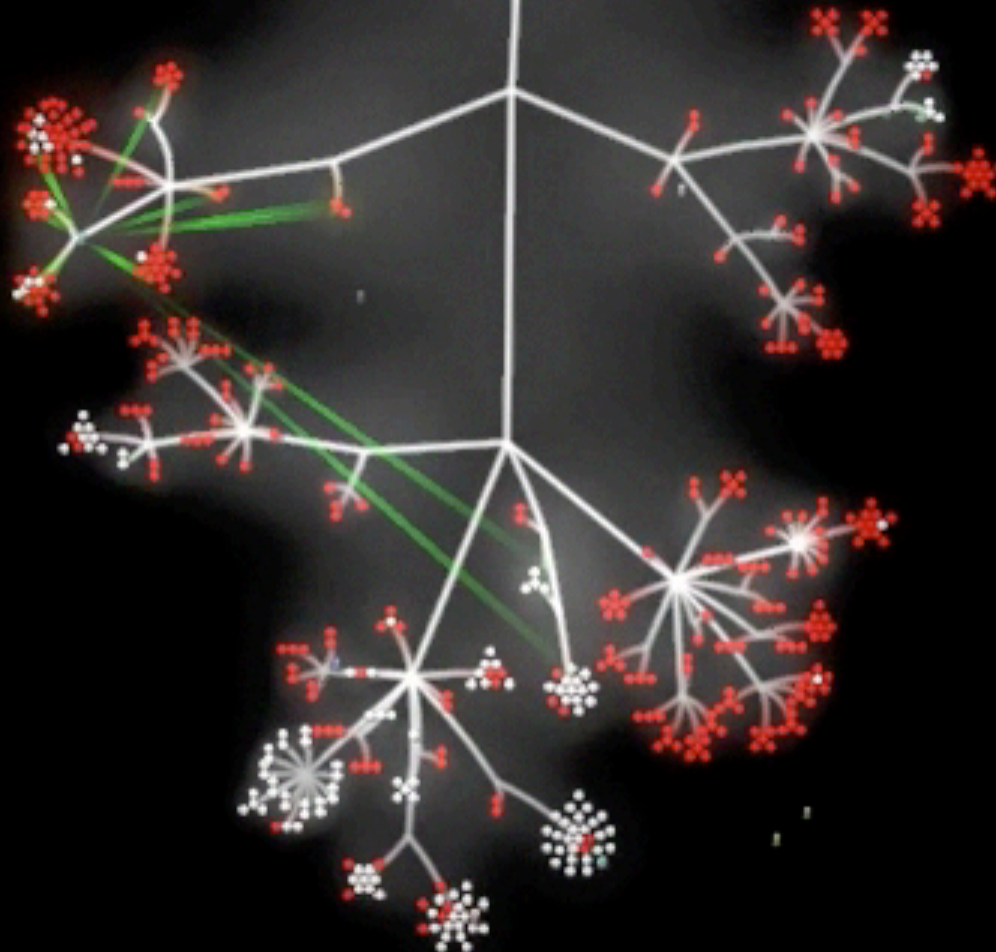


January 2003

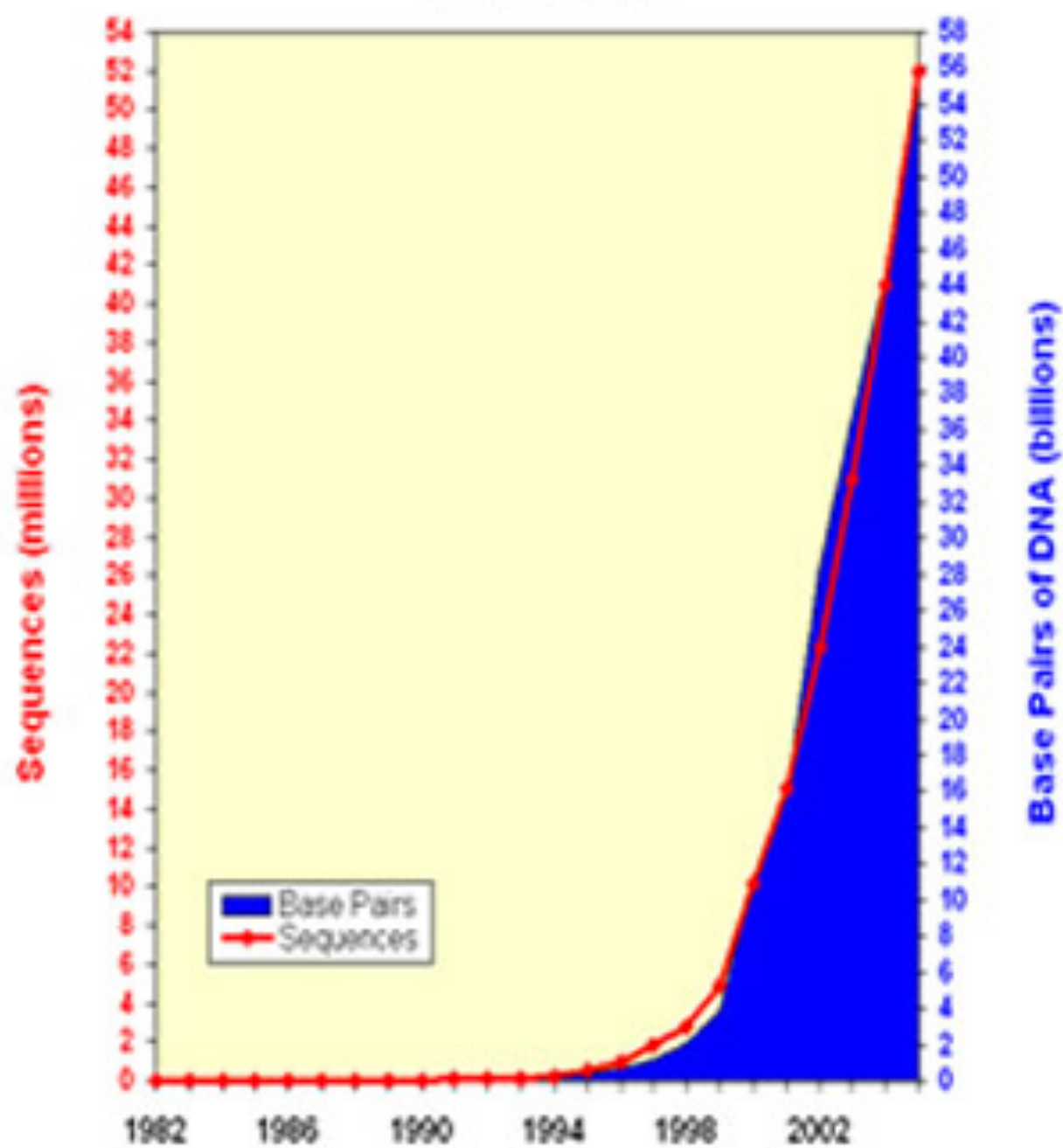
Big Data in Phylogenetics

Dealing with the deluge

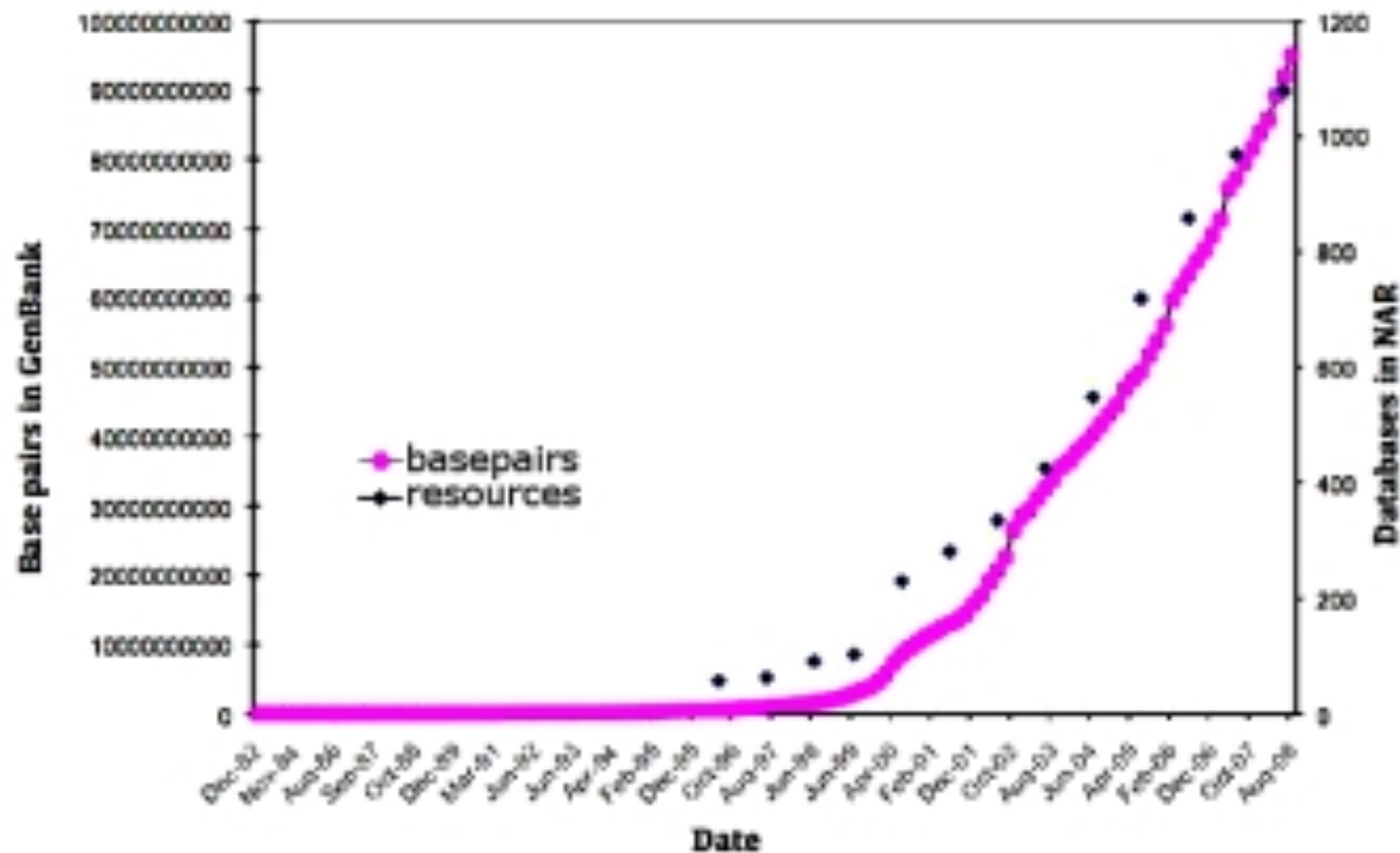


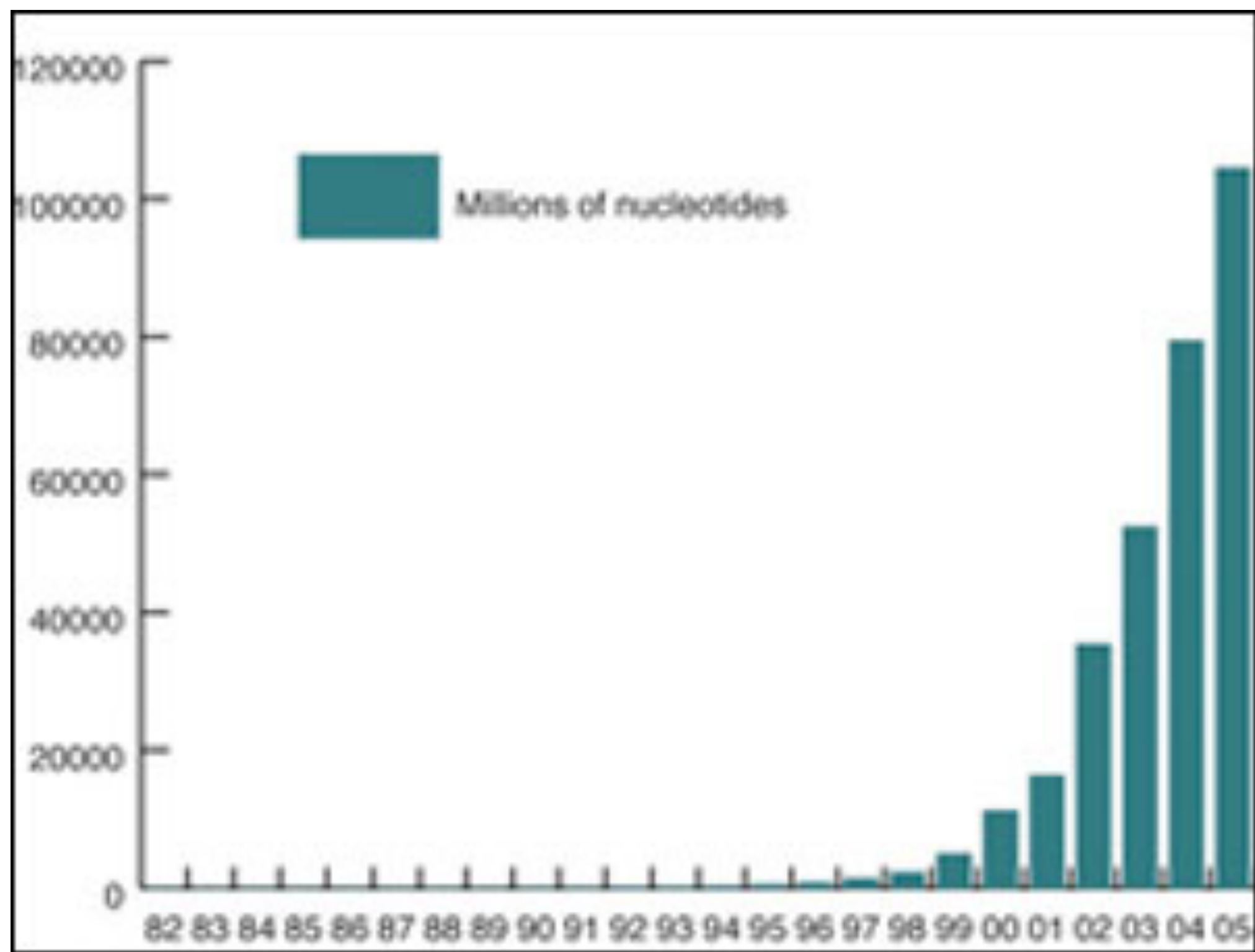
Growth of GenBank

(1982 - 2005)

