UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

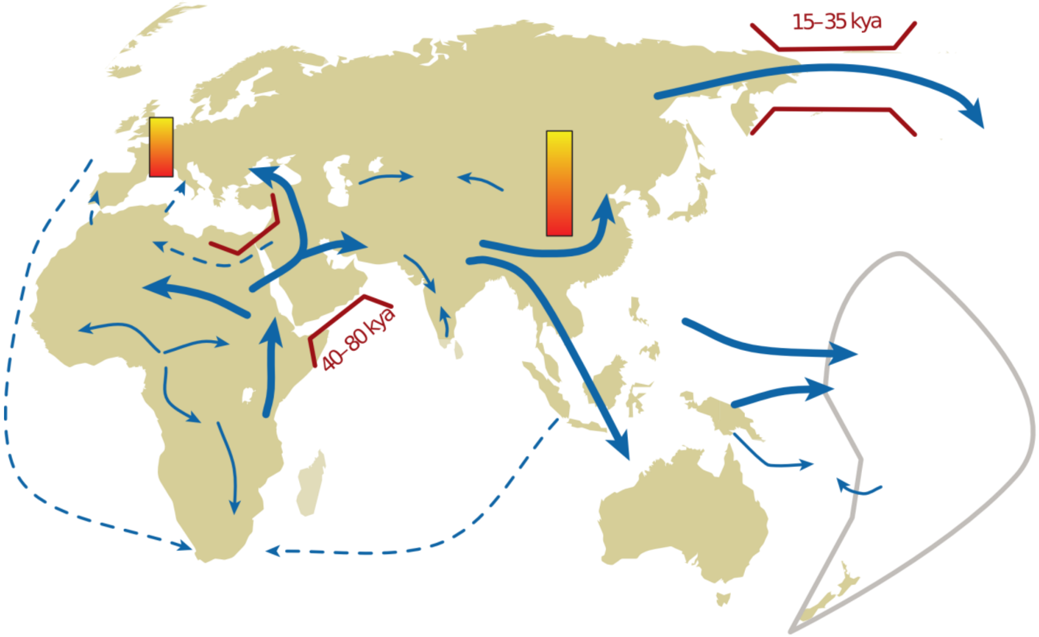
Model based inference of evolutionary histories

Daniel Wegmann
University of Fribourg

Introduction

- The current genetic diversity is the outcome of past evolutionary processes.
- Hence, we can use genetic diversity to tell stories about the past.

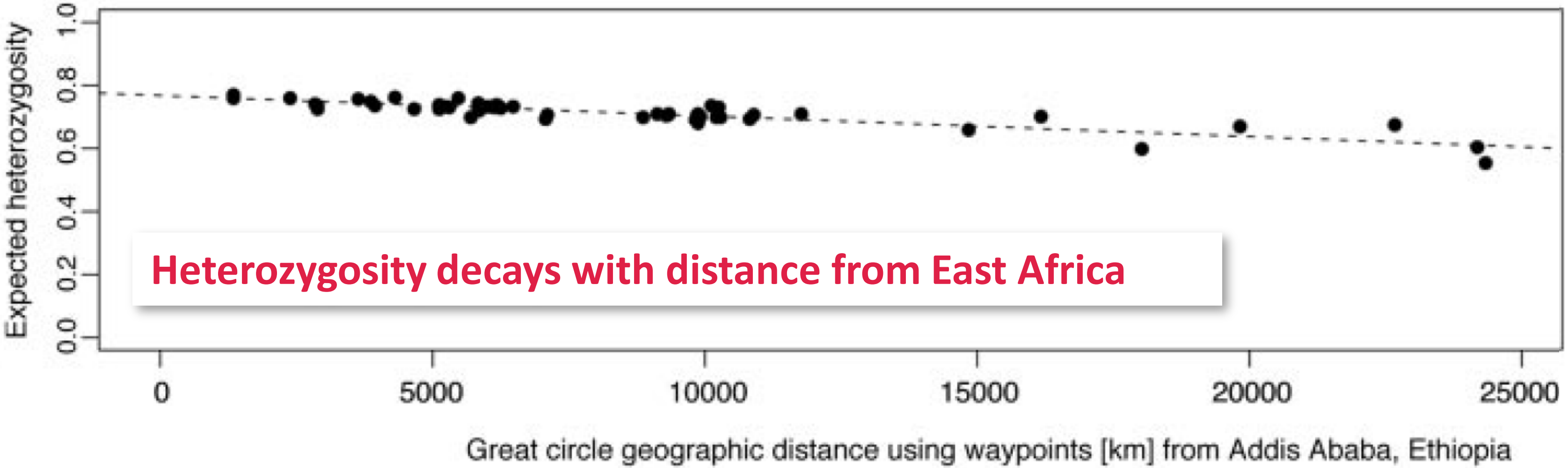
- But this is a **challenging task!**
  - The history of natural populations is usually **complex**.
  - Several evolutionary processes can leave **similar footprints** (bottleneck vs. selection).



Novembre & Ramachandran (2011)

Qualitative inference

- Traditionally, we have relied on qualitative inference
- **Example**: out of Africa expansion via sequential founder effects in humans.



Heterozygosity decays with distance from East Africa

Ramachandran *et al.* (2005)

Model-based inference

- Patterns of genetic diversity may serve as evidence for or against stories of the evolutionary past.
- Such stories are usually vague („Serial founder effects“).
- While the evidence may be strong, the argument remains verbal and is potentially subjective to interpretation.

➡ Model-based inference provides statistical support

Model-based inference

- Patterns of genetic diversity may serve as evidence for or against stories of the evolutionary past.
- Such stories are usually vague („Serial founder effects“).
- While the evidence may be strong, the argument remains verbal and is potentially subjective to interpretation.

➡ Model-based inference provides statistical support

Essentially, all models are wrong, but some are useful.

George E. Box

- Qualitative inference is key when constructing sensible models!

Examples of Model Based Inference

1

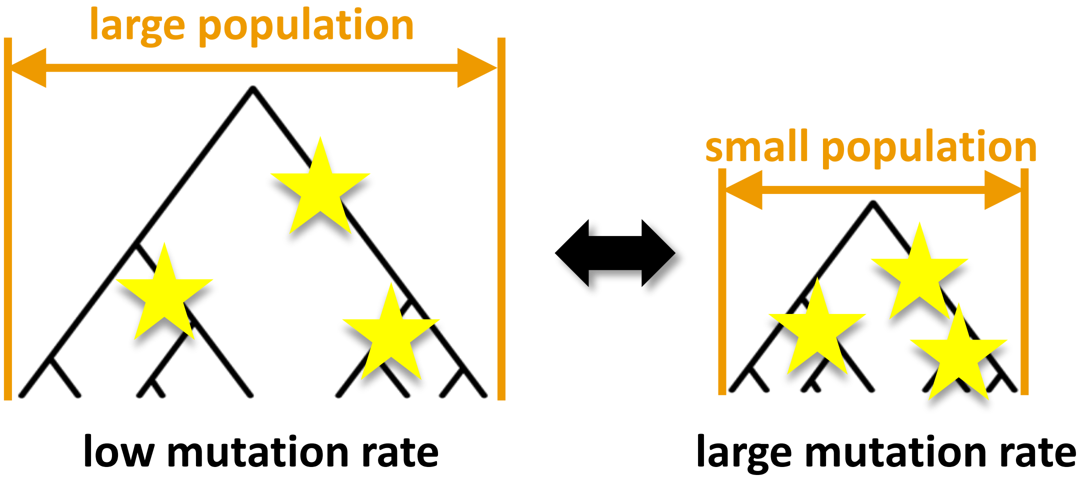
Human mutation rates
using maximum likelihood of summary statistics

2

Demographic histories
using Approximate Bayesian Computation

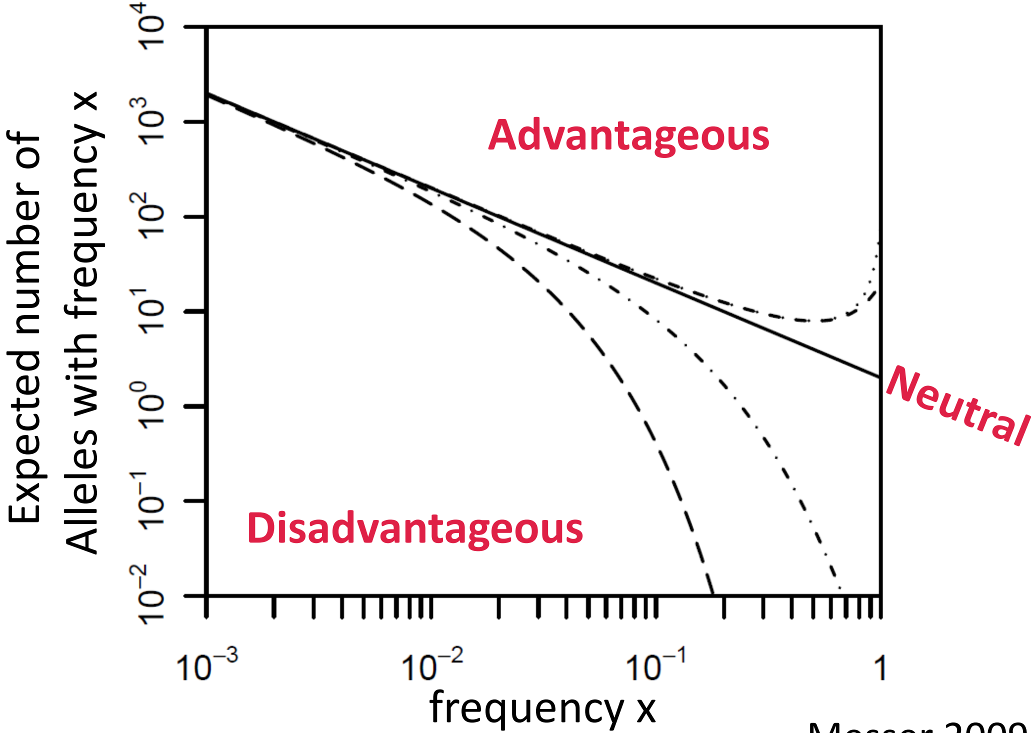
Joint inference of demography and mutation rates

- Estimating **mutation rates** μ from population data is difficult as the number of polymorphisms is **confounded by demography** ...



Joint inference of demography and mutation rates

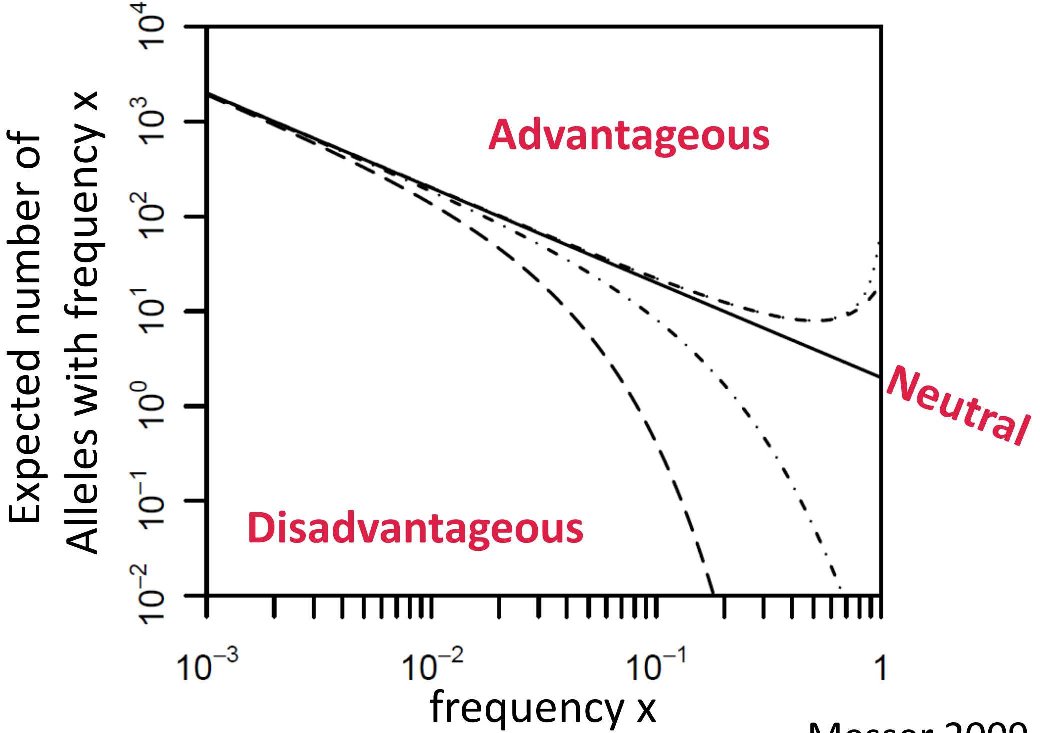
- Estimating **mutation rates** μ from population data is difficult as the number of polymorphisms is **confounded by demography** ...
- ... **and selection.**



Messer 2009

Joint inference of demography and mutation rates

- Estimating **mutation rates** μ from population data is difficult as the number of polymorphisms is **confounded by demography** ...
- ... **and selection**.
- Very rare variants are virtually **unaffected by selection**.



Messer 2009

Joint inference of demography and mutation rates

- Estimating **mutation rates** μ from population data is difficult as the number of polymorphisms is **confounded by demography** ...
- ... **and selection**.

- Very rare variants are virtually **unaffected by selection**.

- If **sample size > population size**, multiple coalescent events occur at a rate largely independent of **N**, making an estimation of μ and **N** possible.

