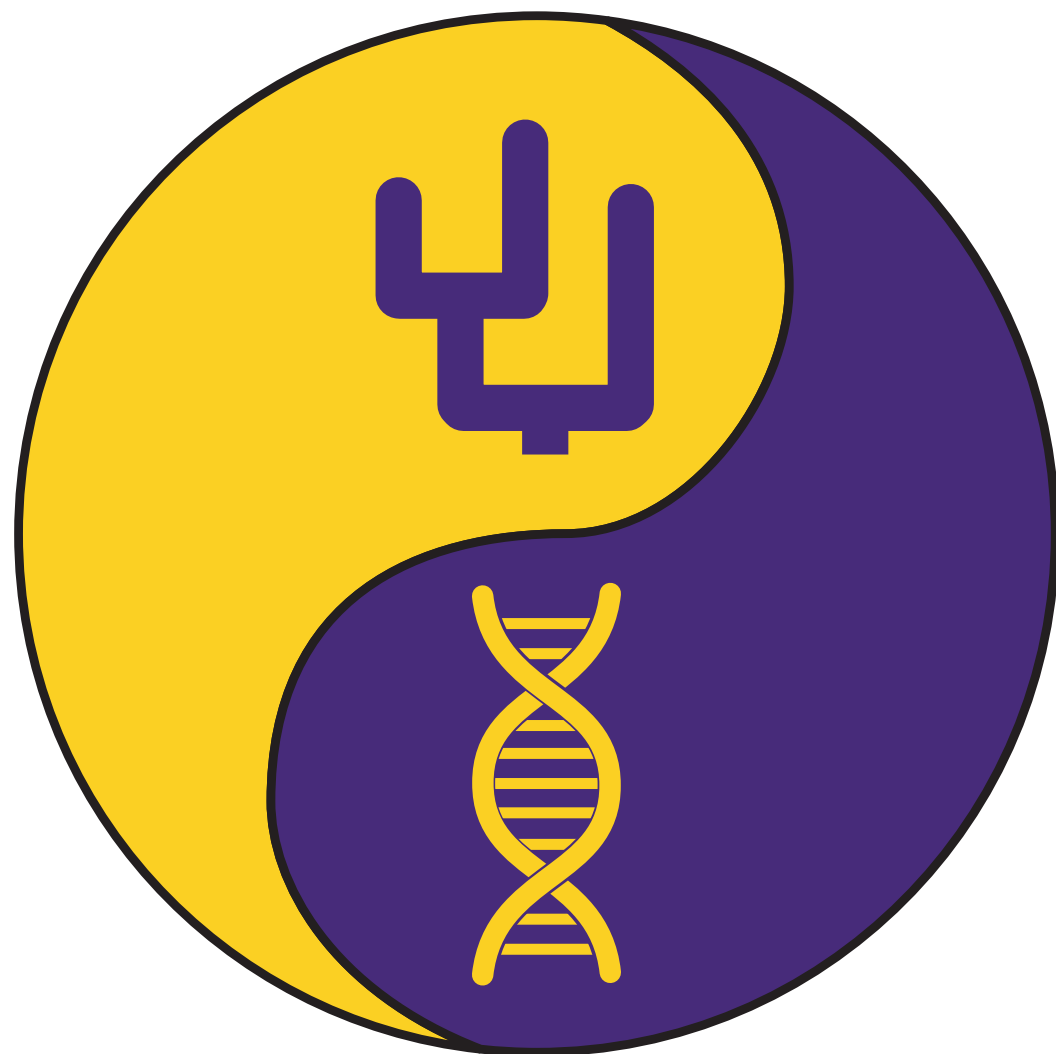


Assessing Model Fit

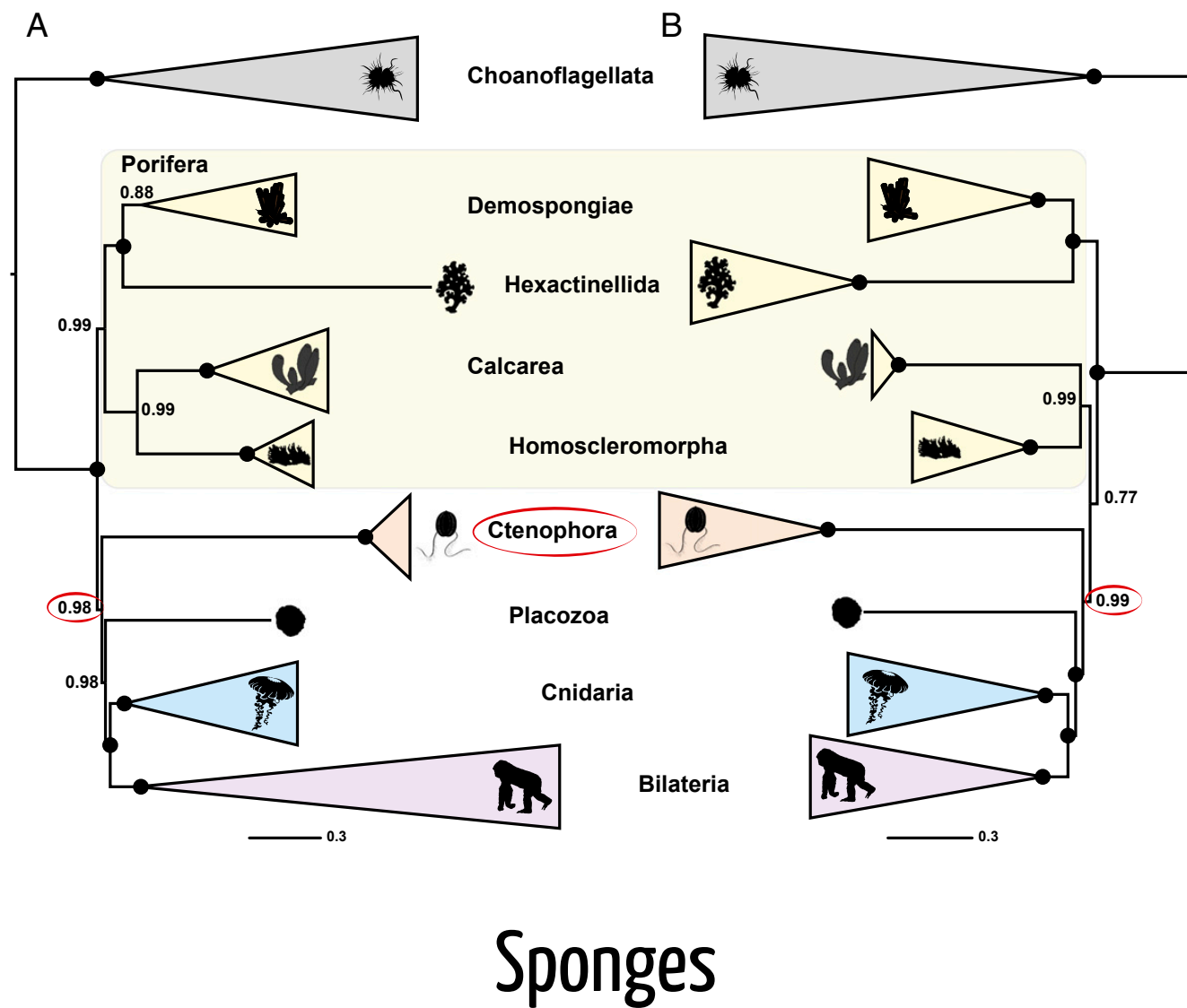
Jeremy M. Brown @jembrown

Robert C. Thomson @rcthomson

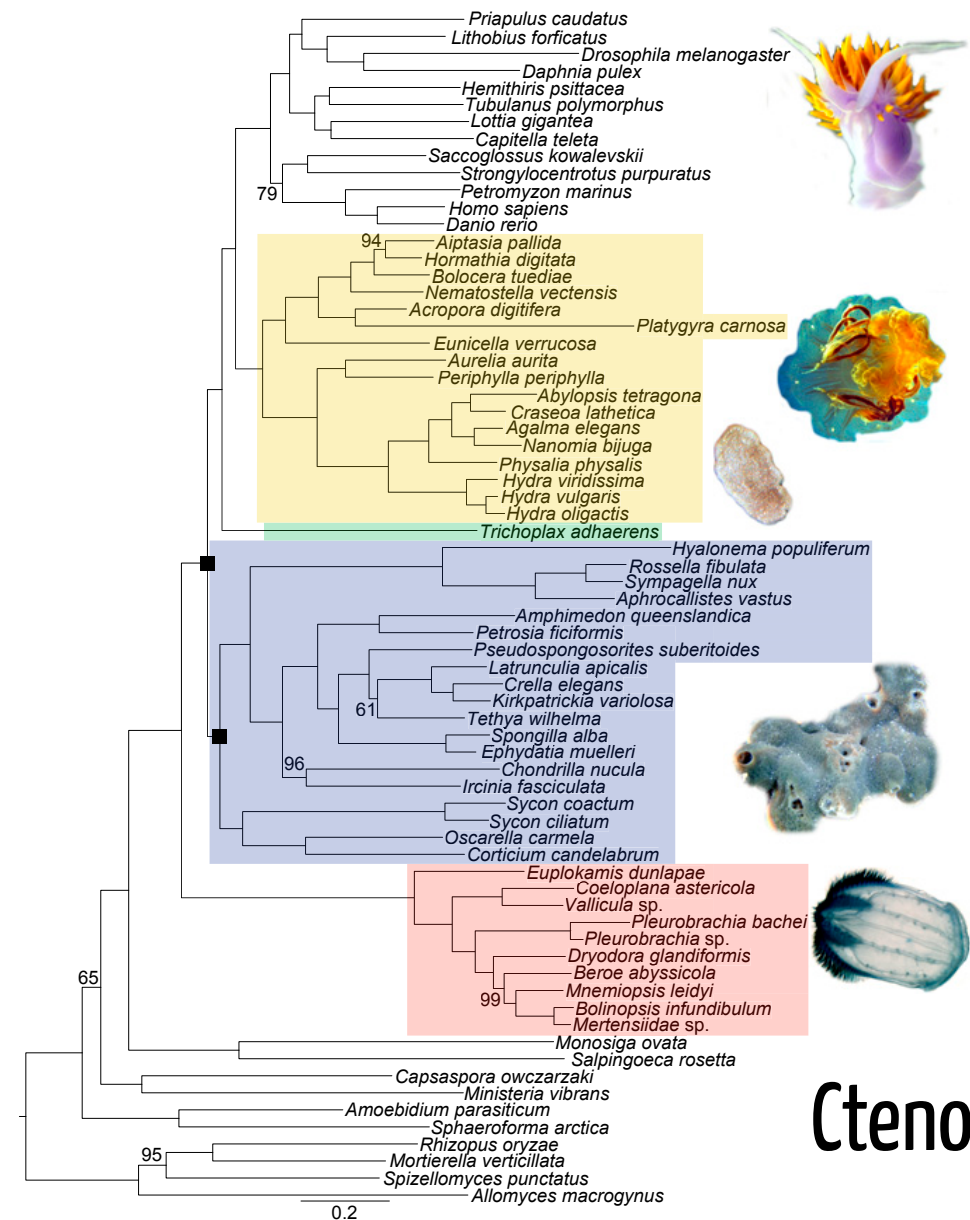


Evidence for Systematic Error (Bias)

Who are the earliest diverging animals?



