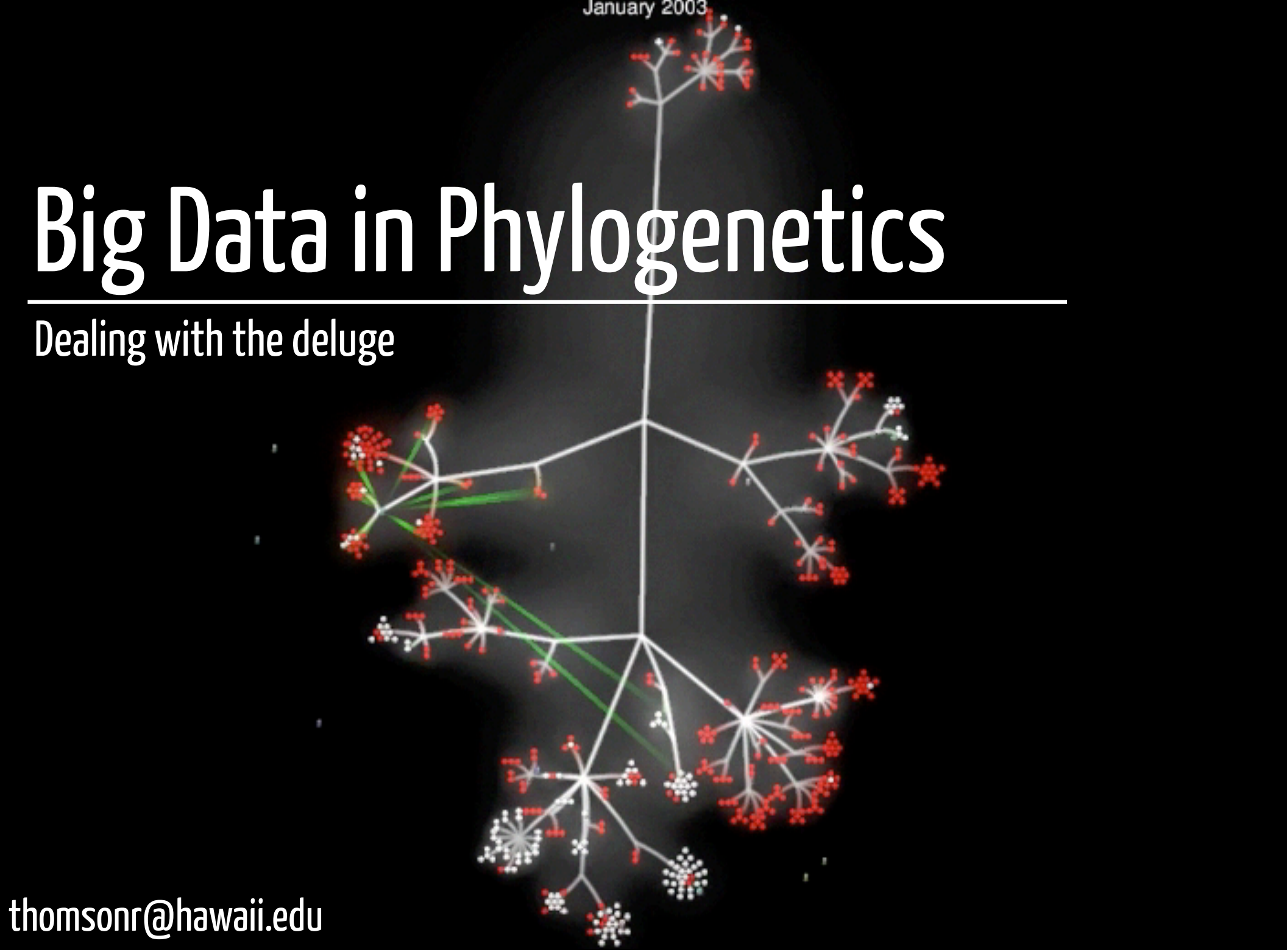
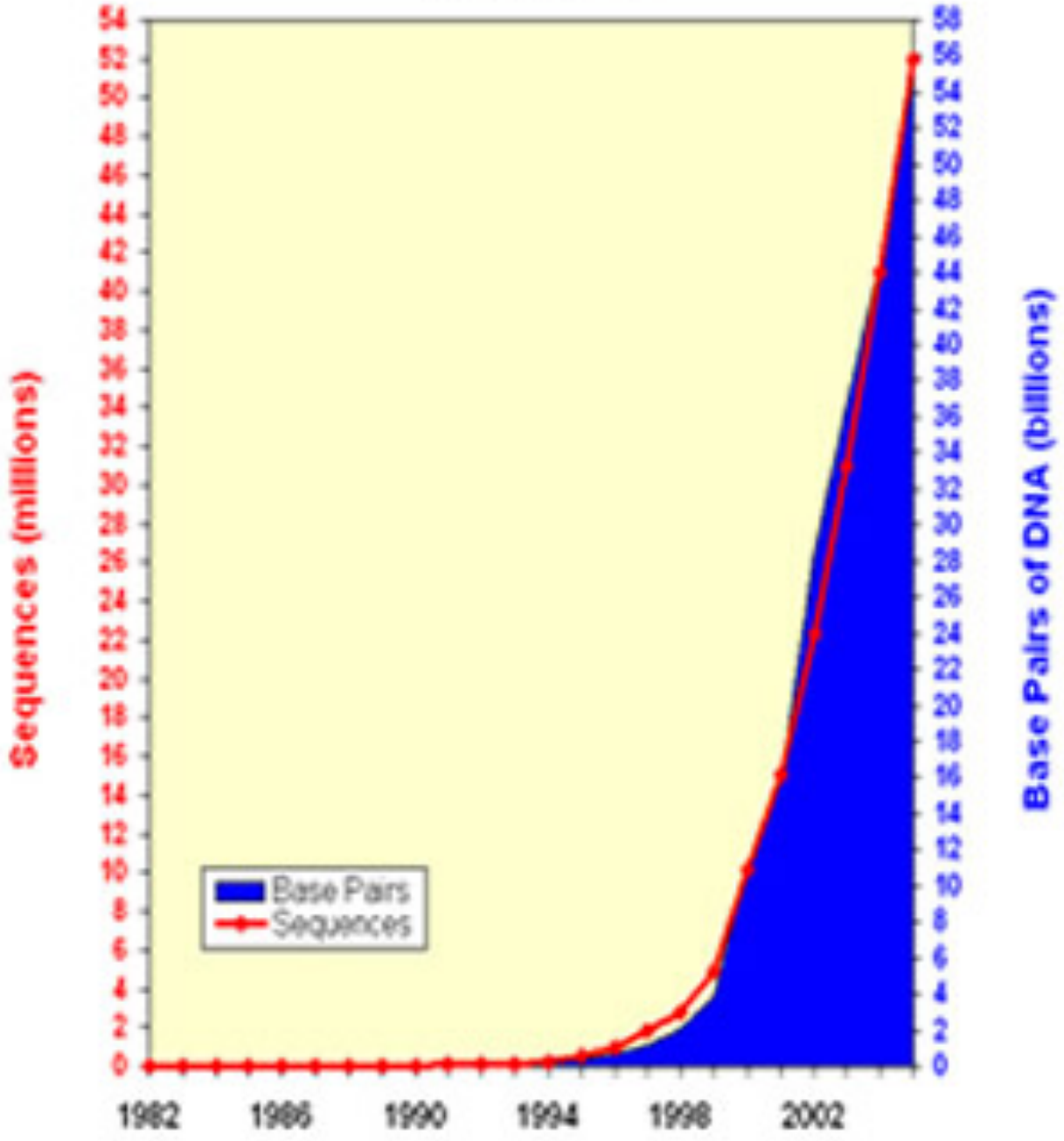


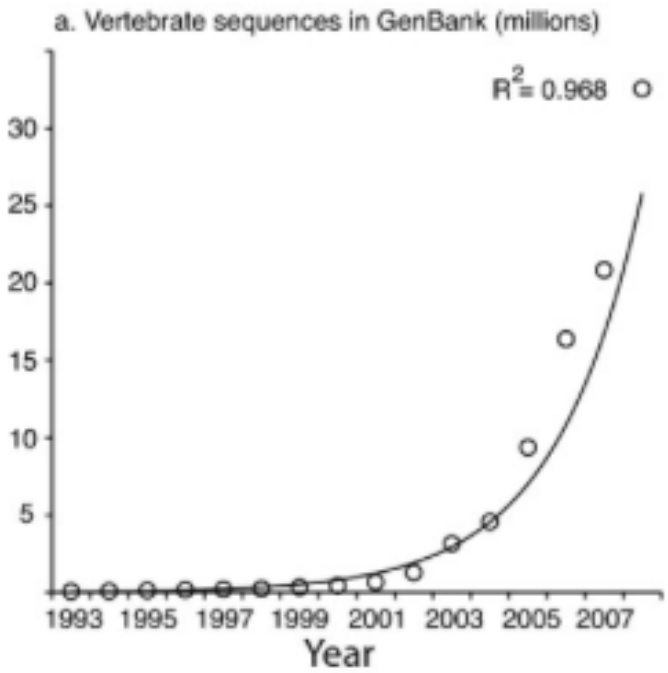
Big Data in Phylogenetics

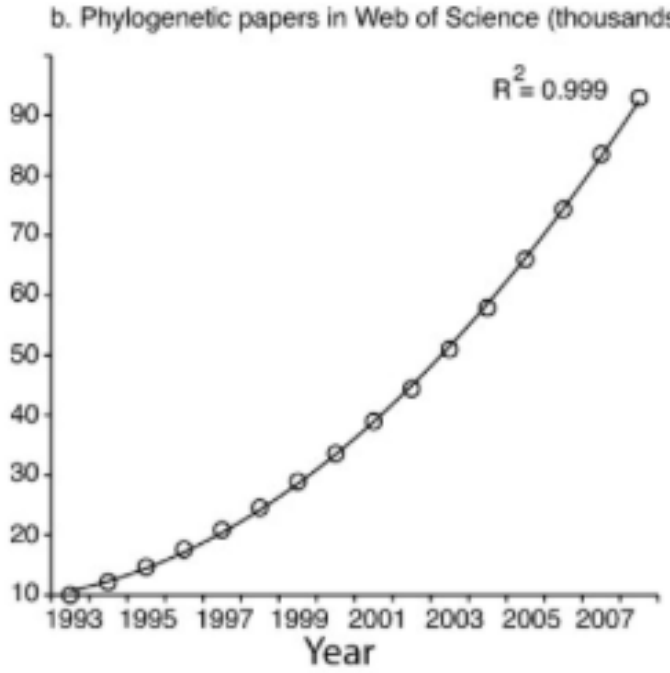
Dealing with the deluge

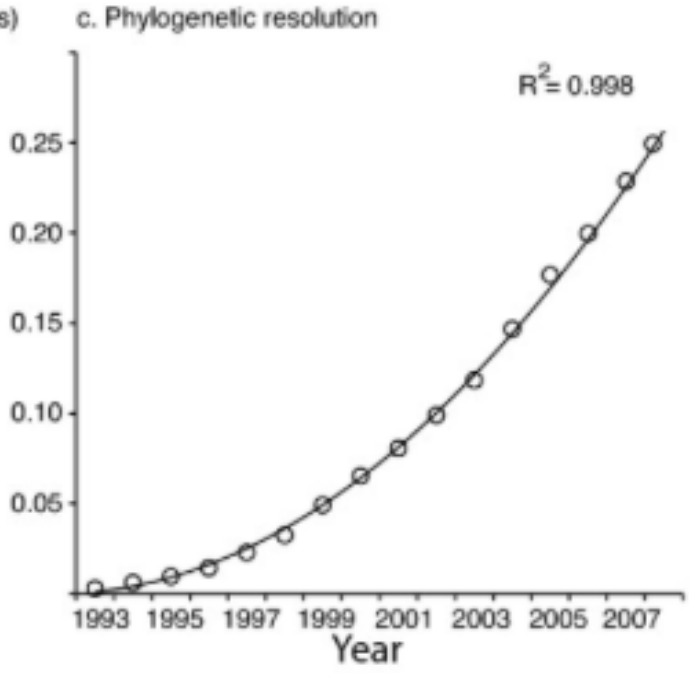
thomsonr@hawaii.edu


Growth of GenBank
(1982 - 2005)
Sequences (millions)
Base Pairs of DNA (billions)
Base Pairs
Sequences
1982
1986
1990
1994
1998
2002

skipping ~30 figures that all show growth of data or decreasing cost



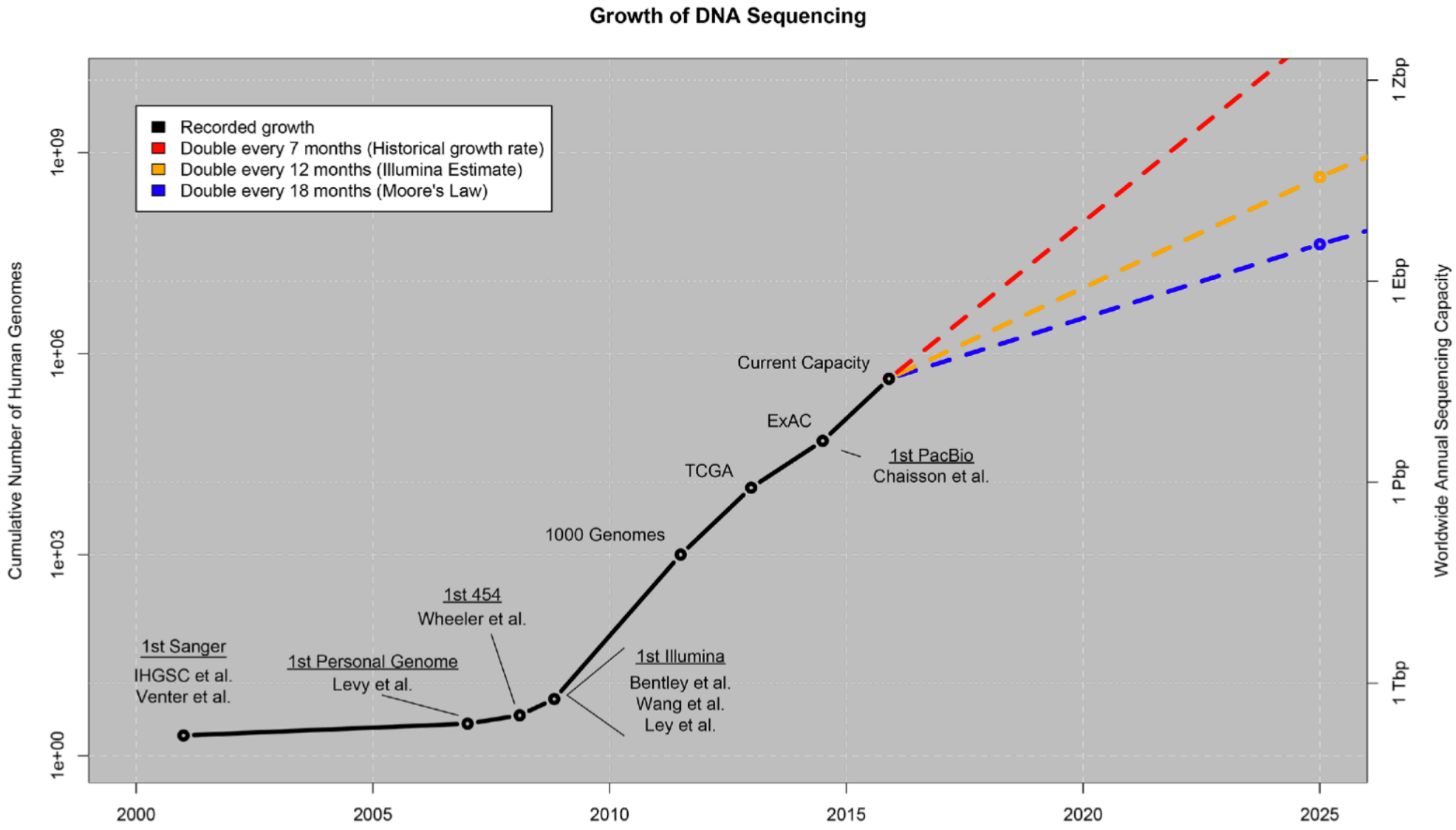




Thomson and Shaffer 2010

PERSPECTIVE

Big Data: Astronomical or Genomical?



Stephens et al. 2015

To what end?

"We think big data is what everyone cares about.
It's not.
It's stories."

- Dr. Jessica Utts
President, American Statistical Association

The goal is to gather 'sufficient' data in order to answer a question 'robustly.'

To what end?

"We think big data is what everyone cares about.
It's not.
It's stories."

- Dr. Jessica Utts
President, American Statistical Association

The goal is to gather 'sufficient' data in order to answer a question 'robustly.'

The question is what is interesting.

To what end?

"We think big data is what everyone cares about.
It's not.
It's stories."

- Dr. Jessica Utts
President, American Statistical Association

The goal is to gather 'sufficient' data in order to answer a question 'robustly.'

The question is what is interesting.

This is no different than it's always been.

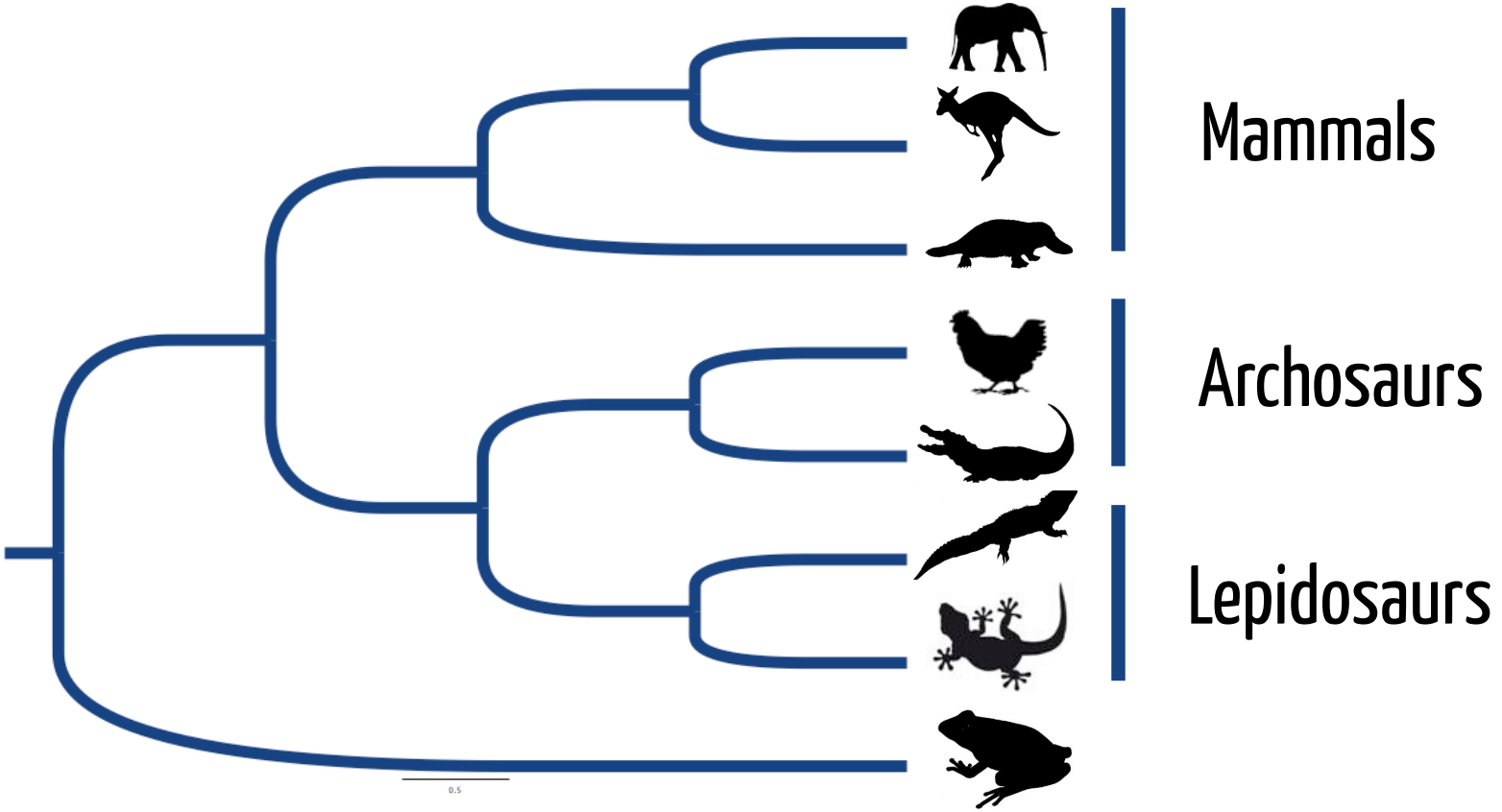
A case study

- A very specific question:
  - What are the phylogenetic affinities of turtles?

- Brings up more general issues:
  - How do we approach difficult phylogenetic problems?
  - How **should** we approach difficult phylogenetic problems?

Turtle Phylogenetics

- Overarching problem:
  - Where do turtles sit in the amniote tree?



