

Topology testing and incongruence

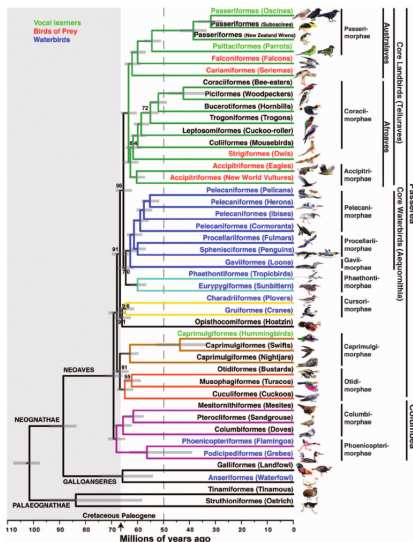
Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced

`ejmctavish@ucmerced.edu`, `twitter:snacktavish`

(With thanks to Mark Holder for slides)

Fig. 1. Genome-scale phylogeny of birds. The dated TBT inferred with ExaML. Branch colors denote well-supported clades in this and other analyses. All BS values are 100% except where noted. Names on branches denote orders (-iformes) and English group terms (in parentheses); drawings are of the specific species sequenced (names in table S1 and fig. S1). Order names are according to (36, 37) (SM6). To the right are superorder (-imorphae) and higher unranked names. In some groups, more than one species was sequenced, and these branches have been collapsed (noncollapsed version in fig. S1). Text color denotes groups of species with broadly shared traits, whether by homology or convergence. The arrow indicates the K-Pg boundary at 66 Ma, with the Cretaceous period shaded at left. The gray dashed line represents the approximate end time (50 Ma) by which nearly all neavian orders diverged. Horizontal gray bars on each node indicate the 95% credible interval of divergence time in millions of years.



Jarvis et al. (2014)

“Underlying this single topology was large-scale incongruence: none of the 14,536 trees from individual loci matched the inferred species tree, and many nodes with 100% bootstrap support appeared in <10% of the gene trees (Jarvis et al. 2014).”
Hahn and Nakhleh (2016)

How can you end up with 100% bootstrap support for relationships in only few of the gene trees?

How can you end up with a consensus tree that doesn't match any of your gene trees?

<u>Taxa</u>	<u>Unrooted binary trees</u>	<u>Rooted binary trees</u>
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	3×10^7
15	7×10^{12}	2×10^{14}
20	2×10^{20}	8×10^{21}
50	3×10^{74}	
100	2×10^{182}	
1,000	2×10^{2860}	
10,000	8×10^{38658}	
1,000,000	1×10^{5866723}	

How do you decide if you have statistical support for one tree over another?

Reasons phylogenetic inference might be wrong

1. Our data might be wrong (“garbage in garbage out”)
2. *Systematic error* – Our inference method might not be sophisticated enough
3. Random error – We might not have enough data – we are misled by sampling error.

(or it could be some combination of these).

Focus of this evening: **How confident can we be in the trees/splits inferred by ML?**

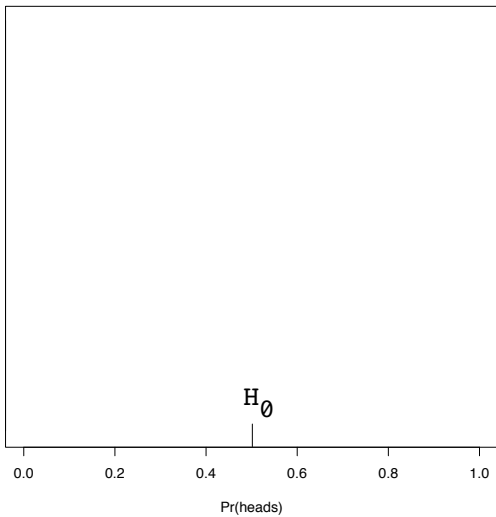
Frequentist hypothesis testing: coin flipping example

