

**Hashtag: #CoME9**

# **EMBO Practical Course on Computational Molecular Evolution**

**Sunday 8th–Thursday 19th May 2016**

**Institute of Marine Biology, Biotechnology and Aquaculture  
(IMBBC), Hellenic Centre for Marine Research (HCMR)  
Crete, Greece**

# Welcome!

# **TIMETABLE FOR DAY I**

```
$ whoami
```

# **8th version of this course**

**Support from:**

**EMBO, HITS (Wellcome), IMBBC, Astir Beach**

**THANKS!**



# Margarita Metallinou fee waiver awards: Thanks to HITS

# Course aims:

- improve understanding of how tools and methods work
- show examples of diverse applications of these approaches
- make you better able to do more appropriate analyses on your data
- give you access to help/advice from many experts
- help you form new professional relationships

# Course outline:

1.fundamentals

2.data

3.phylogeny reconstruction

4.hypothesis testing

5.inferences from population data

voluntary “advanced computing” sessions:

I. Markov simulations

2. Pair-wise sequence alignment in R

new this year - an experiment

# Logistics

- “comfort breaks”
- WiFi
- keeping on time!
- putting up posters
- feedback
- sign-in sheet
- photos
- social media #EMBOcompmol

# Thanks

- Alexis
- Many locals, especially: Eftichia, Cilia, Jacques, Antonis
- Funders
- Trainers
- You!

# Antonis Magoulas

# Introductions

Monday 9th May 2016

EMBO Practical Course on Computational Molecular Evolution

Institute of Marine Biology, Biotechnology and Aquaculture  
(IMBBC), Hellenic Centre for Marine Research (HCMR)  
Crete, Greece

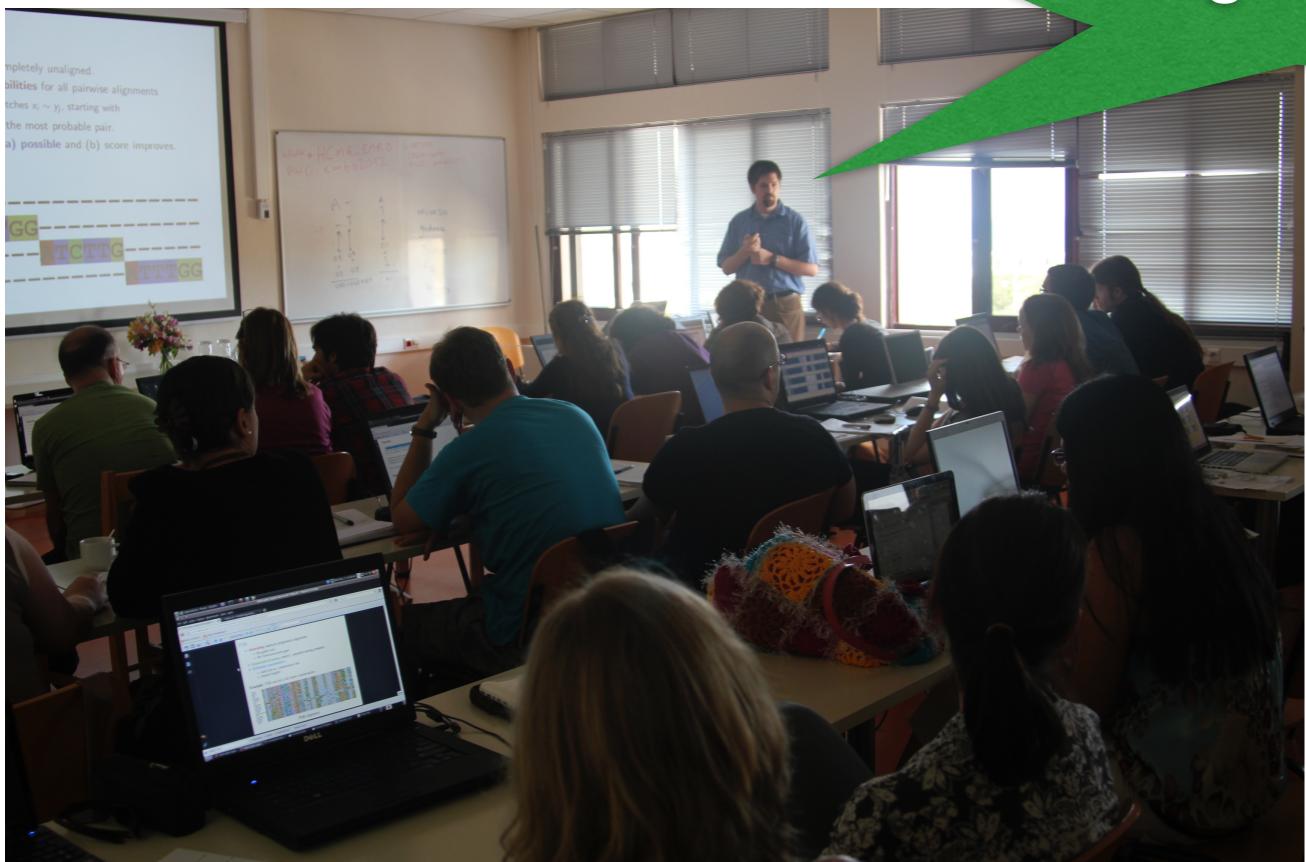
Aidan Budd  
EMBL Heidelberg, Germany

# 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**

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

**so the message of these first slides...**

**is that it's not just fun**

**but also really important**

that you make an **effort** during the course

**to get to know other trainees and the trainers**

**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 “introducing ourselves to someone for the first time” by doing lots of them, quickly
- hopefully makes it easier to start chatting later in the course
- quickly find people you have things in common with

# Speed Dating: Format

---

- meet other participants in many 1:1 chats (3-4 minutes)
- tell each other
  - names
  - where you work
  - research topics
  - look for person you both know or place you've both been

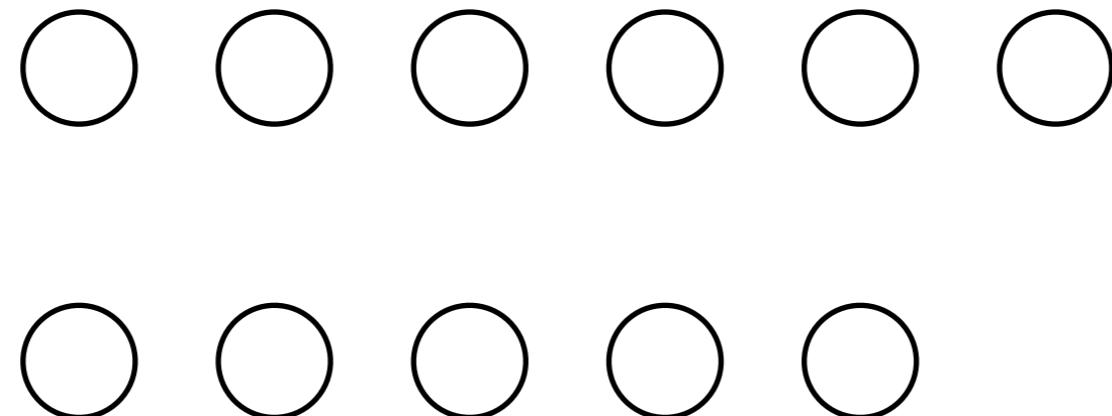
# Speed Dating: Format

---

Stand, awkwardly, in two rows

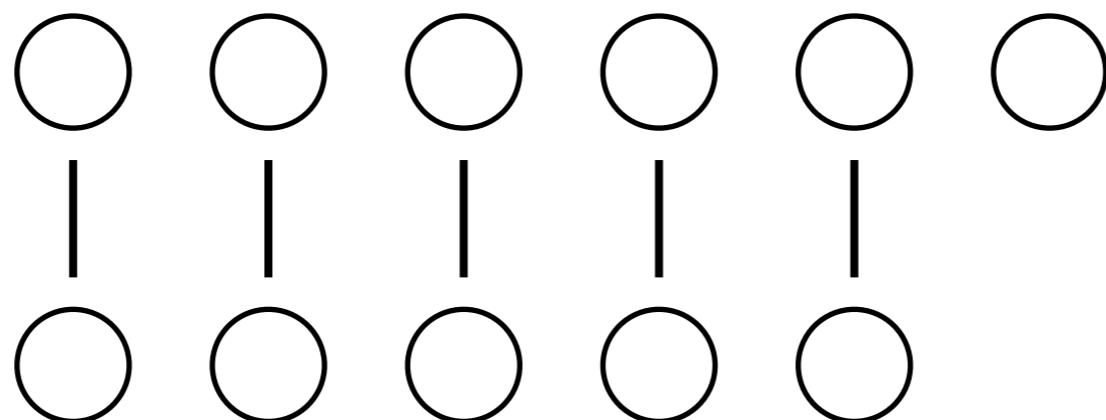
Face one person in the other row

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

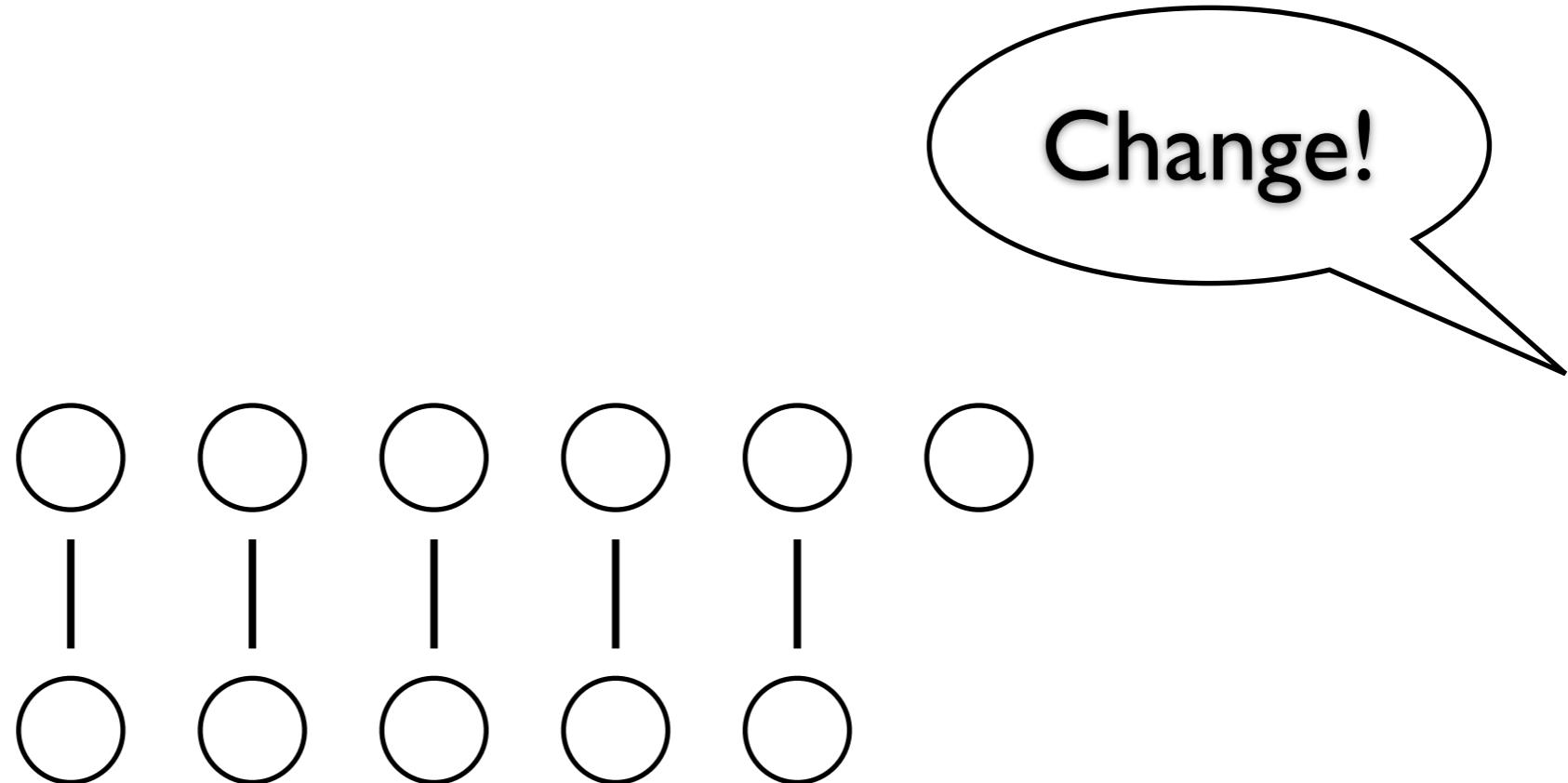


# Speed Dating: Format

**Chat!**

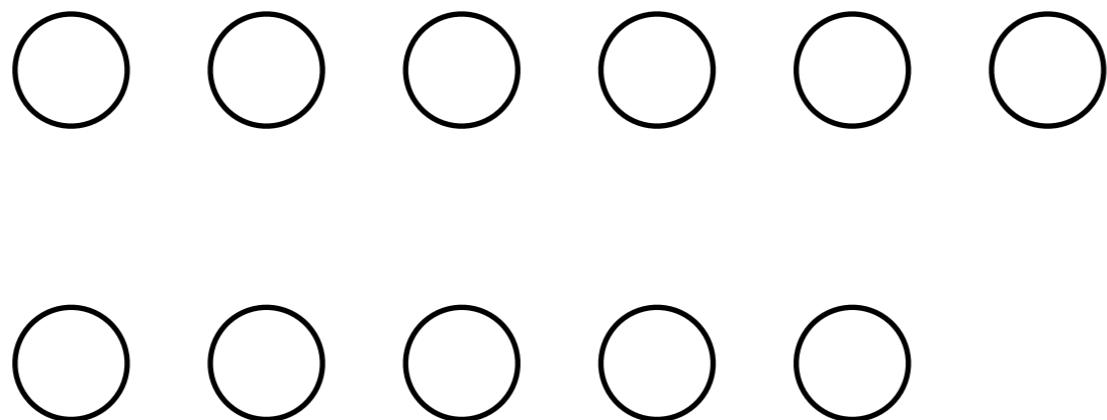


# Speed Dating: Format



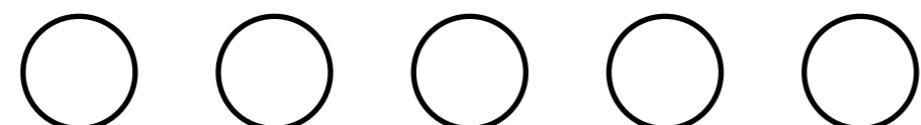
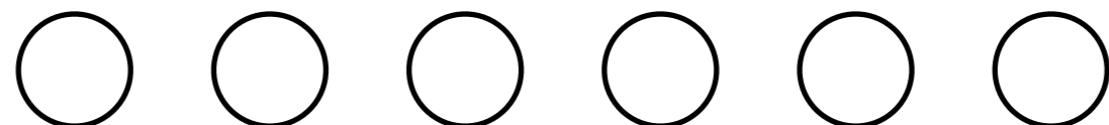
# Speed Dating: Format