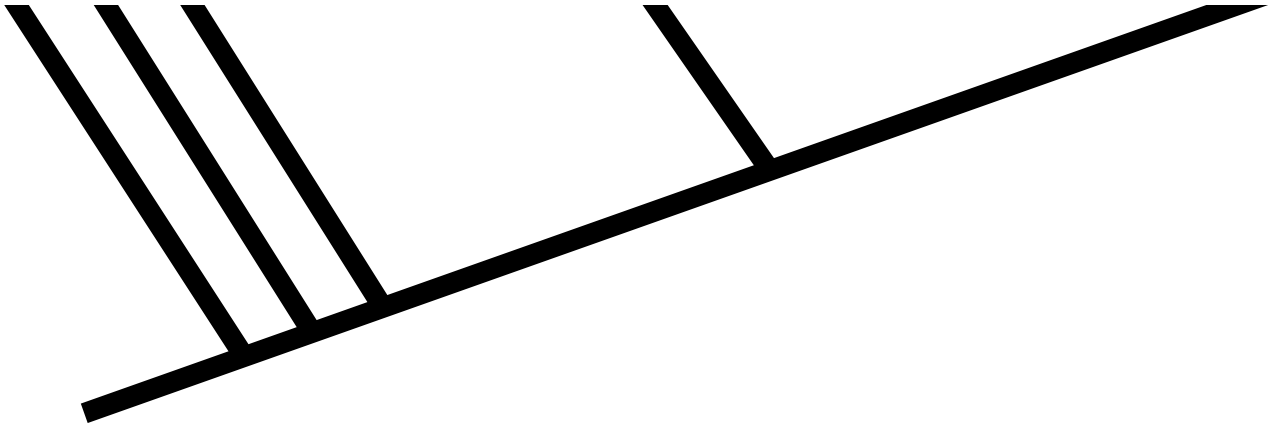
Osteology

- Early approaches relied on osteology (primarily of the skull)

Anapsid

Diapsid

Synapsid

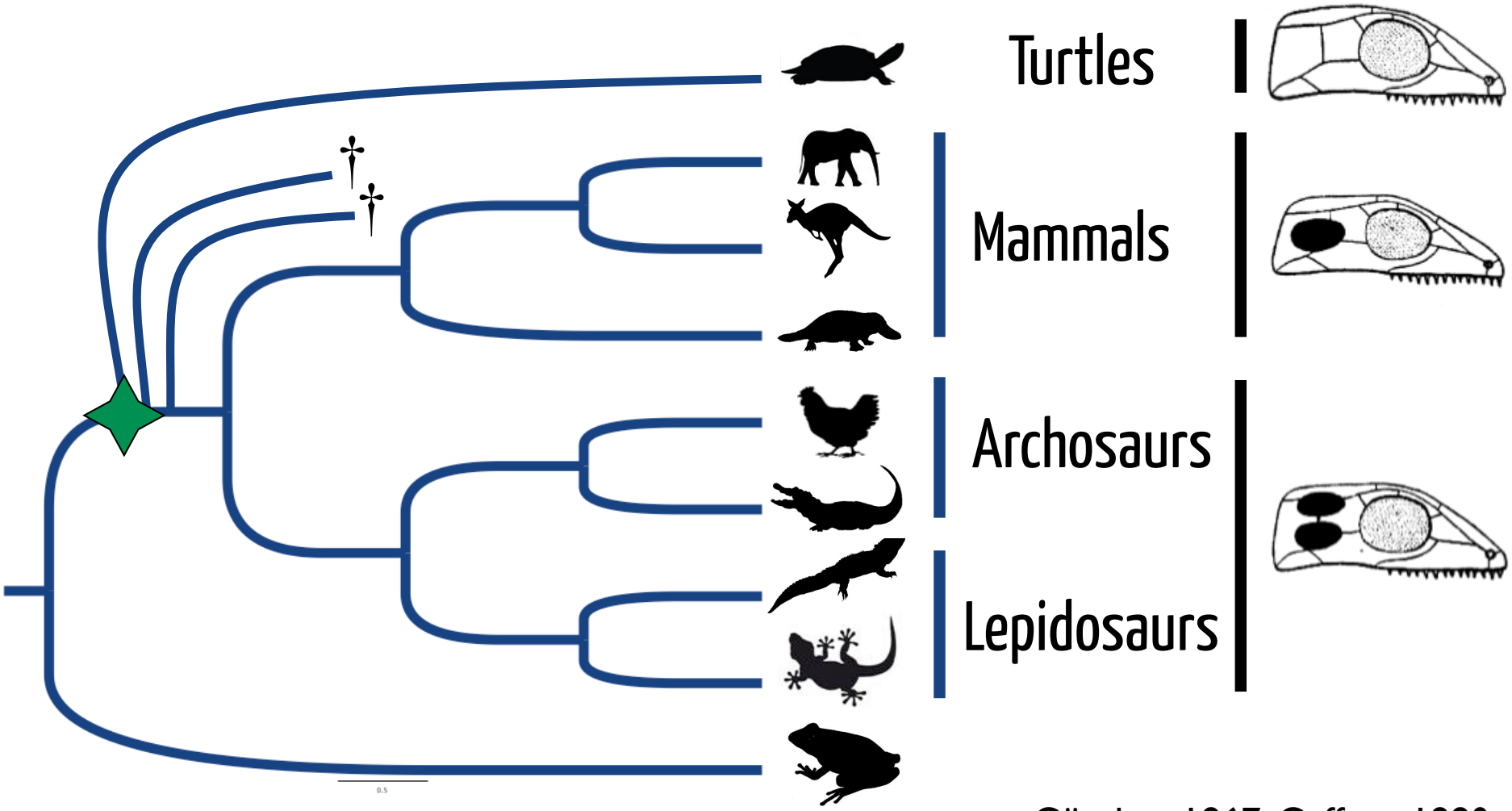


Osteology

- Early appr
  - Anaps

Osteology

- Early approaches relied on osteology (primarily of the skull)

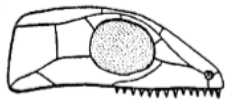


Günther 1867, Gaffney 1980

Osteology

- Primary issue with this hypothesis

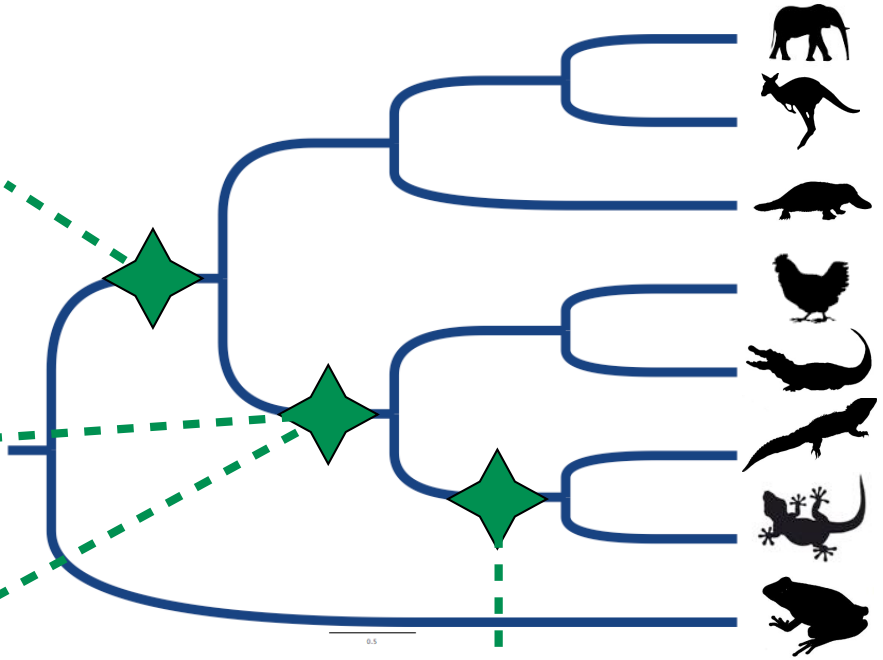
More osteology



Gaffney 1980

Lee 2001 Contrib. to Zool

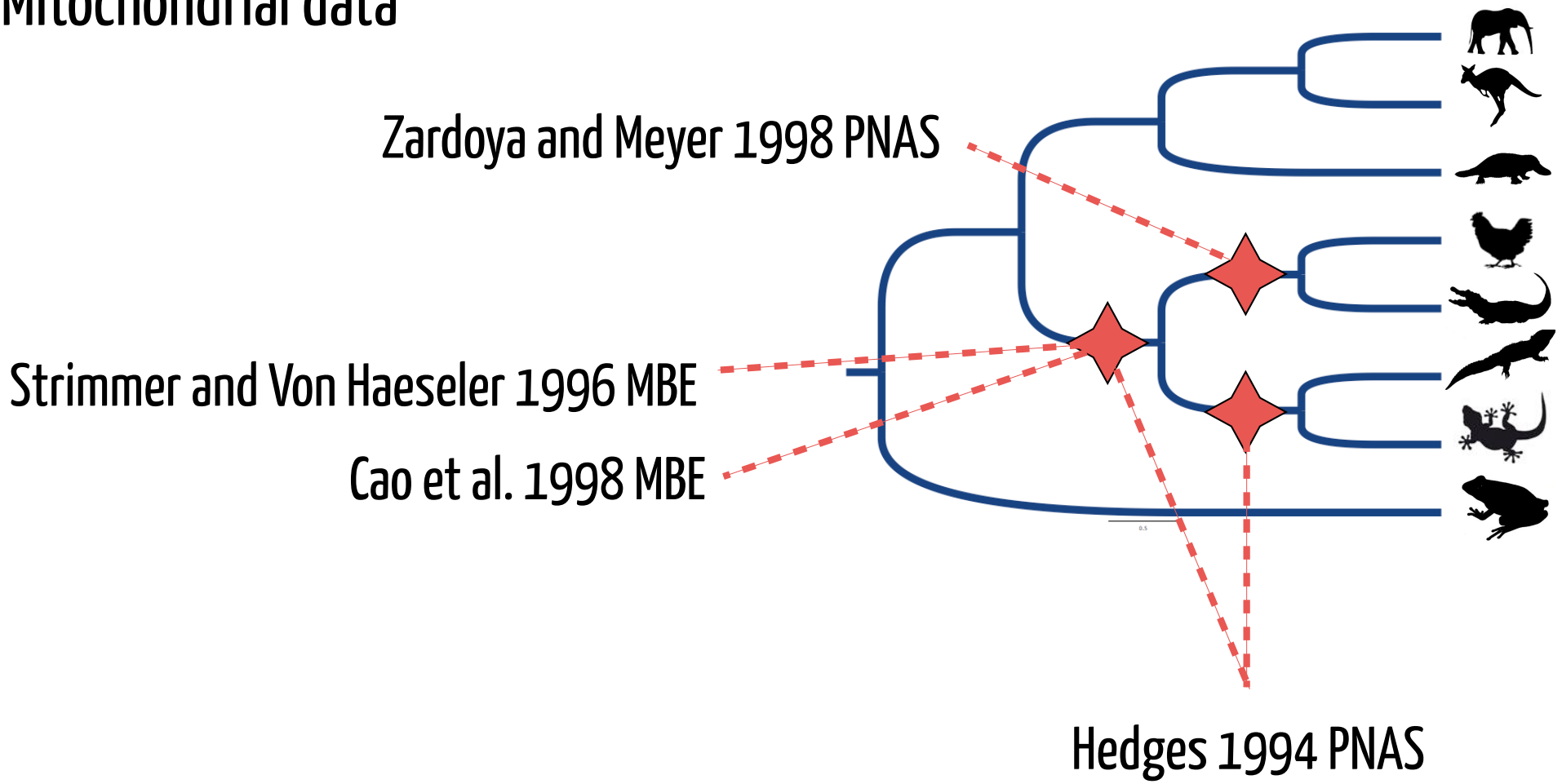
Lyson et al. 2010 Biol. Lett.



