Bayesian Phylogenetics and Molecular Clocks

Simon Ho

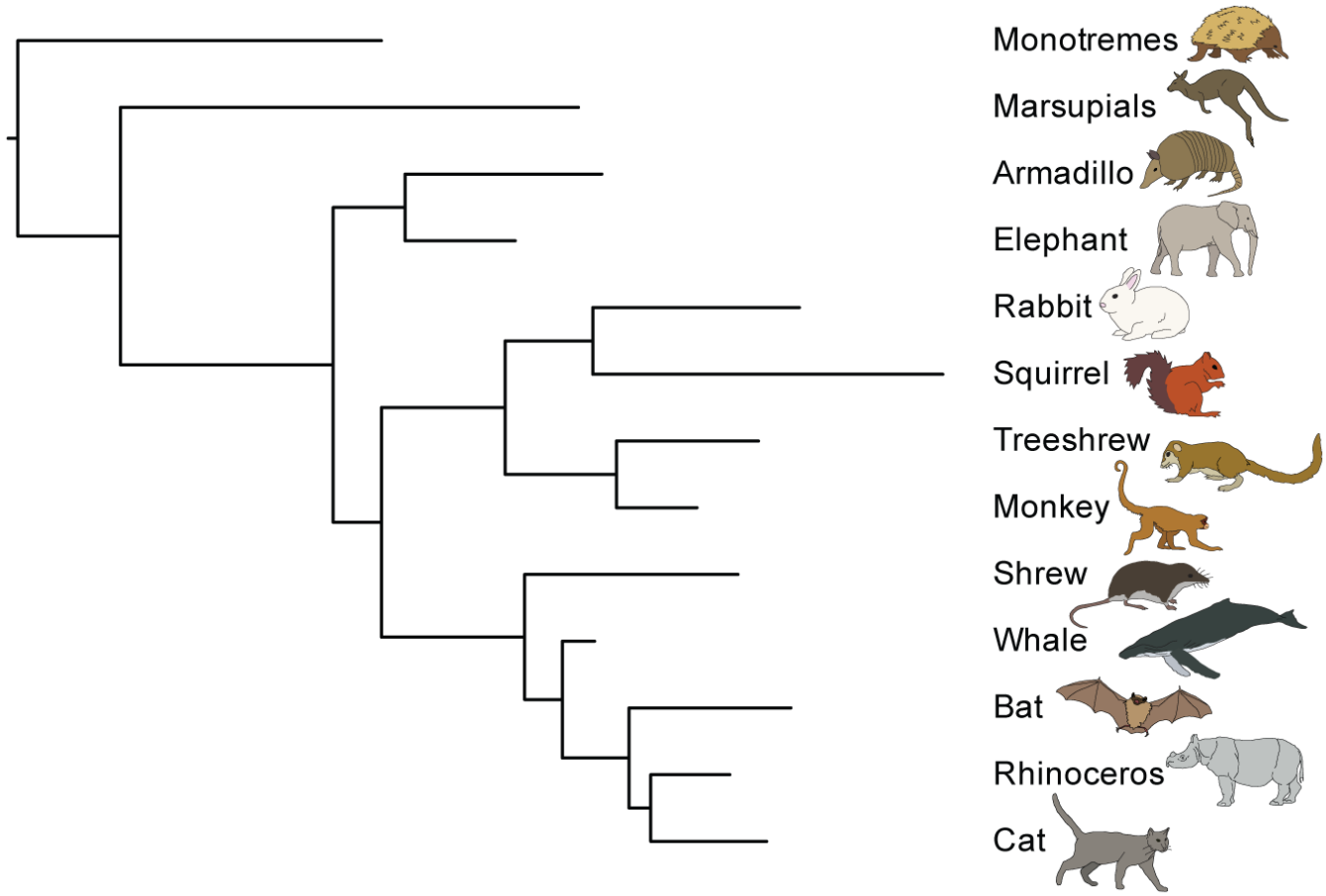
Outline of Part A

1. Molecular phylogenetics
2. Substitution models
3. Bayesian phylogenetics
4. Priors
5. Likelihood
6. Model selection
7. Markov chain Monte Carlo sampling
8. Software

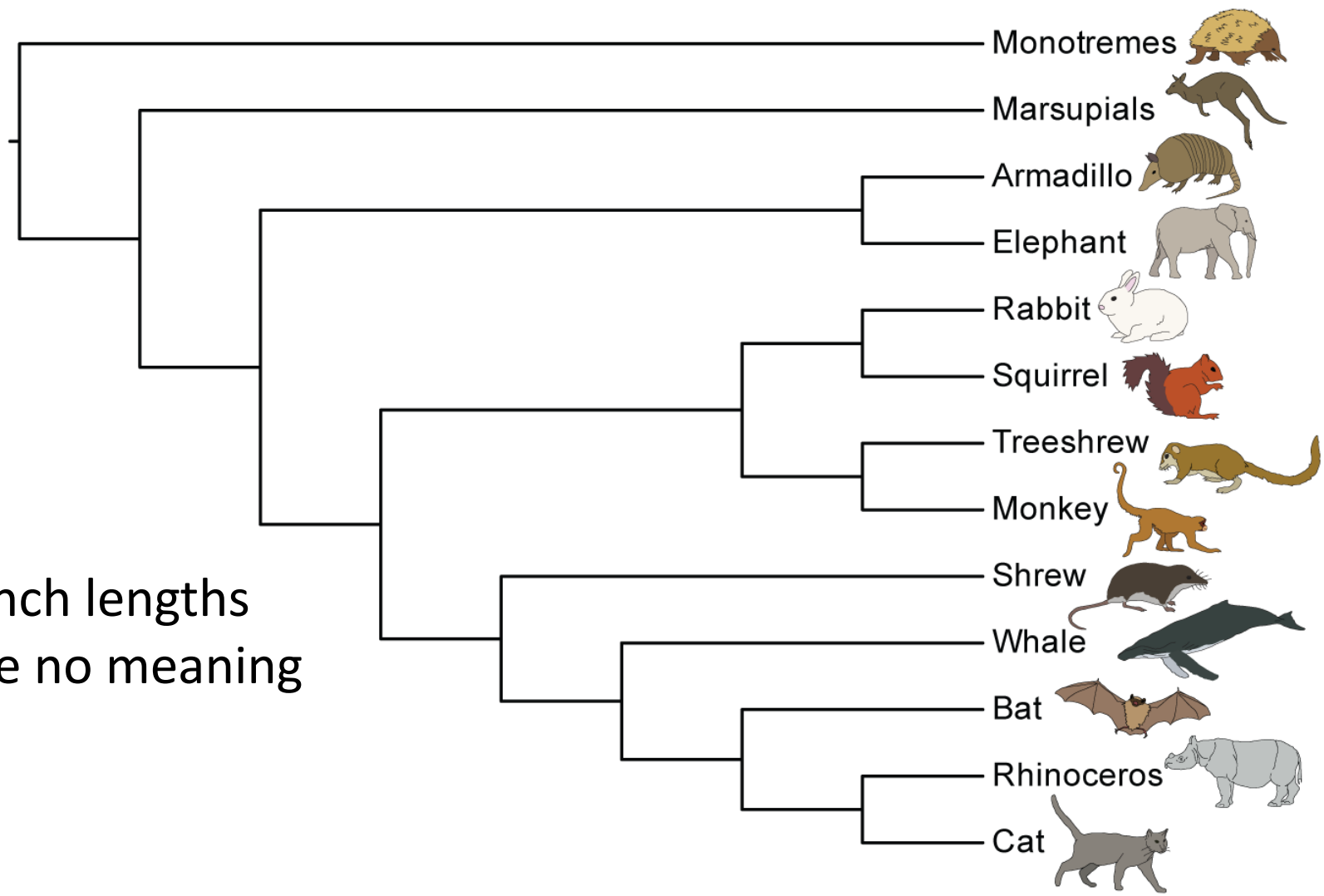
1. Molecular Phylogenetics

Phylogenetic trees

- Topology (relationships)
- Branch lengths (amount of evolutionary change or time)

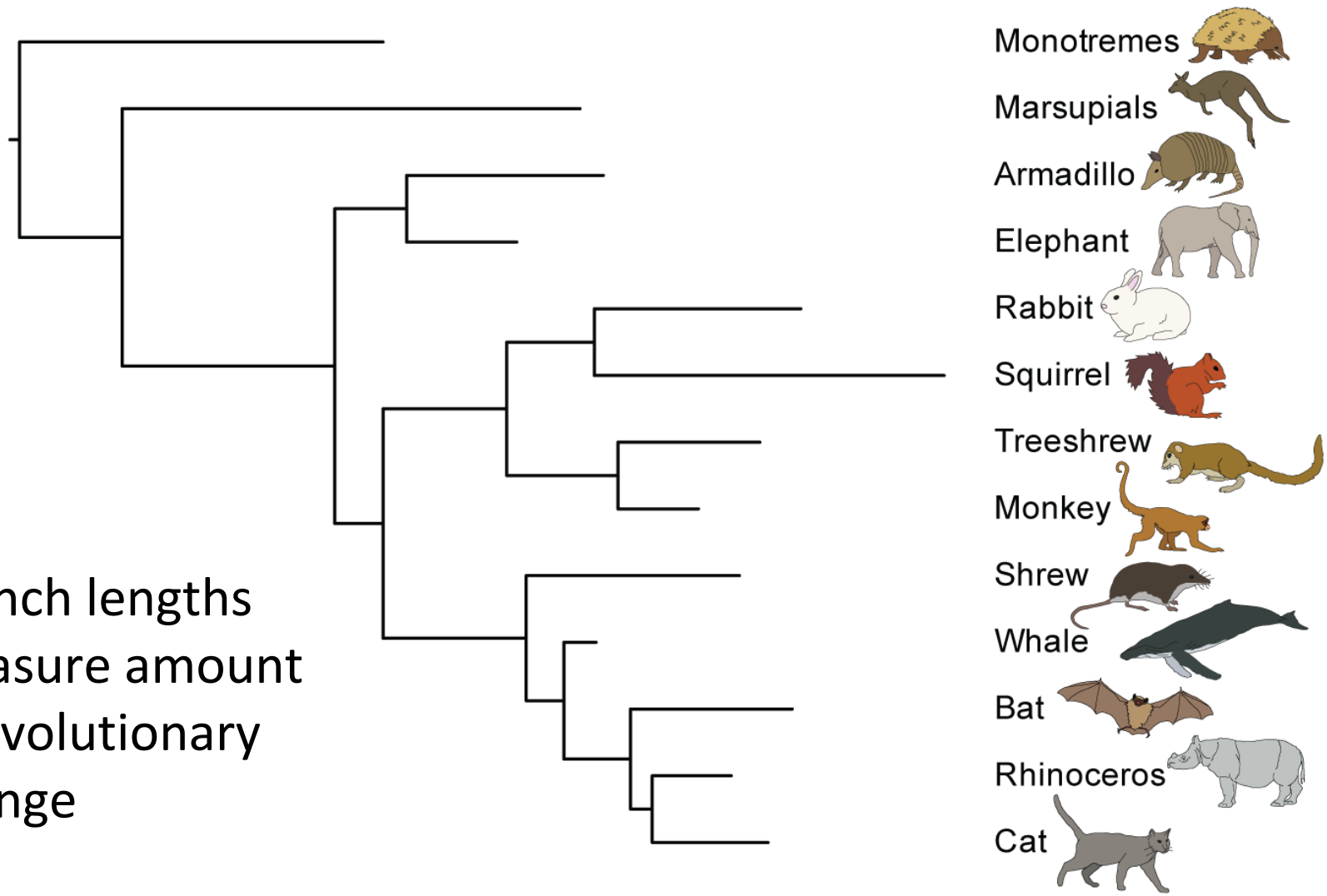


Phylogenetic trees: Cladogram



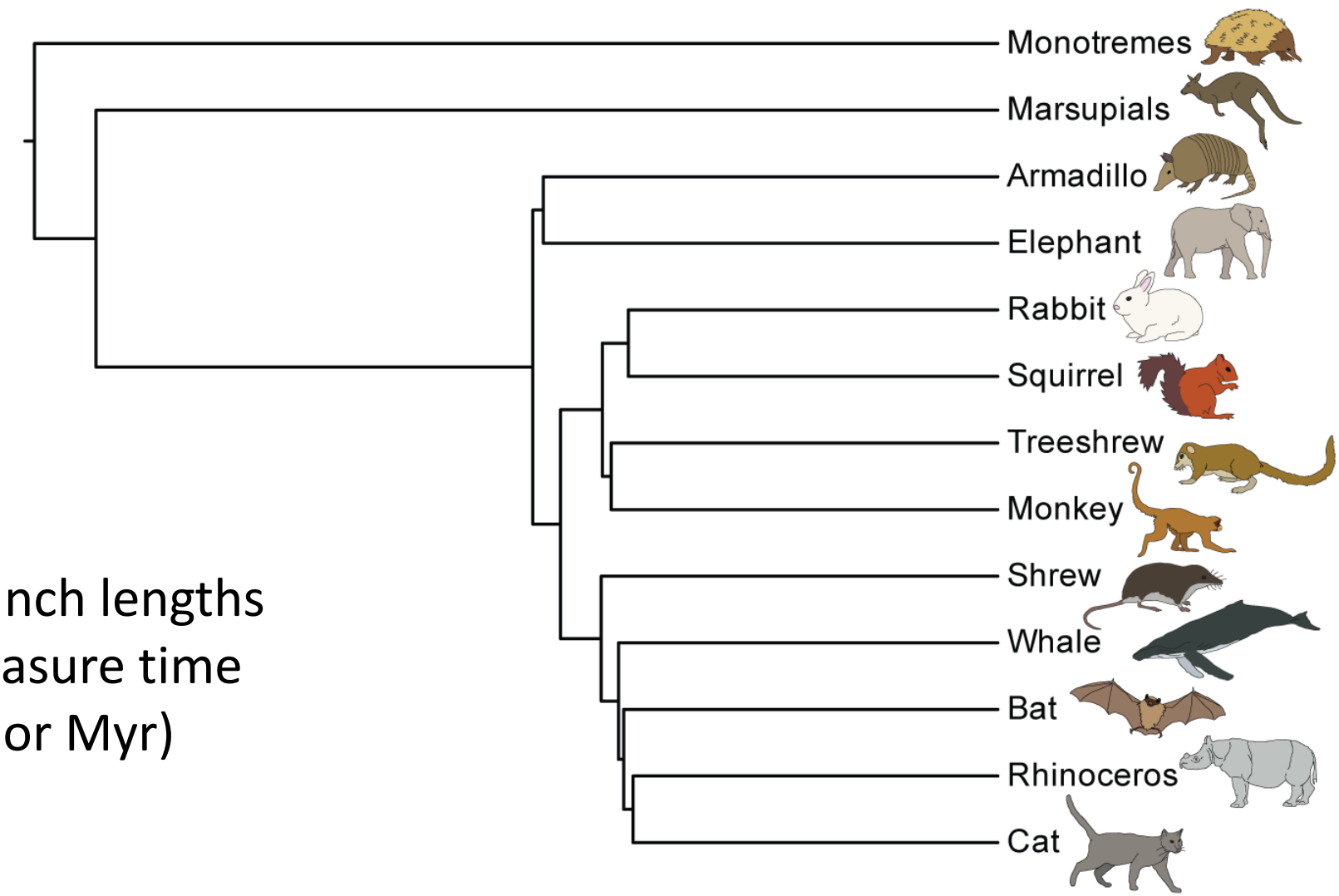
Branch lengths have no meaning

Phylogenetic trees: Phylogram



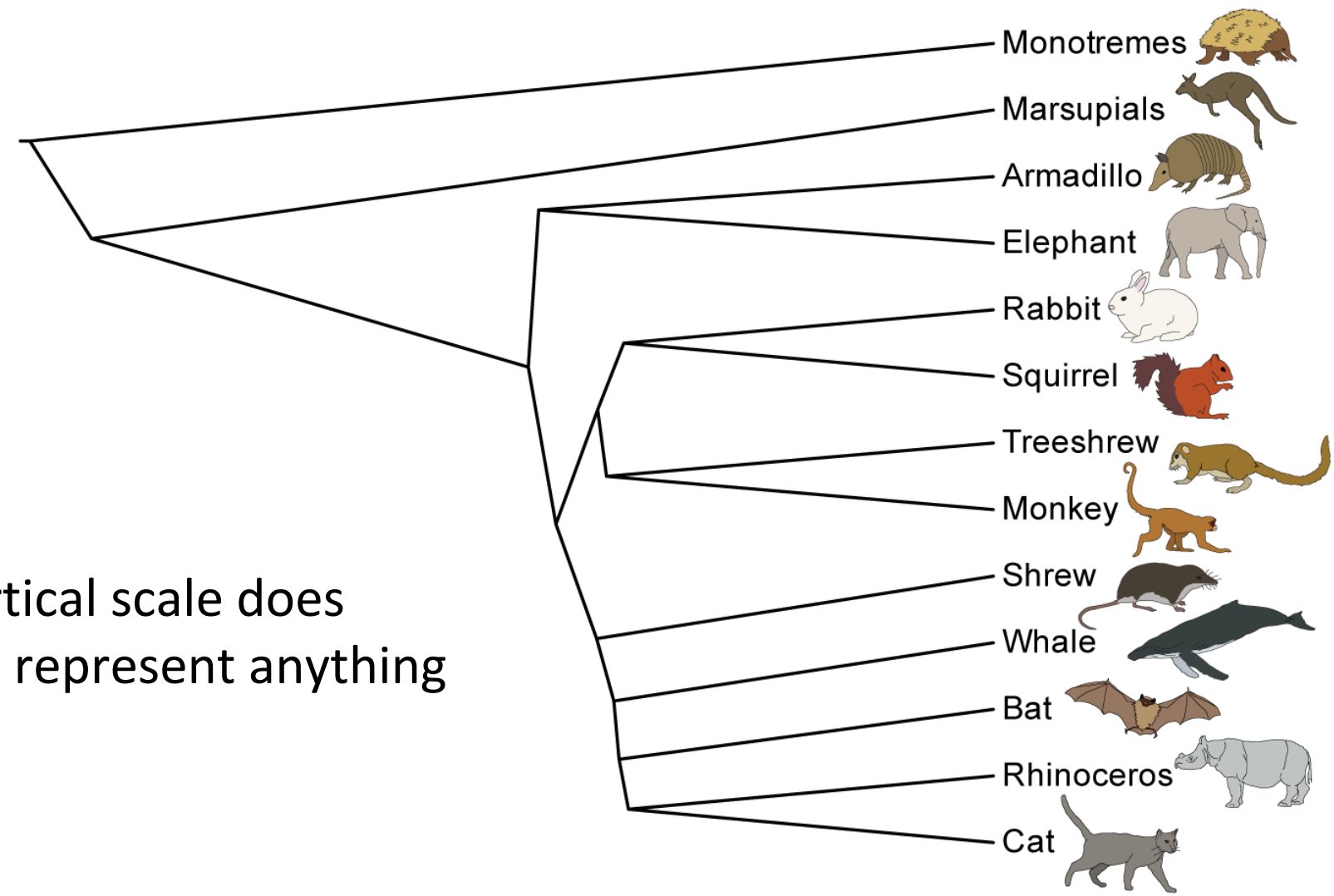
Branch lengths measure amount of evolutionary change

Phylogenetic trees: Chronogram



Branch lengths measure time (yr or Myr)

Phylogenetic trees



Vertical scale does not represent anything

Molecular data

- Microsatellites (repeat numbers)
- Reduced-representation sequences
- Single-nucleotide polymorphisms (SNPs)
- Sequence data
  - Nucleotides
  - Amino acids

Single-nucleotide polymorphisms

- Single sites sampled from throughout the genome
- More common in population studies

- Issues to consider:
  - **Recombination**
    SNPs are usually unlinked so they are likely to have different (gene) trees
  - **Ascertainment bias**
    SNPs are selected for variability and this can mislead estimates of population sizes, mutation rates, and other parameters

Sequence data

AACATTAGT

AACATAGGT

ACCAAAGT

CACAAAT

ATAAACAA

- **Protein-coding genes**
- **Other genes**
- **Non-coding sequences**
  - Intergenic sites
  - Introns

- Often have indels (insertions/deletions)
- Need to align sequences

DNA sequence alignment

AACATTAGT

AACATAGGT

ACCAAAGT

CACAAAT

ATAAACAA

AACATTAGT

AACATAGGT

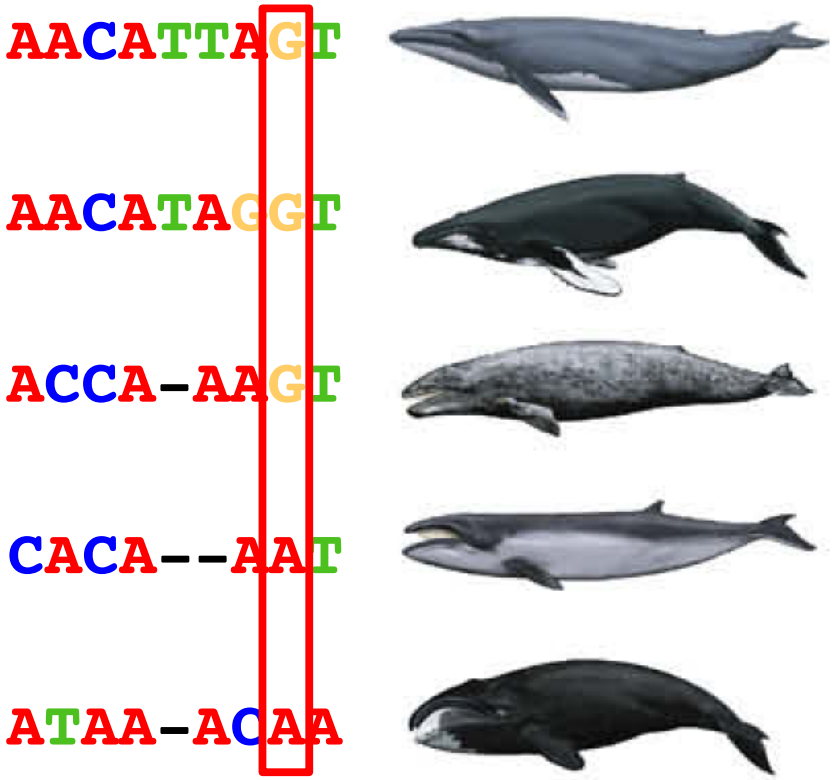
ACCA-AAGT

CACA--AAT

ATAA-ACAA

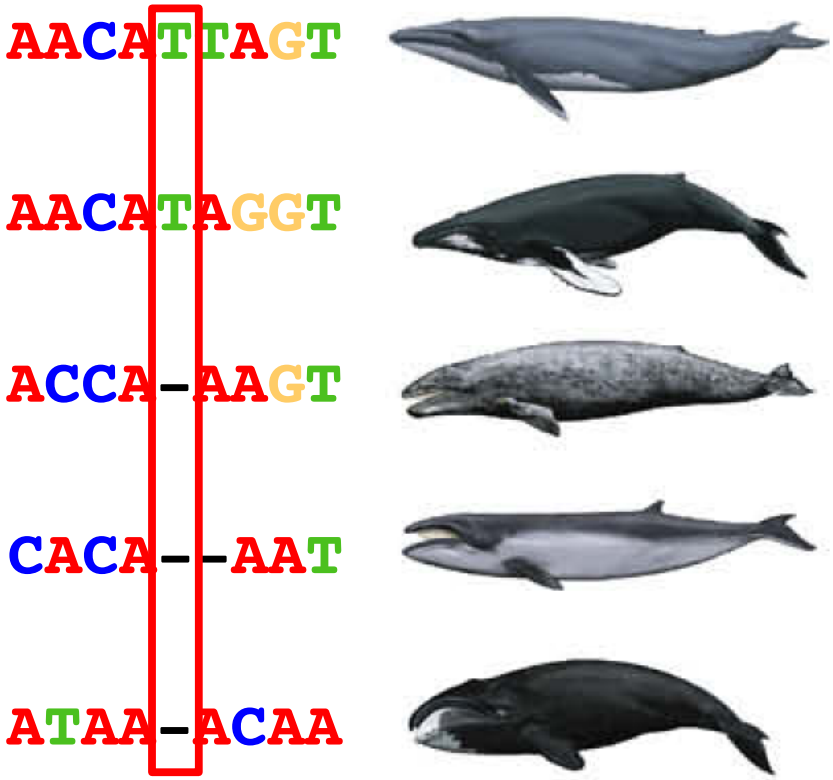
DNA sequence alignment

- Homologous site
- Inherited from the common ancestor of all sequences in the alignment
- The aim of sequence alignment is to maximise the number of sites for which you can infer homology

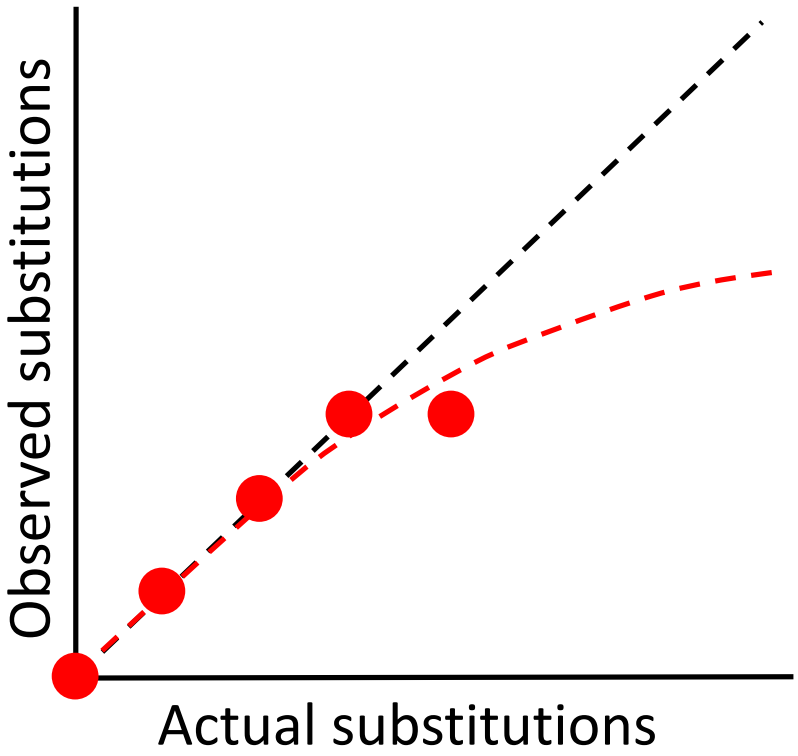


DNA sequence alignment

- Indel – insertion or deletion
- Potentially informative
- Most phylogenetic methods do not really use indel data
- Maximum-likelihood and Bayesian methods typically treat them in the same way as missing data



2. Substitution Models

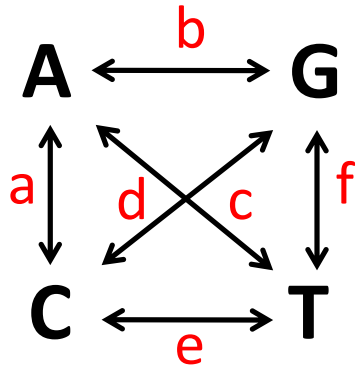



```
A  A  A  A  A
A  T  T  T  T
C  C  G  G  G
A  A  A  A  A
T  T  T  T  T
T  T  T  T  T
A  A  A  A  A
G  G  G  G  G
T  T  T  A  C
```

- Need to correct for multiple substitutions at the same site
- Otherwise: problem known as **long-branch attraction**
  - Long branch = many substitutions
  - Similarities arise by chance
  - Long branches cluster together

Nucleotide substitution models

Rate Matrix

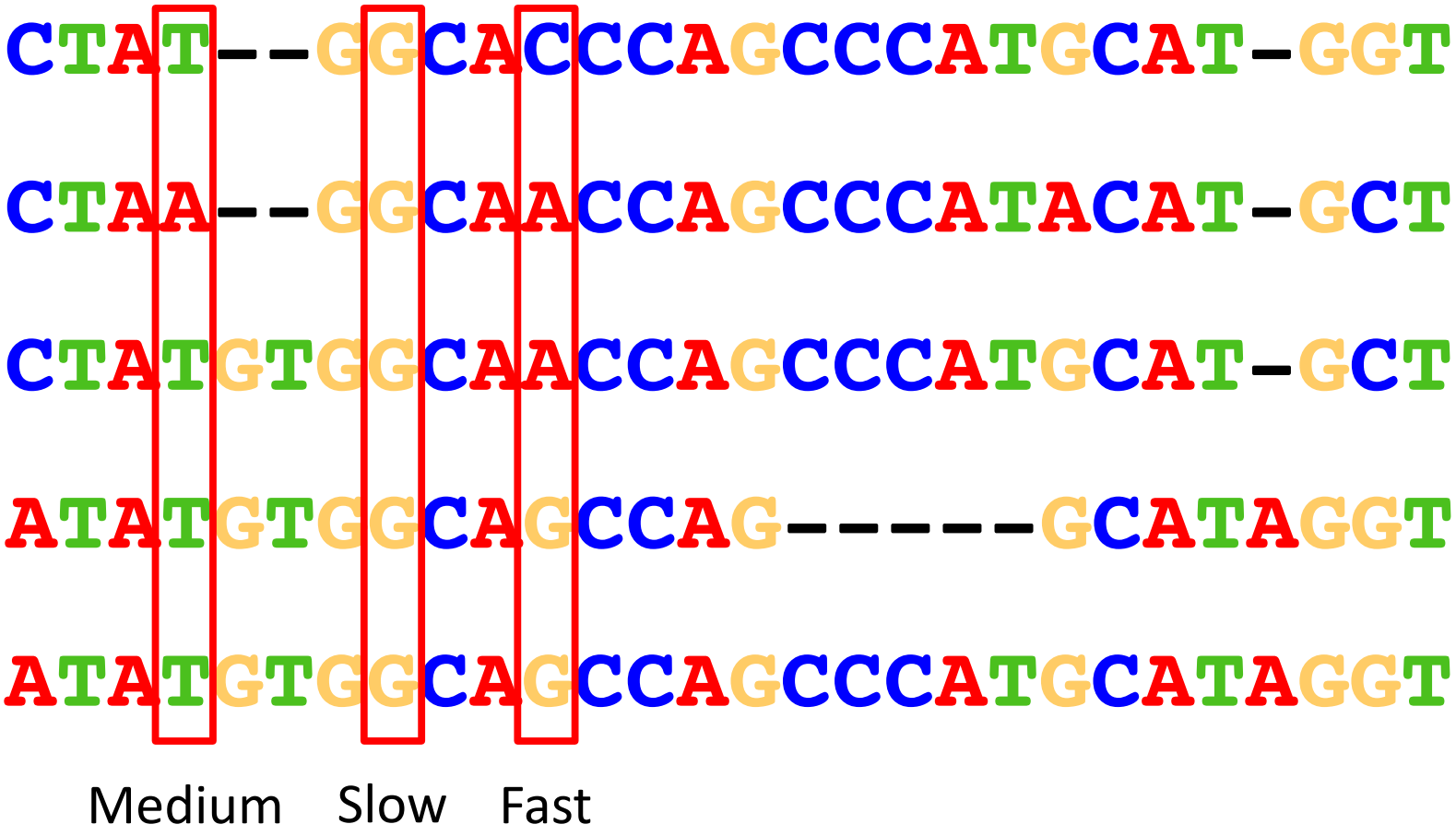


Base Frequencies

$\pi_A + \pi_C + \pi_G + \pi_T = 1$

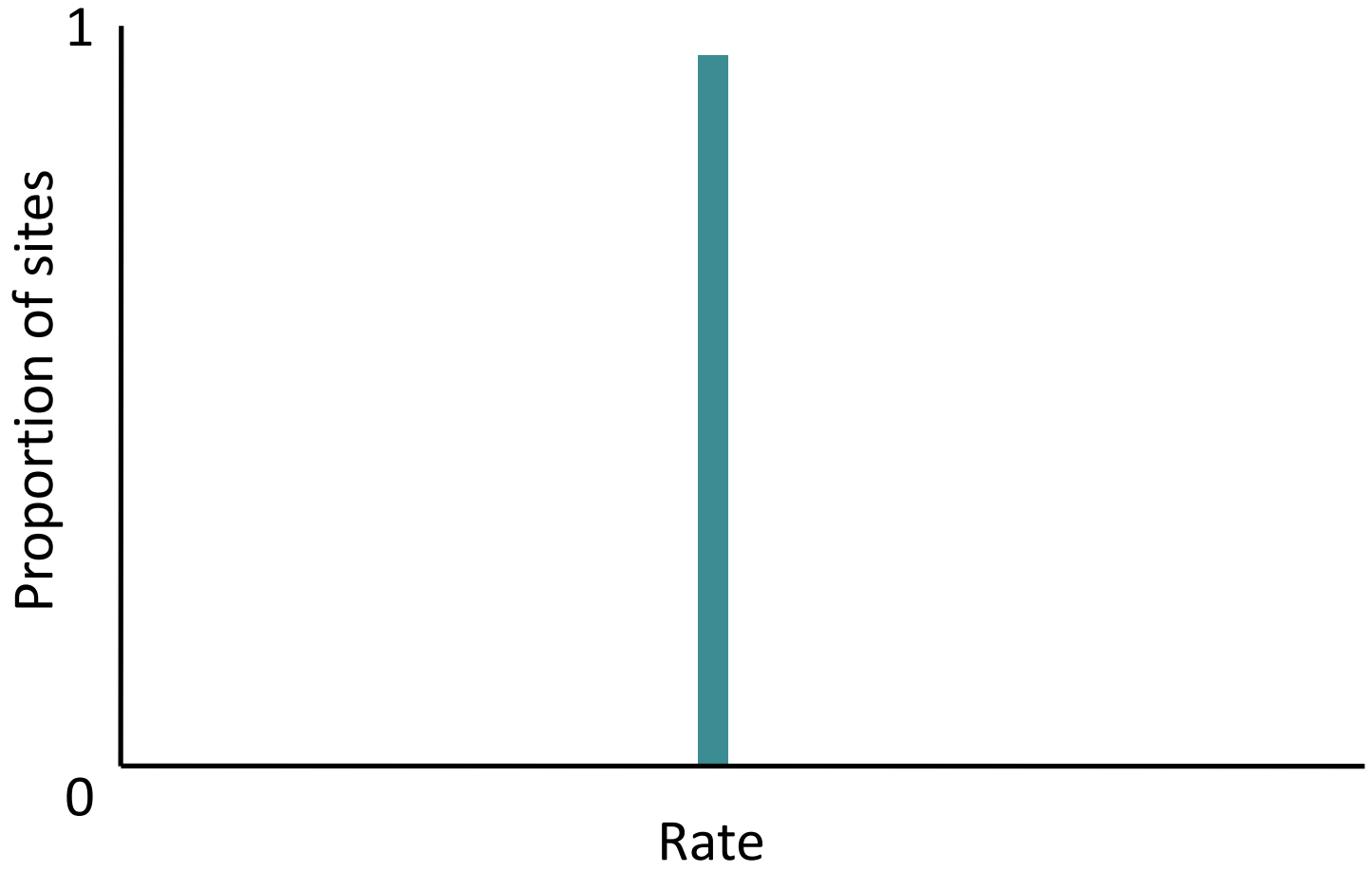
JC	HKY	GTR
a=b=c=d=e=f	a=c=d=f, b=e	a, b, c, d, e, f
$\pi_A=\pi_C=\pi_G=\pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$

