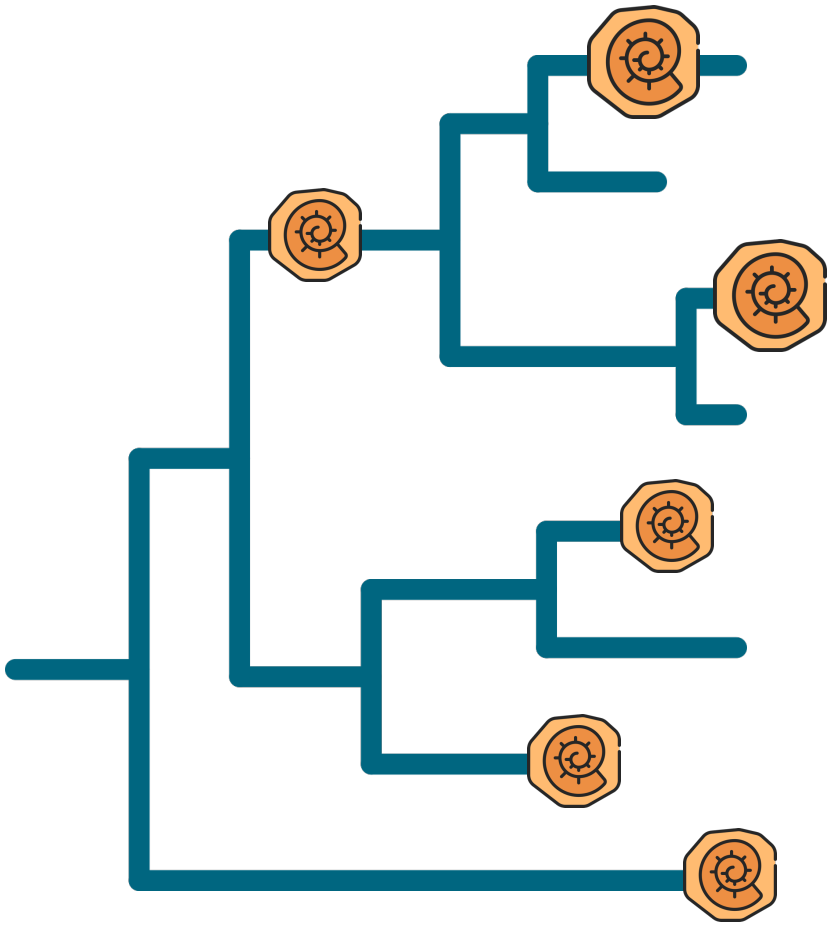


Morphological models and how to choose them

Laura Mulvey

PalAss 2024

Phylogenetics Workshop

Outline

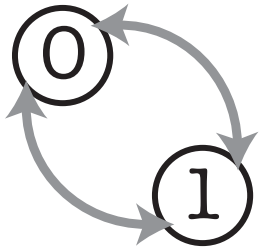
Morphological data

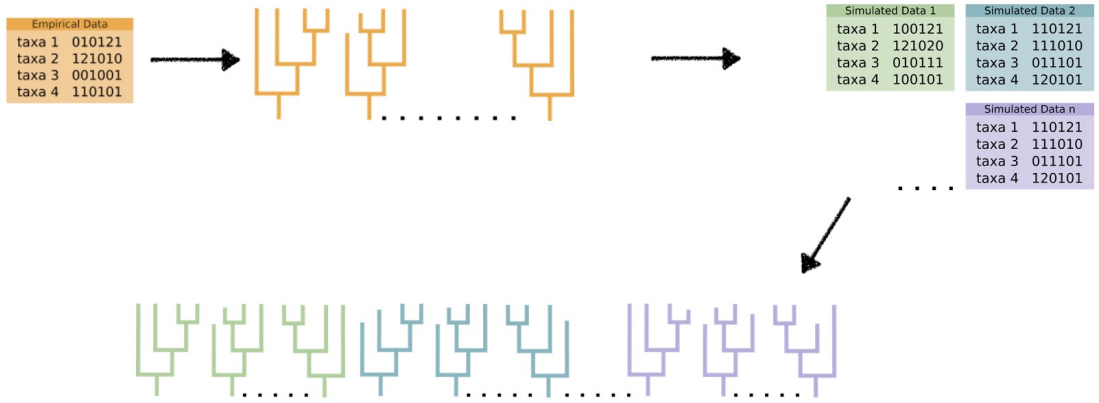
0100201102…….
1220111102…….