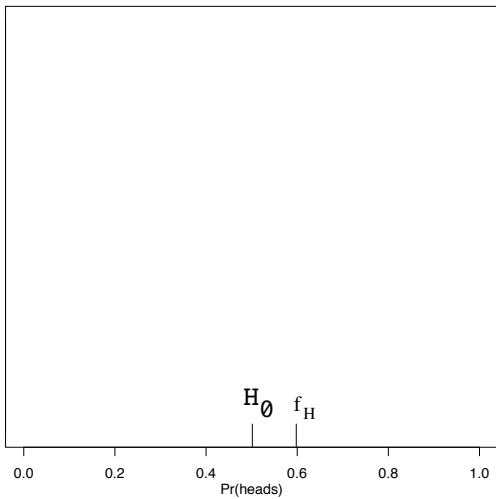
$N = 100$ and $h = 60$

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

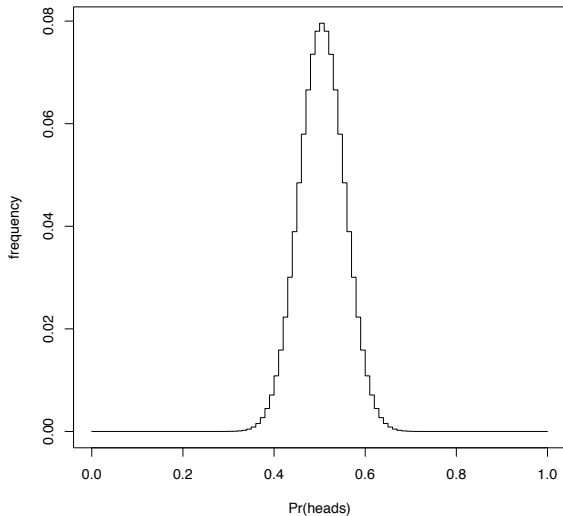
The “recipe” is:

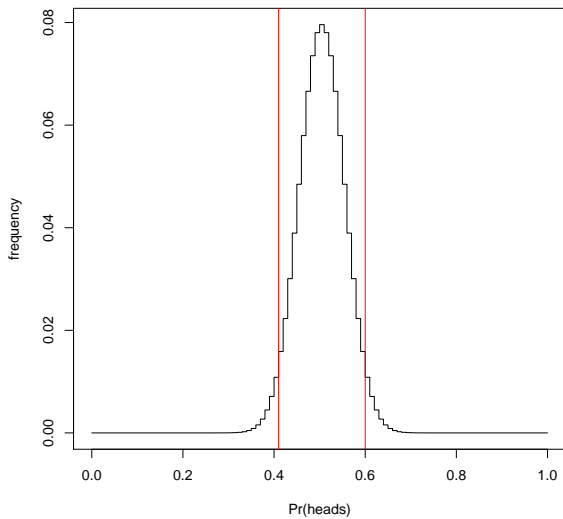
1. Formulate null (H_0) and alternative (H_A) hypotheses.
2. Choose an acceptable Type-I error rate (significance level)
3. Choose a test statistic: f_H = fraction of heads in sample.
 $f_H = 0.6$
4. Characterize the null distribution of the test statistic
5. Calculate the P -value: The probability of 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.





Null distribution





$P\text{-value} \approx 0.058$

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 on the same plot.
- We will see later (during “cartoon time”), that we *can* visualize them both in a parameter space that describes the frequency of different data patterns.

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:

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

Expectation under null:

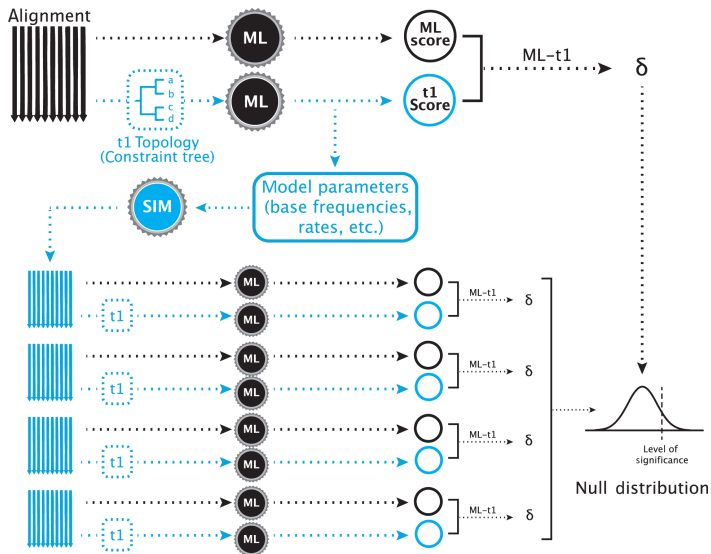
$$\mathbb{E}_{H_0} [\delta(T_1, T_2 \mid X)] = 0$$

How to generate this distribution for trees?

