Measurement error, identifiability, and model adequacy

95% CI
$\pm$ <0.01
$\pm$ 0.281
$\pm \infty$

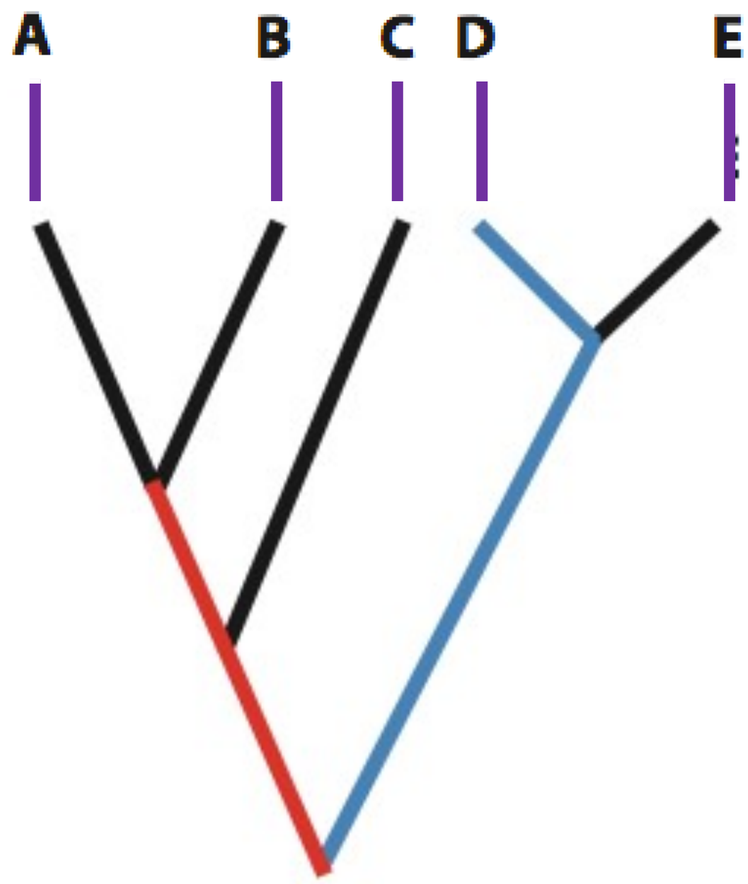
Brian O'Meara
http://www.brianomeara.info
@omearabrian

Learning objectives

- Be afraid
- Be very afraid
- Figure out how to deal with this
- Also, measurement error. Let's start with that….

Measurement error

- Darwin's great insight was that variation is widespread in nature, even within species
- In comparative methods…. we're getting there
- There is variation within species
- There is also uncertainty in measurements for a given individual
    - Giant squid length, tree height, primate socialization frequency, leaf nitrogen, etc.



Variance-covariance matrix

	A	B	C	D	E
A	var(A)	cov(A,B)	cov(A,C)	cov(A,D)	cov(A,E)
B	cov(B,A)	var(B)	cov(B,C)	cov(B,D)	cov(B,E)
C	cov(C,A)	cov(C,B)	var(C)	cov(C,D)	cov(C,E)
D	cov(D,A)	cov(D,B)	cov(D,C)	var(D)	cov(D,E)
E	cov(E,A)	cov(E,B)	cov(E,C)	cov(E,D)	var(E)

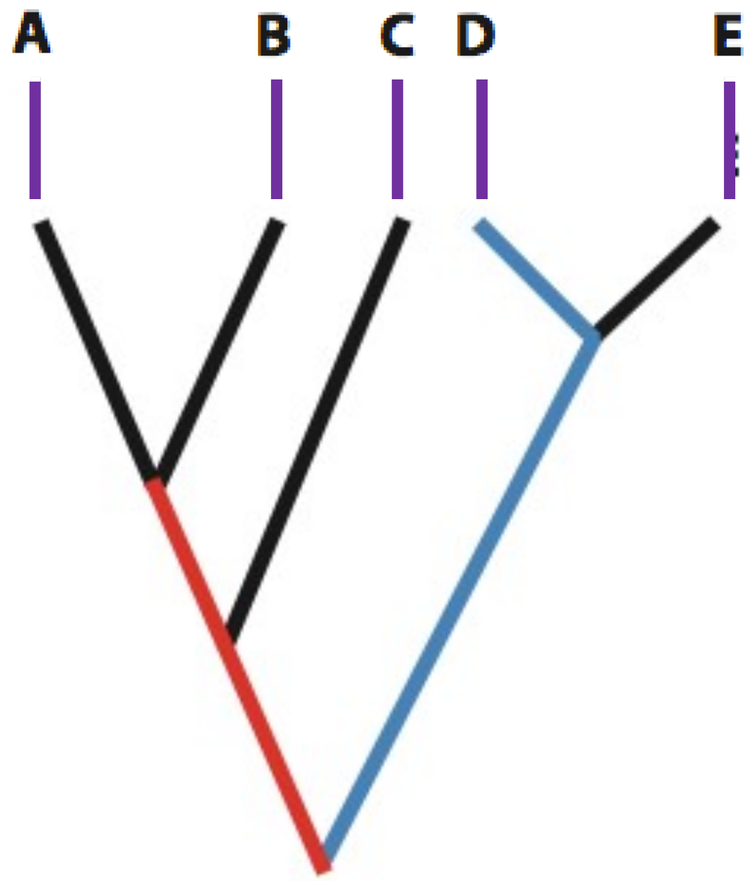
Mean vector

A	μ_A
B	μ_B
C	μ_C
D	μ_D
E	μ_E

Figure 4

Multivariate normal distribution. The figure shows a tree, the tree's variance-covariance matrix, and the vector of means (which, under Brownian motion, would equal the root state). Highlighted are the branches leading to covariance between taxa A and B (*red*) and the branches leading to variance in D (*blue*).

O'Meara 2012



Variance-covariance matrix

	A	B	C	D	E
A	var(A) + error	cov(A,B)	cov(A,C)	cov(A,D)	cov(A,E)
B	cov(B,A)	var(B) + error	cov(B,C)	cov(B,D)	cov(B,E)
C	cov(C,A)	cov(C,B)	var(C) + error	cov(C,D)	cov(C,E)
D	cov(D,A)	cov(D,B)	cov(D,C)	var(D) + error	cov(D,E)
E	cov(E,A)	cov(E,B)	cov(E,C)	cov(E,D)	var(E) + error

Mean vector

A	μ_A
B	μ_B
C	μ_C
D	μ_D
E	μ_E

Figure 4

Multivariate normal distribution. The figure shows a tree, the tree's variance-covariance matrix, and the vector of means (which, under Brownian motion, would equal the root state). Highlighted are the branches leading to covariance between taxa A and B (*red*) and the branches leading to variance in D (*blue*).

