

# Phylogenetic model comparison (by) selection & averaging

Remco R. Bouckaert  
[r.bouckaert@auckland.ac.nz](mailto:r.bouckaert@auckland.ac.nz)

Centre of Computational Evolution, University of Auckland  
Max Planck Institute for the Science of Human History

13 August 2019



- Model averaging
- Model selection
- Model comparison

# What does a model look like

Model = tree prior + site model + clock model + *priors* + *hyper priors*

startpage

phylogenetic models

go

≡

Web Images Videos Advanced

**Difference between Phylogenetic Trees and Cladistics**

- Phylogenetic tree - model of evolutionary history
- Cladistics - sprout or branch
- Studies traits to understand phylogeny
- Cladogram - model of the phylogeny of a species based on shared traits

**Phylogeny**

Ch. 7 & 8

**17.2 The Six Kingdoms**

**EK 1B2**

Phylogenetic trees and cladograms are phylogenetic models (models) of evolutionary history that can be tested

**Classification models**

- Phylogeny = evolutionary history
- Phylogenetic classification reflects evolutionary relationships
- Two models
  - Cladistics
  - Fan model

**Overview**

- Evolution and sequence variation
- Phylogenetic trees
  - The meaning of distance
  - Phylogenetic distance models
  - Constructing trees
    - Sequence alignment

**The Six Kingdoms of Organisms**

- Archaeabacteria
- Eubacteria
- Protists
- Fungi
- Plants
- Animals

**Parts of a phylogenetic tree**

**Generating new models for detecting phylogenetic signals of adaptation and dispersal effects of**

**Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology**

**modern phylogenetic comparative methods and their application in evolutionary biology**



## What does a model look like

## Model Comparison: Which one is better?

## Model Selection: Which one to pick?

Model Averaging: What if you don't want to choose?

Model = tree prior + site model + clock model + *priors* + *hyper priors*

# Model averaging

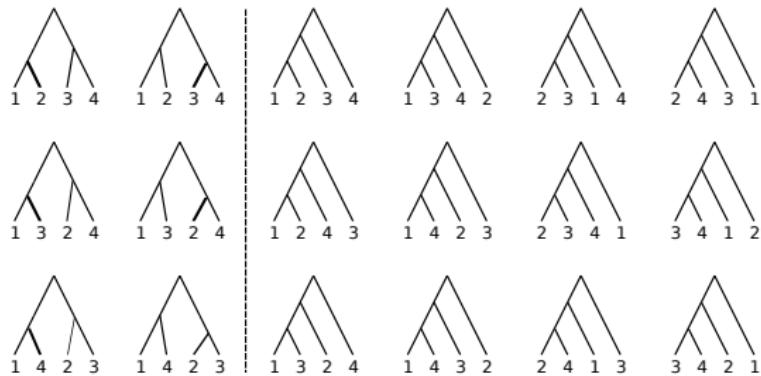
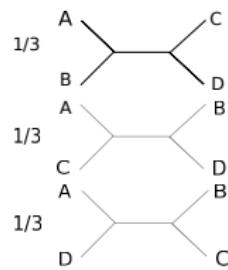
- Posterior:

$$p(\theta|D) = \sum_k \overbrace{P(M_k)}^{\text{model prior}} \underbrace{\pi(\theta|M_k)L(D|M_k, \theta)}_{\text{posterior}}$$

- Models can be substitution models, clock models, tree topologies, etc.
- Accounts for model uncertainty
- Requires specifying another prior  $P(M_k)$

# Model averaging

Already been doing that this yesterday: each tree topology is a model



18 ranked rooted trees –  $\frac{1}{3}$  probability of being balanced

# Model averaging: stochastic variable selection

Use indicator variable to select model

- Example: ancestral state reconstruction using mask matrix  $I$  and rate matrix  $R$

$$I = \begin{pmatrix} - & i_{12} & i_{13} & i_{14} \\ i_{21} & - & i_{23} & i_{24} \\ i_{31} & i_{32} & - & i_{34} \\ i_{41} & i_{42} & i_{43} & - \end{pmatrix} \quad R = \begin{pmatrix} - & r_{12} & r_{13} & r_{14} \\ r_{21} & - & r_{23} & r_{24} \\ r_{31} & r_{32} & - & r_{34} \\ r_{41} & r_{42} & r_{43} & - \end{pmatrix}$$

- Use  $r_{ij}$  if  $i_{ij}$  is true, but use rate 0 if  $i_{ij}$  is false
- Sample  $I$  and all rates in  $R$  throughout MCMC run.
- Use strong prior on number of  $i_{ij} = \text{true}$  to reduce number of non-zero rates

Lemey et al, PLoS Comput Biol, 2009

# Model averaging: stochastic variable selection

Use indicator variable to select model

- Example: ancestral state reconstruction using mask matrix  $I$  and rate matrix  $R$

$$I = \begin{pmatrix} - & i_{12} & i_{13} & i_{14} \\ i_{21} & - & i_{23} & i_{24} \\ i_{31} & i_{32} & - & i_{34} \\ i_{41} & i_{42} & i_{43} & - \end{pmatrix} \quad R = \begin{pmatrix} - & r_{12} & r_{13} & r_{14} \\ r_{21} & - & r_{23} & r_{24} \\ r_{31} & r_{32} & - & r_{34} \\ r_{41} & r_{42} & r_{43} & - \end{pmatrix}$$

- Use  $r_{ij}$  if  $i_{ij}$  is true, but use rate 0 if  $i_{ij}$  is false
- Sample  $I$  and all rates in  $R$  throughout MCMC run.
- Use strong prior on number of  $i_{ij} = \text{true}$  to reduce number of non-zero rates

Lemey et al, PLoS Comput Biol, 2009

Stochastic variable selection:

- Simple to implement
- Potentially inefficient in sampling unused parameters

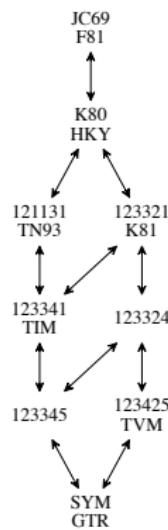
# Model averaging: reversible jump

The probability of acceptance of a (possibly trans-dimensional) proposal is

$$\min\{1, \text{posterior ratio} \times \text{proposal ratio} \times \text{Jacobian}\}$$

- **posterior ratio** is the posterior of the proposed state  $S'$  divided by that of the current state  $S$ ,
- **proposal ratio** the probability of moving from  $S$  to  $S'$  divided by the probability of moving back from  $S'$  to  $S$
- **Jacobian** is the determinant of the matrix of partial derivatives of the parameters in the proposed state with respect to that of the current state

Green, Biometrika, 1995



# Model averaging: reversible jump

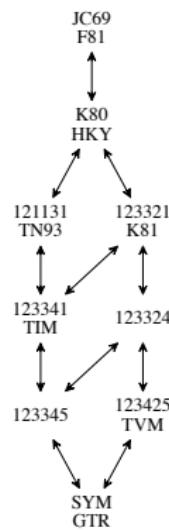
The probability of acceptance of a (possibly trans-dimensional) proposal is

$$\min\{1, \text{posterior ratio} \times \text{proposal ratio} \times \text{Jacobian}\}$$

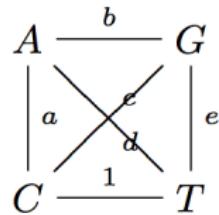
- **posterior ratio** is the posterior of the proposed state  $S'$  divided by that of the current state  $S$ ,
- **proposal ratio** the probability of moving from  $S$  to  $S'$  divided by the probability of moving back from  $S'$  to  $S$
- **Jacobian** is the determinant of the matrix of partial derivatives of the parameters in the proposed state with respect to that of the current state Green, Biometrika, 1995

Reversible jump:

- Hard to implement correctly
- Efficient sampling



# Reversible-jump Based (RB) substitution model for nucleotides

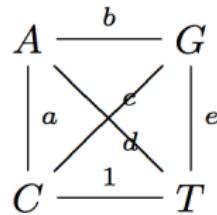


$$R = \begin{vmatrix} - & \alpha & \beta & \gamma \\ \alpha & - & \delta & \epsilon \\ \beta & \delta & - & \omega \\ \gamma & \epsilon & \omega & - \end{vmatrix}$$

frequencies  $\pi$   
 $Q = \pi R$

transition prob.  
 $P(t) = e^{Qt}$

# Reversible-jump Based (RB) substitution model for nucleotides



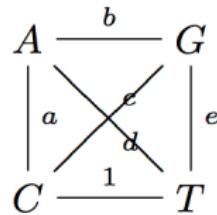
$$R = \begin{vmatrix} - & \alpha & \beta & \gamma \\ \alpha & - & \delta & \epsilon \\ \beta & \delta & - & \omega \\ \gamma & \epsilon & \omega & - \end{vmatrix}$$

frequencies  $\pi$   
 $Q = \pi R$

transition prob.  
 $P(t) = e^{Qt}$

	$\alpha$	$\beta$	$\gamma$	$\delta$	$\epsilon$	$\omega$	# dimensions
F81 (JC69)	1	1	1	1	1	1	0
HKY85 (K80)	a	1	a	a	1	a	1
TN93	a	b	a	a	1	a	2
TIM	a	b	c	c	1	a	3
new	a	b	c	d	1	a	4
GTR (SYM)	a	b	c	d	1	e	5

# Reversible-jump Based (RB) substitution model for nucleotides



$$R = \begin{vmatrix} - & \alpha & \beta & \gamma \\ \alpha & - & \delta & \epsilon \\ \beta & \delta & - & \omega \\ \gamma & \epsilon & \omega & - \end{vmatrix}$$

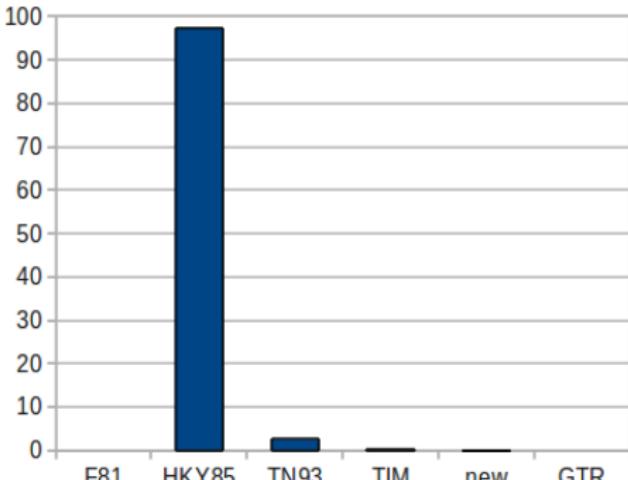
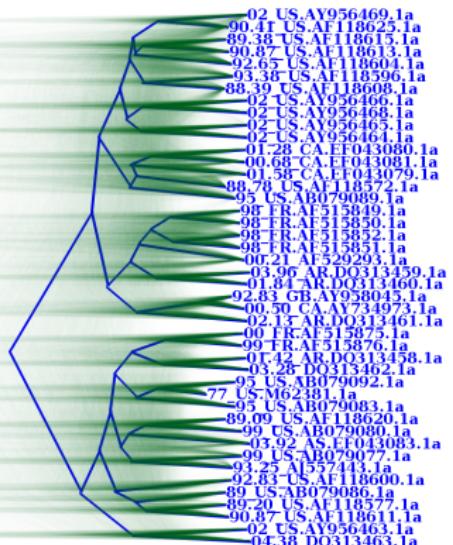
frequencies  $\pi$   
 $Q = \pi R$

transition prob.  
 $P(t) = e^{Qt}$

	$\alpha$	$\beta$	$\gamma$	$\delta$	$\epsilon$	$\omega$	# dimensions
F81 (JC69)	1	1	1	1	1	1	0
HKY85 (K80)	a	1	a	a	1	a	1
TN93	a	b	a	a	1	a	2
TIM	a	b	c	c	1	a	3
new	a	b	c	d	1	a	4
GTR (SYM)	a	b	c	d	1	e	5

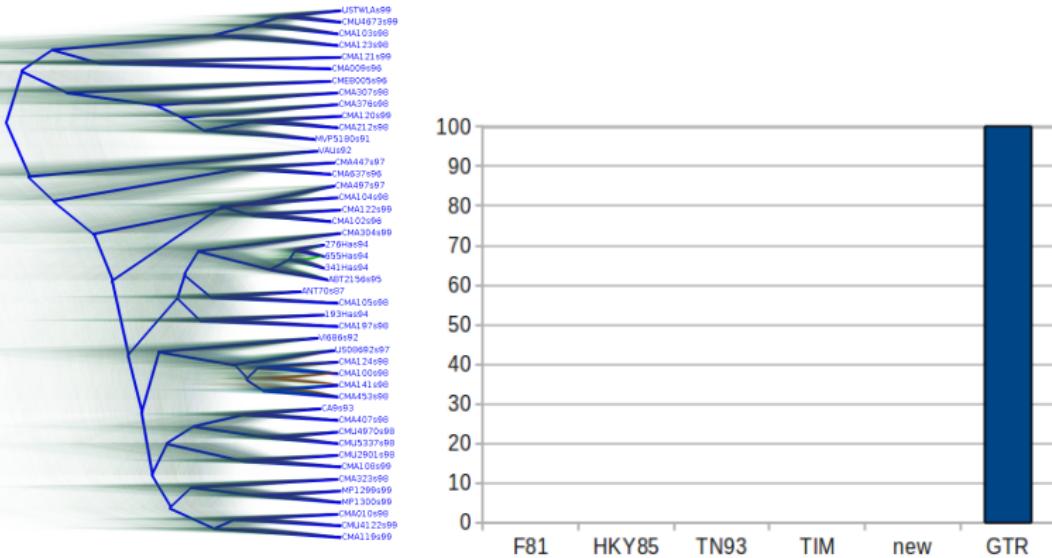
Increase dimension by drawing a new parameter value from  $\Gamma(0.2, 5)$

