# Approximate Bayesian Computation: algorithms, theory and applications

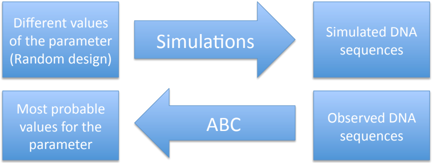
Michael G.B. Blum

Laboratoire TIMC-IMAG, UJF Grenoble, CNRS

MCEB, June 2012

"ABC is a 'democratizing' method in that it will attract, for example, biologists, who enjoy computer simulation but have little background in probability, into converting their favorite simulation into a tool for inference"

Beaumont and Rannala, Nat. Rev. Genet. 2004


Different values of the parameter (Random design)

Simulations

Simulated DNA sequences

Most probable values for the parameter

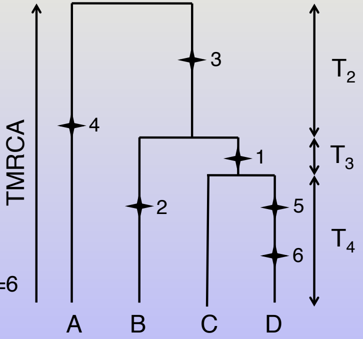
ABC

Observed DNA sequences

# Parameter inference for ABC

# A coalescent example in population genetics
# Estimating the mutation rate $\theta$

| | Segregating sites |
|---|---|
| | 123456 |
| A | 000100 |
| B | 011000 |
| C | 101000 |
| D | 101011 |



Number of segregating sites: S=6

Heterozygosity: H=3.16

# Two approximations in ABC

- Replace *full posterior* $p(\theta|D)$ with *partial posterior* $p(\theta|\mathbf{s}_{obs})$
- Nonparametric estimation of $p(\theta|\mathbf{s}_{obs})$

# Rejection algorithm

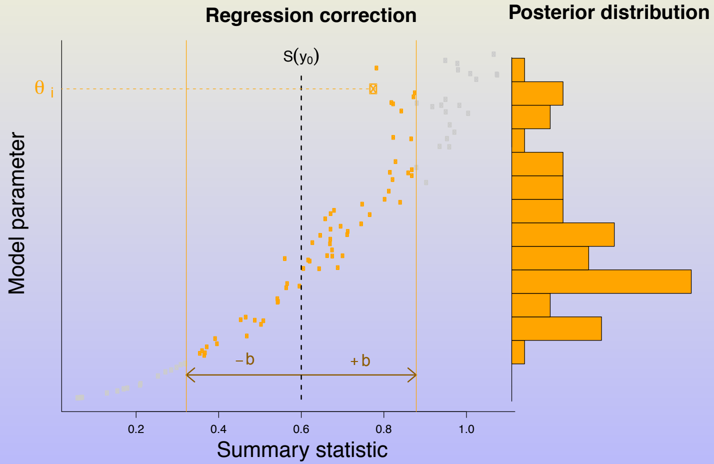
`Pritchard et al., MBE 1999`

- Simulate $n$ values $\theta_i$, $i = 1, \ldots, n$ from the prior $\pi$
- Simulate $n$ (possibly multivariate) summary statistics $\mathbf{s_i}$ according to $p(\mathbf{s_i}|\theta_i)$
- Consider the weighted sample $(\theta_i, W_i)$, $i = 1, \ldots, n$

$$W_i = \begin{cases} 1 \text{ if} \|\mathbf{s}_i - \mathbf{s}_{obs})\| \leq b \\ 0 \text{ otherwise.} \end{cases}$$

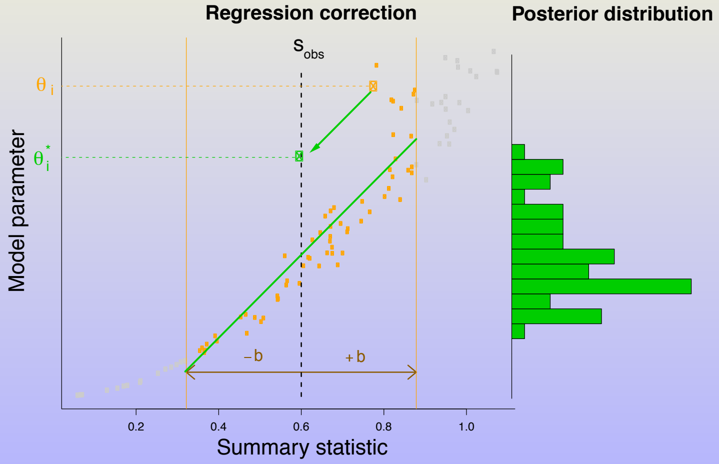
The parameter $b$ is an acceptance threshold.

