Lecture 2.3:
Demographic priors and model selection

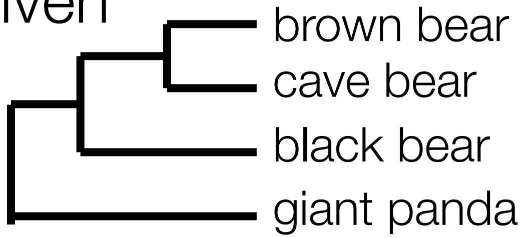
Likelihood revisited

Likelihood of hypothesis H =

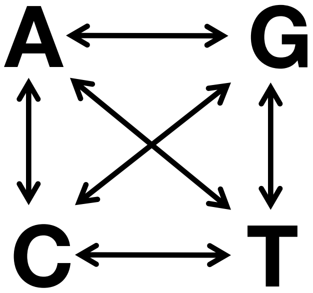
$$\boxed{P(D \mid H)}$$

the probability of the data, given the hypothesis

Given



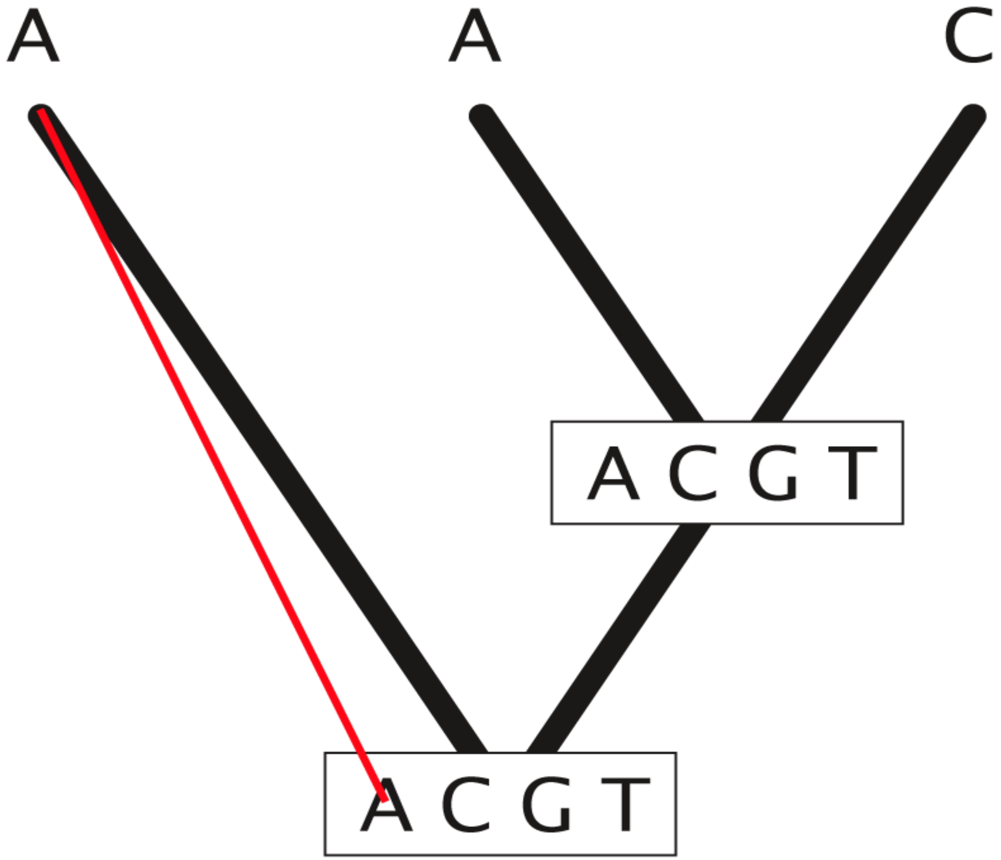
+



Probability of?

Brown bear	CGTTAGTACACT
Cave bear	CGATAGTTCACT
Black bear	CGTTAGTTTACC
Giant panda	CATTGGTTTACT

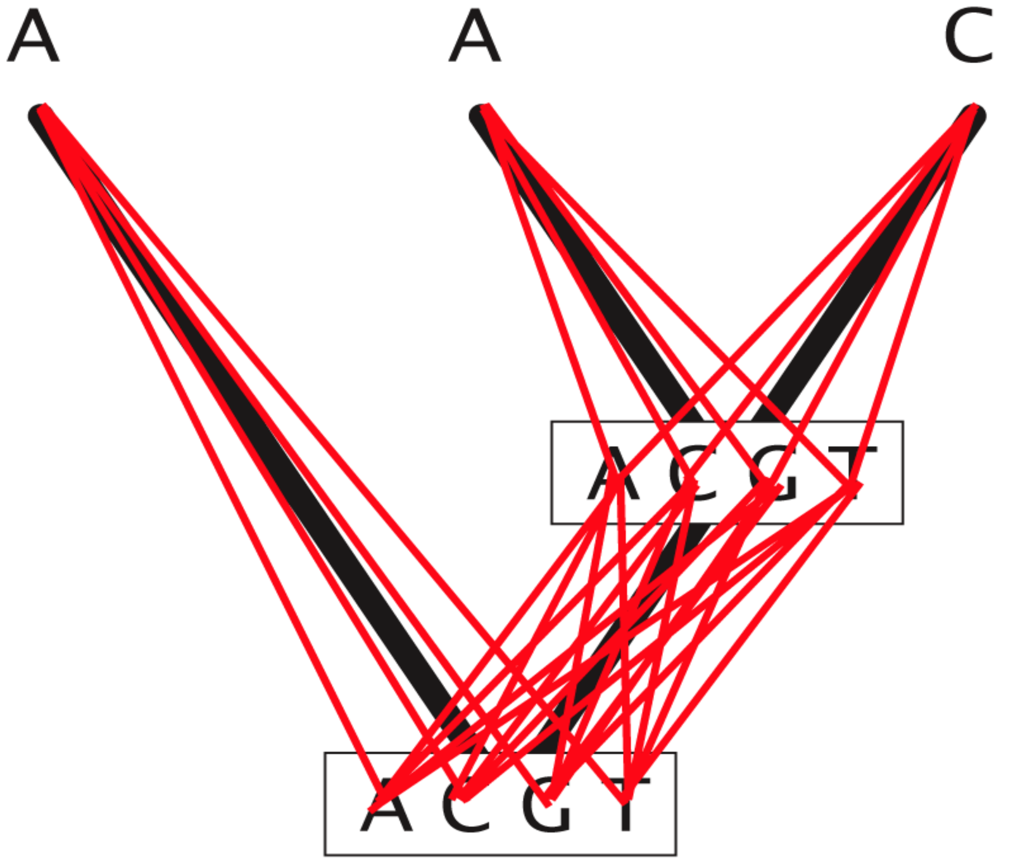
Likelihood revisited



to

		A	C	G	T
from	A	–	a	b	c
	C	a	–	d	e
	G	b	d	–	f
	T	c	e	f	–

Likelihood revisited



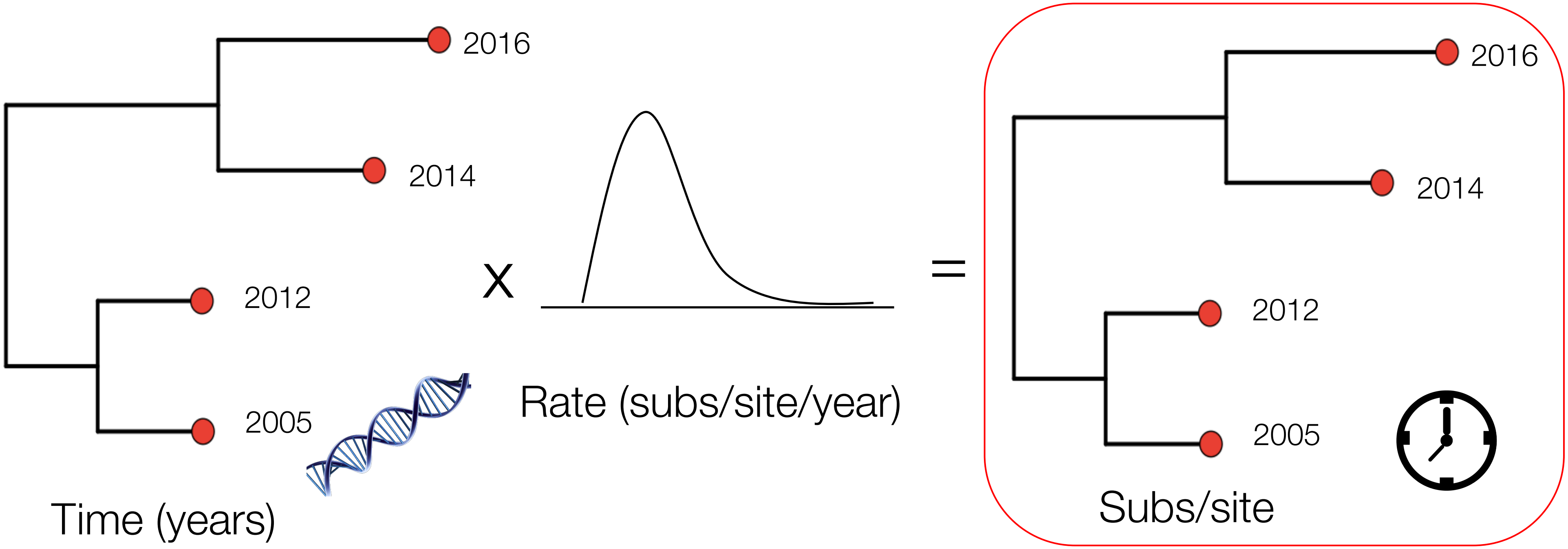
to

from	A	C	G	T
A	–	a	b	c
C	a	–	d	e
G	b	d	–	f
T	c	e	f	–

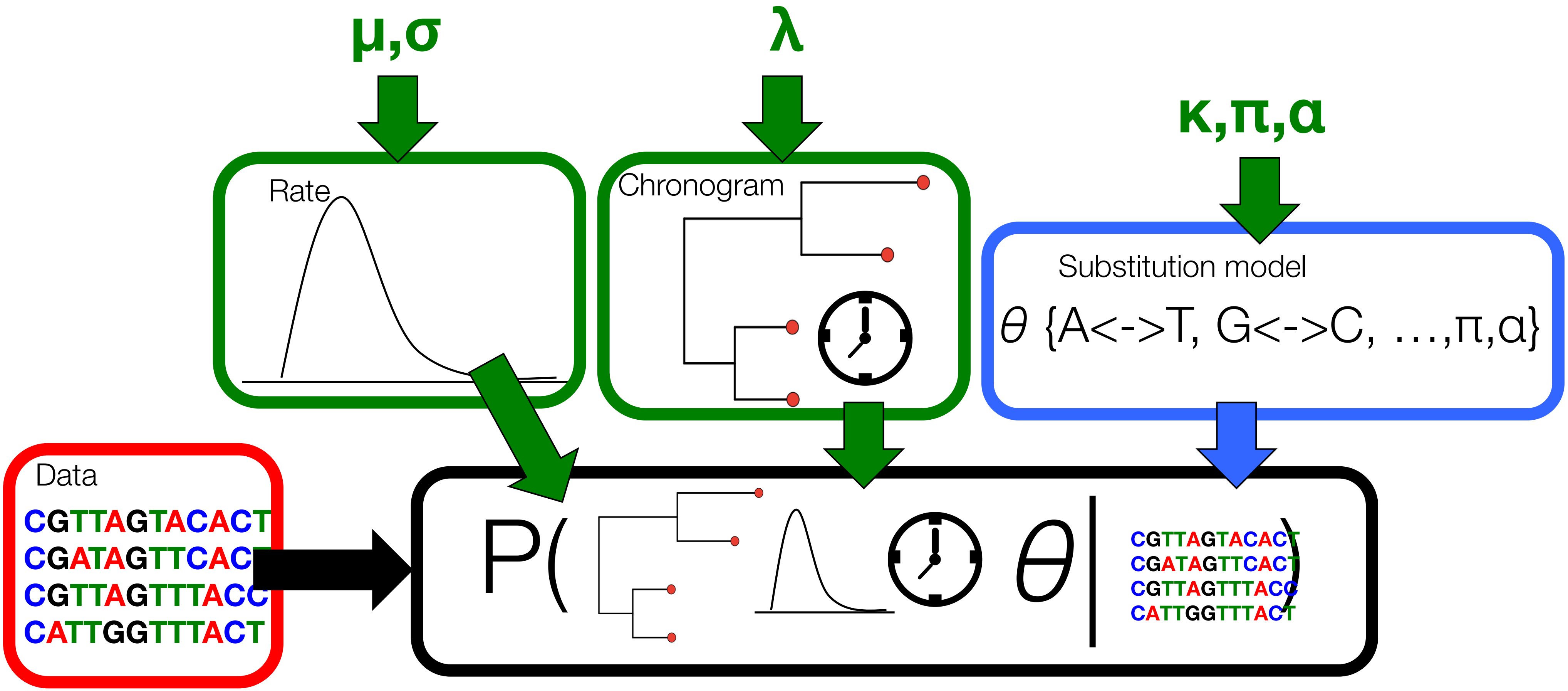
Likelihood = sum of all possible scenarios

