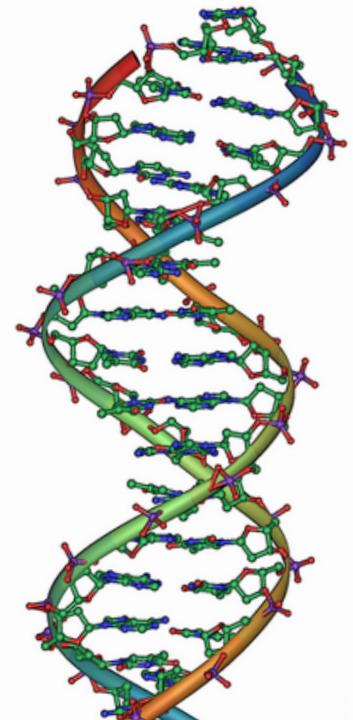
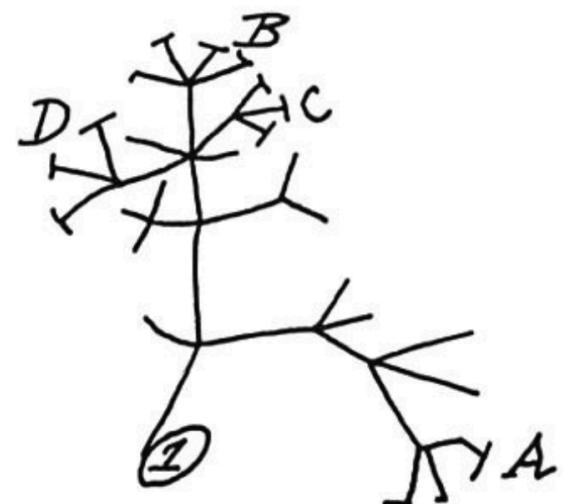


Practical Phylogenetic Inference

Jeremy M. Brown
Dept. of Biological Sciences
Louisiana State University

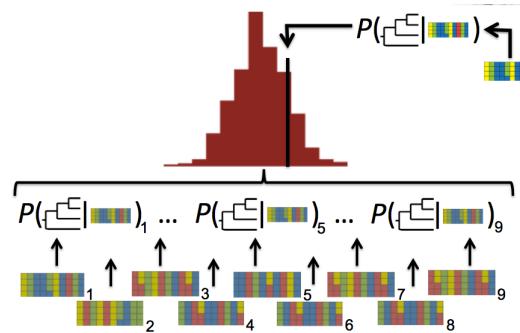


$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta$$

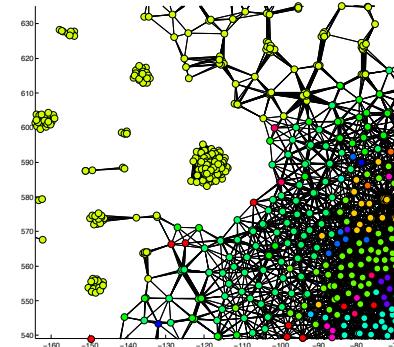
@jembrown
#bodega14

A Bit About My Interests

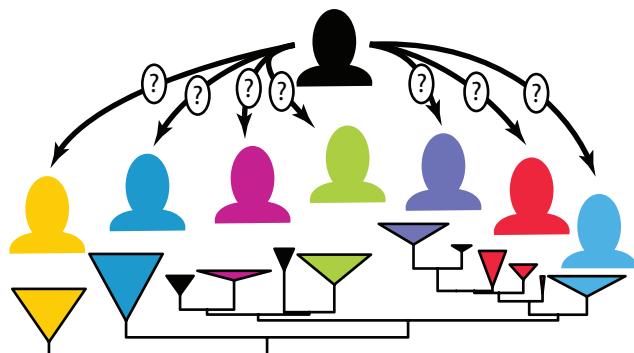
Assessing Model Fit with Posterior Prediction



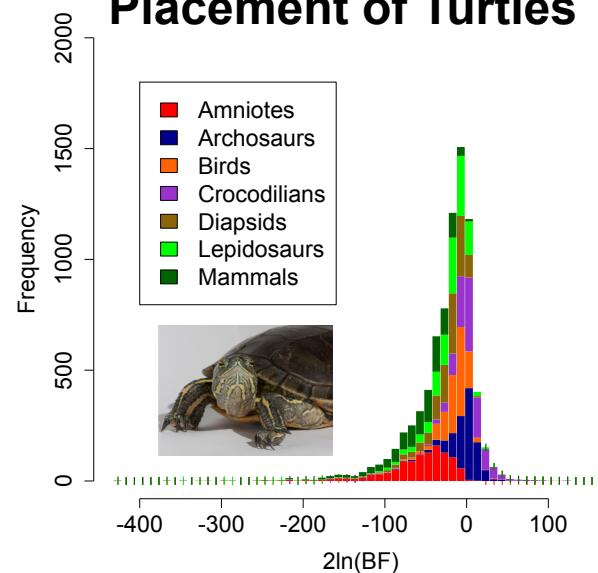
Summarizing Phylogenetic Information Using Networks



Phylogenetic Forensics of HIV Transmission

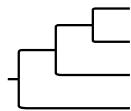


Phylogenomic Information and the Placement of Turtles



The Players

θ = Sequence Evolution Model Parameters

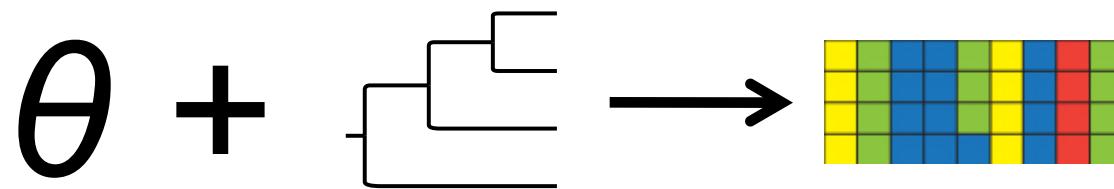


= Tree Topology and Branch Lengths

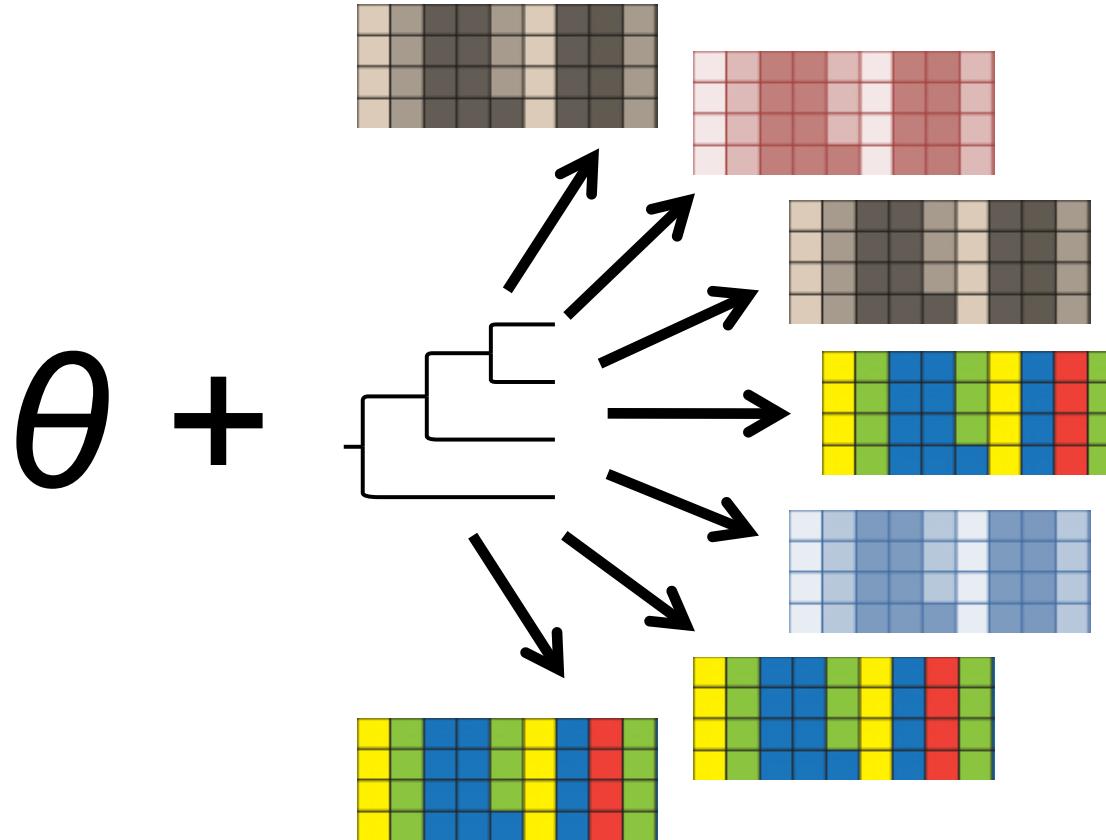


= Sequence Alignment

Simulation View of Models



Simulation View of Models



How frequently is some data observed when datasets are repeatedly generated with a particular tree and set of model parameters?

The Likelihood Function

$$P(\text{█} | \text{└─┐}, \theta)$$

Read as “the probability of the sequence data given a tree and a set of model parameter values”.

The quantity by which the data provide information.

Compares how well different trees and models predict the observed data.

The Likelihood Function

$$P(\text{█} | \text{└─┐}, \theta)$$

Read as “the probability of the sequence data given a tree and a set of model parameter values”.

The quantity by which the data provide information.

Compares how well different trees and models predict the observed data or as a “measure of surprise”.

NOT the same as $P(\text{└─┐} | \text{█})$

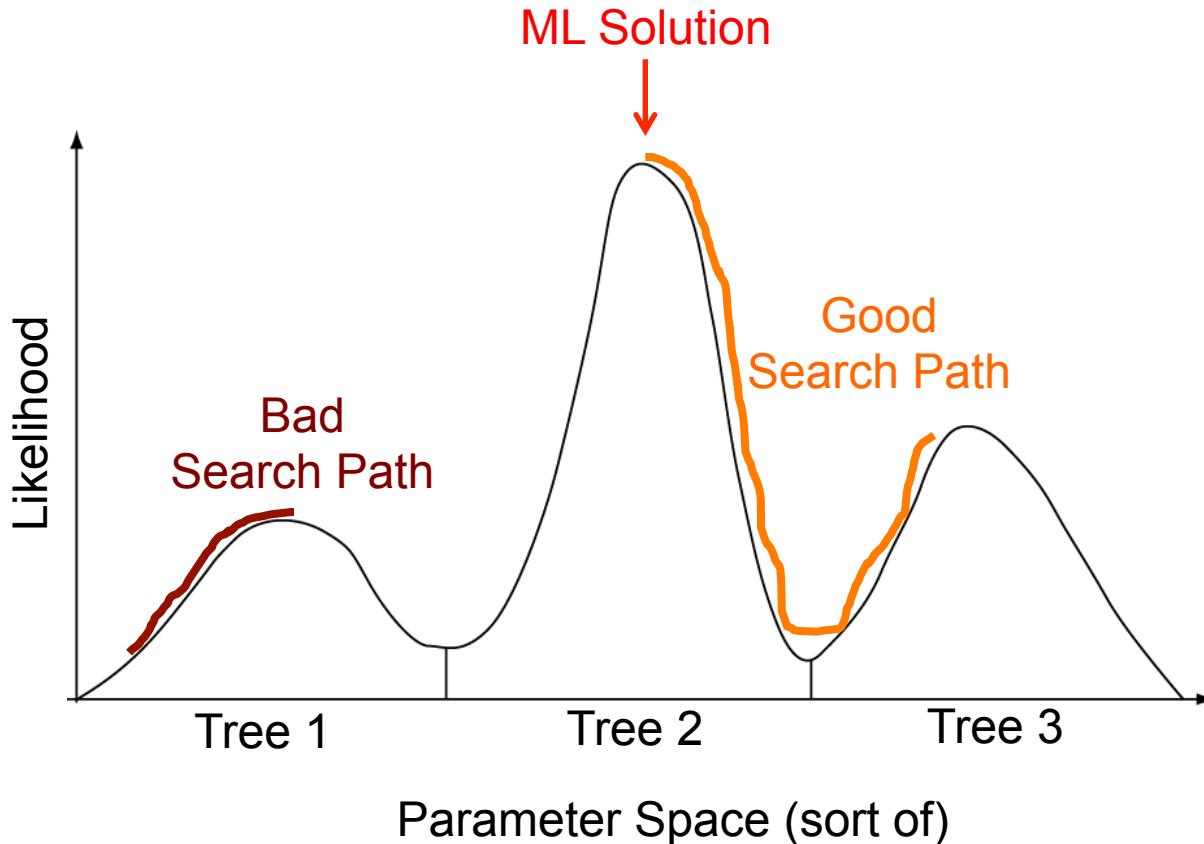
Maximum Likelihood

$$P(\text{grid} | \text{tree}, \theta)$$

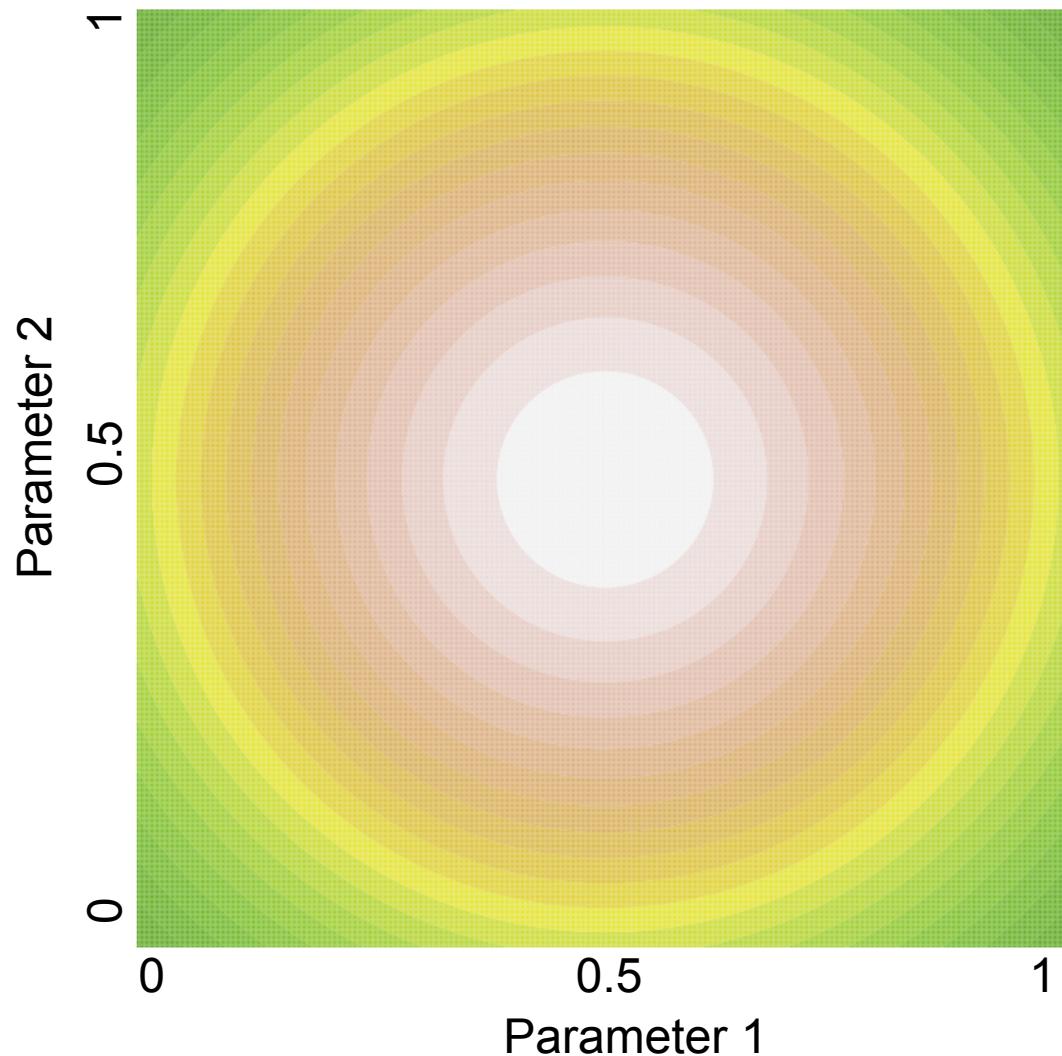
What tree and parameter values give the highest likelihood?

ML scores are just relative, so alone they don't tell us how confident we are in this solution, just that this is the preferred solution.

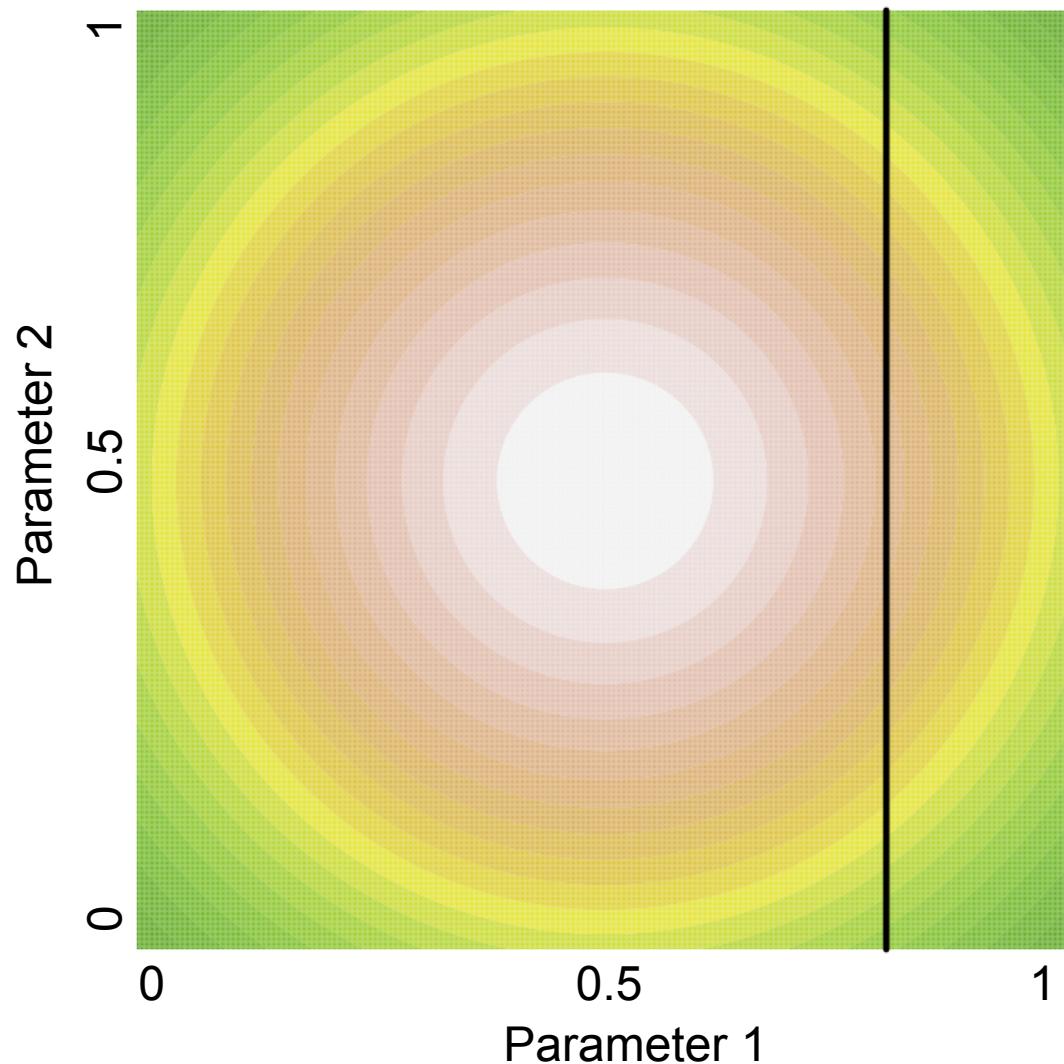
Maximum Likelihood



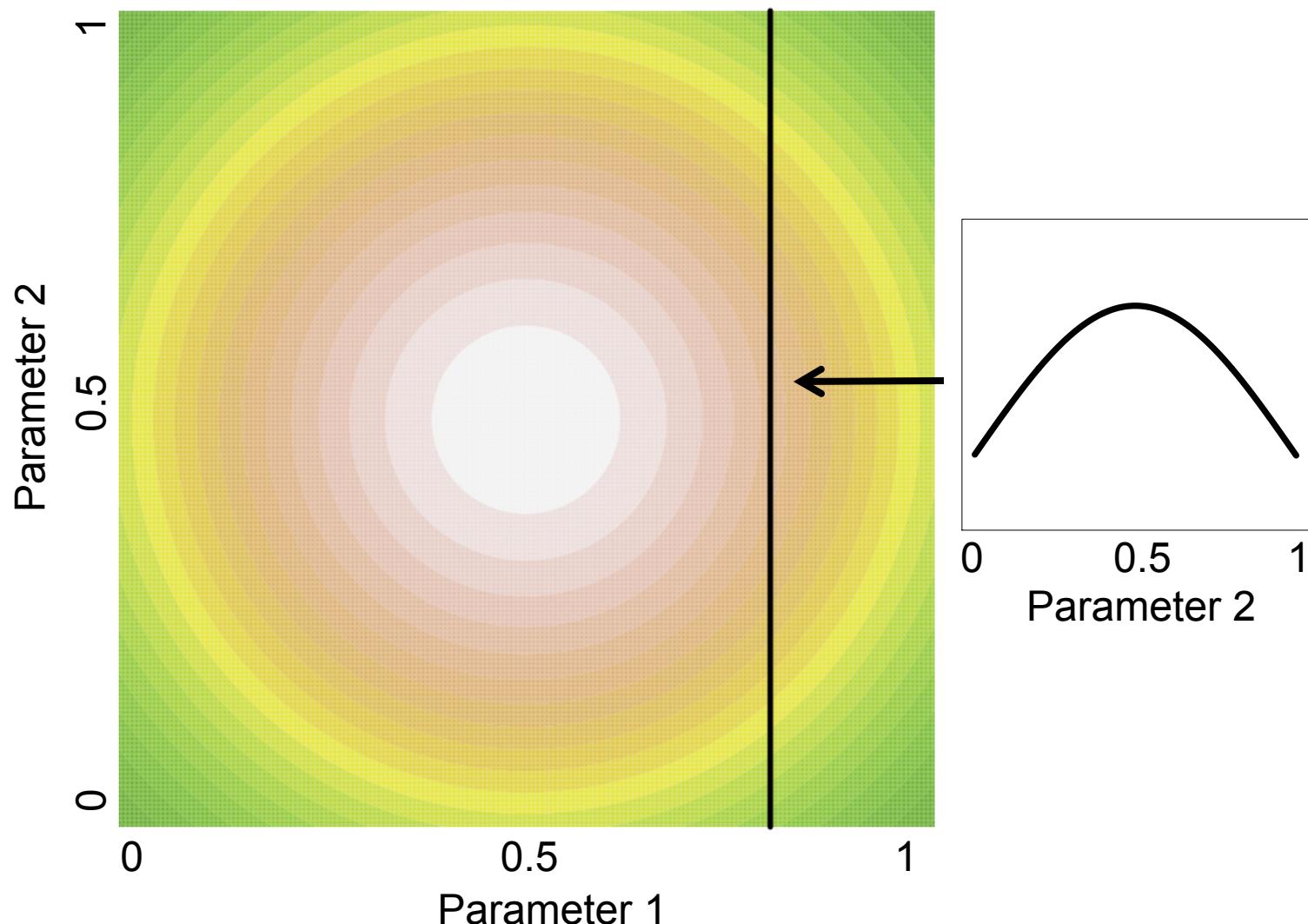
More Parameters = Better ML Score



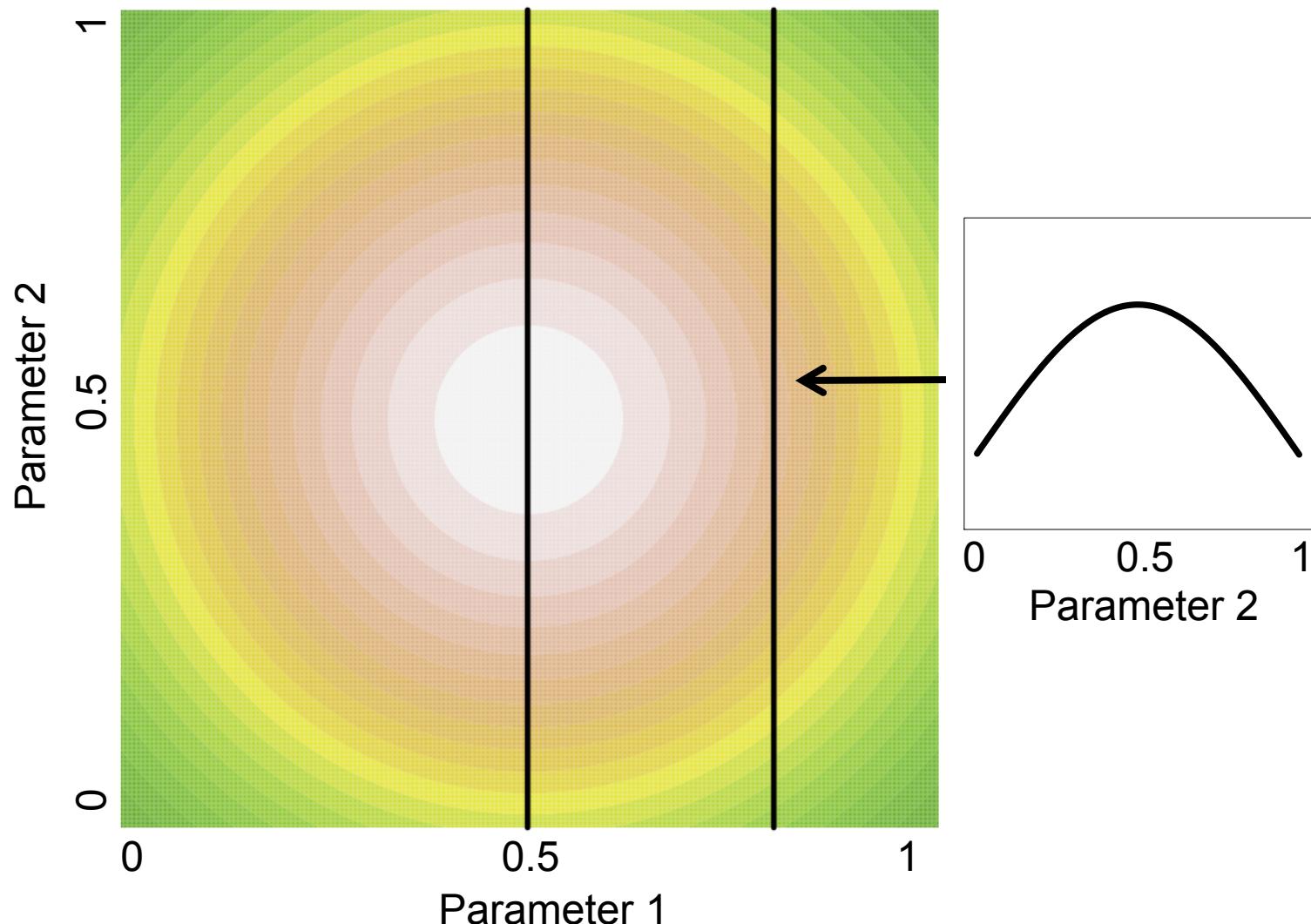
More Parameters = Better ML Score



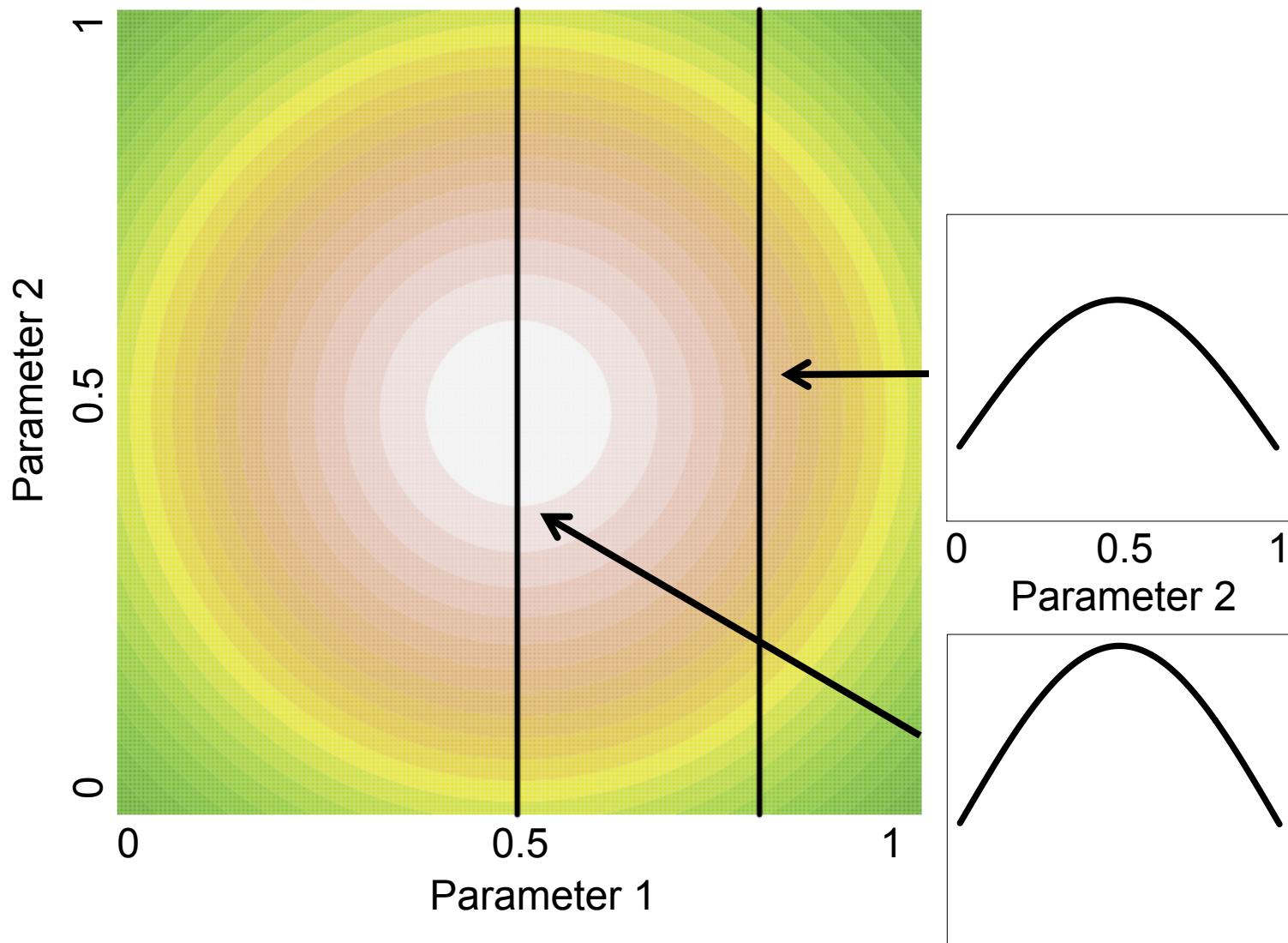
More Parameters = Better ML Score



More Parameters = Better ML Score



More Parameters = Better ML Score



Maximum Likelihood

Usually implicit

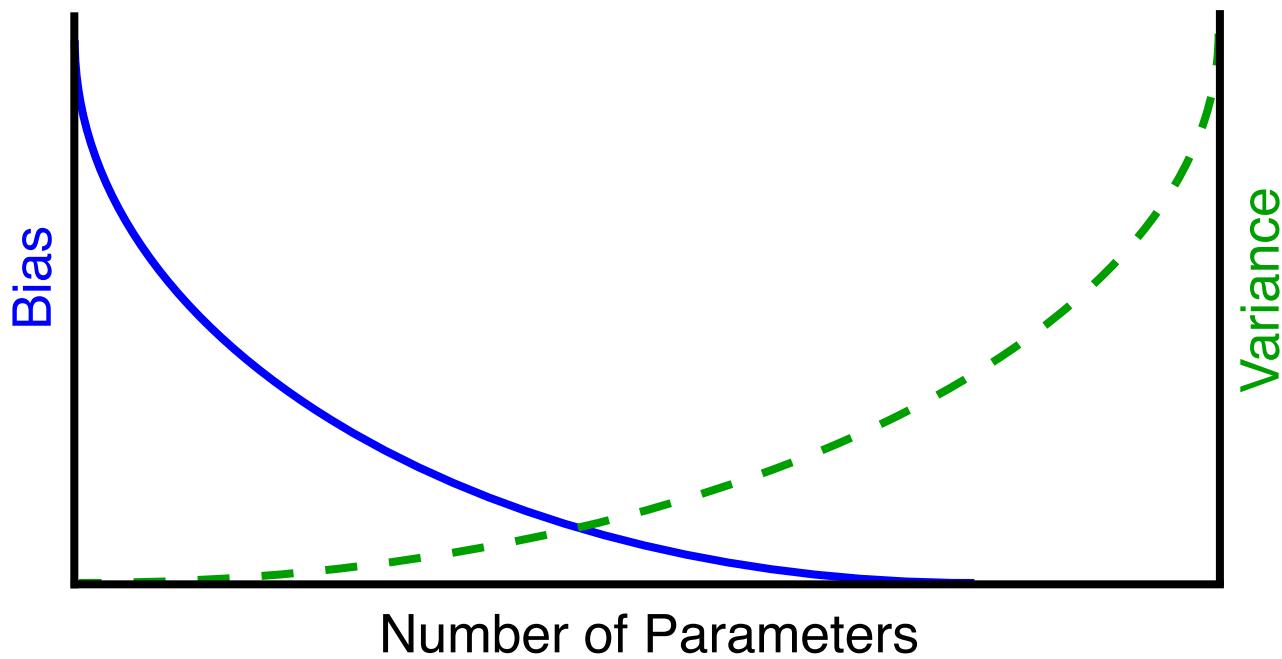
$$P(\text{grid} | \text{tree}, \theta, M)$$



What tree and parameter values give the highest likelihood?

