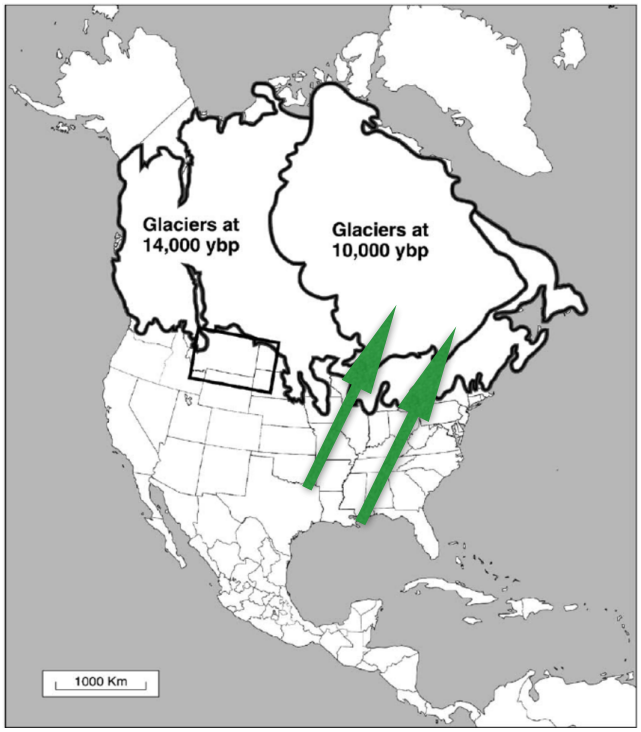
HABC - step by step part 1

Michael Hickerson
City University of New York
City College of New York
American Museum of Natural History

HABC - worked example



- How did a temperate assemblage respond to Holocene warming?

Ecology Letters, (2016) doi: 10.1111/ele.12695

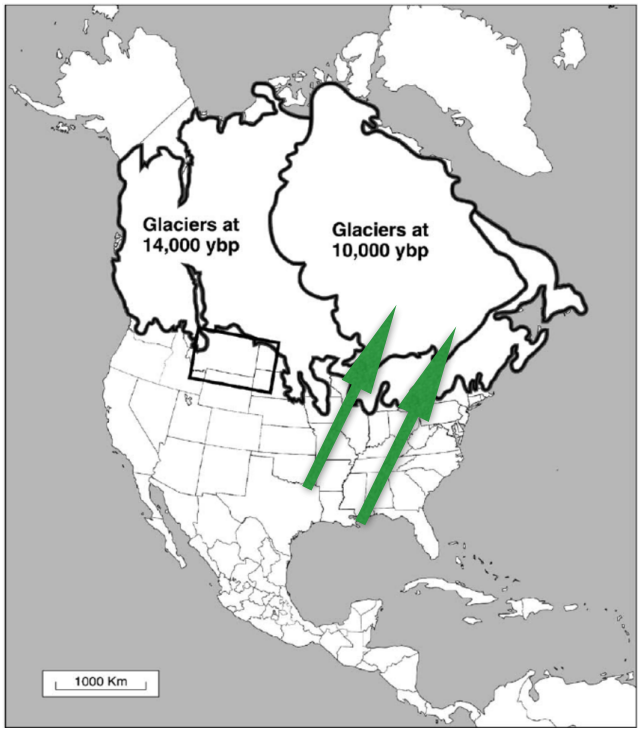
LETTER

Asynchronous demographic responses to Pleistocene climate change in Eastern Nearctic vertebrates

Frank T. Burbrink,[1]* Yvonne L. Chan,[2] Edward A. Myers,[3,4] Sara Ruane,[5] Brian Tilston Smith[6] and Michael J. Hickerson[4,7,8]

Yvonne Chan - U.Hawaii

Brian Smith - AMNH



Frank Burbrink - AMNH

- How did a temperate assemblage respond to Holocene warming?

90 lineages:

birds
mammals
frogs
salamanders
lizards
snakes

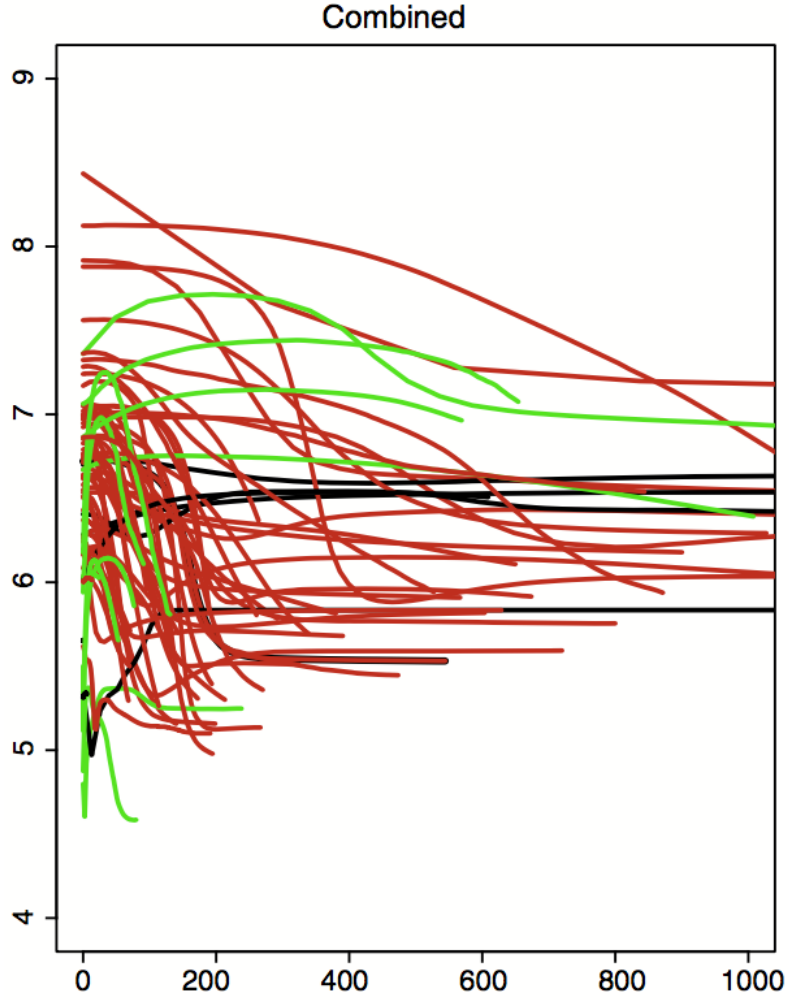
mtDNA

10 - 400 individuals per taxon

hierarchical Bayesian model

> 90 Skyline plots (eastern Nearctic - w/ mtDNA)

Effective Pop Size



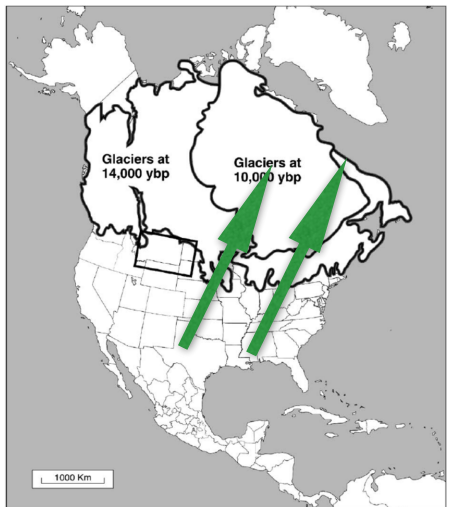
Time (years)

– Impractical for large number of species
– Difficult to statistically compare results

hierarchical Bayesian model

$$P(\zeta,\tau,\phi \mid \textbf{\textit{Data}}) \quad \alpha \quad p(\textbf{\textit{Data}} \mid \zeta,\tau,\phi\,) = p(\textbf{\textit{Data}}|\phi,\tau)\; p(\phi)\; p(\tau\,|\zeta)\; p(\zeta)$$

flexible
within species uncertainty
across species estimates



Chan et al. 2014
Xue and Hickerson 2014
Xue and Hickerson 2017

hierarchical Bayesian model

$$P(\zeta,\tau,\phi \mid \textbf{\textit{Data}}) \;\; \alpha \;\; p(\textbf{\textit{Data}} \mid \zeta,\tau,\phi\,) = p(\textbf{\textit{Data}}|\phi,\tau)\; p(\phi)\; p(\tau\,|\zeta)\; p(\zeta)$$

hyperparameters

ζ = **Co-expansion Proportion**

τ_S = Co-expansion *time*

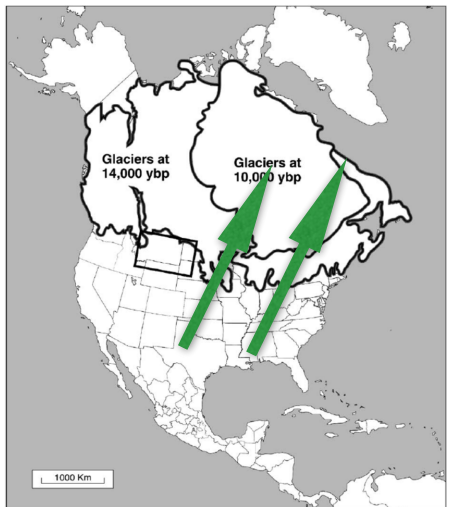
species-specific parameters (ϕ,τ)

τ = expansion times

$\{\tau_1, \ldots, \tau_n\}$

$\phi = N$,expansion magnitudes (independent)

$\{\phi_1, \ldots, \phi_n\}$ $\{N_1, \ldots, N_n\}$



Chan et al. 2014

Xue and Hickerson 2014

hierarchical Bayesian model

$$P(\zeta,\tau,\phi \mid \textbf{Data}) \;\; \alpha \;\; p(\textbf{Data} \mid \zeta,\tau,\phi\,) = p(\textbf{Data}|\phi,\tau)\; p(\phi)\; p(\tau\,|\zeta)\; p(\zeta)$$

hyperparameters

ζ = **Co-expansion Proportion**

τ_S = Co-expansion *time*

species-specific parameters (ϕ,τ)

