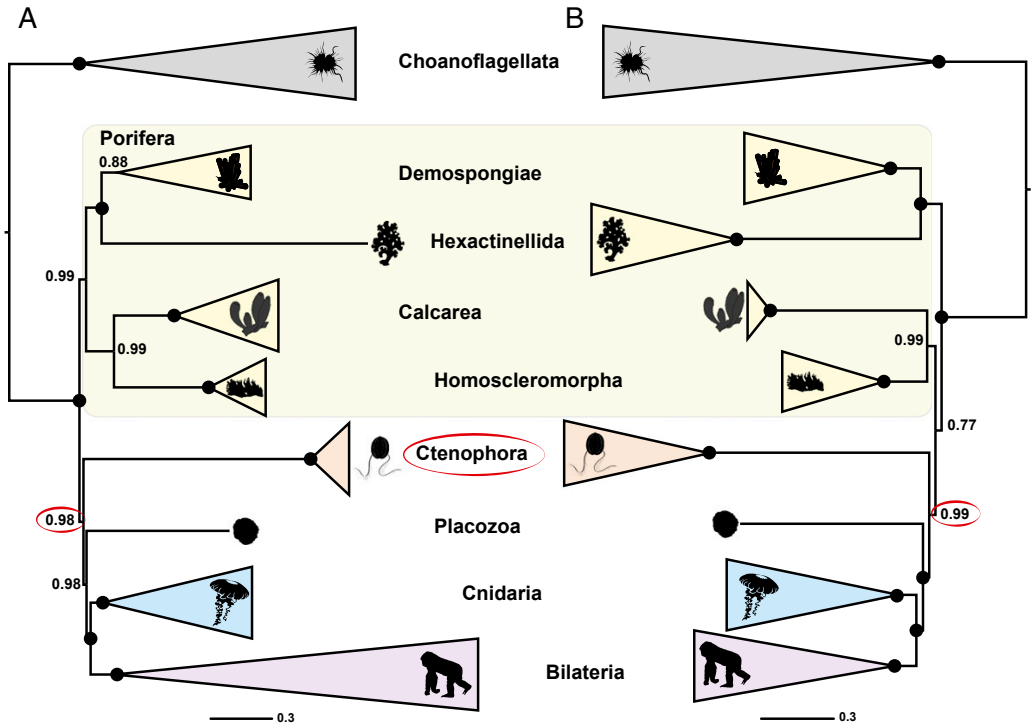
Assessing Model Fit

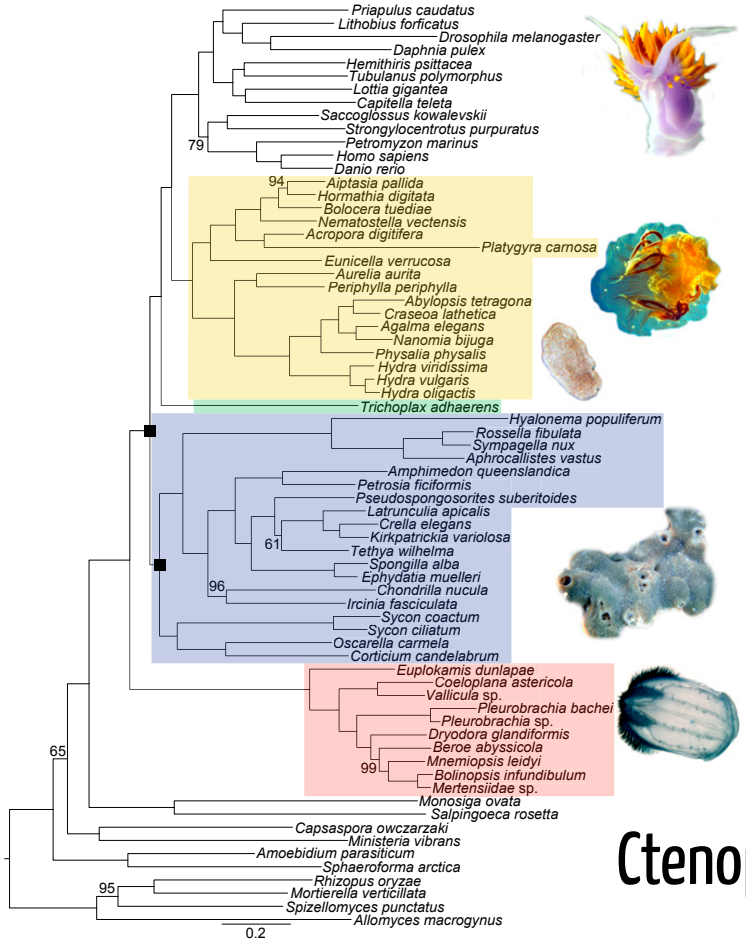
Evidence for Systematic Error (Bias)

Who are the earliest diverging animals?



Sponges

Pisani et al., PNAS, 2015

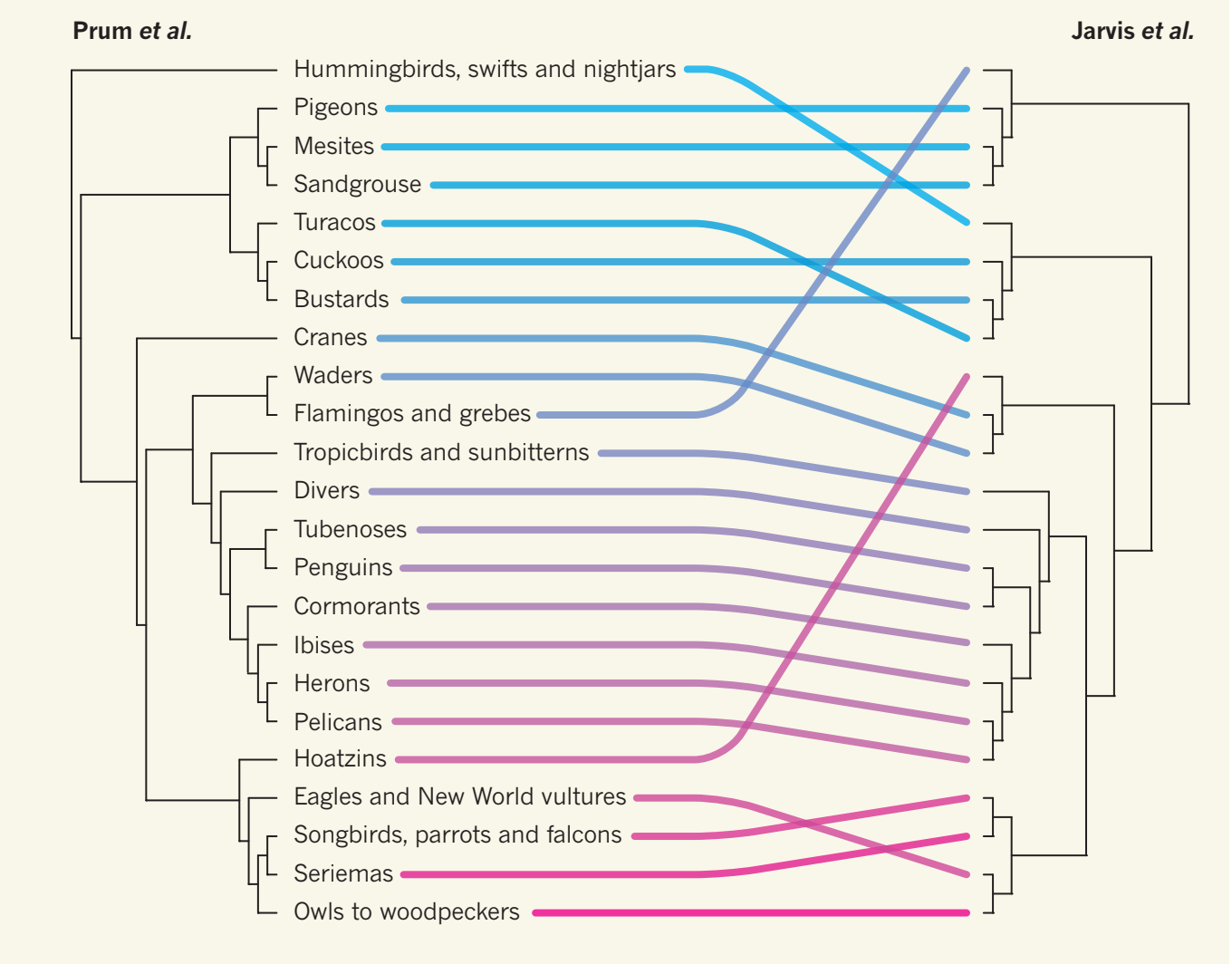


Ctenophores

Whelan et al., PNAS, 2015

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)

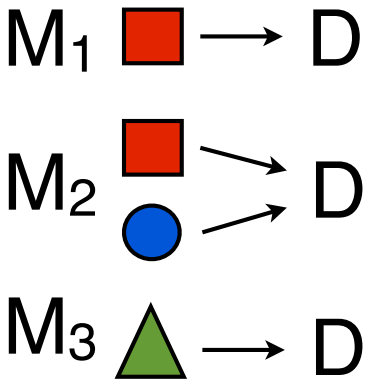
{3.4, 2.1, 5.4, ...}

The Standard Approach

(1) Collect Data (D)

{3.4, 2.1, 5.4, ...}

(2) Define Models



The Standard Approach

(1) Collect Data (D)

{3.4, 2.1, 5.4, ...}

(2) Define Models

M_1 → D

M_2 → D

M_3 → D

(3) Fit Models

M_1 → D : L_1

M_2 → D : L_2

M_3 → D : L_3

The Standard Approach

(1) Collect Data (D)

{3.4, 2.1, 5.4, ...}

(2) Define Models

M_1 → D

M_2 → D

M_3 → D

(3) Fit Models

M_1 → D : L_1

M_2 → D : L_2

M_3 → D : L_3

(4) Compare Models & Choose "Best"

$M_2 > M_1 > M_3$ ← AIC, BIC, BF, LRT ← L_1, L_2, L_3

The Standard Approach

(1) Collect Data (D)

{3.4, 2.1, 5.4, ...}

(2) Define Models

M_1 ■ → D

M_2 ■ ● → D

M_3 ▲ → D

(3) Fit Models

M_1 ■ → D : L_1

M_2 □ ○ → D : L_2

M_3 ▲ → D : L_3

(4) Compare Models & Choose "Best"

$M_2 > M_1 > M_3$ ← AIC, BIC, BF, LRT } L_1, L_2, L_3

(5) Report Inferences from "Best" Model

Nature 2012

The Shapes of our Data

Intro — Blah blah blah blah. Blah blah blah blah blah blah. Blah blah blah important blah blah blah.Science blah blah blah blah blah.

