Phylogenetics – a primer

tobias.warnecke@csc.mrc.ac.uk

What this primer can and can’t do

Alice: "Would you tell me, please, which way I ought to go from here?"

Cat: "That depends a good deal on where you want to get to,”

Alice: "I don't much care where –”

Cat: "Then it doesn't matter which way you go”

Alice: "– so long as I get *somewhere*,”

Cat: "Oh, you're sure to do that, if you only walk long enough."

How do you get to where you want to be?

What this primer can and can't do

"No wise fish would go
anywhere without a porpoise."

Bioinformatics is not a good subject for passive learning.

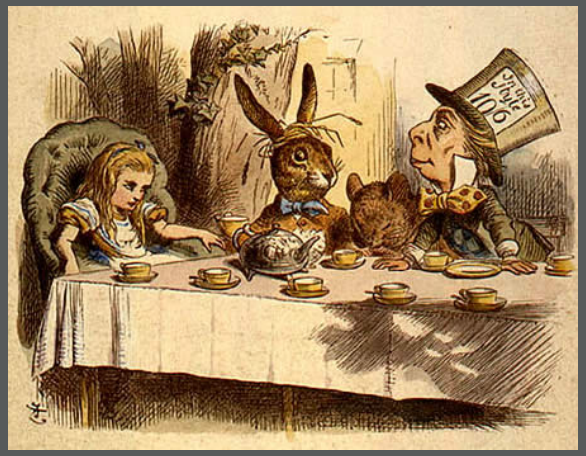
- Learn some basic scripting. Go solve your own problems.
- If you get stuck (badly): Google, Biostars, SeqAnswers, StackOverflow
- If you get stuck (really badly): ask somebody

What's happening?

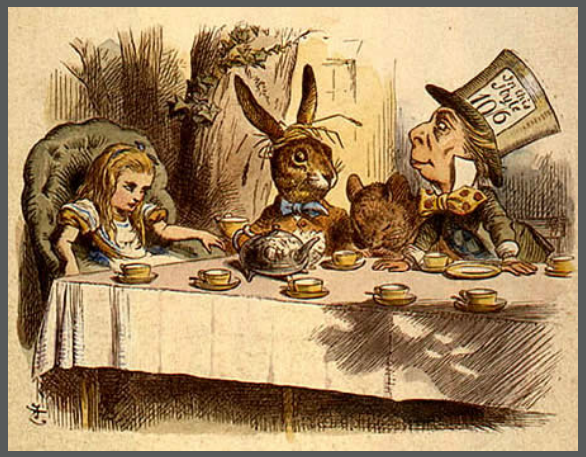
1. Reading

1.5 What's the point?

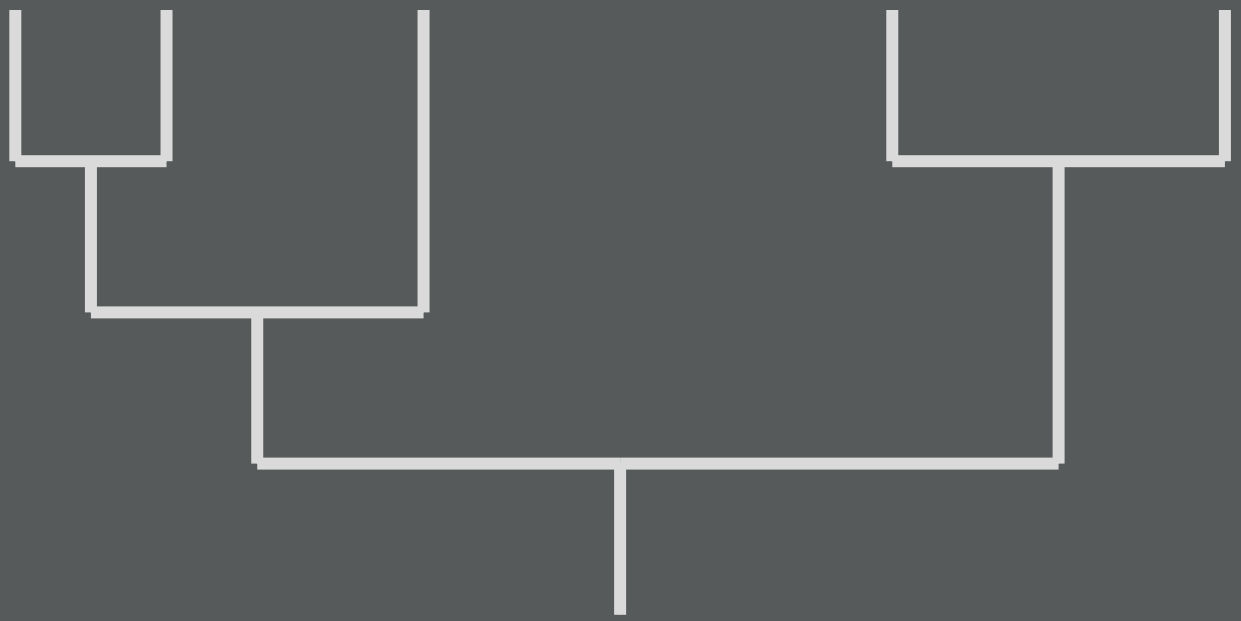
2. Writing



?



A tree



evolutionary similarity

A tree is a representation of relationships

Testing the “Oasis hypothesis (OH)”

OH: “Our music is totally different from Blur!”

1. Go to Youtube and copy links of Oasis song, Blur songs, and some others.

- She’s electric (Oasis)
- Country House (Blur)
- Sunny afternoon (The Kinks)
- My favourite things (Julie Andrews)
- Who let the dogs out? (Baha Men)

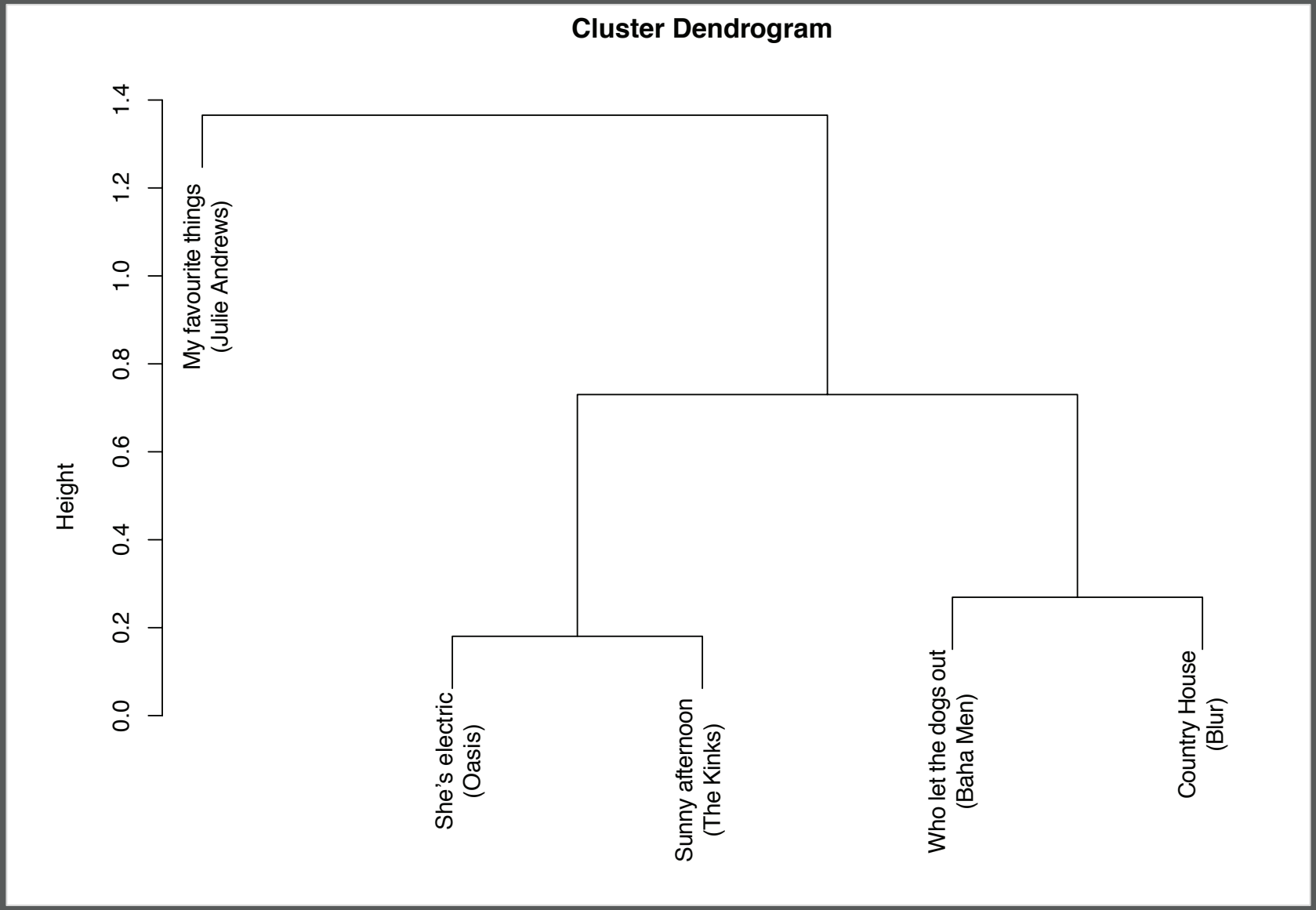
2. Convert to .mp3 (http://www.youtube-mp3.org/)
3. Convert to .wav (http://media.io)
4. Cut out random 10s fragment (Audacity)

Testing the “Oasis hypothesis (OH)”

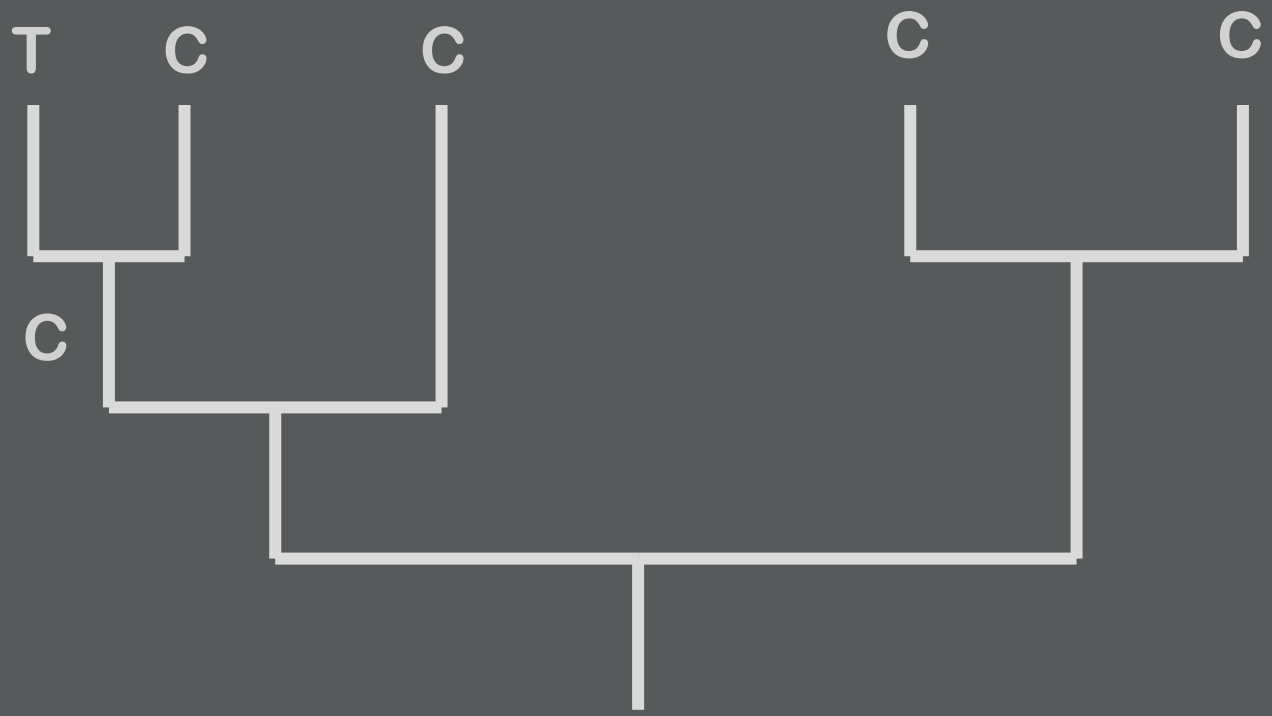
- Use R packages “tuneR” and “seewave” for amplitude modulation analysis
- A recipe can be found here:
  http://www.vesnam.com/Rblog/sortmymusic/
- You end up with a distance (similarity) matrix

```
> d
                  dogs_short.wav electric_short.wav house_short.wav sunny_short.wav
electric_short.wav      0.5813271
house_short.wav         0.2692098          0.7301821
sunny_short.wav         0.4994576          0.1803900       0.6509496
things_short.wav        1.2696424          0.9448022       1.3655203       1.0213808
```

Testing the “Oasis hypothesis (OH)”



