially-explicit ecological model from genomic data: A proof of concept for the Mojave Desert Tortoise. In *bioRxiv* (p. 354530). https://doi.org/10.1101/354530

Jamieson, A., Carmagnini, A., Howard-McCombe, J., Doherty, S., Hirons, A., Dimopoulos, E., Lin, A. T., Allen, R., Anderson-Whymark, H., Barnett, R., Batey, C., Beglane, F., Bowden, W., Bratten, J., Cupere, B. D., Drew, E., Foley, N. M., Fowler, T., Fox, A., … Frantz, L. (2023). Limited historical admixture between European wildcats and domestic cats. *Current Biology*, *33*(21), 4751-4760.e14. https://doi.org/10.1016/j.cub.2023.08.031

Jay, F., Boitard, S., & Austerlitz, F. (2019). An ABC method for whole-genome sequence data: Inferring paleolithic and neolithic human expansions. In *Molecular biology and evolution* (Vol. 36, Issue 7, pp. 1565–1579).

Kaeuffer, R., Pontier, D., Devillard, S., & Perrin, N. (2004). Effective size of two feral domestic cat populations (Felis catus L.): Effect of the mating system. In *Molecular Ecology* (Vol. 13, Issue 2, pp. 483–490).

Kelleher, J., Thornton, K. R., Ashander, J., & Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. In *PLOS Computational Biology* (Vol. 14, Issue 11, p. e1006581). https://doi.org/10.1371/journal.pcbi.1006581

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., & McVean, G. (2019). Inferring whole-genome histories in large population datasets. In *Nature genetics* (Vol. 51, Issue 9, pp. 1330–1338).

Kidger, P., & Garcia, C. (2021). Equinox: Neural networks in JAX via callable PyTrees and filtered transformations. *arXiv Preprint arXiv:2111.00254*.

Kingman, J. F. C. (1982). The coalescent. In *Stochastic Processes and their Applications* (Vol. 13, Issue 3, pp. 235–248). https://doi.org/10.1016/0304-4149(82)90011-4

Langley, P. J. W., & Yalden, D. W. (1977). The decline of the rarer carnivores in Great Britain during the nineteenth century. In *Mammal Review* (Vol. 7, Issues 3–4, pp. 95–116). https://doi.org/10.1111/j.1365-2907.1977.tb00363.x

Li, P., & Vu, Q. D. (2013). Identification of parameter correlations for parameter estimation in dynamic biological models. In *BMC Systems Biology* (Vol. 7, Issue 1, p. 91). https://doi.org/10.1186/1752-0509-7-91

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., & Macke, J. H. (2017). Flexible statistical inference for mechanistic models of neural dynamics. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper\_files/paper/2017/file/addfa9b7e234254d26e9c7f2af1005cb-Paper.pdf

Mellett, C. L., Hodgson, D. M., Lang, A., Mauz, B., Selby, I., & Plater, A. J. (2012). Preservation of a drowned gravel barrier complex: A landscape evolution study from the north-eastern English Channel. In *Marine Geology* (Vols 315–318, pp. 115–131). https://doi.org/10.1016/j.margeo.2012.04.008

Neaves, L. E., & Hollingsworth, P. M. (2013). *The Scottish wildcat (Felis sylvestris): A review of genetic information and its implications for management*.

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. In *Proceedings of the National Academy of Sciences* (Vol. 76, Issue 10, pp. 5269–5273). https://doi.org/10.1073/pnas.76.10.5269

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, *22*(57), 1–64.

Papamakarios, G., Sterratt, D., & Murray, I. (2019). Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows. In K. Chaudhuri & M. Sugiyama (Eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (Vol. 89, pp. 837–848). PMLR. https://proceedings.mlr.press/v89/papamakarios19a.html

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (Vol. 32).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. In *The Journal of machine Learning research* (Vol. 12, pp. 2825–2830).

Peng, B., Chen, H.-S., Mechanic, L. E., Racine, B., Clarke, J., Gillanders, E., & Feuer, E. J. (2015). Genetic Data Simulators and their Applications: An Overview. In *Genetic Epidemiology* (Vol. 39, Issue 1, pp. 2–10). https://doi.org/10.1002/gepi.21876

Pierpaoli, M., Birò, Z. S., Herrmann, M., Hupe, K., Fernandes, M., Ragni, B., Szemethy, L., & Randi, E. (2003). Genetic distinction of wildcat (Felis silvestris) populations in Europe, and hybridization with domestic cats in Hungary. In *Molecular Ecology* (Vol. 12, Issue 10, pp. 2585–2598). https://doi.org/10.1046/j.1365-294X.2003.01939.x

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. In *Molecular biology and evolution* (Vol. 16, Issue 12, pp. 1791–1798).

