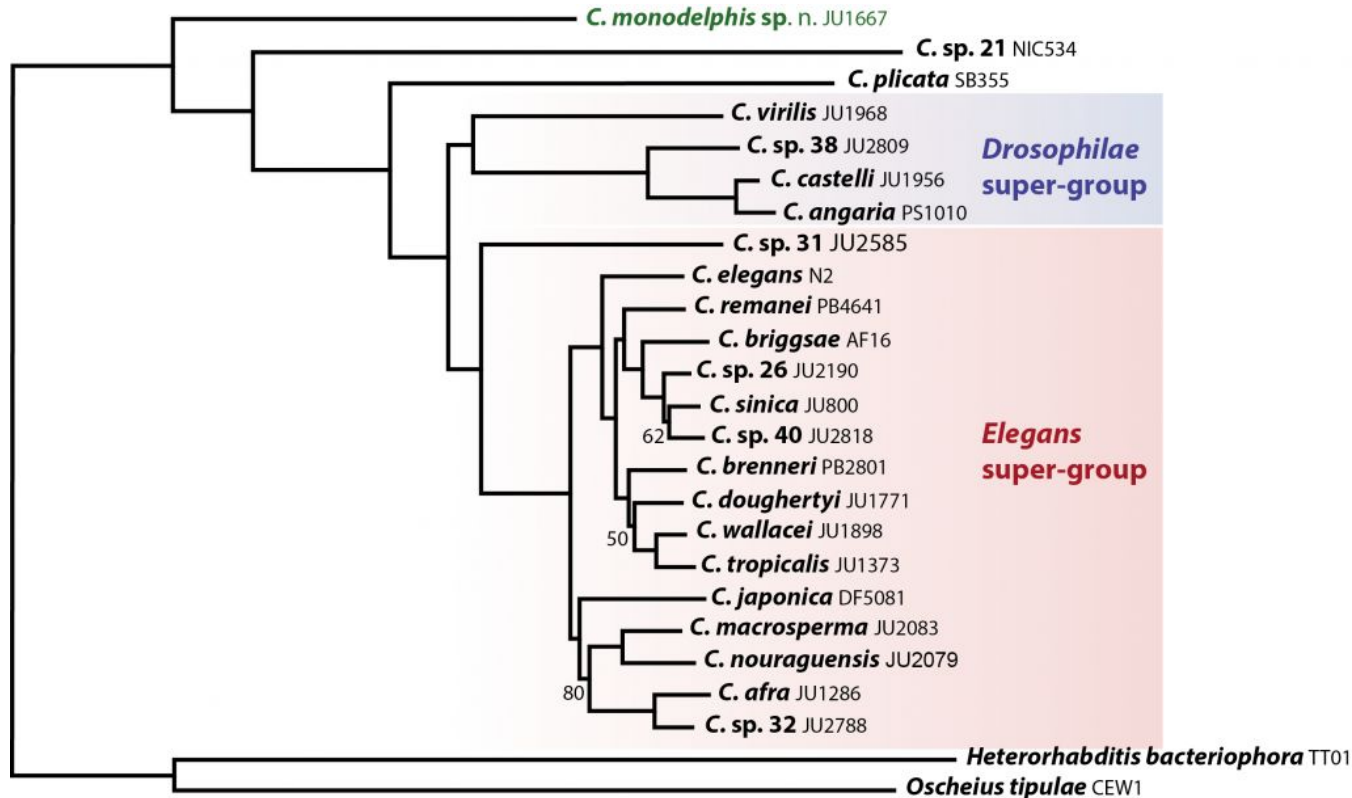


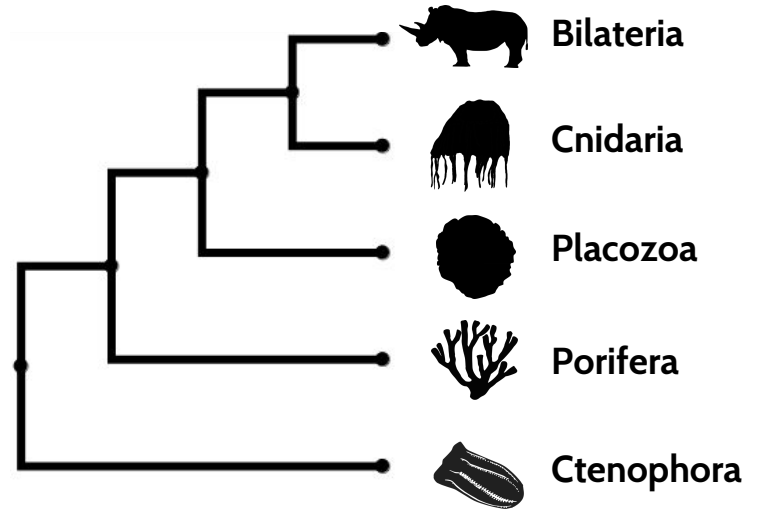
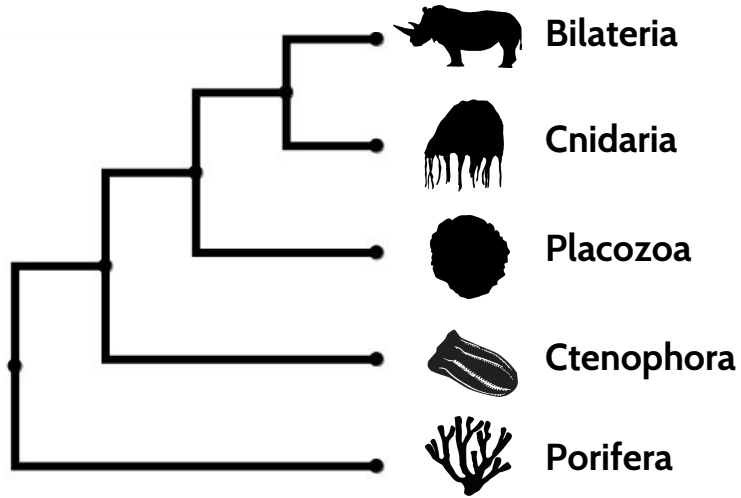
# Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses

Jeremy M. Brown & Robert C. Thomson  
(2016) *Systematic Biology* 66(4):517-530

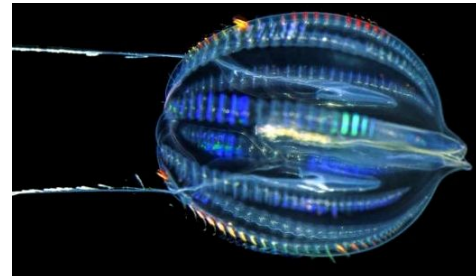
# Phylogenomic analyses yield highly supported topologies



# Ctenophores sister to all other animals?!

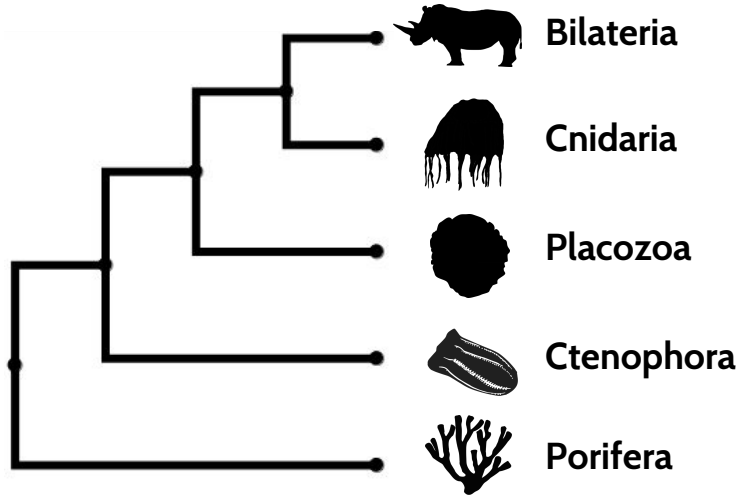


Sponge (Porifera)