Morphological models and RevBayes

Model adequacy

Tutorial

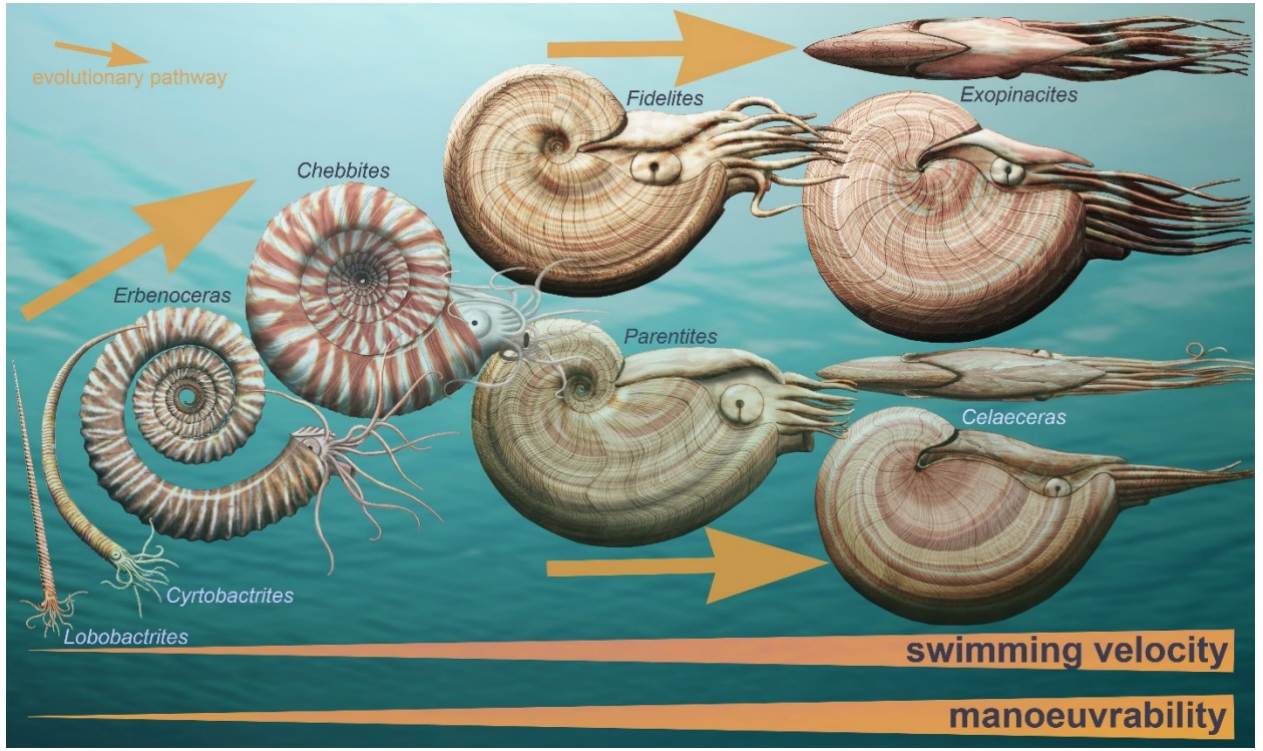




Morphological data

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

Morphological character 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

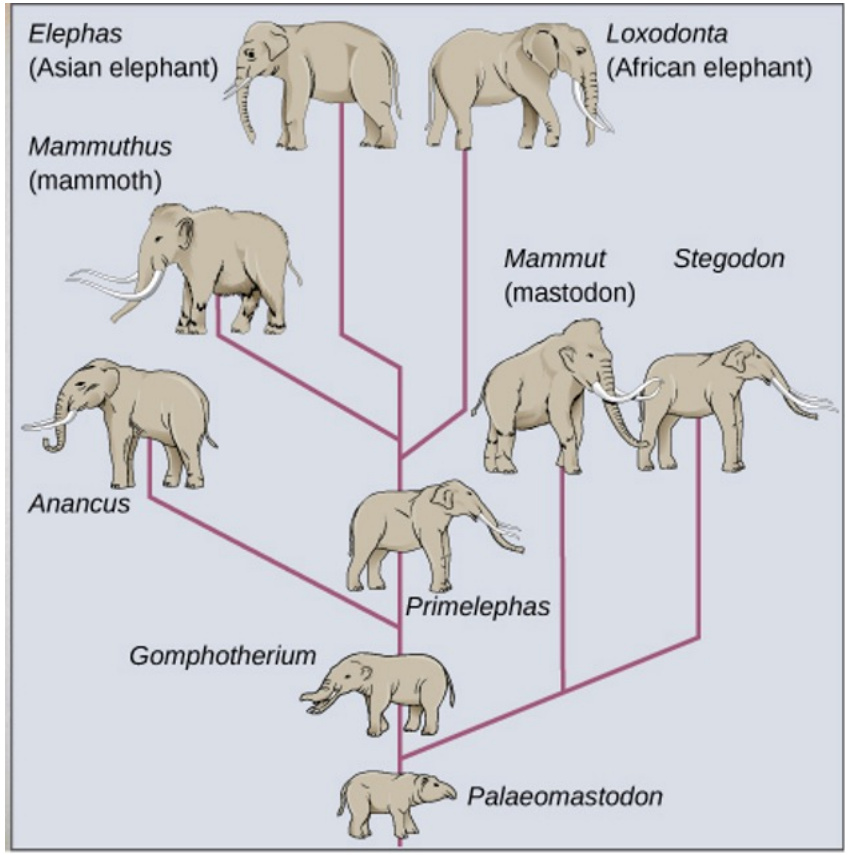


Image from https://www.zoologytalks.com/

Morphological character 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

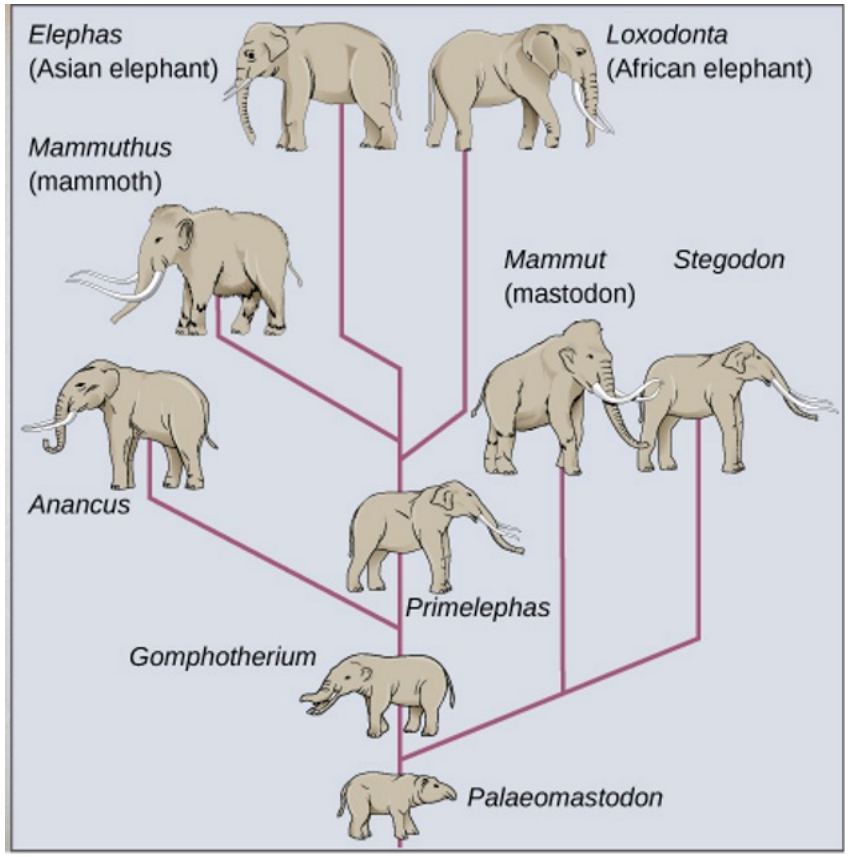
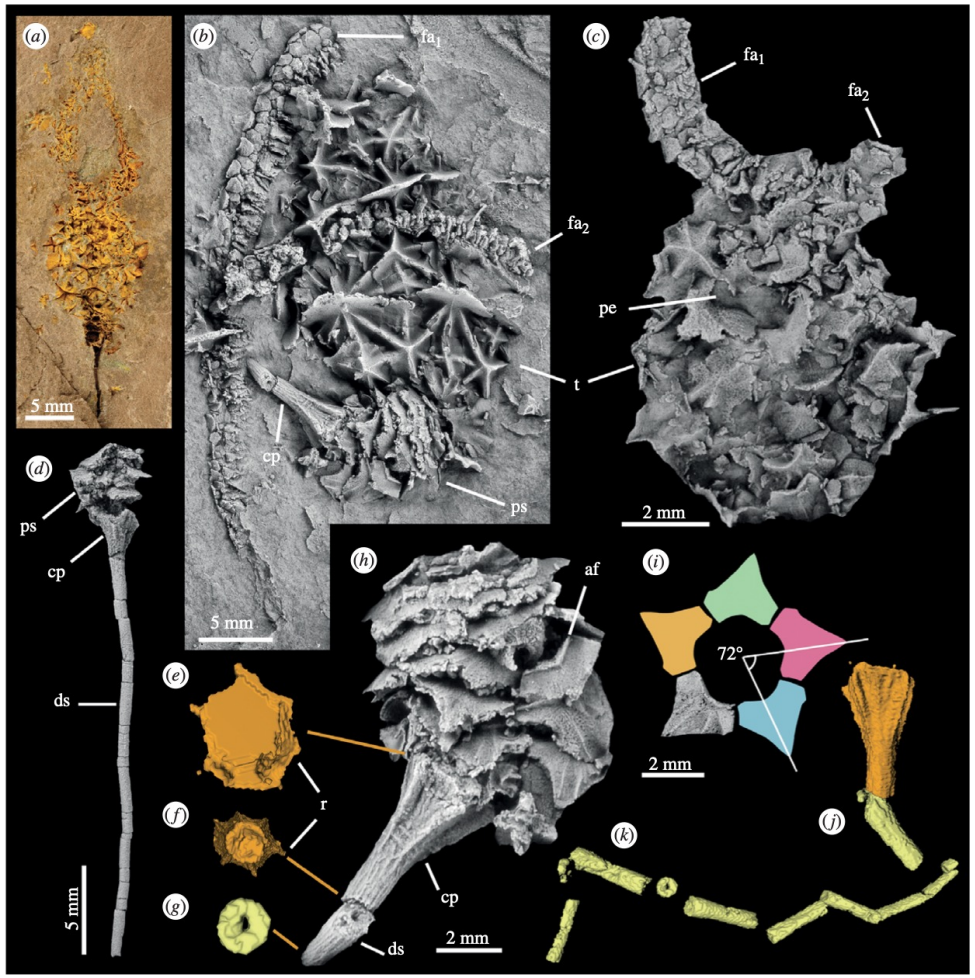


Image from https://www.zoologytalks.com/

Traits →

Taxa ↓

```
0015100 10?00-100--0000000000
0005000 10?200100--0010010000
0025000 10?200100--0?10010000
00?5?00 10?200100?-0???010110
0015000 1012010004301000111111
0015000 1012010104401110111111
??050?? ???20100044 0?11011111
01050?0 10-2100000 501??010110
0002000 1002101003-1110010110
0002000 1002110014411210111111
000201111-2100 10?-??11011121
?103?0?11?1001104-0000010000
1005002 1101000 10--0?00110?20
1005002 0001010 10540?00110020
```



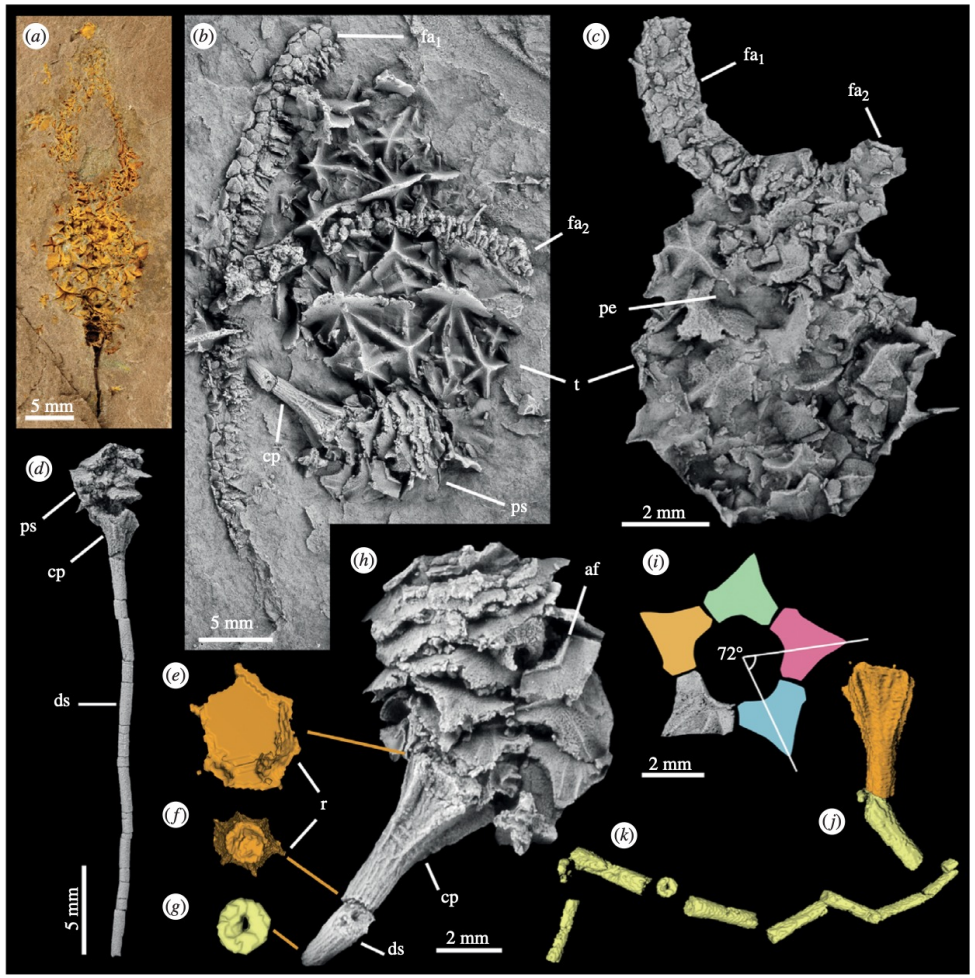
Cambrian stalked echinoderms show unexpected plasticity of arm construction
Zamora & Smith. 2012 Proc B

Traits →

Taxa ↓

```
00151001O?00-100--0000000000
00050001O?200100--0010010000
00250001O?200100--0          00
00?5?001O?200100?-0???01011O
00150001012010004301000111111
00150001012010104401110111111
??050?????2010004O?11011111
01050?01O-210000?501??01011O
00020001002101003-1110010110
00020001002110014411210111111
00020111-2100lO?-??11011121
?103?0?11?1001104-0000010000
10050021101000l0--0?00110?20
10050020001010105402?0011002O
```

Presence/ absence



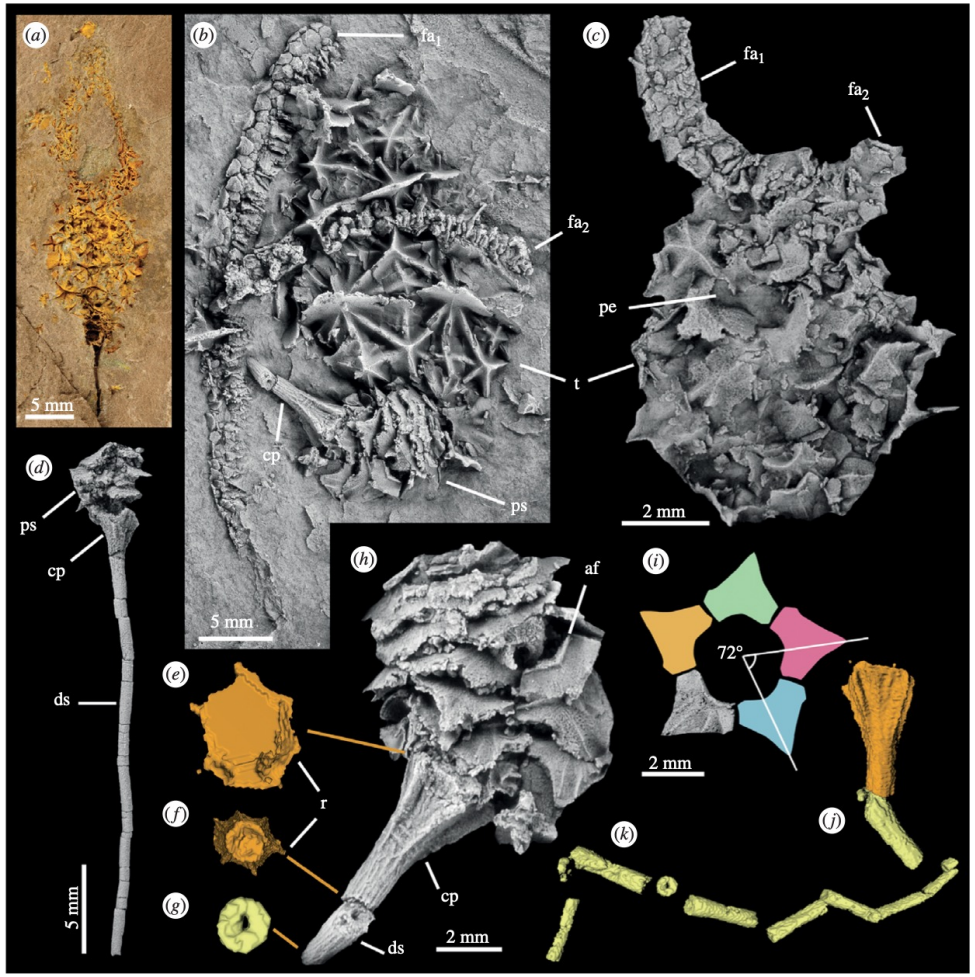
Cambrian stalked echinoderms show unexpected plasticity of arm construction
Zamora & Smith. 2012 Proc B

Traits →

Taxa ↓

Appendage pattern

```
0015100l0?00-100--0000000000
0005000l0?200100--0010010000
0025000l0?200100--0?10010000
00?5200l0?200100?-0???010110
0015000l0?201000430100011111
0015000l012010104401110111111
??050?????201000440?11011111
01050?010-2100000?501??010110
0002000l002101003-1110010110
0002000l002110014411210111111
000201111-2100l0?-??11011121
?103?0?11?1001104-0000010000
1005002110100010--0?00110?20
1005002000l01010540?00110020
```



