Inferring the Demographic History of Wildcats (*Felis silvestris*) through Genetic Data: Simulations and Machine Learning.

Harry Gordon

**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 this inference 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 aimed to expand the previously studied model of evolution 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. After two rounds of inference, we found that SNPE was able to provide informative posterior density estimates for most parameters, but failed to recover informative posteriors for some parameters, especially those describing ancient demography. We propose that this shortcoming was due largely to model misspecification, which is a common pitfall of such an approach. Future studies should aim to improve this fit of the model by reconciling differences between observed and simulated data, and exploring different scenarios of evolution, while using point values for parameters when possible. Our approach provides a valuable and flexible template for inferring models of evolution and demography based on whole genome sequence data.

**Acknowledgements**

I would like to thank my supervisor, Mark Beaumont, Jo Howard-McCombe, and Dan Ward for conceiving such an interesting project, and for so generously giving me their time, patience, and guidance through such a steep learning curve. I would also like to thank Dan Ward and Jack Morton, for their superbly written GitHub repositories, which lay the foundation for much of this project. Thanks also to the University of Bristol for the use of their computing cluster, which was essential in the undertaking of this project.

**Keywords**

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

**Author’s declaration**

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

SIGNED: HARRY GORDON DATE: 09/09/2023

**Contents**

1. Introduction
2. Methods
   1. Model of Wildcat demography
   2. Prior distributions
   3. Genomic data
   4. Coupling forward-time and coalescent simulators
   5. Summary statistics
   6. Model misspecification
   7. SNPE procedure
3. Results
   1. Posterior distributions
   2. *Maximum a posteriori* model
4. Discussion
   1. Conclusions for demography
   2. Choice of summary statistics
   3. Possible model improvements
   4. Genomic data
   5. Simulation approach limitations
   6. Final conclusions

**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 unpiece 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 [1].

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 an elegant solution to this problem by simulating, according to a mathematical model which 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 [2]. As highly 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 ‘Coalescent theory’ are mechanistic models describing the genealogy of a population of individuals, and underpin all simulators of this nature [3-5]. The main applications of these simulations are in evaluating population genomics methods, investigating scenarios of evolution, and simulation-based inference [6]. Since their arrival, these simulators have improved massively in their efficiency and scope of possible models, which has allowed for increasingly complex analyses [7].

In ABC, the posterior is traditionally 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 towards the observed data [2]. Sequential Neural Posterior Estimation (SNPE) harnesses the flexibility of neural networks to carry out this step in the analysis and makes inferences based on all the simulated data without the need for rejection [8]. The neural network is trained to identify the relationships between summary statistics and the parameters for the simulations that produced them, so it can then compute a posterior probability distribution of model parameters that are consistent with a sample of 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 posterior, providing a narrower posterior distribution with each round. 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.

*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 [9]. 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 the other [10]. 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 [11]. 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 [12, 13]. In an effort 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. The second 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 extent of this hybridisation would provide useful information to assist in conservation efforts and the model would support our wider understanding of Wildcats.

Previous studies similar to ours have used an ABC framework with a three-population model and reduced-representation sequencing data to investigate Scottish Wildcat hybridisation, using the forward-time simulator SLiM to simulate 500 generations of recent demography for Domestic cats, Scottish Wildcats and a captive population [14, 15]. A subsequent study optimized this ABC framework for a more detailed form of the same three-population model, this time using a combination of SLiM and a coalescent simulator, msprime [16], to model the demographic history of the three populations from the divergence of Domestics and European Wildcats to the present [17]. This coupled simulation approach was used in our study and will be covered in more detail in the methods section. SNPE is, relatively, a 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 [18, 19] and physics [20, 21]. However, very few studies, if any, have tried to use this approach sequentially to infer a model of evolution in population genomics.

**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 including domestic cats, Scottish *F. silvestris*, and a captive population of Scottish *F. silvestris*. 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 Domestics 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 of 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. 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 [22]. 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 which used BGISEQ and Illumina methods to sequence samples from a variety of sources [13, 23].

