

# Introductions

Monday 5th May 2014

EMBO Practical Course on Computational Molecular Evolution

Institute of Marine Biology, Biotechnology and Aquaculture  
(IMBBC), Hellenic Center for Marine Research) HCMR

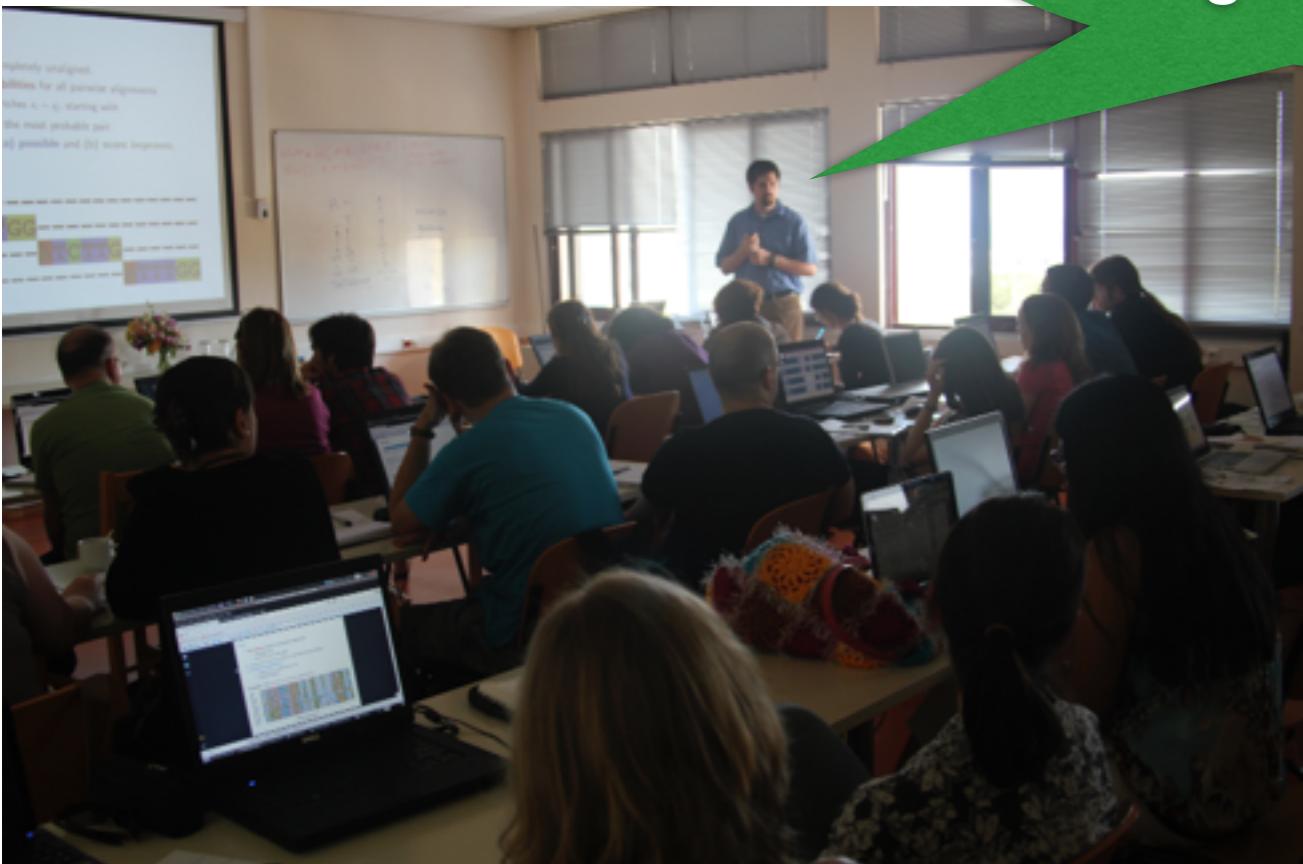
Heraklion, Greece

# What You Get From a Course?

# I. knowledge & understanding

# Knowledge & Understanding

## Presentations and demos



insightful comments about  
alignments and trees...



## Practical exercises and discussions

## **2. useful professional relationships**

**a. with other trainees**

**b. with trainers**



Aidan Budd, EMBL Heidelberg

**both (I estimate, on average) equally valuable**

**so we start with an activity...**

**to help us get to know each other...**

# speed dating

# Speed Dating:Aims

---

- facilitate the (for some of us) awkward “meeting/introducing ourselves to someone for the first time” by doing lots of them in a short time
  - hopefully makes it easier to start talking with each other later in the course
- quickly find people you have things in common with/get on with

# Speed Dating: Format

---

- meet other participants in many 1:1 chats
- tell each other
  - names
  - where you work
  - research topics
  - something surprising about you (e.g. I have three nationalities)
  - try and find someone you know or somewhere you've been that you have in common

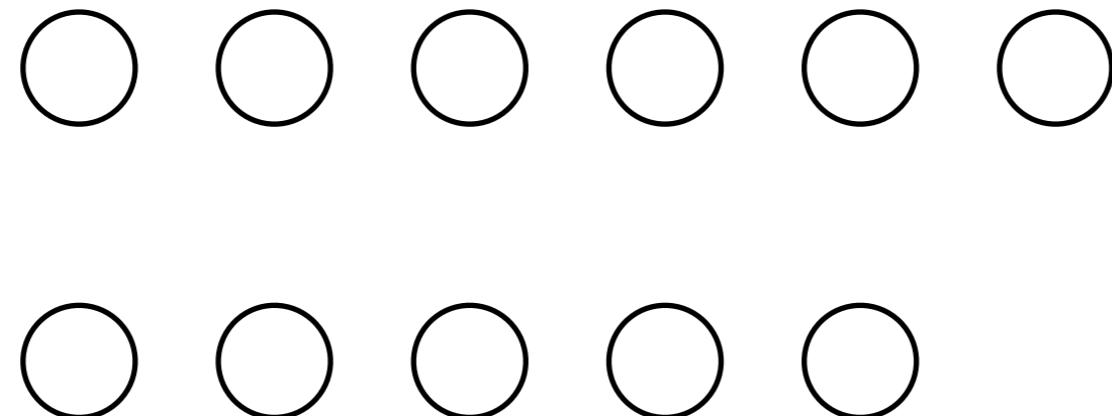
# Speed Dating: Format

---

Stand, awkwardly, in two rows

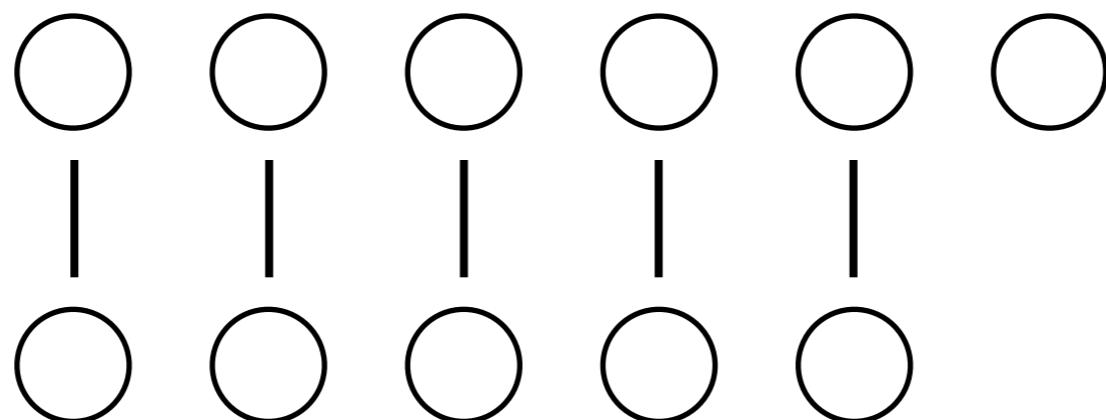
Face one person in the other row

If there's an odd number of you, one person stands alone at one end

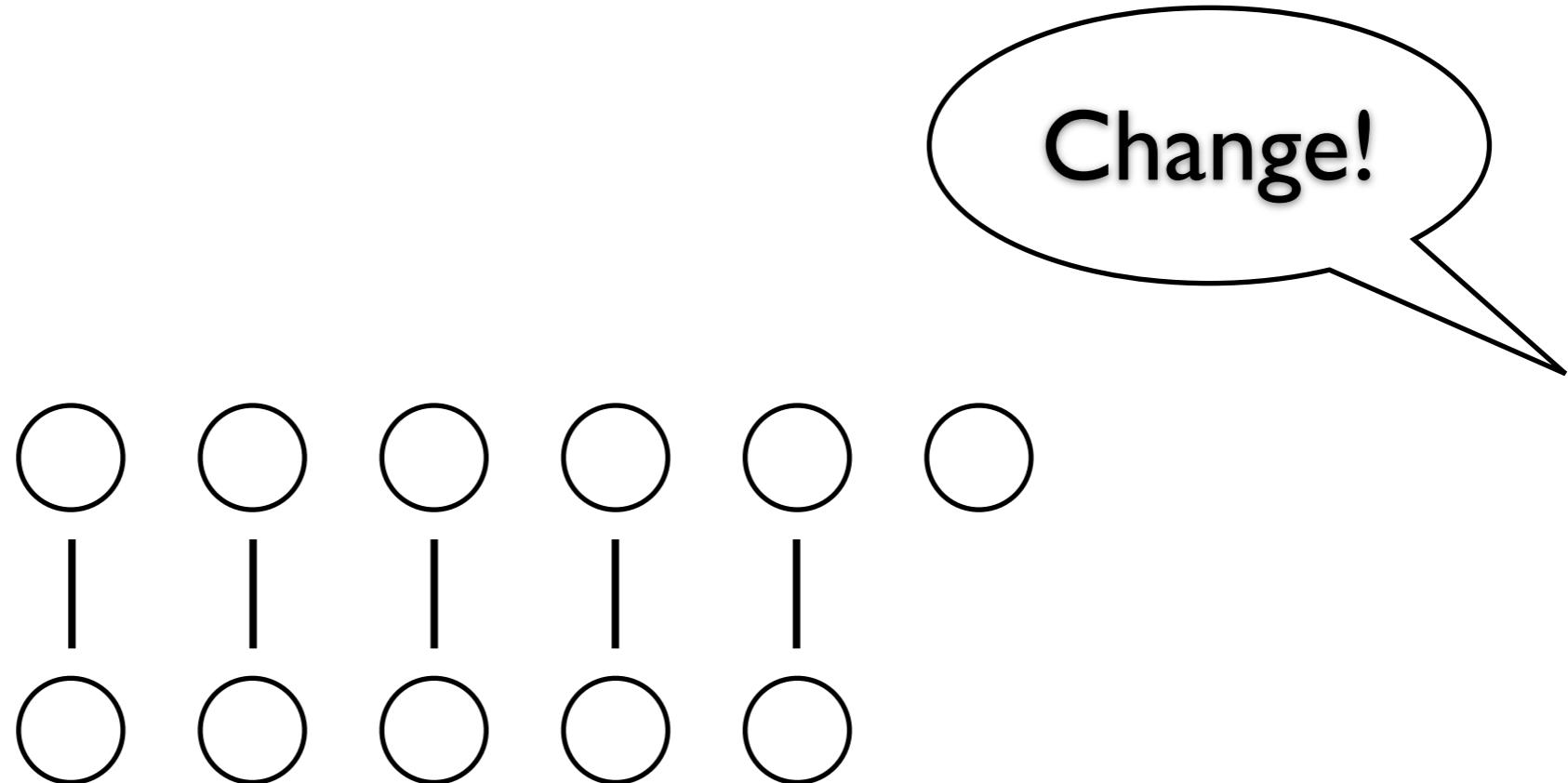


# Speed Dating: Format

**Chat!**

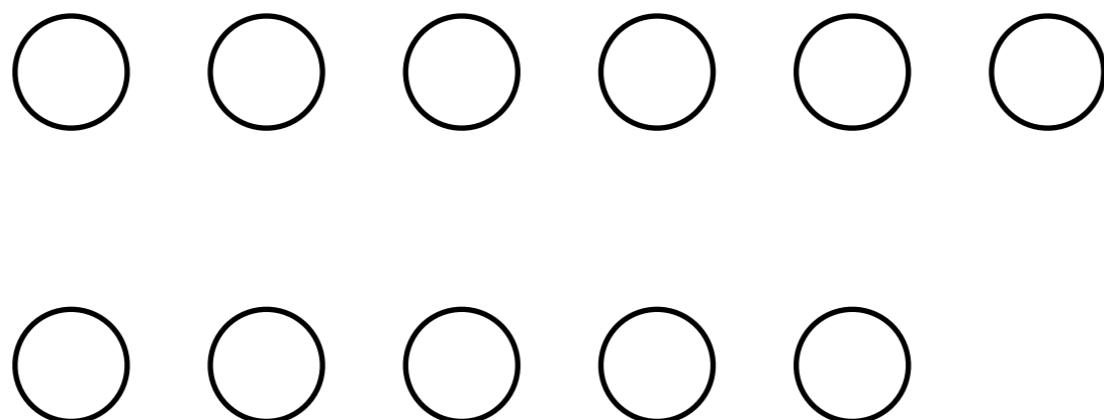


# Speed Dating: Format



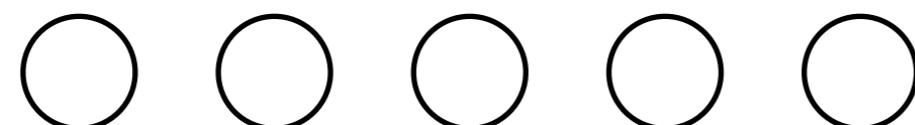
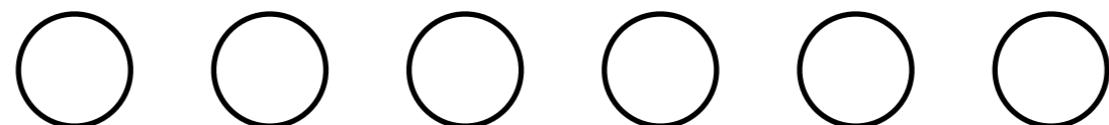
# Speed Dating: Format

---



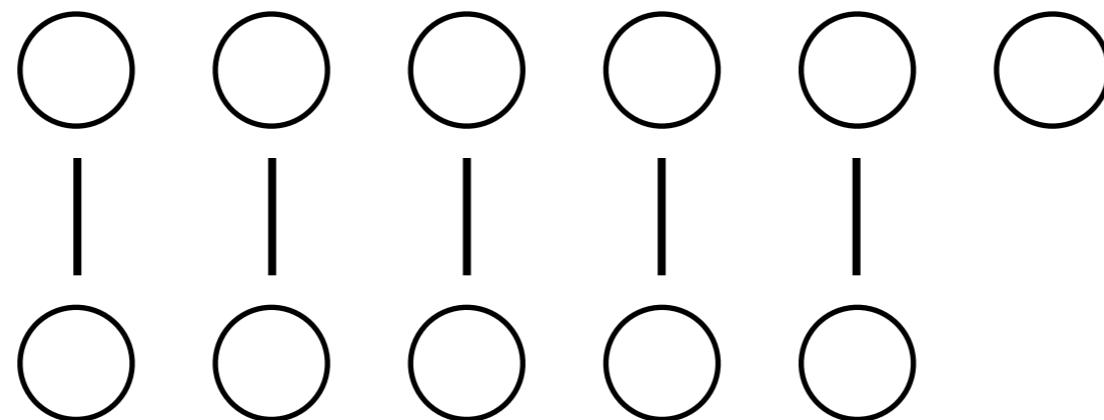
# Speed Dating: Format

---



# Speed Dating: Format

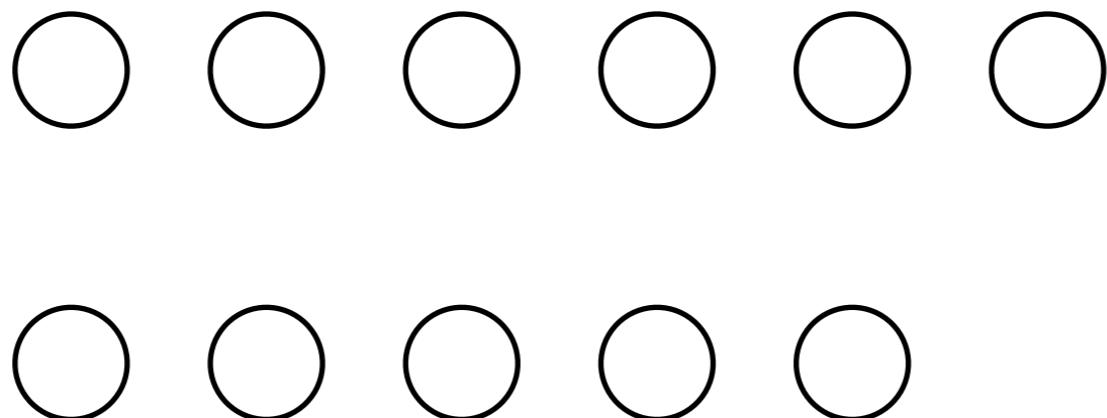
**Chat!**



and repeat until you've met everyone in the other row...

# Speed Dating: Format

---

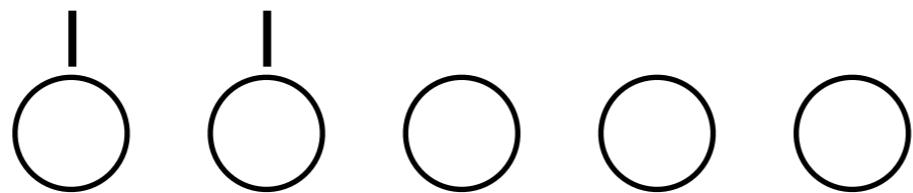
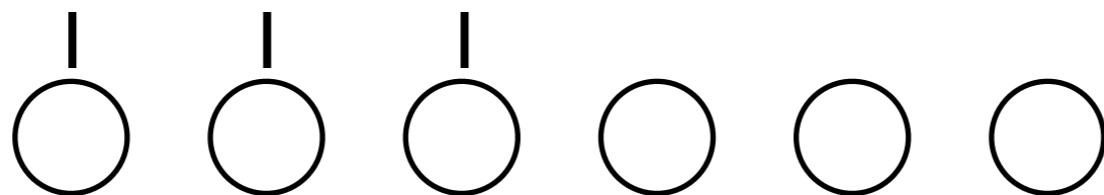


then split each row into two new rows

# Speed Dating: Format

---

**Chat!**



make two new rows

and start again with the chat...

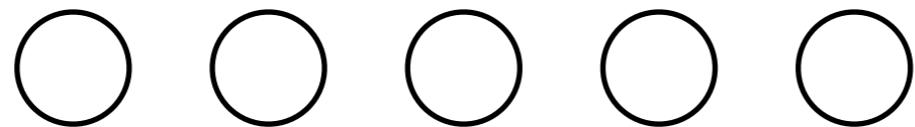
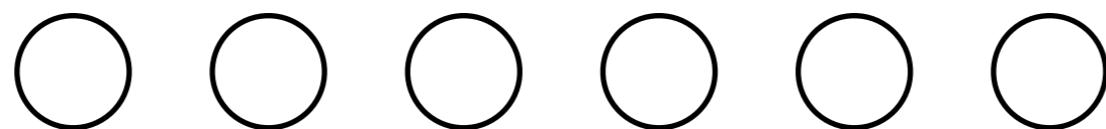
# Speed Dating: Format

---

So... go outside now and form the initial two awkward rows

Face one person in the other row

If there's an odd number of you, one person stands alone at one end



# Session HTML pages

download zip/git clone from:

**<https://github.com/aidanbudd/crete2014>**

“homepage” (links to exercises, instructions, presentations)  
**homepageInterpretingPhylogeniesCrete2014.html**

exercises

**interpretingPhylogeniesCrete2014.html**

# Interpreting Molecular Phylogenetic Trees

Monday 5th May 2014

EMBO Practical Course on Computational Molecular Evolution

Institute of Marine Biology, Biotechnology and Aquaculture  
(IMBBC), Hellenic Center for Marine Research) HCMR

Heraklion, Greece

Aidan Budd  
EMBL Heidelberg, Germany

Laura Emery  
EMBL-EBI, Hinxton, UK

Sarah Parks  
EMBL-EBI, Hinxton, UK

# Introductions and Aims

# Phylogenetics

---

'phylogenetics' is of Greek origin from the term

- **phyle/phylon** (φυλή/φῦλον), meaning "tribe, race"
- **genetikos** (γενετικός), meaning "relative to birth" from **genesis** (γένεσις, "birth").

(Wikipedia)

Give me a word, any word, and I show you how the root of that word is Greek

OK, Mr Portokalos, how about the word "Kimono"?

Hah. Kimono.

Hah. Of course. Kimono is come from the Greek word "Himona" which mean winter...

<http://www.youtube.com/watch?v=VL9whwwTK6I>

<http://www.youtube.com/watch?v=2ALrm3nDGXI>

(My Big Fat Greek Wedding)

# Aims

---

**This session aims to provide:**

An introductory overview of terminology, concepts, and tools associated with phylogenetic trees

An overview of how different stages and tools used in an analysis fit together

# Before we start

---

- Mixture of presentations, demonstrations, discussions, and exercises
- Working in pairs is encouraged
- Please ask questions at any point

# How Do We Interpret Molecular Phylogenetic Trees?

## An Example

# Interpreting Molecular Phylogenetic Trees: An Example

---

Begin by interpreting together a tree from a published article

Tree chosen for its use in a way that is clearly generally interesting/important to everyone i.e. all of you

We're doing this to highlight and explore common features and possible problems we encounter when interpreting trees

# Interpreting Molecular Phylogenetic Trees: An Example

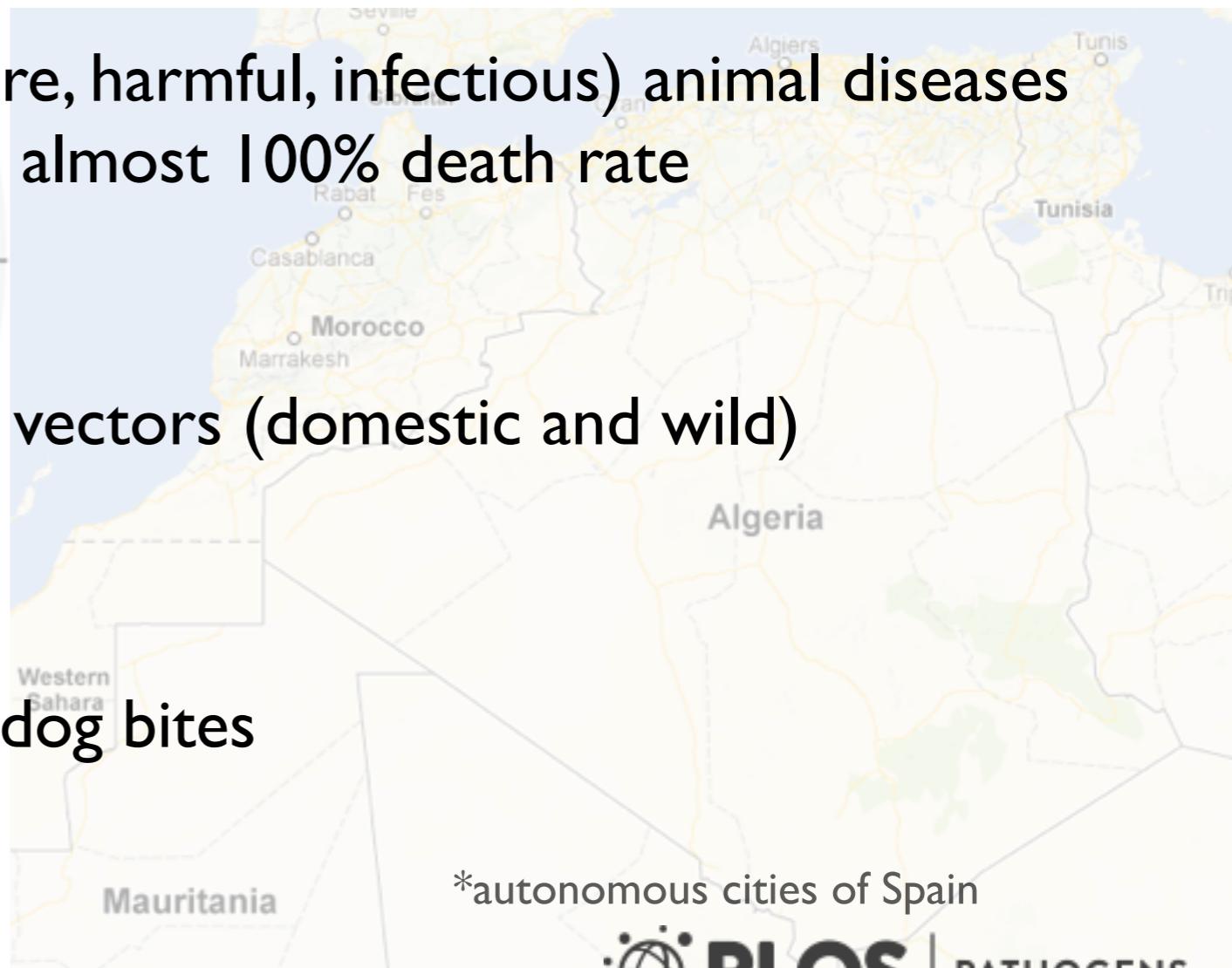
Study aiming to identify factors contributing to pattern and rate of transmission ("transmission dynamics") of rabies virus in North Africa

One of the world's most virulent (severe, harmful, infectious) animal diseases  
Once disease presents in humans almost 100% death rate

99% of human infections linked to dog vectors (domestic and wild)

Transmission via saliva, particularly via dog bites

Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.



\*autonomous cities of Spain

# Interpreting Molecular Phylogenetic Trees: An Example

Rabies is a major public health problem - yearly, worldwide:

55,000 deaths

15,000,000 doses of anti-rabies post-exposure prophylaxis administered

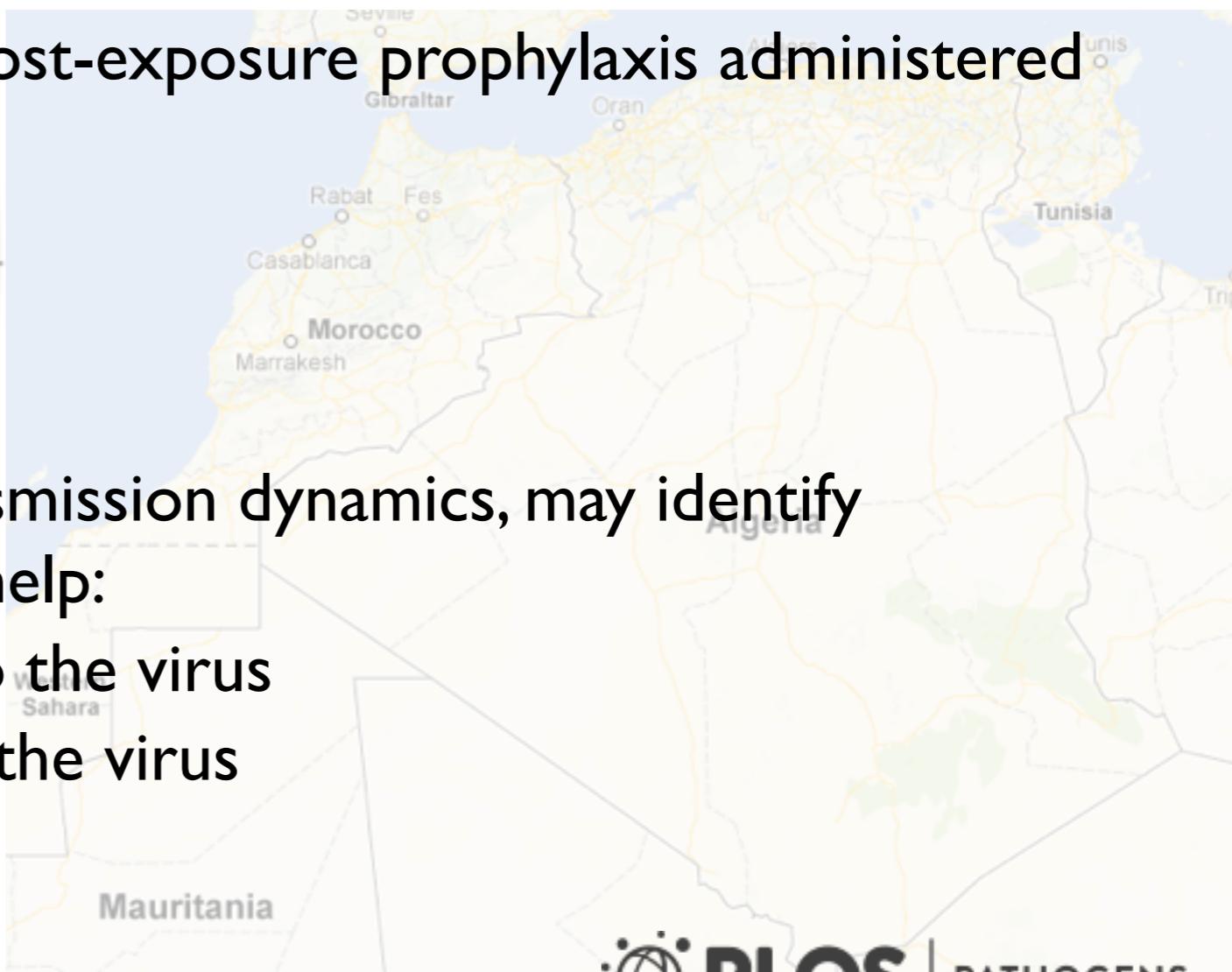
Therefore rabies:

causes significant human suffering  
is a major economic burden

Identifying factors contributing to transmission dynamics, may identify public health interventions that could help:

reduce human suffering related to the virus  
reduce economic cost/burden of the virus

Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.



# Interpreting Molecular Phylogenetic Trees: An Example

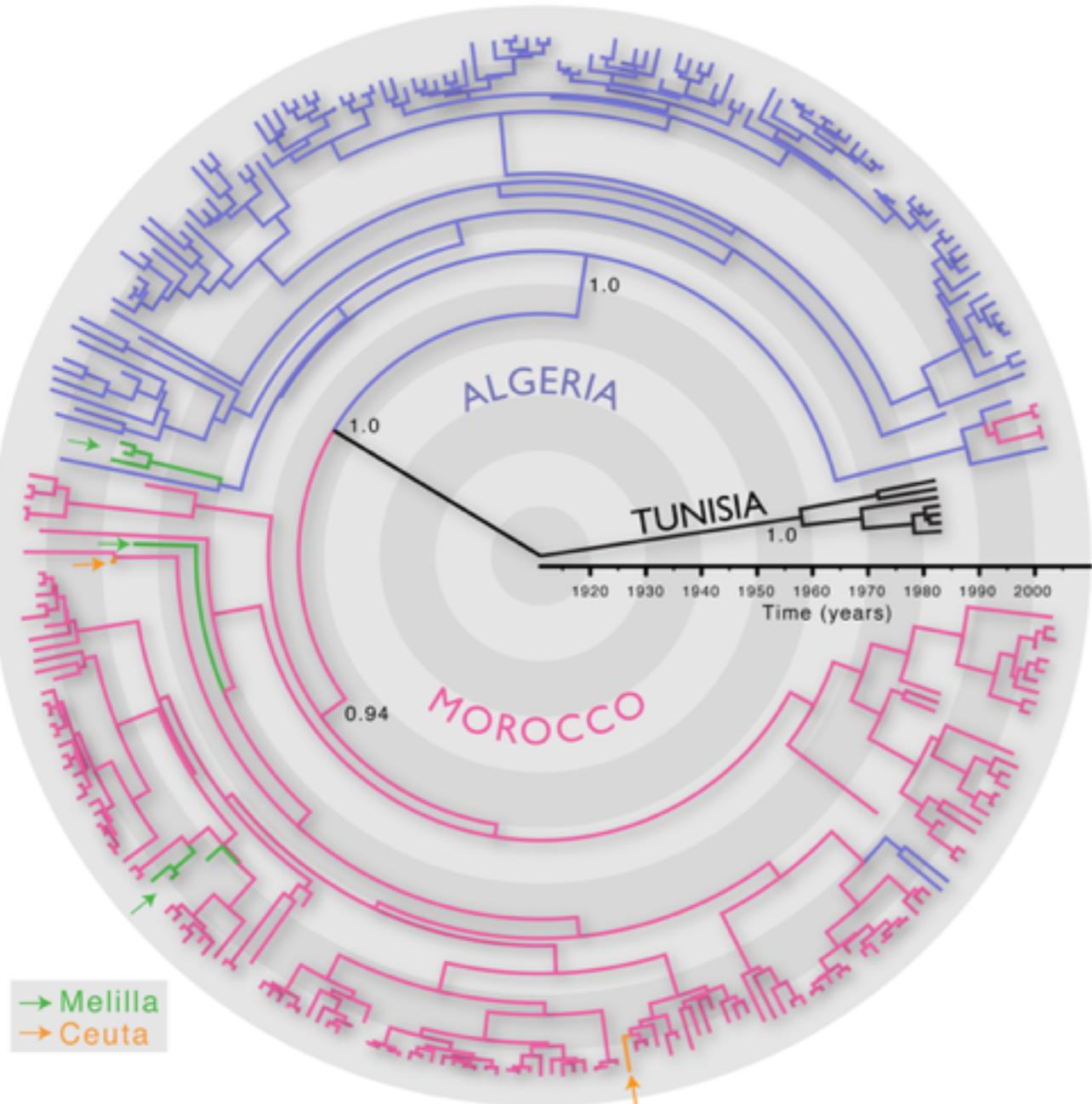
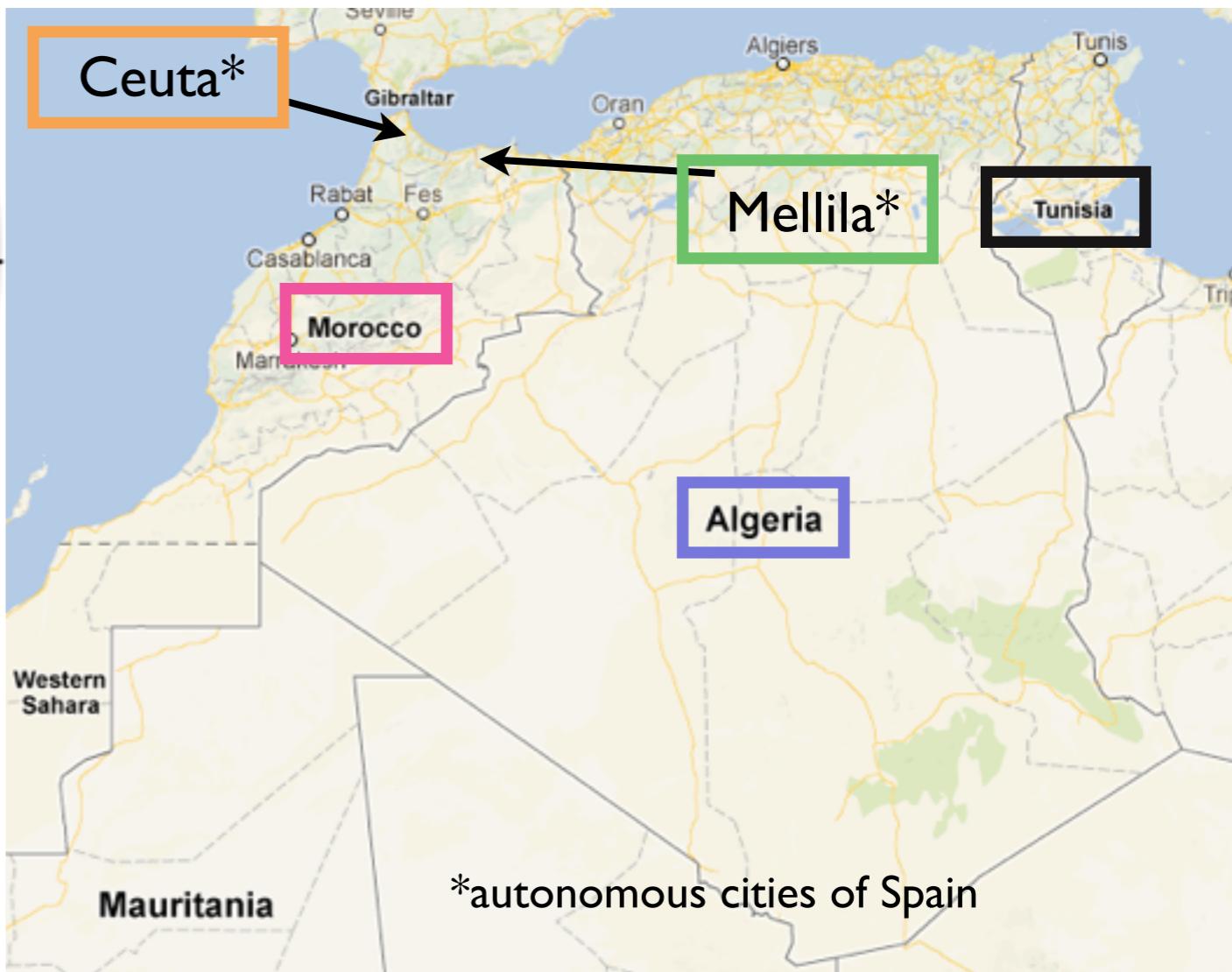


Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.