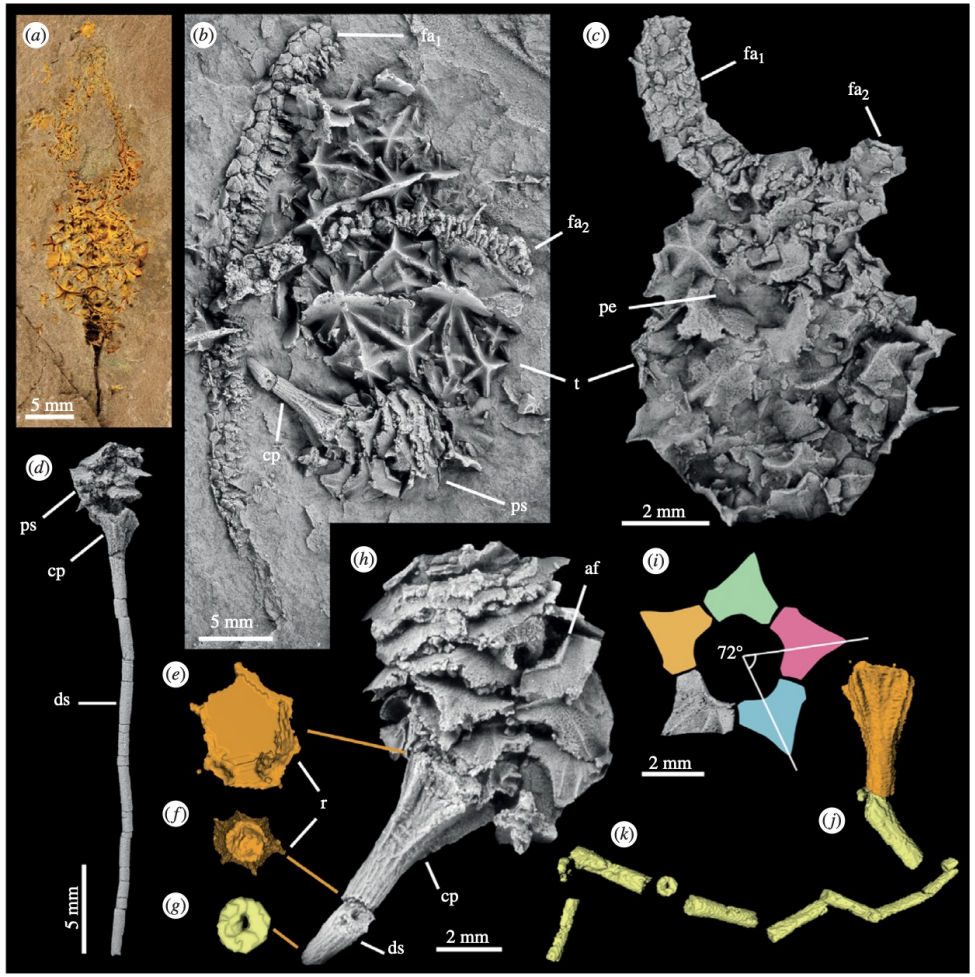
Cambrian stalked echinoderms show unexpected plasticity of arm construction
Zamora & Smith. 2012 Proc B

Traits →

Taxa ↓

```
00151001O?00-100--0000000000
00050001O?200100--0010010000
00250001O?200100--0?10010000
00?5?001O?200100?-0???010110
00150001012010004301000111111
001500010120101044011101111?
??050?????20100044O?11011111
01050?010-210000?501??010110
0002000100210100 3-1110010110
0002000100211001441121011111
00020111-2100107-??11011121
?103?0?11?1001104-0000010000
10050021101000 0--0?00110?20
10050020001010105 40?00110020
```

Missing data



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

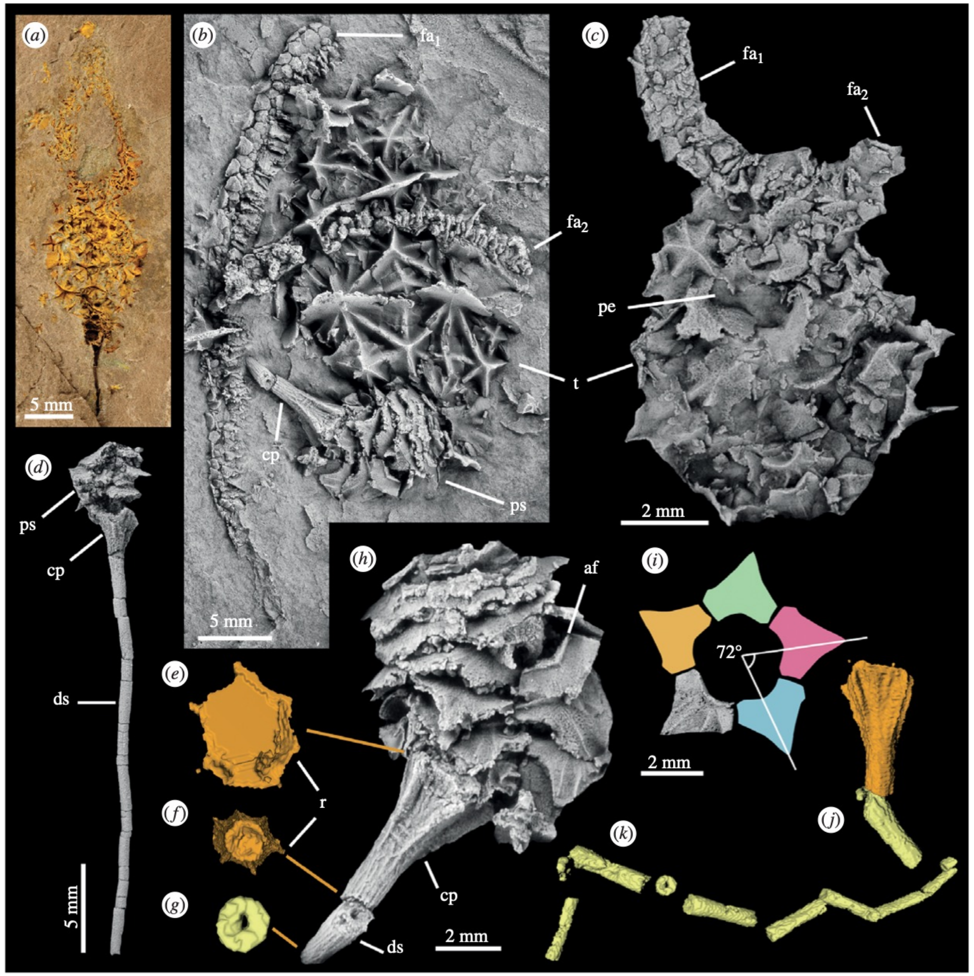
Traits →

Taxa ↓

```
00151001 0?00-100--000000000
```

```
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
00020001002101003-1110010110
00020001      01441121011111
000201111-210010?-??11011121
?103?0?11?1001104-0000010000
1005002110100010--0?00110?20
1005002000101010540?00110020
```

Non-applicable



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

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
Ambiguous	0/1/2	Used when it is not clear which character trait is present in the taxon

Morphological models

Morphological models in RevBayes

Using the source function we can read in different scripts for our analysis.

It can be helpful to have a number of different model components in different scripts that you can switch in and out of an analysis as you want

Here I will explain the assumptions of common morphological models and then provide the code to run them in RevBayes

Make a directory **scripts /**

Mk.rev

MkV.rev

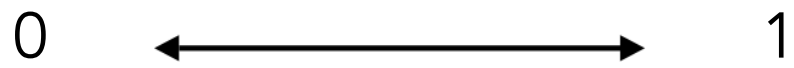
Mk+G.rev

MkP.rev

Mk Model

Assumes equal transition probabilities between states

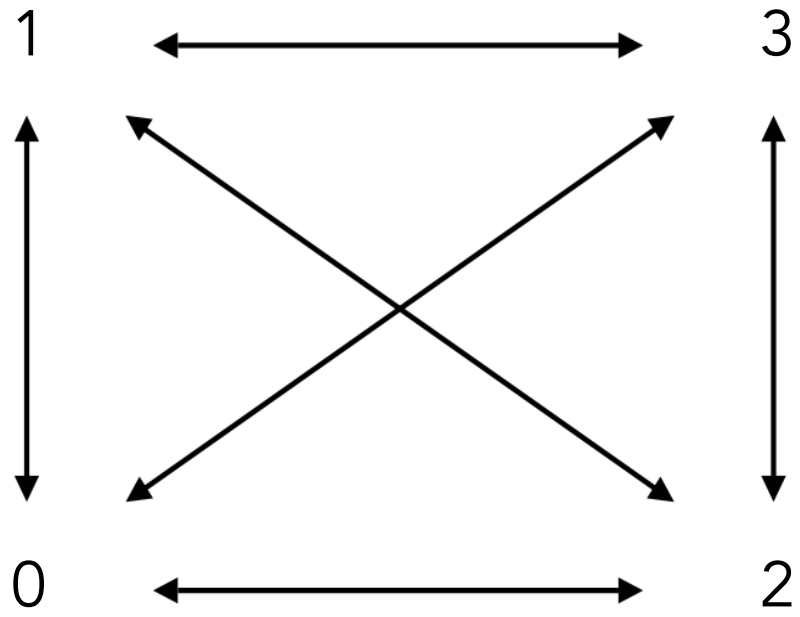
Equal state frequencies



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

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!

Mk

```
## define our Q matrix
Q <- fnJC(5)


## create phylo object
seq ~ dnPhyloCTMC(tree=phylogeny, Q=Q, type="Standard")
seq.clamp(morpho)
```

MkV model

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

though there are plenty……

```
00151O010?00-100--0000000000
000500010?200100--0010010000
002500010?200100--0?10010000
00?5?0010?200100?-0???010110
00150001012010004301000111111
00150001012010104401110111111
??050?????201000440?11011111
01050?010-210000?501??010110
00020001002101003-1110010110
000200010021100144112101111
000201111-210010?-??11011121
?103?0?11?1001104-0000010000
10050021101000l0--0?00110?20
100500200010101O540?00110020
```

MkV model

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

though there are plenty……

All varying characters

```
00151O010?00-100--0000000000
000500010?200100--0010010000
002500010?200100--0?10010000
00?5?0010?200100?-0???010110
00150001012010004301000111111
001500010120101044011101111
??050?????201000440?11011111
01050?010-210000?501??010110
000200010021010003-1110010110
0002000100211001441121011111
000201111-210010?-??11011121
?103?0?11?1001104-0000010000
1005002110100010--0?00110?20
10050020001010105 40?00110020
```

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

```
## define our Q matrix
Q <- fnJC(5)


## create phylo object
seq ~ dnPhyloCTMC(tree=phylogeny, Q=Q, type="Standard", coding = "variable")
seq.clamp(morpho)
```