*2.4. Coupling forward-time and coalescent simulators*

Simulations were carried out using a coupled framework of SLiM (v4.0.1) [24], in forward-time and msprime (v1.2.0) [16] in the coalescent. A key difference between forward-time and coalescent simulators is that, generally speaking, forward-time simulators model the migration of genetic material much more accurately than coalescent simulators, at the cost of longer computation times [25, 26, ]. Therefore, similarly to a recent study using this coupling [17], our approach simulates the recent history of wildcat hybridisation in forward-time, and the more ancient demography of wildcats in the coalescent.

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 [12]. During simulations, SLiM and msprime recorded ancestry using ‘succinct 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 [27]. 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 [28].

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. (add info for subsequent rounds here)

*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 [29], number of segregating sites, Tajima’s D [30], divergence [29], genetic relatedness [31], Patterson’s f statistics [32], Y statistics [33], and Fst [34], all efficiently computed directly from the tree sequence by tskit (v0.5.5) [27].
* PCA median, inter-quartile range, and pairwise distance summary stats using scikit-learn (v1.2.2) [35]. 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 data frame and used as input for SNPE.

*2.6. Addressing model misspecification*

Originally, mutation and recombination rate parameters were treated as fixed point values, obtained from a previous study with a similar approach [17]. However, prior predictive checks comparing summary statistics of simulated and observed data revealed that the observed number of segregating sites was well below the support of the simulation, indicating model misspecification. Thus, prior distributions for mutation rate and recombination rate were added to provide some uncertainty in the model and allow simulation of models producing a wider range of segregating sites which are consistent with the observed data. Furthermore, the first round prior predictive check revealed that 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 [30]. However, this demography is not consistent with current understanding [36, 37] and it would be difficult to replicate this statistic in the simulated data, so this statistic was removed to allow the neural network to estimate a better model free of hugely inflated population sizes (inference including this statistic predicted a current European population of the order 107). All other observed summary statistics fell within the distribution of the simulated data and indicated that the model described the observed data adequately.

Rewrite this paragraph, go through:

* Removal of correlated stats
* Outlier/noise method (ask how to phrase this)
* ‘drop one’ method

*2.7. SNPE procedure*

For inferring the posterior distribution, the Flowjax package (v12.0.1) was used to carry out SNPE [8]. 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 method; carrying out inference without normalisation of the simulated data caused the neural network to fail to converge. The same normalising 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 [38]. The neural network was then trained with the first dataset of simulated data (10,000 simulations) to learn the relationships between summary stats and parameters, . Systematically, 10% of input data was held out for validation in training the neural network. To find the most appropriate inference method to use within SNPE, two density estimation models were compared: the ‘Mixture Density Network’ (MDN) model and the ‘Masked Autoregressive Flow’ (MAF) model [39, 40]. Both methods were used to estimate the posterior distribution after training the network with this initial dataset, and the log probabilities of parameters given the corresponding summary statistics for a held-out validation set of 1000 samples were averaged to obtain a measure of each method’s ability to estimate posterior density. The MAF method had a higher average log probability (-5.98) than the MDN method (-10.82), so was used in the following analysis. After training, the estimator was then provided the observed summary stats from which it estimated a posterior distribution of model parameters that were consistent with our observed data. 10,000 sets of model parameters were then sampled from this posterior, and unscaled and exponentiated to obtain the parameters for the next round of simulation. This simulation and inference procedure was repeated once more, providing the estimator with the posterior distribution from the previous round as a ‘proposal’, which prevented SNPE from overconfidently estimating a new posterior based on resampled data. The basic outline of this approach is illustrated in Figure 2.