# Hepatitis C: HKY massively preferred



44 taxa, 576 sites, Bayesian skyline, 0 gamma, 0 prop invariable

# HIV: only GTR will do



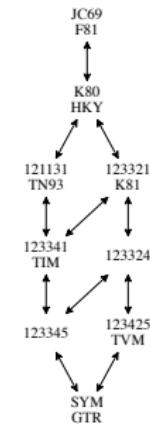
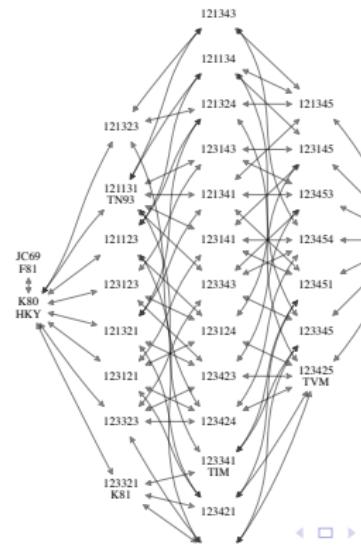
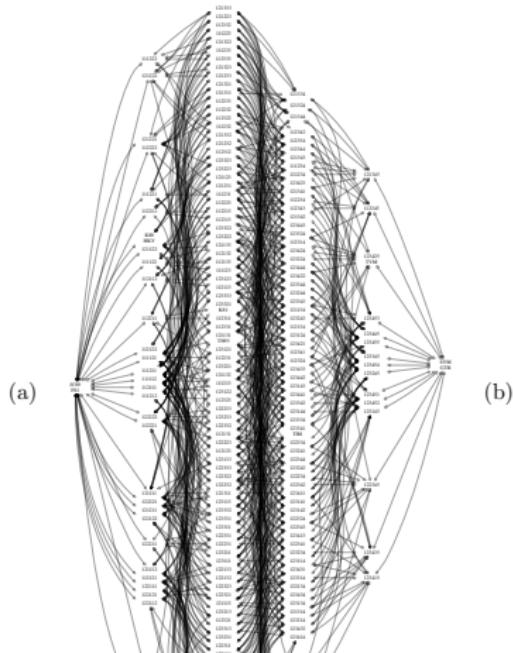
Same results without and with partitioning sites at codon positions {1,2},{3} or {1},{2},{3}

46 taxa, 2382 sites, Coalescent (exp), 0  $\Gamma$ , 0 PI. MP, FPV, animal dna, primates.

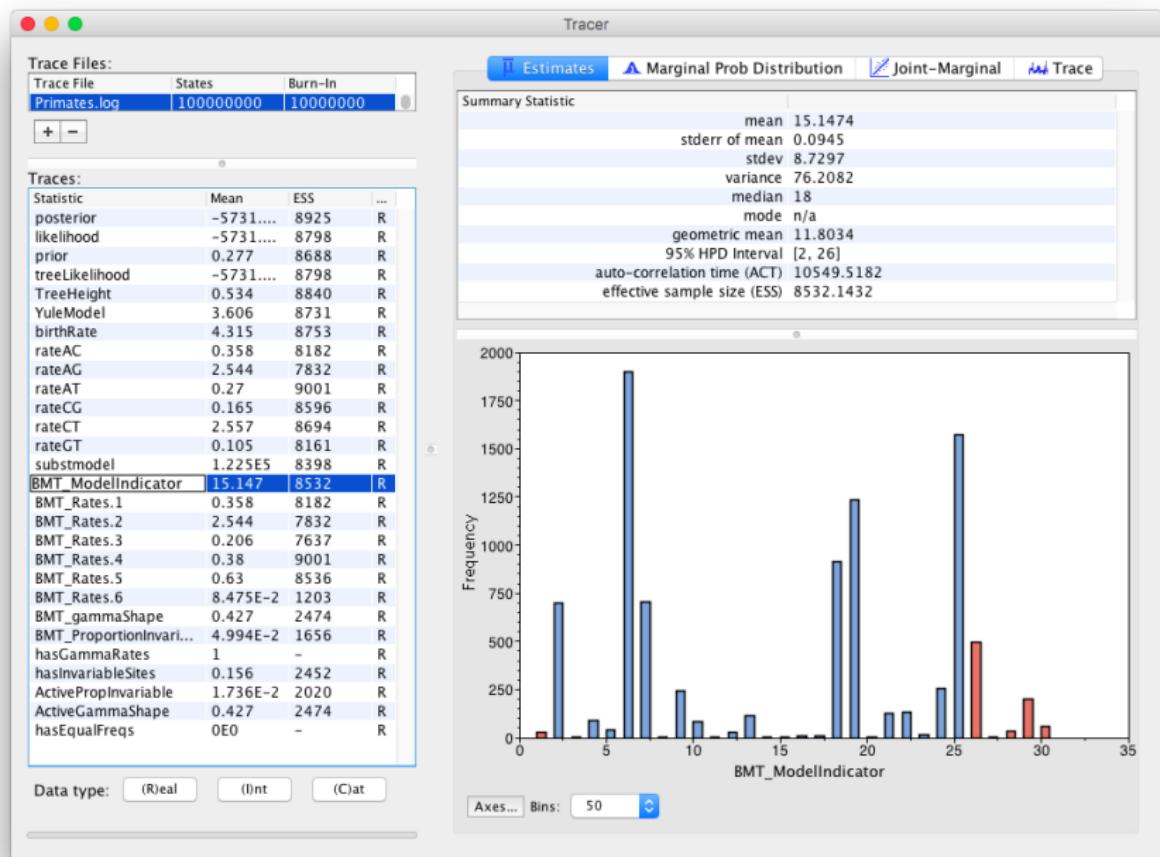
## Model averaging: bModelTest

Bouckaert & Drummond, BMC Evo Bio, 6 Feb 2017

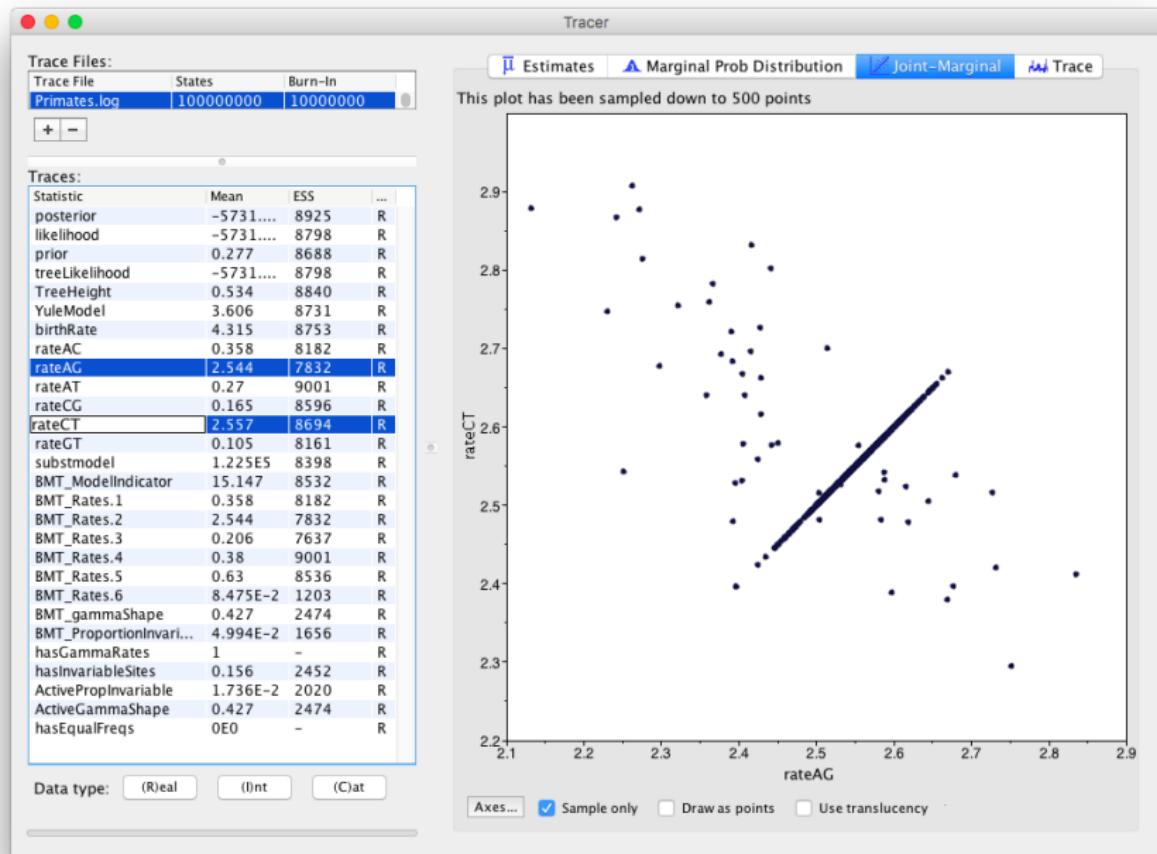
- Averages over substitution models
  - Estimated/fixed frequencies
  - With/without gamma rate heterogeneity
  - With/without gamma proportion invariable sites



# Model averaging: bModelTest



# Model averaging: bModelTest

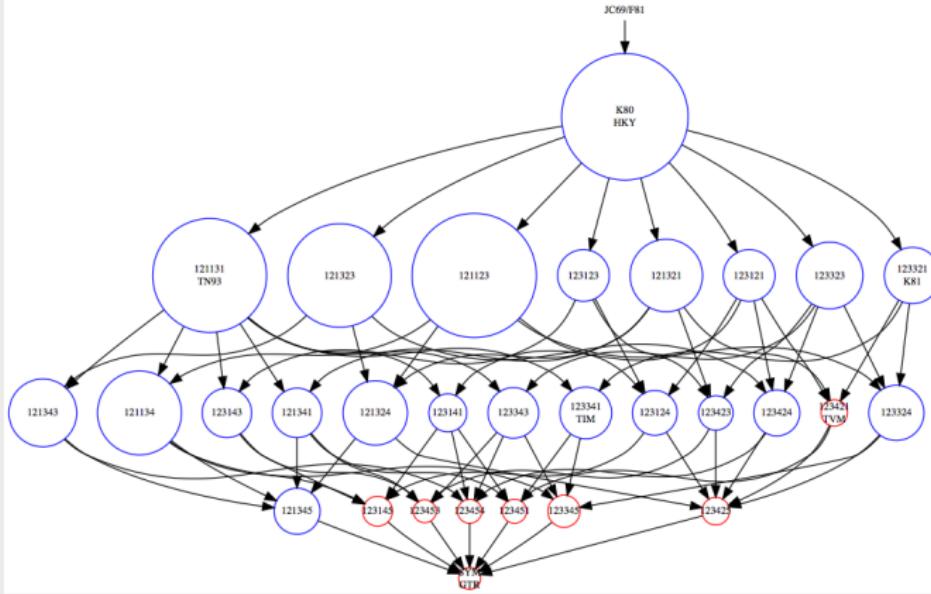


# Model averaging: bModelTest

File: Primates.log item: substmodel

Models with blue circles are inside 95%HPD, red outside, and without circles have 0.00% support.

posterior support	cumulative support	model
13.77%	13.77%	121121
13.21%	26.99%	121123
11.16%	38.15%	121131
9.27%	47.42%	121323
6.05%	53.47%	121134
4.44%	57.91%	121321
3.94%	61.85%	121343
3.78%	65.63%	123323
3.50%	69.13%	121324
2.72%	71.85%	123321
2.61%	74.46%	123324
2.33%	76.79%	123121
2.28%	79.07%	123123
2.28%	81.34%	123341
2.17%	83.51%	123343
2.05%	85.56%	121341
2.05%	87.52%	123143
1.83%	89.45%	123424
1.78%	91.23%	121345
1.72%	92.95%	123124
1.22%	94.17%	123141
1.05%	95.22%	123423
0.89%	96.11%	123345

