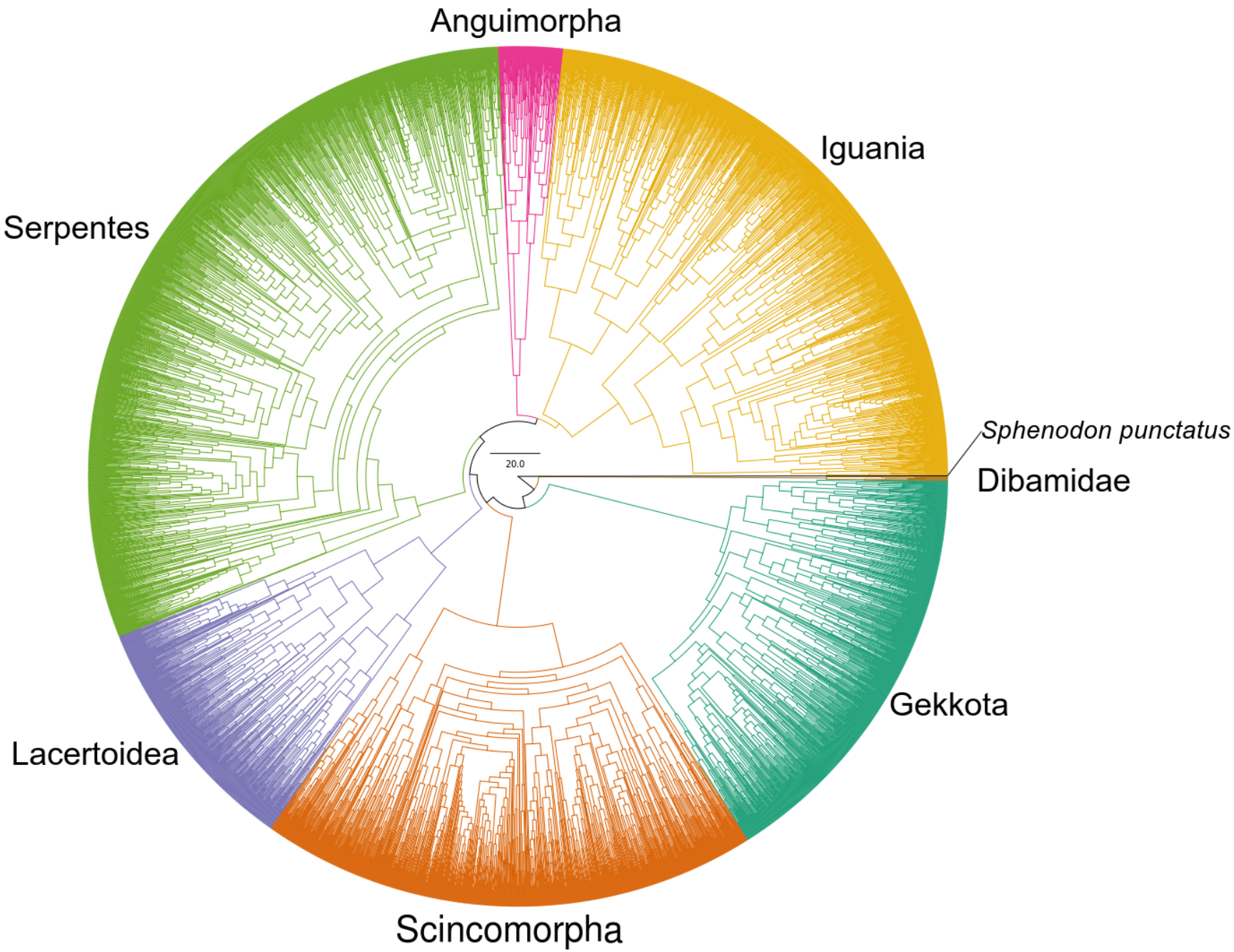
Scientific ethics, tree testing, Open Tree of Life

Workshop on Molecular Evolution 2018

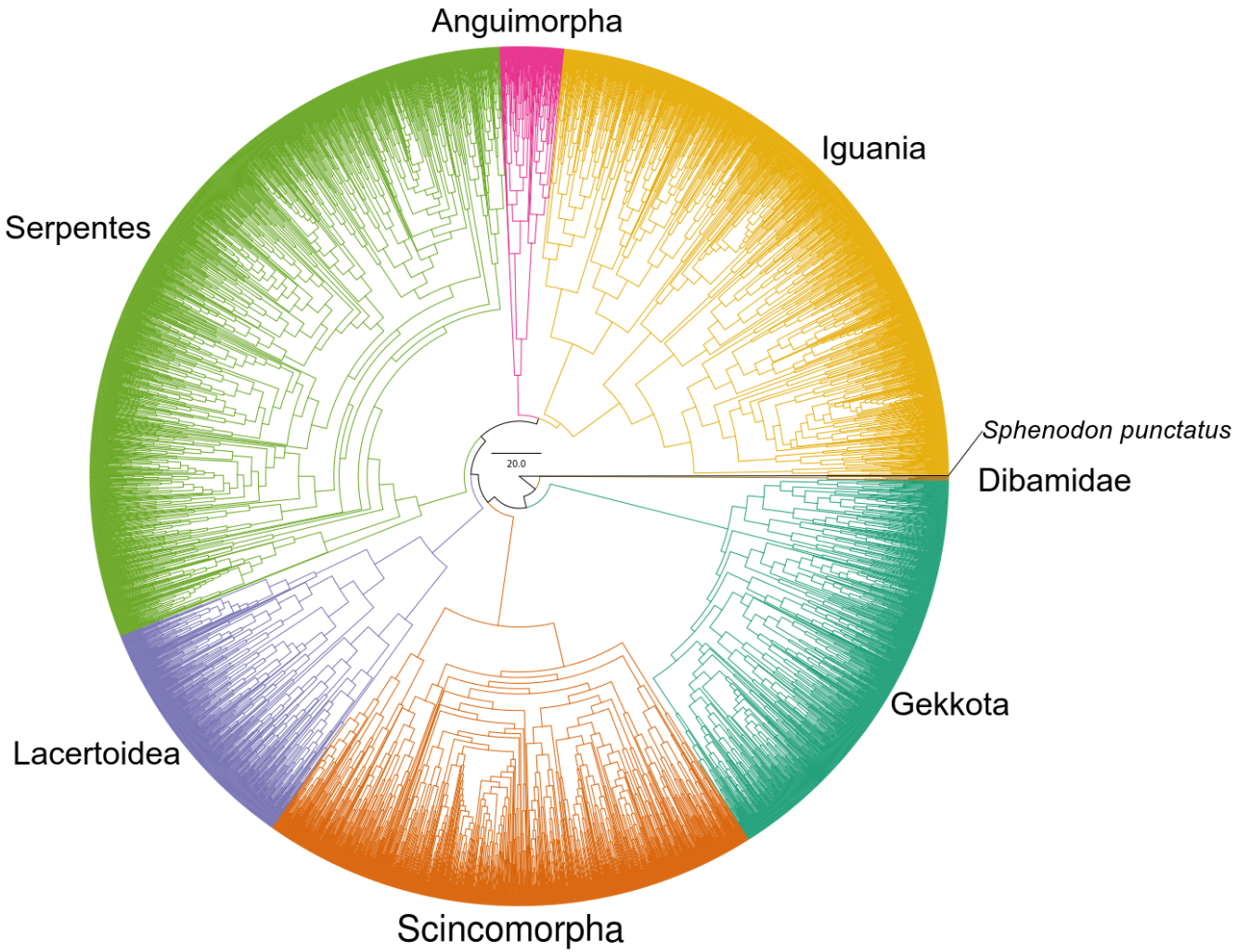
Marine Biological Lab, Woods Hole, MA. USA

Mark T. Holder
University of Kansas

next ≈ 22 slides from David Hillis



Data from Pyron et al., 2014; Figure from Wright et al., 2015



Good:

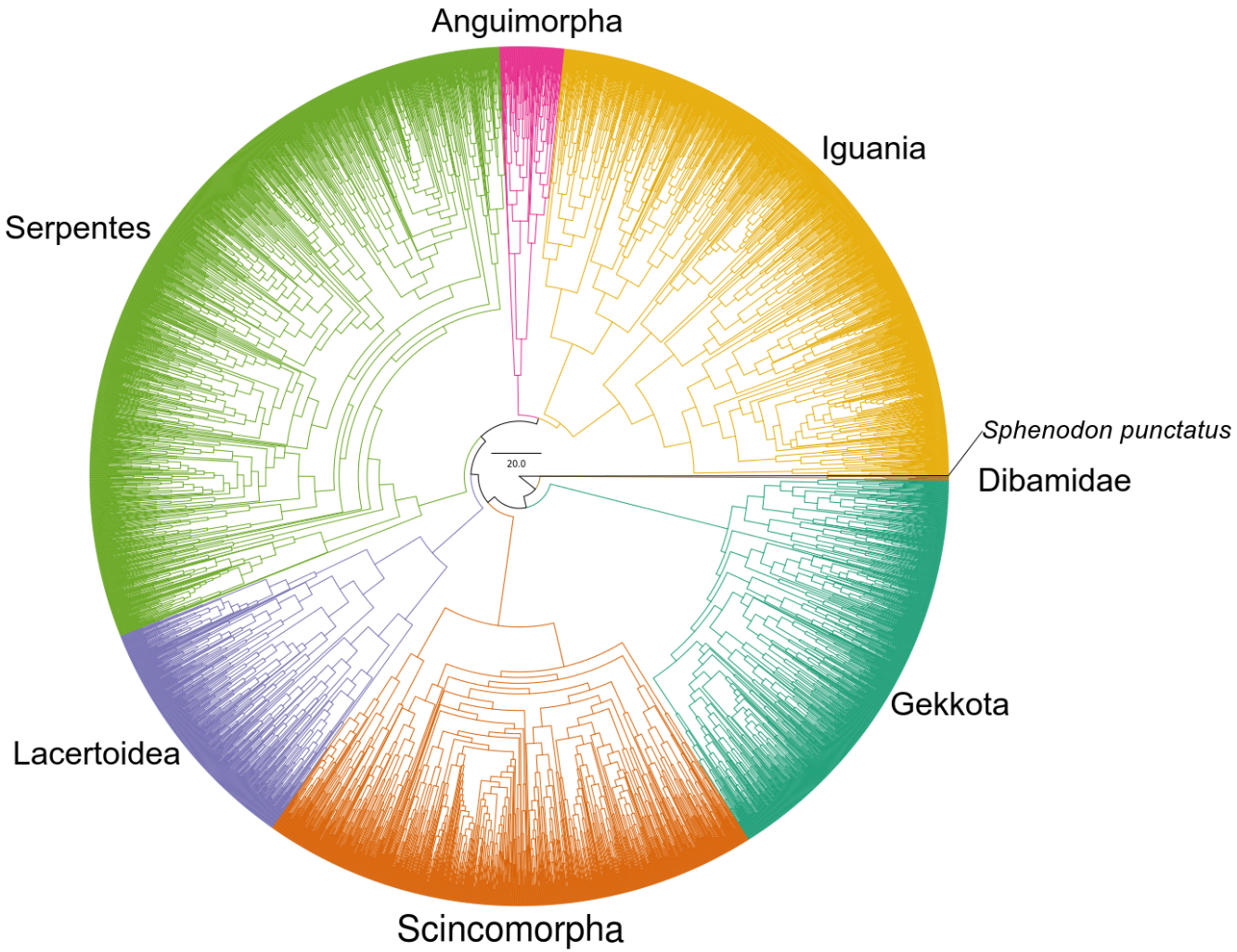
- About 4,000 species
- Multiple nuclear and mitochondrial genes

But worrisome:

- 81% missing data
- Biases in missing data (most taxa only have fragments of mtDNA)

How do biases in missing data affect our models?

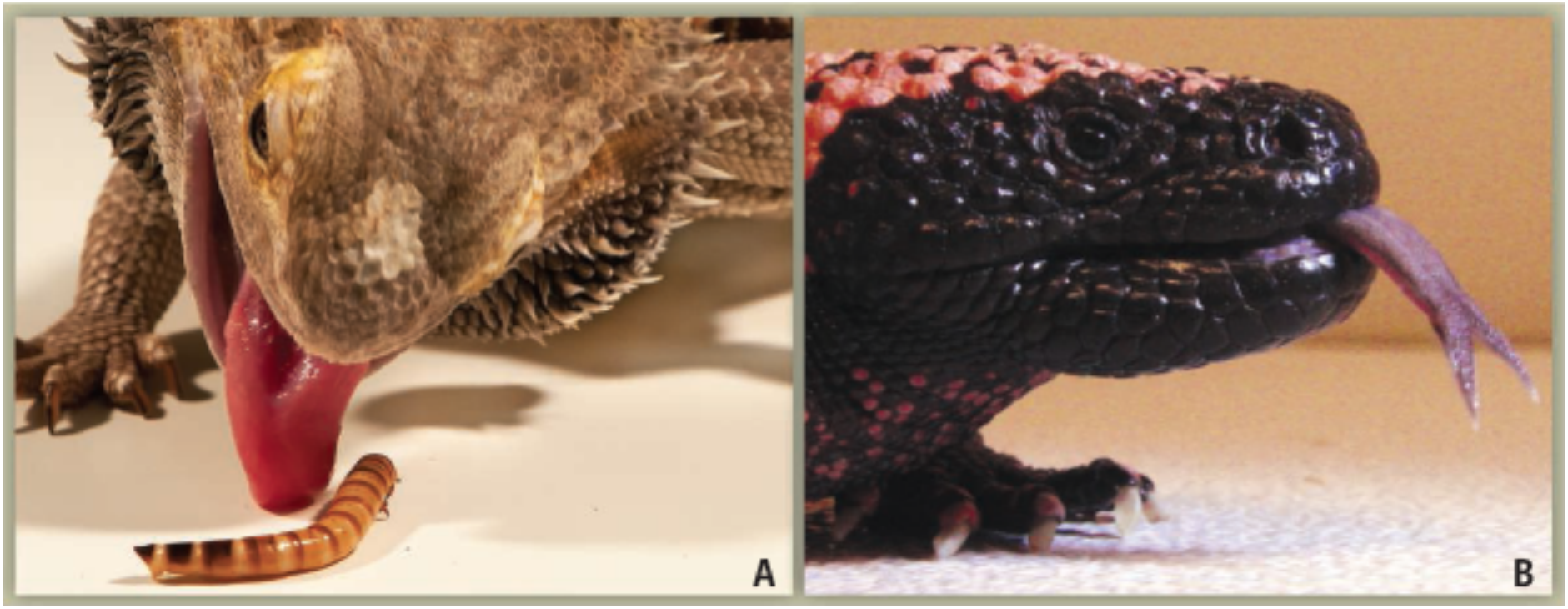
Data from Pyron et al., 2014; Figure from Wright et al., 2015



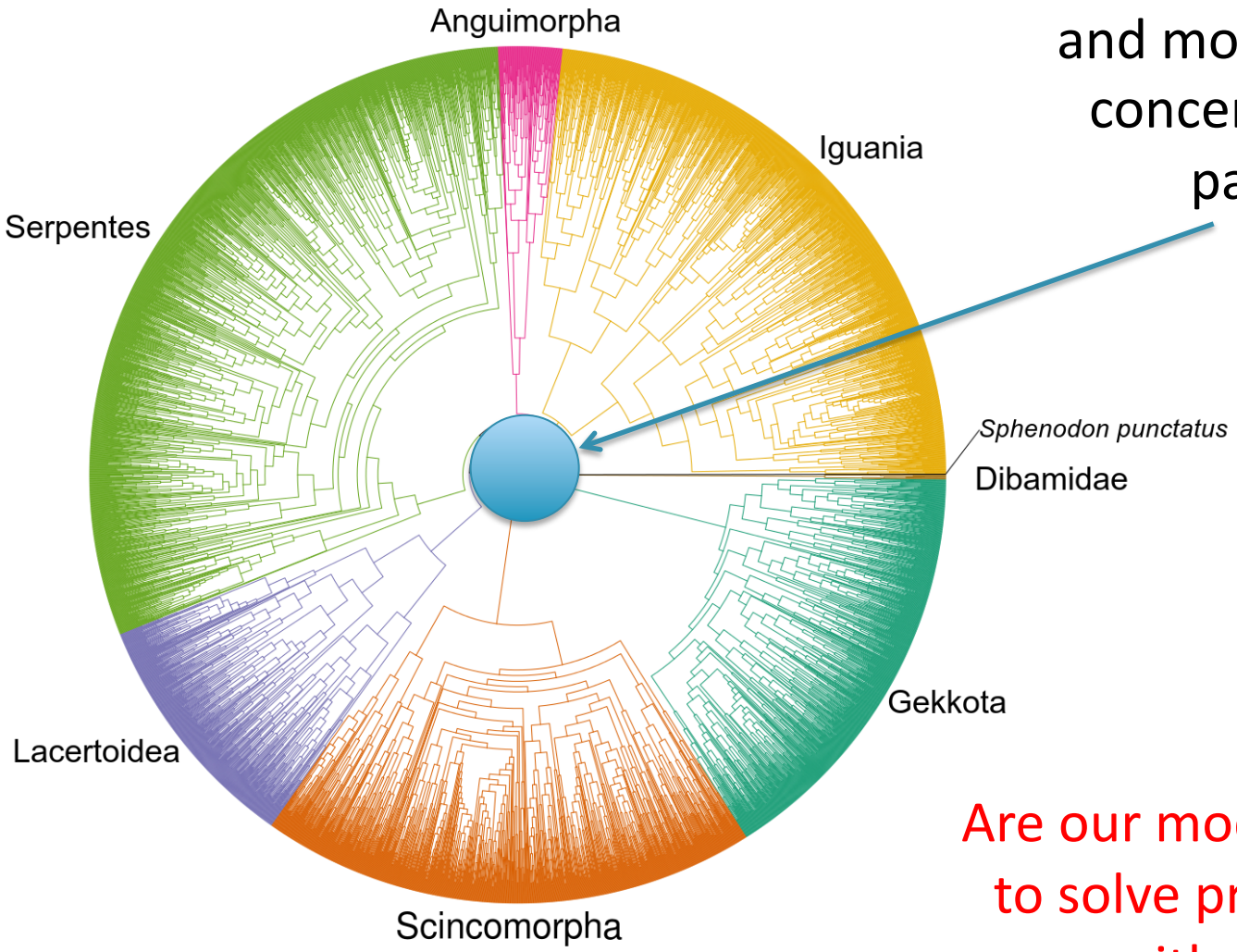
Tree conflicts with morphology, which places iguanians as the sister group of remaining squamates.

How confident should we be in this tree?

Data from Pyron et al., 2014; Figure from Wright et al., 2015


A
B

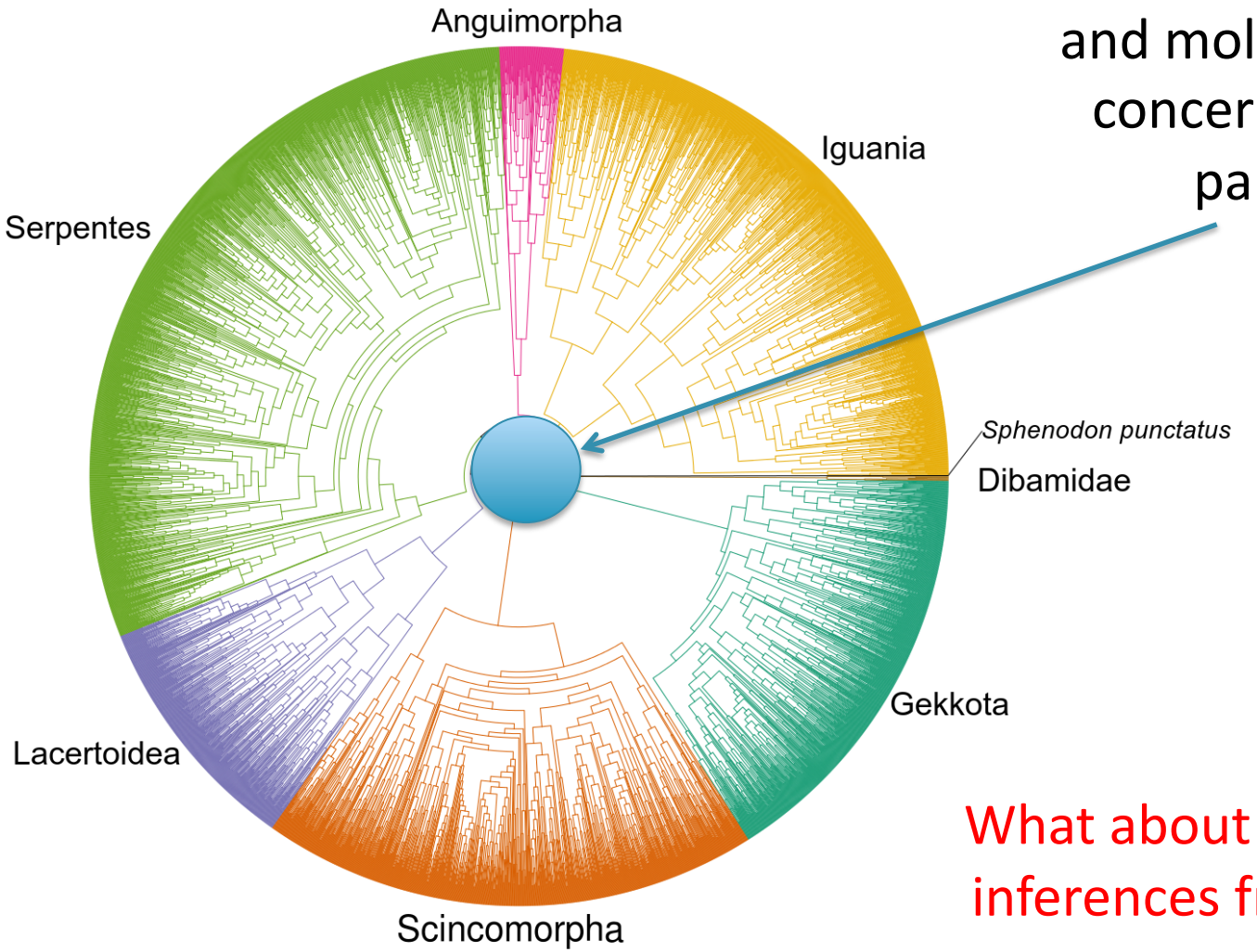
All the arguments between morphologists and molecular systematists concern branches in this part of the tree.



Are our models good enough to solve problems like this, with these data?

Data from Pyron et al., 2014; Figure from Wright et al., 2015

All the arguments between morphologists and molecular systematists concern branches in this part of the tree.



What about our evolutionary inferences from these trees?