Methods — Blah blah blah blah. Blah blah blah liklihood blah 3 models. Blah blah blah AIC blah blah blah.Science blah blah blah blah.

Results — M_2 !

Conclusions — Both squares and circles are important.

The Standard Approach

(1) Collect Data (D)

{3.4, 2.1, 5.4, ...}

(5) Report Inferenc
"Best" Mode

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 rodent[1]. 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

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

Nature 2012

The Shapes of our Data

Intro Blah blah blah blah. Blah blah blah blah blah blah. Blah blah blah important blah blah blah.Science blah blah blah blah blah.

Methods Blah blah blah blah. Blah blah blah liklihood blah 3 models. Blah blah blah AIC blah blah blah.Science blah blah blah blah.

Results M_2 !

Conclusions Both squares and circles are important.

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

Nature 2012

The Shapes of our Data

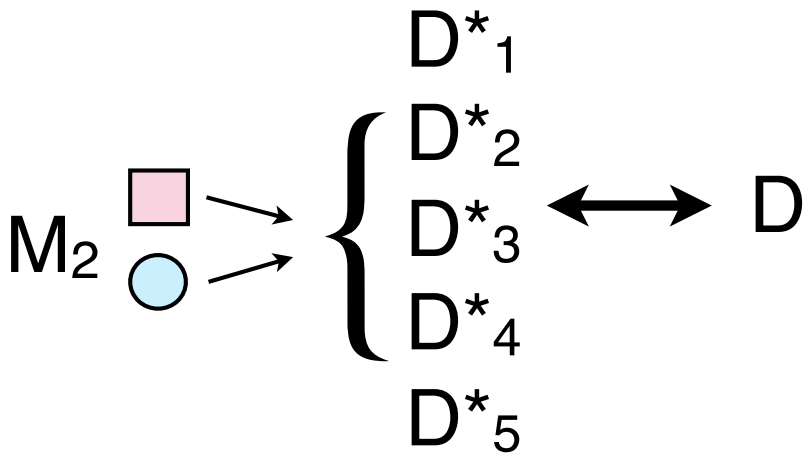
Intro Blah blah blah blah. Blah blah blah blah blah blah. Blah blah blah important blah blah blah.Science blah blah blah blah blah.

Methods Blah blah blah blah. Blah blah blah liklihood blah 3 models. Blah blah blah AIC blah blah blah.Science blah blah blah blah.

Results M_2 !

Conclusions Both squares and circles are important.

(6) Check Fit of Model to Data



The Next Step - Assessing Fit

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

Jack Sullivan[1,2] and David L. Swofford[1]

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.

How might we assess fit?

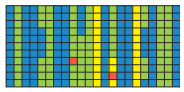
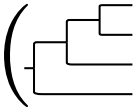
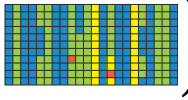
(1) Use our **prior knowledge** to ask if the **data** are reasonable.

(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.

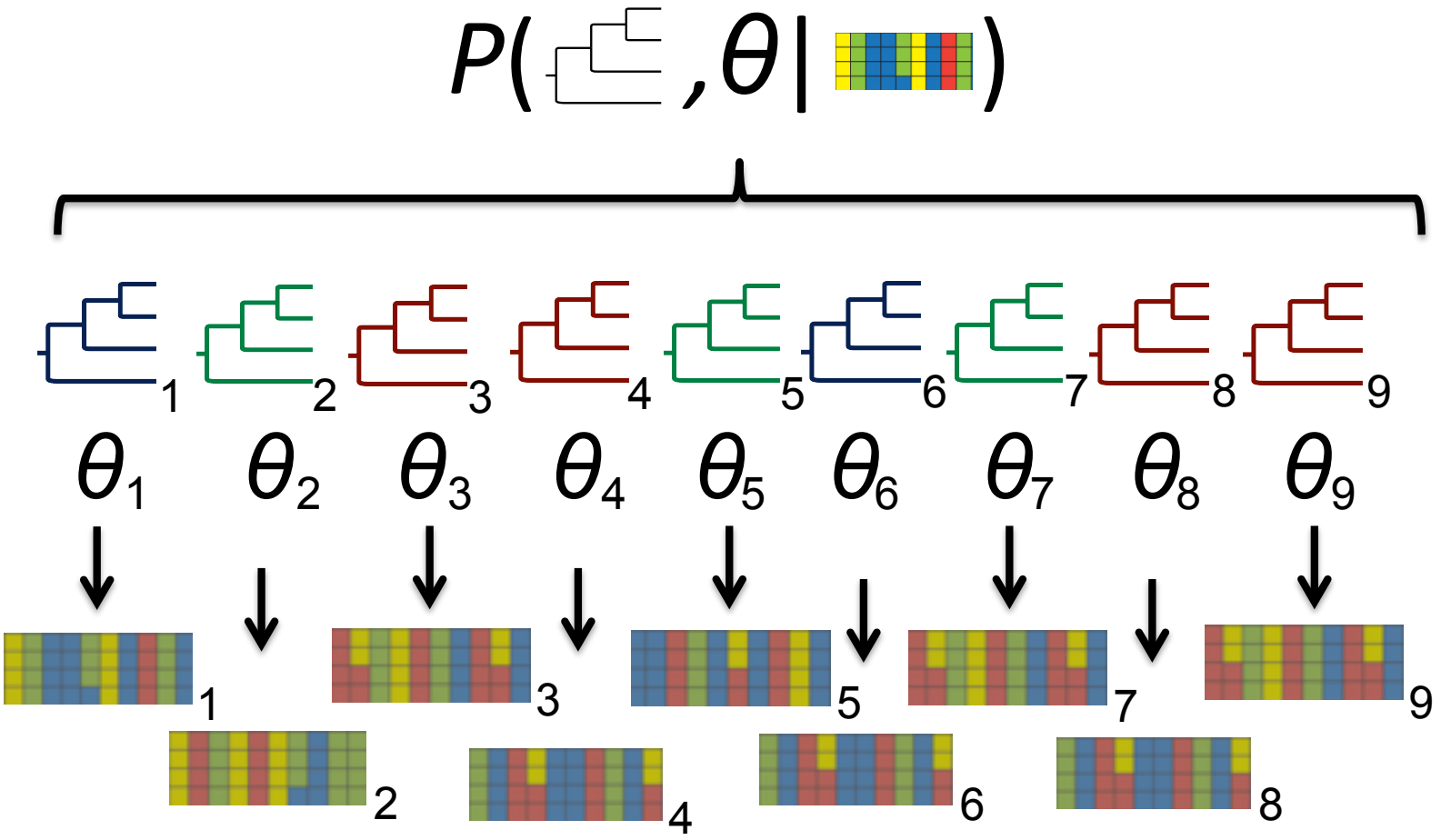
(3) Use your **data** (all or part) to make a **prediction** and see if your prediction matches what you've seen.
(Posterior Prediction and Cross Validation)

Posterior Prediction

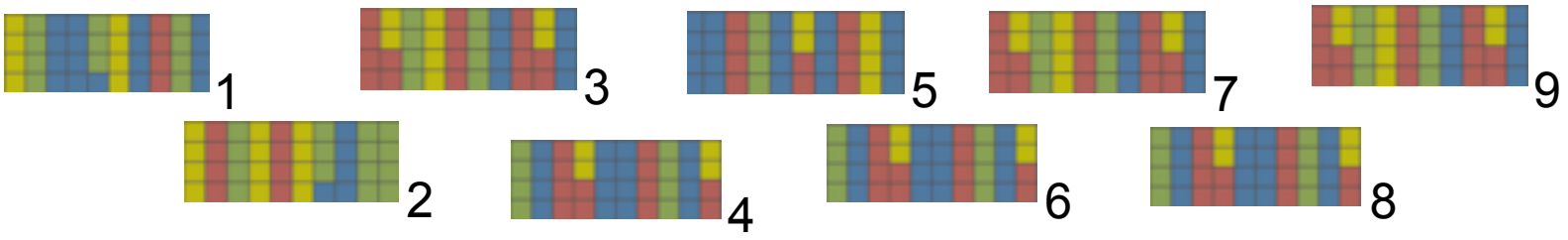
Could  have come from $P($  $, \theta |$  $)$?

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

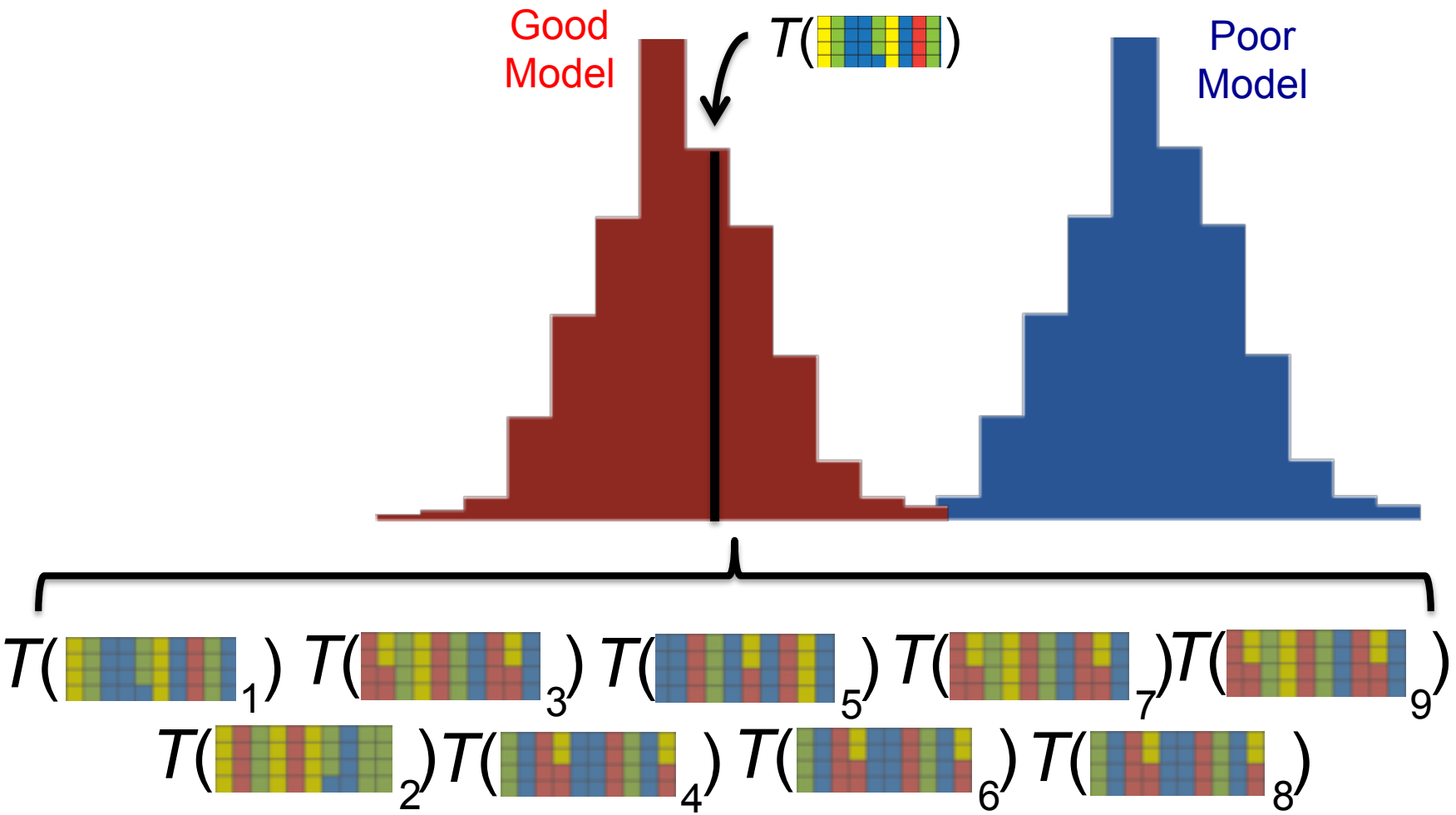
Posterior Prediction



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?' "