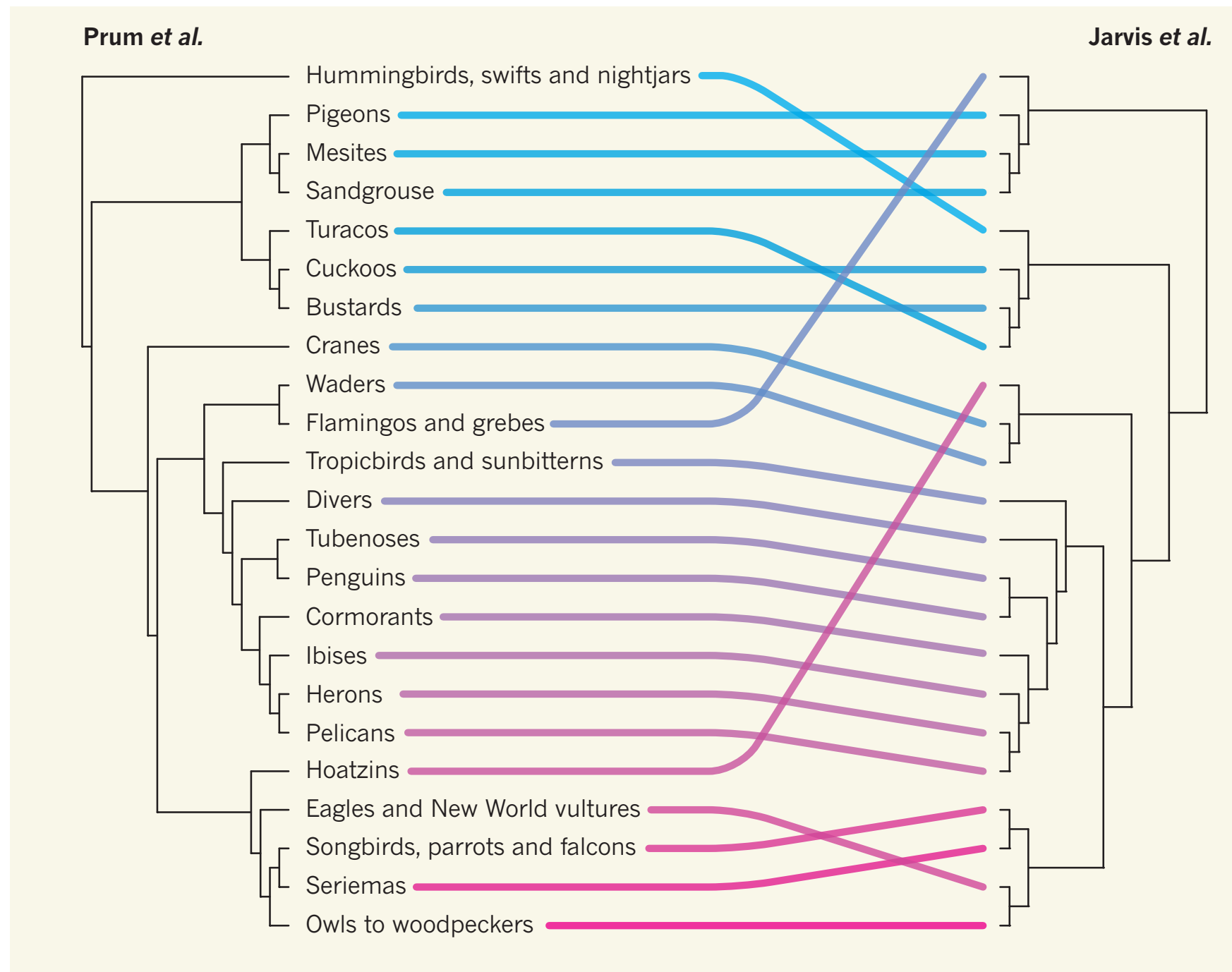
Sponges



Ctenophores

Evidence for Systematic Error (Bias)

Backbone Tree for All Birds



These are enormous datasets, yet they conflict strongly for early divergences.

Thomas. 2015.

Evidence for Systematic Error (Bias)

These are two high-profile examples, but there are many others (we'll talk about turtles later).

When conflict is this strong, stochastic error is not a plausible explanation.

Data is no longer limiting. We are now limited by our ability to accurately extract information from the data.

The Standard Approach

(1) Collect Data (D)



{3.4, 2.1, 5.4, ...}


The Standard Approach



(1) Collect Data (D)



{3.4, 2.1, 5.4, ...}

(2) Define Models

M_1  $\rightarrow D$

M_2   $\rightarrow D$

M_3  $\rightarrow D$


The Standard Approach



(1) Collect Data (D)



{3.4, 2.1, 5.4, ...}


(2) Define Models



M_1  $\rightarrow D$


M_2   $\rightarrow D$

M_3  $\rightarrow D$

(3) Fit Models

M_1  $\rightarrow D : \mathbf{L}_1$

M_2   $\rightarrow D : \mathbf{L}_2$

M_3  $\rightarrow D : \mathbf{L}_3$


The Standard Approach



(1) Collect Data (D)



{3.4, 2.1, 5.4, ...}


(2) Define Models



M_1  $\rightarrow D$


M_2  $\nearrow D$
 $\nearrow D$

M_3  $\rightarrow D$

(3) Fit Models

M_1  $\rightarrow D : L_1$

M_2  $\nearrow D : L_2$
 $\nearrow D$

M_3  $\rightarrow D : L_3$

(4) Compare Models &
Choose “Best”

$M_2 > M_1 > M_3$ \leftarrow $\left. \begin{array}{l} \text{AIC} \\ \text{BIC} \\ \text{BF} \\ \text{LRT} \end{array} \right\} \begin{array}{l} \nearrow L_1 \\ \nearrow L_2 \\ \nearrow L_3 \end{array}$


The Standard Approach



(1) Collect Data (D)



{3.4, 2.1, 5.4, ...}


(2) Define Models



M_1  $\rightarrow D$


M_2  $\nearrow D$
 $\searrow D$

M_3  $\rightarrow D$

(3) Fit Models

M_1  $\rightarrow D : L_1$

M_2  $\nearrow D : L_2$
 $\searrow D : L_2$

M_3  $\rightarrow D : L_3$

(5) Report Inferences from
“Best” Model

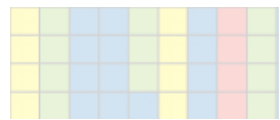
<i>Nature</i>	2012
The Shapes of our Data	
<u>Intro</u>	Blah blah blah. Blah blah blah blah blah. Blah blah blah important blah blah blah. Science blah blah blah blah.
<u>Methods</u>	Blah blah blah. Blah blah likelihood blah 3 models. Blah blah AIC blah blah. Science blah blah blah.
<u>Results</u>	M_2 !
<u>Conclusions</u>	Both squares and circles are important.

(4) Compare Models &
Choose “Best”

$M_2 > M_1 > M_3$ \leftarrow $\left. \begin{array}{l} \text{AIC} \\ \text{BIC} \\ \text{BF} \\ \text{LRT} \end{array} \right\} \begin{array}{l} \nearrow L_1 \\ \nearrow L_2 \\ \nearrow L_3 \end{array}$

The Standard Approach

(1) Collect Data (D)



{3.4, 2.1, 5.4, ...}

nature

The guinea-pig is not a rodent

Anna Maria D'Erchia*†, Carmela Gissi*†,
Graziano Pesole‡, Cecilia Saccone*§ & Ulfur Arnason†

* Dipartimento di Biochimica e Biologia Molecolare, Università di Bari,
70125 Bari, Italy

† Department of Evolutionary Molecular Systematics, University of Lund,
Sölvegatan 29, S-22362 Lund, Sweden



‡ Dipartimento di Biologia DBAF, Università della Basilicata, 30100
Potenza, Italy

§ Centro di Studio sui Mitocondri e Metabolismo Energetico,
Consiglio Nazionale delle Ricerche, 70125 Bari, Italy

IN 1991 Graur *et al.* raised the question of whether the guinea-pig, *Cavia porcellus*, is a rodent¹. They suggested that the guinea-pig and myomorph rodents diverged before the separation between myomorph rodents and a lineage leading to primates and artiodactyls. Several findings have since been reported, both for and against this phylogeny, thereby highlighting the issue of the validity of molecular analysis in mammalian phylogeny. Here we present findings based on the sequence of the complete mitochondrial genome of the guinea-pig, which strongly contradict rodent monophyly. The conclusions are based on the cumulative evidence provided by orthologically inherited genes and the use of three different analytical methods, none of which joins the guinea-pig with myomorph rodents. In addition to the phylogenetic conclusions, we also draw attention to several factors that are important for the validity of phylogenetic analysis based on molecular data.

(3) Fit Models

VI  → D :

VI  → D :
 → D :

VI  → D :

(5) Report Inference
“Best” Model

Nature

The Shapes of our D:

Intro

Blah blah blah. Blah blah blah blah blah. E
blah important blah blah blah. Science blah blah t

Methods

Blah blah blah. Blah blah blah likelihood bla
Blah blah blah AIC blah blah blah. Science blah bl

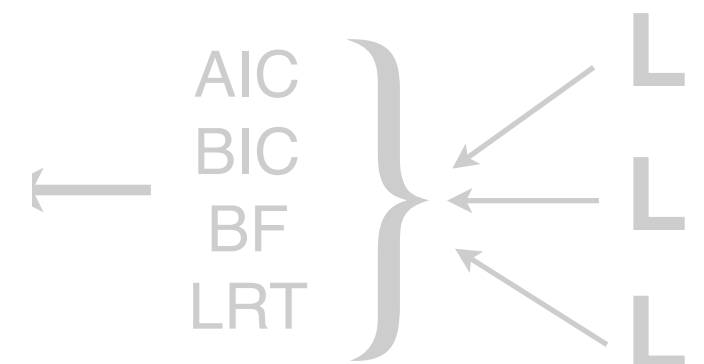
Results

M

Conclusions

Both squares and
circles are importar

Models &
Best”



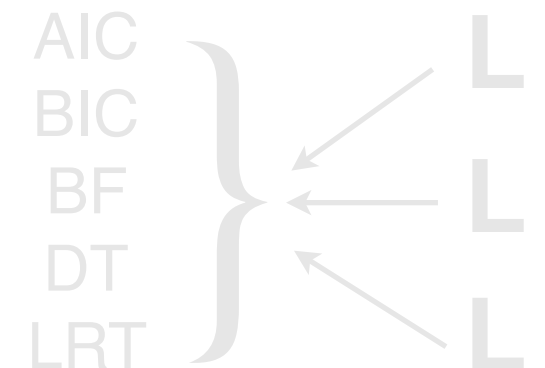
The Next Step - Assessing Fit

We know that none of our models is really true. Can we be sure that the chosen model captures the salient features of the evolutionary process and provides reliable inferences?

(5) Report Inferences from “Best” Model

<i>Nature</i>	2012
The Shapes of our Data	
<u>Intro</u>	Blah blah blah. Blah blah blah blah blah. Blah blah blah important blah blah. Science blah blah blah.
<u>Methods</u>	Blah blah blah. Blah blah likelihood blah 3 models. Blah blah AIC blah blah. Science blah blah.
<u>Results</u>	M₂ !
<u>Conclusions</u>	Both squares and circles are important.

(4) Compare Models & Choose “Best”

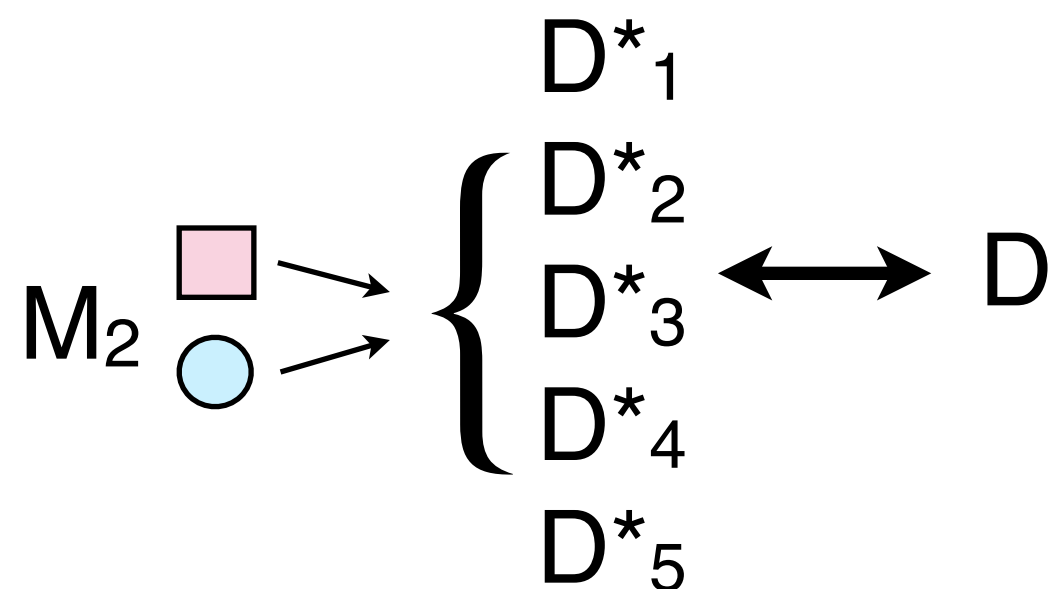


The Next Step - Assessing Fit

We know that none of our models is really true. Can we be sure that the chosen model captures the salient features of the evolutionary process and provides reliable inferences?

(5) Report Inferences from “Best” Model (6) Check Fit of Model to Data

<i>Nature</i>	2012
The Shapes of our Data	
<u>Intro</u>	Blah blah blah. Blah blah blah blah blah. Blah blah blah important blah blah. Science blah blah blah.
<u>Methods</u>	Blah blah blah. Blah blah likelihood blah 3 models. Blah blah AIC blah blah. Science blah blah.
<u>Results</u>	M_2 !
<u>Conclusions</u>	Both squares and circles are important.