Parametric bootstrapping, Swofford, Olson, Wadell, Hillis (SOWH)
(Swofford et al., 1996)

- ▶ Simulate sequence data on alternative tree under estimated parameters
- ▶ Test Statistic = Difference in Likelihood between ML tree and alternative tree
- ▶ Null Distribution = Expected distribution if the T1 hypothesis is assumed to be correct

script at <https://github.com/josephryan/sowhat>



(Church et al., 2015)

Problems with SOWH test:

- ▶ Many tree searches - computationally expensive
- ▶ Over simplified models result in high type 1 error

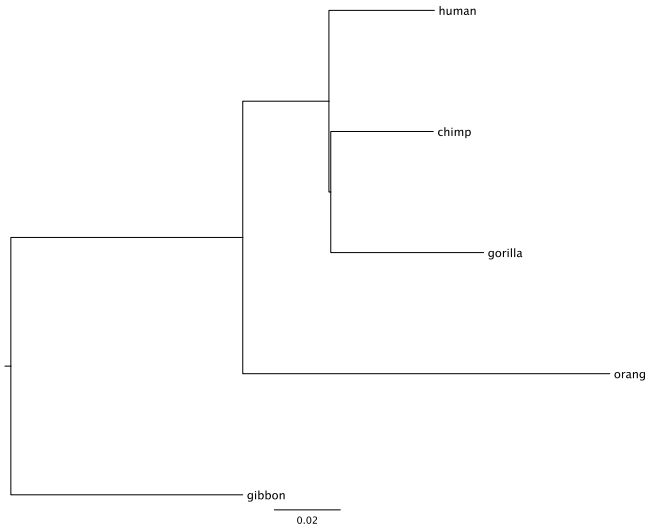
If you reject a null hypothesis based on parametric bootstrapping, the appropriate conclusion is that the degree of support observed is too large to be easily explained by chance assuming that the simulated model of sequence evolution adequately mimics the level of conflicting signal in the true generating process.

McTavish and Holder, Encyclopedia of Evolution, 2016

Non-parametric approach

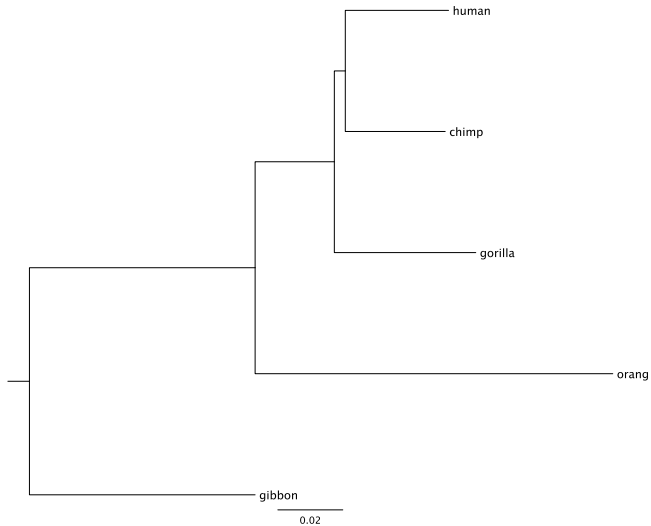
Using 3000 sites of mtDNA sequence for 5 primates

T_1 is ((chimp, gorilla), human)



Using 3000 sites of mtDNA sequence for 5 primates

T_2 is ((chimp, human), gorilla)