Phylogeny of rabies virus sampled from North African dogs

Branches coloured according to measured or inferred geographical location



# Interpreting Molecular Phylogenetic Trees: An Example

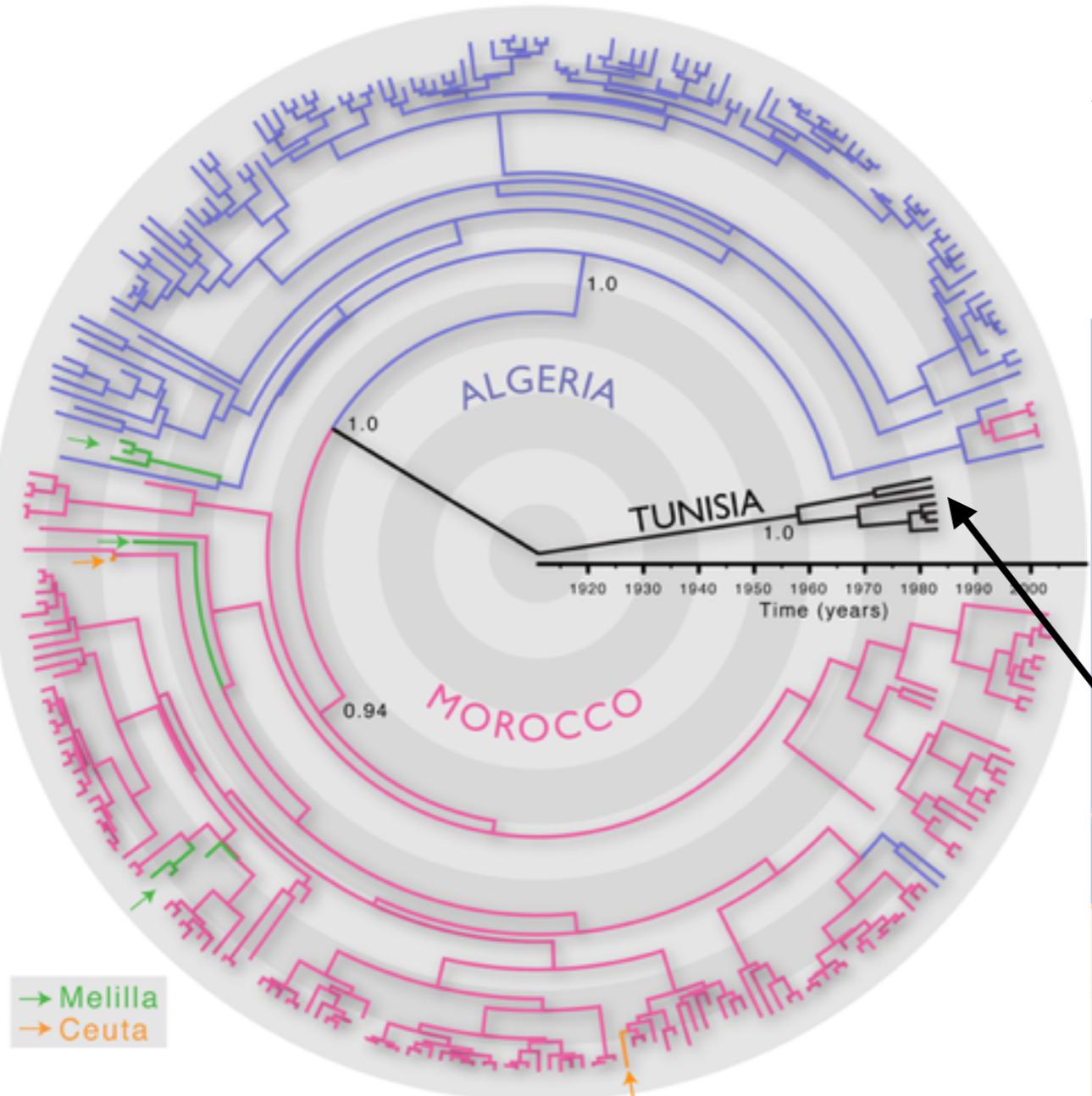
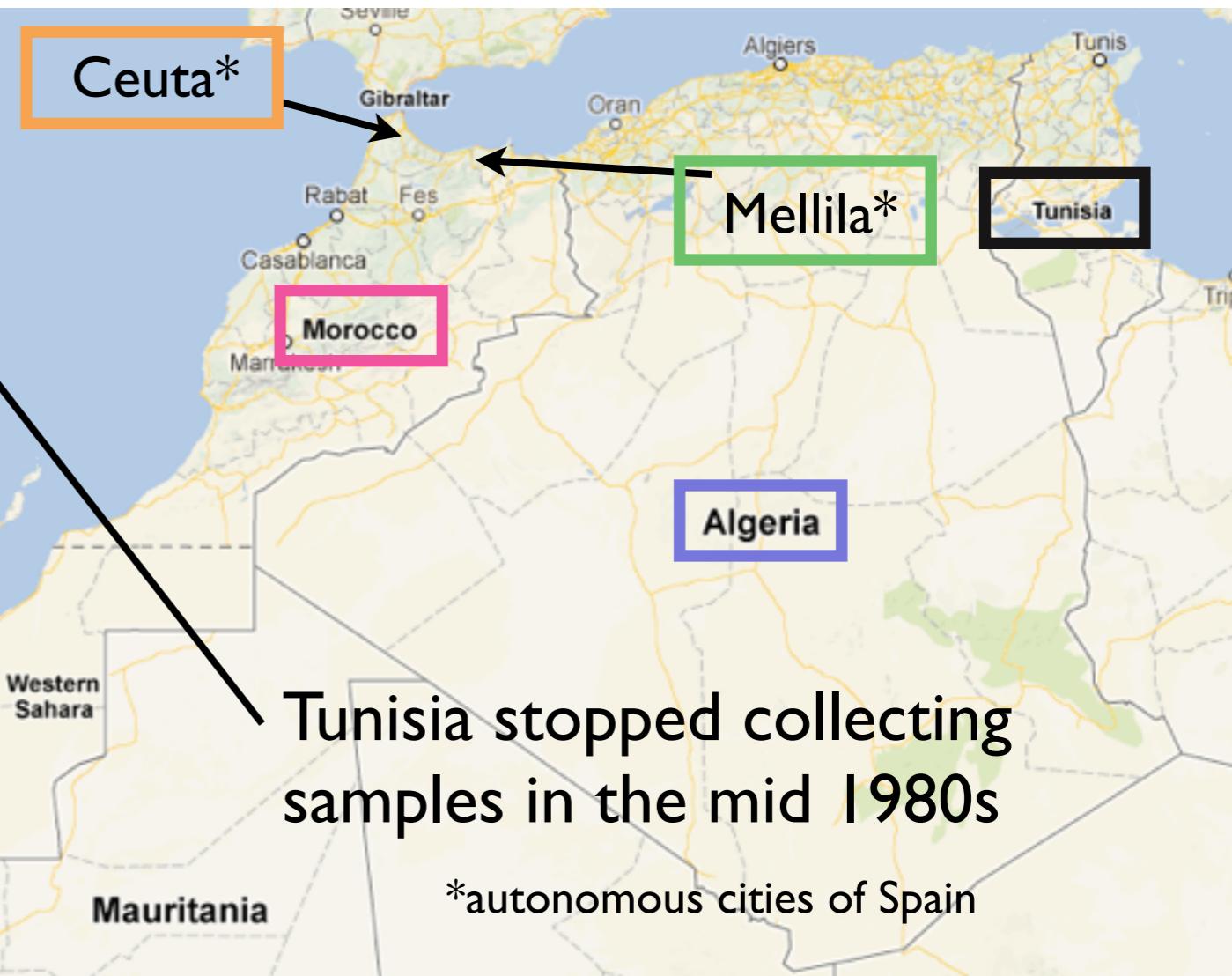


Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.

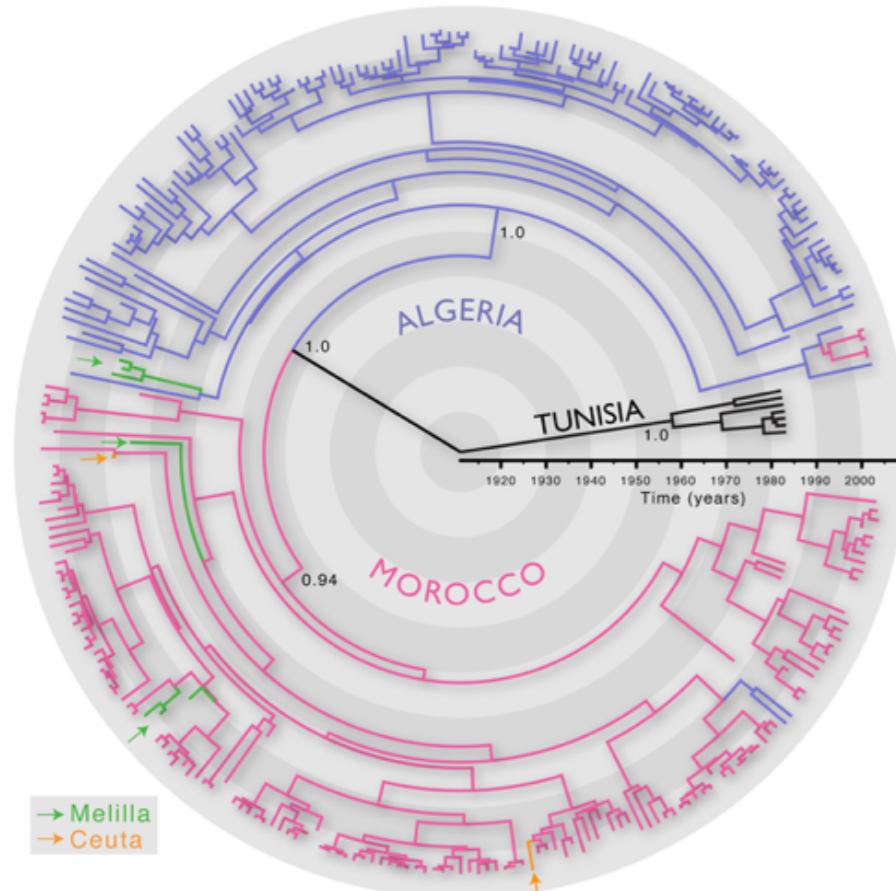
Phylogeny of rabies virus sampled from North African dogs

Branches coloured according to measured or inferred geographical location



\*autonomous cities of Spain

# Interpreting Molecular Phylogenetic Trees: An Example



Does observing this tree make you consider it  
A. more probable  
B. less probable  
that human activity significantly influences the  
dynamics of rabies virus transmission between dogs?

Try and decide, firstly, on your own, without  
discussing with your neighbours!

Then we'll take a vote to see what you think

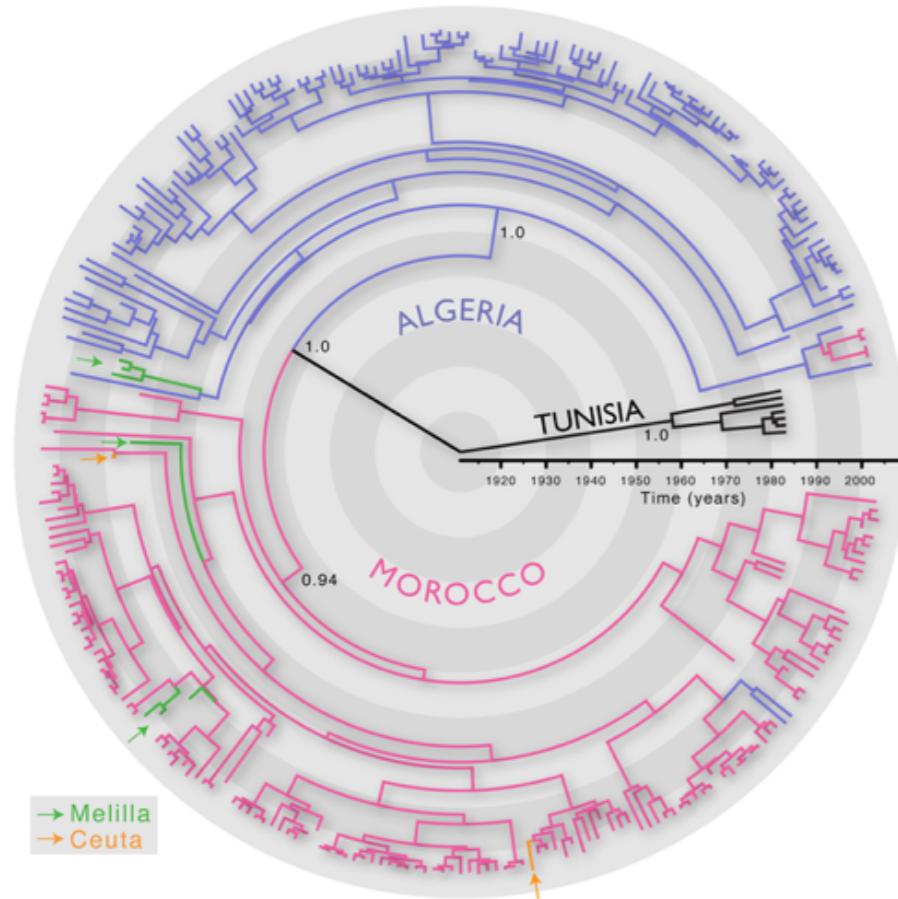
Then discuss the question with your neighbours

Feel uncomfortable that you don't know enough  
about the study/data to decide?

Then make (and keep a note of!) reasonable/  
possible/plausible assumptions about what you don't  
know, then answer assuming these are correct

Don't move to next slide yet!

# Interpreting Molecular Phylogenetic Trees: An Example



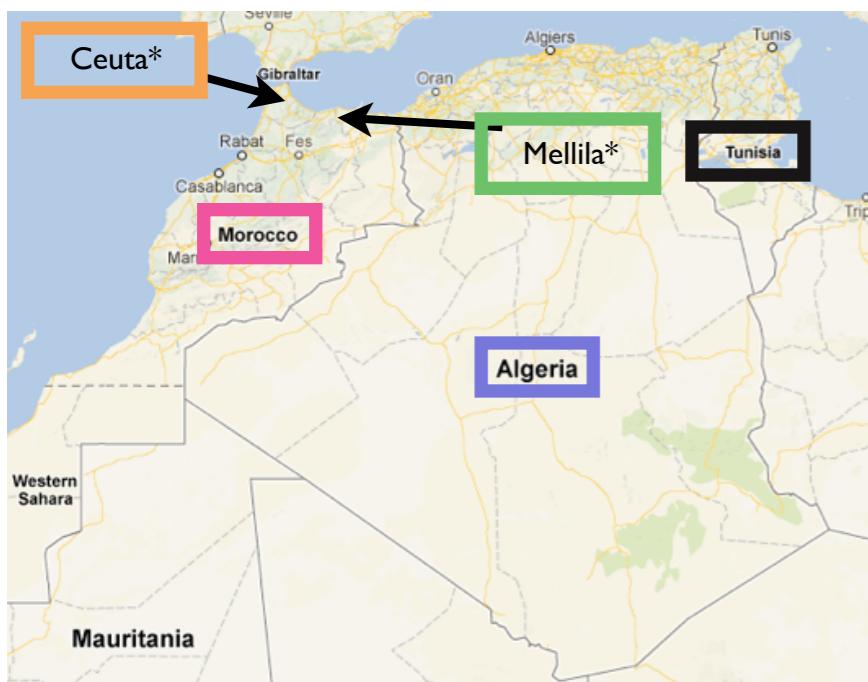
Does observing this tree make you consider it  
A. more probable  
B. less probable  
that human activity significantly influences the  
dynamics of rabies virus transmission between dogs?

On the basis of this tree (and several other analyses)  
the authors conclude that the data supports a tree that  
makes it

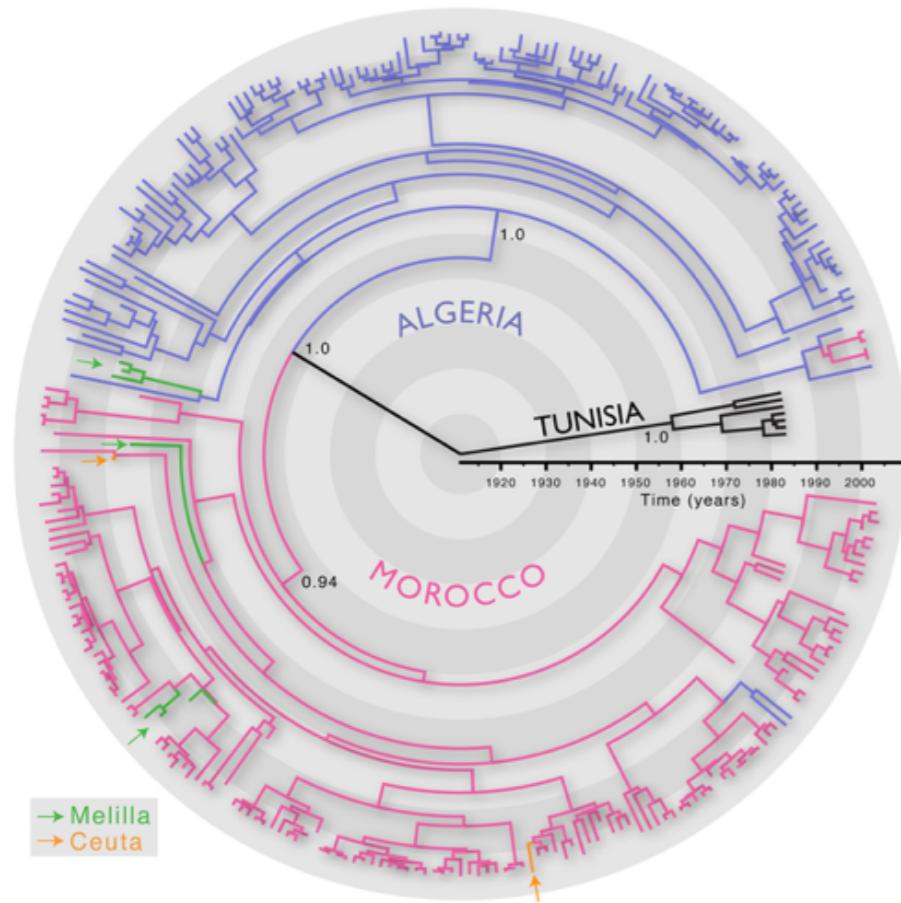
A. more probable  
that human activity significantly influences the dynamics  
of rabies virus transmission between dogs

seen in the rarity of virus transmission across political (i.e.  
at least partially human-activity imposed) borders -

Obvious important implications for public health policy  
e.g. suggests that restricting/regulating dog transport may  
reduce impact of the virus



# Interpreting Molecular Phylogenetic Trees: An Example



this exercise aimed to highlight that:

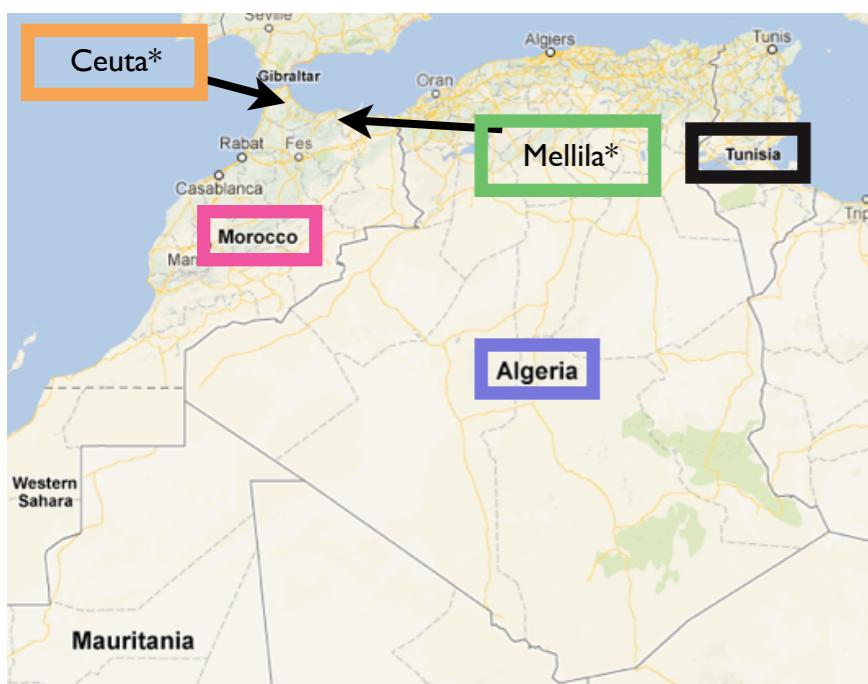
we have to make assumptions/use models to interpret the data - examples of assumptions that could be made while interpreting this tree:

- topology of the tree is correct
- inference of taxon location is correct
- natural geographic (mountains, deserts) do not influence gene flow for the virus

it's useful/important to be aware of what these are and to state them

it's important to present information (the tree, the sample location) in a way that makes the conclusions you want to draw from the analysis clear/obvious

- it's useful spending time thinking/practicing changing how trees are presented/displayed

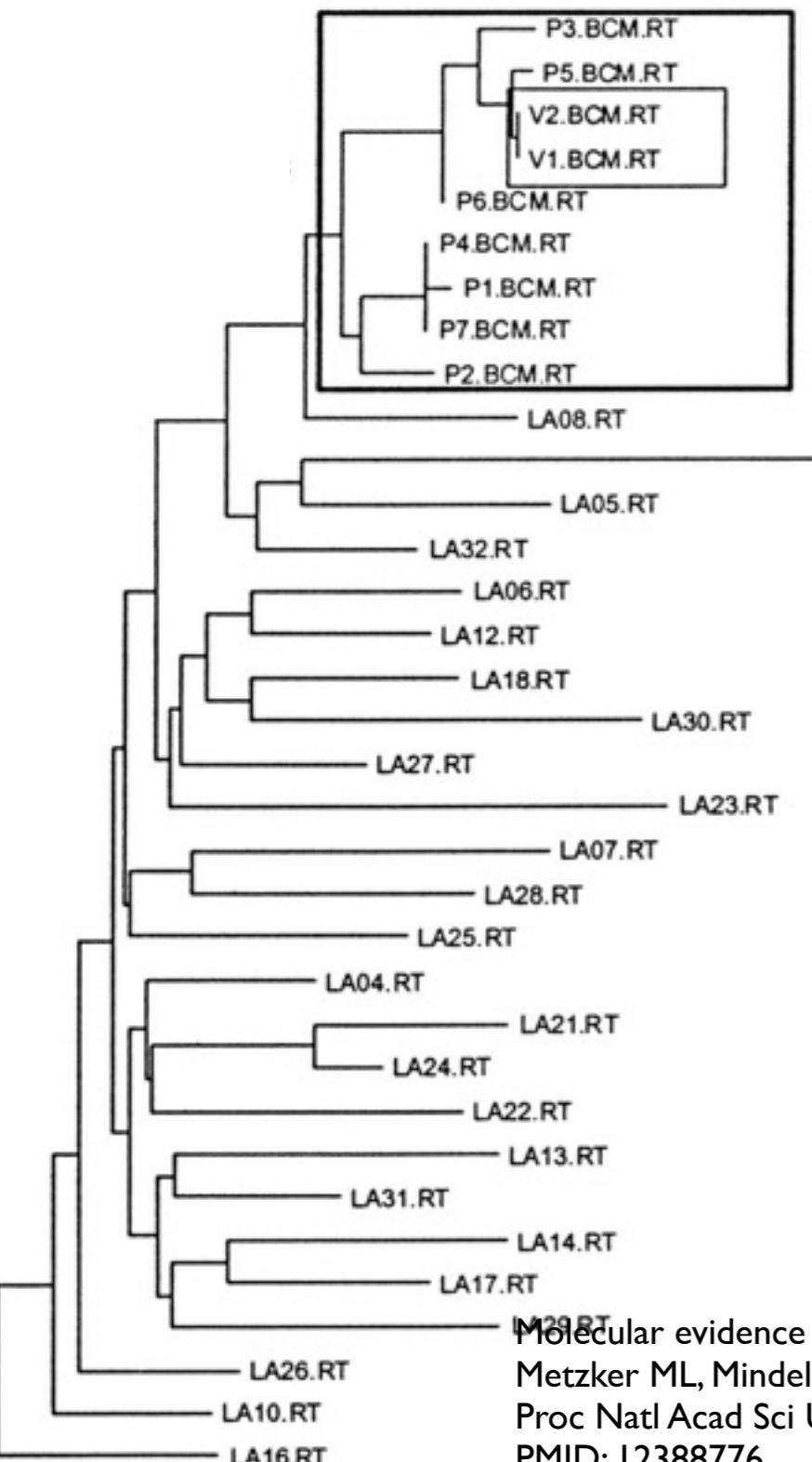


# Other Applications of Phylogenetics

---

- Epidemiology
- Forensics
- Selecting conservation targets
- Monitoring trade in illegal organisms
- Bioinformatics tools - in particular:
  - building MSAs
  - predicting function
- Basic evolutionary research
  - characterising processes of evolutionary transformation
  - estimating patterns of transmission of genetic material

# Interpreting Molecular Phylogenetic Trees: Another Example



"Louisiana gastroenterologist" (Richard J. Schmidt) accused of attempted second degree murder for allegedly injecting a former lover with blood from one of his HIV+ patients

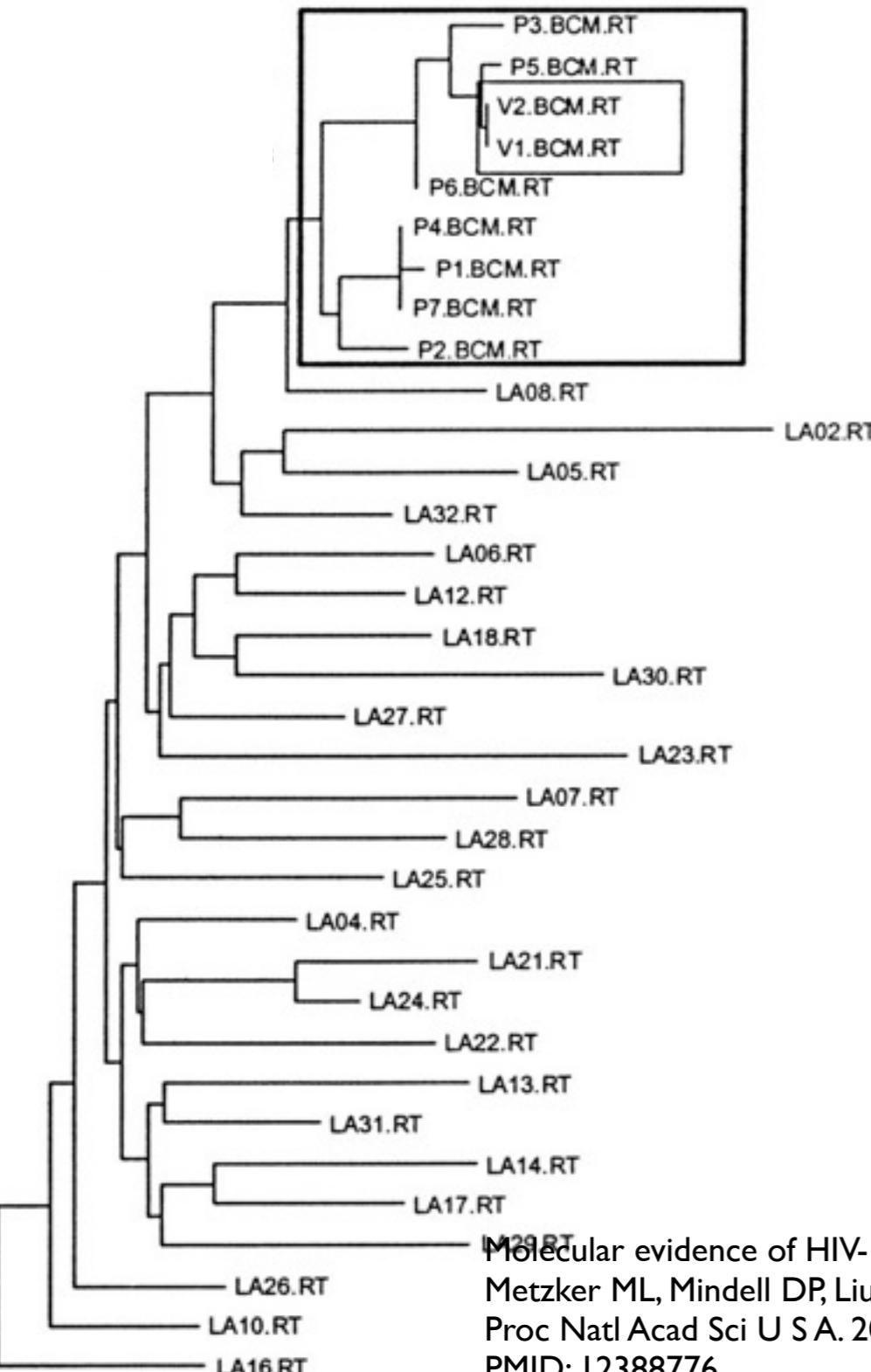
Phylogenetic analyses used as evidence in the trial

Few applications where we care this much (and the accused even more!) that the analysis is done as well as possible

i.e relevant parameters estimated as accurately and precisely as possible

Molecular evidence of HIV-1 transmission in a criminal case.  
Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA, Hillis DM.  
Proc Natl Acad Sci U S A. 2002 Oct 29;99(22):14292-7.  
PMID: 12388776

# Interpreting Molecular Phylogenetic Trees: Another Example



- victim (V)
- gastroenterologist's HIV+ patient (P)
- randomly-selected isolates of contemporary sequences from Louisiana (LA)

Note: several samples taken from patient (7) and victim (2)

Imagine you are a **juror** on this case....

How/would seeing this phylogeny influence the verdict you would choose in this case?