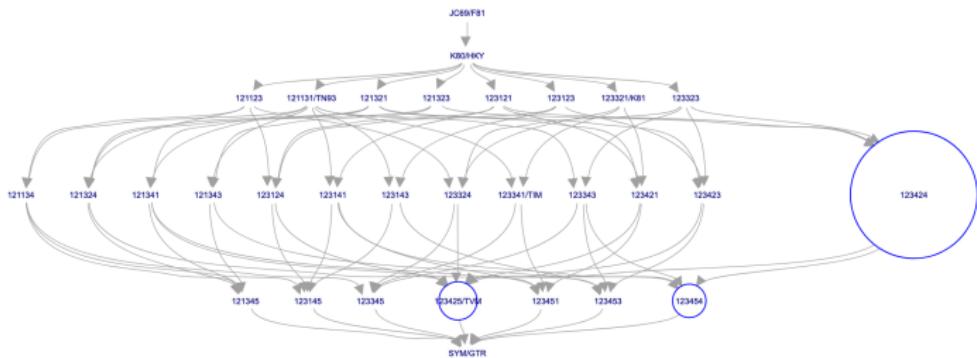


# Model averaging: bModelTest

File: hbvz-bmt-cexp.log item: substmodel

Models with blue circles are inside 95%HPD, red outside, and without circles have at most 0.34% support.

posterior support	cumulative support	model
85.91%	85.91%	123424
7.79%	93.70%	123425
5.96%	99.66%	123454
0.34%	100.00%	123456



## Model averaging: bModelTest model priors

Posterior:

$$p(\theta|D) = \sum_k P(M_k) \pi(\theta|M_k) L(D|M_k, \theta)$$

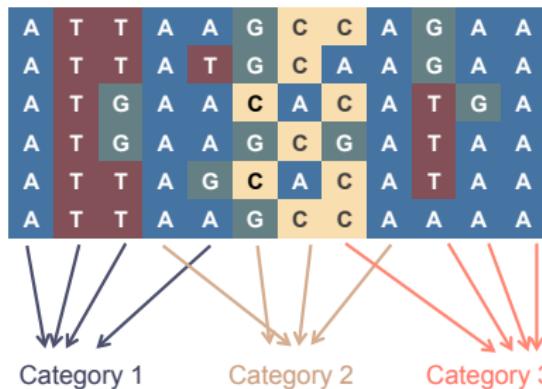
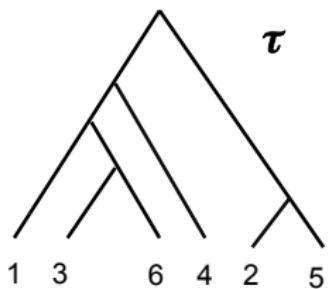
$P(M_k)$  model prior:

- uniform over substitution models in model set
  - ▶ For 31 model set:  $P(JC) = P(HKY) = P(TN) = P(GTR) = \frac{1}{31}$
- uniform over number of parameters in substitution models
  - ▶ For 31 model set:  $P(JC) = P(HKY) = P(GTR) = \frac{1}{6}$  but  
 $P(TN) = \frac{1}{6} \cdot \frac{1}{8} = \frac{1}{48}$

# Bayesian inference

- Posterior distributions of parameters of interest are simulated by an algorithm called Markov chain Monte Carlo (MCMC)

Phylogenetic tree



Substitution model parameters

 $\phi_1^*$ 

Model

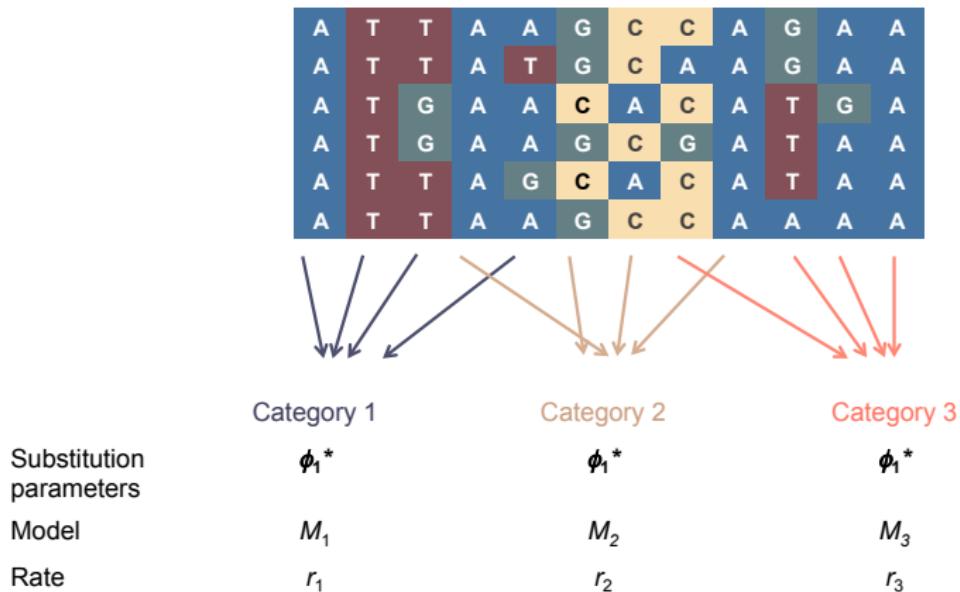
 $M_1$  $\phi_1^*$  $M_2$  $\phi_1^*$  $M_3$  $\phi$ 

Rate

 $r_1$  $r_2$  $r_3$  $r$ Population history or speciation hyperparameters  $\theta$

## Subst BMA

Site rates and substitution patterns share a common Dirichlet Process prior (DPM1)



# Subst BMA

Site rates and substitution patterns have independent Dirichlet Process priors (DPM2)

Substitution parameters

$\phi_1^*$

$\phi_2^*$

Model

$M_1$

$M_2$

Category 1 Category 2

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

Rate

$r_1$

$r_2$

$r_3$

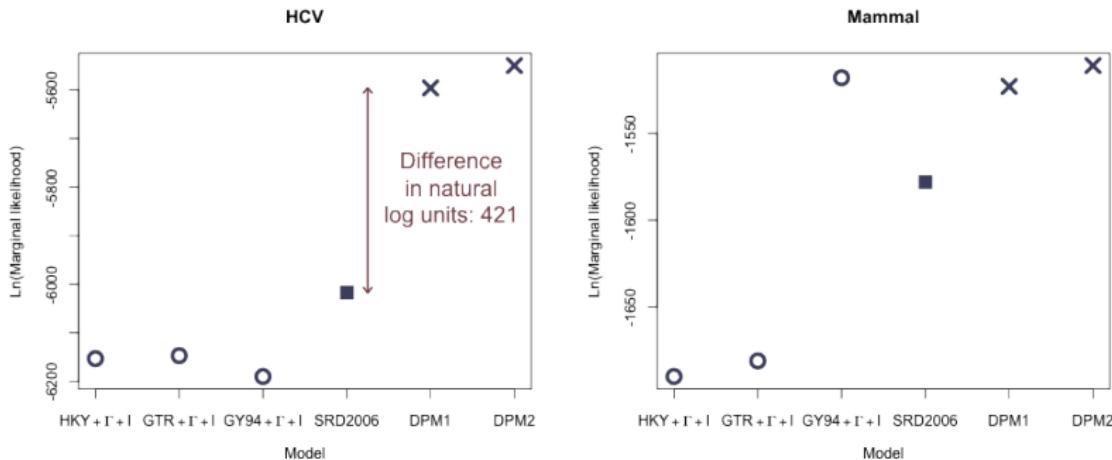
Category 1

Category 2

Category 3

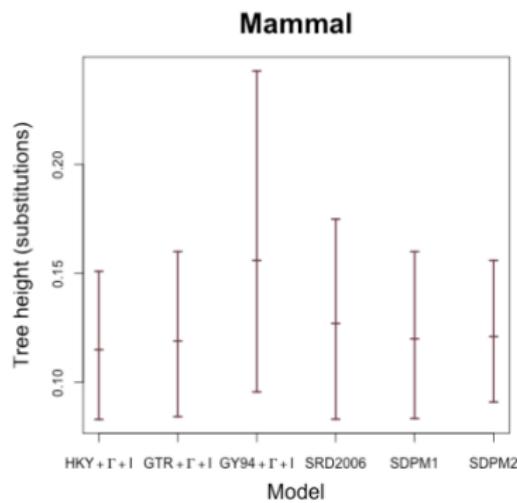
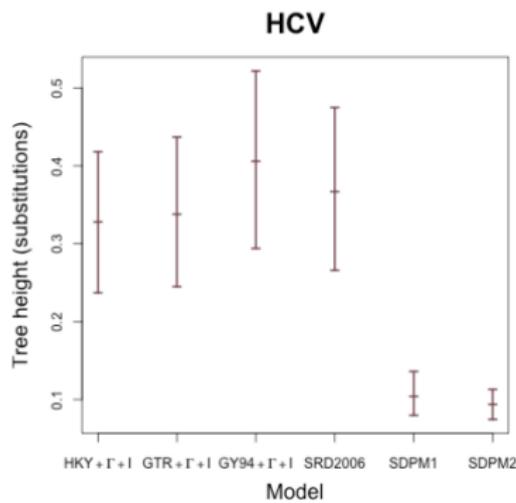
# Marginal Likelihoods

- Use marginal likelihood as a measure of how well the model explains the data.

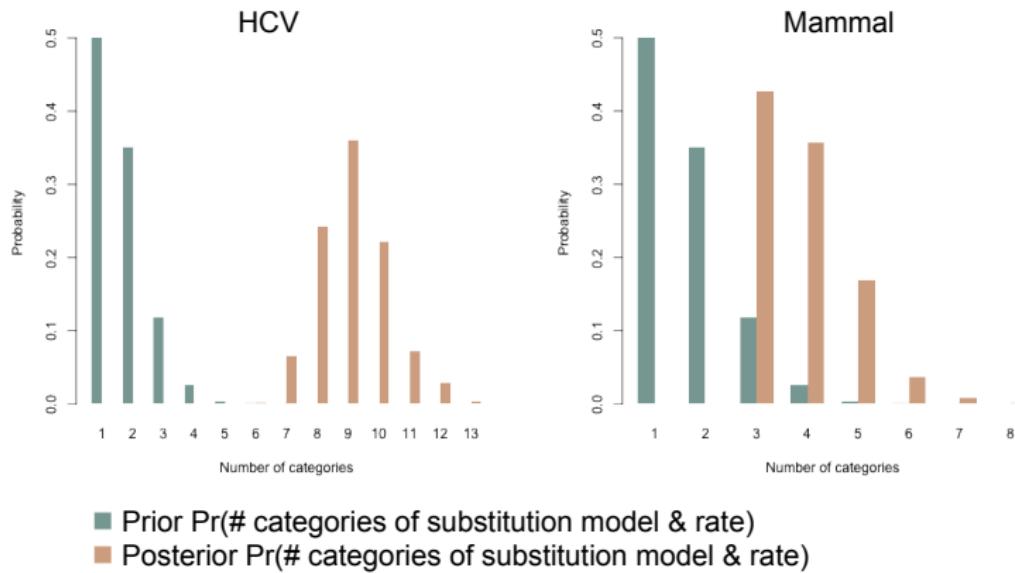


- Substitution pattern assumed homogeneous across sites
- Alignment partition for substitution pattern defined prior to the analysis
- ✗ Alignment partition for substitution pattern estimated

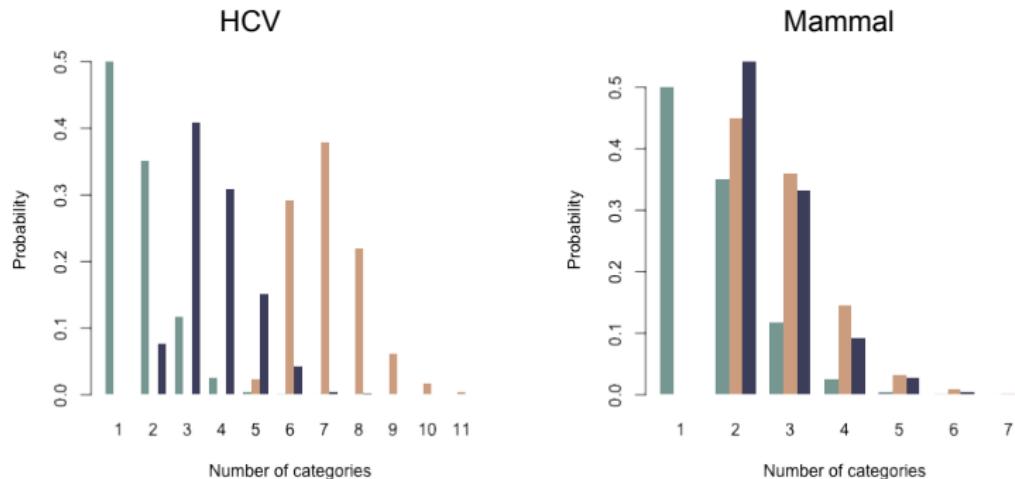
# Subst BMA



## Number of categories for substitution patterns and rates across sites (DPM1)



## Number of categories for substitution patterns and rates across sites (DPM2)



- Prior Pr(# substitution model/rate categories)
- Posterior Pr(# substitution model category)
- Posterior Pr(# rate category)