---



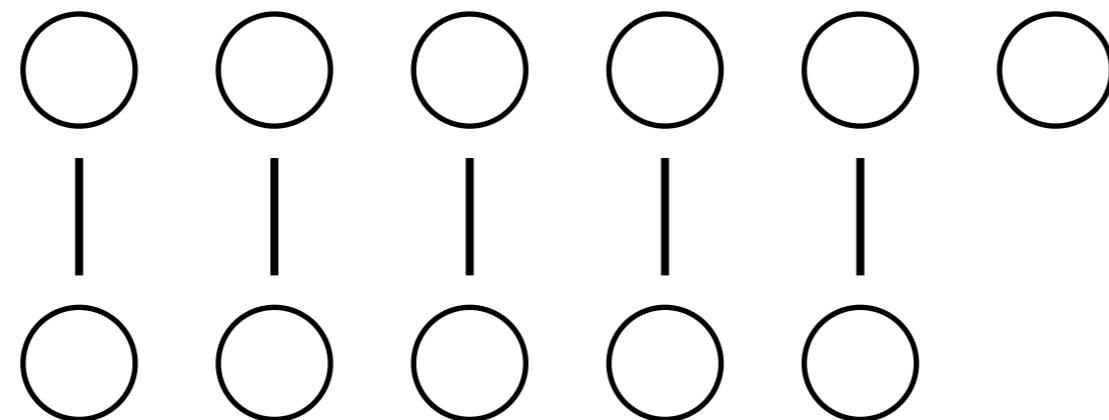
# Speed Dating: Format

---



# Speed Dating: Format

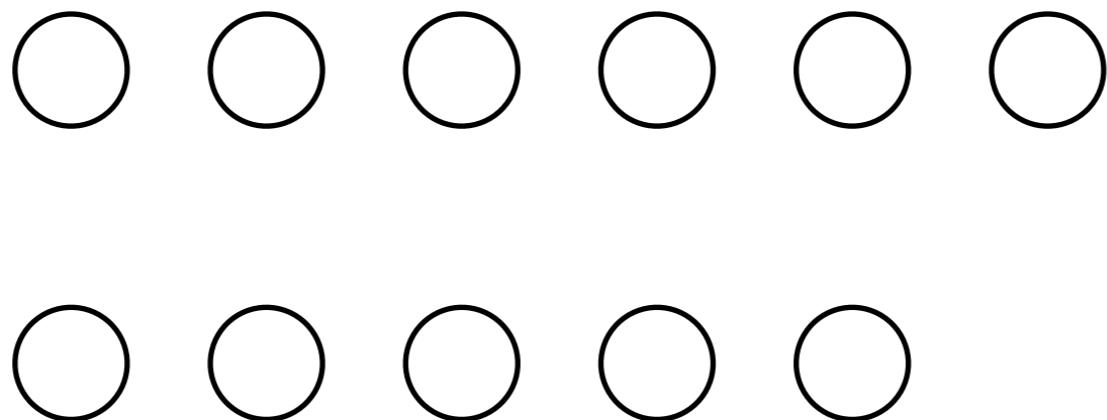
**Chat!**



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

# Speed Dating: Format

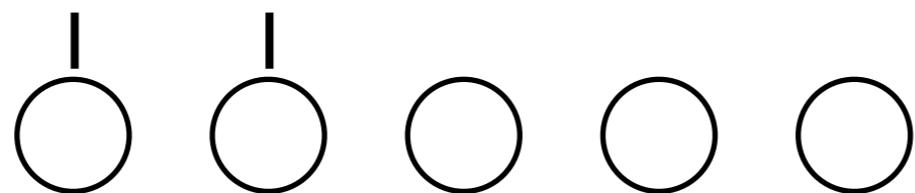
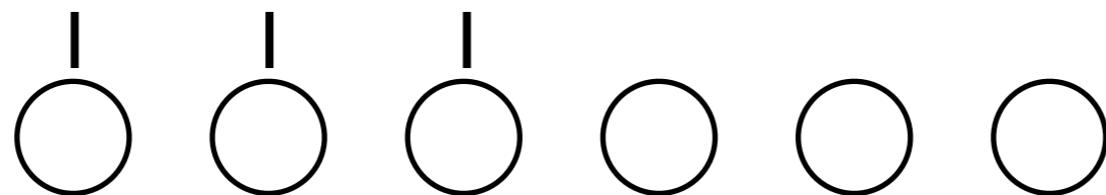
---



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

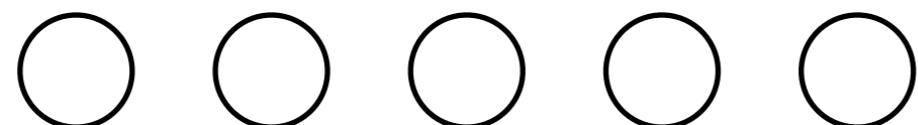
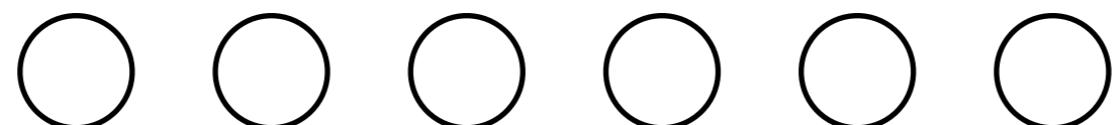
---

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

Note - one line stays still, the other **moves...**



# Session HTML pages

download zip/git clone from:

**<https://github.com/aidanbudd/trainingPhyloIntro>**  
**\$ git clone https://github.com/aidanbudd/trainingPhyloIntro**

“homepage” (links to exercises, instructions, presentations)

**homepageInterpretingPhylogenies.html**

exercises

**interpretingPhylogenies.html**

**Hashtag: #CoME9**

# Interpreting Molecular Phylogenetic Trees

Monday 9th May 2016

Aidan Budd  
EMBL Heidelberg, Germany

# How Do We Interpret Molecular Phylogenetic Trees?

## An Example

# Interpreting Molecular Phylogenetic Trees: An Example

---

Interpreting together a tree from a published article

Highlights a public health application of phylogenetics

an example of accurate tree estimation having a clear positive impact

Helps explore common features and problems 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
- Almost 100% death rate once rabies symptoms start in humans
- 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.



# 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

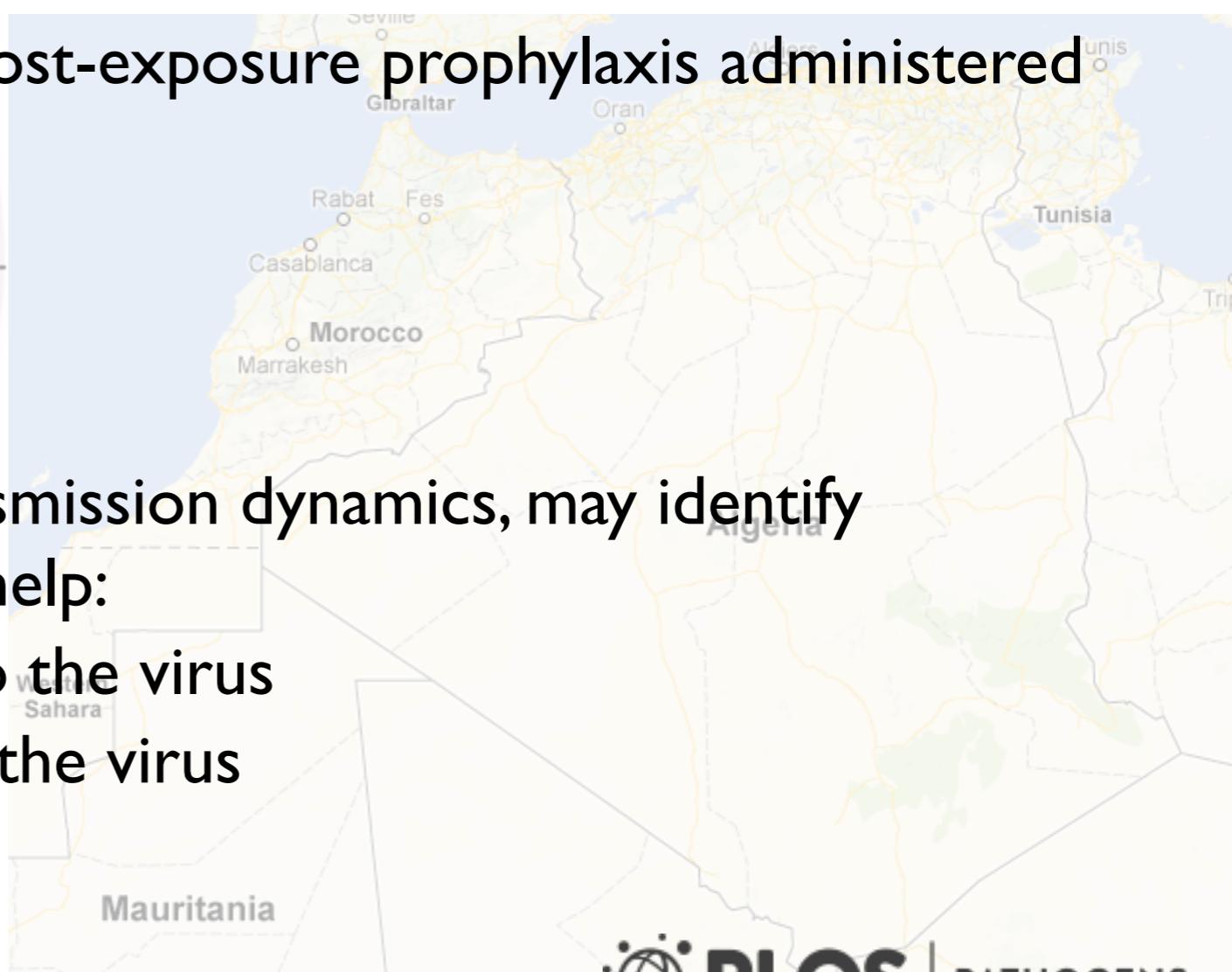
Therefore rabies:

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

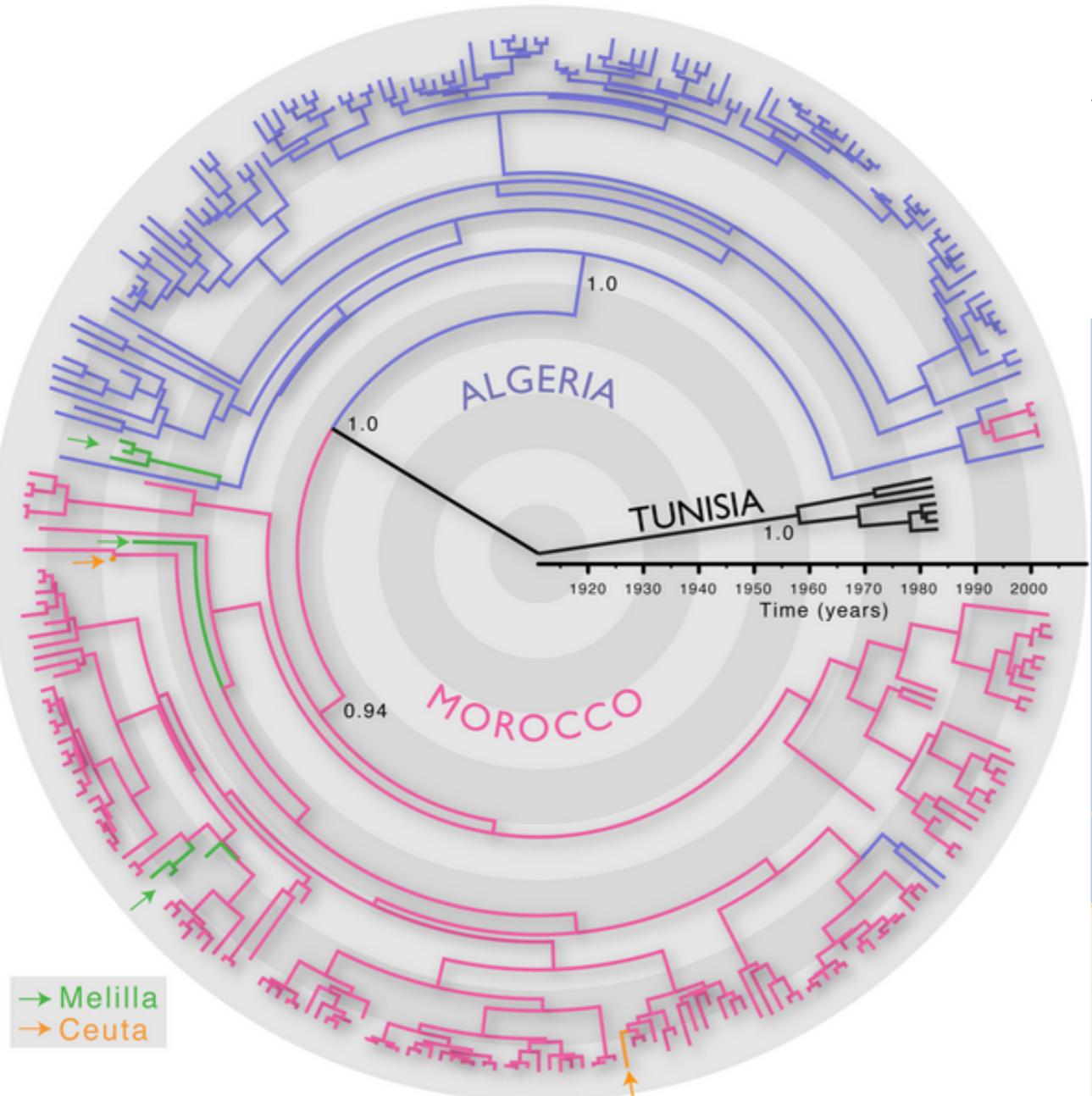
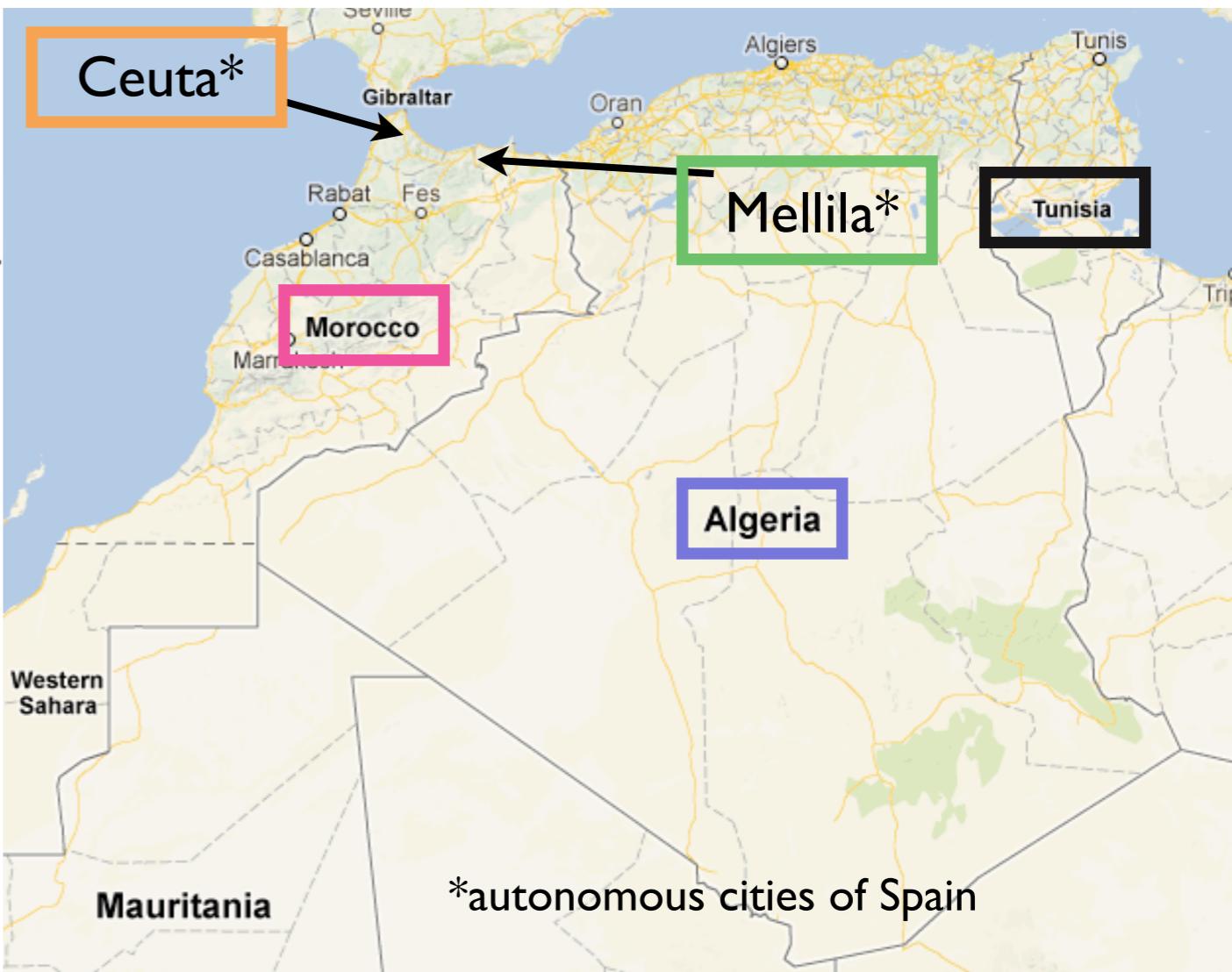


Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.

Phylogeny of rabies virus sampled from North African dogs

Branches coloured according to measured or inferred geographical location



\*autonomous cities of Spain

# Interpreting Molecular Phylogenetic Trees: An Example

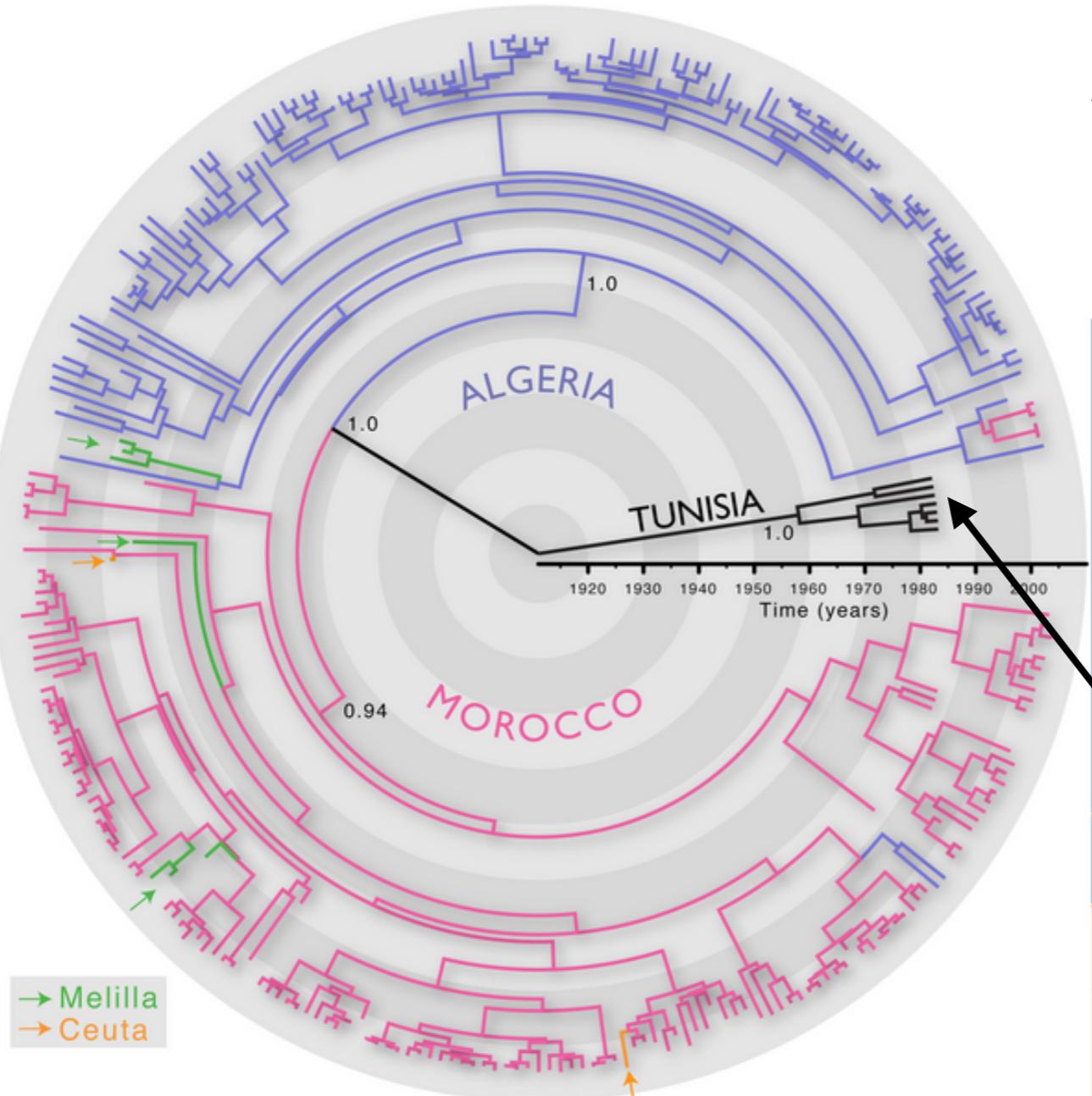
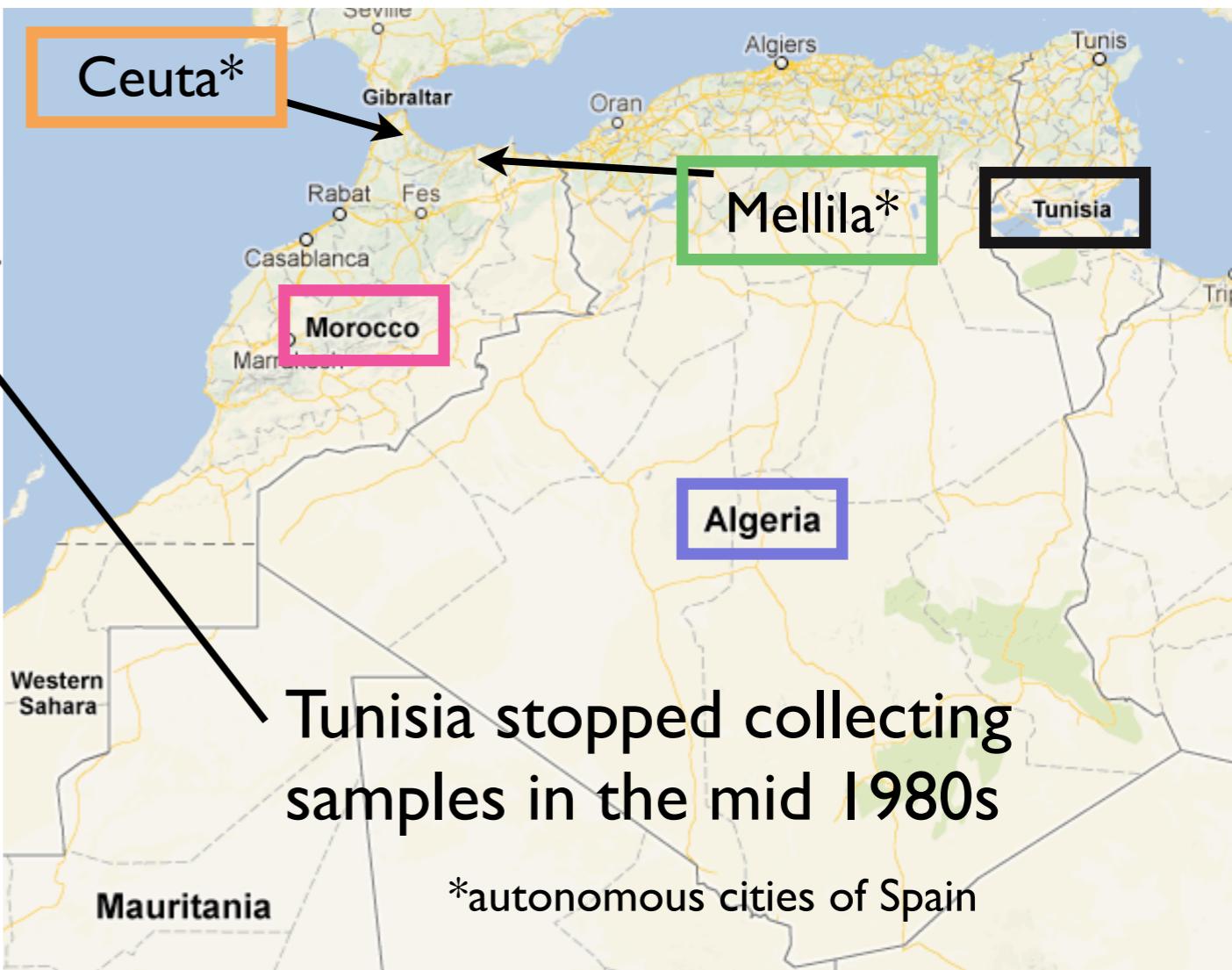


Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.

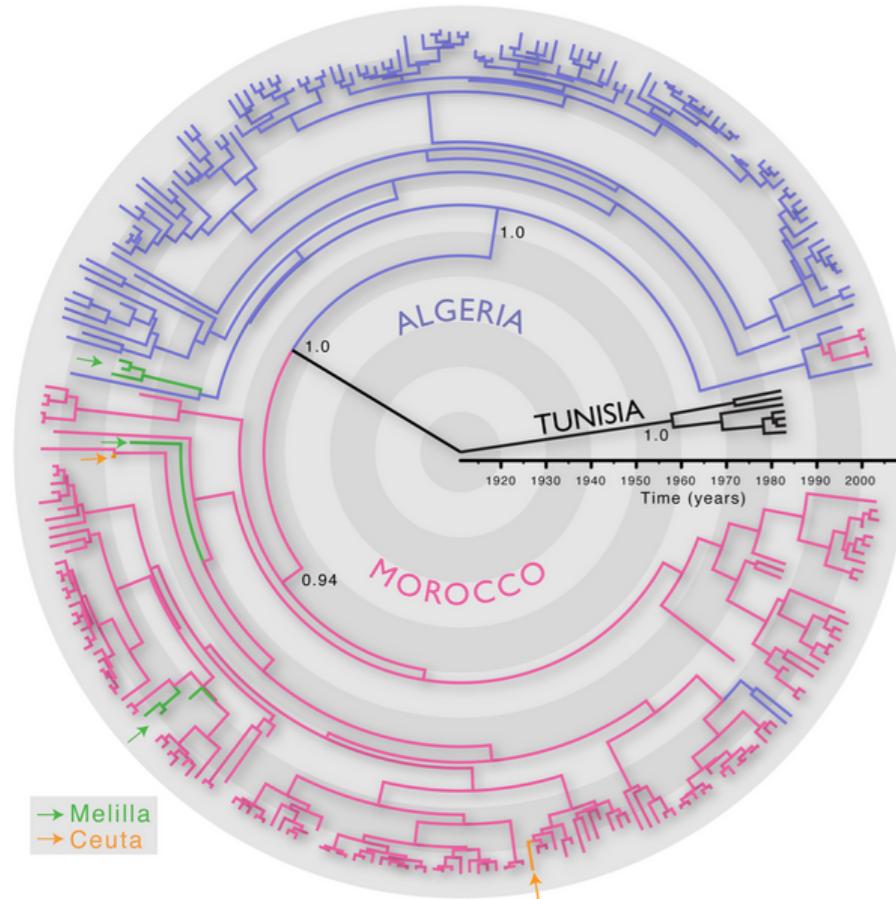
Phylogeny of rabies virus sampled from North African dogs