Messer 2009

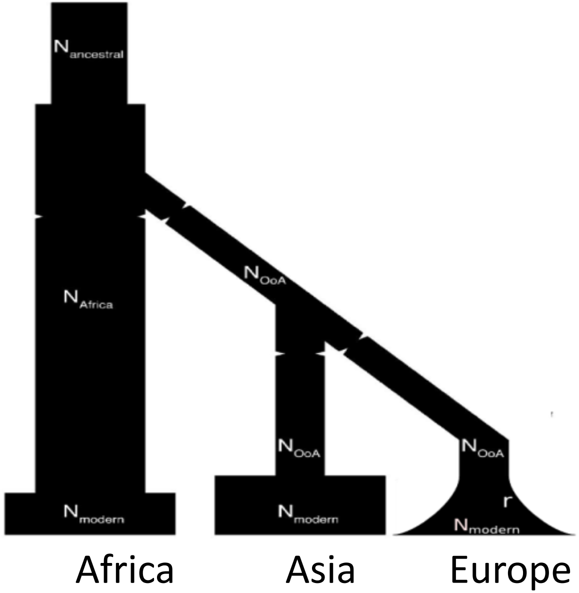
Joint inference of demography and mutation rates

Data Set:

- 202 known or prospective drug target genes sequenced in 12,514 Europeans.
- Median coverage of 27x and a call rate of 90.7%
- Heterozygous and singleton concordance > 99% in 130 sample duplicates.

Model:

- Exponential growth in Europe.
- All other parameters fixed to Schaffner estimates.



Joint inference of demography and mutation rates

- Likelihood: probability of data **D** given parameters **μ,N**

$$P(\mathbf{D} \mid \mu,\mathbf{N})$$

Polymorphisms

Mutation rates & Population sizes

- Maximum-Likelihood: Find **μ,N** that maximize $P(\mathbf{D}|\mu,\mathbf{N})$
- For many evolutionary models, analytical solutions of the likelihood are **very hard** and often **impossible** to obtain
- We will use two tricks:

1) Use **summary statistics** **S** instead of the full data **D**
    - The hope is that $P(\mathbf{D}|\mu,\mathbf{N})$ is proportional to $P(\mathbf{S}|\mu,\mathbf{N})$,
2) Use **simulations** to approximate the likelihood function $P(\mathbf{S}|\mu,\mathbf{N})$

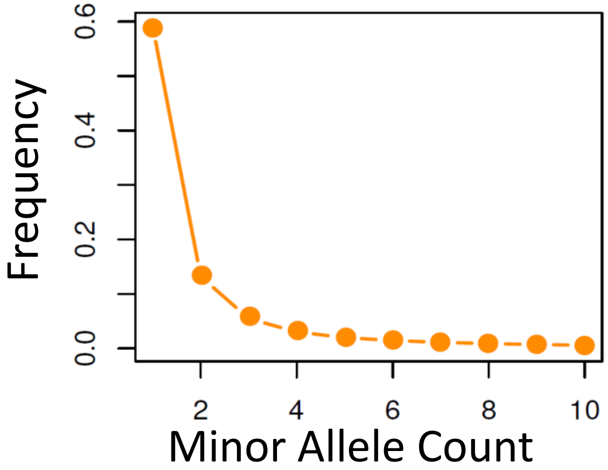
Joint inference of demography and mutation rates

1) Using **Site Frequency Spectrum SFS** instead of the full data **D**

```
AGATTCAC
AGCTTCAT
AGATTCAT
AGATTCAT
AGCTTCGC
```

⋮

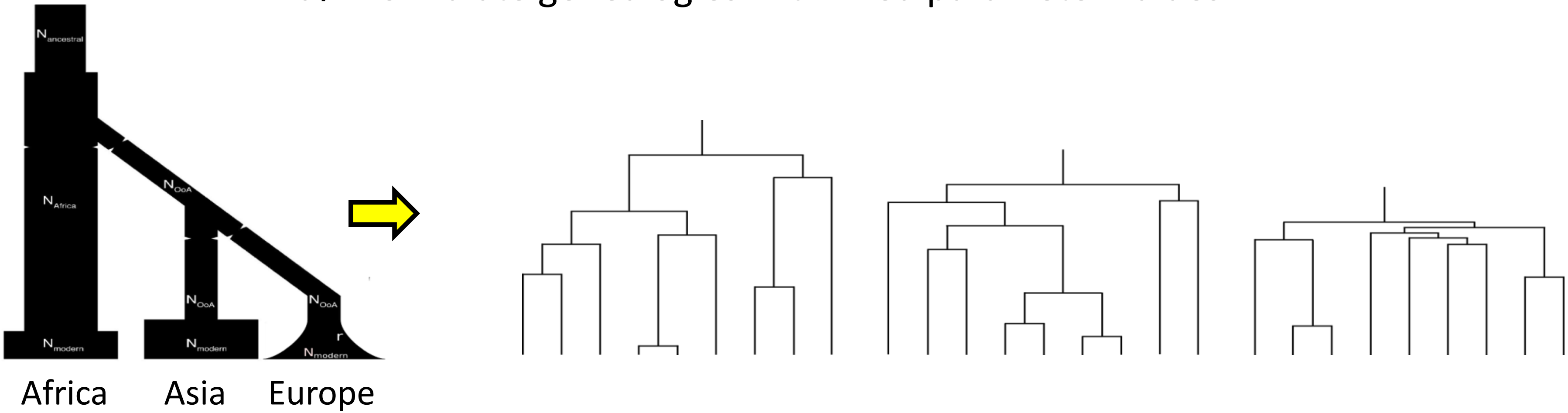
22,000 Sequences of 202 genes



Site Frequency Spectrum SFS

Joint inference of demography and mutation rates

1) Using **Site Frequency Spectrum** **SFS** instead of the full data **D**
2) Using Monte Carlo simulations to approximate P(**SFS**|**μ,N**):
    - a) Simulate genealogies with fixed parameter values



Nielsen 2000; Coventry et al. 2010

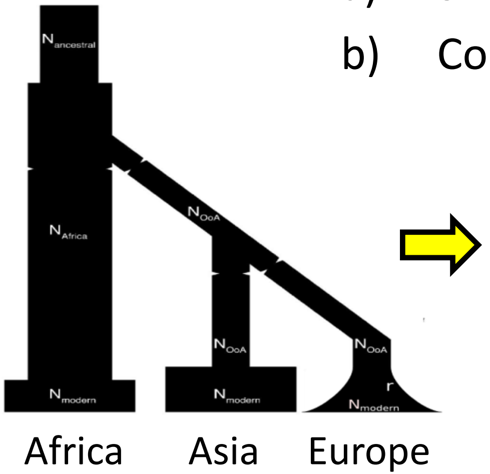
Joint inference of demography and mutation rates

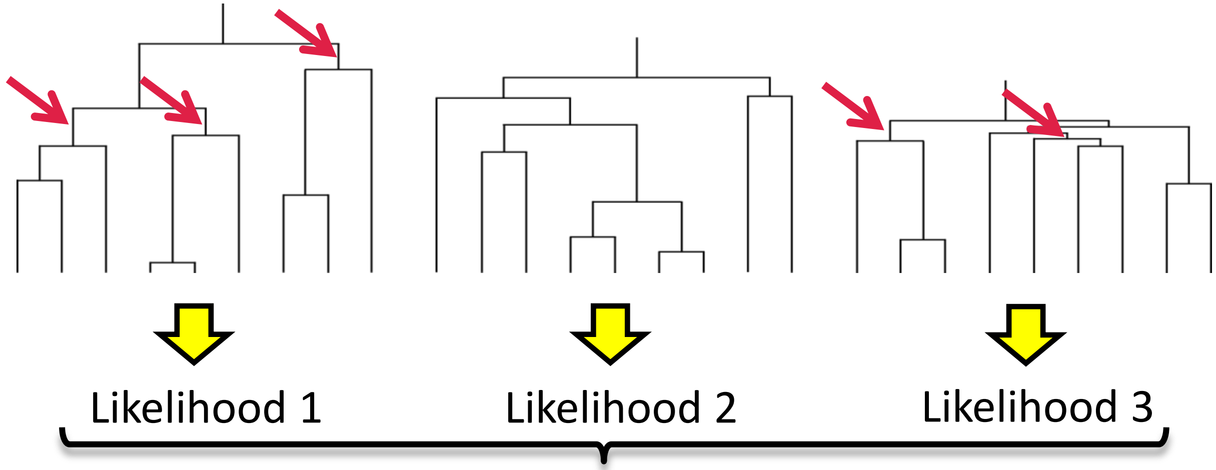
1) Using **Site Frequency Spectrum** **SFS** instead of the full data **D**

2) Using Monte Carlo simulations to approximate P(**SFS**|**μ,N**):

   a) Simulate genealogies with fixed parameter values

   b) Compute average likelihood of the **SFS** across genealogies





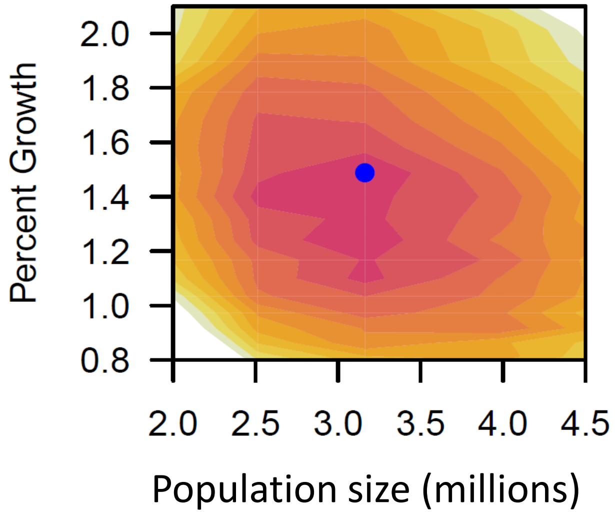
Likelihood 1 Likelihood 2 Likelihood 3

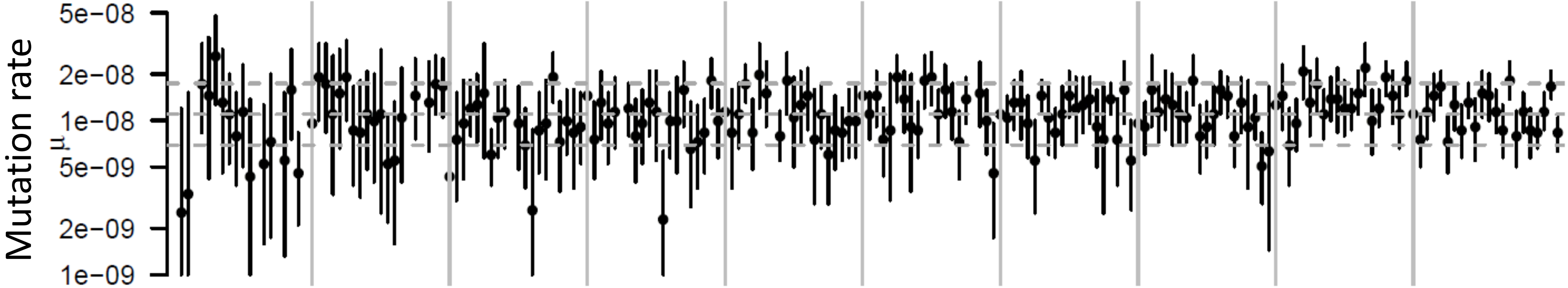
Average Likelihood

Nielsen 2000; Coventry et al. 2010

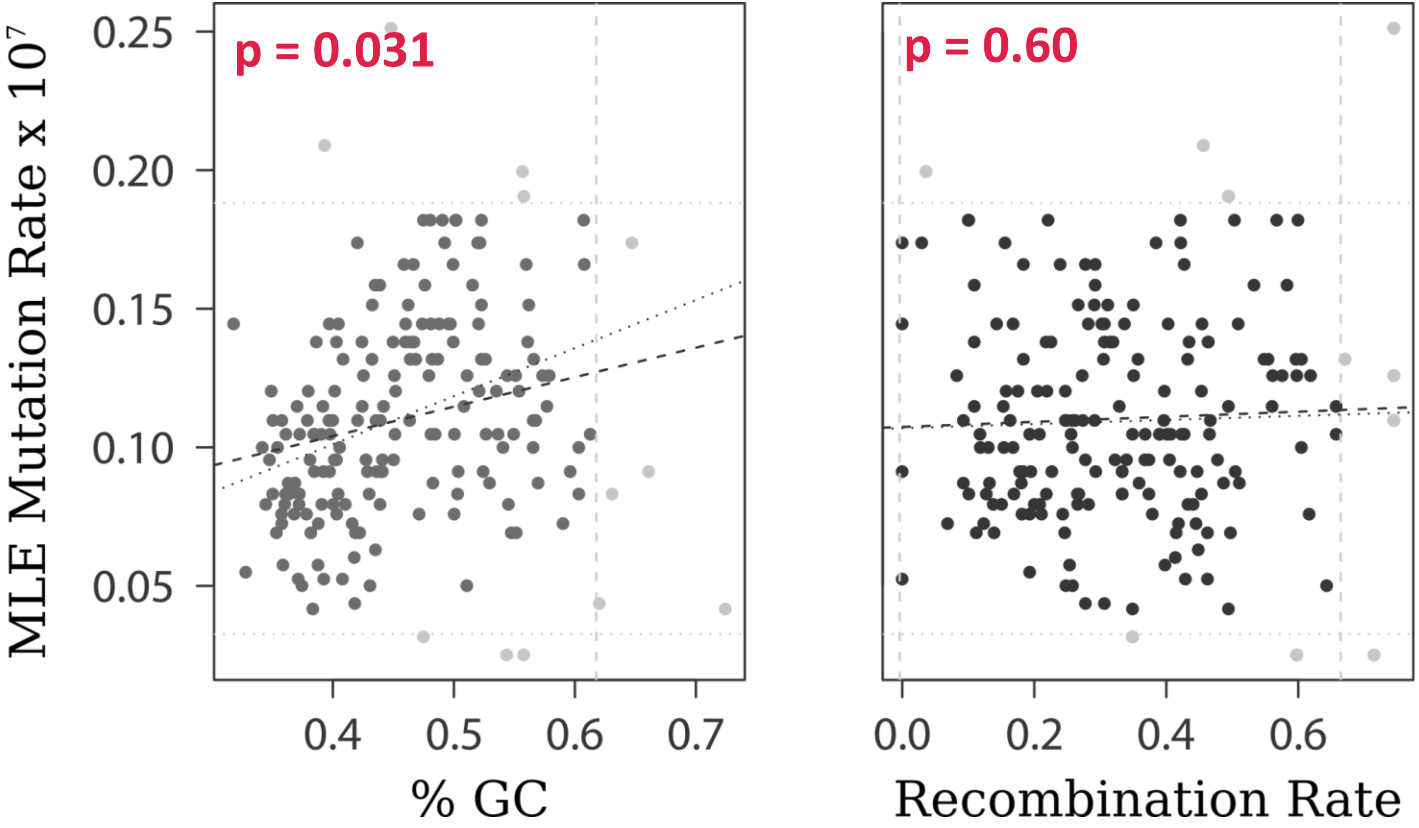
Joint inference of demography and mutation rates