A tree

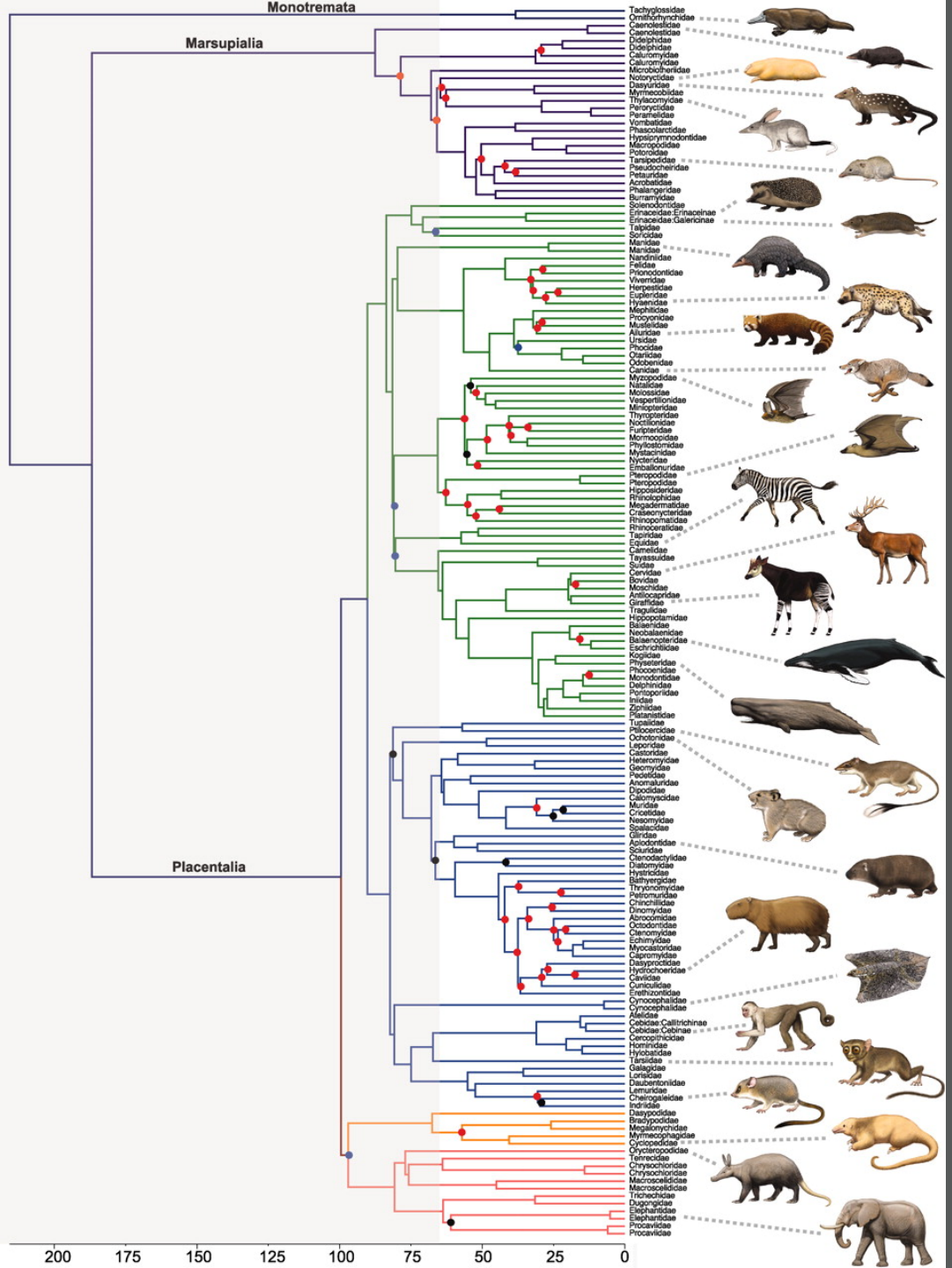


Similarity

best-guess

evolutionary → Character reconstruction

A tree is a representation of relationships



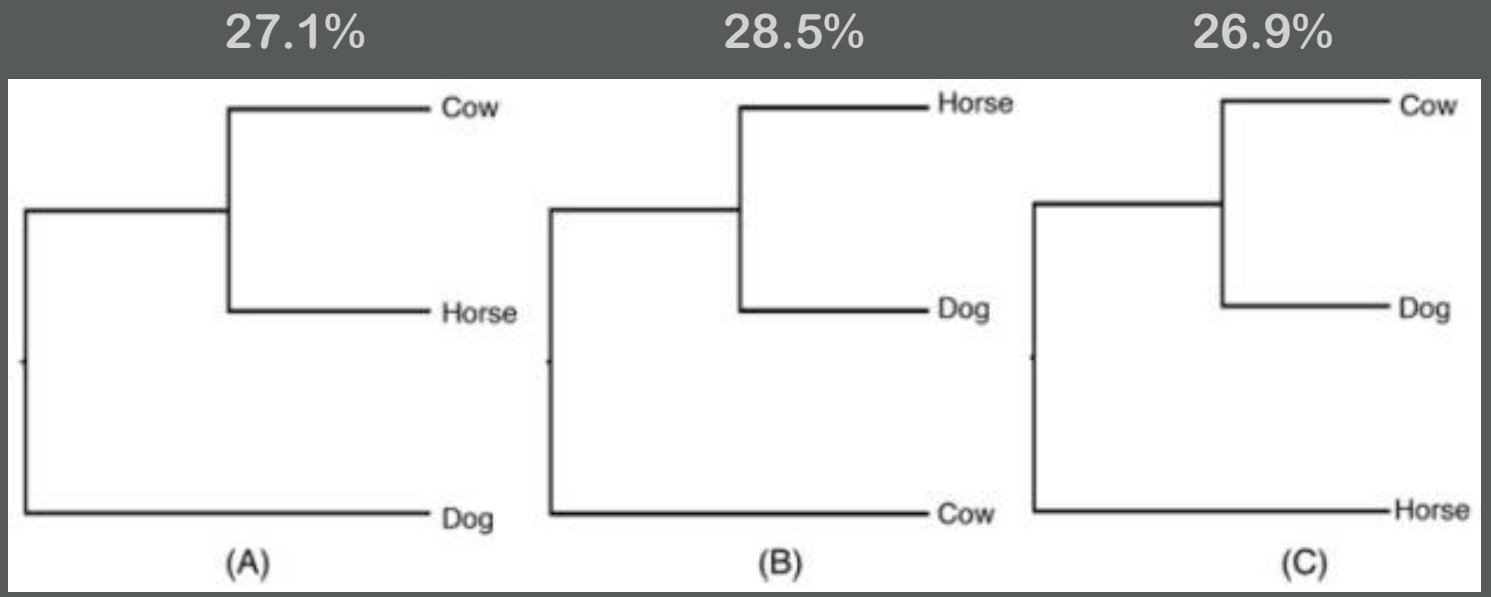
Uncertainty is common

- Nucleotide and protein tree disagree
- Disagrees with previous best tree

Meredith et al (2011) Science

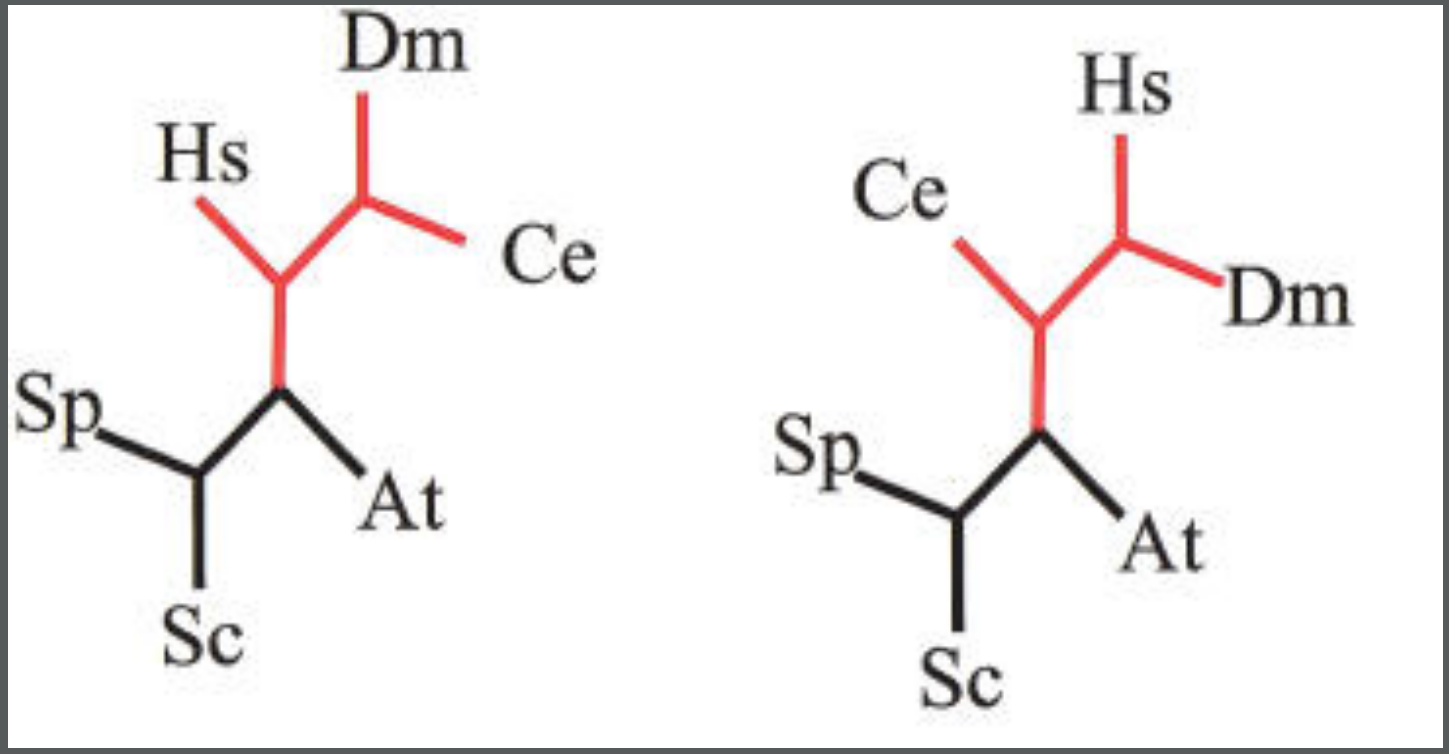
Uncertainty is common – even in unexpected places

Are horses more closely related to dogs than to cows?