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. In *Nature* (Vol. 461, Issue 7263, pp. 489–494). https://doi.org/10.1038/nature08365

Speed, D., & Balding, D. J. (2015). Relatedness in the post-genomic era: Is it still useful? In *Nature Reviews Genetics* (Vol. 16, Issue 1, pp. 33–44). https://doi.org/10.1038/nrg3821

Stajich, J. E., & Hahn, M. W. (2005). Disentangling the Effects of Demography and Selection in Human History. *Molecular Biology and Evolution*, *22*(1), 63–73. https://doi.org/10.1093/molbev/msh252

Steinrücken, M., Kamm, J., Spence, J. P., & Song, Y. S. (2019). Inference of complex population histories using whole-genome sequences from multiple populations. In *Proceedings of the National Academy of Sciences* (Vol. 116, Issue 34, pp. 17115–17120). https://doi.org/10.1073/pnas.1905060116

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. In *Genetics* (Vol. 123, Issue 3, pp. 585–595). https://doi.org/10.1093/genetics/123.3.585

Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics*, *153*(4), 1863–1871. https://doi.org/10.1093/genetics/153.4.1863

Wakeley, J., King, L., Low, B. S., & Ramachandran, S. (2012). Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingman’s Coalescent. *Genetics*, *190*(4), 1433–1445. https://doi.org/10.1534/genetics.111.135574

Ward, D. (2024). *FlowJAX: Distributions and Normalizing Flows in Jax* (Version 12.0.1) [Computer software]. https://doi.org/10.5281/zenodo.10402073

Yamaguchi, N., Kitchener, A., Driscoll, C., & Nussberger, B. (2015). *Felis silvestris. The IUCN red list of threatened species 2015: E. T60354712A50652361*. URL: http://dx. doi. org/10.2305/IUCN. UK.

**Figures and tables**

*Table 1*: Prior distributions for all model parameters. The ‘loc’ parameter of the translated log normal distribution denotes the size of the positive translation. All distributions were implemented using the scipy package (v1.13.0).

|  |  |  |
| --- | --- | --- |
| Parameter | Distribution | Source/Justification |
| N1 – Scottish captive population | Translated Log Normal:  (μ = 4.6, σ = 0.5, loc=10) | Based on posterior from Howard-McCombe *et al.* (2021) |
| N2 – Scottish wild population | Translated Log Normal:  (μ = 8.7, σ = 0.2, loc=30) | Based on posterior from Howard-McCombe *et al.* (2021) |
| N3 – European wild population | Log Normal:  (μ = 9.2, σ = 1) | Adapted from Ward (2021) |
| N4 – domestic population | Log Normal:  (μ = 10.8, σ = 1) | Adapted from Howard-McCombe *et al.* (2021) |
| N5 – African population | Log Normal:  (μ = 9.9, σ = 1) | No good estimates, wide prior taken. |
| N6 – Ancestral European population | Log Normal:  (μ = 10.6, σ = 1) | No good estimates, wide prior taken. |
| N7 – Ancestral African population | Log Normal:  (μ = 11.3, σ = 1) | No good estimates, wide prior taken. |
| T1 – African wildcat divergence | Translated Log Normal:  (μ = 10.8, σ = 1, loc=20000) | Very broad around mtDNA based estimate of ~200,000 years. (Driscoll et al., 2007) |
| T2 – domestic divergence | Translated Log Normal:  (μ = 7.8, σ = 0.2, loc=1000) | Wide around ~9000ya based on archaeological evidence. (Hu et al., 2014) |
| T3 – Scottish divergence | Translated Log Normal:  (μ = 8.2, σ = 0.05, loc=1000) | Flooding of Doggerland ~8kya, although barriers likely existed before this (Mellett et al., 2012). |
| T4 – Captive time | Translated Log Normal:  (μ = 2.7, σ = 0.4, loc=1) | Based on posterior from Howard-McCombe *et al.* (2021) |
| t – Migration duration | Translated Log Normal:  (μ = 2.5, σ = 0.4, loc=1) | Based on posterior from Howard-McCombe *et al.* (2021) |
| M1 – Migration rate: domestic to Scottish | Log Normal:  (μ = -2.5, σ = 0.5) | Based on posterior from Howard-McCombe *et al.* (2021) |
| M2 – Migration rate: Scottish to Captive | Log Normal:  (μ = -2.5, σ = 0.5) | Based on posterior from Howard-McCombe *et al.* (2021) |
| m – Mutation rate | Log Normal:  (μ = -18.4, σ = 1) | Steinrϋcken *et al.* (2019) |
| r – Recombination rate | Translated Log Normal:  (μ = -19.1, σ = 0.4, loc=1e-8) | Steinrϋcken *et al.* (2019) |