ML scores are just relative, so alone they doesn't tell us how confident we are in this solution, just that this is the preferred solution.

ML-based Model Selection



ML-based Model Selection Criteria

- AIC – Penalty from information theory
- BIC – Penalty designed to mimic a posterior
- DT – Penalty based on performance
(e.g., branch-length estimation)
- LRT – Compares observed difference in likelihoods to expected if simple model true.
This test requires nested models.

ML-based Model Selection Criteria

Let's try this with jModelTest!

- (1) Load Data
- (2) Infer Likelihoods
- (3) Calculate Scores

ML Inference

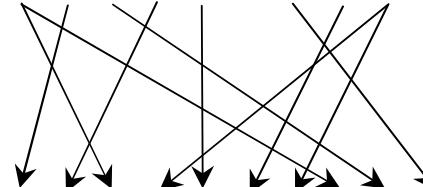


Usually done in Garli or RAxML
for large datasets.

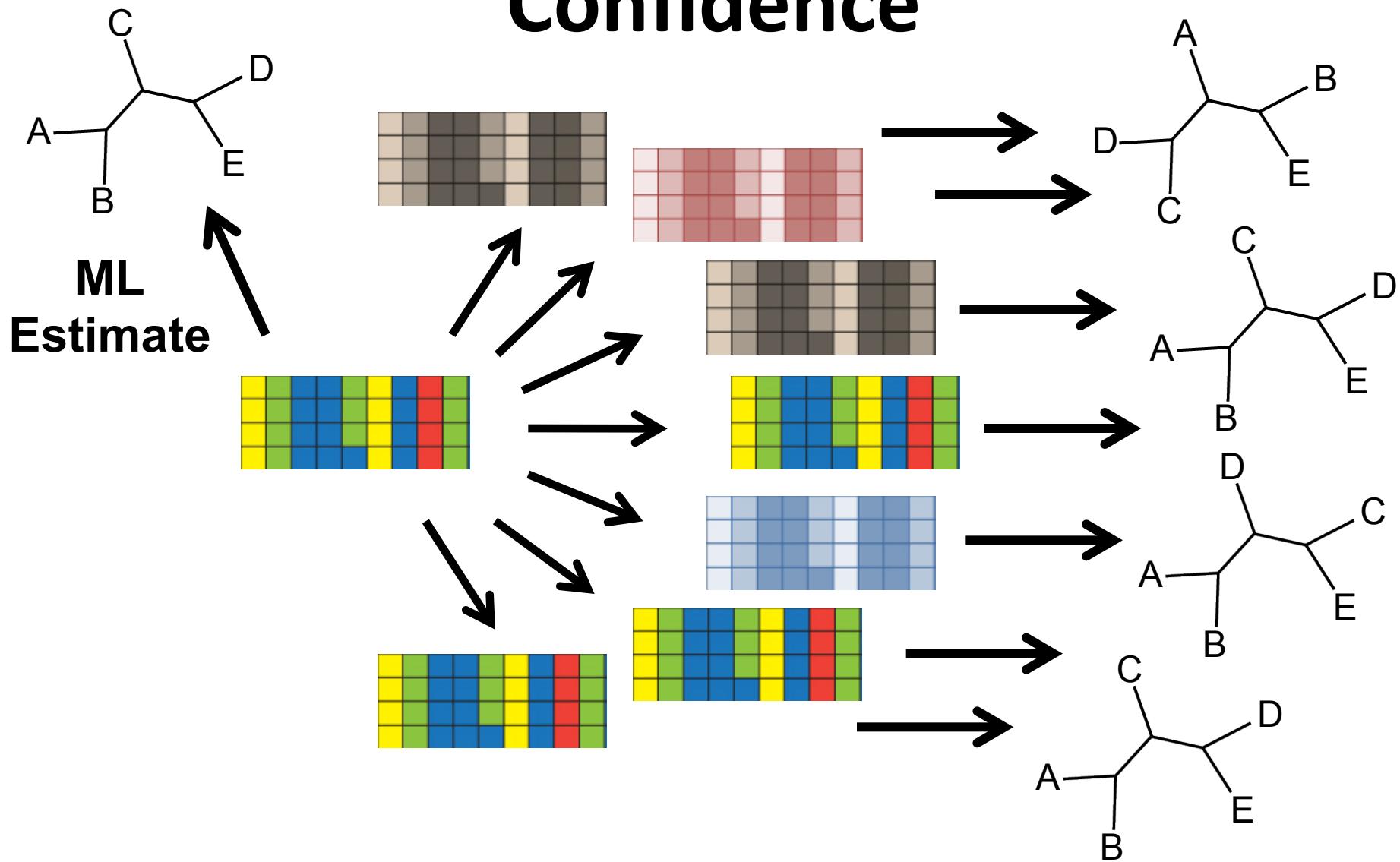
Garli uses a genetic algorithm to
find the best trees.

Bootstrapping to Assess Confidence

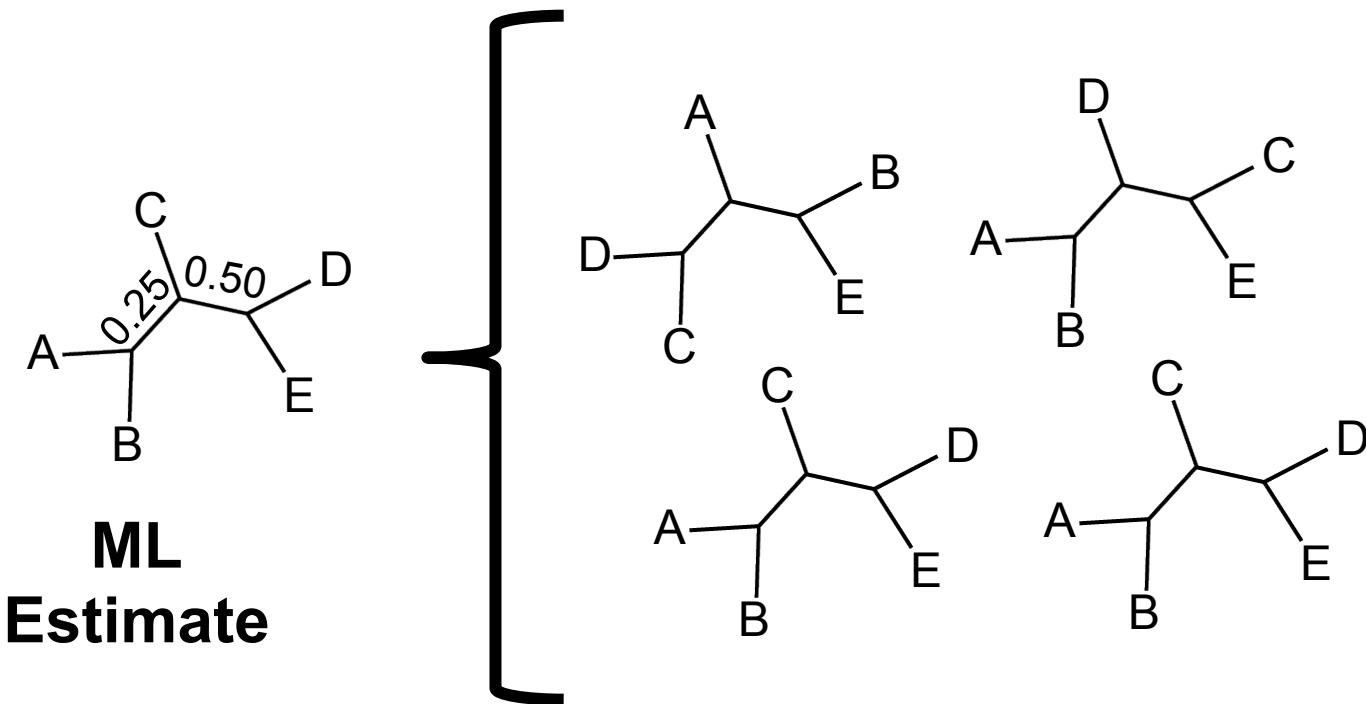
Original alignment	Site	1	2	3	4	5	6	7	8	9	10
	human	N	E	N	L	F	A	S	F	I	A
	chimpanzee	N	E	N	L	F	A	S	F	A	A
	bonobo	N	E	N	L	F	A	S	F	A	A
	gorilla	N	E	N	L	F	A	S	F	I	A
	orangutan	N	E	D	L	F	T	P	F	T	T
	Sumatran	N	E	S	L	F	T	P	F	I	T
	gibbon	N	E	N	L	F	T	S	F	A	T
Bootstrap sample	Site	2	4	1	9	5	8	9	1	3	7
	human	E	L	N	I	F	F	I	N	N	S
	chimpanzee	E	L	N	A	F	F	A	N	N	S
	bonobo	E	L	N	A	F	F	A	N	N	S
	gorilla	E	L	N	I	F	F	I	N	N	S
	orangutan	E	L	N	T	F	F	T	N	D	P
	Sumatran	E	L	N	I	F	F	I	N	S	P
	gibbon	E	L	N	A	F	F	A	N	N	S



Bootstrapping to Assess Confidence



Bootstrapping to Assess Confidence



Bootstrapping to Assess Confidence

Interpretations of the bootstrap:

- Repeatability
- $1 - \text{False Positive Rate from a polytomy}$
- Probability branch is true

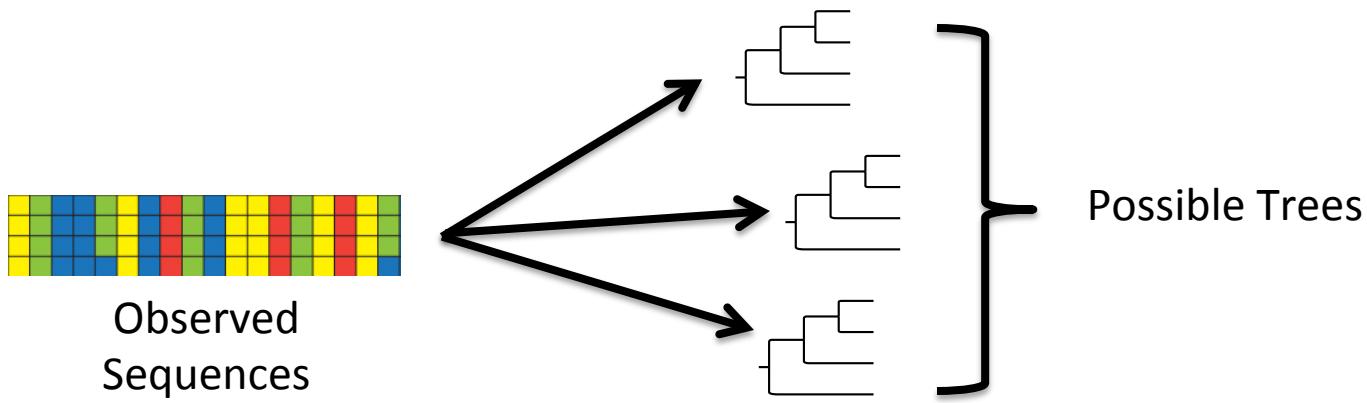
Bootstrapping to Assess Confidence

Interpretations of the bootstrap:

- Repeatability
- $1 - \text{False Positive Rate from a polytomy}$



Bayesian Inference



Posterior Probability: Conditional on observed data (alignment), the probability that a particular tree (or part of a tree) is true.

$$P(\text{Tree} \mid \text{Observed Sequences})$$

Bayes' Theorem

$$\frac{P(\text{ } \cdot \text{ }, \theta) \cdot P(\text{ } \cdot \text{ } | \text{ } \cdot \text{ }, \theta)}{P(\text{ } \cdot \text{ })} = P(\text{ } \cdot \text{ } | \text{ } \cdot \text{ })$$

Prior Probability
↓
 $P(\text{ } \cdot \text{ }, \theta)$

Likelihood
↓
 $P(\text{ } \cdot \text{ } | \text{ } \cdot \text{ }, \theta)$

Posterior Probability
↓
 $P(\text{ } \cdot \text{ } | \text{ } \cdot \text{ })$

Normalizing Constant
↑

Marginalizing

Joint
Posterior
Probability



$$P(\text{data} \mid \theta)$$

Marginalizing

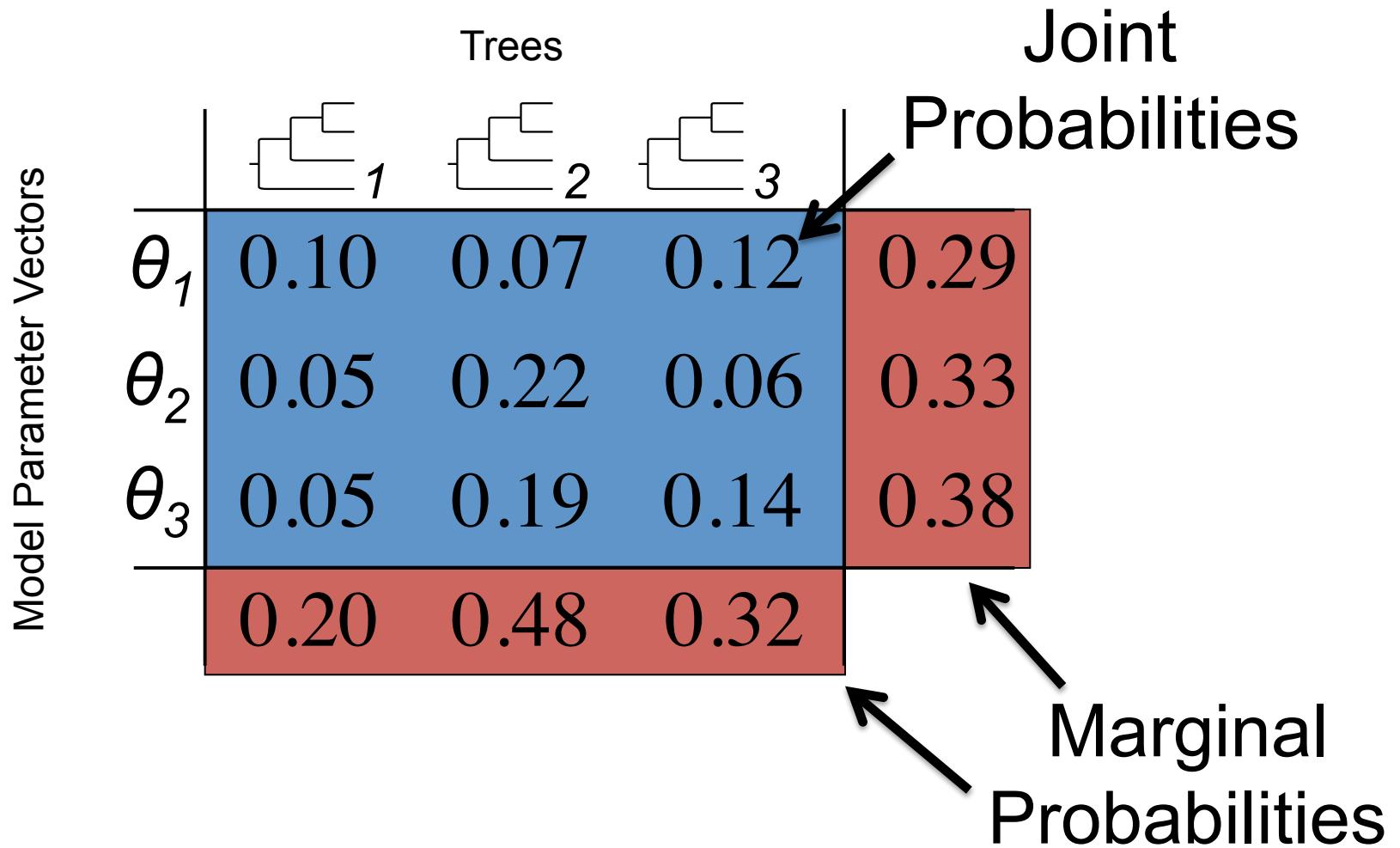
$$\int P(\text{data}, \theta | \text{parameters}) d\theta = P(\text{data} | \text{parameters})$$

Joint Posterior Probability

Marginal Posterior Probability

Integrating across all values of model parameters

Marginalizing



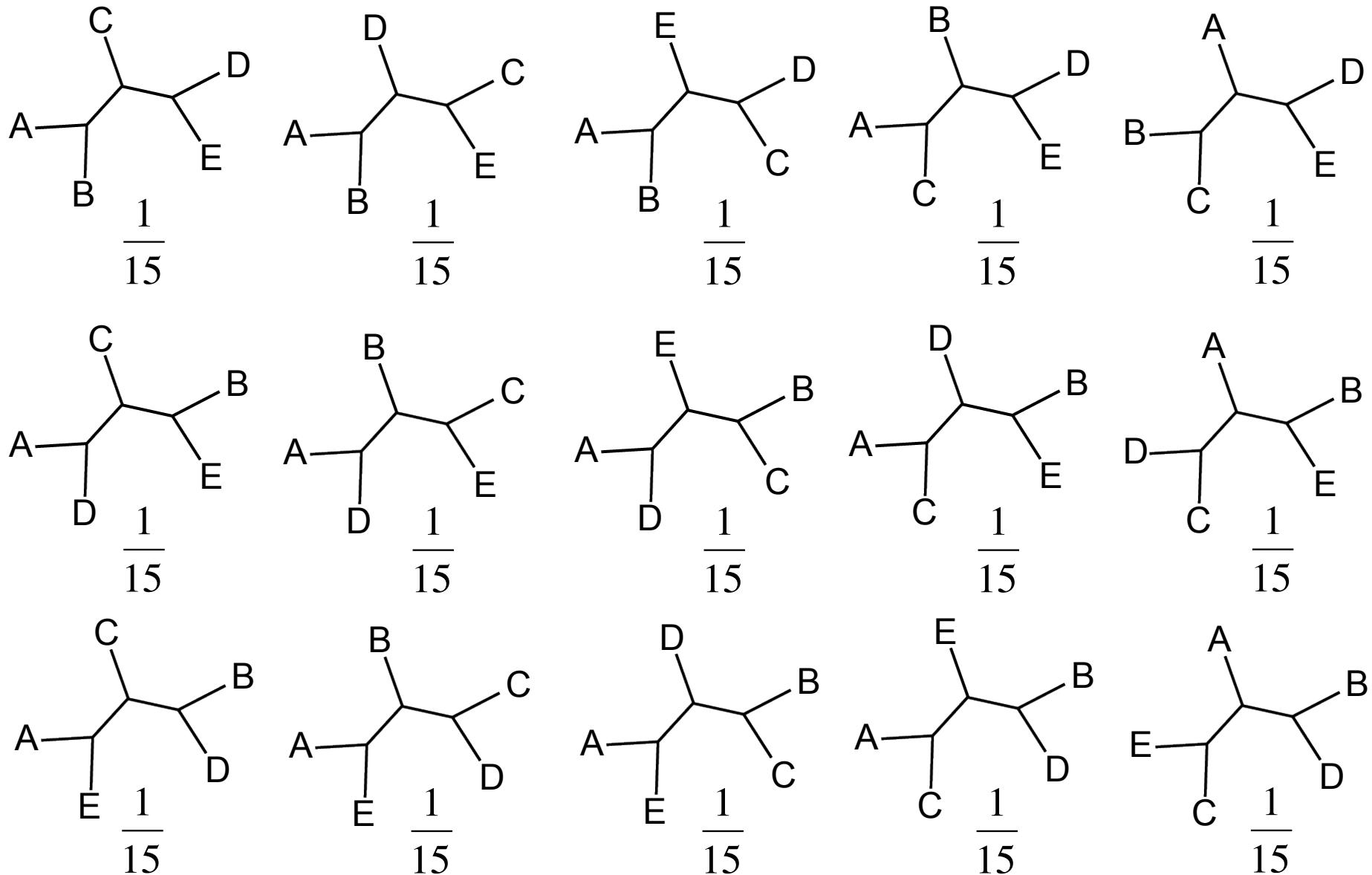
Marginalizing Across Models

		Trees			Joint Probabilities	
		1	2	3		
		0.10	0.07	0.12	0.29	
M_1		0.05	0.22	0.06	0.33	
M_2		0.05	0.19	0.14	0.38	
		0.20	0.48	0.32		
					Marginal Probabilities	

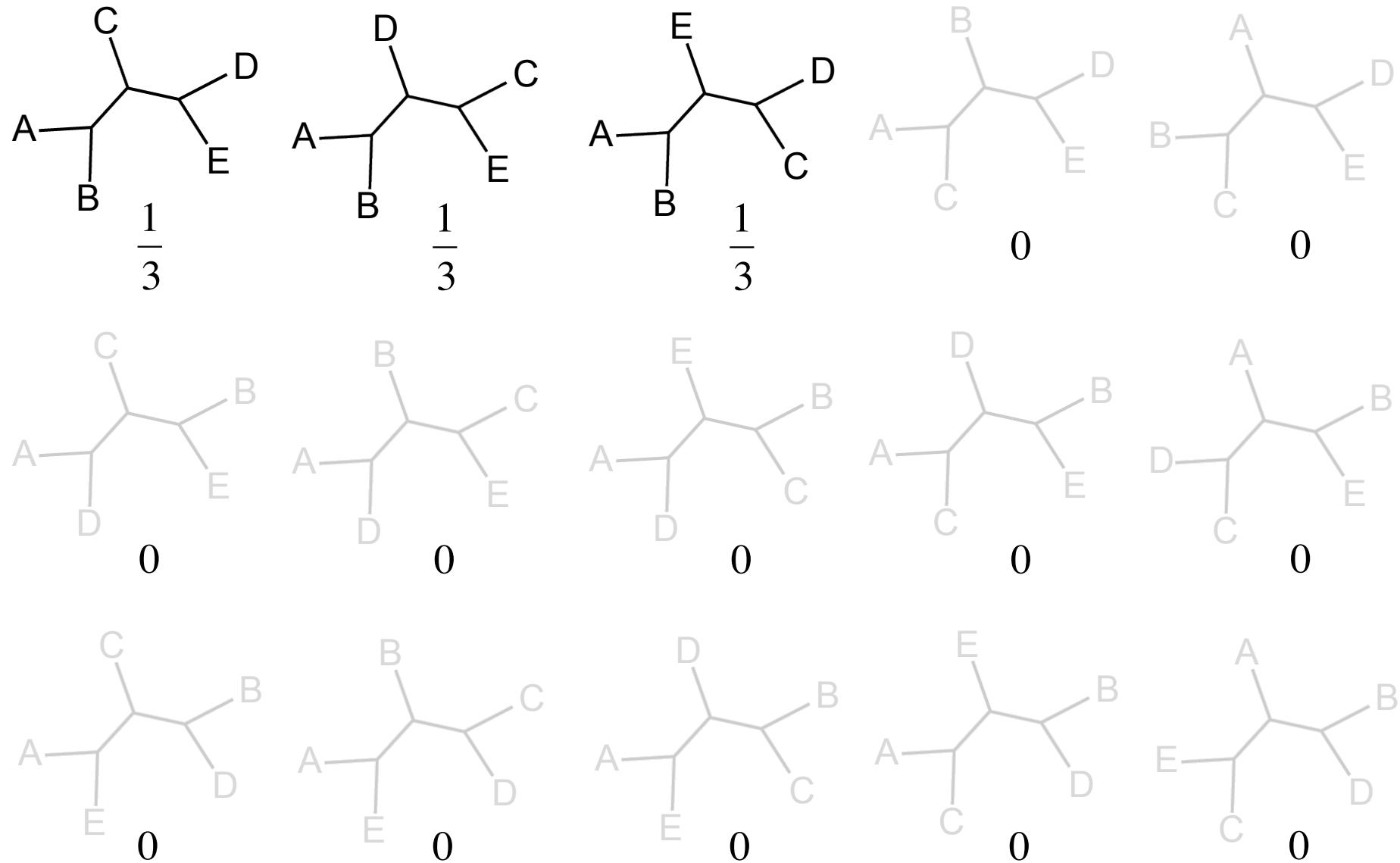
What Priors to Use?

- The controversial part of Bayesian analysis
- Choice can vary by researcher
- Often chosen in an attempt to reflect prior ignorance
- Analysis can be run under several priors to assess sensitivity

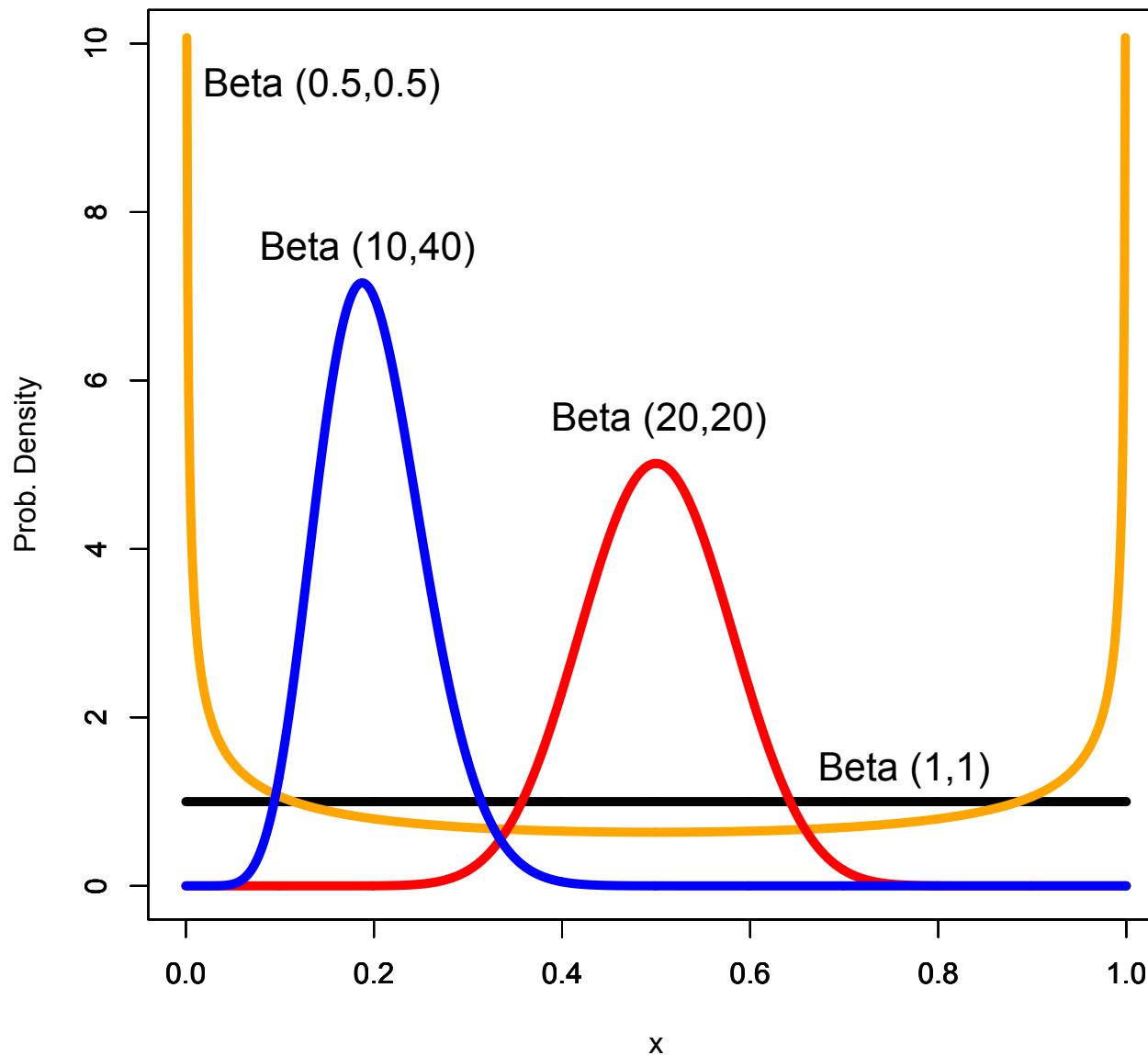
Uniform Topology Prior



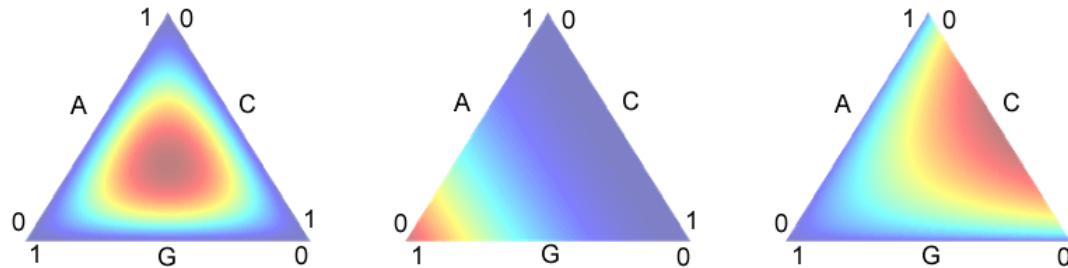
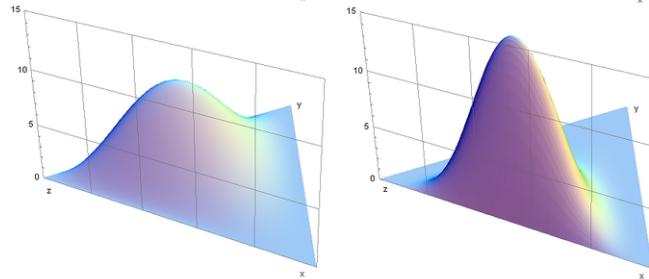
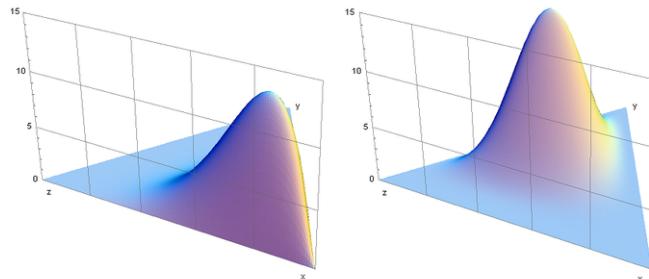
$(AB \mid CDE)$ Constraint Prior