# RBS vs bModelTest vs Subst BMA

	RBS	bModelTest	Subst BMA
Subst models	6	11/31/202	4
Average over Gamma	no	yes	yes
Average over Prop Invariable	no	yes	no
Average over Sites	no	no	yes
Average over Frequencies	no	yes	somewhat
Works with tip dates	yes	yes	yes
Works without tip dates	yes	yes	not yet

# Model selection

When to do model selection:

- answers of interest are not robust for different models
- to test hypotheses (encoded by prior)
- not because the reviewers demand it

Evade it if you can

# Model selection

Measures of fit:

- Super naive: compare posteriors
  - ▶ priors are not normalised, so posteriors cannot be compared
  - ▶ **Never do this!**

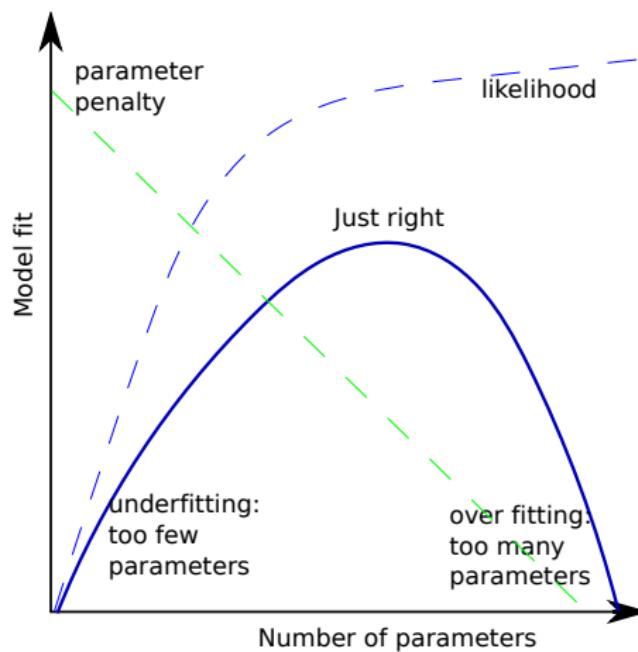
# Model selection

## Measures of fit:

- Super naive: compare posteriors
  - ▶ priors are not normalised, so posteriors cannot be compared
  - ▶ **Never do this!**
- Super naive: compare likelihoods
  - ▶ overparameterisation/overfitting cannot be detected
  - ▶ **Never do this!**

## Model selection: select model with best "fit"

Desirable model fit property 1: likelihood - parameter penalty



Desirable model fit property 2: replicability/low variance

Desirable model fit property 3: easy & cheap to calculate  
(this list is not exhaustive)

# Bayesian model selection: marginal likelihood

Posterior:

$$p(\theta|D, M) = \frac{\overbrace{\pi(\theta|M)}^{\text{prior}} \overbrace{L(D|M, \theta)}^{\text{likelihood}}}{\underbrace{p(D|M)}_{\text{marginal likelihood}}}$$

Marginal likelihood:

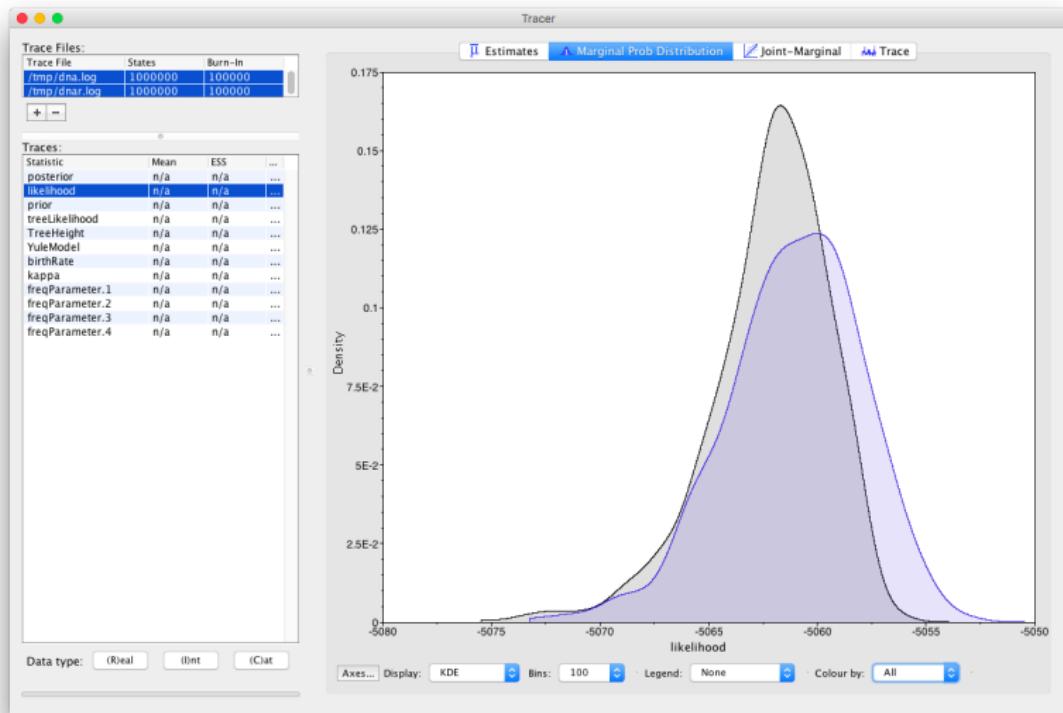
$$p(D|M) = \int_{\theta \in \Theta} \pi(\theta|M) L(D|M, \theta) d\theta$$

integrate/marginalise out  $\theta$

Bayes factor:

$$\frac{p(D|M_1)}{p(D|M_2)}$$

# Model selection: “marginal likelihood” in Tracer



This is the likelihood averaged over samples from the posterior, not prior.

## Model selection: marginal likelihood

Naive: Harmonic mean estimator (HME) of marginal likelihood

$$HME = \left( \frac{1}{n} \sum_{i=1}^n \frac{1}{L(D|\theta_i)} \right)^{-1}$$

where  $\theta_1, \dots, \theta_n$  a sample from the posterior

- Only requires a sample from the posterior
- Conveniently & quickly calculated in Tracer
- High variance estimator  $\Rightarrow$  unreliable

# Model selection: marginal likelihood

Naive: Harmonic mean estimator (HME) of marginal likelihood

$$HME = \left( \frac{1}{n} \sum_{i=1}^n \frac{1}{L(D|\theta_i)} \right)^{-1}$$

where  $\theta_1, \dots, \theta_n$  a sample from the posterior

- Only requires a sample from the posterior
- Conveniently & quickly calculated in Tracer
- High variance estimator  $\Rightarrow$  unreliable
- A post on Dr. Radford Neal's blog <http://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever>

“The total unsuitability of the harmonic mean estimator should have been apparent within an hour of its discovery.”

- **Never do this!**

# AICM model selection

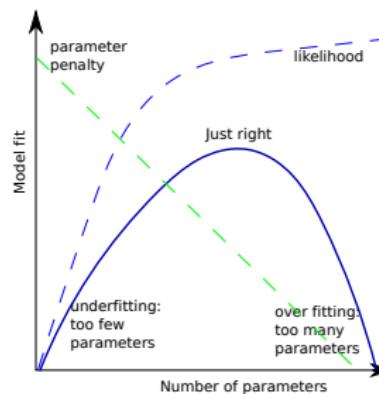
Akaike Information Criterion (AIC) by MCMC (AICM)

AIC:

$$2k - 2 \log L(D|\theta)$$

$k$  = effective number of parameters

$P(D|\theta)$  = likelihood



- Estimate  $k$  from fitting gamma distribution to MCMC sample of likelihood
- Smaller AICM is better (unlike marginal likelihood)
- Conveniently calculated in Tracer (or by model-selection package)

## Path sampling/Stepping stone theory

- Marginal likelihood:

$$p(D) = \int_{\theta} \pi(\theta) L(D|\theta) d\theta$$

hard to estimate directly.

- Define *power posterior* for some tractable reference distribution  $p_w(\theta)$

$$P_{\beta}(\theta|D) = \frac{[L(D|\theta)\pi(\theta)]^{\beta} p_w(\theta)^{1-\beta}}{c_{\beta}}$$

$P_1(\theta|D)$  is the posterior,  $c_1$  the marginal likelihood.

$P_0(\theta|D)$  is the reference distribution,  $c_0 = 1$

# Path sampling/Stepping stone theory

- Marginal likelihood:

$$p(D) = \int_{\theta} \pi(\theta) L(D|\theta) d\theta$$

hard to estimate directly.

- Define *power posterior* for some tractable reference distribution  $p_w(\theta)$

$$P_{\beta}(\theta|D) = \frac{[L(D|\theta)\pi(\theta)]^{\beta} p_w(\theta)^{1-\beta}}{c_{\beta}}$$

$P_1(\theta|D)$  is the posterior,  $c_1$  the marginal likelihood.

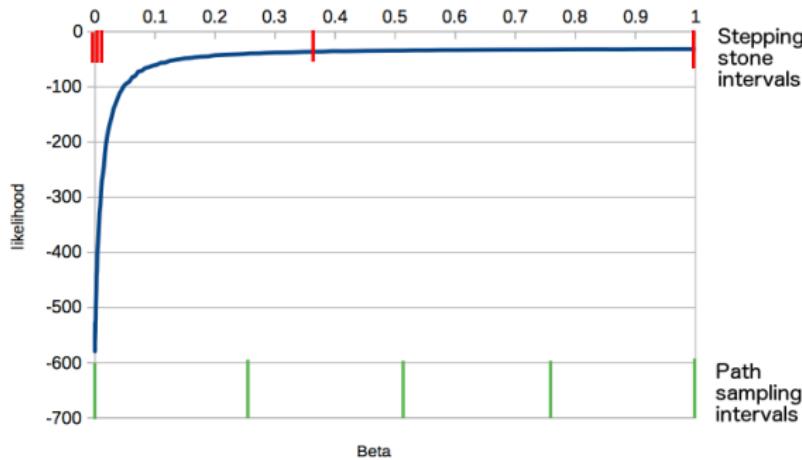
$P_0(\theta|D)$  is the reference distribution,  $c_0 = 1$

- $$\frac{c_{\beta_k}}{c_{\beta_{k-1}}} \approx \frac{1}{n} \sum_{i=1}^n p(D|\theta_{k-1,i})^{\beta_k - \beta_{k-1}}$$

- $$P(D|M) = \frac{c_1}{c_0} = \frac{c_1}{c_{0.3}} \frac{c_{0.3}}{c_{0.1}} \frac{c_{0.1}}{c_{0.01}} \frac{c_{0.01}}{c_0} = \frac{c_1}{\cancel{c_{0.3}}} \frac{\cancel{c_{0.3}}}{\cancel{c_{0.1}}} \frac{\cancel{c_{0.1}}}{\cancel{c_{0.01}}} \frac{\cancel{c_{0.01}}}{c_0}$$

## Model selection: *Stepping stone vs path sampling*

Both use prior as reference distribution ( $p_w(\theta) = \pi(\theta)$ )



Stepping stone uses different intervals (set of  $\beta$  values) from path sampling, but otherwise the same

# Generalised Stepping Stone

Use ‘working distribution’ for  $p_w(\theta)$  based on posterior sample

- For parameters, use empirical distribution based on kernel estimators
- For trees, we need a distribution on topology and branch lengths
- Promises lower variance estimates
- Currently tedious to set up, since working distribution needs to be specified

Holder et al, Bayesian phylogenetics, 2014, Beale et al, Sys Bio, 2016

# Pairwise Stepping Stone

Use posterior of  $M_2$  for  $p_w(\theta)$ .

- Calculates Bayes factor between  $M_1$  and  $M_2$  directly
- + requires fewer steps for accurate estimate
- - does not result in Bayes factors directly

Baele et al, BMC Bioinformatics, 2013

# Path sampling/Stepping stone in practice

Number of steps:

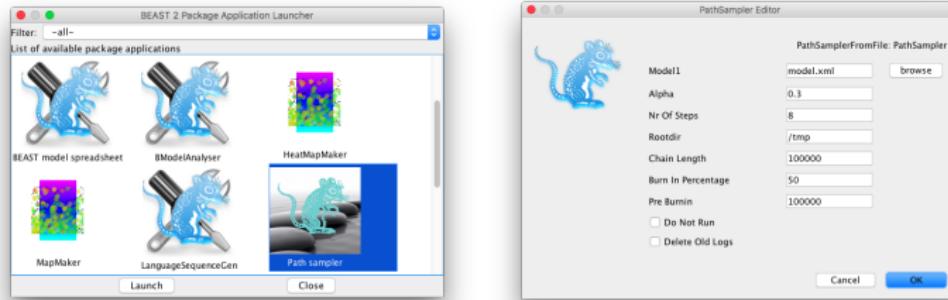
- Start with small nr of steps, say 16 and estimate ML
- increase nr of steps, estimate ML
- continue till ML estimate does not decrease any more

Chain length per step/ESS:

- total chain length at least as long as for posterior
- not all ESSs have to be 200 (errors cancel out)
- run different runs to get impression of variance of estimate
- use logcombiner to combine logs, for final estimate

# Path sampling/Stepping stone in practice

Set up through XML or GUI



Set up through CLI:

- to list BEAST apps:

```
/path/to/applauncher -list
```

- To show PathSampler options:

```
/path/to/applauncher PathSampler -help
```

- To set up PathSampler analysis:

```
/path/to/applauncher PathSampler -nrOfSteps 64 -rootdir  
dir/withs/steps -burnInPercentage 50 -model beast.xml
```

Creates subdirectory structure, one for each step containing all log files.