Data from Pyron et al., 2014; Figure from Wright et al.,2015

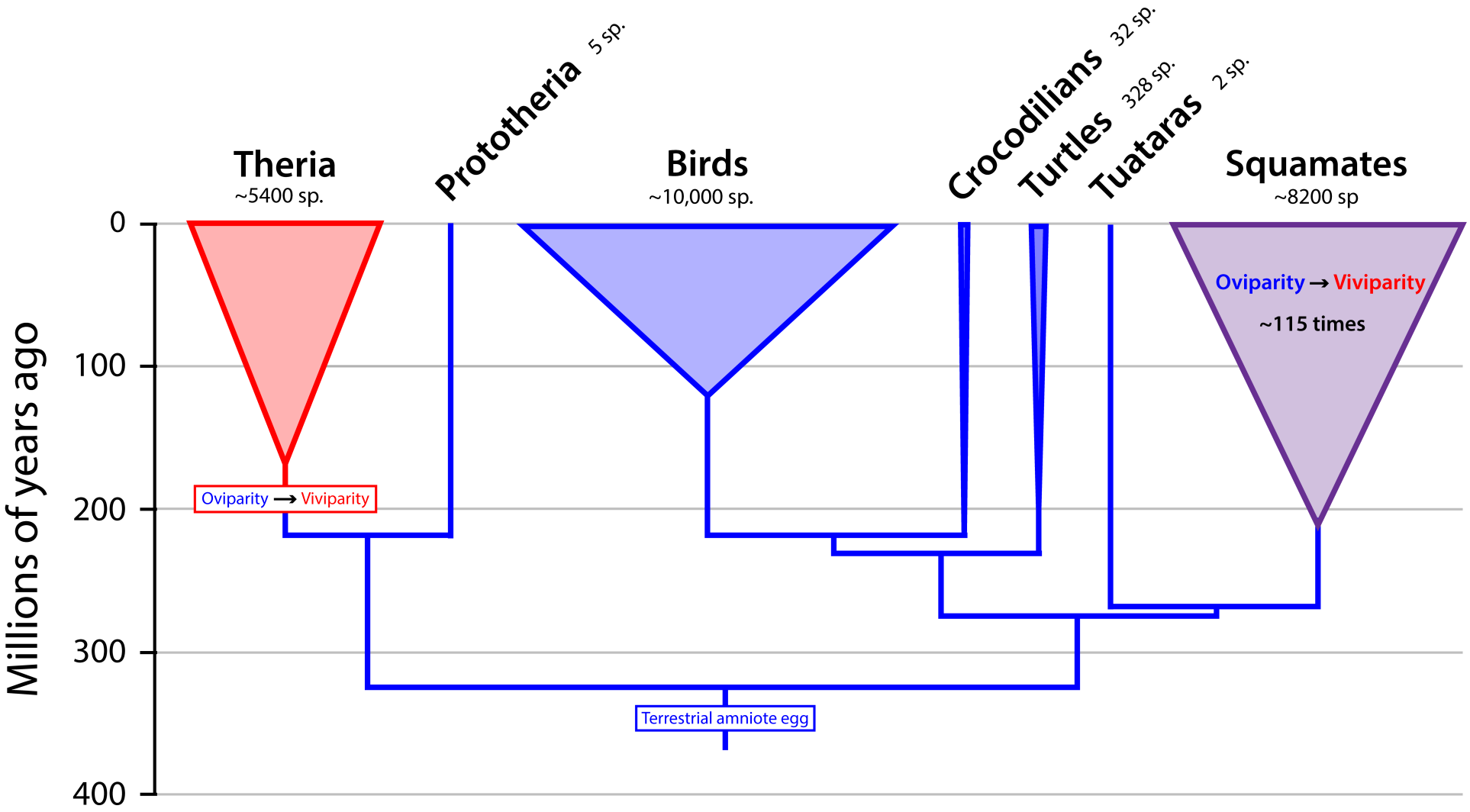
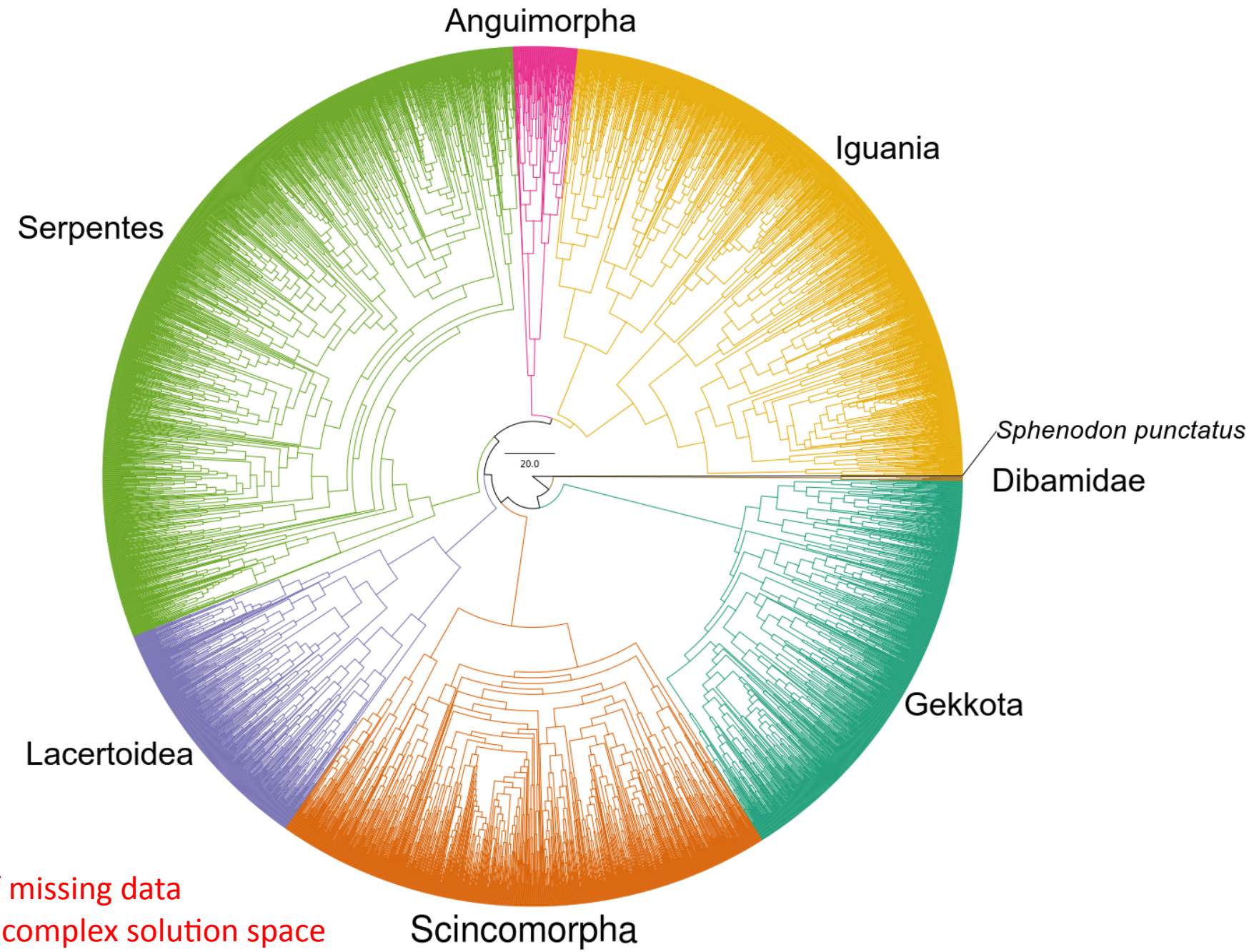


Figure from Wright et al., 2015



- Lots of missing data
- Highly complex solution space

Data from Pyron et al., 2014; Figure from Wright et al., 2015

Best tree so far (improvement of >83,796 ln-L units)

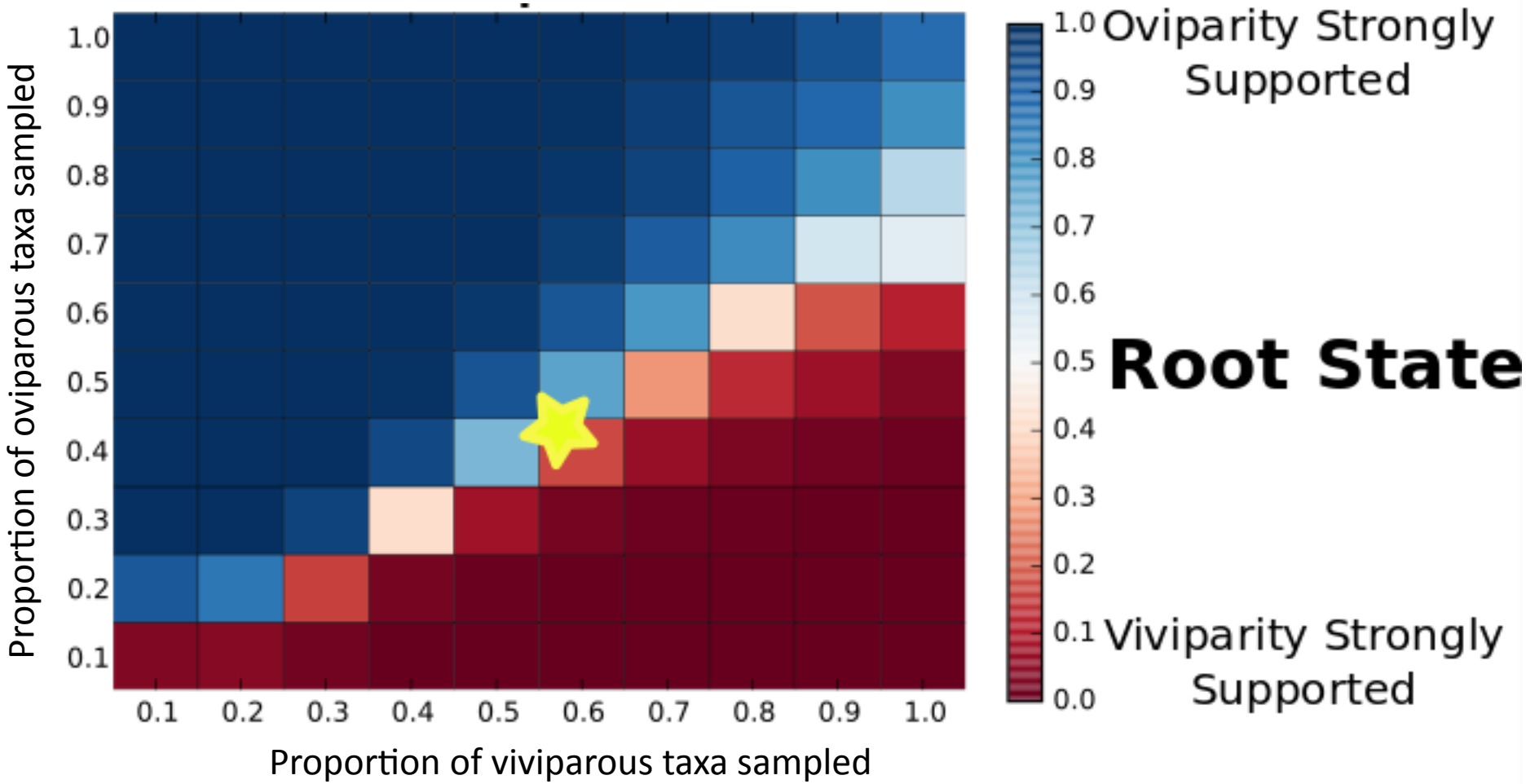


Figure from Wright et al., 2015

Parameter of model (e.g., **α** parameter of **Γ** distribution)

Parameter of model (e.g., **α** parameter of **Γ** distribution)

Assumed or estimated value of parameter leads to inference **Red**.
How confident should we be in answer **Red**?

Parameter of model (e.g., **α** parameter of **Γ** distribution)

Assumed or estimated value of parameter leads to inference **Red**.

How confident should we be in answer **Red**?

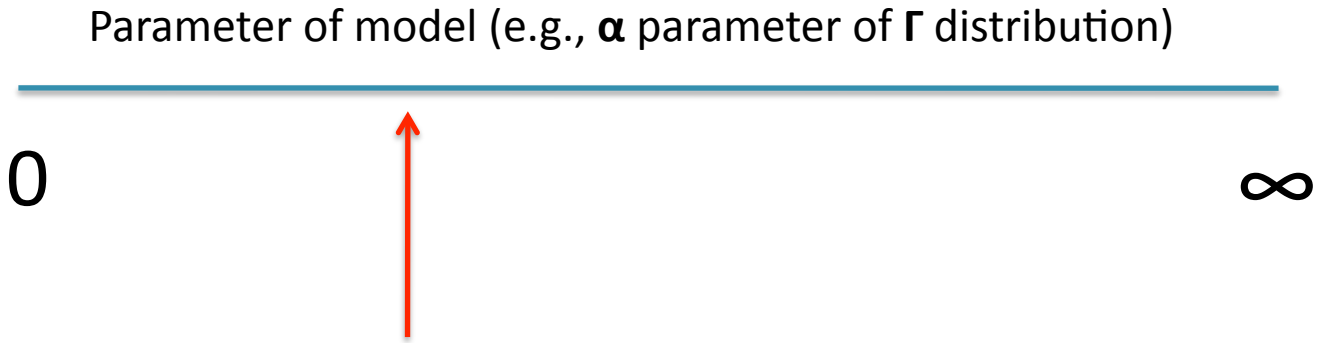
1. We should be concerned with the sampling error that led to this estimated value of **α**. (This is all that many people do).

Parameter of model (e.g., **α** parameter of **Γ** distribution)

Assumed or estimated value of parameter leads to inference **Red**.
How confident should we be in answer **Red**?

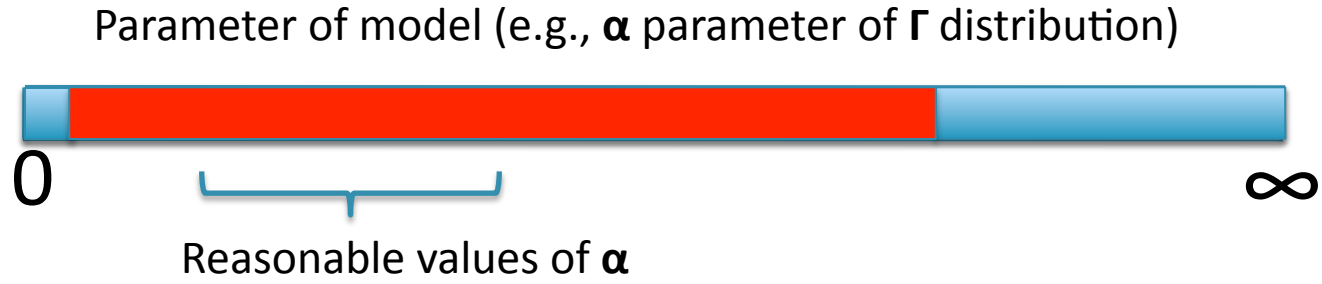
1. We should be concerned with the sampling error that led to this estimated value of **α**. (This is all that many people do).
2. We should be concerned with the sensitivity of the answer to this parameter value.

Parameter of model (e.g., **α** parameter of **Γ** distribution)



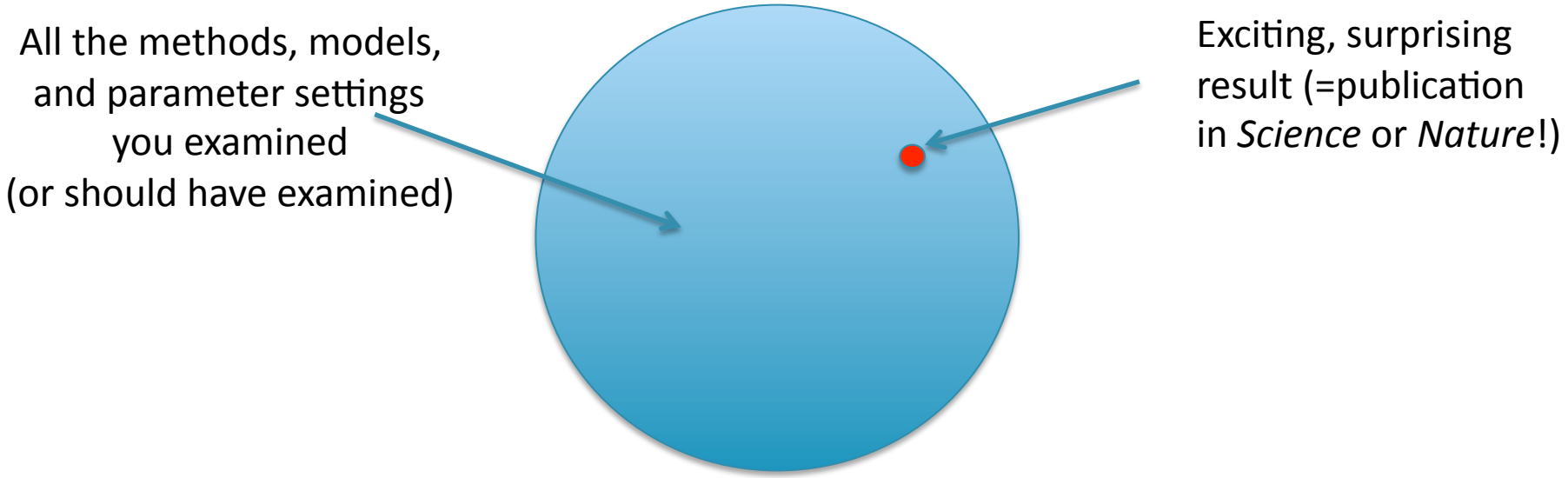
Assumed or estimated value of parameter leads to inference **Red**.
How confident should we be in answer **Red**?

1. We should be concerned with the sampling error that led to this estimated value of **α**. (This is all that many people do).
2. We should be concerned with the sensitivity of the answer to this parameter value.



Ethics of Data Presentation and Analysis

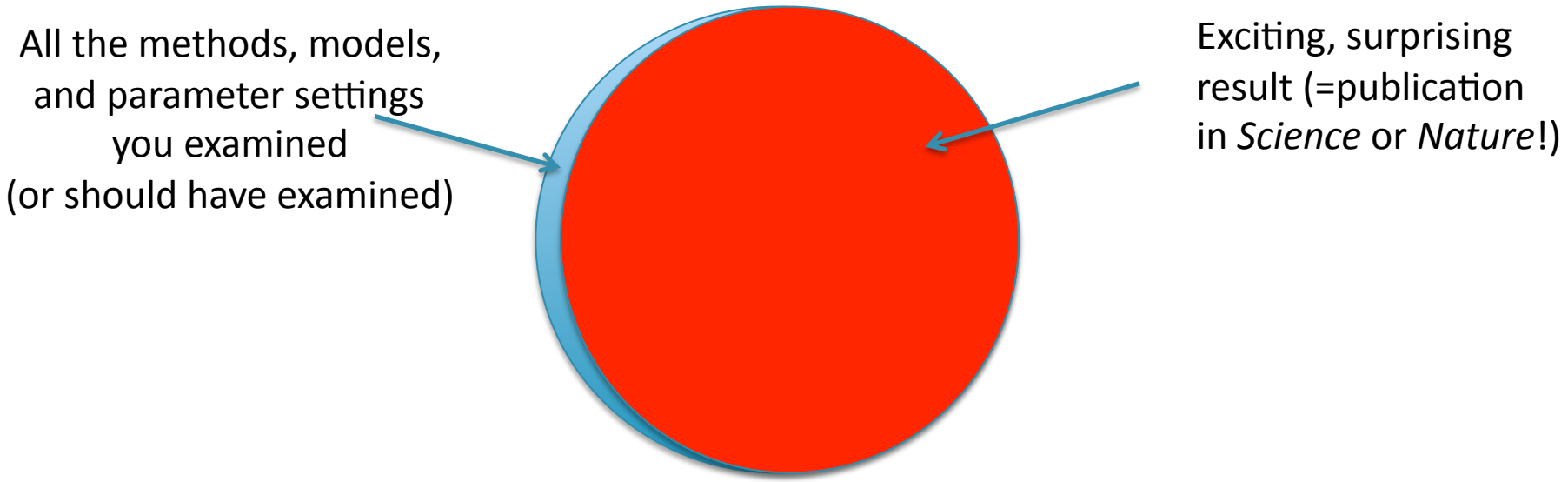
- Assume you analyze your data with multiple models, methods, or parameter settings:



What do you publish?

Ethics of Data Presentation and Analysis

- Assume you analyze your data with multiple models, methods, or parameter settings:



What do you publish?

Ethics of Data Presentation and Analysis

- All data, method descriptions, scripts, and programs must be publicly available in a way that all your analyses can be repeated, checked, and extended

Ethics of Data Presentation and Analysis

- All data, method descriptions, scripts, and programs must be publicly available in a way that all your analyses can be repeated, checked, and extended
  - Not just sequences in GenBank and a statement that you used a particular program for analysis!

Ethics of Data Presentation and Analysis

- All data, method descriptions, scripts, and programs must be publicly available in a way that all your analyses can be repeated, checked, and extended
  - Not just sequences in GenBank and a statement that you used a particular program for analysis!
  - Include alignments, parameter settings, scripts with program settings, and information on the range of methods, models, and parameter settings examined.

Ethics of Data Presentation and Analysis

- Where can you put all this information?
  - Most journals allow online Supplementary Information
  - There may be discipline specific data repositories (such as *TreeBase* for phylogenetic analyses; http:// http://treebase.org)
  - Public, archival databases such as *Dryad*, a digital data repository (http://datadryad.org/)
  - Individual websites are not the best solution, since long-term access and archiving are serious problems

Conclusions 1 - confidence on trees

1. Non-parametric bootstrapping: useful for assessing sampling error, but a little hard to interpret precisely.
   - Susko's aBP gives $1 - aBP \approx P$-value for the hypothesis that a recovered branch is not present in the true tree.
2. "How should we assign a P-value to tree hypothesis?" is surprisingly complicated.
   - Kishino-Hasegawa (KH-Test) if testing 2 (*a priori*) trees.
   - Shimodaira's approximately unbiased (AU-Test) for sets of trees.
   - Parametric bootstrapping (can simulate under complex models)

Conclusions 2 - confidence about evo. hypotheses

If H_0 is about the evolution of a trait:

1. P-value must consider uncertainty of the tree:
   - can be large P over confidence set of trees.
   - Bayesian methods enable prior predictive or posterior predictive P-values.

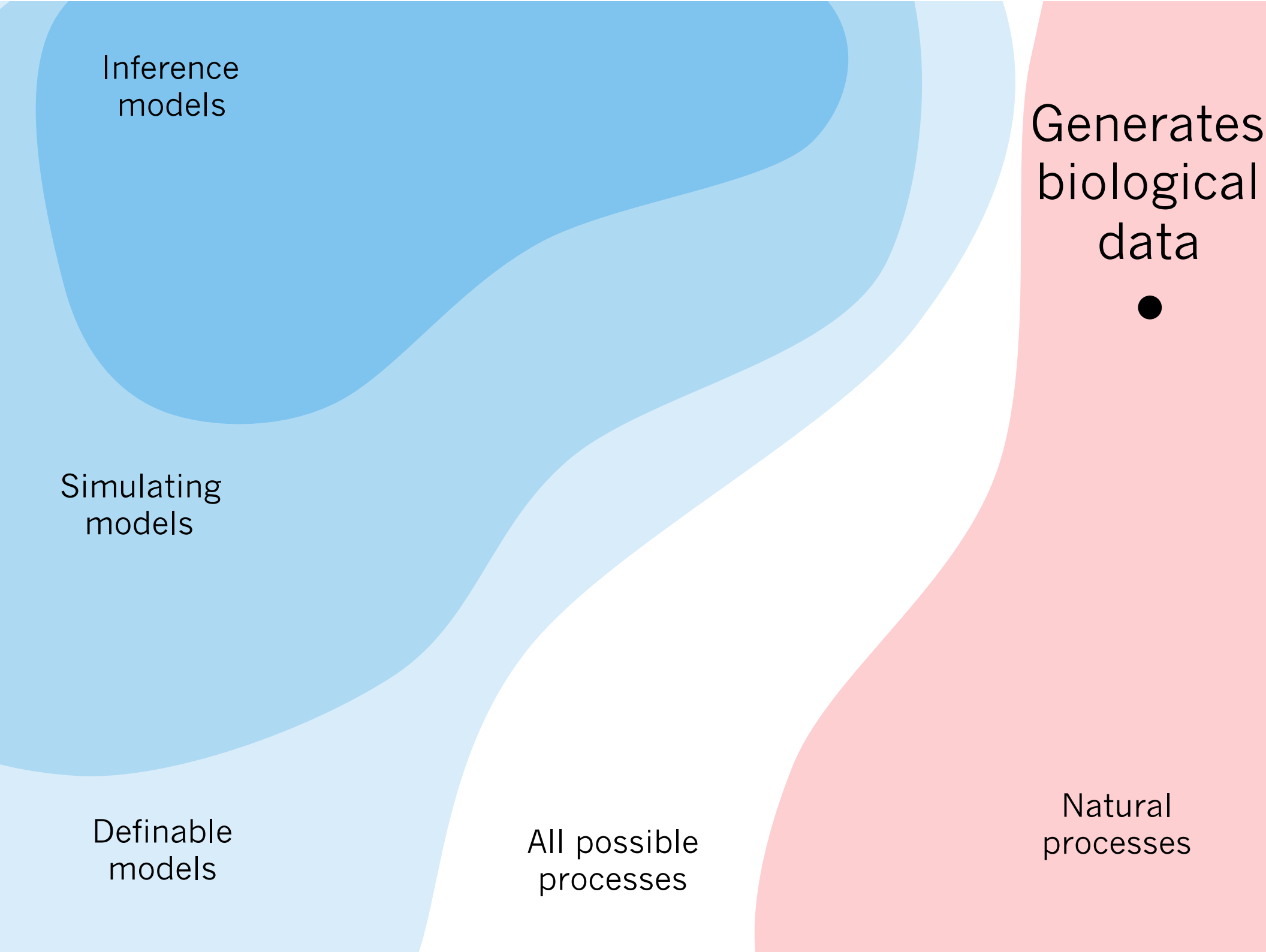
Conclusions 3 - simulate your own null distributions

1. In phylogenetics we often have to simulate data to approximate P-values
2. Designing the simulations requires care to make a convincing argument.