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

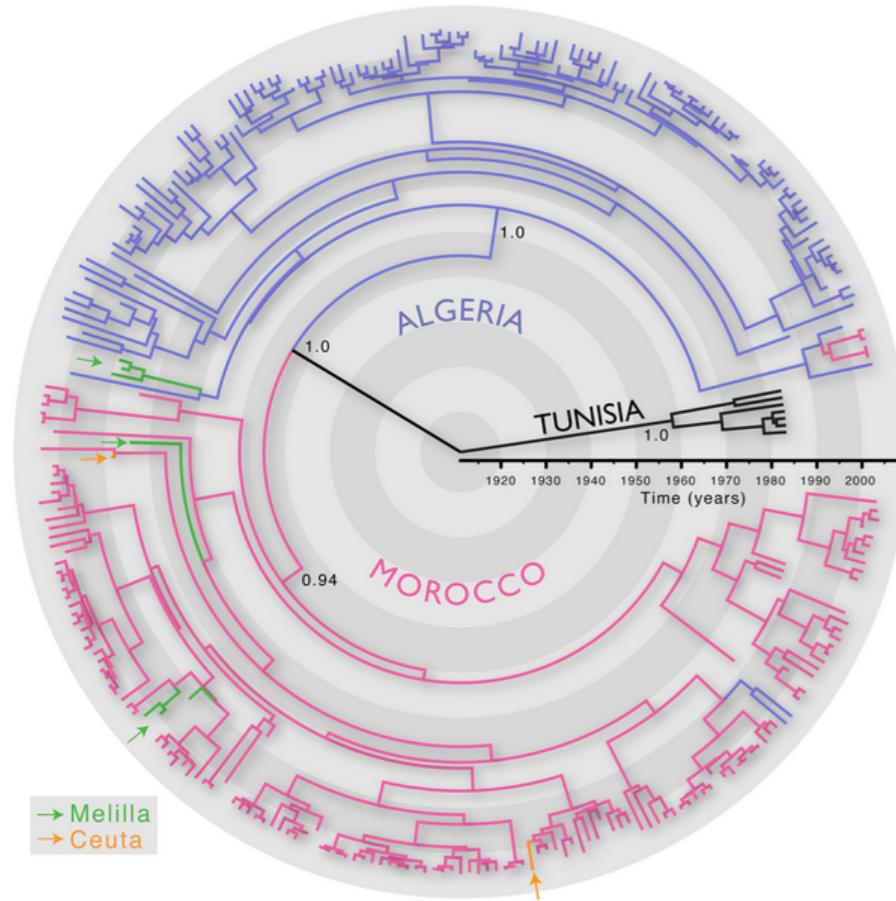
Then discuss the question with your neighbours

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?

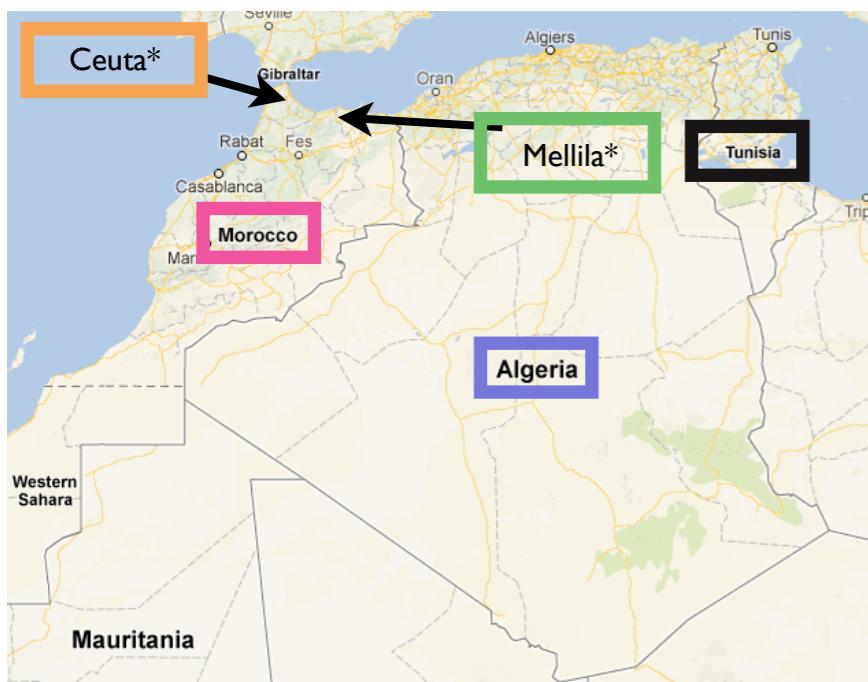
On the basis of this tree (and several other analyses) the authors conclude that the data supports a tree that makes it

A. more probable

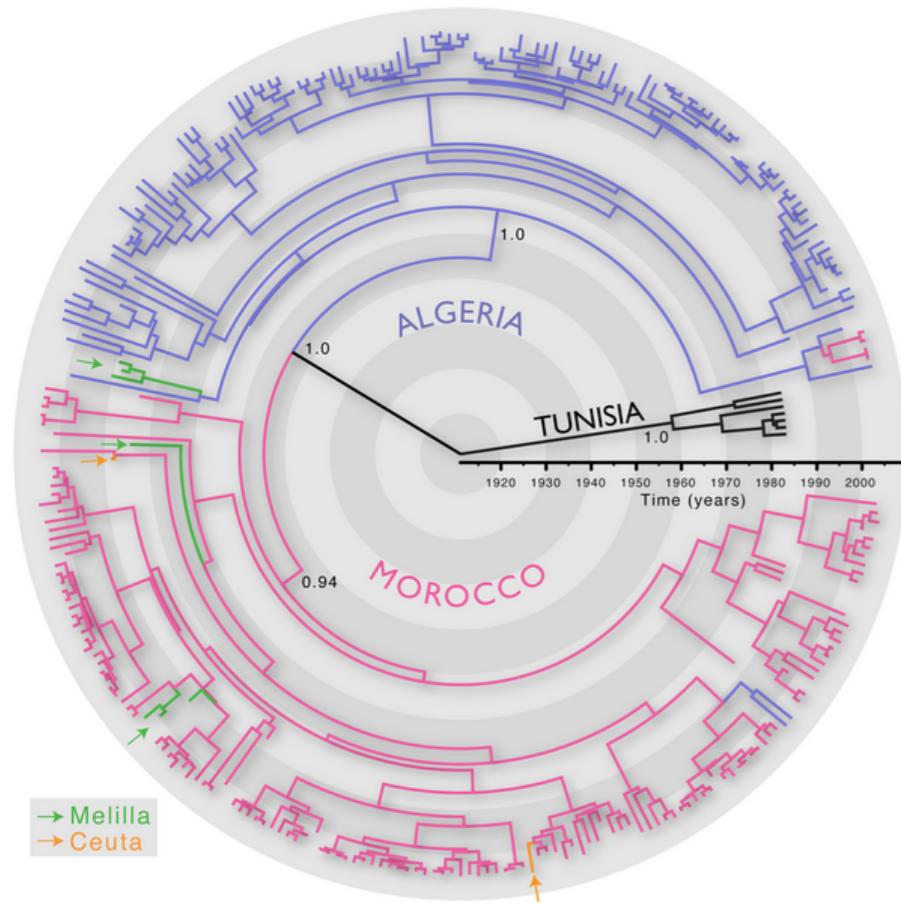
that human activity significantly influences the dynamics of rabies virus transmission between dogs

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

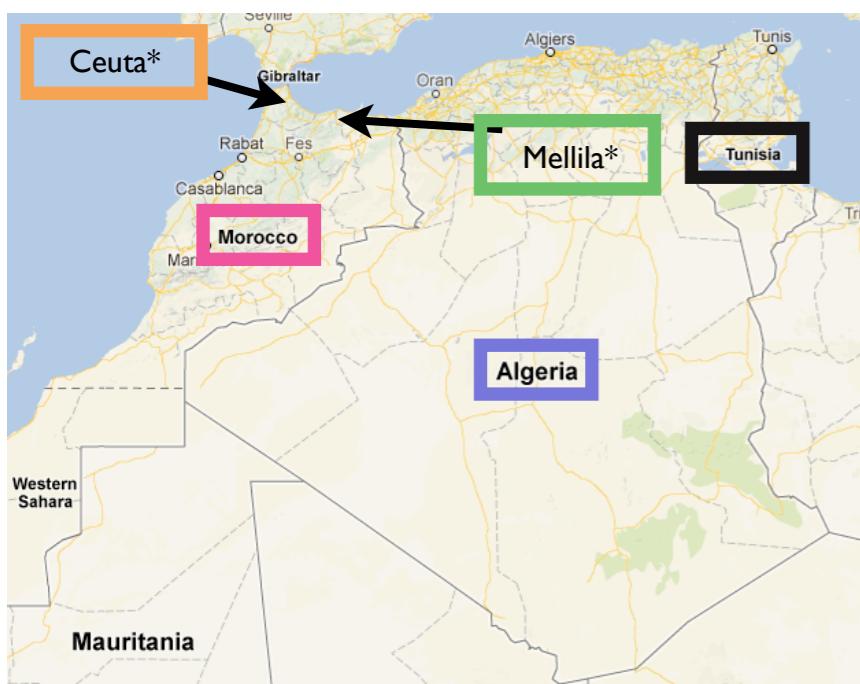


this exercise aimed to highlight that:

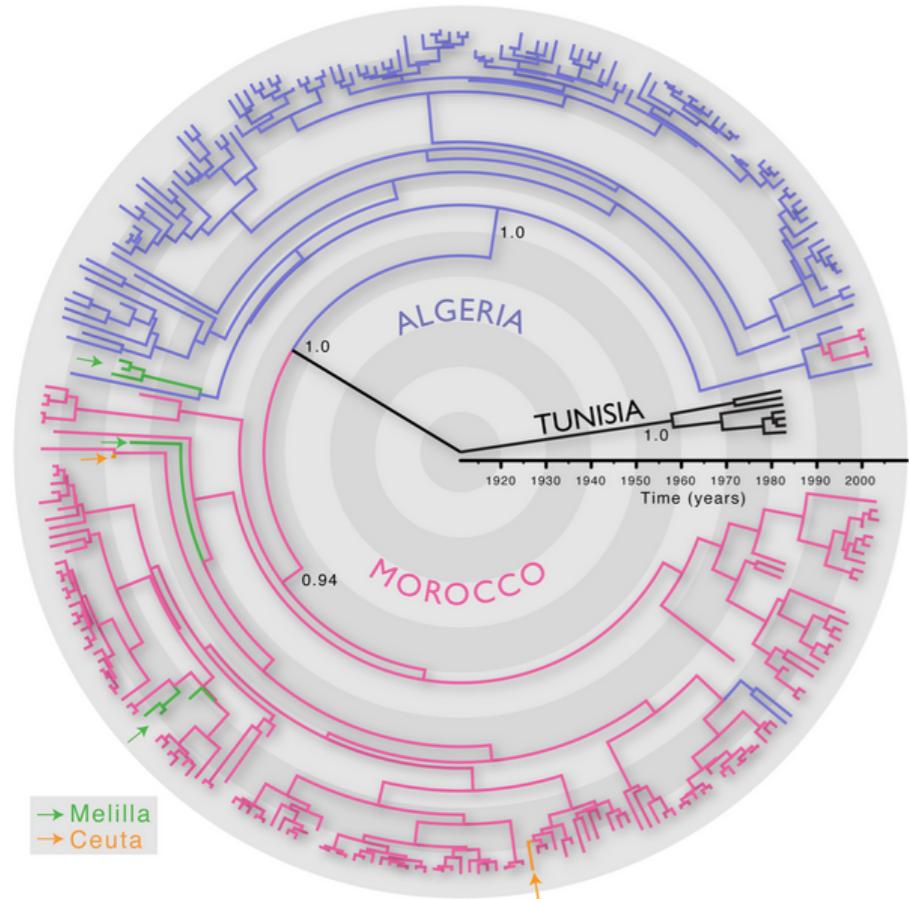
we have to make assumptions/use models to interpret the data - examples of assumptions that could be made while interpreting this tree:

- 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



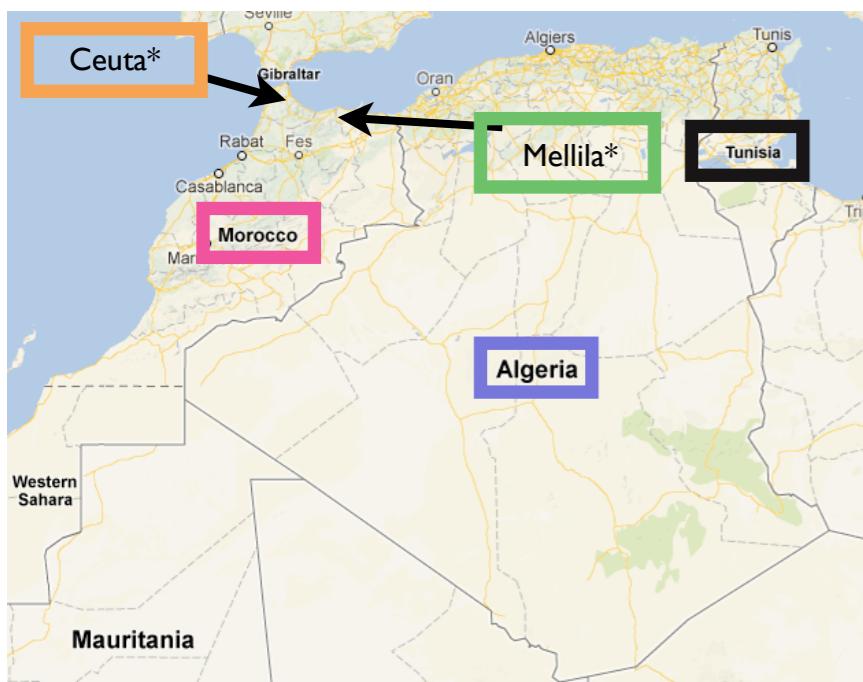
# Interpreting Molecular Phylogenetic Trees: An Example