Among-character rate variation

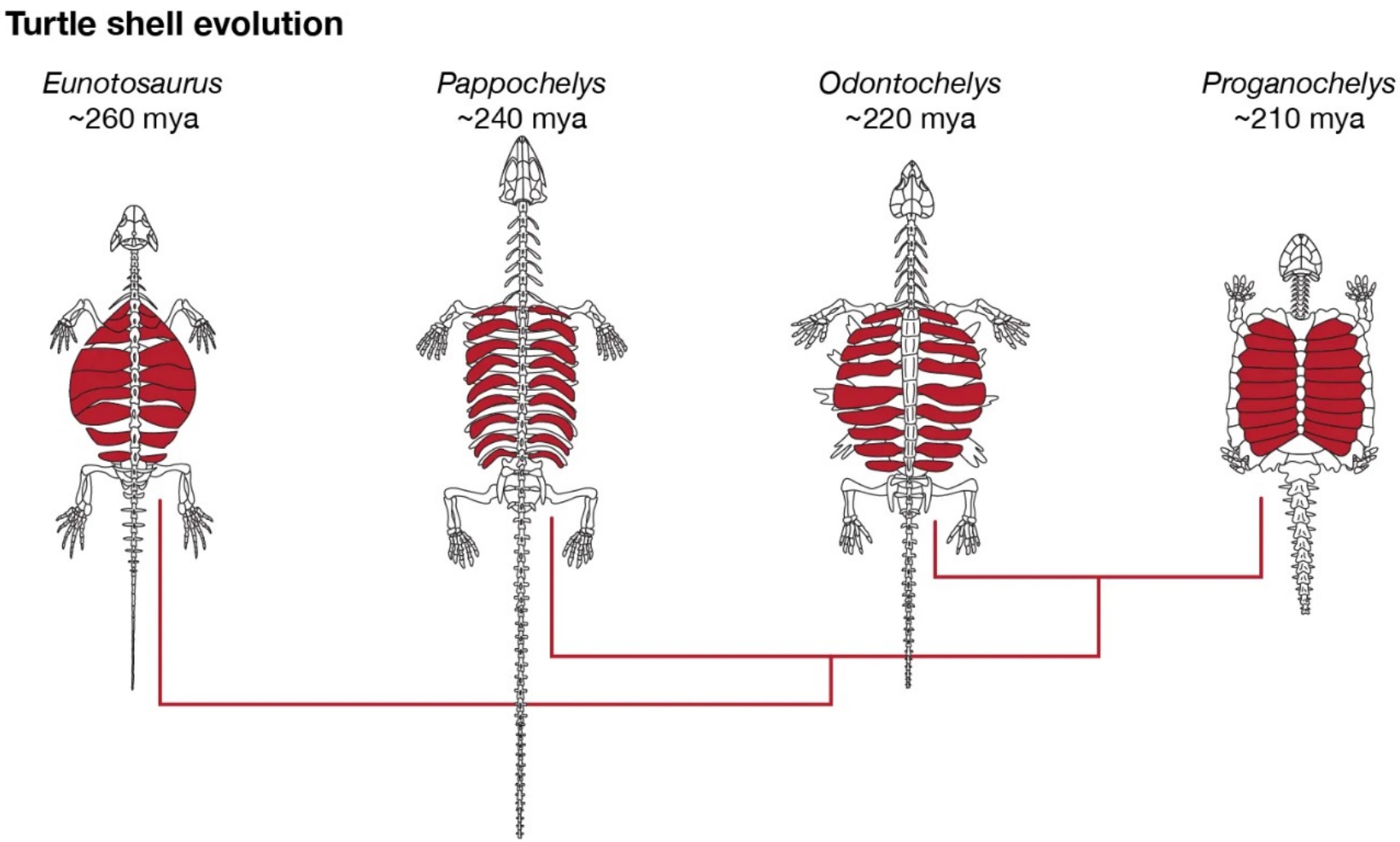


Image source

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

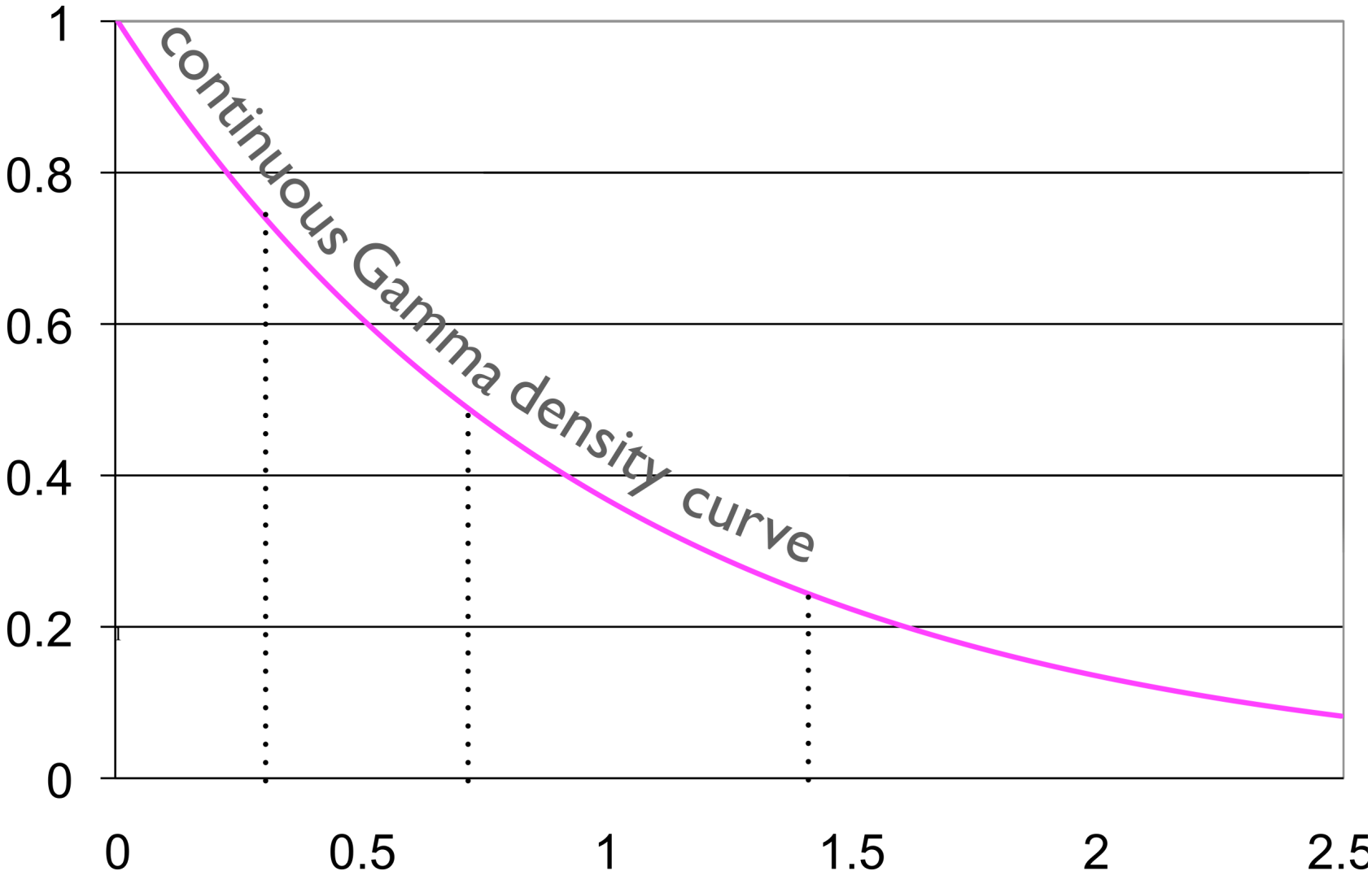
What do we do?

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

Allow these traits to evolve at different rates:

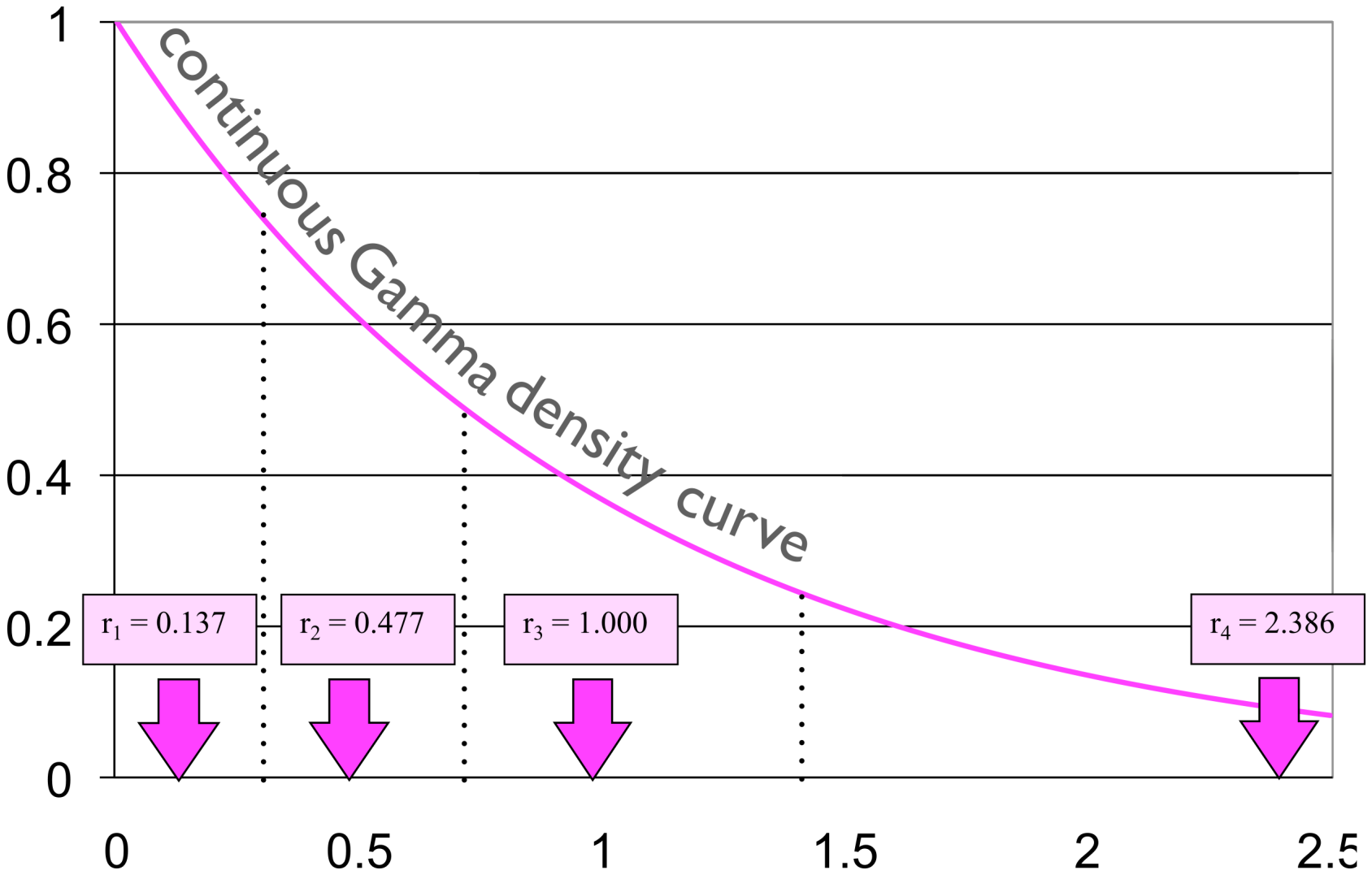
- Specify which traits evolve fast
- **Use a gamma model to account for rate heterogeneity**

Mk + Gamma



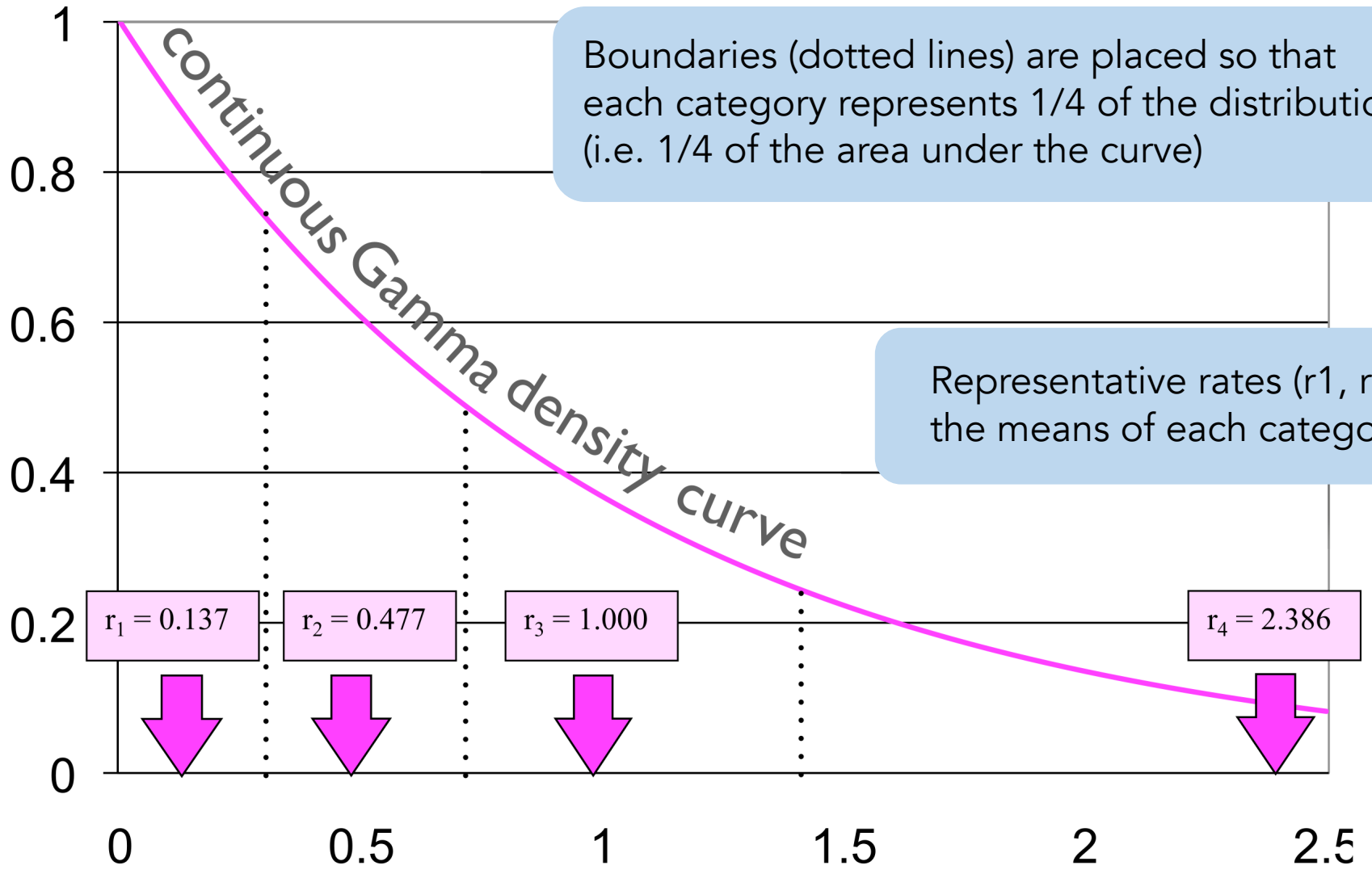
Adapted from Paul Lewis PhyloSeminar

Mk + Gamma



Adapted from Paul Lewis PhyloSeminar

Mk + Gamma



Boundaries (dotted lines) are placed so that each category represents 1/4 of the distribution (i.e. 1/4 of the area under the curve)

Representative rates (r1, r2, r3, r4) are the means of each category

Adapted from Paul Lewis PhyloSeminar

Mk + Gamma

What do we do?

