UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Demographic inference (from NGS data)
& some other things

Daniel Wegmann
University of Fribourg

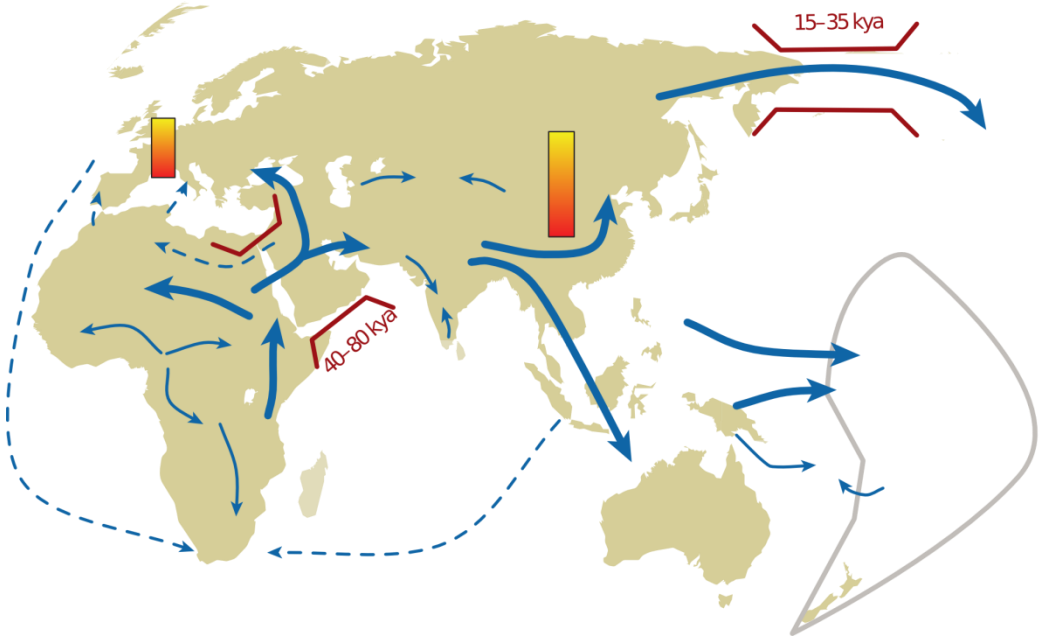
Introduction

- 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.

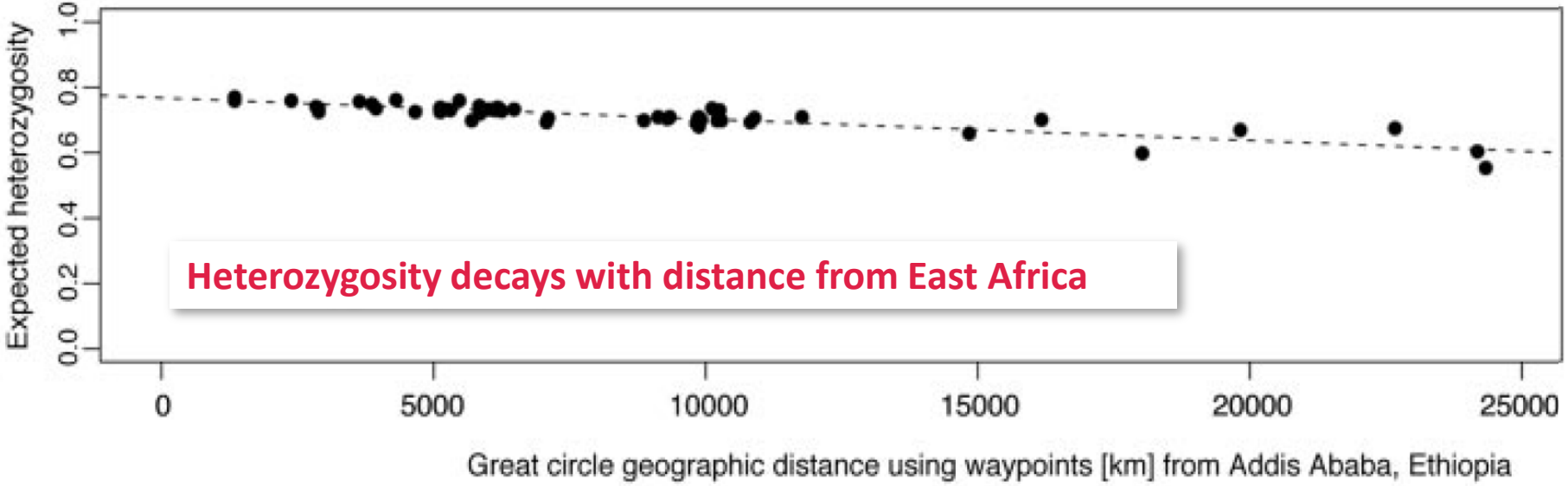
- But this is a **challenging task!**
  - The history of natural populations is usually **complex**.
  - Several evolutionary processes can leave **similar footprints** (bottleneck vs. selection).
  - Loci are not independent, but **correlated realizations** of the same process.



Novembre & Ramachandran (2011)

Qualitative inference

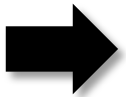
- Traditionally, we have relied on qualitative inference
- **Example**: out of Africa expansion via sequential founder effects in humans.



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**.

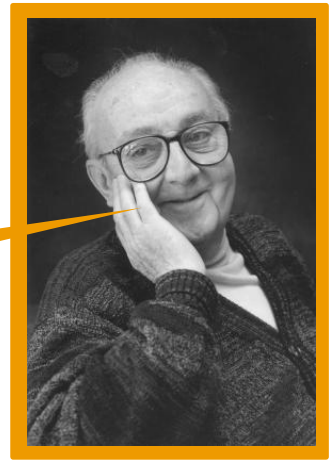


Model-based inference provides **statistical support**

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!

Next Generation Sequencing (NGS) Data

- **HUGE** amounts of data

Next Generation Sequencing (NGS) Data

- **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

Topics of this talk

- **How to avoid the dirty issues of filtering?**
  - Model sequencing errors (e.g. by working with genotype likelihoods)

- **How to increase the number of individuals?**
  - Go low coverage!
  - When accounting for sequencing errors in downstream analysis, coverage can often be very low.

- **How to increase the power to identify selective events?**
  - If possible, include a temporal perspective through multi-generation sampling or ancient DNA

Outline

1. Rejecting a null model using summary statistics
2. Composite-likelihood using coalescent simulations
3. Approximate Bayesian Computation (ABC)
4. Dealing with (ancient) low coverage data
5. Demography & selection from time series data
6. GWAS (in the unlikely case time will permit)

Outline

1. **Rejecting a null model using summary statistics**
2. Composite-likelihood using coalescent simulations
3. Approximate Bayesian Computation (ABC)
4. Dealing with (ancient) low coverage data
5. Demography & selection from time series data
6. GWAS (in the unlikely case time will permit)

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, $\alpha = 0.05$
- 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

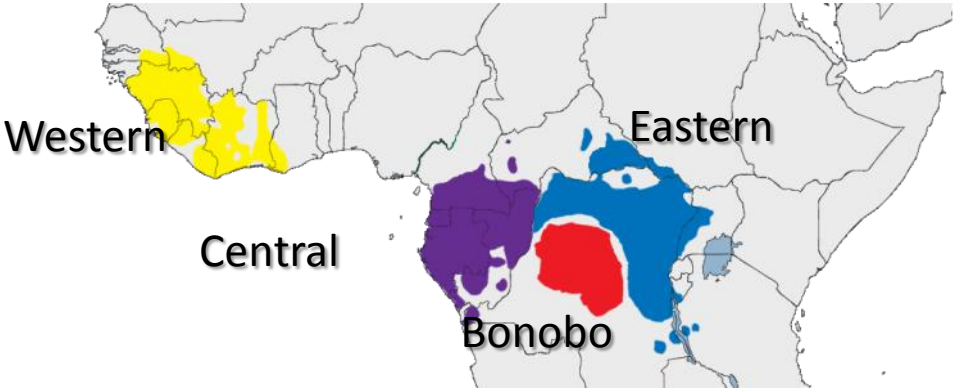


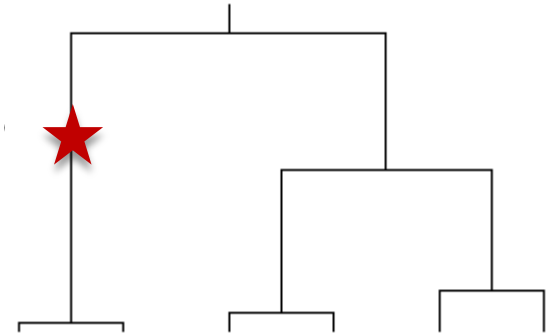
Table 1. Observed Population- and Marker-Specific F_{IS} Values.

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*

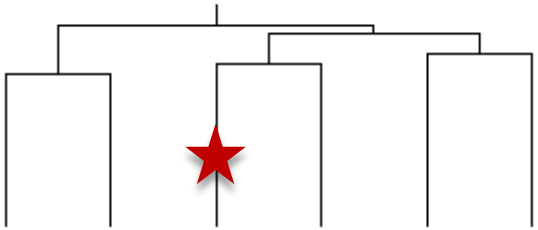
Wegmann & Excoffier, MBE, 2010

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.



Wright-Fisher population



expanding population

Outline

1. Rejecting a null model using summary statistics
2. **Composite-likelihood using coalescent simulations**
3. Approximate Bayesian Computation (ABC)
4. Dealing with (ancient) low coverage data
5. Demography & selection from time series data
6. GWAS (in the unlikely case time will permit)

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

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

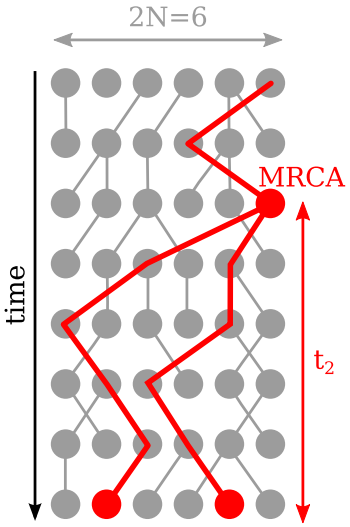
$$P(\mathcal{D}|\Theta, \mu) = \int_G P(\mathcal{D}|G, \mu) P(G|\Theta) dG$$

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.

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

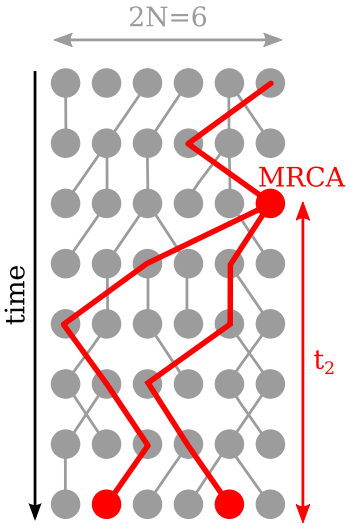


Coalescent theory

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.



Coalescent theory

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

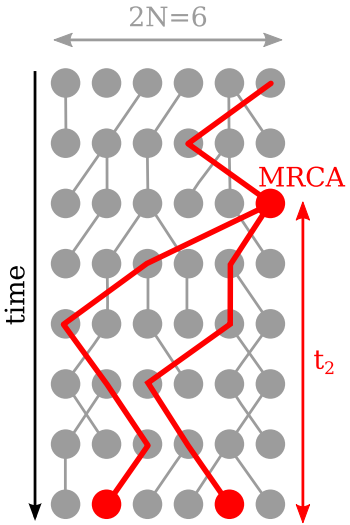
Coalescent event

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

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



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

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

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

Time to MRCA

While the expected time for two lineage to coalesce in $2N$ generations, there is a large variance associated with this expectations.

The probability that two lineages coalesce t generations ago is

$$Pr(t_2 = t) = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}$$

Time to MRCA

While the expected time for two lineage to coalesce in $2N$ generations, there is a large variance associated with this expectations.

The probability that two lineages coalesce t generations ago is

$$Pr(t_2 = t) = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}$$

Using the approximation

$$(1 - x)^t \approx e^{-xt}$$

we can rewrite this as

$$Pr(t_2 = t) = e^{-\frac{1}{2N}(t-1)} \frac{1}{2N}$$

Time to MRCA

While the expected time for two lineage to coalesce in $2N$ generations, there is a large variance associated with this expectations.

The probability that two lineages coalesce t generations ago is

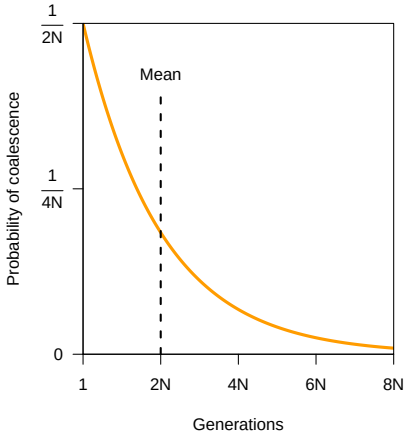
$$Pr(t_2 = t) = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}$$

Using the approximation

$$(1 - x)^t \approx e^{-xt}$$

we can rewrite this as

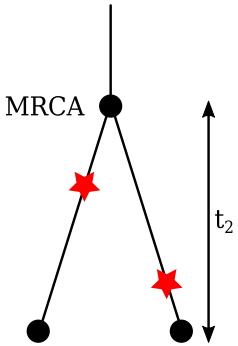
$$Pr(t_2 = t) = e^{-\frac{1}{2N}(t-1)} \frac{1}{2N}$$



Expected number of mutations between two samples

If the mutation rate per generation per lineage is μ, the expected number of mutational differences d_{ij} between two samples i and j is thus

$$E[d_{ij}] = 2 \cdot E[t_2] \cdot \mu = 4N\mu$$

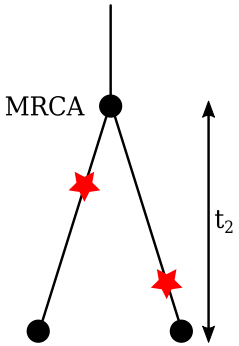


Expected number of mutations between two samples

If the mutation rate per generation per lineage is μ, the expected number of mutational differences d_{ij} between two samples i and j is thus

$$E[d_{ij}] = 2 \cdot E[t_2] \cdot \mu = 4N\mu$$

Note: the population size N and the mutation rate μ have the same effect on genetic diversity!

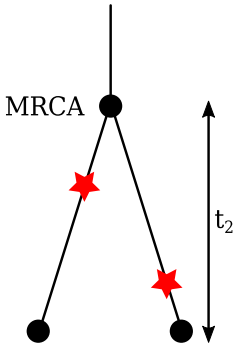


Expected number of mutations between two samples

If the mutation rate per generation per lineage is μ, the expected number of mutational differences d_{ij} between two samples i and j is thus

$$E[d_{ij}] = 2 \cdot E[t_2] \cdot \mu = 4N\mu$$

Note: the population size N and the mutation rate μ have the same effect on genetic diversity!



Definition: θ

$\theta := 4N\mu$ is an important population genetics parameter characterizing the expected genetic diversity of a population.

Infinite Sites Model (ISM)

Assumes that each mutation hits a different base pair. This is realistic if $\theta \ll 1$.

The Tajima estimator of θ under ISM

Since we expect $\theta = 4N\mu$ mutations between two samples, the average number of differences π between individuals is an estimate of θ.

$$\hat{\theta}_T = \pi = \frac{1}{n(n-1)/2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij},$$

where $n(n-1)/2$ is the number of pairs among n samples.

Infinite Sites Model (ISM)

Assumes that each mutation hits a different base pair. This is realistic if $\theta \ll 1$.

The Tajima estimator of θ under ISM

Since we expect $\theta = 4N\mu$ mutations between two samples, the average number of differences π between individuals is an estimate of θ.

$$\hat{\theta}_T = \pi = \frac{1}{n(n-1)/2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij},$$

where $n(n-1)/2$ is the number of pairs among n samples.

Note: While $E[\pi] = \theta$, any particular value of π will unlikely be exactely θ!

Infinite Sites Model (ISM)

Assumes that each mutation hits a different base pair. This is realistic if $\theta \ll 1$.

The Tajima estimator of θ under ISM

Since we expect $\theta = 4N\mu$ mutations between two samples, the average number of differences π between individuals is an estimate of θ.

$$\hat{\theta}_T = \pi = \frac{1}{n(n-1)/2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij},$$

where $n(n-1)/2$ is the number of pairs among n samples.

Note: While $E[\pi] = \theta$, any particular value of π will unlikely be exactely θ!

This is why we call π and *estimator* of θ and write it as $\hat{\theta}$ or $\hat{\theta}_T$ to indicate that it is the *Tajima* estimator.

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 $\binom{k}{2}$.

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

Probability of coalescent

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

Intuitive explanation

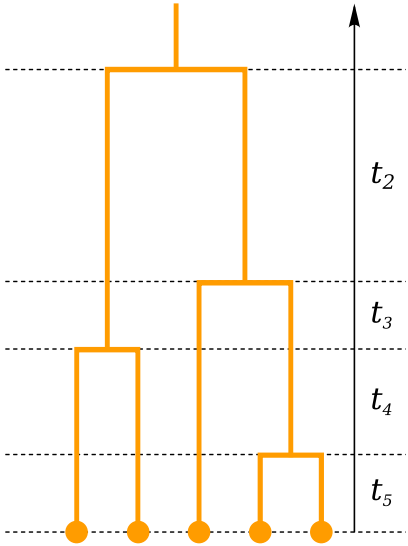
Probability of coalescence among k lineages = probability of coalescence among two lineages $\frac{1}{2N}$ times the number of possible pairs $\binom{k}{2}$.

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)



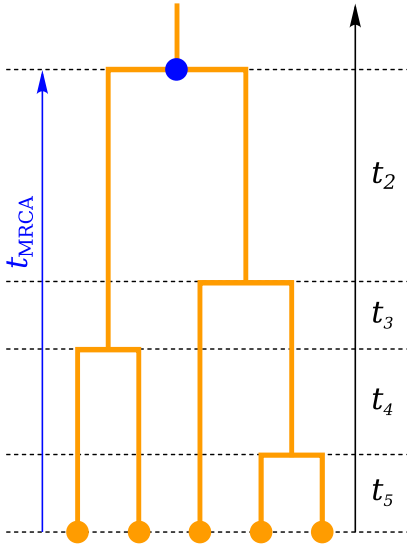
$$\mathbb{E}[t_k] = \frac{1}{\binom{k}{2}\frac{1}{2N}} = \frac{2N}{\binom{k}{2}} = \frac{4N}{k(k-1)}$$

$$\mathbb{E}[t_2] = \frac{2N}{\binom{2}{2}} = 2N$$

$$\mathbb{E}[t_3] = \frac{2N}{\binom{3}{2}} = \frac{2N}{3}$$

$$\mathbb{E}[t_4] = \frac{2N}{\binom{4}{2}} = \frac{2N}{6}$$

$$\mathbb{E}[t_5] = \frac{2N}{\binom{5}{2}} = \frac{2N}{10}$$



Expected total tree height $E[t_{MRCA}]$

$$\mathbb{E}[t_{MRCA}] = \sum_{k=2}^{n} \mathbb{E}[t_k] = \sum_{k=2}^{n} \frac{4N}{k(k-1)} = 4N \sum_{k=2}^{n} \frac{1}{k(k-1)}$$

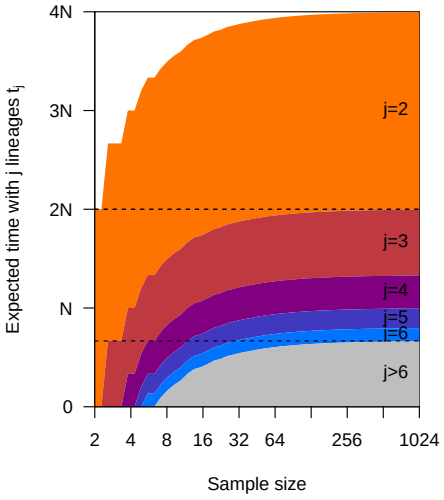
$$= 4N \sum_{k=2}^{n} \frac{k-k+1}{k(k-1)}$$

$$= 4N \left(\sum_{k=2}^{n} \frac{k}{k(k-1)} - \sum_{k=2}^{n} \frac{k-1}{k(k-1)} \right)$$

$$= 4N \left(\sum_{k=2}^{n} \frac{1}{k-1} - \sum_{k=2}^{n} \frac{1}{k} \right)$$

$$= 4N \left(1 - \frac{1}{n} \right)$$

There are only few lineages for most of the genealogy!



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

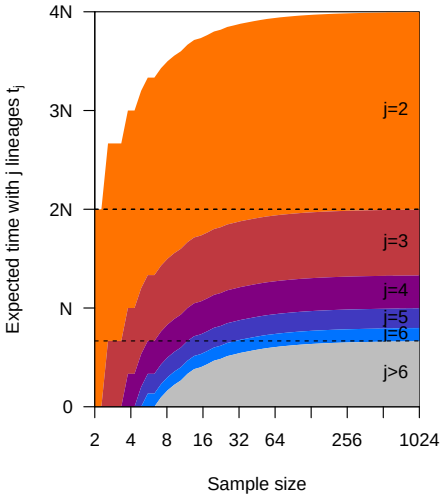
$$\mathbb{E}[t_{MRCA}] = 4N\left(1 - \frac{1}{n}\right)$$

$$\frac{\mathbb{E}[t_k]}{\mathbb{E}[t_{MRCA}]} = \frac{1}{k(k-1)} \frac{1}{1-\frac{1}{n}}$$

When n is large:

$$\frac{\mathbb{E}[t_k]}{\mathbb{E}[t_{MRCA}]} \approx \frac{1}{k(k-1)}$$

There are only few lineages for most of the genealogy!



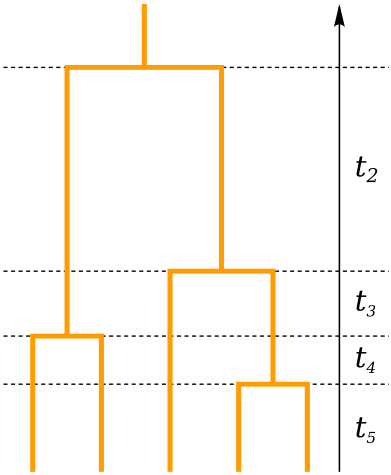
$$\mathbb{E}[t_k] = \frac{2N}{\binom{k}{2}} = \frac{4N}{k(k-1)}$$

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

$$\frac{\mathbb{E}[t_k]}{\mathbb{E}[t_{MRCA}]} = \frac{1}{k(k-1)} \frac{1}{1-\frac{1}{n}}$$

When n is large:

$$\frac{\mathbb{E}[t_k]}{\mathbb{E}[t_{MRCA}]} \approx \frac{1}{k(k-1)}$$



Expected total length L_n of the genealogy

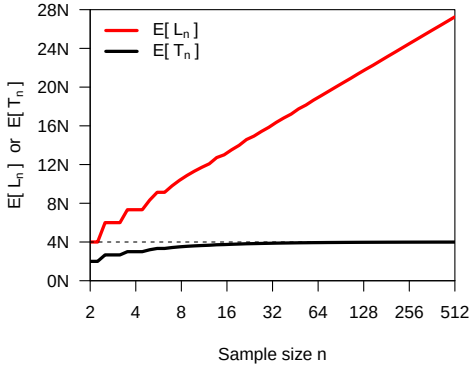
The expected total length of a genealogy of n samples, L_n is the sum of the expected epoch length times the number of lineages during that epoch.

$$\begin{aligned}\mathbb{E}[L_n] &= \sum_{k=2}^{n} \mathbb{E}[t_k]k = \sum_{k=2}^{n} \frac{4N}{k(k-1)}k \\ &= 4N \sum_{k=2}^{n} \frac{1}{k-1} = 4N \sum_{k=1}^{n-1} \frac{1}{k}\end{aligned}$$

Height versus length of a genealogy of n samples

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

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



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

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

Expected number of mutations

The expected number of mutations on a genealogy is given by the product of the mutation rate μ and the expected length of the genealogy $\mathbb{E}[L_n]$:

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

Expected number of mutations

The expected number of mutations on a genealogy is given by the product of the mutation rate μ and the expected length of the genealogy $\mathbb{E}[L_n]$:

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

Infinite Sites Model (ISM)

Assumes that each mutation hits a different base pair. This is realistic if $\theta \ll 1$.

Expected number of mutations

The expected number of mutations on a genealogy is given by the product of the mutation rate μ and the expected length of the genealogy $\mathbb{E}[L_n]$:

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

Infinite Sites Model (ISM)

Assumes that each mutation hits a different base pair. This is realistic if $\theta \ll 1$.