Hou et al (2009) Mol Phyl Evol

Uncertainty is common



Ecdysozoa

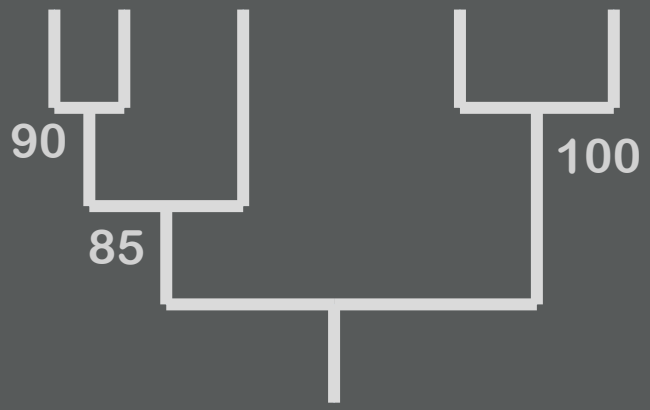
Coelomata

Wolf et al (2004) Genome Res

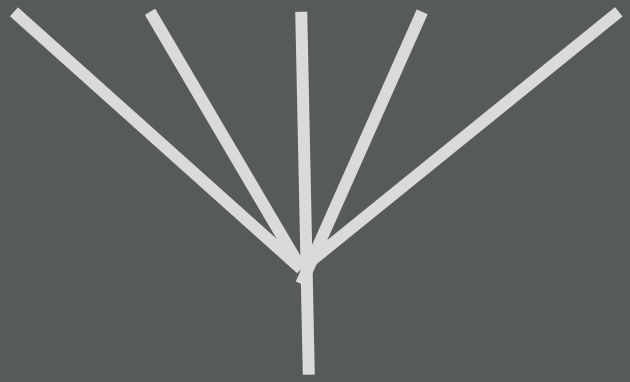
Trees can be treacherous

Representing unresolved relationships

Bootstrap values

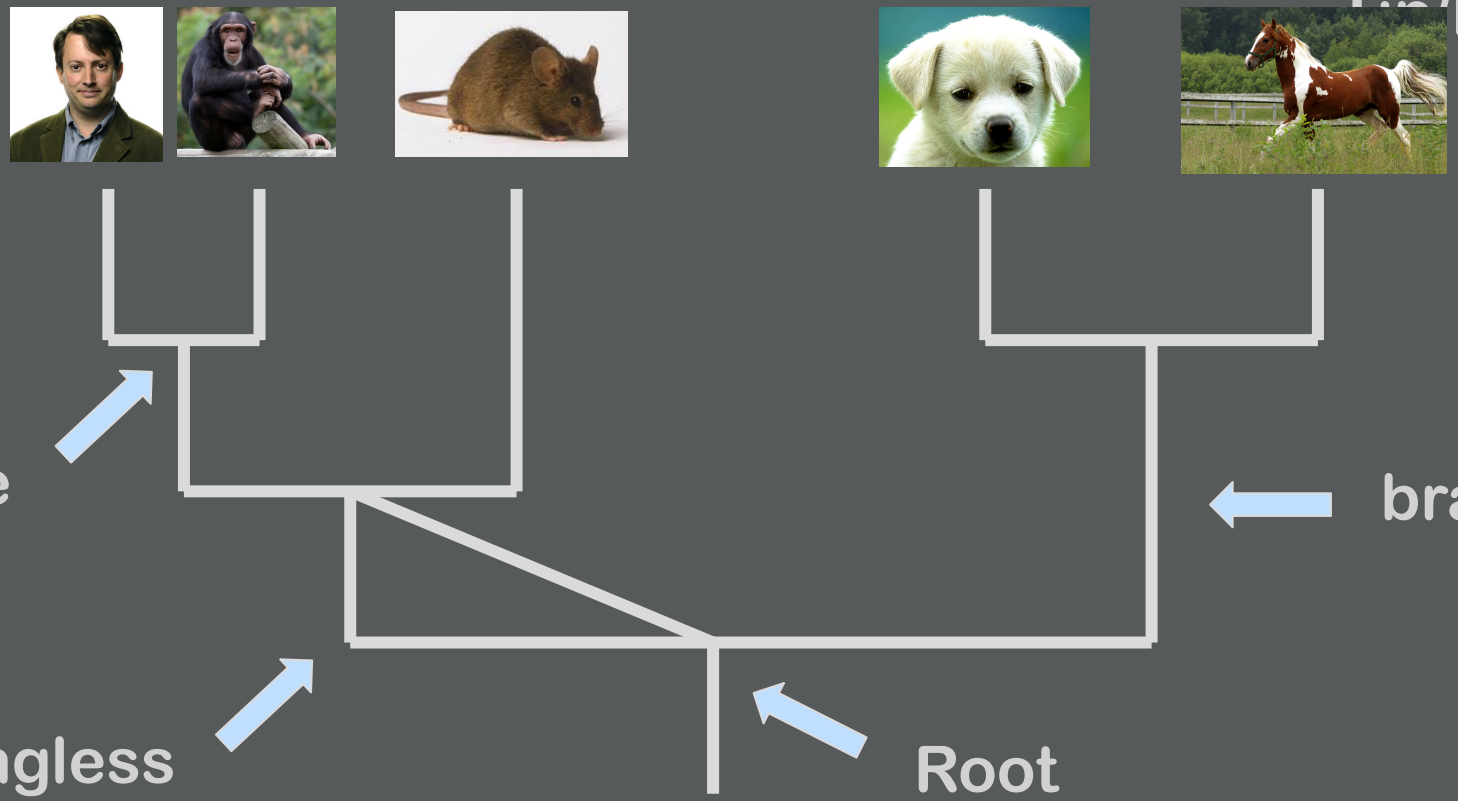


Bifurcating tree

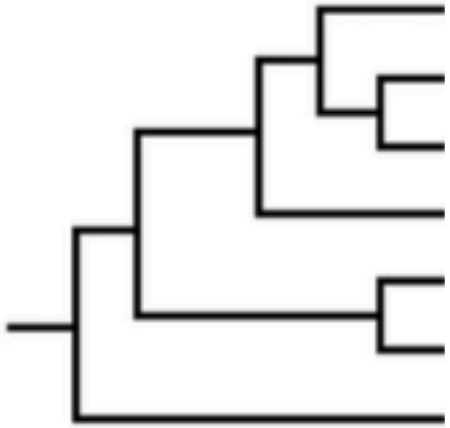


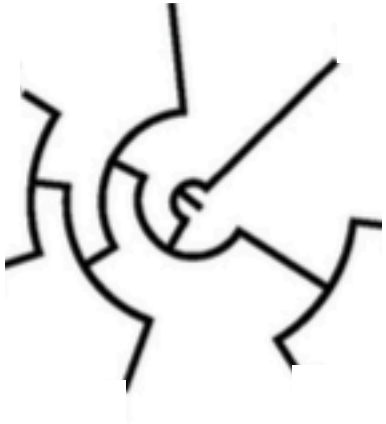
Polytomy

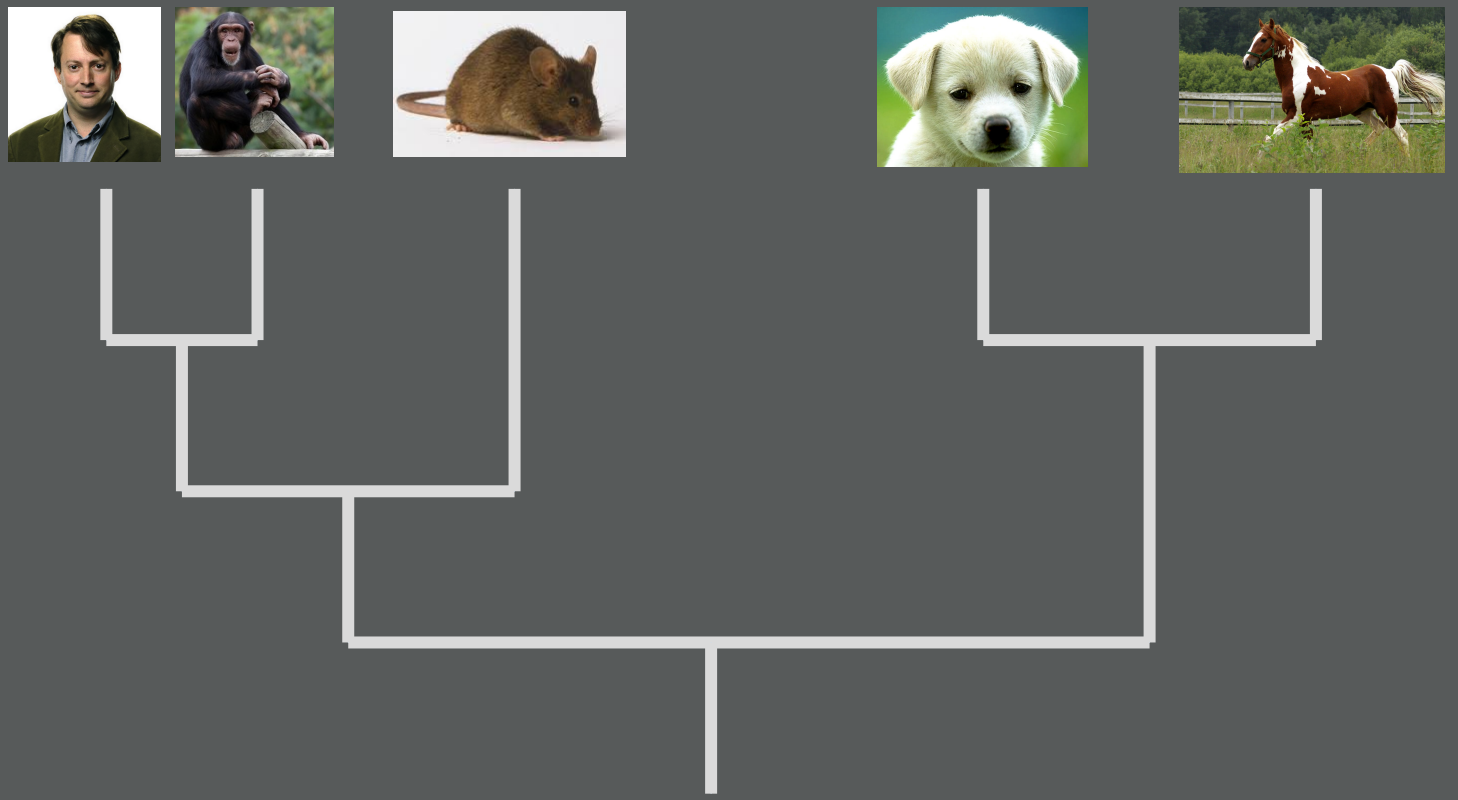
Some inevitable terminology

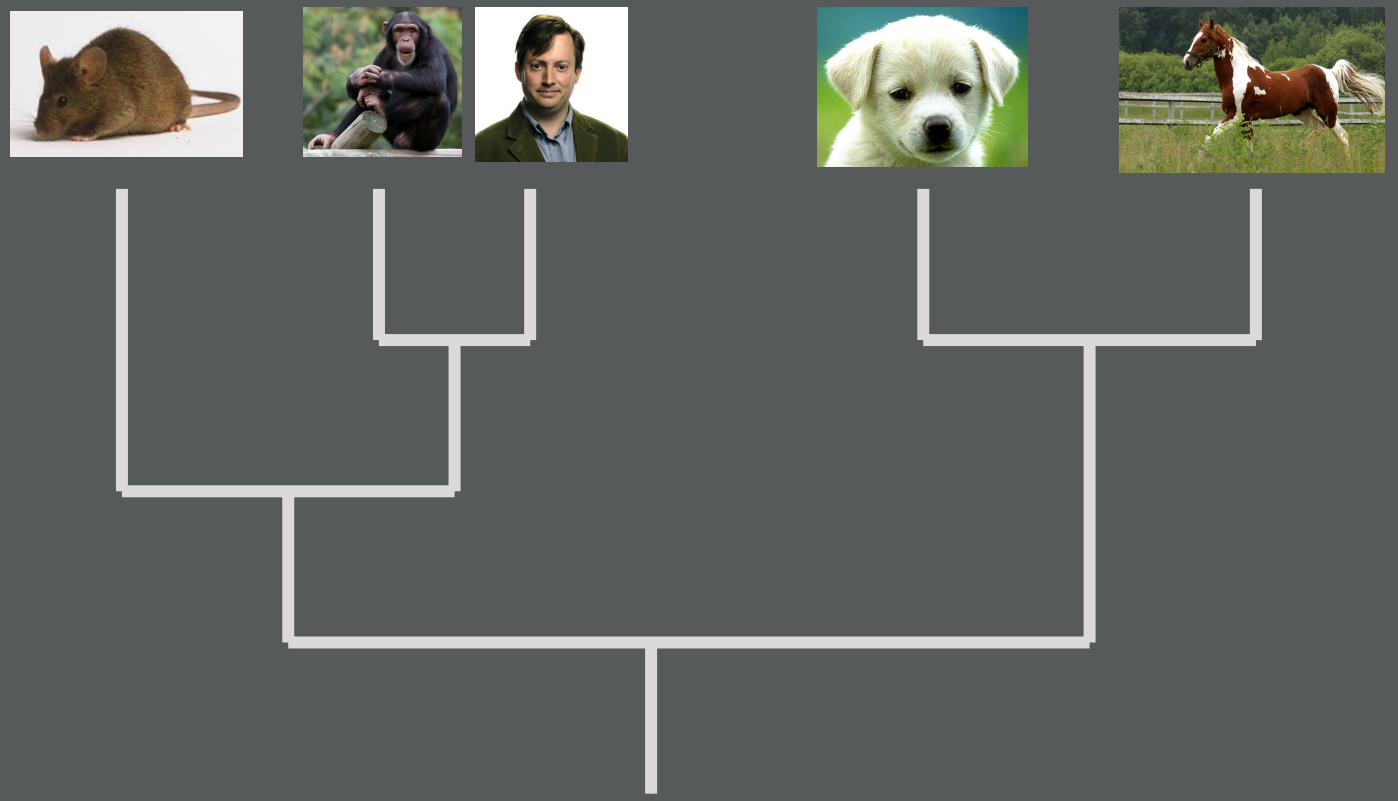


Different ways to represent the same tree



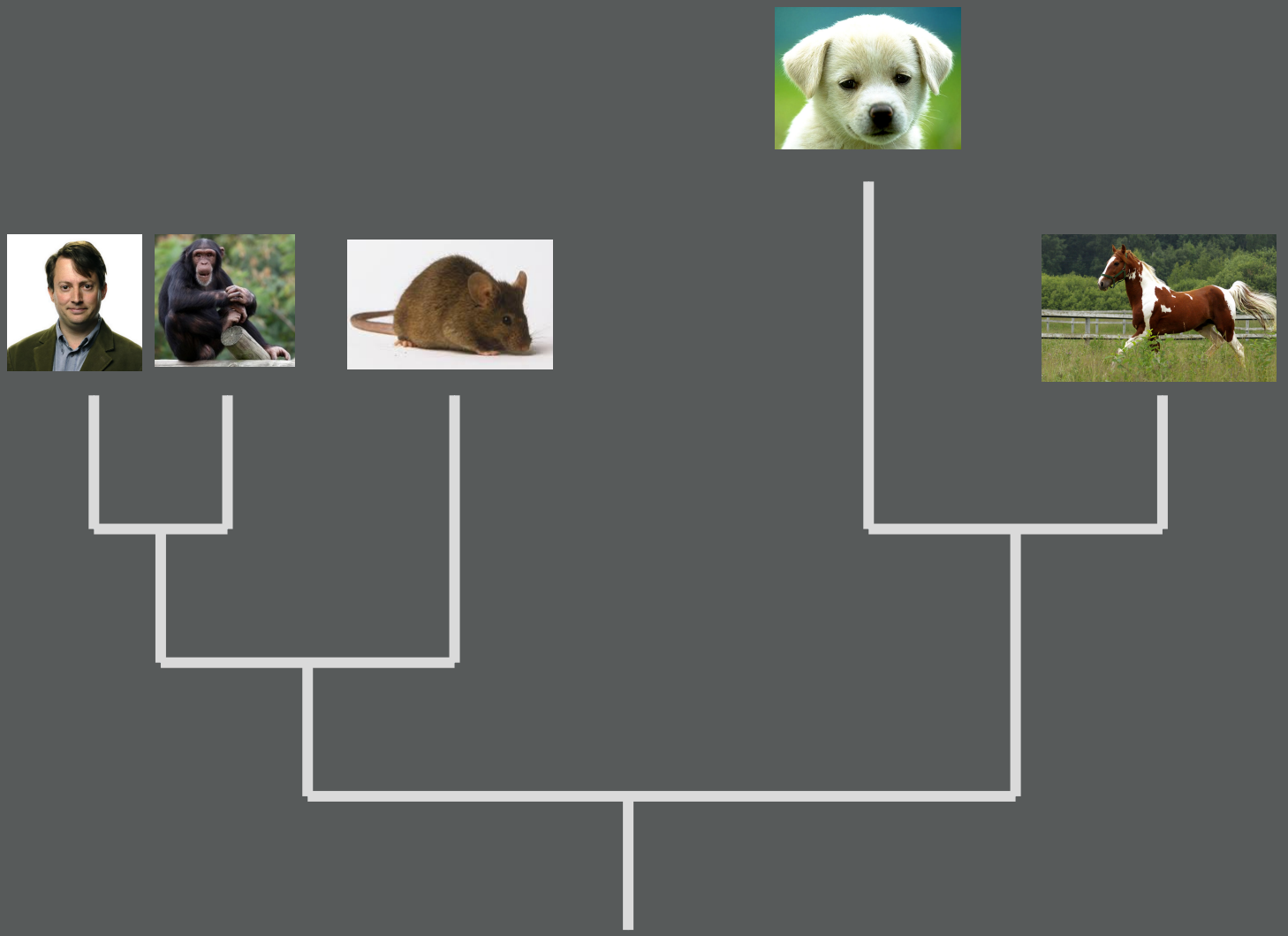






How does this tree differ?

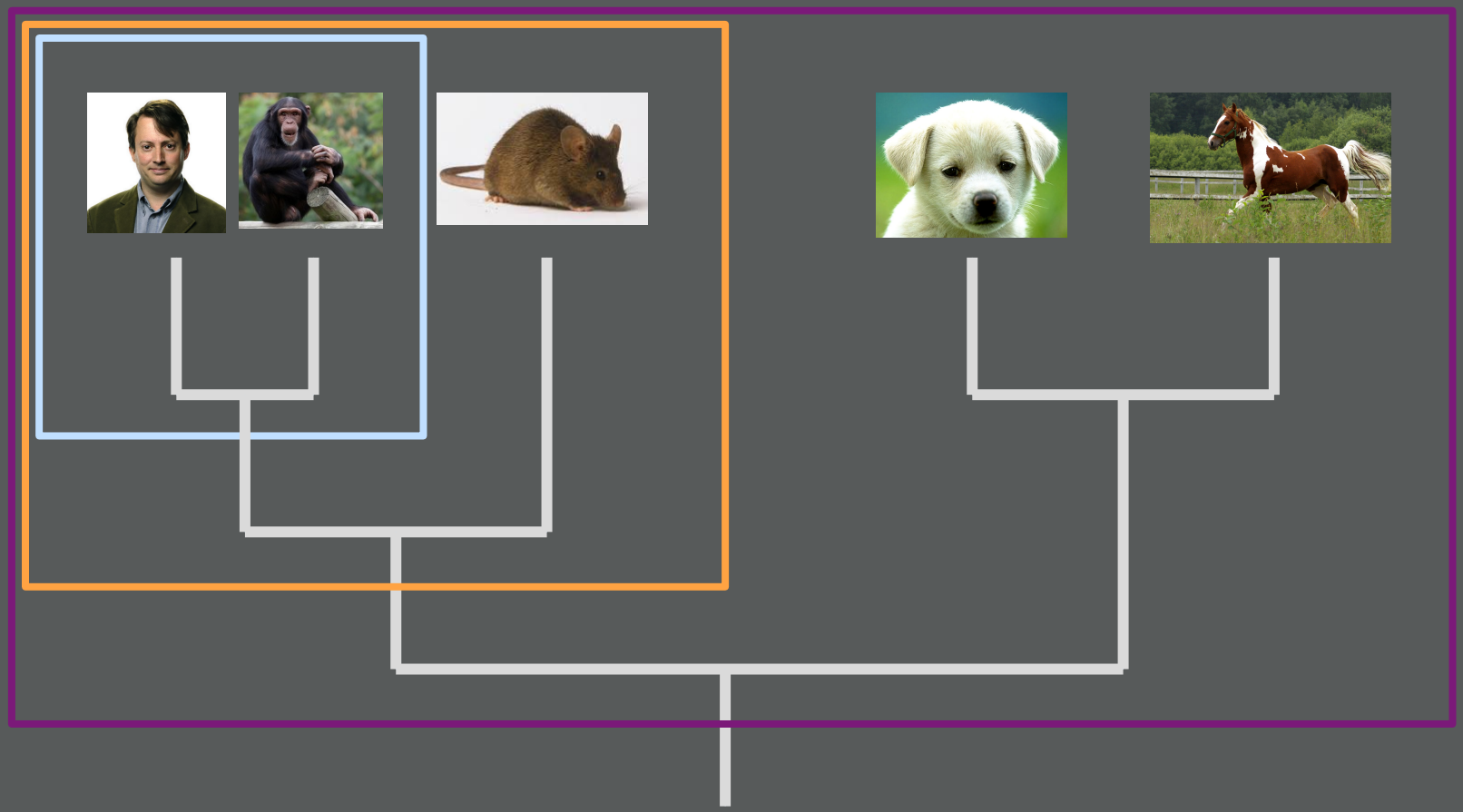
Branches can freely rotate around a node

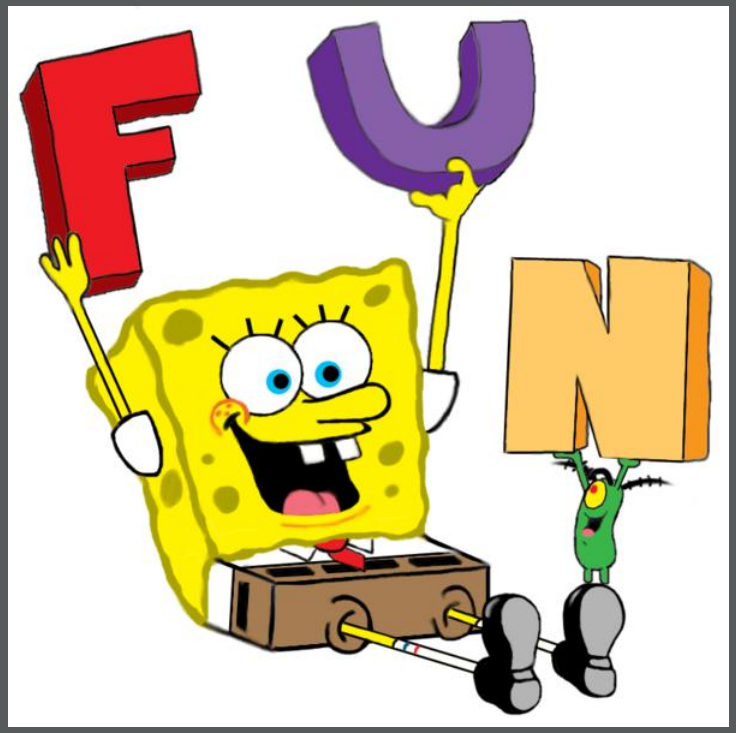


How about *this* tree?

Some trees (cladograms) only show relationship

Clade/monopylum: An ancestor with *all* its dependants

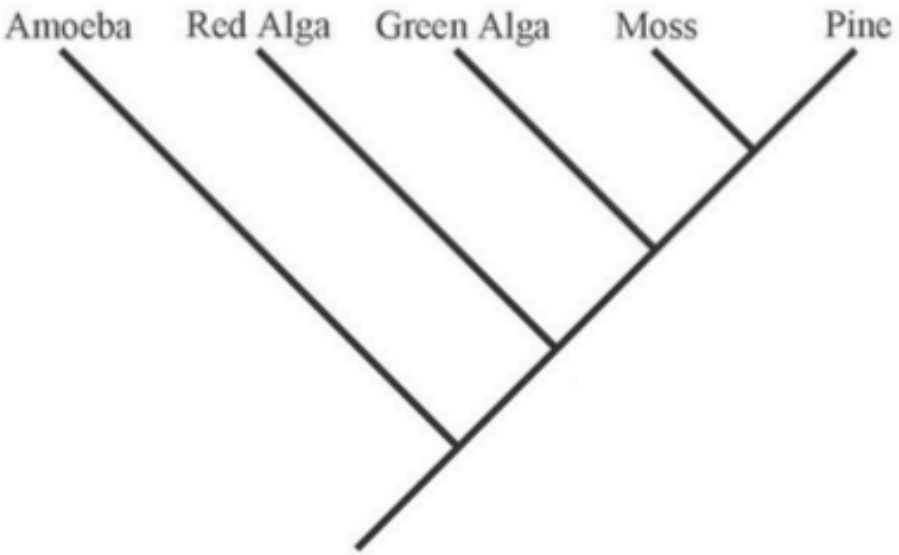




with trees (Part I)

brought to you by Baum et al (2005) Science

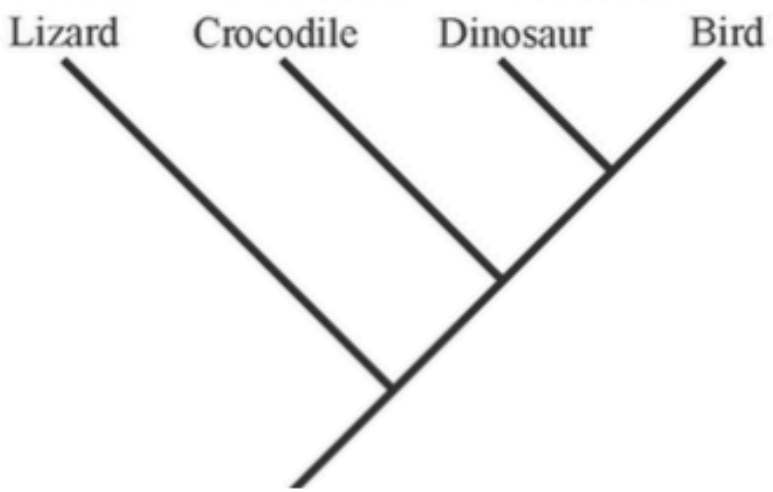
Tree 1



1) By reference to the tree above, which of the following is an accurate statement of relationships?

a) A green alga is more closely related to a red alga than to a moss
b) A green alga is more closely related to a moss than to a red alga
c) A green alga is equally related to a red alga and a moss
d) A green alga is related to a red alga, but is not related to a moss

