UNI
FR

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Coalescent based demographic inference (for NGS)

Daniel Wegmann
University of Fribourg

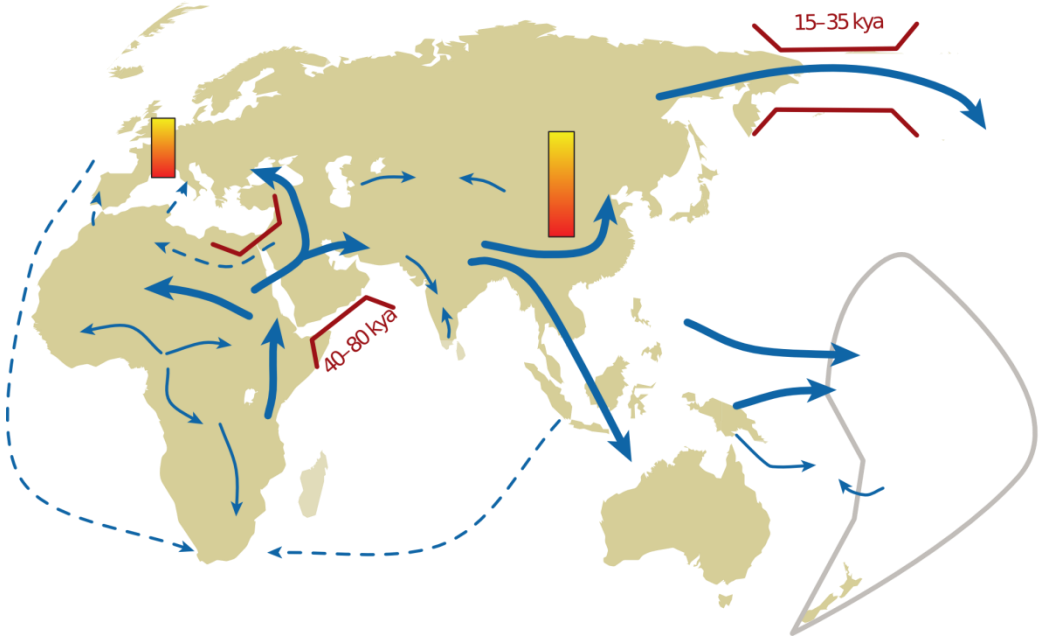
Introduction

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

Introduction

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

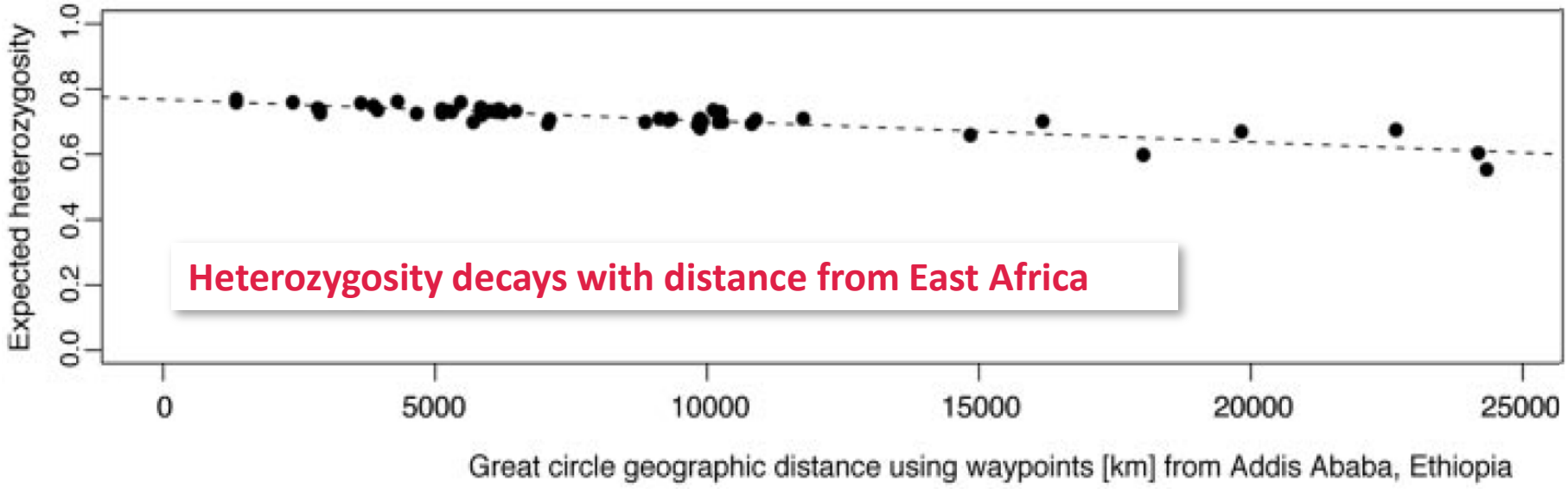
- But this is a **challenging task!**
  - The history of natural populations is usually **complex**.
  - Several evolutionary processes can leave **similar footprints** (bottleneck vs. selection).
  - Loci are not independent, but **correlated realizations** of the same process.



Novembre & Ramachandran (2011)

Qualitative inference

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



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**.

➡ 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**.

➡ 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!

Rejection of a Null Model

- The same as hypothesis testing in frequentist statistics:
  A null model **M** is rejected using a summary statistics ***s*** if $\int_{s=s_{obs}}^{\infty} P(s \mid M) < \alpha$
- By convention, α = 0.05
- Often the Null model is an isolated Wright-Fisher population of constant size

Rejection of a Null Model

- The same as hypothesis testing in frequentist statistics:
  A null model **M** is rejected using a summary statistics ***s*** if $\int_{s=s_{obs}}^{\infty} P(s \mid M) < \alpha$
- By convention, α = 0.05
- Often the Null model is an isolated Wright-Fisher population of constant size

Example: F-Statistics

- F_{ST} may be used to reject a panmictic population in favor of a specific structure
- F_{IS} may be used to reject a panmictic population in favor of non-random mating (inbreeding or substructure)
- The significance of F-Statistics is usually assessed using permutation or randomization approaches.

Rejection of a Null Model: F-Statistics

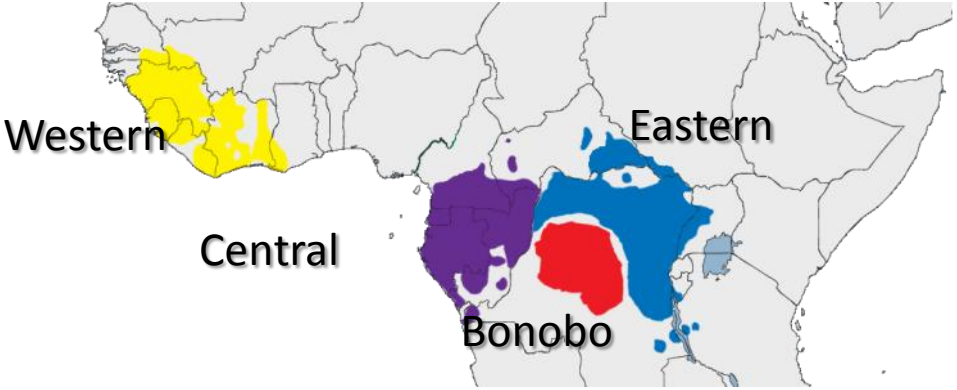


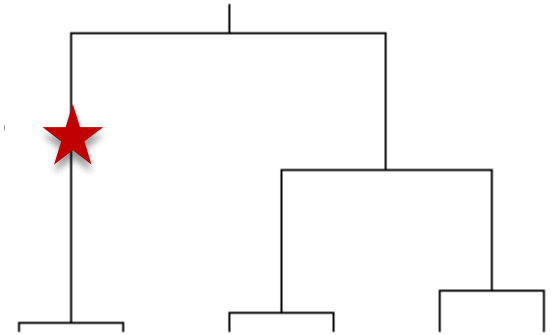
Table 1. Observed Population- and Marker-Specific F_{IS} Values.

Sample	DNA	Microsatellites
Bonobo	−0.054*	0.023
Eastern chimpanzee	0.049*	0.093*
Central chimpanzee	0.111*	0.057*
Western chimpanzee	0.096*	0.026*

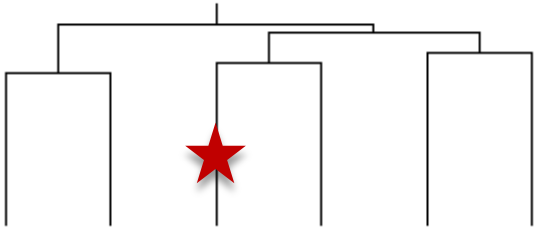
Wegmann & Excoffier, MBE, 2010

Rejection of a Null Model: Tajima's D

- Tajima's ***D*** compares two estimates of $\theta=4N\mu$ for a Wright-Fisher population of constant size:
  - one based on the number segregating sites ***S***
  - one based on the average number of pairwise differences π
- These estimates may differ when assumptions of the Wright-Fisher population are violated.
- An expanding population, for instance, leads to a negative ***D***
- Significance is usually assessed via simulations.



Witgh-Fisher population



expanding population

The Felsenstein Equation

The Likelihood Function

The probability of the data $\mathcal{D}$ given the parameters of the model Θ: $P(\mathcal{D}|\Theta)$

Maximum Likelihood Inference

The maximum likelihood estimates are the values of Θ for which the likelihood $P(\mathcal{D}|\Theta)$ is maximized.

The Felsenstein Equation

The Likelihood Function

The probability of the data $\mathcal{D}$ given the parameters of the model Θ: $P(\mathcal{D}|\Theta)$

Maximum Likelihood Inference

The maximum likelihood estimates are the values of Θ for which the likelihood $P(\mathcal{D}|\Theta)$ is maximized.

Bayesian Statistics

The goal is to infer the probability of the parameters Θ given the data $\mathcal{D}$.

According to probability theory,

$$P(\Theta|\mathcal{D}) = \frac{P(\mathcal{D}|\Theta)P(\Theta)}{P(\mathcal{D})} = \frac{P(\mathcal{D}|\Theta)P(\Theta)}{\int_\Theta P(\mathcal{D}|\Theta)P(\Theta)d_\Theta}$$

Here,

- $P(\Theta)$ is the **prior** probability, the probability of the parameter *before* looking at the data (yes, this is subjective!).
- $P(\Theta|\mathcal{D})$ is the **posterior** probability of the parameter *after* considering the data.

Mutation Model

Likelihood of sequence data given a Genealogy

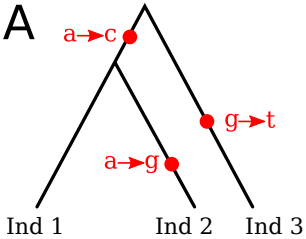
The link between sequencing data $\mathcal{D}$ and some demographic parameters Θ is the underlying, unknown genealogy.