The Phylogenetic hierarchical model

- Bayesian molecular clocks estimate rates and chronograms.
    - Treat substitutions as the product of rates and times



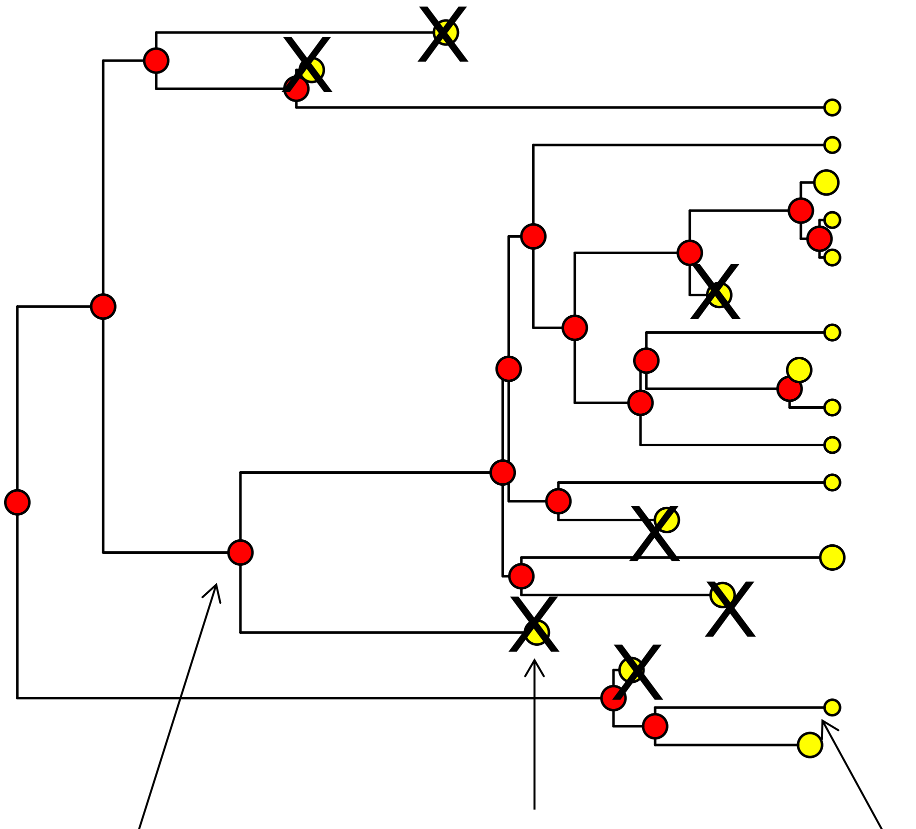
The Phylogenetic hierarchical model



To specify a tree prior, we typically use a branching process

Birth-Death processes

Branching process



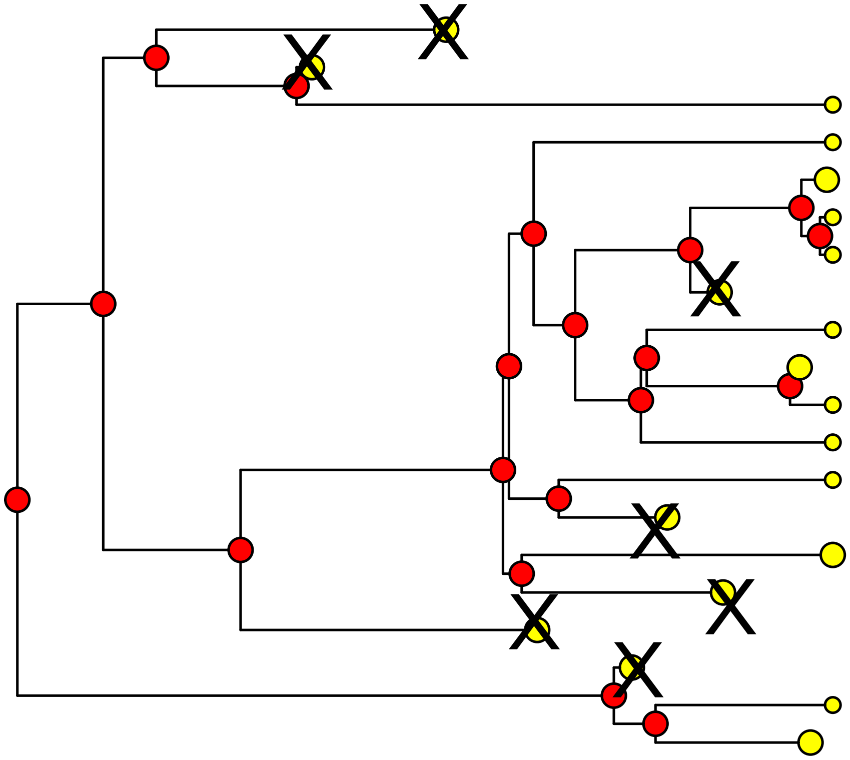
- Model speciation processes
- Branch lengths in units of time (i.e. chronograms)

Branching/S
peciation

Death/Extin
ction

Species/Sample

Birth-Death models



🔴 Speciation/Birth rate (λ)

X Death/Extinction rate (μ)

🟡 Sampling probability of extant species (ρ)