O'Meara 2012

Does this matter? Exercise interlude…

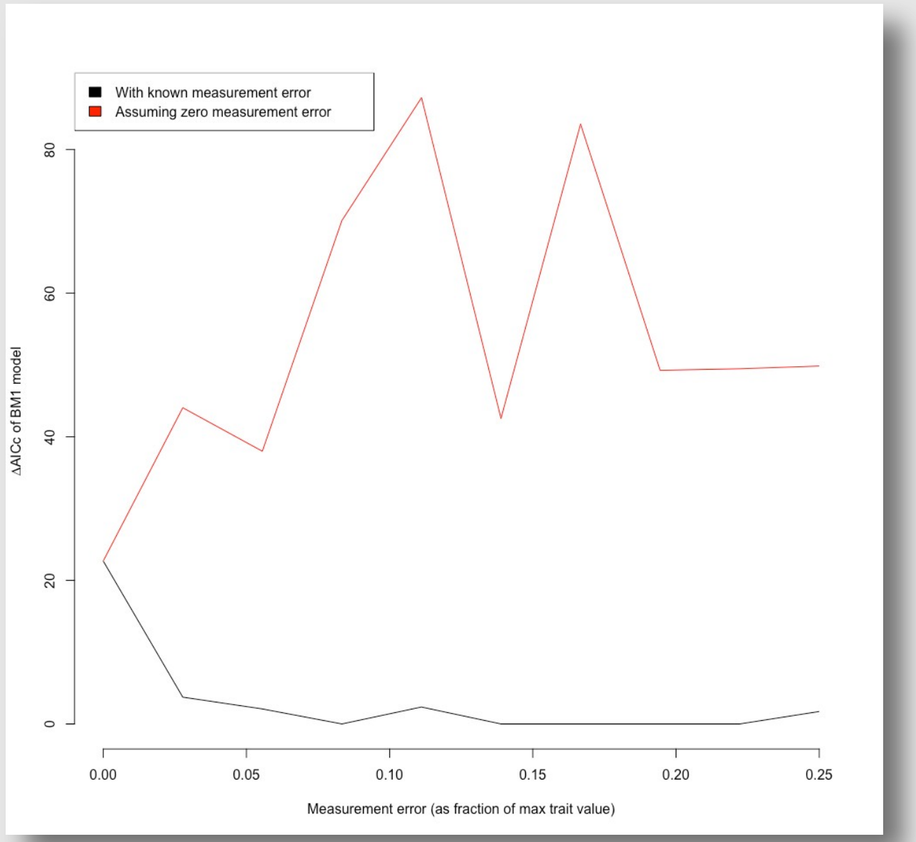
```r
rm(list=ls()) # because by this point in the course your workspace is probably polluted
library(OUwie)
library(plyr)

# First set up the model parameters
data(sim.ex)
alpha=rep(2,2)
sigma.sq=rep(.45, 2)
theta0=10
theta=rep(10,2)

# Now make the data
sim.data<-OUwie.sim(tree,trait,simmap.tree=FALSE,scaleHeight=FALSE,
                    alpha=alpha,sigma.sq=sigma.sq,theta0=theta0,theta=theta)

# compute all combinations
mserr.vector <- seq(from=0, to=.25*max(sim.data$X), length.out=10)
models.vector <- c("BM1", "OU1")
mserr.argument.vector <- c("none", "known")

# combine all conditions
```