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

Slower R (R1,2)

Faster R (R4)

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

Mk + Gamma

```
## define our Q matrix
Q <- fnJC(5)
# Set up Gamma-distributed rate variation.
alpha_morpho ~ dnUniform( 0.0, 1E5 )
rates_morpho := fnDiscretizeGamma( alpha_morpho, alpha_morpho, 4 )

# Moves on the parameters to the Gamma distribution.
moves.append(mvScale(alpha_morpho, lambda=1, weight=2.0))

## create phylo object
seq ~ dnPhyloCTMC(tree=phylogeny, Q=Q, type="Standard", siteRates=rates_morpho)
seq.clamp(morpho)
```

Partitioning Data

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

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

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

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

Partitioning Data

```
n_max_states <- 5
idx = 1
morpho_bystate[1] <- morpho
for (i in 2:n_max_states) {
    morpho_bystate[i] <- morpho
    morpho_bystate[i].setNumStatesPartition(i)
ze i
    nc = morpho_bystate[i].nchar()
d states

    if (nc > 0) {
        q[idx] <- fnJC(i)
        m_morph[idx] ~ dnPhyloCTMC( tree=phylogeny,
                                    Q=q[idx],
                                    nSites=nc,
                                    type="Standard" )

        m_morph[idx].clamp(morpho_bystate[i])

        idx = idx + 1
        #idx
    }
}
```

Further extensions

We can combine any of these extensions to make more complex models

Further extensions

We can combine any of these extensions to make more complex models

We can have ordered characters

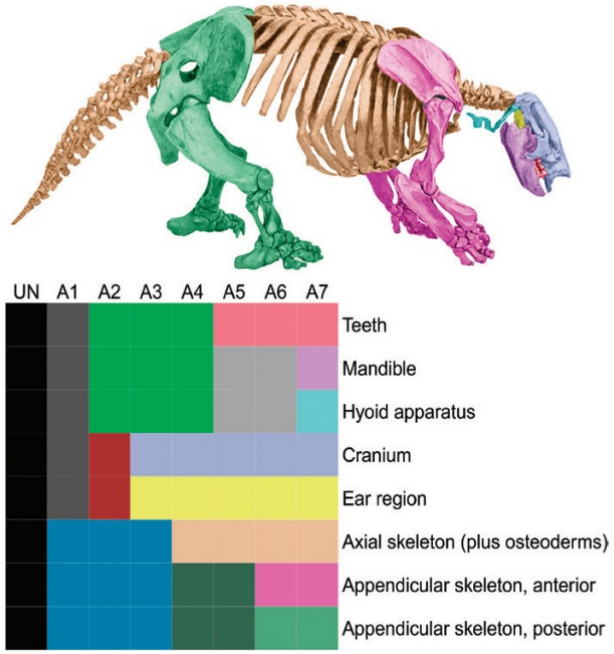
Further extensions

We can combine any of these extensions to make more complex models

0 ⟷ 1 ⟷ 2 ⟷ 3

We can have ordered characters

Anatomical partitioning schemes



Casali et al 2022 Zoological Journal of the Linnean Society

Further extensions

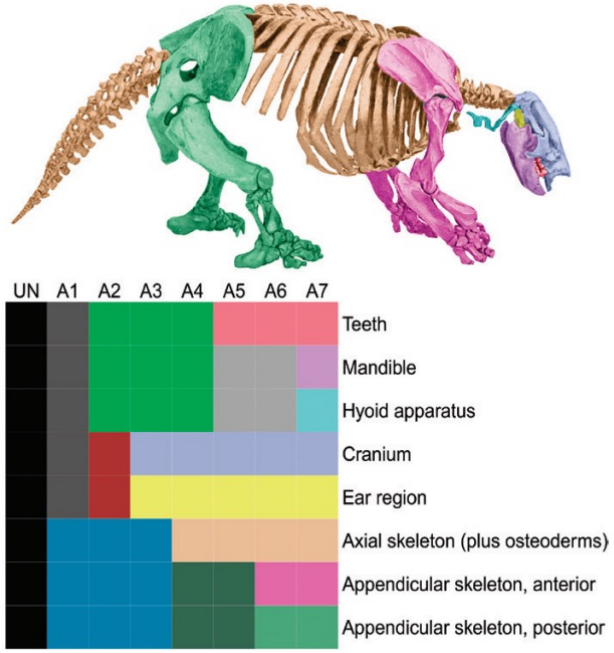
We can combine any of these extensions to make more complex models

0 ⟷ 1 ⟷ 2 ⟷ 3

We can have ordered characters

Anatomical partitioning schemes

Important to consider impact of combining extensions



Casali et al 2022 Zoological Journal of the Linnean Society

On the Mkv Model with Among-Character Rate Variation

Alessio Capobianco, Sebastian Höhna

doi: https://doi.org/10.1101/2024.11.15.623796

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract | Full Text | Info/History | Metrics | Preview PDF

Tomorrow session C

11:15 – 11:30 **Modelling among-character rate variation with the Mkv model: lessons learned and perspectives for morphological phylogenetics**
Alessio Capobianco, Sebastian Höhna

used in likelihood-based morphological phylogenetics often adapt molecular phylogenetics models to the specificities of morphological data. Such is the case for the widely

Choosing a model

Model Selection: Take a bunch of different models and test which is the *best*

Gives the **relative** fit

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

Gives the **absolute** fit

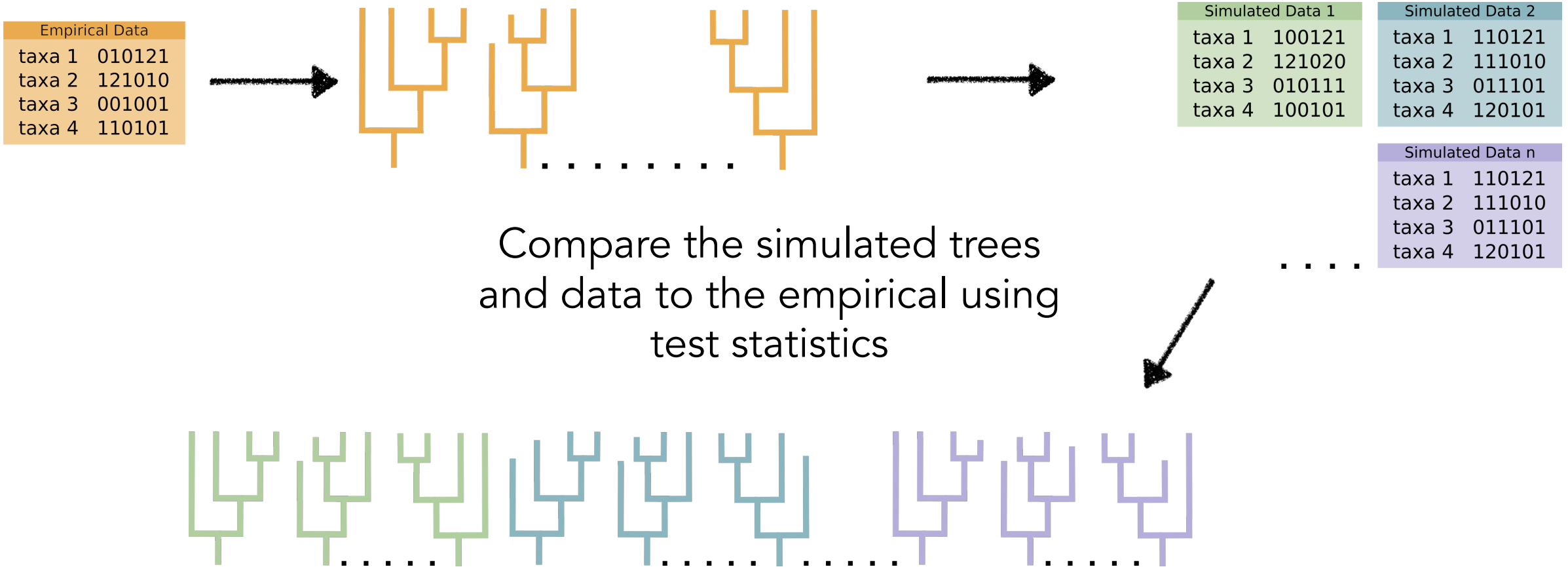
Model Adequacy

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

Gives the **absolute fit**

One approach is **Posterior Predictive Simulations**

Model Adequacy: 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:

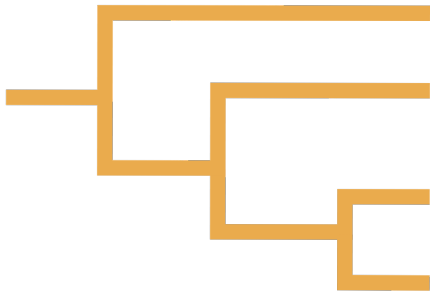
Data-based, inference-based, mixed

Test Statistics

Calculating **consistency index**

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

MCC summary tree



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

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

. . . .

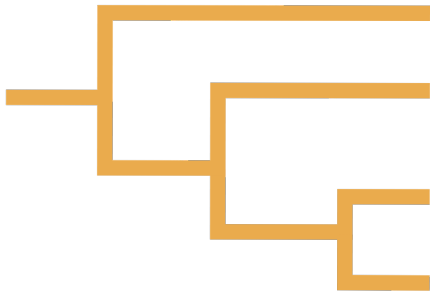
Calculate **a range (500) values** using all simulated data sets

Test Statistics

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

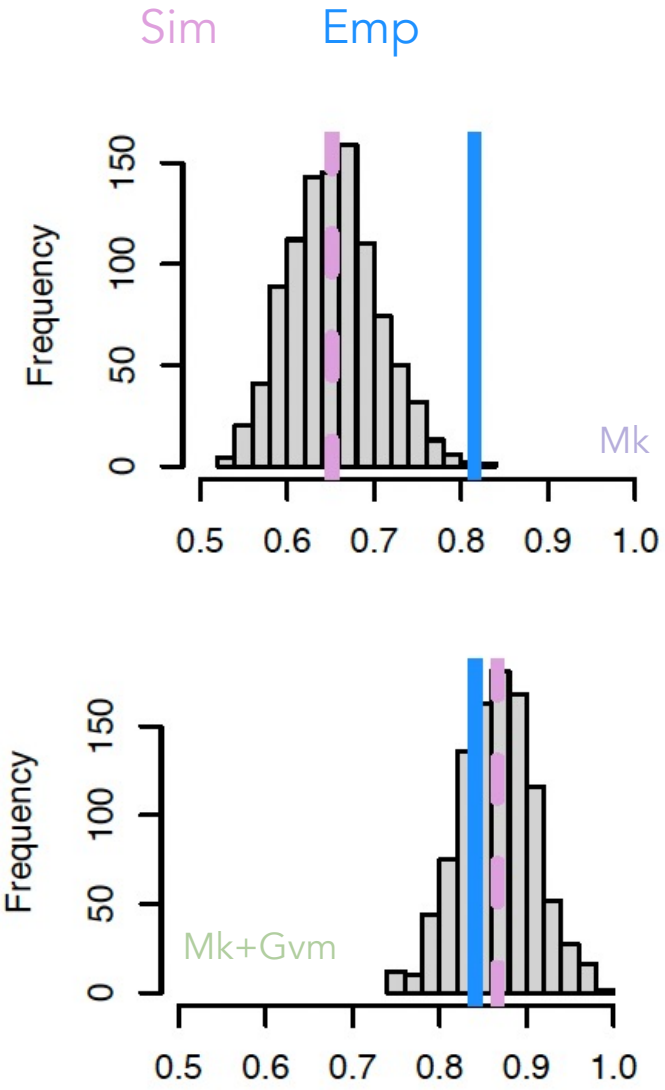
. . . .