$\phi = N$, expansion magnitude

τ = expansion time

A) Synchronous expansion

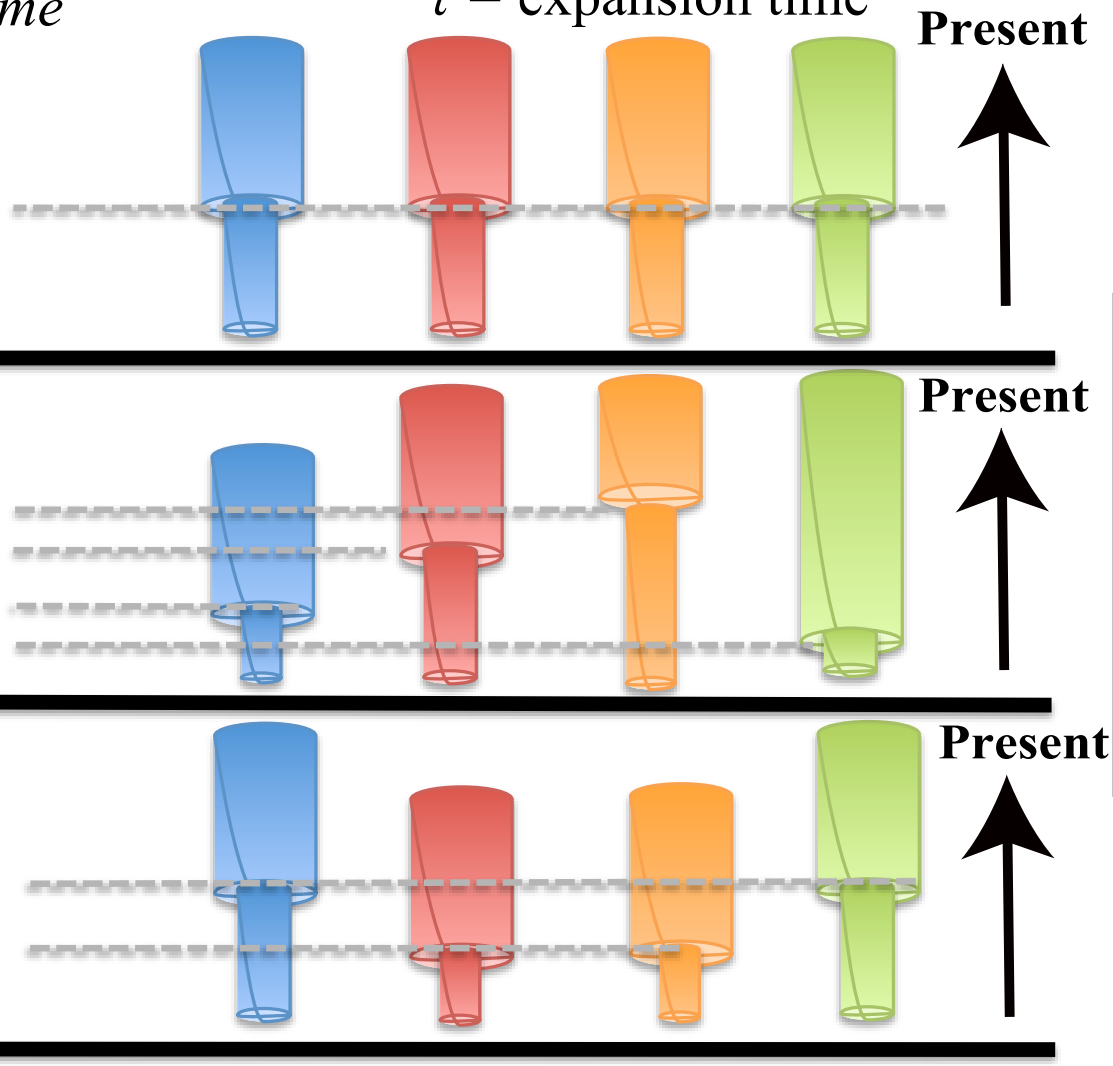
$\zeta = 1.0$

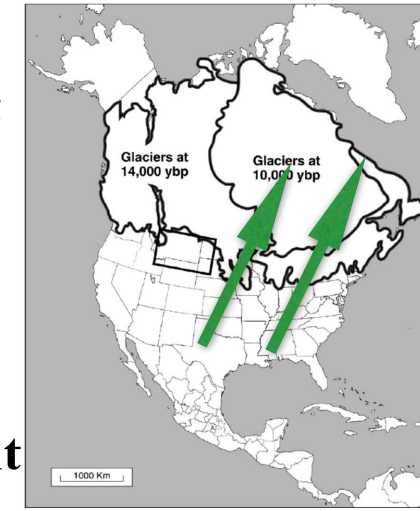
B) Asynchronous expansion

$\zeta = 0.0$

C) Expansion mixture: asynchronous & synchronous

$\zeta = 0.5$





Chan et al. 2014

Xue and Hickerson 2014

$$P(\zeta,\tau,\phi \mid \textbf{\textit{Data}}) \quad \alpha \quad p(\textbf{\textit{Data}} \mid \zeta,\tau,\phi\ ) = p(\textbf{\textit{Data}}|\phi,\tau)\ p(\phi)\ p(\tau\ |\zeta)\ p(\zeta)$$

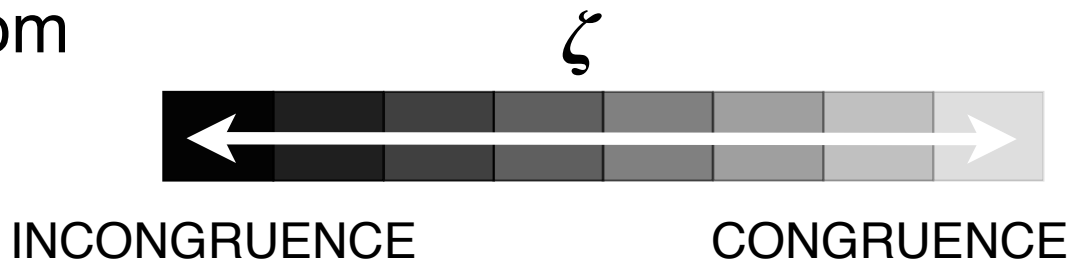
hard to solve analytically (and we want flexibility)

Approximate Bayesian Computation - ABC

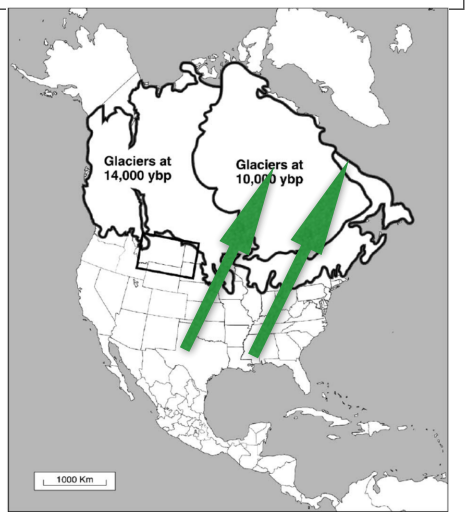
Likelihood of HyperParameter value is inversely proportional to the difference $Data_{simulated} - Data_{observed}$

Approximate Bayesian Computation - ABC

Simulate Genetic **Data** from random values from the hyperprior $p(\phi)\, p(\tau \mid \zeta)\, p(\zeta)$

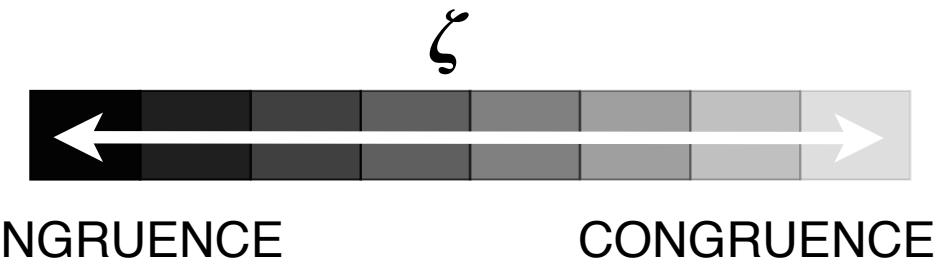


simulate coalescent & mutation across assemblage



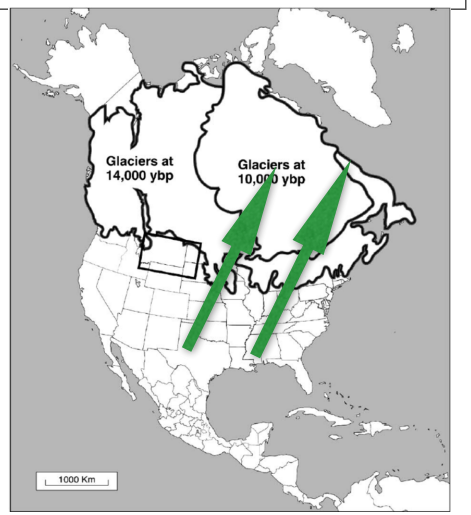
Approximate Bayesian Computation - ABC

Simulate Genetic **Data** from random values from the hyperprior $p(\phi)\, p(\tau \mid \zeta)\, p(\zeta)$



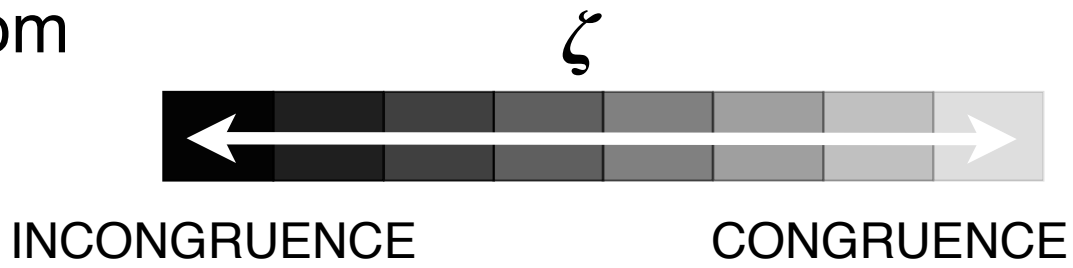
Compress **Data** (summary statistics)

simulate coalescent & mutation across assemblage



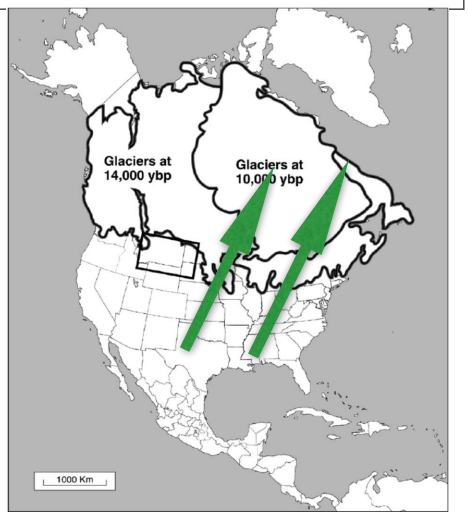
Approximate Bayesian Computation - ABC

Simulate Genetic **Data** from random values from the hyperprior $p(\phi)\, p(\tau \mid \zeta)\, p(\zeta)$