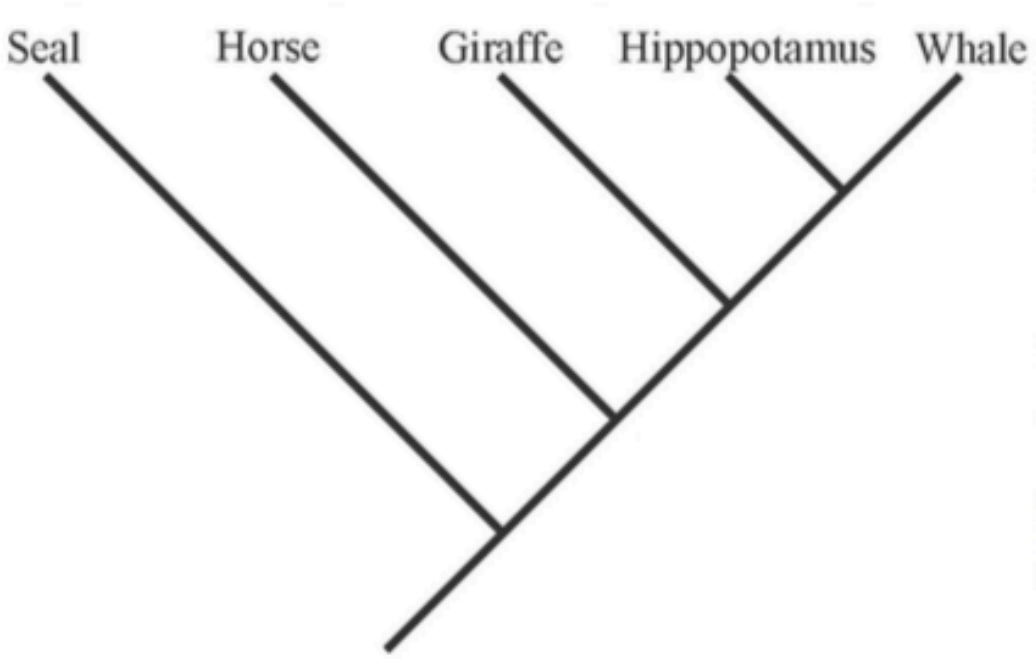
Tree 2



2) By reference to the tree above, which of the following is an accurate statement of relationships?

a) A crocodile is more closely related to a lizard than to a bird
b) A crocodile is more closely related to a bird than to a lizard
c) A crocodile is equally related to a lizard and a bird
d) A crocodile is related to a lizard, but is not related to a bird

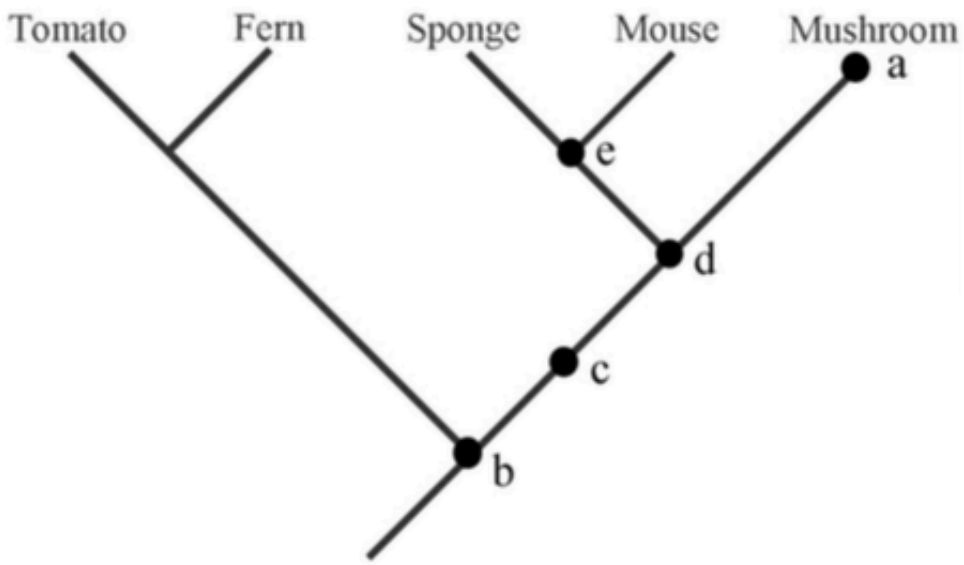
Tree 3



3) By reference to the tree above, which of the following is an accurate statement of relationships?

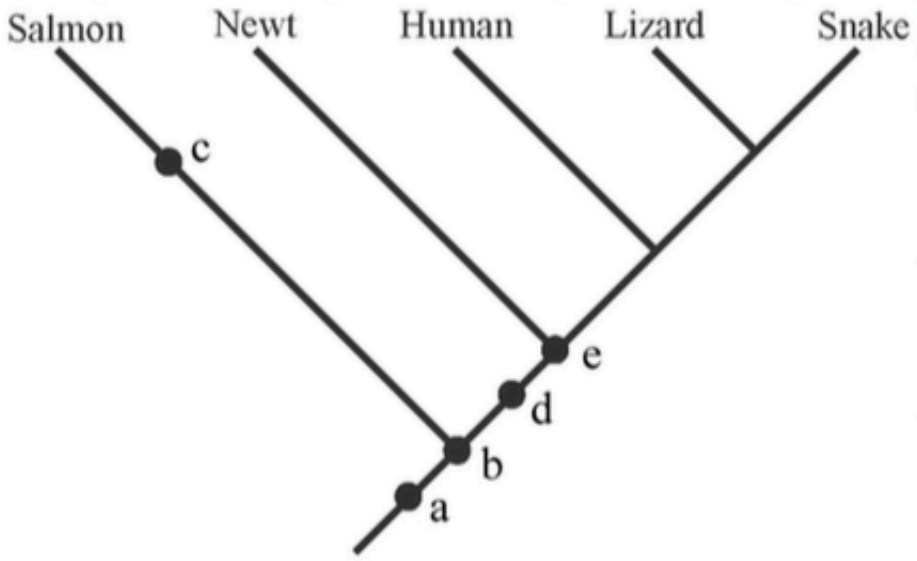
a) A seal is more closely related to a horse than to a whale
b) A seal is more closely related to a whale than to a horse
c) A seal is equally related to a horse and a whale
d) A seal is related to a whale, but is not related to a horse

Tree 4



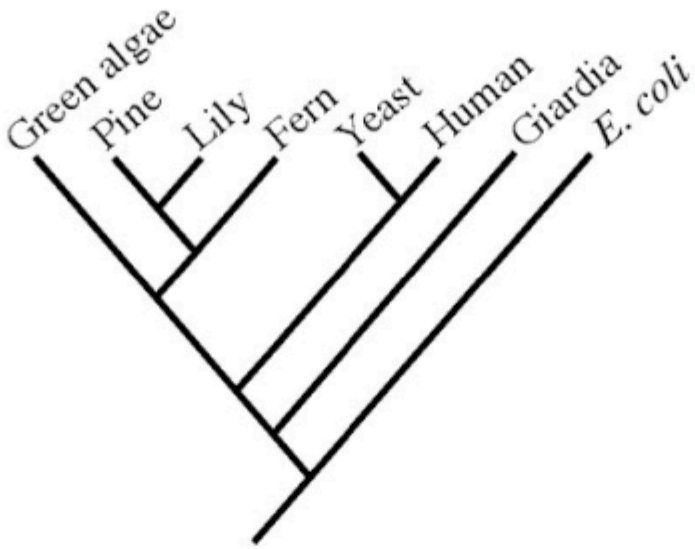
4) Which of the five marks in the tree above corresponds to the most recent common ancestor of a mushroom and a sponge?

Tree 5

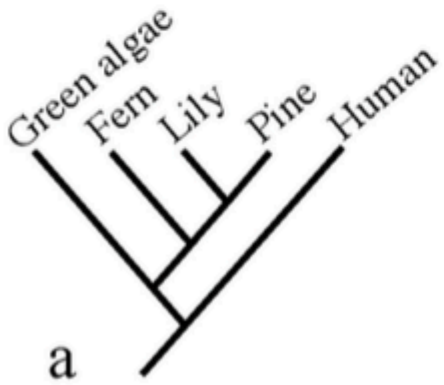


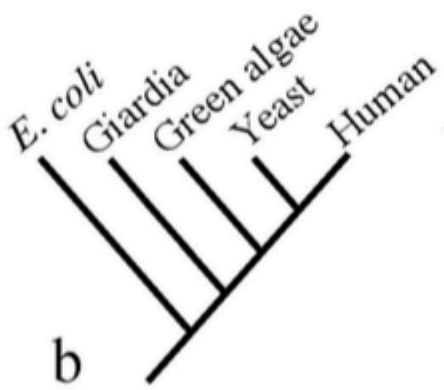
5) If you were to add a trout to the phylogeny shown above, where would its lineage attach to the rest of the tree?

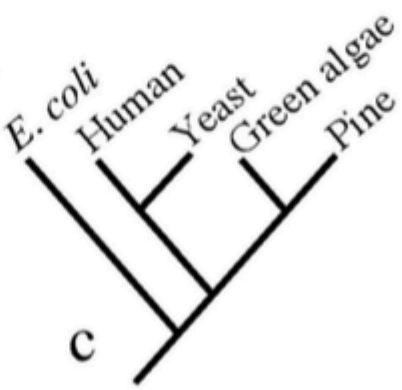
Tree 6

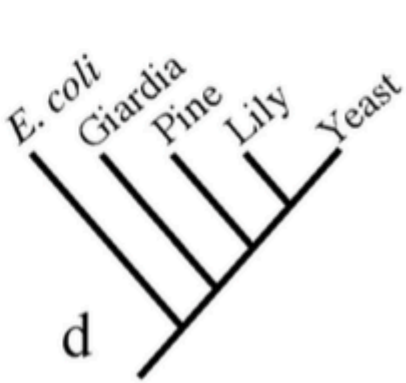


6) Which of trees below is false given the larger phylogeny above?

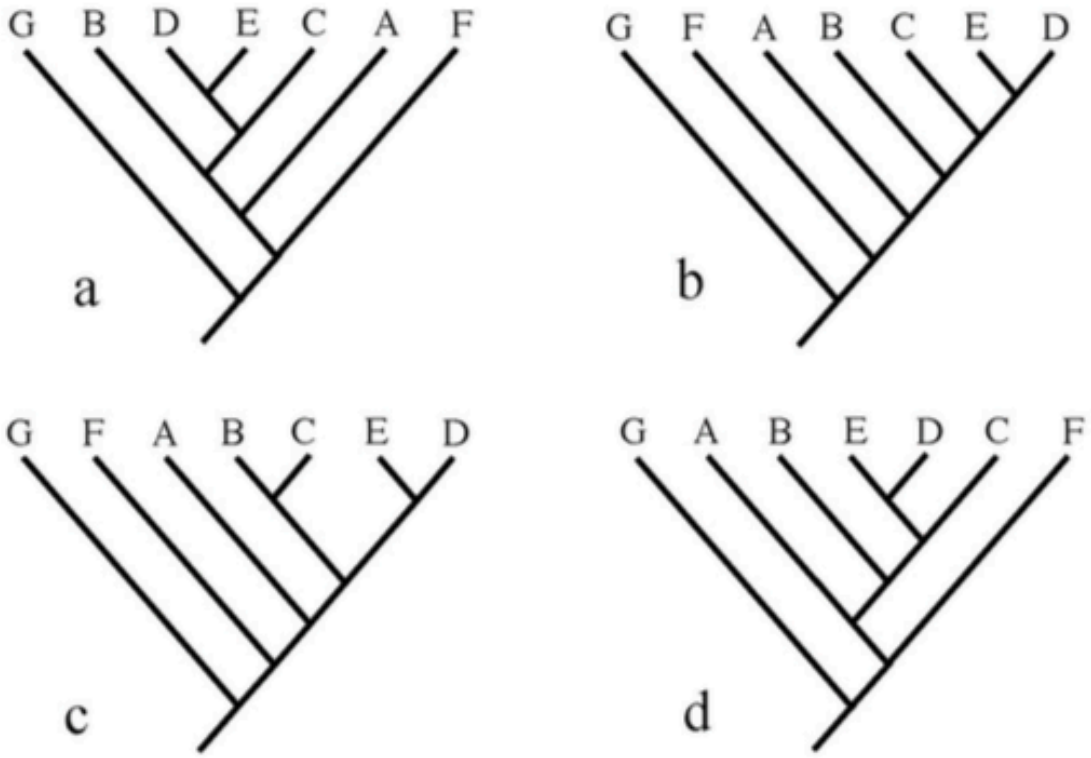






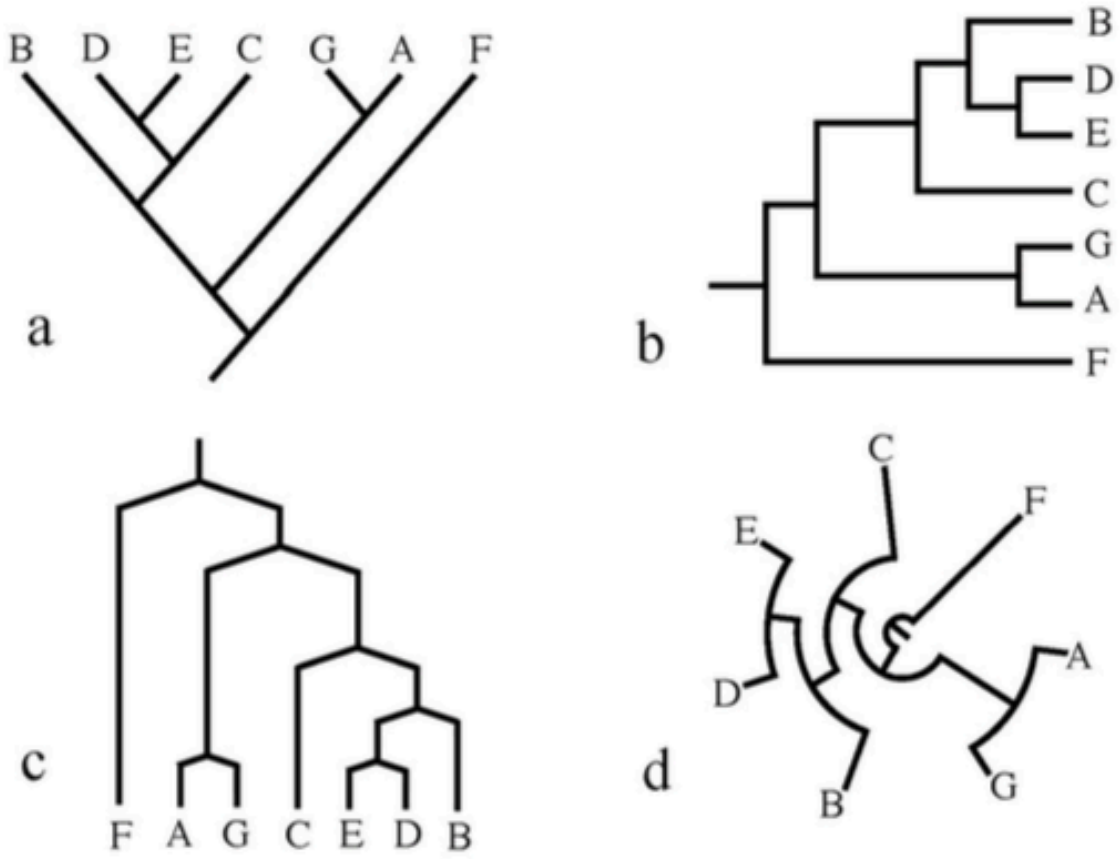


Tree 7



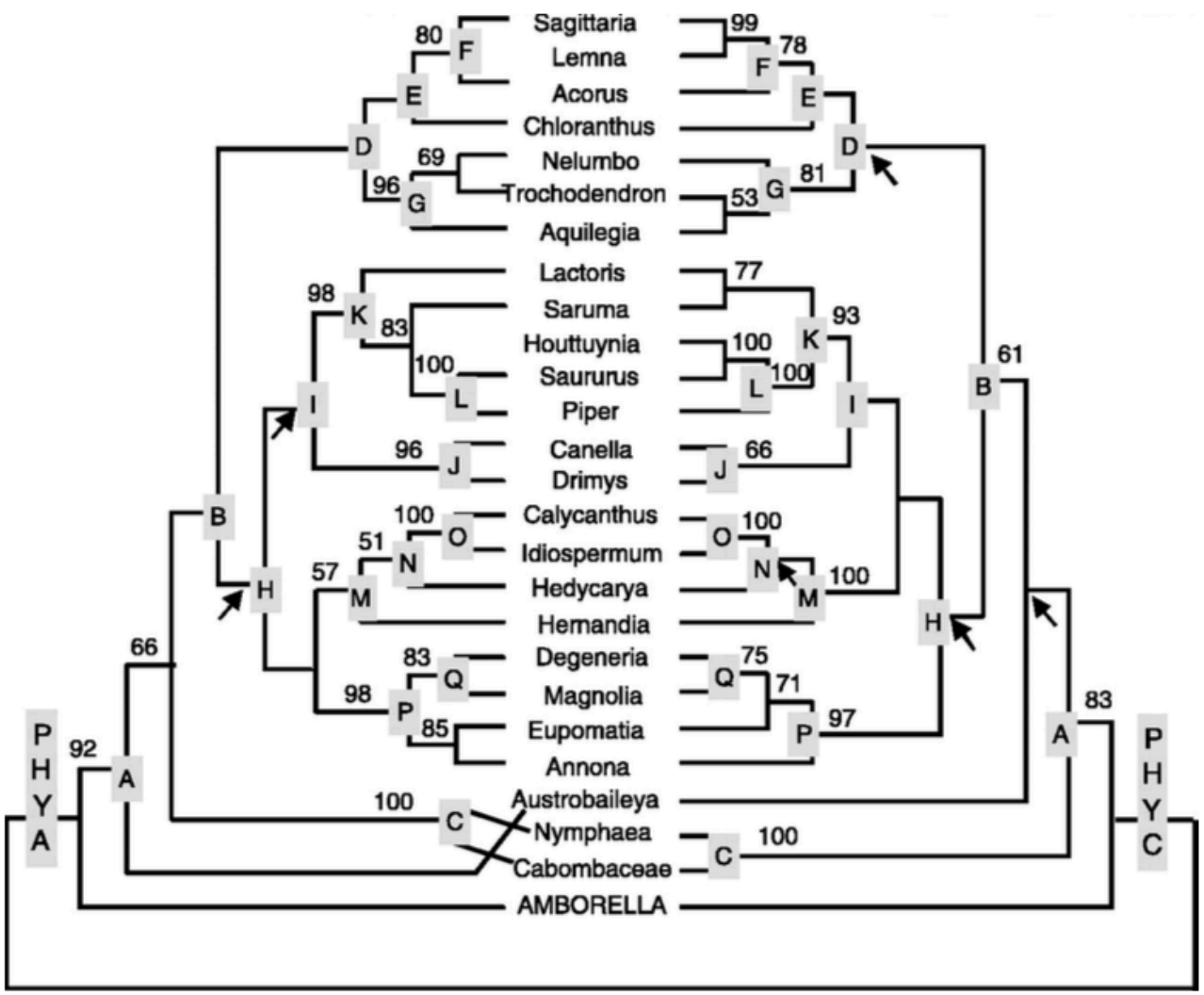
7) Which of the four trees above depicts a different pattern of relationships than the others?