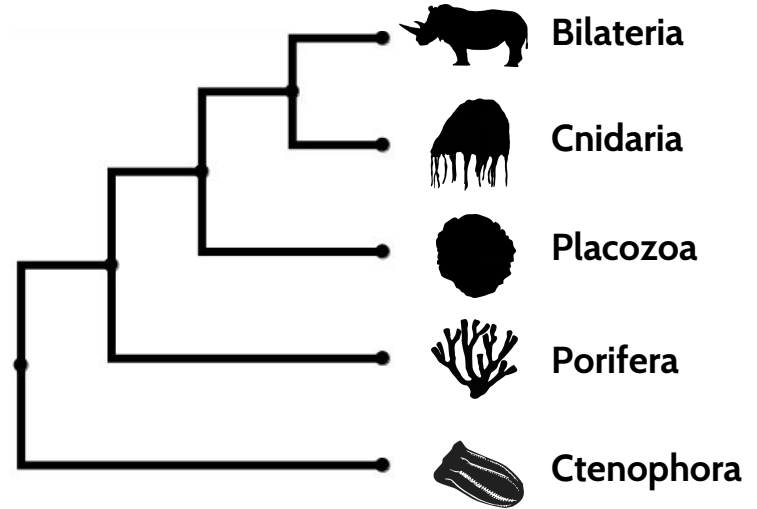


Comb Jellies (Ctenophora)

# Ctenophores sister to all other animals?!



Philippe *et al.* (2009)  
Pick *et al.* (2010)  
Pisani *et al.* (2015)



Dunn *et al.* (2008)  
Chang *et al.* (2015)

# Bayes Factors

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

posterior odds ratio

prior odds ratio

$H_1$  = presence of particular bipartition  
 $H_2$  = absence of particular bipartition  
 $D$  = data

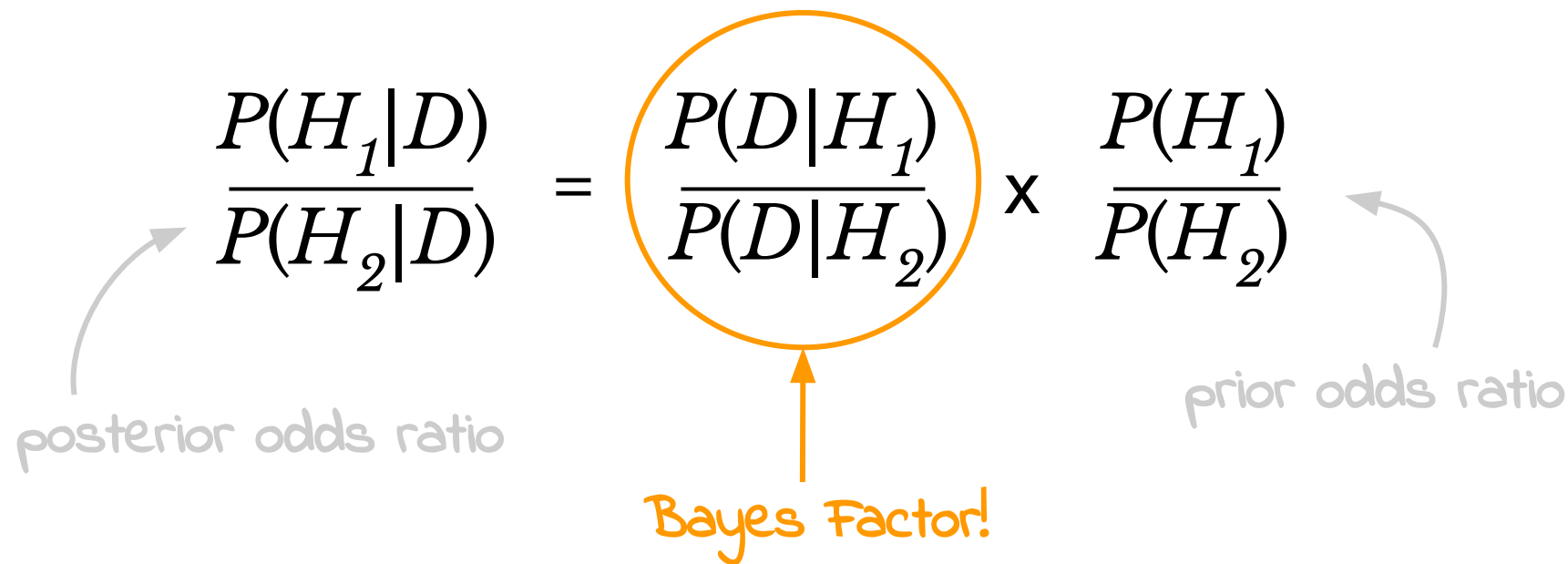
# Bayes Factors

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

posterior odds ratio

Bayes Factor!

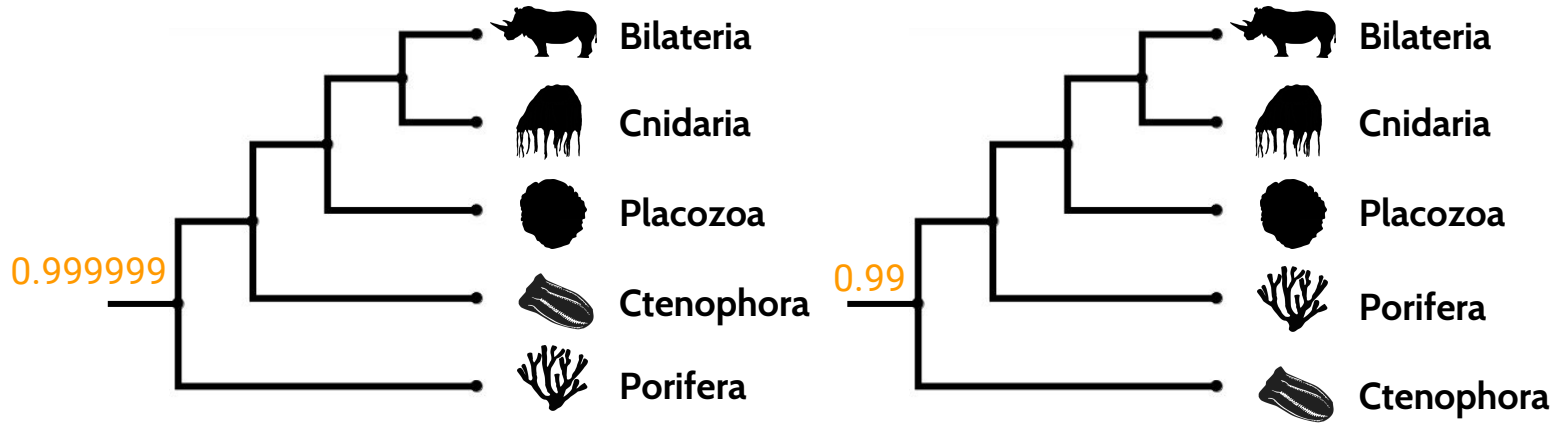
prior odds ratio



$H_1$  = presence of particular bipartition  
 $H_2$  = absence of particular bipartition  
 $D$  = data

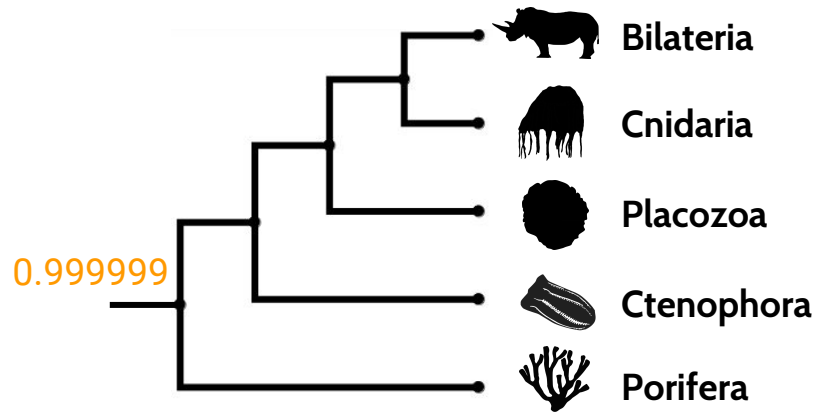
# Benefits of Bayes factors

1. “BFs, and particularly  $\log(\text{BF})$ s, offer a larger numerical range to measure support than posterior probabilities”