These parameters are not identifiable. We typically fix *ρ.*

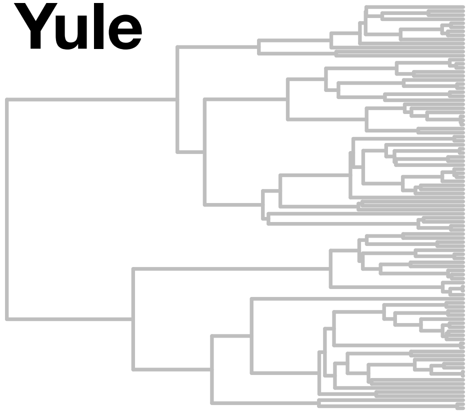
Lineages through time

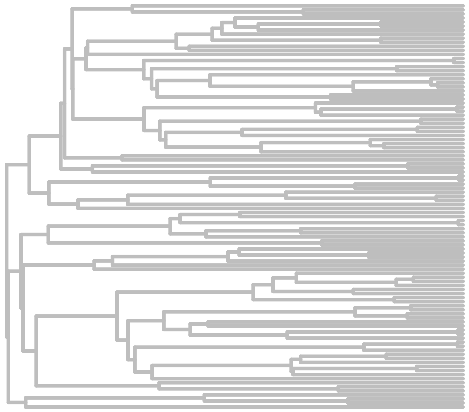
Pure Birth / Yule

$\lambda = 0.8$

$\mu = 0$

$\rho = 1$



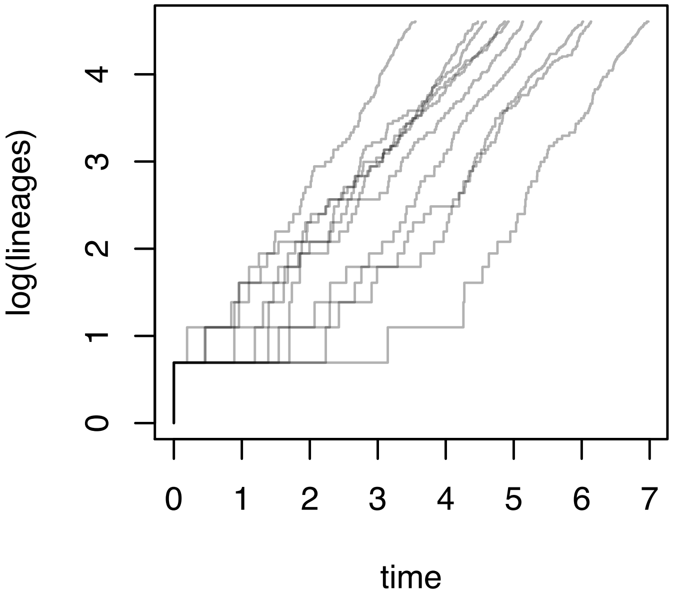


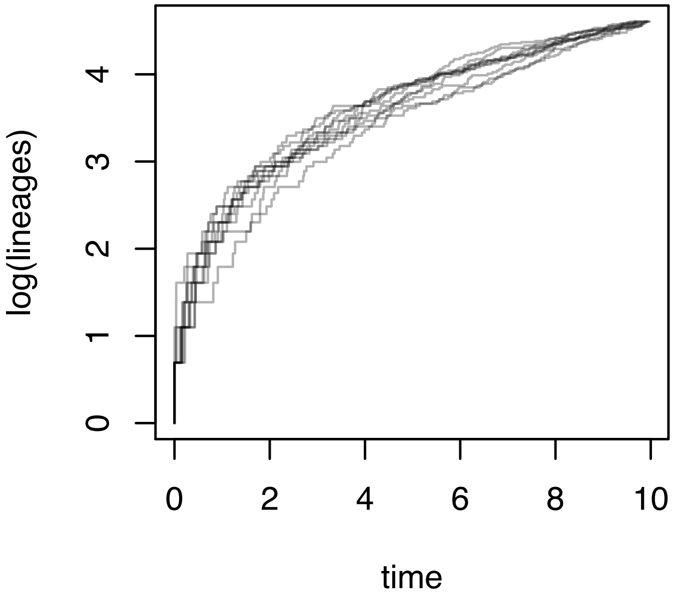
Birth-Death

$\lambda = 1$

$\mu = 0.8$

$\rho = 1$

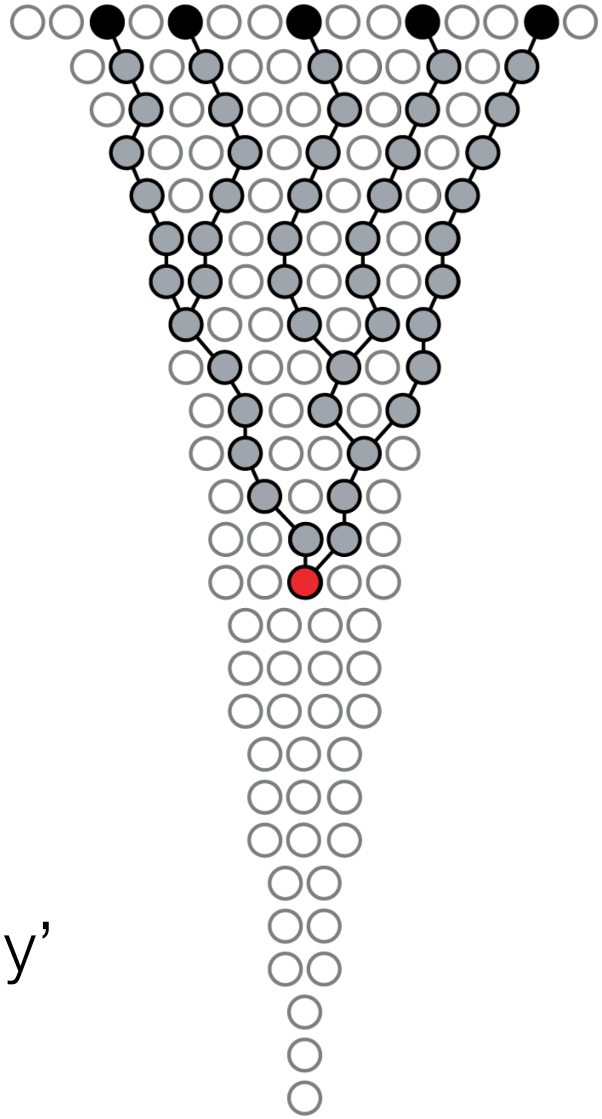




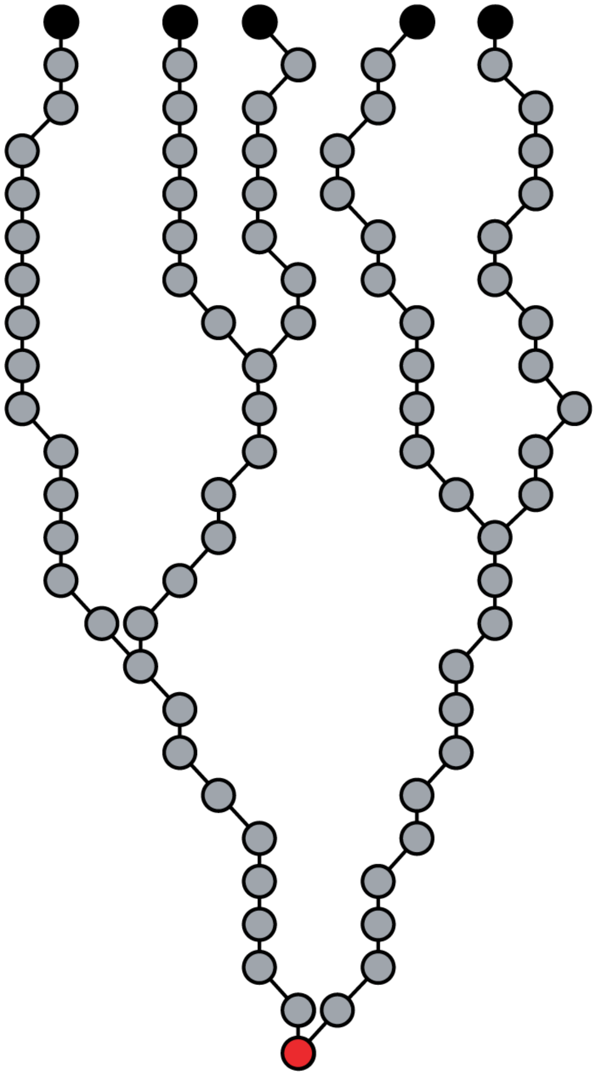
Coalescent Theory

Coalescent theory

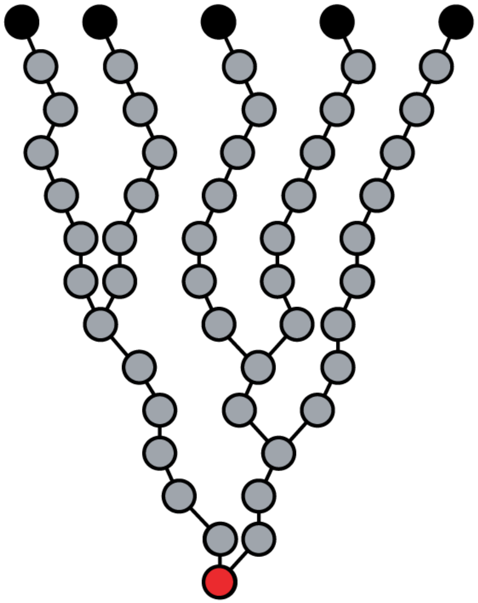
- Coalescent model used to put a prior on the tree
- Time between coalescent events depends on population size
- Different demographic models:
  - Constant population
  - Exponential growth
- Usually used for within-species data
- The tree is sometimes called ‘genealogy’



Shape of the genealogy



Constant size

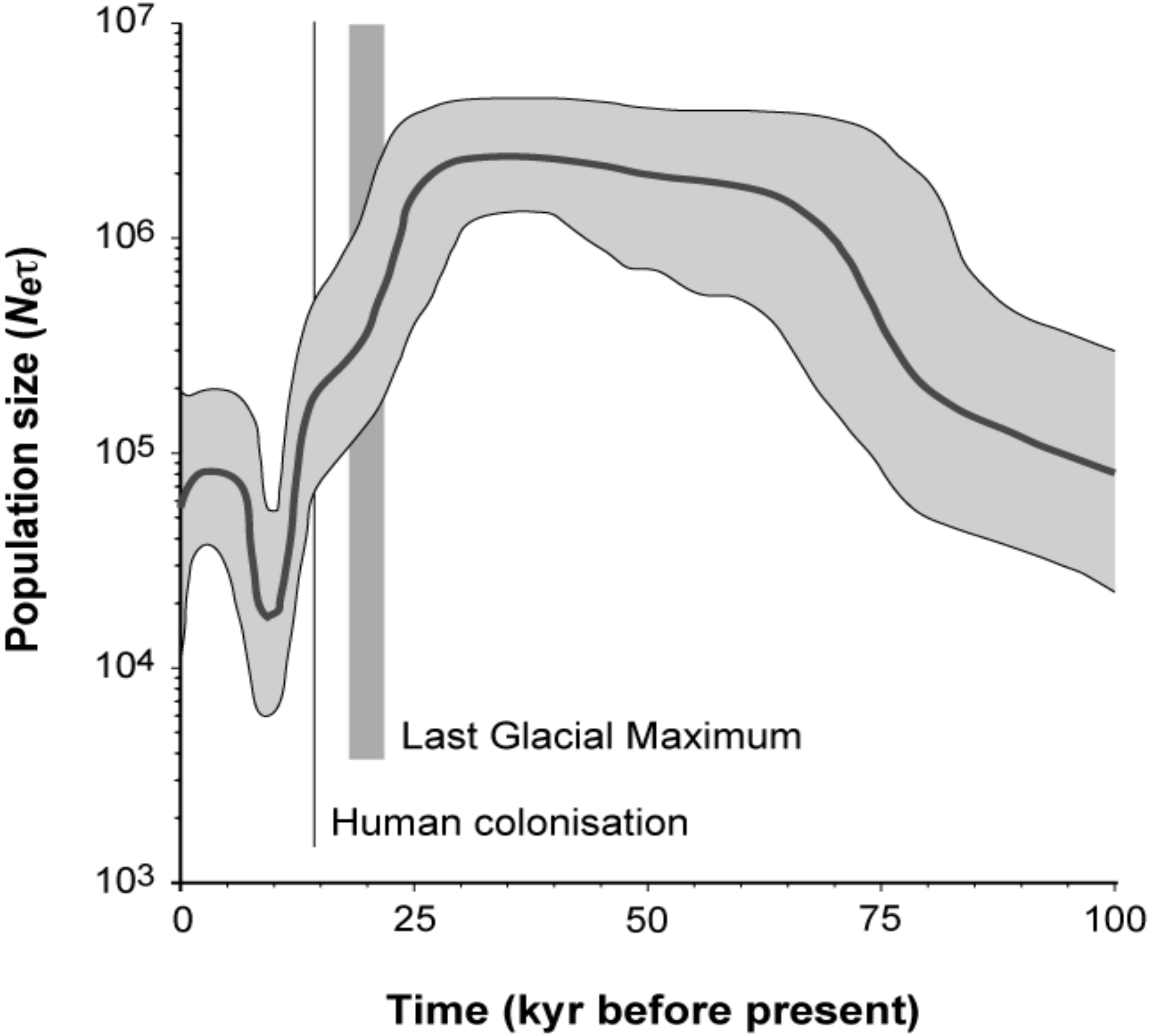


Exponential growth

Demographic history

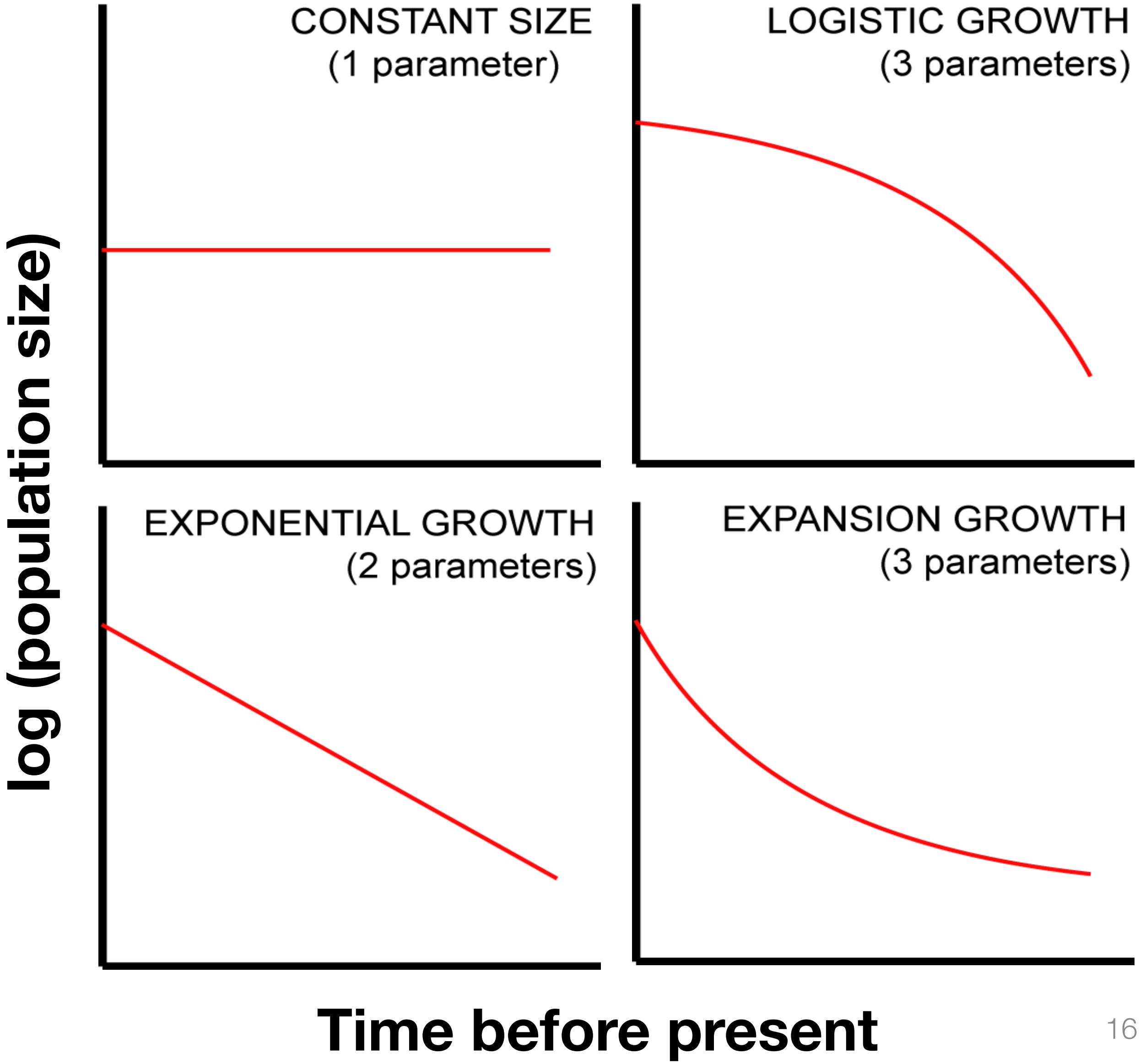
- The demographic history of a population leaves a signature in the DNA of its modern representatives
- Reconstructing this history might be of interest
  - Testing correlations with abiotic factors
  - Examine factors driving population dynamics
  - Tracing transmission and spread of viruses

