Model misspecification in population genomics

Mark A. Beaumont,
The University of Bristol,
Bristol, UK

30 May 2024

Overview of talk

1. Introduction and Motivation
2. Example of misspecification
3. ABC methods for misspecification
4. Misspecification and Neural Posterior Estimation
5. ABC by dropping summary statistics
6. Example applications

Introduction and Motivation

- In the standard likelihood-free posterior

$$p_\epsilon(\theta|s_O) = \int (\theta) f(s|\theta) K_\epsilon(||s - s_O||) ds$$

- In typical likelihood-free settings we sample from the joint distribution

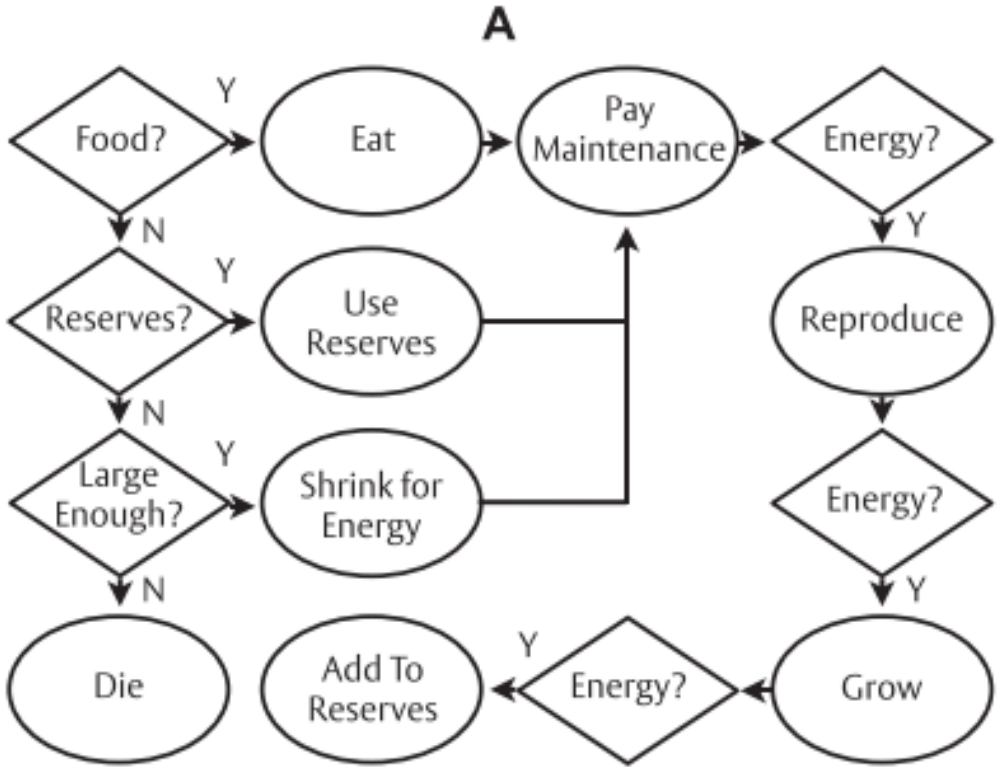
$$(x_i, \theta_i) \sim p(x|\theta)p(\theta)$$

, compute summary $s = S(x)$

- In complex models it is often apparent that $K_\epsilon(||s - s_O||) << K_\epsilon(||s - s_{ref}||)$ for any (s_i, s_{ref}) computed from any finite sample from $p(x|\theta)p(\theta)$
- This model misspecification causes issues both for accurate inference, and for stability of sequential Monte Carlo inference methods.

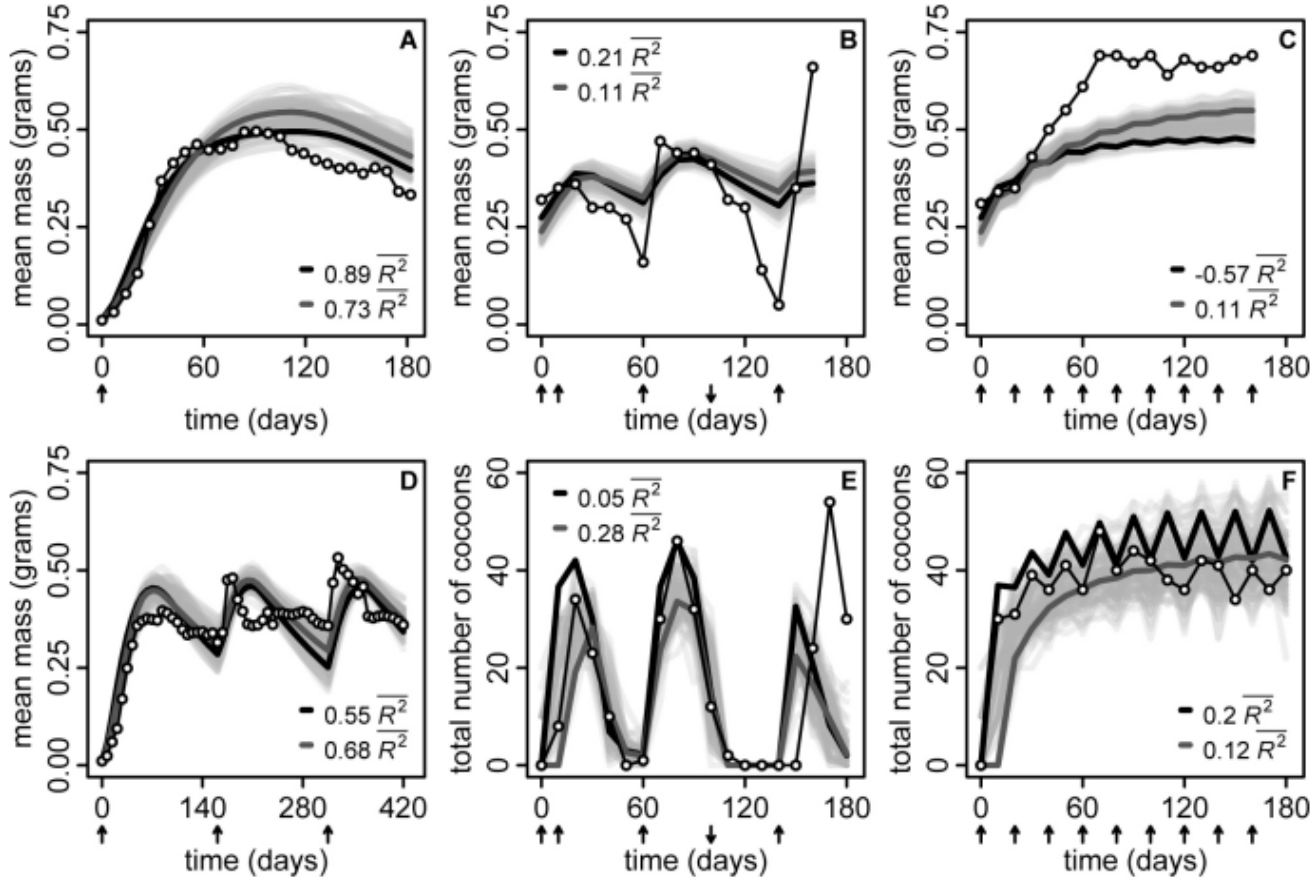
Example of misspecification in Individual-Based Model

- Model for growth of individual earthworms, used to predict population dynamics in lab culture
- Van der Vaart *et al* (Ecological Modelling, 2015)