# Rejection algorithm



# Regression adjustment

Beaumont et al., Genetics 2002



# Linear regression adjustment

- A model of local regression

$$\theta_i|\mathbf{s}_i = m(\mathbf{s}_i) + \epsilon_i$$

- Local linear approximation

$$m(\mathbf{s}_i) = \alpha + \mathbf{s}_i^t\boldsymbol{\beta}$$

- Adjustment

$$\theta_i^* = \hat{m}(\mathbf{s}_{obs}) + \tilde{\epsilon}_i,$$

# Main theorem

`Blum, JASA 2010`

Asymptotic bias of the estimated posterior mean

$j = 0$ rejection, $j = 1$ linear adjustment

$$C_{1,j} b^2$$

Asymptotic variance

$$\frac{C_3}{nb^d}$$

$d$ is the number of the statistics and $n$ is the number of simulations

Overemphasizes the curse of dimensionality because empirical evidence are much more optimistic.

# Comparison between the two estimators with adjustment

When the model

$$\theta_i = m(\mathbf{s}_i) + \epsilon_i$$

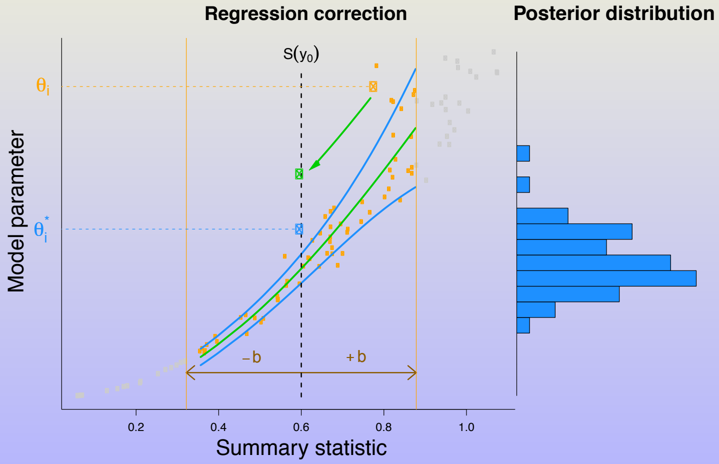
is homoscedastic in the vicinity of $\mathbf{s}_{obs}$, the bias for the estimator with quadratic adjustment is

$$o(b^2).$$

- Transformations of the sum stat to make the model as homoscedastic as possible.
- non-linear adj.
- Non-homoscedastic adjustment

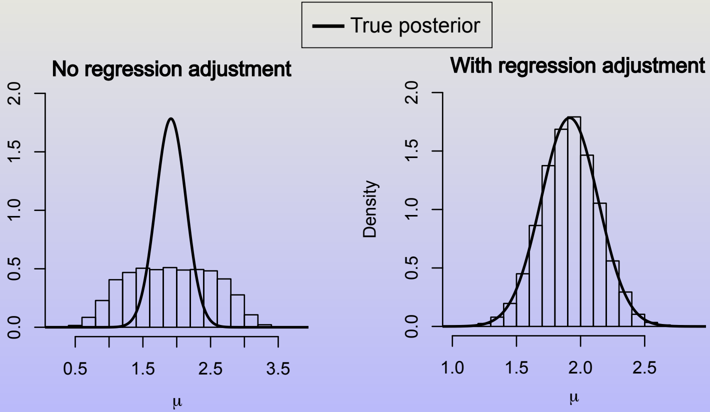
# Regression adjustment for the mean and the variance

