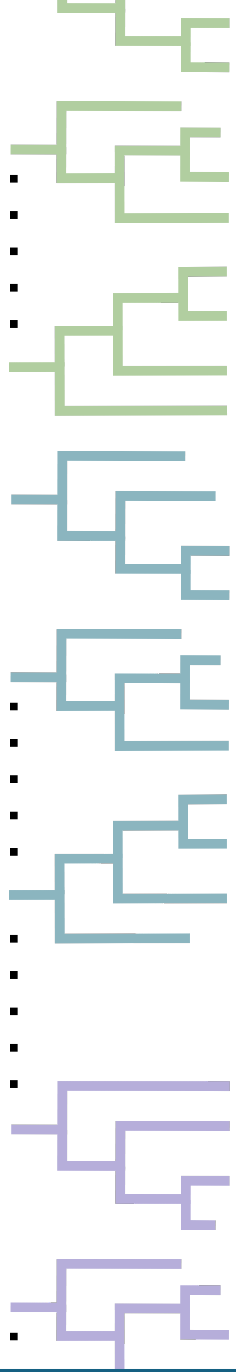
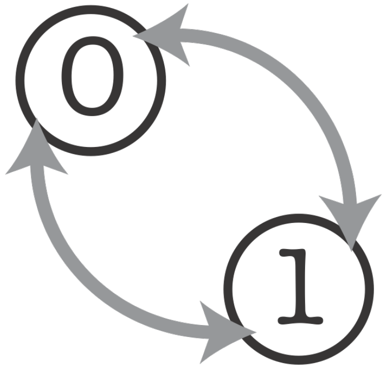
Morphological models and model choice

Laura Mulvey

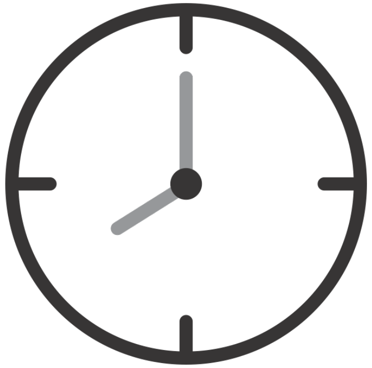


```
101510010?00-100--0000
000500010?200100--0010
102500010?200100--0?10
00?5?0010?200100?-0??1
0015000101201000430101
0015000101201010440111
??050????201000440?11
0015000101201000430101
0015000101201010440111
??050????201000440?11
```

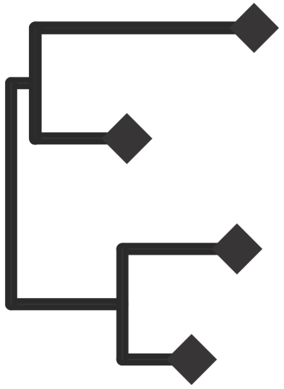
substitution model



clock model



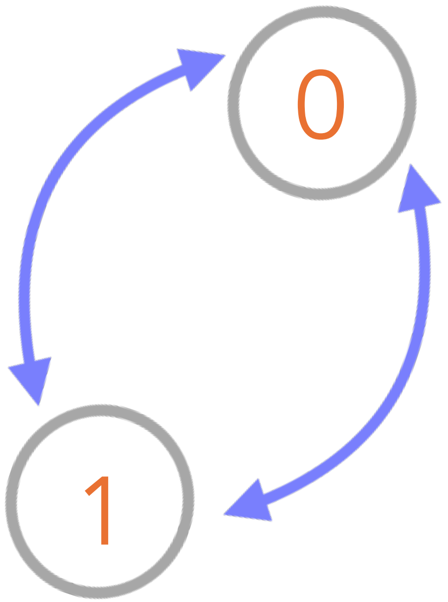
tree model



Warnock & Wright. 2020

Substitution models

A **substitution model** is a mathematical description of how characters change over evolutionary time:
- DNA
- RNA,
- Amino acids,
- Morphology



Mutation = when one state changes to another (e.g., A → G).

Substitution = when that mutation *sticks* in a population and becomes the new normal.
Substitution models describe the *probability* of these changes happening along the branches of a phylogenetic tree.

Rate Matrix

Every substitution model is defined by a **rate matrix (Q)**, which tells us how fast different changes happen

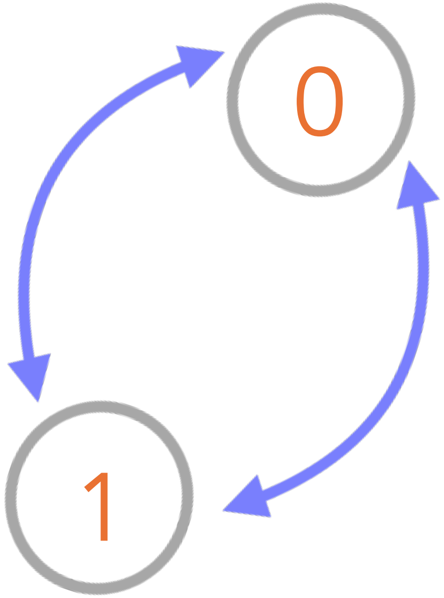
For DNA: 4 possible states (A, C, G, T).

For proteins: 20 states (amino acids).

For morphology: as many states as you code for a character.

Different substitution models are defined using the Q matrix

Rate Matrix



$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} \\ \mu_{10} & -\mu_1 \end{pmatrix}$$

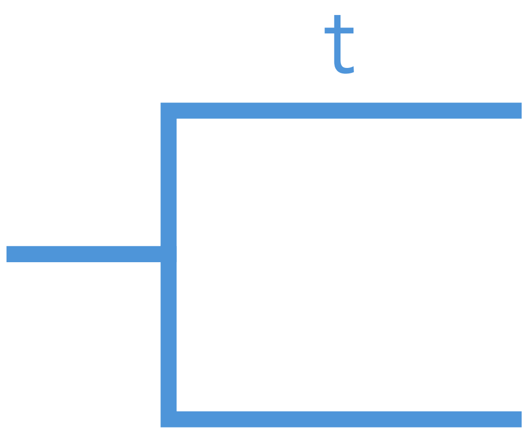
Must sum to zero

Any assumptions about your data can be incorporated through the mathematical expression

Rate Matrix

$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} \\ \mu_{10} & -\mu_1 \end{pmatrix}$$

This tell us the rate, how can we use this in an inference?



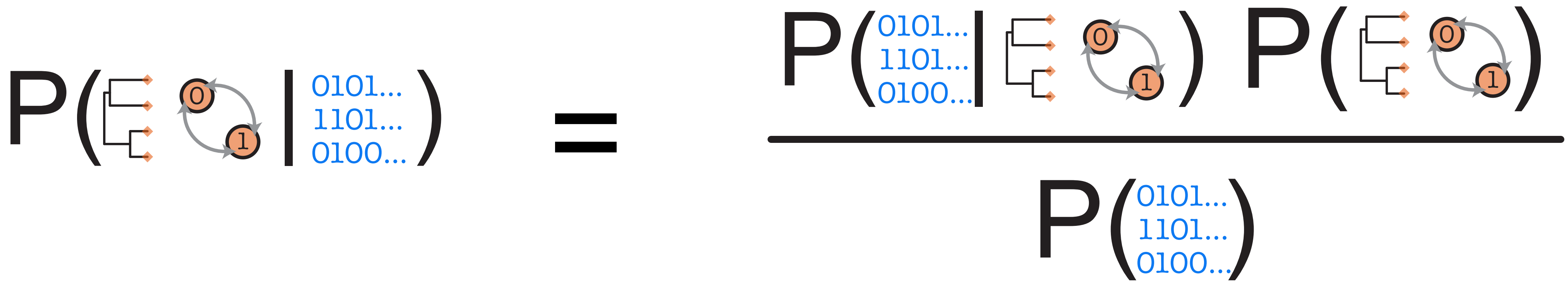
Matrix exponentiation

$$P(t) = e^{Qt}$$

This allows us to calculate the probability of a switch along a branch

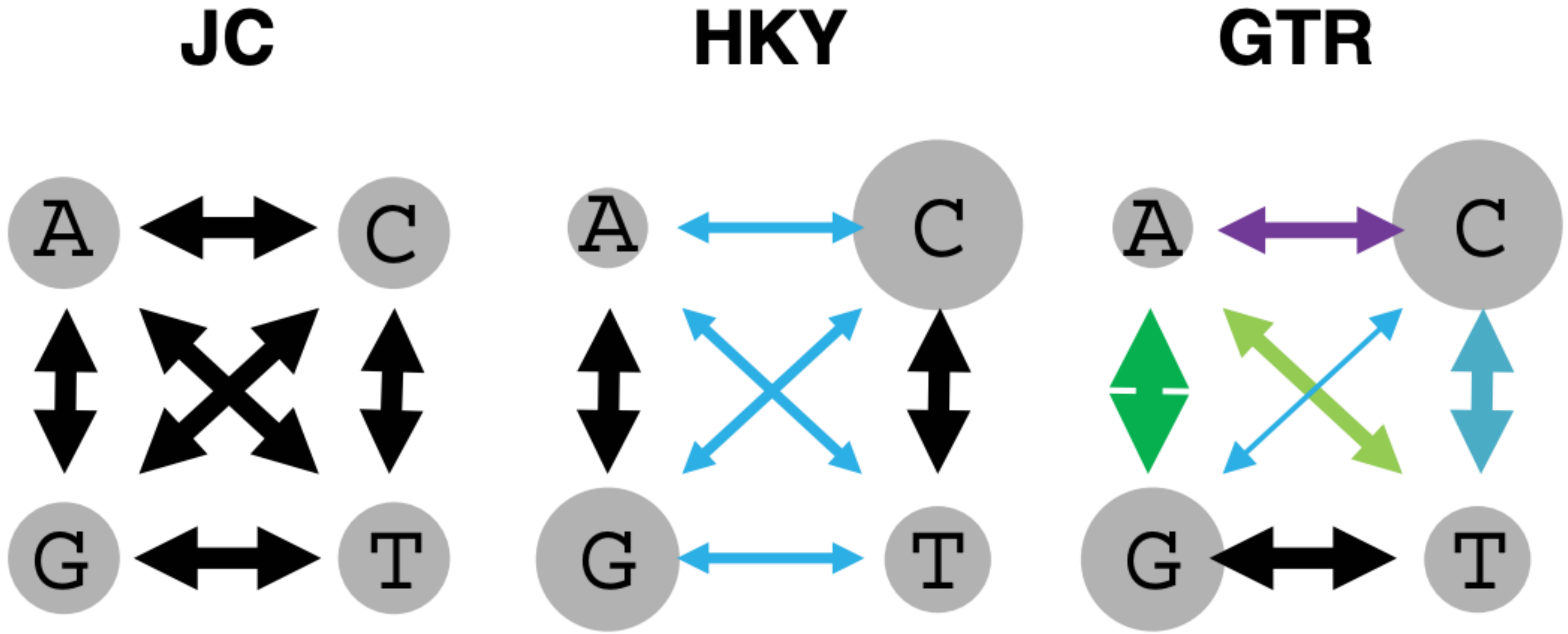
Bayes Theorem

The **rate matrix (Q)** is what gives us the **likelihood term**.
The data at the tips (DNA bases, amino acids, or morphology) have to be explained by the tree.



how probable it is to see your fossil and/or molecular data given a tree.

Substitution models in molecular data

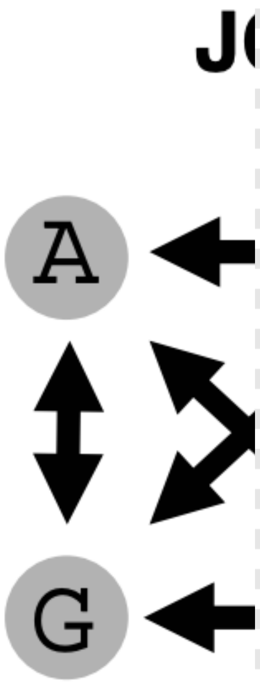


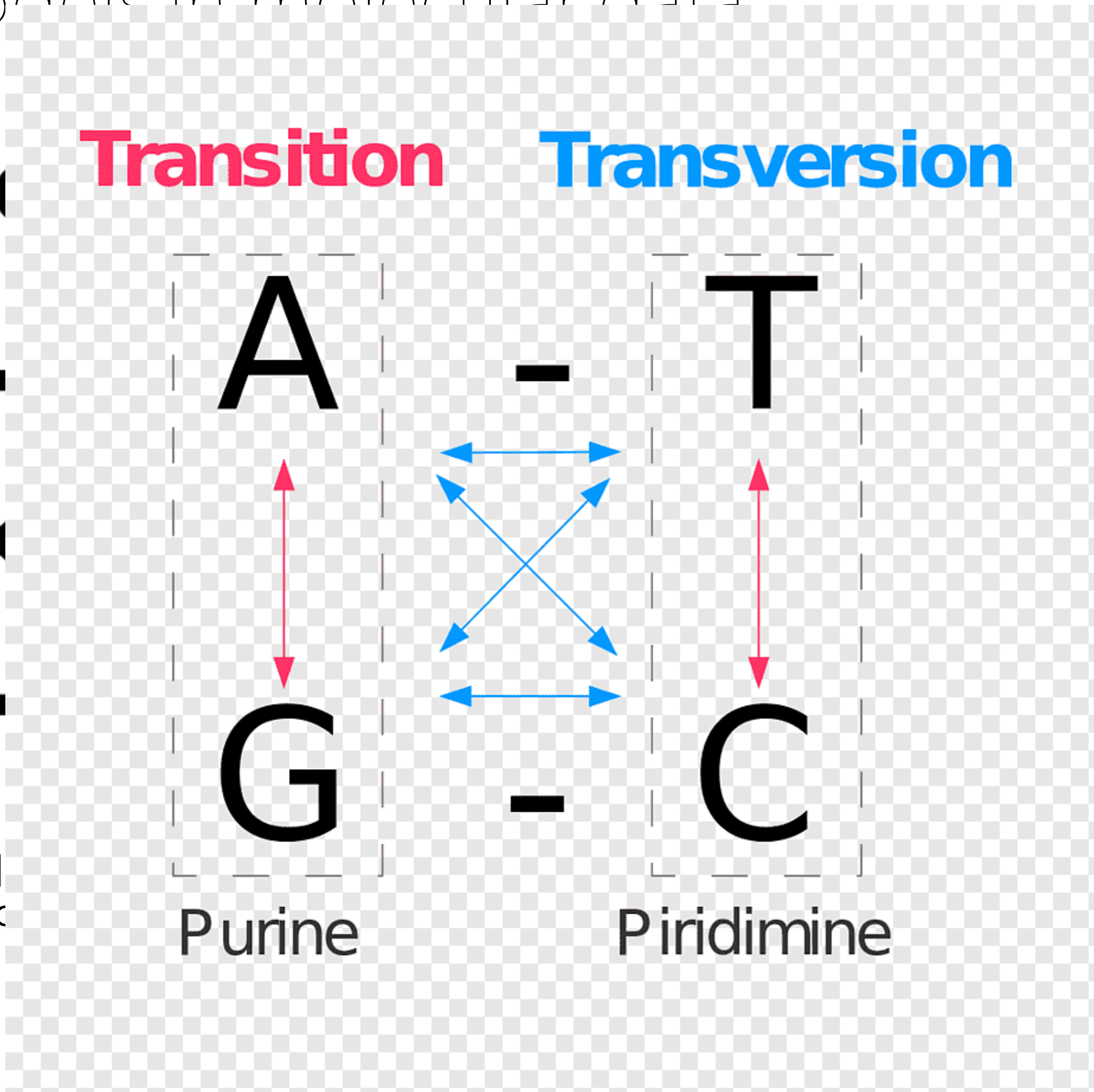
All changes equal
Equal base frequences

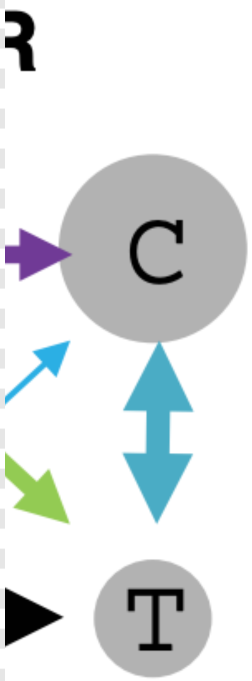
Transitions ≠ Transversions
Unequal base frequencies