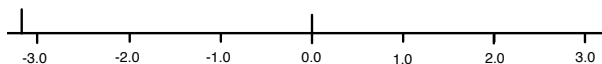
Using 3000 sites of mtDNA sequence for 5 primates

$$T_1 \text{ is } ((\text{chimp}, \text{gorilla}), \text{human}) \quad \ln L(T_1 | X) = -7363.296$$

$$T_2 \text{ is } ((\text{chimp}, \text{human}), \text{gorilla}) \quad \ln L(T_2 | X) = -7361.707$$

$$\delta(T_1, T_2 | X) = -3.18$$

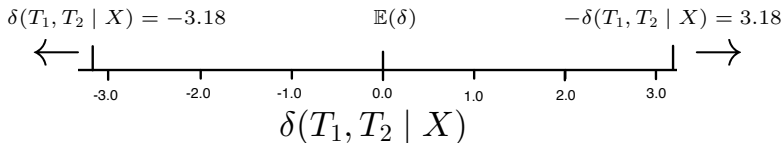
$$\mathbb{E}(\delta)$$



$$\delta(T_1, T_2 | X)$$

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

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



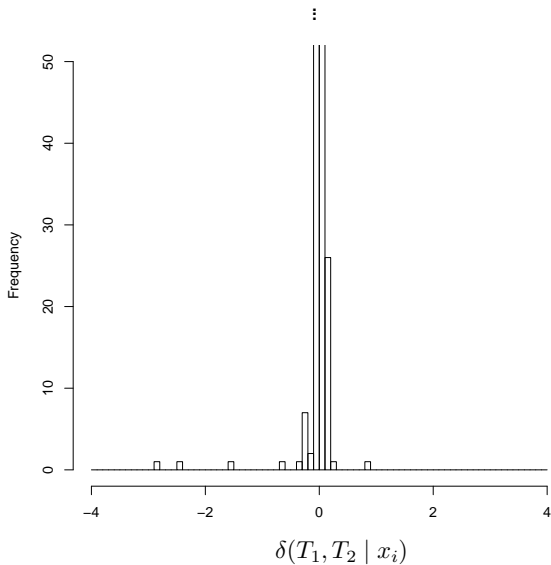
KH Test

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 .

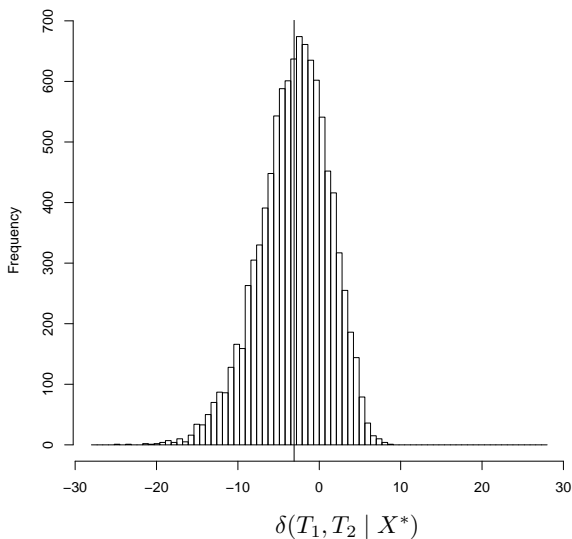


KH Test - the variance of $\delta(T_1, T_2 \mid X)$

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). Or
2. use bootstrapping to generate a cloud of pseudo-replicate $\delta(T_1, T_2 \mid X^*)$ values, and look at their variance.

δ for many (RELL) bootstrapped replicates of the data



RELL bootstrap

Often, the MLE of numerical parameters (including branch lengths) do not change much when we bootstrap.

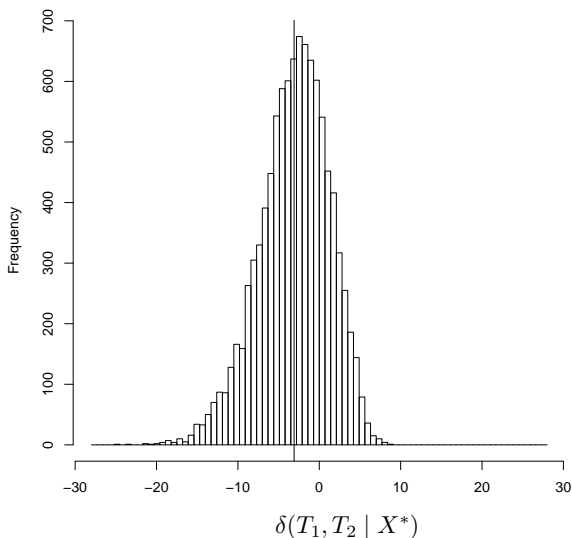
So, we can simply resample the site $\ln L$ values and sum them (rather than reoptimizing parameters).