- Rapid population growth in Europe
- Variable mutation rates across genes ($p < 10^{-16}$)
- Median mutation rate of 1.2×10^{-8}
  - Lower than divergence based estimates (2.5×10^{-8})
  - But in good agreement with recent estimates from pedigrees

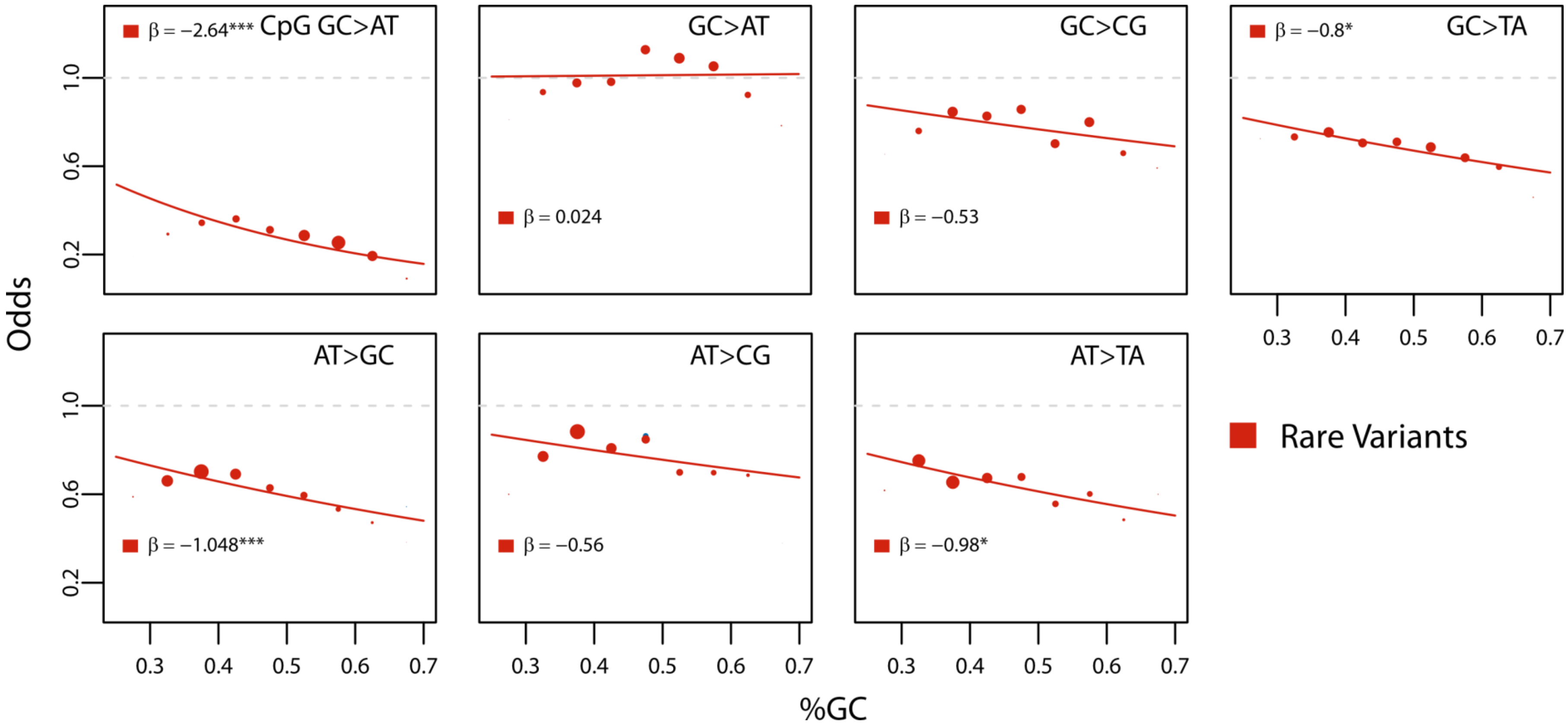




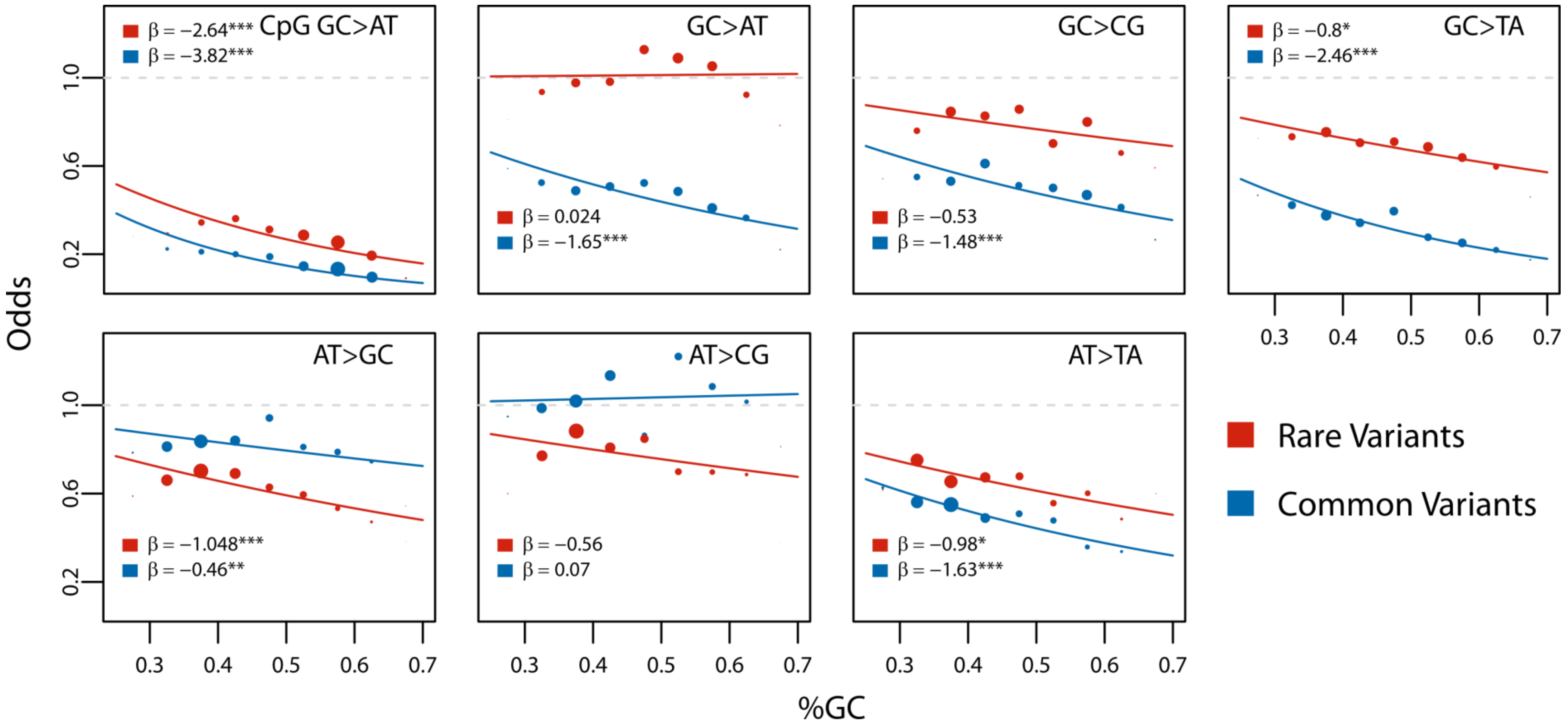
Drivers of mutation rate variation



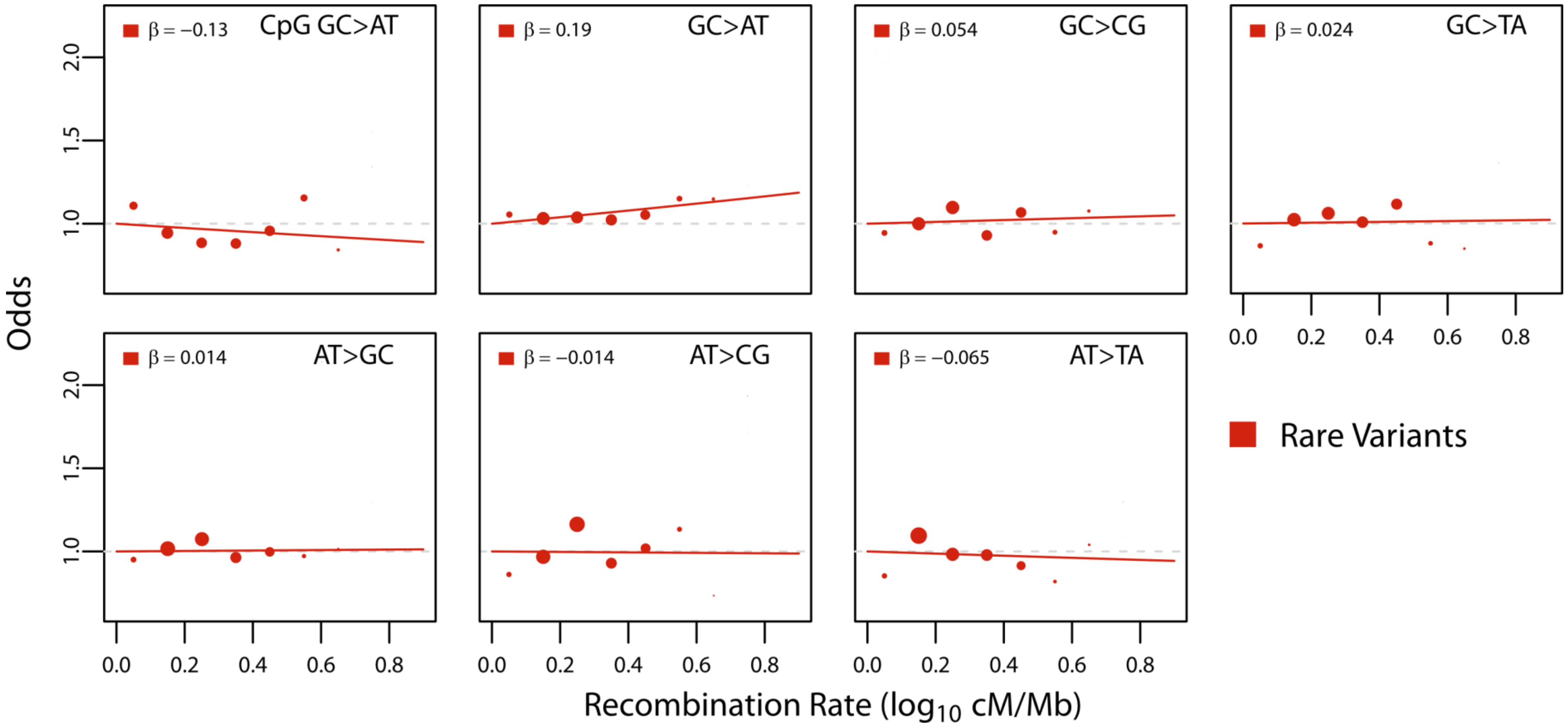
Effect of GC due to CpG sites only



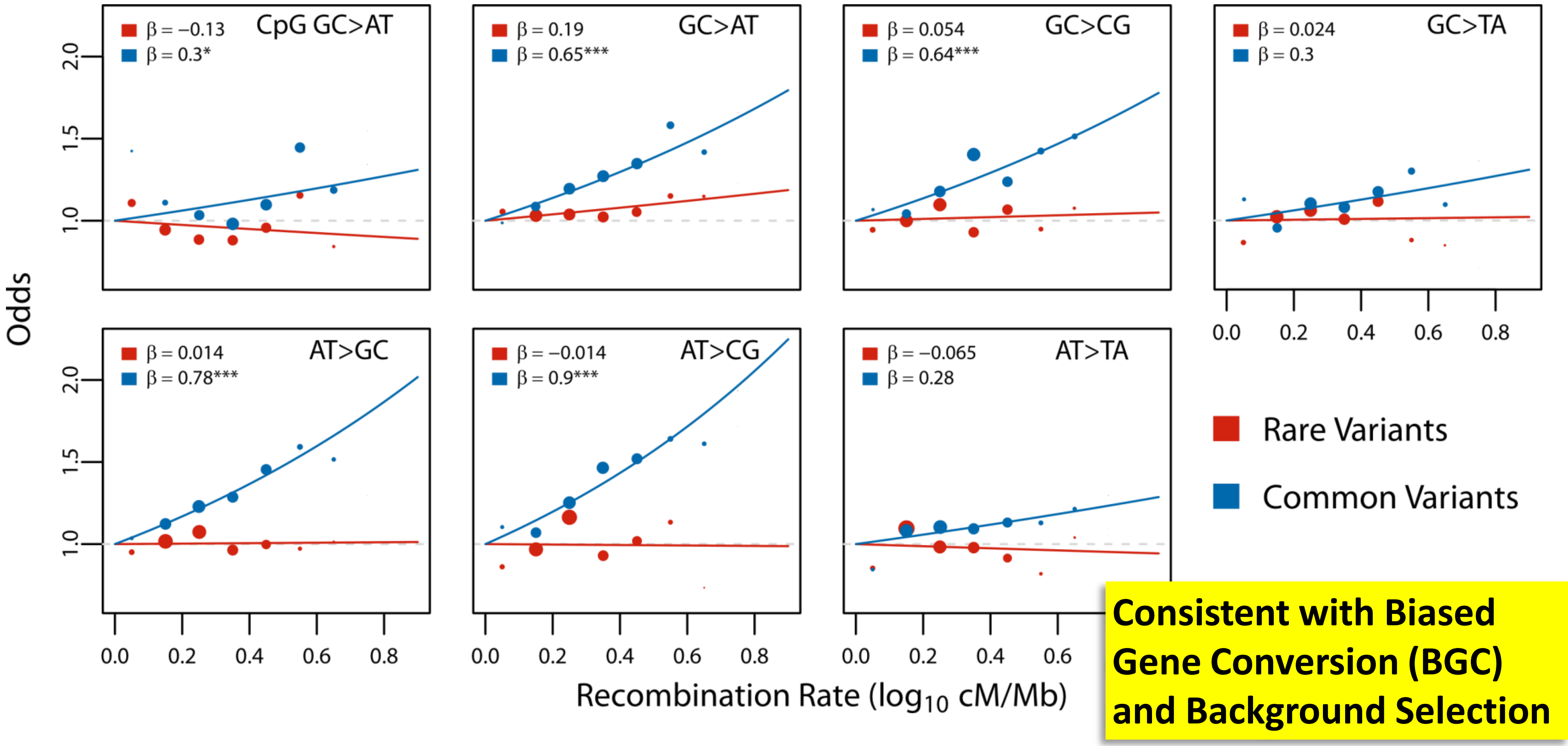
Effect of GC due to CpG sites only



Recombination rate has no effect on mutation rates



Recombination rate has no effect on mutation rates



Consistent with Biased Gene Conversion (BGC) and Background Selection

Examples of Model Based Inference

1 **Human mutation rates**
using maximum likelihood of summary statistics

2 **Demographic histories**
using Approximate Bayesian Computation

Mode of Speciation in Rose Finches

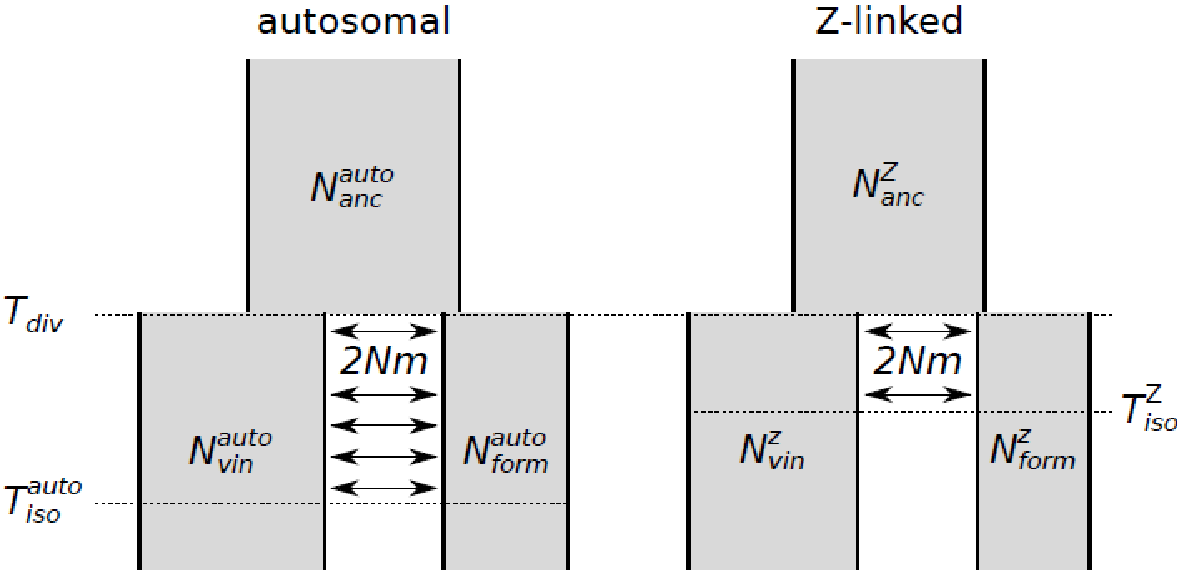
- In the classic view, **geographic isolation** was considered essential for speciation.
- However, recent evidence suggests that local adaptation and speciation may occur in the presence of **gene flow** if ecological selection is strong.

- In Birds, the **Z-chromosome** is known to play a vital role is speciation
    - **Haldanes Rule**: In hybrids, fintness is lower in the hemizygous sex (females)