Every substitution type has its own rate
Unequal base frequencies

Substitution models in molecular data



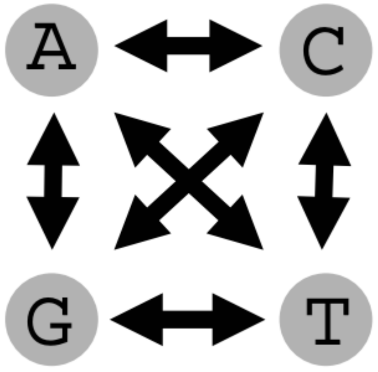




All changes eq
Equal base freq

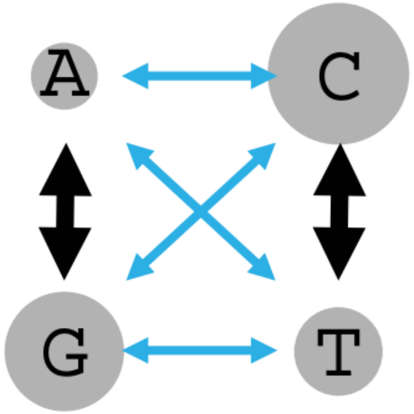
n type has its own rate
quencies

JC



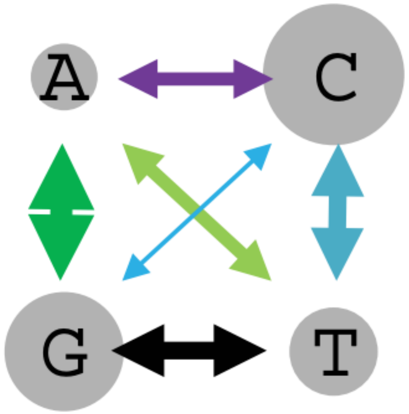
$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

HKY



$$Q = \begin{pmatrix} * & \kappa\pi_G & \pi_C & \pi_T \\ \kappa\pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \kappa\pi_T \\ \pi_A & \pi_G & \kappa\pi_C & * \end{pmatrix}$$

GTR

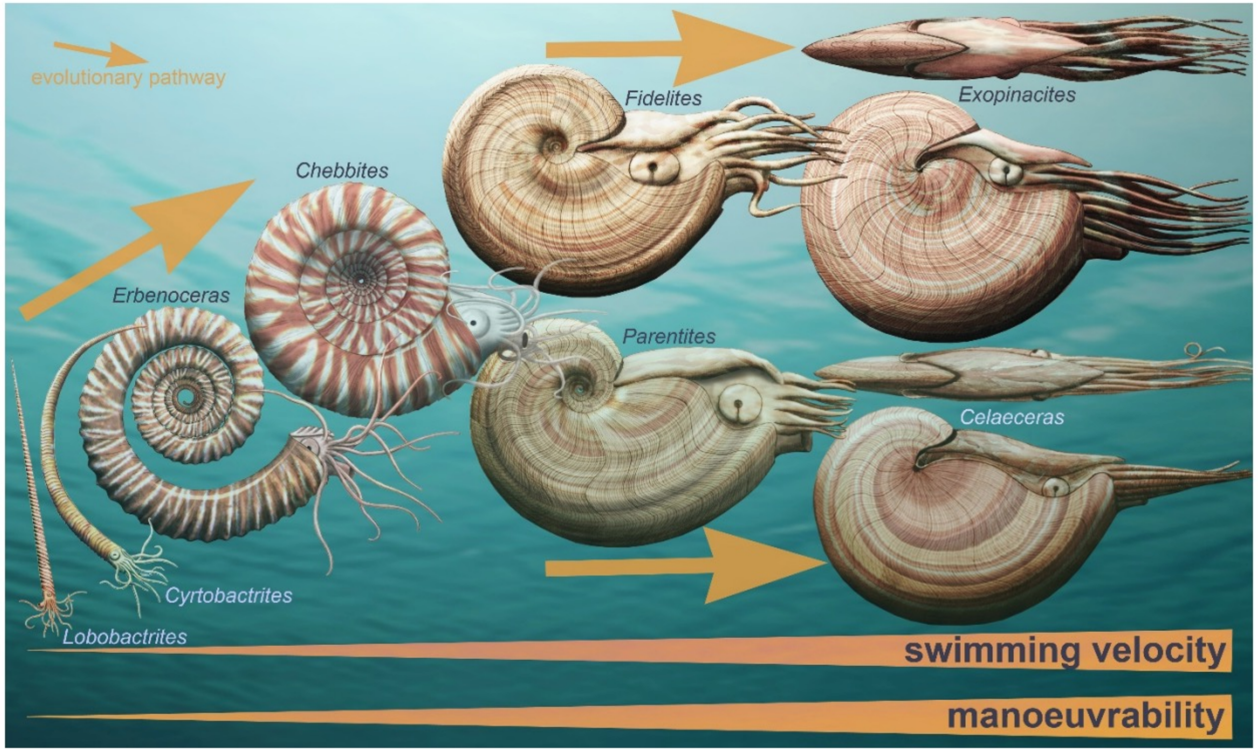


$$Q = \begin{pmatrix} * & \kappa_1\pi_G & \pi_C & \pi_T \\ \kappa_1\pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \kappa_2\pi_T \\ \pi_A & \pi_G & \kappa_2\pi_C & * \end{pmatrix}$$

Morphological data

Morphological data was the original type of information used in phylogenetic analysis

Fossils can be used to provide time calibrations, helps extant phylogeny, allows us to understand evolution through time



Monnet et al 2011 BMC Ecology and Evolution

Types of morphological data

Discrete Characters: Morphological data often consist of discrete characters, such as the presence or absence of certain traits, or more complex multistate traits (e.g., number of limbs, type of leaf, presence of a particular bone structure)

Continuous Characters: Some morphological data can be continuous, such as measurements of body size, length of bones, or other quantitative traits

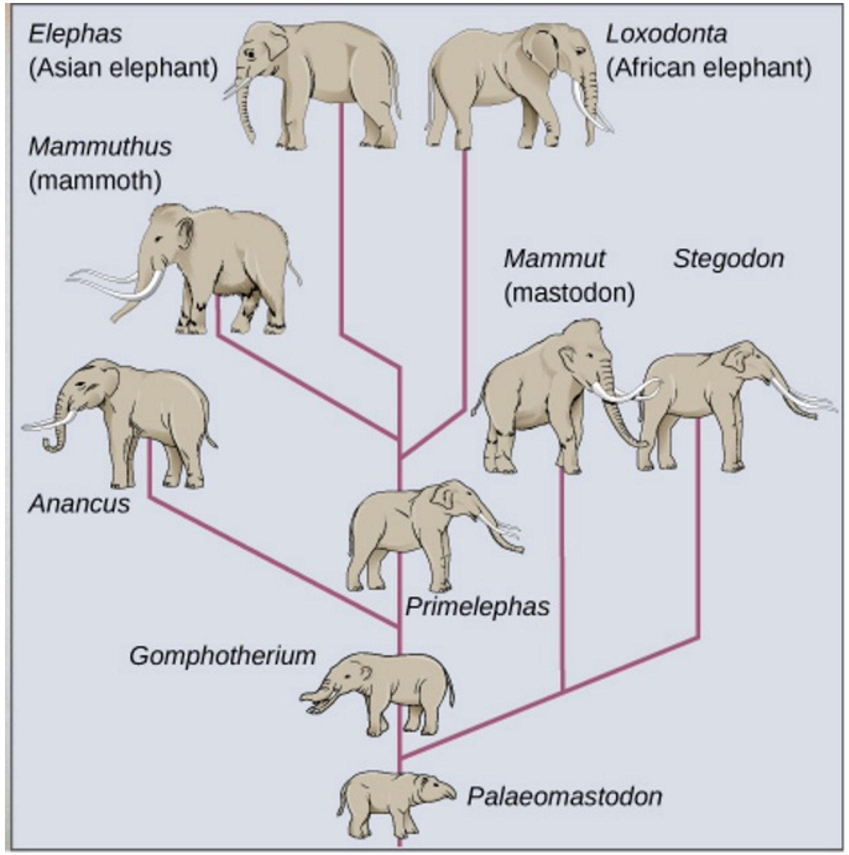


Image source

Types of morphological data

Discrete Characters: Morphological data often consist of discrete characters, such as the presence or absence of certain traits, or more complex multistate traits (e.g., number of limbs, type of leaf, presence of a particular bone structure)

Continuous Characters: Some morphological data can be continuous, such as measurements of body size, length of bones, or other quantitative traits

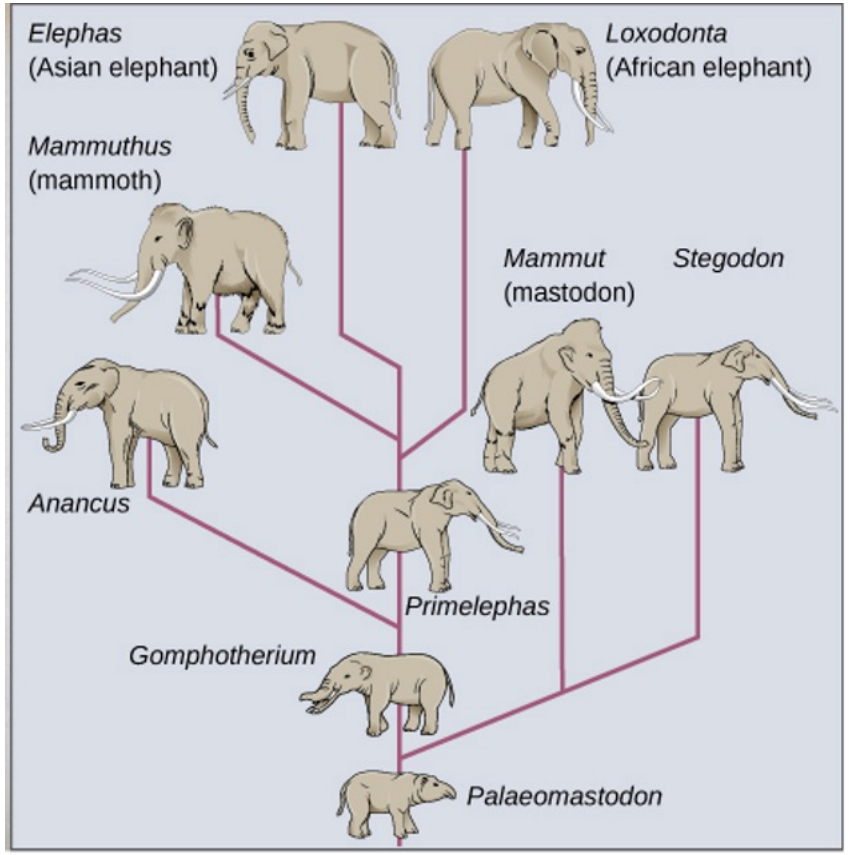
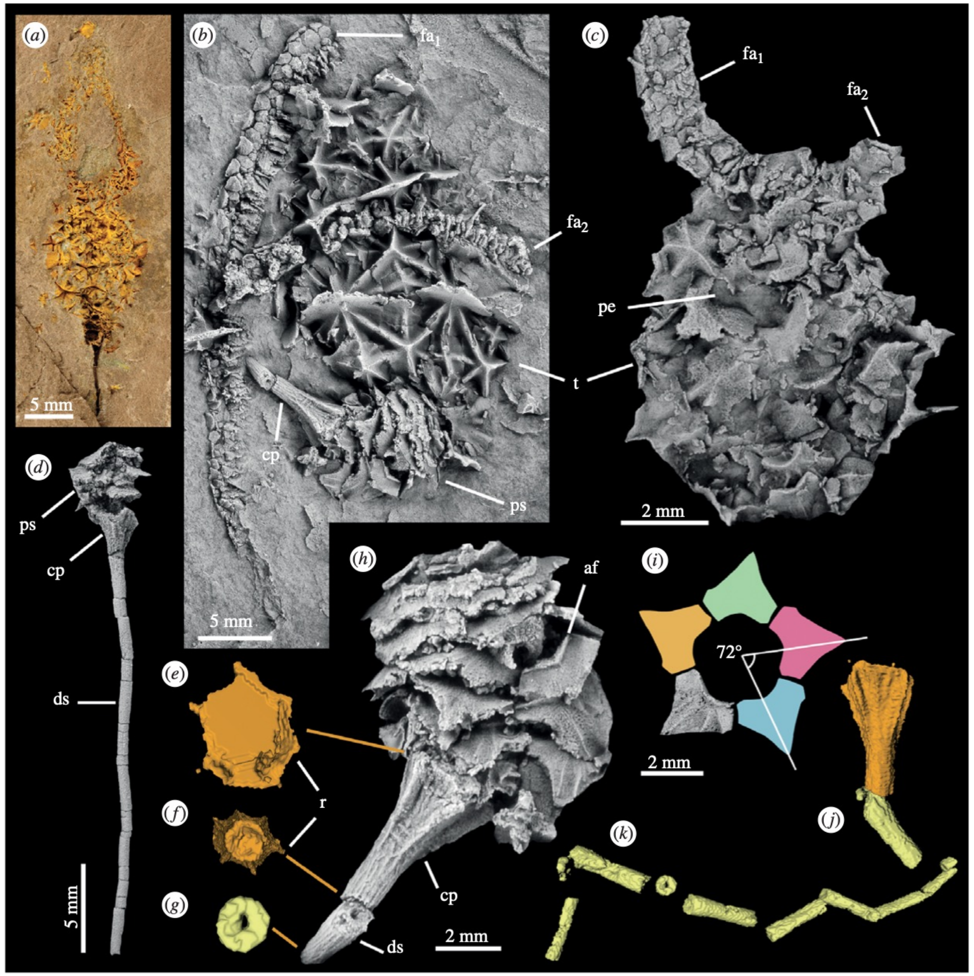


Image from https://www.zoologytalks.com/

	Trait 1 … Trait 28
Taxa 1	0015100010?00-100--000000000
	00050001 0?200100--001001000 0
	00250001 0?200100--0?1001000 0
	00?5?0010?200100?-0???010110
	0015000101201000430100011111
	0015000101201010440111011111
	??050?????201000440?11011111
	01050?010-210000?501??010110
	0002000100210100 3-1110010110
	0002000100211001441121011111
	00020111 1-210010?-??11011121
	?103?0?11?1001104-0000010000
	1005002110100010--0?00110?20
Taxa 14	1005002000101010540?00110020



Cambrian stalked echinoderms show unexpected plasticity of arm construction
Zamora & Smith. 2012 Proc B

Making assumptions about evolution from this data is extremely difficult

Multistate characters

Non-applicable characters

Presence/ absence

```
101510010?00-100--0000000000
000500010?200100--0010010000
102500010?200100--0?10010000
00?5?0010?200100?-0???010110
0015000101201000430100011111
0015000101201010440111011111
??050?????201000440?11011111
```

Often small

No consistent biological meaning

High proportion of missing data

Discrete character data

Binary traits	0 1	Often describes the presence/absence of a trait

Discrete character data

Binary traits	0 1	Often describes the presence/absence of a trait
Multistate traits	0 1 2 3 4 …..	Used to describe more complex traits and can capture greater variation between taxa

Discrete character data

Binary traits	0 1	Often describes the presence/absence of a trait
Multistate traits	0 1 2 3 4 …..	Used to describe more complex traits and can capture greater variation between taxa
Missing characters	?	Used when the specimen is either too decayed to determine whether it has a certain character trait or not, or we are missing the relevant part of the body

Discrete character data

Binary traits	0 1	Often describes the presence/absence of a trait
Multistate traits	0 1 2 3 4 …..	Used to describe more complex traits and can capture greater variation between taxa
Missing characters	?	Used when the specimen is either too decayed to determine whether it has a certain character trait or not, or we are missing the relevant part of the body
Non-applicable	-	Used when the trait is not associated with a taxon. They represent a type of nested coding where the presence of the trait is defined in a different trait

Discrete character data

Binary traits	0 1	Often describes the presence/absence of a trait
Multistate traits	0 1 2 3 4 …..	Used to describe more complex traits and can capture greater variation between taxa
Missing characters	?	Used when the specimen is either too decayed to determine whether it has a certain character trait or not, or we are missing the relevant part of the body
Non-applicable	-	Used when the trait is not associated with a taxon. They represent a type of nested coding where the presence of the trait is defined in a different trait
Polymorphisms	0/1/2	Used when there are variations in a traits within species

Discrete character data

Binary traits	0 1	Often describes the presence/absence of a trait
Multistate traits	0 1 2 3 4 …..	Used to describe more complex traits and can capture greater variation between taxa
Missing characters	?	Used when the specimen is either too decayed to determine whether it has a certain character trait or not, or we are missing the relevant part of the body
Non-applicable	-	Used when the trait is not associated with a taxon. They represent a type of nested coding where the presence of the trait is defined in a different trait
Polymorphisms	0/1/2	Used when there are variations in a traits within species
Uncertain	0/1/2	Used when it is not clear which character trait is present in the taxon

How do we model morphological evolution?