Compress **Data** (summary statistics)

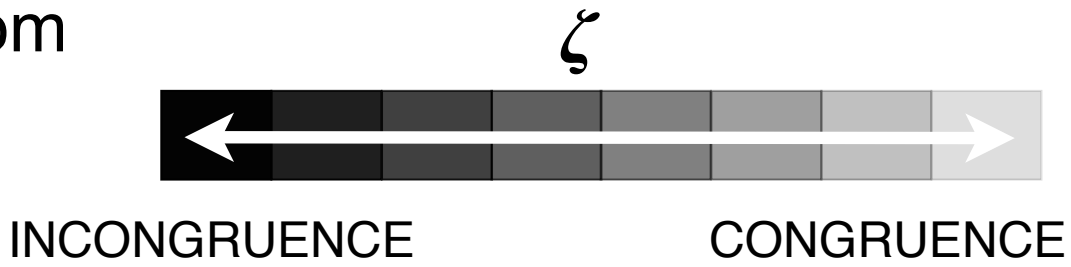
simulate coalescent & mutation across assemblage



Compare Compressed Simulated **Data** to Compressed Observed **Data**

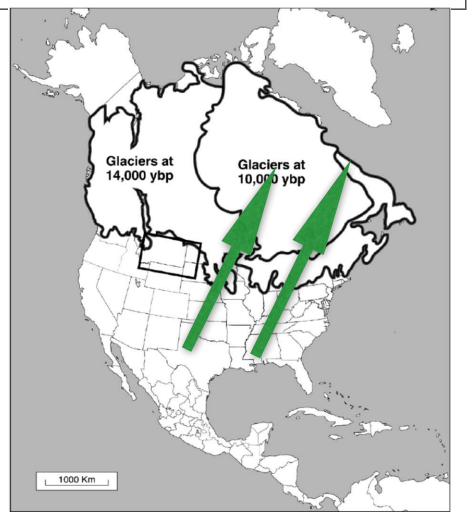
Approximate Bayesian Computation - ABC

Simulate Genetic **Data** from random values from the hyperprior $p(\phi)\, p(\tau \mid \zeta)\, p(\zeta)$



Compress **Data** (summary statistics)

simulate coalescent & mutation across assemblage



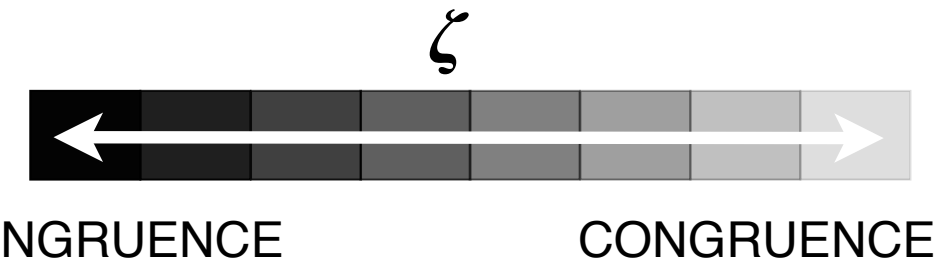
Compare Compressed Simulated **Data** to Compressed Observed **Data**

Keep if the Match is Close Enough

$$Data_{simulated} - Data_{observed} \approx 0\ ?$$

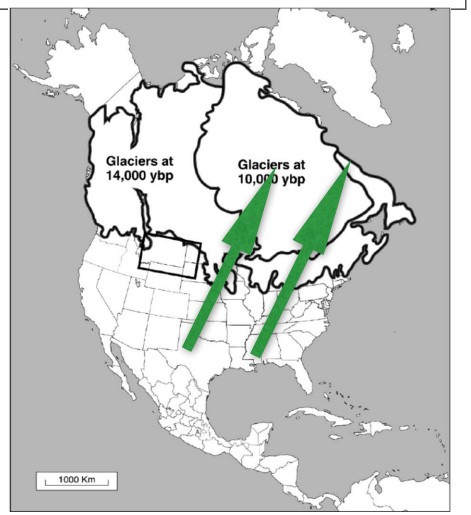
Approximate Bayesian Computation - ABC

Simulate Genetic **Data** from random values from the hyperprior $p(\phi)\, p(\tau \mid \zeta)\, p(\zeta)$



Compress **Data** (summary statistics)

simulate coalescent & mutation across assemblage



Compare Compressed Simulated **Data** to Compressed Observed **Data**

Keep if the Match is Close Enough

$$Data_{simulated} - Data_{observed} \approx 0\ ?$$

Approximate Bayesian Computation - ABC

Simulate Genetic **Data** from random values from the hyperprior $p($

ζ

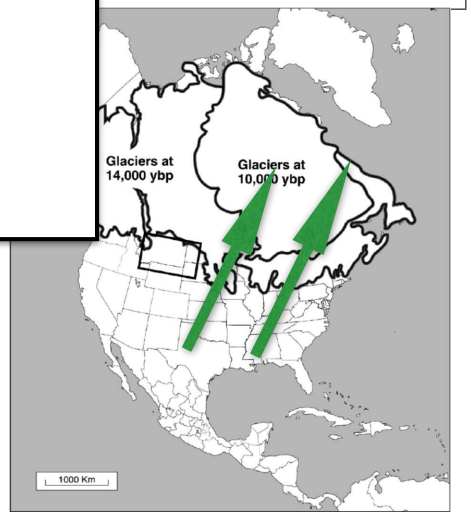
ONGRUENCE

$$P(\Phi \mid \boldsymbol{Data}) = \frac{p(\boldsymbol{Data}|\Phi)\, p(\Phi)}{p(\boldsymbol{Data})}$$

Compress **D**

balescent
across
ge

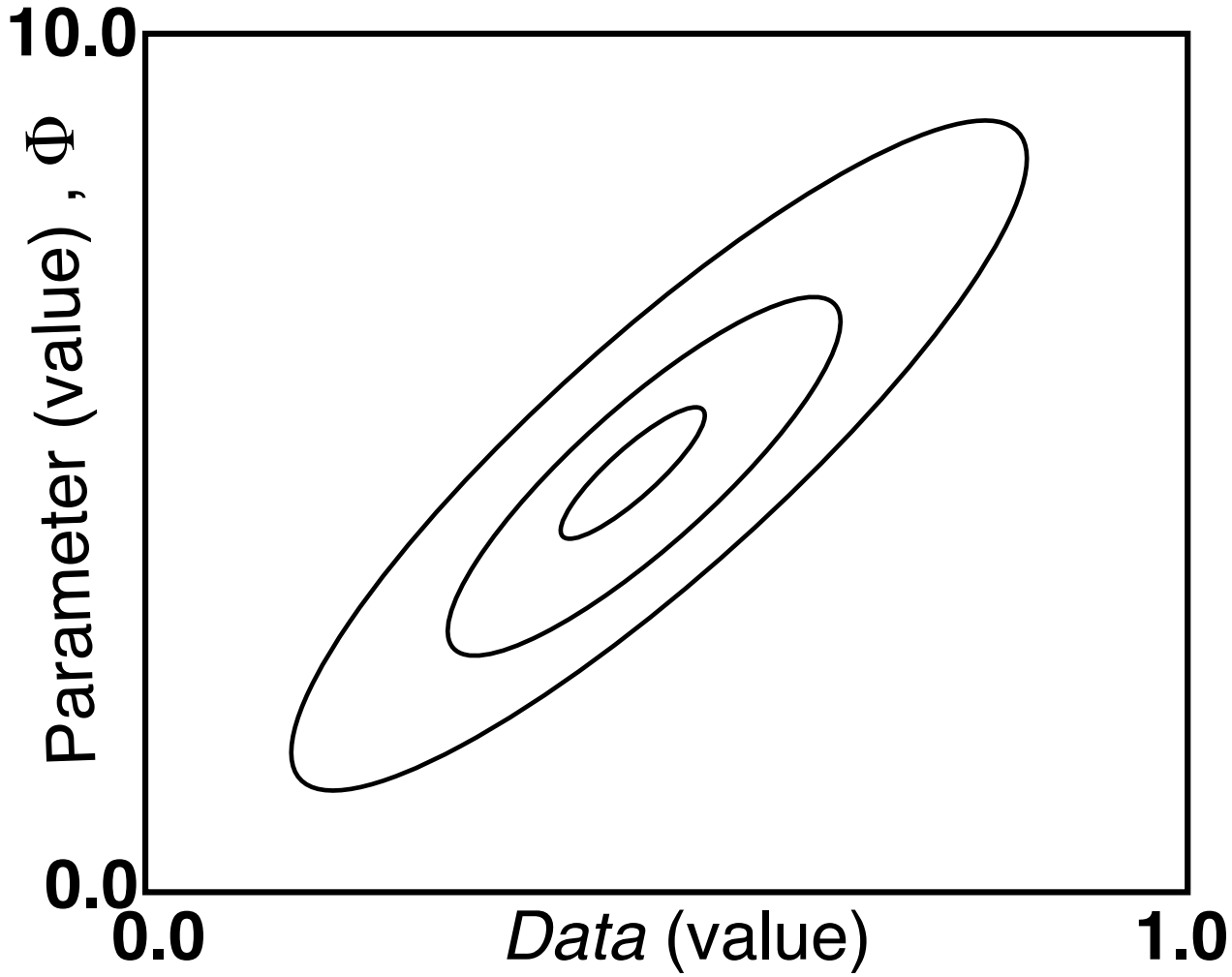
Compare Compr
to Compressed C



Keep if the Match is Close Enough

$$Data_{simulated} - Data_{observed} \approx 0\ ?$$

$$P(\Phi \mid \textbf{\textit{Data}}) = \frac{p(\textbf{\textit{Data}}|\Phi)\, p(\Phi)}{p(\textbf{\textit{Data}})}$$