    - Male **sexually selected traits and female preference** was mapped to the Z-chromosome in several species.

- **Prediction**
  If selection against hybrids is a driving force in speciation, gene flow will be interrupted ealier on the Z-chromosome than on autosomes.

Mode of Speciation in Rose Finches

- Inferring the isolation times for Z-linked and autosomal markers seperately.



Carpodacus vinaceus (Himalaya)

Carpodacus formosa (Taiwan)

Two major difficulties

- For realistic evolutionary models, analytical solutions of the likelihood function are usually **very hard** and often **impossible** to obtain.

- We will use two tricks:

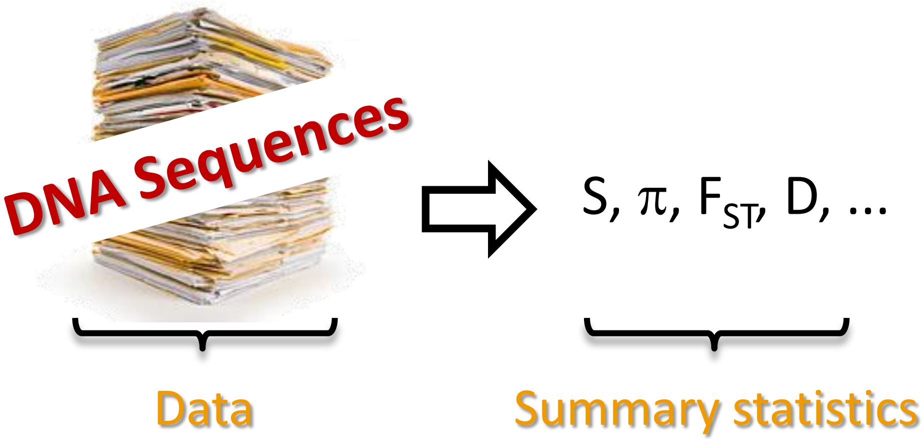
  1) Using **summary statistics** **S** instead of the full data **D**
     - The hope is that $P(\mathbf{D}|\boldsymbol{\theta})$ is proportional to $P(\mathbf{S}|\boldsymbol{\theta})$
  2) Using **simulations** to approximate the likelihood function $P(\mathbf{S}|\boldsymbol{\theta})$

- Apply in a Bayesian setting: $\underbrace{P(\boldsymbol{\theta}|\mathbf{D})}_{\text{Posterior}} \propto \underbrace{P(\mathbf{D}|\boldsymbol{\theta})}_{\text{Likelihood}} \underbrace{P(\boldsymbol{\theta})}_{\text{Prior}}$

➡ Approximate Bayesian Computation (ABC)

Approximate Bayesian Computation ABC

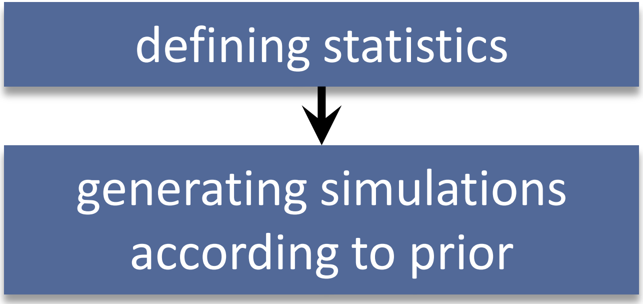
defining statistics



$S, \pi, F_{ST}, D, \ldots$

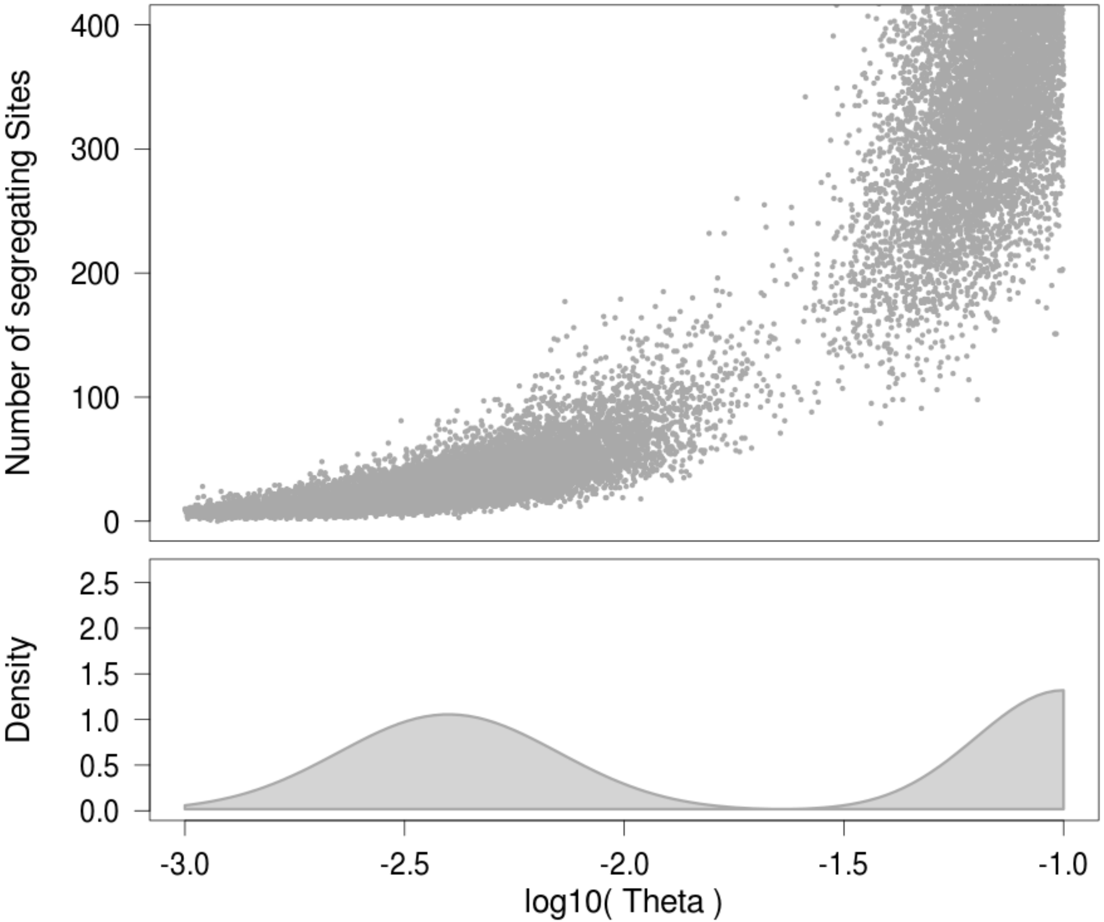
Tavaré *et al.* (1997); Weiss & von Haeseler (1998)