The Next Step - Assessing Fit

Are Guinea Pigs Rodents? The Importance of Adequate Models in Molecular Phylogenetics

Jack Sullivan^{1,2} and David L. Swofford¹

The monophyly of Rodentia has repeatedly been challenged based on several studies of molecular sequence data. Most recently, D'Erchia *et al.* (1996) analyzed complete mtDNA sequences of 16 mammals and concluded that rodents are not monophyletic. We have reanalyzed these data using maximum-likelihood methods. We use two methods to test for significance of differences among alternative topologies and show that (1) models that incorporate variation in evolutionary rates across sites fit the data dramatically better than models used in the original analyses, (2) the mtDNA data fail to refute rodent monophyly, and (3) the original interpretation of strong support for nonmonophyly results from systematic error associated with an oversimplified model of sequence evolution. These analyses illustrate the importance of incorporating recent theoretical advances into molecular phylogenetic analyses, especially when results of these analyses conflict with classical hypotheses of relationships.

KEY WORDS: inconsistency; maximum likelihood; molecular systematics; rodents; rate heterogeneity.

Natur

The

Intro

Bla
blat

Methods

Results

IVI

Conclusions

Both squares and
circles are important.

How might we assess fit?

- (1) Use our **prior knowledge** to ask if the **data** are reasonable.
- (2) Use our **prior knowledge** to ask if **inferences** are reasonable.

Above are “gut checks”. Very useful, but perhaps subjective. Also difficult to have strong priors for complicated data and models.

- (3) Use your **data** (all or part) to make a **prediction** and see if your prediction matches what you’ve seen.
(Posterior Prediction and Cross Validation)

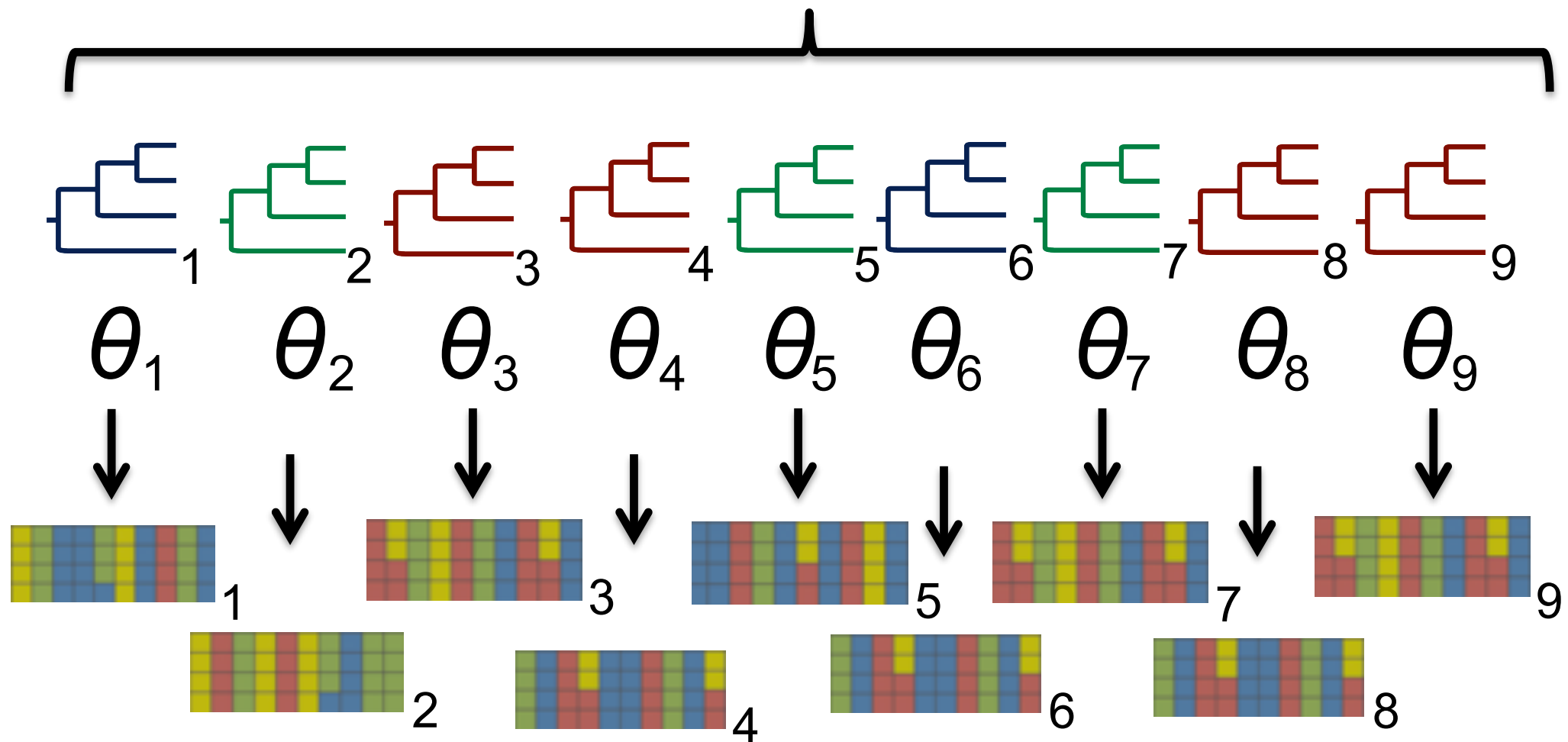
Posterior Prediction

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

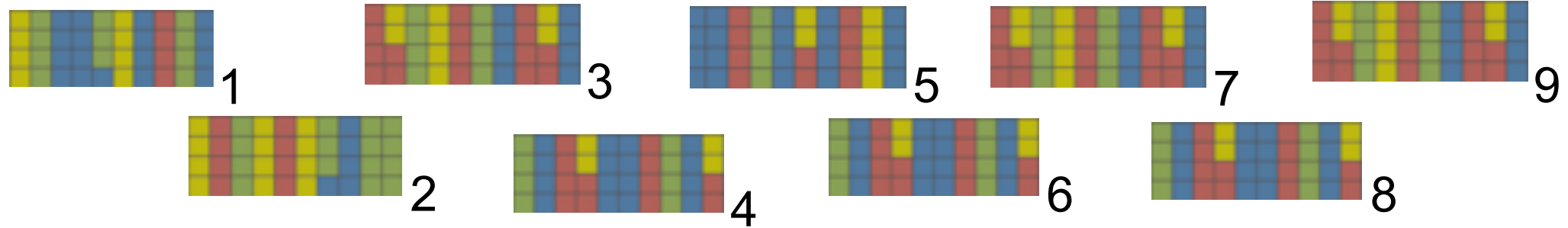
Could the model and priors plausibly have
given rise to the data?

Posterior Prediction

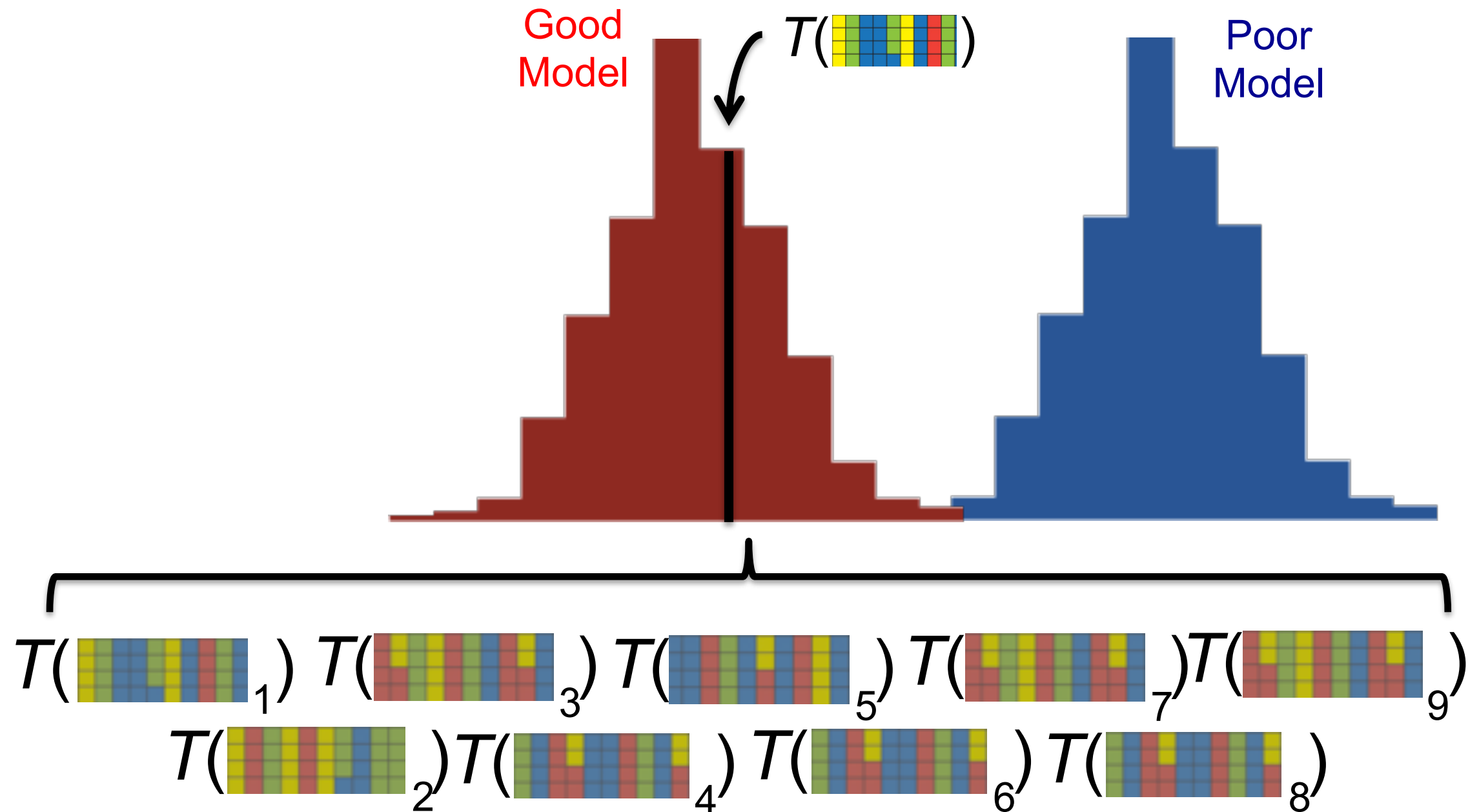
$$P(\text{tree}, \theta | \text{matrix})$$



Posterior Prediction



Posterior Prediction



Posterior Prediction

Previously proposed statistics based on the data:

- Multinomial Likelihood (based on frequencies of site patterns)
- Number of Unique Site Patterns
- Frequency of Invariant Sites
- Heterogeneity of Base Frequencies
- Number of parsimony-inferred “parallel” sites

Posterior Prediction

“We do not like to ask, ‘Is our model true or false?’, since most 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?’ “