**3. Results**

*3.1. Posterior distributions*

After one round of inference, the marginal posteriors estimated were narrower than priors for the majority of parameters (*Fig. 3*). However, the second round of inference yielded posteriors that were almost all wider than the previous round’s estimates. Despite this, posteriors for many parameters were still improvements on our prior distributions. The posterior for *F. lybica* divergence was almost identical to our prior, and was very widely distributed around a posterior mean of 79k generations (*Fig. 3a*). Posteriors for Scottish *F. silvestris* and Domestic divergence times were also very similar to our priors, but were slightly lower and similarly distributed around posterior means of 4400 and 3400 generations respectively (*Fig. 3b, c*). The posterior for divergence of the captive population was wider than our prior and had a posterior mean of 11.7 generations (*Fig. 3d*), the size of this population was also wider than the prior and had a posterior mean of 140 individuals (*Fig. 3e*). The posterior for population size of Scottish *F. silvestris* was narrower than both our prior and first round posterior, narrowly distributed around a mean of 4050 individuals (*Fig. 3f*). Posteriors for European *F. silvestris* and Domestic cat effective population sizes were both slightly narrower than our priors, with means of 15.3k and 111k individuals respectively (*Fig. 3g, h*). The posterior for current *F. lybica* population size was narrower than our prior, and had a posterior mean of 7.9k individuals (*Fig. 3i*). Despite reasonably narrow first-round posteriors, both second-round posteriors for ancestral *F. silvestris* and *F. lybica* population size were far wider than our priors, so provided no reasonable estimates of these parameters (*Fig. 3j, k*). Posteriors for genetic migration parameters were all more narrowly distributed than our priors, with a posterior mean for migration length between Domestics and Wild populations of 11.6 generations (*Fig. 3l*), and posterior means for the rate of migration to wild and captive populations of 3.7% and 6.8% migrants per generation respectively (*Fig. 3m, n*). Posteriors for mutation rate and recombination rate were both considerably wider than our priors, so provide no precise estimates for these parameters (*Fig. 3o, p*). The plot of marginal pairwise distributions for the first round revealed a number of parameters were correlated, and therefore potentially non-identifiable (*Fig. 4*). Correlation of these parameters indicates that the data may lack features that separately describe these parameters, and features determining one parameter will also determine the other [41]. Parameters with high correlation included mutation rate, ancient population sizes, and *F. lybica* divergence.

*3.2. Maximum a posteriori model*

The model with the greatest inferred joint posterior probability, known as *maximum a posteriori* (MAP), was obtained for the first round of inference. This model was simulated, and PCA analysis was conducted on both the genotype array produced by this simulation and the observed data. Principal components 1 and 2 were plotted (*Fig. 5*) for both and show considerable resemblance between them. All populations show a similar relative distribution across PC1 and PC2, with the possible exception of the European *F. silvestris* population, which is clustered further 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**

Although the first round of inference produced marginal posterior distributions of model parameters that were predominantly narrower than our prior distributions indicating the potential for precise inference, many of the marginal distributions estimated from the second batch of sequentially simulated data were wider than the previous rounds estimate, which indicates some level of model misspecification. Despite this, using our model, around half of the investigated parameters were recovered to some extent and estimated with higher density than the prior.

*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 carry some ecological significance (*Fig. 3m, l*). The posterior mean for rate of migration estimated at 4% per generation and the date for onset of this migration estimated at 11.7 generations (~35 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 [14]. 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 [13]. 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*) [14]. This finding suggests that the captive population was established before much of the hybridisation between Wildcats and Domestics occurred, emphasizing the potential 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 [42, 43]. 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*) [44].

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 [45, 46]. 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 [47]. 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 [40]. 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.

*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 [30], 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 [48]. 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 [49].

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 [50]. 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.

*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. 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 [51]. 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.

*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 [17]. 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.

**Supplementary material**

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

**References**

1. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002; 162(4):2025-35.

2. Beaumont MA, Rannala B. The Bayesian revolution in genetics. Nature Reviews Genetics. 2004; 5(4):251-61.

3. Barroso GV, Moutinho AF, Dutheil JY. A Population Genomics Lexicon. In: Dutheil JY, editor. Statistical Population Genomics. New York, NY: Springer US; 2020. p. 3-17.

4. Kingman JFC. The coalescent. Stochastic Processes and their Applications. 1982; 13(3):235-48.

5. Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002; 18(2):337-8.

6. Hoban S. An overview of the utility of population simulation software in molecular ecology. Molecular Ecology. 2014; 23(10):2383-401.

7. Peng B, Chen H-S, Mechanic LE, Racine B, Clarke J, Gillanders E, Feuer EJ. Genetic Data Simulators and their Applications: An Overview. Genetic Epidemiology. 2015; 39(1):2-10.