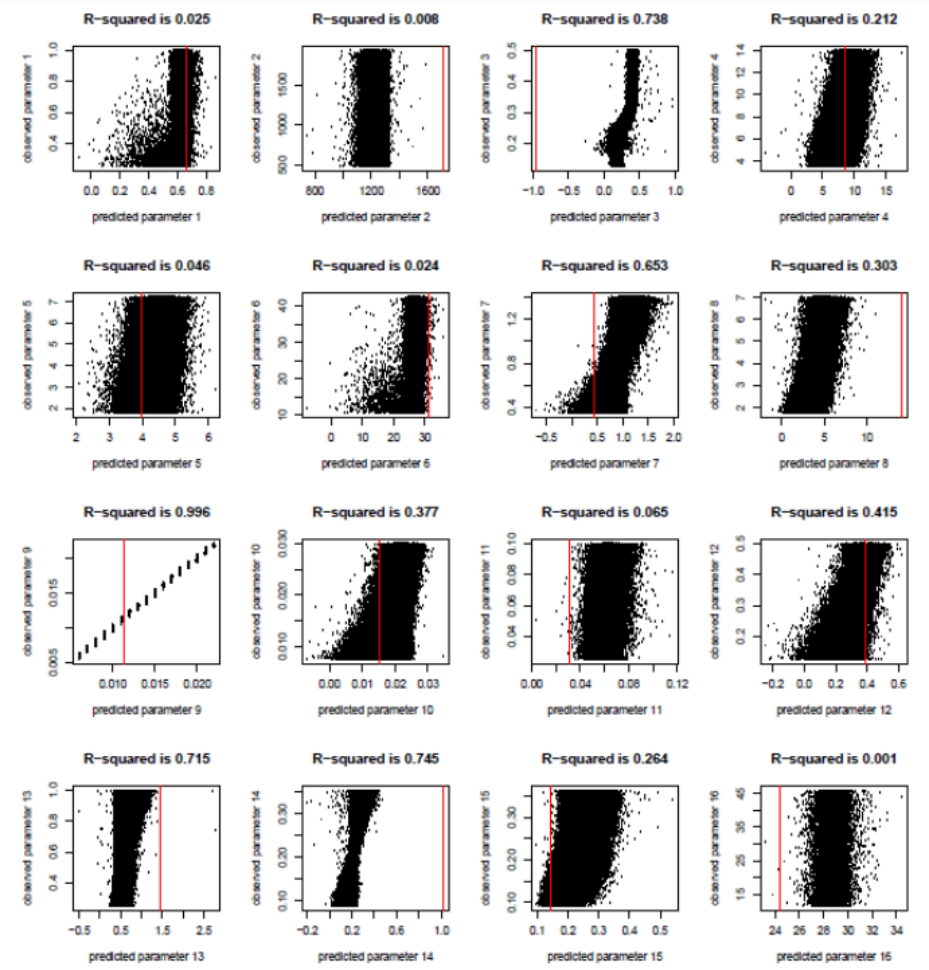
Posterior Predictive Distribution

- Aim is to model emergent phenomena from physiological parameters.
- Different experiments (arrows indicate input of food)



Evidence of misspecification

- Use regression with samples from $p(s, \theta)$ to predict elements of θ from s (Fearnhead & Prangle, 2012).
- Plots shown for different parameters. Red line shows prediction from observation summary S_O



Robust ABC

- Early study of robust ABC inference by Ratmann *et al* (2009), following ideas in Wilkinson (2008).
- Modify standard ABC posterior

$$p_\epsilon(\theta|s_O) = \int p(\theta) f(s|\theta) K_\epsilon(||s - s_O||) ds$$

- by augmenting θ with vector ϵ, with prior $p(\epsilon)$.
- Assume $p(\theta, \epsilon) = p(\theta)p(\epsilon)$ giving

$$p(\theta, \epsilon|s_O) = \int p(\theta, \epsilon) f(s|\theta) K_\epsilon(||s - s_O||) ds$$

- These ideas extended by many researchers, summarised and extended in Frazier *et al* (2020) and Frazier and Drovandi (2021).

Robust Neural Posterior Estimation

Ward *et al*, NIPS (2022)

- Assume that observable data y arises with error from unobserved latent x, with some error model $p(y|x)$
- The joint distribution

$$p(y, x, \theta) = p(y|x)p(x|\theta)p(\theta)$$

  can be equivalently written as $p(y)p(x|y)p(\theta|x)$
- giving

$$p(\theta|y) = \int p(x|y)p(\theta|x)dx$$

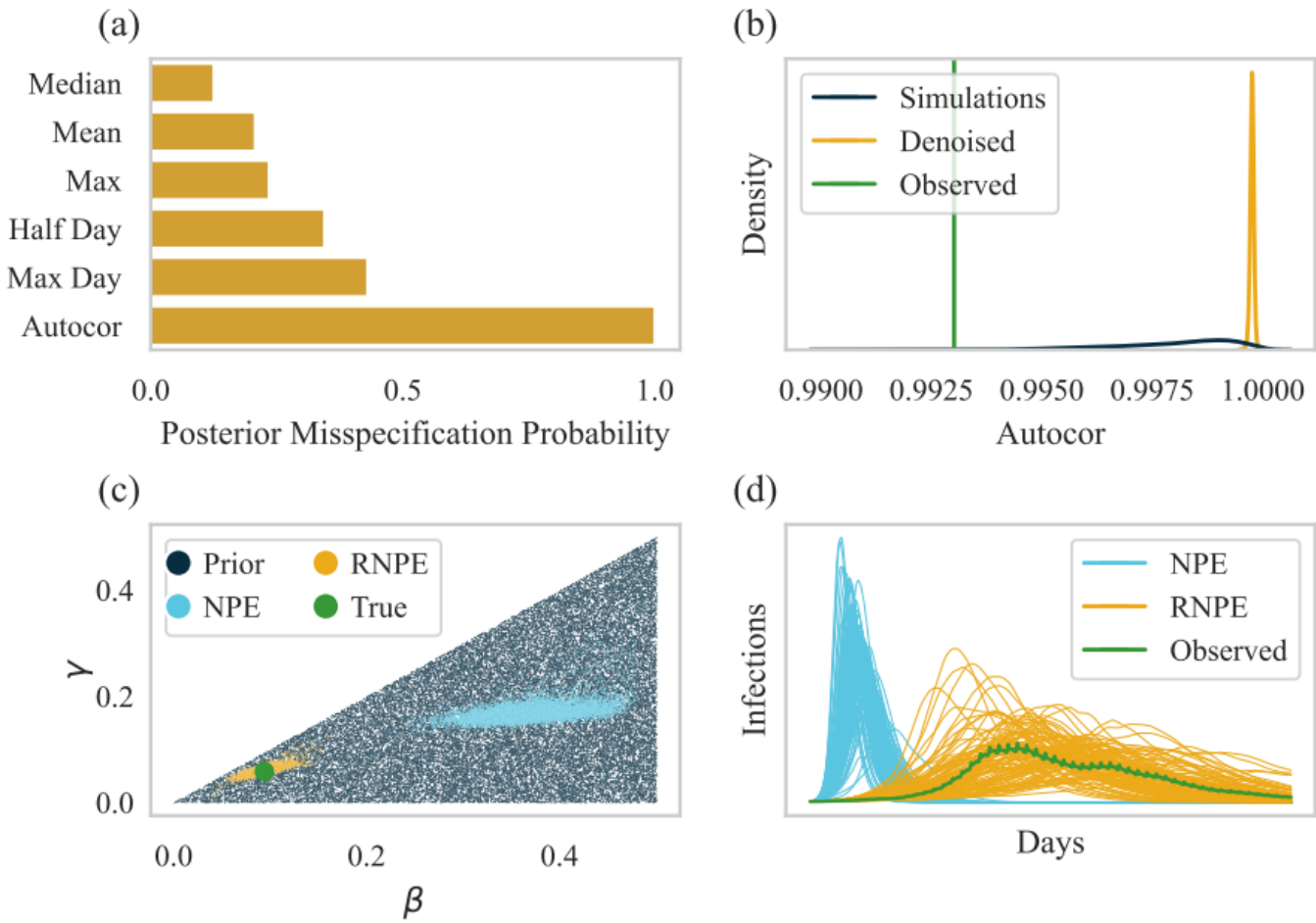
- The distribution $p(\theta|x)$ can be approximated by neural posterior estimation (with a normalising flow), trained with samples from $p(y, x, \theta)$, marginal to y
- To obtain samples of $x_i \sim p(x|y)$ to approximate the integral by Monte Carlo, Ward *et al* used MCMC to sample from $q_\phi(x)p(y|x)$ where $q_\phi(x)$ is a normalising flow, with weights ϕ trained with marginal x_i from $p(y, x, \theta)$

Error model

- Use a spike and slab model.
- Assume $\mathbf{x} \in \mathbb{R}^d$
- Sample $\mathbf{x} \sim q_\phi(\mathbf{x})$
- $z_j \sim \text{Bernoulli}(\rho)$
- $$y_j | x_j, z_j \sim \begin{cases} N(x_j, \sigma^2), & \text{if } z_j = 0 \\ \text{Cauchy}(x_j, \tau), & \text{if } z_j = 1 \end{cases}$$

Example — SIR model with reporting delays

- Model of epidemic spread ('Susceptible - Infected - Removed').
- Infer two parameters: infection rate β and recovery rate γ
- Simulate observations with misspecification (proportion of weekend infections reported on Monday)



Dropping Summary statistics

Motivation to discard summary statistics completely in ABC:

- Complex models are designed to capture only some features of data.
- Many models of misspecification allow for increased uncertainty in the value of some summary statistics (*e.g.* Ratmann *et al*,2009; Ward *et al*, 2022)
- Potential advantages of simplicity (and hence wider use) in the approach.