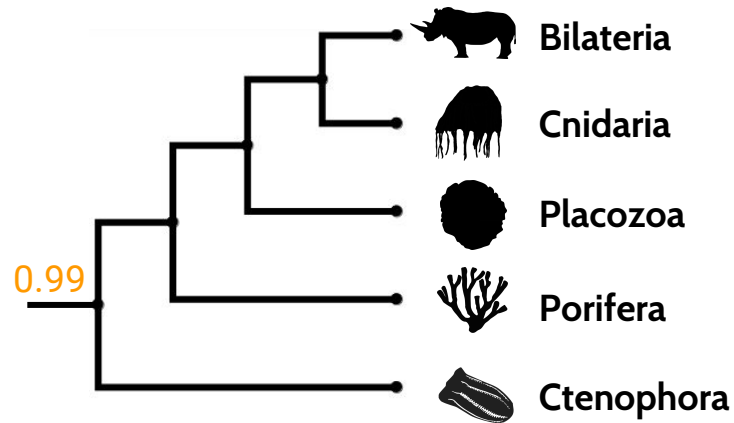


# Benefits of Bayes factors

1. “BFs, and particularly log(BF)s, offer a larger numerical range to measure support than posterior probabilities”



$$\frac{0.999999}{0.000001} = 999,999$$



$$\frac{0.99}{0.01} = 99$$



# Benefits of Bayes factors

1. “BFs, and particularly  $\log(\text{BF})$ s, offer a larger numerical range to measure support than posterior probabilities”
2. “When posterior probabilities are extreme (near 0 or 1), MCMC does not estimate these values with sufficient precision to distinguish between 0.99 and 0.99999...”

# Benefits of Bayes factors

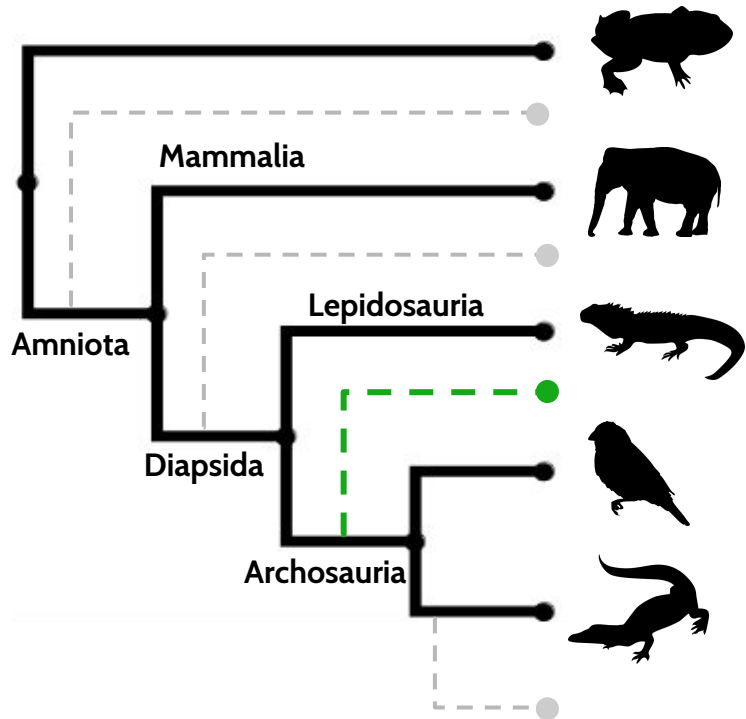
1. “BFs, and particularly  $\log(\text{BF})$ s, offer a larger numerical range to measure support than posterior probabilities”
2. “When posterior probabilities are extreme (near 0 or 1), MCMC does not estimate these values with sufficient precisions to distinguish between 0.99 and 0.99999...”
3. “Bayes factors do not depend on the prior probabilities of the two hypotheses, as opposed to posterior odds ratios...”

“If the hypotheses concern the monophyly of a set of taxa (eg.  $H_1$  requires monophyly and  $H_2$  requires non-monophyly) the standard discrete, uniform prior on topology will tend to favour  $H_2$ , sometimes strongly...”

# Where in the tree of life do turtles sit?



# Where in the tree of life do turtles sit?



Chiari *et al.* (2012)

- Sequenced transcriptomes

Fong *et al.* (2012)

- Sanger sequencing of PCR amplicons

Crawford *et al.* (2012)

- Sequenced “UCEs” (highly conserved nuclear loci)

Shaffer *et al.* (2013)

Wang *et al.* (2013)

- Sequenced new turtle genomes

Lu *et al.* (2013)

- Combined new genomic data

# Methodology

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

To calculate  $2\ln(\text{BFs})$  for well established groups (eg. birds):

$P(D|H_1)$ :

- Constrain birds to be monophyletic
- Constrain all other well-established groups to be monophyletic

$P(D|H_2)$ :

- Constrain bird to be paraphyletic
- Constrain all other well-established groups to be monophyletic

*“We enforced positive constraints on all non-focal clades to provide a more meaningful measure of support (Bergsten et al. 2013)”*

# Methodology

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

To calculate  $2\ln(\text{BFs})$  for the various turtle placement hypotheses (eg. **Turtles sister to Archosaurs**):

$P(D|H_1)$ :

- Constrain turtles to be sister to Archosaurs
- Constrain all other well-established groups to be monophyletic

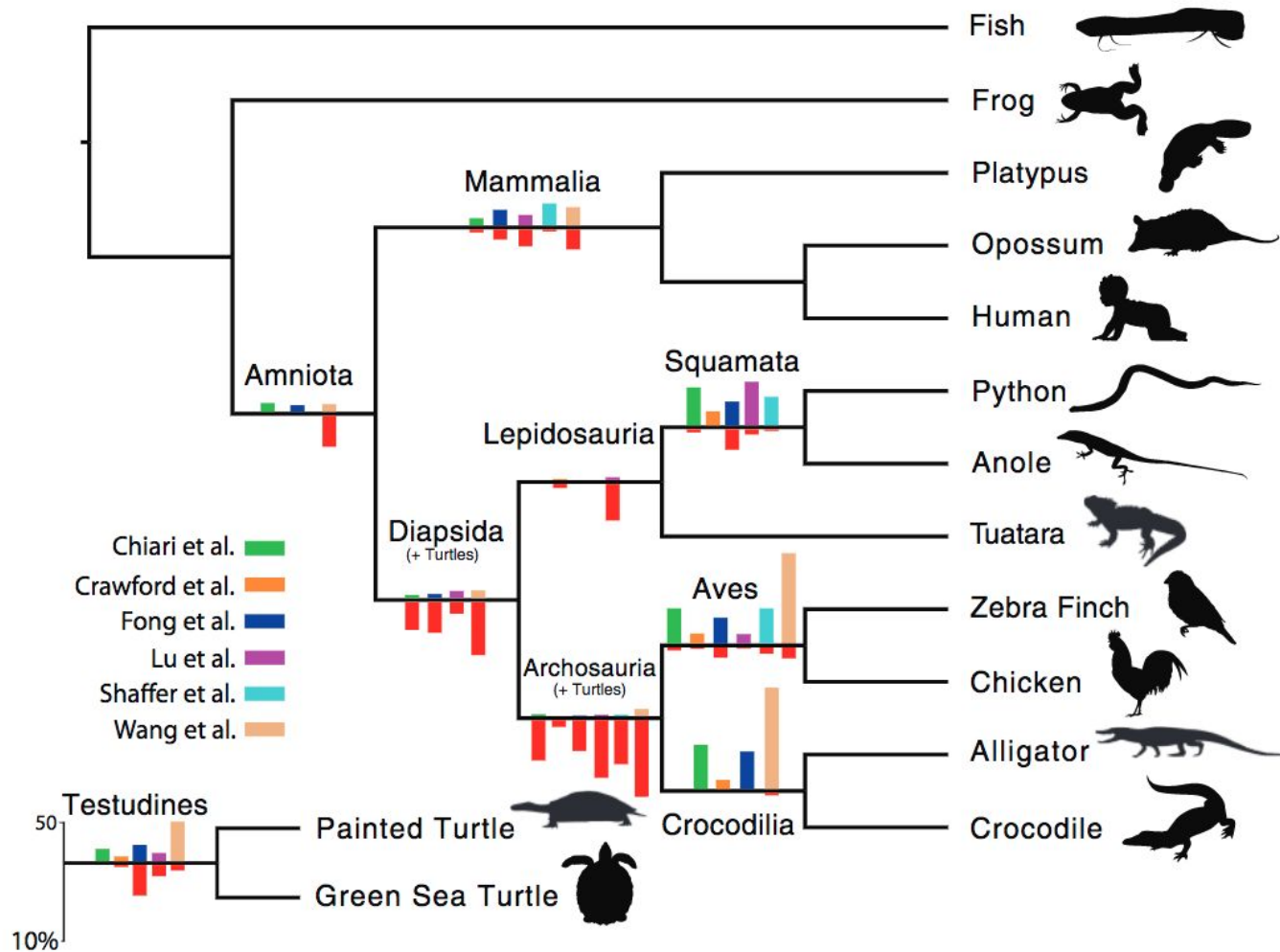
$P(D|H_2)$ :