Approximate Bayesian Computation ABC



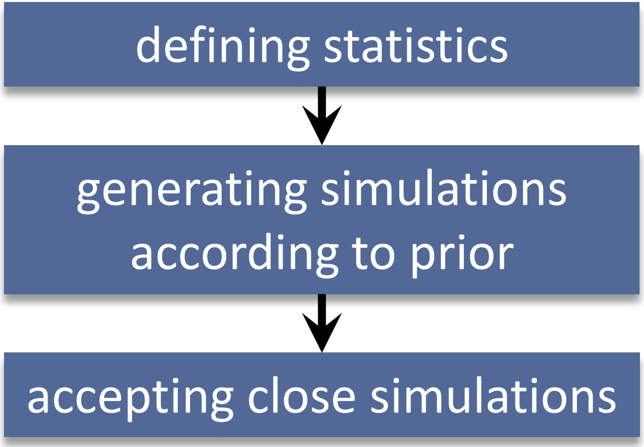
defining statistics

generating simulations according to prior



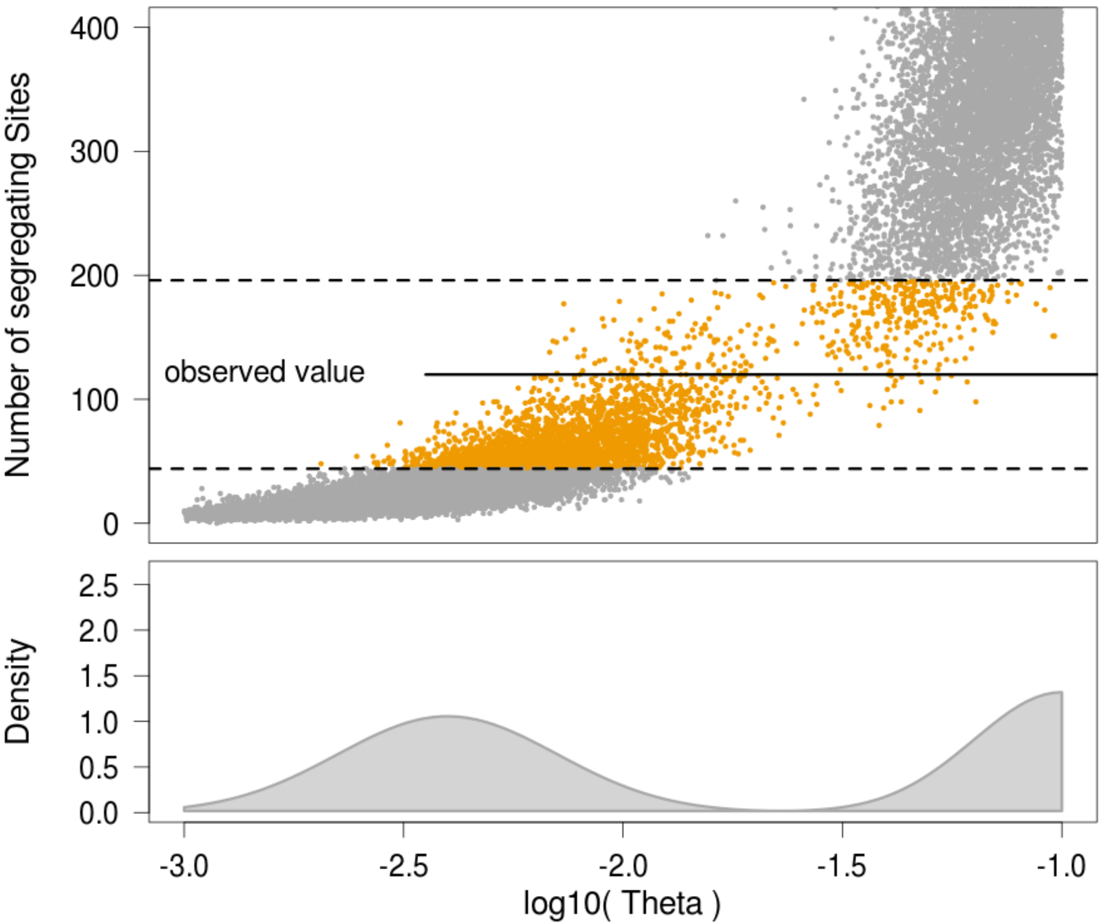
Tavaré *et al.* (1997); Weiss & von Haeseler (1998)

Approximate Bayesian Computation ABC


defining statistics

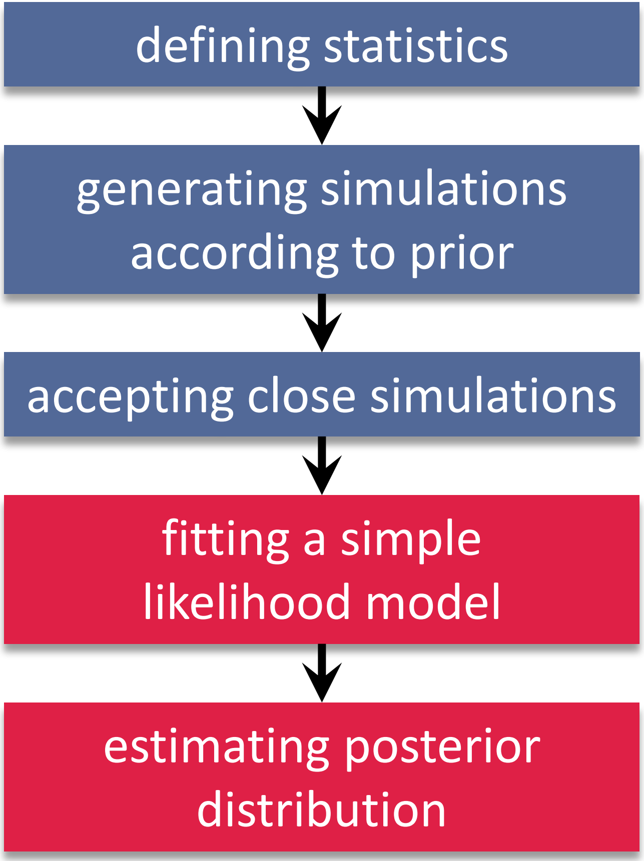
generating simulations according to prior

accepting close simulations



Tavaré *et al.* (1997); Weiss & von Haeseler (1998)

Approximate Bayesian Computation ABC



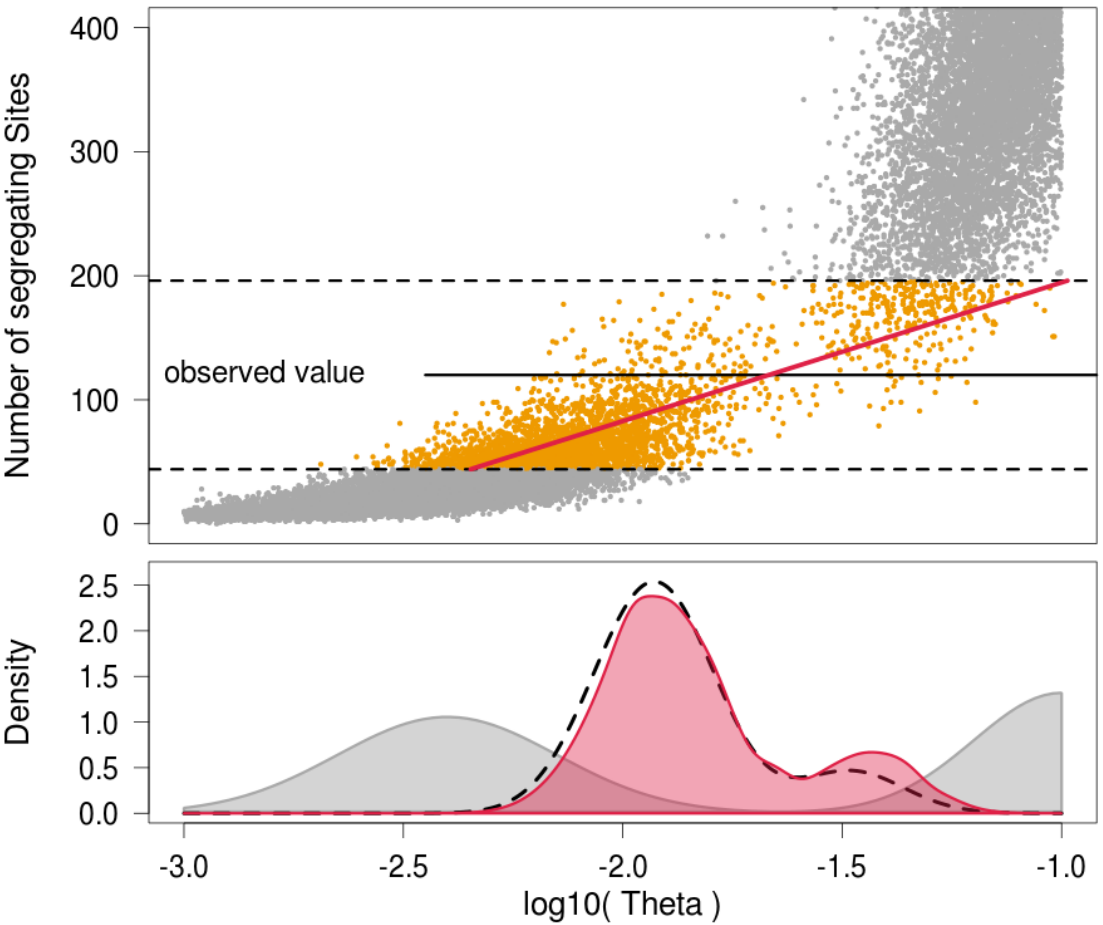
defining statistics

generating simulations according to prior

accepting close simulations

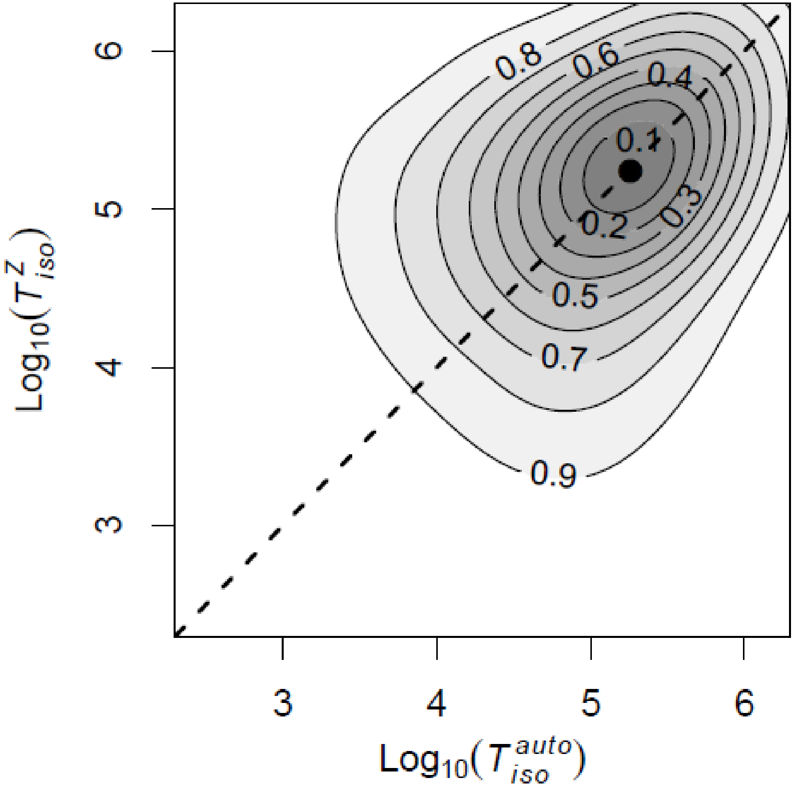
fitting a simple likelihood model

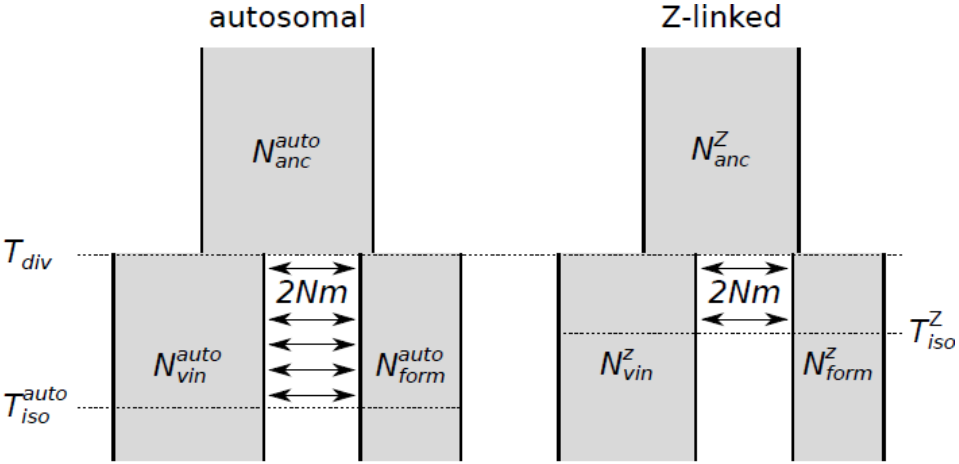
estimating posterior distribution



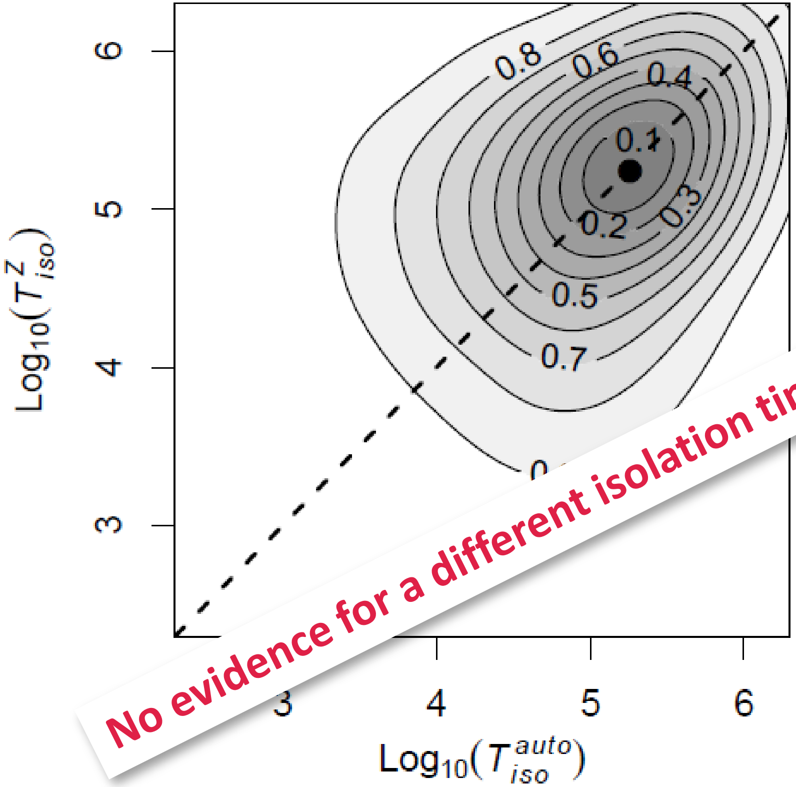
Leuenberger & Wegmann (2010)

