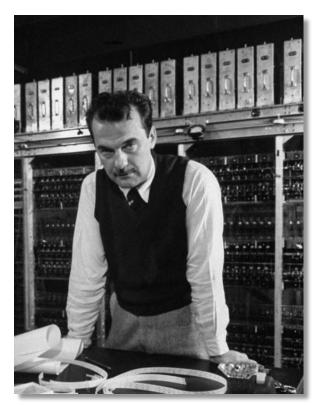
#### Lecture 2.2

# **MCMC Sampling**

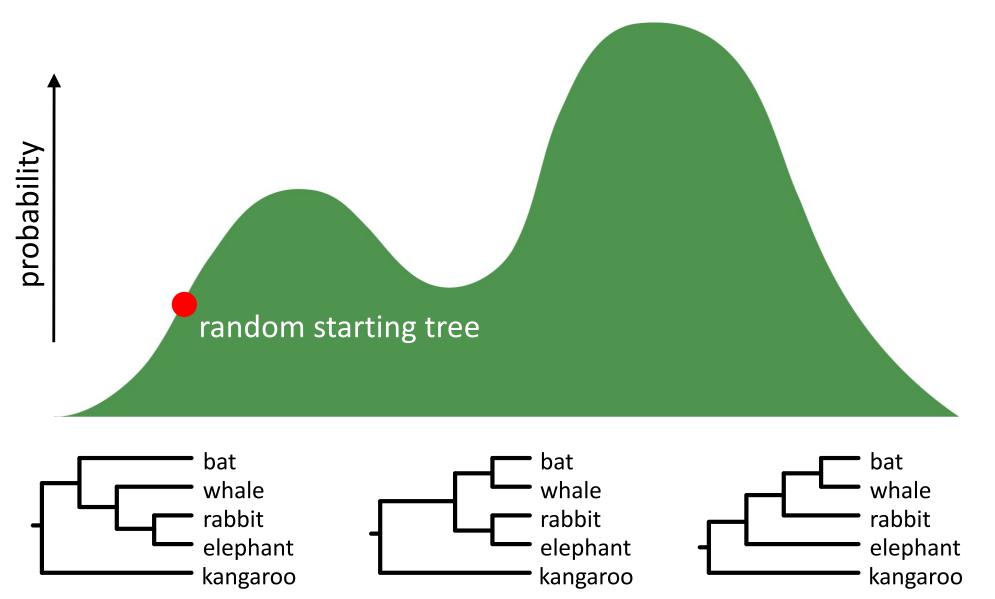
#### **Posterior Distribution**

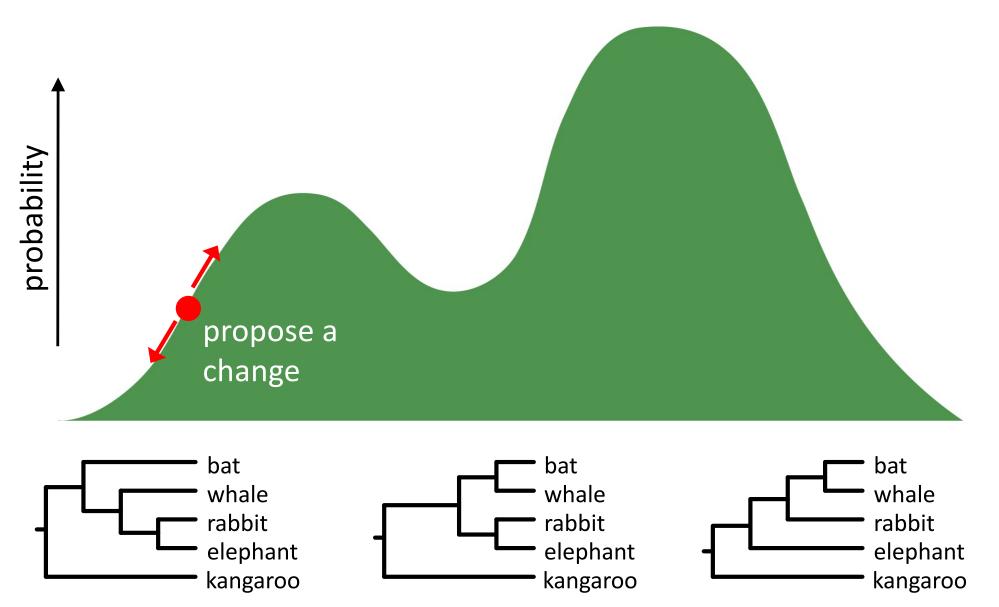
# Estimating the posterior

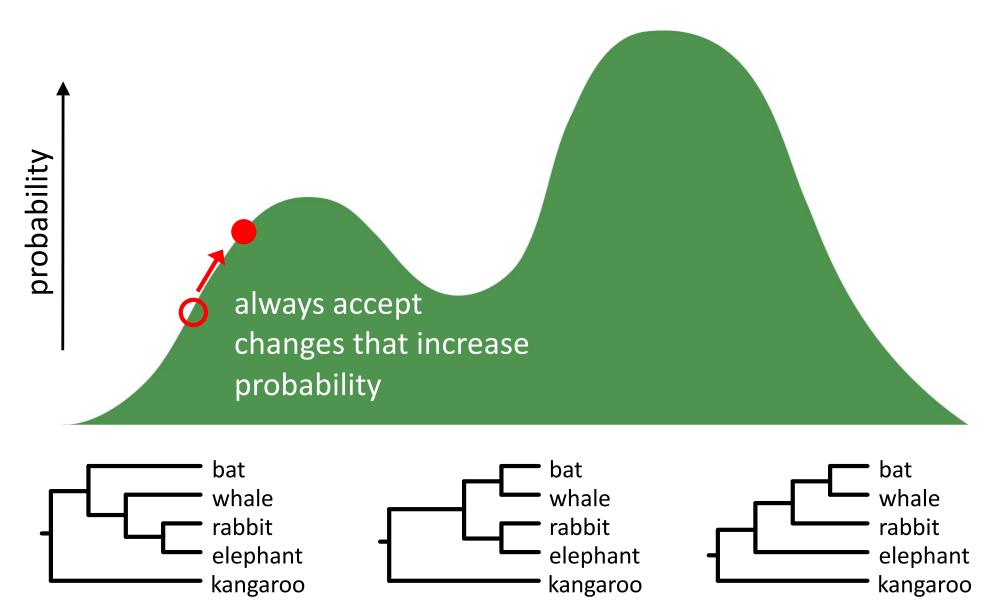
- Impossible to obtain the posterior directly
- Instead, the posterior can be estimated using Markov chain Monte Carlo simulation
- This is usually done using the Metropolis-Hastings algorithm