This is called the RELL bootstrap (Kishino et al., 1990, and Felsenstein). It is not a “safe” replacement for normal bootstrapping (especially on large trees; Stamatakis et al., 2008) when you want to estimate clade support.

But it should be good enough for helping us learn about the standard error of the $\ln L$.

And it is really fast.

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



KH Test - 'centering'

H_0 gives us the expected value:

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

Bootstrapping gives us a reasonable guess of the variance under H_0

By subtracting the mean of the bootstrapped $\delta(T_1, T_2 \mid X^*)$ values, we can create a null distribution.

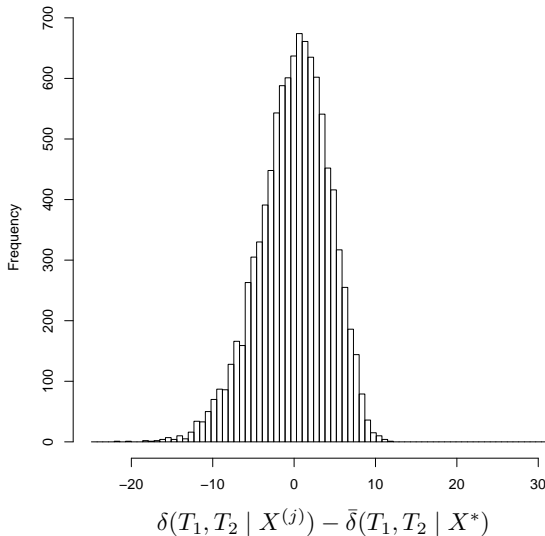
For each of the j bootstrap replicates, we treat

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

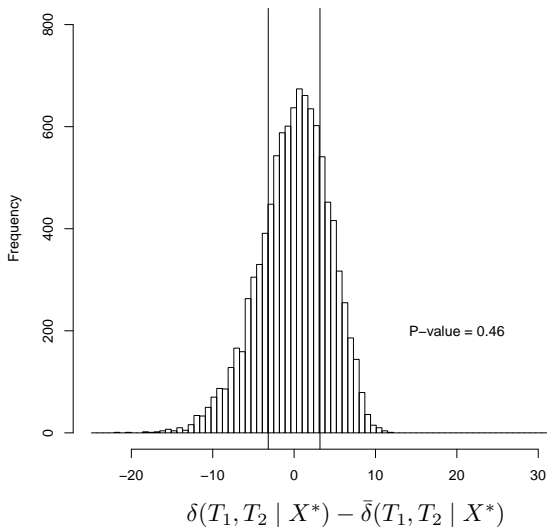
as draws from the null distribution.

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

for many (RELL) bootstrapped replicates of the data



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



Summary - Part 1

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

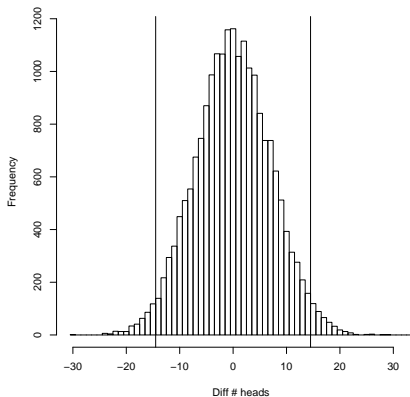
Even when the H_0 is true, we do not expect $2 \left[\ln L(\hat{T}) - \ln L(T_1) \right] = 0$

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.

$$h_1 - h_2$$

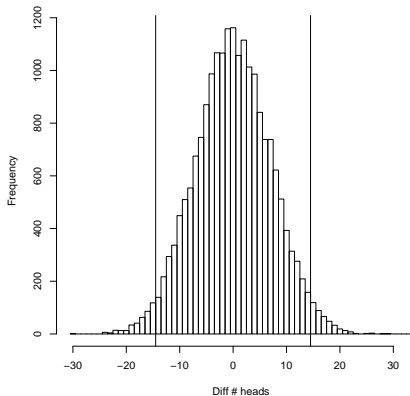
Null dist.: difference in # heads any two competitors