Example: Bison



Demographic models in *BEAST*

1. Choose one of the models that are available



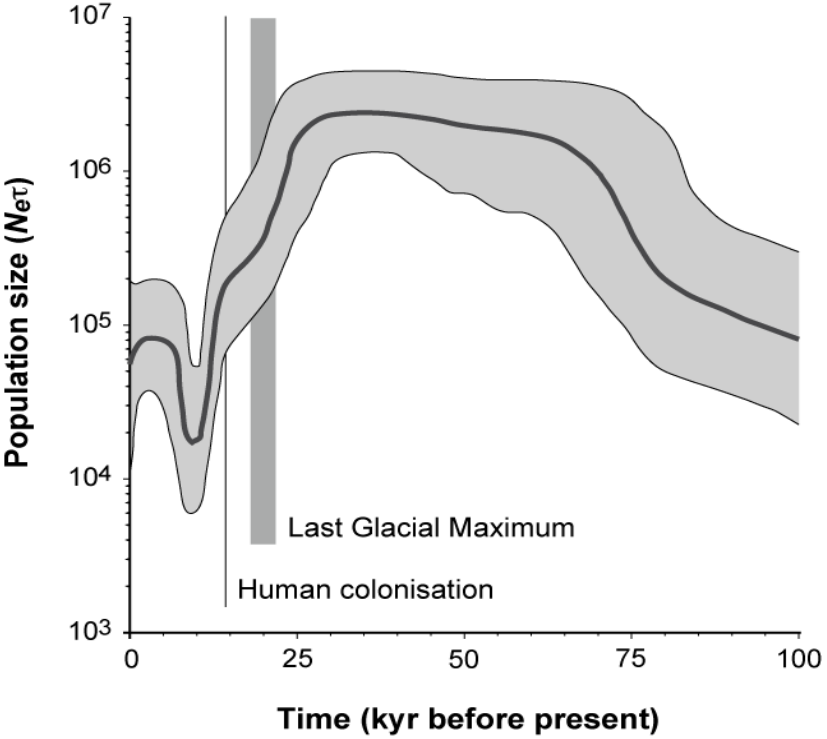
Demographic models in *BEAST*

2. Run the *BEAST* analysis under the assumed model
3. Test between candidate models by:
   - Inspecting estimates of parameters
     (e.g. growth rate in exponential-growth model)
   - Bayes factors

Skyline-plot Methods

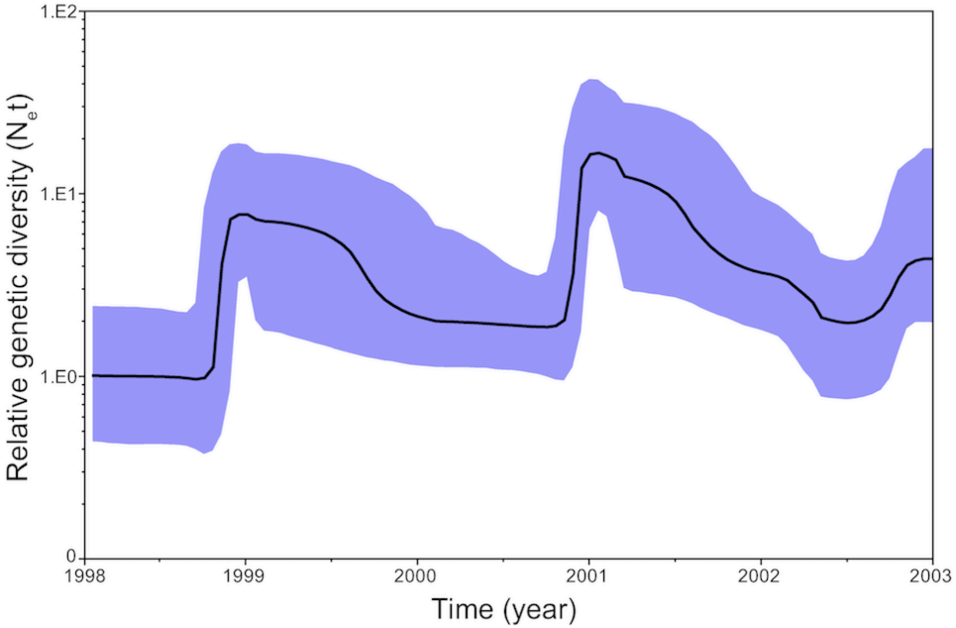
Skyine-plot methods

- In some cases it is inappropriate to limit our investigation to a small range of simple parametric models
- Skyline-plot methods enable the demographic history to be estimated from the sequence data



Data set

- Sequence data
  - One or more (informative) loci
  - Neutrally evolving
  - Non-recombining
  - High-quality sequences
- Sampling from population
  - Random sampling



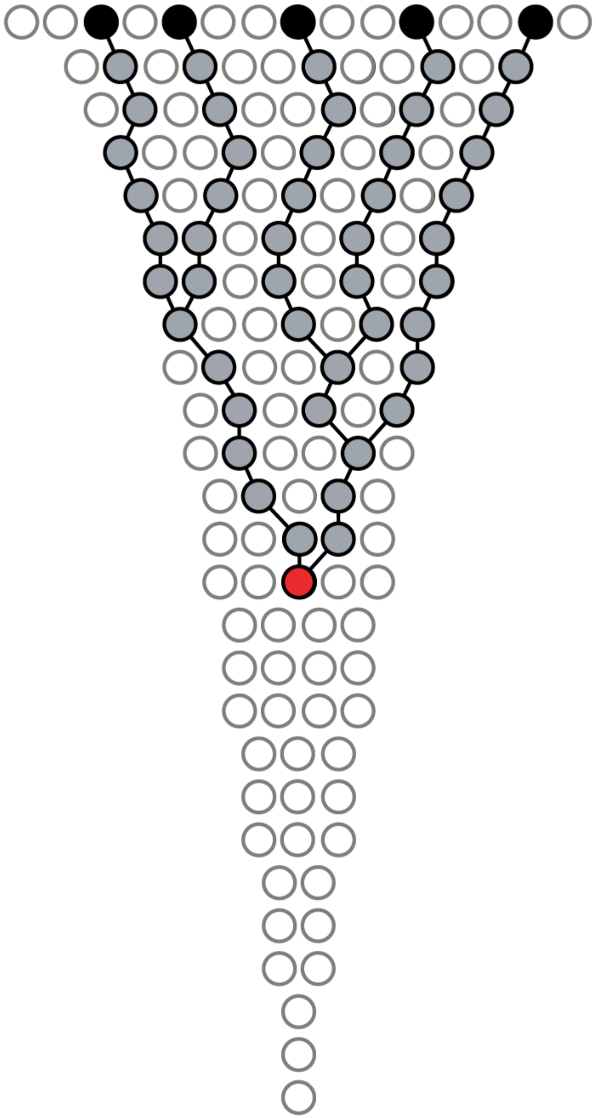
From Holmes and Grenfell 2009 *PLOS. Comp. Biol.*

Skyline-plot methods

- Given a sequence alignment, demographic reconstruction comprises two separable steps:

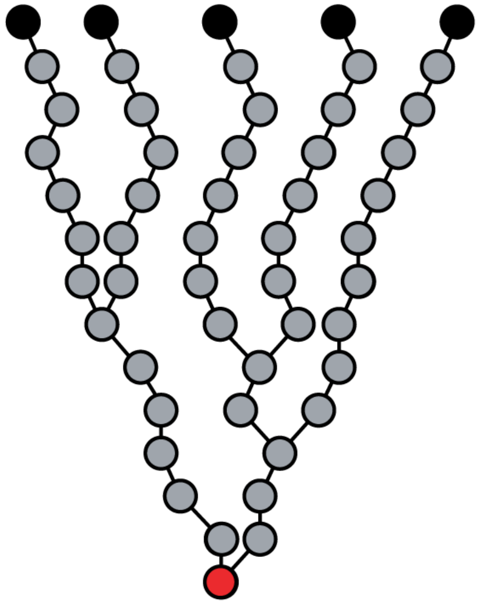
  1. Estimation of the genealogy from the alignment

  2. Estimation of population history from the genealogy



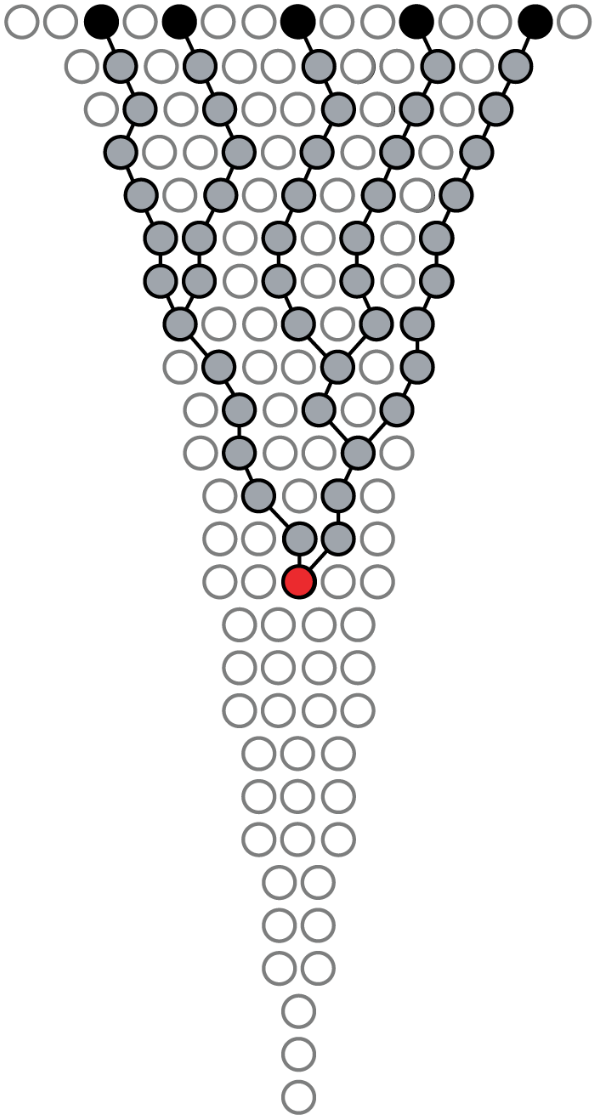
Step 1: Estimation of genealogy

- Genealogy is estimated using a phylogenetic method
- Genealogy needs to be chronogram
  - Branch lengths in time units or in substitutions per site
- Uncertainty in the estimate is referred to as *phylogenetic error*



Step 2: Estimation of demo. history

- Based on coalescent theory
- Coalescent theory quantifies the relationship between the genealogy and demographic history of the sequences
- Uncertainty in the estimate is referred to as *coalescent error*