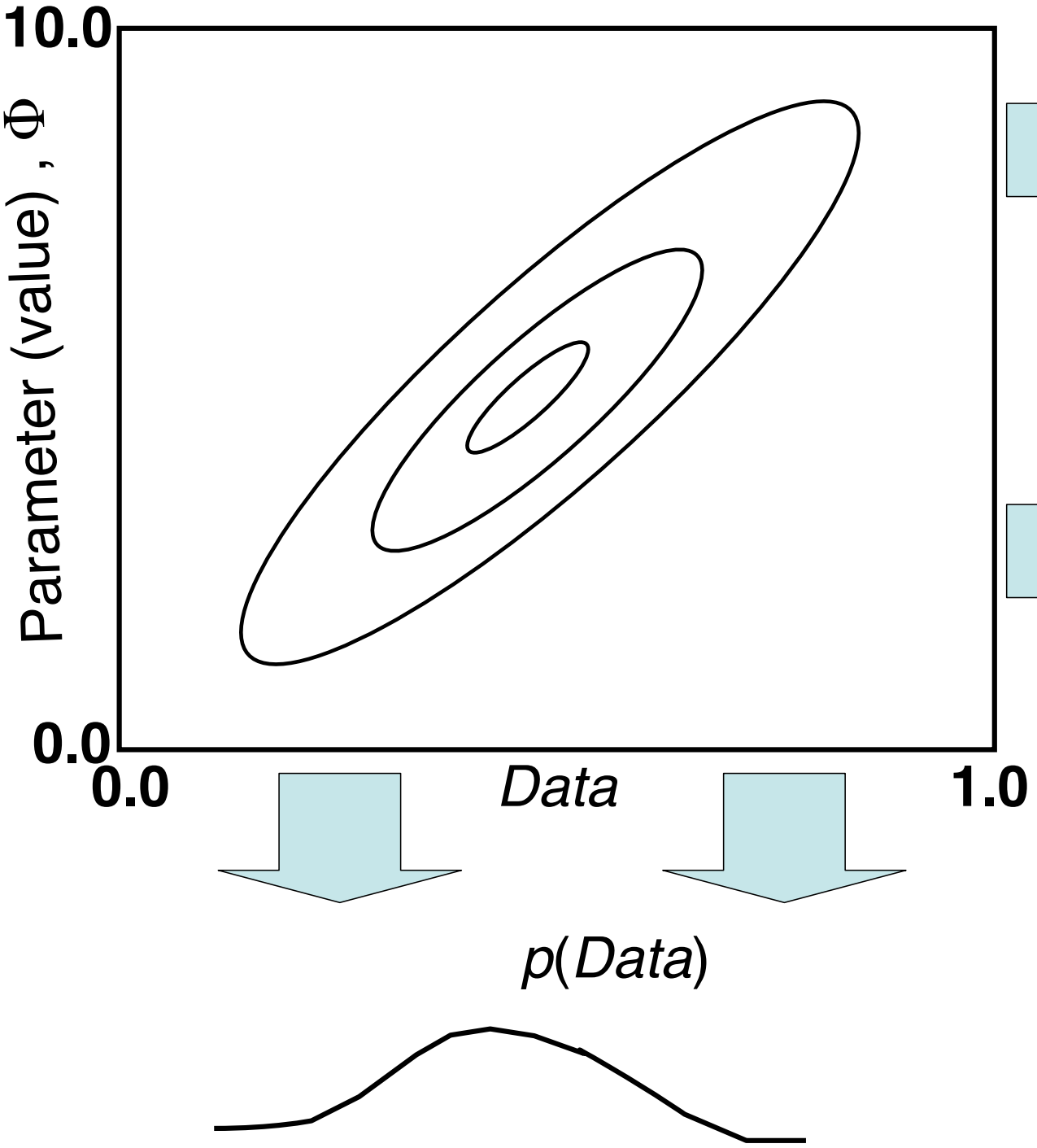


Mark Beaumont's ABC schematic

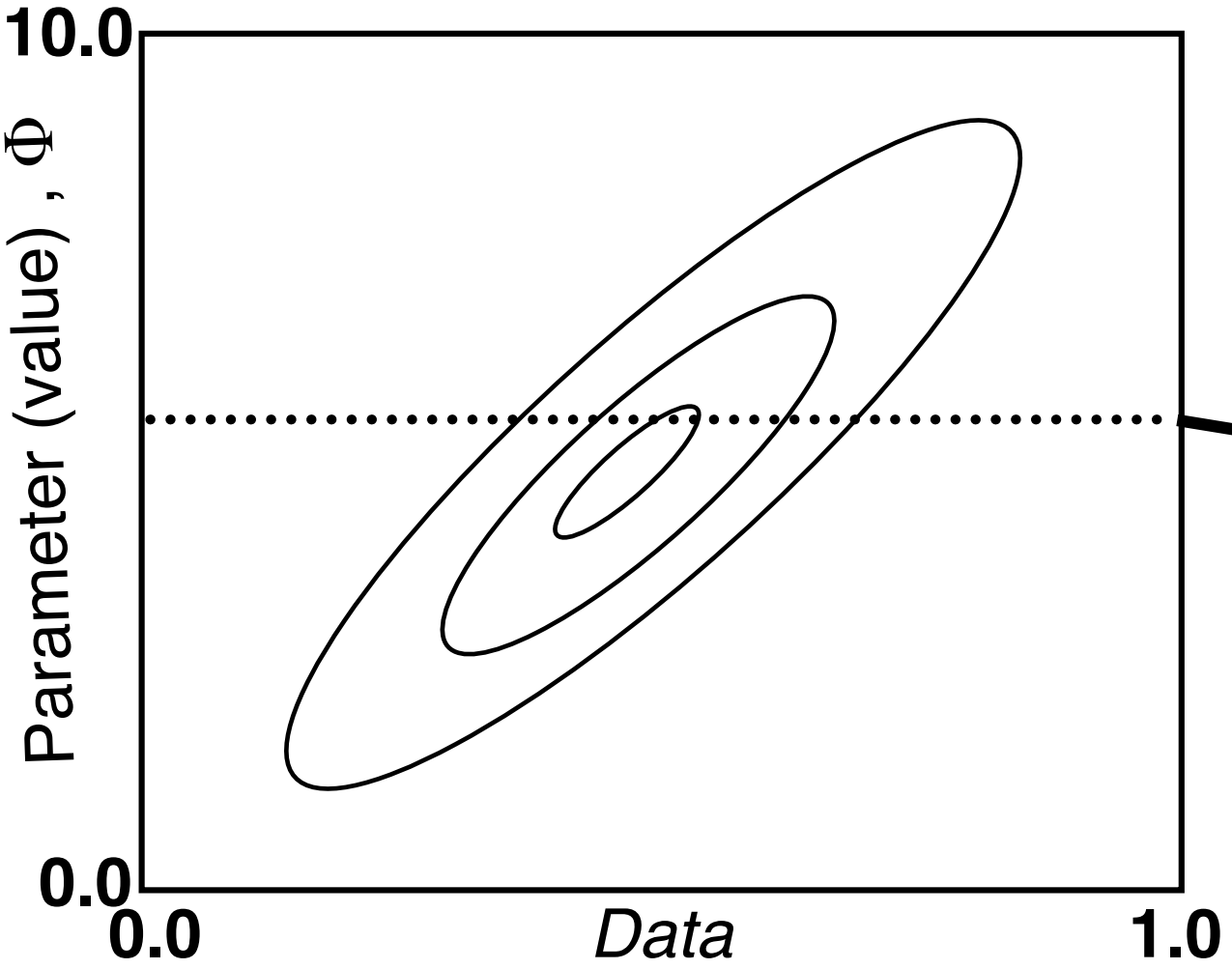
if $Data_{simulated} - Data_{observed} \approx 0$?

then $\Phi_{simulated} \approx \Phi_{observed}$

$$P(\Phi \mid \textbf{\textit{Data}}) = \frac{p(\textbf{\textit{Data}}|\Phi)\, p(\Phi)}{p(\textbf{\textit{Data}})}$$

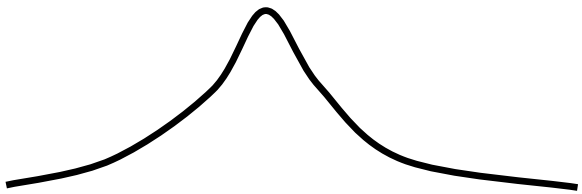


$$P(\Phi \mid \textbf{\textit{Data}}) = \frac{p(\textbf{\textit{Data}}|\Phi)\, p(\Phi)}{p(\textbf{\textit{Data}})}$$