Mk model

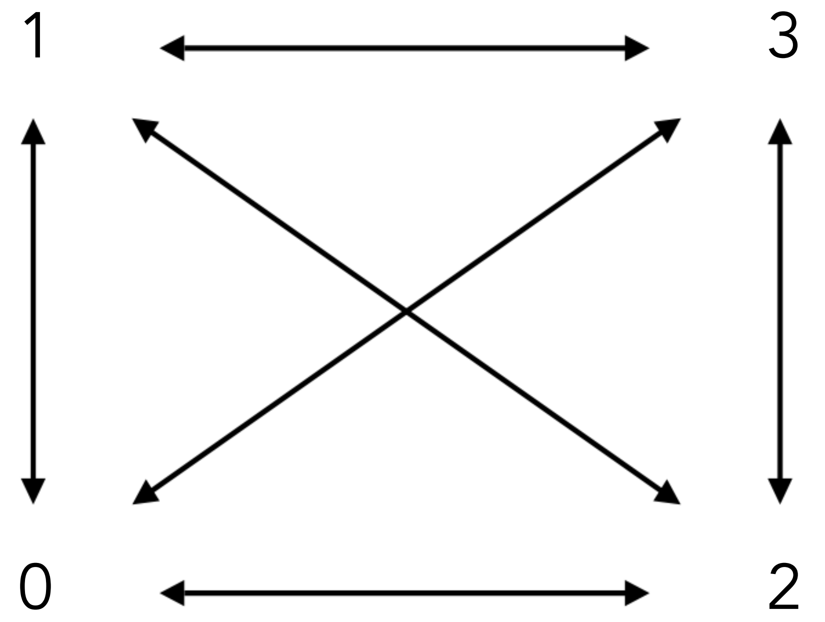
$$0 \longleftrightarrow 1$$

Assumes **equal transition probabilities** between states

$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} \\ \mu_{10} & -\mu_1 \end{pmatrix}$$

Mk model

K can be any number of states



$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} & \mu_{02} & \mu_{03} \\ \mu_{10} & -\mu_1 & \mu_{12} & \mu_{13} \\ \mu_{20} & \mu_{21} & -\mu_2 & \mu_{23} \\ \mu_{30} & \mu_{31} & \mu_{32} & -\mu_3 \end{pmatrix},$$

*4 state here as an example, can be any number from 2!

MkV model

What is one characteristic of morphological data that is extremely different to molecular

```
0015100010?00-100--0000000000
```

```
001510010?00-100--0000000000
000500010?200100--0010010000
002500010?200100--0?10010000
00?5?0010?200100?-0???010110
0015000101201000430100011111
0015000101201010440111011111
??050?????201000440?11011111
01050?010-210000?501??010110
0002000100210100З-1110010110
0002000100211001441121011111
00020111 1-210010?-??11011121
?103?0?11?1001104-0000010000
1005002110100010--0?00110?20
1005002000101010540?00110020
```

MkV model

What is one characteristic of morphological data that is extremely different to molecular

```
001510010?00-100--0000000000
000500010?200100--0010010000
002500010?200100--0?10010000
00?5?0010?200100?-0???010110
0015000101201000430100011111
0015000101201010440111011111
??050?????201000440?11011111
01050?010-210000?501??010110
000200010021010003-111001011 0
```

All varying characters

MkV model

Corrects for **ascertainment bias**

Failing to account for this can lead to **overestimations in branch lengths** and which can further lead to errors in topology!

Condition the likelihood on there only being varying site

$$Pr(D \mid V) = \frac{Pr(D, V)}{Pr(V)}$$

MkV model

	True Branch Length	Mk	Mkv
Percent correct	-	74.0	99.8
Branch A	0.2	241,750 (±349,100)	0.206 (±0.060)
Branch B	0.05	0.43210 (±0.13756)	0.050 (± 0.018)
Branch X	0.05	54.646 (±1,725.3)	0.052 (± 0.023)
Branch C	0.2	143,950 (±228,910)	0.206 (± 0.059)
Branch D	0.05	0.022 (± 0.054)	0.051 (±0.019)

Among character rate variation

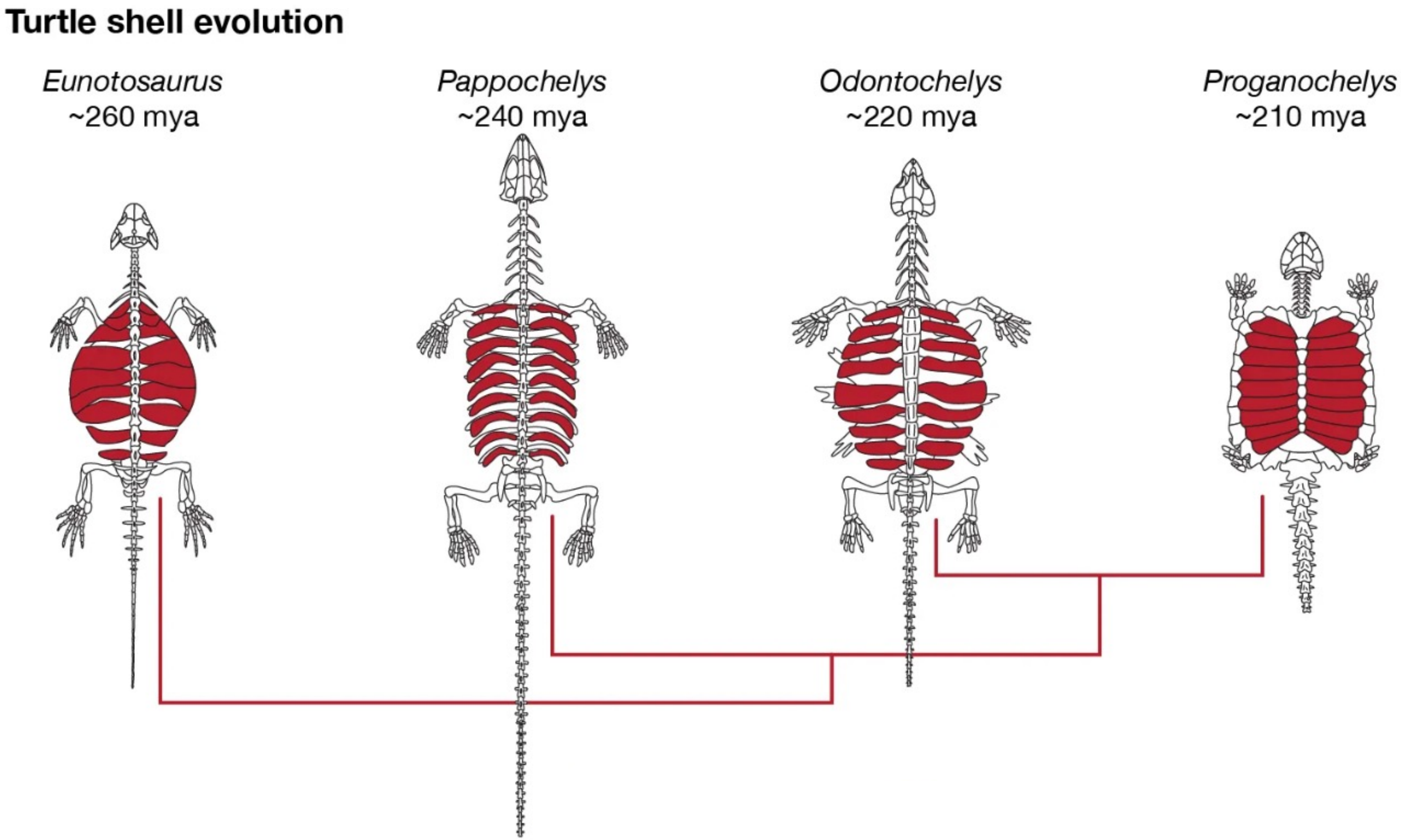


Image source

Among character rate variation

	T1	T2
Taxa A	0	0
Taxa B	0	1
Taxa C	1	2

The transition rate will impact branch lengths

Slow rate of evolution

Fast rate of evolution

Relative to each other!

Among character rate variation

What do we do?

	T1	T2
Taxa A	0	0
Taxa B	0	1
Taxa C	1	2

Allow these traits to evolve at different rates:

- Specify which traits evolve fast
- Use a gamma model to account for rate heterogeneity

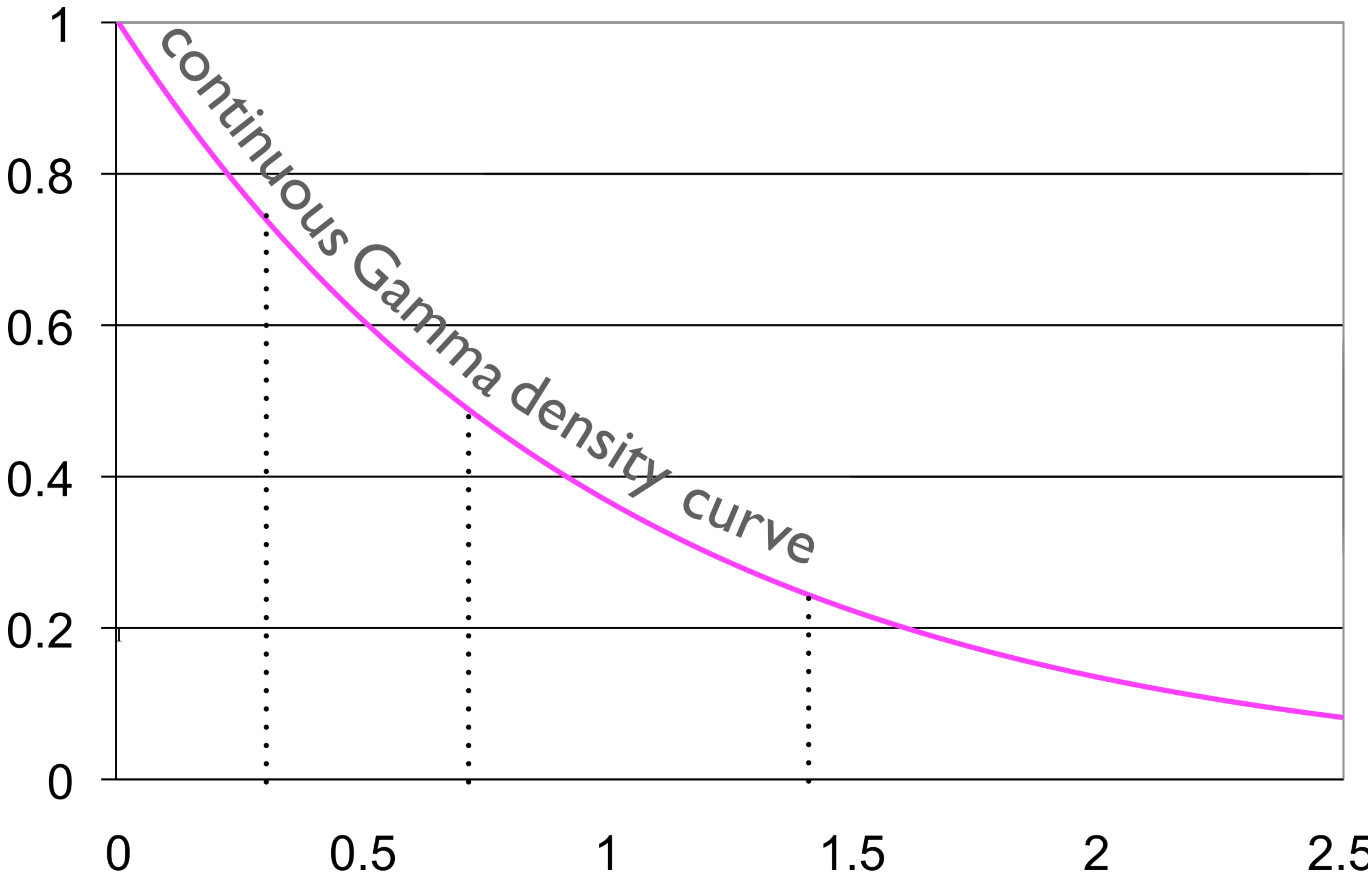
Among character rate variation

What do we do?