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

9. Yamaguchi N, Kitchener A, Driscoll C, Nussberger B. Felis silvestris. The IUCN red list of threatened species 2015: e. T60354712A50652361. URL: http://dx. doi. org/10.2305/IUCN. UK; 2015.

10. Todesco M, Pascual MA, Owens GL, Ostevik KL, Moyers BT, Hübner S, Heredia SM, Hahn MA, Caseys C, Bock DG. Hybridization and extinction. Evolutionary applications. 2016; 9(7):892-908.

11. Breitenmoser U, Lanz T, Breitenmoser-Würsten C. Conservation of the wildcat (Felis silvestris) in Scotland: review of the conservation status and assessment of conservation activities. IUCN SSC Cat Specialist Group. 2019.

12. Beaumont M, Barratt EM, Gottelli D, Kitchener AC, Daniels MJ, Pritchard JK, Bruford MW. Genetic diversity and introgression in the Scottish wildcat. Molecular Ecology. 2001; 10(2):319-36.

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

14. Howard-McCombe J, Ward D, Kitchener AC, Lawson D, Senn HV, Beaumont M. On the use of genome-wide data to model and date the time of anthropogenic hybridisation: An example from the Scottish wildcat. Molecular Ecology. 2021; 30(15):3688-702.

15. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. Molecular Biology and Evolution. 2019; 36(3):632-7.

16. Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG. et al. Efficient ancestry and mutation simulation with msprime 1.0. Genetics. 2022; 220(3).

17. Ward D. Scottish Wildcats and Simulations as a Tool for Demographic Inference: University of Bristol; 2021.

18. Gonçalves PJ, Lueckmann J-M, Deistler M, Nonnenmacher M, Öcal K, Bassetto G, Chintaluri C, Podlaski WF, Haddad SA, Vogels TP, et al. Training deep neural density estimators to identify mechanistic models of neural dynamics. eLife. 2020; 9:e56261.

19. Groschner LN, Malis JG, Zuidinga B, Borst A. A biophysical account of multiplication by a single neuron. Nature. 2022; 603(7899):119-23.

20. Furia CS, Churchill RM. Normalizing flows for likelihood-free inference with fusion simulations. Plasma Physics and Controlled Fusion. 2022; 64(10):104003.

21. Akhmetzhanova A, Mishra-Sharma S, Dvorkin C. Data Compression and Inference in Cosmology with Self-Supervised Machine Learning. arXiv preprint arXiv:230809751. 2023.

22. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156-8.

23. Jamieson A, Carmagnini A, Howard-McCombe J, Doherty S, Hirons A, Dimopolous E, Lin A, Allen R, Anderson-Whymark H, Barnett R. Palaeogenomic Evidence for the Long-Term Reproductive Isolation Between Wild and Domestic Cats.

24. Haller BC, Messer PW. SLiM 4: multispecies eco-evolutionary modeling. The American Naturalist. 2023; 201(5):E127-E39.

25. 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.

26. Cyran KA, Myszor D. Coalescent vs. time-forward simulations in the problem of the detection of past population expansion. J. Appl. Math. Inform. 2008; 2(1):10.

27. Kelleher J, Thornton KR, Ashander J, Ralph PL. Efficient pedigree recording for fast population genetics simulation. PLOS Computational Biology. 2018; 14(11):e1006581.

28. Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. Nature genetics. 2019; 51(9):1330-8.

29. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. 1979; 76(10):5269-73.

30. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123(3):585-95.

31. Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nature Reviews Genetics. 2015; 16(1):33-44.

32. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009; 461(7263):489-94.

33. Jaime A, Peter R, Evan M-M, Shaffer HB. Demographic inference in a spatially-explicit ecological model from genomic data: a proof of concept for the Mojave Desert Tortoise. bioRxiv. 2018:354530.

34. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F ST. Nature Reviews Genetics. 2009; 10(9):639-50.

35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res.. 2011; 12:2825-30.

36. Pierpaoli M, Birò ZS, Herrmann M, Hupe K, Fernandes M, Ragni B, Szemethy L, Randi E. Genetic distinction of wildcat (Felis silvestris) populations in Europe, and hybridization with domestic cats in Hungary. Molecular Ecology. 2003; 12(10):2585-98.