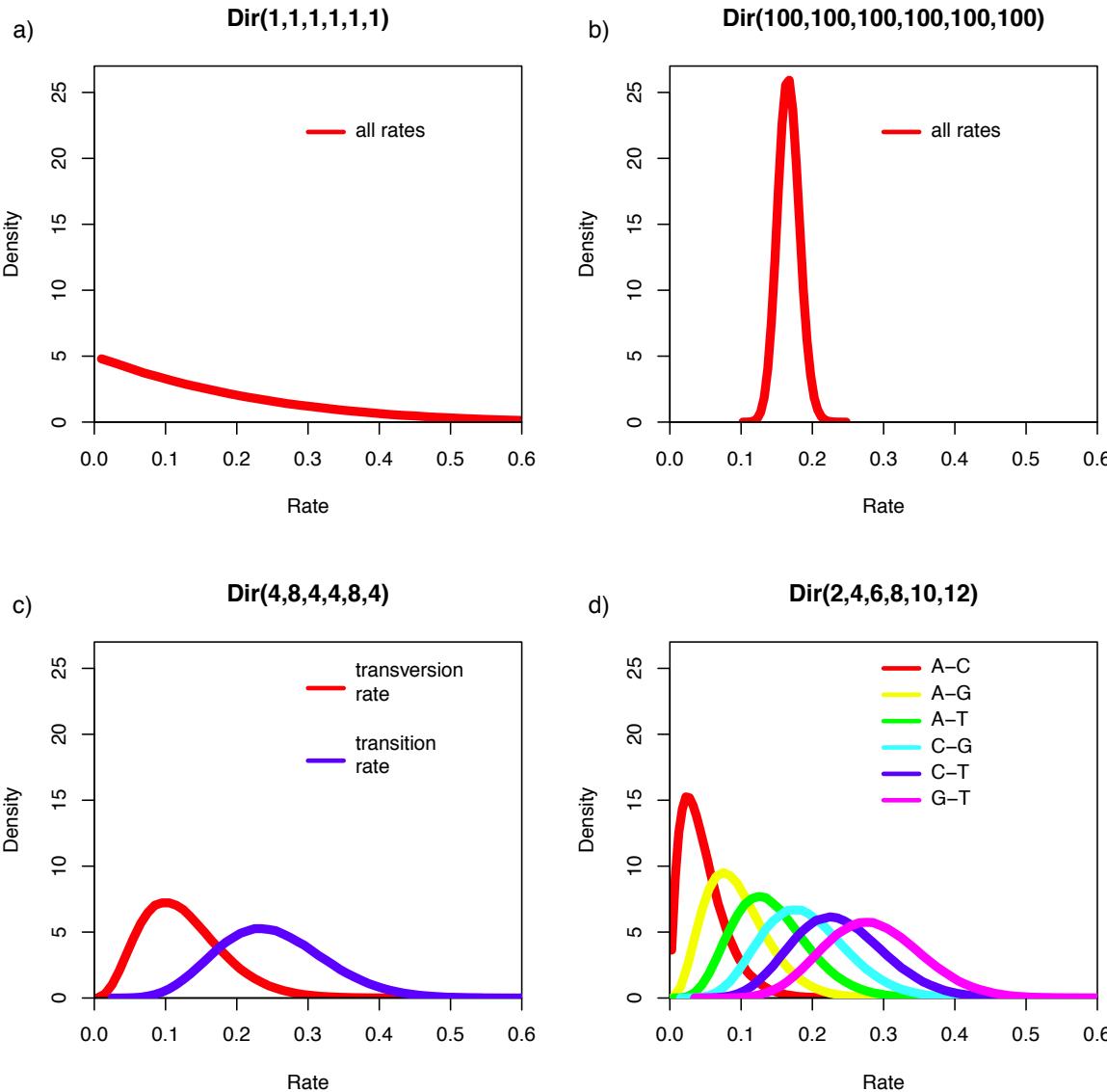
Beta Distributions



Prior on Sets - Dirichlet Distribution

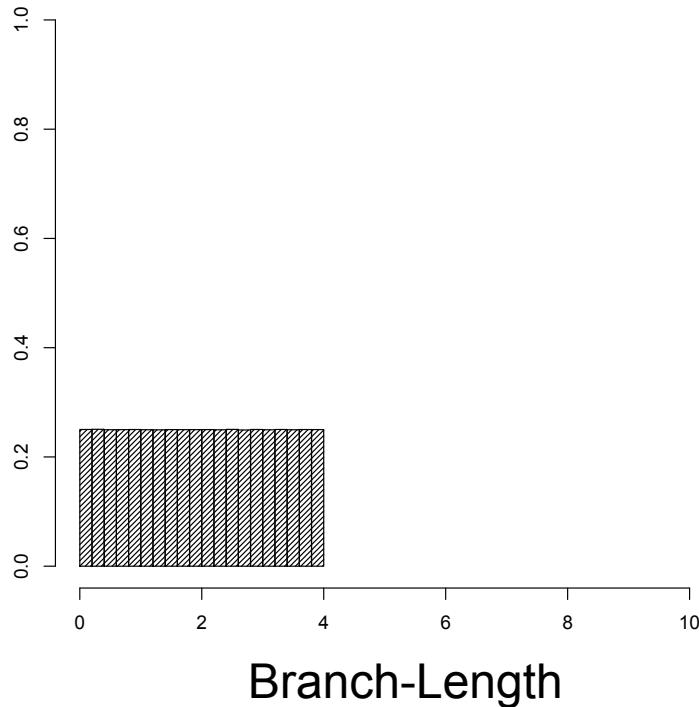


Prior on Sets - Dirichlet Distribution

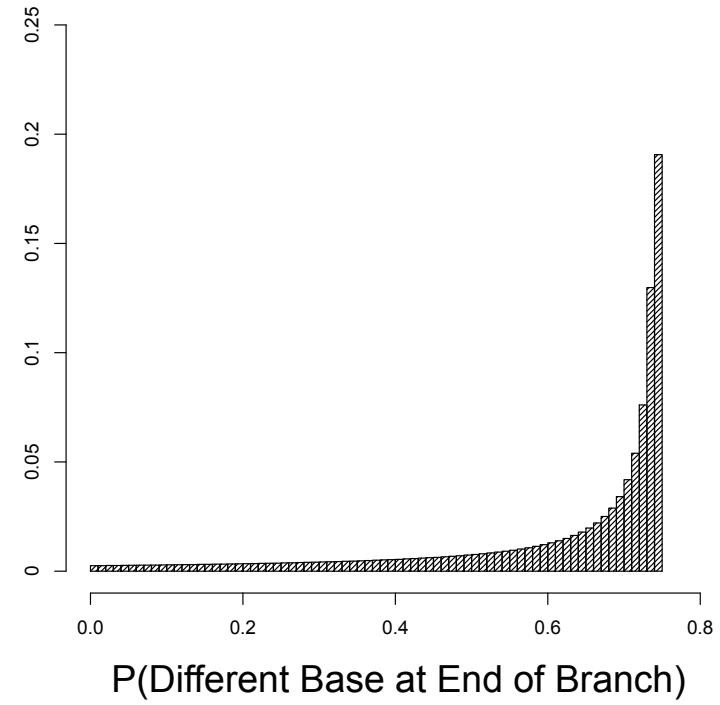


Branch-Length Priors

Uniform(0,4) Branch-Length Prior



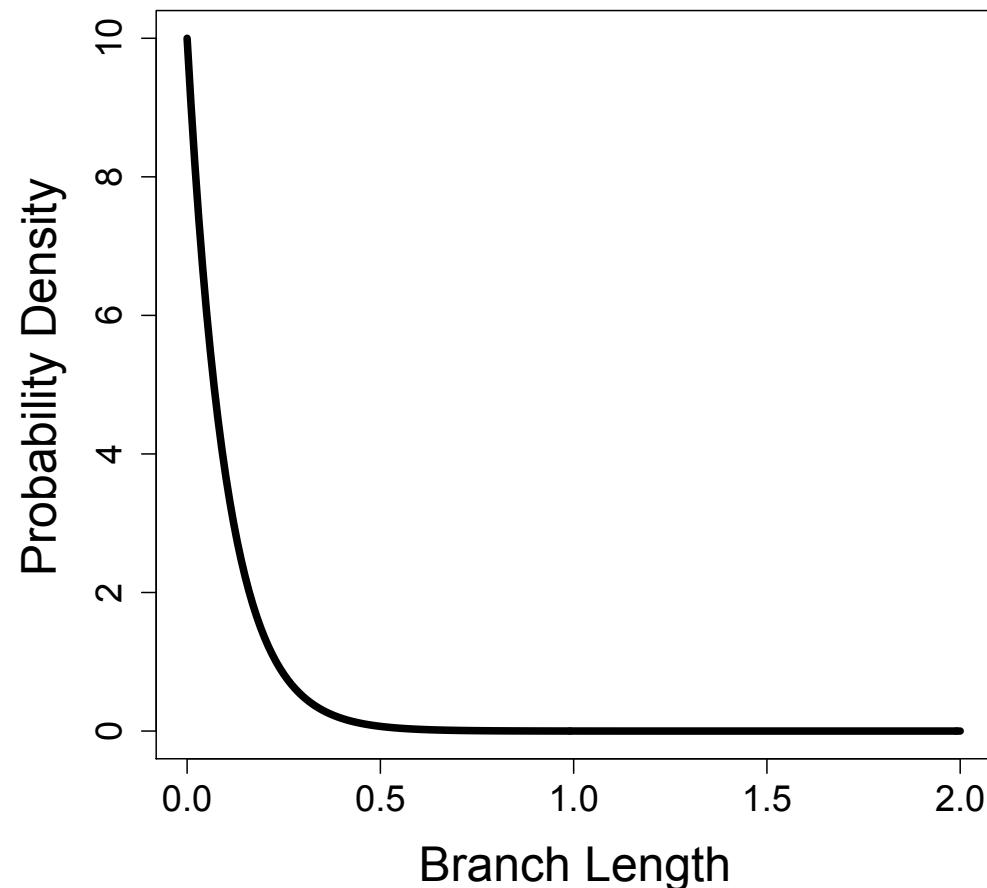
Strongly Informative Prior on Change



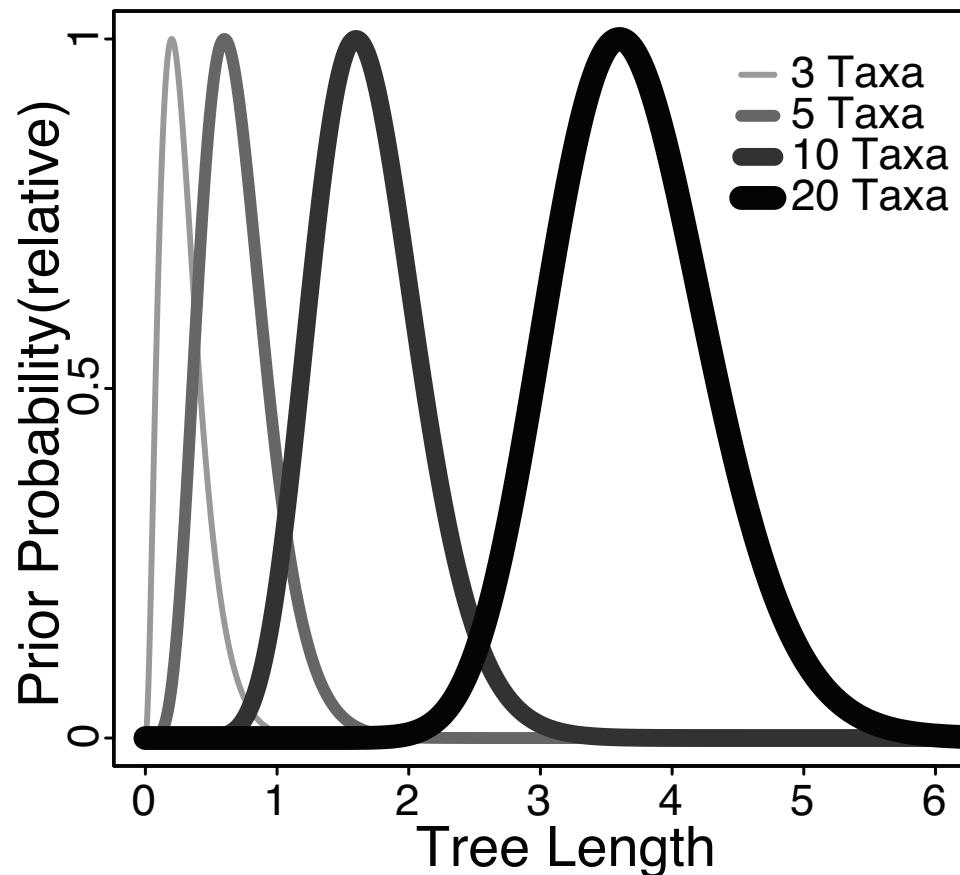
Idea from P.O. Lewis

Branch-Length Priors

Default Exponential Branch-Length Prior ($\lambda=10$, mean=0.1)

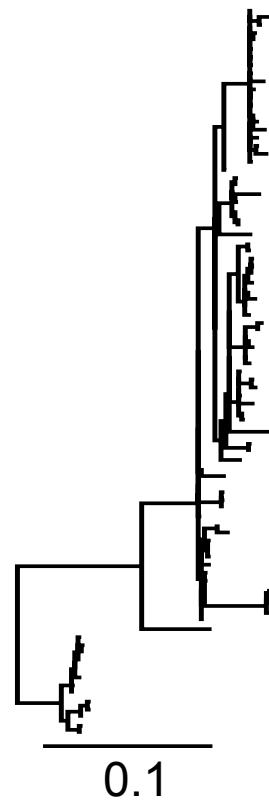
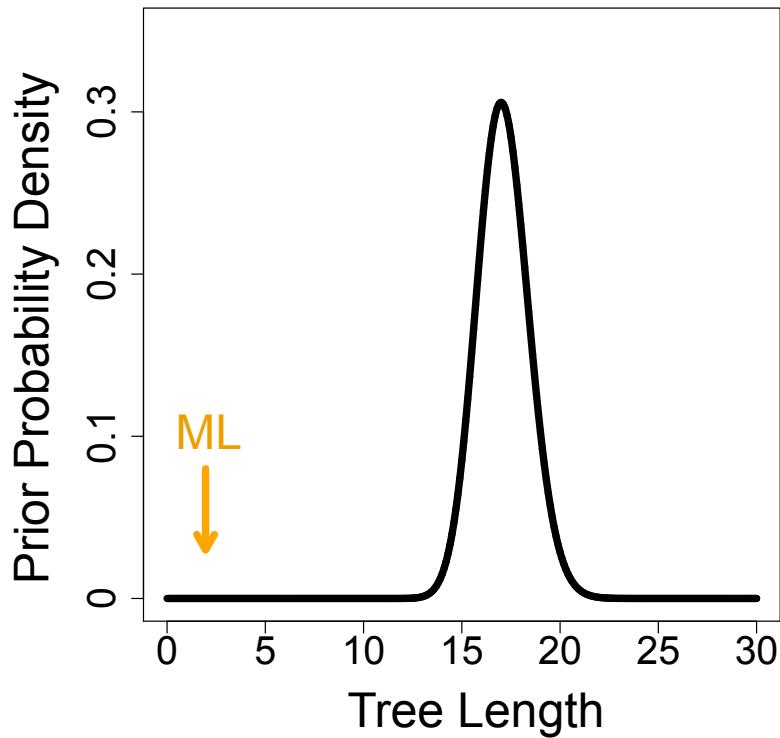


Implied Tree-Length Prior

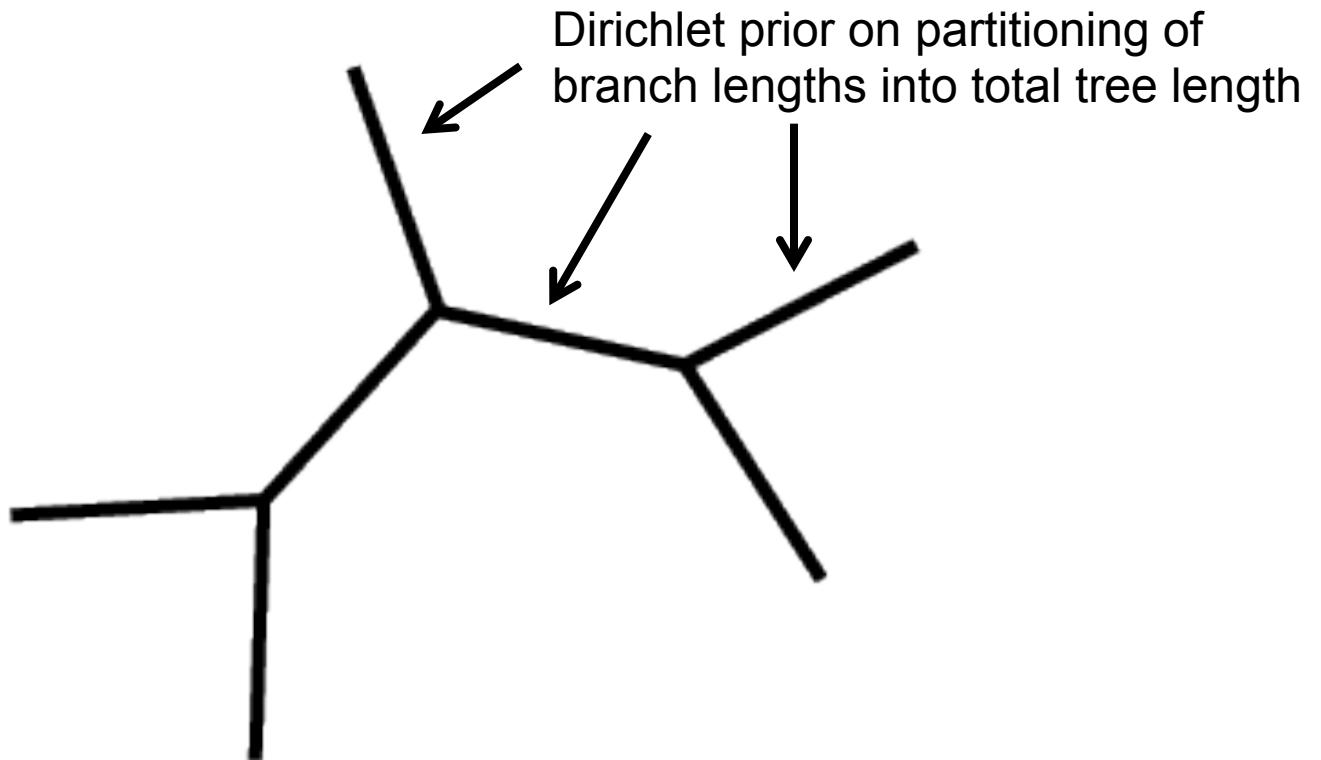


Implied Tree-Length Prior

88 Taxon Tree



Compound Branch-Length Prior



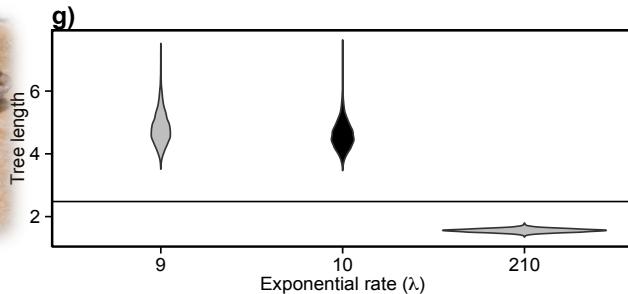
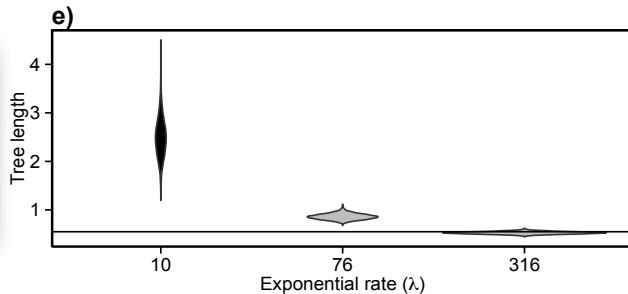
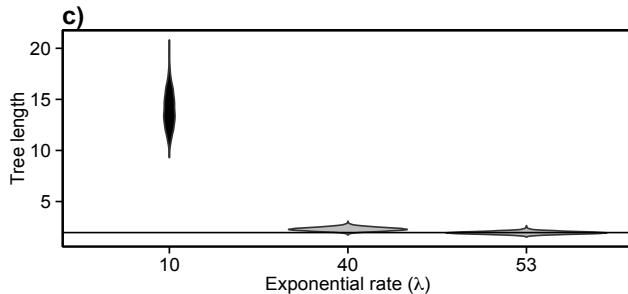
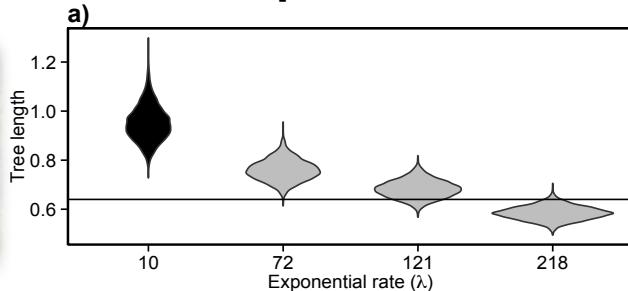
Total Tree Length Prior: (Inv) Gamma

Rannala et al. 2011. Mol. Biol. Evol.
Zhang et al. 2012. Syst. Biol

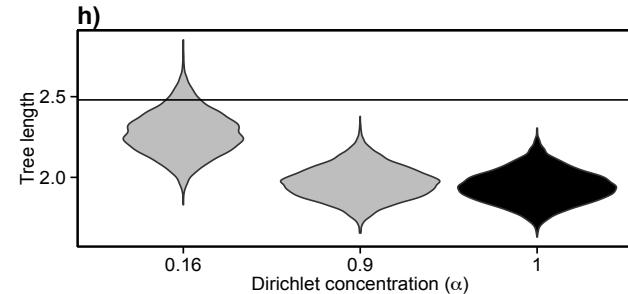
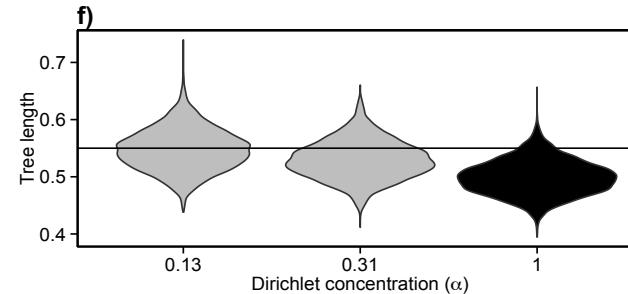
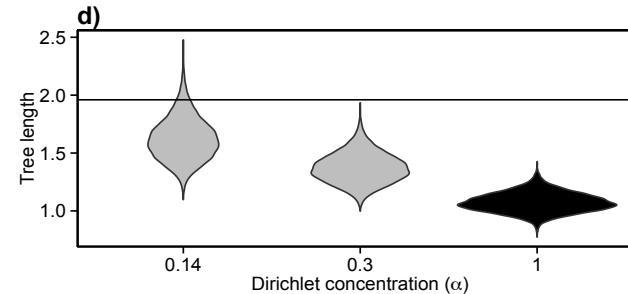
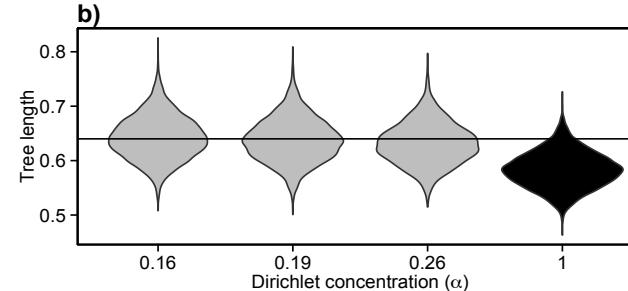
Outside Info Improves Estimates

Dataset	Clams	Frogs
ML TL Estimate	1.96	0.55
Bayes TL Interval (MB Default)	10.7 - 17.7	1.77-3.29
Bayes TL Interval (Informed)	1.15 – 1.43	0.32 – 0.38
Bayes TL Interval (Compound Prior)	0.9 – 1.3	0.44 – 0.57

Exponential



Compound Dirichlet



Brad
Nelson

EmpPrior

<https://code.google.com/p/empprior/>

Searches TreeBase to find other phylogenetic studies that have used the same gene at roughly the same taxonomic depth, infers an ML tree, and fits the parameters of a branch-length distribution using these related data sets.

Written by John Andersen and Brad Nelson

Bayes' Theorem

$$\frac{P(\text{Evidence}, \theta) \cdot P(\text{Image} | \text{Evidence}, \theta)}{P(\text{Image})} = P(\text{Evidence} | \theta, \text{Image})$$

$P(\text{Image})$

No Analytical
Solution

Posterior Odds Ratio

$$\frac{P(\text{[staircase diagram]}_1, \theta_1 | \text{[color bar]})}{P(\text{[staircase diagram]}_2, \theta_2 | \text{[color bar]})}$$

Posterior Odds Ratio

$$\frac{\frac{P(\text{Tree}_1, \theta_1) \cdot P(\text{Data} | \text{Tree}_1, \theta_1)}{P(\text{Data})}}{\frac{P(\text{Tree}_2, \theta_2) \cdot P(\text{Data} | \text{Tree}_2, \theta_2)}{P(\text{Data})}} = \frac{P(\text{Tree}_1, \theta_1 | \text{Data})}{P(\text{Tree}_2, \theta_2 | \text{Data})}$$

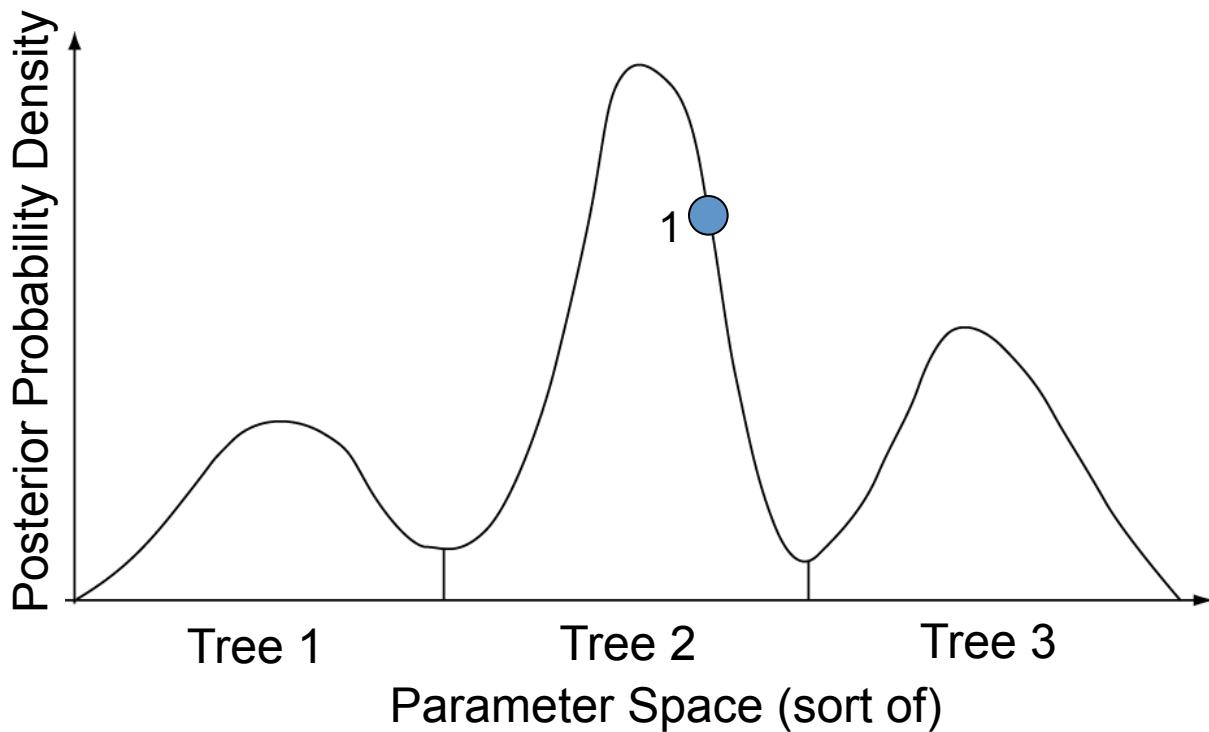
The equation illustrates the posterior odds ratio for two competing phylogenetic trees, Tree 1 and Tree 2, given observed data. The numerator represents the joint probability of the data and each tree under their respective parameter sets, θ_1 and θ_2 . The denominator represents the probability of the data alone, which is shared by both trees. The final ratio is the posterior odds in favor of Tree 1 relative to Tree 2.

Posterior Odds Ratio

$$\frac{P(\text{Tree}_1, \theta_1) \cdot P(\text{Data} | \text{Tree}_1, \theta_1)}{P(\text{Tree}_2, \theta_2) \cdot P(\text{Data} | \text{Tree}_2, \theta_2)} = \frac{P(\text{Tree}_1, \theta_1 | \text{Data})}{P(\text{Tree}_2, \theta_2 | \text{Data})}$$

Markov chain Monte Carlo

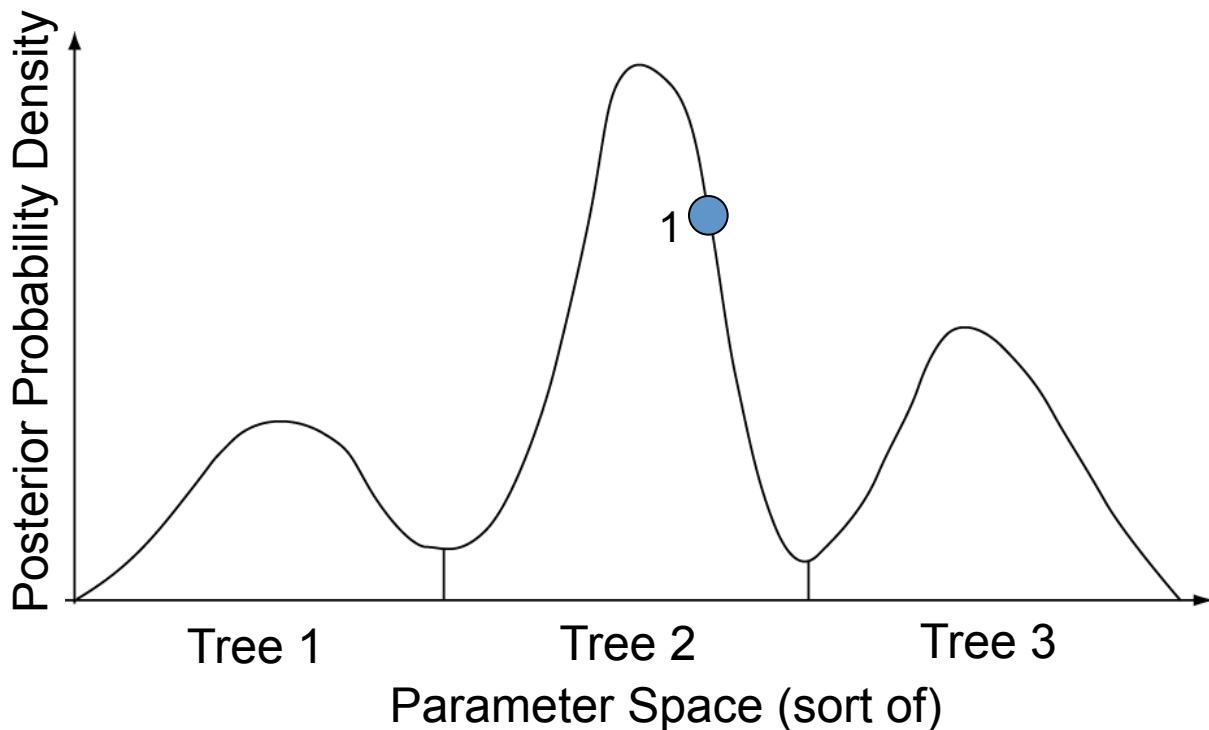
1. Start at an arbitrary point



This slide “borrowed” from F. Ronquist

Markov chain Monte Carlo

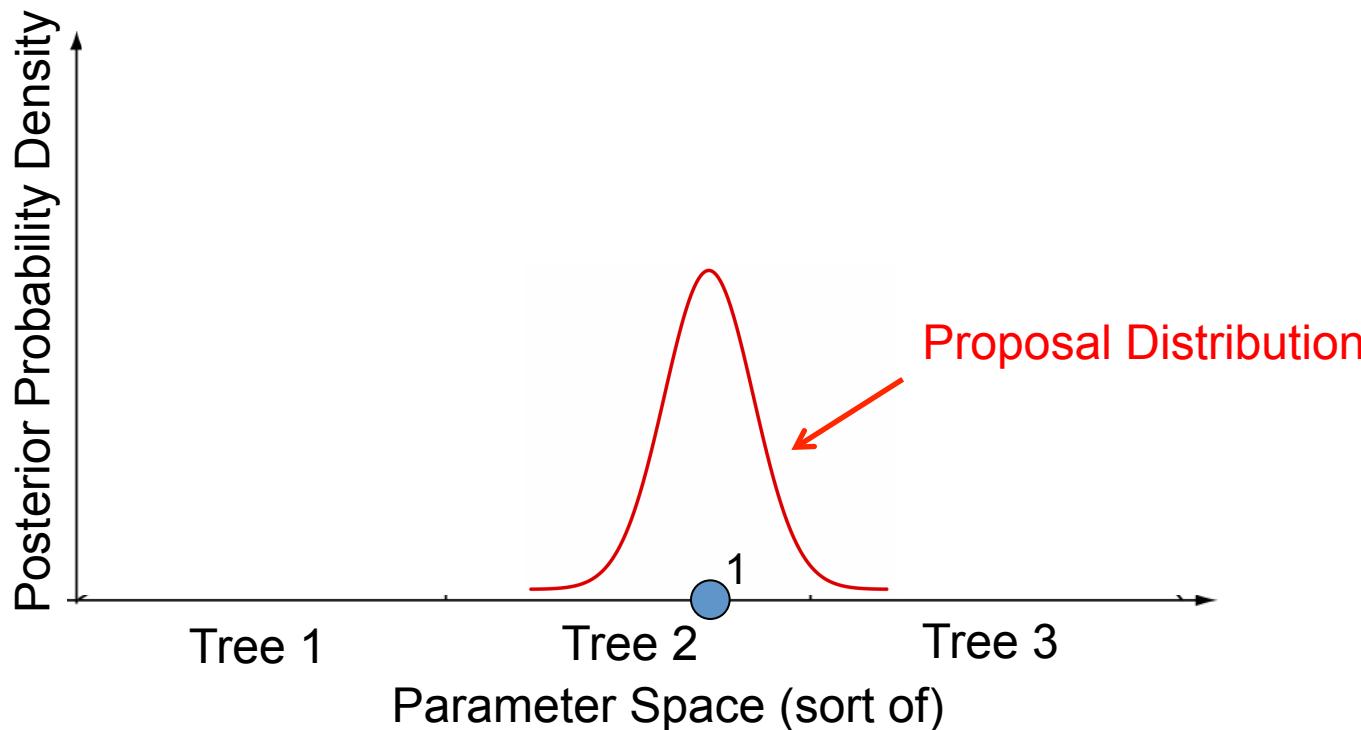
1. Start at an arbitrary point
2. Make a small random move



This slide “borrowed” from F. Ronquist

Markov chain Monte Carlo

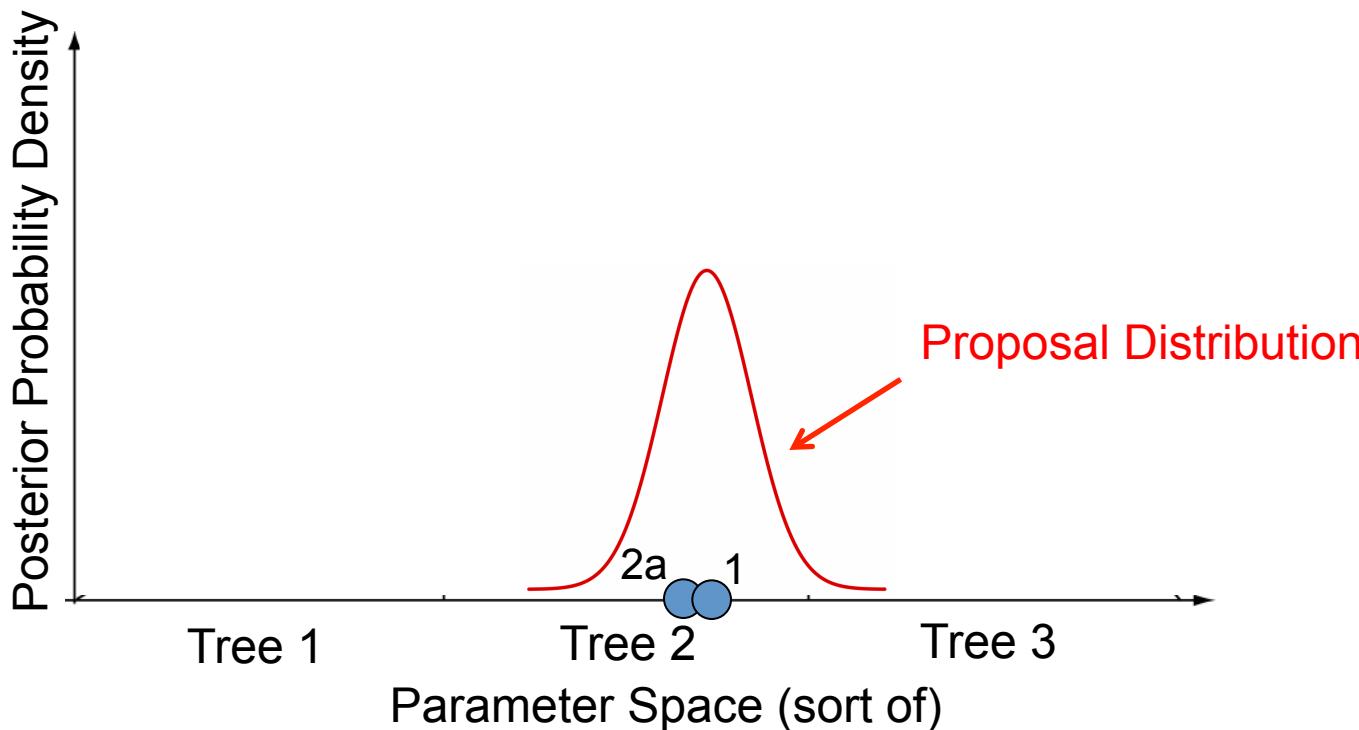
1. Start at an arbitrary point
2. Make a small random move