Reippel and deBraga 1996 Nature

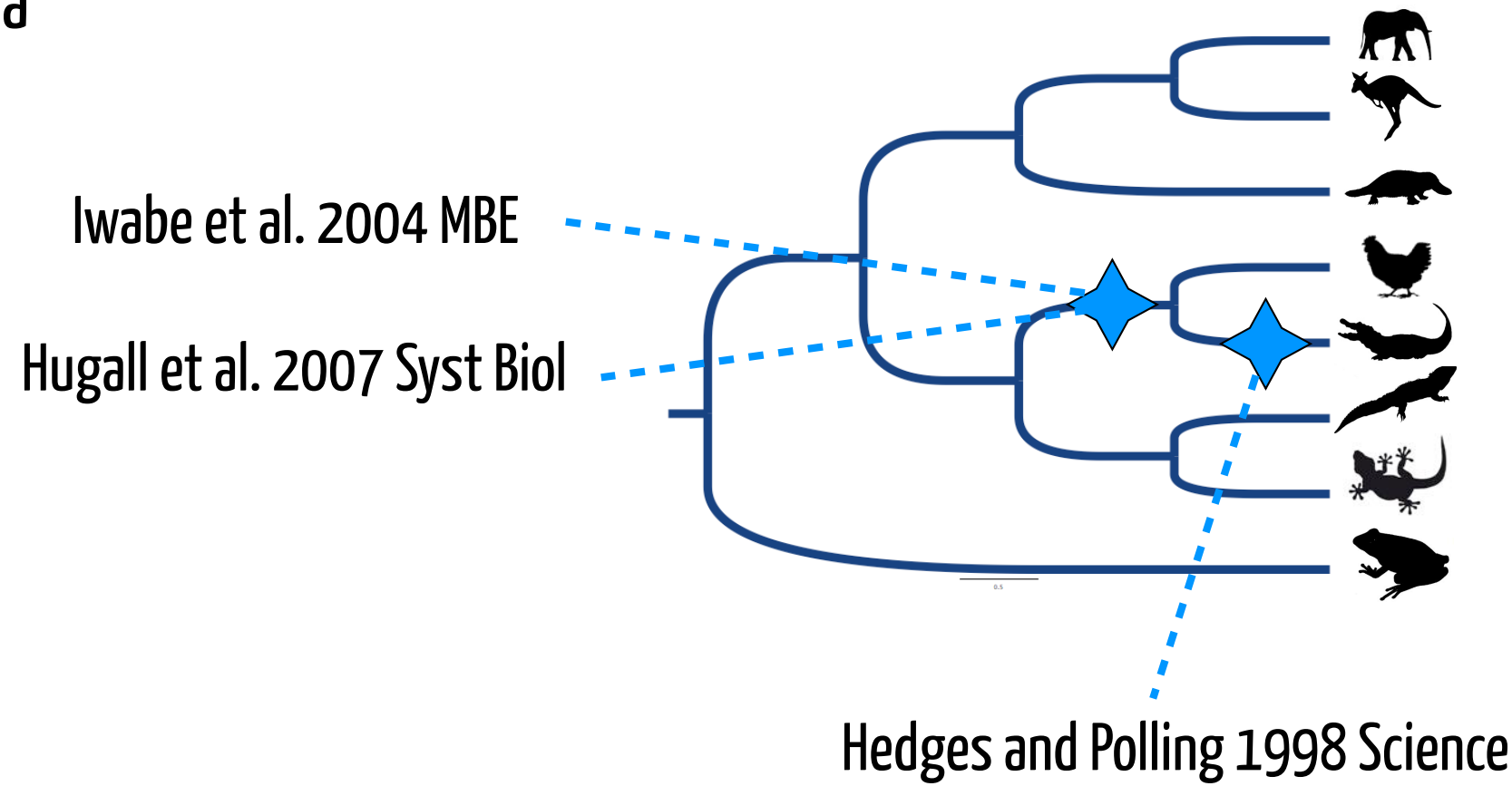
Molecular Information

- Mitochondrial data

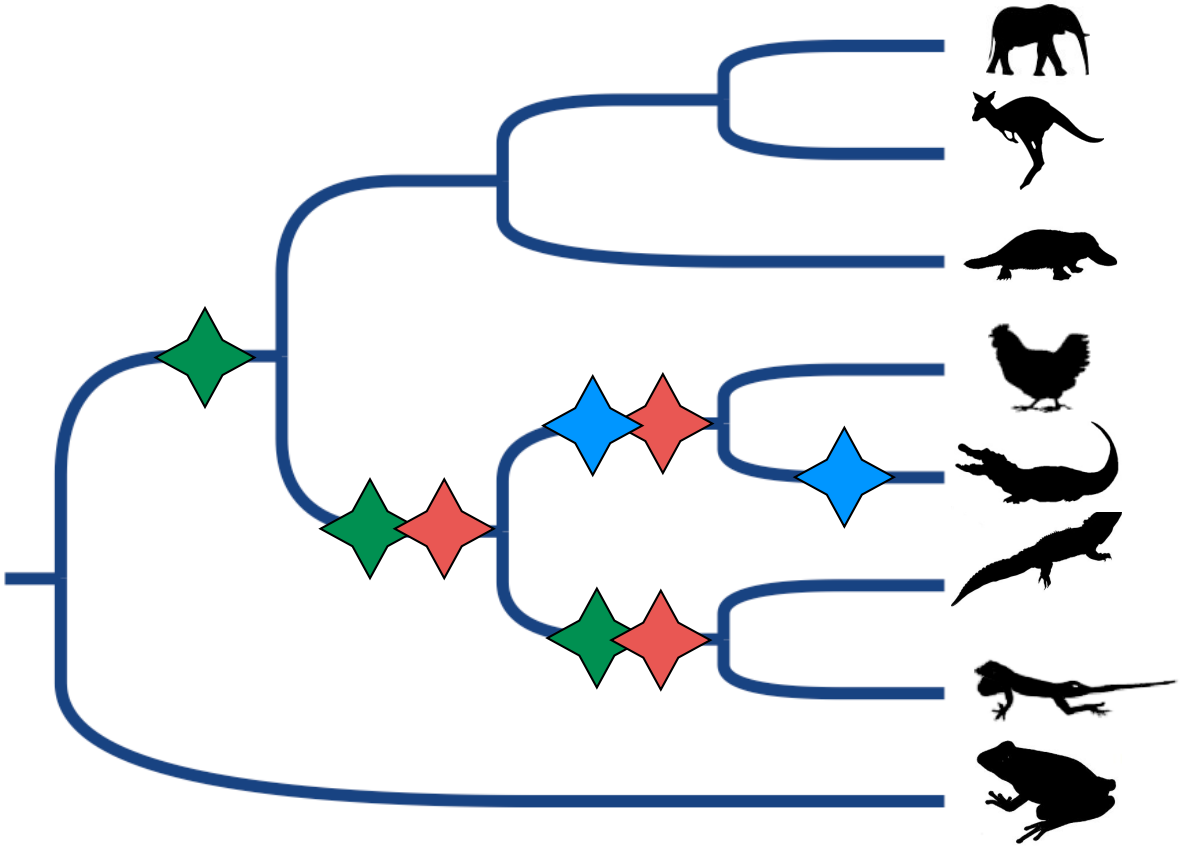


Molecular Information

- Nuclear data



Summary



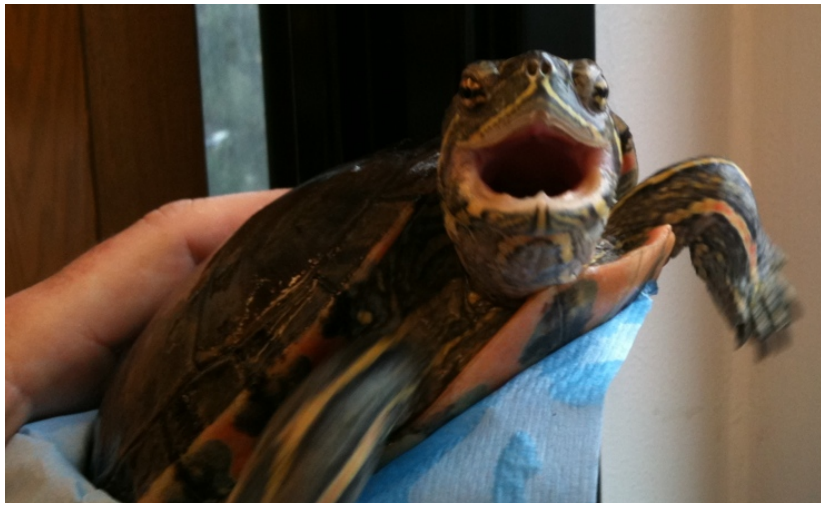
Osteology

mtDNA

nuDNA

Turtle Genomics

- 3 genome consortia
- Several more independent studies

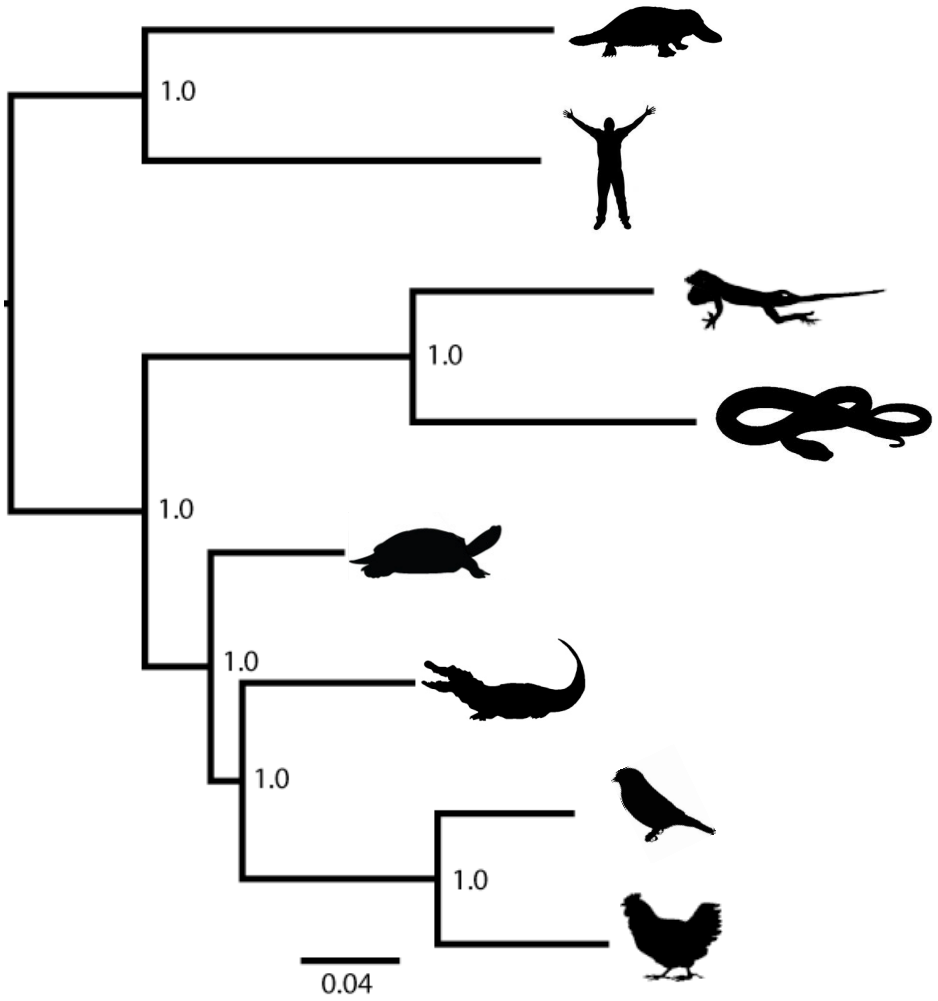


National Human Genome Research Institute

华大基因 BGI

RIKEN

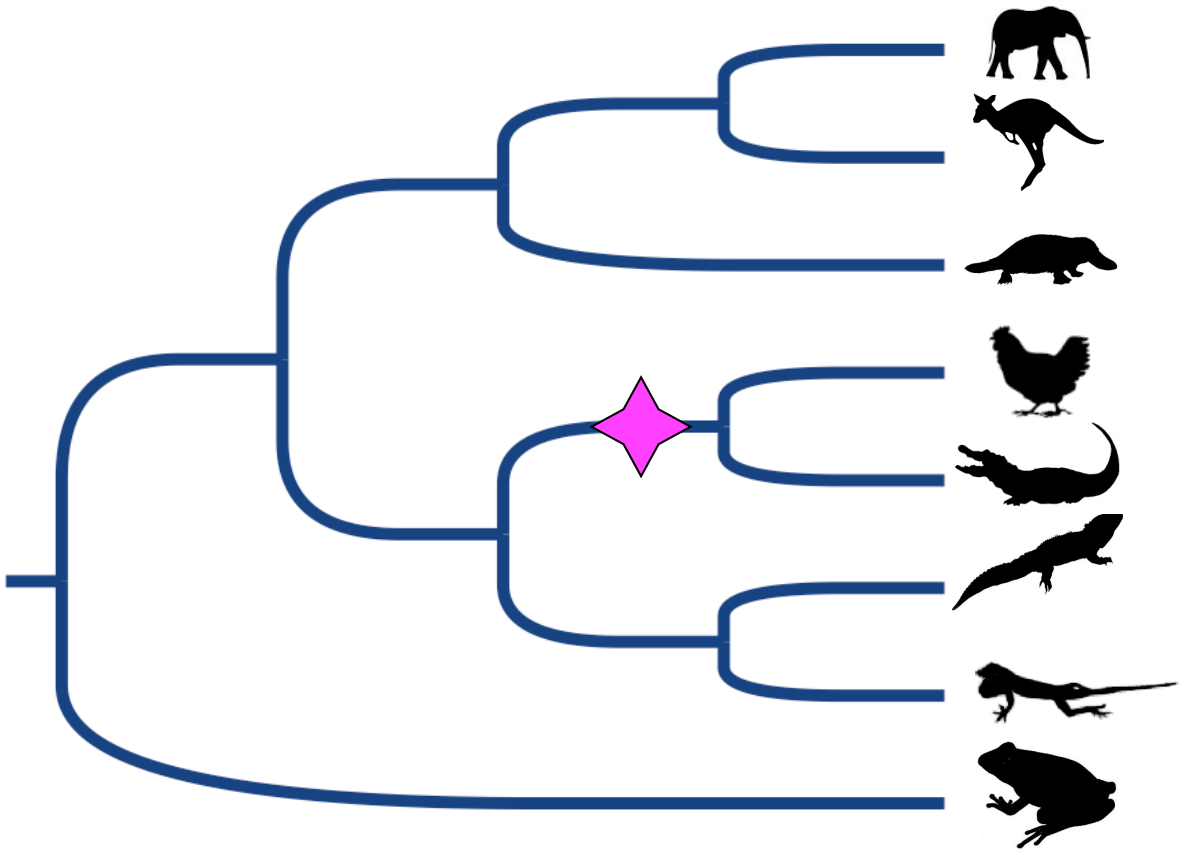
Phylogenomics



Shaffer et al. 2013 Genome Biol

Phylogenomics

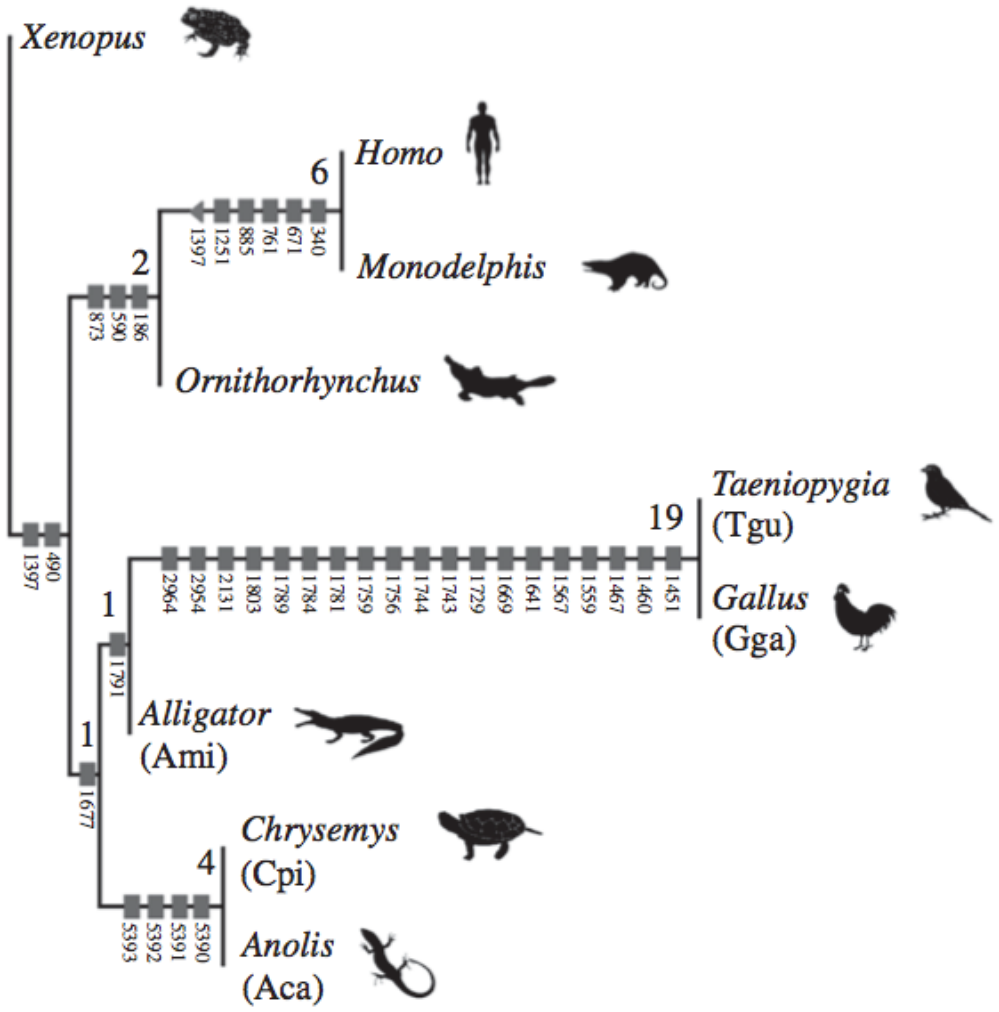
- All analyses agree!



MicroRNA Result

biology letters
Phylogeny
Biol. Lett.
doi:10.1098/rsbl.2011.0477
Published online

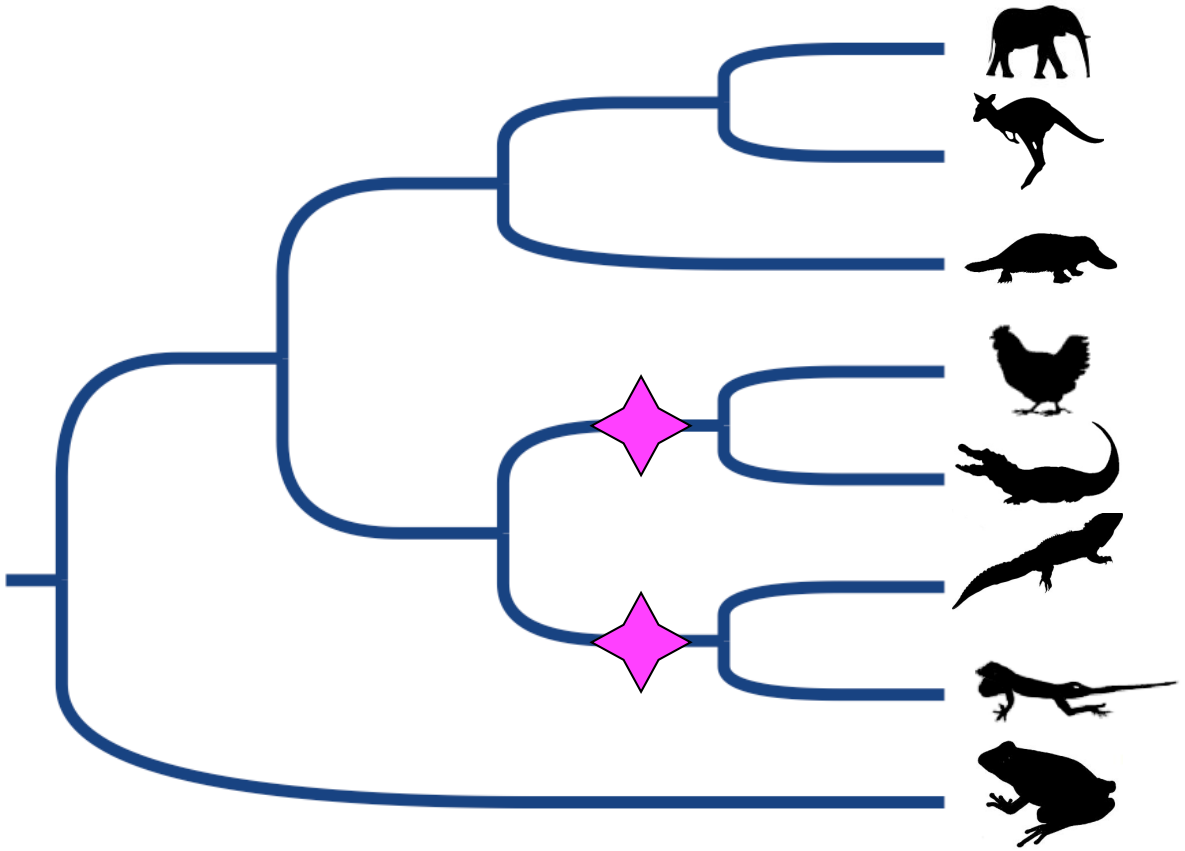
MicroRNAs support a turtle + lizard clade



Lyson et al. 2011 Biol Lett

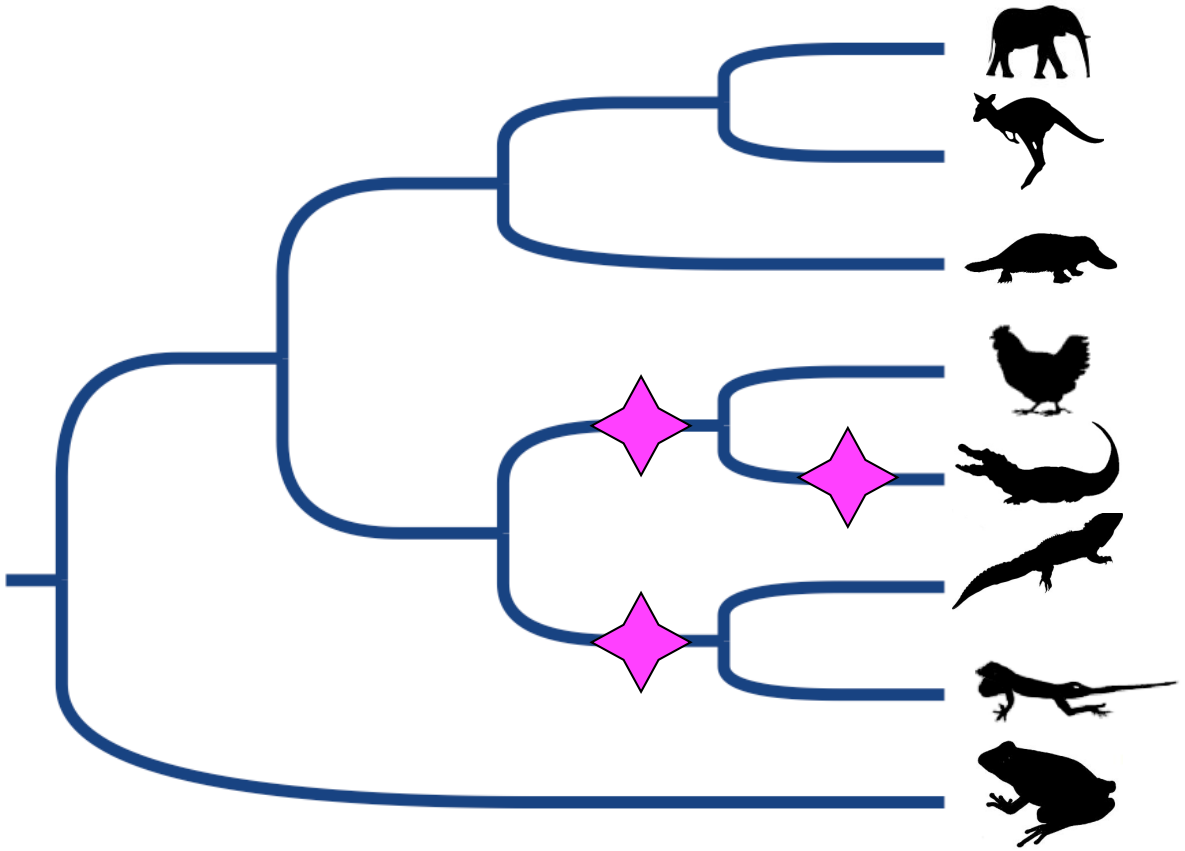
Summary

- Ugh...so what do we do?



Summary

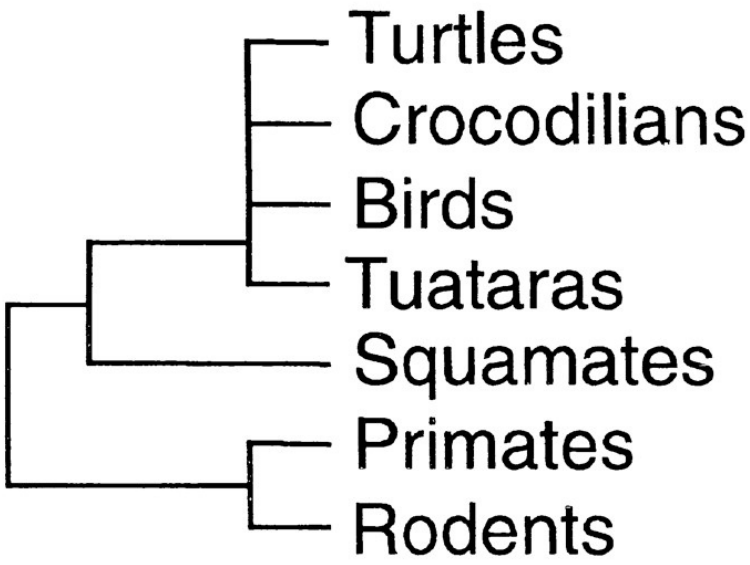
- Ugh…so what do we do?



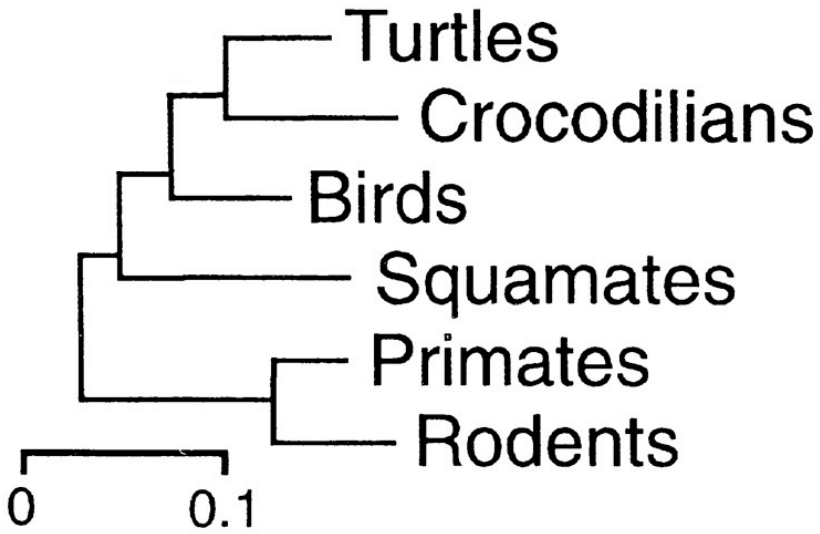
Data in Phylogenetics

- Let's take a step back.
- How have we been approaching this (and most other) phylogenetic questions?

4 nuclear genes



11 nuclear genes

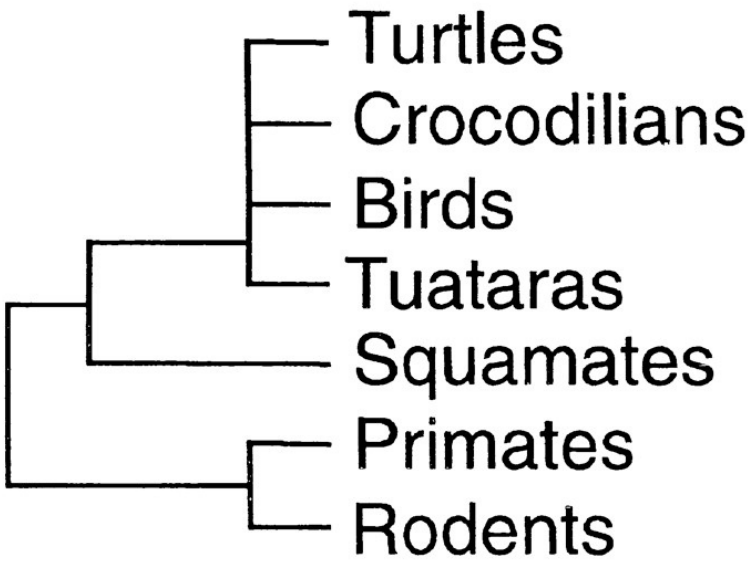


Hedges and Polling 1998 Science

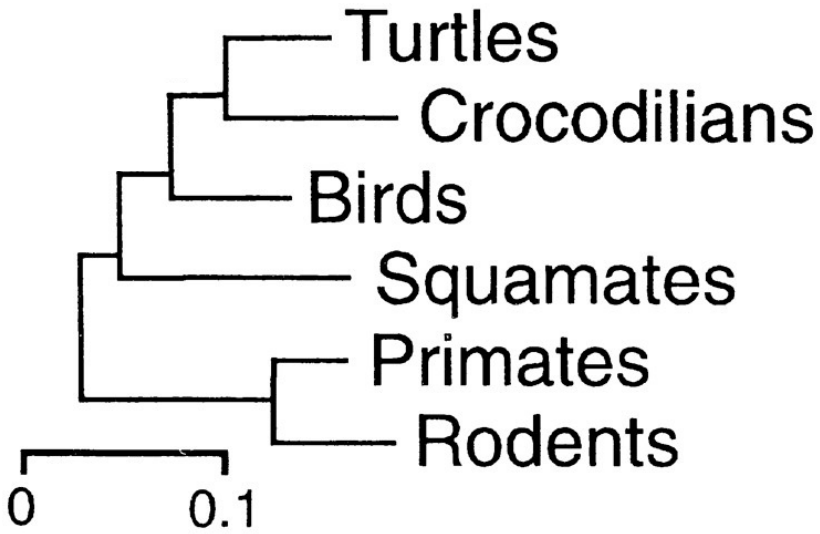
Data in Phylogenetics

- Let's take a step back.
- How have we been approaching this (and most other) phylogenetic questions?