	T1	T2
Taxa A	0	0
Taxa B	0	1
Taxa C	1	2

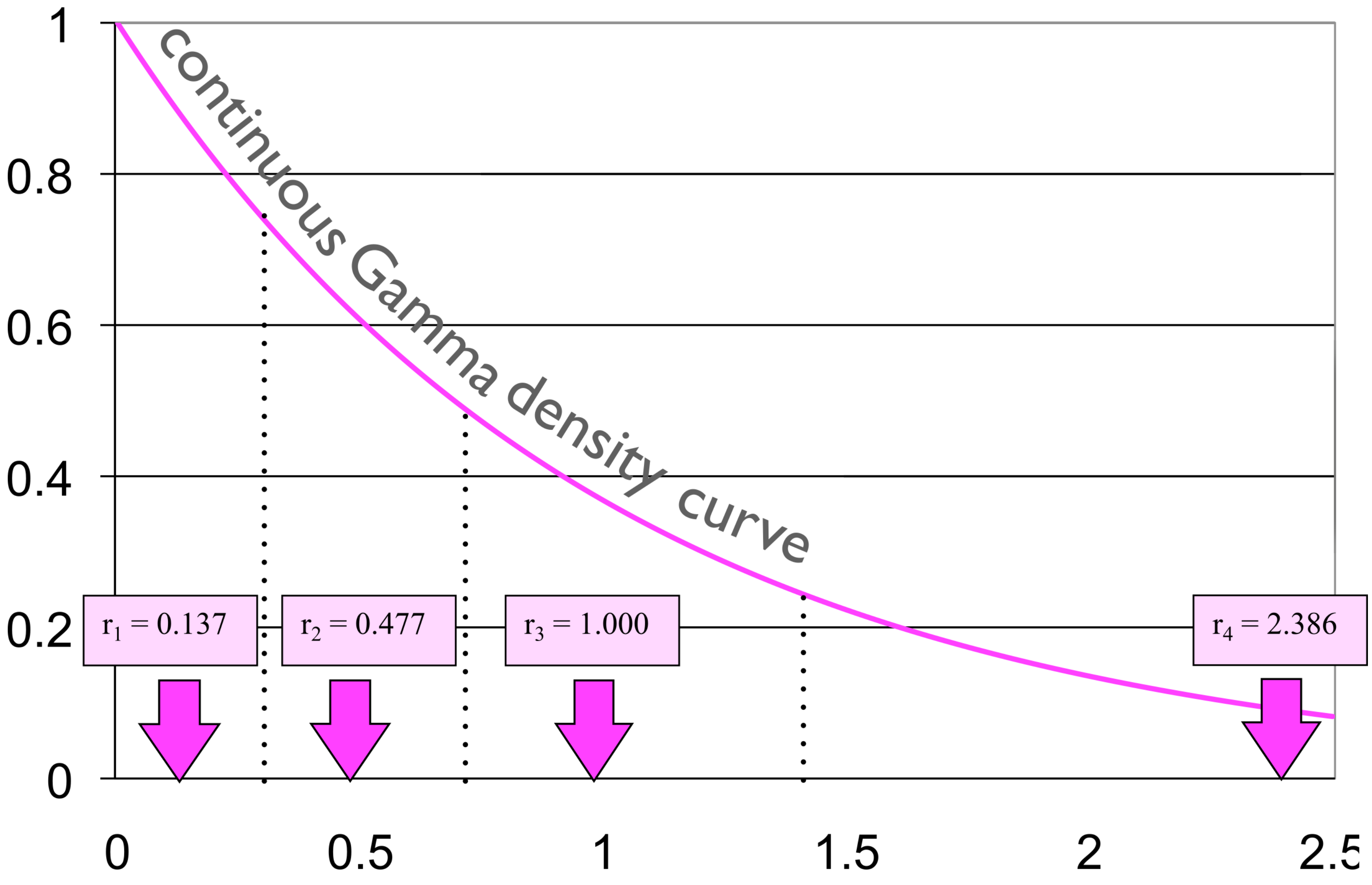
Allow these traits to evolve at different rates:
- Specify which traits evolve fast
- **Use a gamma model to account for rate heterogeneity**

Among character rate variation



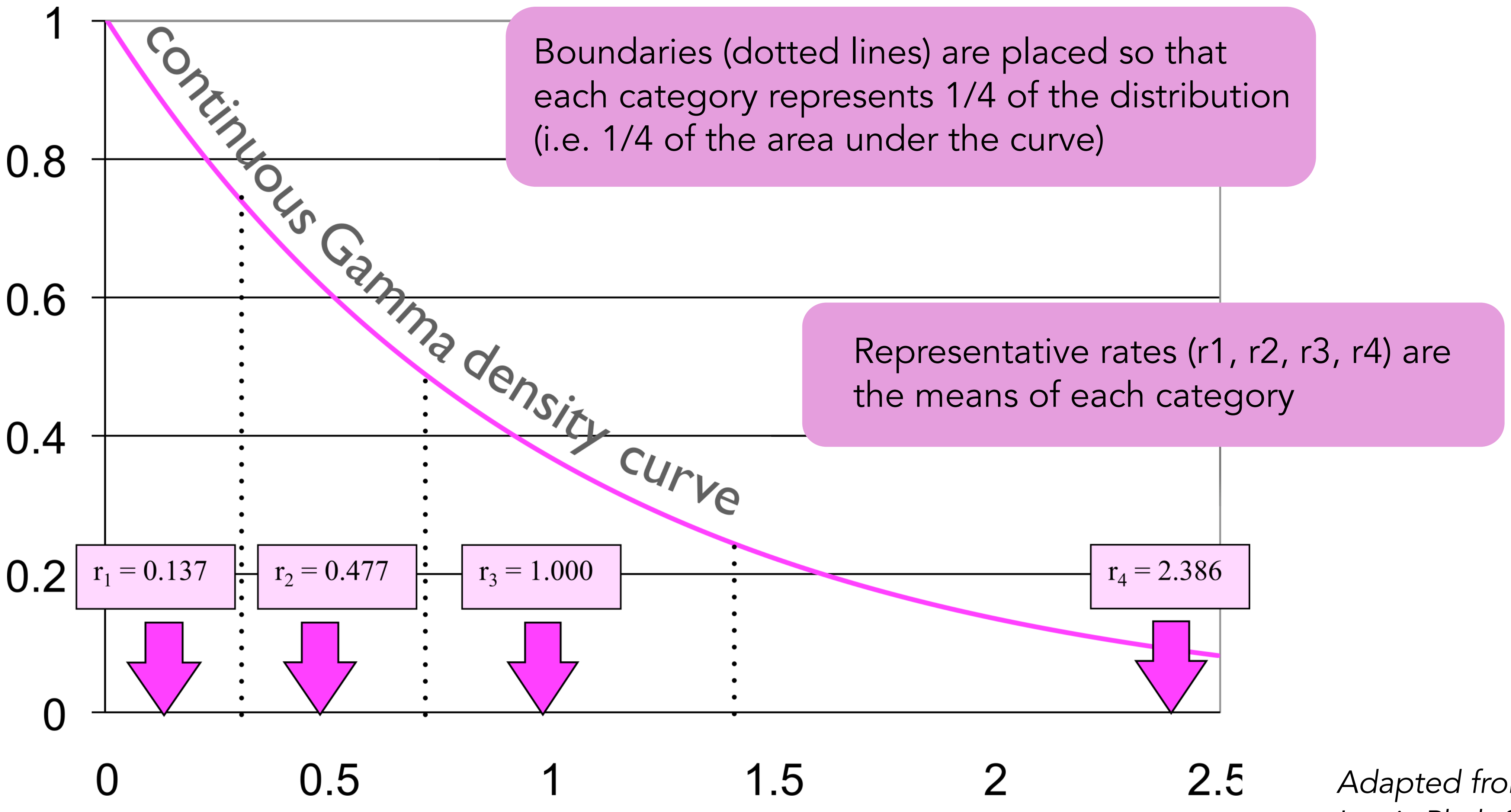
Adapted from Paul Lewis PhyloSeminar

Among character rate variation



Adapted from Paul Lewis PhyloSeminar

Among character rate variation



Adapted from Paul Lewis PhyloSeminar

Among character rate variation

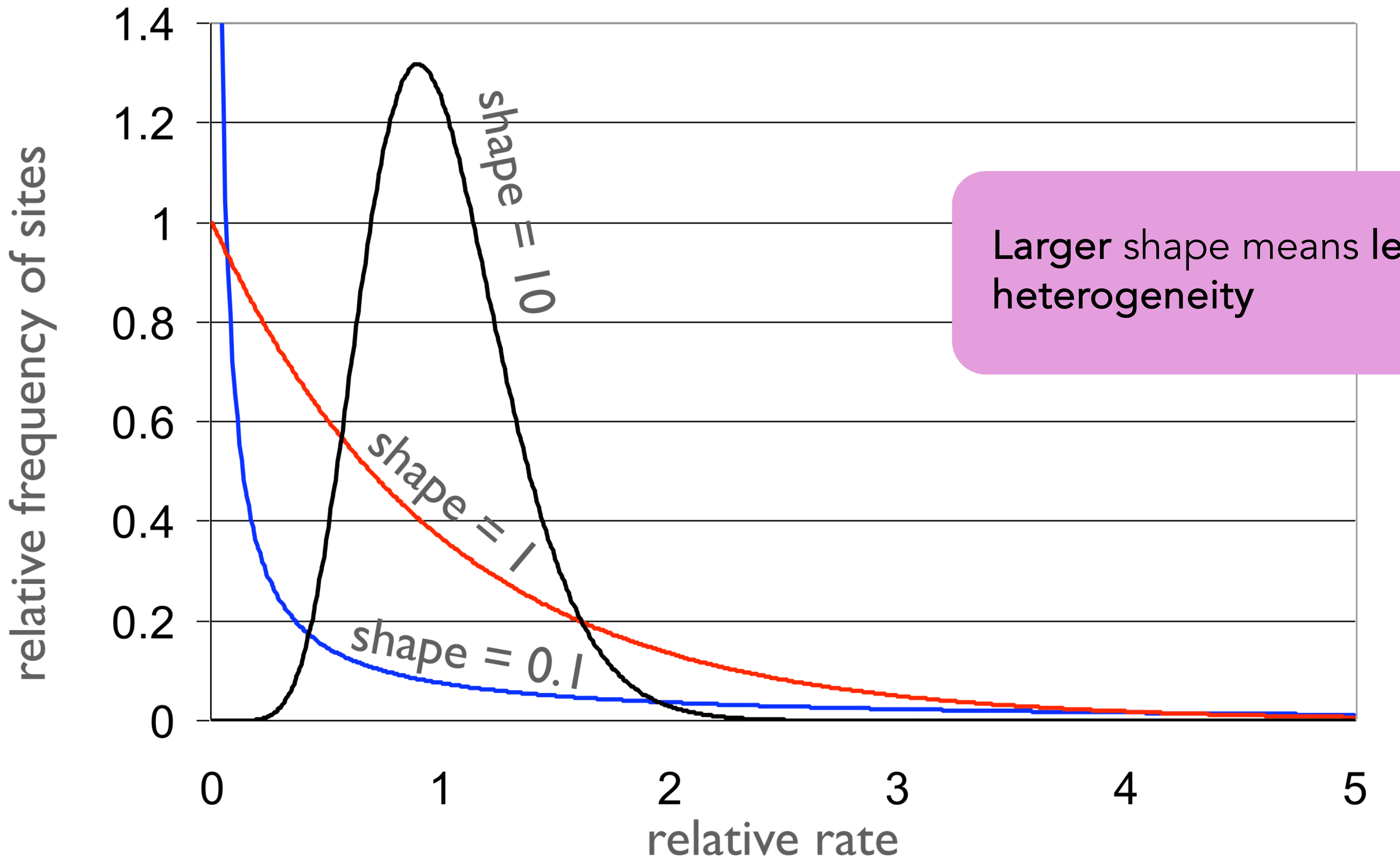
What do we do?

	T1	T2	
Taxa A	0	0	
Taxa B	0	1	
Taxa C	1	2	Faster R (R4)

Slower R (R1,2)

Allow each trait to evolve according to the rates drawn from the gamma distribution
One rate will fit the best and be the most influential for the likelihood calculation

Among character rate variation



Larger shape means **less heterogeneity**

Adapted from Paul Lewis PhyloSeminar

Incorporating ACRV into the matrix

$$\mathbf{Q} = \begin{pmatrix} -\mu_0 & \mu_{01} & \mu_{02} & \mu_{03} \\ \mu_{10} & -\mu_1 & \mu_{12} & \mu_{13} \\ \mu_{20} & \mu_{21} & -\mu_2 & \mu_{23} \\ \mu_{30} & \mu_{31} & \mu_{32} & -\mu_3 \end{pmatrix} \ast_{,} \; r$$

Compute the likelihood for each site

$L_1 = P(\text{site} \mid Q_1)$
$L_2 = P(\text{site} \mid Q_2)$
$L_3 = P(\text{site} \mid Q_3)$
$L_4 = P(\text{site} \mid Q_4)$

For each category:

$Q_1 = r_1 \cdot Q,$
$Q_2 = r_2 \cdot Q,$
$Q_3 = r_3 \cdot Q,$
$Q_4 = r_4 . Q$

Average over categories

$L_{site} = 1/4 \, (L_1 + L_2 + L_3 + L_4)$

This gives the **final likelihood for that site**, accounting for among-site rate variation.

Partitioning data

Grouping together parts of the alignment that have similar characteristics and or may have **evolved together** due to evolutionary pressures

$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} & \mu_{02} & \mu_{03} \\ \mu_{10} & -\mu_1 & \mu_{12} & \mu_{13} \\ \mu_{20} & \mu_{21} & -\mu_2 & \mu_{23} \\ \mu_{30} & \mu_{31} & \mu_{32} & -\mu_3 \end{pmatrix},$$

The **defaults** in many phylogenetic software is to group by maximum observed state size

$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} \\ \mu_{10} & -\mu_1 \end{pmatrix}$$

Partitioning data

When should we partition our data?

Partitioning data

When should we partition our data?

If we have presence (1) absence (0) traits partitioning will always be a logical approach: what would transitioning to state 2 in this scenario even mean?

Partitioning data

When should we partition our data?

If we have presence (1) absence (0) traits partitioning will always be a logical approach: what would transitioning to state 2 in this scenario even mean?

We should be cautious for traits describing a trait – just because we do not observe a state 2 can we be absolutely certain there never was one?

Justifying partitioning schemes is very important as they have a major impact on inference results

Other morphological models

Ordered characters

Ordered characters can be placed in an order so that transitions only occur between adjacent states.

For example, "**intermediate**" species that are somewhere in between limbed and limbless – for example, the "mermaid skinks" (Sirenoscincus) from Madagascar, so called because they lack hind limbs. An ordered model might only allow transitions between limbless and intermediate, and intermediate and limbed; it would be impossible under such a model to go directly from limbed to limbless without first becoming intermediate.

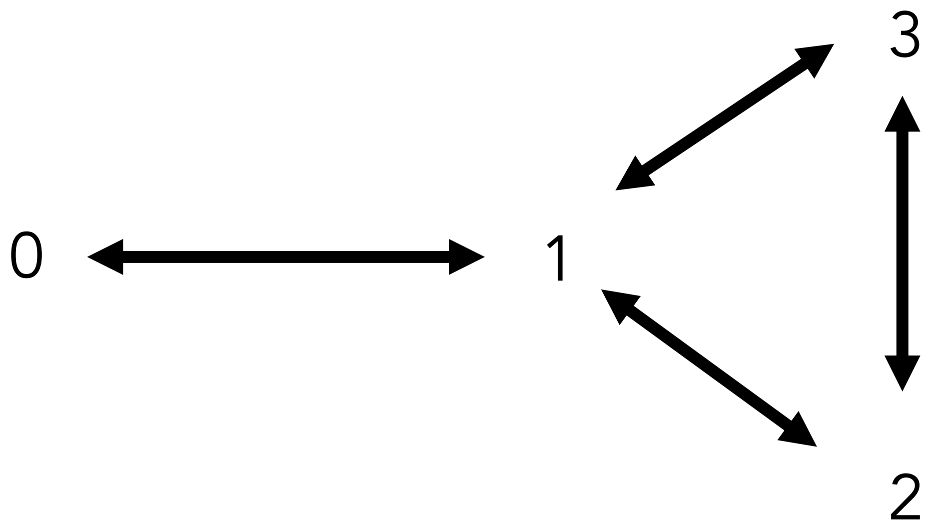
For unordered characters, any state can change into any other state.