37. Langley PJW, Yalden DW. The decline of the rarer carnivores in Great Britain during the nineteenth century. Mammal Review. 1977; 7(3-4):95-116.

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

39. Papamakarios G, Murray I. Fast ε-free inference of simulation models with bayesian conditional density estimation. Advances in neural information processing systems. 2016; 29.

40. Greenberg D, Nonnenmacher M, Macke J, editors. Automatic posterior transformation for likelihood-free inference. P. Mach. Learn. Res. 2019.

41. Li P, Vu QD. Identification of parameter correlations for parameter estimation in dynamic biological models. BMC Systems Biology. 2013; 7(1):91.

42. CATS Report Scotland. Cats Protection; 2021.

43. Kaeuffer R, Pontier D, Devillard S, Perrin N. Effective size of two feral domestic cat populations (Felis catus L.): effect of the mating system. Molecular Ecology. 2004; 13(2):483-90.

44. Eckert I, Suchentrunk F, Markov G, Hartl GB. Genetic diversity and integrity of German wildcat (Felis silvestris) populations as revealed by microsatellites, allozymes, and mitochondrial DNA sequences. Mammalian Biology. 2010; 75:160-74.

45. Driscoll CA, Menotti-Raymond M, Roca AL, Hupe K, Johnson WE, Geffen E, Harley EH, Delibes M, Pontier D, Kitchener AC. The Near Eastern origin of cat domestication. Science. 2007; 317(5837):519-23.

46. Mellett CL, Hodgson DM, Lang A, Mauz B, Selby I, Plater AJ. Preservation of a drowned gravel barrier complex: A landscape evolution study from the north-eastern English Channel. Marine Geology. 2012; 315-318:115-31.

47. Aeschbacher S, Beaumont MA, Futschik A. A novel approach for choosing summary statistics in approximate Bayesian computation. Genetics. 2012; 192(3):1027-47.

48. Neaves LE, Hollingsworth PM. The Scottish wildcat (Felis silvestris): a review of genetic information and its implications for management. 2013.

49. Beugin MP, Salvador O, Leblanc G, Queney G, Natoli E, Pontier D. Hybridization between Felis silvestris silvestris and Felis silvestris catus in two contrasted environments in France. Ecology and Evolution. 2020; 10(1):263-76.

50. Cannon P, Ward D, Schmon SM. Investigating the impact of model misspecification in neural simulation-based inference. arXiv preprint arXiv:220901845. 2022.

51. Jay F, Boitard S, Austerlitz F. An ABC method for whole-genome sequence data: inferring paleolithic and neolithic human expansions. Molecular biology and evolution. 2019; 36(7):1565-79.

52. Hu Y, Hu S, Wang W, Wu X, Marshall FB, Chen X, Hou L, Wang C. Earliest evidence for commensal processes of cat domestication. Proc. Natl. Acad. Sci.. 2014; 111(1):116-20.

53. Steinrücken M, Kamm J, Spence JP, Song YS. Inference of complex population histories using whole-genome sequences from multiple populations. Proc. Natl. Acad. Sci. 2019; 116(34):17115-20.

**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.

|  |  |  |
| --- | --- | --- |
| Parameter | Distribution | Source/Justification |
| N1 – Scottish captive population | Translated Log Normal:  (μ = 4.6, σ = 0.5, loc=10) | Howard-McCombe *et al.* (2021) |
| N2 – Scottish wild population | Translated Log Normal:  (μ = 8.7, σ = 0.2, loc=30) | Howard-McCombe *et al.* (2021) |
| N3 – European wild population | Log Normal:  (μ = 9.2, σ = 1) | Adapted from Ward (2021) [17] |
| 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 divergence | Translated Log Normal:  (μ = 10.8, σ = 1, loc=20000) | Very broad around mtDNA based estimate of ~200,000 years.  Driscoll, (2007) [45] |
| 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) [52] |
| 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) [46] |
| T4 – Captive time | Translated Log Normal:  (μ = 2.7, σ = 0.4, loc=1) | Howard-McCombe *et al.* (2021) |
| t – Migration duration | Translated Log Normal:  (μ = 2.5, σ = 0.4, loc=1) | Howard-McCombe *et al.* (2021) |
| M1 – Migration rate: Domestic to Scottish | Log Normal:  (μ = -2.5, σ = 0.5) | Howard-McCombe *et al.* (2021) |
| M2 – Migration rate: Scottish to Captive | Log Normal:  (μ = -2.5, σ = 0.5) | Howard-McCombe *et al.* (2021) |
| m – Mutation rate | Log Normal:  (μ = -18.4, σ = 1) | Steinrϋcken *et al.* (2019) (53) |
| 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 two previous related studies [14, 17]. 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