4 nuclear genes



11 nuclear genes

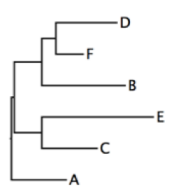


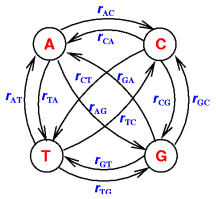
A data centric view

Hedges and Polling 1998 Science

Phylogenomics

- Inferences result from <u>both</u> data and the model

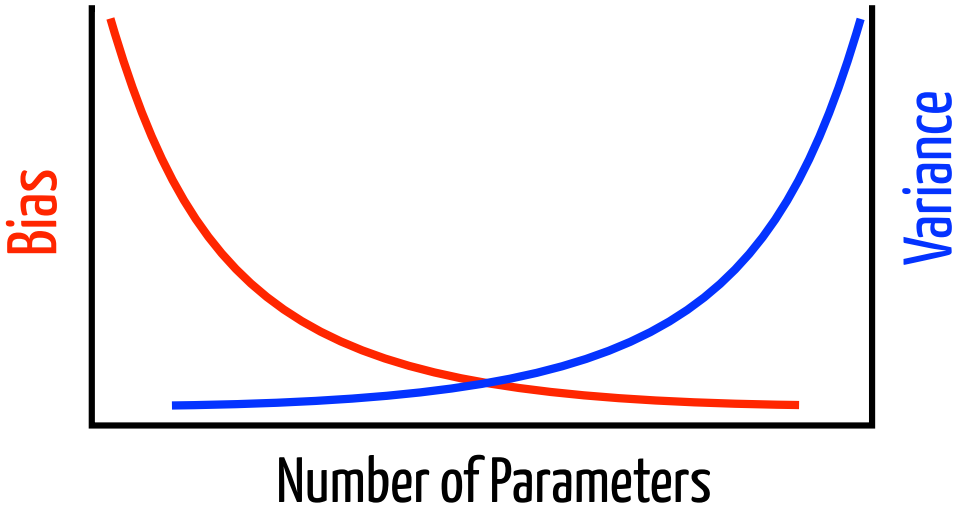
TAGGT
TTGGT
Prob(TTGGT | 

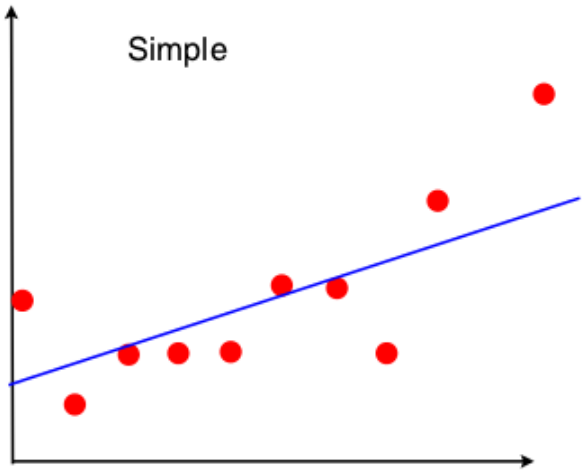


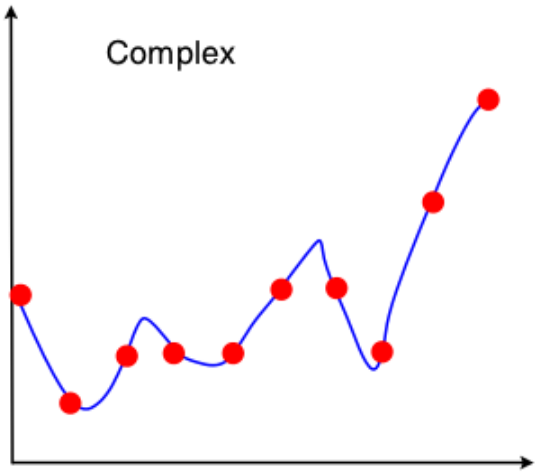
Frequencies = (A, C, T,G))
TTCTT
TTCTT
TTCAT

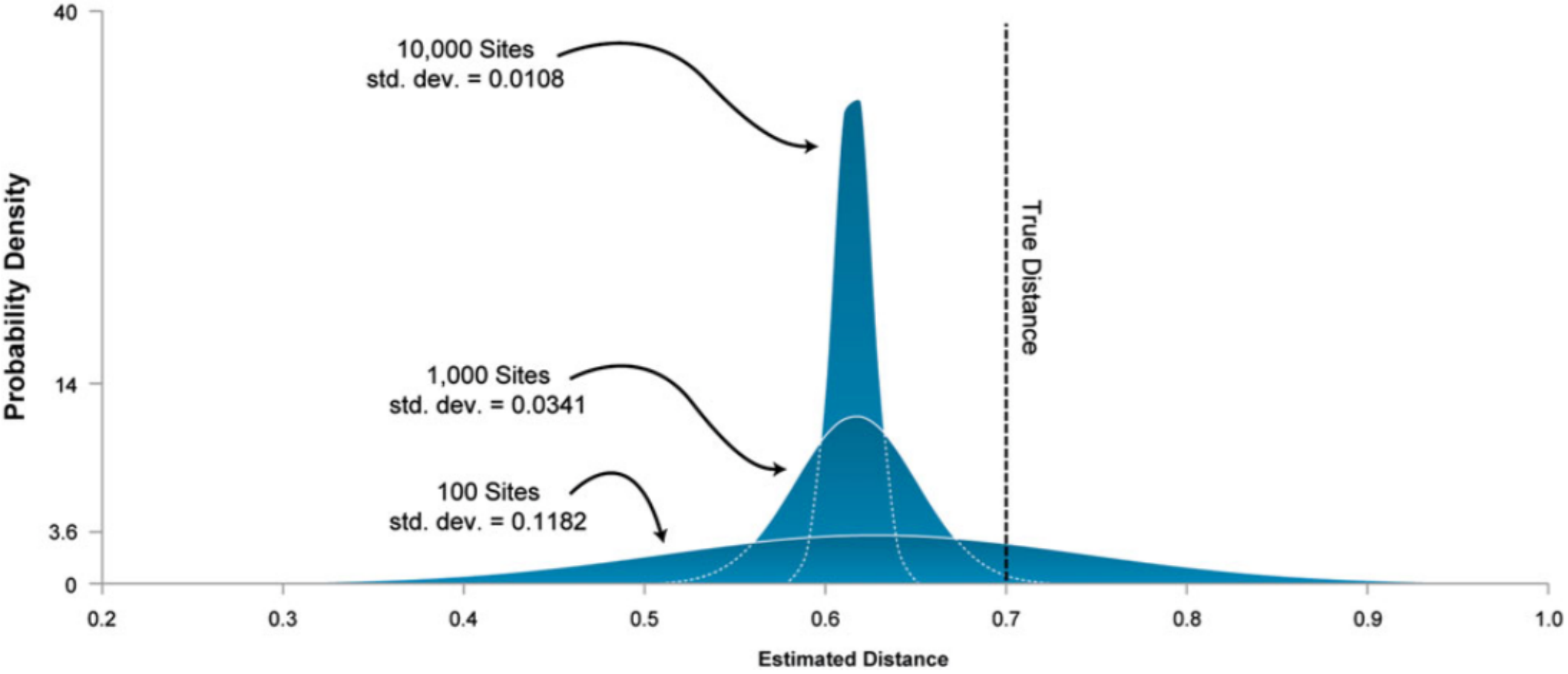
Why does this matter?

- In developing a statistical model for a problem, we inevitably make a tradeoff





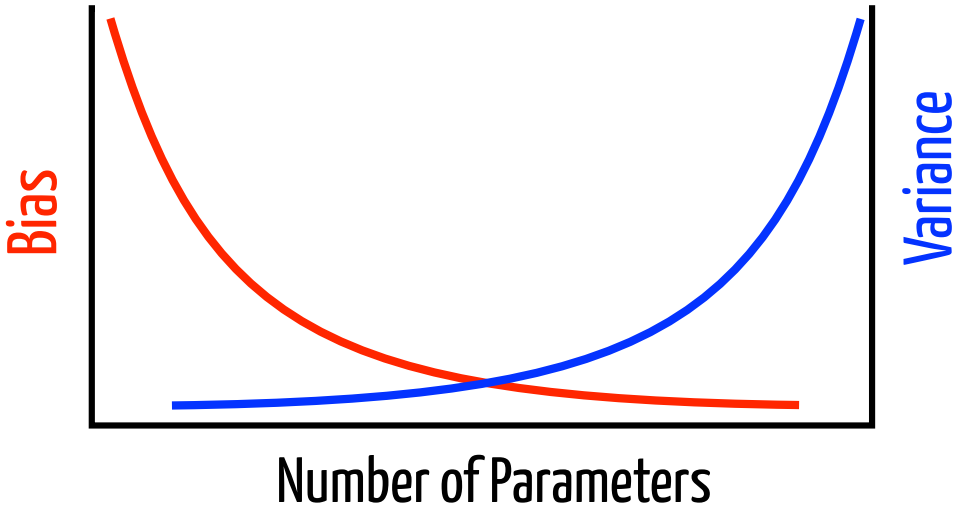


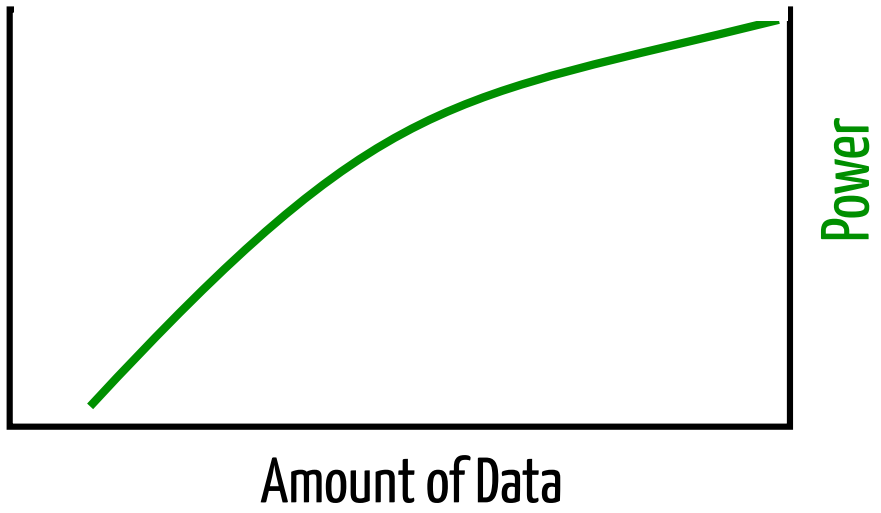


Kumar et al. 2012 MBE

Why does this matter?

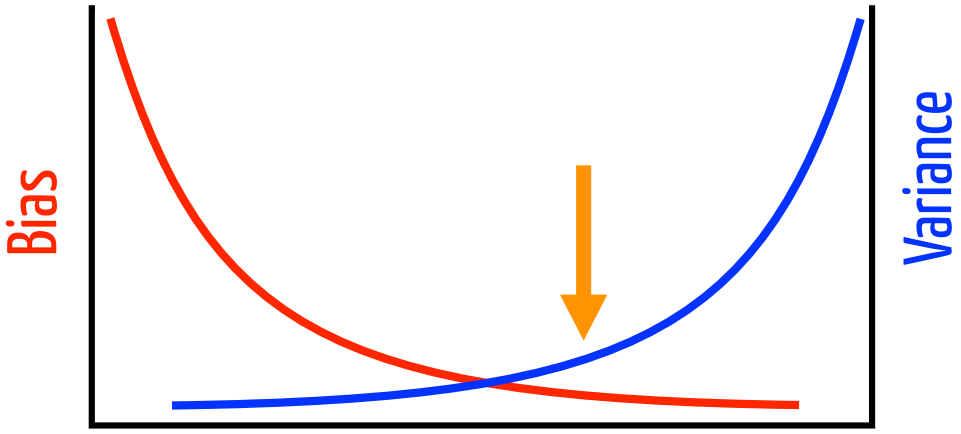
- In developing a statistical model for a problem, we inevitably make a tradeoff

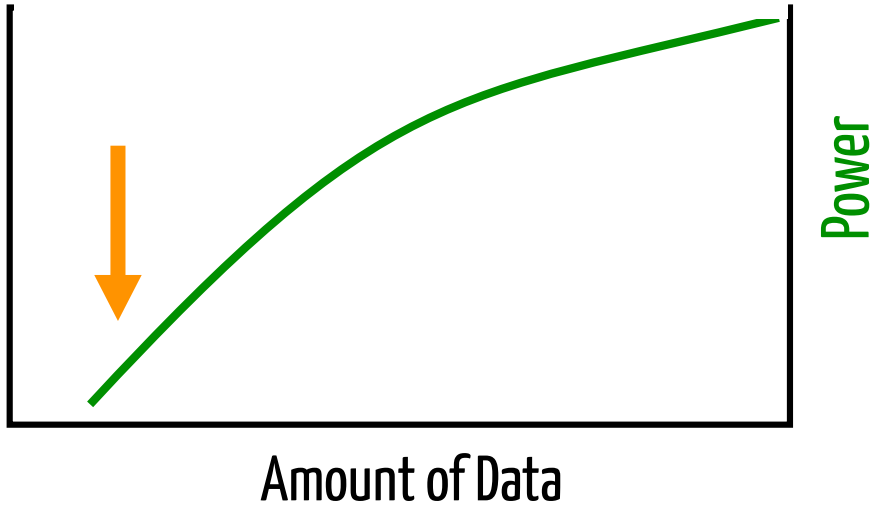




Why does this matter?

- In developing a statistical model for a problem, we inevitably make a tradeoff





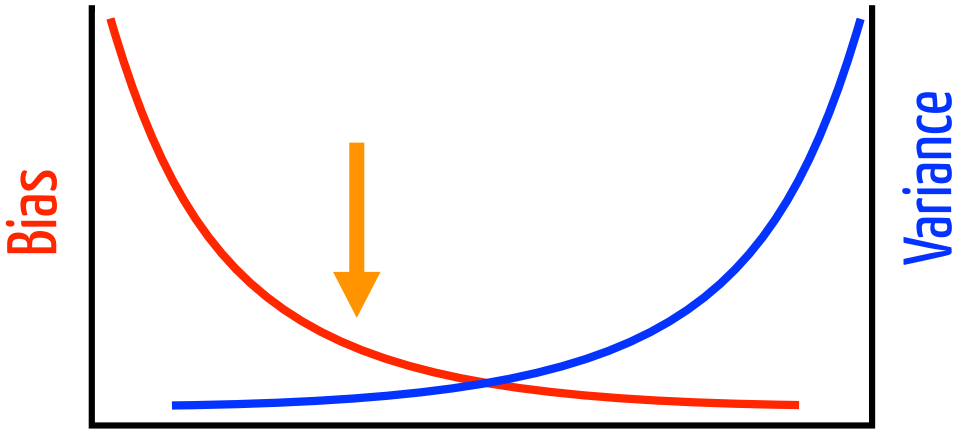
1 gene

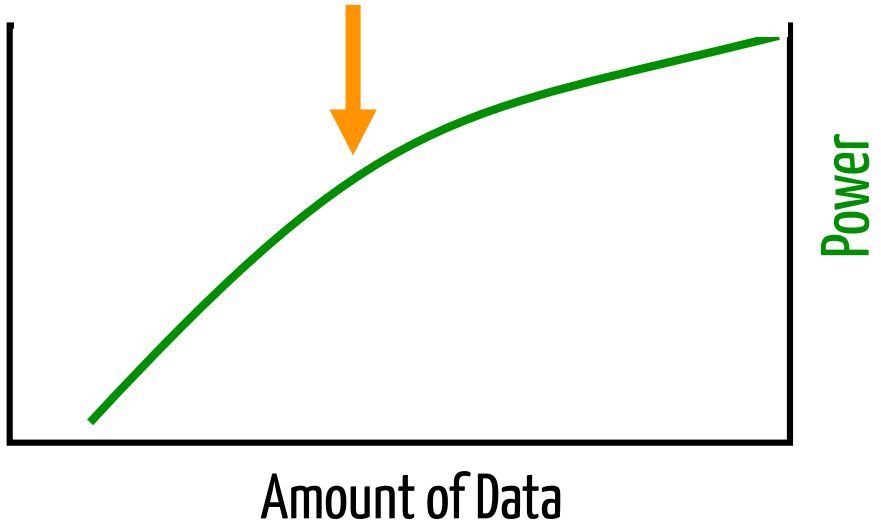
10 genes

1000 genes

Why does this matter?

- In developing a statistical model for a problem, we inevitably make a tradeoff





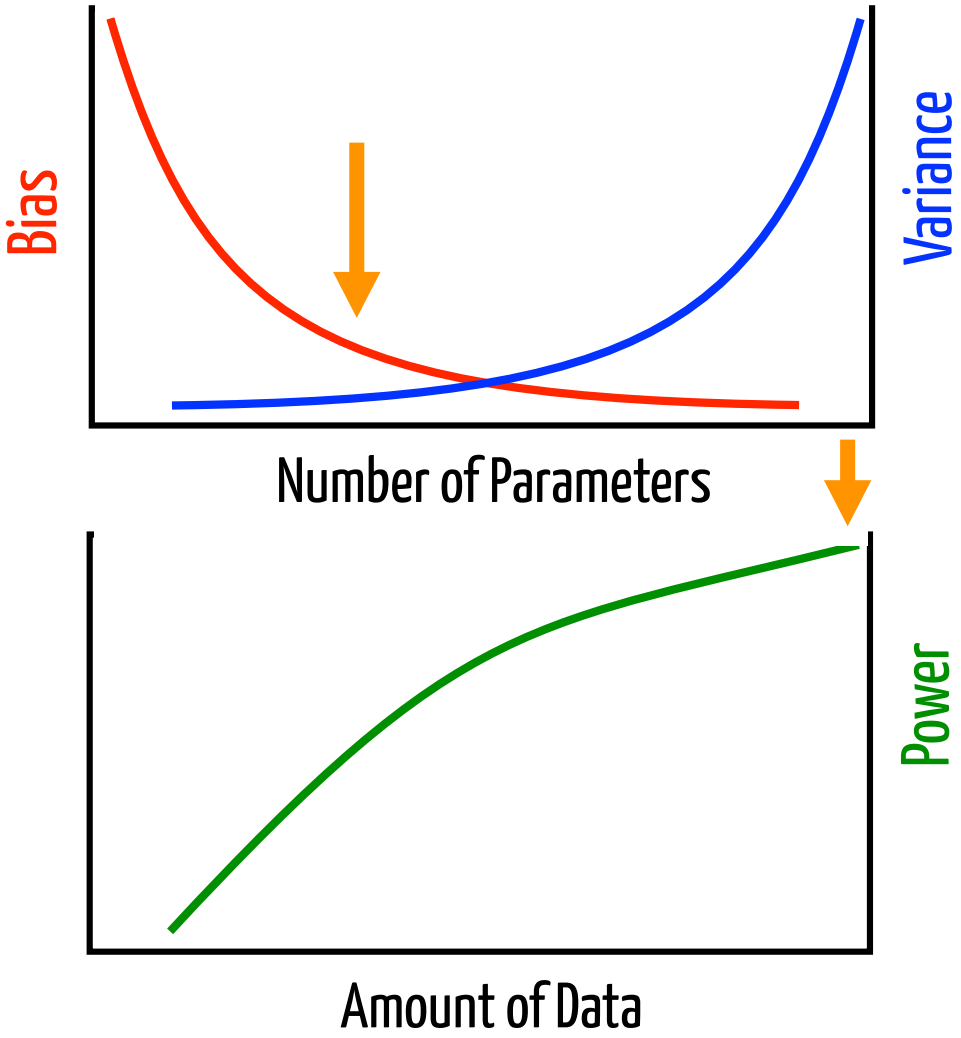
1 gene

10 genes

1000 genes

Why does this matter?