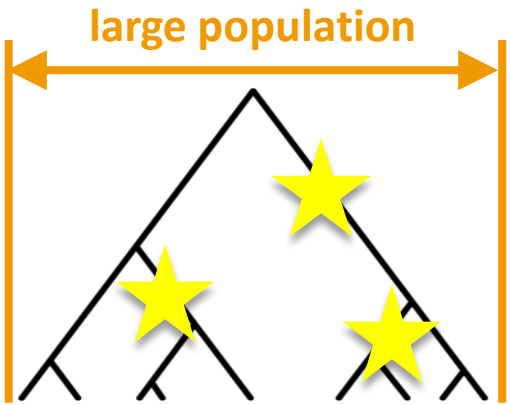
The Watterson estimator of θ under ISM

Under ISM, the number of mutations $=$ to the number of polymorphic (called segregating) sites S among samples. S can thus be used to estimate θ:

$$\mathbb{E}[S] = \theta \sum_{k=1}^{n-1} \frac{1}{k} \qquad \Rightarrow \qquad \hat{\theta}_W = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

Joint inference of demography and mutation rates

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



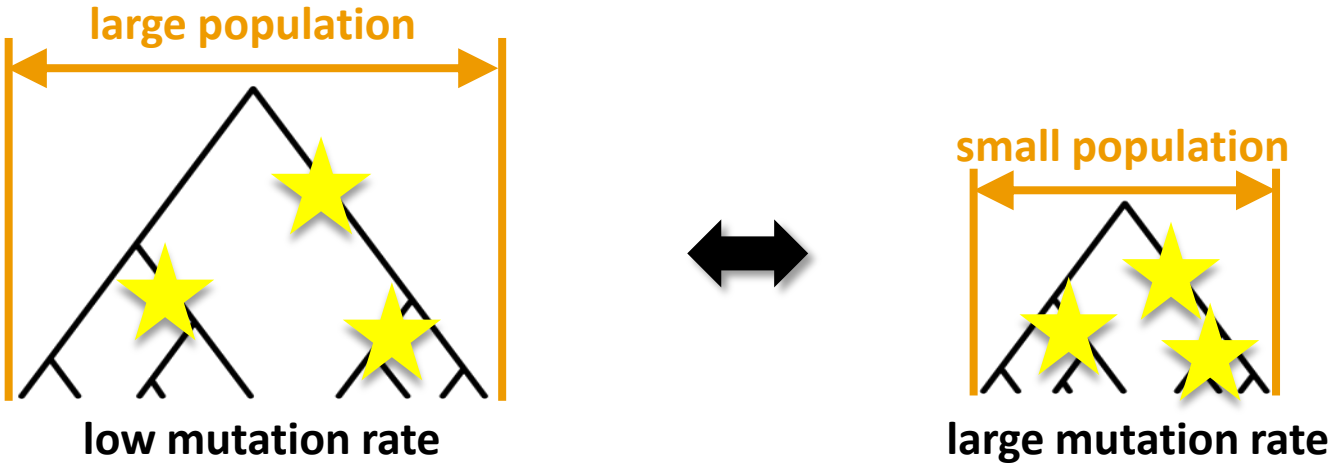
low mutation rate

large mutation rate

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:

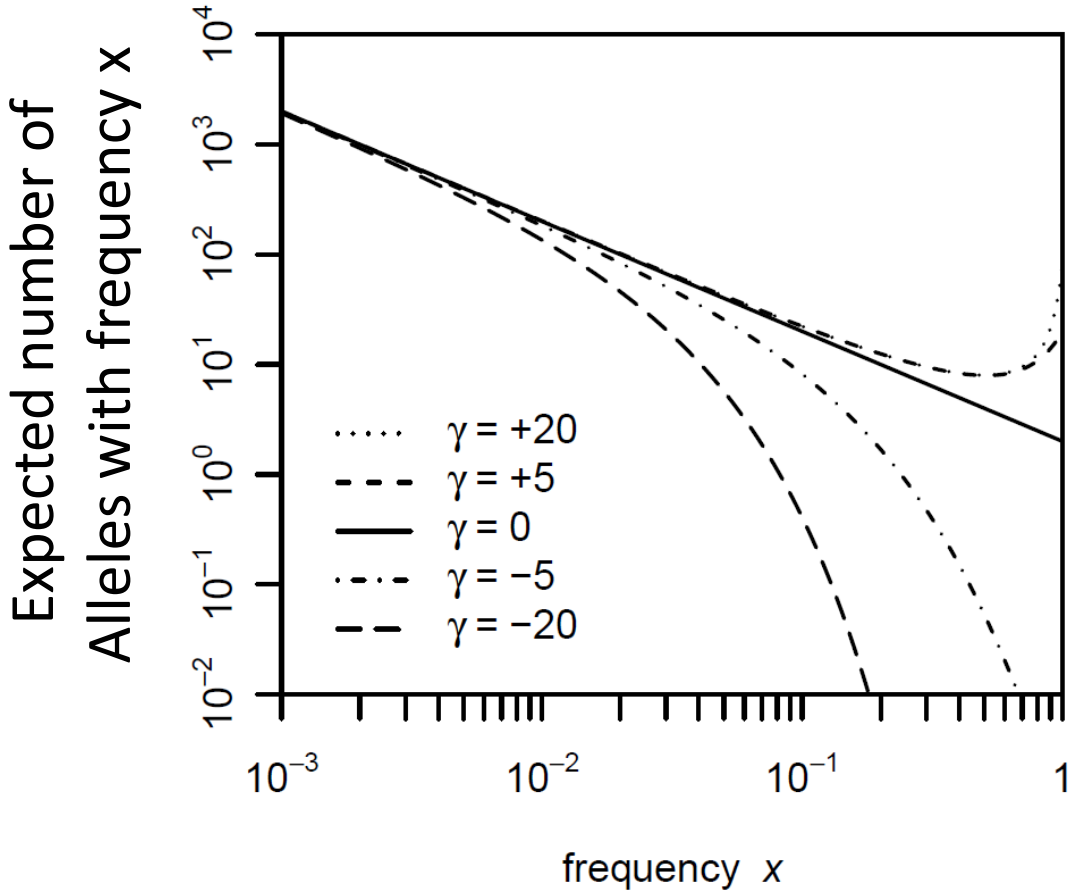
- 202 known or prospective drug target genes
- 14,002 individuals, of which 12,514 Europeans
- Median coverage of 27x and a call rate of 90.7%

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

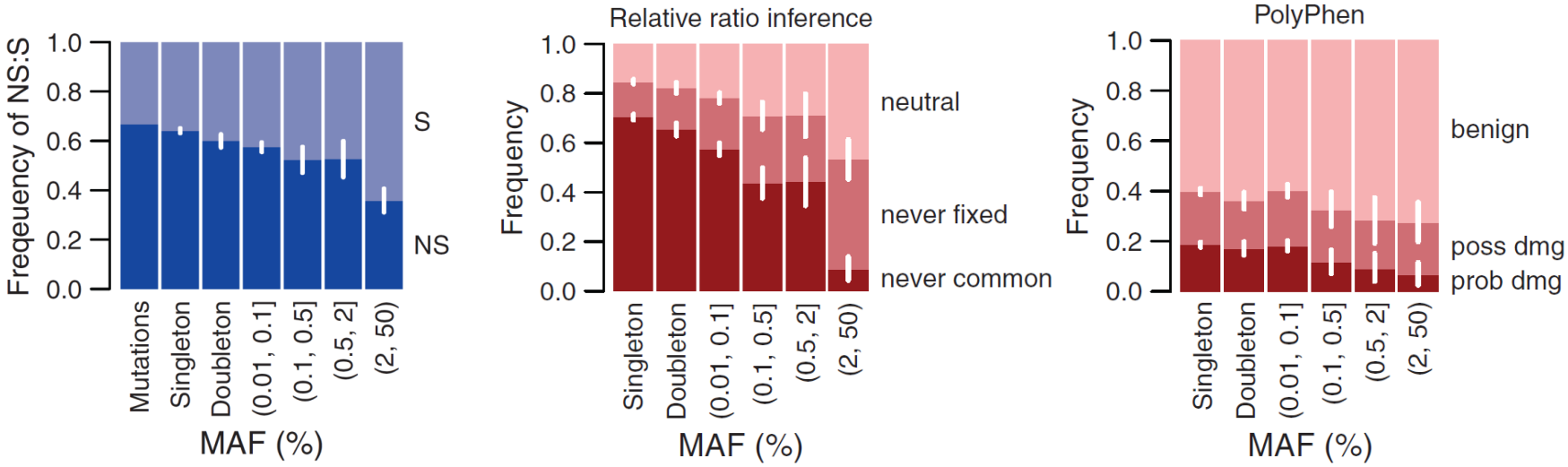
Neutral alleles

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**

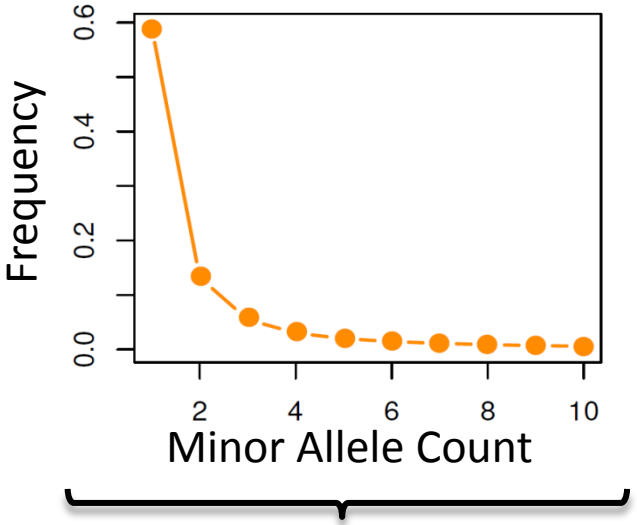
$$P(\mathbf{D} \mid \mu, N)$$

Polymorphisms — **Mutation rates & Population sizes**

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

Joint inference of demography and mutation rates

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



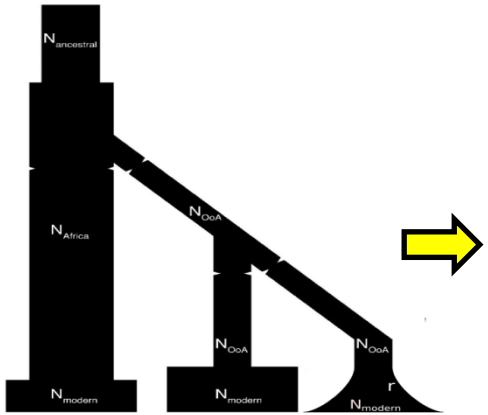
22,000 Sequences of 202 genes

Site Frequency Spectrum SFS

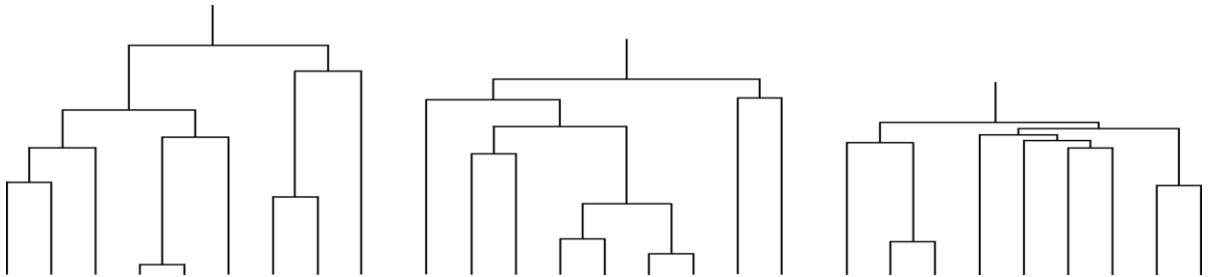
Joint inference of demography and mutation rates

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

- Simulate genealogies with fixed parameter values



Africa Asia Europe



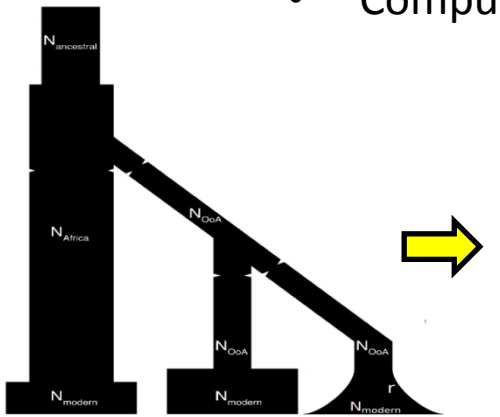
- Exponential growth in Europe
- All other parameters fixed to Schaffner estimates

Nielsen 2000; Coventry et al. 2010

Joint inference of demography and mutation rates

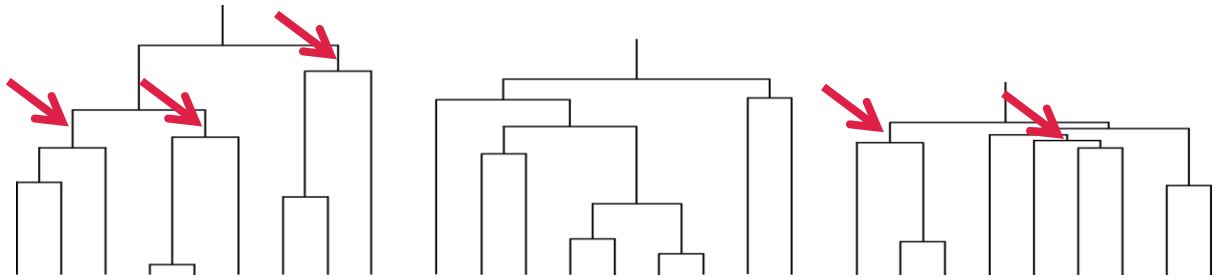
Using **Monte Carlo simulations** to approximate *P*(**SFS**|**μ,N**):

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



Africa Asia Europe

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



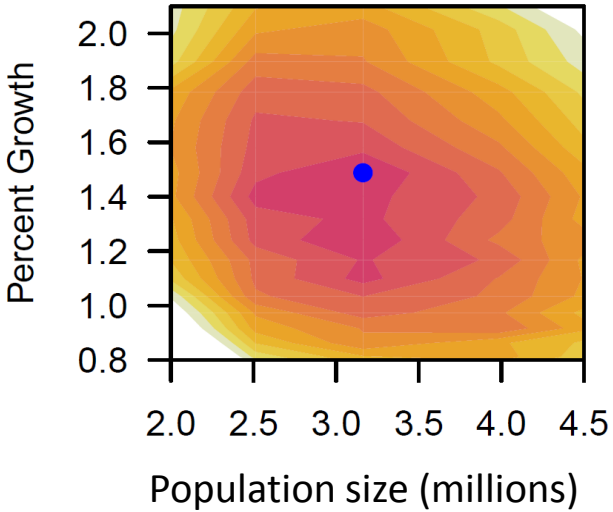
Likelihood 1 Likelihood 2 Likelihood 3

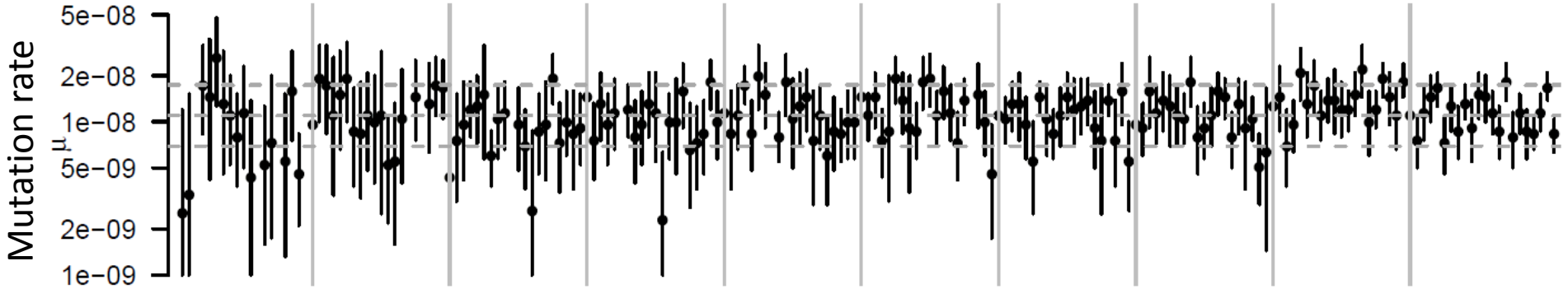
Average Likelihood

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





Outline

1. Rejecting a null model using summary statistics
2. Composite-likelihood using coalescent simulations
3. **Approximate Bayesian Computation (ABC)**
4. Dealing with (ancient) low coverage data
5. Demography & selection from time series data
6. GWAS (in the unlikely case time will permit)

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.

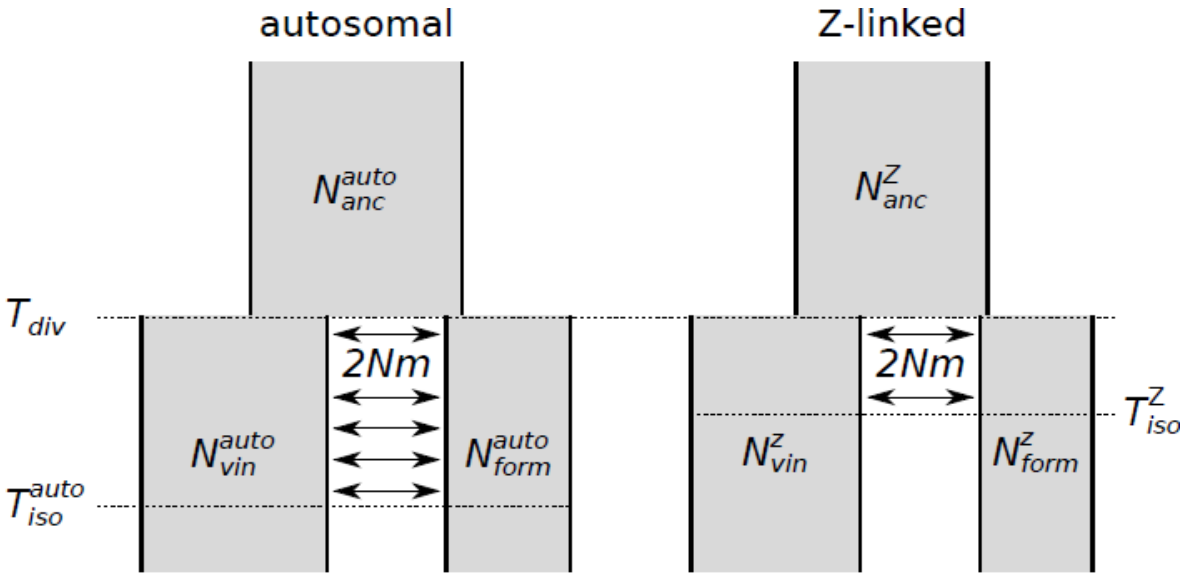
Mode of Speciation in Rose Finches

- 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 $P(\mathbf{D}|\theta)$ is proportional to $P(\mathbf{S}|\theta)$
  2) Using **simulations** to approximate the likelihood function $P(\mathbf{S}|\theta)$

- Apply in a Bayesian setting:

$$\underbrace{P(\theta \mid \mathbf{D})}_{\text{Posterior}} \propto \underbrace{P(\mathbf{D} \mid \theta)}_{\text{Likelihood}} \underbrace{P(\theta)}_{\text{Prior}}$$

➡ Approximate Bayesian Computation (ABC)

Approximate Bayesian Computation ABC

defining statistics

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

Data

Summary statistics

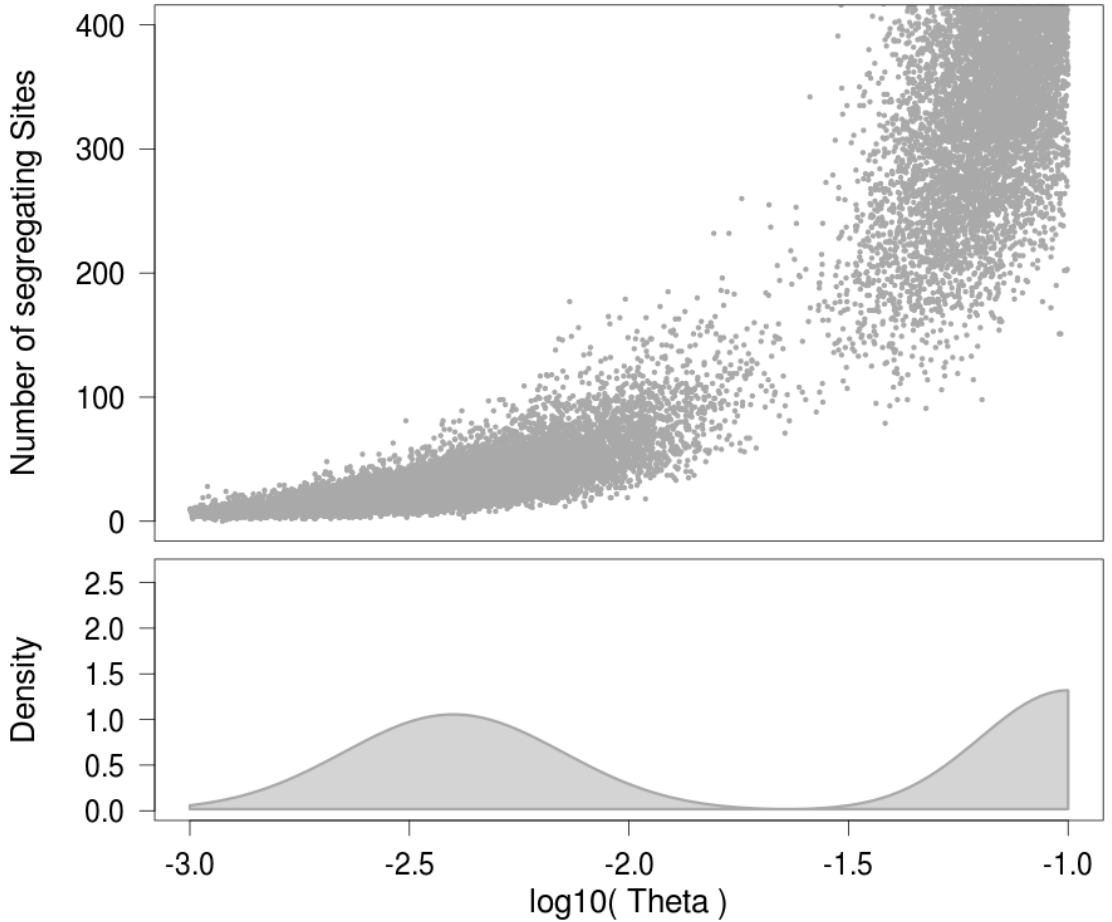
Tavaré *et al.* (1997); Weiss & von Haeseler (1998)

Standard ABC Algorithm

defining statistics

↓

generating simulations according to prior



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

Approximate Bayesian Computation ABC

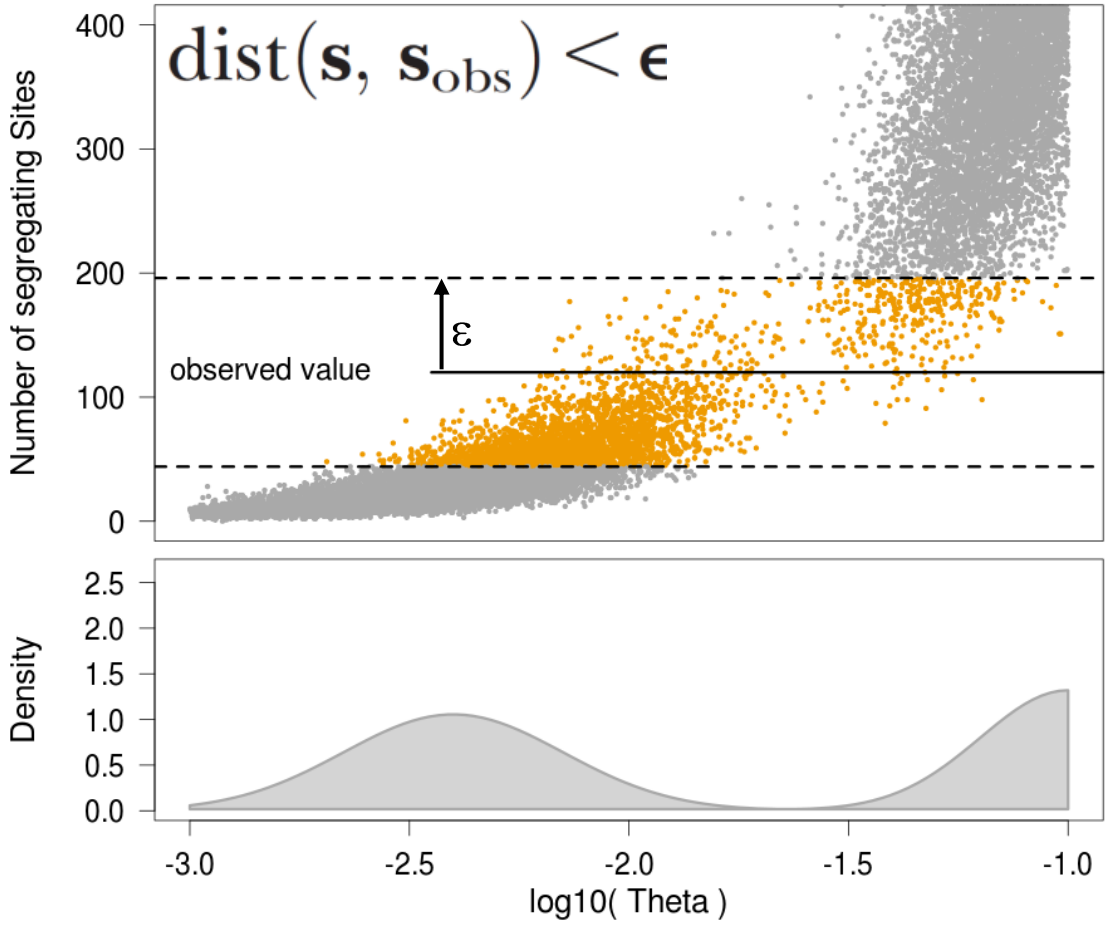
defining statistics

↓

generating simulations according to prior

↓

accepting close simulations



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

Approximate Bayesian Computation ABC

defining statistics

↓

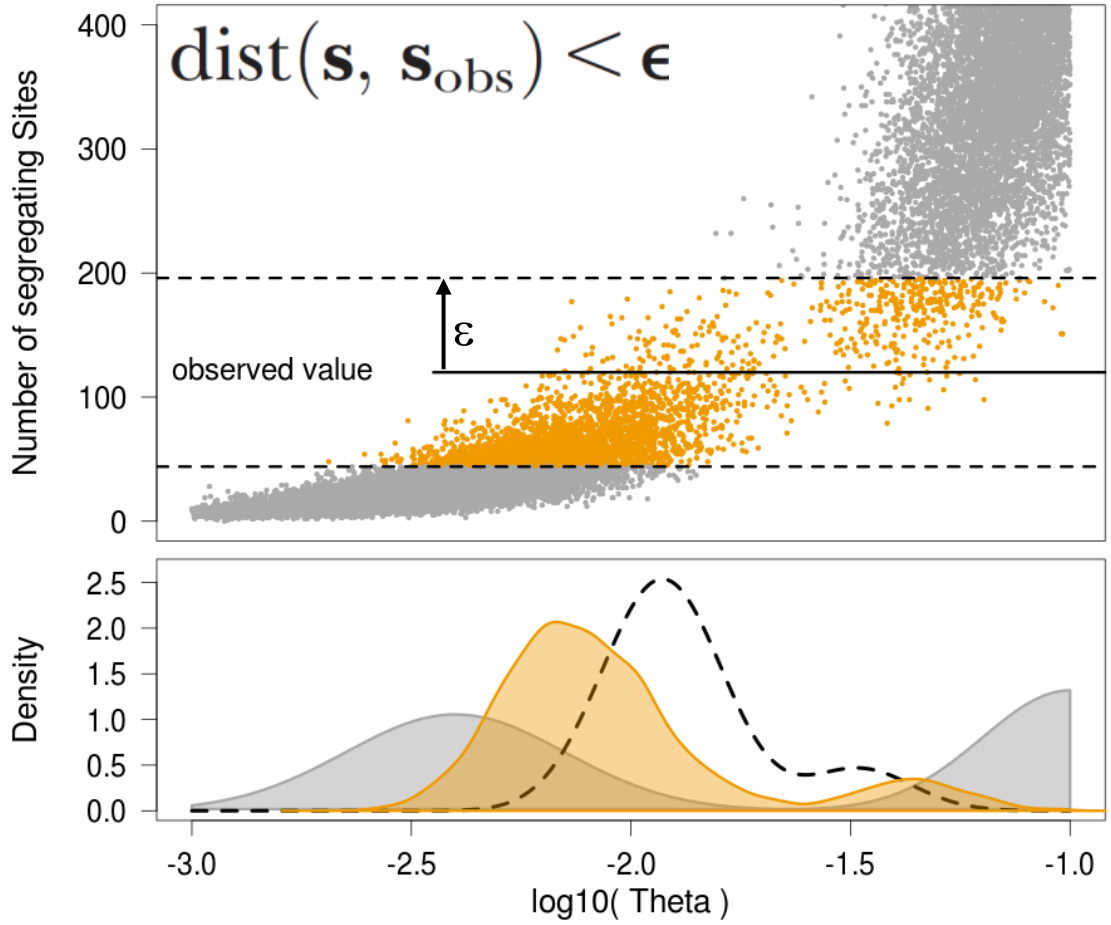
generating simulations according to prior