`Blum and François, Stat and Comput 2010`



# ABC with and without adjustment

## Estimating the mean in a Gaussian sample



# How to check that ABC works when you do not know the posterior

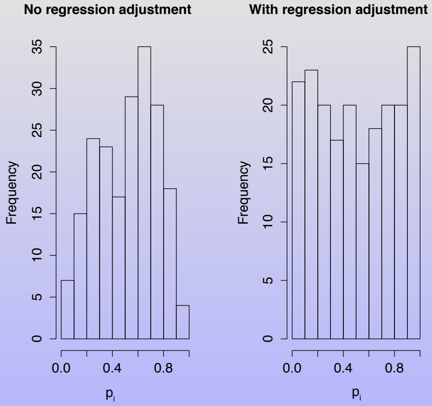
`Cook et al. 2006 J. Comp. Graph. Stat.`

- Take a $(\theta_i, \mathbf{s_i})$ drawn from $\pi(\theta_i)p(\mathbf{s_i}|\theta_i)$.
- Perform ABC with $\mathbf{s}_{obs} = \mathbf{s}_i$.
- Compute the proportion $p_i$ of posterior samples smaller than $\theta_i$.

If the algorithm provide samples from $p(\theta_i|\mathbf{s_i})$, $p_i$ should be uniformly distributed.

# How to check that ABC works when you do not know the posterior

`Cook et al. 2006 J. Comp. Graph. Stat.`



# Adaptive ABC

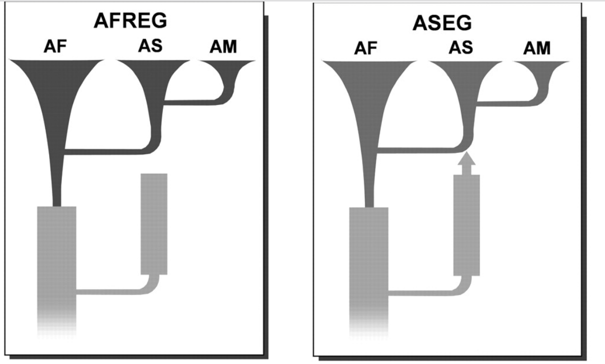
`Sisson et al., PNAS 2007; Beaumont et al, Biometrika 2009; Del Moral et al. Stat and Comput 2011`

Multi-step algorithms that sample $\theta$ from updated distributions that get closer and closer to the posterior distribution.

# Model selection and related criticisms

# Distinguishing between models An exemple in human evolution

`Fagundes et al., PNAS 2007`



# Rejection algorithm

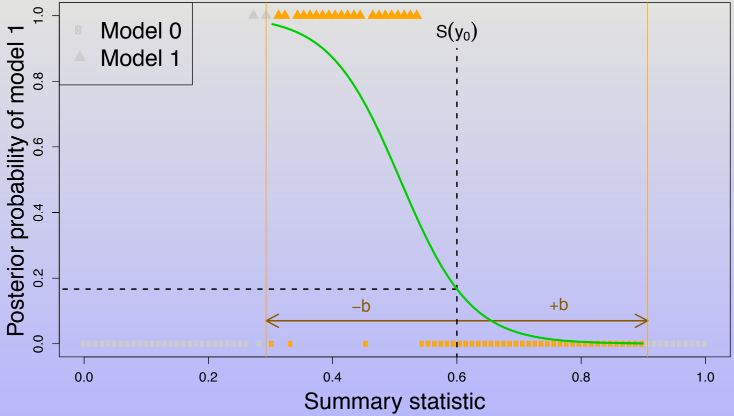
`Pritchard et al., MBE 1999`

- Simulate the same number of simulations under each model $\mathcal{M}_k$, $k = 1, \dots, K$.
- Accept the simulations for which $\|\mathbf{s}_i - \mathbf{s}_{obs}\| \leq b$.

The proportion of accepted simulations under each model $k = 1, \dots, K$ is an estimate of the posterior distribution $p(\mathcal{M}_k|\mathbf{s}_{obs})$.

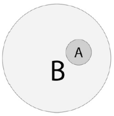
# Model selection with logistic regression

`Beaumont 2008`



# Criticisms about model selection with ABC

`Templeton, PNAS 2010`



‘The probability of the nested special case must be less than or equal to the probability of the general model within which the special case is nested’.