This slide “borrowed” from F. Ronquist

Markov chain Monte Carlo

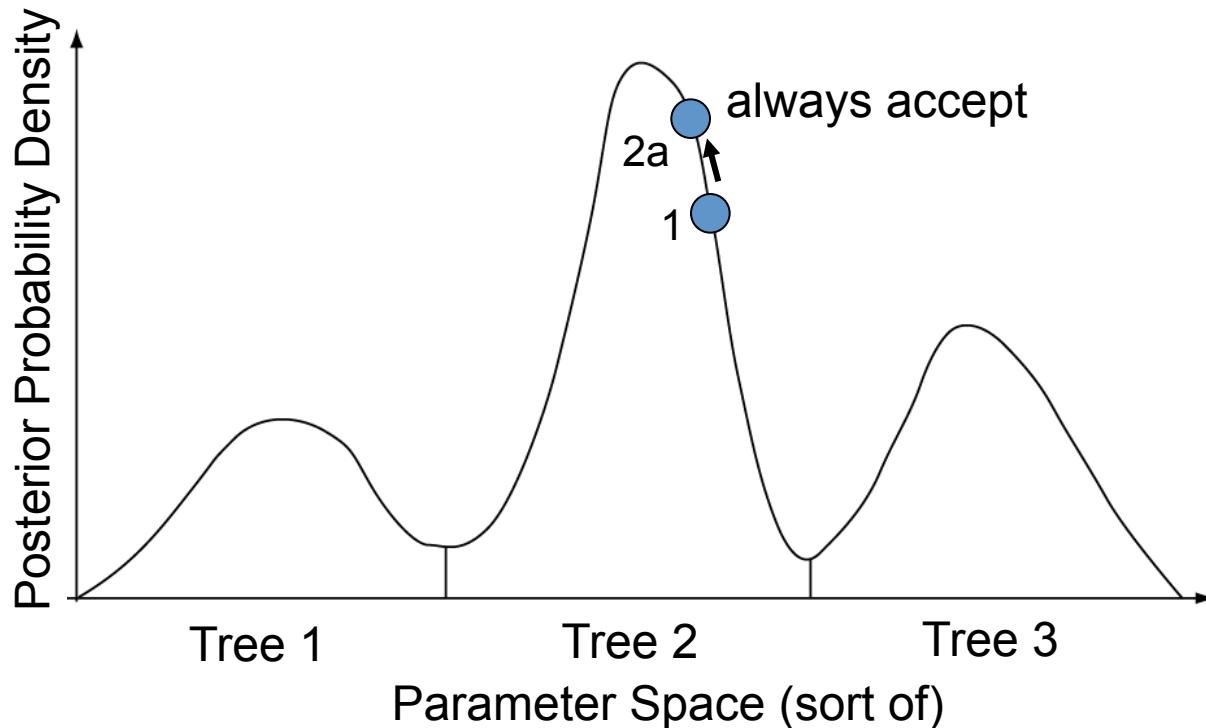
1. Start at an arbitrary point
2. Make a small random move



This slide “borrowed” from F. Ronquist

Markov chain Monte Carlo

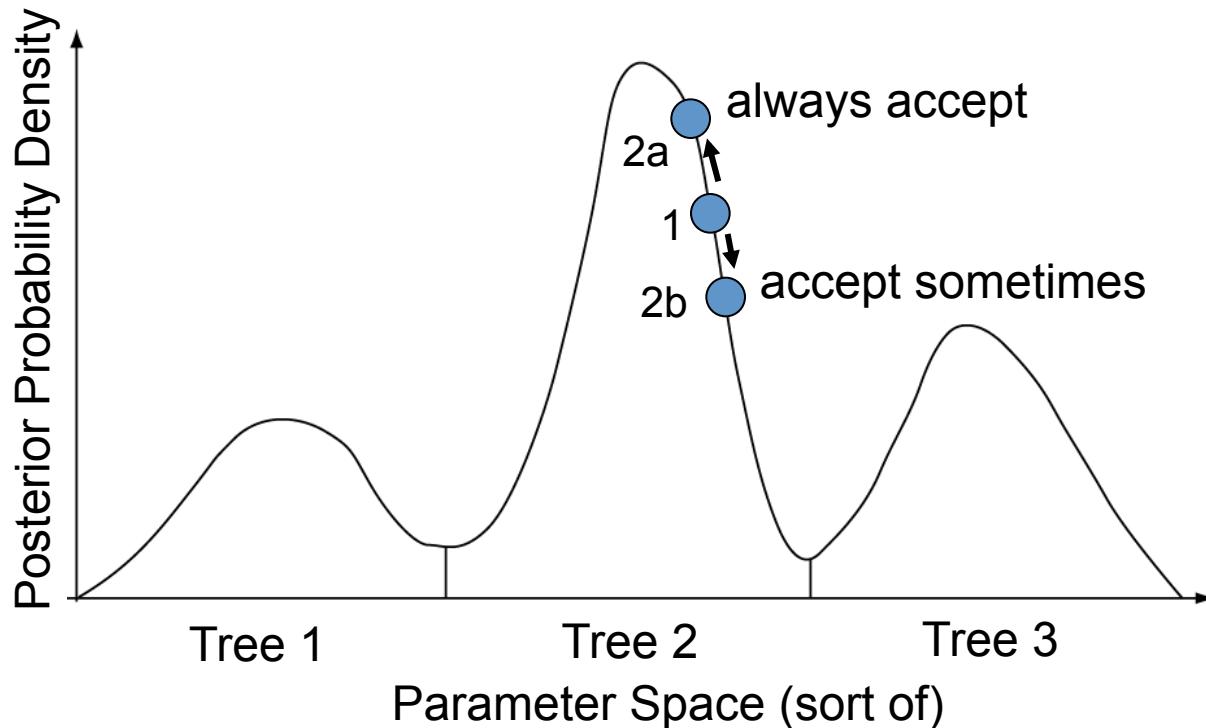
1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (r) of new state to old state:
 - a) $r > 1 \rightarrow$ new state accepted
 - b) $r < 1 \rightarrow$ new state accepted with probability r . If new state not accepted, stay in the old state



This slide “borrowed” from F. Ronquist

Markov chain Monte Carlo

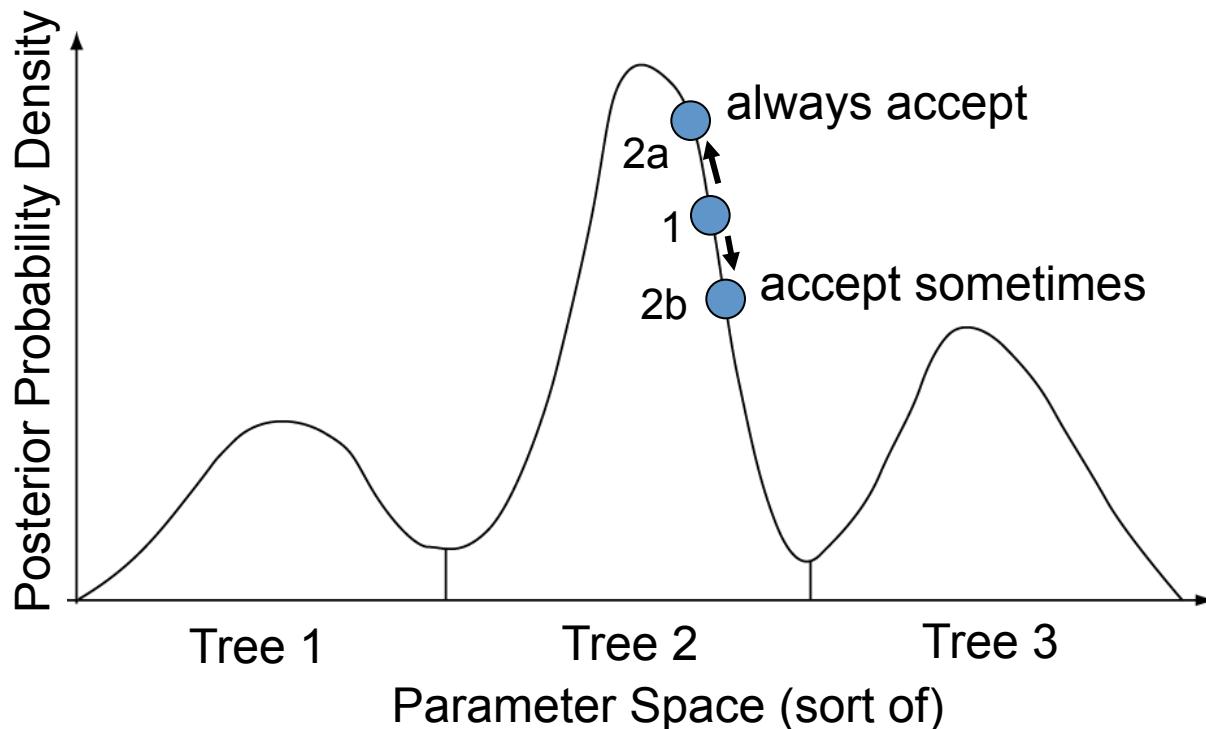
1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (r) of new state to old state:
 - a) $r > 1 \rightarrow$ new state accepted
 - b) $r < 1 \rightarrow$ new state accepted with probability r . If new state not accepted, stay in the old state



This slide “borrowed” from F. Ronquist

Markov chain Monte Carlo

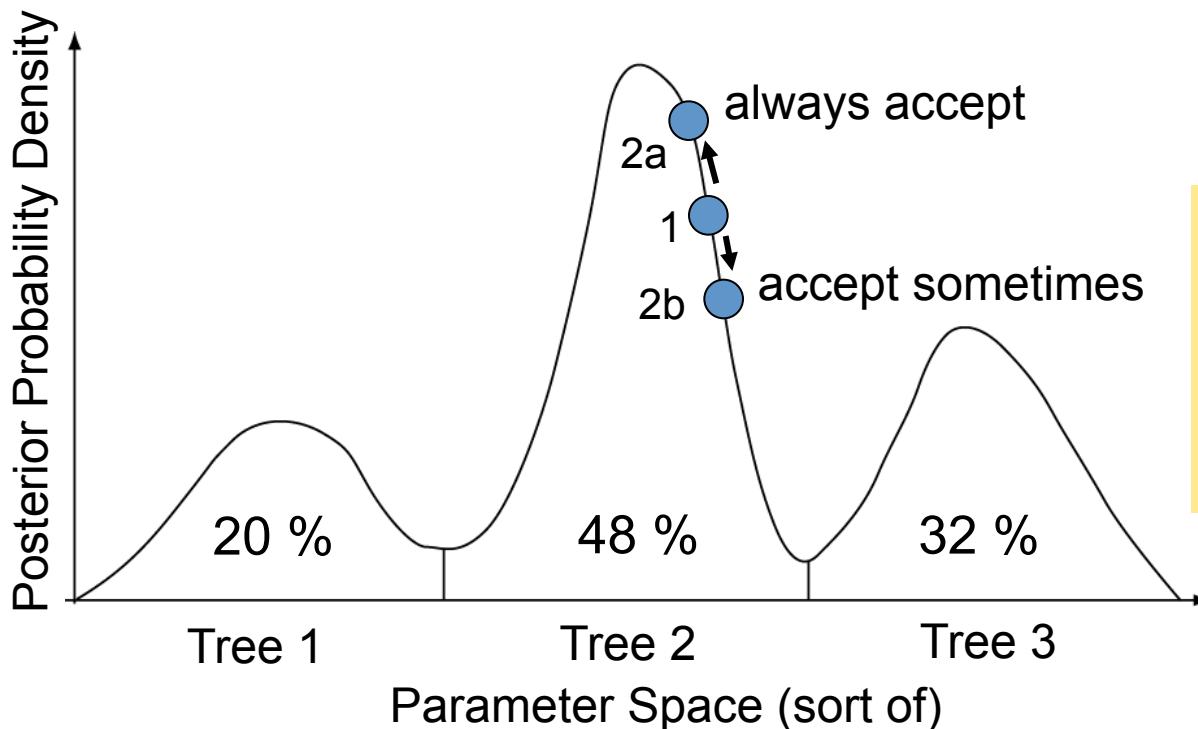
1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (r) of new state to old state:
 - a) $r > 1 \rightarrow$ new state accepted
 - b) $r < 1 \rightarrow$ new state accepted with probability r . If new state not accepted, stay in the old state
4. Go to step 2 a BUNCH ($\times 10,000$'s – $\times 10,000,000$'s)



This slide “borrowed” from F. Ronquist

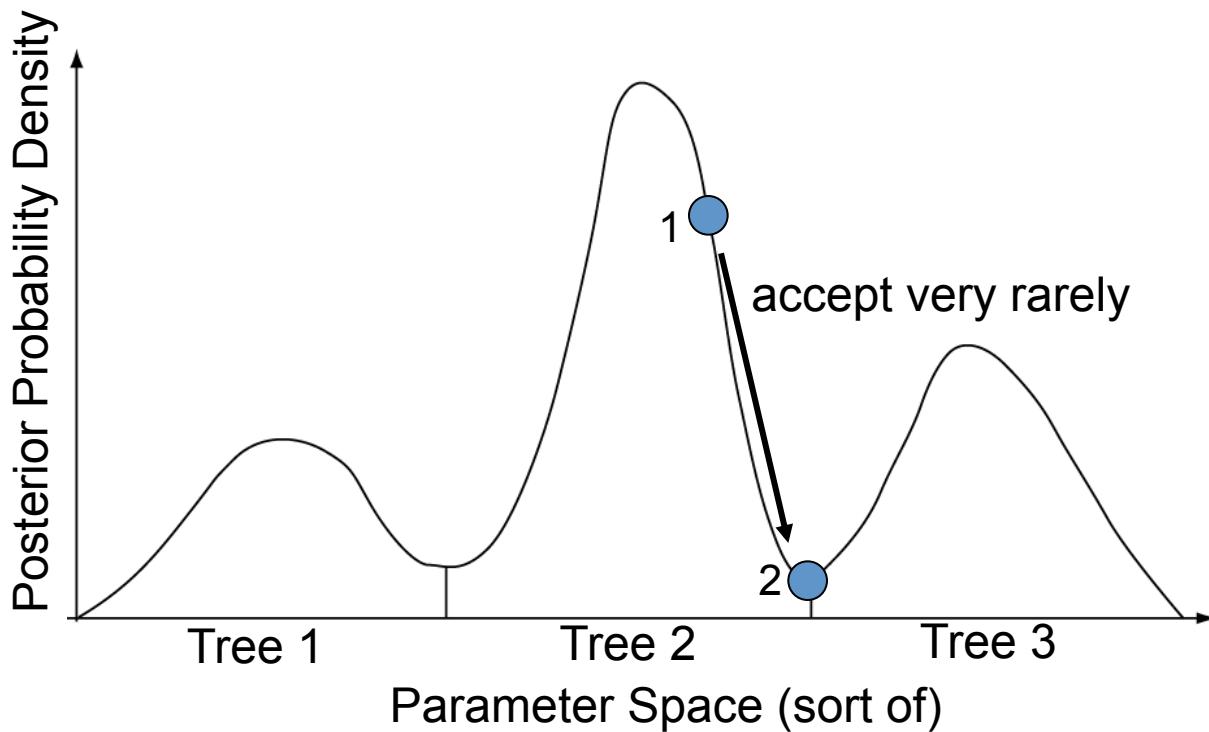
Markov chain Monte Carlo

1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (r) of new state to old state:
 - a) $r > 1 \rightarrow$ new state accepted
 - b) $r < 1 \rightarrow$ new state accepted with probability r . If new state not accepted, stay in the old state
4. Go to step 2 a BUNCH ($\times 10,000$'s – $\times 10,000,000$'s)



The proportion of time the MCMC procedure samples from a particular parameter region is an estimate of that region's posterior probability density

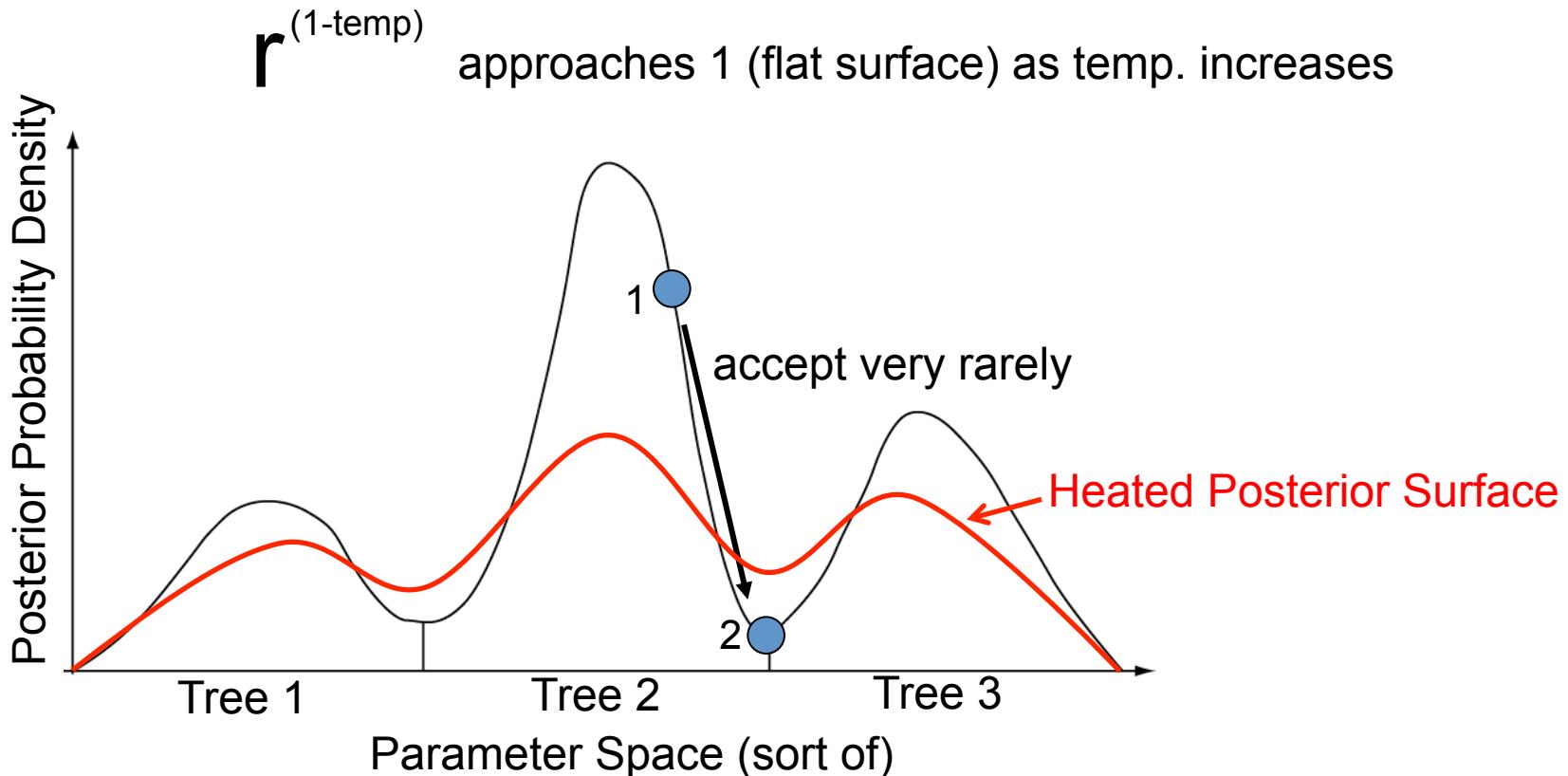
Metropolis Coupling



This slide “borrowed” from F. Ronquist

Metropolis Coupling

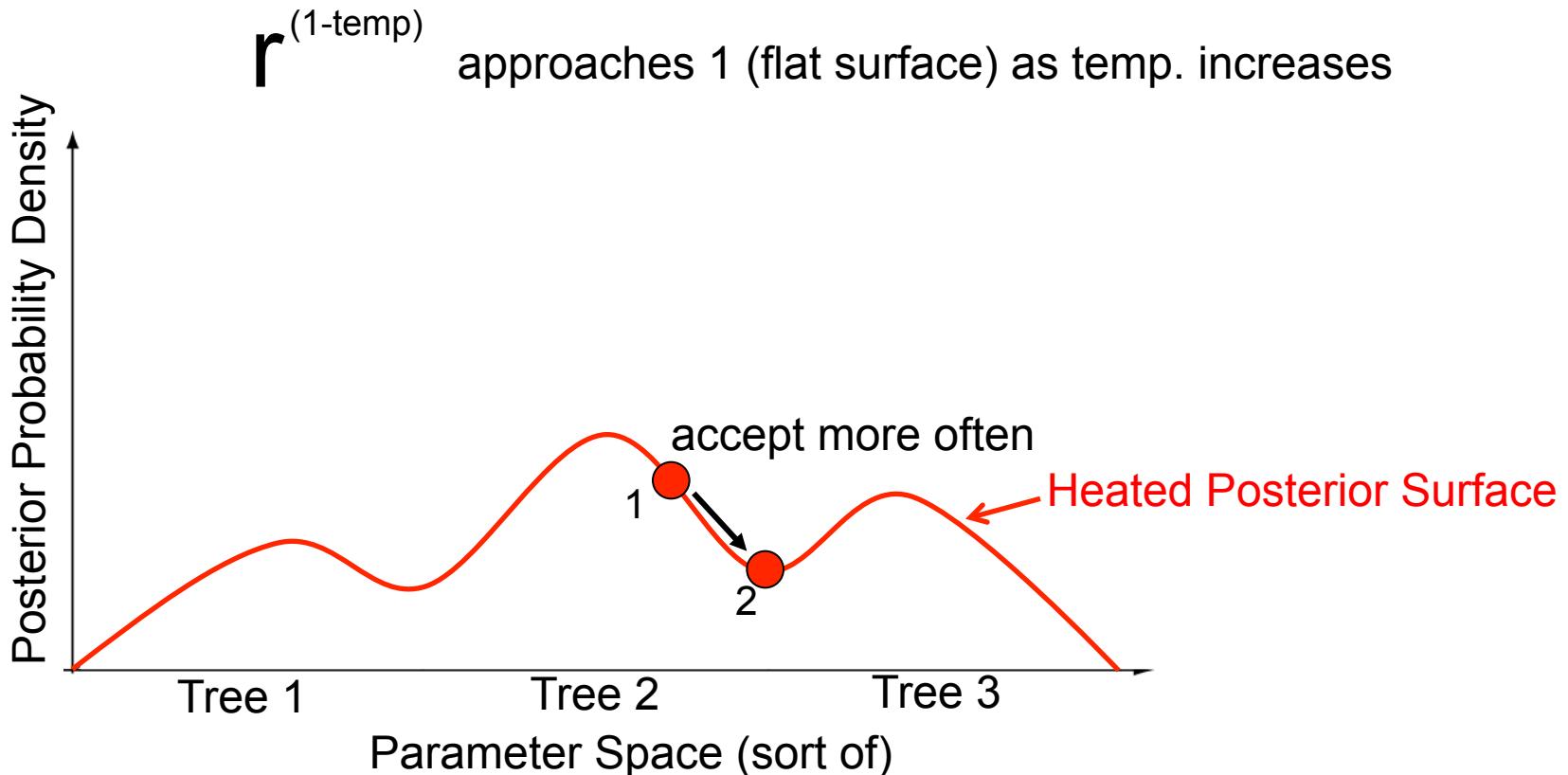
- Same rules as regular MCMC, but now there are multiple chains with different ‘temperatures’.
- ‘Heated’ chains sample a ‘melted’ version of the posterior
- Only difference is that heated chains raise the ratio of posterior densities to (1-temp) when deciding whether to accept a move.



This slide “borrowed” from F. Ronquist

Metropolis Coupling

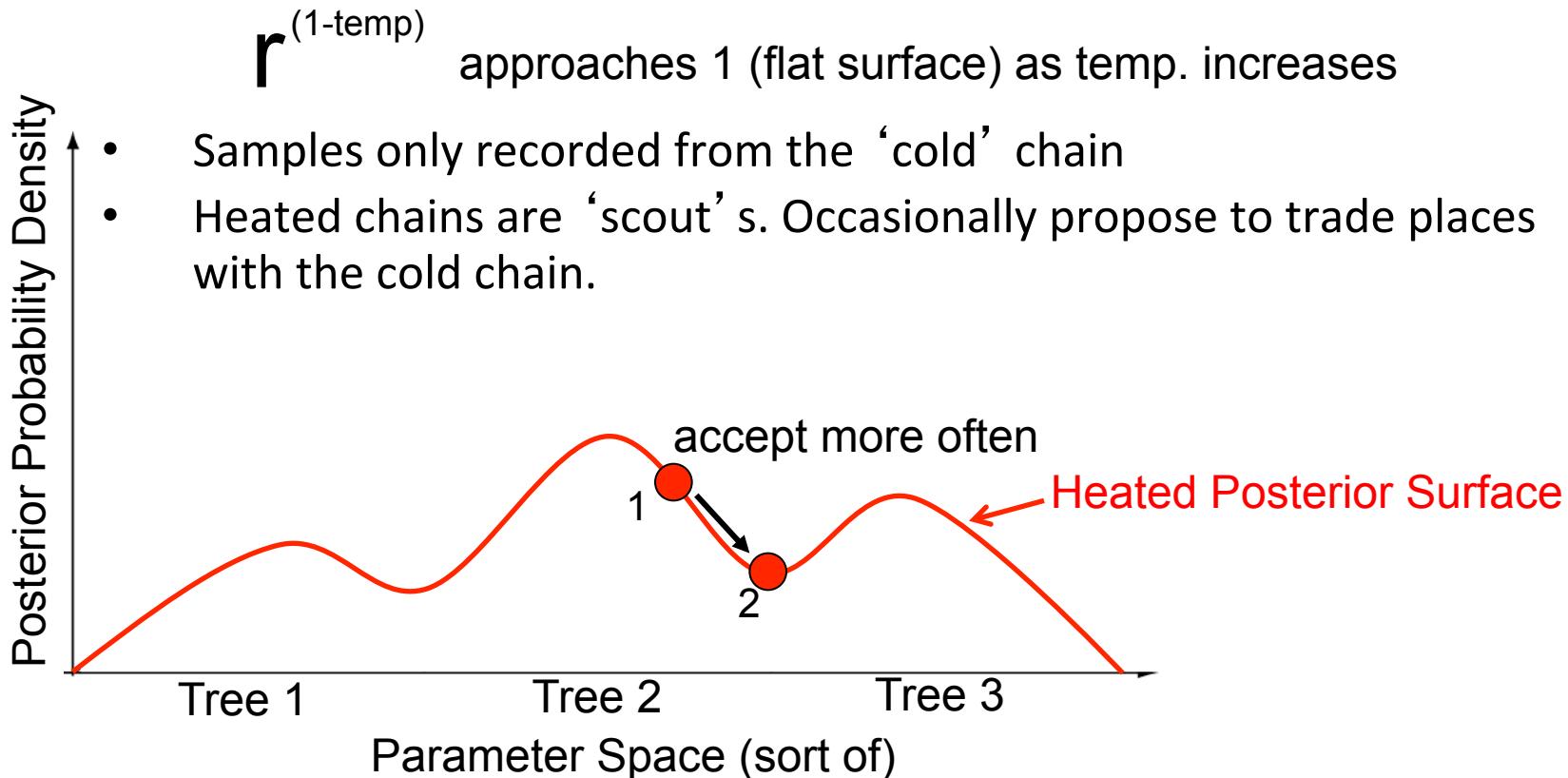
- Same rules as regular MCMC, but now there are multiple chains with different ‘temperatures’.
- ‘Heated’ chains sample a ‘melted’ version of the posterior
- Only difference is that heated chains raise the ratio of posterior densities to $(1-\text{temp})$ when deciding whether to accept a move.



This slide “borrowed” from F. Ronquist

Metropolis Coupling

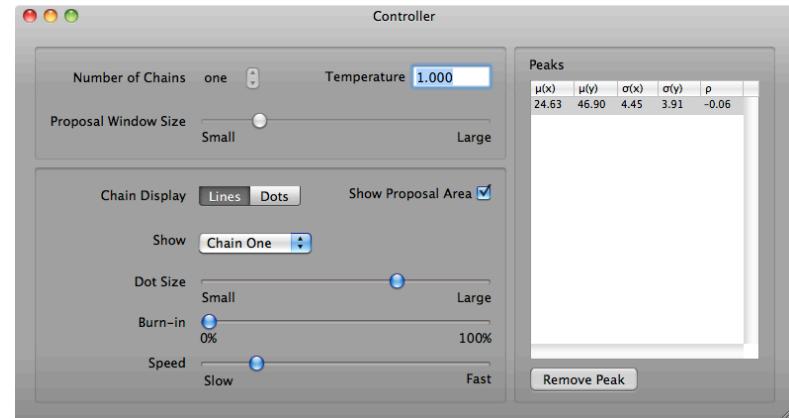
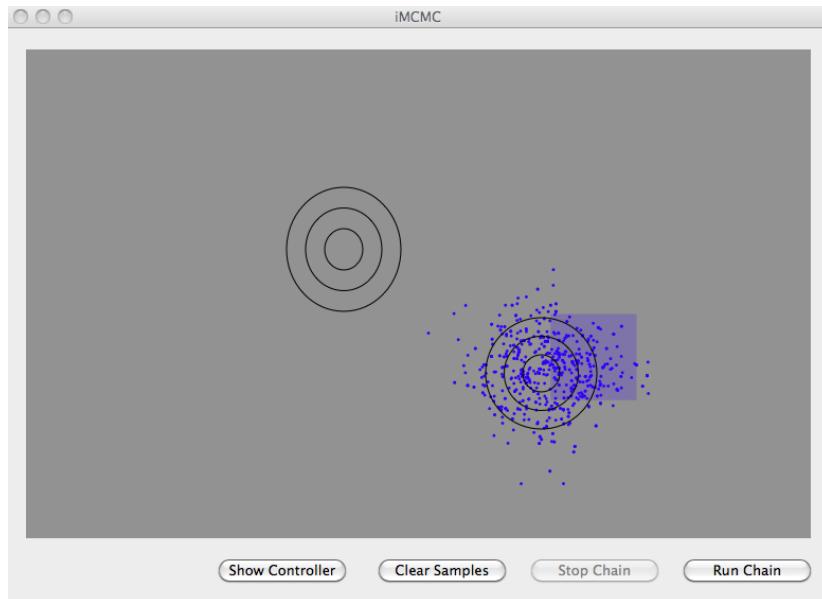
- Same rules as regular MCMC, but now there are multiple chains with different ‘temperatures’.
- ‘Heated’ chains sample a ‘melted’ version of the posterior
- Only difference is that heated chains raise the ratio of posterior densities to (1-temp) when deciding whether to accept a move.