↓

accepting close simulations

↓

estimating posterior distribution



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

Approximate Bayesian Computation ABC

defining statistics

↓

generating simulations according to prior

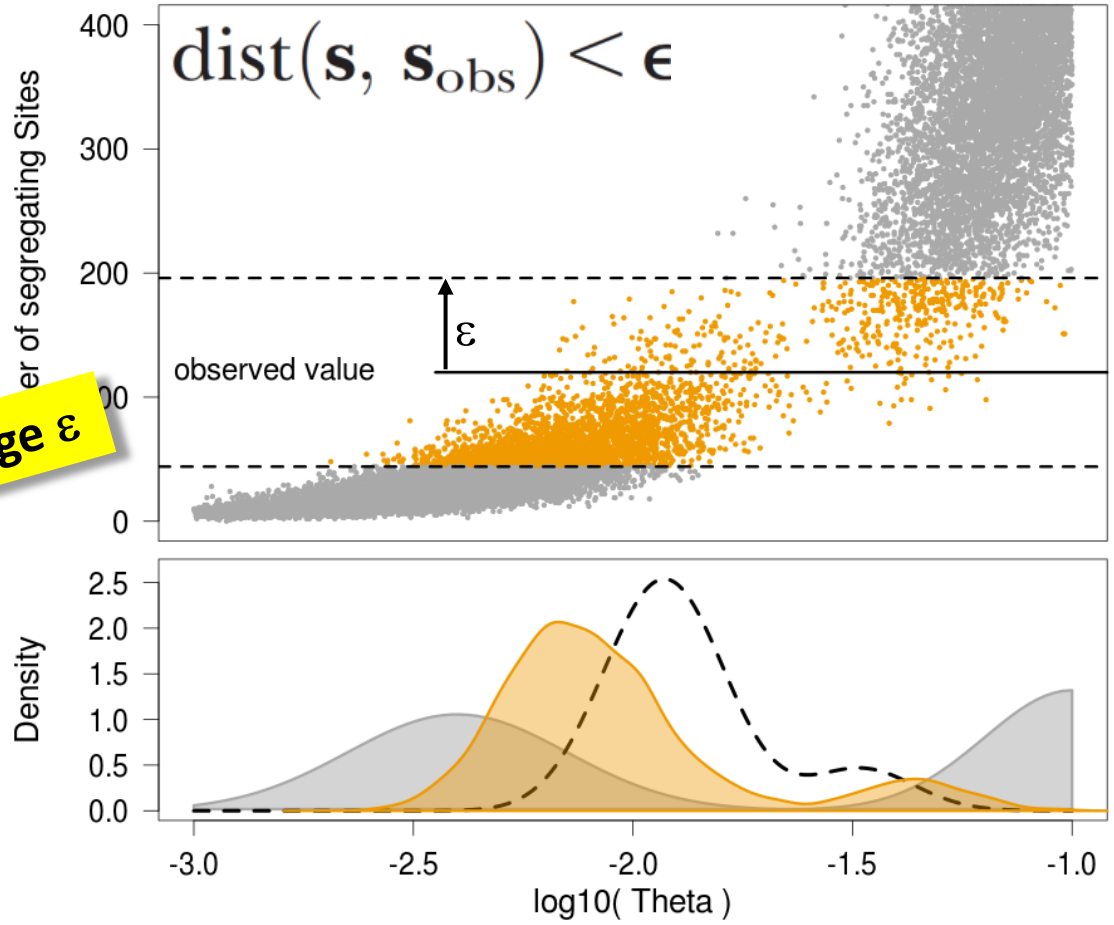
↓

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

↓

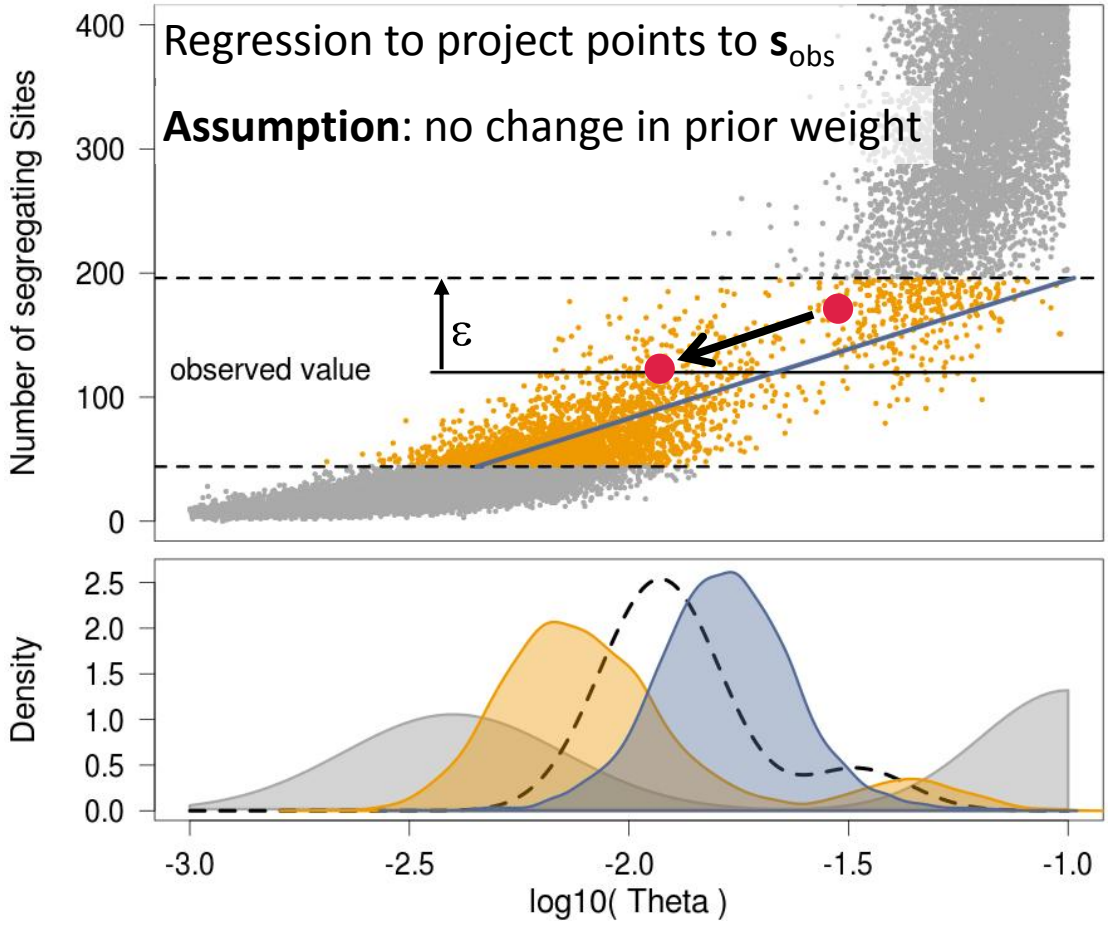
accepting close simulations

↓

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

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

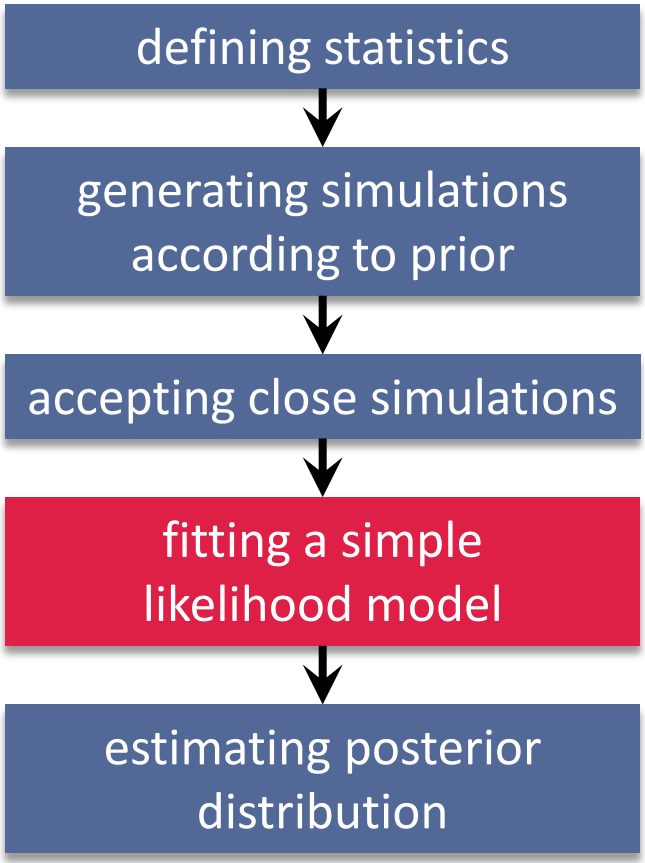
- and $\pi_\epsilon(\boldsymbol{\theta})$ the „truncated prior”

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

Chris Leuenberger

Leuenberger & Wegmann (2010)

ABC-GLM



Leuenberger & Wegmann (2010)

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

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

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

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

ABC-GLM

defining statistics

↓

generating simulations according to prior

↓

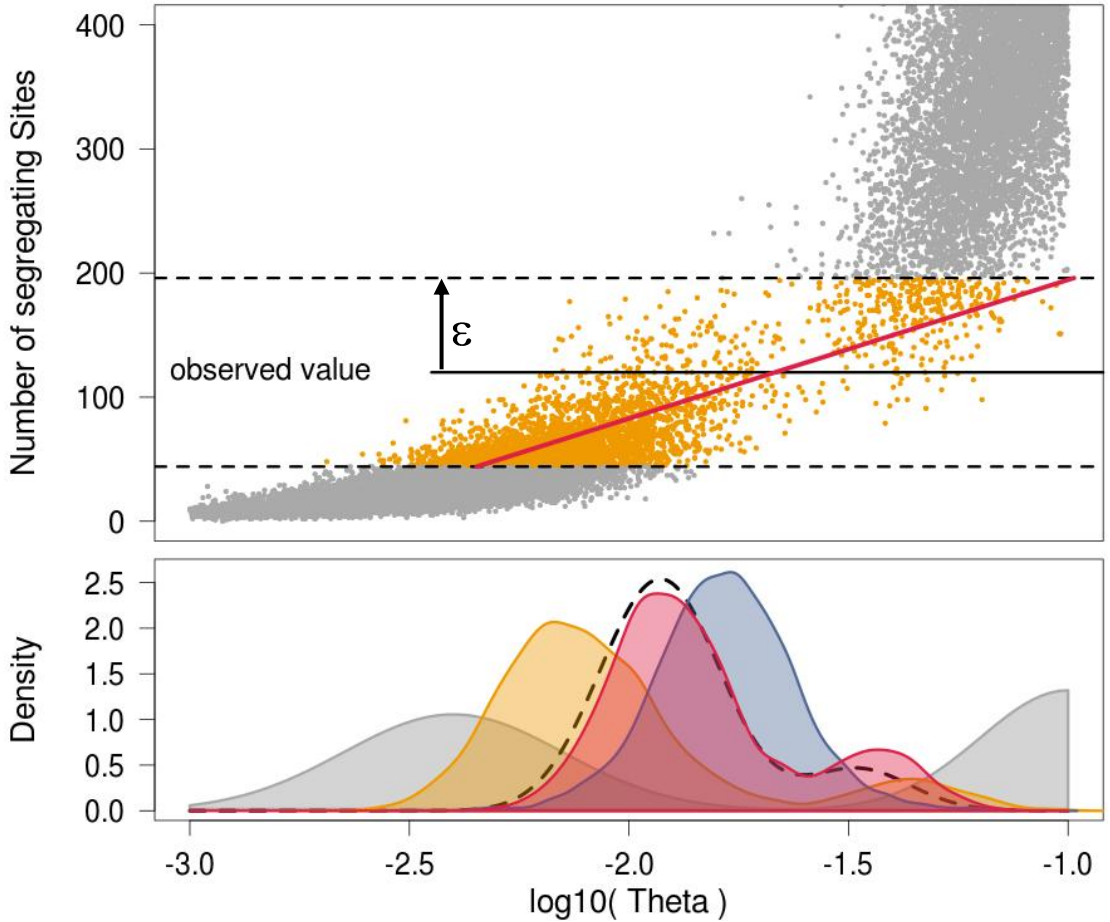
accepting close simulations

↓

fitting a simple likelihood model

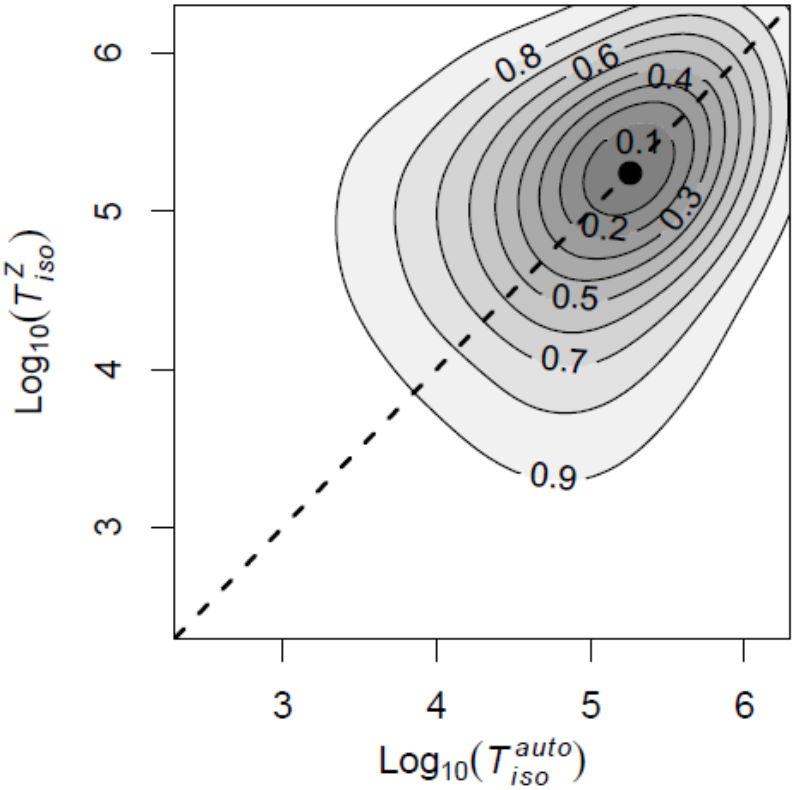
↓

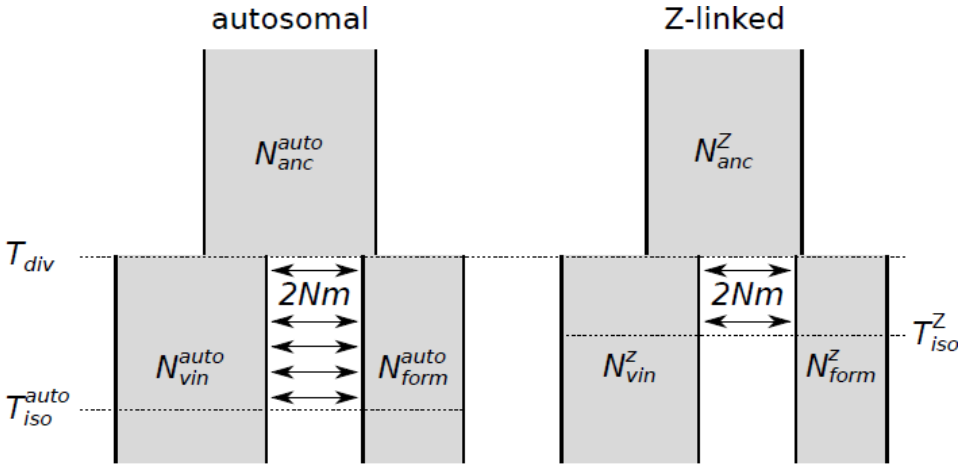
estimating posterior distribution



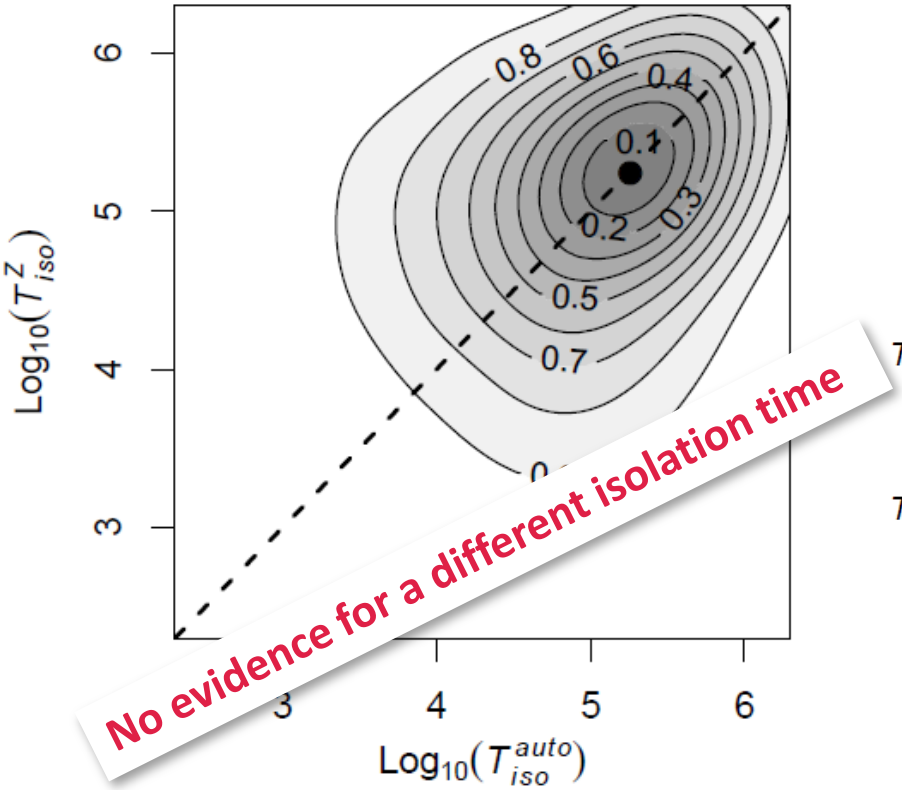
Leuenberger & Wegmann (2010)

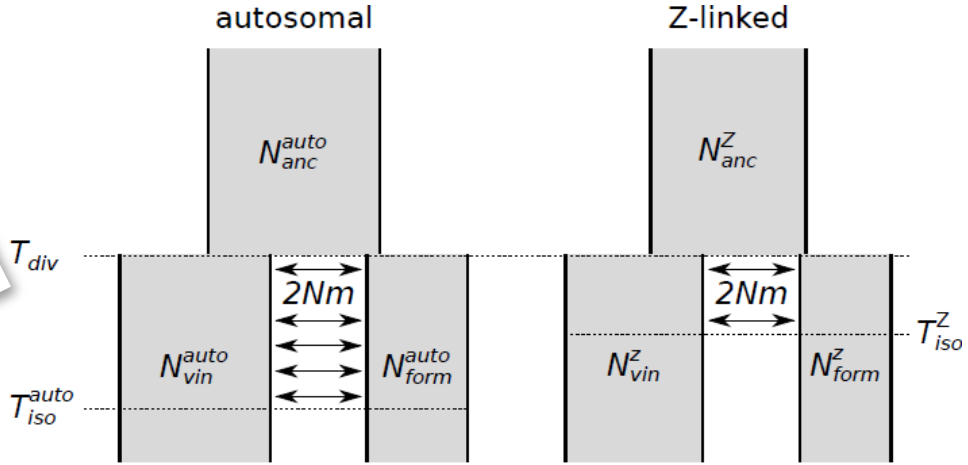
Mode of Speciation in Rose Finches



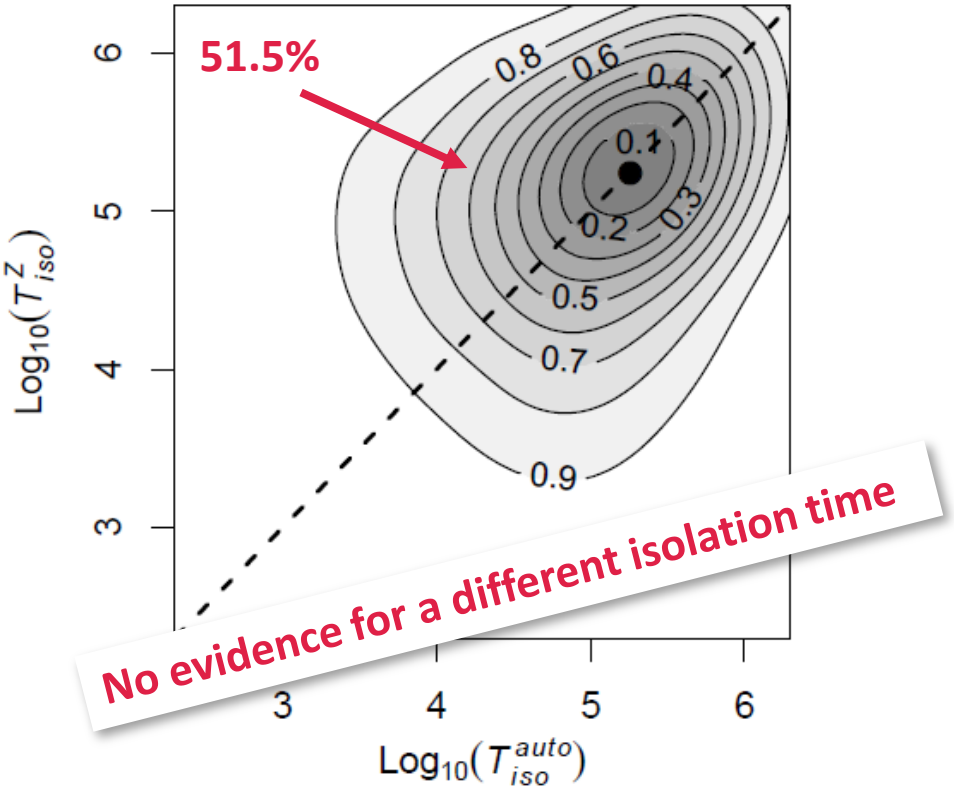


Mode of Speciation in Rose Finches

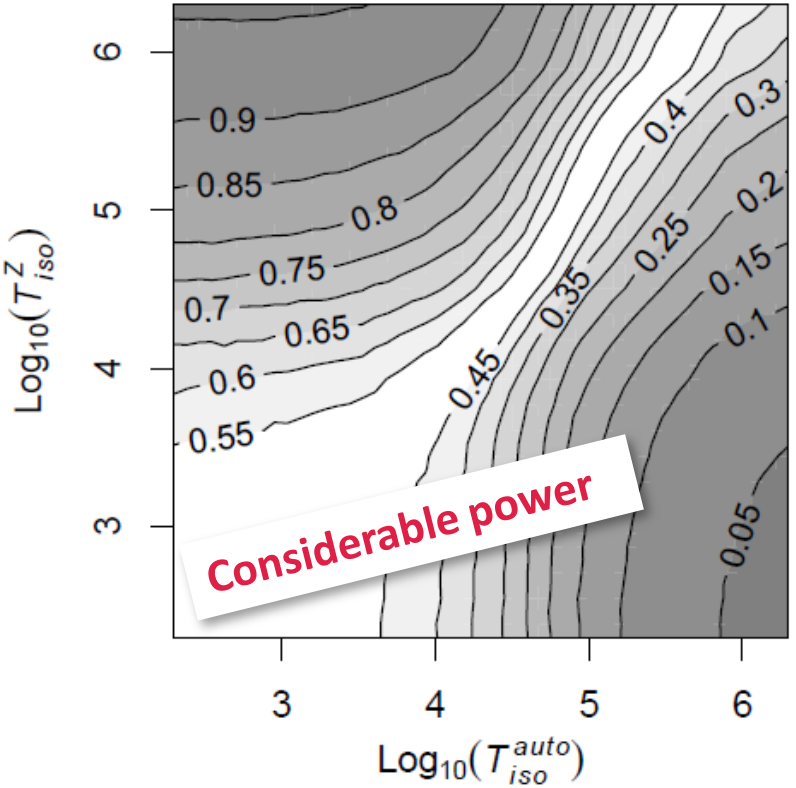




Mode of Speciation in Rose Finches



Joint posterior asymmetry observed in simulated data sets

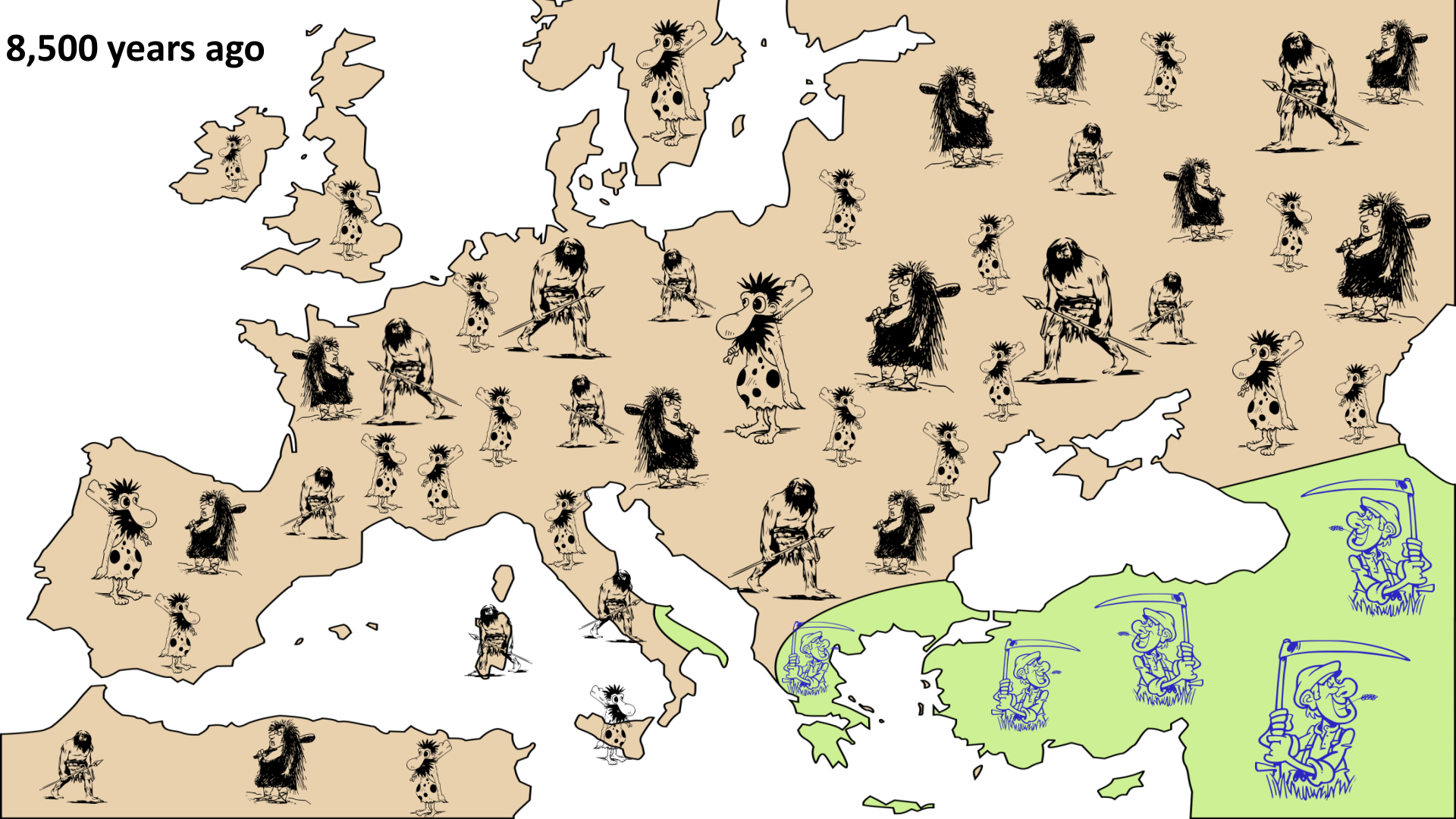


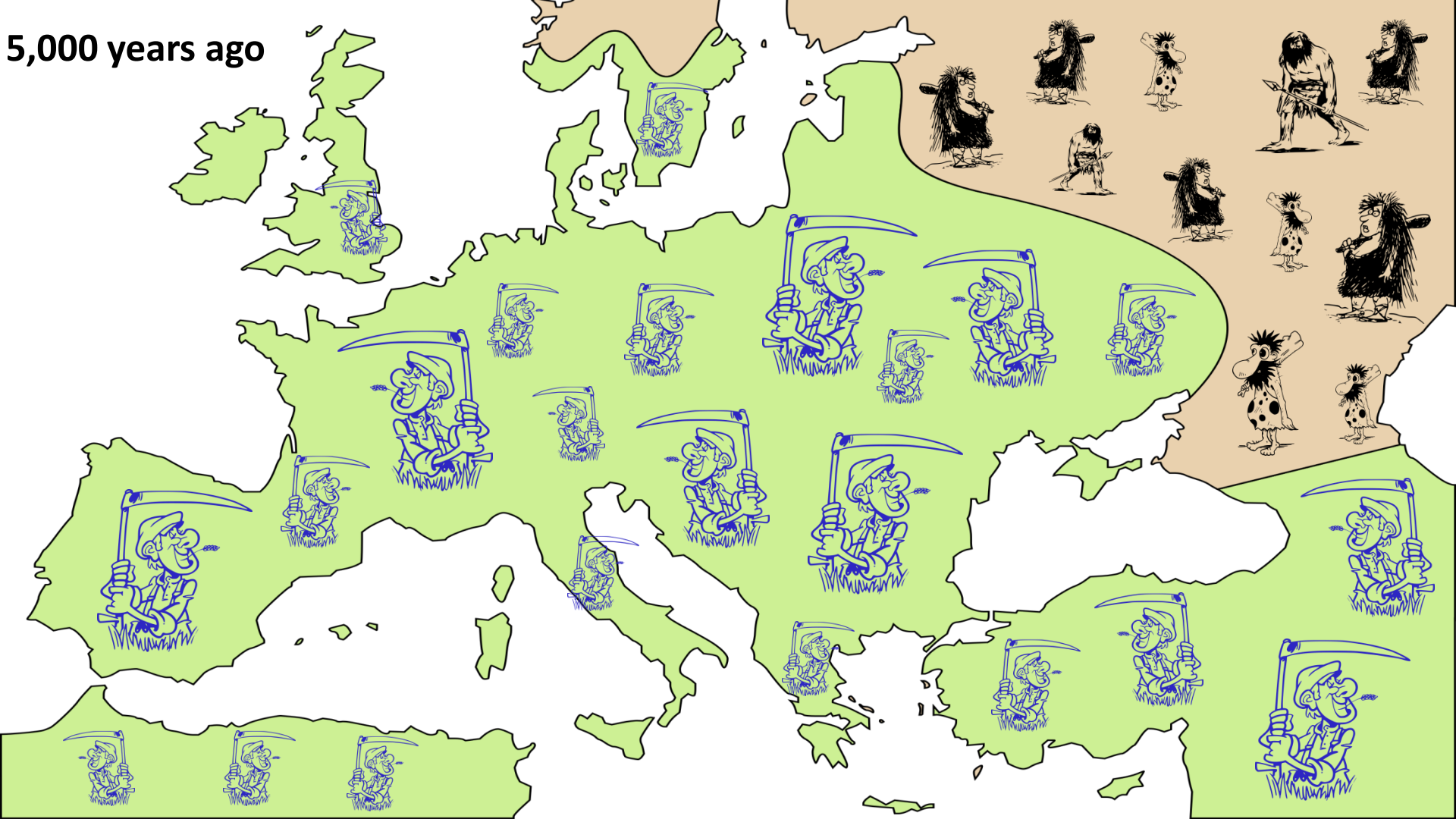
Outline

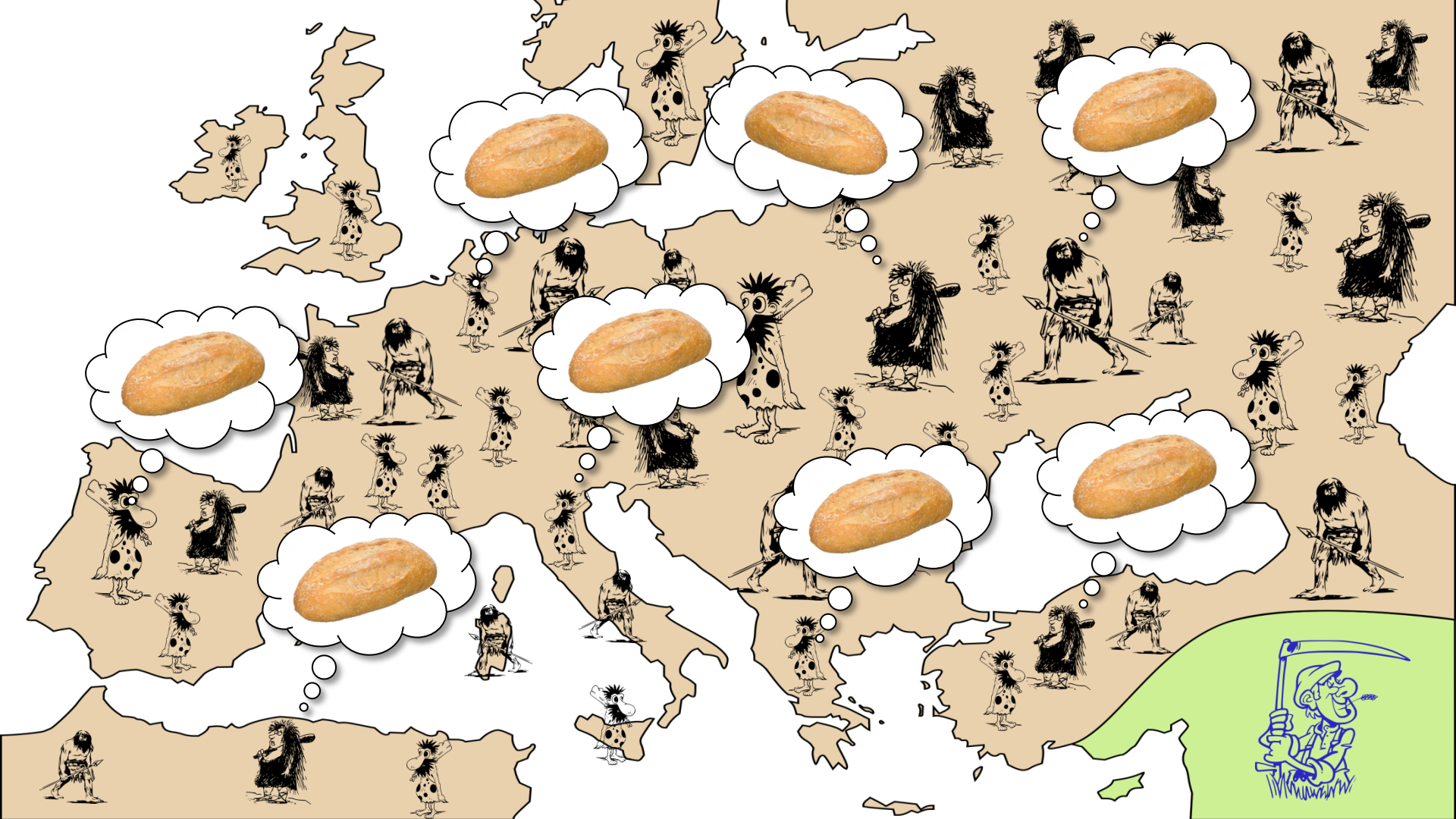
1. Rejecting a null model using summary statistics
2. Composite-likelihood using coalescent simulations
3. Approximate Bayesian Computation (ABC)
4. **Dealing with (ancient) low coverage data**
5. Demography & selection from time series data
6. GWAS (in the unlikely case time will permit)