Ordered characters

All characters ordered:

$0 \longleftrightarrow 1 \longleftrightarrow 2 \longleftrightarrow 3$

Specific characters ordered:

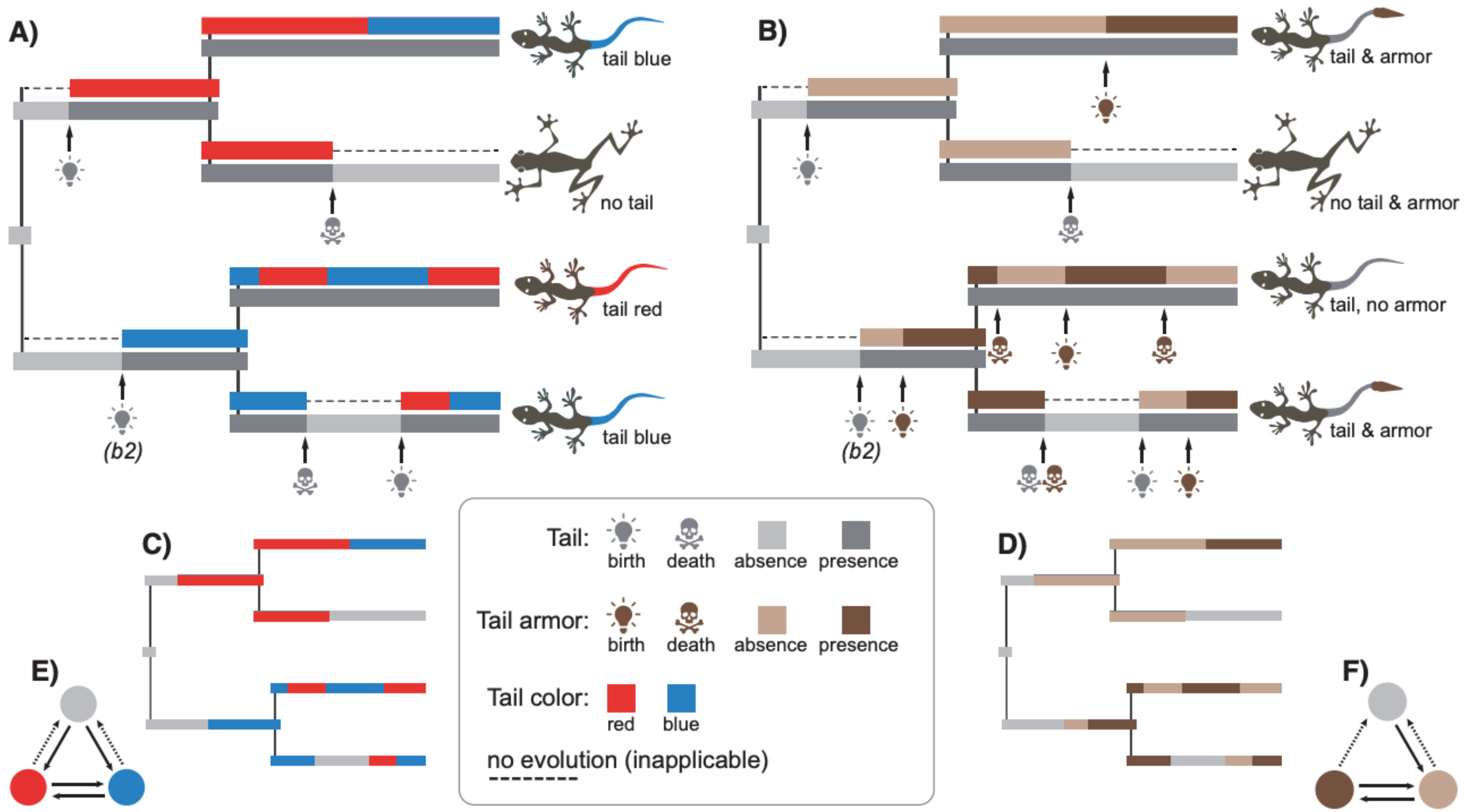


Phylogenetic Comparative Methods Harmon [2019]

Embedded dependency model

Markov models for phylogenetic inference with anatomically dependent (inapplicable) morphological characters

Non-applicable characters only considered when they are present (1)



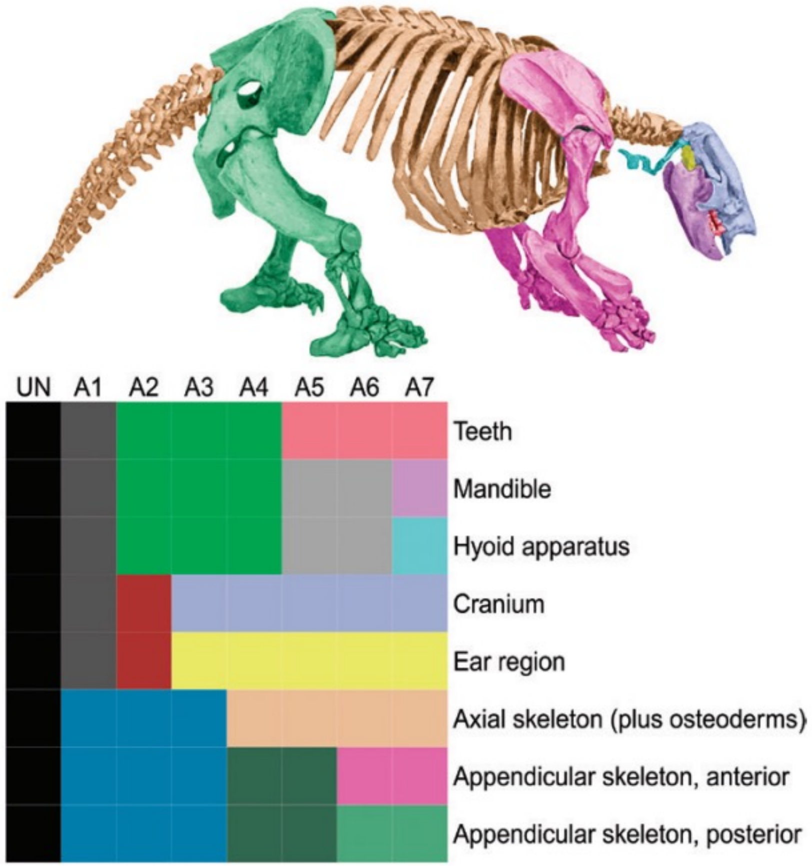
Tarasov 2023 *Systematic Biology*

Alternative partitioning schemes

Reassessing the phylogeny and divergence times of sloths (Mammalia: Pilosa: Folivora)

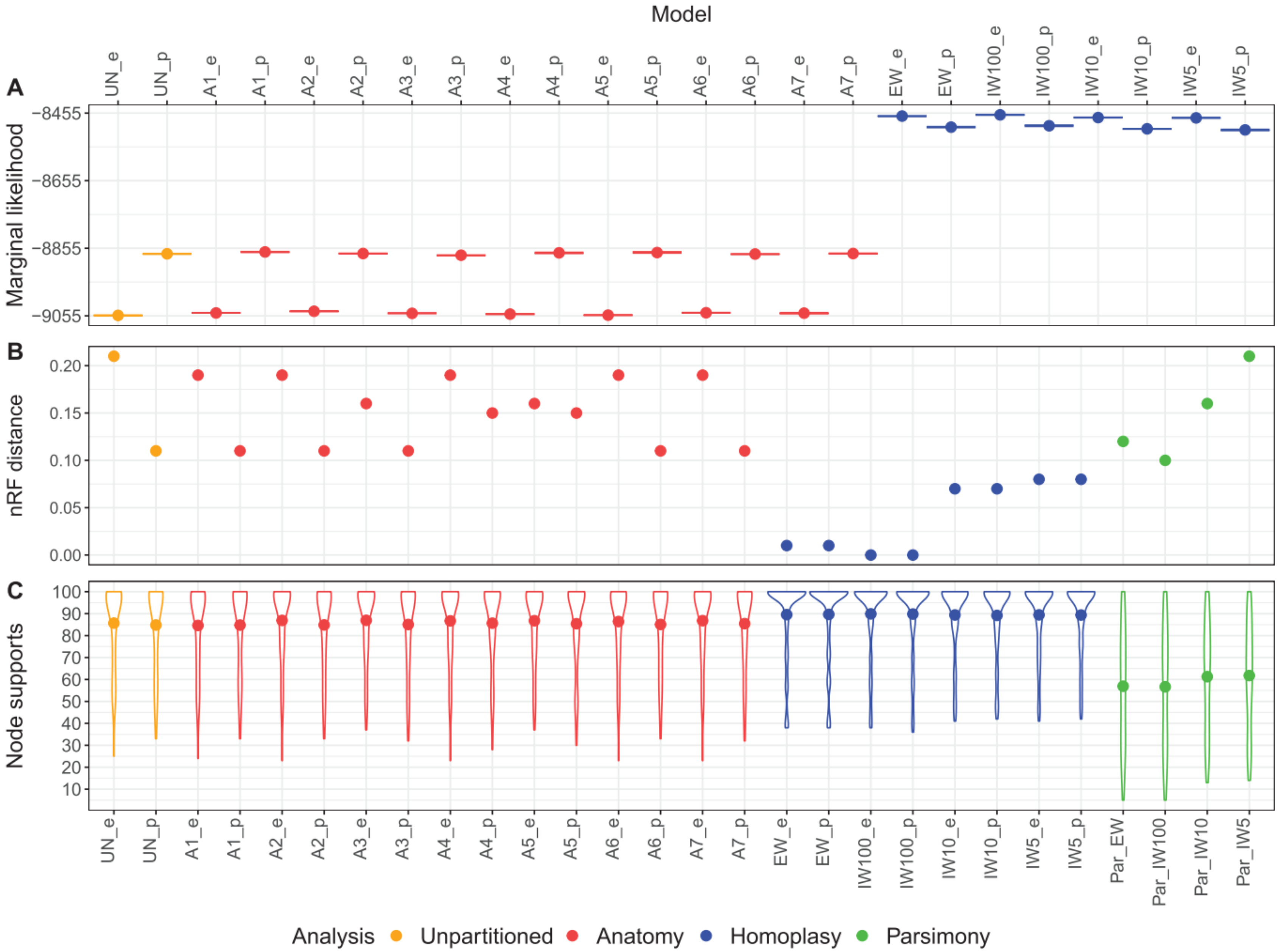
Characters can be groups based on anatomical region

Other criteria such as the degree of homoplasy present in a character was explored in this study – and found to be a better fit using Bayes factors



Casali et al 2022 Zoological Journal of the Linnean Society

Alternative partitioning schemes



Casali et al 2022 Zoological Journal of the Linnean Society

Challenges with morphological data

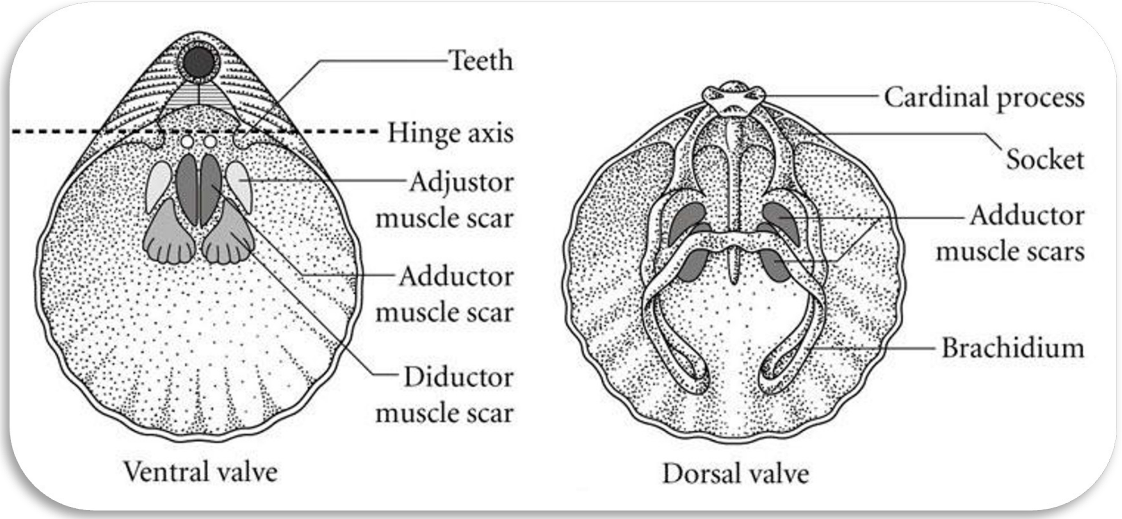
Generalising assumptions across different traits is often not possible
Modelling special characters in matrices

Character correlation occurs when two or more characters are **not independent**. Functional/developmental linkage: Traits are biologically linked. Example: The length of finger bones may be correlated with the length of the hand.

```
101510010?00-100--0000000000
000500010?200100--0010010000
102500010?200100--0?10010000
00?5?0010?200100?-0???010110
0015000101201000430100011111
```

Challenges with morphological data

Morphological matrices are often quite small:
- Collection is very time consuming
- Number of characters available can be very small depending on the group



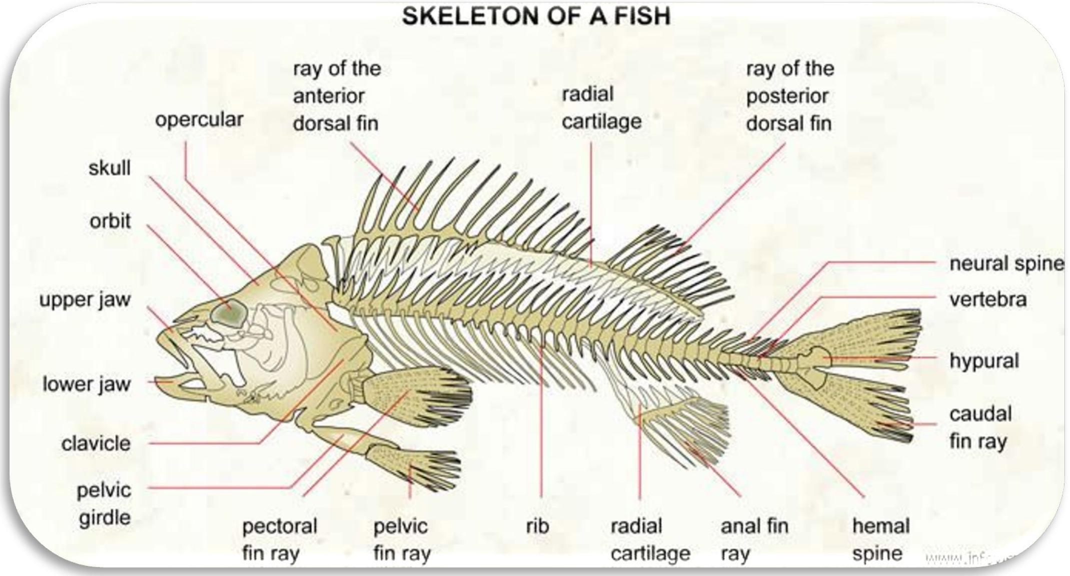
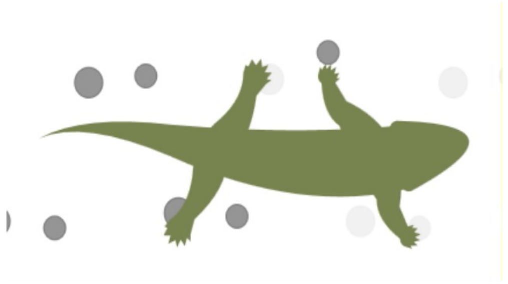


