29th Annual Workshop in Molecular Evolution

Director: David Hillis

Introduction: David M. Hillis, University of Texas

Marine Biological Lab, Woods Hole

2016

Molecular Evolution

- Using biomolecules to understand evolutionary history and evolutionary processes

- Phylogenetic trees are critical for studying molecular evolution, because many biological phenomena of interest can be modeled as bifurcating processes

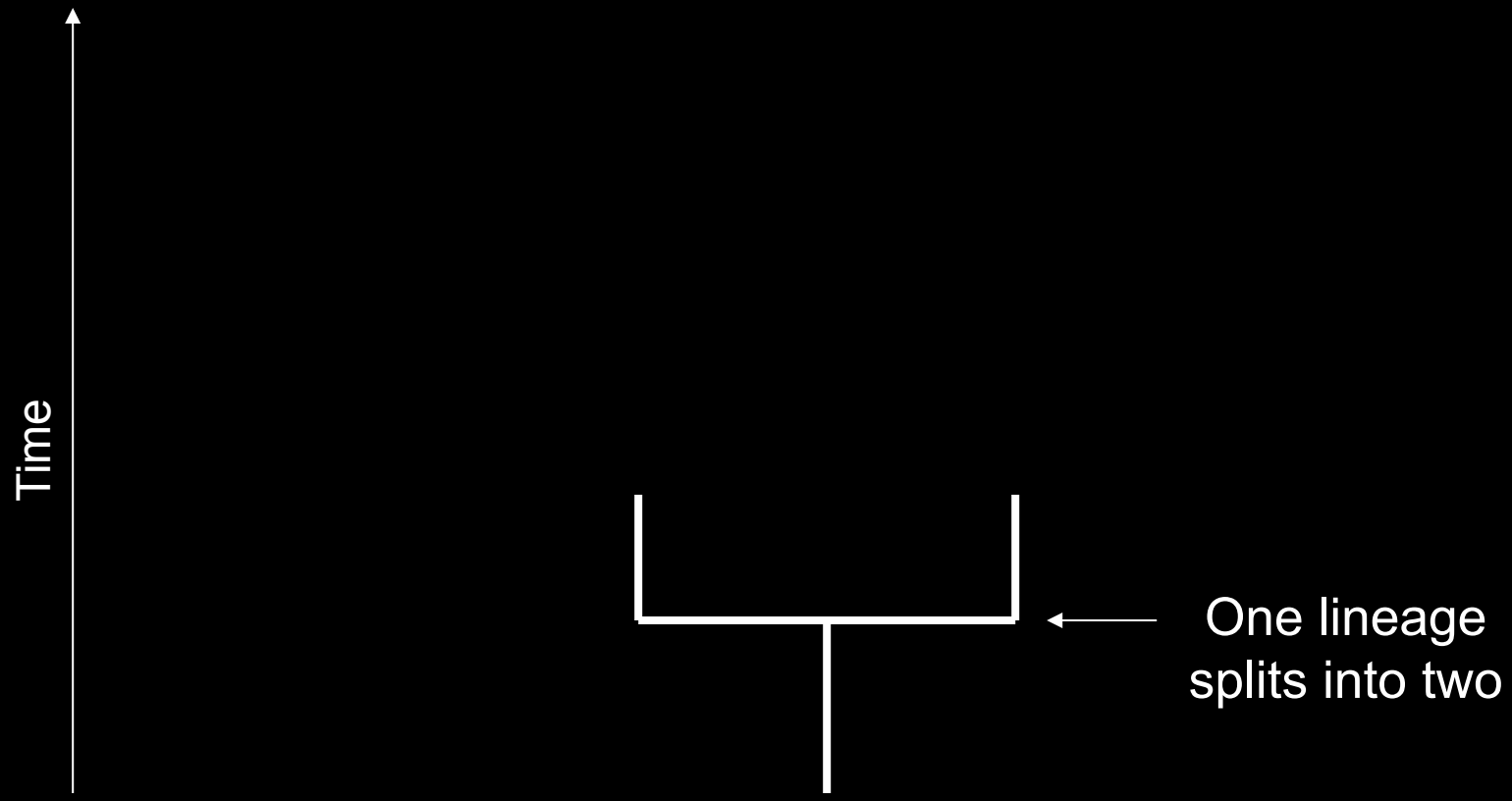
Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.

Consider an ancestral lineage
(e.g., descendants from one HIV virus)

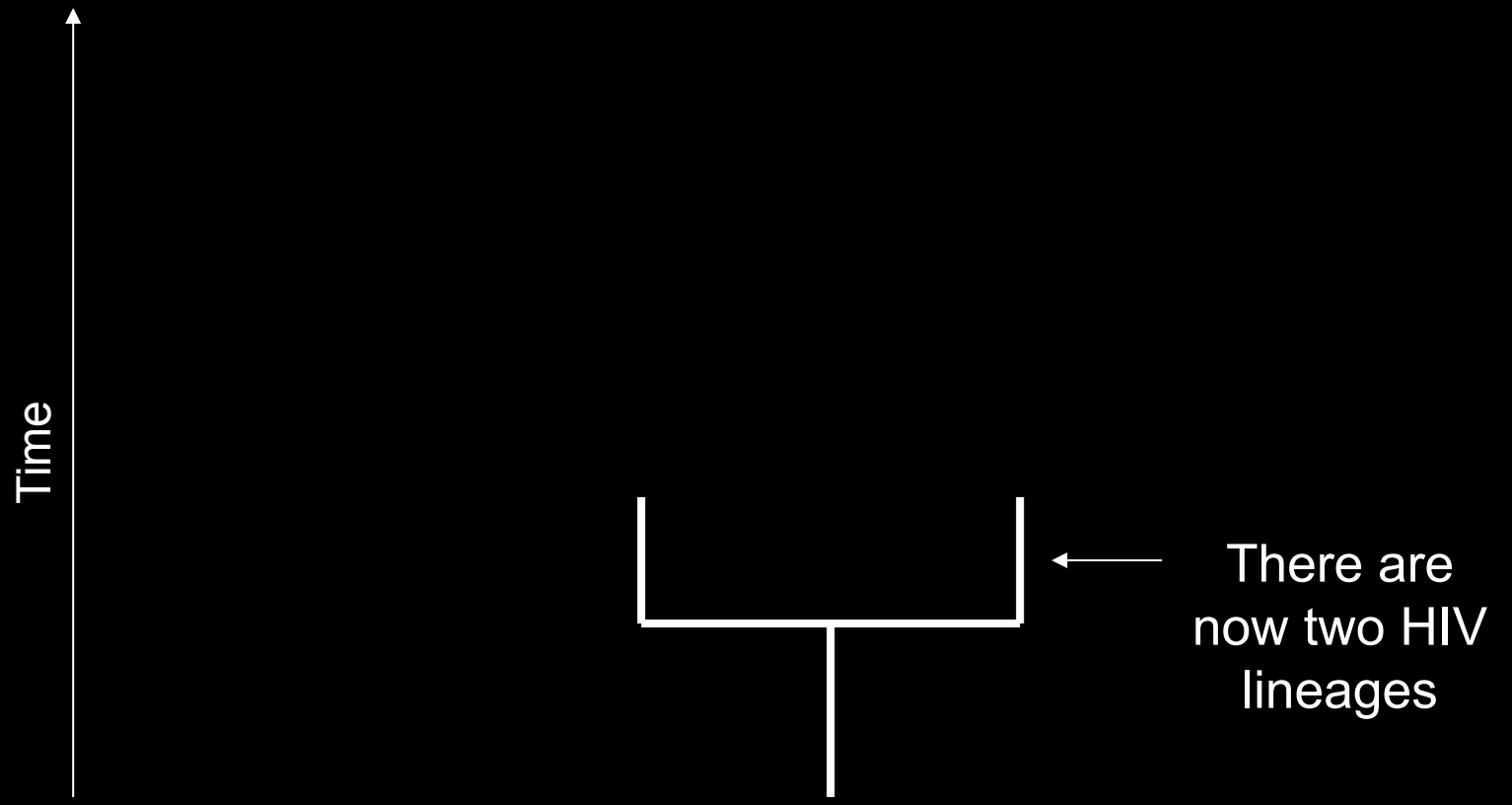
Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



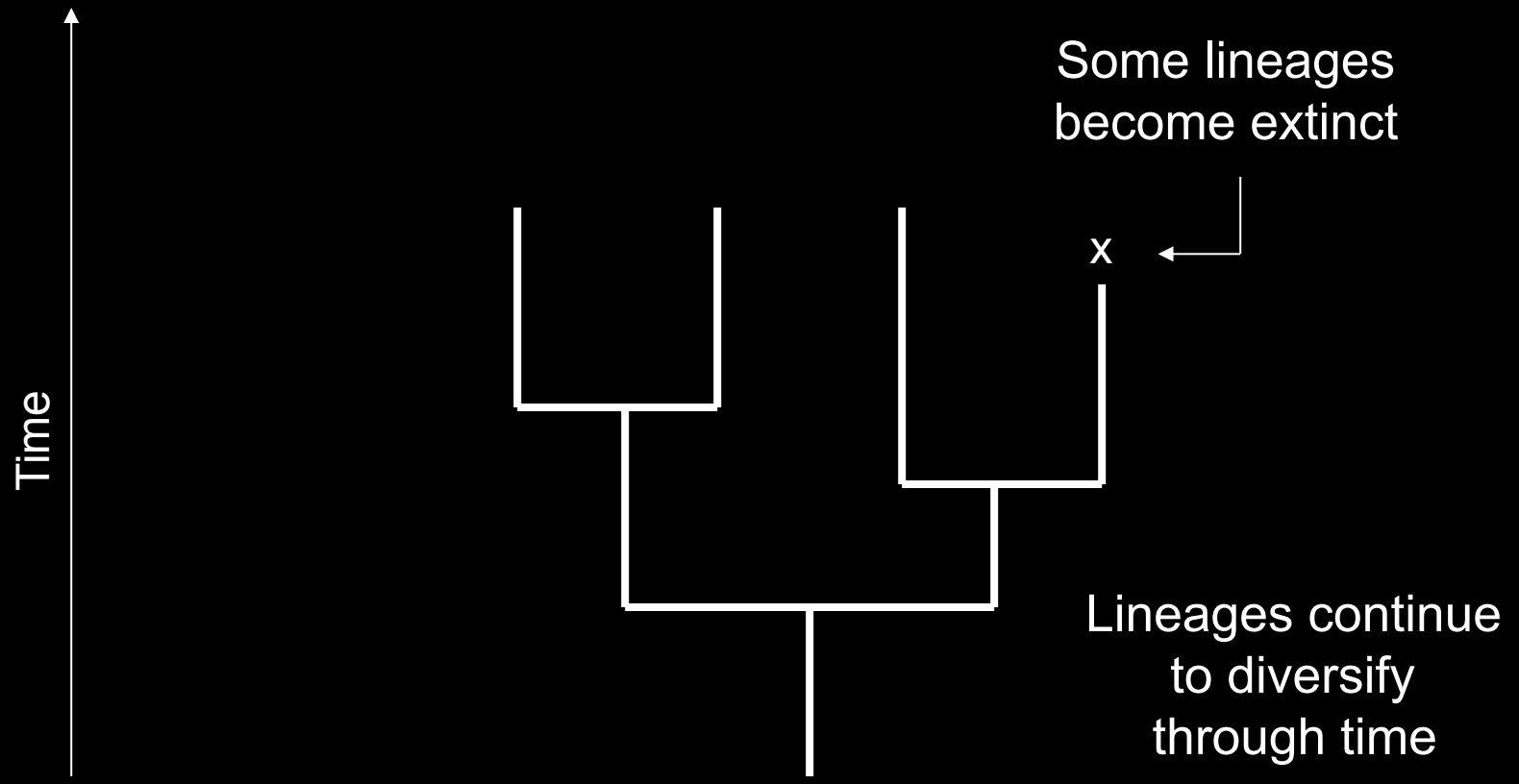
Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



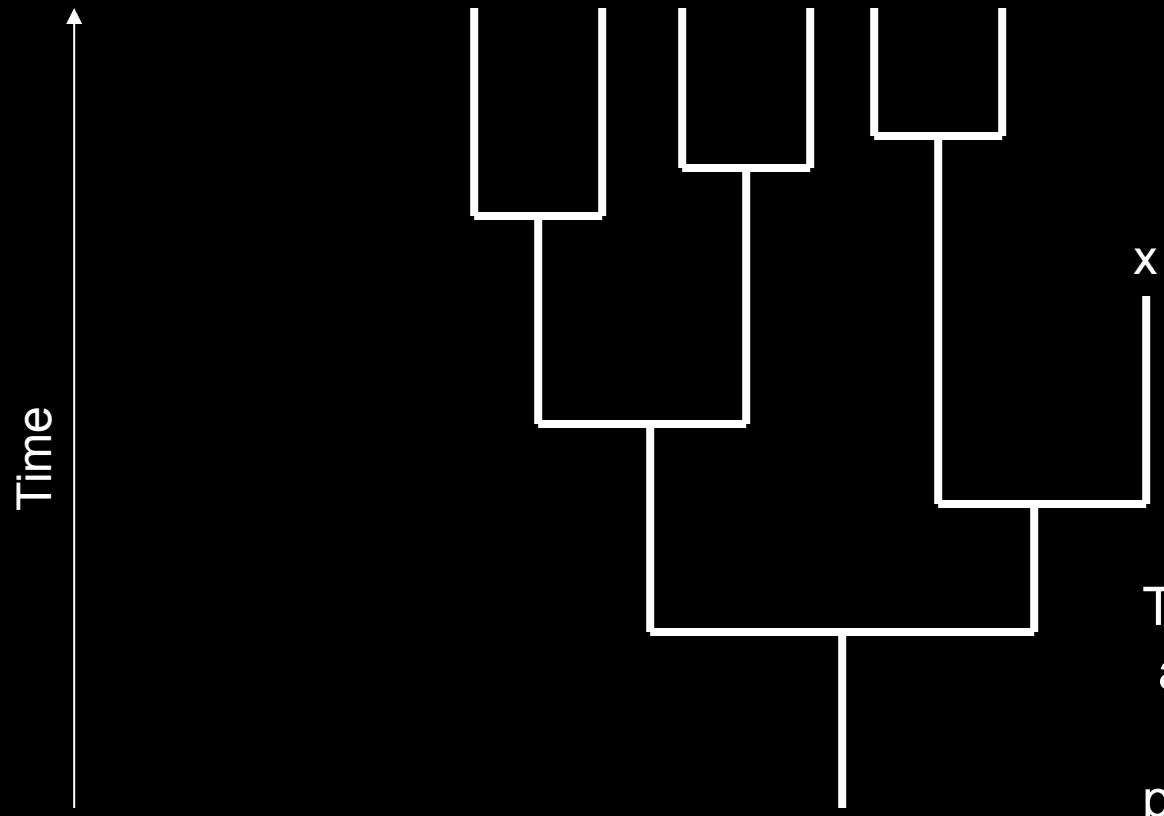
Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



Phylogeny

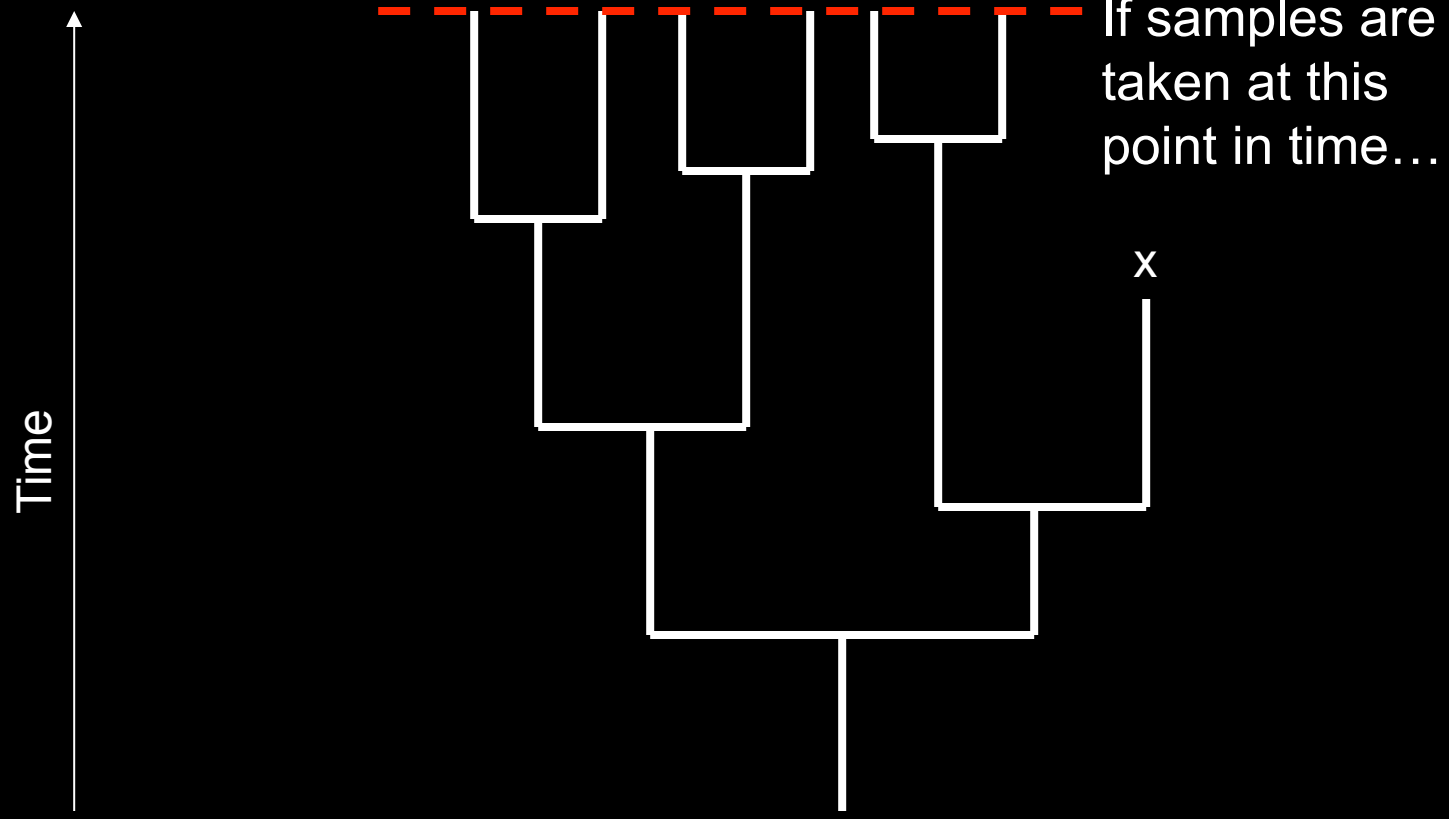
Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



The relationships among lineages produce a phylogenetic tree

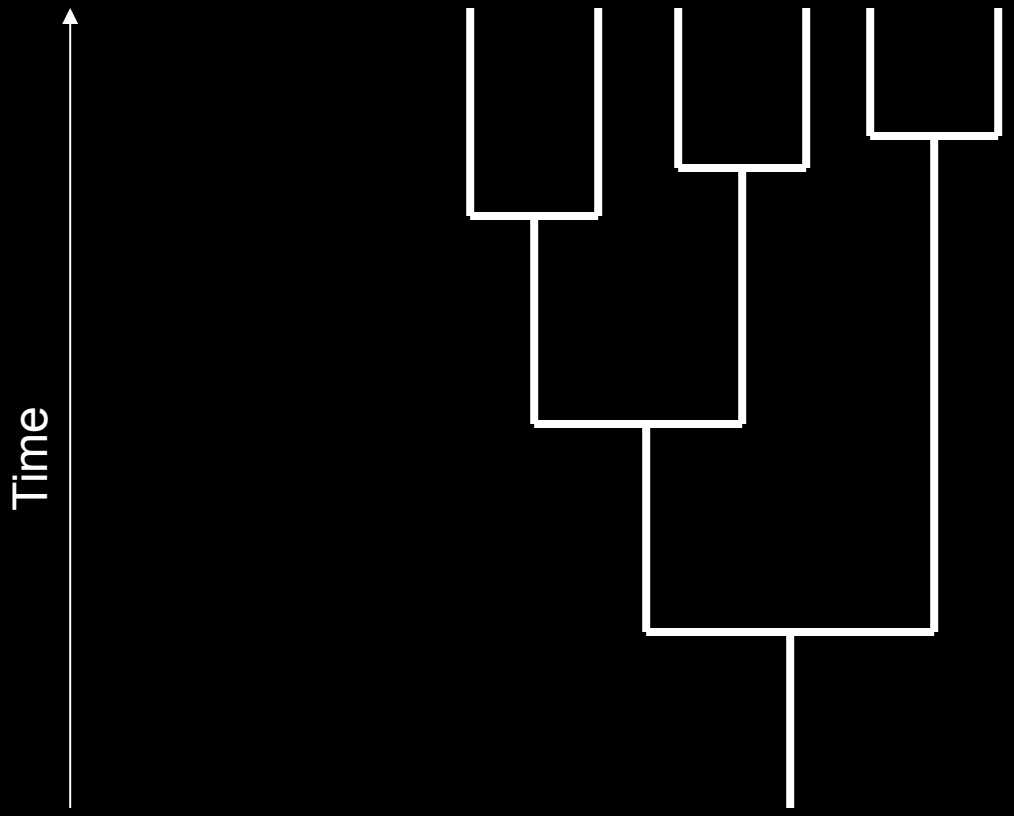
Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.

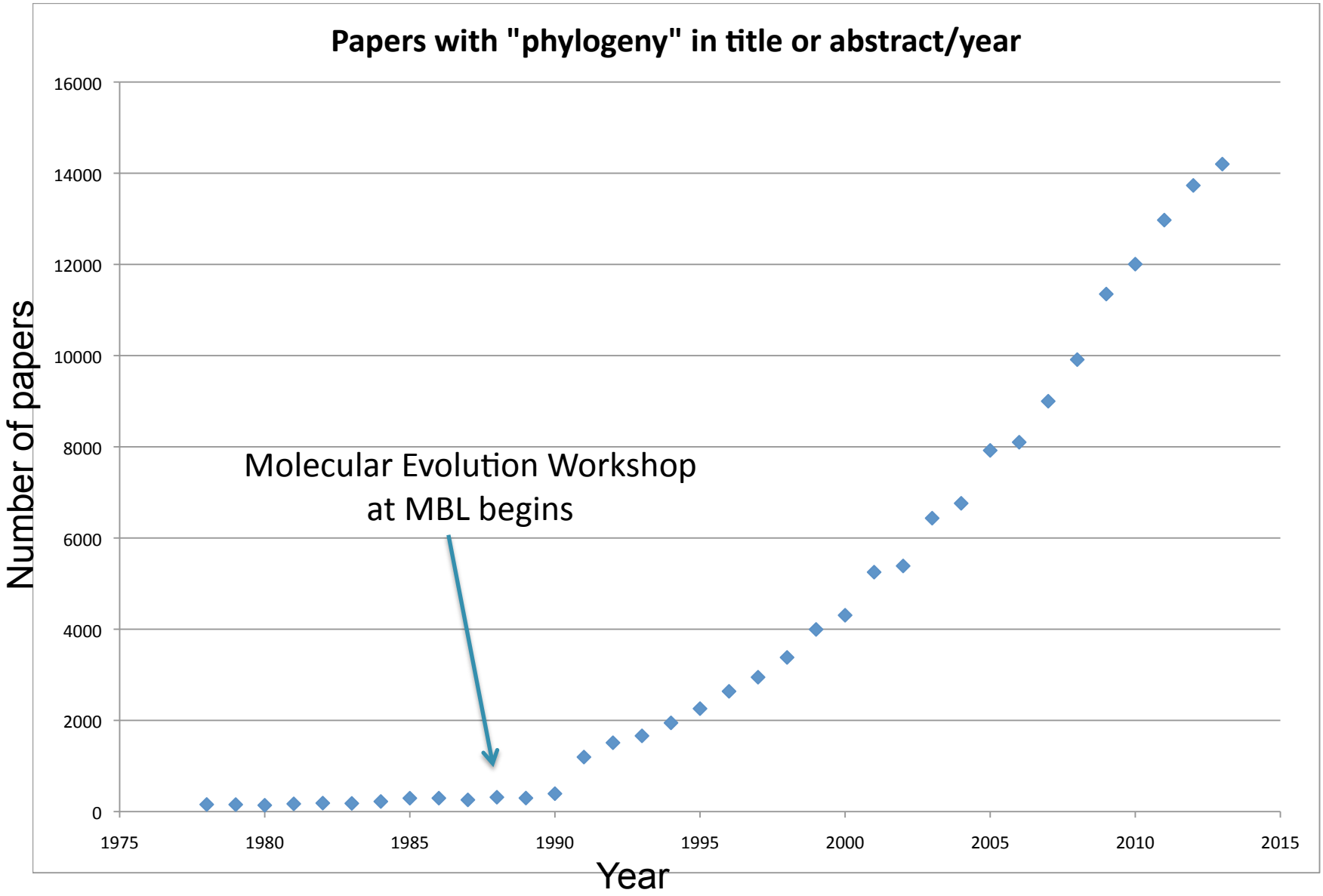


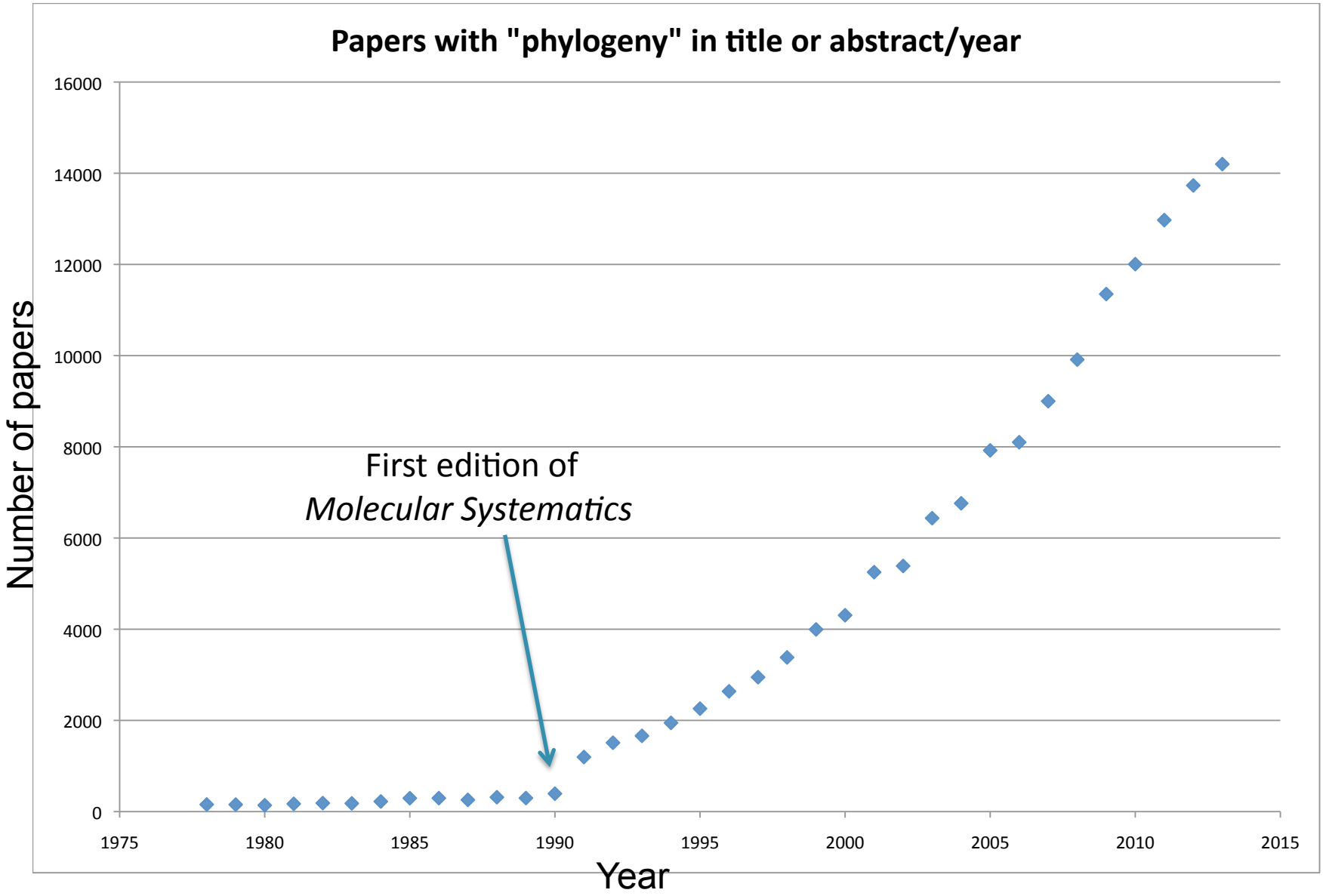
Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



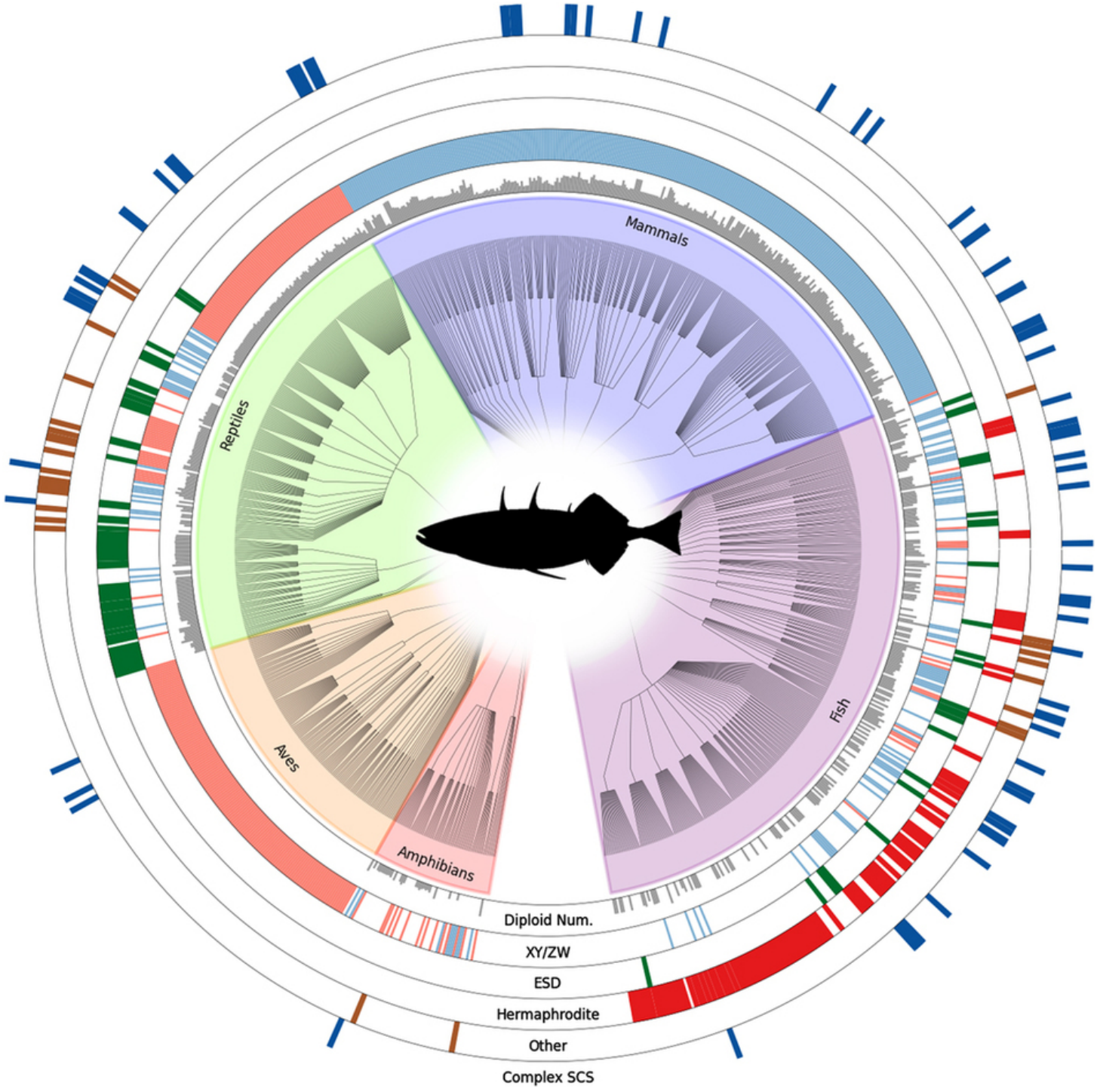
…then (hopefully, with appropriate data, models of evolution, and methods of analysis) the reconstructed phylogenetic tree will look like this

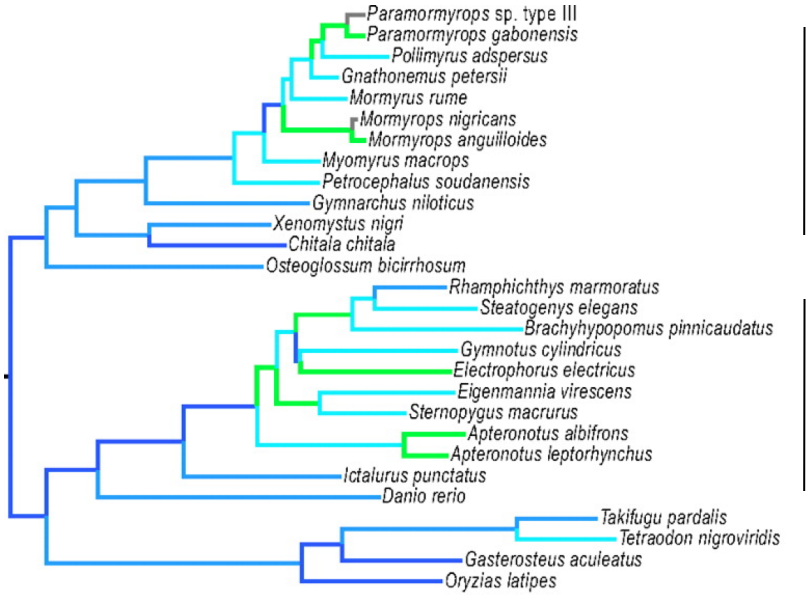

Papers with "phylogeny" in title or abstract/year
Number of papers
16000
14000
12000
10000
8000
6000
4000
2000
0
Molecular Evolution Workshop
at MBL begins
1975
1980
1985
1990
1995
2000
2005
2010
2015
Year


Papers with "phylogeny" in title or abstract/year
Number of papers
16000
14000
12000
10000
8000
6000
4000
2000
0
1975
1980
1985
1990
1995
2000
2005
2010
2015
Year
First edition of
Molecular Systematics

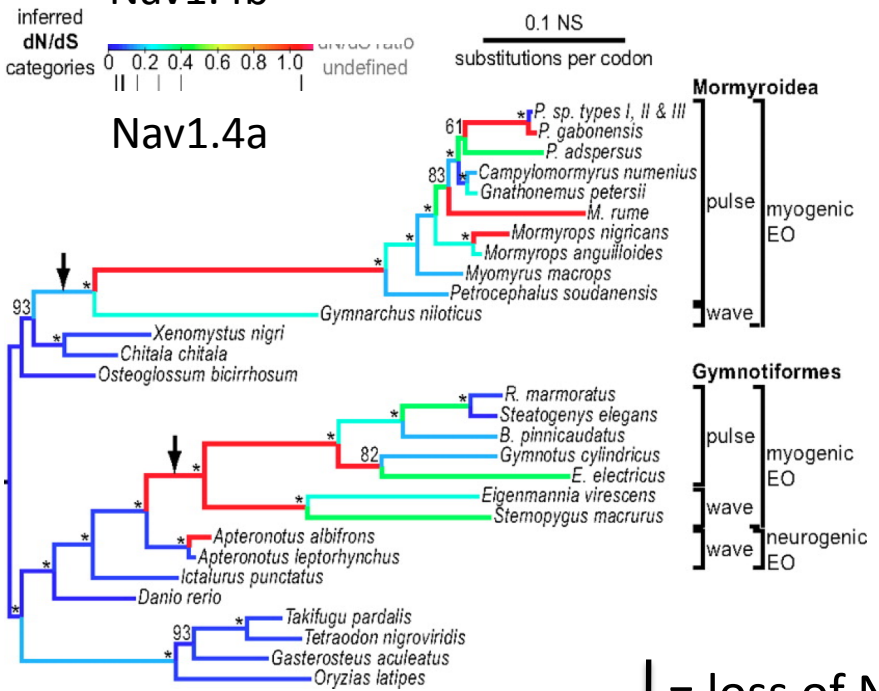
Major Uses of Phylogenetic Trees

- The relationships are often informative about evolution (where genes came from, how and where viruses are transmitted, the origins of particular structures or processes)
- Trees allow estimation of time for various evolutionary events
- Trees facilitate analysis of evolutionary processes, such as selection, gene flow, and speciation
- Trees allow appropriate comparative biology (identify appropriate comparisons)
- Trees are informative about ancestral states and state transitions


Mammals
Reptiles
Aves
Amphibians
Fish
Diploid Num.
XY/ZW
ESD
Hermaphrodite
Other
Complex SCS



Nav1.4b



↓= loss of Nav1.4a expression in muscle

29th Workshop on Molecular Evolution

Focus:
70% models and methods of analysis
10% on data acquisition
20% on applications

Why this emphasis on models and methods of analysis?

Where were/are the problems?

- 1970s and 1980s: we badly needed more data
- 1990s and 2000s: we badly needed better sampling of taxa (also more loci)
- In the 2010s, we are finally getting some pretty good data sets. But are our models and methods up to the task of analyzing the data?
- Are we over-confident in the conclusions from existing models and methods?

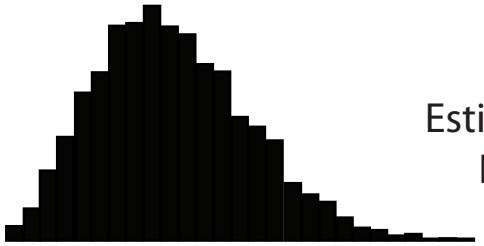
What are the sources of error in statistical inference?

Population

Draw sample from (unknown) true distribution

Stochastic sampling error (decreasing with increased sample size)

Sample

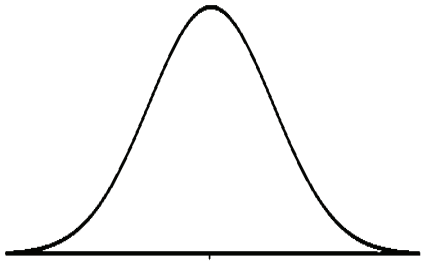


Estimate of true distribution: Better with more data

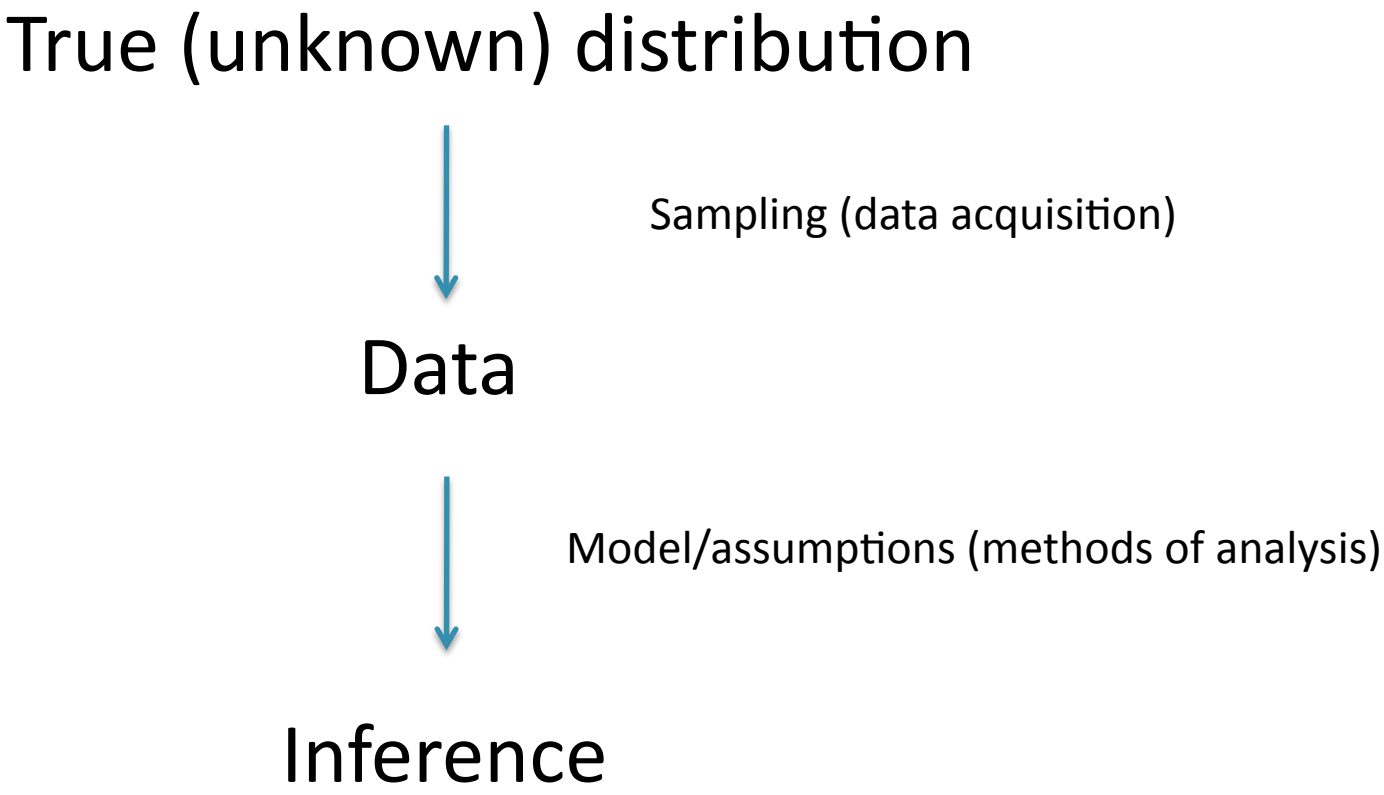
Inferences about true distribution

Systematic error in model assumptions (does not decrease with increasing sample size)

Inference



Our estimate of the true distribution (includes sampling error and systematic error)



Some people seem to think that with enough data (say, whole genomes from everything), we can stop worrying about inference error. Why are they mistaken?

True (unknown) distribution

If sample is large enough, sampling error may indeed approach 0%

Data

But even if we have 100% confidence that our data are representative of the underlying population, we may still have 0% confidence in our model assumptions

Inference

True (unknown) distribution

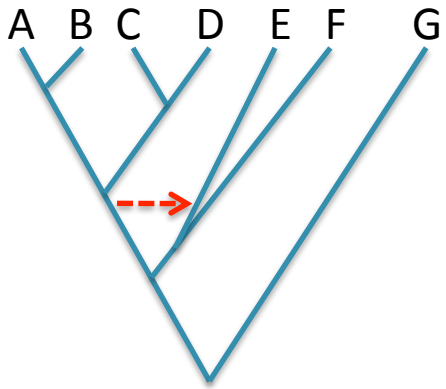
If sample is large enough, sampling error may approach 0%

Data

Even if we have 100% confidence that our data are representative of the underlying population, we may still have 0% confidence in our model assumptions

Inference

(100% confidence in sample) X (0% confidence in model assumptions) = 0% confidence in inference

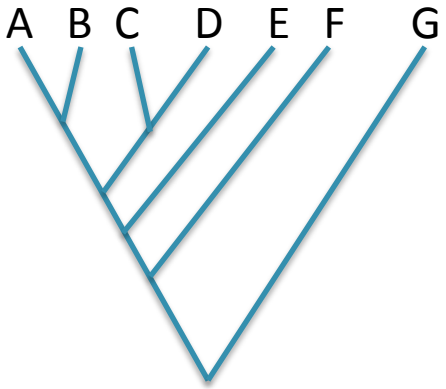


True evolutionary history and processes (unknown)

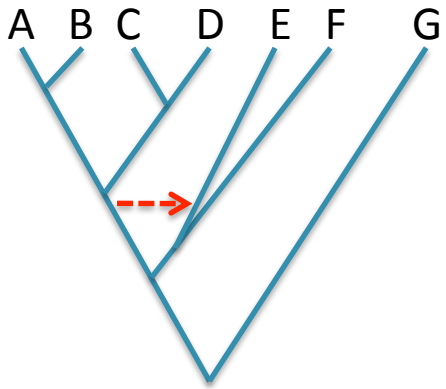
Universe of possible genomes of A-G, given true phylogeny and evolutionary processes

Samples drawn from genomes A-G (subject to sampling error, decreasing with increasing data)

Estimate phylogeny, based on model assumptions (subject to systematic error)



If we collect enough data, we will have 100% bootstrap proportions or Bayesian posterior probabilities for our inference.

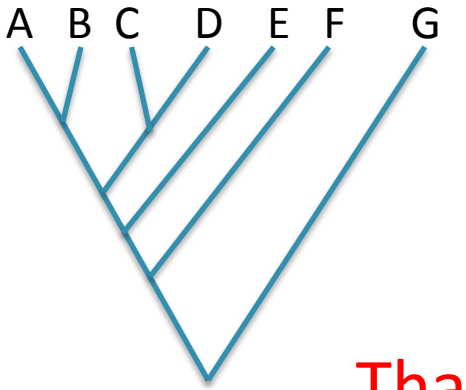


True evolutionary history and processes (unknown)

Universe of possible genomes of A-G, given true phylogeny and evolutionary processes

Samples drawn from genomes A-G (subject to sampling error, decreasing with increasing data)

Estimate phylogeny, based on model assumptions (subject to systematic error)



If we collect enough data, we will have 100% bootstrap proportions or Bayesian posterior probabilities for our inference.

That doesn't mean it is correct

Sampling error: Relatively easy to assess

Bootstrapping, Bayesian posterior probabilities
(This is what most people do, because it is easy.)

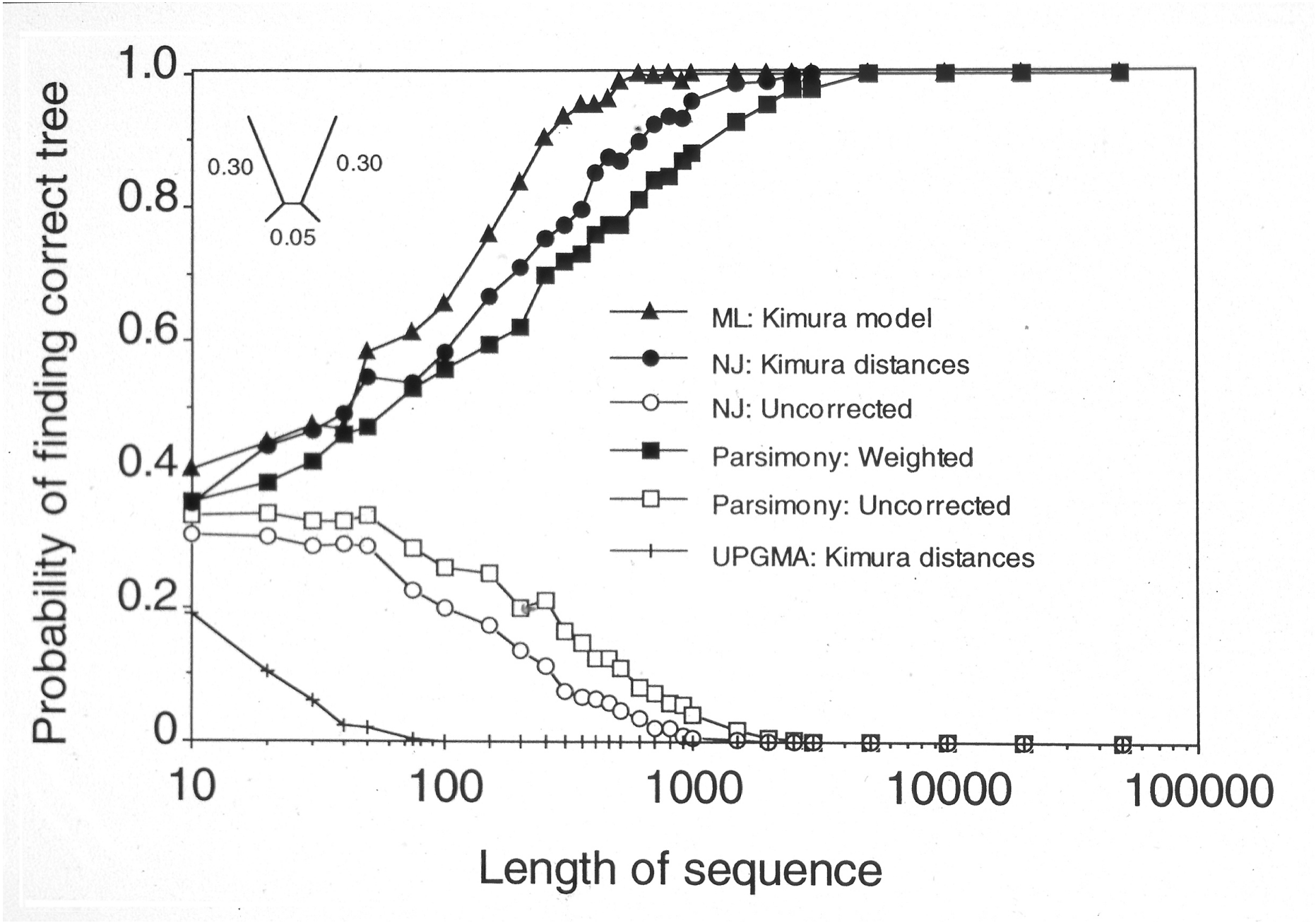
Systematic error: Harder to assess

Examine sensitivity to model assumptions;
develop and compare more complex models
(This is what many people ignore, because it is hard)

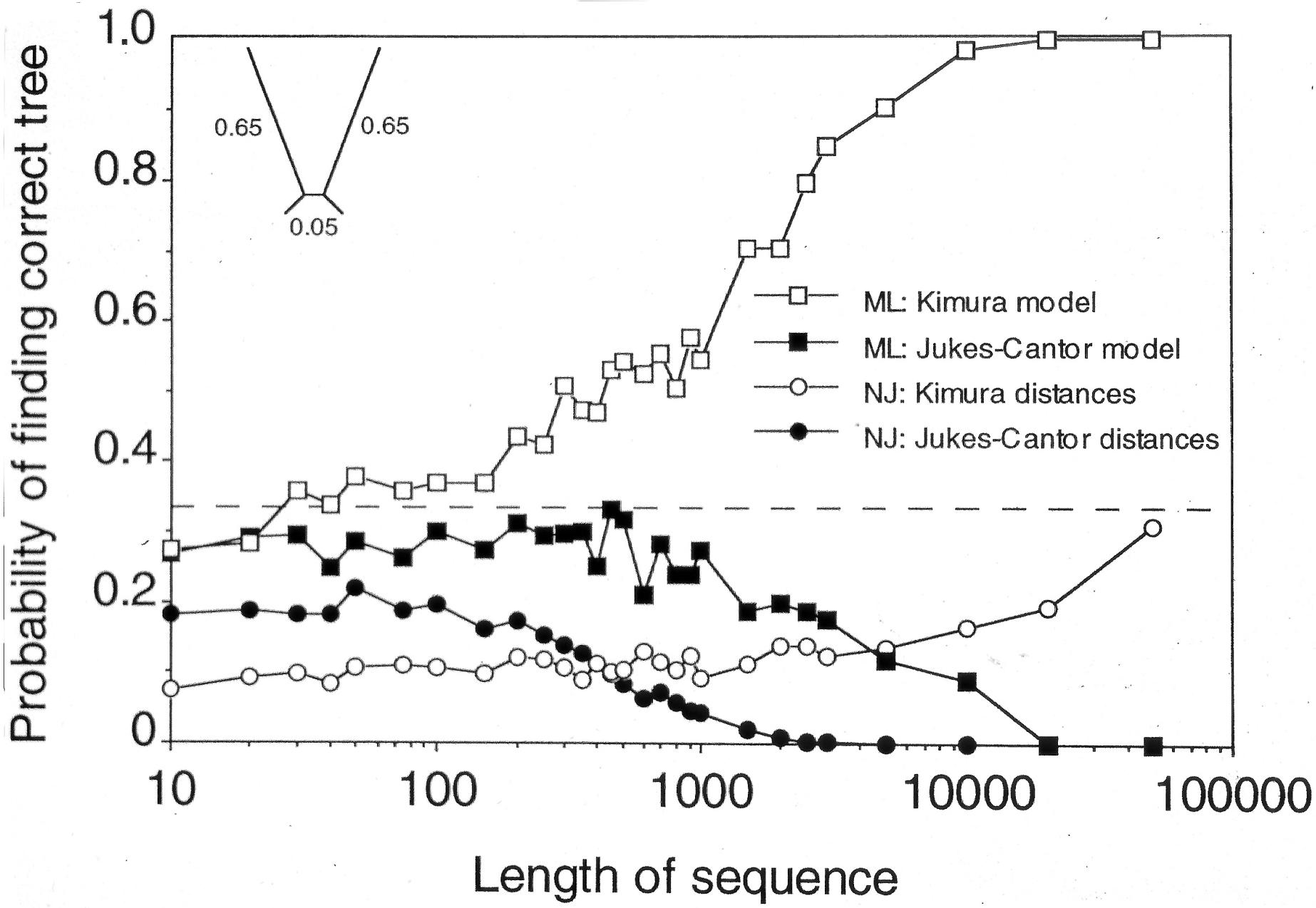
What do we mean by more data? More sequences, or more taxa?

- Yes: both are very important
- Early days of molecular systematics: very short sequences, very few genes, lots of sampling error

Emphasis was on increasing sequence length, because we didn't have enough data for our simple models



Hillis, D. M., J. P. Huelsenbeck, and D. L. Swofford. 1994. Hobgoblin of Phylogenetics? Nature 369:363-364.



Hillis, D. M., J. P. Huelsenbeck, and D. L. Swofford. 1994. Hobgoblin of Phylogenetics? Nature 369:363-364.

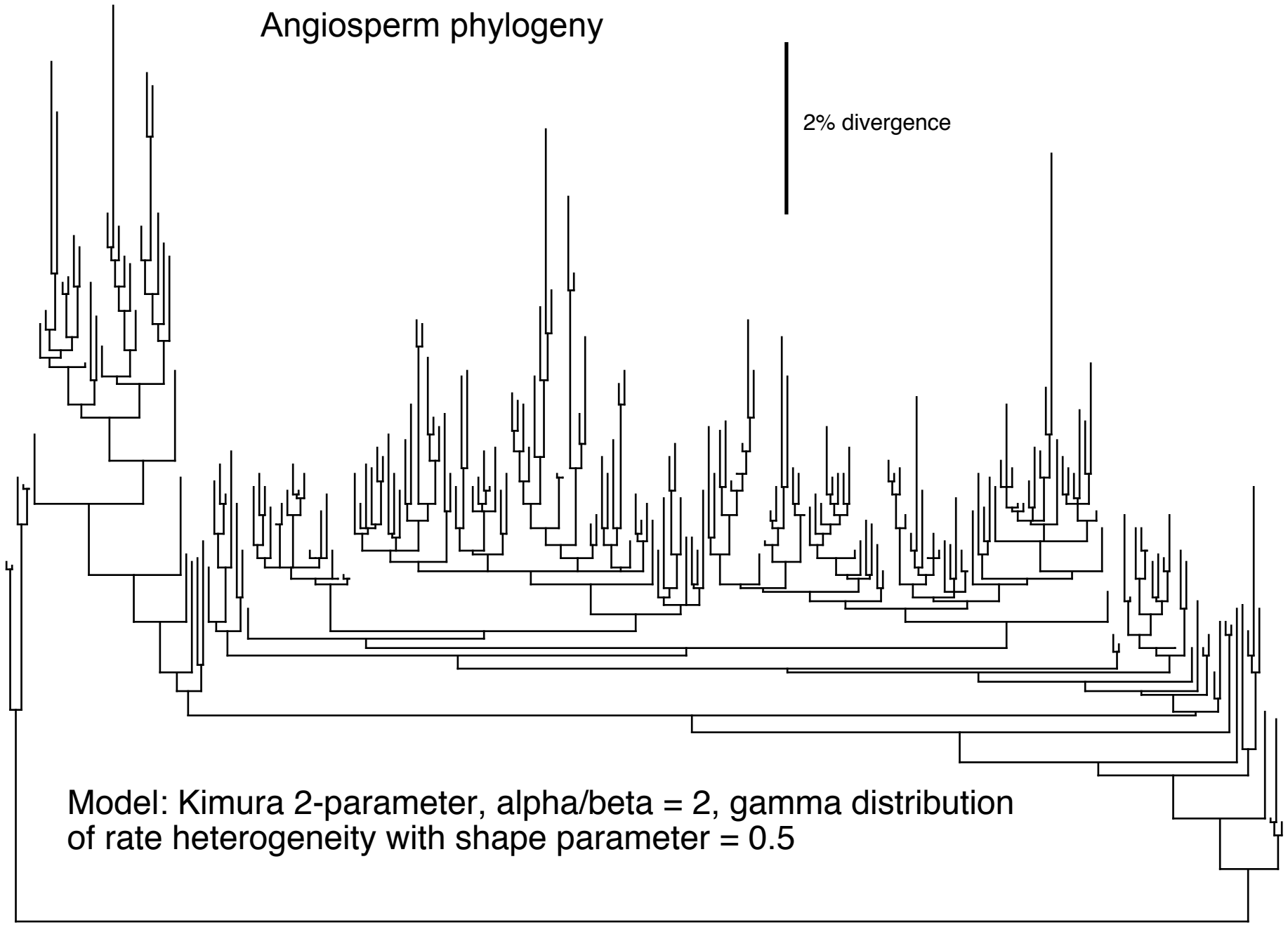
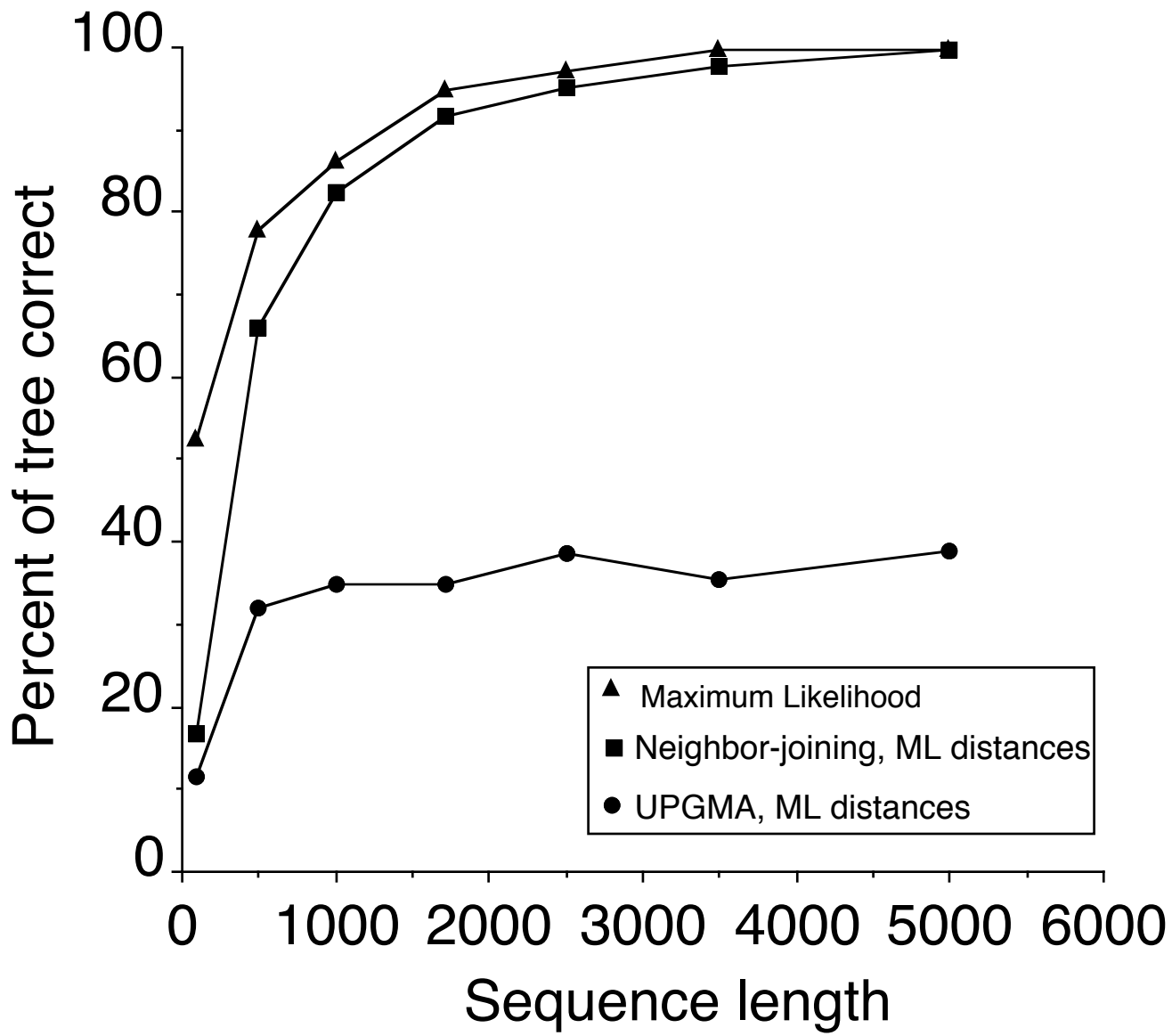


Figure from Hillis, 1996. Inferring complex phylogenies. Nature 383:130-131.

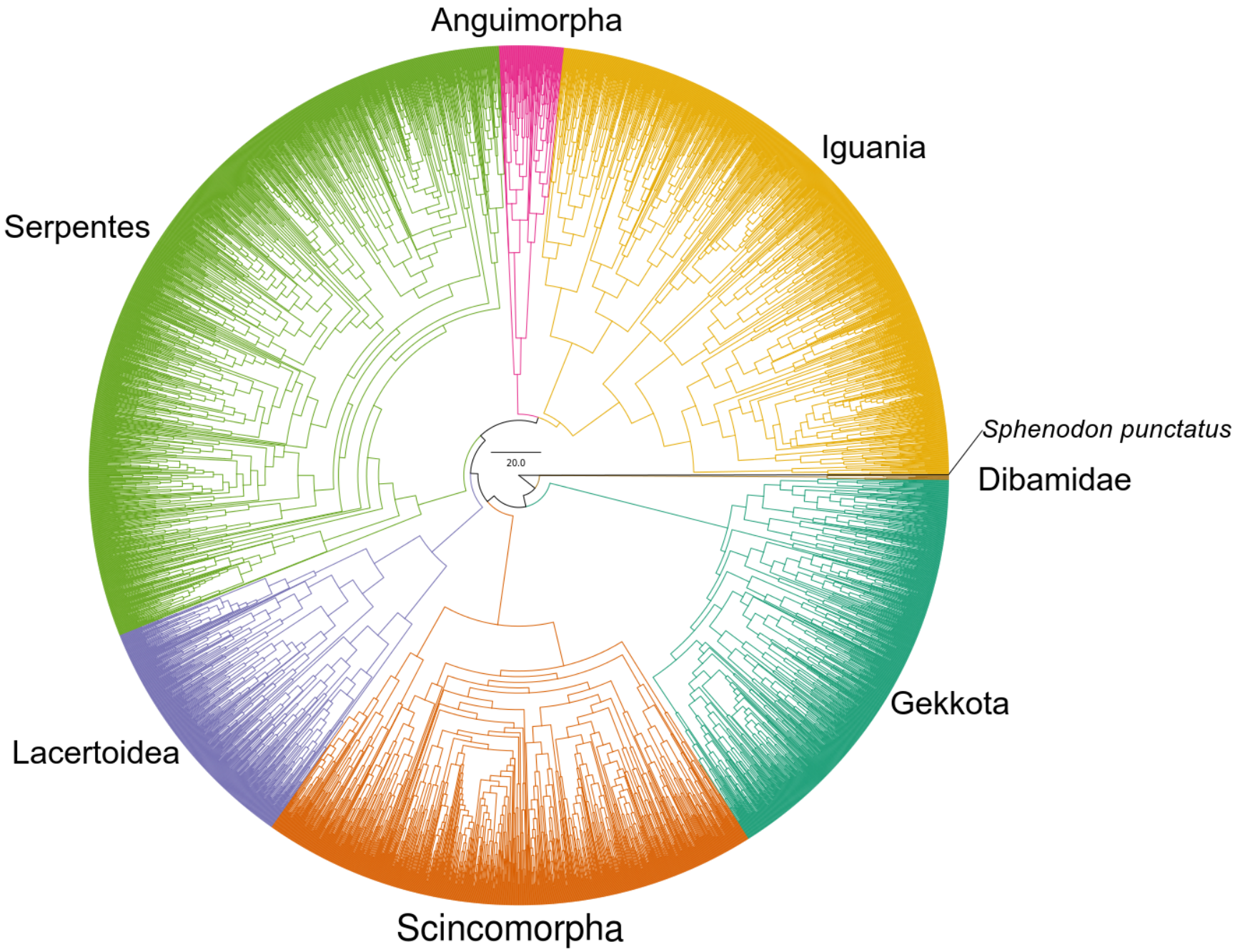

100
80
60
40
20
0
Percent of tree correct
0
1000
2000
3000
4000
5000
6000
Sequence length
Maximum Likelihood
Neighbor-joining, ML distances
UPGMA, ML distances

Dense sampling of taxa greatly reduces systematic error

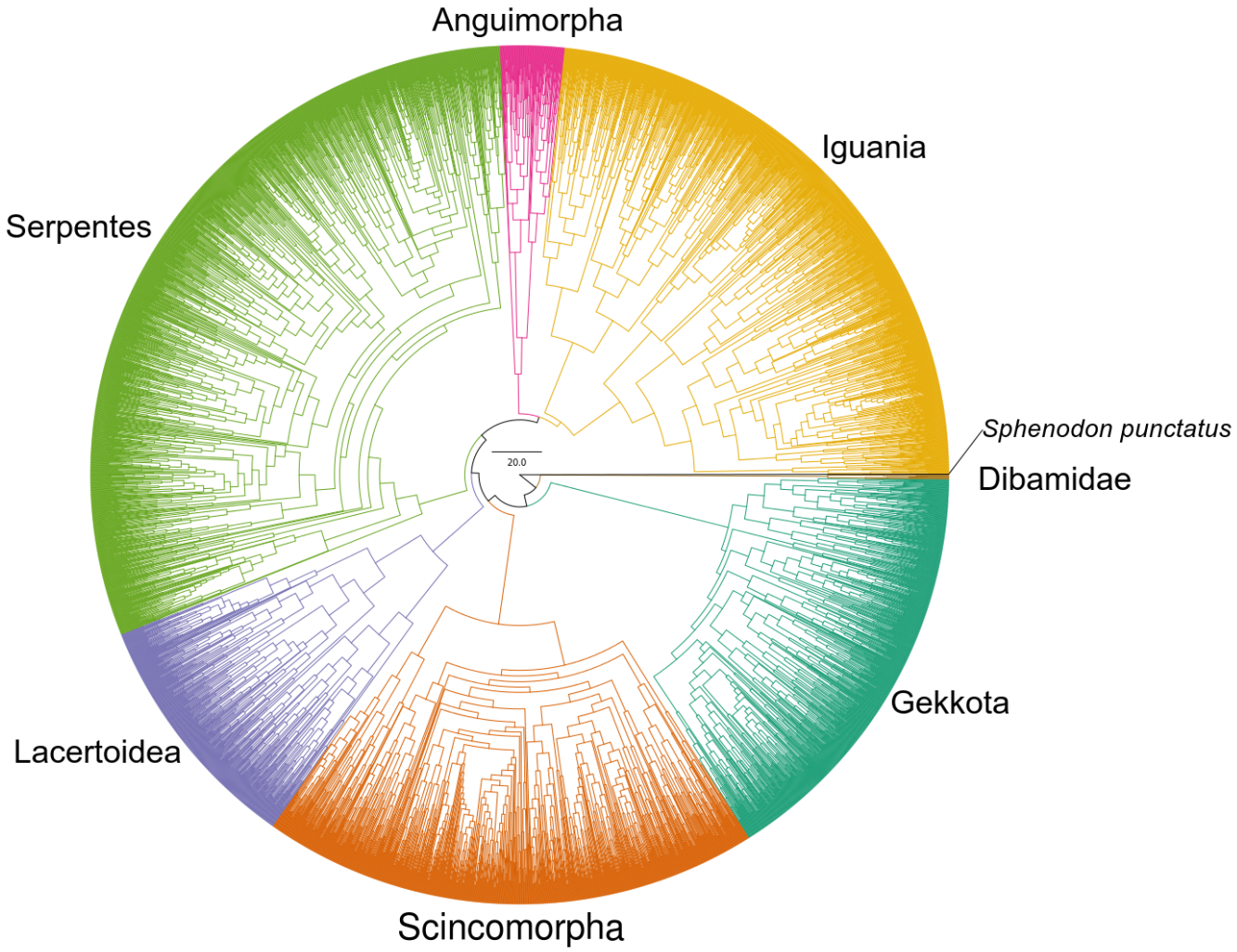
- Using many loci is also critical, to sort out differences between gene trees and species trees
- But even with lots of genes, poor taxon sampling results in large systematic error (fewer taxa means more reliance on overly simplistic models to infer multiple changes along long branches)

Where are we today?

- Many studies use a large sample of genes
- Many studies sample taxa very broadly
- (Often with LOTS of missing data)
- Sampling error tends to be very low
- Bootstrap proportions and BPPs are often near 100%
- But are we over-confident in our model assumptions? (100 x 0 = 0)



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



Good:

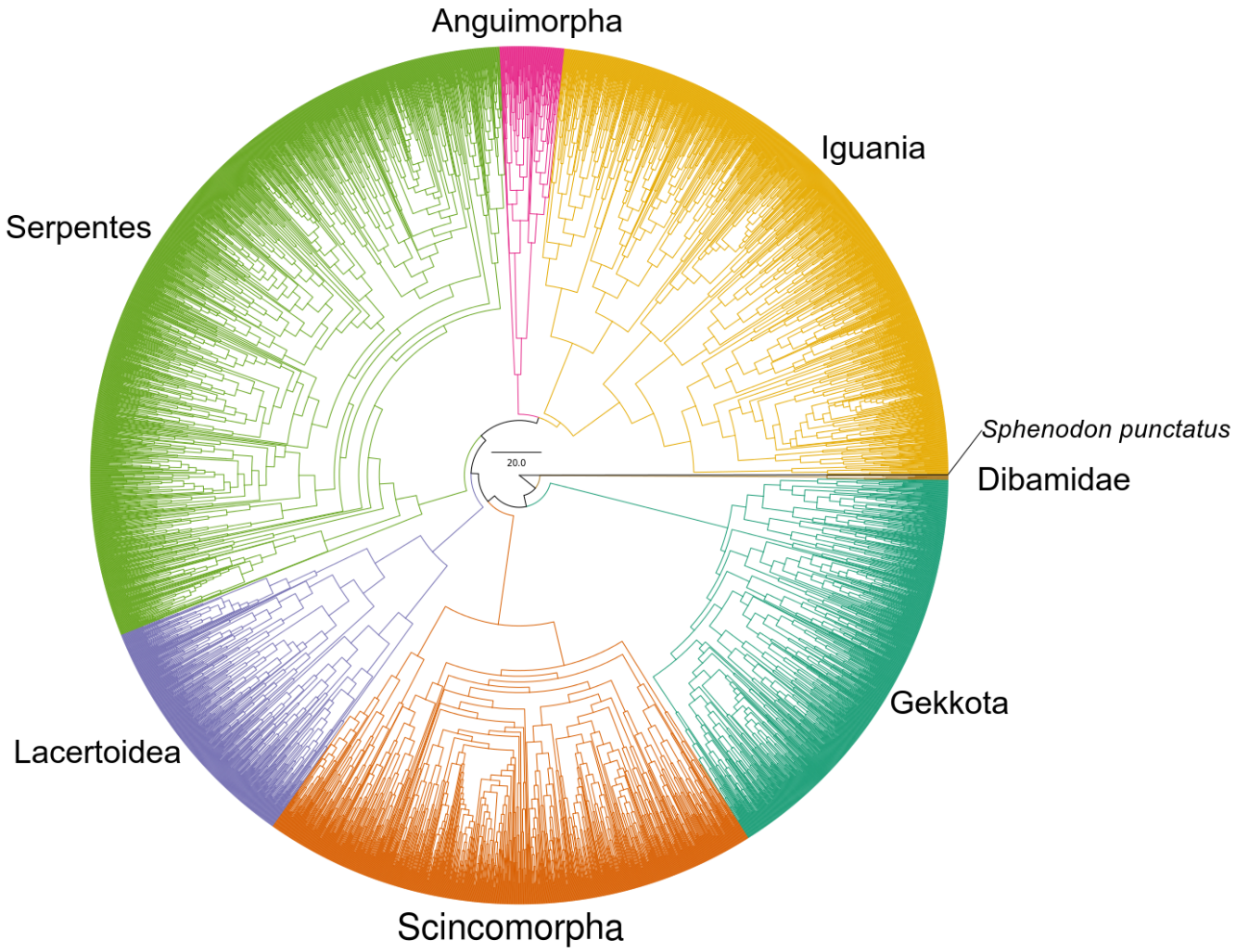
- About 4,000 species
- Multiple nuclear and mitochondrial genes

But worrisome:

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

How do biases in missing data affect our models?

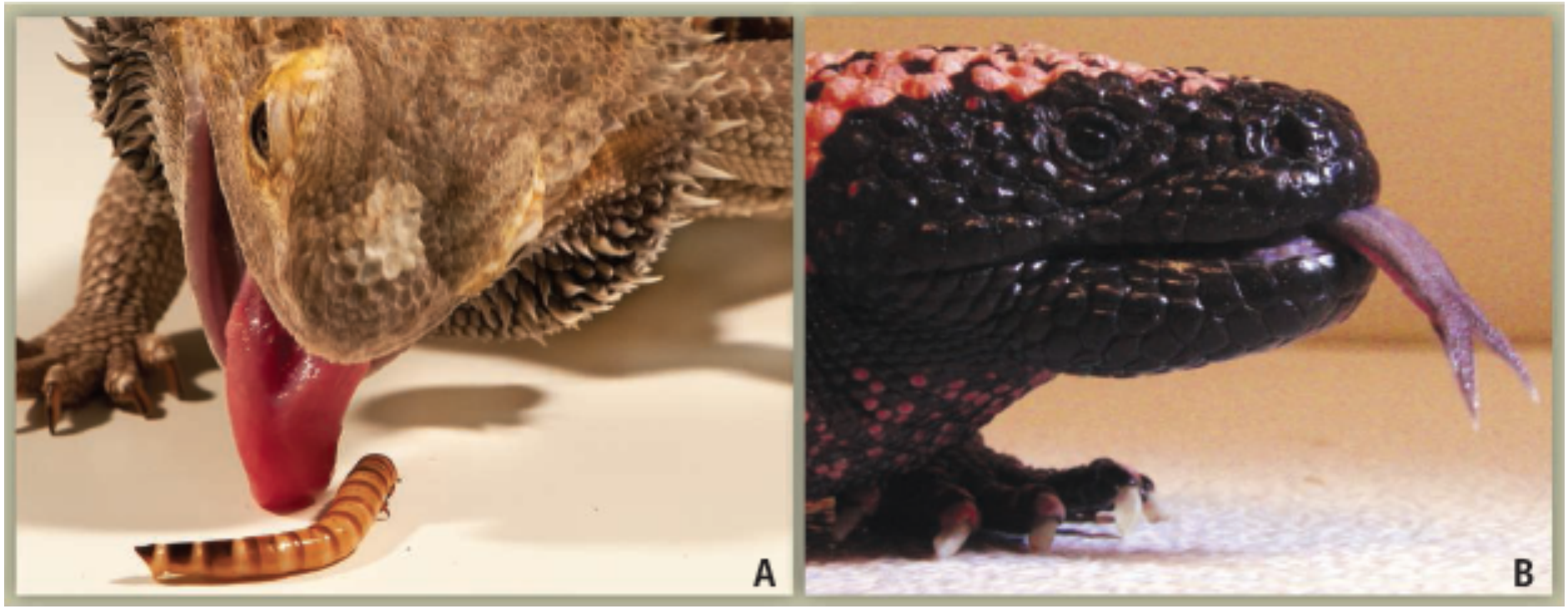
Data from Pyron et al., 2014; Figure from Wright et al., 2015



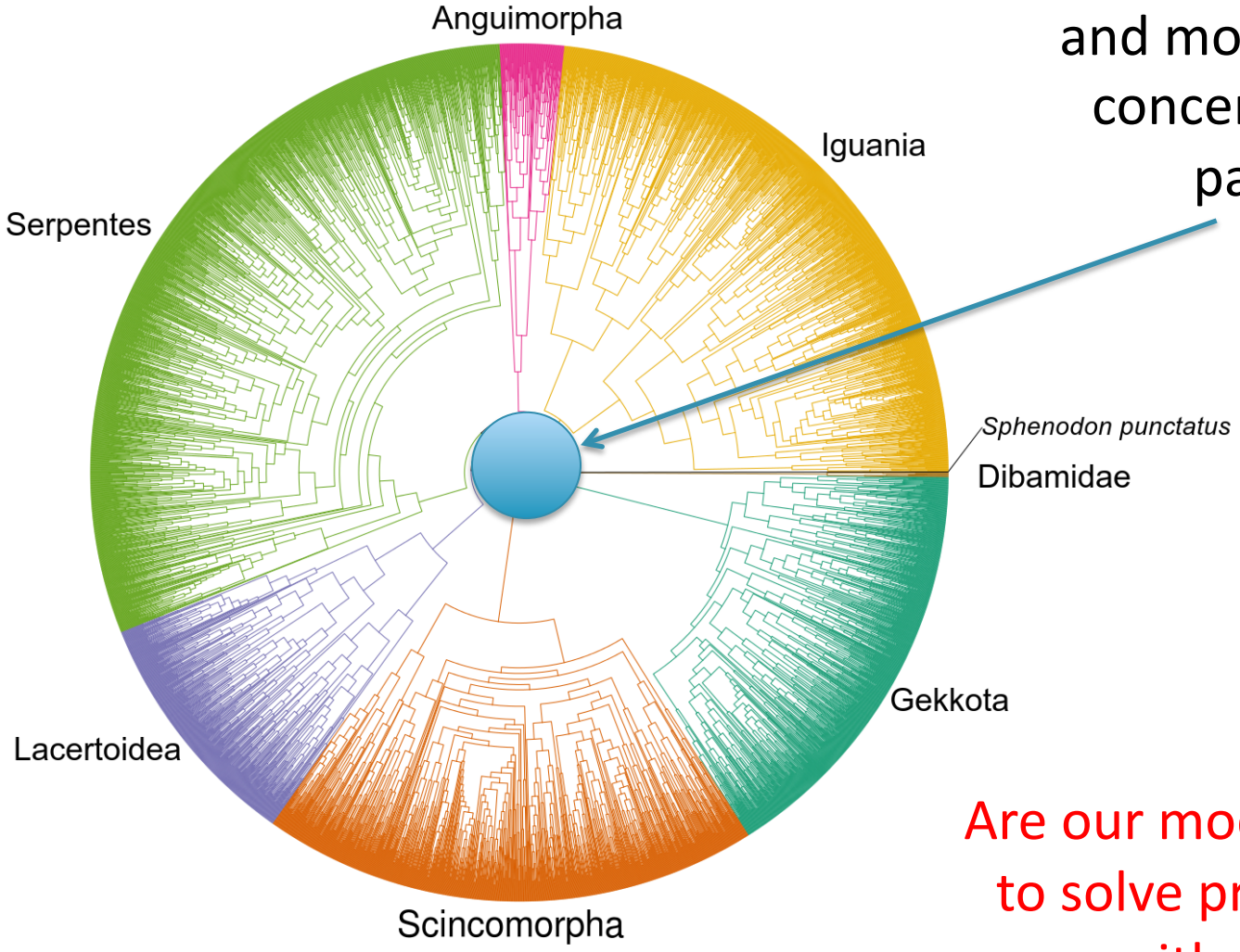
Tree conflicts with morphology, which places iguanians as the sister group of remaining squamates.

How confident should we be in this tree?

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


A
B

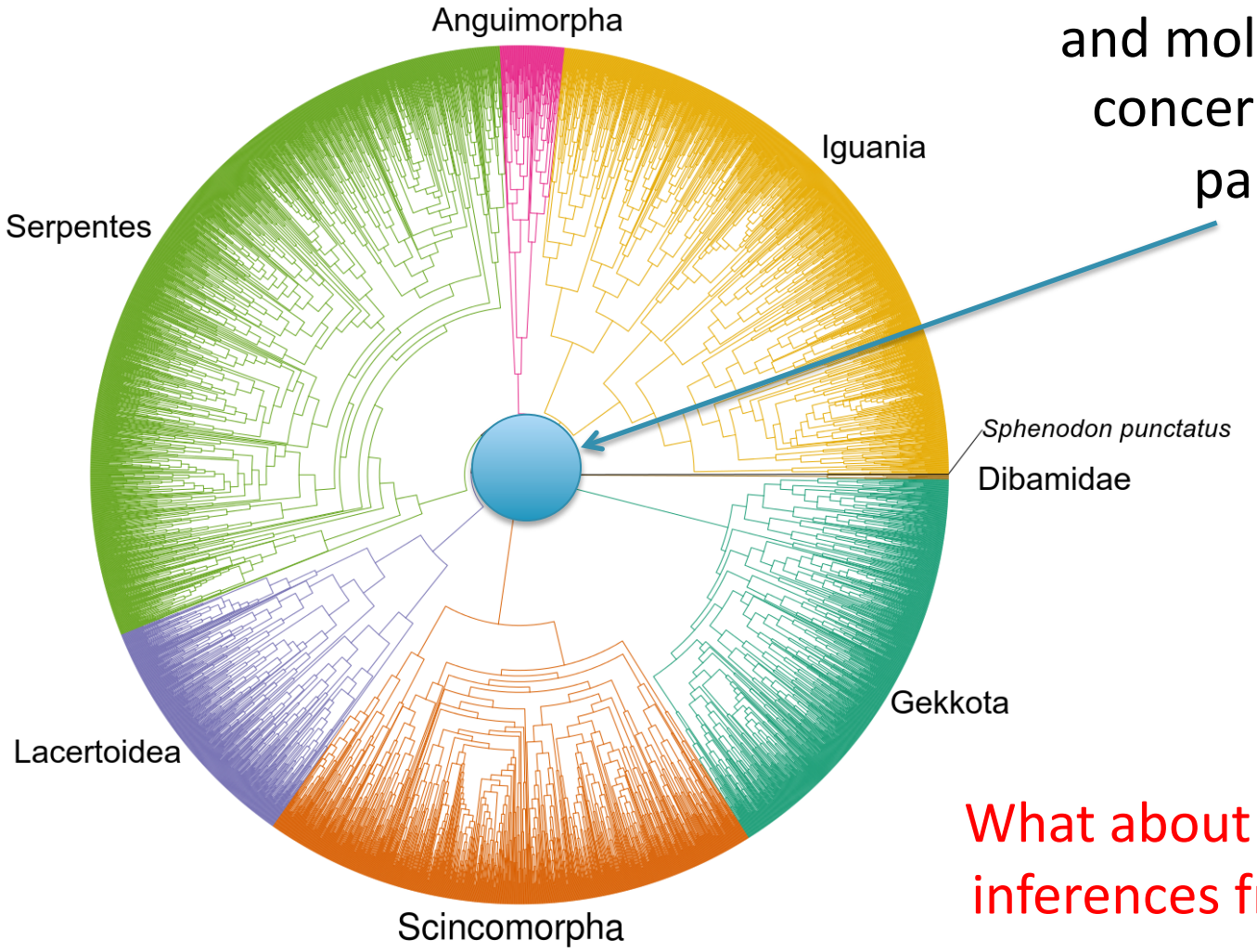
All the arguments between morphologists and molecular systematists concern branches in this part of the tree.



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

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

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



What about our evolutionary inferences from these trees?

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

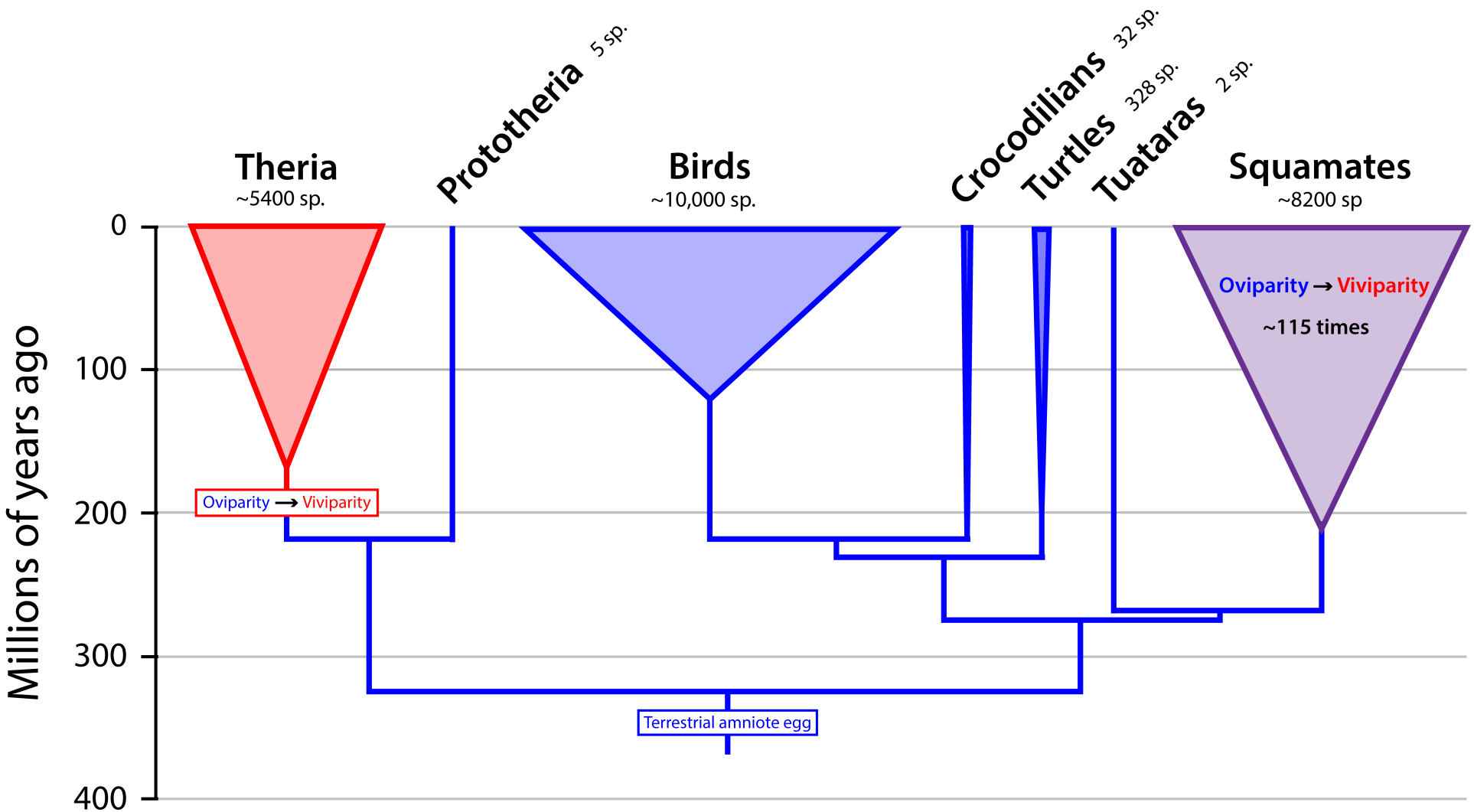
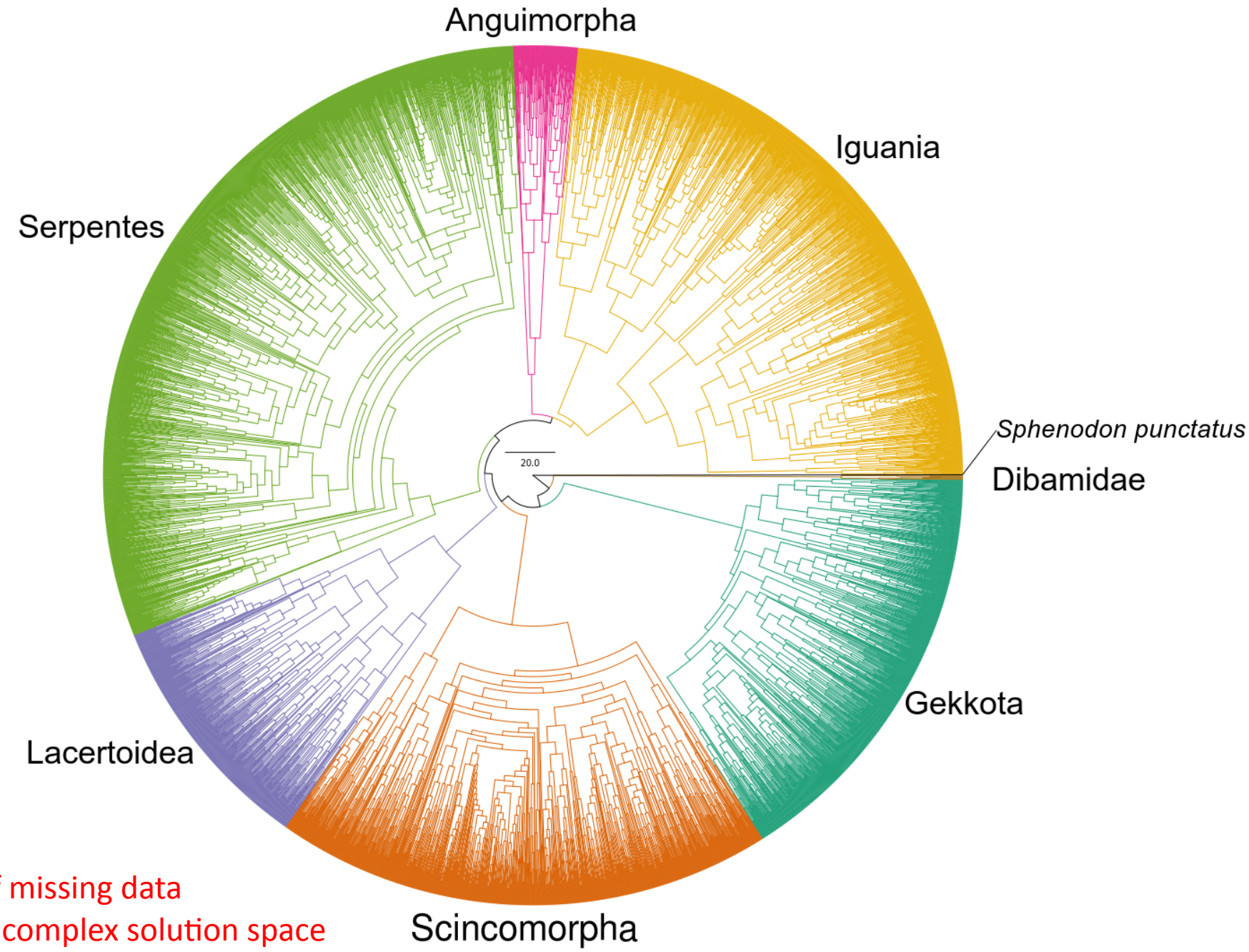


Figure from Wright et al., 2015



- Lots of missing data
- Highly complex solution space

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

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

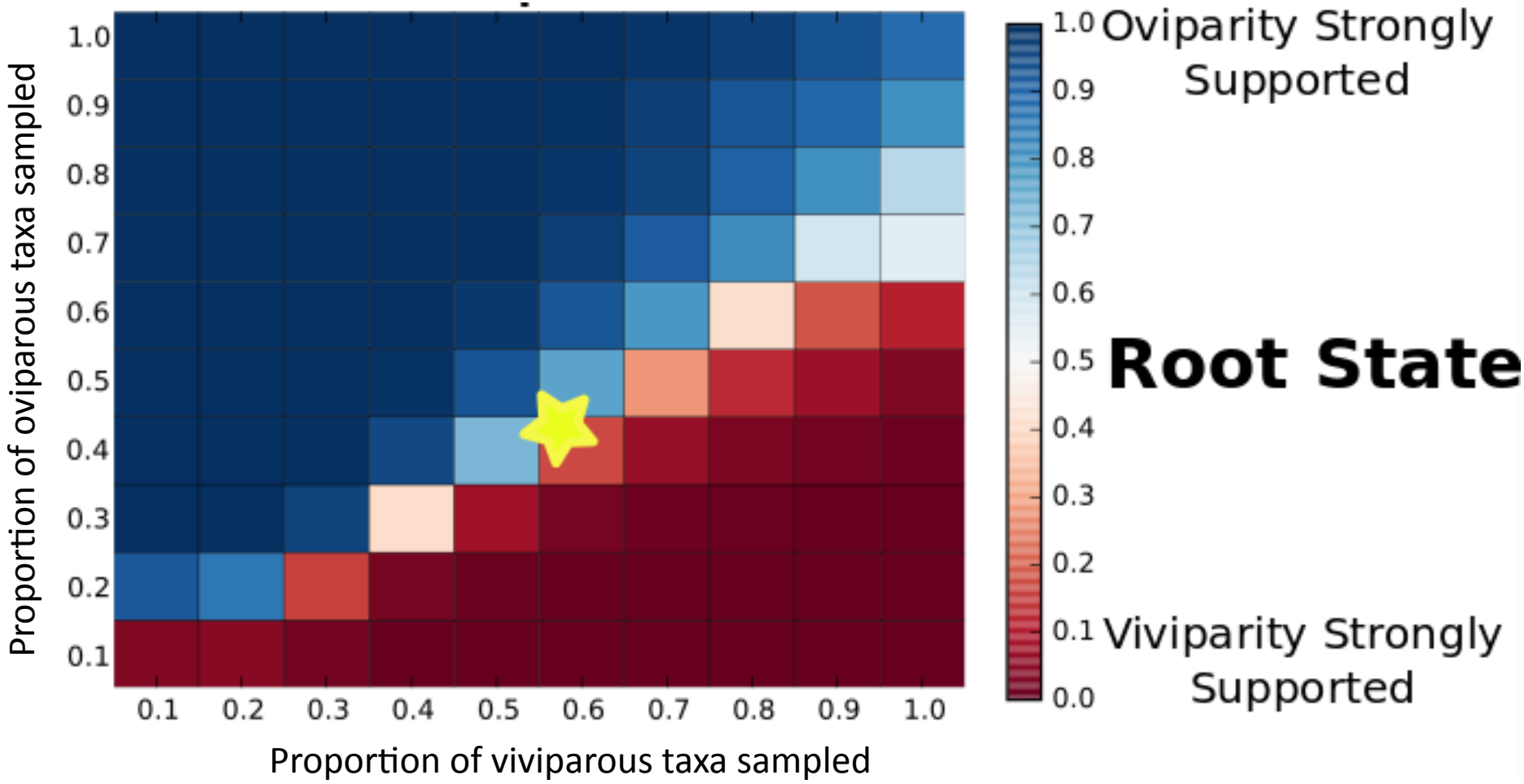
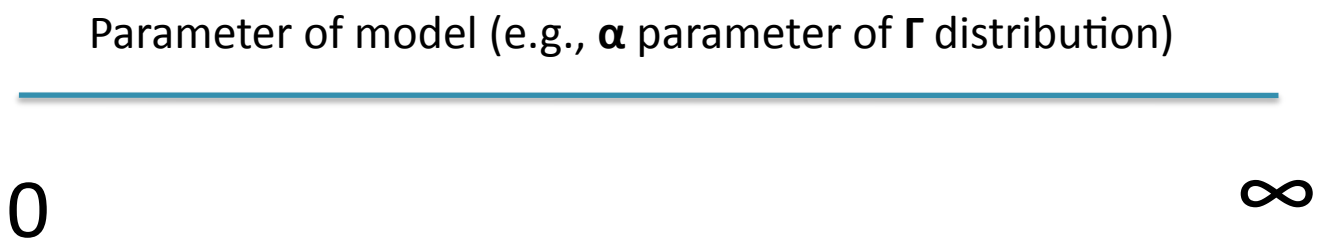
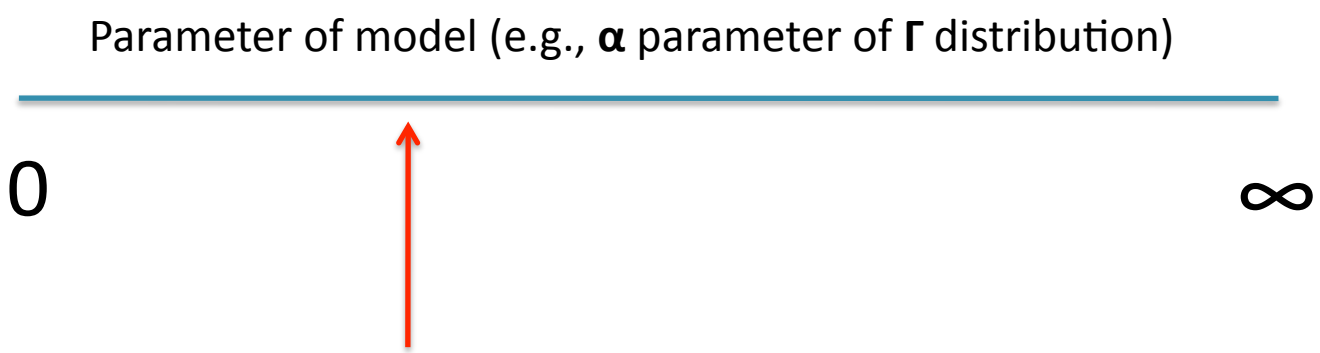


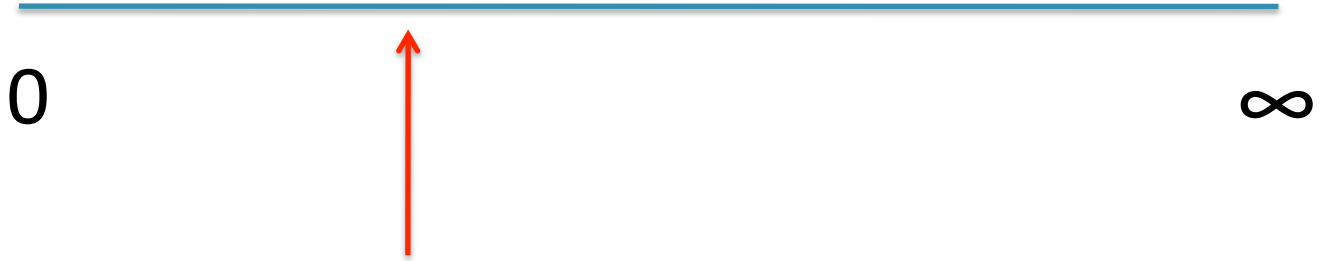
Figure from Wright et al., 2015


Parameter of model (e.g., α parameter of Γ distribution)
0
∞



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

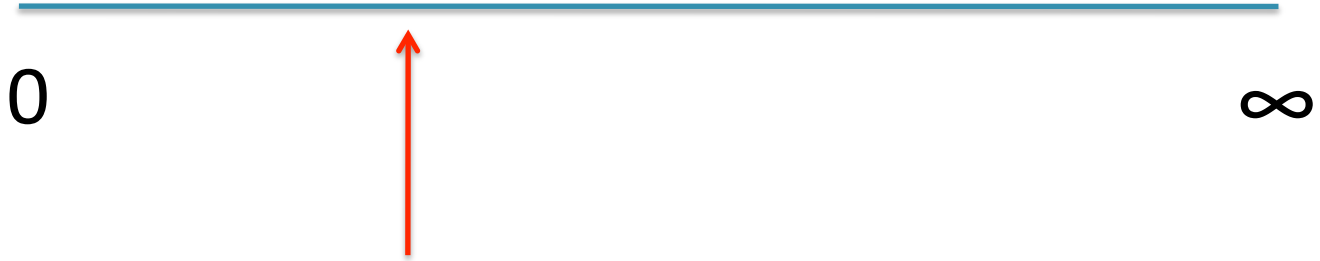
Parameter of model (e.g., **α** parameter of **Γ** distribution)



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

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

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

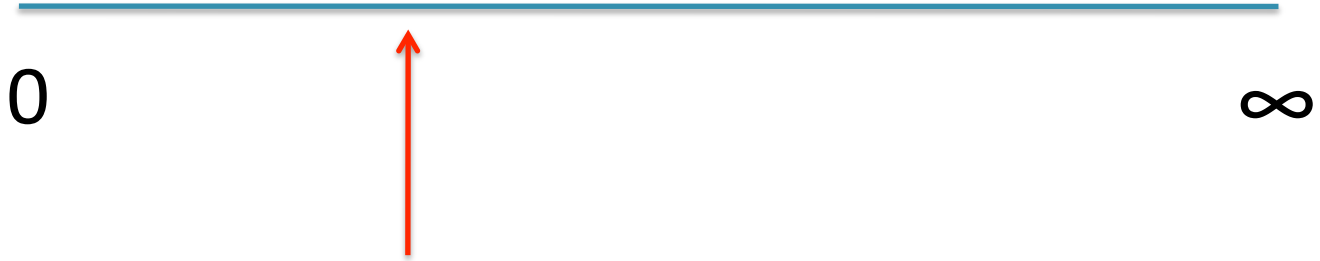


Assumed or estimated value of parameter leads to inference **Red**.
How confident should we be in answer **Red**?

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

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

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



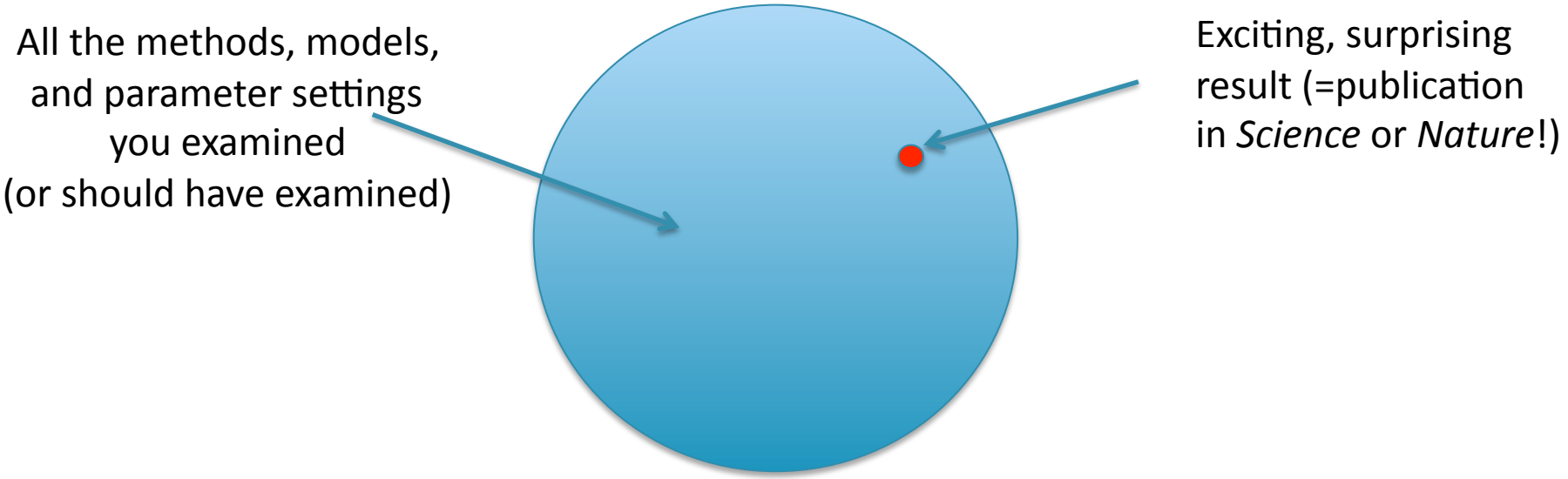
Assumed or estimated value of parameter leads to inference **Red**.
How confident should we be in answer **Red**?

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

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

Ethics of Data Presentation and Analysis

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

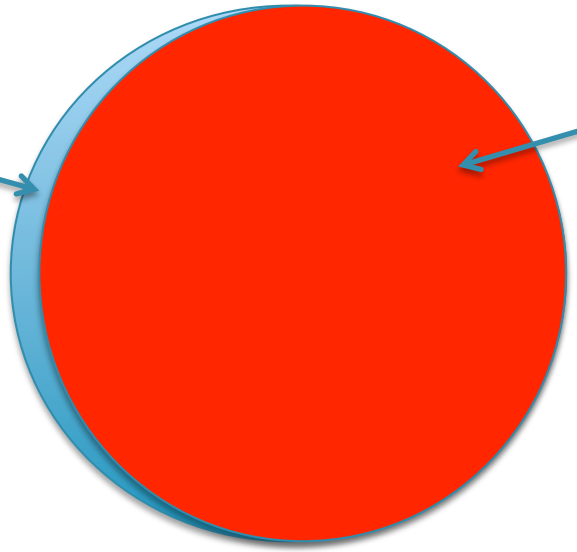


What do you publish?

Ethics of Data Presentation and Analysis

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

All the methods, models, and parameter settings you examined (or should have examined)



Exciting, surprising result (=publication in *Science* or *Nature*!)

What do you publish?

How can our models be improved?

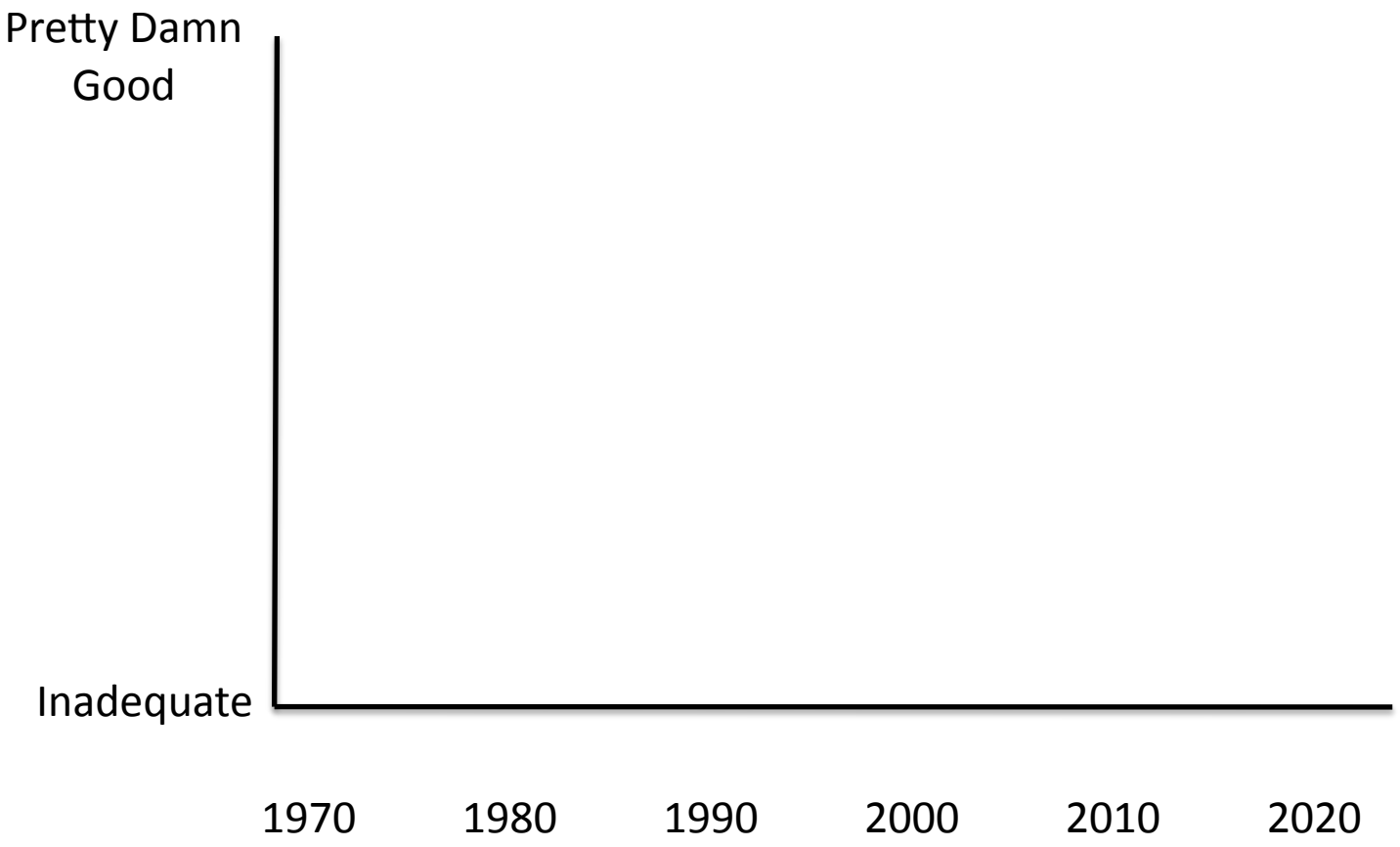
- Pretty much done:

  Simple reversible models of nucleotide substitution under stationary conditions (GTR family of models) and bifurcating evolutionary lineages

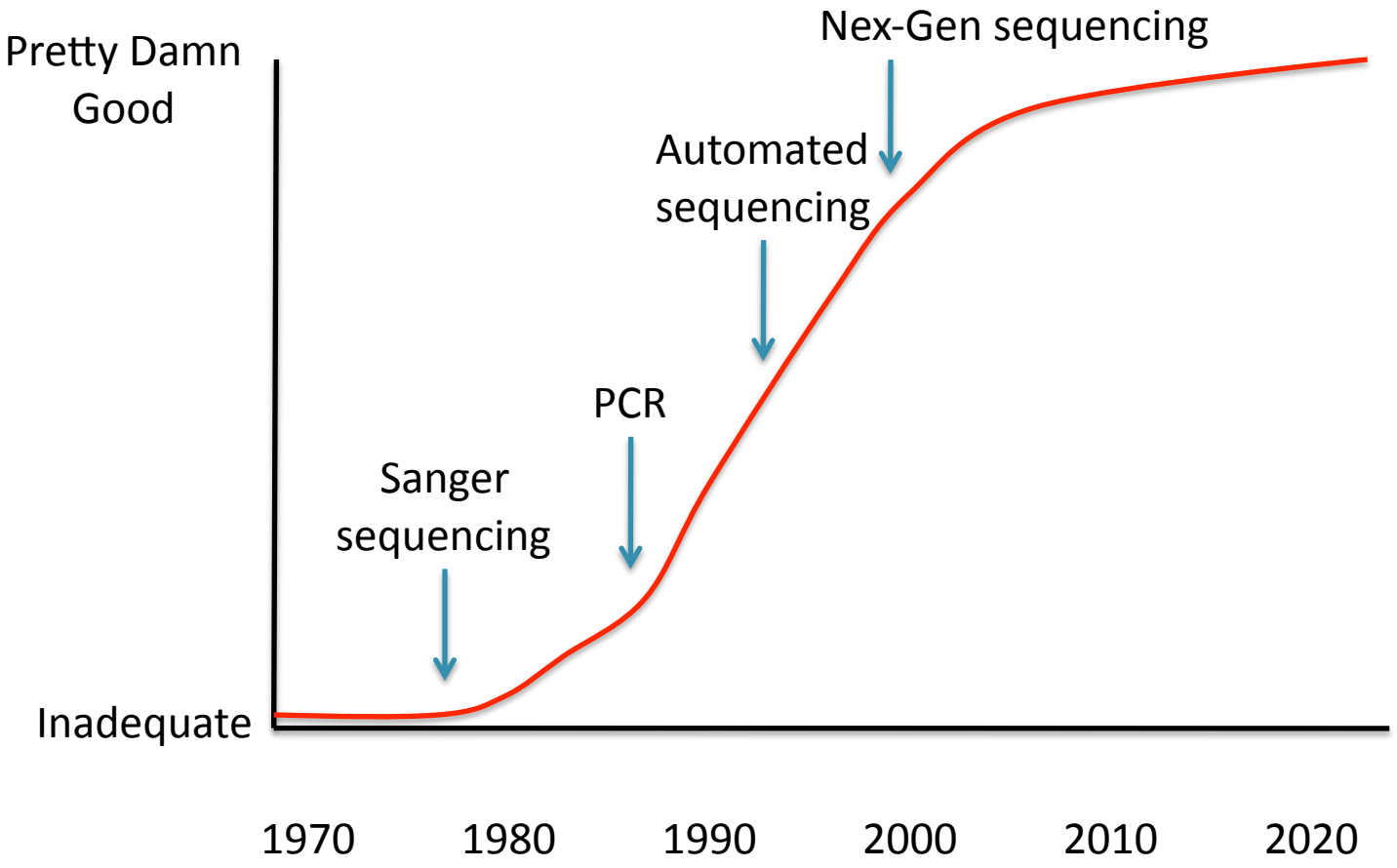
How can our models be improved?

Still lots of work to do:

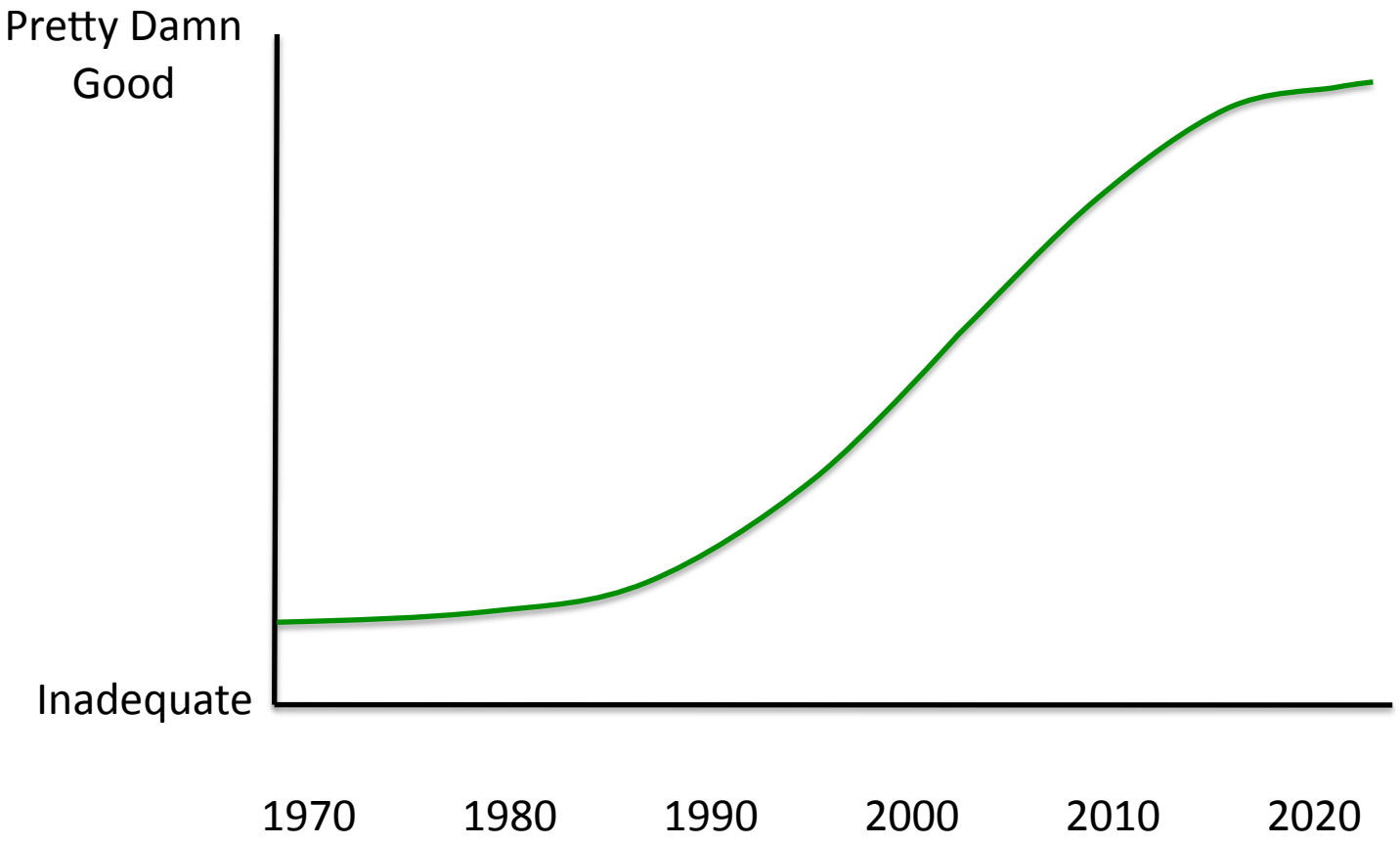
- Non-stationary models
- Codon biases
- Interactions among sites and genes
- Orthology, paralogy, and xenology
- Alignment within homologous genes
- Reticulation, lateral gene transfer
- Models for morphology and non-sequence data
- Methods for assessing confidence in our solutions


Pretty Damn Good
Inadequate
1970
1980
1990
2000
2010
2020

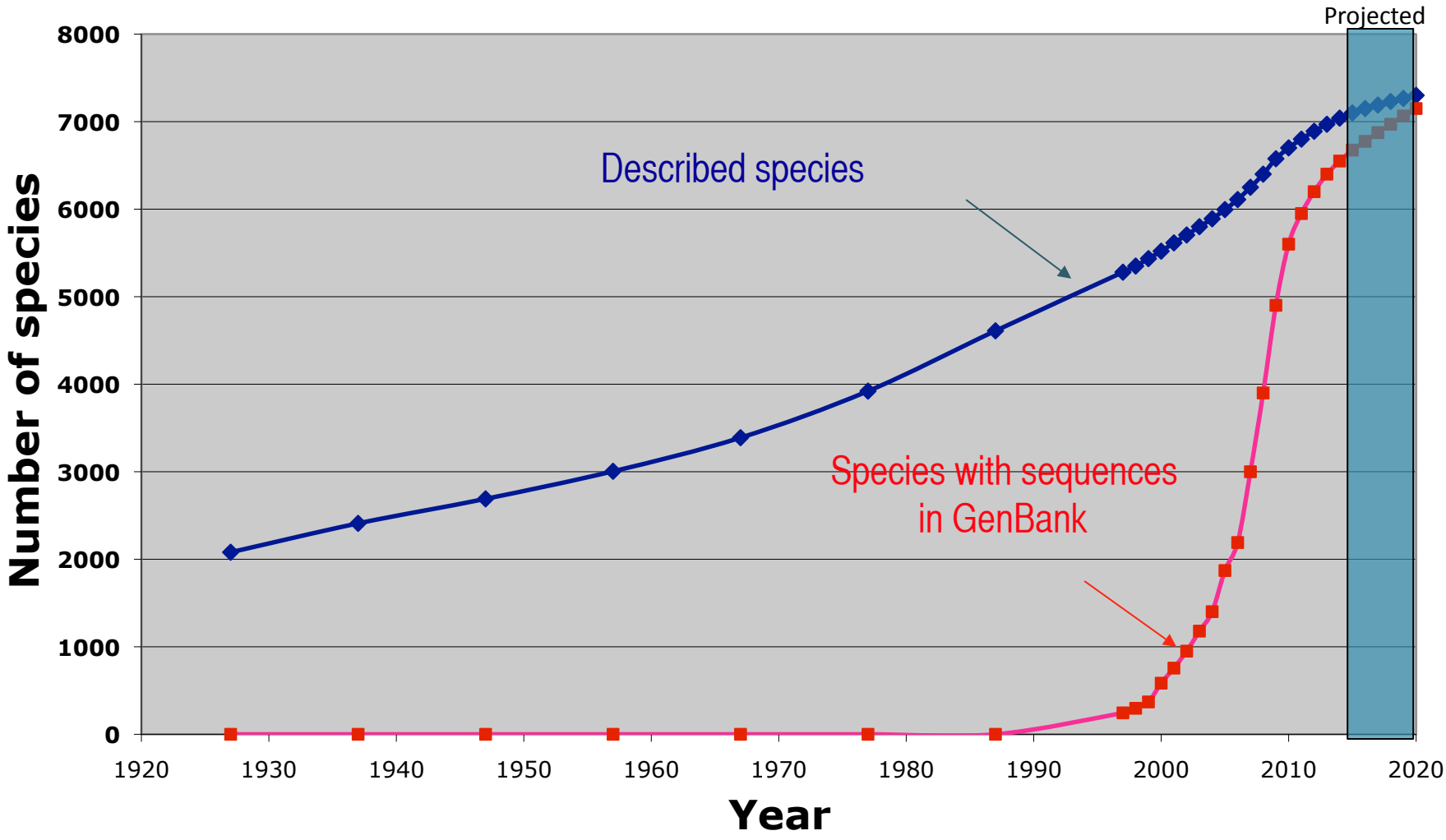
Sequence length and number of loci



Taxon sampling
(at least for some genes in some taxa)

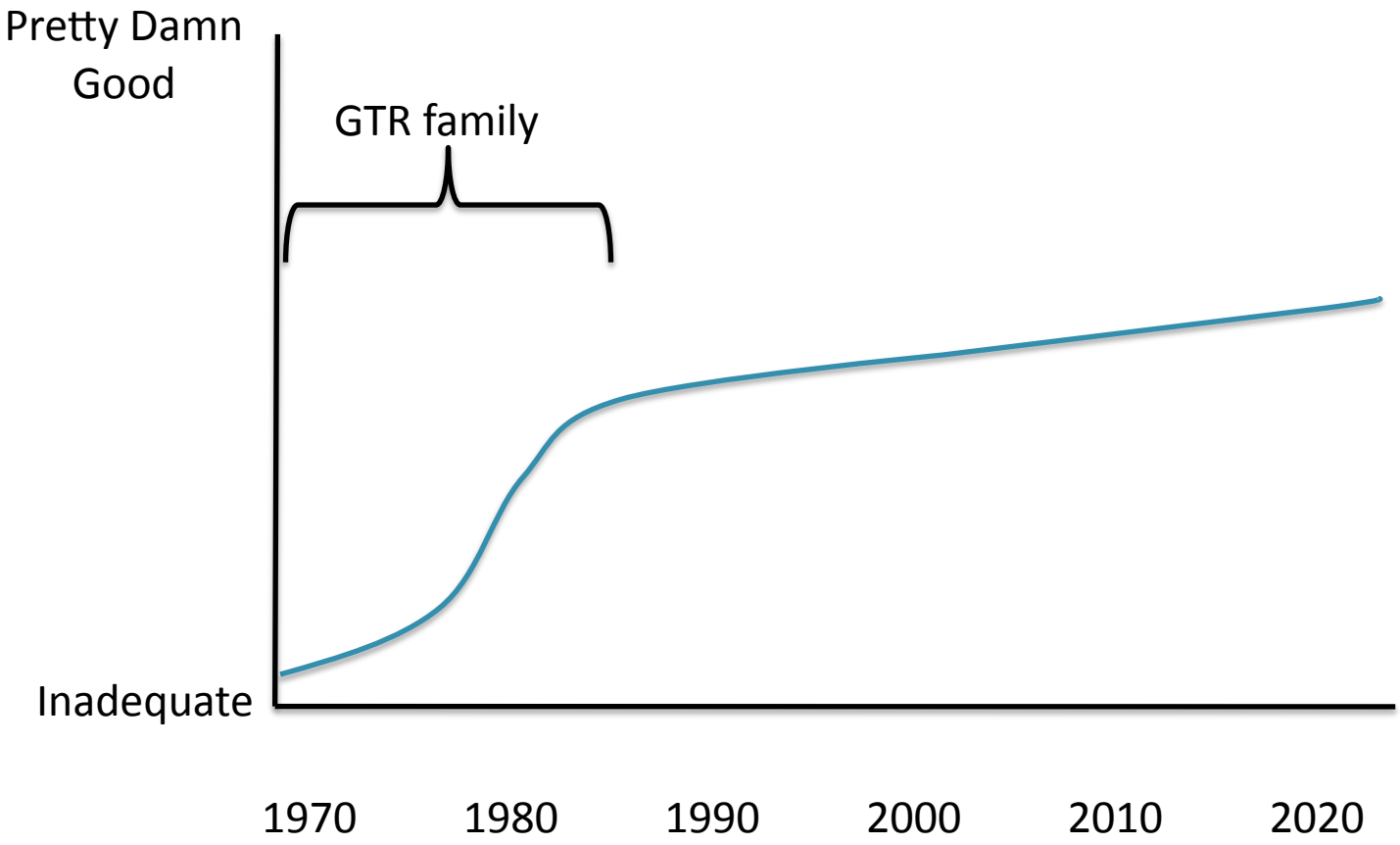


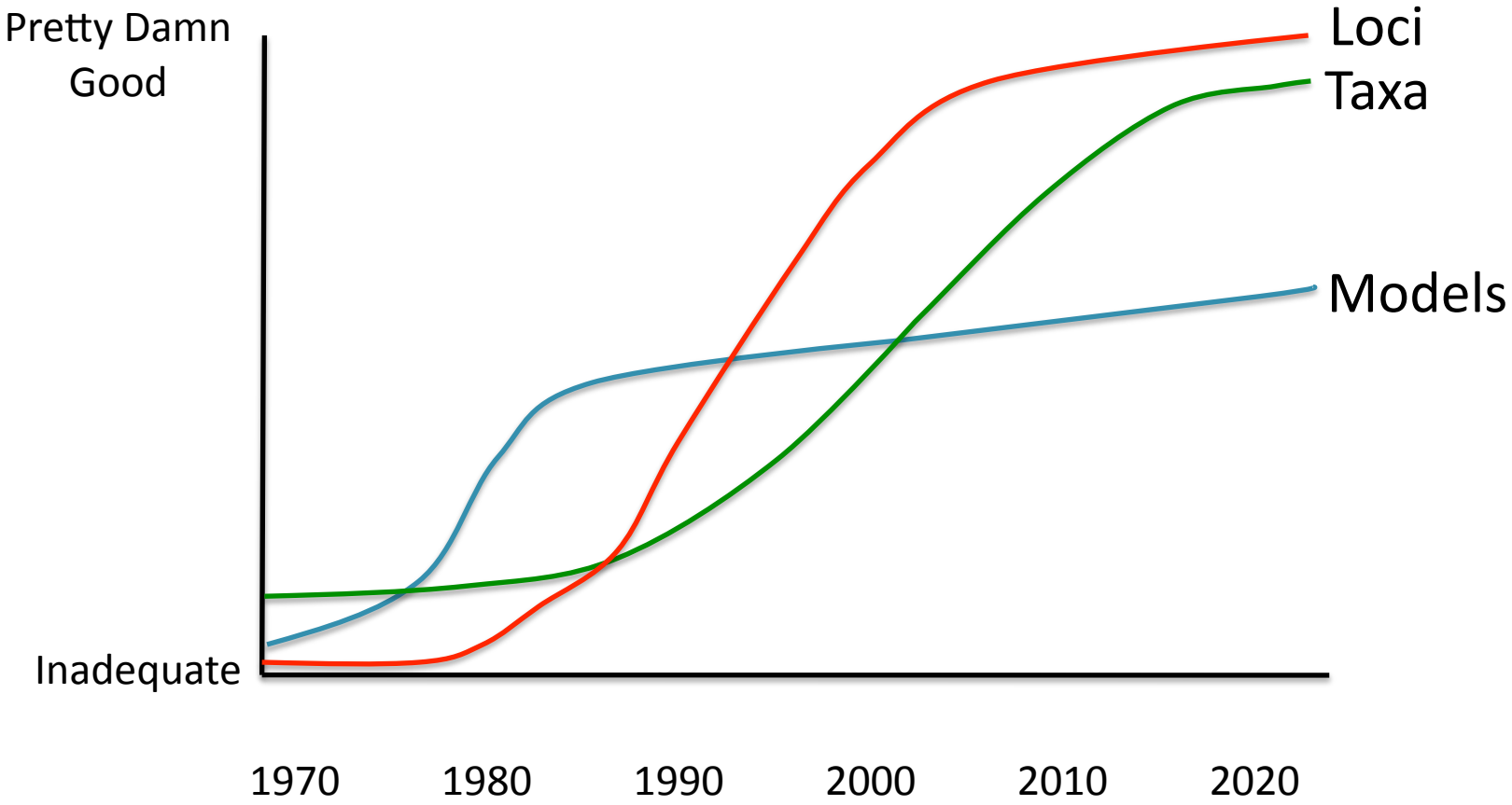
Amphibian species and sequences



From Hillis, 2010. *in* Evolution Since Darwin: The First 150 Years, Sinauer.

Models of sequence evolution




Pretty Damn Good
Inadequate
Loci
Taxa
Models
1970
1980
1990
2000
2010
2020

Do we need more data, or better analyses (models and methods)?

- The short answer is “yes, we need both”
- 1970s and 1980s: we badly needed more data
- 1990s and 2000s: we badly needed denser sampling of taxa (also more loci)
- In the 2010s, we are finally getting some pretty good data sets. But are our models up to the task of analyzing the data we have?
- Are we over-confident in the conclusions from the models that we have?

Ethics of Data Presentation and Analysis

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

Ethics of Data Presentation and Analysis

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

Ethics of Data Presentation and Analysis

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

Ethics of Data Presentation and Analysis

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

An Introduction to Phylogenetic Methods

- An optimality criterion defines how we measure the fit of data to a given solution

- Tree-searching is a separate step; this is how we search through possible solutions (which we then evaluate with the chosen optimality criterion)

Phylogenetic methods

Three main classes of optimality criteria:

- Nonparametric methods: parsimony and related approaches
- Semi-parametric methods: pairwise distance approaches
- Parametric methods: Likelihood and Bayesian approaches

Advantages of each

Parsimony methods

- Widely applicable to many discrete data types (often used to combine analyses of different data types)
- Requires no explicit model of evolutionary change
- Computationally relatively fast
- Relatively easy interpretation of character change
- Perform well with many data sets

Pairwise distance methods

- Can be used with pairwise distance data (e.g., non-discrete characters)
- Can incorporate an explicit model of evolution in estimation of pairwise distances
- Computationally relatively fast (especially for single-point estimates)

Likelihood-based methods

- Fully based on explicit model of evolution
- Most efficient method under widest set of conditions
- Consistent (converges on correct answer with increasing data, as long as assumptions are met)
- Most straightforward statistical assessment of results; probabilistic assessment of ancestral character states

Disadvantages of each

Parsimony methods

- No explicit model of evolution; often less efficient
- Nonparametic statistical approaches for assessing results often have poorly understood properties
- Can provide misleading results under some fairly common conditions
- Do not provide probablistic assessment of alternative solutions

Pairwise distance methods

- Model of evolution applied locally (to pairs of taxa), rather than globally
- Statistical interpretation not straightforward
- Can provide misleading results under some fairly common conditions (but not as sensitive as parsimony)
- Do not provide probablistic assessment of alternative solutions

Likelihood-based methods

- Require an explicit model of evolution, which may not be realistic or available for some data types
- Computationally most intense

Parsimony Criterion

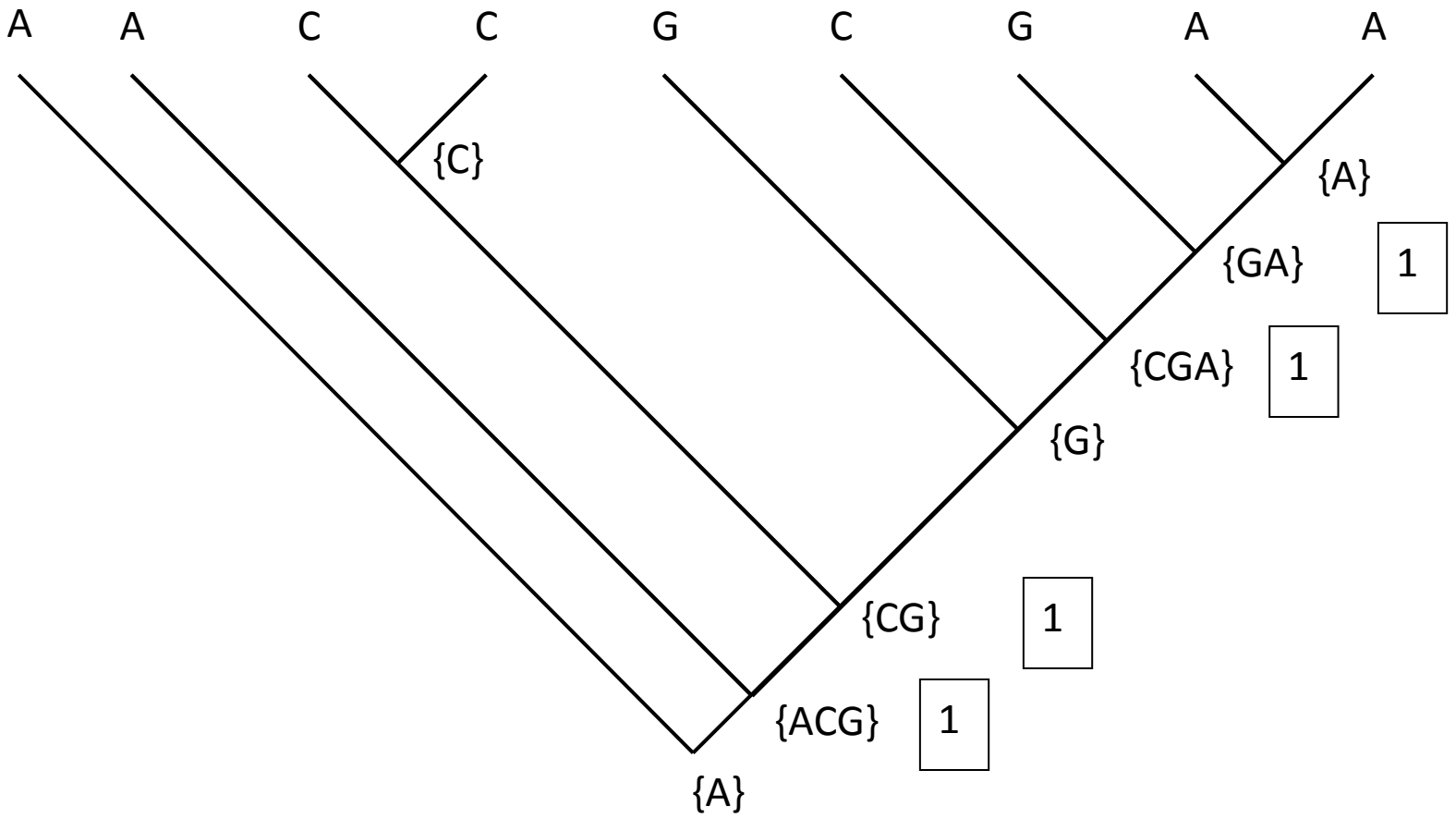
- Under the parsimony criterion, the optimal tree (the shortest or minimum-length tree) is the one that minimizes the sum of the lengths of all characters in terms of evolutionary steps (a step is a change from one character-state to another).
- For a given tree, find the length of each character, and sum these lengths; this is the tree length.
- The tree with the minimum length is the most-parsimonious tree.
- The most parsimonious tree provides the **best fit** of the data set under the parsimony criterion.

Optimal versus True Tree

There is no guarantee that any criterion will necessarily identify the “true” tree. These are simply criteria for choosing which tree best fits a dataset

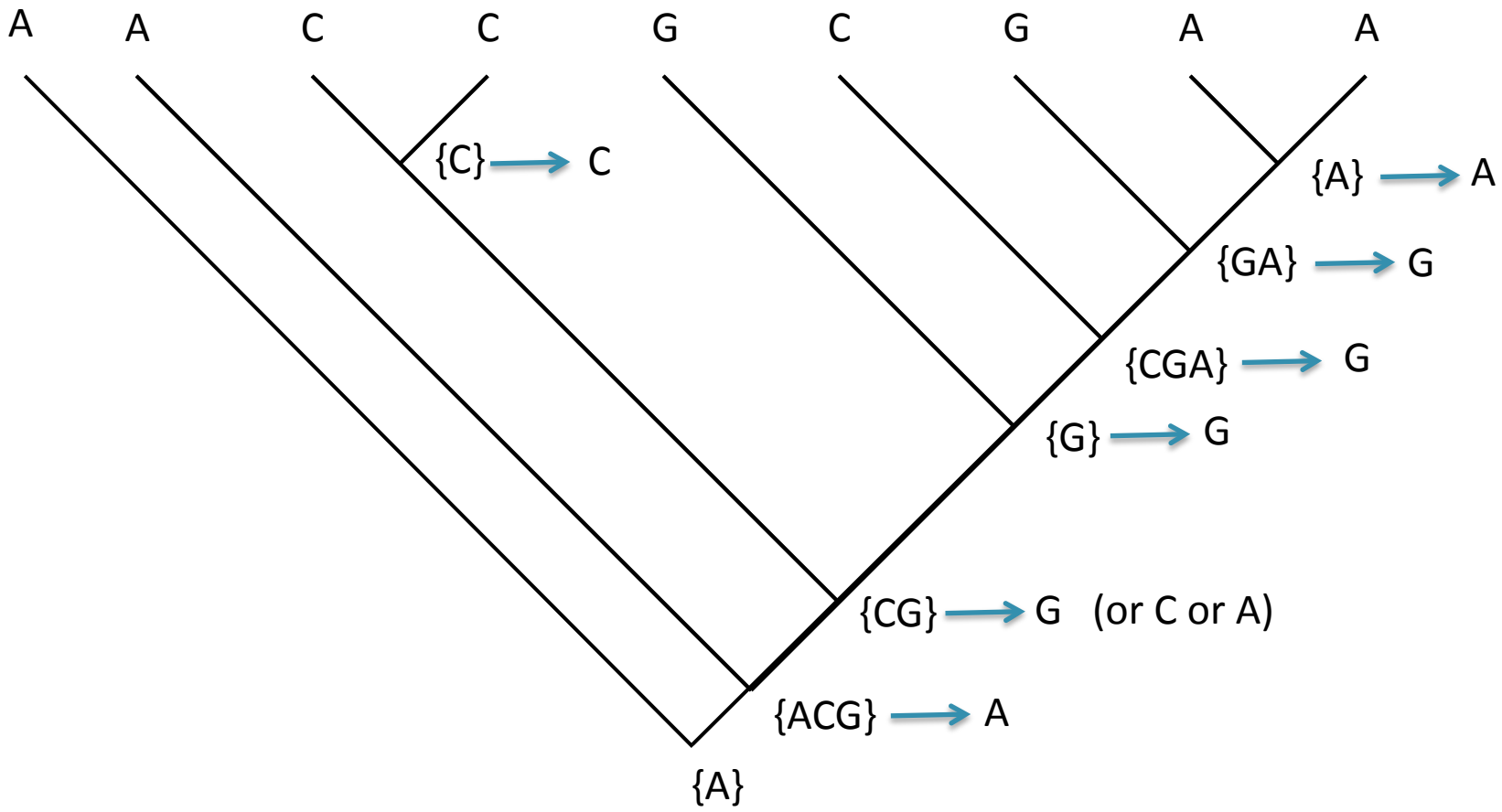
Optimization

- In parsimony, this involves minimizing the number of changes of a character across a tree (the *length* of the character)
- The optimization involves estimating the state at all internal nodes.
- It is possible for a character to have more than one best optimization.



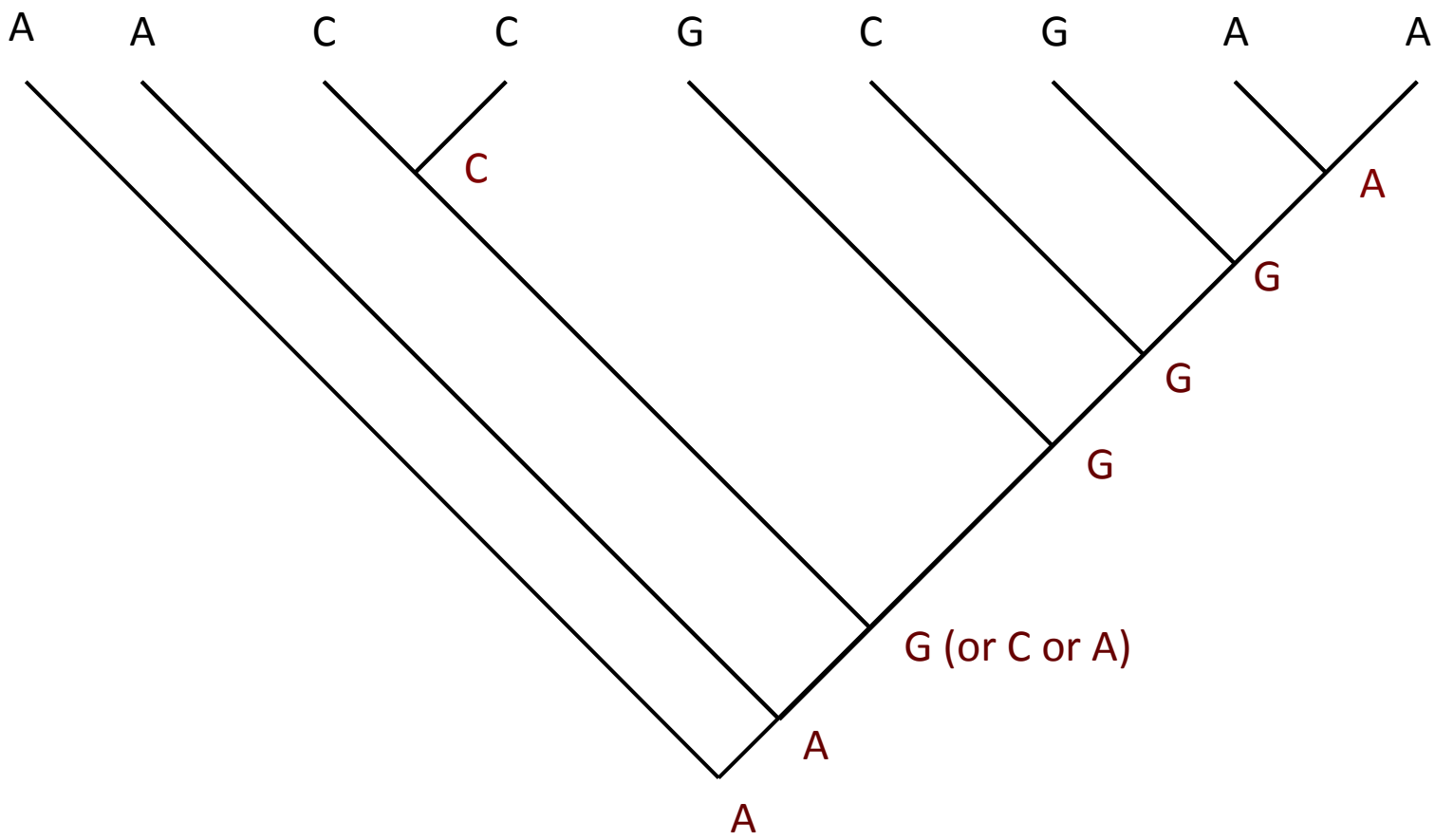
Downpass (postorder traversal)

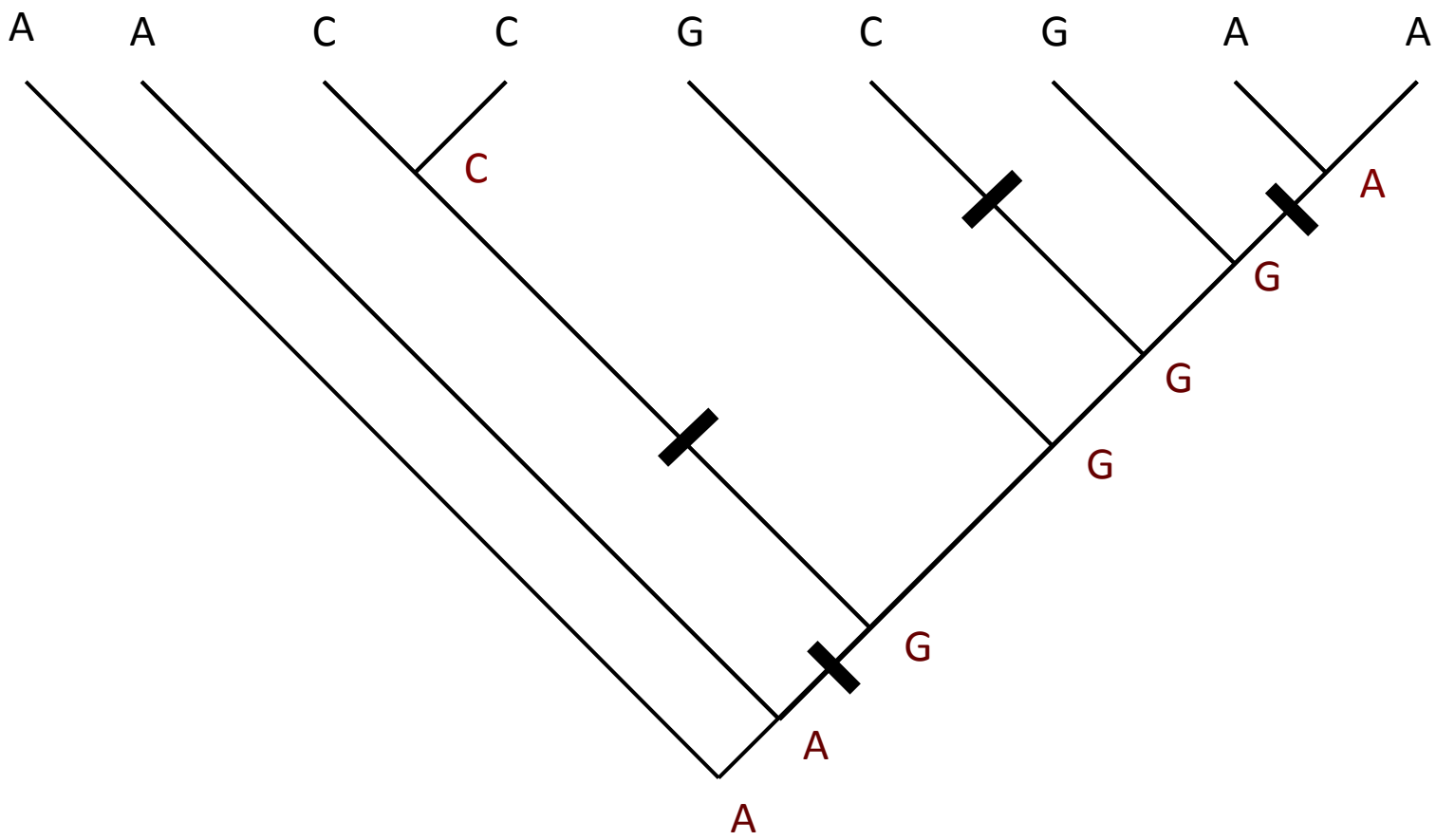
Length = 4

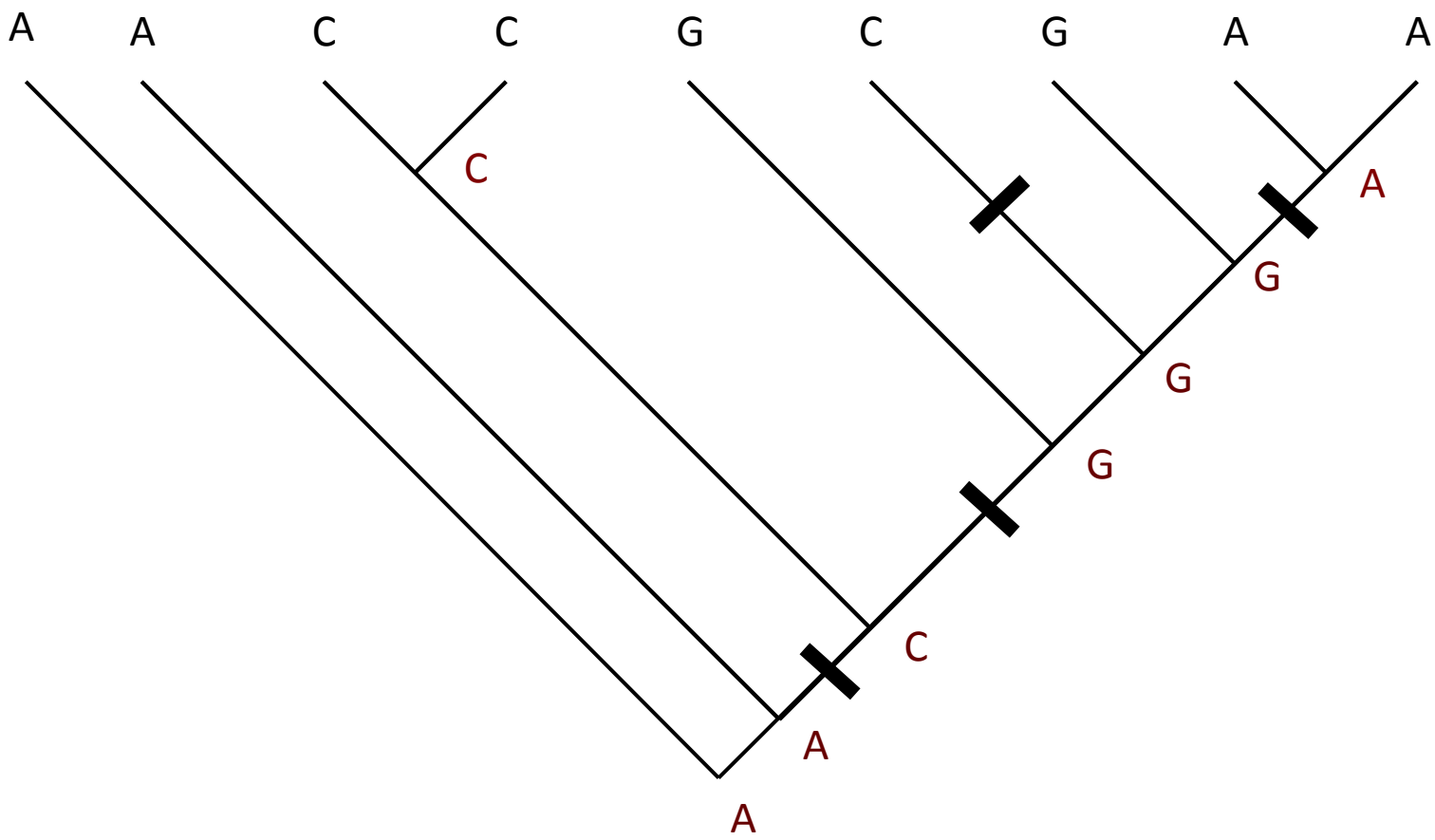


Up-Pass (preorder traversal)

Length = 4


A
A
C
C
G
C
G
A
A
C
A
G
G
G (or C or A)
A
A


A
A
C
C
G
C
G
A
A
C
A
G
G
G
G
A


A
A
C
C
G
C
G
A
A
C
A
C
G
G
A
A

Weighted Parsimony

- Transformations among character-states do not need to be weighted equally
- Can account for different weights between transitions and transversions, for example
- A way to approximately incorporate some aspects of models of evolution

Pairwise distances

- Distances summarize character differences between objects (terminals, taxa).
- Pairwise distances are computationally quick to calculate.
- Character differences cannot be recovered from distances, because different combinations of character states can yield the same distance.
- Characters cannot be compared individually, as in discrete character analyses.
- The distances in a matrix are not independent of each other, and errors are often compounded in fitting distances to a tree.

	Characters				
Taxa	1	2	3	4	5
one	A	G	C	G	A
two	A	G	C	G	T
three	C	T	C	G	T
four	C	T	C	A	A

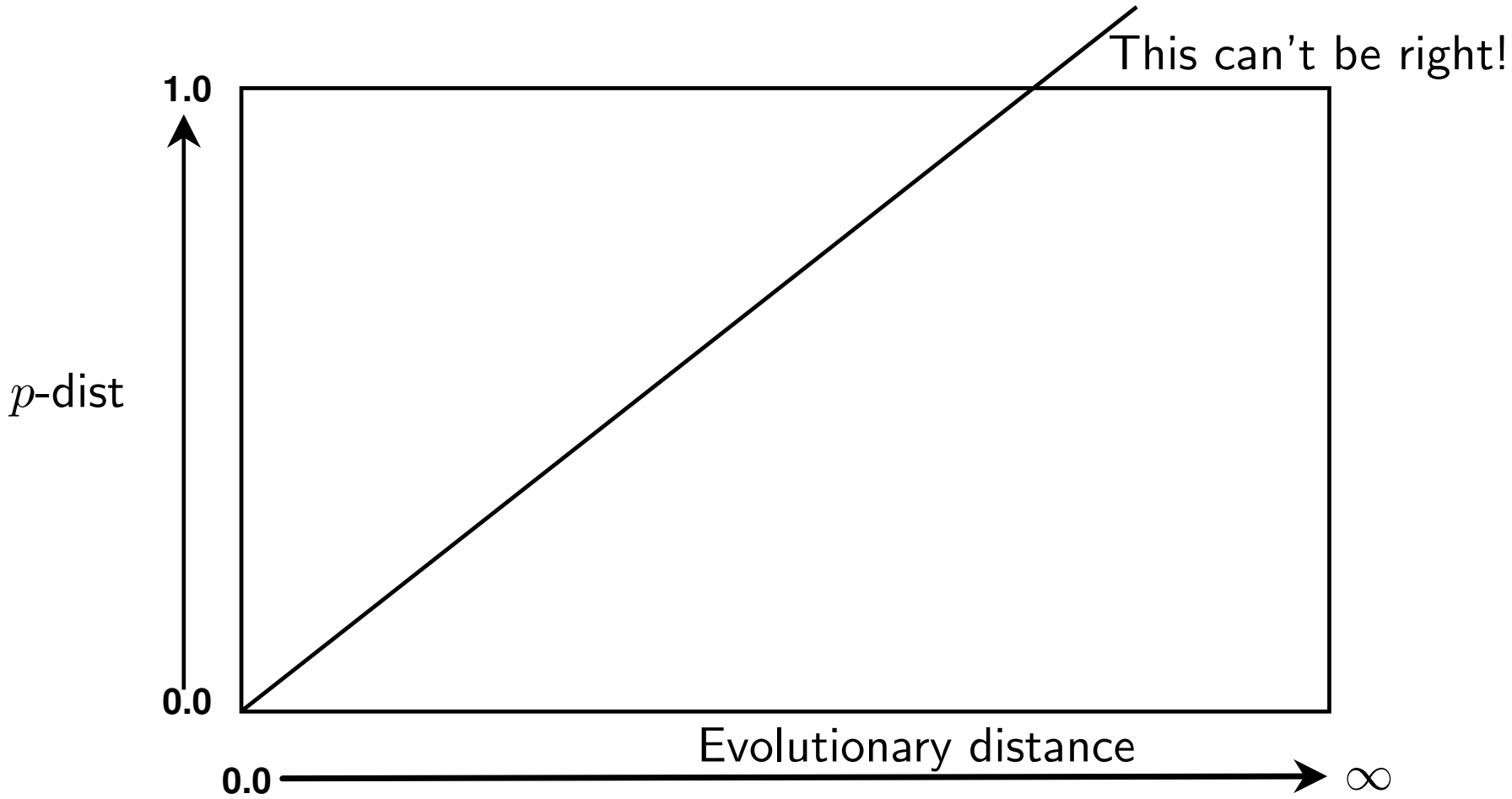
Start with a data matrix of the usual form

	Characters				
Taxa	1	2	3	4	5
one	A	G	C	G	A
two	A	G	C	G	T
three	C	T	C	G	T
four	C	T	C	A	A

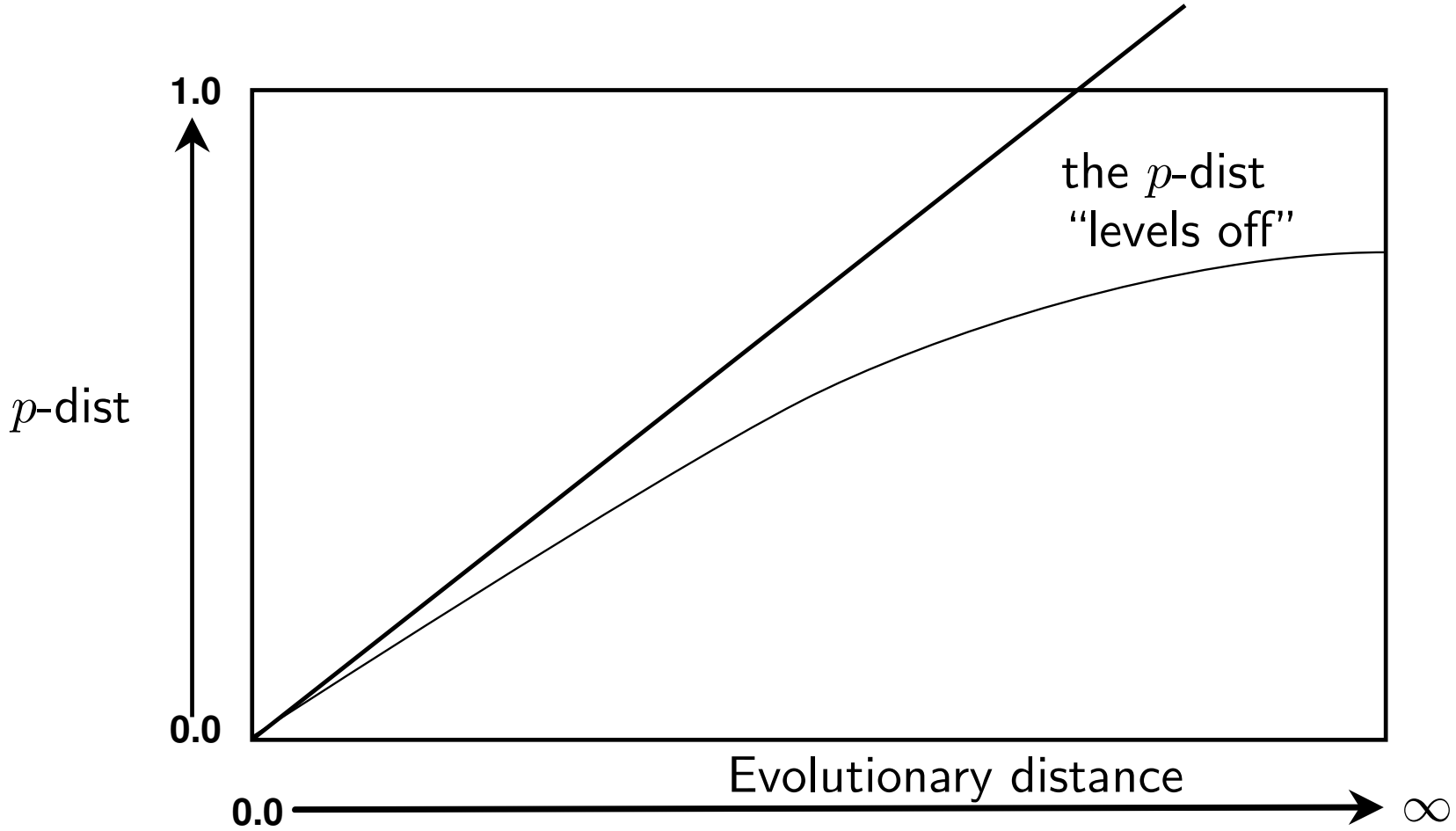
	one	two	three	four
one	-	.2	.6	.6
two		-	.4	.8
three			-	.4
four				-

Compute a distance matrix of observed *proportional* distances

Intuition of sequence divergence vs evolutionary distance



Sequence divergence vs evolutionary distance



	Characters				
Taxa	1	2	3	4	5
one	A	G	C	G	A
two	A	G	C	G	T
three	C	T	C	G	T
four	C	T	C	A	A

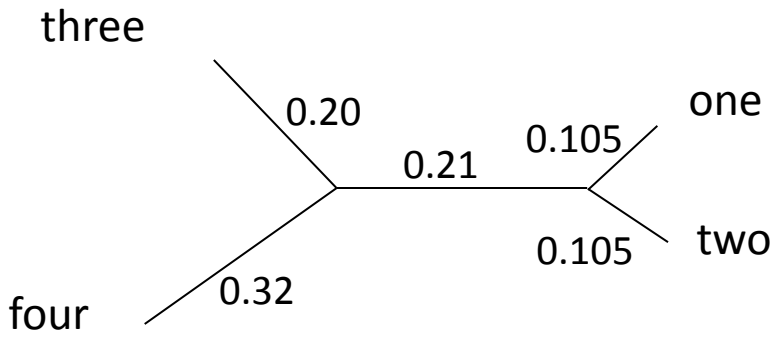
	one	two	three	four
one	-	.2	.6	.6
two		-	.4	.8
three			-	.4
four				-

Transform the distance matrix (d_{ij}) into a matrix of evolutionary (corrected) distances, using a model of evolution to account for superimposed changes (reversal, convergences, multiple changes, etc.). This is where the parametric model is applied (to many separate 2-taxon "trees").

Model of evolution

	one	two	three	four
one	-	.21	.63	.63
two		-	.42	.85
three			-	.42
four				-

Find the tree and branch lengths that result in the best match (using an objective function) between the corrected distance matrix (d_{ij}) and the patristic distance matrix (p_{ij}) (the matrix of path-length distances)



	one	two	three	four
one	-	.21	.515	.635
two		-	.515	.635
three			-	.52
four				-

	one	two	three	four
one	-	.21	.63	.63
two		-	.42	.85
three			-	.42
four				-

Optimality Criteria using Pairwise Distances

- Two commonly used objective functions:
  - Fitch-Margoliash
  - Minimum Evolution

- The general strategy is to find a set of patristic distances (path-length distances) for the branches so as to minimize the difference between the evolutionary distances and the patristic distances.

Pairwise Distance Methods

- Fitch-Margoliash family

$$Fit = \sum_{1 \le i < j < n} \omega_{ij} \left| d_{ij} - p_{ij} \right|^{\alpha}$$

i = taxon i
j = taxon j, up to n
d = evolutionary distance
p = patristic or tree distance
w = weight
Exponent: 2 = least squares
1 = absolute difference

Common weights

$w_{ij} = 1$

$w_{ij} = 1/d_{ij}$

$w_{ij} = 1/d^2_{ij}$

Pairwise Distance Methods

- Minimum Evolution

$$Fit = \sum_{1 \le i < j < n} \omega_{ij} \left| d_{ij} - p_{ij} \right|^{\alpha}$$

1. Use w = 1 and alpha = 2 to fit branch lengths l_i
2. Pick the tree that minimizes the sum of the branch lengths, *L*, over all branches (this is parsimony in spirit):

$$L = \sum_{i=1}^{2n-3} l_i$$

Algorithmic Methods for Distance Trees

- UPGMA--unweighted pair-group method using arithmetic means

  (not widely used anymore…assumes

  equal rates of change)
- Neighbor-joining--an approximation method for the minimum evolution criterion

Likelihood

- Imagine that we are given a coin, and flip it *n* times, getting *h* heads: these are our data (*D*)
- We can explore various hypotheses (*H*) about the coin, which may have implicit and explicit components:

  – The coin has a p_h probability of landing on heads
  – The coin has a heads side and a tails side
  – Successive flips of the coin are independent
  – The flipping process is fair
  – etc.

Coin flipping

- The likelihood (*L*) is proportional to the probability of observing our data, given our hypothesis:

$$L(H \mid D) \propto P(D \mid H)$$

- The probability of getting the outcome *h* heads on *n* flips is given by the binomial distribution:

$$P(h,n \mid p_h) = \binom{n}{h} (p_h)^h (1 - p_h)^{n-h}$$

(Likelihood score)

Coin flipping

- The expression $\binom{n}{h}$ gives the binomial coefficients, or the number of different ways to (for example) get 4 heads in 10 flips

- We can ignore that term to look at the probability of a particular sequence of heads and tails (to make it more like the case of a particular observed sequence of nucleotides)

Coin flipping

- Let's try applying this to some data
  - Dataset 1 : A particular sequence of
    H T H T T T H T T H
- Assume a particular hypothesis
  - Try $p_h = 0.5$
- This gives us a likelihood score of

$$L(p_h = 0.5 \mid obs) = (0.5)^4 (0.5)^6 = 0.000976563$$

Coin flipping

- What does the likelihood score tell us about the likelihood of our hypothesis? In isolation, nothing, because the score is dependent on the particular data set. The score will get smaller as we collect more data (flip the coin more times).

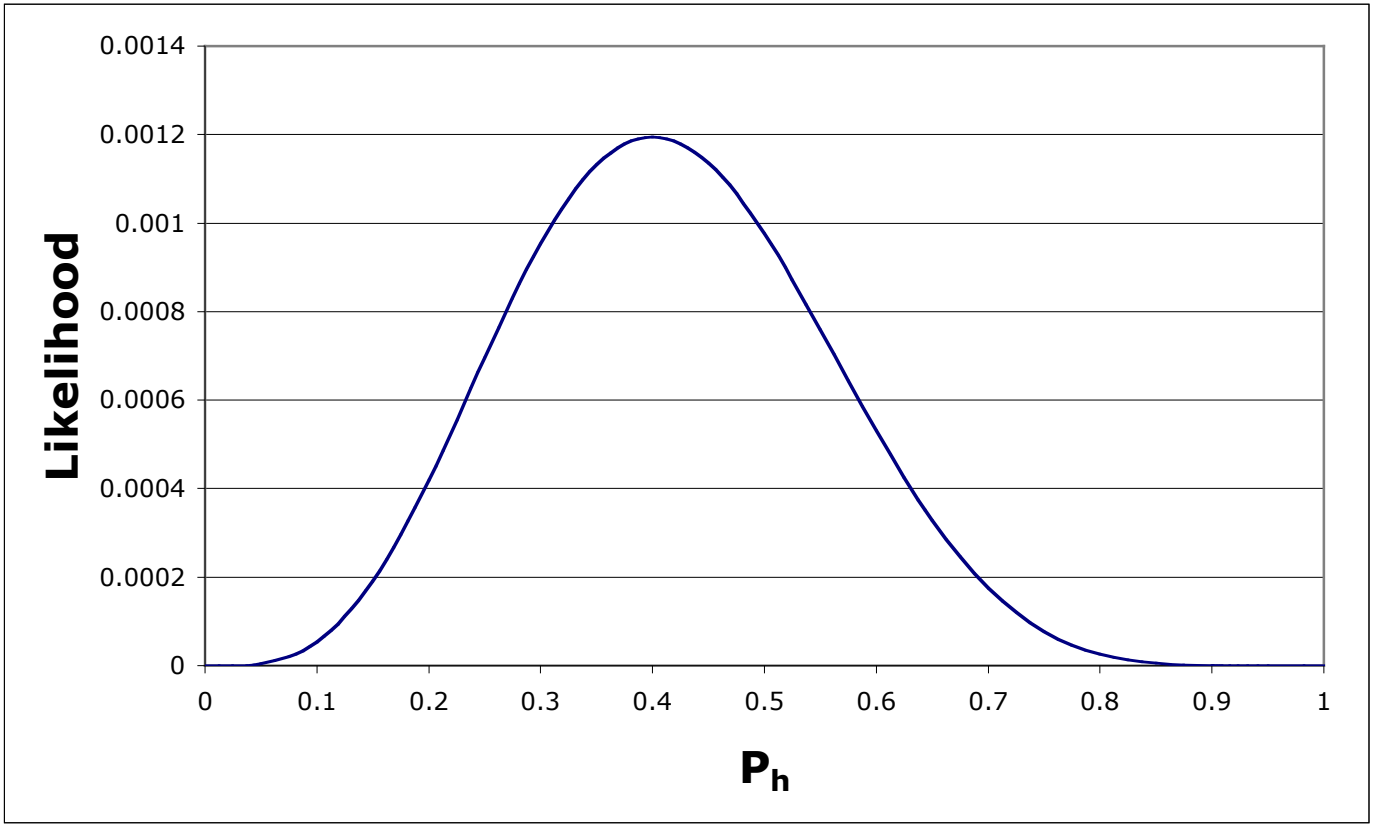
- Only the *relative* likelihood scores for various hypotheses, evaluated using the same data, are useful to us.

- What are some other models?

$$L(p_h = 0.6 \mid obs) = (0.6)^4 (0.4)^6 = 0.000530842$$

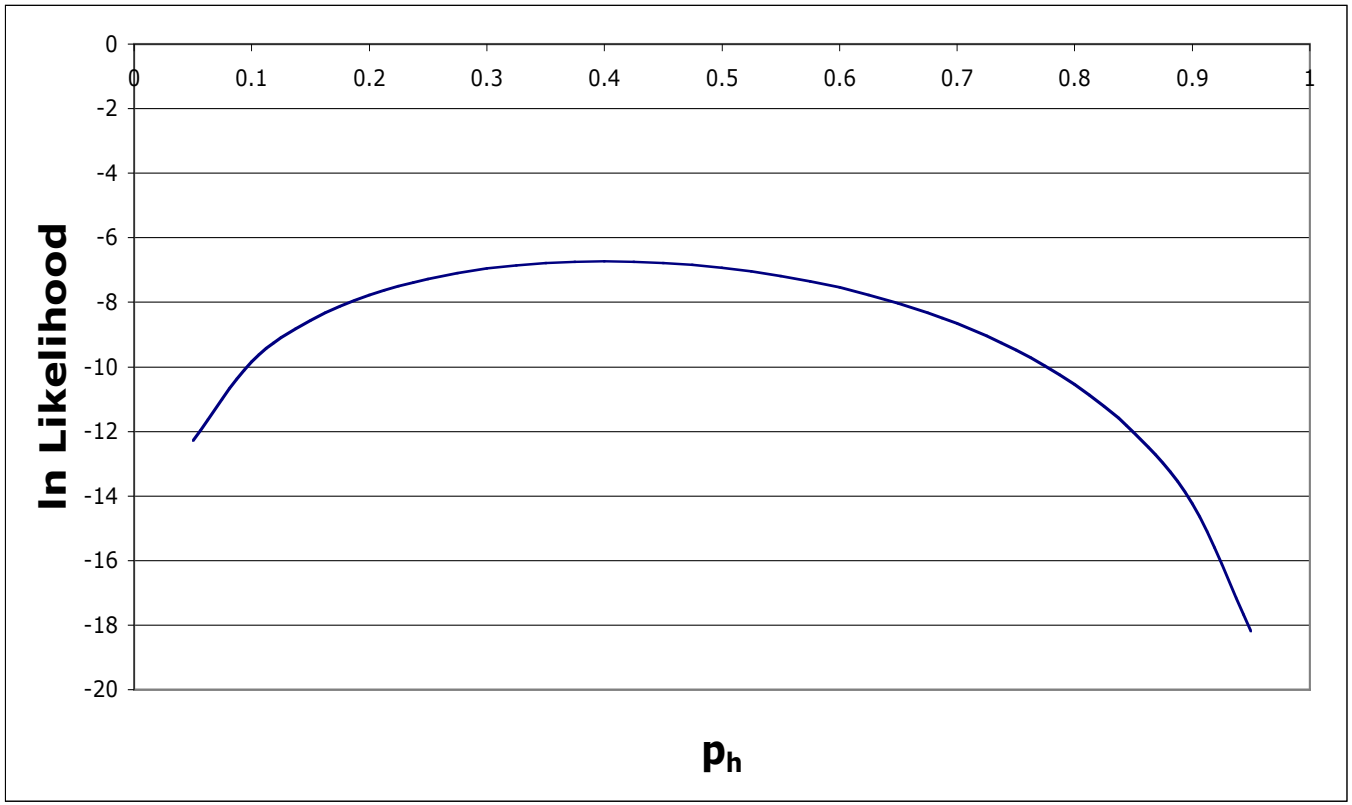
$$L(p_h = 0.4 \mid obs) = (0.4)^4 (0.6)^6 = 0.001194394$$

The likelihood surface



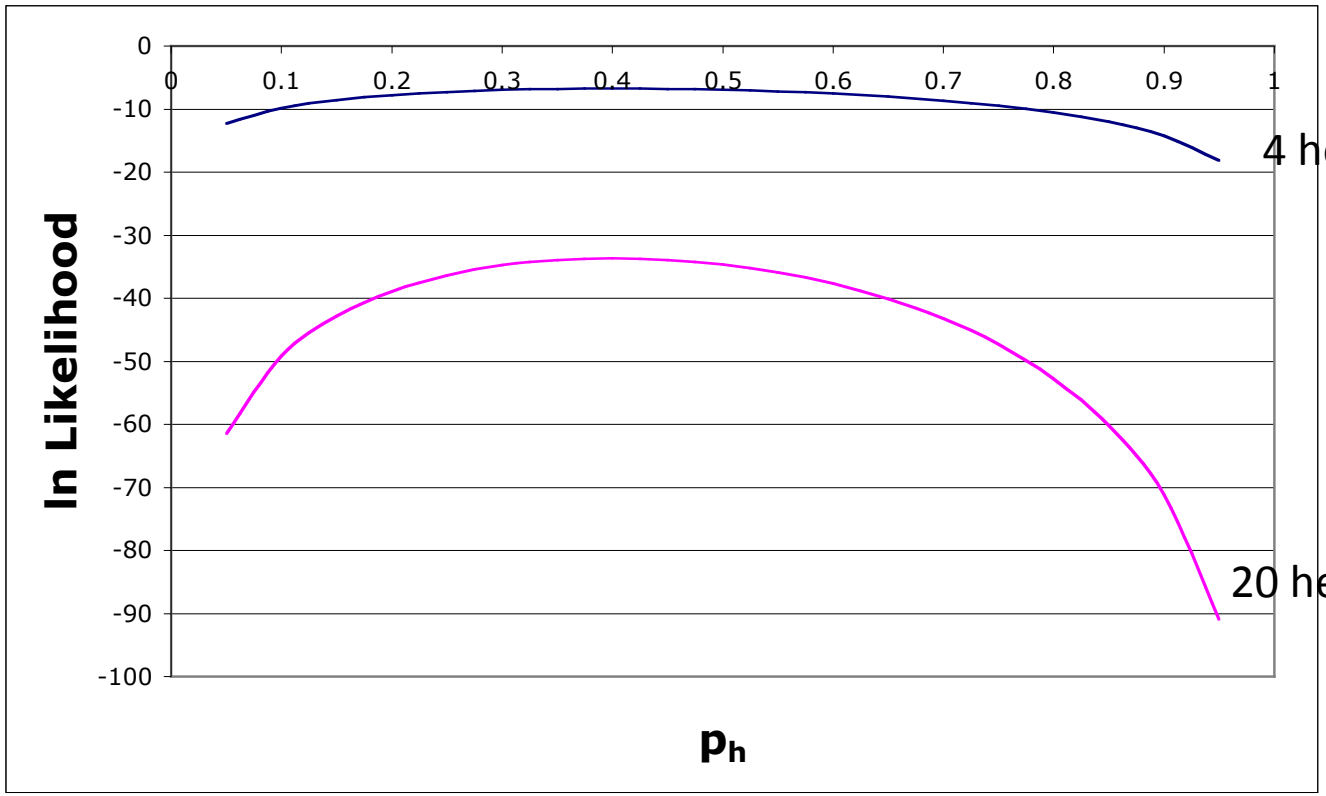
Data: H T H T T T H T T H

The log likelihood surface



Data: H T H T T T H T T H

Other data


0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
ln Likelihood
p_h
4 heads in 10 flips
20 heads in 50 flips

Likelihood

- Likelihood ($H|D$) is proportional to P ($D|H$)
- Components of the hypothesis can be explicit and implicit
- Only relative likelihoods are important in evaluating hypotheses
- The point on the likelihood curve that maximizes the likelihood score (the MLE) is our best estimate given the data at hand
- Likelihood scores shouldn't be compared between datasets
- More data lead to more peaked surfaces (i.e., better ability to discriminate among hypotheses)

Likelihood in Phylogenetics

- In phylogenetics, the data are the observed characters (e.g., DNA sequences) as they are distributed across taxa
- The hypothesis consists of the tree topology, a set of specified branch lengths, and an explicit model of character evolution.
- Calculating the likelihood score for a tree requires a very large number of calculations

Bayesian Approaches

- Take prior information into account

$$P(A|B) = \frac{P(B|A)\ \ P(A)}{P(B)}$$

(Bayes Theorem)

Reverend Thomas Bayes, 1701-1761

Bayesian Approaches

- Take prior information into account

$$P(A \mid B) = \frac{P(B \mid A) \quad P(A)}{P(B)}$$

Posterior probability

Bayesian Approaches

- Take prior information into account

$$P(A|B) = \frac{P(B|A)}{P(B)} P(A)$$

The support that *B* provides for *A*

Bayesian Approaches

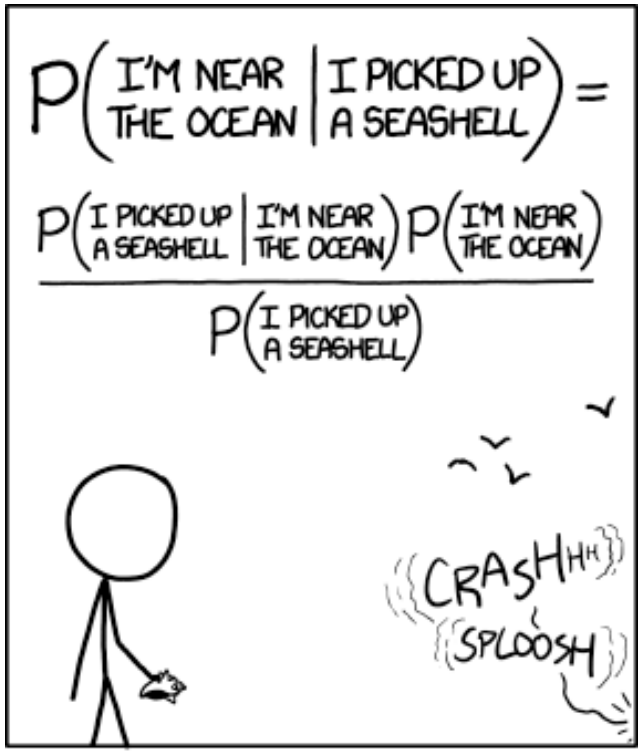
- Take prior information into account

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Prior information about *A*
(the initial expectation for *A*)

Bayesian Approaches

- An example:



What determines the accuracy of phylogenetic estimates?

- Optimality criterion (likelihood-based methods have best performance, as long as assumptions of model are met)

What determines the accuracy of phylogenetic estimates?

- Optimality criterion (likelihood-based methods have best performance, as long as assumptions of model are met)
- Model selection (finding an appropriate model of evolution for the data at hand)

What determines the accuracy of phylogenetic estimates?

- Optimality criterion (likelihood-based methods have best performance, as long as assumptions of model are met)
- Model selection (finding an appropriate model of evolution for the data at hand)
- Thoroughness of tree search (finding best solutions)

What determines the accuracy of phylogenetic estimates?

- Optimality criterion (likelihood-based methods have best performance, as long as assumptions of model are met)
- Model selection (finding an appropriate model of evolution for the data at hand)
- Thoroughness of tree search (finding best solutions)
- Sampling density of genes (more genes provide more information, and allow resolution of species trees from gene trees)

What determines the accuracy of phylogenetic estimates?

- Optimality criterion (likelihood-based methods have best performance, as long as assumptions of model are met)
- Model selection (finding an appropriate model of evolution for the data at hand)
- Thoroughness of tree search (finding best solutions)
- Sampling density of genes (more genes provide more information, and allow resolution of species trees from gene trees)
- Sampling density of taxa (more thorough taxon sampling produces better estimates of parameters and results in better estimates of trees)