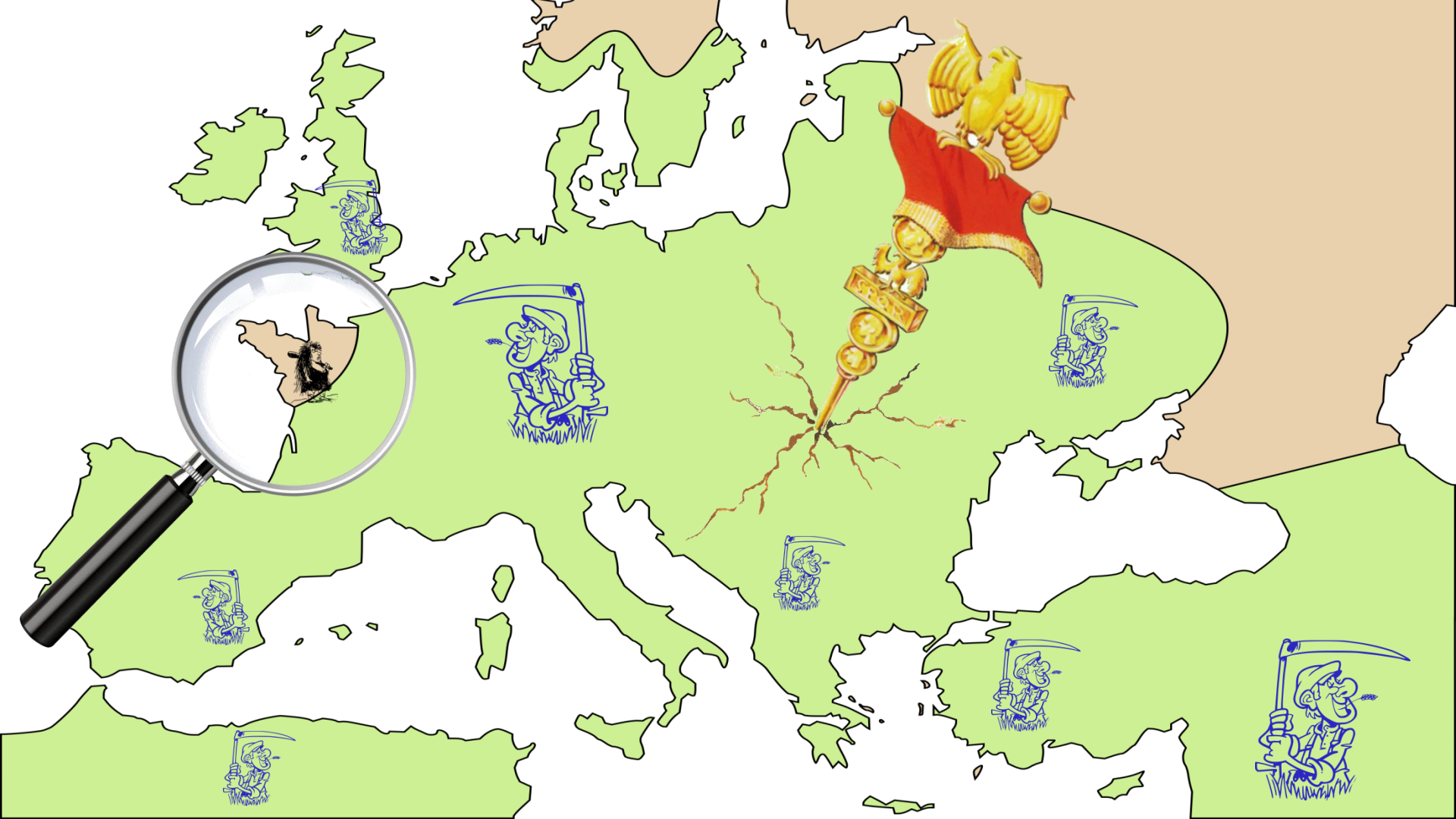

12,000 years ago

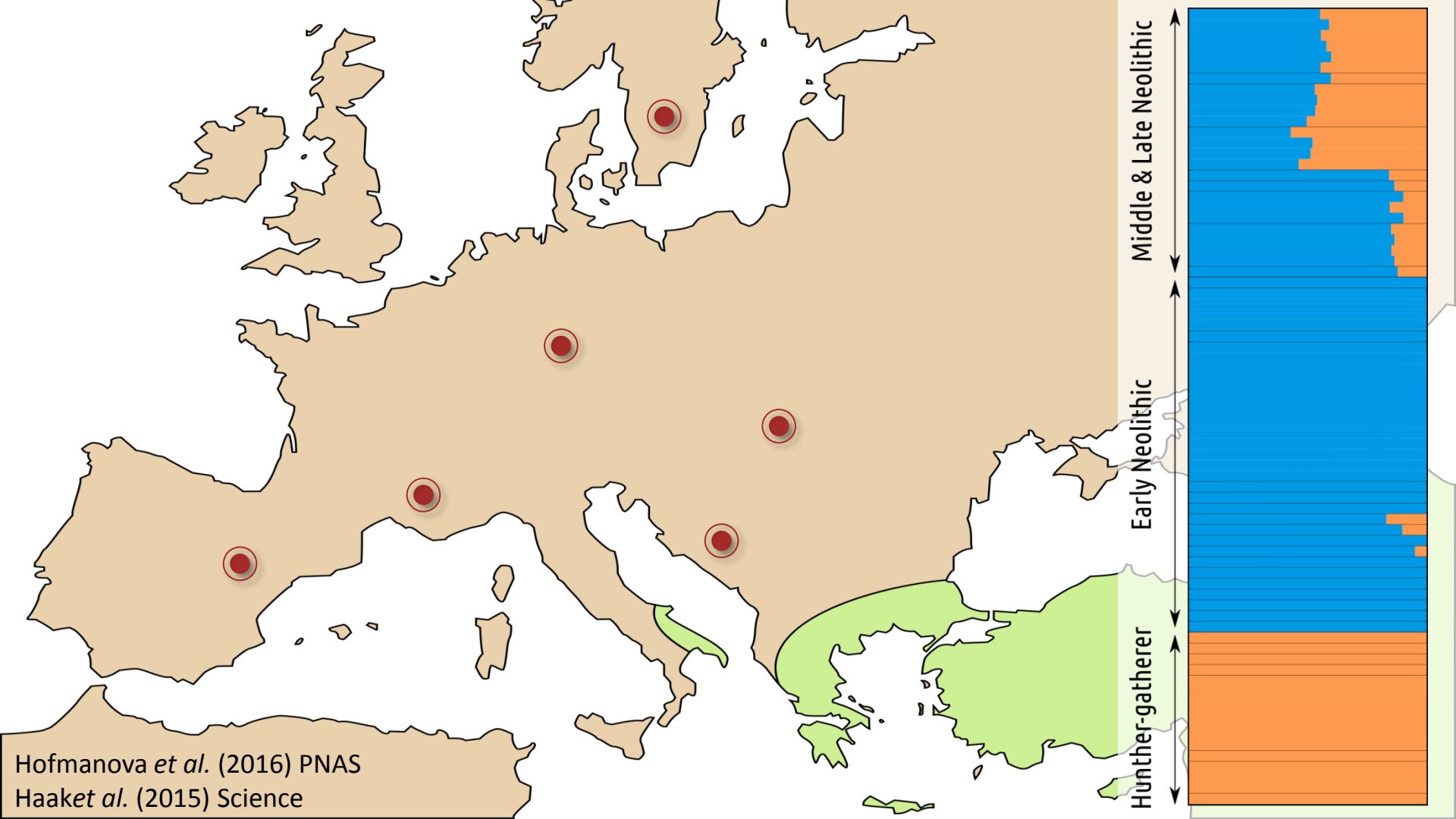

11,000 years ago


8,500 years ago


5,000 years ago




GREECE



Hofmanova *et al.* (2016) PNAS
Haak*et al.* (2015) Science

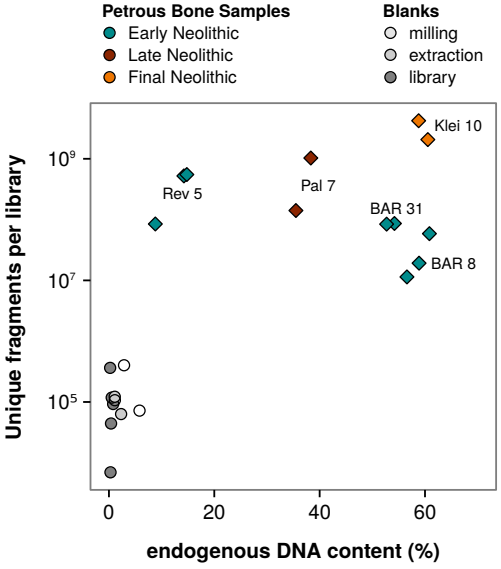
Joachim Burger

Zuzana Hofmanova

Farnaz Broushaki

Low DNA content

Even our best 5 samples contained large amounts of exogenous DNA.

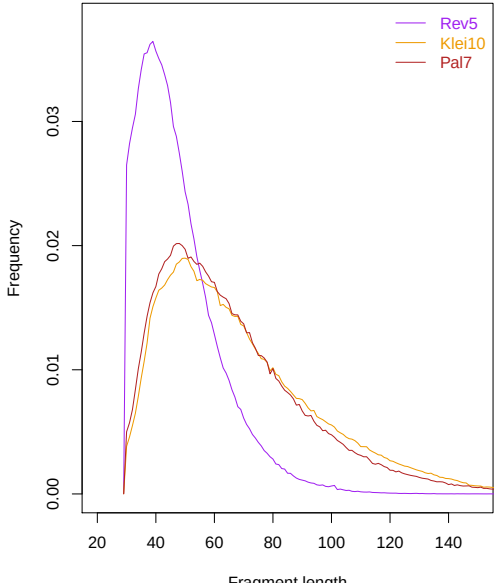


Low DNA content

Even our best 5 samples contained large amounts of exogenous DNA.

Short fragments

Most fragments sequenced were < 50bp.

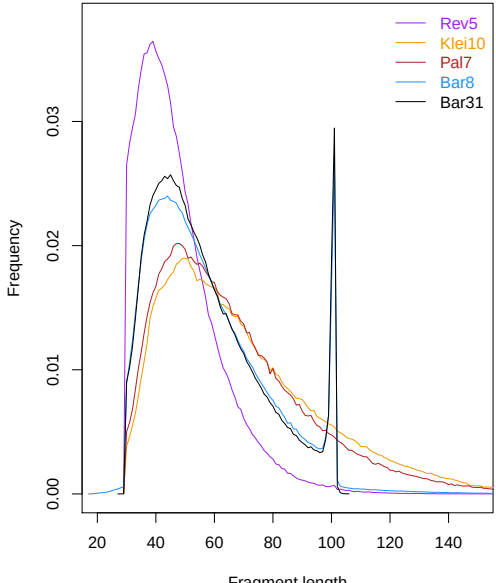


Low DNA content

Even our best 5 samples contained large amounts of exogenous DNA.

Short fragments

Most fragments sequenced were < 50bp.



Low DNA content

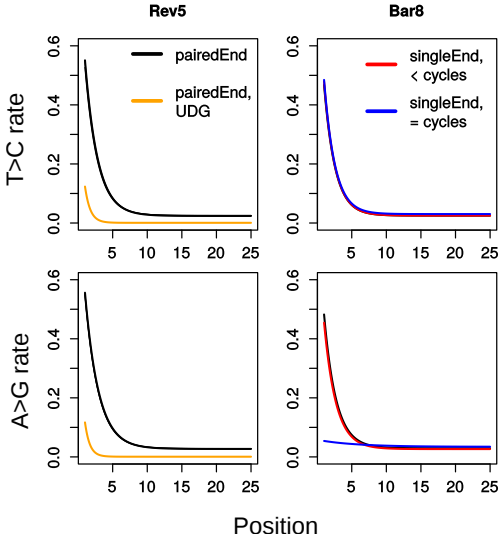
Even our best 5 samples contained large amounts of exogenous DNA.

Short fragments

Most fragments sequenced were < 50bp.

Post-Mortem Damage (PMD)

Over time, cytosins deaminate, resulting in C $\rightarrow$ T (or G $\rightarrow$ A) errors after PCR.



Low sequencing depth and high error rates == ambiguous genotyping

- A major problem with ancient DNA is the low depth ($<$1x) of many samples, ...
- ... and the prevalence of sequencing errors and PMD.
- Clearly, estimating diversity from called genotypes is a bad idea.

Estimating heterozygosity from low depth data

Low sequencing depth and high error rates == ambiguous genotyping

- A major problem with ancient DNA is the low depth ($<$1x) of many samples, ...
- ... and the prevalence of sequencing errors and PMD.
- Clearly, estimating diversity from called genotypes is a bad idea.

How to turn noisy data into biological insight?

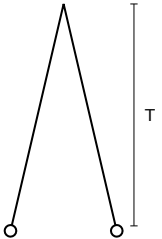
1) Model error rates (noise) explicitly
2) Infer hierarchical parameters to **integrate out uncertainty** and **aggregate information** across loci.

- 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

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

where g_i denotes the hidden genotype.



Athanasios Kousathanas

Vivian Link

Felsenstein's substitution model (1981)

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

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

Likelihood of individual bases

- $\mathbb{P}(d_j = A | g = T, \epsilon_j, D_{jA}) = \frac{\epsilon_j}{3}$
- $\mathbb{P}(d_j = T | g = C, \epsilon_j, D_{jT}) = D_{jg}(1 - \epsilon_i) + (1 - D_{jg})\frac{\epsilon_j}{3}$

d_j: base in read j
g: hidden genotype
ϵ_i: error rate
D_{jg}: generalized PMD prob.

Modeling / Inferring PMD

We infer PMD patterns from the data by fitting a generalization of the exponential decay function proposed by Skoglund *et al.* (2014)

Emission probabilities

Likelihood of individual bases

- $\mathbb{P}(d_j = A | g = T, \epsilon_j, D_{jA}) = \frac{\epsilon_i}{3}$
- $\mathbb{P}(d_j = T | g = C, \epsilon_j, D_{jT}) = D_{jg}(1 - \epsilon_i) + (1 - D_{jg})\frac{\epsilon_i}{3}$

d_j: base in read j
g: hidden genotype
ϵ_i: error rate
D_{jg}: generalized PMD prob.

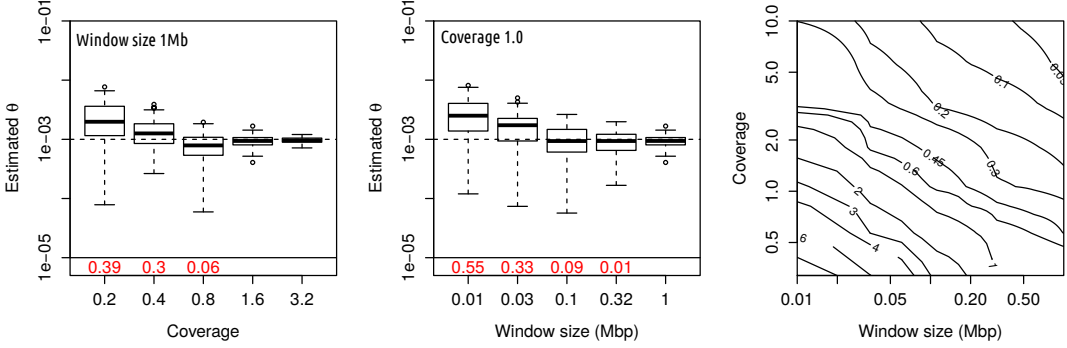
Modeling / Inferring PMD

We infer PMD patterns from the data by fitting a generalization of the exponential decay function proposed by Skoglund *et al.* (2014)

Likelihood of data at one site

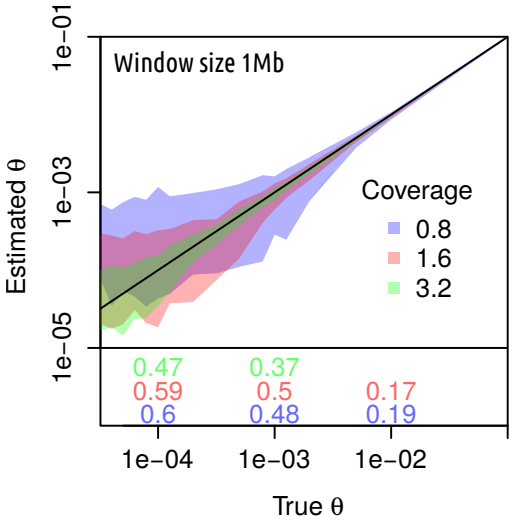
$$\mathbb{P}(\mathbf{d}_i | g, \epsilon_i) = \prod_{i=1}^{N} \mathbb{P}(d_j | g, \epsilon_j, D_{jg})$$

Power analysis through simulations



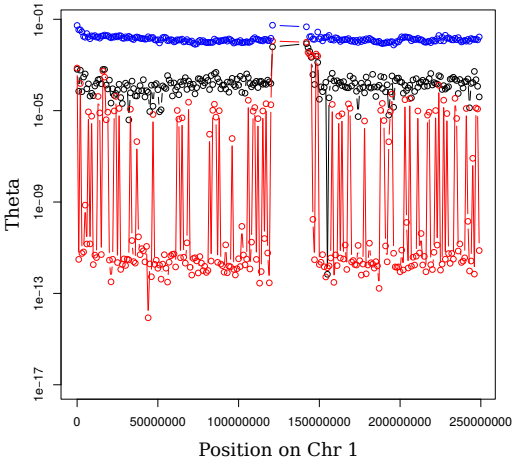
- Relatively high power to infer θ within a 1Mb window even at very low depth.
- Increasing window size or sequencing depth has very similar effects.

Power analysis through simulations



- At low theta, we often infer no diversity at all.
- Higher sequencing depth or larger windows required for small θ values.

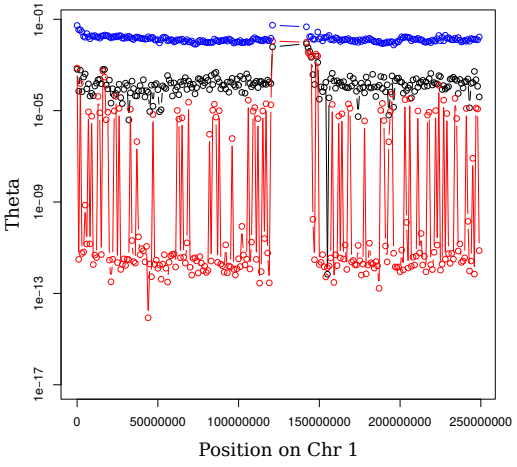
Application to ancient Greek samples



Three Greek samples

- Samples are about 10,000 years old.
- Sequencing depths from 0.8 to 3.5x.
- Expected theta for humans: $\sim 10^{-3}$.

Application to ancient Greek samples



Three Greek samples

- Samples are about 10,000 years old.
- Sequencing depths from 0.8 to 3.5x.
- Expected theta for humans: $\sim 10^{-3}$.

Why are these estimates so different?

Recalibration using X-linked data

X-linked data of males is informative about base qualities, as there should be no polymorphisms. We assume:

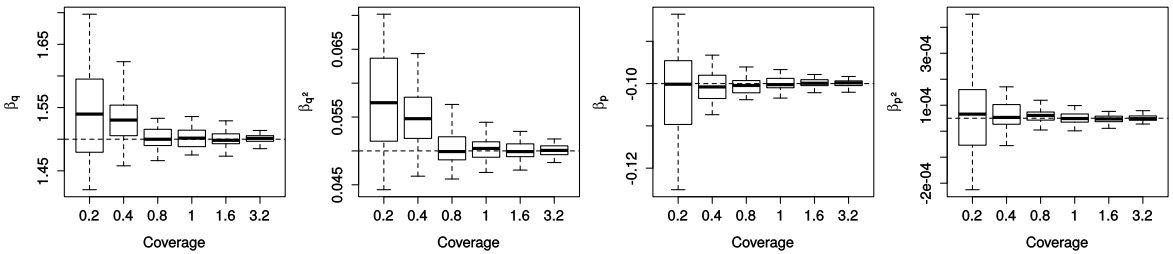
$$\mathbb{P}(d_i|\boldsymbol{\beta}) = \sum_g \prod_{j=1}^{n_i} [\mathbb{P}(d_i|g, \epsilon_{ij}, D_{jg})]\, \mathbb{P}(g|\boldsymbol{\pi}), \quad g = A, C, G, T$$

where

$$\epsilon_{ij} = \epsilon_{ij}(\mathbf{q_{ij}}, \boldsymbol{\beta}) = \frac{\exp(\eta_{ij}(\boldsymbol{\beta}))}{1 + \exp(\eta_{ij}(\boldsymbol{\beta}))}; \qquad \eta_{ik}(\boldsymbol{\beta}) = \beta_0 + \sum_{l=1}^{L} q_{ijl}\beta_l.$$

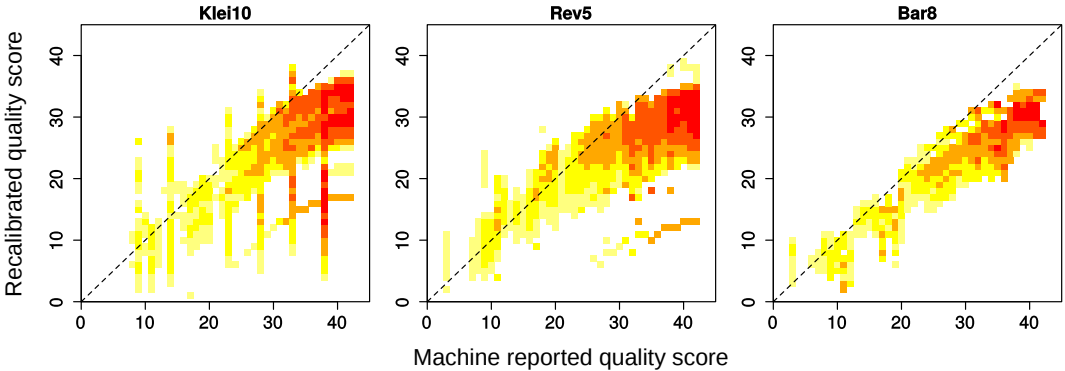
$\mathbf{q_{ij}} = (q_{ij1}, \ldots, q_{ijL})$ is a vector of information (raw quality score, the position within the read, two-base contexts, ...), and
$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_L)$ are the parameters of the model we infer using an EM algorithm.

Power analysis using simulations



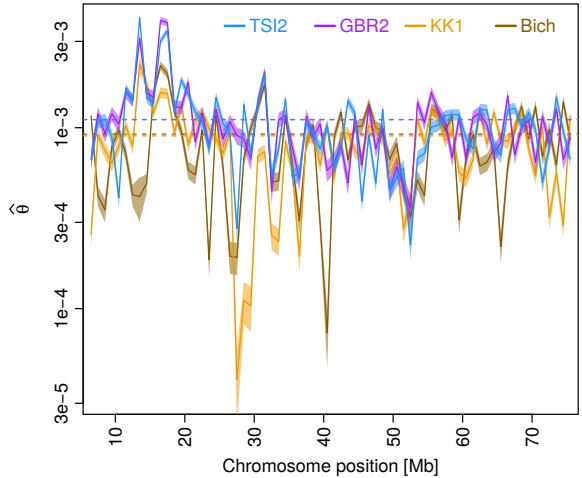
- High power to infer recalibration model from 1Mb at 2x or 10Mb at 1x.
- Sufficient power to apply it also to e.g. Ultra Conserved elements or mtDNA.

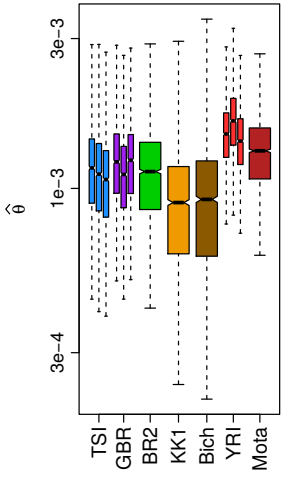
Application to ancient samples

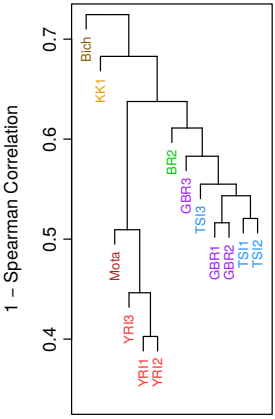


- Recalibration of Greek samples revealed a general overestimation in quality.
- The complex pattern is a result of library-specific recalibration.