A criterion for dropping summary statistics

- Given n samples $x_i \sim p(x, \theta)$ summarise to a d-dimensional vector $s = S_d(x)$ (similarly for observations $s_O = S_d(x_o)$).
- Use some method to approximate prior predictive distribution of summary statistics $p(s)$ from sample.
- Assume we require s_O to lie within the approximate 95% Highest Density Region (HDR).
- Rank densities $p(s)$ for all points $\{s_i, s_O\}$ from largest to smallest, with rank $j = 1 \ldots n + 1$
- Accept s_O if $p(s_O) > p(s_j)$ when $j = 0{\cdot}95(n + 1)$
- Otherwise, drop component $1, \ldots, d$ from $S_d(\cdot)$
- Each time re-rank densities $p(s)$ for all points $\{s_i, s_O\}$
- Choose to drop the component giving the largest rank improvement for $p(s'_O)$ with $s'_O = S_{d-1}(x_O)$.
- Repeat procedure until $p(s_O) > p(s_j)$ when $j = 0{\cdot}95(n + 1)$

k-NN density estimation

- Assume $x \in \mathbb{R}^d$
- with $k = 1$ (nearest neighbour) out of n observations
- Estimated density at point x, $\hat{p}(x) = \frac{1}{nV(d)r(x)^d}$
- where $r(x)$ is the nearest neighbour distance at point x (Euclidean)
- $V(d)$ is volume of unit ball in d dimensions.

Obtaining approximate prior HDR in ABC framework

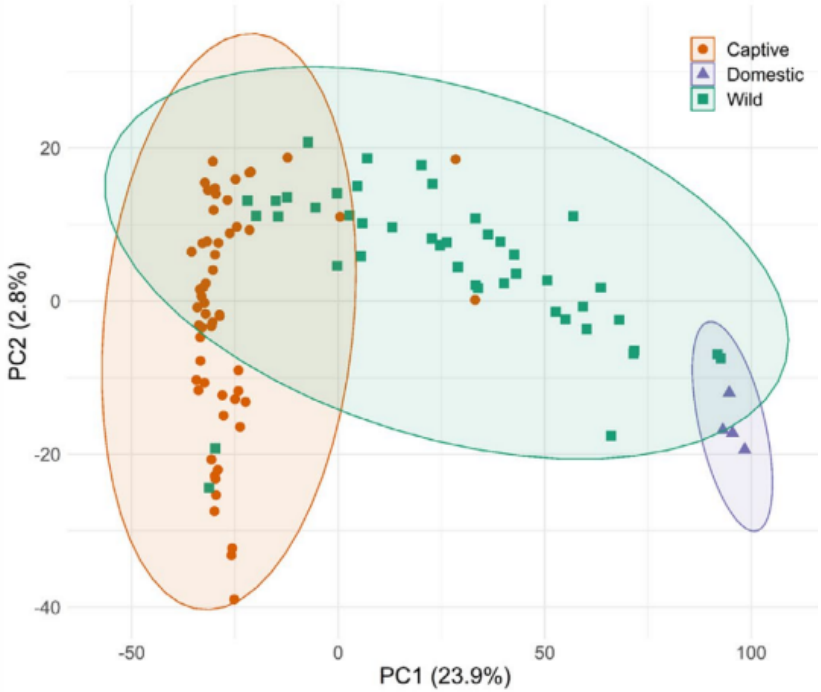
- Assume we have n simulations from the prior predictive distribution: $\theta_i \sim \pi(\theta)$ $x_i \sim p(x|\theta_i)$
- Summarise ith point as $s_i = S_d(x_i)$
- In principle we could compute $\hat{p}(s_i) = \frac{1}{nV(d)r(s_i)^d}$
- Rank $\hat{p}(s_i)$ from largest to smallest.
- Approximate *e.g.* 95% highest density region (HDR) given by points with k-NN density not ranked less than $0{\cdot}95n$.

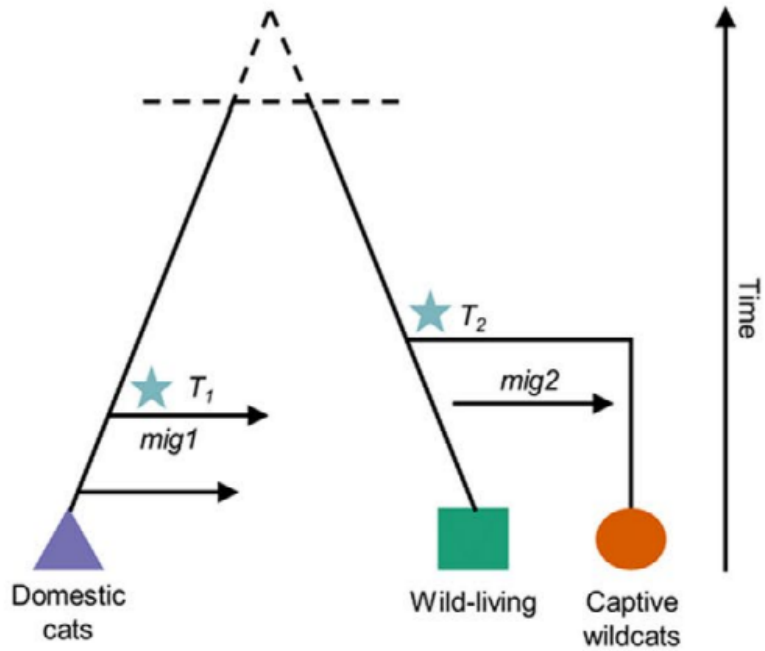
Choosing summary statistics

- Note that $\hat{p}(s_i)$ is monotonically increasing with $\frac{1}{r(s_i)}$.
- *I.e* rank distances $r(s)$ for all points $\{s_i, s_O\}$ from smallest to largest, with rank $j = 1, \ldots, n+1$.
- Accept s_O if $r(s_O) < r(s_j)$ when $j = 0{\cdot}95(n+1)$.
- Otherwise, drop component $1, \ldots, d$ from $S_d(\cdot)$
- Each time re-rank distances $r(s)$ for all points $\{s_i, s_O\}$
- Choose to drop the component giving the largest rank improvement for $r(s'_O)$ with $s'_O = S_{d-1}(x_O)$
- Repeat procedure until $r(s_O) < r(s_j)$ when $j = 0{\cdot}95(n+1)$.