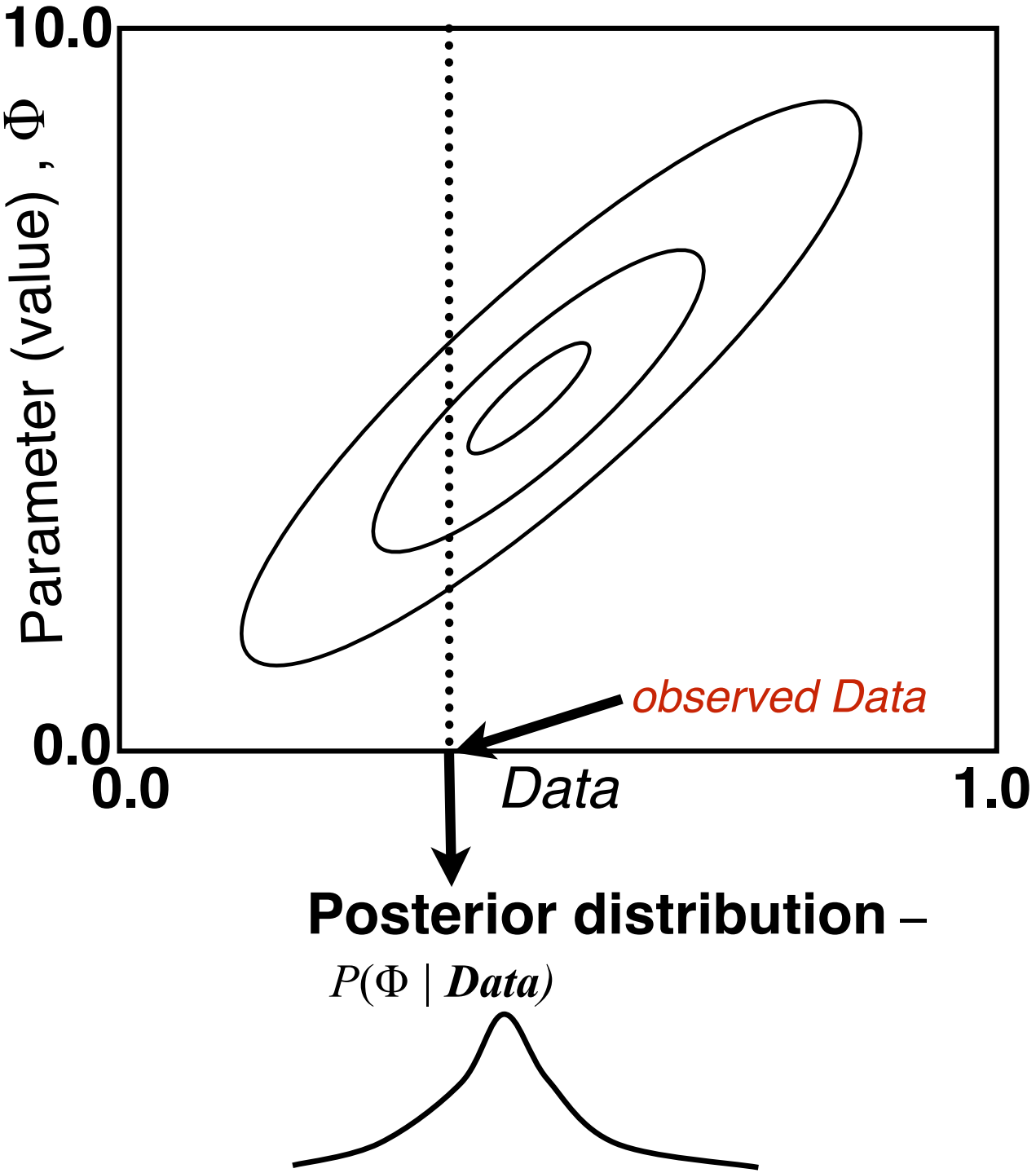


Likelihood–

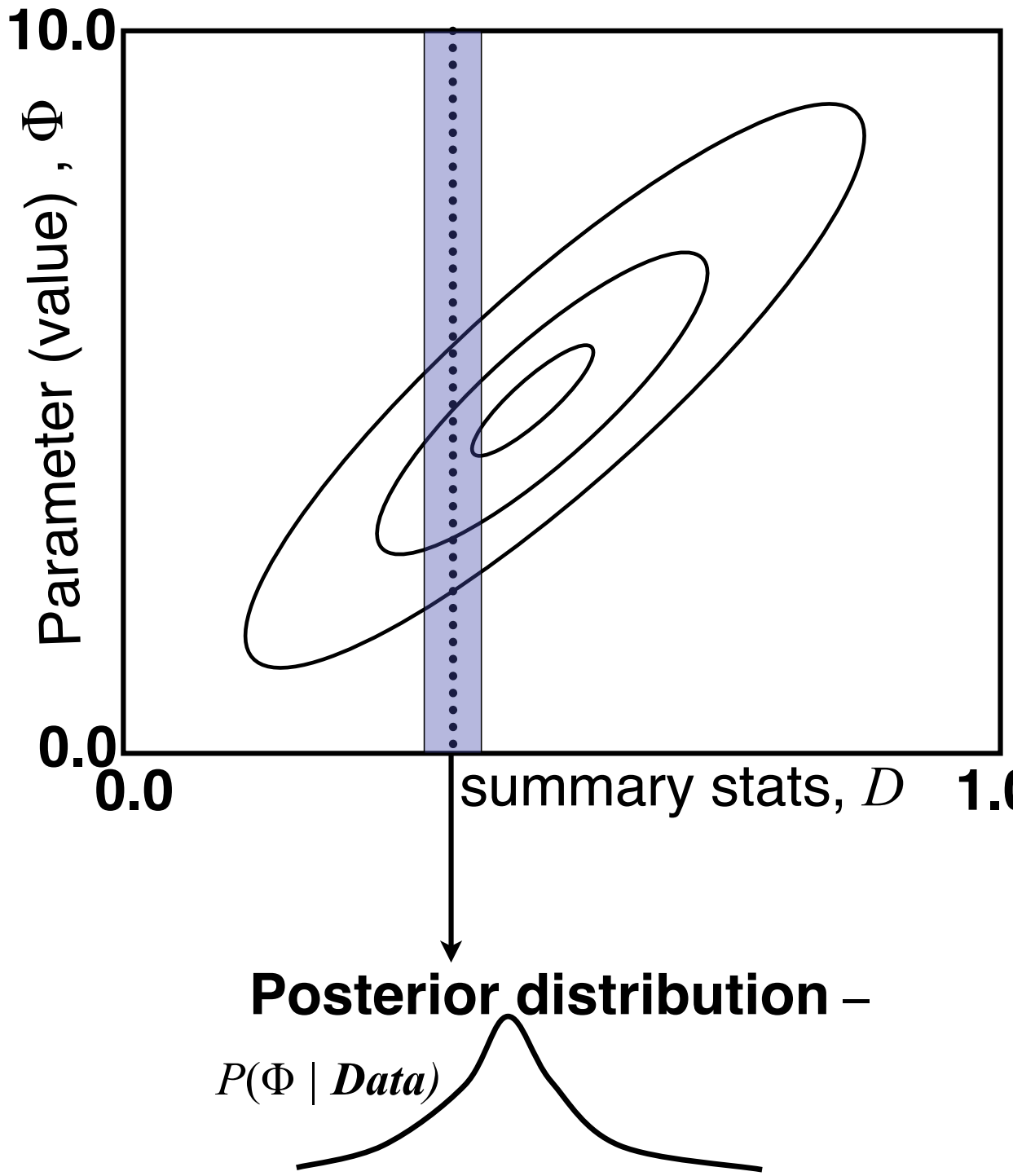
$P(\textbf{\textit{Data}} \mid \Phi)$



$$P(\Phi \mid \textbf{\textit{Data}}) = \frac{p(\textbf{\textit{Data}}|\Phi)\, p(\Phi)}{p(\textbf{\textit{Data}})}$$



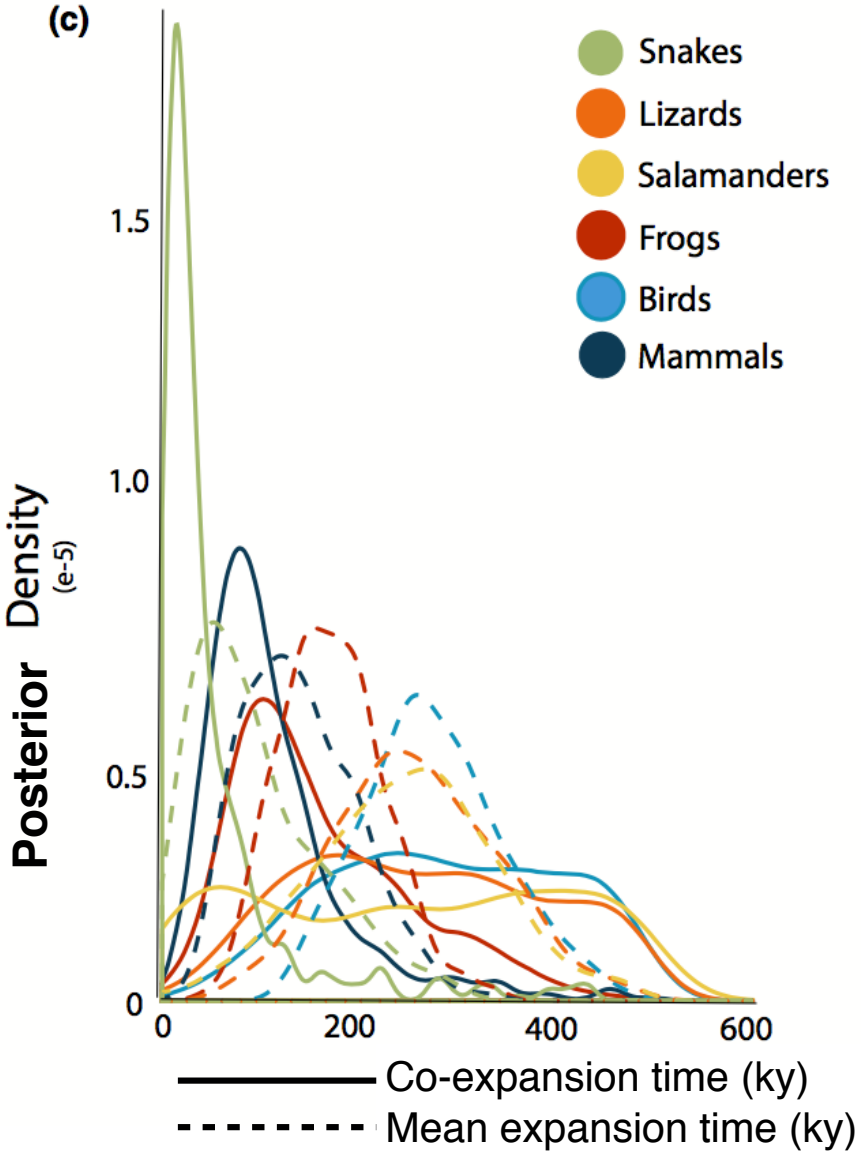
$$P(\Phi \mid \textbf{\textit{Data}}) = \frac{p(\textbf{\textit{Data}}|\Phi)\, p(\Phi)}{p(\textbf{\textit{Data}})}$$

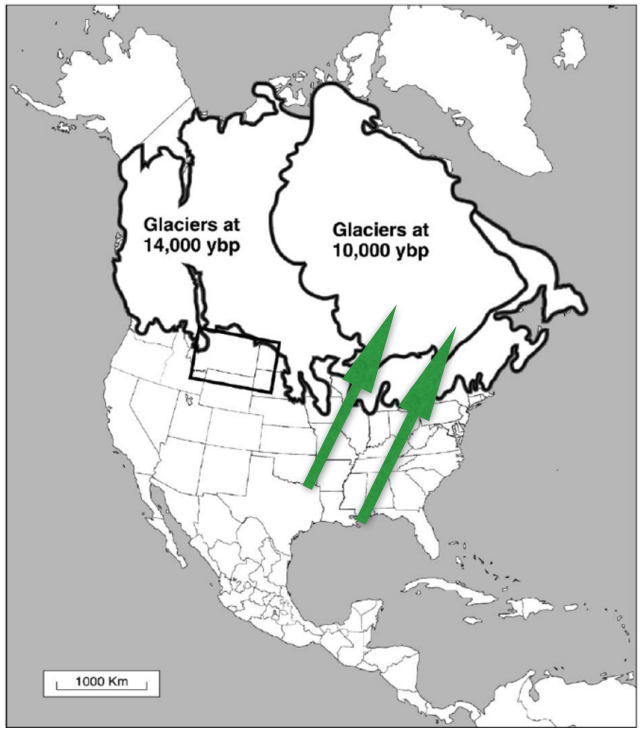


By converting *Data* to summary stats D, hitting the data exactly is not required. Instead we we need to only hit $\| D_i - D_{obs} \| < \varepsilon$ within a "threshold"