Application to ancient samples



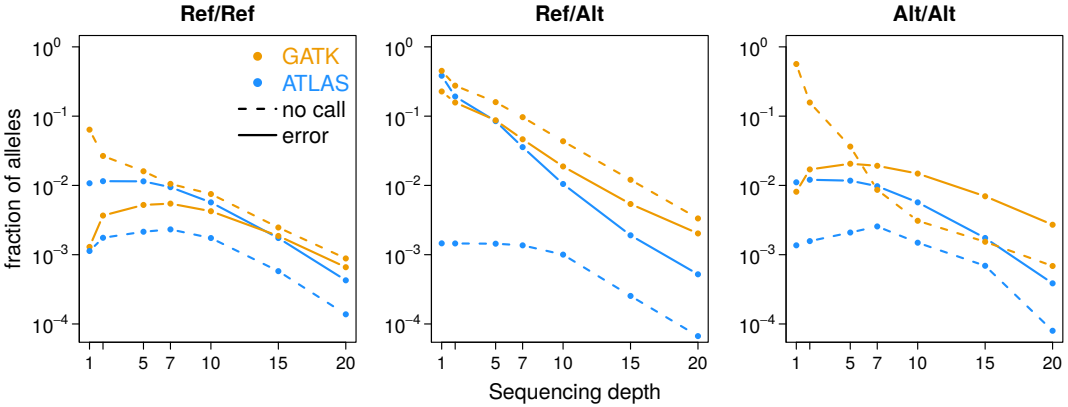




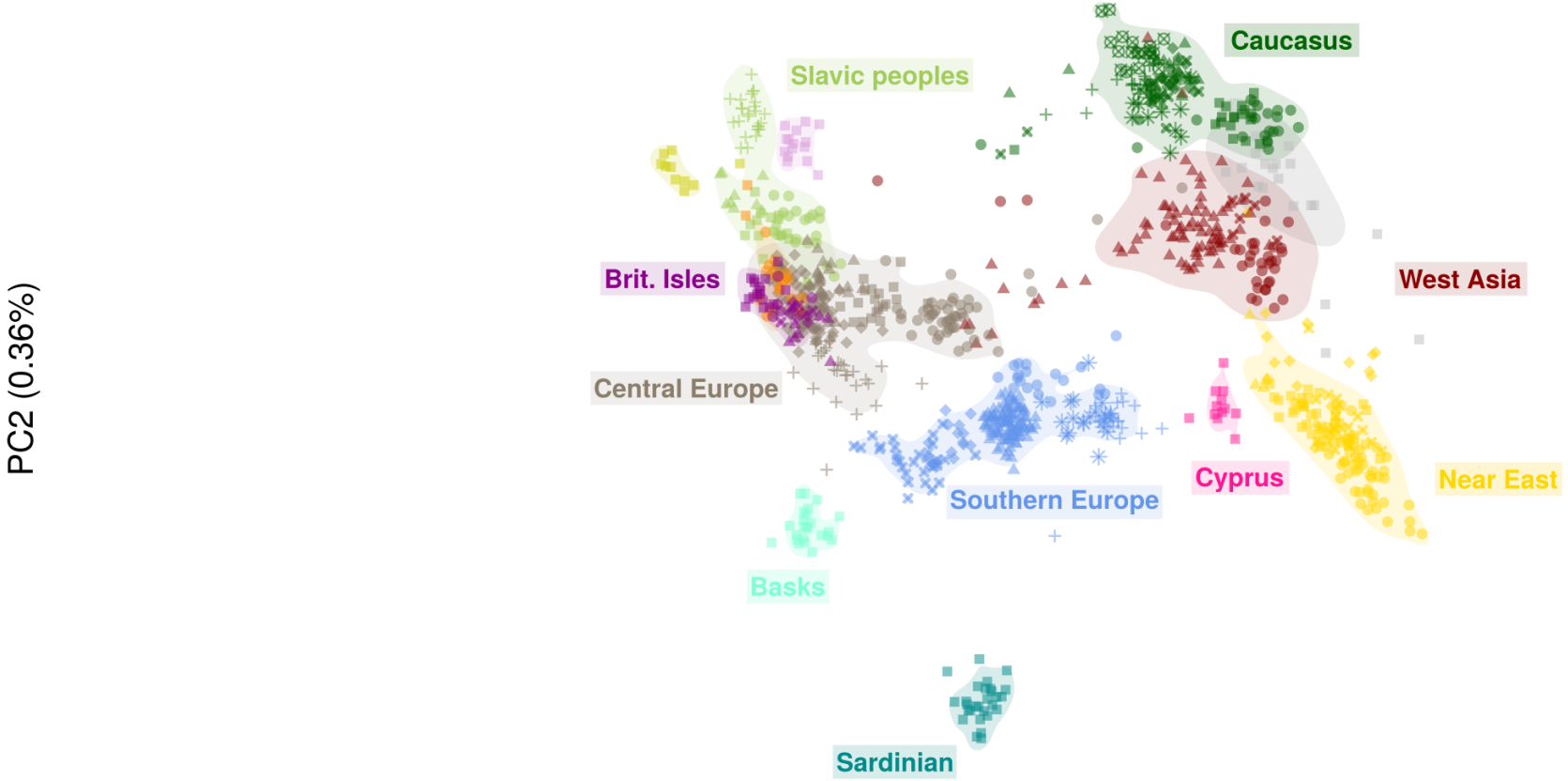
- Diversity of ancient samples is similar to moderns, but hunter-gatherers differ.

Calling ancient genotypes

- The model lends itself readily for unbiased genotype calling.
- In simulations, our caller outperforms GATK.

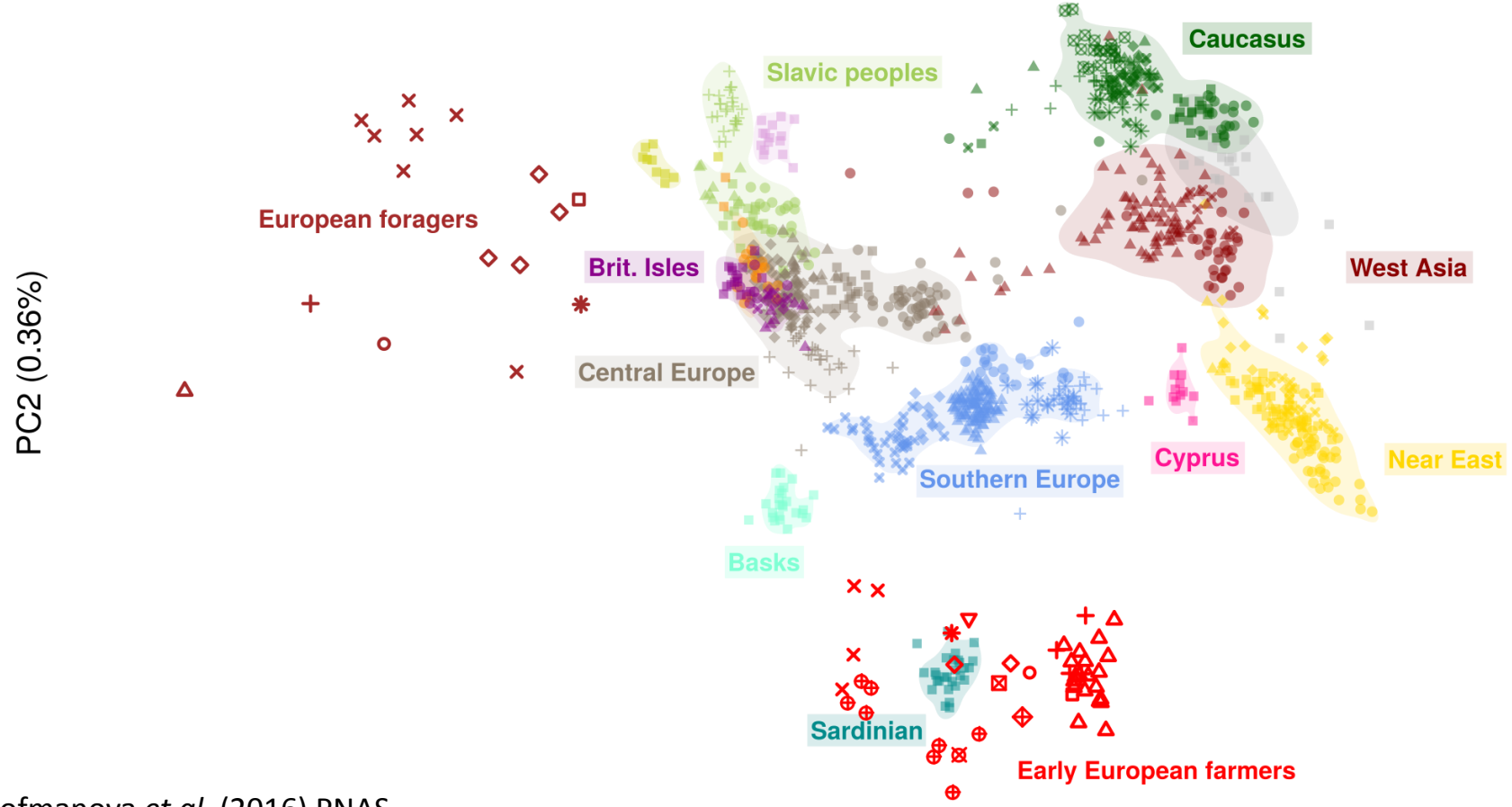


PCA of modern and ancient samples



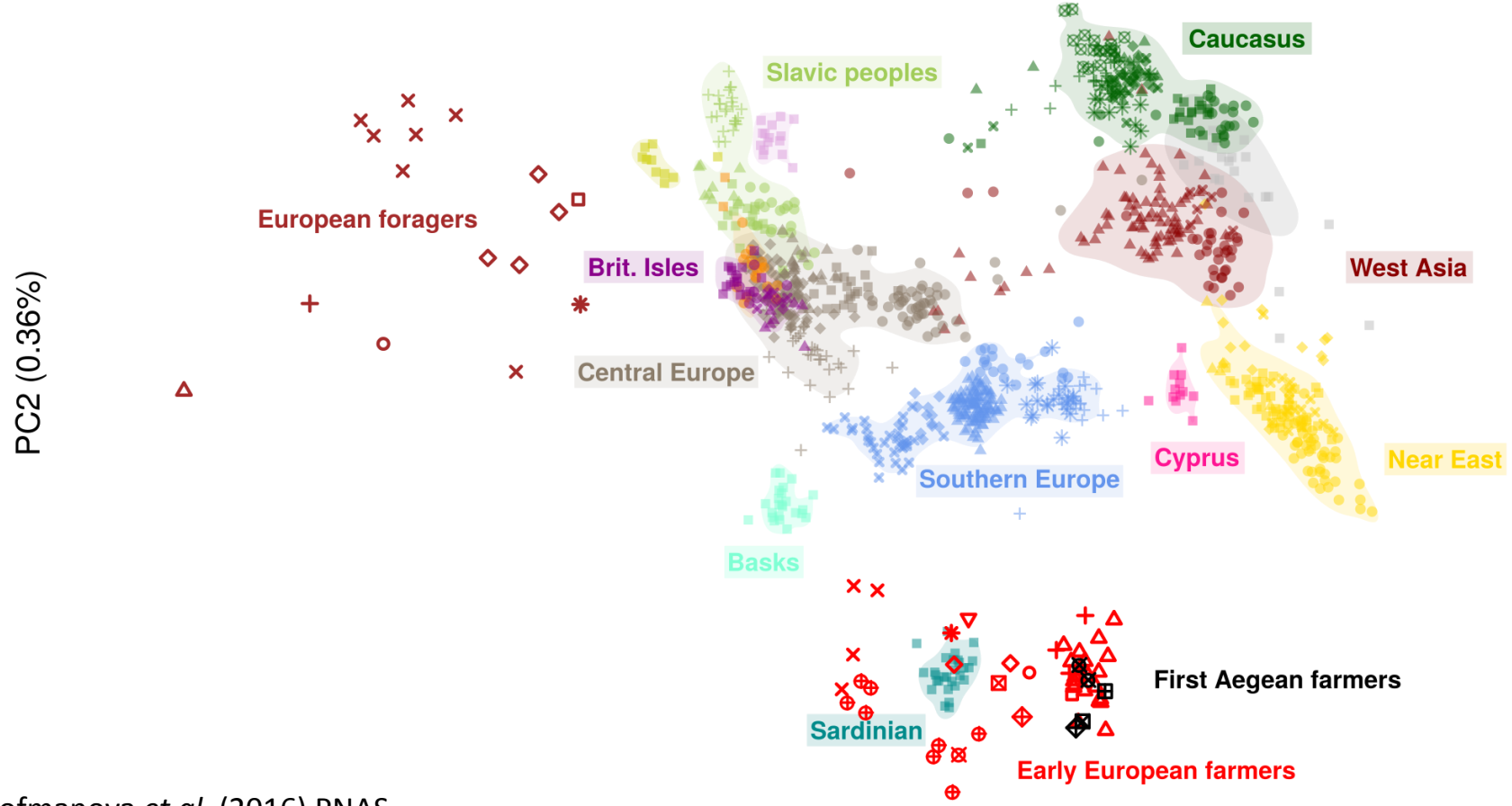
Hofmanova *et al.* (2016) PNAS
Broushaki *et al.* (2016) Science

PCA of modern and ancient samples



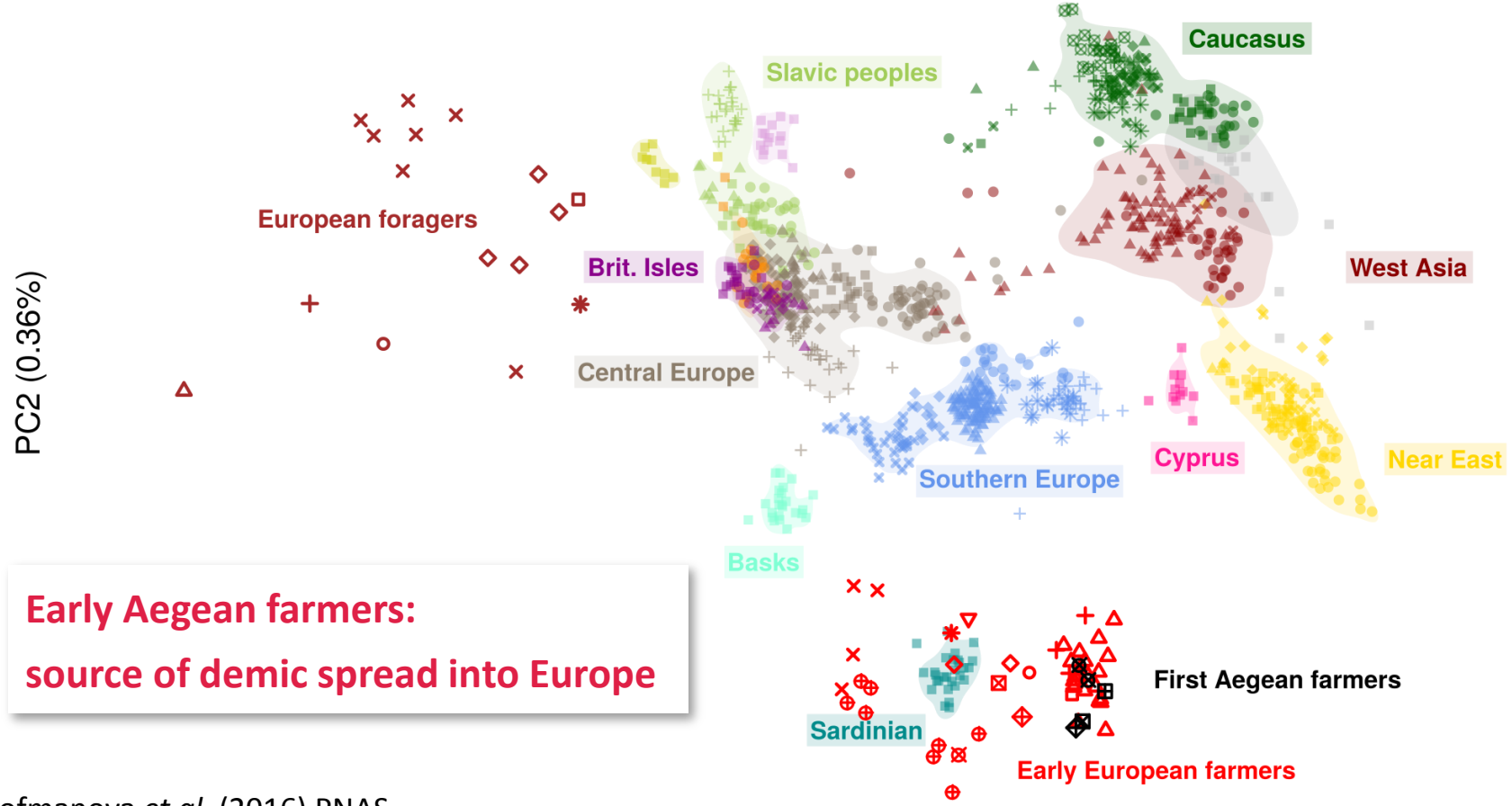
Hofmanova *et al.* (2016) PNAS
Broushaki *et al.* (2016) Science

PCA of modern and ancient samples



Hofmanova *et al.* (2016) PNAS
Broushaki *et al.* (2016) Science

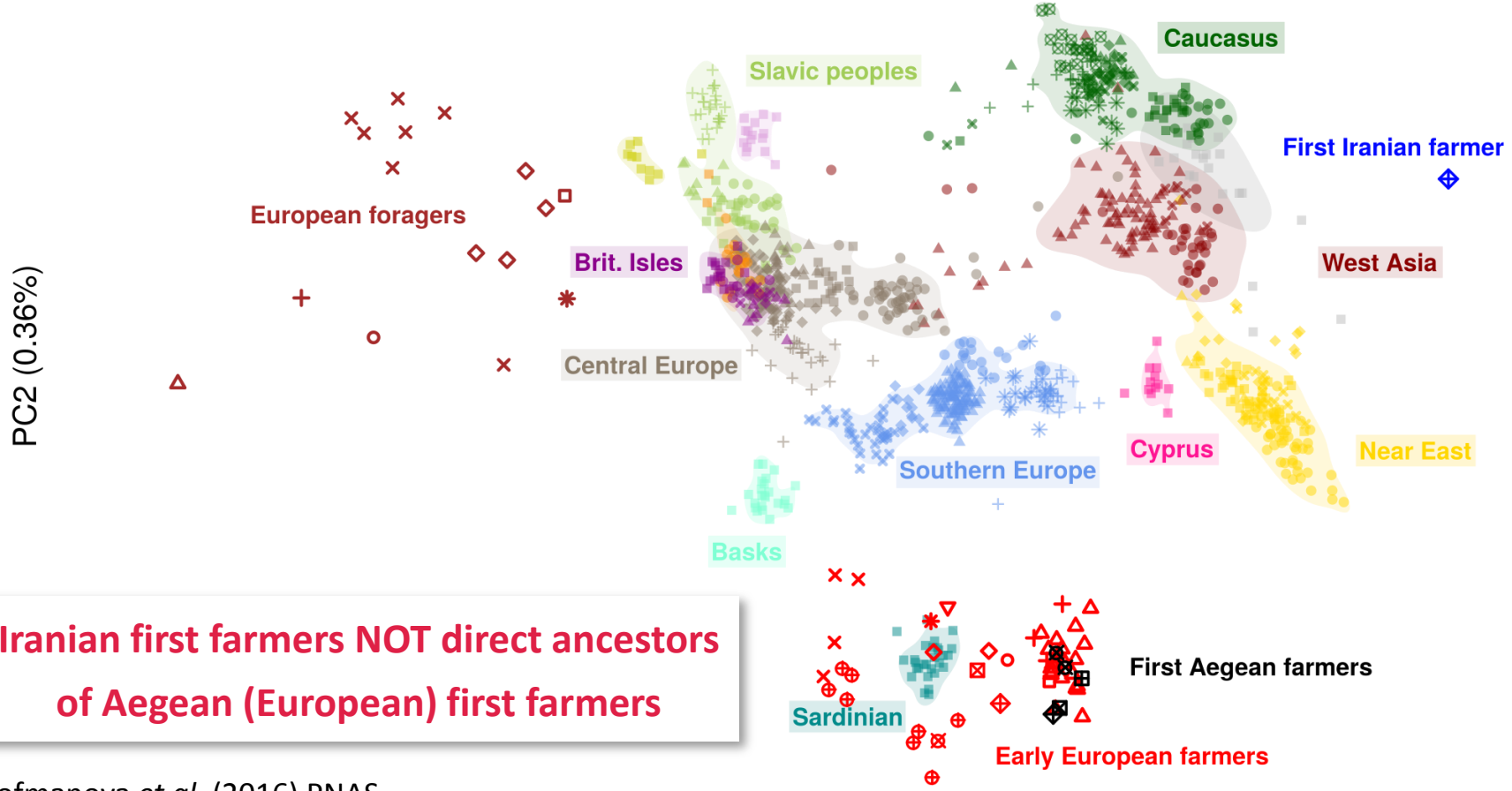
PCA of modern and ancient samples



Early Aegean farmers:
source of demic spread into Europe

Hofmanova *et al.* (2016) PNAS
Broushaki *et al.* (2016) Science

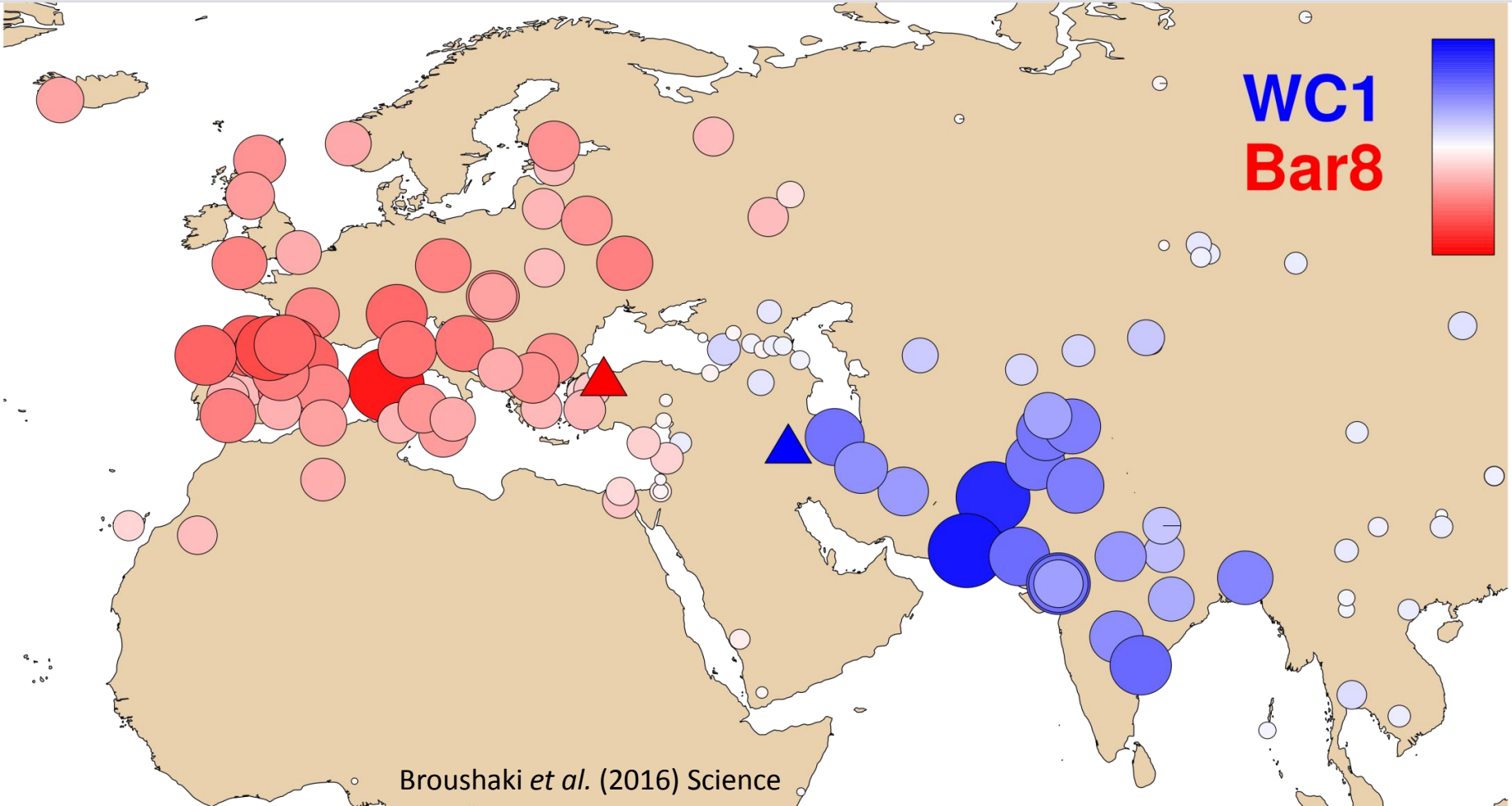
PCA of modern and ancient samples



Iranian first farmers NOT direct ancestors of Aegean (European) first farmers

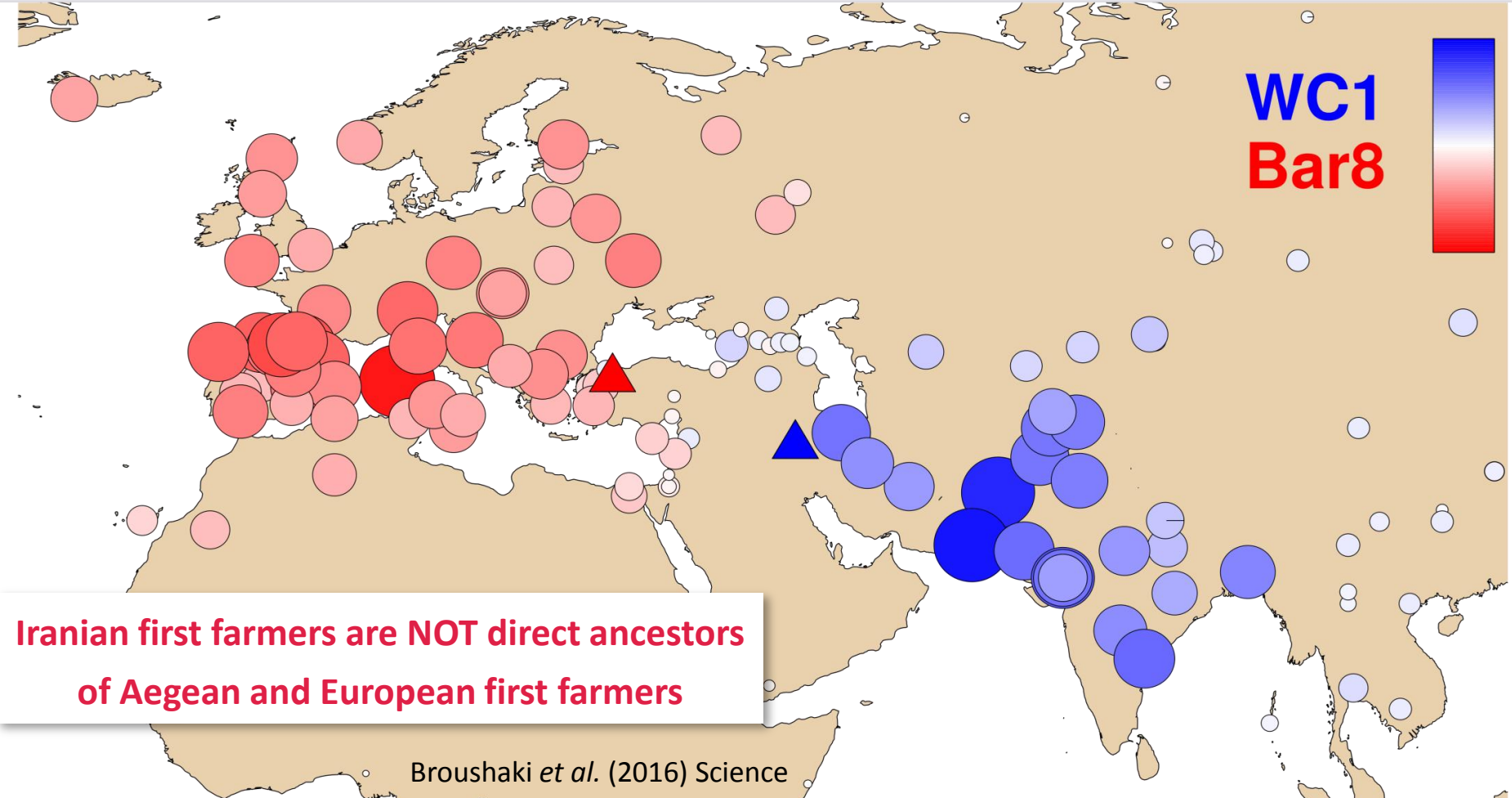
Hofmanova *et al.* (2016) PNAS
Broushaki *et al.* (2016) Science