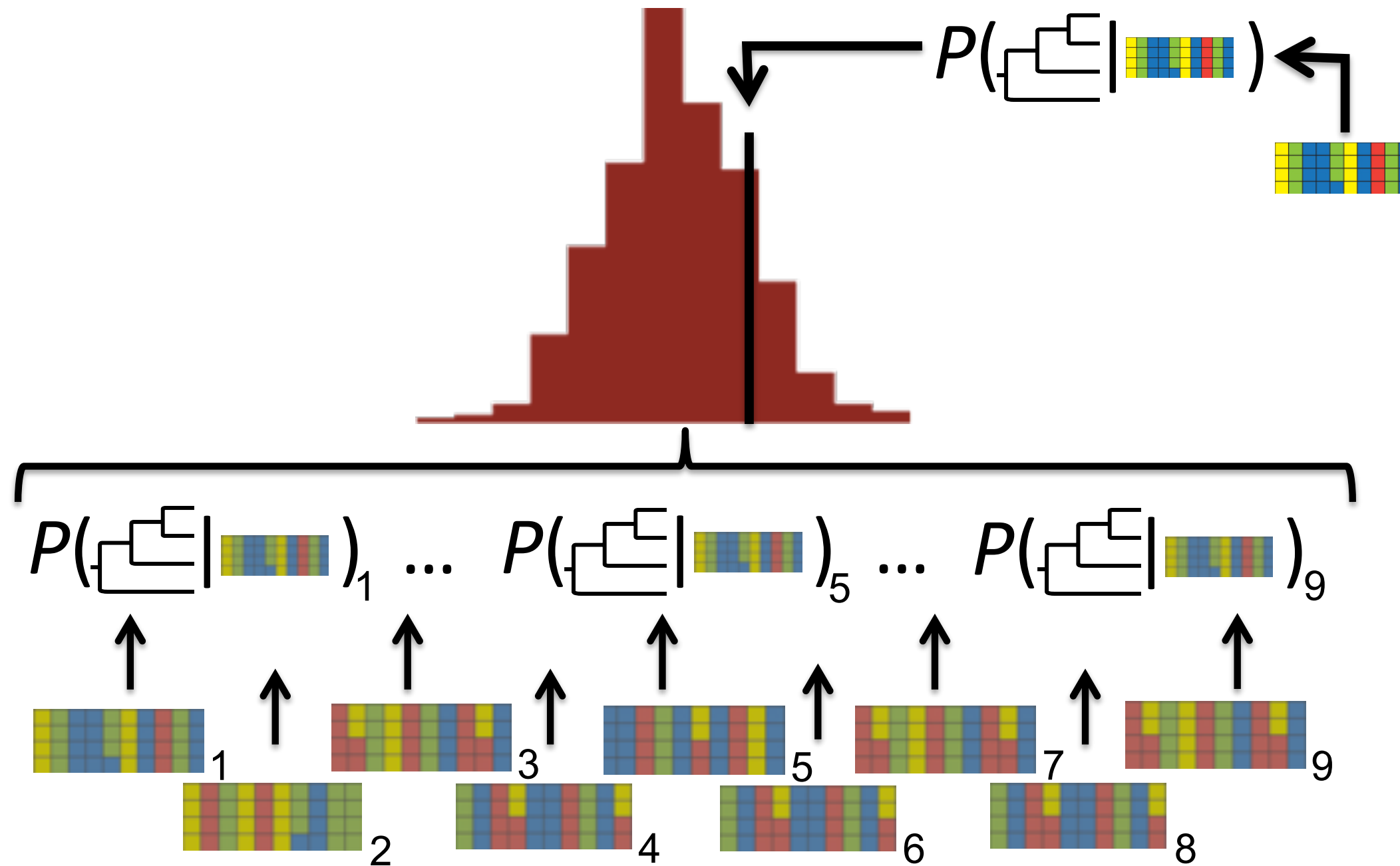
- Gelman, Carlin, Stern, and Rubin

Bayesian Data Analysis

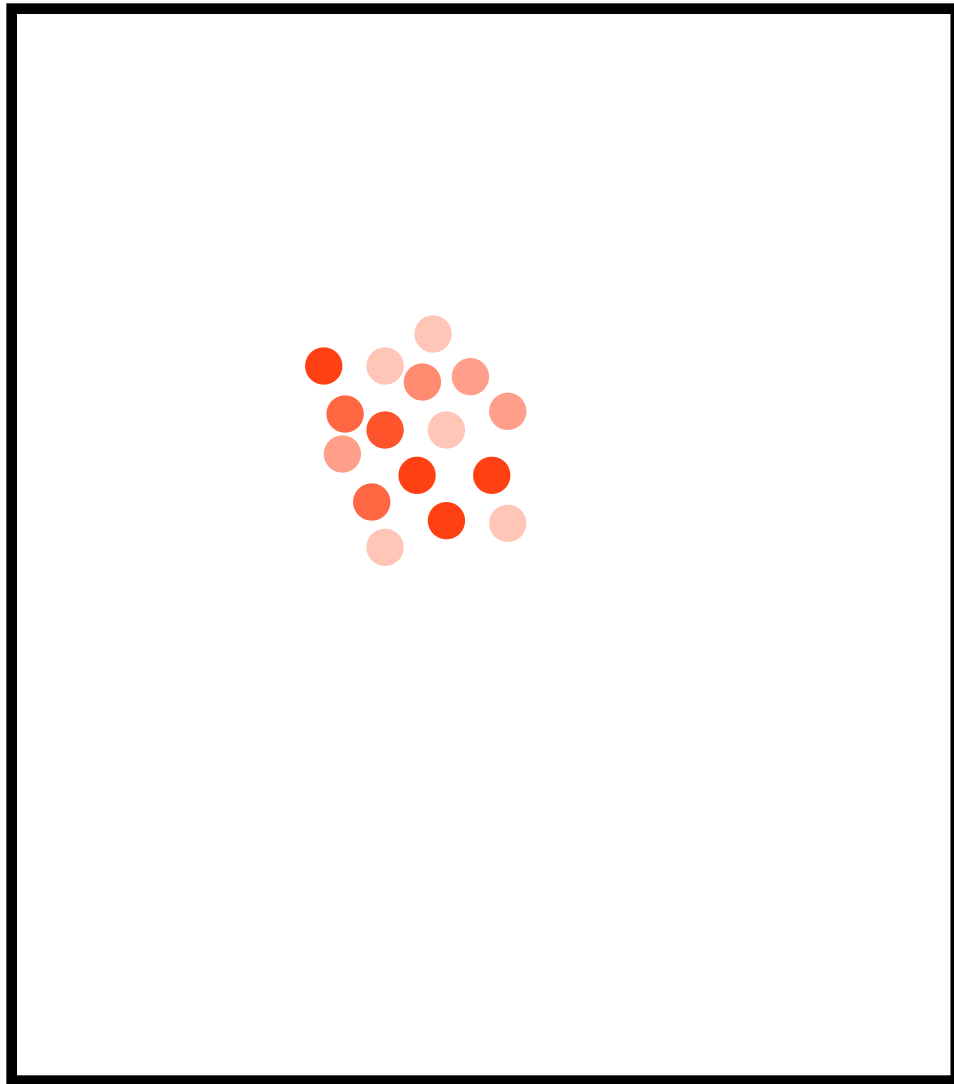
Posterior Prediction

What about using the inferences provided
by our data as a test statistic(s)?