- Constrain turtles to in all other hypothesised places (eg. sister to crocodilians, birds, etc.)
- Constrain all other well-established groups to be monophyletic

***“We enforced positive constraints on all non-focal clades to provide a more meaningful measure of support (Bergsten et al. 2013)”***

# Methodology

- Trees for each of the 6 concatenated datasets (GTR+G+I)
- Posterior probability distributions of tree topologies for each gene in each dataset (model estimated using AIC)
- $2\ln(\text{BF})$ s for the monophyly of all well-established groups
- $2\ln(\text{BFs})$  for the all suggested placements of turtles with respect to well-established groups



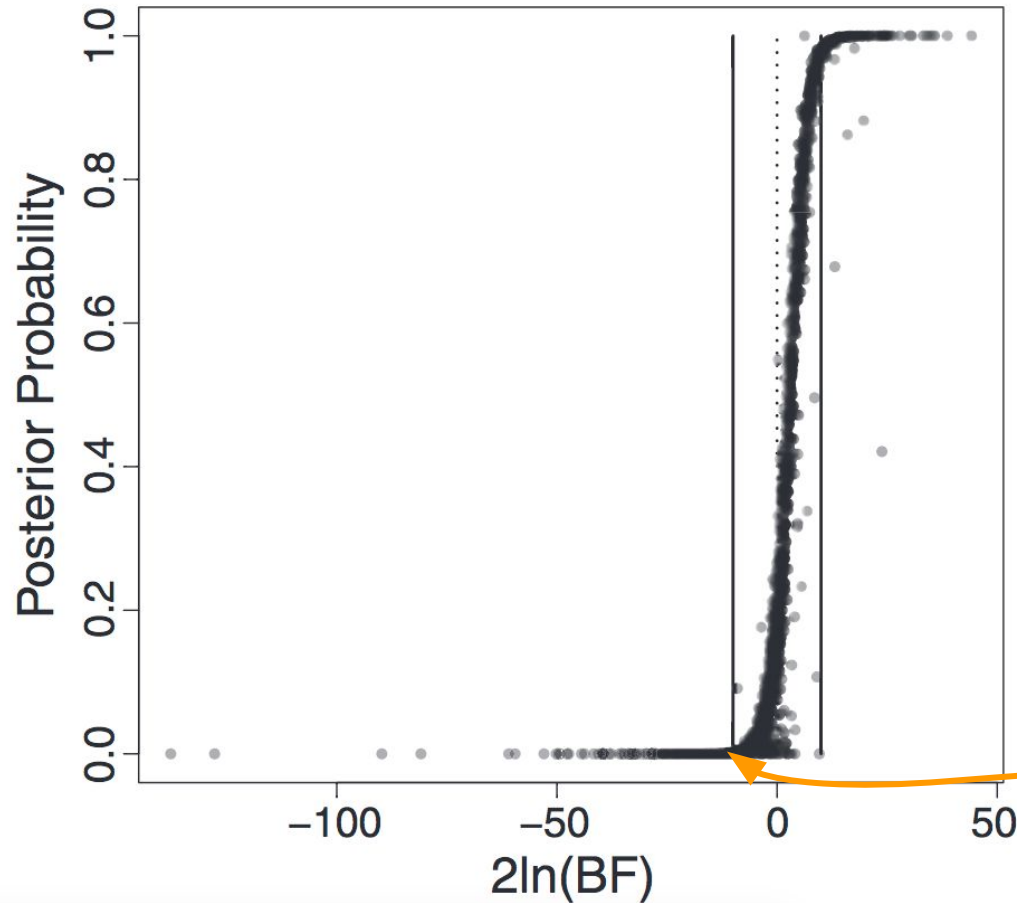
**Figure 1:**  
Colored, upward bars on each branch give the median  $2\ln(\text{BF})$  value across genes supporting that relationship for each data set.

Red, downward bars show the percentage of genes in each data set that strongly reject ( $2\ln(\text{BF}) < -10$ ) each clade.

For Archosauria and Diapsida, we provide values for the monophyly of these groups along with turtles, since most studies suggest that turtles are a member of these clades.



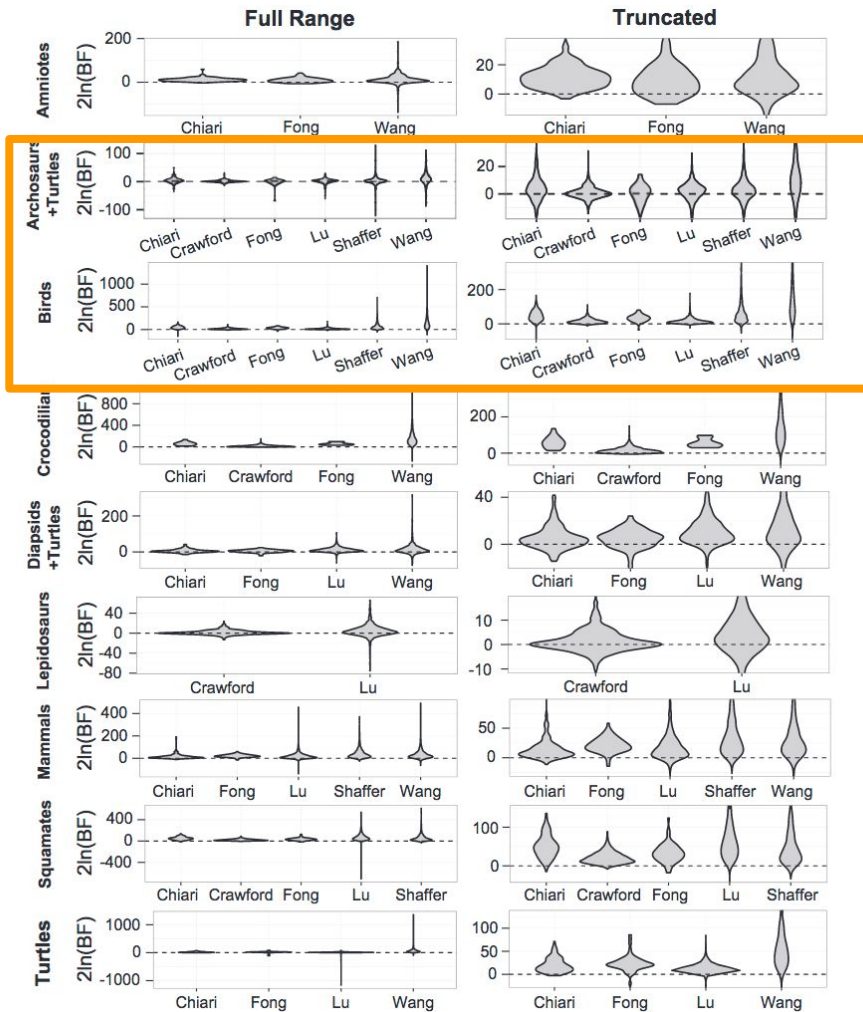
## Estimates of PPs provided by MCMC obscure variation in strength of support/rejection