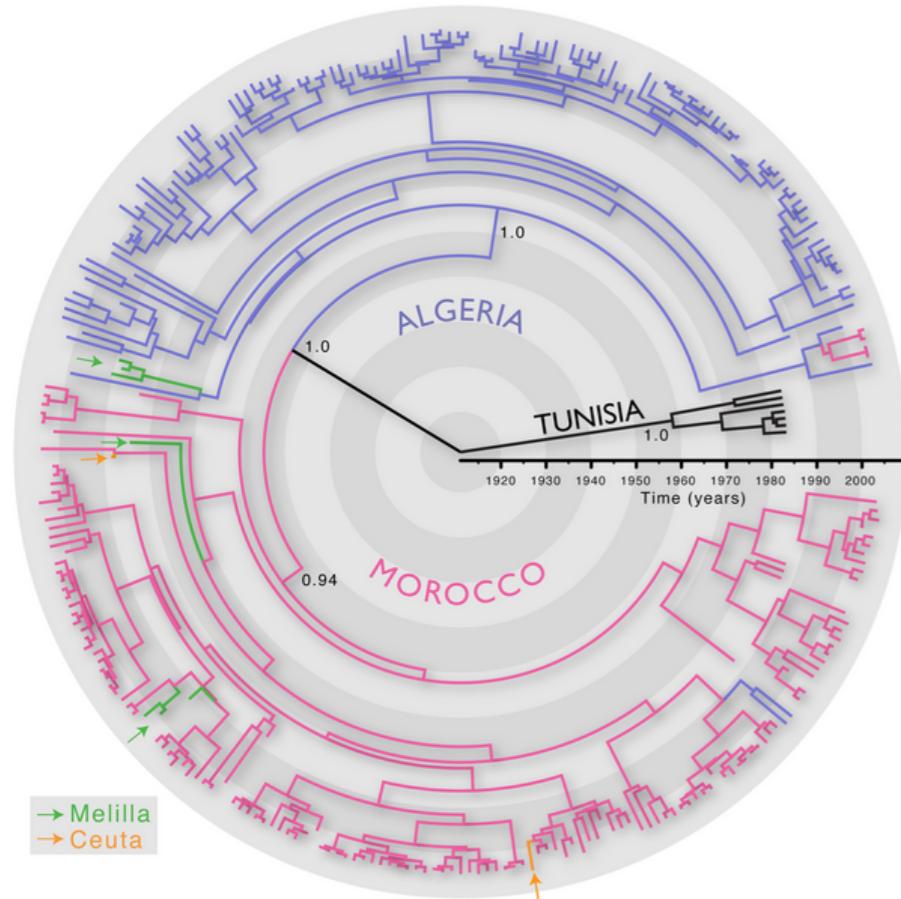
this exercise aimed to highlight that:

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

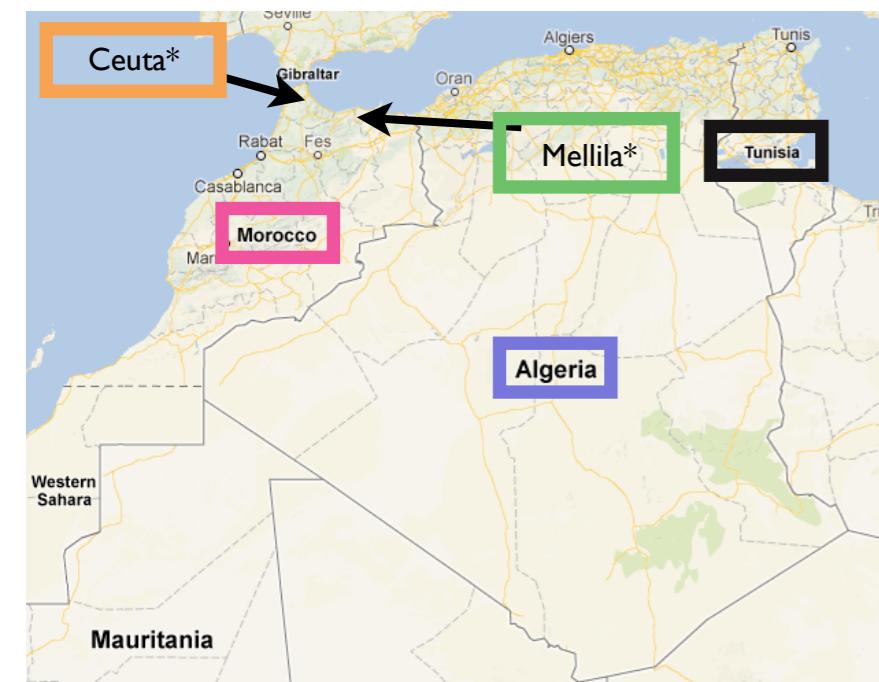


# Interpreting Molecular Phylogenetic Trees: An Example



this exercise aimed to highlight that:

phylogenetic trees can help making important, evidence-based, decisions



# Interpreting Molecular Phylogenetic Trees: Another Example

---

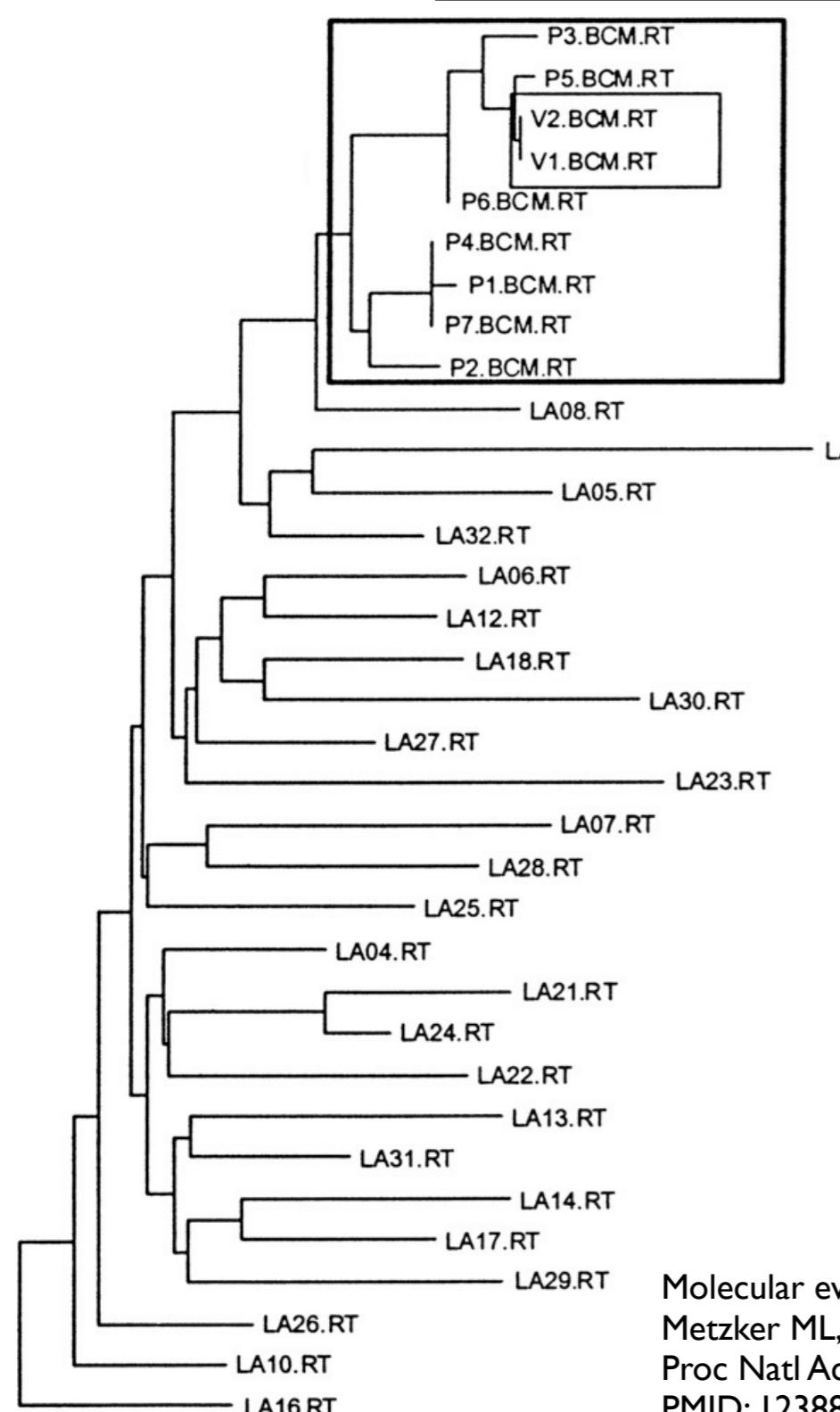
Interpreting together a tree from a published article

Highlights a forensic application of phylogenetics

an example in which the accuracy with which a tree is estimated has high stakes

Helps explore common features and problems when interpreting trees

# 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

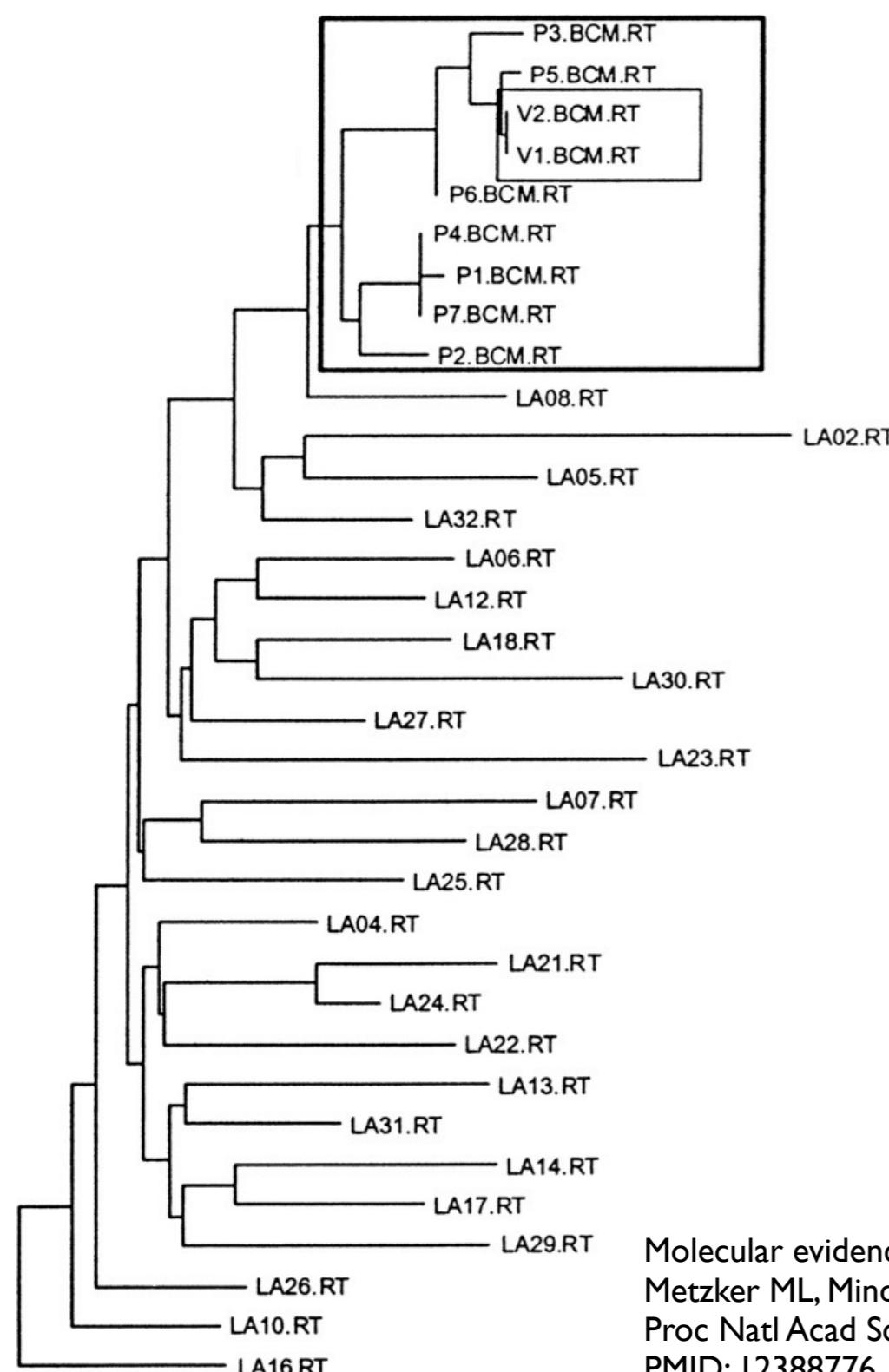
Phylogenetic analyses used as evidence in the trial

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



You are a juror on this case....

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

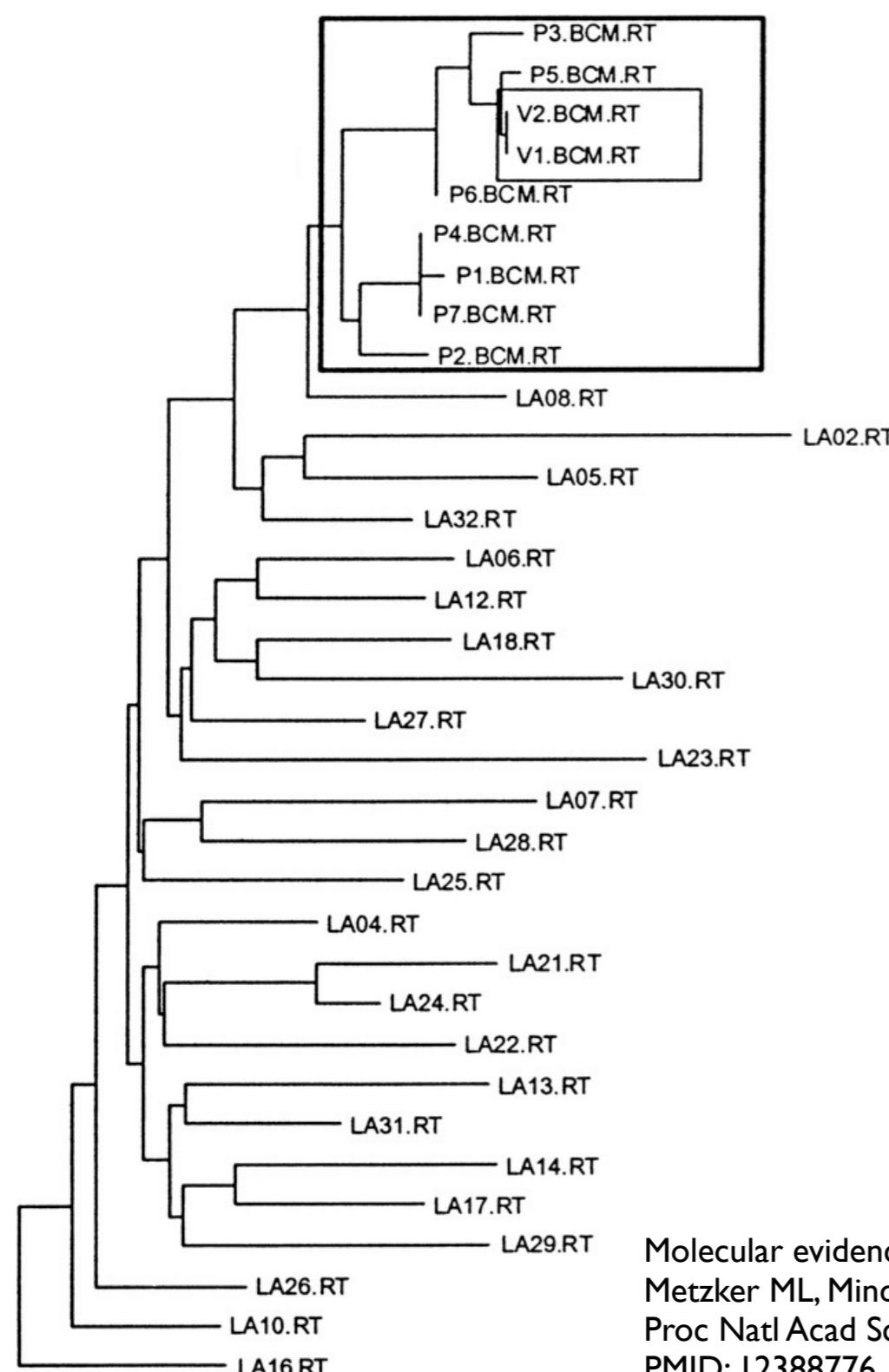
1. More likely to choose "guilty"
2. More likely to choose "not guilty"
3. Would not influence your choice of verdict

Which questions would you want the witness presenting the tree to be asked...