# Path sampling/Stepping stone in practice

To set up on a HPC cluster

- Set up locally, using 'doNotRun' flag = true
- Move steps to cluster, and run steps in parallel there
- Estimate ML using PathSampleAnalyser

```
/path/to/appauncher PathSampleAnalyser -nrOfSteps 64  
-rootdir dir/withs/steps -burnInPercentage 50
```

# Path sampling/Stepping stone in practice

## Trouble shooting

- ESS too small for a step: resume runs for that step
- Infinite likelihoods caused by numeric instability: improper priors – use proper priors instead
- -Infinite likelihoods: priors too wide – narrow priors
- Inspect log files in step directory to see which parameter escapes, so which prior to adjust

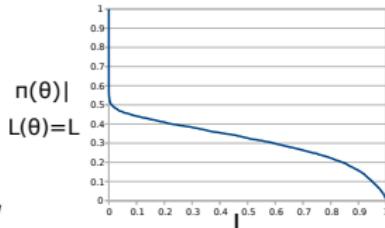
# Nested Sampling Theory

$$\mathcal{Z} = \int_{\theta} \pi(\theta) L(\theta) d\theta$$

# Nested Sampling Theory

$$\mathcal{Z} = \int_{\theta} \pi(\theta) L(\theta) d\theta$$

$$= \int_{L=0}^{\infty} L \left( \int_{\theta, L(\theta)=L} \pi(\theta) d\theta \right) dL$$

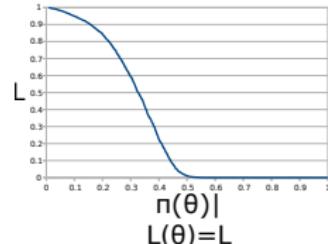
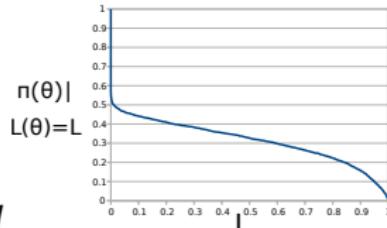


# Nested Sampling Theory

$$\mathcal{Z} = \int_{\theta} \pi(\theta) L(\theta) d\theta$$

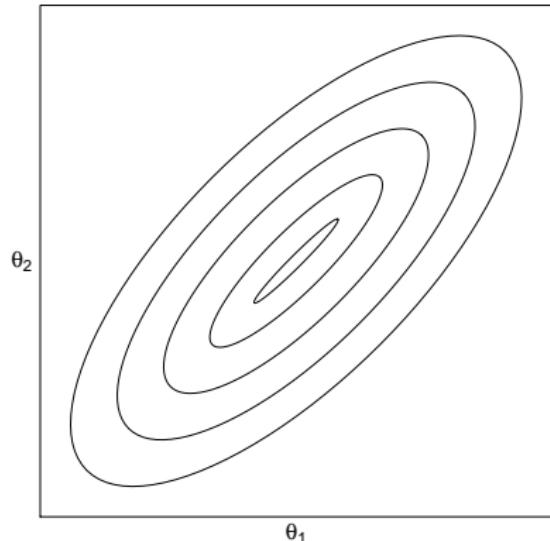
$$= \int_{L=0}^{\infty} L \left( \int_{\theta, L(\theta)=L} \pi(\theta) d\theta \right) dL$$