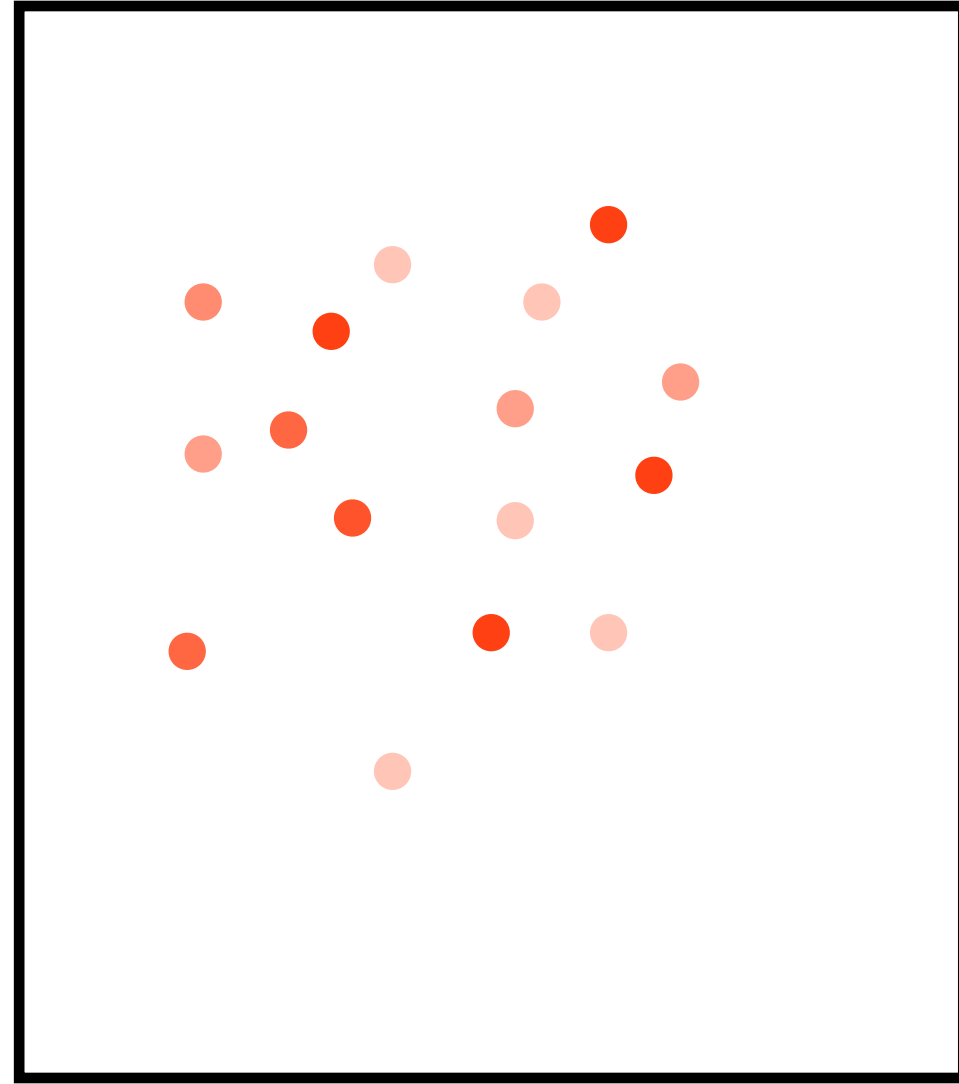
Posterior Prediction



Topology Test Statistics

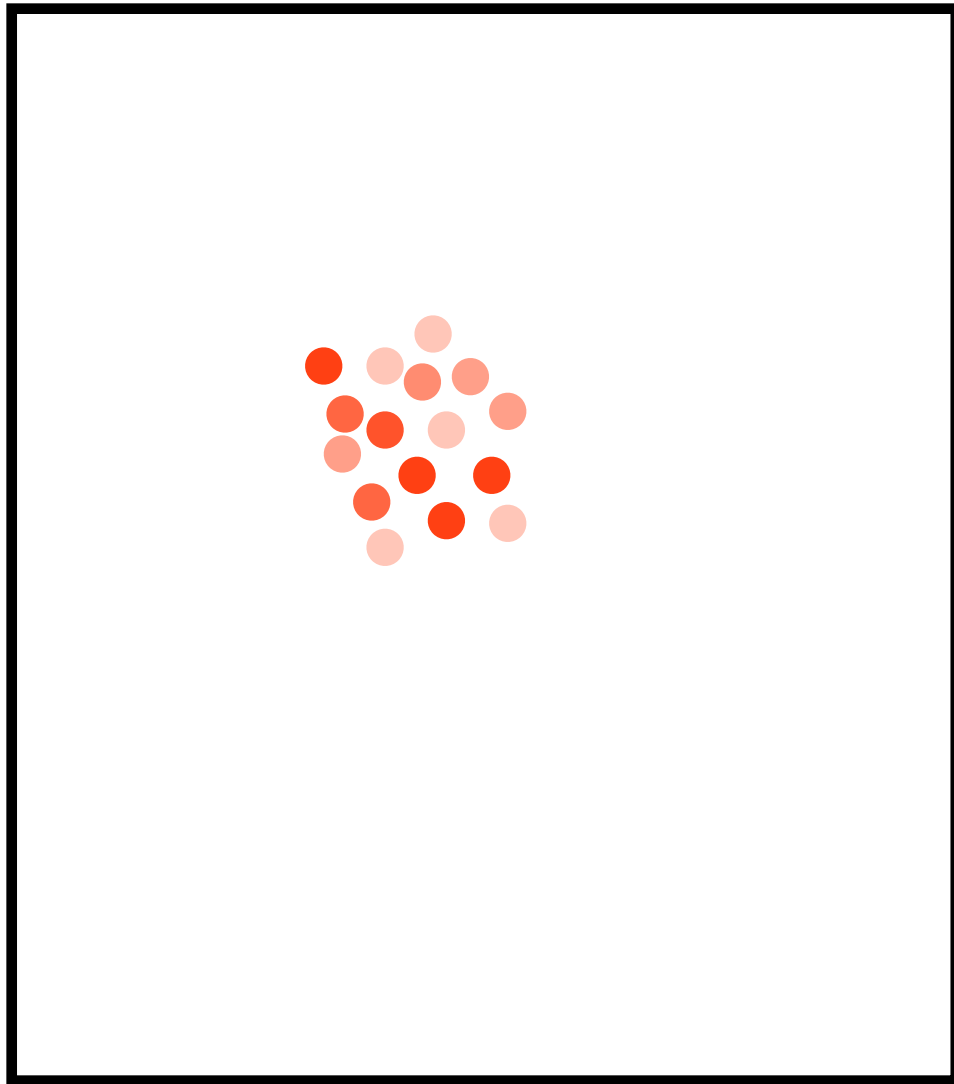


Tree Space

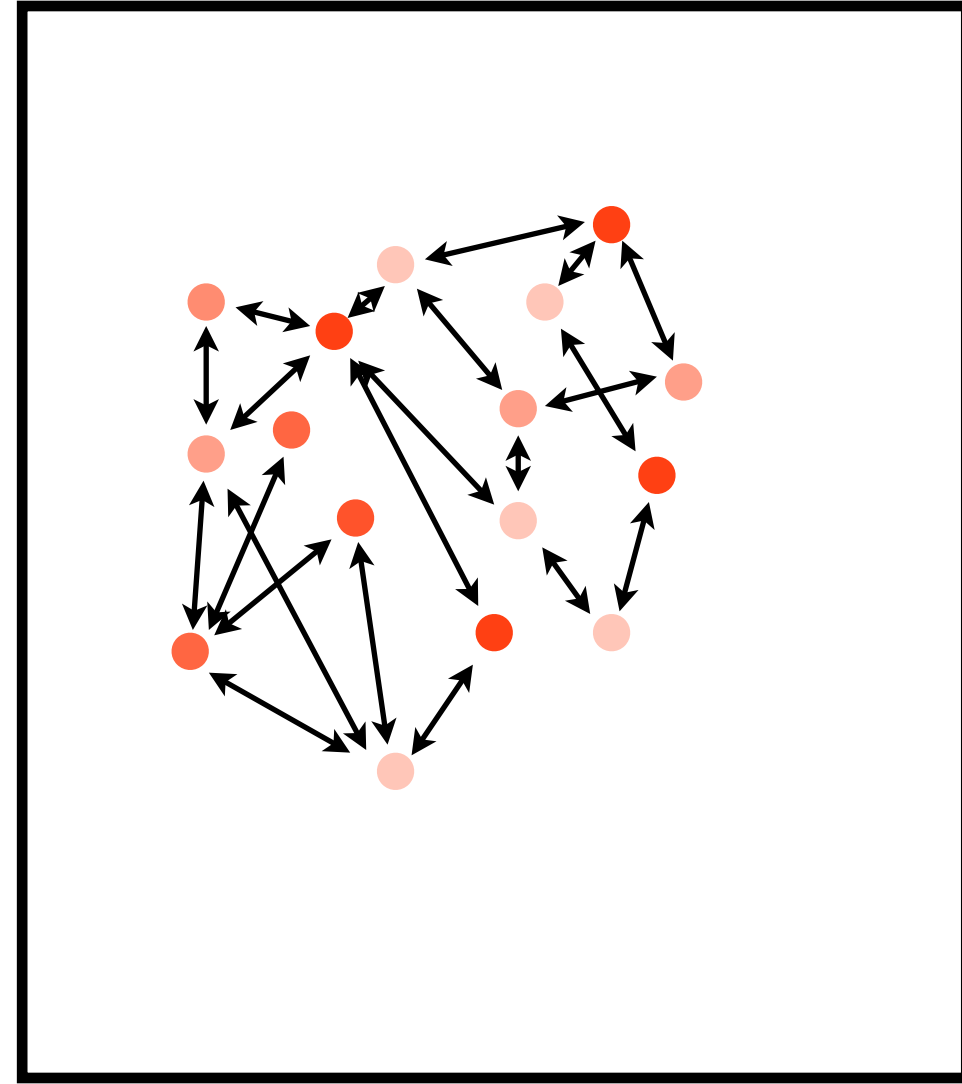


Tree Space

Topology Test Statistics

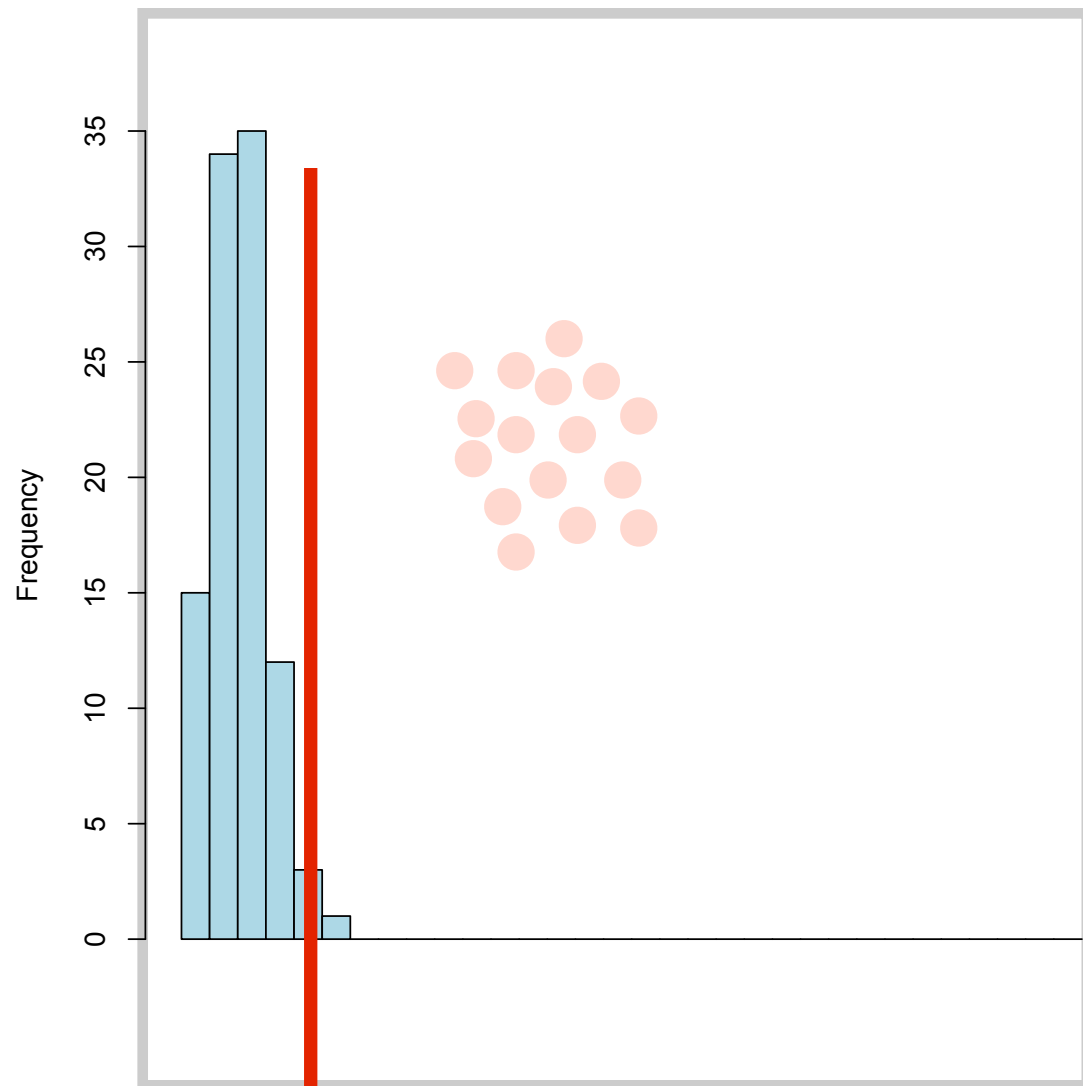


Tree Space

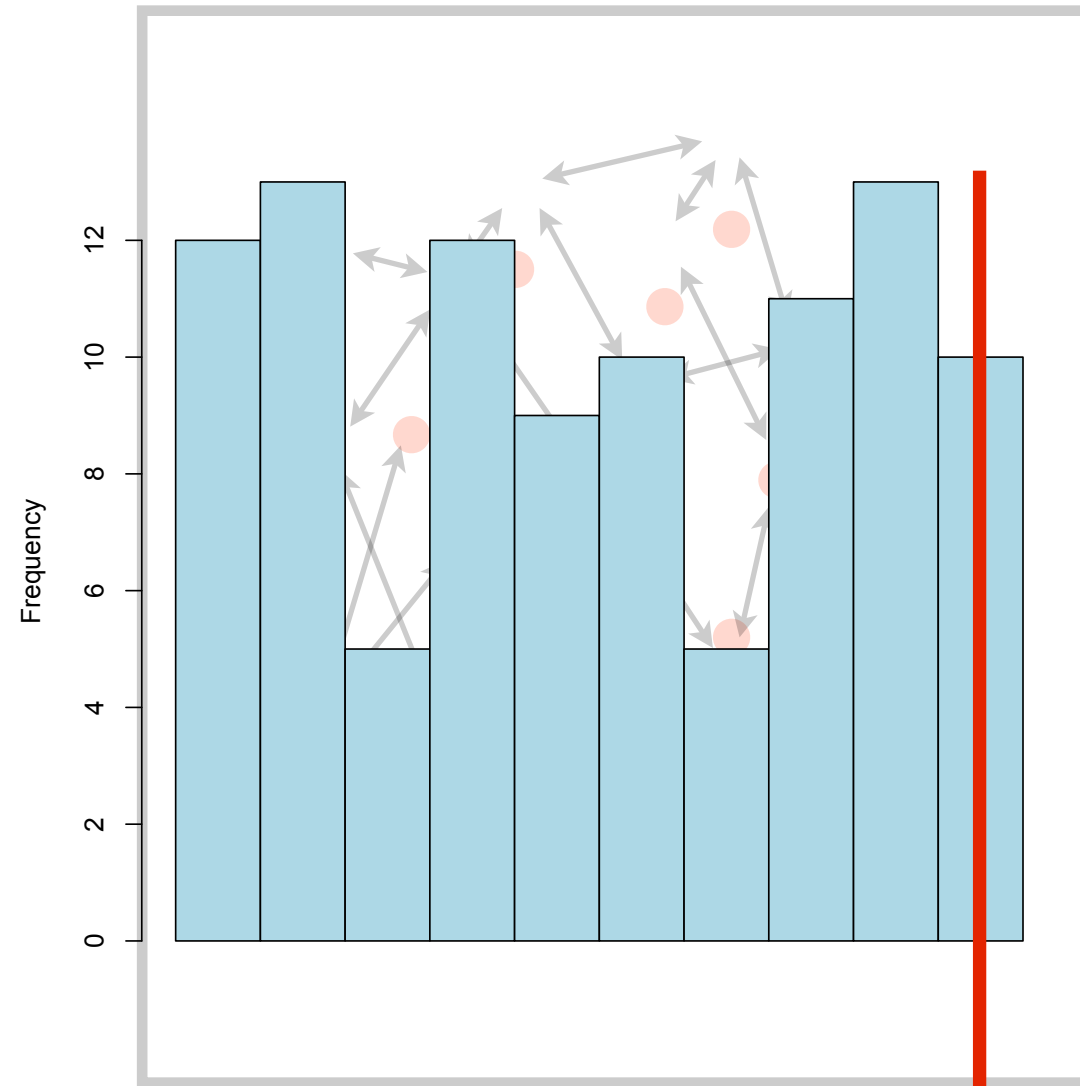


Tree Space

Topology Test Statistics



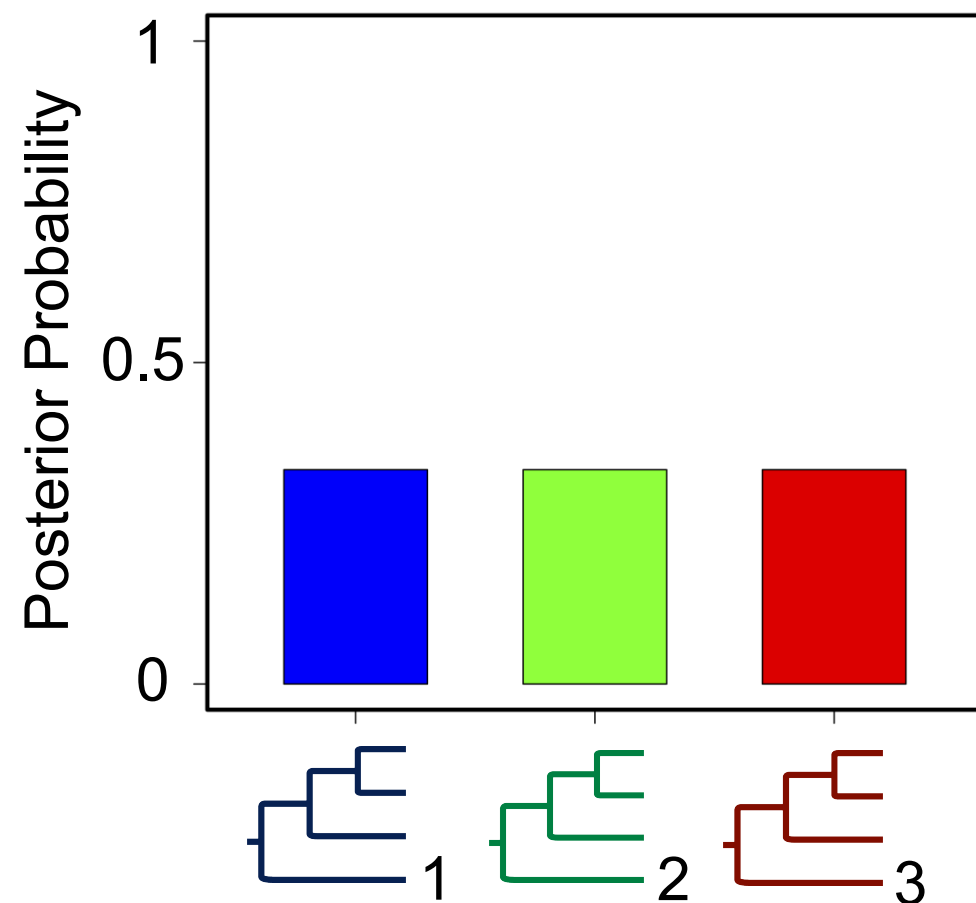
Tree-to-Tree Distance