**Figure 2:**

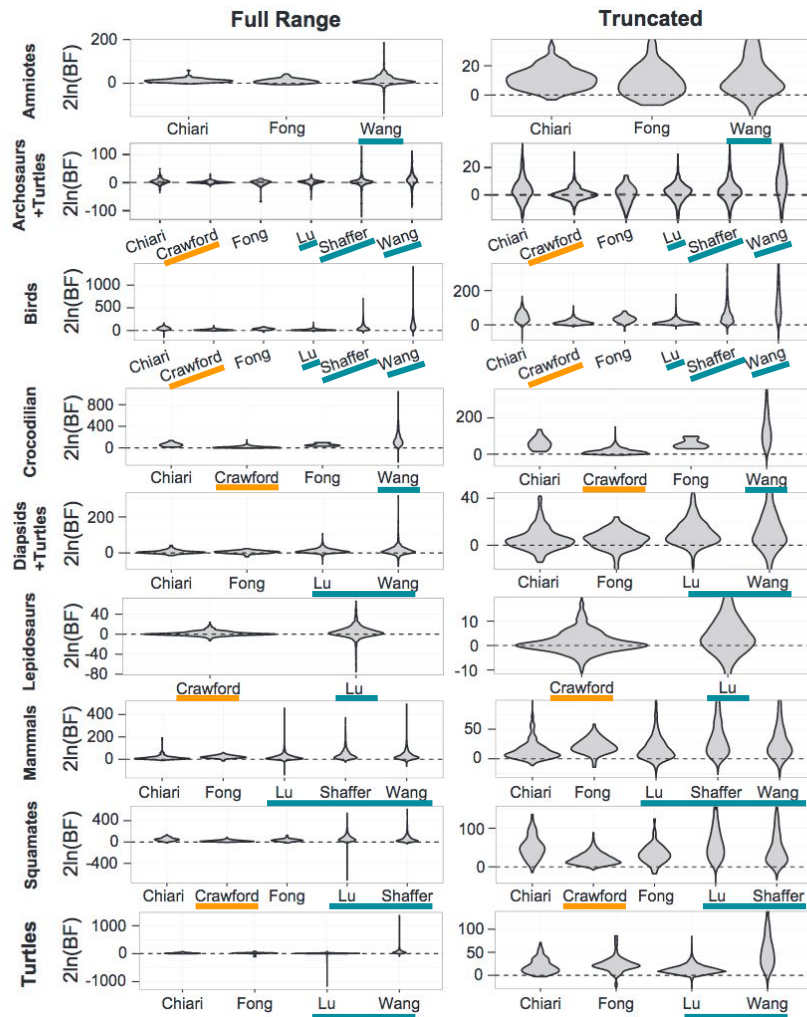
A comparison of MCMC posterior probability estimates to  $2\ln(\text{BF})$  values for **archosaur monophyly** across all genes in the Shaffer data set.

“Genes strongly rejecting the monophyly of well-established groups should certainly be considered suspect...”



**Figure 3:**  
Summary of BF support for major clades  
in the amniote phylogeny.

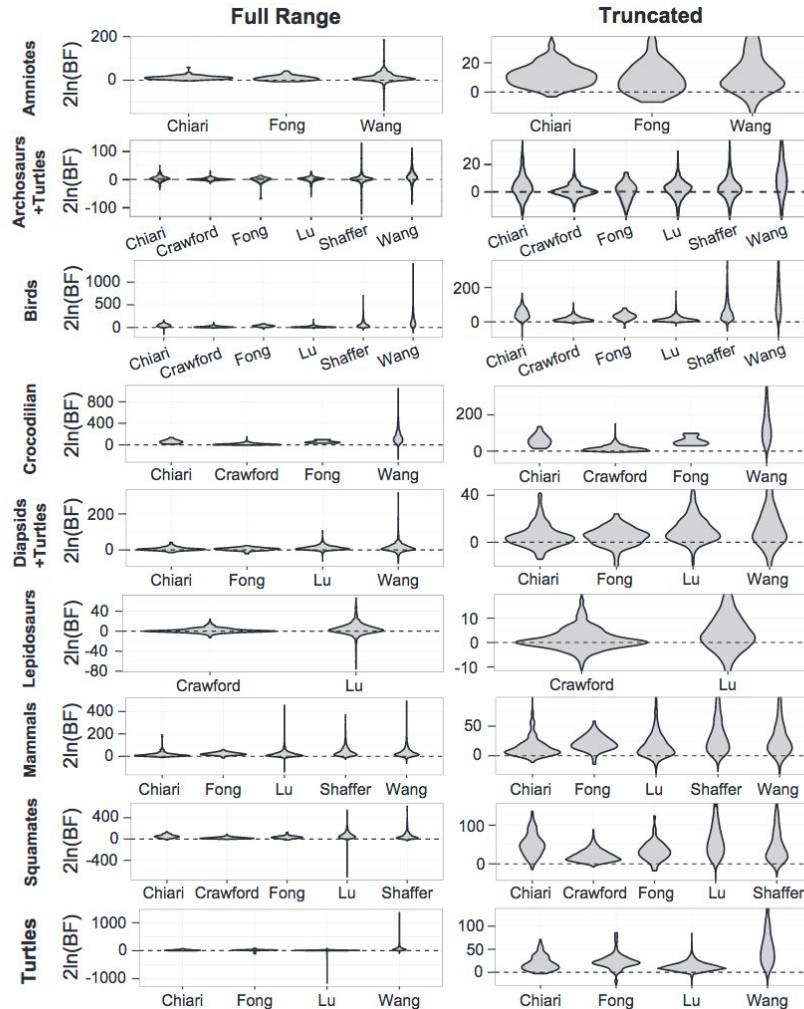
“Despite being supported with indistinguishable PP estimates, major clades varied extensively in the strength of support provided by individual genes...”



**Figure 3:**  
Summary of BF support for major clades  
in the amniote phylogeny.

Crawford *et. al.* = UCEs

Wang *et. al.* + Lu *et. al.* = Loci from draft genomes



“When performing concatenated inference, genes with large BF’s can have immense influence on overall phylogenetic estimate.

Eg.  
Outlier genes with  $\log(\text{BF})$  values of  $>200$  were present in several datasets.

$2\ln(\text{BF})$  of 200 == BF of  $2.7 \times 10^{43}$ , a level of certainty which is difficult to put into words

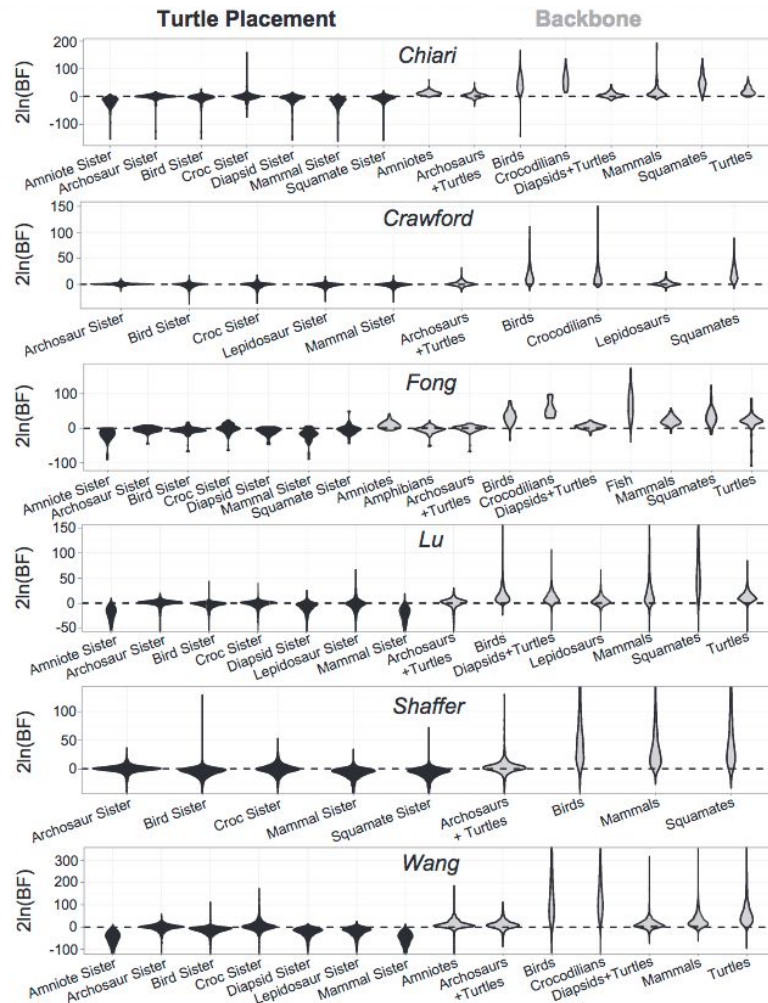
If this gene was unreliable for some reason, it would take **20 genes** with a  $2\ln(\text{BF}) = -10$  to counterbalance the outliers effect..