1. More likely to choose "guilty"
2. More likely to choose "not guilty"
3. Would not influence your choice of verdict

# Rooted Phylogenies

Terminology and Concepts

# Definitions

---

Phylogeny terminology and concepts "definition" exercise  
(see [homepage/InterpretingPhylogeniesCrete2014.html](#))

To begin looking at how we think about phylogenies, we'll explore together how we define several fundamental phylogenetic concepts

Try to write, on your own, definitions of:

- phylogenetic tree
- branch (of a phylogenetic tree)
- root (of a rooted phylogenetic tree)

**Writing forces you to be explicit about what you mean, and can help identify issues you are uncertain about**

Then compare your definitions with your neighbour, and write together a consensus definition for each term

# Definitions

---

Phylogeny terminology and concepts "definition" exercise  
(see [homepage/InterpretingPhylogeniesCrete2014.html](#))

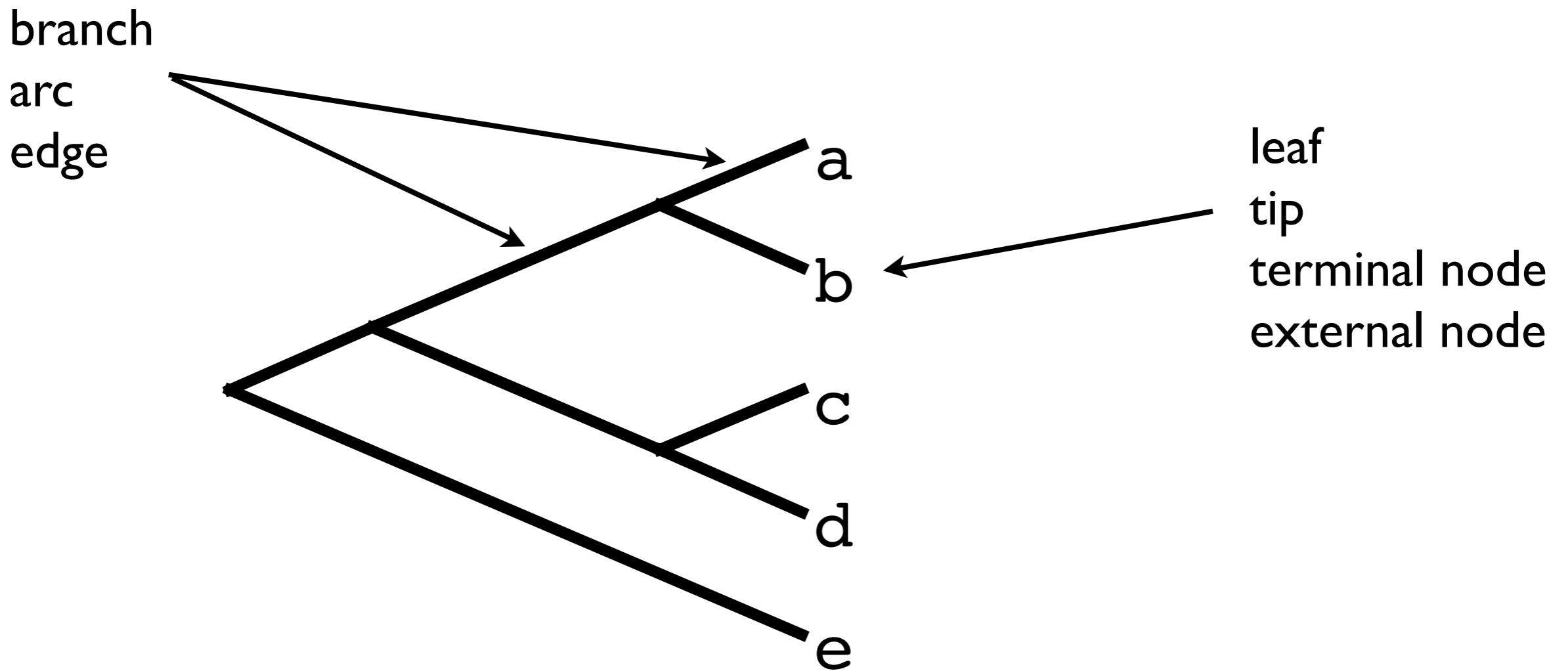
below are some suggested definitions for these terms

**phylogenetic tree:** A description of a path of transmission of genetic information between a set of taxa.

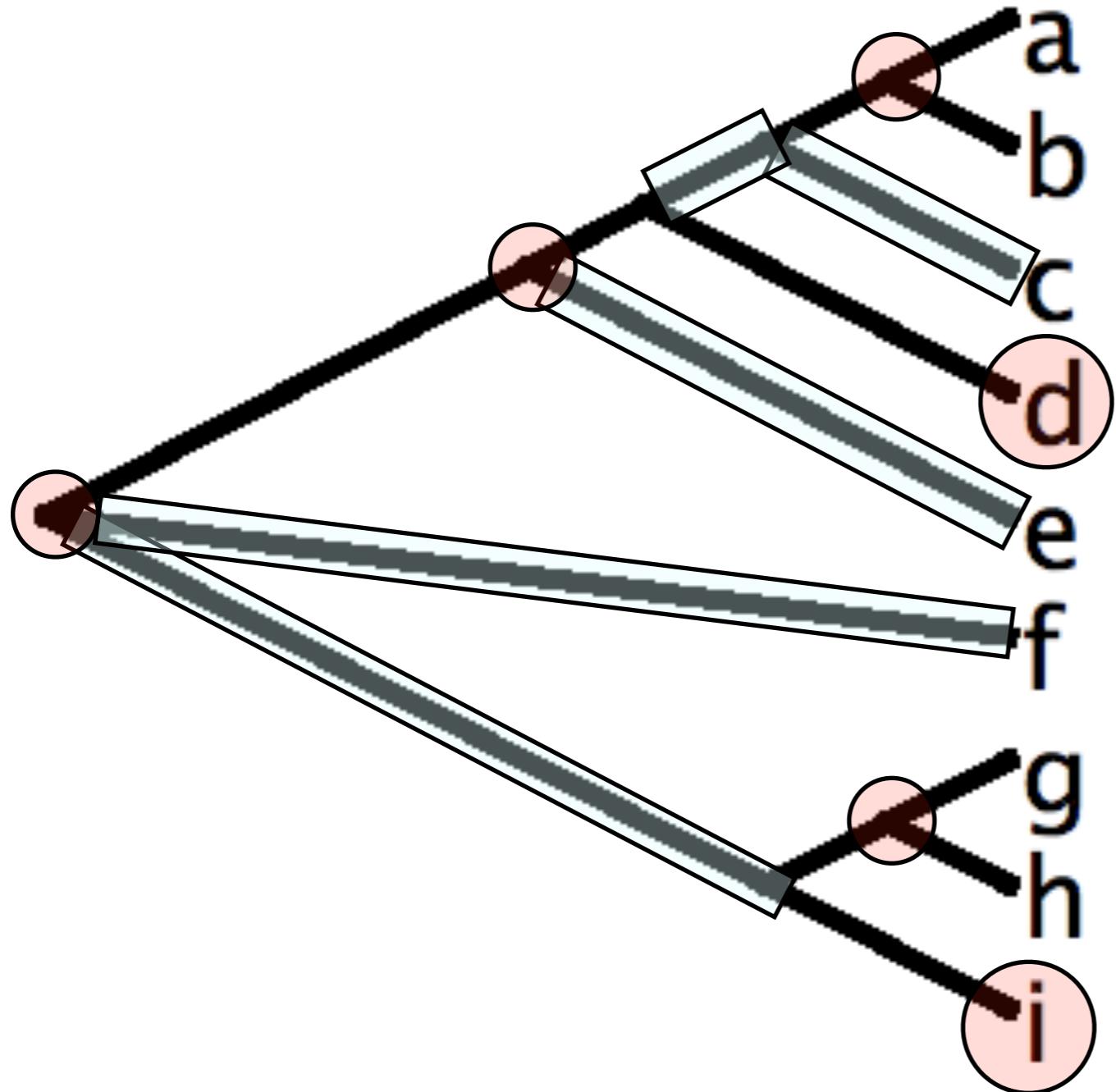
**branch (of a phylogenetic tree):** Lineages of taxonomic units that link nodes within a phylogenetic tree.

**root (of a rooted phylogenetic tree):** In a rooted tree, the node that represents the most recent common ancestor taxon of all other taxa in the tree.

# Alternative Tree-Related Terminologies



# Trees: Branches and Nodes



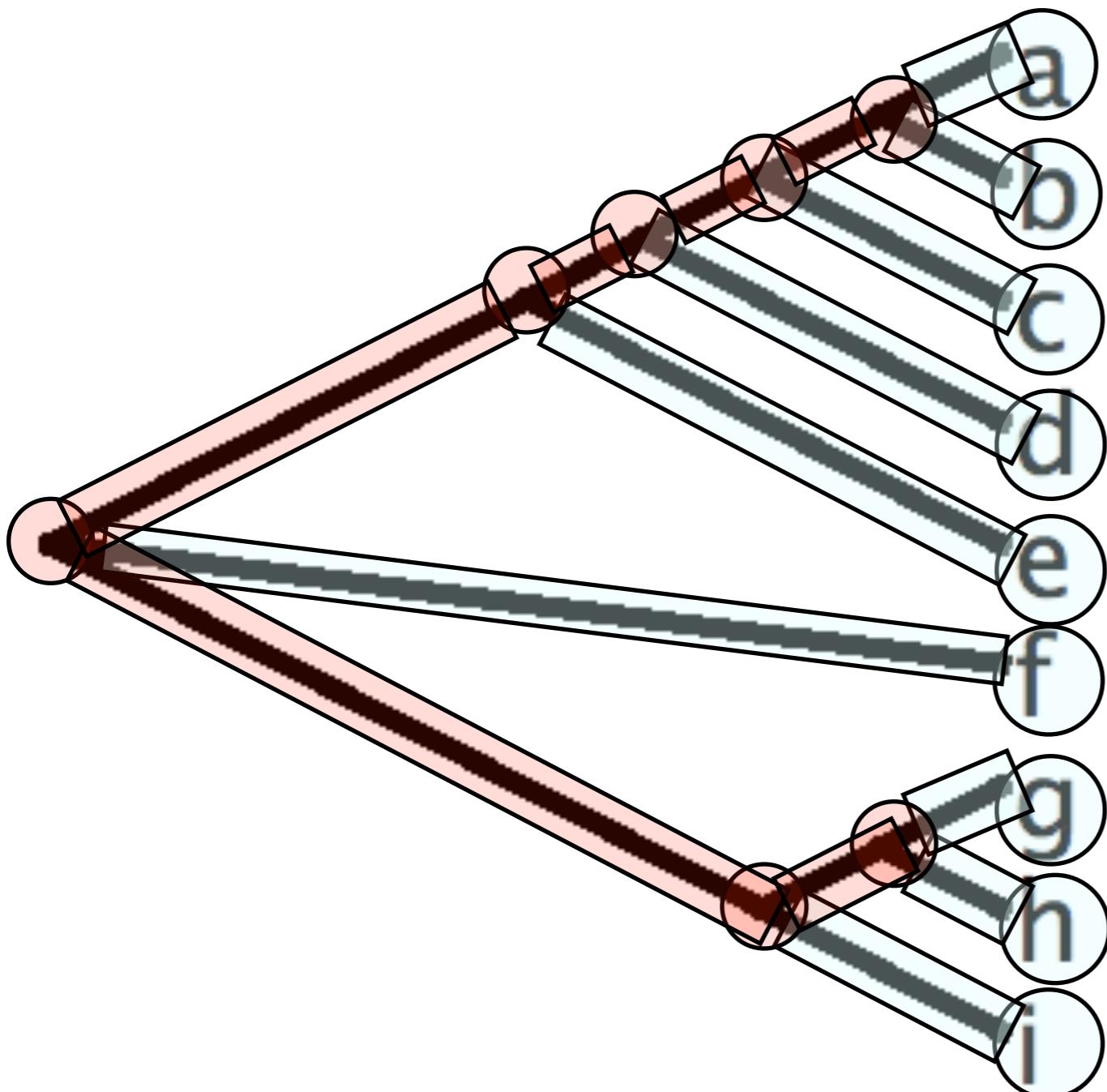
Trees consist of:

branches

nodes (ends of branches)

# Internal/External Nodes/Branches

Branches and Nodes are either:



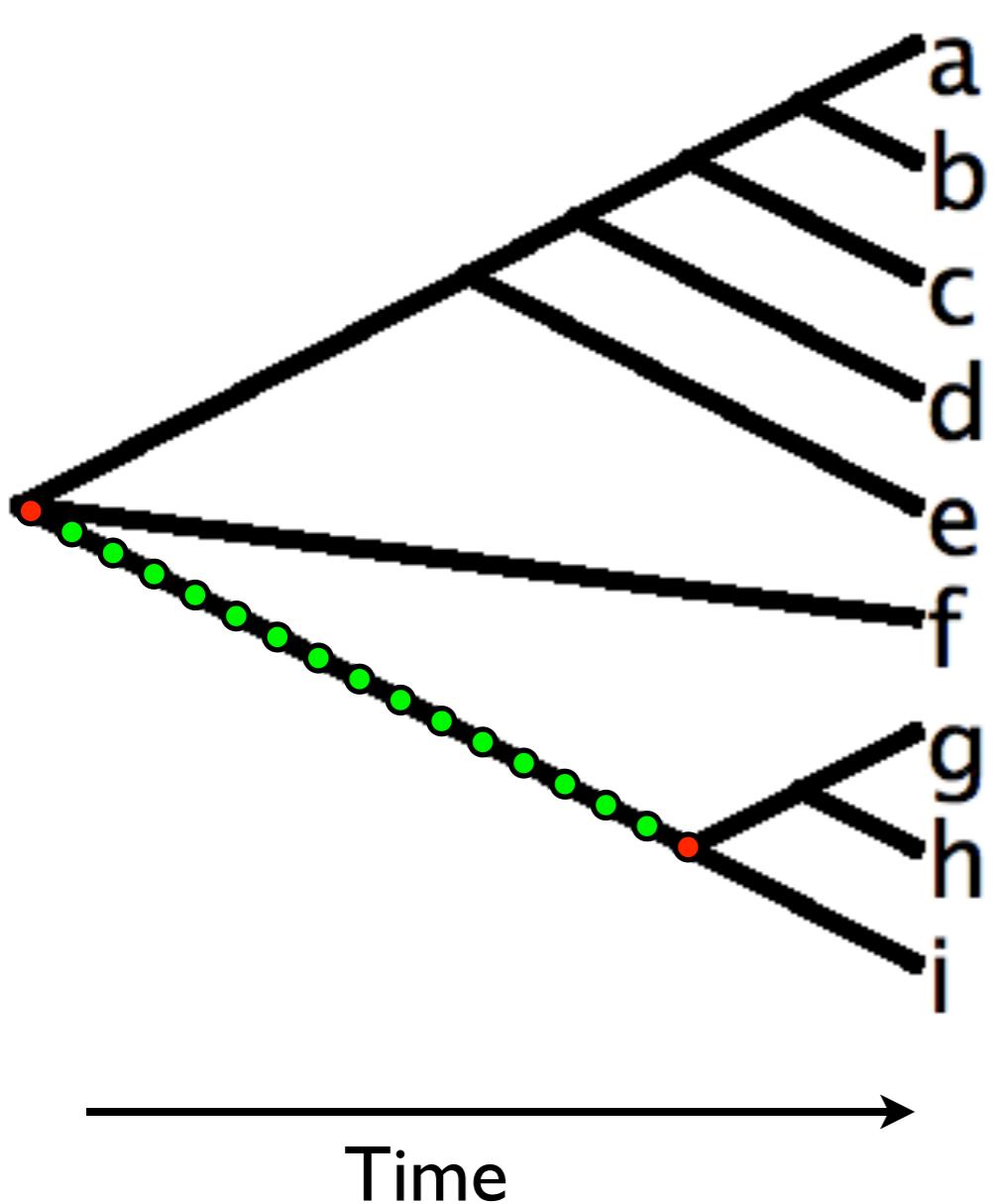
**internal/interior**

**Node** - at the intersection of two or more branches  
**Branch** - links two internal nodes

**external/terminal**

**Node** - associated with an extant sequence/OTU (operational taxonomic unit)  
**Branch** - links an external and an internal node

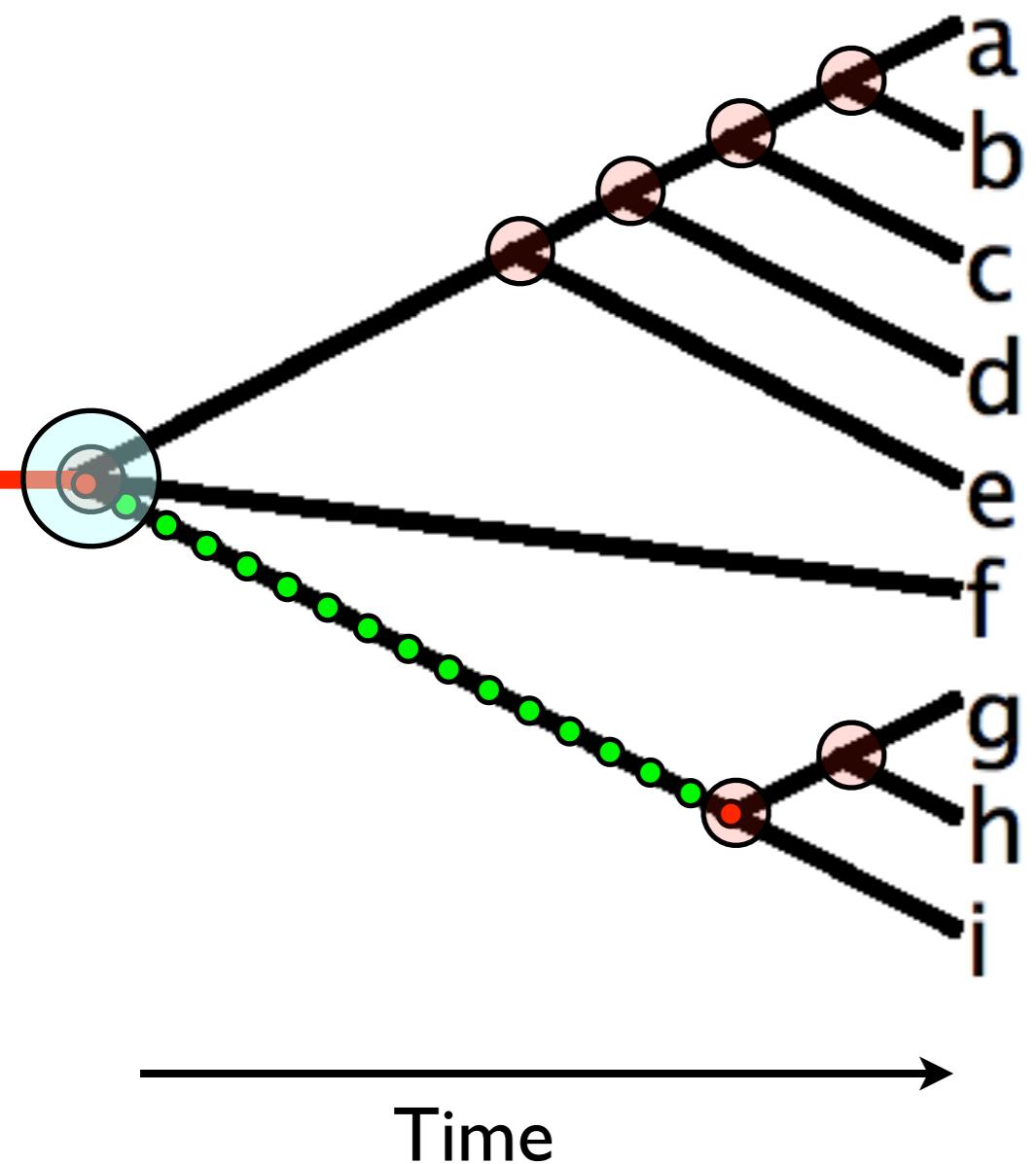
# Branches



## Branches

- represent successive generations of “taxa”
- ‘later’ taxa have “earlier” taxa as their ancestors
- i.e. a lineage
- time flows from the base of the tree to the tips

# Internal Nodes



## Internal Nodes

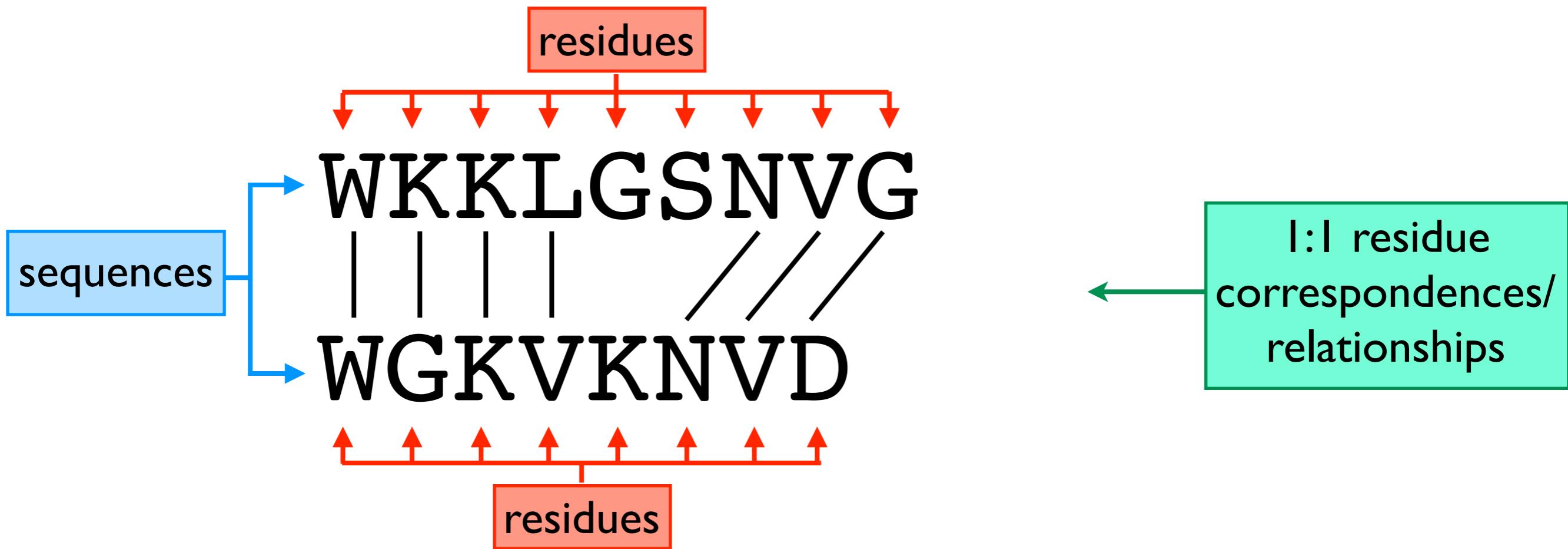
- represent hypothetical ancestral taxa/sequences/organisms
- i.e. HTUs - hypothetical taxonomic units

## Root (Root Node)

- A "special" internal node
- The most recent common ancestor of all OTUs
- Usually implies many other **less recent common ancestors**

# Alignments

# "Anatomy" of a Sequence Alignment



## I:I residue correspondences/relationships

Correspondences between

- a single residue in one sequence and
- a single residue in another sequence

# "Anatomy" of a Sequence Alignment



I:I residue correspondences/relationships

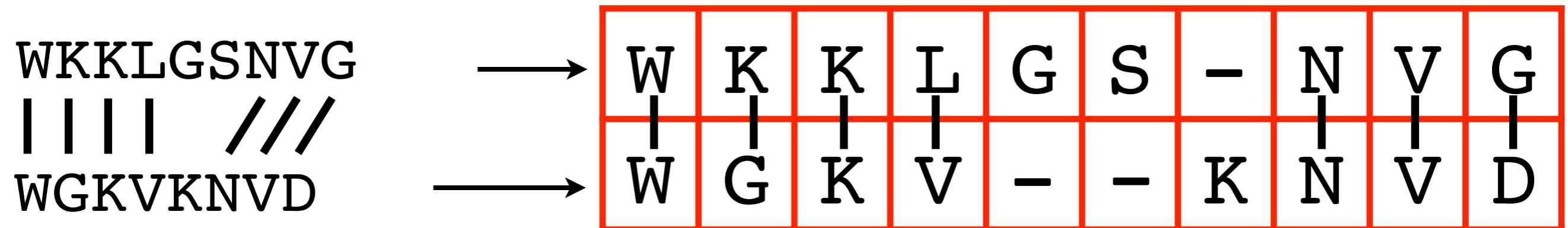
Residue has no equivalent in the top sequence

i.e. no residue in the top sequence has a I:I relationship with this residue

Could perhaps say there is a "I:2" relationship between this residue and these residues

However, alignments focus on I:I relationships

# Sequence Alignment Within a Grid



Often represented using a **grid/matrix**:

One sequence per row

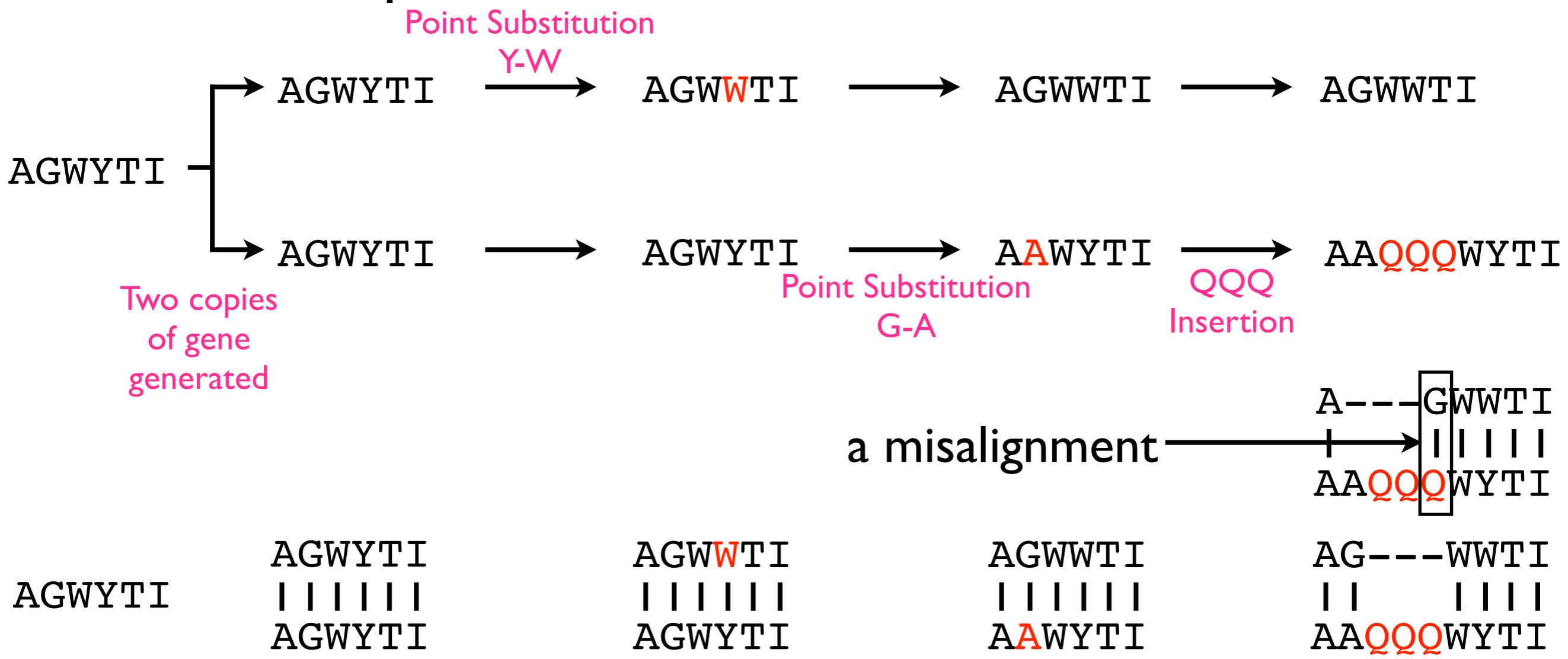
Residues in the same column are 'equivalent'

Gap characters (usually "-") indicate that the sequence contains no residues 'equivalent' to other residues in that column

# Evolutionary "Equivalence"

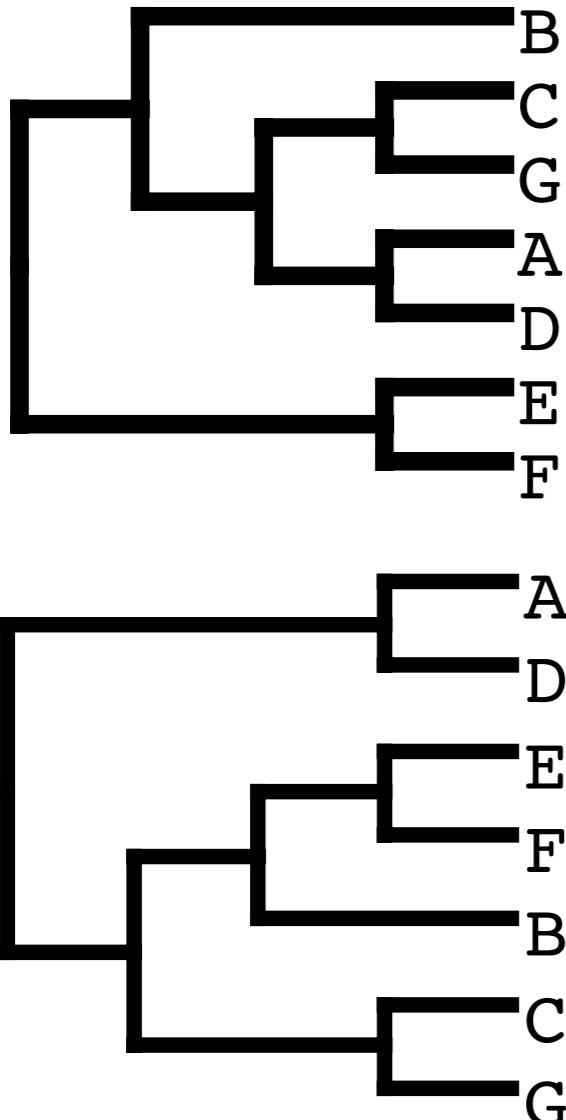
Residues are "evolutionarily equivalent" when:

- they are derived from the same residue in an ancestral sequence
- the only mutations experienced during divergence from this ancestral residue were **point substitutions**



# Branch Lengths

# Unscaled Trees



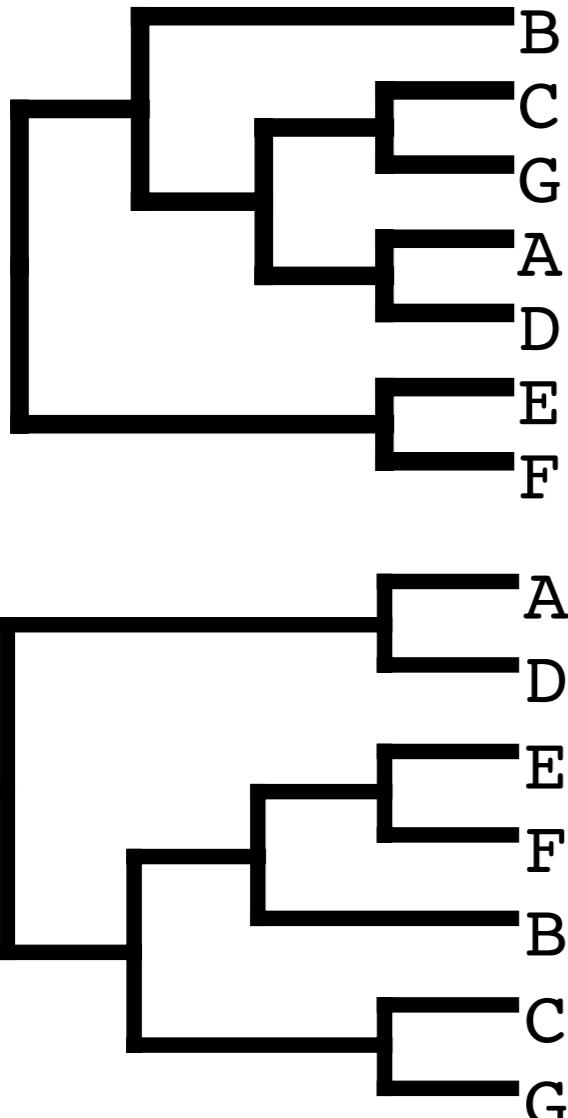
Branch lengths provide no information

Branch lengths usually chosen to align OTU labels

Re-rooting the tree typically changes the choice of branch lengths

Same unscaled unrooted tree

# Scaled Trees



Same unscaled  
unrooted tree

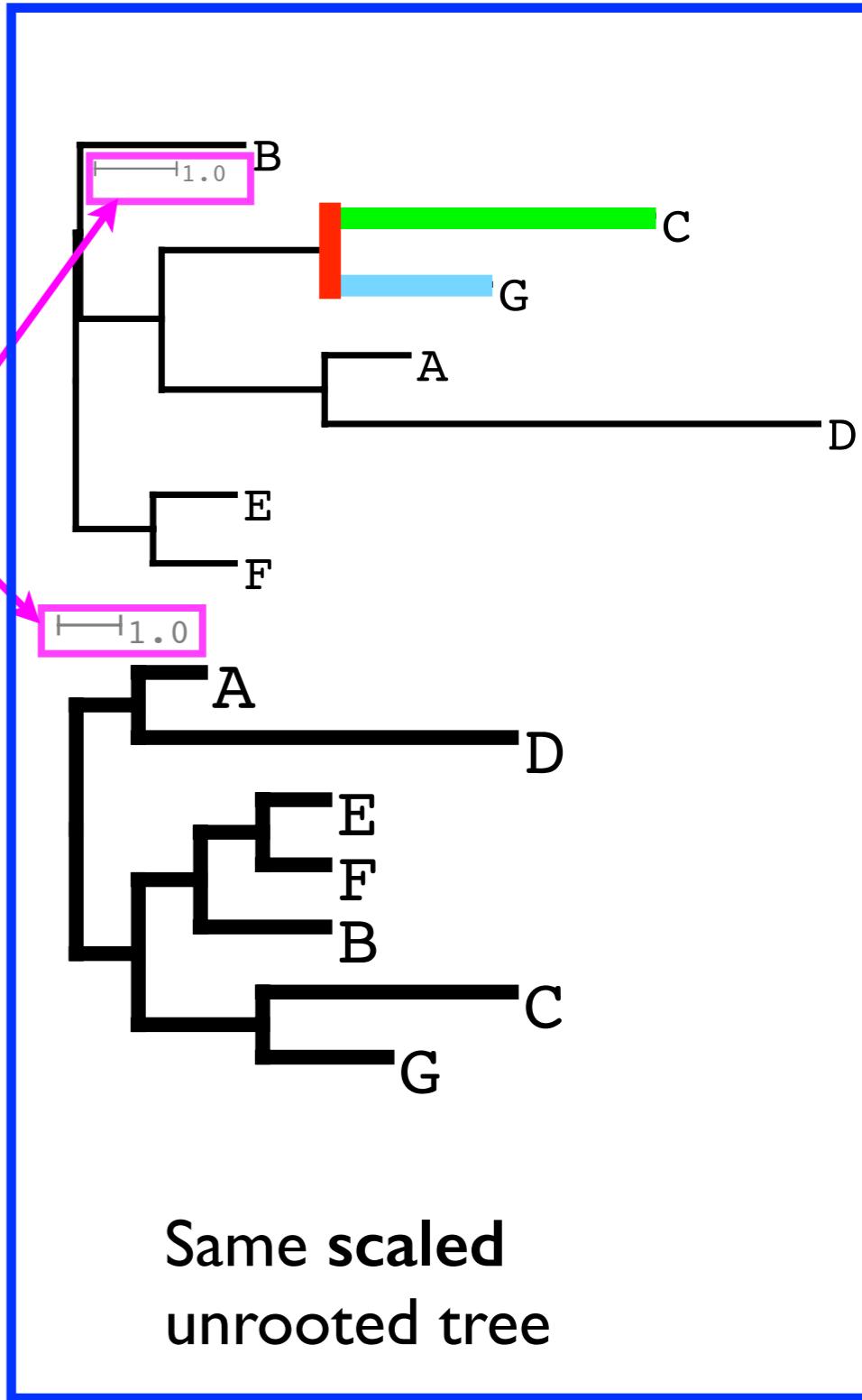
Branch length usually represents some measure of the difference/distance between TUs at ends of the branch

Tree should be presented together with a scale bar

For rectangular trees, “node lines” are NOT branches! Their length provides no indication of intertaxa difference/distance!

i.e. distance between taxa C and G is the sum of the green and cyan lines (it does NOT include the length of the red line!)