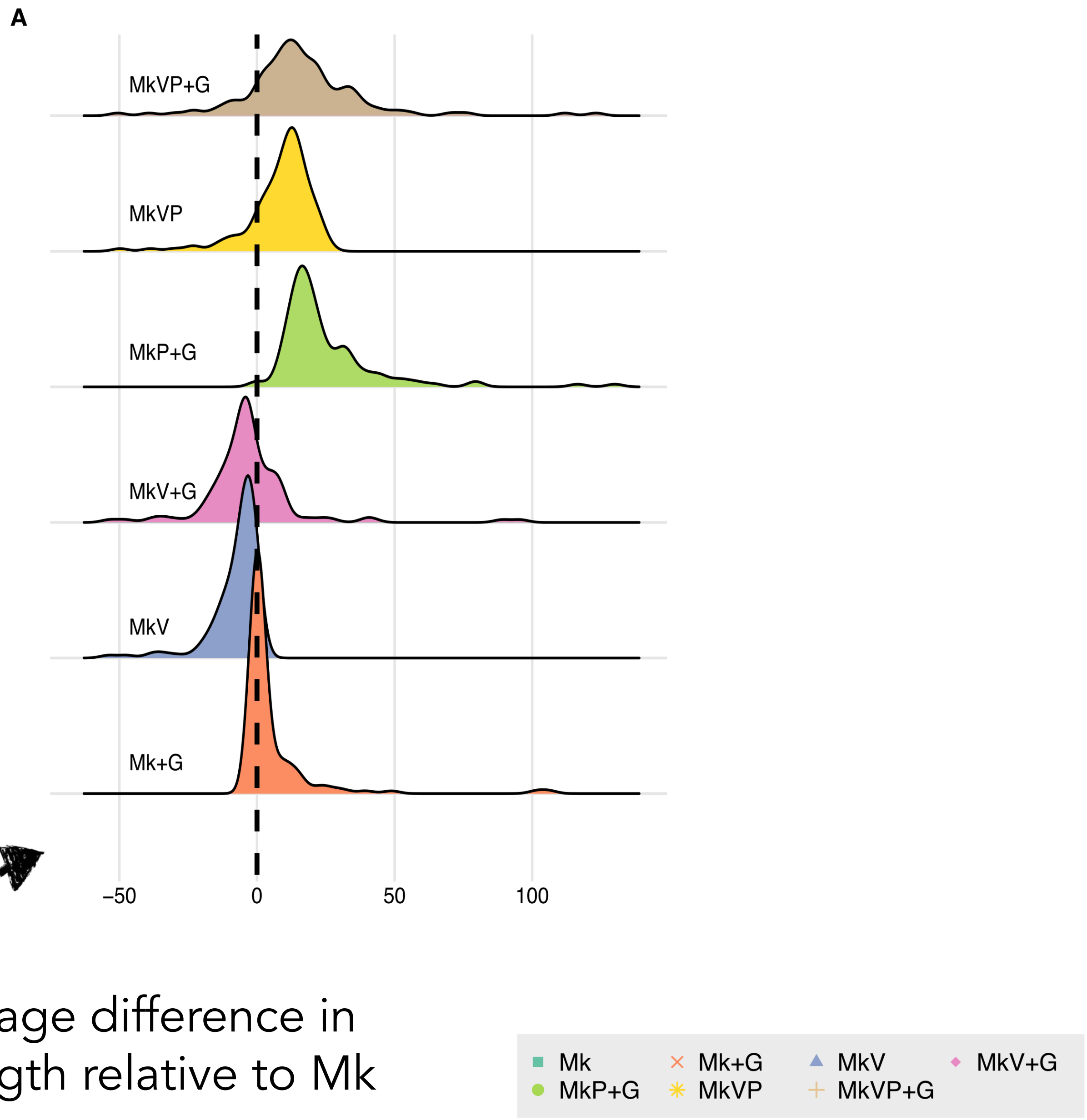
Image source

Impact of model on key parameter estimates

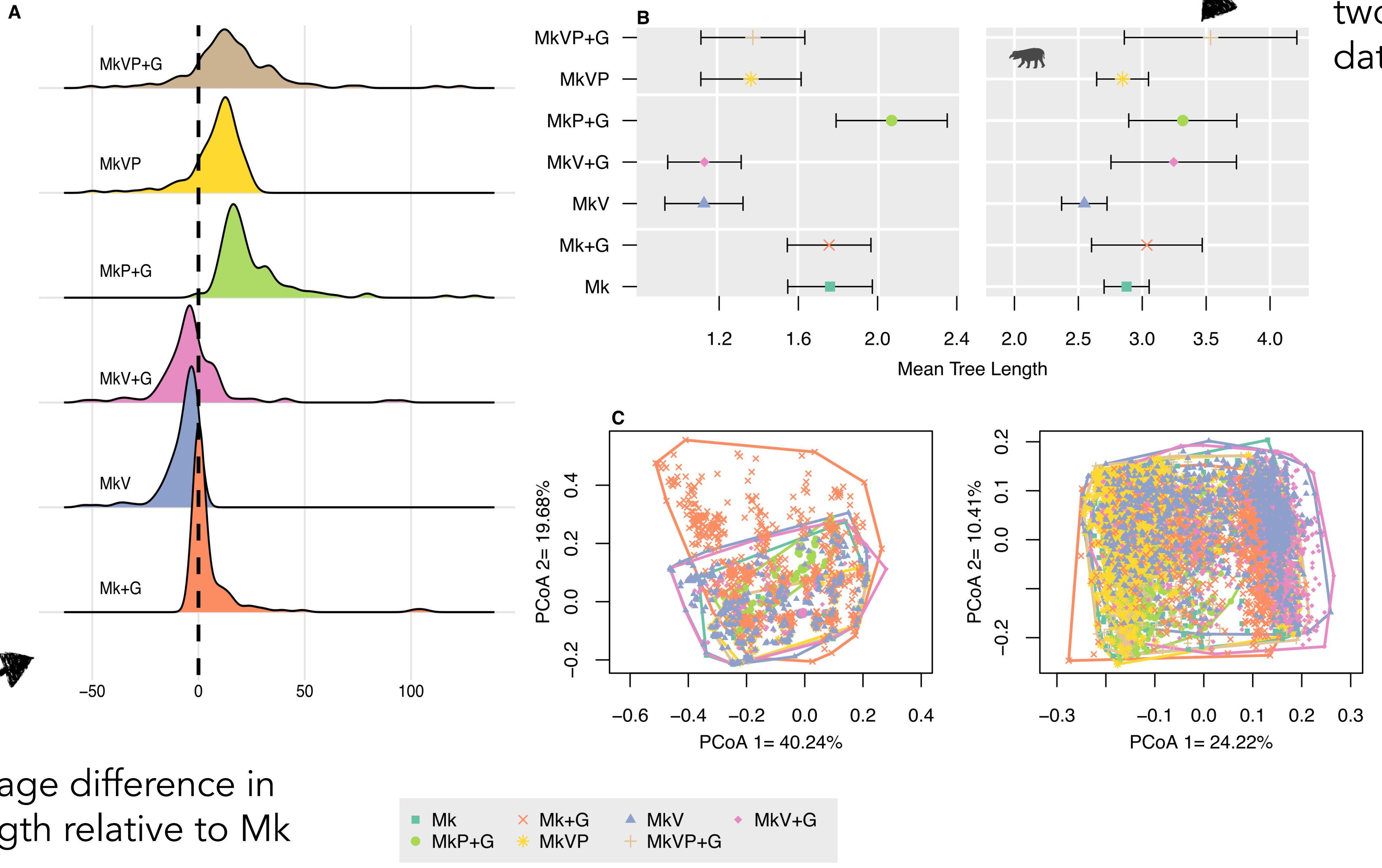
Example of 114 empirical **tetrapod** matrices



Looked at the impact on:
- branch lengths **(evolutionary distances)**
- Tree topology **(species relationships)**



Percentage difference in tree length relative to Mk model



Tree length of two different data sets

Rf distances of two data sets

Percentage difference in tree length relative to Mk model

How do we choose a model?

Model selection

Bayes factors are commonly used to determine the **relative fit** between model.

It relies on comparing the marginal likelihoods approximated from different models.

The ML measures the average fit of a model to our data.

We use MCMC to avoid calculating this number as it is computationally expensive and often not directly possible.

Model selection

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})\, P(\text{model})}{P(\text{data} \mid \text{model})}$$

Posterior

Likelihood

Priors

Marginal
likelihood

Marginal likelihood

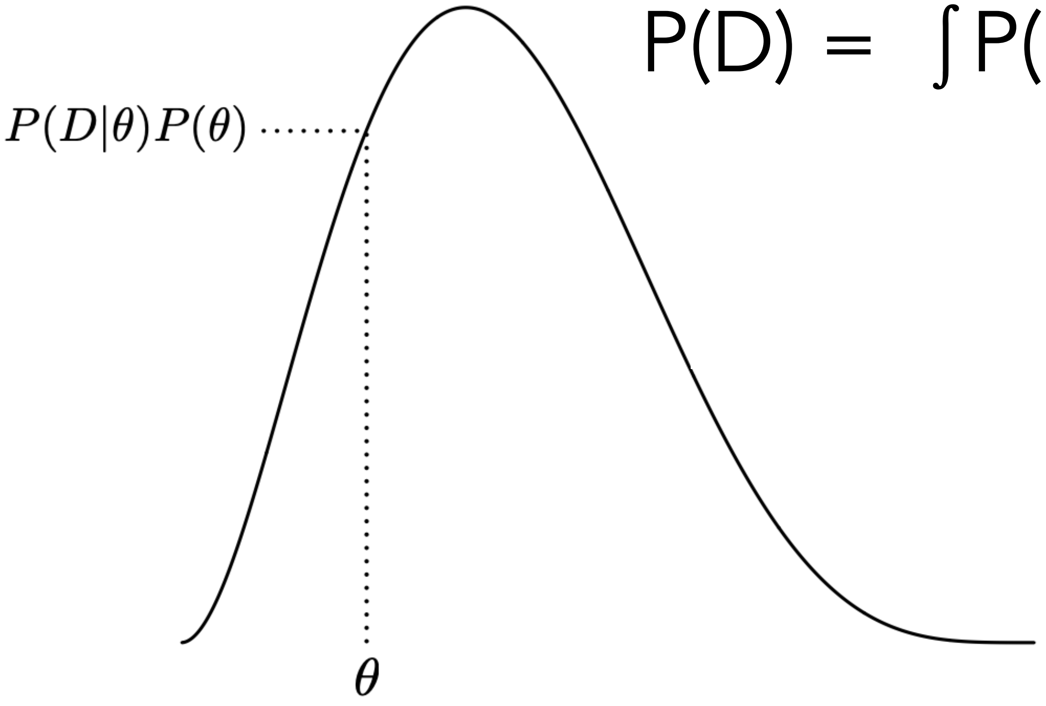
Marginal probability of the data (denominator in Bayes' rule) is the expected value of the likelihood with respect to the prior distribution.

If likelihood measures model fit, then the marginal likelihood measures the average fit of the model to the data over all parameter values.

What is the expected value?

Marginal likelihood

P (data | model)



$$P(D) = \int P(D|\theta)P(\theta)\, d(\theta)$$

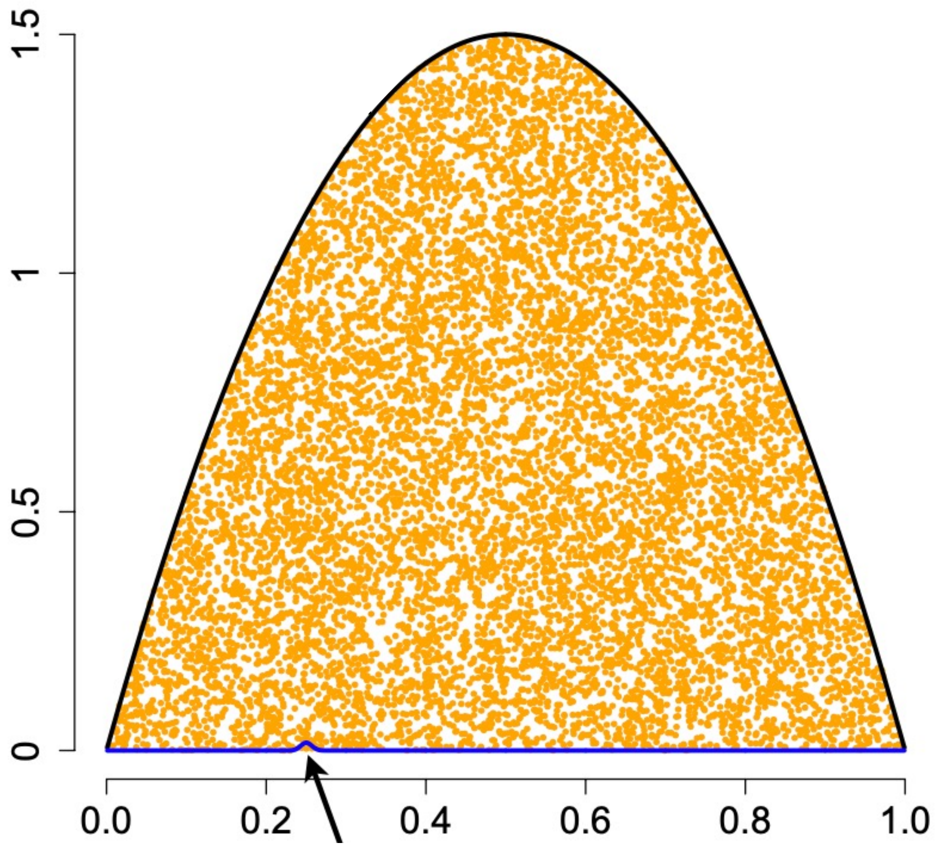
The marginal likelihood is used to evaluate the overall fit of the model to the data, integrating over all parameter values.

Adapted from Paul Lewis PhyloSeminar

Marginal likelihood

P (data | model)

Very small, single number between the posterior distribution and the prior



Adapted from Paul Lewis PhyloSeminar

Approximating the marginal likelihood

There are two common algorithms to do this:

- Stepping stone
- Path sampling

Both of these approaches are computationally expensive

Stepping-stone algorithms are like a series of MCMC simulations that iteratively sample from a specified number of distributions that are discrete steps between the posterior and the prior probability distributions.