Rate variation across sites



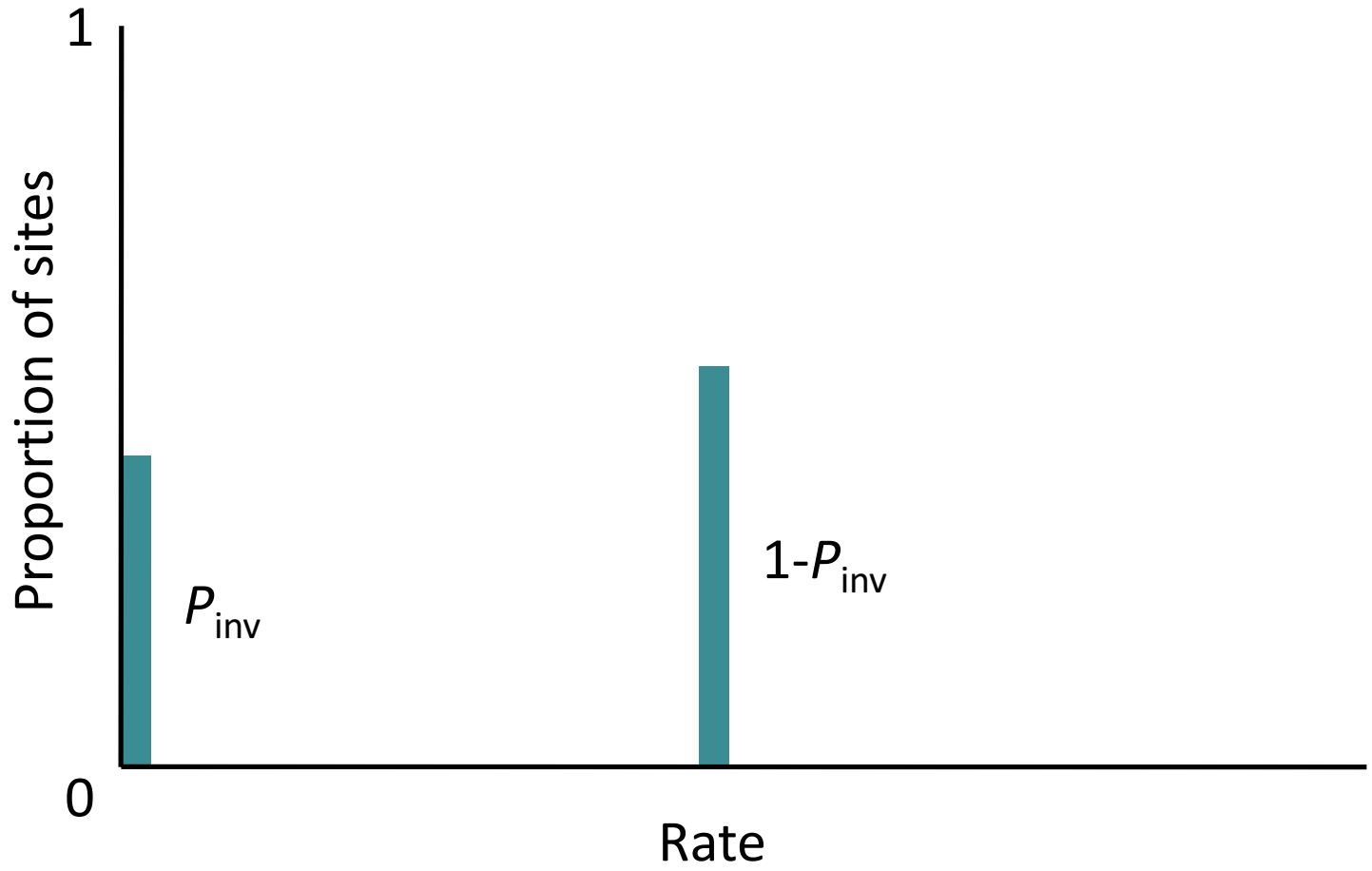
Rate variation across sites

- Equal rates among sites



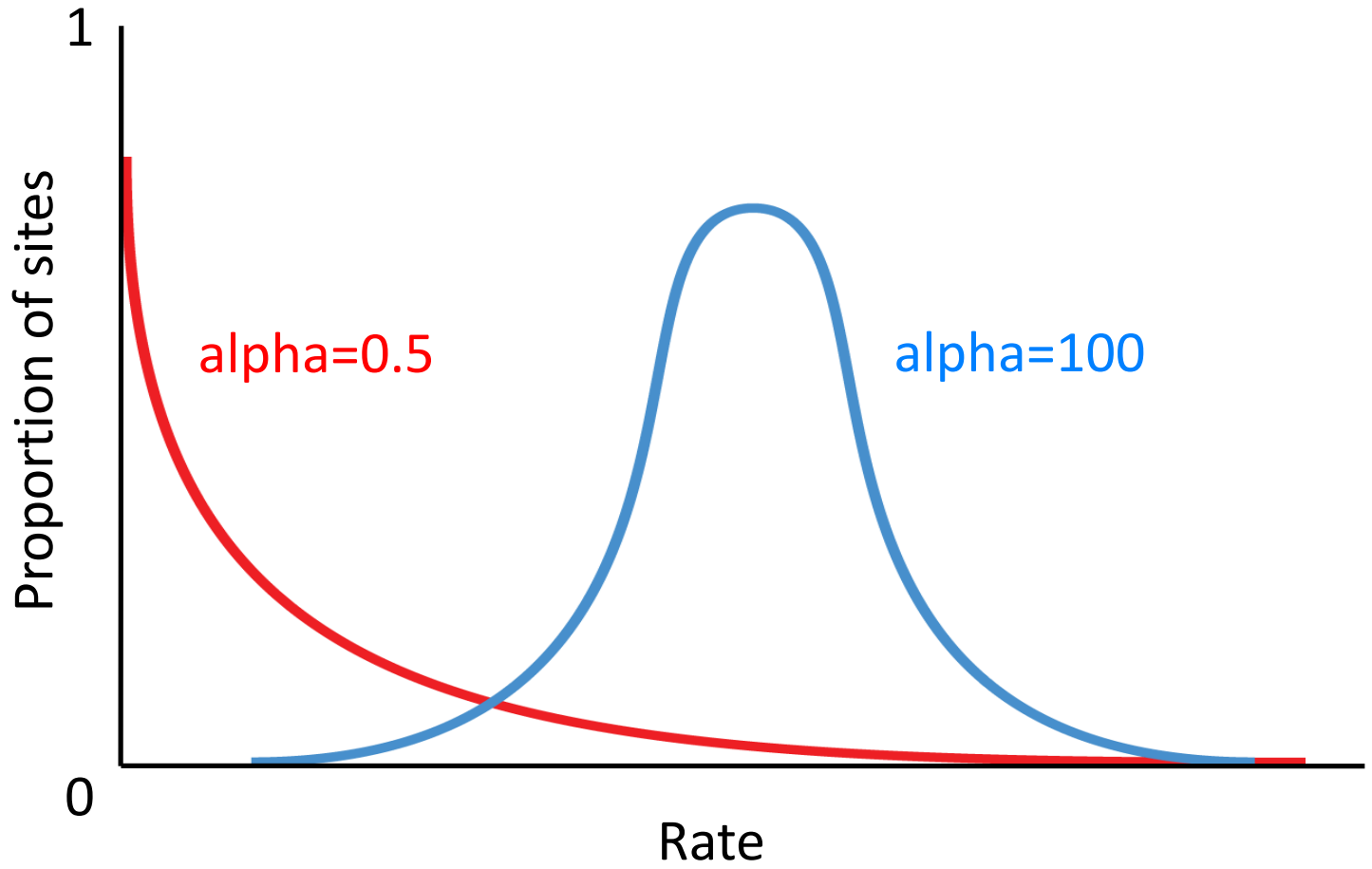
Rate variation across sites

- Proportion of invariable sites (**+I** models)



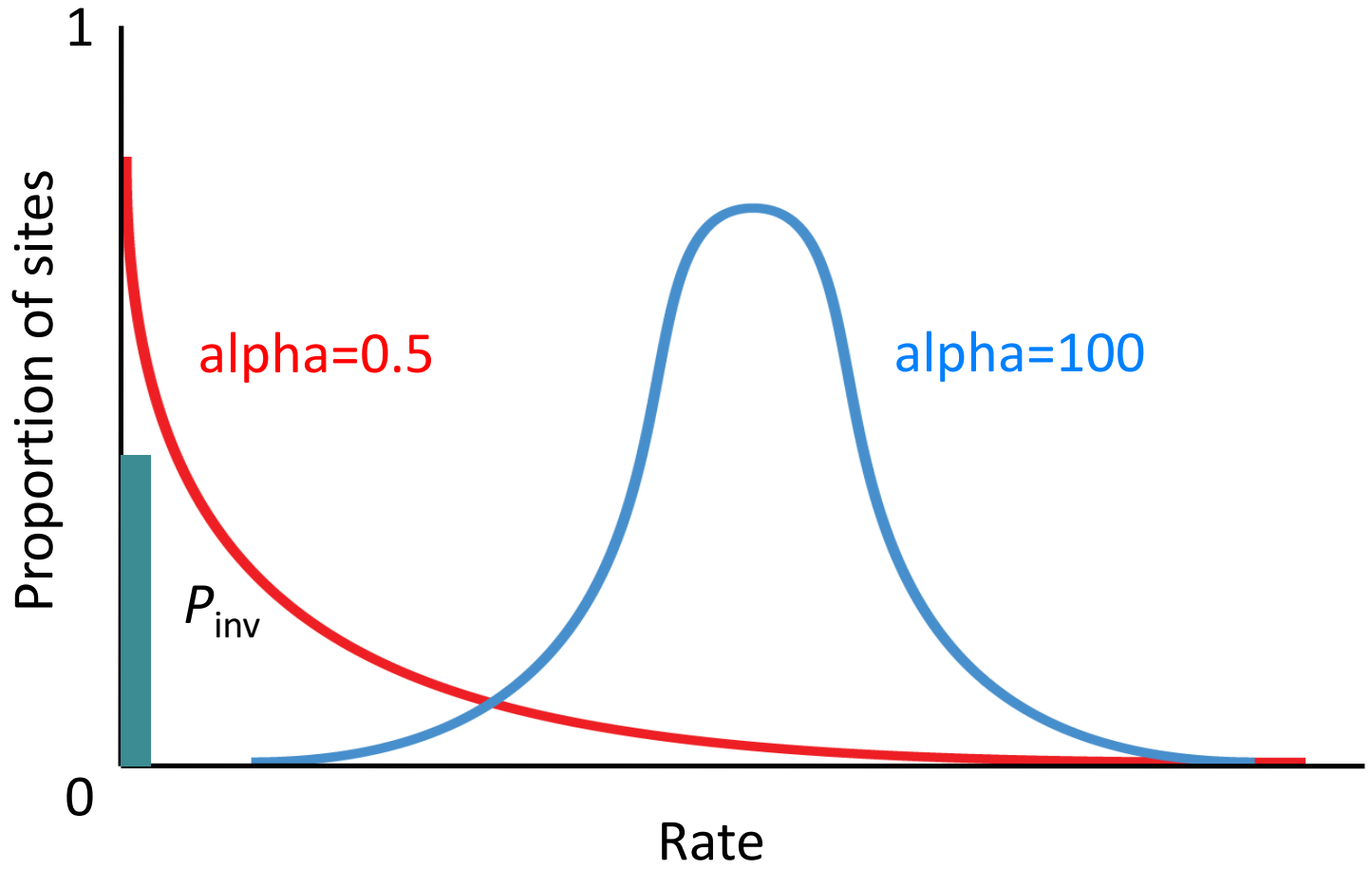
Rate variation across sites

- Gamma-distributed rate variation across sites (**+G** models)



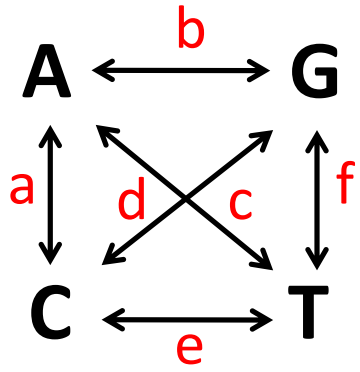
Rate variation across sites

- Gamma-distributed rate variation across sites and a proportion of invariable sites (**+G+I** models)



Nucleotide substitution models

Rate Matrix



Base Frequencies

$\pi_A + \pi_C + \pi_G + \pi_T = 1$

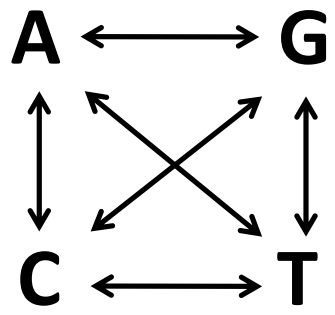
Site Rates

+ I + G

JC	**HKY**	**GTR**	**GTR+I+G**
a=b=c=d=e=f	a=c=d=f, b=e	a, b, c, d, e, f	a, b, c, d, e, f
$\pi_A=\pi_C=\pi_G=\pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$
			I, G

Nucleotide substitution models

Rate Matrix



Base Frequencies

$\pi_A + \pi_C + \pi_G + \pi_T = 1$

Site Rates

+ I + G

Number of models

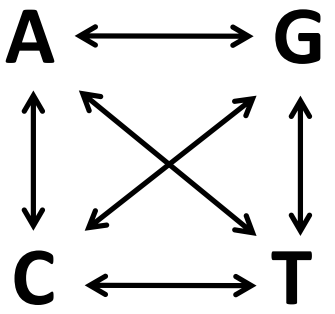
203 x 15 x 4 = 12,180

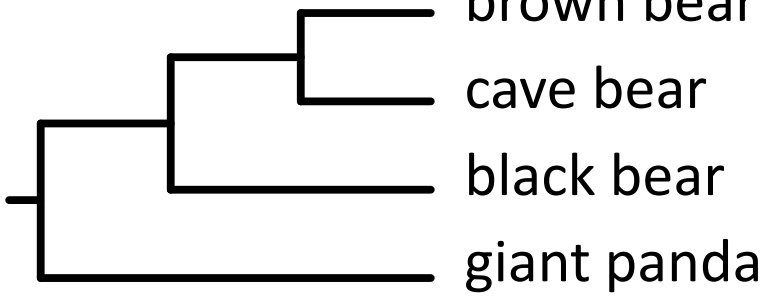
In phylogenetics, we typically consider a small subset of these

Fundamental assumptions

- Stationary
- Reversible
- Homogeneous
- Independent across sites

π_A π_C π_G π_T

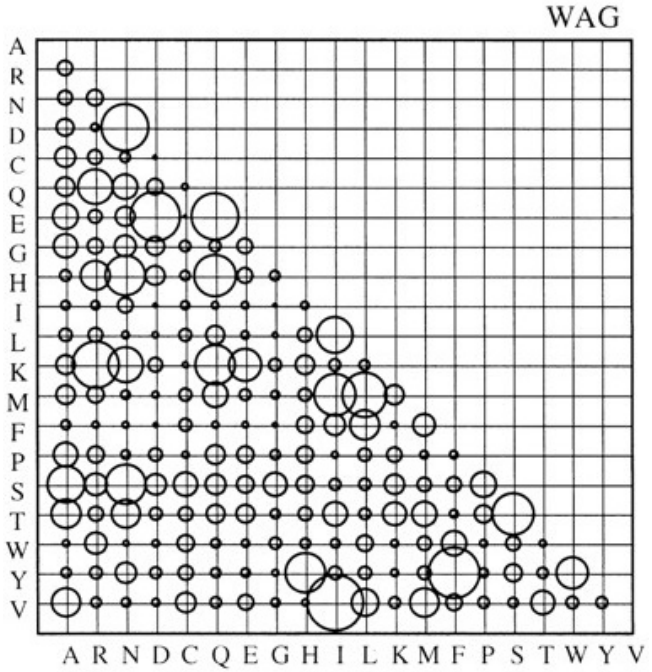




brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

Amino acid substitution matrices

- 20x20 matrix of substitution probabilities
- Too many parameters to estimate
  - GTR model for DNA: 6 parameters
  - GTR model for proteins: 190 parameters
- Estimate substitution probabilities using large data set
  - PAM
  - BLOSUM
  - JTT
  - WAG

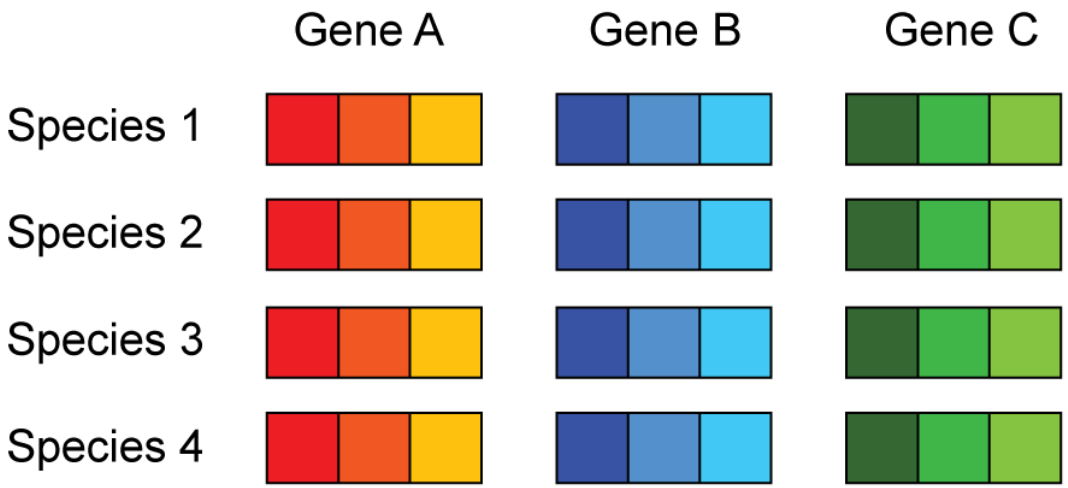


Other models

- Codon models
  - Extends 4-nucleotide state space to 64-codon state space
- Microsatellite data
  - Stepwise mutation model
- Morphological data
  - Generalisation of the JC substitution model
- Language data

Data partitioning

- Separate substitution model for each gene and codon position?



- **Biological**
  - Genome
  - Genes
  - Codon positions
  - RNA stems *vs* loops
  - Hydrophobic *vs* hydrophilic
- **Statistical**

3. Bayesian Phylogenetics

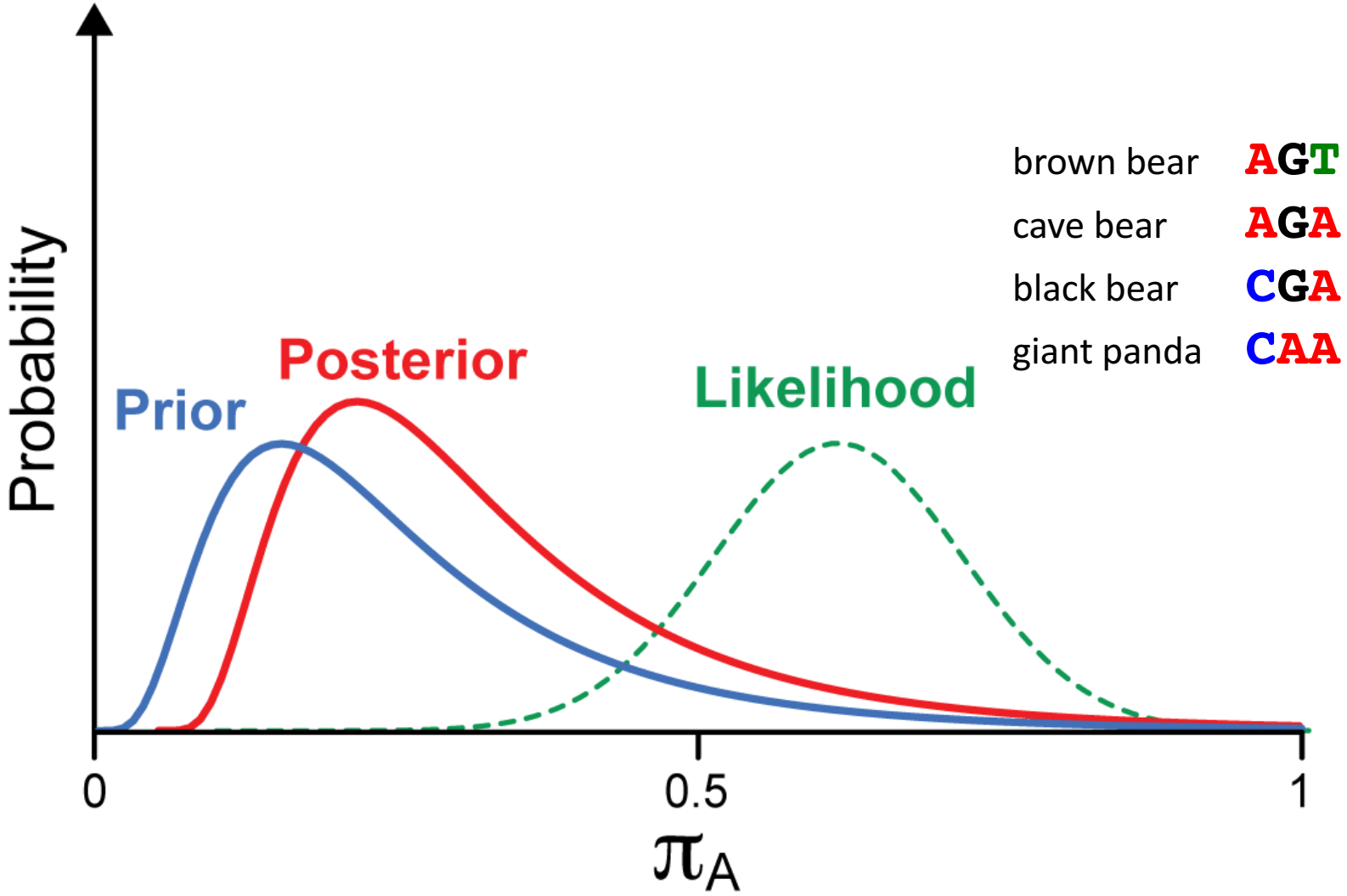
Bayesian phylogenetic analysis

- Bayesian phylogenetic analysis was developed in the mid 1990s
- Now one of the most widely used methods

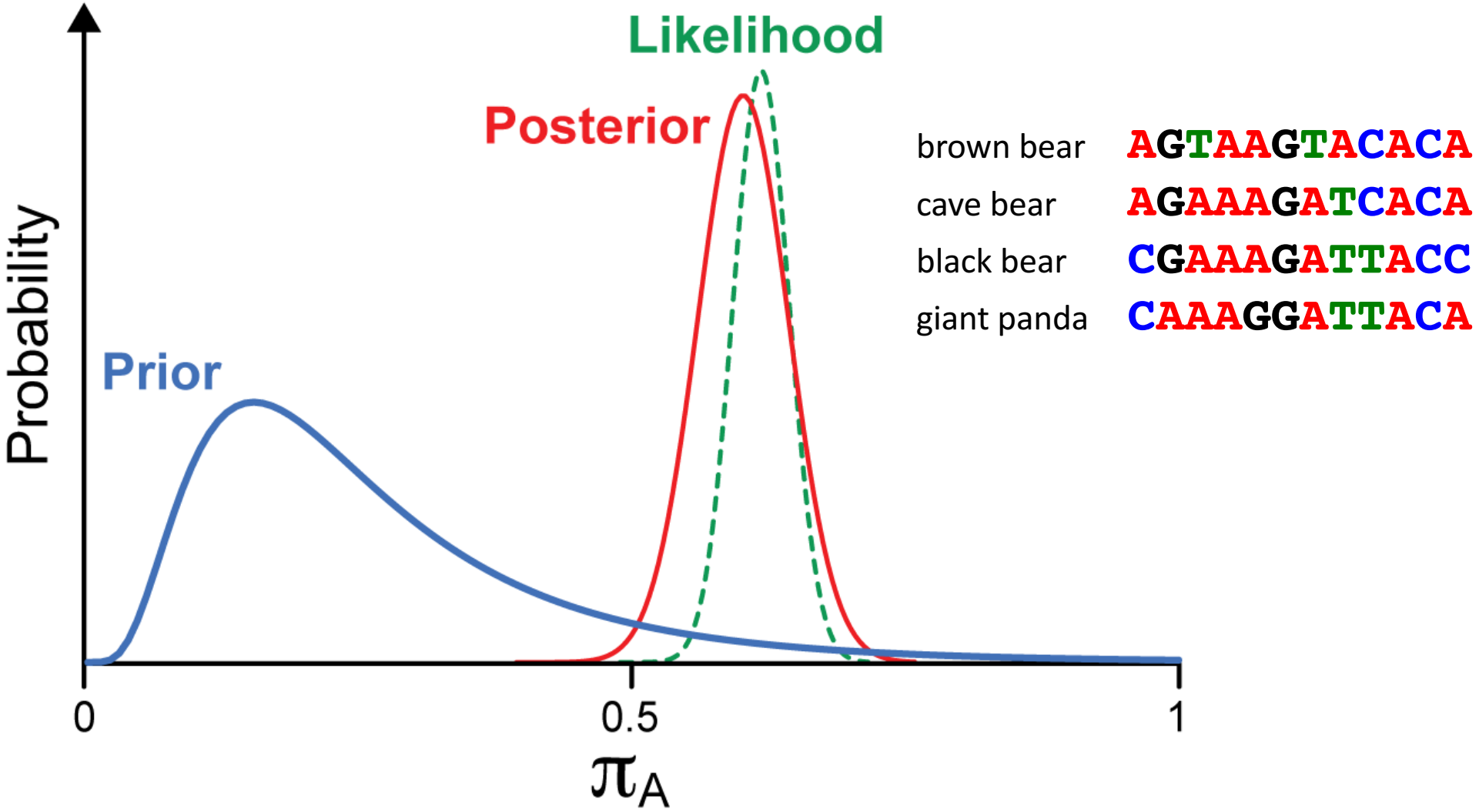
The Bayesian paradigm

- Parameters have **distributions**
- Before observing the data, each parameter has a **prior distribution**
- The **likelihood** of the data is computed
- The prior distribution is combined (updated) with the likelihood to yield the **posterior distribution**

Simple example



Simple example

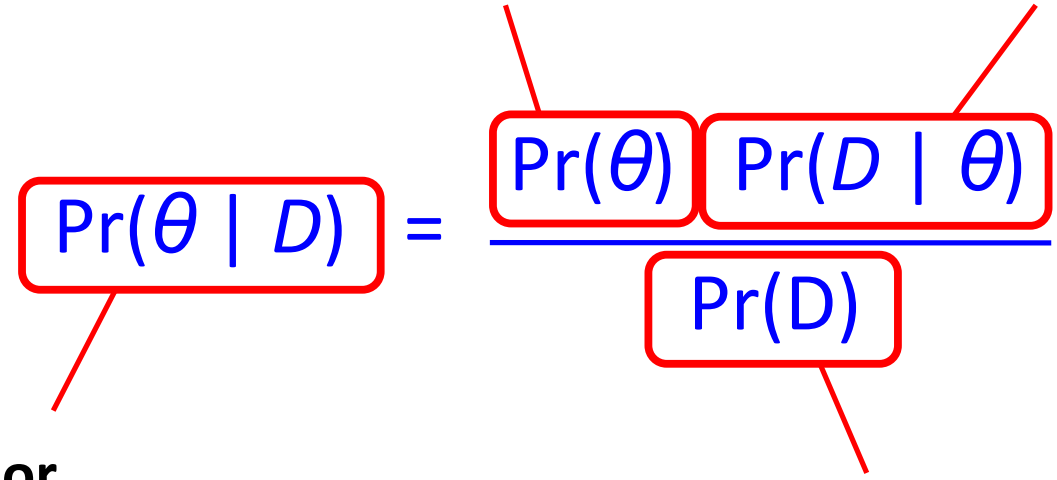


brown bear	AGTAAGTACACA
cave bear	AGAAAGATCACA
black bear	CGAAAGATTACC
giant panda	CAAAGGATTACA

Bayesian inference

Prior
Specified by user,
independent of data

Likelihood
Calculated from data



Posterior
This is what we
want to estimate

normalising constant
marginal likelihood of the data
model likelihood

4. Priors

Priors

- Priors are chosen in the form of probability distributions
- Reflect our prior expectations (and uncertainty) about values of parameters (without knowledge of the data)
  - Past observations
  - Personal beliefs
  - Use of a biological model
- Uninformative priors
  - Uniform prior
  - Jeffreys prior
  - Reference prior