Example Application: Modelling Hybridisation in Scottish Wildcat

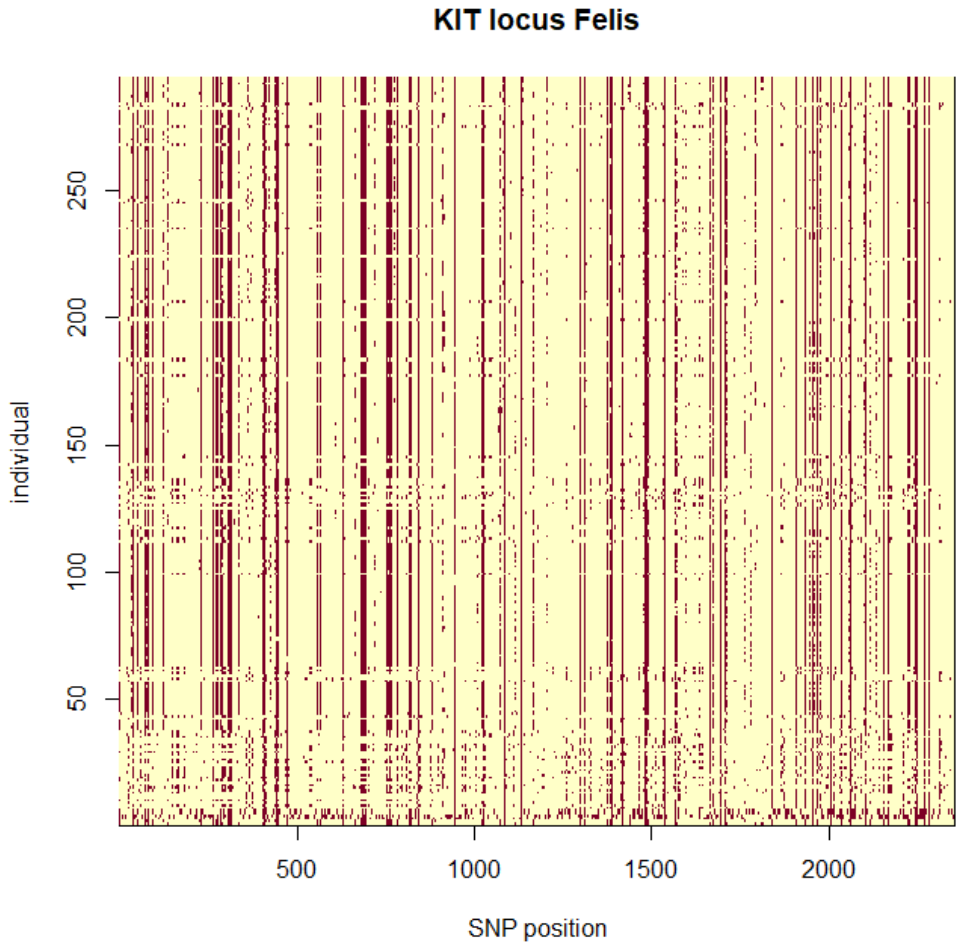
- Aim: to model history of hybridisation in Scottish wildcat
- Data: Single nucleotide polymorphism (SNP) data from wild-living cats in Scotland.





Howard-McCombe *et al* (2021) *Molecular Ecology*.

Digression — how do you carry out PCA on genome data?



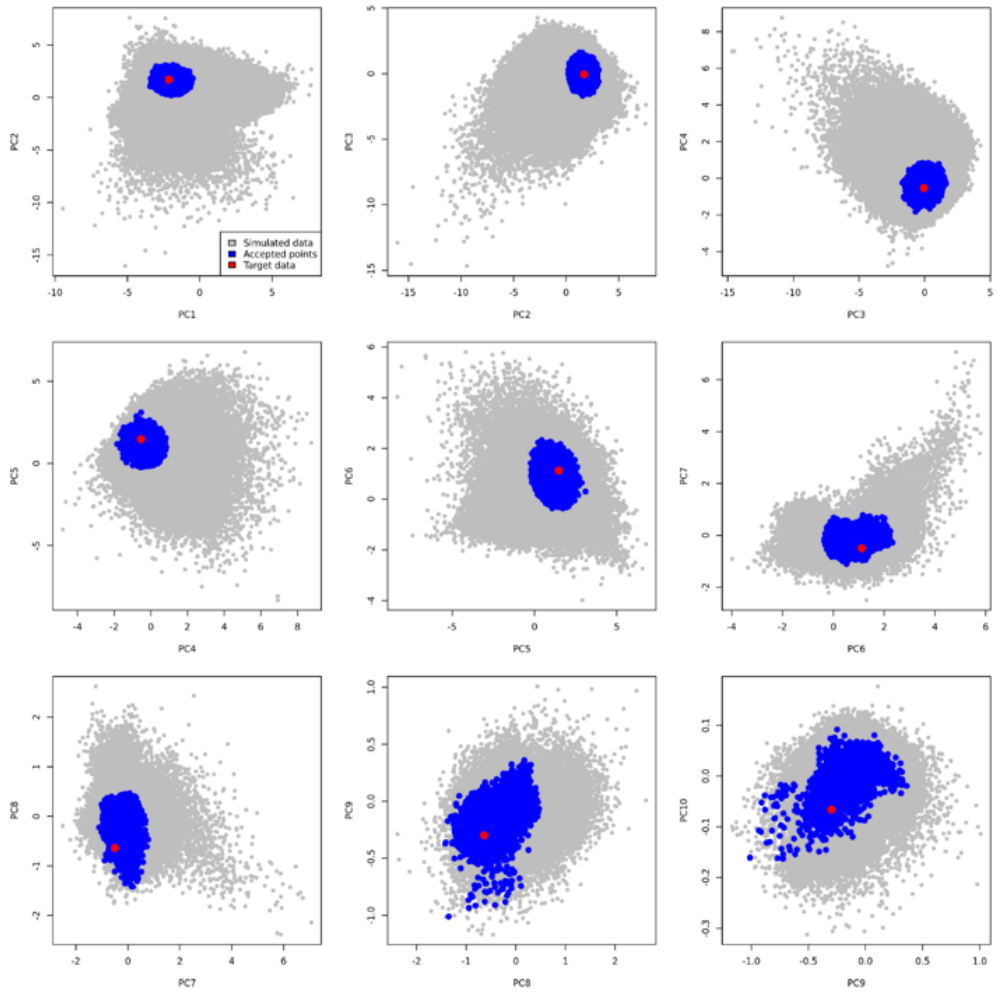
- SNPs scored as a matrix of 0s and 1s (0 means the same as reference sample — a cat called Cinnamon)
- Apply SVD to scaled matrix and obtain PCs (first two tend to mirror geography/demographic history).

Dropping Summary Statistics

- We summarised data using 22 summary statistics.
- 14 summary statistics related to PCA plot (made invariant to reflection).
- 8 summary statistics dropped with approximate HDR method (95% threshold)
- 5 out of 8 dropped summaries related to shape of clusters within PCA
- 8 PCA-related summaries retained — all related to overall shape of PCA.

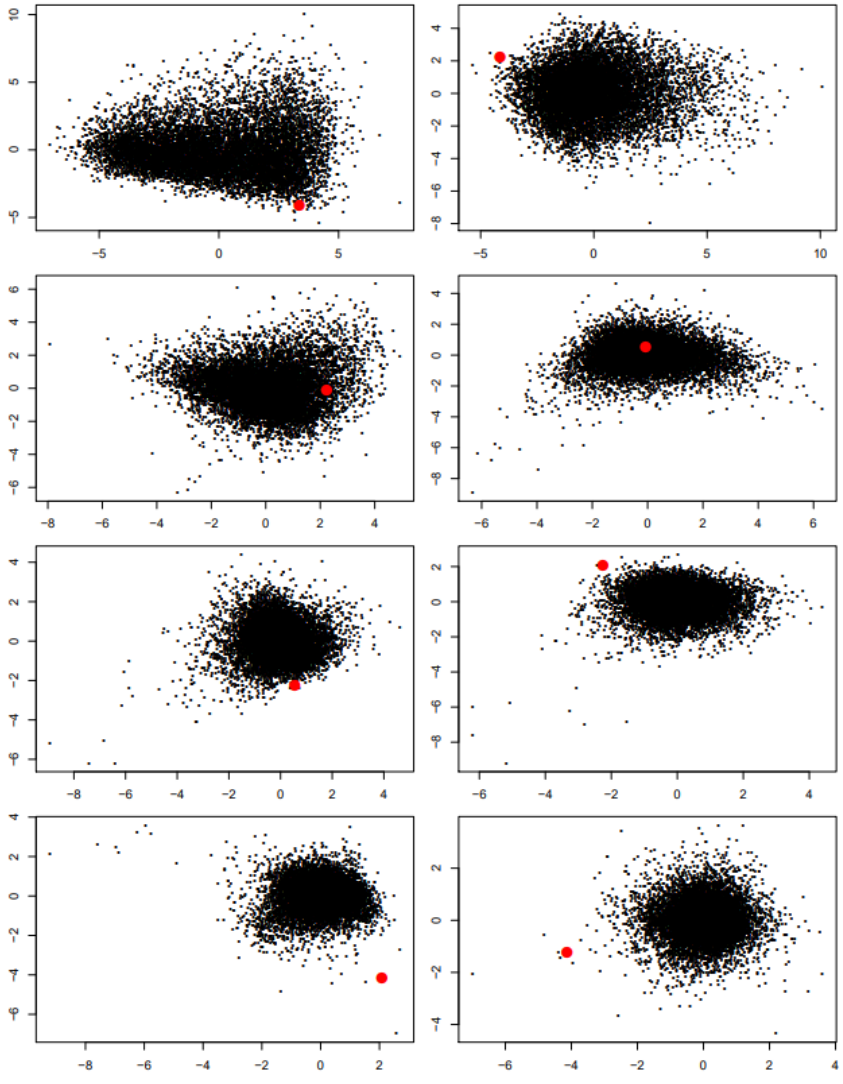
Final Model Fit

- Prior predictive distribution of summary statistics after dropping discrepant summaries.
- Pairwise plots of successive pairs of PCs from PCA rotation.