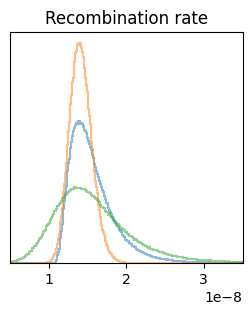
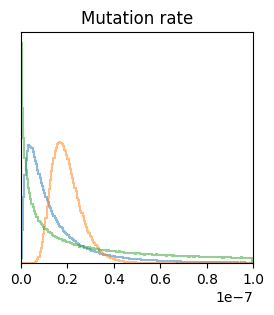
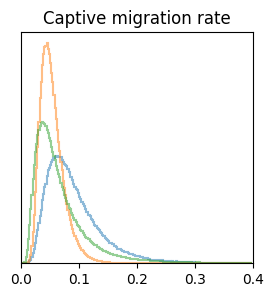
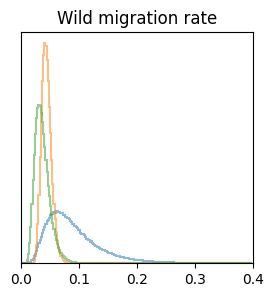
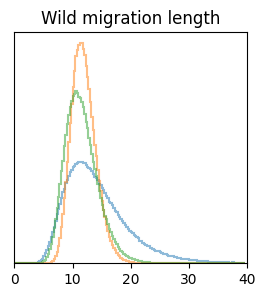
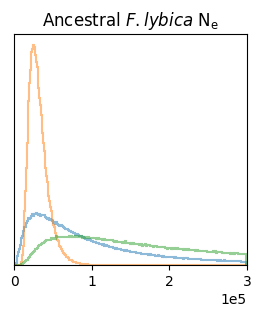
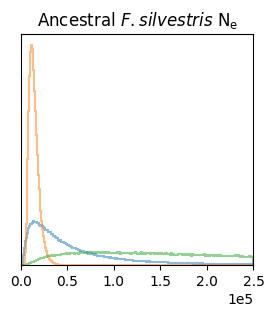
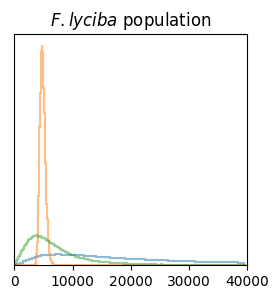
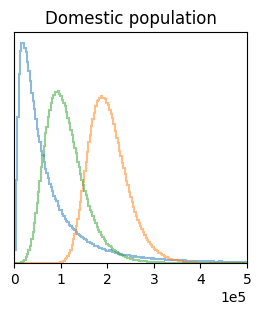
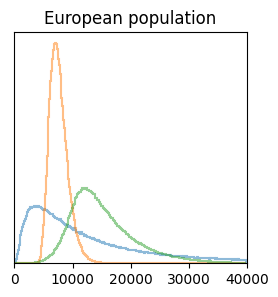
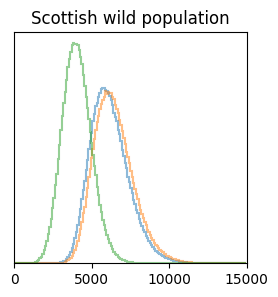
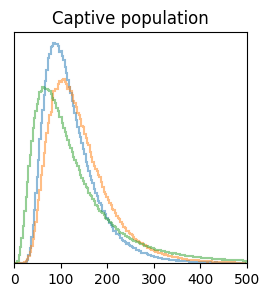
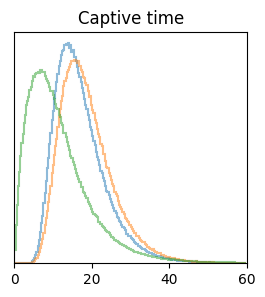
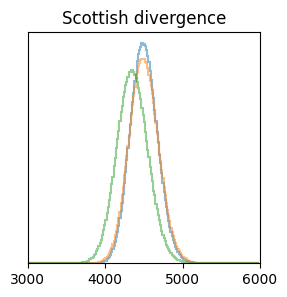
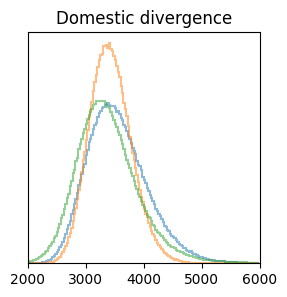
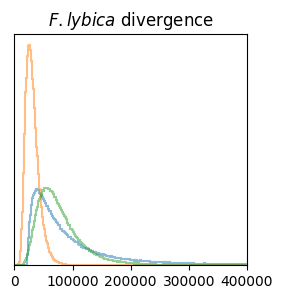