- In a proper Bayesian framework, two hypotheses represent non-nested events.
- Coin tossing example. $\mathcal{M}_0 : q = 0.5$, $\mathcal{M}_1 : q \rightsquigarrow \mathcal{U}(0,1)$. 9 heads out of 20 tosses.

$$p(\mathcal{M}_0|9/20) \approx 3.36 p(\mathcal{M}_1|9/20)$$

# Criticisms about model selection with ABC

`Robert et al., PNAS 2011`

- ‘The algorithm involves an unknown loss of information induced by the use of insufficient summary statistics’.
- Assume that **s** is sufficient for parameter inference in model 0 and model 1.

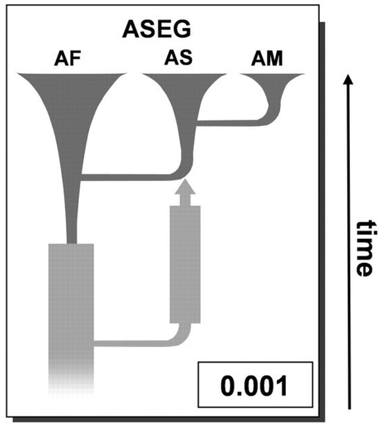
$$\frac{p(\mathcal{M}_0|\mathbf{s}_{obs})}{p(\mathcal{M}_1|\mathbf{s}_{obs})} = g(D)\frac{p(\mathcal{M}_0|D)}{p(\mathcal{M}_1|D)},$$

with $g(D)$ possibly different from 1.

# Some elements of answers if your NSF reviewer keeps bothering you with this.

- The criticism pertains to situation where **s** is sufficient for parameter inference in model 0 and model 1. Exercise : For coalescent models, try to find one for a constant-size population model and a bottleneck model.
- In approximate Bayesian Computation, we target $p(\mathcal{M}|\mathbf{s})$ instead of $p(\mathcal{M}|D)$.
- An important question to address might be ‘Does **s** contain enough information to distinguish between models’.

# Model selection with ABC : the right answer ?



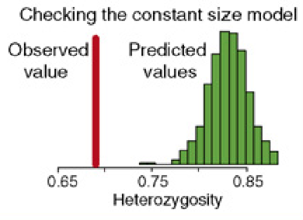
Fagundes et al., PNAS 2007

# A deviance criterion for model selection with ABC

François and Laval, SAGMB 2011

Estimators of $p(\mathcal{M}|\mathbf{s})$ ignore regression adjustments on parameter samples

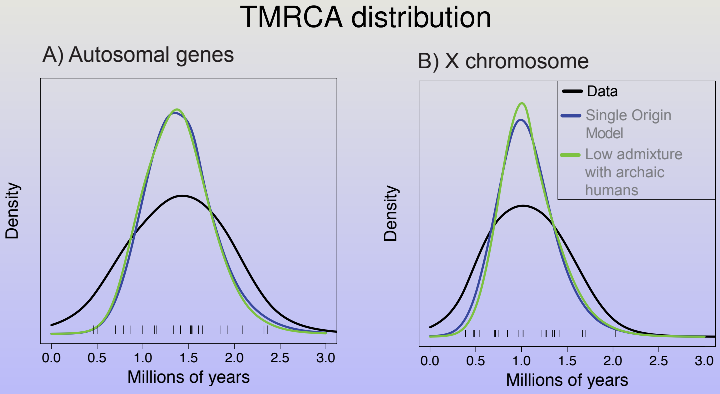
$$\mathrm{DIC} = E_{\mathrm{Post}}[\mathrm{deviance}] + \text{effective number of parameters}$$