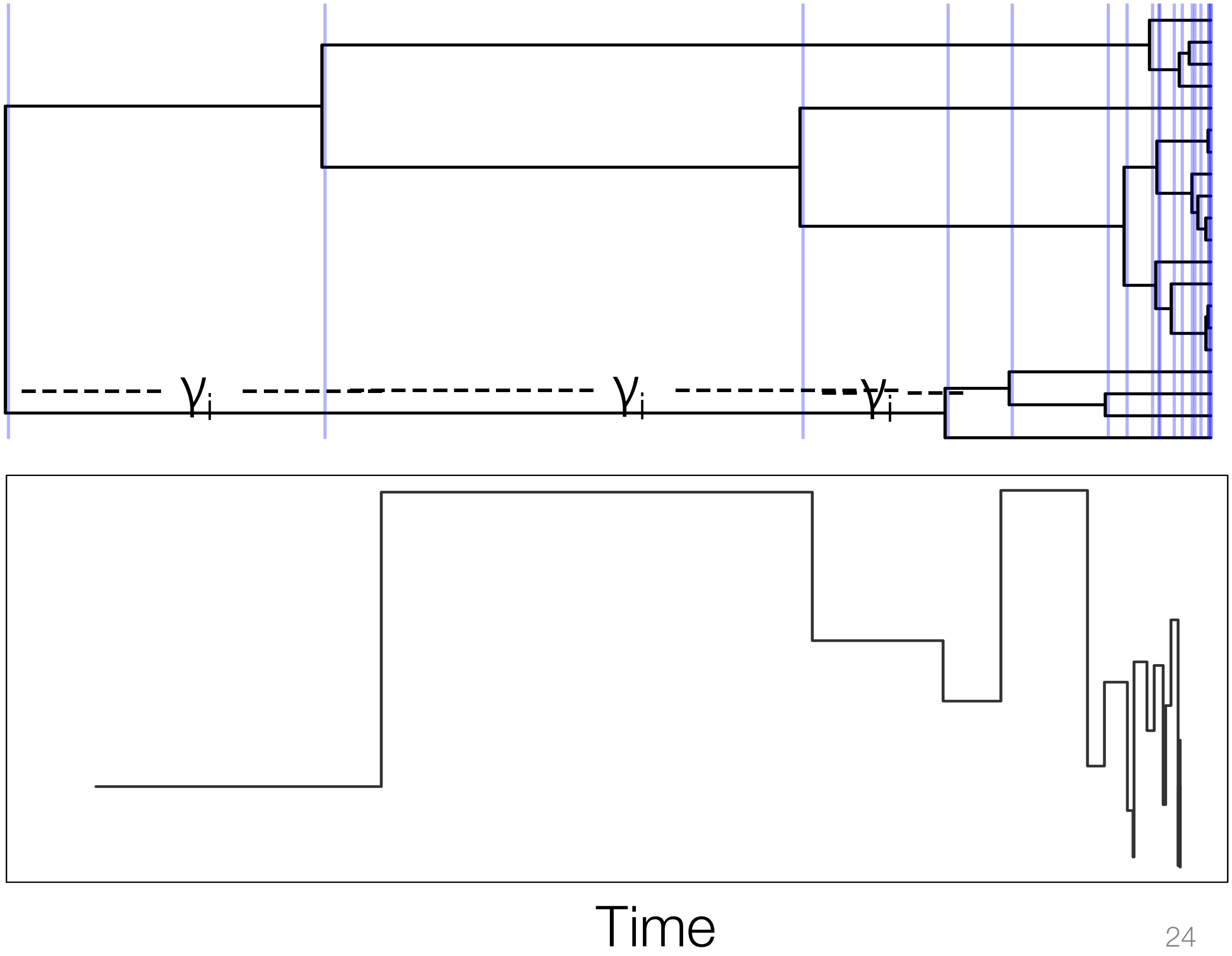
Classic skyline

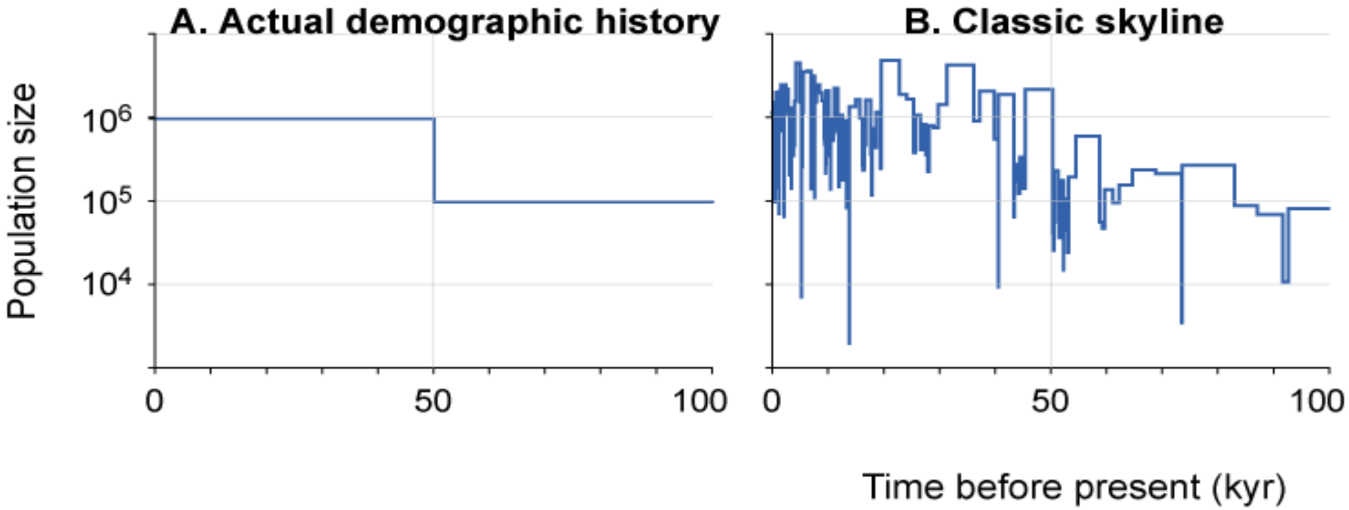
$N_i = \gamma_i \, i \, (i - 1) / 2$

i: number of lineages

γ_i: coalescent intervals

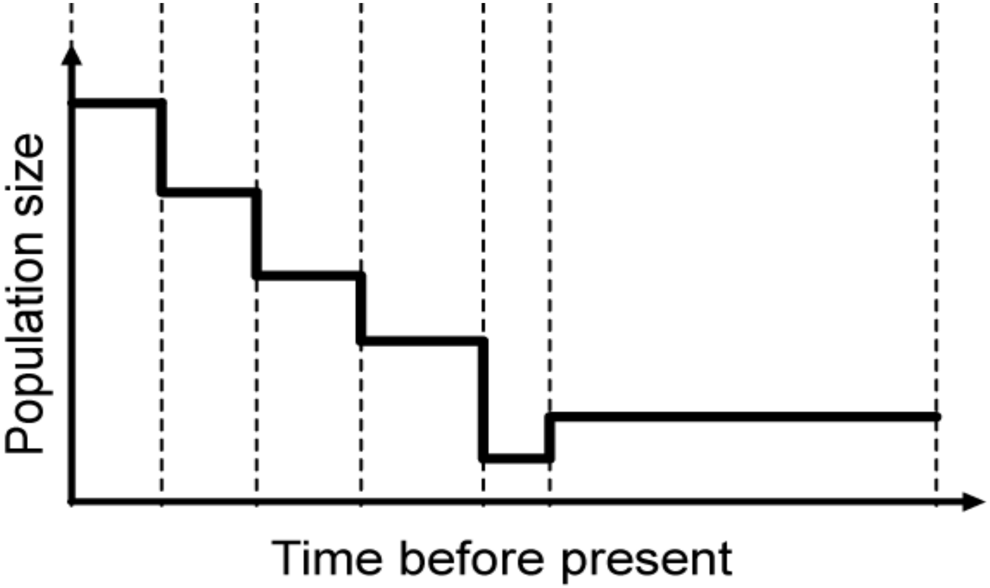
Population effective size (N_e)

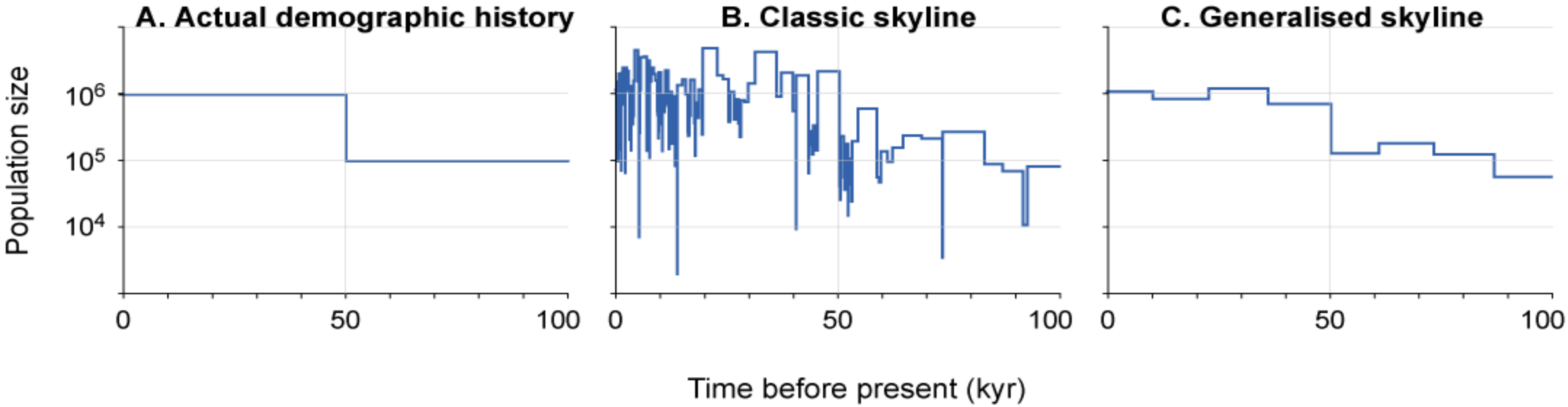



A. Actual demographic history
B. Classic skyline
Population size
10^6
10^5
10^4
0
50
100
0
50
100
Time before present (kyr)

Generalised skyline

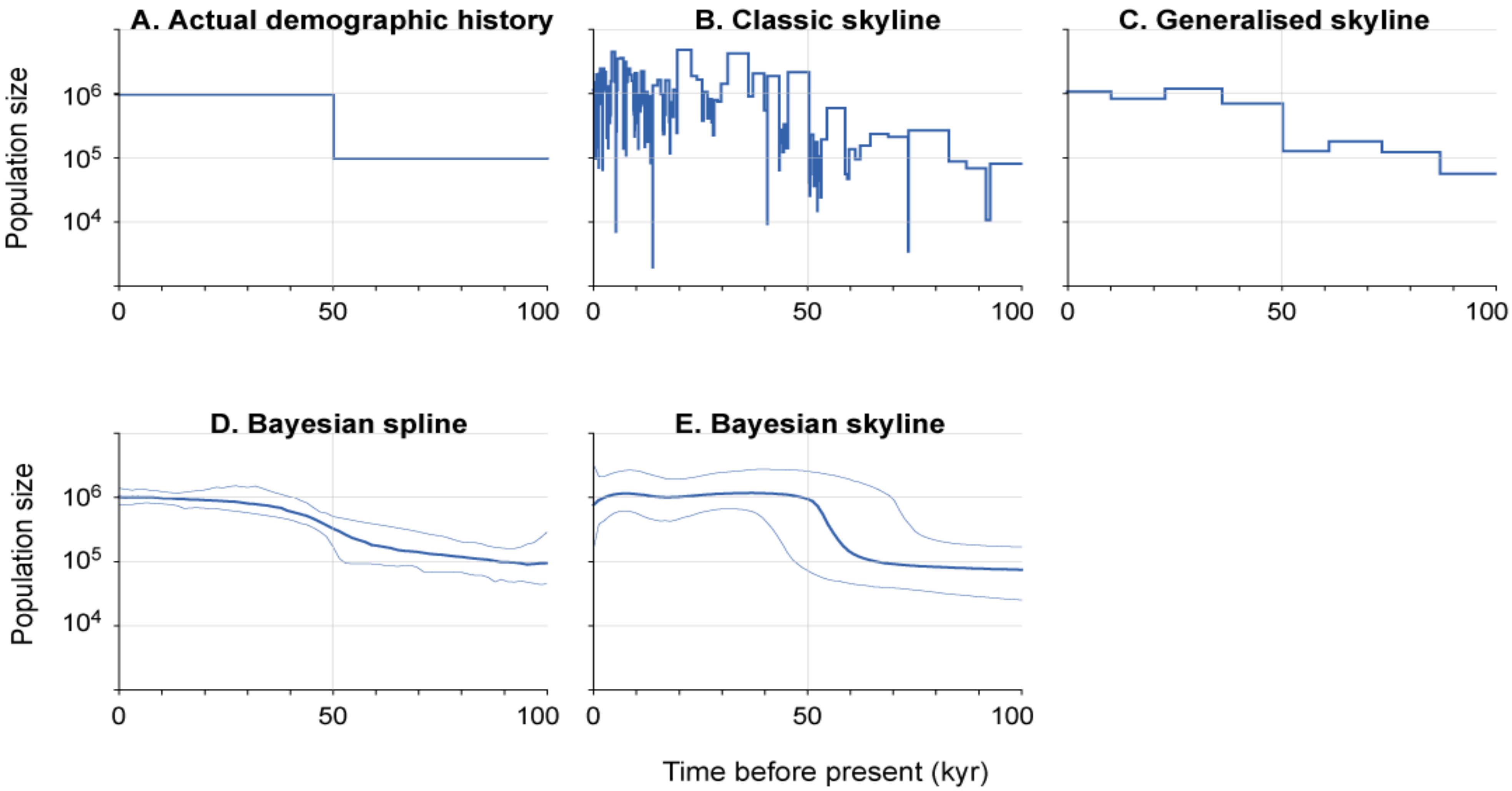
- Smaller coalescent intervals are grouped together
- Optimal number of groups determined statistically using the Akaike information criterion