100

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.

*Figure 3*: Density plots of the prior (P), first round posterior (R1), and second round posterior (R2) distributions for model parameters. Approximately 9500 simulated samples were used in training for each round of SNPE. Ne = effective population size.



**(a)**

**(b)**

**(c)**

**(d)**

**(e)**

**(f)**

**(g)**

**(h)**

**(i)**

**(j)**

**(k)**

**(l)**

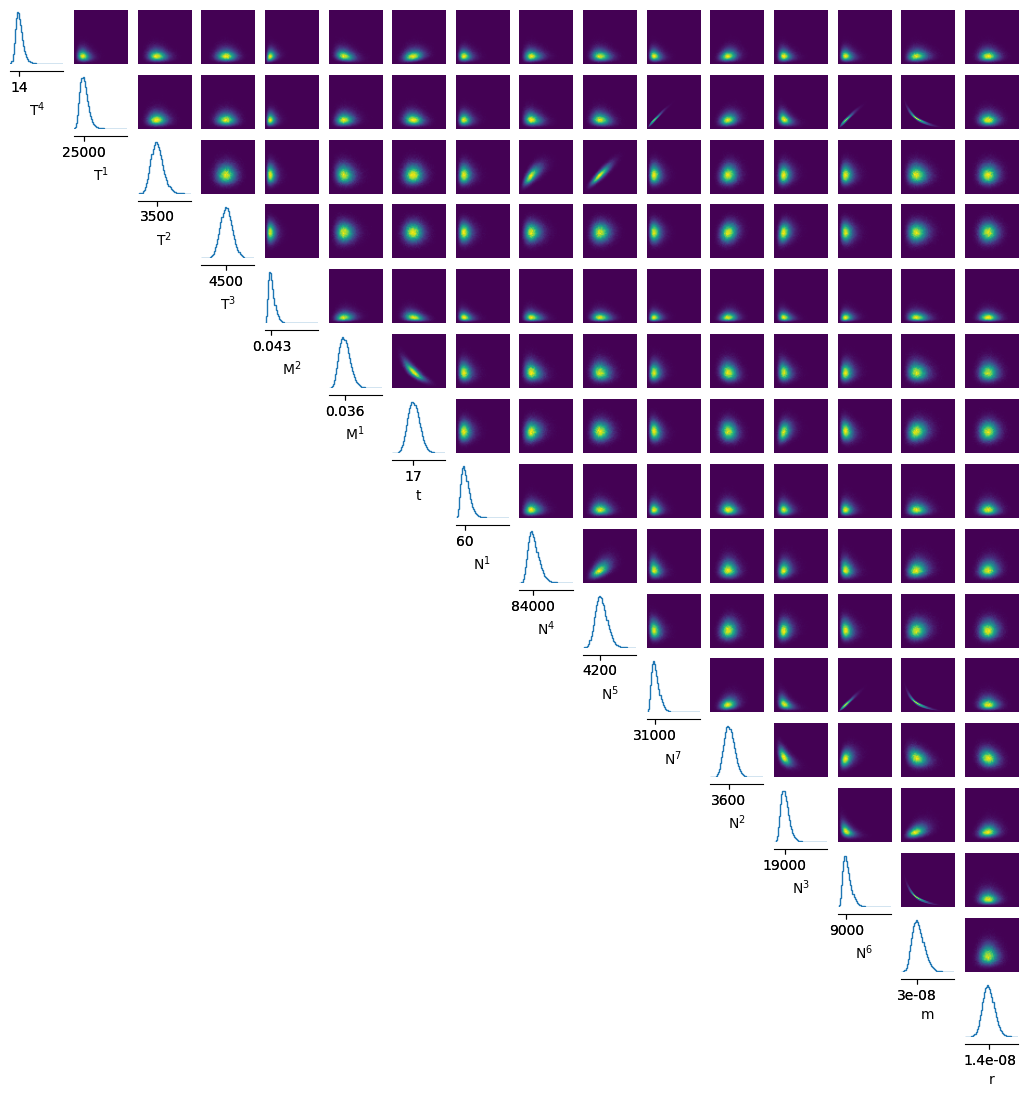
**(m)**

**(n)**

**(o)**

**(p)**

*Figure 4*: Density plots of pairwise marginal distributions between parameters after the first round of inference. The linear shapes of some marginal distributions indicate correlation of