Tree 8



8) Which of the four trees above depicts a different pattern of relationships than the others?

Tree 9

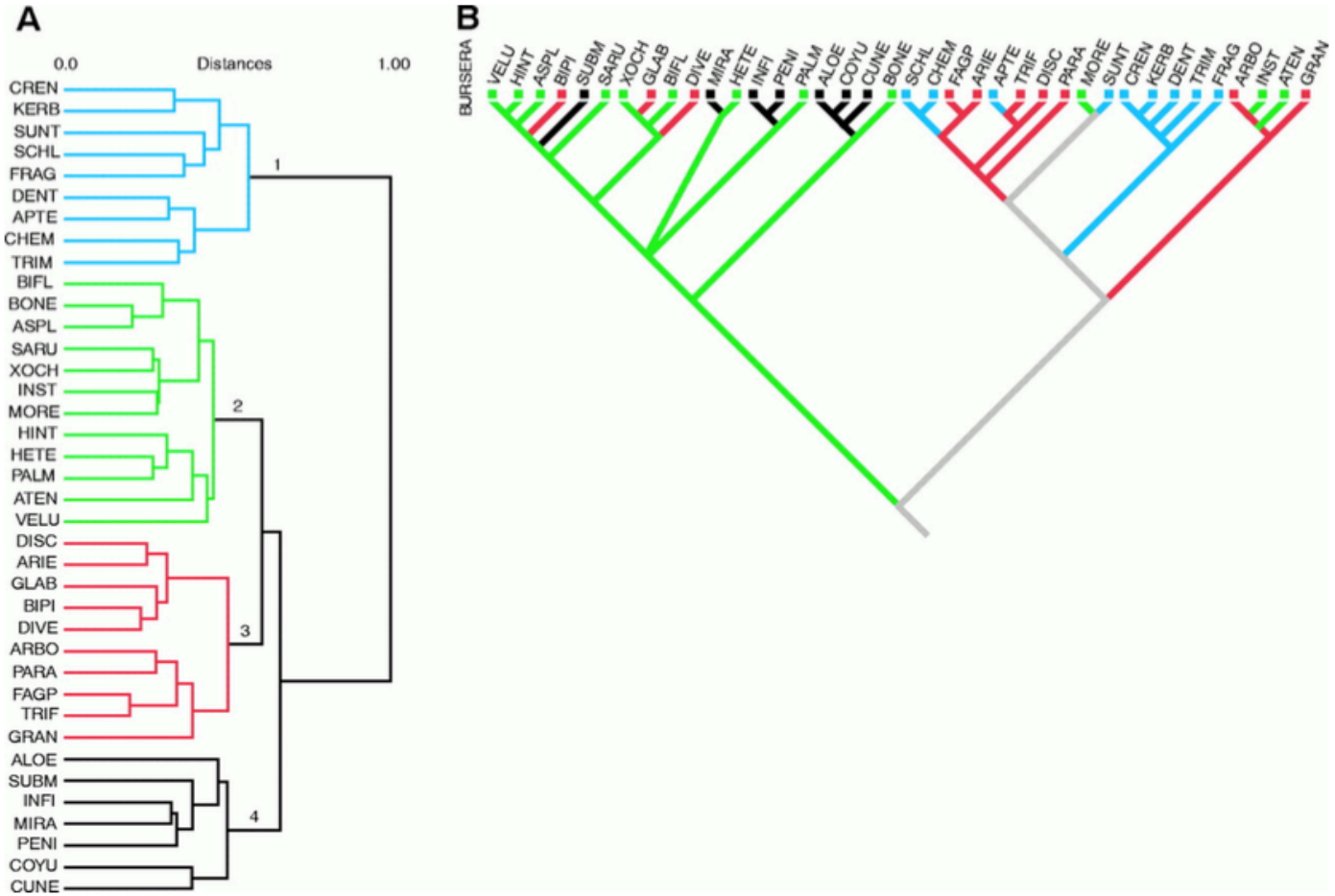


S. Mathews, M. J. Donoghue. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* **286**, 947 (1999).

1) The figure above shows the phylogeny estimated for a sample of flowering plants (angiosperms) from *PHYTOCHROME A* and *PHYTOCHROME C*, a pair of genes that duplicated prior to the origin of the angiosperms. Which of the following sets of taxa constitute a clade (=monophyletic group) on one gene tree but not on the other?

a) *Degeneria-Magnolia-Eupomatia*
b) All angiosperms except *Amborella*
c) *Austrobaileya-Nymphaea*-Cabombaceae
d) *Nelumbo-Trochodendron-Aquilegia*

Tree 10

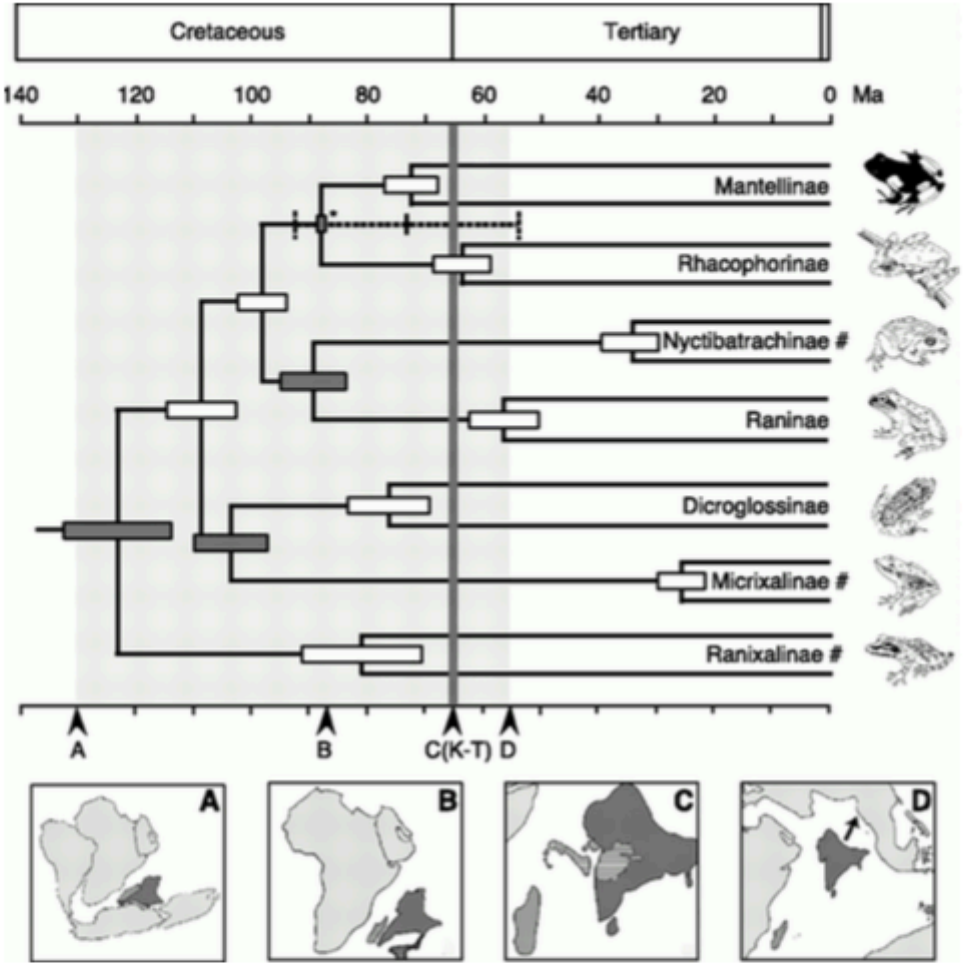


J. X. Becerra. Insects on plants: macroevolutionary chemical trends in host use. *Science* **276**, 253 (1997).

2) The dendrogram on the left clusters plant species by chemical similarity; each of the four main chemical groups is indicated with a different color. This tree does not depict descent relationships, just degree of chemical similarity. On the right, the evolution of these chemical types is reconstructed on a phylogeny of the plants (this does depict inferred evolutionary relationships). The colors correspond to the chemical groups on the left, and the gray branches indicate uncertainty in character reconstruction. What does a comparison of these two figures tell us about the evolution of plant secondary chemistry?

a) The four groups of chemically similar species each constitutes a distinct evolutionary lineage
b) The group colored "black" has the most advanced chemical defenses
c) The red (3) and blue (1) chemical groups are most distantly related
d) The chemical groups have each been gained and/or lost multiple times in evolution

Tree 11

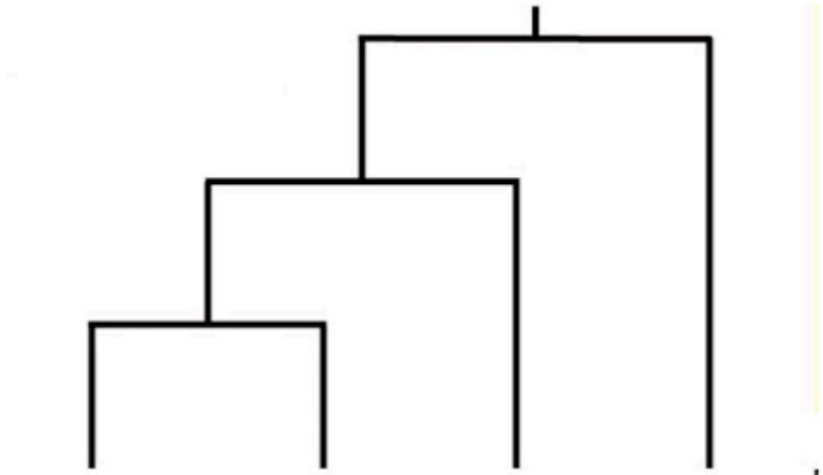


F. Bossuyt, M. C. Milinkovitch. Amphibians as indicators of early tertiary "out-of-India" dispersal of vertebrates. *Science* **292**, 93 (2001).

3) This tree depicts inferred relationships among some major frog groups with branches drawn proportional to absolute time. Error bars on internal nodes depict confidence intervals on the dates of estimated nodes. Assuming this tree and the associated ages are correct which of the following statements is true?

a) No individual living before 70 million years ago is an ancestor of Raninae
b) Raninae and Dicroglossinae shared a common ancestor about 75 million years ago
c) The divergence of Raninae and Nyctibatrachinae occurred more recently than the 85 million year old separation of India from Madagascar
d) The last common ancestor of Micrixalinae and Dicroglossinae lived before India and Madagascar separated (85 million years ago)

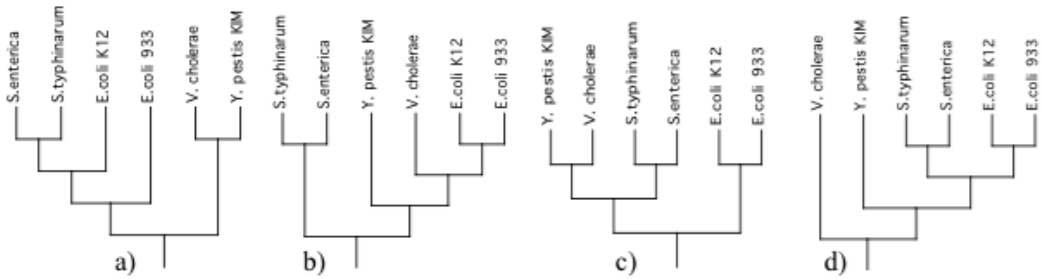
Tree 12



Genome A	Genome B	Genome C	Genome D
E. coli 933	*E. coli K12*	*S. typhimurium*	*Y. pestis KIM*
S. enterica	*S. typhimurium*	*E. coli K12*	*Y. pestis KIM*
E. coli K12	*S. typhimurium*	*Y. pestis KIM*	*V. cholerae*

V. Daubin, N. A. Moran, H. Ochman. Phylogenetics and the cohesion of bacterial genomes. *Science* **301,** 829 (2003).
[The above is only a portion of the figure].

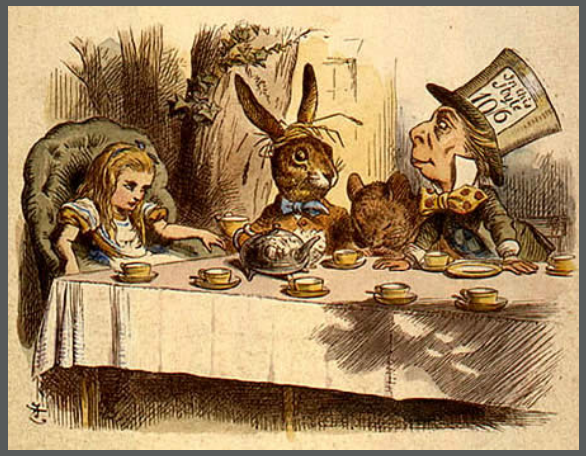
5) Each row in the table above lists a set of four bacterial taxa whose relationship follows the topology shown. Thus each row can be read as a four-taxon tree. Which of the four trees below is compatible with the information in the three rows of the table?



Reading trees – key points

- A tree is a representation of relationships – often based on some measure of distance.
- To find the most recent common ancestor, find the first shared node
- In some trees the ancestor has been reconstructed
- Trees are best guesses – there can be considerable uncertainty (more about how we can quantify this later)

10 minute break



Why bother?