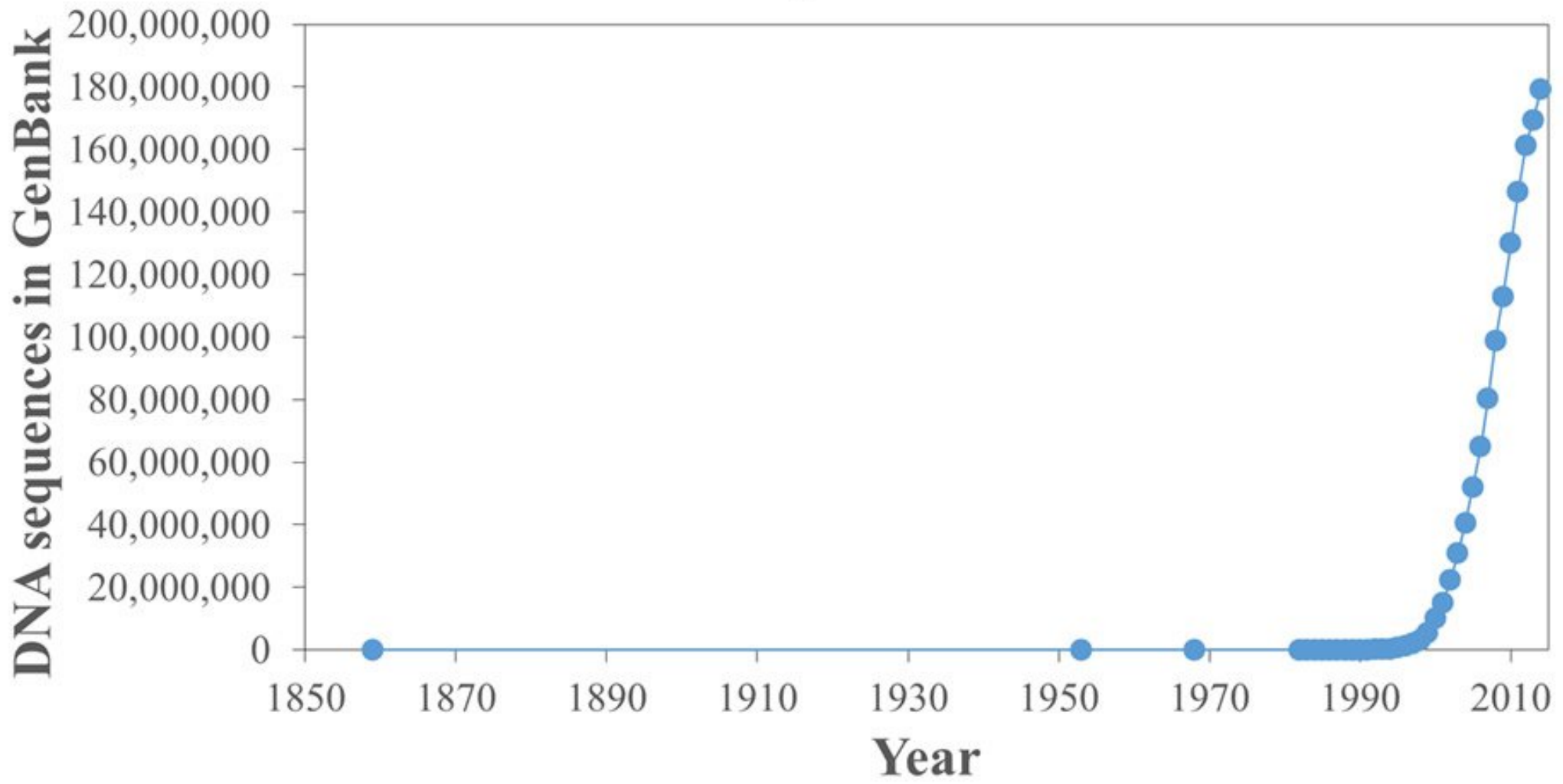


Growth of Sequences & Databases





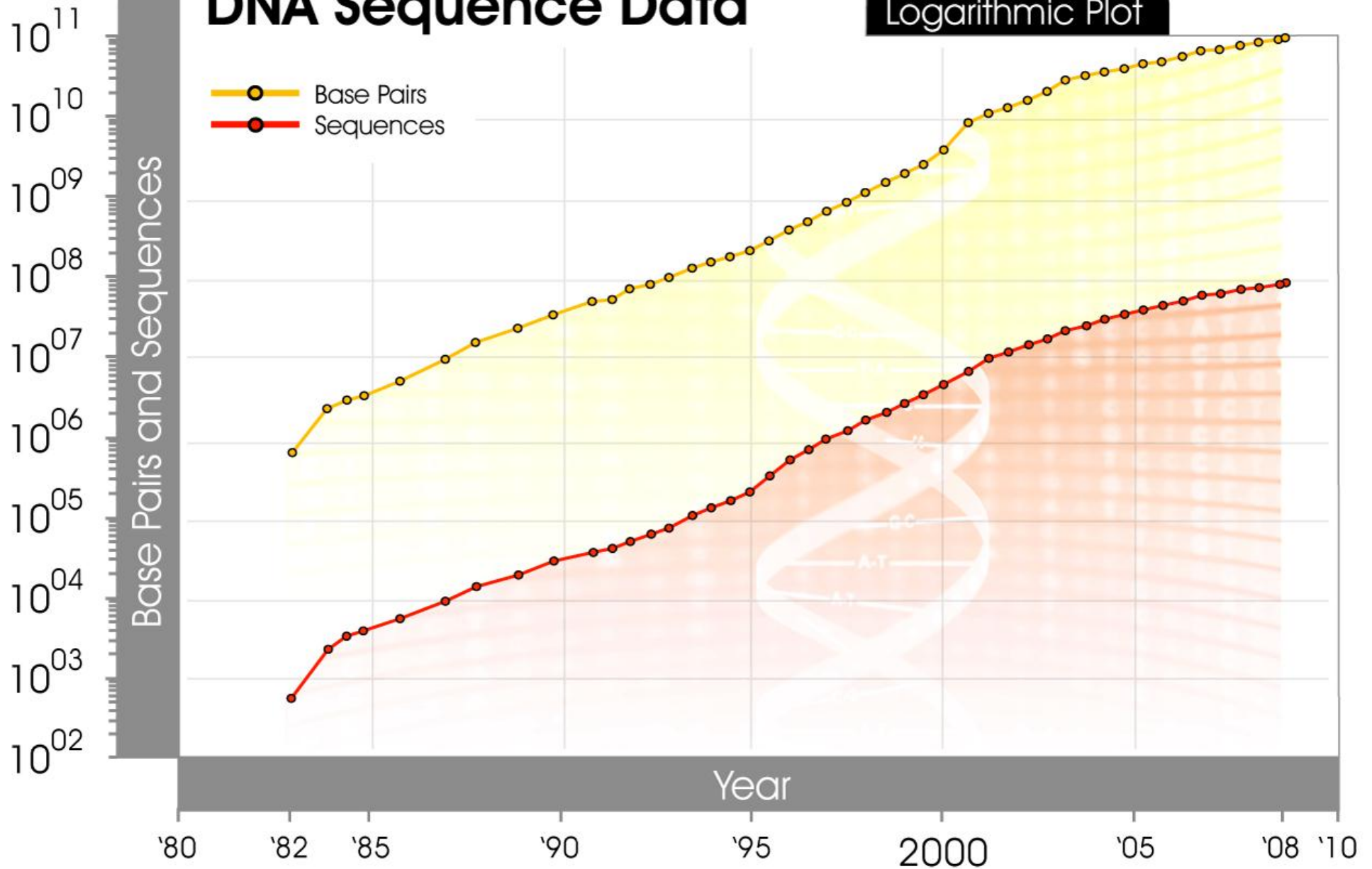
Growth of DNA Sequence Information



Growth in Genbank

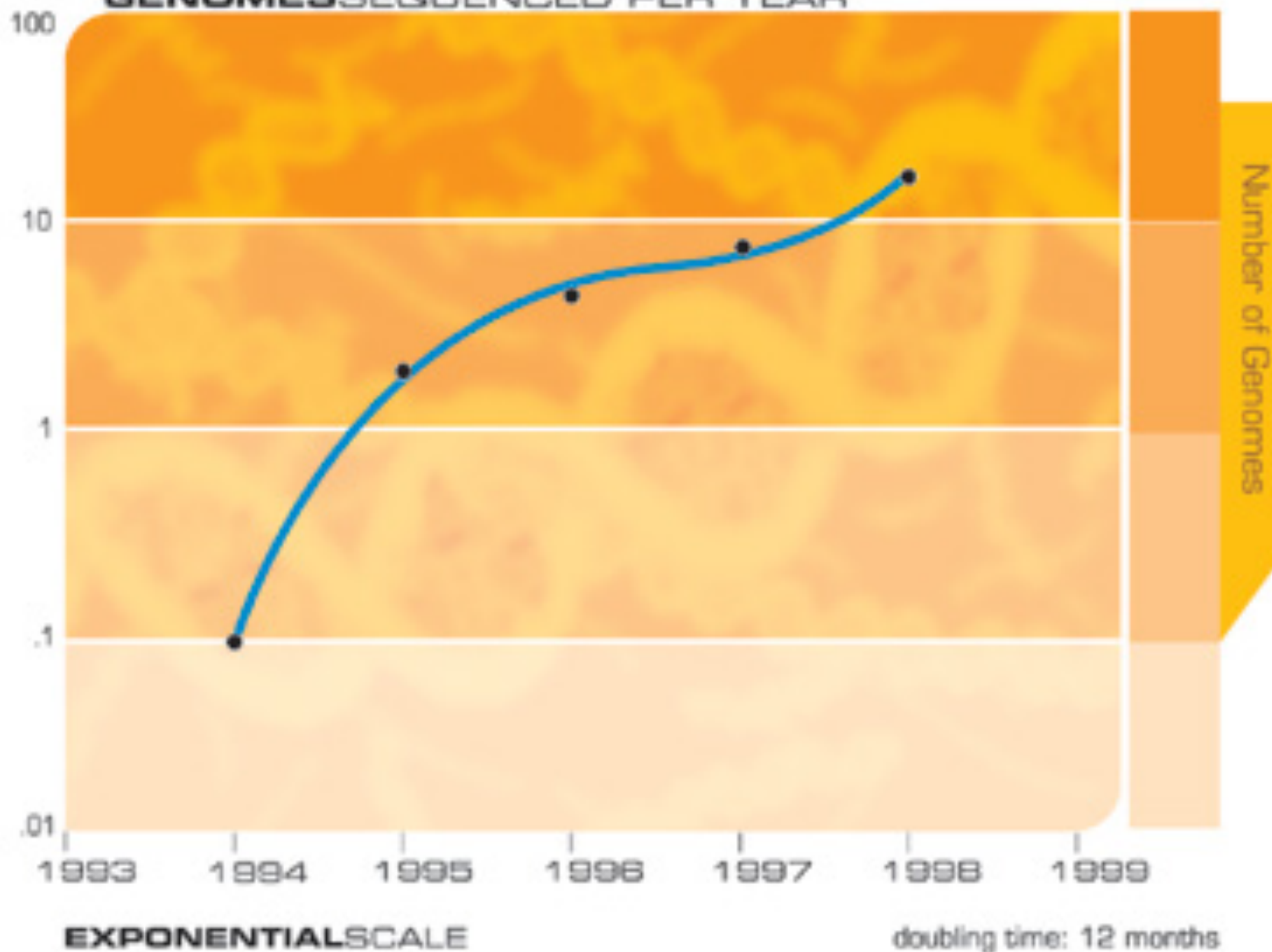
DNA Sequence Data

Logarithmic Plot

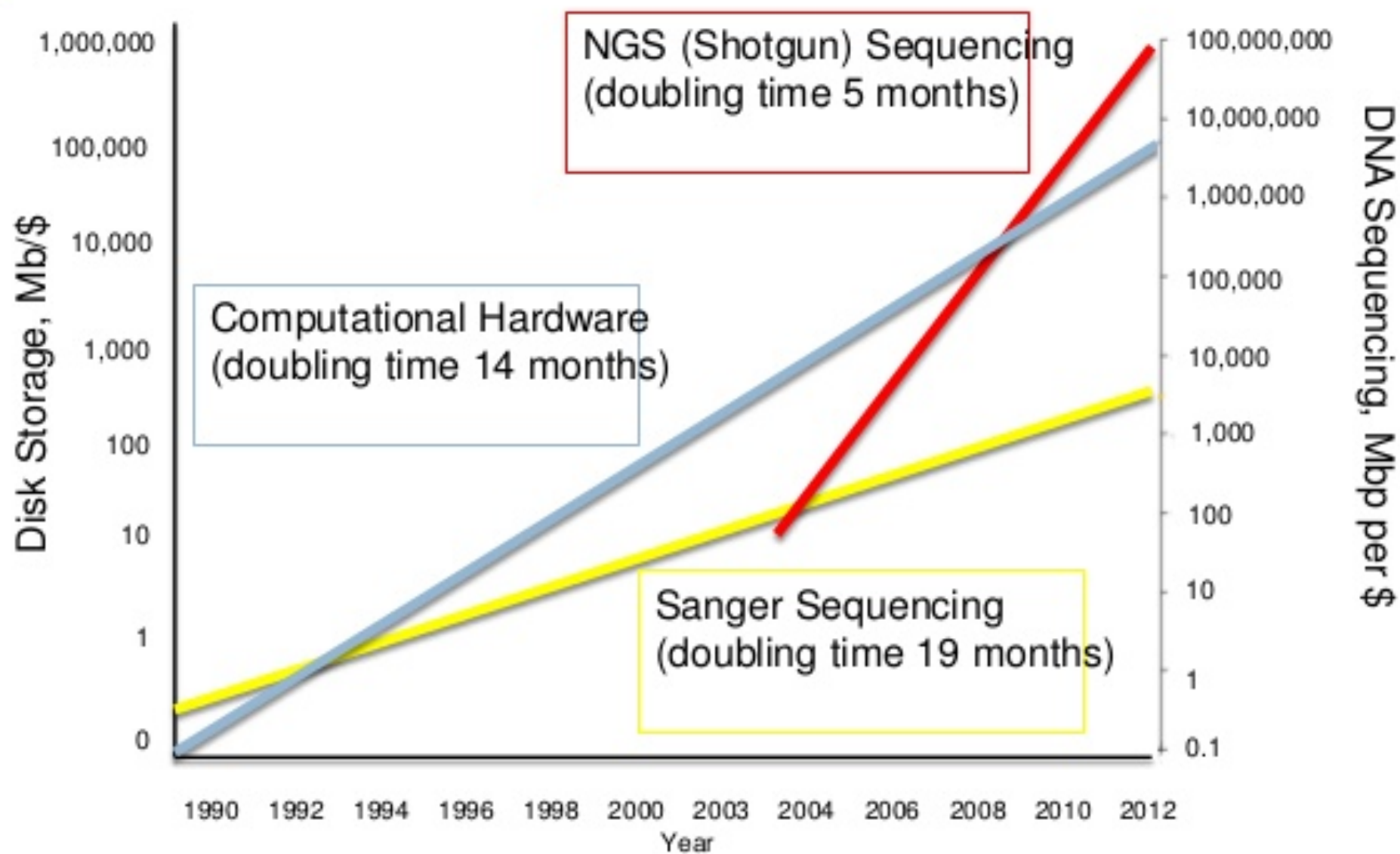




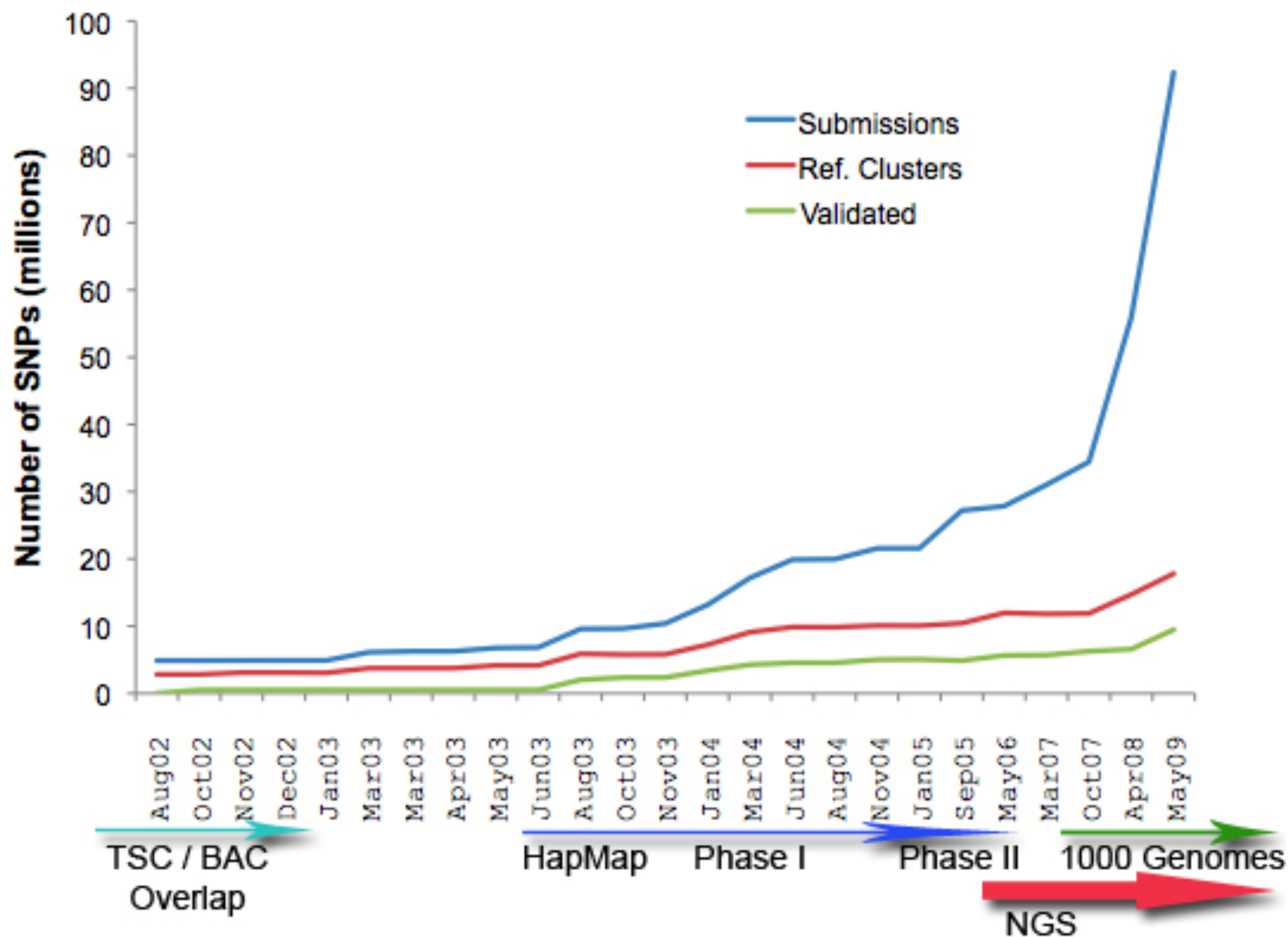
GENOMES SEQUENCED PER YEAR



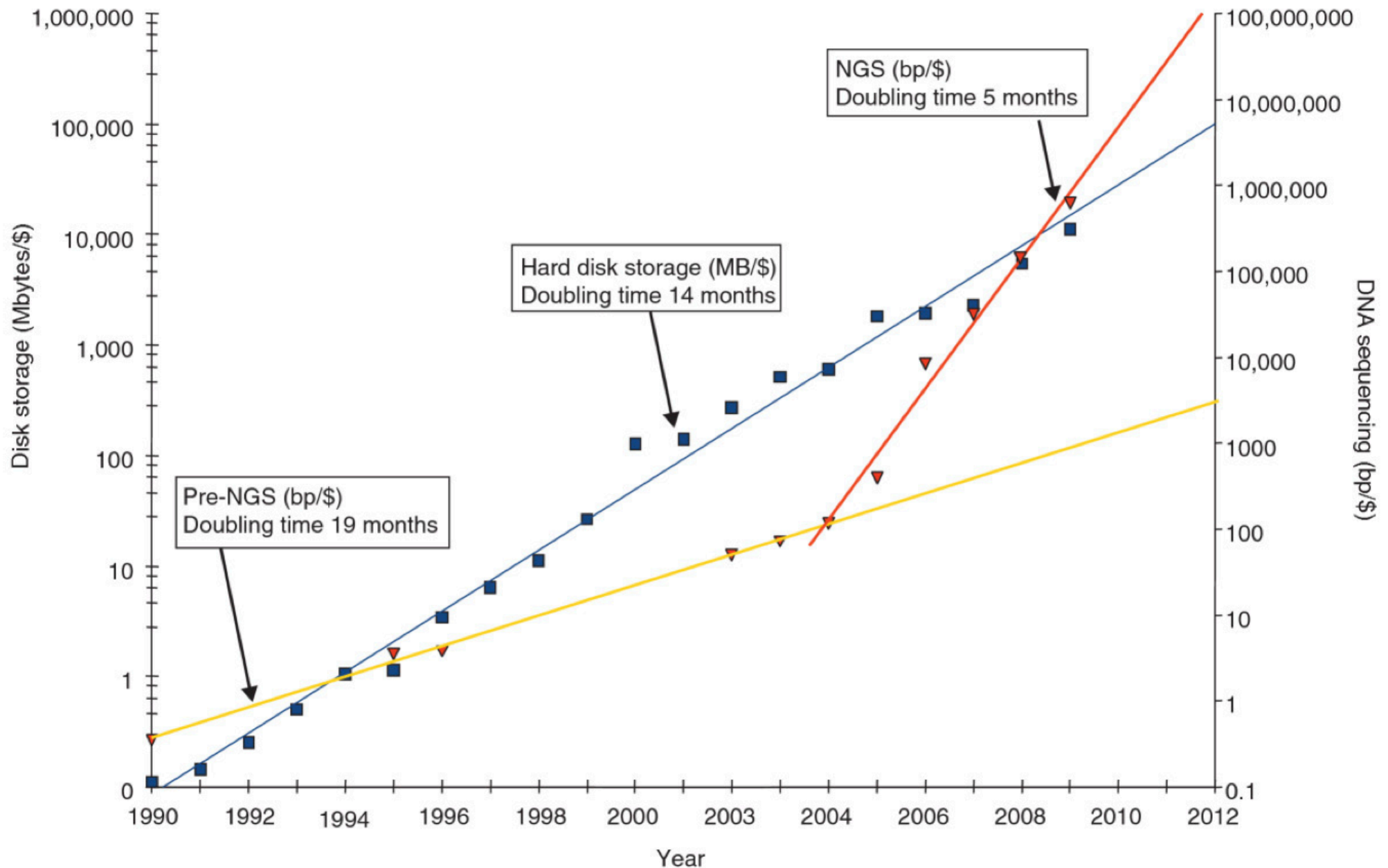
The era of big data in biology



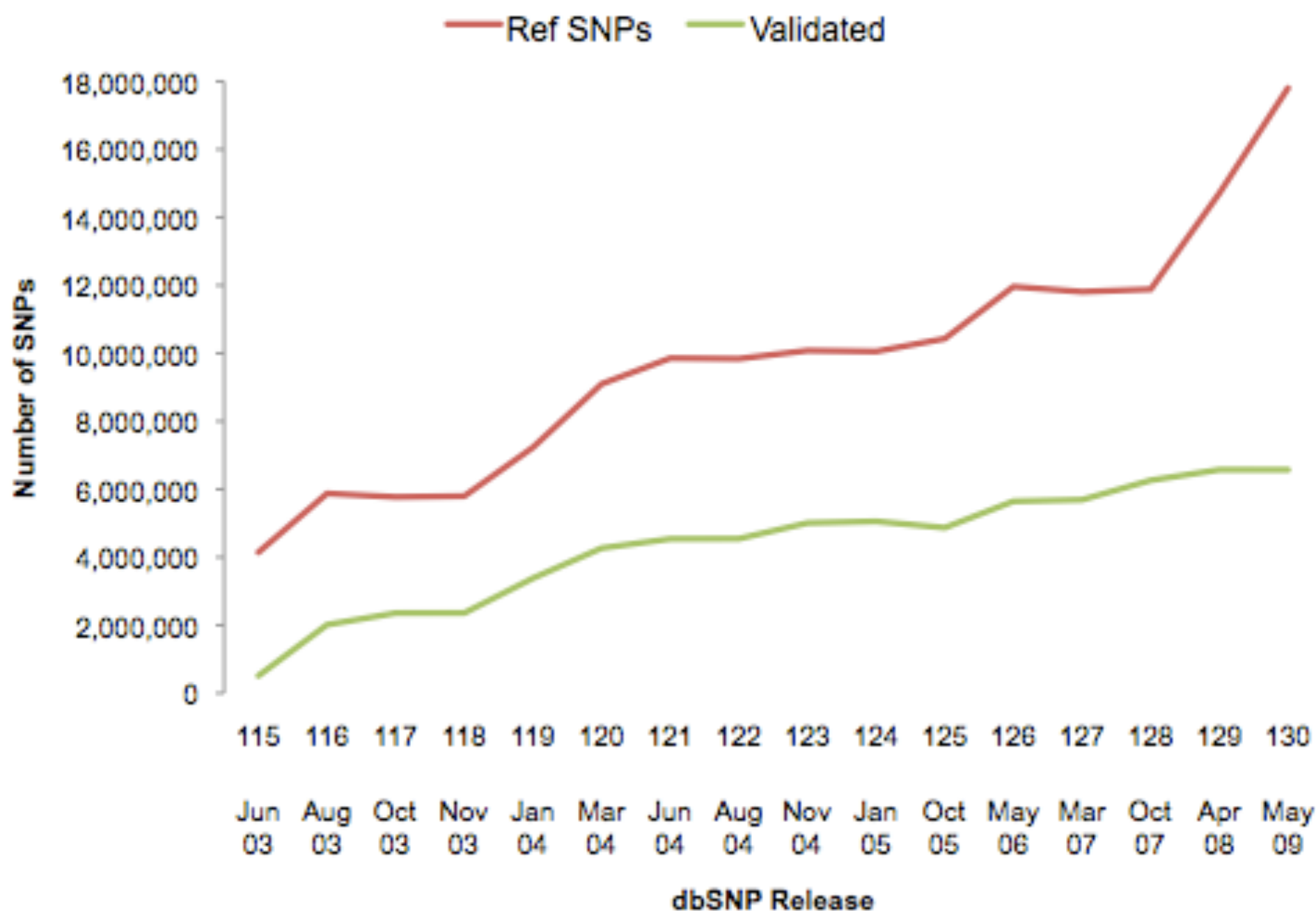
Growth of dbSNP, 2002-2009

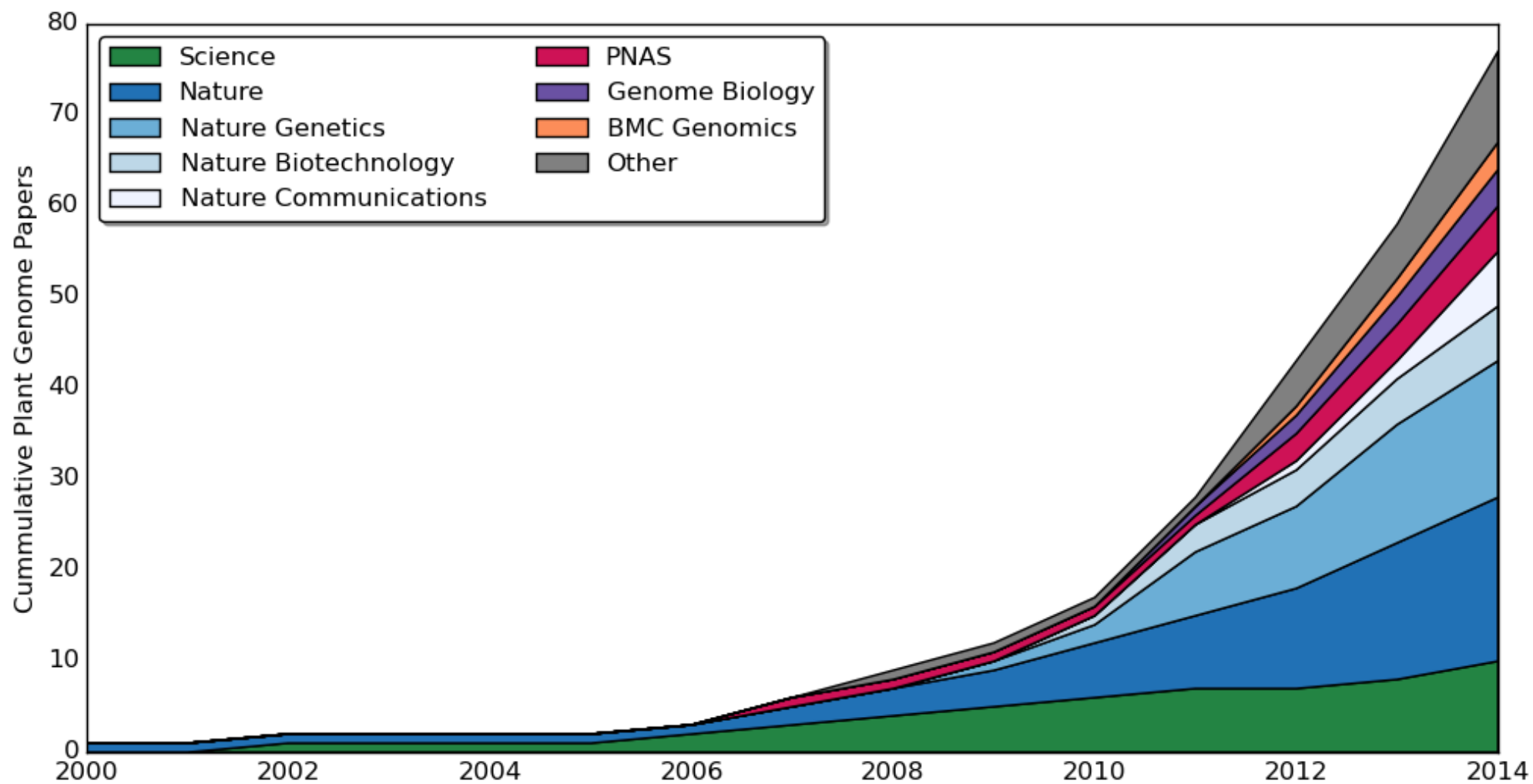


NextGen Sequencing a Game-Changer



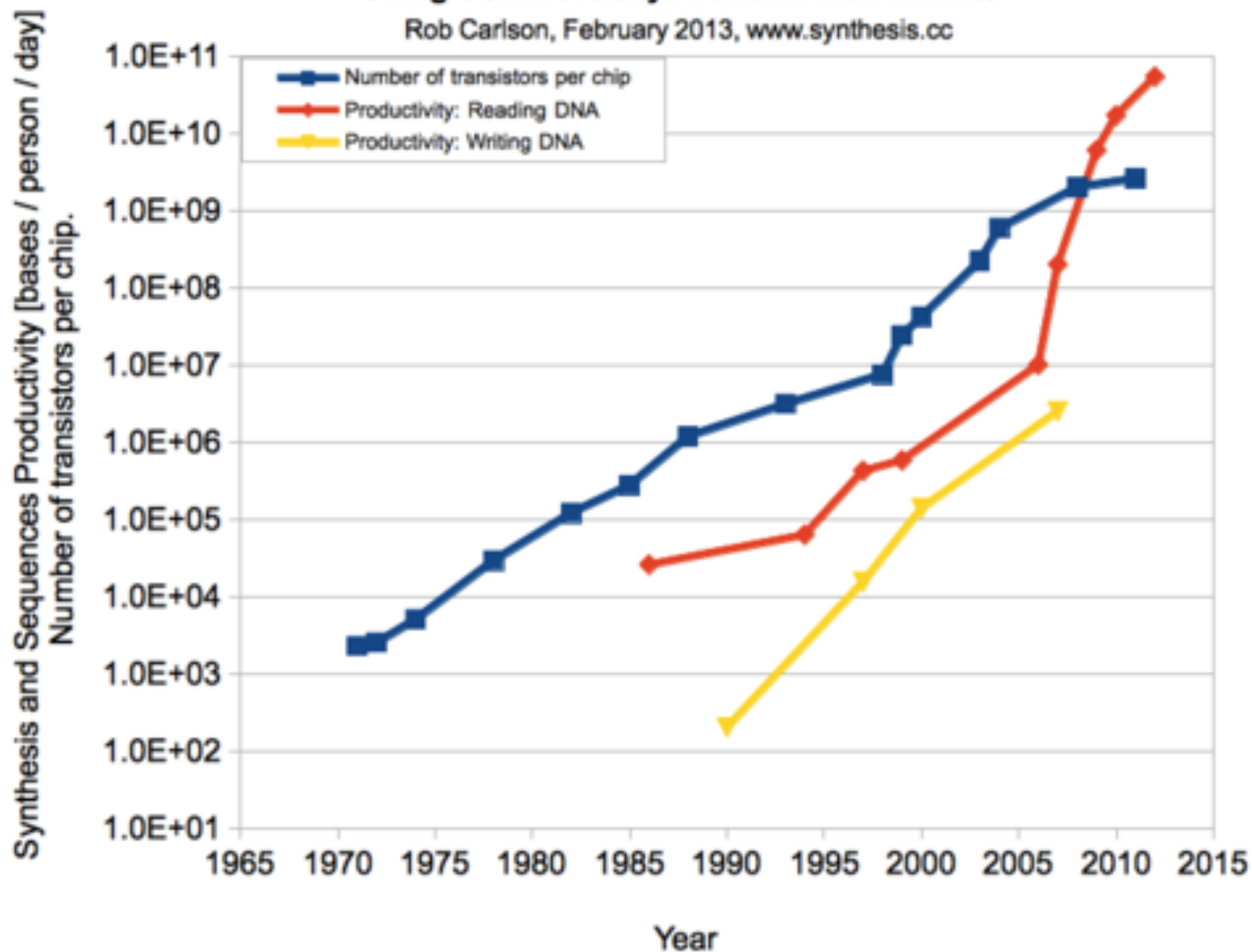
Growth of dbSNP (2003-2009)

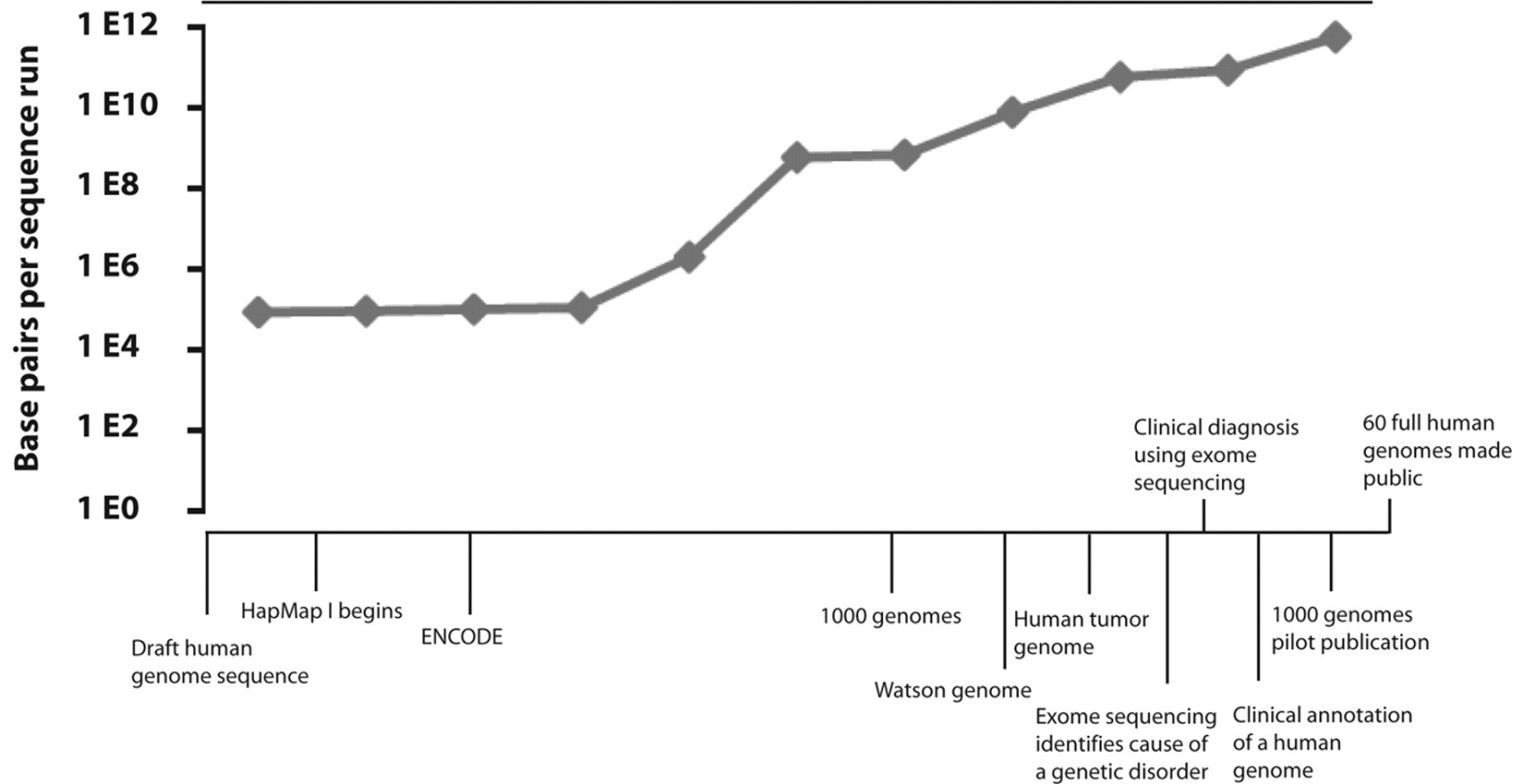
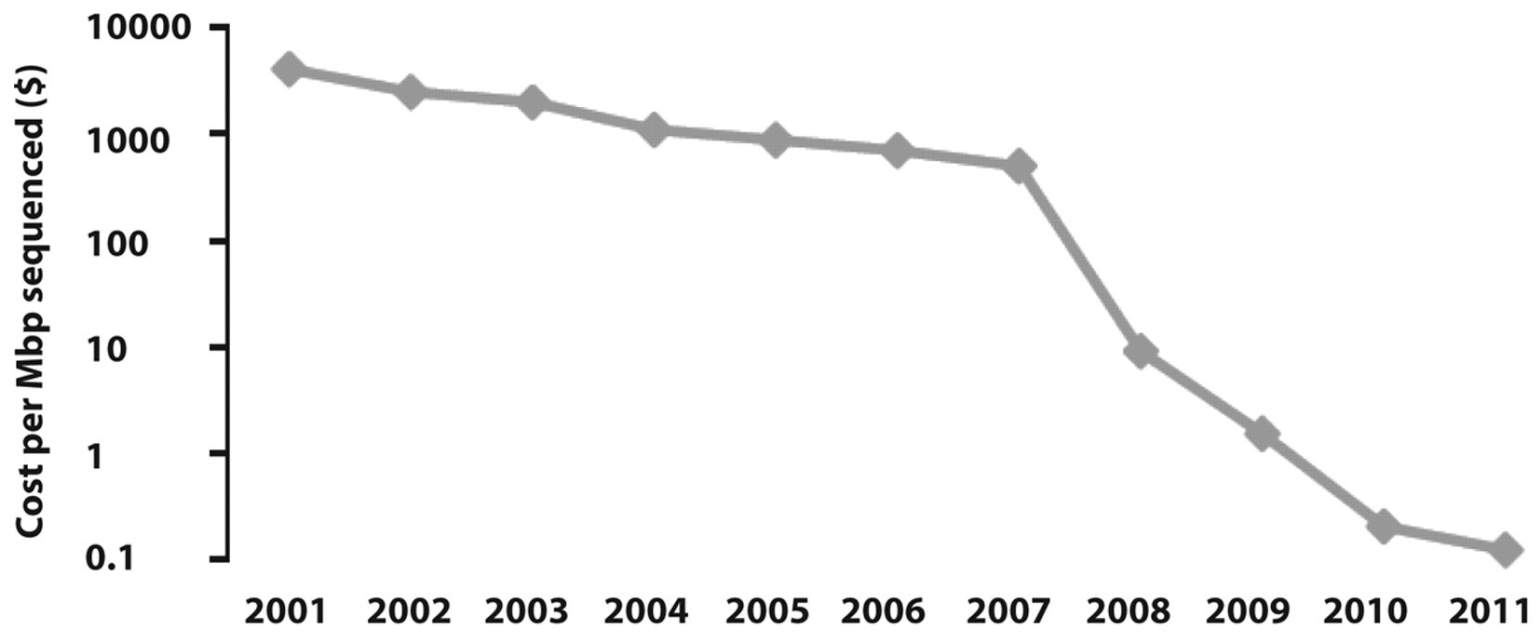


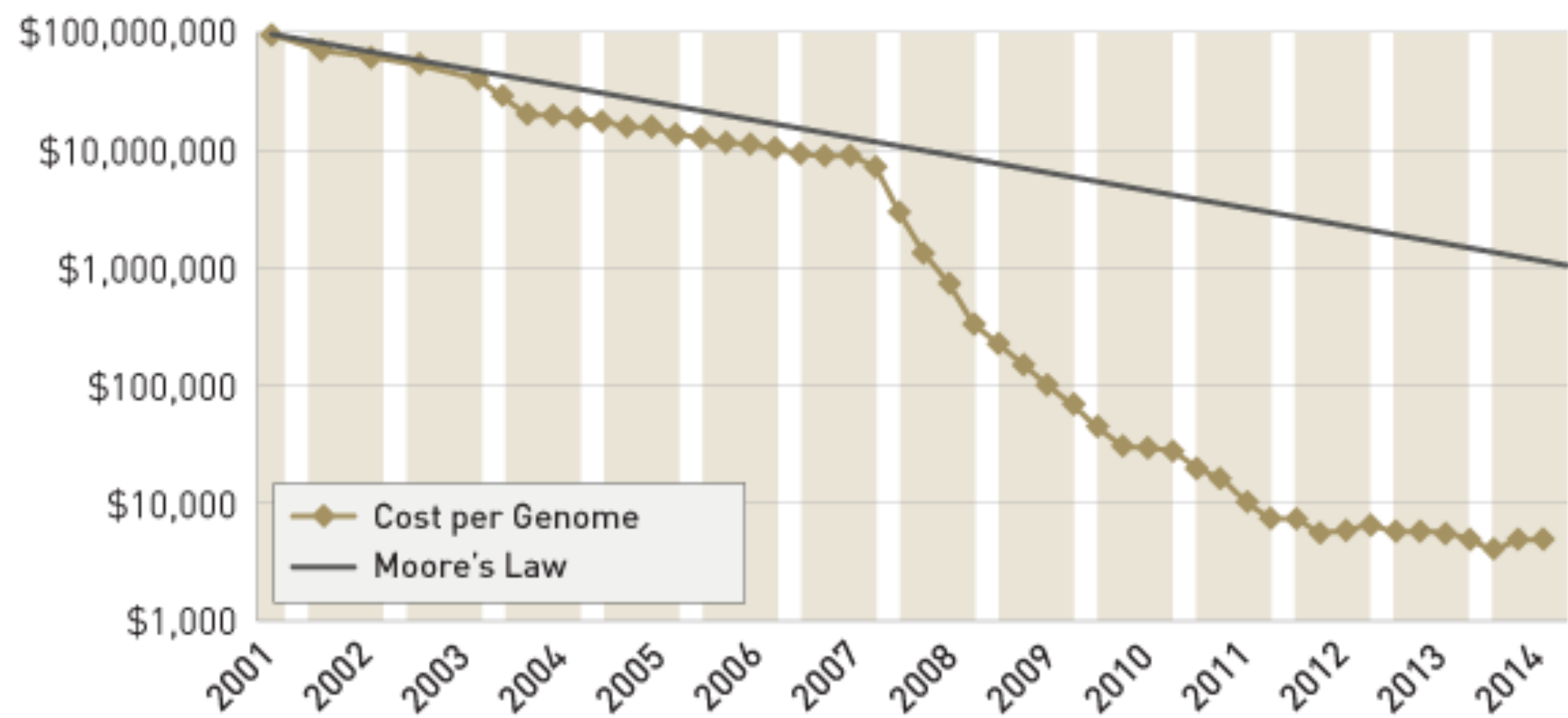


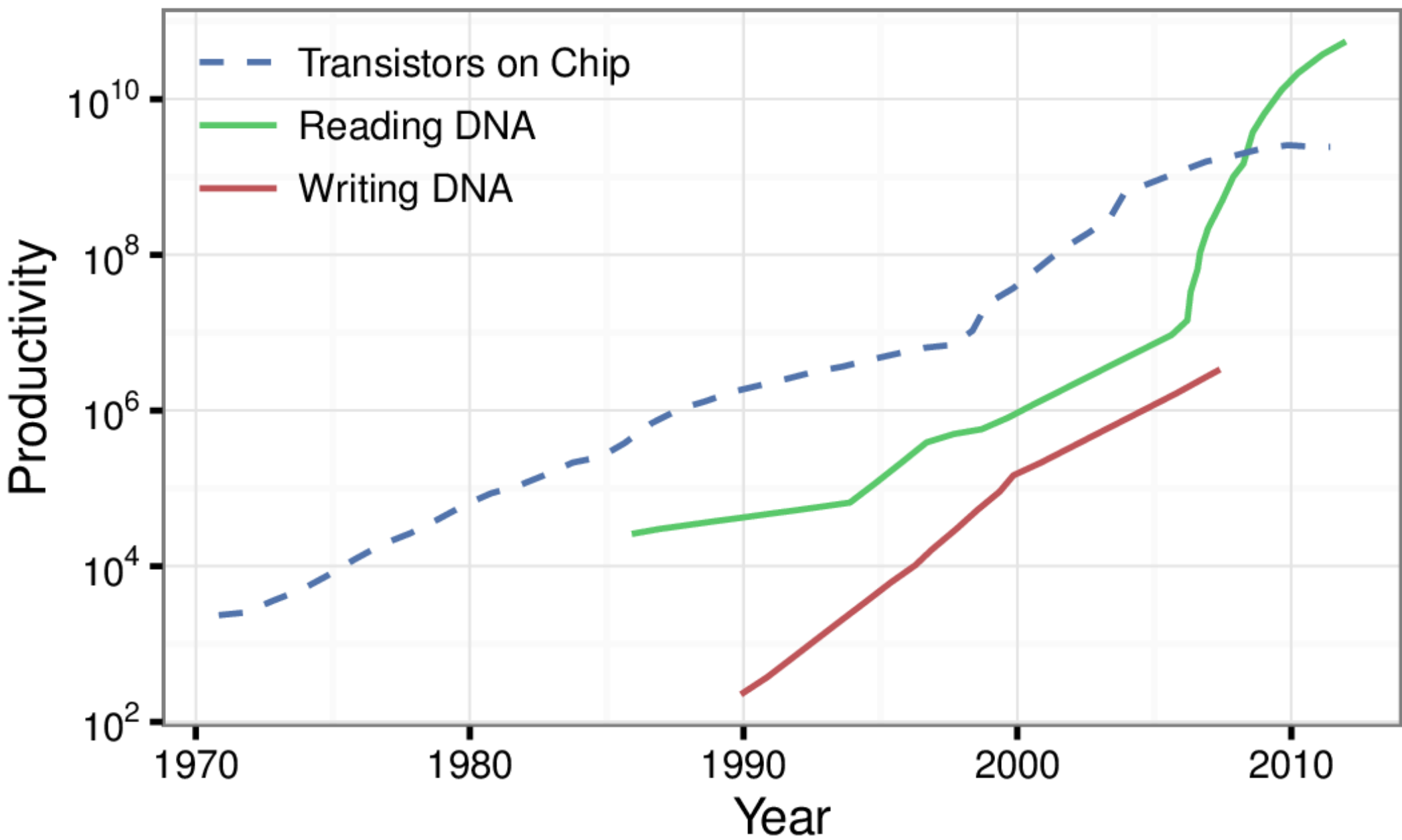
Productivity in DNA Synthesis and Sequencing Using Commercially Available Instruments