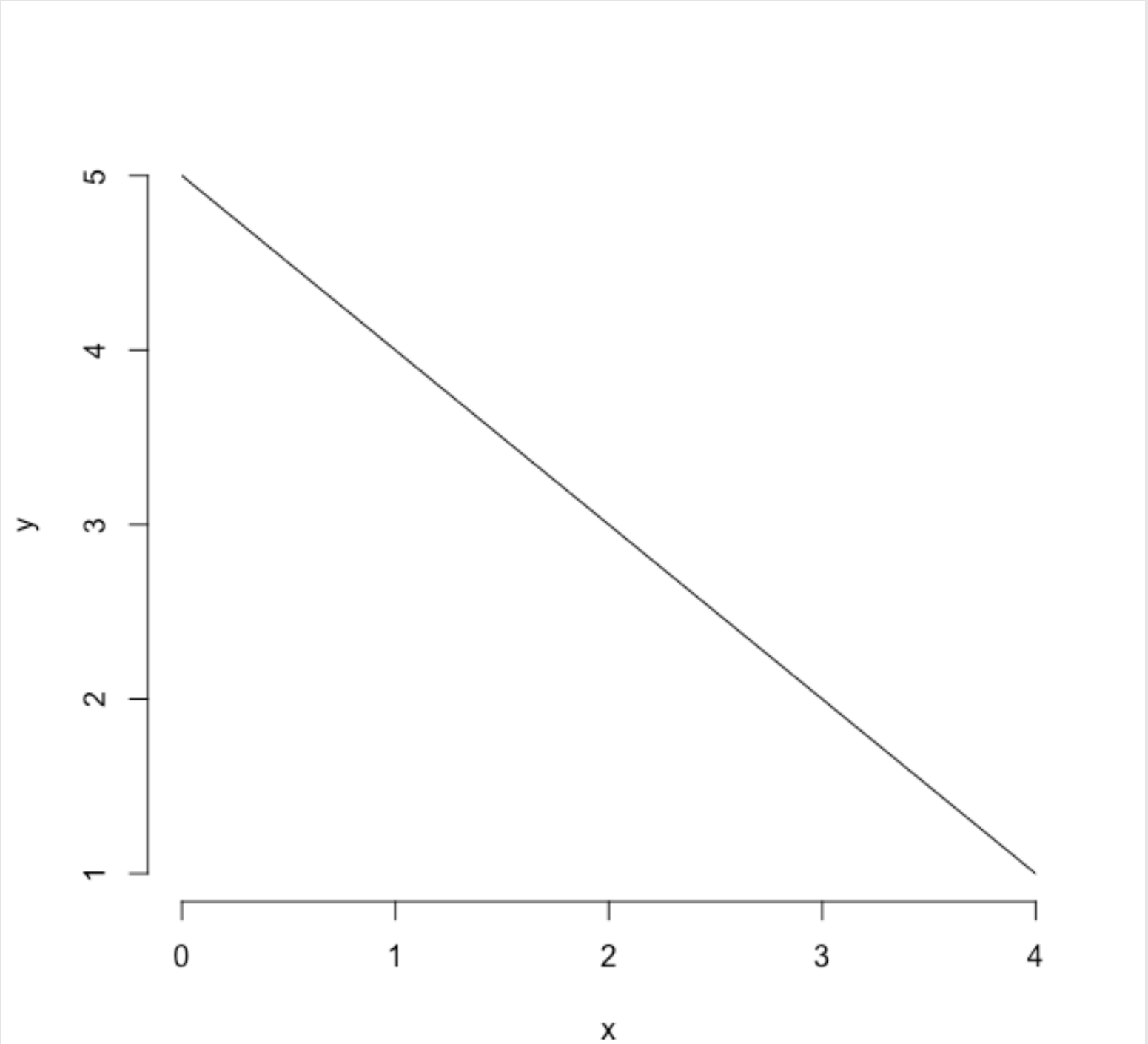


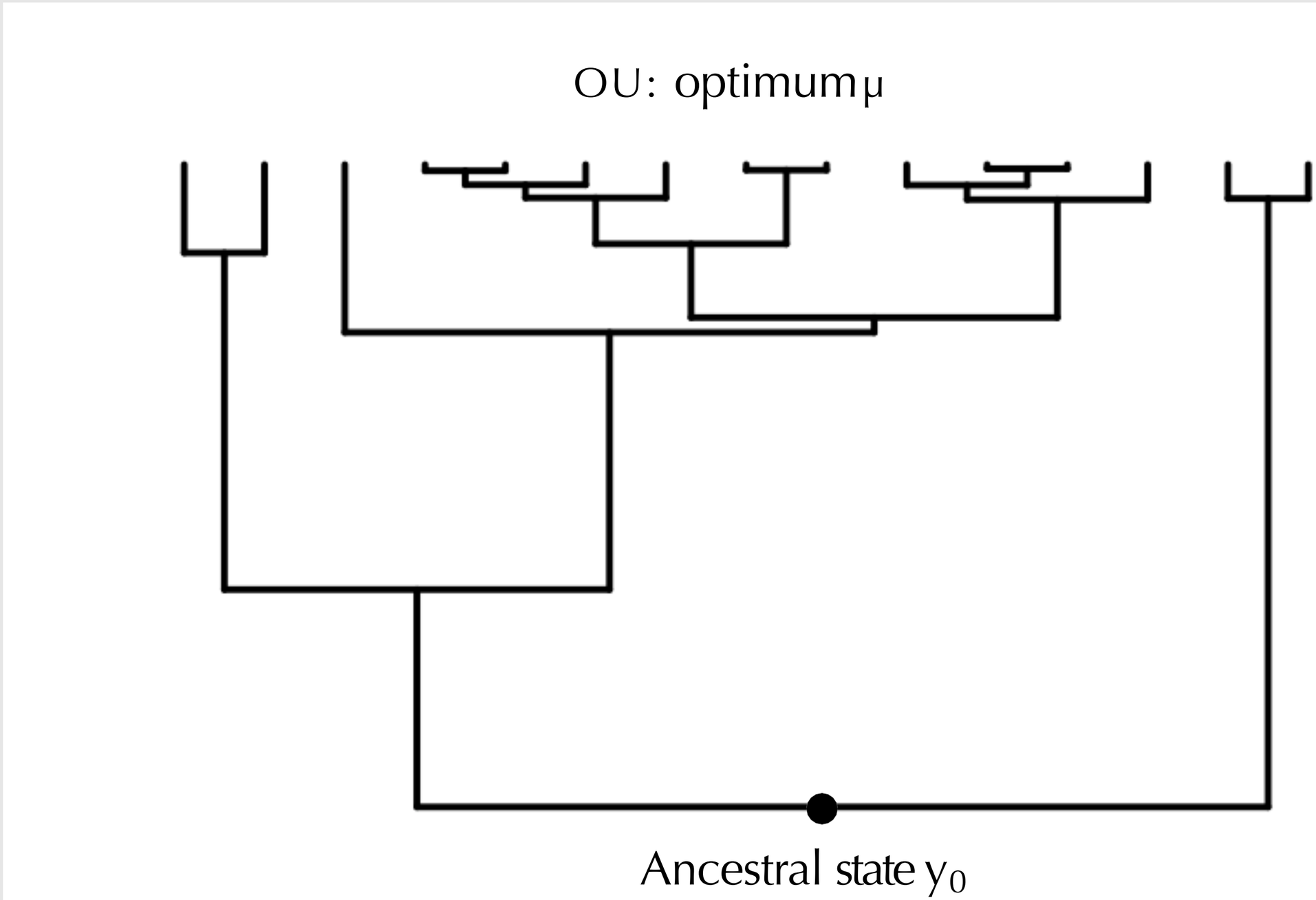
Identifiability

$$x + y = 5$$

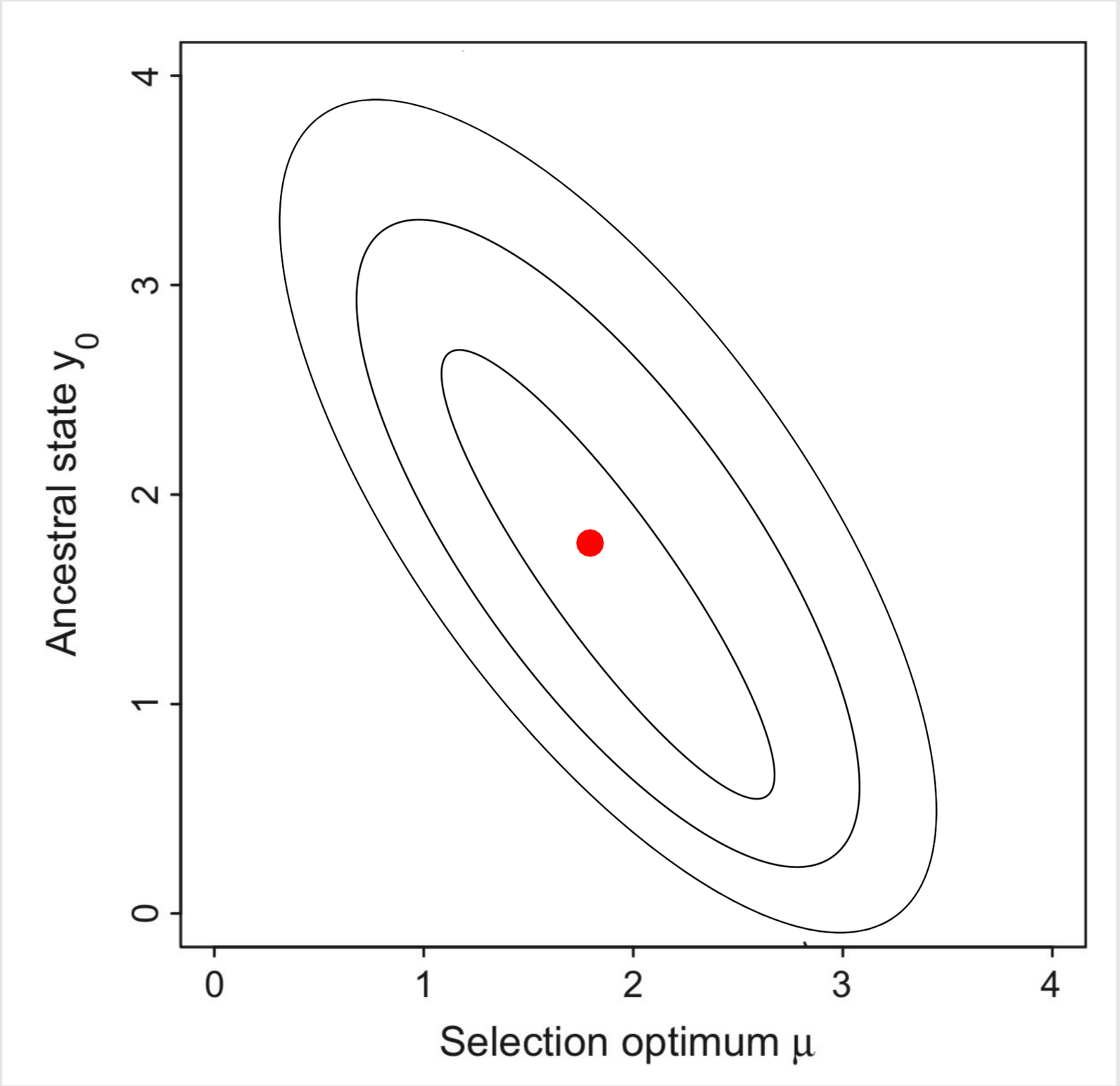
What is x?