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

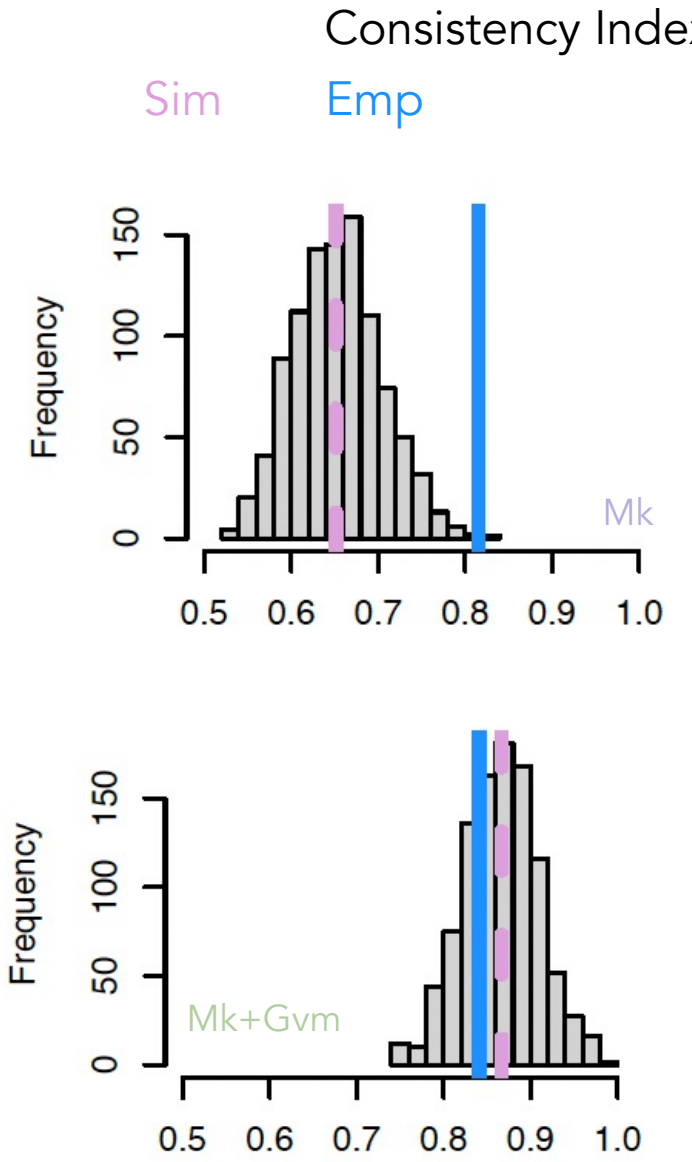
Are these values significantly different from each other?

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

Consistency Index



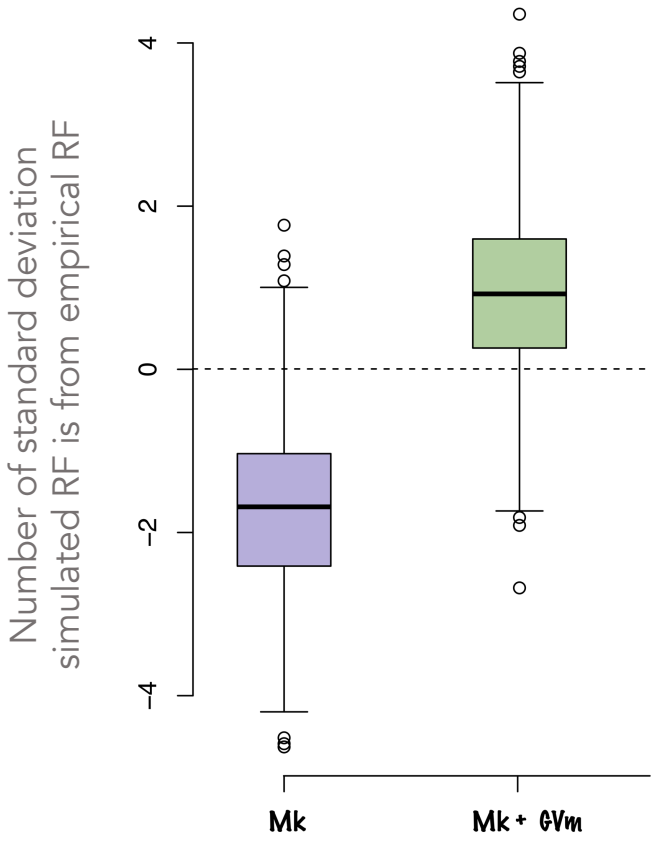
Histograms showing the range of RF values for all the simulated data

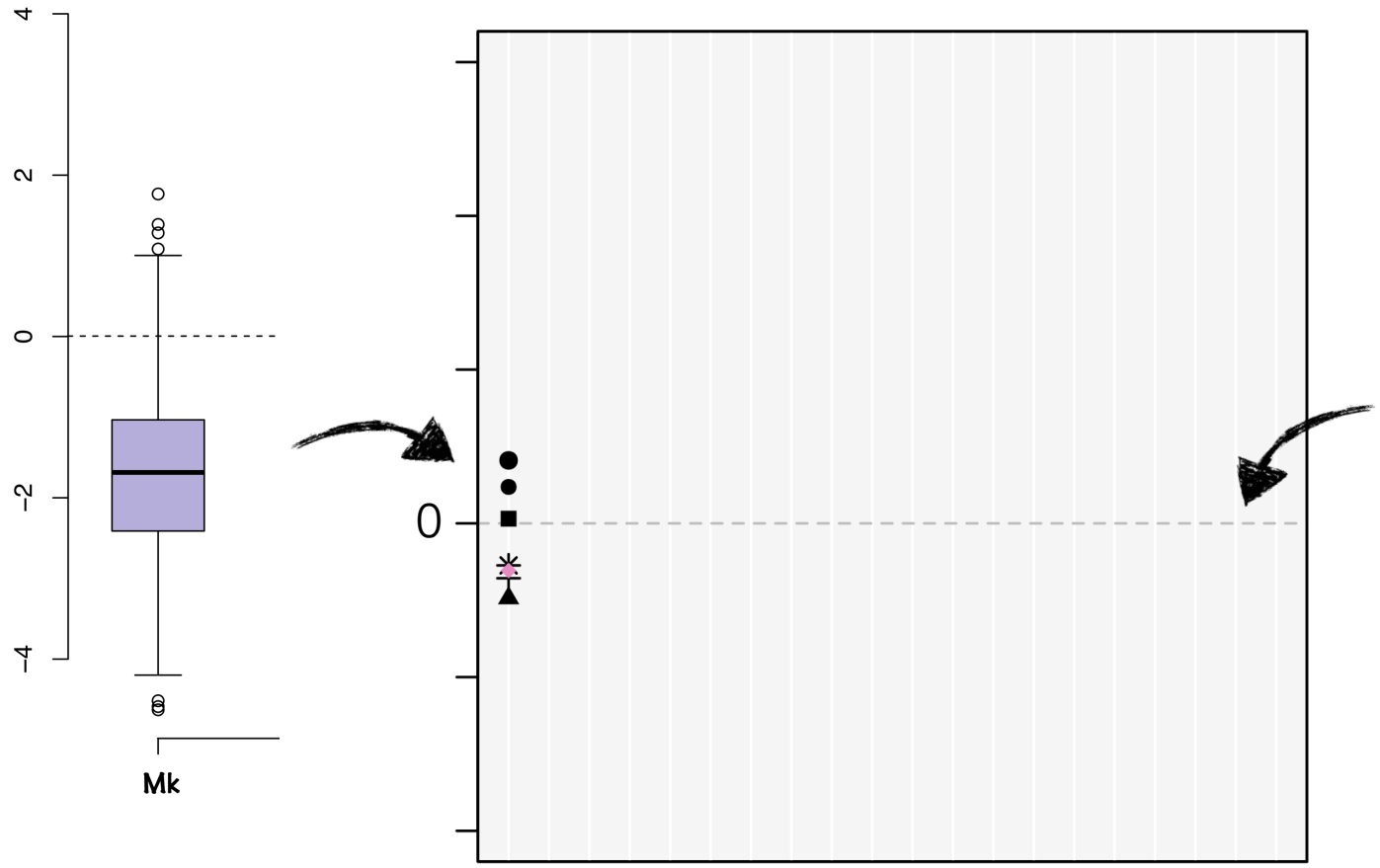


Histograms showing the range of RF values for all the simulated data

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

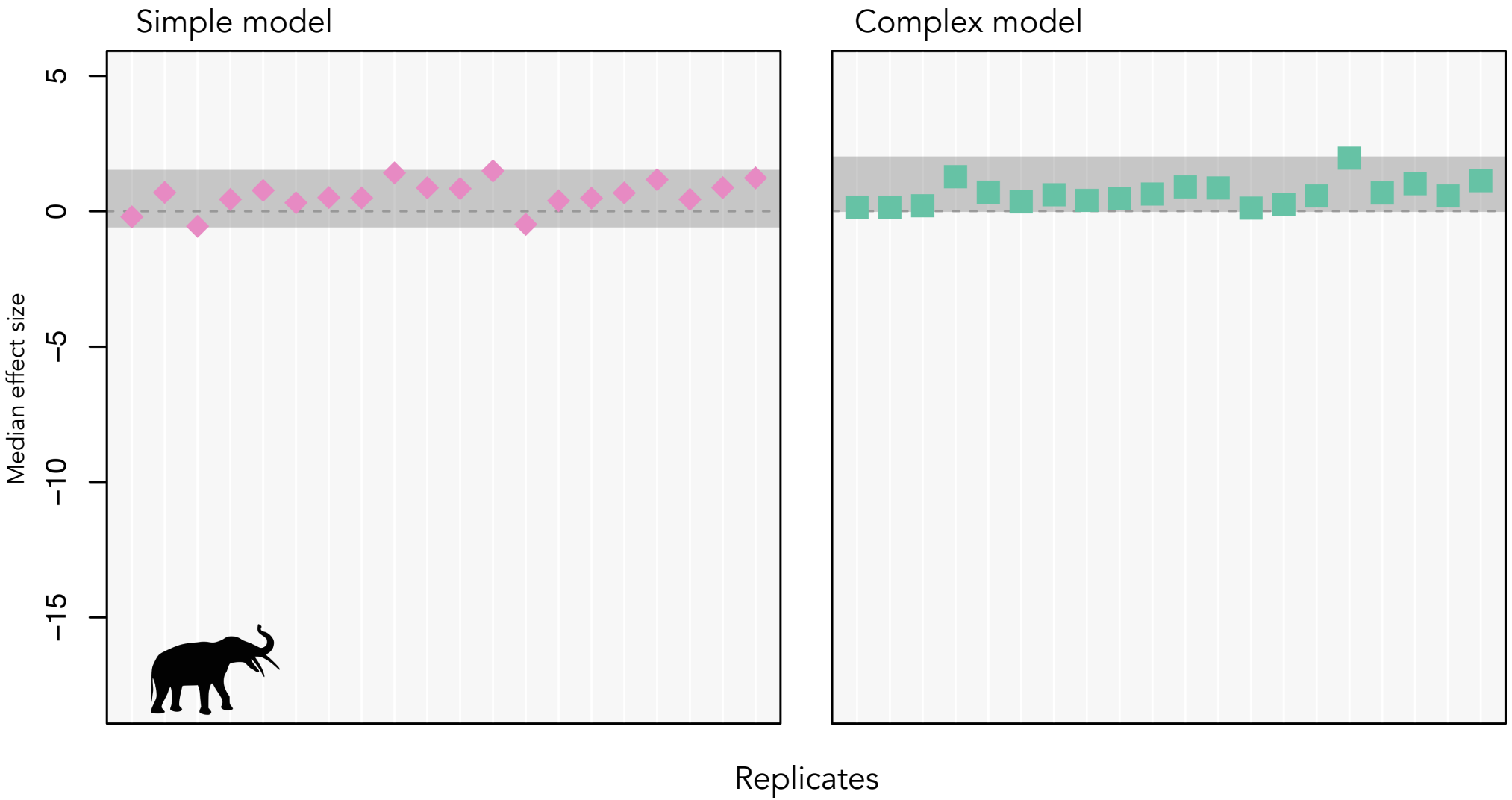
$$\frac{\text{Empirical TS - SimTs}}{\text{Sd(All Sim TS)}}$$