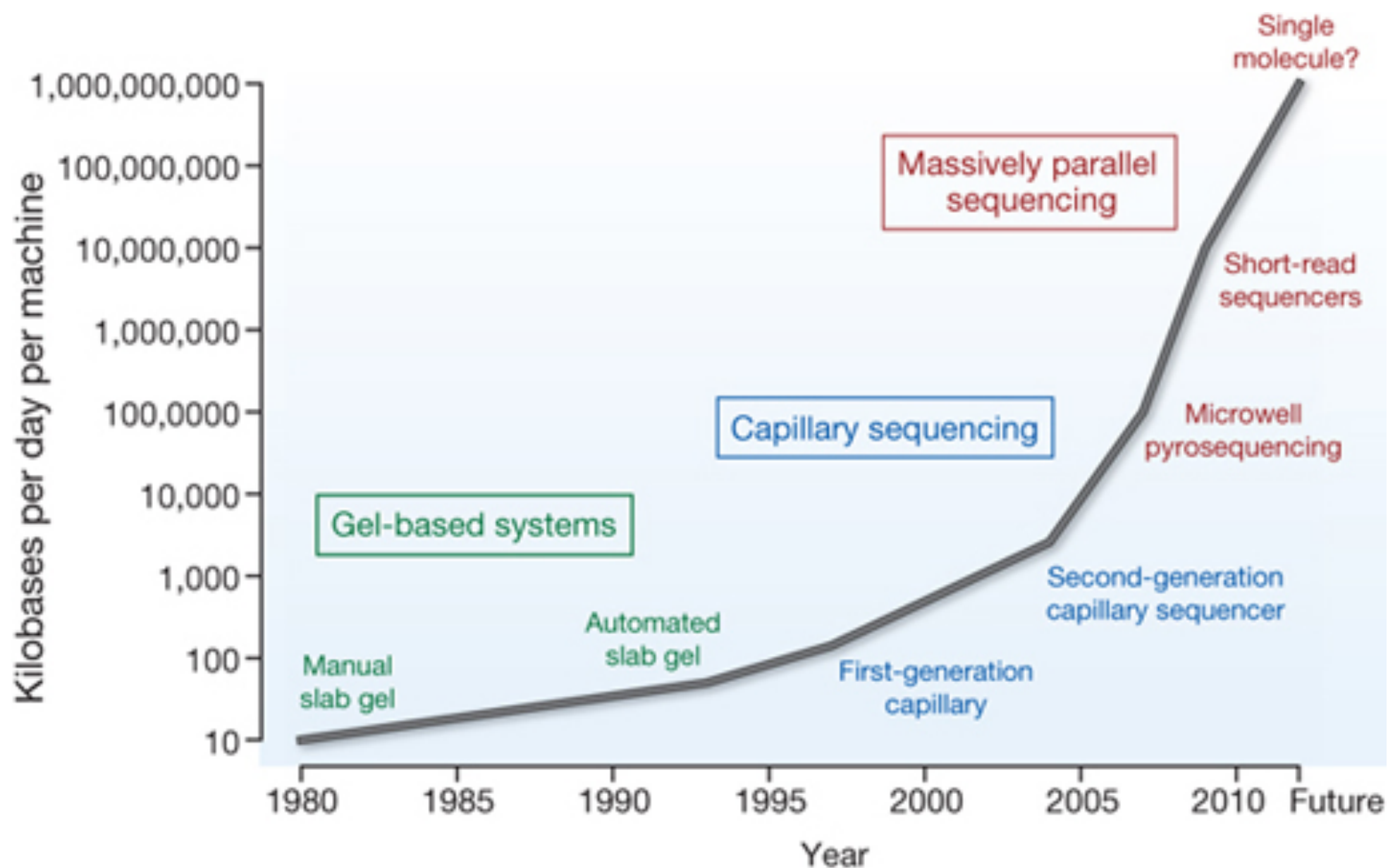
Rob Carlson, February 2013, www.synthesis.cc



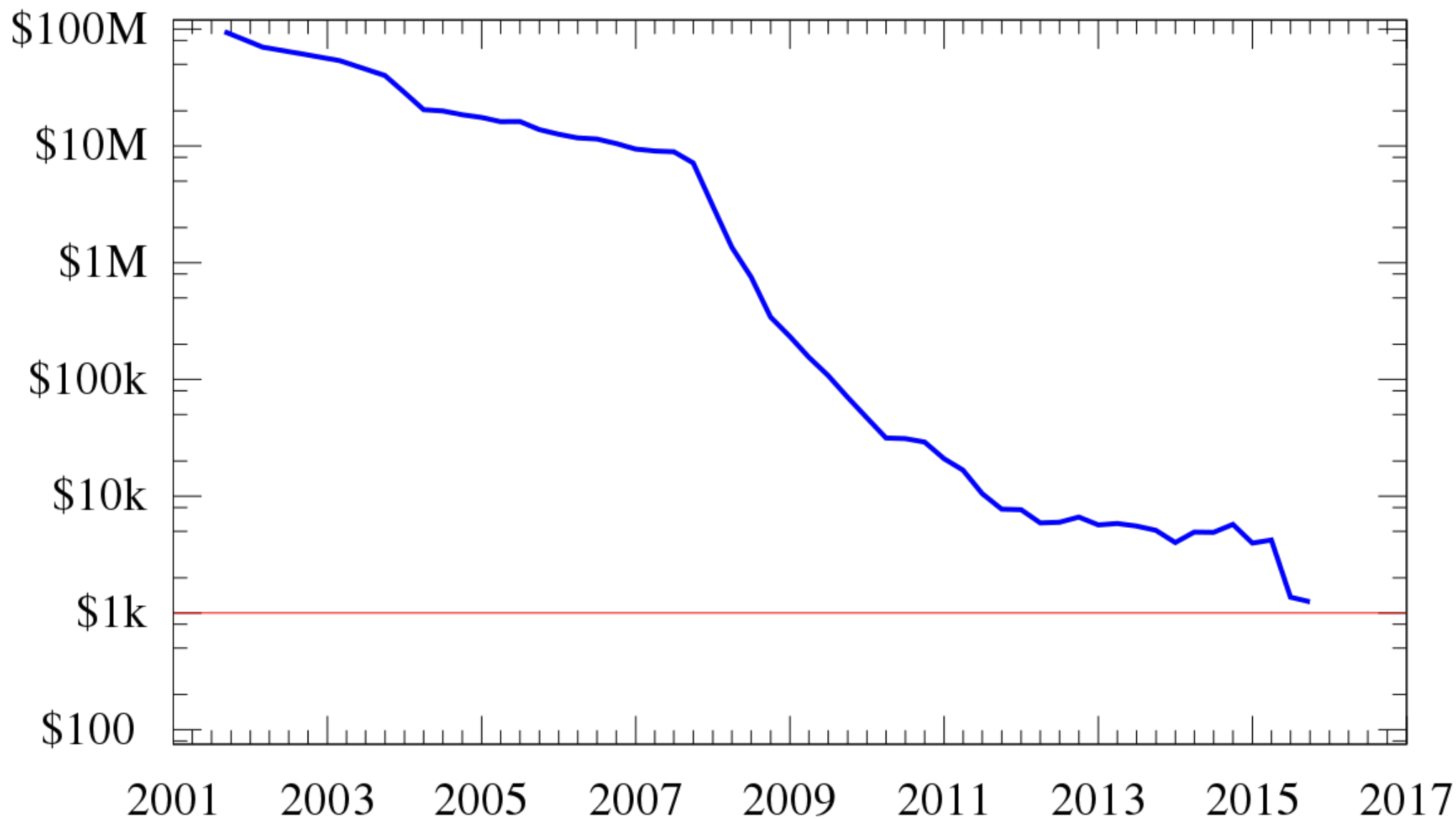




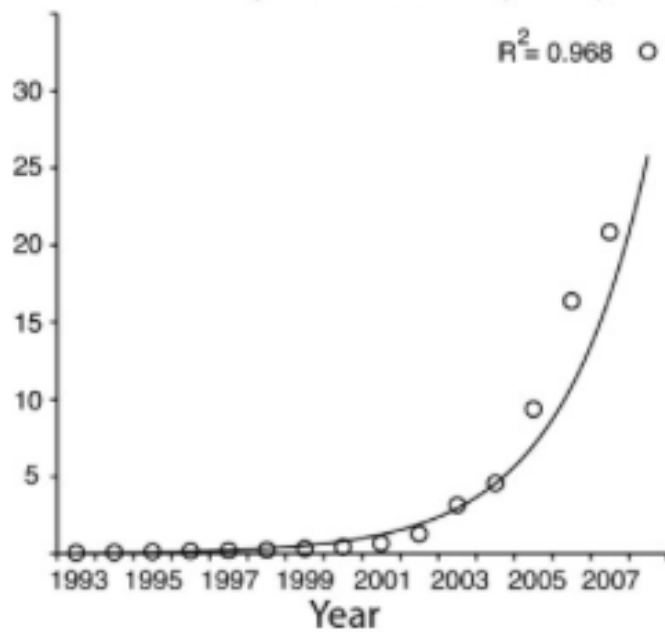




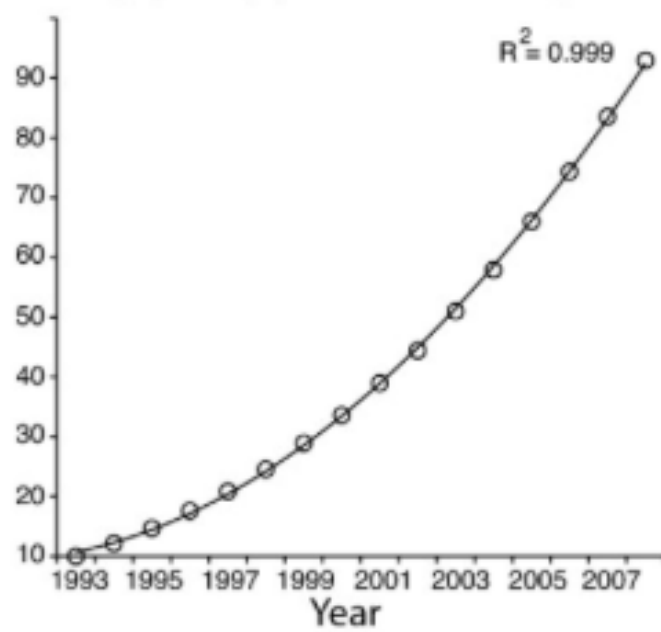
Cost to sequence a human genome (USD)



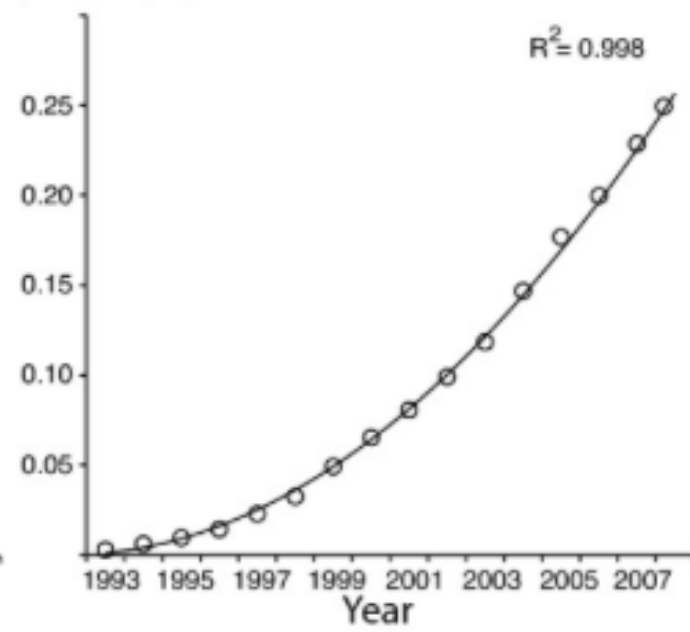
a. Vertebrate sequences in GenBank (millions)

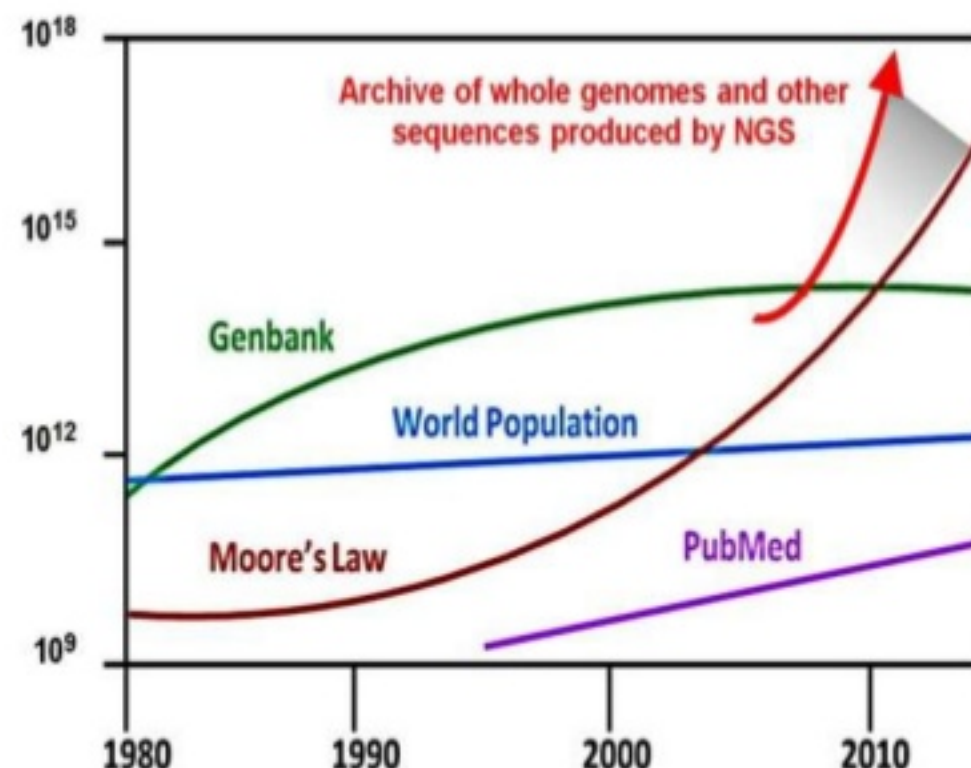
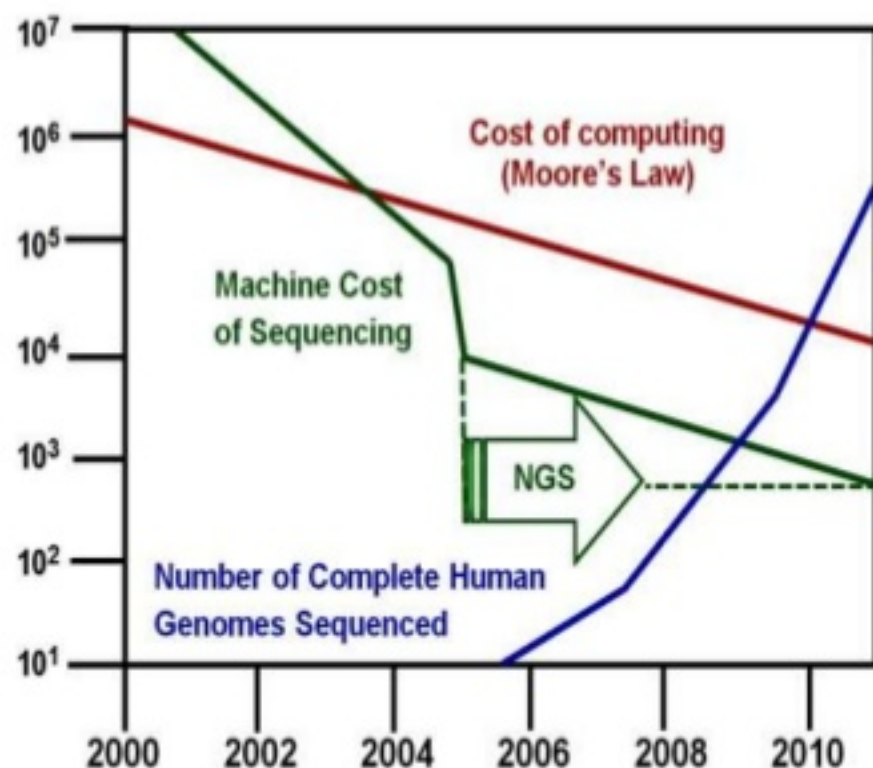


b. Phylogenetic papers in Web of Science (thousands)



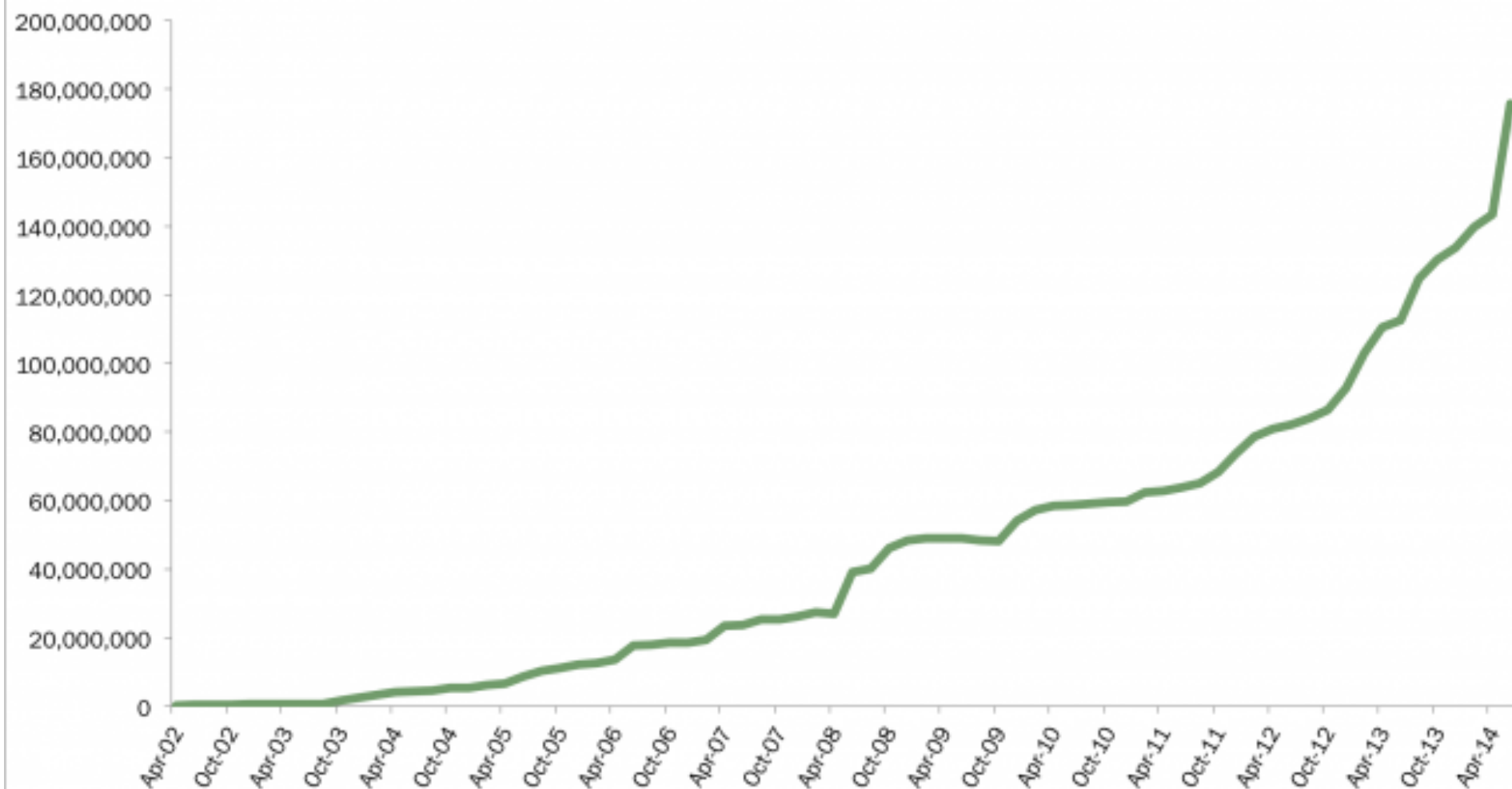
c. Phylogenetic resolution

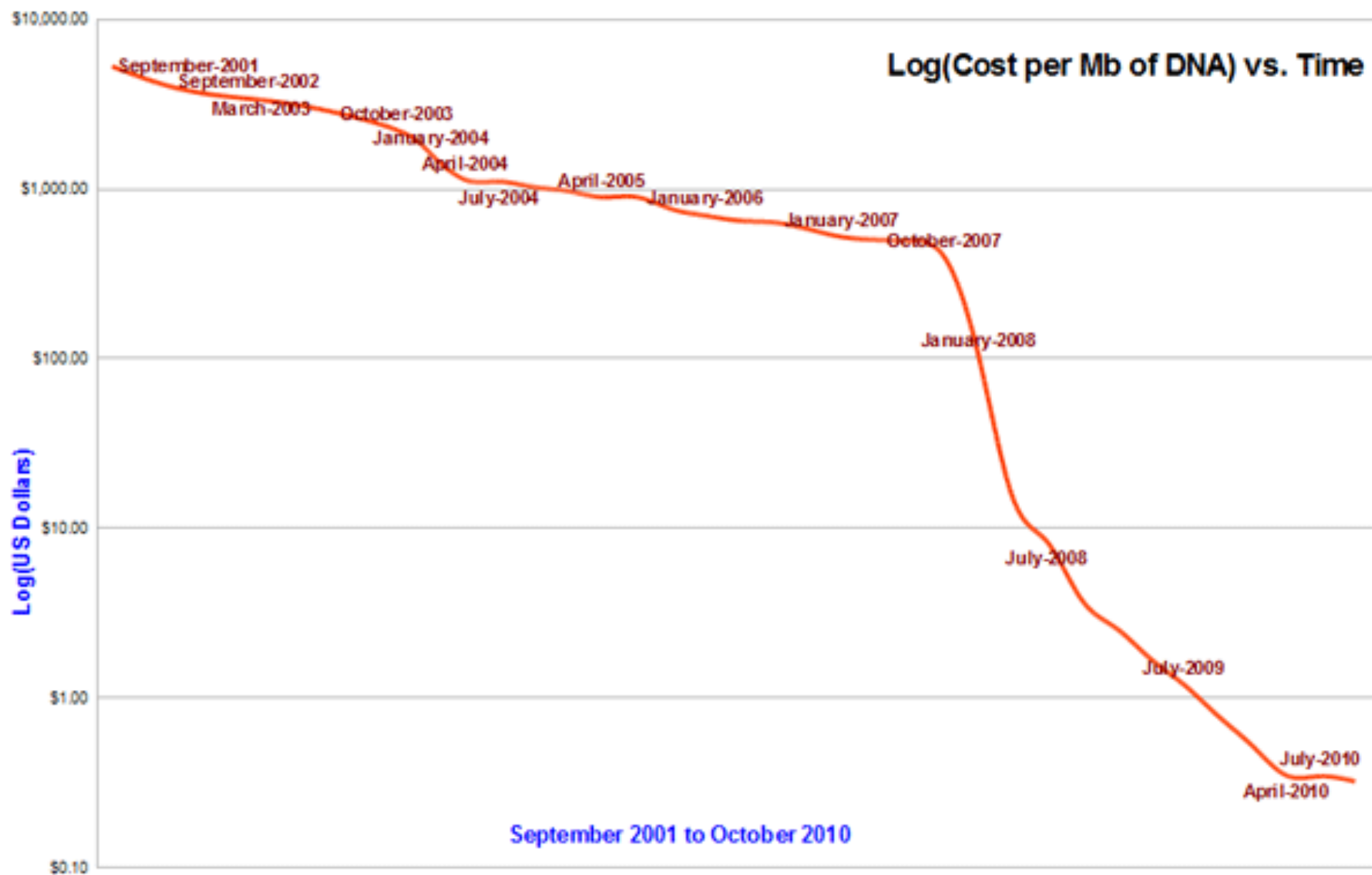


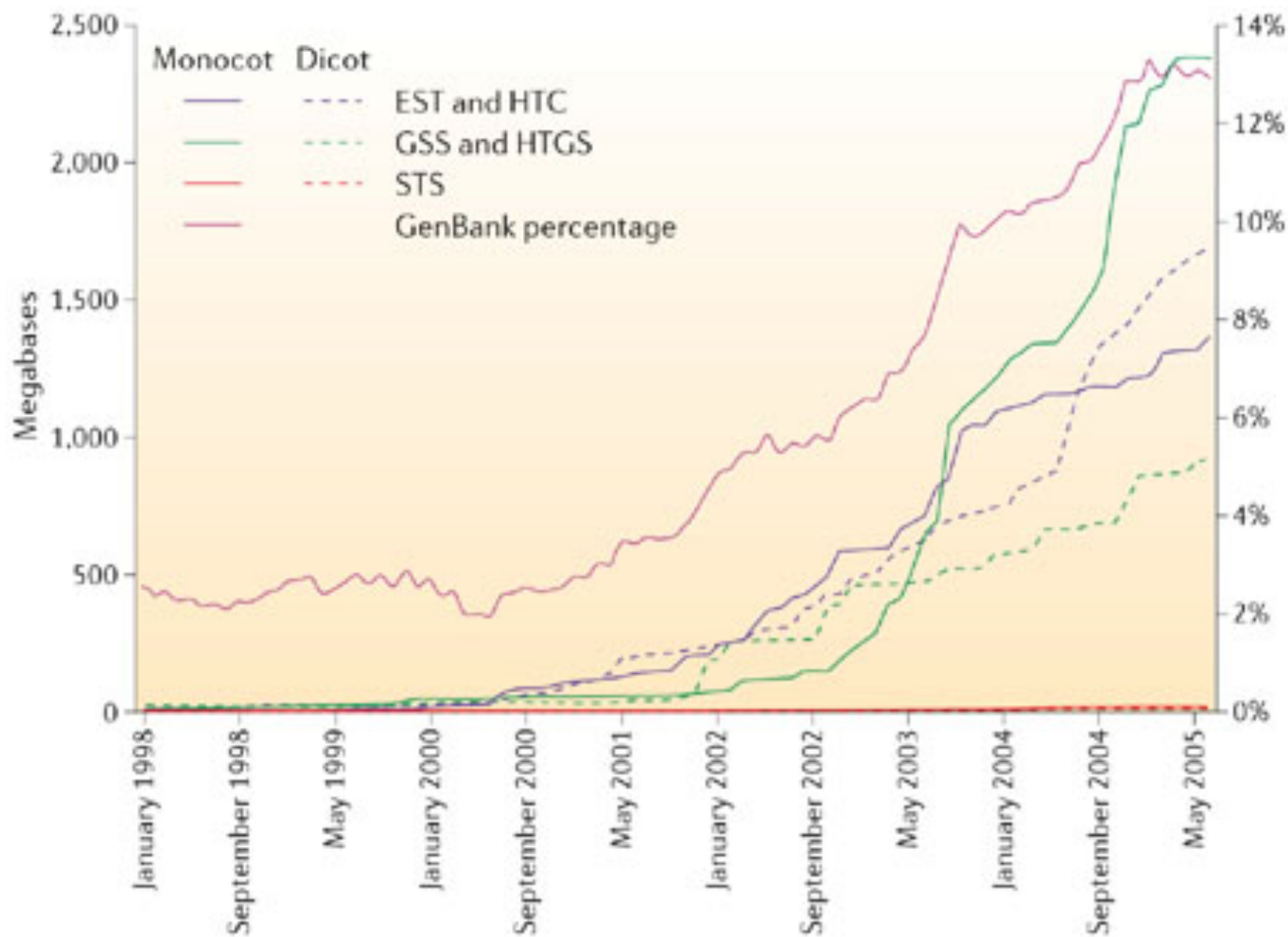


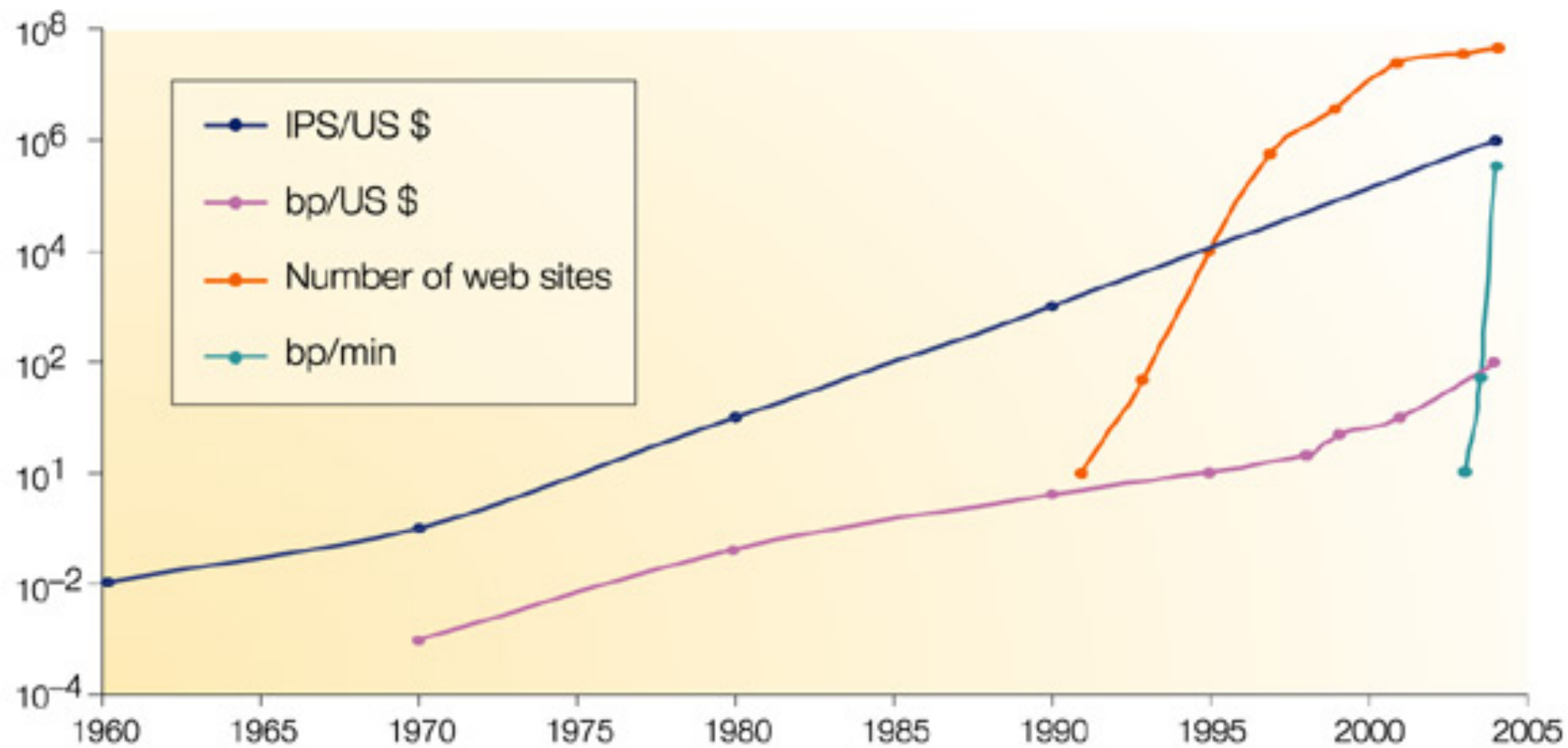
**As the cost of DNA sequencing falls,
the growth of human genome data becomes exponential**