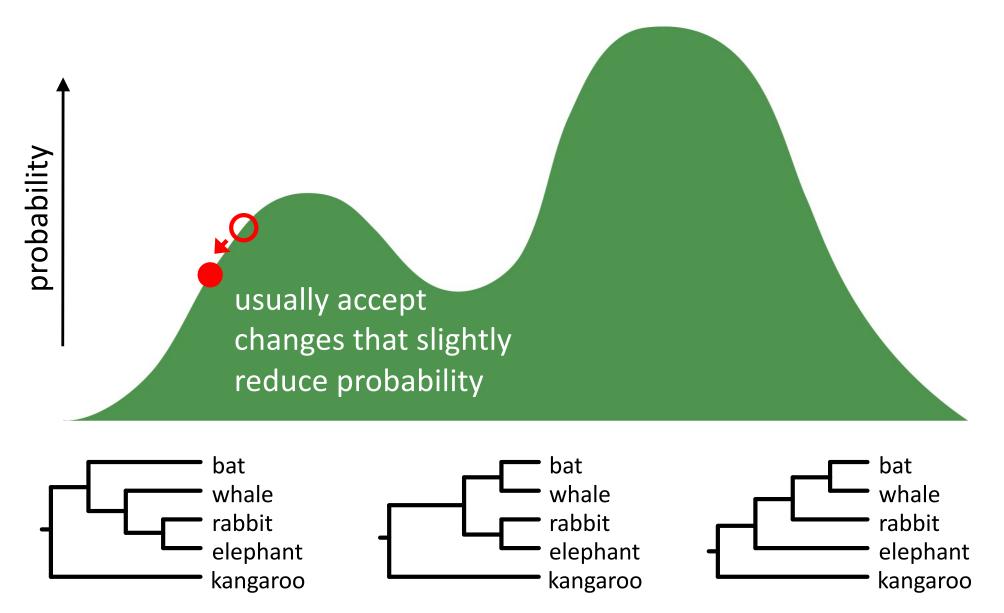


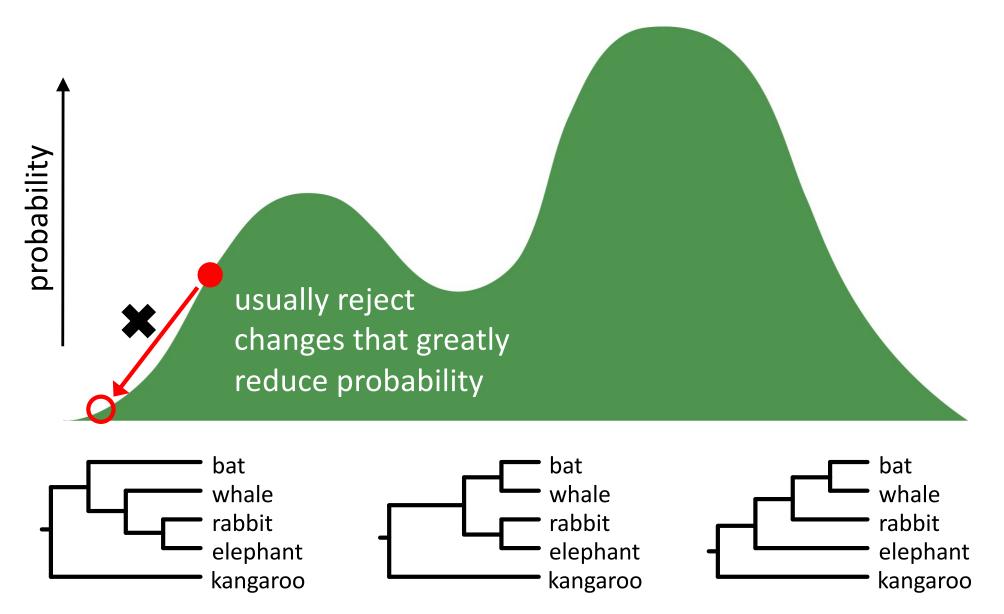
Nicholas Metropolis Los Alamos, 1953

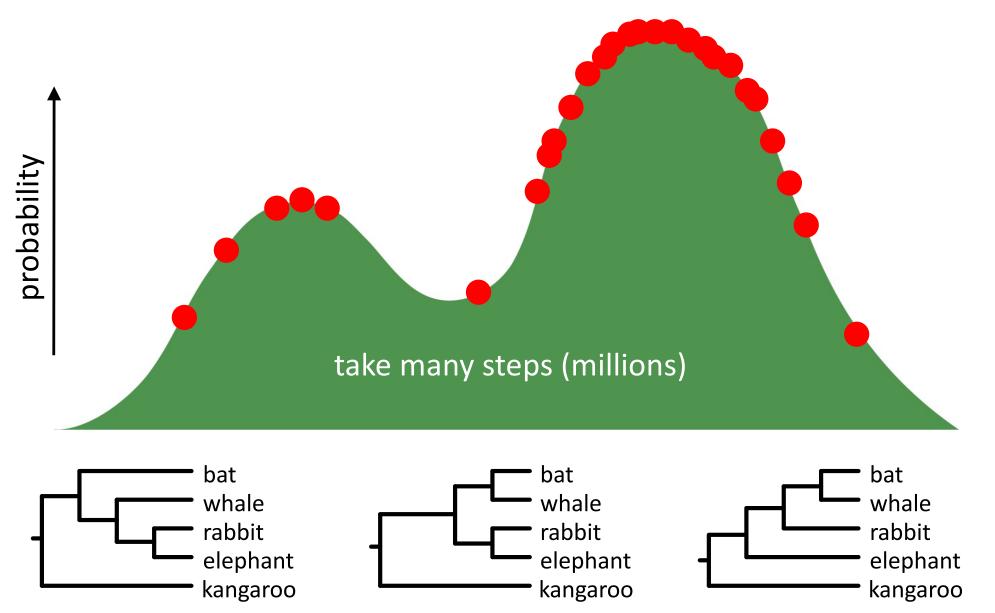




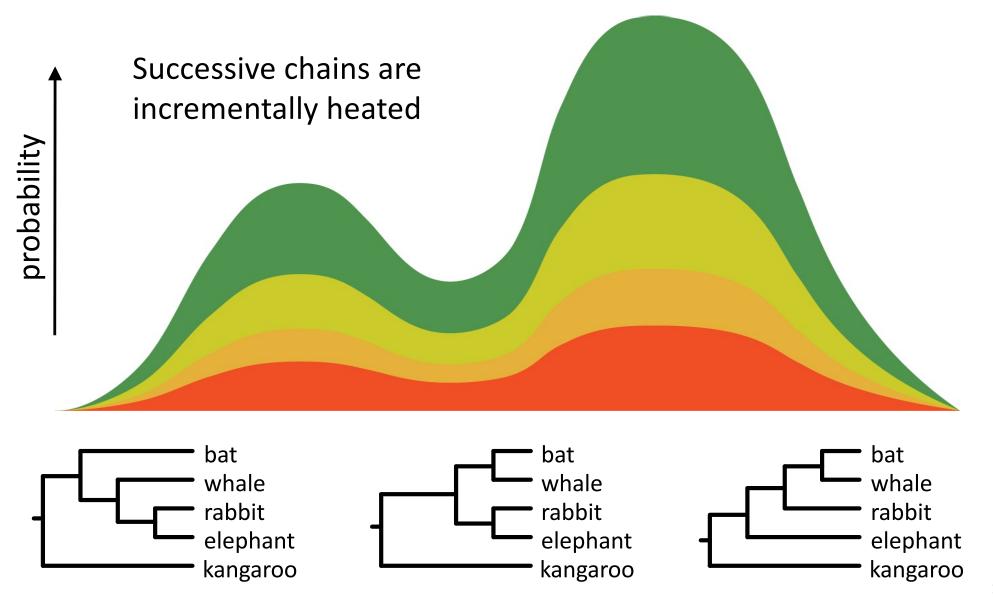