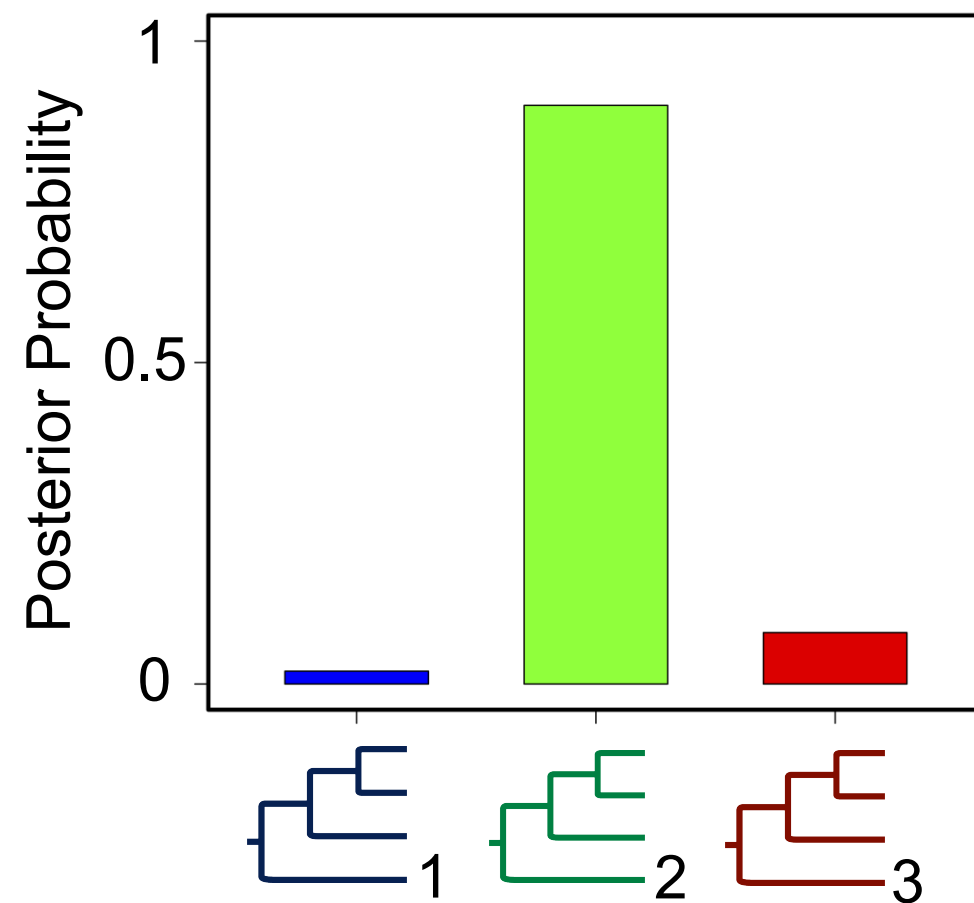
Tree-to-Tree Distance

Topology Test Statistics

Higher Entropy



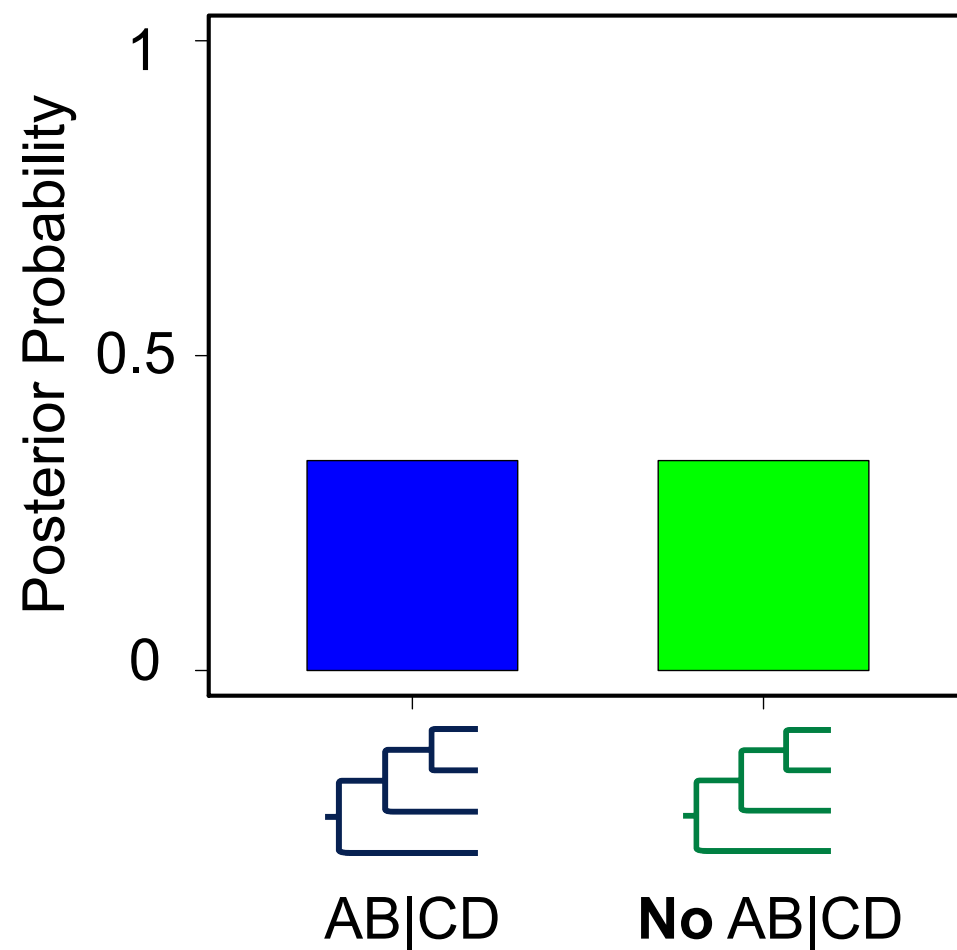
Lower Entropy



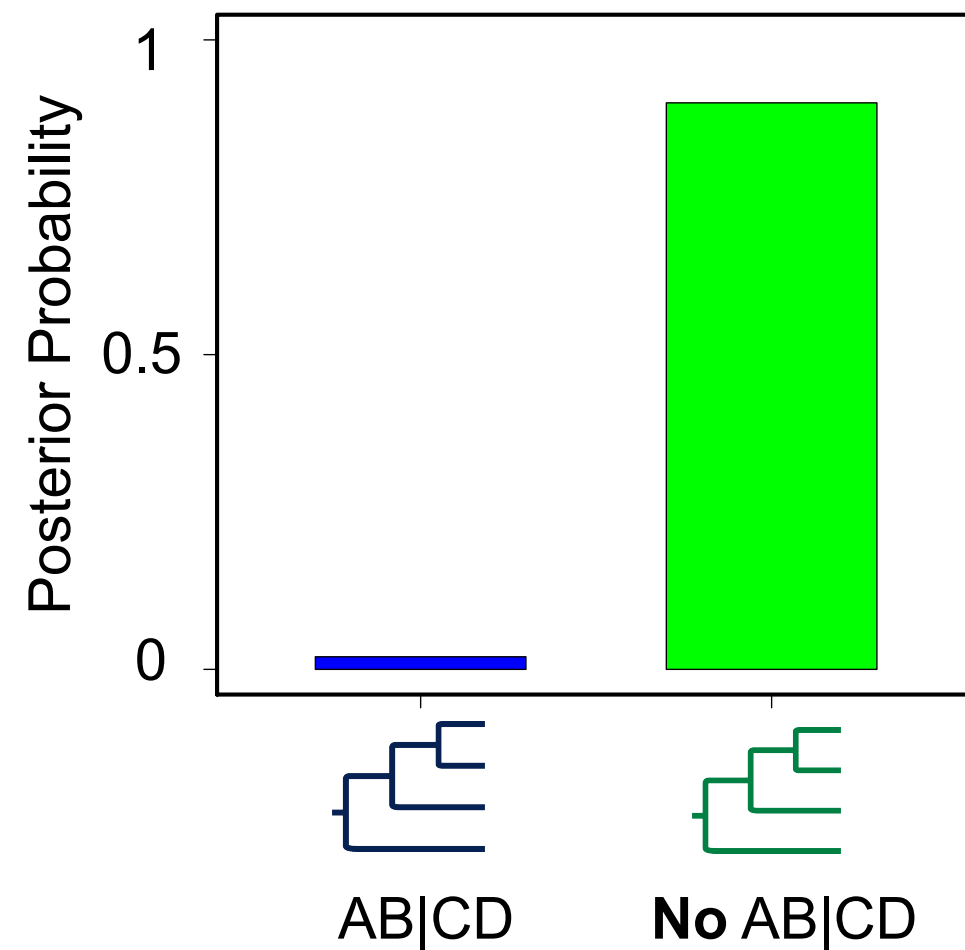
Branch-Specific Test Statistics

(not yet in RevBayes)

Higher Entropy

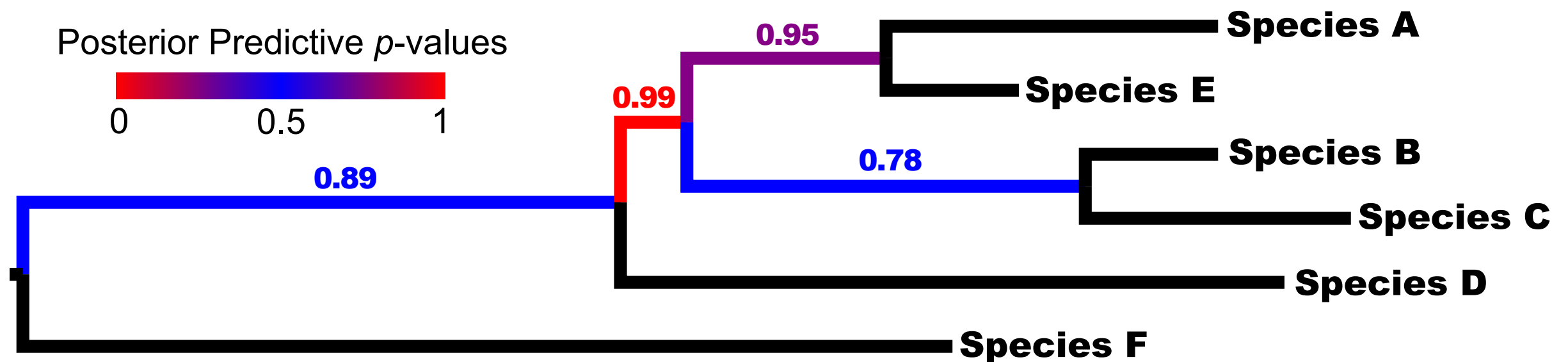


Lower Entropy

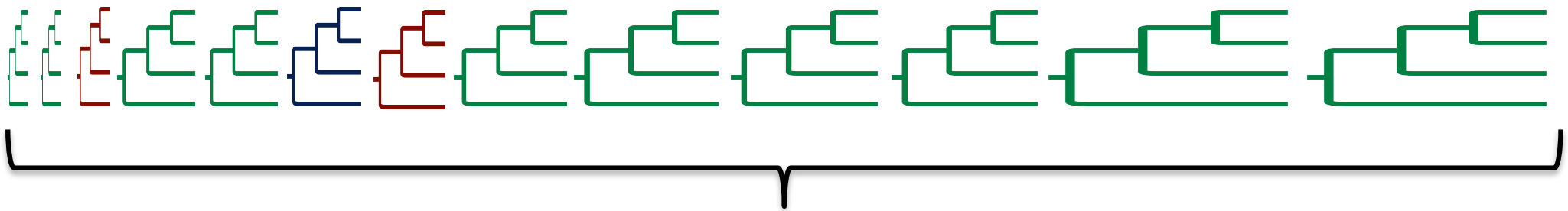


Branch-Specific Test Statistics

(not yet in RevBayes)