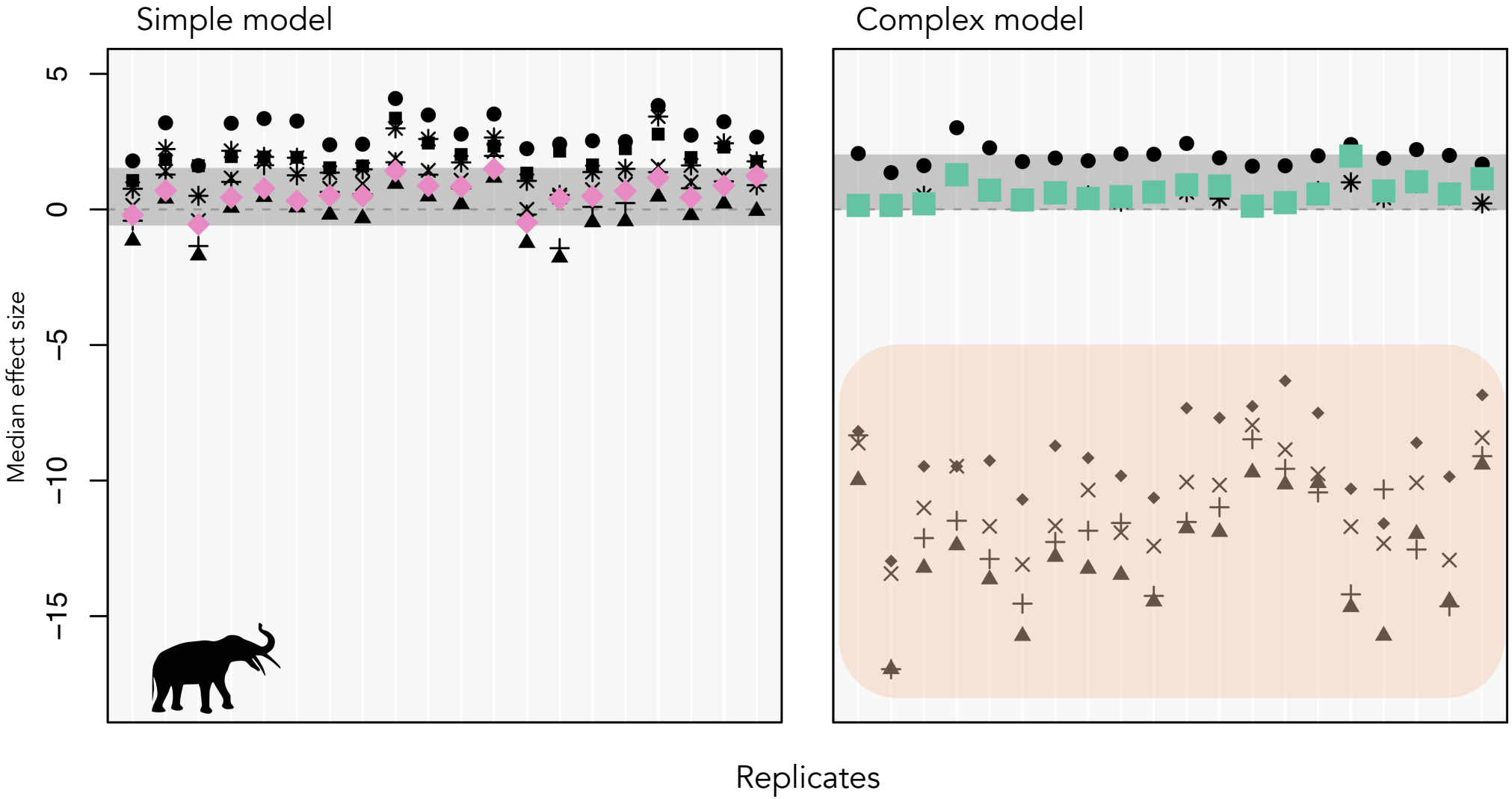


Closer to zero the better the model – no different between the input data and simulated data

Consistency Index



Consistency Index

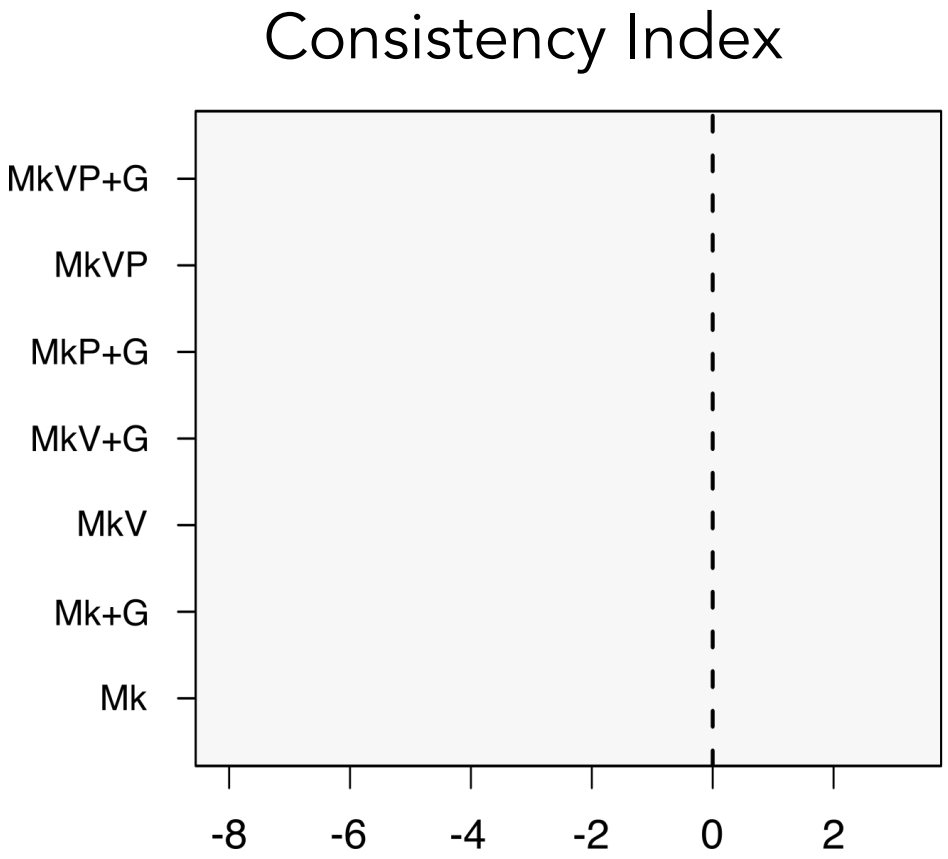


Are current morphological models adequate for empirical data sets?

Are current morphological models adequate for empirical data sets?

12 taxa
51 chars
4 states

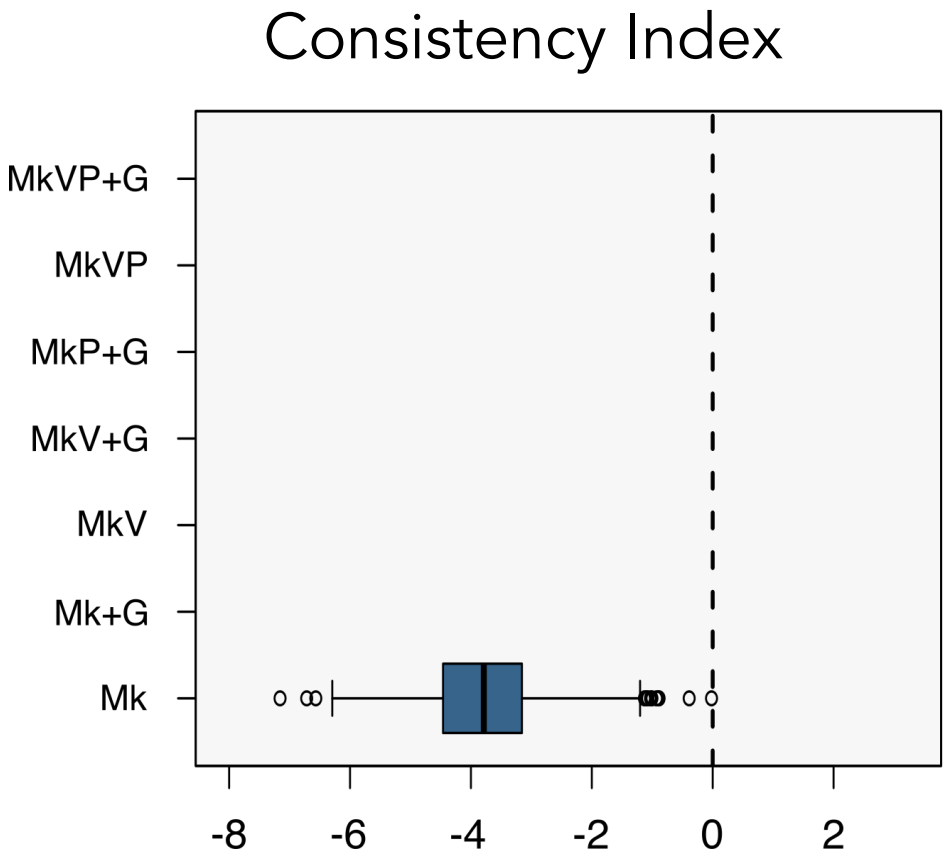
Agnolin et al 2007



Are current morphological models adequate for empirical data sets?

12 taxa
51 chars
4 states

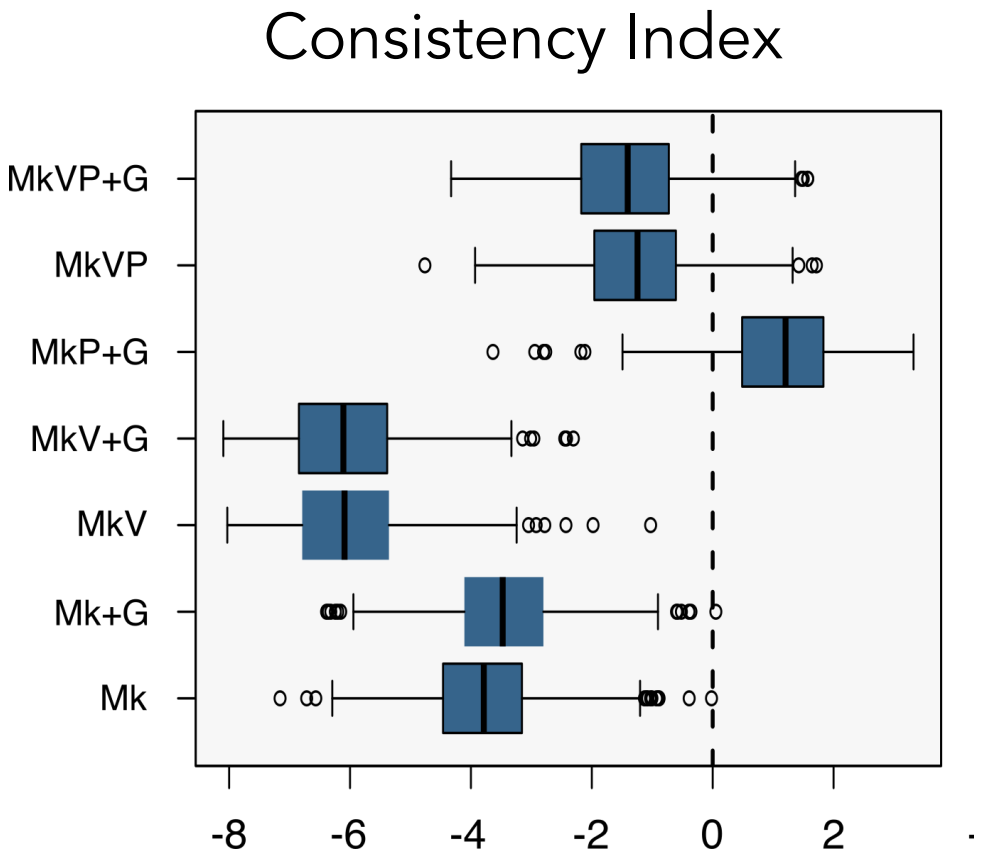
Agnolin et al 2007



Are current morphological models adequate for empirical data sets?