C —————— G



Same scaled  
unrooted tree

# Branch Lengths

Usually an ESTIMATE of the EXPECTED/AVERAGE number of substitutions per site between two sequences

SeqA	I	K	T	I	I	L	K	W	W	S	P
SeqB	I	K	T	I	V	K	W	D	S	P	

If we assume:

- All identical residues between two sequences have not experienced substitutions
- All different residues have experienced one substitution

Mean/Average No. Substitutions =  $2/10 = 0.2$

SeqA —————<sup>0.2</sup>———— SeqB

# Branch Lengths

Usually an ESTIMATE of the EXPECTED/AVERAGE number of substitutions per site between two sequences

SeqA	I	K	T	I	I	L	K	W	W	S	P
SeqB	I	K	T	I	V	K	W	D	S	P	

Branch-length estimate depends on SUBSTITUTION MODEL

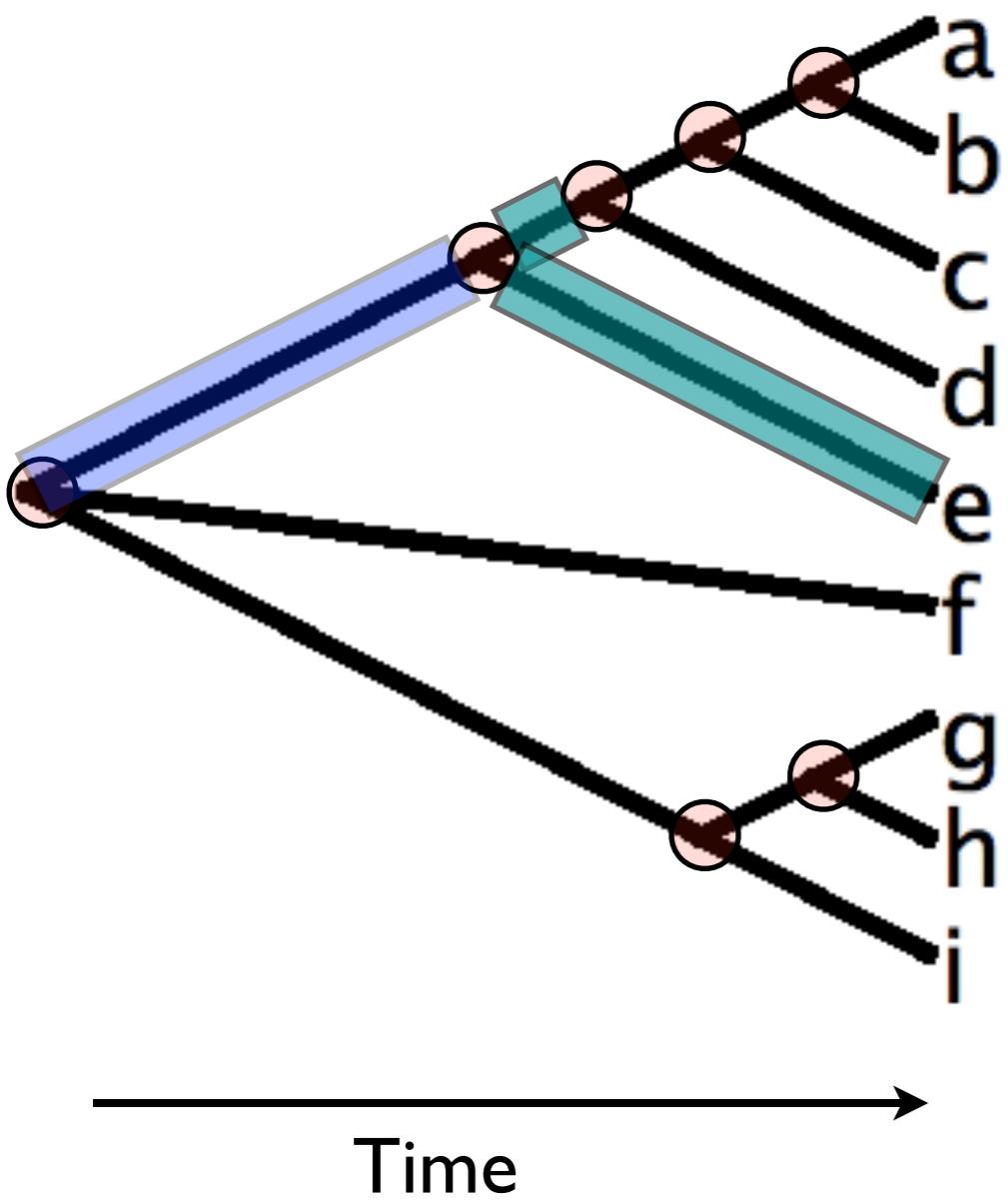
Further assumptions of this model

- All alignment positions/residues evolve (are substituted at) the same rate
- All residues substitute to all other residues at the same rate i.e. A->G at same frequency as A->W

SeqA ————— 0.2 SeqB

# More Rooted Tree Terminology

# Parent/Daughter Branches

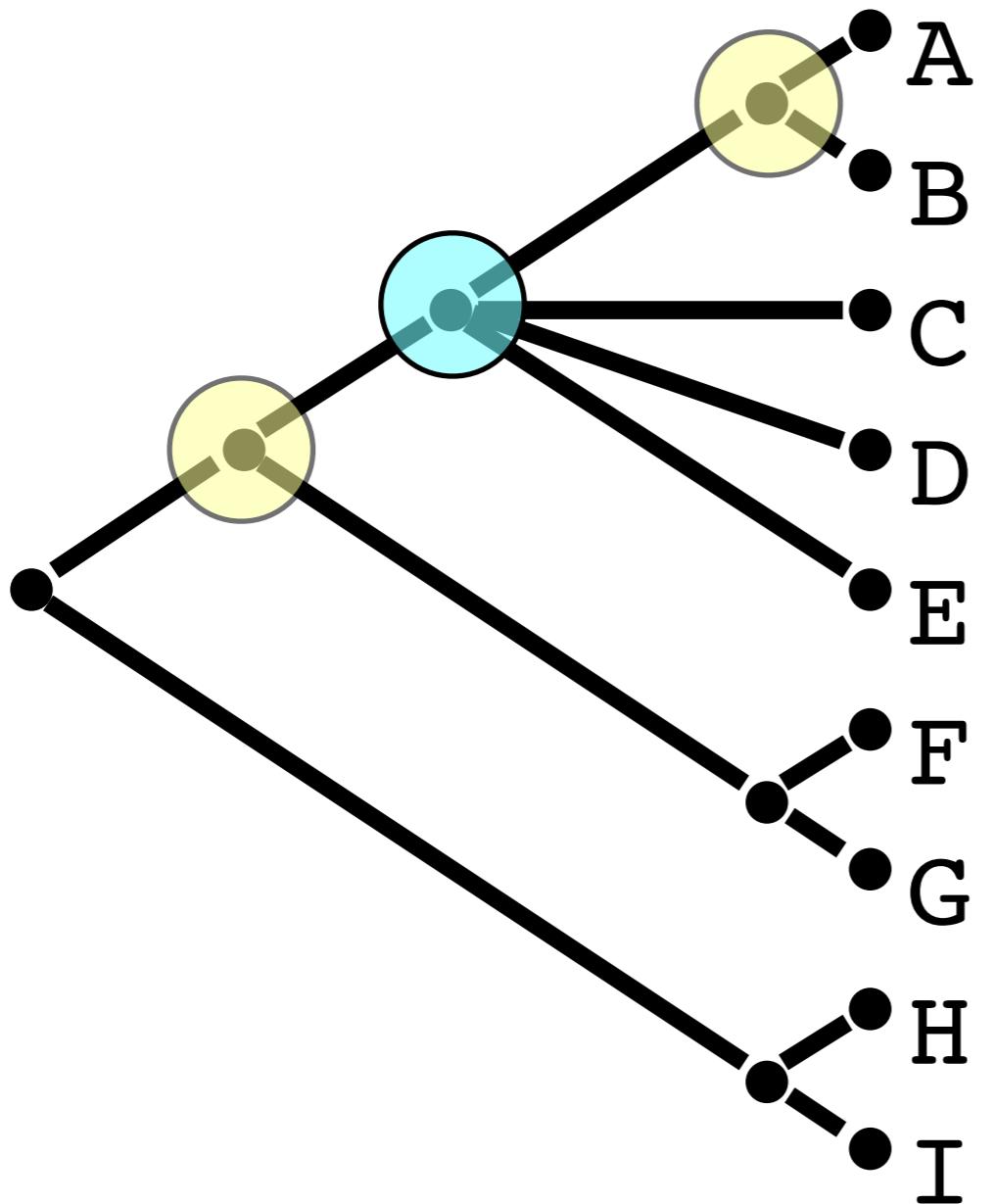


parental/ancestral  
lineages/branches

diverge into

multiple daughter  
lineages/branches

# Polytomies



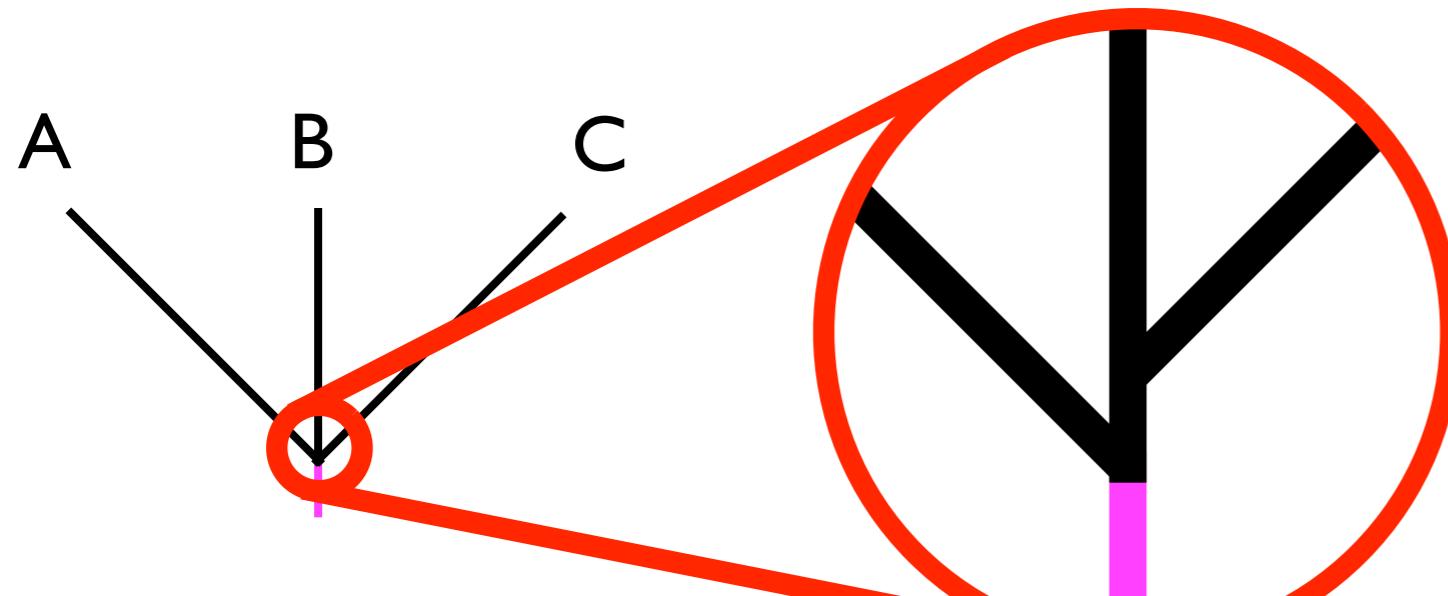
## Polytomies

Internal nodes associated with more than two daughter branches

Internal nodes with two daughter branches are bifurcations

How many bifurcations on the tree? (a) 4 (b) 5 (c) 6

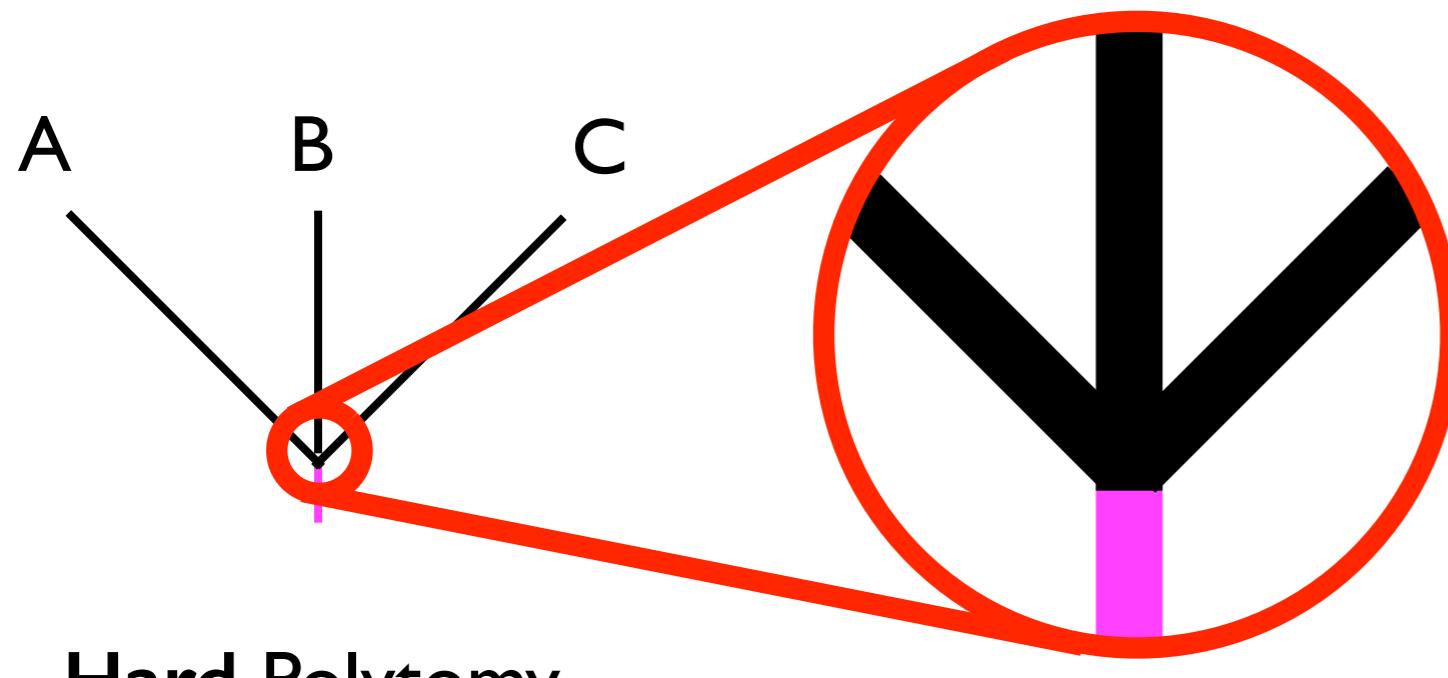
# Interpreting Polytomies



Soft Polytomy

Lineages only bifurcate - internal lineages so short that no identifiable change/evolution occurred along them

Thus true pattern of lineage divergence cannot be resolved



Hard Polytomy

Ancestral lineage diverged into 3+ lineages simultaneously

**NB:** Some software only accepts bifurcating trees

# Relatedness

# Relatedness (in the context of phylogenetic trees)

---

Inferring patterns of relatedness is often one of the main aims of evolutionary tree estimation.

"relatedness" in this context has a specific meaning, as exemplified here:

*"the more recently species share a common ancestor, the more closely related they are" \**

As "relatedness" has other meanings in other contexts, there can be some confusion about it's meaning in a **phylogenetic** context

As many trees are estimated to inform ideas about patterns of relatedness, we need a clear understanding of how the term is used in this context

Thus, in the next slides, we will look at several examples of how the word is used when describing phylogenetic relationships

\* Evolution. The tree-thinking challenge.

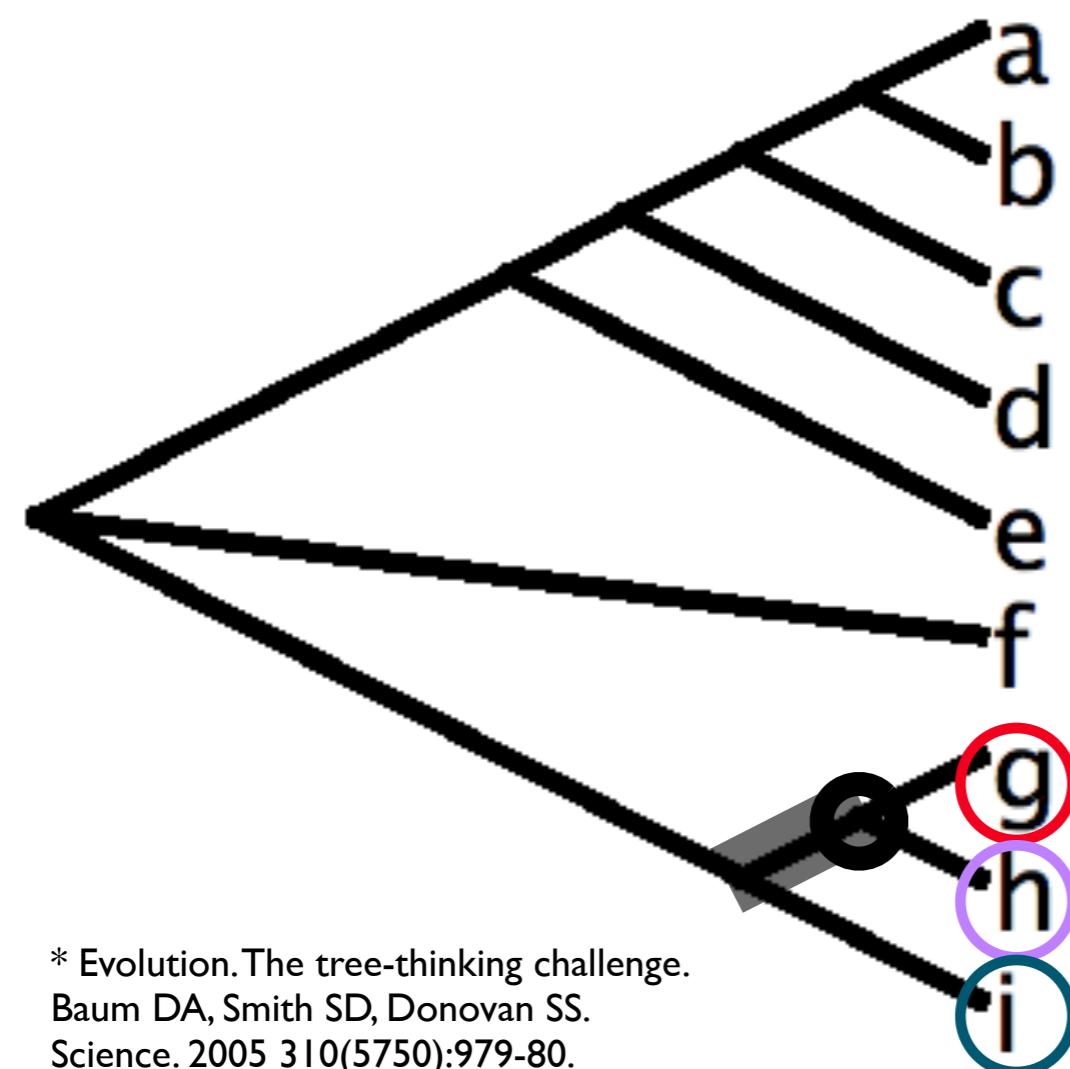
Baum DA, Smith SD, Donovan SS.

Science. 2005 310(5750):979-80.

PMID: 16284166

# Relatedness (in the context of phylogenetic trees)

"the more recently species share a common ancestor, the more closely related they are" \*

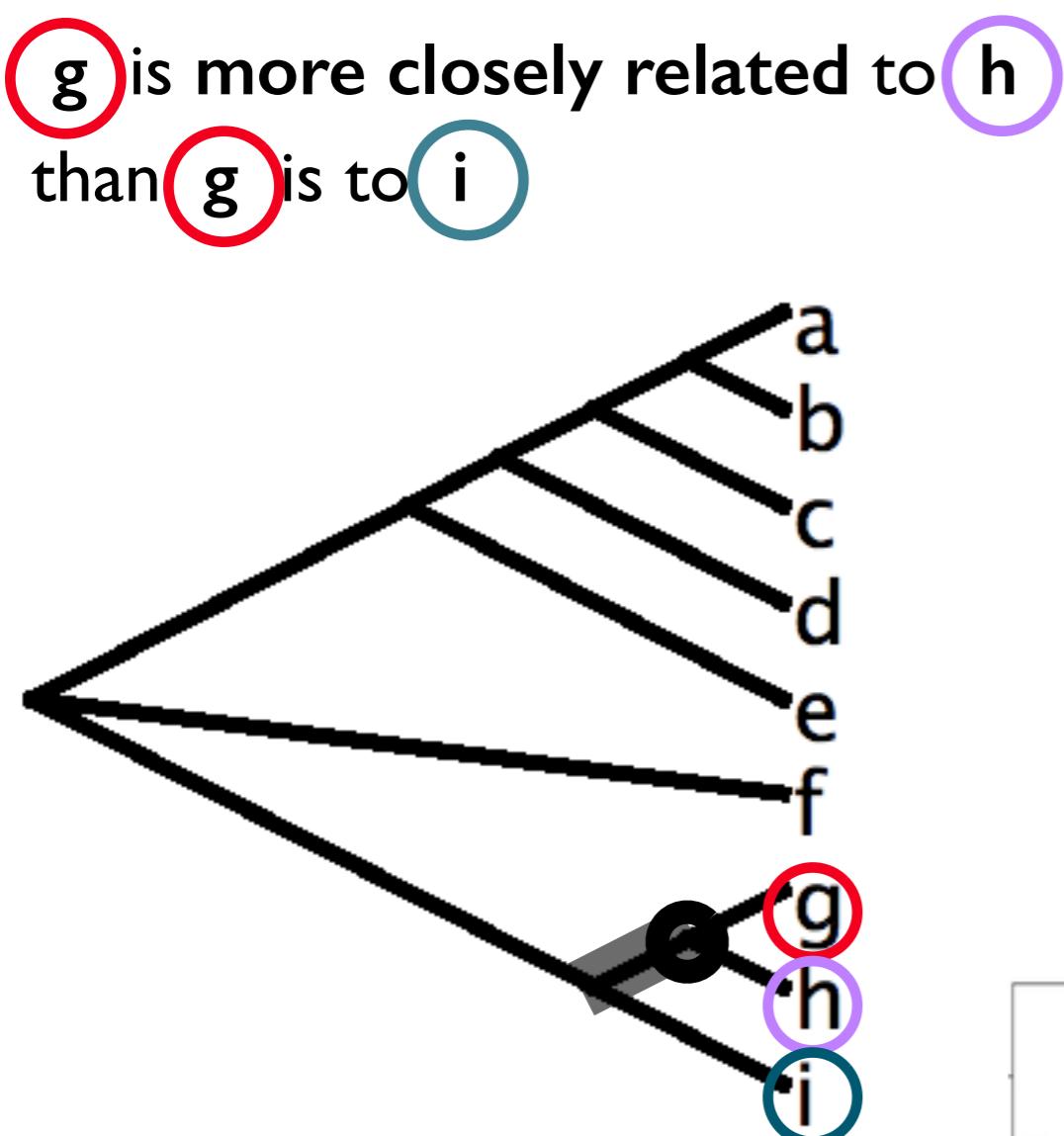


g is more closely related to h  
than g is to i  
because g and h share common ancestors  
that neither share with i  
i.e. degree of relatedness  
is associated with the extent of ancestry  
(i.e. the number of ancestors) taxa share  
with each other

\* Evolution. The tree-thinking challenge.  
Baum DA, Smith SD, Donovan SS.  
Science. 2005 310(5750):979-80.  
PMID: 16284166

# Relatedness (in the context of phylogenetic trees)

"the more recently species share a common ancestor, the more closely related they are" \*



... Plesiorycteropus is more closely related to tenrecoids than to tubulidentates ..

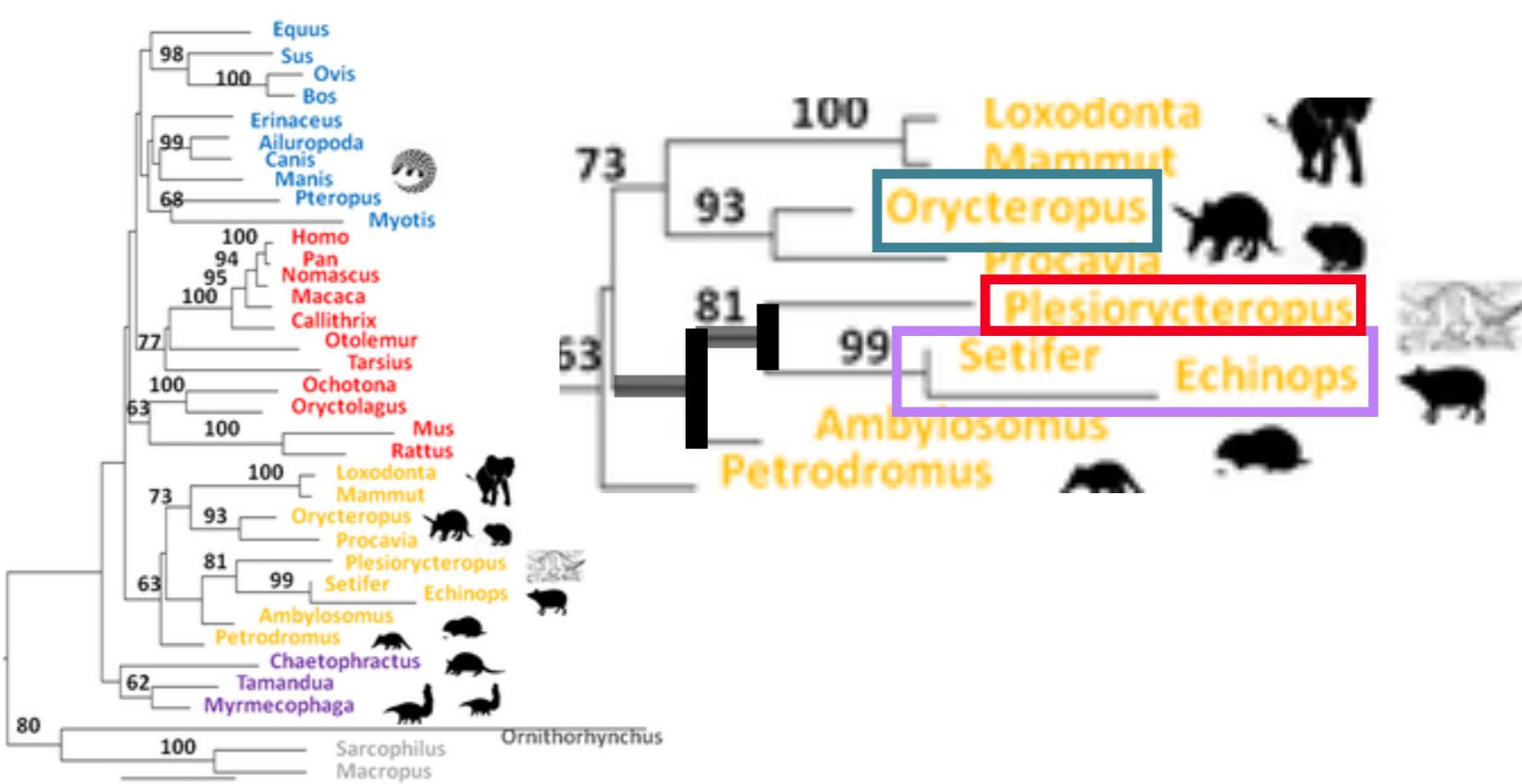


Figure 4. Phylogenetic analyses of Plesiorycteropus collagen (I) sequences obtained by LC-MS in comparison to previously postulated closest relatives.

\* Evolution. The tree-thinking challenge.  
Baum DA, Smith SD, Donovan SS.  
Science. 2005 310(5750):979-80.  
PMID: 16284166

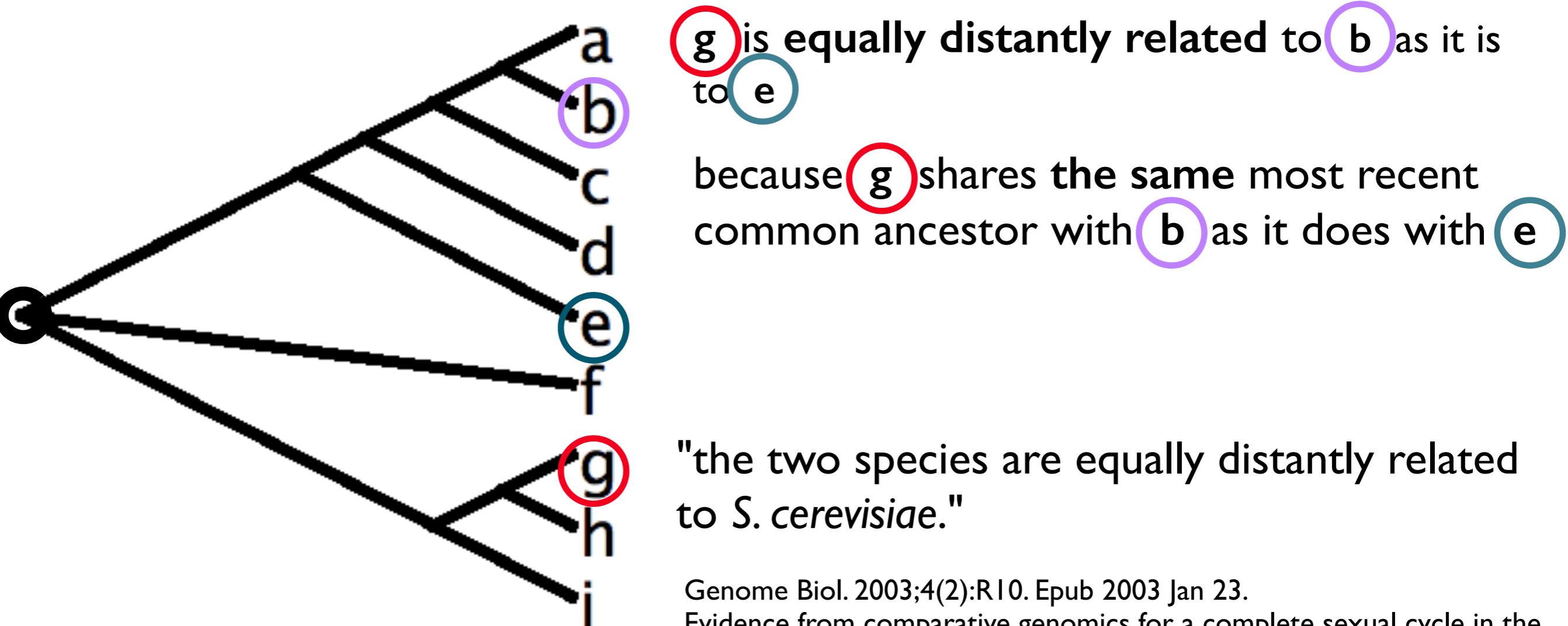
Buckley M (2013) A Molecular Phylogeny of Plesiorycteropus Reassigns the Extinct Mammalian Order 'Bibymalagasia'. PLoS ONE 8(3): e59614. doi:10.1371/journal.pone.0059614

<http://www.plosone.org/article/info:doi/10.1371/journal.pone.0059614>

Aidan Budd, EMBL Heidelberg

# Relatedness (in the context of phylogenetic trees)

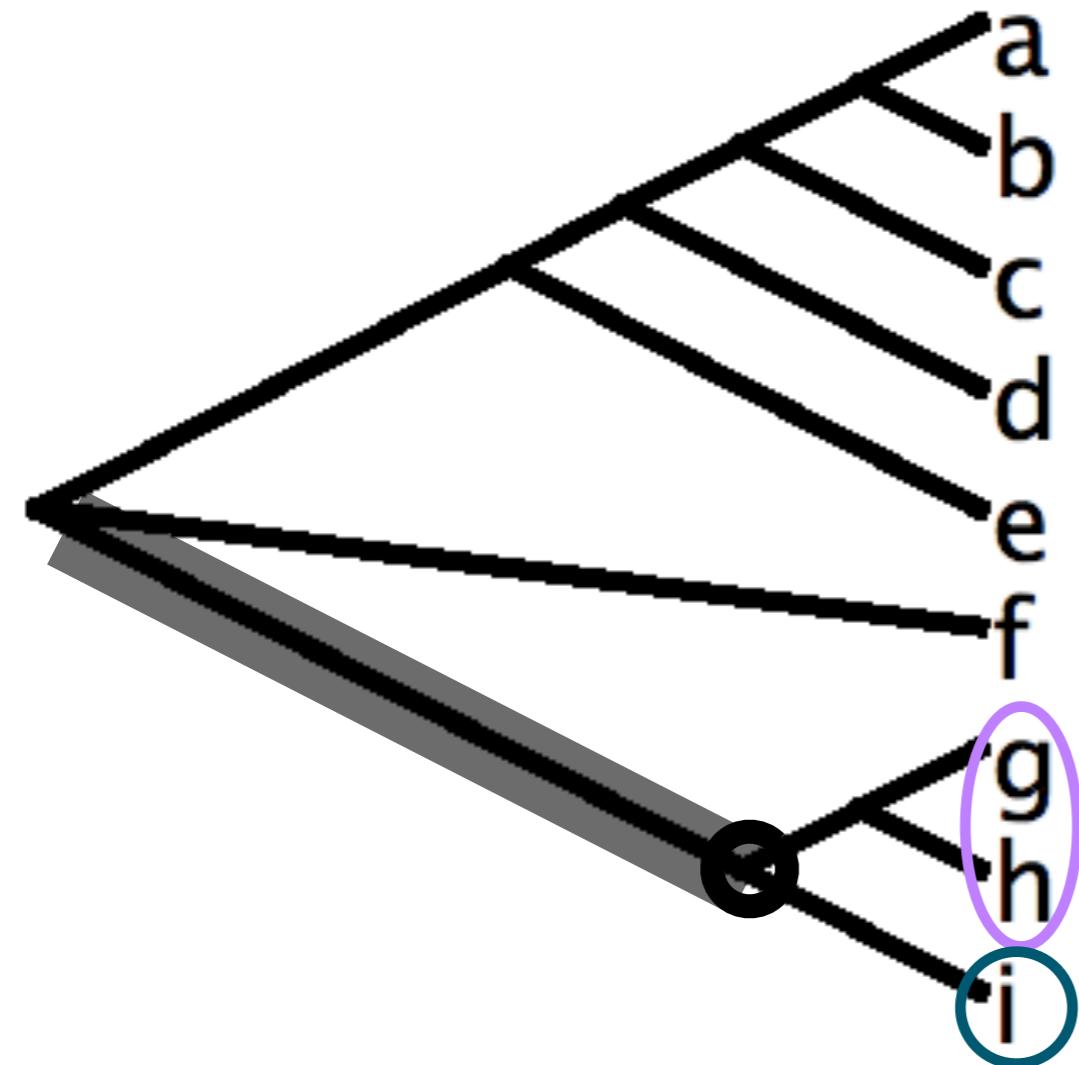
of all the OTUs represented in this tree



Genome Biol. 2003;4(2):R10. Epub 2003 Jan 23.  
Evidence from comparative genomics for a complete sexual cycle in the 'asexual' pathogenic yeast *Candida glabrata*.  
Wong S, Fares MA, Zimmermann W, Butler G, Wolfe KH.

# Relatedness (in the context of phylogenetic trees)

of all the OTUs represented in this tree



i is most closely related to g and h  
(i.e. i is the *sister group* of g and h... which is equivalent to saying g and h are the sister group of i)

because i shares common ancestors with g and h that it does not share with any other OTUs in the tree

"PEPV was confirmed to [...] be most closely related to Turkeypox virus (TKPV), Ostrichpox virus (OSPV) and Pigeonpox virus (PGPV)."

Virol J. 2009 May 8;6:52. doi: 10.1186/1743-422X-6-52.