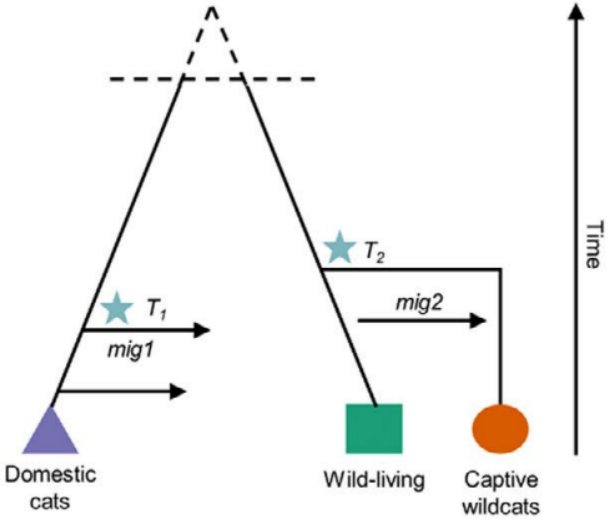


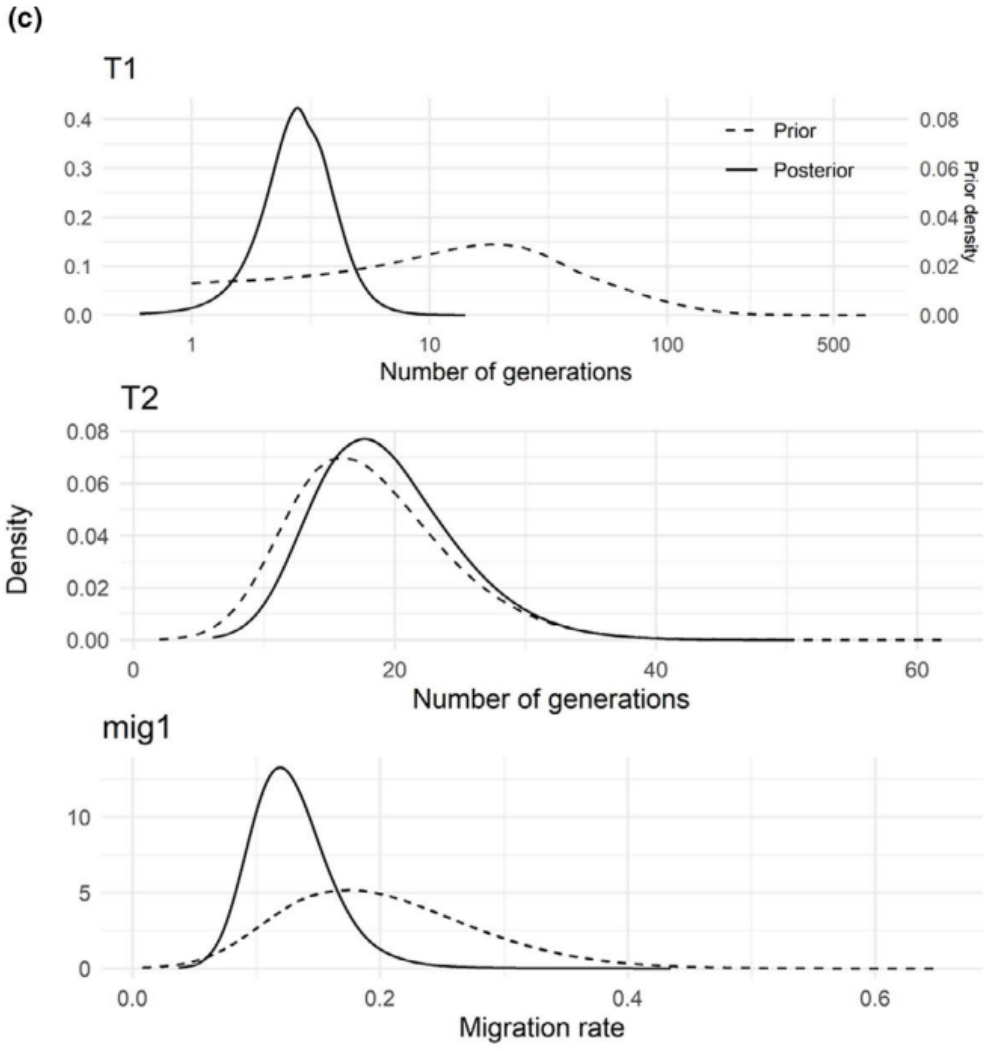
Early Model Fit

Unpublished early PCA plots from project (pairwise for first 9 PCs). Red dot corresponds to observation.