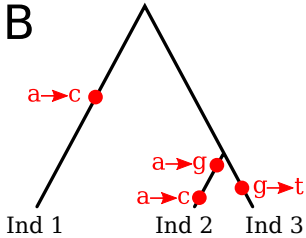
Given a genealogy G_i and a mutation model μ, the likelihood of the data is straight forward to calculate.

```
Ind 1  :  aagacacaga gatagaccag
Ind 2  :  aagacgcaga gatagaccag
Ind 3  :  aagacacaga tatagacaag
```

Assuming all mutations to occur with rate μ:

$$P(\mathcal{D}|G_i,\mu) = \prod_{b\in\{\text{Branches}\}} P(\#\text{ mutations on } b|\text{length}(b),\mu)$$





The Felsenstein Equation

The Felsenstein Equation

Calculating $P(\mathcal{D}|\Theta)$ requires to integrate over *all possible genealogies* and weighting each by their probability.

$$P(\mathcal{D}|\Theta, \mu) = \int_G P(\mathcal{D}|G, \mu) P(G|\Theta) dG$$

The Felsenstein Equation

The Felsenstein Equation

Calculating $P(\mathcal{D}|\Theta)$ requires to integrate over *all possible genealogies* and weighting each by their probability.

$$P(\mathcal{D}|\Theta, \mu) = \int_G P(\mathcal{D}|G, \mu) P(G|\Theta) dG$$

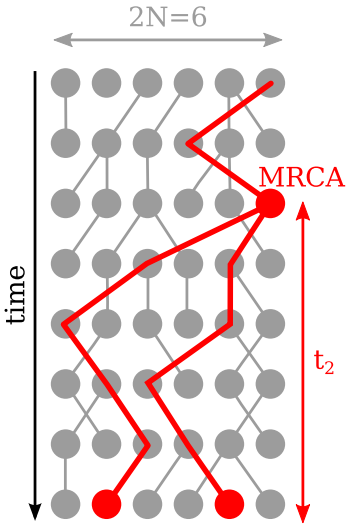
The Felsenstein Equation in practice

Unfortunately, this integral is impossible to solve analytically in all but some extremely simple models.

In practice, we thus approximate this integral using a random sample of coalescent trees.

$$P(\mathcal{D}|\Theta, \mu) \approx \frac{1}{N} \sum_{i=1}^{N} P(\mathcal{D}|G_i, \mu) \quad \text{where} \quad g_i \sim P(G|\Theta)$$

Primer in Coalescent Theory



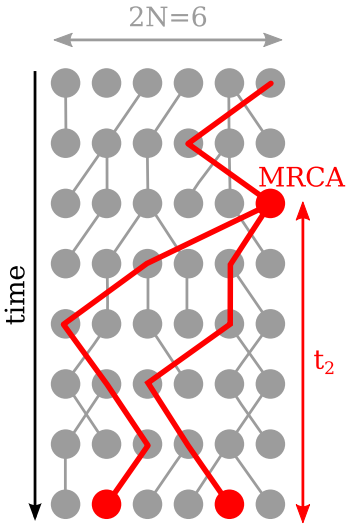
Coalescent theory

A population genetic theory that considers the history of a sample **backward in time**.

Coalescent event

If two sampled lineages have the same parent in the previous generation.

Primer in Coalescent Theory



Coalescent theory

A population genetic theory that considers the history of a sample **backward in time**.

Coalescent event

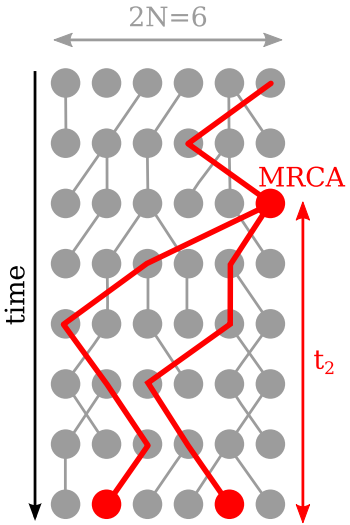
If two sampled lineages have the same parent in the previous generation.

Probability to coalesce

Under random mating in a constant population, two lineages coalesce in the previous generation with probability

$$Pr(2 \text{ individuals coalesce}) = \frac{1}{2N}$$

Primer in Coalescent Theory



Coalescent theory

A population genetic theory that considers the history of a sample **backward in time**.

Coalescent event

If two sampled lineages have the same parent in the previous generation.

Probability to coalesce

Under random mating in a constant population, two lineages coalesce in the previous generation with probability

$$Pr(2 \text{ individuals coalesce}) = \frac{1}{2N}$$

Expected time t_2 until two lineages coalesce (time to Most Recent Common Ancestor, MRCA): $E[t_2] = 2N$ generations.

Coalescence with multiple samples

Probability of coalescent

$$Pr(\text{at least one coalescent event}) = \binom{k}{2}\frac{1}{2N} = \frac{k(k-1)}{4N}$$

Intuitive explanation

Probability of coalescence among k lineages = probability of coalescence among two lineages $\frac{1}{2N}$ times the number of possible pairs $\binom{k}{2}$.

Coalescence with multiple samples

Probability of coalescent

$$Pr(\text{at least one coalescent event}) = \binom{k}{2}\frac{1}{2N} = \frac{k(k-1)}{4N}$$

Intuitive explanation

Probability of coalescence among k lineages = probability of coalescence among two lineages $\frac{1}{2N}$ times the number of possible pairs $\binom{k}{2}$.

Expected time t_k until k lineages coalesce

Coalescence with multiple samples

Probability of coalescent

$$Pr(\text{at least one coalescent event}) = \binom{k}{2}\frac{1}{2N} = \frac{k(k-1)}{4N}$$

Intuitive explanation

Probability of coalescence among k lineages = probability of coalescence among two lineages $\frac{1}{2N}$ times the number of possible pairs $\binom{k}{2}$.

Expected time t_k until k lineages coalesce

The expected waiting time until an event occurs the first time is given by the inverse of the probability of the event!

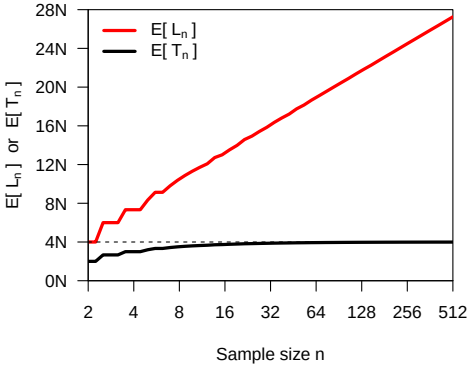
$$E[t_k] = \frac{1}{\binom{k}{2}\frac{1}{2N}} = \frac{2N}{\binom{k}{2}} = \frac{4N}{k(k-1)}$$

Expected genealogy of n samples (lineages)

Height versus length of a genealogy of n samples

$$E[T_n] = 4N\left(1 - \frac{1}{n}\right)$$

$$E[L_n] = 4N\sum_{k=1}^{n-1}\frac{1}{k}$$

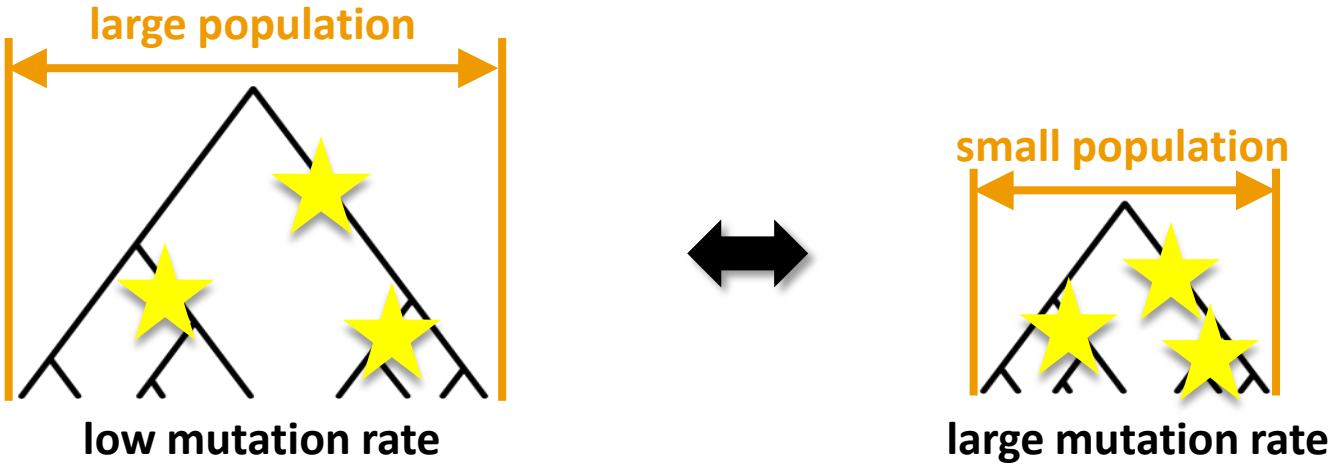


Note: Adding additional samples does increase the expected tree height only marginally, but increases the tree length a lot.

Actually, doubling of the sample size increases the tree length by about 1.5 N.

Joint inference of demography and mutation rates

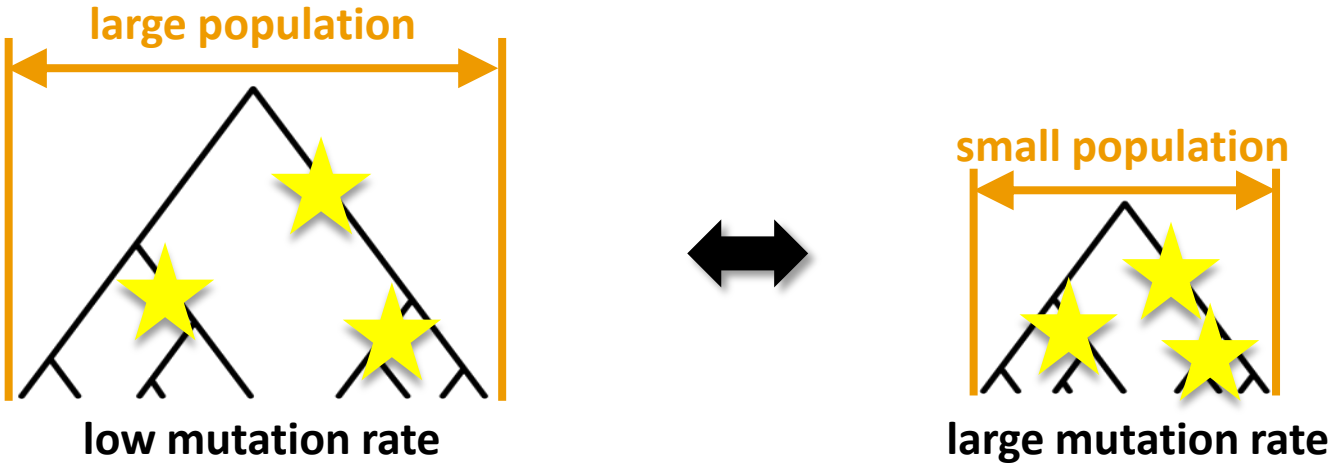
- Mutation rate μ and population size **N** have **similar effects** on genetic diversity.