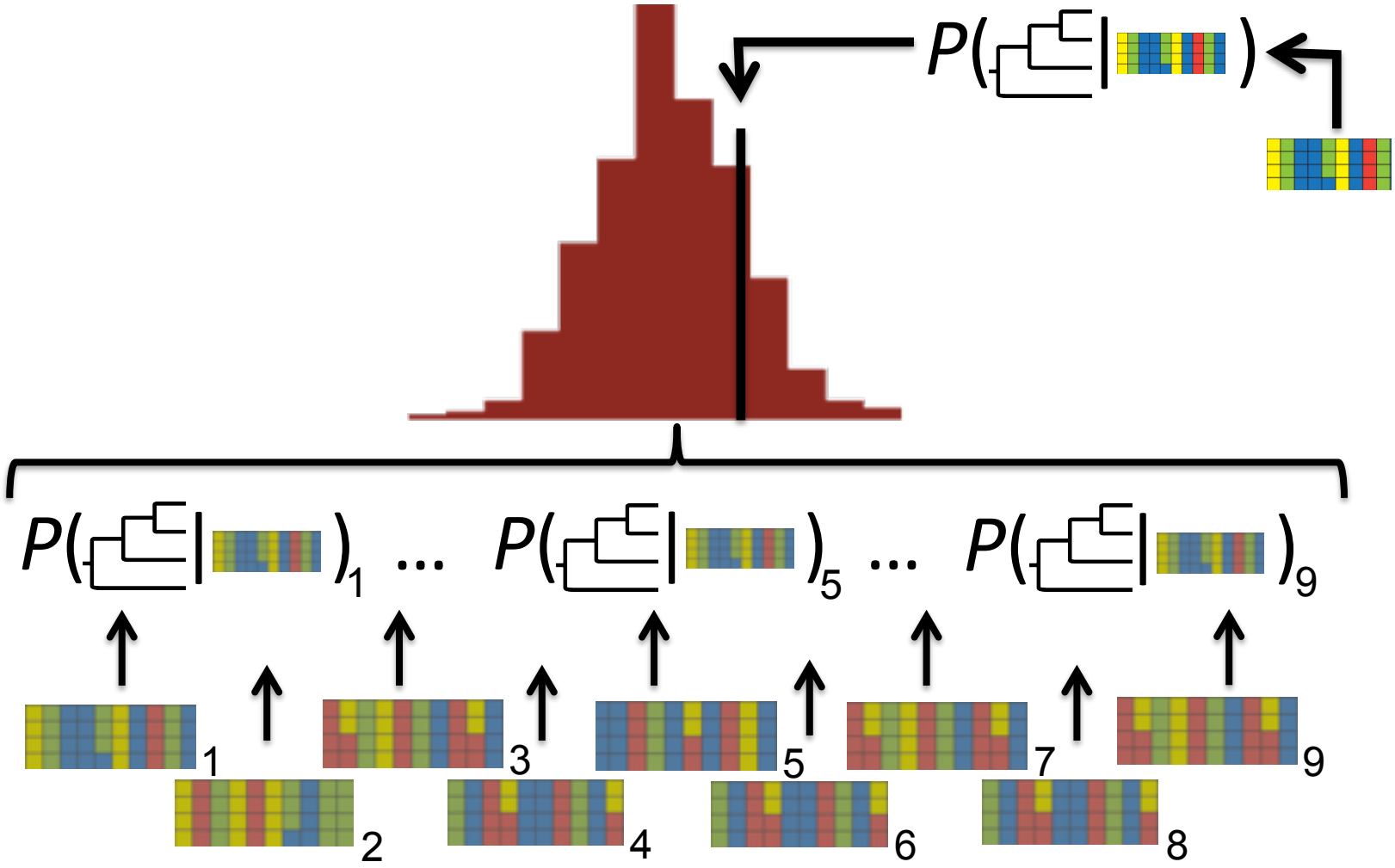
- Gelman, Carlin, Stern, and Rubin

Bayesian Data Analysis

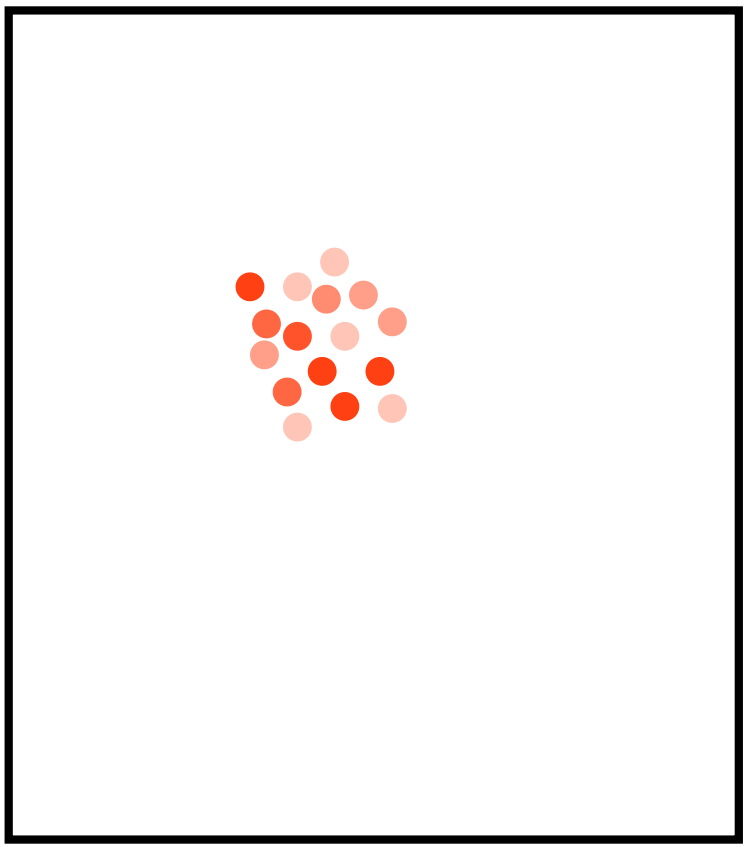
Posterior Prediction

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

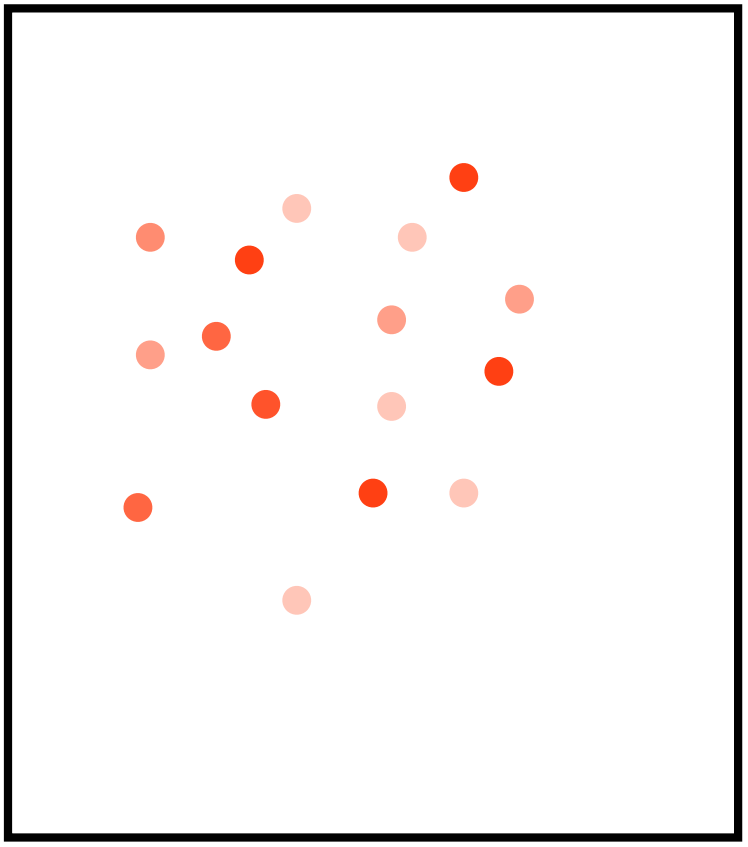
Posterior Prediction



Topology Test Statistics

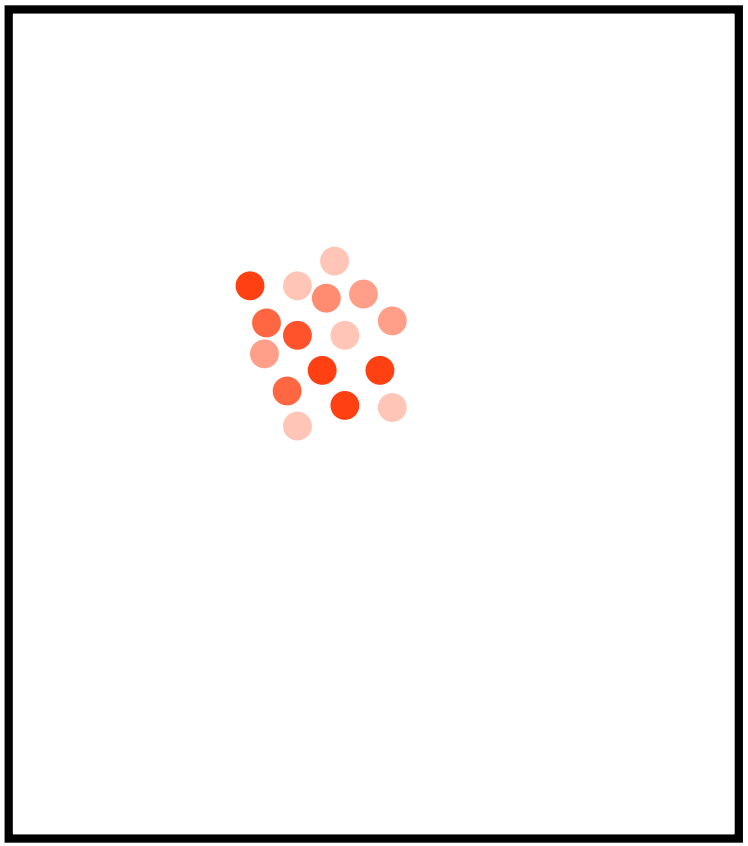


Tree Space

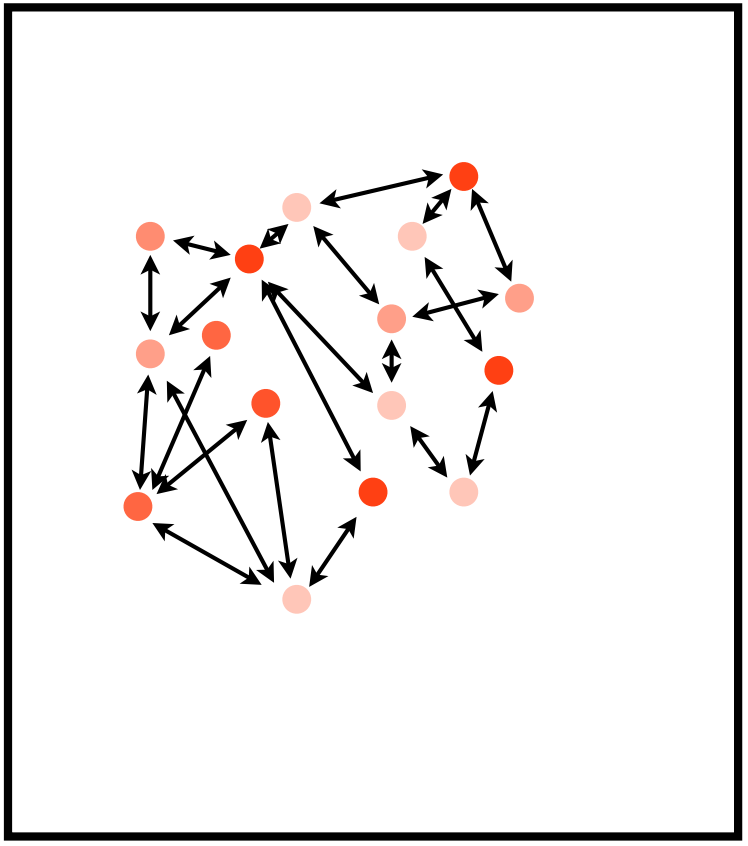


Tree Space

Topology Test Statistics

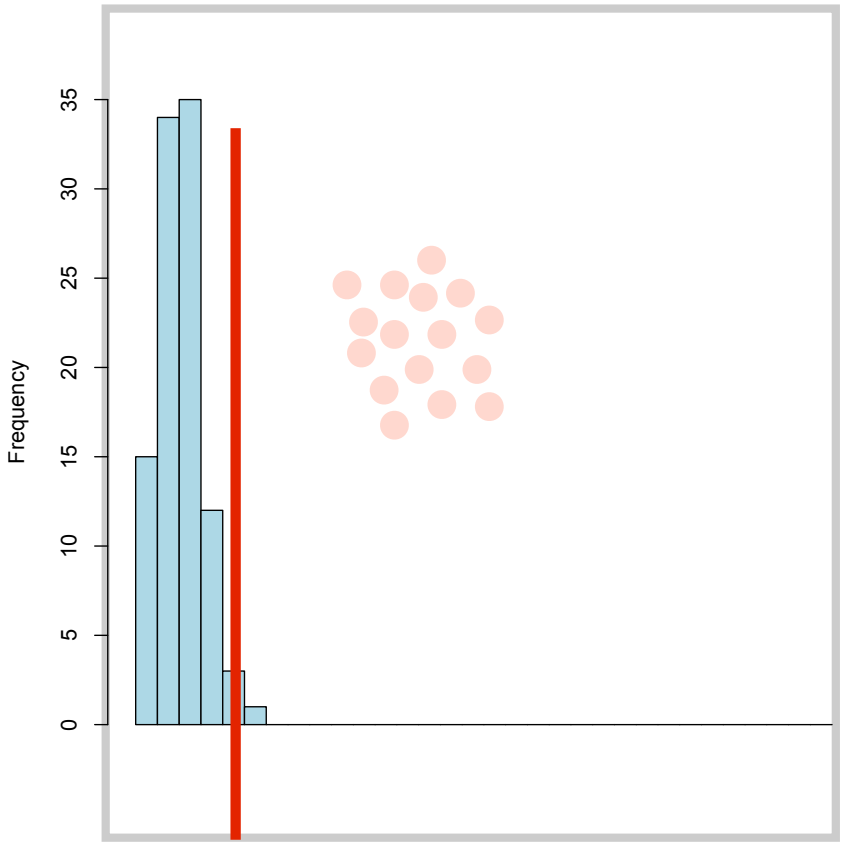