We have some parametric bootstrapping labs on the course wiki:

https://molevol.mbl.edu/index.php/ParametricBootstrappingLab

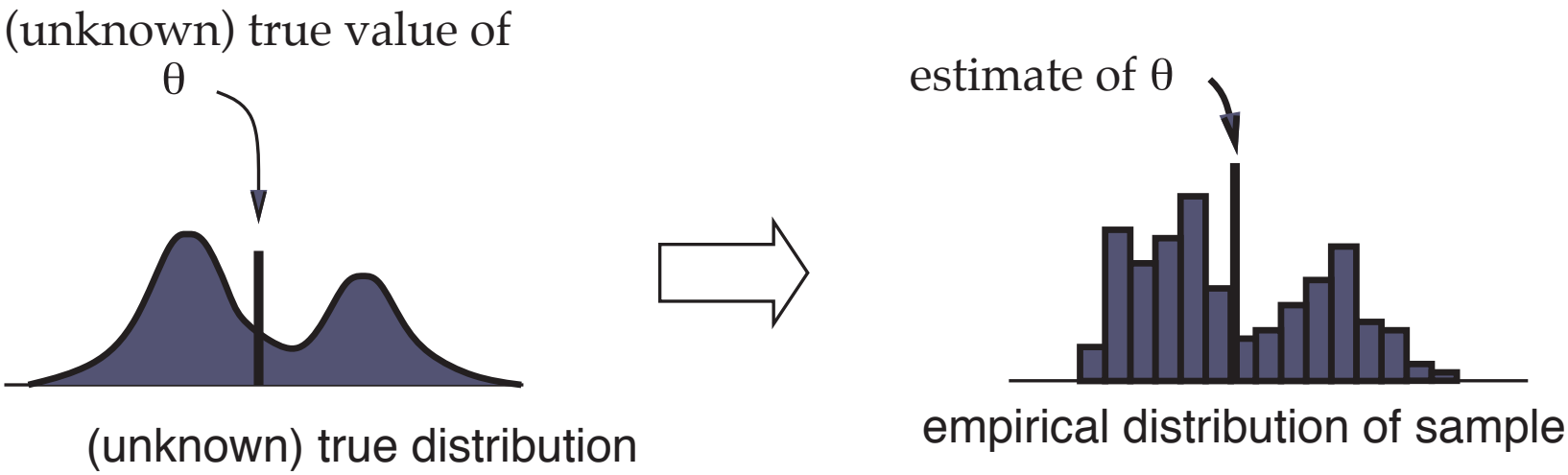
next slide by Landis


Inference models
Simulating models
Definable models
All possible processes
Natural processes
Generates biological data

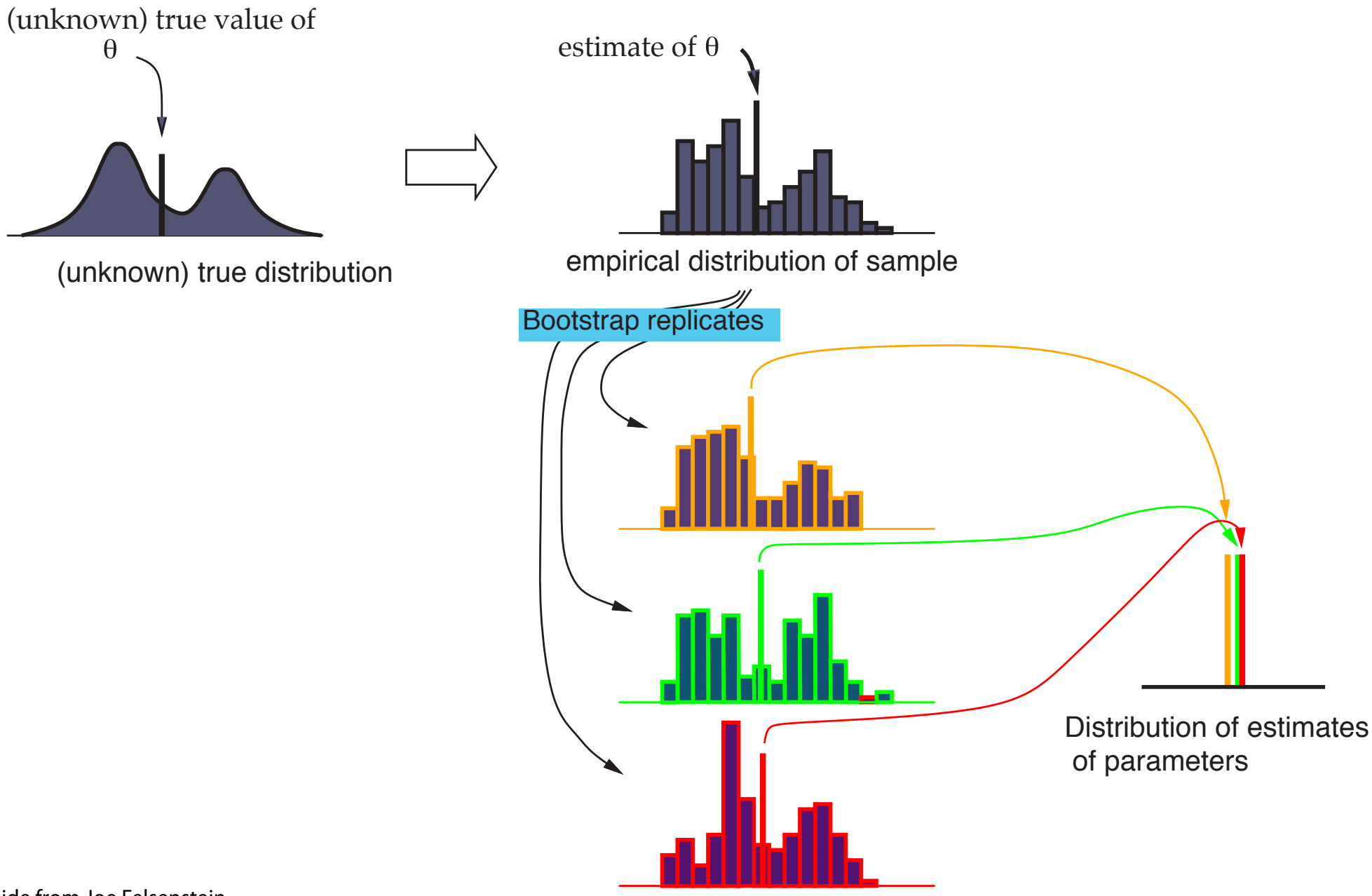
Reasons phylogenetic inference might be wrong

1. *Systematic error* – Our inference method might not be sophisticated enough
2. *Random error* – We might not have enough data – we are misled by sampling error.

The bootstrap

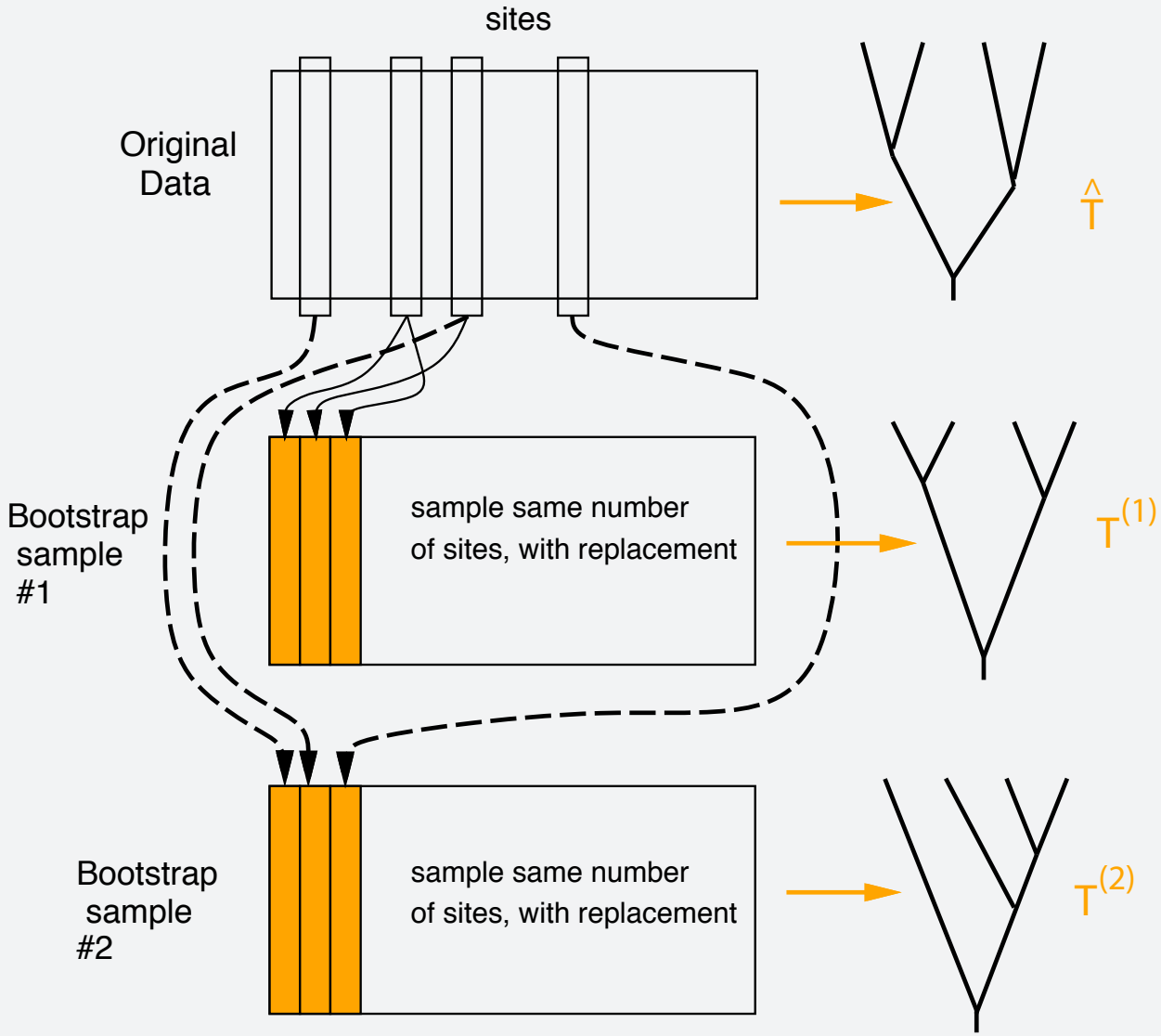


The bootstrap



Slide from Joe Felsenstein

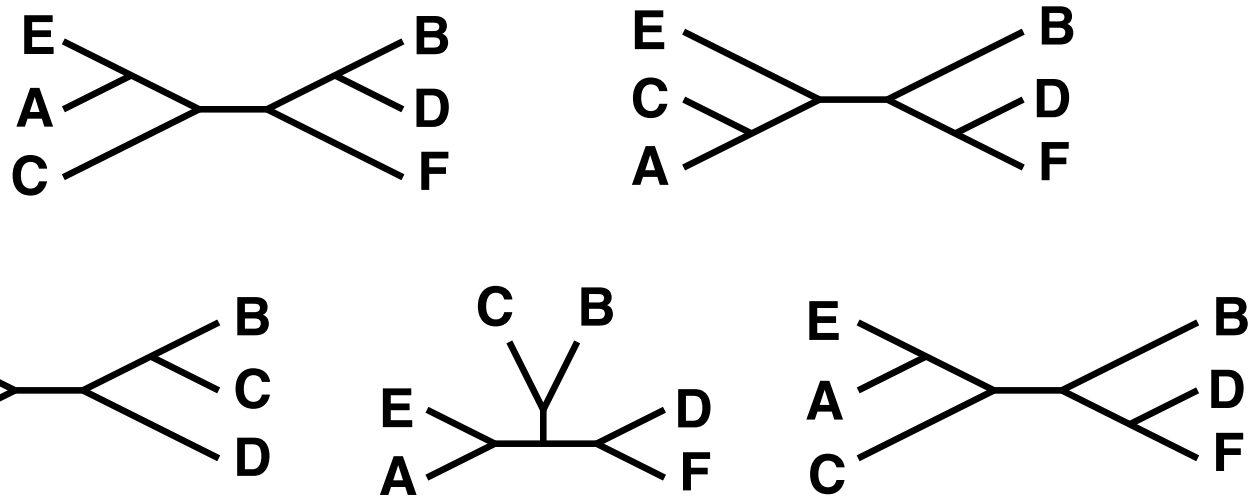
The bootstrap for phylogenies



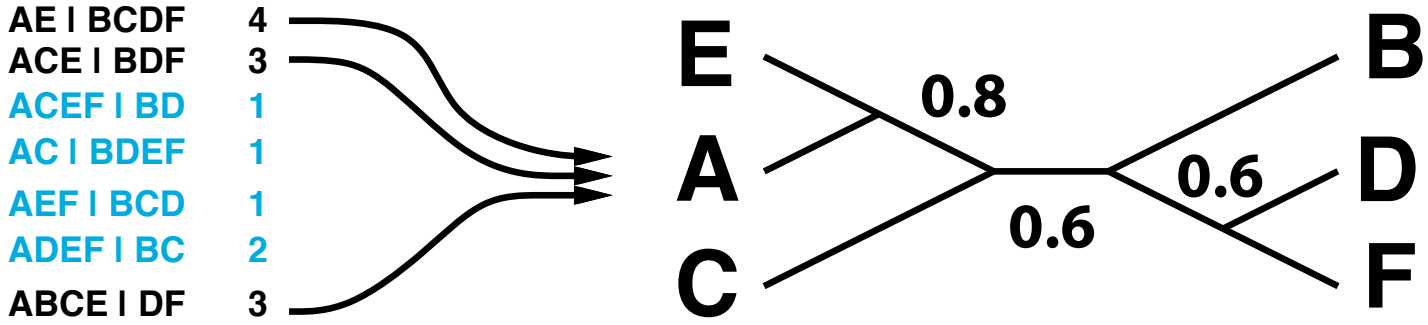
Slide from Joe Felsenstein

The majority-rule consensus tree

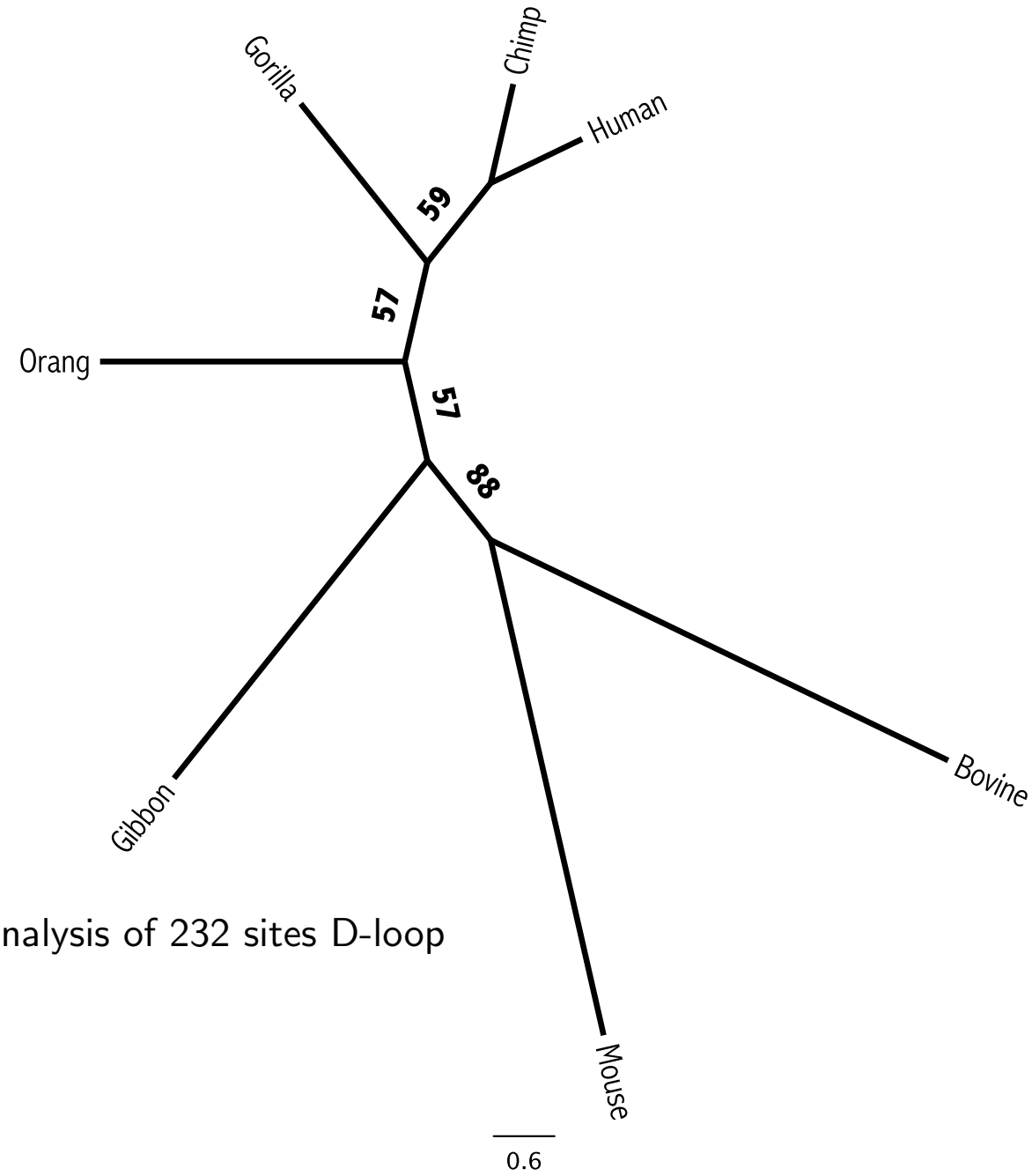
Trees:



How many times each partition of species is found:



Slide from Joe Felsenstein



From Hasegawa's analysis of 232 sites D-loop

http://phylo.bio.ku.edu/mephytis/boot-sample.html

http://phylo.bio.ku.edu/mephytis/parsimony.html

http://phylo.bio.ku.edu/mephytis/bootstrap.html

Bootstrapping for branch support

- Typically a few hundred bootstrap, pseudoreplicate datasets are produced.

- Less thorough searching is faster, but will usually artificially lower bootstrap proportions (BP). However, Anisimova et al. (2011) report that RAxML's rapid bootstrap algorithm may inflate BP.

- "Rogue" taxa can lower support for many splits – you do not have to use the majority-rule consensus tree to summarize bootstrap confidence statements; See also (Lemoine et al., 2017)

Bootstrap proportions have been characterized as providing:

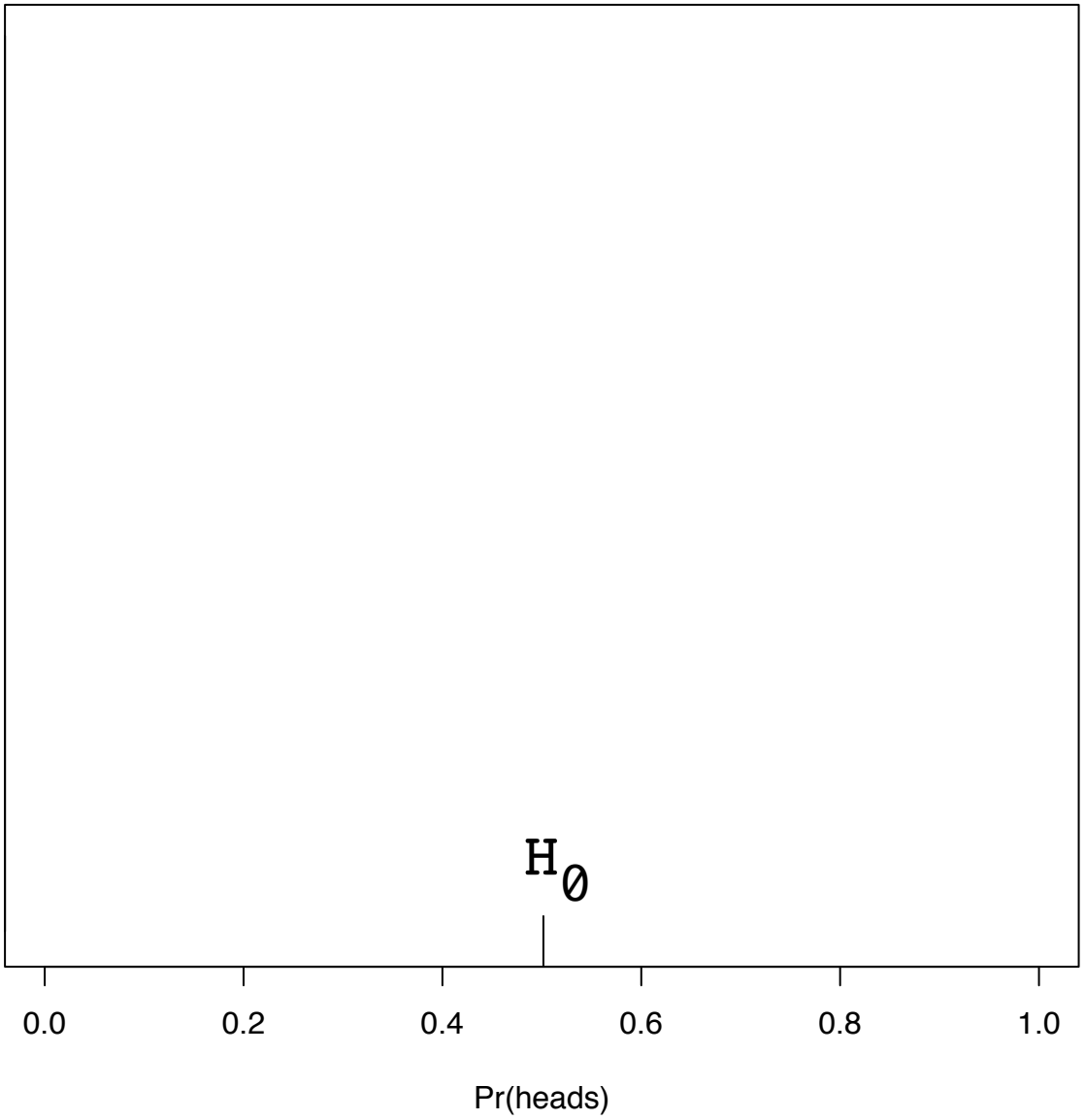
- a measure of repeatability,
- an estimate of the probability that the tree is correct (and bootstrapping has been criticized as being too conservative in this context),
- the P-value for a tree or clade

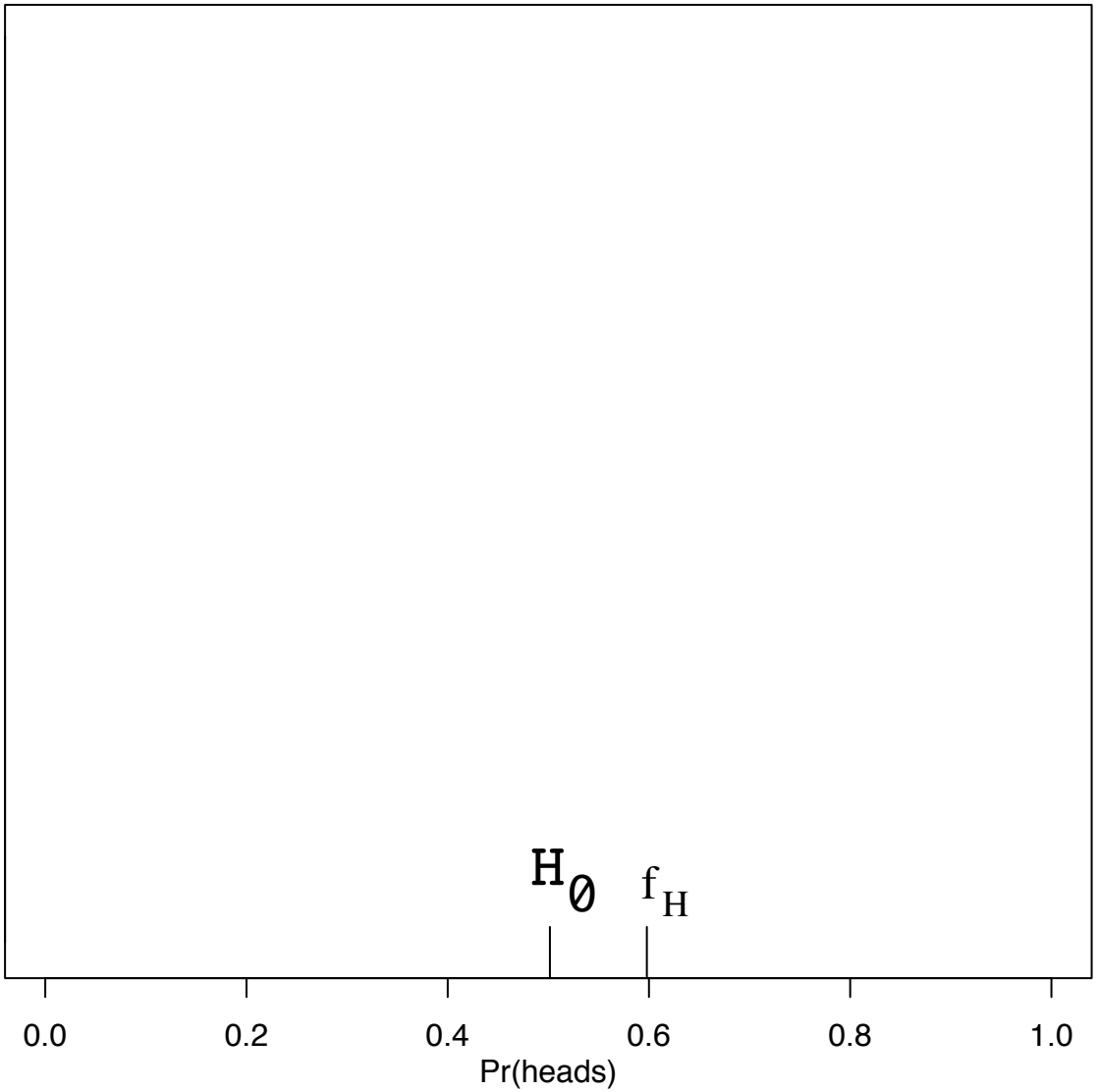
Frequentist hypothesis testing: coin flipping example

$N = 100$ and $h = 60$
Can we reject the fair coin hypothesis? $H_0 : \Pr(\text{heads}) = 0.5$