Continuous distributions

- Uniform
- Normal

→ Used to specify prior distributions of various continuous parameters

- Exponential
- Lognormal
- Gamma

→ Used to specify prior distributions of continuous parameters that cannot take negative values

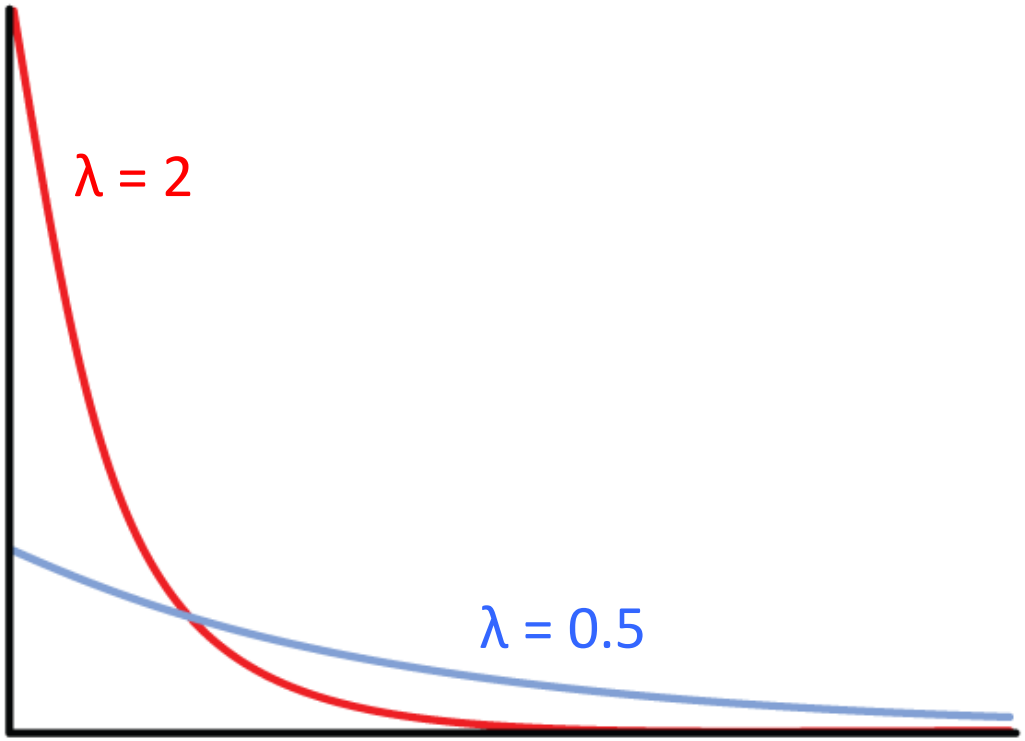
- Beta

Continuous distributions

- Uniform
- Normal
- **Exponential**
- Lognormal
- Gamma
- Beta

Parameters

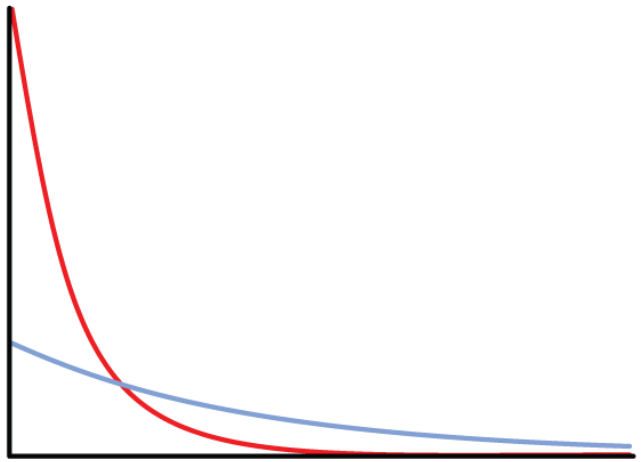
- λ = rate of decay



Hyperpriors

- Prior distributions can have unknown parameters
- Assign priors to these parameters
  - Hyperparameters
- Usually only 2 or 3 levels because effect becomes unimportant

- Example:
  Uniform prior on λ in exponential prior on mutation rate



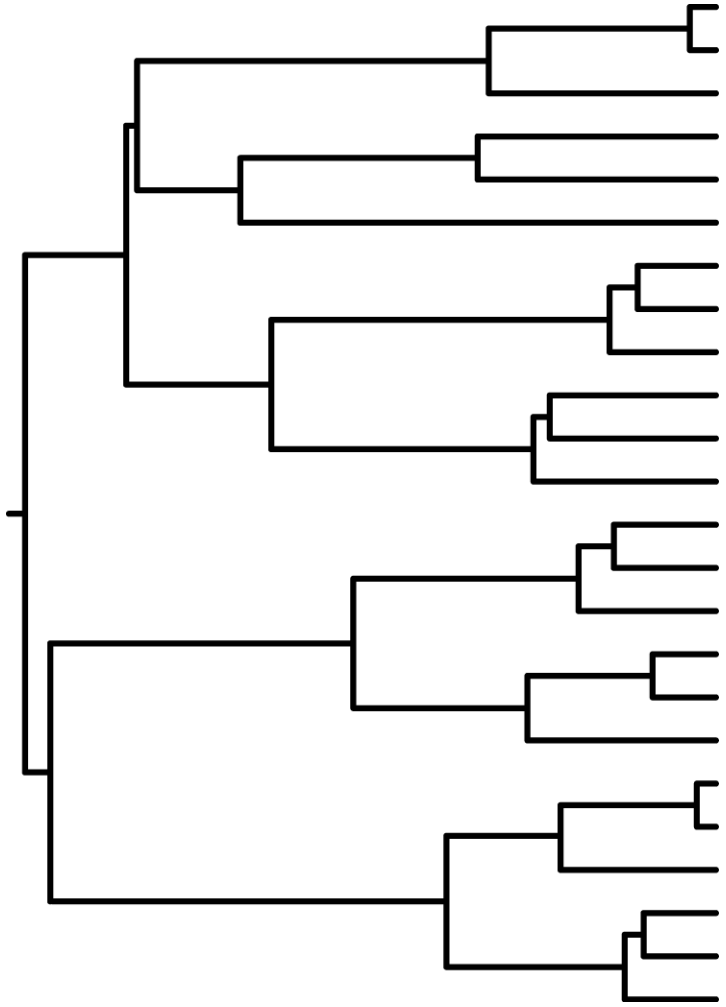
Tree priors

1. Use a **flat prior** for tree topology (*MrBayes*)
   - All trees have equal probability
   - Also need a prior for branch lengths or node times

2. Use a **biological model** to generate prior distribution (*BEAST* and *MrBayes*)
   - Among species: speciation model
   - Within species: coalescent model

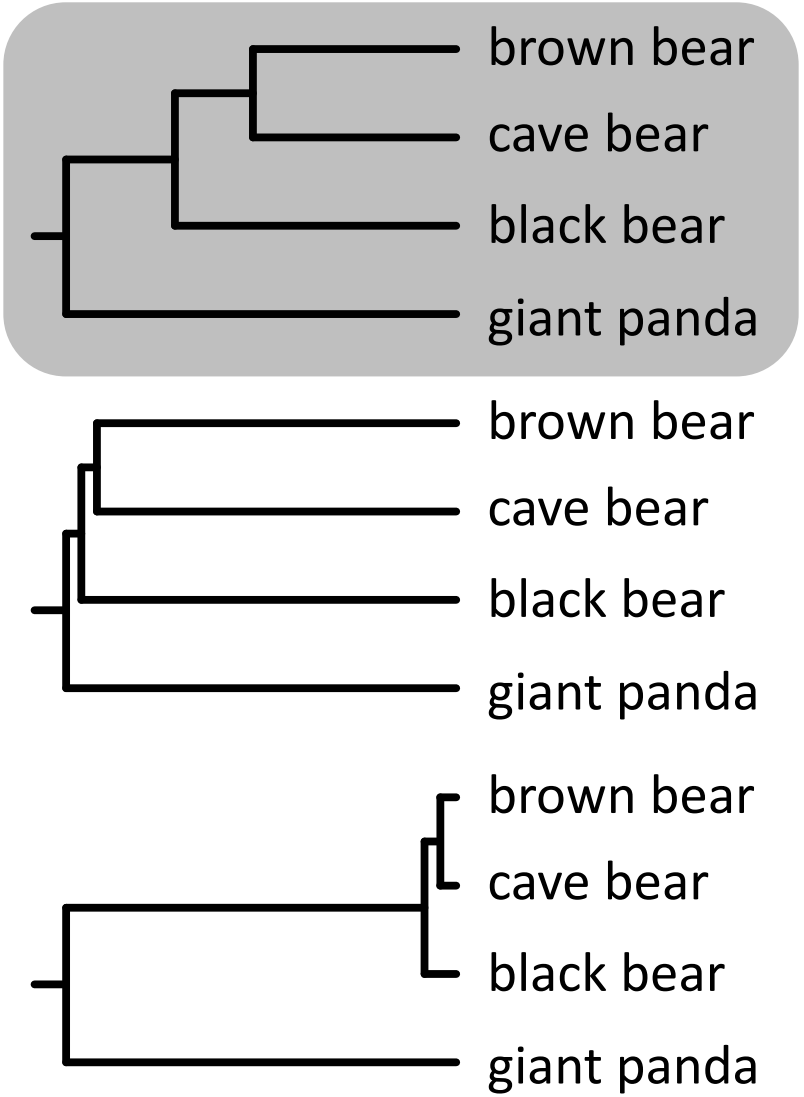
Tree priors: Speciation model

- Tree shape described by a stochastic branching process
- **Yule process**
  - The root lineage splits into two
  - Lineages split at a constant rate
  - Simulates speciation process
- **Birth-death process**
  - Allow lineages to go extinct



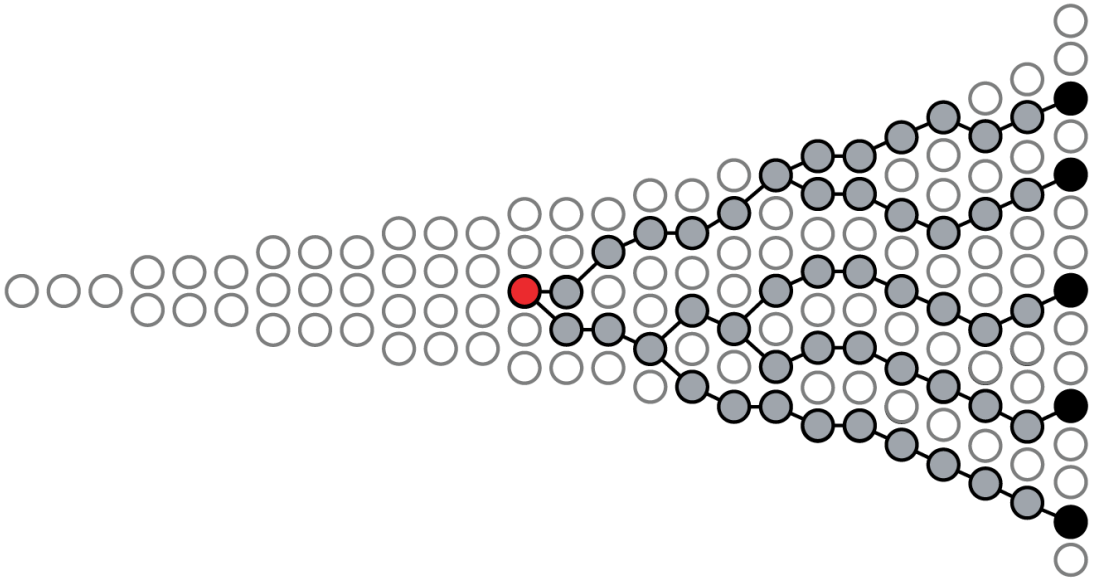
Tree priors: Speciation model

- Tree shape described by a stochastic branching process
- **Yule process**
  - The root lineage splits into two
  - Lineages split at a constant rate
  - Simulates speciation process
- **Birth-death process**
  - Allow lineages to go extinct



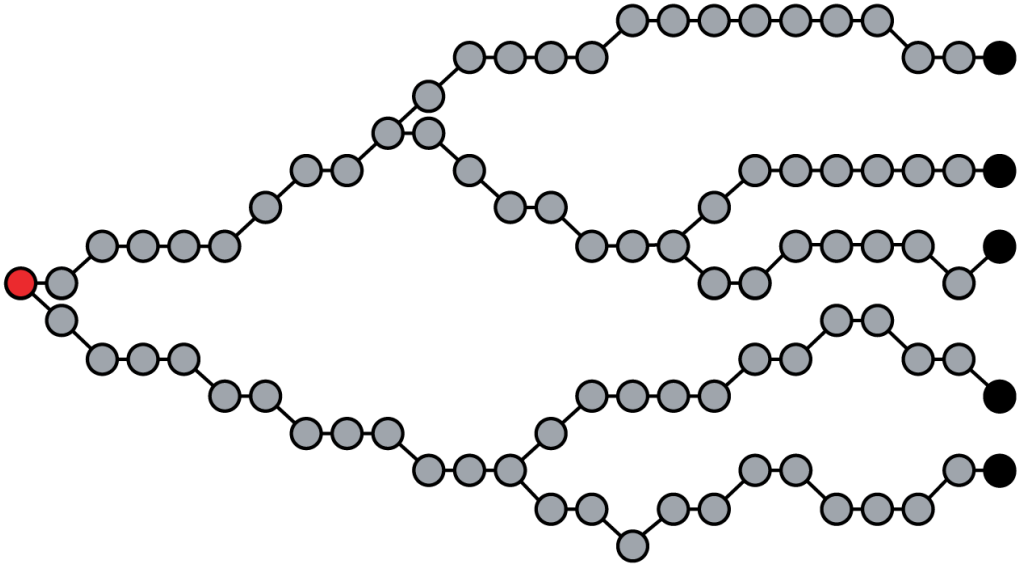
Tree priors: Coalescent model

- Coalescent model used to put a prior on the tree
- Time between coalescent events depends on population size

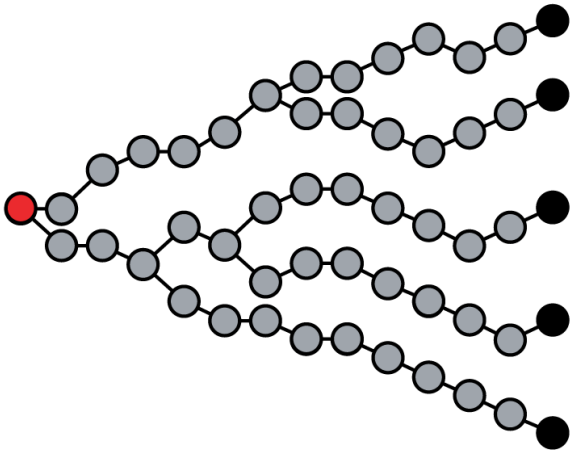


Tree priors: Coalescent model

Constant size



Exponential growth



5. Likelihood

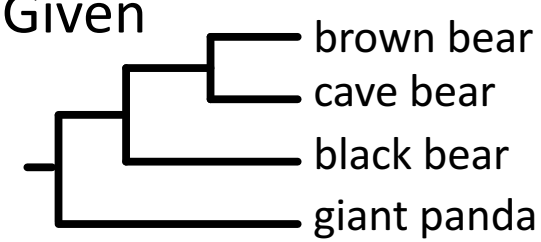
Bayesian inference

Prior
Specified by user, independent of data

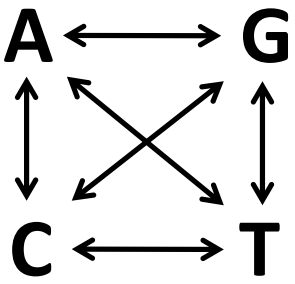
Likelihood
Calculated from data

$$\Pr(\theta \mid D) = \frac{\Pr(\theta)\,\Pr(D \mid \theta)}{\Pr(D)}$$

Given



\+



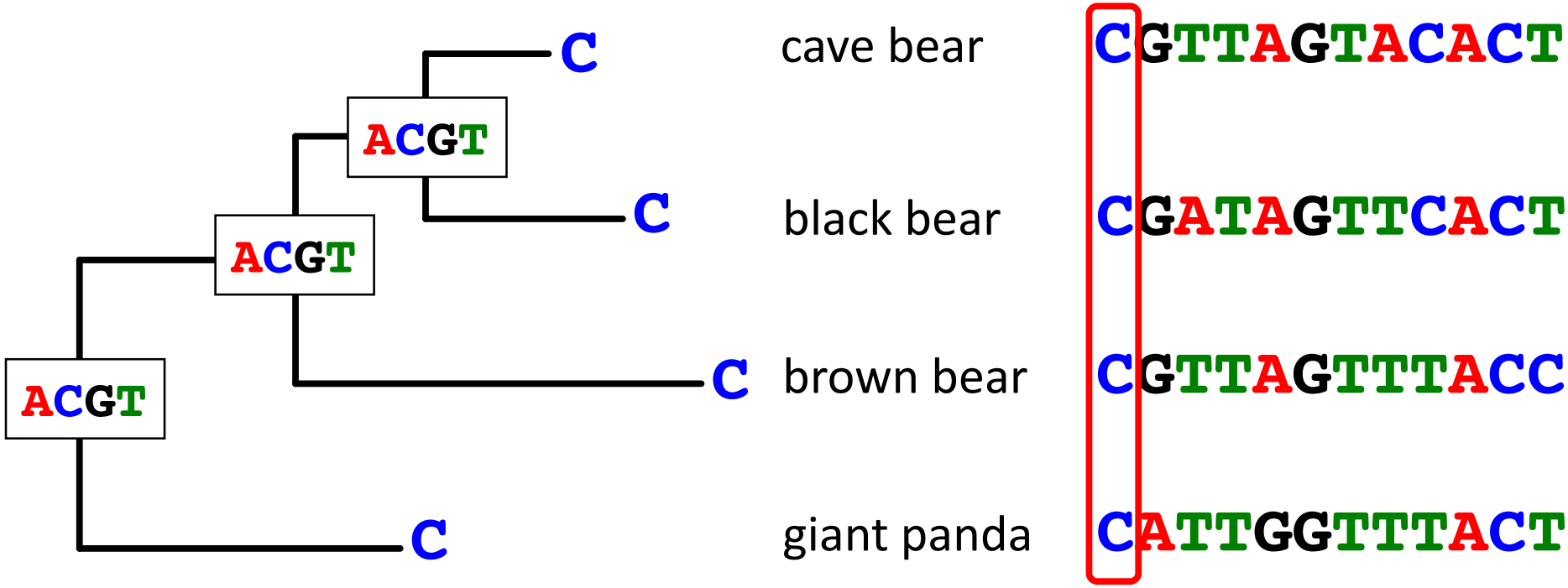
Probability of?

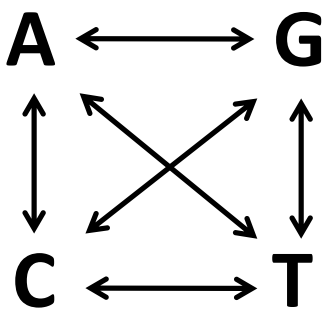
brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

Likelihood

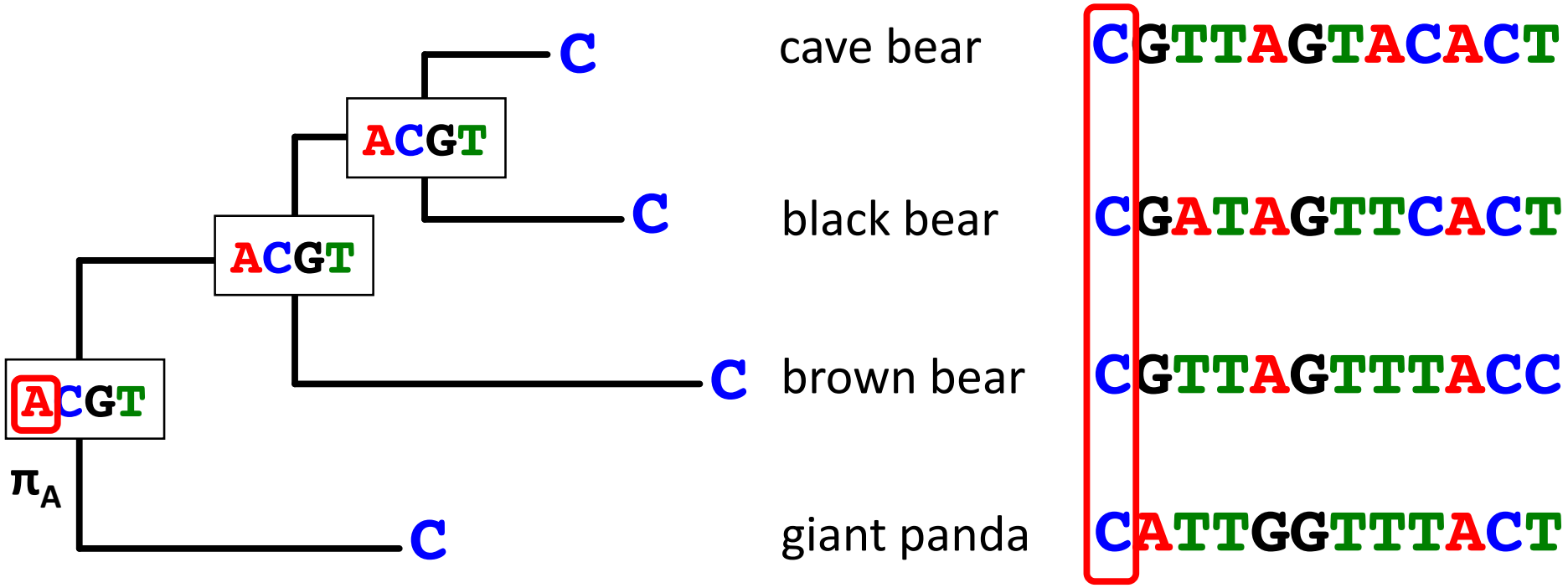
cave bear	CGTTAGTACACT
black bear	CGATAGTTCACT
brown bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

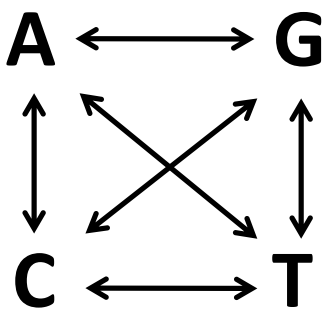
Likelihood


C
ACGT
C
ACGT
C
ACGT
C
cave bear
black bear
brown bear
giant panda
CGTTAGTACACT
CGATAGTTCACT
CGTTAGTTTACC
CATTGGTTTACT

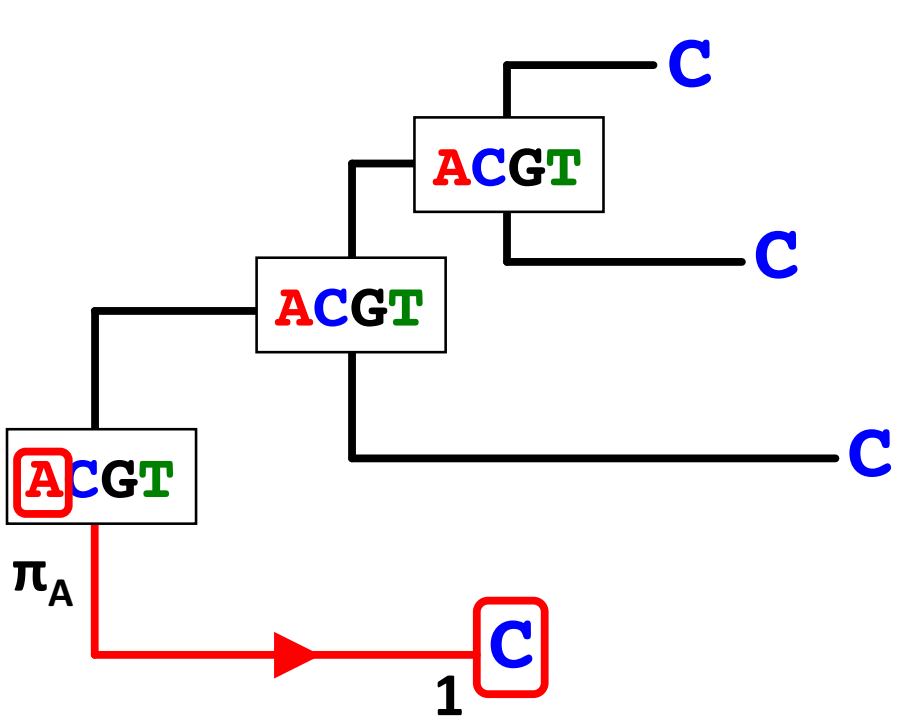

A
G
C
T

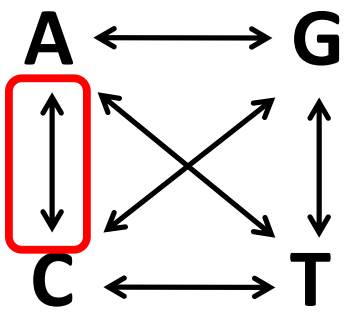
Likelihood


C
ACGT
C
ACGT
C
ACGT
π_A
C
cave bear
black bear
brown bear
giant panda
CGTTAGTACACT
CGATAGTTCACT
CGTTAGTTTACC
CATTGGTTTACT