Chromopainter analysis



Broushaki *et al.* (2016) Science

Chromopainter analysis



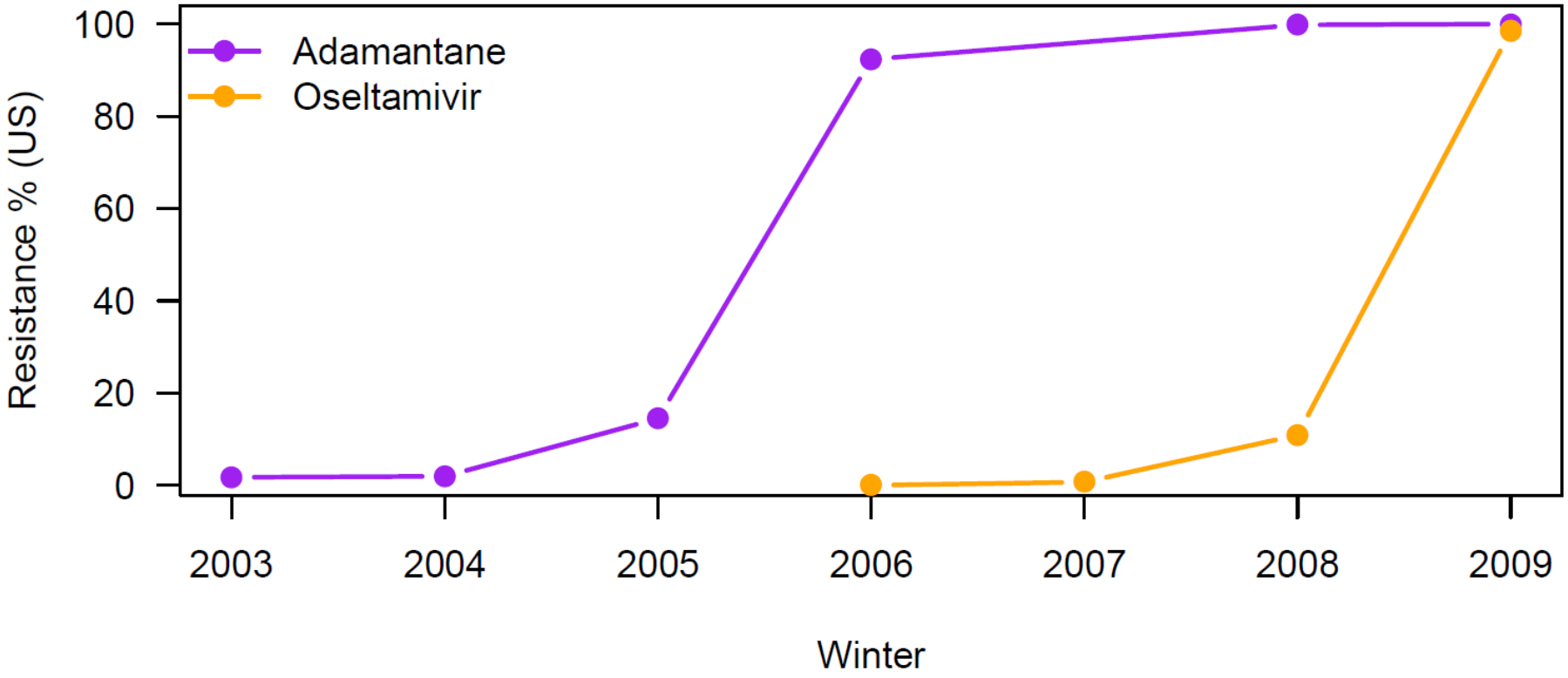
Iranian first farmers are NOT direct ancestors of Aegean and European first farmers

Broushaki *et al.* (2016) Science

Outline

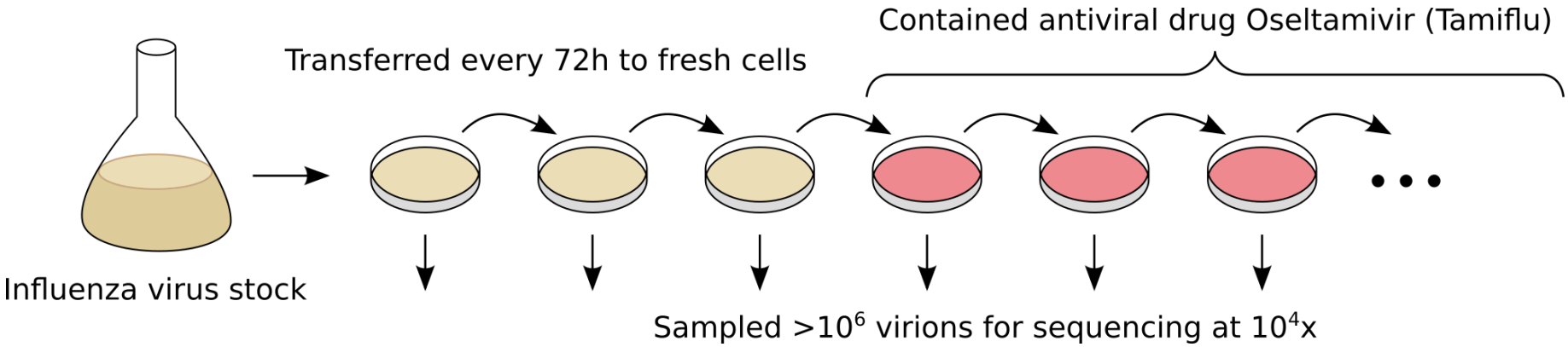
1. Rejecting a null model using summary statistics
2. Composite-likelihood using coalescent simulations
3. Approximate Bayesian Computation (ABC)
4. Dealing with (ancient) low coverage data
5. **Demography & selection from time series data**
6. GWAS (in the unlikely case time will permit)

Influenza rapidly evolved resistance against novel drugs



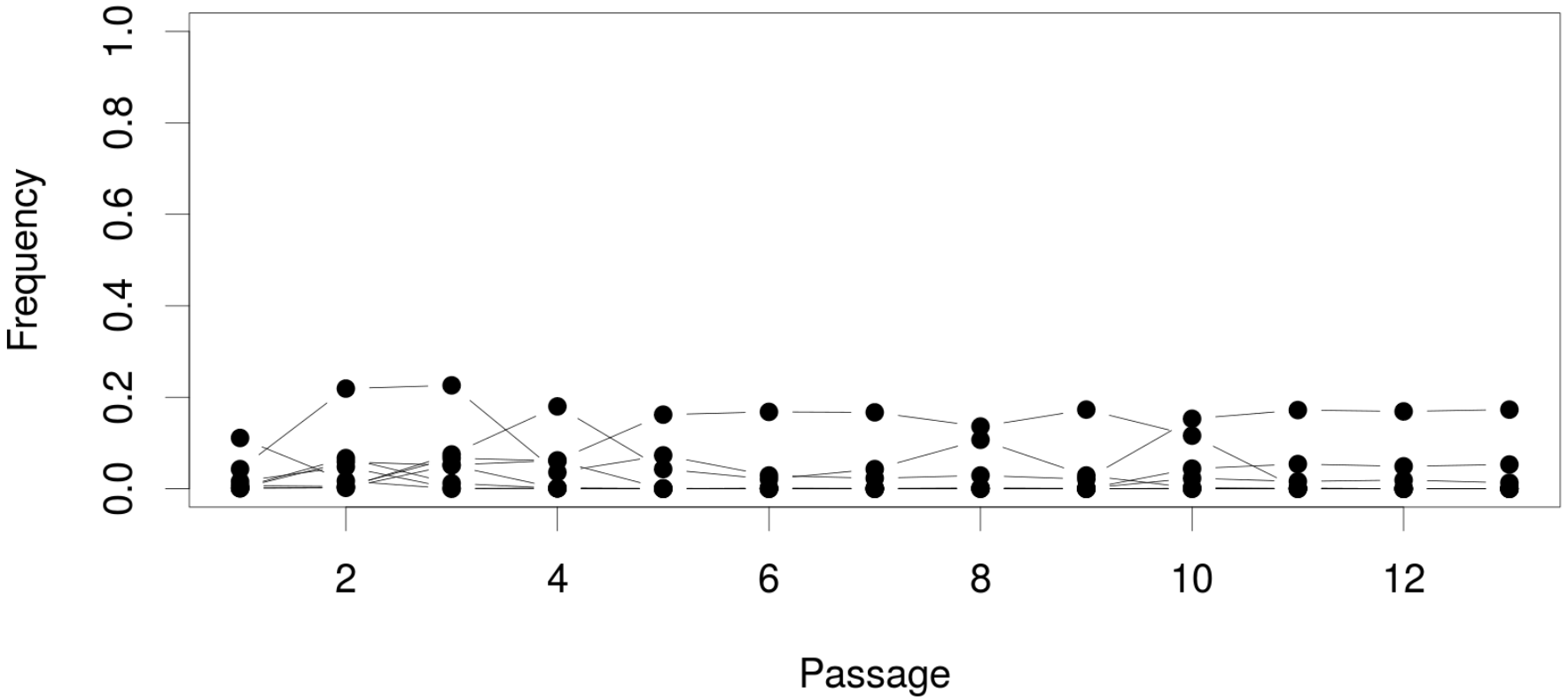
Weinstock & Zuccotti (2009)

Experimental evolution in the lab

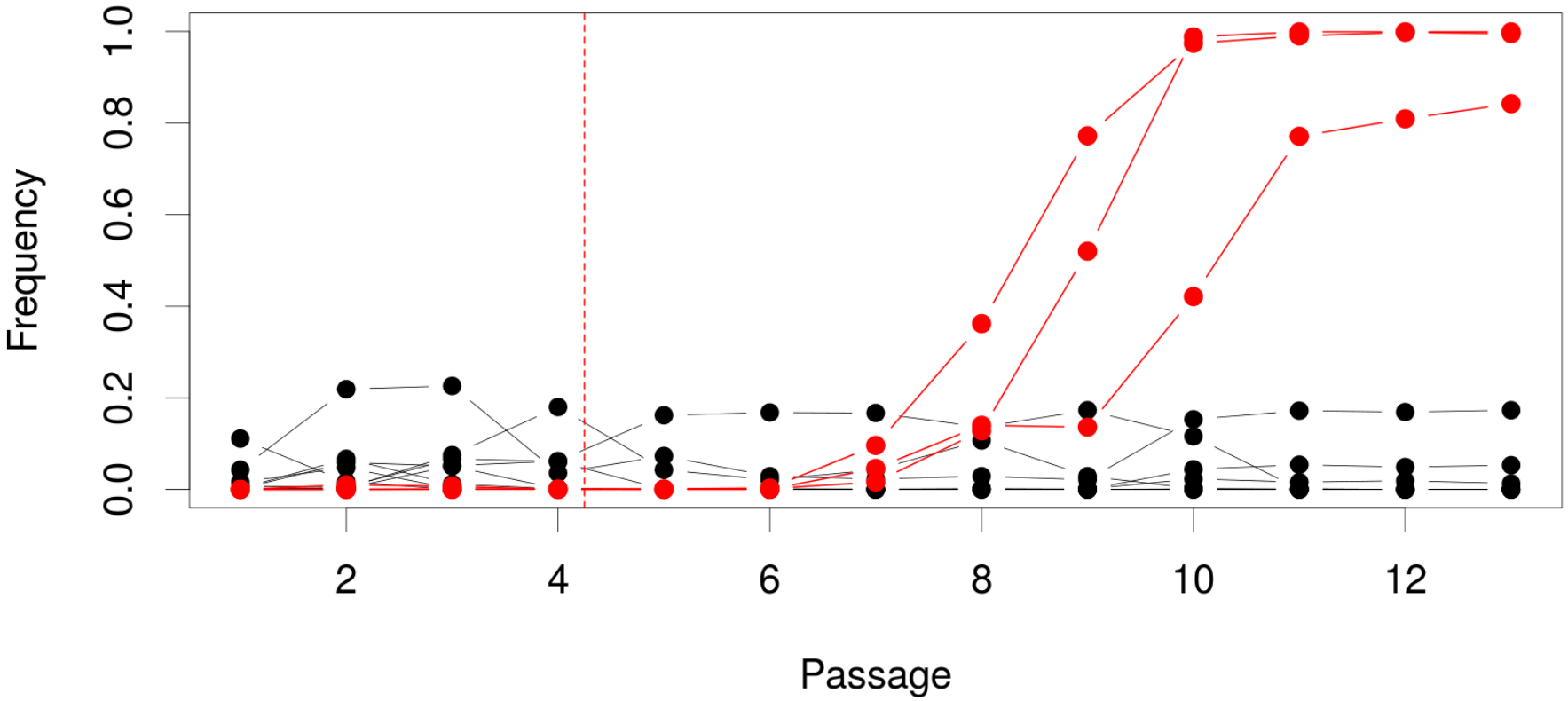


Foll et al. (2014)

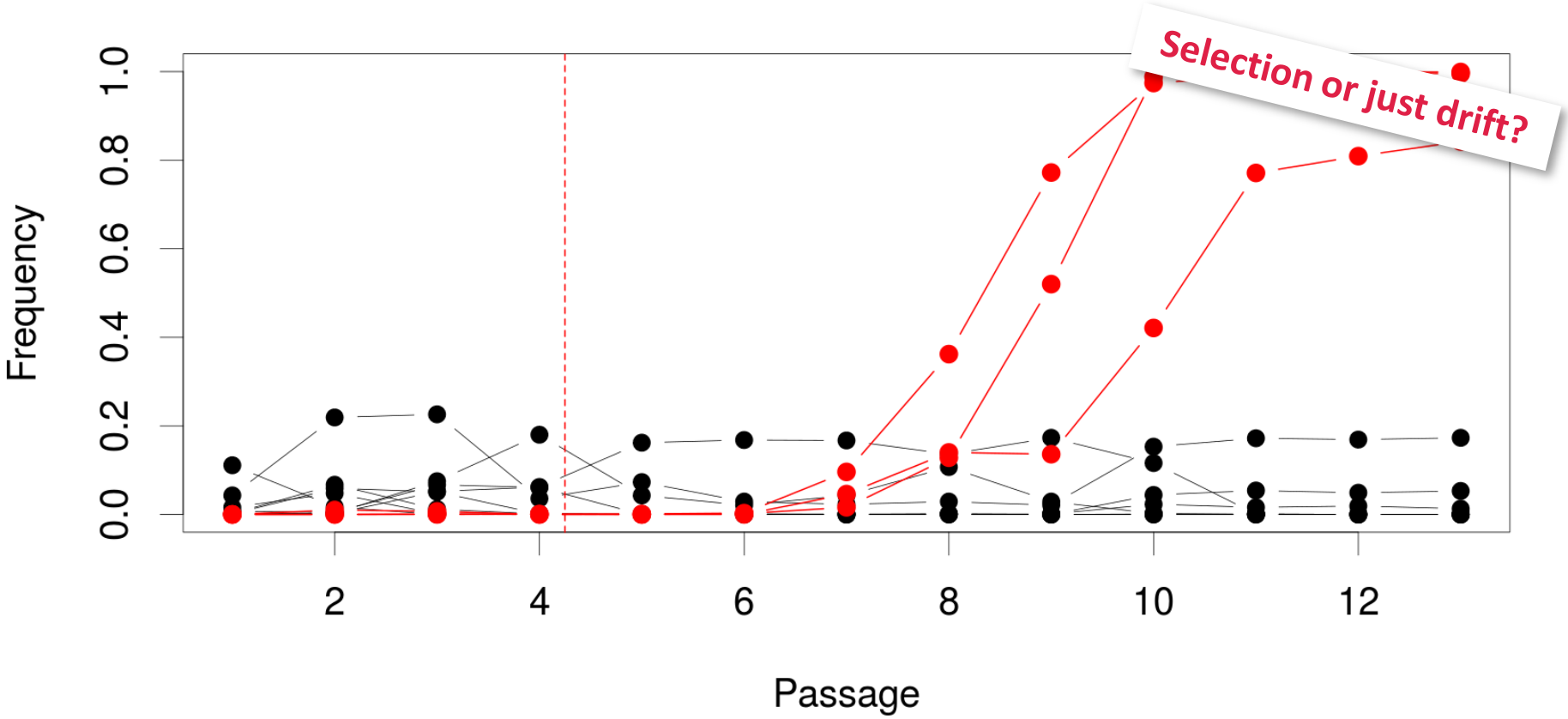
Experimental evolution in the lab



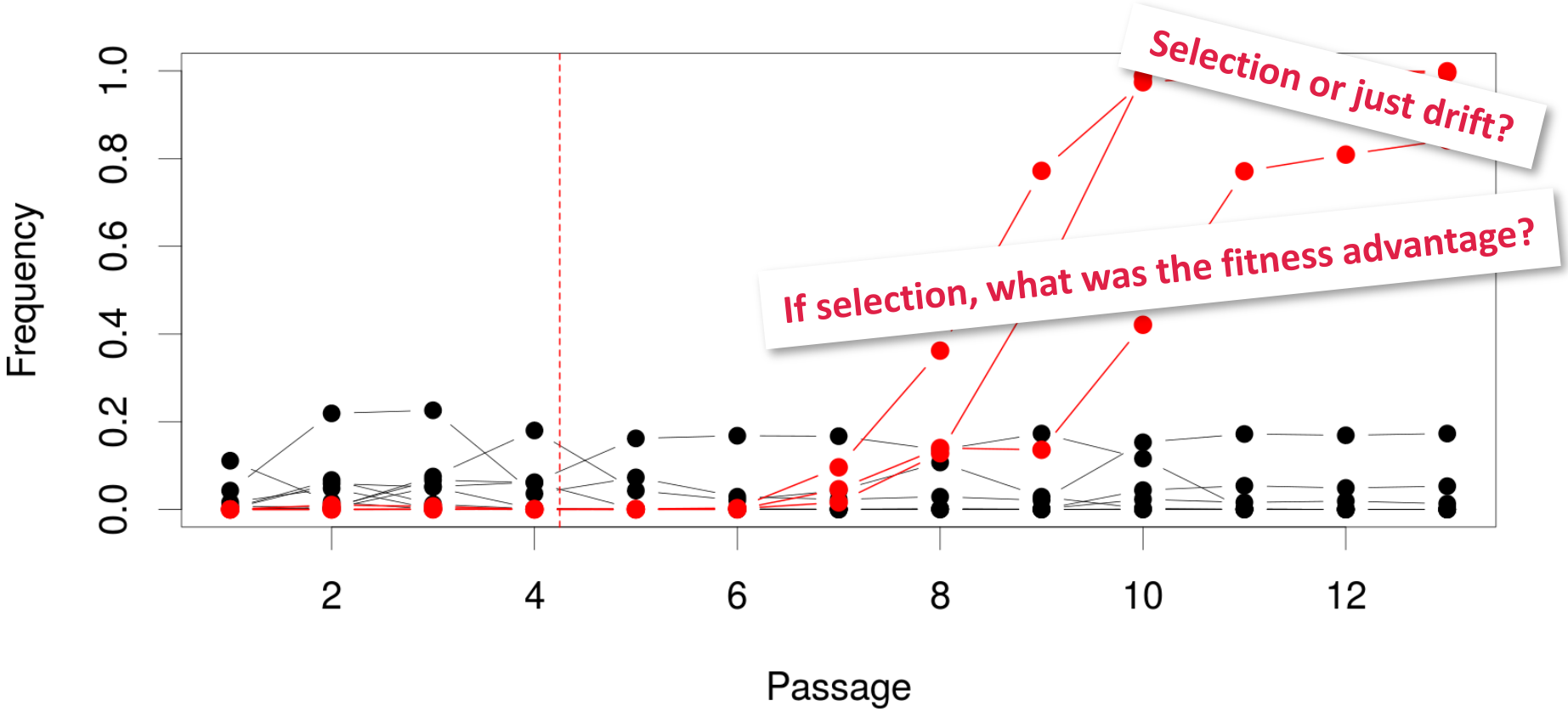
Experimental evolution in the lab



Experimental evolution in the lab

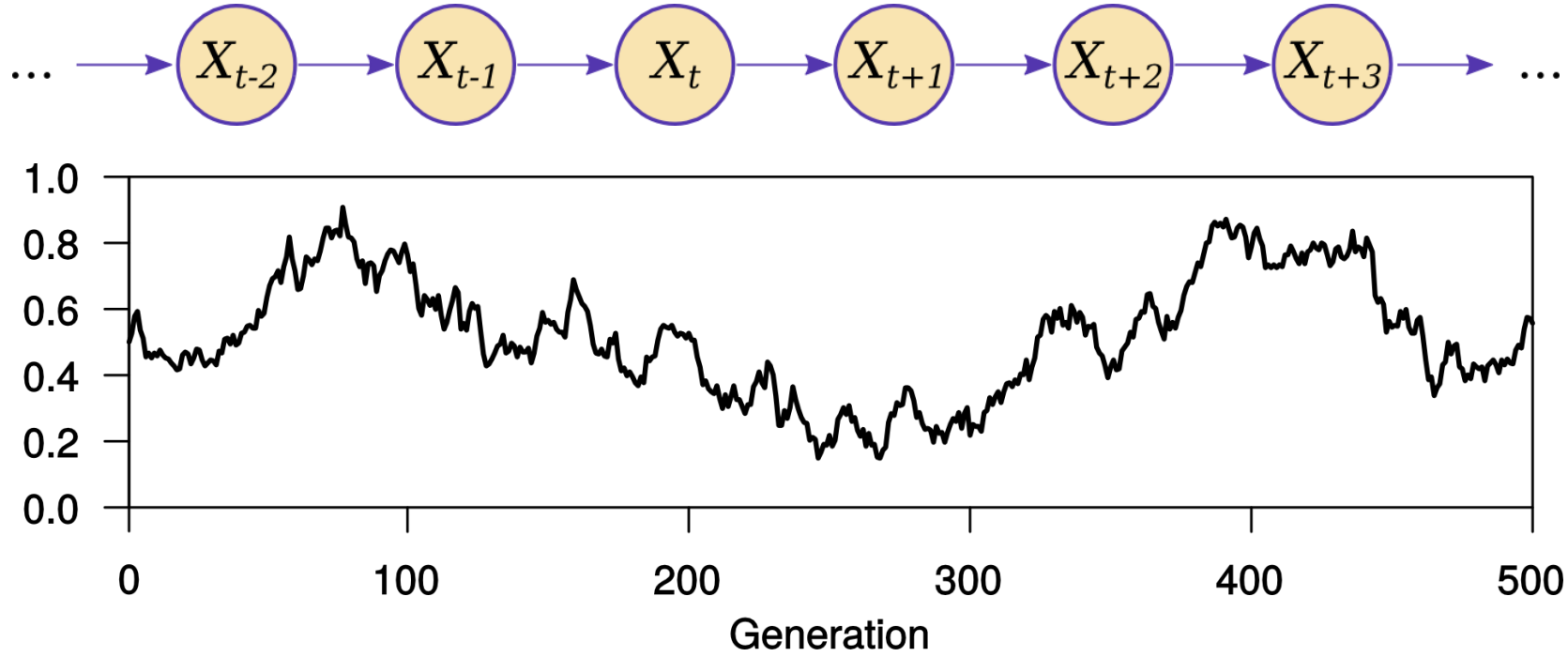


Experimental evolution in the lab



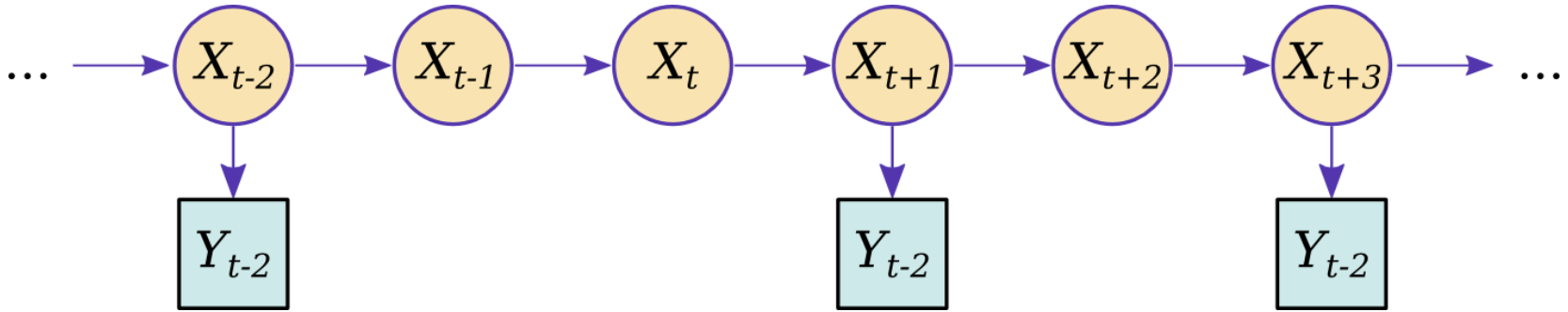
Wright-Fisher Model

- A classic model incorporating **random genetic drift** and **selection**.
- It is a **first order Markov** model with binomial transition probabilities.



Wright-Fisher Model

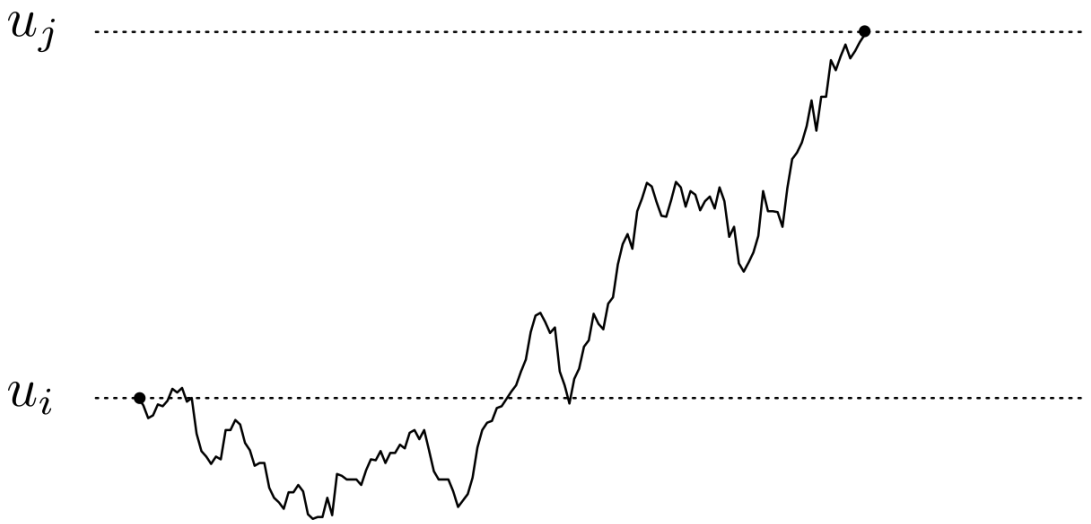
- A classic model incorporating **random genetic drift** and **selection**.
- It is a **first order Markov** model with binomial transition probabilities.



- Time-series samples are naturally modeled as a **hidden Markov model** (HMM) with binomial emission probabilities.
- **Computationally prohibitive** as the forward or backward variable require $(N+1)^2$ multiplications per generation / sample point!

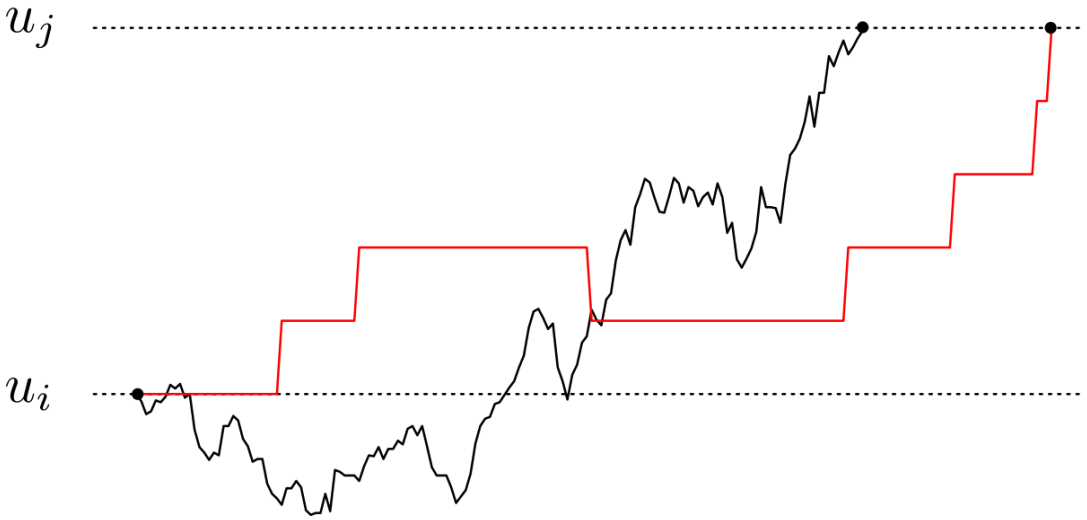
Approximate Wright-Fisher

- To render HMM inference computationally feasible, one would like to have a **coarse grid** on the allele frequencies.
- However, the resulting process is no longer first-order Markovian.



Approximate Wright-Fisher

- We seek to approximate a continous diffusion process ***X*** with a discrete-state Markov process ***U*** that captures the main characteristics.