Bayes factors

$$B_{01} = \frac{P(D \mid M_0)}{P(D \mid M_1)} = \frac{\text{Marginal likelihood for model } M_0}{\text{Marginal likelihood for model } M_1}$$

Bayes factors

$$B_{01} = \frac{P(D \mid M_0)}{P(D \mid M_1)} = \frac{\text{Marginal likelihood for model } M_0}{\text{Marginal likelihood for model } M_1}$$

Marginal likelihoods are often on the log scale so the Bayes factor can be calculated as:

$$\log B_{01} = \log P(D \mid M_0) - \log P(D \mid M_1)$$

Bayes factors

Strength of evidence	$BF(M0,M1)$	$log(BF(M0,M1))$
Negative (supports M_1)	<1	<0
Barely worth mentioning	1 to 3.2	0 to 1.16
Substantial	3.2 to 10	1.16 to 2.3
Strong	10 to 100	2.3 to 4.6
Decisive	>100	>4.6

Table source

Issues with Bayes factors for morphological data

The way we **partition data** for morphological data is different to molecular

010023
201102
112131

Unpartitioned everything in Q-matrix of size 4

Partitioning the data puts characters into correctly sizes Q-matrix

The way we **partition data** for morphological data is different to molecular

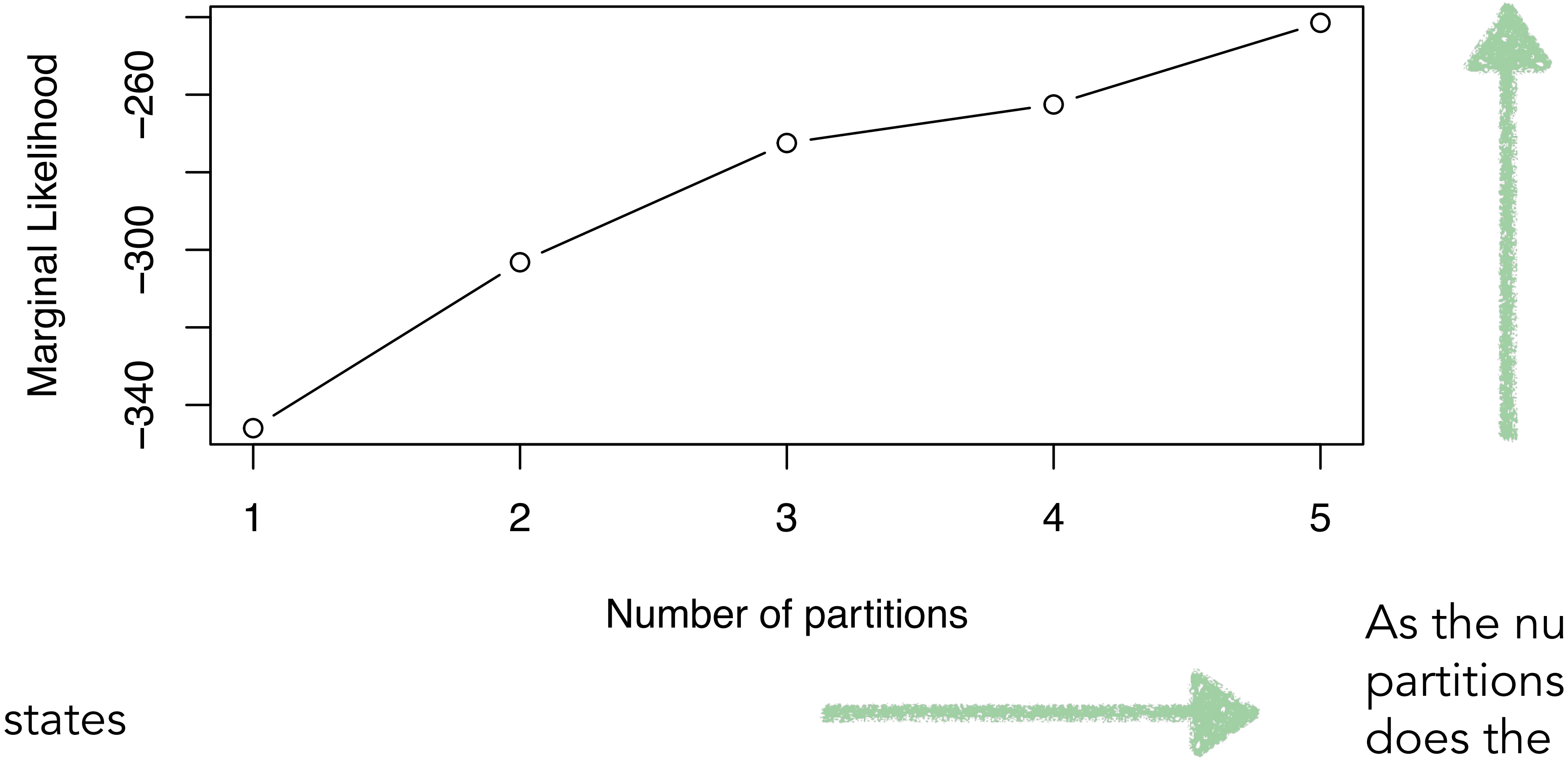
010023
201102
112131

10	00	23
01	21	02
11	12	31

Unpartitioned everything in Q-matrix of size 4

Partitioning the data puts characters into correctly sizes Q-matrix

Issues with Bayes factors for morphological data



As the number of partitions increases, so does the likelihood

Data set with 6 states

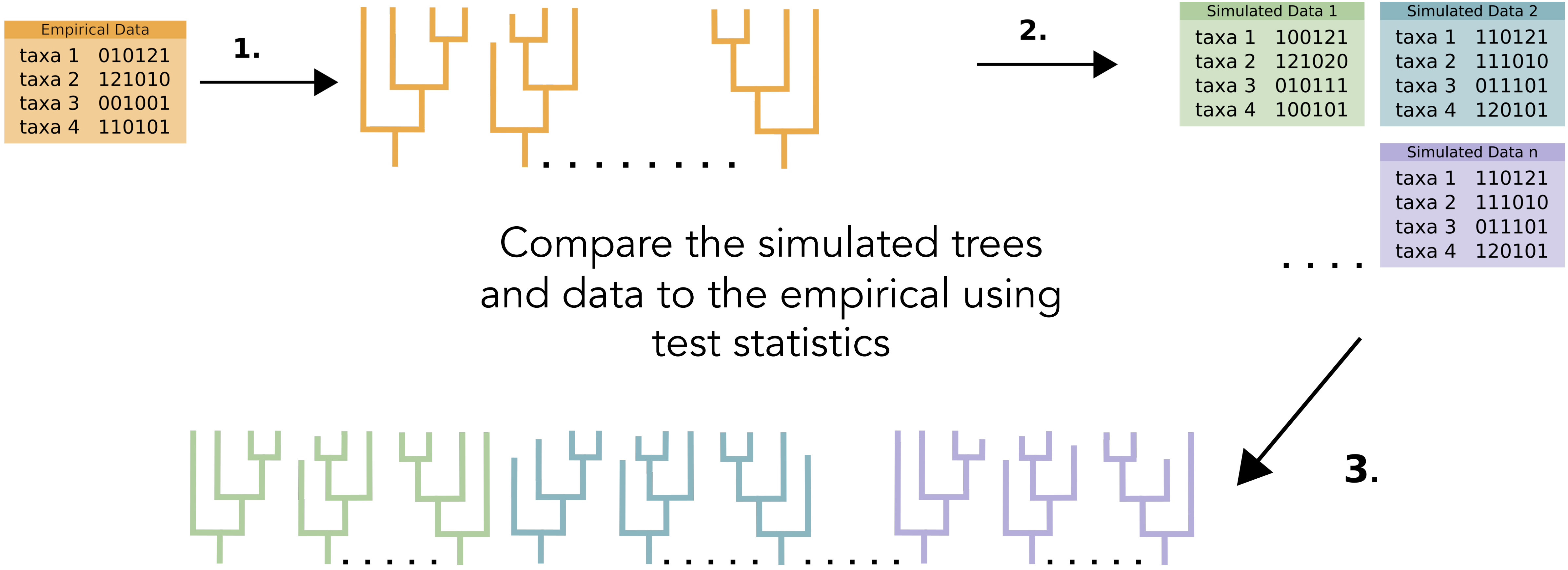
Model adequacy

Assess whether a model is capturing the evolutionary dynamics that generated the data

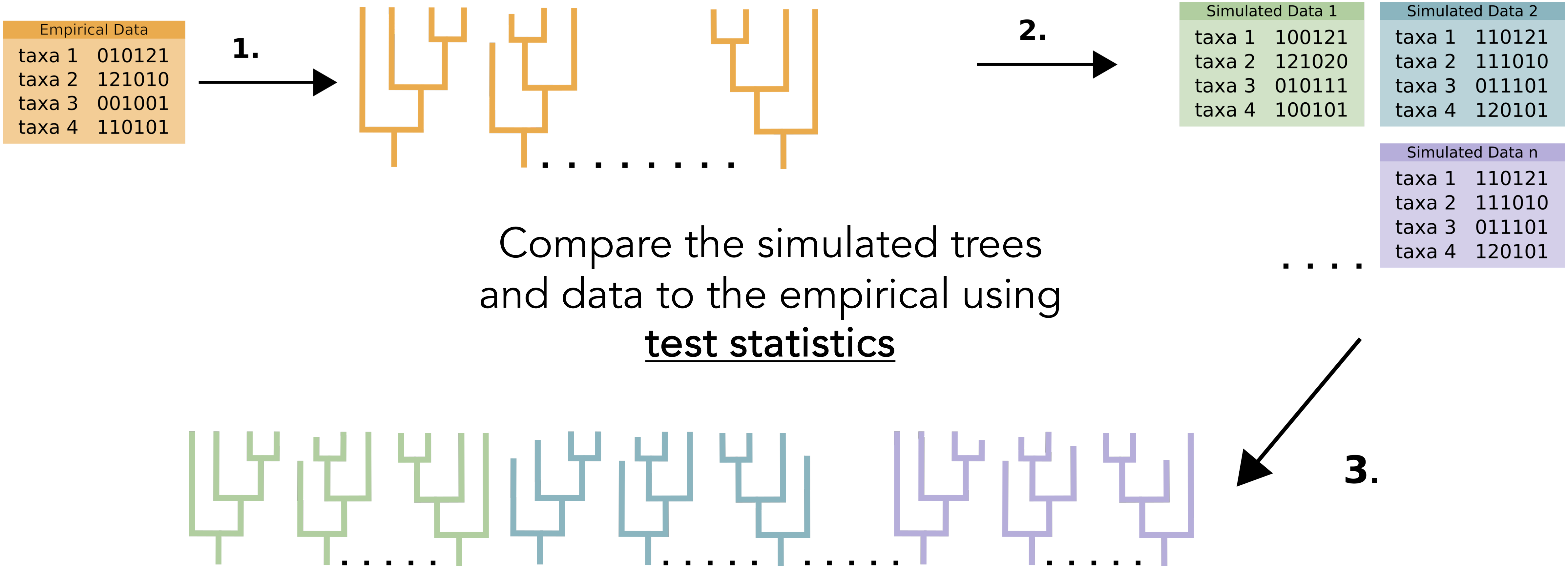
Gives the **absolute fit**

One approach is **Posterior Predictive Simulations**

Posterior Predictive Simulations



Posterior Predictive Simulations



Test statistics

A test statistic is a **numerical summary** of data.
A value that captures the characteristic of you data.

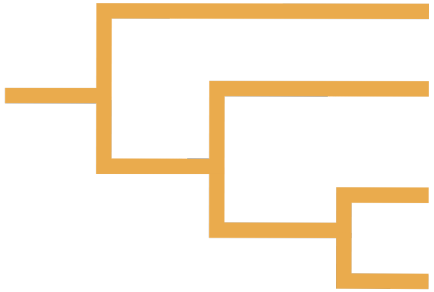
For PPS we have 3 categories:
Data-based, inference-based, mixed

Test statistics: CI

Calculating **consistency index**

Empirical Data	
taxa 1	010121
taxa 2	121010
taxa 3	001001
taxa 4	110101

MCC summary tree



Simulated Data 1	
taxa 1	100121
taxa 2	121020
taxa 3	010111
taxa 4	100101

Simulated Data 2	
taxa 1	110121
taxa 2	111010
taxa 3	011101
taxa 4	120101

. . . .

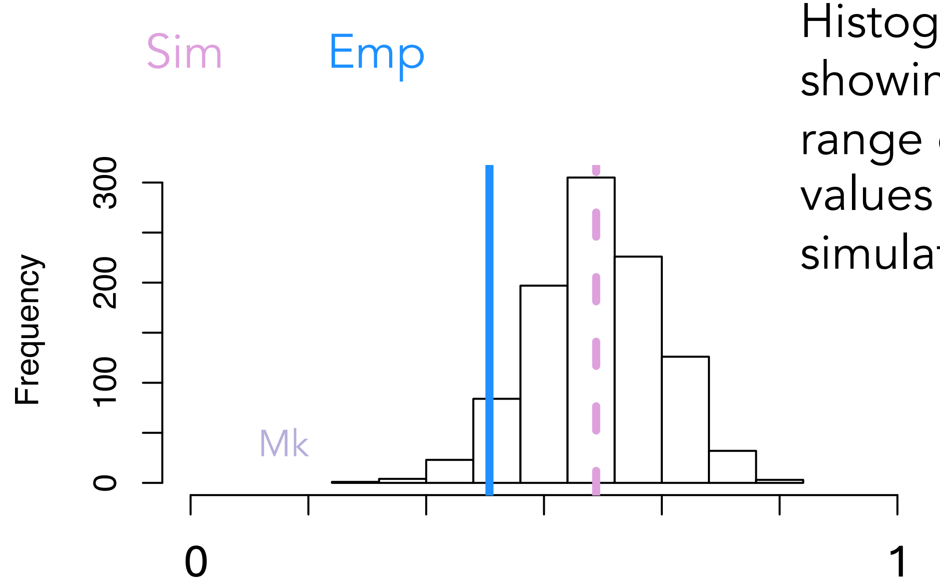
Simulated Data n	
taxa 1	110121
taxa 2	111010
taxa 3	011101
taxa 4	120101

Calculate **one value** for the empirical data set

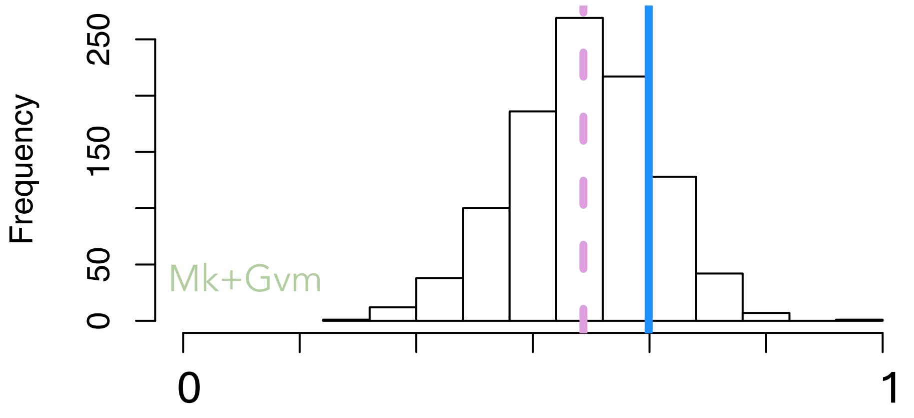
consistency index: measure of homoplasy (convergent evolution)