$$x + y = 5$$

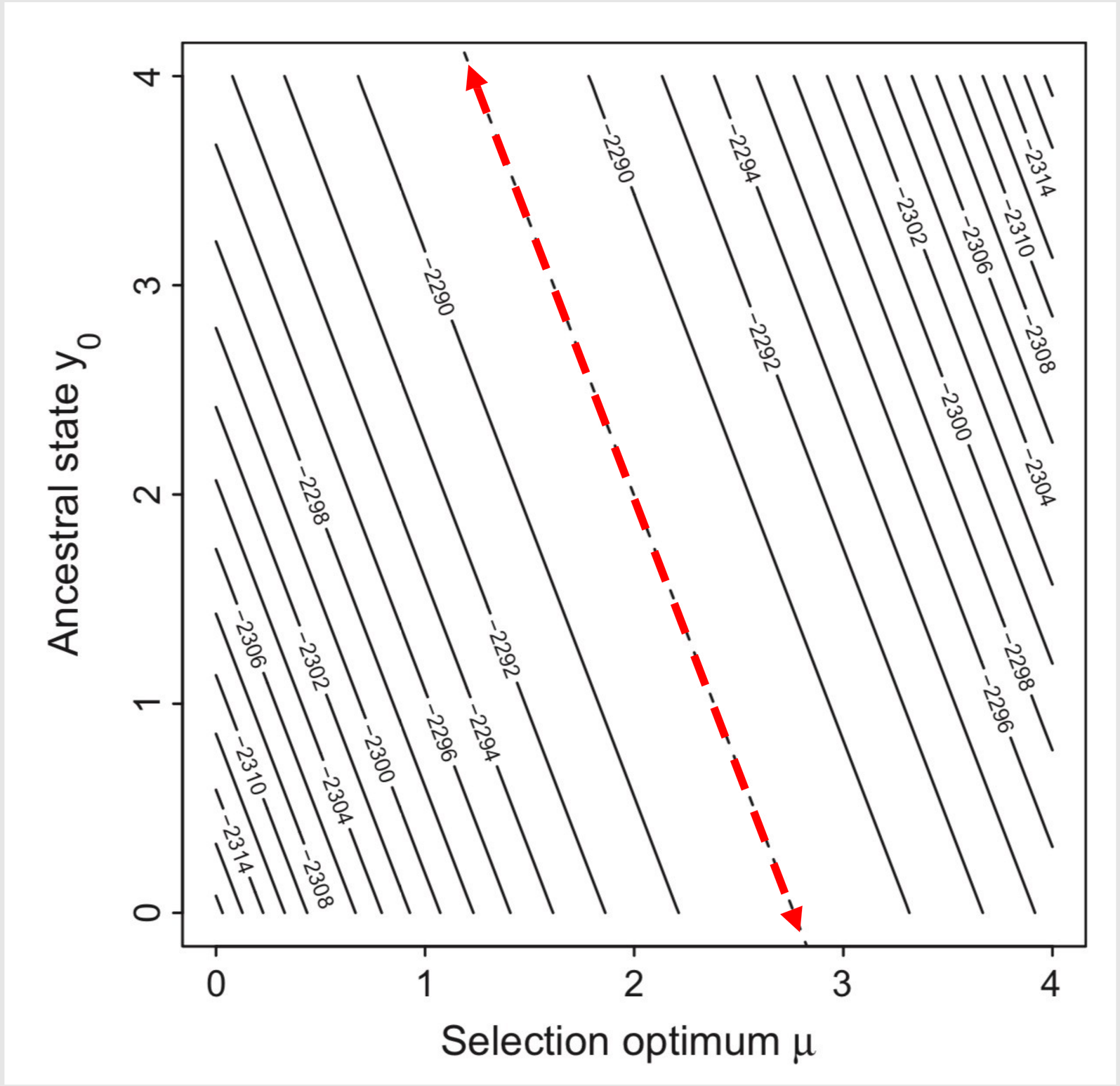



OU: optimum μ
Ancestral state y_0

This is what we want



This is what we get



Ho and Ané, 2014

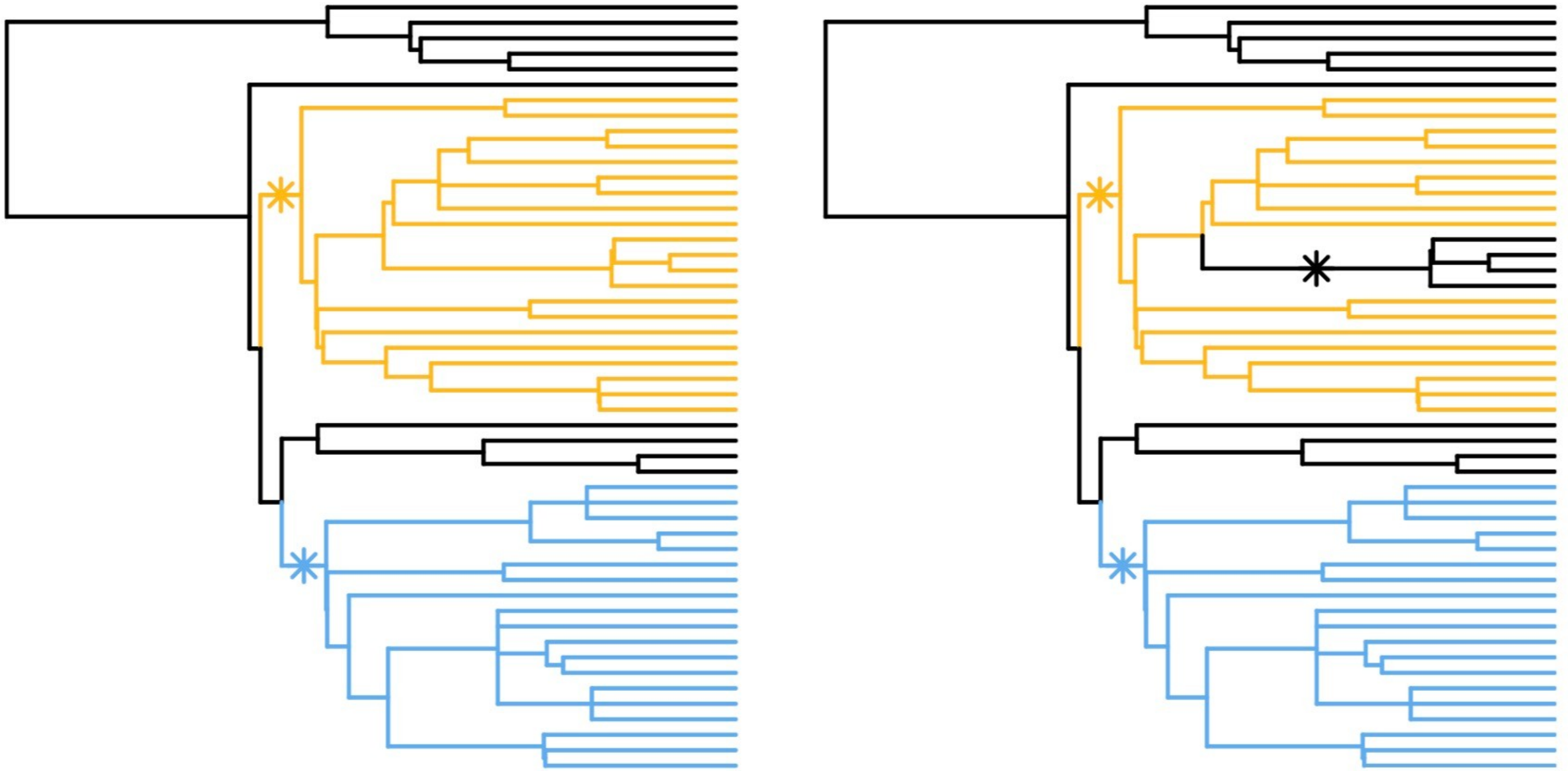


Fig. 2. Edges are 'painted' according to their selection regime, with one optimum μ_ℓ for each colour. Unidentifiability case (left): every selection regime forms a connected component. Identifiability case (right): one regime (black) covers two disconnected parts in the tree.

Ho and Ané, 2014

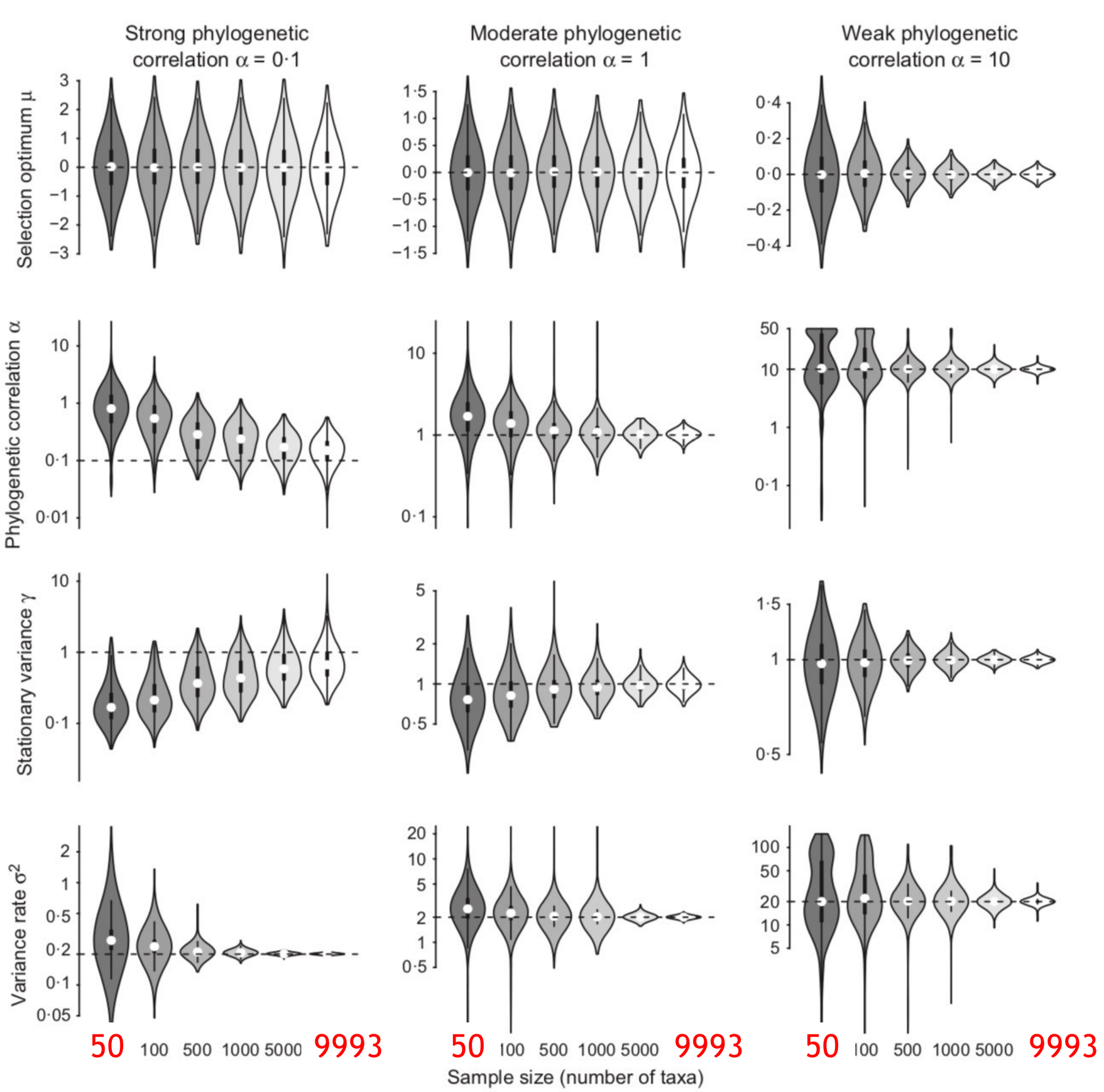


Fig. 3. Violin plots showing the distribution of the MLEs of μ, γ, α and $\sigma^2 = 2\alpha\gamma$ on subtrees from the 9993-species bird phylogeny (Jetz et al. 2012 with 2000 simulations at each sample size. Each column corresponds to one set of true parameter values: μ = 0; γ = 1; α = 0·1 (left, $t_{1/2}$ = 6· compared to the tree height T = 1), 1 (middle, $t_{1/2}$ = 0·69), 10 (right, $t_{1/2}$ = 0·069) and $\sigma^2 = 2\alpha\gamma$.