- In developing a statistical model for a problem, we inevitably make a tradeoff



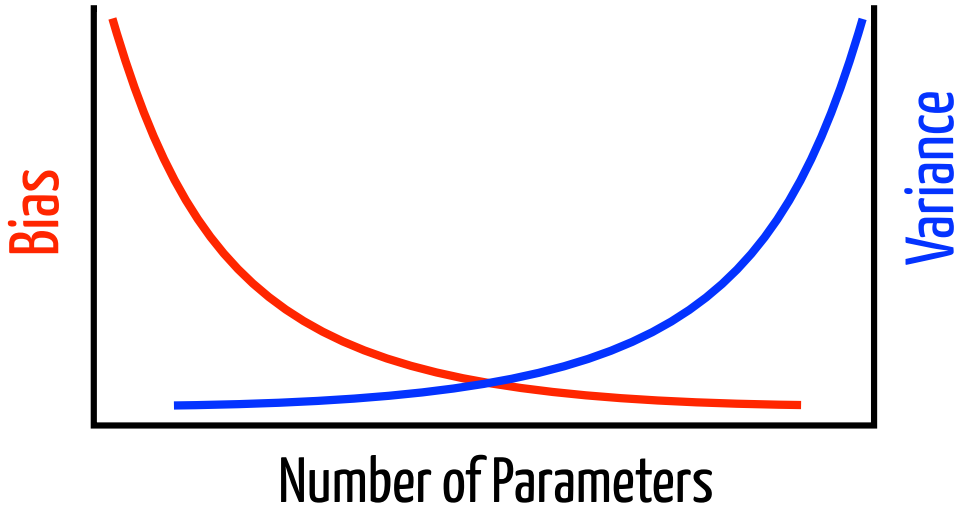
1 gene

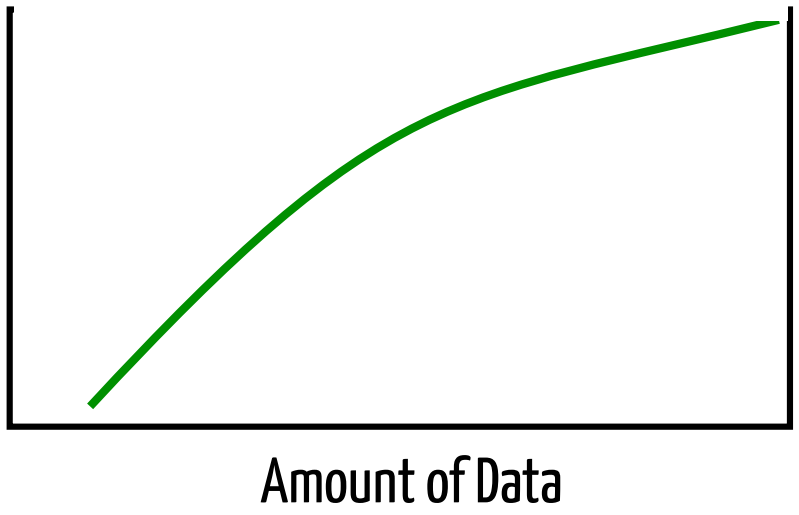
10 genes

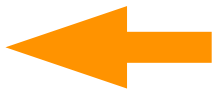
1000 genes

Why does this matter?

- The point.



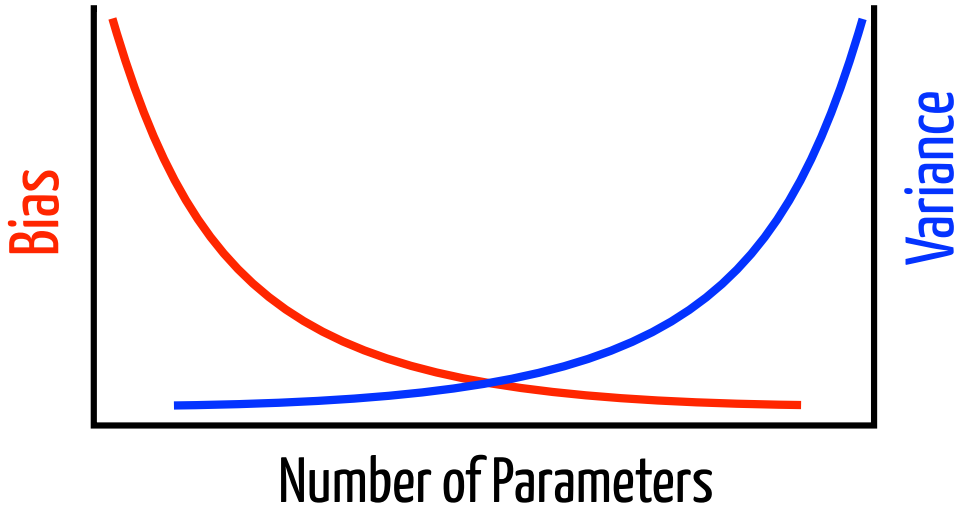


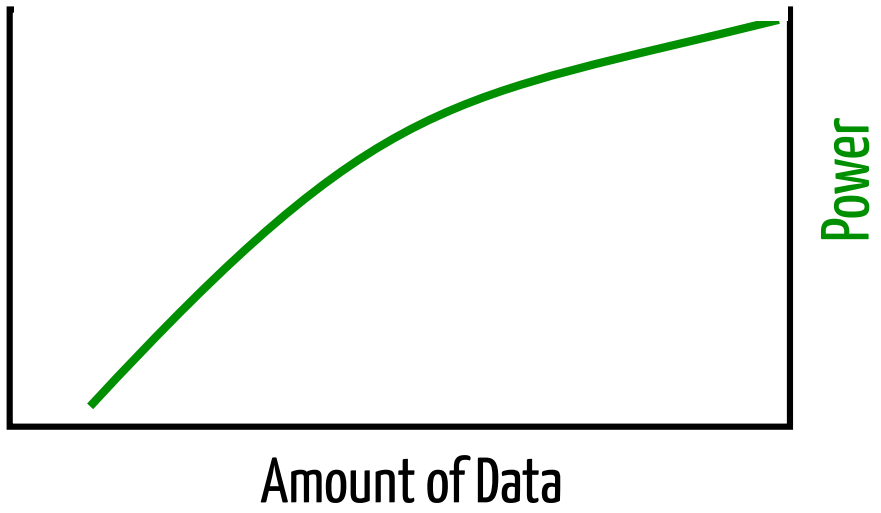


Our data centric view focuses on this

Why does this matter?

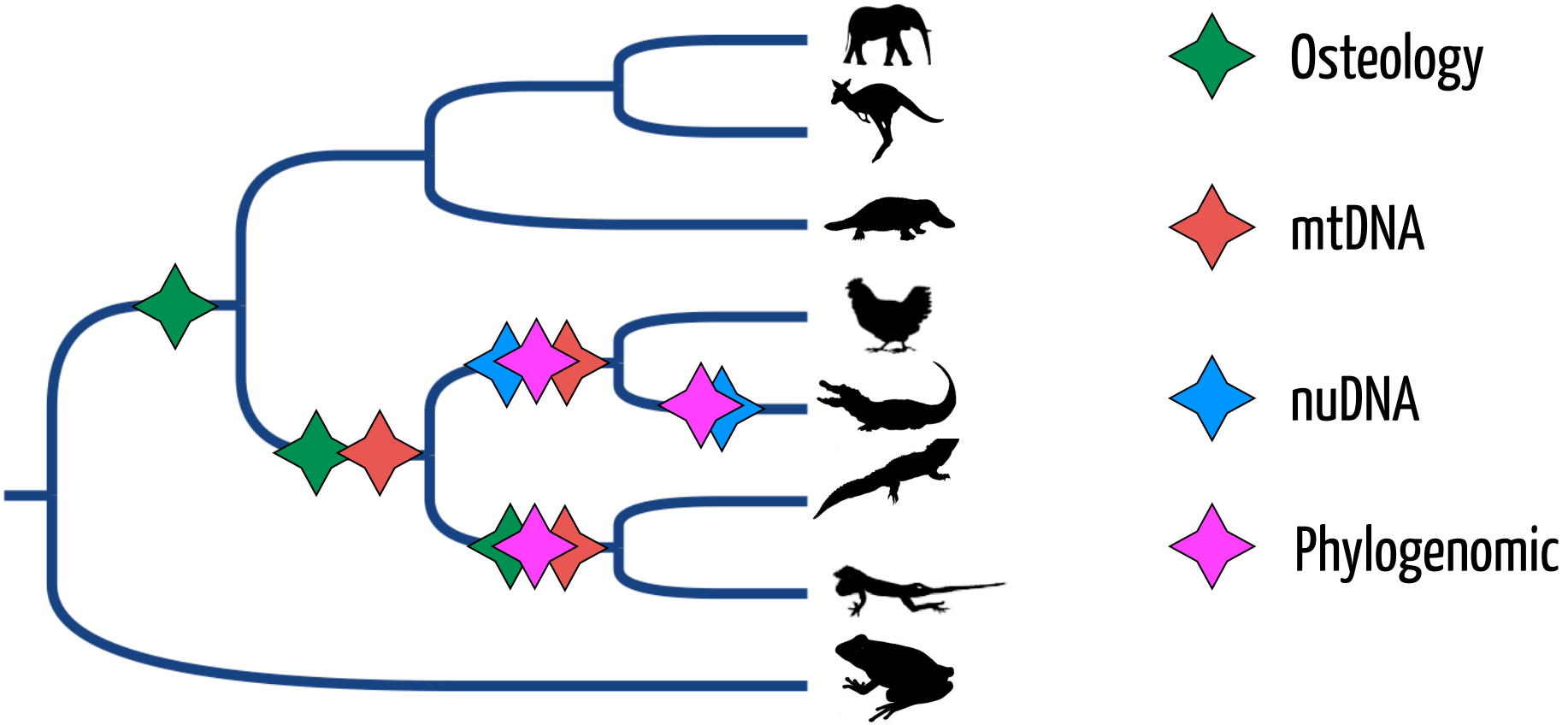
- The point.



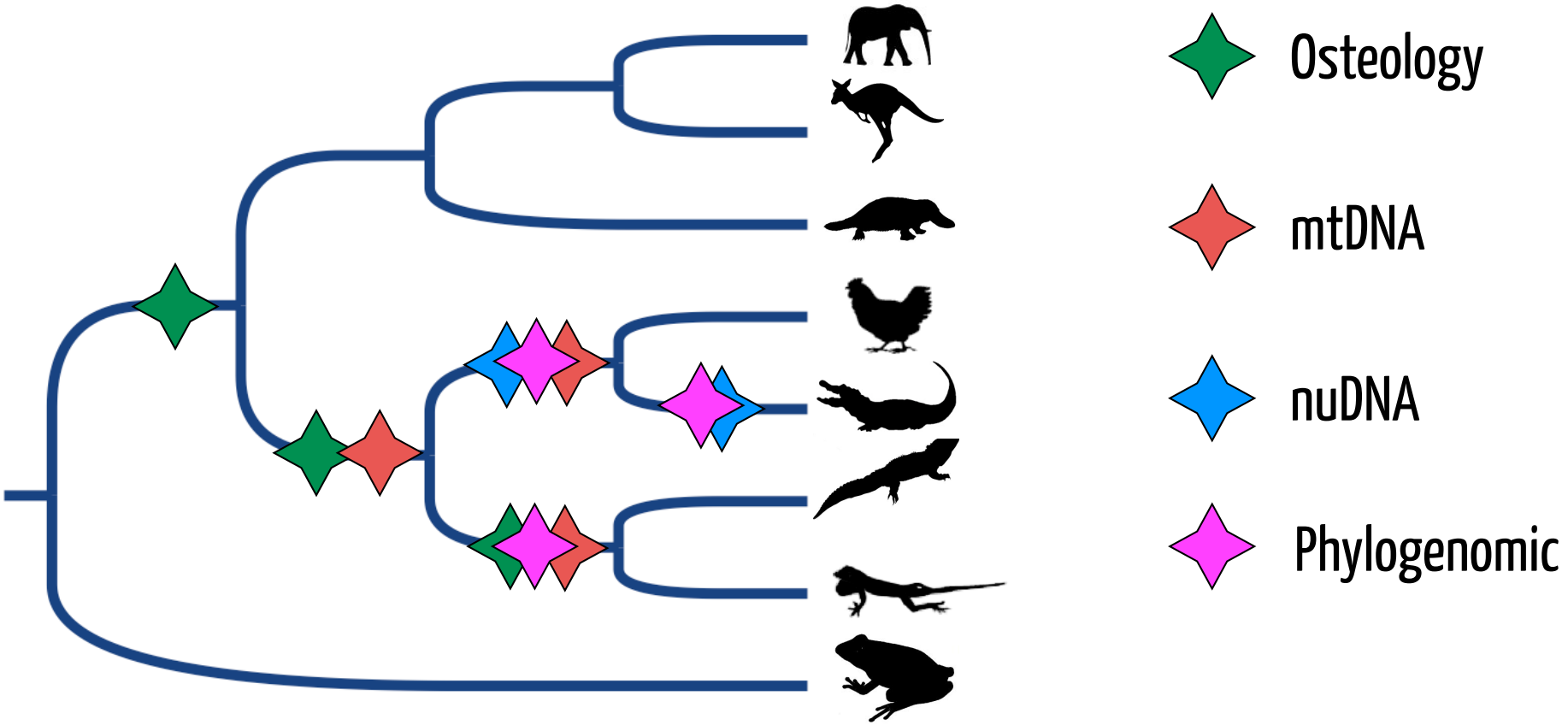


But this is our bigger problem

How do we know it's a bigger problem?

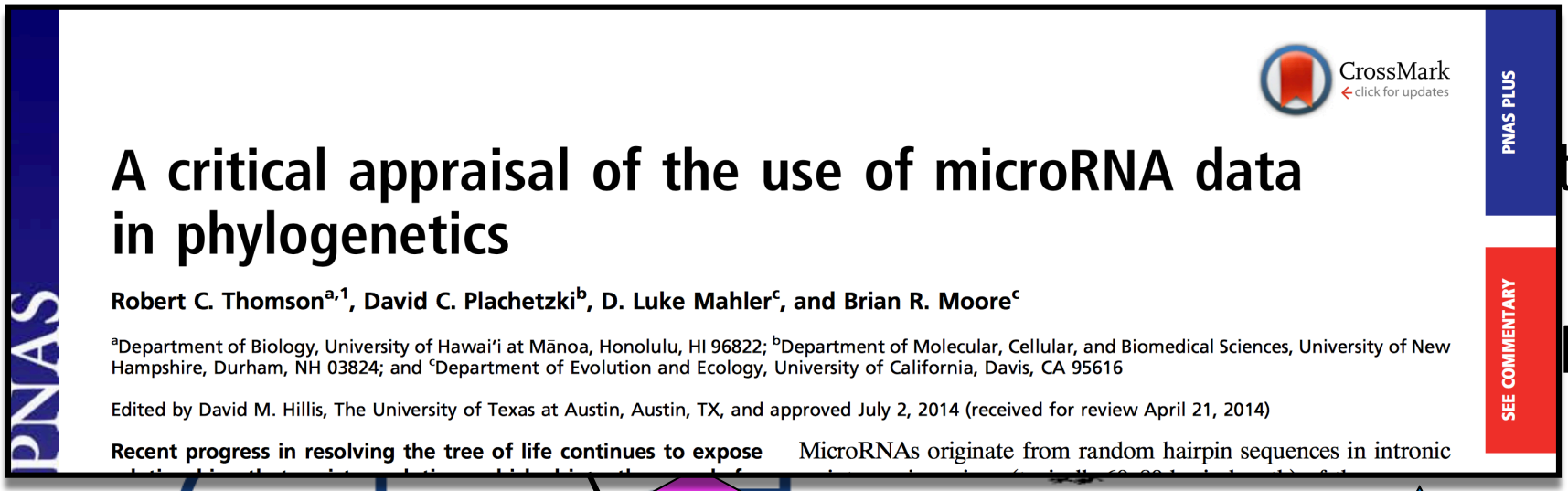


How do we know it's a bigger problem?



Where's the disagreement coming from?

How do we know it's a bigger problem?



PNAS

CrossMark click for updates

PNAS PLUS

SEE COMMENTARY

A critical appraisal of the use of microRNA data in phylogenetics

Robert C. Thomson[a,1], David C. Plachetzki[b], D. Luke Mahler[c], and Brian R. Moore[c]

[a]Department of Biology, University of Hawai'i at Mānoa, Honolulu, HI 96822; [b]Department of Molecular, Cellular, and Biomedical Sciences, University of New Hampshire, Durham, NH 03824; and [c]Department of Evolution and Ecology, University of California, Davis, CA 95616

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved July 2, 2014 (received for review April 21, 2014)

Recent progress in resolving the tree of life continues to expose MicroRNAs originate from random hairpin sequences in intronic

teology

DNA

nuDNA

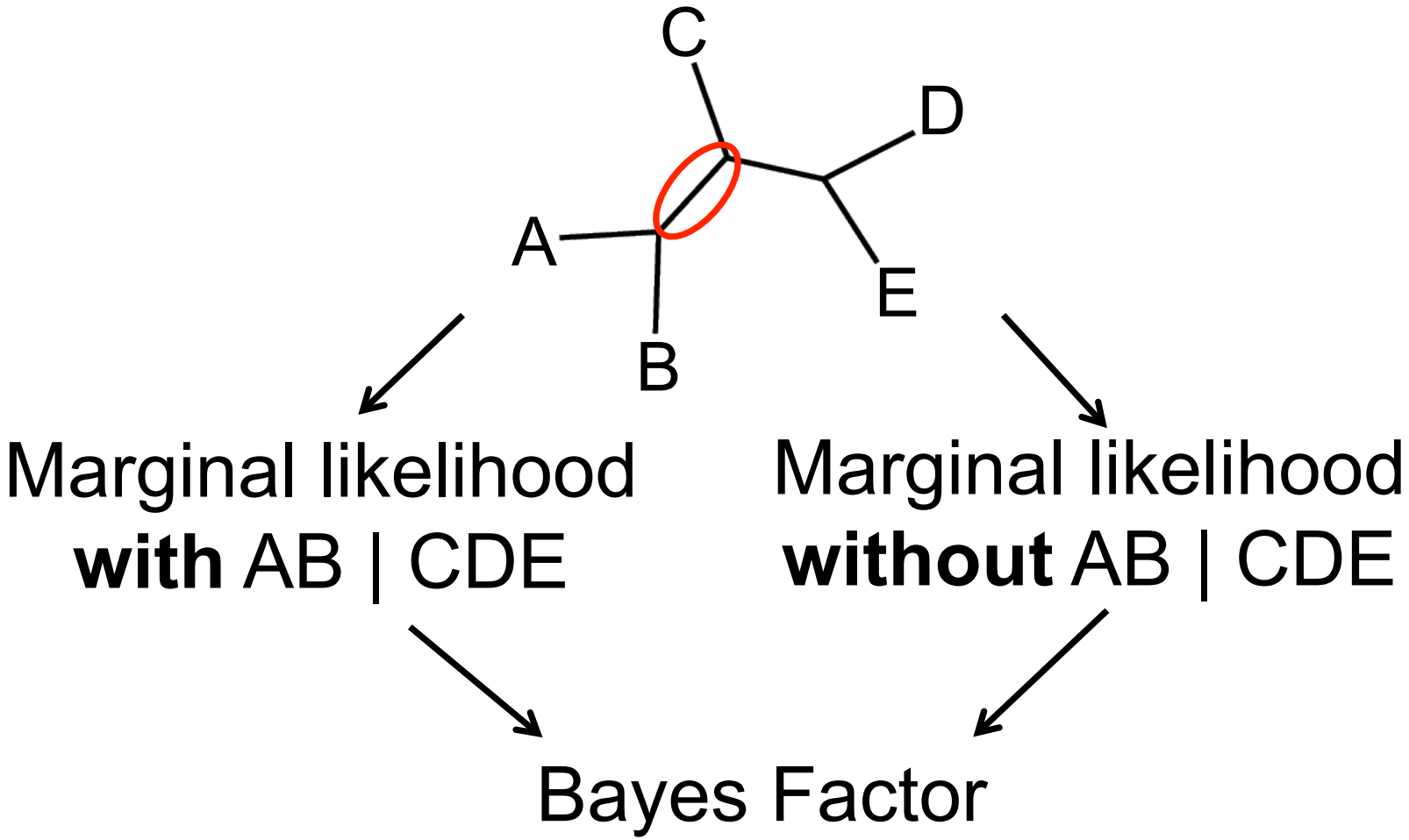
Phylogenomic

Where's the disagreement coming from?