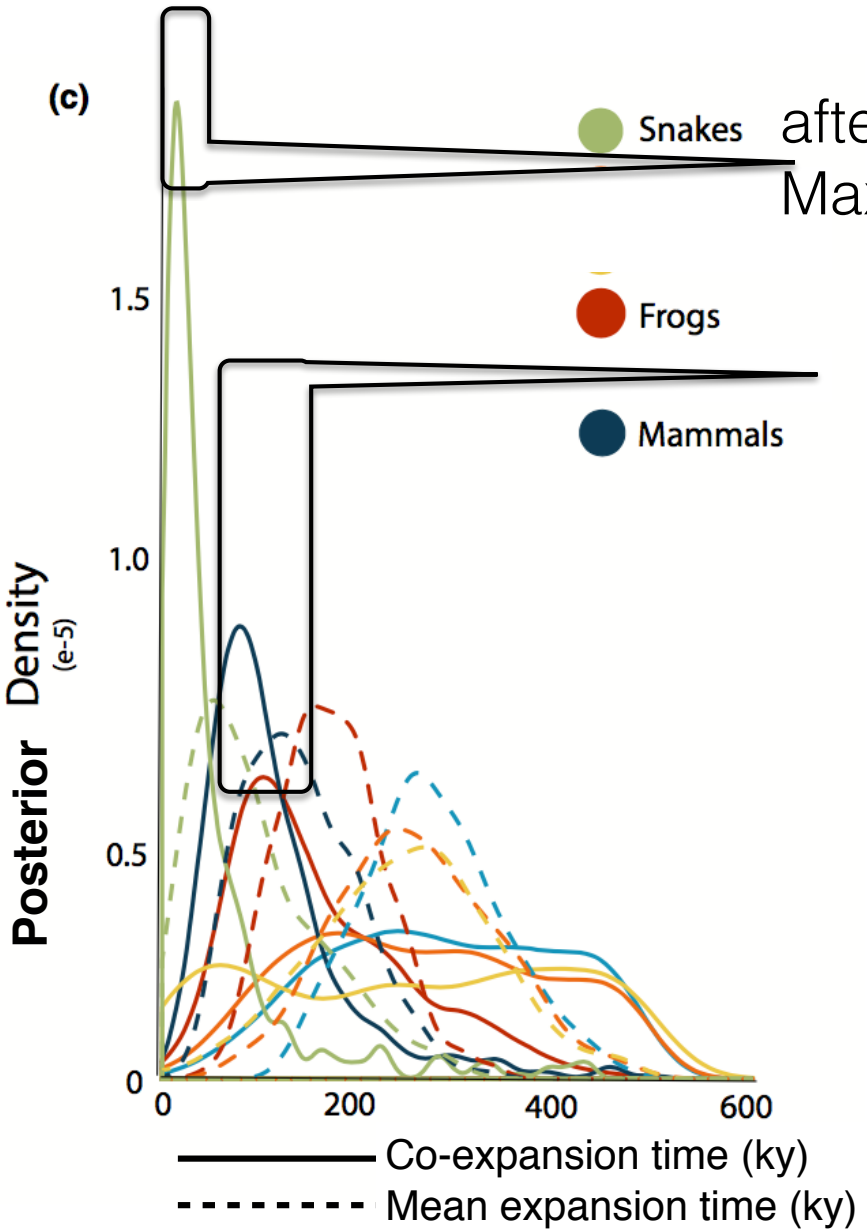
Results (hABC)

overall, extremely asynchronous, but ….





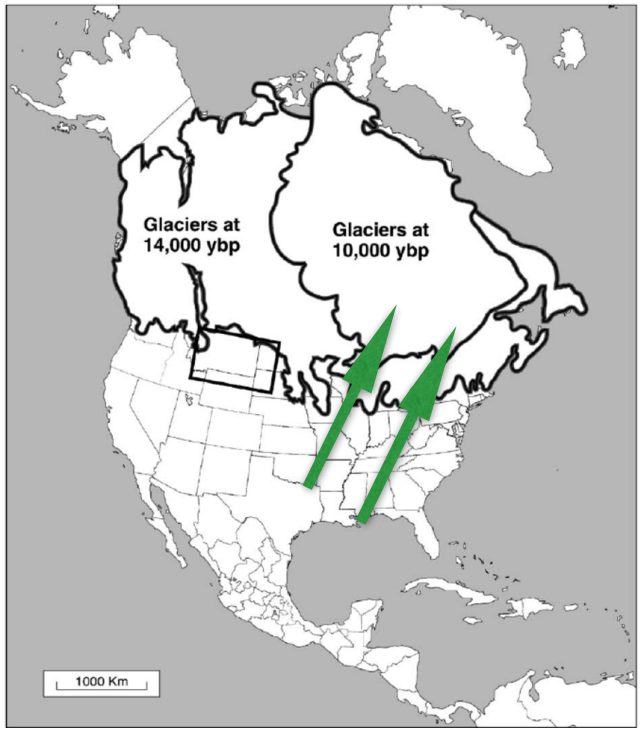
Results (hABC)



after last Glacial Maximum?

after penultimate Glacial Maximum?

Strong to moderate recent co-expansion in **Snakes**, **Frogs** and **Mammals**



Development of ABC inferential strategy

1. Summary statistic Selection

2. Prior Selection & Model Checking

3. Estimator Validation with PODS (pseudo-observed data sets)

4. Estimation

Development of ABC inferential strategy

1. Summary statistic Selection

2. Prior Selection & Model Checking

3. Validation

4. Estimation

2. Prior Selection

How do we choose the prior?

1. look at the data

2. graphical checks from trial runs

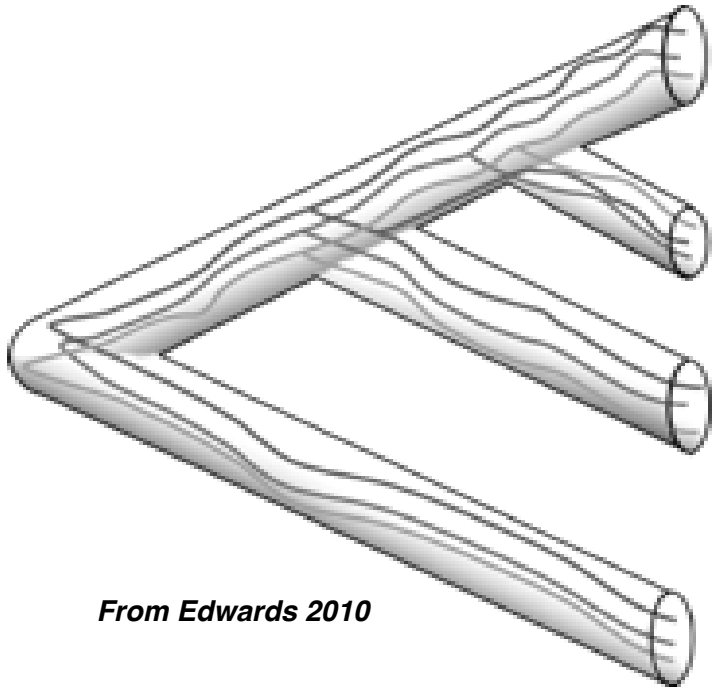
3. Bayesian model averaging

(Hickerson et al. *in review* - response to Oaks et al. 2013)

2. Prior Selection

How do we choose the prior?

1. look at the data



From Edwards 2010

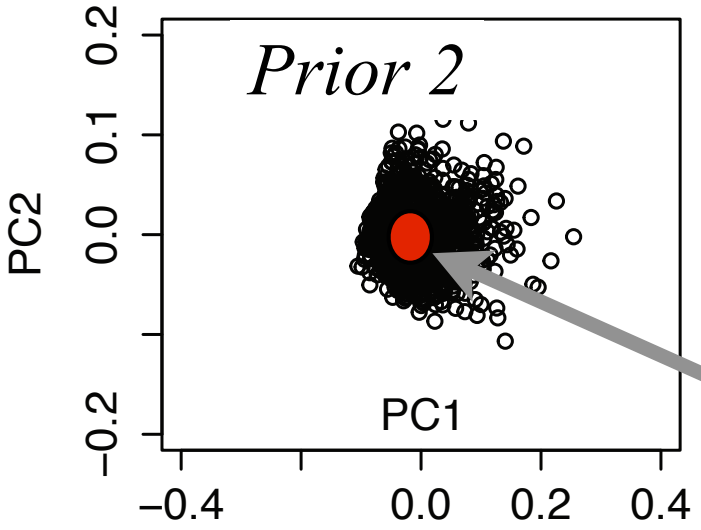
use gene tree divergences as the maximum

(population divergence < gene divergence)

2. Prior Selection

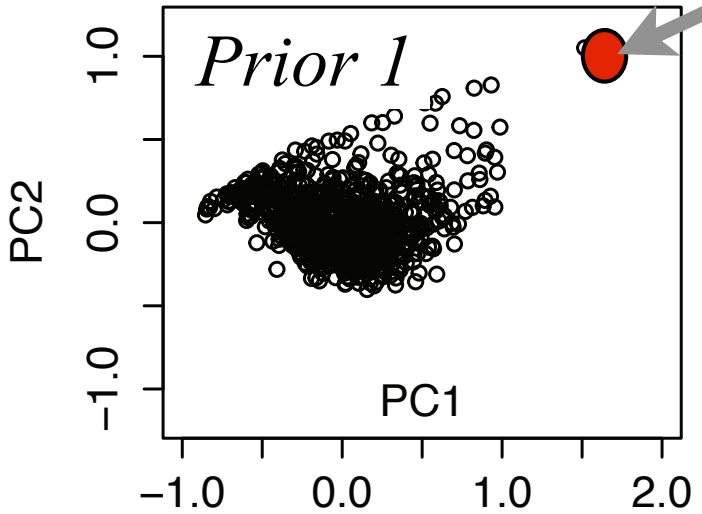
How do we choose the prior?

2. graphical checks from trial runs (small prior samples)



well sampled prior

observed data



poorly sampled prior

(Hickerson et al. 2014)

2. Prior Selection

How do we choose the prior?