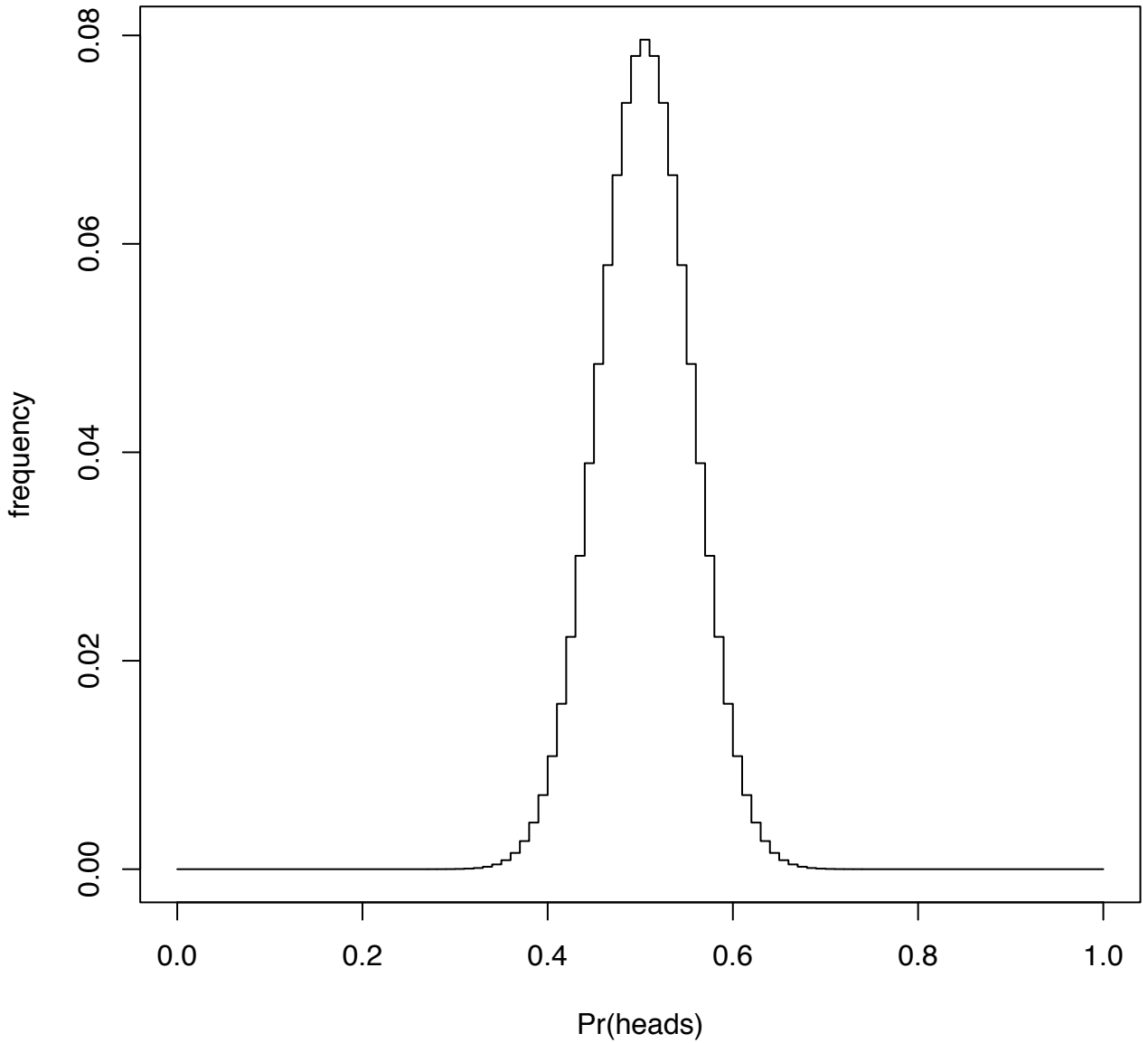
The "recipe" is:

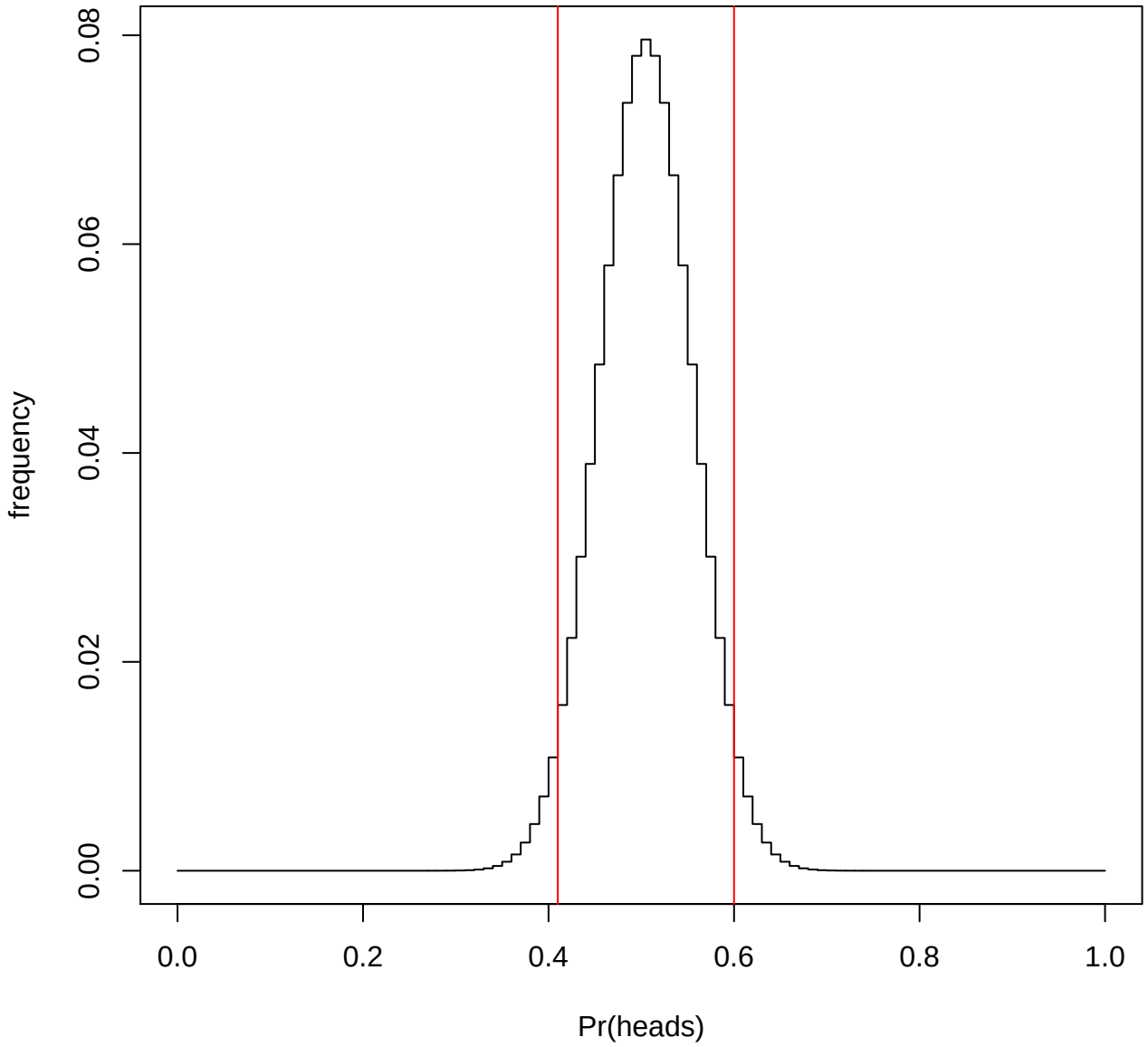
1. Formulate null (H_0) and alternative (H_A) hypotheses.
2. Choose an acceptable Type-I error rate (significance level)
3. Choose a test statistic: $f_H =$ fraction of heads in sample. $f_H = 0.6$
4. Characterize the null distribution of the test statistic
5. Calculate the P-value: The probability of a test statistic value more extreme than f_H arising *even if* H_0 *is true*.
6. Reject H_0 if P-value is $\leq$ your Type I error rate.


H_0
0.0
0.2
0.4
0.6
0.8
1.0
Pr(heads)


H0
fH
0.0
0.2
0.4
0.6
0.8
1.0
Pr(heads)

Null distribution





P-value ≈ 0.058

Making similar plots for tree inference is hard.

- Our parameter space is trees and branch lengths.
- Our data is a matrix of characters.
- It is hard to put these objects in the same space. You can do this "pattern frequency space".

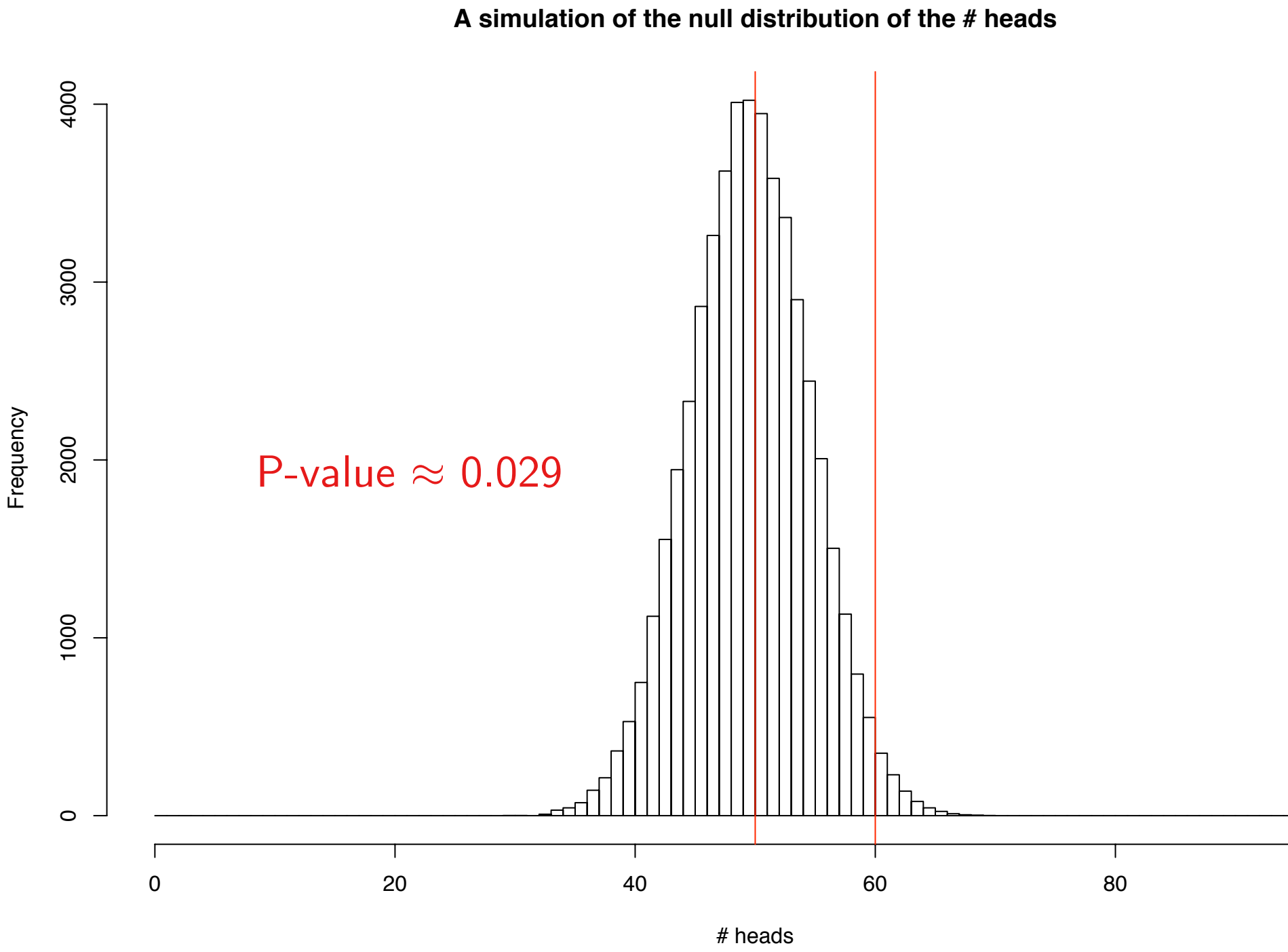
Some cartoons of projections of this space are posted at:
http://phylo.bio.ku.edu/slides/pattern-freq-space-cartoons.pdf

coin flipping

$N = 100$ and $H = 60$

Can we reject the hypothesis of a fair coin?

We can use simulation to generate the null distribution (we could actually use the binomial distribution to analytically solve this one)...


A simulation of the null distribution of the # heads
P-value ≈ 0.029
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
heads

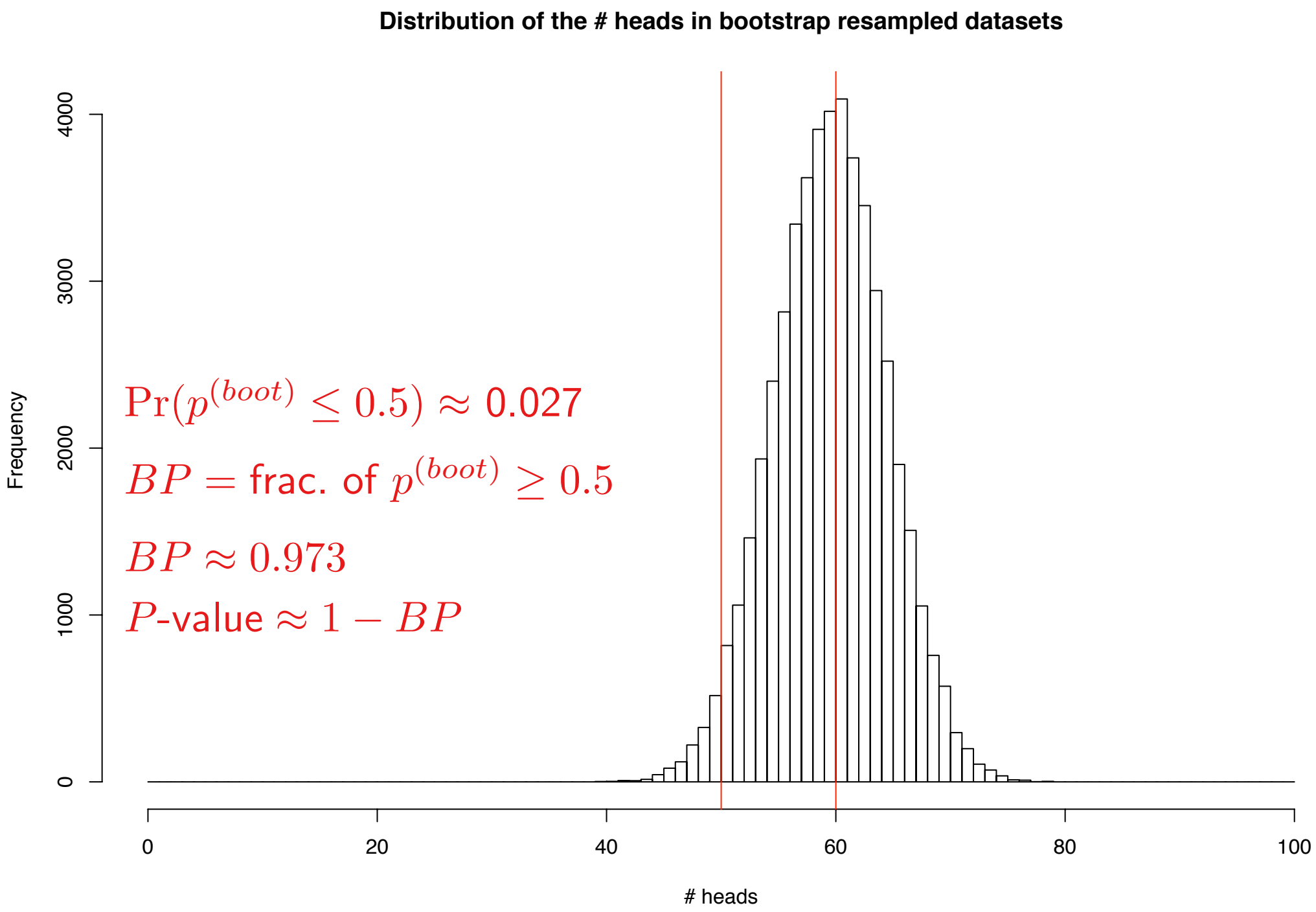
We discussed how bootstrapping gives us a sense of the variability ofour estimate

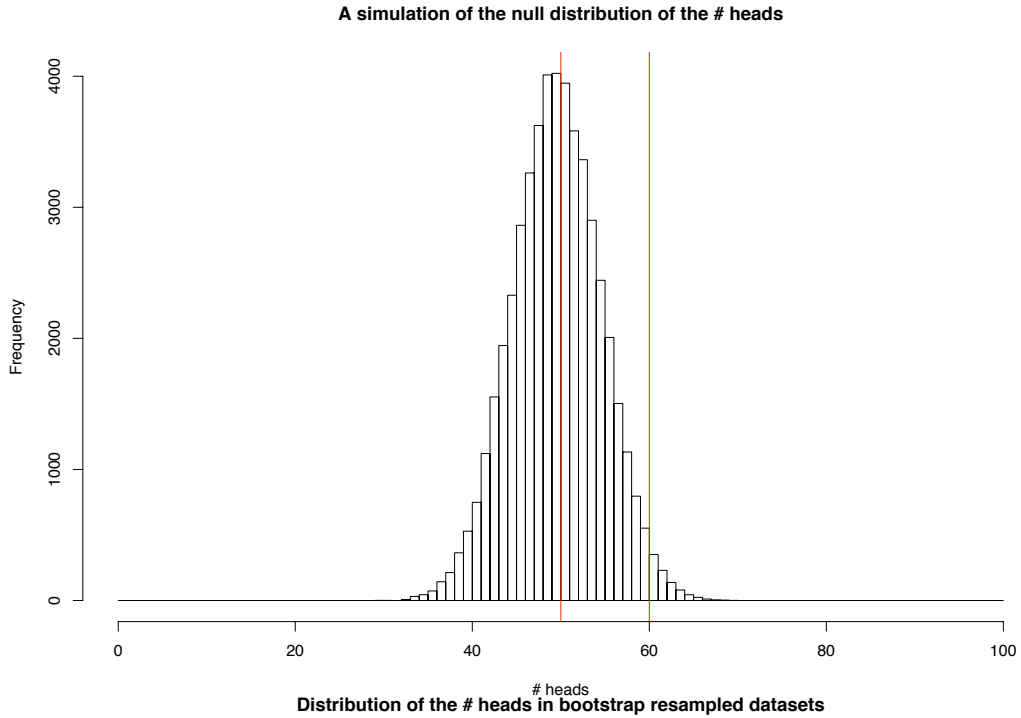
It can also give a tail probability for $\Pr(f_H^{(boot)} \leq 0.5)$

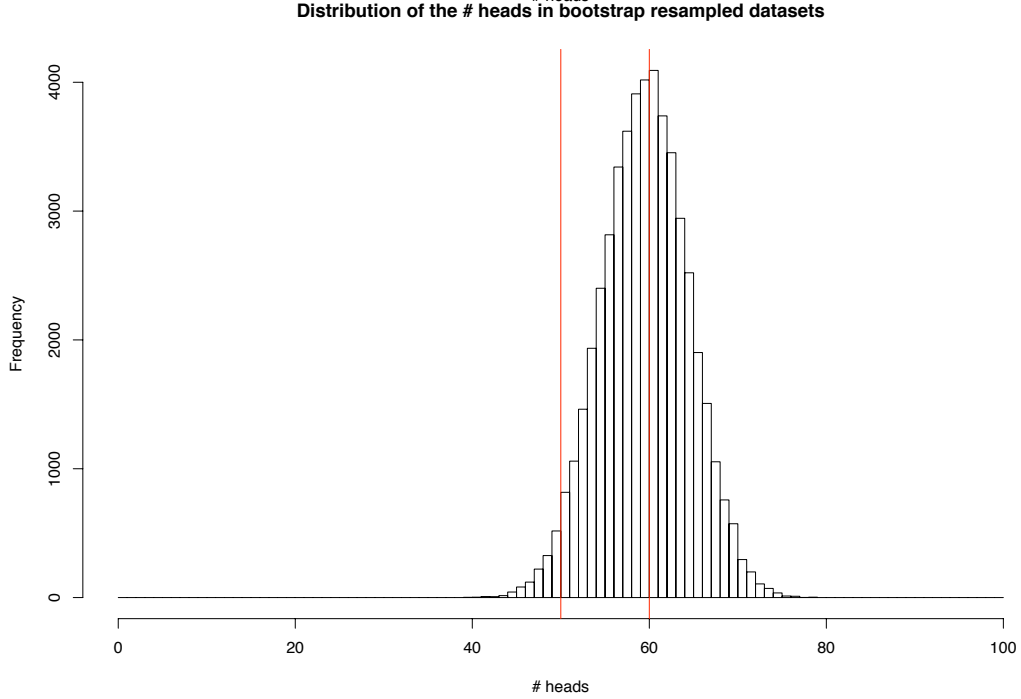
Amazingly (for many applications):

$$\Pr(\hat{f}_H \geq 0.6 \mid \text{null is true}) \approx \Pr(f_H^{(boot)} \leq 0.5)$$

In other words, the P-value is approximate by the fraction of bootstrap replicates consistent with the null.


Distribution of the # heads in bootstrap resampled datasets
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
heads
$\Pr(p^{(boot)} \leq 0.5) \approx 0.027$
$BP =$ frac. of $p^{(boot)} \geq 0.5$
$BP \approx 0.973$
P-value $\approx 1 - BP$


A simulation of the null distribution of the # heads
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
heads


Distribution of the # heads in bootstrap resampled datasets
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
heads

- When you decide between trees, the boundaries between tree hypotheses can be curved

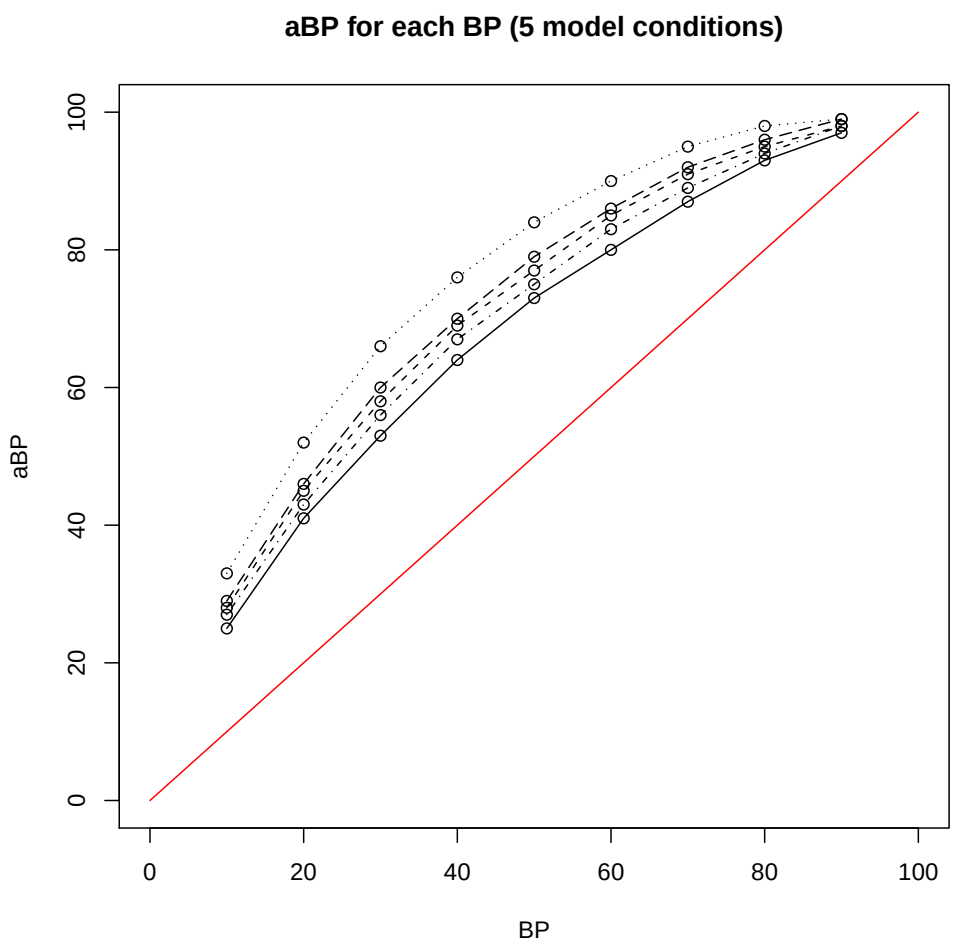
- When the boundary of the hypothesis space is curved, 1 - BP can be a poor approximation of the P-value. – Efron et al. (1996)

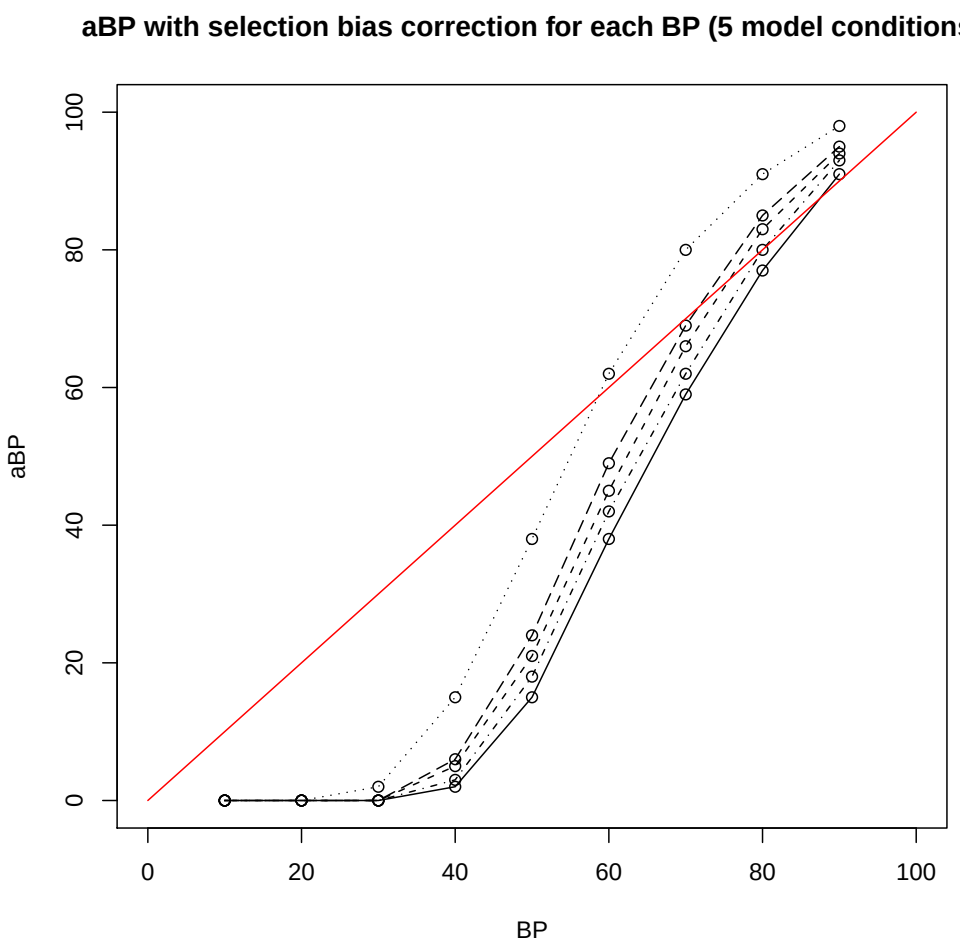
- Efron et al. (1996) proposed a computationally expensive multi-level bootstrap (which has not been widely used).