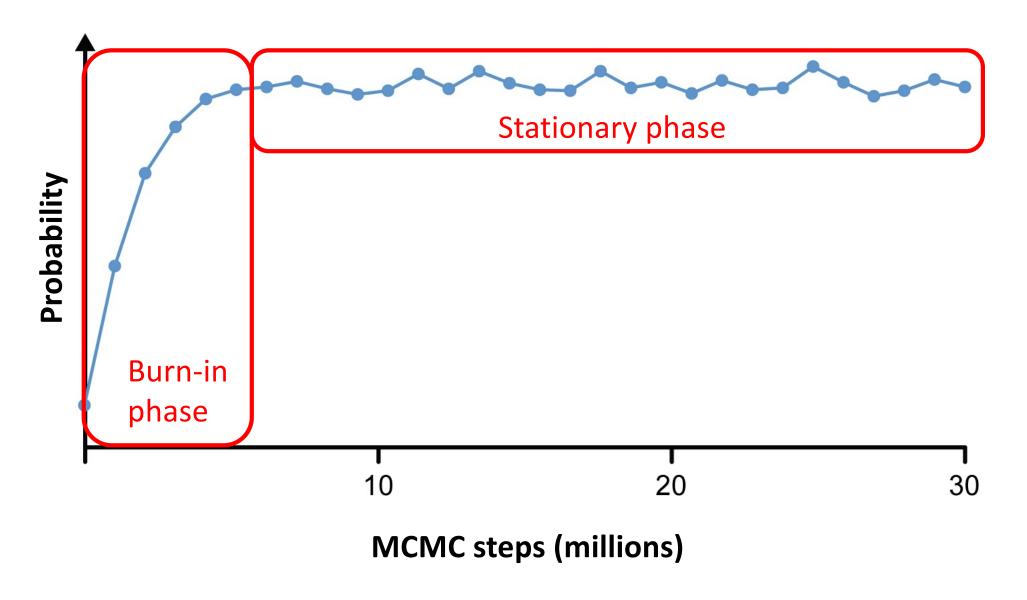


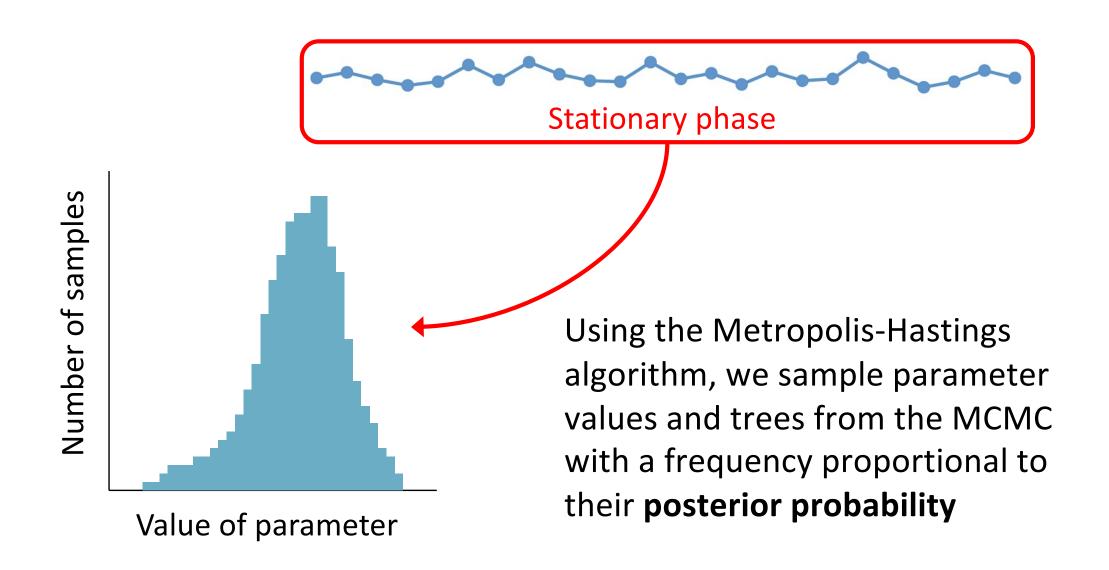


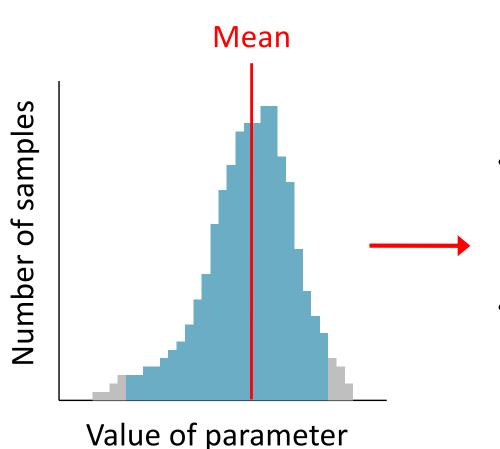
# Metropolis-coupled MCMC



- Output from a Bayesian phylogenetic analysis:
  - A list of the parameter values visited by the Markov chain (.p file in MrBayes, .log file in BEAST)
  - A list of the trees visited by the Markov chain (.t file in MrBayes, .trees file in BEAST)