$$= \int_{X=0}^1 \mathcal{L}(X) dX$$

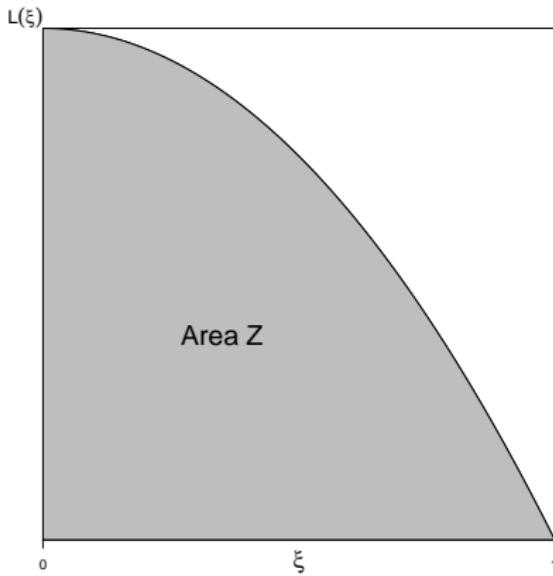


where  $\mathcal{L}(X)$  inverse likelihood

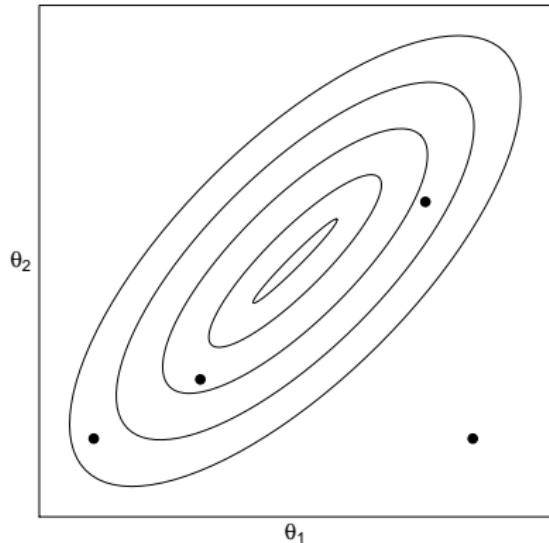
# Nested sampling



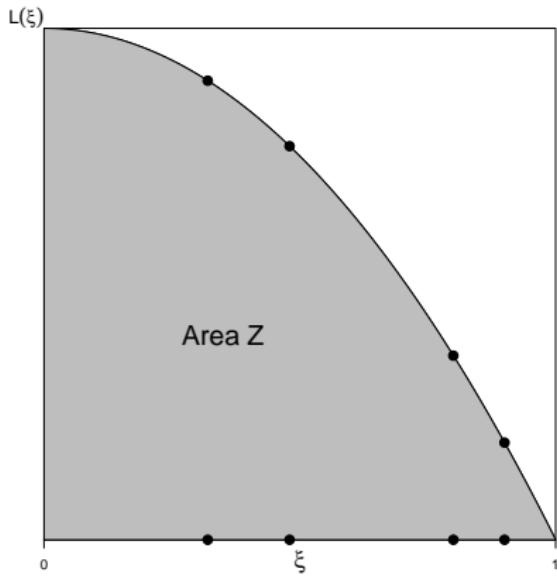
images courtesy Patricio Russell



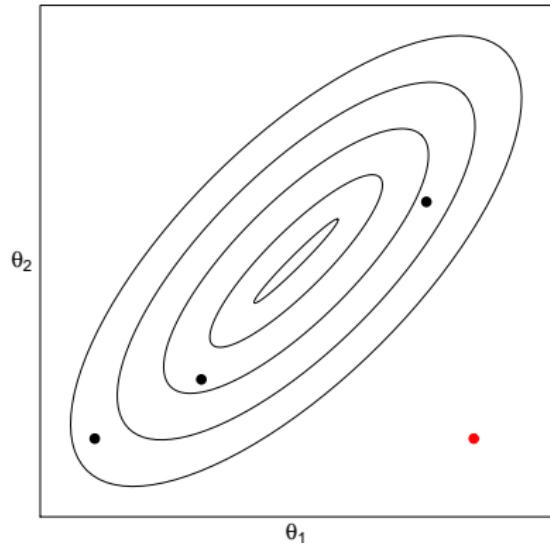
# Nested sampling



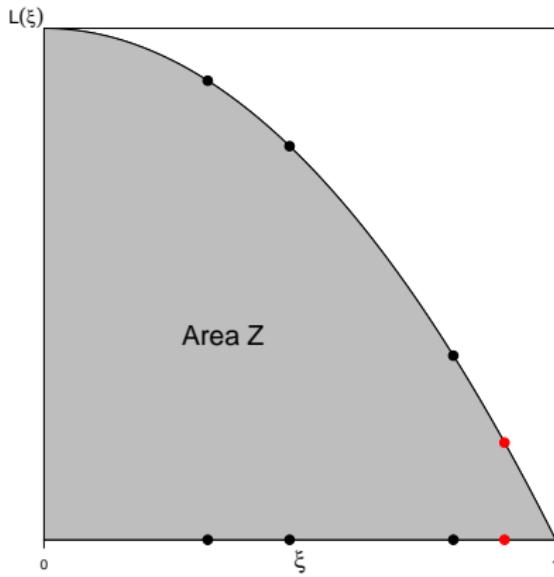
images courtesy Patricio Russell



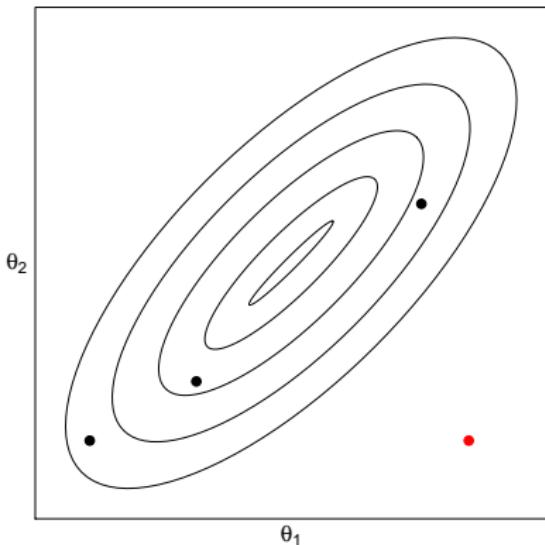
# Nested sampling



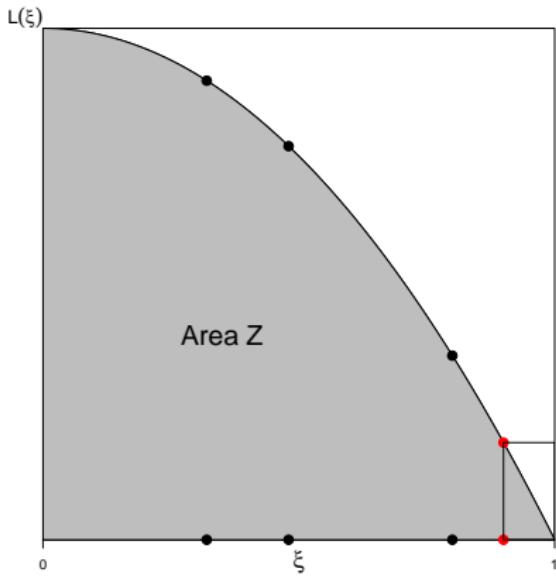
images courtesy Patricio Russell



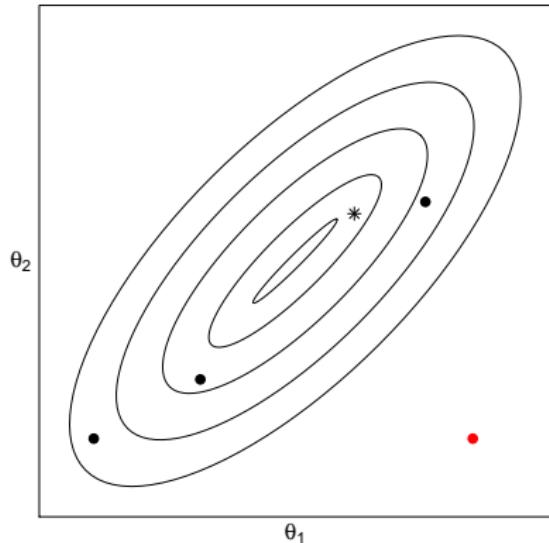
# Nested sampling



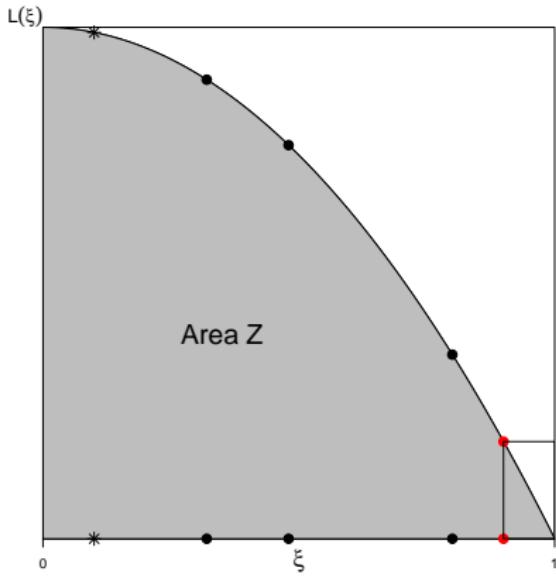
images courtesy Patricio Russell



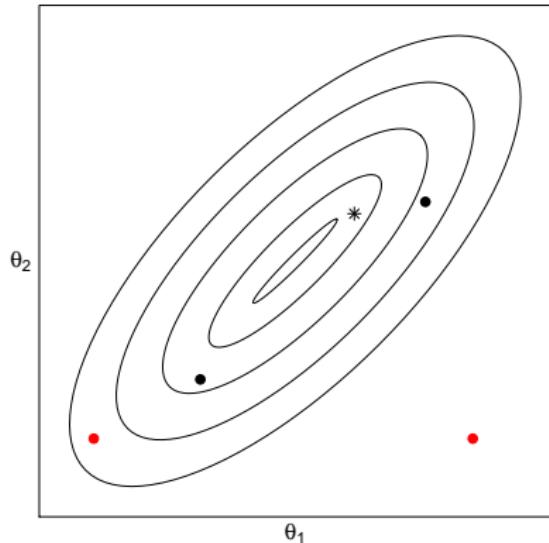
# Nested sampling



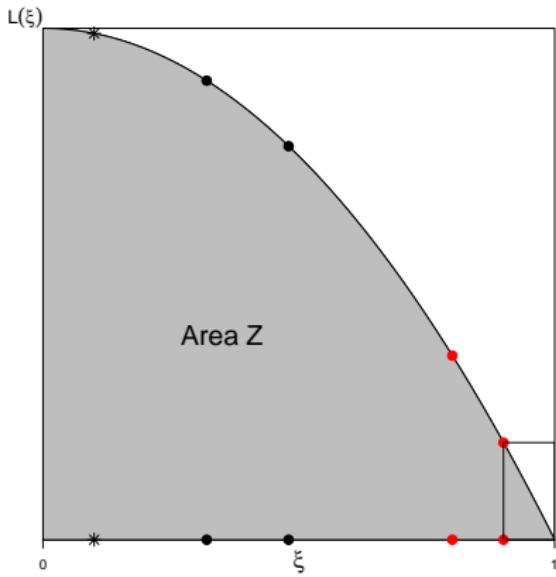
images courtesy Patricio Russell



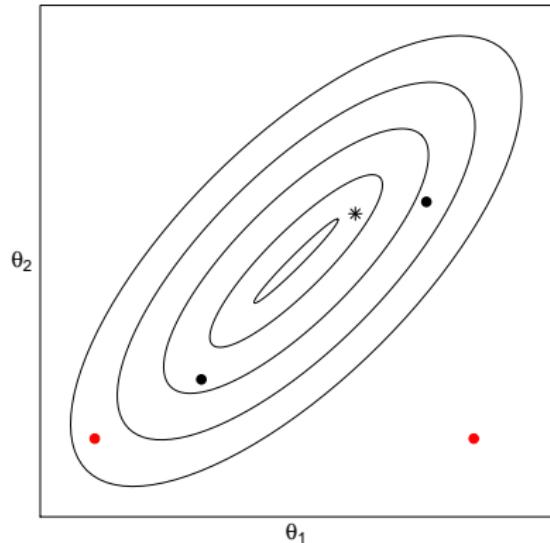
# Nested sampling



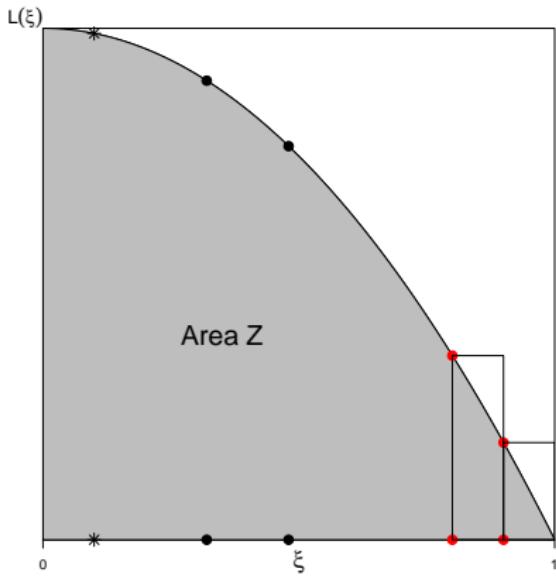
images courtesy Patricio Russell



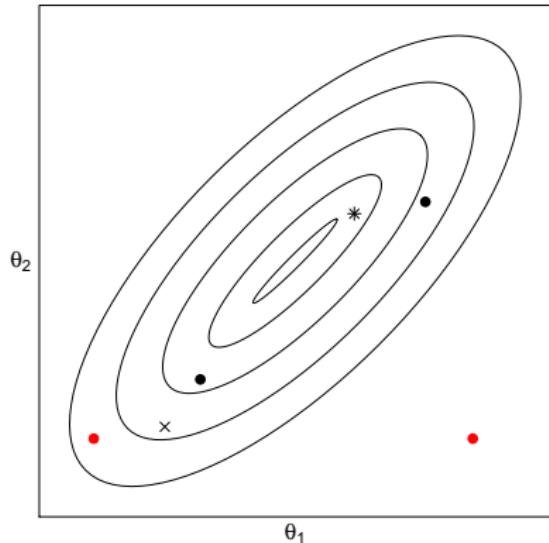
# Nested sampling



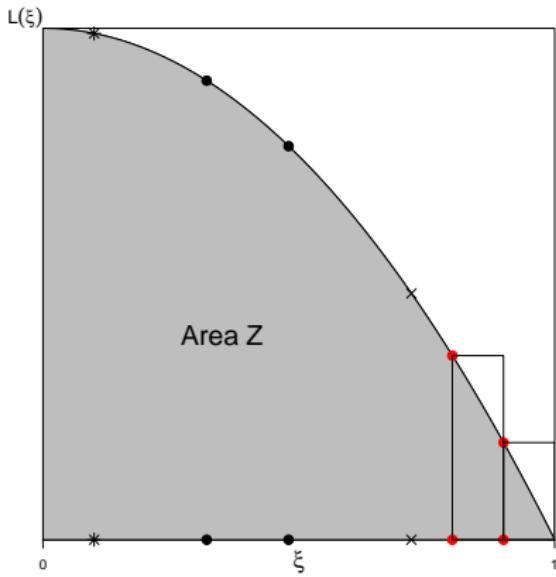
images courtesy Patricio Russell



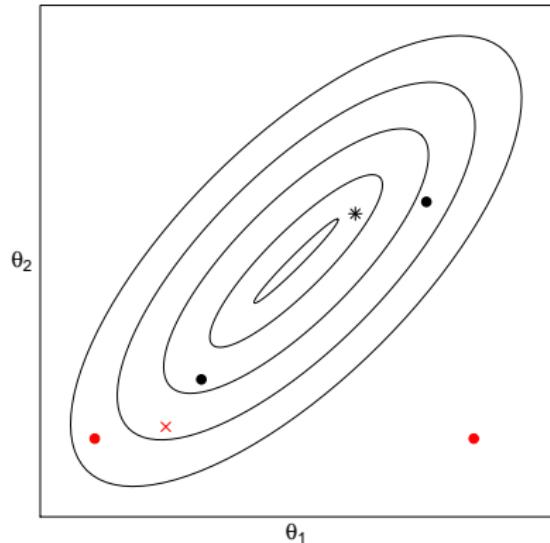
# Nested sampling



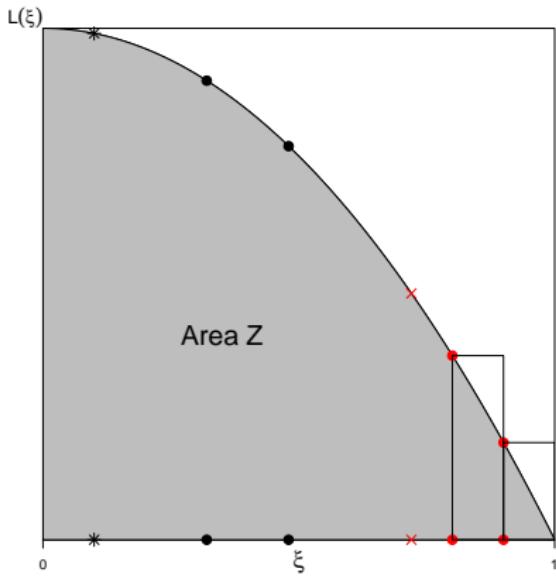
images courtesy Patricio Russell



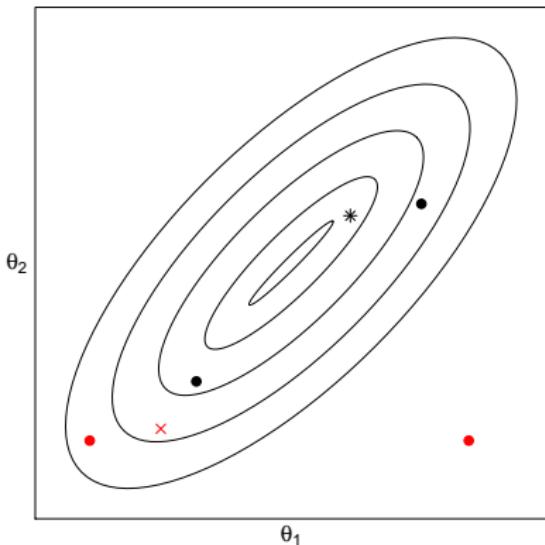
# Nested sampling



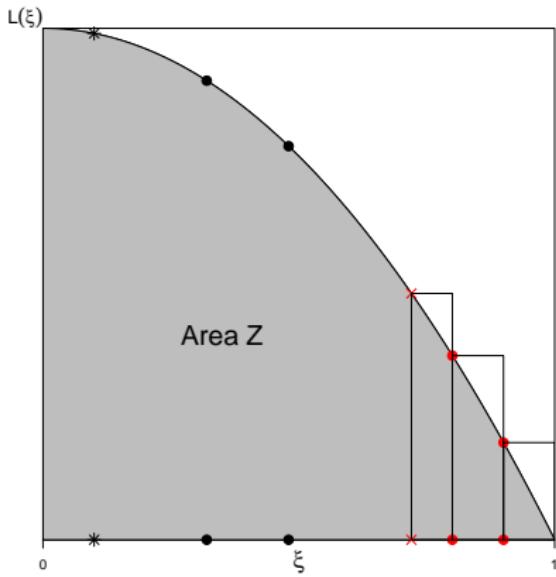
images courtesy Patricio Russell



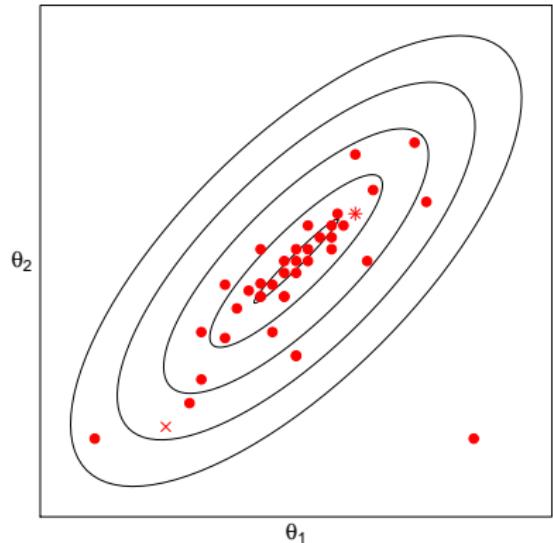
# Nested sampling



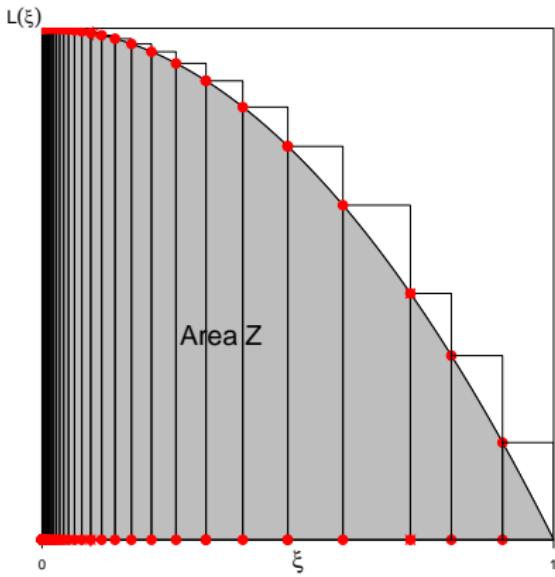
images courtesy Patricio Russell



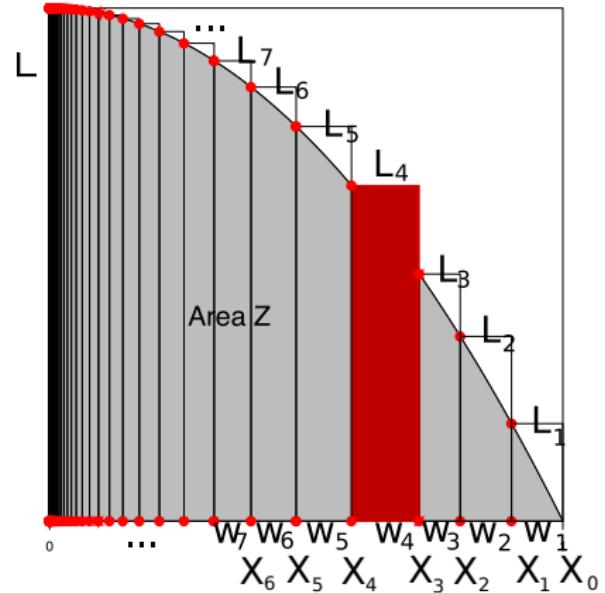
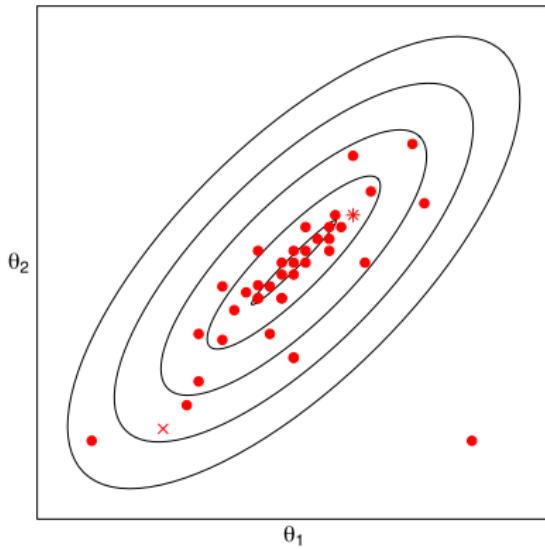
# Nested sampling



images courtesy Patricio Russell



Area under curve is  $ML = \mathcal{Z}$ , with  $N$  active points



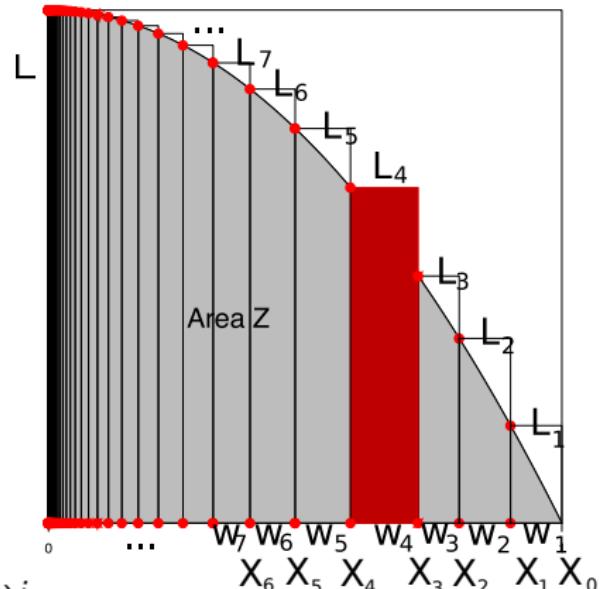
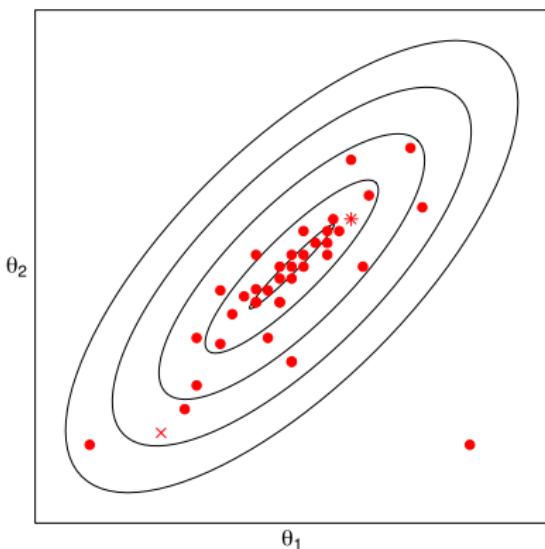
$$X_0 = 1$$

$X_i$  = proportion of prior mass with likelihood at least  $L_i$

$$w_i = X_i - X_{i-1}$$

$$\mathcal{Z} \approx \sum_i w_i L_i$$

## Defining $X_i$ with $N$ active points

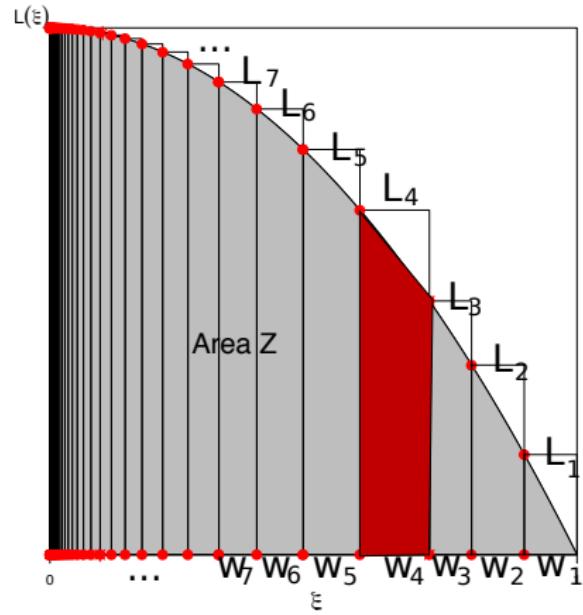
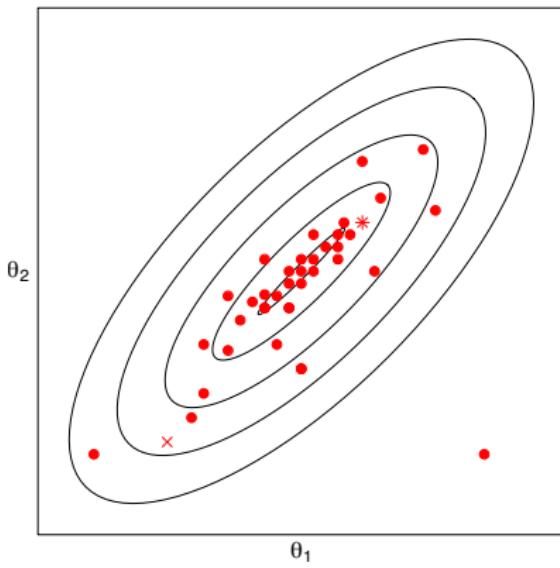


1. Arithmetic mean:  $X_i = (\frac{N}{N+1})^i$
2. Geometric mean:  $X_i = e^{-\frac{i}{N}} <=$  fast, most popular
3. Stochastic:  $X_i = \beta(1, N)X_{i-1}$ ,  $X_0 = 1 <=$  allows SD estimate

$$w_i = X_i - X_{i-1}$$

$$\mathcal{Z} \approx \sum_i w_i L_i$$

# Use trapezium rule for more accurate ML estimate



$$\mathcal{Z} = \sum \dots$$

## Nested sampling with $N$ active points

Assign weights to 'saved points'

$$E\{w_i\} = e^{-(i-1)/N} - e^{-i/N}$$

Estimate of marginal likelihood

$$\mathcal{Z} = \sum_i w_i L_i$$

## Nested sampling with $N$ active points

Assign weights to 'saved points'

$$E\{w_i\} = e^{-(i-1)/N} - e^{-i/N}$$

Estimate of marginal likelihood