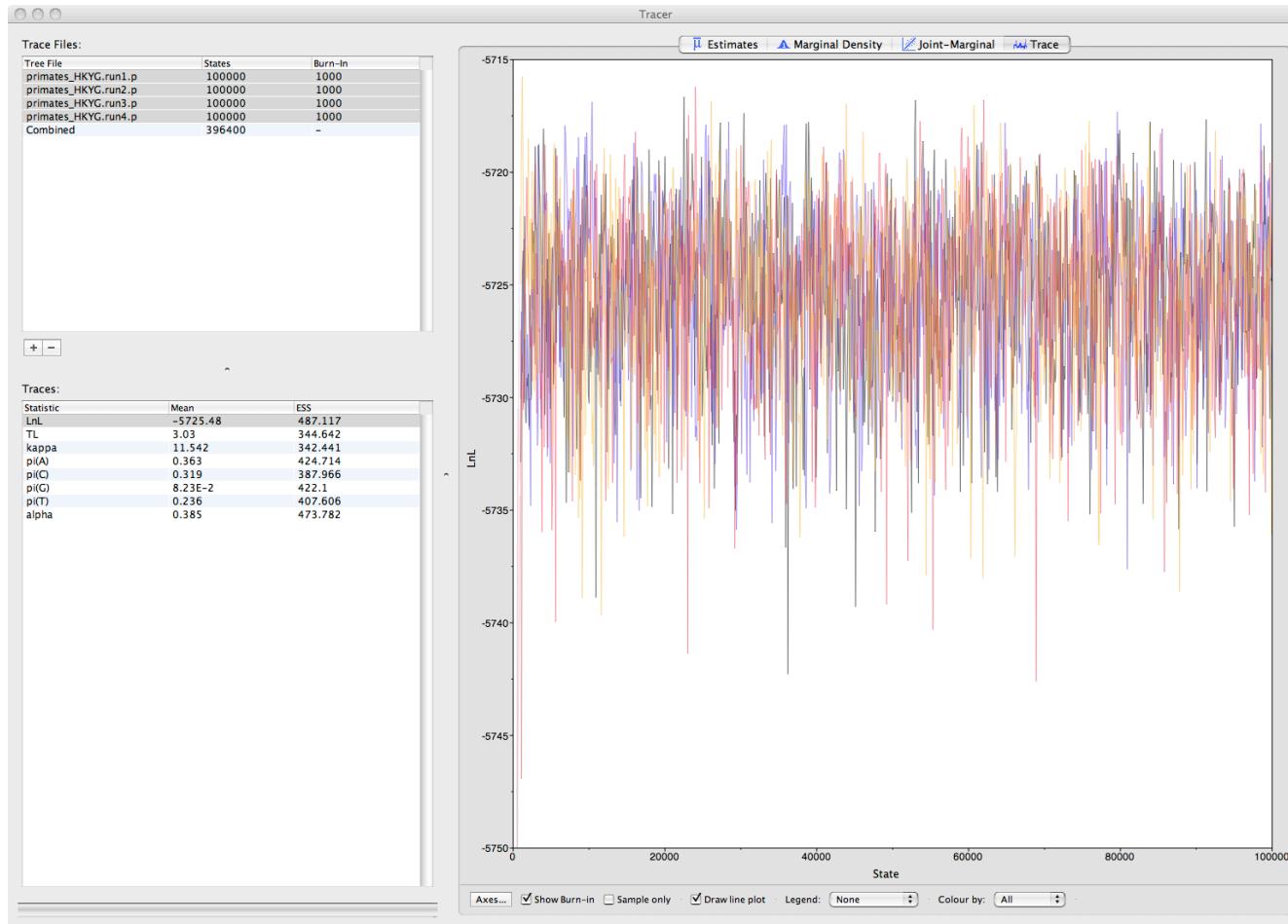
This slide “borrowed” from F. Ronquist

Toy MCMC Exploration

- MCRobot – PC (Lewis)
 - <http://www.eeb.uconn.edu/people/plewis/software.php>
- iMCMC – Mac (Huelsenbeck)
 - <http://fisher.berkeley.edu/cteg/software.html>



Convergence of Scalars - Tracer



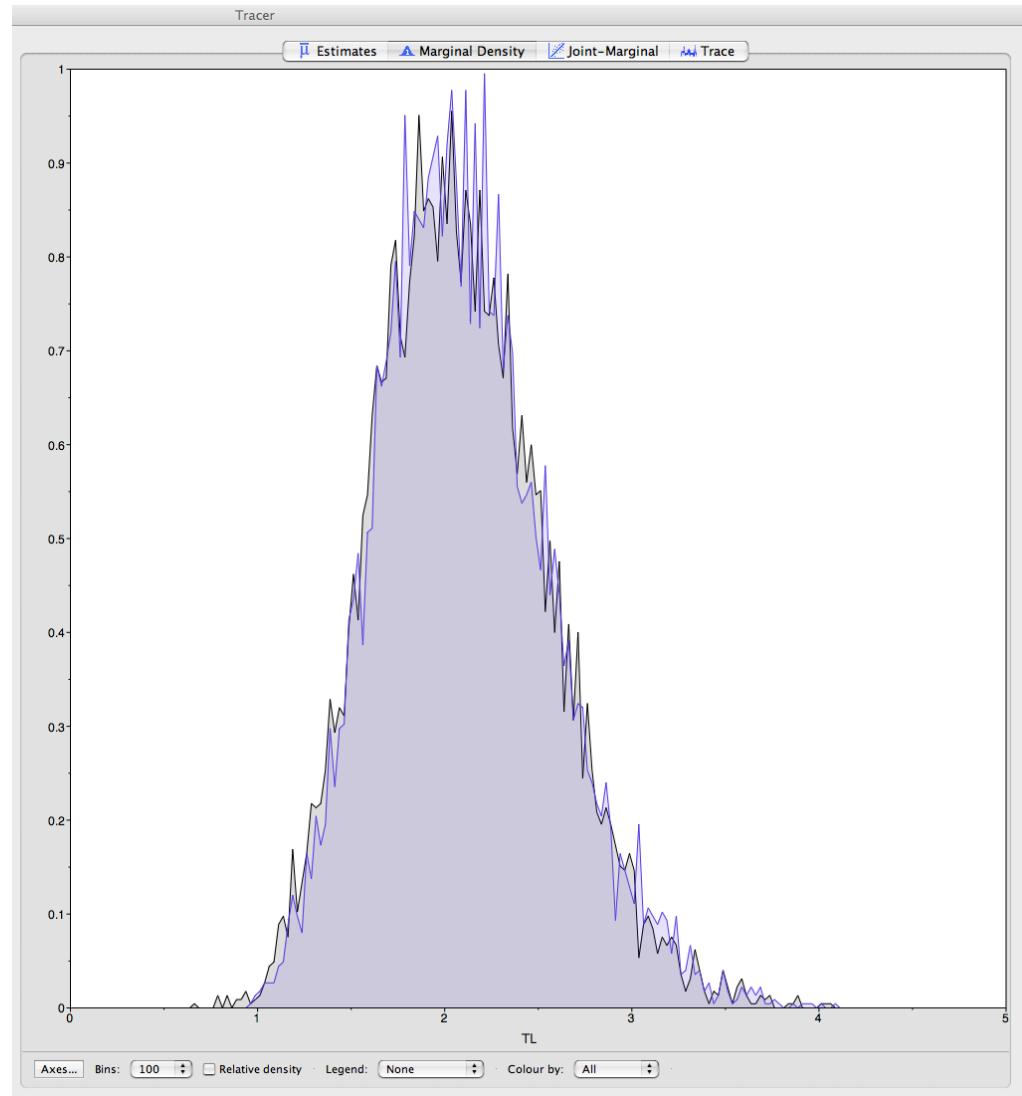
Running on Empty



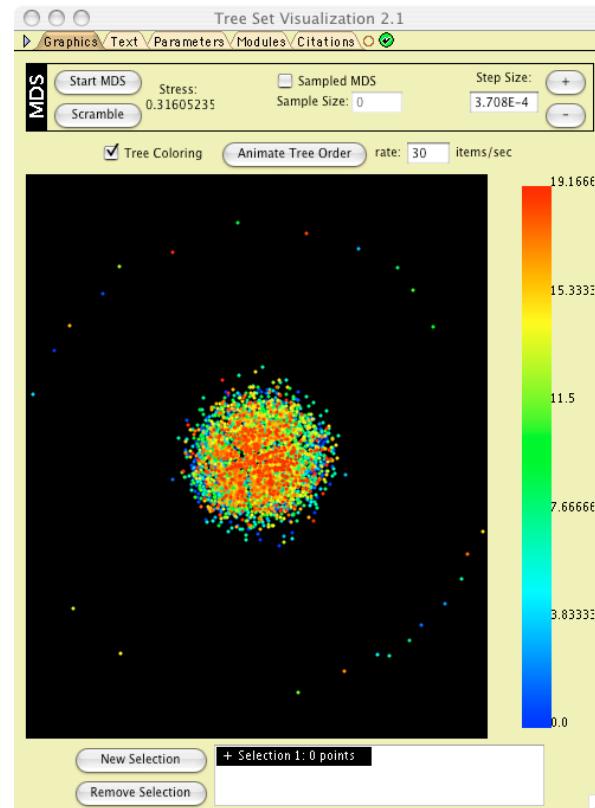
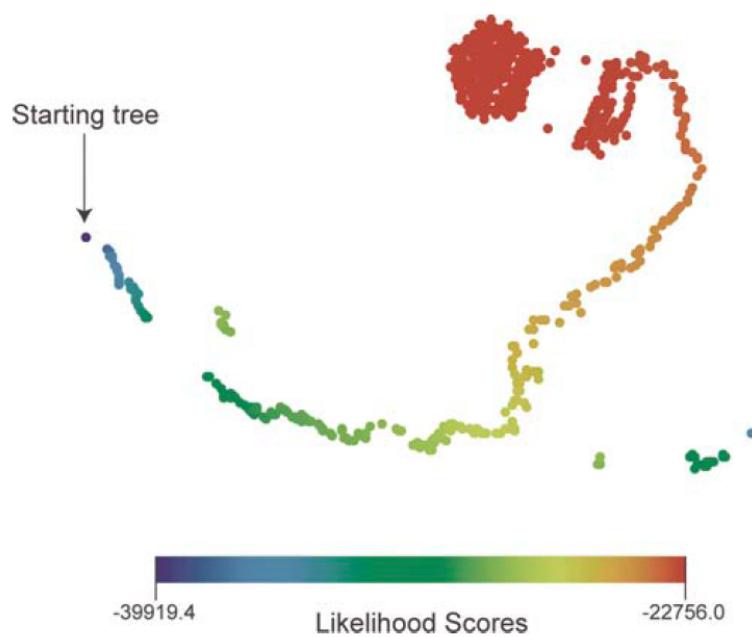
```
#NEXUS
begin data;
dimensions ntax=12 nchar=5;
format datatype=dna interleave=no gap=- missing=?;
matrix
Tarsius_syrichta    ??????
Lemur_catta          ??????
Homo_sapiens         ??????
Pan                  ??????
Gorilla              ??????
Pongo                ??????
Hylobates             ??????
Macaca_fuscata       ??????
M_mulatta            ??????
M_fascicularis      ??????
M_sylvanus           ??????
Saimiri_sciureus    ??????
;
end;
```

Or now in MrBayes:

mcmc data=no



Topological Convergence - TreeSetViz



TreeSetViz: <http://comet.lehman.cuny.edu/treeviz/>

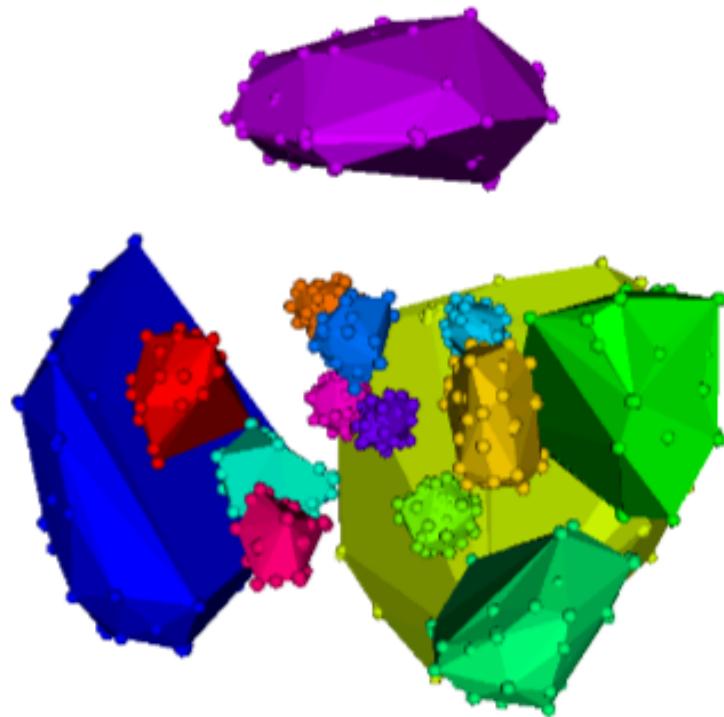
Mesquite: <http://www.mesquiteproject.org/mesquite/mesquite.html>

Hillis et al., Analysis and Visualization of Tree Space, *Syst. Biol.*, 54(3): 471-482.

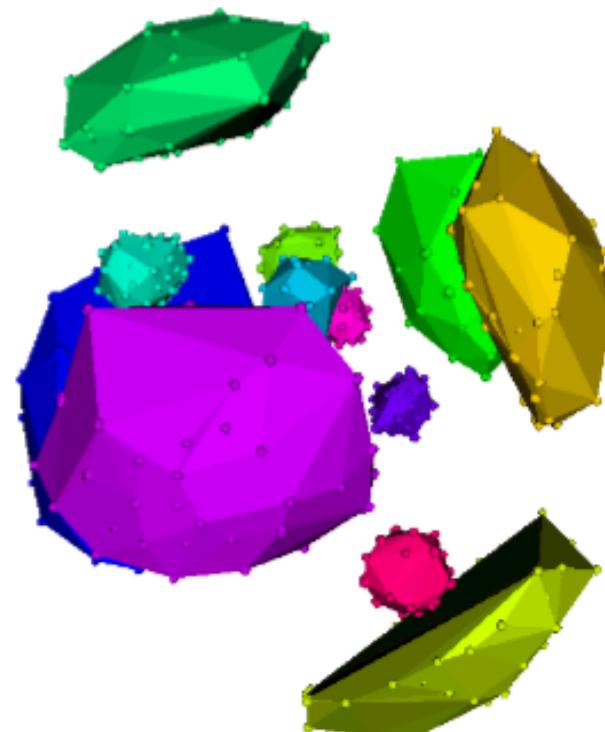
Topological Convergence - TreeScaper

Colors = Trees from Different Genes

Salamanders



Mammals



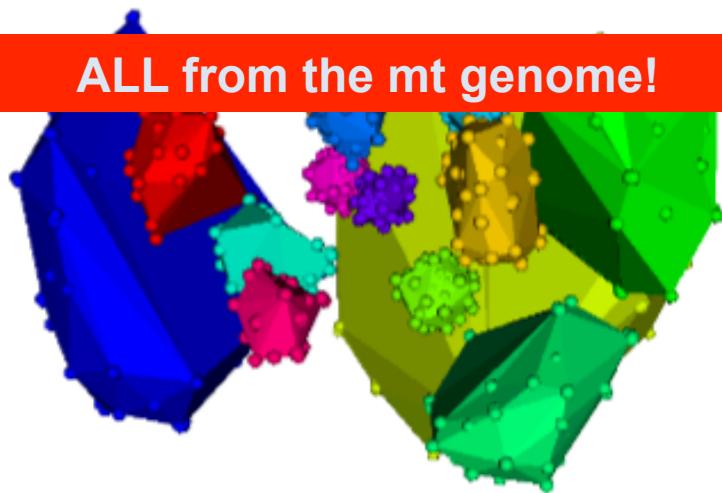
Topological Convergence - TreeScaper

Colors = Trees from Different Genes

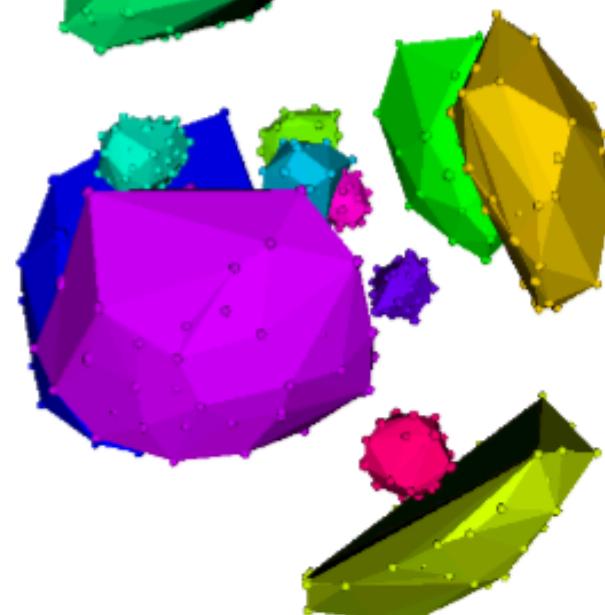
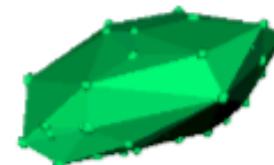
Salamanders

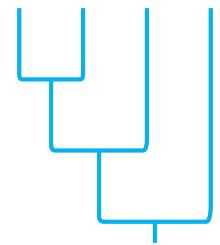


ALL from the mt genome!



Mammals

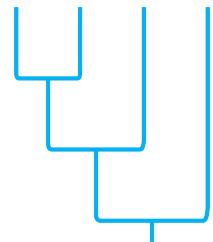




MrBayes - Installation

Available at: <http://mrbayes.sourceforge.net/>

- Executable versions for PC and Mac
- Source code for compilation (command-line)
- Serial, Parallel, and GPU-enabled (via Beagle)
- Nice manual (in progress), tutorials, command reference
- Current version: 3.2.2
 - Complete re-write (RevBayes) on its way



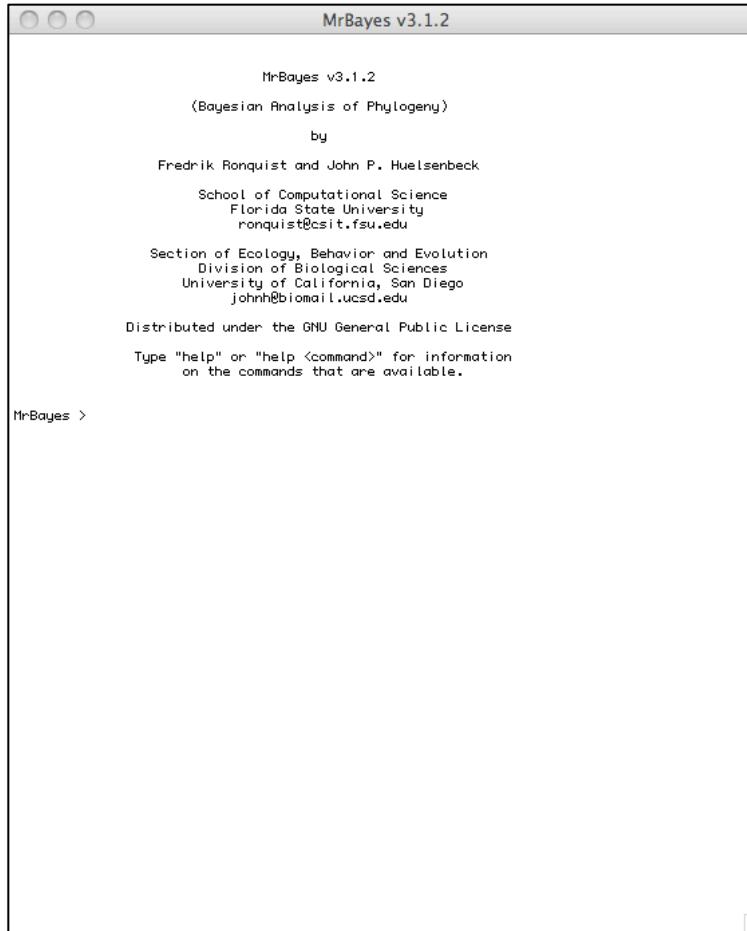
MrBayes – Fire It Up

- Double click on icon or use command line in Unix

> ./mb

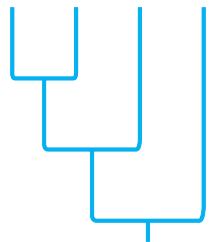
- or -

> mb





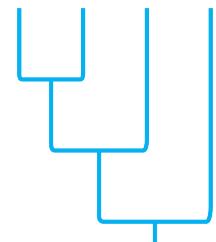
MrBayes – Command Line



- If compiled locally, you can run MrBayes from the command line (e.g., Terminal for Mac OS X)
`> mb mb_cmds.nex`
- Facilitates batch analyses, especially on a computing cluster



MrBayes with Beagle



If you have an NVIDIA graphics card, you can take advantage of GPU-based parallelization via the Beagle library. Especially helpful for codon and amino acid models.

How To:

- Check your graphics card (on Mac: “About This Mac”)
- Download and install CUDA drivers from NVIDIA
[\(http://www.nvidia.com/Download/index.aspx?lang=en-us\)](http://www.nvidia.com/Download/index.aspx?lang=en-us)
- Check that MrBayes sees it (*showbeagle*)
- Turn it on (set `beagledevice=gpu` `usebeagle=yes`)

See if you can get MrBayes up and running.

Try loading your data.

Set a useful output file name
("mcmcp filename=*data_JC*").

Then type "mcmc".

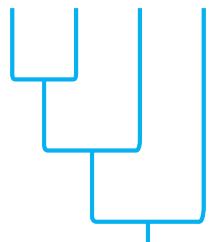
Look at output files.

Run sump, then sumt.

Bayesian Model Selection and Adequacy



Don't Choose!



Let the analysis sample different models for you (called reversible jump – RJ)

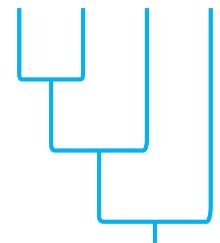
To sample sub-models of GTR:

MrBayes > lset nst=mixed

Model	Posterior Probability	Standard Deviation	Min. Probability	Max. Probability
<hr/>				
gtrsubmodel[121123]	0.195	0.033	0.172	0.219
gtrsubmodel[121324]	0.089	0.005	0.086	0.093
gtrsubmodel[121323]	0.079	0.009	0.073	0.086
gtrsubmodel[121321]	0.060	0.037	0.033	0.086
gtrsubmodel[123324]	0.060	0.000	0.060	0.060
gtrsubmodel[121121]	0.053	0.009	0.046	0.060
<hr/>				



Posterior Odds Ratio



Prior
Odds

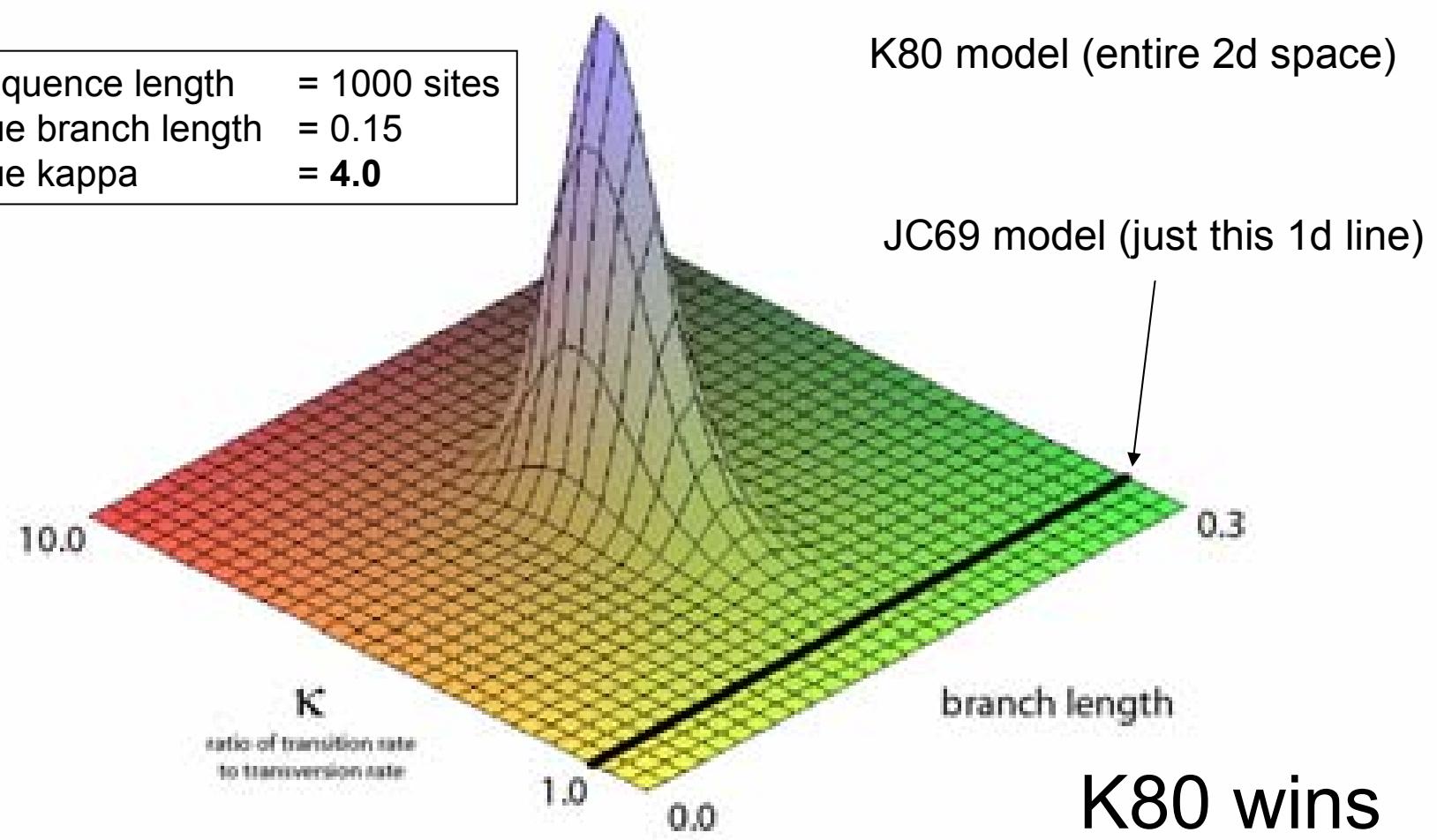
Bayes
Factor

Posterior
Odds

$$\frac{P(M_1) \cdot P(\text{[color grid]} | M_1)}{P(M_2) \cdot P(\text{[color grid]} | M_2)} = \frac{P(M_1 | \text{[color grid]})}{P(M_2 | \text{[color grid]})}$$

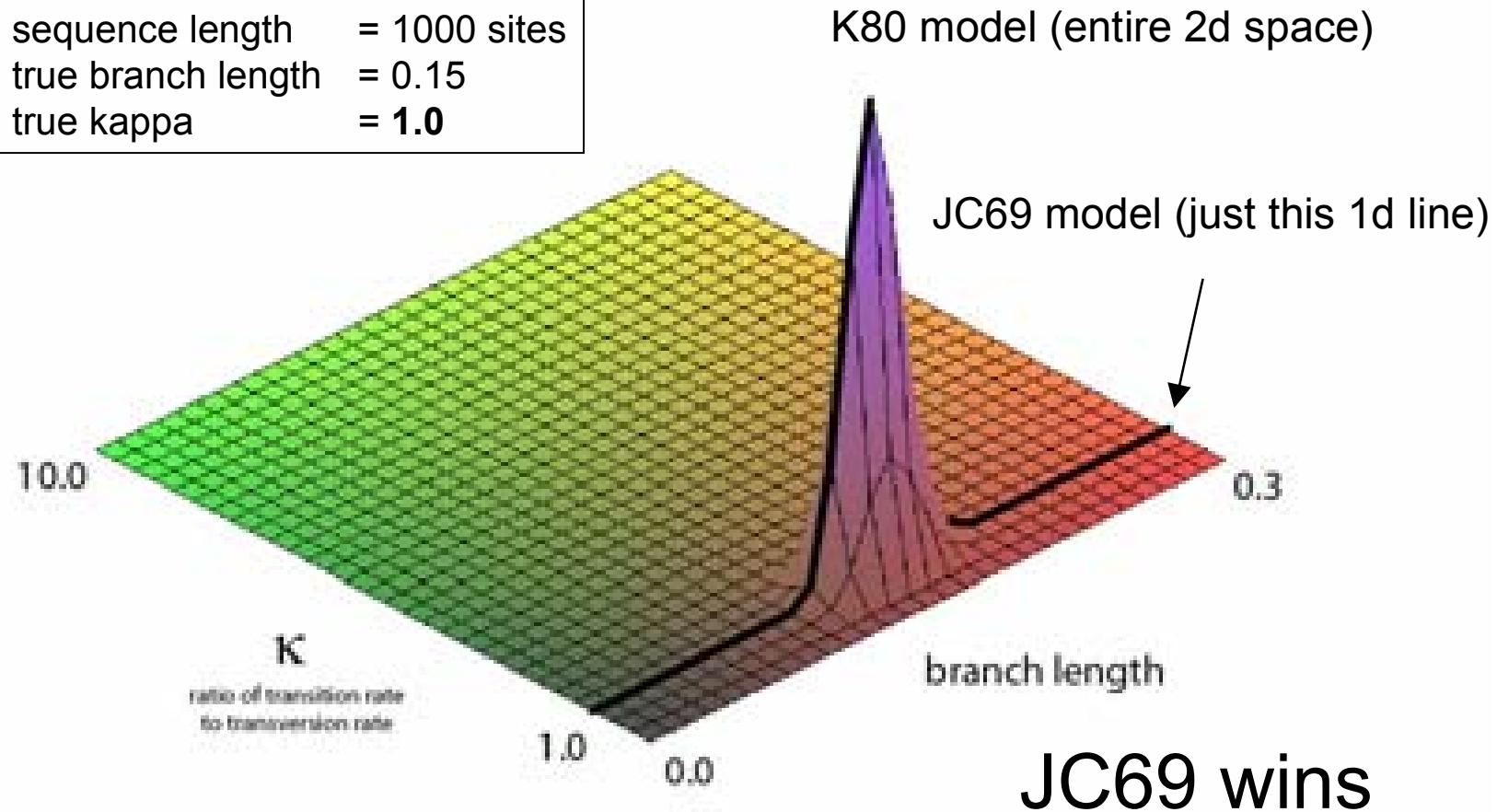
Marginal Likelihood of a Model

sequence length	= 1000 sites
true branch length	= 0.15
true kappa	= 4.0



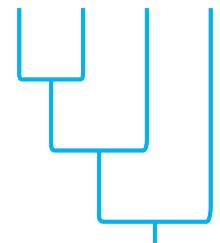
Marginal Likelihood of a Model

sequence length	= 1000 sites
true branch length	= 0.15
true kappa	= 1.0





Posterior Odds Ratio



Prior
Odds

Bayes
Factor

Posterior
Odds

$$\frac{P(M_1) \cdot P(\text{[color grid]} | M_1)}{P(M_2) \cdot P(\text{[color grid]} | M_2)} = \frac{P(M_1 | \text{[color grid]})}{P(M_2 | \text{[color grid]})}$$

Bayes Factor

Posterior Odds

Bayes Factor

$$\frac{P(\text{grid pattern} | M_1)}{P(\text{grid pattern} | M_2)} = \frac{\frac{P(M_1 | \text{grid pattern})}{P(M_2 | \text{grid pattern})}}{\frac{P(M_1)}{P(M_2)}}$$

Prior Odds

Bayes Factor

Bayes Factor

$$\frac{P(\text{grid} | M_1)}{P(\text{grid} | M_2)} = \frac{\text{Posterior Odds}}{\text{Prior Odds}}$$

$P(M_1 | \text{grid})$
 $P(M_2 | \text{grid})$

What should be true
in the case of equal
prior probabilities?

Prior
Odds

Interpreting Bayes Factors

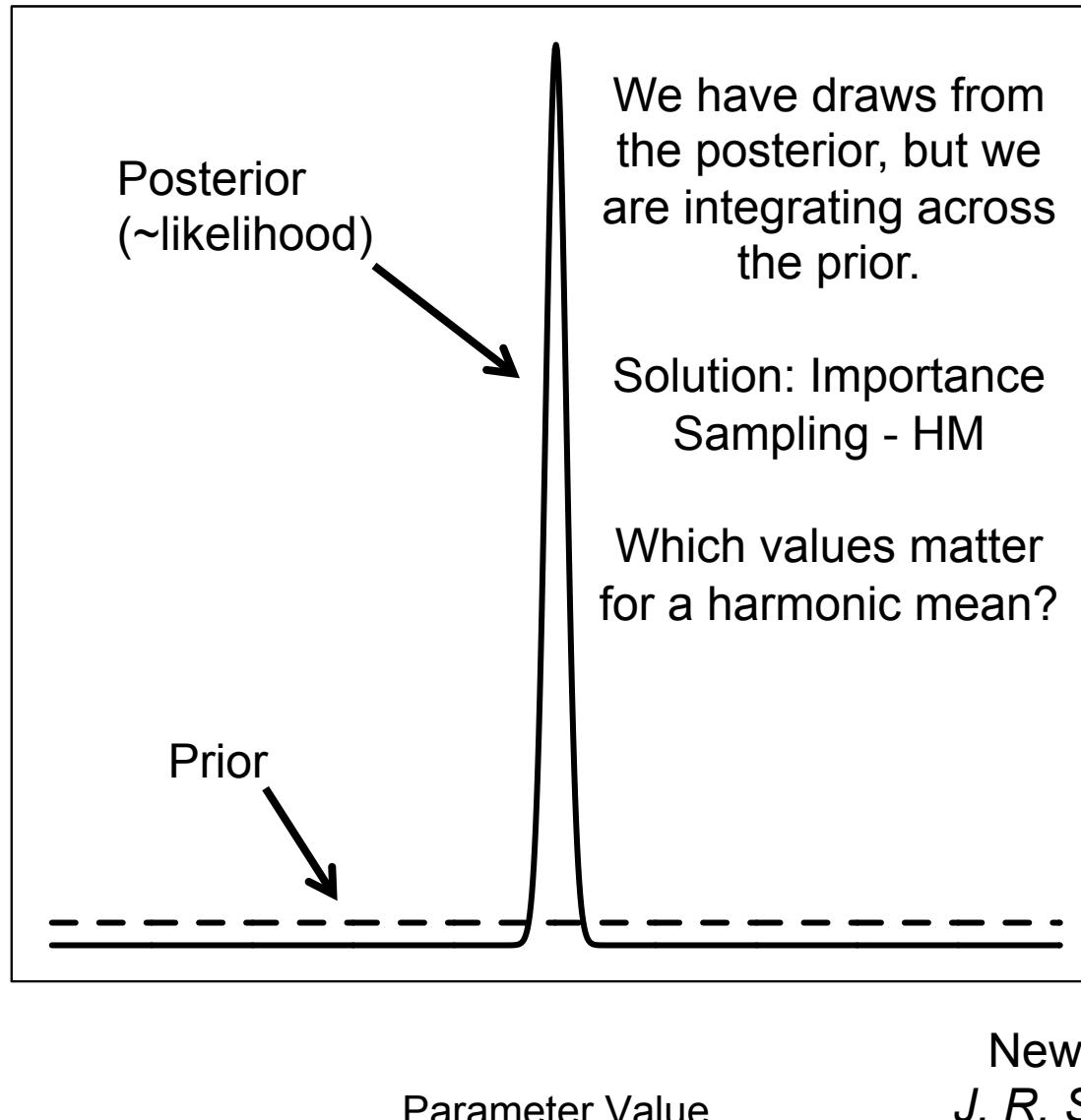
- BFs greater than 1 indicate support for model 1 (numerator), less than 1 indicate support for model 2 (denominator)
- $\log(BF)$ values greater than 0 support model 1 and less than 0 support model 2
- Practically, $\log(BF)$ values of > 5 or < -5 are often used to indicate very strong support for the corresponding model.

Estimating Bayes Factors

- With RJ, use posterior & prior odds ratios
- Can also calculate marginal likelihoods directly
 - Harmonic mean (easy, but inaccurate)
 - Stepping stone (more accurate, but harder)
 - Thermodynamic integration (similar to SS)
 - MORE SOON!