- Shimodaira (2002) used the same theoretical framework to devise a (more feasible) Approximately Unbiased (AU) test of topologies.
  - Multiple scales of bootstrap resampling (80% of characters, 90%, 100%, 110%...) are used to detect and correct for curvature of the boundary.
  - Implemented in the new versions of PAUP*

Susko (2010) adjusted BP – aBP

- Susko agrees with curvature arguments of Efron et al. (1996) and Shimodaira (2002), **but** points out that they ignore the **sharp point** in parameter space around the polytomy.
- He correct bootstrap proportions: $1 - aBP$ accurately estimates the P-value.
- The method uses the multivariate normal distributions the based on calculations about the curvature of the *likelihood* surface.
- You need to perform a different correction when you know the candidate tree *a priori* versus when you are putting BP on the ML tree.
- BP may **not** be conservative when you correct for selection bias.

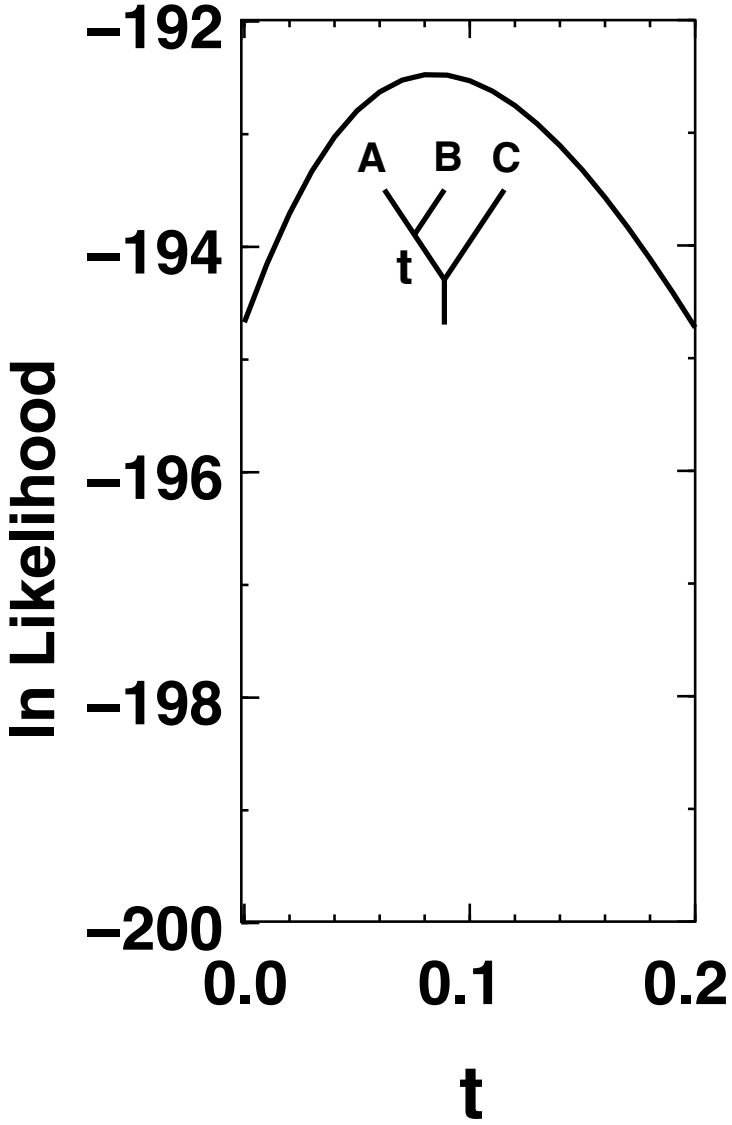

aBP for each BP (5 model conditions)
aBP
BP
0
20
40
60
80
100


aBP with selection bias correction for each BP (5 model conditions)
aBP
BP
0
20
40
60
80
100

Conclusions – bootstrapping

1. Non-parametric bootstrapping proportions help us see which branches have so little support that they could be plausibly explained by sampling error.
2. BPs are a little hard to interpret precisely.
3. Susko has and adjustment ("aBP") so that $1 - aBP \approx P$-value for the hypothesis that a recovered branch is not present in the true tree.

Can we test trees using the LRT?



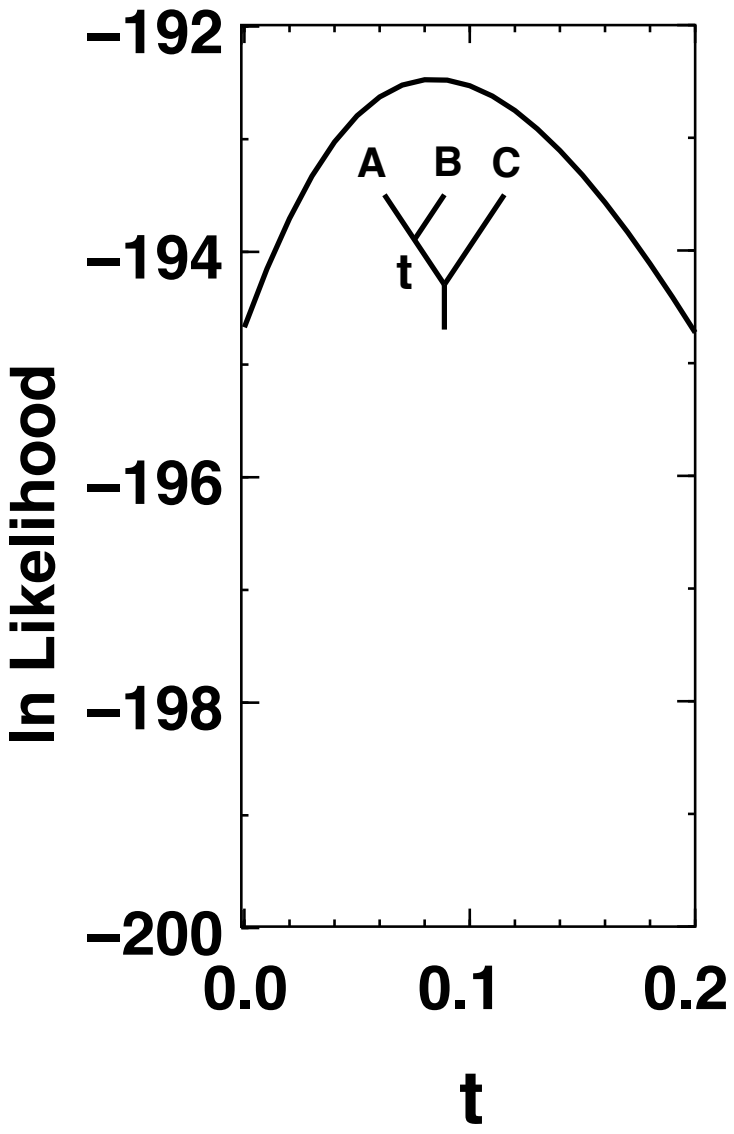
Slide from Joe Felsenstein

1. Should we calculate the LRT as:

$$\delta_i = 2\left[\ln L(t = \hat{t}, T_i \mid X) - \ln L(t = 0, T_i \mid X)\right]$$

2. And can we use the χ^2_1 distribution to get the critical value for δ?

Can we test trees using the LRT?



Slide from Joe Felsenstein

1. Should we calculate the LRT as:

$$\delta_i = 2\left[\ln L(t = \hat{t}, T_i \mid X) - \ln L(t = 0, T_i \mid X)\right]$$

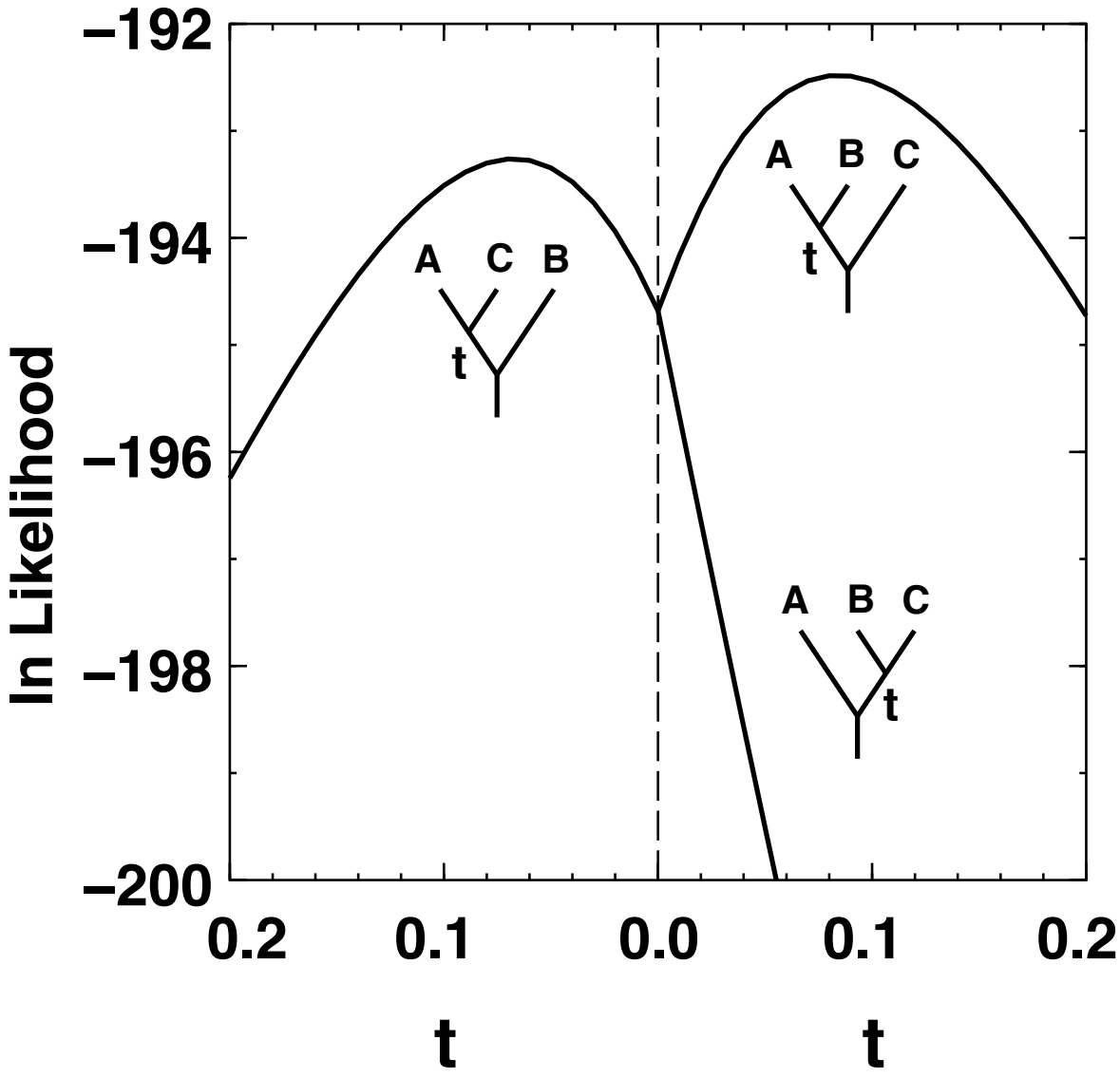
No. $t = 0$ might not yield the best alternative $\ln L$

2. And can we use the χ^2_1 distribution to get the critical value for δ ?

No. Constraining parameters at boundaries leads to a mixture such as: $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$

See Ota et al. (2000).

Can we test trees using the LRT?

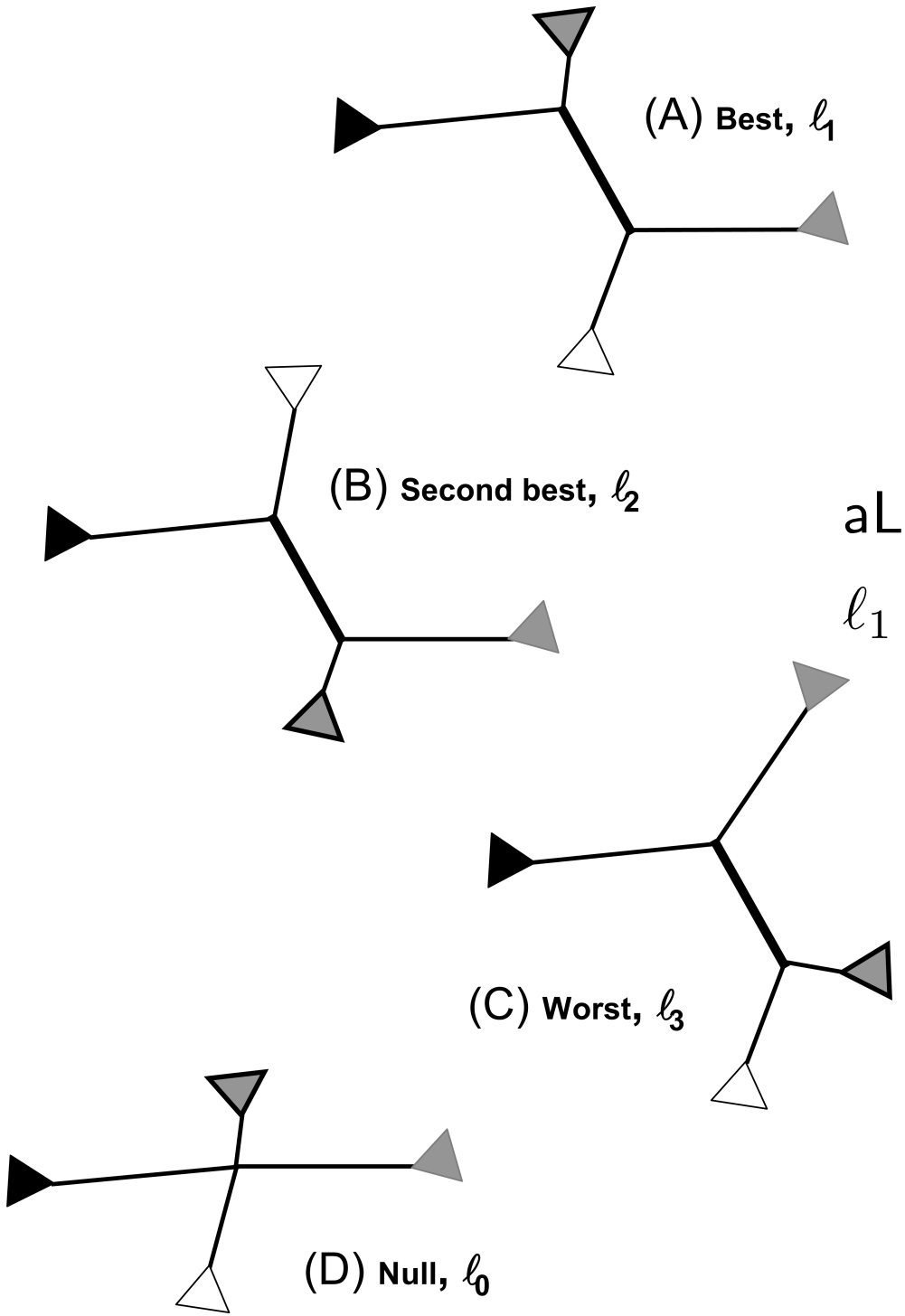


No, tree hypotheses are not nested!

Slide from Joe Felsenstein

aLRT of Anisimova and Gascuel (2006)

- For a **branch** j, calculate $\delta_j^\dagger$ as twice the difference in $\ln L$ between the optimal tree (which has the branch) and the best NNI neighbor.
- This is very fast.
- They argue that the null distribution for each LRT around the polytomy follows a $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ distribution
- The introduce Bonferroni-correction appropriate for correcting for the selection of the best of the three resolutions.
- They find aLRT to be accurate and powerful in simulations, but Anisimova et al. (2011) report that it rejects too often and is sensitive to model violation.



$$\text{aLRT} = 2\left[\ln \ell_1 - \ln L(T_2 \mid X)\right]$$

$$\ell_1 = L(T_1 \mid X)$$

Image from Anisimova and Gascuel (2006)

aBayes Anisimova et al. (2011)

$$\text{aBayes}(T_1 \mid X) = \frac{\Pr(X \mid T_1)}{\Pr(X \mid T_1) + \Pr(X \mid T_2) + \Pr(X \mid T_3)}$$

Simulation studies of Anisimova et al. (2011) show it to have the best power of the methods that do not have inflated probability of falsely rejecting the null.

It is sensitive to model violation.

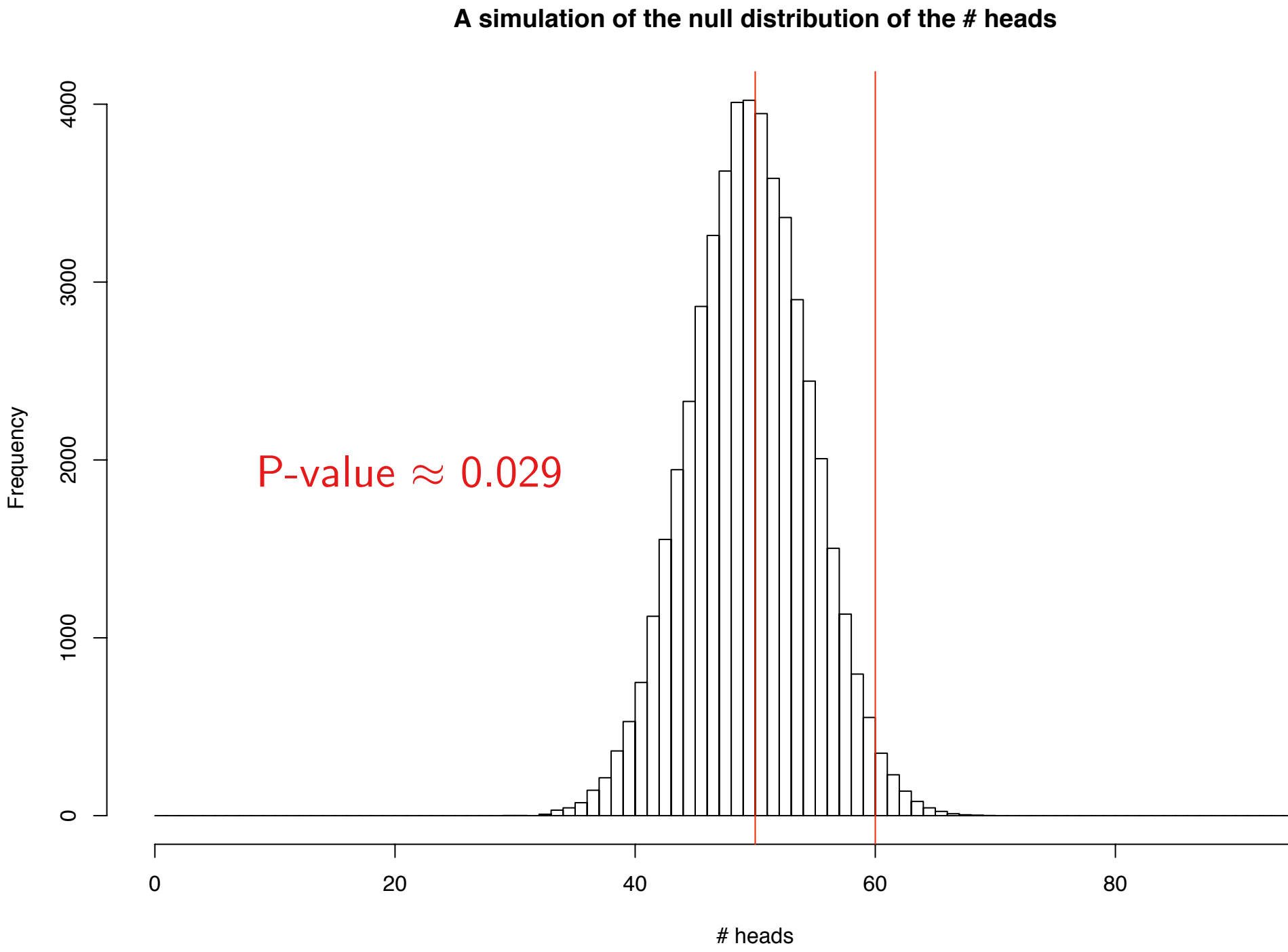
This is similar to "likelihood-mapping" of Strimmer and von Haeseler (1997)

coin flipping example (again, for inspiration)

$N = 100$ and $H = 60$

Can we reject the hypothesis of a fair coin?

We can use simulation to generate the null distribution (we could actually use the binomial distribution to analytically solve this one)...


A simulation of the null distribution of the # heads
P-value ≈ 0.029
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
heads

The simplest phylogenetic test would compare two trees

Null: If we had no sampling error (infinite data) T_1 and T_2 would explain the data equally well.

Test Statistic:

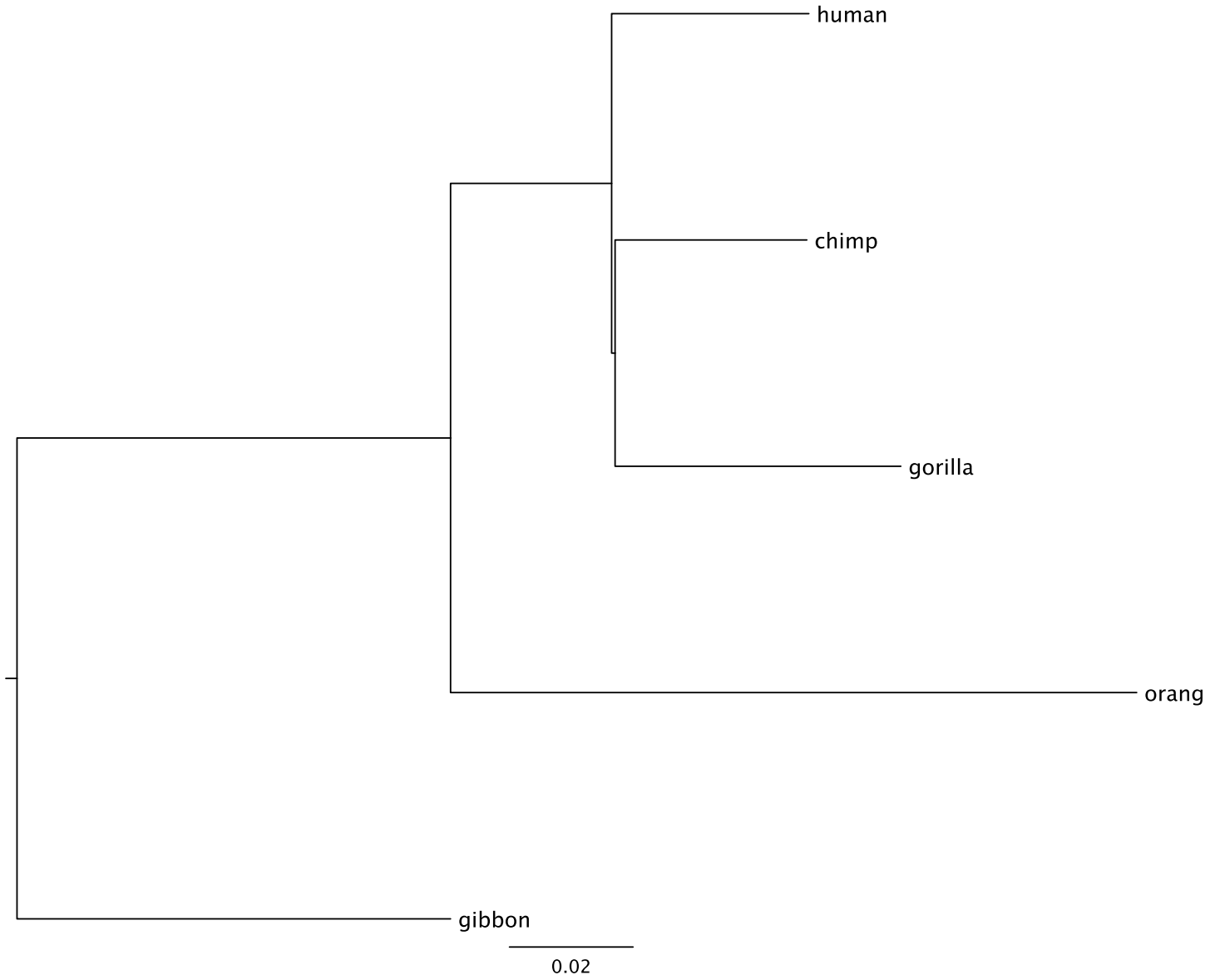
$$\delta(T_1, T_2 \mid X) = 2\left[\ln L(T_1 \mid X) - \ln L(T_2 \mid X)\right]$$