‘Big data’ turtle studies

- Chiari et al. (2012)
  - 248 transcriptomic loci
  - 12 taxa
- Crawford et al. (2012)
  - 1,145 UCEs
  - 10 taxa
- Fong et al. (2012)
  - 75 Sanger-sequenced loci
  - 129 taxa
- Lu et al. (2013)
  - 1,638 transcriptomic and genomic loci
  - 11 taxa
- Shaffer et al. (2013)
  - 1,955 genomic loci
  - 8 taxa
- Wang et al. (2013)
  - 1,113 genomic loci
  - 12 taxa

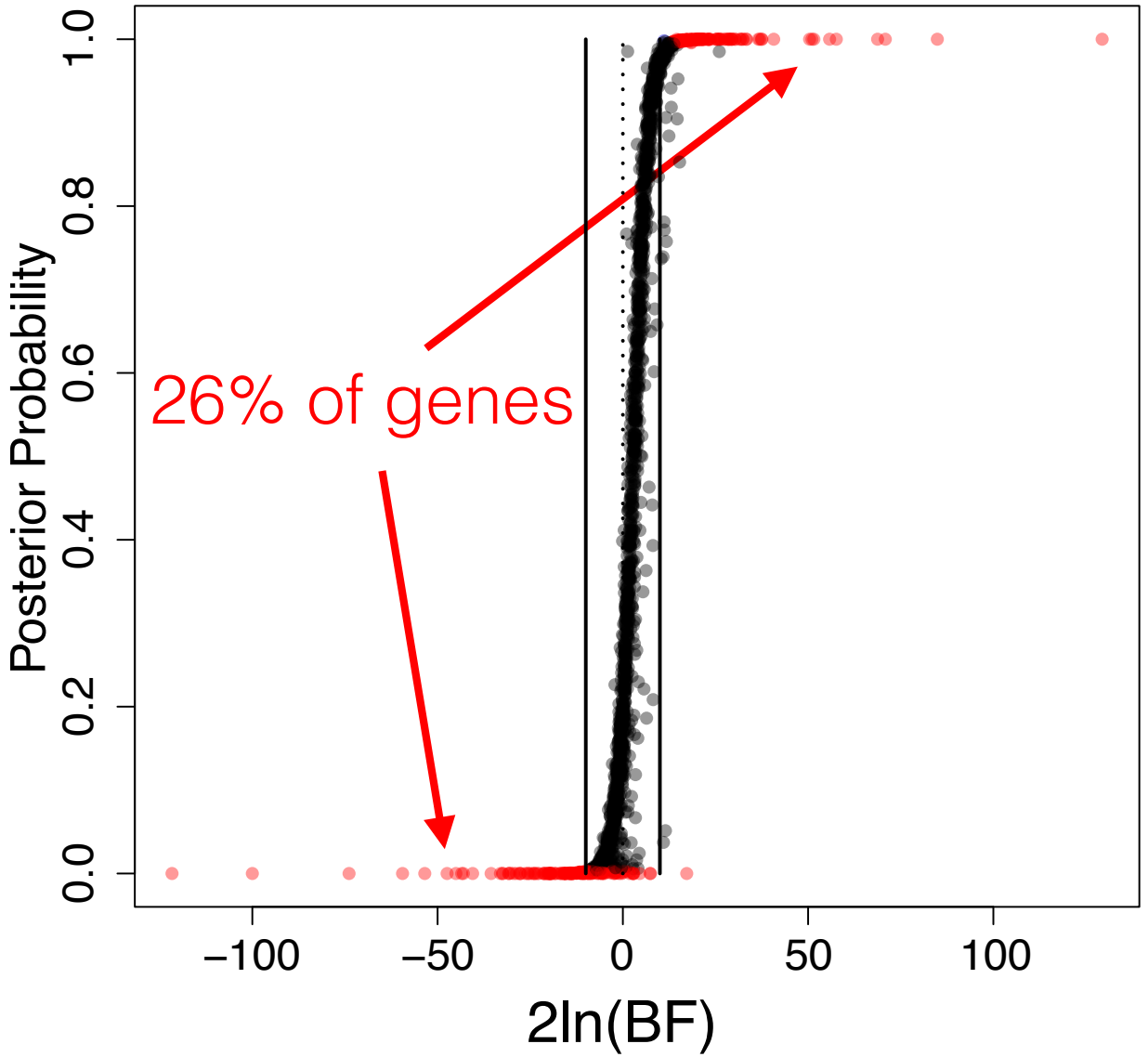
Bipartition Bayes Factors



A Note on Extreme Probabilities

Archosaur + Turtle Monophyly

Shaffer et al.

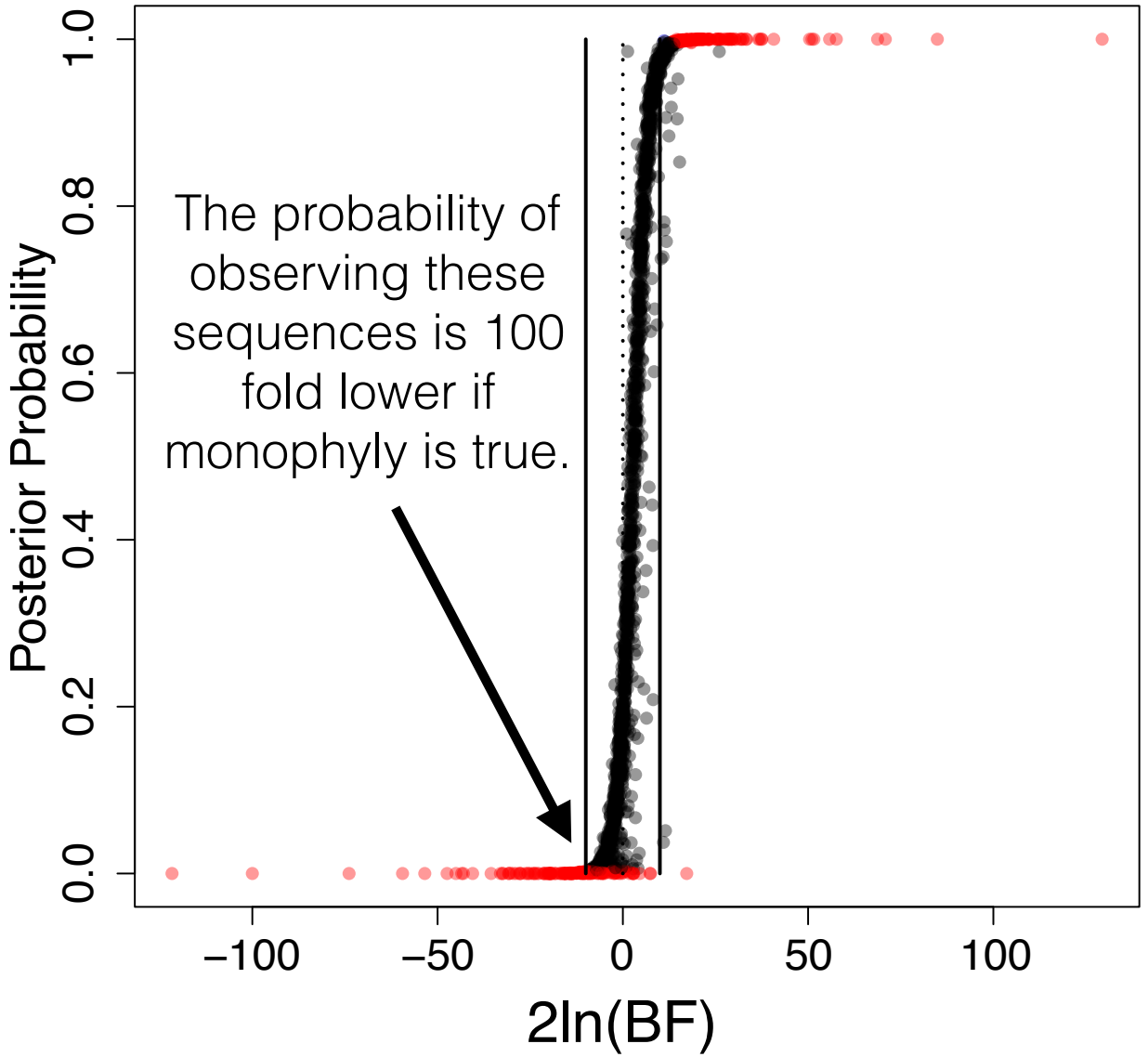


Brown and Thomson 2017

A Note on Extreme Probabilities

Archosaur + Turtle Monophyly

Shaffer et al.

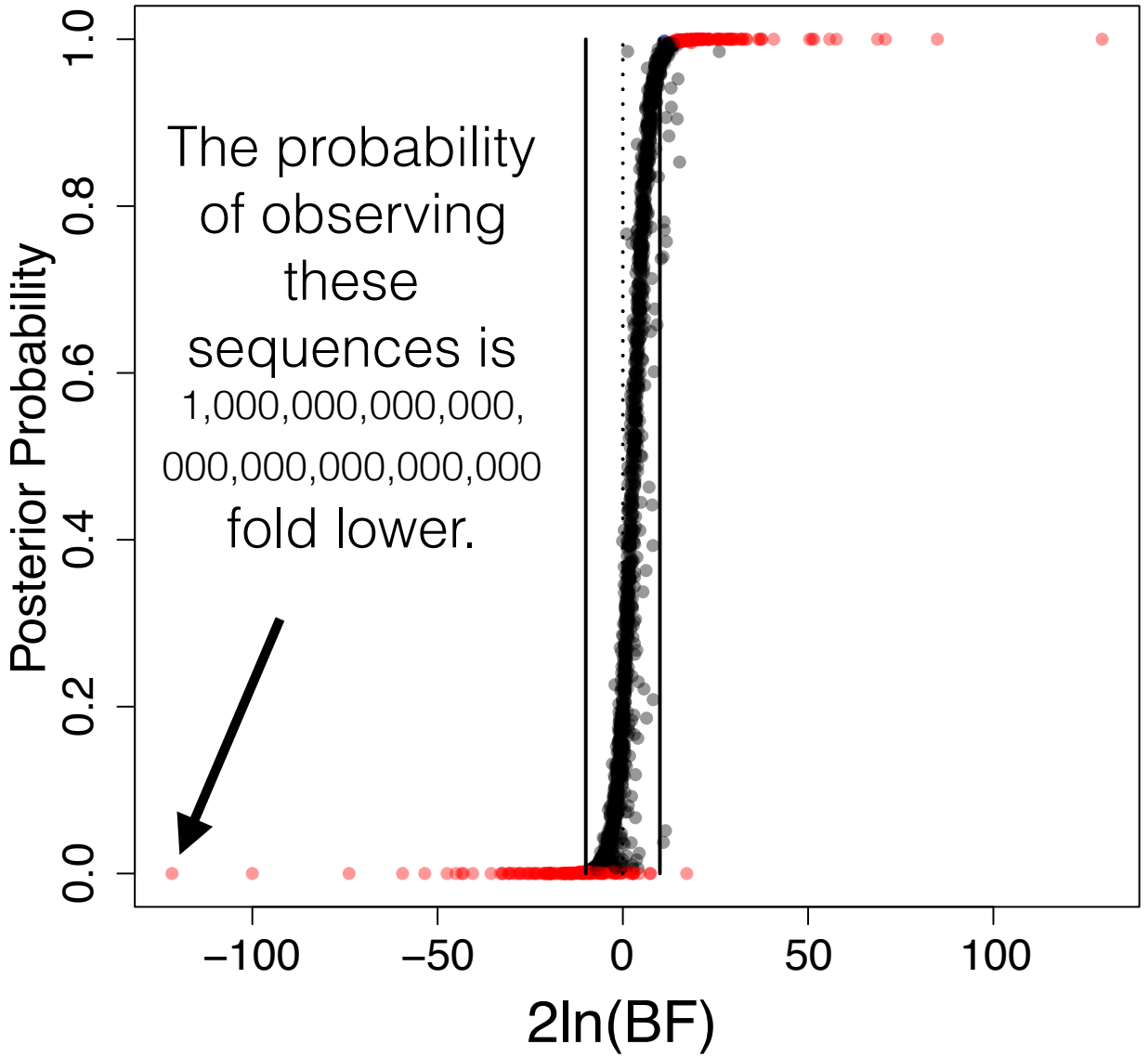


Brown and Thomson 2017

A Note on Extreme Probabilities

Archosaur + Turtle Monophyly

Shaffer et al.



Brown and Thomson 2017

A Note on Extreme Probabilities

1/1,000,000,000,000,000,000,000,000,000

That's 27 zeroes!

If you played a lottery every minute with that chance of winning, you still probably wouldn't win, unless you played for…

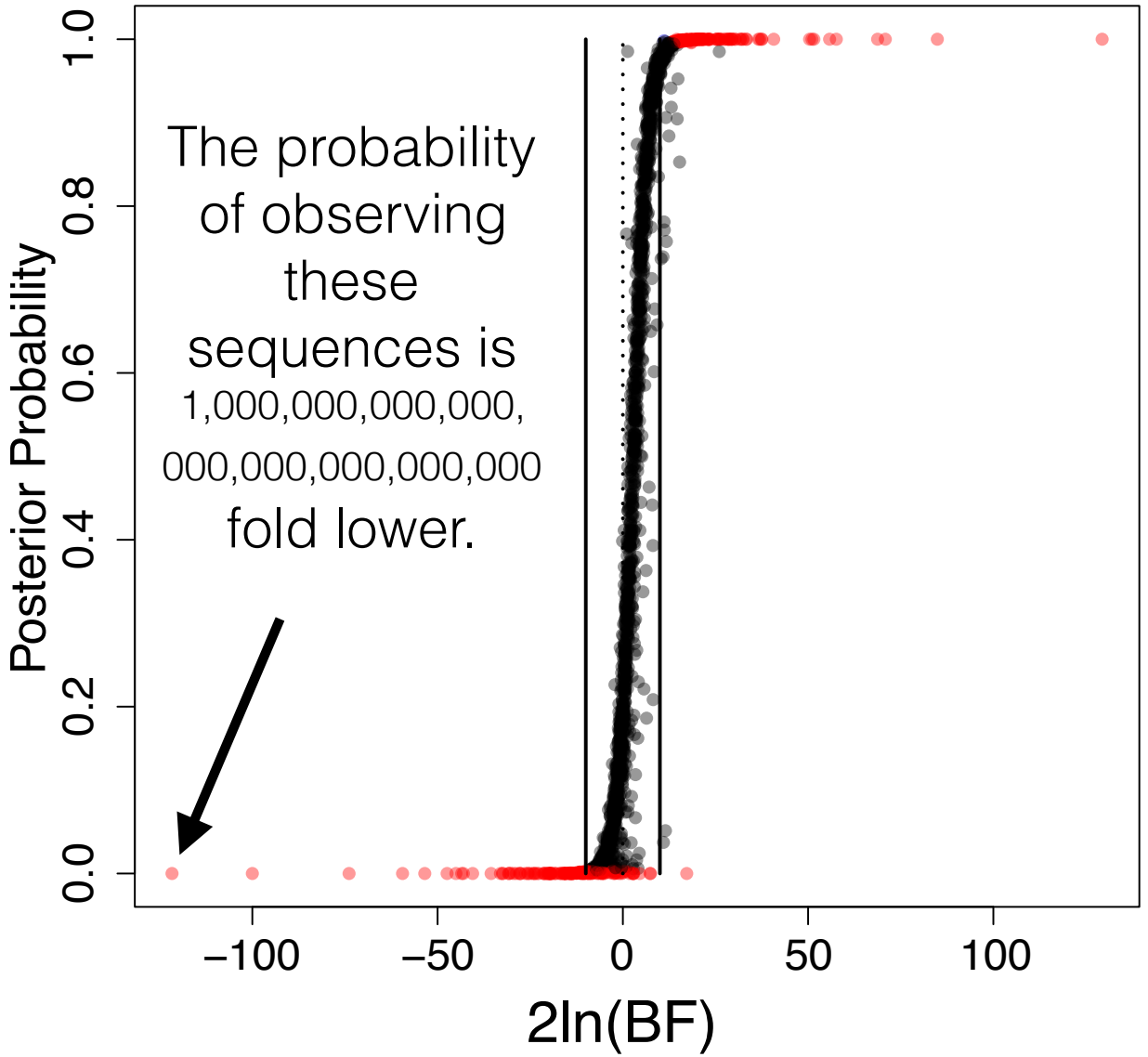
the age of the universe*190,258,751,903

Jeremy's Analogy

A Note on Extreme Probabilities

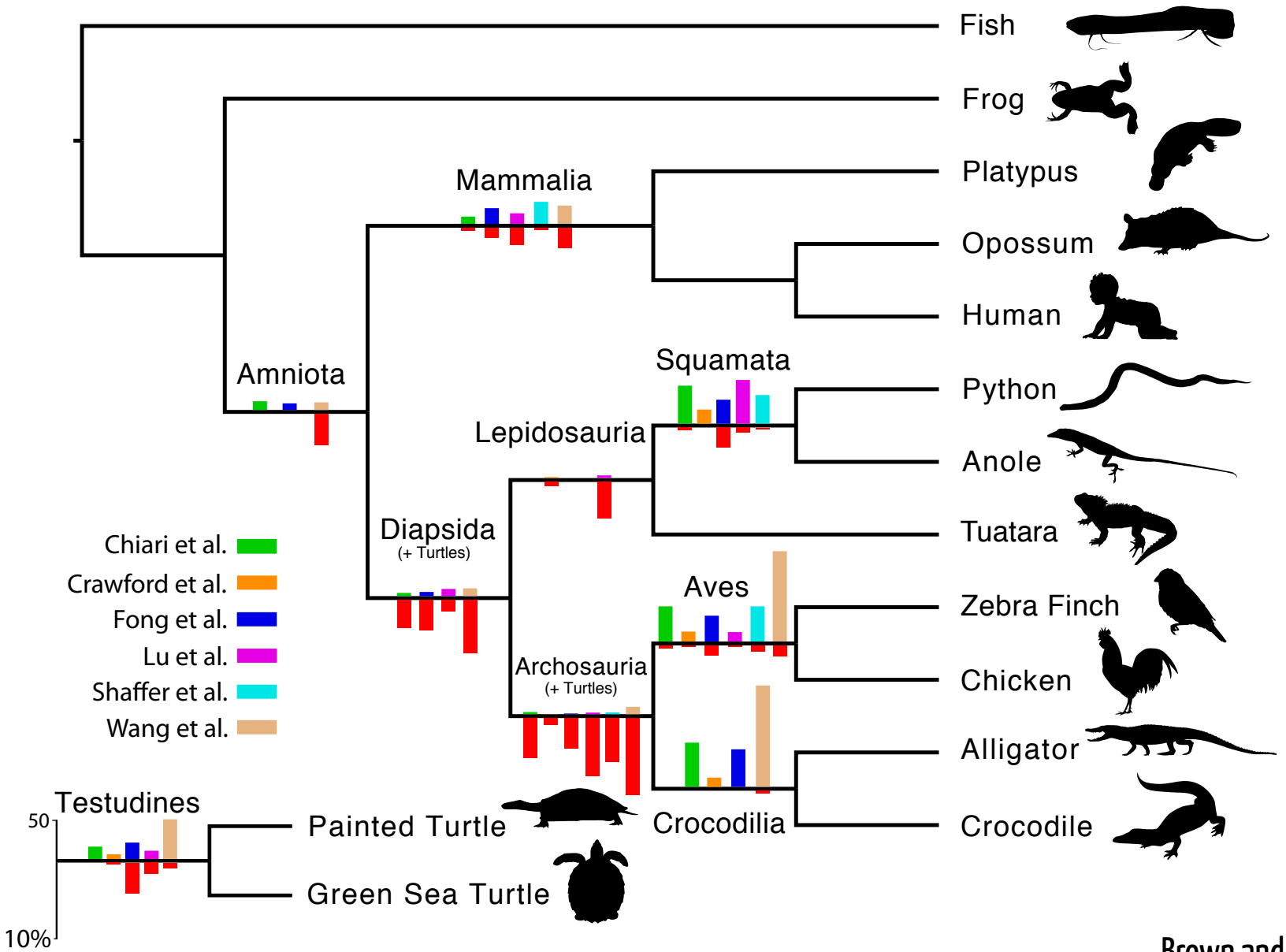
Archosaur + Turtle Monophyly

Shaffer et al.