A
G
C
T

Likelihood





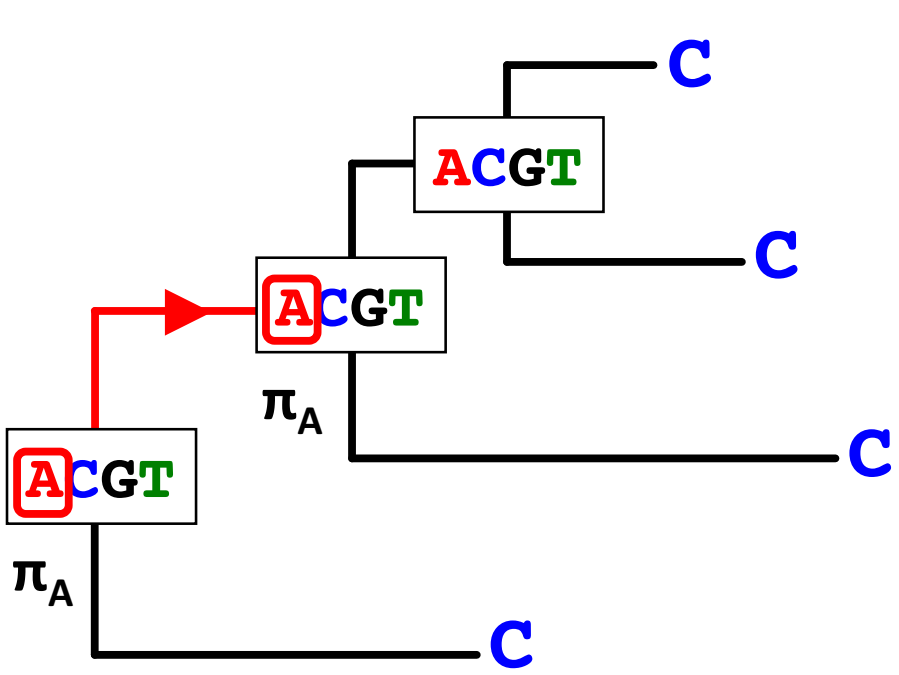
cave bear CGTTAGTACACT

black bear CGATAGTTCACT

brown bear CGTTAGTTTACC

giant panda CATTGGTTTACT

Likelihood

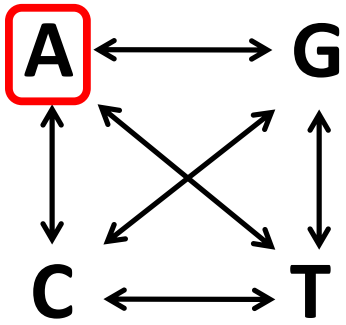


cave bear CGTTAGTACACT

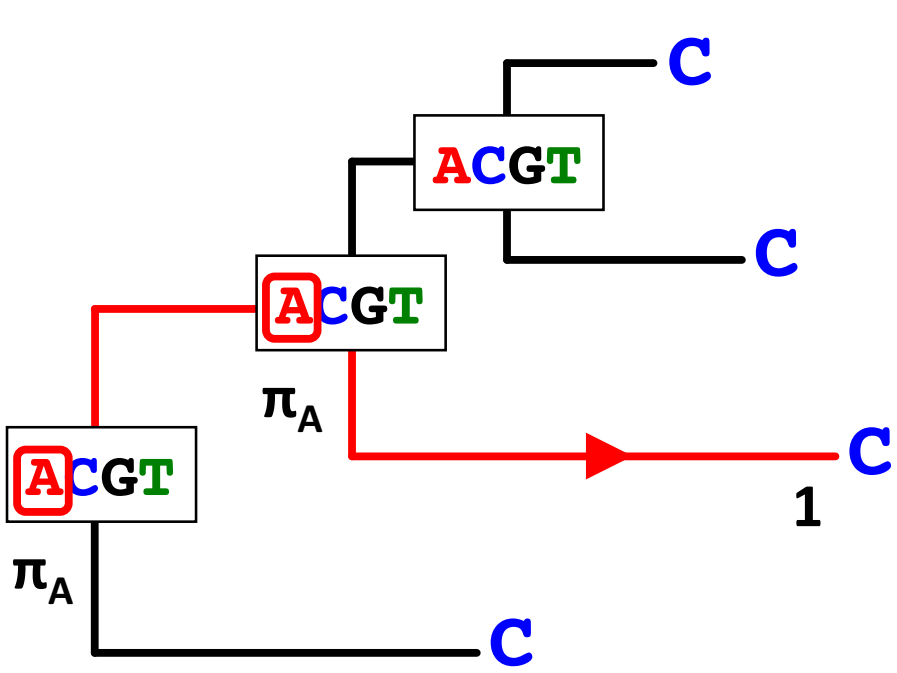
black bear CGATAGTTCACT

brown bear CGTTAGTTTACC

giant panda CATTGGTTTACT



Likelihood

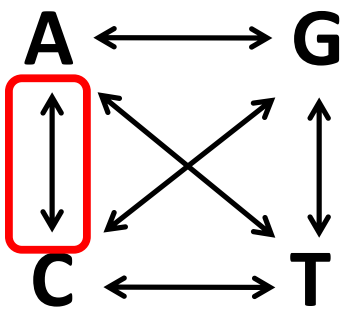


cave bear CGTTAGTACACT

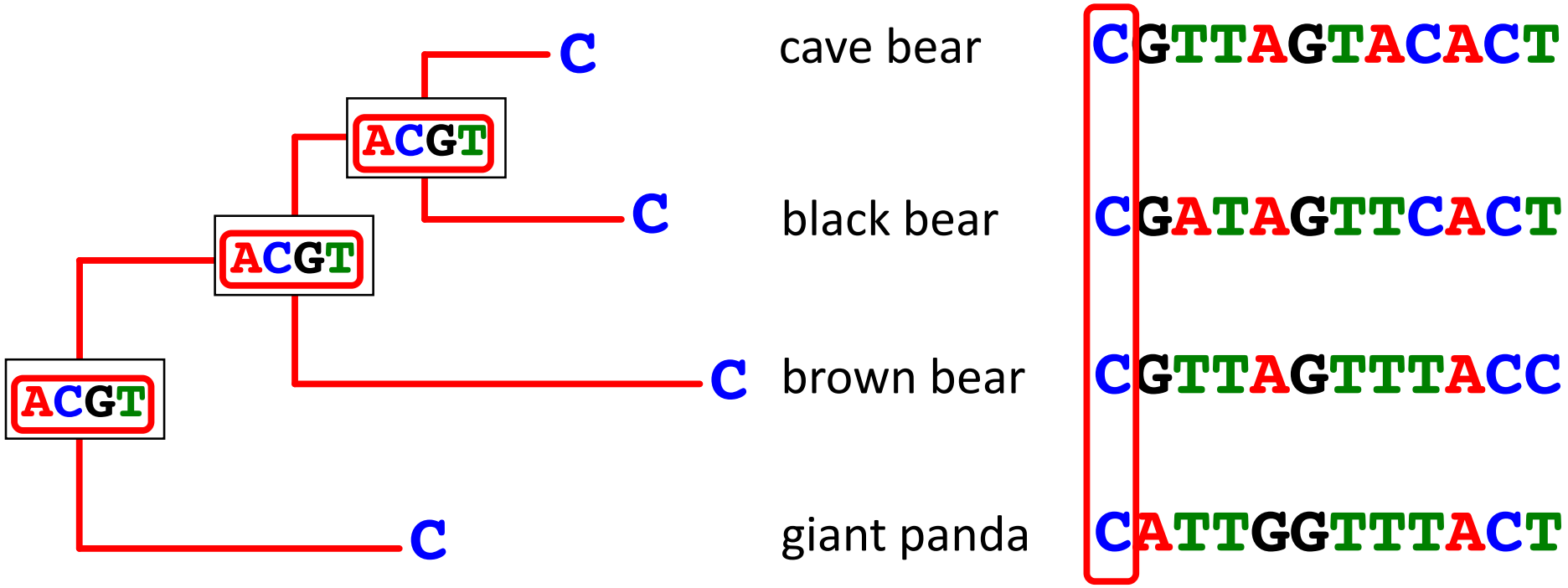
black bear CGATAGTTCACT

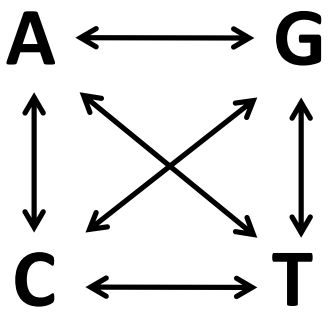
brown bear CGTTAGTTTACC

giant panda CATTGGTTTACT



Likelihood



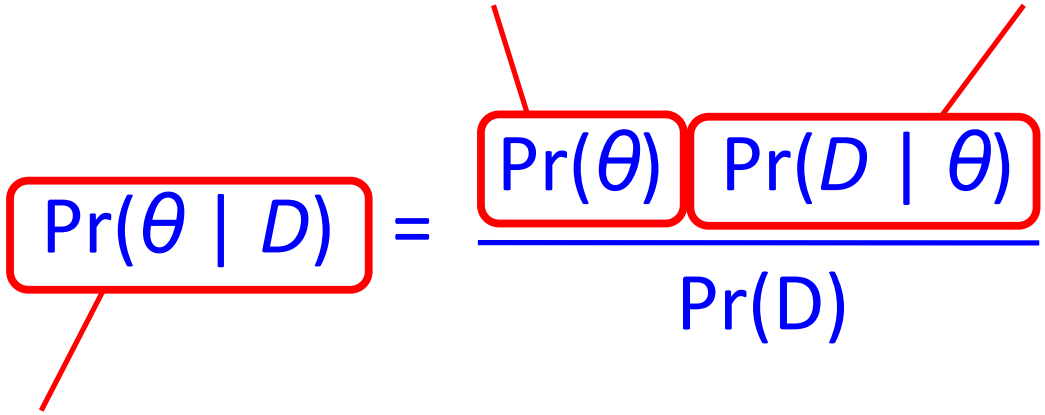


Likelihood is summed over all possibilities

Bayesian inference

Prior
Specified by user, independent of data

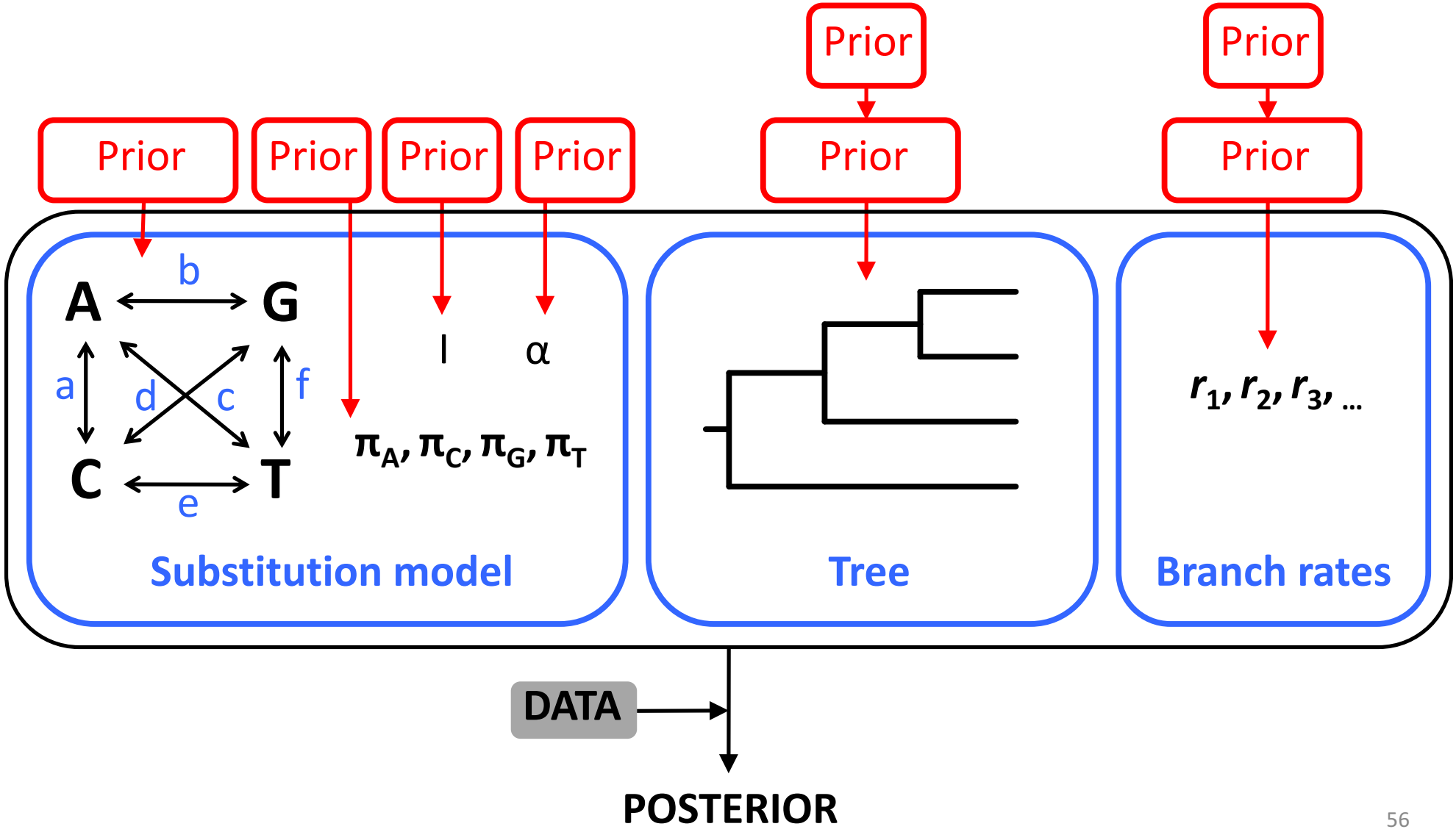
Likelihood
Calculated from data



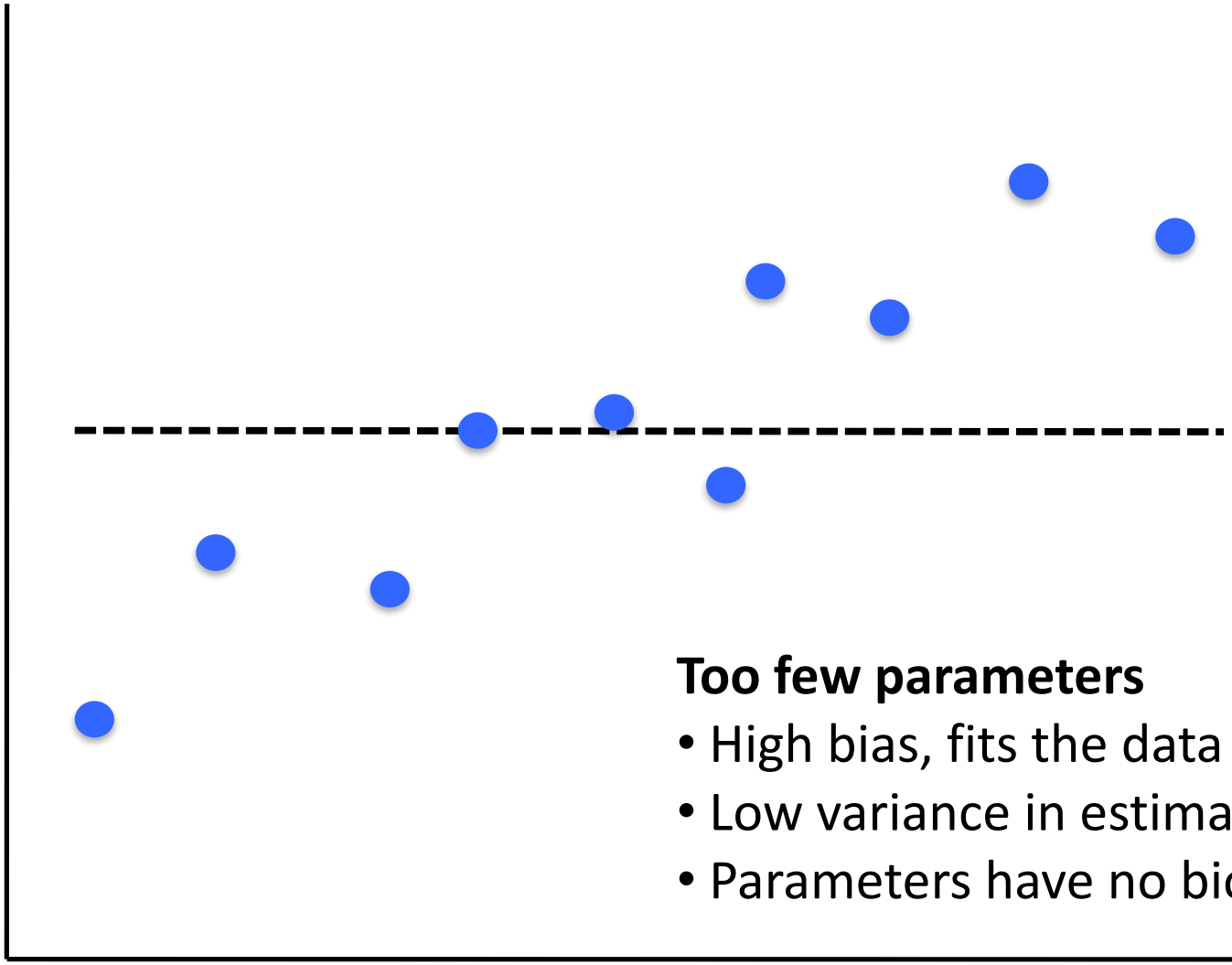
Posterior
This is what we want to estimate

6. Model Selection

Bayesian hierarchical model



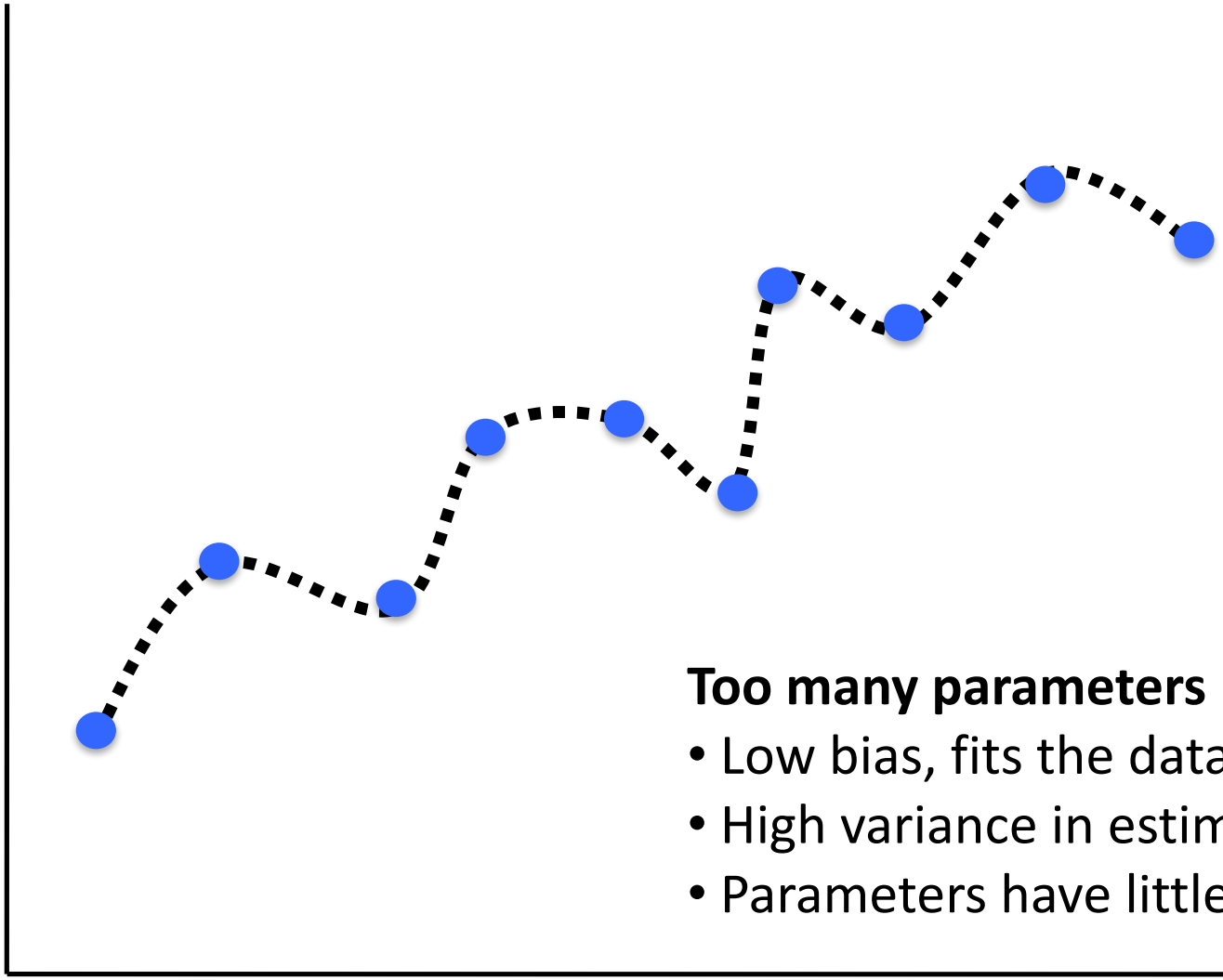
Model selection



Too few parameters

- High bias, fits the data poorly
- Low variance in estimate
- Parameters have no biological meaning

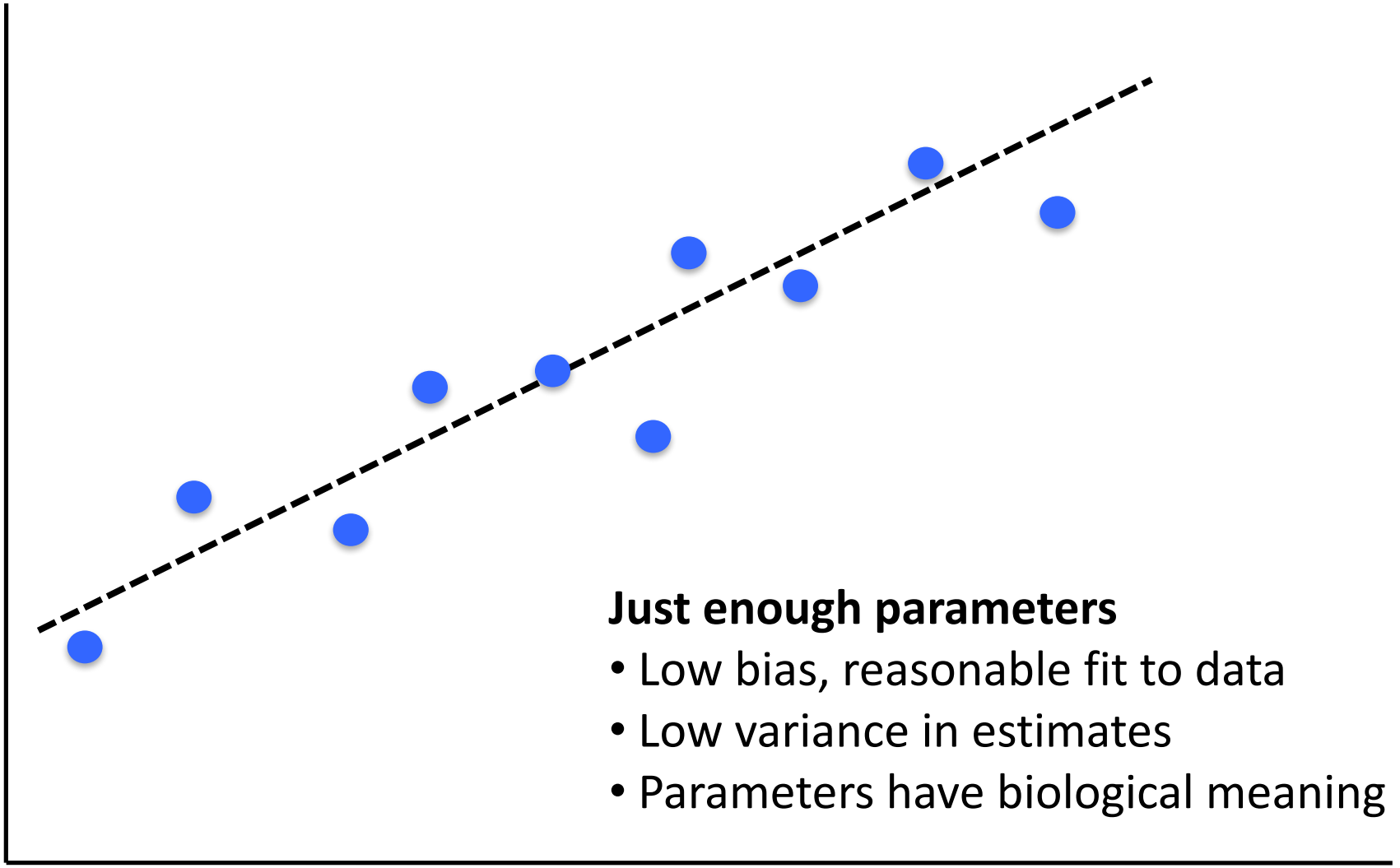
Model selection



Too many parameters

- Low bias, fits the data very well
- High variance in estimates
- Parameters have little biological meaning

Model selection



Just enough parameters

- Low bias, reasonable fit to data
- Low variance in estimates
- Parameters have biological meaning

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

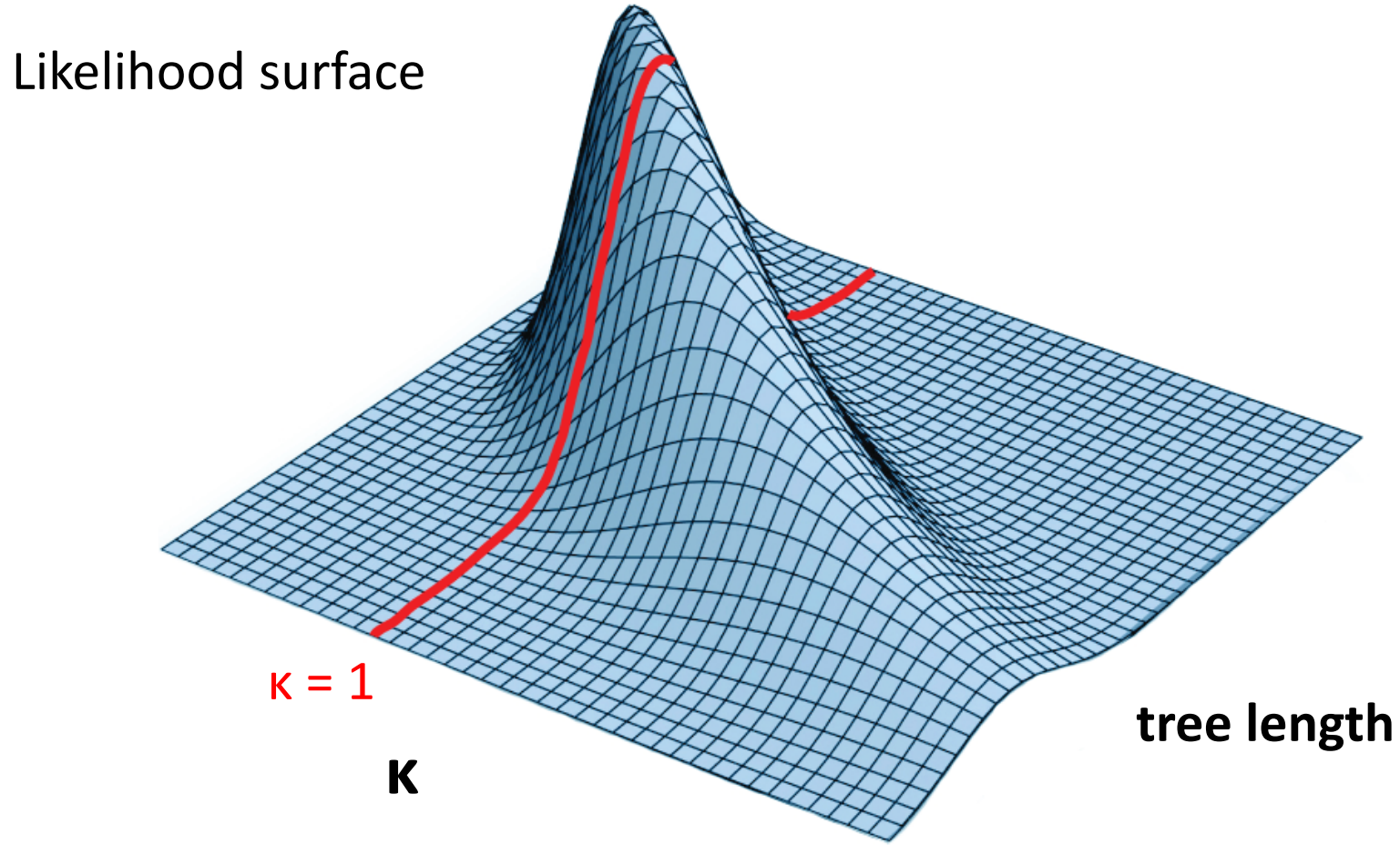
- Compare marginal likelihoods of competing models
- Ratio of marginal likelihoods is the **Bayes factor**

$$\mathbf{BF} = \frac{\mathbf{Pr(D|M_1)}}{\mathbf{Pr(D|M_2)}}$$

$$\mathbf{logBF = logPr(D|M_1) - logPr(D|M_2)}$$

- Models do not need to be nested
- Do not need to correct for number of parameters

Bayesian model selection



Bayesian model selection

- Interpreting the Bayes factor

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 & Raftery (1995) *J Am Stat Assoc*

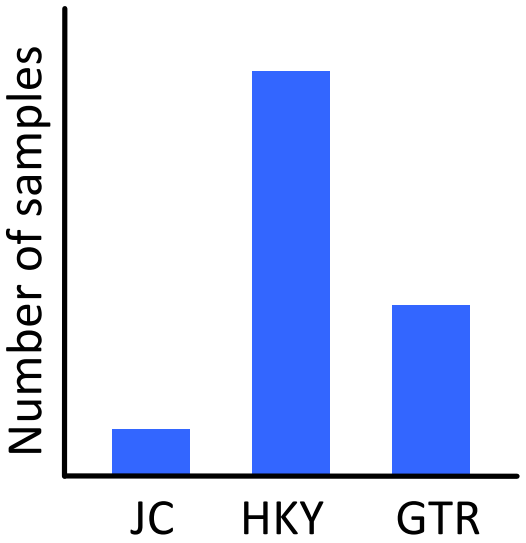
Model adequacy

- Current practice is to compare models from a set of candidates
- Model selection involves comparing the relative fit to the data

What if all of the models being considered are inadequate?

Model averaging

- Treat models as random variables
- Useful when there is uncertainty in model choice
- Sample models with frequencies equal to their posterior probabilities



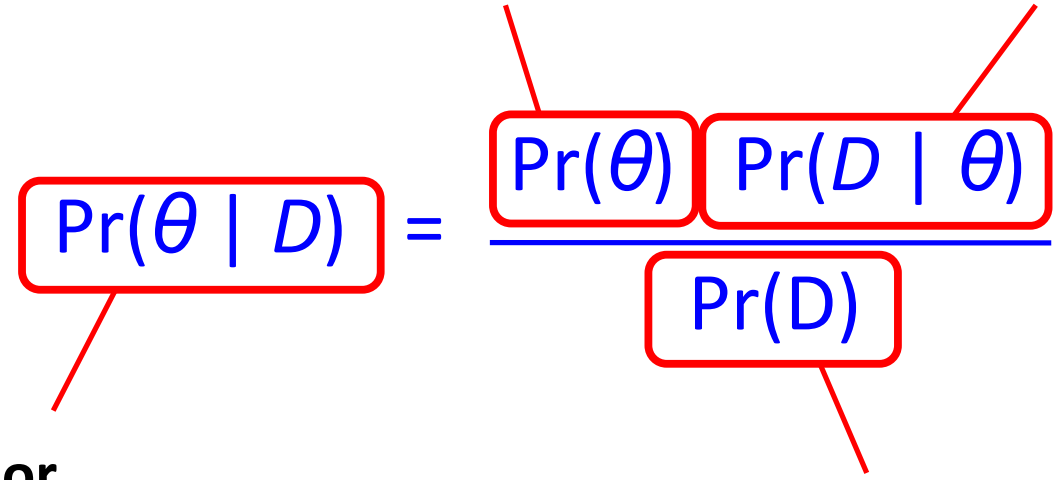
- Parameter estimates are weighted by the probabilities of the models that are considered

7. Markov Chain Monte Carlo Sampling

Bayesian inference

Prior
Specified by user,
independent of data

Likelihood
Calculated from data

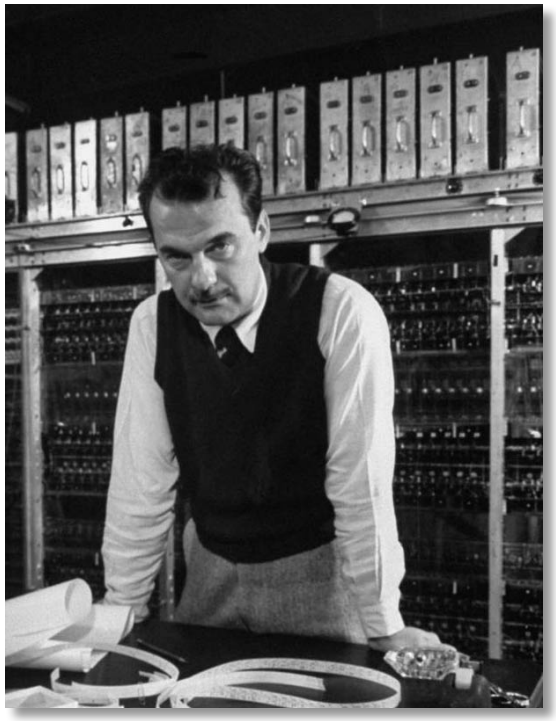


Posterior
This is what we
want to estimate

normalising constant
marginal likelihood of the data
model likelihood

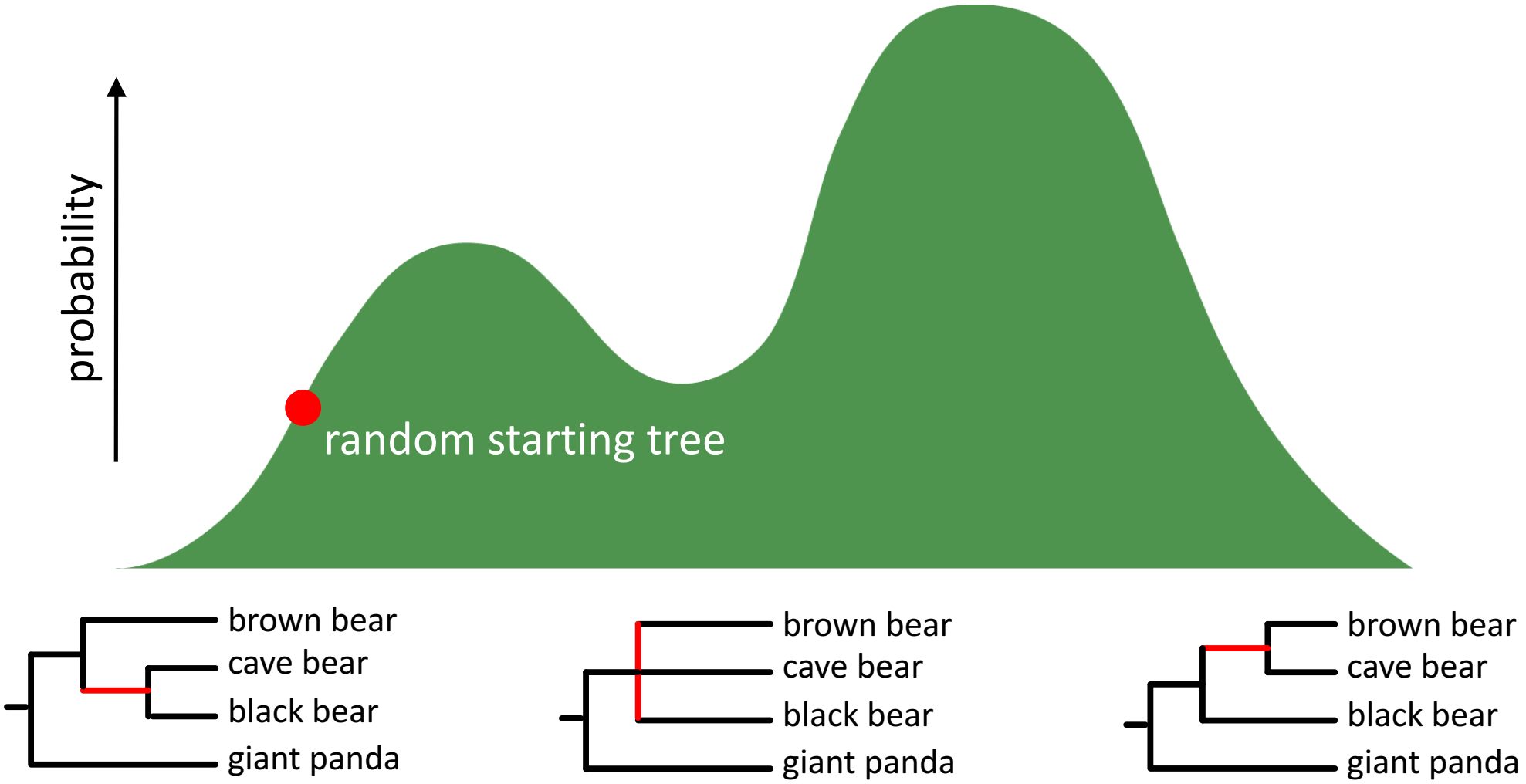
Estimating the posterior

- Impossible to obtain the posterior directly
- Instead, the posterior can be estimated using **Markov chain Monte Carlo simulation**
- This is usually done using the **Metropolis-Hastings algorithm**