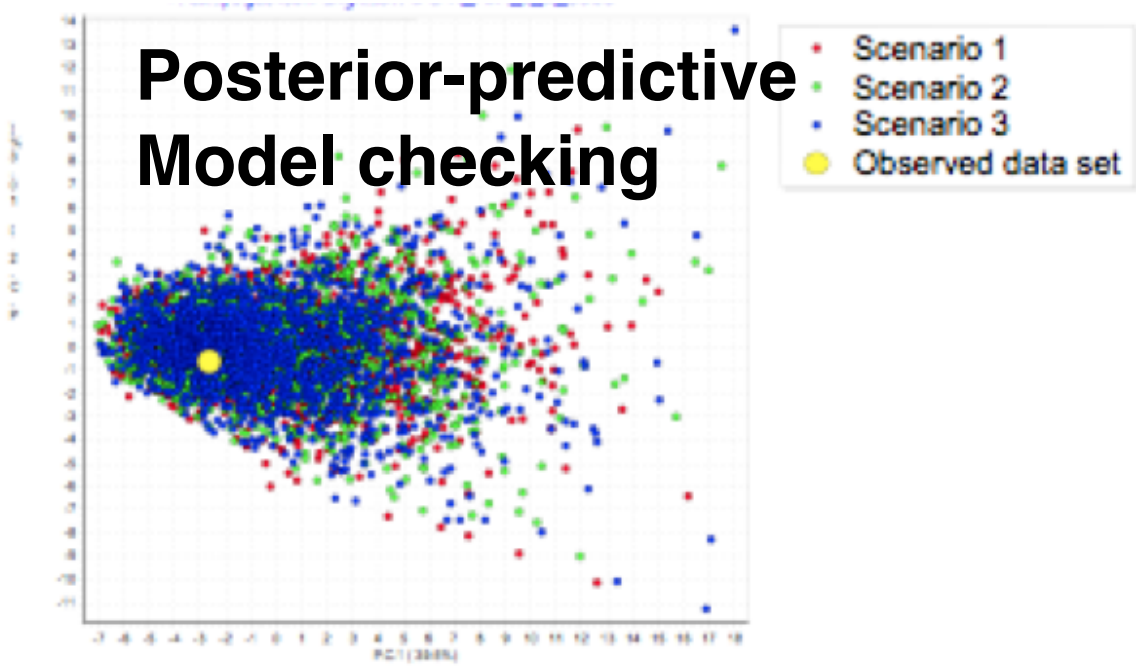
3. Bayesian model averaging

Sample from **multiple** candidate priors

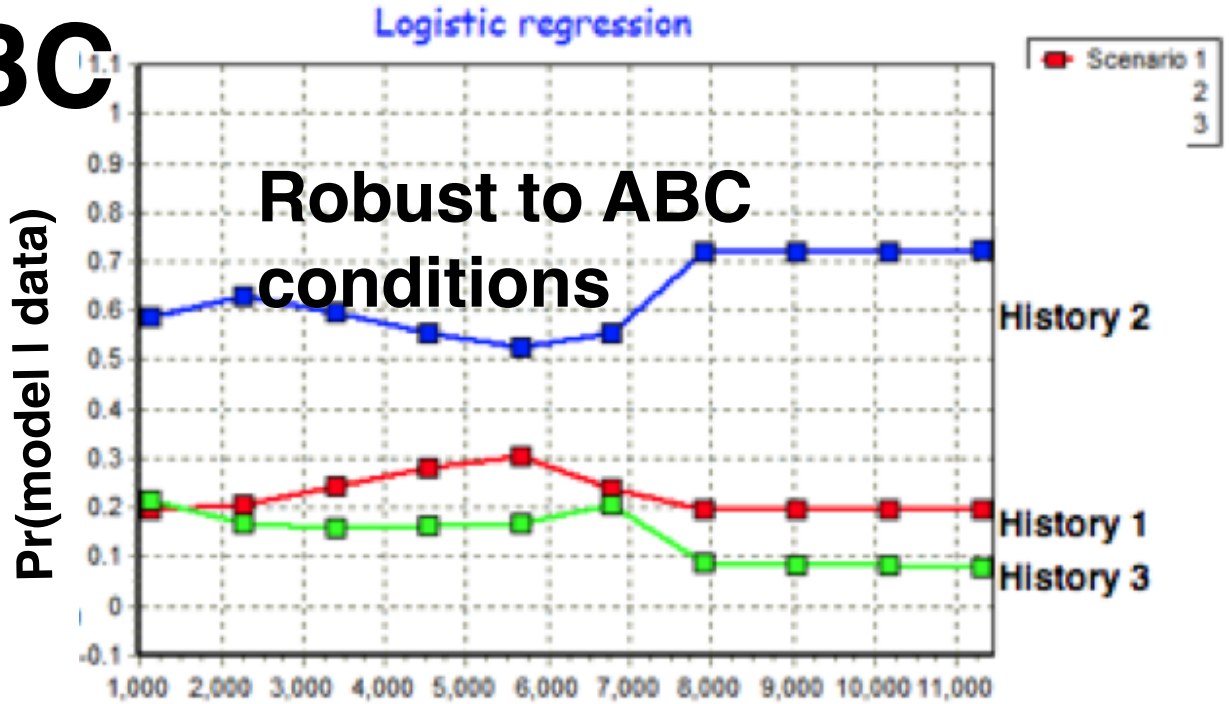
$$\{M_1, \ldots, M_8\}$$

Do regular ABC estimation

ABC posterior estimate is weighted by posterior strength of each prior



DIY-ABC



Proportion of accepted simulations

Development of ABC inferential strategy

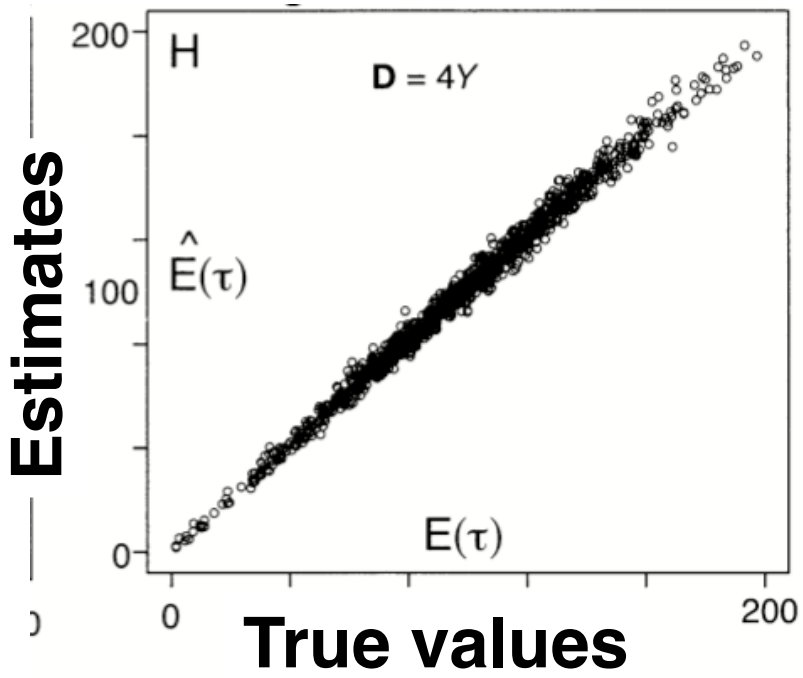
1. Summary statistic Selection
2. Prior Selection and Model Checking
3. **Estimator Validation with PODS (pseudo-observed data sets)**
4. Estimation

3. Estimator Validation with PODS (pseudo-observed data sets)

1 simulate “reference table” (the prior sample)

2 simulate PODS (pseudo-observed data sets with recorded parameter values)