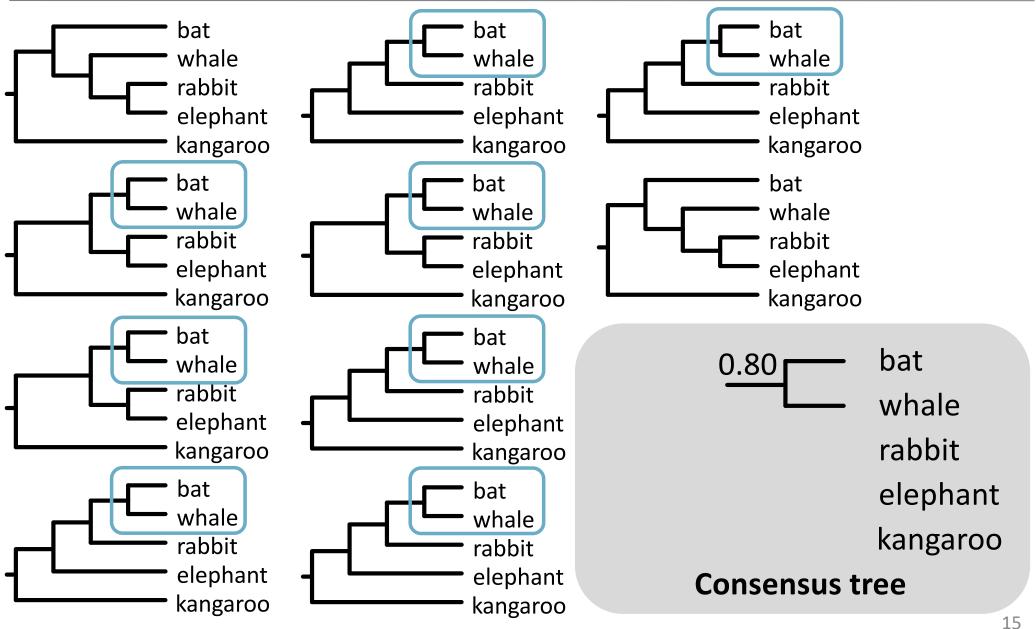


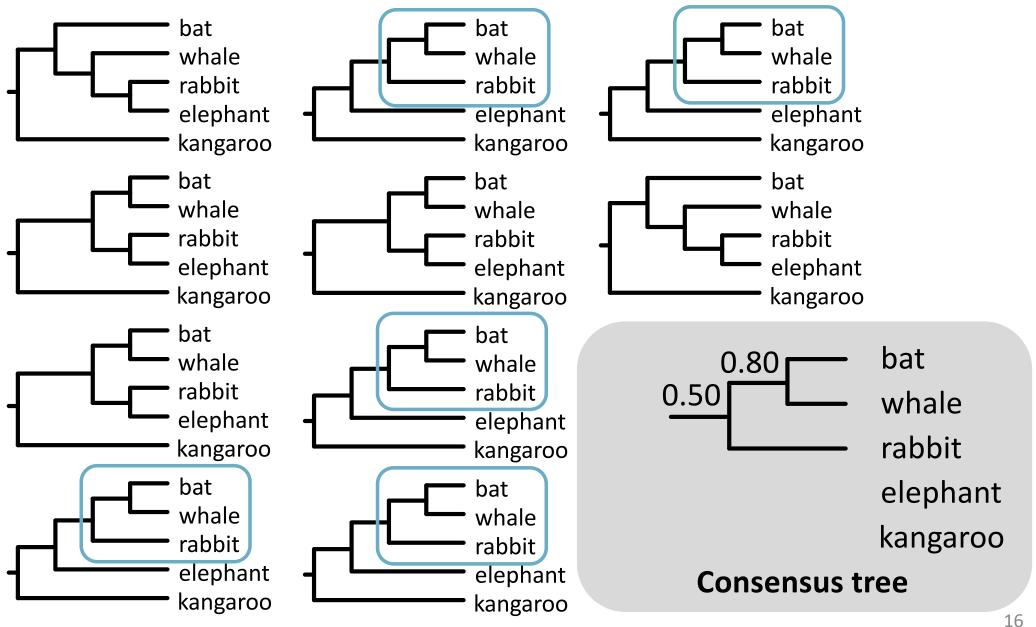
Take the mean of the sampled values

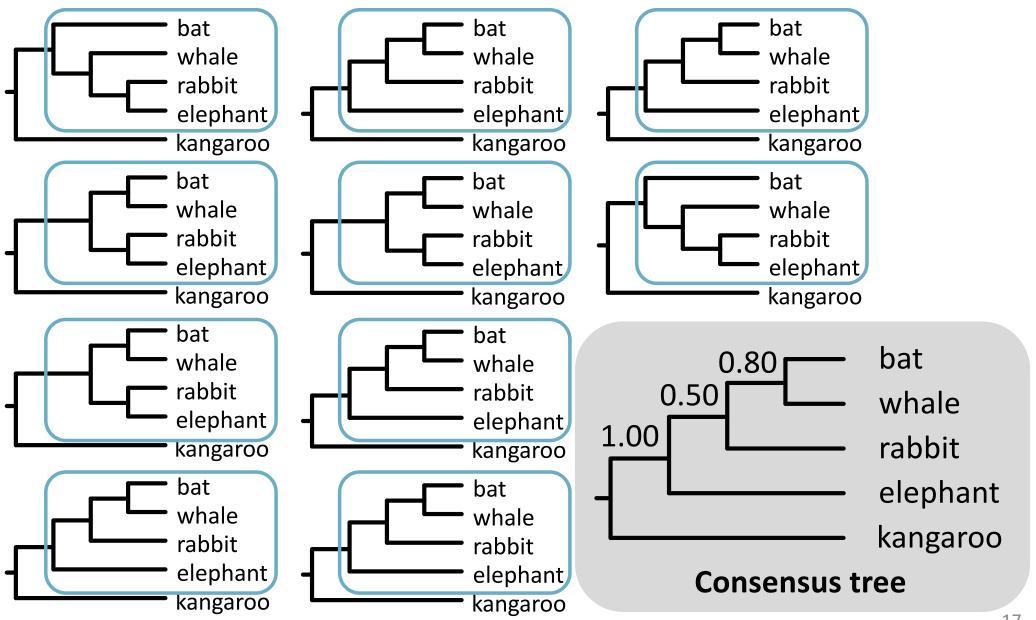
Mean posterior estimate

 Take the 'central' 95% of the sampled values

95% credibility interval







Majority-rule consensus tree (MrBayes)
 Shows all nodes with posterior probability >0.50

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

Maximum clade credibility (MCC) tree (BEAST/TreeAnnotator)
 Sampled tree with highest sum or product of posterior node
 probabilities