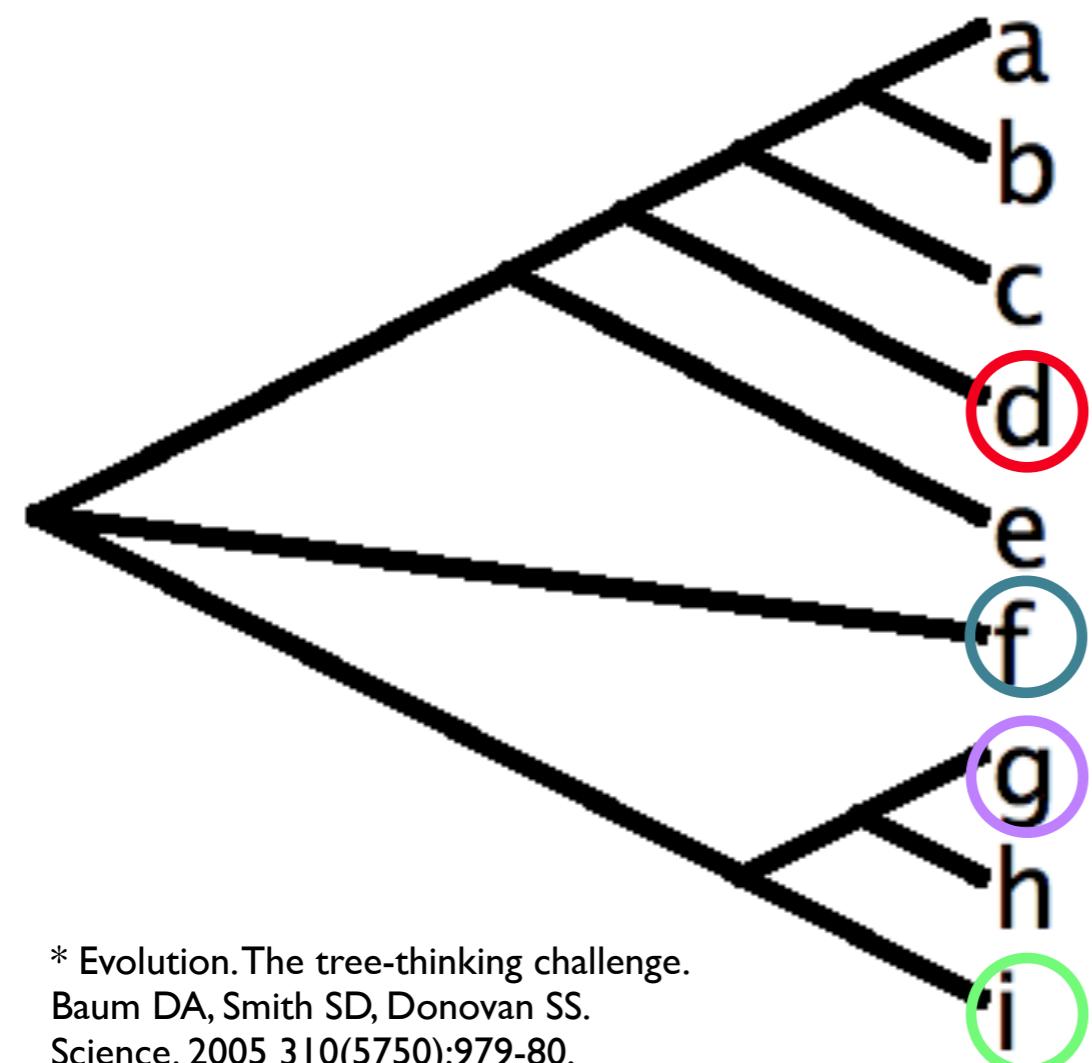
Phylogenetic analysis of three genes of Penguinpox virus corresponding to Vaccinia virus G8R (VLTF-1), A3L (P4b) and H3L reveals that it is most closely related to Turkeypox virus, Ostrichpox virus and Pigeonpox virus.

Carulei O, Douglass N, Williamson AL.

# Relatedness (in the context of phylogenetic trees)

"the more recently species share a common ancestor, the more closely related they are" \*

Which of the following statements is correct, given this tree?



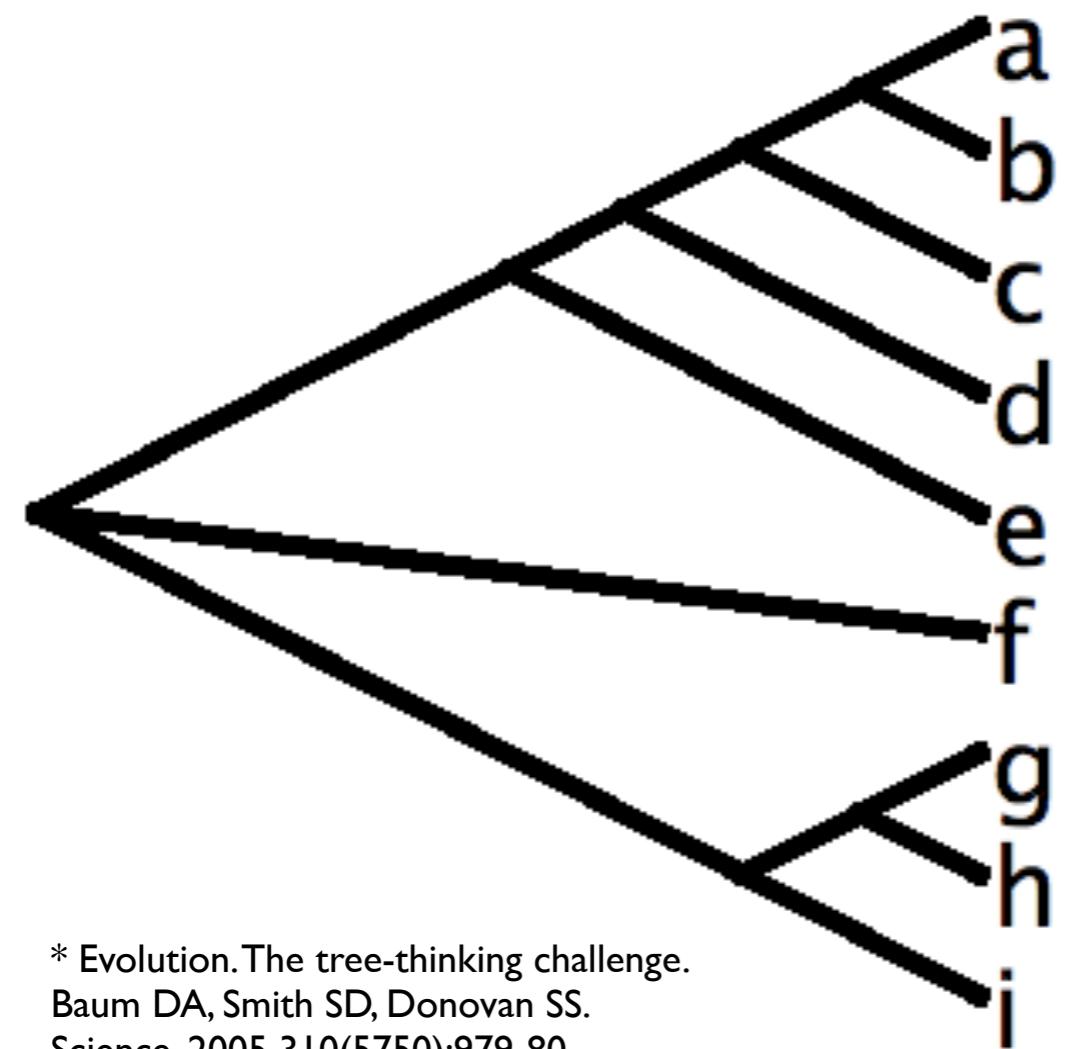
1. **d** is more closely related to than to **f** or **i** g
2. **d** is more closely related to than to **g** or **i** f
3. **d** is more closely related to than to **g** or **f** i

\* Evolution. The tree-thinking challenge.  
Baum DA, Smith SD, Donovan SS.  
Science. 2005 310(5750):979-80.  
PMID: 16284166

# Relatedness (in the context of phylogenetic trees)

"the more recently species share a common ancestor, the more closely related they are" \*

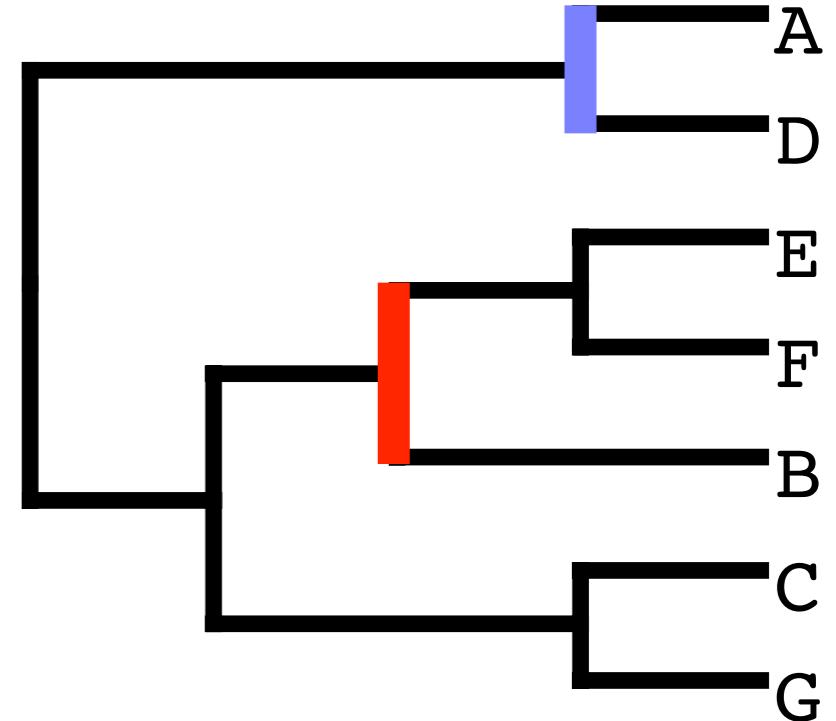
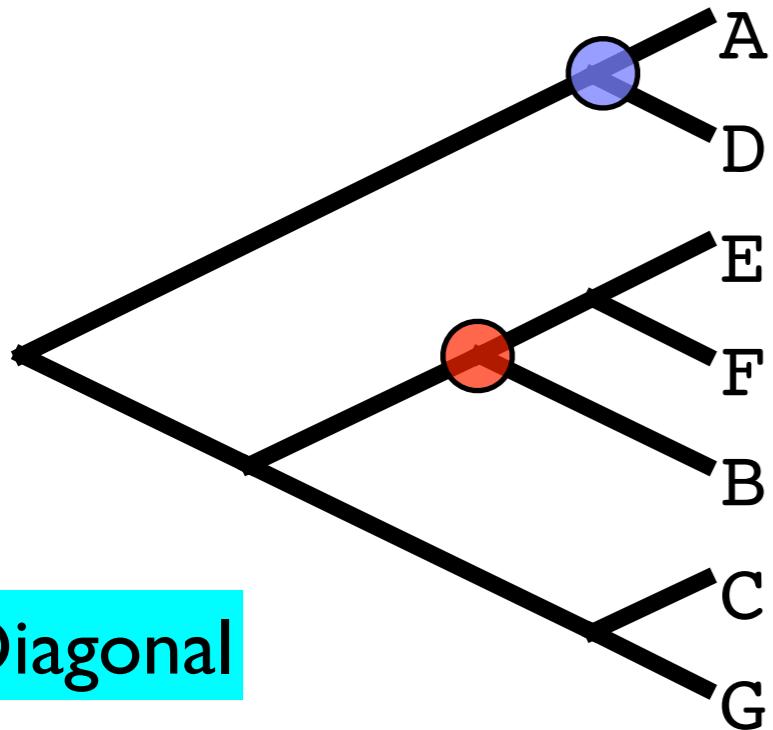
Why spend so much time discussing "relatedness" with you?



Many analyses aim to test whether particular "relatedness" statements are supported by the data - thus crucial that the statements are understood correctly, which is not always easy

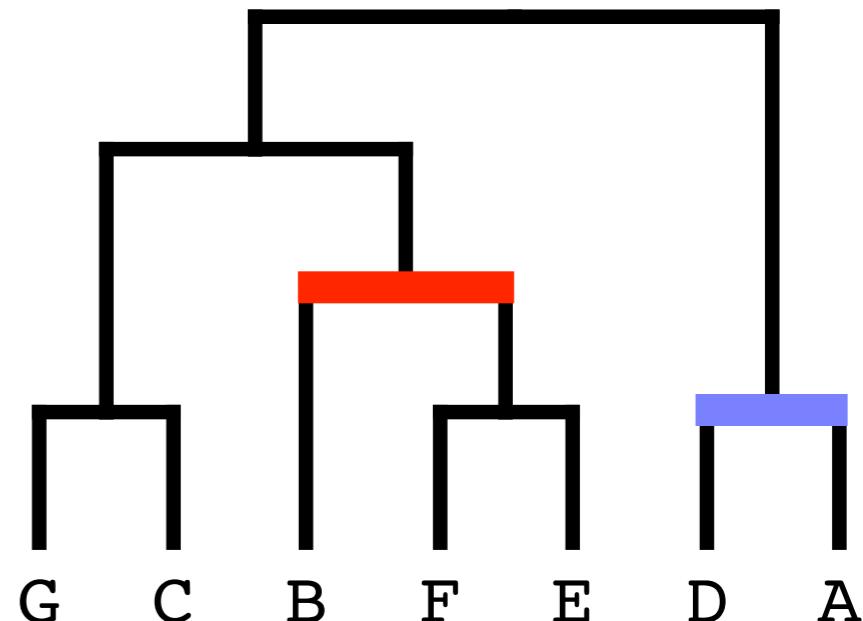
\* Evolution. The tree-thinking challenge.  
Baum DA, Smith SD, Donovan SS.  
Science. 2005 310(5750):979-80.  
PMID: 16284166

# Tree Representations



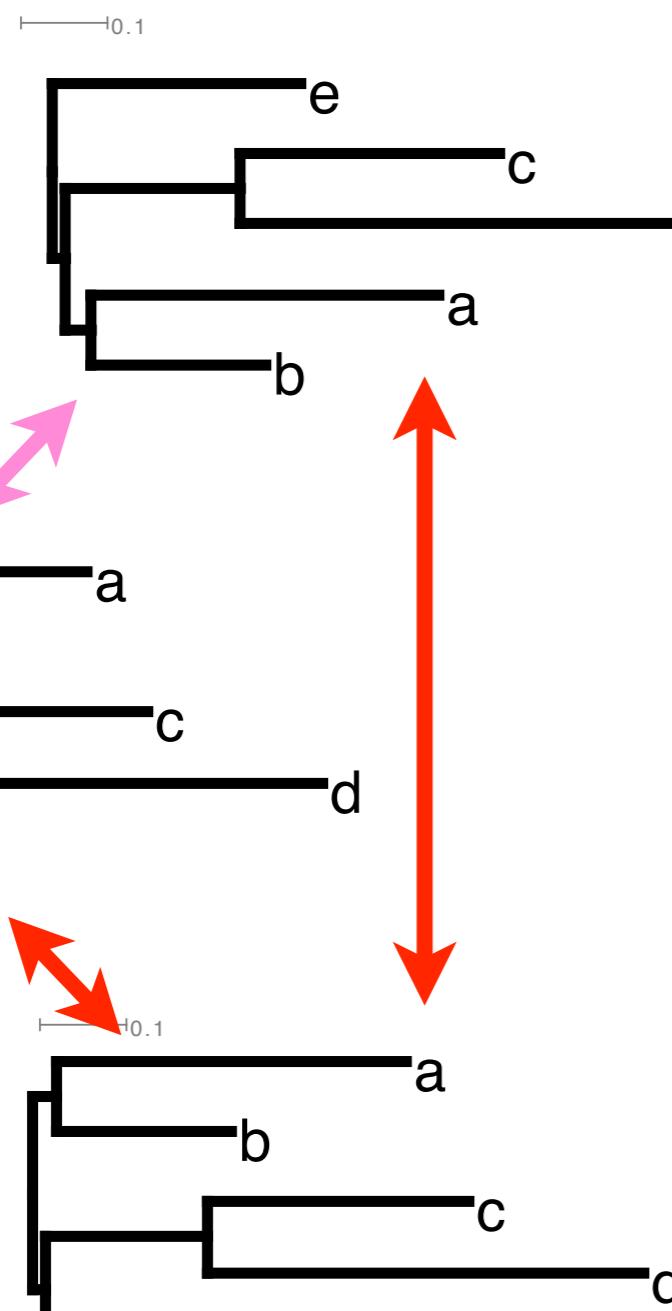
Most rooted tree figures use a “rectangular” rather than a “diagonal” representation

Rectangular trees represent internal nodes with lines perpendicular to lines representing the branches



Rectangular

# Tree Topology



Trees with **identical topologies**...  
... describe the same set of "relatedness statements" between taxa  
i.e. any (true!) statement such as  
*"c is more closely related to a than c is to e"*  
is true for all trees with identical topologies

↔ identical topologies  
↔ different topologies

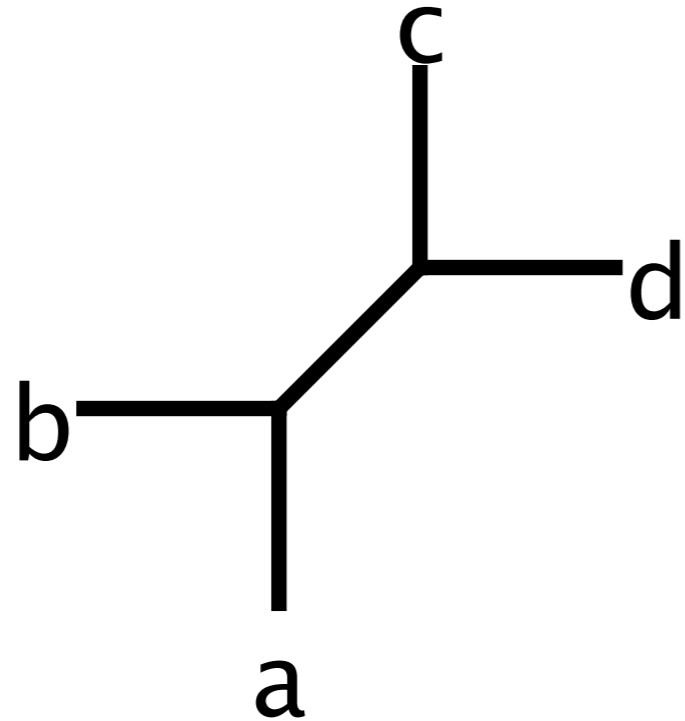
Trees with **different topologies**...  
... describe different sets of "relatedness statements" between taxa

# Unrooted Phylogenies

# Unrooted Trees

There's no root on the tree...

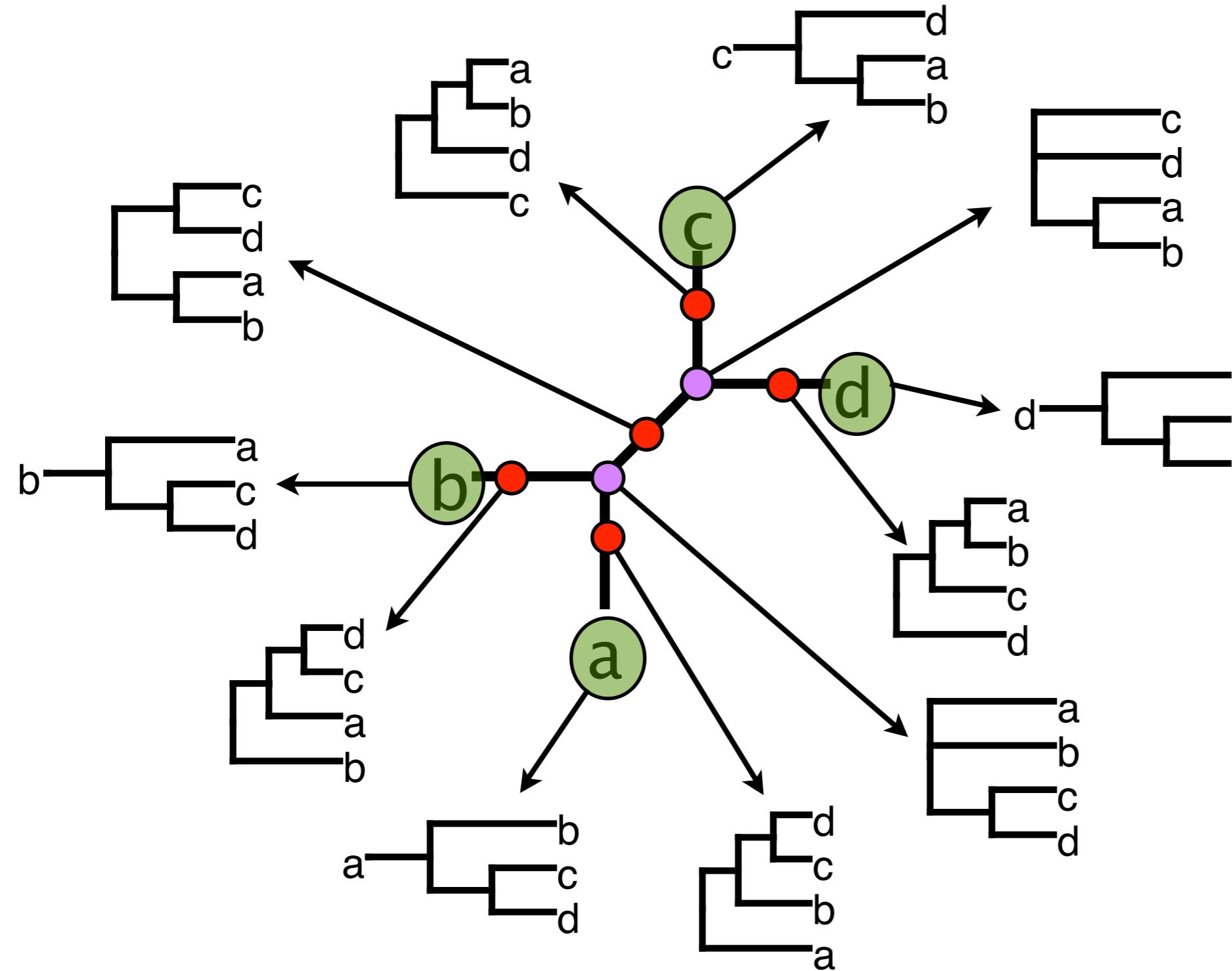
...which is usually interpreted as meaning that these taxa are related by a rooted tree but we don't know where the root is



Many applications of phylogenies require a rooted tree

But many tree estimation tools yield only unrooted trees!

# Unrooted → Rooted



There are multiple **rooted tree topologies** for a given unrooted tree topology

Unrooted trees can be rooted on their:

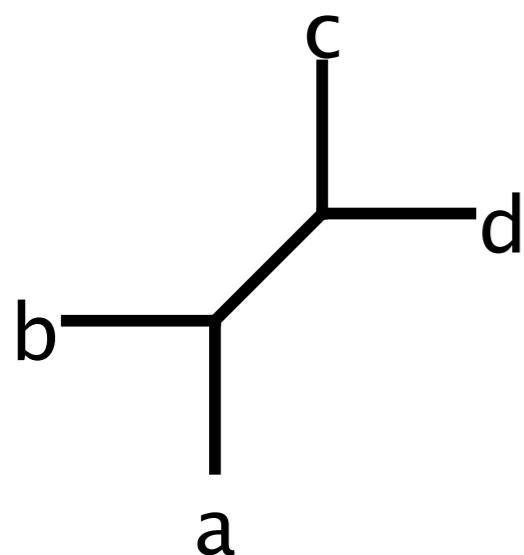
- **branches**
- **interior nodes**
- **terminal nodes**

# Quiz

Which of these statements is true, given the unrooted tree topology shown below?

**d** more closely related to:

1. **a** than it is to **b** or **c**
2. **b** than it is to **a** or **c**
3. **c** than it is to **a** or **b**



None of these is true, given this unrooted tree topology, under all possible rootings of the tree

Indeed, no rooted topology contains the relationships described in 1. and 2.

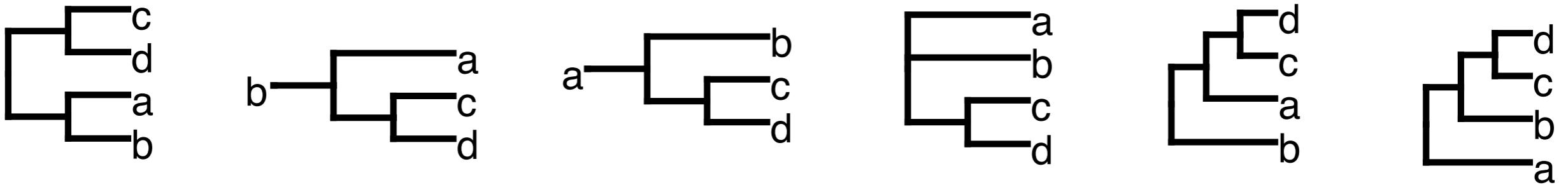
And, while 3. is true for some of the rooted trees, in others it is not

Draw the set of rooted tree topologies in which statement 3. is:

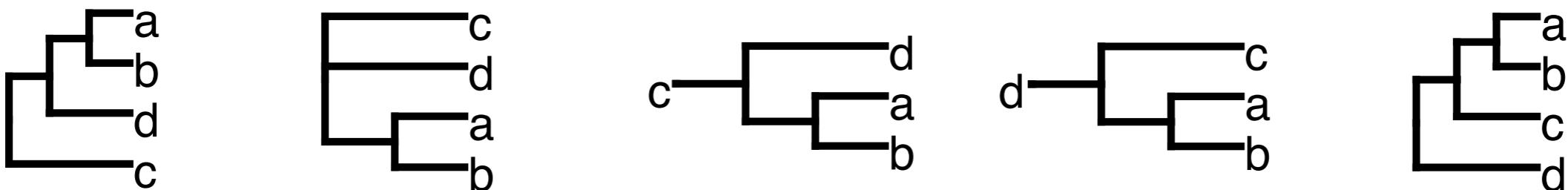
- true
- false

# Quiz

d more closely related to c than it is to a or b

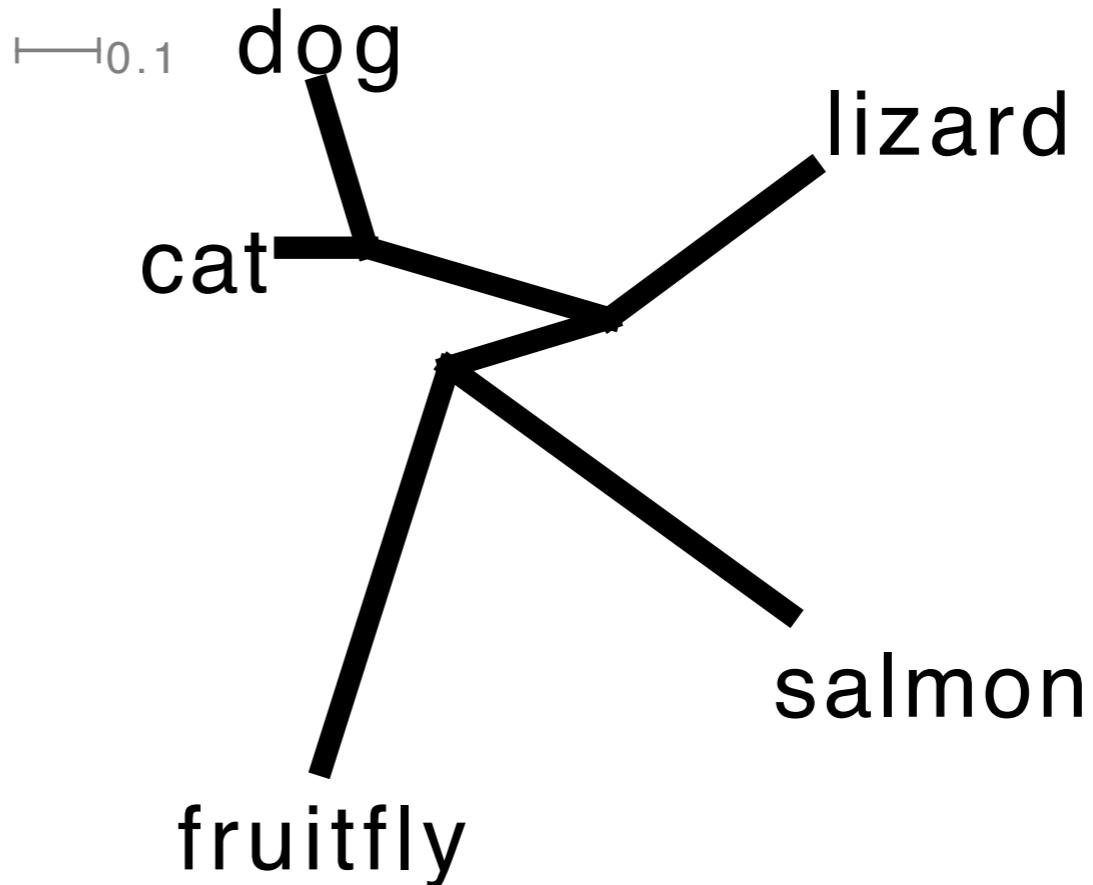


d **not** more closely related to c than it is to a or b



# Unrooted → Rooted

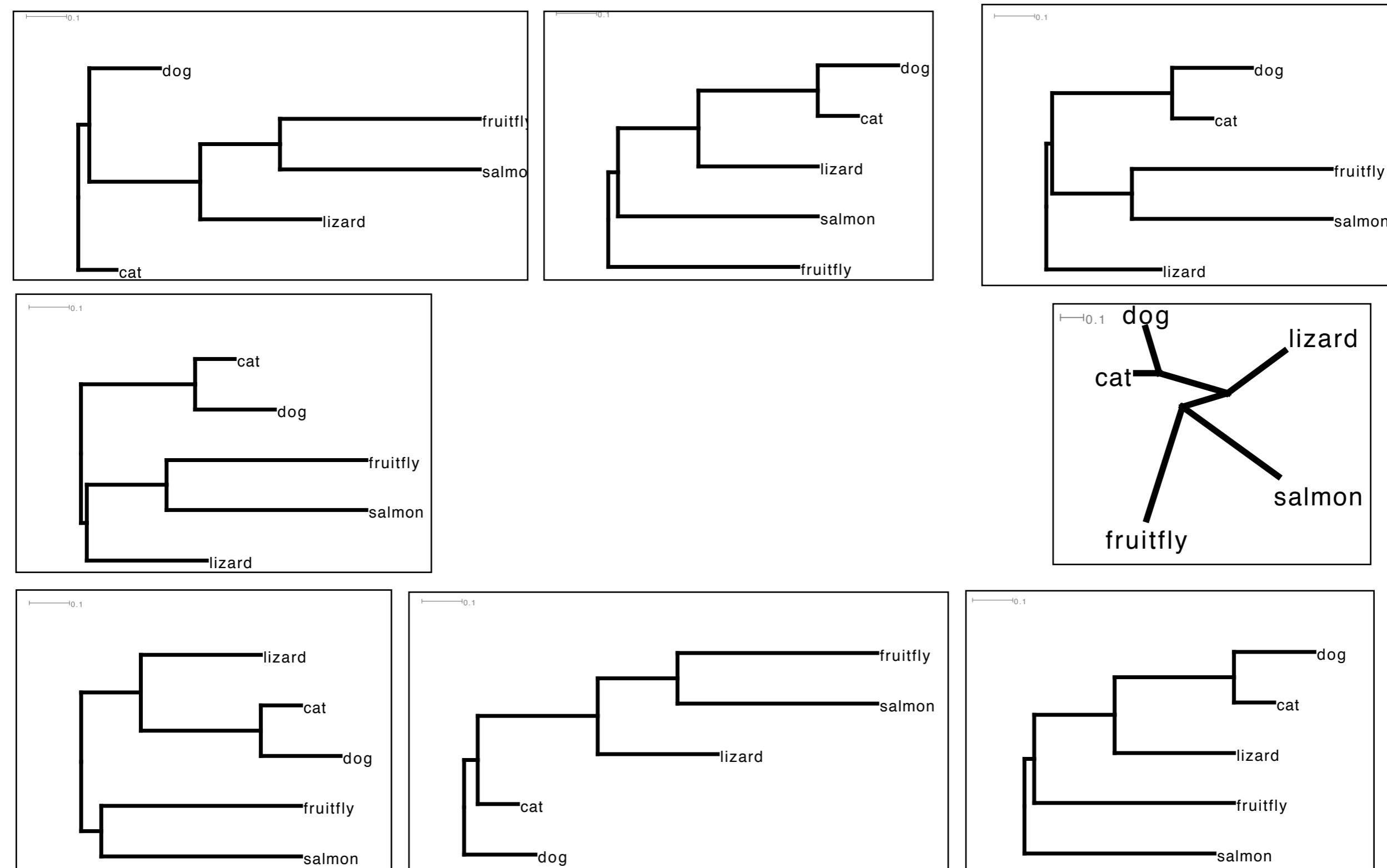
This unrooted tree can be rooted to yield several different rooted topologies.



How many different rooted topologies exist by placing the root on a/an:

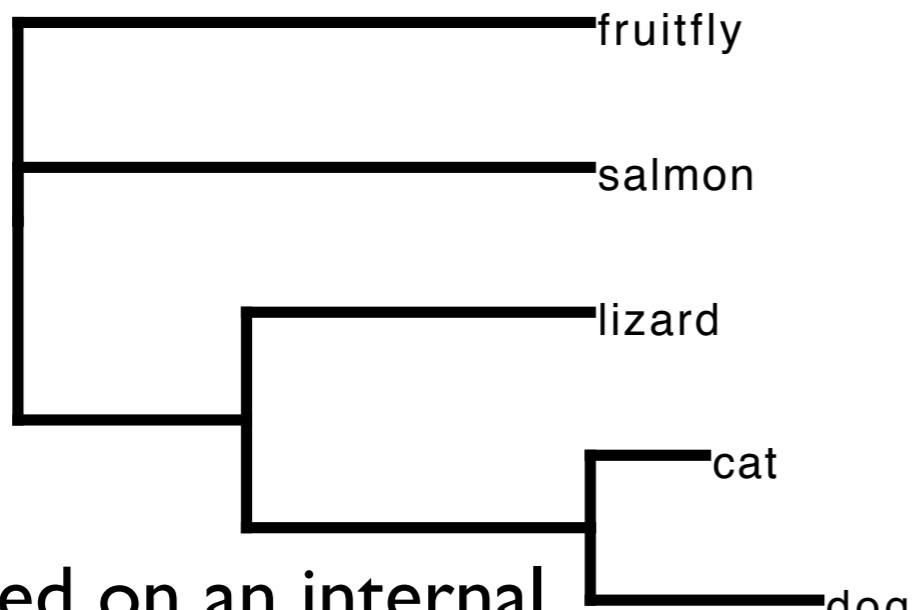
- branch? Draw all of these.
- terminal node? Draw one of these.
- internal node? Draw one of these.