Mode of Speciation in Rose Finches

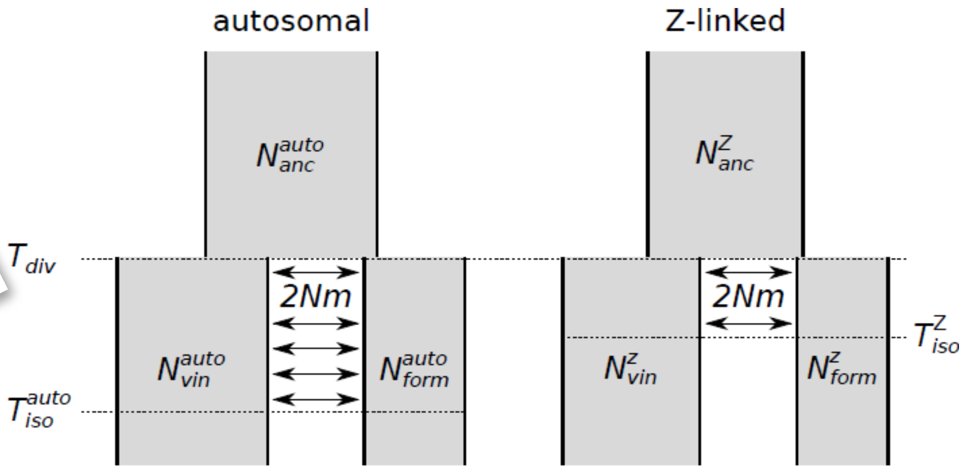




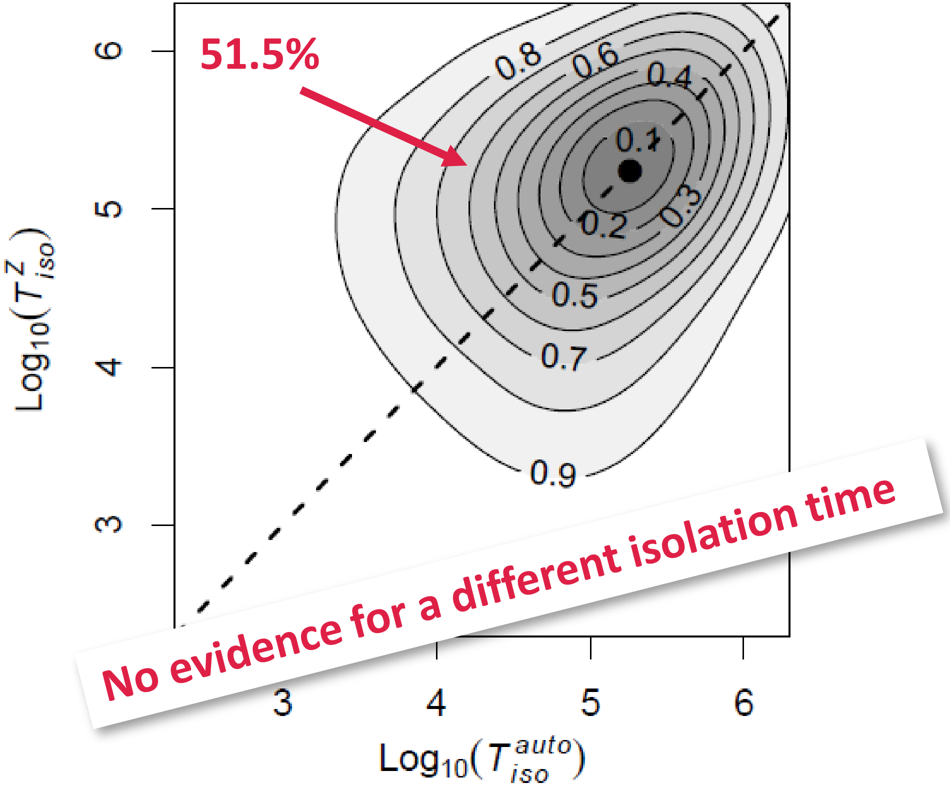
Mode of Speciation in Rose Finches



No evidence for a different isolation time



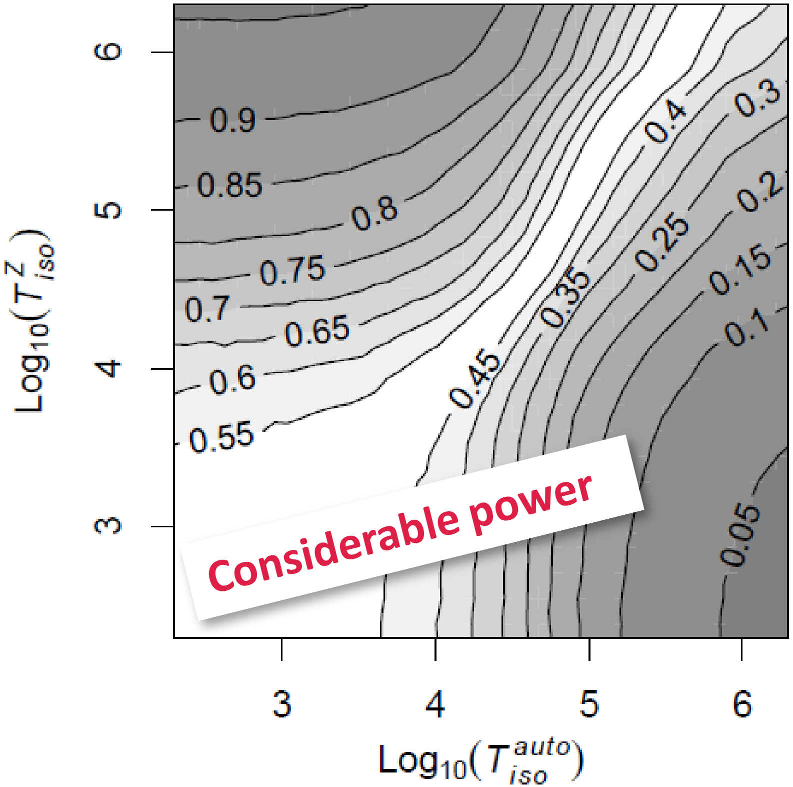
Mode of Speciation in Rose Finches



51.5%

No evidence for a different isolation time

Joint posterior asymmetry observed in simulated data sets



Considerable power

Conclusions

- While preferred, model based inference of evolutionary histories is challenging due to the **stochasticity** and **complexity** of realistic models.

- As a consequence, we often rely on **numerical approaches** (e.g. simulations).
  - It may help to replace the full data with **summary statistics**.
  - Approximate Bayesian Computation is an **extremely flexible** but crude approach.

- **On the bright side**:
  Such techniques allow us to estimate what we are interested in, rather than requiring us to shift to problems, for which analytical solutions are available.

Acknowledgements

John Novembre
U Chigaco

Matt Nelson
GSK

Shou-Hsien Li
Taiwan Normal U

Chris Leuenberger
U Fribourg

Mutation rates
Li Li
Liling Warren
Claudio Verzilli
Keith Nangle
Karen King
John Whittaker
Jennifer Aponte
Vincent Mooser
Stephanie Chissoe
Sebastian Zöllner
Matt Zawistowski
Peng Zhang
Yancy Lo
Shyam Gopalakrishnan
Jun Li
Goncalo Abecasis

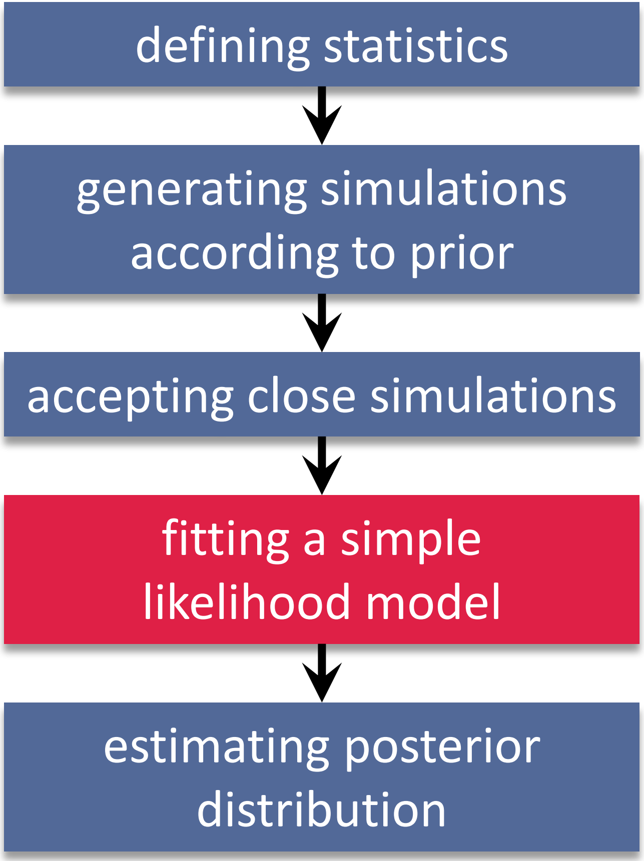
Rose Finches
Riva Chu
Jui-Hua Chu
Chia-Fen Yeh
Rong-Chien Lin
Xiao-Jun Yang
Fu-Min Lei
Cheng-Te Yao
Fa-Sheng Zhou

UNI
FR

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

FNSNF
SWISS NATIONAL SCIENCE FOUNDATION

ABC-GLM


defining statistics

generating simulations according to prior

accepting close simulations

fitting a simple likelihood model

estimating posterior distribution

- It is easy to show that

$$\pi(\boldsymbol{\theta} \mid \mathbf{s}_{\mathrm{obs}}) \propto f_\epsilon(\mathbf{s}_{\mathrm{obs}} \mid \boldsymbol{\theta})\,\pi_\epsilon(\boldsymbol{\theta})$$

- where $f_\epsilon(\mathbf{s} \mid \boldsymbol{\theta})$ is the truncated likelihood

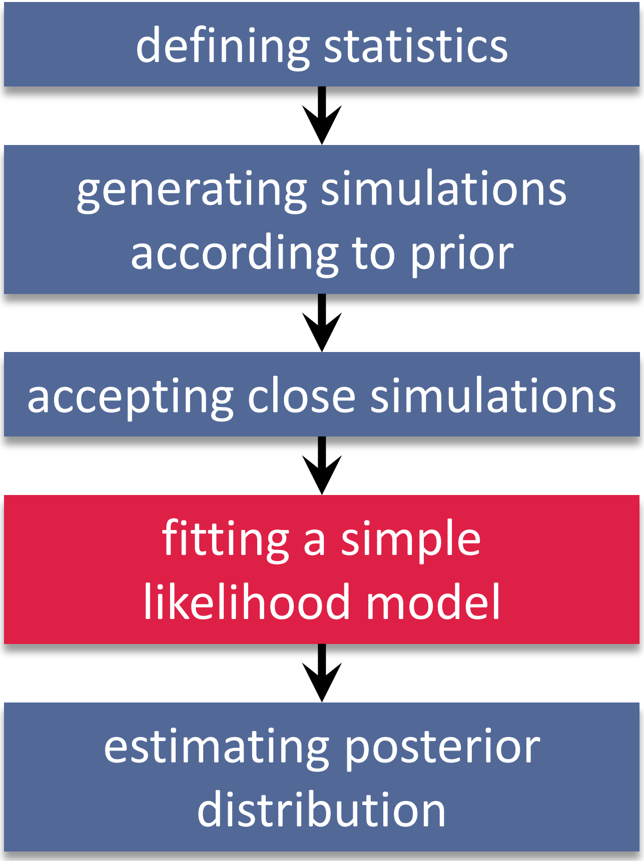
$$f_\epsilon(\mathbf{s} \mid \boldsymbol{\theta}) \propto \mathrm{Ind}(\mathbf{s} \in \underbrace{\mathcal{B}_\epsilon(\mathbf{s}_{\mathrm{obs}})}_{\{\mathbf{s} \in \mathbb{R}^n \mid \mathrm{dist}(\mathbf{s},\, \mathbf{s}_{\mathrm{obs}}) < \boldsymbol{\epsilon}\}}) \cdot f_{\mathcal{M}}(\mathbf{s} \mid \boldsymbol{\theta})$$

- and $\pi_\epsilon(\boldsymbol{\theta})$ the „truncated prior”

$$\pi_\epsilon(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{B}_\epsilon} f_{\mathcal{M}}(\mathbf{s} \mid \boldsymbol{\theta})\, ds$$

Leuenberger & Wegmann (2010)