Ho and Ané, 2014

Table 3. Parameter estimates and their associated 95% confidence interval (CI) for the OU_{MVA} model, the model that best fit the genome size data. Each CI was obtained by multiplying each approximate standard error by the critical value in the *t*-distribution where the cumulative probability is equal to 0.975 (i.e., $t(0.975, \infty) = 1.96$).

	Herb		Woody	
	Estimate	95% CI	Estimate	95% CI
α	3.85	$\pm$ 0.955	<0.001	$\pm$ <0.01
σ^2	2.51	$\pm$ 0.376	0.531	$\pm$ 0.281
θ	0.618	$\pm$ 0.143	<0.001	$\pm \infty$

Beaulieu, Jhwueng,
Boettiger, O'Meara 2012

Picking models

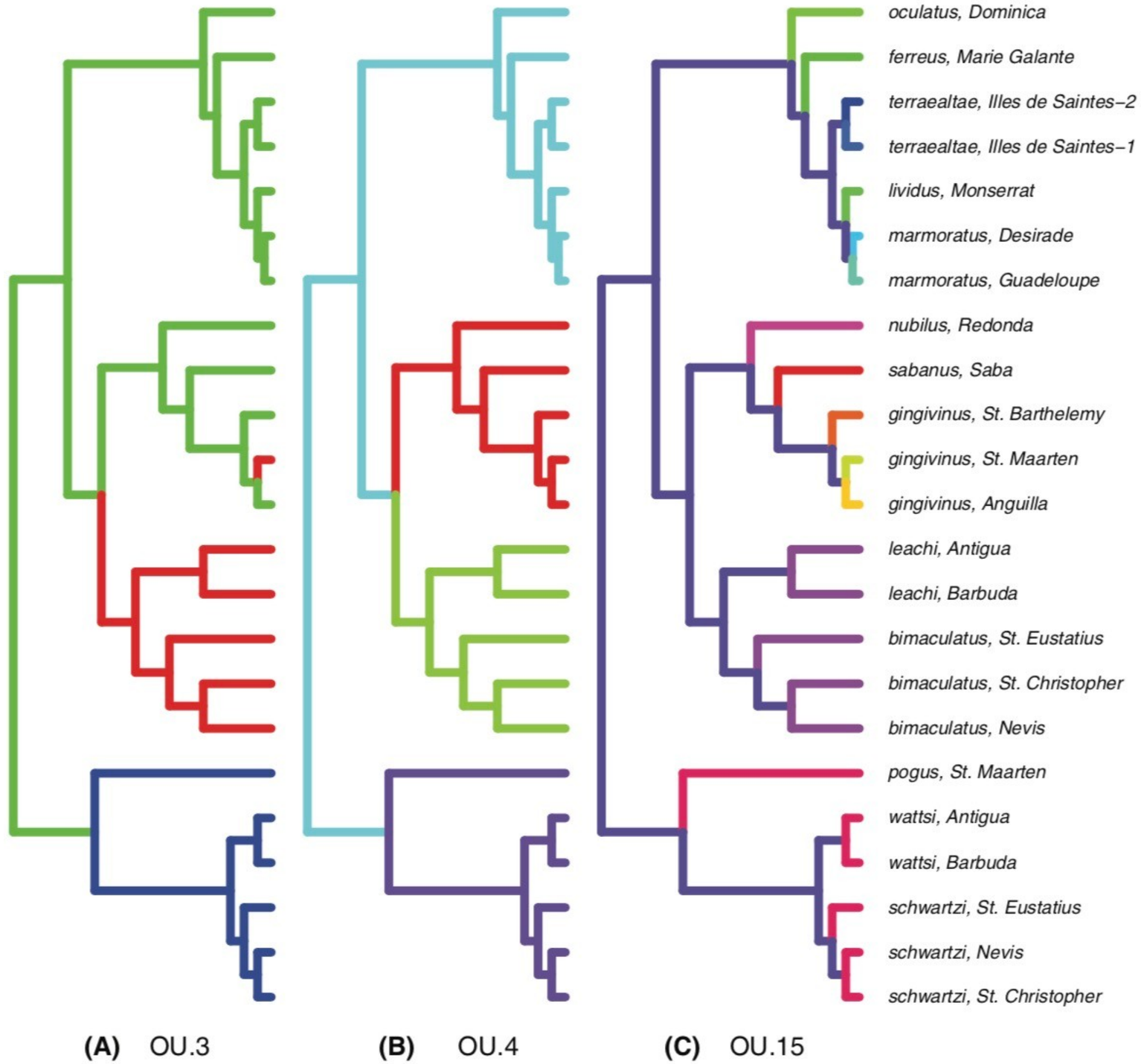


Figure 3. "Paintings" of the *Anolis* phylogeny specifying which branches are assumed to have a common value of the trait optimum θ for three different models: (A) OU.3, with three possible optima from Butler and King (2004); (B) OU.4, with four possible optima; and (C) OU.15, with a unique optimum for each branch in the upper two clades.The remaining models, BM and OU.1, fit the same parameters across the entire phylogeny and so are not shown . Estimated model parameters for each are shown in Table 1.