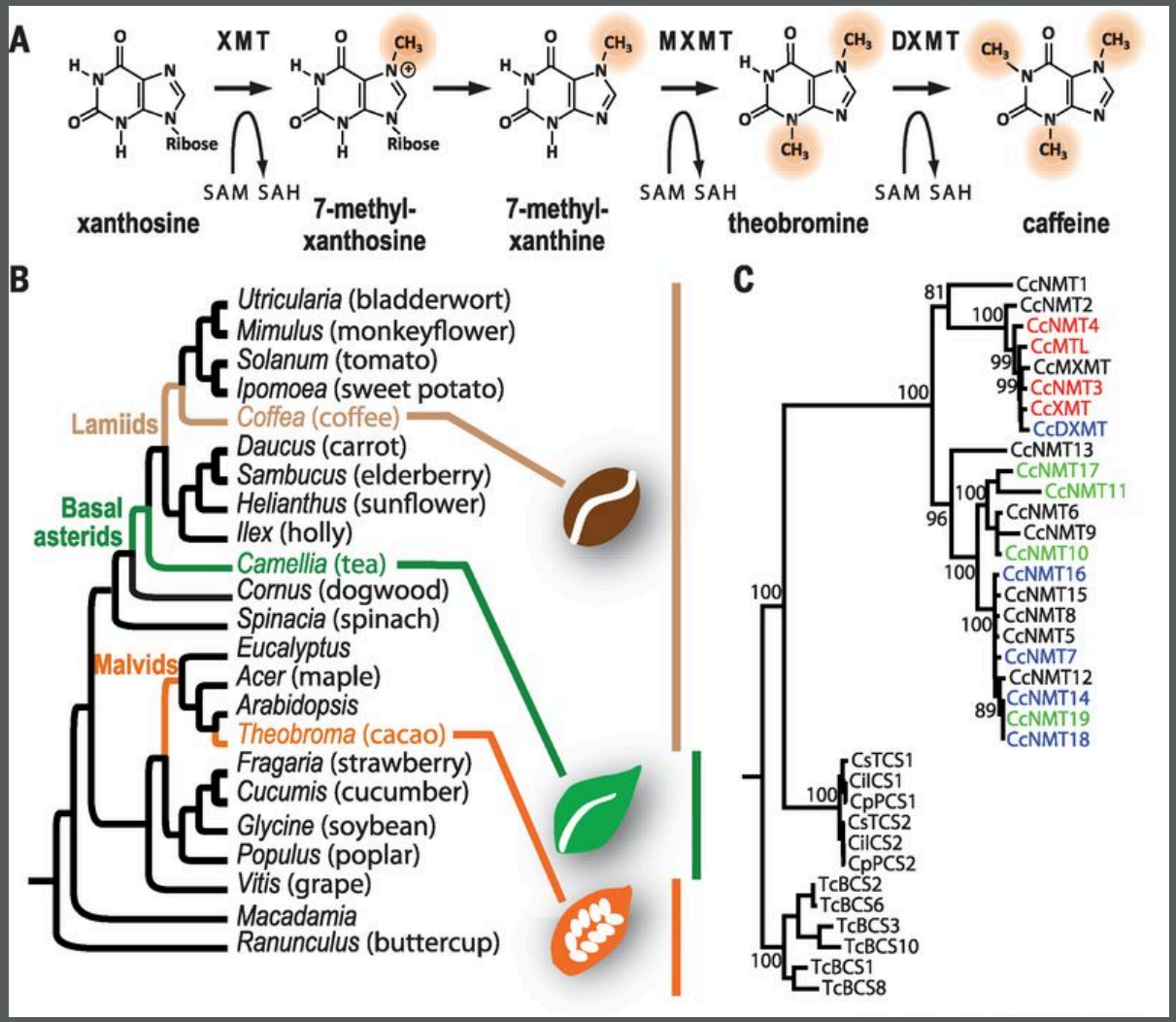
Trees can help us to understand

- Relationship between species
- How phenotypic traits map onto evolutionary history

Trees can help us to understand

- evolutionary processes and their timing

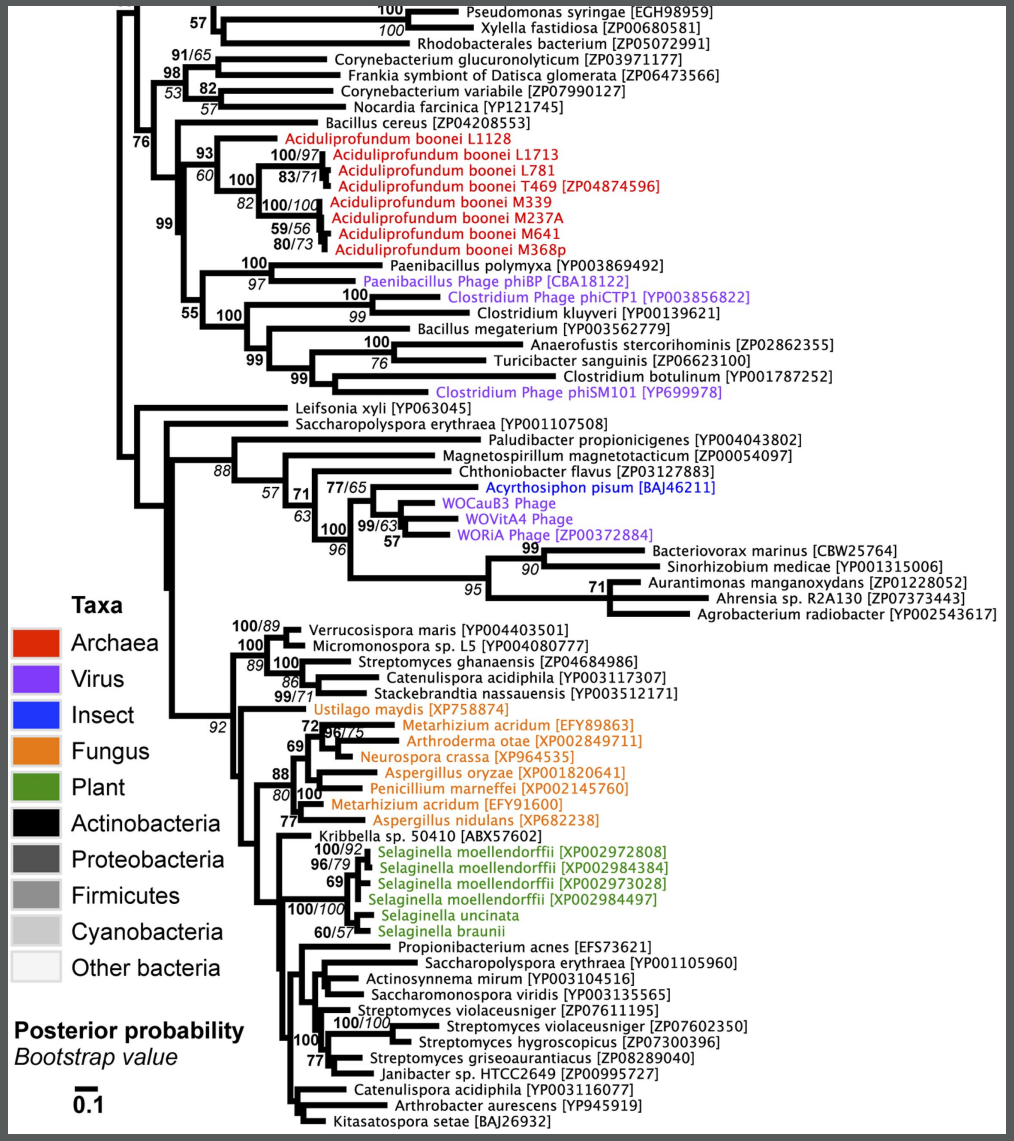
Gene gain and loss



A polyphyletic origin of caffeine

Denoeud et al (2014) Science

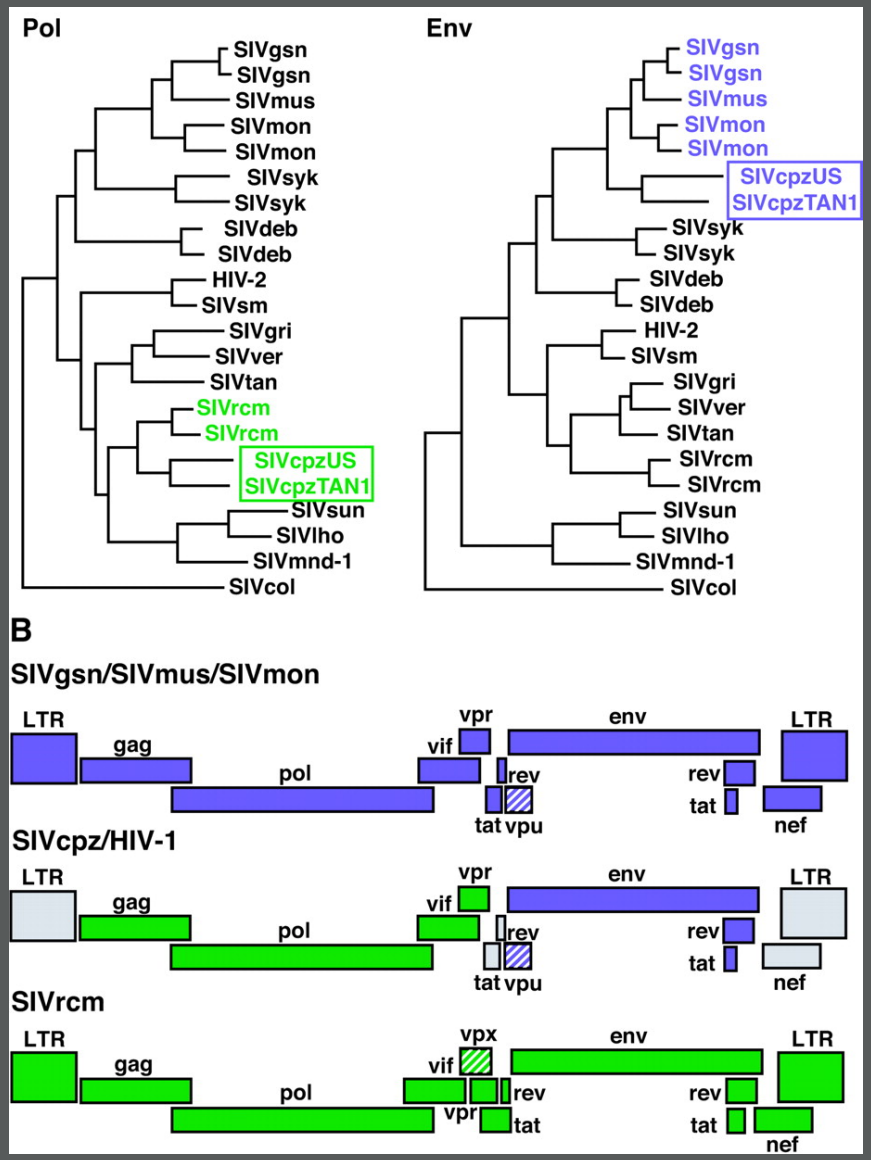
Horizontal gene transfer



GH25 muramidase

Metcalf et al (2014) eLife

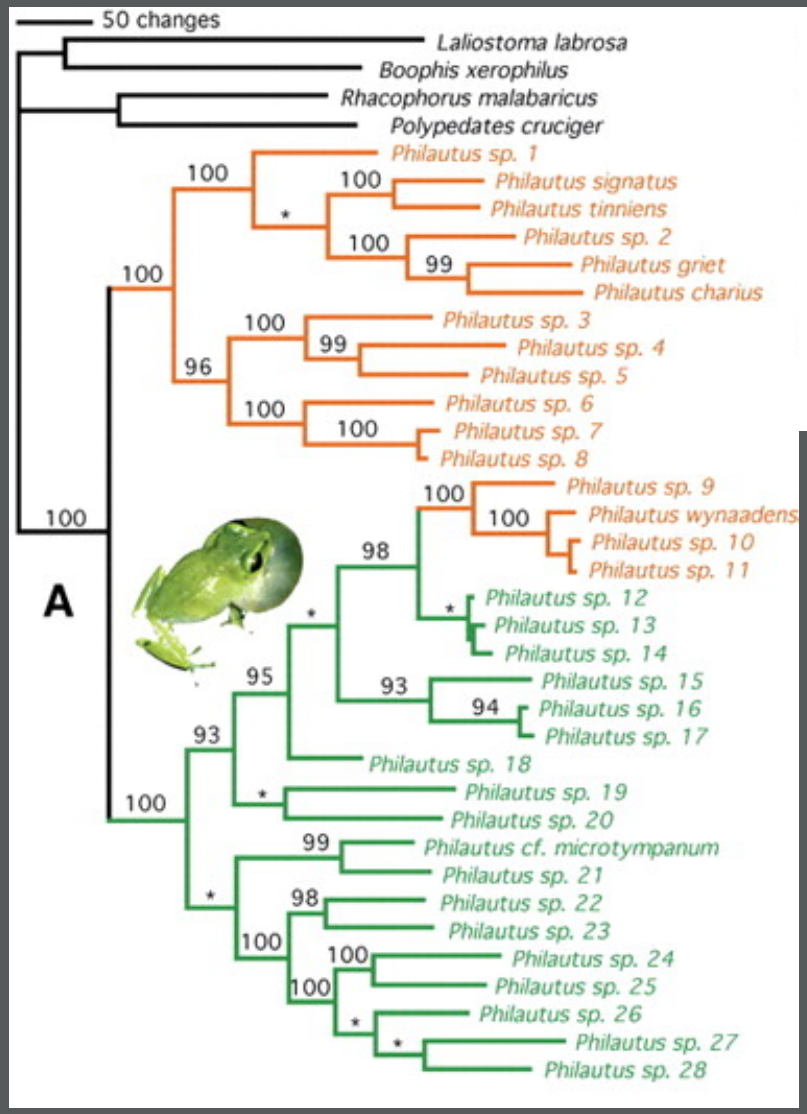
Recombination



Recombination between different strains of simian HIV (SIV)

Sharp et al (2005) J Virol

Migration



Southern India

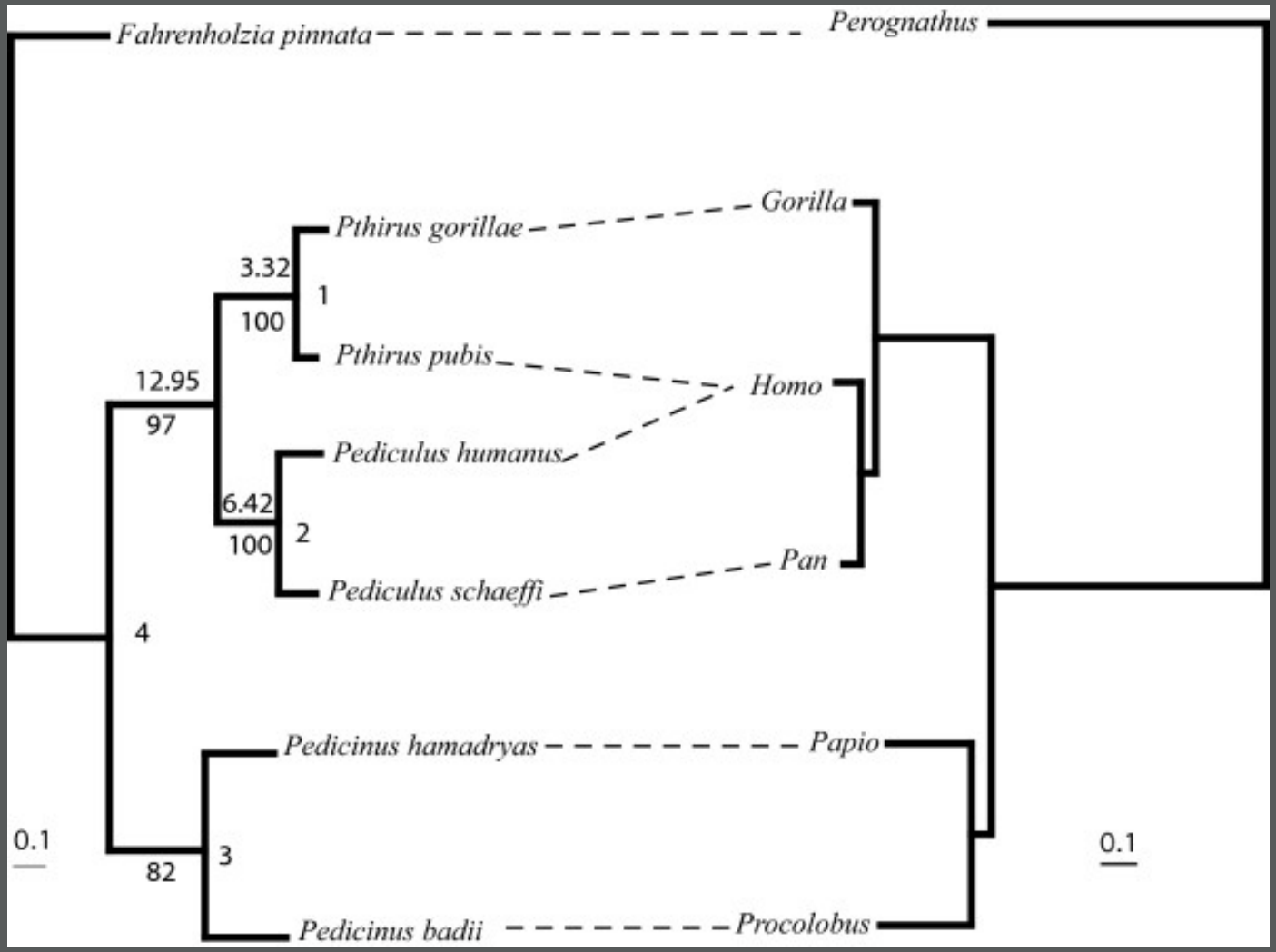
Sri Lanka

Phylogeography

Bossuyt et al (2004) Science

Co-evolution

Lice



Primates

Reed et al (2007) BMC Biol

Mutation (polarity)