ABC-GLM


defining statistics
generating simulations according to prior
accepting close simulations
fitting a simple likelihood model
estimating posterior distribution

$$\pi(\boldsymbol{\theta} \mid \mathbf{s}_{\text{obs}}) \propto f_{\epsilon}(\mathbf{s}_{\text{obs}} \mid \boldsymbol{\theta})\pi_{\epsilon}(\boldsymbol{\theta})$$

Assume GLM (estimate via OLS)

$$\mathbf{s} \mid \boldsymbol{\theta} = \mathbf{C}\boldsymbol{\theta} + \mathbf{c}_0 + \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_s)$$

From retained sample using Gaussian peaks

$$\pi_{\epsilon}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{j=1}^{N}\phi(\boldsymbol{\theta} - \boldsymbol{\theta}^{j}, \boldsymbol{\Sigma}_{\theta})$$

Leuenberger & Wegmann (2010)

Hybridizing ABC with Full Likelihood: ABC-GLM

- Given data $\mathcal{D} = \{D_l, S_{abc}\}$ where D_l and S_{abc} are independent, the posterior is given by $\pi(\boldsymbol{\theta}|\mathcal{D}) \propto f(\mathcal{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = f(D_l|\boldsymbol{\theta})f(S_{abc}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Hybridizing ABC with Full Likelihood: ABC-GLM

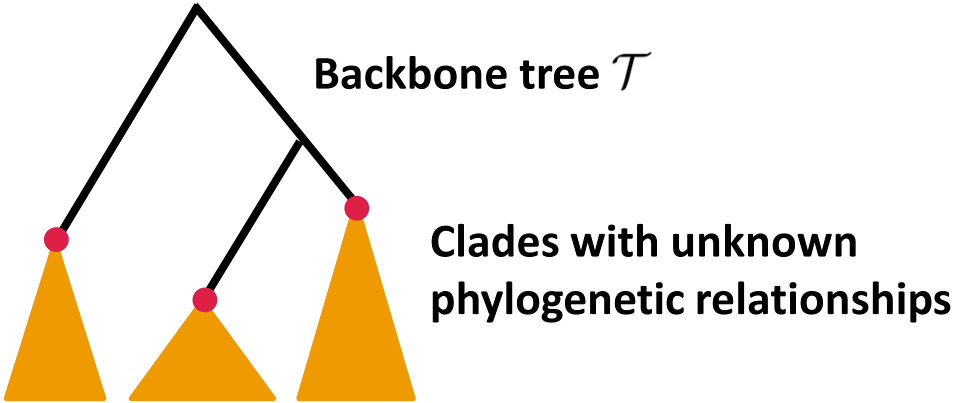
- Given data $\mathcal{D} = \{D_l, S_{abc}\}$ where D_l and S_{abc} are independent, the posterior is given by $\pi(\boldsymbol{\theta}|\mathcal{D}) \propto f(\mathcal{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = f(D_l|\boldsymbol{\theta}) f(S_{abc}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

- Since $\dfrac{f(\boldsymbol{S}_{abc}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_\Pi f(\boldsymbol{S}_{abc}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} = \dfrac{f_\epsilon(\boldsymbol{S}_{abc}|\boldsymbol{\theta})\pi_\epsilon(\boldsymbol{\theta})}{\int_\Pi f_\epsilon(\boldsymbol{S}_{abc}|\boldsymbol{\theta})\pi_\epsilon(\boldsymbol{\theta})d\boldsymbol{\theta}}$,

  which implies that $f(\boldsymbol{S}_{abc}|\boldsymbol{\theta}) = \dfrac{f_\epsilon(\boldsymbol{S}_{abc}|\boldsymbol{\theta})\pi_\epsilon(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \cdot c(\boldsymbol{S}_{abc})$,

  the posterior is given by $\pi(\boldsymbol{\theta}|D_l, \boldsymbol{S}_{abc}) \propto f(D_l|\boldsymbol{\theta}) f_\epsilon(\boldsymbol{S}_{abc}|\boldsymbol{\theta})\pi_\epsilon(\boldsymbol{\theta})$

Hybridizing ABC with Full Likelihood


Backbone tree $\mathcal{T}$
Clades with unknown phylogenetic relationships

$$\mathbb{P}(\mathbf{D}|\mathcal{O}, \mathbf{s_0^2}, \beta, \delta, \mathcal{T})$$

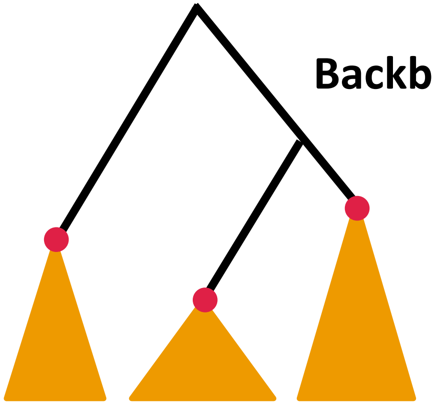
Trait values
mean and variance within clade

Brownian model
$\mathcal{O}$ = root state of trait
$\mathbf{s_0^2}$ = rate of trait evolution

Phylogenetic birth-death process
β = species birthrate
δ = species death rate

Slater *et al.* (2011)

Hybridizing ABC with Full Likelihood



Backbone tree $\mathcal{T}$

Clades with unknown phylogenetic relationships

$$\mathbb{P}(\mathbf{D}|\mathcal{O},\mathbf{s}_0^2,\beta,\delta,\mathcal{T}) = \sum_{\mathbf{T}\in\mathbf{\Omega}} \mathbb{P}(\mathbf{D}|\mathcal{O},\mathbf{s}_0^2,\mathbf{T})\cdot\mathbb{P}(\mathbf{T}|\beta,\delta,\mathcal{T})$$

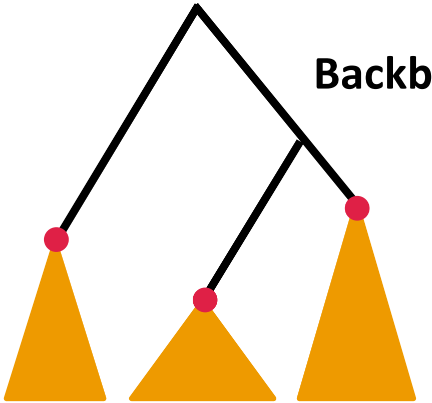
Trait values
mean and variance within clade

Brownian model
$\mathcal{O}$ = root state of trait
$\mathbf{s}_0^2$ = rate of trait evolution

Phylogenetic birth-death process
β = species birthrate
δ = species death rate

Slater *et al.* (2011)

Hybridizing ABC with Full Likelihood



Backbone tree $\mathcal{T}$

Clades with unknown phylogenetic relationships

$$\mathbb{P}(\mathbf{D}|\mathcal{O},\mathbf{s}_{\mathbf{0}}^{\mathbf{2}},\beta,\delta,\mathcal{T}) = \sum_{\mathbf{T}\in\boldsymbol{\Omega}} \overbrace{\mathbb{P}(\mathbf{D}|\mathcal{O},\mathbf{s}_{\mathbf{0}}^{\mathbf{2}},\mathbf{T})}^{\text{ABC-MCMC}} \cdot \overbrace{\mathbb{P}(\mathbf{T}|\beta,\delta,\mathcal{T})}^{\text{Metropolis-Hastings}}$$

Trait values
mean and variance within clade

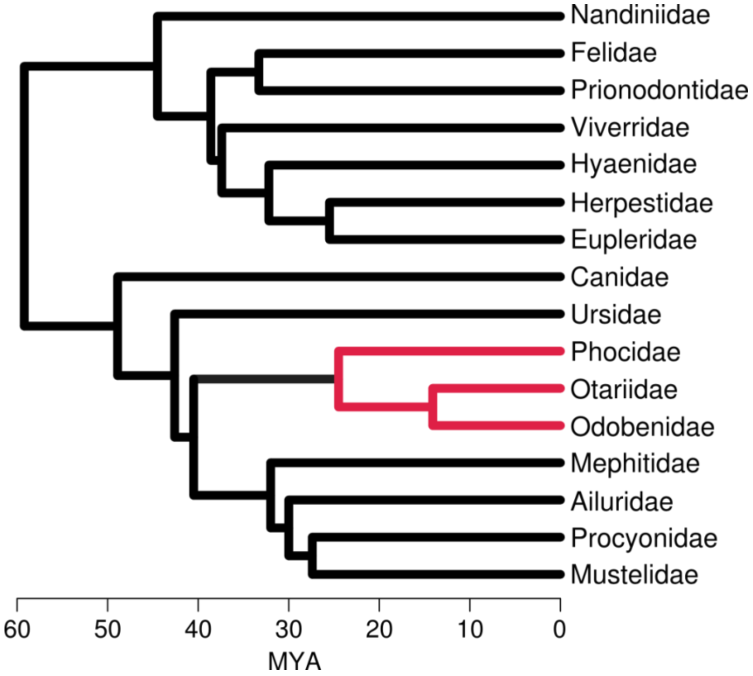
Brownian model
$\mathcal{O}$ = root state of trait
$\mathbf{s}_{\mathbf{0}}^{\mathbf{2}}$ = rate of trait evolution

Phylogenetic birth-death process
β = species birthrate
δ = species death rate

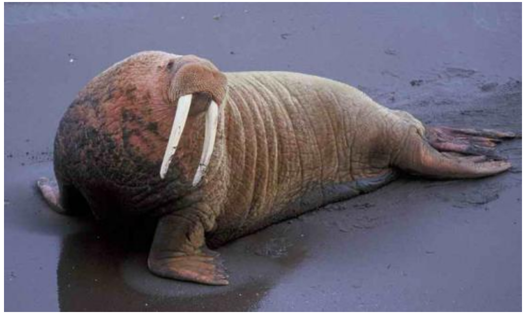
Slater *et al.* (2011)

Application to Body Size Evolution in Carnivora

- Several members of the semiaquatic **Pinnipedia** attain very large body sizes.
- Did body size evolve faster among **Pinnipedia** than all other Carnivora?



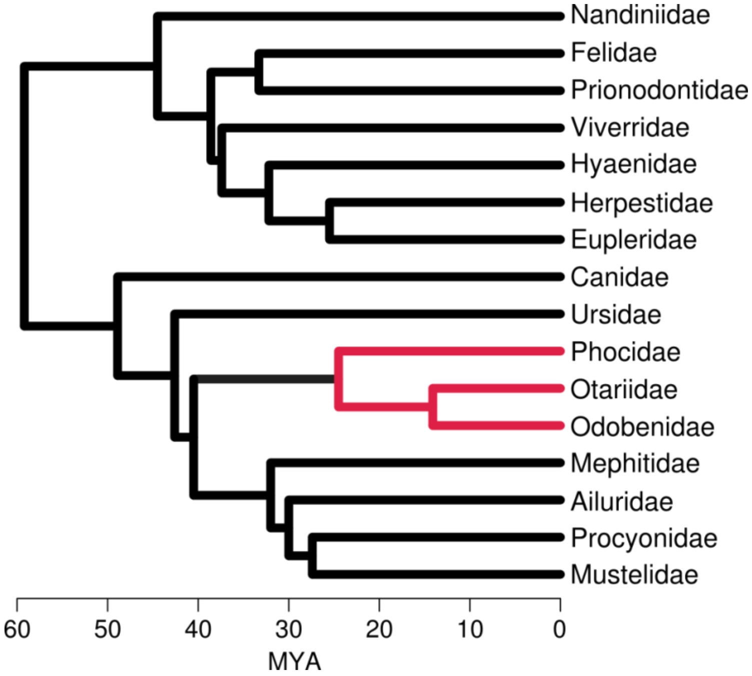
Southern Elephant Seal
up to 4,000 Kg

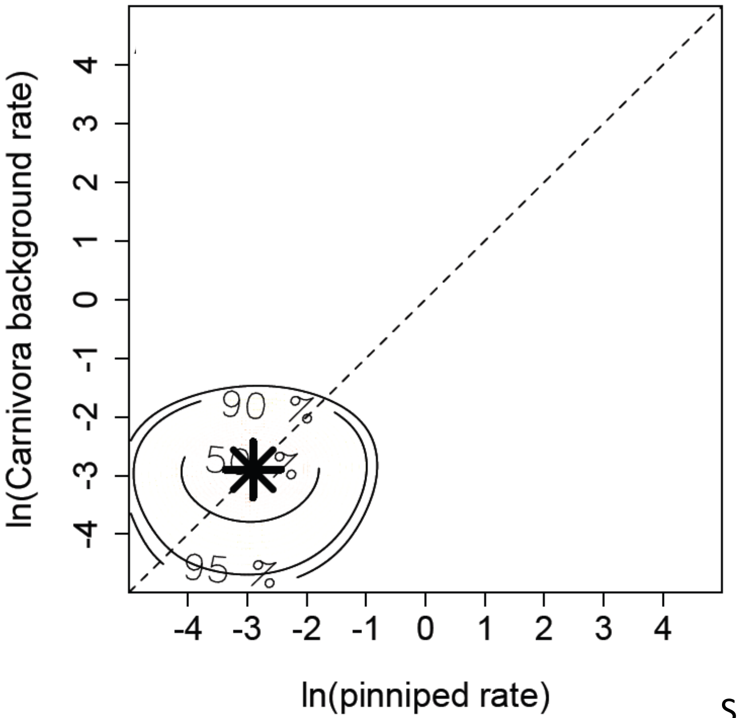


Walrus
up to 1,800 Kg

Slater *et al.* (2011)

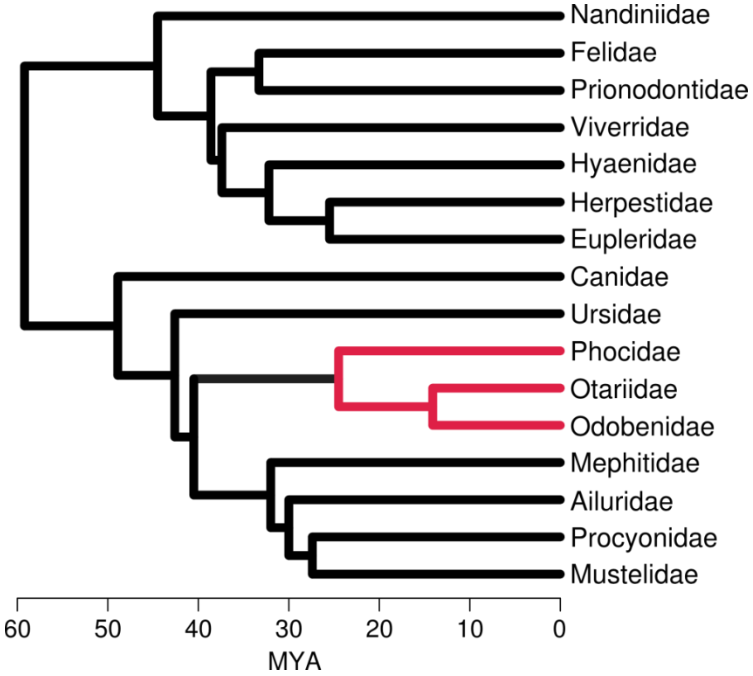
- Several members of the semiaquatic **Pinnipedia** attain very large body sizes.
- Did body size evolve faster among **Pinnipedia** than all other Carnivora?