Number of Whole Genome Sequencing Projects
in GenBank Database Over Time



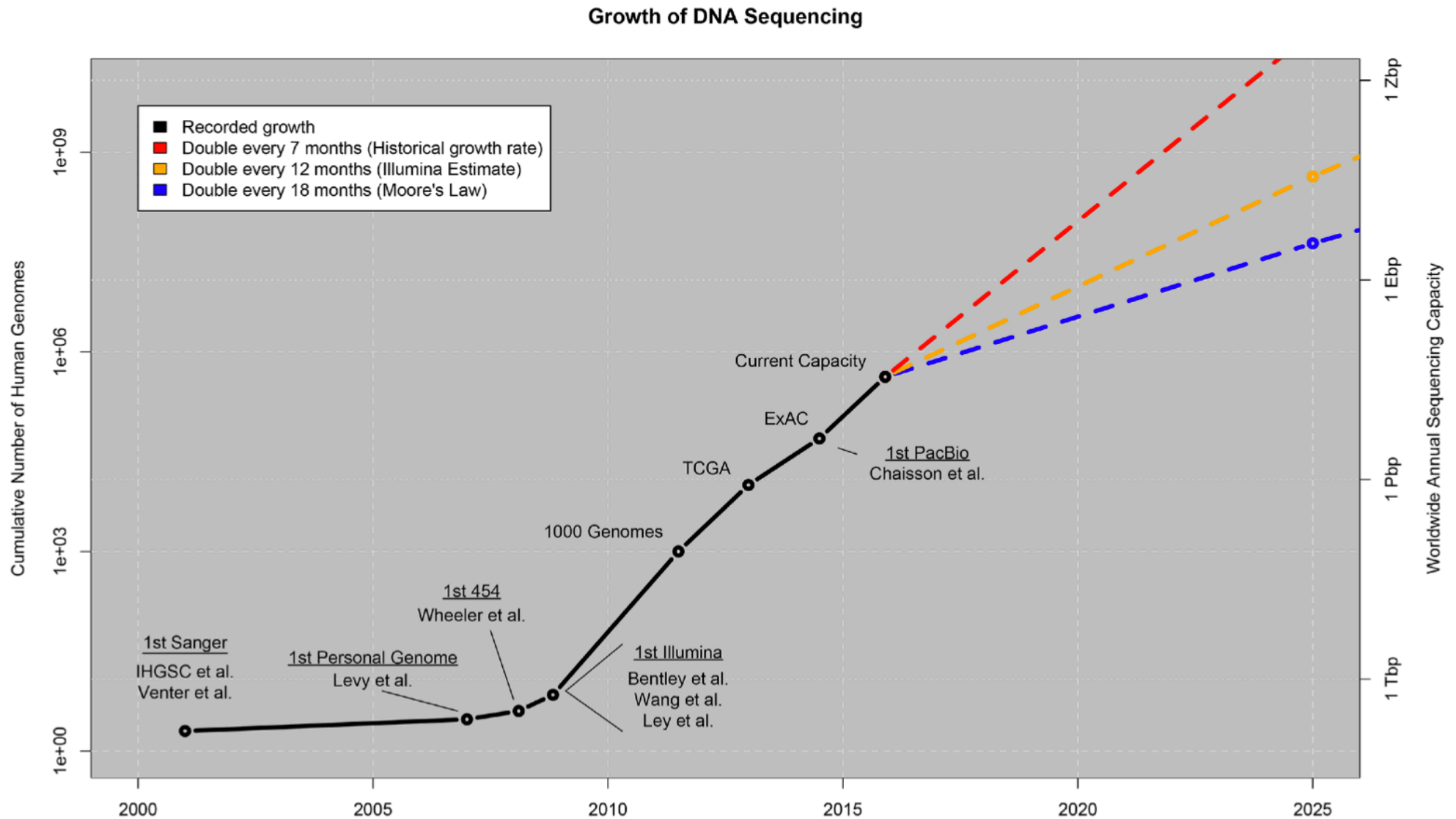






PERSPECTIVE

Big Data: Astronomical or Genomical?



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.

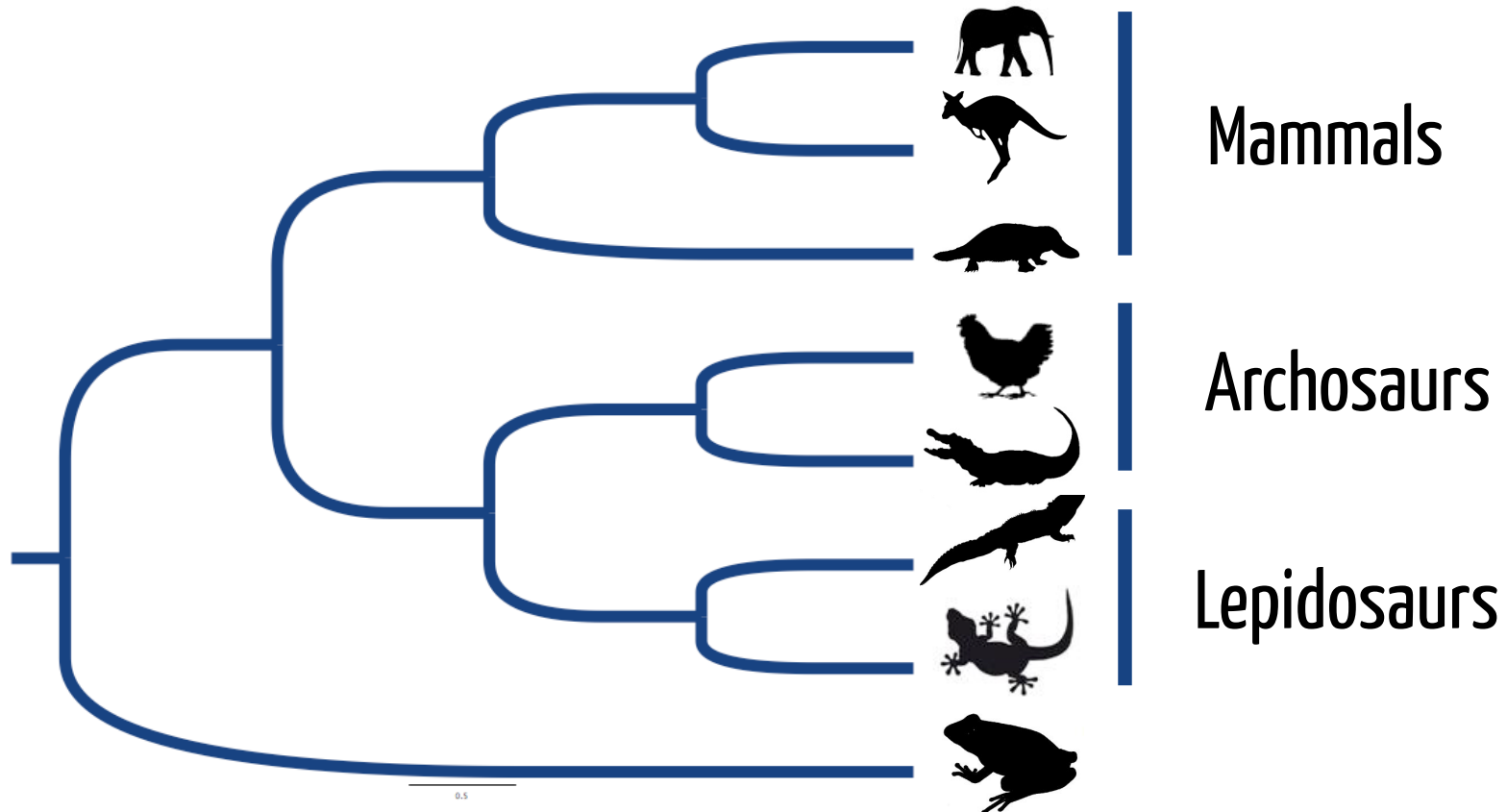
This is no different than it’s always been.

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)

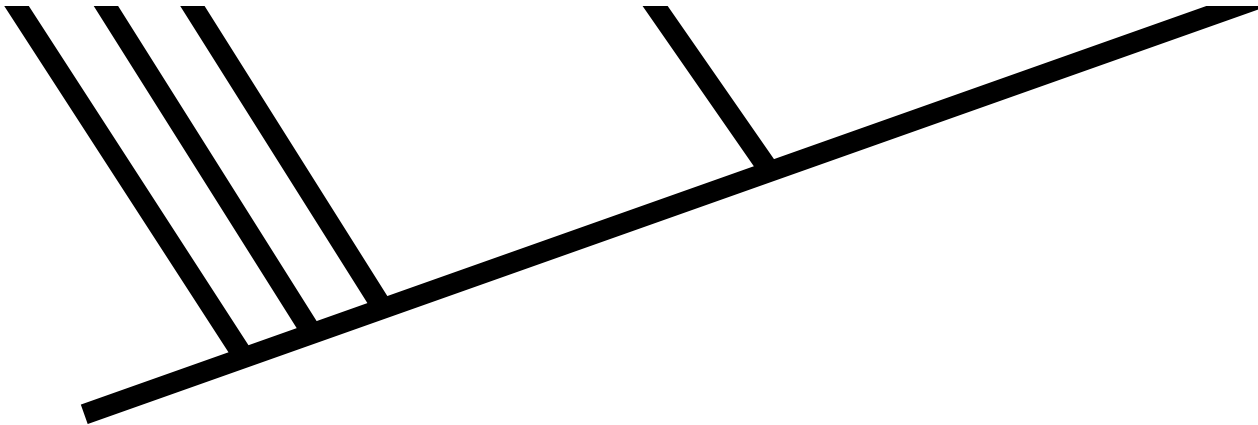
Anapsid



Diapsid



Synapsid



Osteology

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

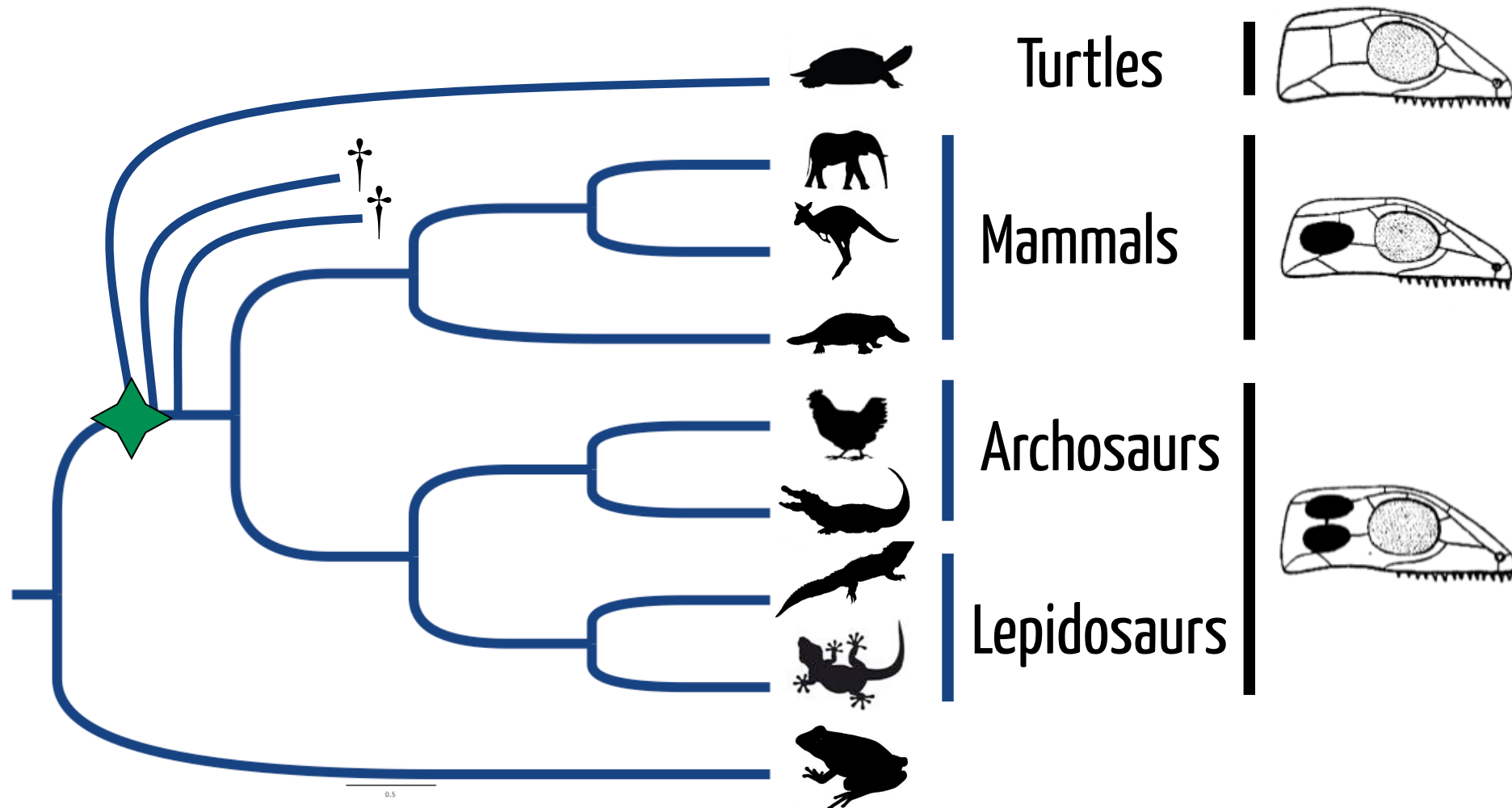
Anaps

id



Osteology

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



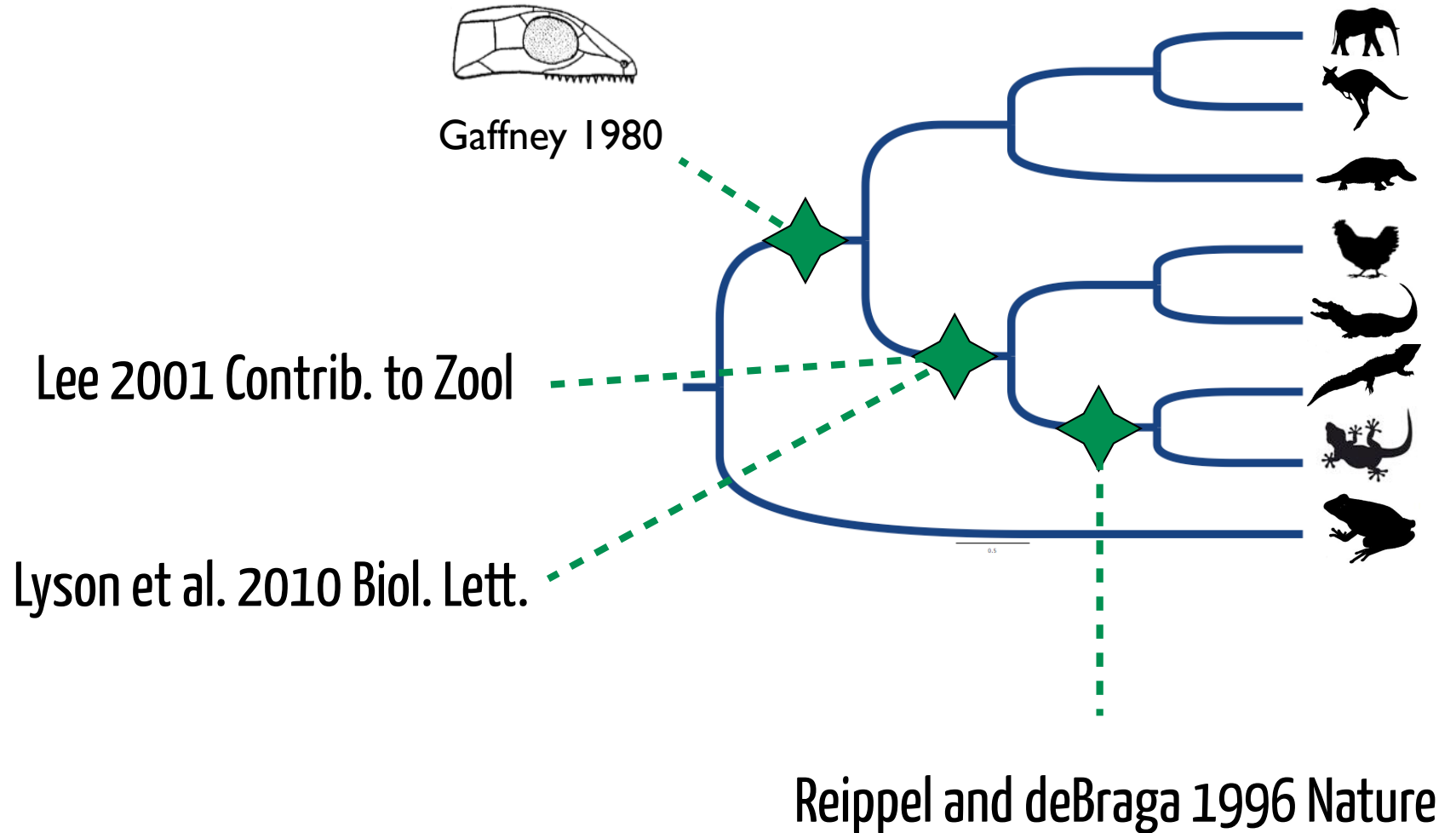
Günther 1867, Gaffney 1980

Osteology

- Primary issue with this hypothesis

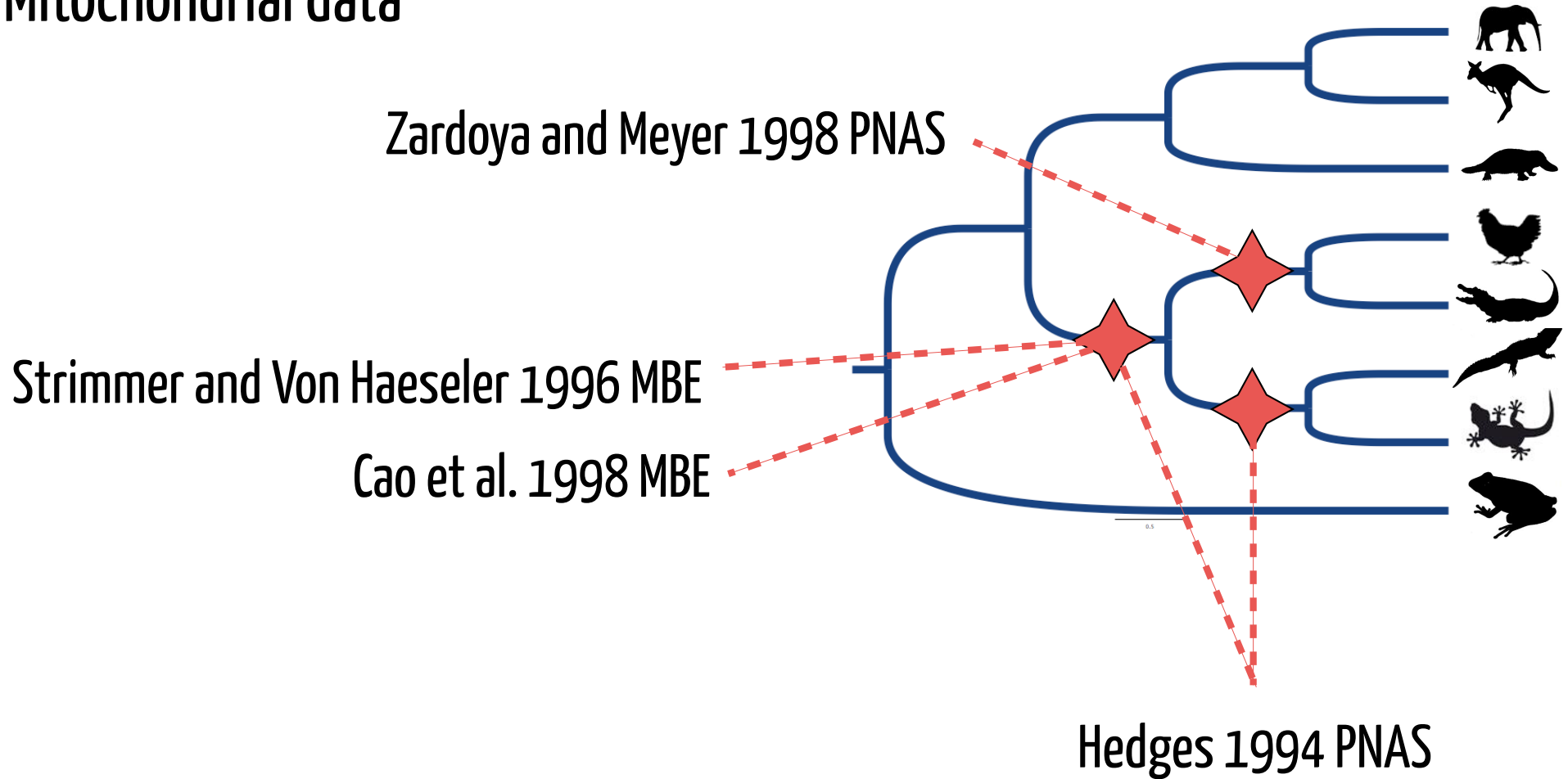


More osteology



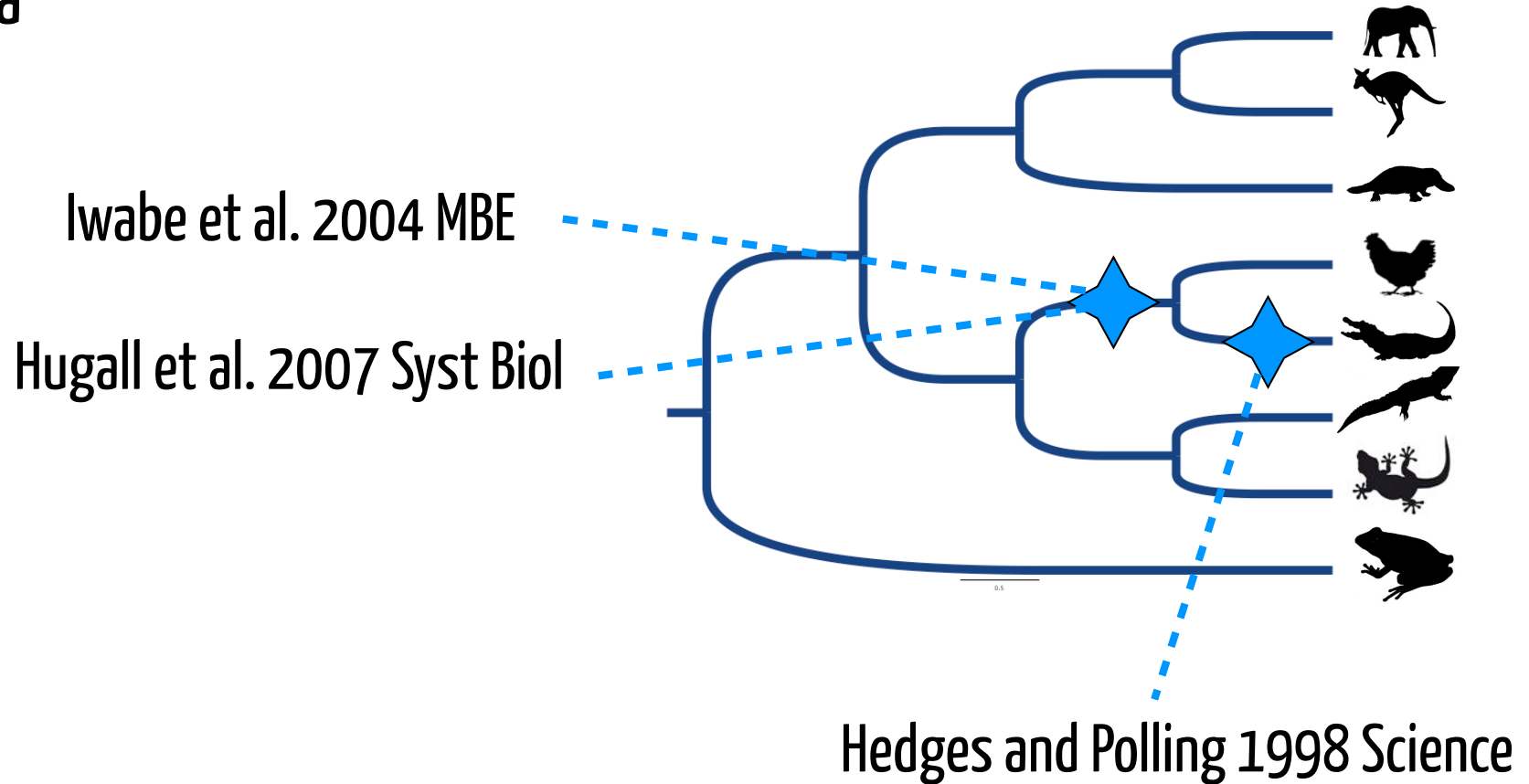
Molecular Information

- Mitochondrial data

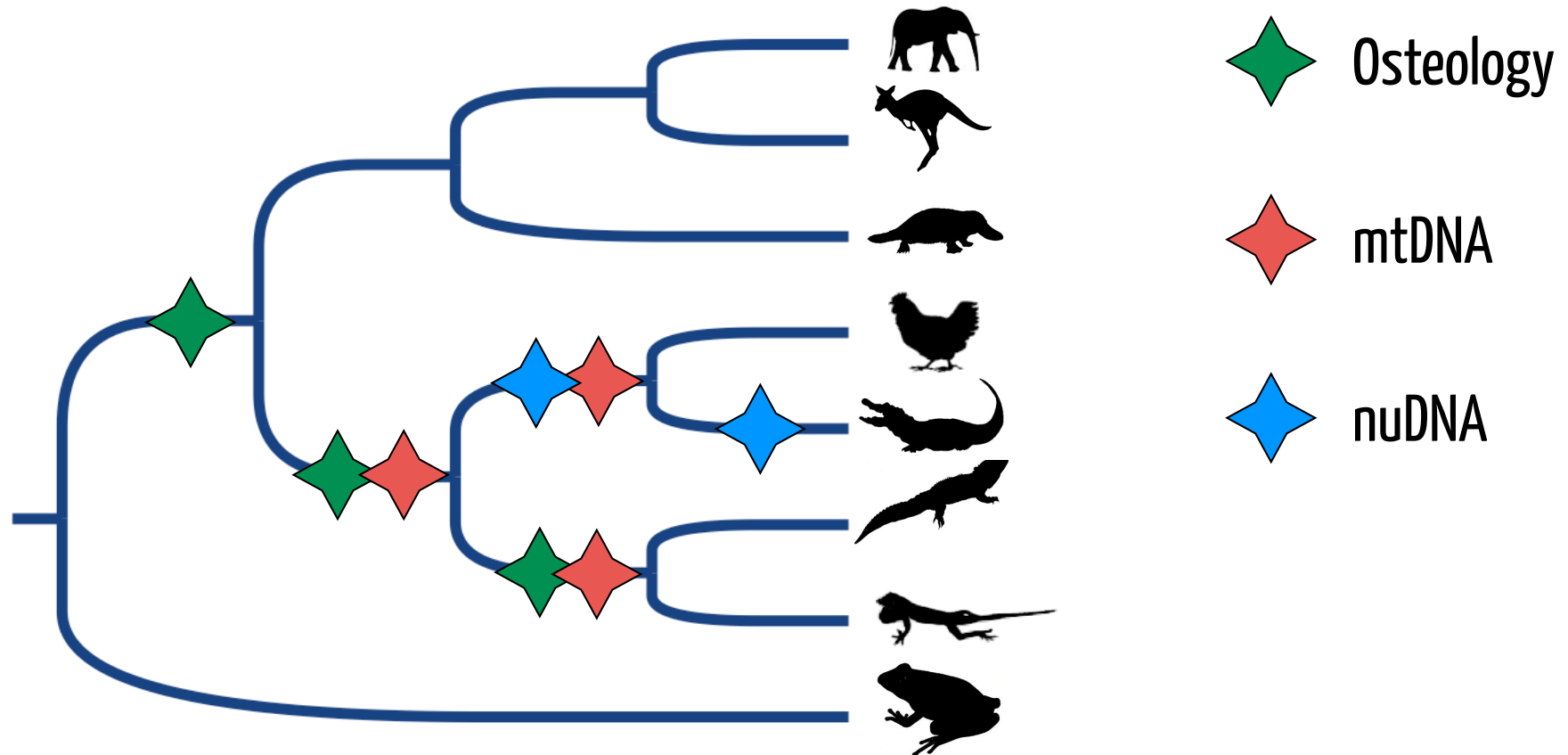


Molecular Information

- Nuclear data



Summary



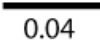
Turtle Genomics

- 3 genome consortia
- Several more independent studies



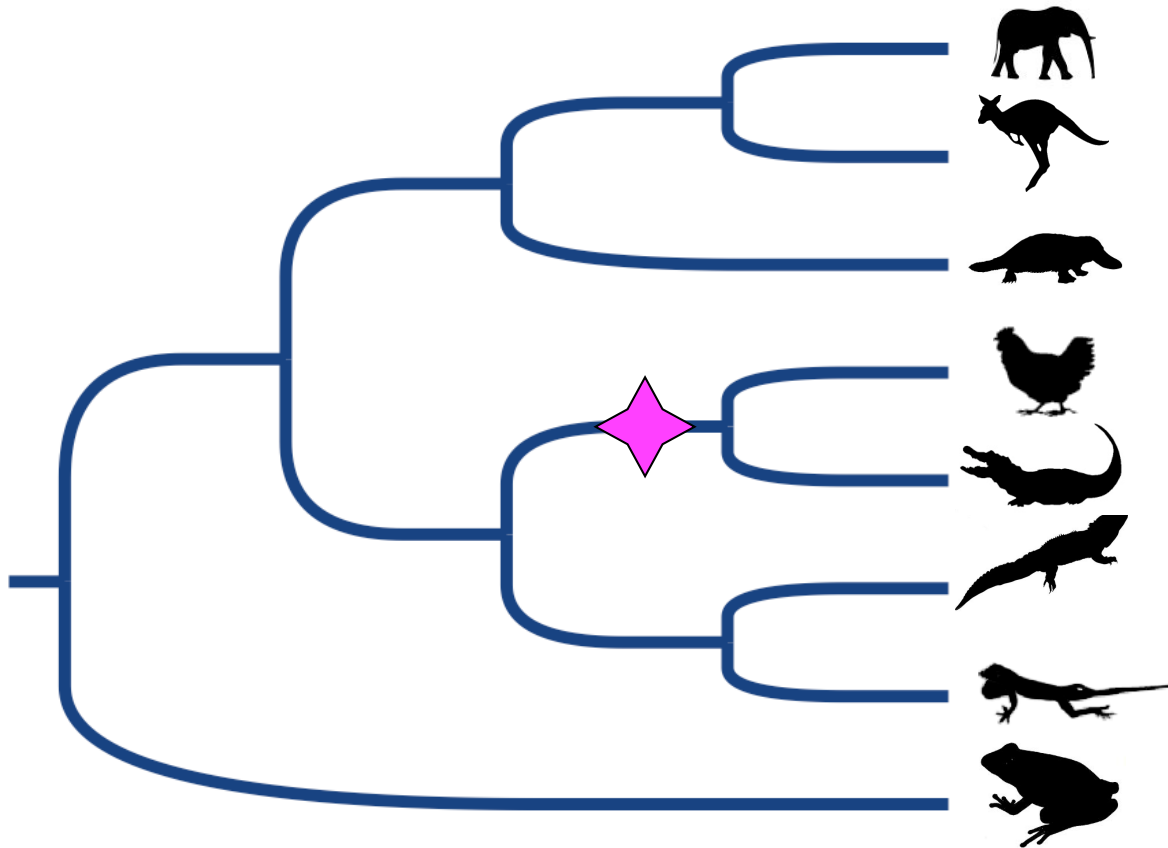
National Human
Genome Research
Institute





Phylogenomics

- All analyses agree!