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

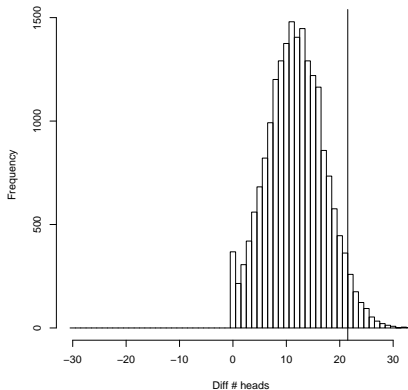
$$h_1 - h_2$$

Null dist.: difference in # heads any two competitors



$$\max(h) - h_1$$

Null dist.: difference highest – random competitor



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

You have to specify of a **set of candidate trees** - inclusion in this set **must not** depend on the dataset to be analyzed.

The null hypothesis is that all members of the candidate set have the same expected score.

The test makes worst-case assumptions, so the SH test **is conservative**.

SH test candidate set selection

- Should be all trees that you would have seriously entertained before seeing the data (considering a subset of trees for computational convenience can invalidate the test).
- Using all trees is safe.
- If a tree has low $\ln L$ and low variance of site-log-likelihoods then it can probably be safely removed without affecting the P -values of other trees¹

¹Because such a tree would be unlikely to ever be the tree that is the determines the maximum displacement from the centered value, $m^{(j)}$.

SH Test details

- For each tree T_i in the candidate set calculate $\delta(\hat{T}, T_i | X)$
- Bootstrap to generate $\ln L(T_i | X^{(j)})$ for each bootstrap replicate j .
- For each tree T_i , use the mean, $\ln \bar{L}(T_i | X^*)$, over all bootstrap replicates to center the bootstrapped collection of log-likelihoods:

$$c_i^{(j)} = \ln L(T_i | X^{(j)}) - \ln \bar{L}(T_i | X^*)$$

- For each bootstrap replicate, j , pick the highest value from the centered distributions (this mimics the selection bias):

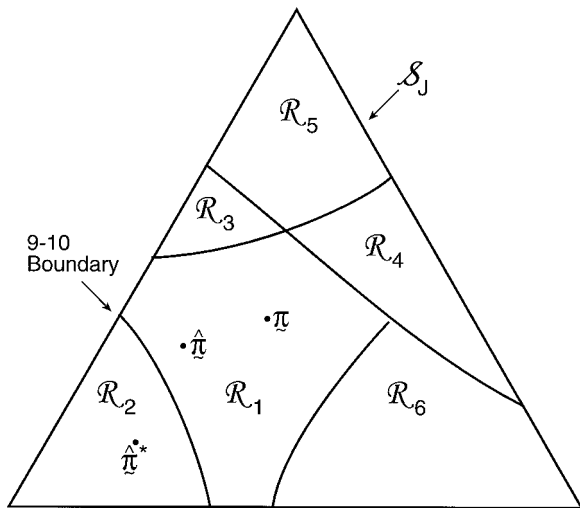
$$m^{(j)} = \max \left[c_i^{(j)} \right] \text{ over all } i$$

- Then for each tree and replicate, you get a sample from the null $\delta_i^{(j)} = m^{(j)} - c_i^{(j)}$
- P -value for tree T_i is approximated by the proportions of bootstrap reps for which:

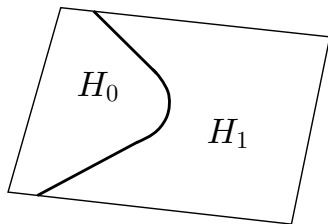
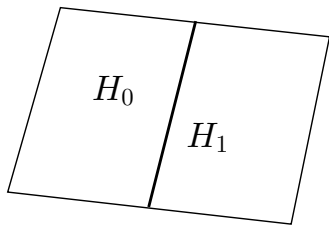
$$\delta_i^{(j)} \leq \delta(\hat{T}, T_i | X)$$

- 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)

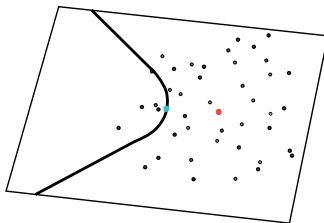
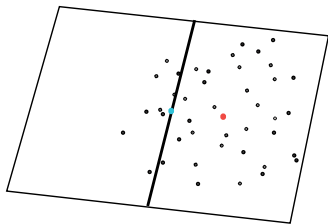
Efron et al. (1996) view of tree space



Imagine hypothesis tests of locations with different border shapes:



Similar dataset with point estimates (red dot) in H_1
Green dot is the hardest set of locations in H_0 to reject.