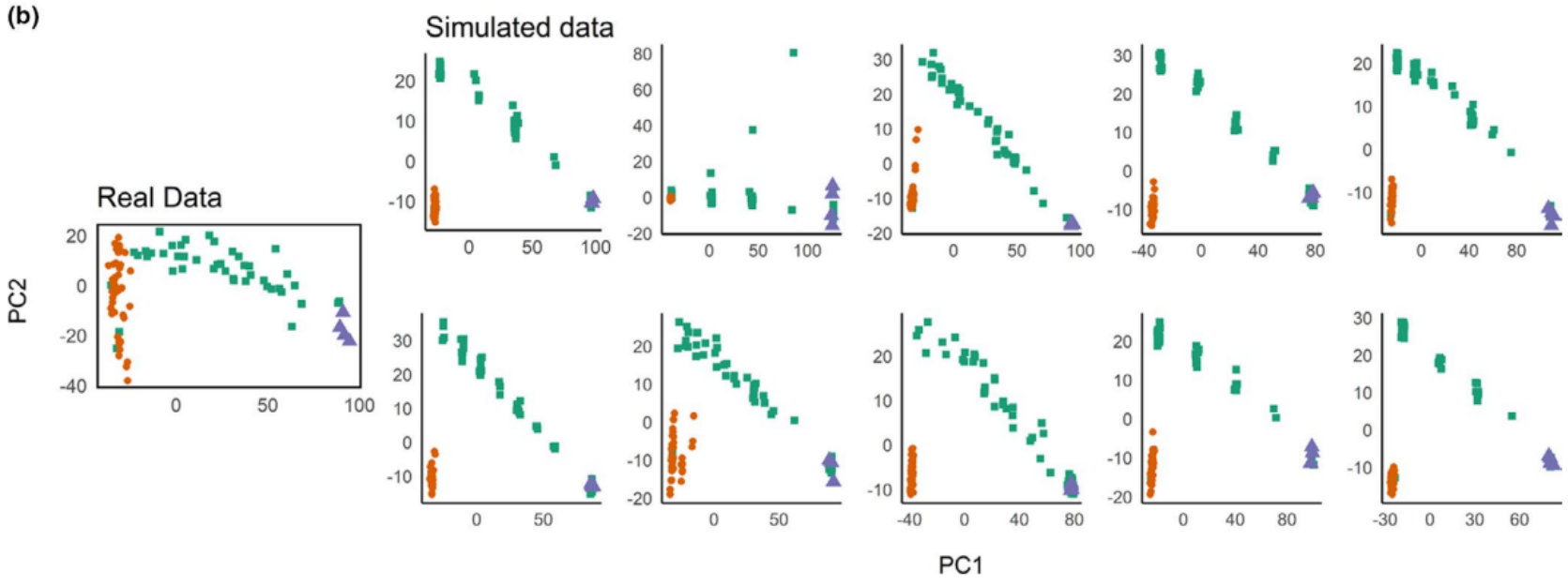


Parameter Estimates





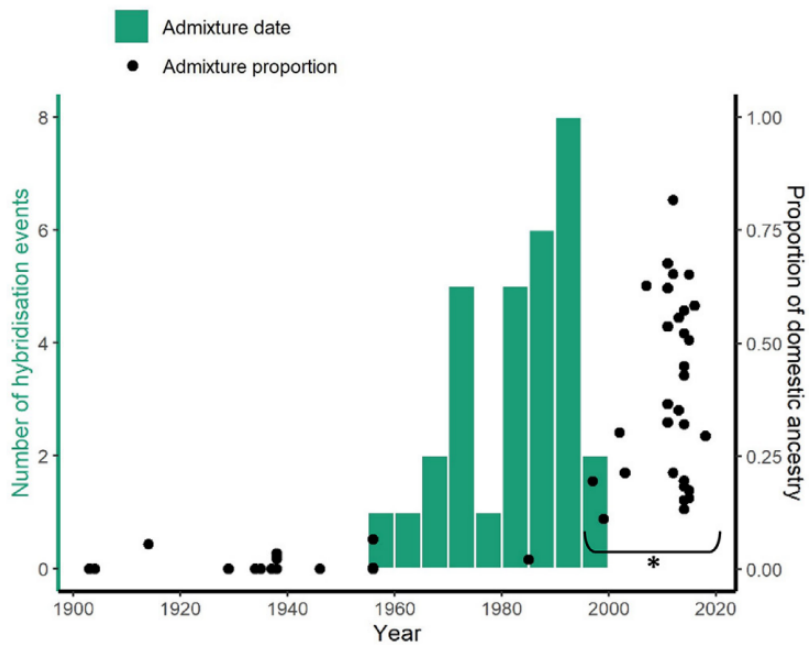
Posterior Predictive Plots



- The model captures the broad shape of plots.
- The spread of hybridisation is well modelled.
- The relationship with domestic cats is well modelled.
- The clustering of captive cats is poorly modelled.

Comparison with Whole-Genome Local Ancestry Estimates

- Using whole genome data we applied a local ancestry modelling approach, implemented in Mosaic.
- Loosely can be considered a non-parametric method.
- Enables sections of genome arising from different populations to be identified.
- Allows timescale of hybridisation to be estimated.



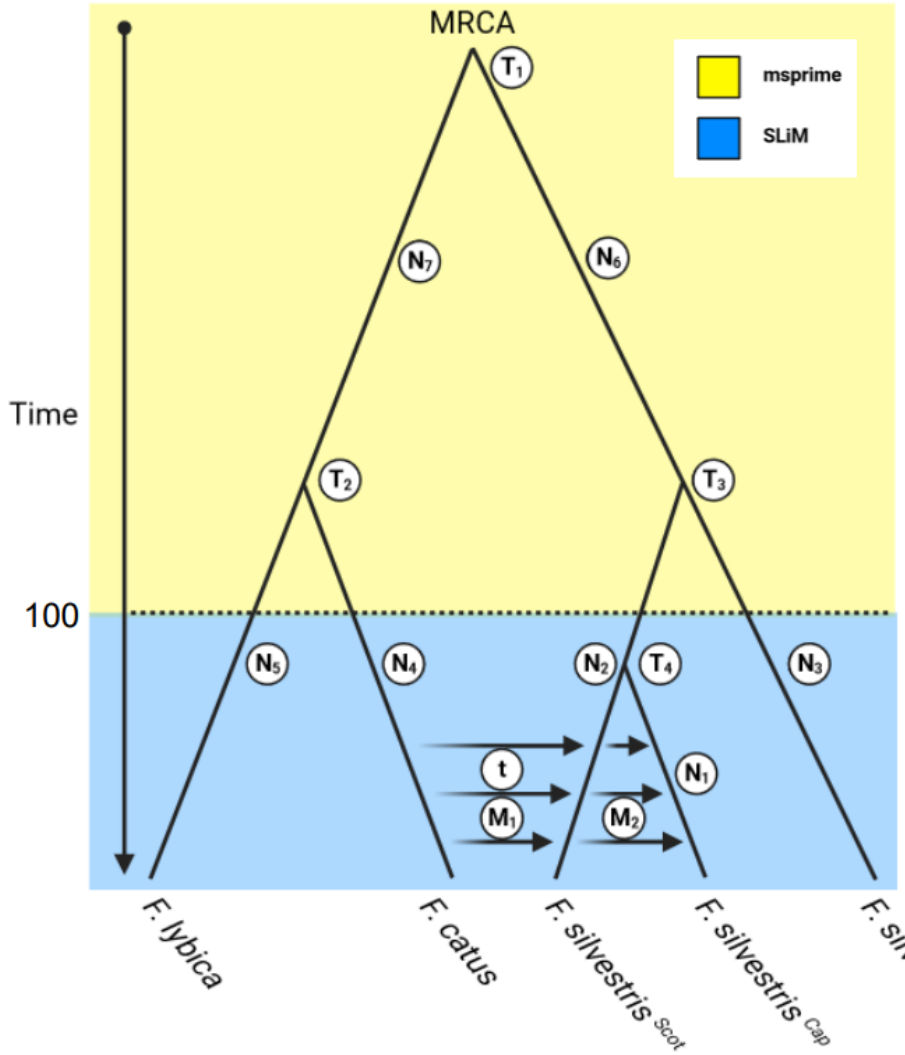
Howard-McCombe *et al* (2023) *Current Biology*.

Application to SBI example

Aims:

- Model whole-genome data using msprime (Kelleher *et al*, 2016) and SLiM (Haller and Messer, 2023)
- Chose a 45Mb chromosome
- Consider more populations
- Date divergence times of European populations
- Use default SNPE from SBI package (Tejero *et al*, 2020)

Demographic Model



N_1: Population size: Captive Scottish *F. silvestris*

N_2: Population size: Scottish *F. silvestris*

N_3: Population size: European *F. silvestris*

N_4: Population size: *F. catus*

N_5: Population size: *F. lybica*

N_6: Ancestral *F. silvestris* population size

N_7: Ancestral *F. lybica* population size

T_1: Divergence of *F. silvestris* and *F. lybica*

T_2: Divergence of *F. catus* (Domestication)

T_3: Divergence of Scottish population

T_4: Captive population established

t: Migration duration: Domestic → Scottish wild

M_1: Migration rate: Domestic → Scottish wild

M_2: Migration rate: Scottish wild → Scottish captive

(Harry Gordon MSc project, in collaboration with Dan Ward, Jo Howard-McCombe, Dan Lawson.)

Fitting with Sequential Neural Posterior Estimation (SNPE)

- Use SNPE-C (Greenberg *et al*, 2019)
- Idea of (S)NPE is to train a neural network $F(\phi, x)$ to approximate conditional density $p(\theta|x)$ by $q_{F(\phi,x)}(\theta)$.
- Train network with (x_i, θ_i) $p(x|\theta)p(\theta)$
- Minimize loss $\mathcal{L}(\phi) = -\sum \log q_{F(\phi,x_j)}(\theta_j)$
- Two NN models compared: Mixture Density Network (MDN) model and the Masked Autoregressive Flow (MAF)
- Higher log-probability with MAF, which was used for subsequent analyses.