MicroRNA Result

biology
letters

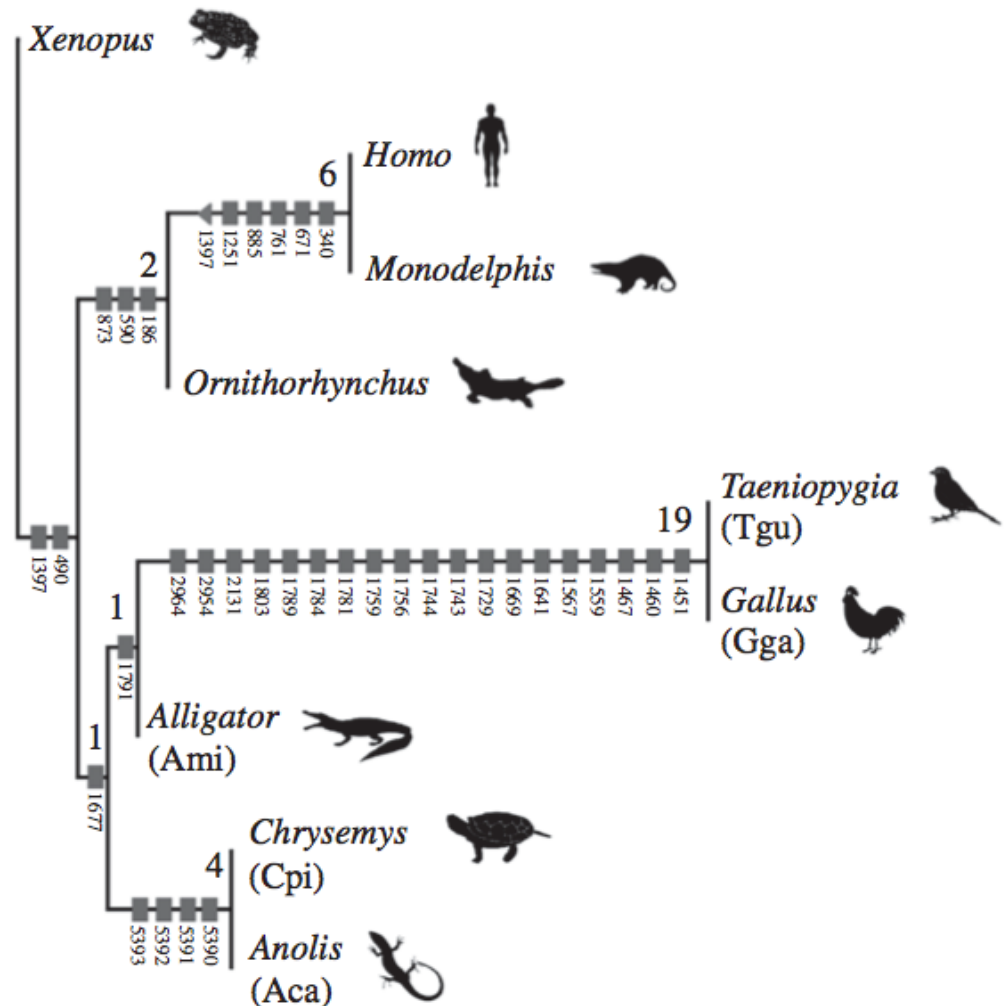
Biol. Lett.

doi:10.1098/rsbl.2011.0477

Published online

Phylogeny

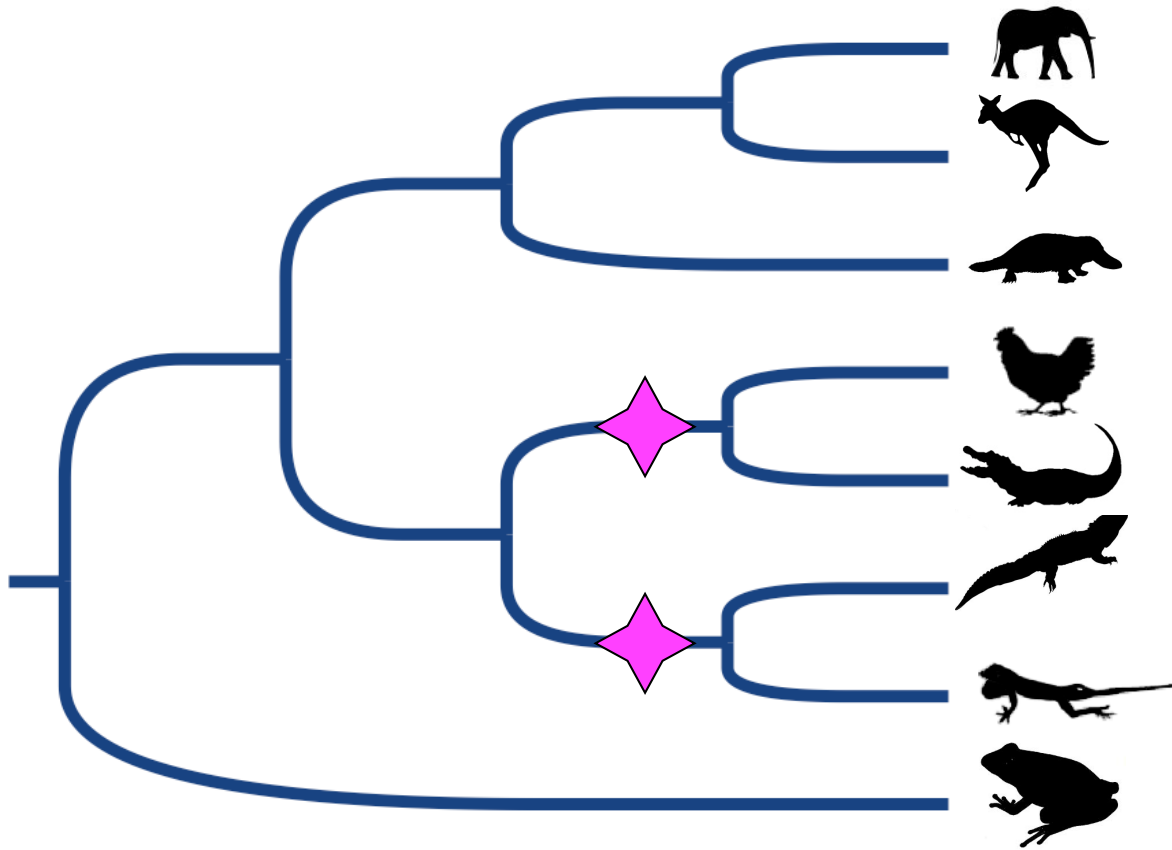
MicroRNAs support a turtle + lizard clade



Lyson et al. 2011 Biol Lett

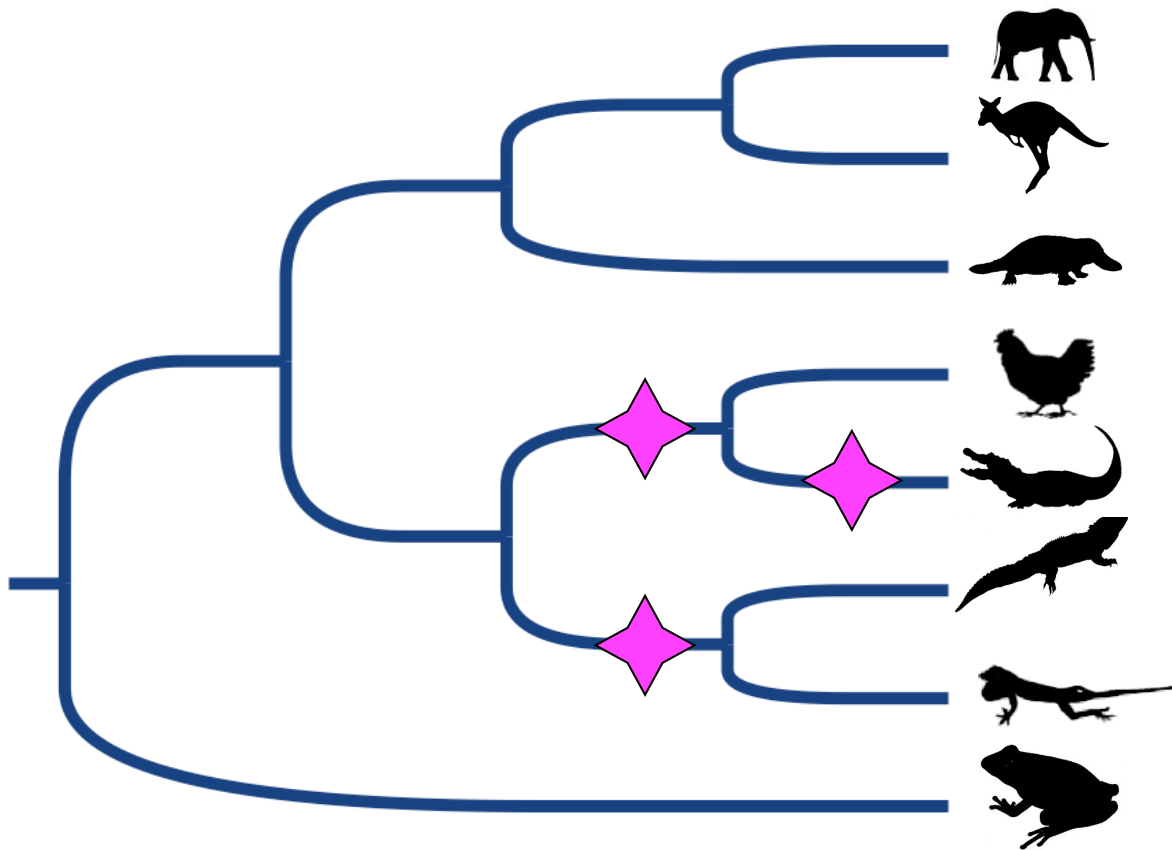
Summary

- Ugh...so what do we do?



Summary

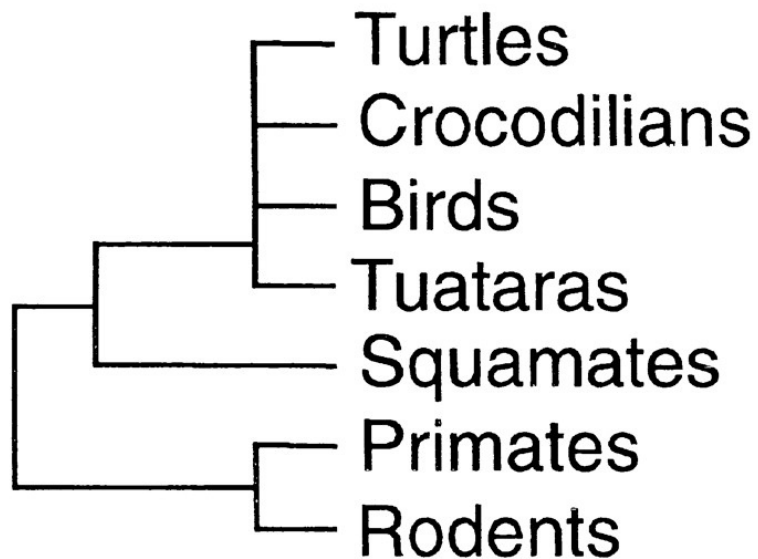
- Ugh...so what do we do?



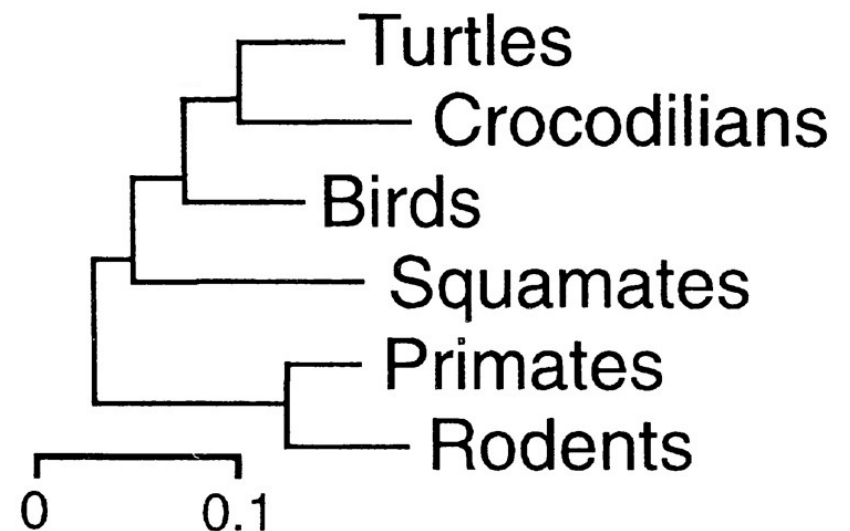
Data in Phylogenetics

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

4 nuclear genes



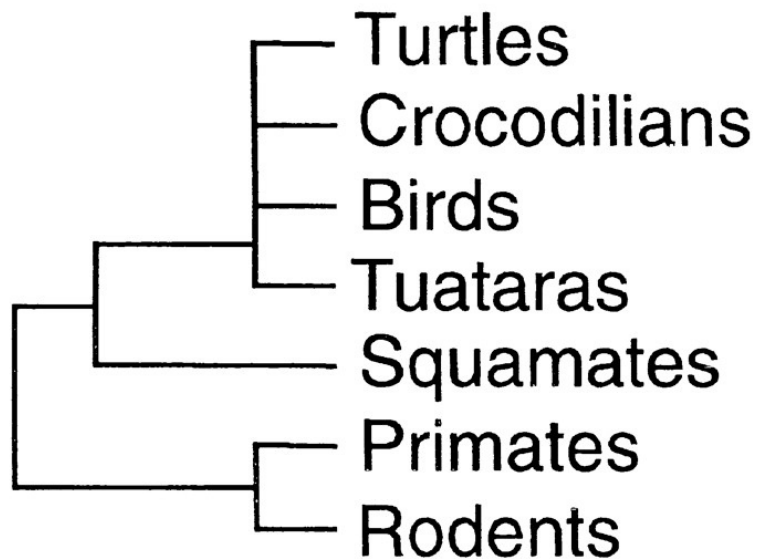
11 nuclear genes



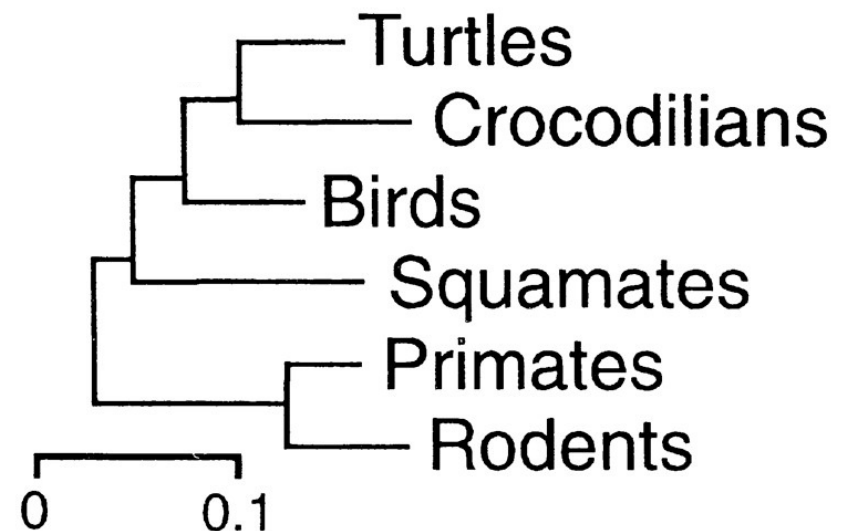
Data in Phylogenetics

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

4 nuclear genes



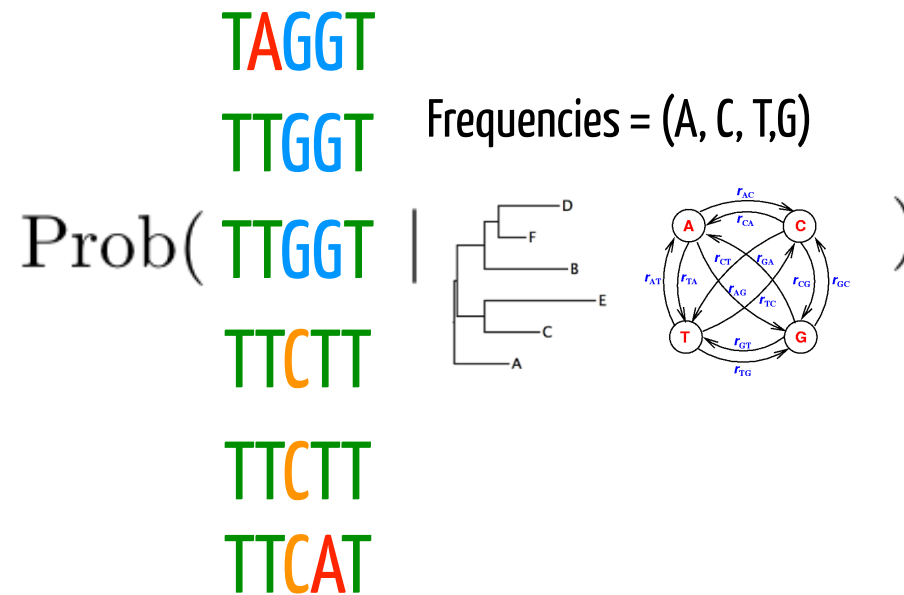
11 nuclear genes



A data centric view

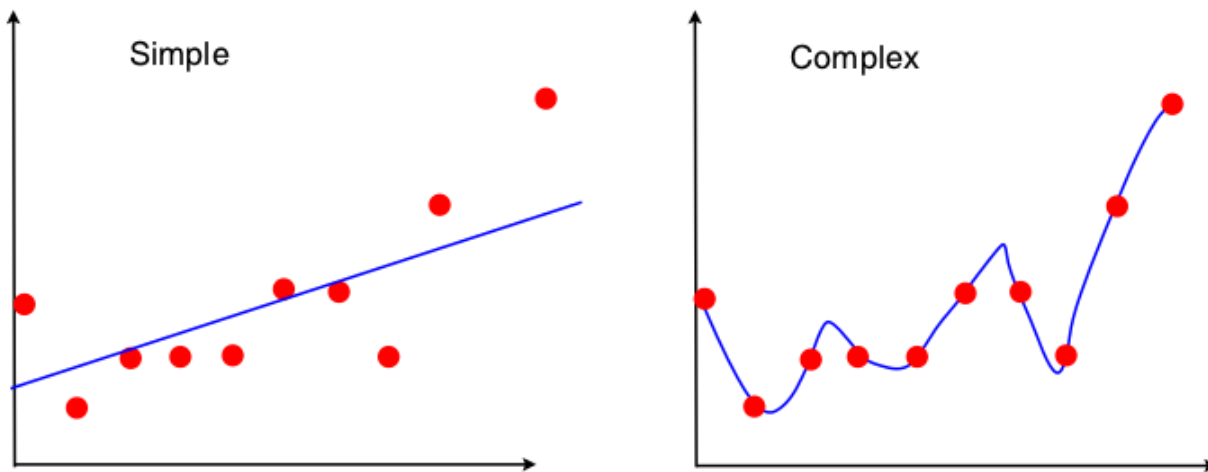
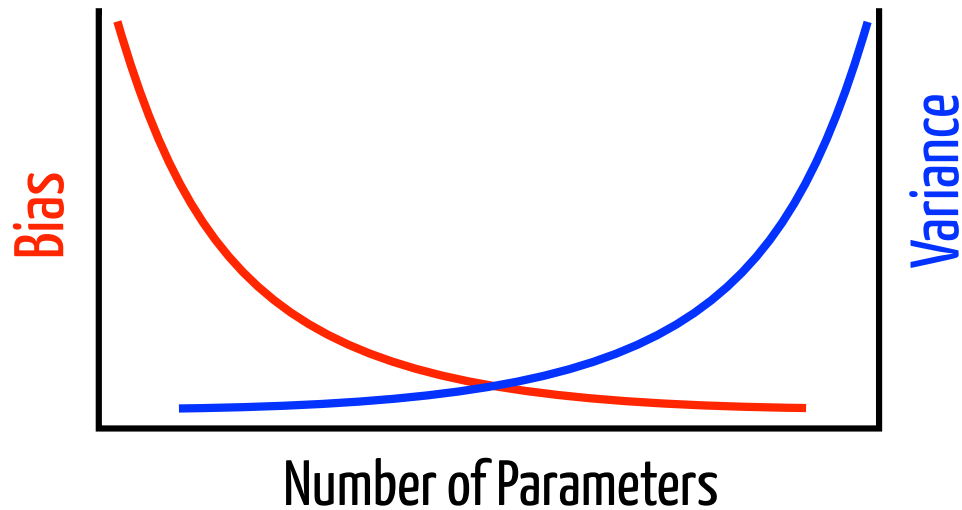
Phylogenomics

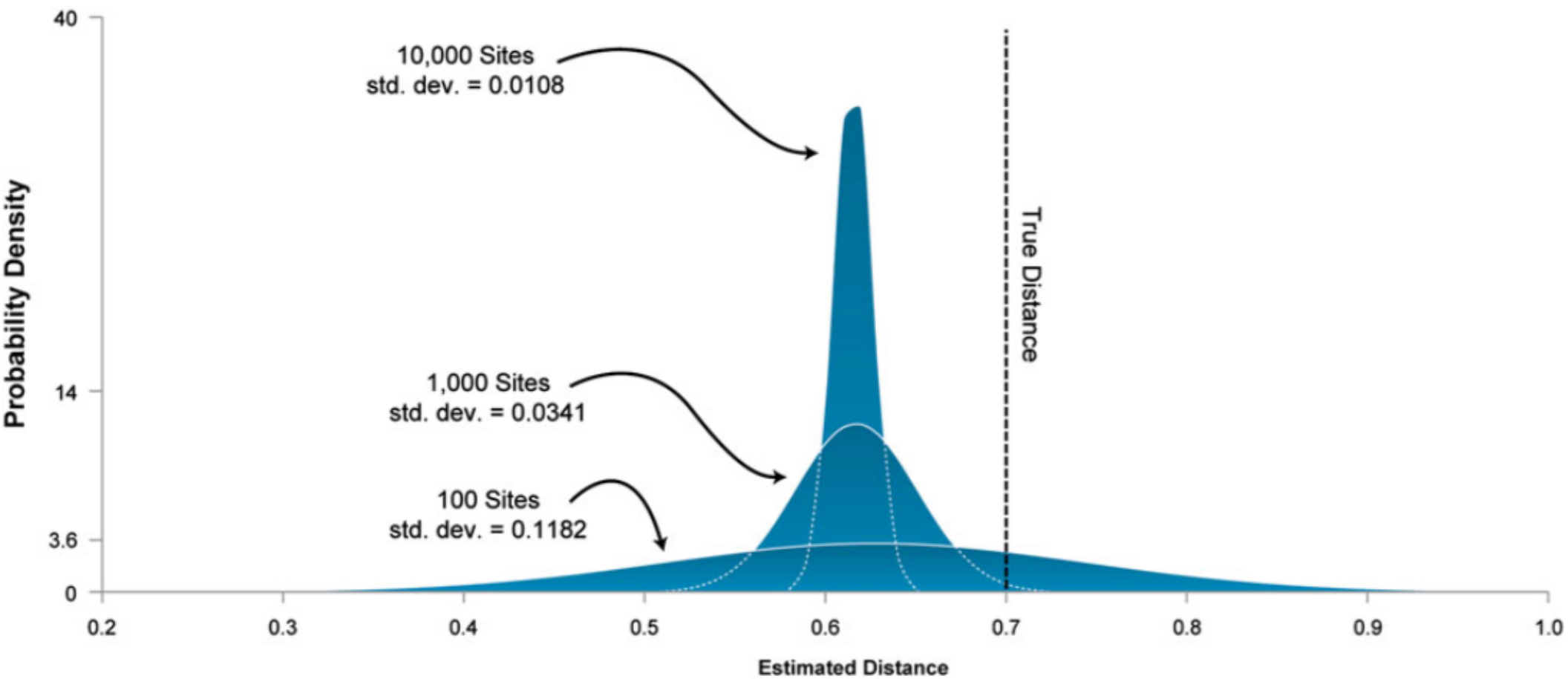
- Inferences result from both data and the model



Why does this matter?

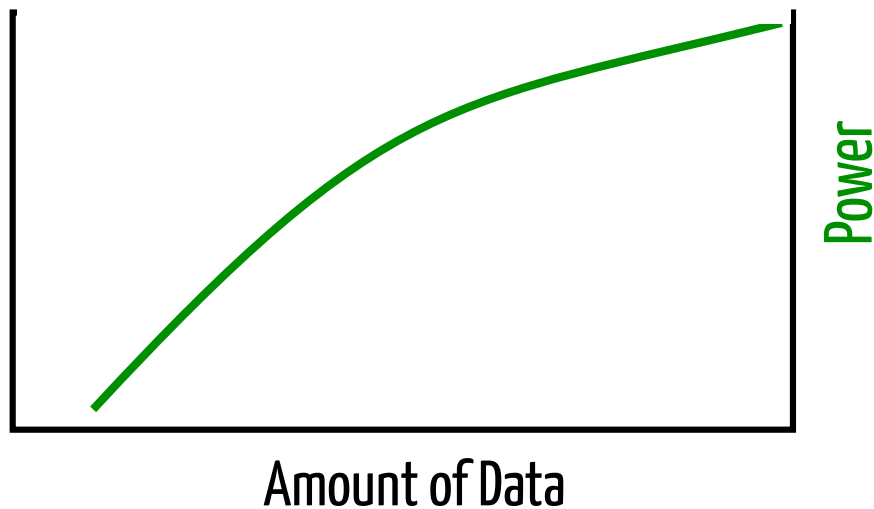
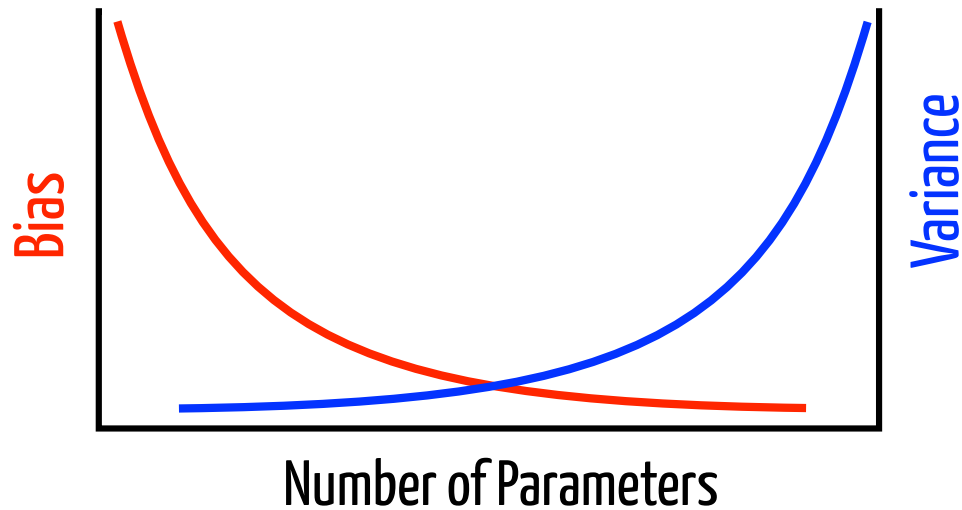
- In developing a statistical model for a problem, we inevitably make a tradeoff





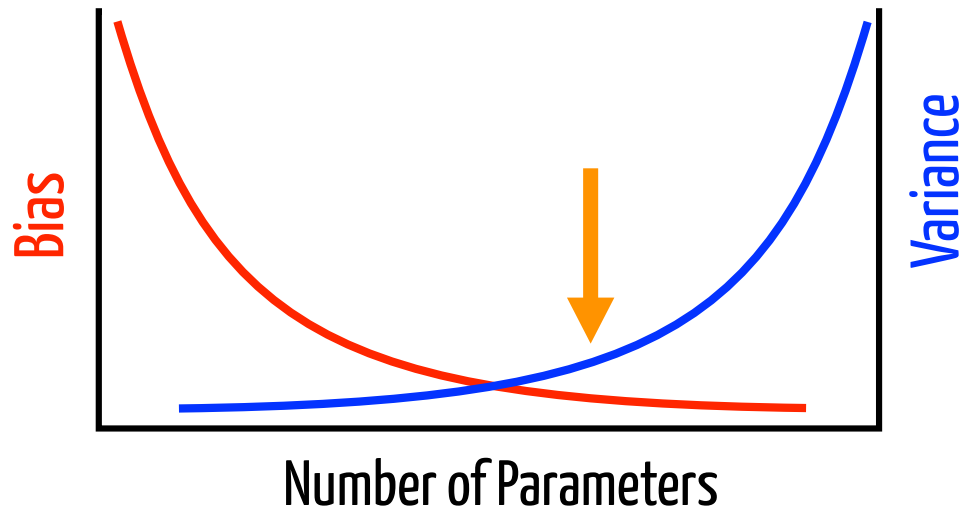
Why does this matter?