...to make the data as useful as possible for judging the guilt/non-guilt of the defendant?

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

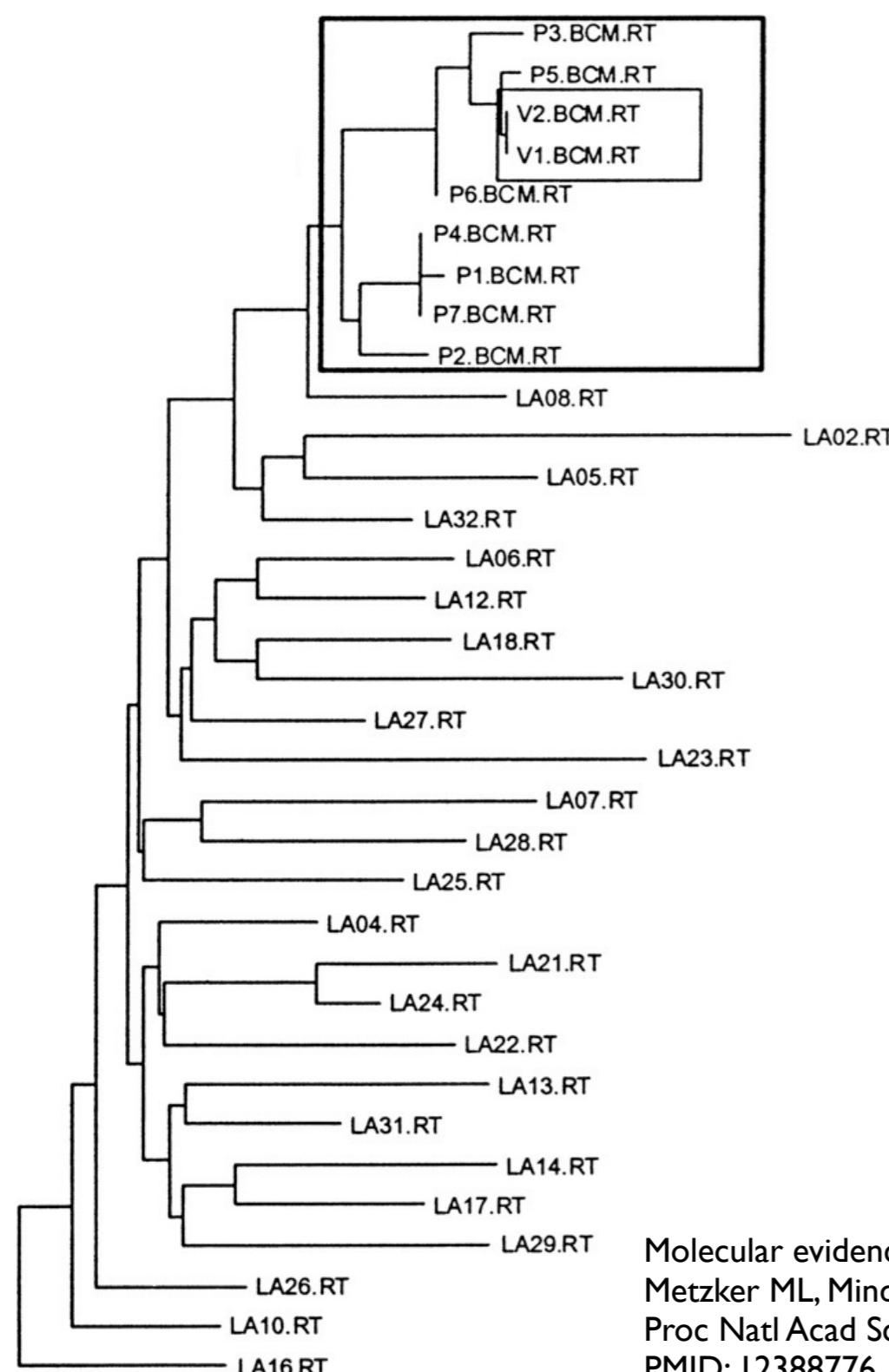


Information about where and how OTUs were sampled impacts how we interpret a tree

The way we label OTUs in a tree affects how quickly and easily we can interpret it - choosing labels to highlight key features of your data relevant to your questions of interest can help make it easier to interpret

Scale bars (when a tree's branches have a 'length') give important information for our expectations of how accurate we think a tree might be

# Interpreting Molecular Phylogenetic Trees: Another Example



Estimates of random error associated with parameters in the tree is also useful information for assessing how well we think the data supports a given hypothesis

The “quality” of the data (here it’s a nucleotide sequence alignment) used to estimate parameters (here we’re most interested in the tree topology parameter) also impacts the confidence we have in an parameter value being accurate

Describing parameters we’d expect under different hypotheses (e.g. sketching tree shapes associated with guilt or innocence) can help when interpreting results

# 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.

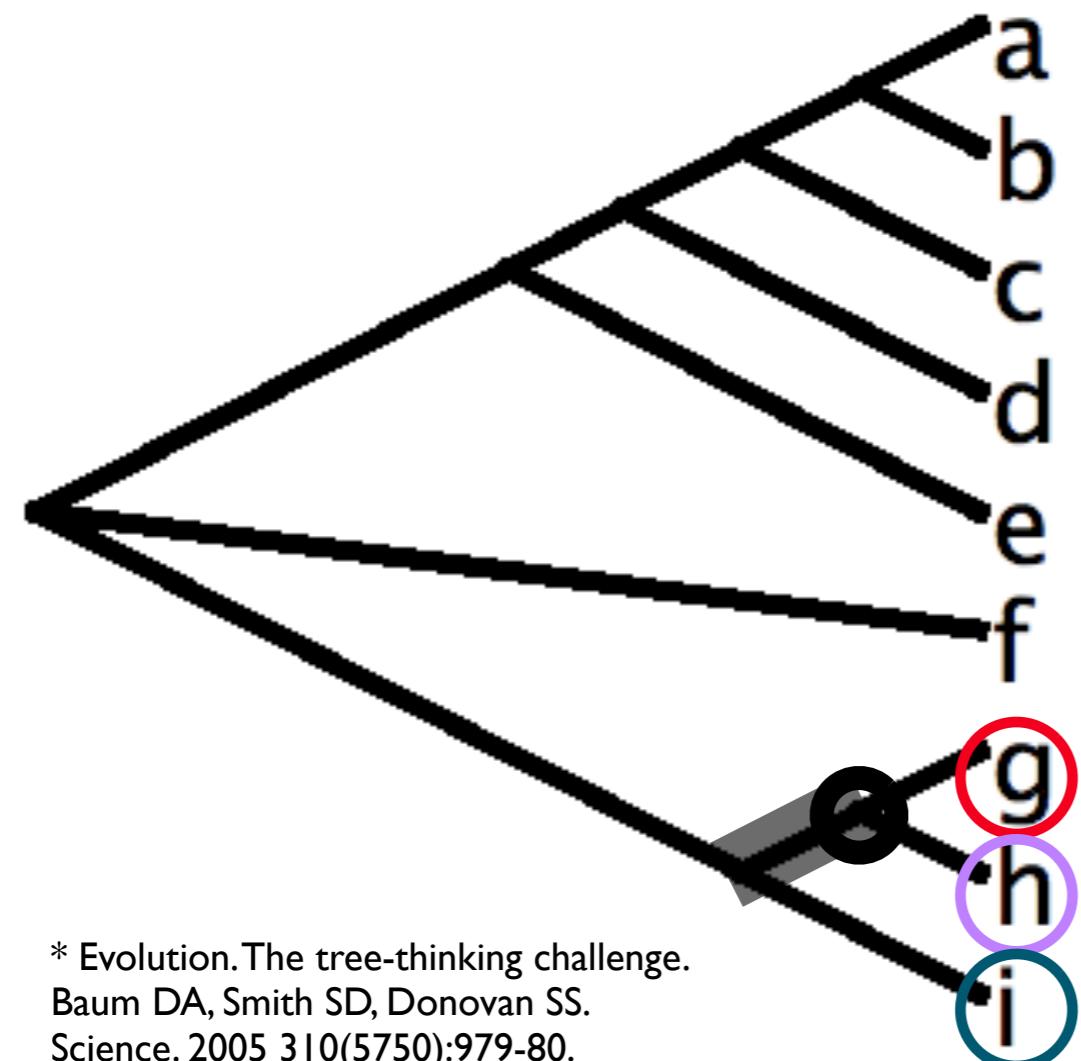
Baum DA, Smith SD, Donovan SS.

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" \*

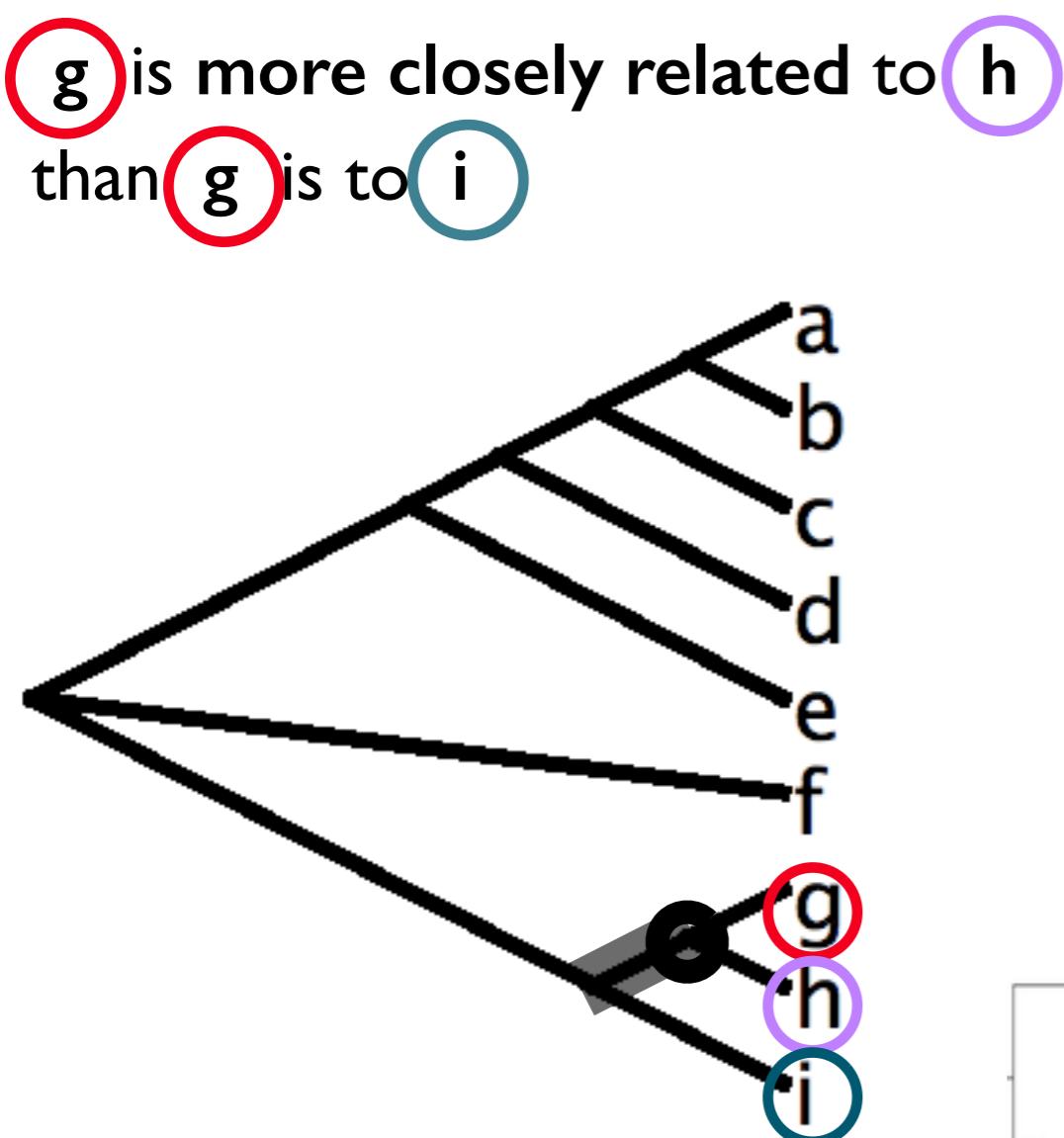


g is more closely related to h  
than g is to i  
because g and h share common ancestors  
that neither share with i  
i.e. degree of relatedness  
is associated with the extent of ancestry  
(i.e. the number of ancestors) taxa share  
with each other