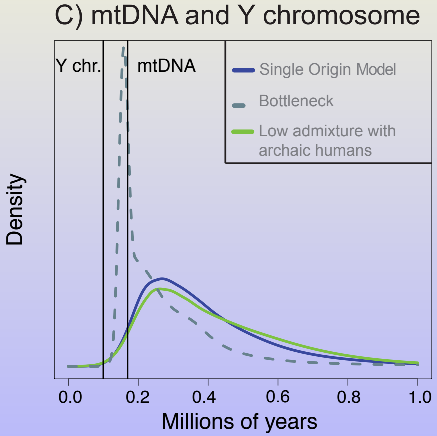
# 2-3 examples

# Example 1 : Models of origins for modern humans

Blum and Jakobsson, MBE 2011

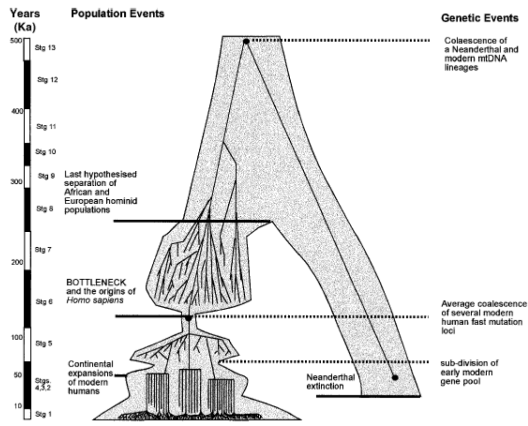


# Models of origins for modern humans



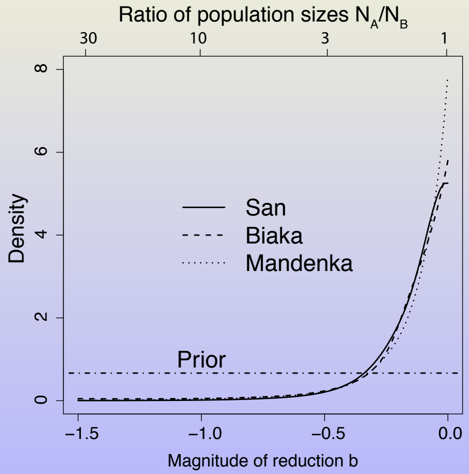
# Example 1 : Testing the human ‘speciation’ bottleneck