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



Why does this matter?

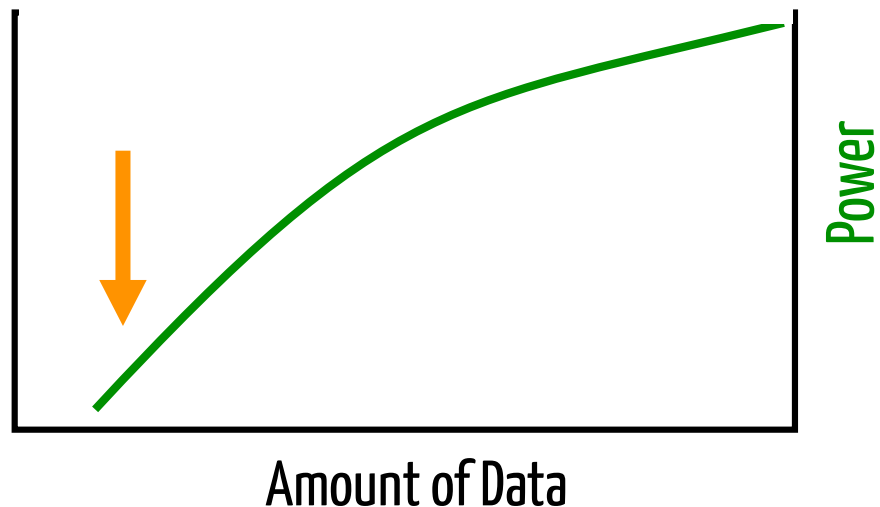
- In developing a statistical model for a problem, we inevitably make a tradeoff



1 gene

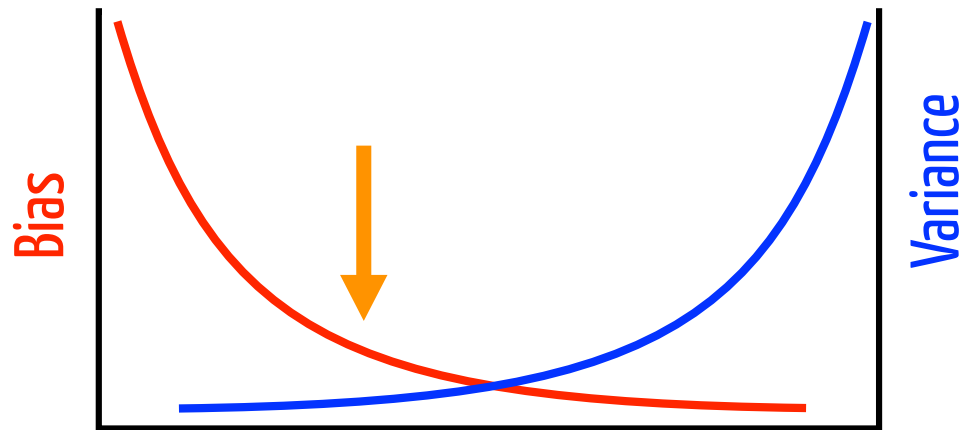
10 genes

1000 genes



Why does this matter?

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

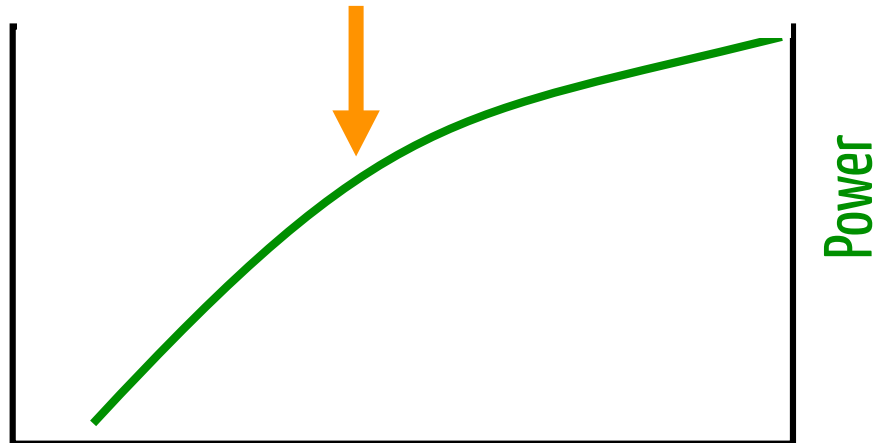


Number of Parameters

1 gene

10 genes

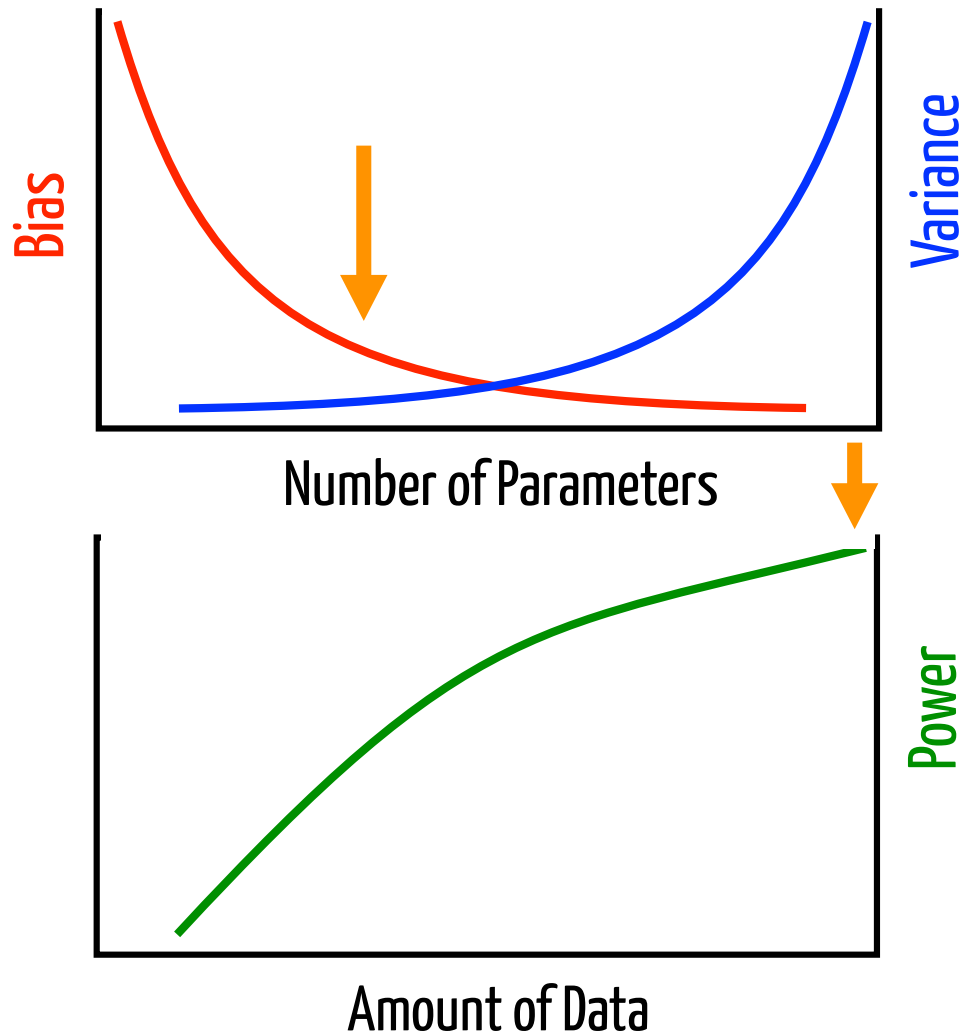
1000 genes



Amount of Data

Why does this matter?

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



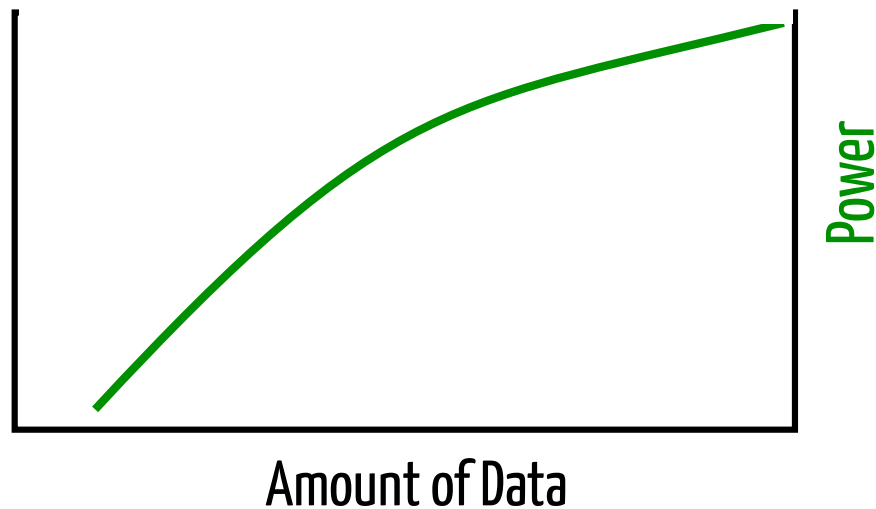
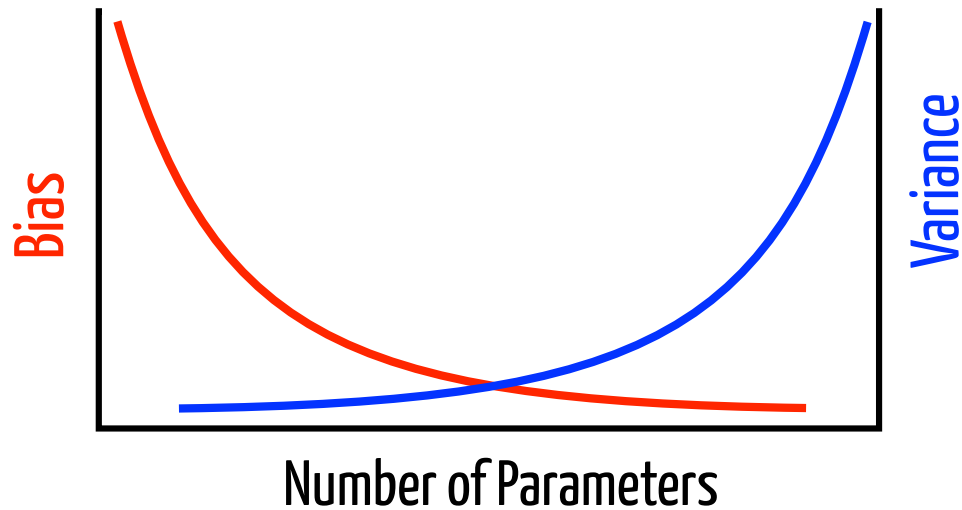
1 gene

10 genes

1000 genes

Why does this matter?

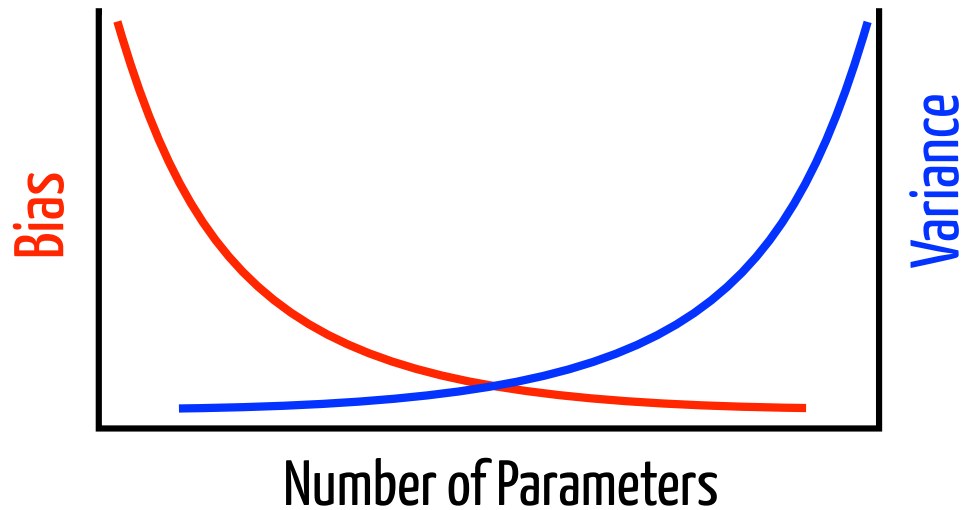
- The point.



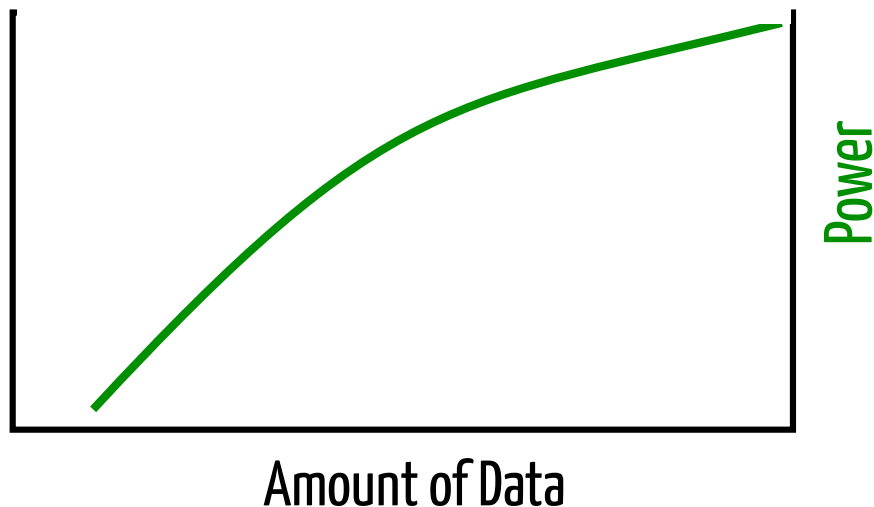
Our data centric
view focuses on this

Why does this matter?

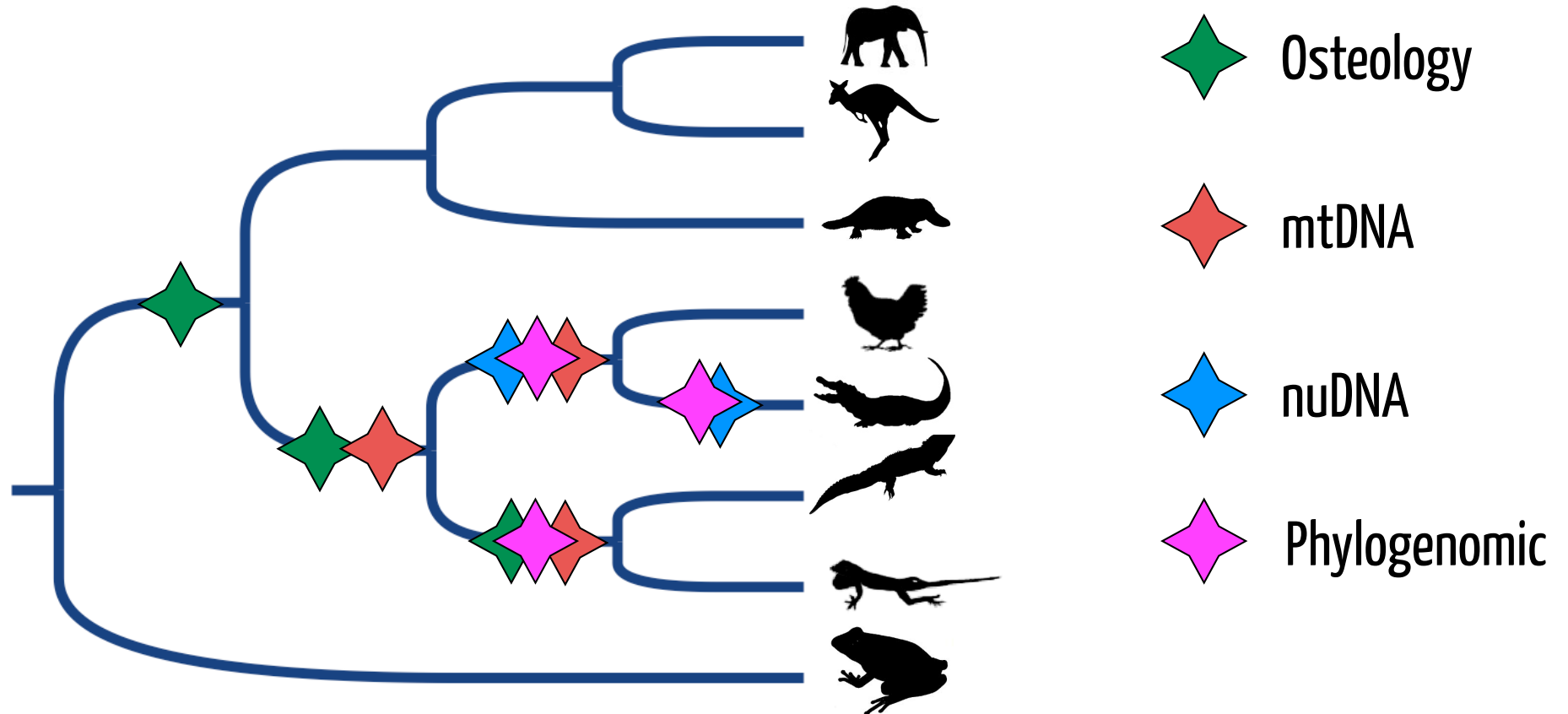
- The point.



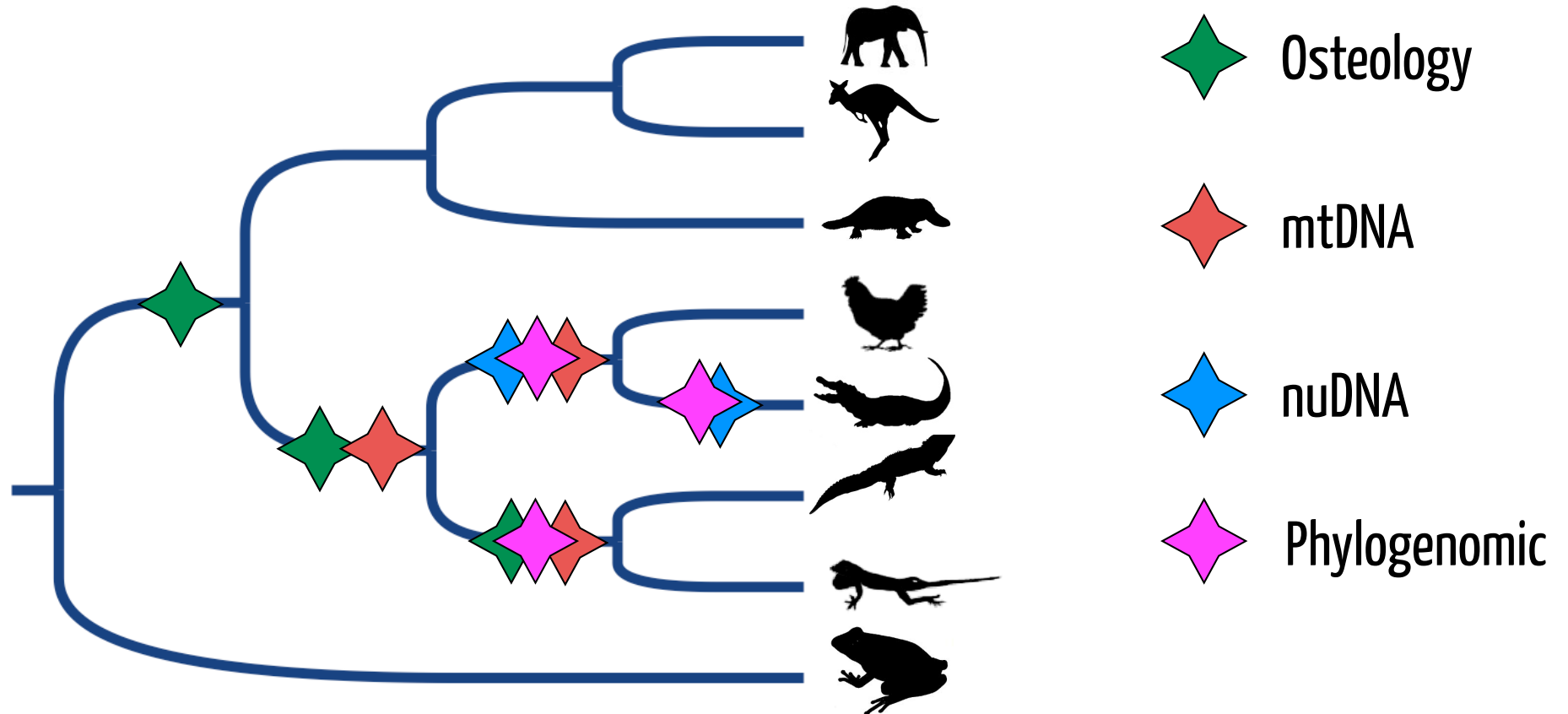
But this is our
bigger problem



How do we know it's a bigger problem?

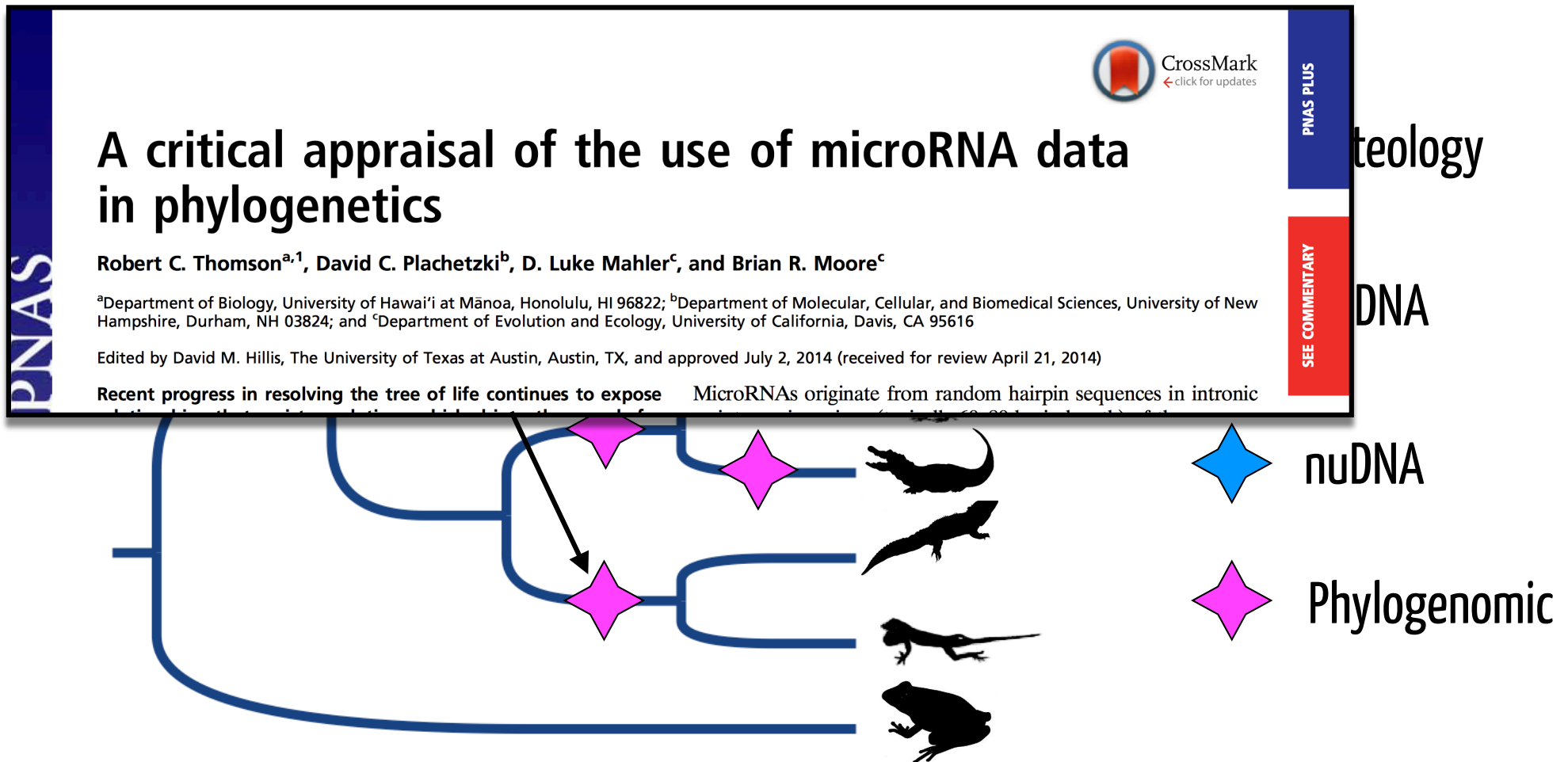


How do we know it's a bigger problem?



Where's the disagreement coming from?

How do we know it's a bigger problem?

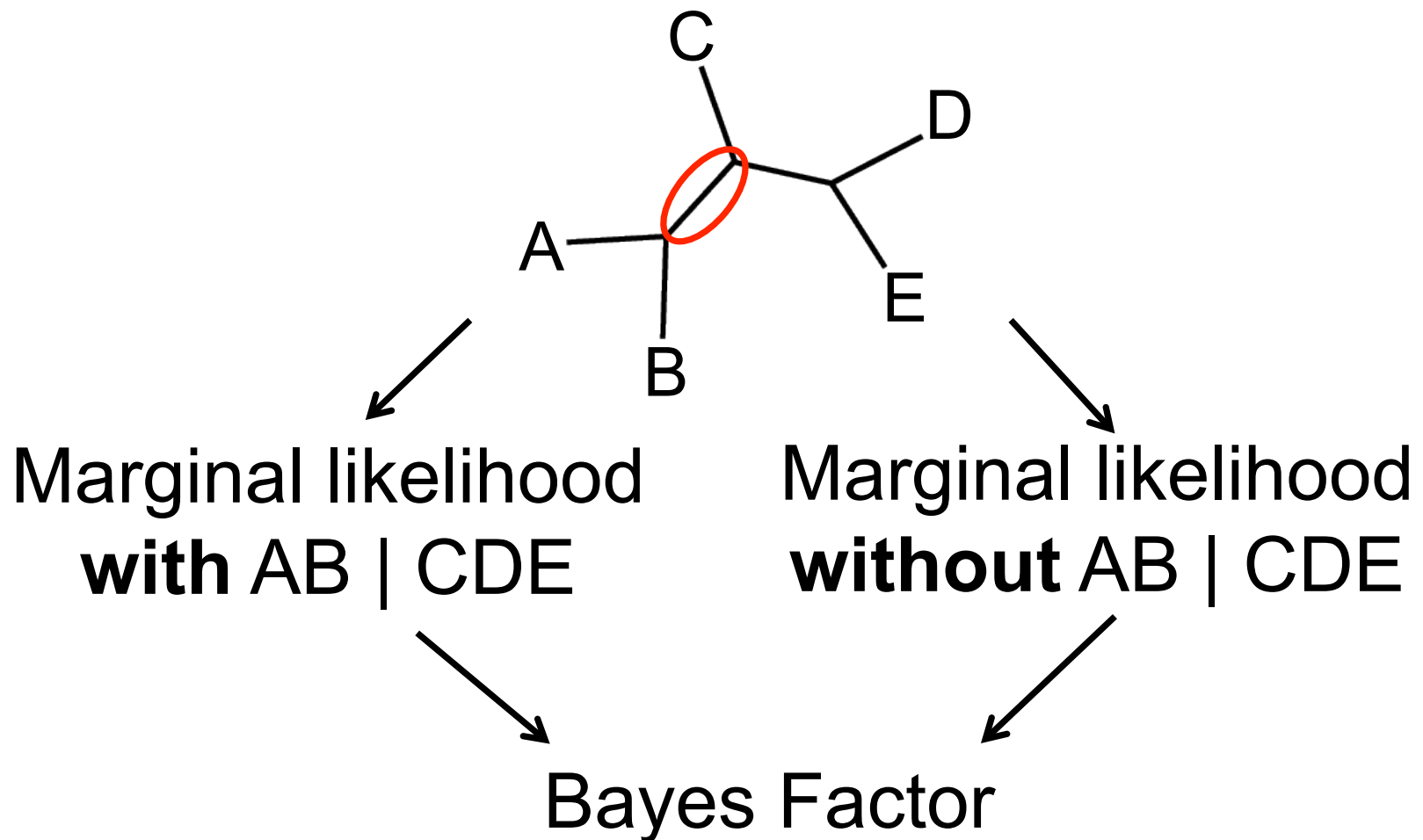


Where's the disagreement coming from?