parameters. T4 = captive time, T1 = *F. lybica* divergence, T2 = Domestic divergence, T3 = Scottish Divergence, M2 = captive migration rate, M1 = wild migration rate, t = migration length, N1 = captive population, N4 = domestic population, N5 = *F. lybica* population, N7 = ancestral *F. lybica* population, N2 = Scottish *F. silvestris* population, N3 = European *F. silvestris*, N6 = ancestral *F. silvestris* population, m = mutation rate, r = recombination rate.



T4

T1

T2

T3

M2

M1

t

N1

N4

N5

N7

N2

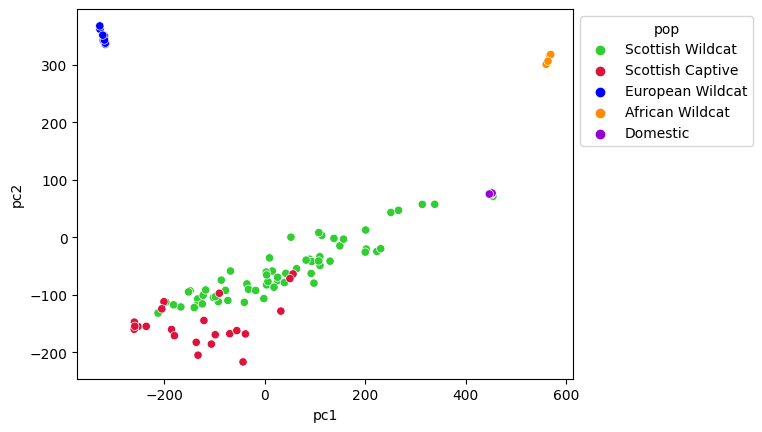
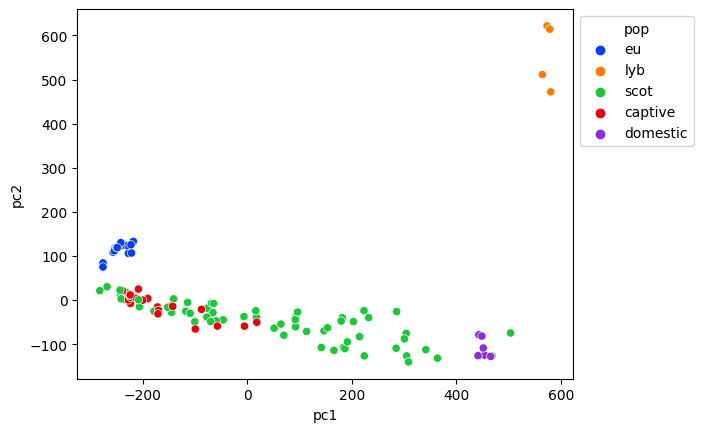
N3

N6

m

r

*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).



**(b)**

**(a)**