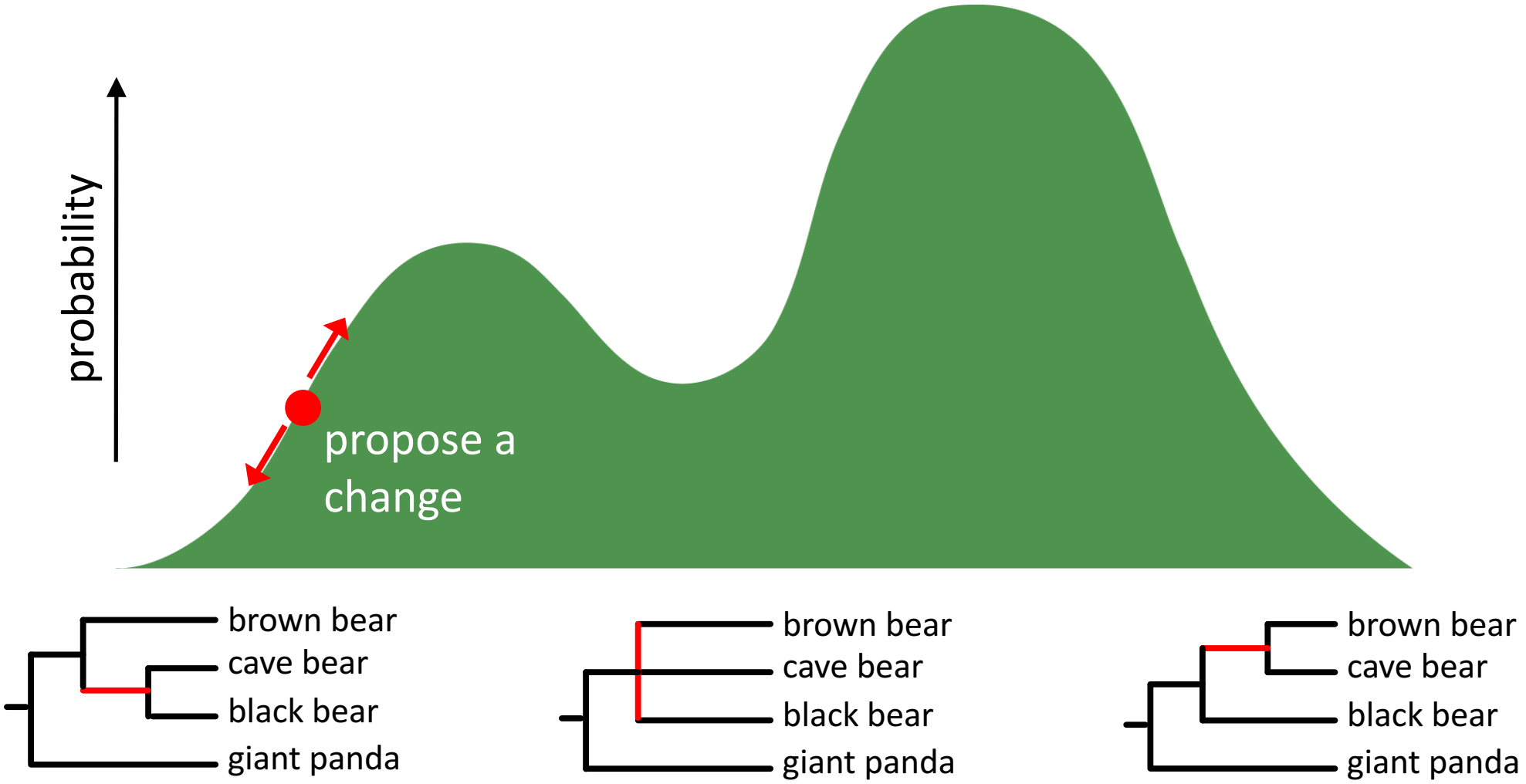


Nicholas Metropolis
Los Alamos, 1953

MCMC simulation

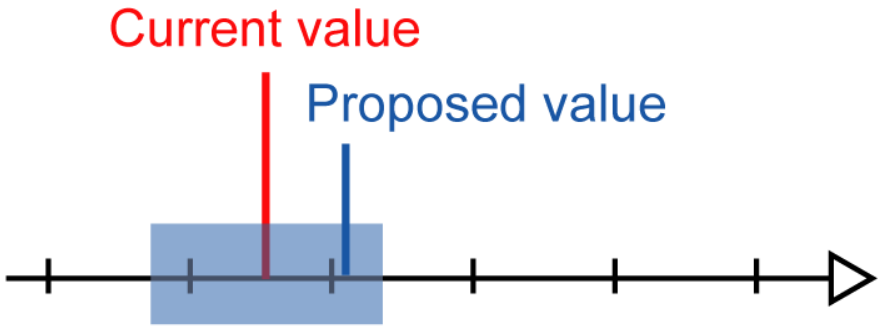


MCMC simulation



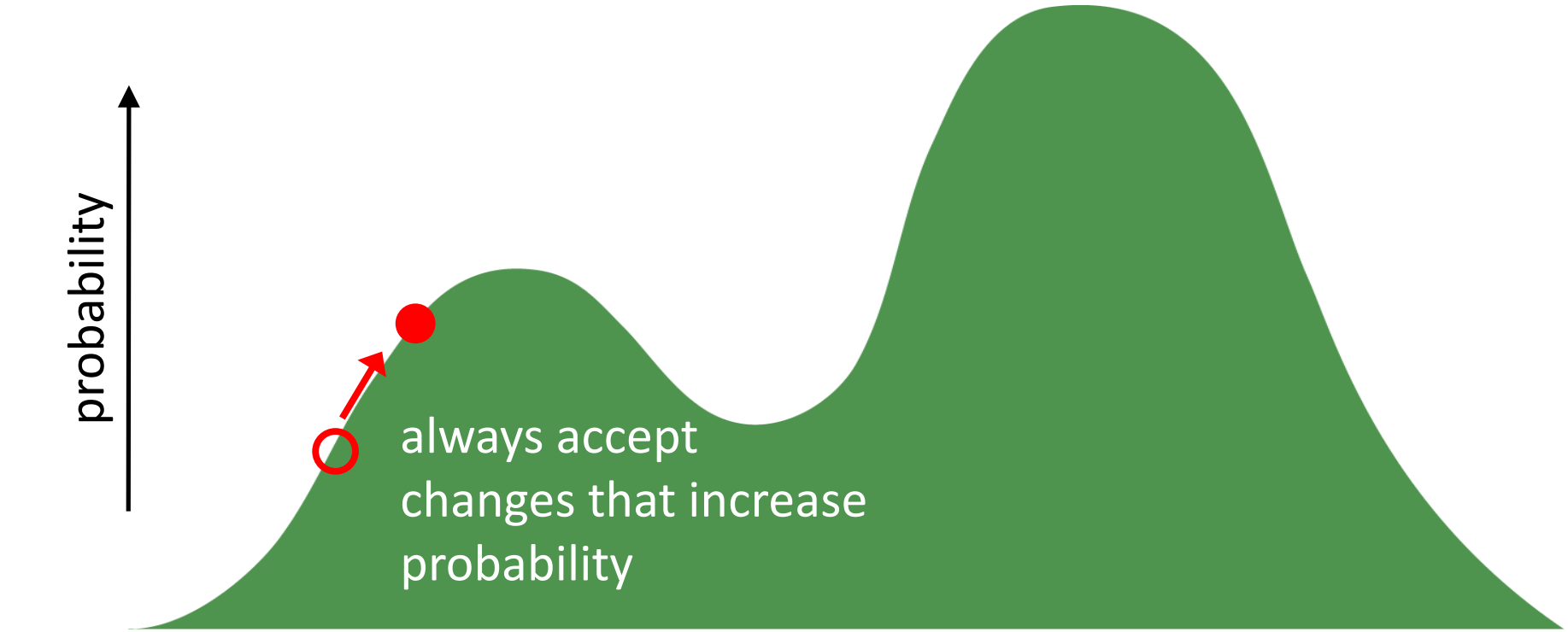
MCMC simulation

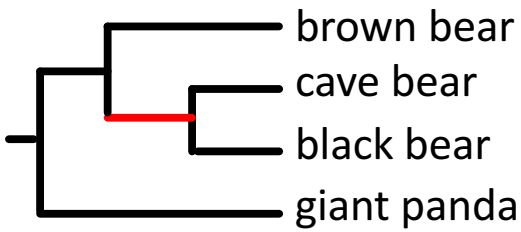
- Proposals for continuous parameters
  - Sliding window
  - Proportional scaling

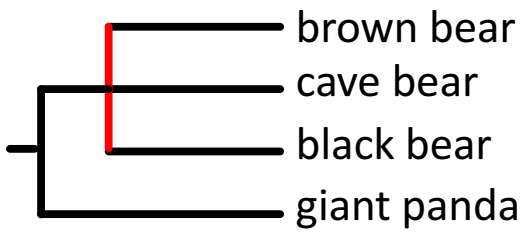


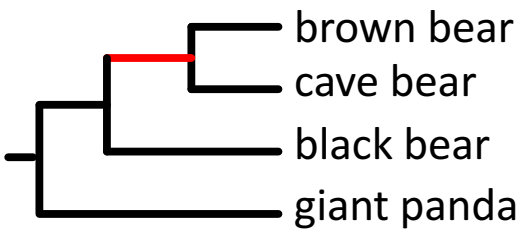
- Proposals for tree
  - Subtree slide
  - Narrow and wide exchange
  - Wilson-Balding rearrangement

MCMC simulation

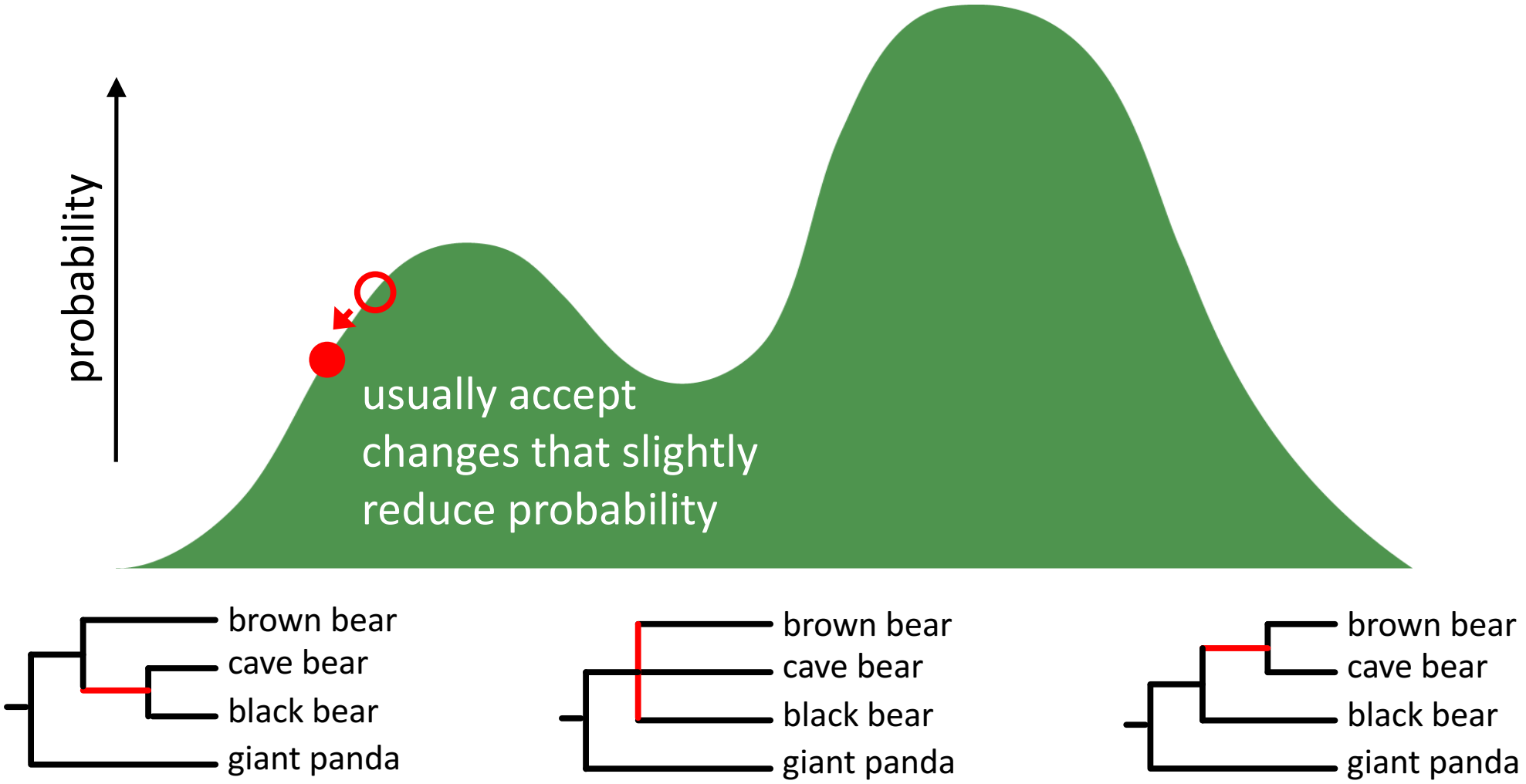




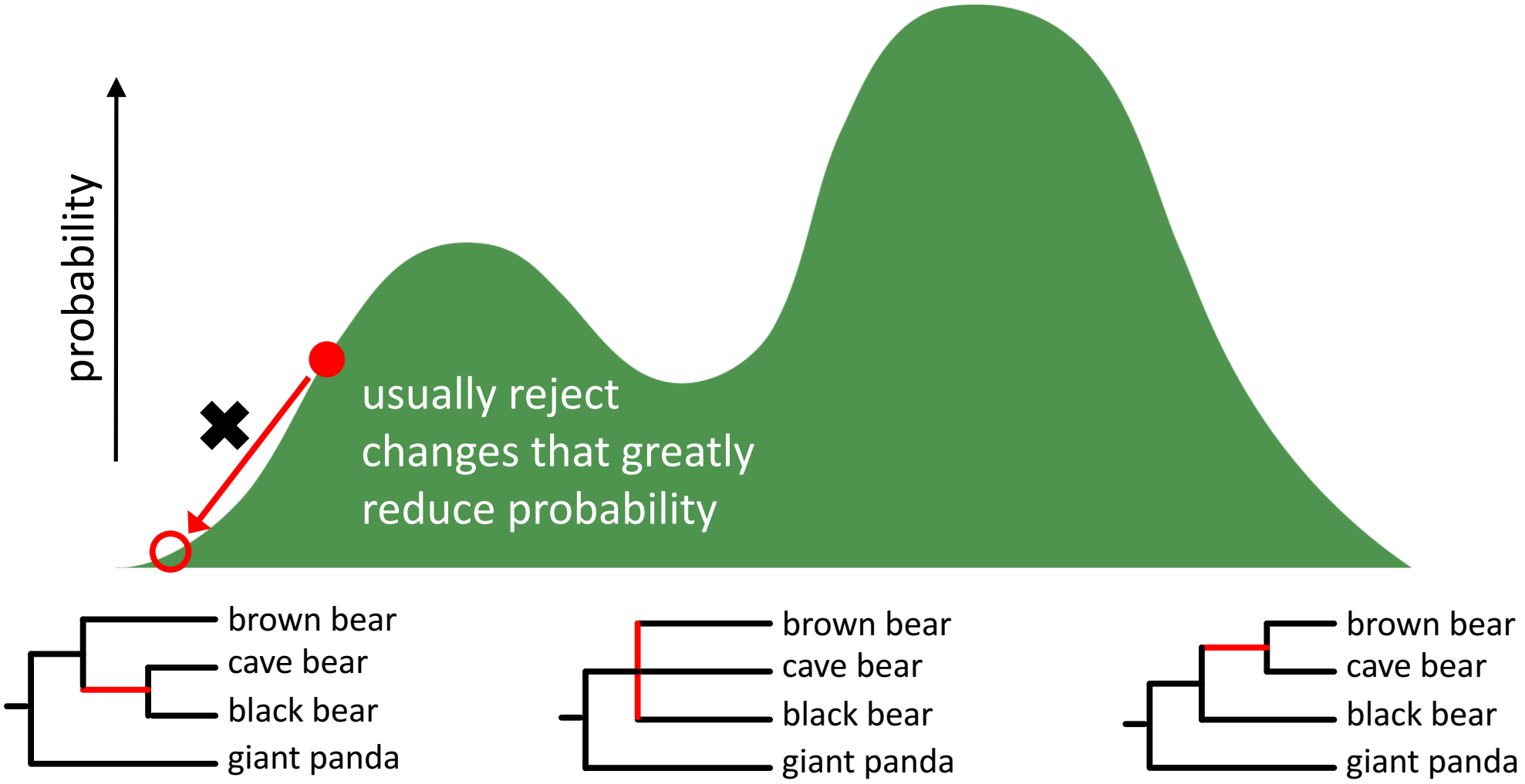




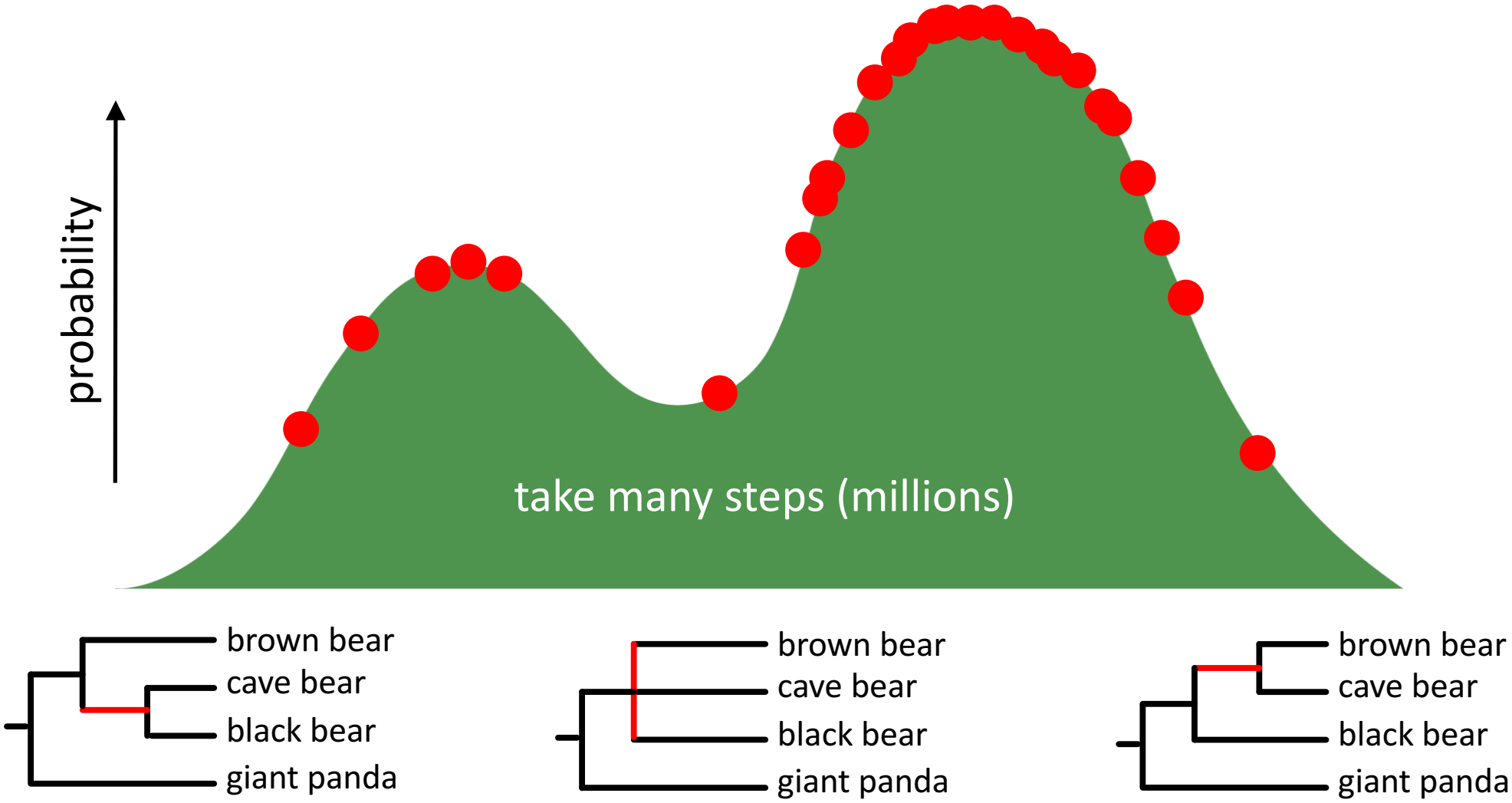
MCMC simulation



MCMC simulation



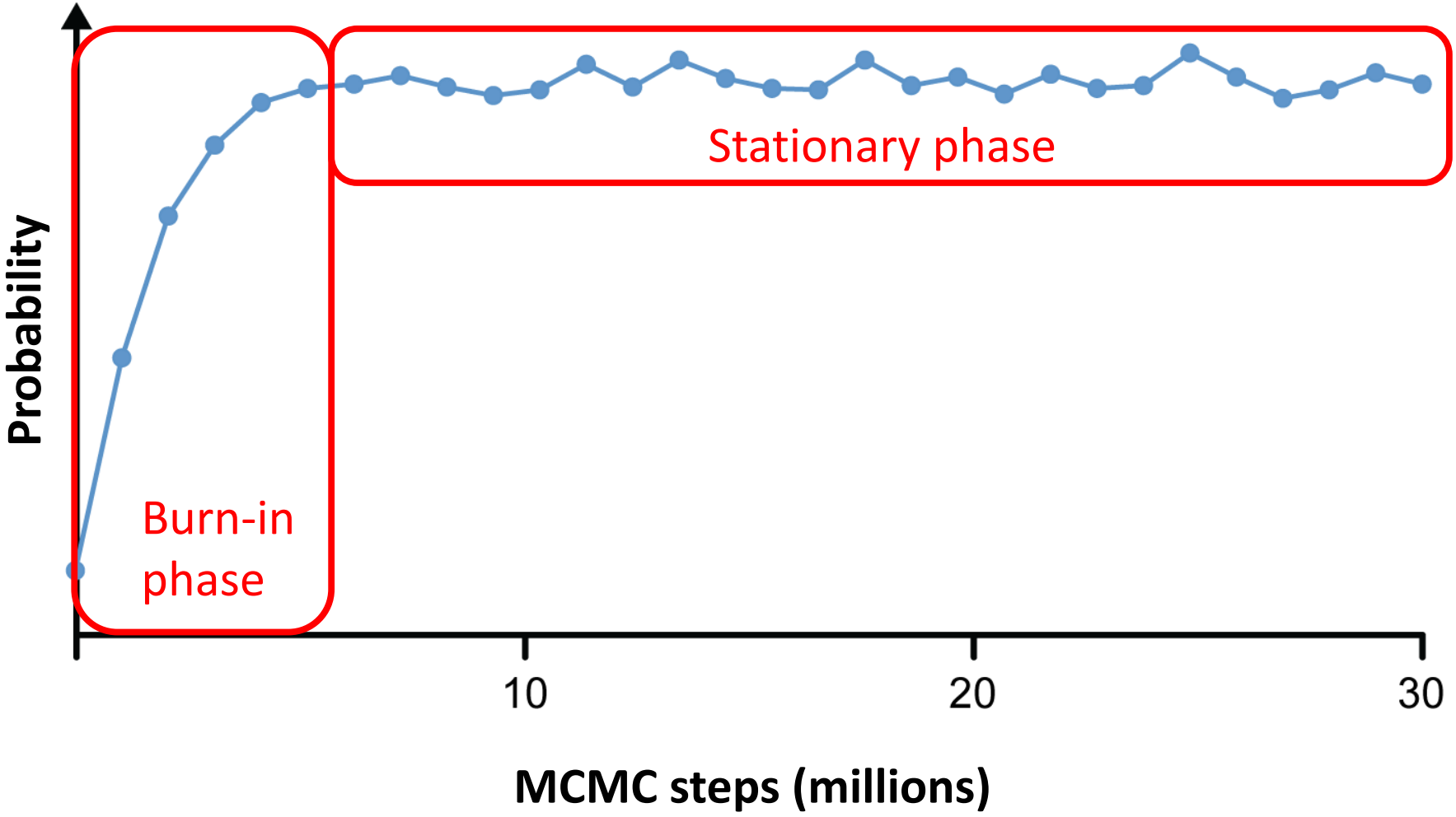
MCMC simulation



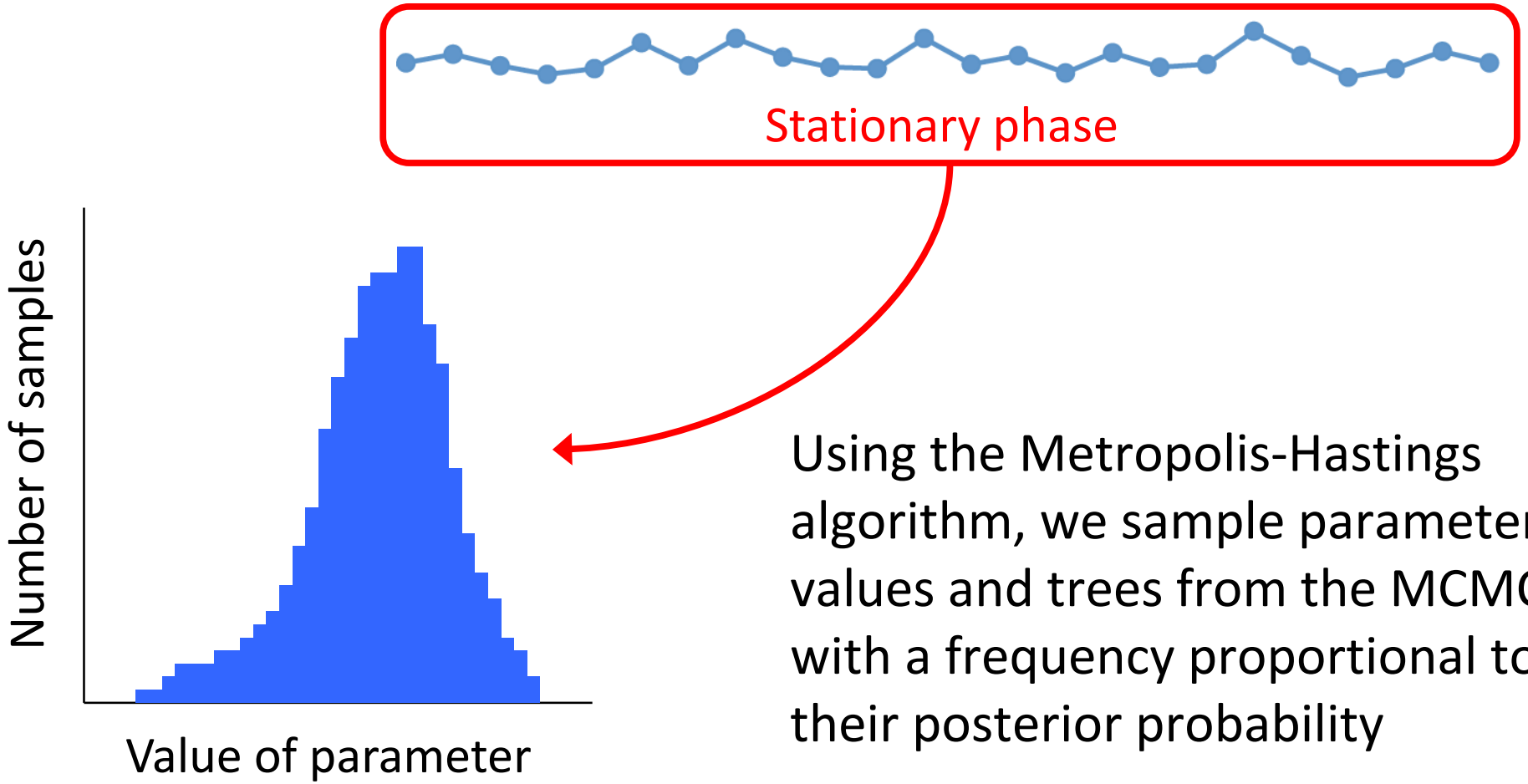
Samples from the MCMC

- Output from a Bayesian phylogenetic analysis:
  - A list of the parameter values visited by the Markov chain (.log)
  - A list of the trees visited by the Markov chain (.trees)

Samples from the MCMC

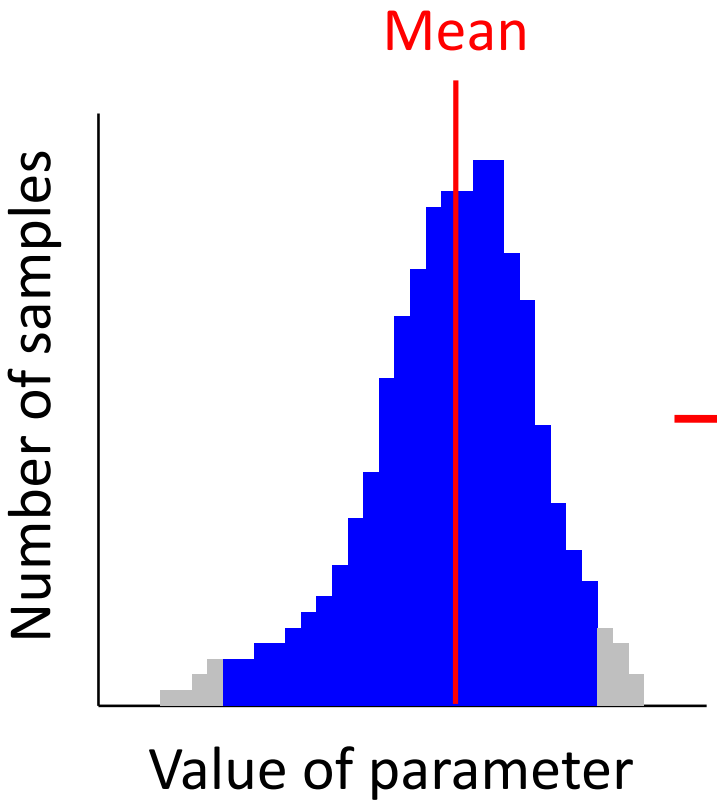


Samples from the MCMC



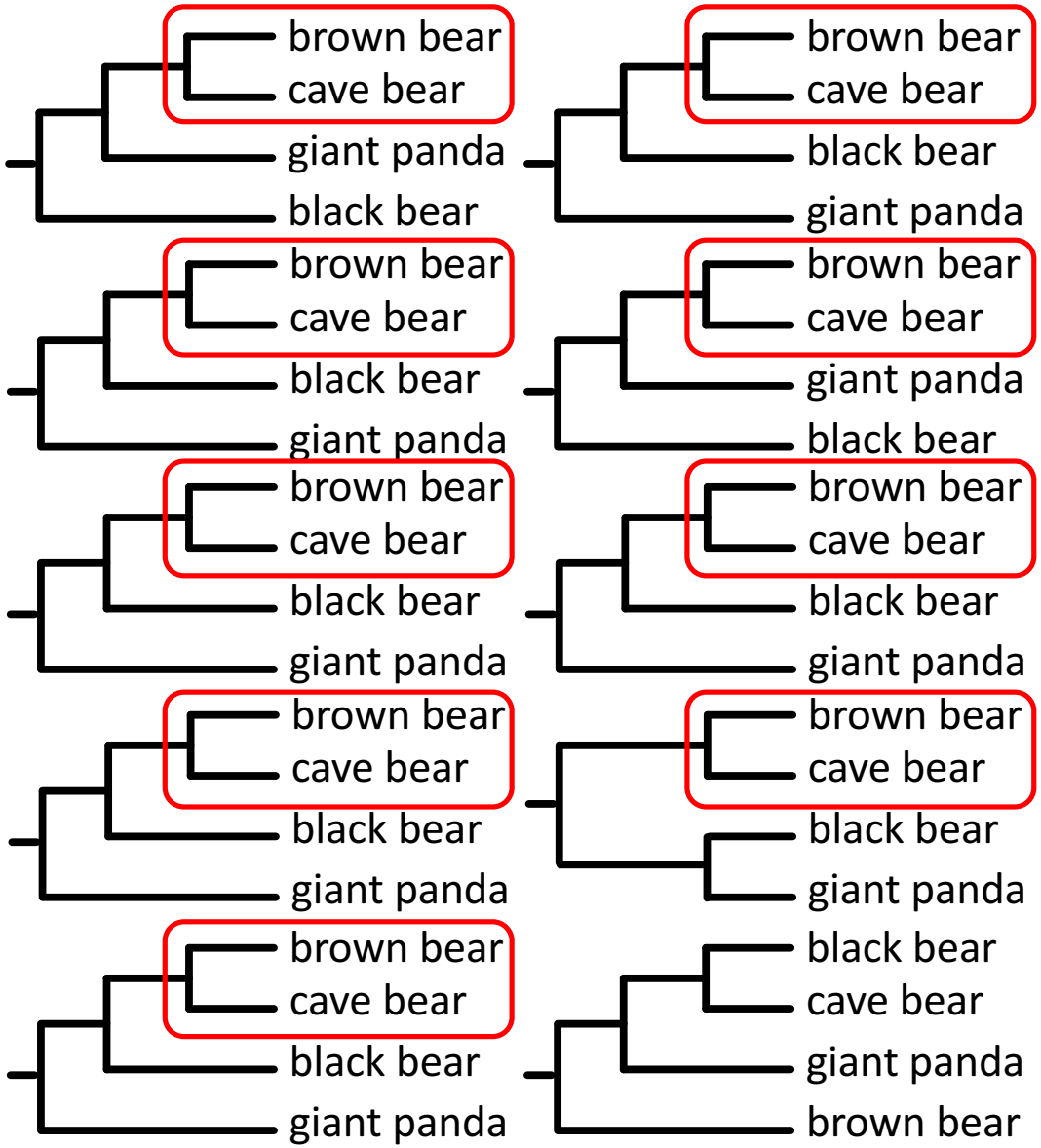
Using the Metropolis-Hastings algorithm, we sample parameter values and trees from the MCMC with a frequency proportional to their posterior probability