'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

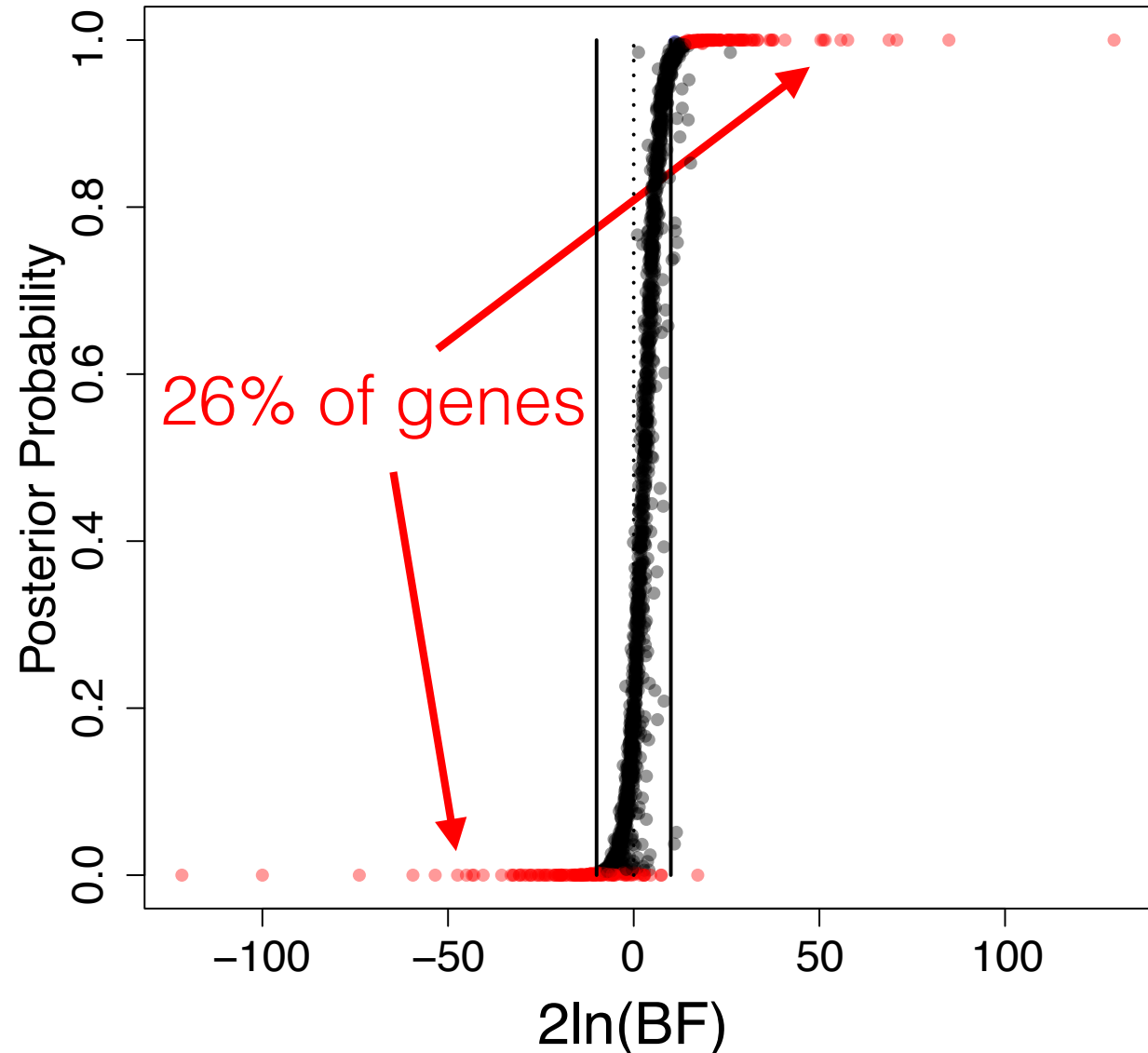
Bipartition Bayes Factors



A Note on Extreme Probabilities

Archosaur +
Turtle
Monophyly

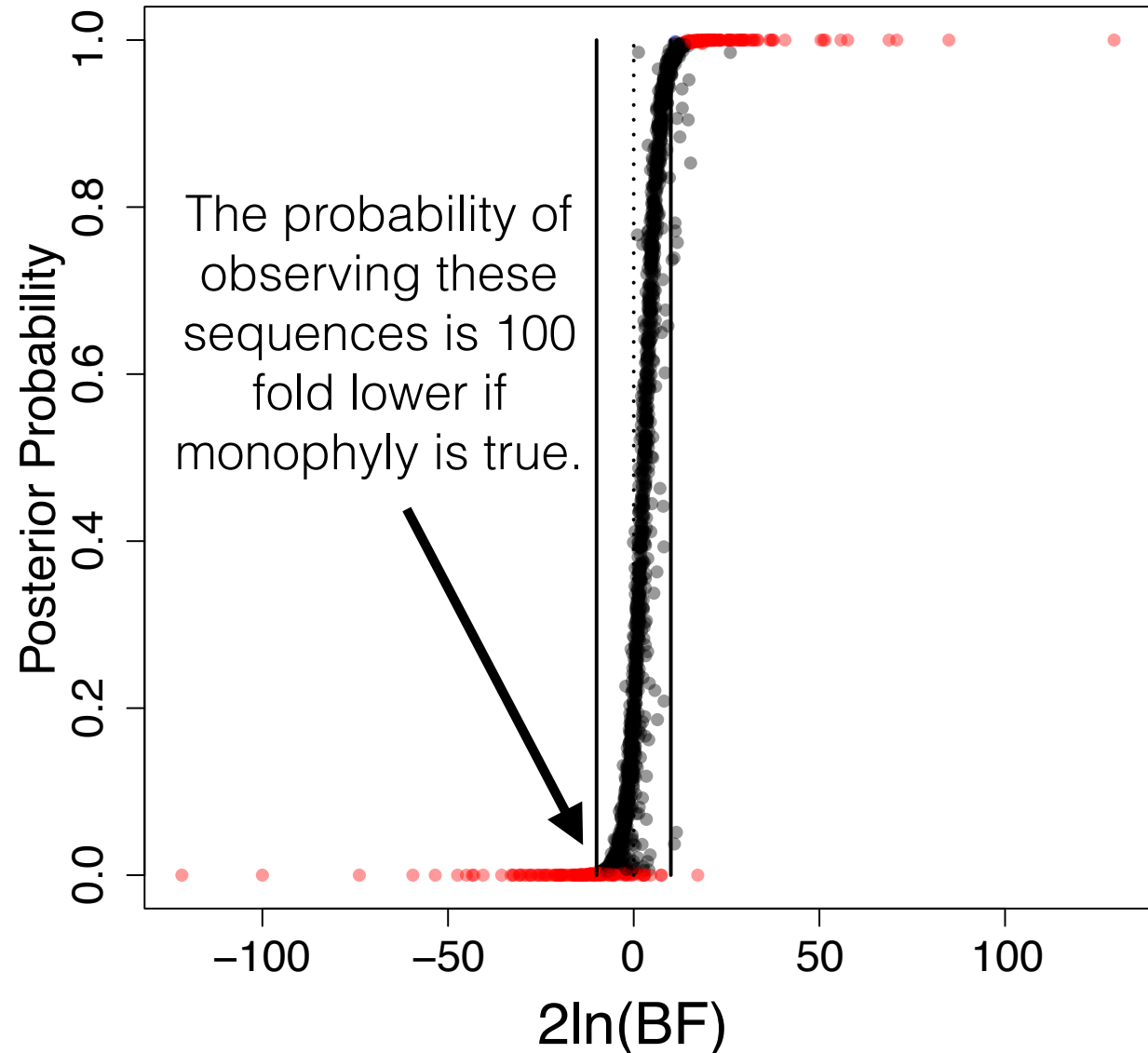
Shaffer et al.



A Note on Extreme Probabilities

Archosaur +
Turtle
Monophyly

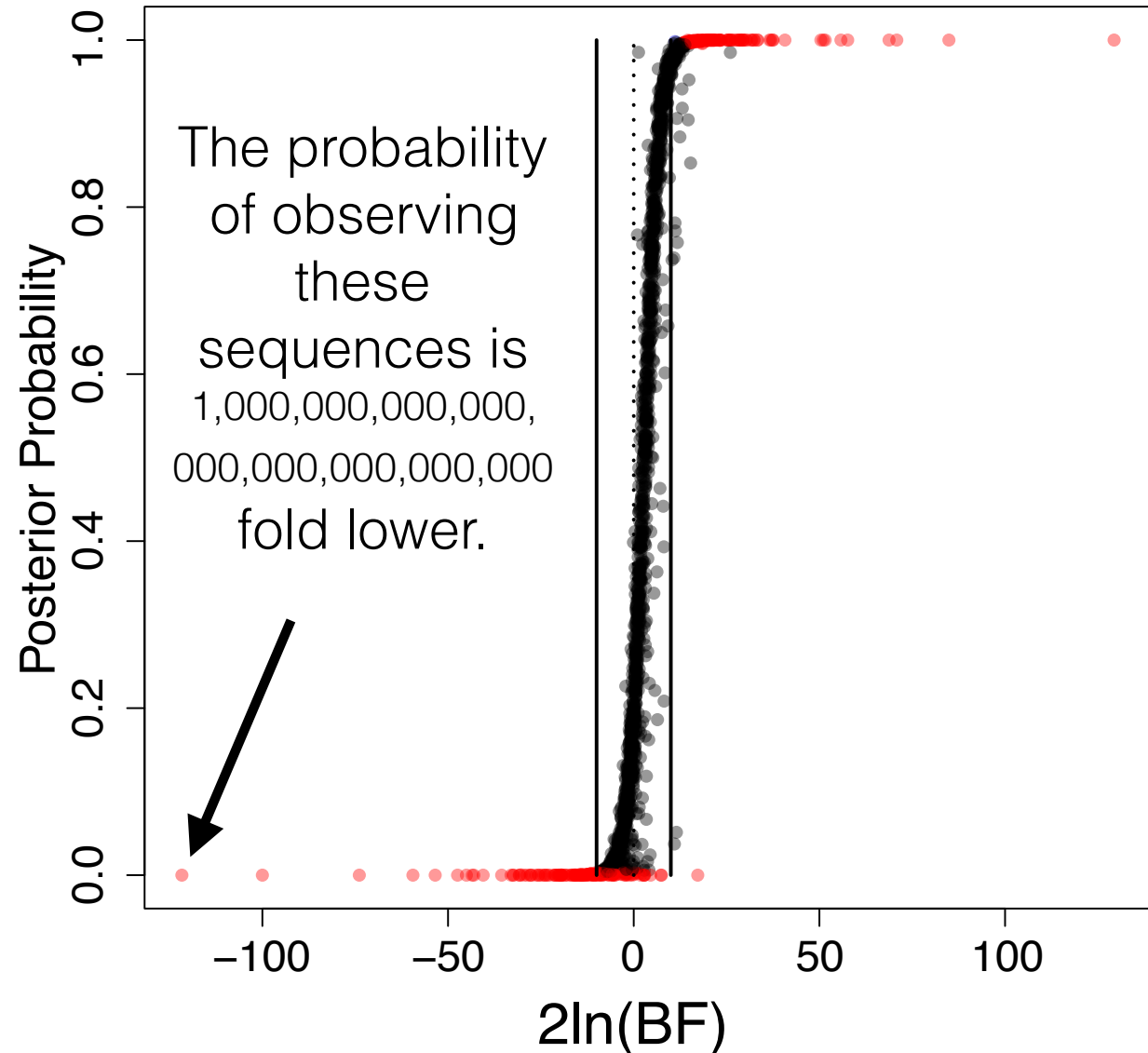
Shaffer et al.



A Note on Extreme Probabilities

Archosaur +
Turtle
Monophyly

Shaffer et al.



A Note on Extreme Probabilities

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

That's 27 zeroes!

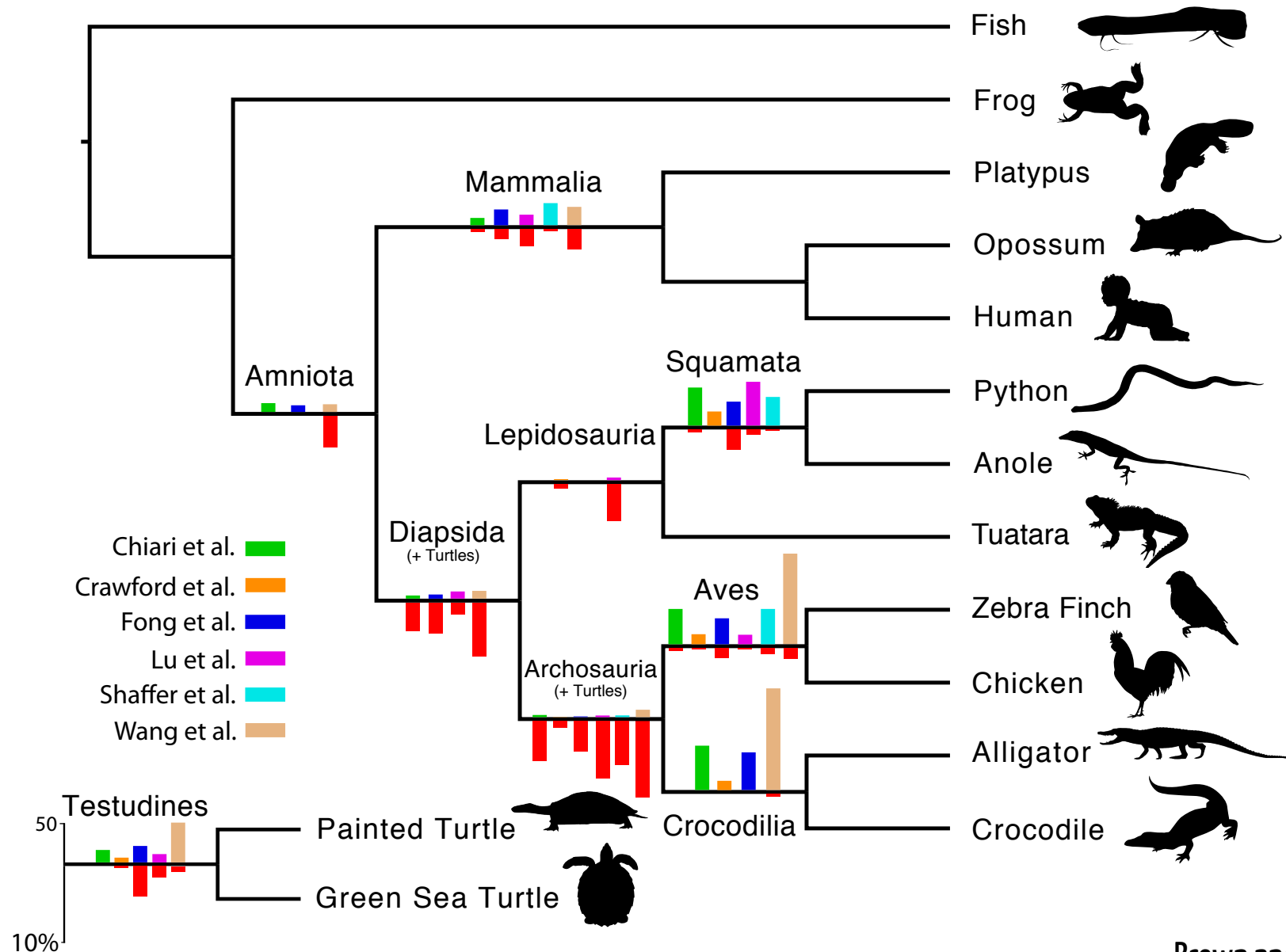
If you played a lottery every minute with that chance of winning, you still probably wouldn't win, unless you played for...

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

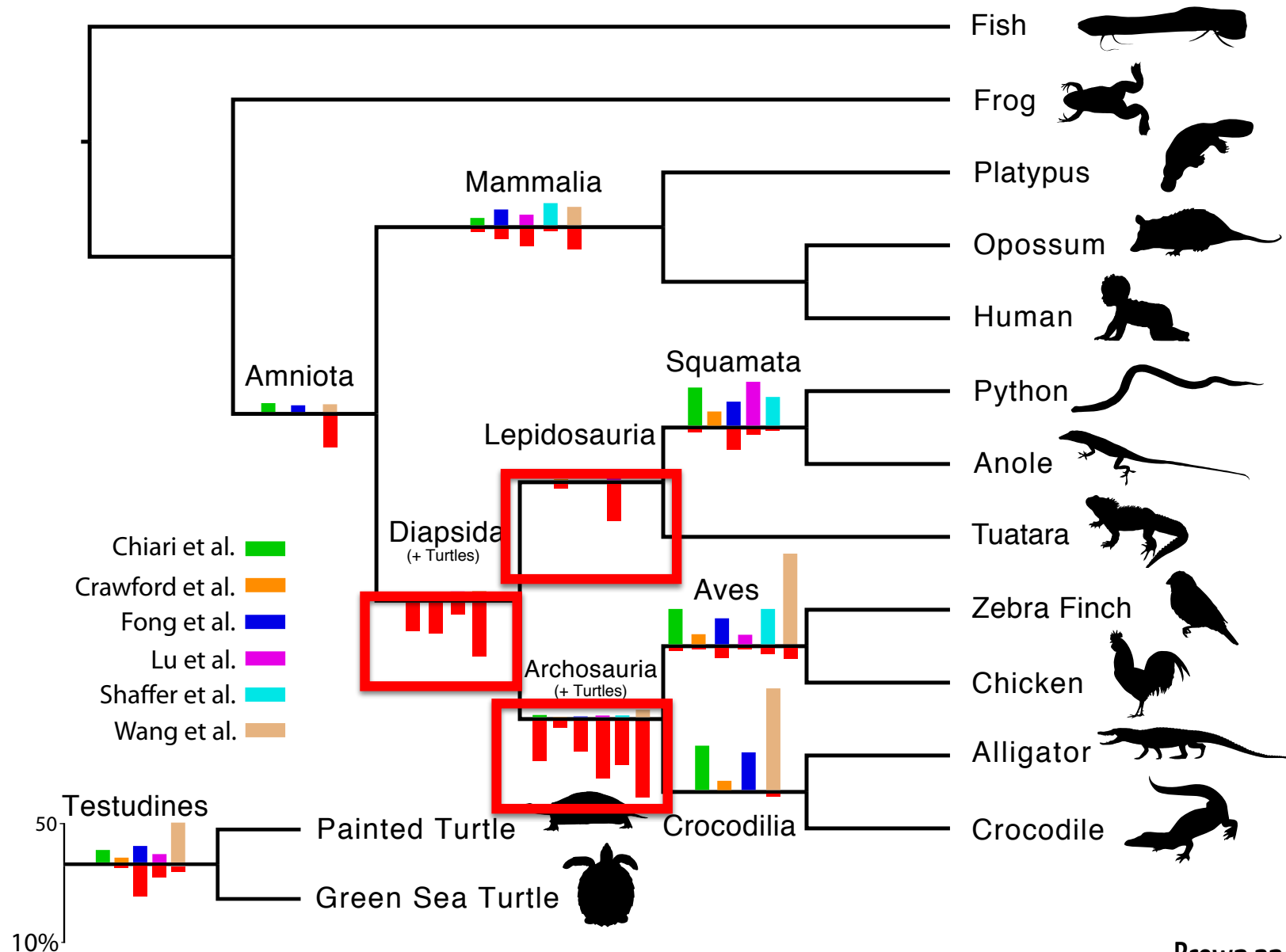
Shaffer et al.



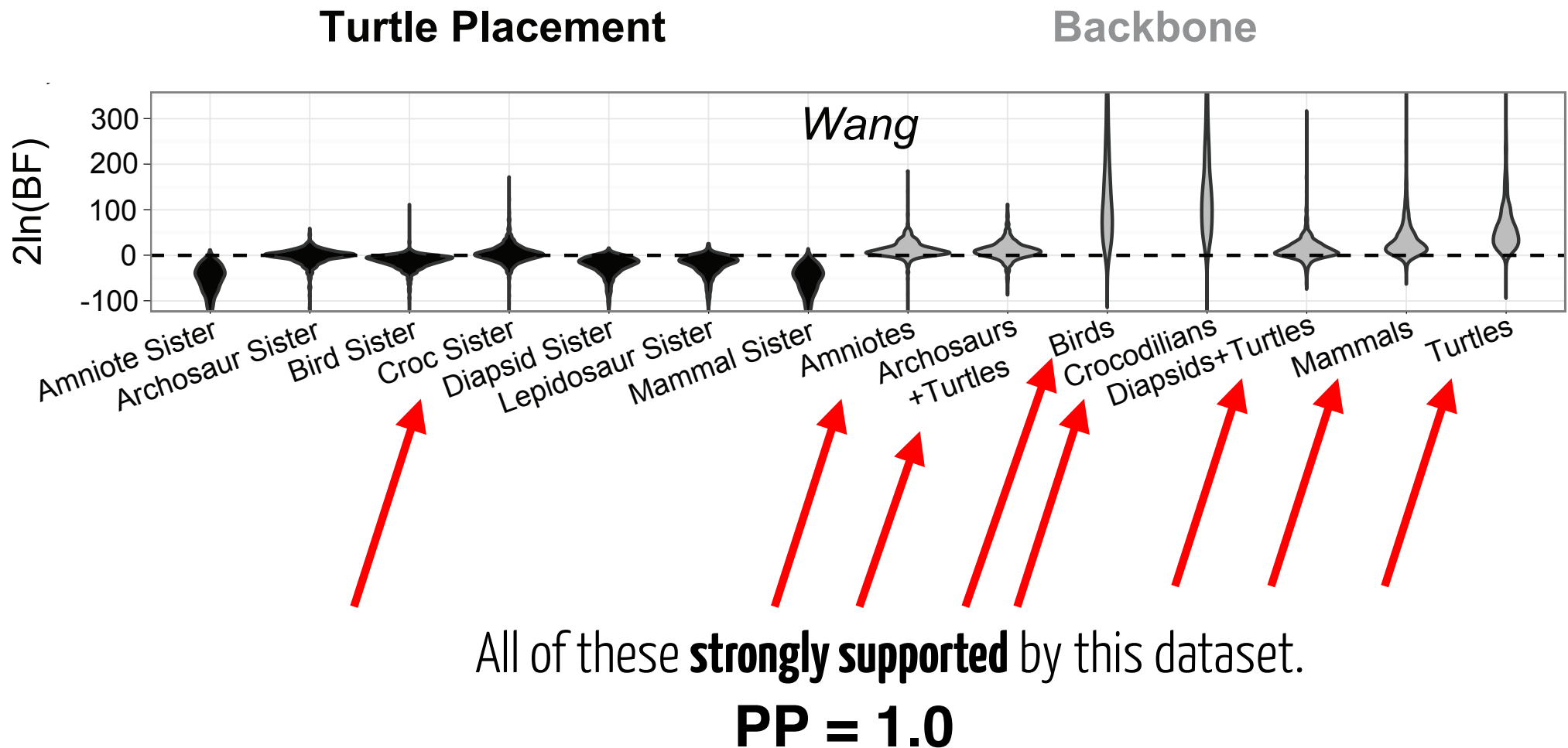
Support varies across branches of the tree



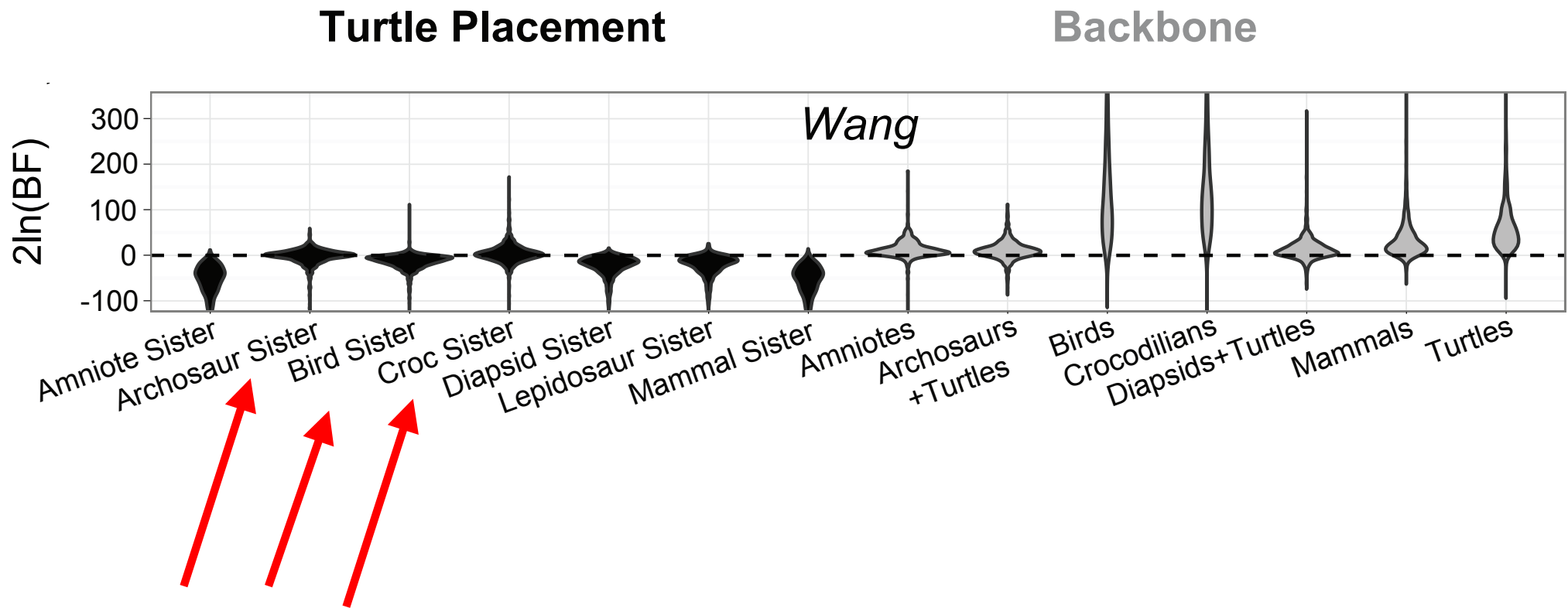
Support varies across branches of the tree



Support varies across branches of the tree

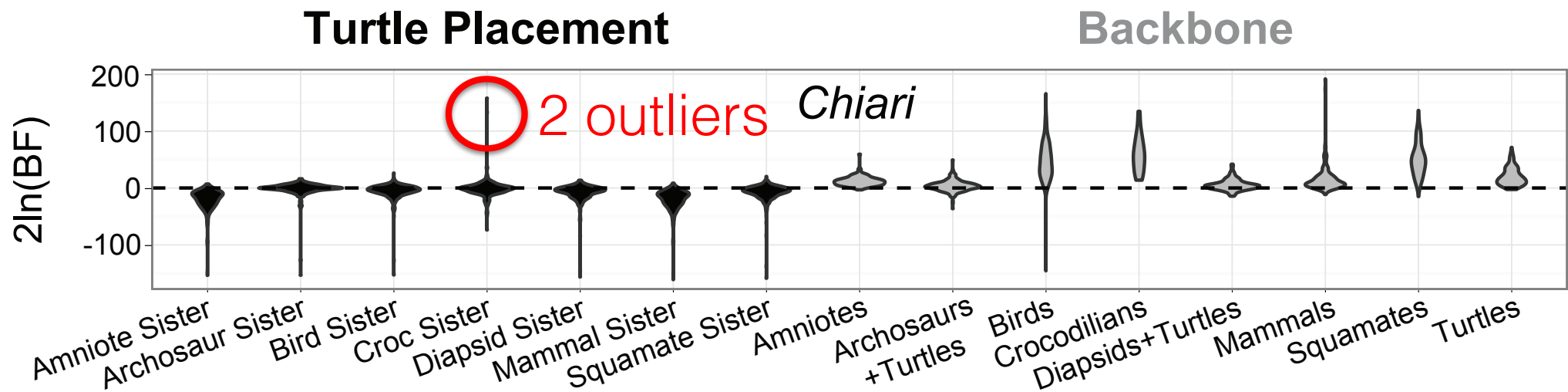


Support varies across branches of the tree



Equivocation about turtle placement **across genes**

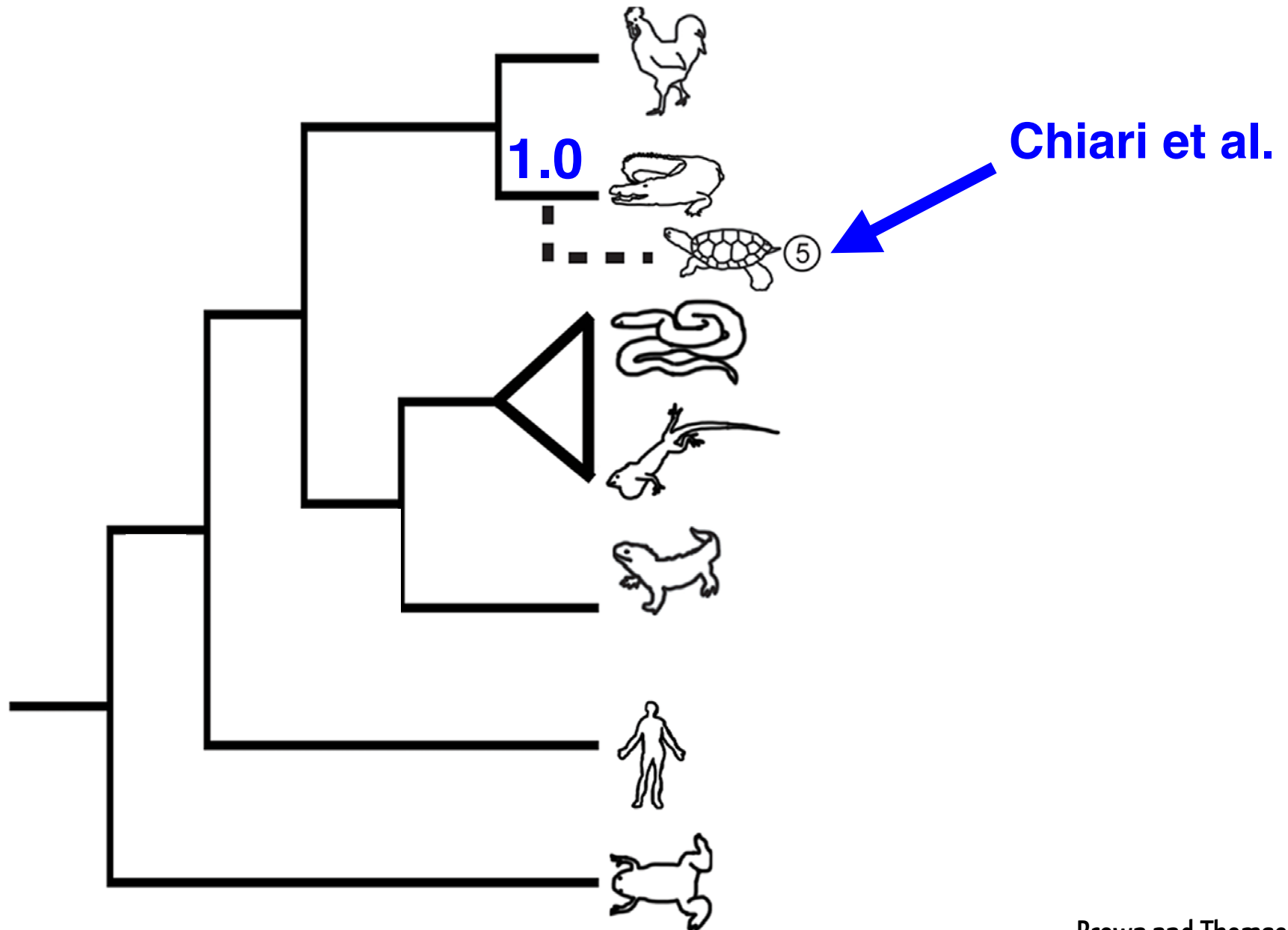
Support varies across branches of the tree



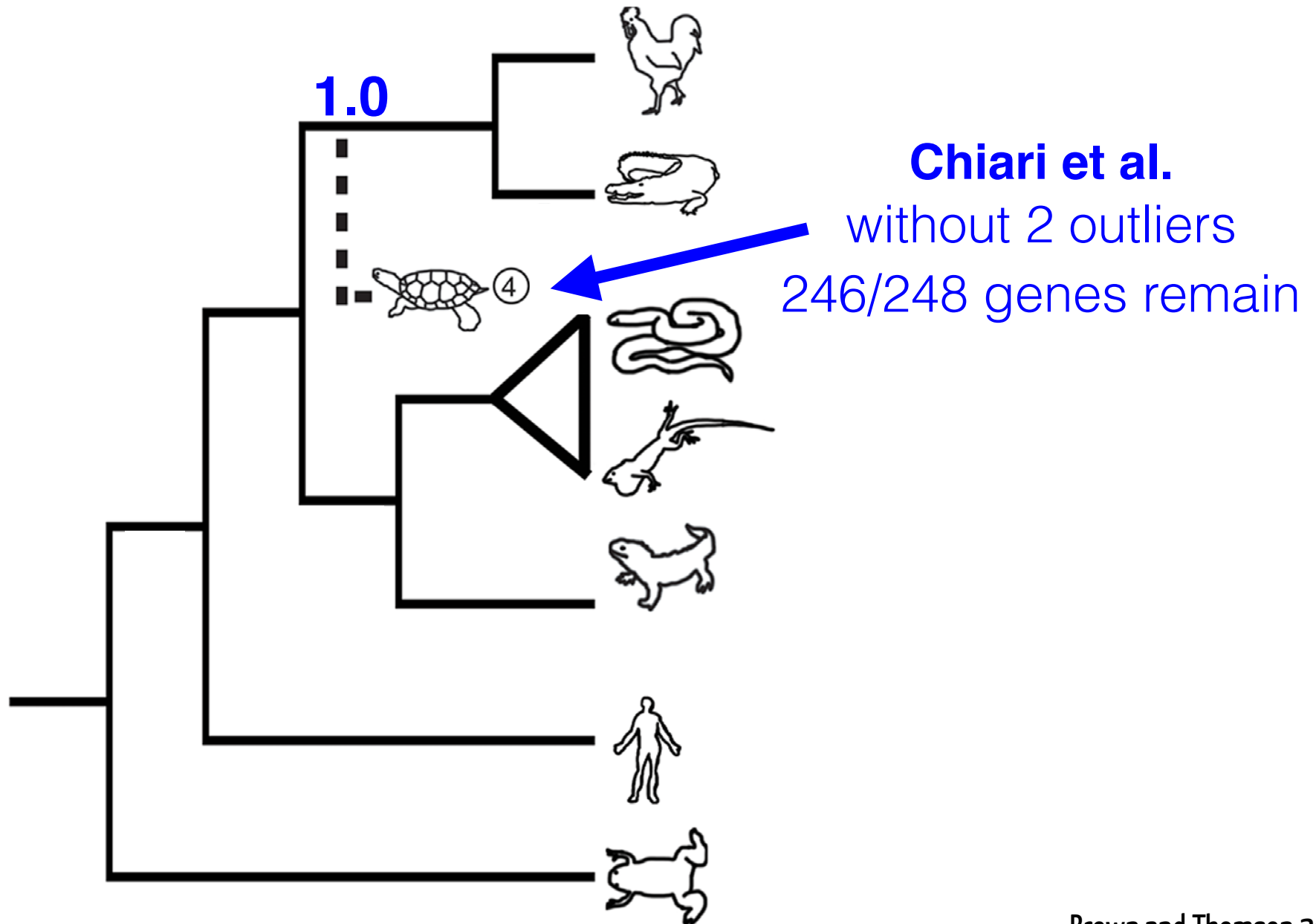
This dataset supports turtles as sister to crocodilians.
But what's up with these outliers?
How influential are they?