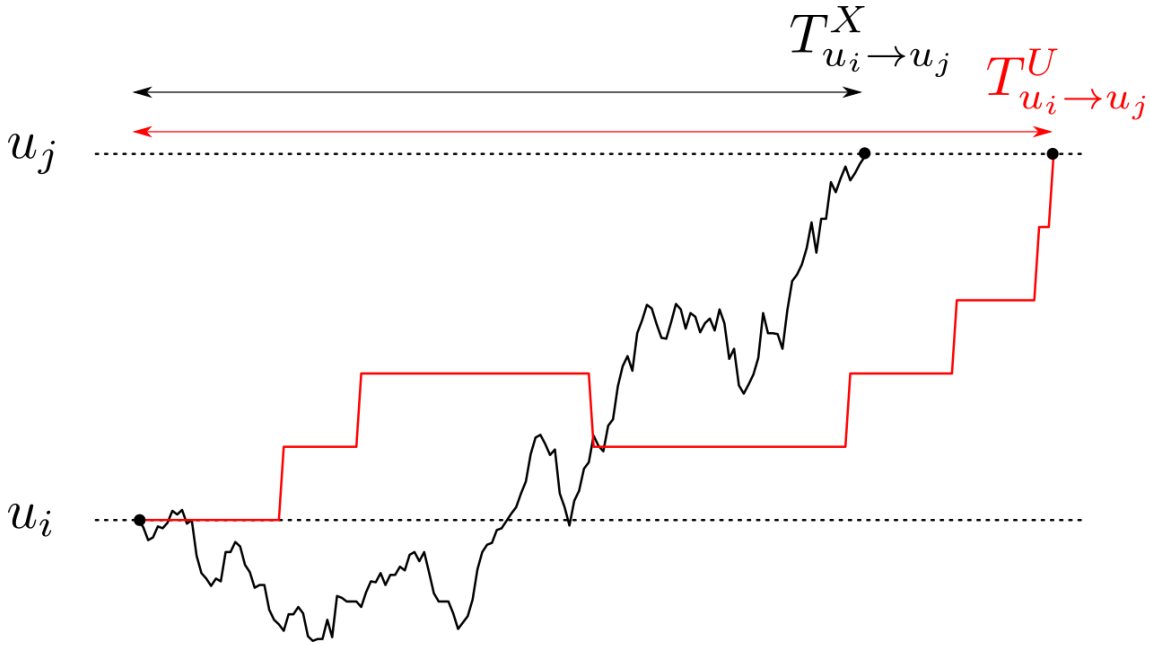
Anna Ferrer Admetlla

Chris Leuenberger

Ferrer-Admettla et al. (2016)

Approximate Wright-Fisher

- We seek to approximate a continous diffusion process ***X*** with a discrete-state Markov process ***U*** that captures the main characteristics.
- **Proposal:** the *mean transition time approximation* $\mathbb{E}\left[T^{U}_{u_i \to u_j}\right] = \mathbb{E}\left[T^{X}_{u_i \to u_j}\right]$



Anna Ferrer Admetlla

Chris Leuenberger

Ferrer-Admettla et al. (2016)

Mean Transition Time Approximation

Approximate the **transition probability** $k \rightarrow k+1$ as

$\mathbb{P}(k+1|k) = \frac{\mathbb{P}(T_{k-1} > T_{k+1})}{\mathbb{E}(T_{k-1} \wedge T_{k+1})}$, where

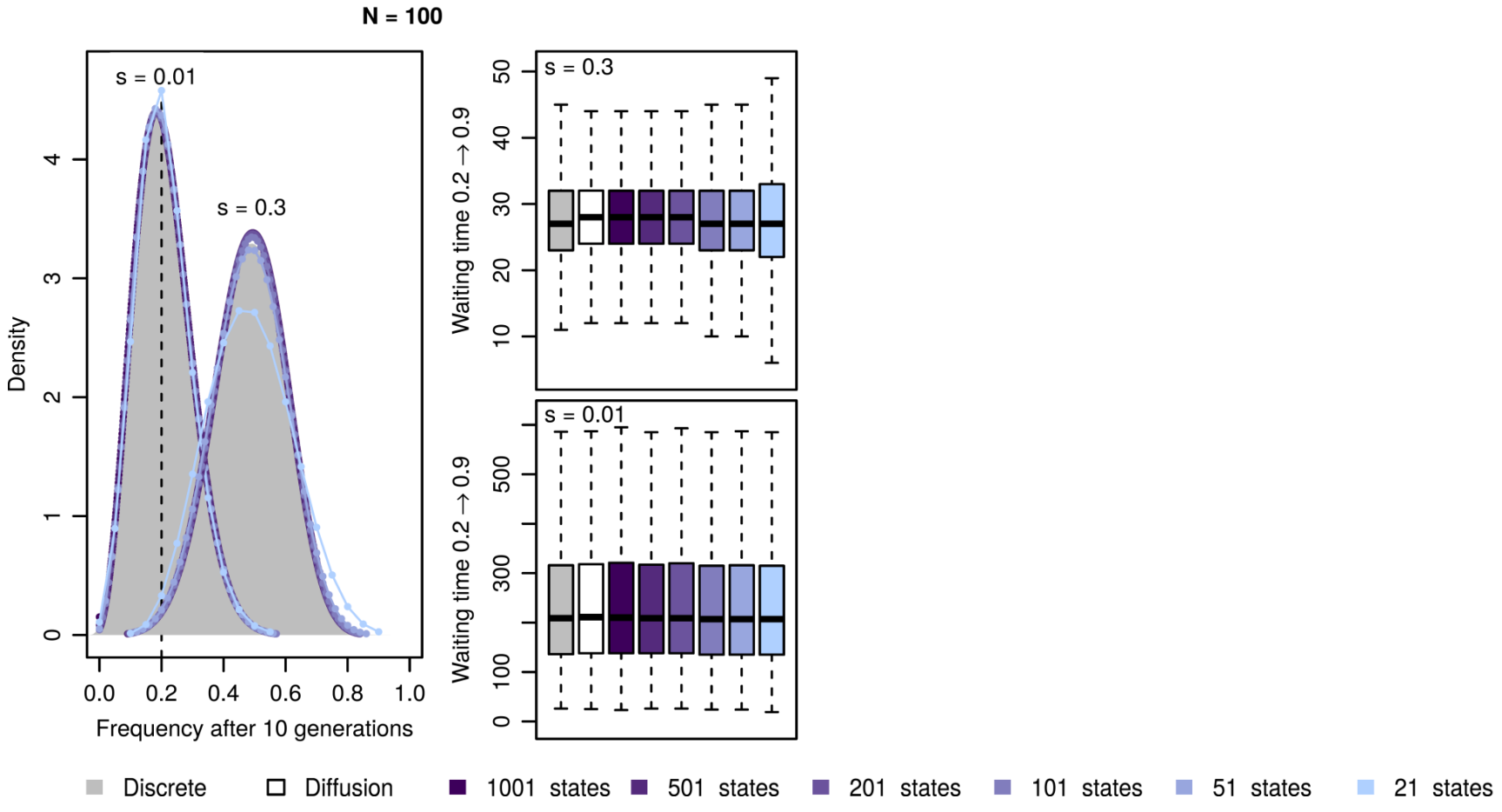
$\mathbb{P}(T_{k-1} > T_{k+1})$: probability the diffusion process exits at the upper limit.

$\mathbb{E}(T_{k-1} \wedge T_{k+1})$: expected time a diffusion process now at u_k remains in (u_{k-1}, u_{k+1}).

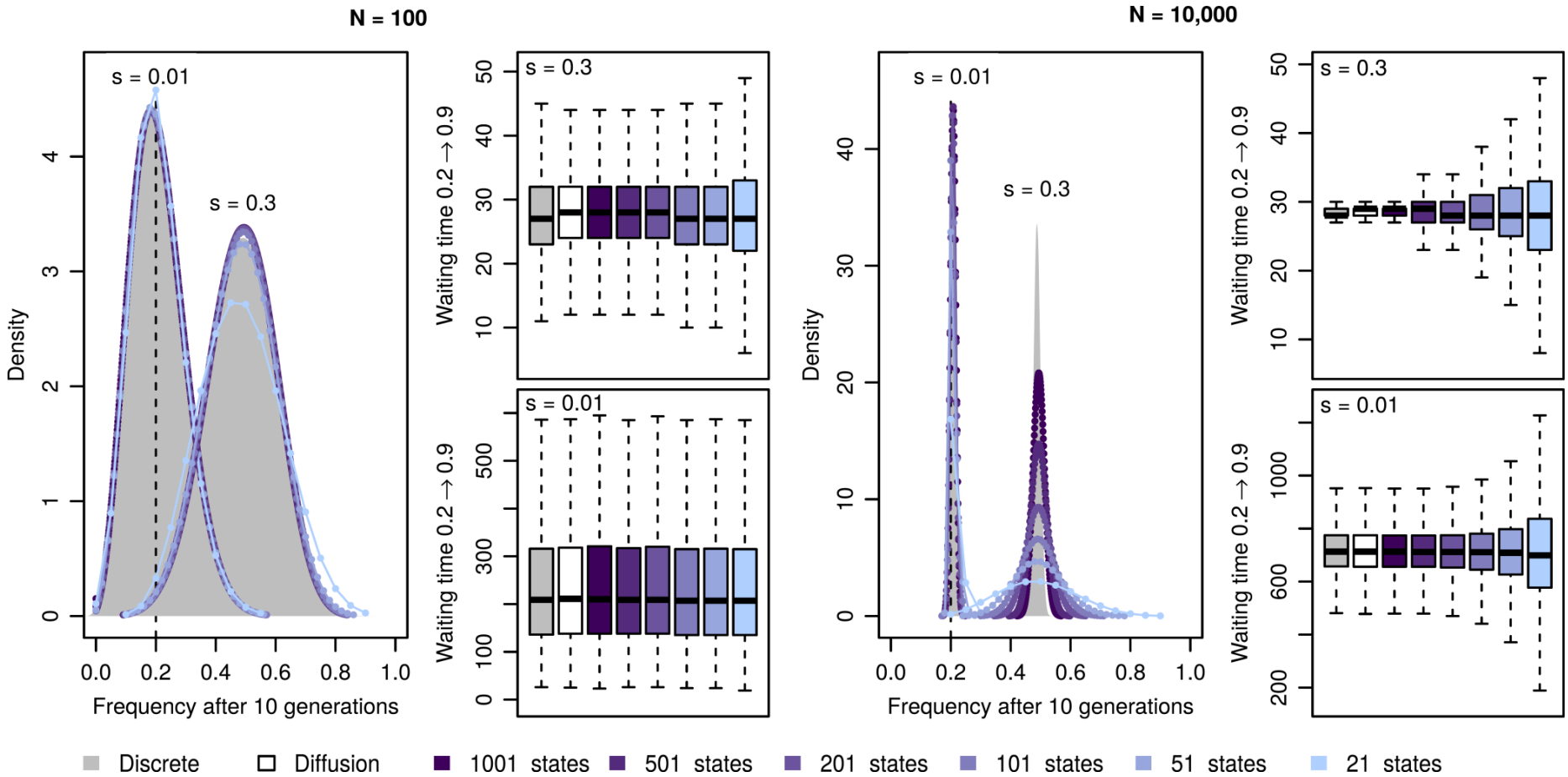
These quantities can be calculated using Green's functions from the diffusion approximation $Lf = \frac{x(1-x)}{4N}\frac{d^2}{dx^2}f + \frac{sx(1-x)}{1+sx}\frac{d}{dx}f$

Ferrer-Admettla et al. (2016)

Mean Transition Time Approximation

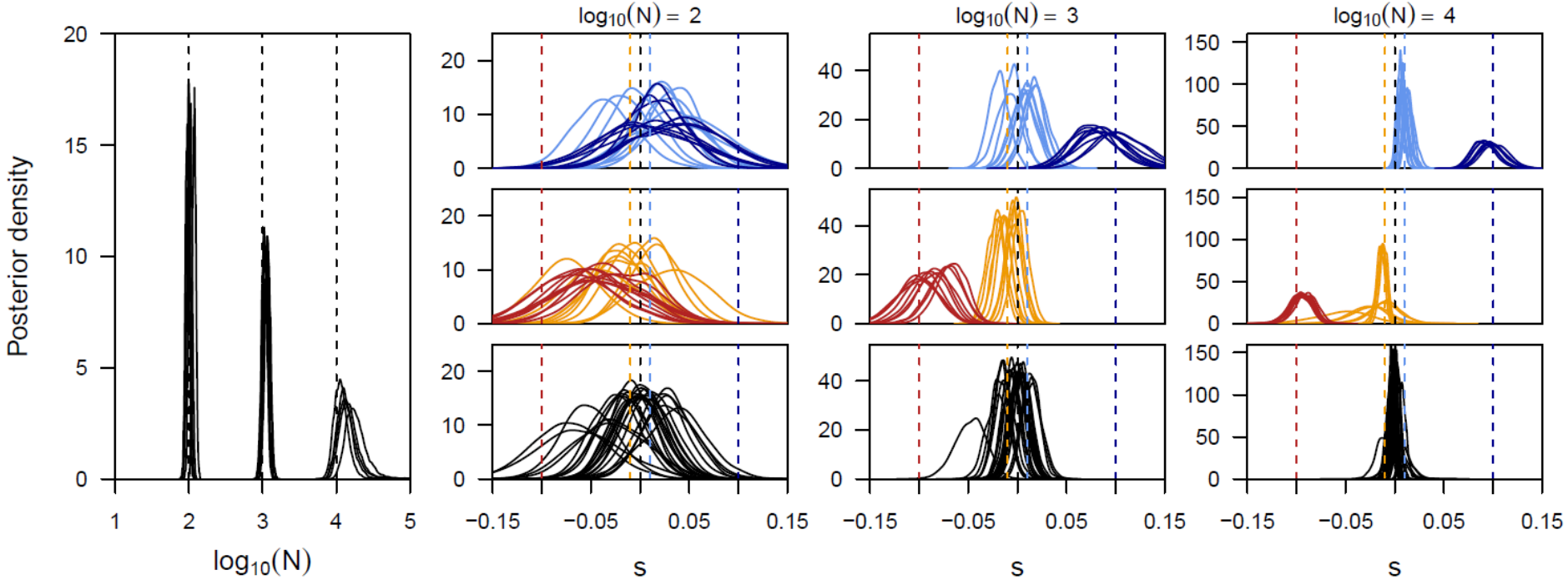


Mean Transition Time Approximation



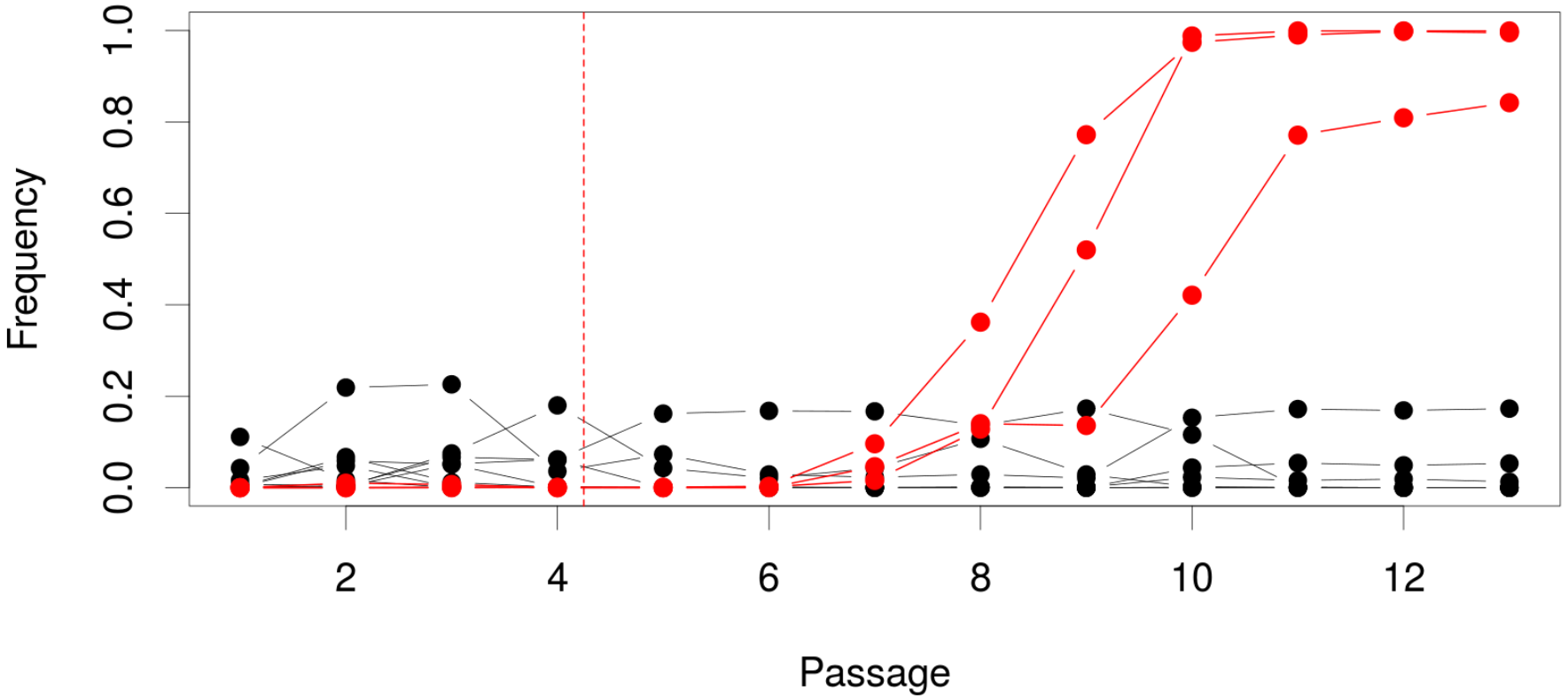
Bayesian inference

- Embed in an **MCMC** scheme to estimate **N** and locus specific **s**.



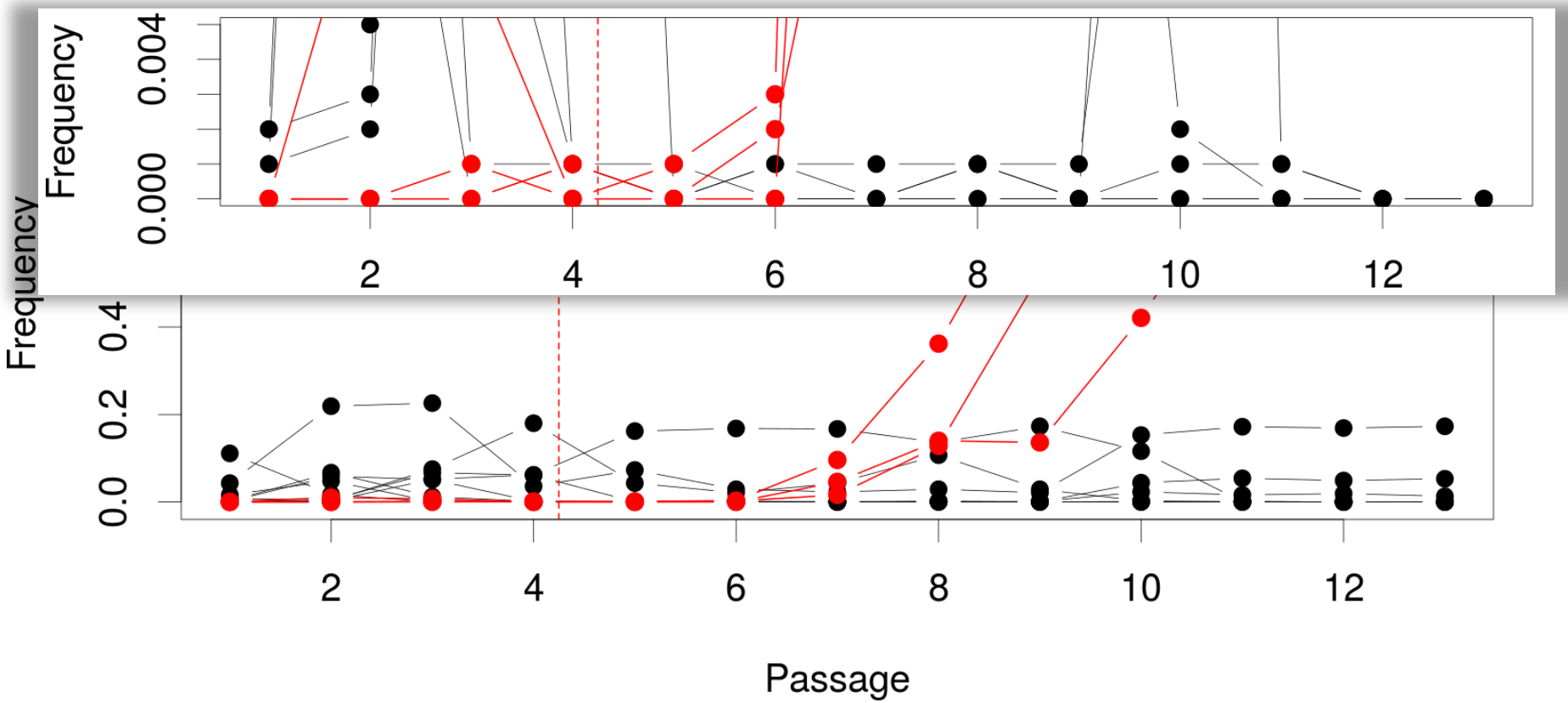
Ferrer-Admettla et al. (2016)

Working with NGS == high error rates



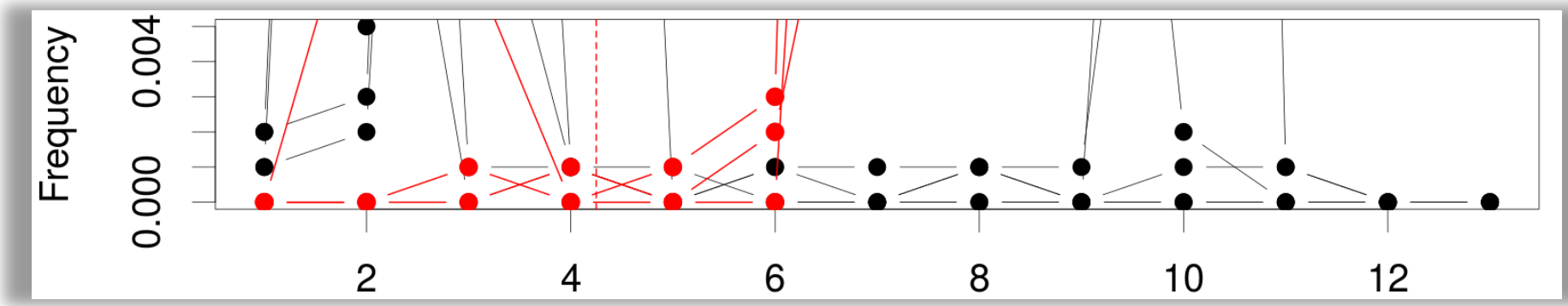
Ferrer-Admettla et al. (2016)

Working with NGS == high error rates



Ferrer-Admettla et al. (2016)

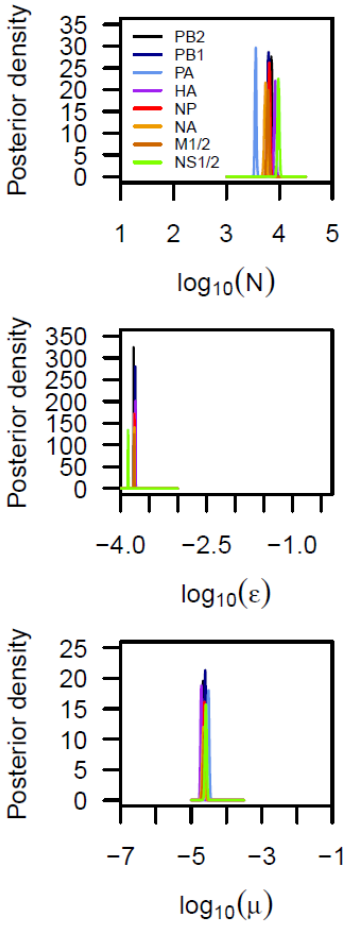
Working with NGS == high error rates

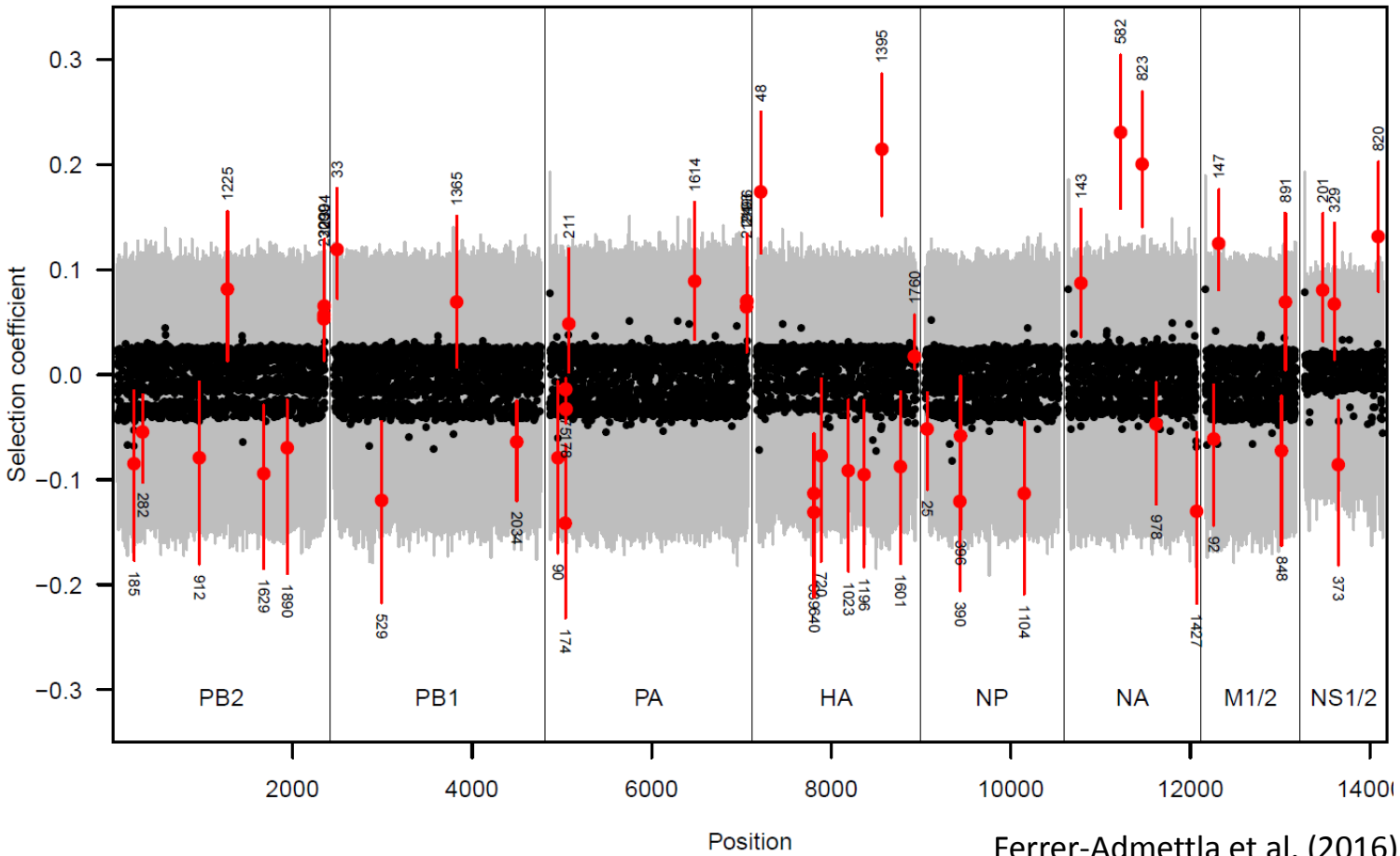


- Mutations staying at **very low but non-zero** frequencies for an extended time are **highly unlikely** given the low effective population size.
- But these patterns can be explained by (a combination of):
  - **Sequencing errors**
  - **Recurrent mutation**

Ferrer-Admettla et al. (2016)

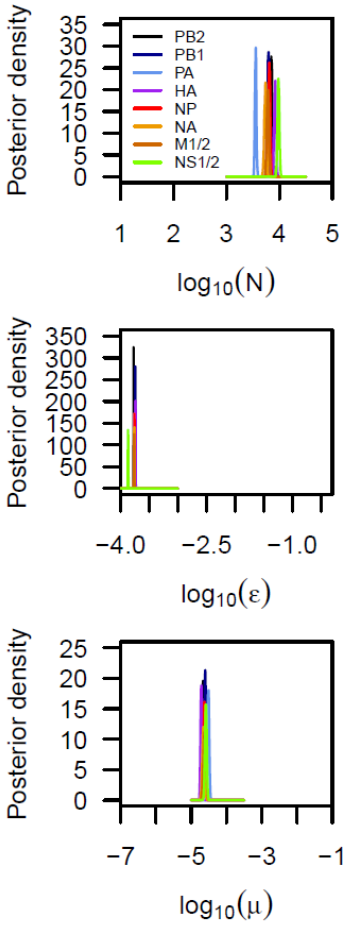
Application to Influenza data

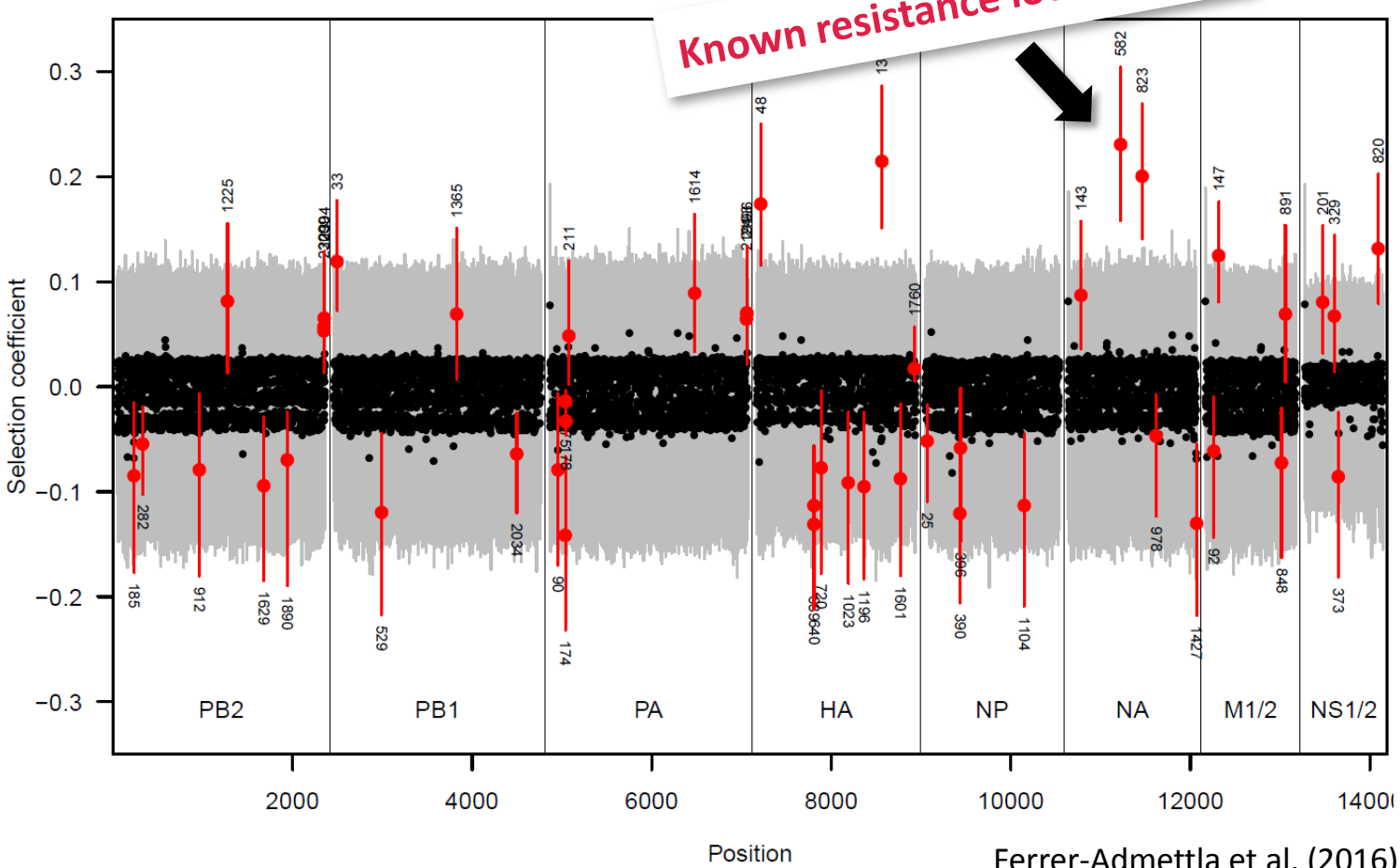




Ferrer-Admettla et al. (2016)

Application to Influenza data





Ferrer-Admettla et al. (2016)

Outline

1. Rejecting a null model using summary statistics
2. Composite-likelihood using coalescent simulations
3. Approximate Bayesian Computation (ABC)
4. Dealing with (ancient) low coverage data
5. Demography & selection from time series data
6. **GWAS (in the unlikely case time will permit)**

Likelihood function

$$\mathbb{P}(\mathbf{d}|\S) = \prod_{l=1}^{L} \mathbb{P}(\mathbf{d}_l|\S) = \prod_{l=1}^{L} \left[\sum_{j=0}^{2k} S_j \mathbb{P}(\mathbf{d}_l|f_l = j) \right],$$

where $\mathbf{d} = \{\mathbf{d}_1, \ldots, \mathbf{d}_L\}$ is a vector of the sequencing data at L loci, $\S$ is the SFS and f_l the sample allele frequency at locus l.