**Identification and careful scrutiny of outliers based on BF’s might be a valuable tool for ensuring the accuracy of concatenated analyses.**

# Genes rejecting well-established relationships

TABLE 2. Tallies of genes rejecting well-established backbone relationships across data sets, with the final row giving the number of genes that reject at least one of the relationships listed in the other rows

	Chiari et al. 2012 248 genes 16 taxa	Crawford et al. 2012 1145 genes 10 taxa	Fong et al. 2012 75 genes 110 taxa	Lu et al. 2013 1638 genes 11 taxa	Shaffer et al. 2013 1955 genes 8 taxa	Wang et al. 2013 1113 genes 12 taxa
Amniota	0	—	0	—	—	45
Archosauria (+ Testudines)	13	10	3	123	112	111
Aves	1	2	1	2	16	16
Crocodylia	0	0	0	—	—	7
Diapsida (+ Testudines)	9	—	3	25	—	77
Lepidosauria	—	6	—	78	—	—
Mammalia	1	—	1	36	5	29
Squamata	1	0	2	11	4	—
Testudines	0	3	3	24	—	8
<b>Total</b>	24 (9.7%)	21 (1.8%)	10 (13.3%)	262 (16.0%)	132 (6.8%)	246 (22.1%)

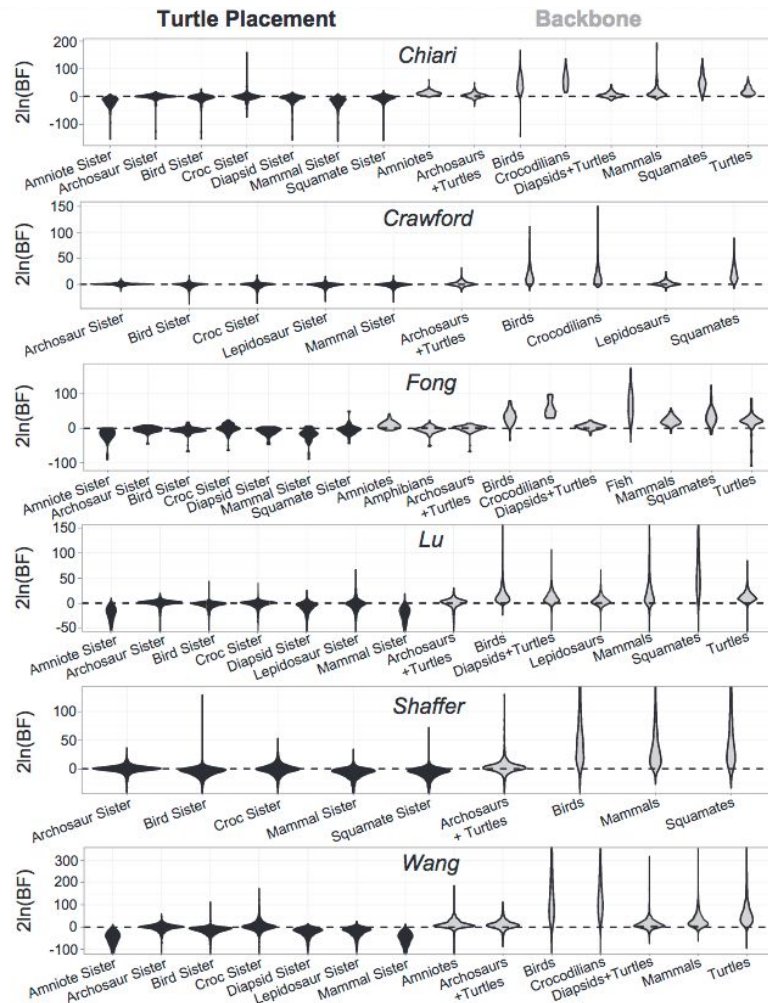


**Figure 4:**

Summary of BF support for both the placement of turtles and the monophyly of major amniote clades, grouped by data set.

Note that the y-axis is truncated for the data sets of Lu *et al.* (2013), Shaffer *et al.* (2013), and Wang *et al.* (2013) to highlight differences in central tendencies of the distributions and minimize the influence of outliers.





## Author's concatenation analyses:

Archosaur-sister: Crawford, Lu, Schaffer

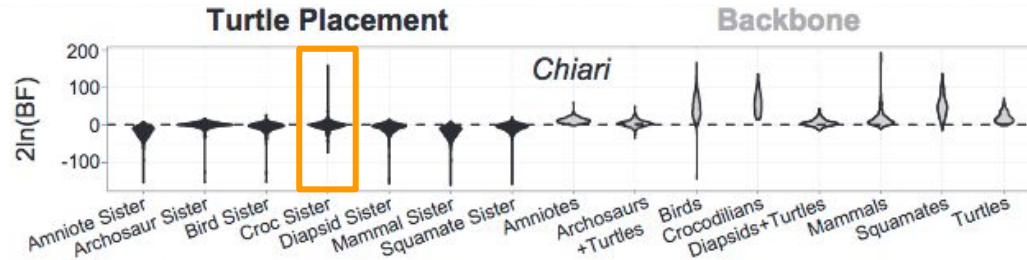
Crocilian-sister: Chiari, Fong, Wang

## Published analysis:

All six supported Archosaur-sister but:

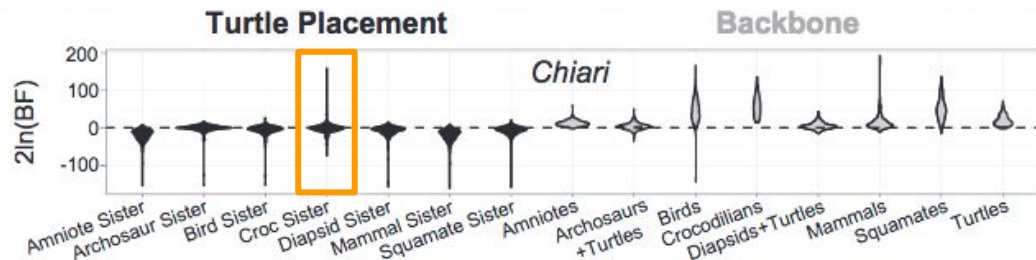
- Fong *et al.* obtained different results depending on different gene and taxon sampling
- Wang *et al.* removed all third-codon positions from alignment
- Chiari *et al.* originally recovered Croc-sister, but by removing all third-codon positions/using more sophisticated models, recovered Archosaur-sister