$$\mathcal{Z} = \sum_i w_i L_i$$

Estimate of standard deviation of marginal likelihood

$$sd(\log \mathcal{Z}) = \sqrt{\frac{H}{N}}$$

where the information  $H \approx \sum_i w_i \frac{L_i}{\mathcal{Z}} \log \frac{L_i}{\mathcal{Z}}$

## Nested sampling with $N$ active points

Assign weights to 'saved points'

$$E\{w_i\} = e^{-(i-1)/N} - e^{-i/N}$$

Estimate of marginal likelihood

$$\mathcal{Z} = \sum_i w_i L_i$$

Estimate of standard deviation of marginal likelihood

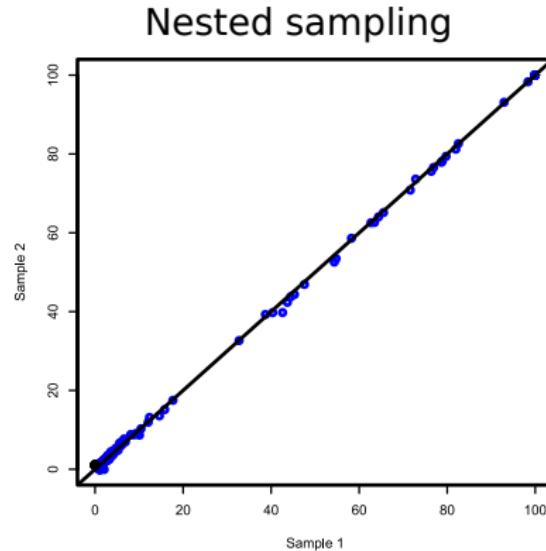
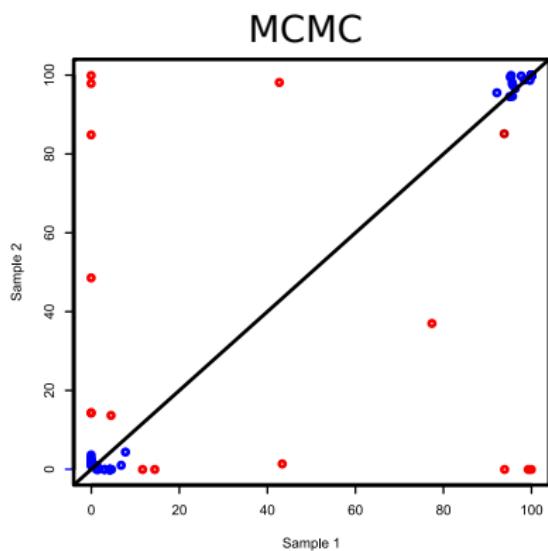
$$sd(\log \mathcal{Z}) = \sqrt{\frac{H}{N}}$$

where the information  $H \approx \sum_i w_i \frac{L_i}{\mathcal{Z}} \log \frac{L_i}{\mathcal{Z}}$

Sample from posterior by sampling saved points according to weights  $\frac{w_i L_i}{\mathcal{Z}}$

# DS1: where MCMC fails

- DS1 data set with tree islands
- MCMC has trouble moving between islands
- Consequently MCMC/stepping stone/path sampling fails



## NS in practice: Setting up an analysis

Requires NS package

Set up analysis in BEAUTi, then edit XML and replace

```
<run id="mcmc" spec="MCMC" chainLength="100000000">
```

with

```
<run id="mcmc" spec="beast.gss.NS" chainLength="20000"  
    particleCount="1" subChainLength="5000" epsilon="1e-12">
```

More info:

<https://github.com/BEAST2-Dev/nested-sampling/wiki/How-to-use-NS>

## NS in practice: Are we there yet?

- Run multiple times (like MCMC).
- Check  $\mathcal{Z}$  estimates are compatible

$$|\log \mathcal{Z}_1 - \log \mathcal{Z}_2| \leq 2\sqrt{(SD_1^2 + SD_2^2)}$$

- If not, run with longer sub chain length
- Check  $\mathcal{Z}$  estimates are compatible with shorter runs/not systematically biased in multiple runs

Note, nested sampling under estimates  $\mathcal{Z}$ . Longer sub chain length results in lower  $\mathcal{Z}$  if not converged yet.

## NS in practice: Model selection with Nested sampling

Given  $M_1$  and  $M_2$ : say HKY vs GTR

- ① estimate log marginal likelihoods  $\mathcal{Z}_1$  for  $M_1$  and  $\mathcal{Z}_2$  for  $M_2$ 
  - NS provides standard deviations  $SD_1$  and  $SD_2$  for **log** marginal likelihoods
- ② if  $|\log \mathcal{Z}_1 - \log \mathcal{Z}_2| \geq 2\sqrt{SD_1^2 + SD_2^2}$  calculate Bayes factor  
 $BF = \log \mathcal{Z}_1 - \log \mathcal{Z}_2$ . Done!
- ③ else if  $\sqrt{SD_1^2 + SD_2^2} < 3(?)$  then  $M_1$  and  $M_2$  cannot be distinguished.  
Done!
- ④ else, run NS with more particles. How many? Use  $SD = \sqrt{\frac{H}{N}}$  so  
 $N = SD^2/H$  for desired  $SD$  ( $H$  from NS run) and goto (2)

## NS in practice: Pitfalls

- Subchain length too short – run with different length, compare whether estimates differ
- Epsilon too large, causing early stopping, underestimate of  $\mathcal{Z}$
- ...