A. Actual demographic history
B. Classic skyline
C. Generalised skyline
Population size
10^6
10^5
10^4
0
50
100
0
50
100
0
50
100
Time before present (kyr)

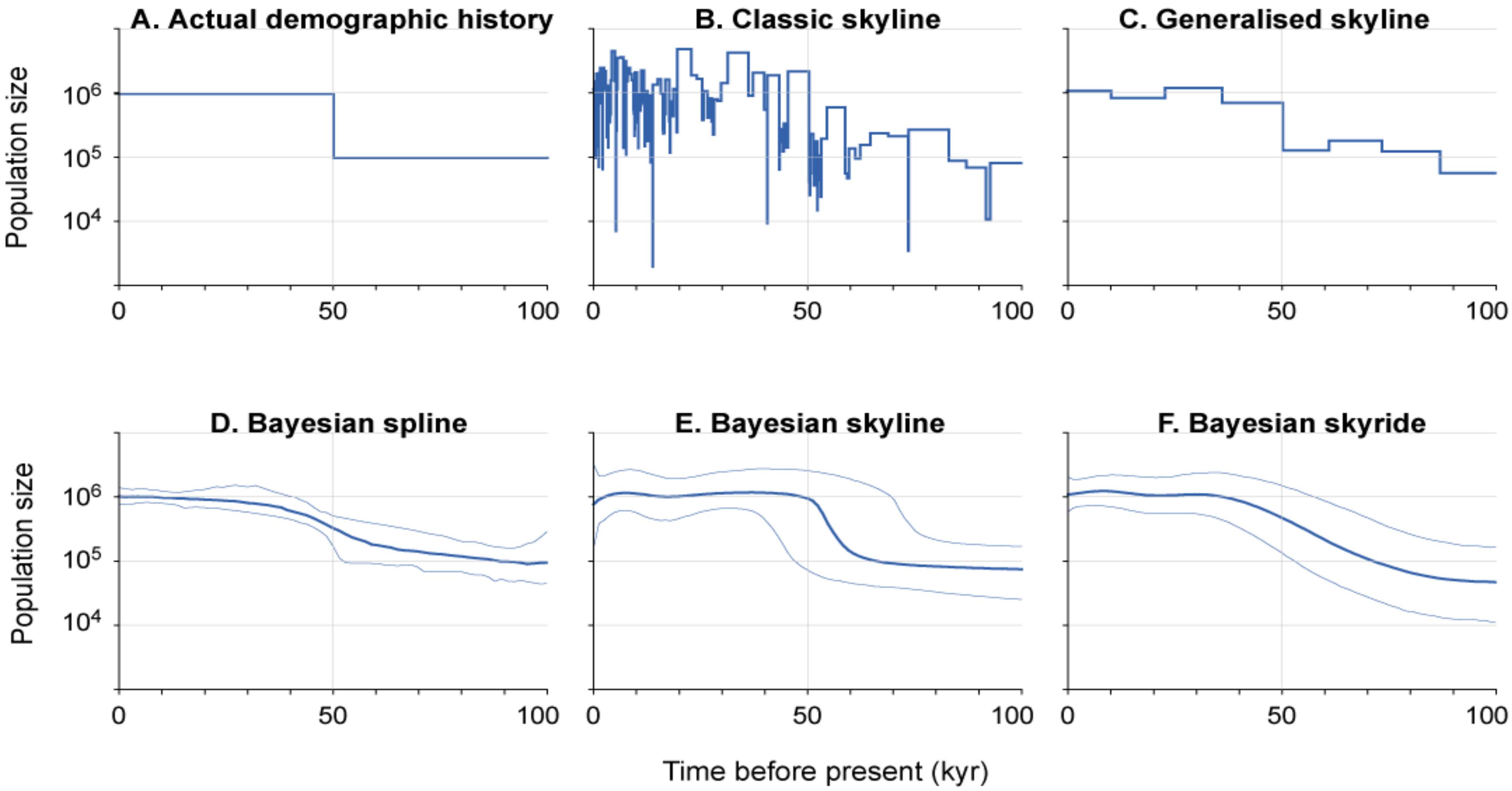
Bayesian skyline

- There is often substantial uncertainty in the estimate of the genealogy (phylogenetic error)
- Bayesian skyline plot allows co-estimation of genealogy, node times, and demographic history
- Based on the generalised skyline plot, but the number of groups needs to be chosen *a priori*
- Successive population sizes have some degree of correlation
- Final demographic plot is averaged over phylogenetic uncertainty


A. Actual demographic history
B. Classic skyline
C. Generalised skyline
D. Bayesian spline
E. Bayesian skyline
Population size
10^6
10^5
10^4
0
50
100
Time before present (kyr)

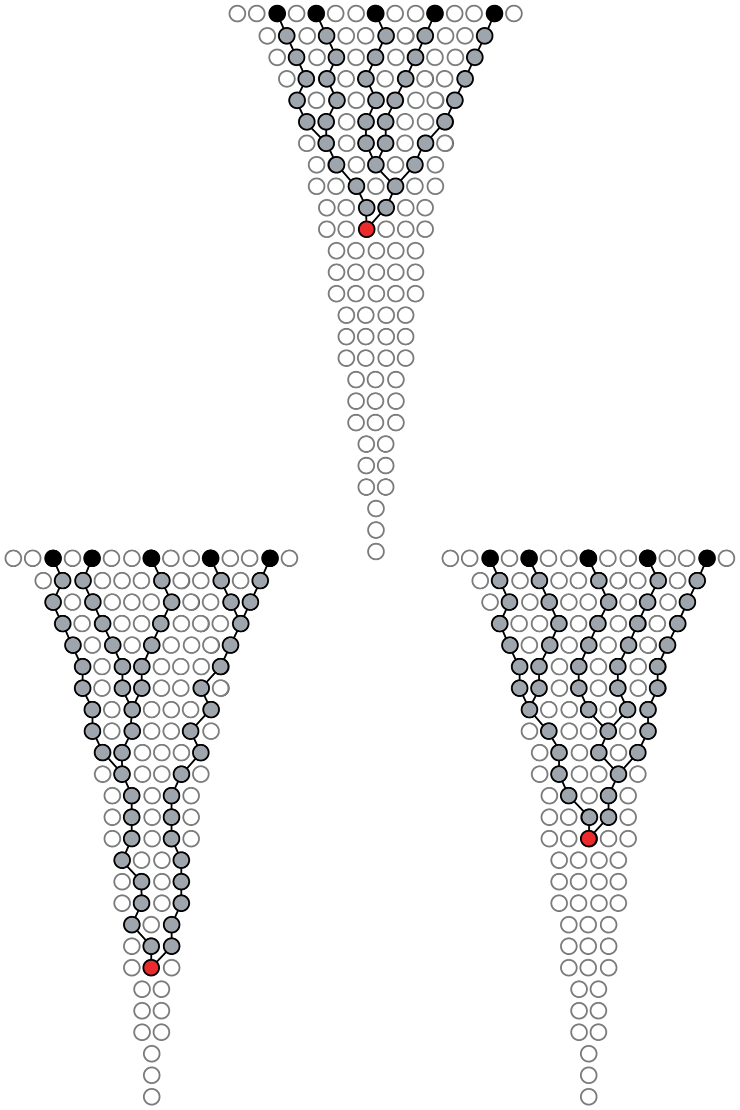
Bayesian skyride

- Extension of the Bayesian skyline plot
- Assumes that population size changes gradually
    - ‘Time-aware’ prior on population size
    - Population-size changes between intervals are smoothed


A. Actual demographic history
B. Classic skyline
C. Generalised skyline
D. Bayesian spline
E. Bayesian skyline
F. Bayesian skyride
Population size
10^6
10^5
10^4
0
50
100
Time before present (kyr)

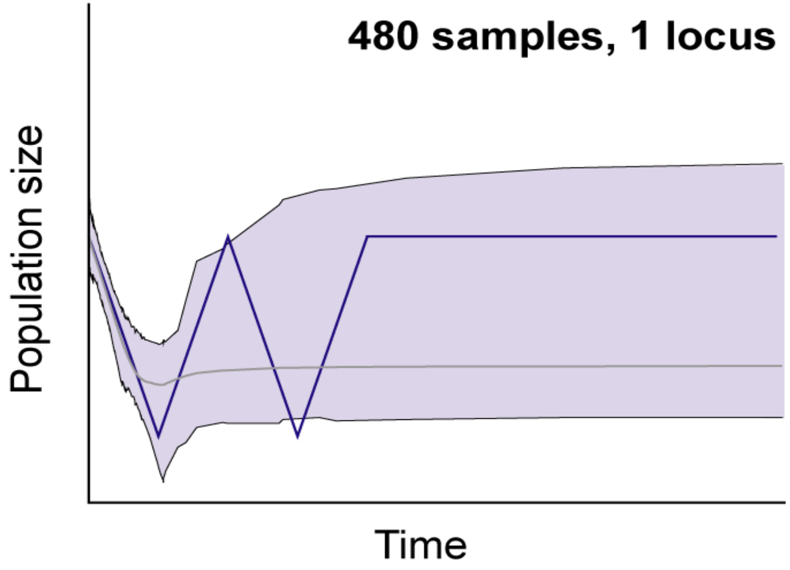
Extended Bayesian skyline

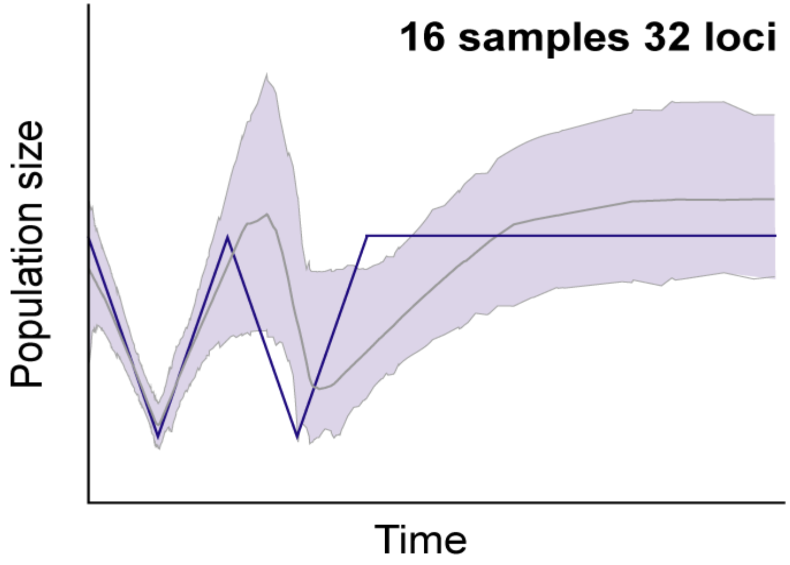
- Substantial coalescent error associated with reconstructing demographic history from a single genealogy/locus
- Any single genealogy is only one realisation of a stochastic process