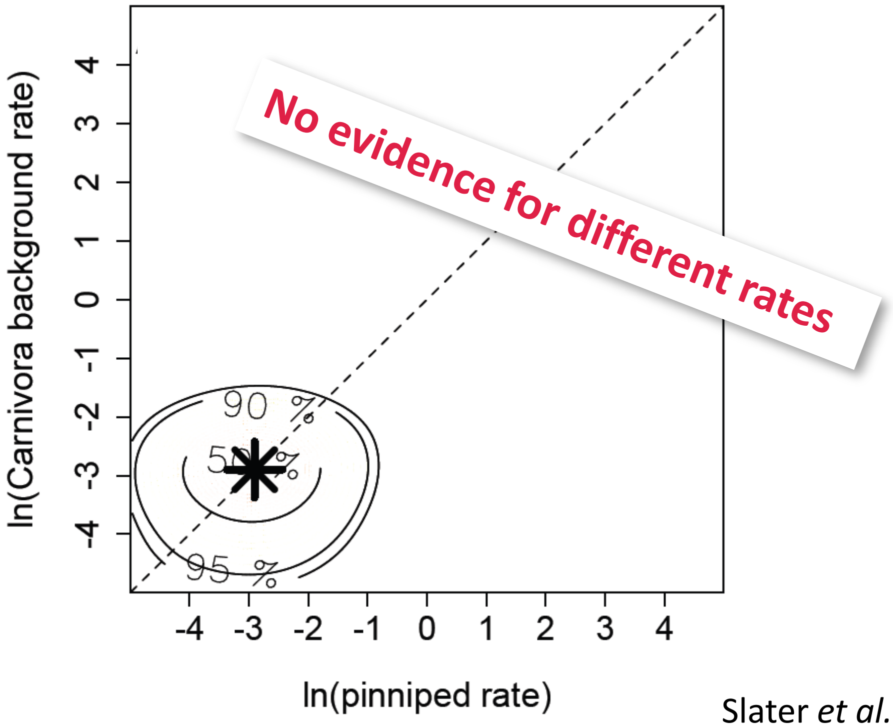




Slater *et al.* (2011)

- Several members of the semiaquatic **Pinnipedia** attain very large body sizes.
- Did body size evolve faster among **Pinnipedia** than all other Carnivora?





No evidence for different rates

Slater *et al.* (2011)

ABC with Independent Loci

- Often, loci are assumed to be independent

$$\mathbf{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_n\}$$

- We can thus estimate the truncated likelihood as

$$P_\varepsilon(\mathbf{S} \mid \theta) = P_\varepsilon(\mathbf{S}_1 \mid \theta) \cdot P_\varepsilon(\mathbf{S}_2 \mid \theta) \cdot \ldots \cdot P_\varepsilon(\mathbf{S}_n \mid \theta)$$

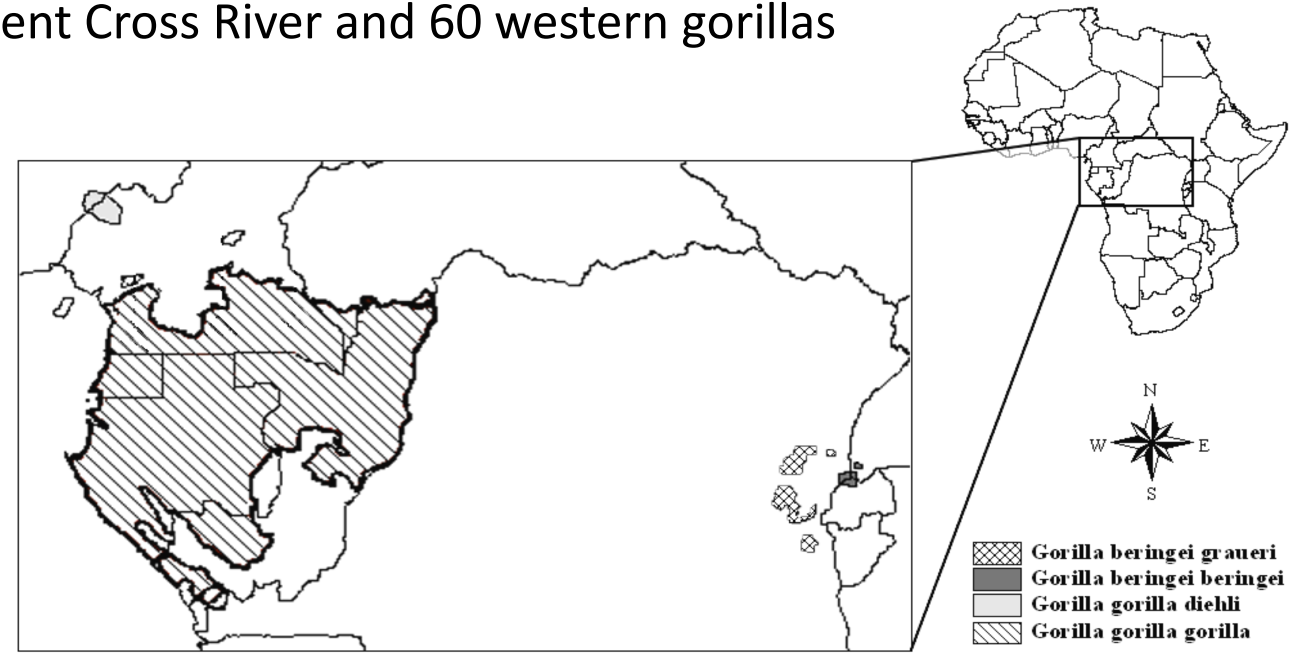
- The likelihood is estimated from simulations of a single locus!

➡ Massive reduction in computation time

Thalmann & Wegmann et al. (2011)

Application to Cross River Gorillas

- Highly endangered subspecies with < 300 individuals
- 7 microsatellites
- 11 ancient, 68 current Cross River and 60 western gorillas

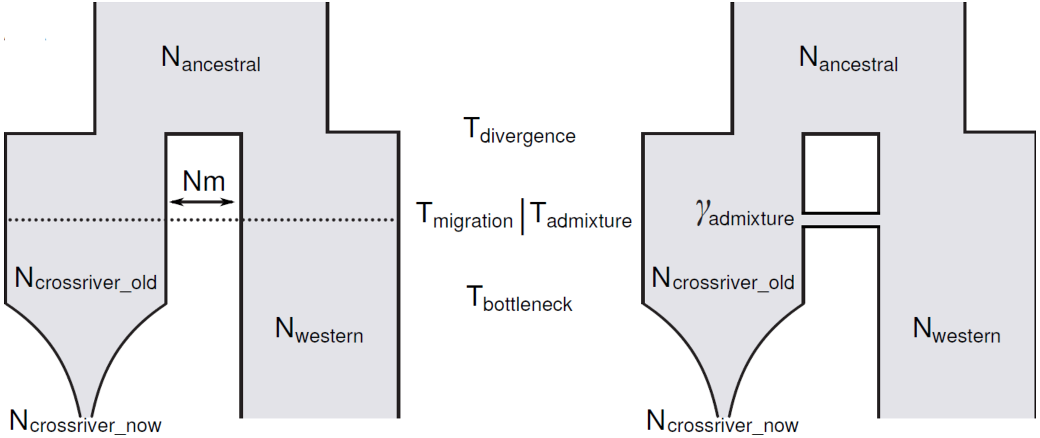


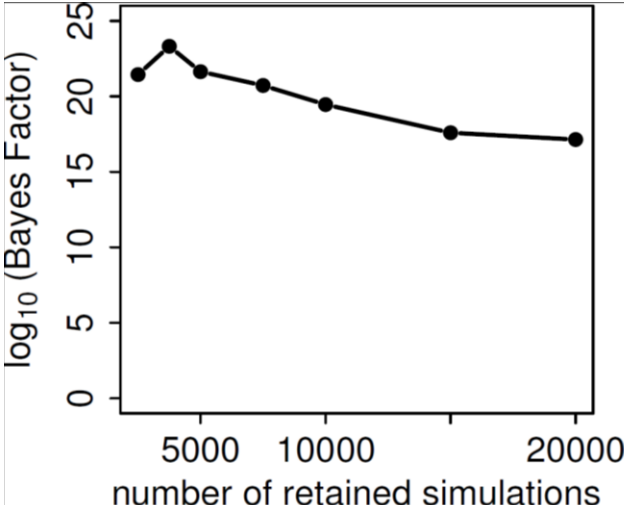
Thalmann & Wegmann et al. (2011)

Application to Cross River Gorillas

- Highly endangered subspecies with < 300 individuals
- 7 microsatellites
- 11 ancient, 68 current Cross River and 60 western gorillas

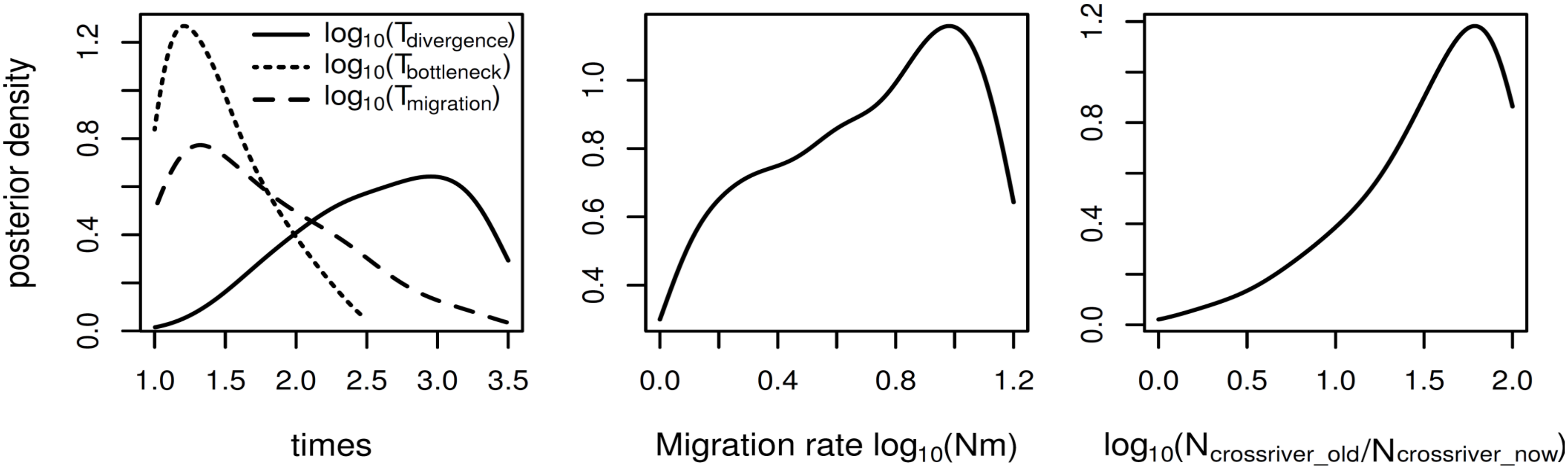
- Population split with gene flow more likely than admixture





Thalmann & Wegmann et al. (2011)

Application to Cross River Gorillas



- Old divergence, followed by high levels of gene flow
- Gene flow ceased only recently, probably at onset of strong bottleneck in Cross River gorillas (~45 times)

Thalmann & Wegmann et al. (2011)

Composite Likelihood ABC

- Concept can easily be extended to models with locus specific parameters:

$$\theta = \{N, \mu_1, \mu_2, \ldots, \mu_n\}$$

- In which case we can estimate the truncated likelihood as

$$f_\varepsilon(S \mid \theta) = f_\varepsilon(S_1 \mid N, \mu_1) \cdot \ldots \cdot f_\varepsilon(S_2 \mid N, \mu_n)$$