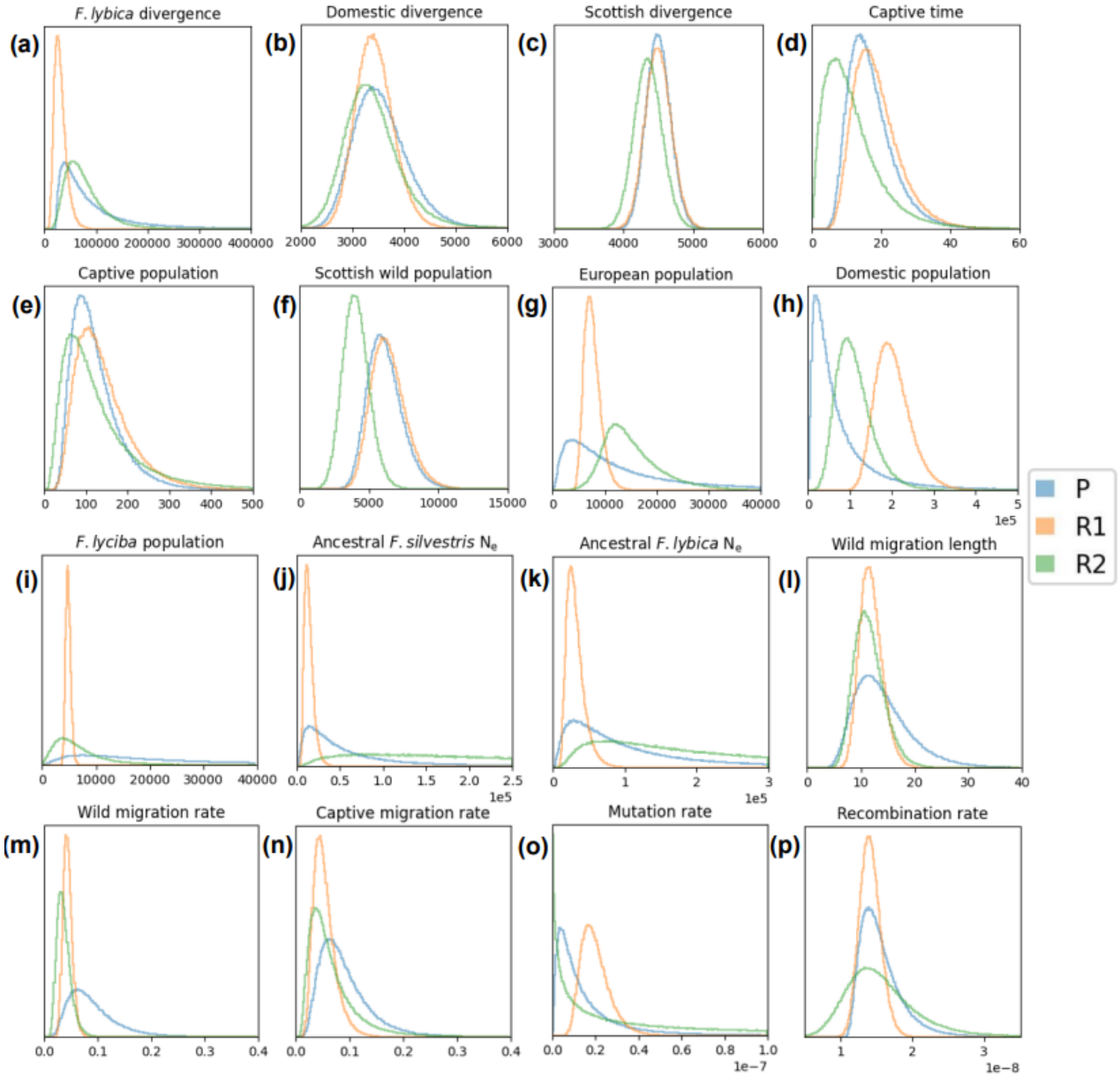
Summary Statistics

- 135 summary statistics computed
- Measures of genetic diversity and between-population divergence.
- Similar summary statistics to Howard-McCombe et al from PCA clustering patterns.

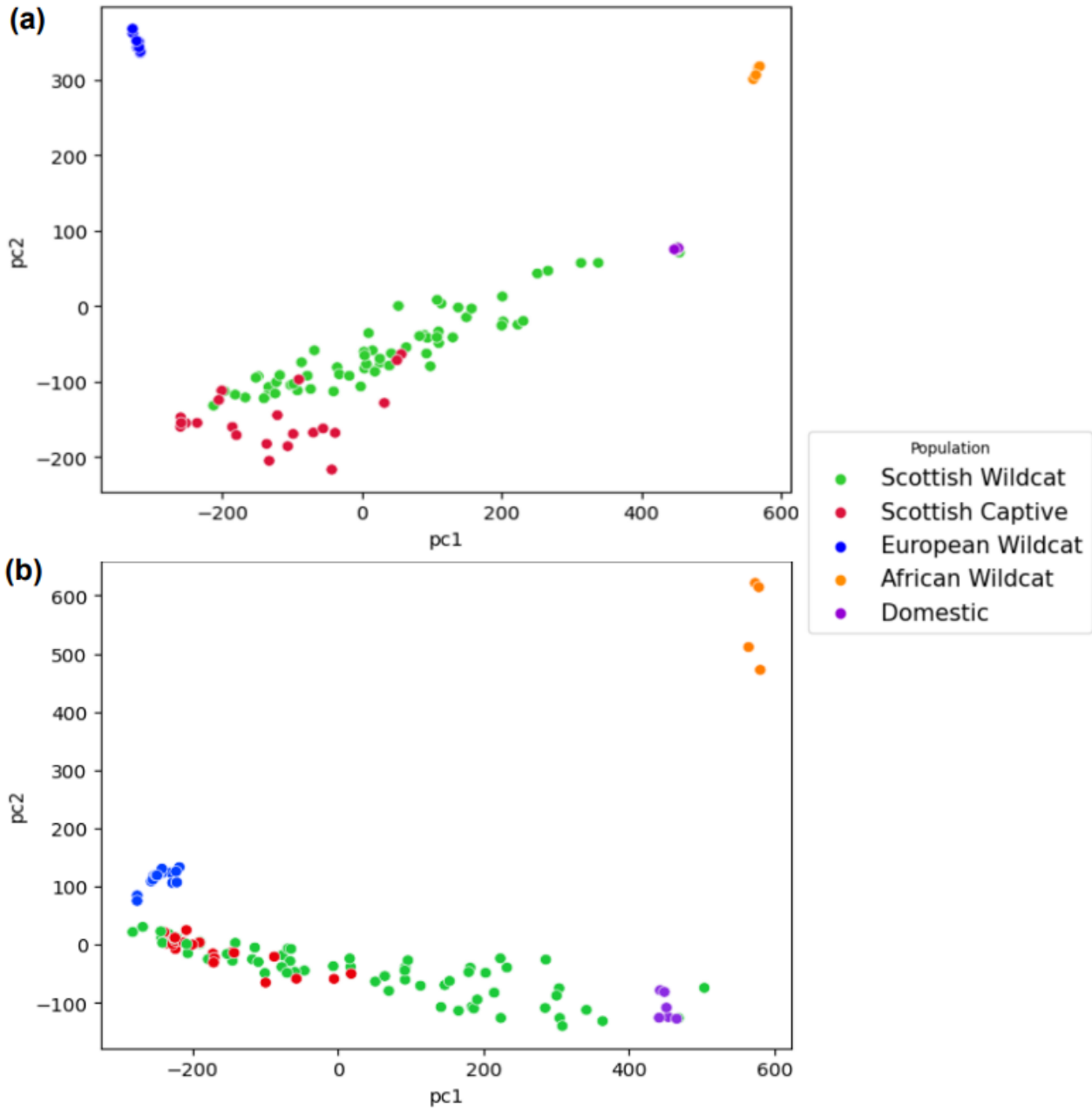
Computational Issues

- To simulate 45 Mb genome for 112 individuals from 5 populations takes 20 minutes to > 2 hours (simulations discarded if taking more than 4 hours).
- Able to use up to 400 cores on HPC
- Limited to training sets from $p(x, \theta)$ of ~ 10000 points for each round.
- Aim is to use sequential NPE to make inference more efficient.
- Compromised by presence of misspecification.