Extended Bayesian skyline

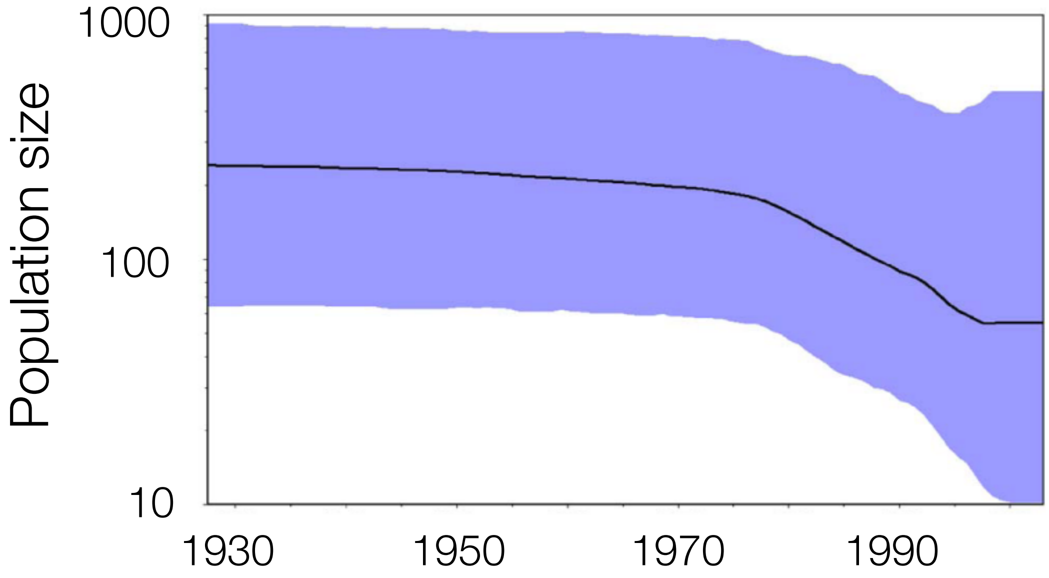
- Extended Bayesian skyline allows simultaneous analysis of multiple loci
  - Distinct substitution models
  - Distinct substitution rates
  - Distinct genealogies
- Advantages of multiple loci
  - Reduce coalescent error
  - Increase power to resolve bottleneck events





Evaluating support

- Is an inferred demographic pattern is meaningful?
  - Bayes factors
  - Visual inspection of skyline plot
  - Number of change points (eBSP)



Toscana virus
Zehender *et al*. (2009)
Infect Genet Evol, 9: 562-566

Bayesian Model Selection

Bayesian model selection

- Bayesian model selection is usually based on the marginal probability of the data, conditioned on the model:

$$\mathbf{Pr(D|M)}$$

- This is a weighted average of the likelihood
- Weights are given by the prior distribution

Marginal likelihood of the model

Bayesian model selection

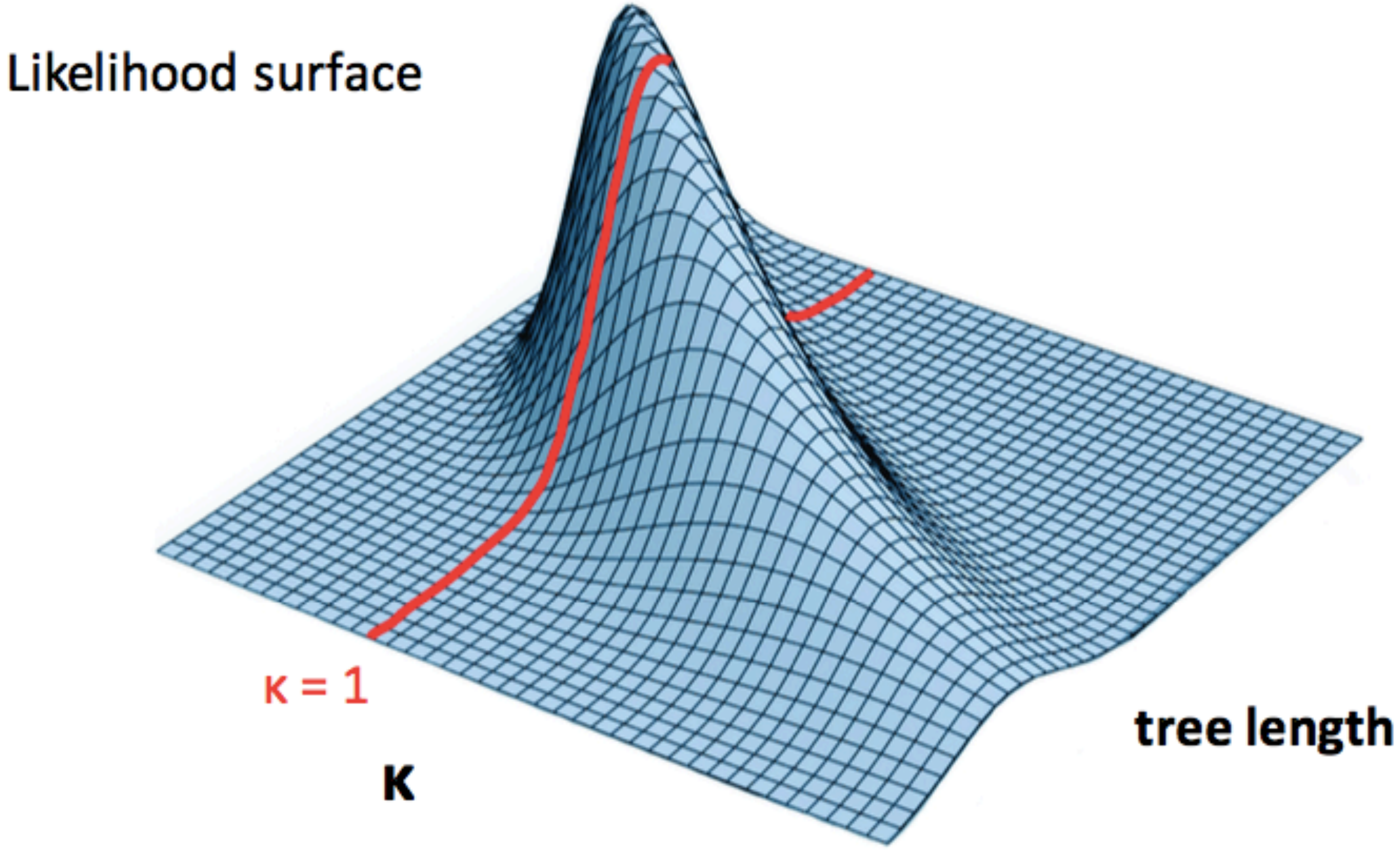
- Compare marginal likelihoods of competing models
- Ratio of marginal likelihoods is known as the Bayes factor:

$$\frac{\mathbf{Pr(D|M1)}}{\mathbf{Pr(D|M2)}}$$

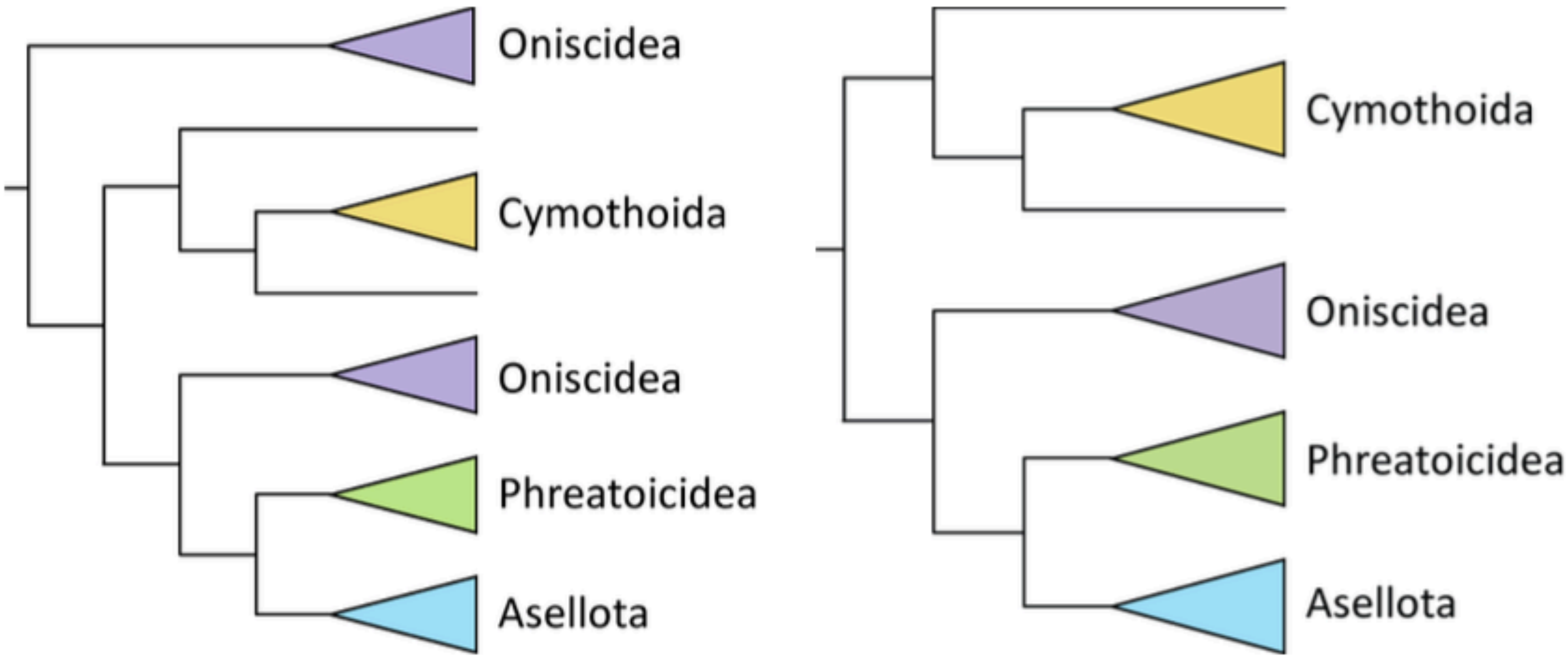
$$\mathbf{Log(BF) = Pr(D|M1) - Pr(D|M2)}$$

- Models do not need to be nested
- No need to correct for the number of parameters