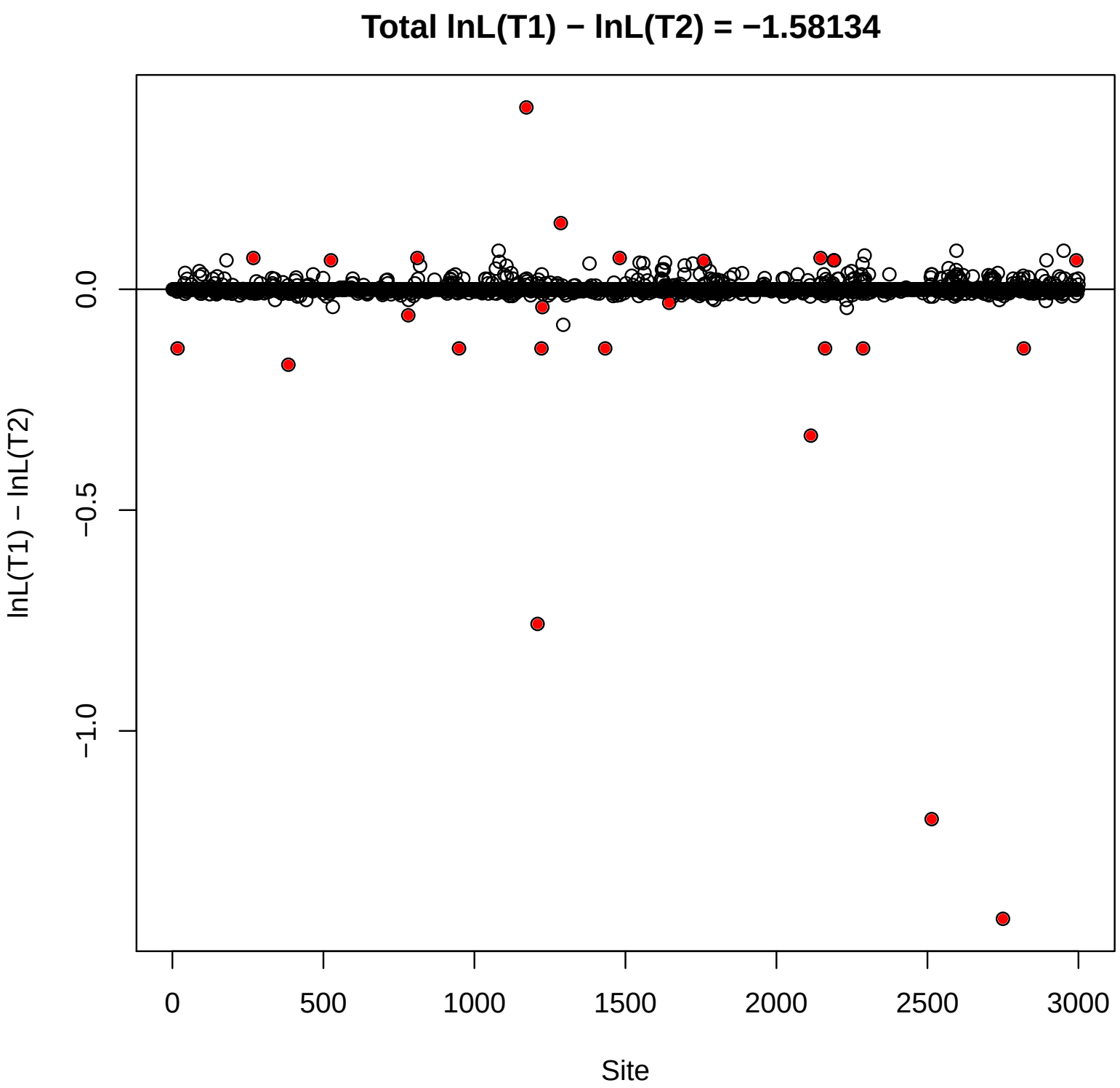
Expectation under null:

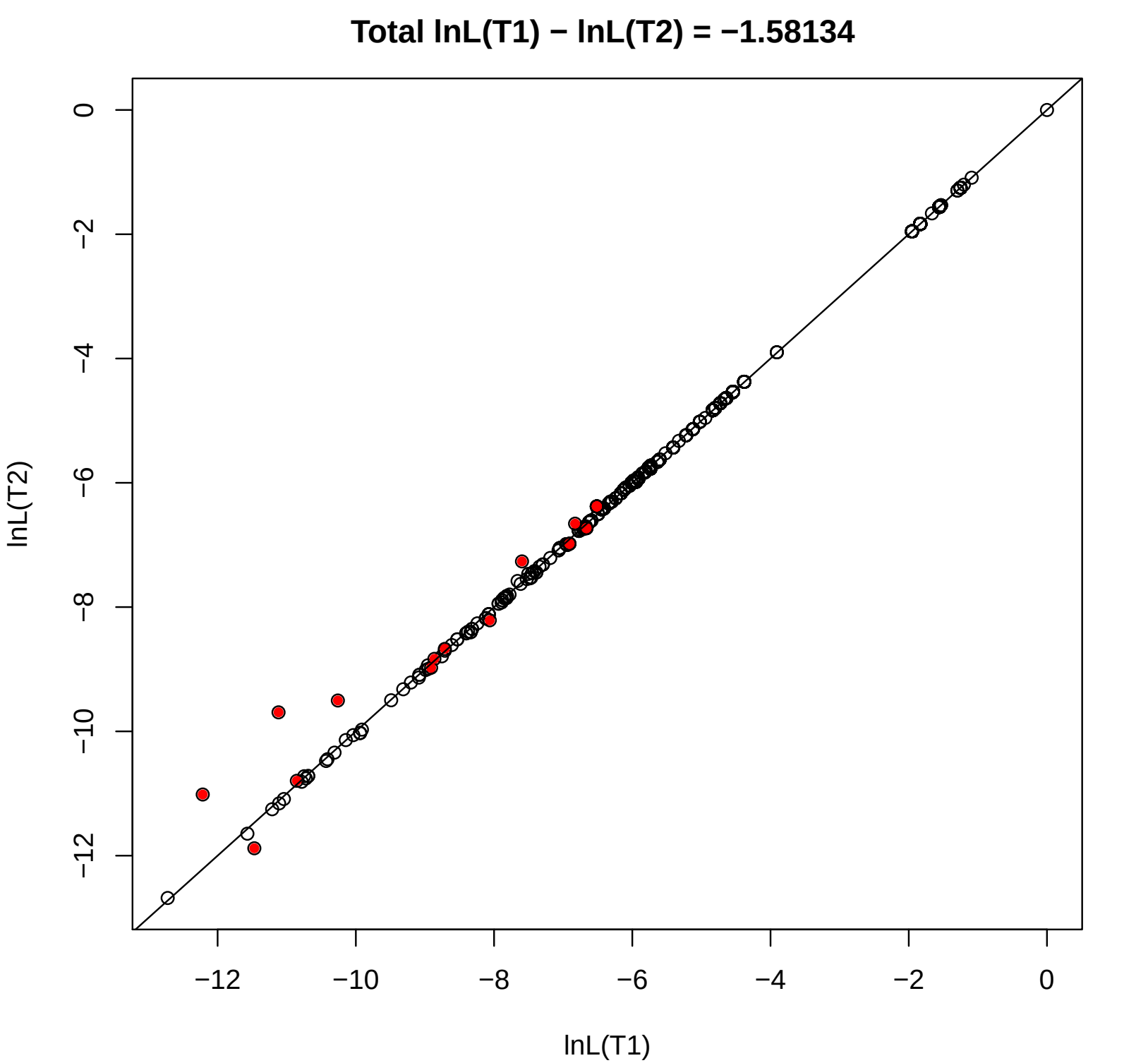
$$\mathbb{E}_{H_0}\left[\delta(T_1, T_2 \mid X)\right] = 0$$

Using 3000 sites of mtDNA sequence for 5 primates

T_1 is ((chimp, gorilla), human)




Total lnL(T1) − lnL(T2) = −1.58134
lnL(T1) − lnL(T2)
0.0
−0.5
−1.0
0
500
1000
1500
2000
2500
3000
Site

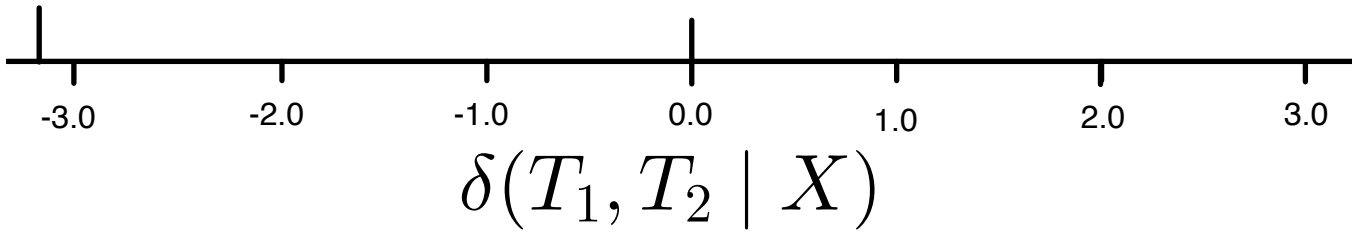

Total lnL(T1) − lnL(T2) = −1.58134
lnL(T2)
lnL(T1)
0
−2
−4
−6
−8
−10
−12
−12
−10
−8
−6
−4
−2
0

Using 3000 sites of mtDNA sequence for 5 primates

T_1 is ((chimp, gorilla), human) $\quad \ln L(T_1 \mid X) = -7363.296$

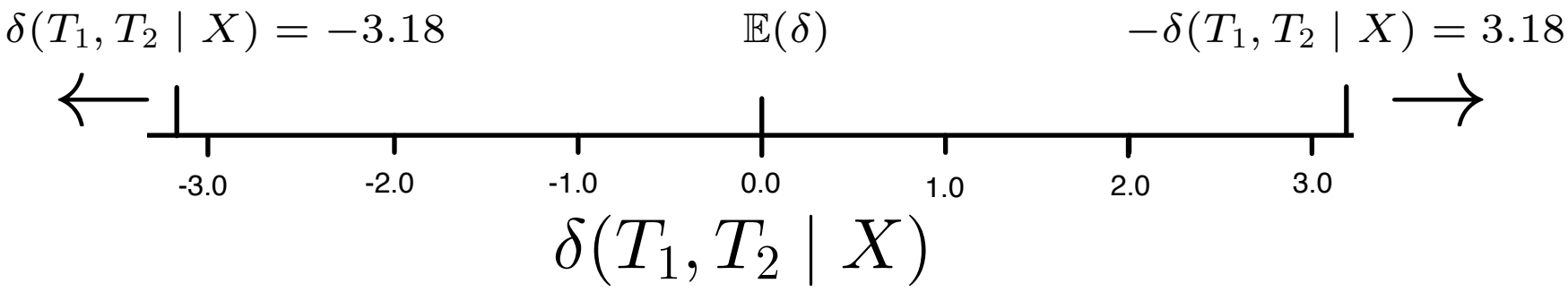
T_2 is ((chimp, human), gorilla) $\quad \ln L(T_2 \mid X) = -7361.707$

$\delta(T_1, T_2 \mid X) = -3.18$ $\quad \mathbb{E}(\delta)$



To get the P-value, we need to know the probability:

$$\Pr\left(\left|\delta(T_1, T_2 \mid X)\right| \geq 3.18 \middle| H_0 \text{ is true}\right)$$



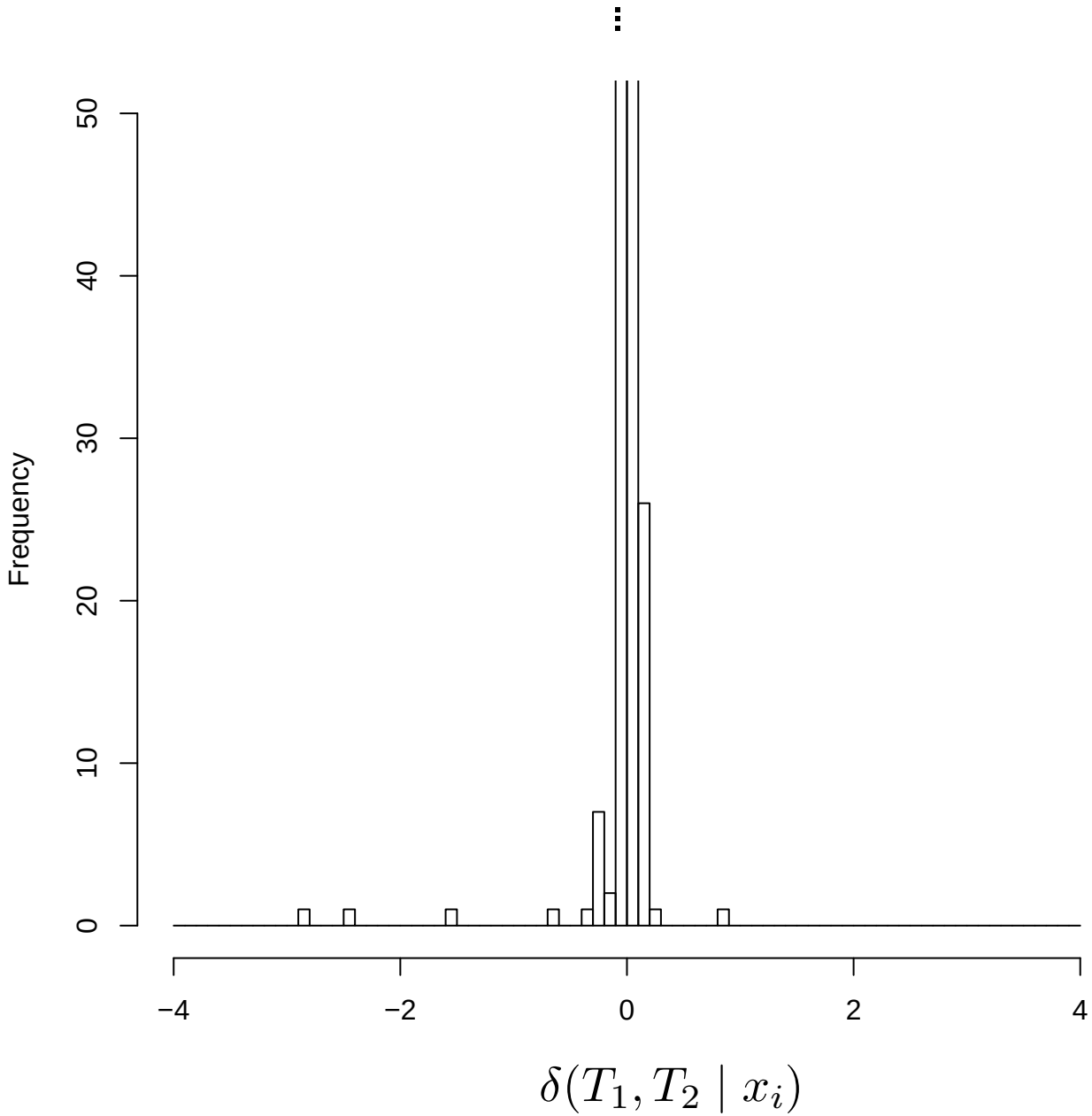
KH Test

1. Examine the difference in $\ln L$ for each site: $\delta(T_1, T_2 \mid X_i)$ for site i.
2. Note that the total difference is simply a sum:

$$\delta(T_1, T_2 \mid X) = \sum_{i=1}^{M} \delta(T_1, T_2 \mid X_i)$$

3. The variance of $\delta(T_1, T_2 \mid X)$ will be a function of the variance in "site" $\delta(T_1, T_2 \mid X_i)$ values.

$\delta(T_1, T_2 \mid X_i)$ for each site, i.



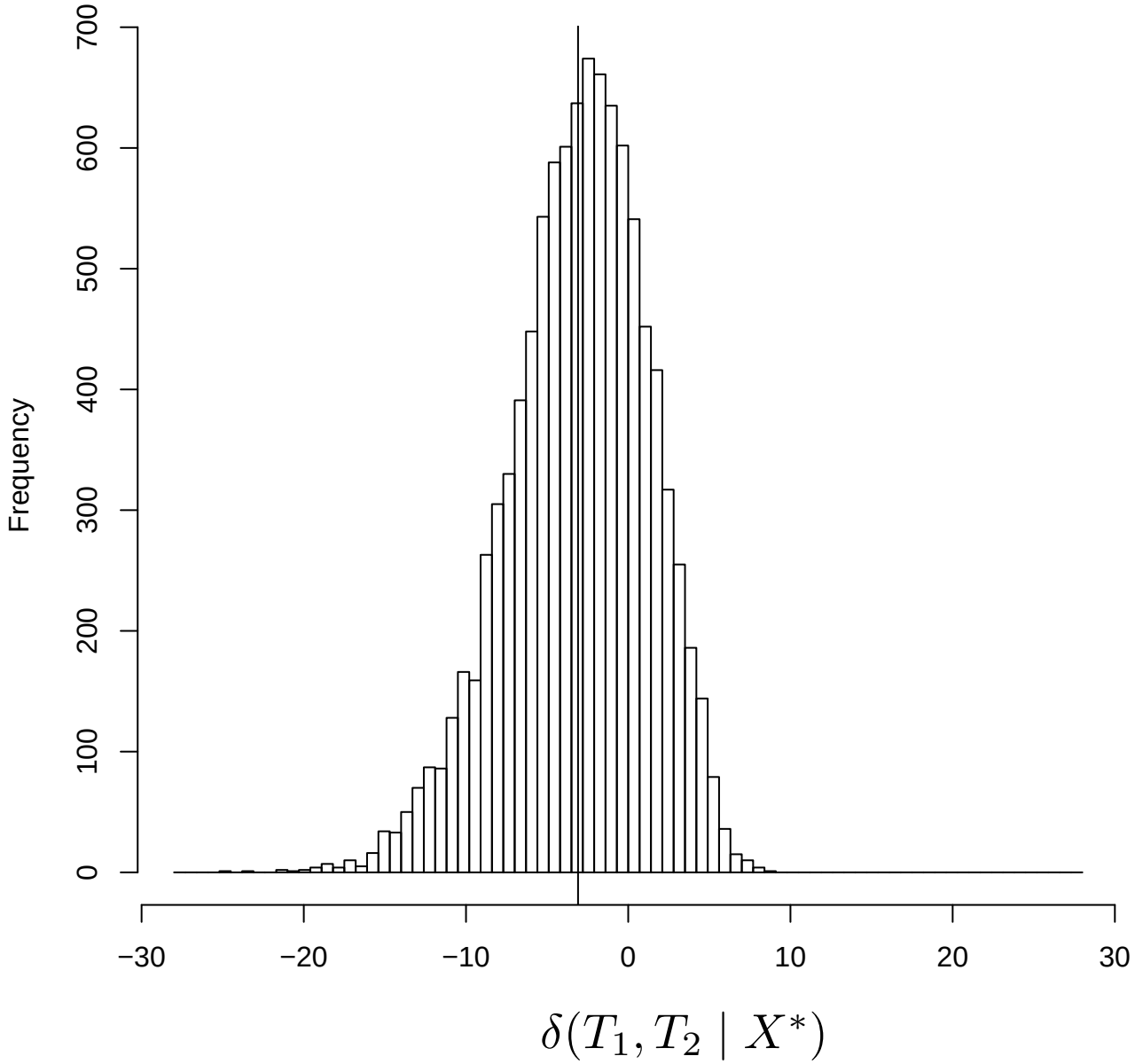
KH Test - the variance of $\delta(T_1, T_2 \mid X)$

To approximate variance of $\delta(T_1, T_2 \mid X)$ under the null, we could:

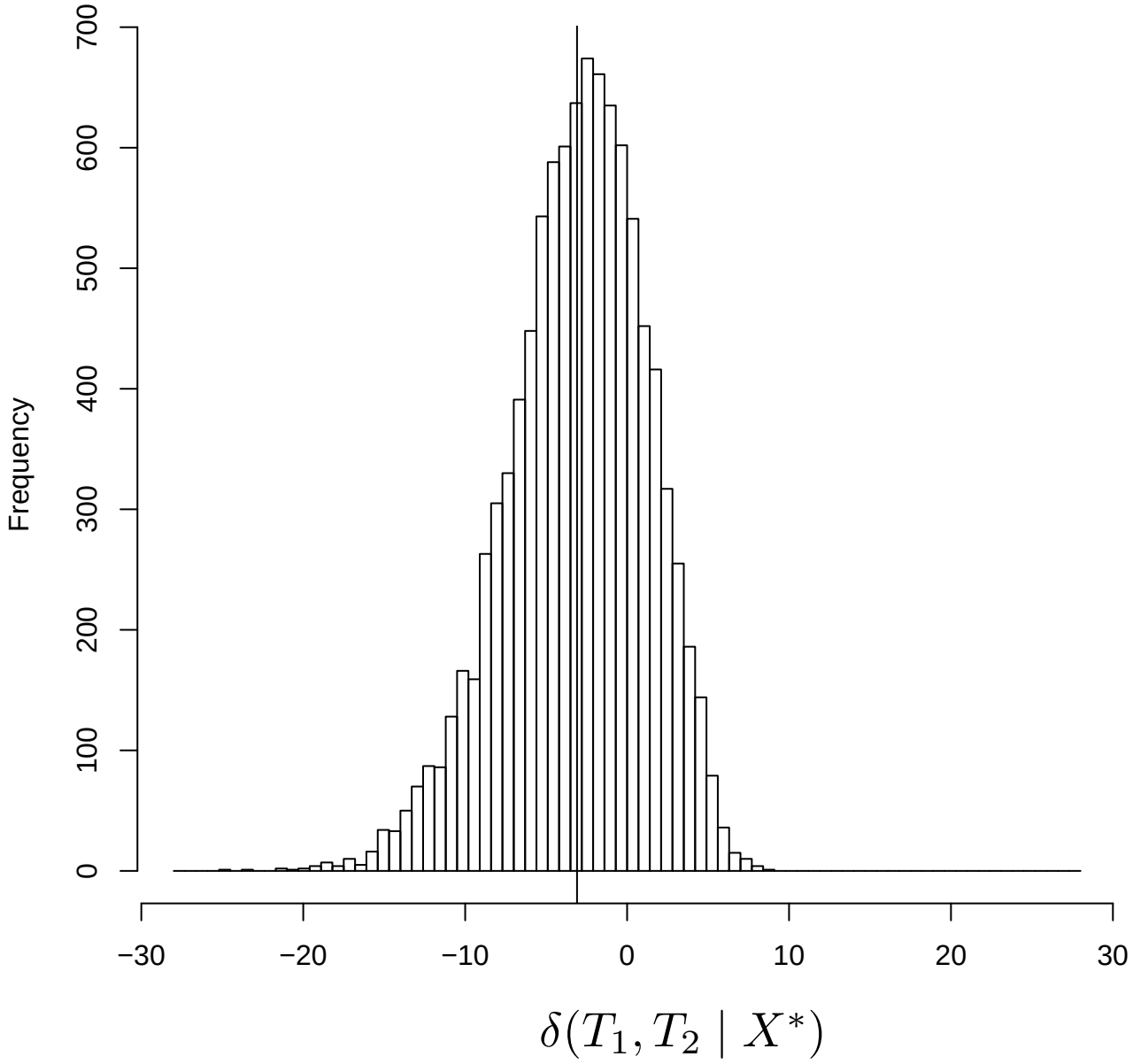
1. use assumptions of Normality (by appealing to the Central Limit Theorem[1]). Or
2. use bootstrapping to generate a cloud of pseudo-replicate $\delta(T_1, T_2 \mid X^*)$ values, and look at their variance.

[1] Susko (2014) recently showed that this is flawed and too conservative.

δ for many (RELL) bootstrapped replicates of the data



The (RELL) bootstrapped sample of statistics.
Is this the null distribution for our δ test statistic?