Harmonic Mean Estimator

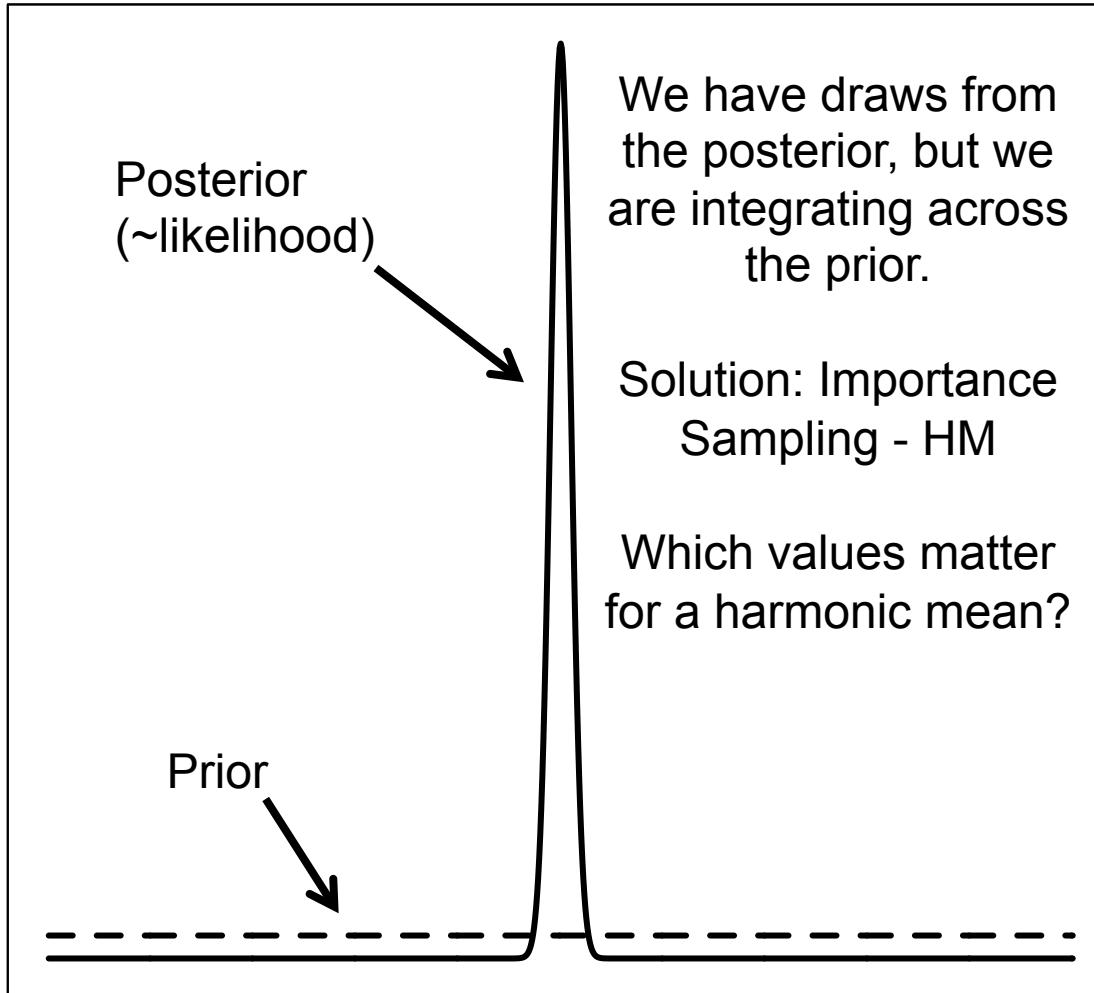
Probability Density



Newton and Raftery. 1994.
J. R. Statistic. Soc. B 56:3-48.

Harmonic Mean Estimator

Probability Density



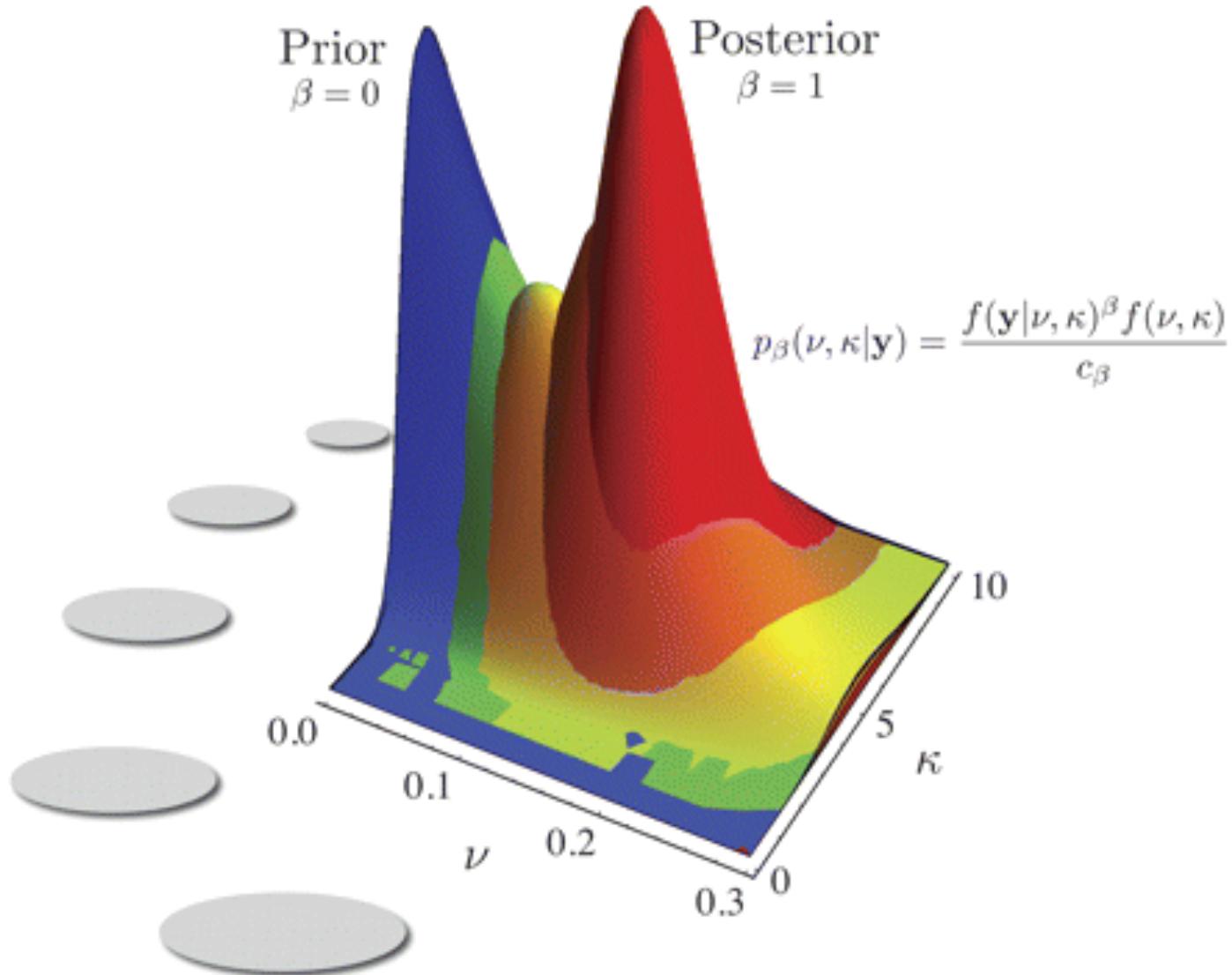
Look up the HM estimate of the marginal likelihood in the .lstat file from your MrBayes run with a Jukes-Cantor model (after running sump) and record it.

Newton and Raftery. 1994.
J. R. Statistic. Soc. B 56:3-48.

Harmonic Mean

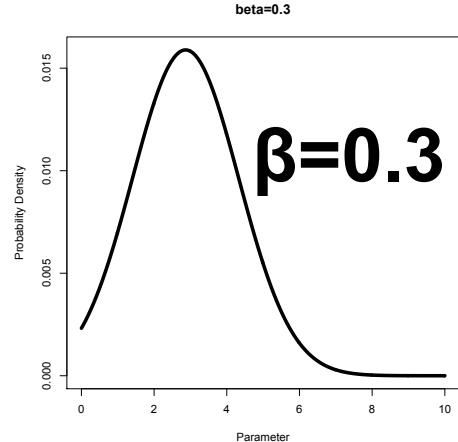
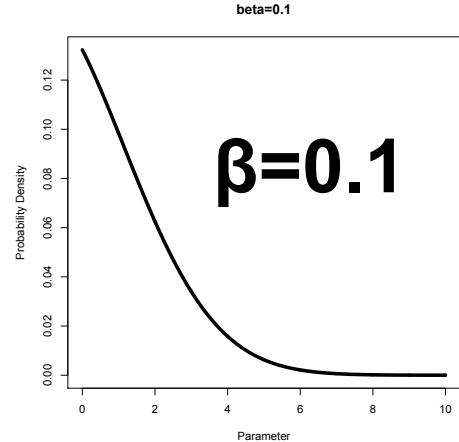
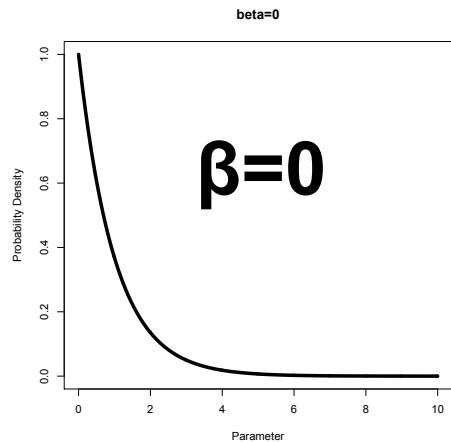
$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}}$$

Stepping Stone Sampling

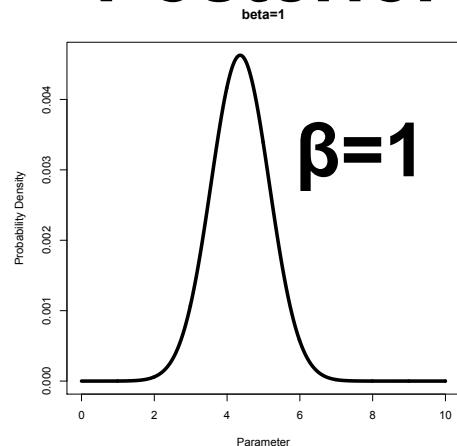
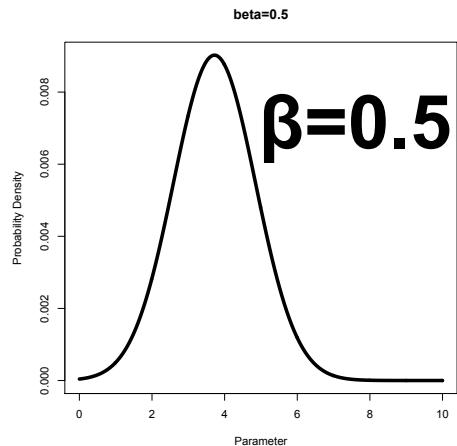


Power Posteriors

Prior

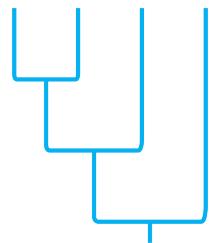


Posterior





MrBayes – Stepping Stone



`ss` and `ssp`

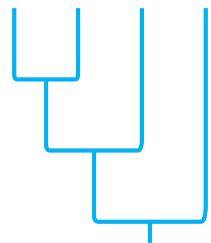
`ssp` allows you to set the parameters of the stepping stone process without running. `ss` does the same, but starts the run.

Try the `ss` command with a UCE locus...

MrBayes > ss



MrBayes – Stepping Stone



sumss

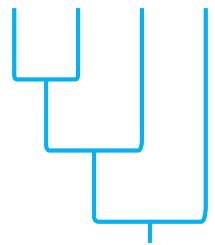
sumss allows you to examine plots of likelihoods across steppingstone steps to look for ‘temperature lag’ (step-specific burn-in).

Try the *sumss* command with a UCE locus...

MrBayes > sumss

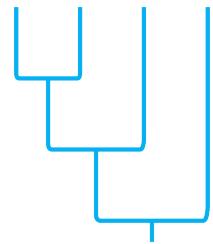


MrBayes – Stepping Stone



Now try calculating the marginal likelihood of a Jukes-Cantor model with stepping stone and record it (see end of ss log file).

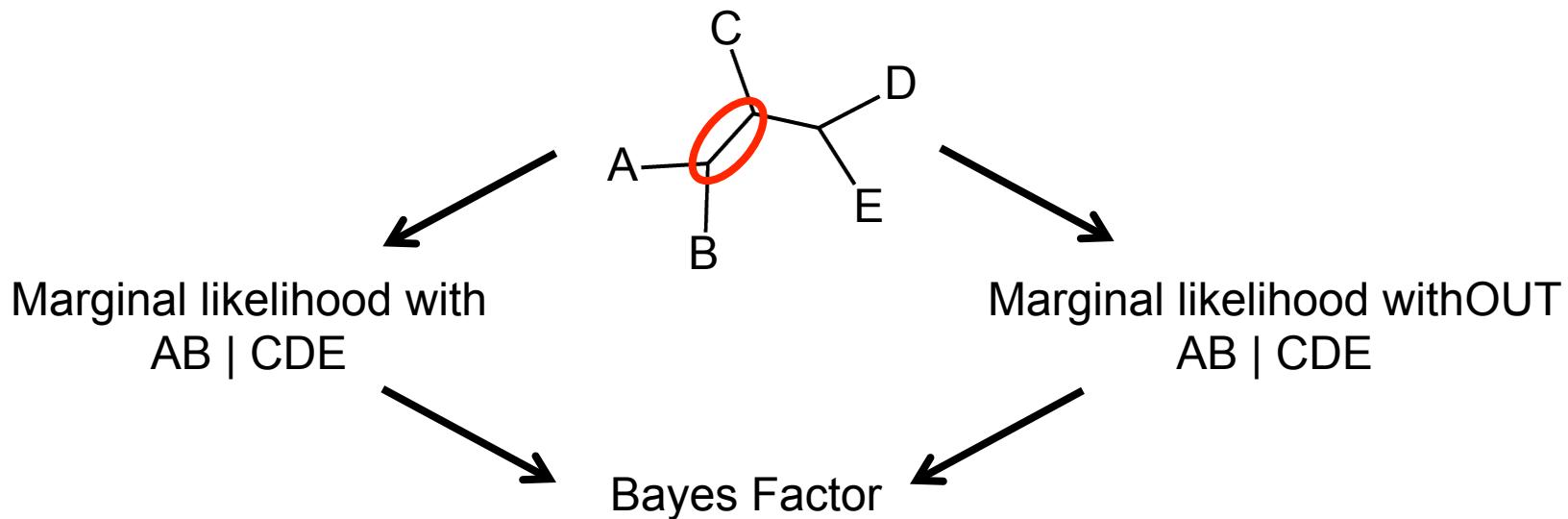
Now do it for GTR.



Topology BFs

Can also use stepping stone to calculate marginal likelihoods for sets of trees subject to particular constraints.

For instance, comparing the marginal likelihoods for a set of trees that all contain a branch to the set of trees that all do NOT contain that branch would give you a Bayes factor supporting that branch.





Model Choice v “Adequacy”

Which of the available models will perform best?

vs.

Is any given model sufficient to provide unbiased
inferences?



Model Choice v “Adequacy”

Or, better, **Plausibility**

Which of the available models will perform best?

vs.

Is any given model sufficient to provide unbiased
inferences?

Posterior Prediction



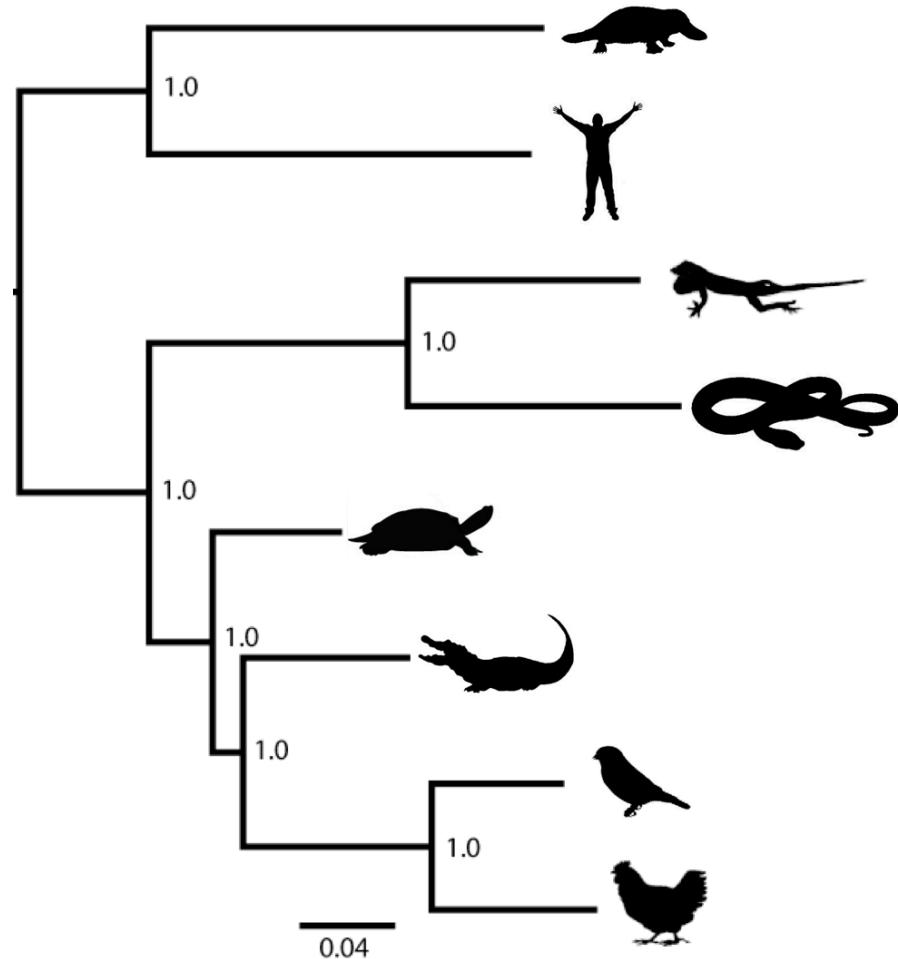
Could  have come from $P(\text{ } \text{ } \text{ } , \theta | \text{ } \text{ } \text{ })$?

Could the model and priors plausibly have given rise to the data?

If not, phylogenies may be **biased**

Big Data = Strong Support

1,955 Genes
Bayesian Tree



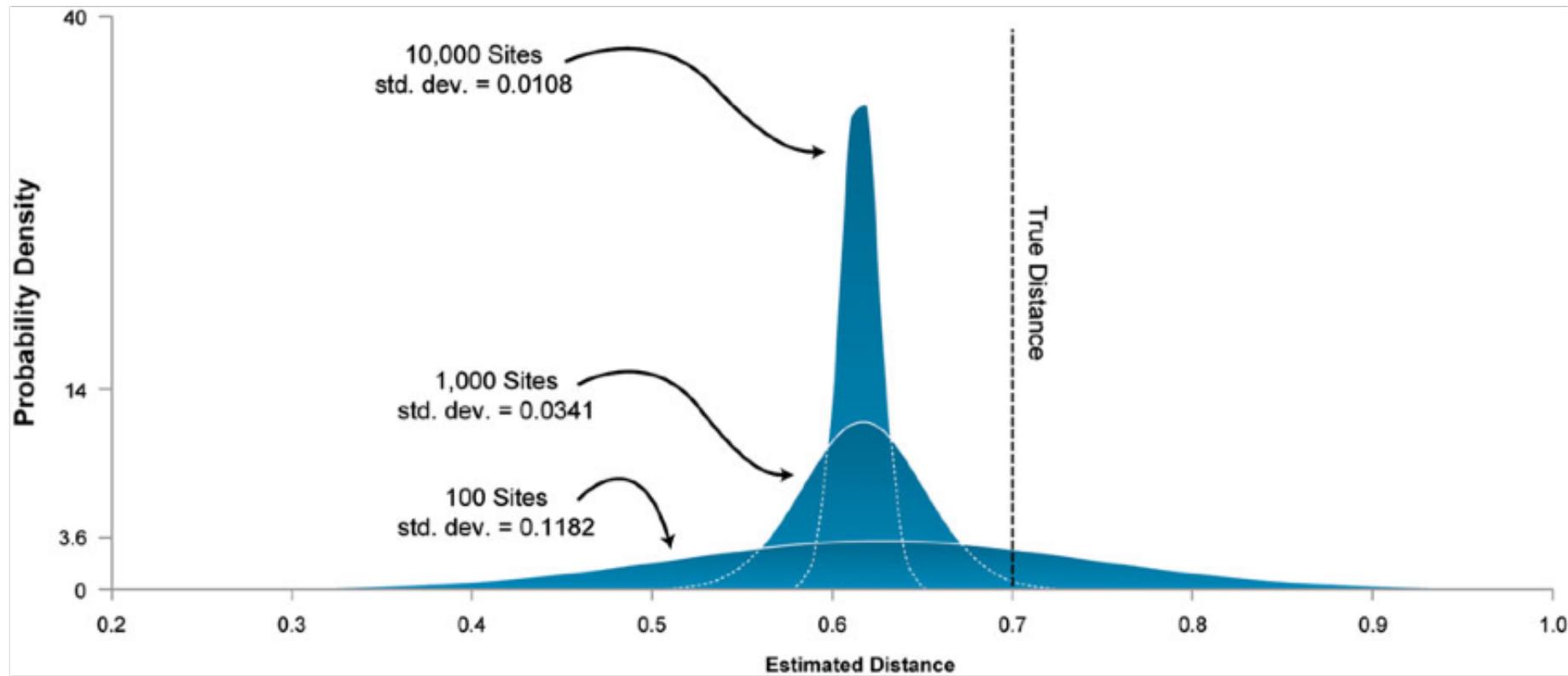
From Bob Thomson (Painted Turtle Genome Project)

Big Data = Strong Support

1,955 Genomes
Bayesian

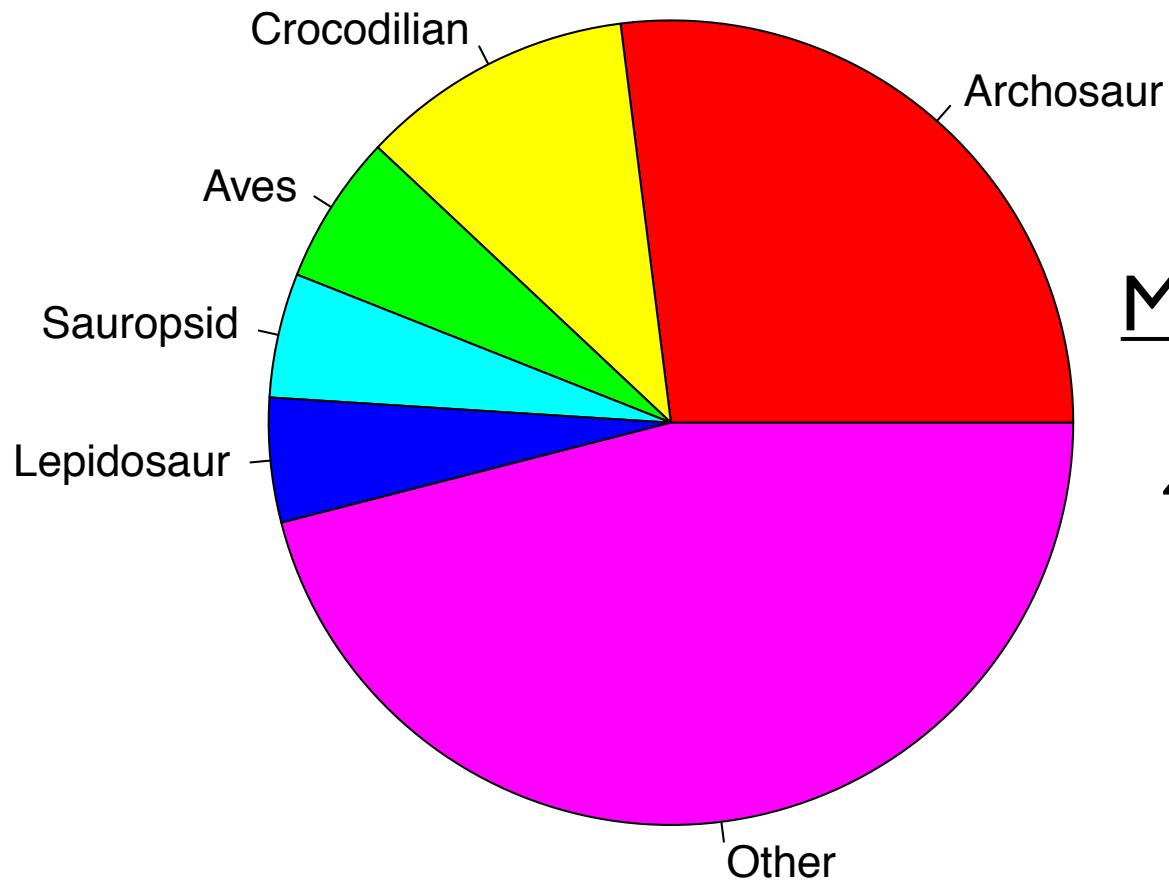


From Bob Thomson (Painted Turtle Genome Project)



Kumar et al. 2012. *Mol. Biol. Evol.*

Diff' Genes - Diff' Trees



Many genes, few taxa

1,144 Genes

4-8 Taxa Per Gene

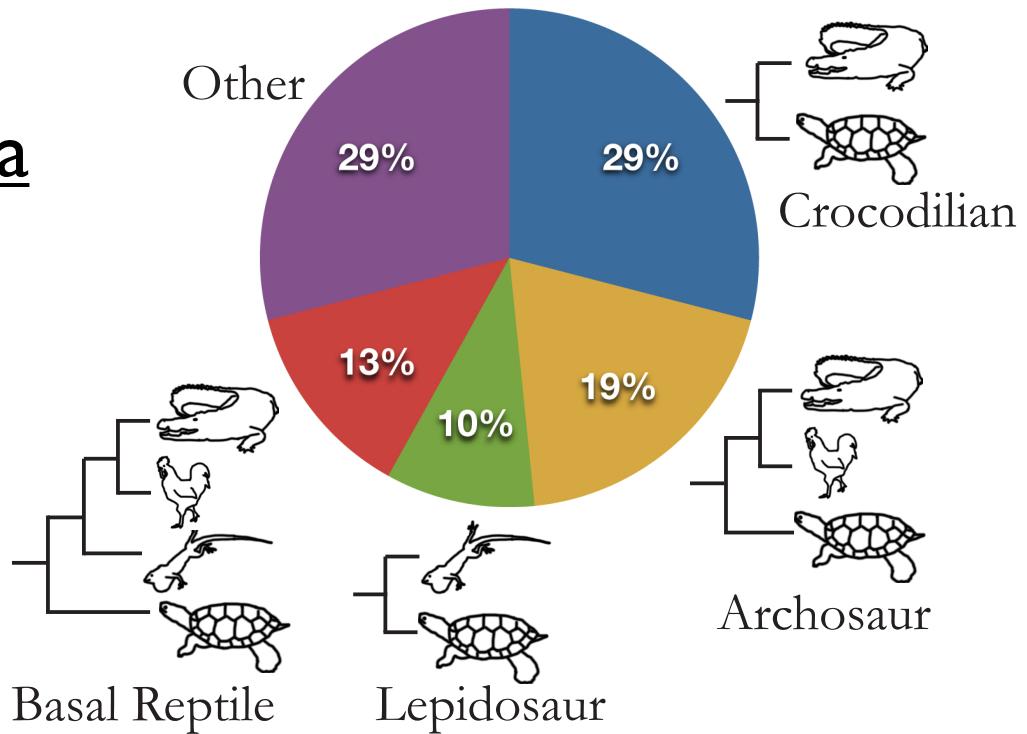
Diff' Genes - Diff' Trees

Few genes, many taxa

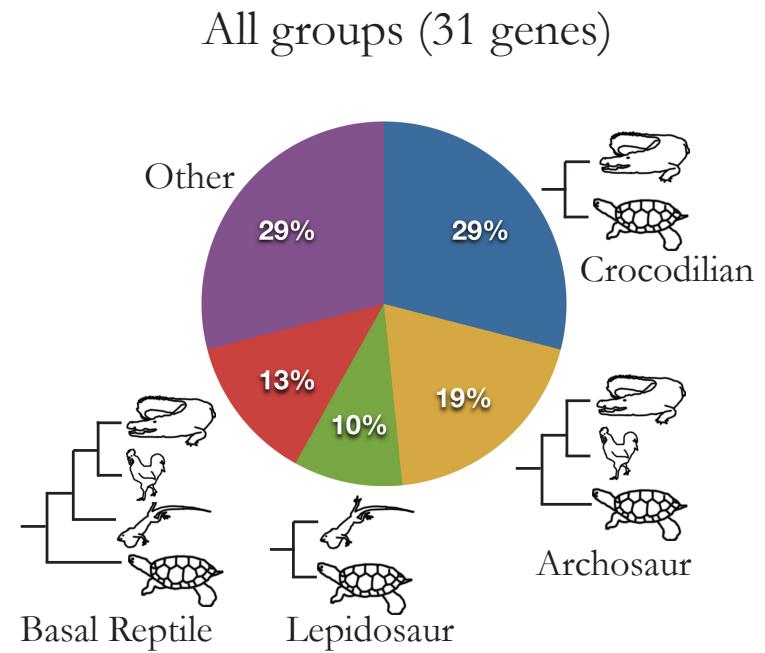
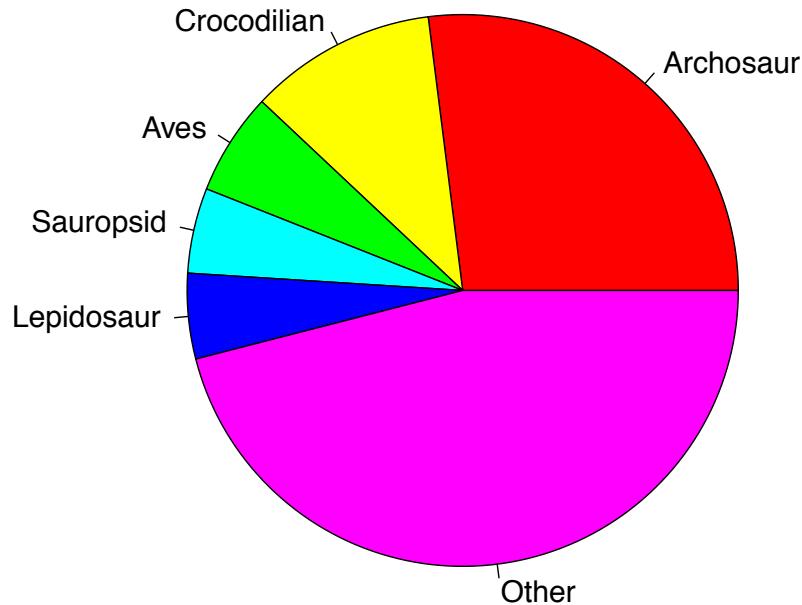
129 Taxa

31 genes

All groups (31 genes)



Diff' Genes - Diff' Trees



Biological or Methodological Variation?