Boettiger, Coop, & Ralph 2012

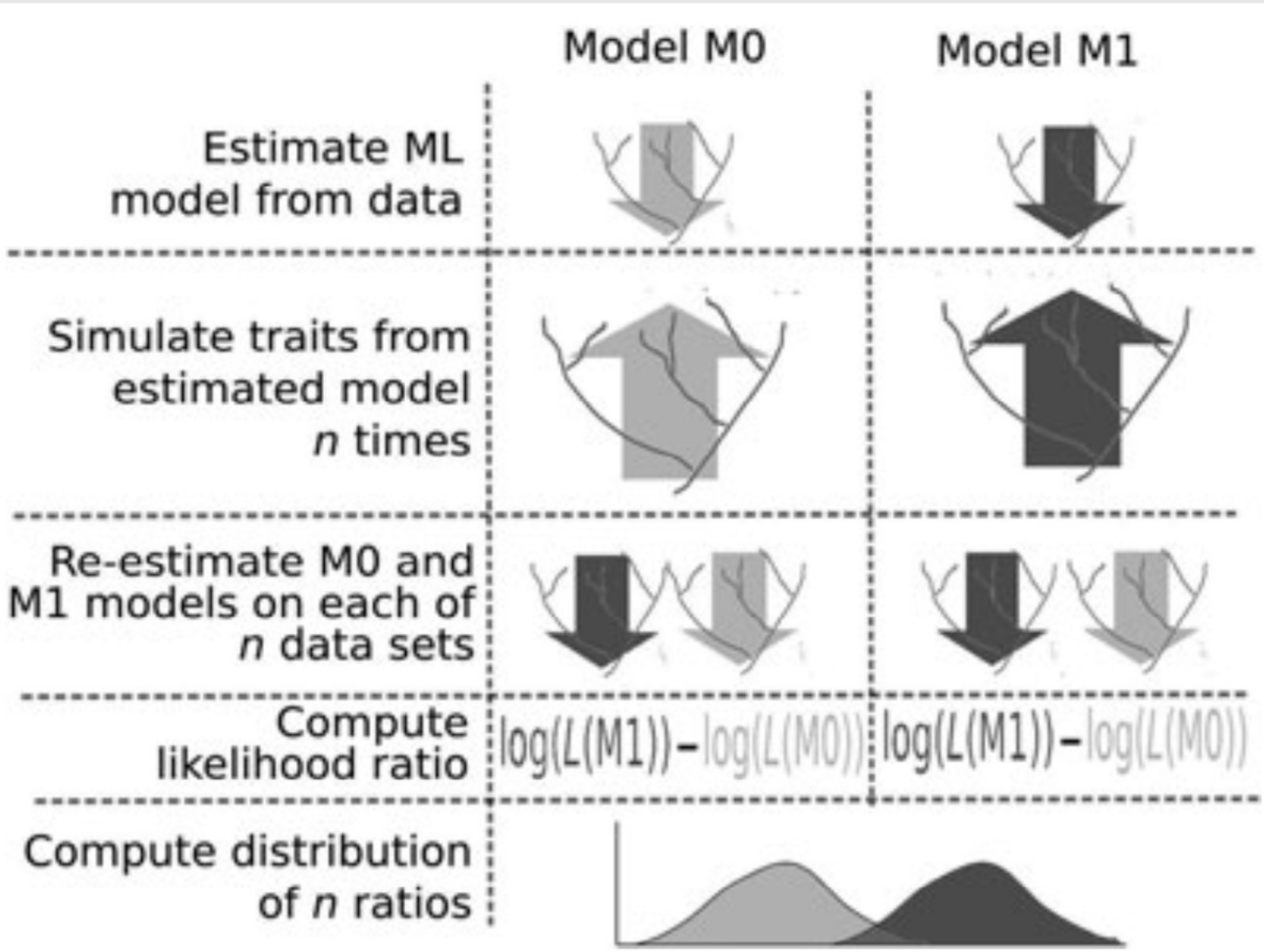


Figure 2. Conceptual diagram of the Monte Carlo method for model choice. First, parameters for both models are estimated from the original data . Then, *n* simulated datasets are created from each model at these parameters, and on each dataset, the parameters for both models are reestimated and the likelihood ratio statistic is computed. The collection of likelihood ratio statistics generates the corresponding distribution. This involves a process of 4*n* fits by maximum likelihood, instead of only two fits required for information criteria.

Boettiger, Coop, & Ralph 2012

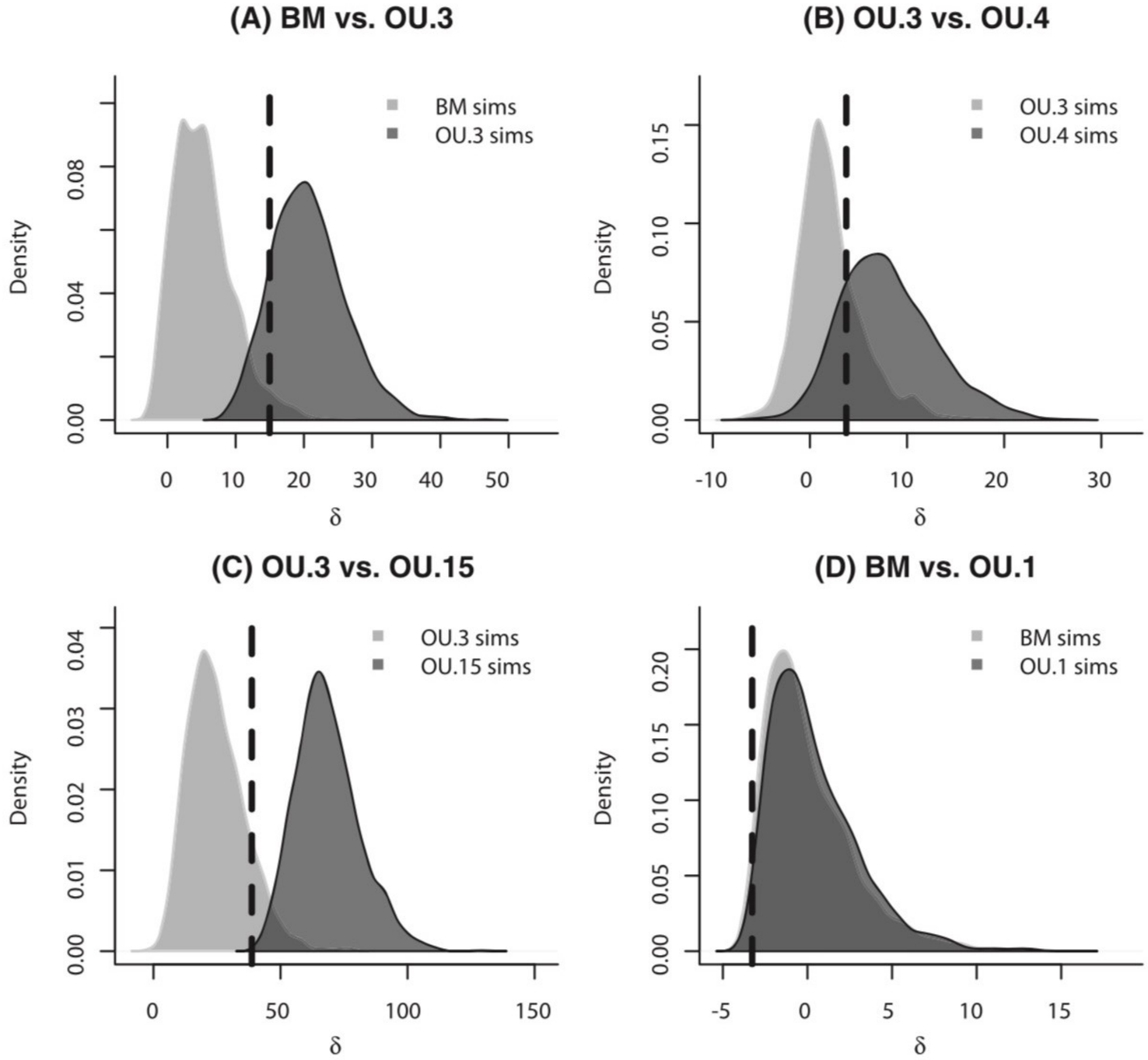
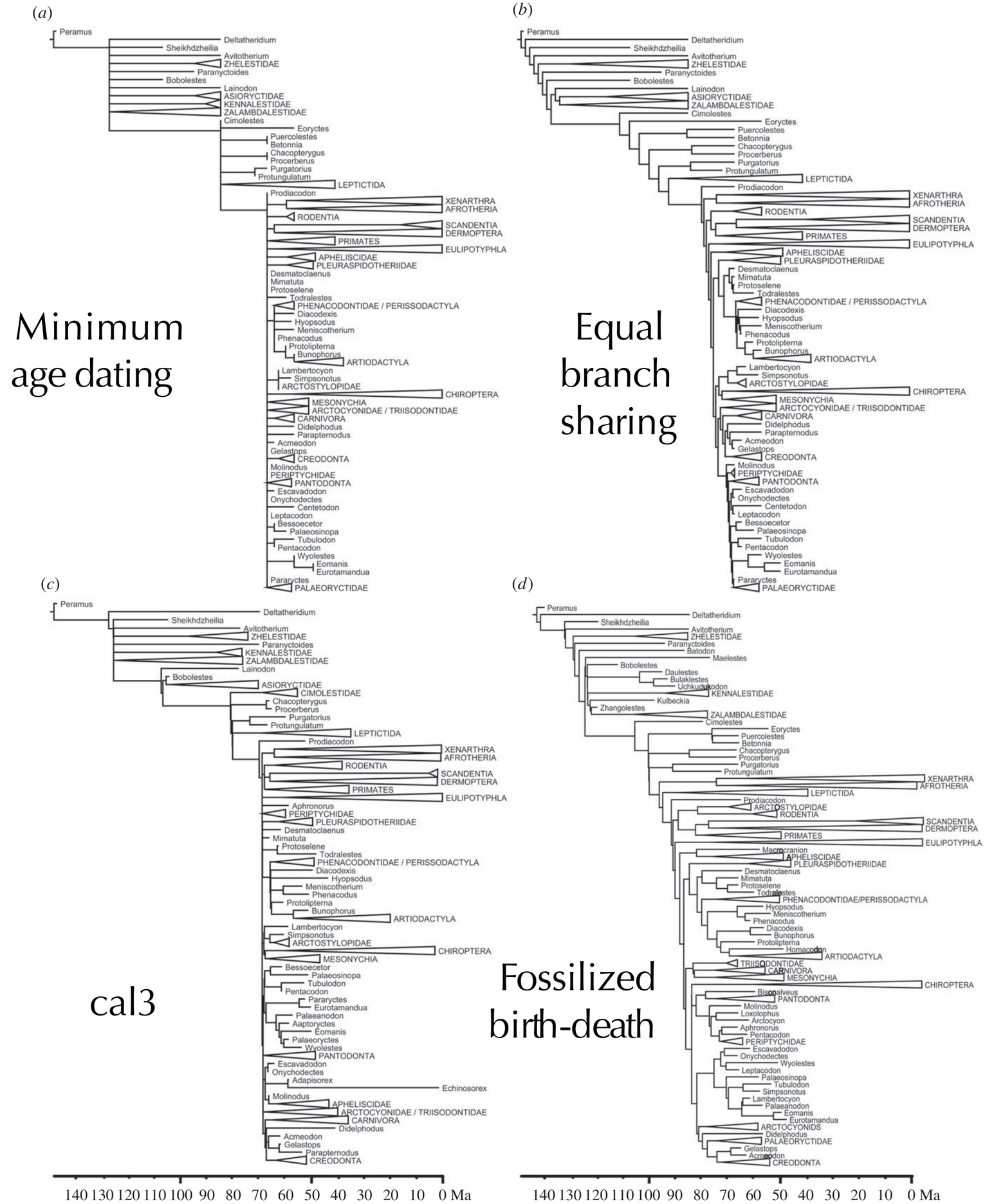
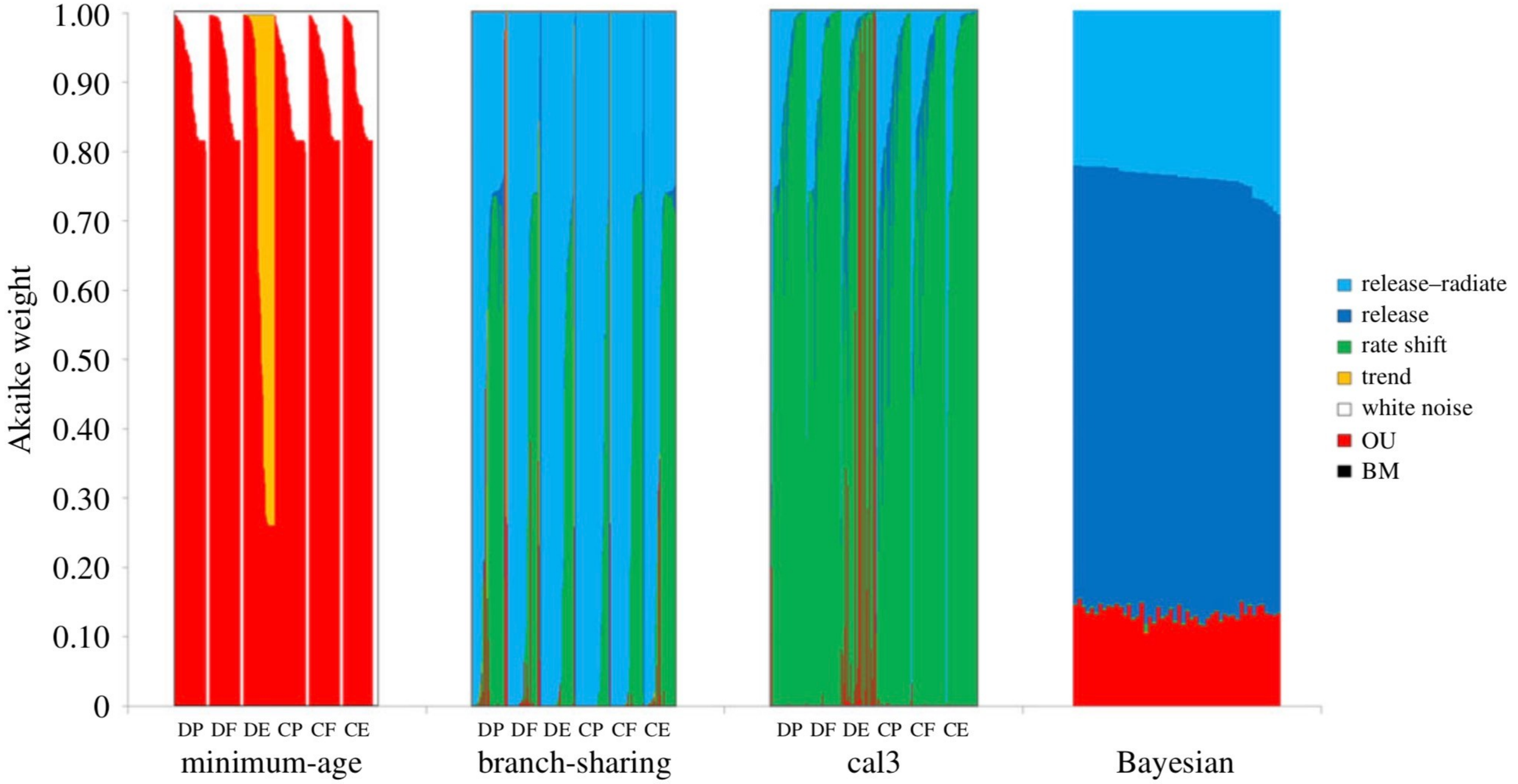