# All Topologies Rooted on a Branch of the Unrooted Tree



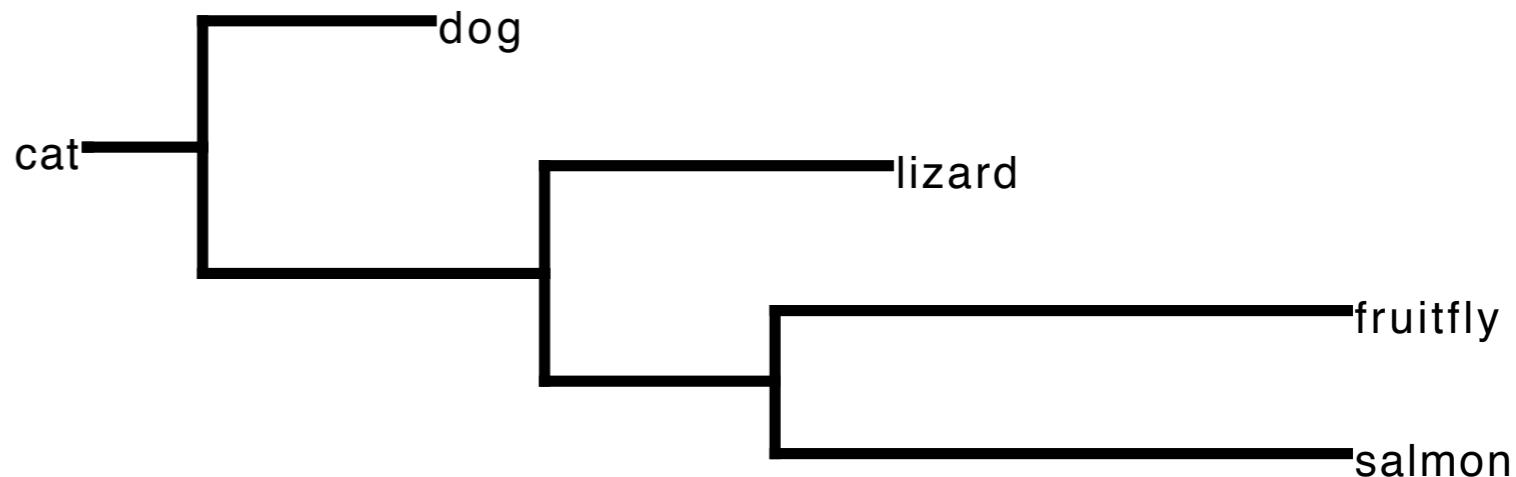
# Roots on Terminal and Internal Nodes

— 0.1



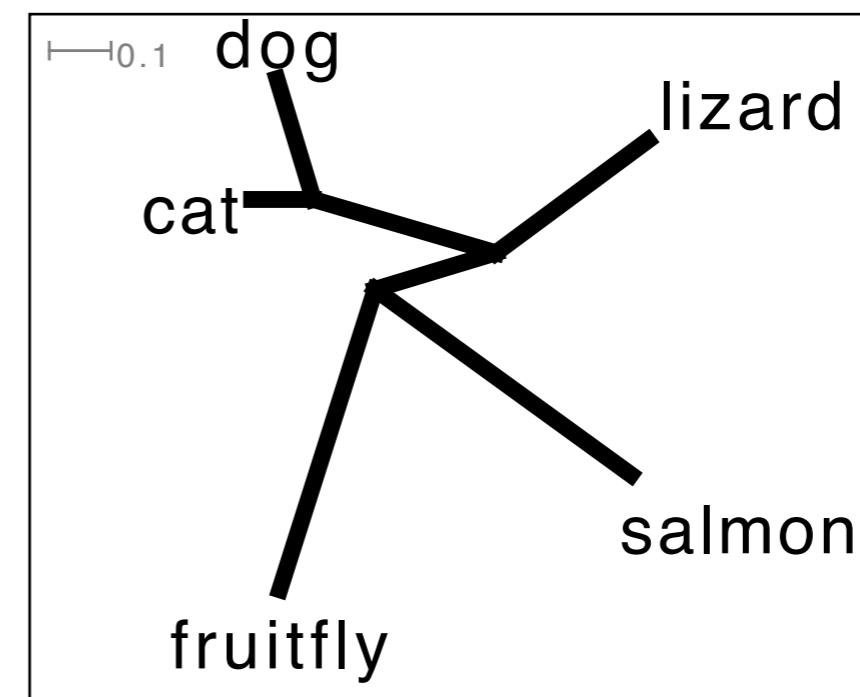
Rooted on an internal  
node of unrooted tree

— 0.1



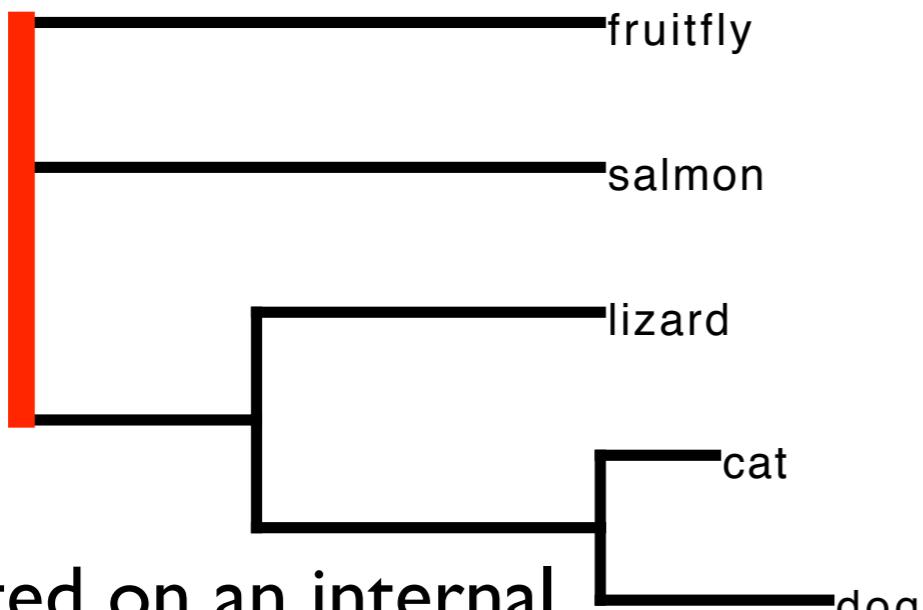
Rooted on a terminal  
node of unrooted tree

On the unrooted tree image to the  
right, label the two nodes on which  
the two above trees are rooted



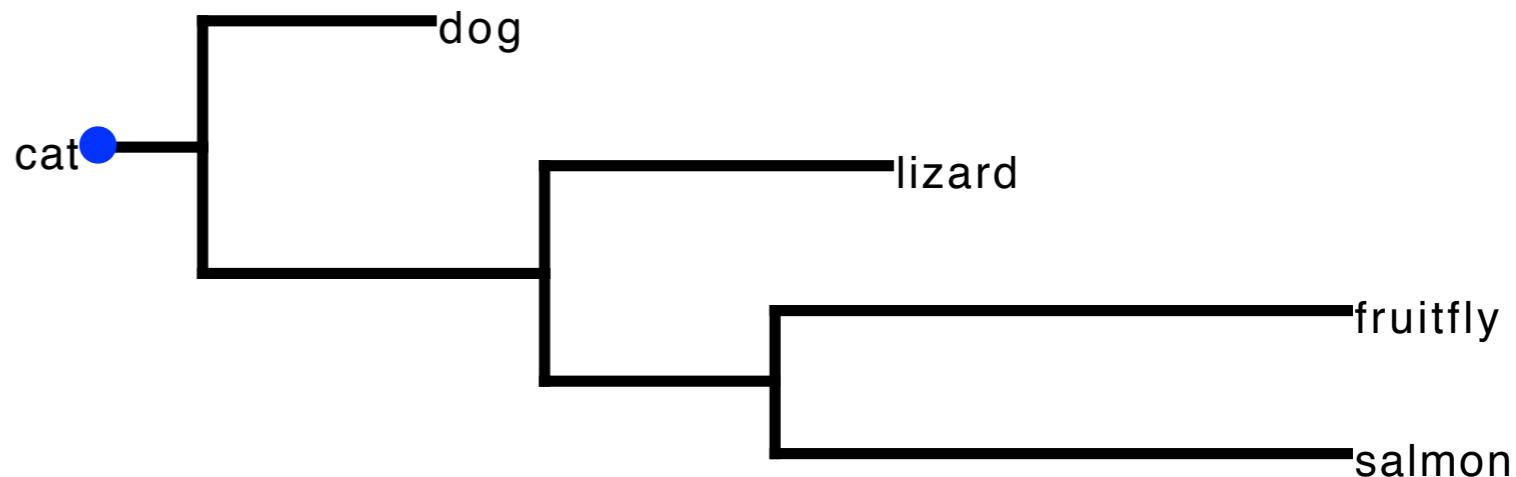
# Roots on Terminal and Internal Nodes

— 0.1



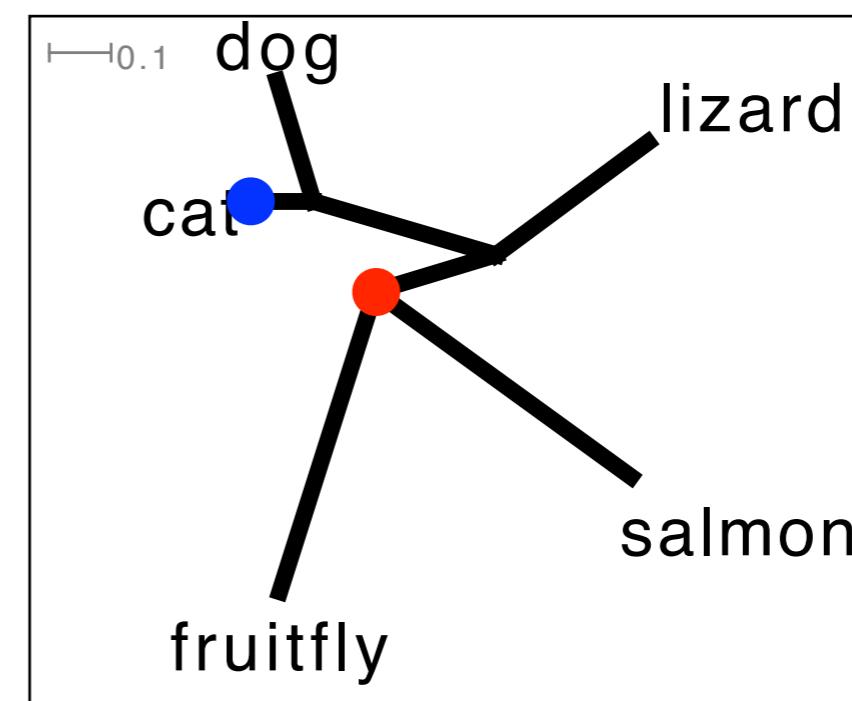
Rooted on an internal  
node of unrooted tree

— 0.1



Rooted on a terminal  
node of unrooted tree

On the unrooted tree image to the  
right, label the two nodes on which  
the two above trees are rooted

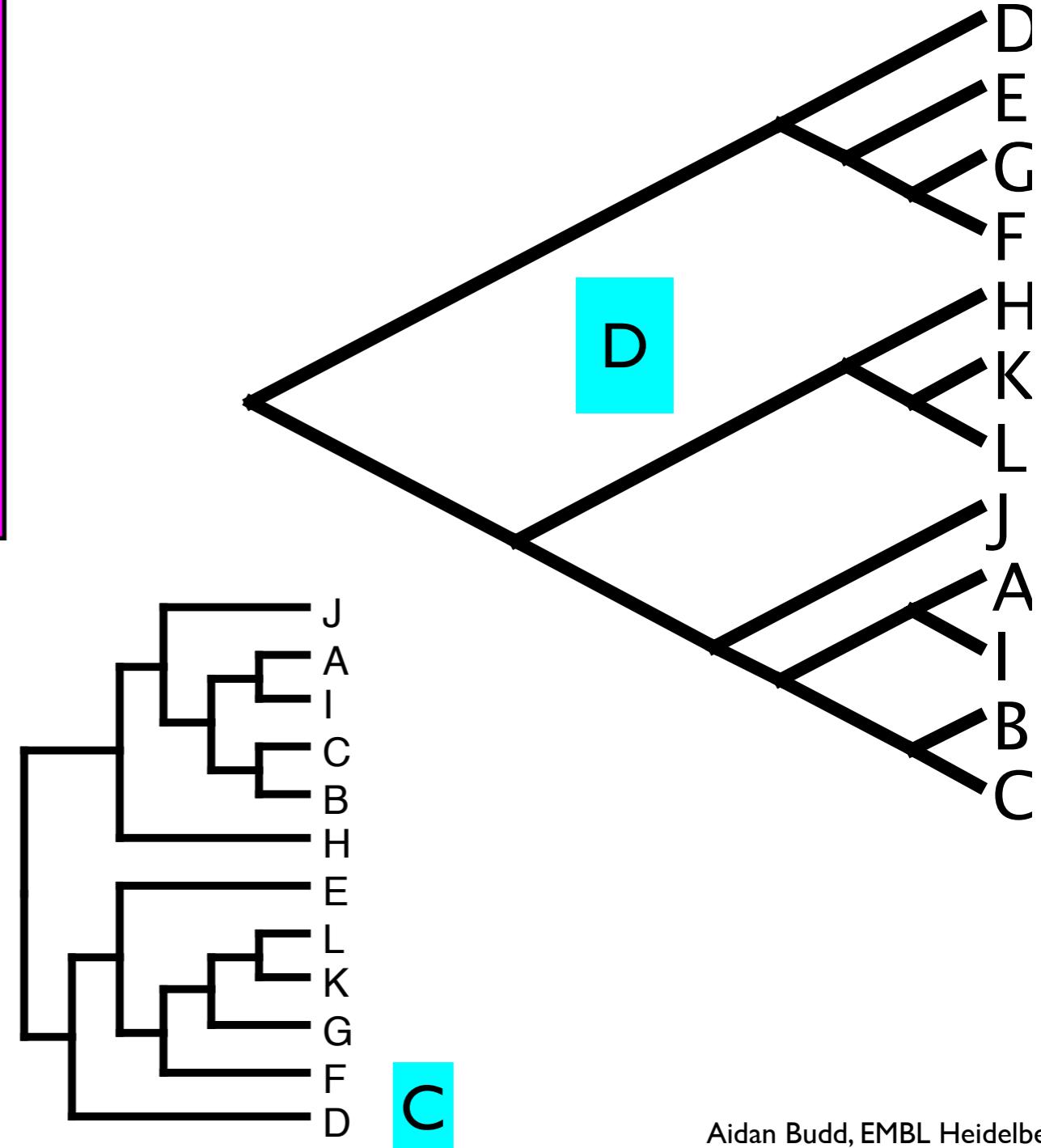
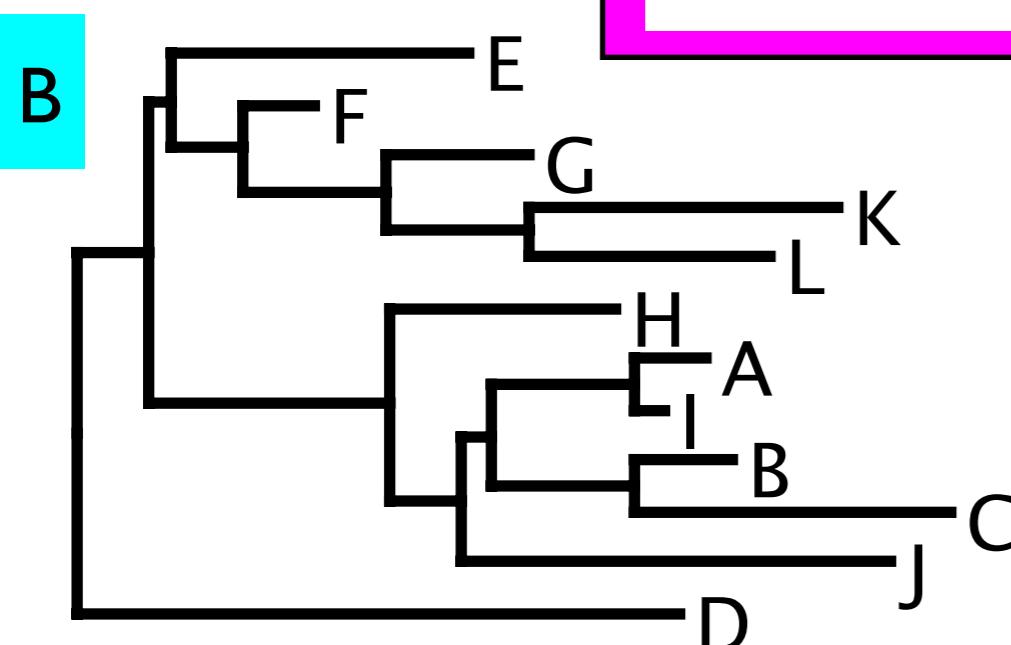
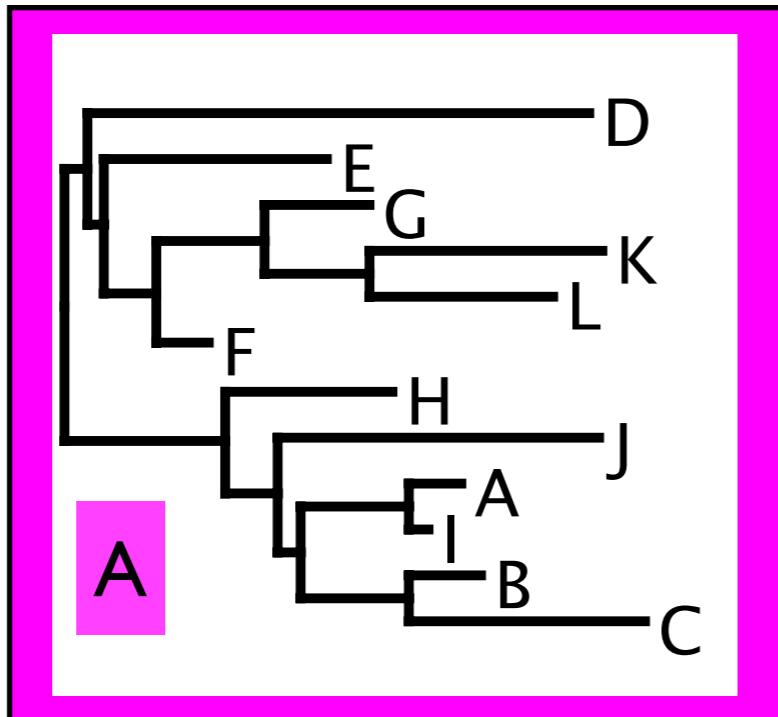


# **NEWICK** format and tree visualisation

Laura Emery

# Quiz: recognise identical topologies

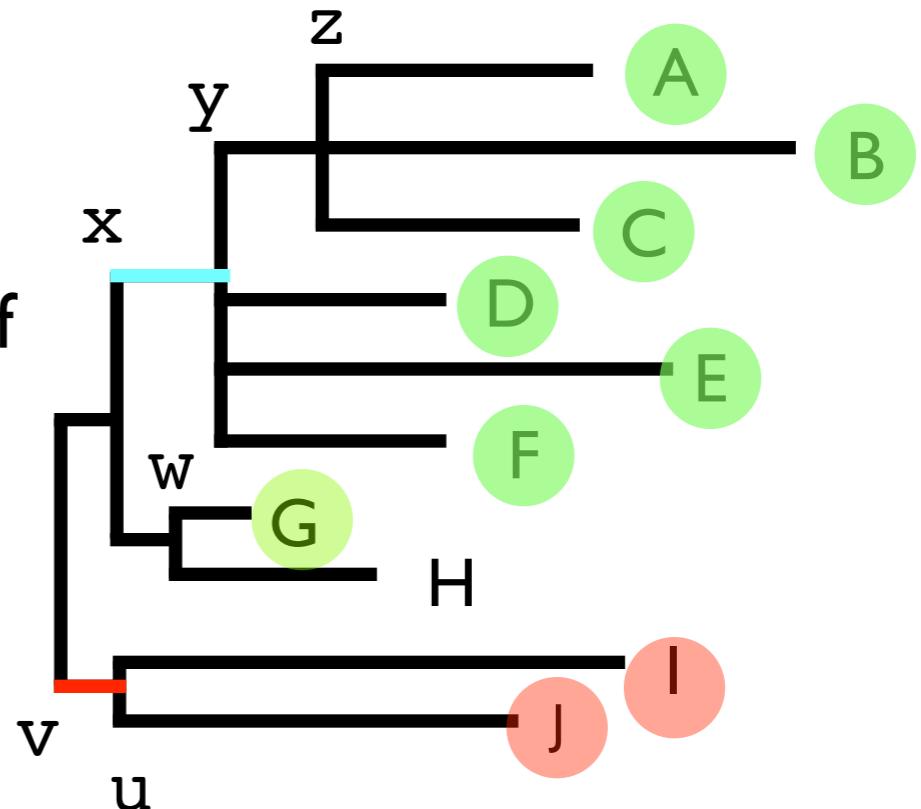
Which of the trees has the same TOPOLOGY and ROOT as tree A?



# Clades

**Clade:**

A set of OTUs that includes all **descendants** of a given internal ancestral/internal branch



**Clades:**

Branch **xy** specifies the clade **ABCDEF**

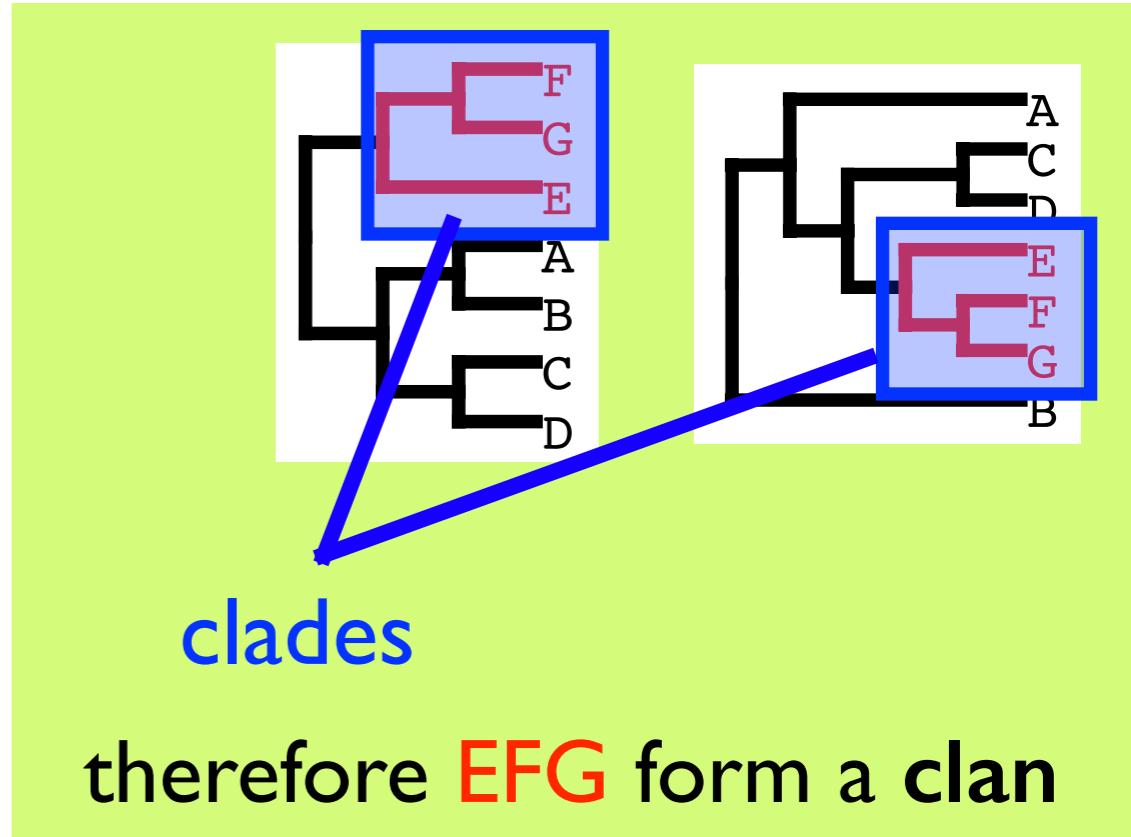
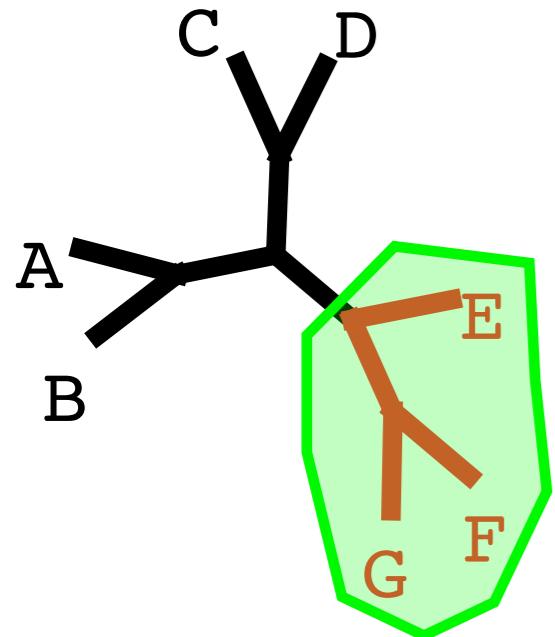
**IJ** is a clade as there is a branch **vu** for which has only **IJ** as its descendants

~~Clades:~~

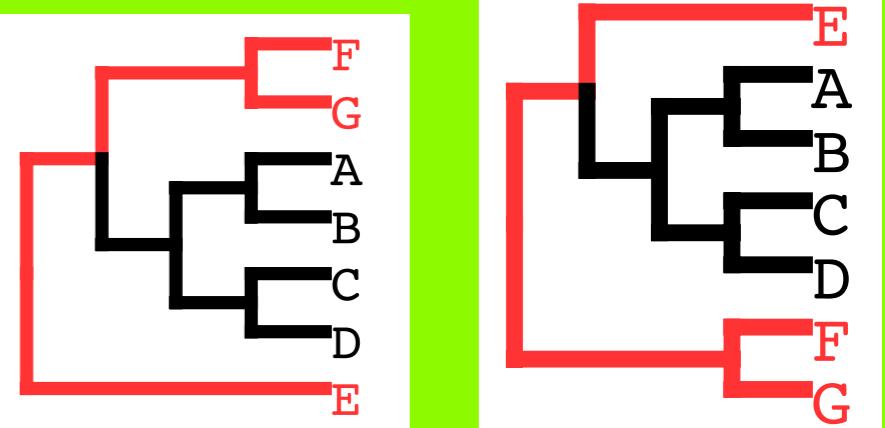
**ABCDEFG** - no branch has ALL and ONLY these taxa as descendants

# Clans

Group of OTUs are a **clan** if there is at least one rooted phylogeny where they form a clade.



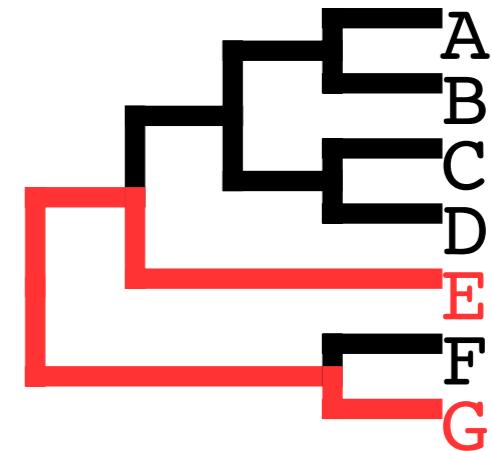
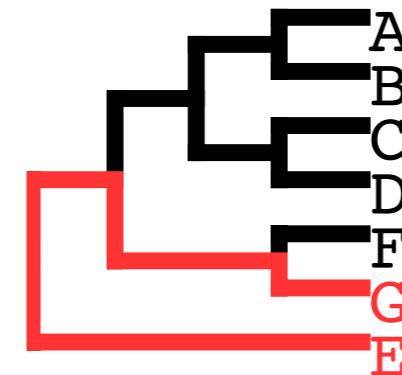
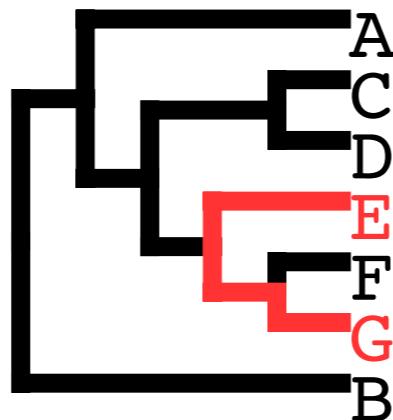
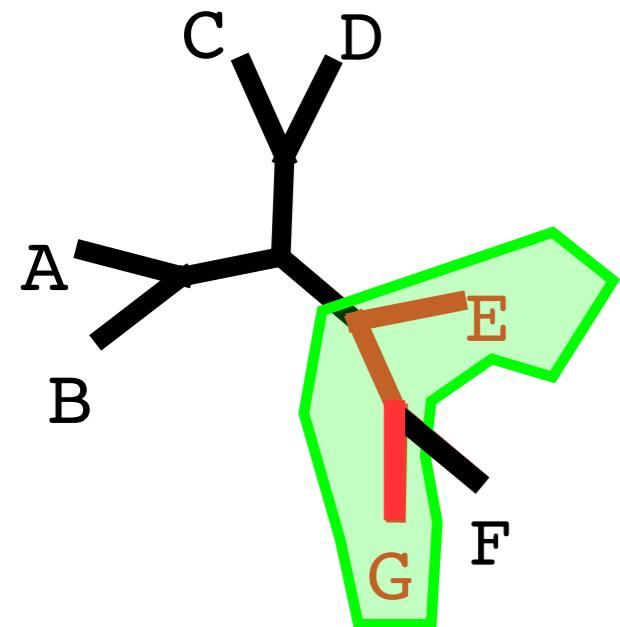
However! Under some rootings **EFG** does not form a clade



Of clades and clans: terms for phylogenetic relationships in unrooted trees.  
Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM.  
Trends Ecol Evol. 2007 Mar;22(3):114-5.  
PMID: 17239486

# Clans

Group of OTUs are a **clan** if there is at least one rooted phylogeny where they form a monophyletic group/clade.

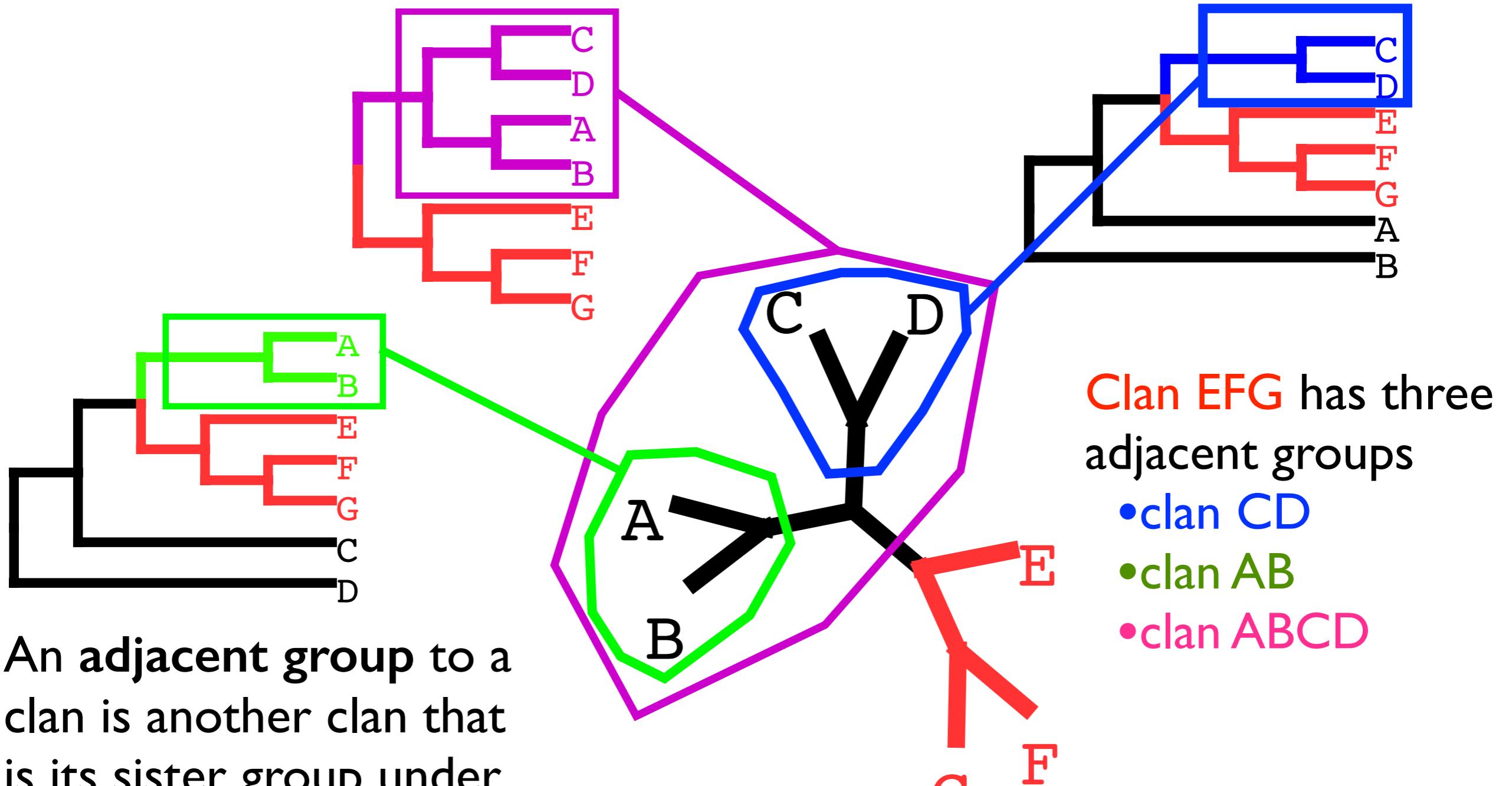


...