$\mathbb{P}(\mathbf{d}_l|S_j)$ is the *site allele frequency likelihood* calculated from individual data as

$$\mathbb{P}(\mathbf{d}_l|S_j) = \sum_{\mathcal{G}} \left[\mathbb{P}(\mathcal{G}|f_l = j) \prod_{i=1}^{k} \mathbb{P}(d_{li}|\mathcal{G}_i) \right],$$

where the sum runs over all genotype configurations $\mathcal{G}$ and $\mathbb{P}(d_{li}|\mathcal{G}_i)$ are the genotype likelihoods calculated using standard approaches.

- ANGSD implements a numerical algorithm to maximize $\mathbb{P}(\mathbf{d}|\S)$.
- As input ANGSD requires the *site allele frequency likelihoods* directly.

Extending ANGSD to pool-seq data

From pool-seq data, we can calculate the *site frequency likelihoods* as

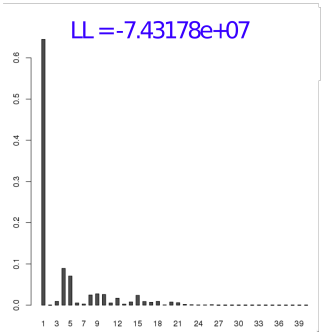
$$\mathbb{P}(\mathbf{d}_l|S_j) = \sum_{r=1}^{R_l} \mathbb{P}(d_{lr}|f_l = j),$$

where the sum runs over all R_l reads at locus l,

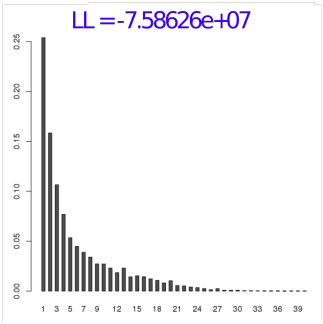
$$\mathbb{P}(d_{lr}|f_l = j) = \begin{cases} f_l(1-\epsilon_{lr}) + (1-f_l)\epsilon_{lr} & \text{if } d_{lr} \text{ is a derived base} \\ (1-f_l)(1-\epsilon_{lr}) + f_l\epsilon_{lr} & \text{if } d_{lr} \text{ is an ancestral base} \end{cases},$$

and ϵ_{lr} is the error rate associated with the specific read.

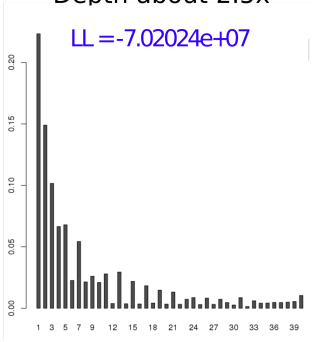
poolStat + ANGSD



MLE Allele frequencies
(VQ < 20 = invariant)

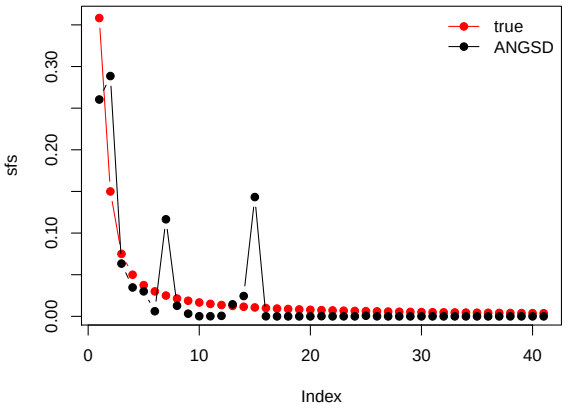


ANGSD on individuals
Depth about 2.5x



- Obviously, the MLE estimates of the SFS are not very helpful.
- Note: this is probably not a bug of ANGSD as the likelihodo does get maximized.
- Note: Issue also when using individual low-coverage data...

Simulated data



- Issue also found with simulated data.
- Likelihood **is** better at inferred SFS!

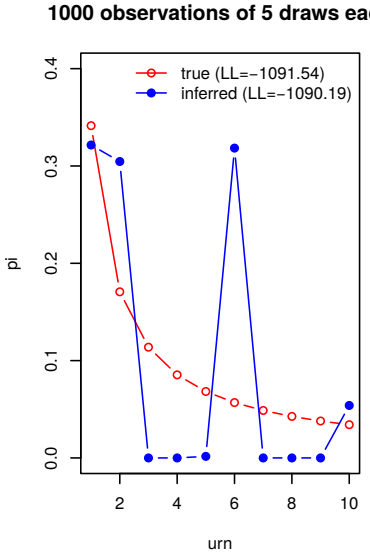
Let's think of urns...

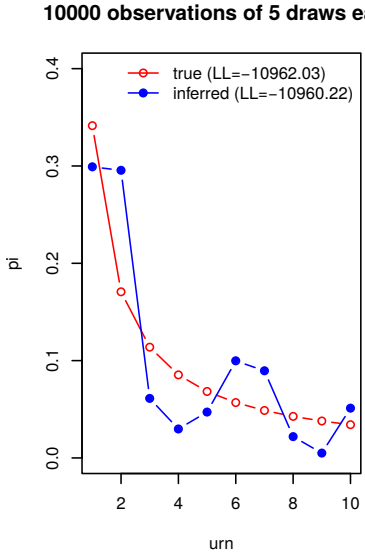
A simple urn example

- Consider $N + 1$ urns that contain N balls each, of which 0 to N are black and the other white.
- In each experiment $l = 1, \ldots, L$ an urn i is randomly chosen with probability $\pi_i, i = 0, \ldots, N$.
- From that urn, m_l balls a drawn with replacement and the number of black balls d_l is recorded.
- The goal is now infer $\boldsymbol{\pi} = \{\pi_0, \ldots, \pi_N\}$ from tis data.

This can easily be solved using an EM algorithm.

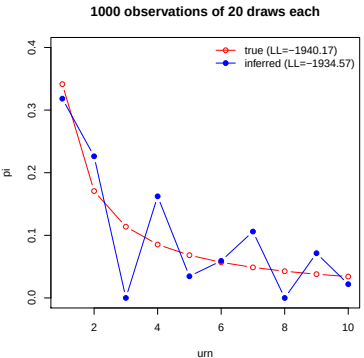
Some simulations

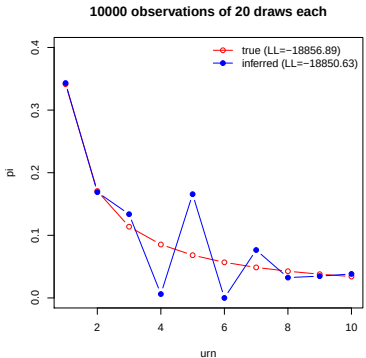


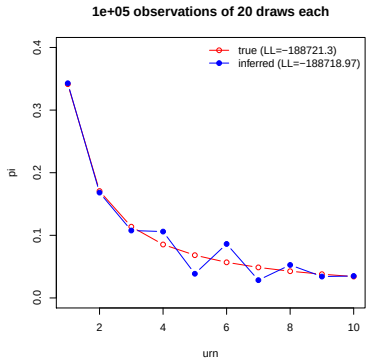


- Again: the MLE is just not a helpful estimator!

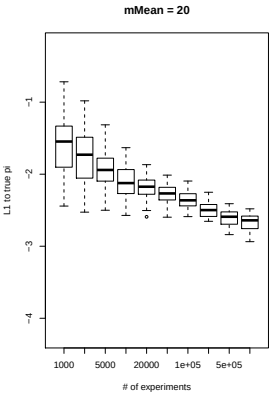
Some simulations

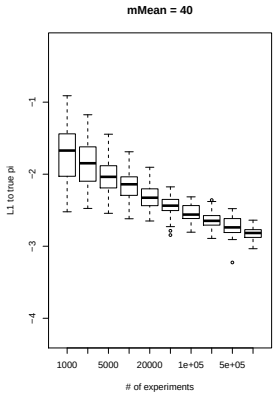


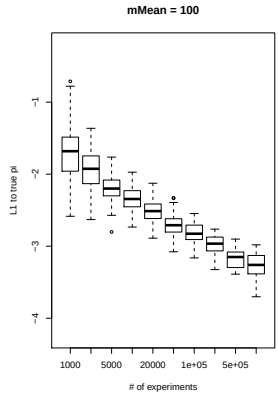


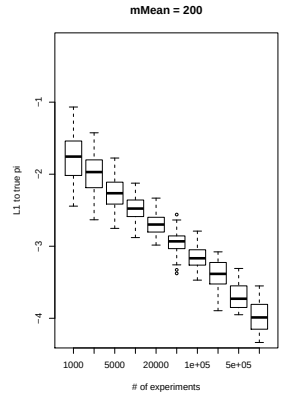


Some simulations









Proposed solution

- Since the data seems not to be sufficient to accurately infer all π_i individually, we may need to make additional model assumptions.
- For this we propose the mixture model:

$$\mathbb{P}(d_l|\boldsymbol{\pi}, \alpha, \beta, m_l) = \pi_0 \mathbb{P}_0(d_l) + \pi_1 \mathbb{P}_1(d_l|\alpha, \beta, m_l) + \pi_2 \mathbb{P}_2(d_l), \pi_0 + \pi_1 + \pi_2 = 1,$$

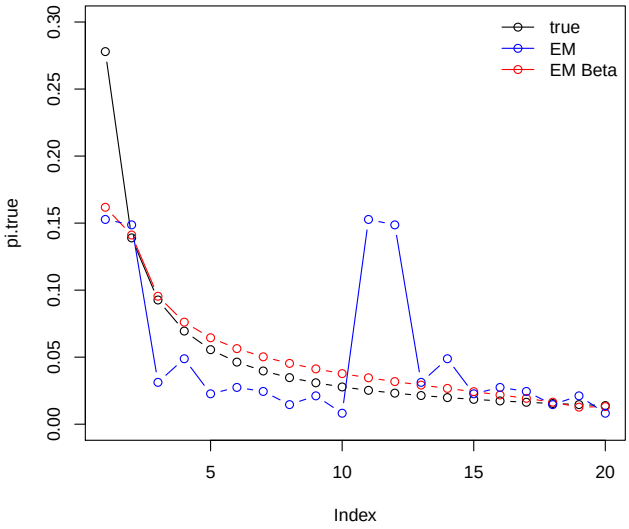
where

$$\mathbb{P}_0(d_l) = \begin{cases} 1 & \text{if } d_l = 0 \\ 0 & \text{if } d_l > 0 \end{cases}, \qquad \mathbb{P}_2(d_l) = \begin{cases} 1 & \text{if } d_l = m_l \\ 0 & \text{if } d_l < m_l \end{cases},$$

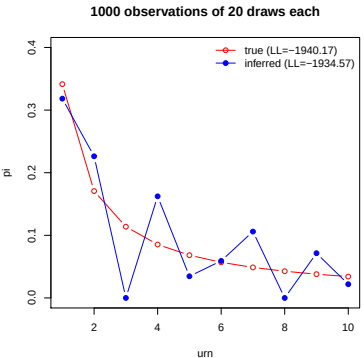
and $\mathbb{P}_1(d_l|\alpha, \beta, m_l) = \mathrm{B}(d_l|\alpha, \beta, m_l)$ is following the Beta distribution.

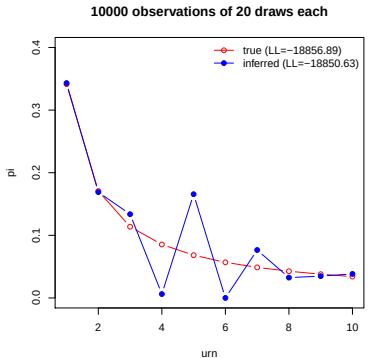
Some first trials...

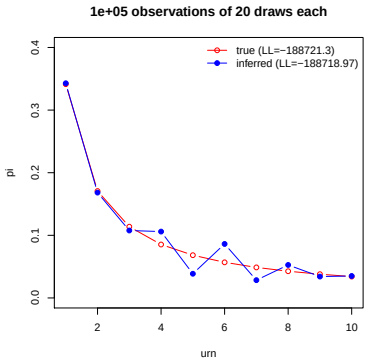
Simulations of 10,000 sites, stupid optimization algorithm.



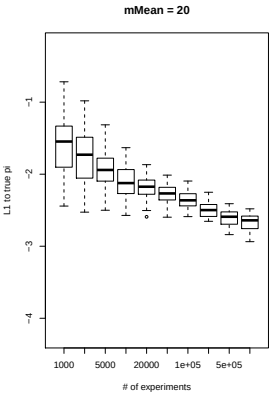
Some simulations

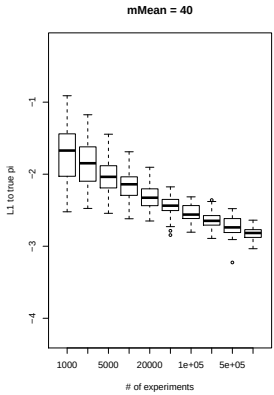


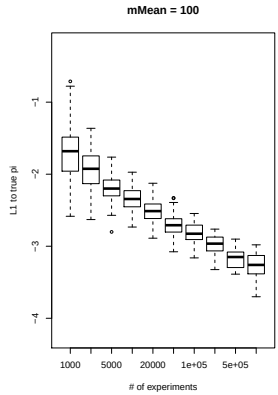


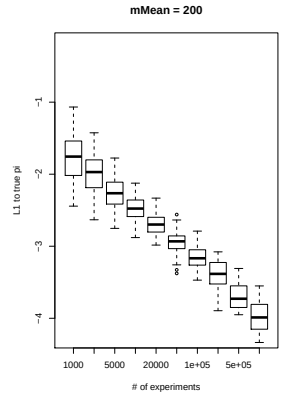


Some simulations









Simple error model

Observed / true	0	1	2
0	$(1-\epsilon)^2$	$2\epsilon(1-\epsilon)$	ϵ^2
1	$\varepsilon(1-\varepsilon)$	$(1-\varepsilon)^2+\varepsilon^2$	$\varepsilon(1-\varepsilon)$
2	ϵ^2	$2\epsilon(1-\epsilon)$	$(1-\epsilon)^2$

The likelihood of this model is given by

$$\mathbb{P}(\mathbf{g}|\epsilon) = \prod_{i=1}^{I}\prod_{l=1}^{L}\sum_{g=0}^{2}\mathbb{P}(\gamma_{il}=g)\prod_{j=1}^{r_i}\mathbb{P}(g_{il}^{(j)}|\gamma_{il}=g,\epsilon_{c_{il}^{(j)}}),$$

where $\mathbf{g}=\{\mathbf{g}_1,\ldots,\mathbf{g}_I\}$, $\mathbb{P}(\gamma_{il}=g)=f_{ig}$ is the frequency of genotype g among all sites typed in individual i.

The model is learned using an EM algorithm we implemented in the software *Tiger*.

Simple error model

Observed / true	0	1	2
0	$(1-\epsilon)^2$	$2\epsilon(1-\epsilon)$	ϵ^2
1	$\varepsilon(1-\varepsilon)$	$(1-\varepsilon)^2+\varepsilon^2$	$\varepsilon(1-\varepsilon)$
2	ϵ^2	$2\epsilon(1-\epsilon)$	$(1-\epsilon)^2$

The likelihood of this model is given by

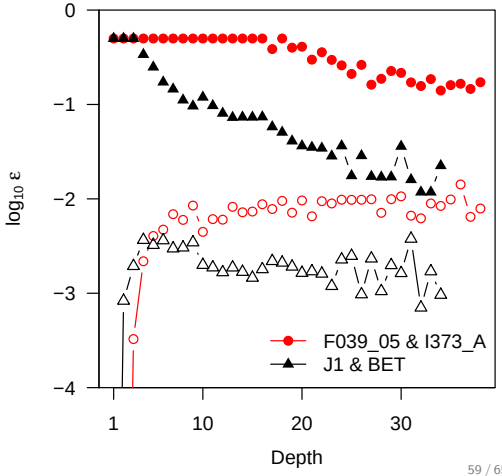
$$\mathbb{P}(\mathbf{g}|\epsilon) = \prod_{i=1}^{I}\prod_{l=1}^{L}\sum_{g=0}^{2}\mathbb{P}(\gamma_{il}=g)\prod_{j=1}^{r_i}\mathbb{P}(g_{il}^{(j)}|\gamma_{il}=g,\epsilon_{c_{il}^{(j)}}),$$

where $\mathbf{g}=\{\mathbf{g}_1,\ldots,\mathbf{g}_I\}$, $\mathbb{P}(\gamma_{il}=g)=f_{ig}$ is the frequency of genotype g among all sites typed in individual i.

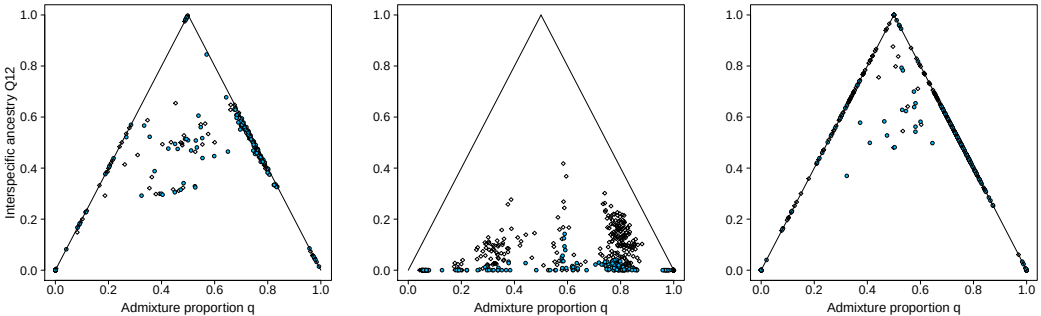
The model is learned using an EM algorithm we implemented in the software *Tiger*.

Application to RAD data of Poplars

Data generated by Floragenex©!!!

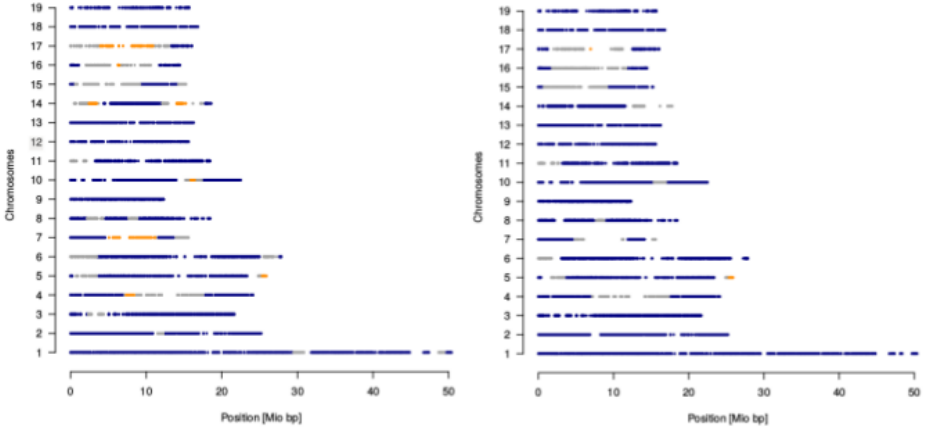


Effect on ancestry inference (Entropy)



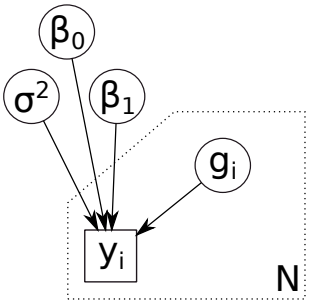
Truth partially recovered after correcting genotype likelihoods.

Effect on ancestry inference (RASPberry)



Truth partially recovered after correcting genotype likelihoods.

Taking genotype uncertainty into account

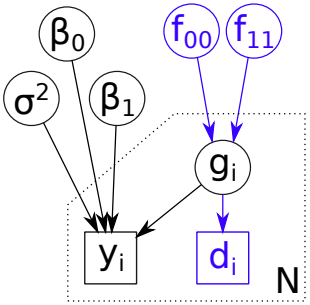


Likelihood if genotype is known:

$$\mathbb{P}(\mathbf{y}|\beta_0, \beta_1, \sigma^2, \mathbf{g}) = \prod_{i=1}^{N} \mathbb{P}(d_i|\beta_0, \beta_1, \sigma^2, g_i)$$

where

$$\mathbb{P}(d_i|\beta_0, \beta_1, \sigma^2, g_i) = \Phi(d_i; \beta_0 + \beta_1 g, \sigma^2)$$



Likelihood if genotype is known:

$$\mathbb{P}(\mathbf{y}|\beta_0, \beta_1, \sigma^2, \mathbf{g}) = \prod_{i=1}^{N} \mathbb{P}(d_i|\beta_0, \beta_1, \sigma^2, g_i)$$

where

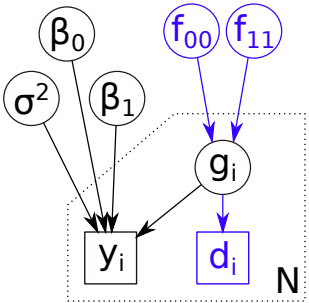
$$\mathbb{P}(d_i|\beta_0, \beta_1, \sigma^2, g_i) = \Phi(d_i; \beta_0 + \beta_1 g, \sigma^2)$$

Vivian Link

Likelihood if genotype is unknown:

$$\mathbb{P}(\mathbf{d}, \mathbf{y}|\beta_0, \beta_1, \sigma^2, f) = \prod_{i=1}^{N} \sum_{g} \mathbb{P}(y_i|\beta_0, \beta_1, \sigma^2, g_i)\mathbb{P}(d_i|g_i)\mathbb{P}(g_i|f)$$

$\rightarrow$ EM algorithm



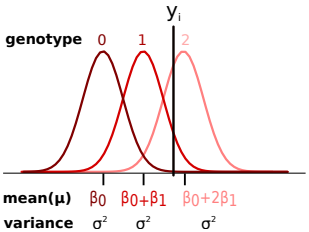
Likelihood if genotype is known:

$$\mathbb{P}(\mathbf{y}|\beta_0, \beta_1, \sigma^2, \mathbf{g}) = \prod_{i=1}^{N} \mathbb{P}(d_i|\beta_0, \beta_1, \sigma^2, g_i)$$

where

$$\mathbb{P}(d_i|\beta_0, \beta_1, \sigma^2, g_i) = \Phi(d_i; \beta_0 + \beta_1 g, \sigma^2)$$

Vivian Link

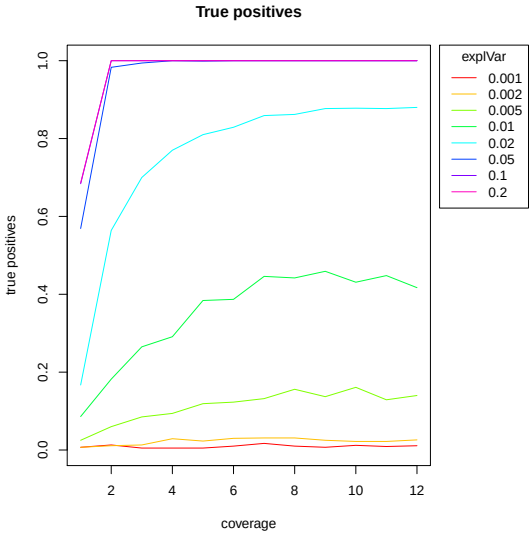


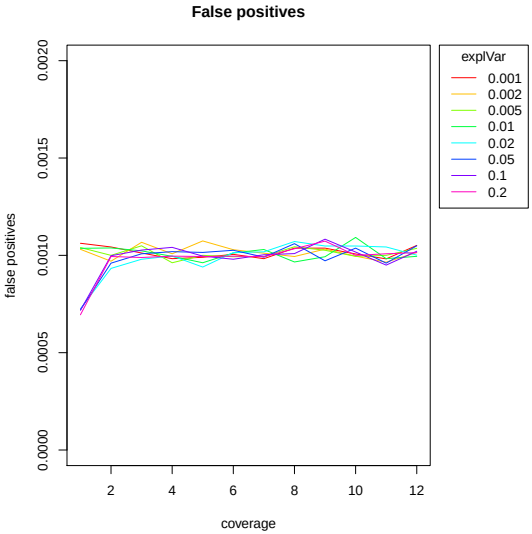
Likelihood if genotype is unknown:

$$\mathbb{P}(\mathbf{d}, \mathbf{y}|\beta_0, \beta_1, \sigma^2, f) = \prod_{i=1}^{N} \sum_{g} \mathbb{P}(y_i|\beta_0, \beta_1, \sigma^2, g_i)\mathbb{P}(d_i|g_i)\mathbb{P}(g_i|f)$$

$\rightarrow$ EM algorithm

Does it work?


True positives
true positives
coverage
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
12
explVar
0.001
0.002
0.005
0.01
0.02
0.05
0.1
0.2


False positives
false positives
coverage
0.0000
0.0005
0.0010
0.0015
0.0020
2
4
6
8
10
12
explVar
0.001
0.002
0.005
0.01
0.02
0.05
0.1
0.2

Summary & Outlook

- 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.

Summary & Outlook

- 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.

- Sequencing errors are an unavoidable part of NGS data - deal with it!
- Filtering introduces biases, generally towards too little diversity.

➡ **avoid calling genotypes** and use tools that **integrate over genotype uncertainty**.

- Such methods also harvest the power of low coverage sequencing, which allows us to invest in biological rather than technical replicates.

FNSNF
Swiss National Science Foundation