Wakeley and Takahashi 2002

Joint inference of demography and mutation rates

- Mutation rate μ and population size **N** have **similar effects** on genetic diversity.



- If **sample size > effective population size:**
  - the effect of the population size is affecting the number of singletons only
  - which rensers estimation of μ and **N** individually possible.

Wakeley and Takahashi 2002

Deep resequencing data set

John Novembre

Matt Nelson

Data set:

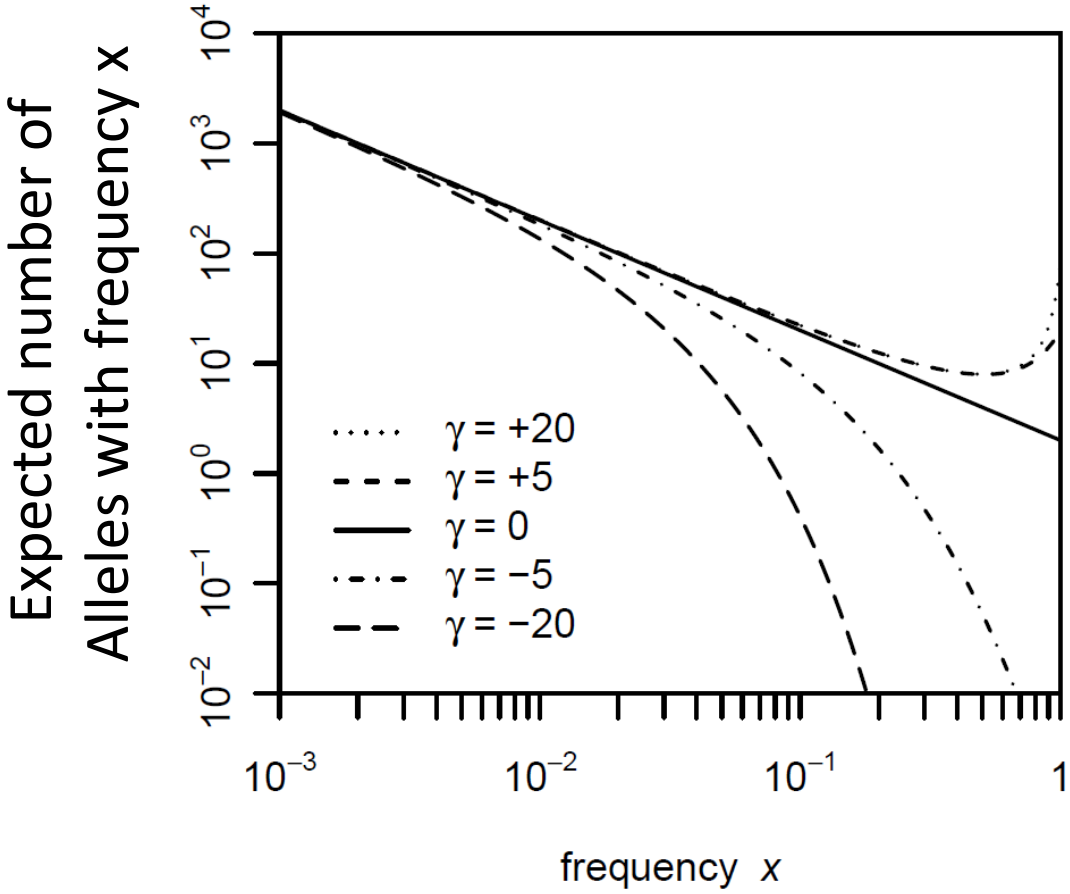
- 202 known or prospective drug target genes
- 14,002 individuals, of which 12,514 Europeans
- Median coverage of 27x and a call rate of 90.7%

Extensive quality control

- Heterozygous concordance
  - 99.1% in 130 sample duplicates
  - 99.0% in comparison to 1000G Trios
- Singleton concordance
  - 98.5% in 130 sample duplicates
  - 98.3% of 245 validated via Sanger

Wegmann & Nelson *et al.* 2012

Rare variants are only weakly affected by selection



Advantageous alleles

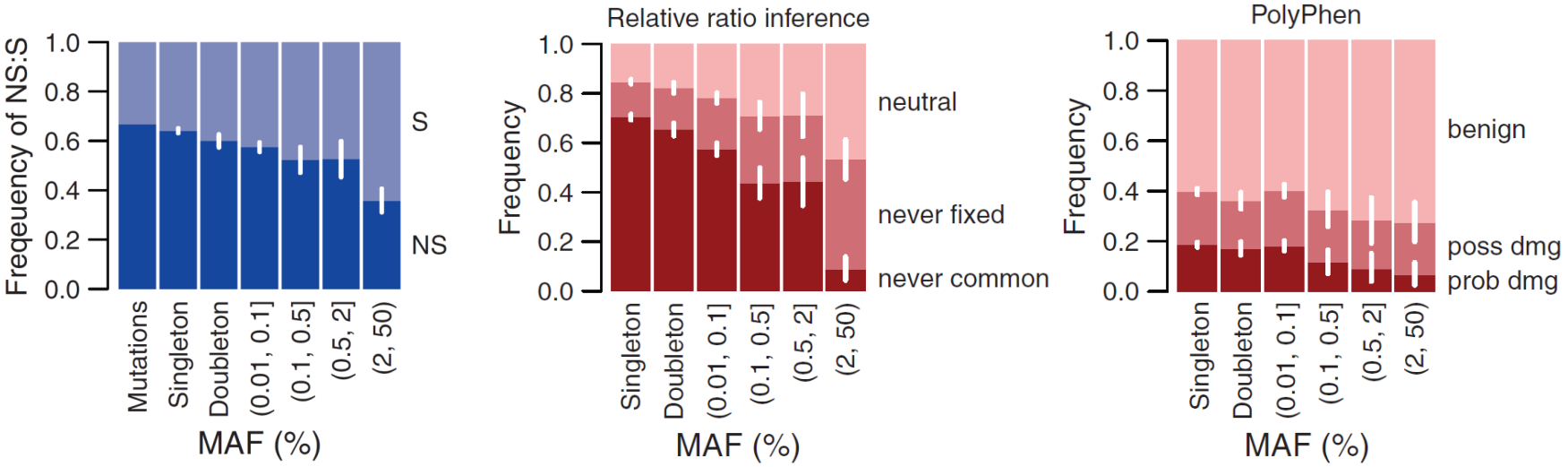
Neutral alleles

Disadvantageous alleles

Messer 2009

Phenotypic Effect of Rare Variants

- Rare variants have a strong, negative impact on the phenotype



- 85% of NS mutations are deleterious enough never to get fixed
- 75% never to never get common (MAF of 5%)
- Similar patterns found by PolyPhen

Wegmann & Nelson *et al.* 2012

Joint inference of demography and mutation rates

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

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

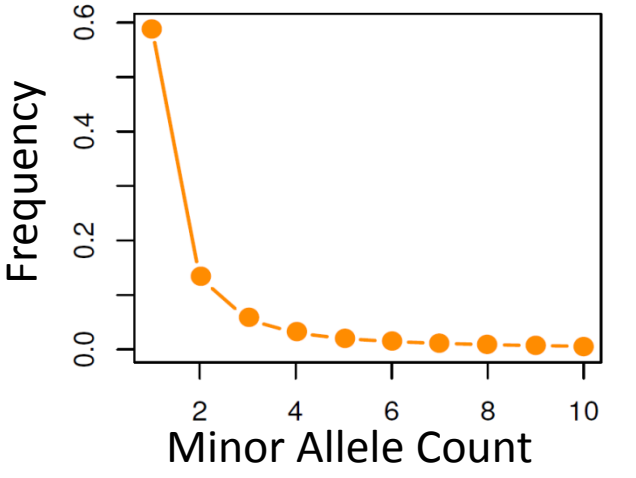
Polymorphisms **Mutation rates & Population sizes**

- Maximum-Likelihood: Find **μ,N** that maximize $P(\mathbf{D} \mid \boldsymbol{\mu},\mathbf{N})$
- 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 $P(\mathbf{D} \mid \boldsymbol{\mu},\mathbf{N})$ is proportional to $P(\mathbf{S} \mid \boldsymbol{\mu},\mathbf{N})$,
  2) Use **simulations** to approximate the likelihood function $P(\mathbf{S} \mid \boldsymbol{\mu},\mathbf{N})$

Joint inference of demography and mutation rates

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

```
AGATTCAC
AGCTTCAT
AGATTCAT
AGATTCAT
AGCTTCGC
   ⋮
```



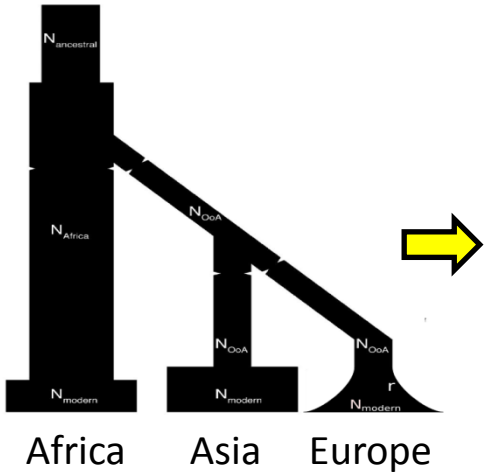
22,000 Sequences of 202 genes

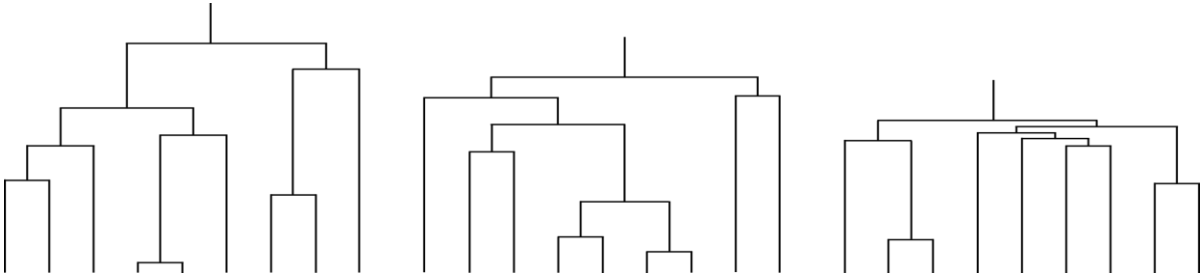
Site Frequency Spectrum SFS

Joint inference of demography and mutation rates

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

- Simulate genealogies with fixed parameter values





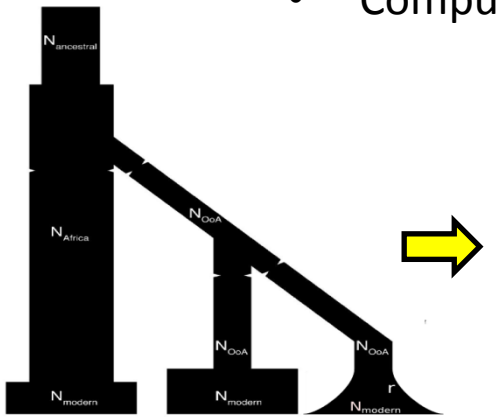
- Exponential growth in Europe
- All other parameters fixed to Schaffner estimates