Bayesian model selection



Bayesian model selection



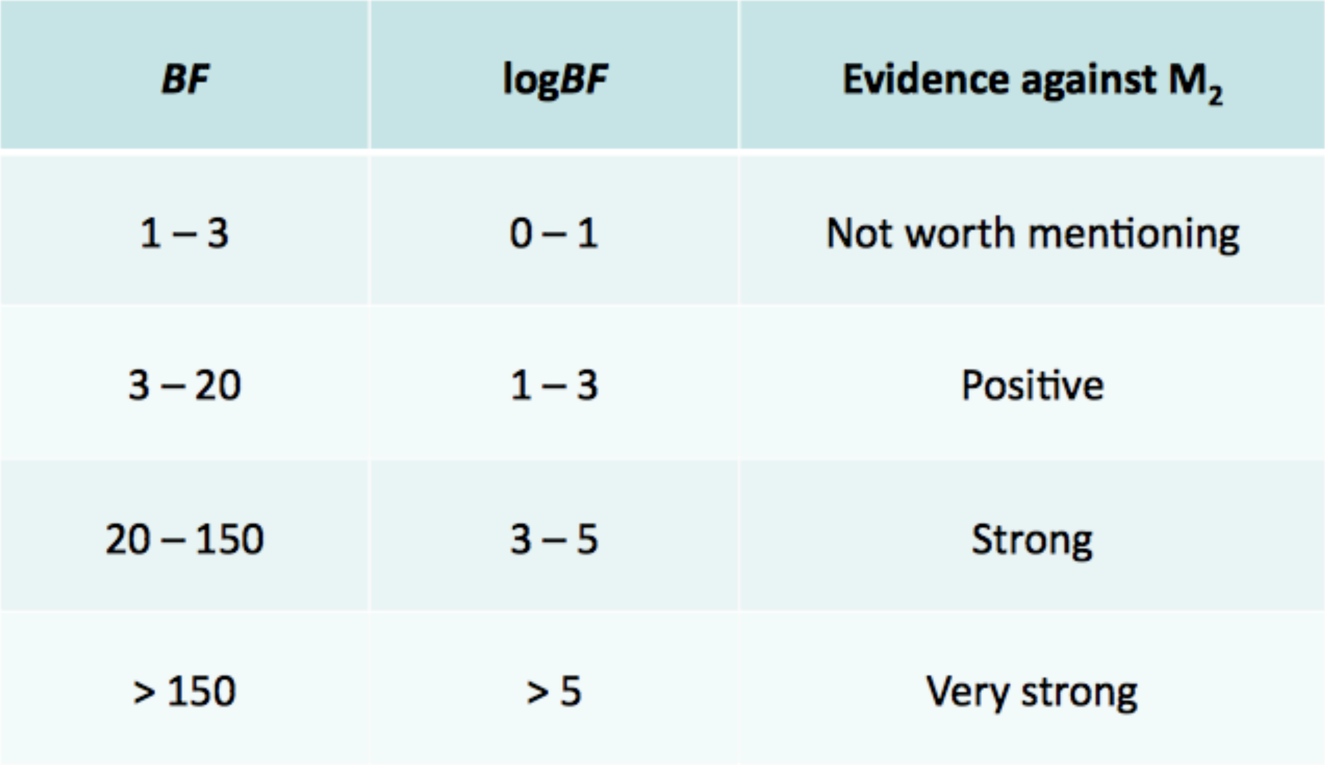
marginal log*L* = -13085

marginal log*L* = -13089

log*BF* = 4

Bayesian model selection

- Interpreting Bayes Factors



BF	log*BF*	Evidence against M_2
1 – 3	0 – 1	Not worth mentioning
3 – 20	1 – 3	Positive
20 – 150	3 – 5	Strong
> 150	> 5	Very strong

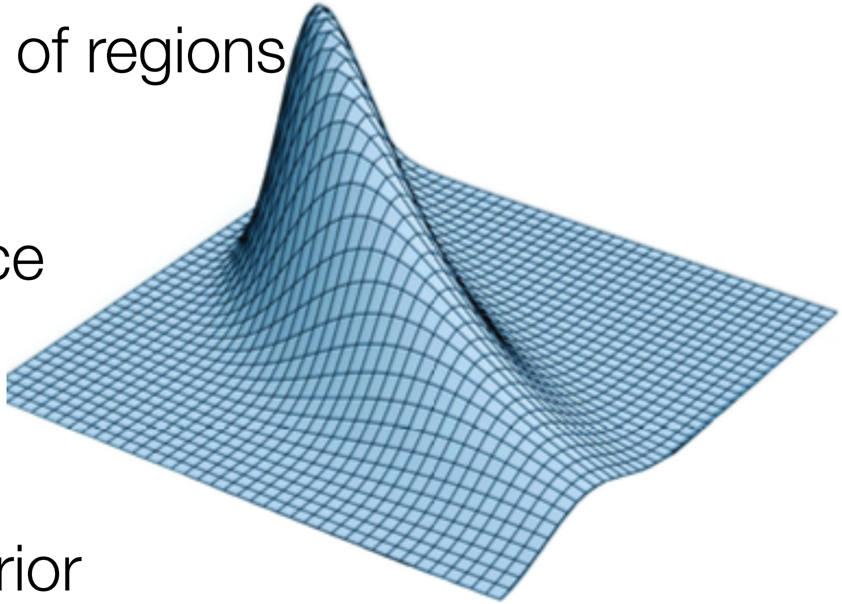
Kass and Raftery (1995) *J Am Stat Assoc*

Estimating the marginal likelihood

- **Harmonic mean estimator**
  - Can be calculated from likelihood values sampled from the MCMC
  - Easy to calculate from standard MCMC output
  - Sensitive to extreme values
  - Does not sufficiently penalise excessive parameters

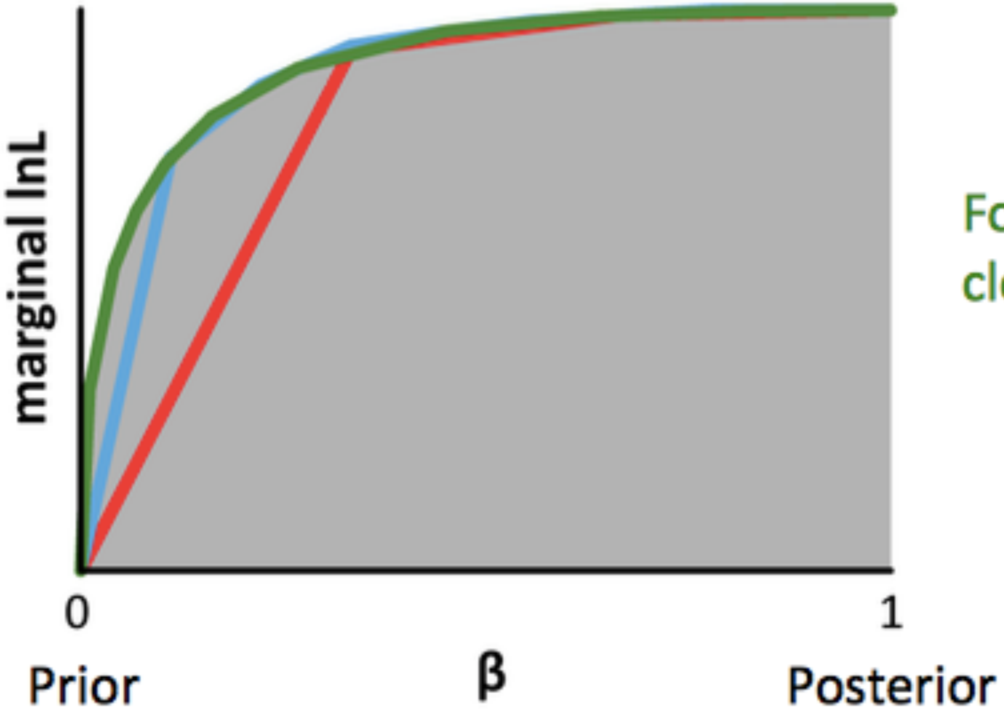
Estimating the marginal likelihood

- The MCMC tends to sample more from regions with high likelihood
- But the marginal likelihood is integrated over the entire likelihood surface
- We need an accurate representations of regions with low likelihood
- Use methods to distort the acceptance ratio of the MCMC to explore regions with low likelihood
- Use a quantity, β, to weight the posterior



Estimating the marginal likelihood

- Generate MCMC samples from a series of densities that lie between the prior and the posterior



Focus on values of β close to zero (the prior)

Estimating the marginal likelihood

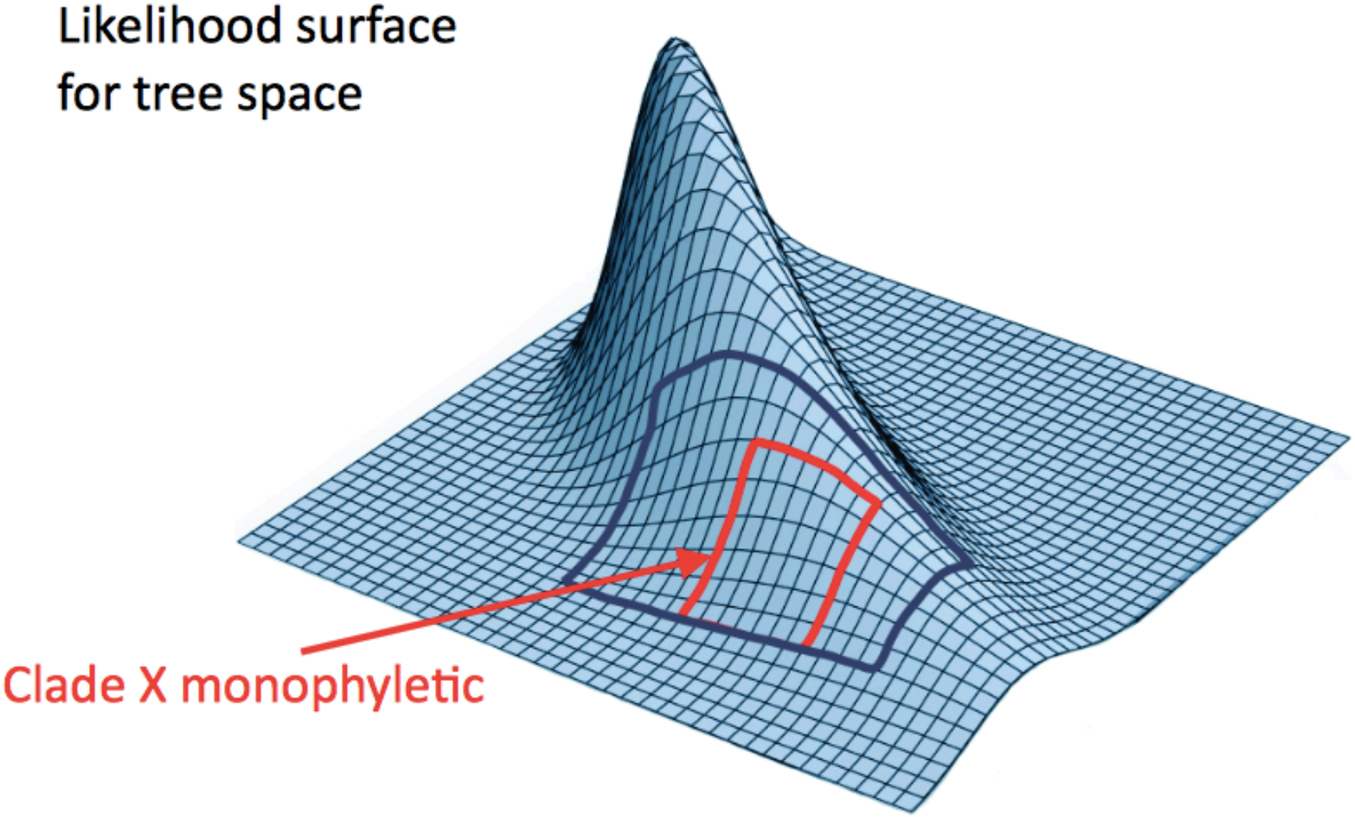
- Three methods
  - Path sampling (thermodynamic integration)
  - Stepping-stone sampling
  - Generalised stepping-stone

- Very slow (each β value entails a whole MCMC analysis)
- More reliable than harmonic-mean estimators

Problems with Bayes factors

- Bayes factors are unreliable when there are improper priors
- Bayesian model selection can be sensitive to the choice of prior distributions
- **Lindely’s paradox**
  - Occurs when frequentist and Bayesian approaches support different models
  - One model is typically more specific than the other
  - In phylogenetics, it is most commonly observed in topology testing

Lindely’s paradox



Go to **Practical 3: Molecular dating using BEAST**