Tree Space

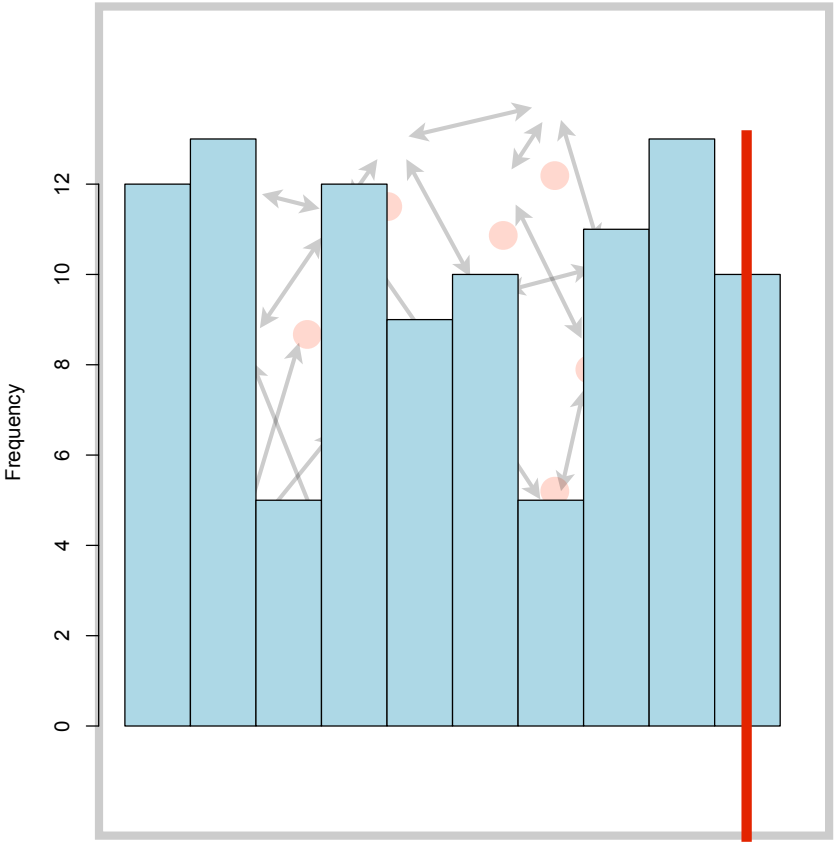


Tree Space

Topology Test Statistics

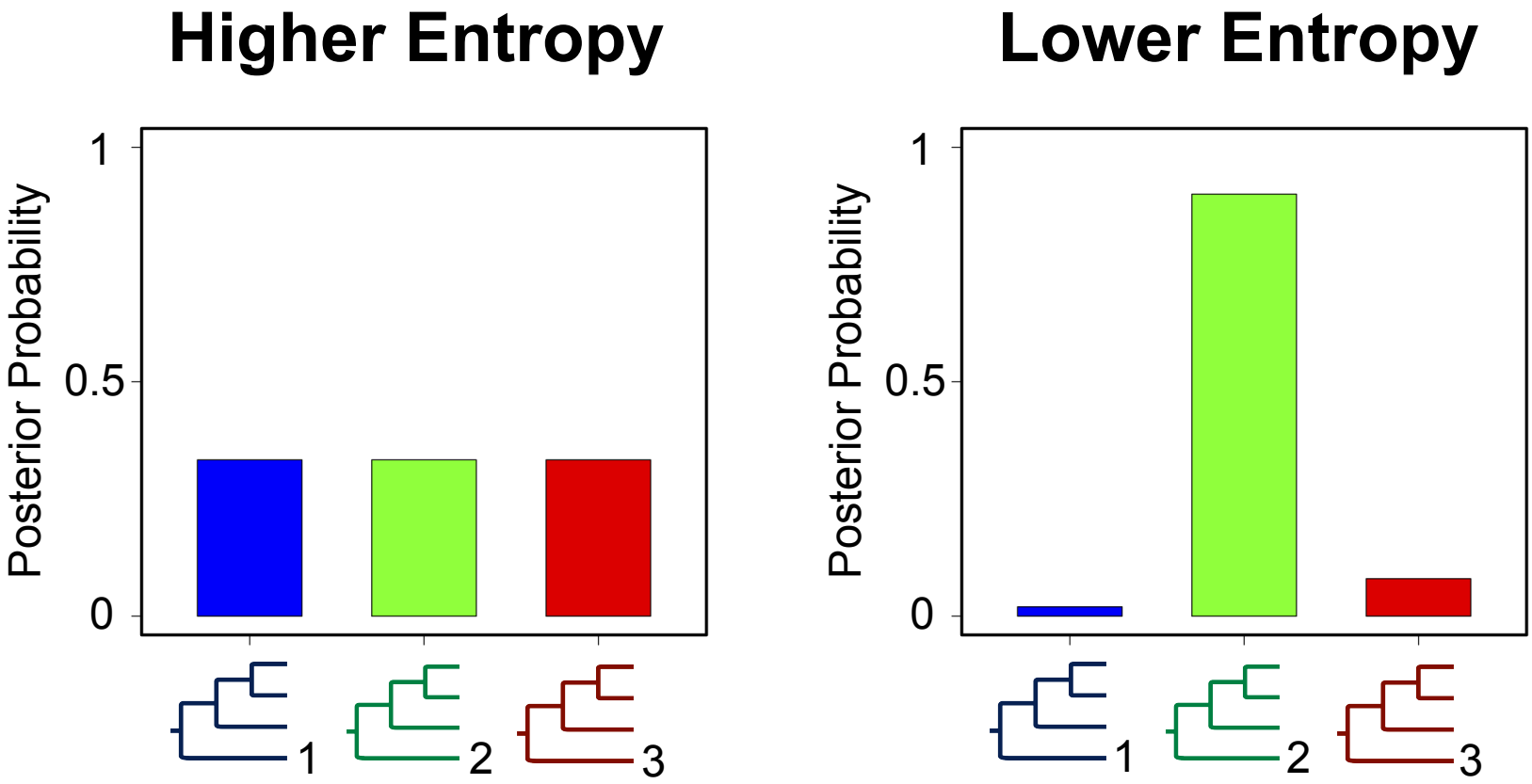


Tree-to-Tree Distance



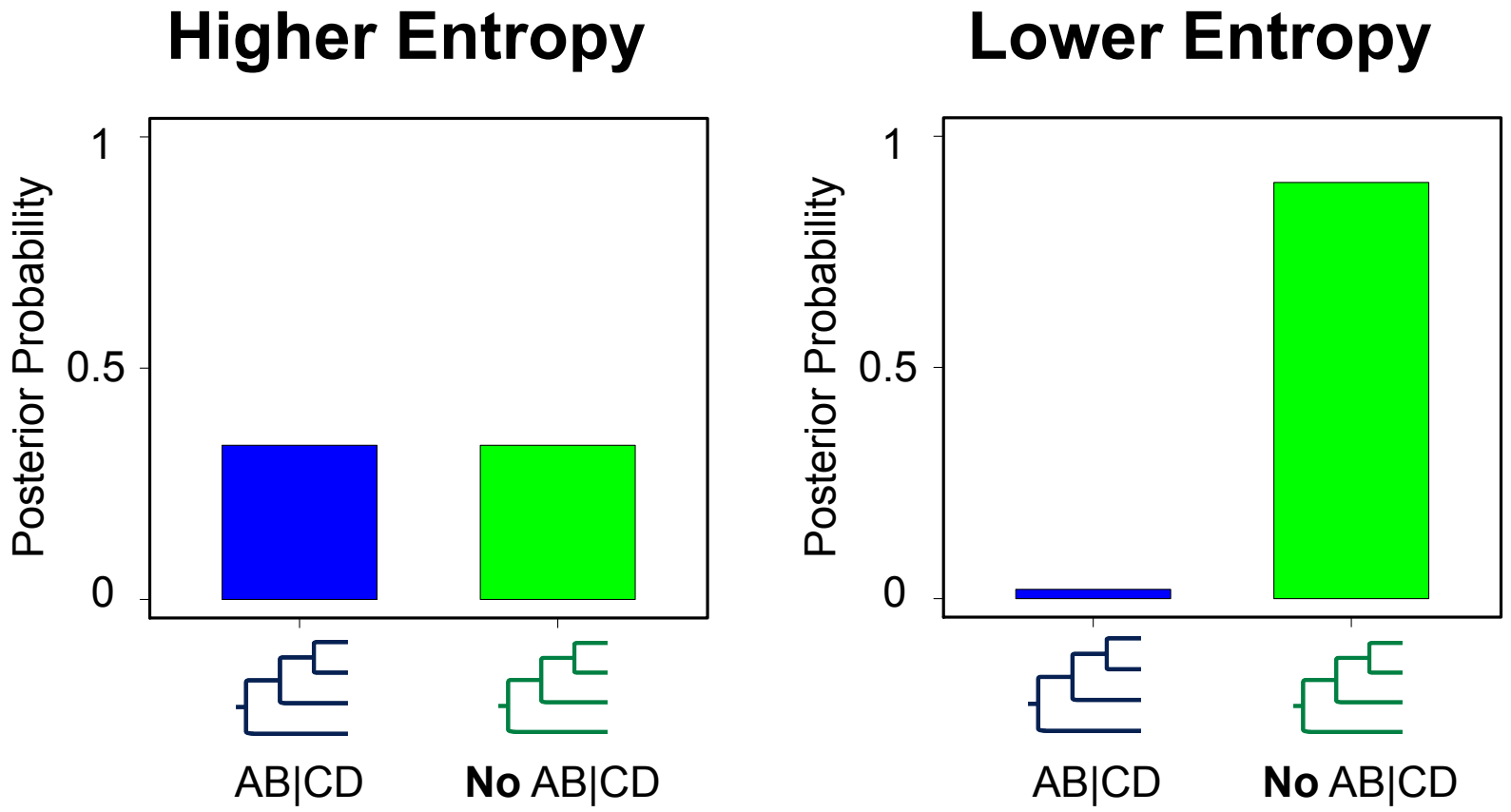
Tree-to-Tree Distance

Topology Test Statistics



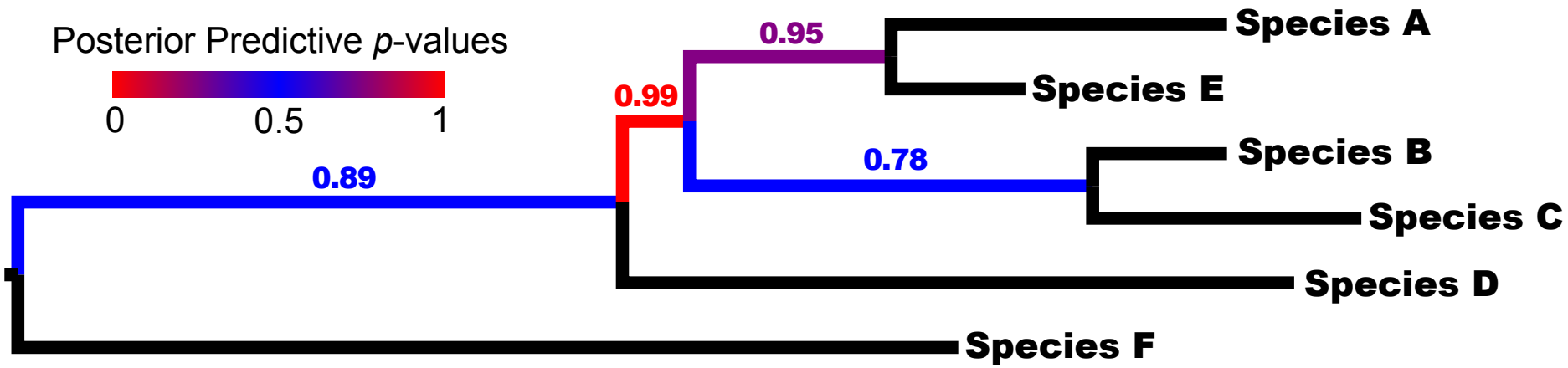
Branch-Specific Test Statistics

(not yet in RevBayes)

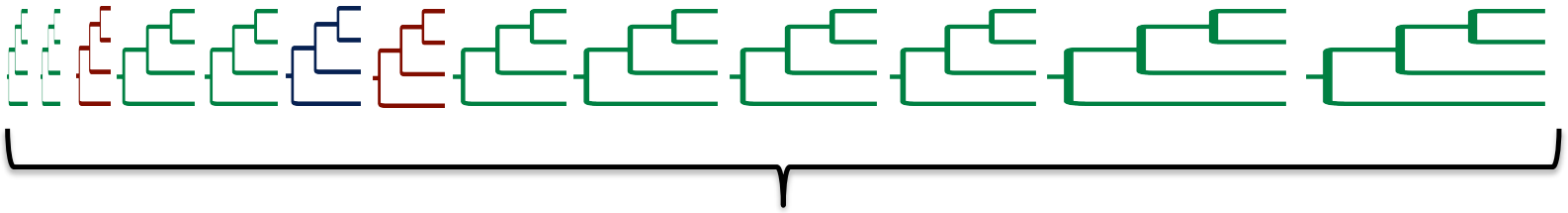


Branch-Specific Test Statistics

(not yet in RevBayes)