NO rooted trees place EG in a monophyletic group  
Therefore **EG is not a clan**

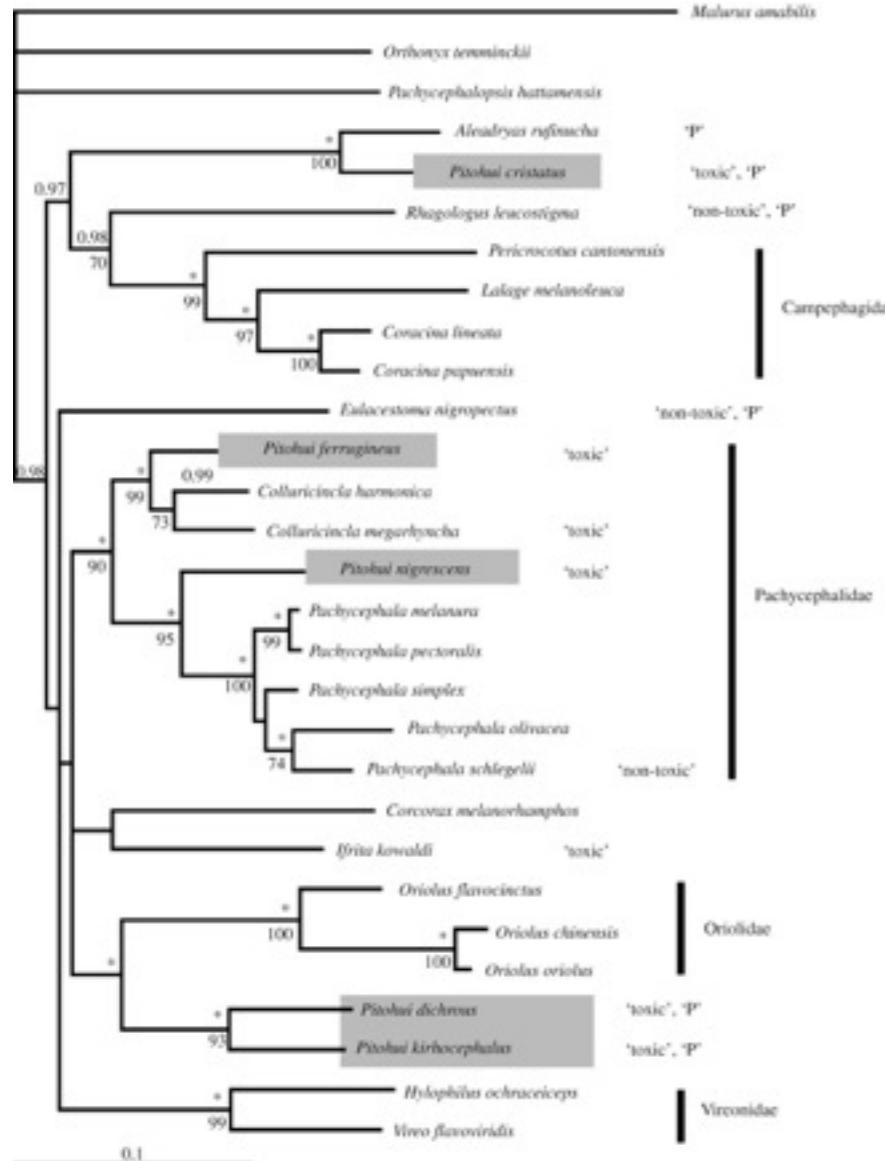
Of clades and clans: terms for phylogenetic relationships in unrooted trees.  
Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM.  
Trends Ecol Evol. 2007 Mar;22(3):114-5.  
PMID: 17239486

# Adjacent Groups



Of clades and clans: terms for phylogenetic relationships in unrooted trees.  
Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM.  
Trends Ecol Evol. 2007 Mar;22(3):114-5.  
PMID: 17239486

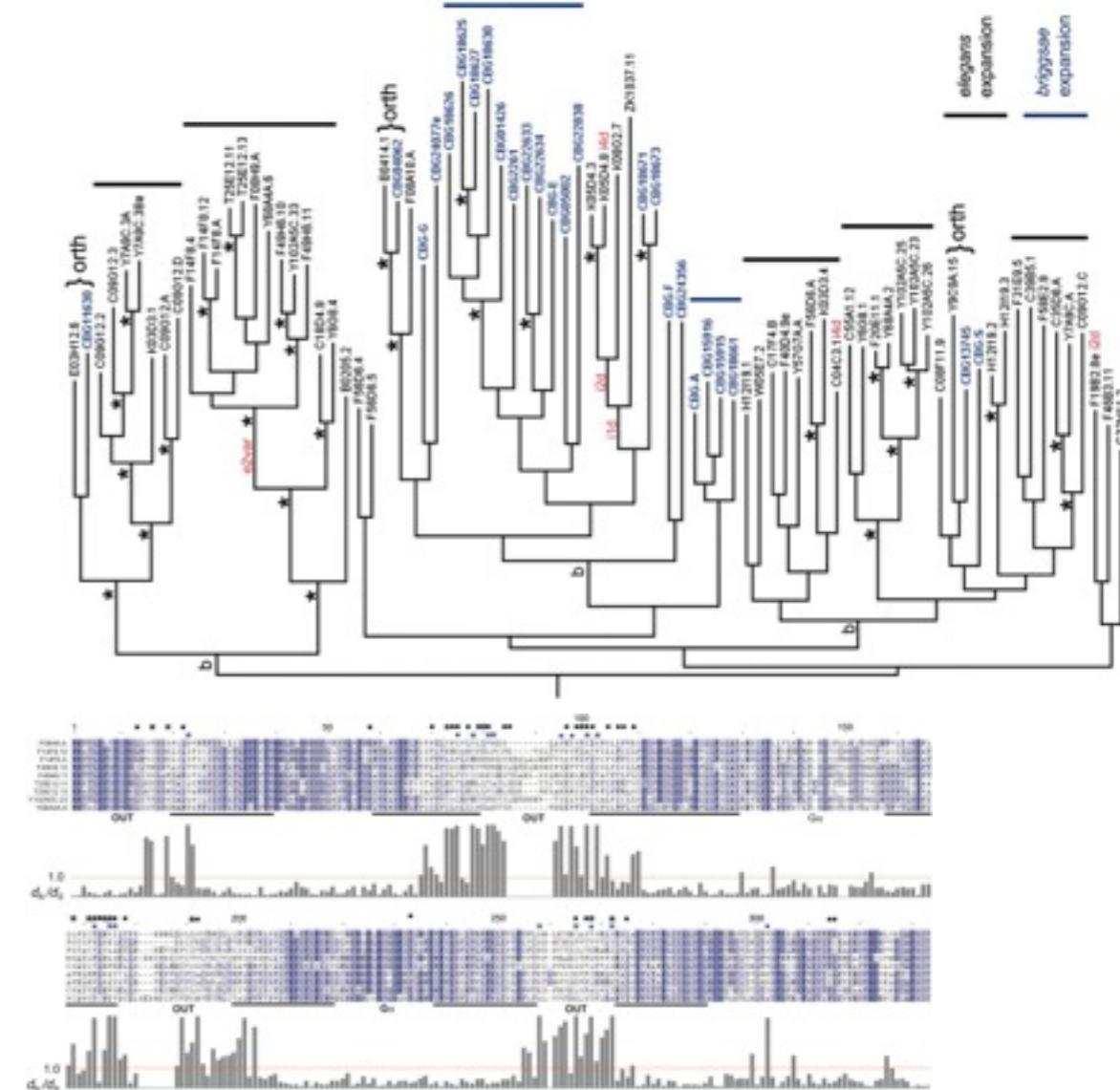
# Unrooted Trees are Sometimes Sufficient



Under all rootings, poisonous members of the order are non-monophyletic

Polyphyletic origin of toxic Pitohui birds suggests widespread occurrence of toxicity in corvoid birds.

Jönsson KA, Bowie RC, Norman JA, Christidis L, Fjeldså J. Biol Lett. 2008 Feb 23;4(1):71-4. PMID: 18055416



Sites of positive selection identified using reversible models of codon substitution

Adaptive evolution in the SRZ chemoreceptor families of *Caenorhabditis elegans* and *Caenorhabditis briggsae*.

Thomas JH, Kelley JL, Robertson HM, Ly K, Swanson WJ. Proc Natl Acad Sci U S A. 2005 Mar 22;102(12):4476-81. PMID: 15761060

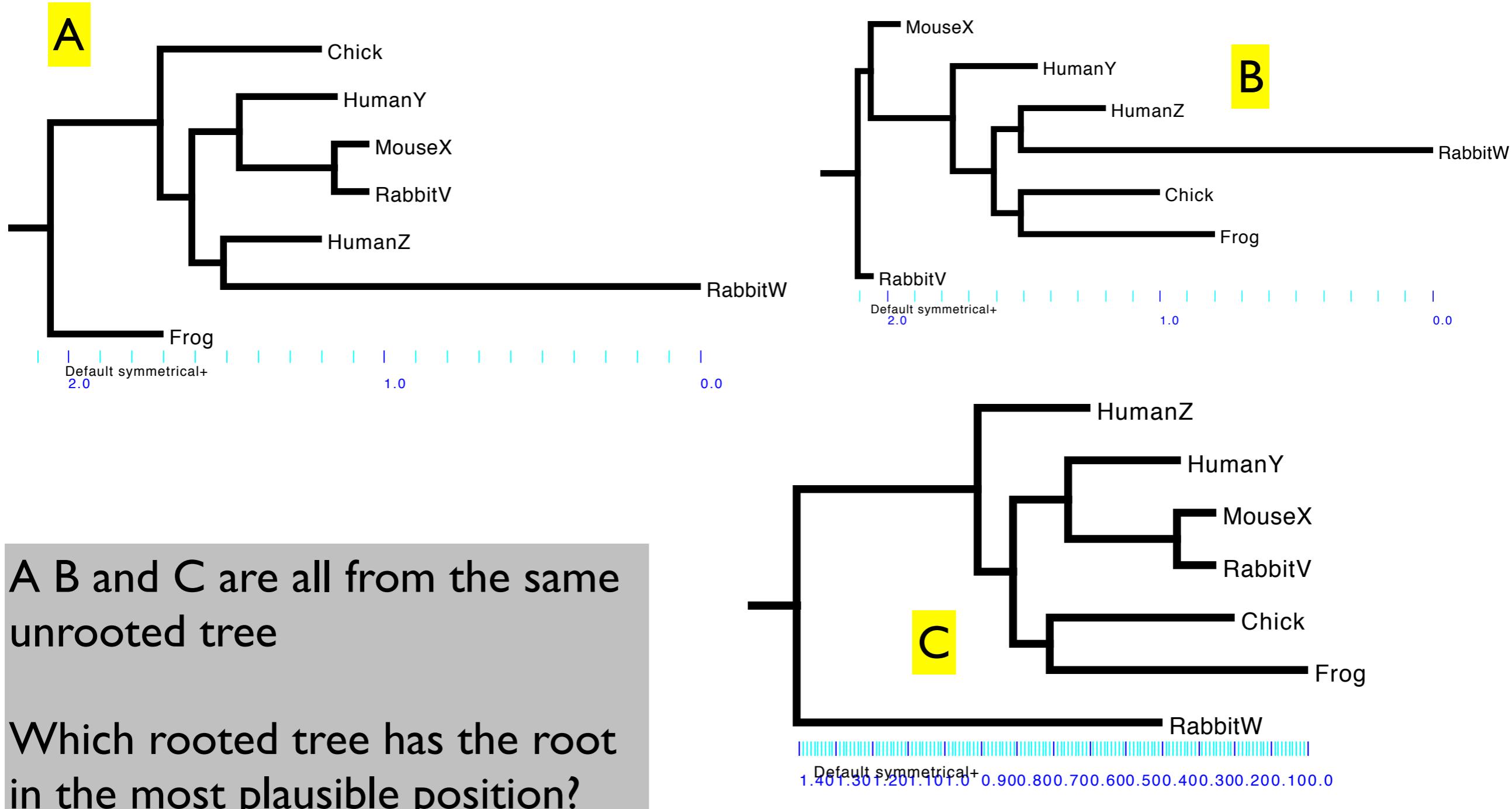
Aidan Budd, EMBL Heidelberg

# Mathematical Models of Sequence Evolution

Sarah Parks

# Rooting Trees/Reconciling Gene/ Species Trees

# Where does the root go?



# Example Phylogeny Estimation Workflow

# Phylogenetic Workflows

---

- Every analysis is different
- There is no “one size fits all” approach
- There are, however, common “phases”/“stages” to many analyses
- We present here **example** analyses to highlight these phases
- Two examples workflows are given
- The first focuses more on common concepts
- The second more on common pragmatic/practical aspects of analyses

# Example Phylogeny Estimation Workflow

## I. Focusing on concepts

# Statistical Estimation of Phylogeny: An Outline

## Statistical paradigm

pose substantive question

develop stochastic model with parameters that, if known, would answer the question.

collect observations that are informative about model parameters.

find the best estimate of parameters conditioned on the observations at hand using some criterion.

## Statistical phylogenetic paradigm

what if the phylogeny of a group of organisms?

develop phylogenetic model with tree (and branch lengths) and a Markov model describing how traits change over tree.

construct a data matrix (e.g., of DNA sequences) sampled from the group of organisms.

find the best estimate of phylogeny using maximum likelihood criterion or Bayesian inference criterion.

Huelsenbook

Brian R. Moore, UC Davis

# Example Phylogeny Estimation Workflow

---

1. Pose a substantive question
2. Build a model involving parameters that, if known, could answer the question
3. Collect observations informative about the model parameter(s)
4. Find best estimate(s) of the parameter(s), conditioned on these observations
5. Estimate sampling/random error associated with parameter estimate [this formulation of the problem inspired by Brian R Moore's slides - thanks Brian!]
6. Answer your question using these parameter estimates

# Example Phylogeny Estimation Workflow

---

1. Pose a substantive question
2. Build a model involving parameters that, if known, could answer the question
3. Collect observations informative about the model parameter(s)
4. Find best estimate(s) of the parameter(s), conditioned on these observations
5. Estimate sampling/random error associated with parameter estimate
6. Answer your question using these parameter estimates

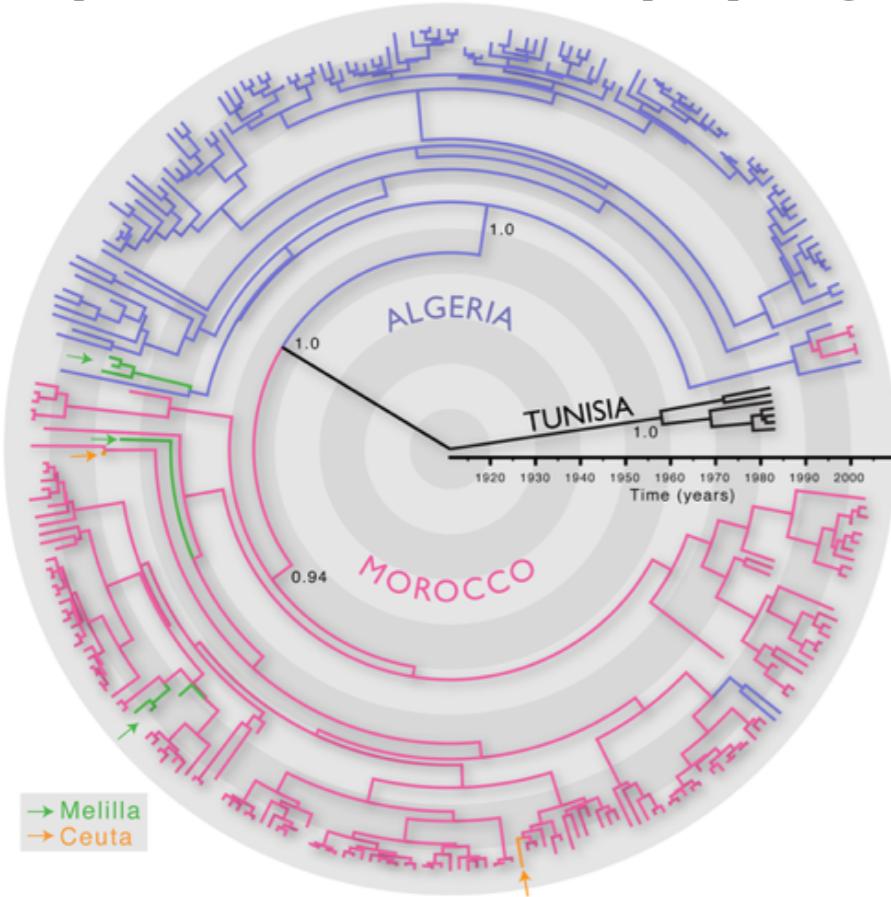
# Example Phylogeny Estimation Workflow

## I. Pose a substantive question

For example:

Can we identify factors promoting rabies virus transmission that could be addressed via public-health measures?

Crucially: a substantive question such that **knowledge** (or rather estimation) of parameters in a phylogenetic model can inform our answer



In this case, we looked earlier at how the topology parameter (i.e. set of "relatedness statements" estimated from the data) of a phylogenetic model the evolution of dog rabies viruses from north Africa informs our belief in the significance of certain factors in the spread of the virus

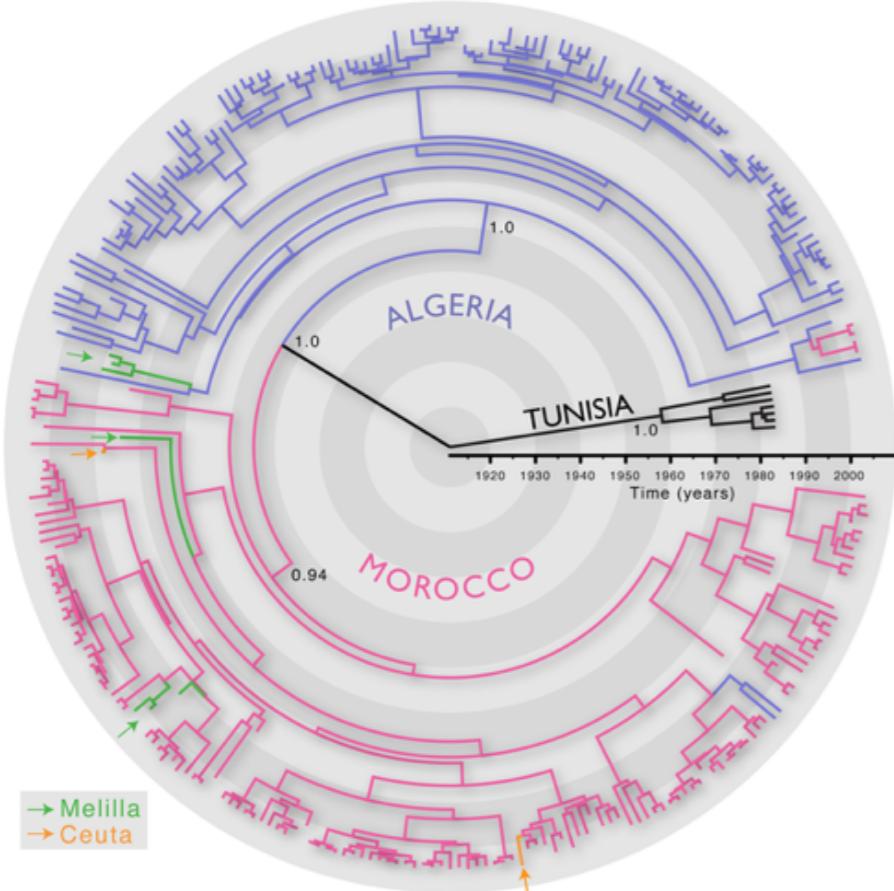
# Example Phylogeny Estimation Workflow

## I. Pose a substantive question

For example:

Can we identify factors promoting rabies virus transmission that could be addressed via public-health measures?

Crucially: a substantive question such that **knowledge** (or rather estimation) of parameters in a phylogenetic model can inform our answer



Reformulating/recasting the question in terms of such parameters can help guide our analysis (e.g. helping us decide which data to collect)

For example, in this case:

Are virus samples that are closely located, but in different countries, relatively closely or distantly related to each other?

# Example Phylogeny Estimation Workflow

---

1. Pose a substantive question
2. Build a model involving parameters that, if known, could answer the question
3. Collect observations informative about the model parameter(s)
4. Find best estimate(s) of the parameter(s), conditioned on these observations
5. Estimate sampling/random error associated with parameter estimate
6. Answer your question using these parameter estimates

## 2. Build a model involving parameters that, if known, could answer the question

Olivier Gascuel – Phylogenetic models – ISCB-ASBCB Casablanca 2013



### The full probabilistic model

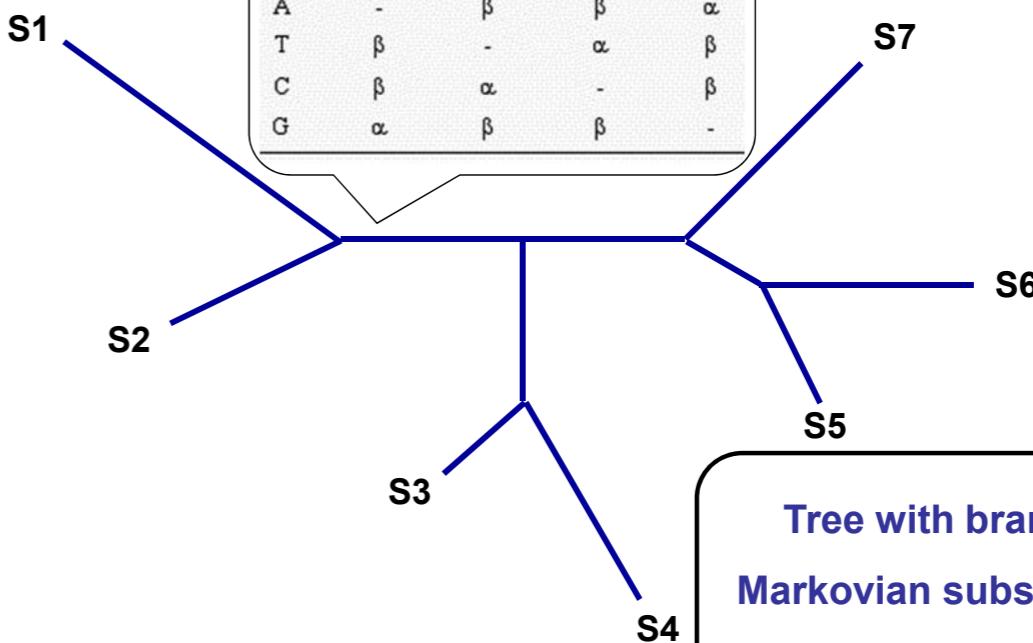
- A tree topology (to be estimated,  $n^n$ )
- Branch lengths (to be estimated,  $2n-3$ )
- A substitution model (to be (partly) estimated, 1, 3, 4, ...208 ...)
- A distribution of site rates (to be estimated, 1, 2, ...)

Olivier Gascuel – Phylogenetic models – ISCB-ASBCB Casablanca 2013



### The full probabilistic model

	A	T	C	G
A	-	$\beta$	$\beta$	$\alpha$
T	$\beta$	-	$\alpha$	$\beta$
C	$\beta$	$\alpha$	-	$\beta$
G	$\alpha$	$\beta$	$\beta$	-



Tree with branch lengths  
Markovian substitution model  
Site rate model

## 2. Build a model involving parameters that, if known, could answer the question

Olivier Gascuel – Phylogenetic models – ISCB-ASCB Casablanca 2013

The full probabilistic model

- A tree topology (to be estimated)
- Branch lengths (to be estimated)
- A substitution model (to be estimated)
- A distribution of site rates (to be estimated)

Olivier Gascuel – Phylogenetic models – ISCB-ASCB Casablanca 2013

The full probabilistic model

A	T	C	G
A	-	$\beta$	$\alpha$
T	$\beta$	-	$\beta$
C	$\beta$	$\alpha$	-
G	$\alpha$	$\beta$	-

Tree with branch lengths  
Markovian substitution model  
Site rate model

For north African rabies analysis, a parameter of interest was the rooted tree topology

But could also be other parameters e.g. identifying positive selection (omega parameter in certain codon-based substitution models above a particular value) that are of most interest

# Example Phylogeny Estimation Workflow

---

1. Pose a substantive question
2. Build a model involving parameters that, if known, could answer the question
3. **Collect observations informative about the model parameter(s)**
4. Find best estimate(s) of the parameter(s), conditioned on these observations
5. Estimate sampling/random error associated with parameter estimate
6. Answer your question using these parameter estimates

# Example Phylogeny Estimation Workflow

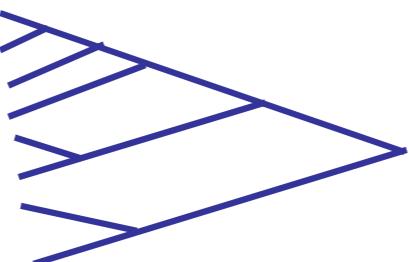
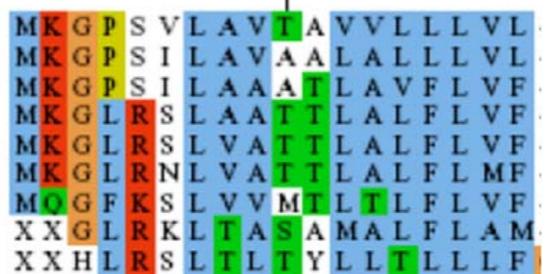
## 3. Collect observations informative about the model parameter(s)

Olivier Gascuel – Phylogenetic models – ISCB-ASBCB Casablanca 2013



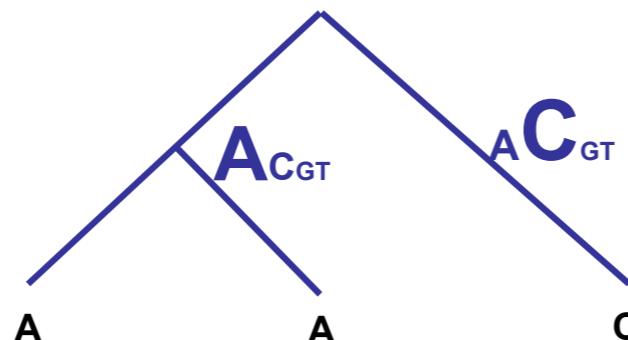
### Modeling sequence evolution: standard assumptions

MOUSE  
RAT  
RABBIT  
HUMAN  
DOG  
ELEPHANT  
COW  
CHICKEN  
FUGU



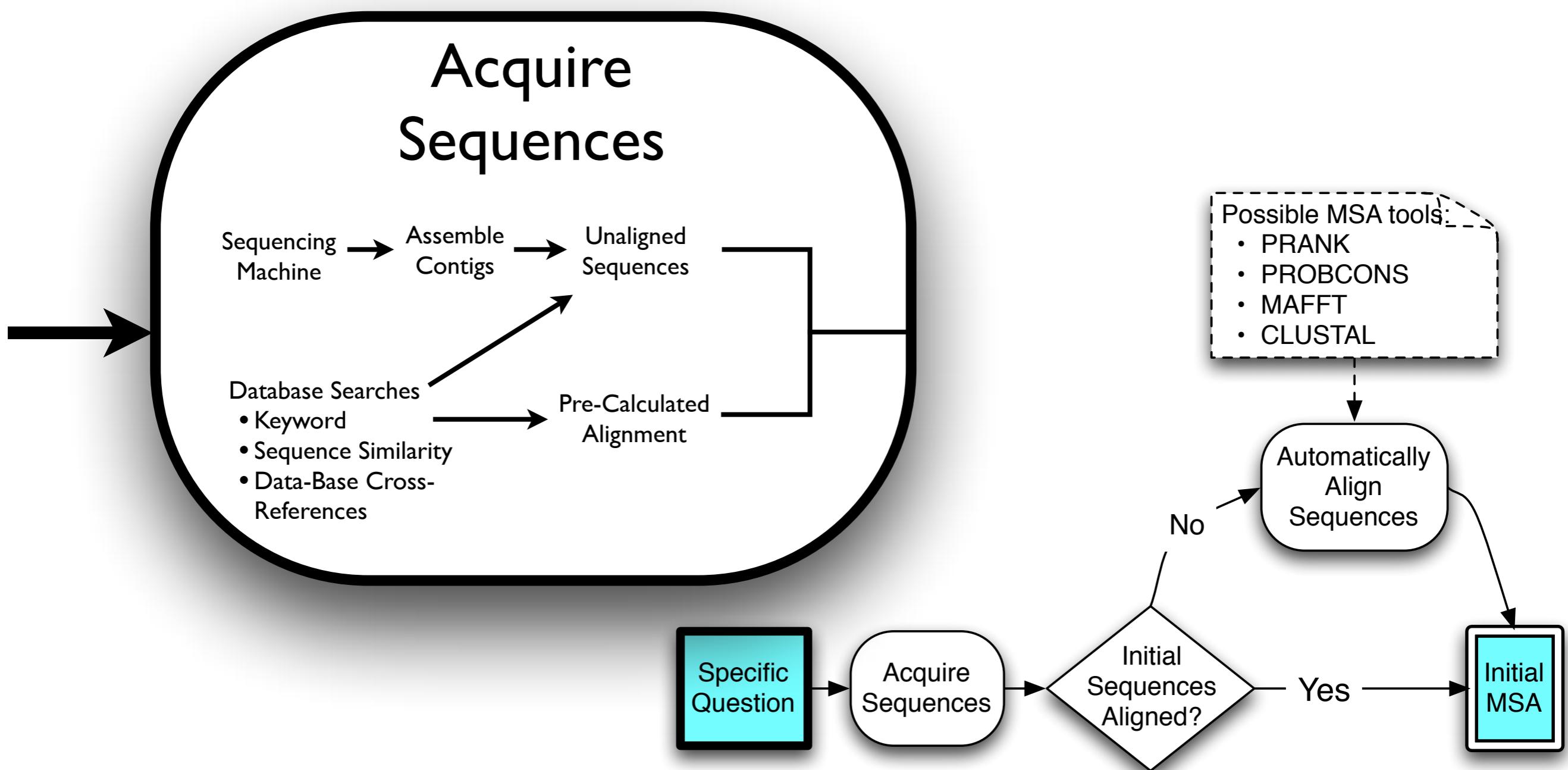
e.g. build a multiple sequences alignment of north African dog rabies sequences

We aim at explaining the data (alignment) using a probabilistic scenario of the evolution of each of the sites along a phylogeny



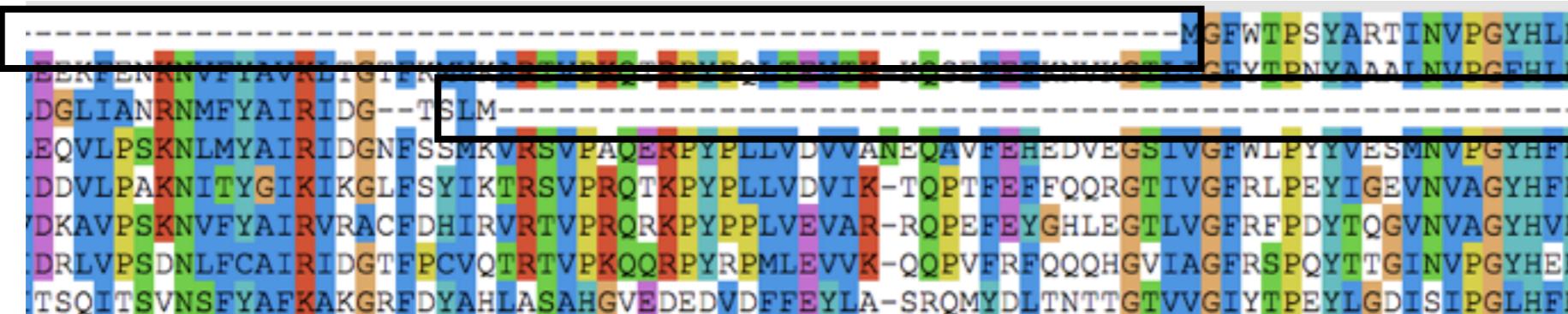
# Example Phylogeny Estimation Workflow

## 3. Collect observations informative about the model parameter(s)



# Example Phylogeny Estimation Workflow

## Unusual Sequences



# Short/fragmented sequences



With CLUSTALX “”Quality”->”Show Low-Scoring Segments” switched on

# Unusual pattern of "conservation"

# Example Phylogeny Estimation Workflow

---

1. Pose a substantive question
2. Build a model involving parameters that, if known, could answer the question
3. Collect observations informative about the model parameter(s)
4. Find best estimate(s) of the parameter(s), conditioned on these observations
5. Estimate sampling/random error associated with parameter estimate
6. Answer your question using these parameter estimates

# Example Phylogeny Estimation Workflow

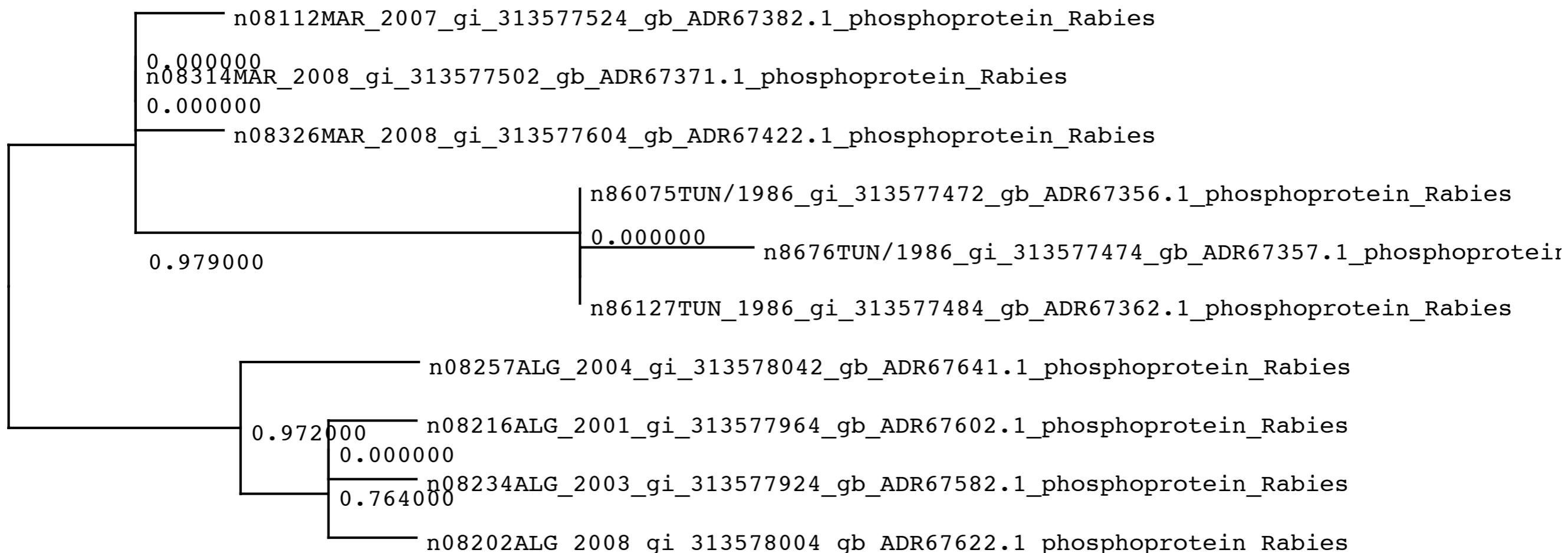
---

1. Pose a substantive question
2. Build a model involving parameters that, if known, could answer the question
3. Collect observations informative about the model parameter(s)
4. Find best estimate(s) of the parameter(s), conditioned on these observations
5. Estimate sampling/random error associated with parameter estimate
6. Answer your question using these parameter estimates

# Example Phylogeny Estimation Workflow

## 6. Answer your question using these parameter estimates

H<sub>0</sub>.0010



# Example Phylogeny Estimation Workflow

---

## Demo and Exercises

We'll follow a demonstration, and you'll have a chance to try this kind of phylogenetic workflow yourself, using the