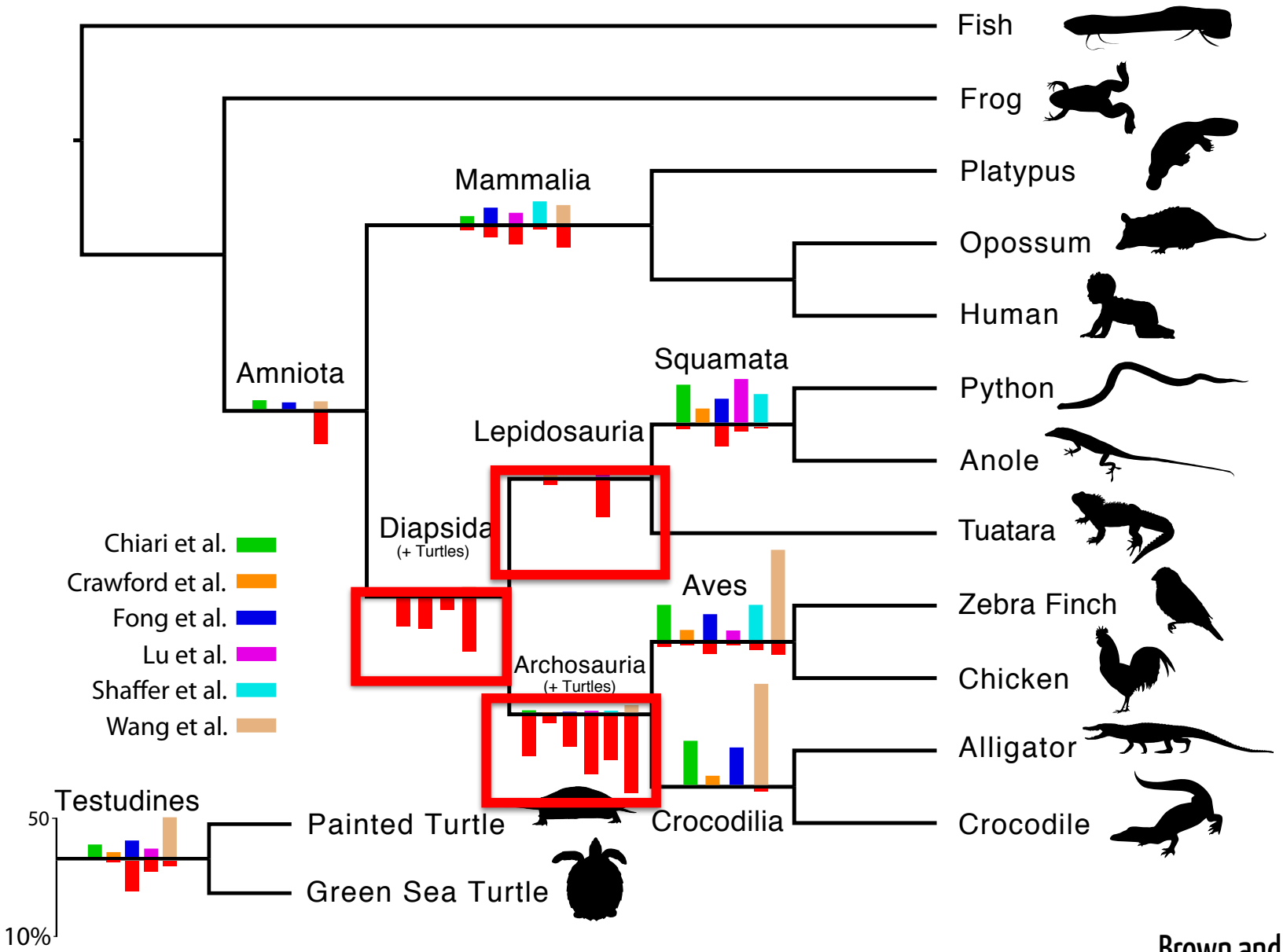
Brown and Thomson 2017

Support varies across branches of the tree



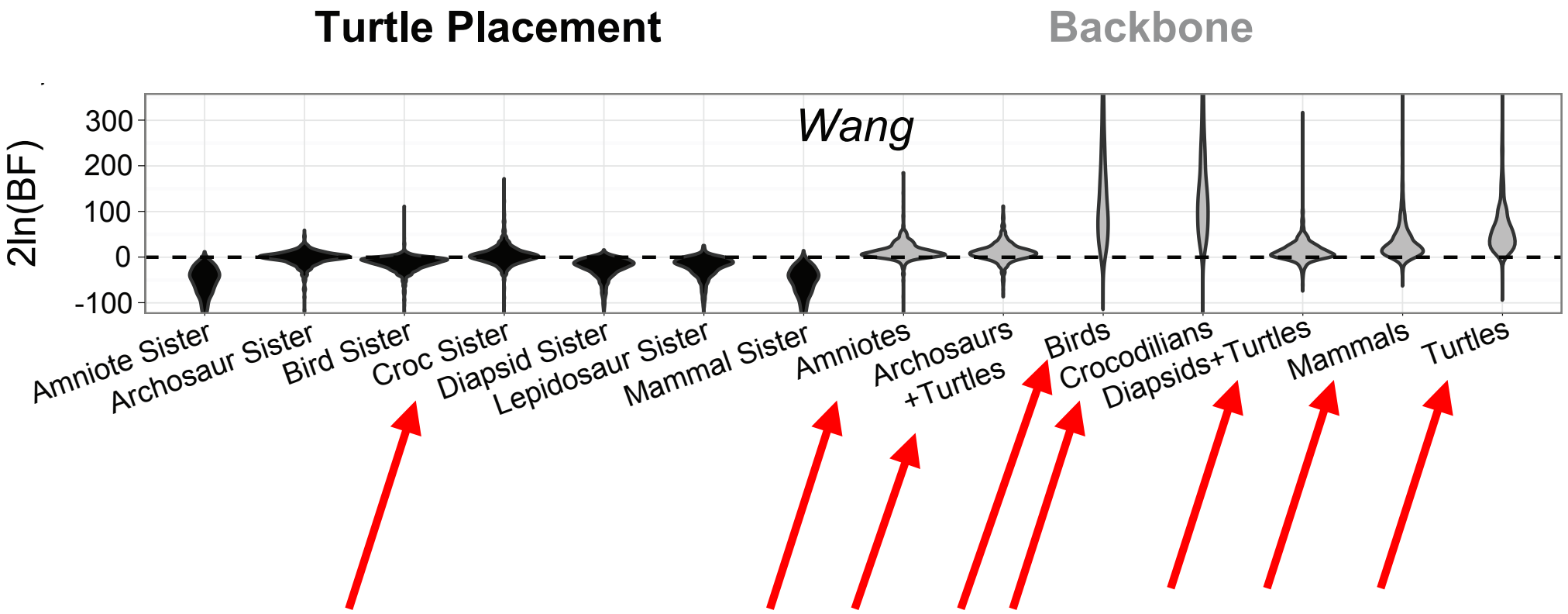
Brown and Thomson 2017

Support varies across branches of the tree



Brown and Thomson 2017

Support varies across branches of the tree

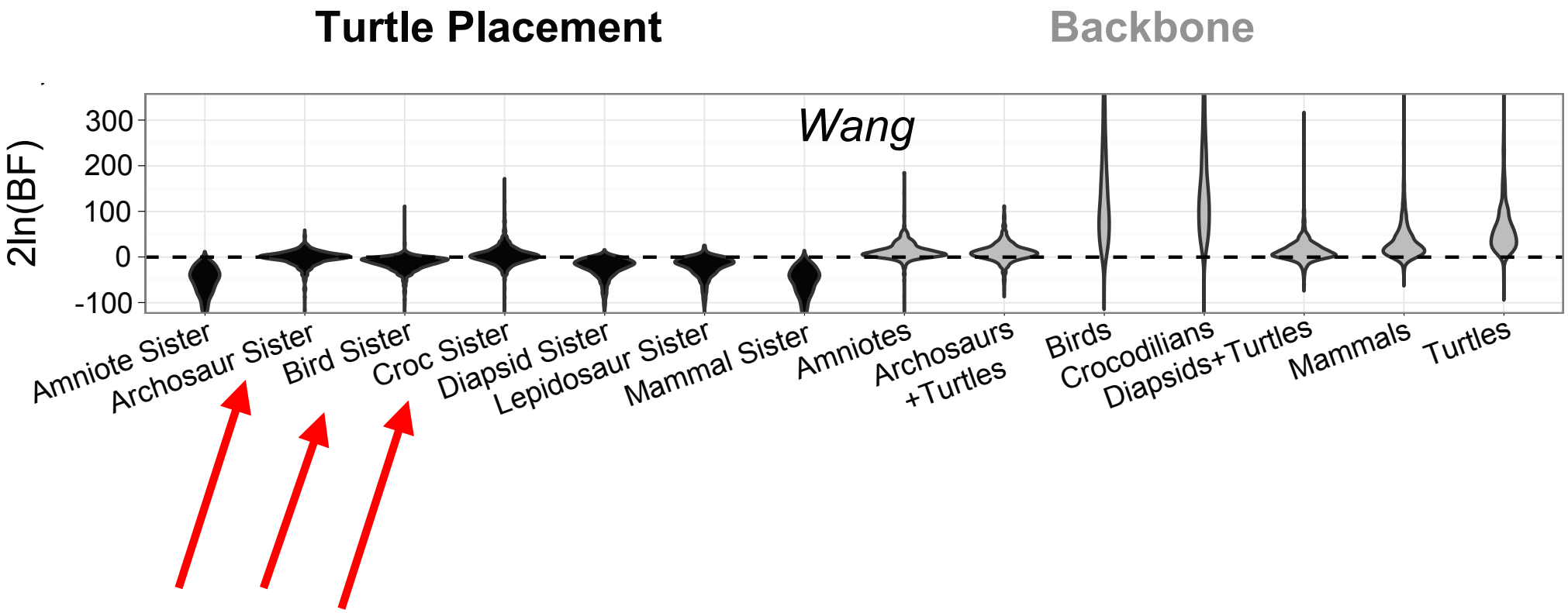


All of these **strongly supported** by this dataset.

PP = 1.0

Brown and Thomson 2017

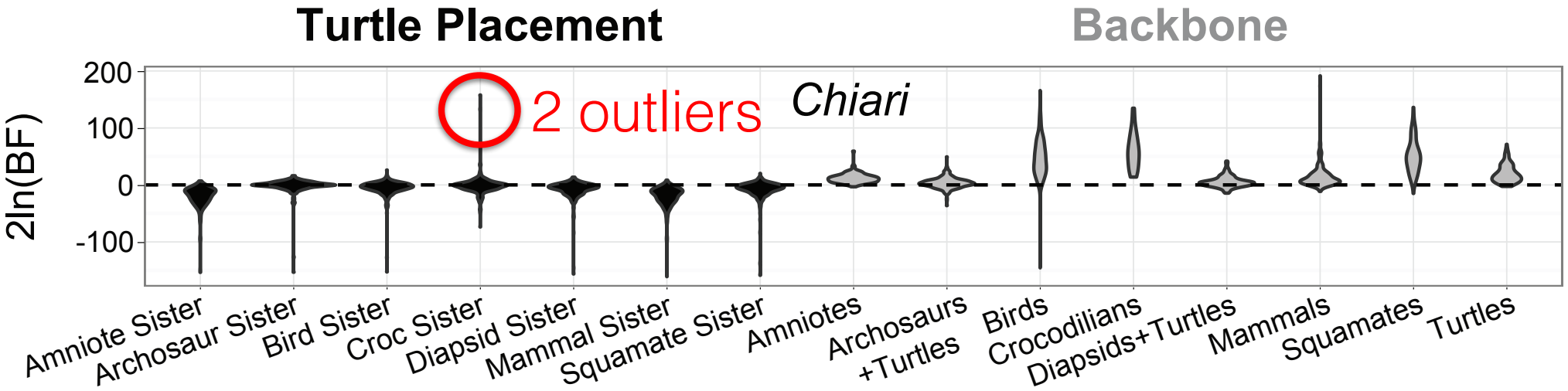
Support varies across branches of the tree



Equivocation about turtle placement **across genes**

Brown and Thomson 2017

Support varies across branches of the tree

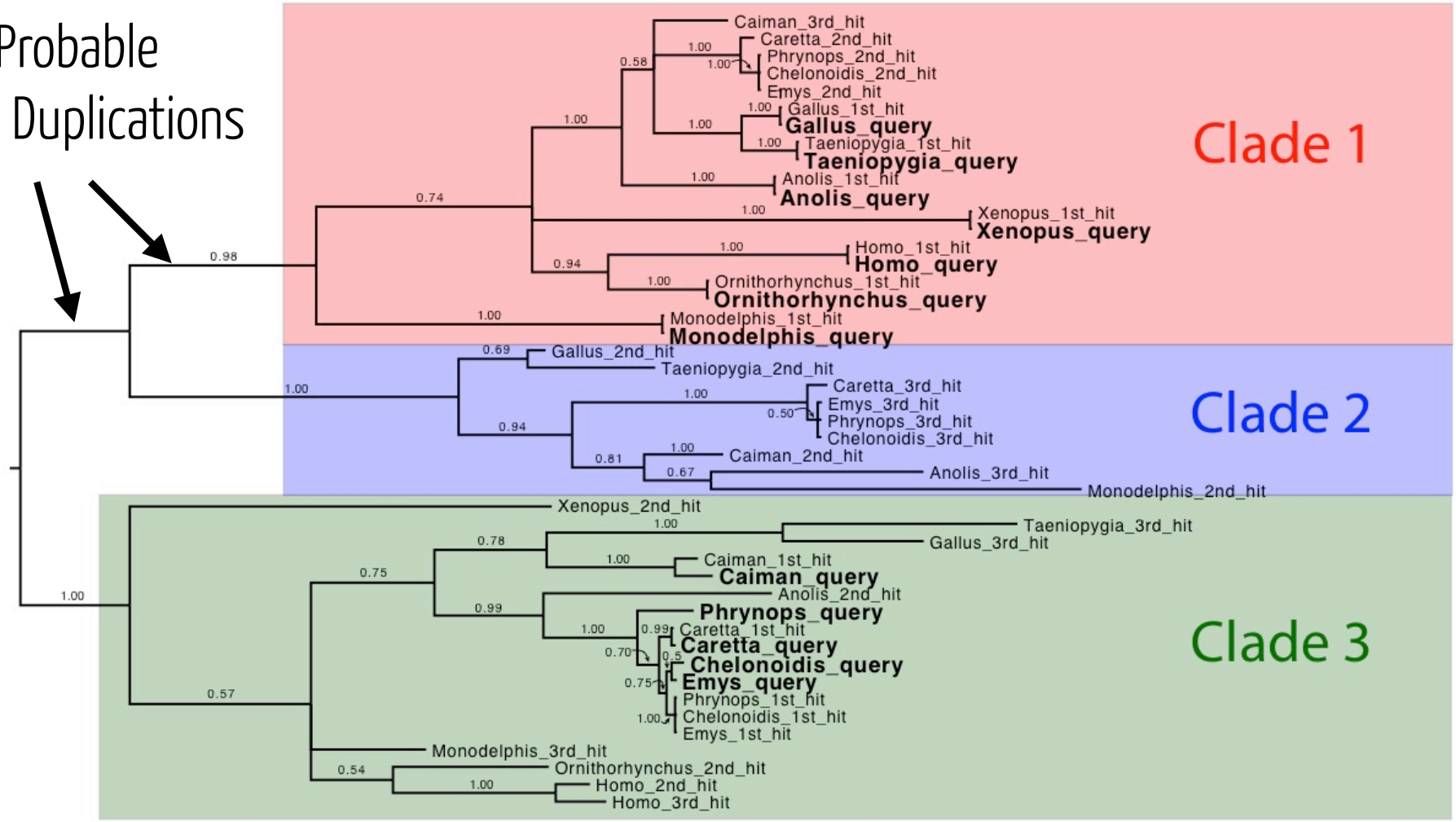


This dataset supports turtles as sister to crocodilians.
But what's up with these outliers?
How influential are they?