Sjödin et al., MBE 2012



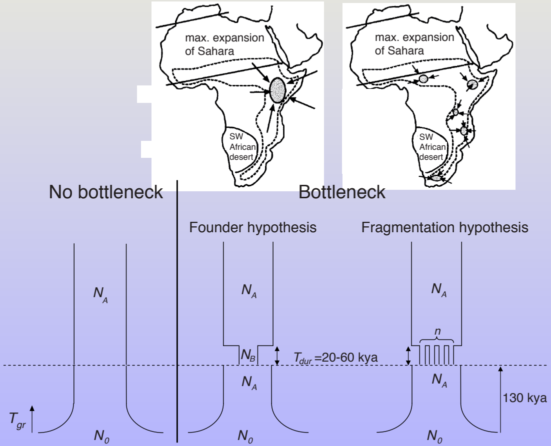
Lahr and Foley 1998

# Estimating the strength of the bottleneck

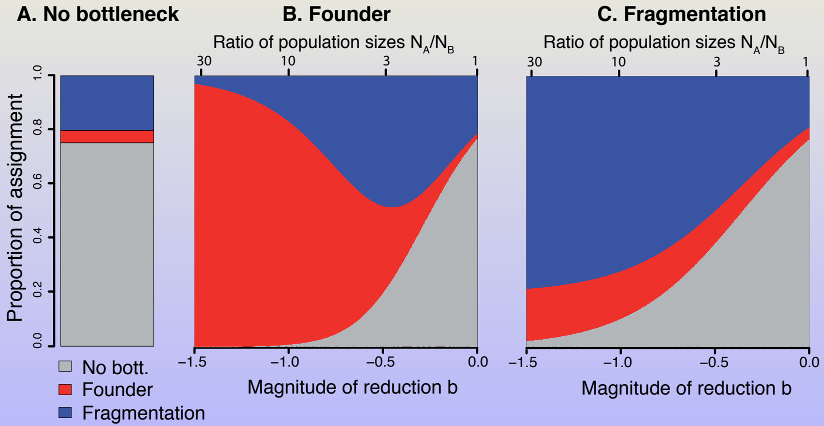


# Support for a ‘no-bottleneck’ model against 2 bottleneck models

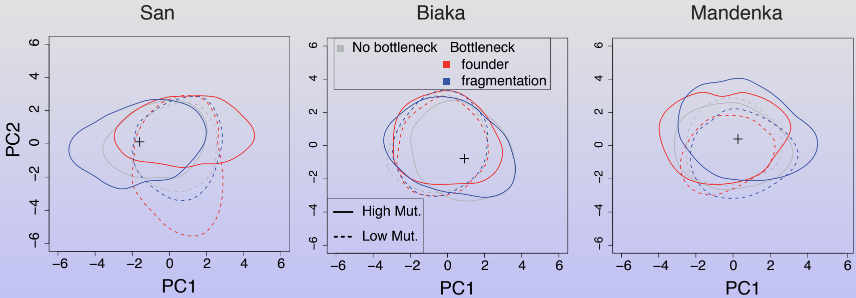
$$Pr(no\ bottleneck|\mathbf{s_{obs}}) \geq 79\%$$