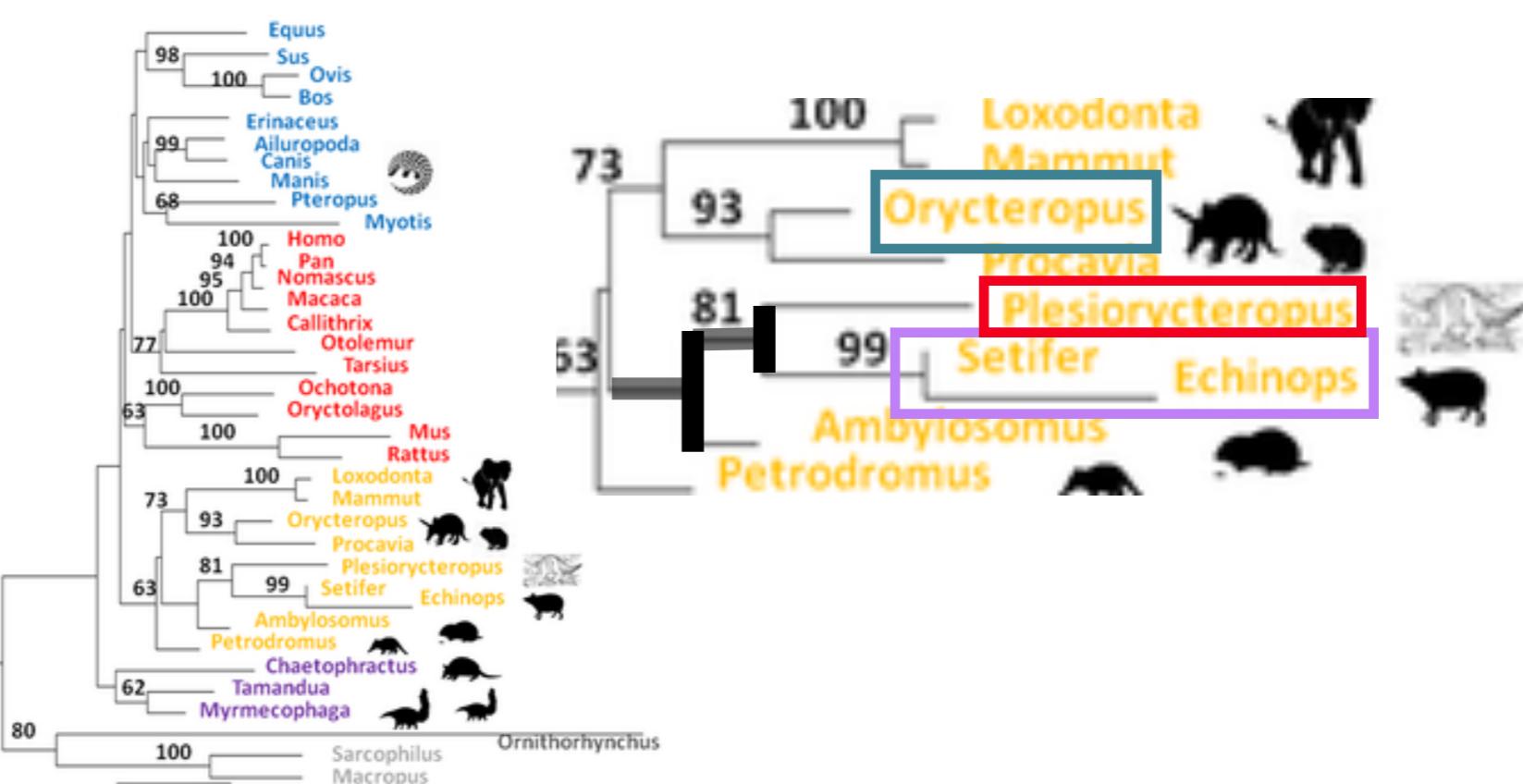
\* Evolution. The tree-thinking challenge.  
Baum DA, Smith SD, Donovan SS.  
Science. 2005 310(5750):979-80.  
PMID: 16284166

# 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 ..



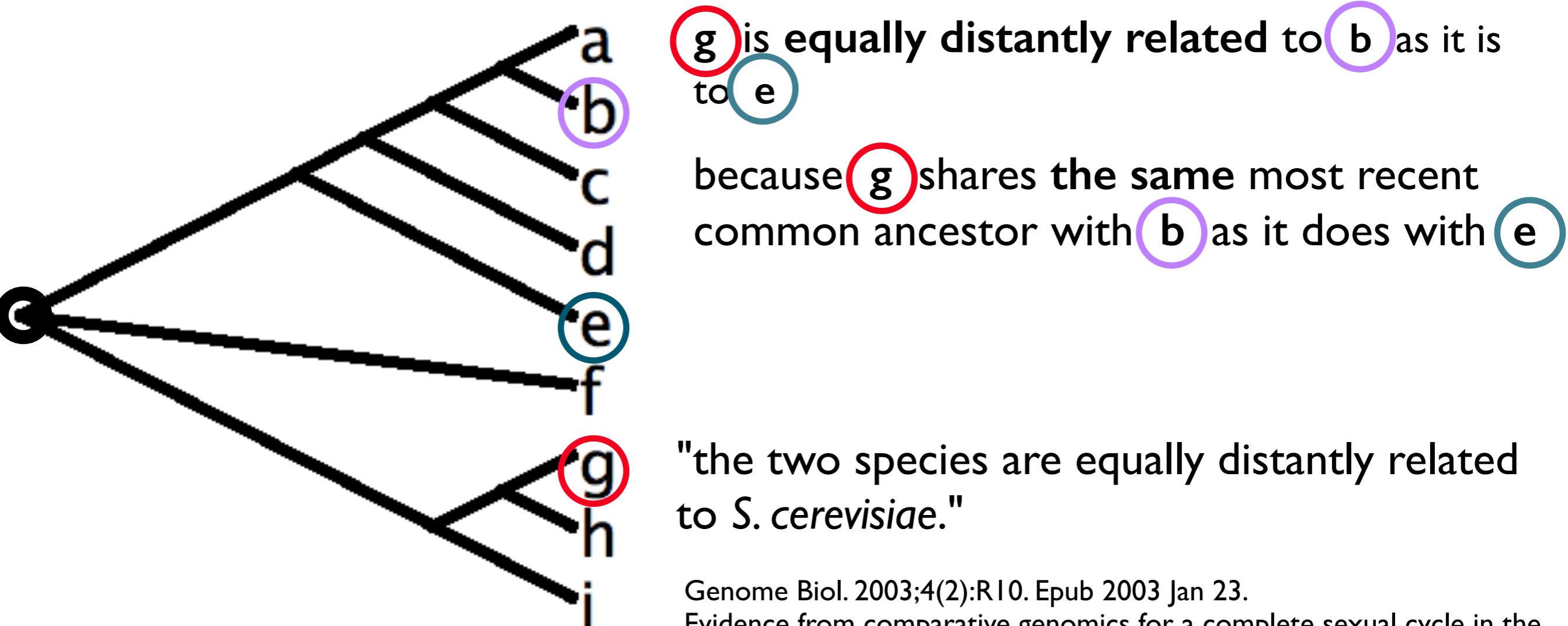
\* Evolution. The tree-thinking challenge.  
Baum DA, Smith SD, Donovan SS.  
Science. 2005 310(5750):979-80.  
PMID: 16284166

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

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>

# Relatedness (in the context of phylogenetic trees)

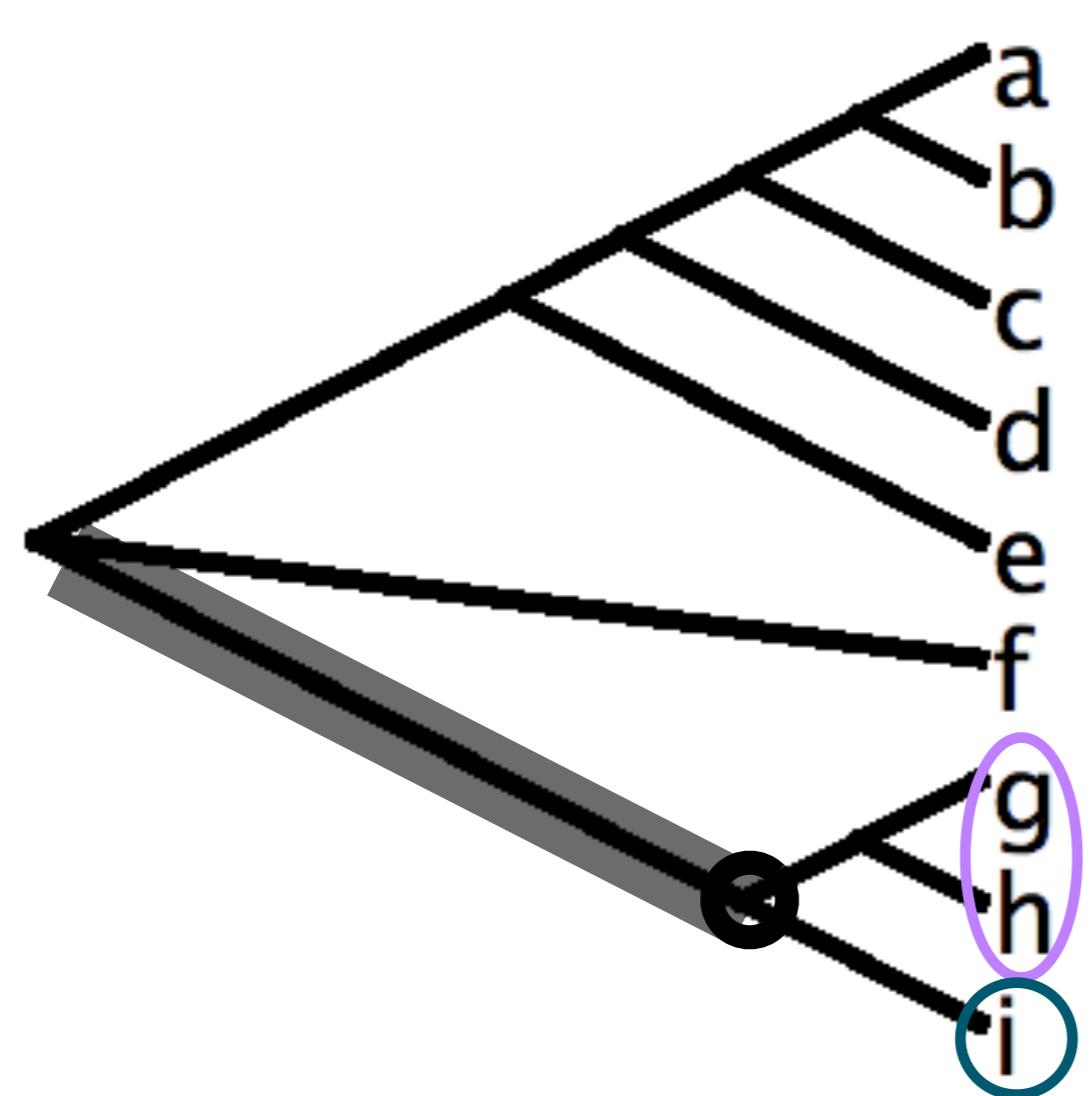
of all the OTUs represented in this tree



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.

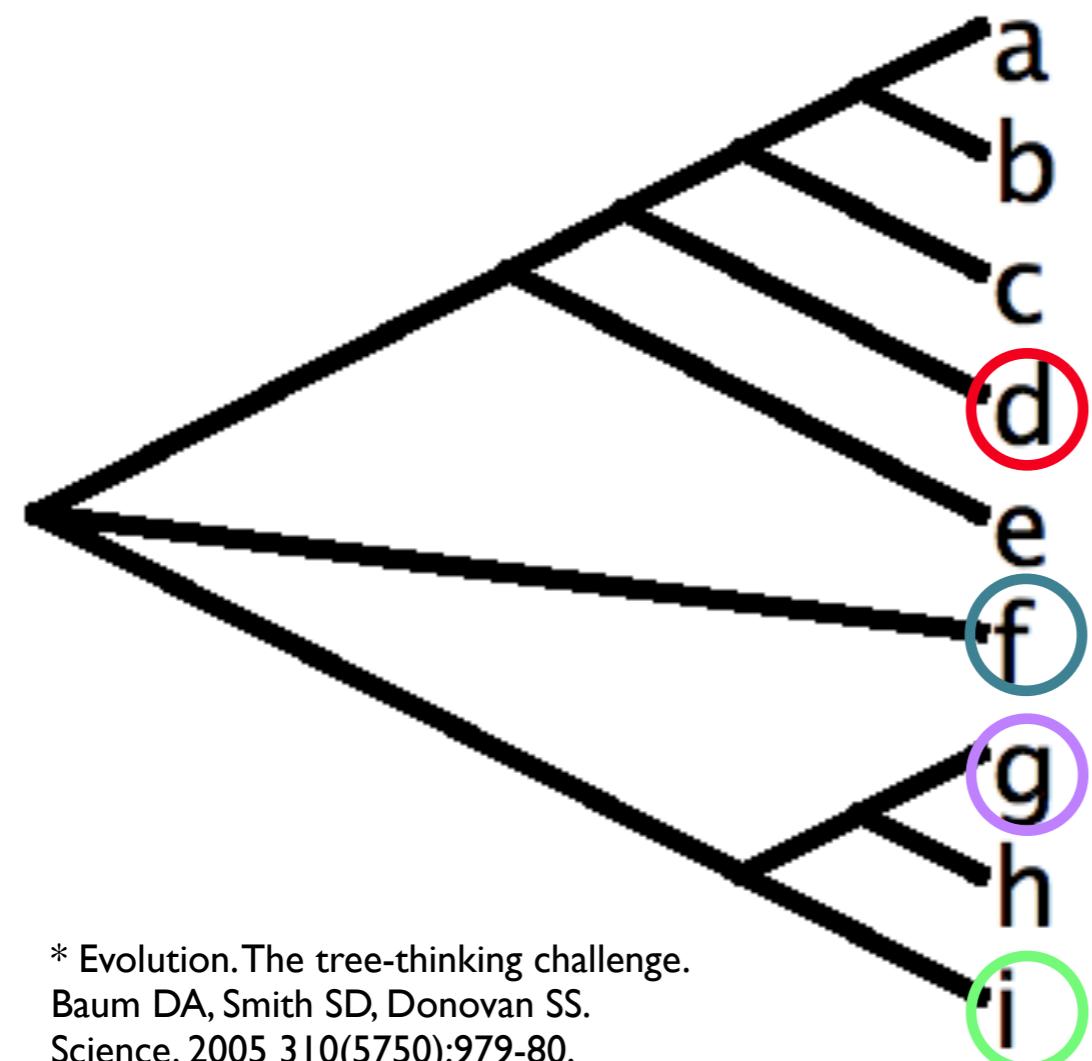
Phylogenetic analysis of three genes of Penguinpox virus corresponding to Vaccinia virus G8R (VLTF-1), A3L (P4b) and H3L reveals that it is most closely related to Turkeypox virus, Ostrichpox virus and Pigeonpox virus.