Brown and Thomson 2017

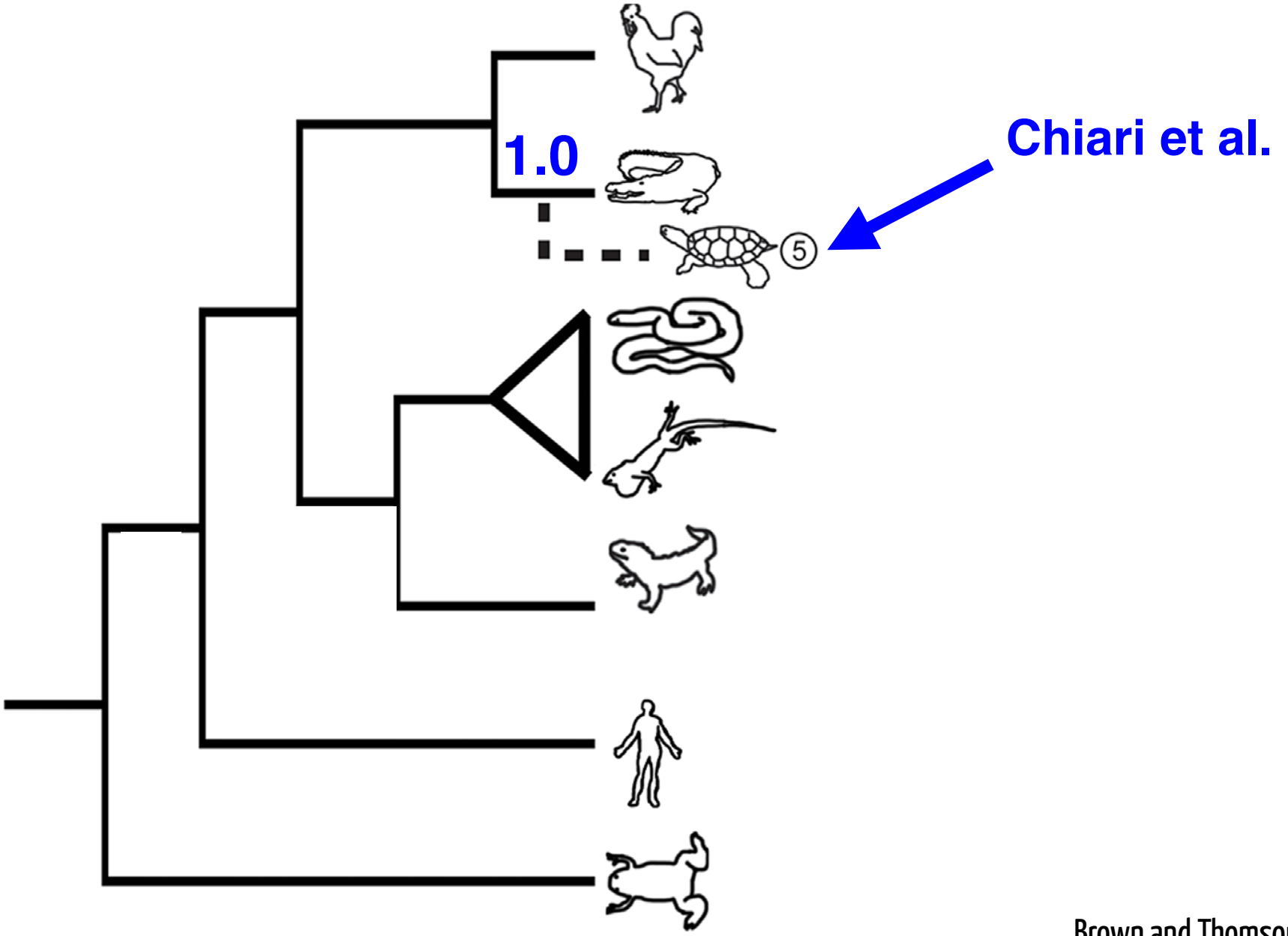
Both look like paralogs

Probable
Gene Duplications



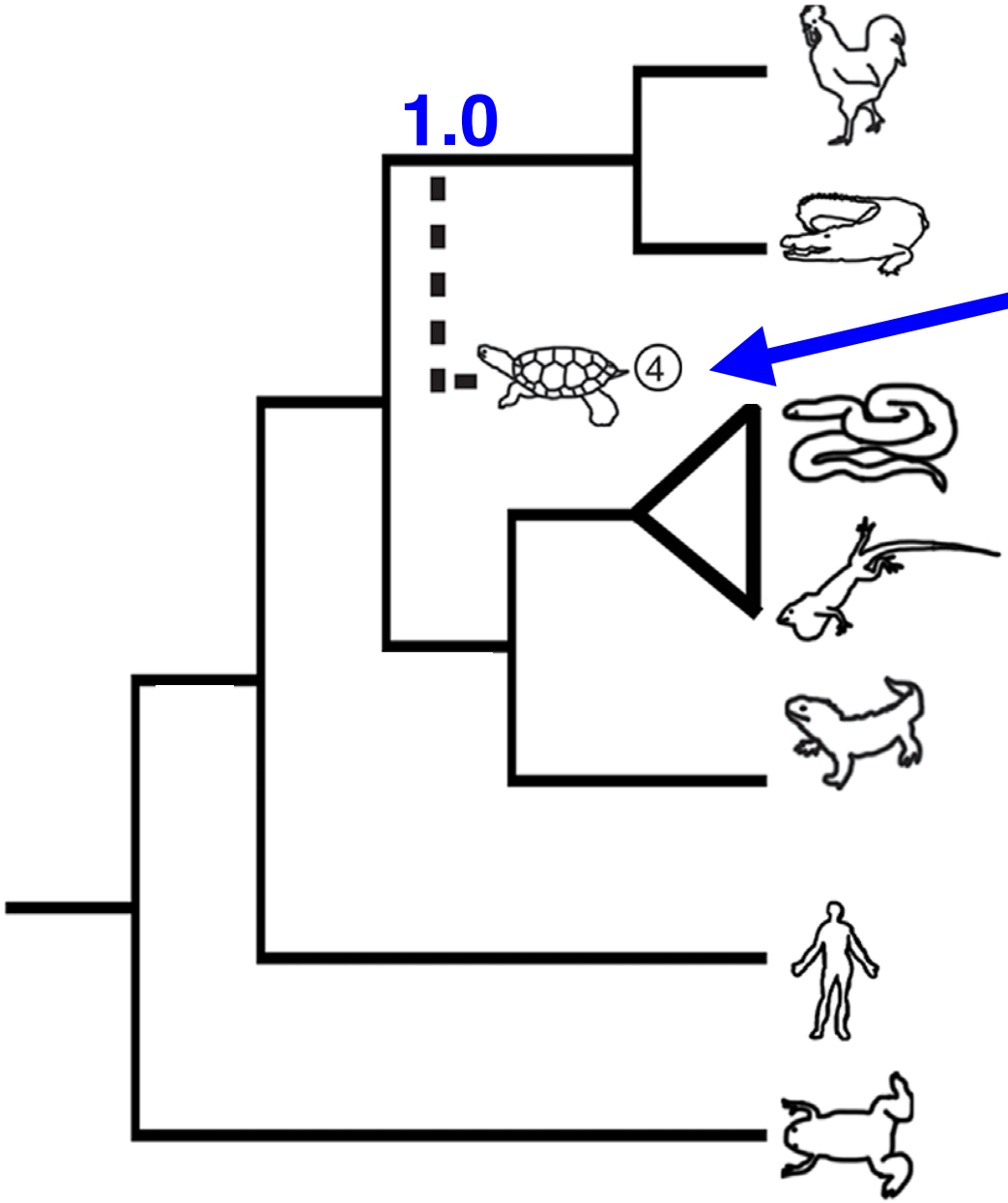
Brown and Thomson 2017

Strong influence



Brown and Thomson 2017

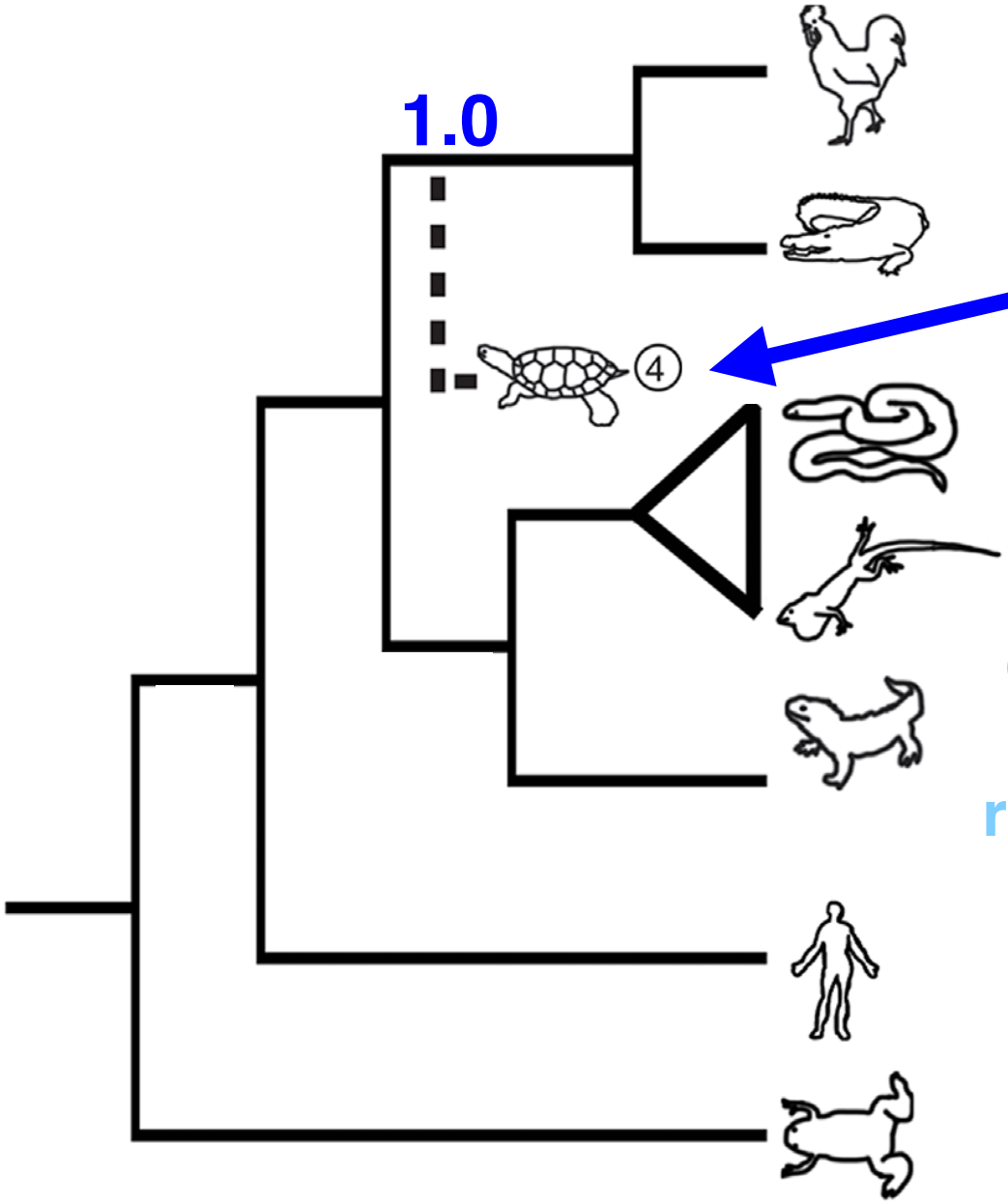
Strong influence



Chiari et al.
without 2 outliers
246/248 genes remain

Brown and Thomson 2017

Strong influence



Chiari et al.
without 2 outliers
246/248 genes remain

An **identical switch** occurs for the dataset of Wang et al. when **removing ~1% of genes**.

Brown and Thomson 2017

Take homes

- More data does not necessarily lead to more accuracy, or to consensus
- A lot of phylogenomic **progress** is actually about figuring out how to **model data well**, not collect more data per se

Some Possible Ways Forward

- **Embrace** the computational **challenge**

Embrace the computation

- Analyses need not finish quickly
- Advances in computation help a lot here
  - parallel architectures and code
  - fast computation libraries
  - availability of compute resources
  - new methods on the horizon (HMC, IDR)

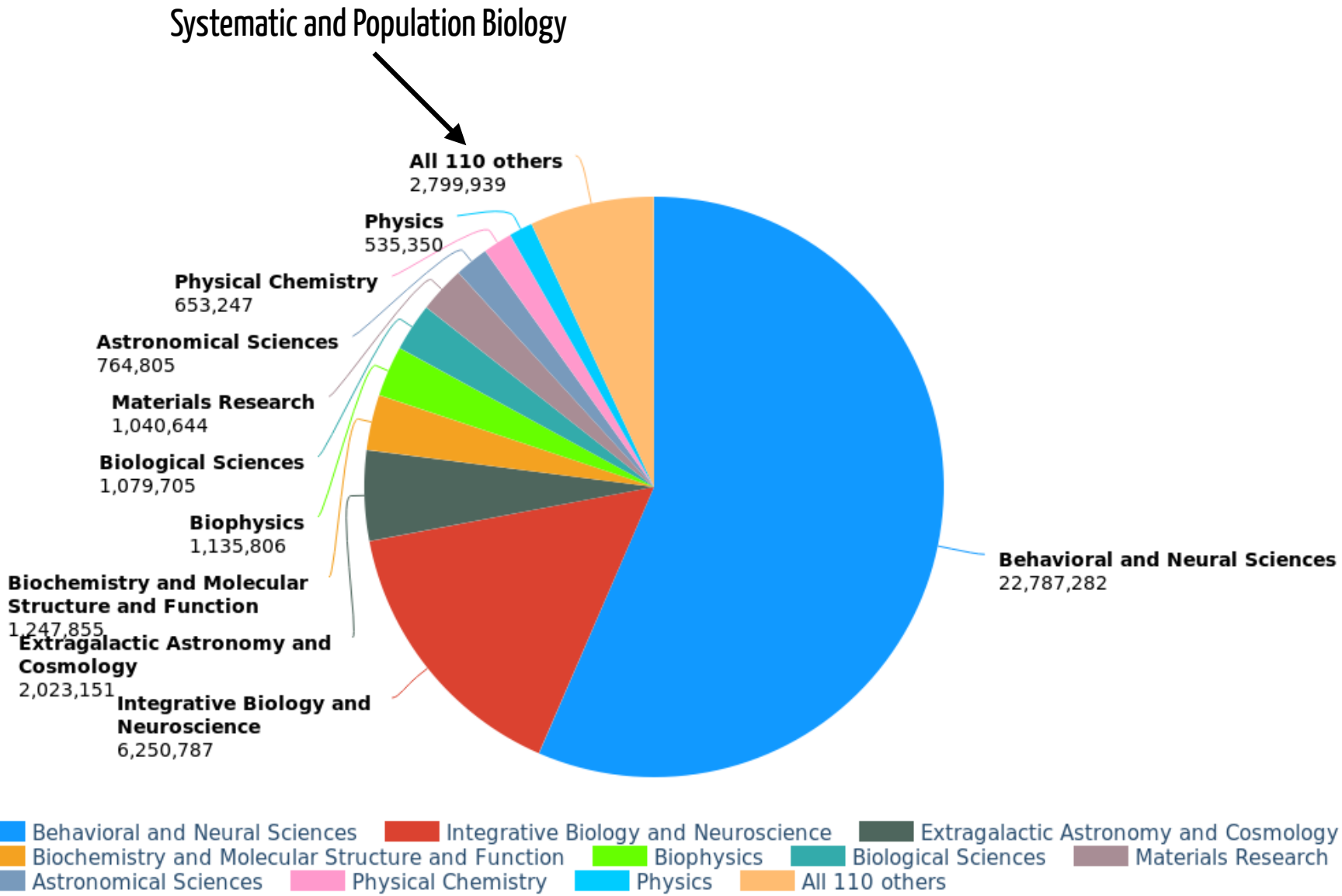
Embrace the computation

Embrace the computation

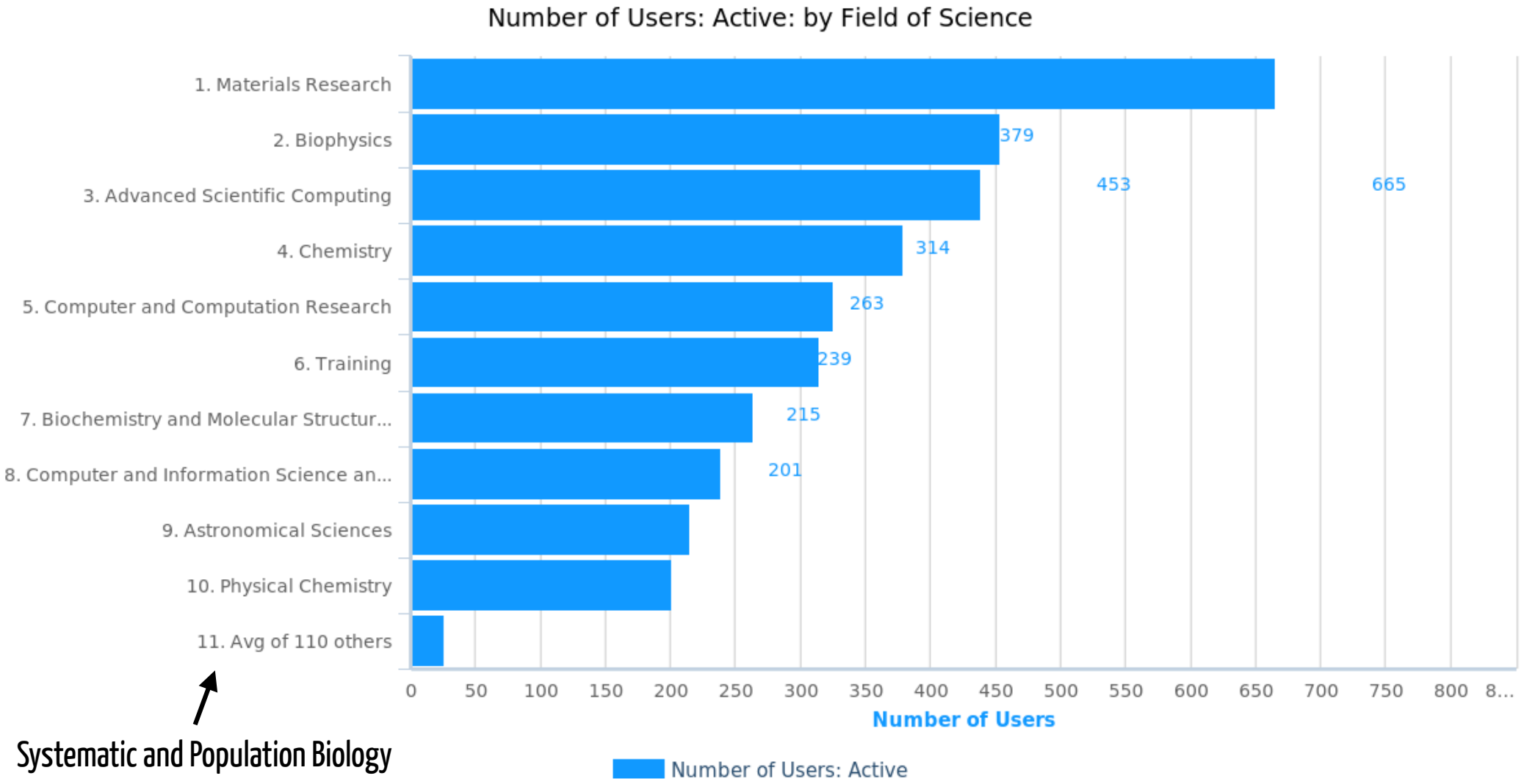
Compute Resources

Name	Status	CPUs	Peak TFlops	Utilization	Running Jobs	Queued Jobs	Other Jobs
Stampede User Guide	Healthy	102400	9600.0	61%	562	2310	161
Comet User Guide	Healthy	47616	2000.0	87%	1487	360	379
XStream User Guide	Healthy	1300	1001.7	82%	303	85	415

XSEDE jobs by field for 2016



XSEDE users by field for 2016



Systematic and Population Biology

New Tools on the Horizon

- More complex models
- faster estimation of marginal likelihoods. e.g., Inflated Density Ratio
- More efficient sampling. e.g., Hamiltonian Monte Carlo

A Comparison of two Markov Chain Monte Carlo samplers

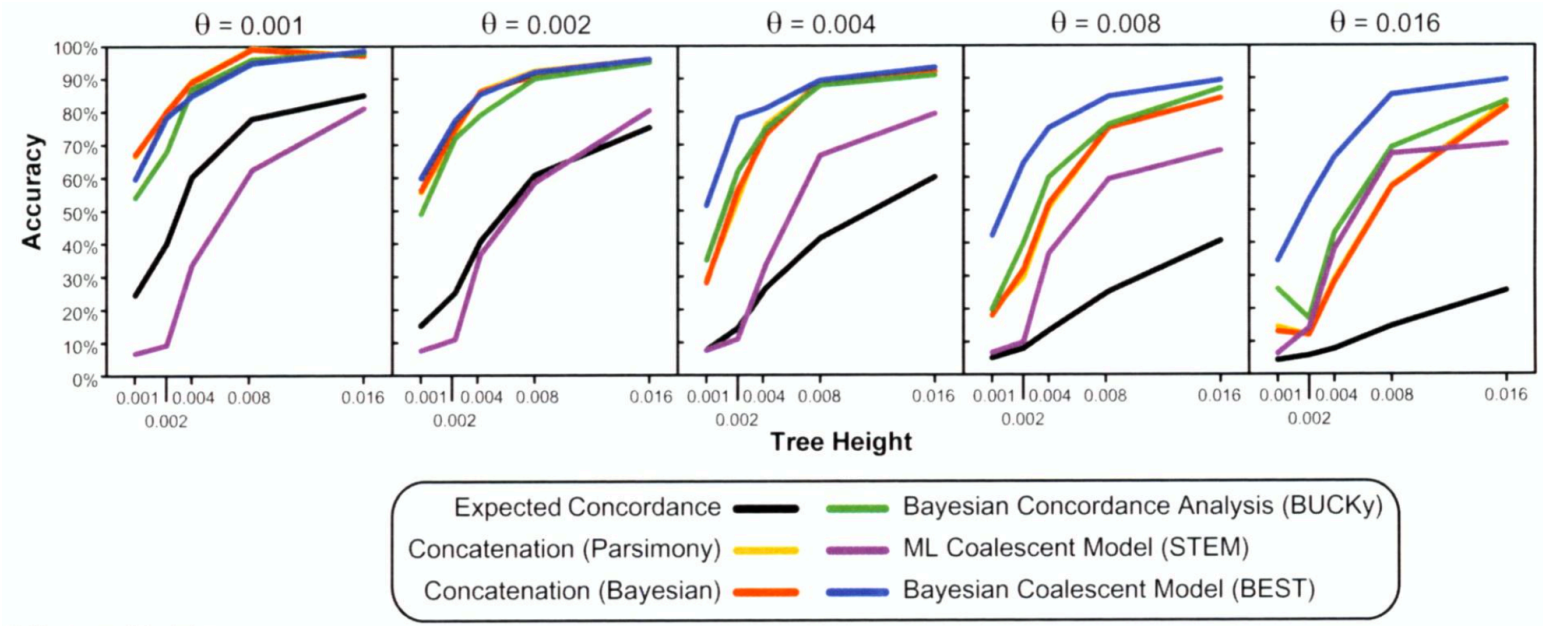
Tamara Broderick (UC Berkeley), David Duvenaud (Cambridge)

Some Possible Ways Forward

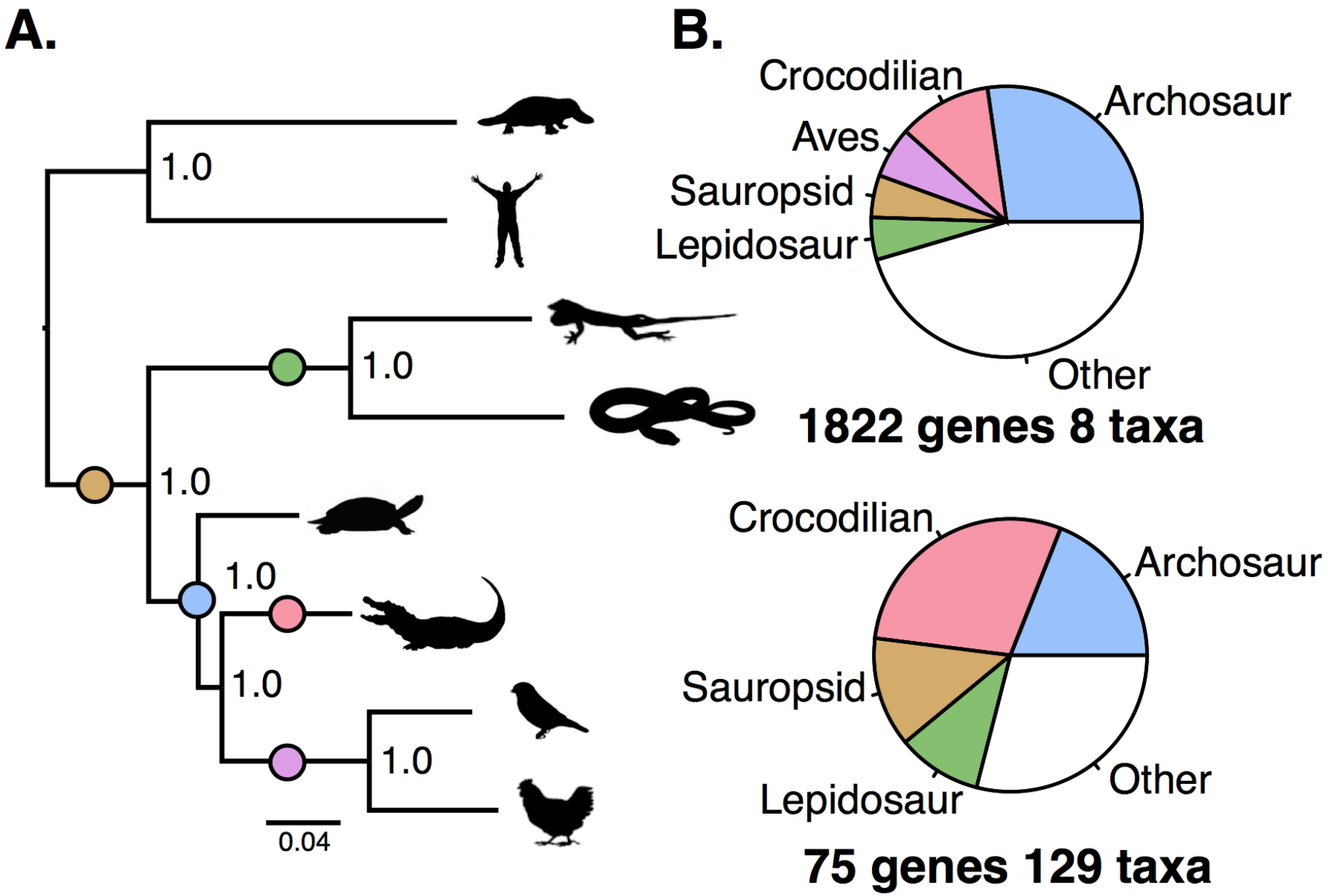
- **Embrace** the computational **challenge**
- **Get very picky** about our data. Careful and detailed data exploration is your friend.

Some Possible Ways Forward

- **Embrace** the computational **challenge**
- **Get very picky** about our data. Careful and detailed data exploration is your friend.
- **Carefully consider tradeoffs** between speed and approximation



Leaché and Rannala 2010



Brown and Thomson unpublished figure