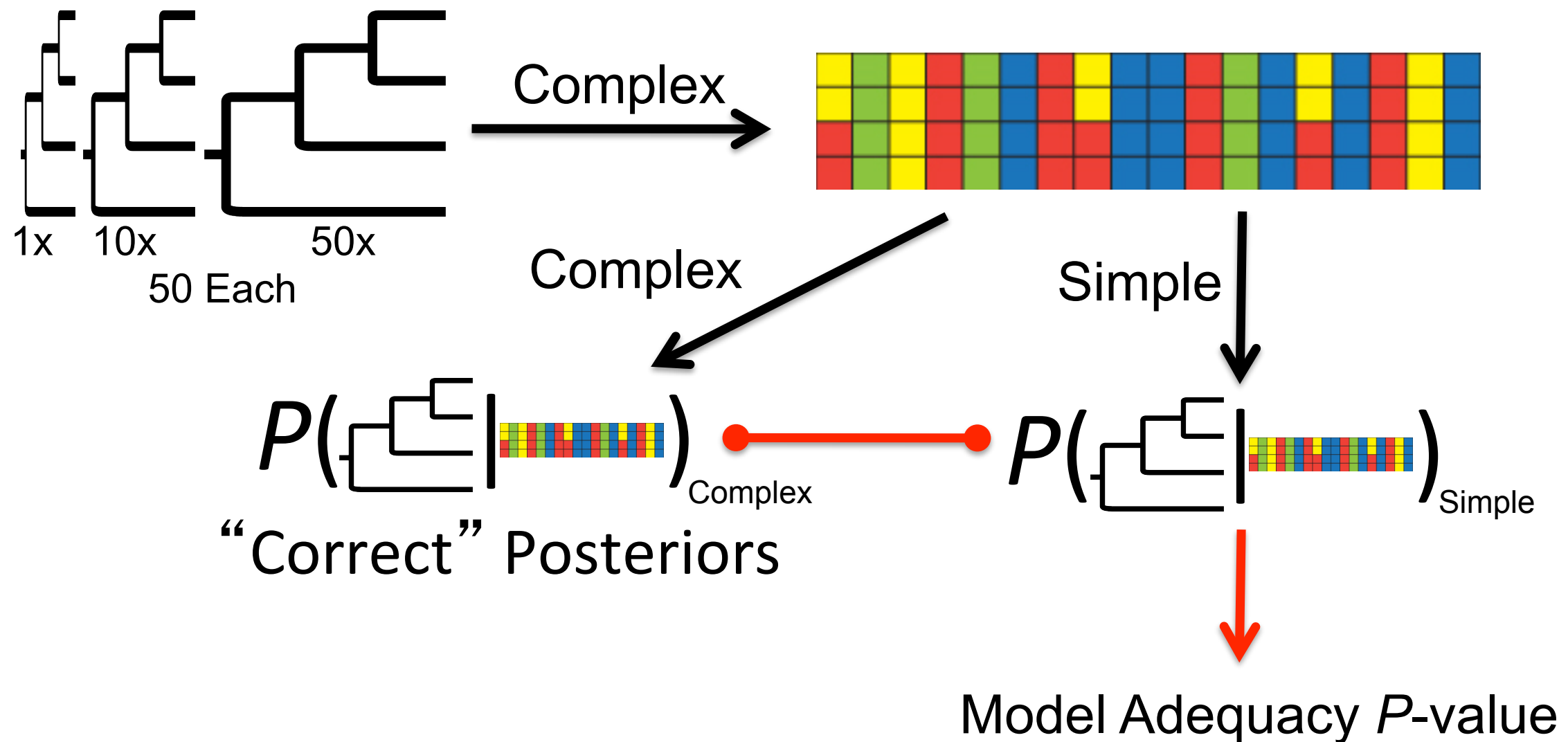
Branch-length Test Statistics



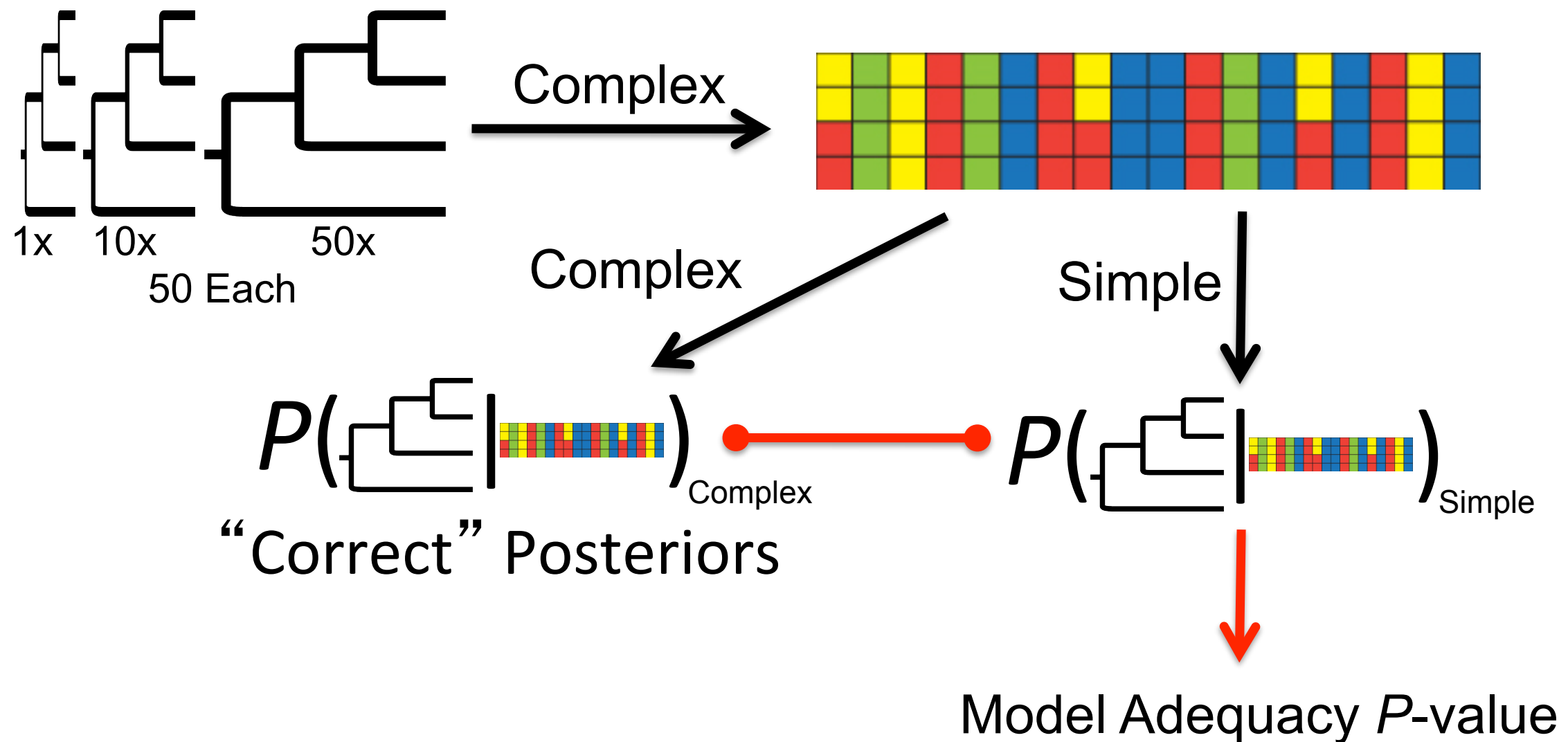
Mean Tree Length = 3.15
Variance in Tree Length = 2.30

Marginalizing across topologies

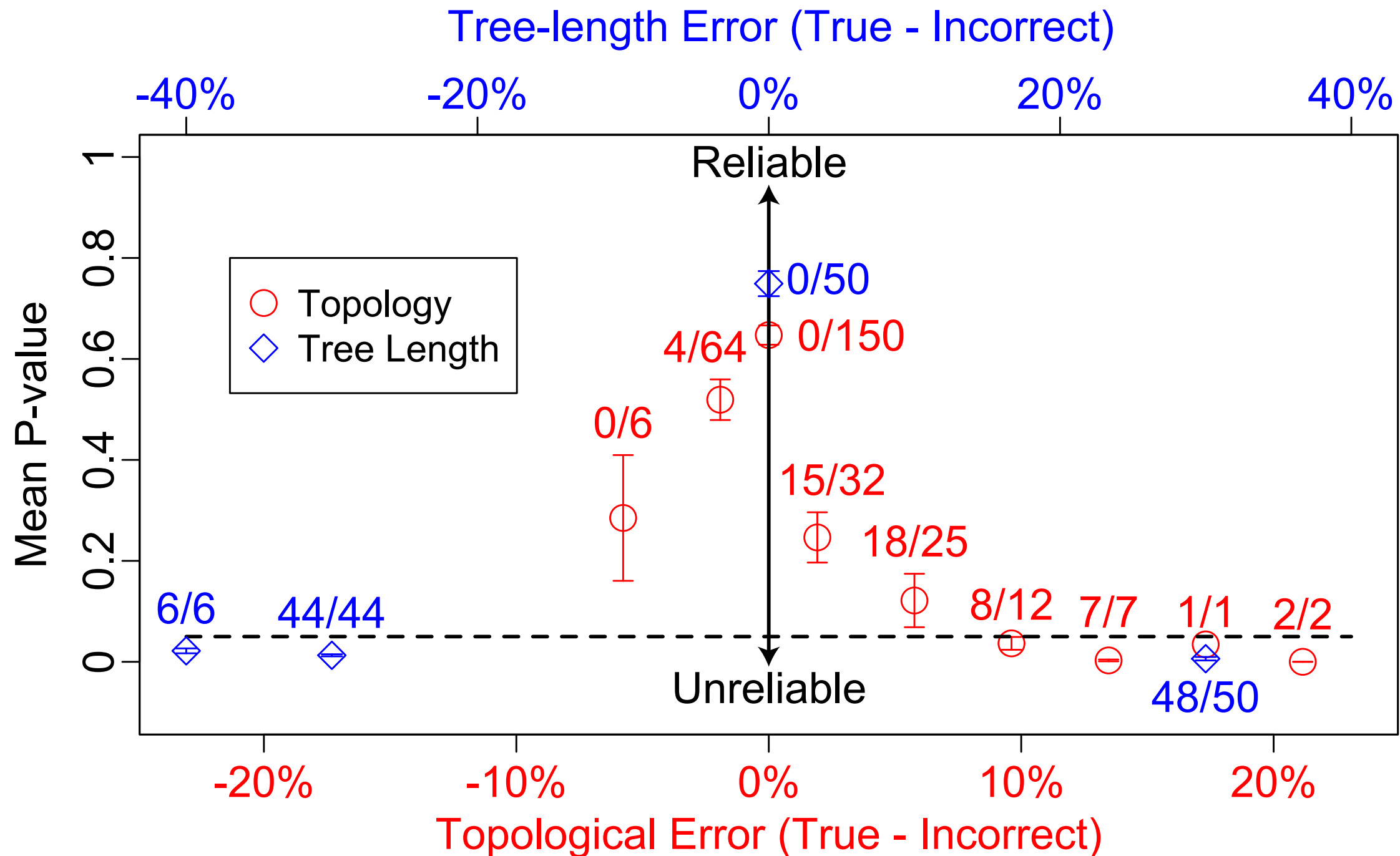
Motivating Results - Simulation



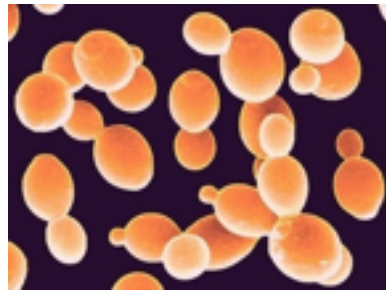
Motivating Results - Simulation



Motivating Results - Simulation



Motivating Results - Empirical



microbewiki.kenyon.edu

Yeast

343 orthologs

18 taxa

Hess & Goldman (2011)

Amniotes

1,145 orthologs

10 taxa

Crawford et al. (2012)