Calculate **a range (500) values** using all simulated data sets

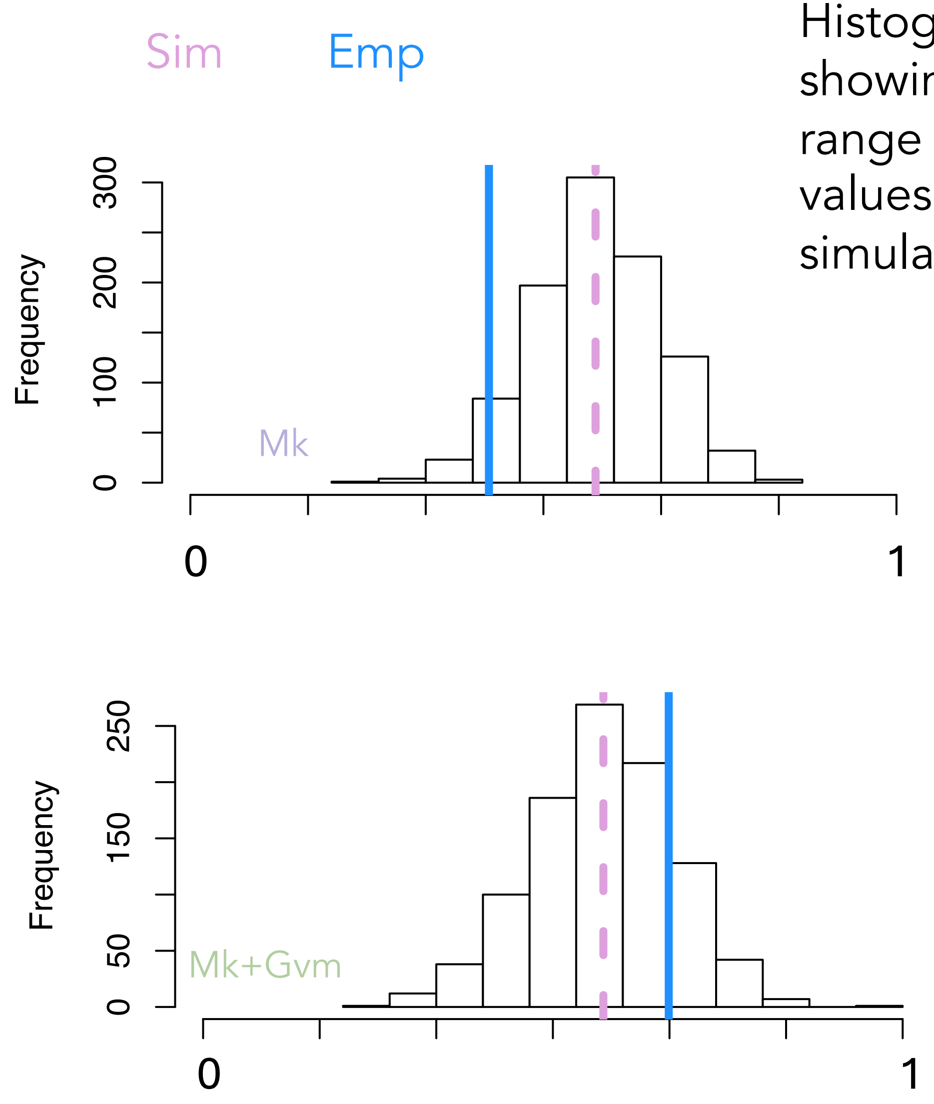
Test statistics: CI



Histograms showing the range of CI values for all the simulated data



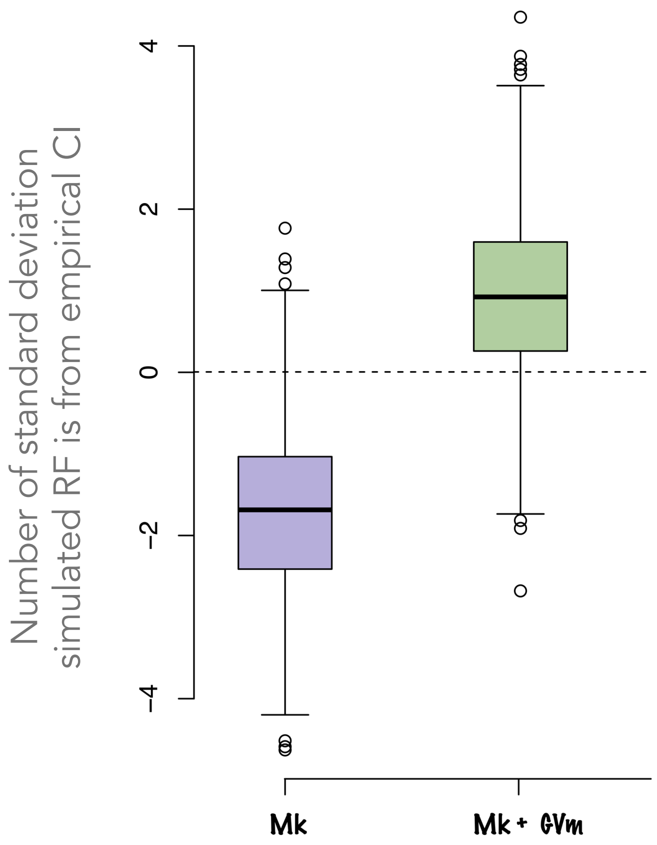
Test statistics: CI



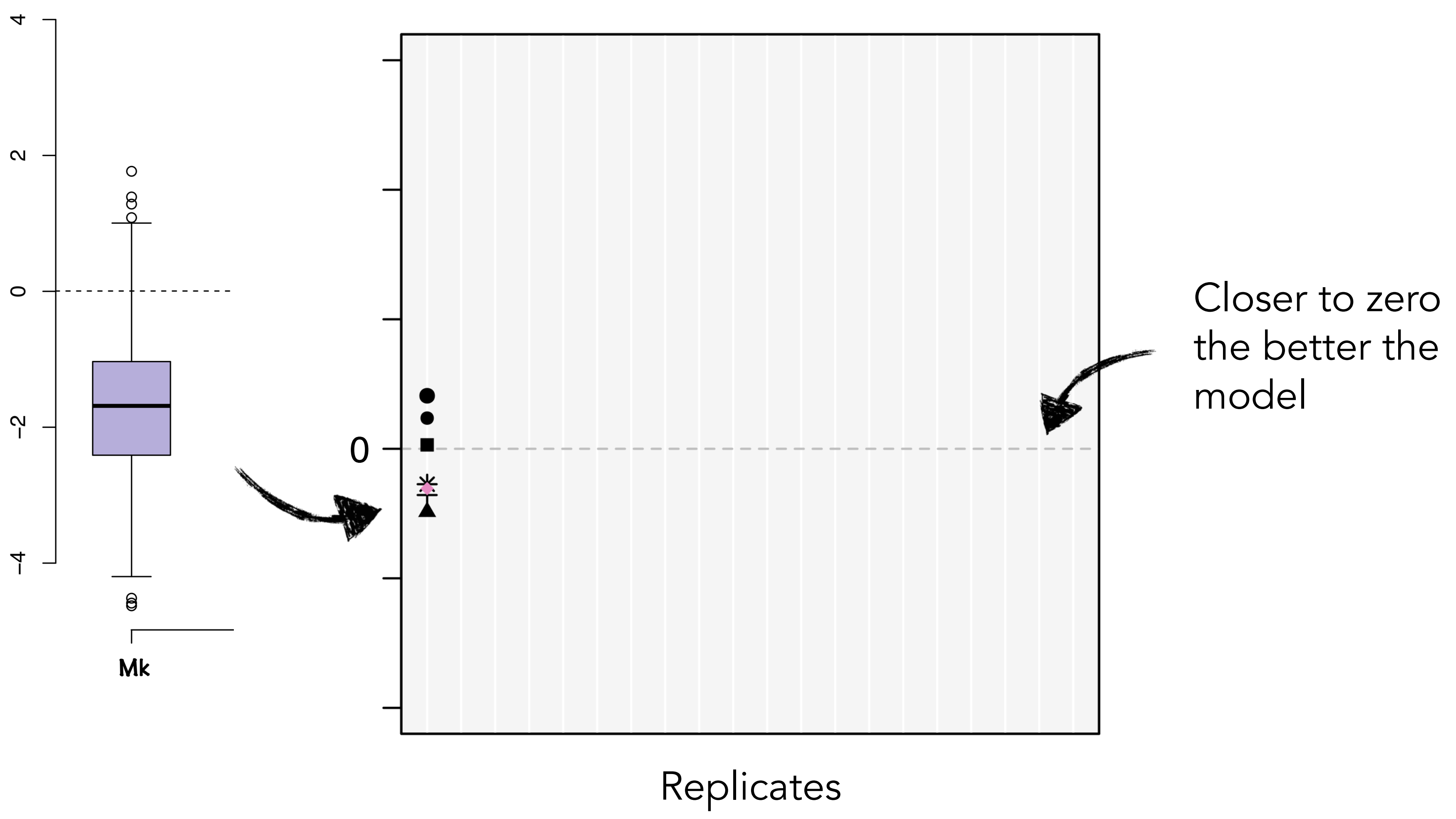
Histograms showing the range of CI values for all the simulated data

We can use this to calculate **effect sizes**

$$\frac{\text{Empirical TS - SimTs}}{\text{Sd(All Sim TS)}}$$



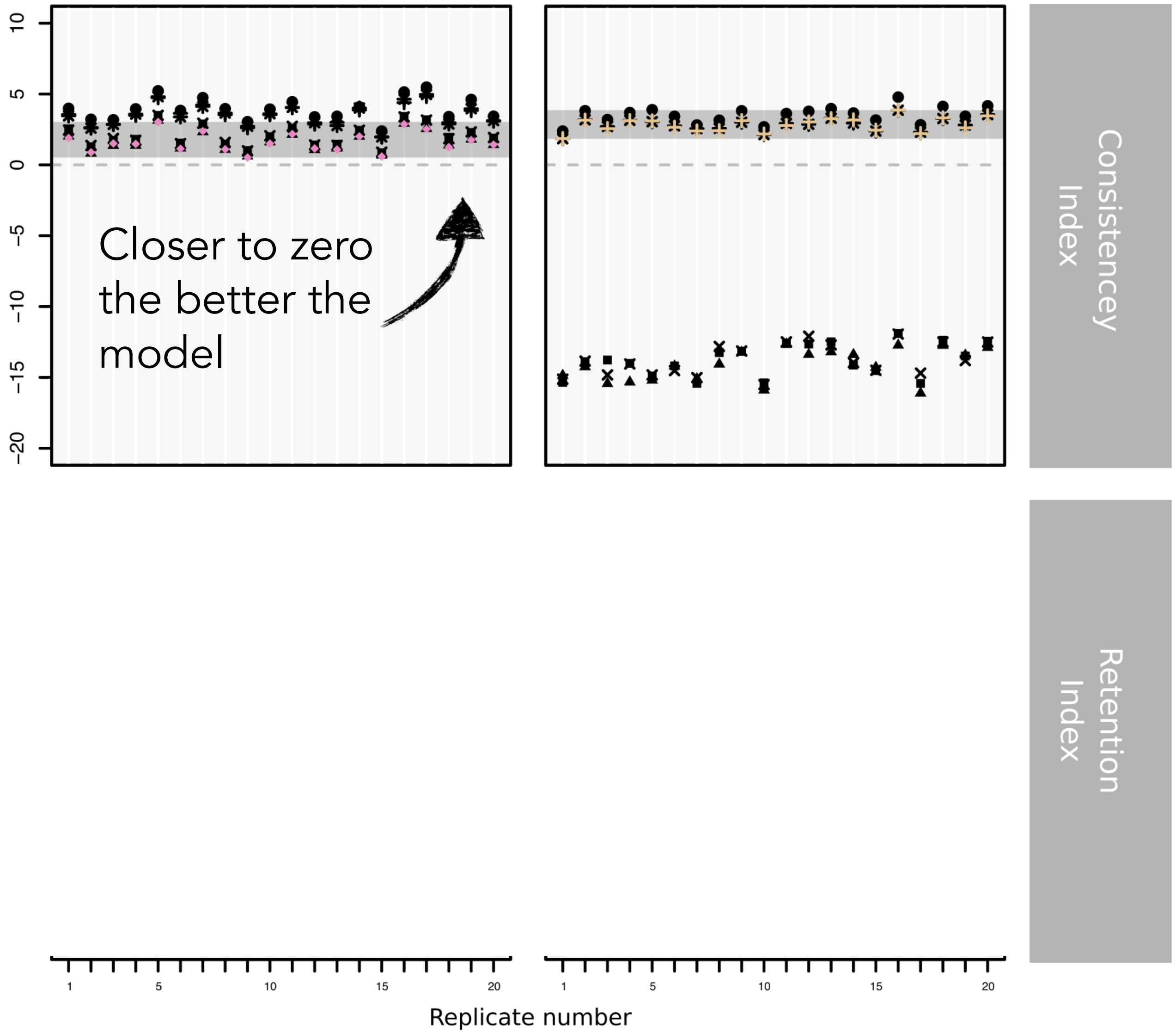
Effect sizes



Test statistics: CI

Simulated under the MkV+G model:

Simulated under the MkVP+G model:



These test statistics **are informative** about the correct model

We do see the correct model consistently closest to zero

Empirical data sets

MkVP+G

MkVP

MkP+G

MkV+G

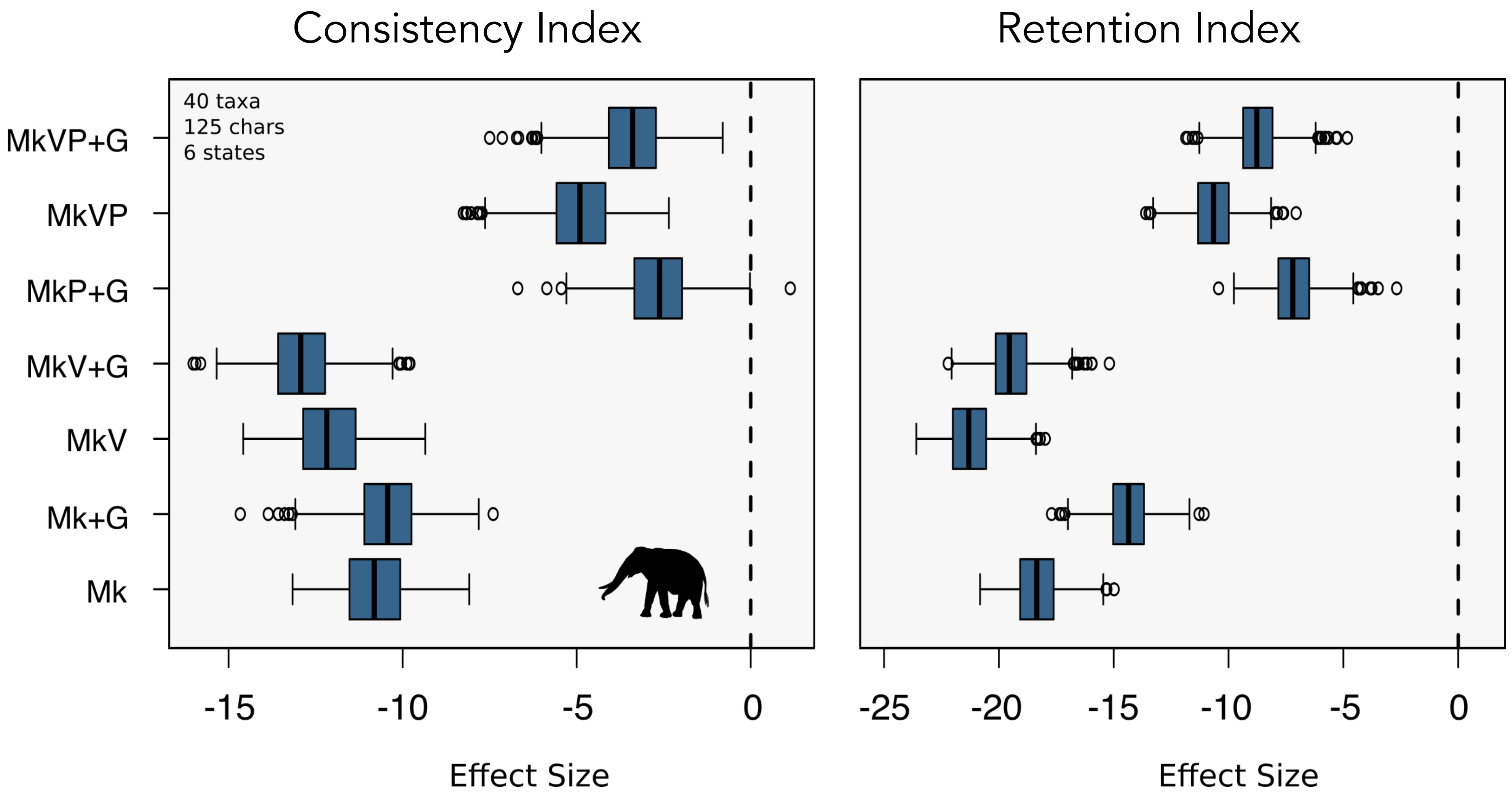
MkV

Mk+G

Mk

ound 3
odels that
re
dequate

Empirical data sets



No models are adequate

Model adequacy exercise