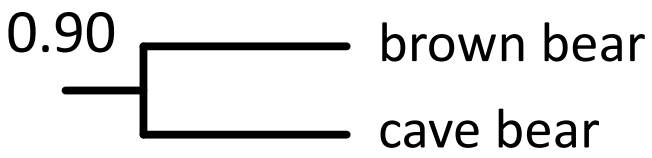
Samples from the MCMC



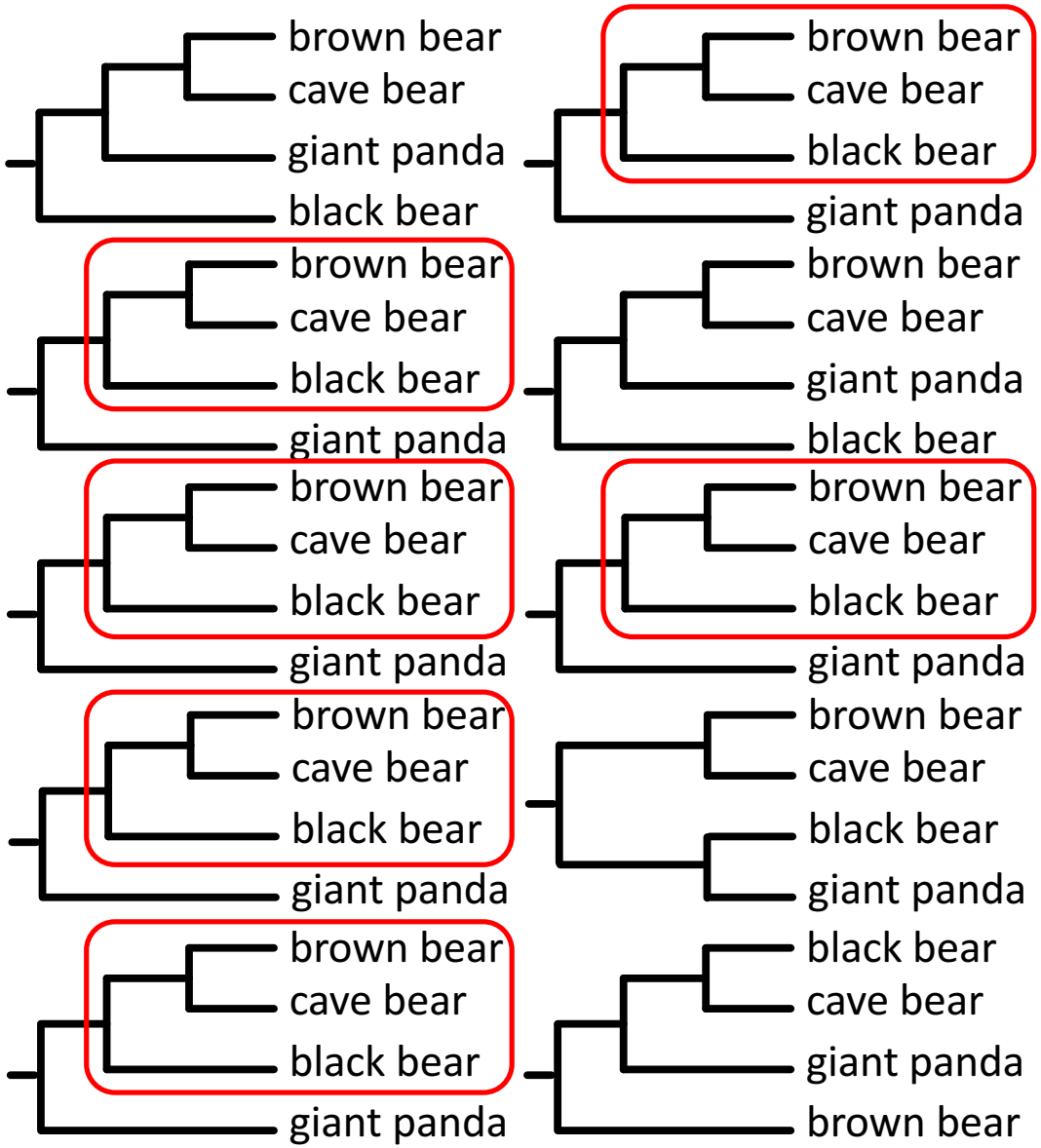
- Take the mean of the sampled values
  Mean posterior estimate
- Take the minimum interval spanning 95% of the sampled values
  95% highest posterior density interval

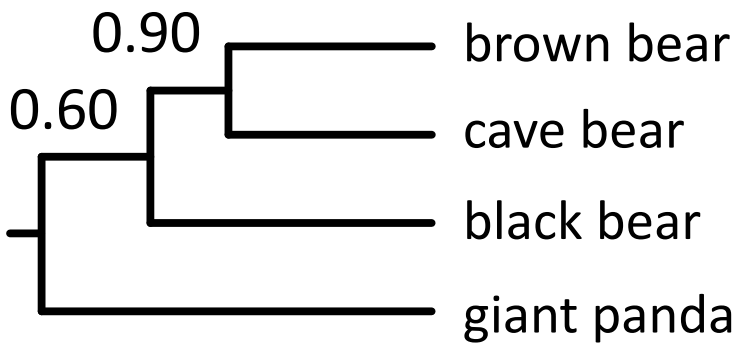
Samples from the MCMC





Samples from the MCMC





Samples from the MCMC

- **Majority-rule consensus tree**
  Shows all nodes with posterior probability >0.50

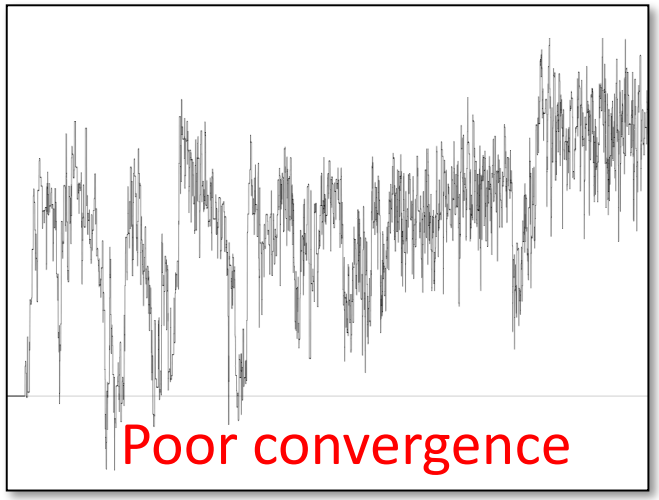
- **Maximum a posteriori (MAP) tree**
  Sampled tree with highest posterior probability

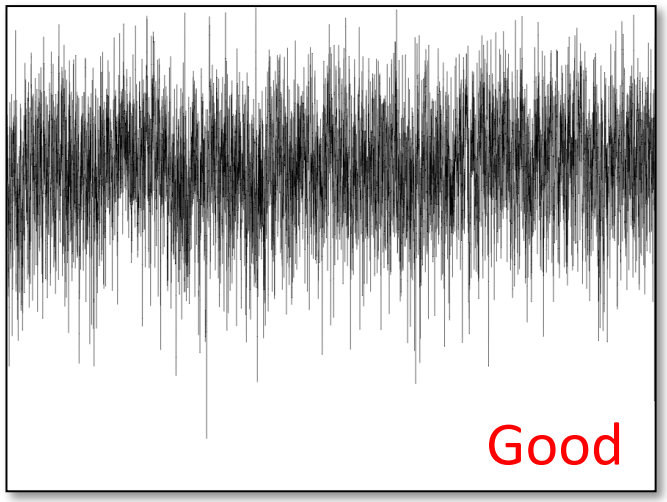
- **Maximum clade credibility (MCC) tree**
  Sampled tree with highest product of posterior node probabilities

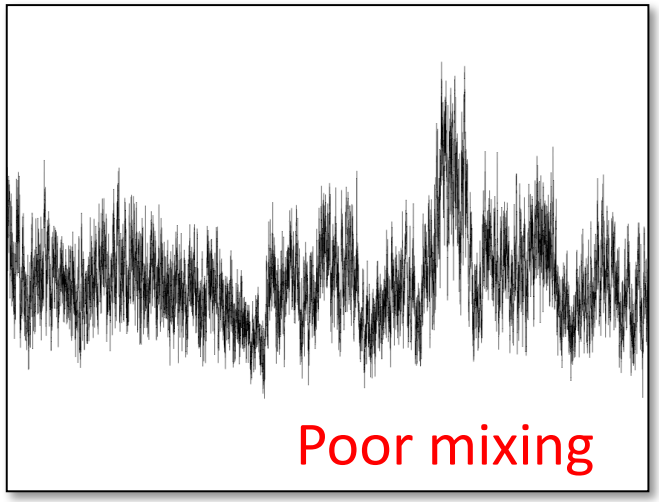
Diagnostics

1. **Convergence**
   Are we drawing samples from the stationary distribution?

2. **Sufficient sampling**
   Have we drawn enough samples to allow a reliable estimate of the posterior distribution?







8. Software

BEAST 1

- **B**ayesian **E**volutionary **A**nalysis by **S**ampling **T**rees
- Analyse population- or species-level data
- Simultaneous estimation of tree and node times
- Range of clock models
- Range of tree priors and demographic models

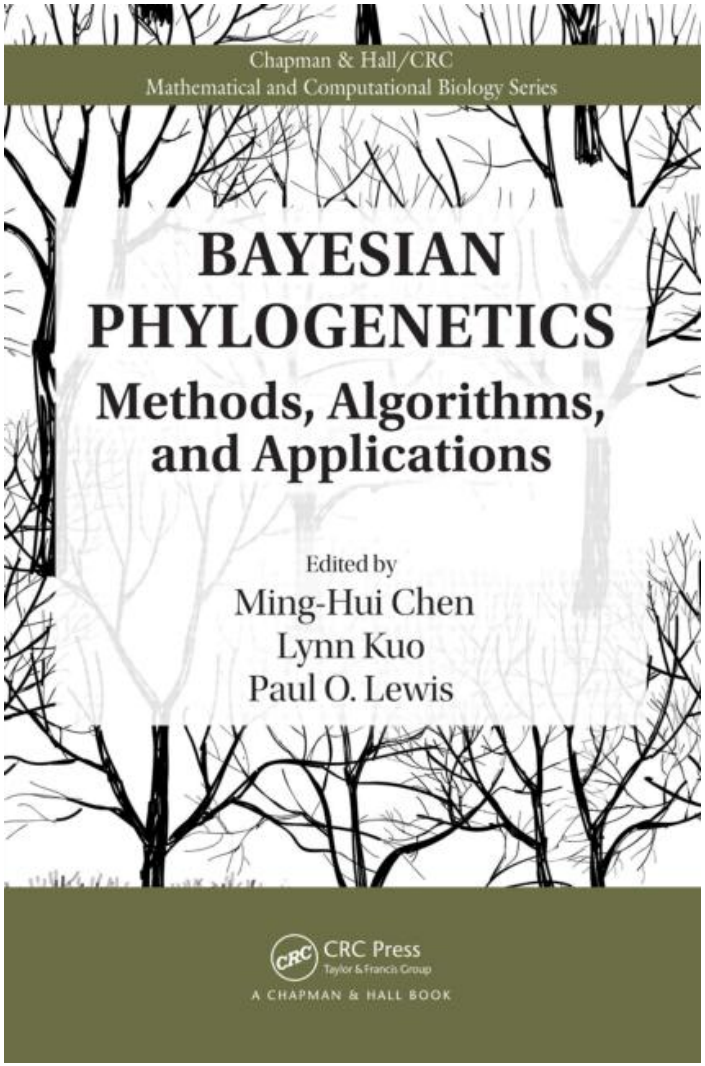
Beast2

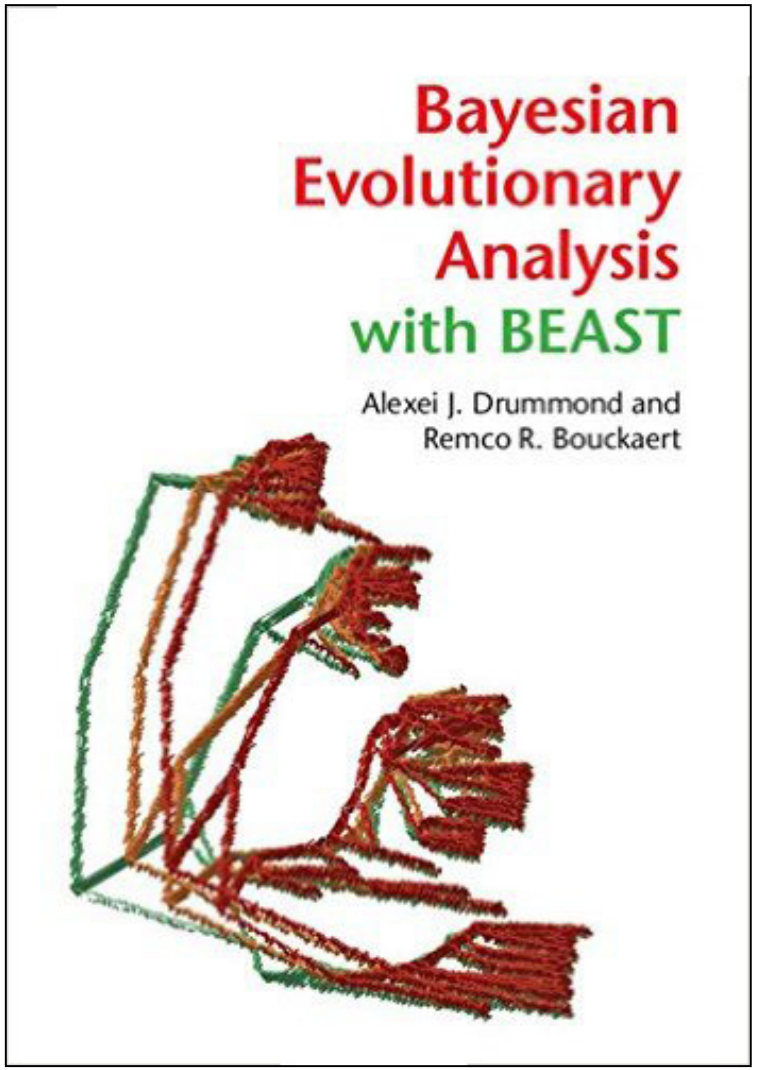
Bayesian evolutionary analysis by sampling trees

- Re-write of *BEAST* to increase modularity
- Users can extend *BEAST* by adding packages
- Lacks some of the clock models and demographic models from *BEAST* 1
- Additional tree priors not available in *BEAST* 1
- Capacity to perform simulations

For a comparison of *BEAST* 1 and 2:
http://beast2.org/beast-features/

Useful references





Outline of Part B

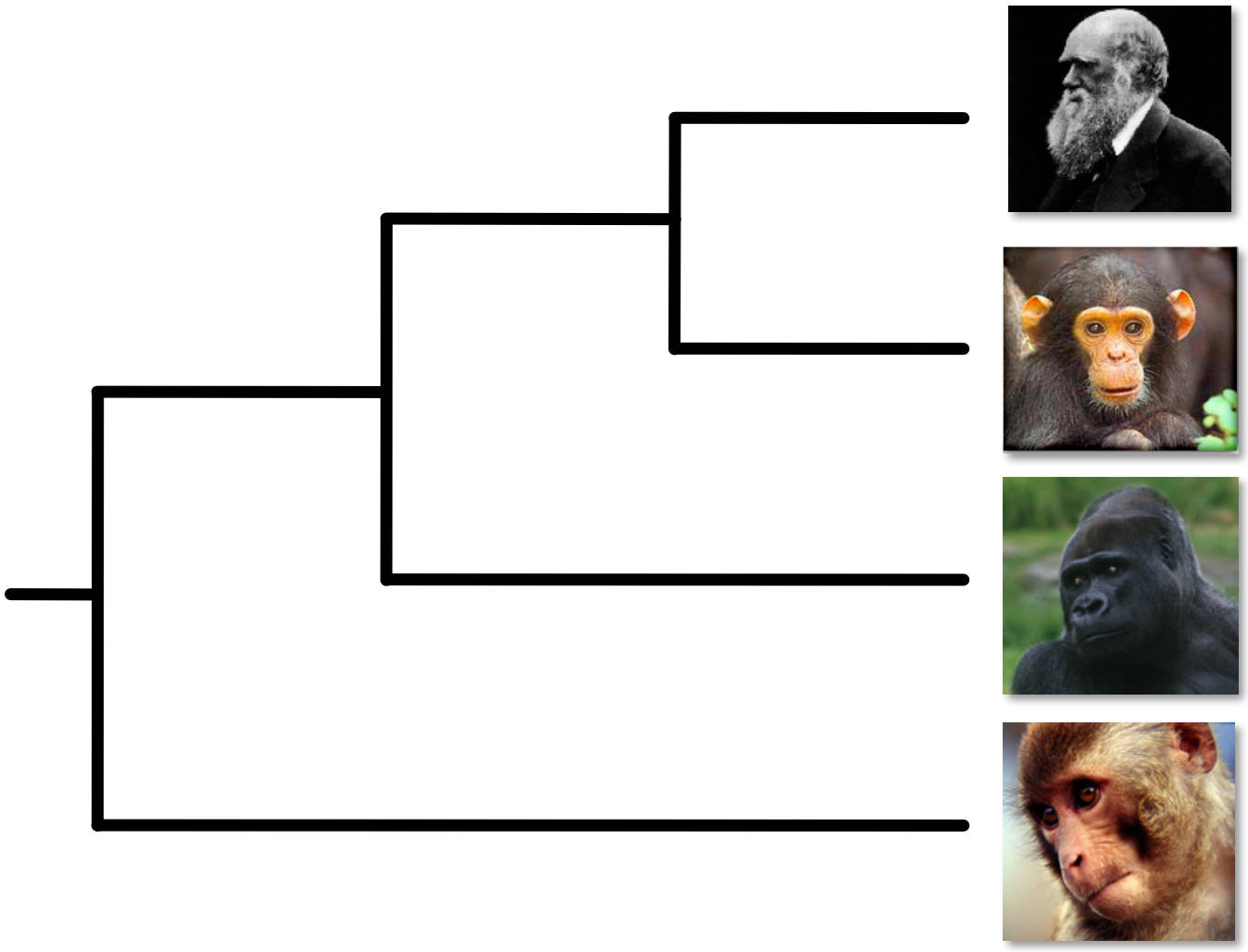
1. Molecular dating
2. Models of rate variation
3. Calibrating the molecular clock

1. Molecular Dating

Evolutionary timescales

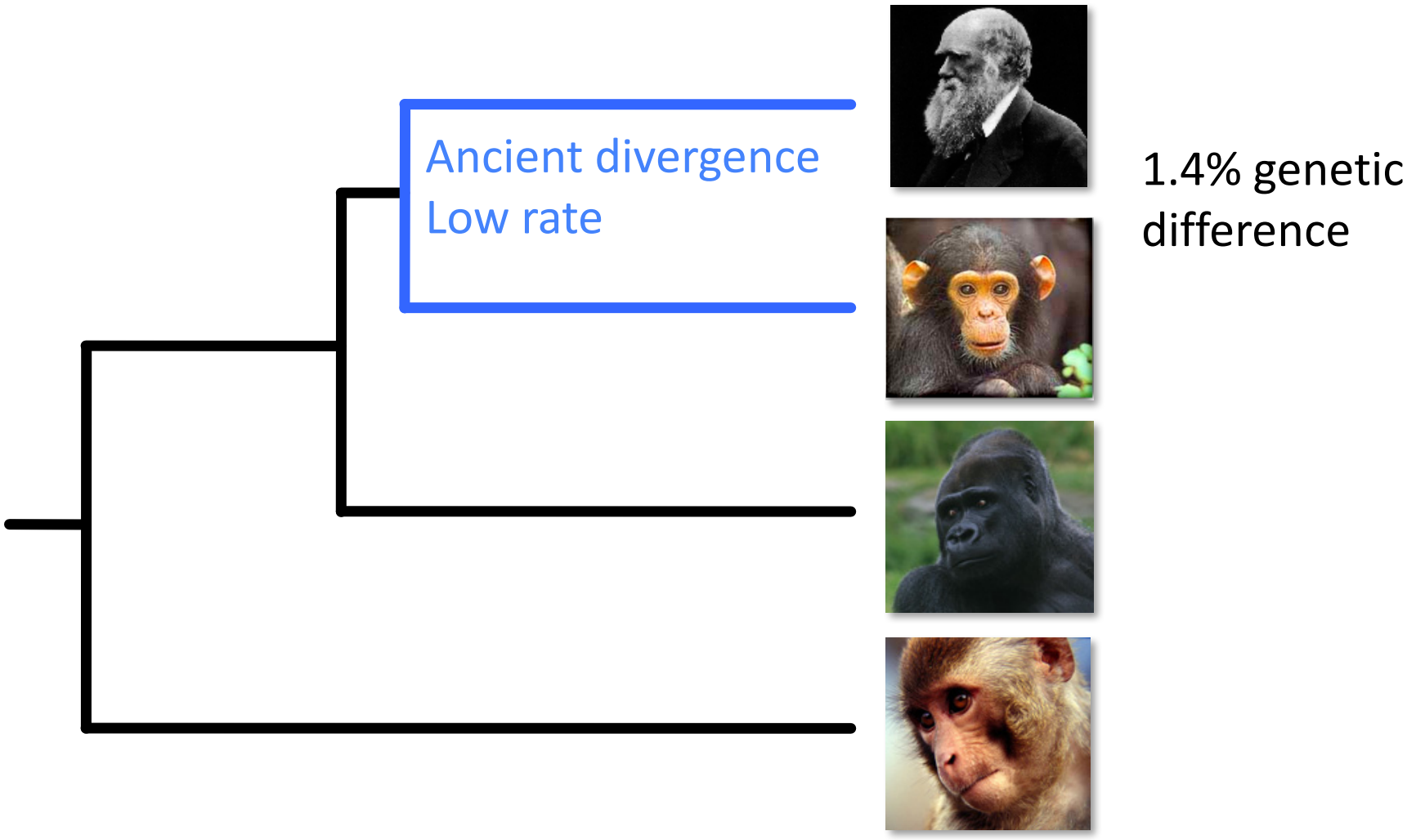
dos Reis *et al*. (2012) *Proc Roy Soc B*

The molecular clock

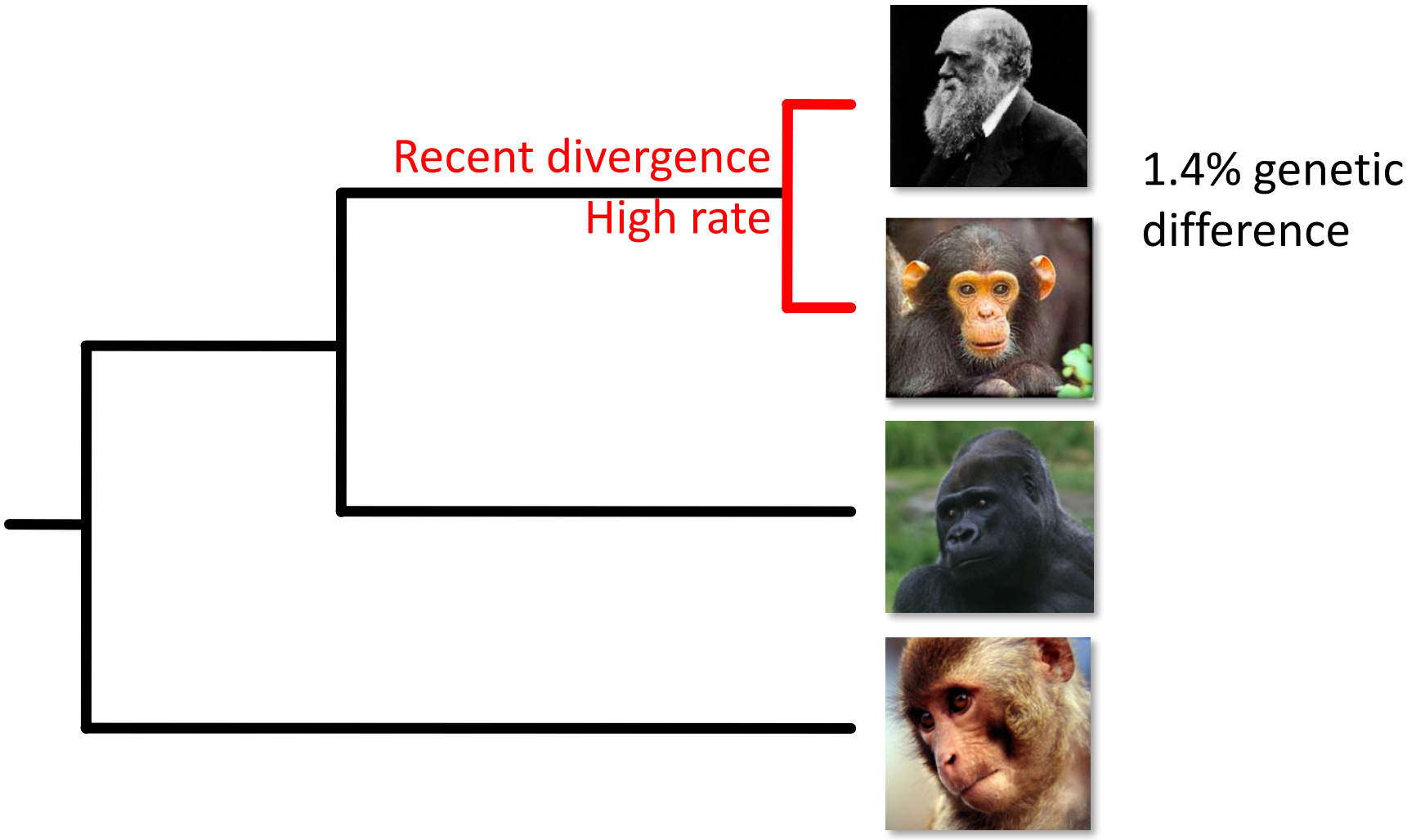


1.4% genetic difference

Calibrating the molecular clock



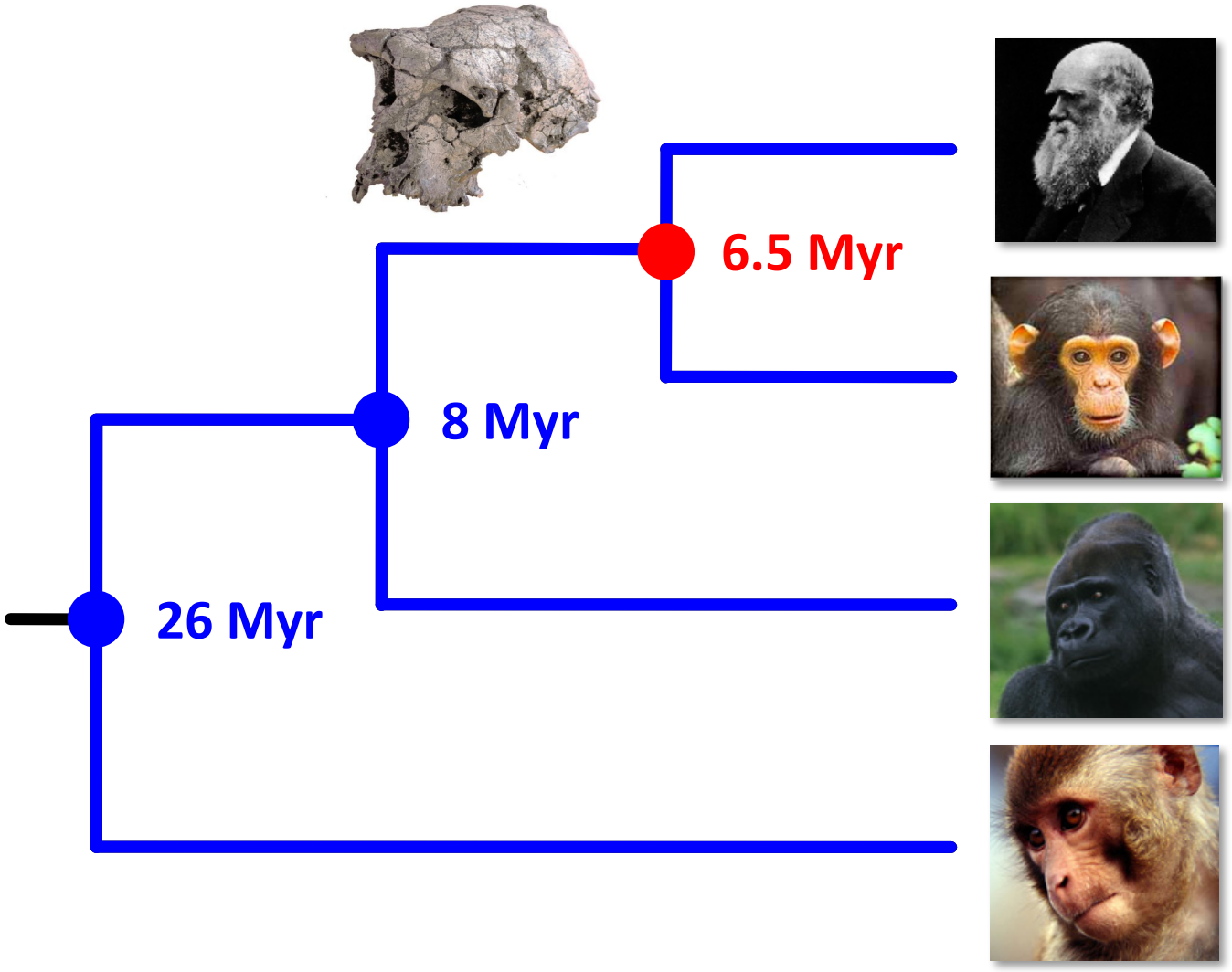
Calibrating the molecular clock



Calibrating the molecular clock

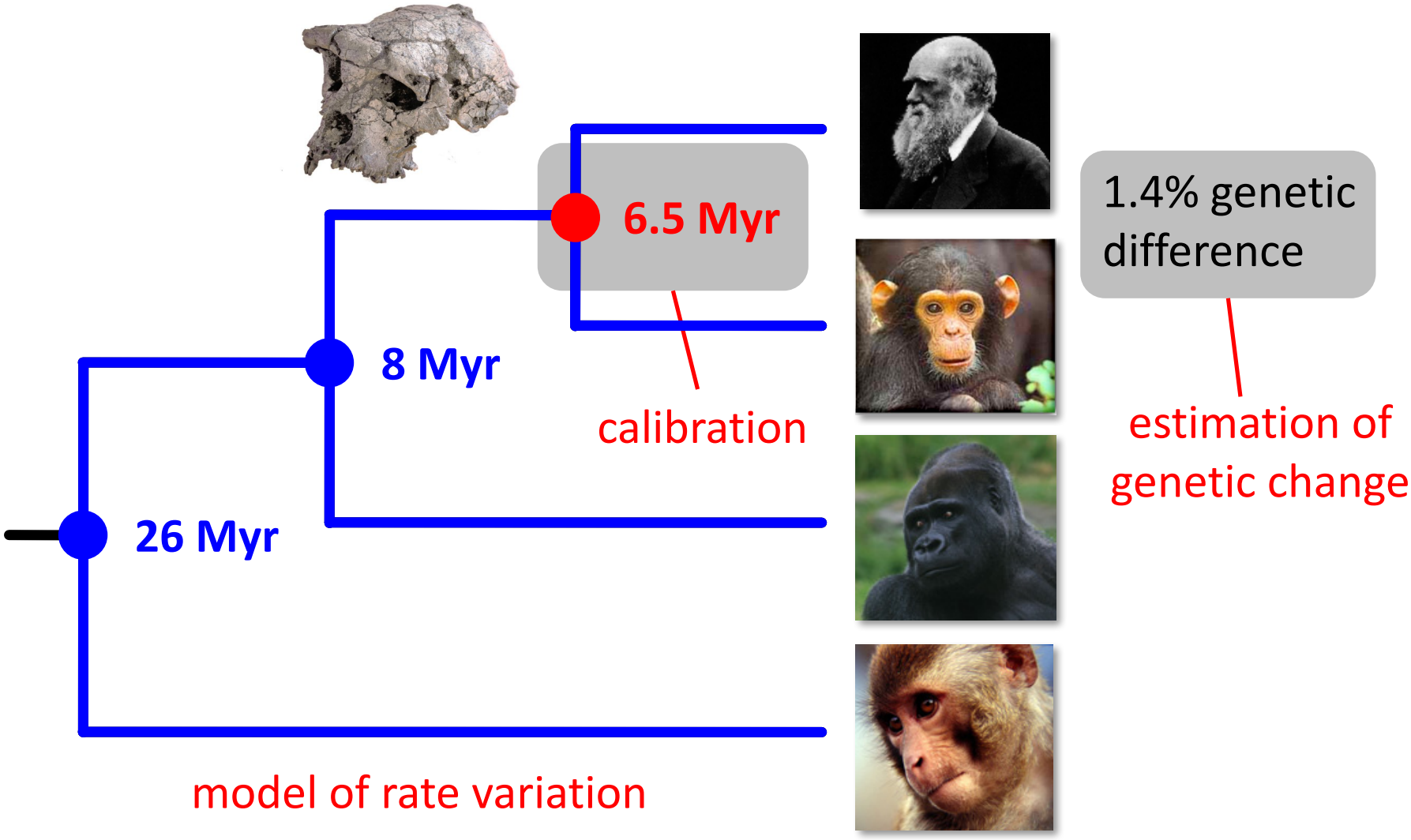
- Rates and times are **non-identifiable**
- Likelihood only depends on their product
  - Branch lengths in substitutions per site
- To separate rate and time, we need (prior) information about one or the other