Branch-length Test Statistics

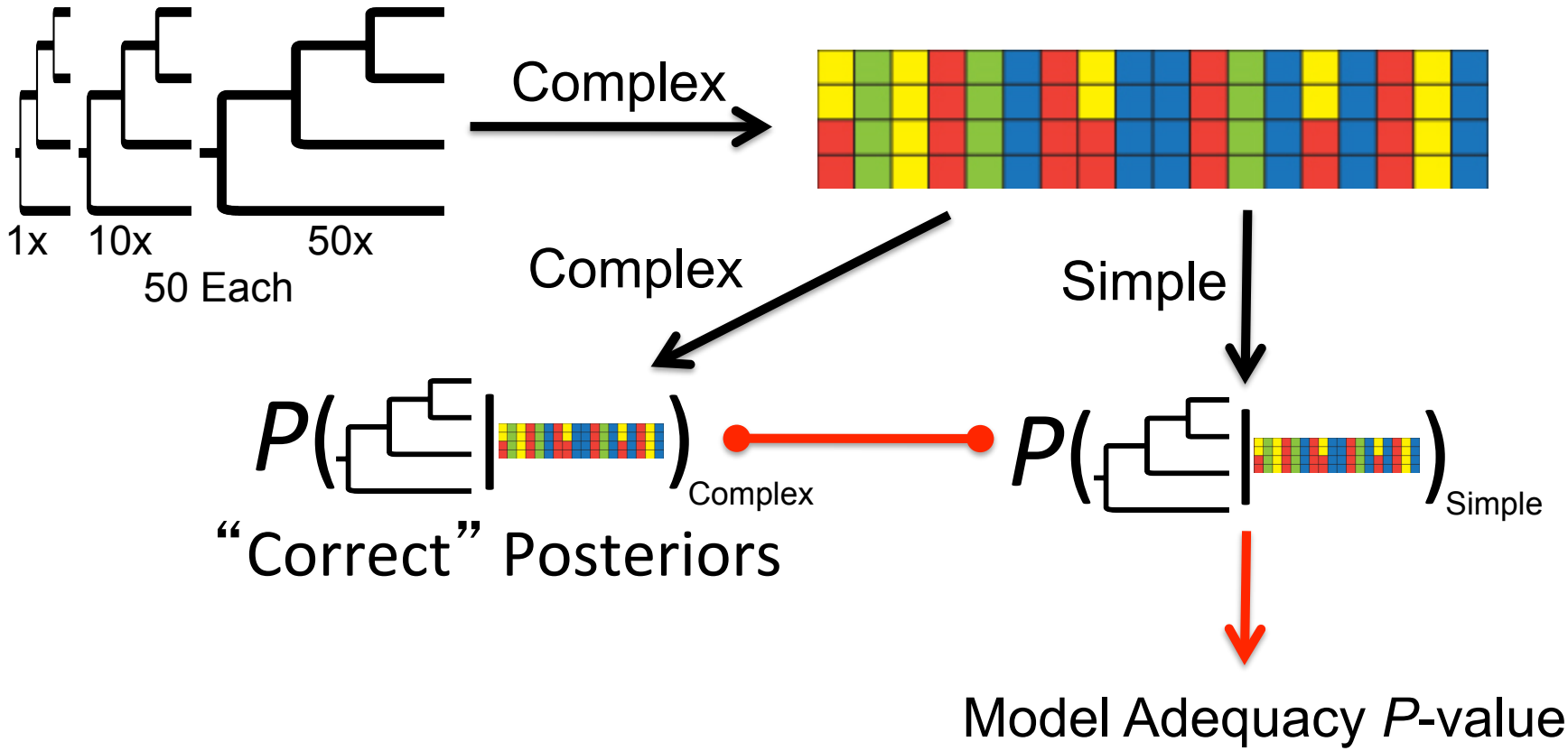


Mean Tree Length = 3.15

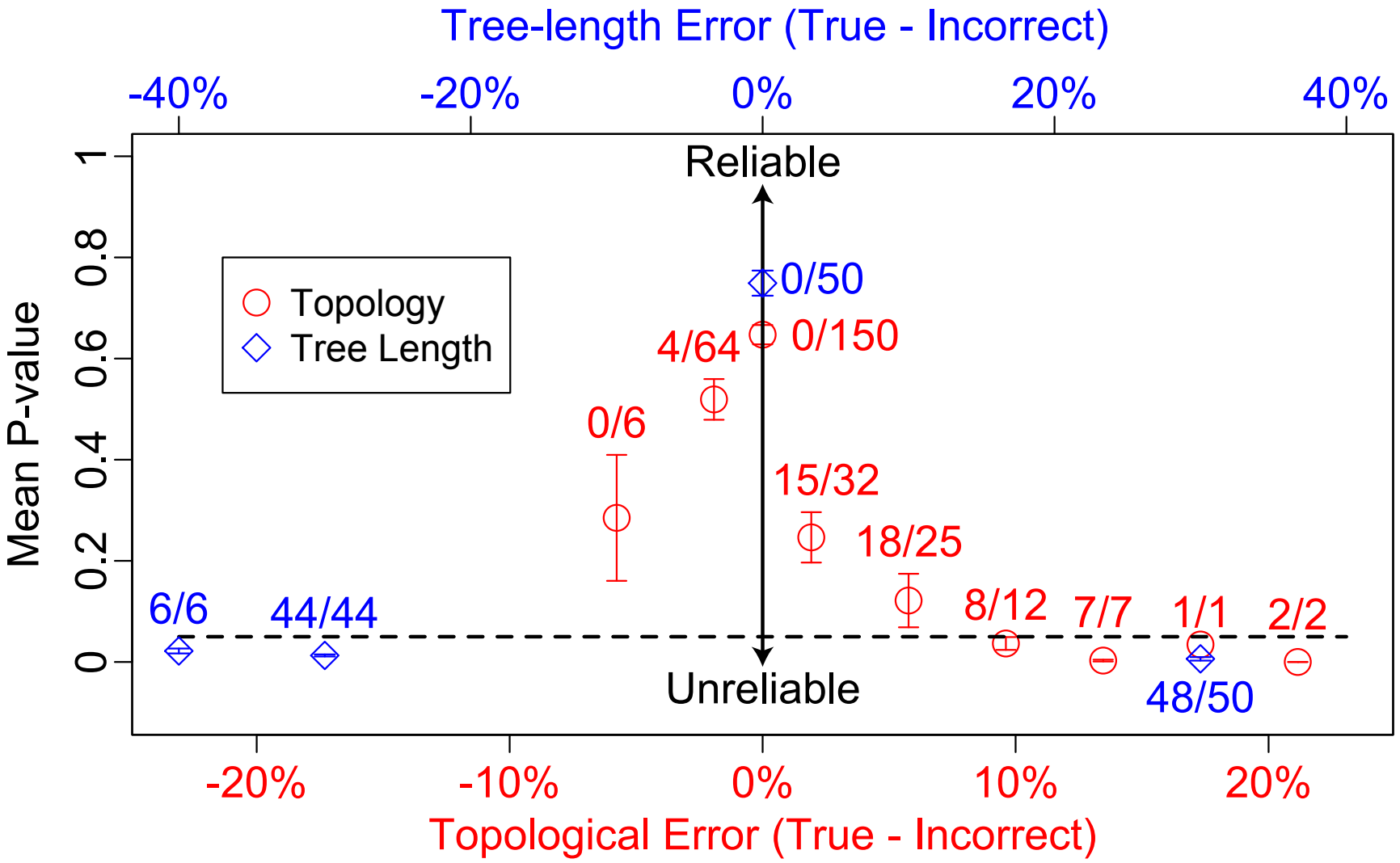
Variance in Tree Length = 2.30

Marginalizing across topologies

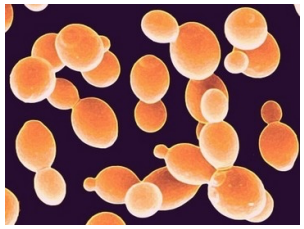
Motivating Results - Simulation



Motivating Results - Simulation



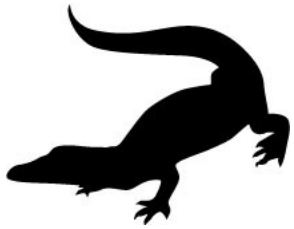
Motivating Results - Empirical



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%

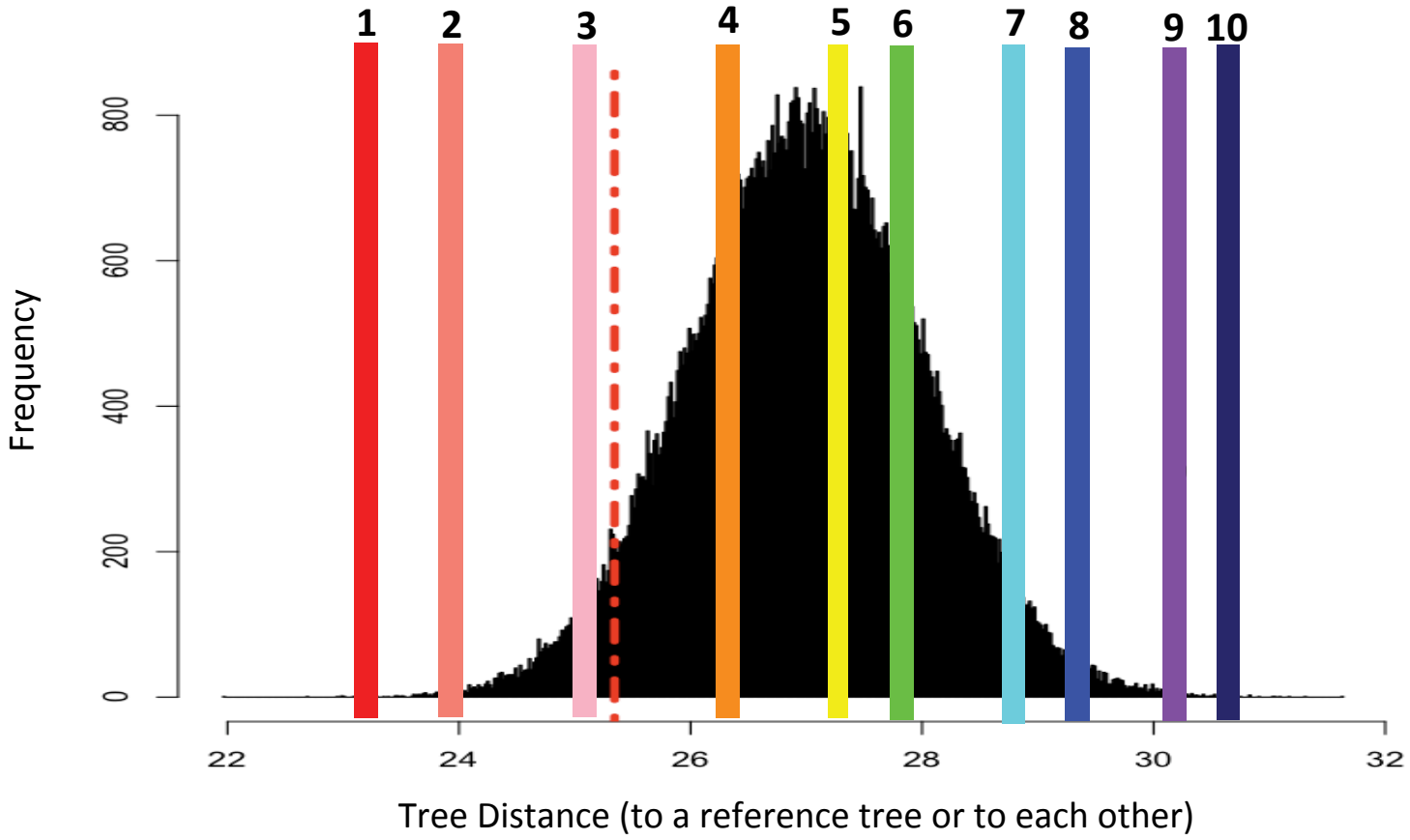
Doyle et al., 2015, Syst Biol

Motivating Results - Empirical

What might we expect from ideal filtering approaches?