# Identifying and removing suspect loci



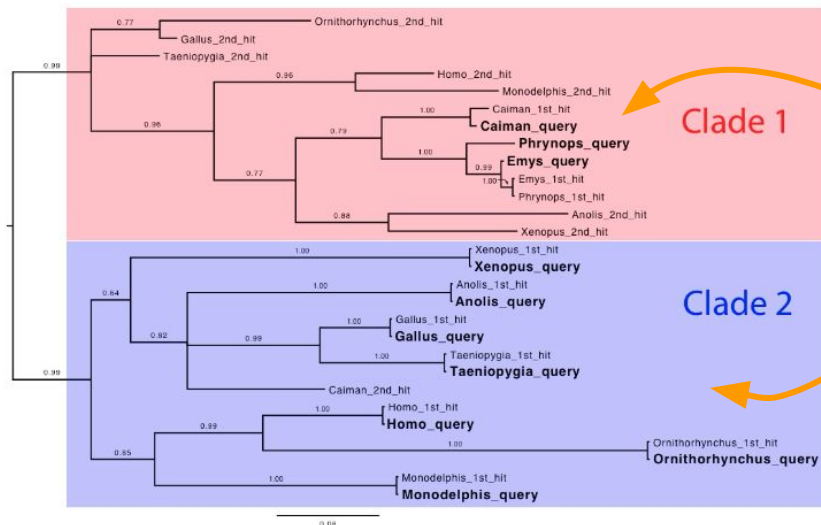


# Identifying and removing suspect loci

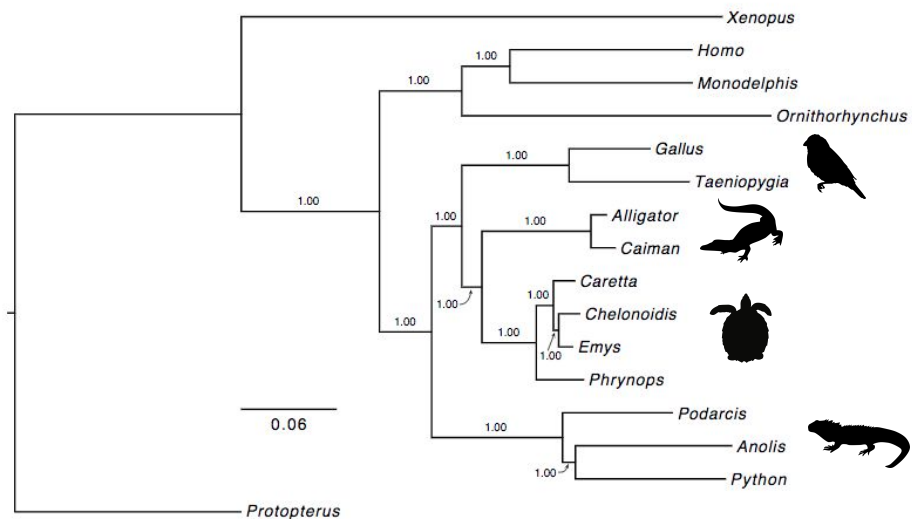


**Figure S4**

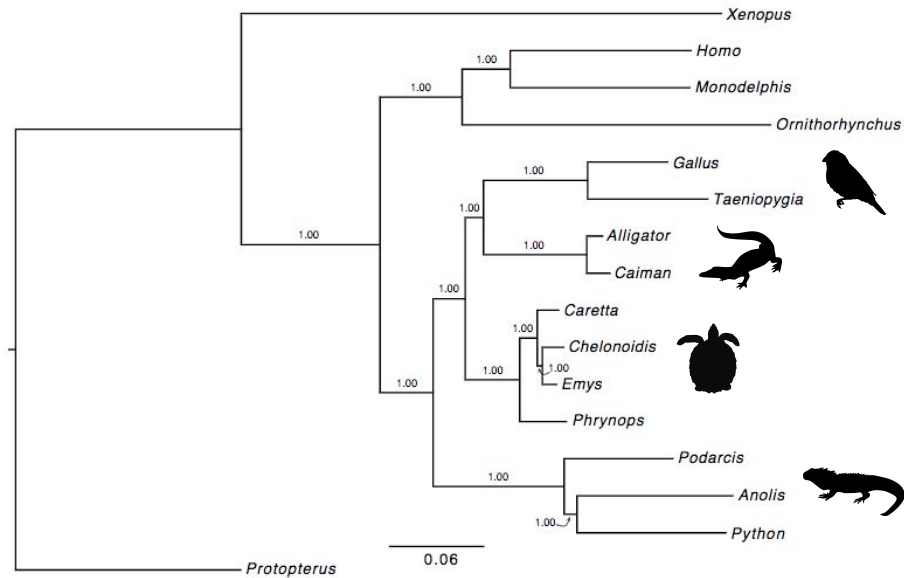
Majority-rule consensus tree of Chiari *et al.*'s alignment 11434 with added BLASTn hits from closely related reference genomes



# Identifying and removing suspect loci



248 genes from Chiari *et. al.*



246 genes from Chiari *et. al.*  
(excluding 2 orthogroups containing paralogues)

# Future Prospects

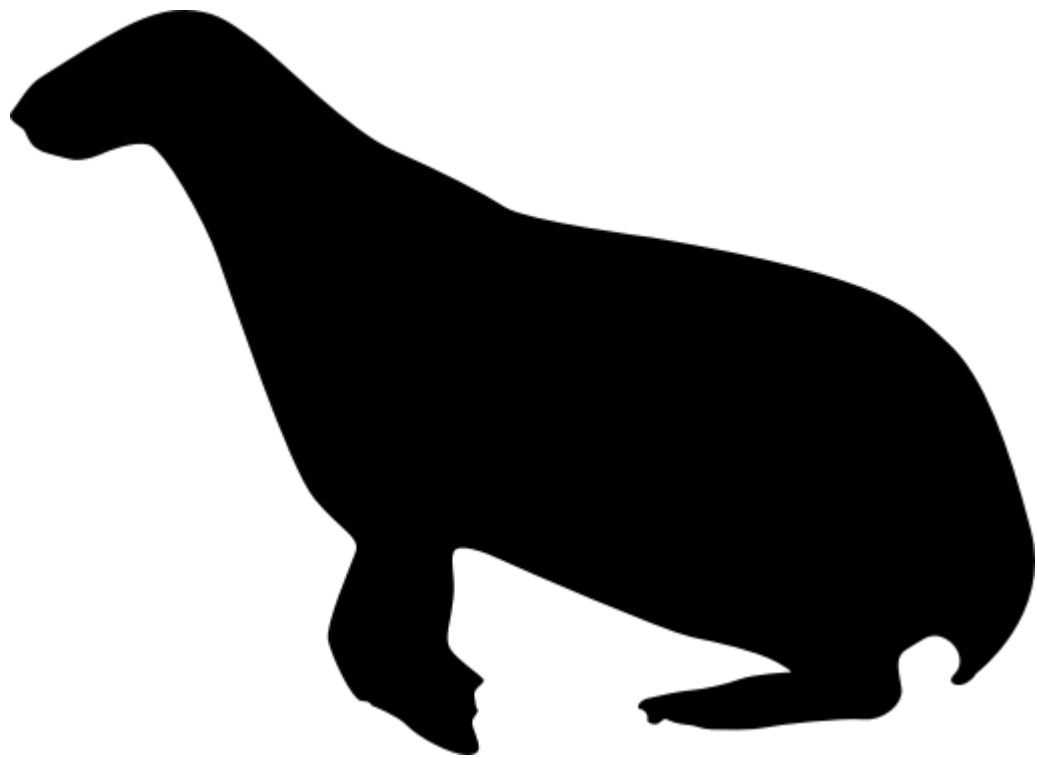
Current methods of estimating BFs are computationally intensive:

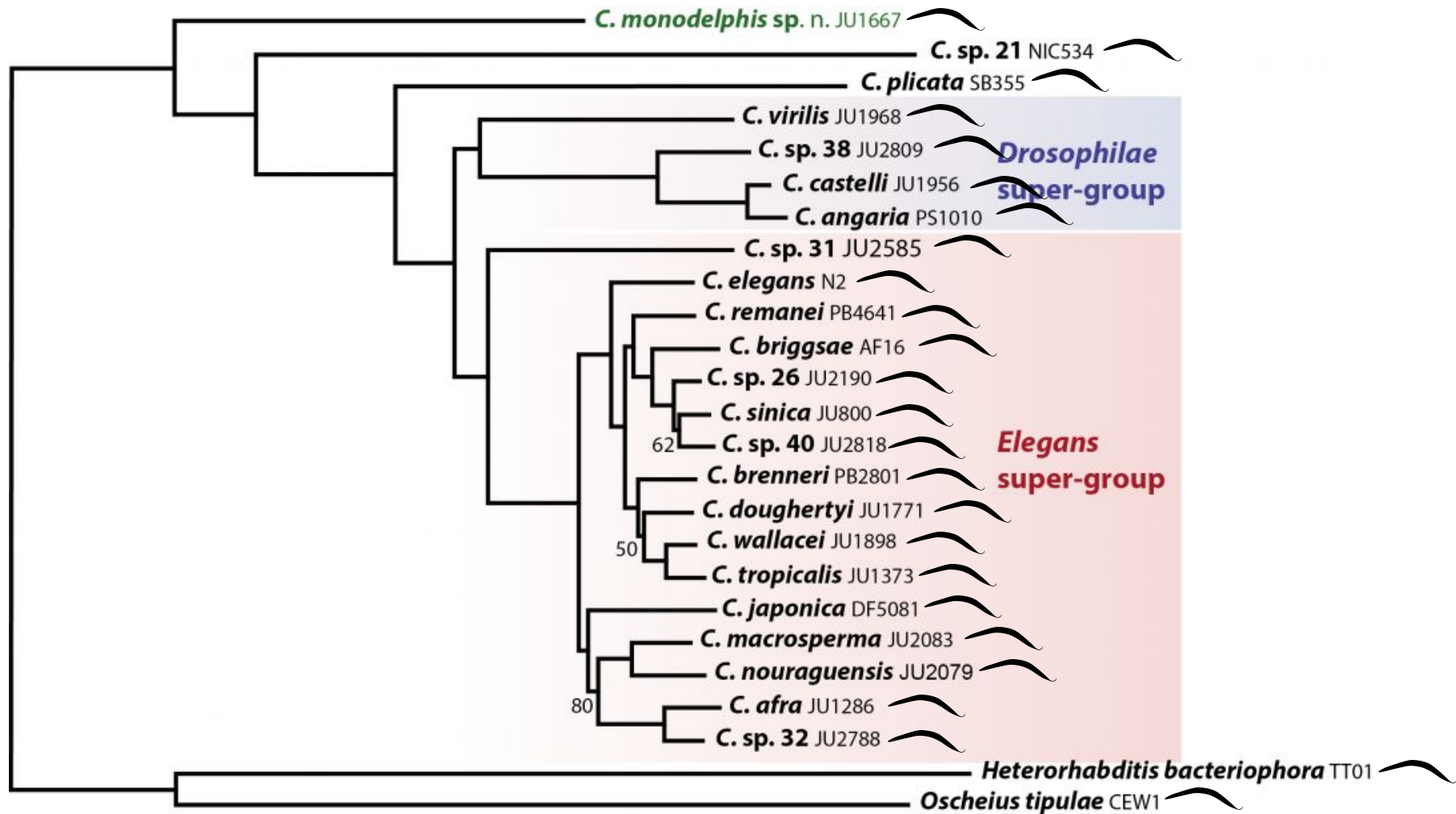
- “Specially constructed MCMC analyses species to each topologies hypothesis...”
- An unrooted tree with  $n$  tips requires  $2(n-3)$  independent MCMC analyses

Faster (while still reasonably accurate) methods (eg. the inflated density ratio; Arima & Tardella, 2014) to estimate marginal likelihoods may become more widely available soon

# Summary

- Bayes factors uncover variation that posterior probabilities obscure
- Different datasets, and approaches to obtain them, can differ in the amount of information they contain
- Some phylogenetic relationships are simply harder to resolve than others
- Data from suspect genes (eg. paralogues) can significantly change topologies





### Base Frequency $\chi^2$ Statistic

Dataset	Amniota	Archosauria +Testudines	Aves	Crocodylia	Diapsida +Testudines	Lepidosauria	Mammalia	Squamata	Testudines
Chiari	0.114	0.018	-0.012	-0.091	0.0317	***	0.0434	-0.0524	-0.041
Crawford	***	0.072	0.165	0.151	***	0.027	***	0.063	0.105
Fong	0.105	0.203	0.194	0.095	0.029	***	0.112	0.097	0.229
Lu	***	0.009	-0.038	***	0.127	-0.013	0.193	0.136	0.046
Shaffer	***	-0.010	0.254	***	***	***	0.312	0.395	***
Wang	-0.118	0.161	0.452	0.543	0.142	***	0.326	***	0.460

### Clockness Likelihood Ratio

Dataset	Amniota	Archosauria +Testudines	Aves	Crocodylia	Diapsida +Testudines	Lepidosauria	Mammalia	Squamata	Testudines
Chiari	-0.018	-0.271	0.049	-0.017	-0.196	***	-0.065	-0.016	-0.248
Crawford	***	-0.212	0.113	-0.145	***	0.040	***	0.067	-0.080
Fong	-0.044	0.024	-0.103	-0.190	-0.219	***	0.075	0.090	0.067
Lu	***	-0.003	0.022	***	0.069	0.204	0.252	0.209	0.040
Shaffer	***	-0.194	0.468	***	***	***	0.449	0.554	***
Wang	-0.190	-0.006	0.301	0.217	-0.009	***	0.148	***	0.187

### Alignment Certainty (Heads Or Tails)

Dataset	Amniota	Archosauria +Testudines	Aves	Crocodylia	Diapsida +Testudines	Lepidosauria	Mammalia	Squamata	Testudines
Chiari	0.068	0.046	-0.076	0.195	-0.068	***	0.074	0.102	0.071
Crawford	***	0.002	-0.079	-0.180	***	-0.017	***	-0.100	-0.087
Fong	-0.001	0.053	-0.013	-0.024	-0.088	***	-0.105	0.040	0.065
Lu	***	-0.123	-0.006	***	-0.133	-0.043	-0.170	-0.151	-0.012
Shaffer	***	-0.078	-0.340	***	***	***	-0.326	-0.351	***
Wang	0.038	-0.103	-0.242	-0.249	-0.044	***	-0.193	***	-0.180

### Percent Missing Data

Dataset	Amniota	Archosauria +Testudines	Aves	Crocodylia	Diapsida +Testudines	Lepidosauria	Mammalia	Squamata	Testudines
Chiari	-0.160	-0.135	-0.021	0.057	0.072	***	0.034	0.036	-0.262
Crawford	***	0.097	0.085	0.101	***	0.036	***	0.050	0.057
Fong	0.348	0.131	0.145	-0.095	0.048	***	-0.029	-0.225	0.043
Lu	***	0.109	-0.147	***	0.188	0.006	0.186	-0.055	-0.086
Shaffer	***	0.061	0.311	***	***	***	0.326	0.325	***
Wang	-0.034	-0.056	-0.100	-0.070	0.032	***	-0.010	***	0.023

### Rate of Evolution

Dataset	Amniota	Archosauria +Testudines	Aves	Crocodylia	Diapsida +Testudines	Lepidosauria	Mammalia	Squamata	Testudines
Chiari	0.105	0.075	0.119	0.073	0.054	***	0.077	-0.061	0.097
Crawford	***	0.036	0.262	0.371	***	0.039	***	0.116	0.184
Fong	-0.044	0.139	0.075	0.119	0.16	***	0.224	0.067	0.128
Lu	***	0.052	-0.002	***	0.056	-0.035	0.184	0.080	0.010
Shaffer	***	-0.040	0.108	***	***	***	0.189	0.169	***
Wang	-0.221	0.023	0.176	0.242	0.045	***	0.197	***	0.134

## Table S2

Rank correlation coefficients between  $2\ln(\text{BF})$  support values and various characteristics of genes that have been hypothesized to influence the reliability of phylogenetic signal.

Blue cells indicate significant positive correlations and red cells indicate significant negative correlations.

The strength of the color is proportional to the correlation, when the value is significantly different from 0.