*Figure 1*: The schematic for our investigated model of wildcat demography, adapted from a previous study (Howard-McCombe et al., 2021). Arrows indicate direction of flow of migration. The part of the model highlighted in blue was simulated in forward-time by SLiM, and the part highlighted in yellow by the coalescent simulator msprime. The parameters used are described to the right of the diagram. The time axis is in units of generations and is not to scale.



**N1**: Population size: Captive Scottish *F. silvestris*

**N2**: Population size: Scottish *F. silvestris*

**N3**: Population size: European *F. silvestris*

**N4**: Population size: *F. catus*

**N5**: Population size: *F. lybica*

**N6**: Ancestral *F. silvestris* population size

**N7**: Ancestral *F. lybica* population size

**T1**: Divergence of *F. silvestris* and *F. lybica*

**T2**: Divergence of *F. catus* (domestication)

**T3**: Divergence of Scottish population

**T4**: Captive population established

**t**: Migration duration: domestic →Scottish wild

**M1**: Migration rate: domestic →Scottish wild

**M2**: Migration rate: Scottish wild →Scottish captive



100

A diagram of a process

Description automatically generated*Figure 2*: A simplified flowchart of the SNPE inference procedure, where the ‘Genomic data’ is the observed data. θ denotes model parameters used to simulate data, x.

A collage of graphs

Description automatically generated

*Figure 3*: KDE density plots of the prior (P) in blue/dashed, and first round posterior (R1) in red/solid. 10,000 simulated samples were used in training for each round of SNPE. Ne = effective population size.

A graph of a graph

Description automatically generated with medium confidence *Figure 4*: Pairplot of joint posterior distributions between parameters that show correlation. (maybe should be in supplementary)

*A screenshot of a graph

Description automatically generated*

*Figure 5*: PCA scatterplot of principal components 1 and 2 for the first round *maximum a posterioiri* simulation (a) and the observed genomic data (b).

**Supplementary material**

Table S1: The summary statistics removed using various methods for detecting statistics contributing to model misspecification.

|  |  |
| --- | --- |
| Method for removing statistics | Statistics removed |
| Limit correlation between summary statistics to 0.99. | y1\_domestic, y1\_scot  y1\_captive  y1\_eu  y1\_lyb  y3\_domestic\_captive\_lyb  divergence\_domestic\_eu  y3\_domestic\_scot\_lyb  y3\_domestic\_eu\_lyb  divergence\_captive\_lyb  segregating\_sites\_lyb  segregating\_sites\_eu  divergence\_scot\_lyb  f4\_domestic\_scot\_captive\_lyb  f4\_domestic\_captive\_eu\_lyb  f4\_domestic\_scot\_eu\_lyb  diversity\_domestic  y3\_domestic\_scot\_captive  y3\_scot\_captive\_eu  f2\_domestic\_eu  y2\_domestic\_lyb  y3\_scot\_captive\_lyb  divergence\_captive\_eu,pc1\_dist\_domestic\_eu |
| Outlier detection absolute noise <0.8. | tajimas\_d\_eu  relatedness\_domestic\_lyb  pc1\_iqr\_domestic  pc1\_iqr\_eu  pc1\_iqr\_lyb  pc2\_iqr\_domestic  pc2\_iqr\_lyb  relatedness\_captive\_lyb  pc1\_median\_eu  pc2\_dist\_domestic\_lyb |
| Iterative algorithm dropping stats to improve percentile of observed data. | pc2\_iqr\_eu  relatedness\_domestic\_eu  tajimas\_d\_lyb  pc1\_dist\_captive\_eu  fst\_scot\_lyb  pc2\_iqr\_scot  pc1\_dist\_scot\_lyb  pc1\_dist\_eu\_lyb |