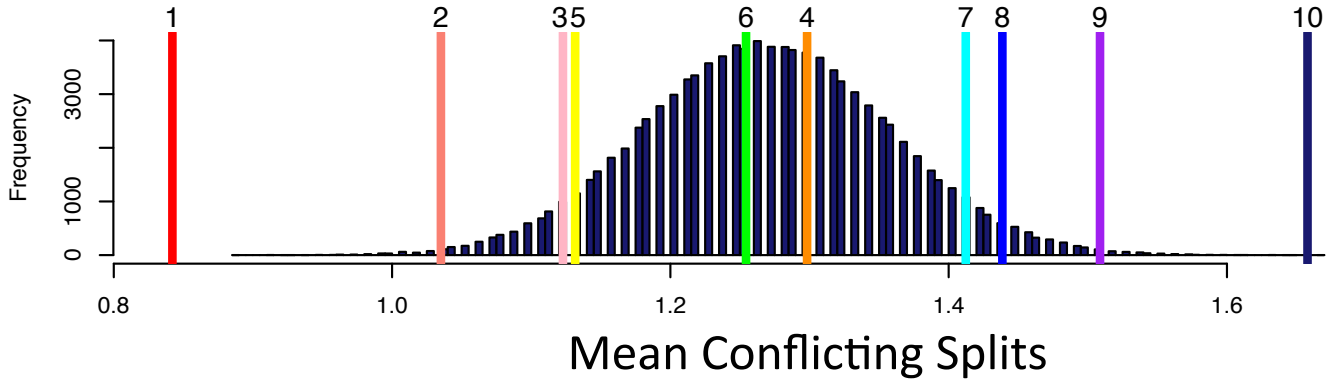
Perfect association between decile membership and tree distance

