27th Annual Workshop in Molecular Evolution

Co-Directors: David Hillis and Mitch Sogin

Introduction: David M. Hillis, University of Texas

Marine Biological Lab, Woods Hole

2014

(Thanks to Mark Holder, Paul Lewis, and Derrick Zwickl for some of these ideas and slides)

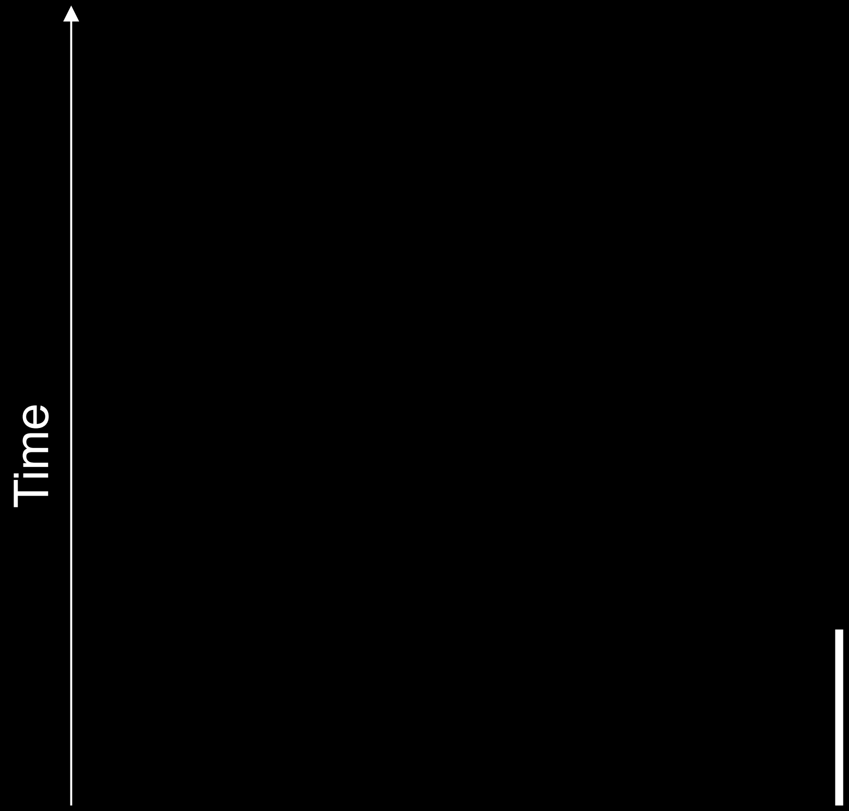
Molecular Evolution

- Using biomolecules to understand evolutionary history and evolutionary processes

- Phylogenetic trees are important to molecular evolution, because many biological phenomena of interest can be modeled as bifurcating processes

Phylogeny

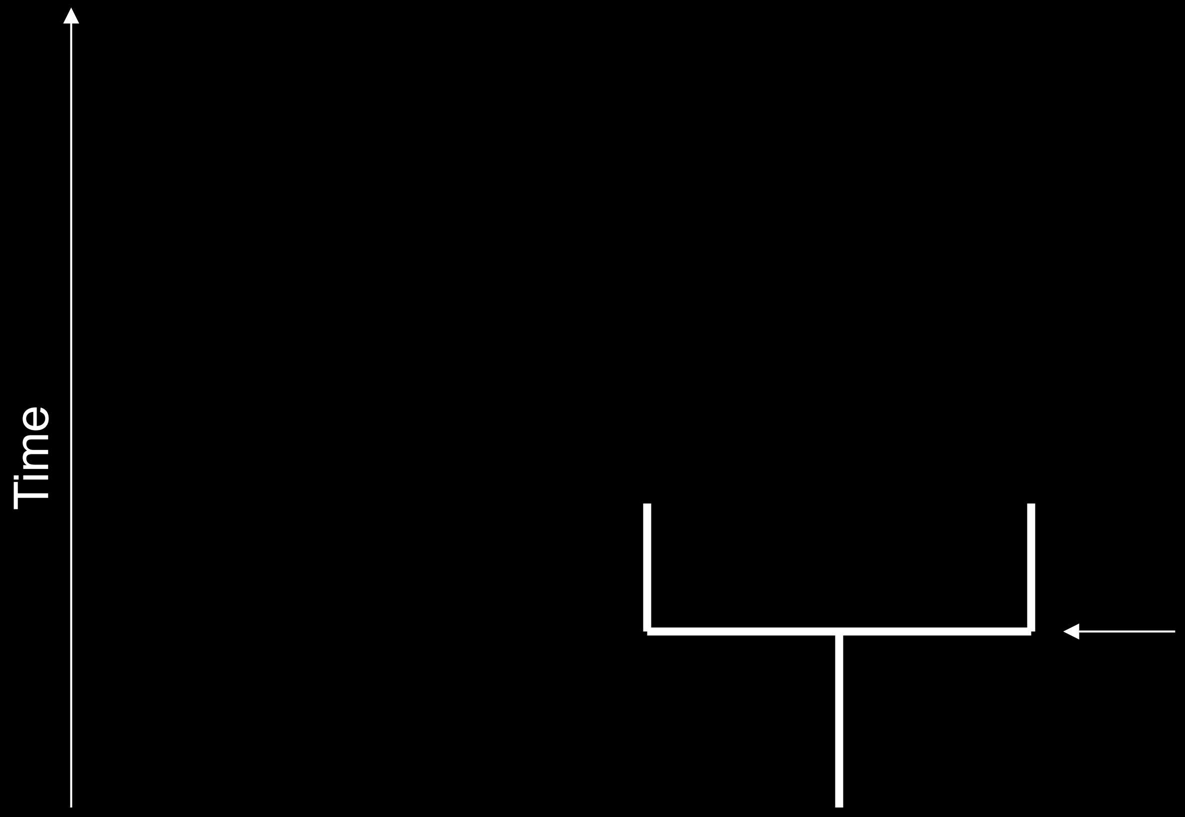
Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



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

Phylogeny

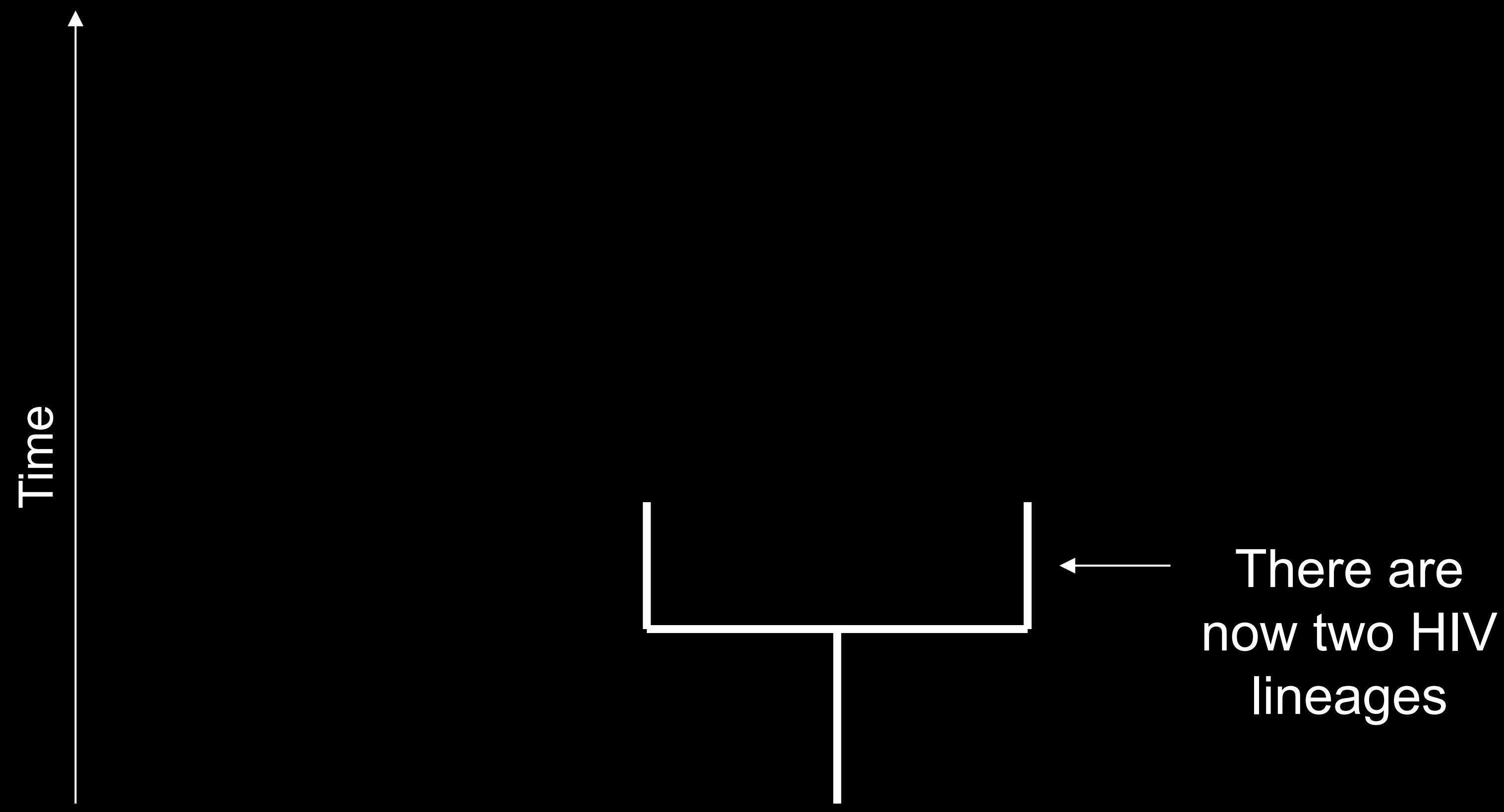
Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



One lineage splits into two

Phylogeny

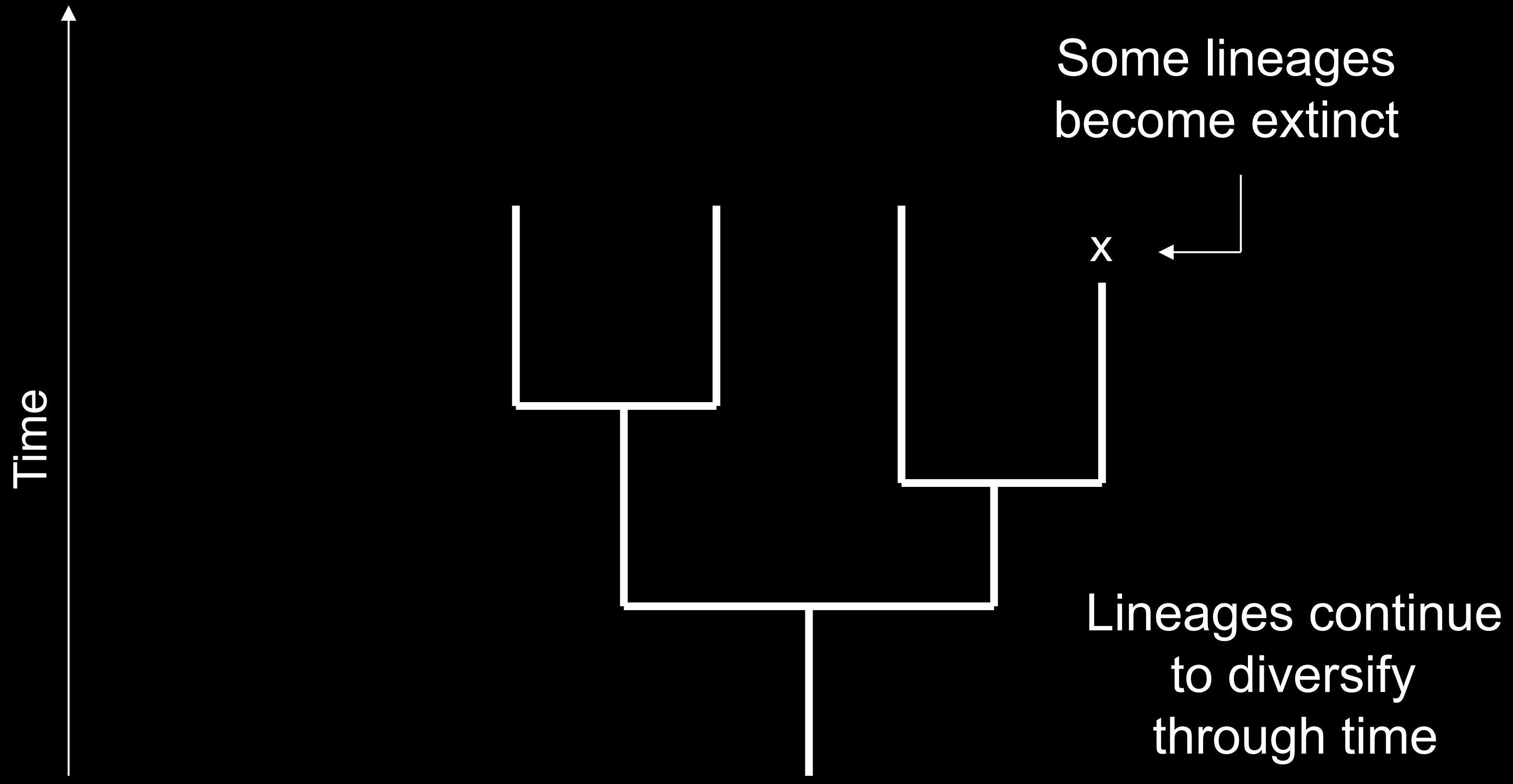
Evolutionary relationships among lineages, such as genes, individuals, populations, species, etc.



There are now two HIV lineages

Phylogeny

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

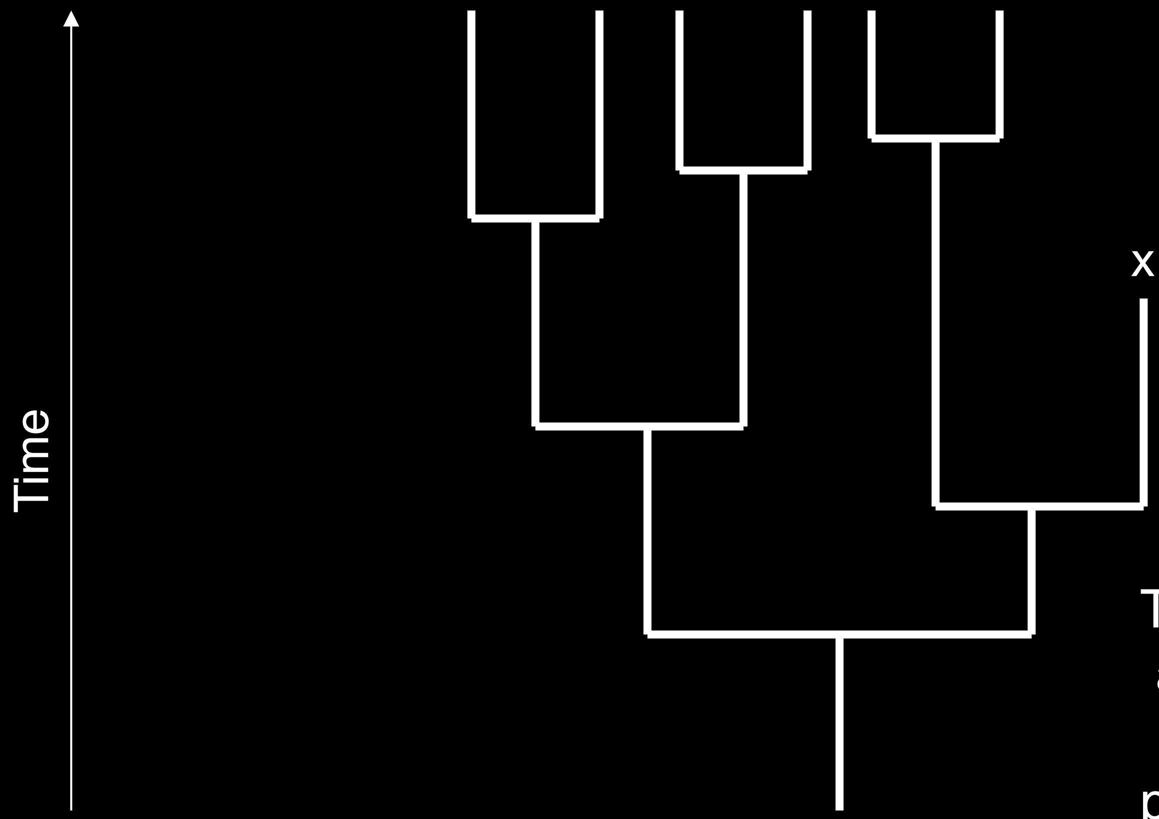


Some lineages become extinct

Lineages continue to diversify through time

Phylogeny

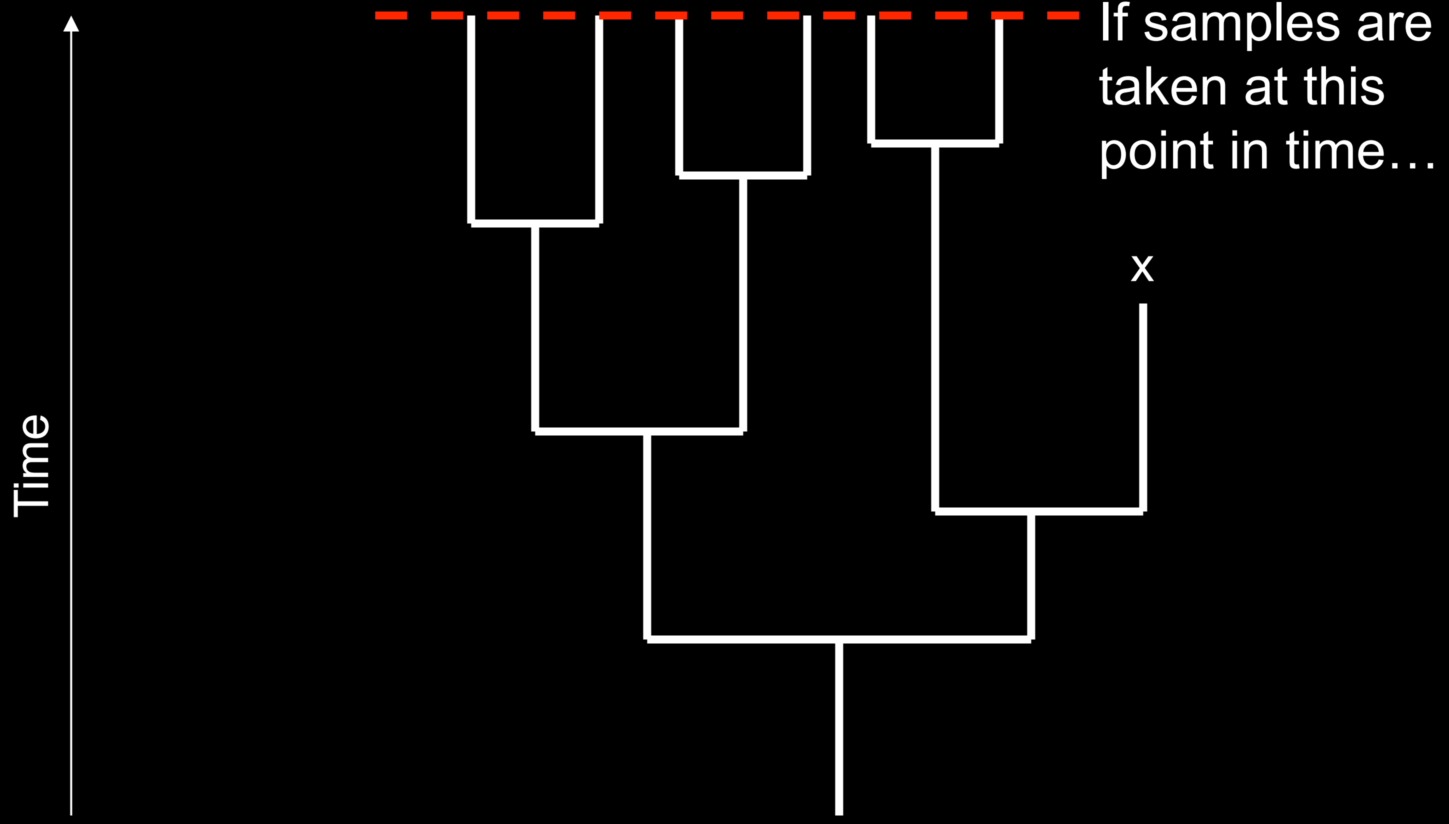
Evolutionary relationships among lineages, such as genes, individuals, populations, species, etc.



The relationships among lineages produce a phylogenetic tree

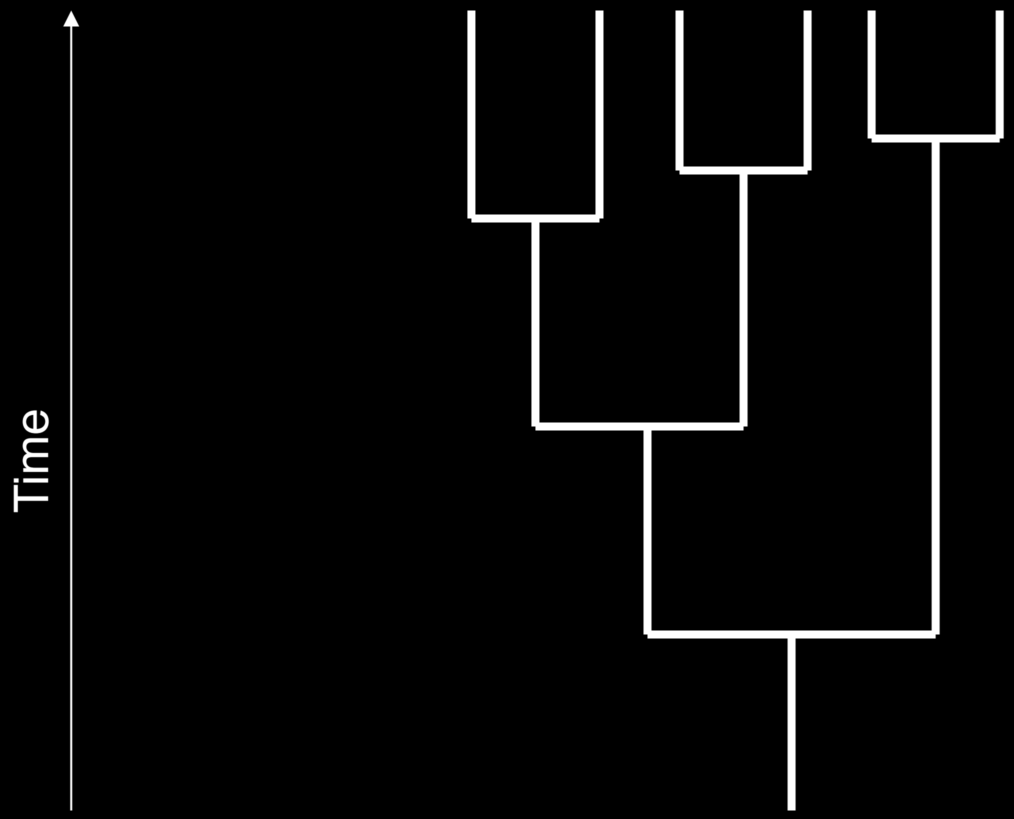
Phylogeny

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

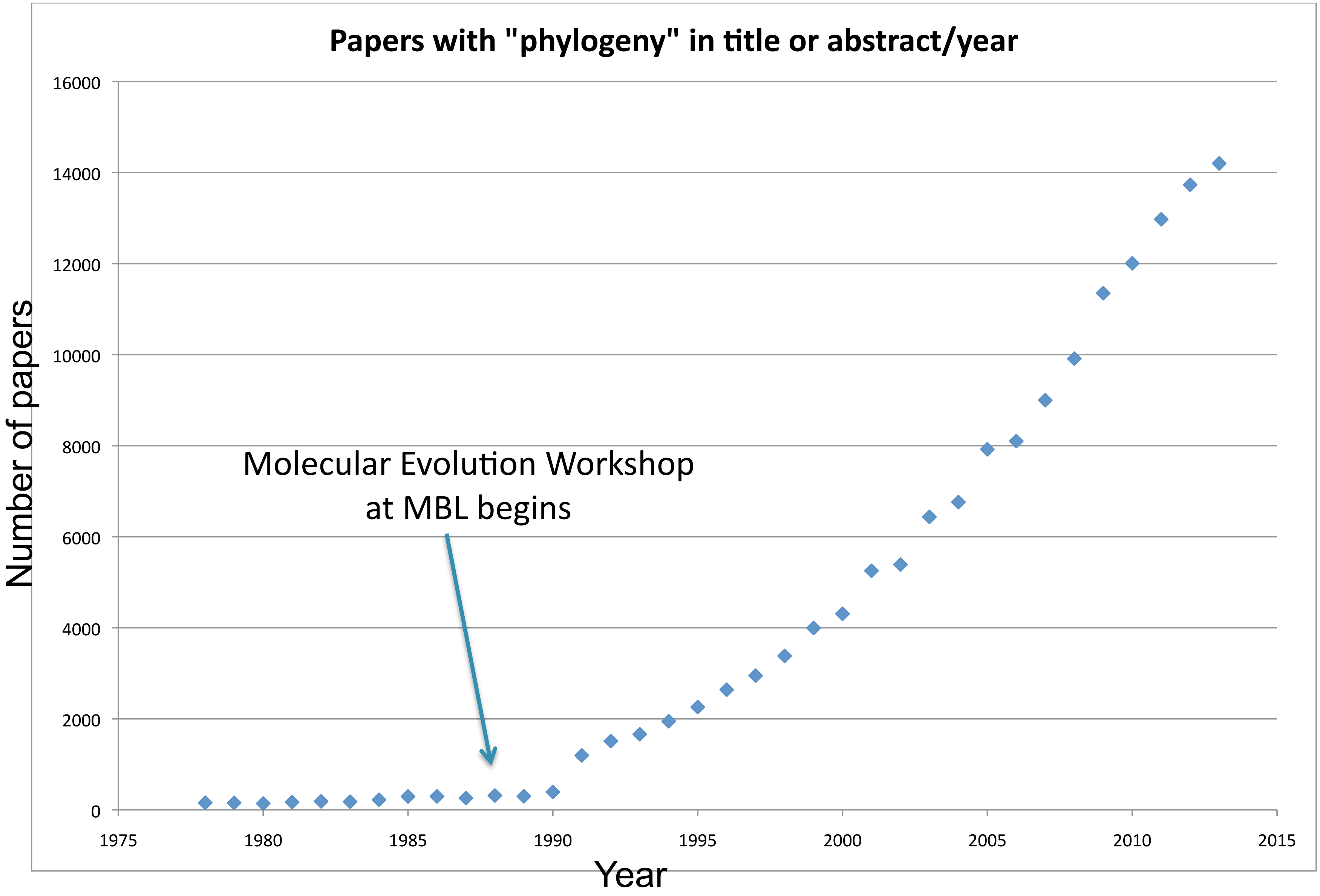


Phylogeny

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

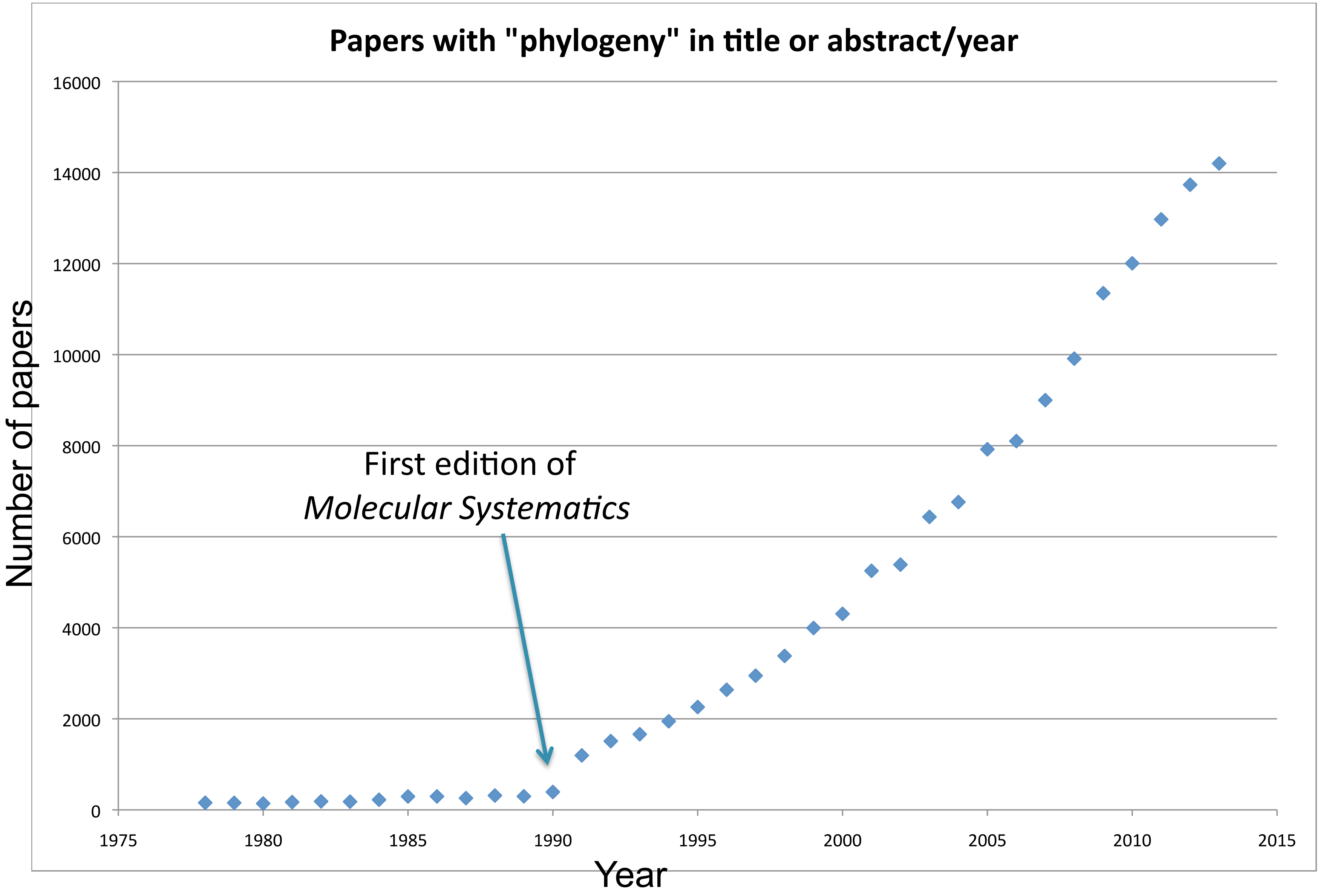


…then the reconstructed phylogenetic tree will look like this



Papers with "phylogeny" in title or abstract/year

Molecular Evolution Workshop
at MBL begins



Papers with "phylogeny" in title or abstract/year

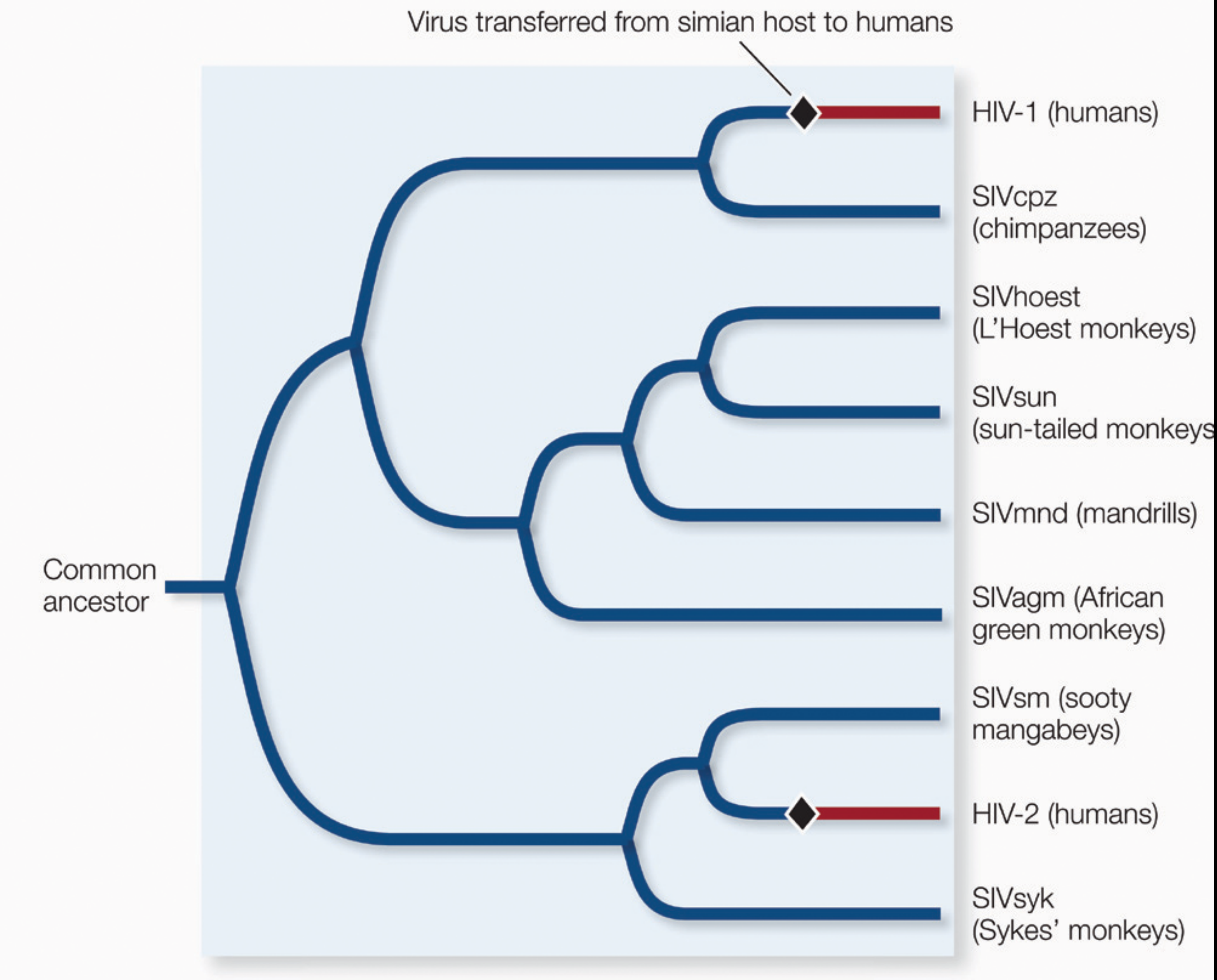
First edition of *Molecular Systematics*

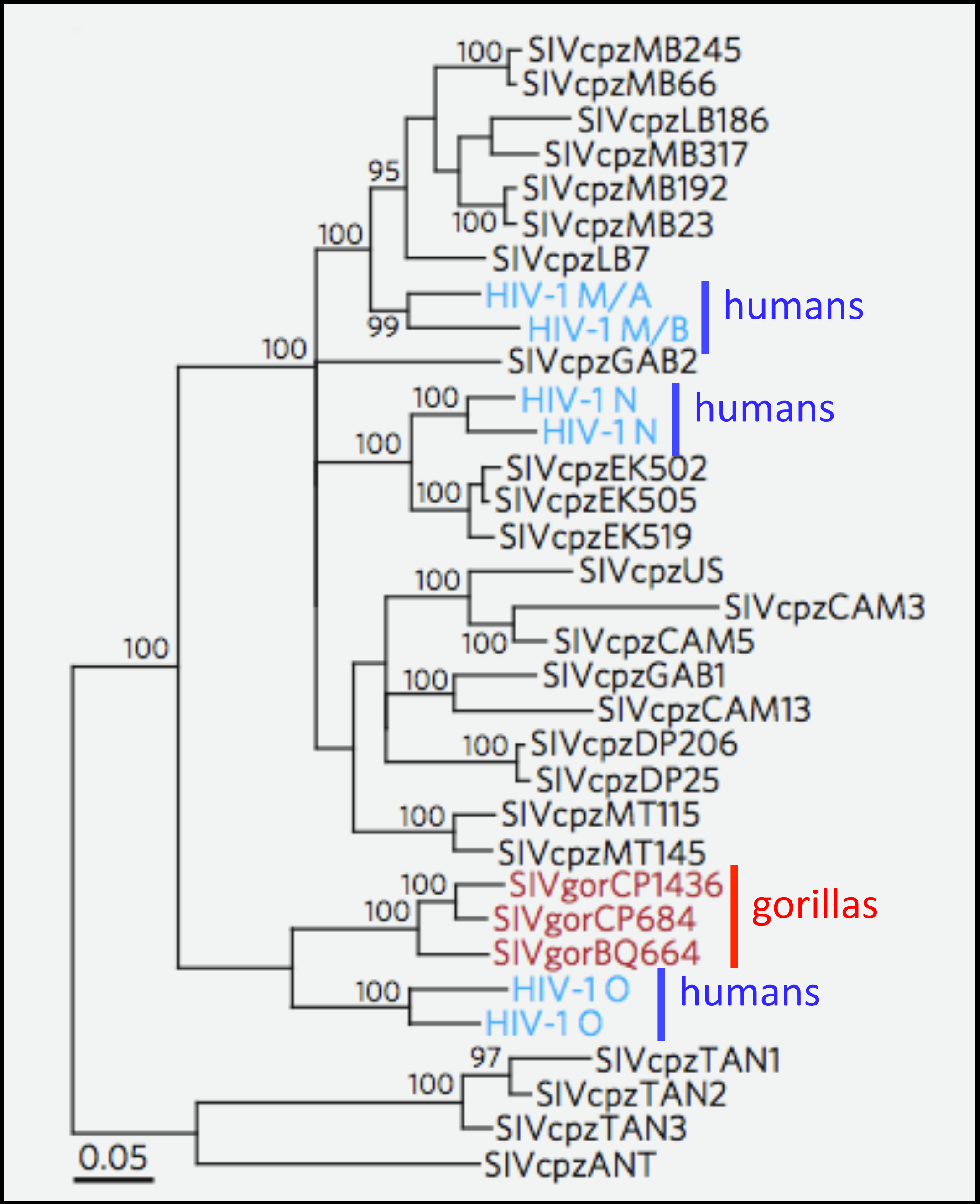
Major Uses of Phylogenetic Trees

- 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
- Trees allow appropriate comparative biology (identify appropriate comparisons)
- Trees are informative about ancestral states and state transitions

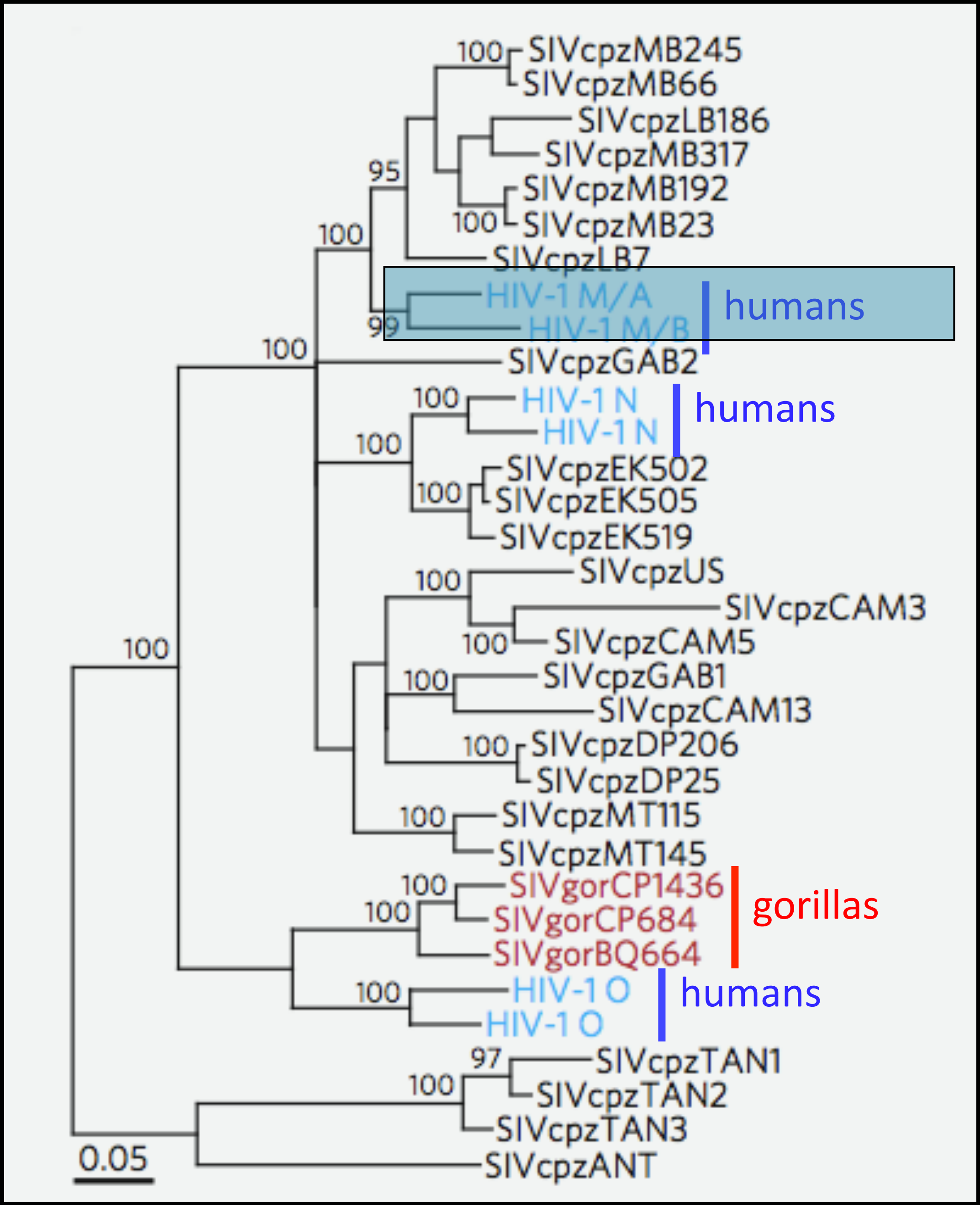
Origins of Emerging Diseases

- Where did HIV come from?
- How did it enter human populations?
- When did it enter human populations?
- How can we prevent similar diseases from entering human populations?


Virus transferred from simian host to humans
HIV-1 (humans)
SIVcpz (chimpanzees)
SIVhoest (L'Hoest monkeys)
SIVsun (sun-tailed monkeys
SIVmnd (mandrills)
SIVagm (African green monkeys)
Common ancestor
SIVsm (sooty mangabeys)
HIV-2 (humans)
SIVsyk (Sykes' monkeys)

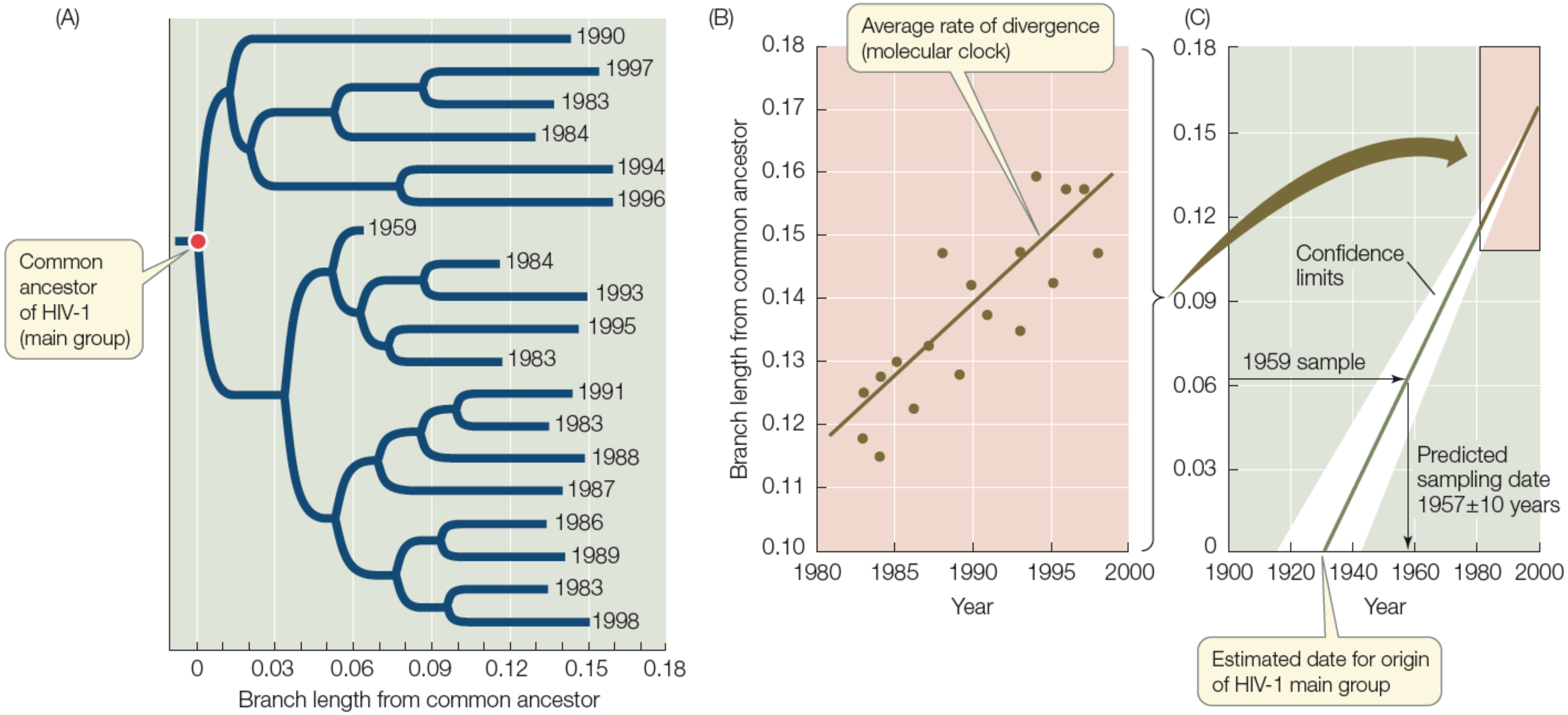


Van Heuverswyn et al., *Nature* 444: 164 (2006)

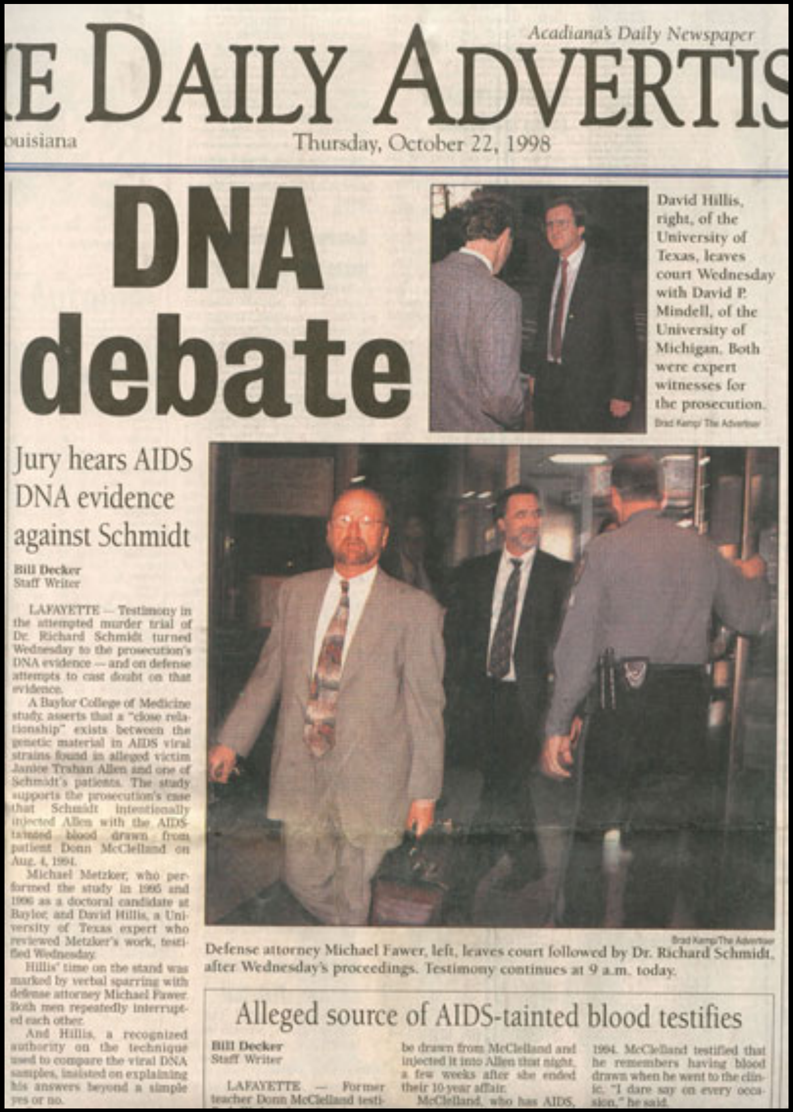


Van Heuverswyn et al., *Nature* 444: 164 (2006)

[redacted]



[redacted]



Acadiana's Daily Newspaper

E DAILY ADVERTIS

ouisiana

Thursday, October 22, 1998

DNA debate

David Hillis, right, of the University of Texas, leaves court Wednesday with David P. Mindell, of the University of Michigan. Both were expert witnesses for the prosecution.

Brad Kemp/ The Advertiser

Jury hears AIDS DNA evidence against Schmidt

Bill Decker
Staff Writer

LAFAYETTE — Testimony in the attempted murder trial of Dr. Richard Schmidt turned Wednesday to the prosecution's DNA evidence — and on defense attempts to cast doubt on that evidence.

A Baylor College of Medicine study asserts that a "close relationship" exists between the genetic material in AIDS viral strains found in alleged victim Janice Trahan Allen and one of Schmidt's patients. The study supports the prosecution's case that Schmidt intentionally injected Allen with the AIDS-tainted blood drawn from patient Donn McClelland on Aug. 4, 1994.

Michael Metzker, who performed the study in 1995 and 1996 as a doctoral candidate at Baylor, and David Hillis, a University of Texas expert who reviewed Metzker's work, testified Wednesday.

Hillis' time on the stand was marked by verbal sparring with defense attorney Michael Fawer. Both men repeatedly interrupted each other.

And Hillis, a recognized authority on the technique used to compare the viral DNA samples, insisted on explaining his answers beyond a simple yes or no.

Brad Kemp/The Advertiser

Defense attorney Michael Fawer, left, leaves court followed by Dr. Richard Schmidt, after Wednesday's proceedings. Testimony continues at 9 a.m. today.

Alleged source of AIDS-tainted blood testifies

Bill Decker
Staff Writer

LAFAYETTE — Former teacher Donn McClelland testi-

be drawn from McClelland and injected it into Allen that night, a few weeks after she ended their 10-year affair.

McClelland, who has AIDS,

1994. McClelland testified that he remembers having blood drawn when he went to the clinic. "I dare say on every occasion," he said.

Phylogenetic analysis can be used to trace viral infections through a human population

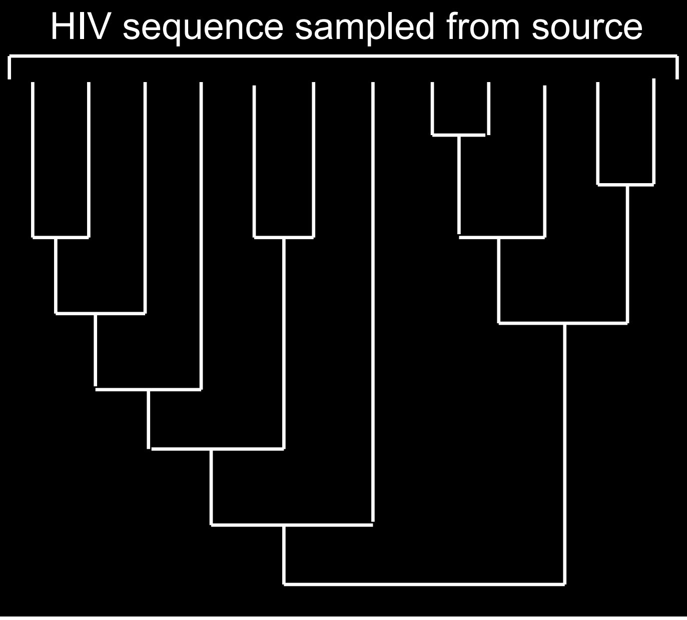
- Origins of HIV, SARS and other viruses transmitted between animals and humans
- Global virus diversity for vaccine trials
- Epidemiological studies
- Identification of new diseases
- Forensic uses

HIV transmission

Viral transmission events may be traced back through time among individuals in a population.

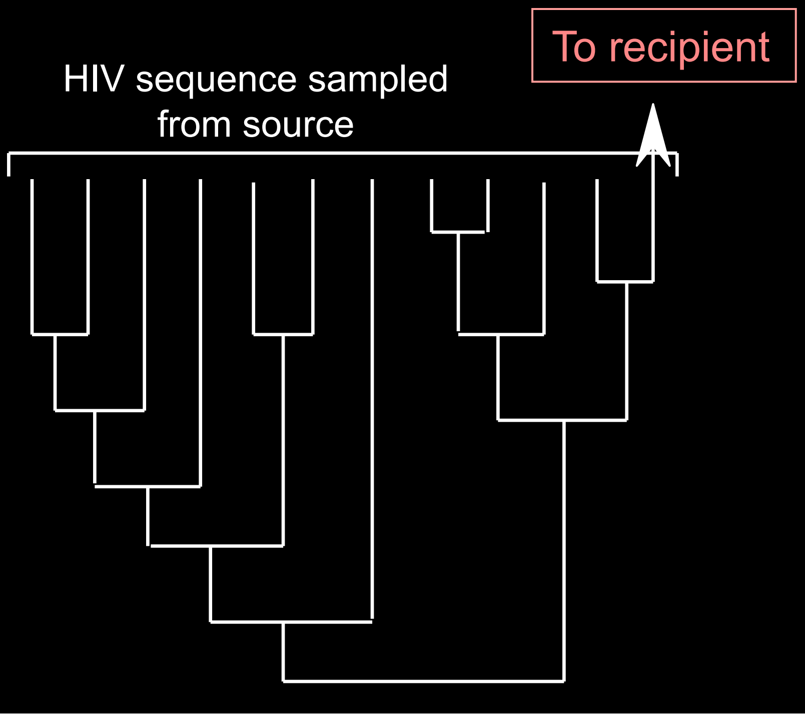
To imagine how this is possible, start by considering the diversity of HIV within one infected individual:

Time 1: Prior to Transmission event



At the transmission event, the HIV in the recipient represents a small subset of the HIV present in the source:

Time 2: The transmission event

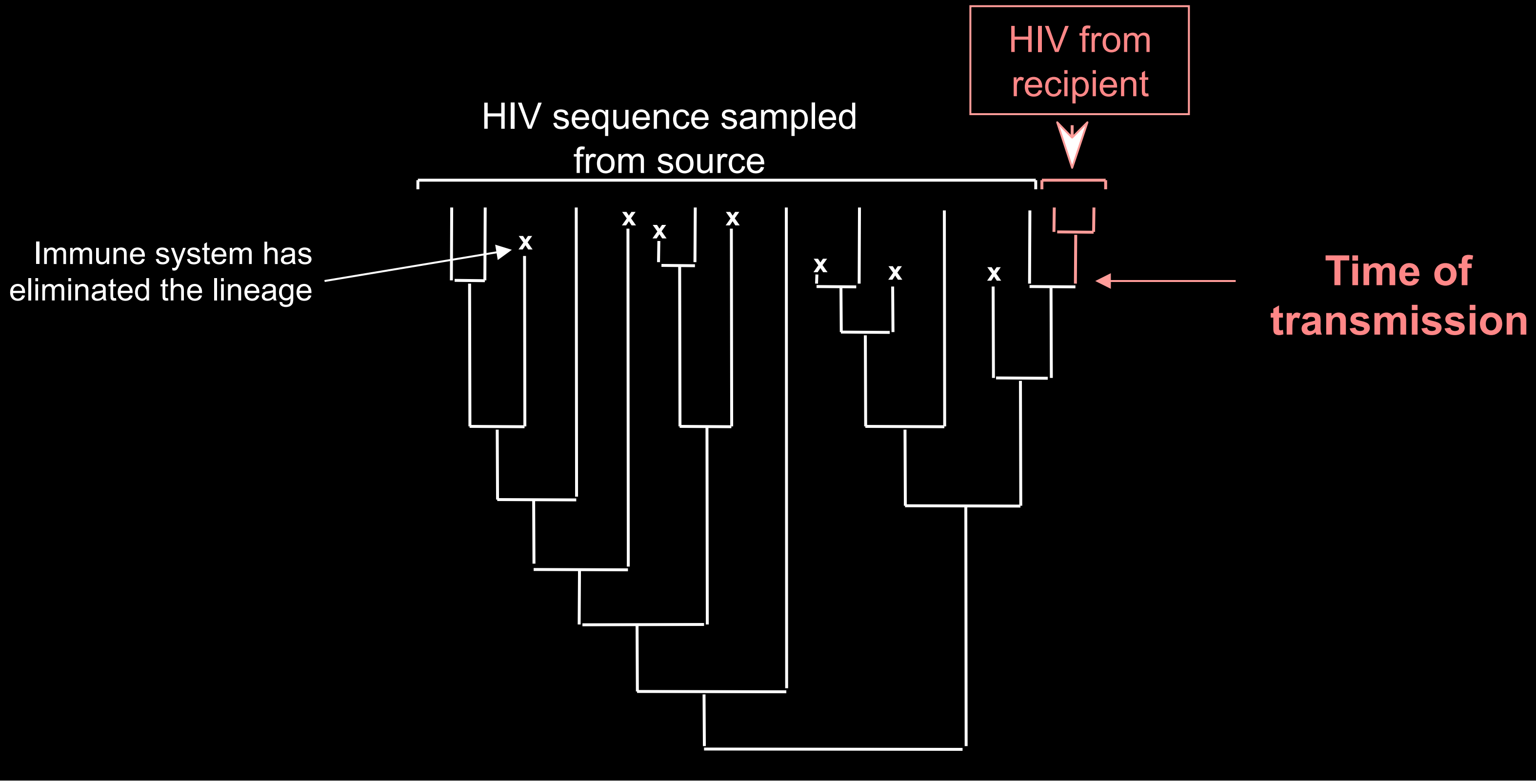


To recipient

HIV sequence sampled from source

As time passes, HIV lineages in the source and recipient diversify, and other lineages become extinct.

Time 3: Shortly after transmission event

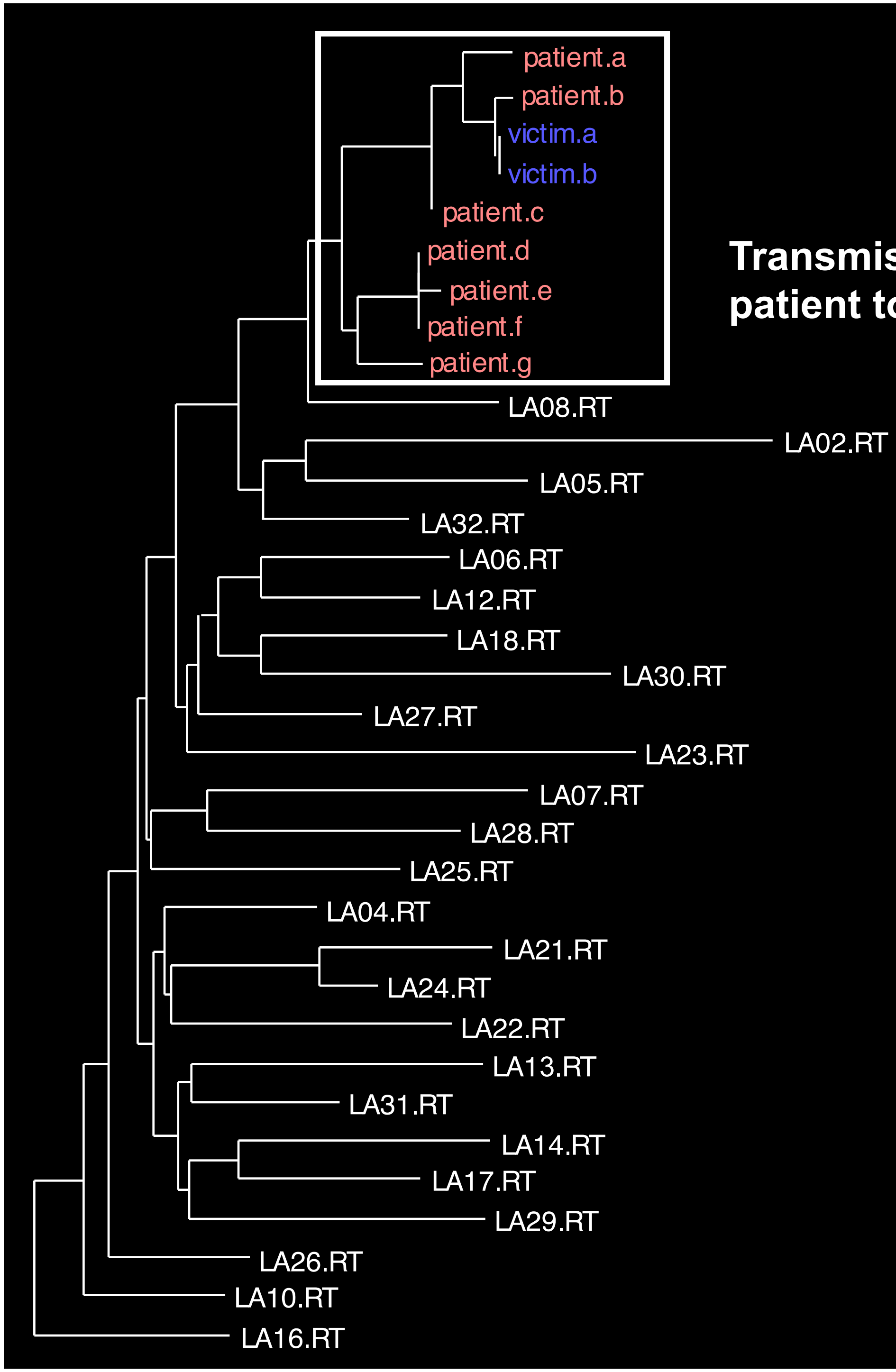


HIV from recipient

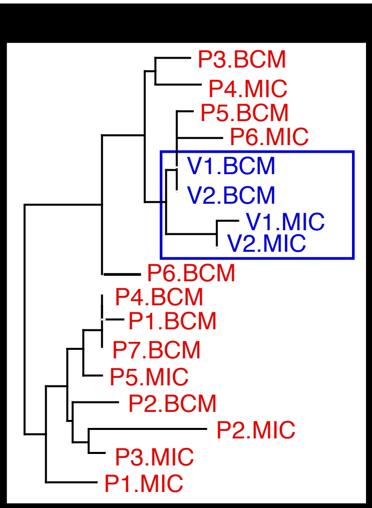
HIV sequence sampled from source

Immune system has eliminated the lineage

Time of transmission



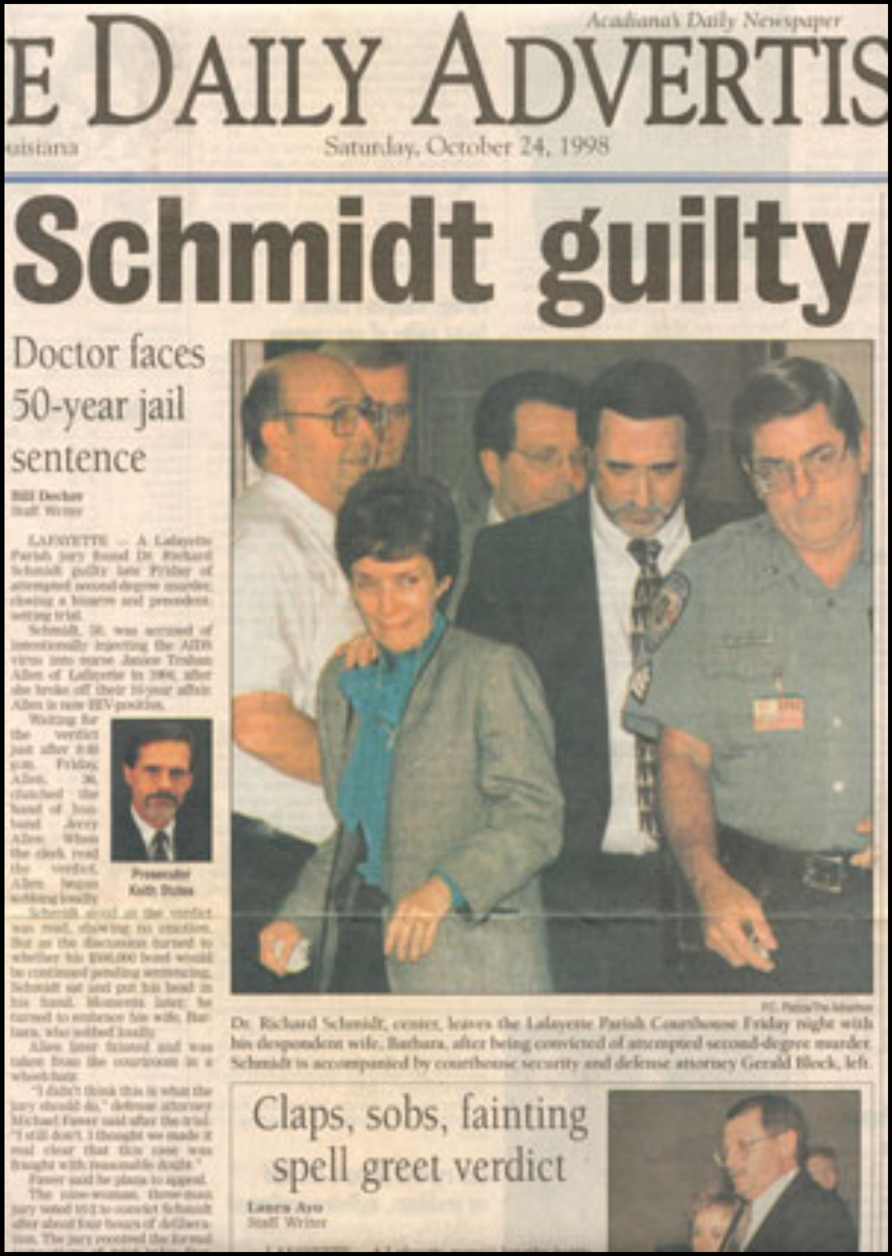
Transmission from patient to victim



New sequences added

Phylogenetic evidence:
pol sequences

Metzker et al., 2002
(PNAS 99:14292-14297)



Acadiana's Daily Newspaper

E DAILY ADVERTIS

uisiana

Saturday, October 24, 1998

Schmidt guilty

Doctor faces 50-year jail sentence

Bill Decker
Staff Writer

LAFAYETTE — A Lafayette Parish jury found Dr. Richard Schmidt guilty late Friday of attempted second-degree murder, closing a bizarre and precedent-setting trial.

Schmidt, 50, was accused of intentionally injecting the AIDS virus into nurse Janice Trahan Allen of Lafayette in 1994, after she broke off their 10-year affair. Allen is now HIV-positive.

Waiting for the verdict just after 8:40 p.m. Friday, Allen, 36, clutched the hand of husband Jerry Allen. When the clerk read the verdict, Allen began sobbing loudly.

Prosecutor Keith Stutes

Schmidt stood as the verdict was read, showing no emotion. But as the discussion turned to whether his $500,000 bond would be continued pending sentencing, Schmidt sat and put his head in his hand. Moments later, he turned to embrace his wife, Barbara, who sobbed loudly.

Allen later fainted and was taken from the courtroom in a wheelchair.

"I didn't think this is what the jury should do," defense attorney Michael Fawer said after the trial. "I still don't. I thought we made it real clear that this case was fraught with reasonable doubt."

Fawer said he plans to appeal.

The nine-woman, three-man jury voted 10-2 to convict Schmidt after about four hours of deliberation. The jury received the formal

P.C. Piazza/The Advertiser

Dr. Richard Schmidt, center, leaves the Lafayette Parish Courthouse Friday night with his despondent wife, Barbara, after being convicted of attempted second-degree murder. Schmidt is accompanied by courthouse security and defense attorney Gerald Block, left.

Claps, sobs, fainting spell greet verdict

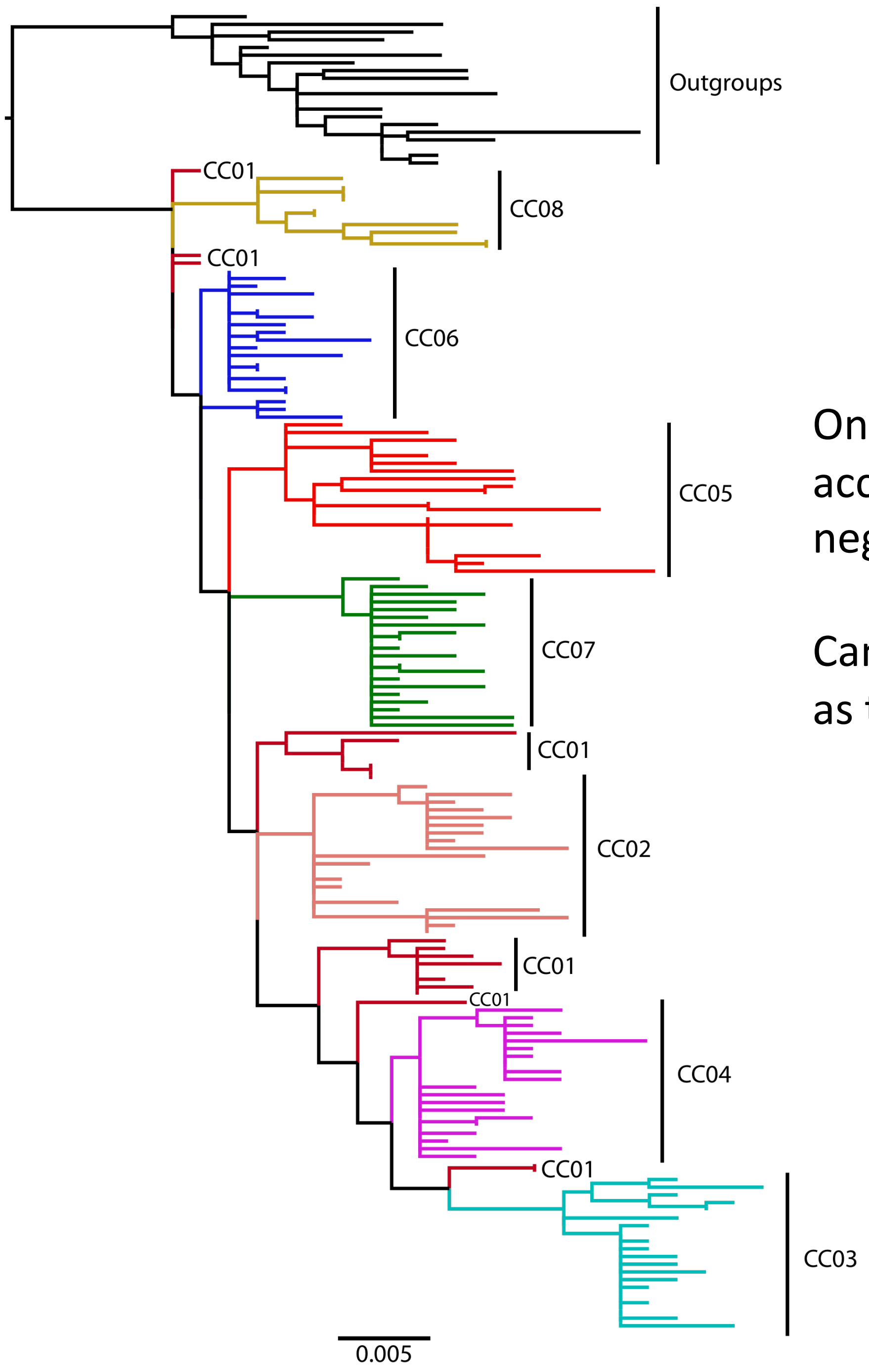
Laura Ayo
Staff Writer

LAFAYETTE —

- Schmidt was convicted of attempted murder; currently serving term of 50 years of hard labor

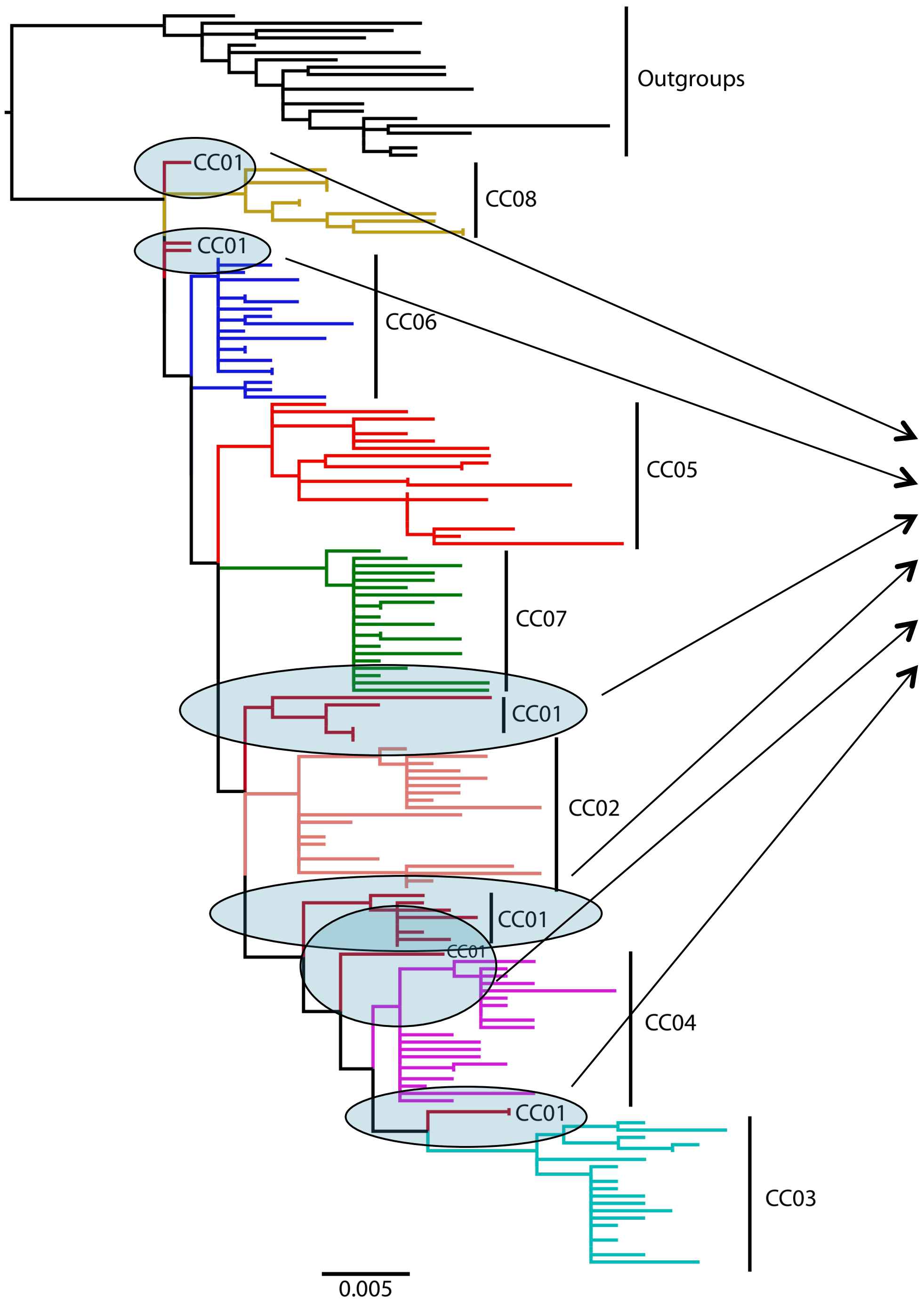
- First use of phylogenetic analysis in U.S. criminal case

- Phylogenetics can be used to trace infections of human pathogens among individuals

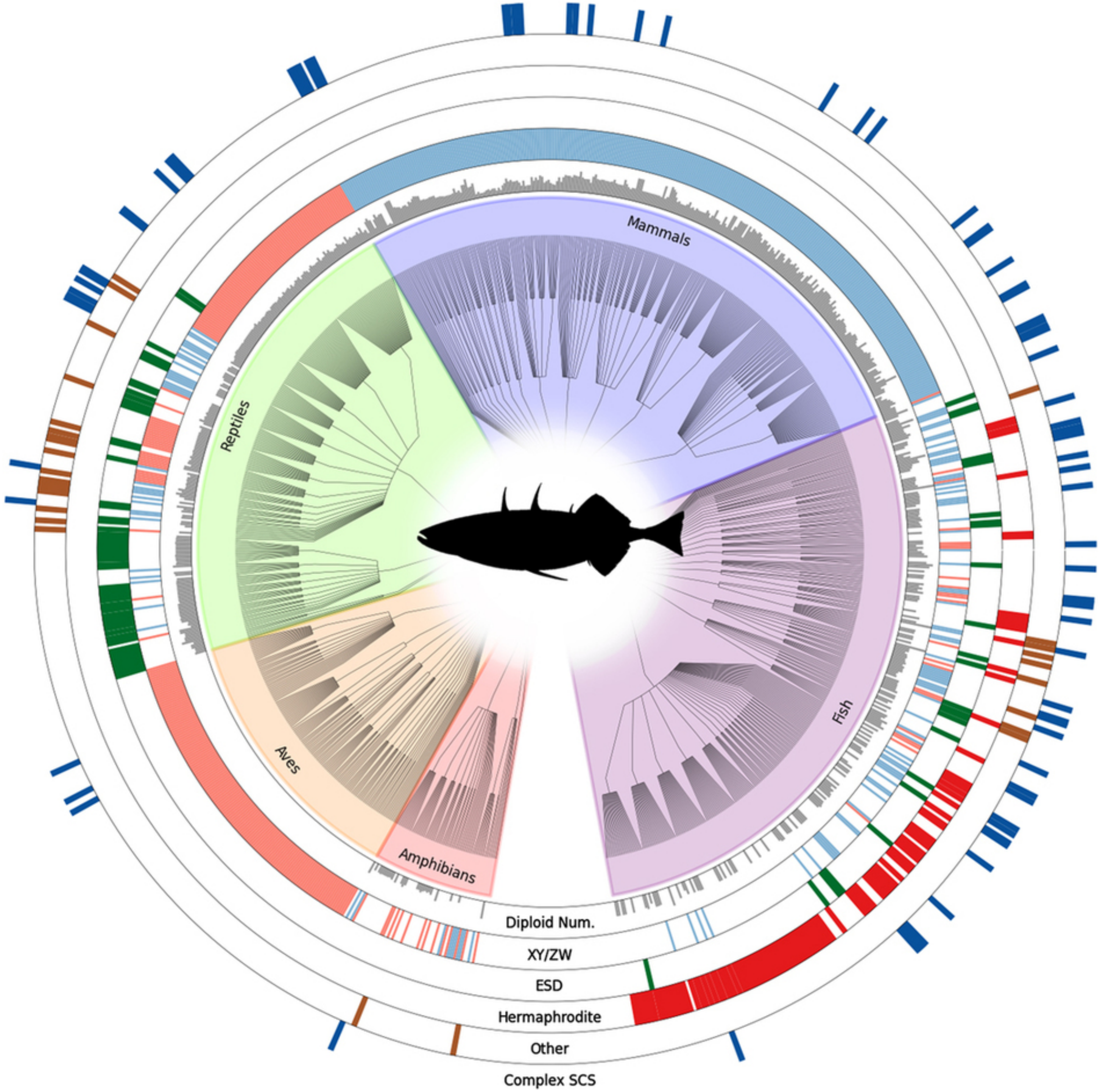


One of these individuals is accused of knowingly and negligently infecting the others.

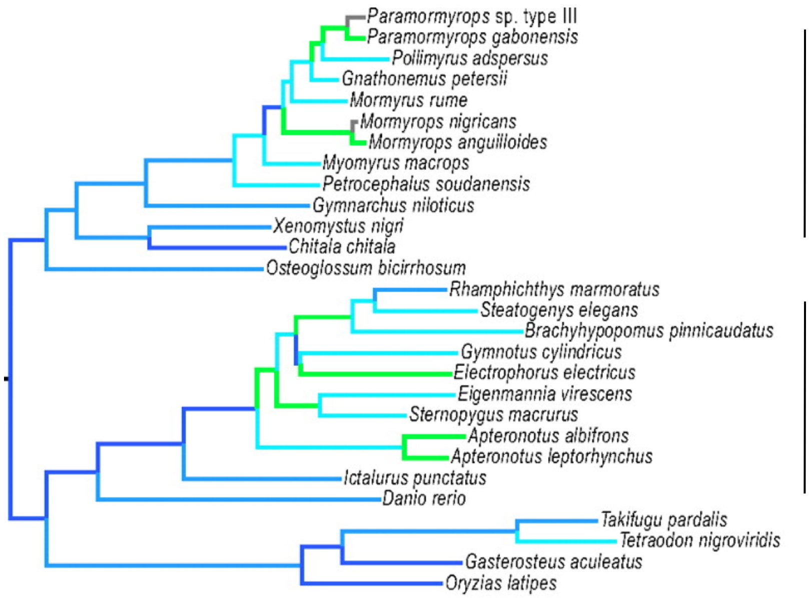
Can one person be identified as the source of the infections?



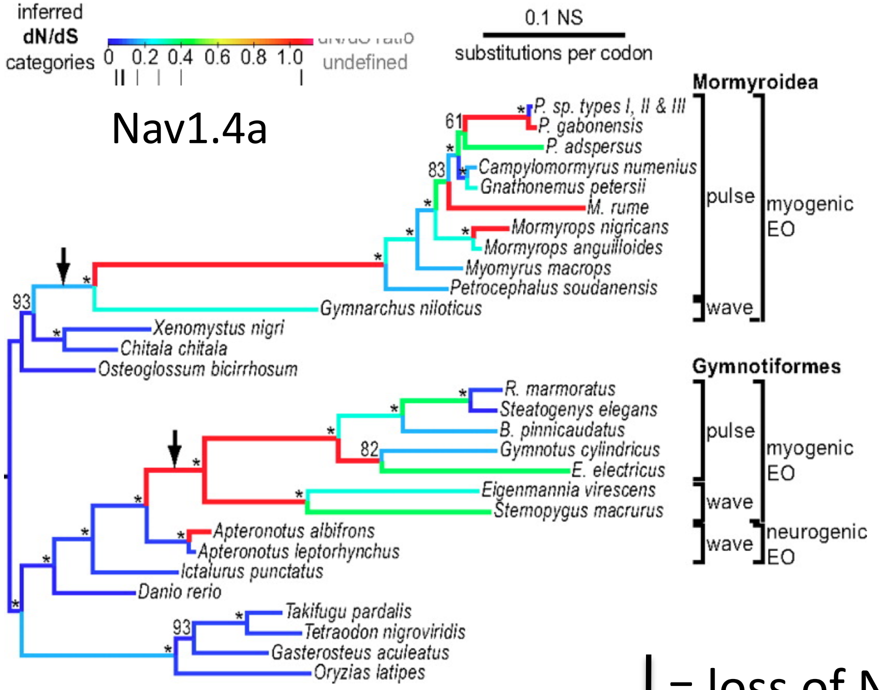
One individual (CC01) is paraphyletic to all the rest. At the trial, CC01 was revealed to be the defendant, who was accused of six counts of motivated assault. He was found guilty by the jury in May 2009.



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



Nav1.4b



Nav1.4a

↓= loss of Nav1.4a expression in muscle

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

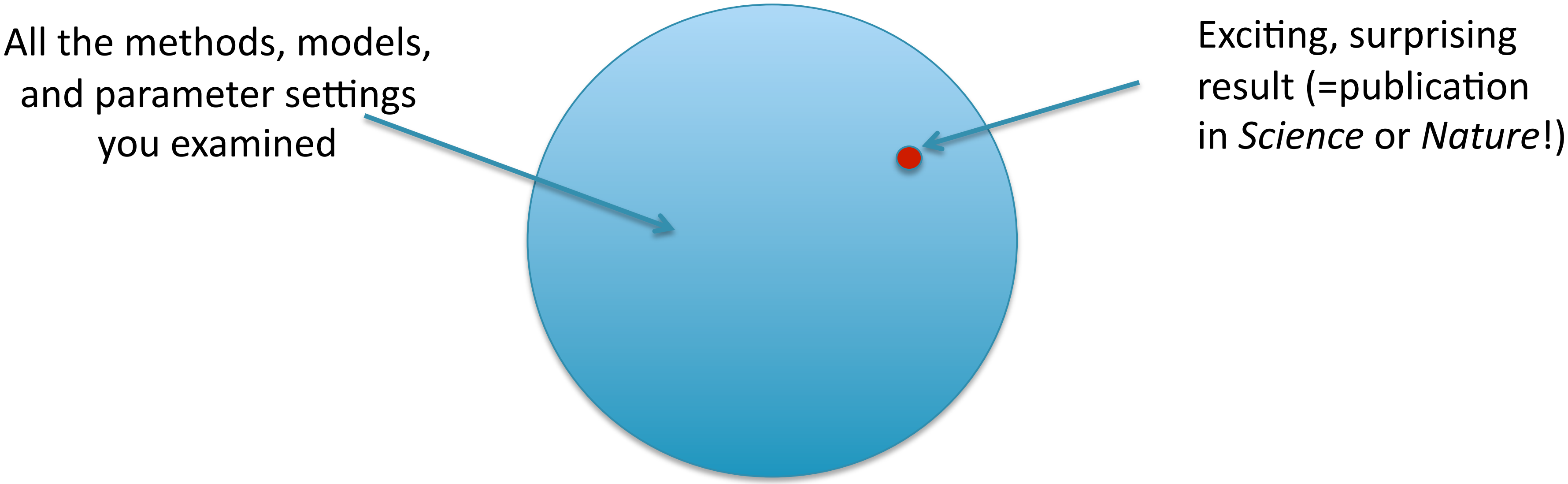
- All data, method descriptions, scripts, and programs must be publicly available in a way that all your analyses can be repeated, checked, and extended
  - Not just sequences in GenBank and a statement that you used a particular program for analysis!
  - Include alignments, parameter settings, scripts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

Ethics of Data Presentation and Analysis

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



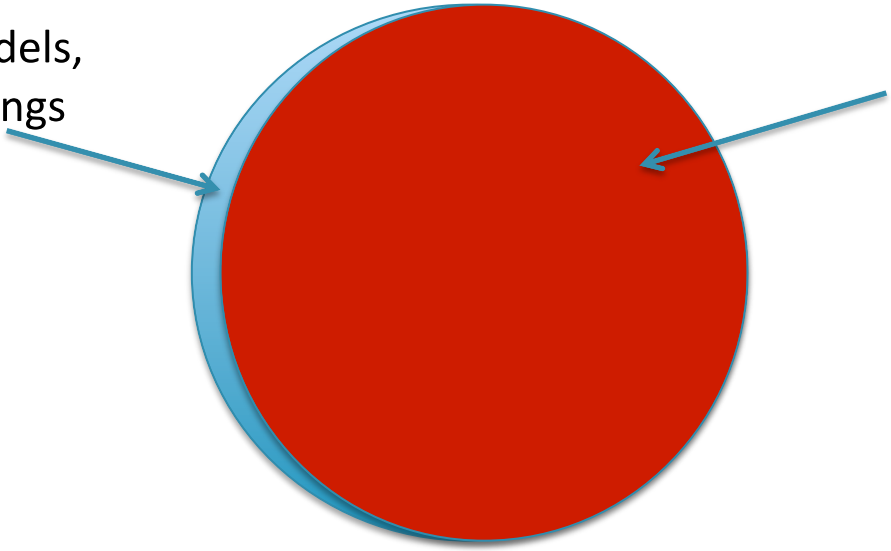
All the methods, models, and parameter settings you examined

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

What do you publish?

Ethics of Data Presentation and Analysis

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

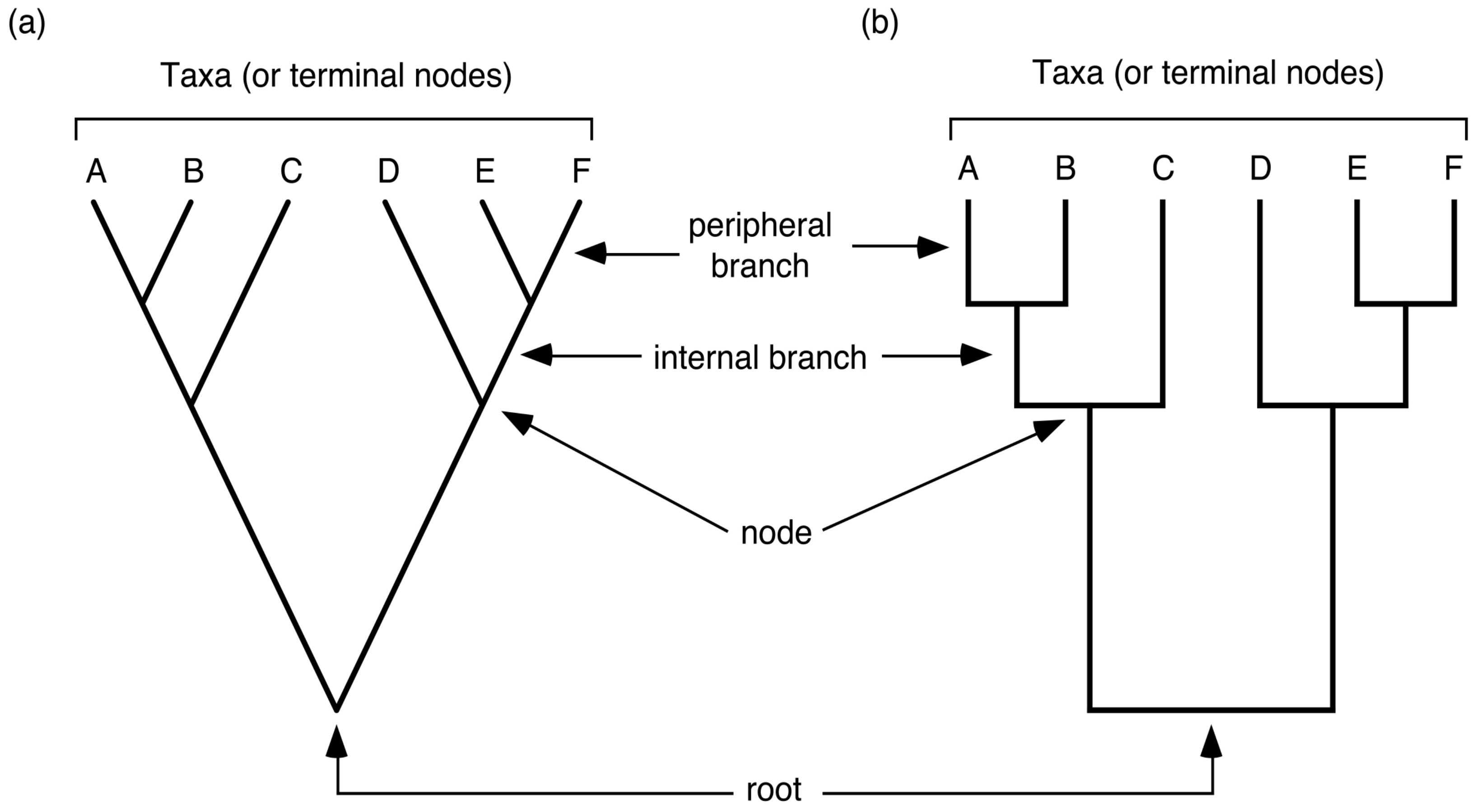


All the methods, models, and parameter settings you examined

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

What do you publish?

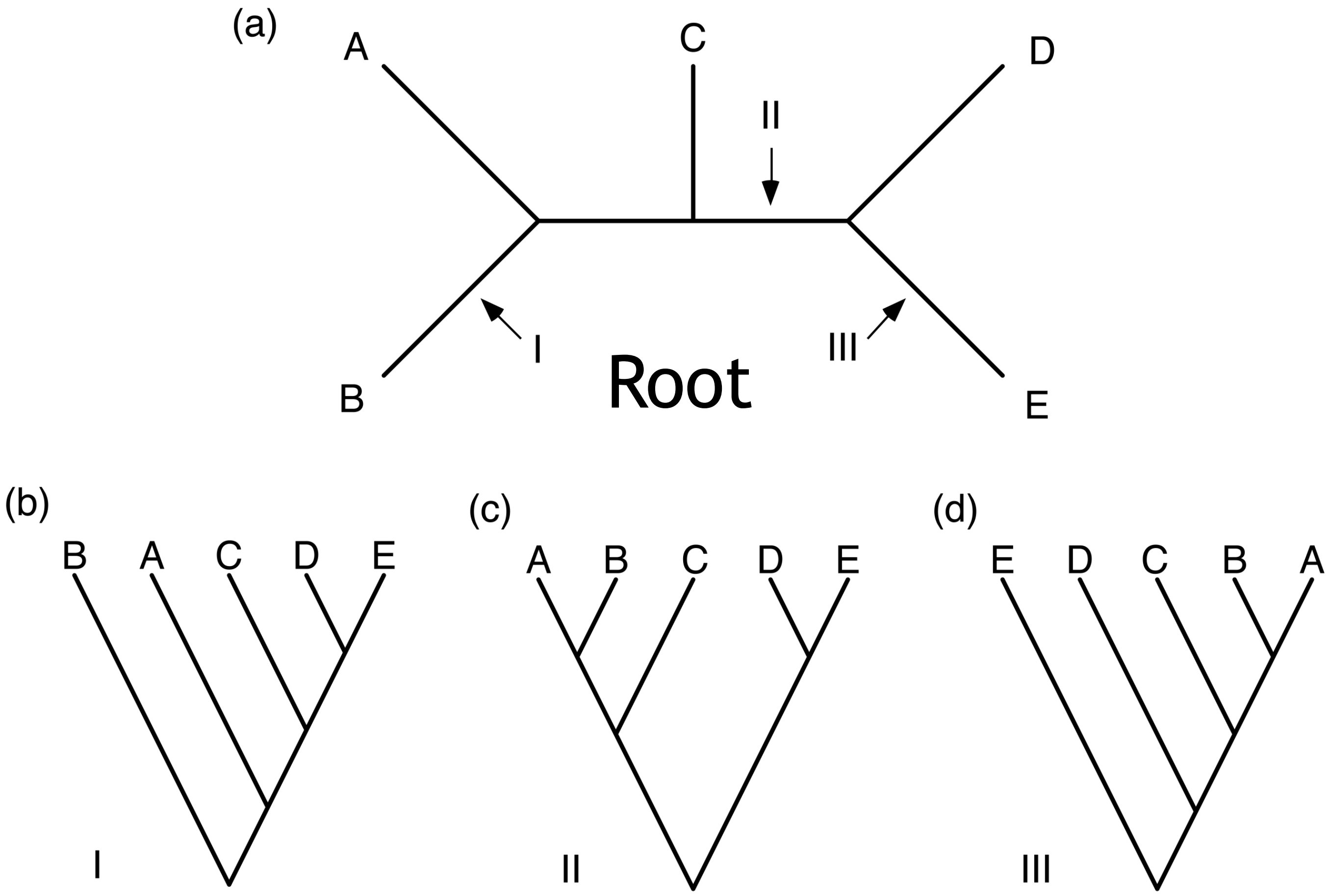
Terminology for Trees



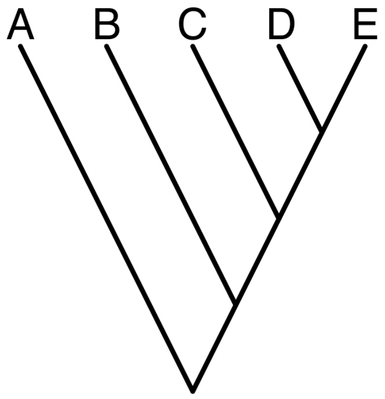
Rooting Trees

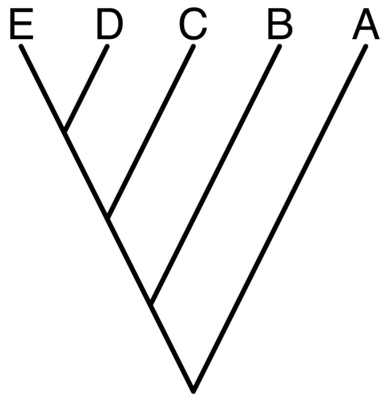
- When trees are used to indicate direction of time, they are *rooted*
- One node is identified as the *root*; this can be an existing node or a new node.
- An *unrooted* tree is uninformative with respect to direction of time.

Rooting Trees

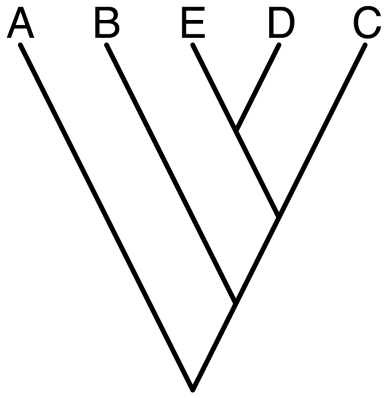


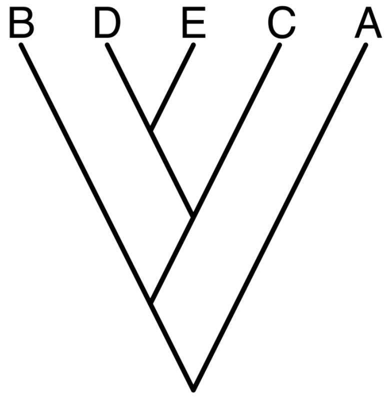
Which are Different?





(A,(B,(C,(D,E))))





(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is ‘(‘ then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node’s parent and add another child;
- if the next symbol is a ‘)’, then move back to the current node’s parent.

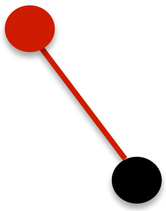
(A,(B,(C,(D,E))))

(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is '(' then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node's parent and add another child;
- if the next symbol is a ')', then move back to the current node's parent.

(A,(B,(C,(D,E))))

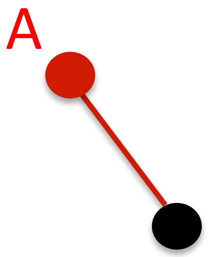


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is '(' then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node's parent and add another child;
- if the next symbol is a ')', then move back to the current node's parent.

(A,(B,(C,(D,E))))

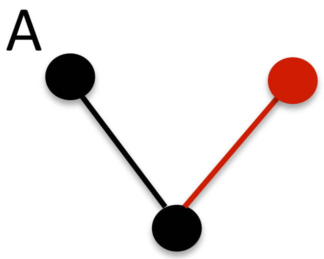


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is '(' then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node's parent and add another child;
- if the next symbol is a ')', then move back to the current node's parent.

(A,(B,(C,(D,E))))

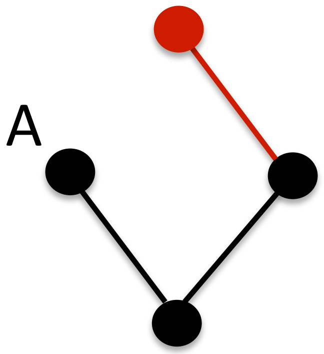


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is '(' then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node's parent and add another child;
- if the next symbol is a ')', then move back to the current node's parent.

(A,(B,(C,(D,E))))

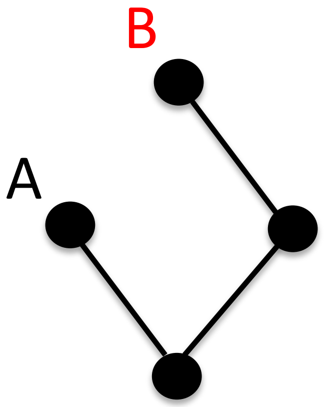


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is '(' then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node's parent and add another child;
- if the next symbol is a ')', then move back to the current node's parent.

(A,(B,(C,(D,E))))

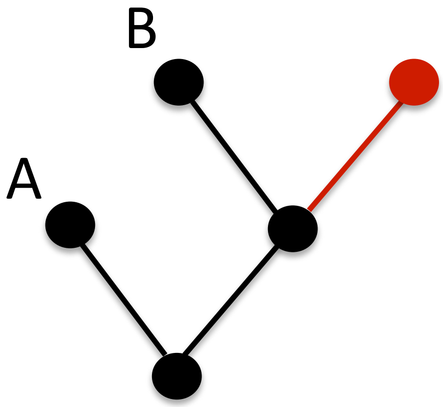


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is '(' then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node's parent and add another child;
- if the next symbol is a ')', then move back to the current node's parent.

(A,(B,(C,(D,E))))

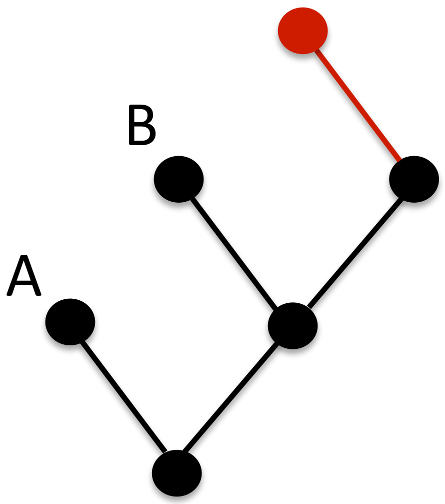


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is '(' then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node's parent and add another child;
- if the next symbol is a ')', then move back to the current node's parent.

(A,(B,(C,(D,E))))

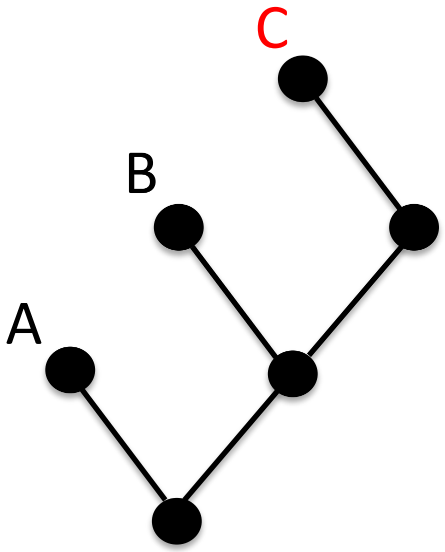


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is ‘(‘ then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node’s parent and add another child;
- if the next symbol is a ‘)’, then move back to the current node’s parent.

(A,(B,(C,(D,E))))

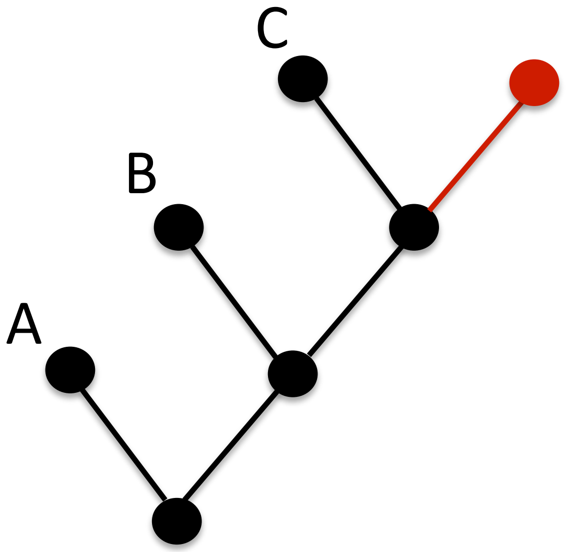


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is ‘(‘ then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node’s parent and add another child;
- if the next symbol is a ‘)’, then move back to the current node’s parent.

(A,(B,(C,(D,E))))

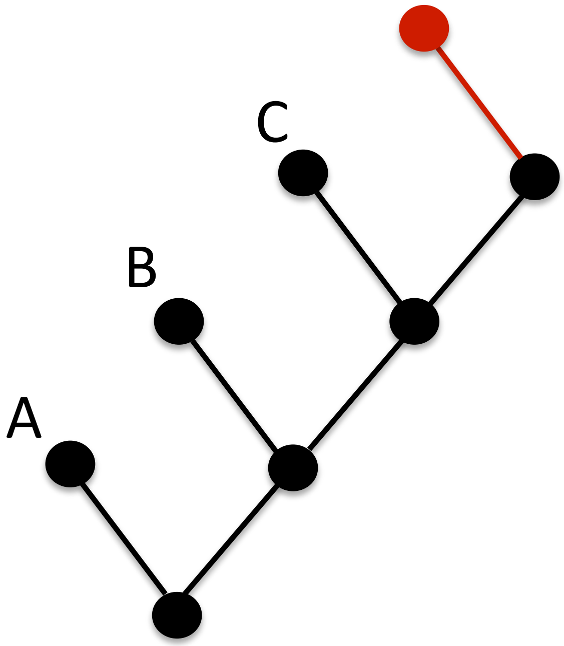


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is '(' then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node's parent and add another child;
- if the next symbol is a ')', then move back to the current node's parent.

(A,(B,(C,(D,E))))

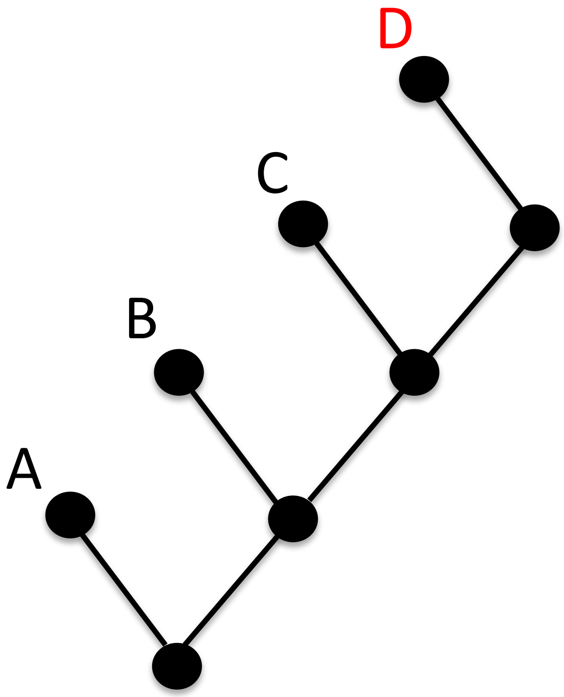


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is ‘(‘ then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node’s parent and add another child;
- if the next symbol is a ‘)’, then move back to the current node’s parent.

(A,(B,(C,(D,E))))

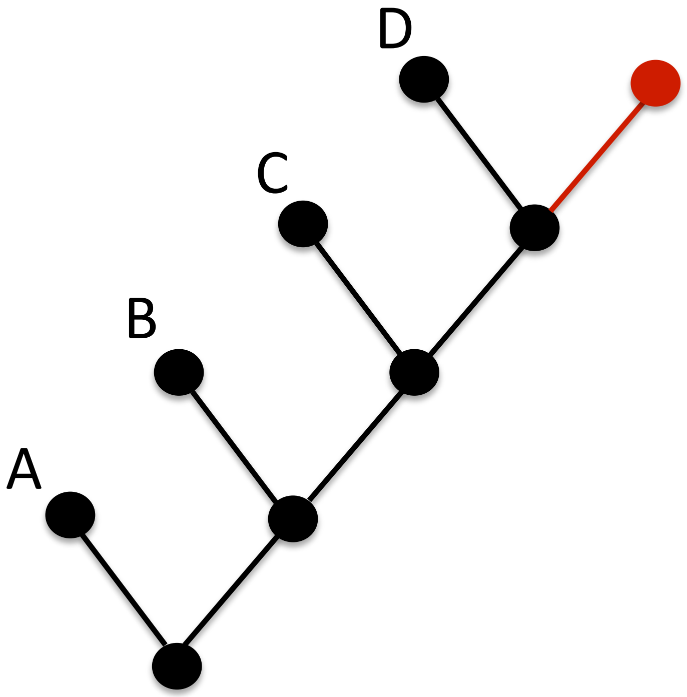


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is '(' then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node's parent and add another child;
- if the next symbol is a ')', then move back to the current node's parent.

(A,(B,(C,(D,E))))

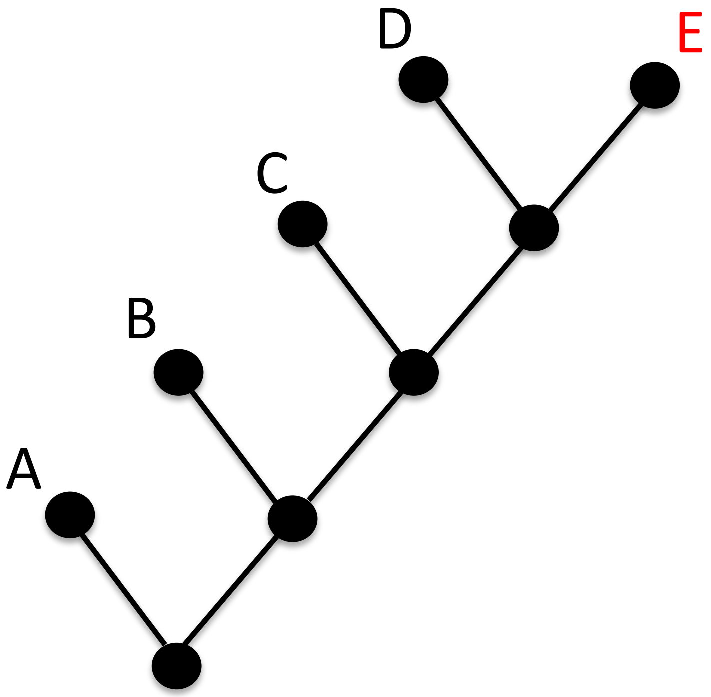


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is ‘(‘ then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node’s parent and add another child;
- if the next symbol is a ‘)’, then move back to the current node’s parent.

(A,(B,(C,(D,E))))

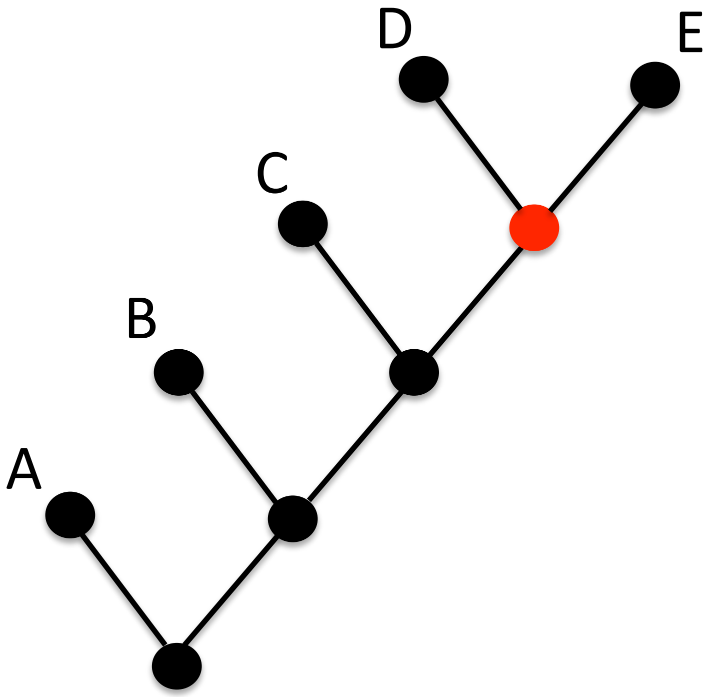


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is ‘(‘ then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node’s parent and add another child;
- if the next symbol is a ‘)’, then move back to the current node’s parent.

(A,(B,(C,(D,E))))

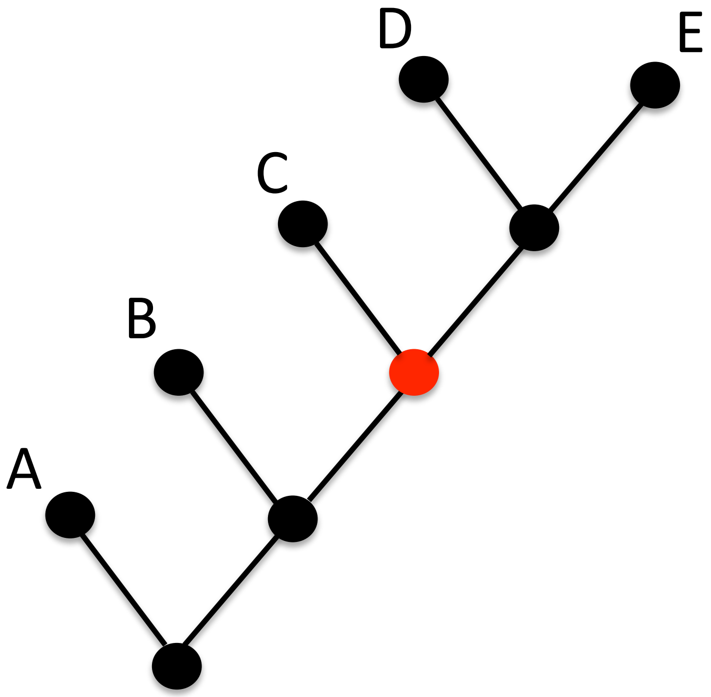


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is '(' then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node's parent and add another child;
- if the next symbol is a ')', then move back to the current node's parent.

(A,(B,(C,(D,E))))

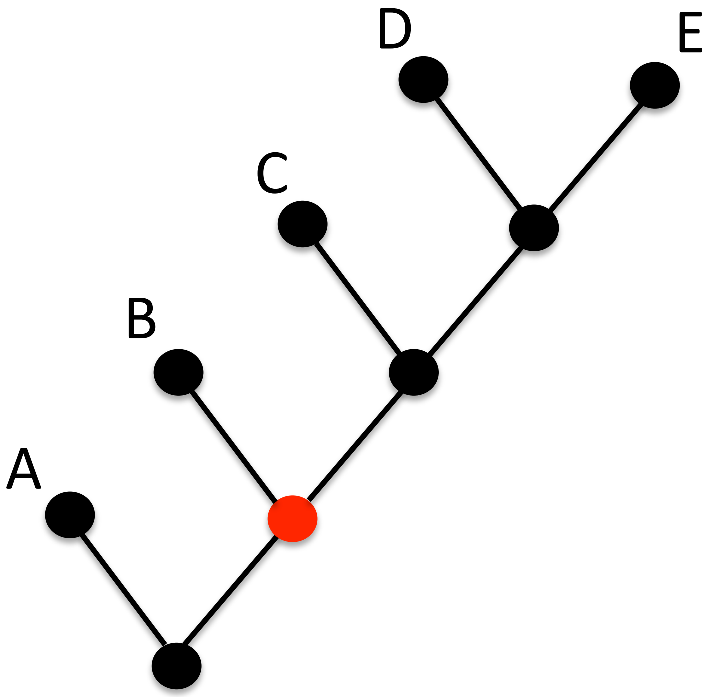


(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is ‘(‘ then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node’s parent and add another child;
- if the next symbol is a ‘)’, then move back to the current node’s parent.

(A,(B,(C,(D,E))))

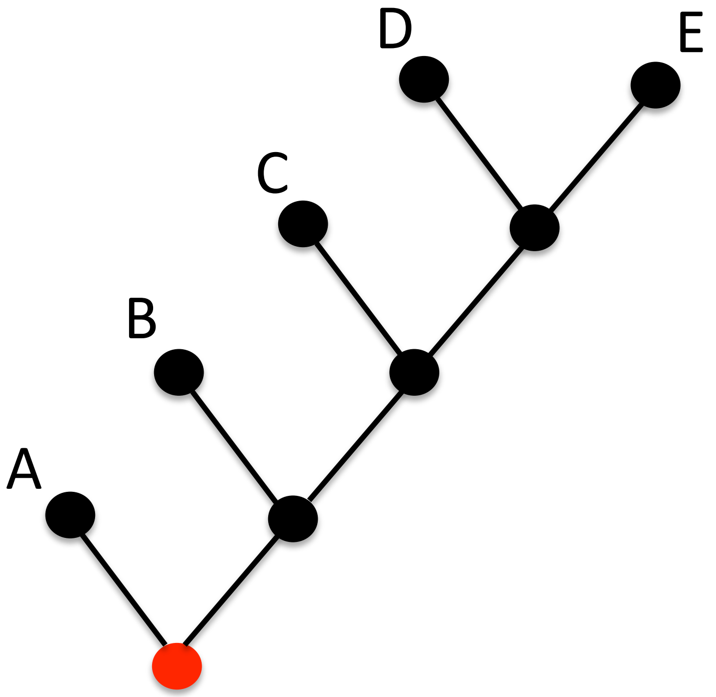


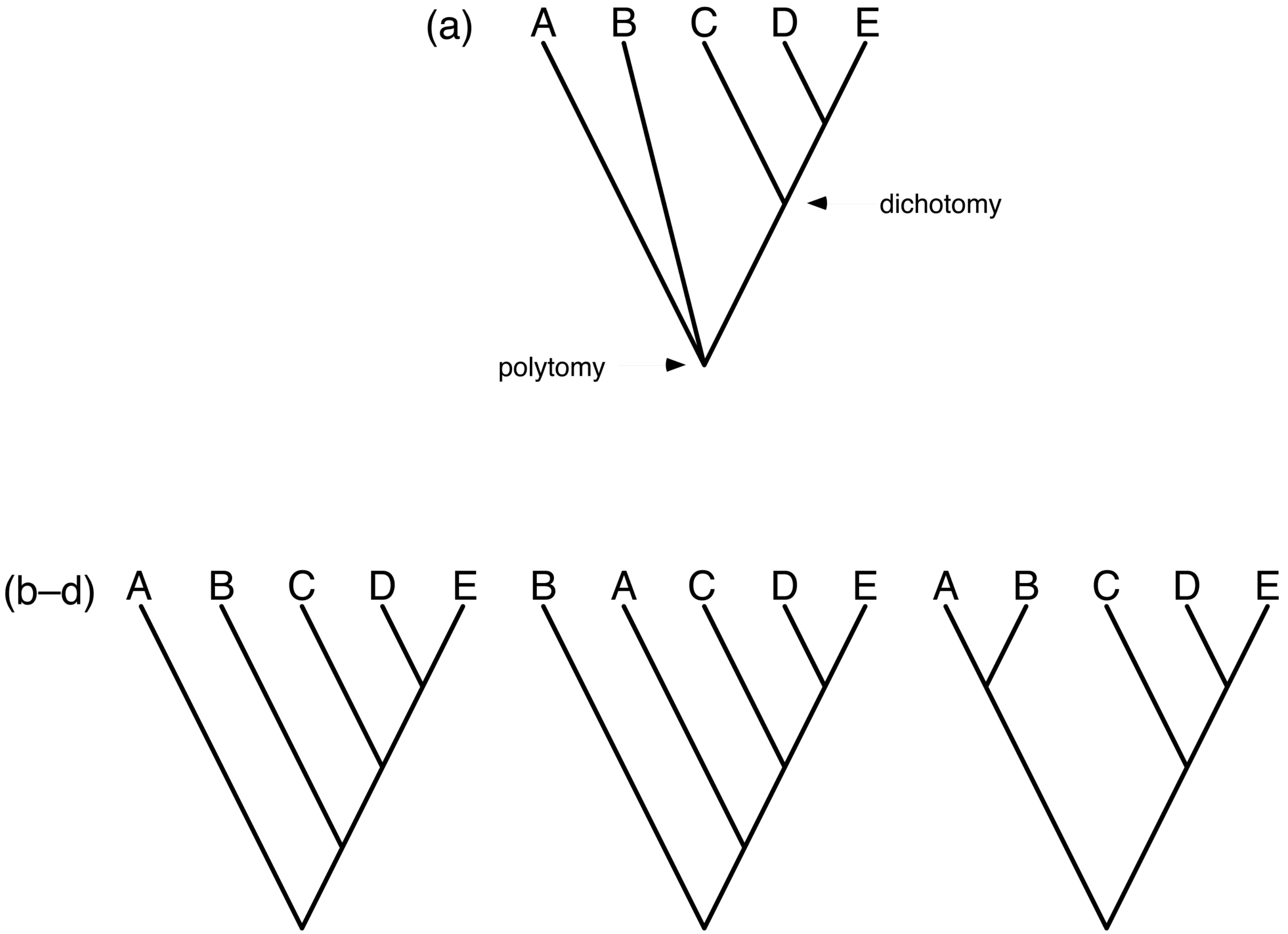
(A,(B,(C,(D,E)))) is called Newick or New Hampshire notation for the tree.

You can read it by following the rules:

- start at a node;
- if the next symbol is ‘(‘ then add a child to the current node and move to this child;
- if the next symbol is a label, then label the current node;
- if the next symbol is a comma, then move back to the current node’s parent and add another child;
- if the next symbol is a ‘)’, then move back to the current node’s parent.

(A,(B,(C,(D,E))))




(a)
A
B
C
D
E
dichotomy
polytomy
(b–d)
A
B
C
D
E
B
A
C
D
E
A
B
C
D
E

How Big Do Trees Get?

An unrooted, bifurcating tree with T terminal nodes has T – 2 internal nodes. It also has T – 3 internal branches and T peripheral branches, for a total of 2T – 3 branches. Adding a root to the tree also adds a branch, since the root divides one branch into two.

The number of labeled, unrooted binary trees is

$$N_U = \prod_{i=3}^{T} (2i - 5)$$

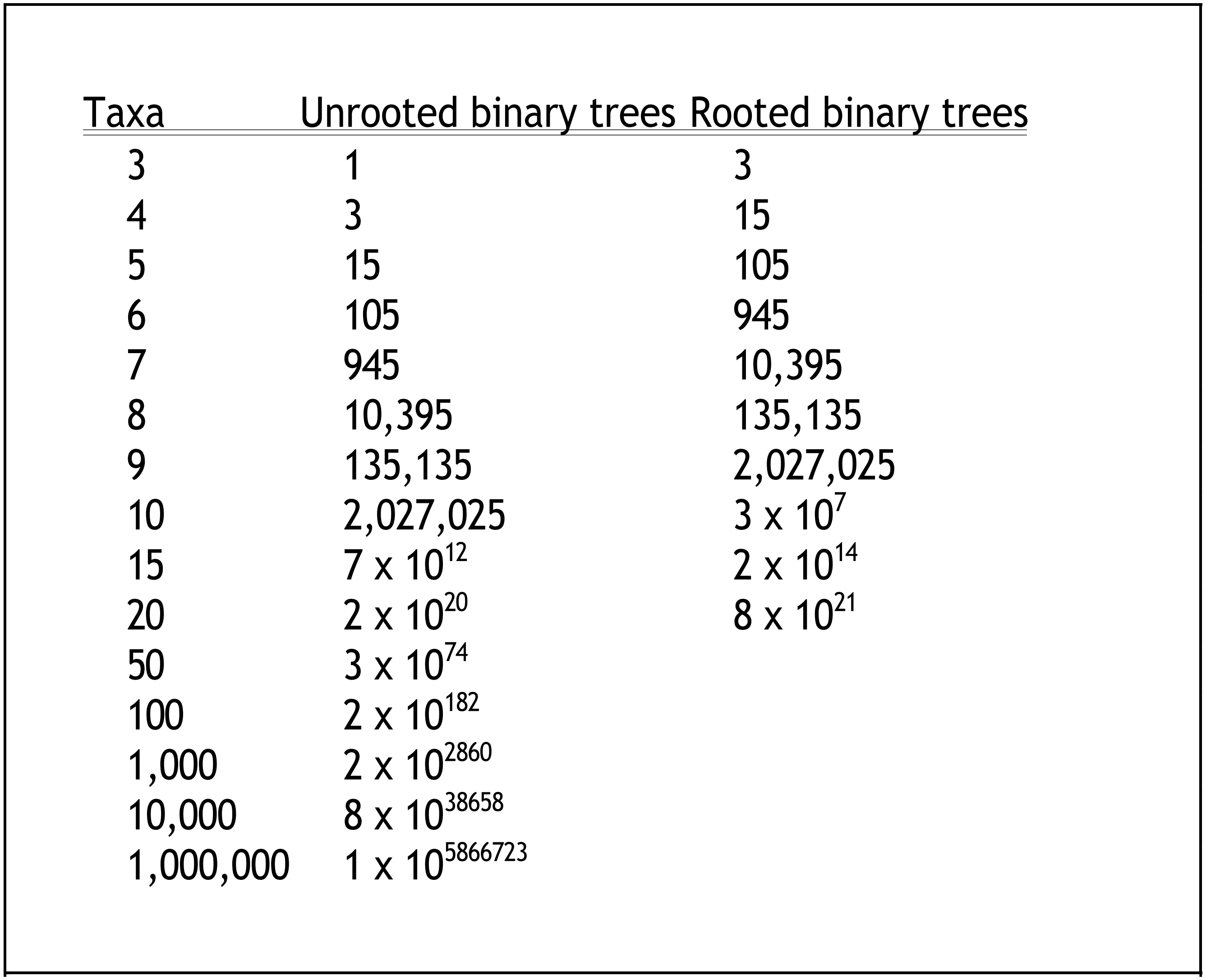
which expands to (2 · 3 – 5)(2 · 4 – 5)(2 · 5 – 5)…(2 · T - 5).

The number of labeled, rooted binary trees is

$$N_R = \prod_{i=2}^{T} (2i - 3)$$

which expands to (2 · 2 – 3)(2 · 3 – 3)(2 · 4 – 3)…(2 · T – 3).

Number of Trees



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

What do branch lengths represent?

- They may be drawn of arbitrary length, if not specified
- Or: The estimated, expected, or (rarely) observed amount of character change, measured in some units (e.g., number of substitutions per site)
- Or: They may also indicate time, as estimated from a molecular clock analysis

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

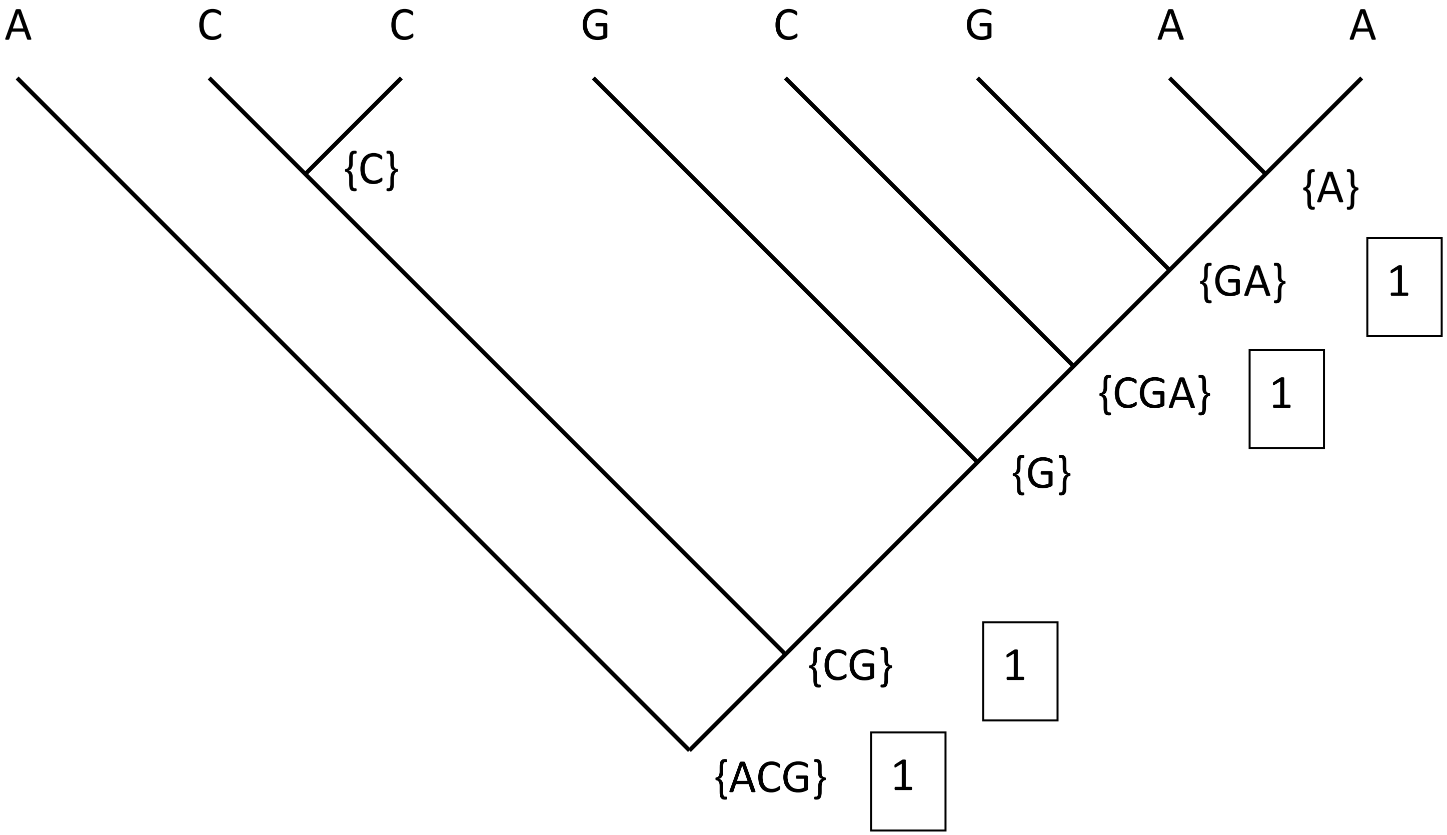
- Under the parsimony criterion, the optimal tree (the shortest or minimum-length tree) is the one that minimizes the sum of the lengths of all characters in terms of evolutionary steps (a step is a change from one character-state to another).
- For a given tree, find the length of each character, and 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

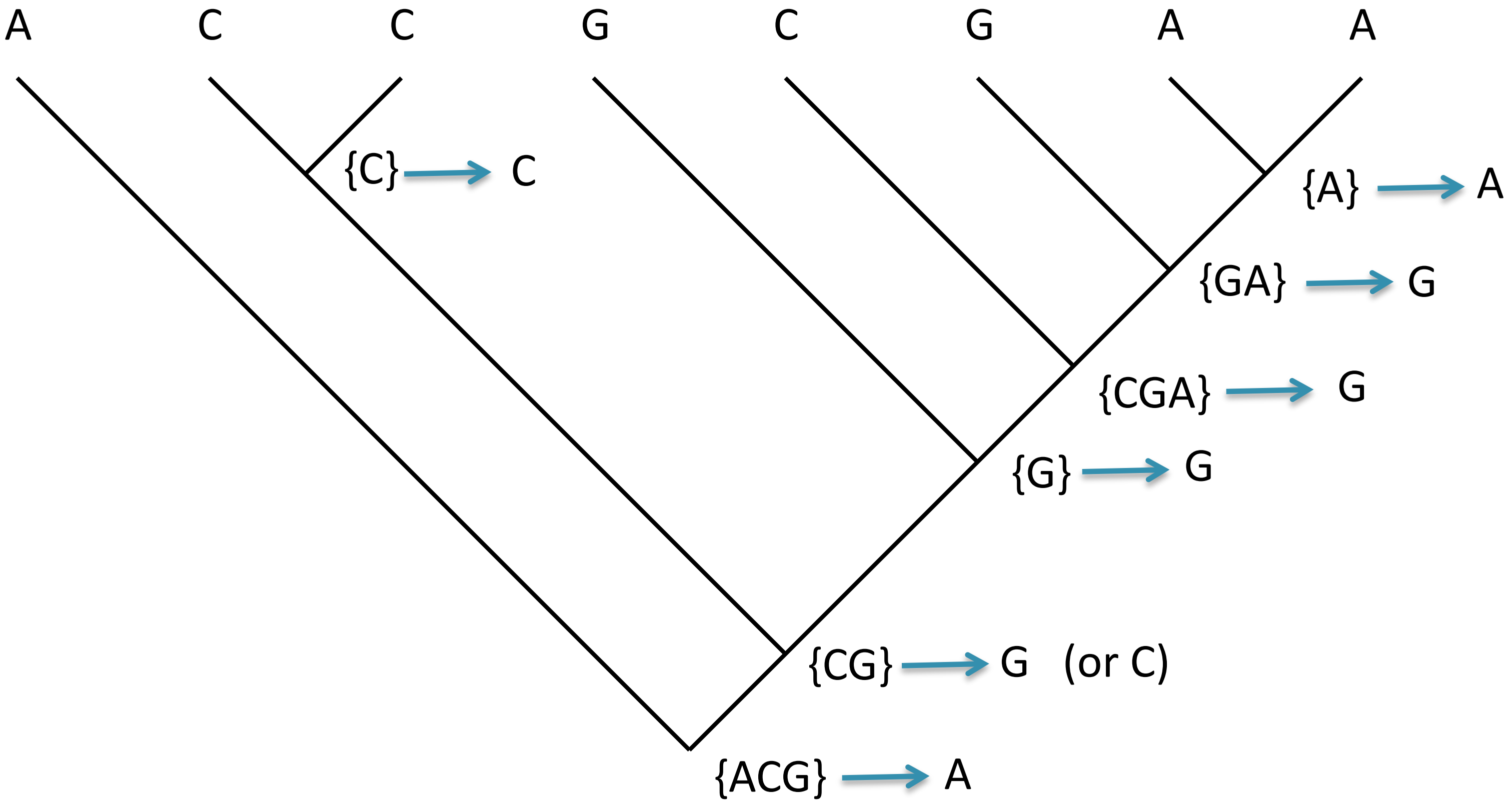
Optimization

- 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.



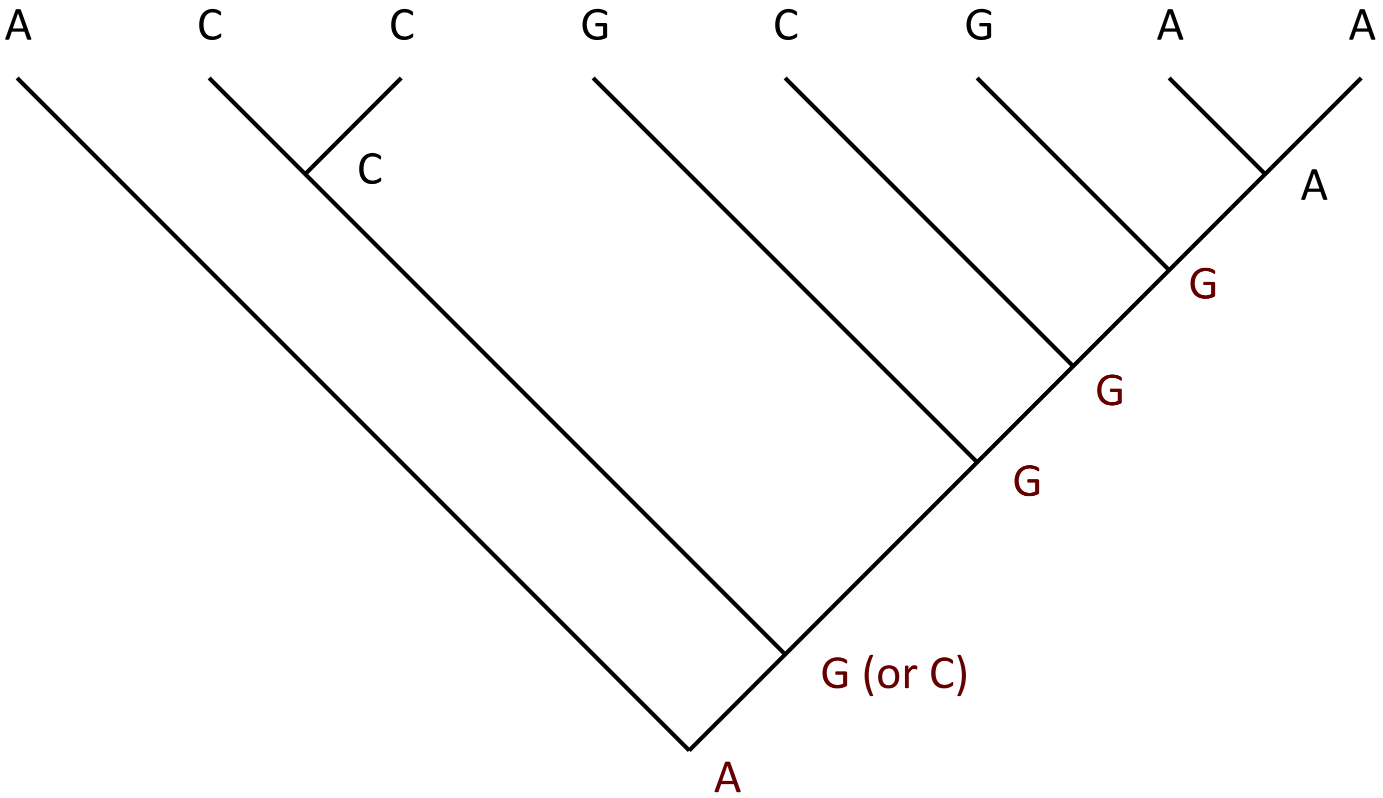
Downpass (postorder traversal)

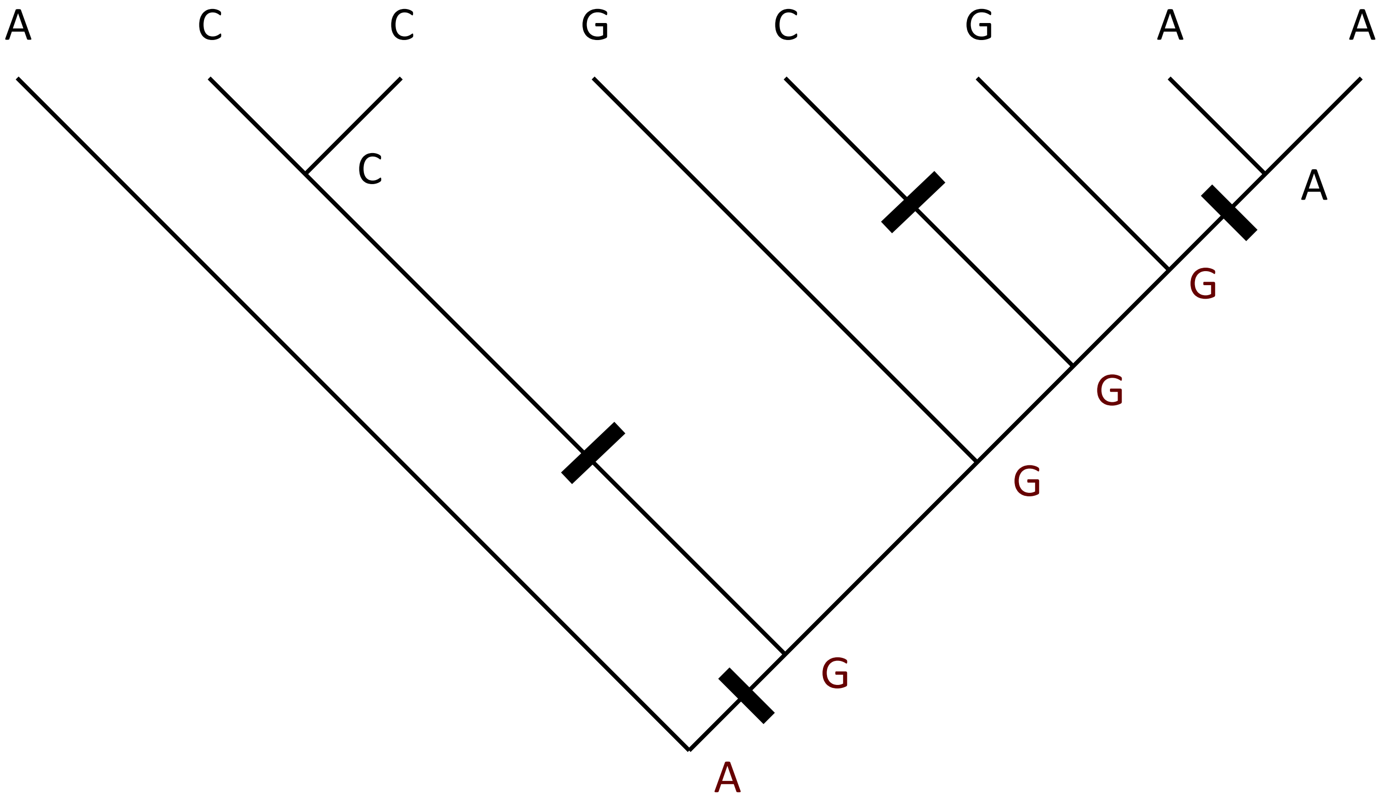
Length = 4

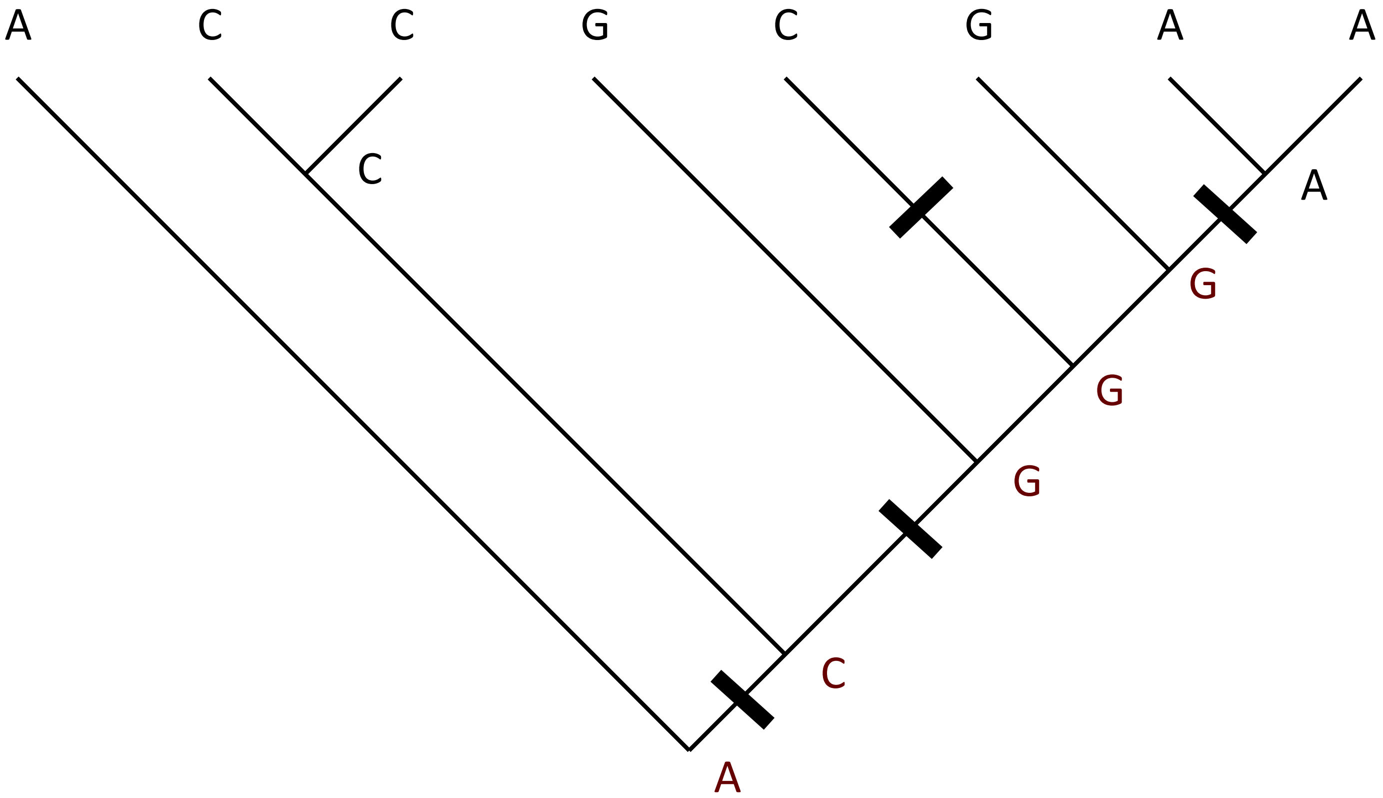


Up-Pass (preorder traversal)

Length = 4


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


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


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

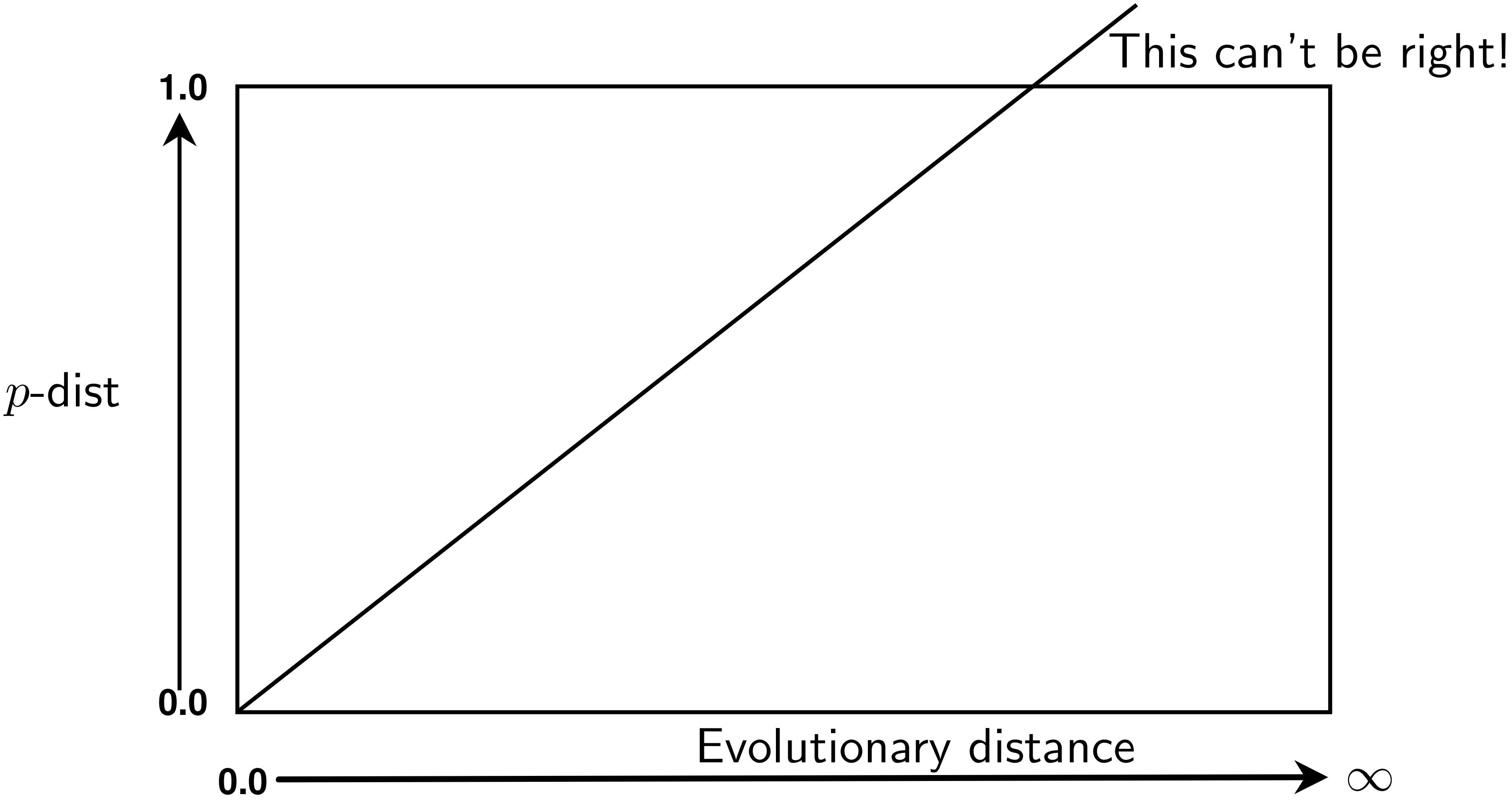
	Characters				
Taxa	1	2	3	4	5
one	A	G	C	G	A
two	A	G	C	G	T
three	C	T	C	G	T
four	C	T	C	A	A

→

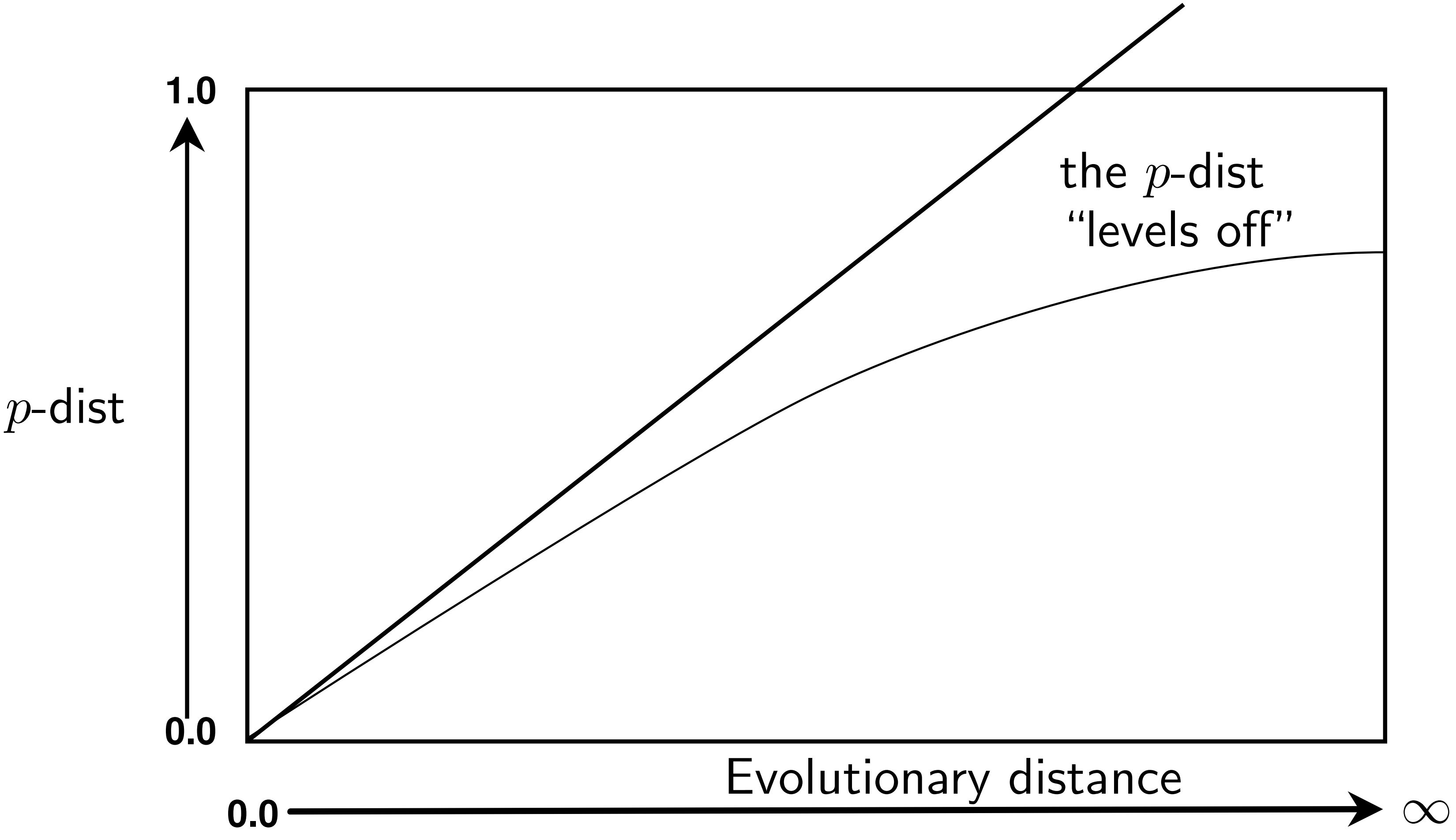
	one	two	three	four
one	-	.2	.6	.6
two		-	.4	.8
three			-	.4
four				-

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

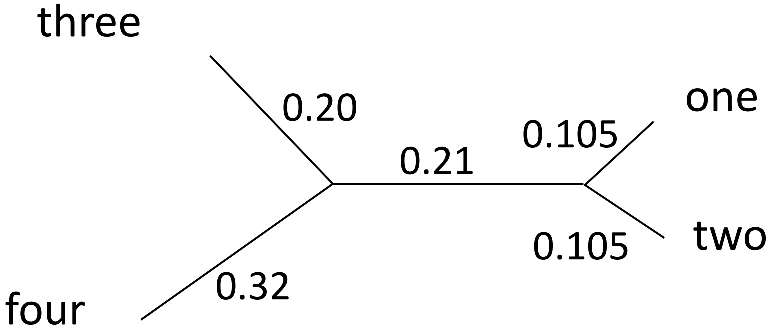
	one	two	three	four
one	-	.2	.6	.6
two		-	.4	.8
three			-	.4
four				-

Model of evolution

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

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…requires

  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).

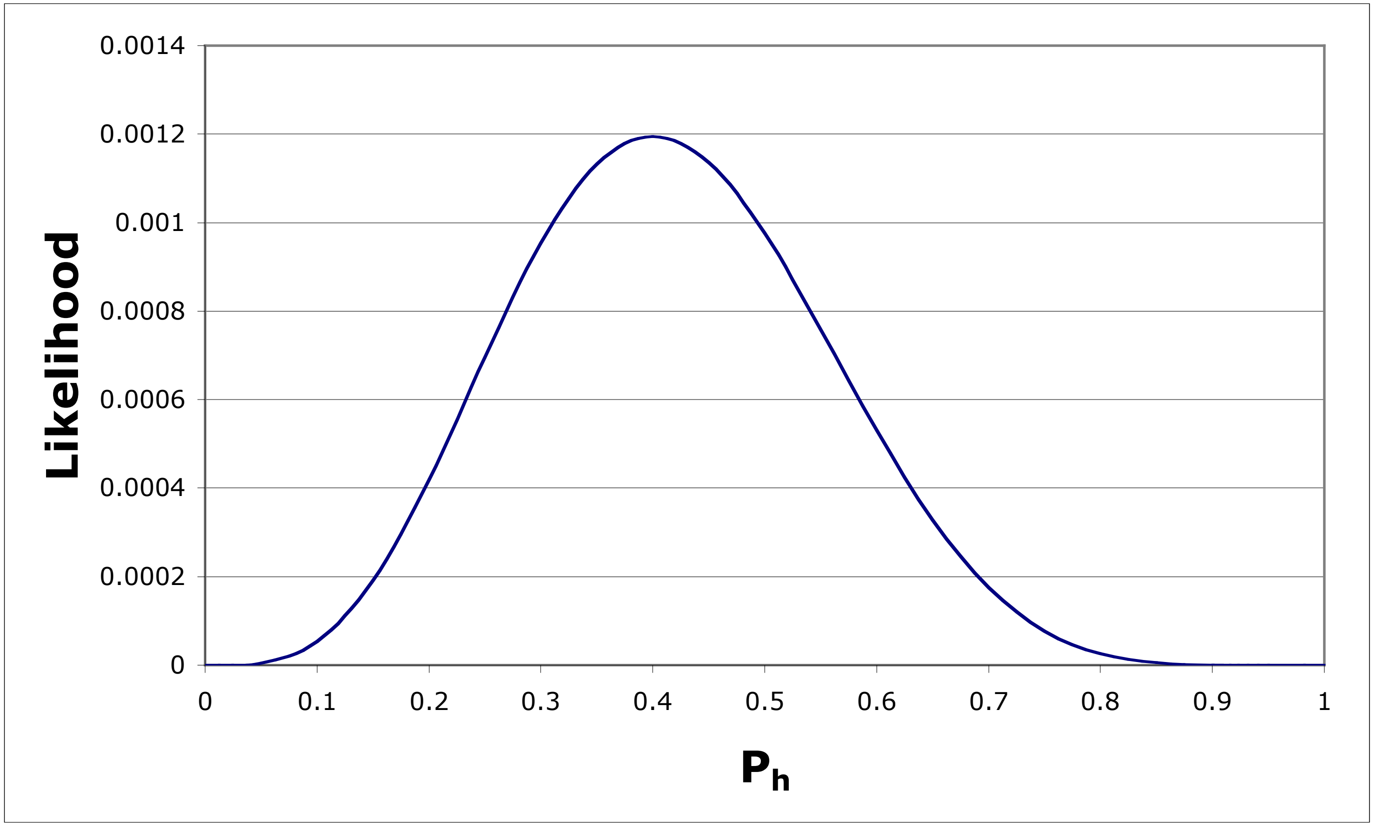
- Only the *relative* likelihood scores for various hypotheses, evaluated using the same data, are useful to us.

- What are some other models?

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

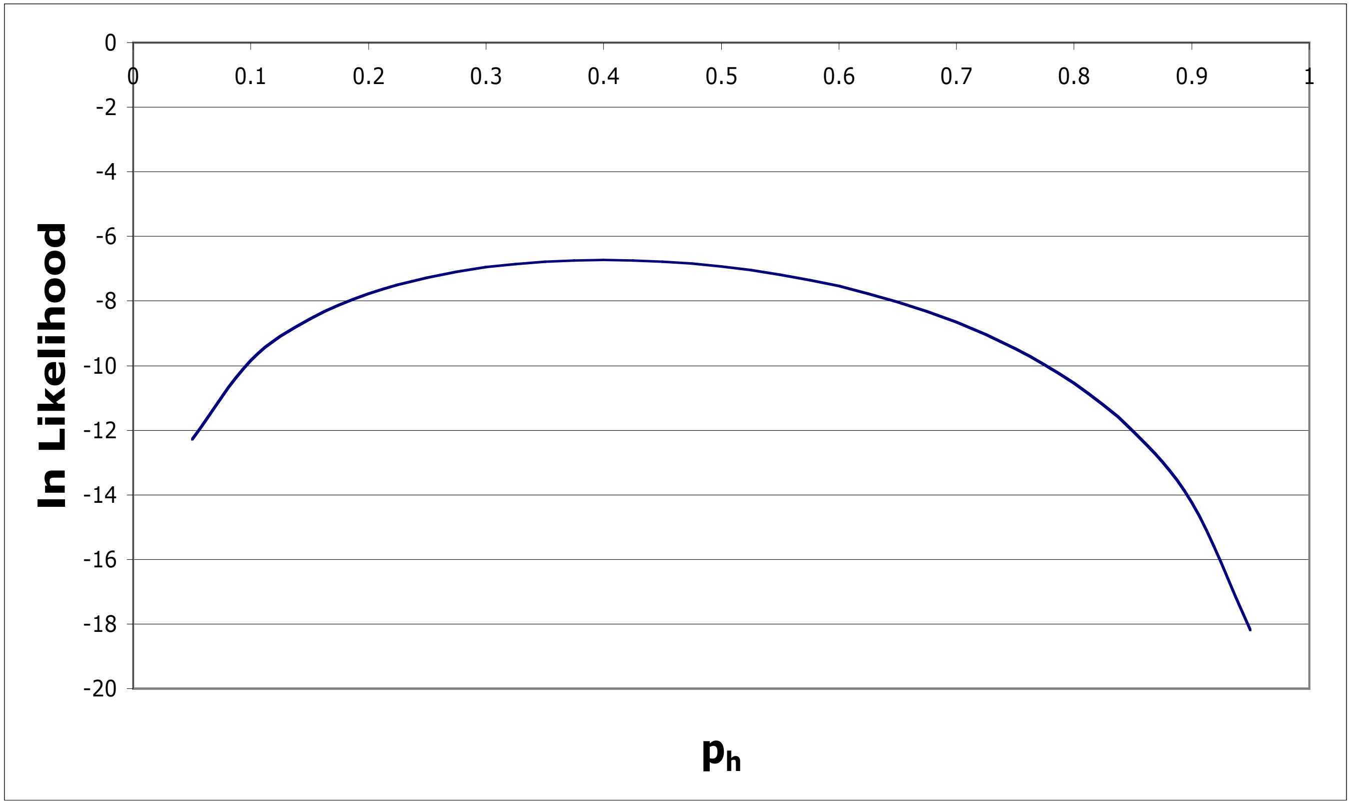
$$L(p_h = 0.4 \mid obs) = (0.4)^4 (0.6)^6 = 0.001194394$$

The likelihood surface



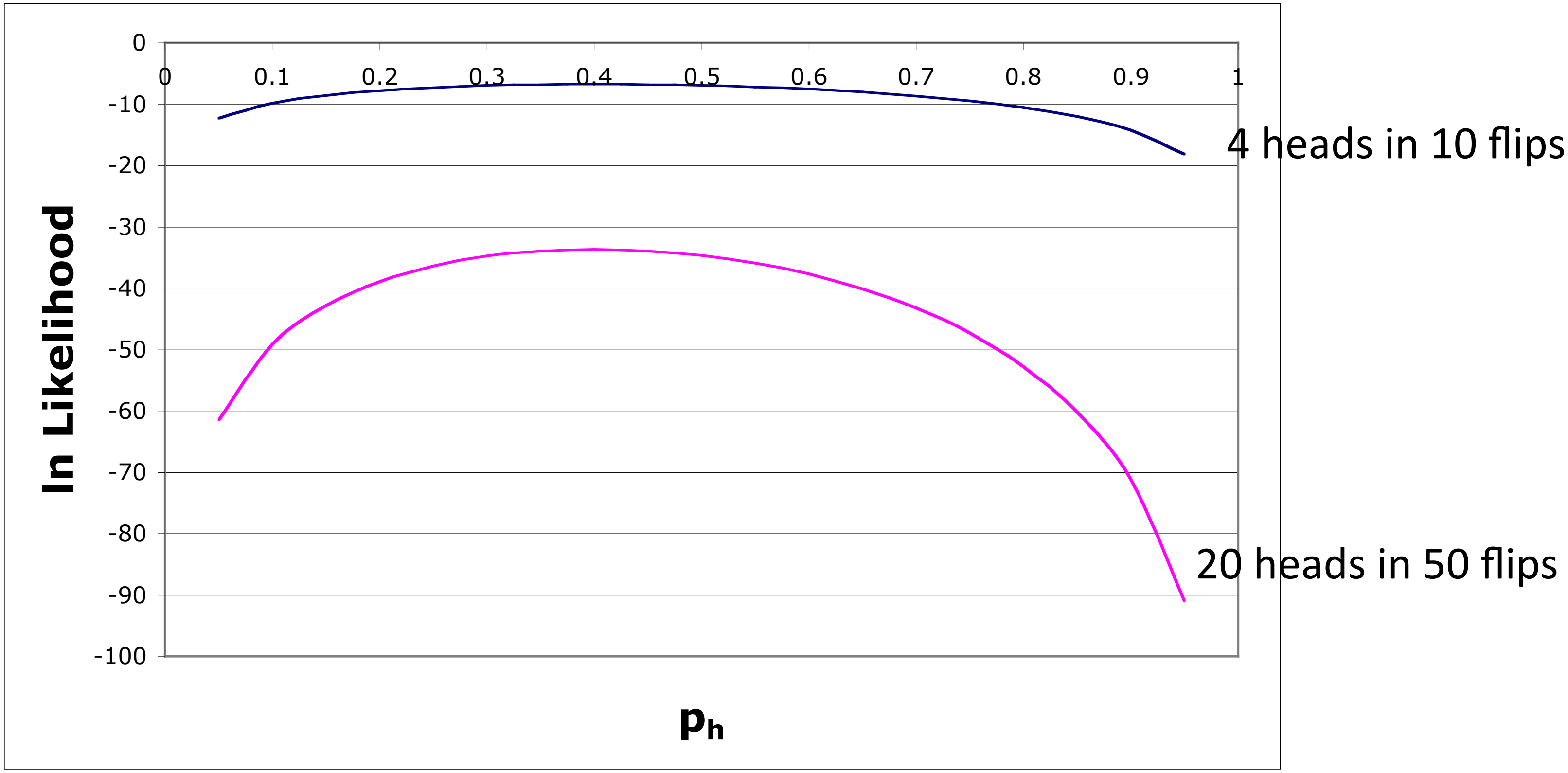
Data: H T H T T T H T T H

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

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

Posterior probability

Bayesian Approaches

- Take prior information into account

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

The support that *B* provides for *A*

Bayesian Approaches

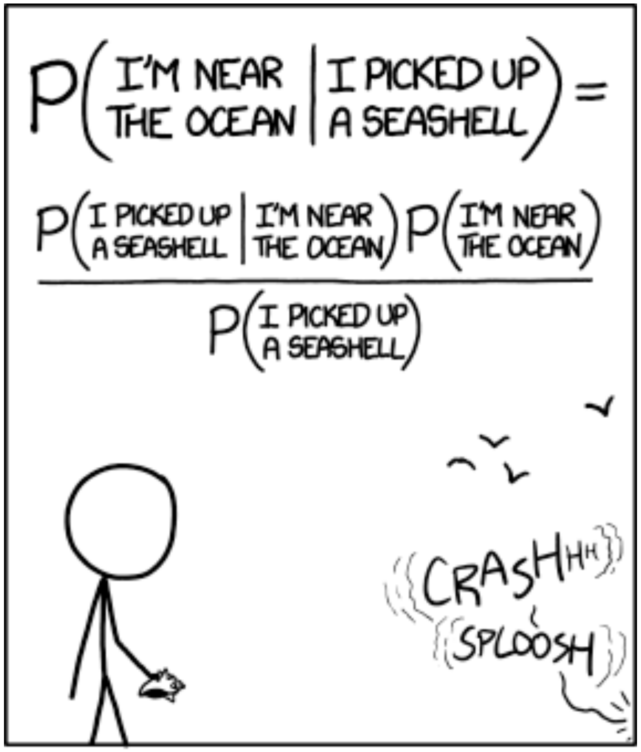
- Take prior information into account

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

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

Bayesian Approaches

- An example:


P(I'M NEAR THE OCEAN | I PICKED UP A SEASHELL) =
P(I PICKED UP A SEASHELL | I'M NEAR THE OCEAN) P(I'M NEAR THE OCEAN)
P(I PICKED UP A SEASHELL)
CRASHHH
SPLOOSH

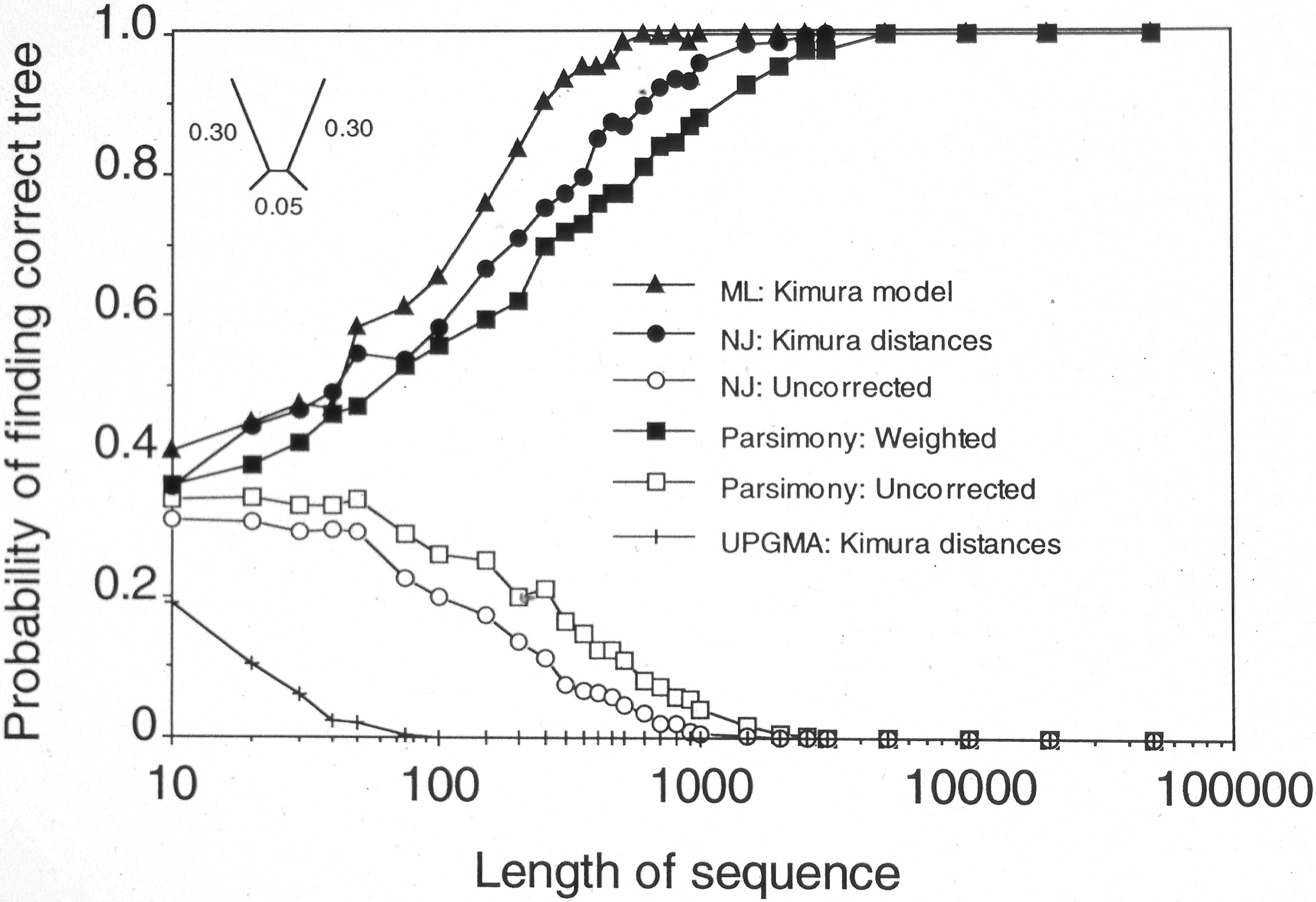
STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

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)


Probability of finding correct tree
1.0
0.8
0.6
0.4
0.2
0
10
100
1000
10000
100000
Length of sequence
0.30
0.30
0.05
ML: Kimura model
NJ: Kimura distances
NJ: Uncorrected
Parsimony: Weighted
Parsimony: Uncorrected
UPGMA: Kimura distances

What determines the accuracy of phylogenetic estimates?

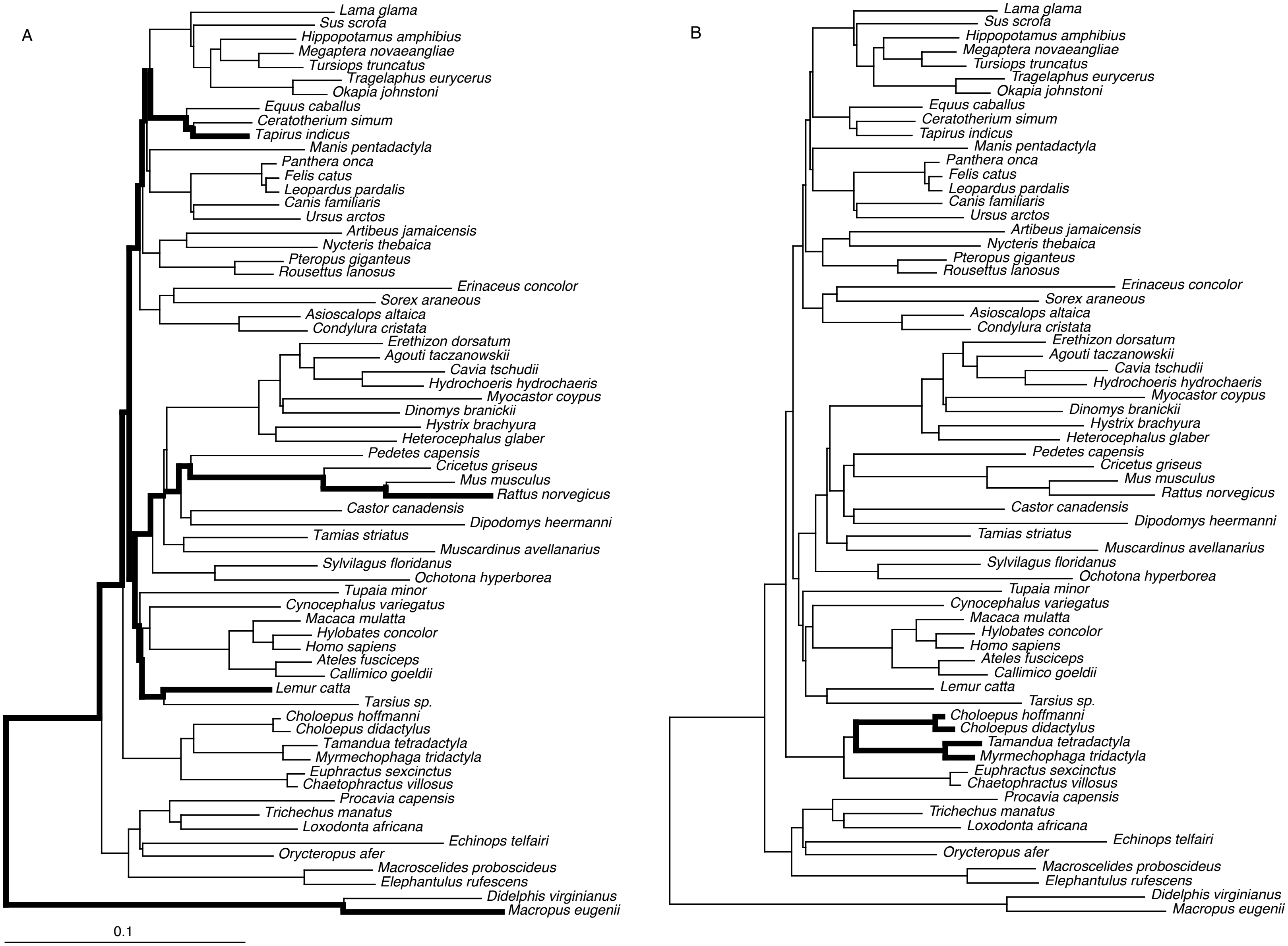
- Optimality criterion (likelihood-based methods have best performance, as long as assumptions of model are met)
- Model selection (finding an appropriate model of evolution for the data at hand)
- Sampling density of genes (more genes provide more information, and allow resolution of species trees from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

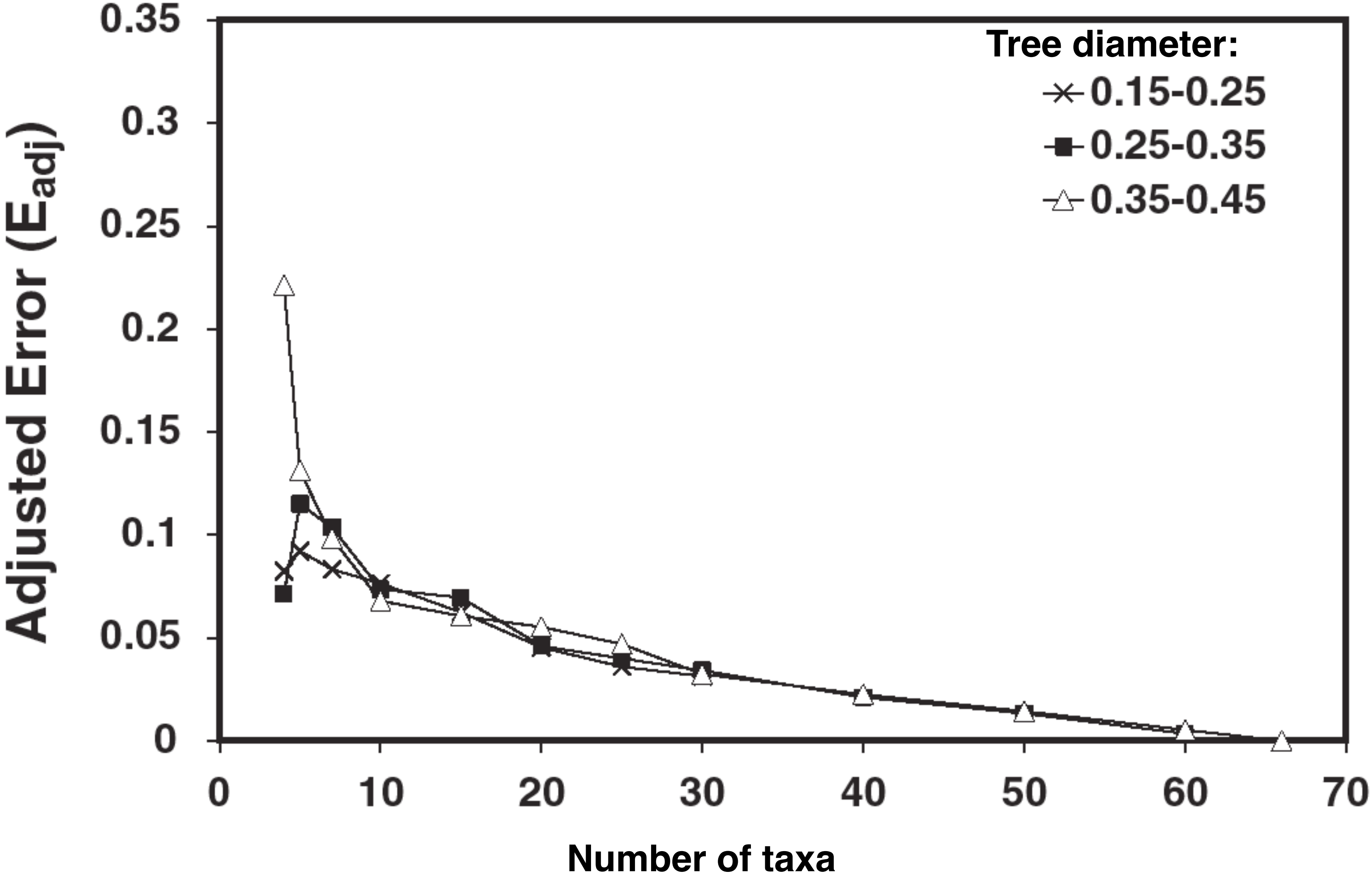
What determines the accuracy of phylogenetic estimates?

- 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)
- Sampling density of genes (more genes provide more information, and allow resolution of species trees from gene trees)
- 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)
- Sampling density of genes (more genes provide more information, and allow resolution of species trees from gene trees)
- Thoroughness of tree search (finding best solutions)
- Sampling density of taxa (more thorough taxon sampling produces better estimates of parameters and results in better estimates of trees)


A
Lama glama
Sus scrofa
Hippopotamus amphibius
Megaptera novaeangliae
Tursiops truncatus
Tragelaphus eurycerus
Okapia johnstoni
Equus caballus
Ceratotherium simum
Tapirus indicus
Manis pentadactyla
Panthera onca
Felis catus
Leopardus pardalis
Canis familiaris
Ursus arctos
Artibeus jamaicensis
Nycteris thebaica
Pteropus giganteus
Rousettus lanosus
Erinaceus concolor
Sorex araneous
Asioscalops altaica
Condylura cristata
Erethizon dorsatum
Agouti taczanowskii
Cavia tschudii
Hydrochoeris hydrochaeris
Myocastor coypus
Dinomys branickii
Hystrix brachyura
Heterocephalus glaber
Pedetes capensis
Cricetus griseus
Mus musculus
Rattus norvegicus
Castor canadensis
Dipodomys heermanni
Tamias striatus
Muscardinus avellanarius
Sylvilagus floridanus
Ochotona hyperborea
Tupaia minor
Cynocephalus variegatus
Macaca mulatta
Hylobates concolor
Homo sapiens
Ateles fusciceps
Callimico goeldii
Lemur catta
Tarsius sp.
Choloepus hoffmanni
Choloepus didactylus
Tamandua tetradactyla
Myrmechophaga tridactyla
Euphractus sexcinctus
Chaetophractus villosus
Procavia capensis
Trichechus manatus
Loxodonta africana
Echinops telfairi
Orycteropus afer
Macroscelides proboscideus
Elephantulus rufescens
Didelphis virginianus
Macropus eugenii
0.1
B
Lama glama
Sus scrofa
Hippopotamus amphibius
Megaptera novaeangliae
Tursiops truncatus
Tragelaphus eurycerus
Okapia johnstoni
Equus caballus
Ceratotherium simum
Tapirus indicus
Manis pentadactyla
Panthera onca
Felis catus
Leopardus pardalis
Canis familiaris
Ursus arctos
Artibeus jamaicensis
Nycteris thebaica
Pteropus giganteus
Rousettus lanosus
Erinaceus concolor
Sorex araneous
Asioscalops altaica
Condylura cristata
Erethizon dorsatum
Agouti taczanowskii
Cavia tschudii
Hydrochoeris hydrochaeris
Myocastor coypus
Dinomys branickii
Hystrix brachyura
Heterocephalus glaber
Pedetes capensis
Cricetus griseus
Mus musculus
Rattus norvegicus
Castor canadensis
Dipodomys heermanni
Tamias striatus
Muscardinus avellanarius
Sylvilagus floridanus
Ochotona hyperborea
Tupaia minor
Cynocephalus variegatus
Macaca mulatta
Hylobates concolor
Homo sapiens
Ateles fusciceps
Callimico goeldii
Lemur catta
Tarsius sp.
Choloepus hoffmanni
Choloepus didactylus
Tamandua tetradactyla
Myrmechophaga tridactyla
Euphractus sexcinctus
Chaetophractus villosus
Procavia capensis
Trichechus manatus
Loxodonta africana
Echinops telfairi
Orycteropus afer
Macroscelides proboscideus
Elephantulus rufescens
Didelphis virginianus
Macropus eugenii


Adjusted Error (E_{adj})
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
70
Number of taxa
Tree diameter:
0.15-0.25
0.25-0.35
0.35-0.45