ACCAGAT

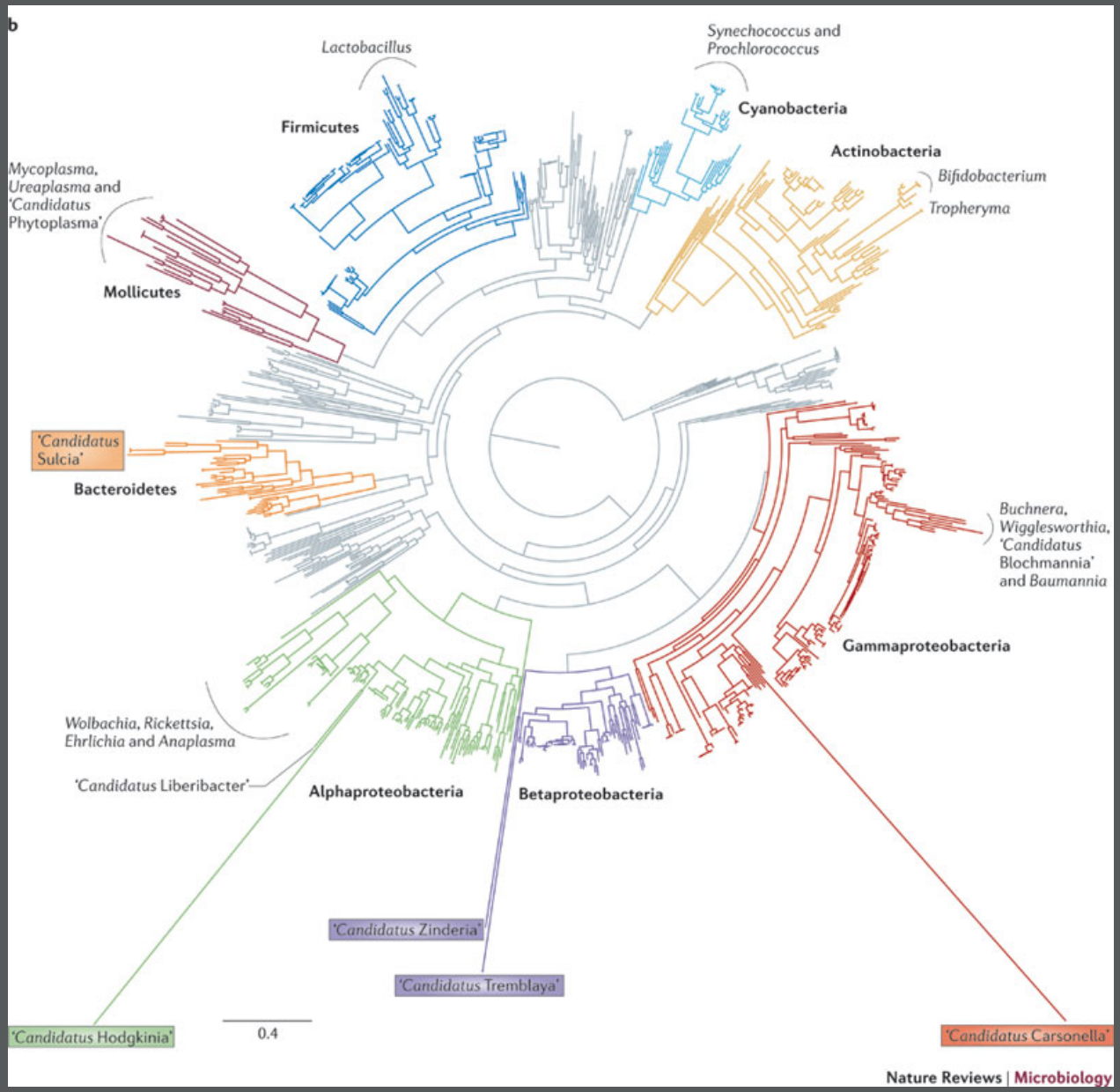
ACCAGAC

ACCAGAC

Human-specific
Substitution
C ➡ T

ACCAGAC

Rates of evolution and selection



Fast-evolving endosymbionts

McCutcheon & Moran (2012) Nat Rev Microb

Any questions on this part?

Making a tree – basic steps

1. **Get** a sequence (not as easy as it sounds…)

2. **Find** similar sequences, hoping that they are

   - Related by descent (**orthologs**) or
   - Related by duplication (**paralogs**) or
   - Not related by chance

3. **Align** them

4. **Tidy** up the alignment

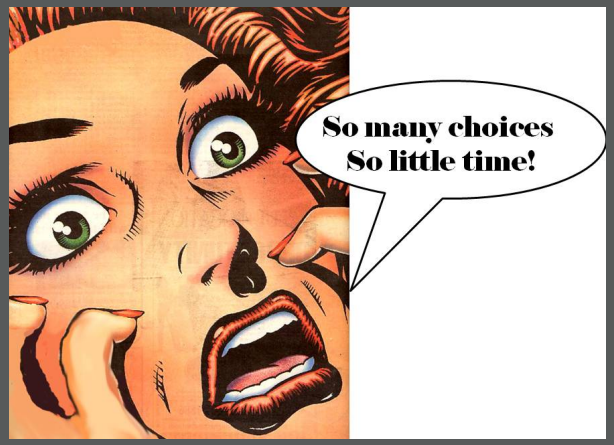
5. **Make** a tree! → **Evaluate** the tree

   ↓

   Use the tree

Get a sequence

- Where do you go?



Let's retrieve the
KDM2A
gene in Ensembl

www.ensembl.org

Single gene analyses:

- Ensembl (vertebrates, some others)
- Model organism databases
  - Flybase, wormbase, ecocyc
- UCSC (a bit of everything)
- Uniprot

Bulk analyses:

- Biomart (interactive retrieval)
- UCSC
- NCBI genomes (pre-parsed files)
  - GenBank, Refseq

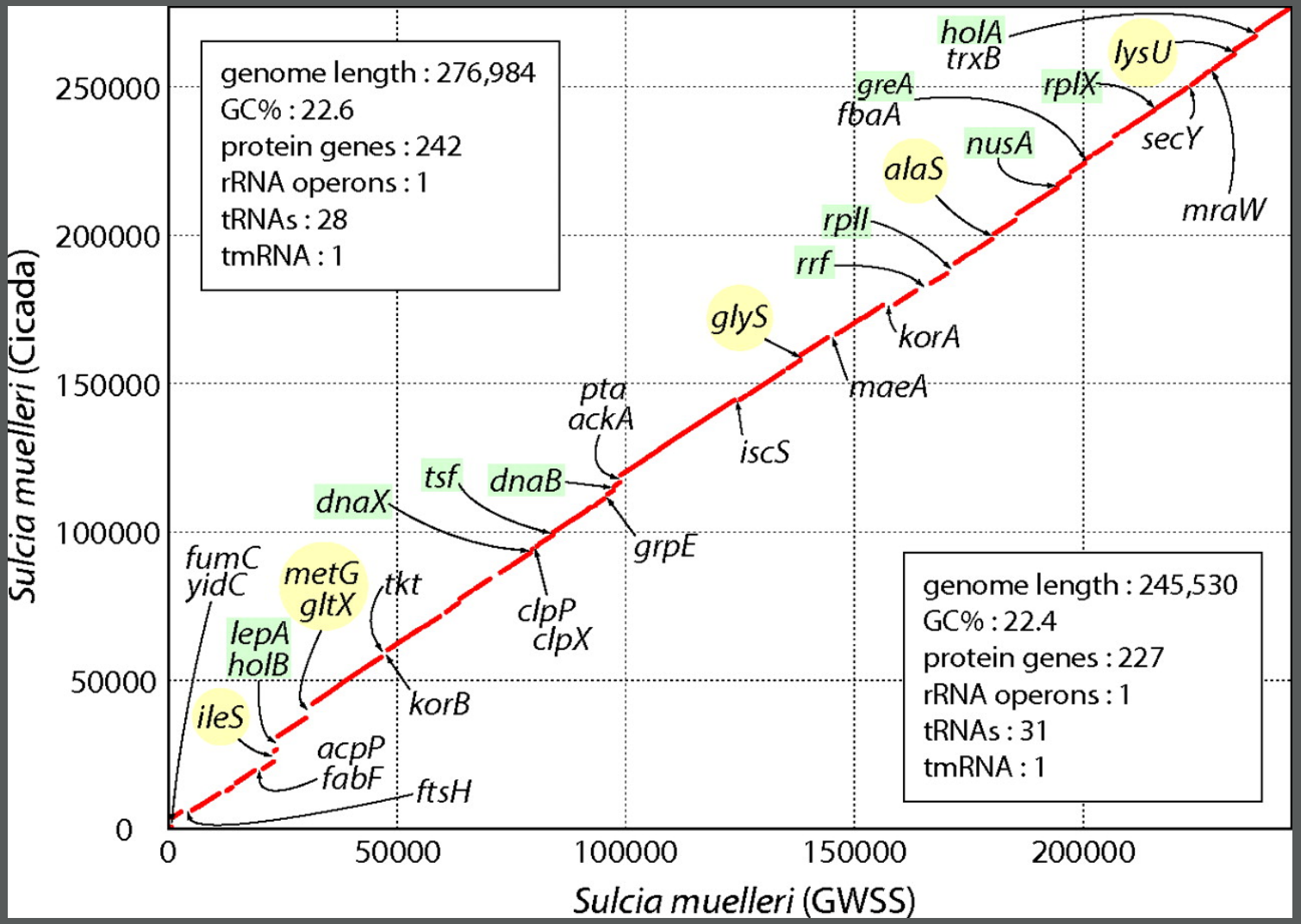
Get a sequence – some pitfalls

- DNA or RNA or protein?
- Which isoform?
- Cross-referencing annotations
  - E.g. Refseq versus ENSG versus…
- Bugs or “known issues”
  - E.g. retrieving a large number of sequences from biomart

Finding orthologs in other species

- There are two main methods to find orthologs for tree building
- 1. Similarity searches
- 2. Synteny (gene order)

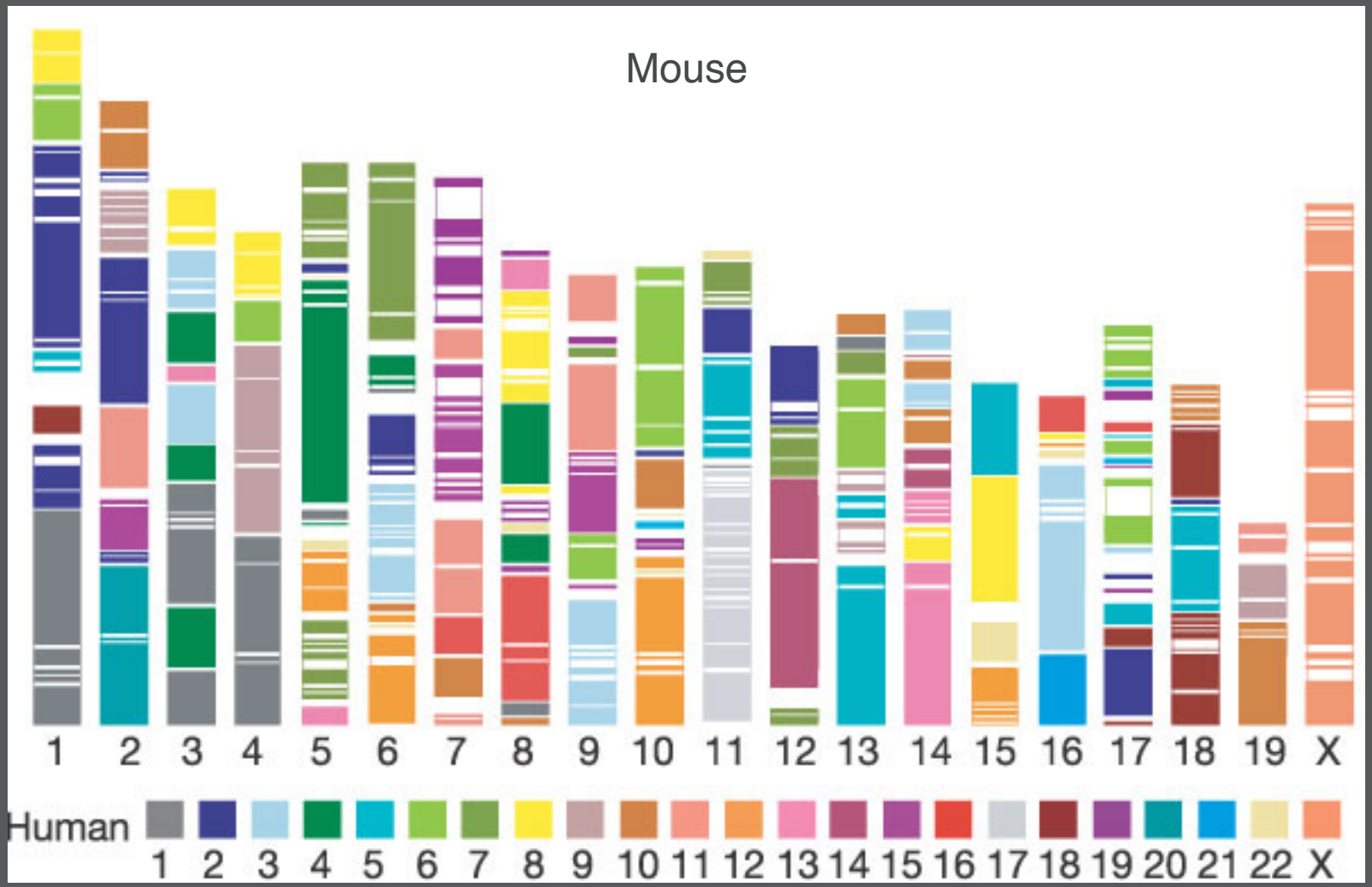
High levels of synteny exist in some clades