KH Test - 'centering'

$$\mathbb{E}_{H_0}\left[\delta(T_1, T_2 \mid X)\right] = 0$$

Bootstrapping gives us a reasonable guess of the variance under H_0

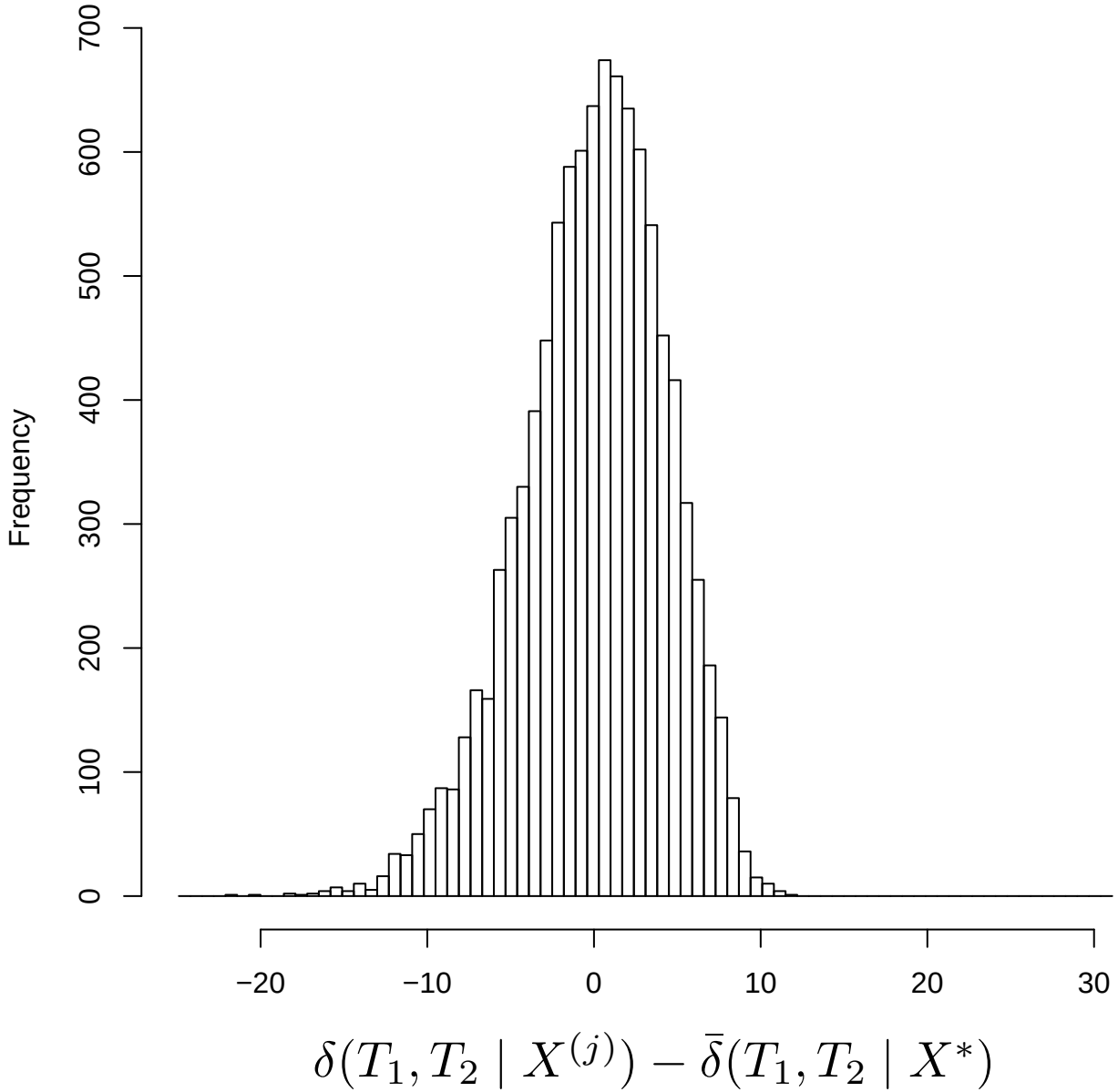
Subtracting the mean of the bootstrapped $\delta(T_1, T_2 \mid X^*)$ values gives the null distribution.

For each of the j bootstrap replicates, we treat

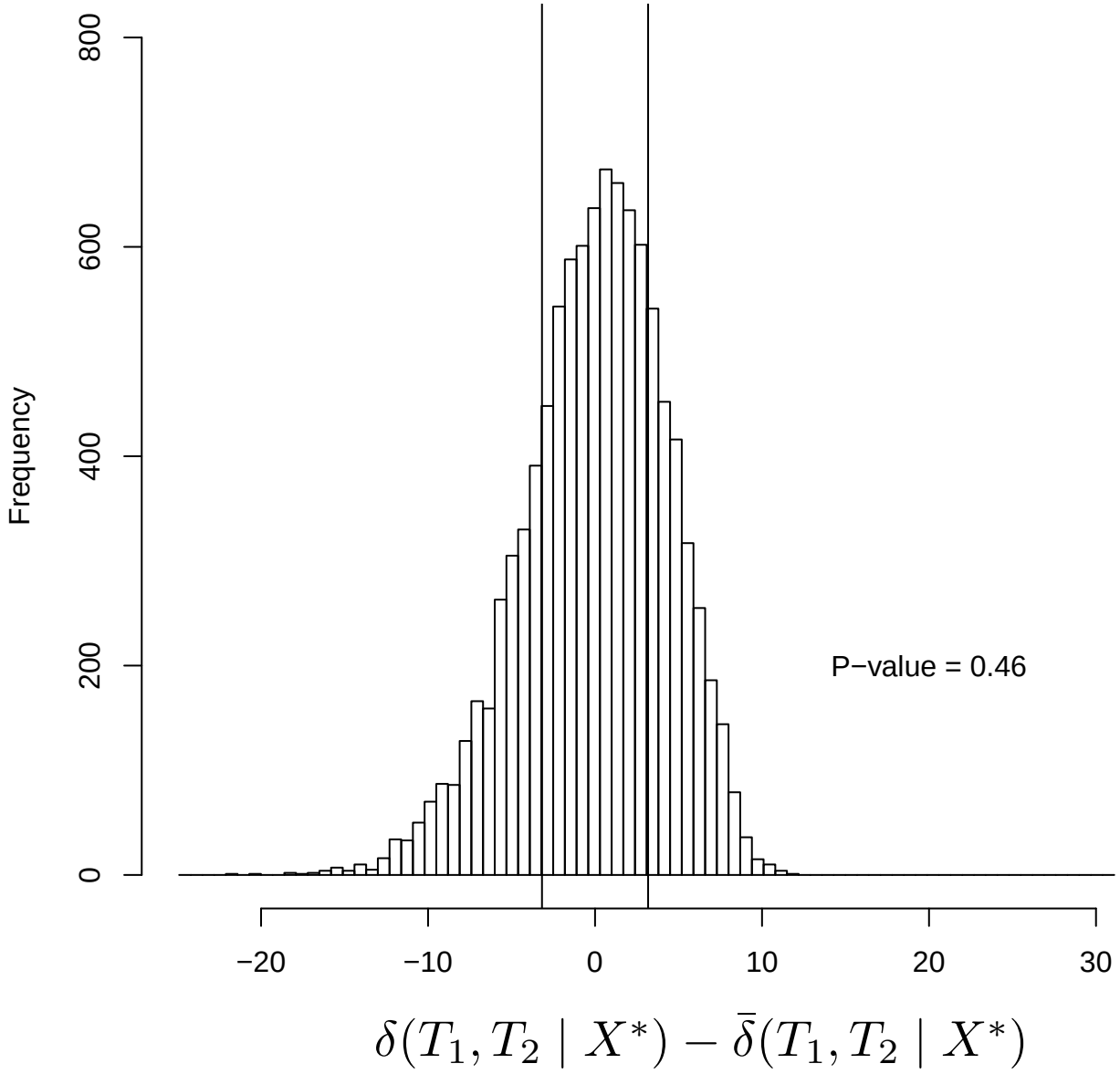
$$\delta(T_1, T_2 \mid X^{*j}) - \bar{\delta}(T_1, T_2 \mid X^*)$$

as draws from the null distribution.

$\delta(T_1, T_2 \mid X^{(j)}) - \bar{\delta}(T_1, T_2 \mid X^*)$
for many (RELL) bootstrapped replicates of the data



Approximate null distribution with tails (absolute value ≥ 3.18) shown



Other ways to assess the null distribution of the LR test statistic

- Bootstrapping then centering LR, and
- Using normality assumptions.

are both clever and cute solutions.

They are too conservative (Susko, 2014) - more complicated calculations from the Normal [KHns] or mixtures of χ^2 distributions [chi-bar].

They do not match the null distribution under any model of sequence evolution.

Mini-summary

- $\delta(T_1, T_2 \mid X) = 2\left[\ln L(T_1 \mid X) - \ln L(T_2 \mid X)\right]$ is a powerful statistic for discrimination between trees.
- We can assess confidence by considering the variance in signal between different characters.
- Bootstrapping helps us assess the variance in $\ln L$ that we would expect to result from sampling error.

Scenario

1. A (presumably evil) competing lab scoops you by publishing a tree, T_1, for your favorite group of organisms.

2. You have just collected a new dataset for the group, and your ML estimate of the best tree, T_2, differ's from T_1.

3. A KH Test shows that your data **significantly** prefer T_2 over T_1.

4. You write a (presumably scathing) response article.

Should a *Systematic Biology* publish your response?

What if start out with only one hypothesized tree, and we want to compare it to the ML tree?

The KH Test is **NOT** appropriate in this context (see Goldman et al., 2000, for discussion of this point)

Multiple Comparisons: lots of trees increases the variance of $\delta(\hat{T}, T_1 \mid X)$

Selection bias: Picking the ML tree to serve as one of the hypotheses invalidates the centering procedure of the KH test.

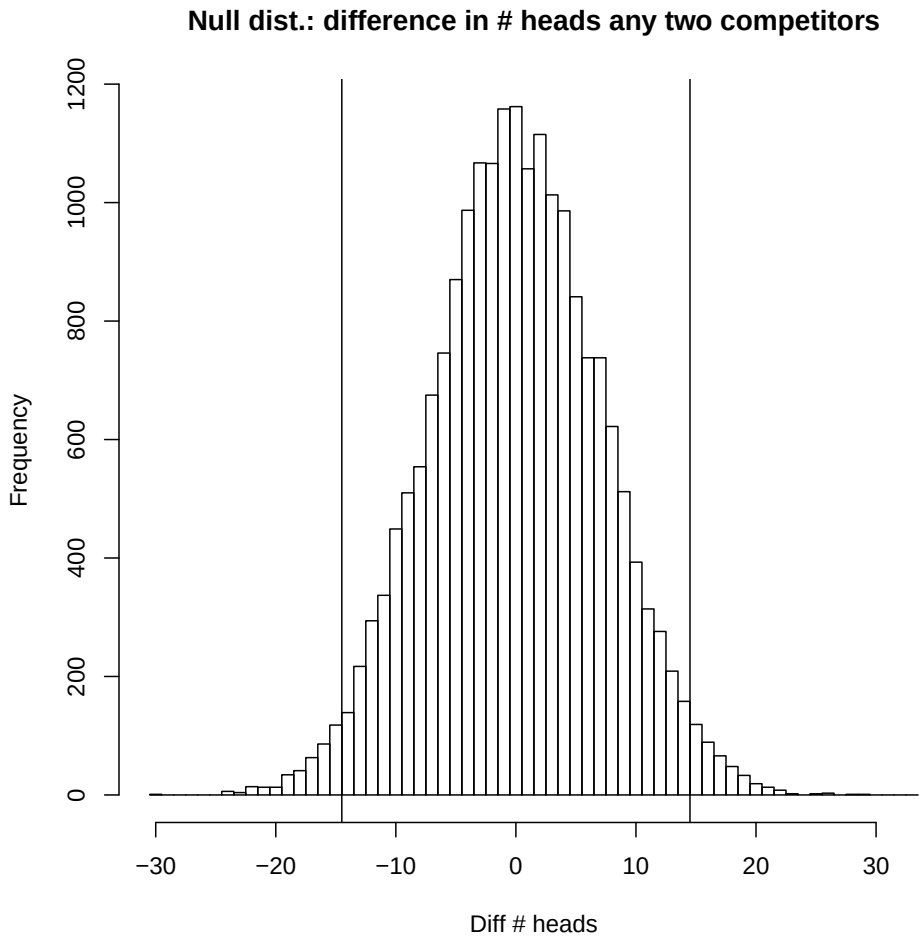
Using the ML tree in your test introduces selection bias

Even when the H_0 is true, we do not expect $2\left[\ln L(\hat{T}) - \ln L(T_1)\right] = 0$

Imagine a competition in which a large number of equally skilled people compete, and you compare the score of one competitor against the highest scorer.

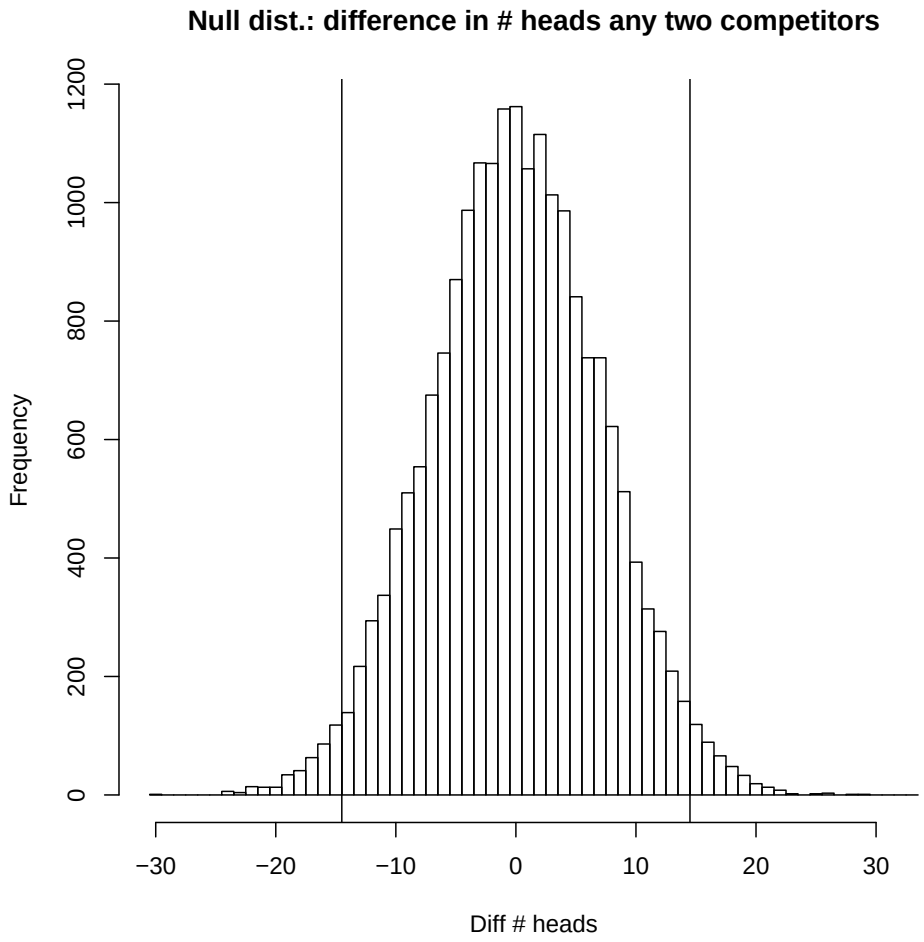
Experiment: 70 people each flip a fair coin 100 times and count # heads.

$$h_1 - h_2$$

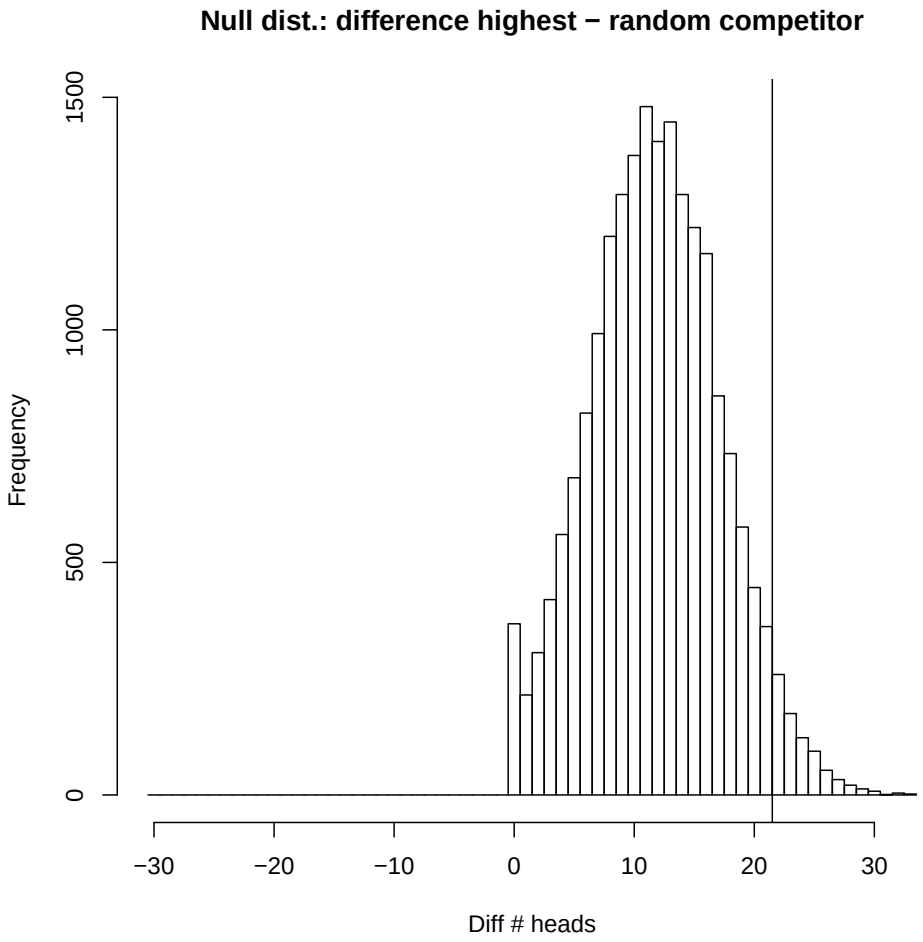


Experiment: 70 people each flip a fair coin 100 times and count # heads.

$h_1 - h_2$



$\max(h) - h_1$



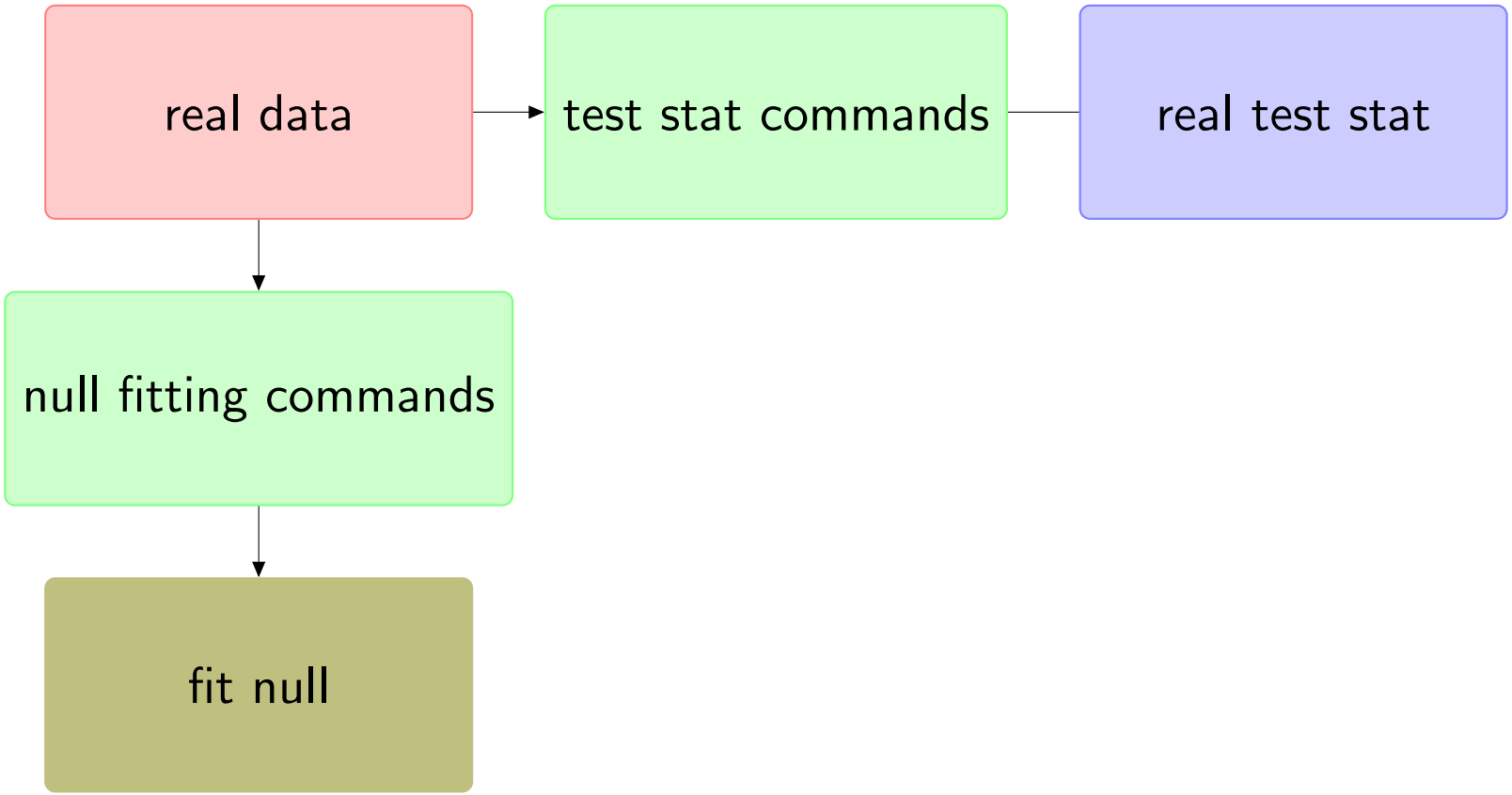
Shimodaira and Hasegawa proposed the SH test which deals the "selection bias" introduced by using the ML tree in your test

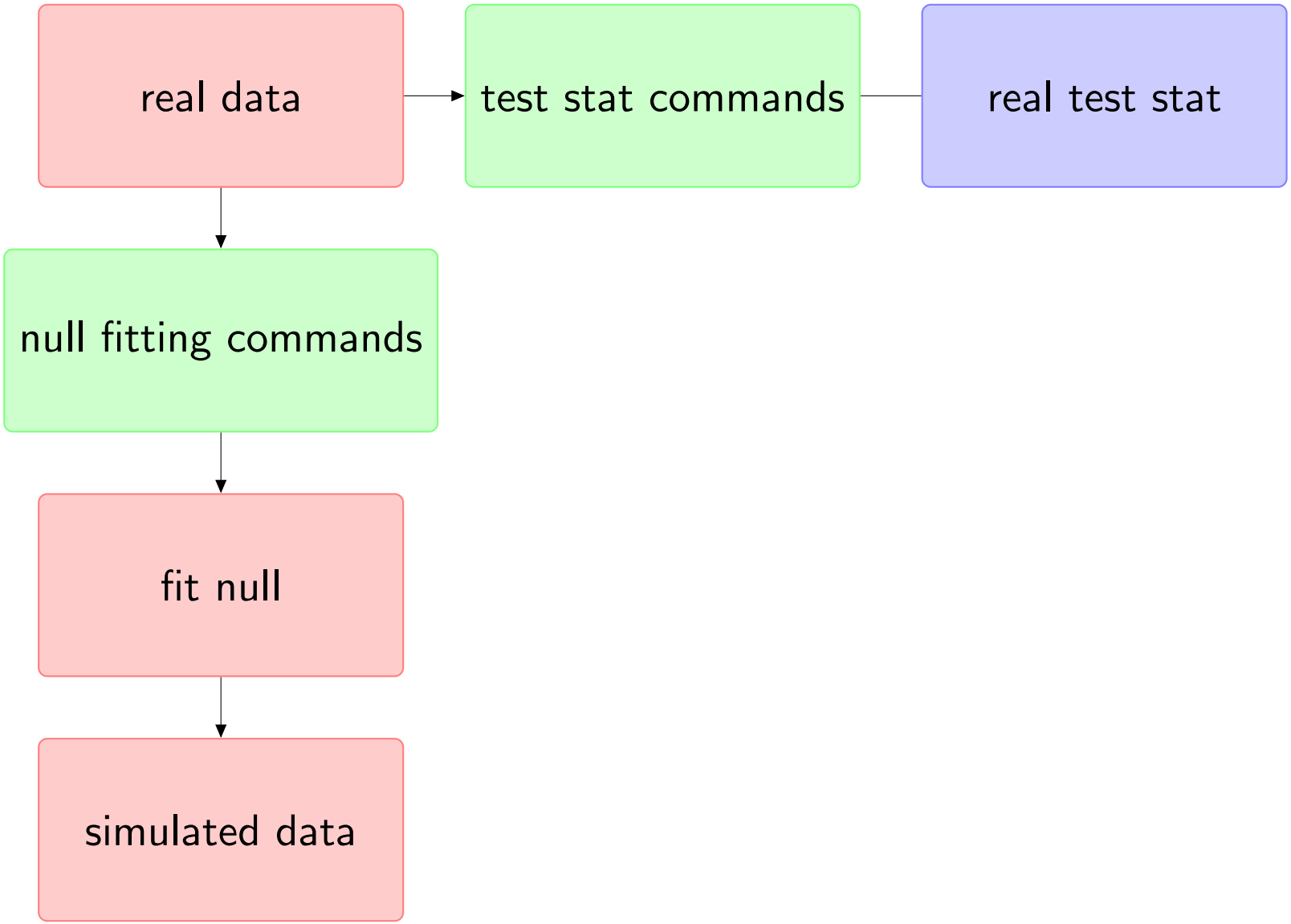
Requires **set of candidate trees** - these **must not** depend on the dataset to be analyzed.