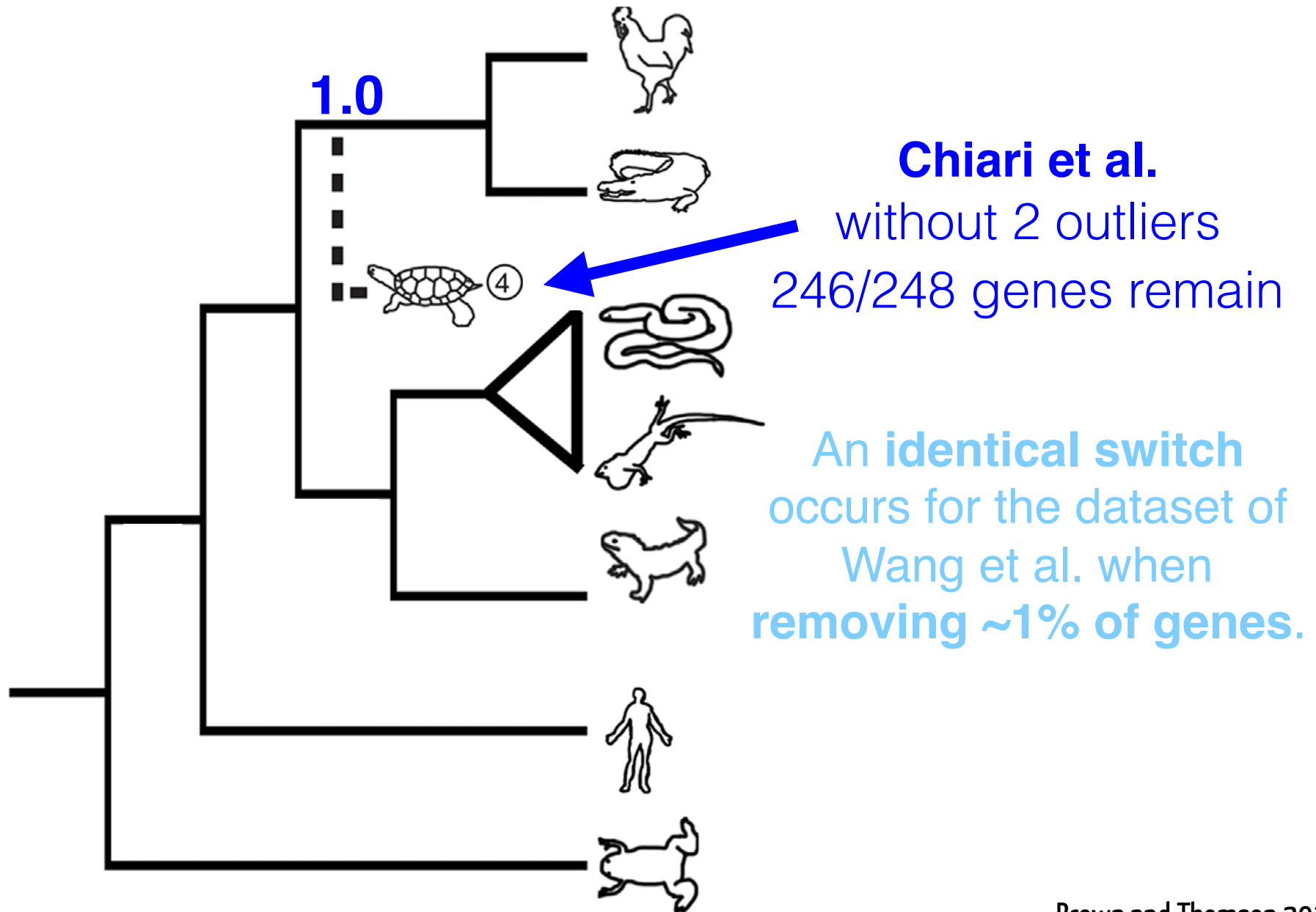
Strong influence



Strong influence



Strong influence



A troubling, but common, result

- More recent papers build on this result and find similar patterns:

New Results

Site and gene-wise likelihoods unmask influential outliers in phylogenomic analyses

Joseph F. Walker, Joseph W. Brown, Stephen A. Smith

doi: <https://doi.org/10.1101/115774>

Article

Contentious relationships in phylogenomic studies can be driven by a handful of genes

Xing-Xing Shen, Chris Todd Hittinger & Antonis Rokas 

Inadvertent Paralog Inclusion Drives Artifactual Topologies and Timetree Estimates in Phylogenomics

Karen Siu-Ting,^{*,1,2,3} María Torres-Sánchez,^{‡,4} Diego San Mauro,⁴ David Wilcockson,⁵ Mark Wilkinson,⁶ Davide Pisani,⁷ Mary J. O'Connell,^{8,9} and Christopher J. Creevey^{*,1}

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 collecting more data per se

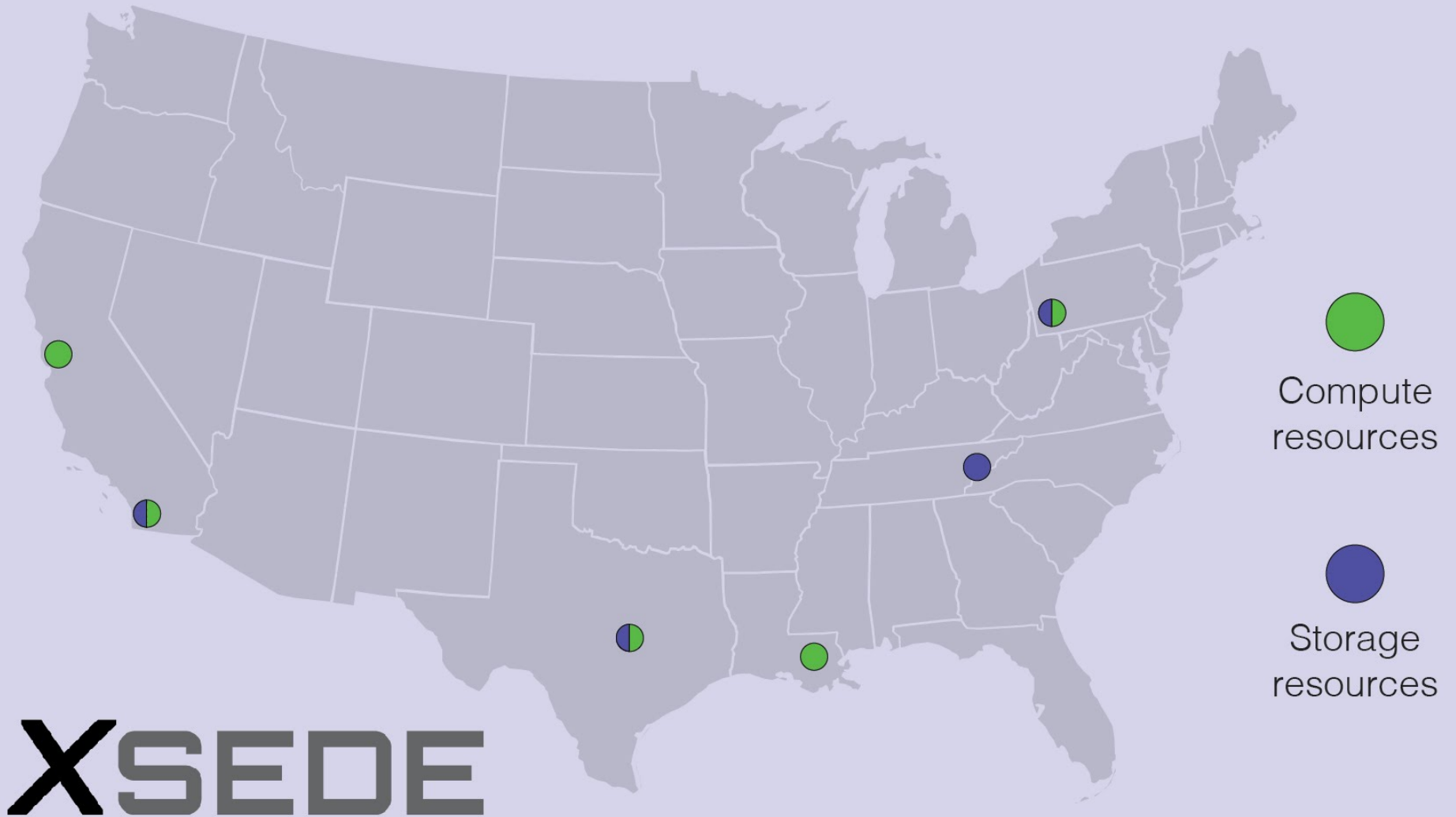
Some Possible Ways Forward

- **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
 - continual methodological improvement










Embrace the computation



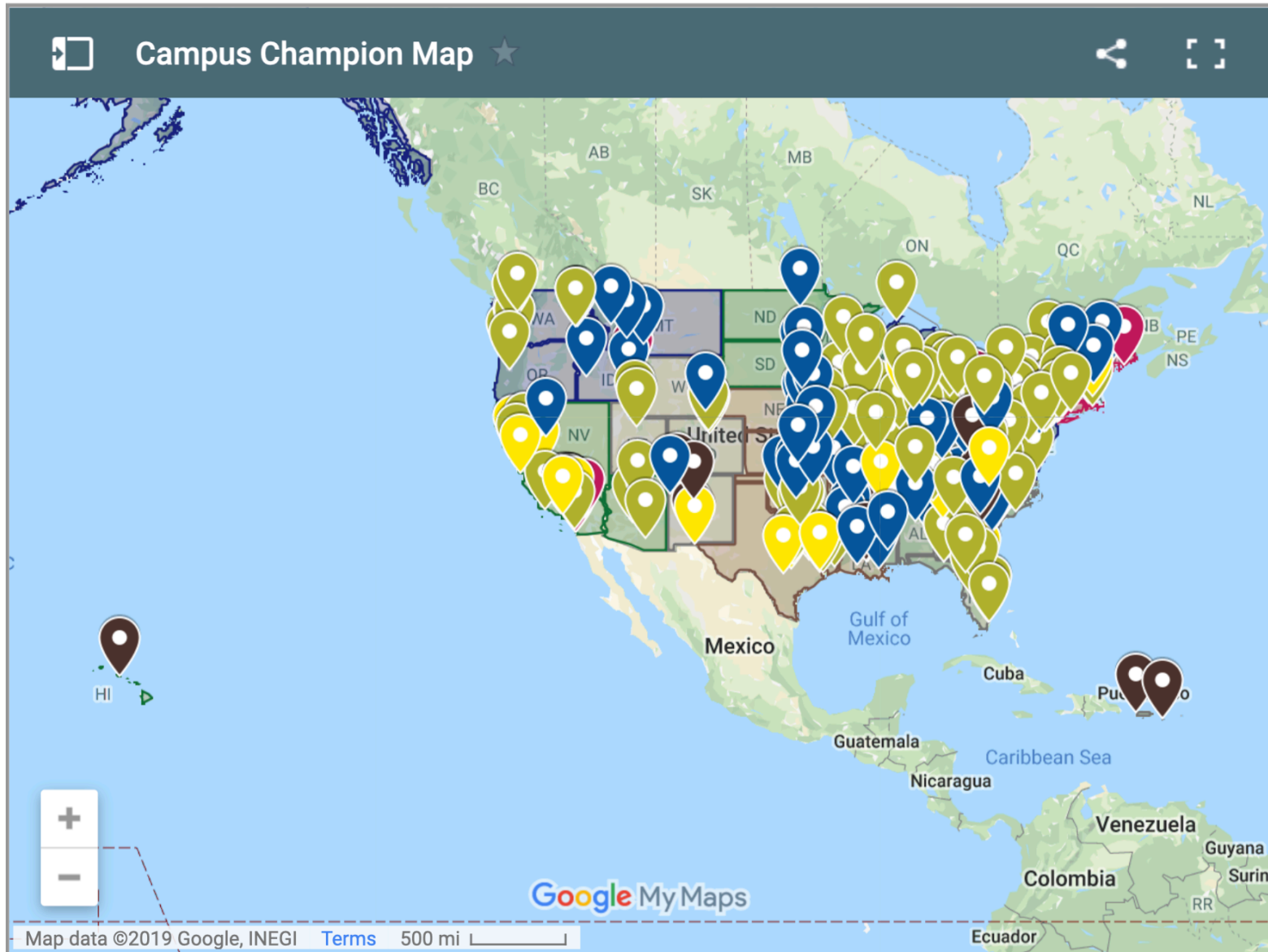
Embrace the computation

⚙️ Compute Resources

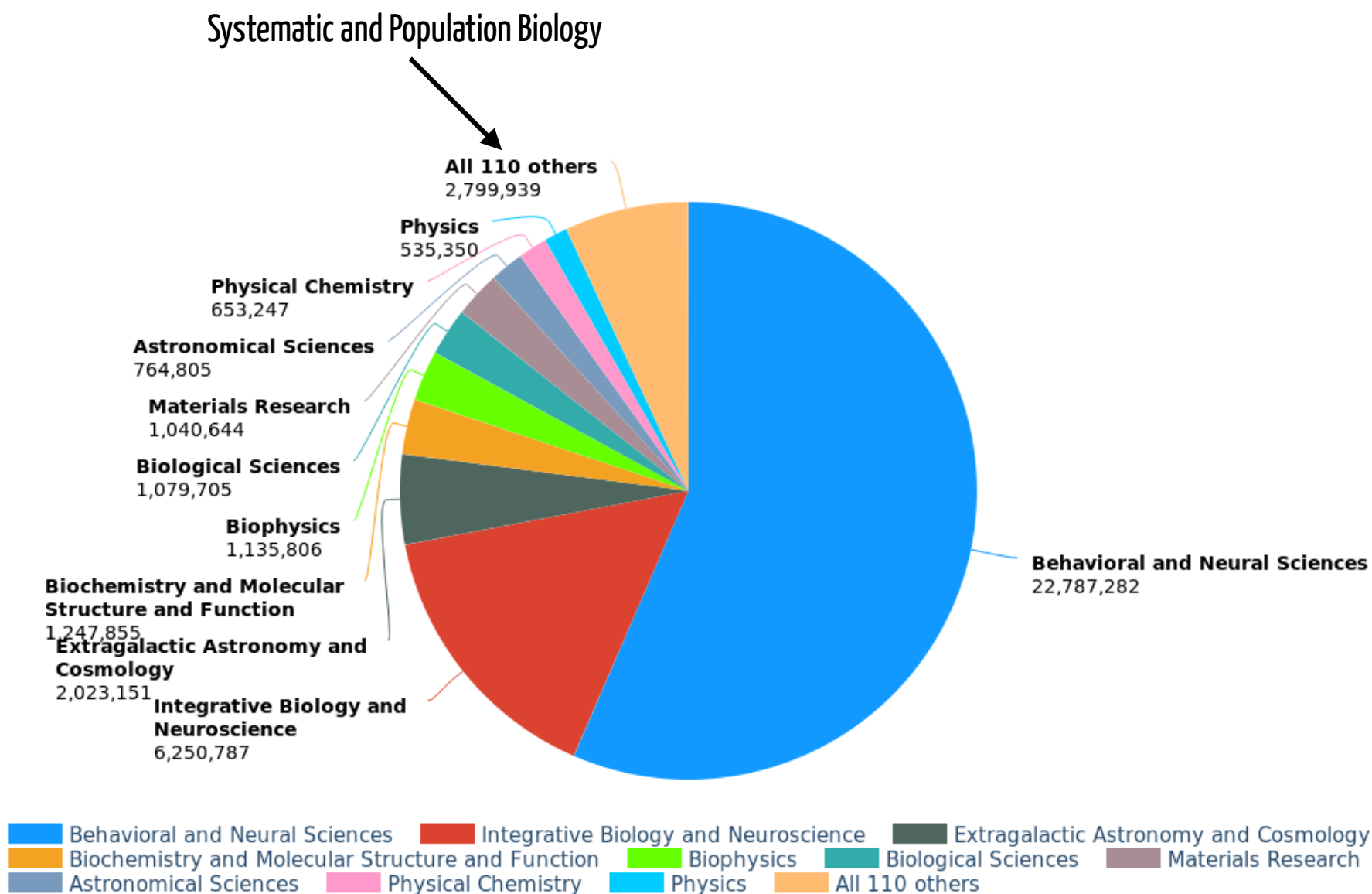


Name	Status	CPU	Peak TFlops	Utilization	Running Jobs	Queued Jobs	Other Jobs
Stampede2  User Guide 	✓ Healthy	368280	12800.0		1006	900	381
Comet  User Guide 	✓ Healthy	46752	2000.0		1866	16	82
SuperMIC  User Guide 	✓ Healthy	7200	925.0		56	0	0

Embrace the computation

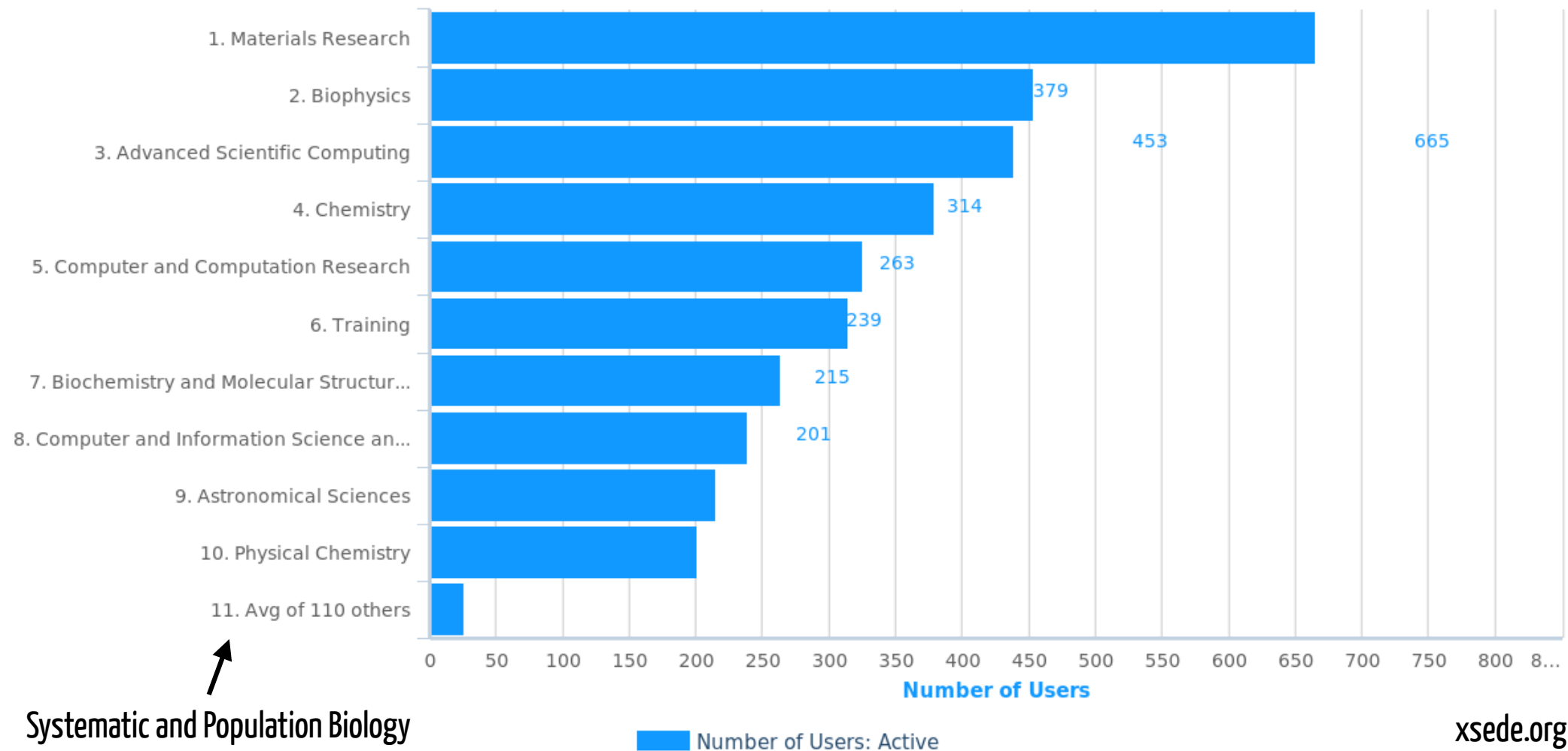


XSEDE jobs by field for 2016



XSEDE users by field for 2016

Number of Users: Active: by Field of Science



New Tools on the Horizon

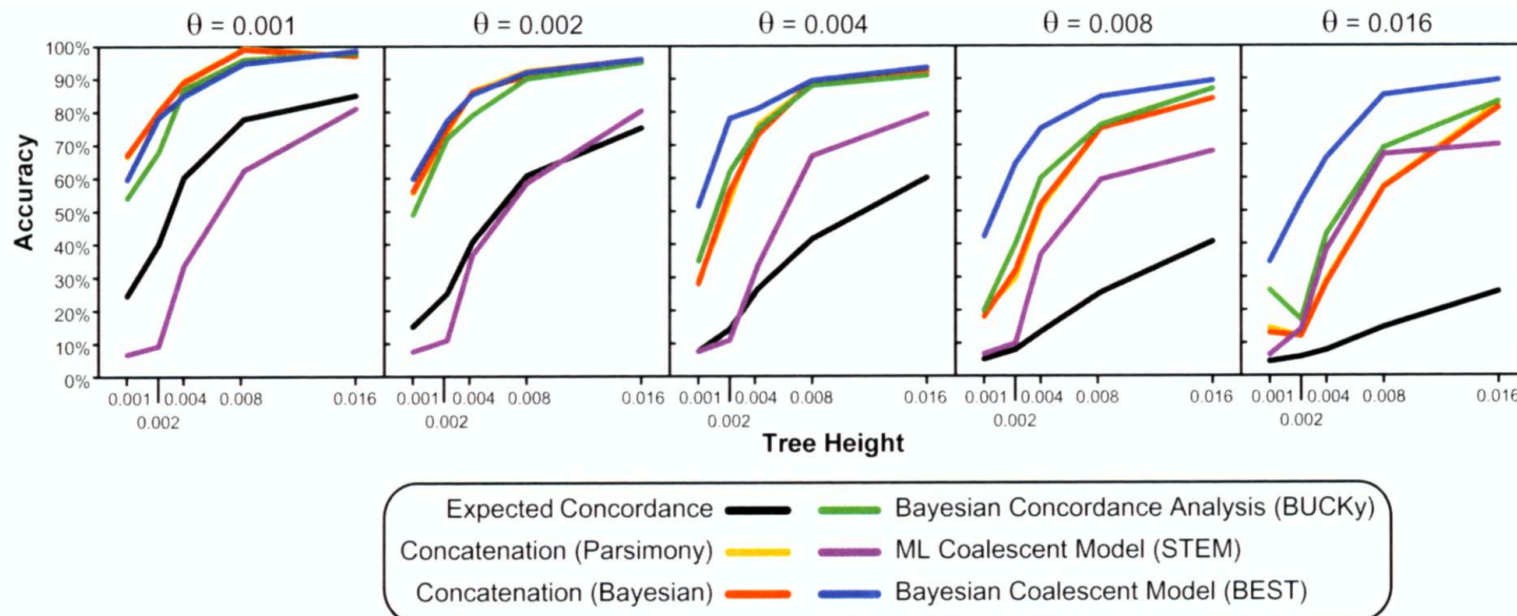
- More complex models
- More efficient sampling. e.g., Hamiltonian Monte Carlo
- More efficient implementations of existing methods

Some Possible Ways Forward

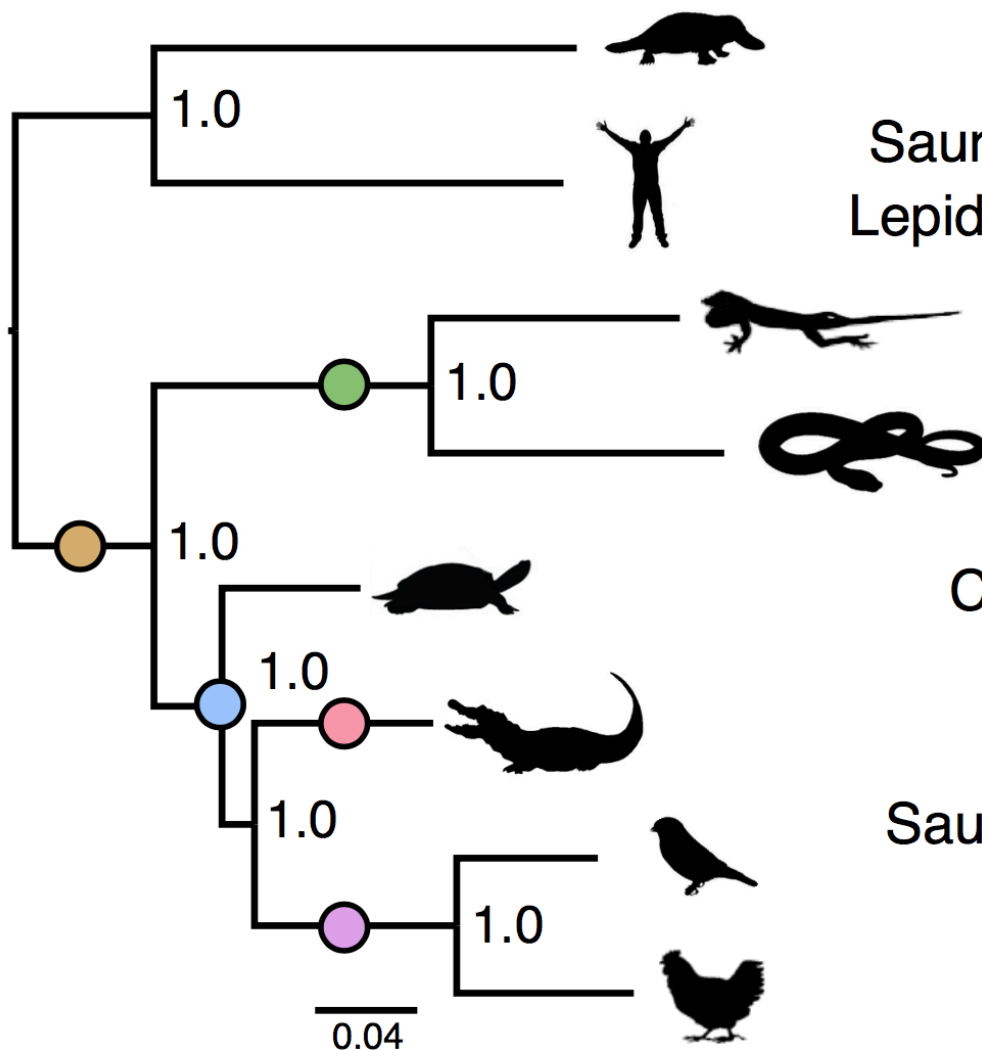
- **Embrace** the computational **challenge**
- **Get very picky** about our data. Careful and detailed data exploration is your friend.

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



A.



B.