12 taxa
51 chars
4 states

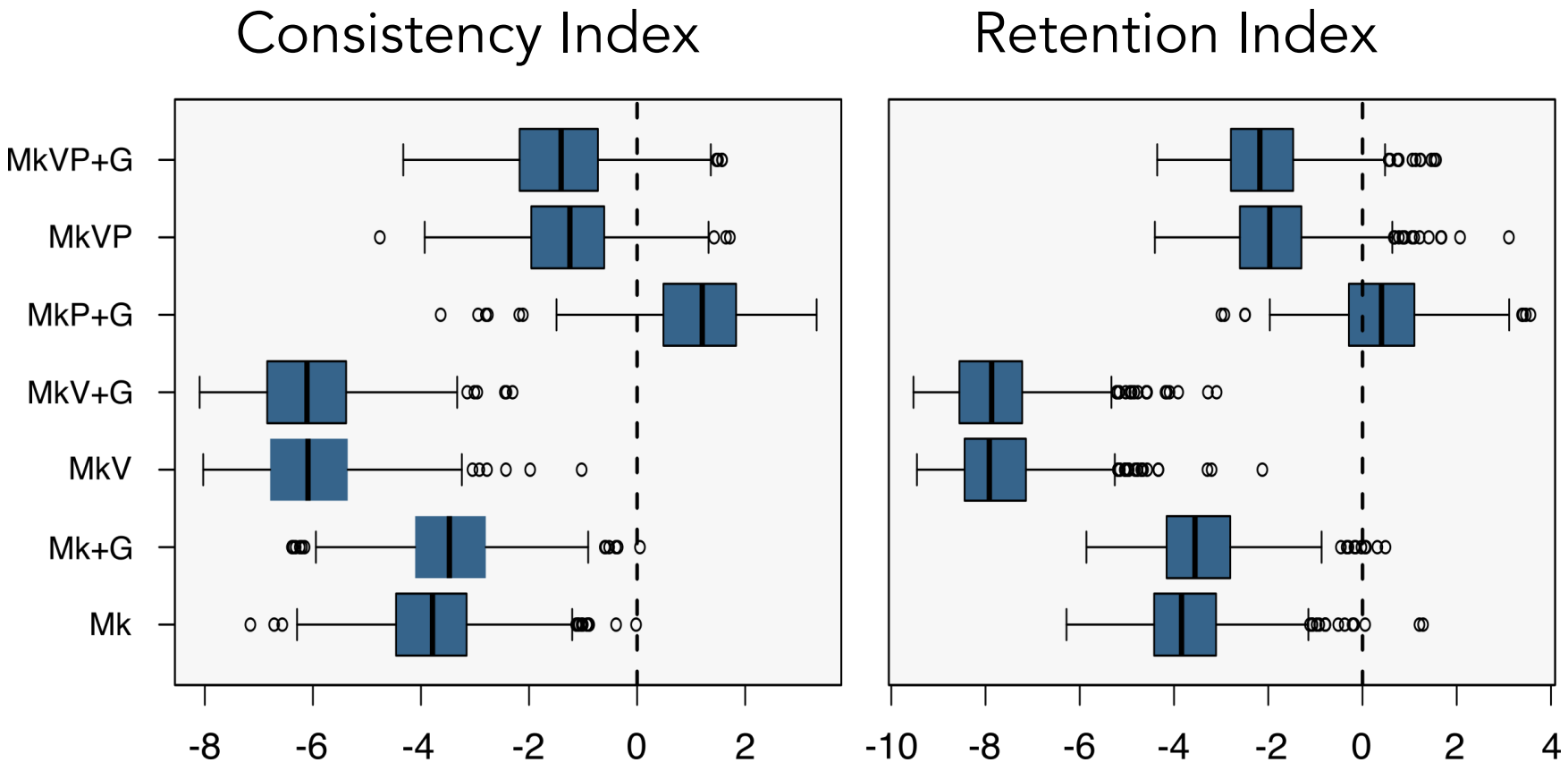
Agnolin et al 2007



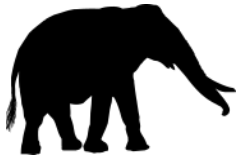
Are current morphological models adequate for empirical data sets?

12 taxa
51 chars
4 states

Agnolin et al 2007

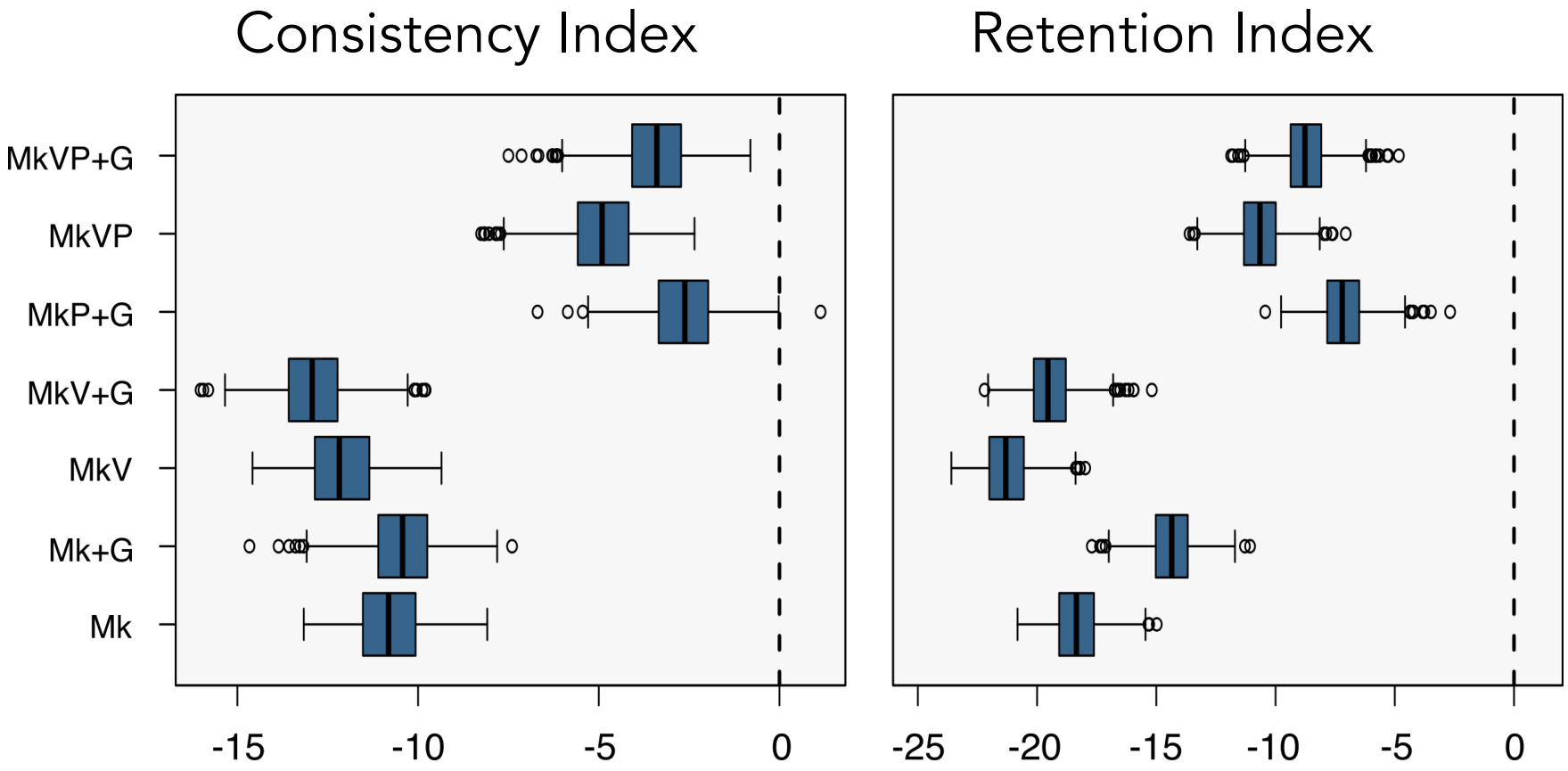


Are current morphological models adequate for empirical data sets?

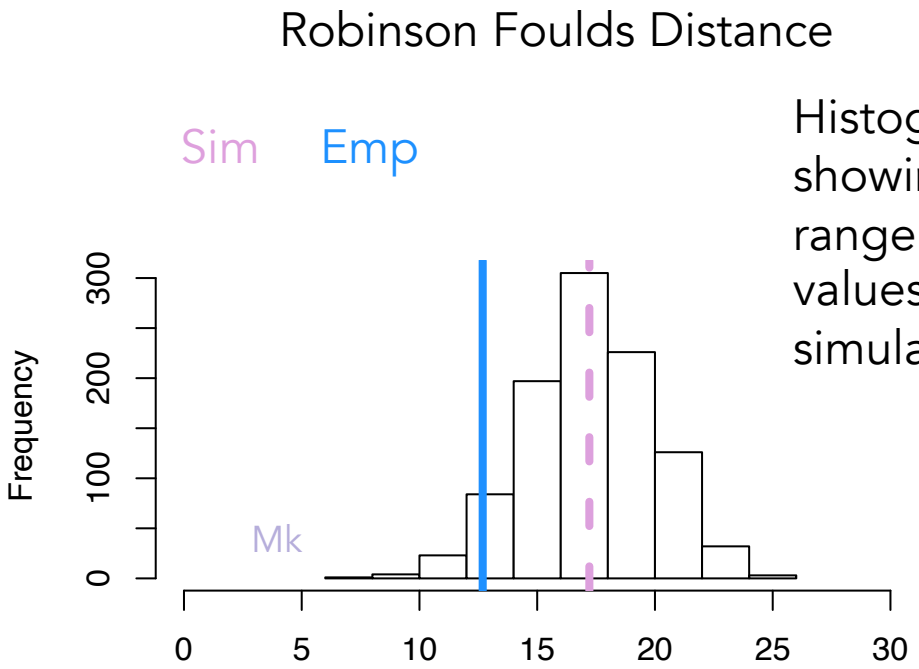


40 taxa
125 chars
6 states

Shoshani et al 2006

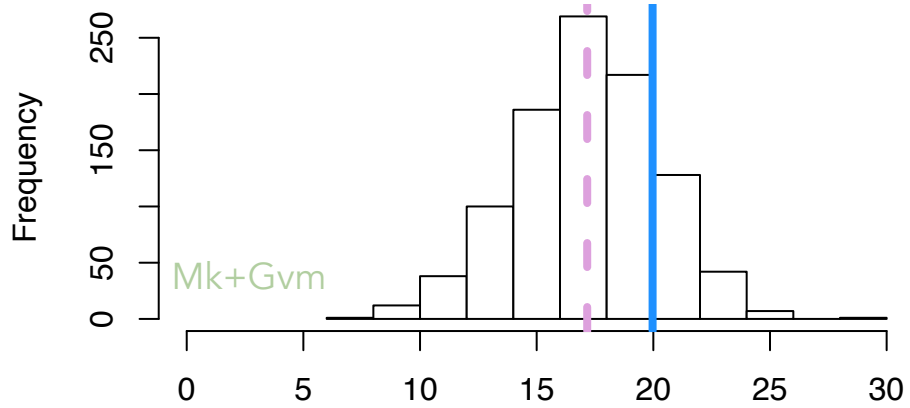


Posterior Predictive P-values



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

Are these values significantly different from each other?



We will also calculate the P-values in R (look at the midpoint value)

Model adequacy exercise