Posterior Prediction



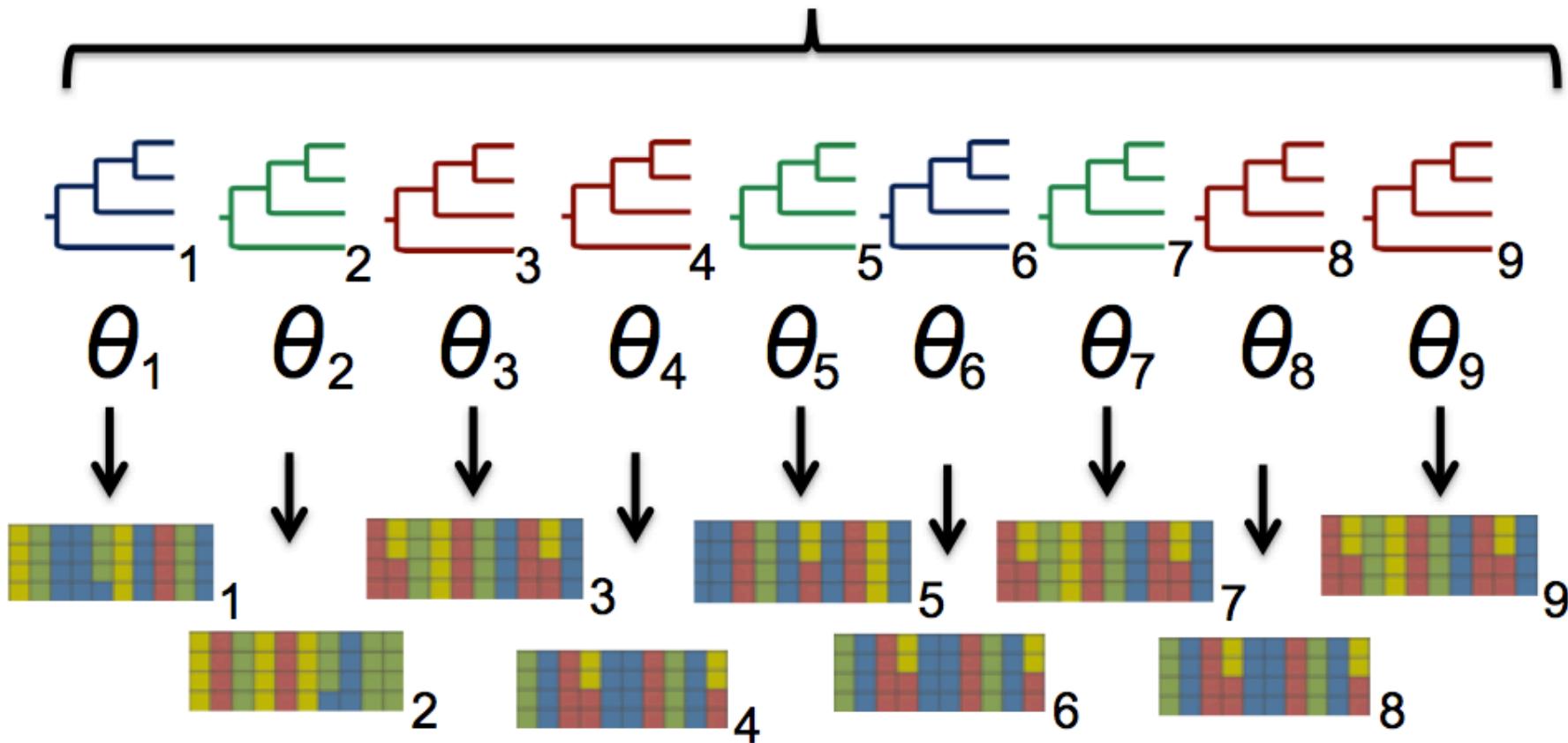
“We do not like to ask, ‘Is our model true or false?’, since probability models in most analyses will not be perfectly true...The more relevant question is, ‘Do the model’s deficiencies have a noticeable effect on the substantive inferences?’ ”

-A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin
Bayesian Data Analysis

Posterior Prediction



$$P(\text{ } \text{ } \text{ } \text{ } , \theta | \text{ } \text{ } \text{ } \text{ })$$



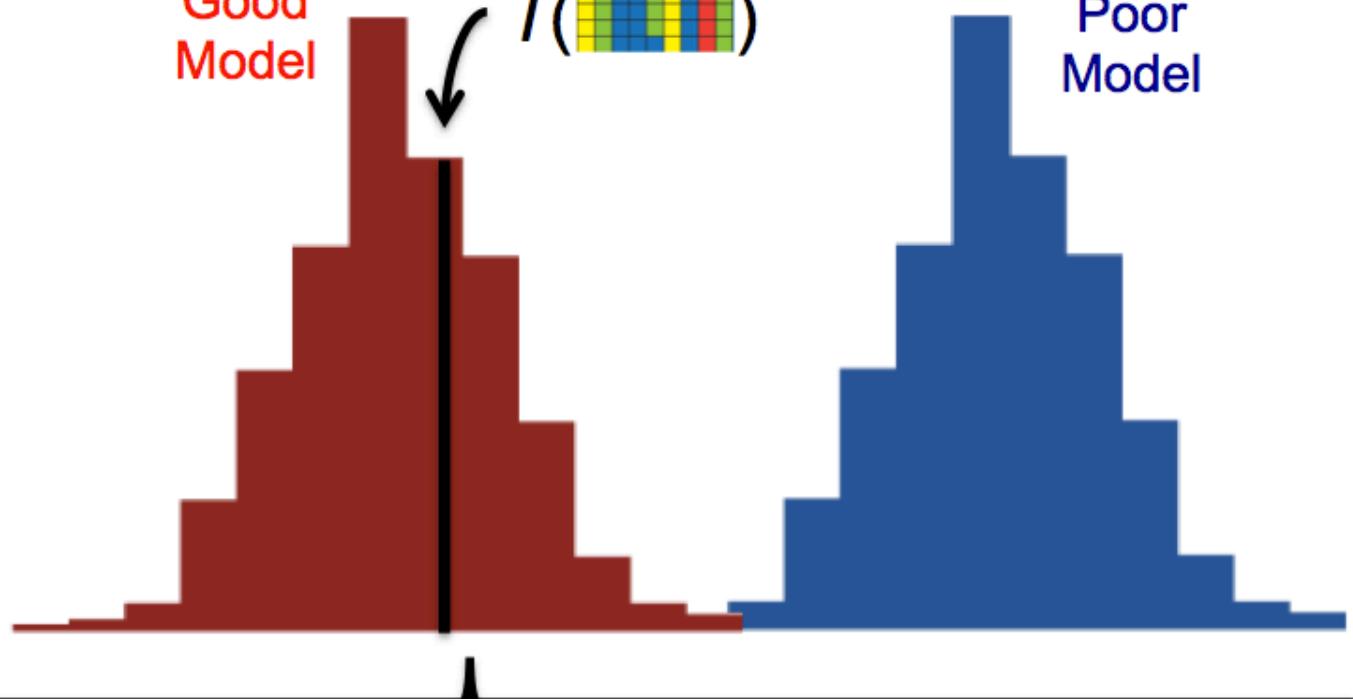
Posterior Prediction



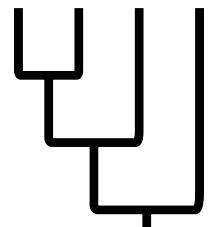
Good
Model

$T(\text{grid})$

Poor
Model

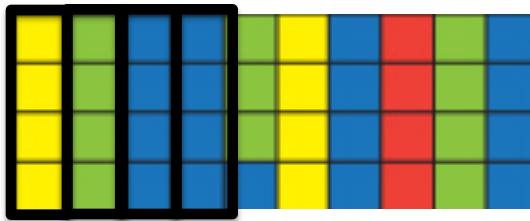


$T(\text{grid}_1) T(\text{grid}_3) T(\text{grid}_5) T(\text{grid}_7) T(\text{grid}_9)$
 $T(\text{grid}_2) T(\text{grid}_4) T(\text{grid}_6) T(\text{grid}_8)$

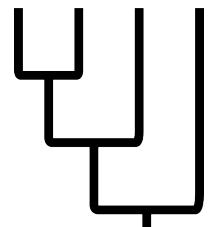


Posterior Predictive Simulation

Previously proposed statistic*:
Multinomial Likelihood

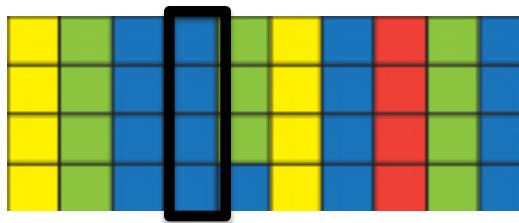


Based on the frequency of different site patterns



Posterior Predictive Simulation

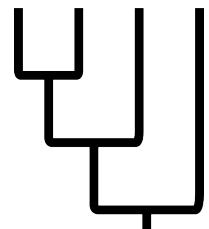
Previously proposed statistic*:
Multinomial Likelihood



Based on the frequency of different site patterns

Tests if the assumed and generating models produce data with similar site pattern frequencies

Intuitively appealing, but very sensitive to branch-length biases.
Can reject adequacy, **even when inferred phylogeny is correct**

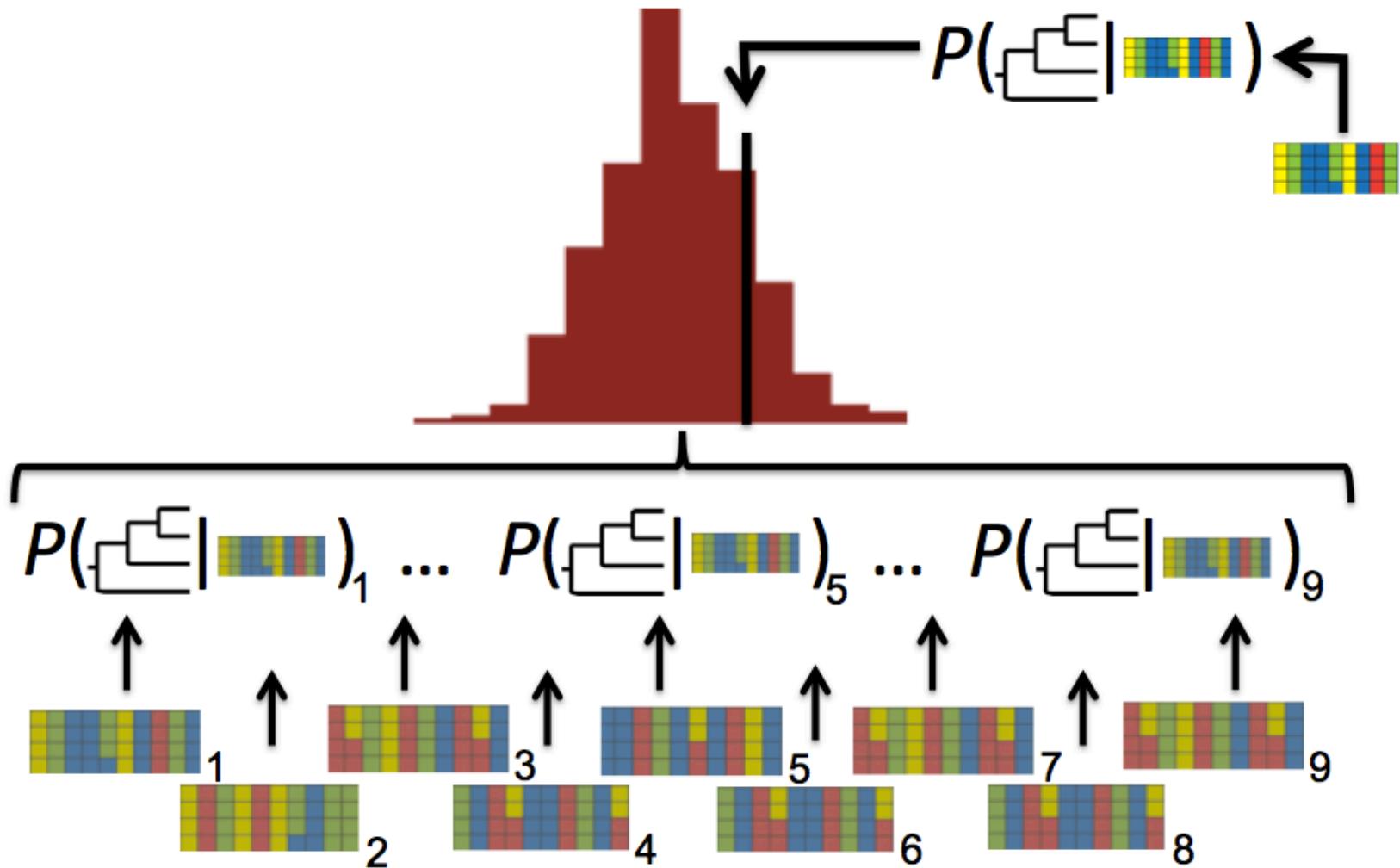


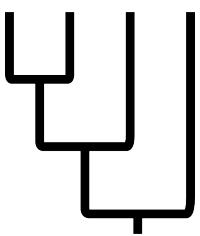
Posterior Predictive Simulation

Previously proposed statistics:

- Multinomial Likelihood
- Number of Unique Site Patterns
- Frequency of Invariant Sites
- Heterogeneity of Base Frequencies
- Number of parsimony-inferred
“parallel” sites

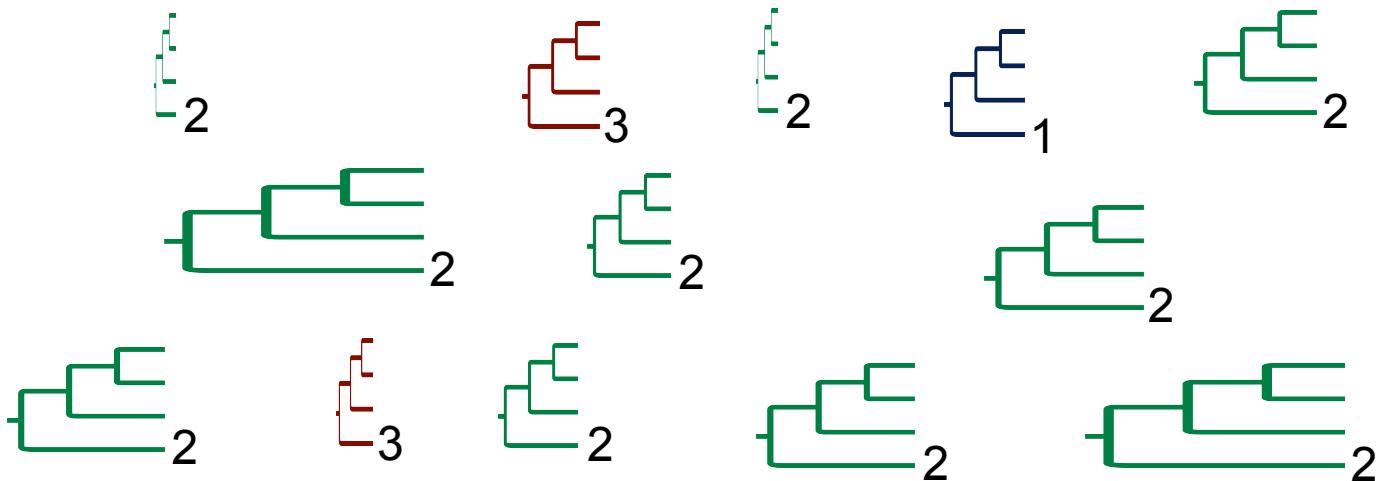
Posterior Prediction

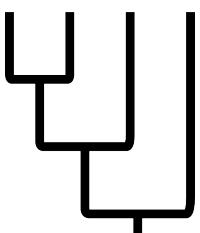




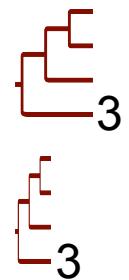
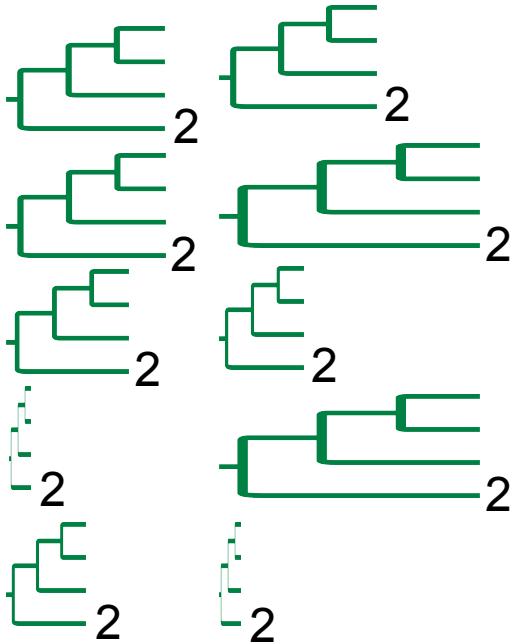
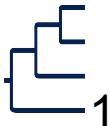
Marginal Test Quantities

$$P(\text{ } \left| \begin{array}{c} \text{ } \\ \text{ } \end{array} \right. \text{ })$$





Marginal Test Quantities - Topology



1/13

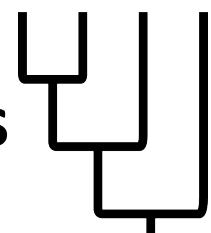
10/13

2/13

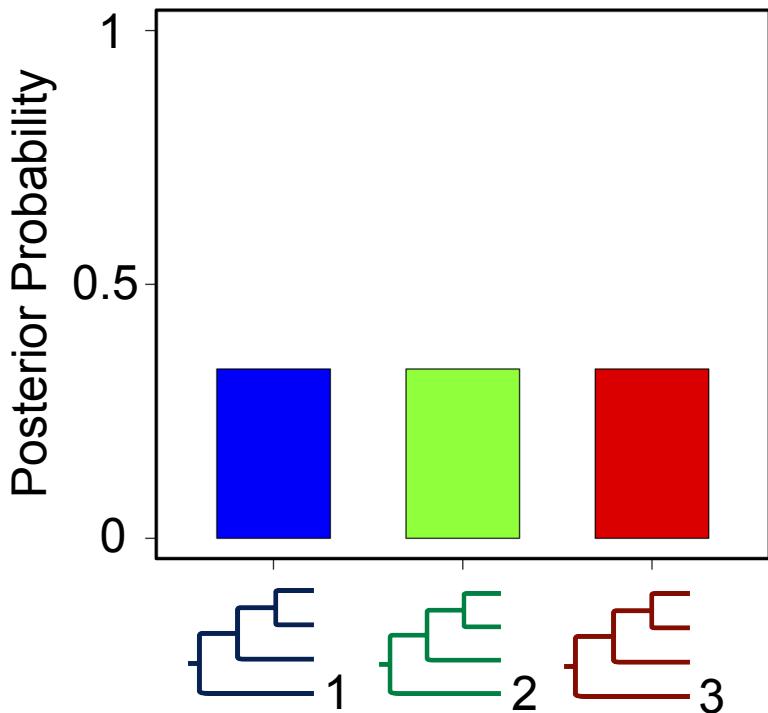
Integrating across variation in branch lengths (nuisance parameter)



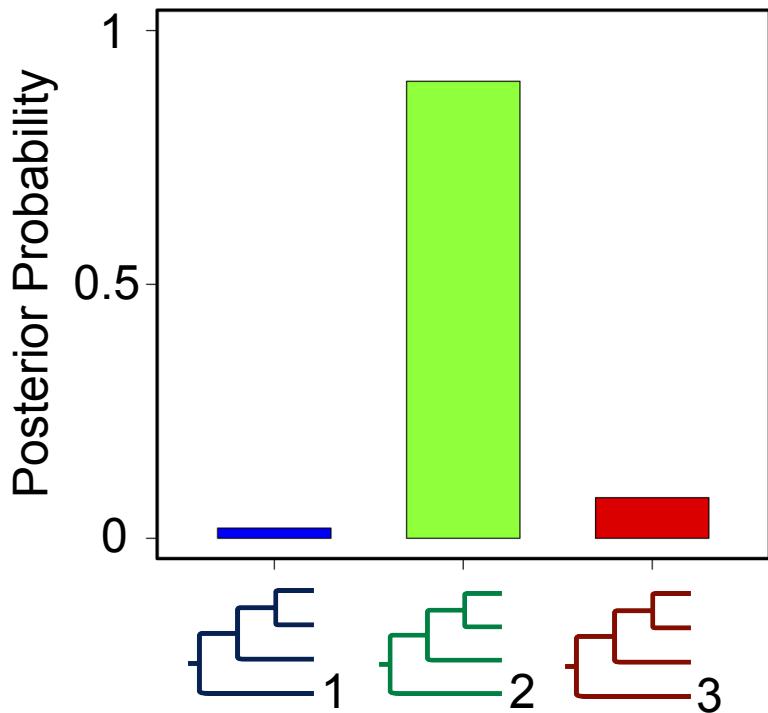
Entropy Quantifies Support Across Topologies



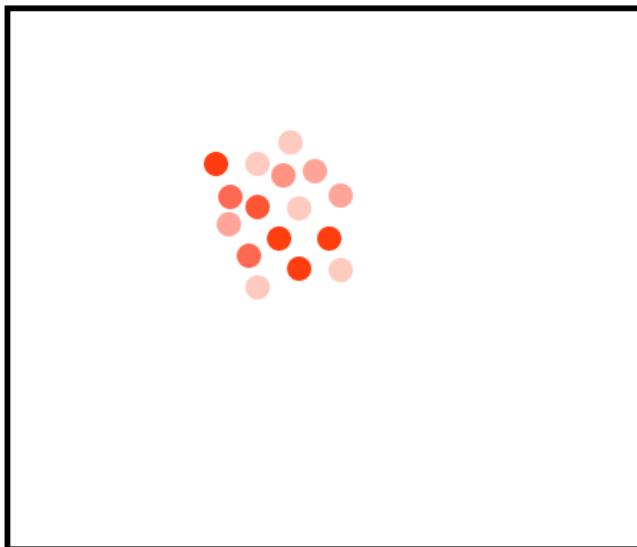
Higher Entropy



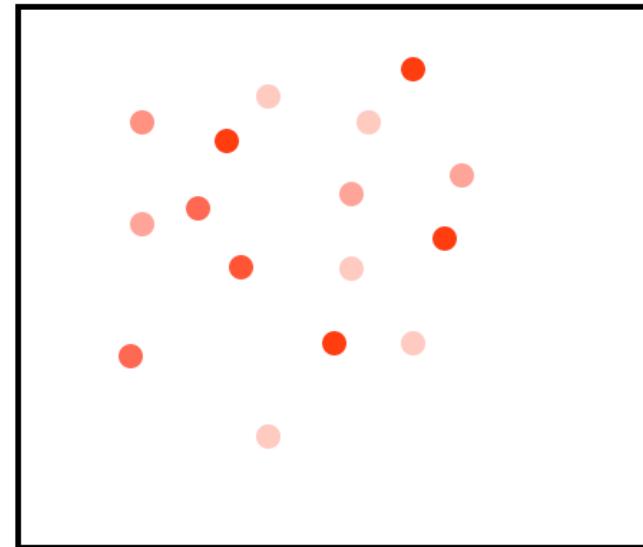
Lower Entropy



Topological Clustering Statistics

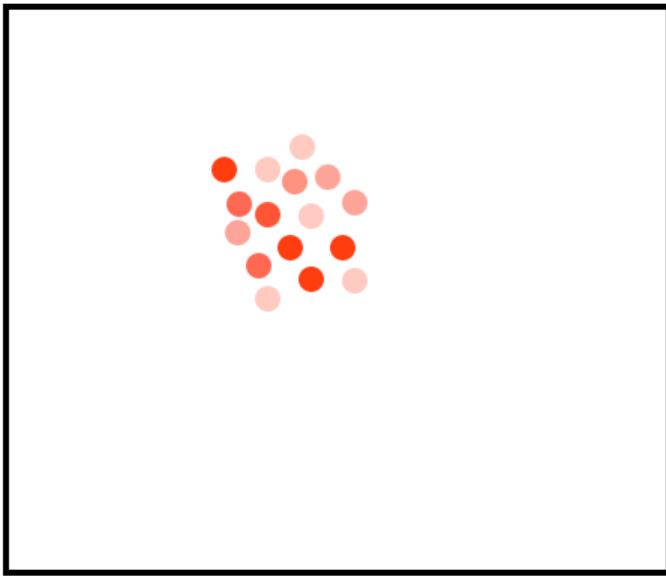


Tree Space

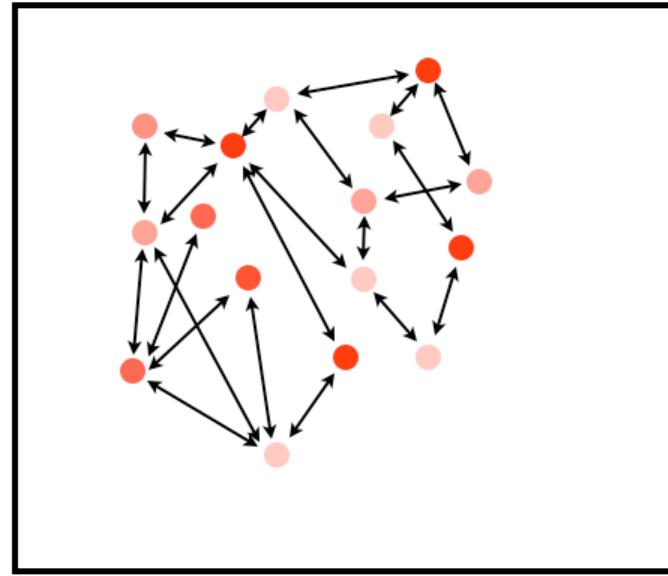


Tree Space

Topological Clustering Statistics

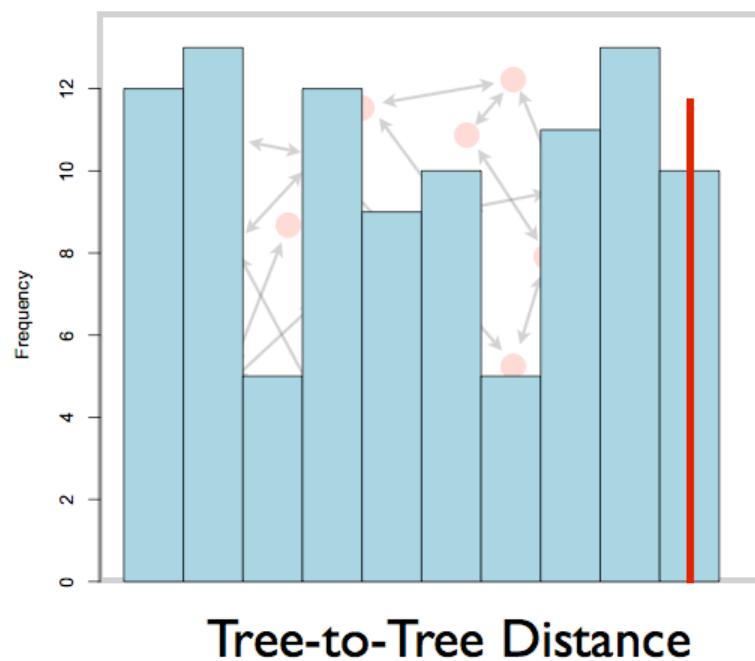
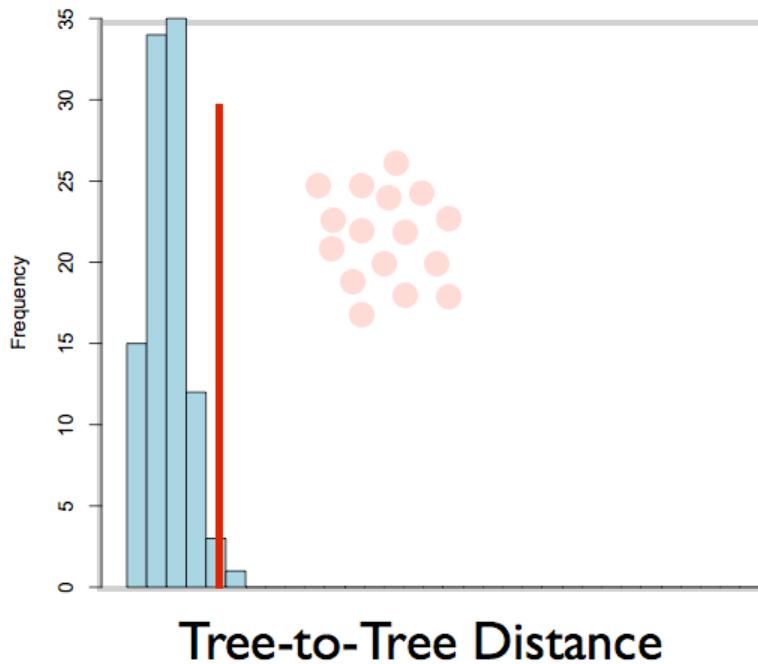


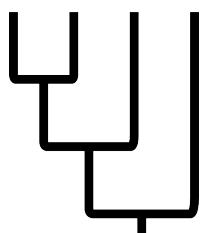
Tree Space



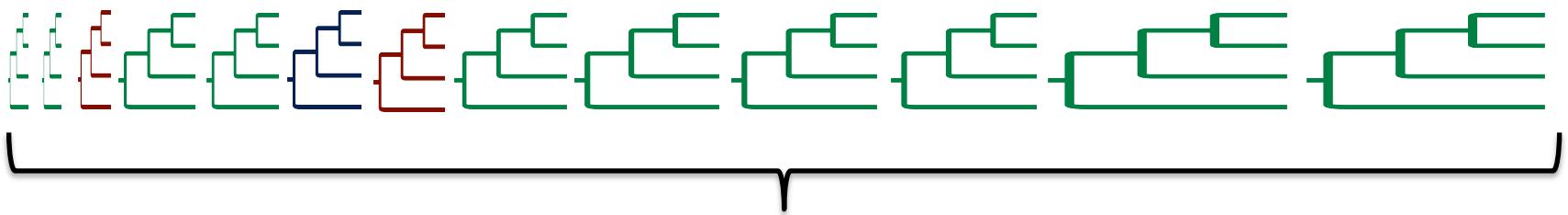
Tree Space

Topological Clustering Statistics





Marginal Test Quantities – Tree Length



Mean Tree Length = 3.15

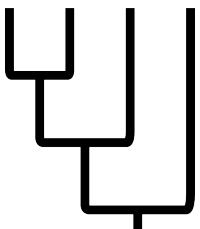
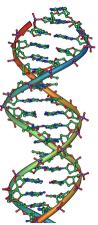
Integrating across topologies (nuisance parameter)

Posterior Prediction



What kinds of inferences might we care about?

- Overall topological inference
- Branch-specific support (posteriors)
- Tree (or branch) length
- Support from individual sequence positions
(Identify specific biased sites)



Try It Out With AMP

1. Perform Bayesian data analysis
(e.g., MrBayes)



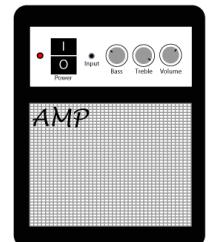
2. Simulate posterior predictive data
(e.g., PuMA or MAPPS)



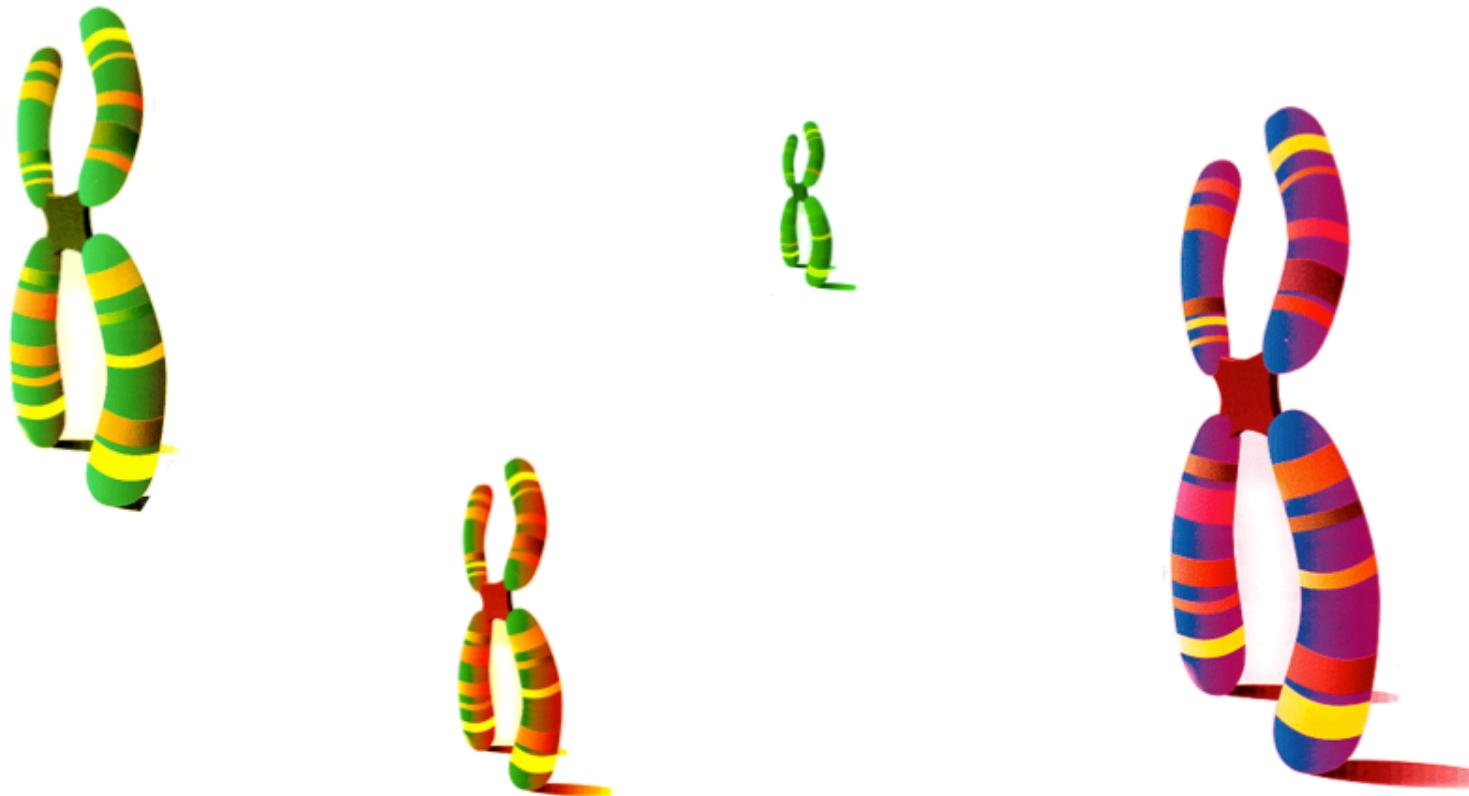
3. Analyze posterior predictive data sets
(e.g., MrBayes)



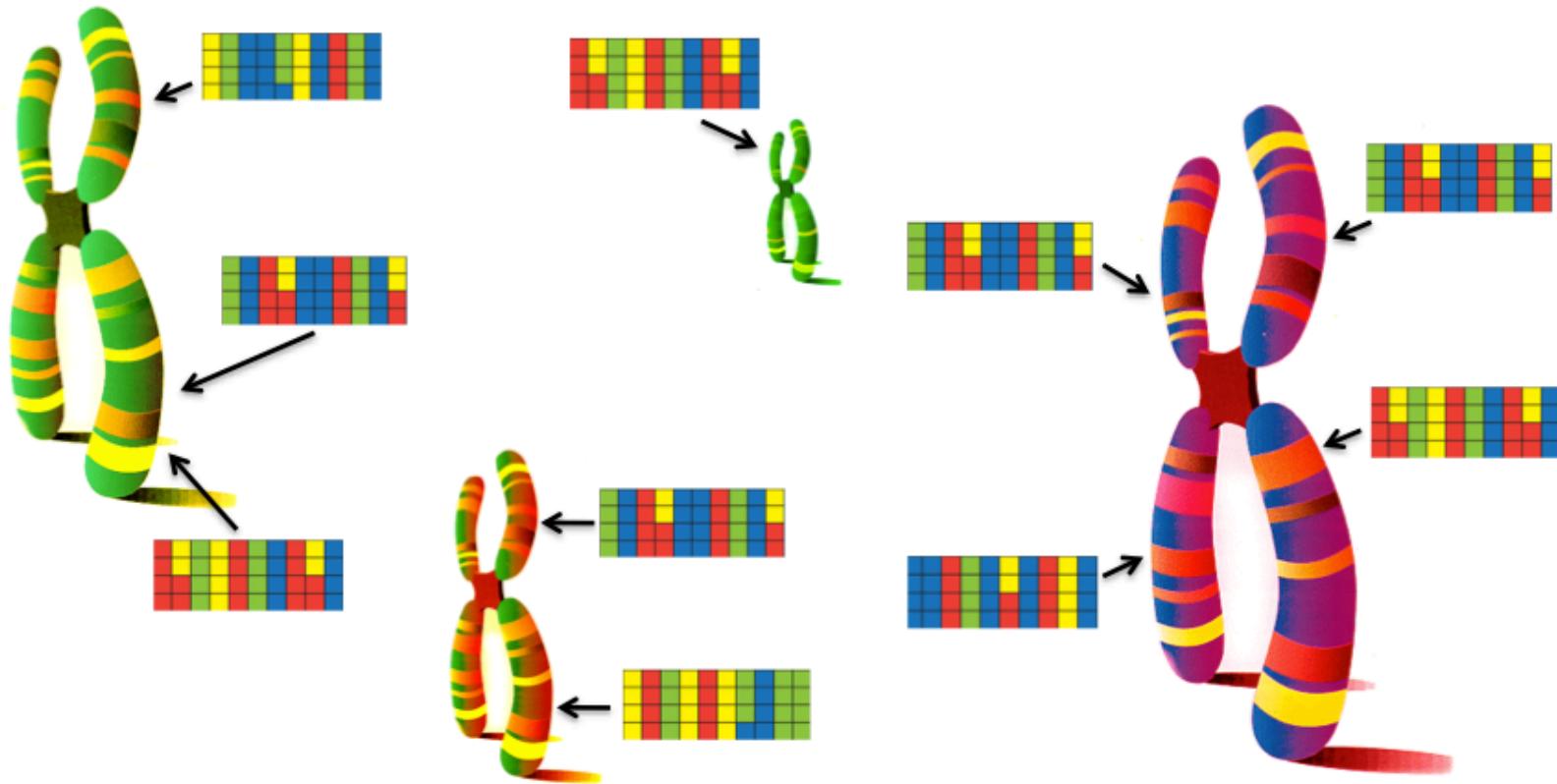
4. Calculate marginal posterior predictive *P*-values
AMP (<http://code.google.com/p/phylo-amp>)