The molecular clock



1.4% genetic difference

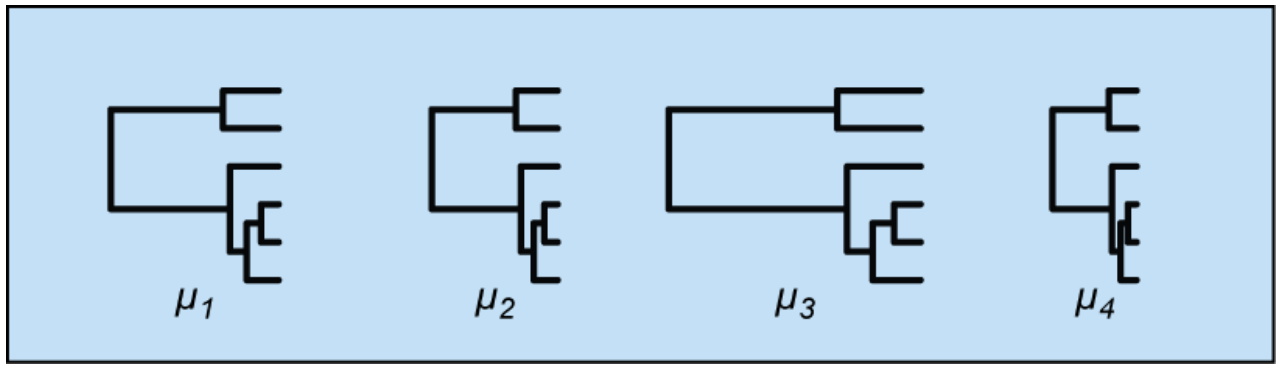
The molecular clock



2. Models of Rate Variation

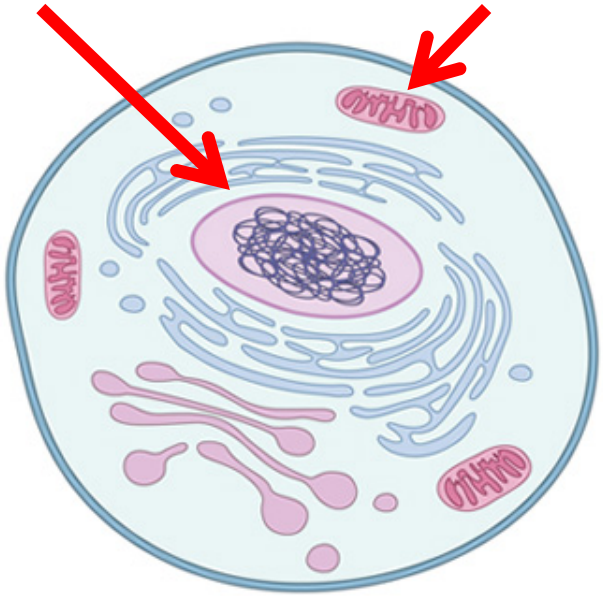
Rate variation across loci

- Across loci (**gene effects**)



Slowly evolving

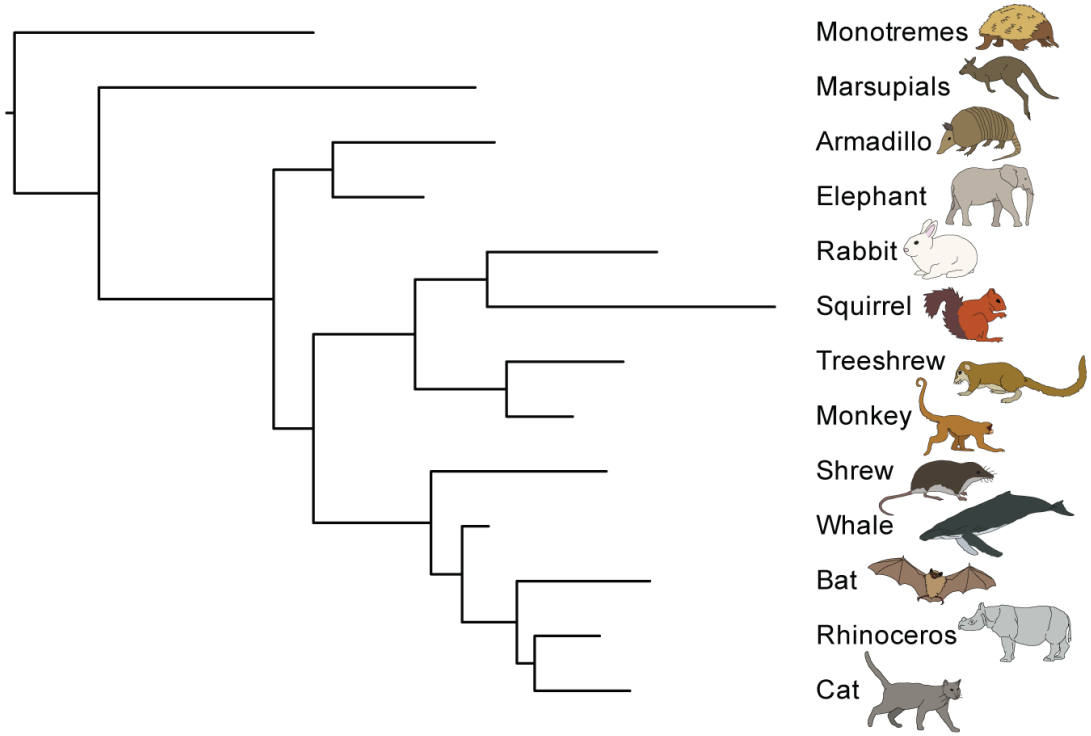
Rapidly evolving



Ho & Duchêne (2014) *Mol Ecol*

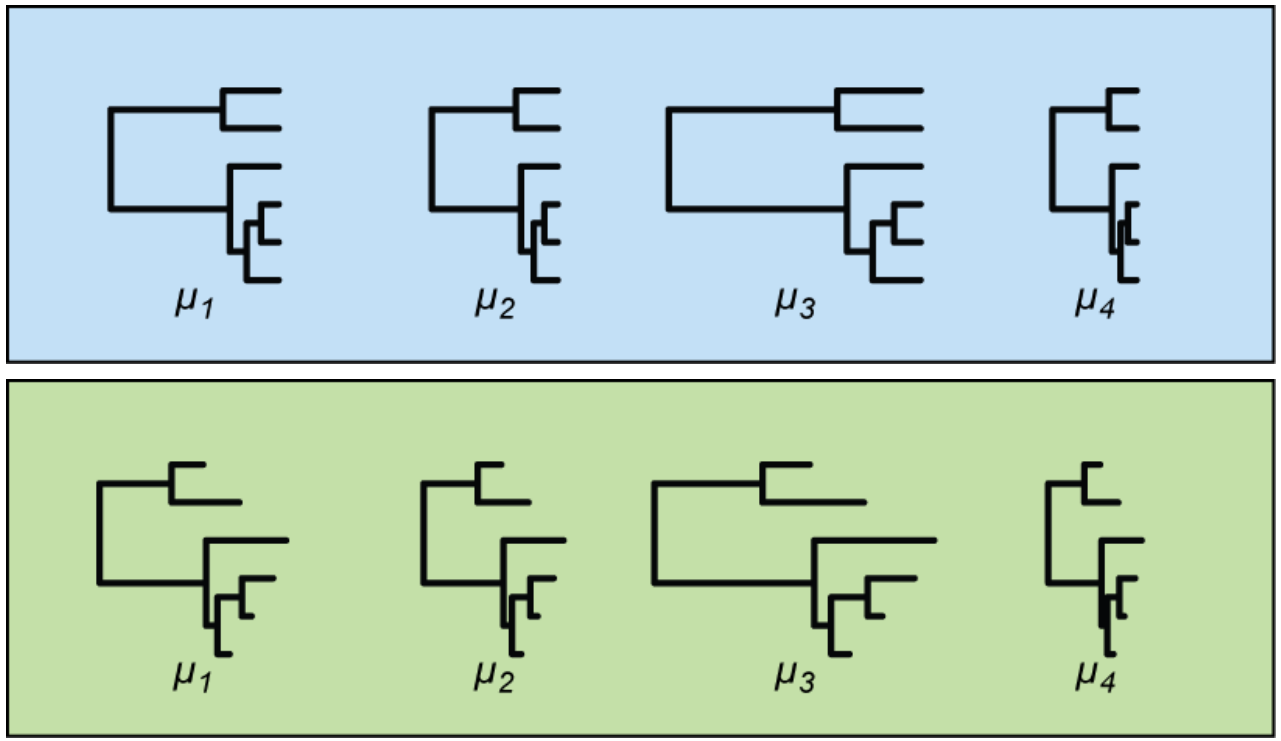
Rate variation across lineages

- Rates vary among lineages because of differences in:
  - Exposure to mutagens
  - Metabolic rate
  - Generation time
  - Population size
  - Strength and direction of selection



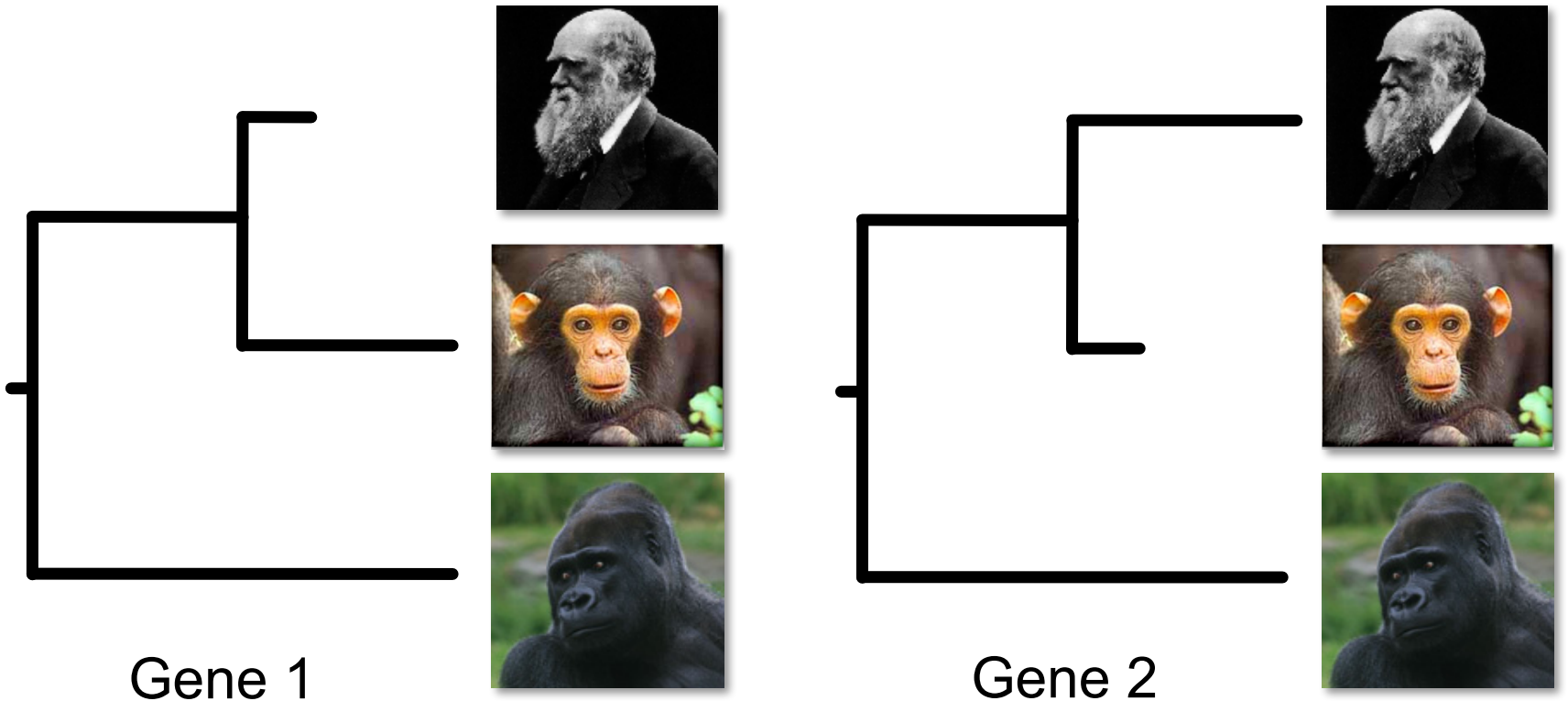
Rate variation across lineages

- Across loci (**gene effects**)
- Across lineages (**lineage effects**)



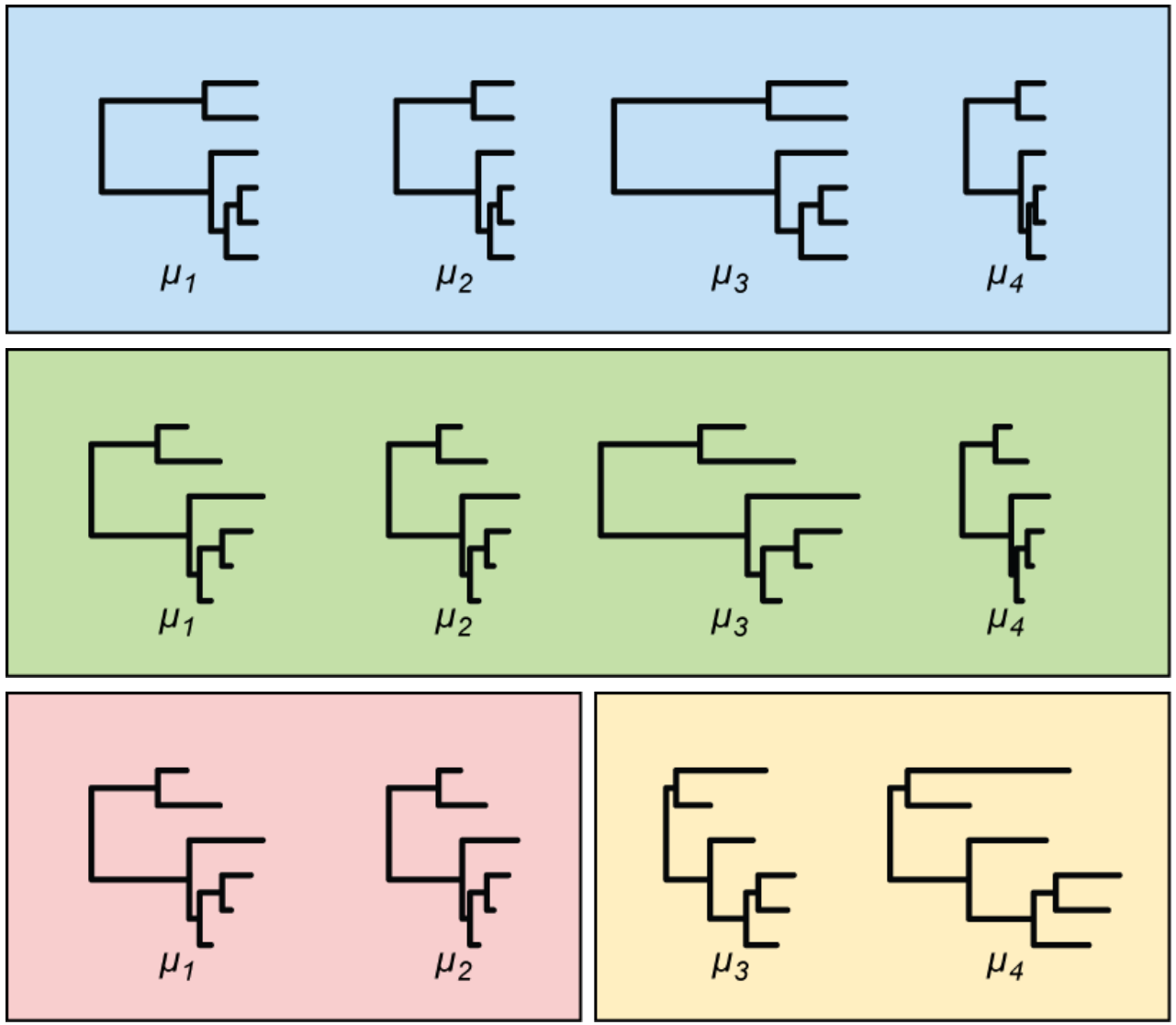
Ho & Duchêne (2014) *Mol Ecol*

Gene-by-lineage interactions



Evolutionary rate variation

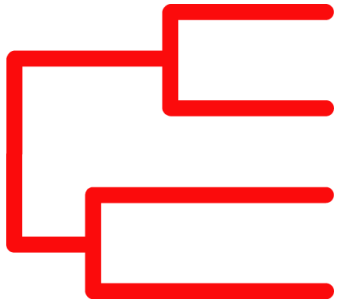
- Across loci (**gene effects**)
- Across lineages (**lineage effects**)
- Gene-by-lineage interaction (**residual effects**)



Ho & Duchêne (2014) *Mol Ecol*

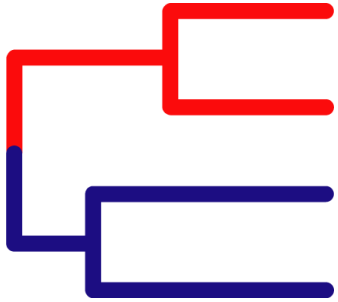
Molecular-clock models

Strict or ‘global’ molecular clock



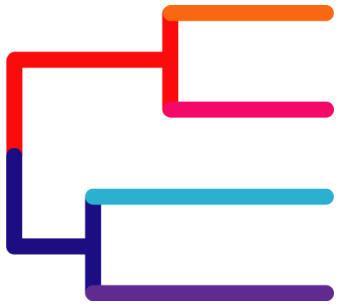
$k = 1$

Local clocks



$1 < k < n$

Relaxed clocks



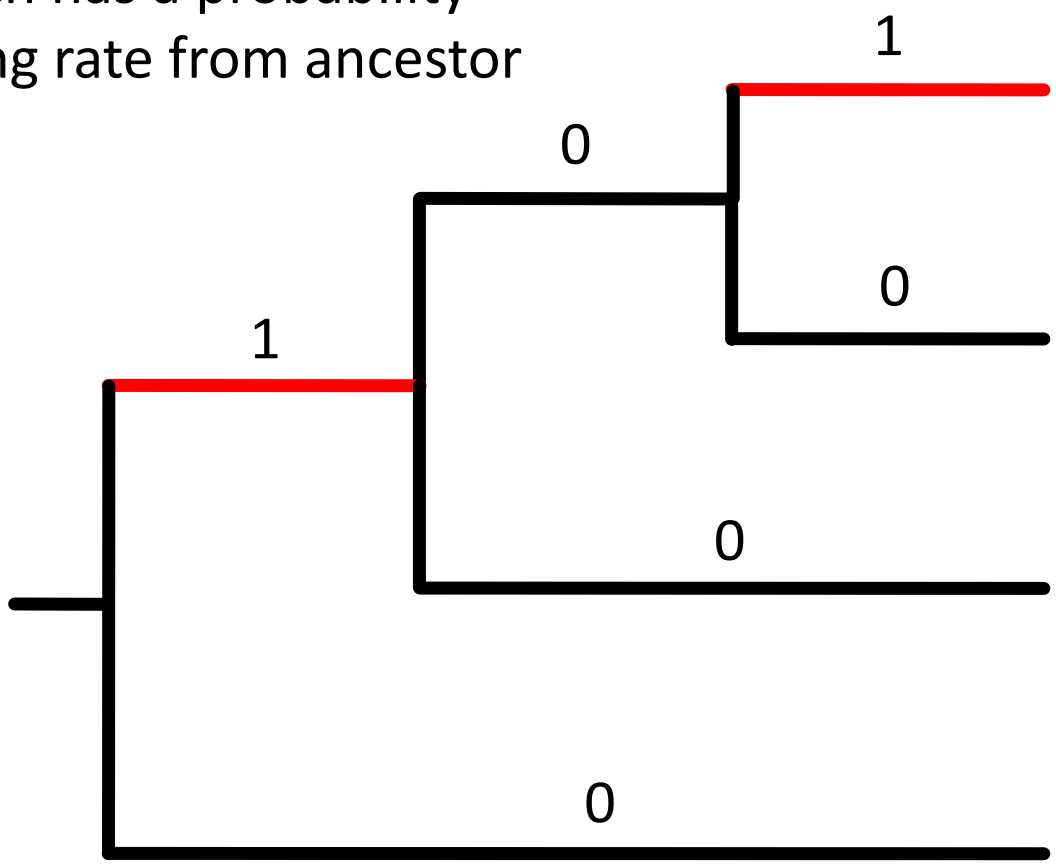
$k = n$

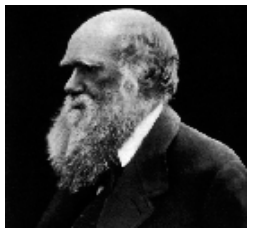
Local clocks

- Small number of rates
  - More than 1 rate (*i.e.*, not a strict clock)
  - Fewer than number of branches (*i.e.*, not a relaxed clock)
- **Local clock**
  - Same rate shared by neighbouring branches
- **Discrete clock**
  - Small number of branch rates, distributed across tree

Random local clock

Each branch has a probability
of inheriting rate from ancestor

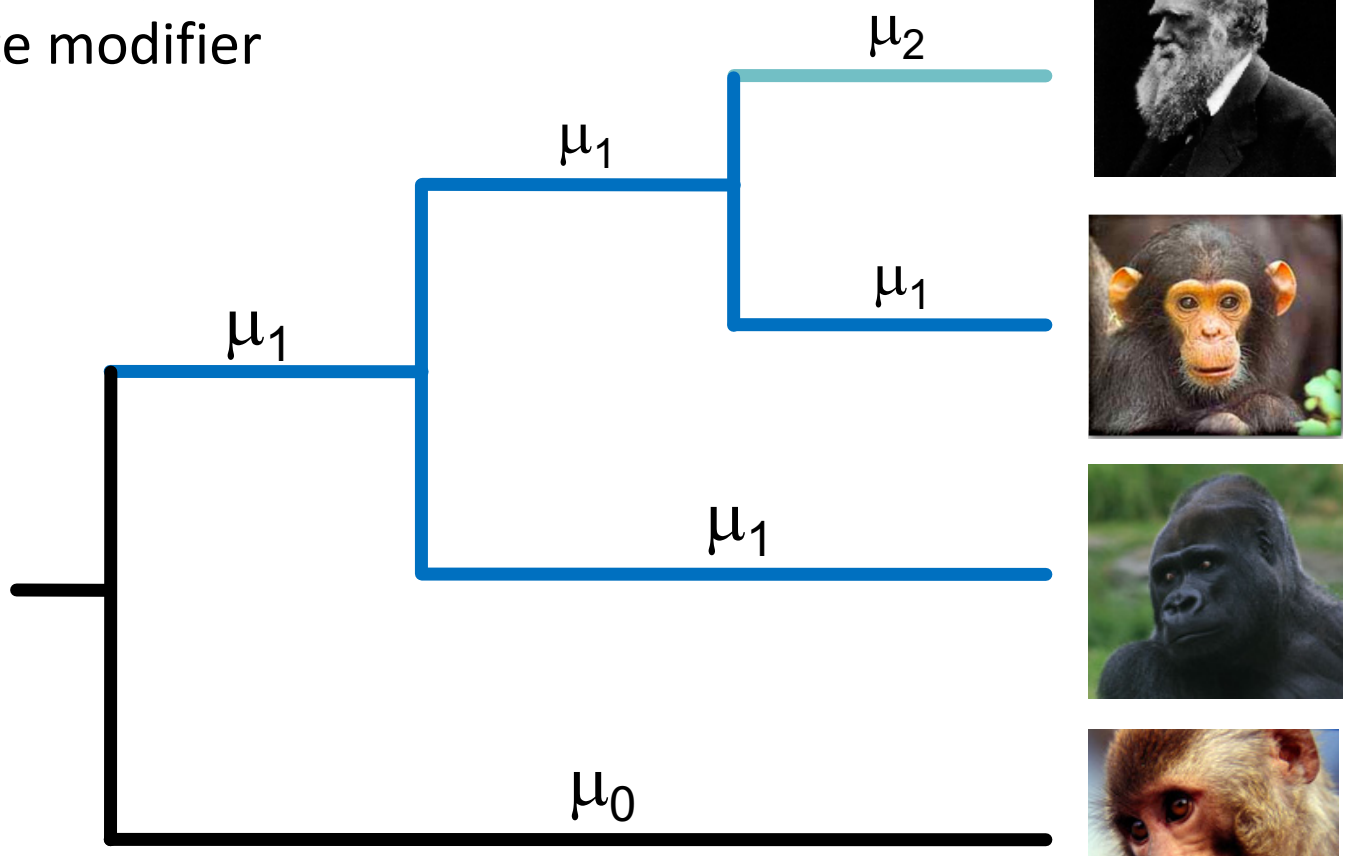




Includes possibility of zero rate changes (= strict clock)

Random local clock

Otherwise multiplied by a relative rate modifier



Note that tree is jointly estimated

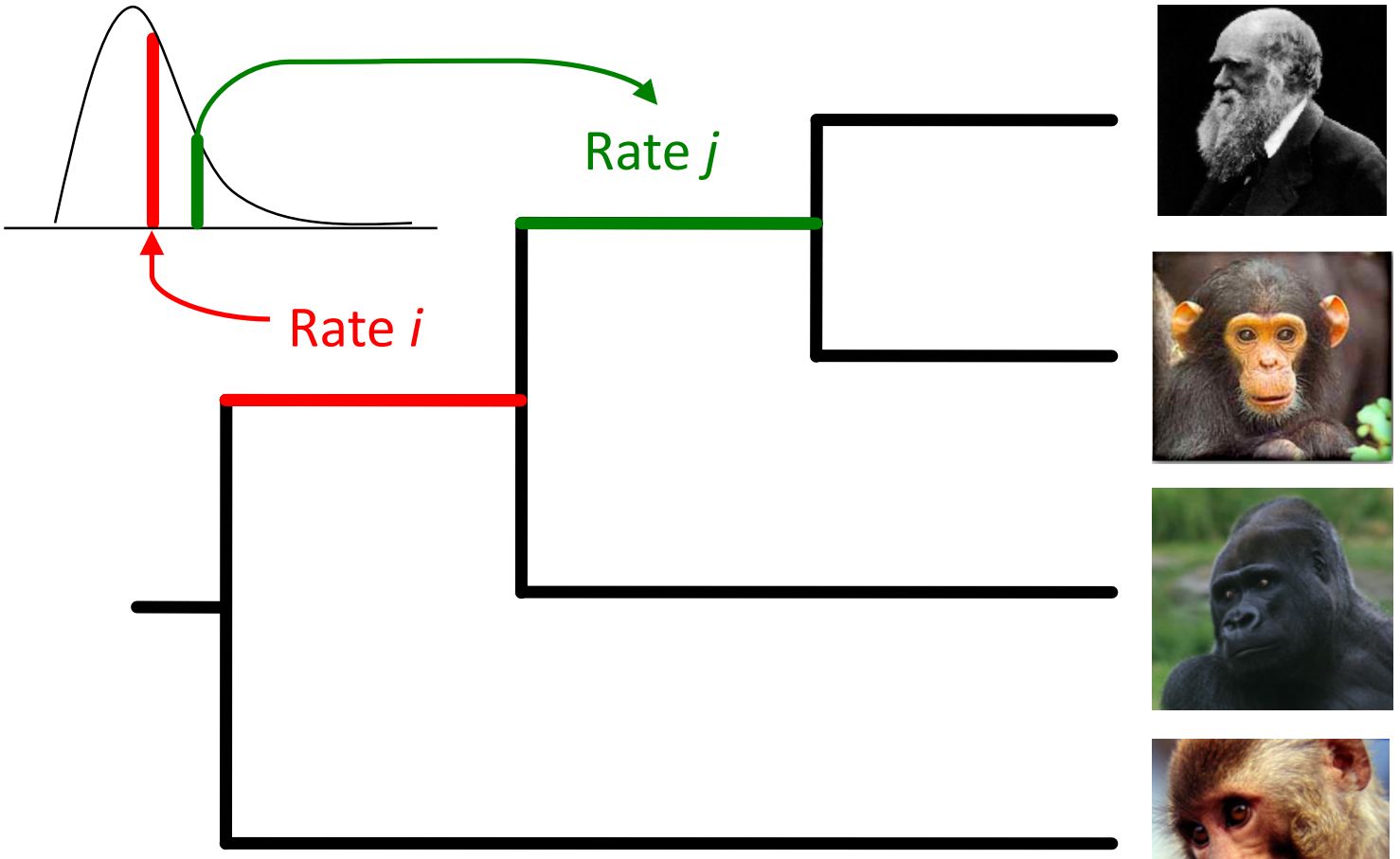
Bayesian relaxed clocks

- Allow a different rate in each branch
- Rates can be autocorrelated or uncorrelated
  - **Autocorrelated**
    rates in neighbouring branches are related
  - **Uncorrelated**
    rates identically and independently distributed among branches