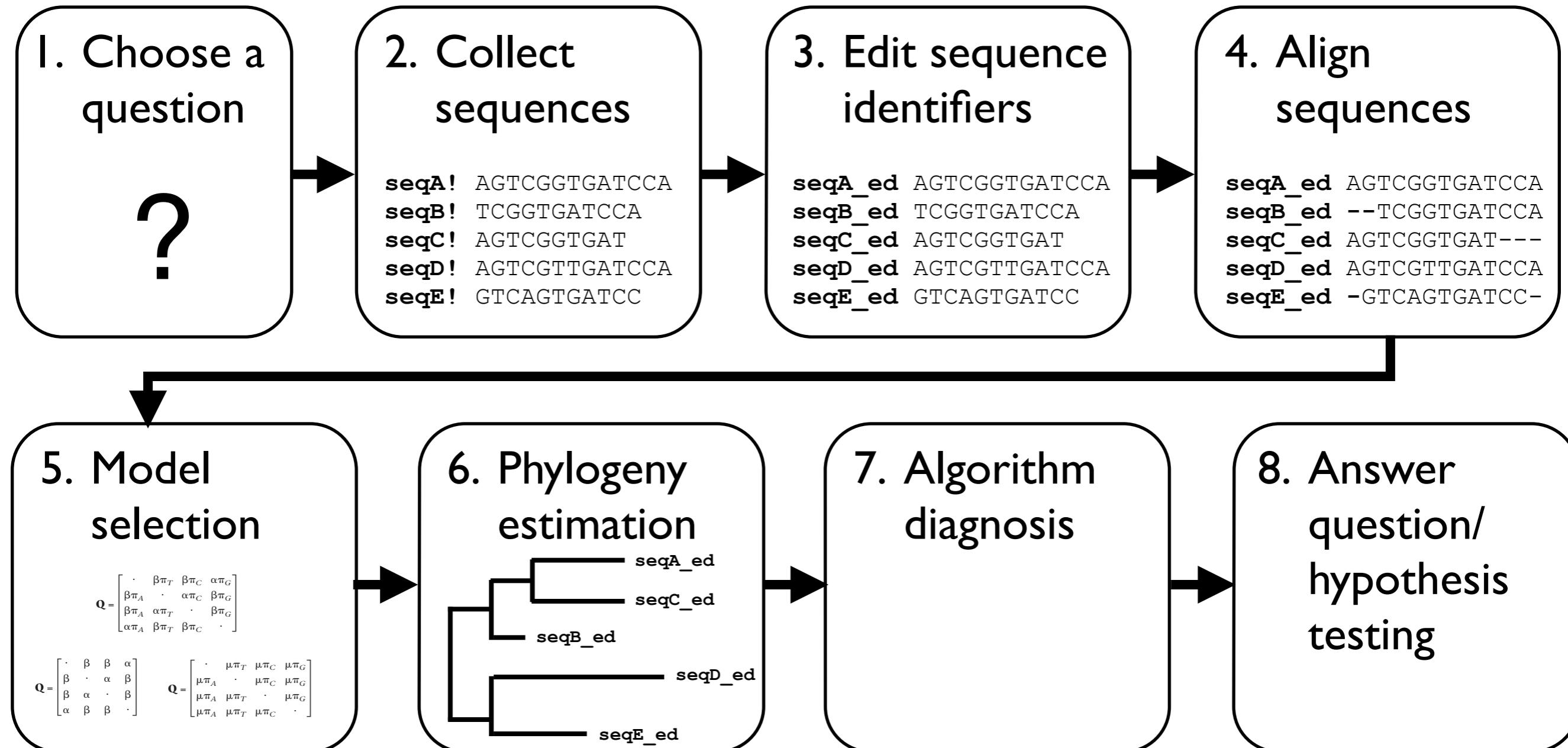
- "Conceptual" **demonstration** with North African dog rabies viruses
- "Conceptual" **exercise** with Louisiana gastroenterologist example

described in this HTML document **interpretingPhylogeniesCrete2014.html**

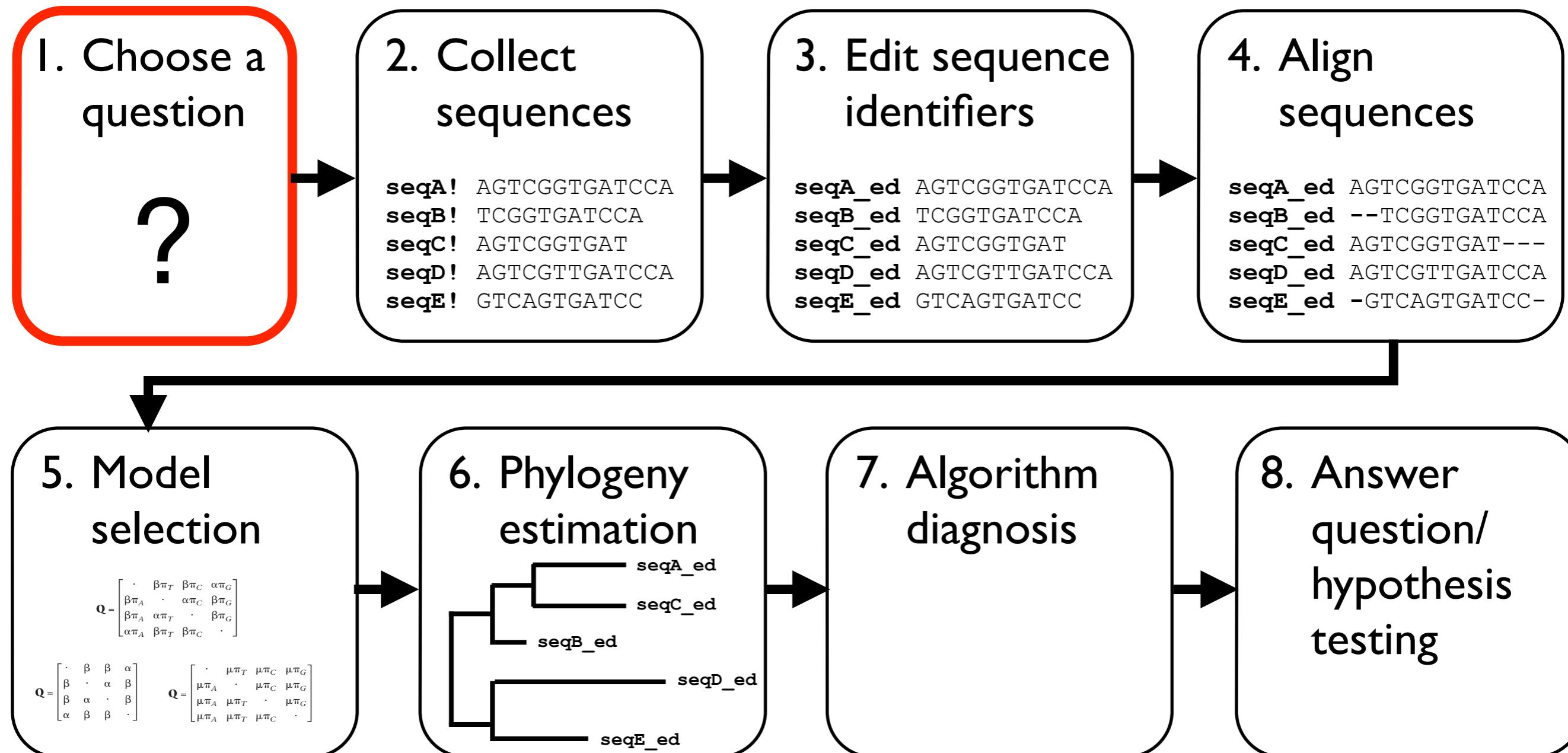
# Example Phylogeny Estimation Workflow

## I. Focusing on practical/pragmatic issues

# Example Phylogeny Estimation Workflow

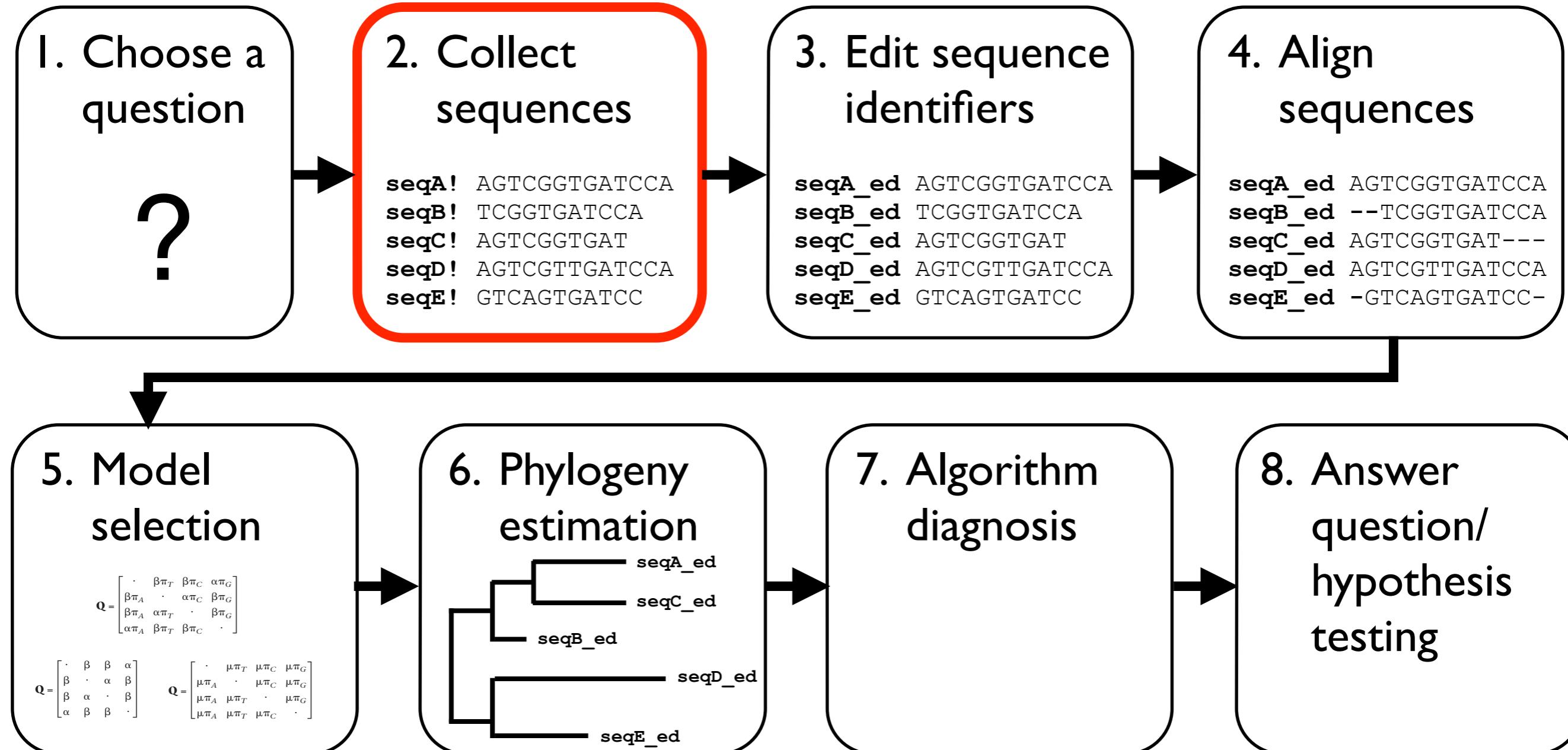


# Example Phylogeny Estimation Workflow



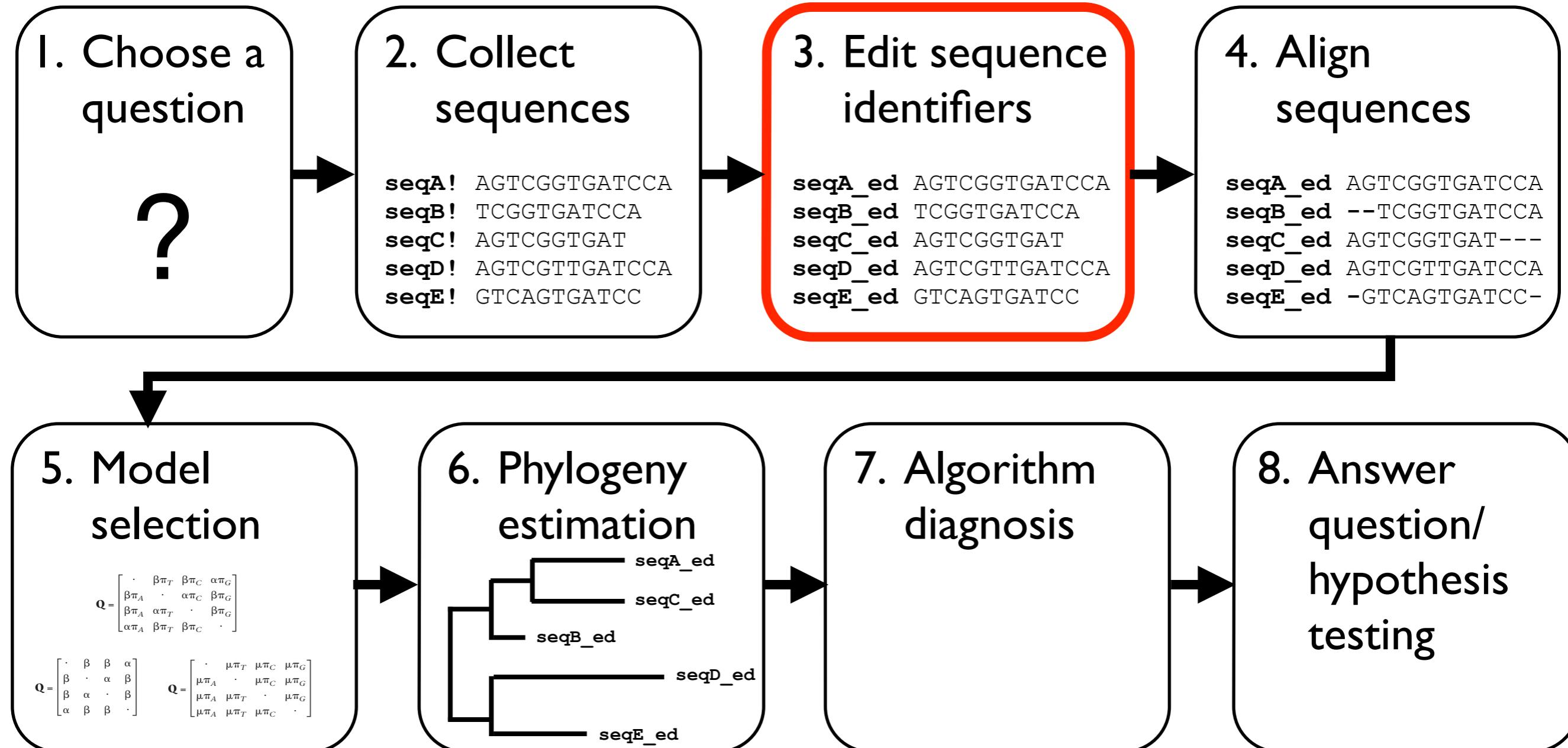
I. Choose a question: me, today, using your brain

# Example Phylogeny Estimation Workflow



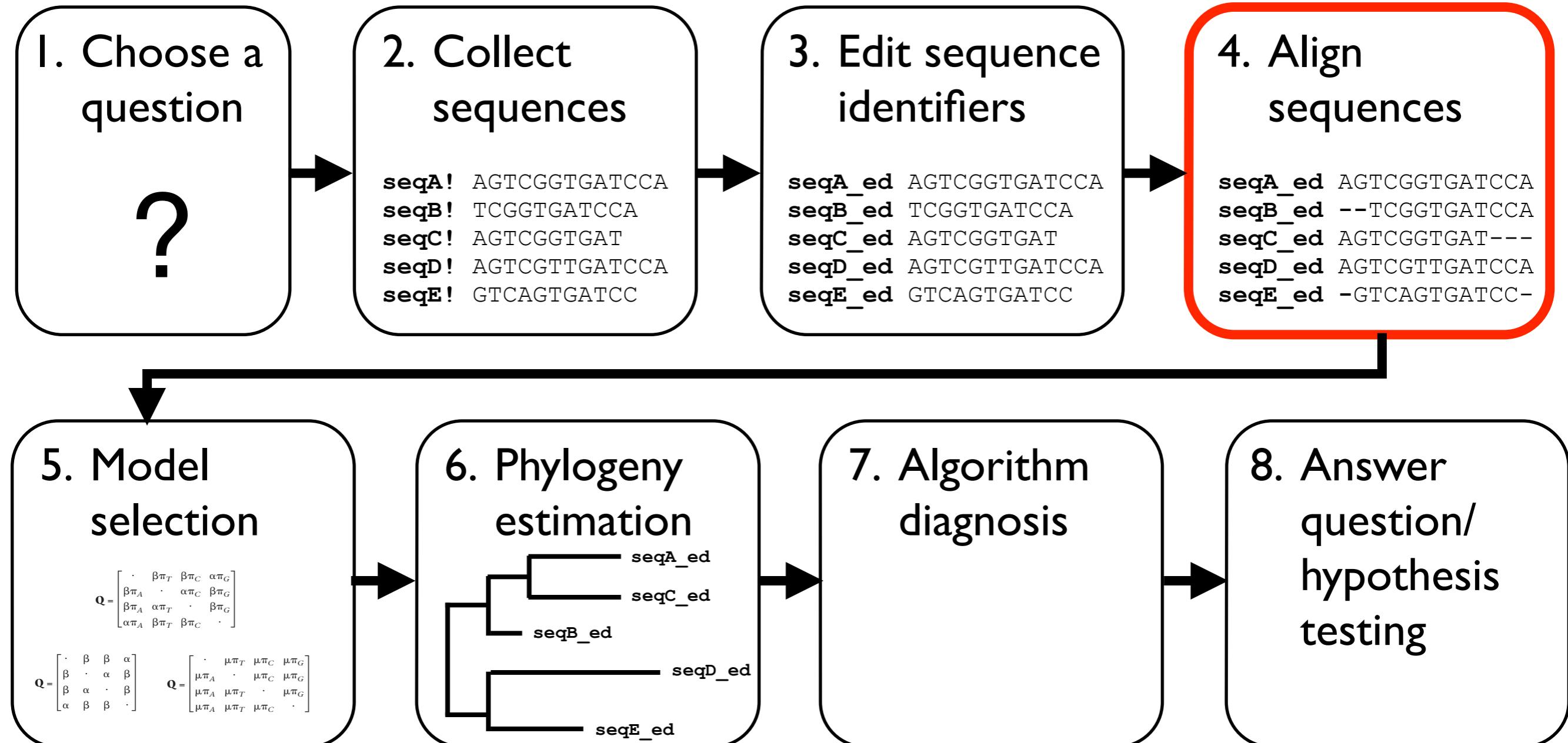
2. Collect sequences: Stephen, DATES, BLAST, XXXXX

# Example Phylogeny Estimation Workflow



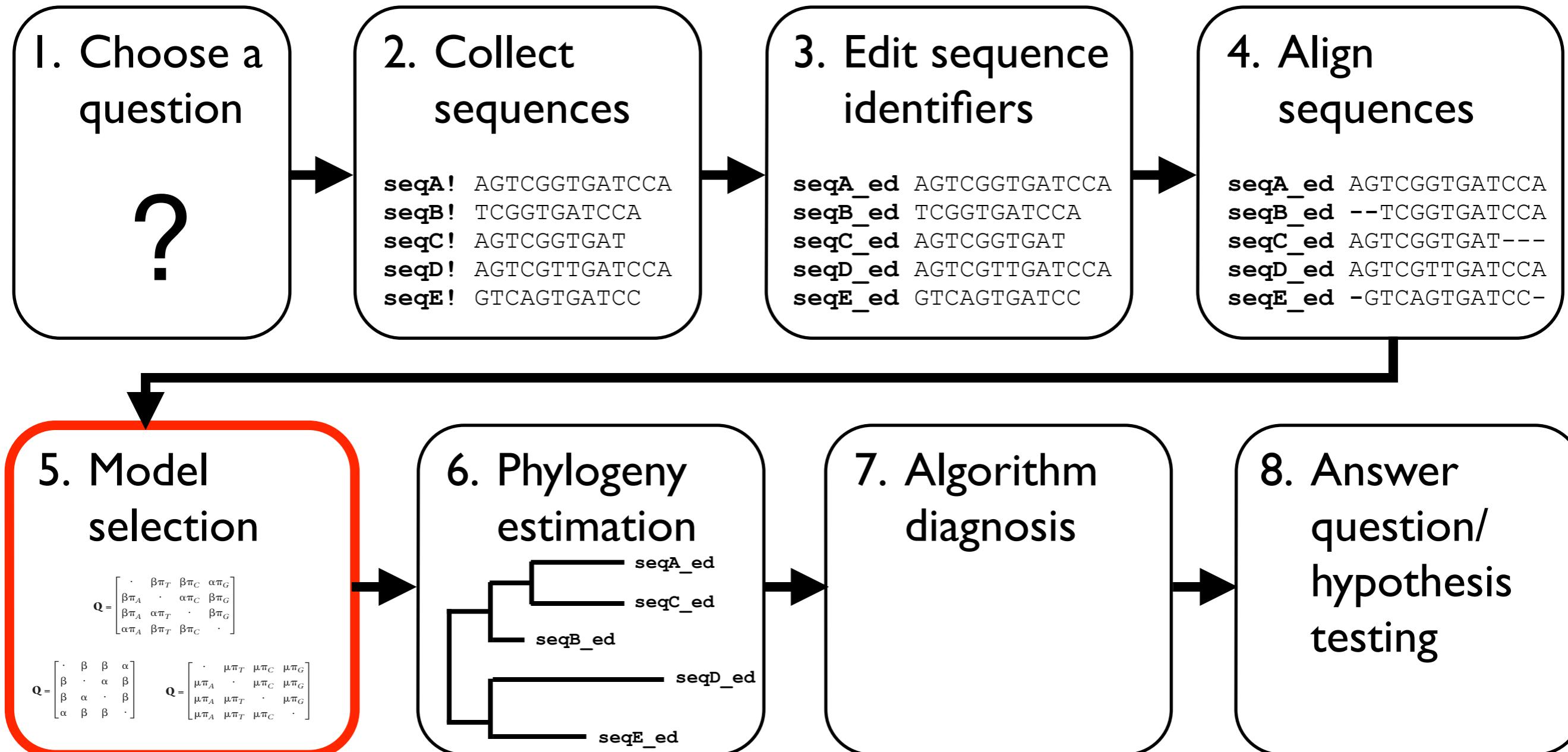
3. Edit sequence identifiers: me, today, text editors or scripting tools

# Example Phylogeny Estimation Workflow



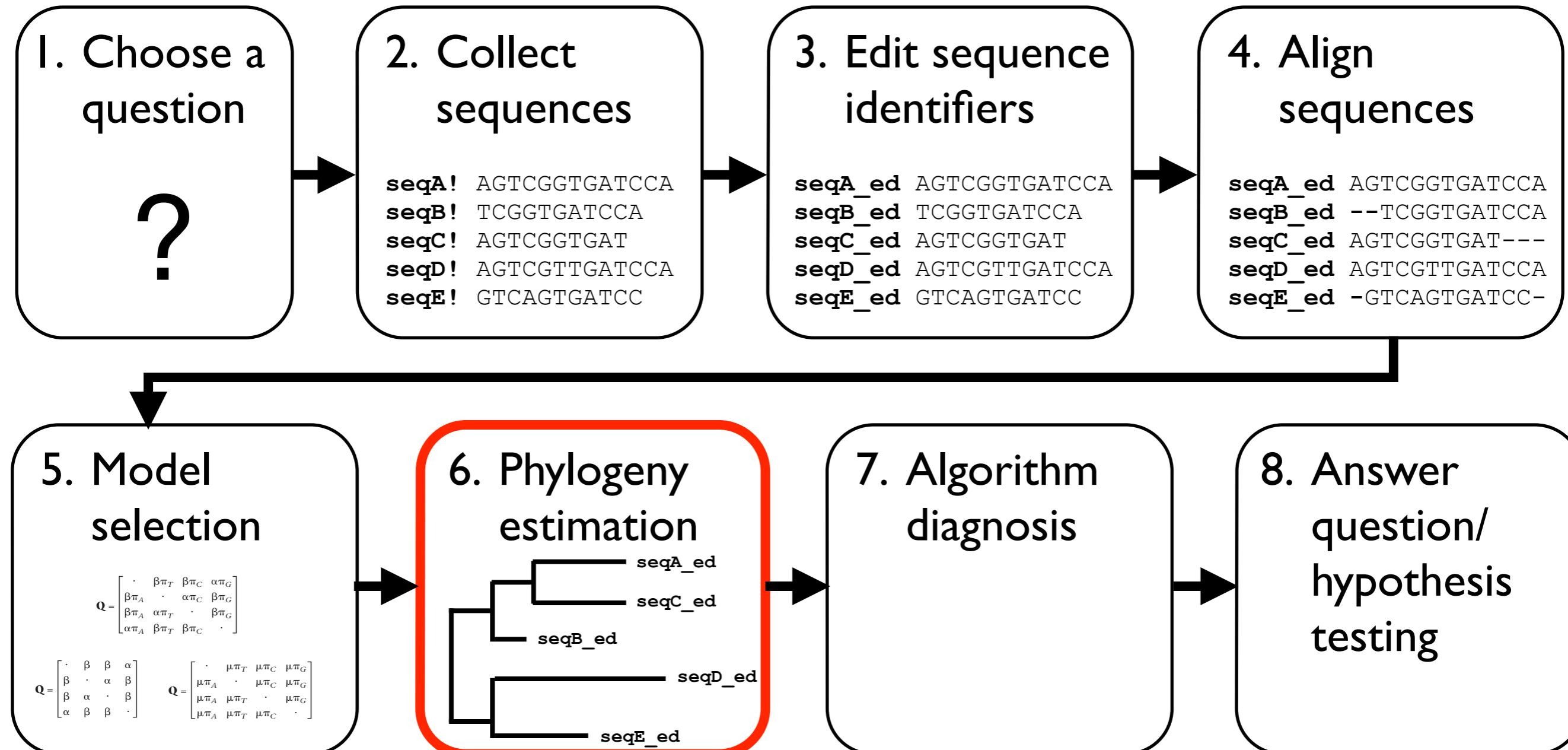
4. Align sequences: Ben, DATE, BaliPhy Prank, muscle, Probcons, mafft etc.

# Example Phylogeny Estimation Workflow



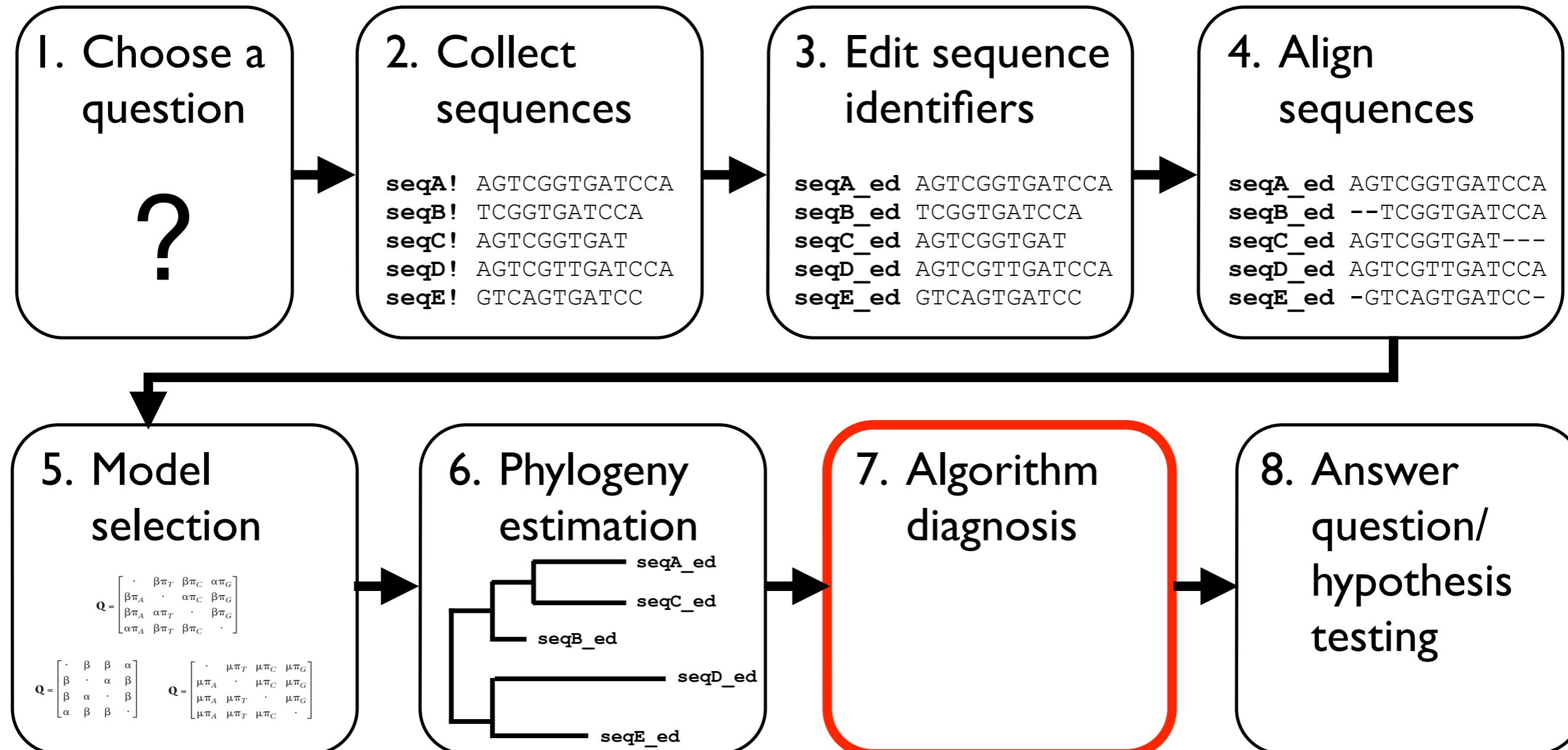
5. Model selection: Nick and Ziheng, DATE, PAML, jModelTest

# Example Phylogeny Estimation Workflow



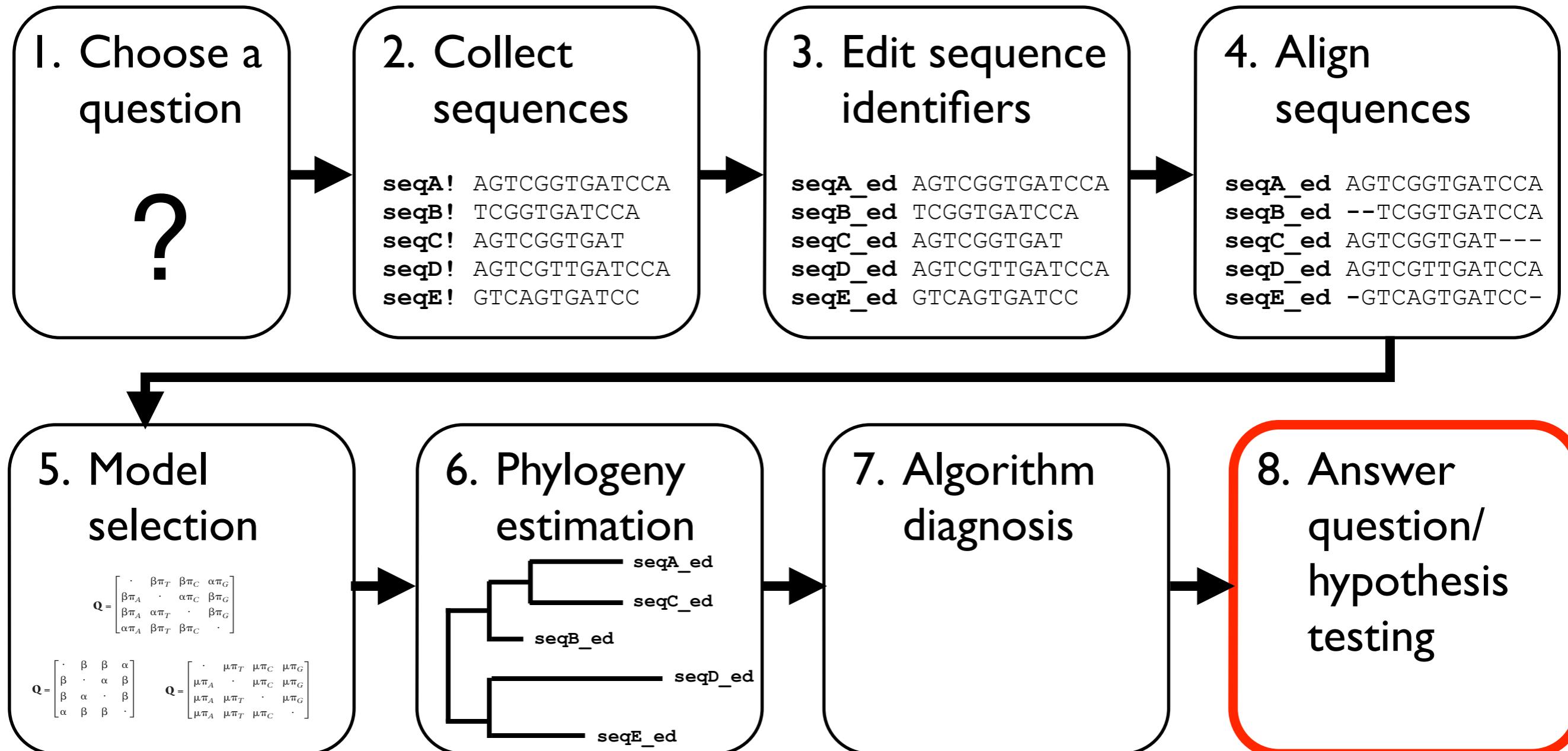
6. Phylogeny estimation: **Maria**, **Brian**, **Olivier**, **DATE**, **RAxML**, **PhyML**, etc.

# Example Phylogeny Estimation Workflow



7. Algorithm diagnosis: Maria, Brian, Olivier, DATE, Tracer, AWTY, etc.

# Example Phylogeny Estimation Workflow



## 7. Algorithm diagnosis

## 8. Answer question/ hypothesis testing

8. Answer question/hypothesis testing: Nick, Ziheng, Maria, Brian, Olivier, DATES, aLRTs, AIC, Bayes factors

- 
1. Write down a substantive question that can be informed by estimating the value of parameter(s) of a phylogenetic model
  2. Collect an appropriate set of sequences
  3. Edit sequence identifiers to be unique, compatible with analysis tools, and meaningful (in the context of allowing us to easily answer our question(s) of interest when examining resulting phylogenetic trees and other results)
  4. Align sequences
  5. Model Selection
  6. Phylogeny estimation
  7. Algorithm Diagnosis
  8. Examine the results of your analyses and address your substantial question of interest