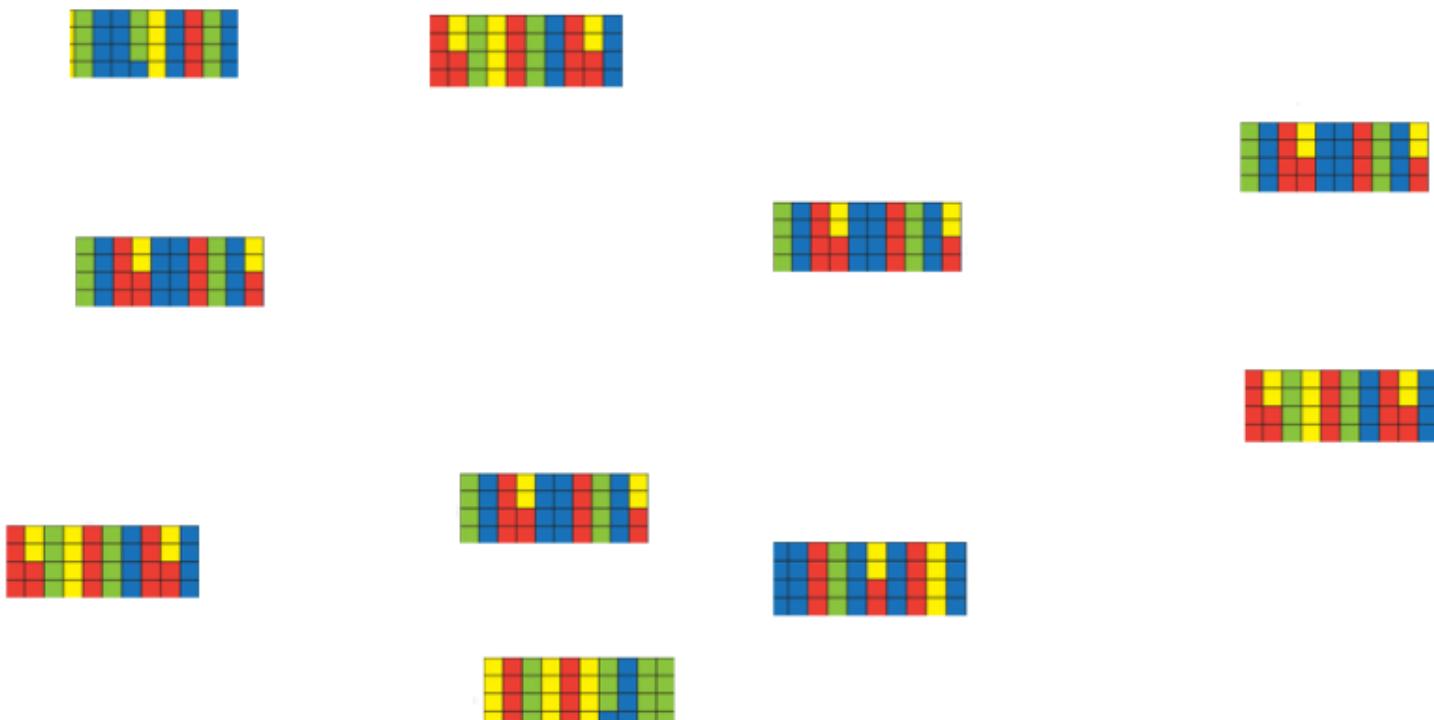
Bottom-Up Phylogenomics



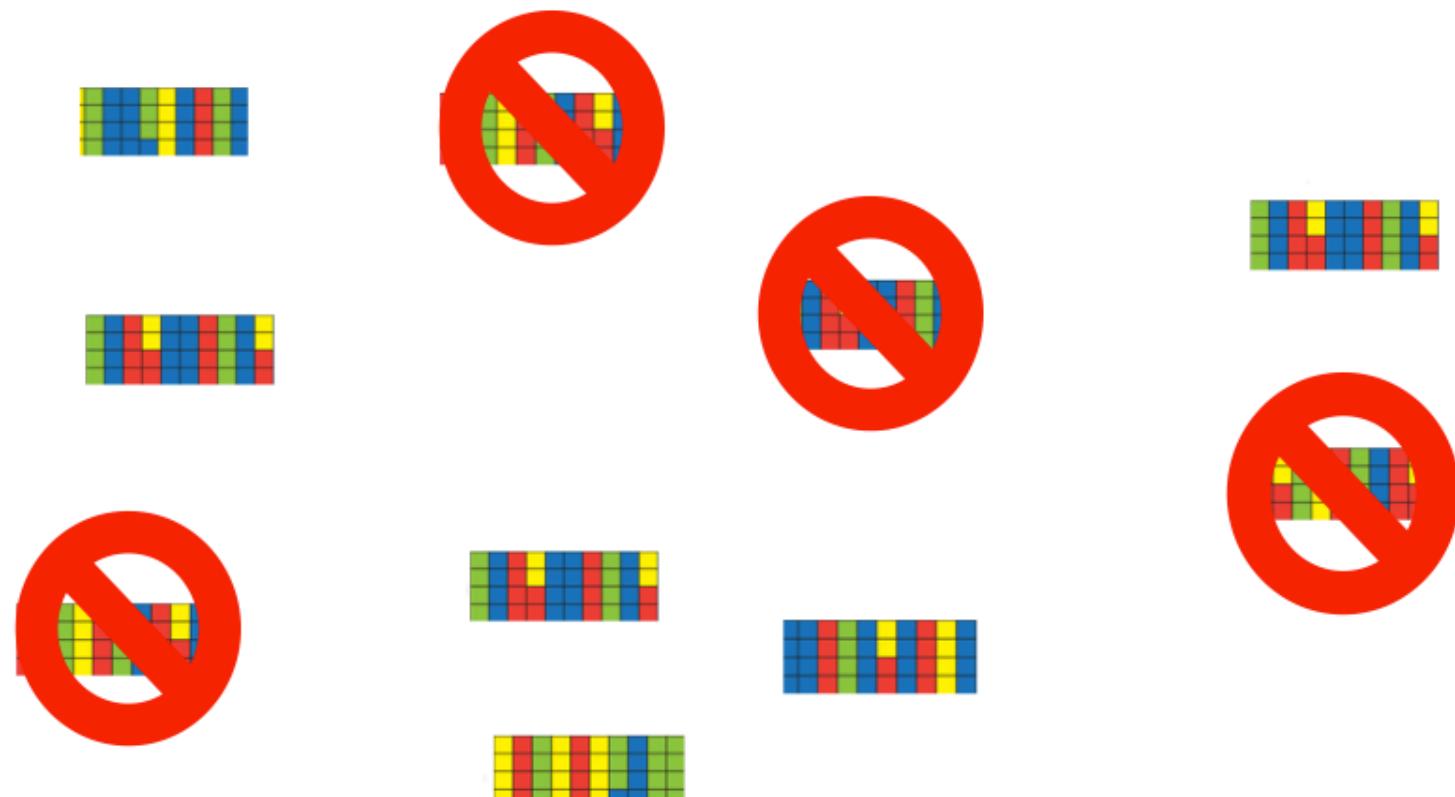
Bottom-Up Phylogenomics



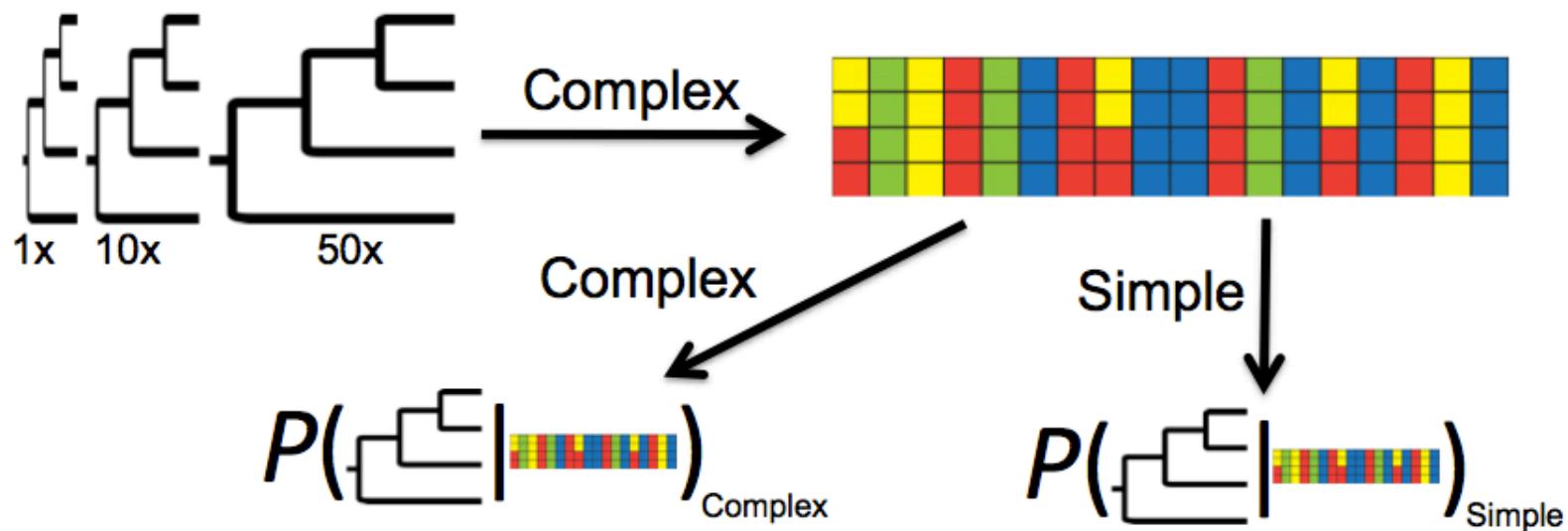
Bottom-Up Phylogenomics

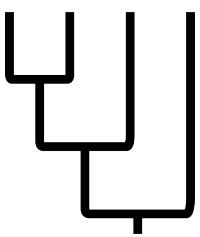


Bottom-Up Phylogenomics

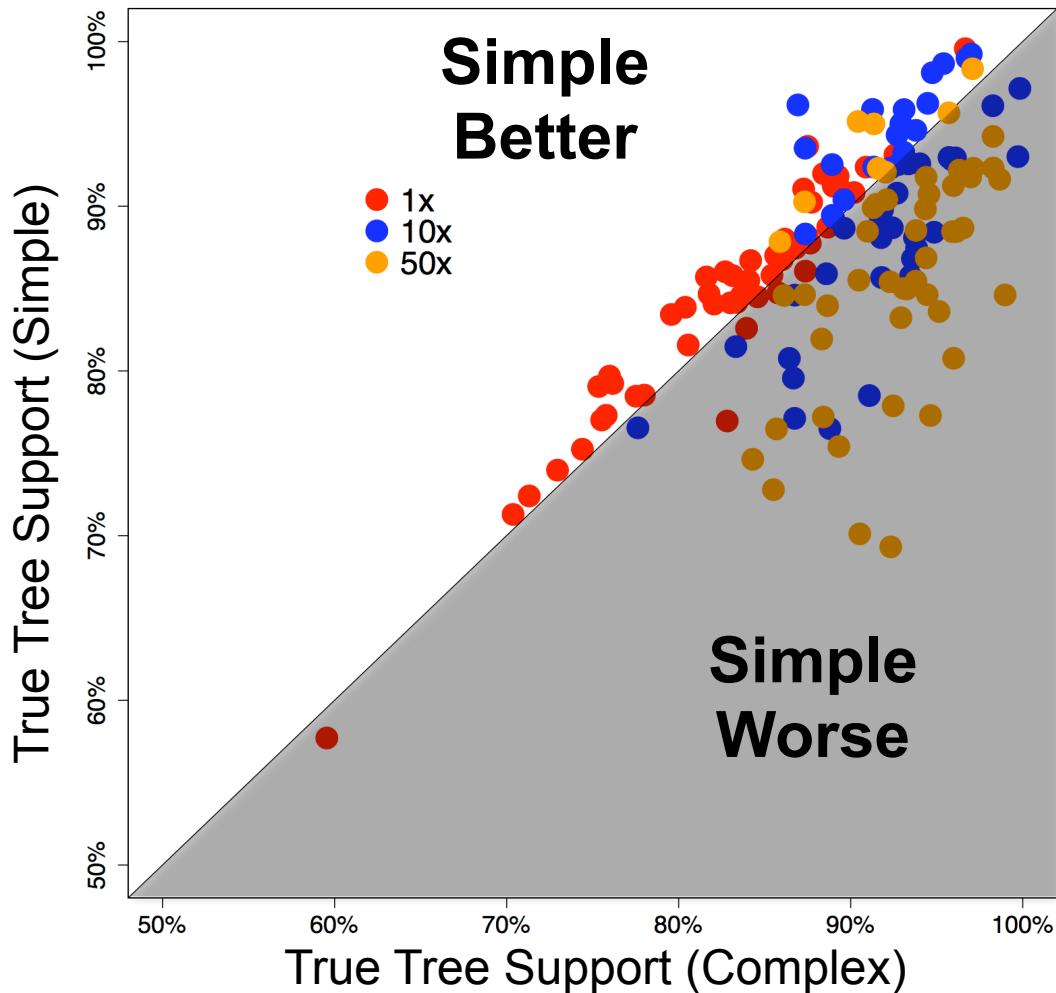


Simulation Test



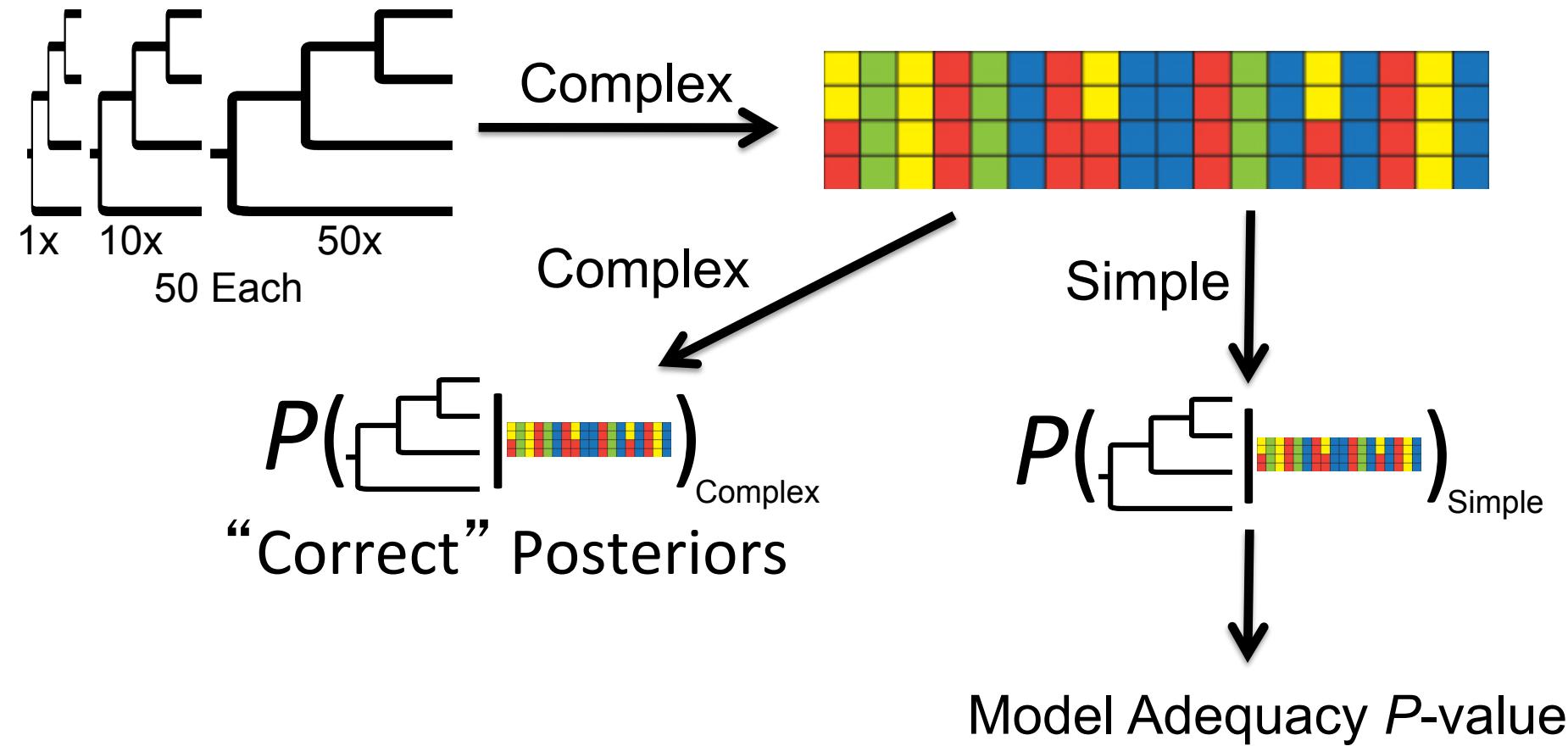
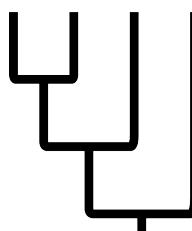


Simple Model Biased with Longer Trees



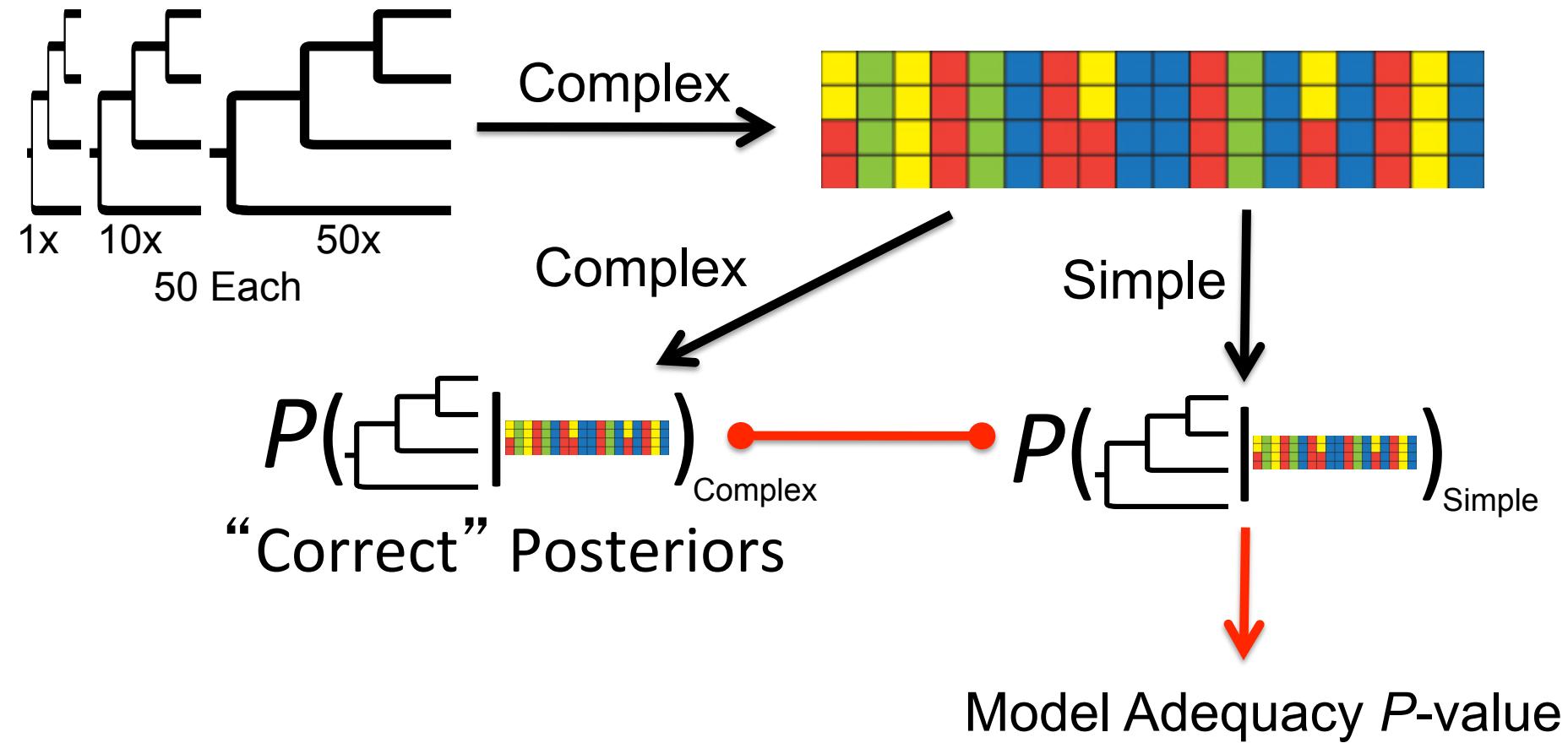
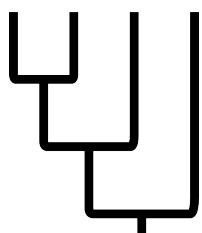


Testing Performance with Simulated Data

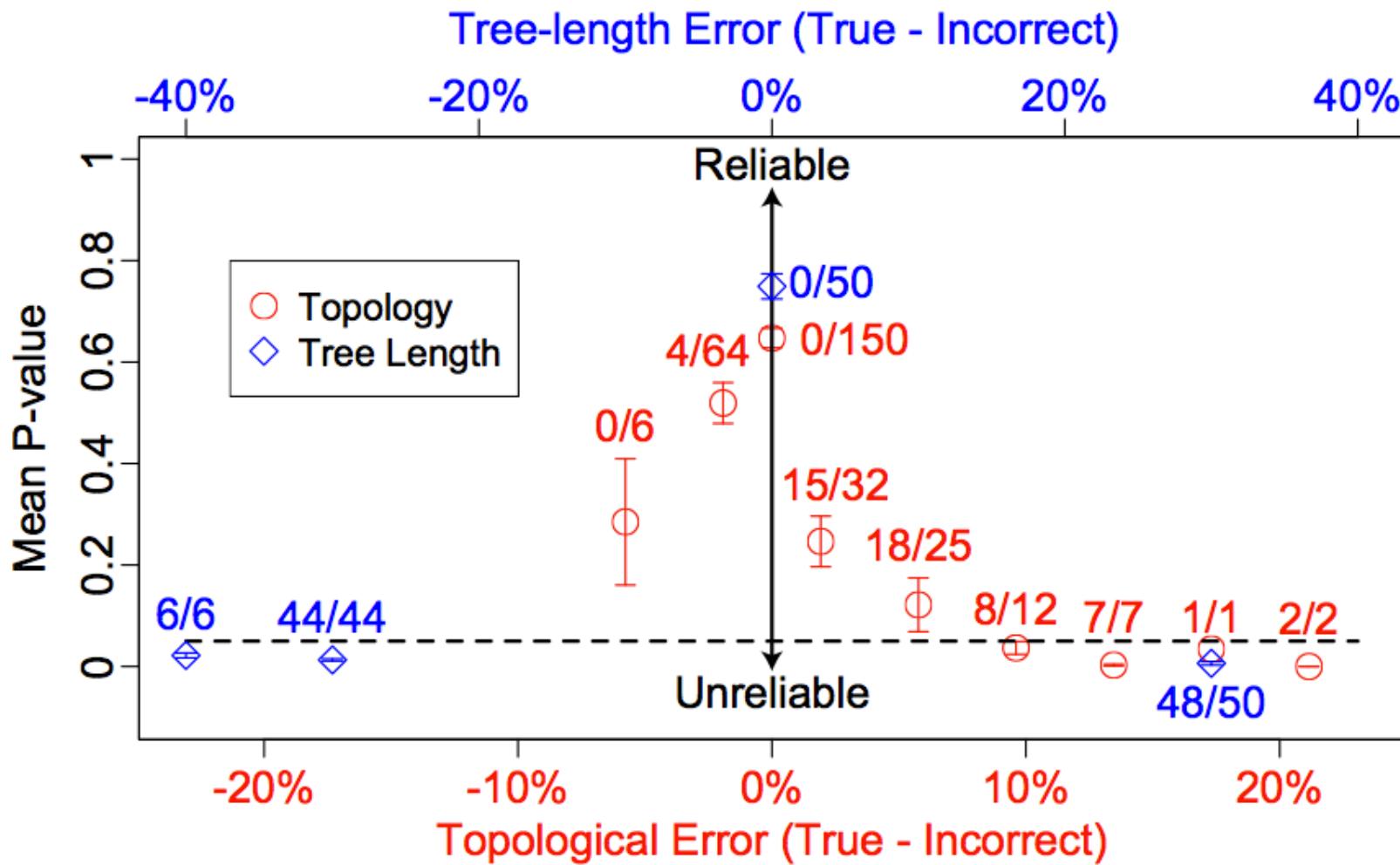




Testing Performance with Simulated Data

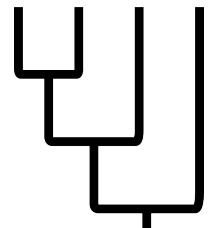


Simulation Test





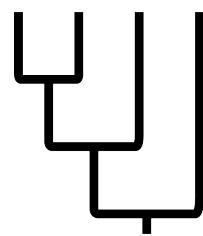
Treelength Test Quantity Detects Biased Branch-Length Inference



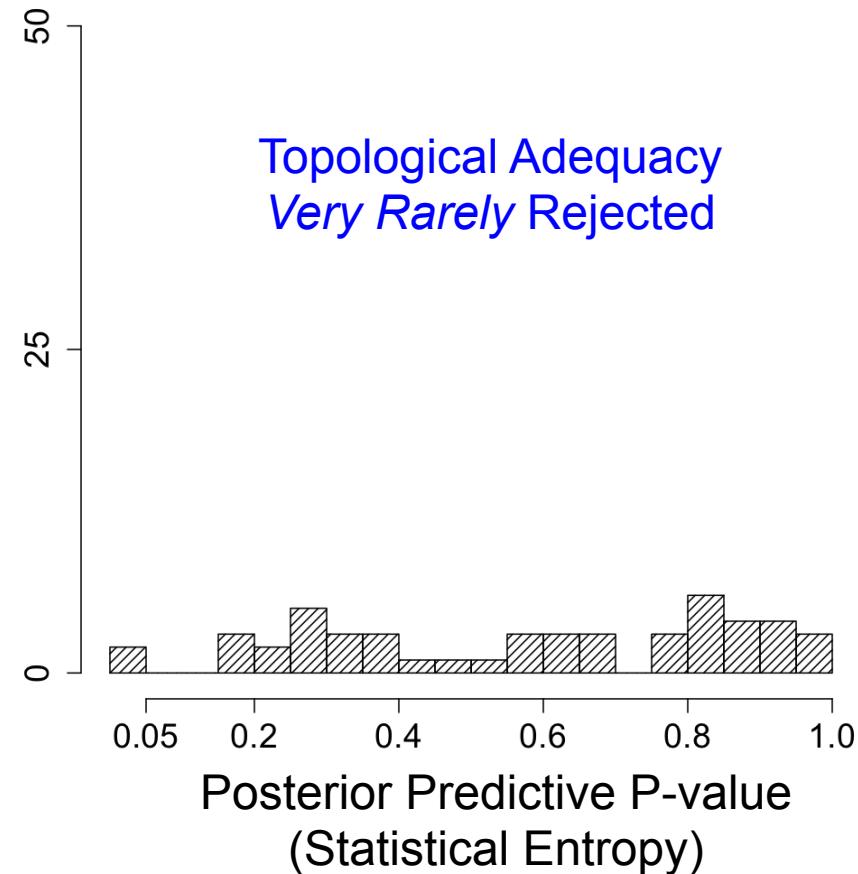
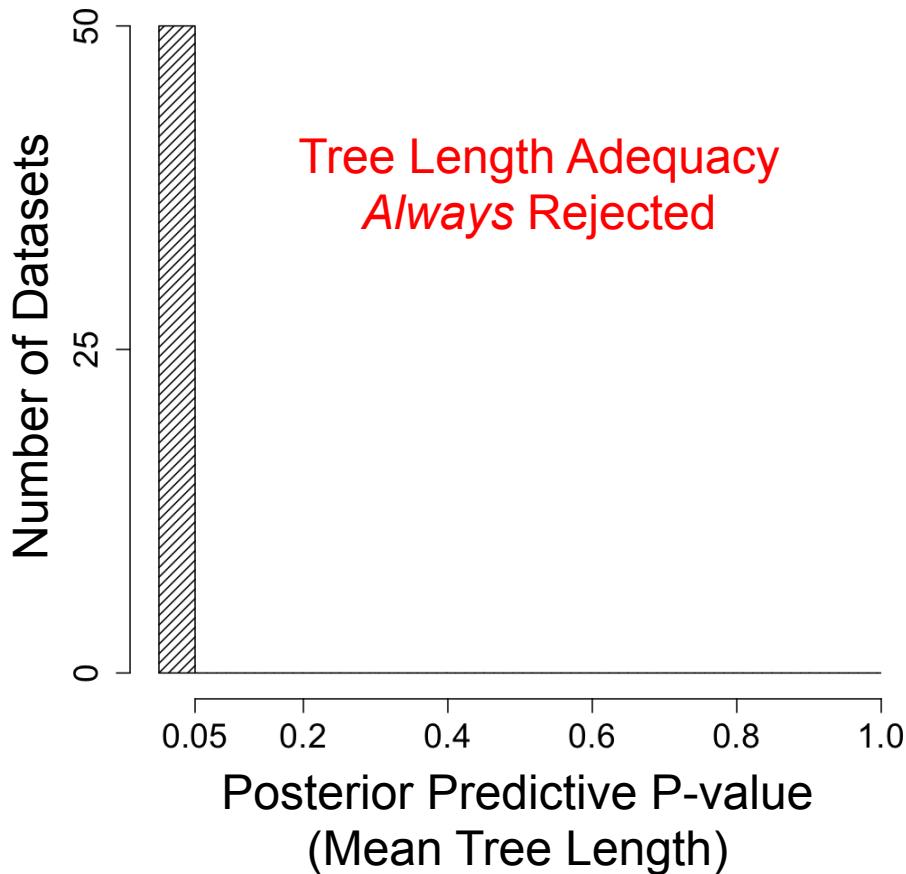
- 50 replicate datasets
- Bipartition posteriors nearly identical to true branch-length prior
- Tree Lengths overestimated



Treelength Test Quantity Detects Biased Branch-Length Inference



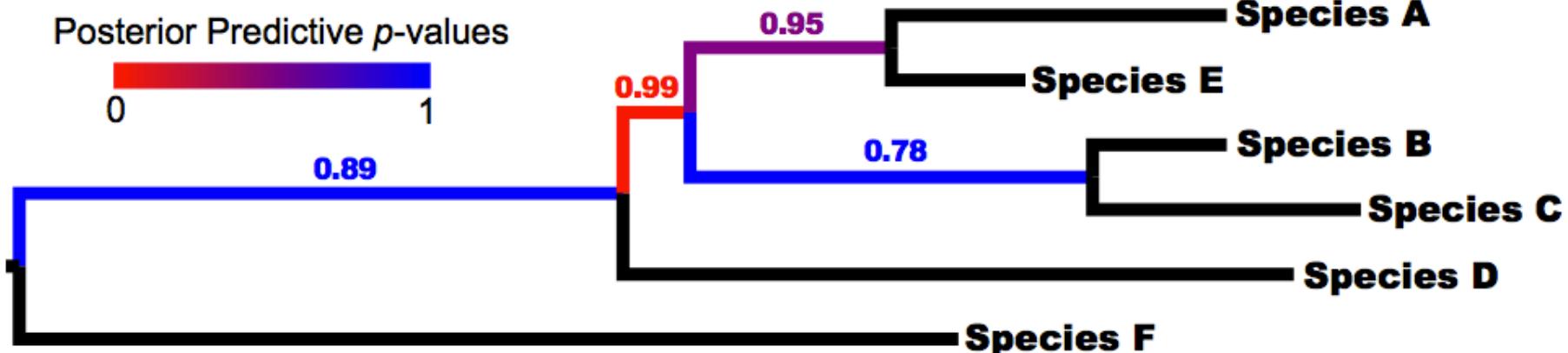
- 50 replicate datasets
- Bipartition posteriors nearly identical to true branch-length prior
- Tree Lengths overestimated



Posterior Prediction



Posterior Predictive p -values



Likelihood Ratio

Posterior Prediction

Topology 1

Empirical

Topology 2

Sequence Position

The End