3 make ABC estimates from the PODS



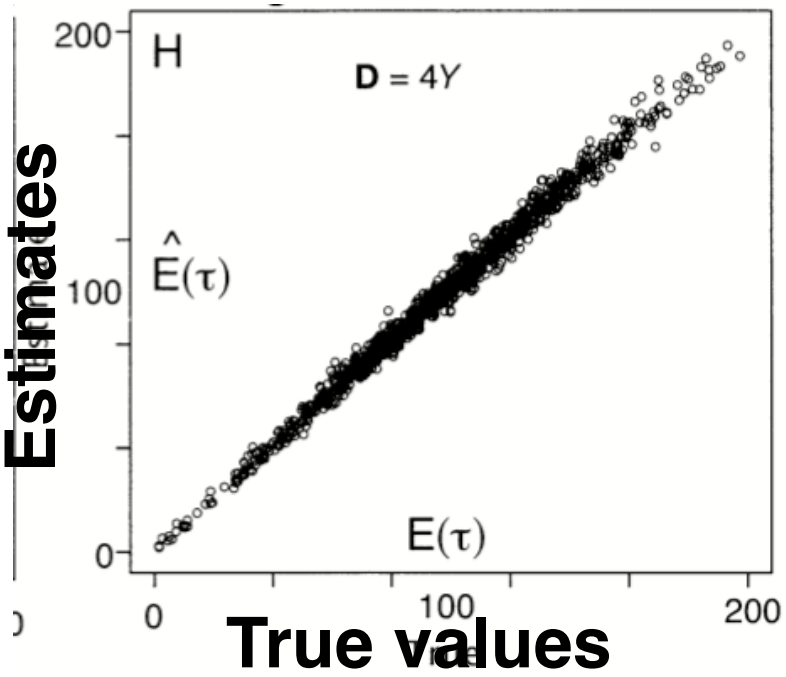
requires CPU

CUNY high performance computing facility

> 300,000 Credit card numbers later

Beware wwww.citibank.com

Simulation validation of ABC

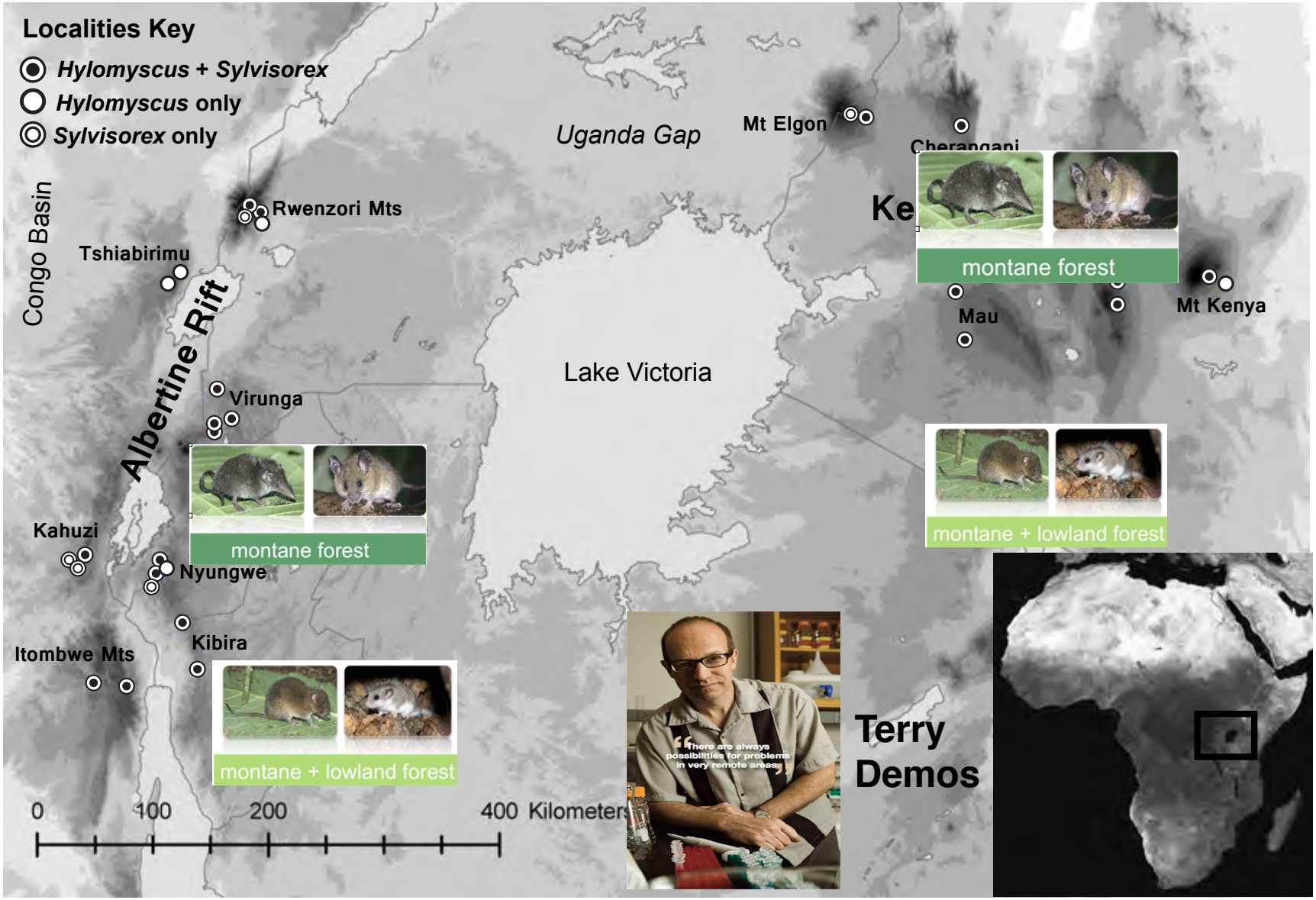


Software - MTML - msBayes

6. PODS validation and/or power analysis

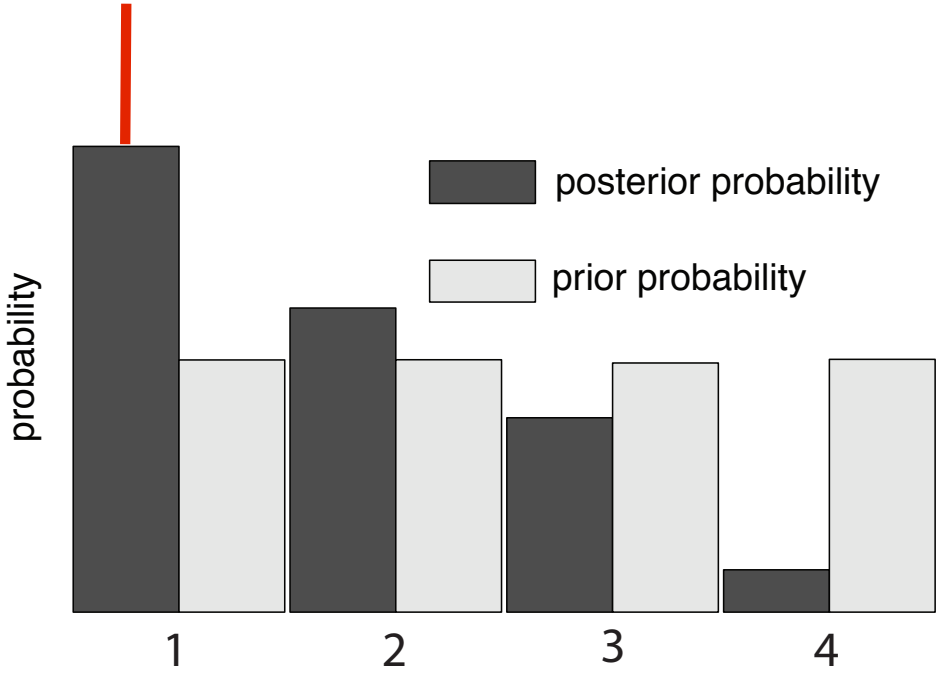
Software - MTML - msBayes

example - 4 East/West species pairs of African rodents

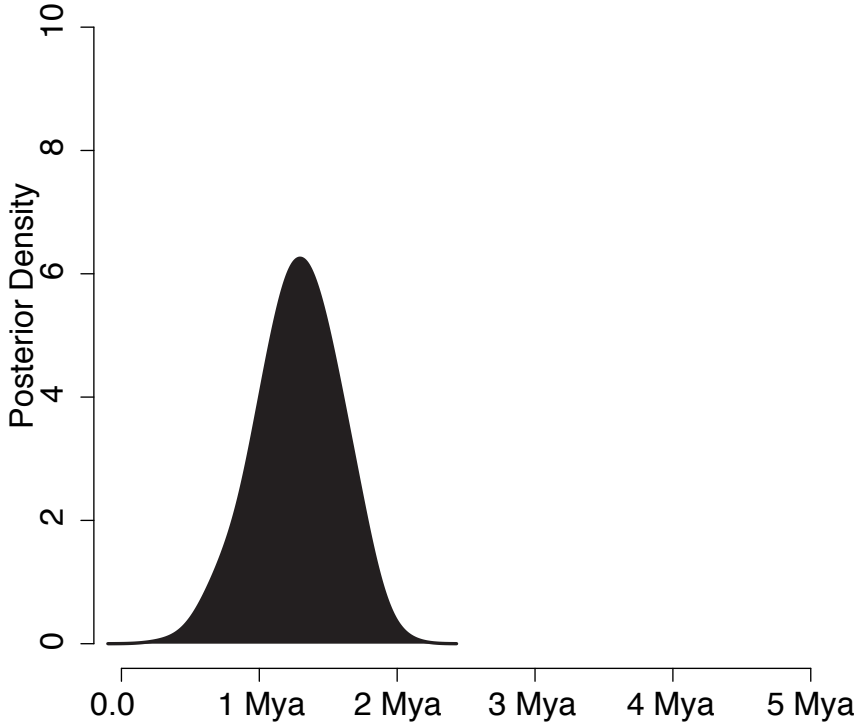


Software - MTML - msBayes

divergence times conditional on Ψ with
highest posterior probability (Ψ=1 in this case)



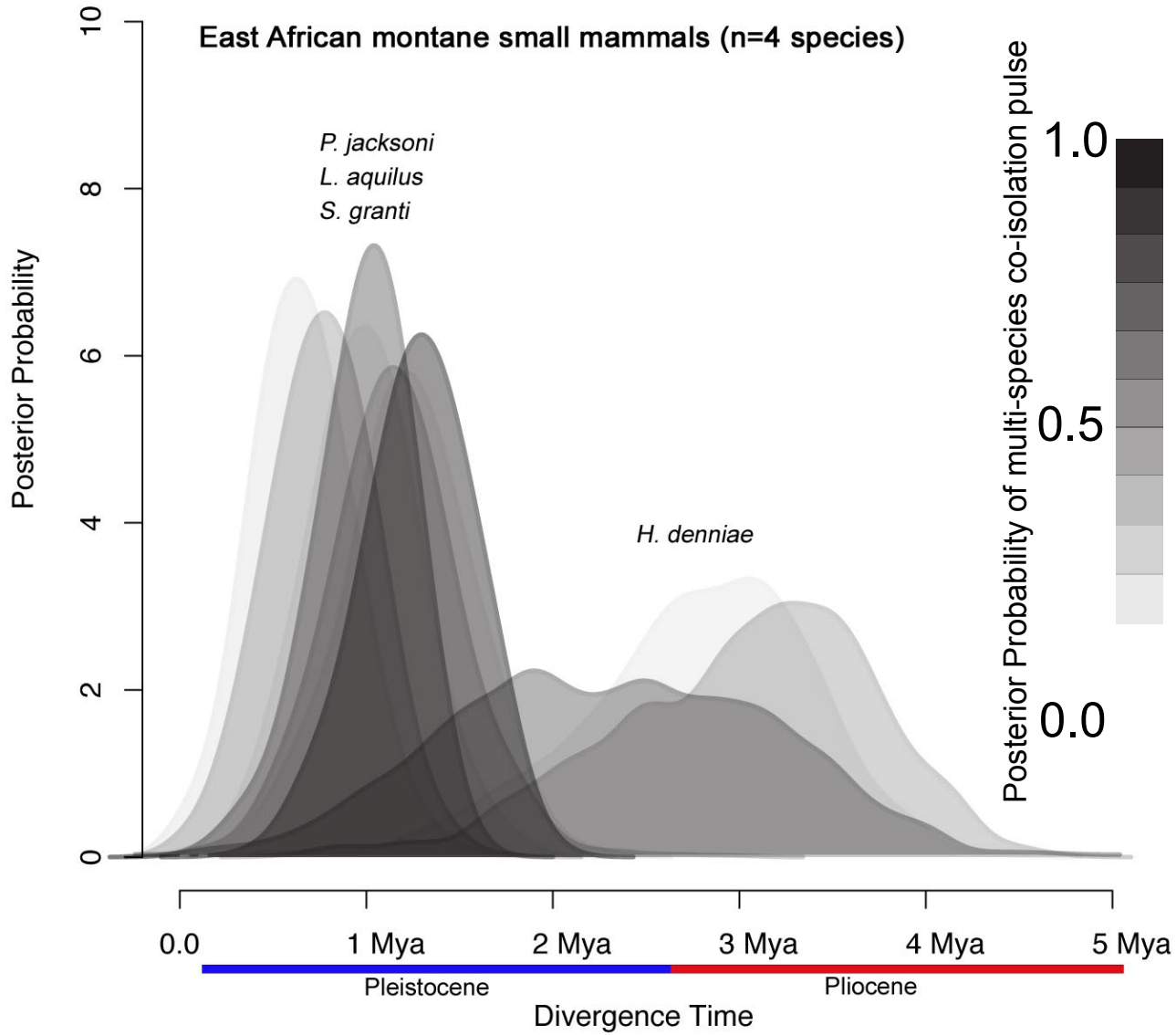
Ψ (number of divergence times)



"simultaneous" divergence time
for 4 pairs

Software - MTML - msBayes

estimate divergence times averaged across posterior for number of times (Ψ)



Thank you!