Nielsen 2000; Coventry et al. 2010

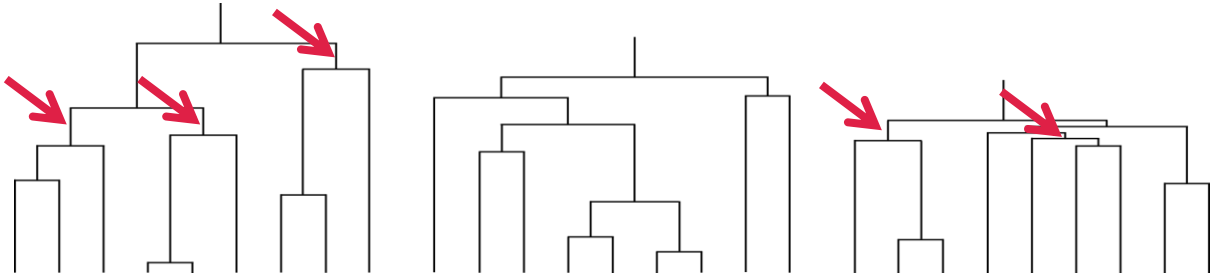
Joint inference of demography and mutation rates

Using **Monte Carlo simulations** to approximate $P(\mathbf{SFS} \mid \mu, \mathbf{N})$:

- Simulate genealogies with fixed parameter values
- Compute average likelihood of the SFS across genealogies



Africa Asia Europe



Likelihood 1 Likelihood 2 Likelihood 3

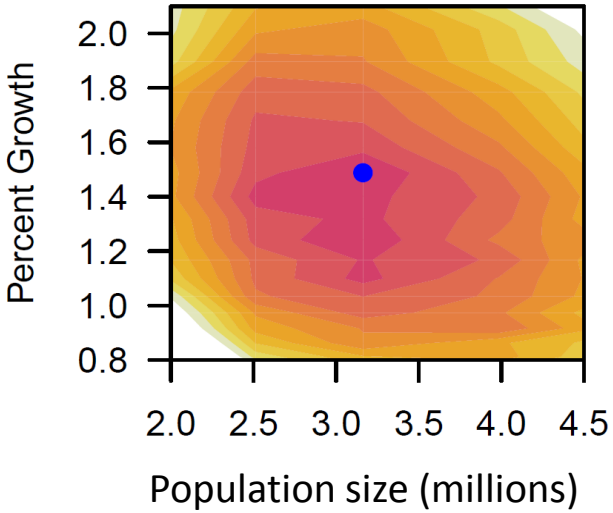
Average Likelihood

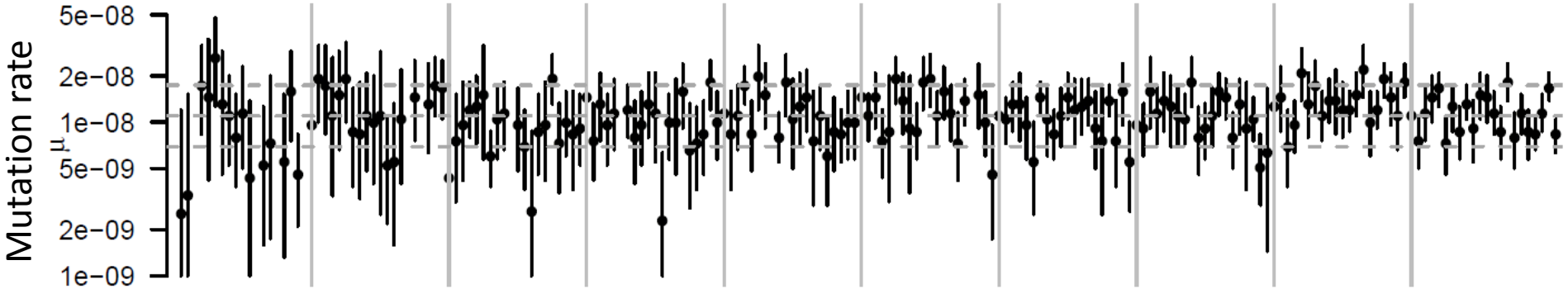
- Exponential growth in Europe
- All other parameters fixed to Schaffner estimates

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





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.

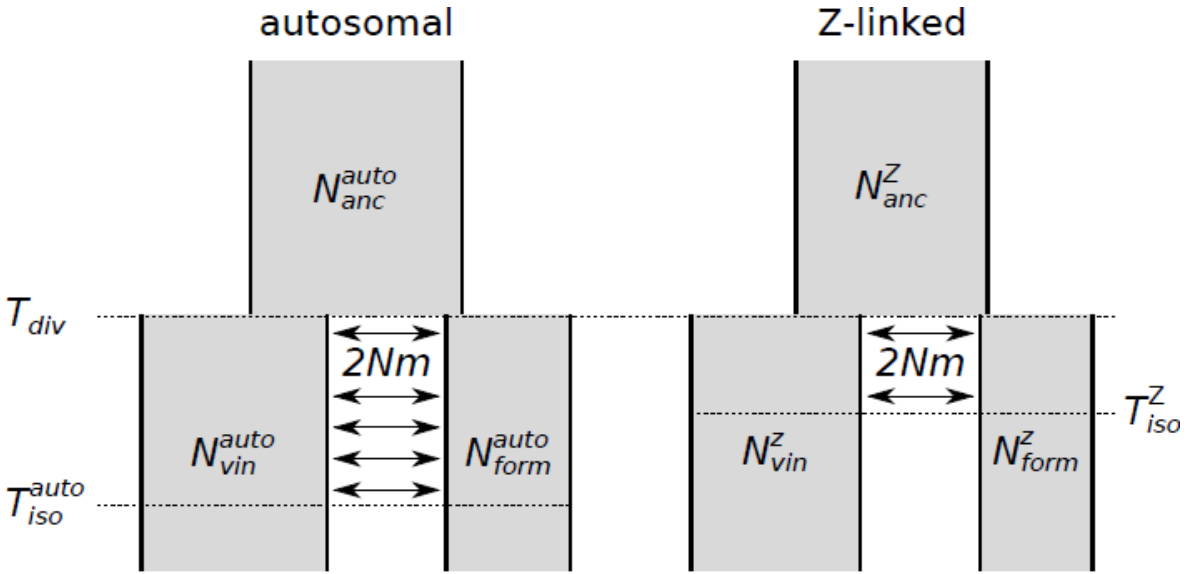
- **Prediction**
  If selection against hybrids is a driving force in speciation, gene flow will be interrupted ealier on the Z-chromosome than on autosomes.

Mode of Speciation in Rose Finches

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

Shou-Hsien Li

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

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

Data

Summary statistics

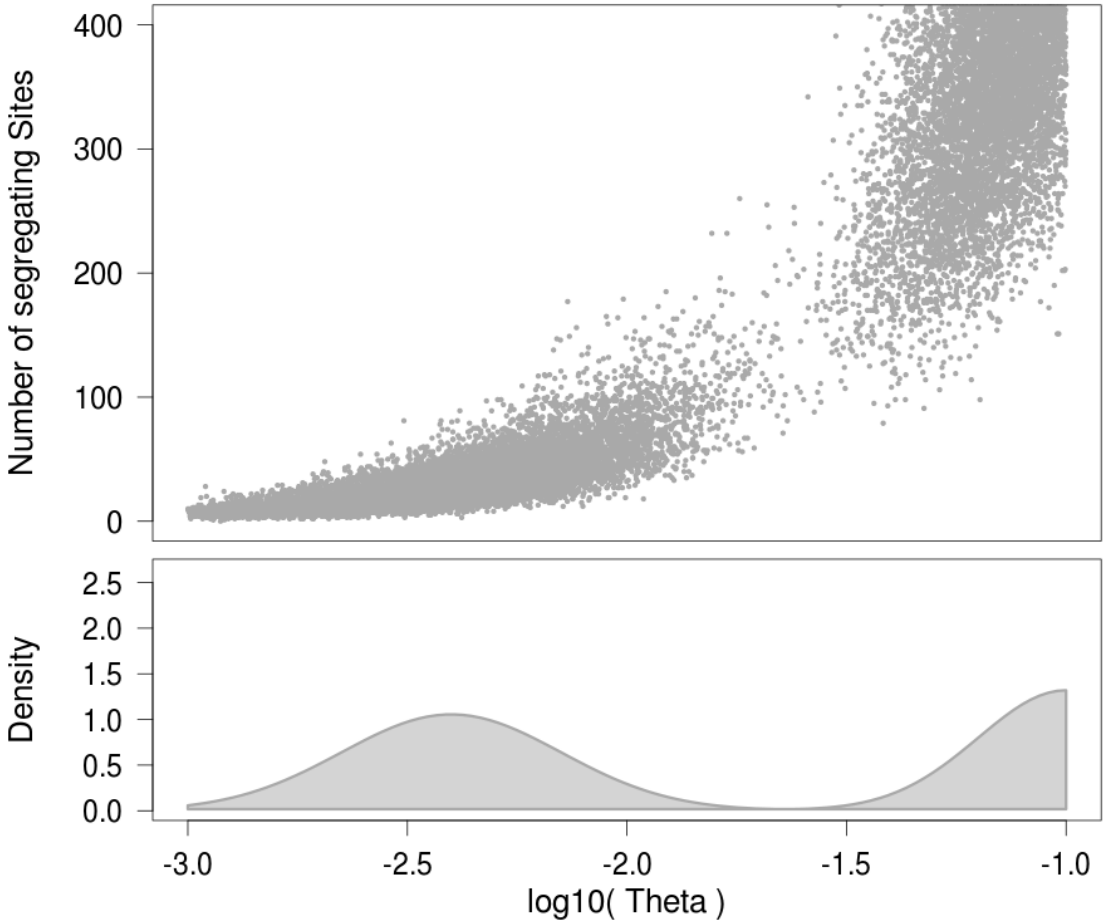
Tavaré *et al.* (1997); Weiss & von Haeseler (1998)