# Is it possible to distinguish between models ?

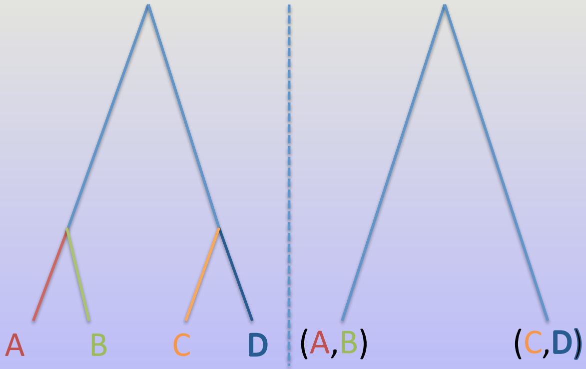


# Goodness of fit
Posterior predictive checks + PCA

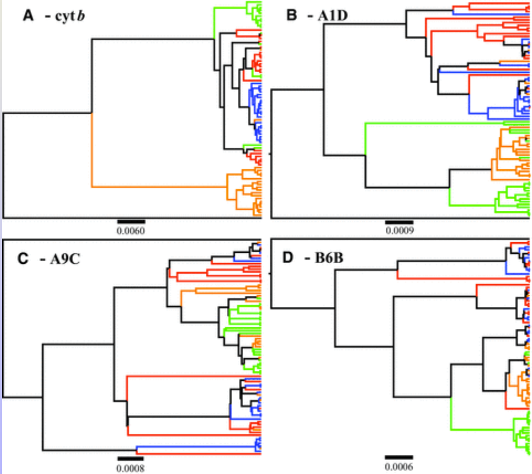


# Example 2 : Species delimitation with ABC

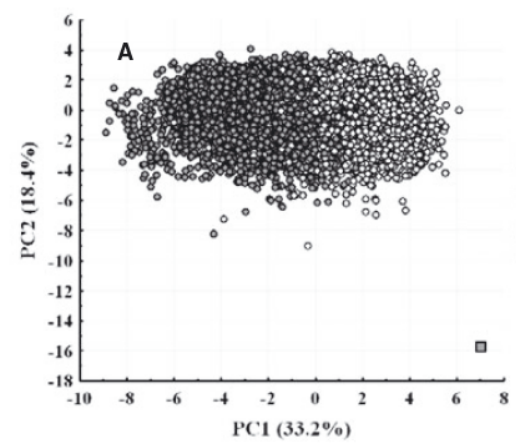
Camargo et al., Evolution 2012



# Gene trees of loci sampled for species delimitation analyses



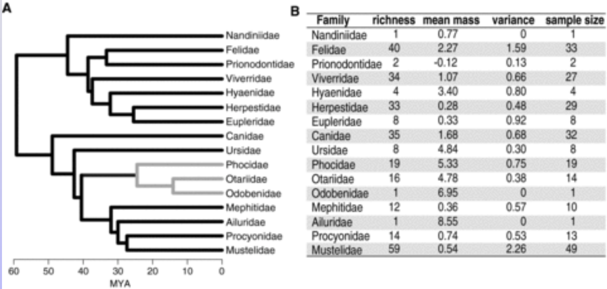
# Prior predictive checks



# Example 3 : Fitting models of continuous trait evolution

`Slater et al., Evolution 2012`

Time-calibrated phylogeny of Carnivora used to estimate rates of trait evolution



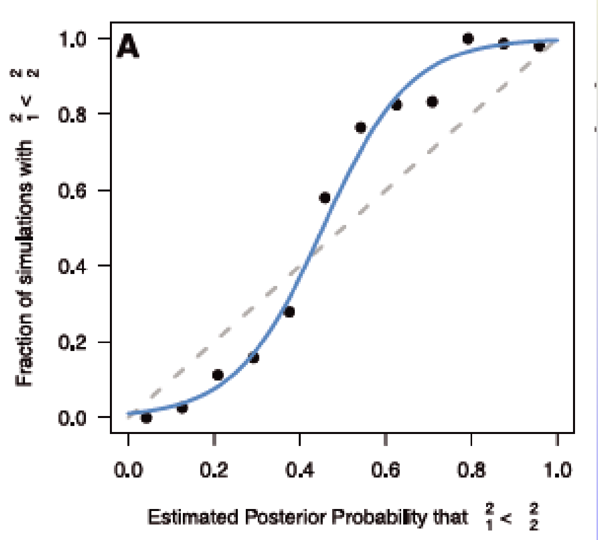
| Family | richness | mean mass | variance | sample size |
|---|---|---|---|---|
| Nandiniidae | 1 | 0.77 | 0 | 1 |
| Felidae | 40 | 2.27 | 1.59 | 33 |
| Prionodontidae | 2 | -0.12 | 0.13 | 2 |
| Viverridae | 34 | 1.07 | 0.66 | 27 |
| Hyaenidae | 4 | 3.40 | 0.80 | 4 |
| Herpestidae | 33 | 0.28 | 0.48 | 29 |
| Eupleridae | 8 | 0.33 | 0.92 | 8 |
| Canidae | 35 | 1.68 | 0.68 | 32 |
| Ursidae | 8 | 4.84 | 0.30 | 8 |
| Phocidae | 19 | 5.33 | 0.75 | 19 |
| Otariidae | 16 | 4.78 | 0.38 | 14 |
| Odobenidae | 1 | 6.95 | 0 | 1 |
| Mephitidae | 12 | 0.36 | 0.57 | 10 |
| Ailuridae | 1 | 8.55 | 0 | 1 |
| Procyonidae | 14 | 0.74 | 0.53 | 13 |
| Mustelidae | 59 | 0.54 | 2.26 | 49 |

# Comparing a two-rate model $\mathcal{M}_2$ to a one-rate model $\mathcal{M}_1$

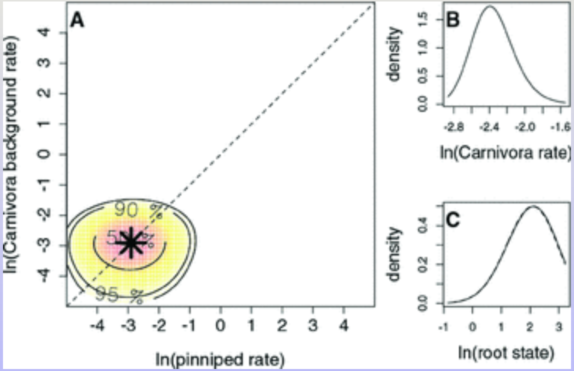
If $p(\mathcal{M}_1|\mathbf{s}) = x\%$, there is a probability of $x\%$ that **s** was generated from $\mathcal{M}_1$ (and $1 - x\%$ that **s** was generated from $\mathcal{M}_2$).

`Cook et al. 2006 in a model selection framework.`

# Checking the ‘consistency’ of the Bayes factor



# Comparing pinniped and terrestrial carnivore body size evolutionary rates



# Conclusion

- ABC incorporates all aspects of Bayesian data analysis : formulation, fitting and model selection, and improvement of a model through model checking
  `Csilléry et al., TREE 2010.`
- To address issues related to model selection
  1. Ability to distinguish between models
  2. ‘Consistency’ of the Bayes factor
  3. Buy a bottle of wine to the reviewer
- The *R* package **abc** implements several ABC algorithms.

# Colleagues


Olivier François


Mattias Jakobsson


Katalin Csilléry


Per Sjödin