rho (r_s) = 1



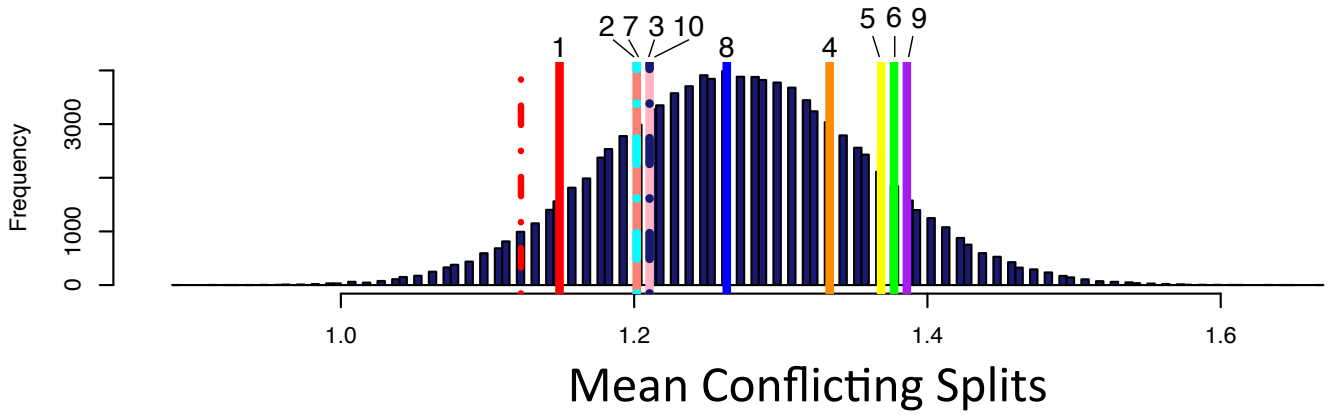
Doyle et al., 2015, Syst Biol

Motivating Results - Empirical



Posterior Predictive Filtering

r_s=0.964, P=2.2x10^{-16}

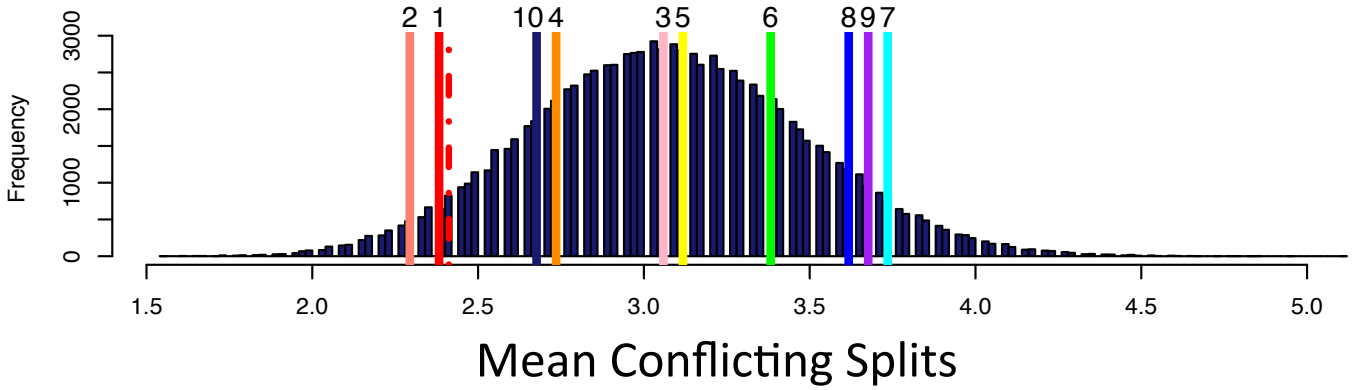


Rate Filtering

r_s=0.482, P=0.0793

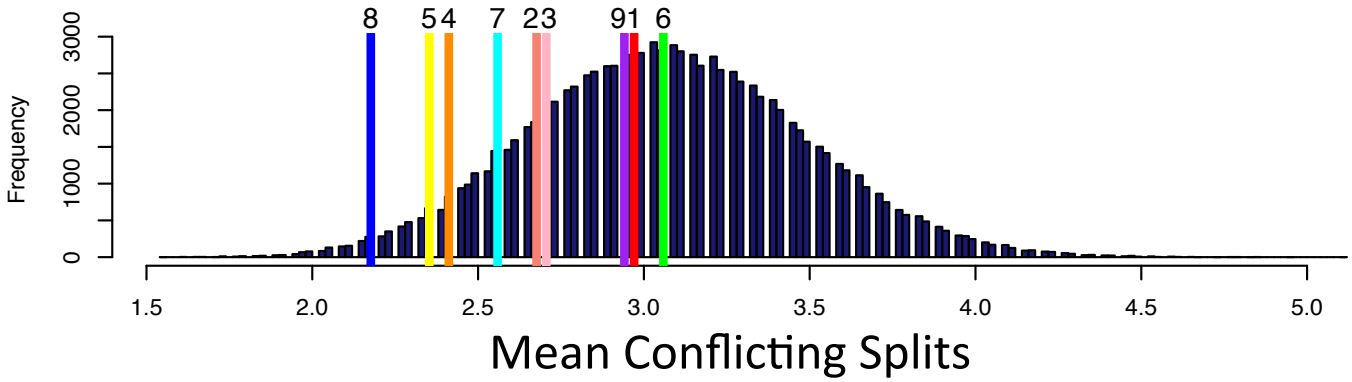
Doyle et al., 2015, Syst Biol

Motivating Results - Empirical



Posterior Predictive Filtering

r_s=0.600, P=0.03656



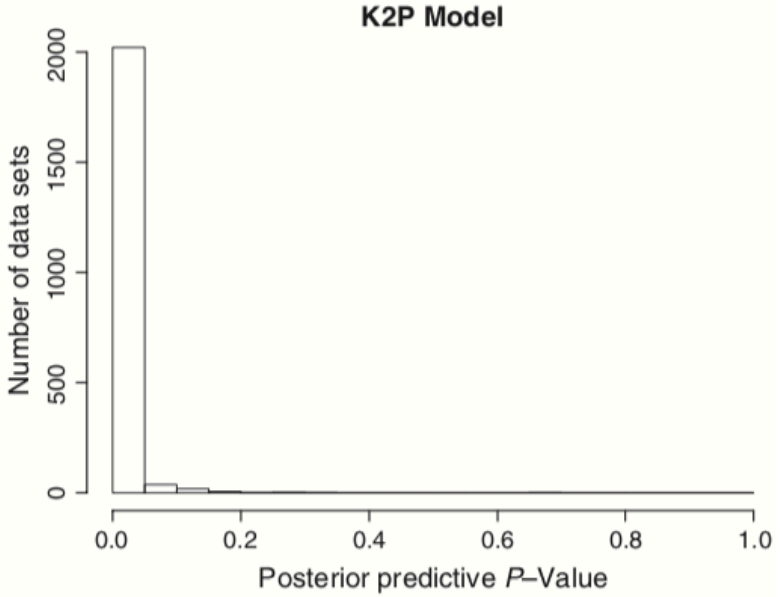
Rate Filtering

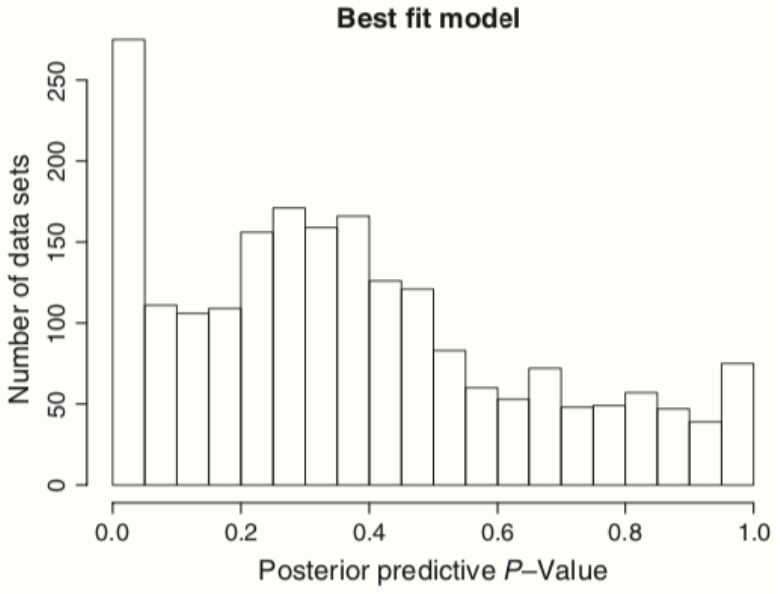
r_s=0.103, P=0.3925

Doyle et al., 2015, Syst Biol

Motivating Results - Barcodes

Barcode data
2,083 alignments
269,929 sequences
from ibol.org



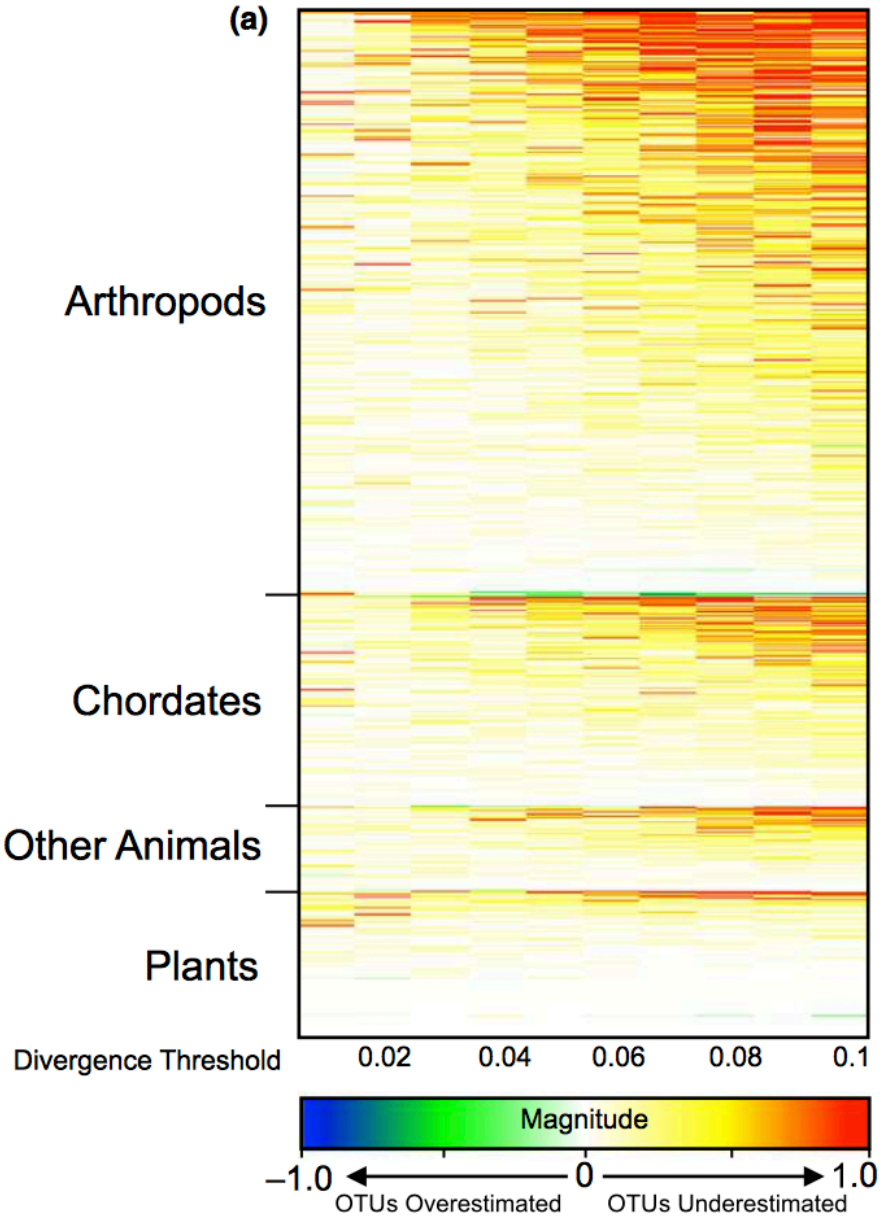


Barley & Thomson 2016 Mol Ecol

Motivating Results - Barcodes

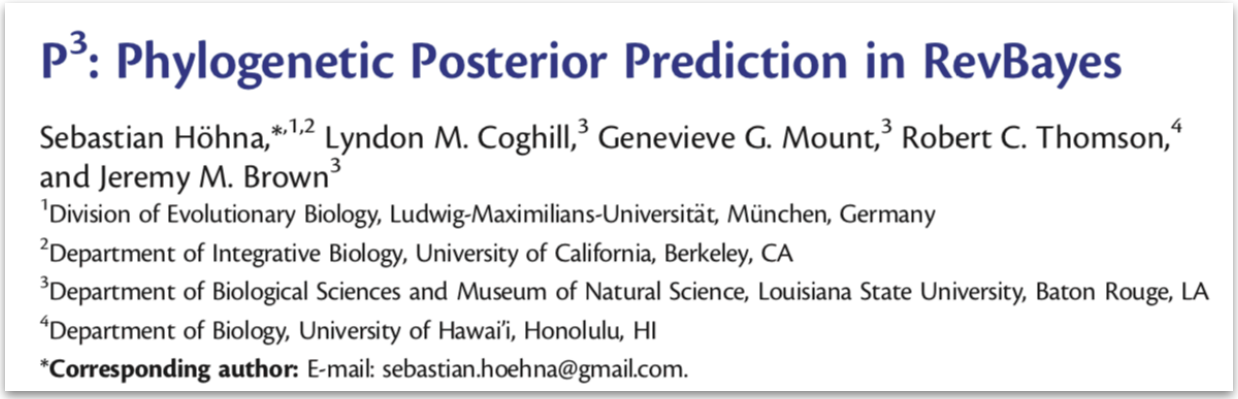
international BARCODE OF LIFE

Barcode data
2,083 alignments
269,929 sequences
from ibol.org



Barley & Thomson 2016 Mol Ecol

Active Development!



P^3: Phylogenetic Posterior Prediction in RevBayes

Sebastian Höhna,*[,1,2] Lyndon M. Coghill,[3] Genevieve G. Mount,[3] Robert C. Thomson,[4] and Jeremy M. Brown[3]

[1]Division of Evolutionary Biology, Ludwig-Maximilians-Universität, München, Germany

[2]Department of Integrative Biology, University of California, Berkeley, CA

[3]Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA

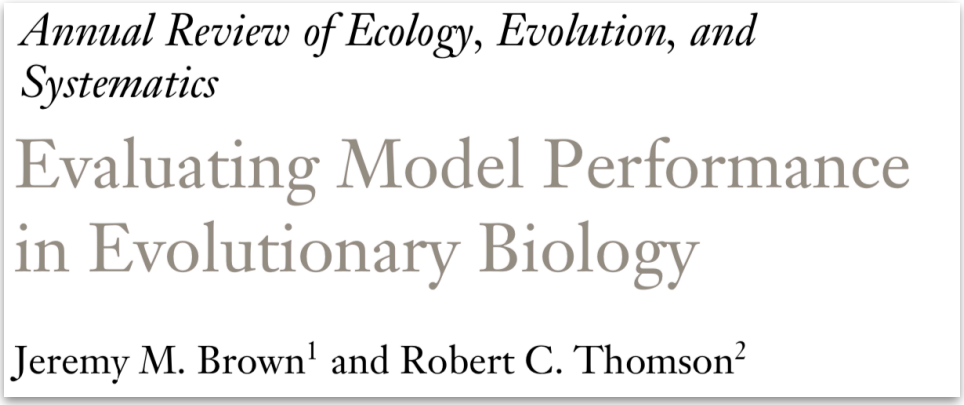
[4]Department of Biology, University of Hawai'i, Honolulu, HI

***Corresponding author:** E-mail: sebastian.hoehna@gmail.com.

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