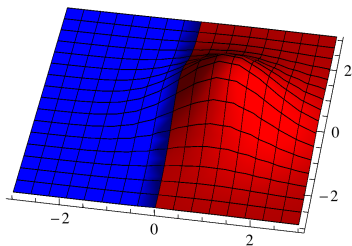
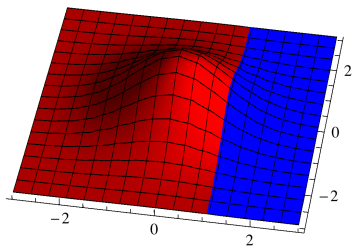


In the straight border case, symmetry implies that:

The actual P -value (blue region)

$$\approx 1 - BP$$

($1 - BP$ is the blue below)

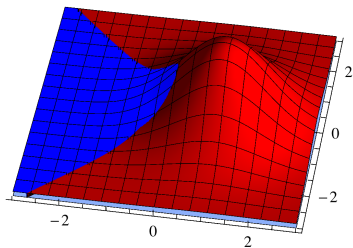
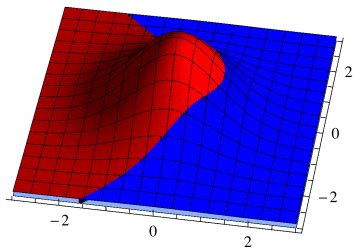


In the curved border case, the symmetry breaks down:

The actual P -value (blue region)

$$\neq 1 - BP$$

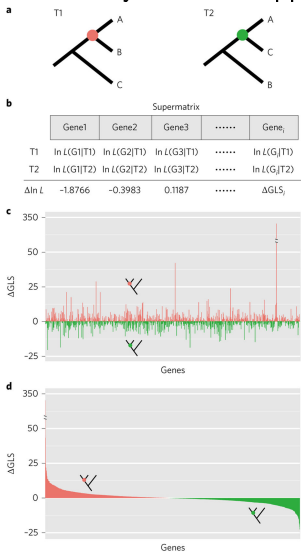
($1 - BP$ is the blue below)



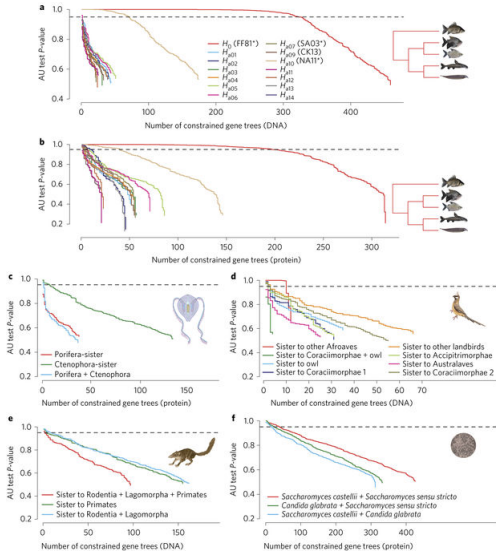
- Efron et al. (1996) proposed a computationally expensive multi-level bootstrap (which has not been widely used).
- Shimodaira (2002) used the same theoretical framework to devise a (more feasible) Approximately Unbiased (AU) test of topologies.
 - Multiple scales of bootstrap resampling (80% of characters, 90%, 100%, 110%...) are used to detect and correct for curvature of the boundary.
 - Implemented in the new versions of PAUP*

How to apply these tests when we are dealing with many many gene trees?