Standard ABC Algorithm

defining statistics

↓

generating simulations according to prior



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

Approximate Bayesian Computation ABC

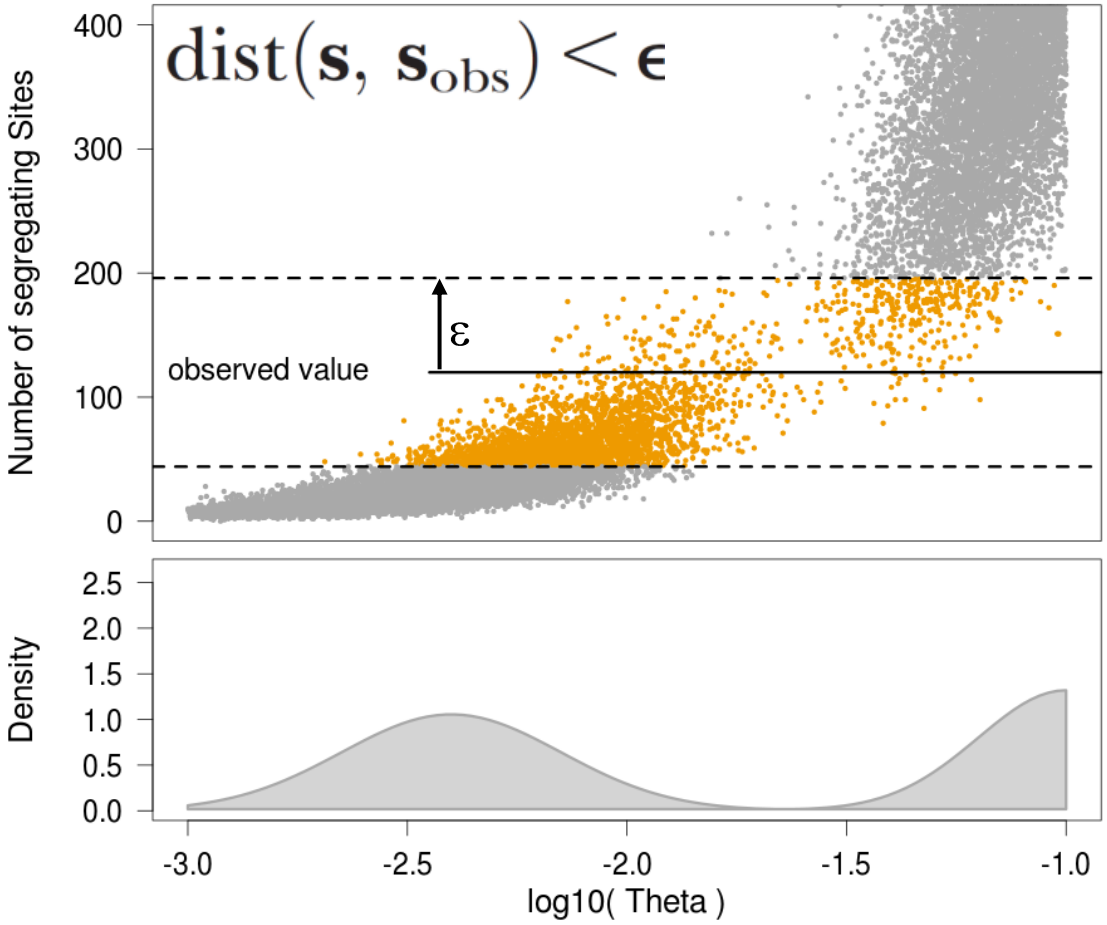
defining statistics

↓

generating simulations according to prior

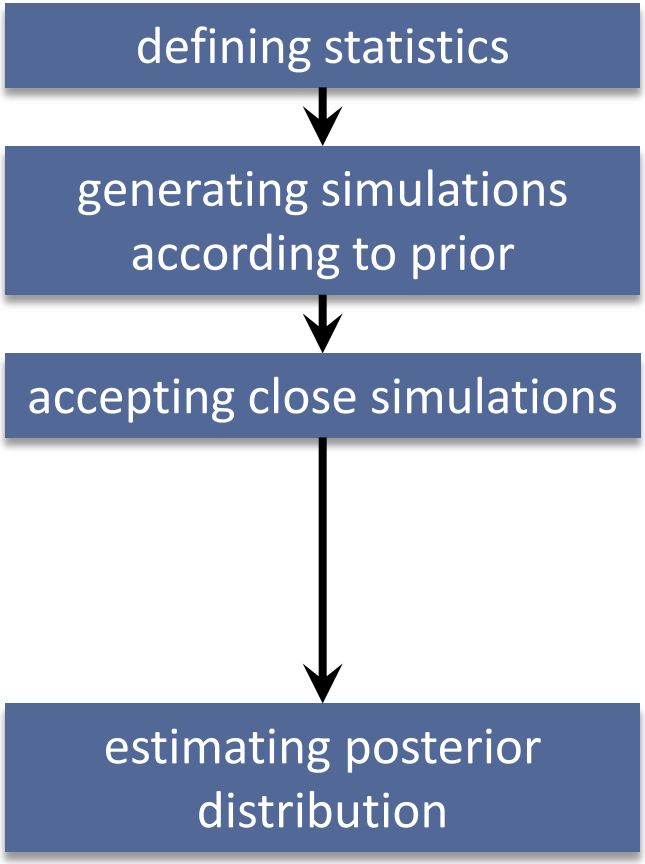
↓

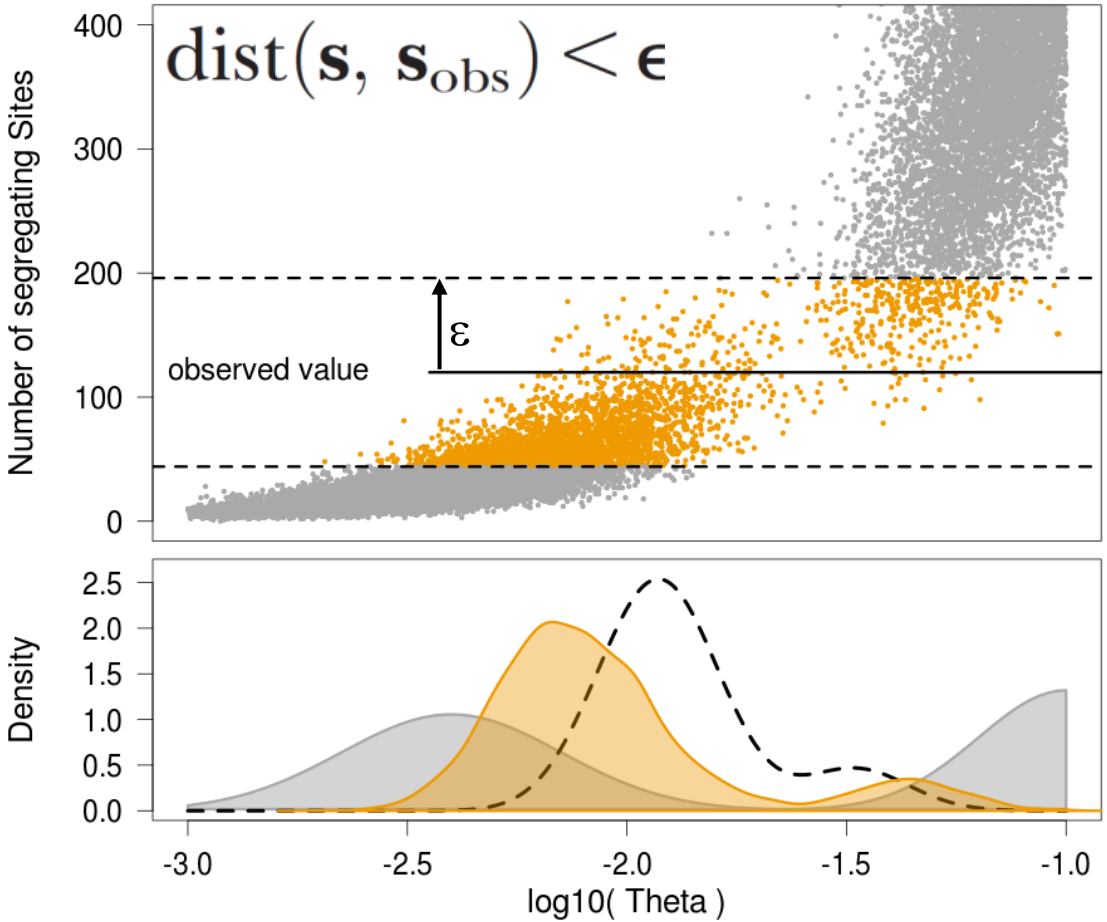
accepting close simulations



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

Approximate Bayesian Computation ABC





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

Approximate Bayesian Computation ABC

defining statistics

↓

generating simulations according to prior

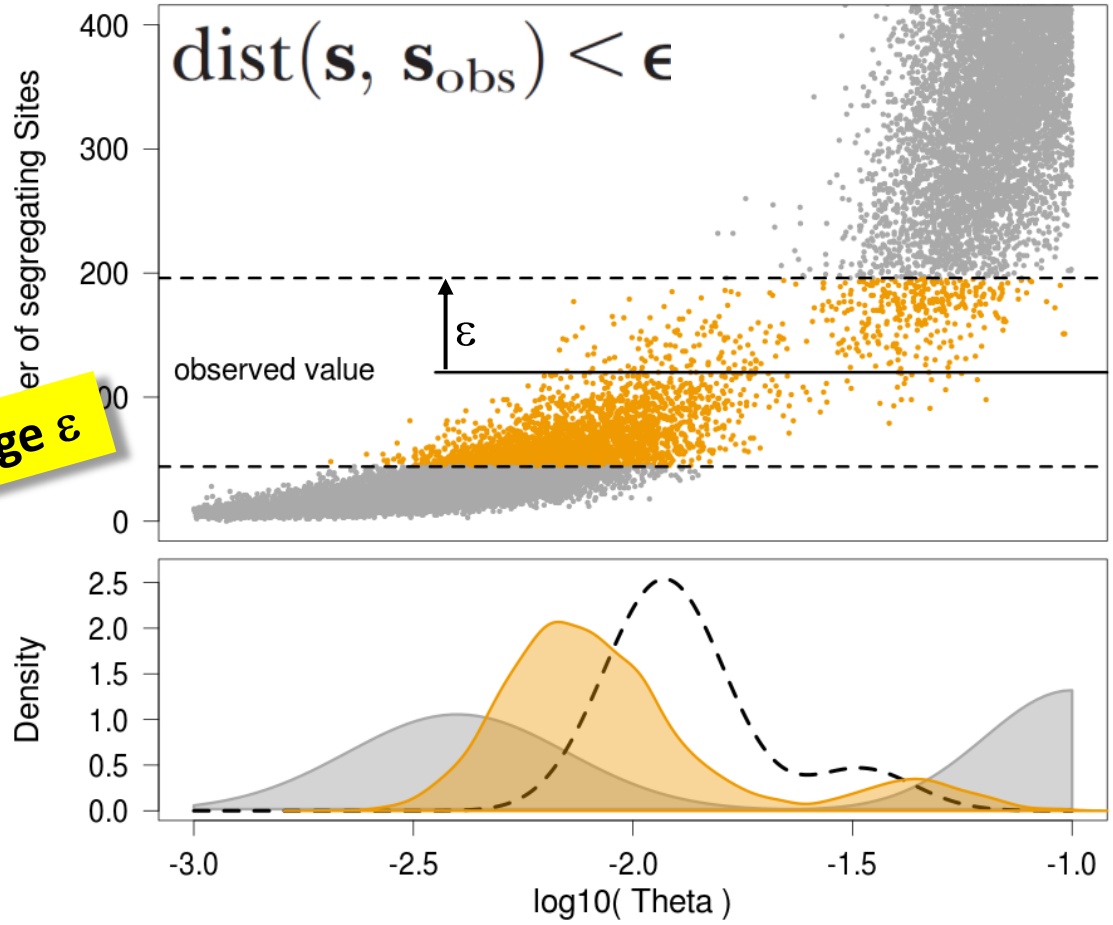
↓

accepting close simulations

↓

estimating posterior distribution

Low absolute likelihoods imply large ε



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

Approximate Bayesian Computation ABC

defining statistics

↓

generating simulations according to prior

↓

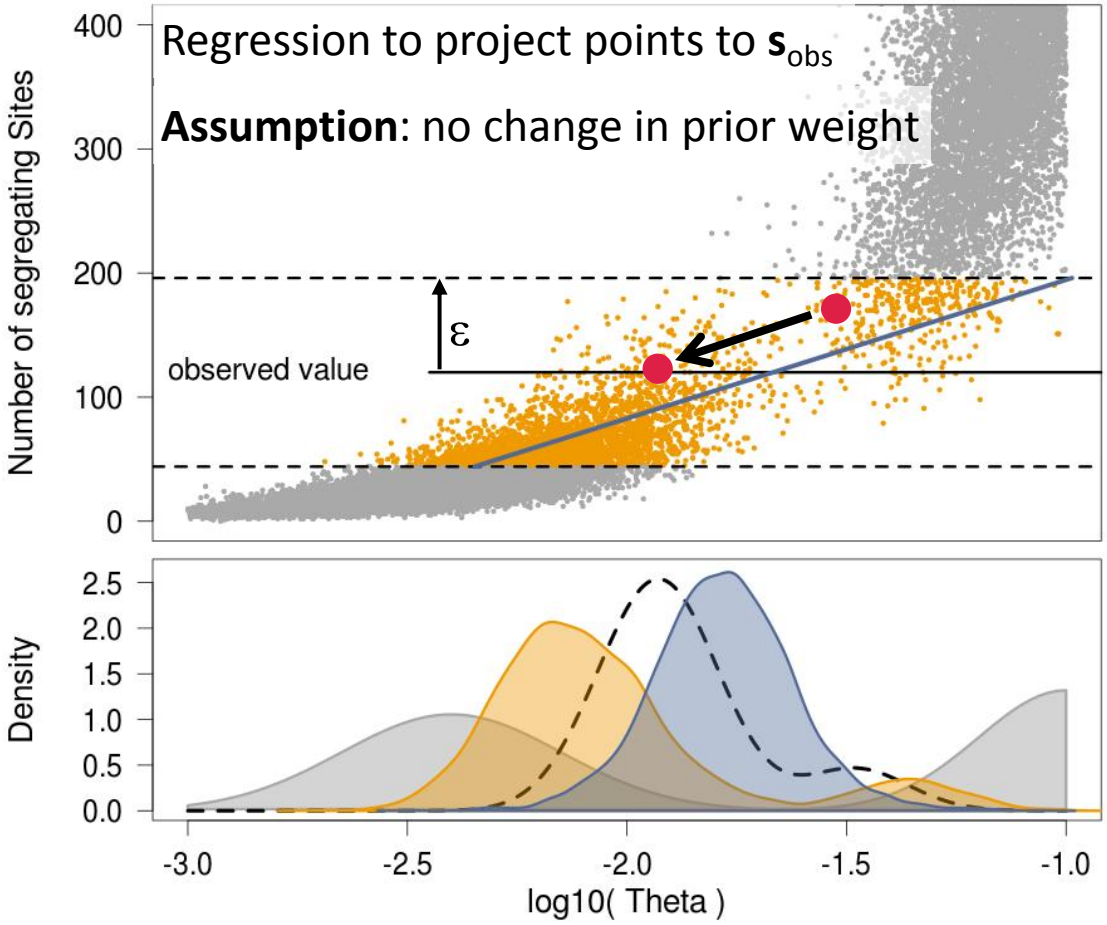
accepting close simulations

↓

post sampling regression adjustment

↓

estimating posterior distribution



Beaumont *et al.* (2002); Blum & François (2009)

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

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

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

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

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

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

Chris Leuenberger

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)

Note: other models could be used, GLM was chosen due to laziness...