## NS in practice: Parallel implementation

- Maintaining shared pool to selected starting point => behaves like single thread  $N$  particle
- Runtime scales linear with nr or particles  $N$
- $N$  single particle runs can be combined => embarrassingly parallel
- Little communication required, so can be forked out over different CPUs

# Stepping Stone and Nested Sampling work on any model

provided the prior is proper

## Model selection summary

Stepping Stone and Nested Sampling work on any model

**provided the prior is proper**

Improper priors do not integrate to one, e.g.,  $1/X$ , uniform( $0, \infty$ )

# Model selection summary

- Naive: Harmonic mean estimator (HME) of marginal likelihood
  - ▶ High variance estimator  $\Rightarrow$  unreliable
  - ▶ Never do this!
- AICM: Akaike information criterion for MCMC
  - ▶ Computational convenient
  - ▶ High variance estimator  $\Rightarrow$  unreliable (but better than HME)
  - ▶ only use when stepping stone/nested sampling is not feasible
- Path sampling/Stepping stone:
  - ▶ computationally expensive
  - ▶ most stable marginal likelihood estimation we got (so far)
  - ▶ use this if you can
- Nested sampling:
  - ▶ Provides estimate of ML + its variance
  - ▶ Computation (inverse) proportional to accuracy of estimate
  - ▶ Can choose accuracy as desired

# Model comparison/Bayesian hypothesis testing

- Through model selection:
  - ▶ compare Bayes factors based on ML estimates
- Through model averaging: post-hoc analysis
  - ▶ compare Bayes factors based on empirical estimates from prior and posterior samples

Bayes factor:

$$\frac{p(D|M_1)}{p(D|M_2)} \text{ estimated by } \frac{\frac{\text{empirical posterior}(M_1)}{\text{empirical prior}(M_1)}}{\frac{\text{empirical posterior}(M_2)}{\text{empirical prior}(M_2)}}$$

Obtain sample from prior for  $M_1$  and  $M_2$ .

Obtain sample from posterior for  $M_1$  and  $M_2$ .

# Bayes Factors

$BF$ range	$\ln(BF)$ range			$\log_{10}(BF)$ range			Interpretation
1 – 3	0	–	1.1	0	–	0.5	hardly worth mentioning
3 – 20	1.1	–	3	0.5	–	1.3	positive support
20 – 150	3	–	5	1.3	–	2.2	strong support
> 150	>	5		>	2.2		overwhelming support

Kass & Raftery, JASA, 1995

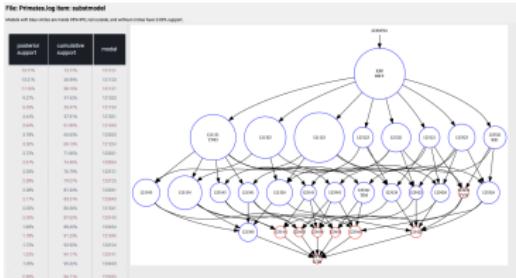
## Model comparison of topologies

- Clade support for two alternative clades –  $M_1$ : clade  $A$  is present  $M_2$ : clade  $B$  is present
- From sample from prior:  $P(M_1) = 0.3$   $P(M_2) = 0.4$
- From sample from posterior:  $P(D|M_1) = 0.6$   $P(D|M_2) = 0.1$
- Bayes factor

$$\frac{p(D|M_1)}{p(D|M_2)} = \frac{\frac{posterior(M_1)}{prior(M_1)}}{\frac{posterior(M_2)}{prior(M_2)}} = \frac{\frac{0.6}{0.3}}{\frac{0.1}{0.4}} = 8$$

- Positive support for  $M_1$  for clade  $A$  vs clade  $B$

# Model comparison of substitution models



- bModelTest –  $M_1$ : HKY vs  $M_2$ : GTR
  - From sample from prior:  $P(M_1) = \frac{1}{31}$   $P(M_2) = \frac{1}{31}$
  - From sample from posterior:  $P(D|M_1) = 0.1377$   $P(D|M_2) = 0.006$
  - Bayes factor

$$\frac{p(D|M_1)}{p(D|M_2)} = \frac{\frac{posterior(M_1)}{prior(M_1)}}{\frac{posterior(M_2)}{prior(M_2)}} = \frac{0.1377/\frac{1}{31}}{0.006/\frac{1}{31}} = 22.95$$

- Strong support for  $M_1$ : HKY

# Model comparison of root age

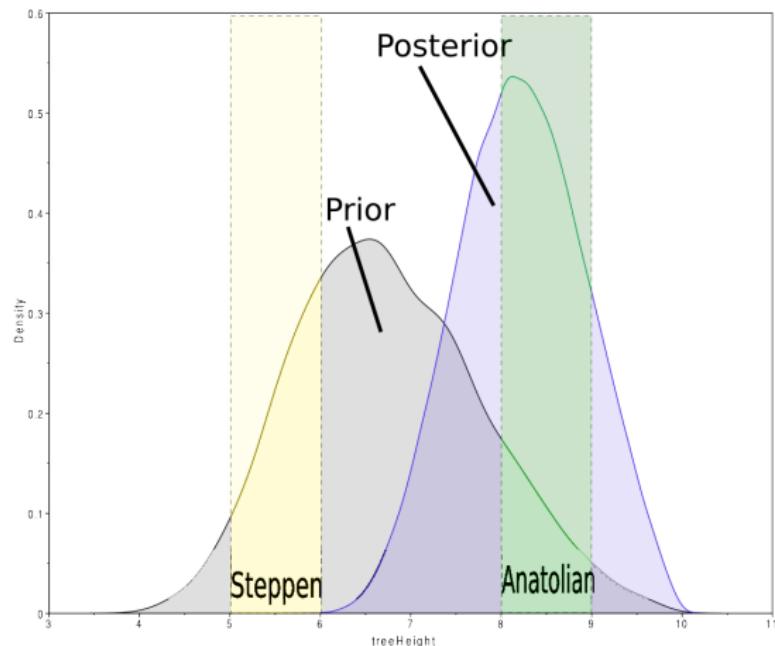
Origin of Indo-European: Two competing theories



Steppen 5000 – 6000BP, Anatolian 8000 – 9000BP

# Model comparison of root age

Origin of Indo-European: Two competing theories

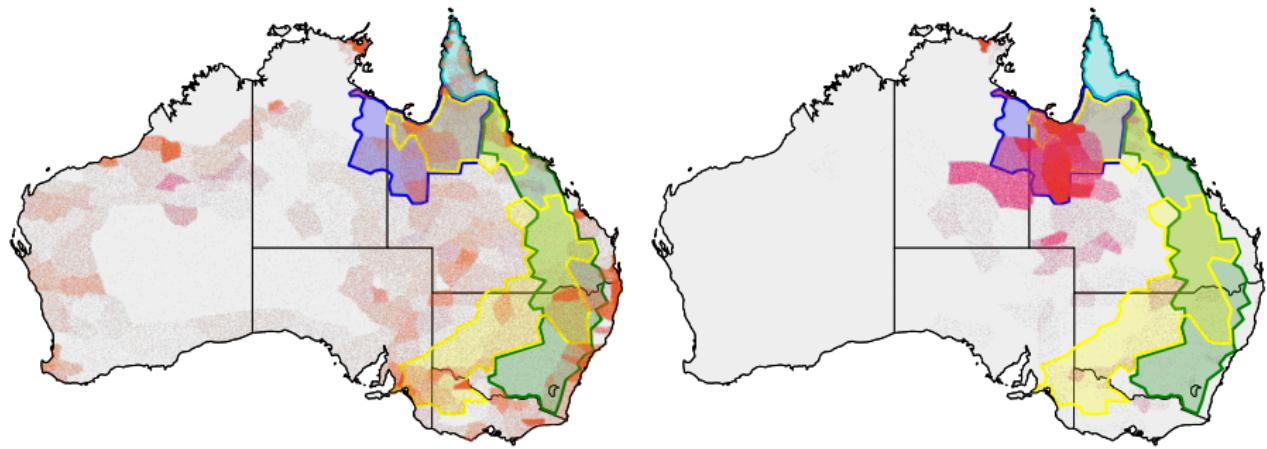


Root height Prior = 6.8 [4.9, 8.8] Posterior 8.2 [6.9, 9.6]

Bayes Factor  $\gg 100$  in favour of Anatolian hypothesis

# Model comparison of root location

Root location: Pama Nyungang  $M_1$   $M_2$   $M_3$   $M_4$



Bayes Factors:

	$M_1$	$M_2$	$M_3$	$M_4$
$M_1$	—	6.22	76.66	486.34
$M_2$	0.16	—	12.33	78.23
$M_3$	0.01	0.08	—	6.34
$M_4$	0.00	0.01	0.16	—

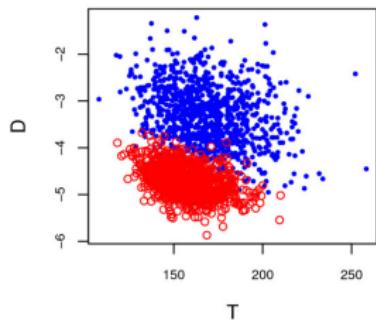
# Model adequacy: Is this model adequate for my data?

Posterior predictive simulation:

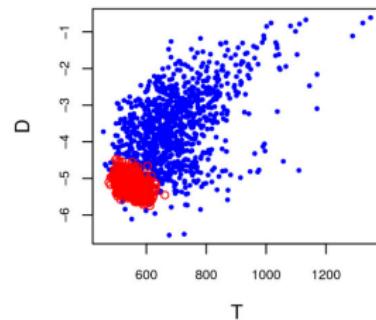
- Sample from posterior
- Generate data for each posterior sample
- Calculate summary statistics on generated data and actual data
  - ▶ frequency of nucleotides (for different substitution models)
- Use loss function to compare statistics
  - ▶ Deviance loss function  $L: \sum_{x \in \{A, C, G, T\}} \log\left(\frac{\pi_x}{\hat{\pi}_x}\right)$
  - ▶ Criterion  $E\{L\} + L$
- Automates goldilocks zone:
  - ▶ If model is too simple, goodness of fit of mean statistic will be low
  - ▶ If model is too complex, variability of statistic will be high

# Model adequacy: test of neutrality

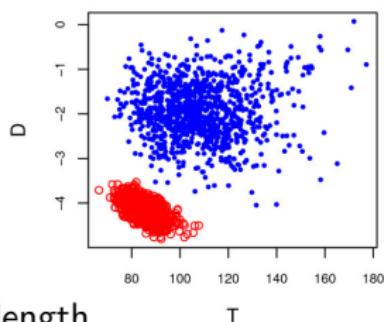
Dengue 4



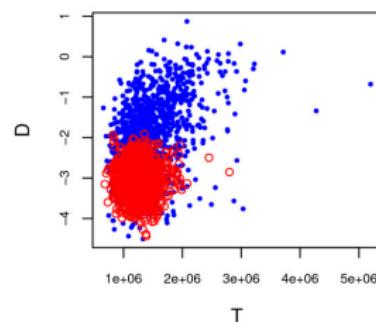
HRSV



Influenza A



Brown Bear



T = total tree length

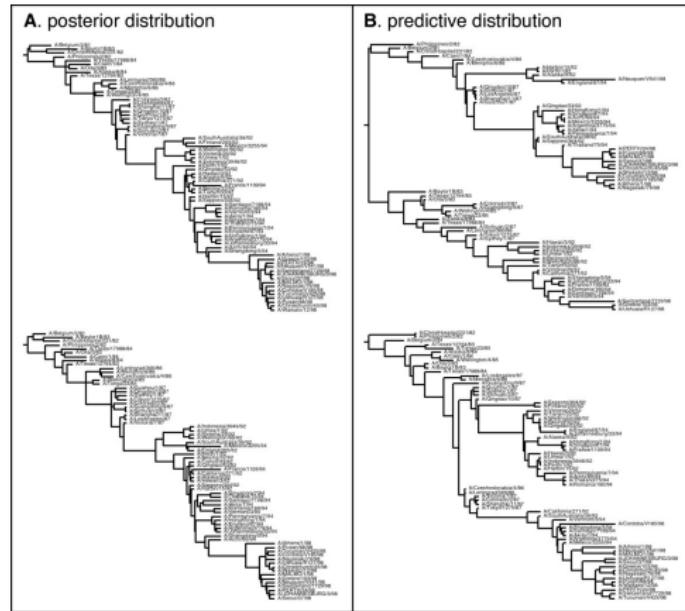
T

D = Normalised difference between external branch lengths and total tree length.

Drummond & Suchard, BMC Genetics, 2008

# Model adequacy: test of neutrality

## Tree shape human influenza A virus



Note the shorter tree length and absence of deep splits in the posterior trees.

Drummond & Suchard, BMC Genetics, 2008

# Summary model comparison, selection & averaging

Questions?

Thanks: Patricio Russell, Jessie Wu for slides