Figure 4. Distributions of the likelihood ratio statistic of equation (3) for four different model comparisons. In each case, the lighter distribution shows the distribution of δ values obtained by bootstrapping under the simpler of the two models, whereas the darker distribution shows the distribution under the more complicated of the two models. A total of 2000 replicates are used for each distribution. The dashed vertical line indicates the observed value of δ when the models are fit to the *Anolis* dataset. (A) BM versus OU.3: the observed likelihood ratio is much more likely under OU.3. (B) OU.3 versus OU.4: here the distributions overlap more, indicating that the data are less informative about this more subtle comparison. (C) OU.3 versus OU.15: these distributions have little overlap and the observed ratio falls clearly in the range of the simpler model. We can conclude that this support for OU.3 is not merely due to lack of power. (D) BM versus OU.1: the data contain almost no information to distinguish between these two models at the estimated (small) level of selection α.



Halliday & Goswami
2016



Release or relHere, we fit models of body mass evolution onto dated phylogenies of Cretaceous and Palaeogene mammals, comparing the effect of dating method on interpretation of evolutionary model. Among traditional palaeontological dating approaches, an Ornstein–Uhlenbeck model with high alpha parameters is recovered as best-fitting when minimum-age dating is used, while branch-sharing methods are highly sensitive to topology.
ease–radiate models are preferred when Bayesian fossilized birth–death method is used, but when using stochastic cal3 dating of trees, a model of increased evolutionary rate without a release in constraint at the Cretaceous–Palaeogene boundary has highest support.

Halliday & Goswami 2016

Model adequacy

I observe: 1121212213131121

Model A

Model B

I observe: 1121212213131121

Model A: Prob of data =
likelihood = $(1/6)^{15} = 2.1e\text{-}12$
NegLnL = 26.9

Model B: Prob of data =
likelihood = $(1/8)^{15} = 2.8e\text{-}14$
NegLnL = 31.9

I observe: 1121212213131121

Model
A

Model
B

I've told you that simulating data is one way to see if a model is adequate. Talk to your neighbors about what this could mean in this case. Code it up, run it. See what you find.