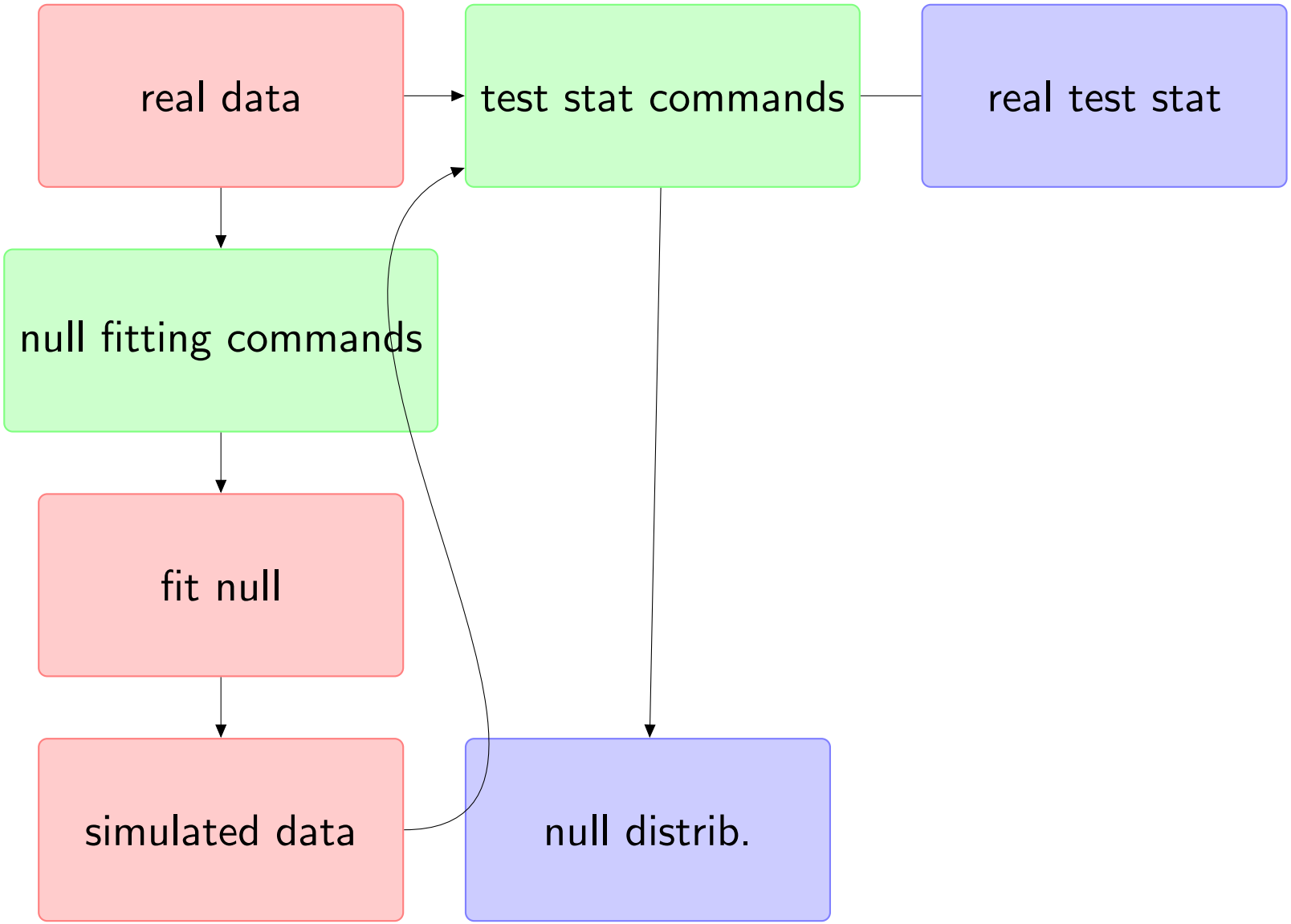
H_0: each tree in the candidate set is as good as the other trees.

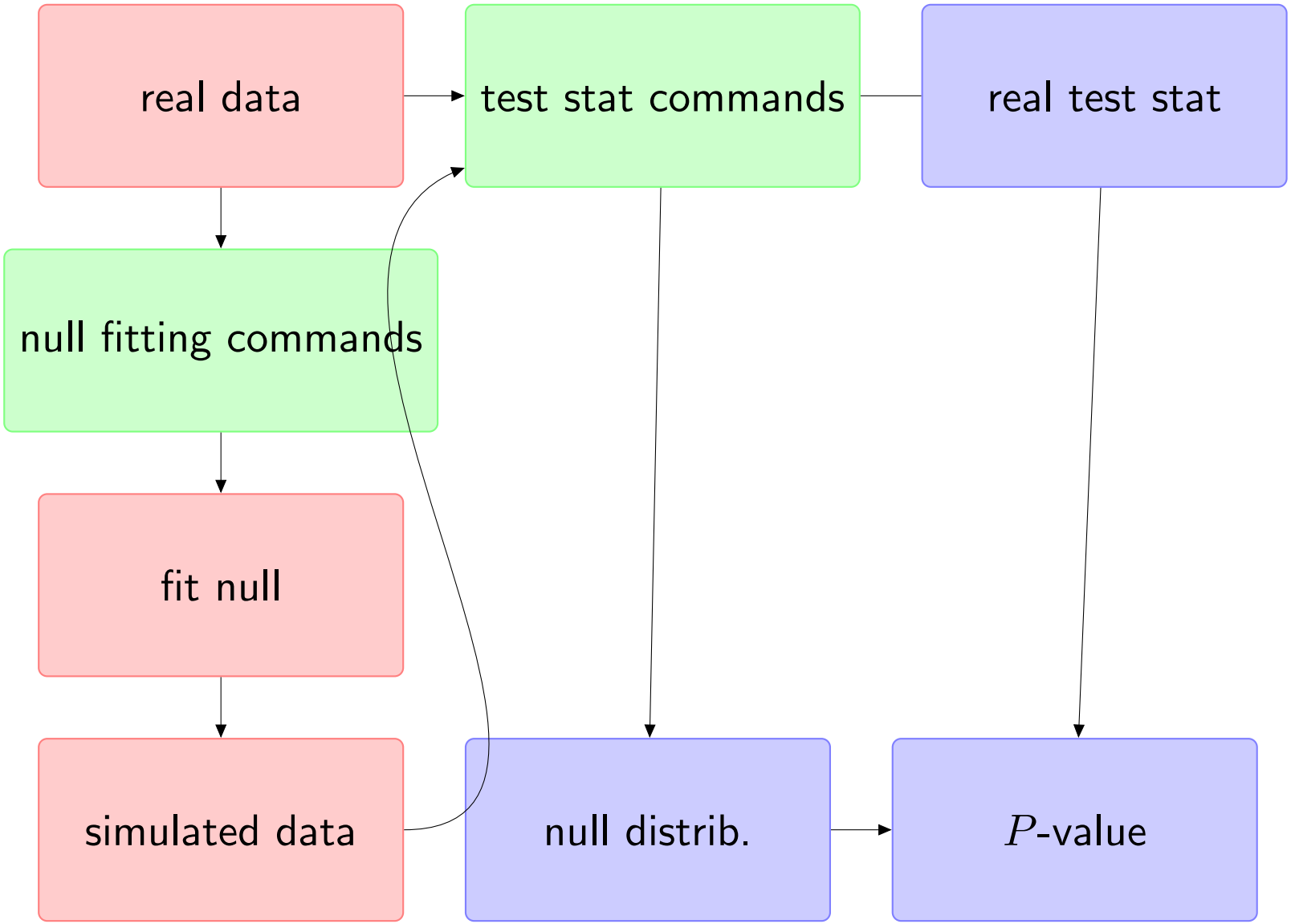
The test makes worst-case assumptions - it **is conservative**.

AU test is less conservative (still needs a candidate set)


real data
test stat commands
real test stat
null fitting commands
fit null


real data
test stat commands
real test stat
null fitting commands
fit null
simulated data


real data
test stat commands
real test stat
null fitting commands
fit null
simulated data
null distrib.


real data
test stat commands
real test stat
null fitting commands
fit null
simulated data
null distrib.
P-value

Parametric bootstrapping to generate the null distribution for the LR statistic

1. find the best tree and model pair that are consistent with the null,

2. Simulate many datasets under the parameters of that model,

3. Calculate $\delta^{(j)} = 2\left[\ln L(\hat{T}^{(j)} \mid X^{(j)}) - \ln L(\hat{T}_0^{(j)} \mid X^{(j)})\right]$ for each simulated dataset.

   - the (j) is just an index for the simulated dataset,
   - $\hat{T}_0^{(j)}$ is the tree under the null hypothesis for simulation replicate j

Parametric bootstrapping

This procedure is often referred to as SOWH test (in that form, the null tree is specified *a priori*).

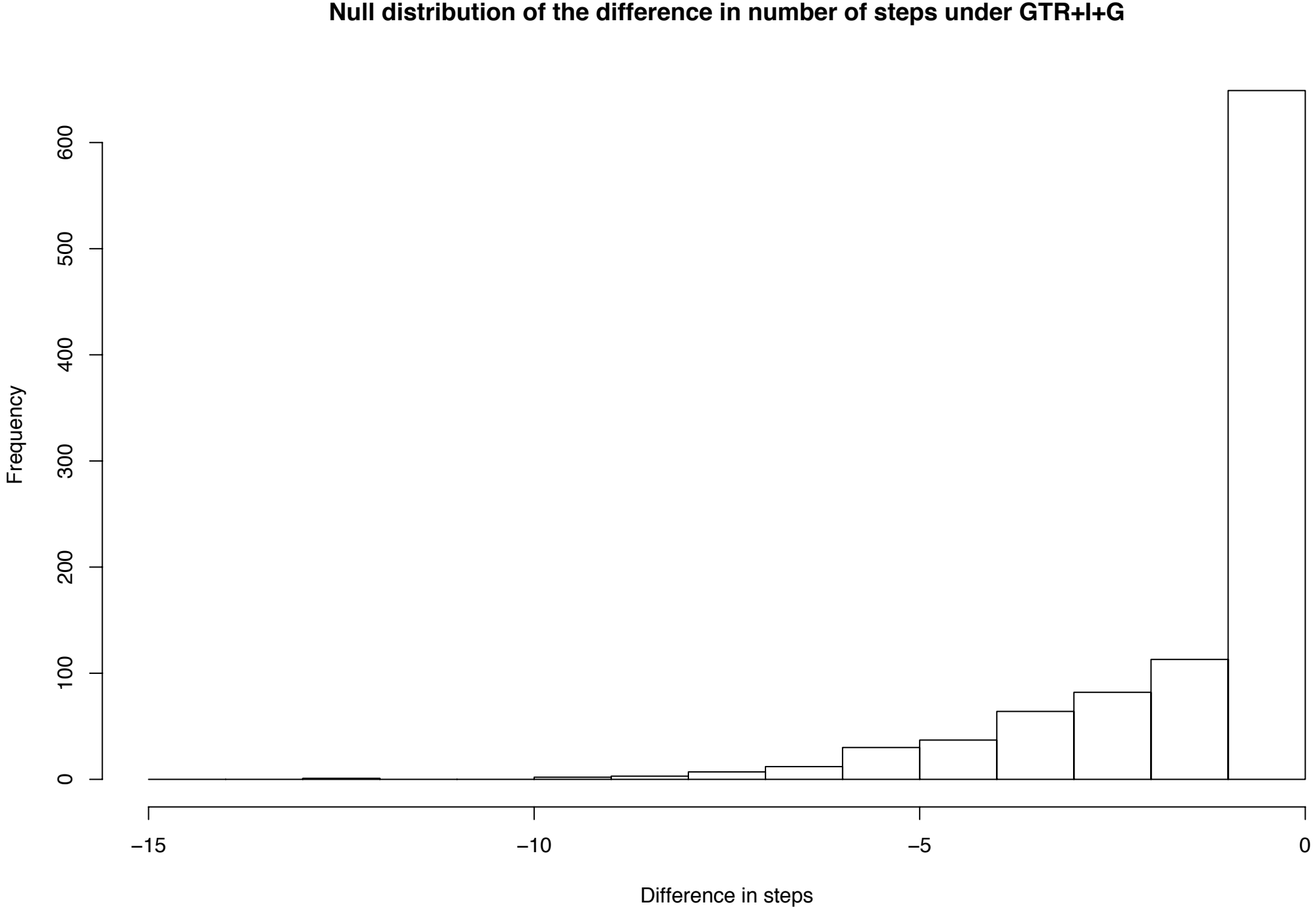
Huelsenbeck et al. (1996) describes how to use the approach as a test for monophyly.

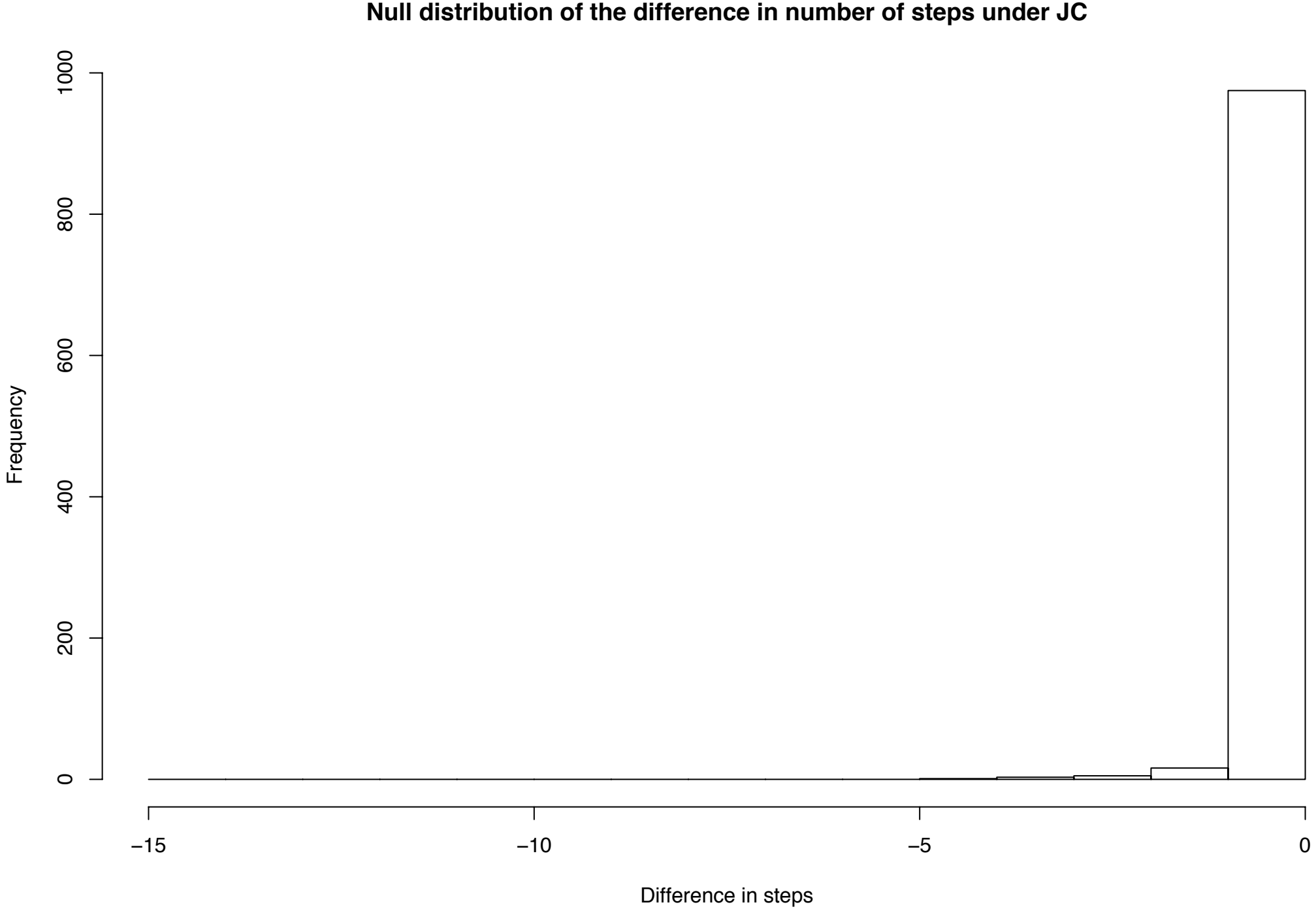
Intuitive and powerful, but not robust to model violation (Buckley, 2002).

Can be done manually[2] or via SOWHAT by Church et al. (2015). Optional demo here.

Susko (2014): collapse optimize null tree with 0-length contraints for the branch in question (to avoid rejecting too often)

[2] instructions in https://molevol.mbl.edu/index.php/ParametricBootstrappingLab


Null distribution of the difference in number of steps under GTR+I+G
Frequency
0
100
200
300
400
500
600
−15
−10
−5
0
Difference in steps

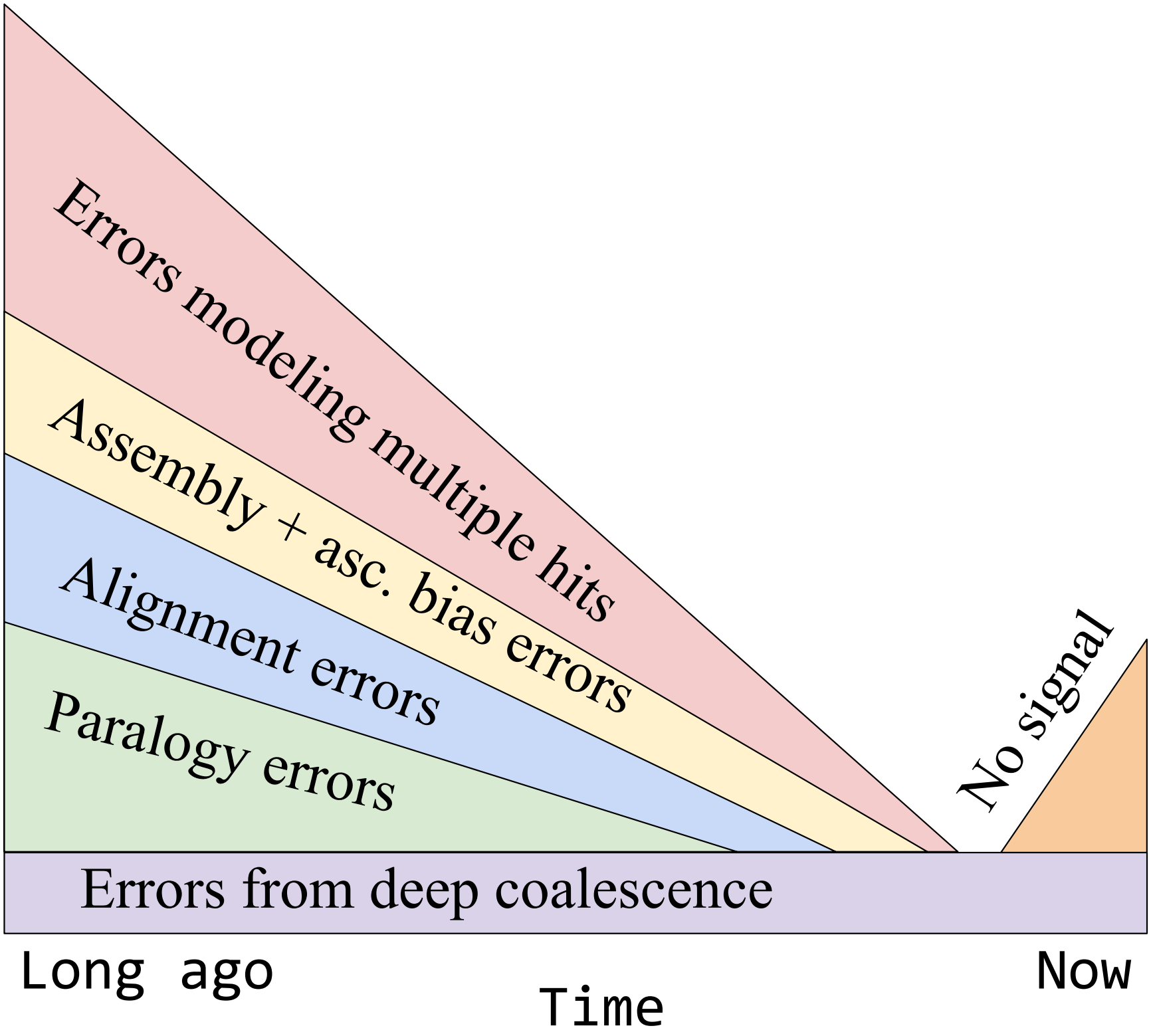

Null distribution of the difference in number of steps under JC
Frequency
0
200
400
600
800
1000
−15
−10
−5
0
Difference in steps

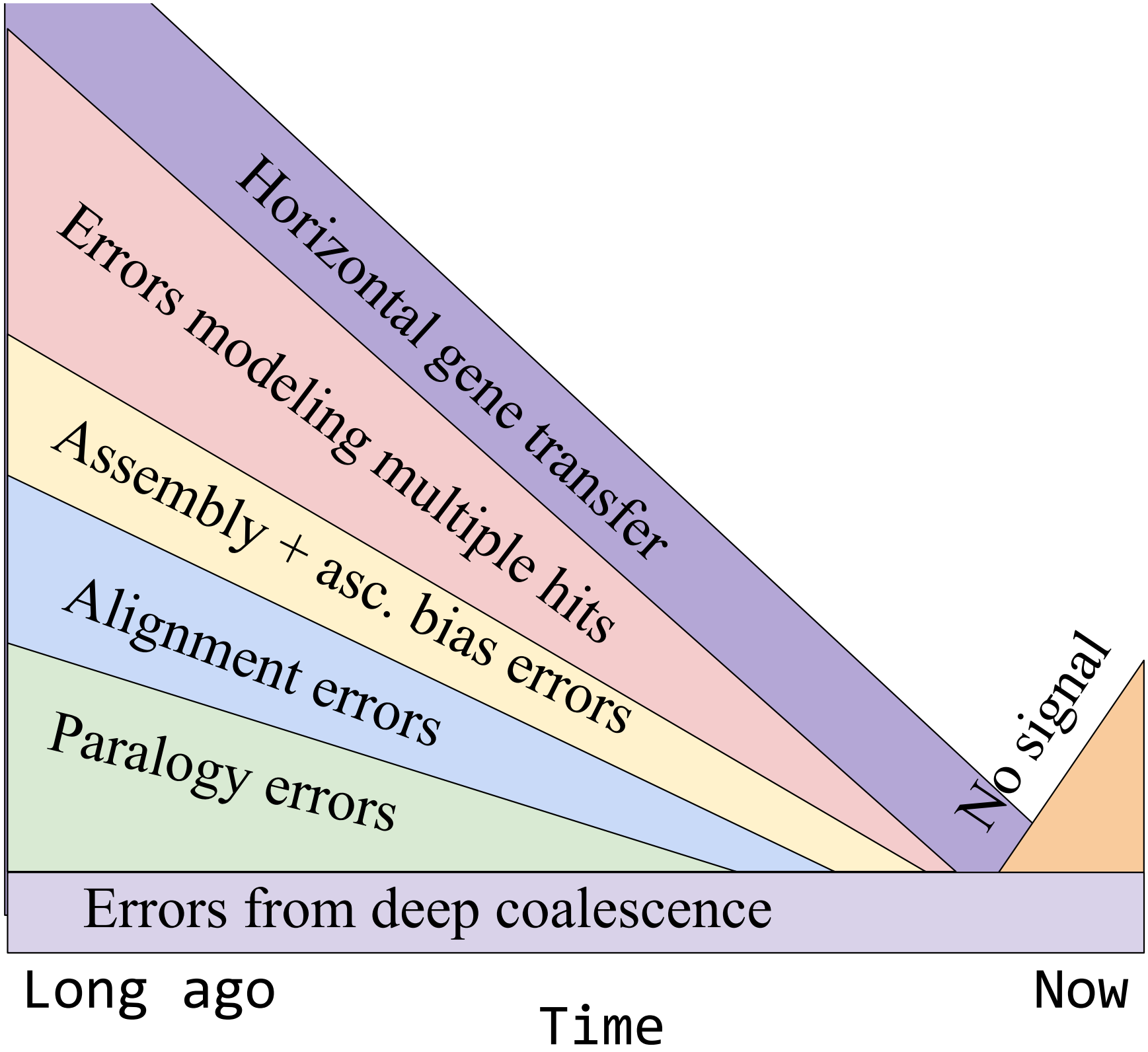
We often don't want to test tree topologies

- If we are conducting a "comparative method" we have to consider phylogenetic history,
- ideally we would integrate out the uncertainty in the phylogeny

Tree is a "nuisance parameter"

You have to think about what sources of error are most relevant for **your** data!


Errors modeling multiple hits
Assembly + asc. bias errors
Alignment errors
Paralogy errors
No signal
Errors from deep coalescence
Long ago
Time
Now


Horizontal gene transfer
Errors modeling multiple hits
Assembly + asc. bias errors
Alignment errors
Paralogy errors
No signal
Errors from deep coalescence
Long ago
Time
Now

Open Tree of Life slides

References

Anisimova, M. and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology*, 55(4):539–552.

Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C., and Gascuel, O. (2011). Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Systematic Biology*.

Buckley, T. R. (2002). Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Systematic Biology*, 51(3):509–523.

Church, S. H., Ryan, J. F., and Dunn, C. W. (2015). Automation and evaluation of the sowh test with sowhat. *Systematic Biology*, 64(6):1048–1058.

Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Science, U. S. A.*, 93:13429–13434.

Goldman, N., Anderson, J. P., and Rodrigo, A. G. (2000). Likelihood-based tests of topologies in phylogenetics. *Systematic Biology*, 49:652–670.

Huelsenbeck, J., Hillis, D., and Nielsen, R. (1996). A Likelihood-Ratio Test of Monophyly. *Systematic Biology*, 45(4):546.

Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., De Oliveira, T., and Gascuel, O. (2017). Boosting felsenstein phylogenetic bootstrap. *bioRxiv*.

Ota, R., Waddell, P. J., Hasegawa, M., Shimodaira, H., and Kishino, H. (2000). Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Molecular Biology and Evolution*, 17(5):798–803.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51(3):492–508.

Strimmer, K. and von Haeseler, A. (1997). Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 94(13):6815–6819.

Susko, E. (2010). First-order correct bootstrap support adjustments for splits that allow hypothesis testing when using maximum likelihood estimation. *Molecular Biology and Evolution*, 27(7):1621–1629.

Susko, E. (2014). Tests for two trees using likelihood methods. *Molecular Biology and Evolution*.

Wright, A. M., Lyons, K. M., Brandley, M. C., and Hillis, D. M. (2015). Which came first: The lizard or the egg? robustness in phylogenetic reconstruction of ancestral states. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 324(6):504–516.