Estimated divergence time: >200mya

McCutcheon et al (2009) PNAS

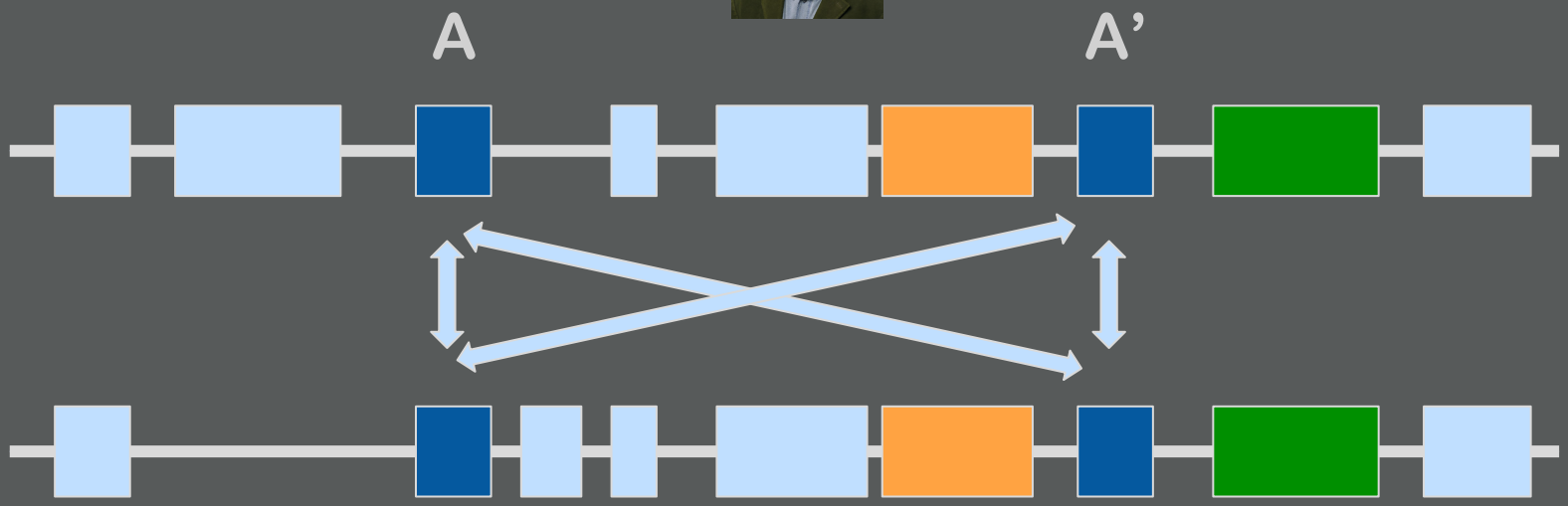
…but not others



Estimated divergence time: 60-80mya

Chinwalla et al (2002) Nature

Synteny can be useful as a tie-breaker



Finding orthologs in other species

1. The lazy way: using existing orthology databases
    - OrthoDB, metaphors, treefam, eggnog, inparanoid,…

2. From scratch: using BLAST

    with/against

- Proteins
- DNA/RNA

Query

e.g. KDM2a

Protein

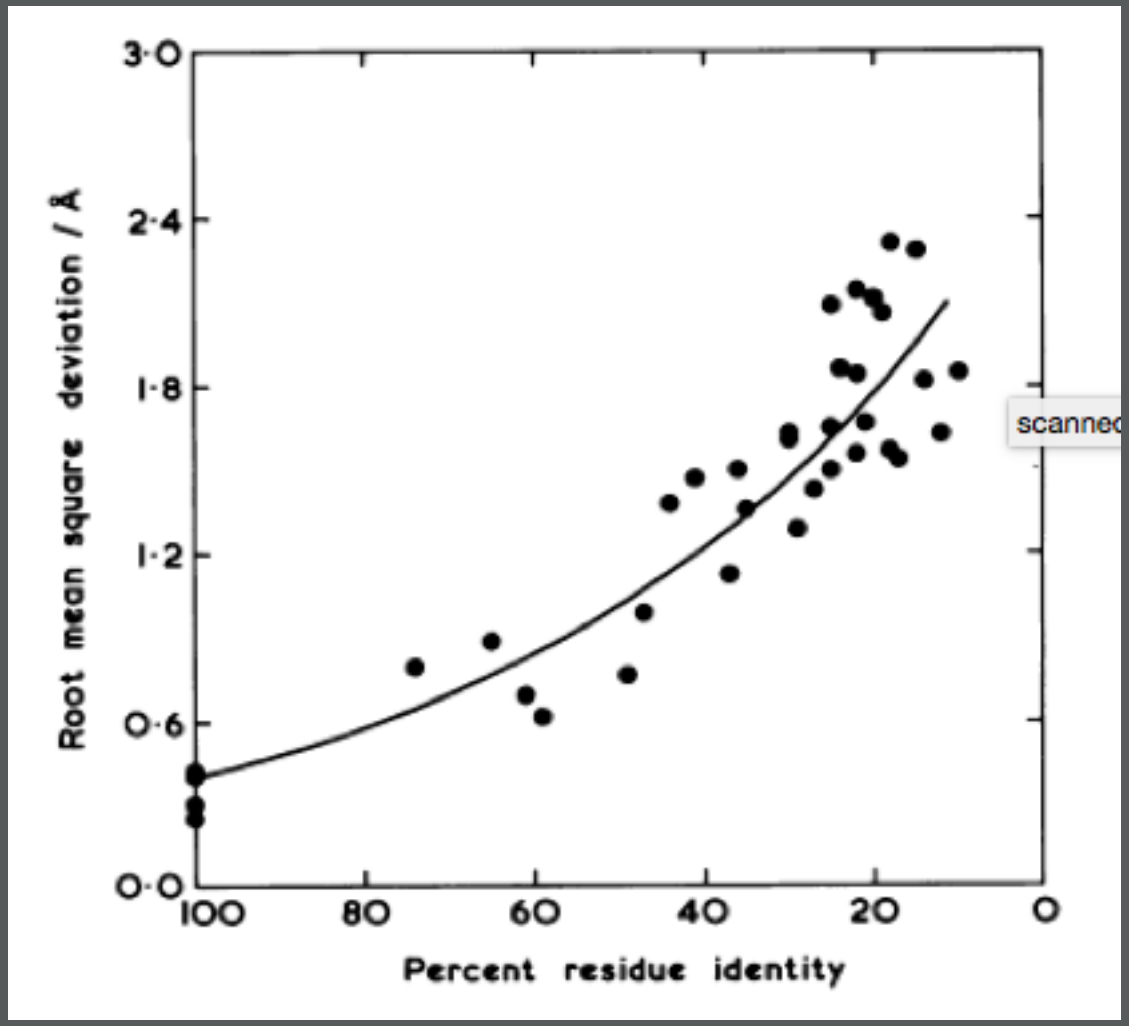
Protein

Protein

Protein

Genome

Conservation: Structure > protein sequence > DNA sequence



Rule of thumb guidance:

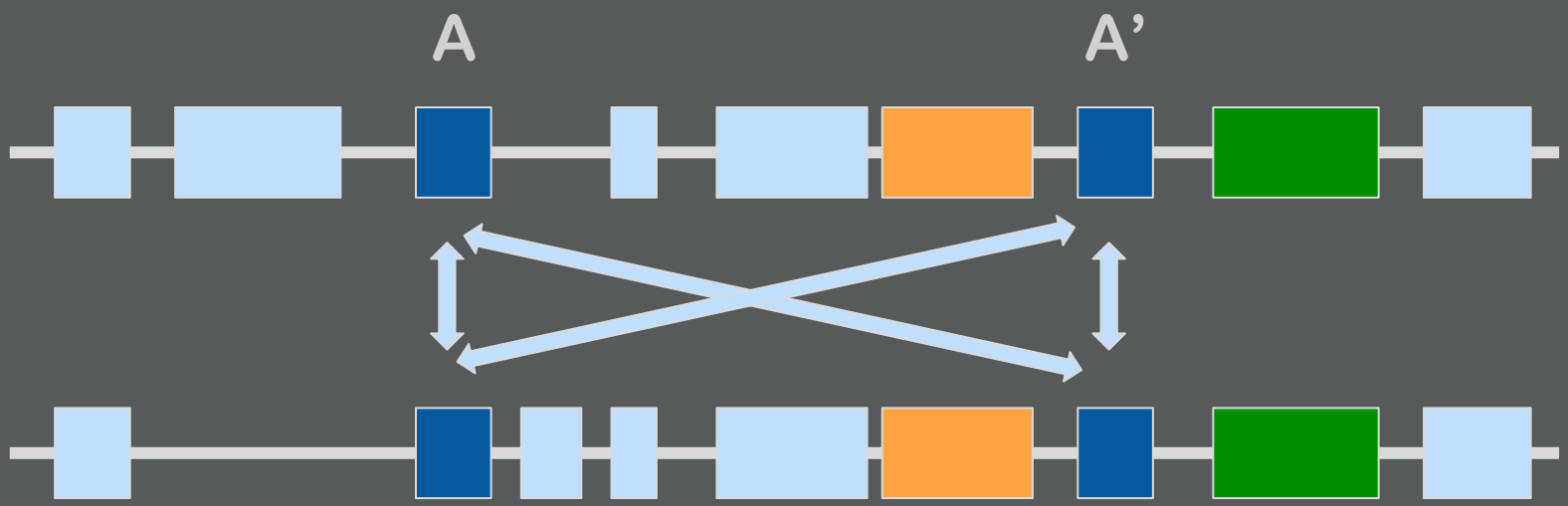
BLAST protein against protein database if you can (blastp)

Then protein against DNA

Then DNA against DNA

Chotia & Lesk (1986) EMBO J

Reciprocal best hits



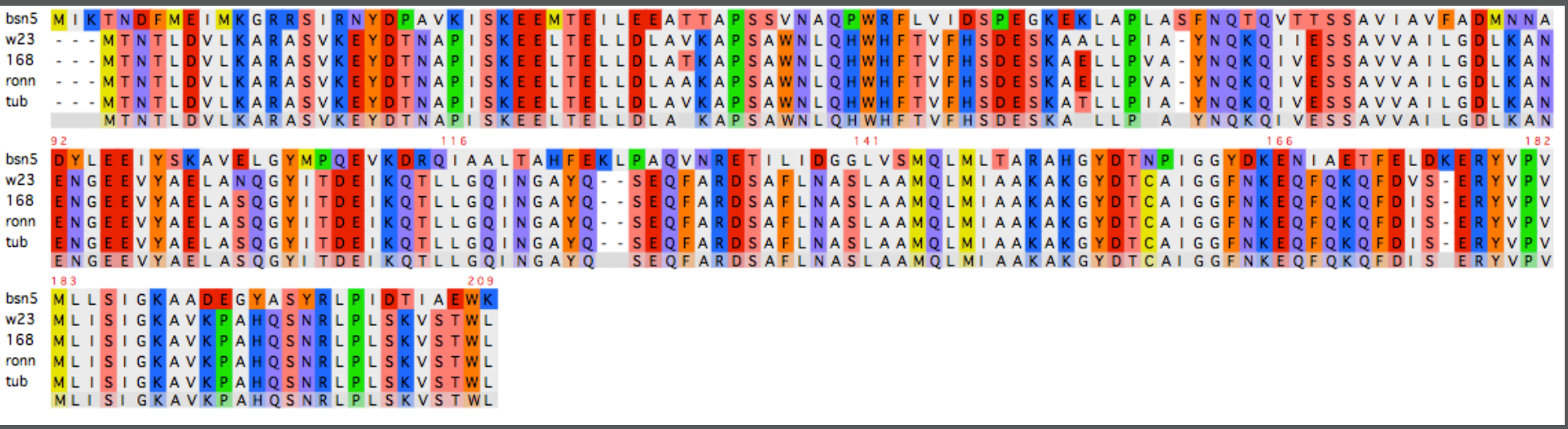
BLAST will find similar sequences and there may be more than one

Reciprocal best hit increases your confidence that you’ve found a true ortholog


much to do
So many choices
So little time!

Align!

Characters in trees are compared by **alignment column**



Columns are the basis to construct (dis)similarity/distance matrices and to reconstruct changes along the tree

Normally, **columns** are considered **independent**

Character in a column can be 1 **nucleotide**, 1 **amino acid** but also 1 **codon**

Horses for courses

Different aligners are good for different things and there are usually speed/accuracy trade-offs

- Single genes/proteins: **MUSCLE, T-Coffee (and it's derivatives)**
- Single genes/proteins but a lot of them: **MAFFT, ClustalOmega**
- Whole genomes: e.g. **progressiveMauve** (bacteria)

ClustalW still in widespread use but not very accurate.

Alignment and processing in R

```
# 1. Load libraries

>require("seqinr")
>require("muscle")
>require("ape")

# 2a. load nucleotide fasta
>cds<-seqinr::read.fasta(file =
"Drosophila_notch_CDS.fa", seqtype ="DNA", as.string=F,
forceDNAtolower=T)

# 2b. Translate

>protein<-getTrans(cds)

# 3. Save protein file

>seqinr::write.fasta(protein, names(cds),
file="Drosophila_notch_protein.fa", open="w")
```

Alignment and processing in R

```
# 4. Align

>muscle::muscle("Drosophila_notch_protein.fa",
out="Drosophila_notch_protein_aligned.fa")
>detach("package:muscle", unload=TRUE)

# 5. reverse translate
>reverse.align("Drosophila_notch_CDS.fa",
"Drosophila_notch_protein_aligned.fa", input.format =
"fasta", "Drosophila_notch_CDS_aligned.fa")
```

Let's have a look at the alignment → jalview.org

Questions:

1. Are the alignments any good?
2. Which is better?
3. Any areas that are more different than others?
4. What are the lengths of the two alignments?

Building a simple distance tree

```
# 6. Read the alignment

>protein_alignment <-
read.alignment("Drosophila_notch_protein_aligned.fa",for
mat="fasta")

# 7. Build a distance matrix
#no evolutionary model assumed here!!
#to see what's under the hood, see ?dist.alignment

>protein_dist<-dist.alignment(protein_alignment)

# 8. Make a tree (finally!)

>protein_tree <- nj(protein_dist)
#nj stands for neighbour-joining

# 9. Have a look
>plot(protein_tree)
```

Building a simple distance tree

1. Try doing the same for nucleotide alignment

```
# 10. Build nucleotide tree

>nucleotide_alignment <-
read.alignment("Drosophila_notch_CDS_aligned.fa",format=
"fasta")
>nucleotide_dist<-dist.alignment(nucleotide_alignment)
>nucleotide_tree <- nj(nucleotide_dist)
>plot(nucleotide_tree)

#Explore the phylo.plot function for more rendering
options

#Writing your trees:
# a) graphical: e.g. pdf()
# b) in a text format (Newick): e.g. write.tree()
```

Bootstrapping

Questions:

1. Are there any differences between the nucleotide and protein trees?
2. Which one is better?

```
# 11. Bootstrapping

boot_nuc<- boot.phylo(nucleotide_tree,
as.matrix.alignment(nucleotide_alignment), function(x)
nj(dist.dna(as.DNAbin(x))))

# plot the tree:
>plot.phylo(nucleotide_tree,type="p”)
>nodelabels(boot_nuc,cex=0.7)
>nucleotide_tree$node.label <- boot_nuc
```

Basic principle: Sample with replacement from the alignment columns (Felsenstein method)

Building rooted trees

Questions:

1. Are the trees rooted or unrooted?

2. Why are they unrooted?

We know from the Drosophila phylogeny that *D. pseudoobscura* is the outgroup to the other species.

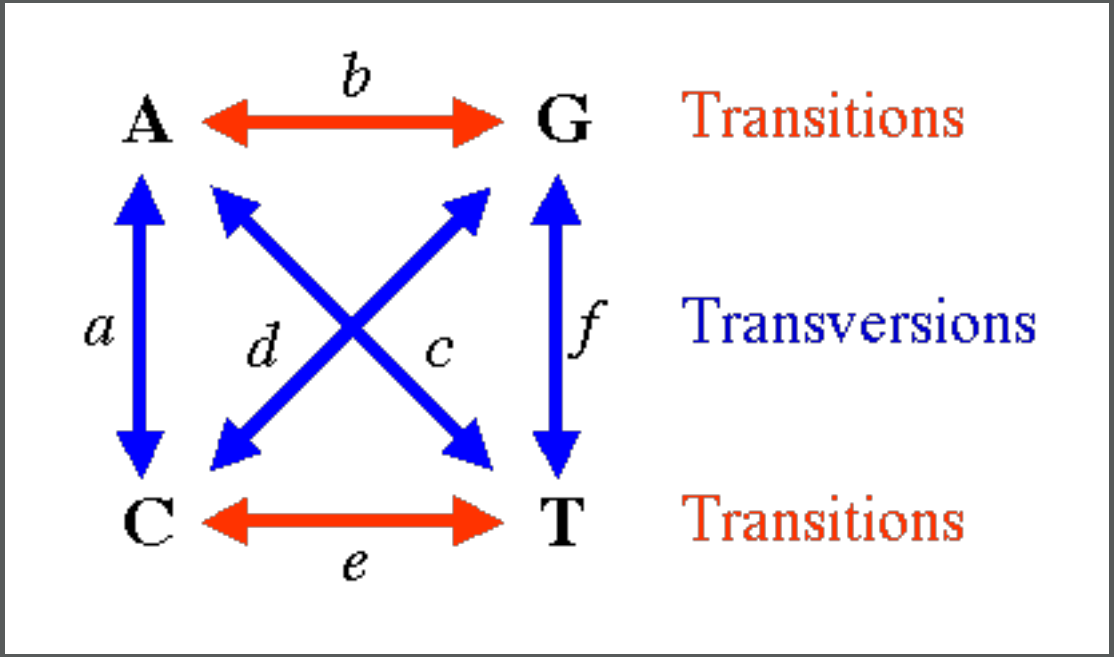
To build a rooted tree, use the root() function:

```
# 12. Building a rooted tree

>rooted_protein_tree <- root(protein_tree, "GA28528-
PA_Dpse", r=TRUE)
>rooted_nucleotide_tree <- root(nucleotide_tree,
"GA28528-PA_Dpse", r=TRUE)
```

Building fancier trees

Most tree building models make some assumptions about the nature of nucleotide changes:



The simplest model: Jukes-Cantor (JC69)

Probability of one nucleotide changing into another is the same

$$M_{JC} = \begin{array}{c|cccc} & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \\ \mathrm{A} & 1-\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \mathrm{C} & \alpha/3 & 1-\alpha & \alpha/3 & \alpha/3 \\ \mathrm{G} & \alpha/3 & \alpha/3 & 1-\alpha & \alpha/3 \\ \mathrm{T} & \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha \end{array}$$

A bit more complex: Kimura's two-parameter model (K80)

There are different probabilities for transitions and transversions

$$M_{K2P} = \begin{array}{c|cccc} & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \\ \mathrm{A} & 1-\alpha-\beta & \beta & \alpha & \beta \\ \mathrm{C} & \beta & 1-\alpha-\beta & \beta & \alpha \\ \mathrm{G} & \alpha & \beta & 1-\alpha-\beta & \beta \\ \mathrm{T} & \beta & \alpha & \beta & 1-\alpha-\beta \end{array}$$

Building a fancier tree in R

```
# 11. Building a tree with a specific substitution model

>nucleotide_dist_fancy<-
dist.dna(as.DNAbin(nucleotide_alignment), model="F84")
>nucleotide_tree_fancy <- nj(nucleotide_dist_fancy)
```

More specialized tree building programs

- PhyML (good online pipeline at phylogeny.lirmm.fr)
- RaXML
- MrBayes
- …

have a large number of models to choose from. It's often not obvious which one is best. You can choose based on prior knowledge, stick with the default, or let an algorithm choose for you (ModelTest)

Making a tree – basic steps

1. **Get** a sequence

2. **Find** similar sequences, hoping that they are

   - Related by descent (**orthologs**) or
   - Related by duplication (**paralogs**) or
   - Related by some other process (convergent evolution?)
   - Not related by chance

3. **Align** them

4. **Tidy** up the alignment

5. **Make** a tree! → **Evaluate** the tree

   ↓

   Use the tree