Posterior Distributions



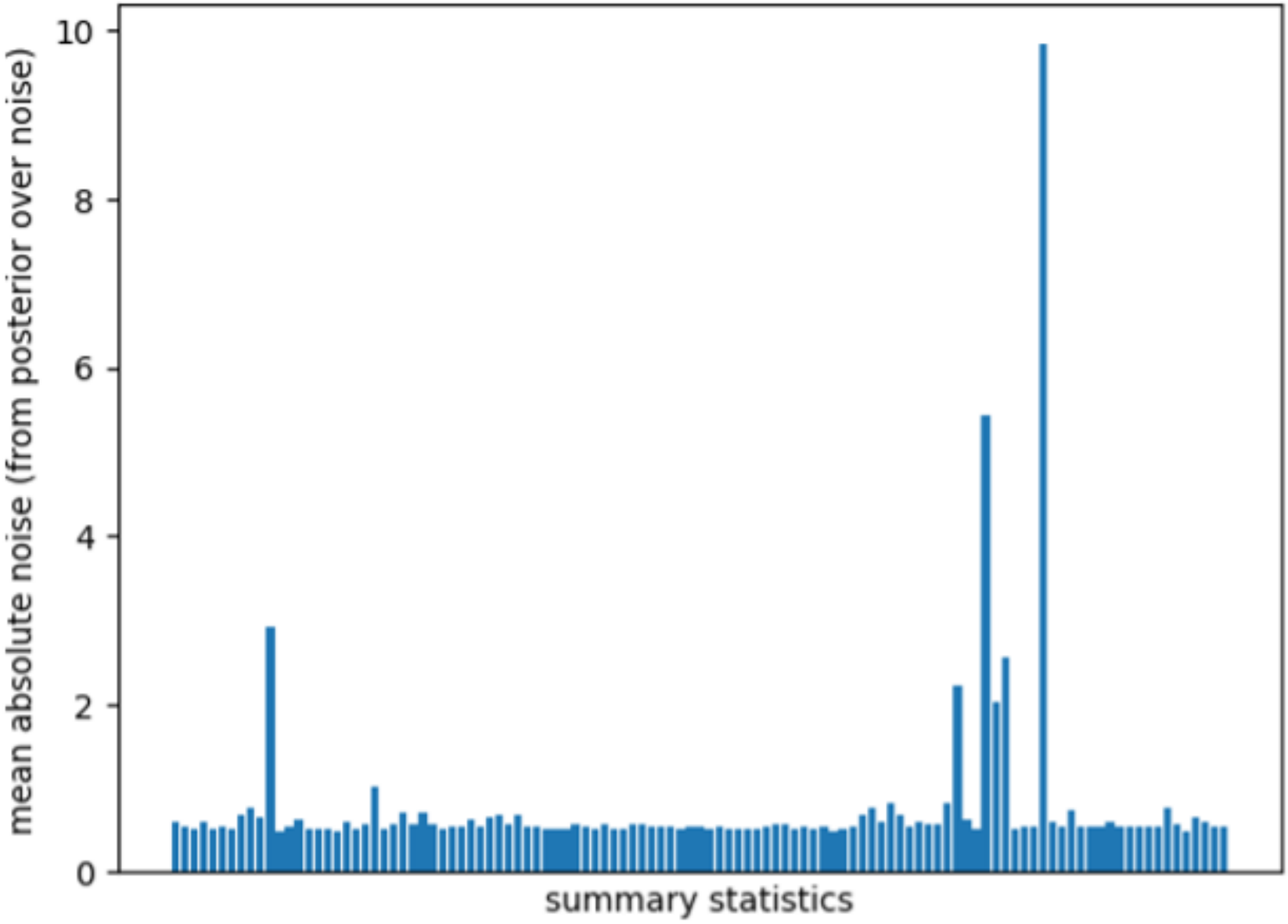
Posterior Predictive Distribution



Approaches to Dropping Summary Statistics

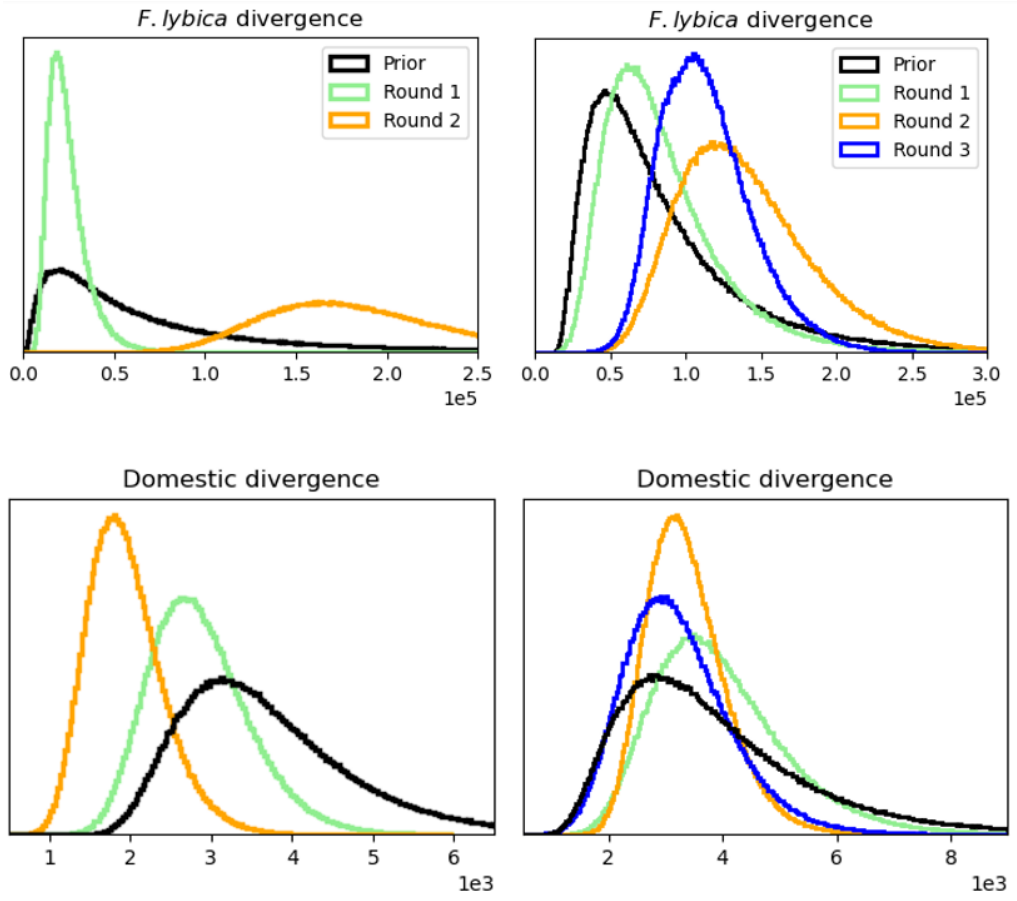
- 23 Summary statistics were dropped because of high correlations ($r > 0{\cdot}99$)
- Use of nearest-neighbour method suggested to drop only 2 summary statistic before reaching a 95% cutoff
- A variant of Ward *et al* (2021) was used (proposed by Dan Ward) where we assume $s_O = s_o + \epsilon$; train a flow to approximate $p(s)$ (using samples from the prior predictive); define a prior over the noise Laplace$(0, 1)$; then infer $p(\epsilon|S_O)$ using MCMC.
- This removed a further 10 summaries
- However, computing HDR from the flow-based estimate of $p(s_O)$ suggests that observations are at $> 0{\cdot}99$ quantile, so further work is needed

Example with Ward's method



Example of current status

- Currently able to carry out 4 rounds of simulation without substantial divergence
- Left plots shows original case; on the right after removing problematic summary statistics. Example of 2 parameters shown.



Current Project Aims

- Pursue the HDR quantile idea, but using flow-based estimate of $p(s_i)$ and $p(s_O)$ rather than nearest-neighbour method.
- Examine posterior predictive distributions for further rounds of SNPE.

Citations I

[1] Paul Fearnhead and Dennis Prangle. "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 74.3 (2012), pp. 419–474.

[2] David T Frazier and Christopher Drovandi. "Robust approximate Bayesian inference with synthetic likelihood". In: *Journal of Computational and Graphical Statistics* 30.4 (2021), pp. 958–976.

[3] David T Frazier, Christian P Robert, and Judith Rousseau. "Model misspecification in approximate Bayesian computation: consequences and diagnostics". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.2 (2020), pp. 421–444.

Citations II

[4] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. "Automatic posterior transformation for likelihood-free inference". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2404–2414.

[5] Benjamin C Haller and Philipp W Messer. "SLiM 4: multispecies eco-evolutionary modeling". In: *The American Naturalist* 201.5 (2023), E127–E139.

[6] Jo Howard-McCombe et al. "Genetic swamping of the critically endangered Scottish wildcat was recent and accelerated by disease". In: *Current Biology* 33.21 (2023), pp. 4761–4769.

[7] Jo Howard-McCombe et al. "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* 30.15 (2021), pp. 3688–3702.

Citations III

[8] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. "Efficient coalescent simulation and genealogical analysis for large sample sizes". In: *PLoS computational biology* 12.5 (2016), e1004842.

[9] Oliver Ratmann et al. "Model criticism based on likelihood-free inference, with an application to protein network evolution". In: *Proceedings of the National Academy of Sciences* 106.26 (2009), pp. 10576–10581.

[10] Alvaro Tejero-Cantero et al. "SBI–A toolkit for simulation-based inference". In: *arXiv preprint arXiv:2007.09114* (2020).

[11] Elske van der Vaart et al. "Calibration and evaluation of individual-based models using Approximate Bayesian Computation". In: *Ecological Modelling* 312 (2015), pp. 182–190.

Citations IV

[12] Daniel Ward et al. "Robust neural posterior estimation and statistical model criticism". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 33845–33859.

[13] RICHARD D WILKINSON. "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error". In: *arXiv preprint arXiv:0811.3355* (2008).

[14] Richard David Wilkinson. "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error". In: *Statistical applications in genetics and molecular biology* 12.2 (2013), pp. 129–141.