phylopic.org

Lowest

10%



DECILE 1
34 Genes

DECILE 2
34 Genes

DECILE 3
34 Genes

DECILE 4
34 Genes

DECILE 5
34 Genes

DECILE 6
34 Genes

DECILE 7
34 Genes

DECILE 8
34 Genes

DECILE 9
34 Genes

DECILE 10
34 Genes



Highest
10%

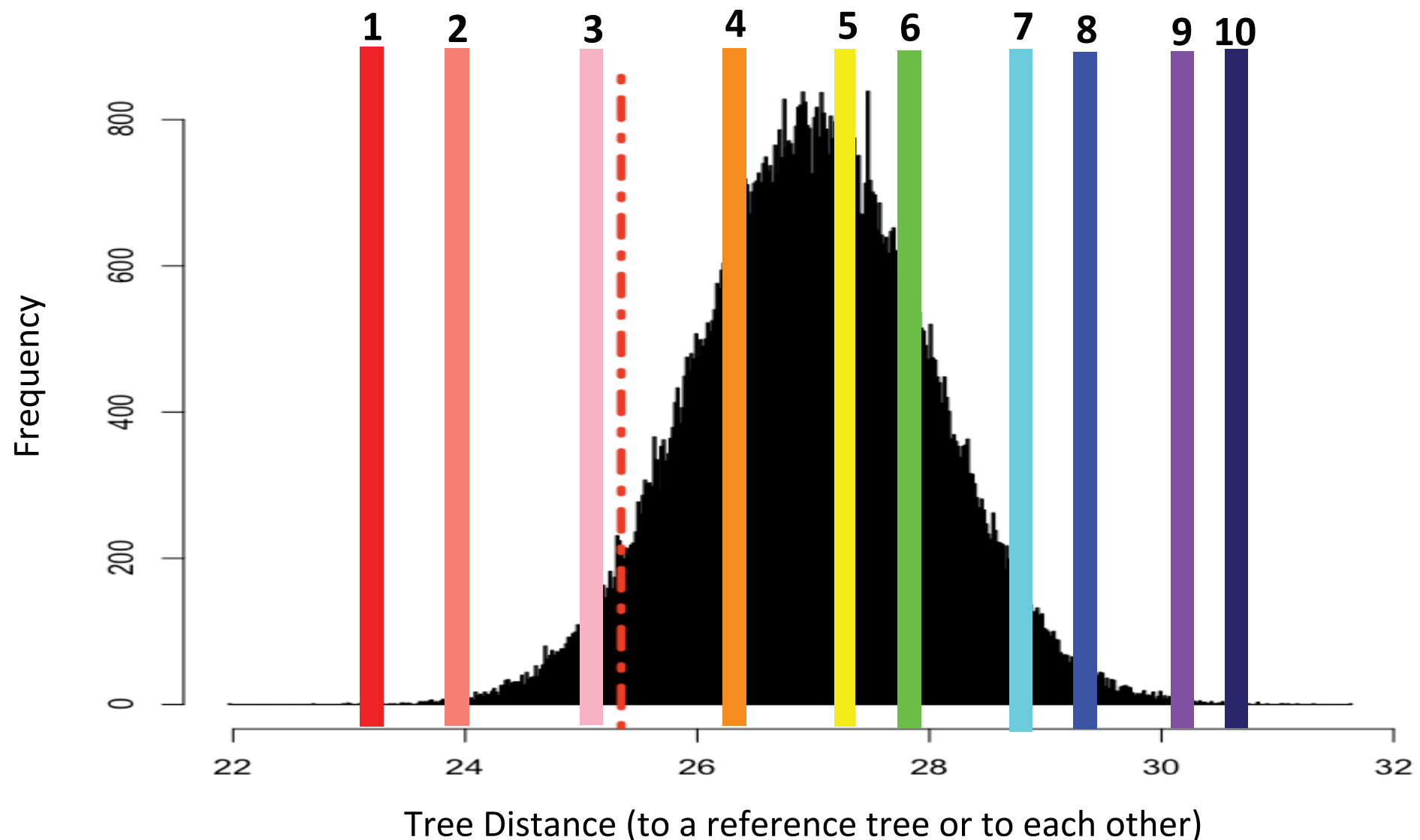


Motivating Results - Empirical

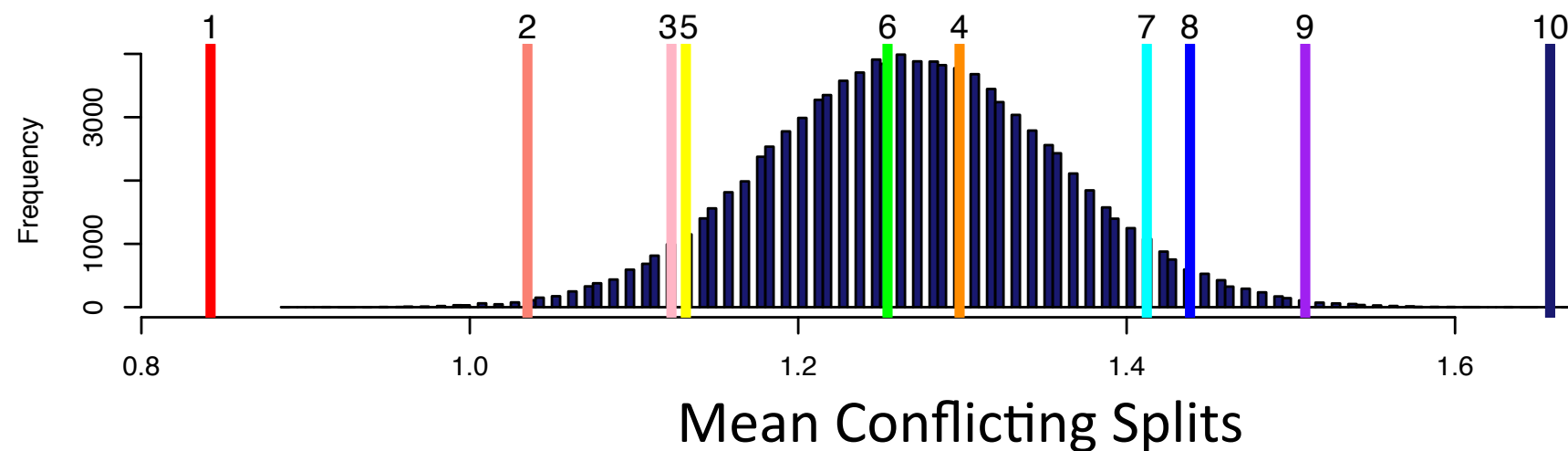
What might we expect from ideal filtering approaches?

Perfect association between decile membership and tree distance

$$\rho(r_s) = 1$$

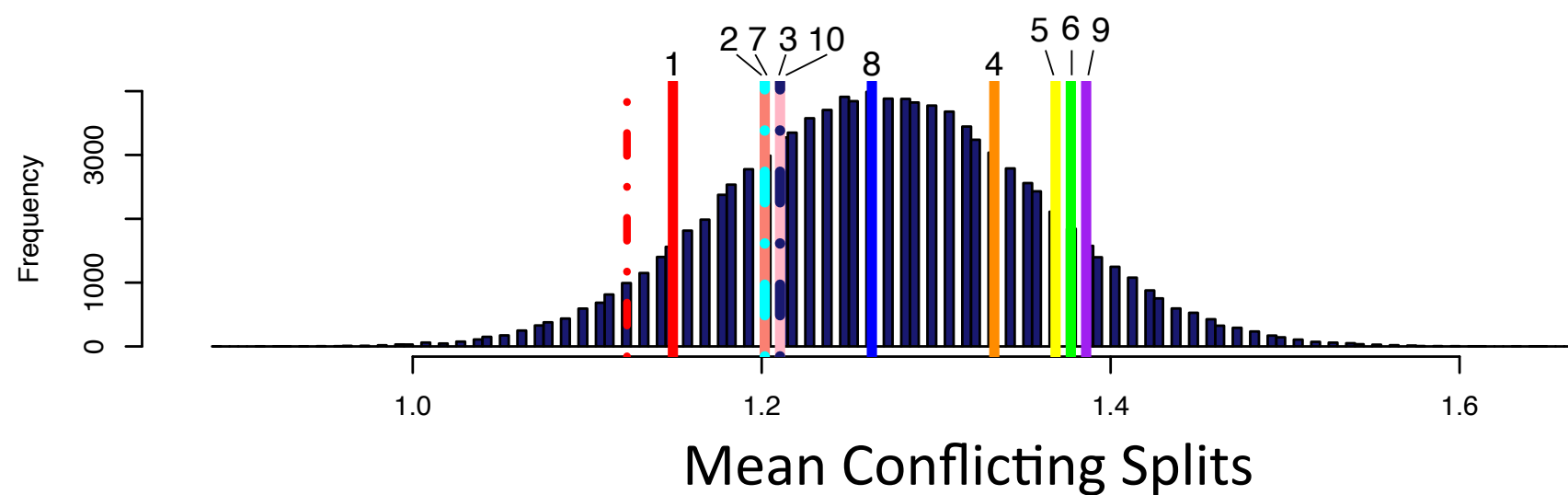


Motivating Results - Empirical



Posterior Predictive Filtering

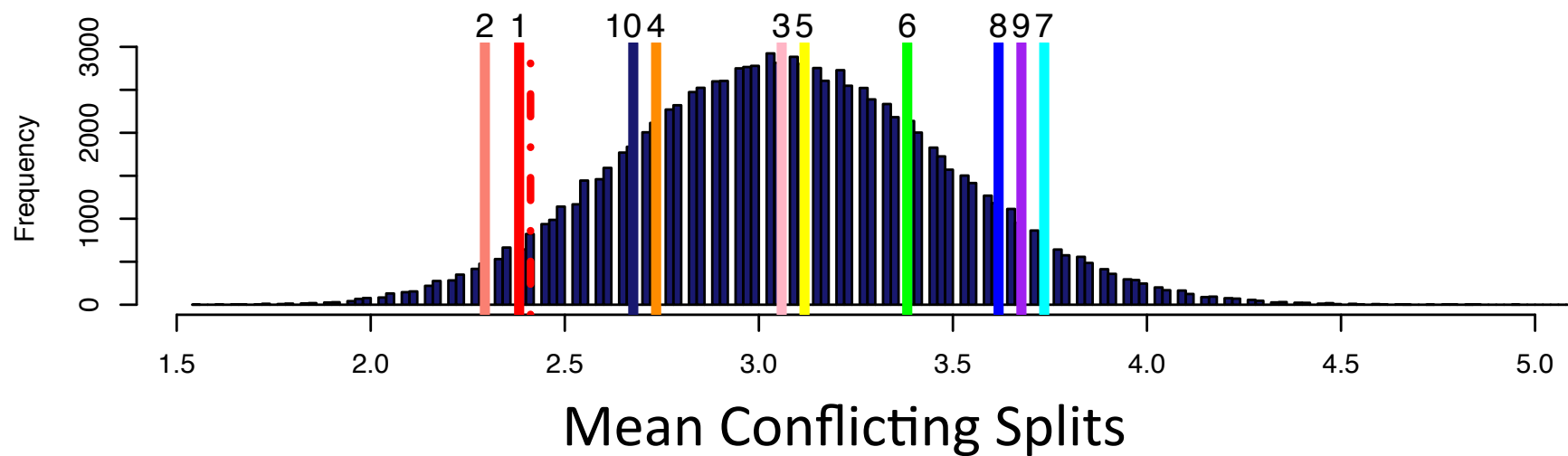
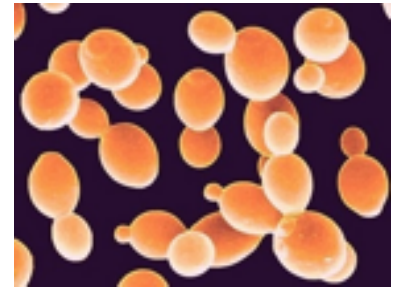
$$r_s=0.964, P=2.2 \times 10^{-16}$$



Rate Filtering

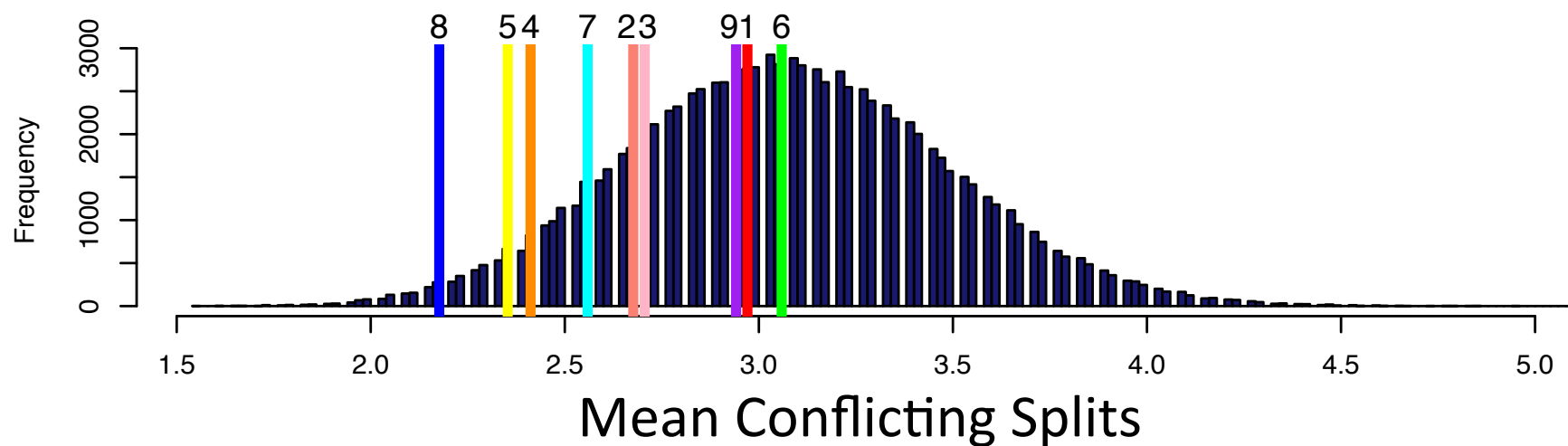
$$r_s=0.482, P=0.0793$$

Motivating Results - Empirical



Posterior Predictive Filtering

$$r_s=0.600, P=0.03656$$



Rate Filtering

$$r_s=0.103, P=0.3925$$

Active Development!

Our current inference-based statistics are computationally intense (lots of MCMC). We are:

- working on faster approximations for inference statistics
- conducting baseline simulation studies to establish power
- making the workflow easier and faster (including HPC)

Thoughts on Interpretation

- Assessing model fit is probably most useful with big data
- Not meant to be a hypothesis test. We can **always** reject the fit of a model in a strict sense. All models are abstractions.
- Based on the aspects of our model that don't fit well, think about how to structure new models. Remember, with RevBayes you can design your own new models!