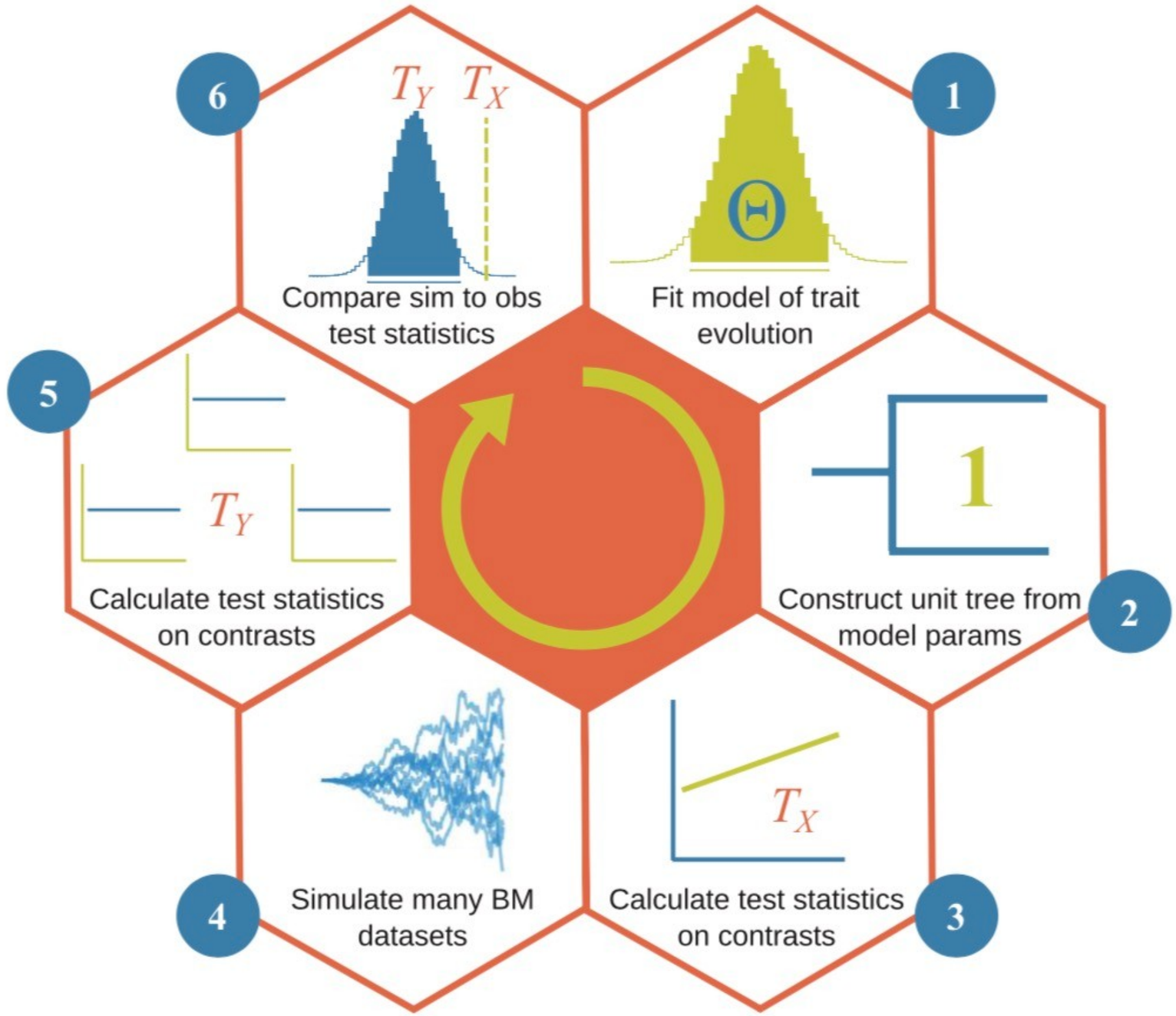


Figure 1: Schematic diagram representing our approach for assessing model adequacy. 1, Fit a model of trait evolution to the data. 2, Use the estimated model parameters to build a unit tree. 3, Compute the contrasts from the data on the unit tree and calculate a set of test statistics $\mathcal{T}_X$. 4, Simulate a large number of data sets on the unit tree, using a Brownian motion (BM) model with $\sigma^2 = 1$. 5, Calculate the test statistics on the contrasts of each simulated data set $\mathcal{T}_Y$. 6, Compare the observed and simulated test statistics. If the observed test statistic lies in the tails of the distribution of simulated test statistics, the model can be rejected as inadequate. The rotational circle in the center of the diagram indicates that assessing model adequacy is an iterative process. If a model is rejected as inadequate, the next step is to propose a new model and repeat the procedure.

Pennell, FitzJohn, Cornwell, Harmon 2015

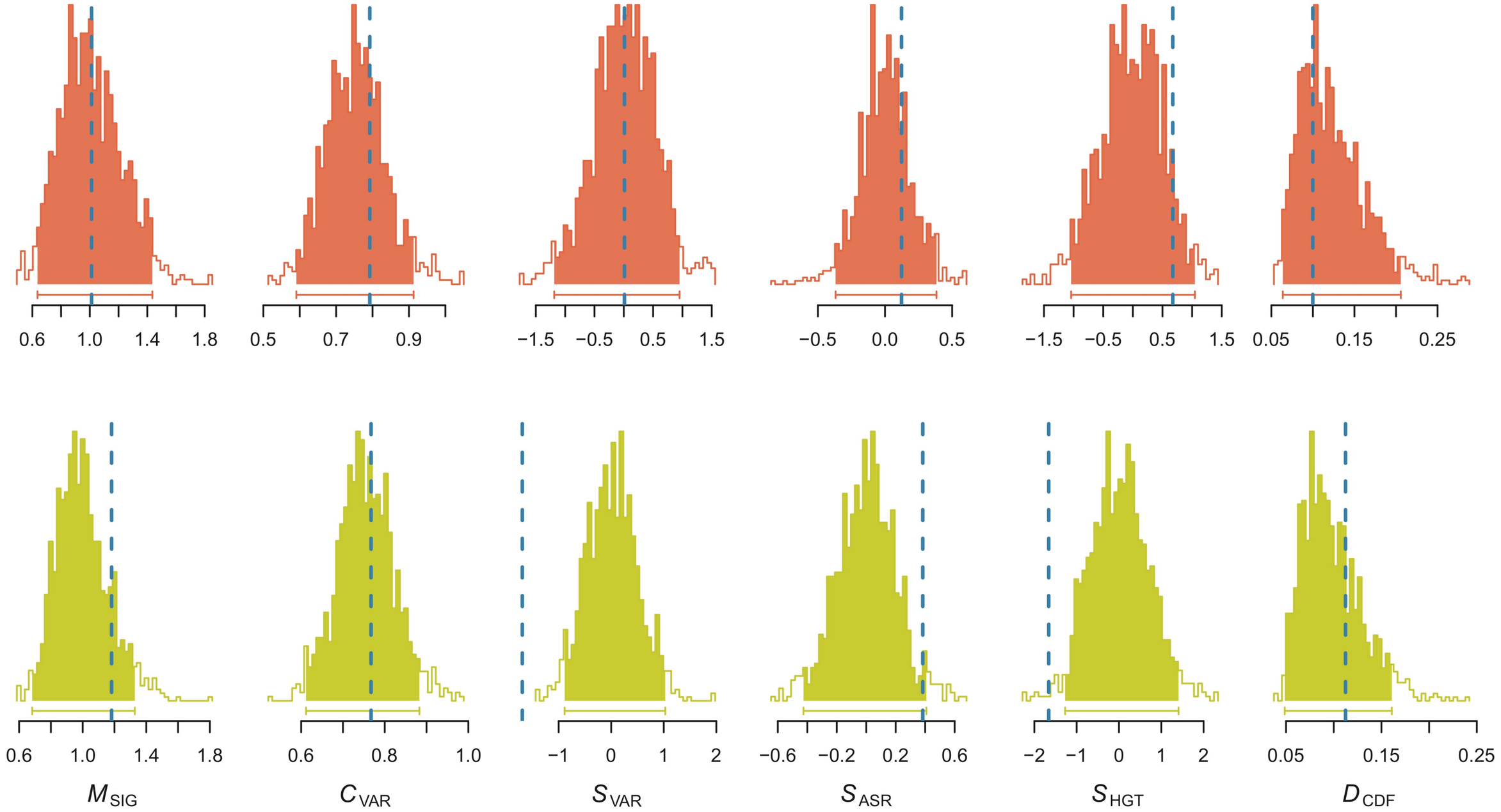


Figure 2: Illustration of our approach to model adequacy. We fitted three models (Brownian motion, Ornstein-Uhlenbeck [OU], and early burst) to seed mass data from two different tree families, the Meliaceae (top, red) and the Fagaceae (bottom, yellow). In both cases, an OU model (analyzed here) was strongly supported when fitted with maximum likelihood. The plotted distributions are the test statistics (M_{SIG}, C_{VAR}, S_{VAR}, S_{ASR}, S_{HGT}, D_{CDF}) calculated from the contrasts of the simulated data; the bars underneath the plots represent 95% of the density. The dashed vertical lines are the values of the test statistics calculated on the contrasts of the observed data. Using our test statistics, an OU model appears to be an adequate model for the evolution of seed mass in the Meliaceae; for all of the test statistics, the observed test statistic lies in the middle of the distribution of simulated test statistics. For the Fagaceae, the slopes of the contrasts against their expected variances S_{VAR} and node height S_{HGT} are much lower than the expectations under the model.

Pennell, FitzJohn, Cornwell, Harmon 2015

- What would you do if you find your model is inadequate? Ideally, make a better model. What if you can't? Should we #JustSayNo and not continue, or #BeBest with the models we have?

- What would you do if you find your results depend on dating or some other aspect of your tree?

- What would you do if your regimes are arranged such that there's not enough information to estimate optima?