ABC-GLM

defining statistics

↓

generating simulations according to prior

↓

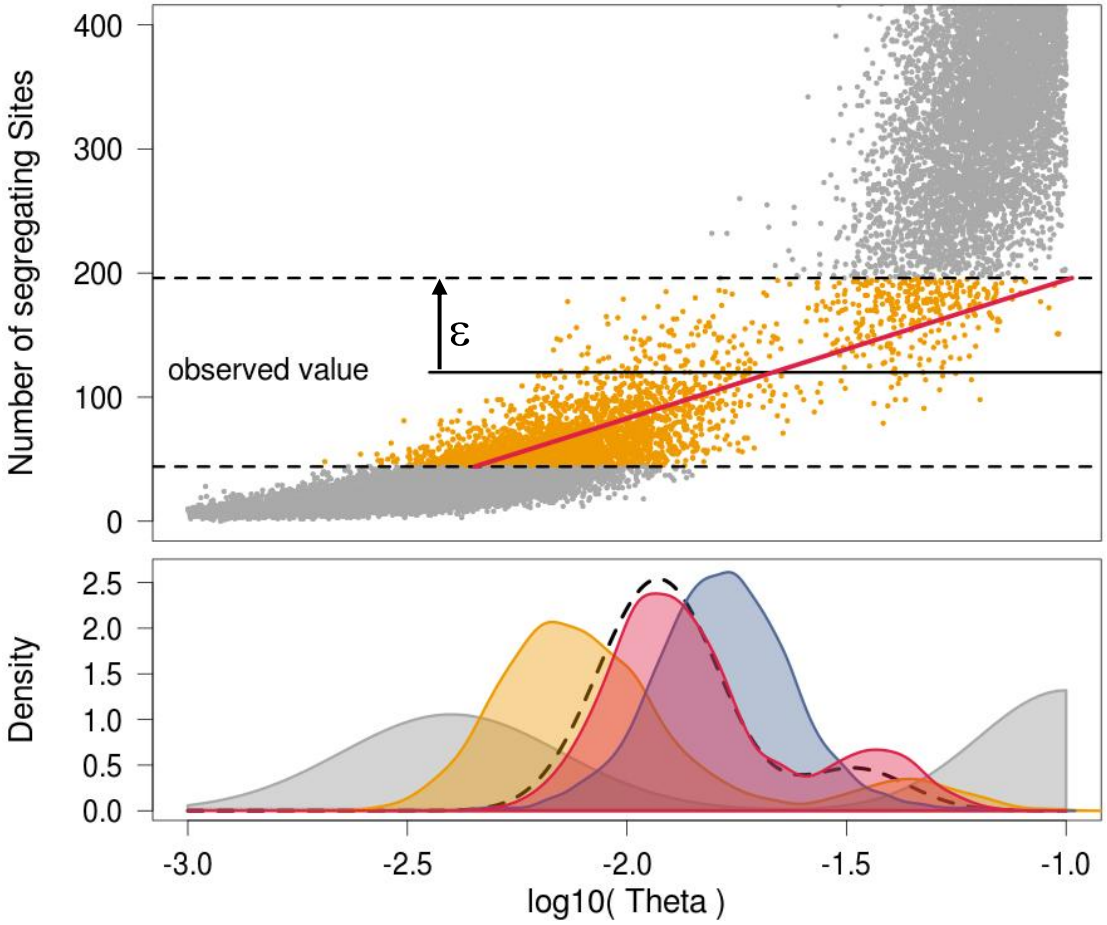
accepting close simulations

↓

fitting a simple likelihood model

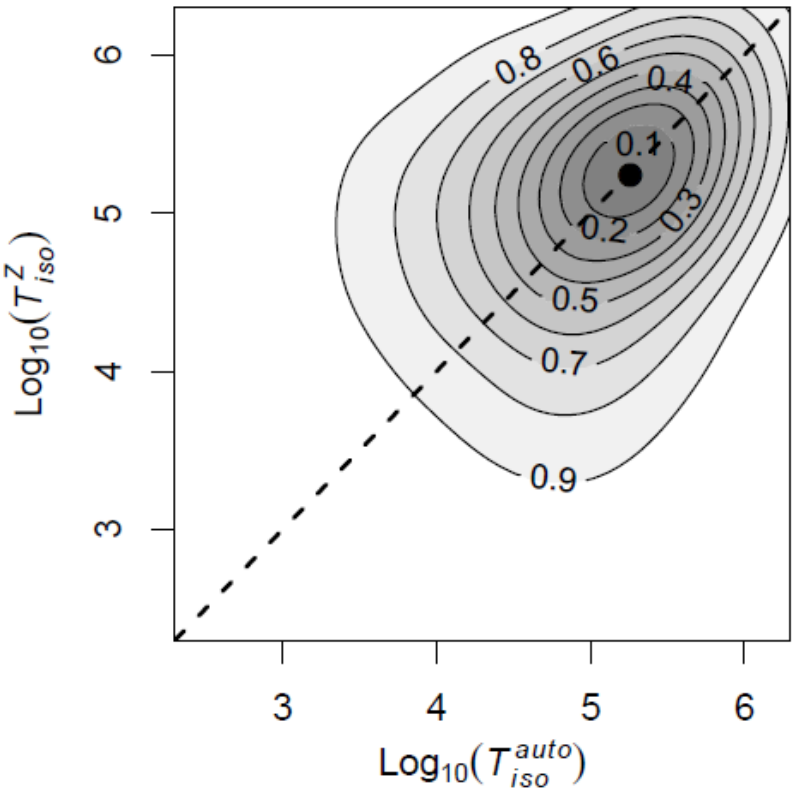
↓

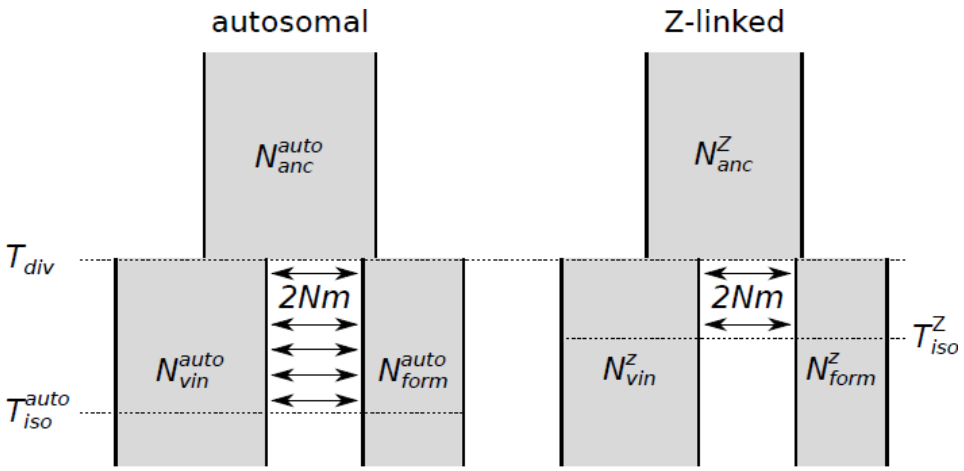
estimating posterior distribution



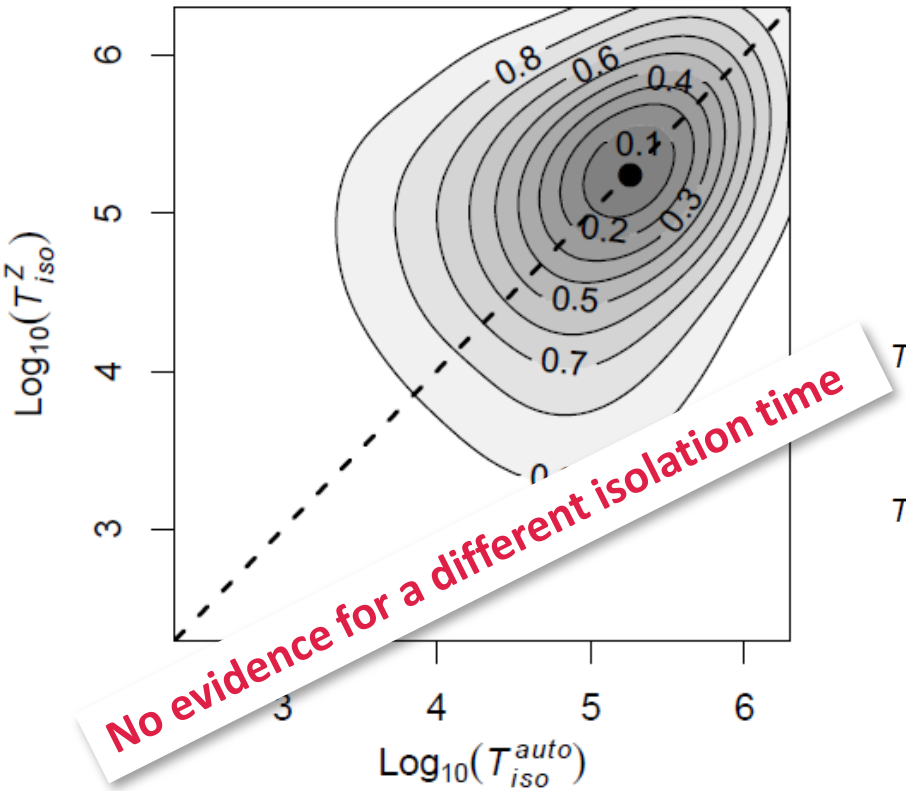
Leuenberger & Wegmann (2010)

Mode of Speciation in Rose Finches

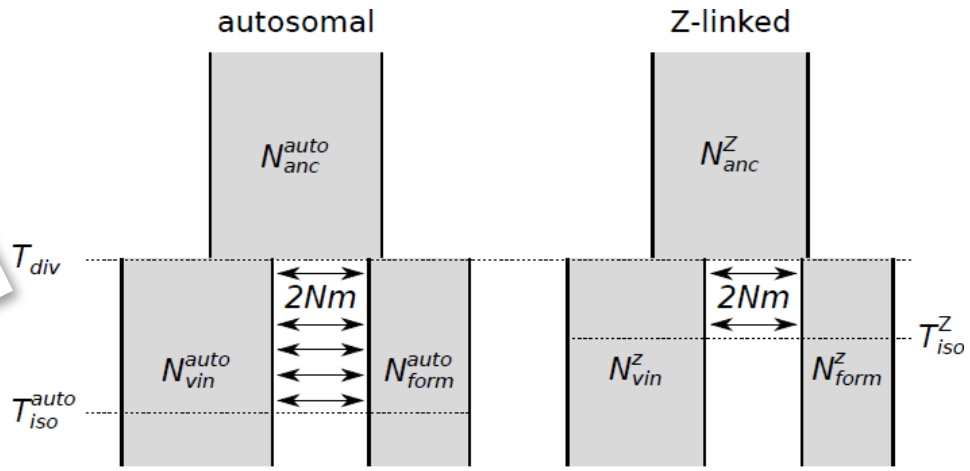




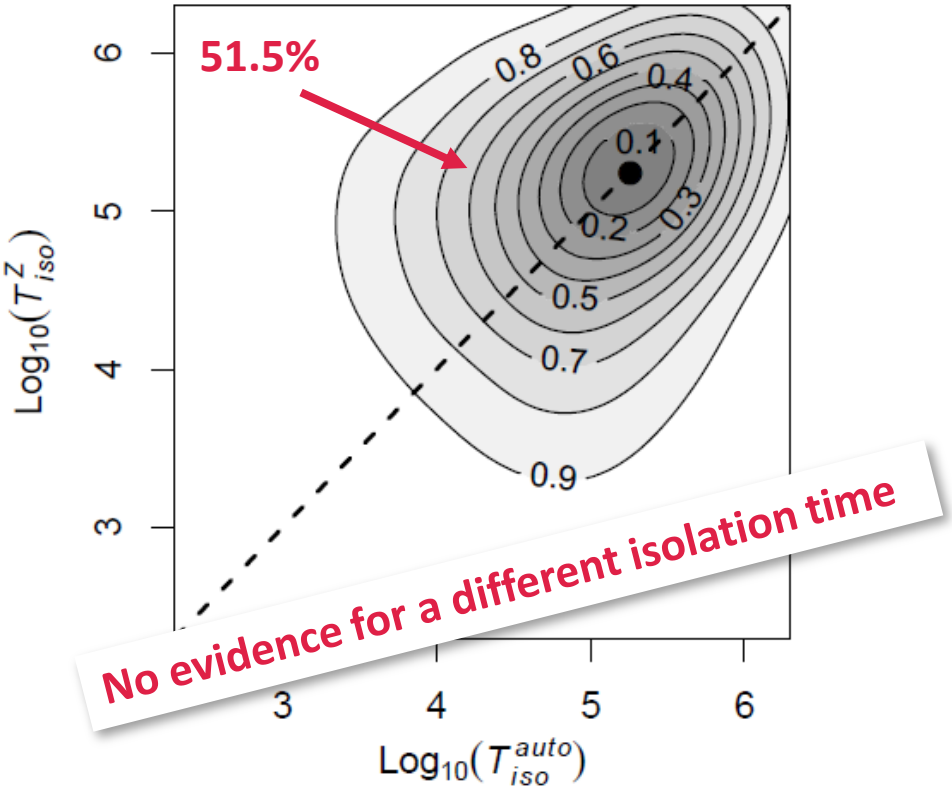
Mode of Speciation in Rose Finches



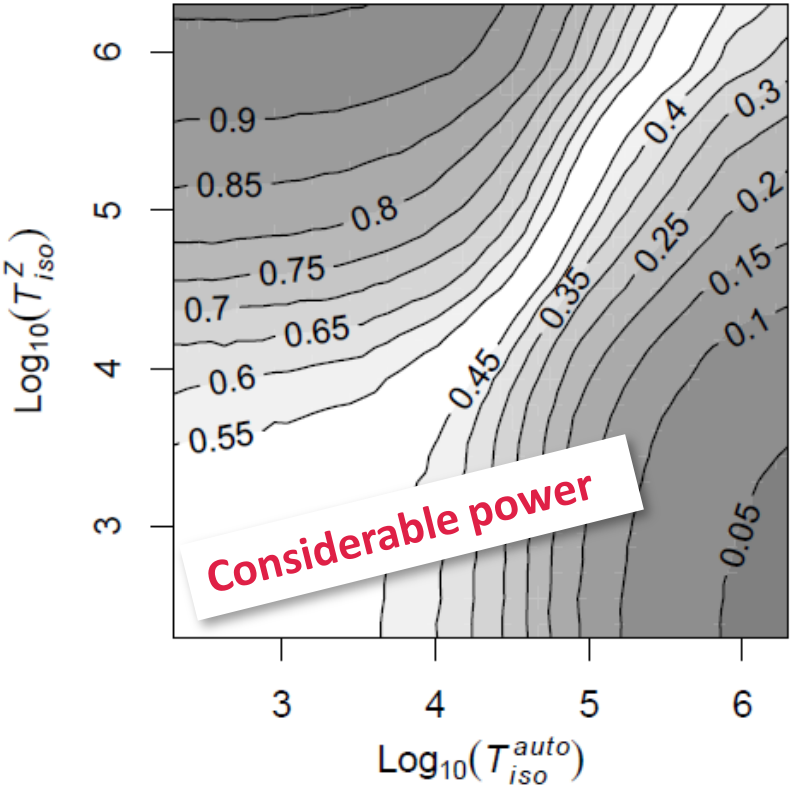
No evidence for a different isolation time



Mode of Speciation in Rose Finches



Joint posterior asymmetry observed in simulated data sets



Next Generation Sequencing (NGS)

- **HUGE** amounts of data

Next Generation Sequencing (NGS)

- **HUGE** amounts of data 

- Some **new challenges** for our inference:
  1. High error rates
     - False-positives without filtering - Biases with filtering
  2. Often only few individuals
     - Difficulty in inferring recent events, bias through specific histories
  3. Tight marker spacing
     - Influence of the genomic location (e.g. genic vs non-genic)
     - Linkage = markers are no longer independent

How to a avoid the dirty issues of filtering?