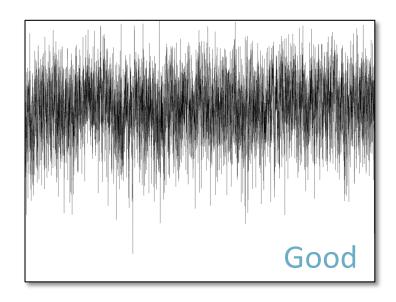
### Diagnostics

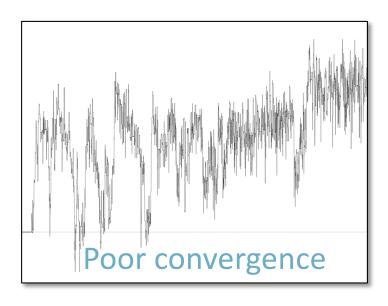
#### 1. Convergence

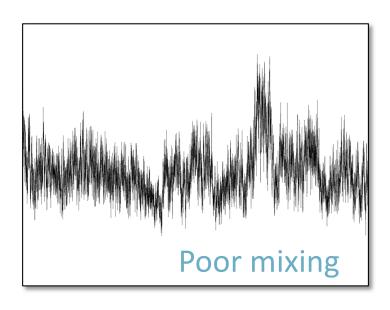
Are we drawing samples from the stationary distribution?

#### 2. Sufficient sampling

Have we drawn enough samples to allow a reliable estimate of the posterior distribution?







## Convergence

- Run at least 2 independent chains
- Likelihoods should be similar
- Estimates of model parameters should be similar

# Sufficient sampling

- Effective sample size (ESS)
   Have we drawn enough independent samples to produce a reliable estimate of the posterior distribution?
- ESS is preferably >200 for each parameter
- ESS can be increased by:
  - Increasing the length of the MCMC
     (and decreasing the frequency of sampling)
  - Modifying the MCMC proposals

# **Advantages and Problems**

## Advantages

- Able to implement complex (highly parameterised) models
- Estimating node support is straightforward
- Posterior probabilities have an intuitive interpretation
- Can incorporate independent information (in the prior)

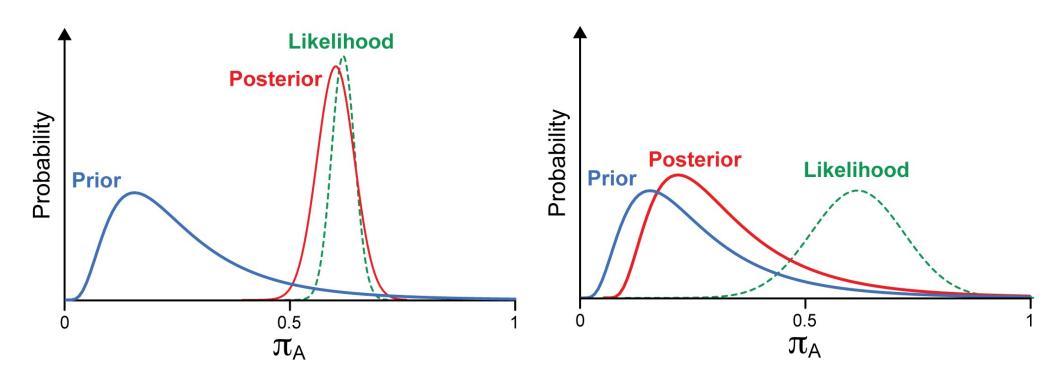
## Nuisance parameters

- Integrate over 'nuisance' parameters
- Marginal distribution of a parameter of interest

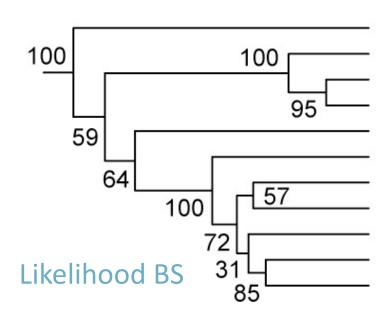
	Tree 1	Tree 2	Tree 3	
Branch lengths 1	0.10	0.07	0.12	0.29
Branch lengths 2	0.05	0.22	0.06	0.33
Branch lengths 3	0.05	0.19	0.14	0.38
Joint probabilities	0.20	0.48	0.32	Marginal probabilities

# Influence of priors

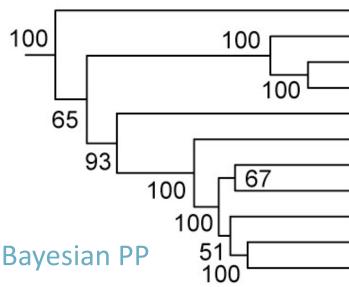
- Sensitivity of the posterior to the prior
- This problem can occur if the data are uninformative, the prior is strong, or both