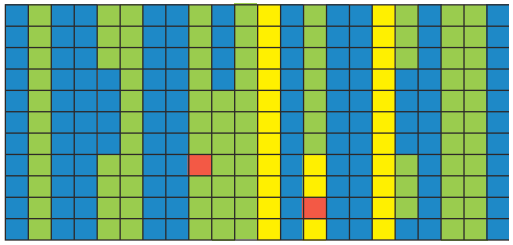


- Assessing model fit is probably most useful with big data
- Not meant to be a hypothesis test. We can **always** reject the fit of a model in a strict sense. All models are abstractions.
- Based on the aspects of our model that don't fit well, think about how to structure new models. Remember, with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

Tutorial

- Assessing Phylogenetic Reliability Using RevBayes and P^3
  - Data and Inference versions
  - Assess adequacy of JC and GTR on example data

Model Selection v Absolute Fit

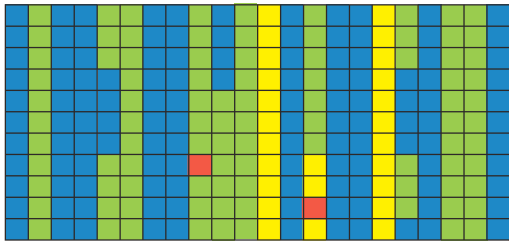


$M_1 \quad M_2$

$M_3 \quad M_4$

M_5

Model Selection v Absolute Fit

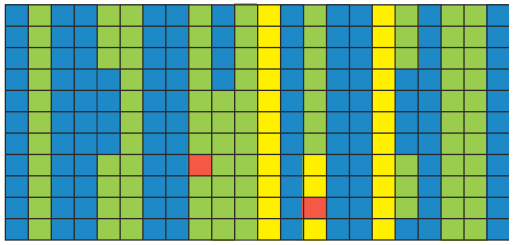


M_1 M_2

M_3 M_4

M_5

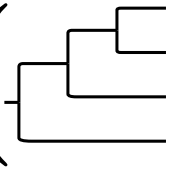
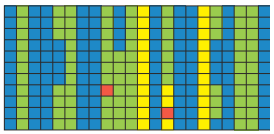
Model Selection v Absolute Fit



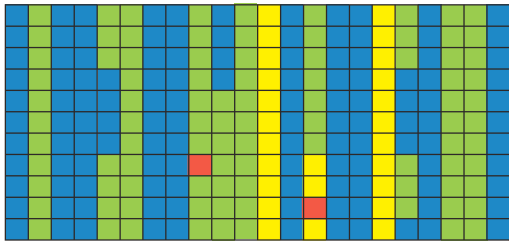
M_1 M_2

M_3 M_4

M_5

$$P(\text{  }, \theta | \text{  }, M_4)$$

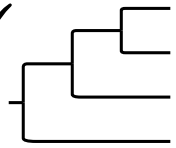
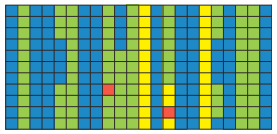
Model Selection v Absolute Fit



M_1 M_2

M_3 M_4

M_5

$$P\left(\text{  }, \theta \middle| \text{  }, M_4\right)$$

nature

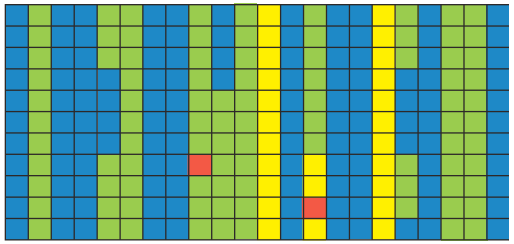
The guinea-pig is not a rodent

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 rodent[1]. 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.

Model Selection v Absolute Fit

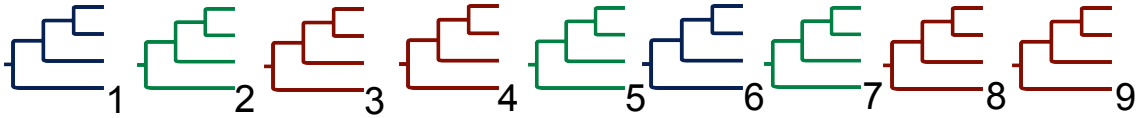


M_1 M_2

M_3 M_4

M_5

$$P(\text{tree}, \theta \mid \text{image}, M_4)$$



θ_1 θ_2 θ_3 θ_4 θ_5 θ_6 θ_7 θ_8 θ_9