Model errors!

How to a avoid the dirty issues of filtering?

Model errors!

1. Use ANGSD to infer the Site Frequency Spectrum (SFS)
2. Use ngsTools for pop gen statistics
3. Use our tools to do GWAS, infer heterozygoisty, ...
4. Write your own tools!

Estimating heterozygosity from low converge data

Low coverage data == ambiguous genotyping

- A major problem with ancient DNA is the low coverage (<1x) of many samples.
- Clearly, estimating heterozygosity from called genotypes is a bad idea.

Estimating heterozygosity from low converge data

Low coverage data == ambiguous genotyping

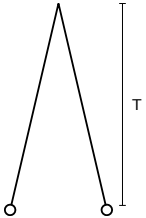
- A major problem with ancient DNA is the low coverage (<1x) of many samples.
- Clearly, estimating heterozygosity from called genotypes is a bad idea.

ANGSD: Analysis of next generation Sequencing Data

- ANGSD implements an algorithm to infer allele frequency distributions (SFS) without calling genotypes.
- BUT: ANGSD requires *a priori* knowledge on the major and minor allele and does not handle *post mortem damage* PDM.

Proposed model

- The goal is to estimate $\theta = 2T\mu$, while integrating over the uncertainty of the genotypes.
- Genotype frequencies shall depend on the unknown base frequencies $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$.



Likelihood

$$L(\theta, \boldsymbol{\pi}) = \mathbb{P}(\boldsymbol{d}|\theta, \boldsymbol{\pi}) = \prod_{i=1}^{I} \sum_{g} \mathbb{P}(d_i|g_i = g)\mathbb{P}(g_i = g|\theta, \boldsymbol{\pi})$$

where g_i denotes the hidden genotype.

Athanasios Kousathanas

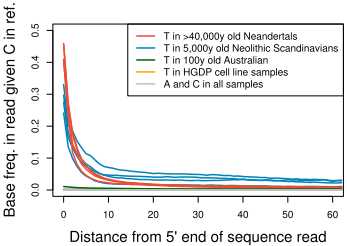
Substitution model

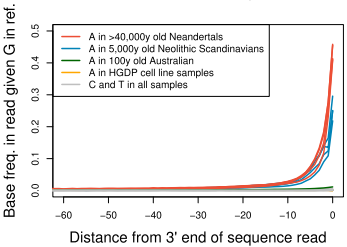
Felsenstein's substitution model (1981)

The probability of observing a specific genotype $g_i = kl$ given base frequencies $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ and the substitution rate θ is given by

$$\mathbb{P}(g_i = kl|\theta, \boldsymbol{\pi}) = \begin{cases} \pi_k q_{kk}(2T) = \pi_k(e^{-\theta} + \pi_k(1 - e^{-\theta})) & \text{if } k = l, \\ \pi_k q_{kl}(2T) = \pi_k \pi_l (1 - e^{-\theta}) & \text{if } k \neq l. \end{cases}$$

Post Mortem Damage

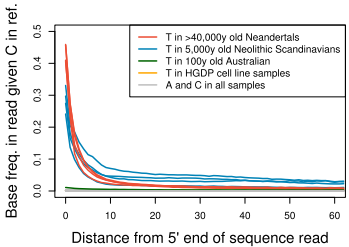


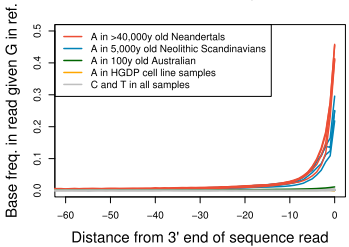


Post Mortem Damage

- Over time, Cytosins are deaminating.
- On Illumina (and 454) platforms, such deaminations result C $\rightarrow$ T (or G $\rightarrow$ A errors after PCR).

Post Mortem Damage





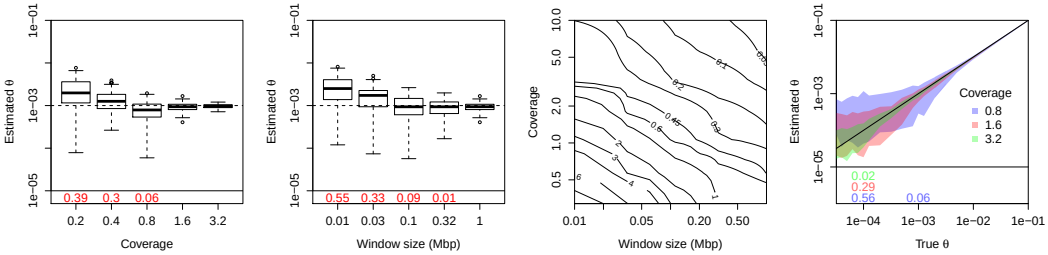
Post Mortem Damage

- Over time, Cytosins are deaminating.
- On Illumina (and 454) platforms, such deaminations result C $\rightarrow$ T (or G $\rightarrow$ A errors after PCR).

Modeling PDM

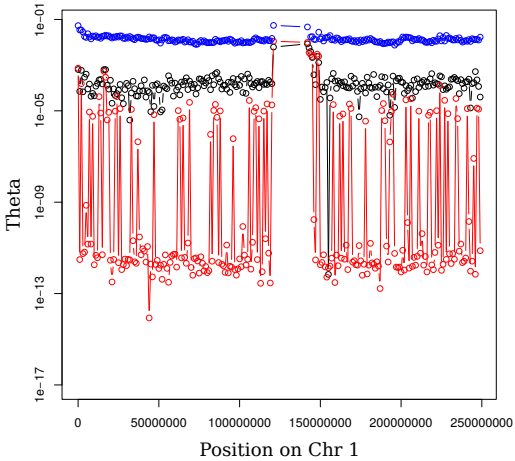
- Following Skoglund *et al.* (2014), we will model the probability of PMD to decay exponentially from the end of the read.

Power analysis through simulations



- Relatively high power to infer θ within a 1Mb window even at very low coverages.
- Higher coverages are required for smaller windows or lower θ values.

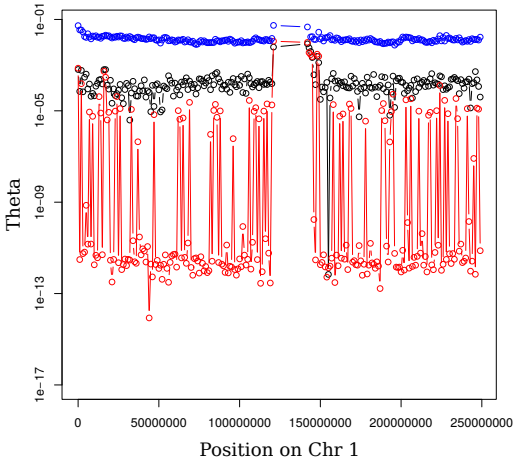
Application to ancient Greek samples



Three Greek samples

- Samples are about 10,000 years old.
- Coverages range from 0.8 to 3.5.
- Expected theta for humans: $\sim 2 \cdot 10^{-3}$

Application to ancient Greek samples



Three Greek samples

- Samples are about 10,000 years old.
- Coverages range from 0.8 to 3.5.
- Expected theta for humans: $\sim 2 \cdot 10^{-3}$

Why are these estimates so different?

Error rate recalibration

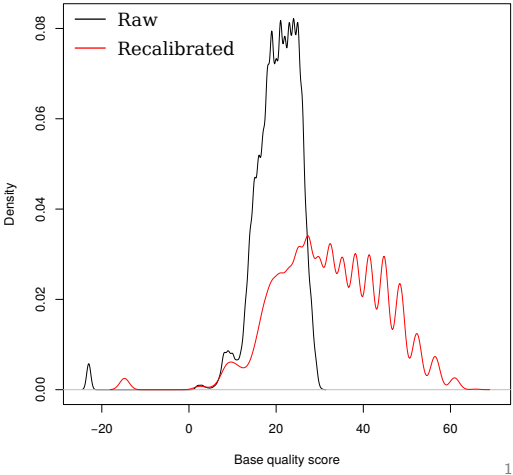
Recalibration using X-linked data

- One sample is male, and hence haploid for all X-linked markers.
- X-linked data is informative about error rate recalibration, as there should be no polymorphisms.
- Adapting the emission probabilities of the model for haploid genotypes is straight forward, which allows us to infer a error-rate recalibration scheme.

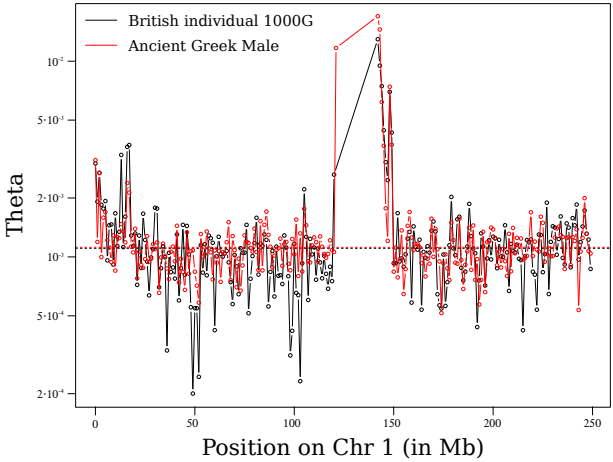
Error rate recalibration

Recalibration using X-linked data

- One sample is male, and hence haploid for all X-linked markers.
- X-linked data is informative about error rate recalibration, as there should be no polymorphisms.
- Adapting the emission probabilities of the model for haploid genotypes is straight forward, which allows us to infer a error-rate recalibration scheme.



It works!



- Very similar estimates from ancient and modern samples!
- Suggesting that error rate recalibration is essential.

Conclusions

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

- As a consequence, we often rely on **approximate inference schemes** ...
  - It may help to replace the full data with **summary statistics**.
  - Approximate Bayesian Computation is an **extremely flexible** but crude approach.

- ... or **approximate models.**
  - Approximating models such that they fit standard inference schemes.

- **Model errors!**
  When properly dealing with genotype uncertainty, there is no need to filter!