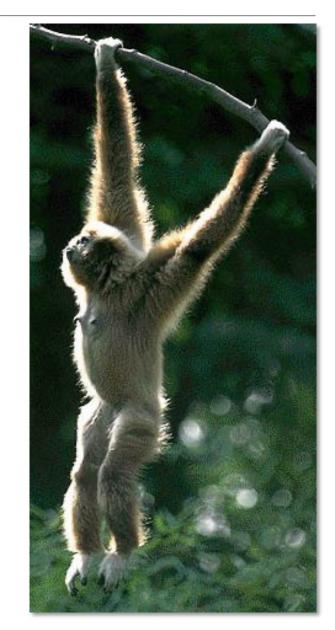
# Node support



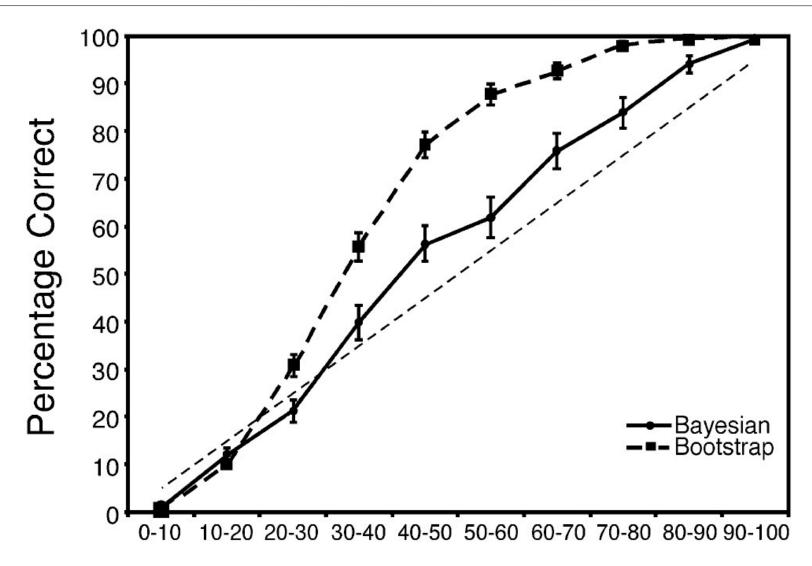
Symphalangus syndactylus
Nomascus concolor
Nomascus gabriellae
Nomascus leucogenys
Hoolock hoolock
Hylobates pileatus
Hylobates agilis
Hylobates muelleri
Hylobates klossii
Hylobates lar



Symphalangus syndactylus Nomascus concolor Nomascus gabriellae Nomascus leucogenys Hoolock hoolock Hylobates pileatus Hylobates agilis Hylobates moloch Hylobates klossii Hylobates lar



## Node support



Percentage Support

#### BEAST 1

- Bayesian Evolutionary Analysis by Sampling Trees
- Analyse population- or species-level data
- Simultaneous estimation of tree and node times
- Range of clock models
- Range of tree priors and demographic models





- Re-write of BEAST to increase modularity
- Users can extend BEAST by adding packages
- Additional tree priors not available in BEAST 1
- Capacity to perform simulations

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





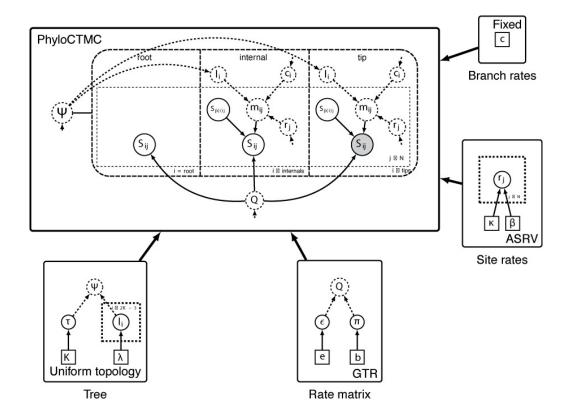
## **MrBayes**

- Primarily designed for species-level data
- Simultaneous estimation of tree and node times
- Range of clock models
- Range of tree priors
- Multiple chains and MCMC diagnostics



# RevBayes

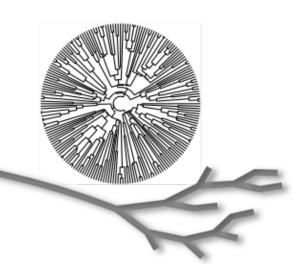
- Uses its own R-like language, Rev
- Interactive construction of graphical model
- Flexible and can be used for simulation and inference











- Analyses of large data sets on computing clusters
- Available priors similar to those in older versions of MrBayes
- Limited options, no molecular dating
- Likelihood component adapted from RAxML

#### Useful references