Genes may differ in support for topologies



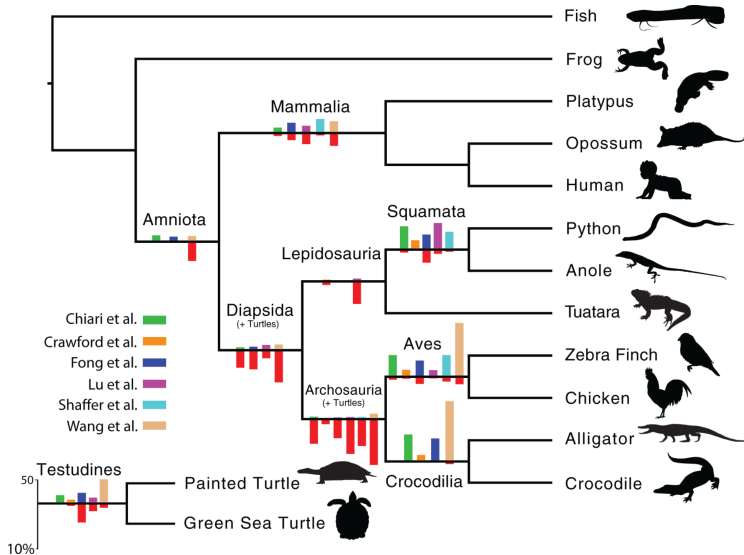
Shen et al. (2017)



Arcila et al. (2017)

Bayes factors capture degree of support for two alternative hypotheses

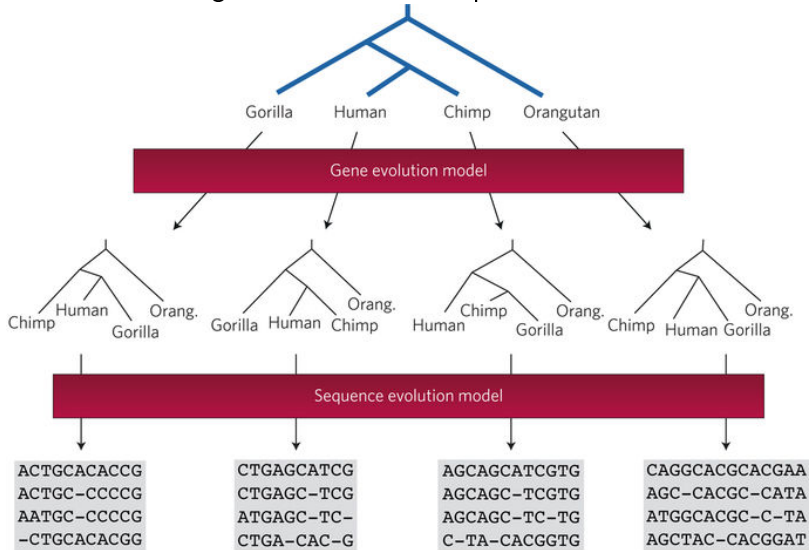
$$\textit{BayesFactor} = \frac{P(D|H_1)}{P(D|H_2)}$$




Brown and Thomson (2017)

major outliers may suggest data problems - e.g. paralogous genes being treated as homologous
Brown and Thomson (2017)

Should we expect gene trees to match species tree?



Mirarab (2017)

- Arcila, D., Ortí, G., Vari, R., Armbruster, J. W., Stiasny, M. L. J., Ko, K. D., Sabaj, M. H., Lundberg, J., Revell, L. J., and Betancur-R, R. (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution*, 1(2):0020.
- Brown, J. M. and Thomson, R. C. (2017). Bayes Factors Unmask Highly Variable Information Content, Bias, and Extreme Influence in Phylogenomic Analyses. *Systematic Biology*, 66(4):517–530.
- 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.
- Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17.
- Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., Fonseca, R. R. d., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, 

- D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V., Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdocimi, F., Samaniego, J. A., Velazquez, A. M. V., Alfaro-Núñez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jönsson, K. A., Johnson, W., Koepfli, K.-P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R., Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun, W., Gilbert, M. T. P., and Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331.
- Mirarab, S. (2017). Phylogenomics: Constrained gene tree inference. *Nature Ecology & Evolution*, 1(2):0056.
- Shen, X.-X., Hittinger, C. T., and Rokas, A. (2017). Contentious

relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*, 1(5):0126.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference.