Phylogenetic terminology and applications

Emily Jane McTavish

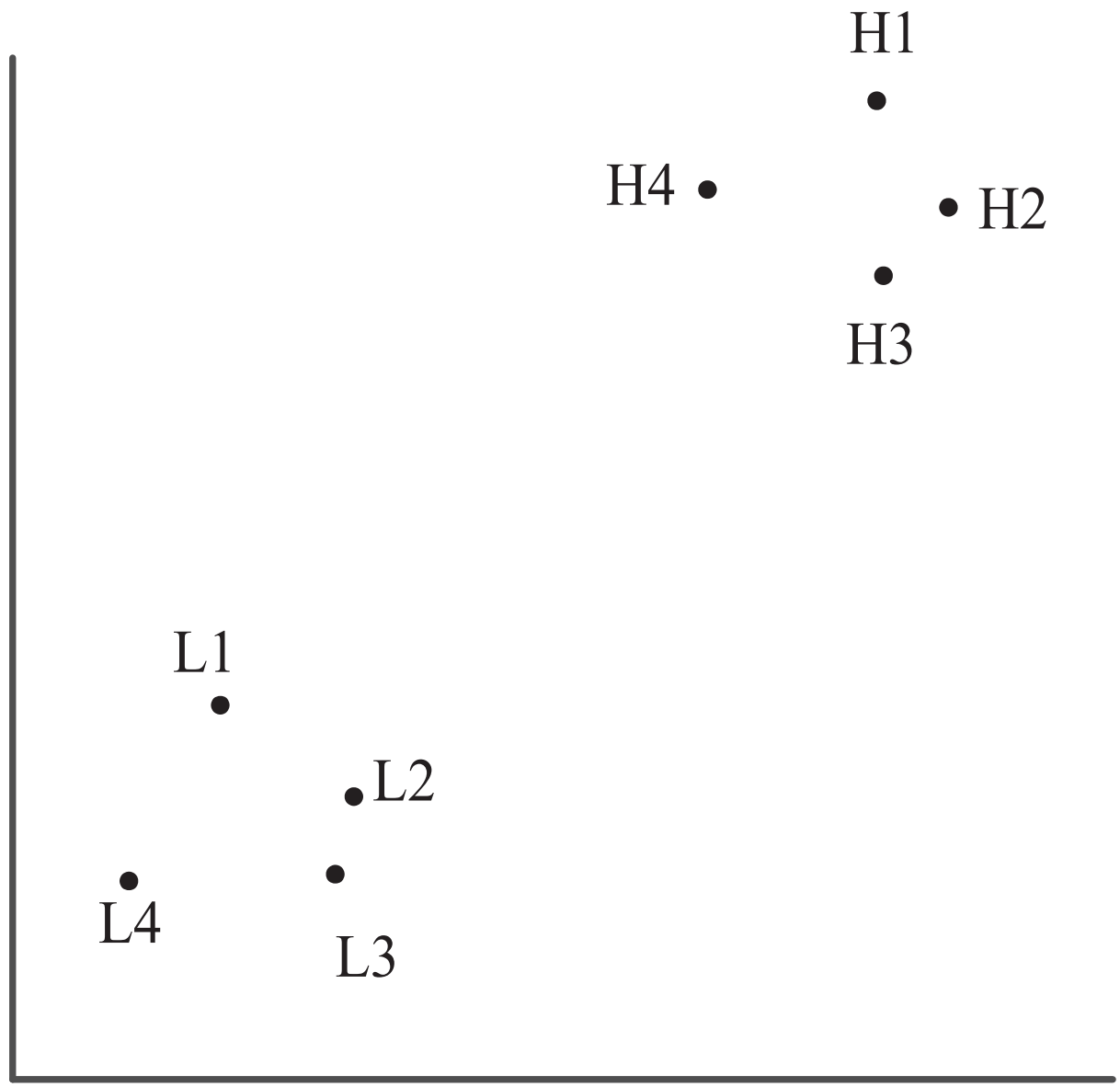
Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

(With thanks to Mark Holder, Paul Lewis, Joe Felsenstein, and David Hillis for slides)

Phylogenies describe shared ancestry
and
inform our understanding of evolutionary processes

Simple test of Bergmann's rule: comparing latitude and mass (I made these data up)
lat. offset = degrees north of the 49th parallel.

species	lat. offset	mass
L1	3.1	5.9
L2	5.4	4.3
L3	5.1	3.1
L4	1.8	3.6
H1	13.5	15.2
H2	14.6	13.5
H3	13.6	12.4
H4	10.8	13.7

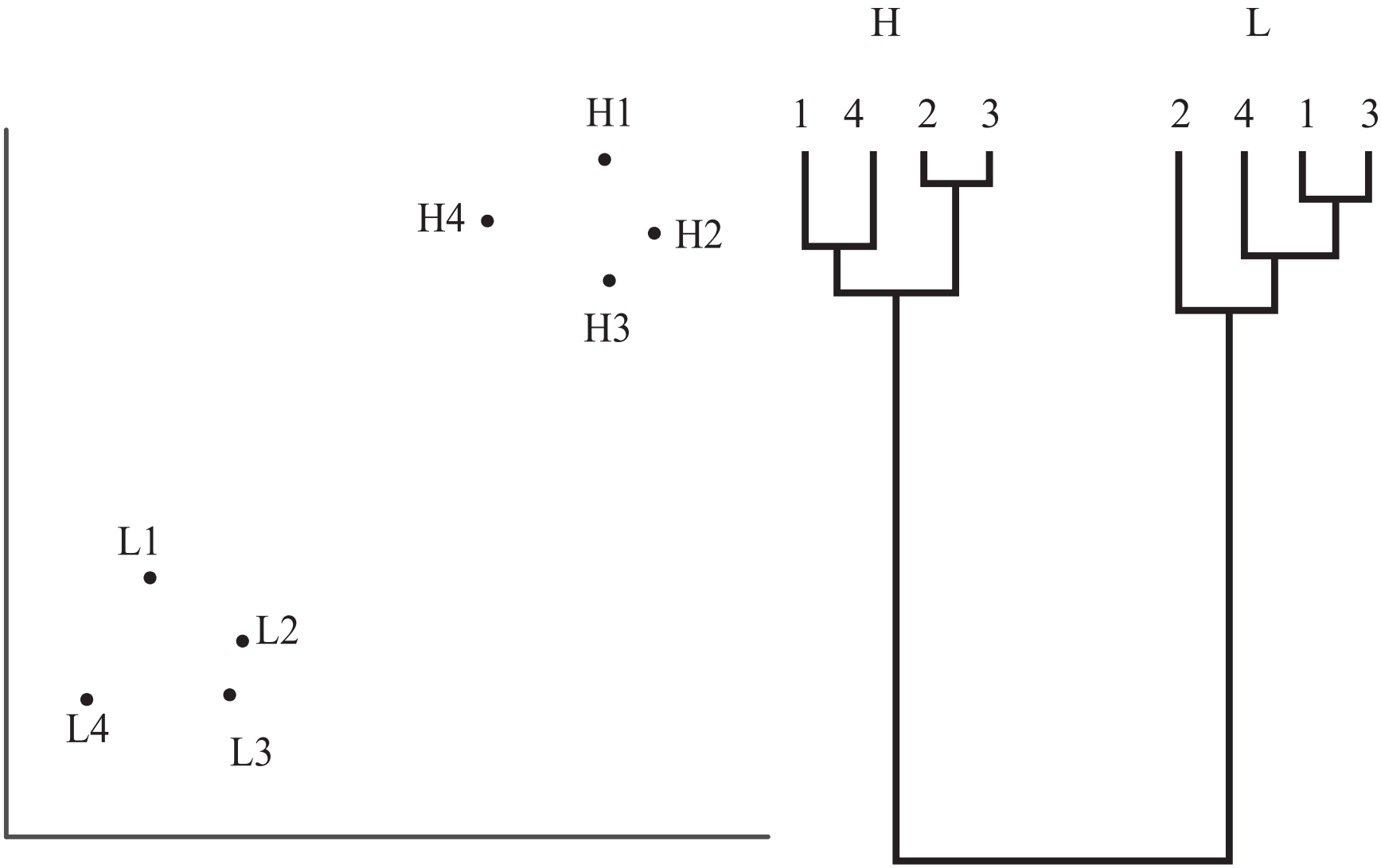

H1
H4
H2
H3
L1
L2
L4
L3

(cue cartoon videos)

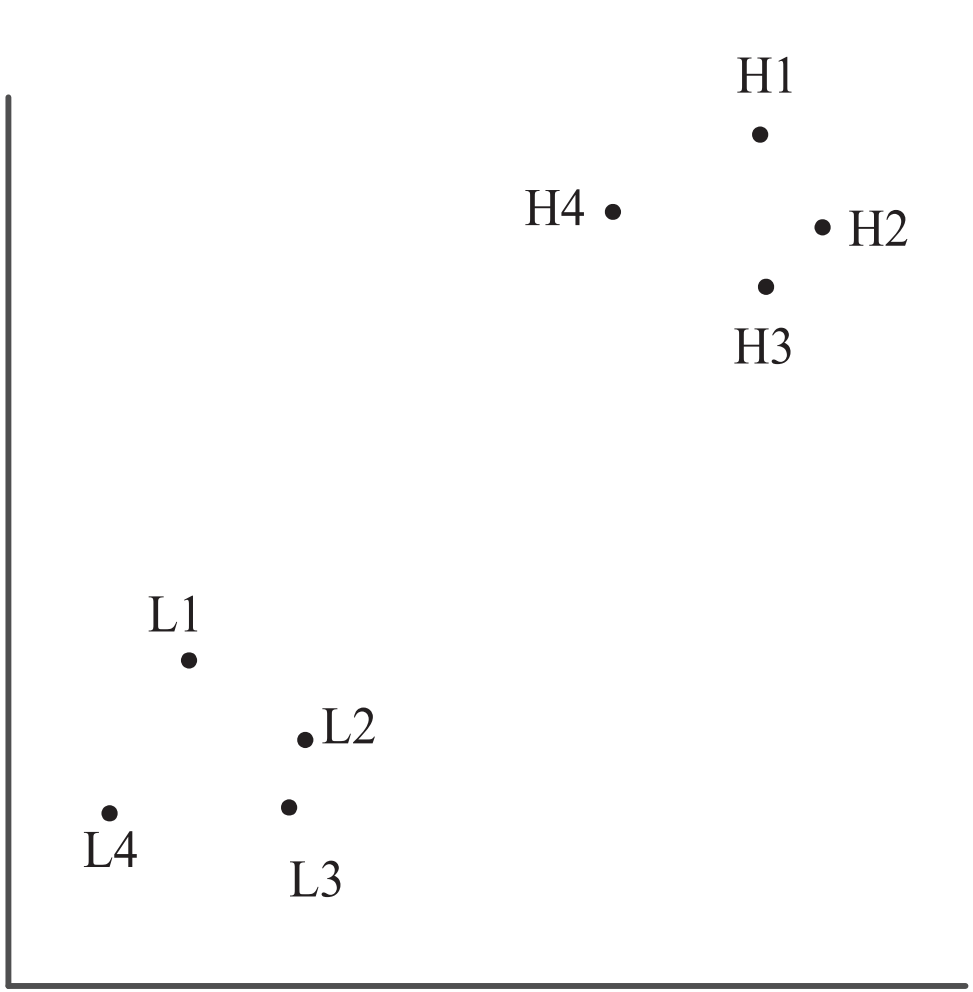
See http://phylo.bio.ku.edu/slides/no-correl-anim.mov

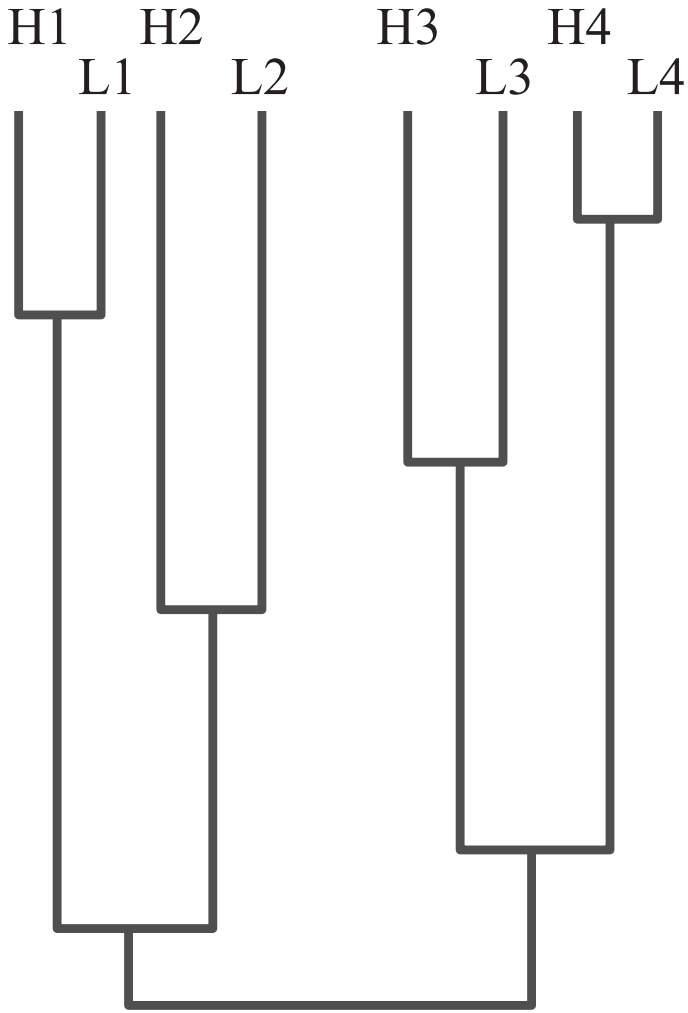
and http://phylo.bio.ku.edu/slides/correl-anim2.mov

No (or little) evidence for correlation



Evidence for correlation

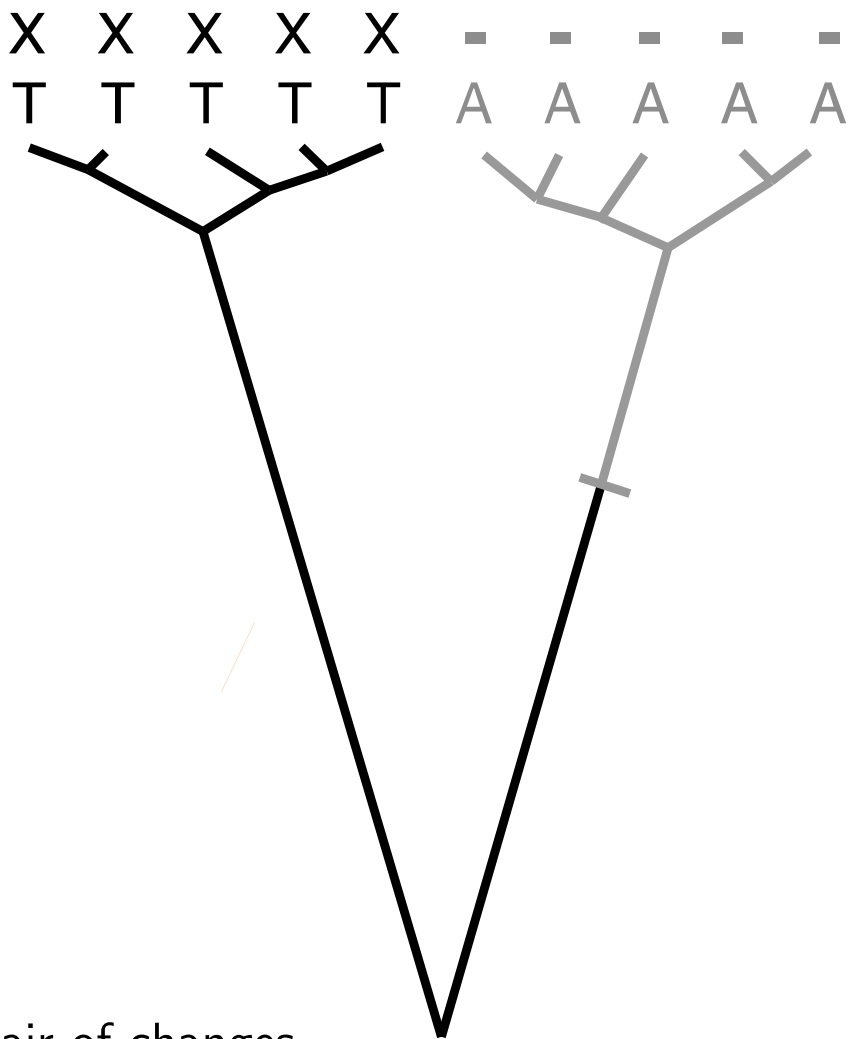




Do desert green algae use xanthophyll to protect against excessive light intensities?

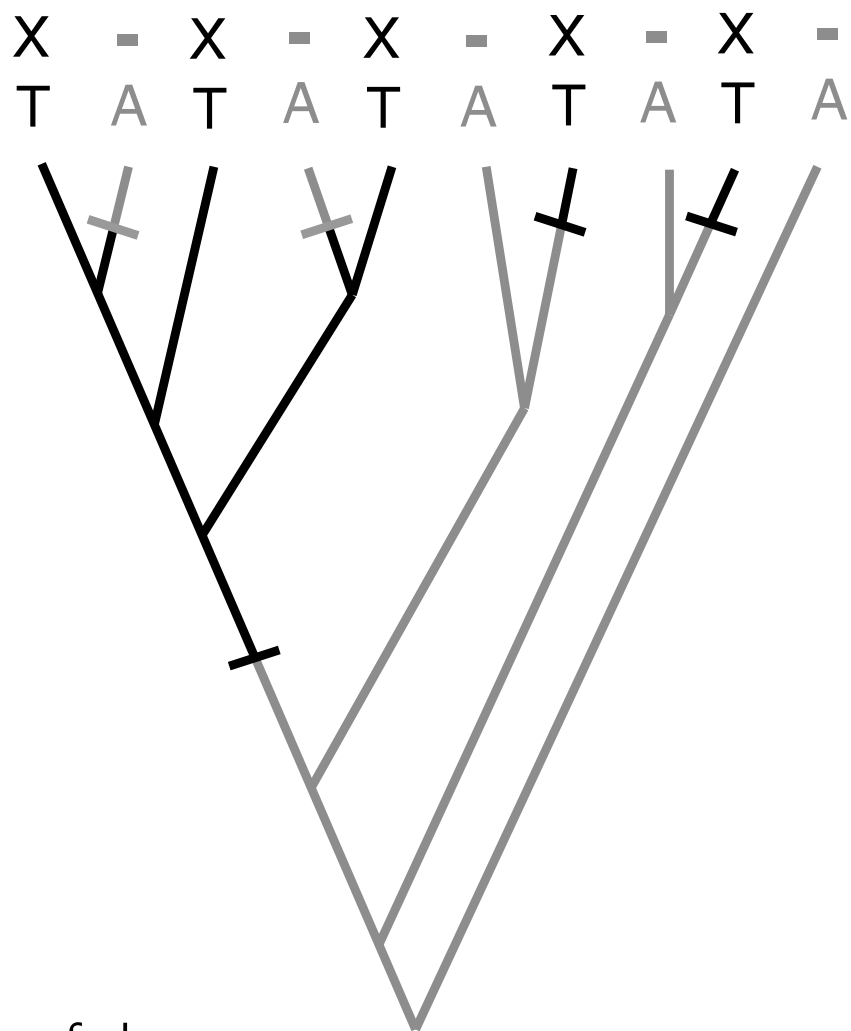
Species	Habitat	Photoprotection
1	terrestrial	xanthophyll
2	terrestrial	xanthophyll
3	terrestrial	xanthophyll
4	terrestrial	xanthophyll
5	terrestrial	xanthophyll
6	aquatic	none
7	aquatic	none
8	aquatic	none
9	aquatic	none
10	aquatic	none

Phylogeny reveals the events that generate the pattern



1 pair of changes.

Coincidence?



5 pairs of changes.

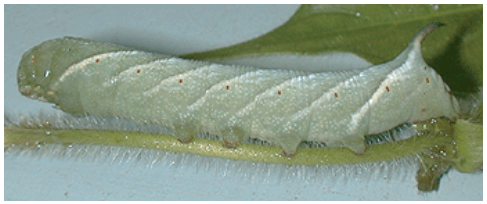
Much more convincing

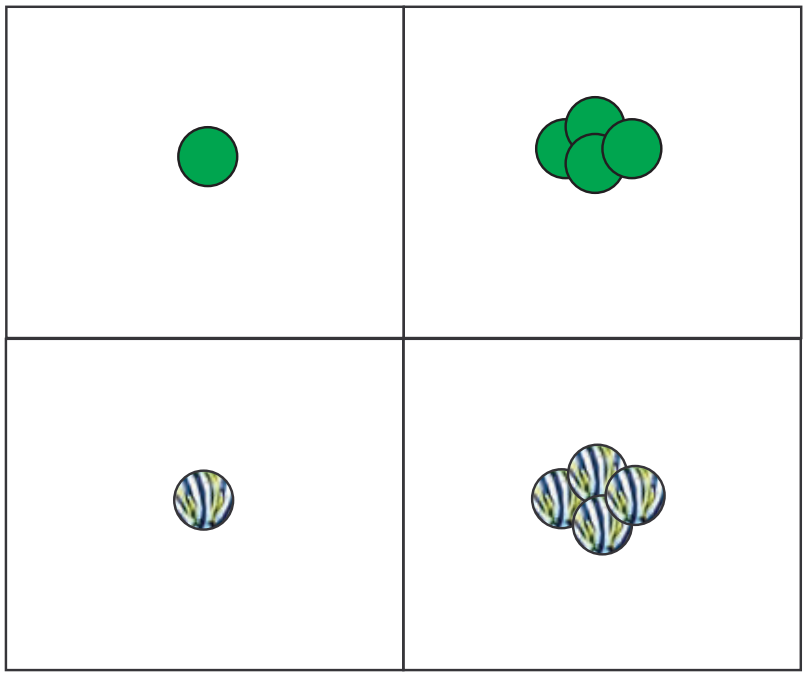
Inferring Process from Pattern

Hypothesis:

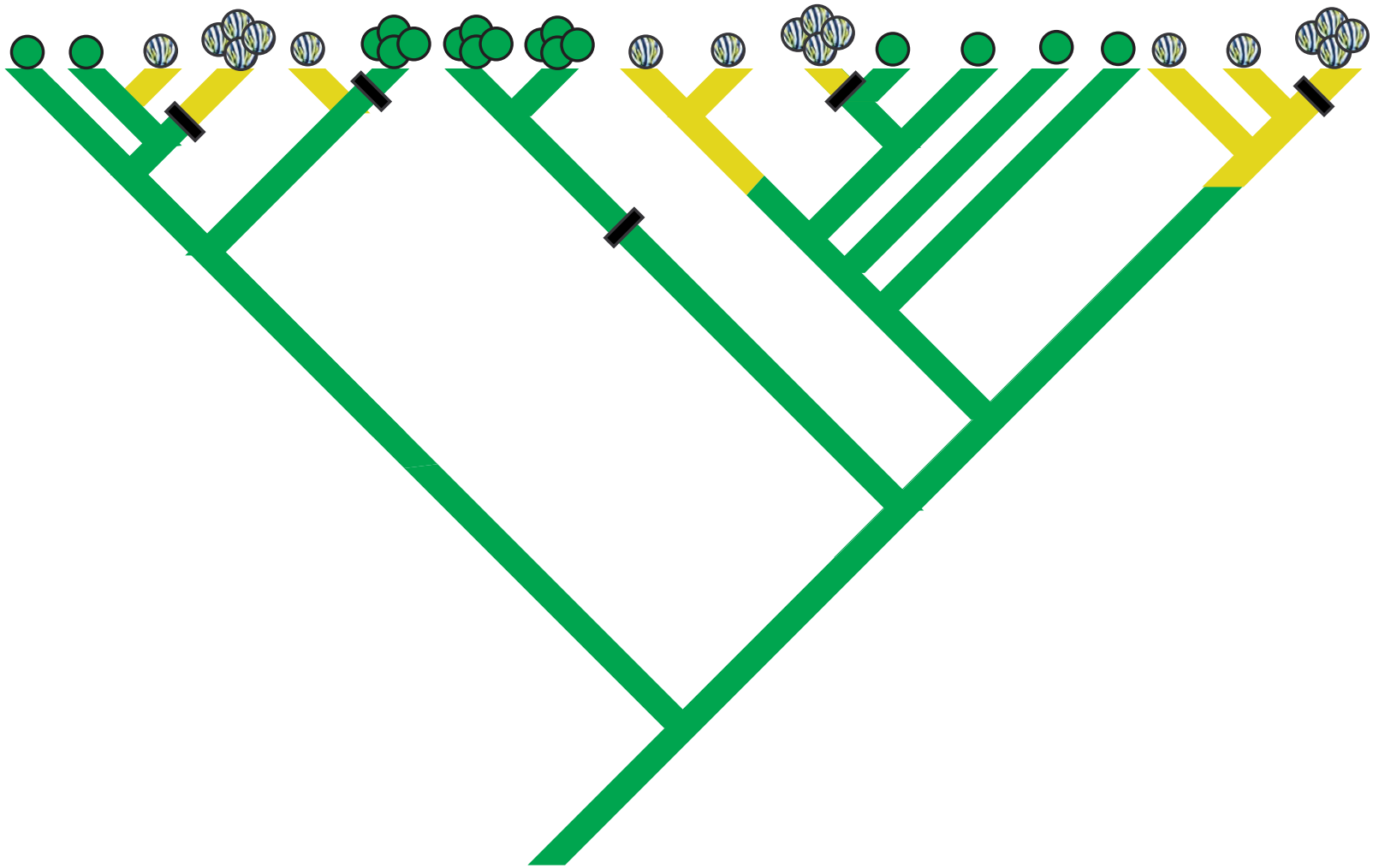
Gregariousness should arise more frequently in unpalatable organisms than in tasty ones (Sillén-Tullberg, 1988)

Inferring Process from Pattern



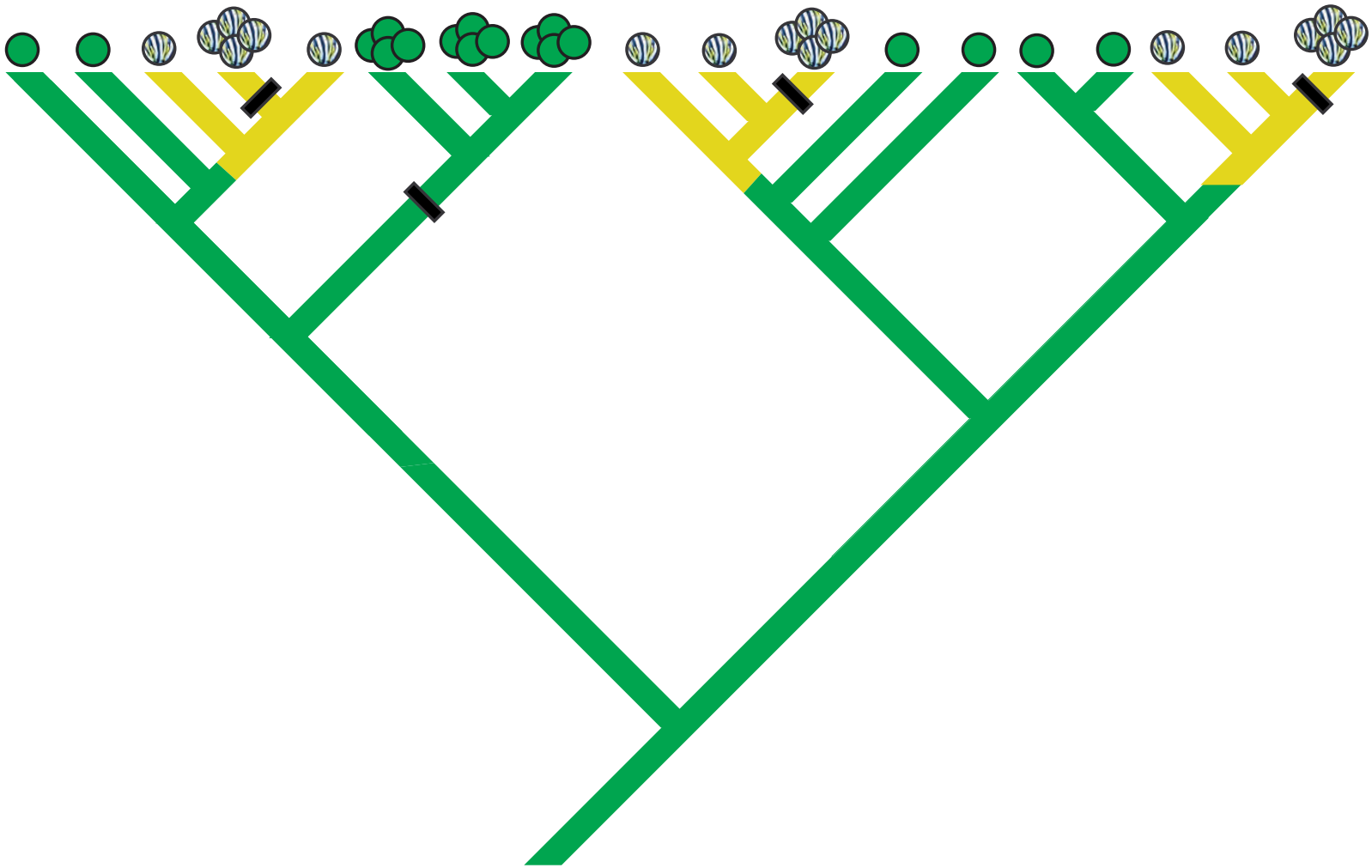


Sillén-Tullberg (1988), Dyer and Gentry (2002), Hill (2001)



One possible outcome:

No clear evidence of associations between traits



Cartoon of the real results (Sillén-Tullberg, 1988)

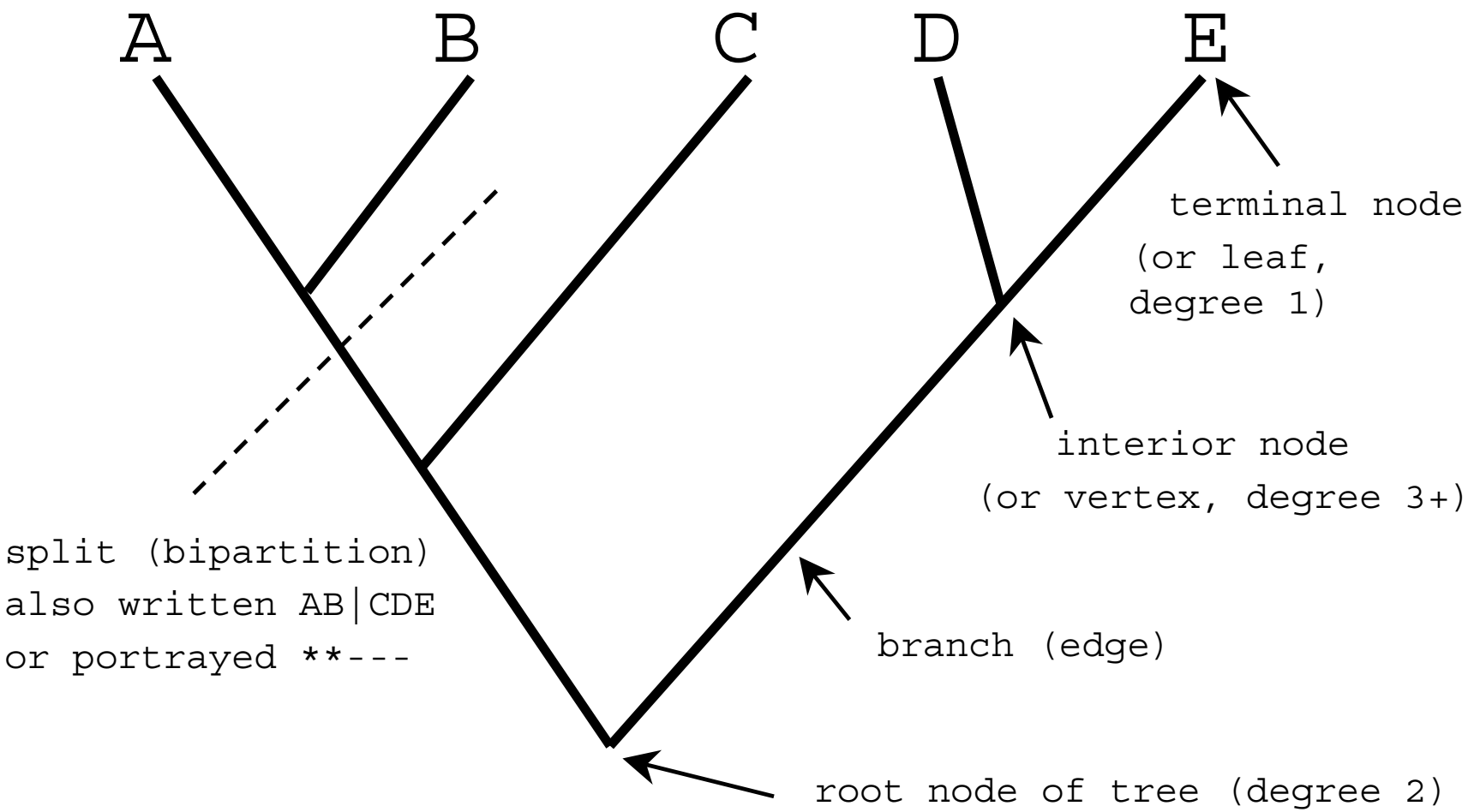
Aposematic species are more likely to evolve gregarious larvae

Importance of phylogeny

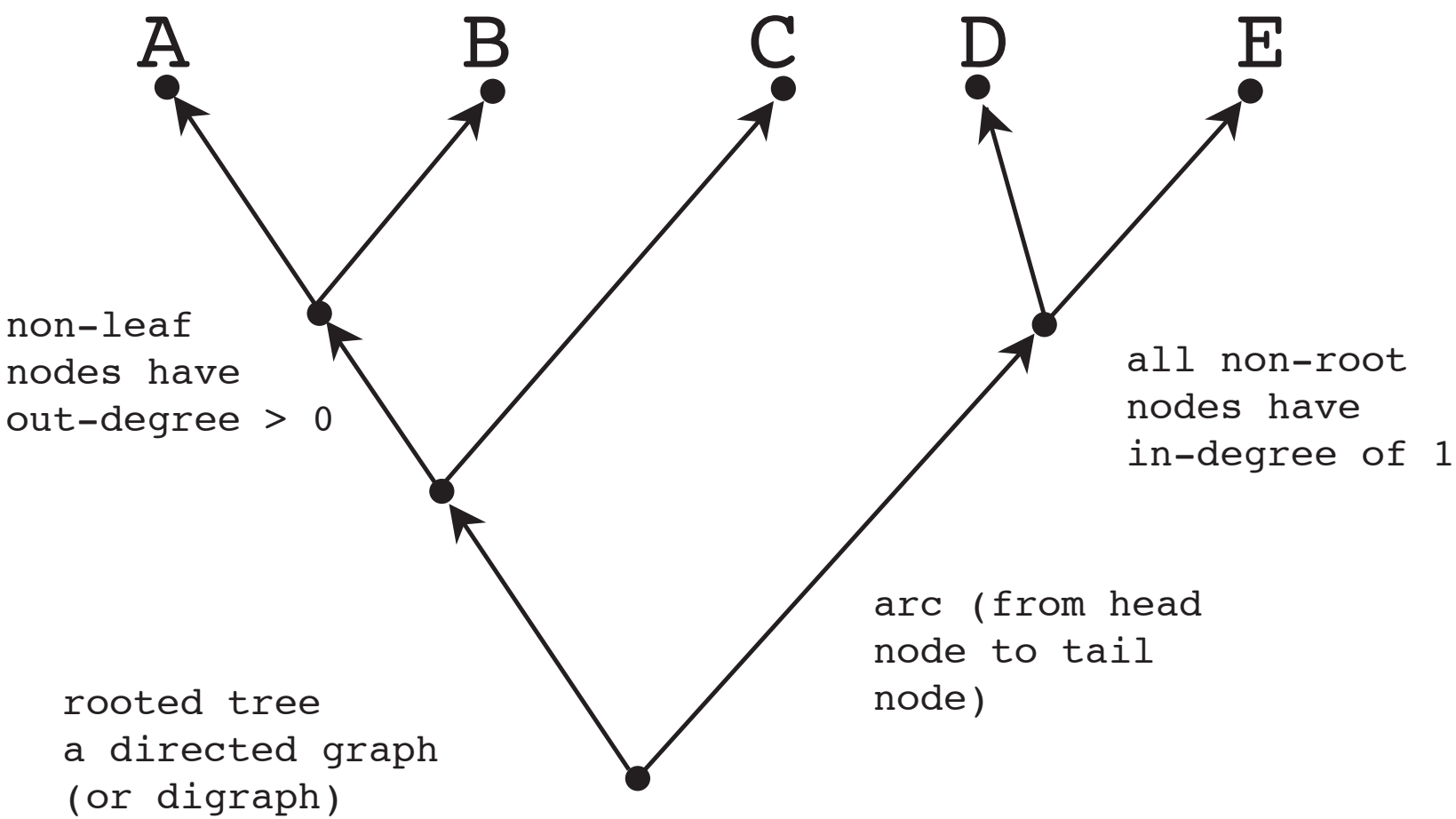
The previous slides had identical patterns of traits if the phylogeny is ignored.

Without knowledge of the tree, no conclusion would be reached.

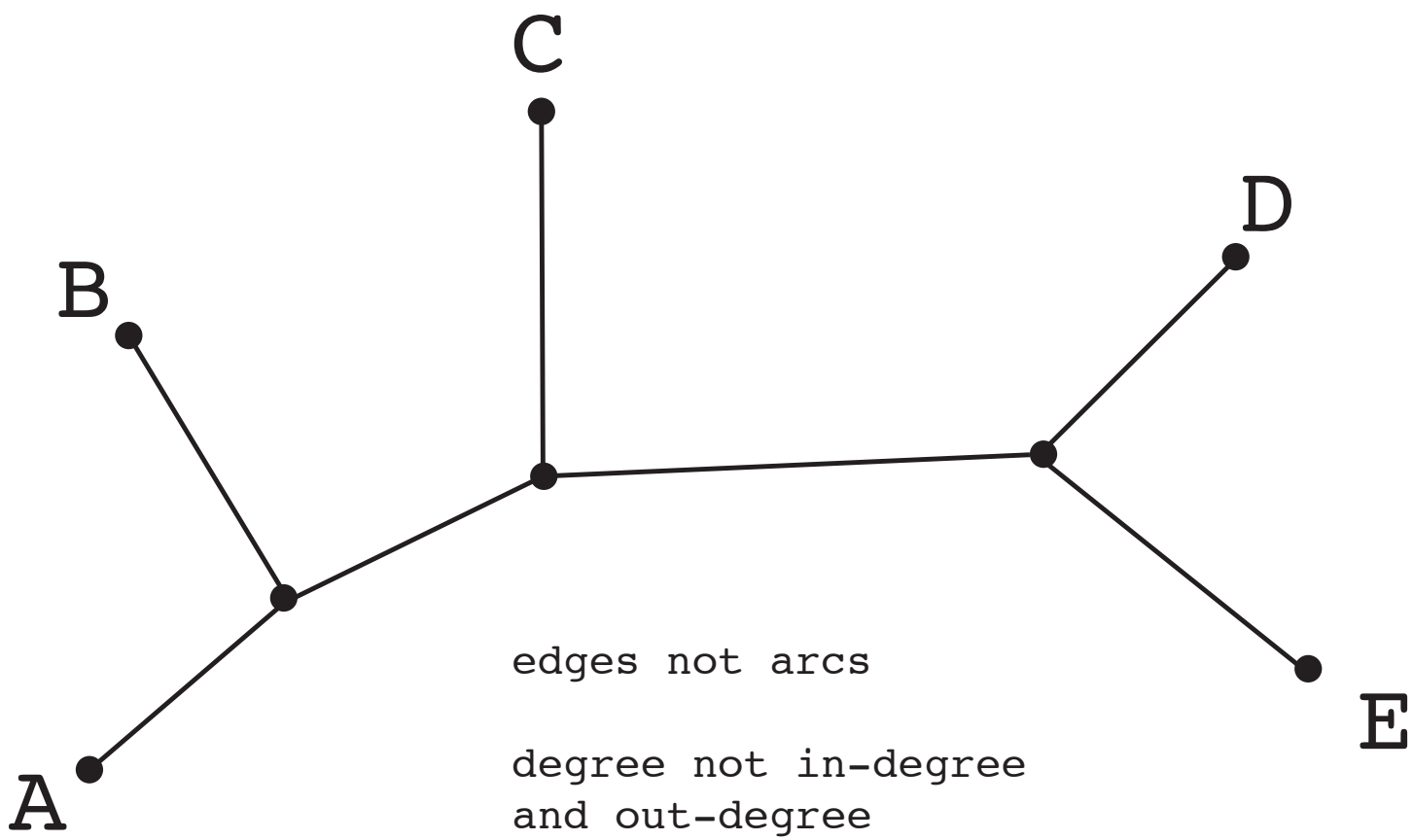
Tree terminology



Rooted tree terminology



Rooted tree terminology



Tree terms

A tree is a connected, acyclic graph.

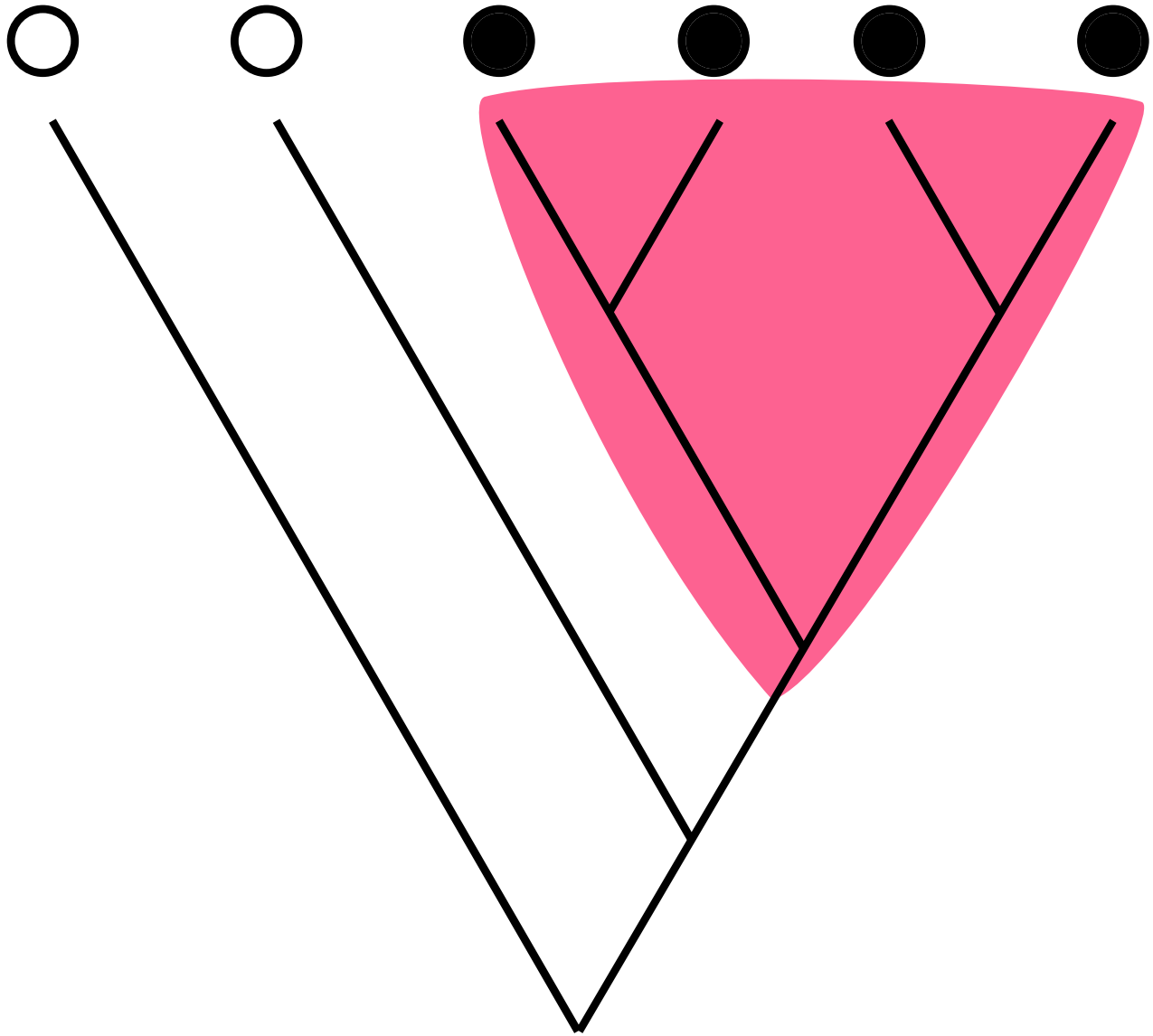
A rooted tree is a connected, acyclic directed graph.

A polytomy or multifurcation is a node with a degree > 3 (in an unrooted tree), or a node with an out-degree > 2 (in a rooted tree).

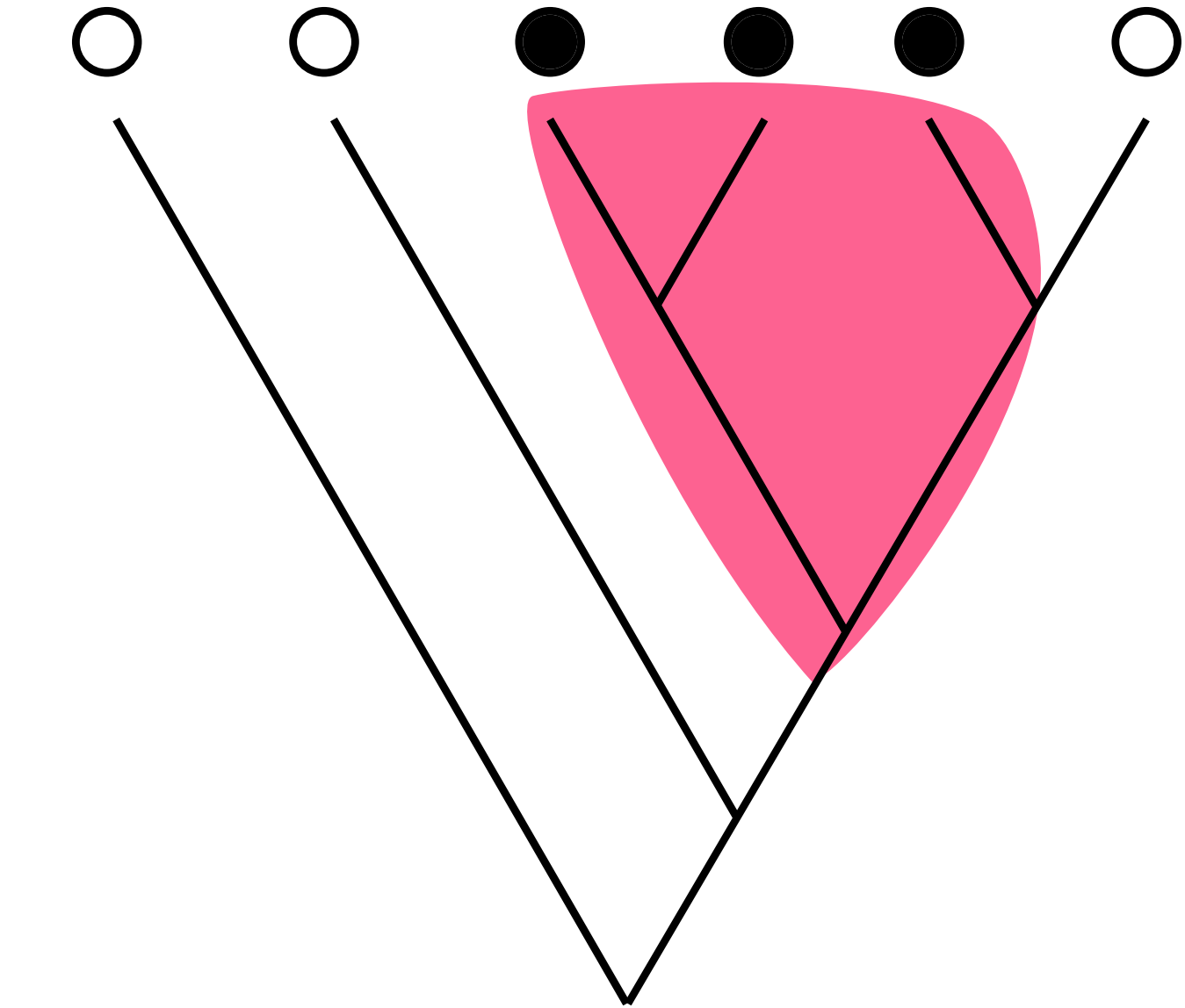
Collapsing an edge means to merge the nodes at the end of the branch (resulting in a polytomy in most cases).

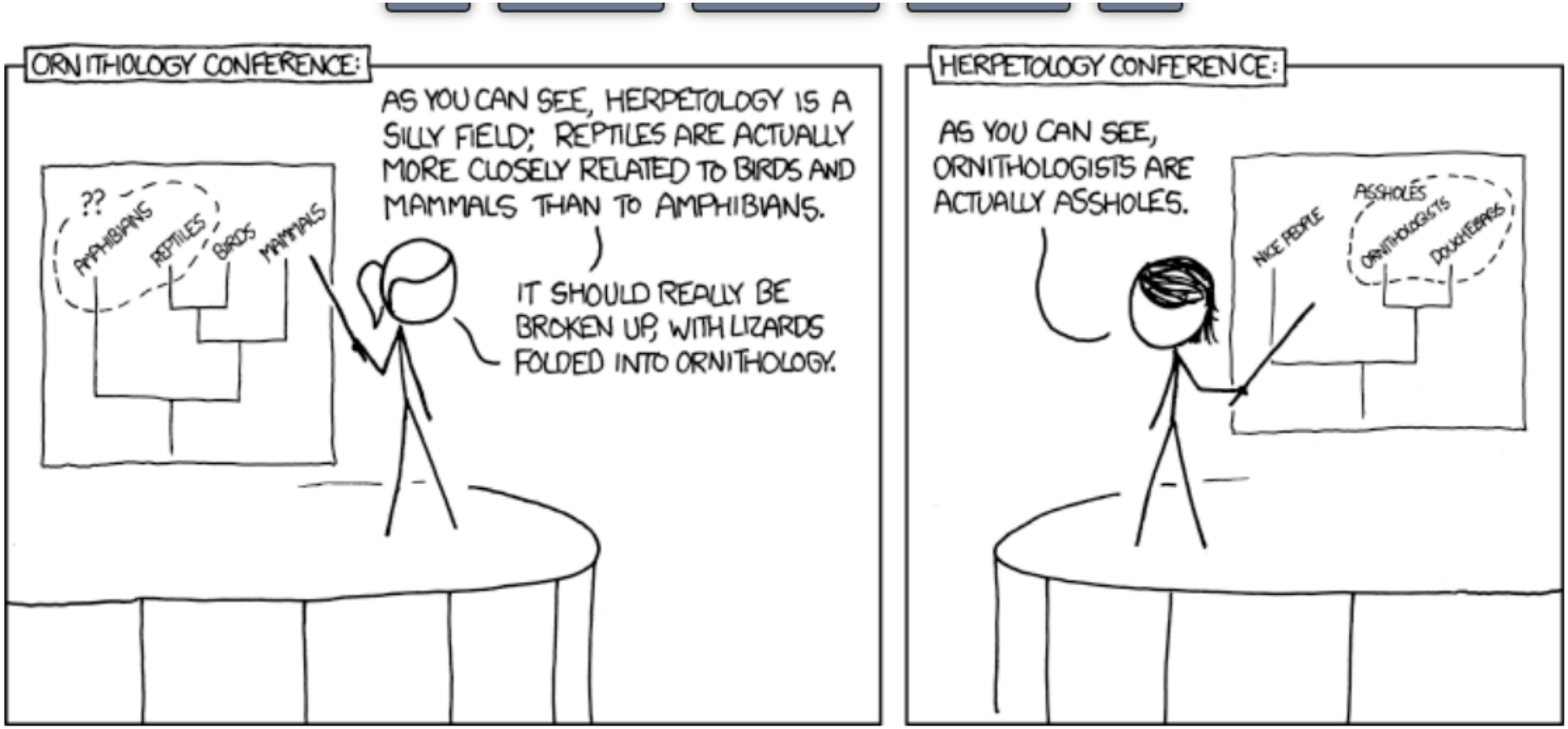
Refining a polytomy means to "break" the node into two nodes that are connected by an edge.

Monophyletic groups ("clades"): the basis of phylogenetic classification



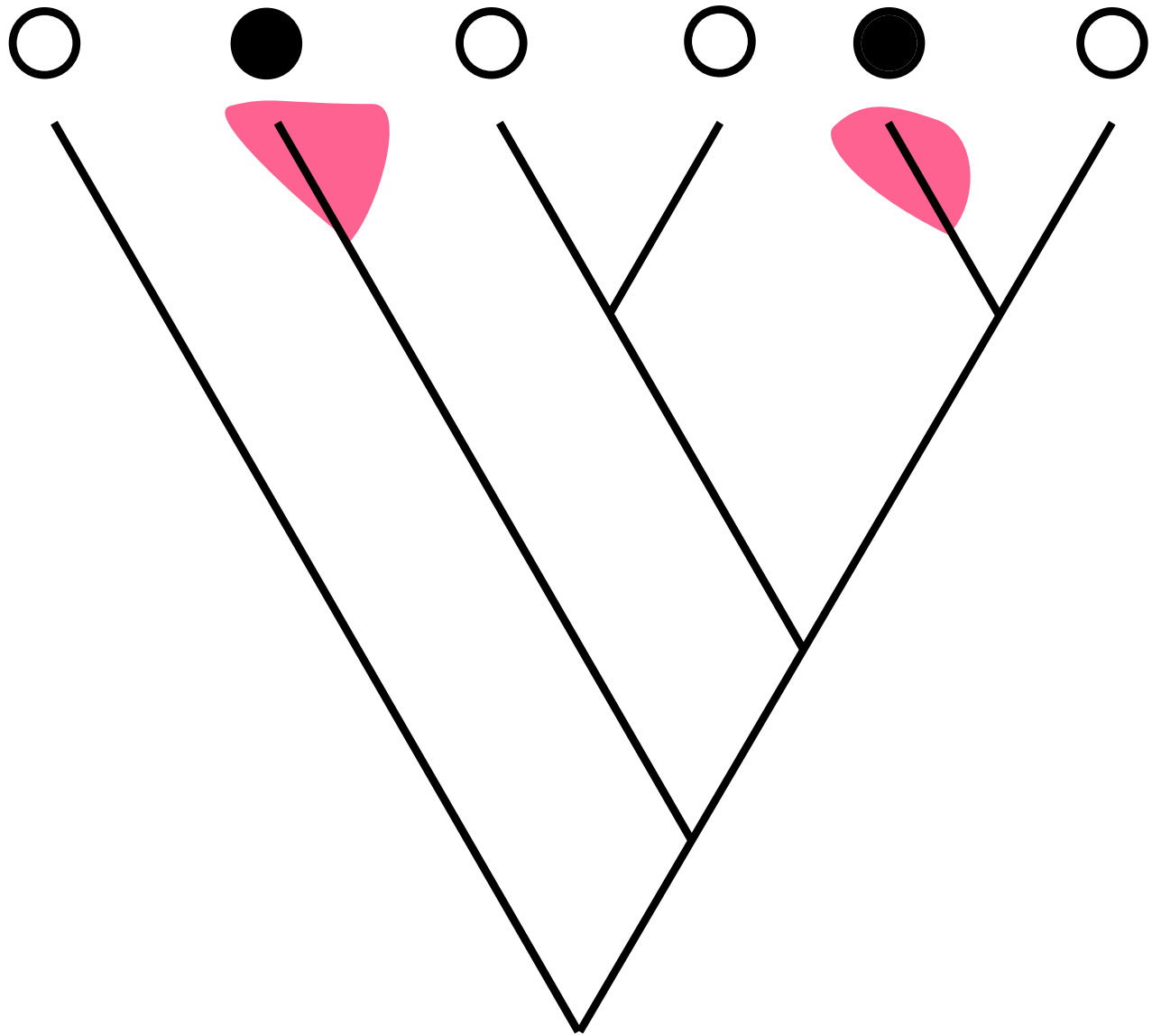
Paraphyletic groups: error of omitting some species

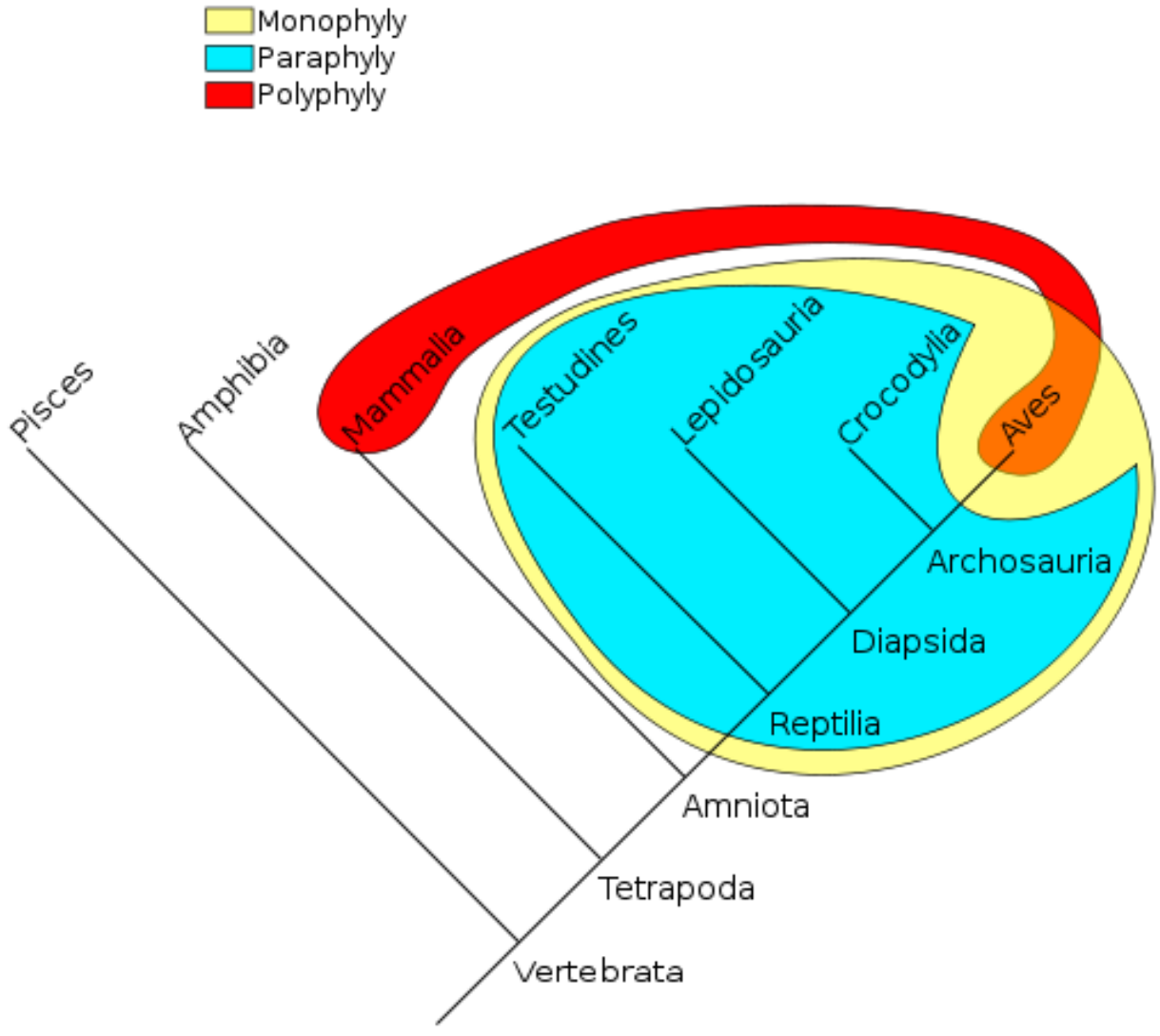




https://xkcd.com/867/

Polyphyletic groups: error of grouping "unrelated" species



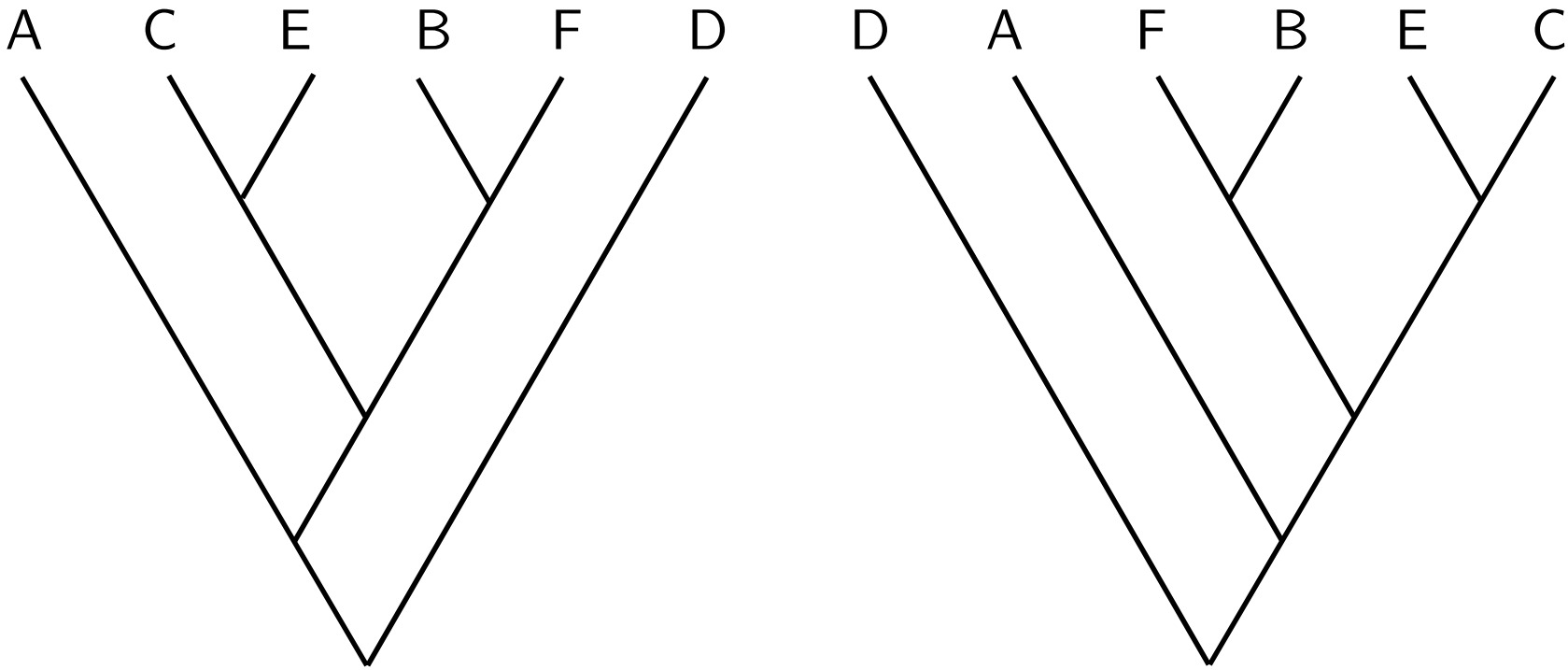


from wikipedia

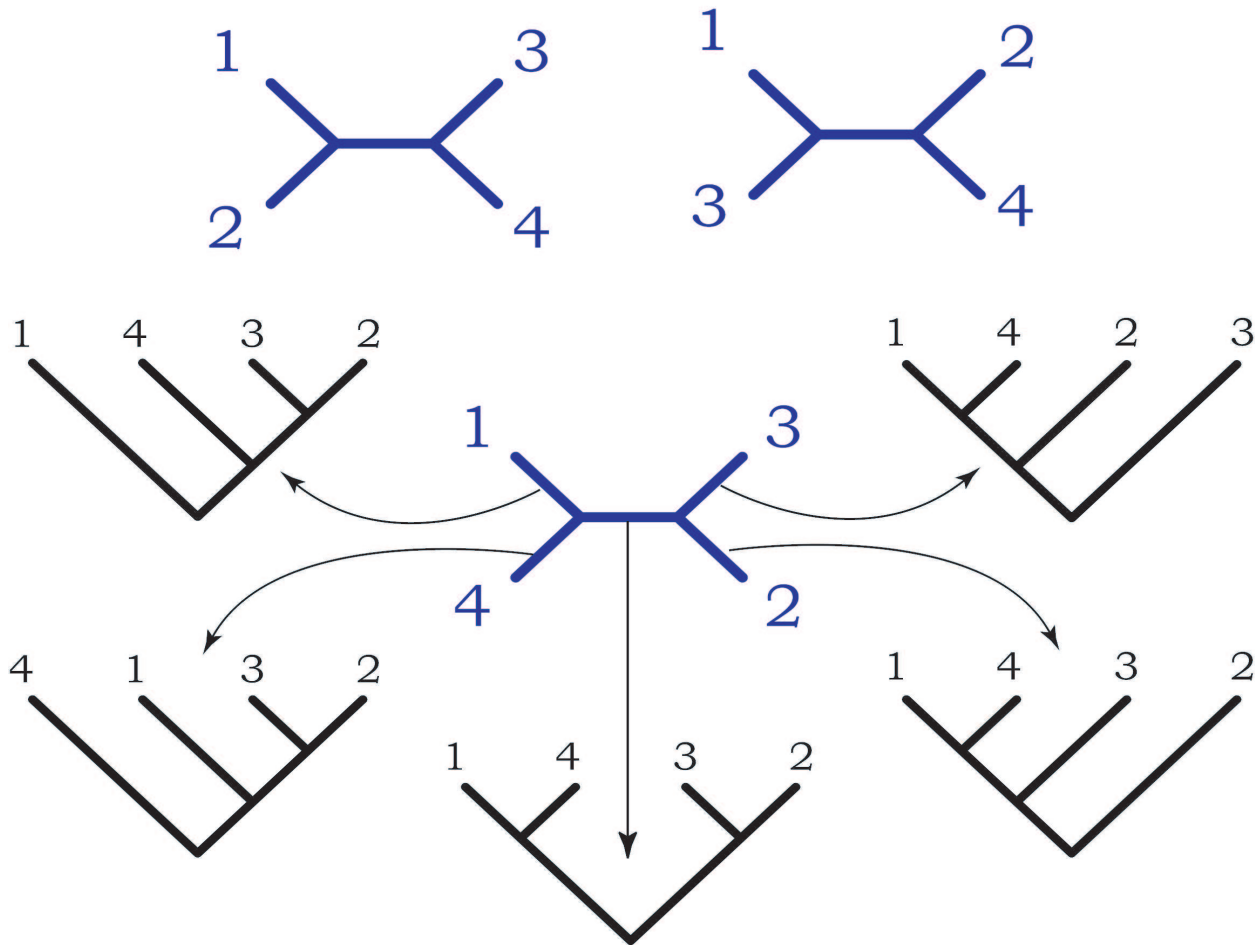
more terms:

- sister taxa: taxa or monophyletic groups which share a most recent common ancestor
- outgroup: taxon that is determined *a priori* to be sister to all other taxa in the analysis. Used for rooting tree

Branch rotation does not matter



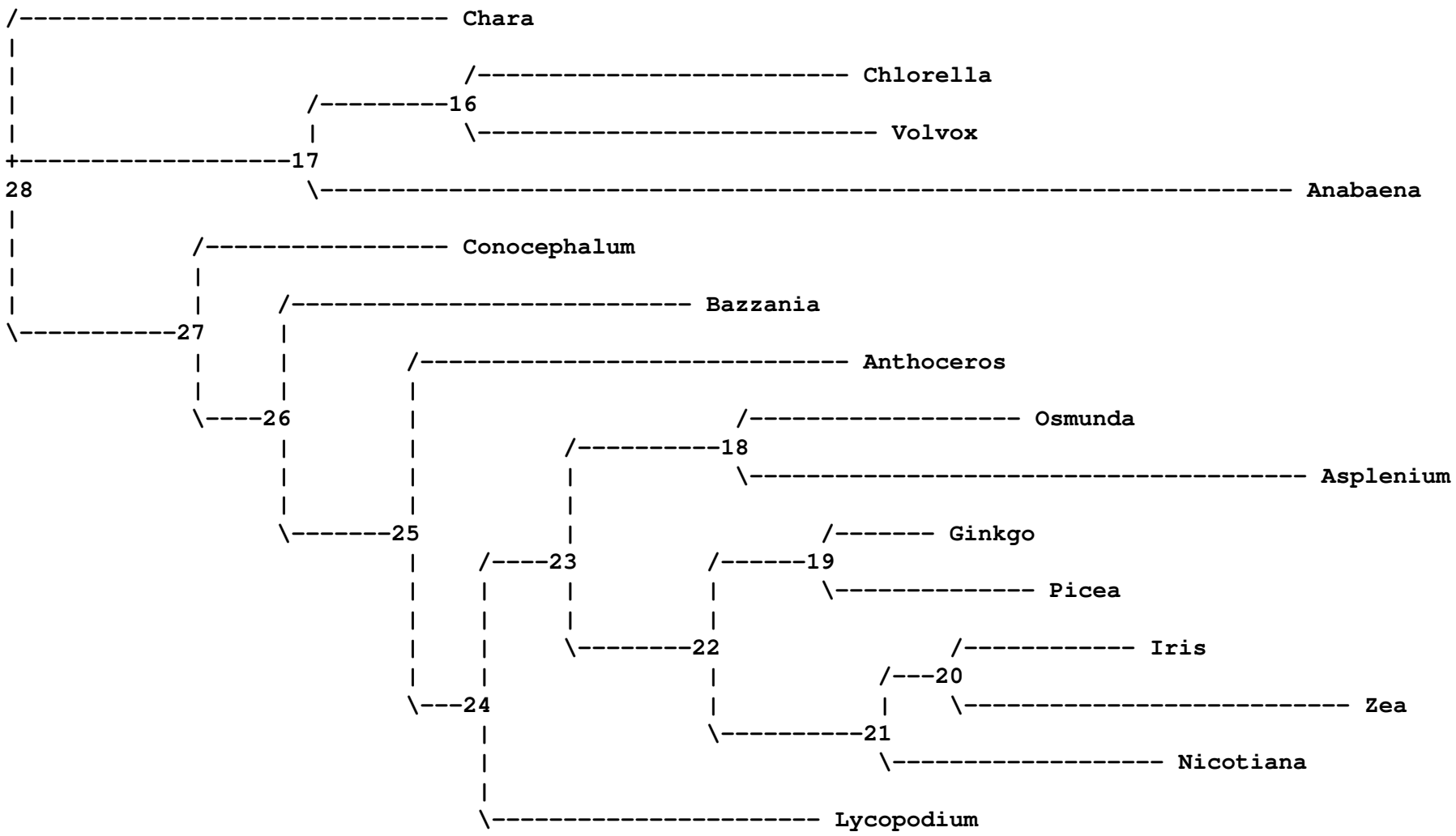
Rooted vs unrooted trees



Splits

- It is useful to think of unrooted trees in terms of 'splits'
- Each branch in an unrootred tree splits the taxa into two groups.
- Membership in those groups can be denoted by ** vs ..
- e.g. a split between 1+2 and 3+4 can be summarized as
- 1234
- ** ..

Warning: software often displays unrooted trees like this:



a brief digression into newick tree file format

Newick's Lobster House was the site of an historic 1986 meeting at which a standard was devised for storing descriptions of phylogenetic trees as strings. (Photo from Paul Lewis)

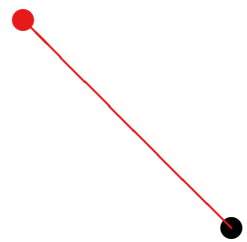
Note: ((1,2),3,4) is referred to as Newick or New Hampshire notation for the tree.

You can read it by following the rules:

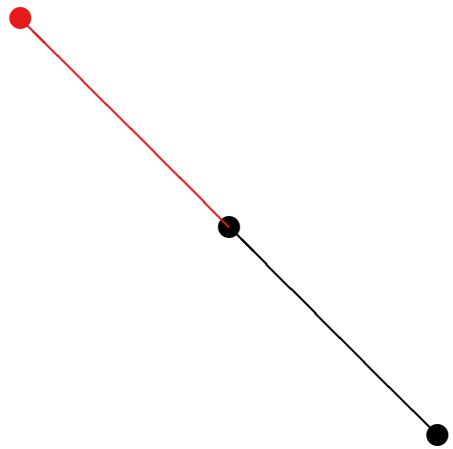
- start at a node,
- if the next symbol is '(' then add a child to the current node and move to this child,
- if the next symbol is a label, then label the node that you are at,
- if the next symbol is a comma, then move back to the current node's parent and add another child,
- if the next symbol is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

((1,2),3,4)

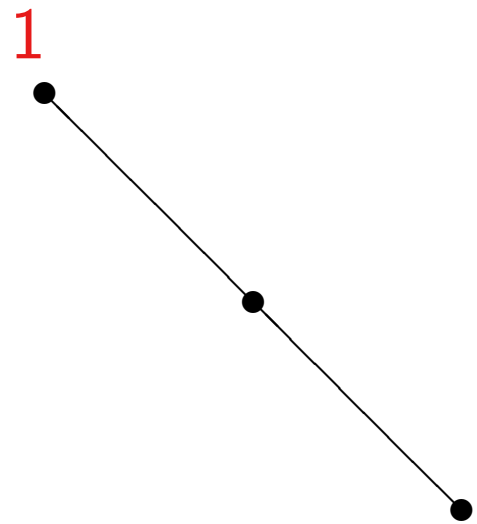
((1,2),3,4)



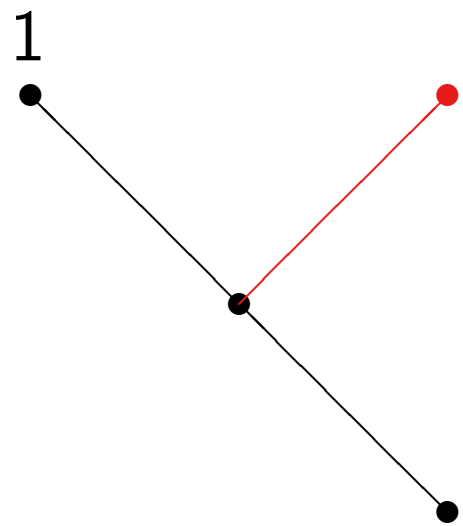
((1,2),3,4)



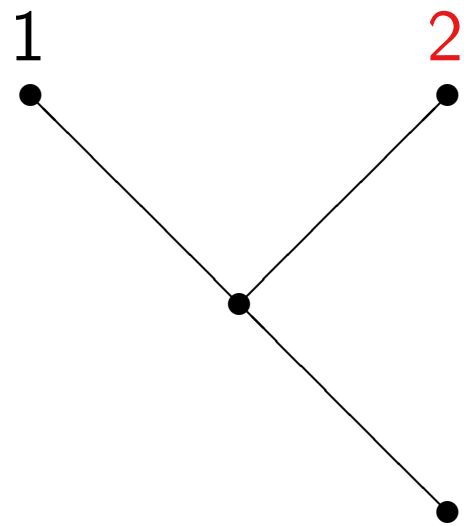
((1,2),3,4)



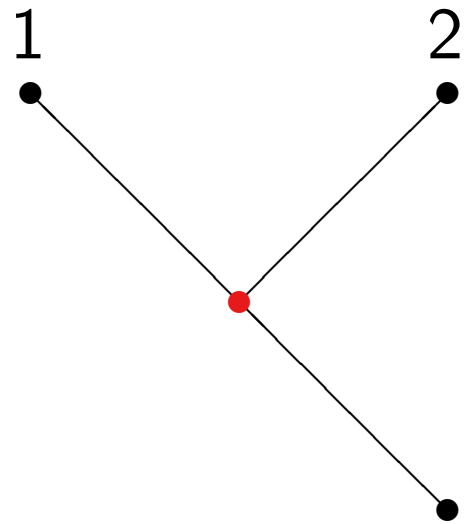
((1,2),3,4)



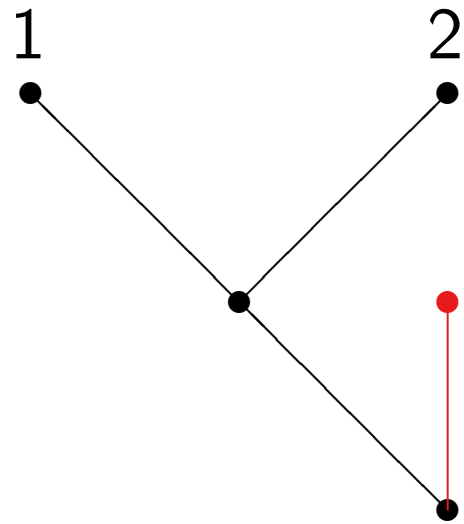
((1,2),3,4)



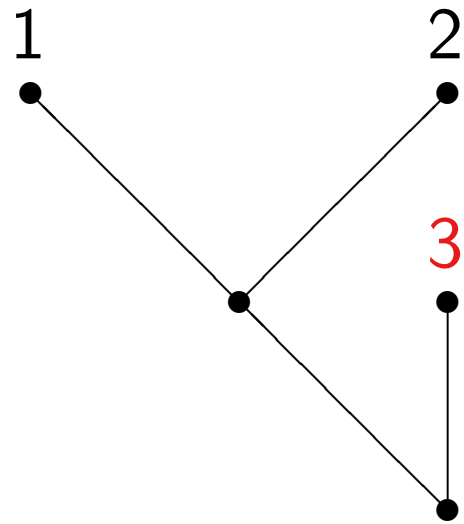
((1,2),3,4)



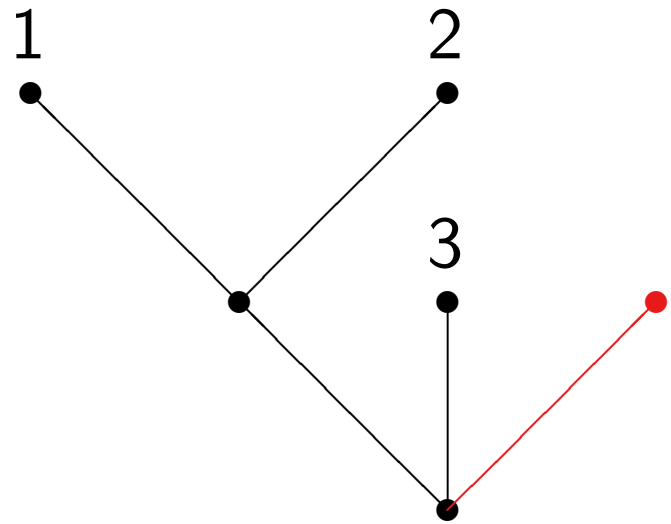
((1,2),3,4)



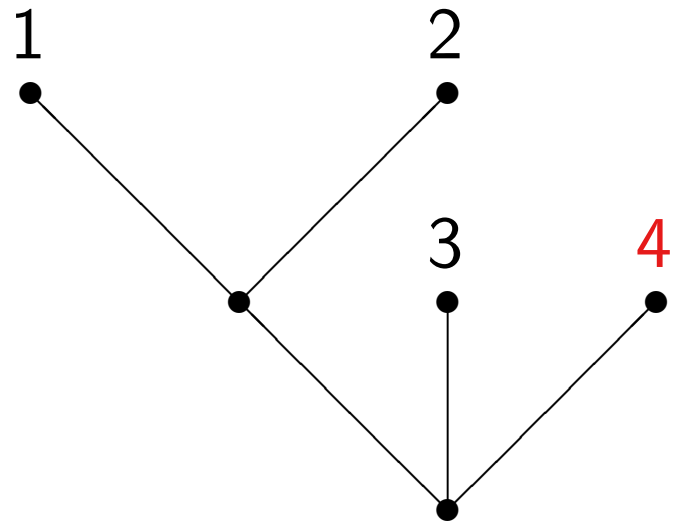
((1,2),3,4)



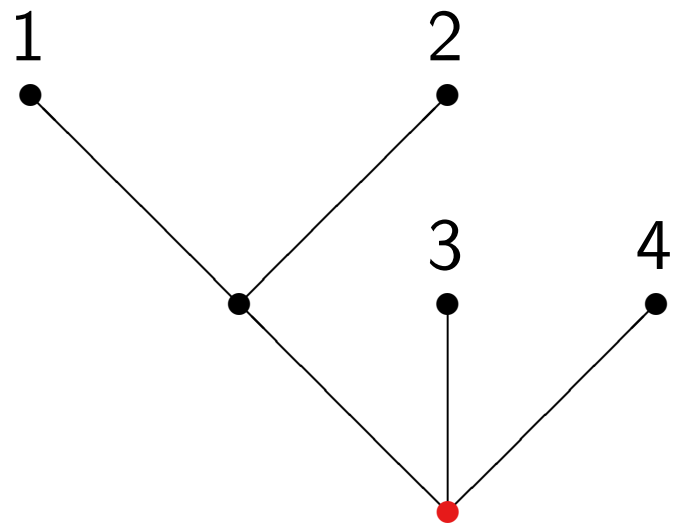
((1,2),3,4)



((1,2),3,4)



((1,2),3,4)



Newick

- Parenthetical tree format
- Rooted vs. unrooted trees are not differentiated
- Some programs interpret polytomy at root as ‘unrooted
- Branches and nodes not well differentiated
- A name can contain and characters except blanks, colons, semicolons, parentheses, and square brackets

Nexus

- Starts with #nexus
- Can contain blocks of alignments, trees, commands, and more!
- Blocks between ‘begin’ and ‘end’
- Trees in Newick format, prepended with [&U] unrooted or [&R] rooted

Nexus

```
#nexus
...
begin taxa;
  dimensions ntax=5;
  taxlabels
    Giardia
    Thermus
    Deinococcus
    Sulfolobus
    Haobacterium
  ;
end;
```

```
#nexus
...
begin data;
  dimensions ntax=5 nchar=54;
  format datatype=dna missing=? gap=-;
  matrix
    Ephedra       TTAAGCCATGCATGTCTAAGTATGAACTAATTCCAAACGGTGAAACTGCGGATG
    Gnetum        TTAAGCCATGCATGTCTATGTACGAACTAATC-AGAACGGTGAAACTGCGGATG
    Welwitschia   TTAAGCCATGCACGTGTAAGTATGAACTAGTC-GAAACGGTGAAACTGCGGATG
    Ginkgo        TTAAGCCATGCATGTGTAAGTATGAACTCTTTACAGACTGTGAAACTGCGAATG
    Pinus         TTAAGCCATGCATGTCTAAGTATGAACTAATTGCAGACTGTGAAACTGCGGATG
                 [----+--10|----+--20|----+--30|----+--40|----+--50|----]
  ;
end;
```

http://hydrodictyon.eeb.uconn.edu/eebedia/index.php/Phylogenetics:_NEXUS_Format

Nexus

```
#nexus
...
begin trees;
  translate
    1 Ephedra,
    2 Gnetum,
    3 Welwitschia,
    4 Ginkgo,
    5 Pinus
  ;
  tree one = [&U] (1,2,(3,(4,5));
  tree two = [&U] (1,3,(5,(2,4));
end;
```

```
#nexus
...
begin sets;
  charset trnL_intron = 562-4226;
  taxset gnetales = Ephedra Gnetum Welwitschia;
end;
```

http://hydrodictyon.eeb.uconn.edu/eebedia/index.php/Phylogenetics:_NEXUS_Format

NeXML

- Phylogenetic data as XML
- Can capture all information from Nexus
- Full semantic annotation
- Easily extensible

NeXML

Computer readable, but not very human readable

```
 <otu about="#otu99" id="otu99" label="Parupeneus barberinoides">
  <meta datatype="xsd:string" property="ot:originalLabel" xsi:type="nex:LiteralMeta">Parupeneus
  <meta datatype="xsd:int" property="ot:ottId" xsi:type="nex:LiteralMeta">758968</meta>
  <meta datatype="xsd:string" property="ot:ottTaxonName" xsi:type="nex:LiteralMeta">Parupeneus b
 </otu>
</otus>
<trees about="#trees1" id="trees1" otus="otus1">
 <tree about="#tree1" id="tree1" label="Untitled (tree1)" xsi:type="nex:FloatTree">
  <meta datatype="xsd:string" property="ot:branchLengthDescription" xsi:type="nex:LiteralMeta"/>
  <meta datatype="xsd:string" property="ot:branchLengthMode" xsi:type="nex:LiteralMeta">ot:undef
  <meta datatype="xsd:string" property="ot:curatedType" xsi:type="nex:LiteralMeta">Bayesian infe
  <meta datatype="xsd:string" property="ot:inGroupClade" xsi:type="nex:LiteralMeta">node2</meta>
  <meta datatype="xsd:string" property="ot:nodeLabelMode" xsi:type="nex:LiteralMeta"/>
  <meta datatype="xsd:string" property="ot:nodeLabelTimeUnit" xsi:type="nex:LiteralMeta"/>
  <meta datatype="xsd:string" property="ot:outGroupEdge" xsi:type="nex:LiteralMeta"/>
  <meta datatype="xsd:string" property="ot:specifiedRoot" xsi:type="nex:LiteralMeta">node1</meta
  <meta datatype="xsd:boolean" property="ot:unrootedTree" xsi:type="nex:LiteralMeta">false</meta
  <node about="#node1" id="node1" root="true"/>
  <node about="#node2" id="node2"/>
  <node about="#node144" id="node144"/>
  <node about="#node145" id="node145"/>
  <node about="#node146" id="node146"/>
  <node about="#node147" id="node147"/>
  <node about="#node148" id="node148"/>
  <node about="#node149" id="node149"/>
  <node about="#node150" id="node150"/>
  <node about="#node151" id="node151"/>
  <node about="#node152" id="node152"/>
  <node about="#node153" id="node153"/>
  <node about="#node154" id="node154"/>
  <node about="#node155" id="node155" otu="otu72">
   <meta datatype="xsd:boolean" property="ot:isLeaf" xsi:type="nex:LiteralMeta">true</meta>
  </node>
  <node about="#node156" id="node156" otu="otu73">
   <meta datatype="xsd:boolean" property="ot:isLeaf" xsi:type="nex:LiteralMeta">true</meta>
  </node>
  <node about="#node157" id="node157" otu="otu74">
   <meta datatype="xsd:boolean" property="ot:isLeaf" xsi:type="nex:LiteralMeta">true</meta>
  </node>
  <node about="#node158" id="node158"/>
  <node about="#node159" id="node159" otu="otu75">
   <meta datatype="xsd:boolean" property="ot:isLeaf" xsi:type="nex:LiteralMeta">true</meta>
  </node>
```

Phylip (sequence data format)

- First line must be two integers: <number of taxa> <number of sites>
- Sequence ID followed by spaces up to 10 char.
- No duplicate names
- Relaxed phylip up to 250 characters followed by a space

```
      5    42
Turkey     AAGCTNGGGC ATTTCAGGGT GAGCCCGGGC AATACAGGGT AT
Salmo gairAAGCCTTGGC AGTGCAGGGT GAGCCGTGGC CGGGCACGGT AT
H. SapiensACCGGTTGGC CGTTCAGGGT ACAGGTTGGC CGTTCAGGGT AA
Chimp      AAACCCTTGC CGTTACGCTT AAACCGAGGC CGGGACACTC AT
Gorilla    AAACCCTTGC CGGTACGCTT AAACCATTGC CGGTACGCTT AA
```

Phylip interleaved

```
  5    42
Turkey    AAGCTNGGGC ATTTCAGGGT
Salmo gairAAGCCTTGGC AGTGCAGGGT
H. SapiensACCGGTTGGC CGTTCAGGGT
Chimp     AAACCCTTGC CGTTACGCTT
Gorilla   AAACCCTTGC CGGTACGCTT

GAGCCCGGGC AATACAGGGT AT
GAGCCGTGGC CGGGCACGGT AT
ACAGGTTGGC CGTTCAGGGT AA
AAACCGAGGC CGGGACACTC AT
AAACCATTGC CGGTACGCTT AA
```

Phylip sequential

```
  5    42
Turkey    AAGCTNGGGC ATTTCAGGGT
GAGCCCGGGC AATACAGGGT AT
Salmo gairAAGCCTTGGC AGTGCAGGGT
GAGCCGTGGC CGGGCACGGT AT
H. SapiensACCGGTTGGC CGTTCAGGGT
ACAGGTTGGC CGTTCAGGGT AA
Chimp     AAACCCTTGC CGTTACGCTT
AAACCGAGGC CGGGACACTC AT
Gorilla   AAACCCTTGC CGGTACGCTT
AAACCATTGC CGGTACGCTT AA
```

Fasta (sequence data format)

- Description line before each sequence starts with (">") symbol in the first column

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.|len=368
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

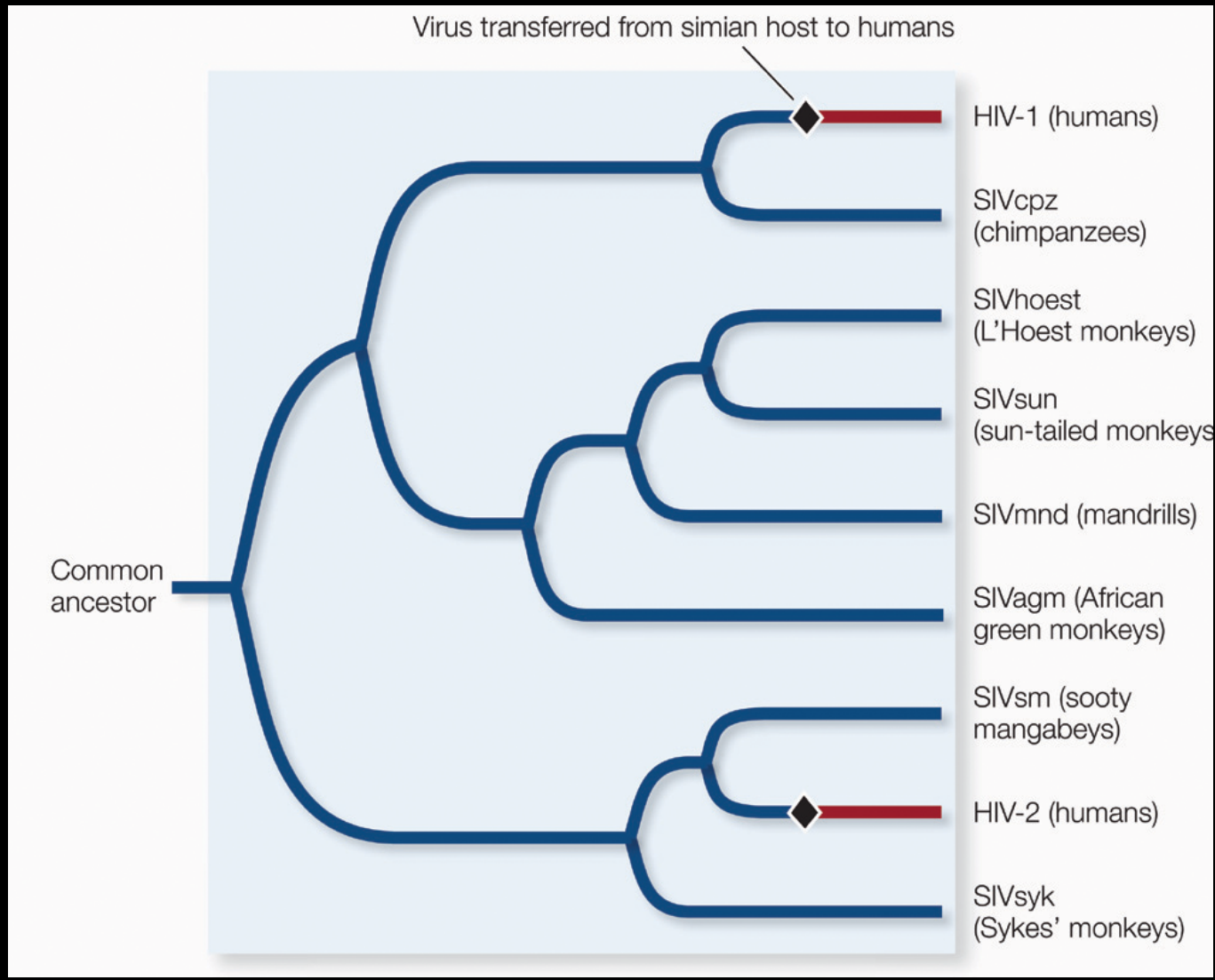
DIY

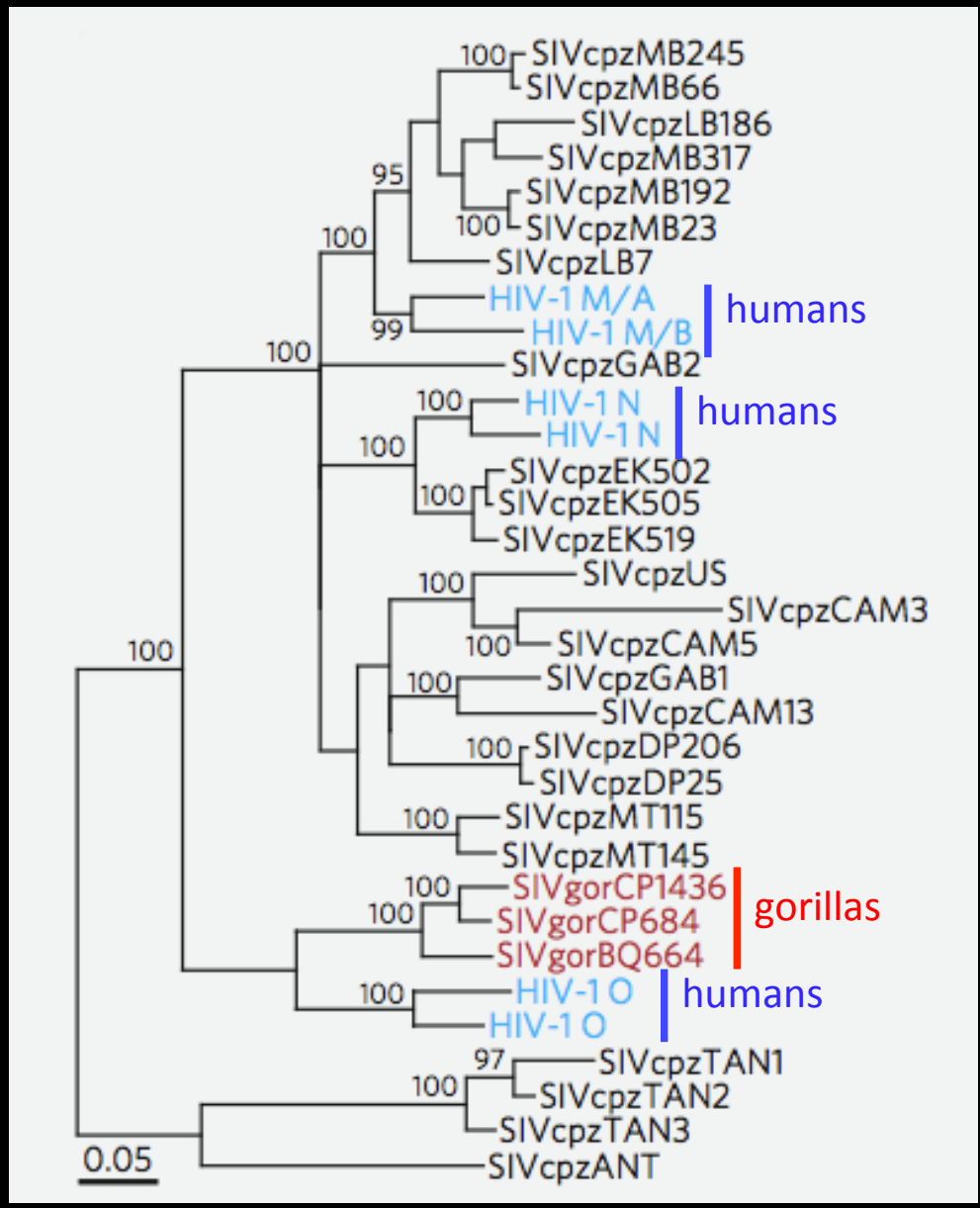
Create a newick tree file in your text editor with the content:
`(((C,(D,E)),(F,G),A),B);`
Save it as 'example.tre'.

- Draw the tree by hand
- Write down all the splits in ..** format.
- Load the tree in a tree viewer (e.g. phylo.io). Re-root the tree. What rootings make the following true? Which cannot be true?
  - A is more closely related to G than it is to C
  - (C,D,E) is sister to (A,B,F,G)
  - (C,D) is sister to (A,B,E,F,G)
  - (C,D,E) is a paraphyletic group
  - (C,D,E) is a monophyletic group
  - (A,B,C) is a monophyletic group

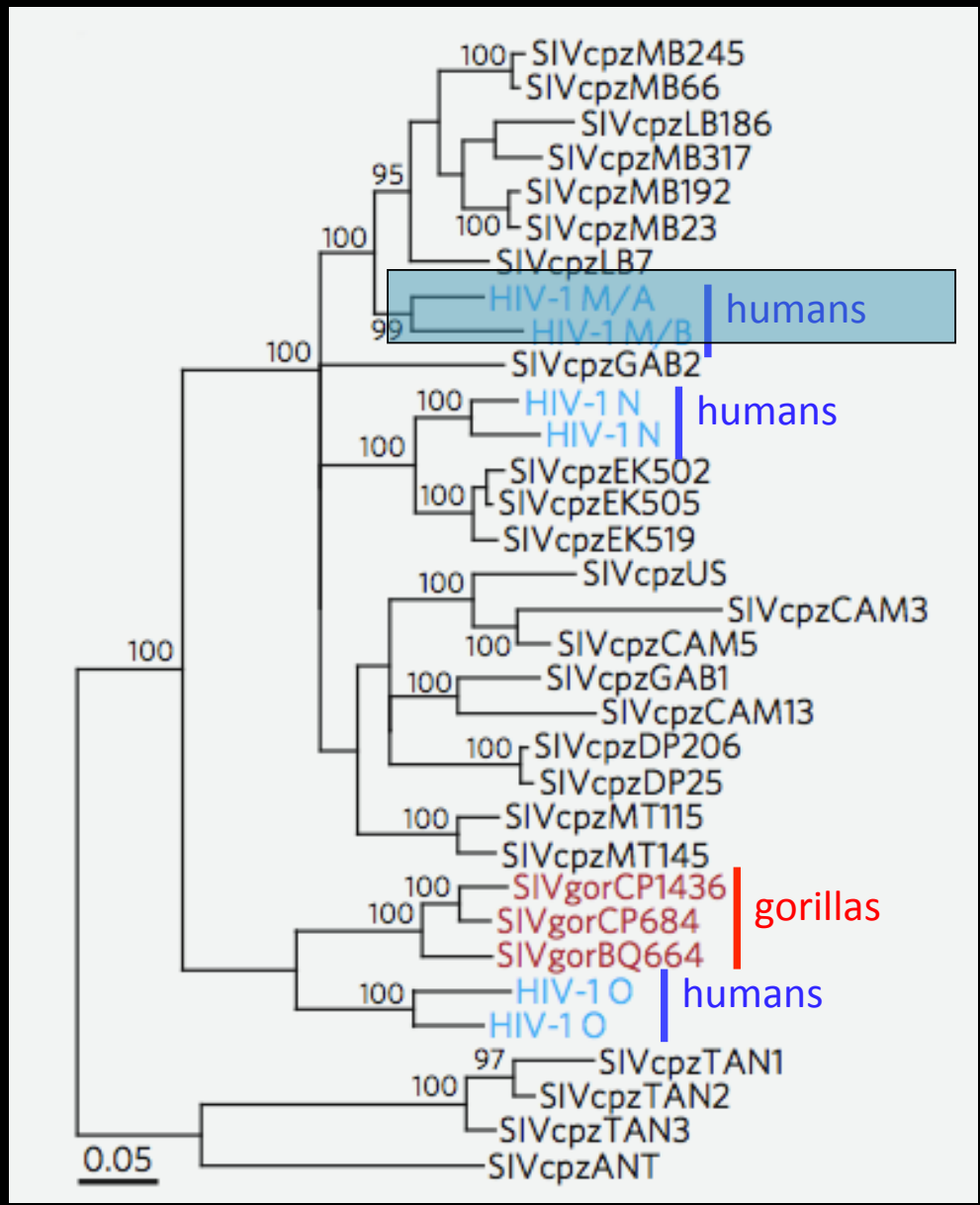
Origins of Emerging Diseases

- 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?

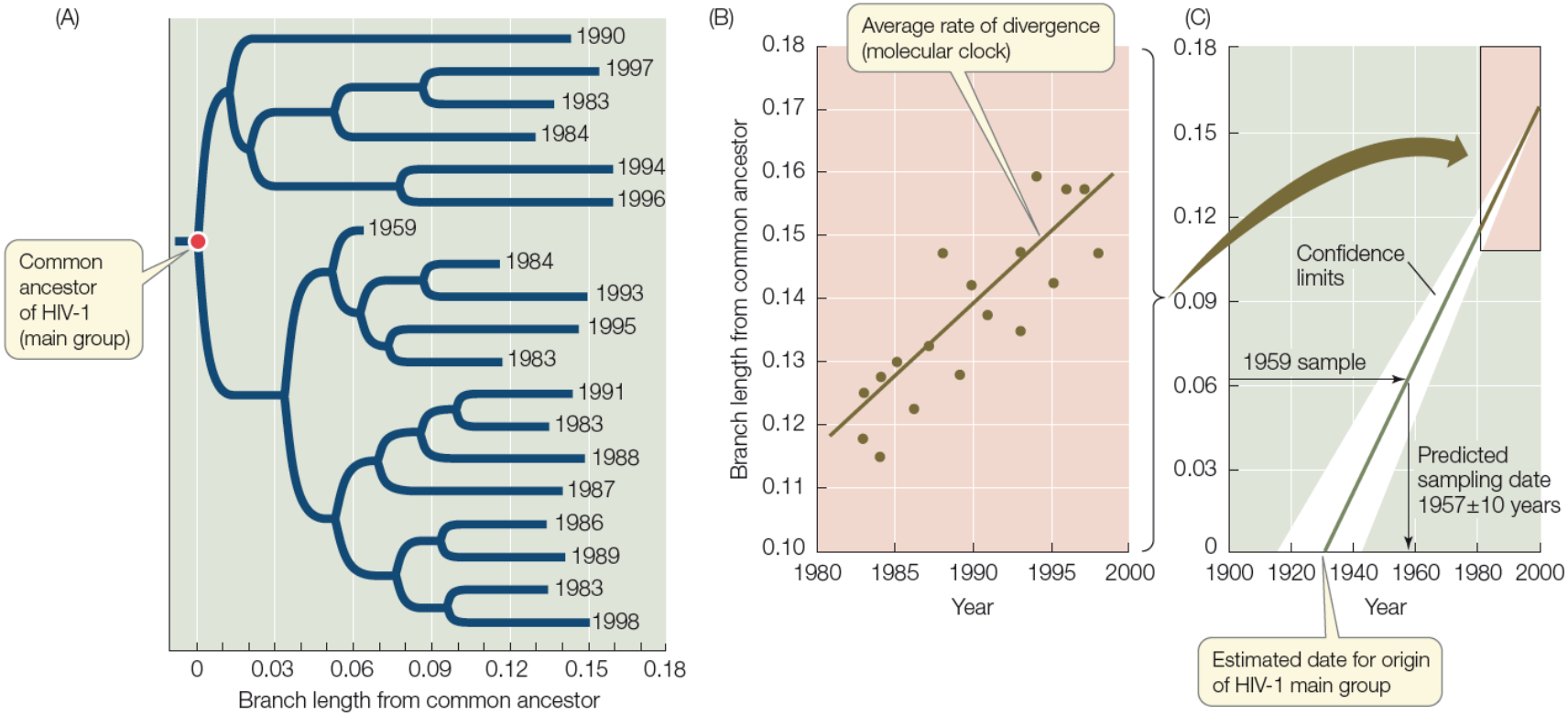

Virus transferred from simian host to humans
HIV-1 (humans)
SIVcpz
(chimpanzees)
SIVhoest
(L'Hoest monkeys)
SIVsun
(sun-tailed monkeys
SIVmnd (mandrills)
Common
ancestor
SIVagm (African
green monkeys)
SIVsm (sooty
mangabeys)
HIV-2 (humans)
SIVsyk
(Sykes' monkeys)

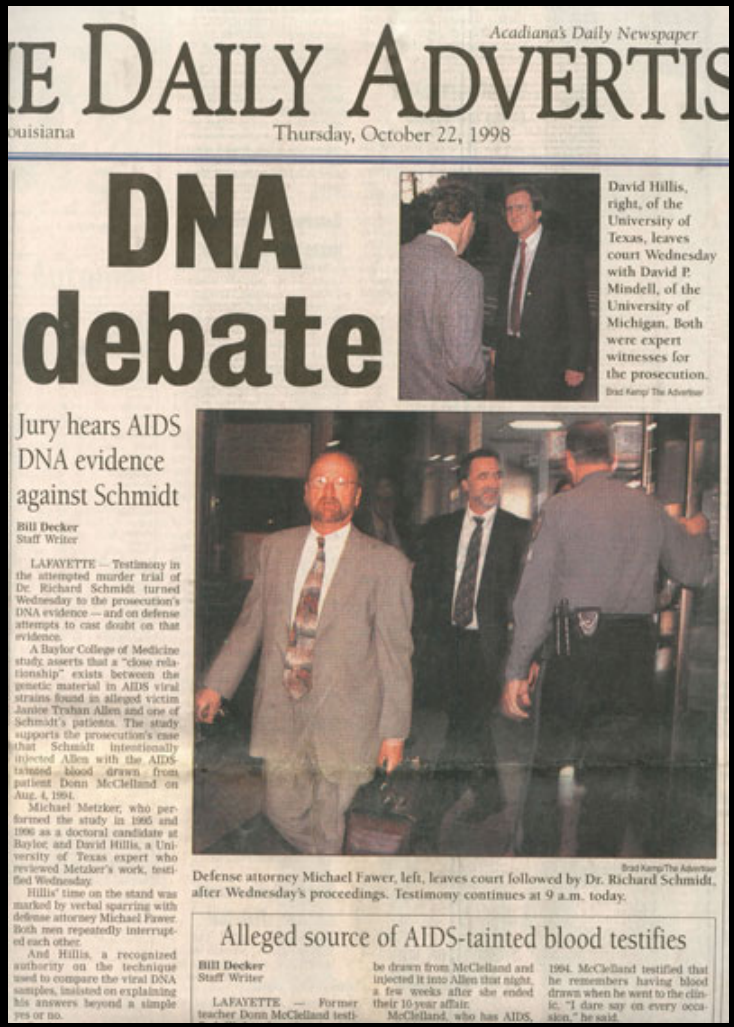


Van Heuverswyn et al., *Nature* 444: 164 (2006)



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


(A)
1990
1997
1983
1984
1994
1996
1959
1984
1993
1995
1983
1991
1983
1988
1987
1986
1989
1983
1998
Common ancestor of HIV-1 (main group)
0
0.03
0.06
0.09
0.12
0.15
0.18
Branch length from common ancestor
(B)
Average rate of divergence (molecular clock)
Branch length from common ancestor
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
1980
1985
1990
1995
2000
Year
(C)
0.18
0.15
0.12
0.09
0.06
0.03
0
Confidence limits
1959 sample
Predicted sampling date 1957±10 years
1900
1920
1940
1960
1980
2000
Year
Estimated date for origin of HIV-1 main group



E DAILY ADVERTIS
uisiana
Acadiana's Daily Newspaper
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

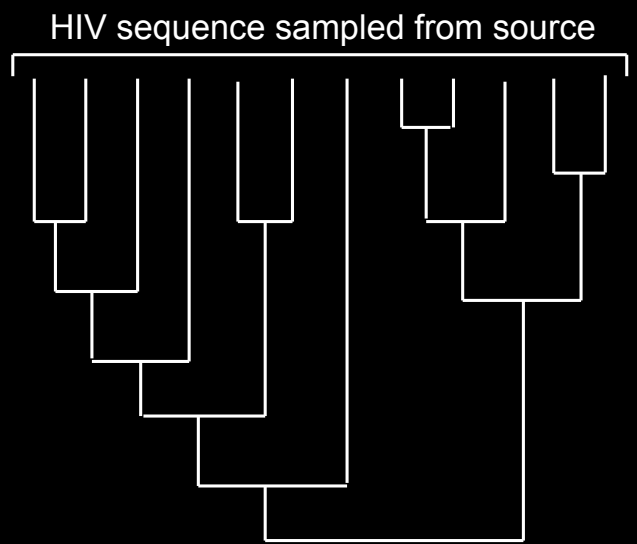
- **Origins of HIV, SARS and other viruses transmitted between animals and humans**
- **Global virus diversity for vaccine trials**
- **Epidemiological studies**
- **Identification of new diseases**
- **Forensic uses**

HIV transmission

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

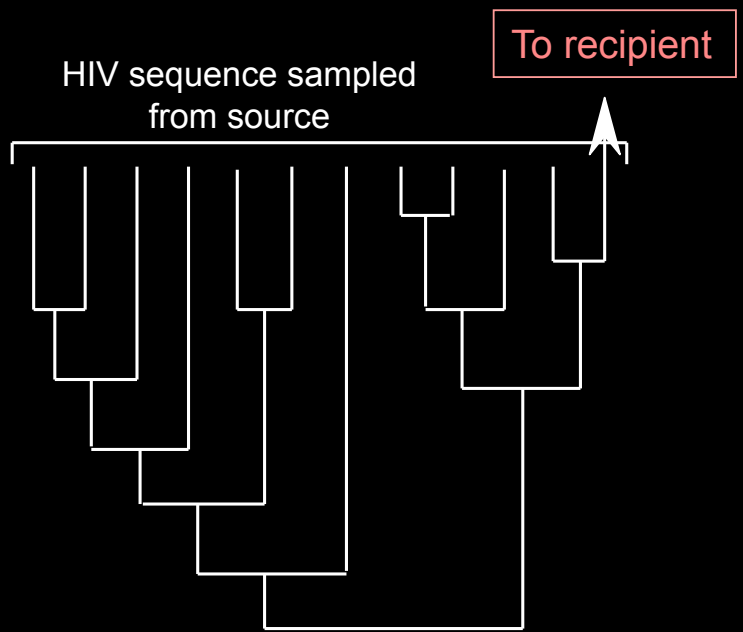
To imagine how this is possible, start by considering the diversity of HIV within one infected individual:

Time 1: Prior to Transmission event

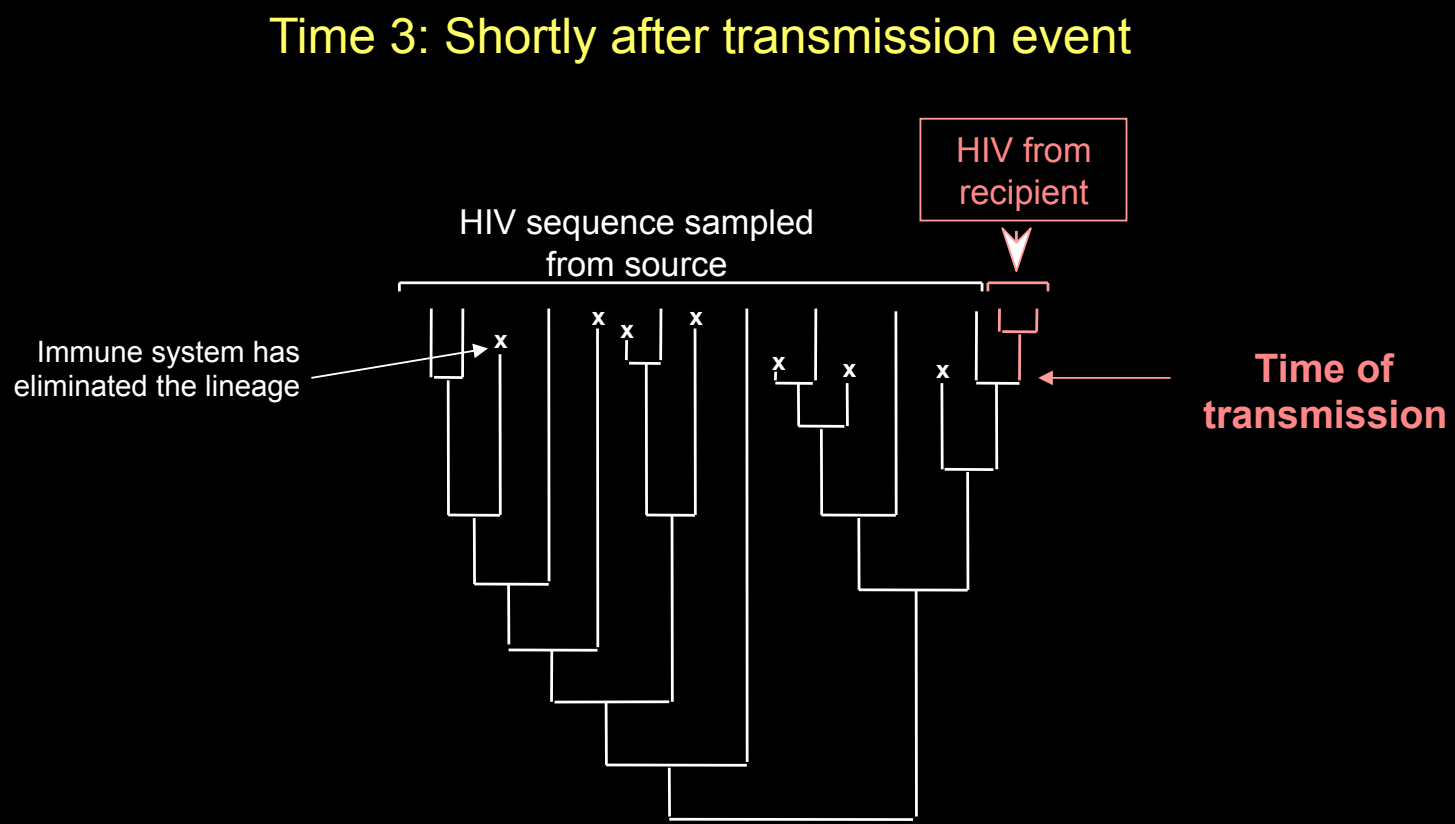


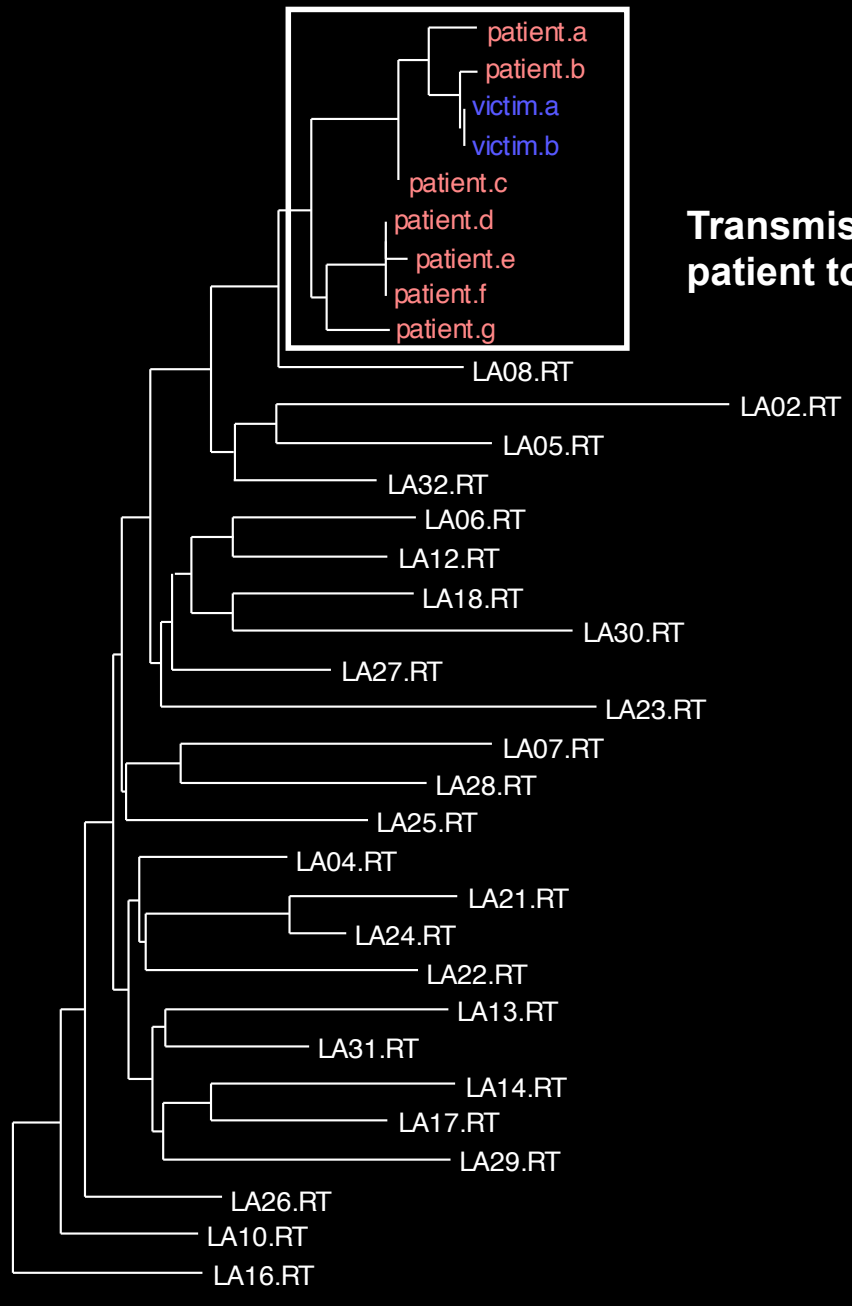
At the transmission event, the HIV in the recipient represents a small subset of the HIV present in the source:

Time 2: The transmission event

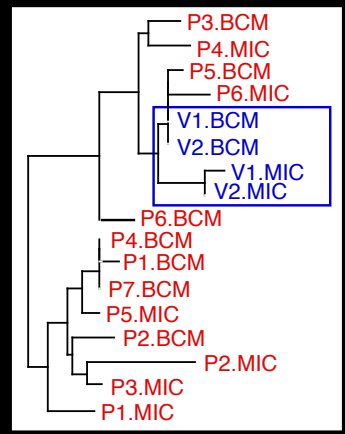


As time passes, HIV lineages in the source and recipient diversify, and other lineages become extinct.





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, 38, was accused of intentionally injecting the AIDS virus into nurse Janice Trahan Allen of Lafayette in 1996, 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.

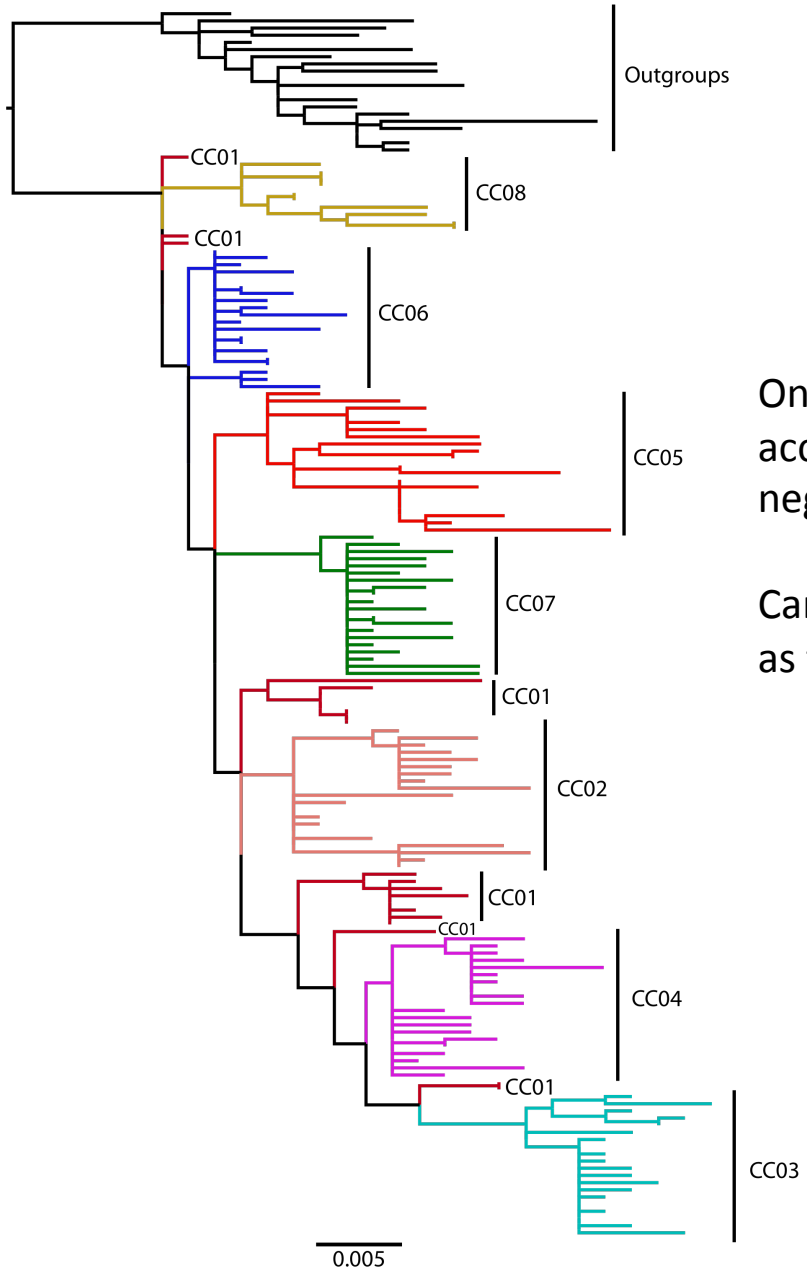
The nine-woman, three-man jury voted 10-2 to convict Schmidt after about four hours of deliberation. The jury received the formal

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

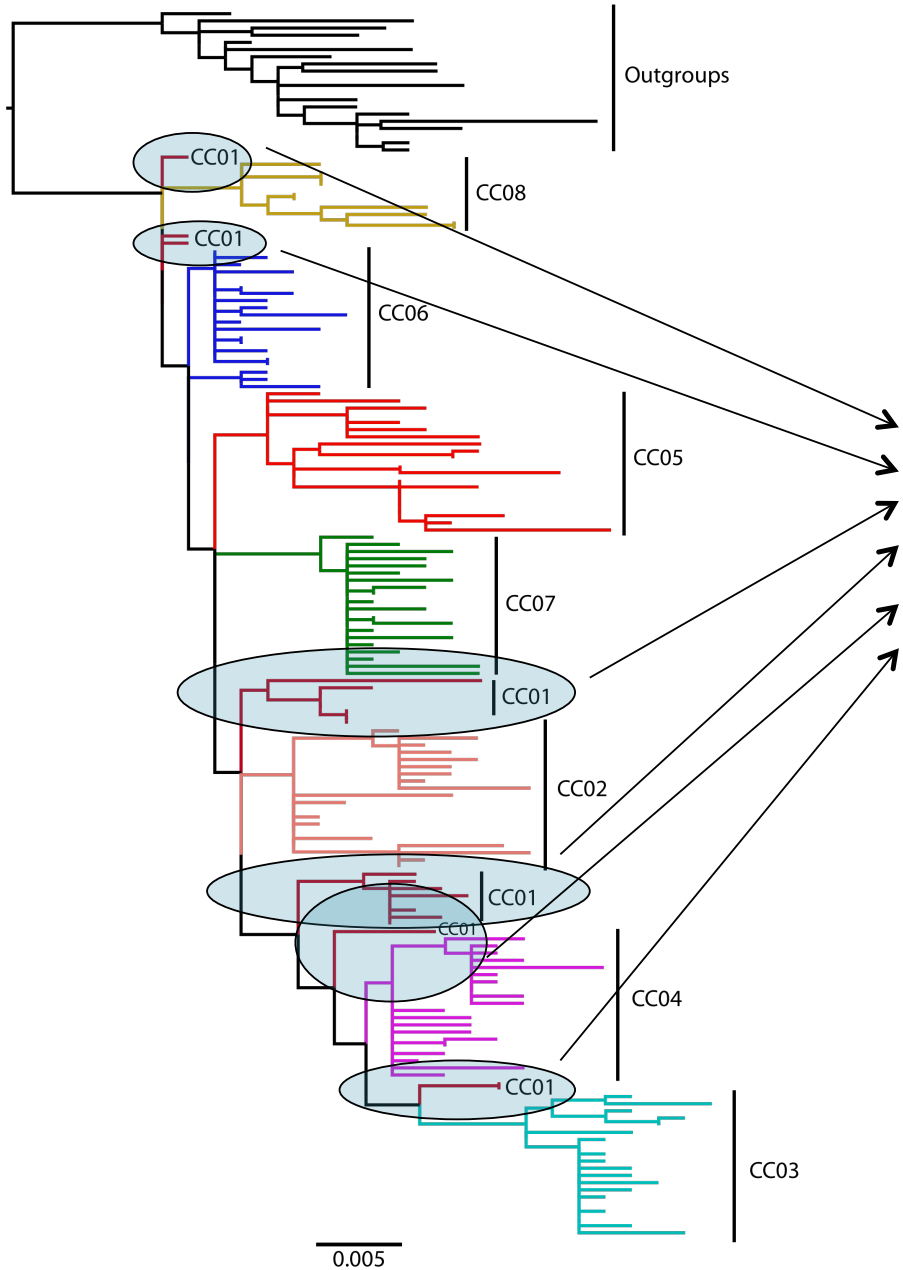
Laura Ayo
Staff Writer

- 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.

Changing the rooting of this phylogeny would change inferences!

Phylogenies can reveal suprising patterns

Phylogenies can reveal suprising patterns

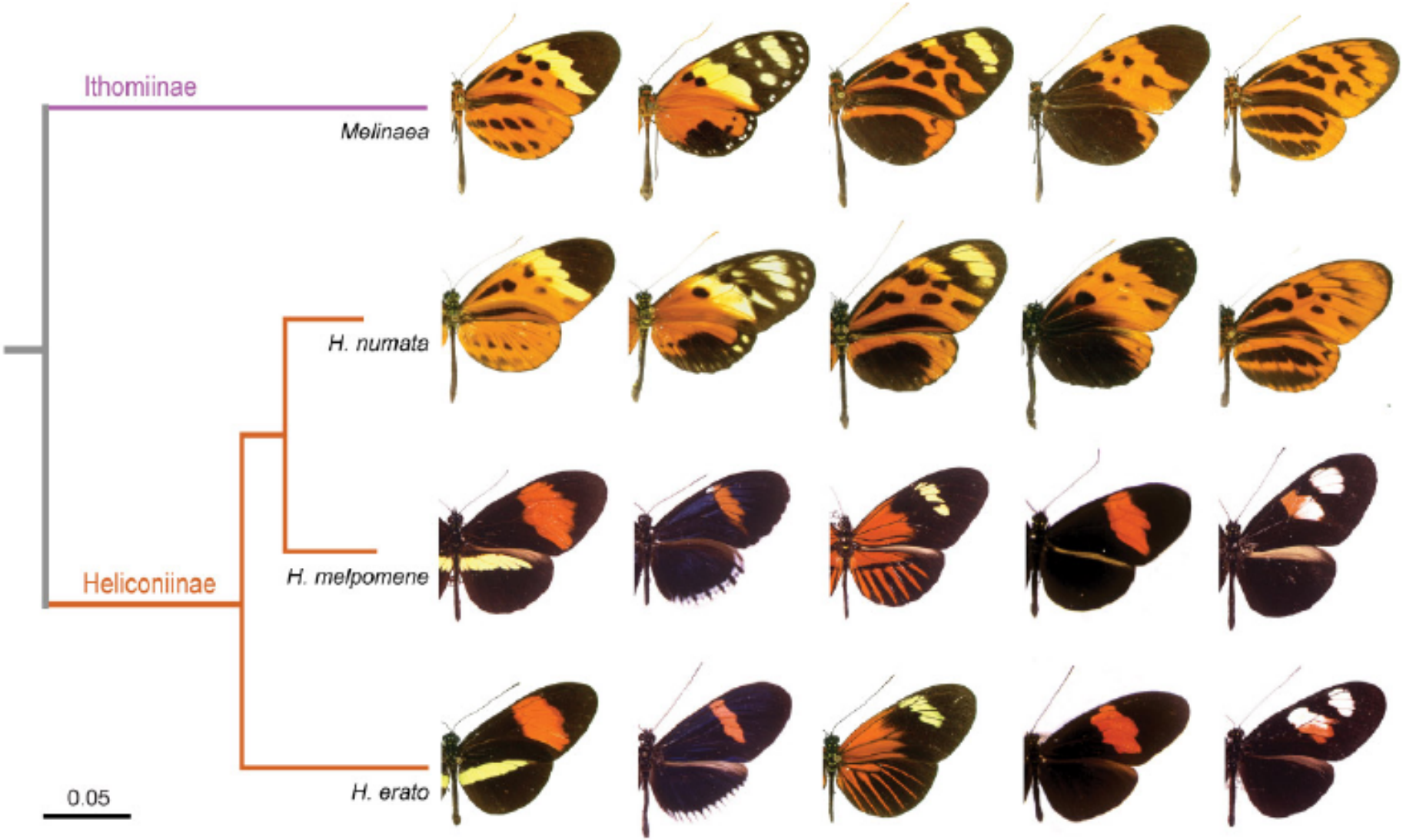


Figure by Mathieu Joron: http://xyala.cap.ed.ac.uk/joron/

What evolutionary processes can drive these patterns?

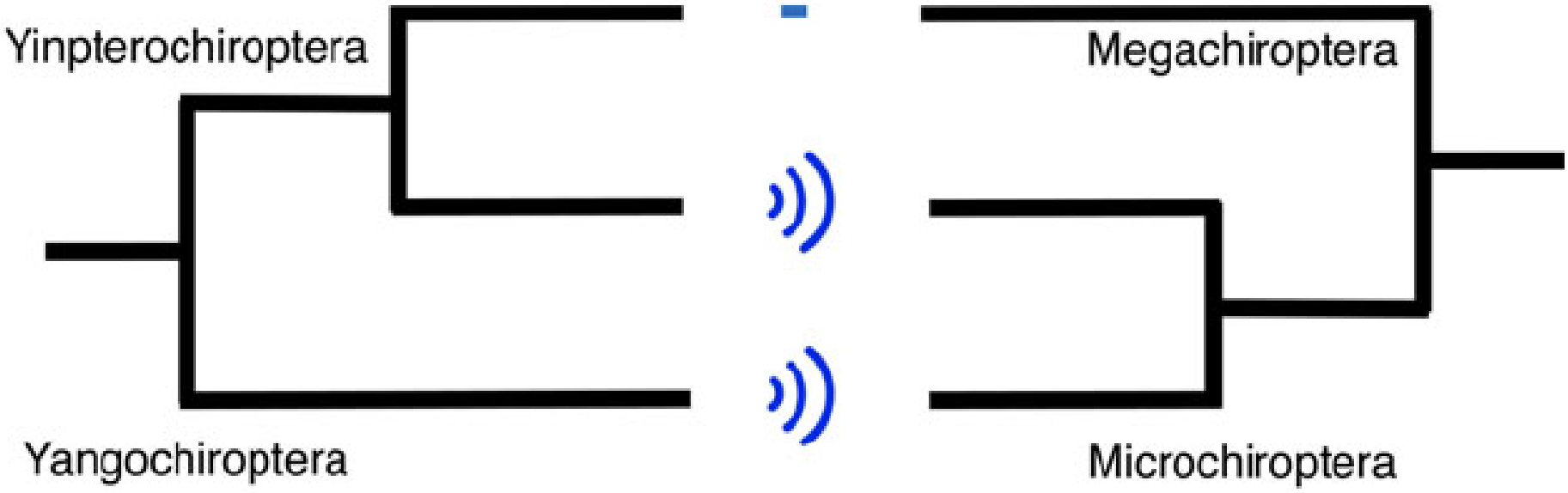
- Convergence
- Horizontal gene transfer
- Incomplete lineage sorting
- ?

What evolutionary processes can drive these patterns?

- Convergence
- Horizontal gene transfer
- Incomplete lineage sorting
- ?

We will discuss how to recognize and (try to) differentiate these processes.

Different data can drive different conclusions



Species relationships between echolocating and nonecholocating bats (after Teeling 2009). Left: inferences from DNA sequence data.
Right: traditional species relationships inferred from morphological characters (and limited sequence data). (Hahn and Nakhleh, 2016)

Estimating a tree from character data

Tree construction:

- strictly algorithmic approaches - use a "recipe" to construct a tree
- optimality based approaches - choose a way to "score" a trees and then search for the tree that has the best score.

Expressing support for aspects of the tree:

- bootstrapping,
- testing competing trees against each other,
- posterior probabilities (in Bayesian approaches).

Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17.

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., Whibley, A., Becuwe, M., Baxter, S. W., Ferguson, L., Wilkinson, P. A., Salazar, C., Davidson, C., Clark, R., Quail, M. A., Beasley, H., Glithero, R., Lloyd, C., Sims, S., Jones, M. C., Rogers, J., Jiggins, C. D., and ffrench Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, 477(7363):203–206.