Relaxed clocks

- We know that life-history characteristics:
  - Have effects on rates of molecular evolution
  - Are usually heritable to some degree
- Treat molecular rate as a heritable trait
- Relaxed clocks generally assume that closely related species share similar rates

Bayesian relaxed clocks



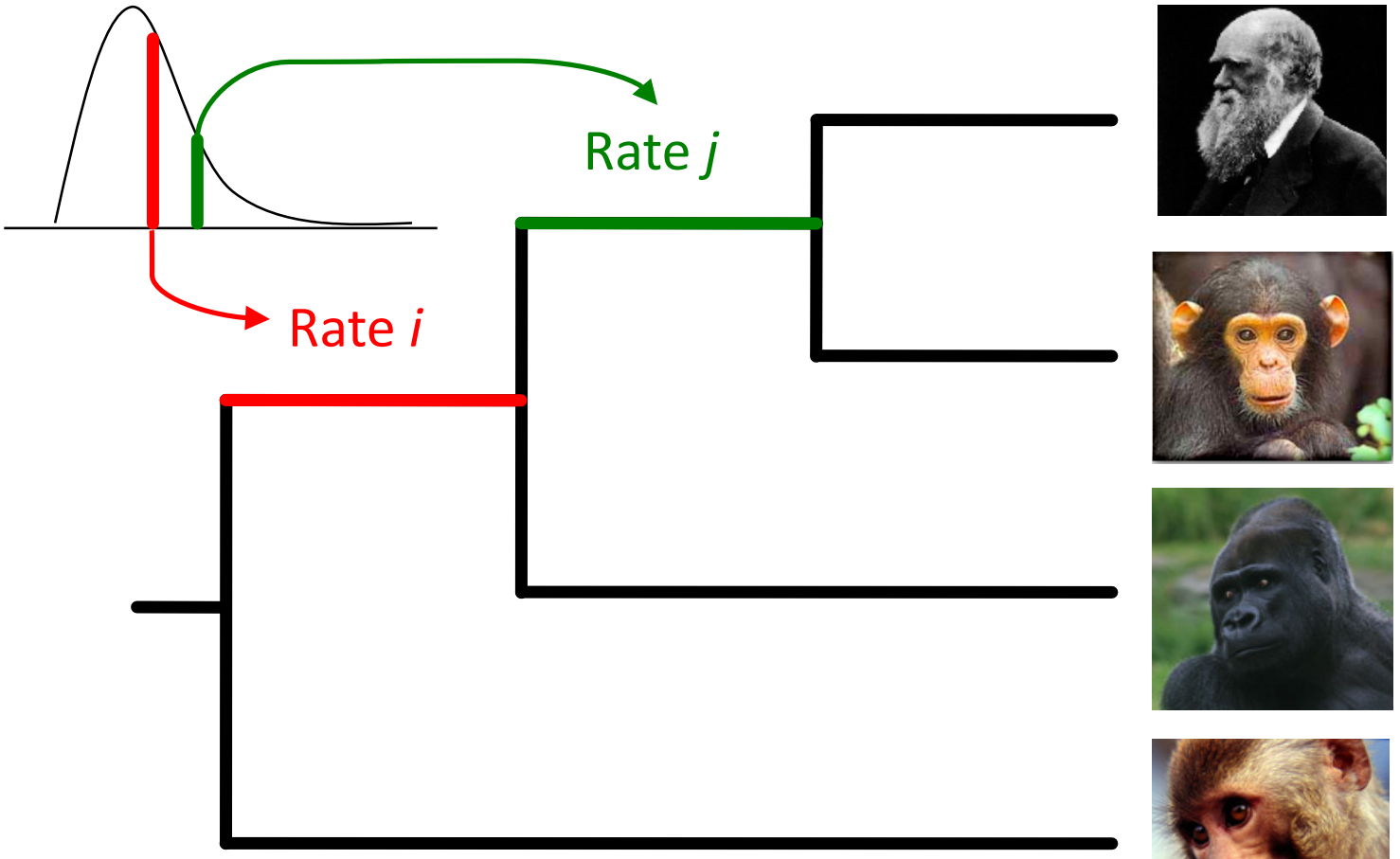
Autocorrelated lognormal relaxed clock

Autocorrelated relaxed clocks

- Continuous
  - r_j ~ Lognormal(μ_i, vt_i)
  - r_j ~ Gamma(α, λ)
  - Cox-Ingersoll-Ross
  - Ornstein-Uhlenbeck
- Episodic
  - r_j ~ Exponential(1/r_i)

Variance of distribution depends on time duration of branch

Bayesian relaxed clocks

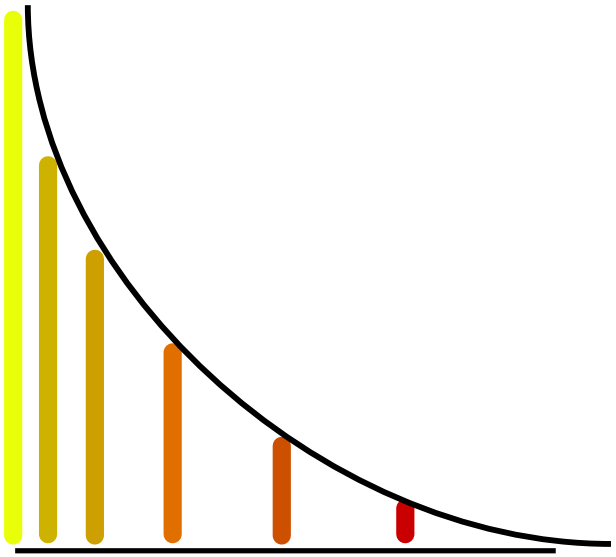


Uncorrelated lognormal relaxed clock

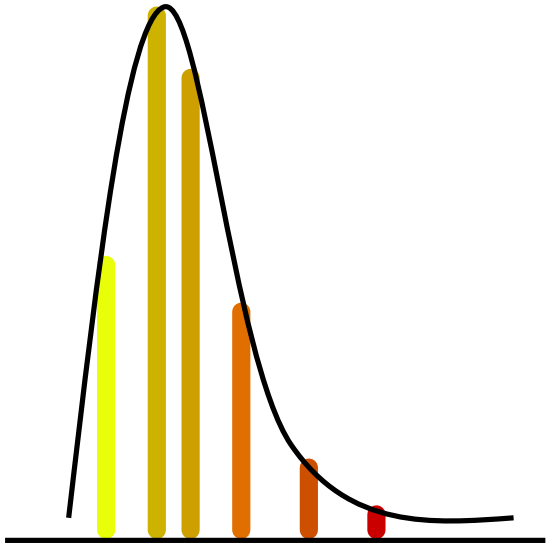
Uncorrelated relaxed clock

- Models available in *BEAST*

Exponential distribution
Most rates are quite low



Lognormal distribution
Most rates cluster around mean

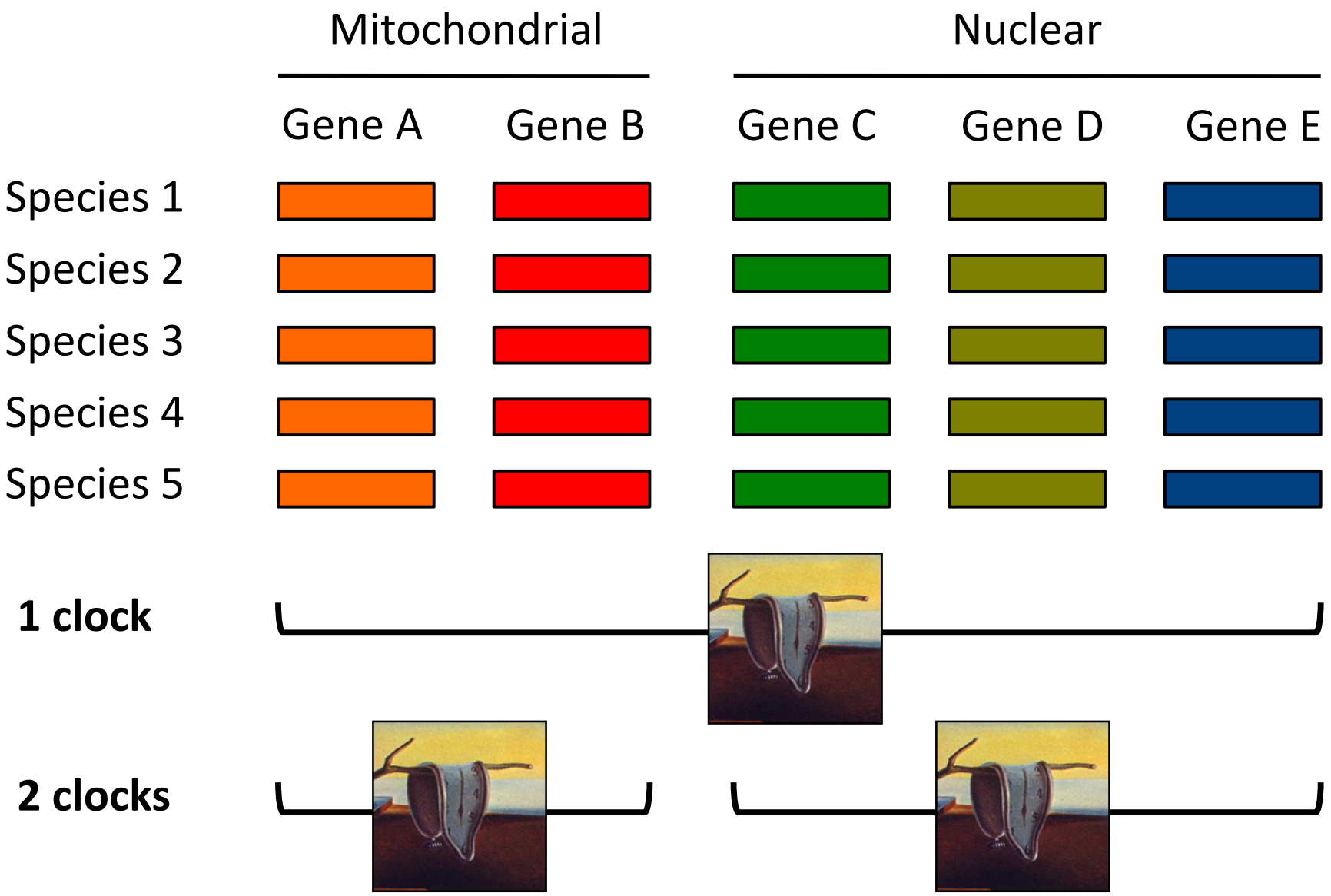


Drummond *et al.* (2006) *PLOS Biol*

Bayesian relaxed clocks

- In the uncorrelated lognormal relaxed clock, two statistics can be obtained:
  1. **Coefficient of variation of rates**
     Measures the rate variation among branches
     A value of 0 indicates clocklike evolution
  2. **Covariance of rates**
     Measures autocorrelation of rates between adjacent branches

Multiple clock models



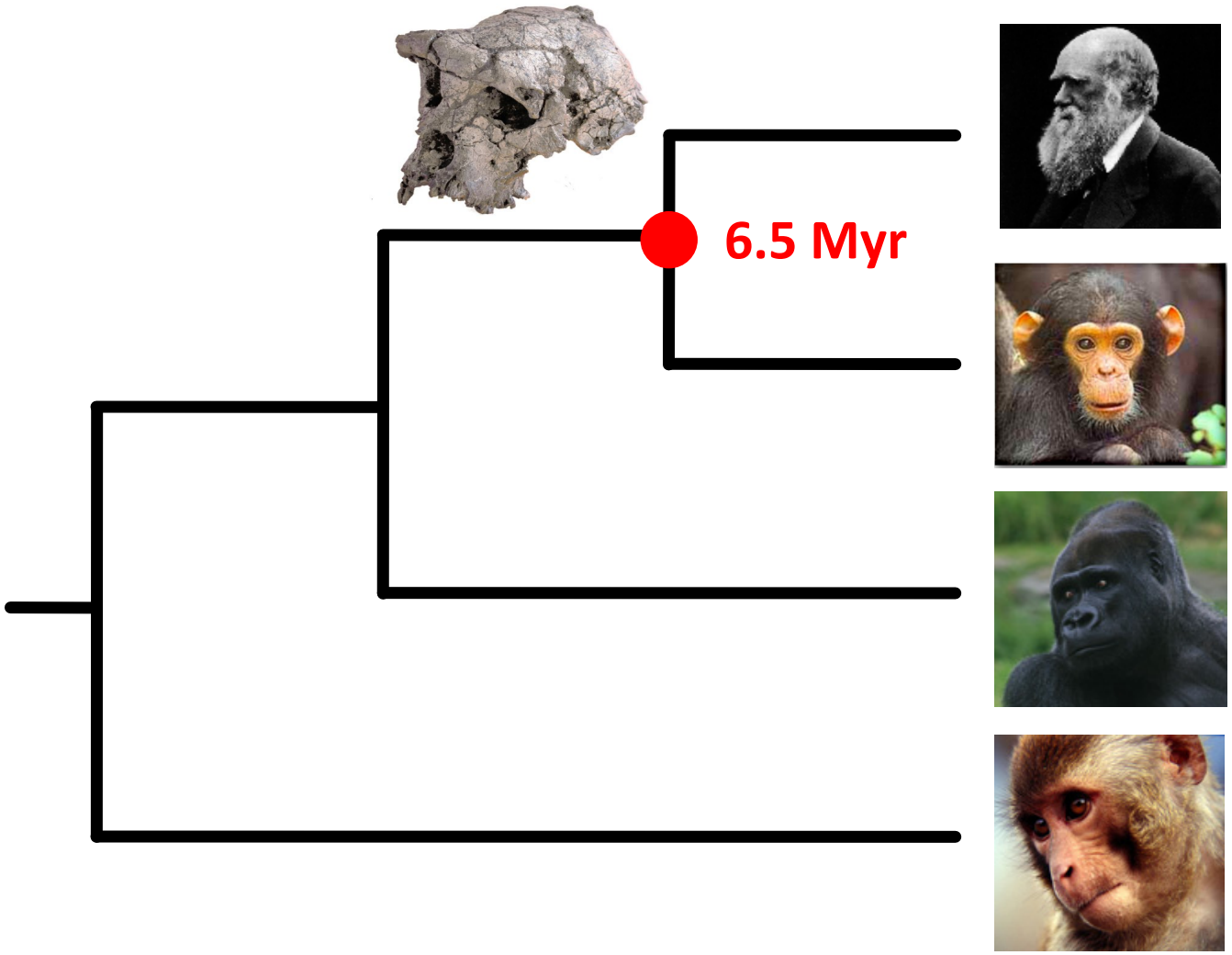
3. Calibrating the Molecular Clock

Calibrating the molecular clock

- Information about **substitution rate**
  - Use to fix rate or to specific prior distribution of rate

- Information about **node times**
  - Fossil record
  - Biogeography
  - Sampling times
  - Documented pedigree

Calibration: Fossil record

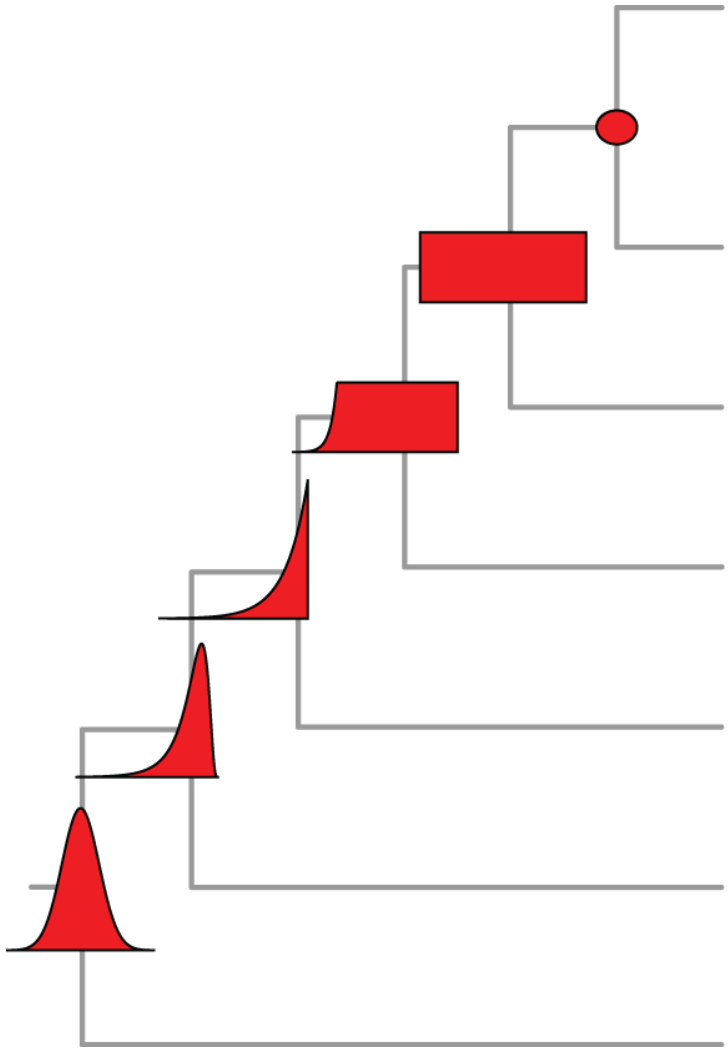


Calibration: Fossil record

1. **Use fossil data to inform priors on node times**
   - Minimum age of a node based on oldest fossil assignable to any of its descendent lineages
   - Prior distribution of node age specified by user

2. **Use fossil directly in the analysis**
   - Model diversification process use fossil occurrence data
   - Include fossil taxa in the data matrix (total-evidence dating)

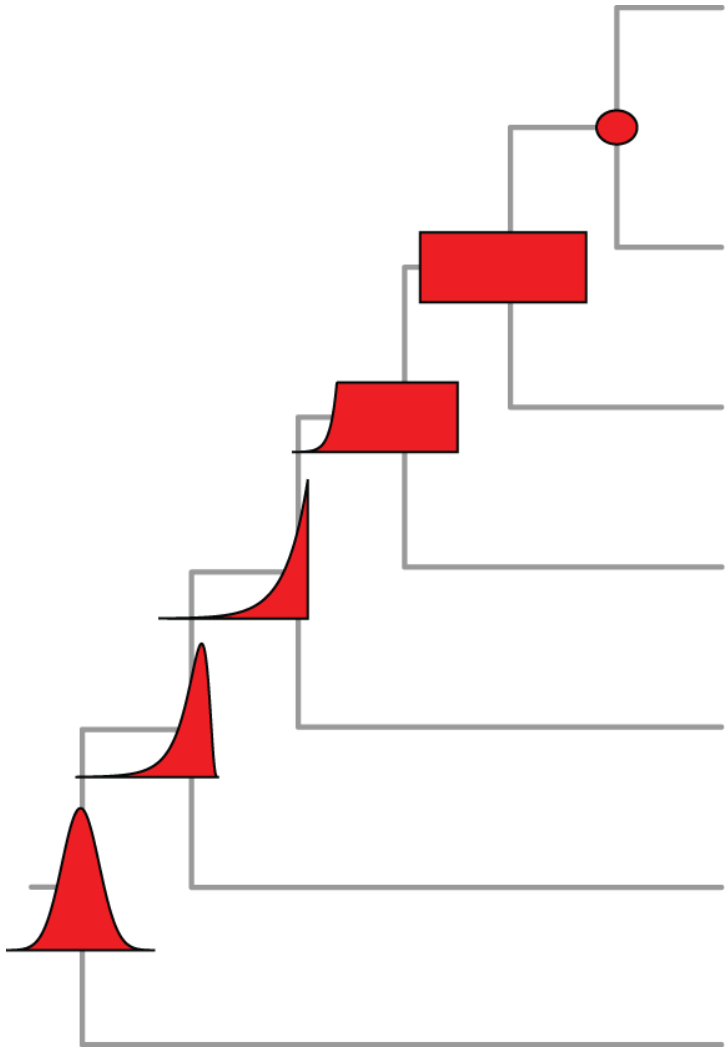
Calibrations



Uniform prior

- Combination of hard minimum and maximum bounds
- Does not effectively use information at hand
- Difficult to choose useful maximum bounds

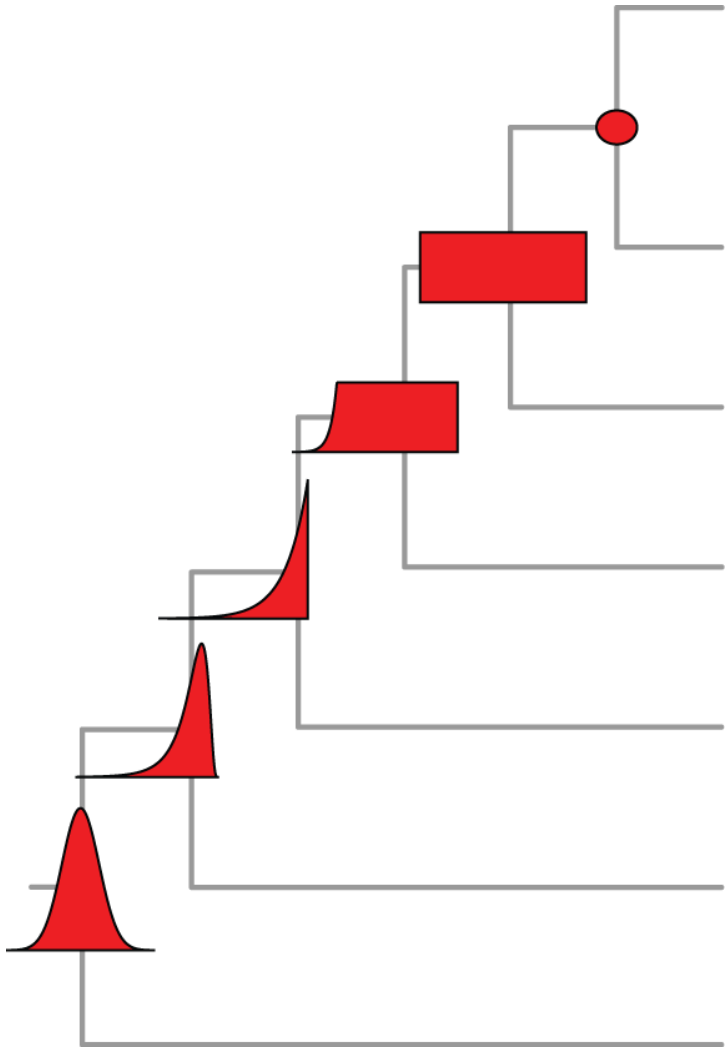
Calibrations

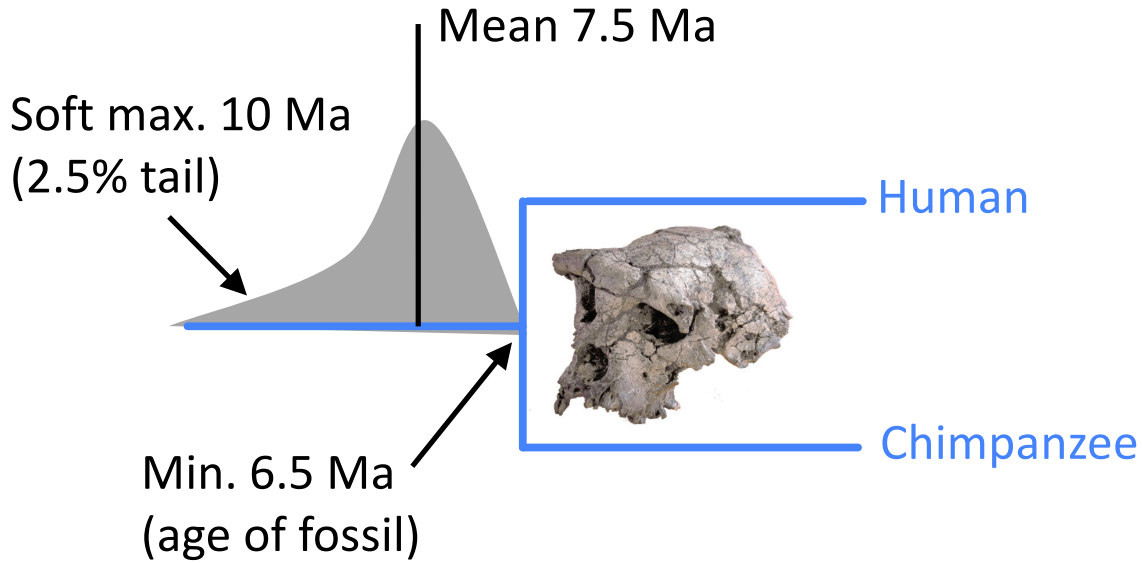


Exponential prior

- Need 2 values: minimum and mean
- Strong assumption about relationship of fossil taxon to internal node

Calibrations



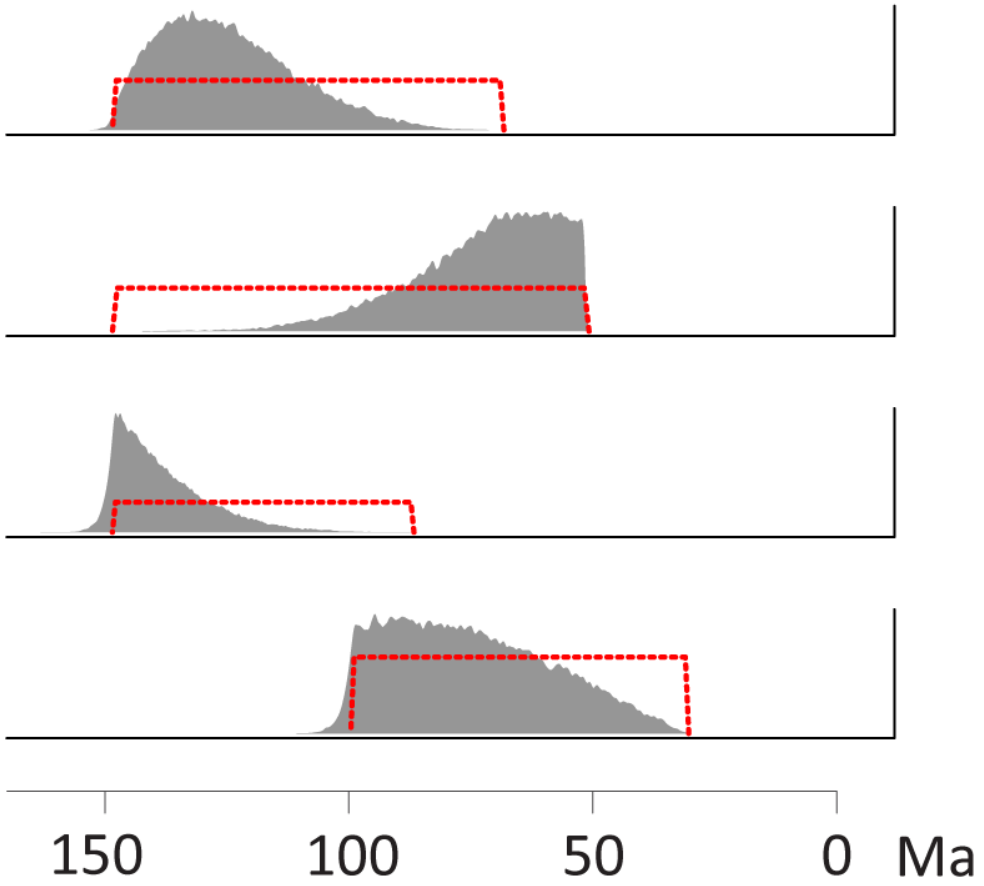


Lognormal prior

- Need 3 values: minimum, mean, and stdev
- Perhaps the most appropriate for fossils

Multiple calibrations

- Priors on node ages are the joint product of the tree prior and the user-specified calibration priors
- These priors can interact
- Marginal priors can differ from user-specified priors



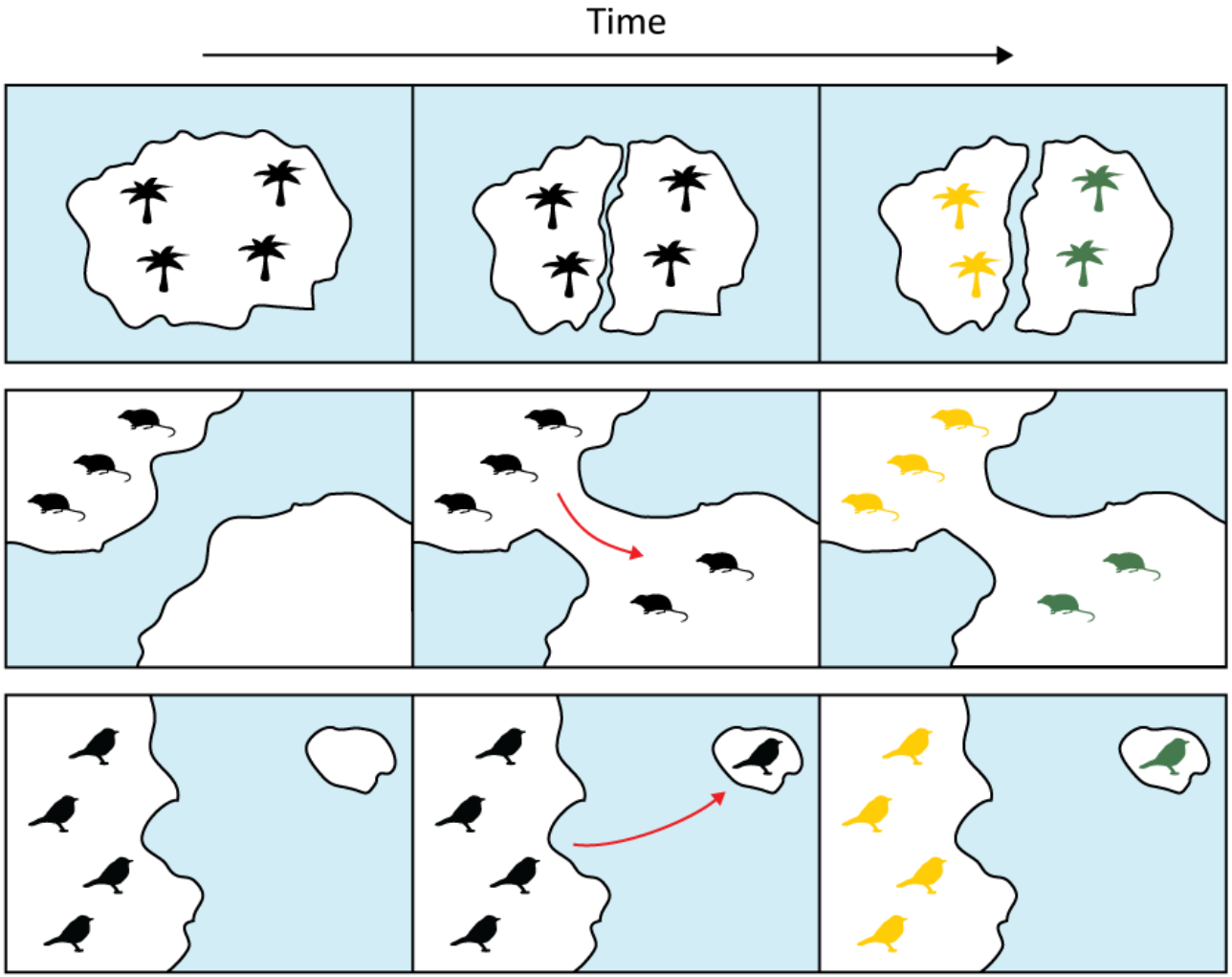
Warnock *et al*. (2014) *Proc Roy Soc B*

Biogeographic calibrations

Vicariance

Geodispersal

Biological dispersal



Sampling times