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?



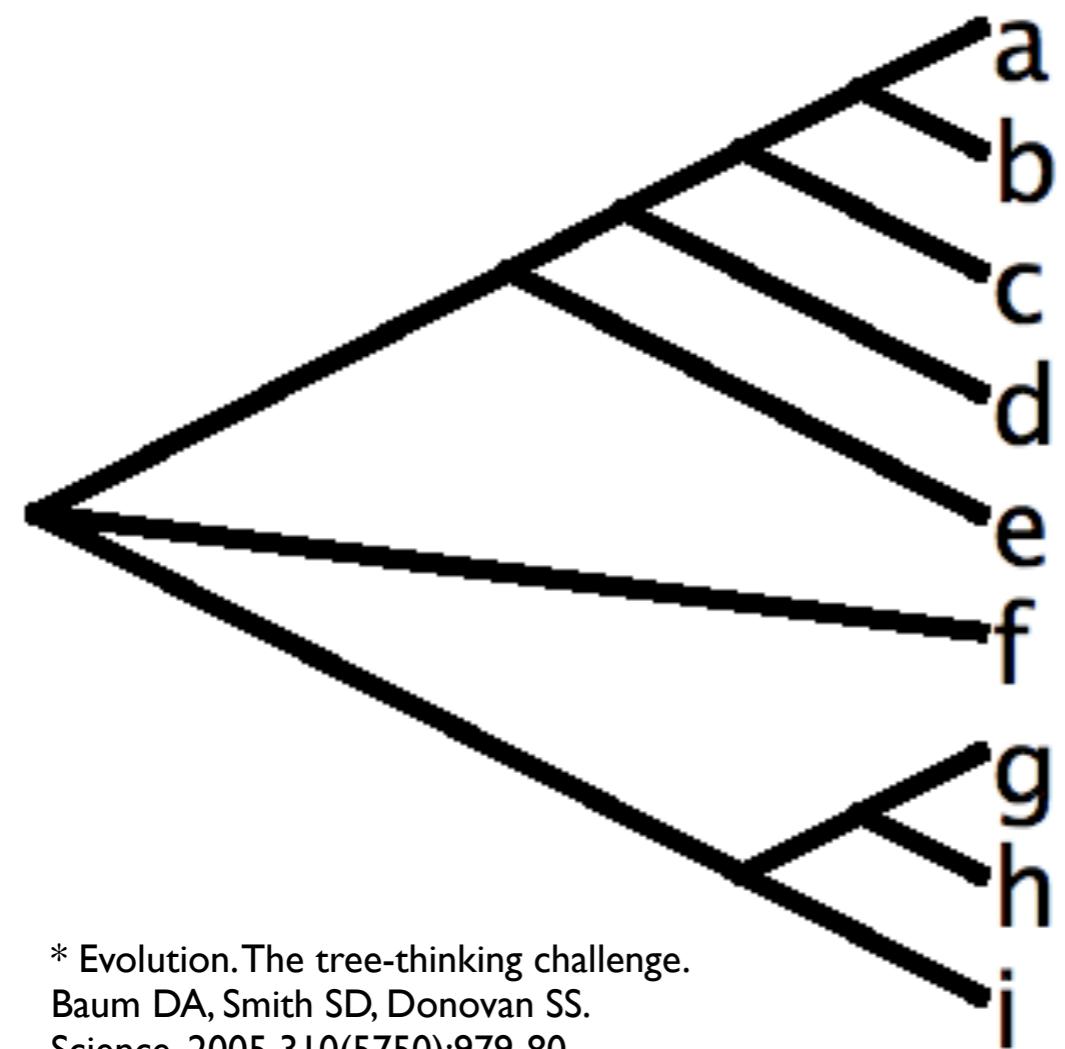
1. **d** is more closely related to than to **f** or **i** g
2. **d** is more closely related to than to **g** or **i** f
3. **d** is more closely related to than to **g** or **f** i

\* 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" \*

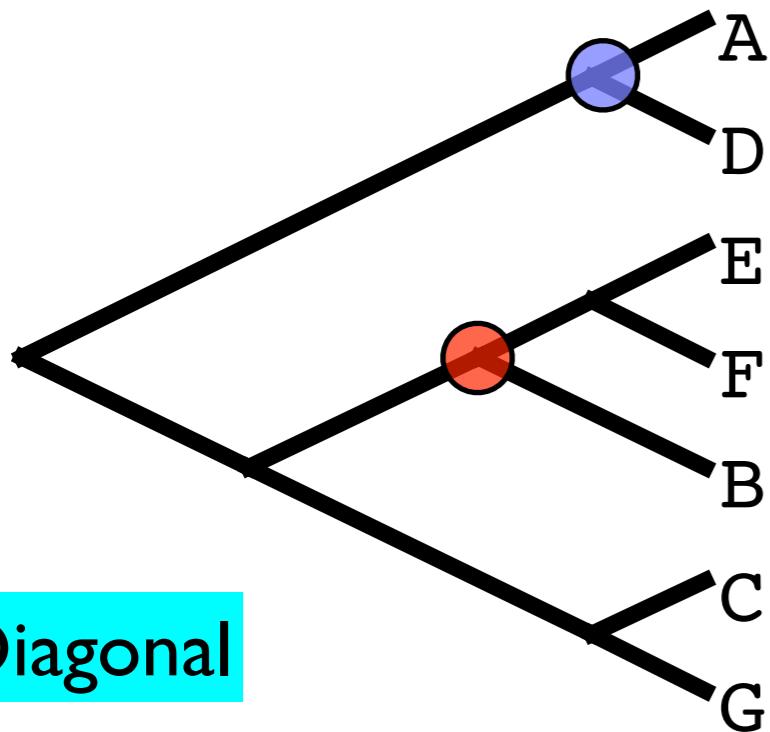
Why spend so much time discussing "relatedness" with you?



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

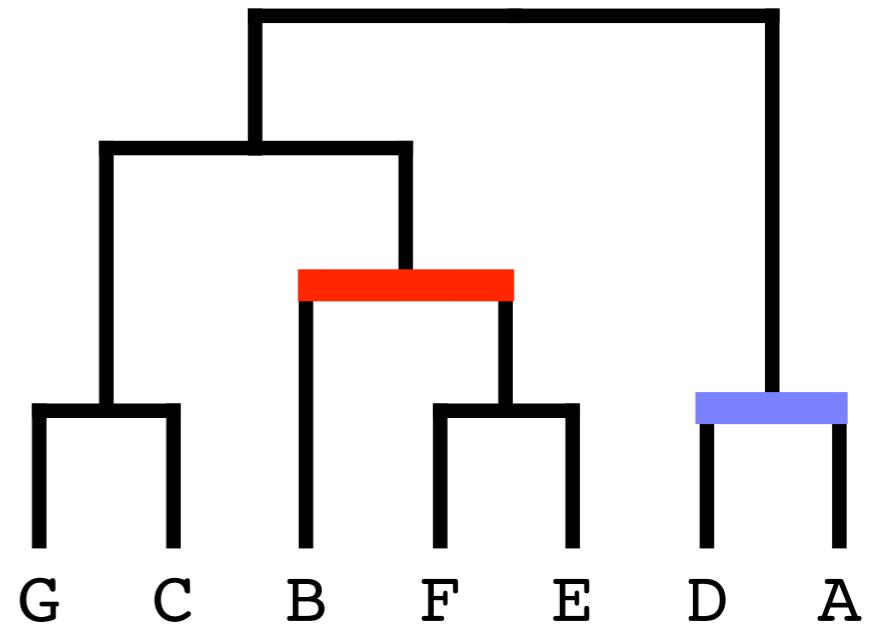
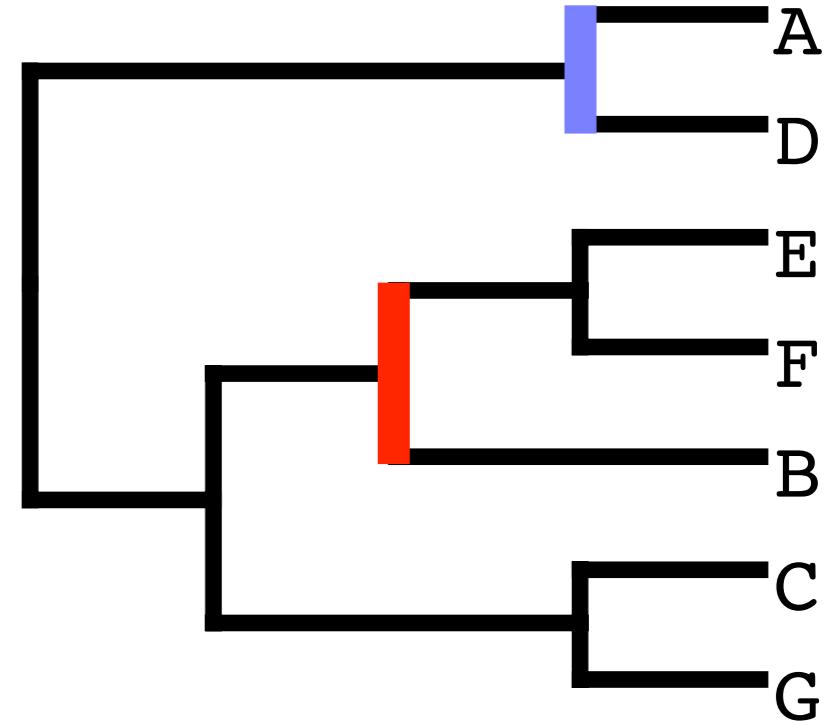
\* Evolution. The tree-thinking challenge.  
Baum DA, Smith SD, Donovan SS.  
Science. 2005 310(5750):979-80.  
PMID: 16284166

# Tree Representations



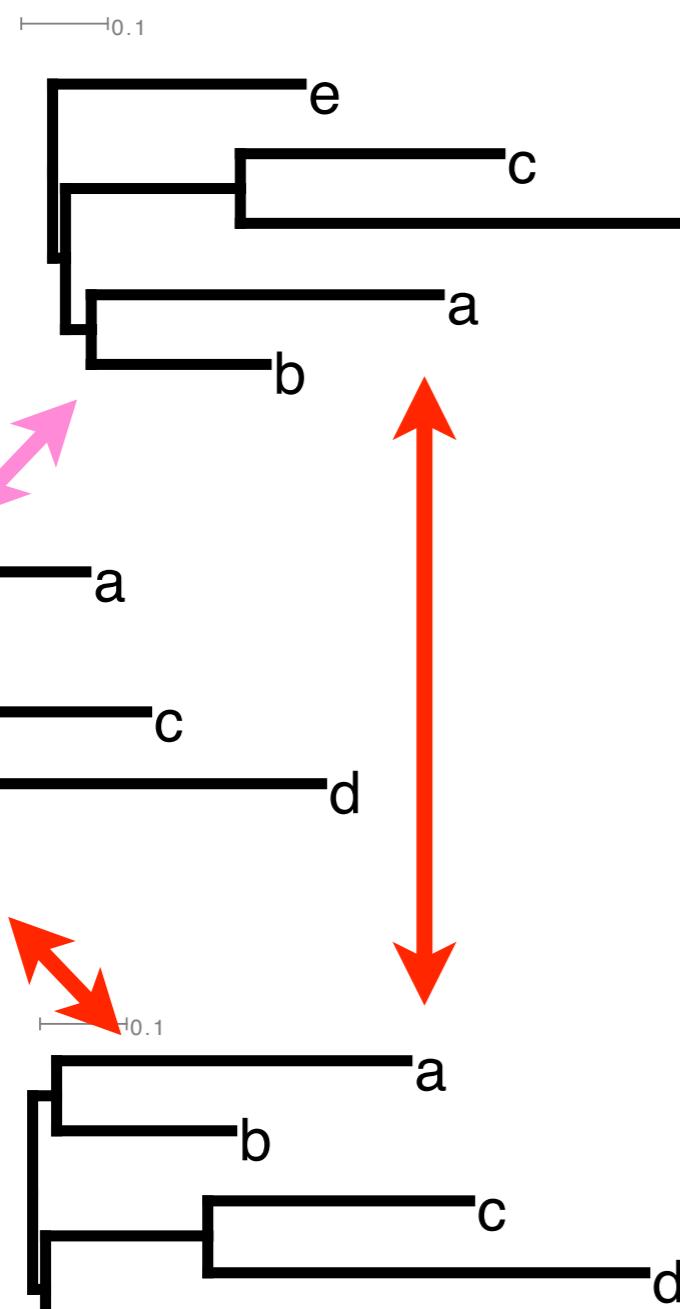
Most rooted tree figures use a “rectangular” rather than a “diagonal” representation

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

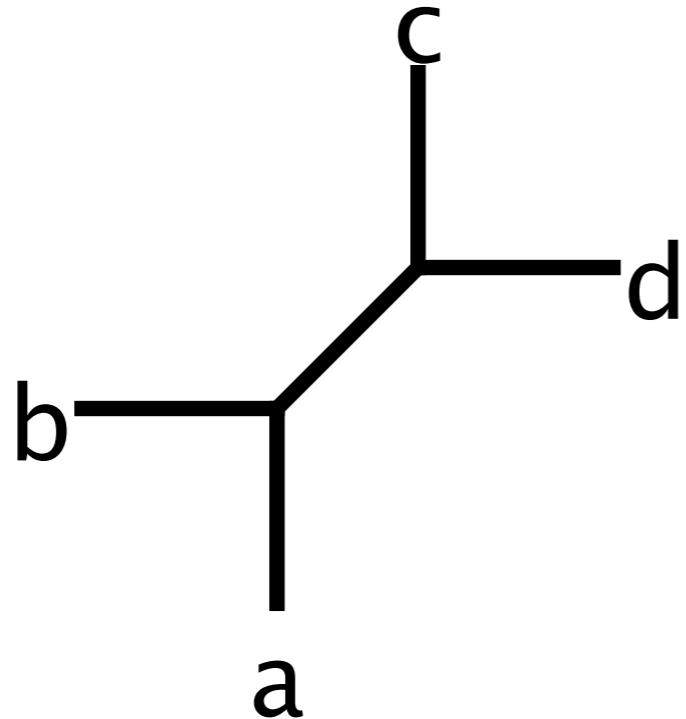
Trees with **different topologies**...  
... describe different sets of "relatedness statements" between taxa

# Unrooted Phylogenies

# Unrooted Trees

There's no root on the tree...

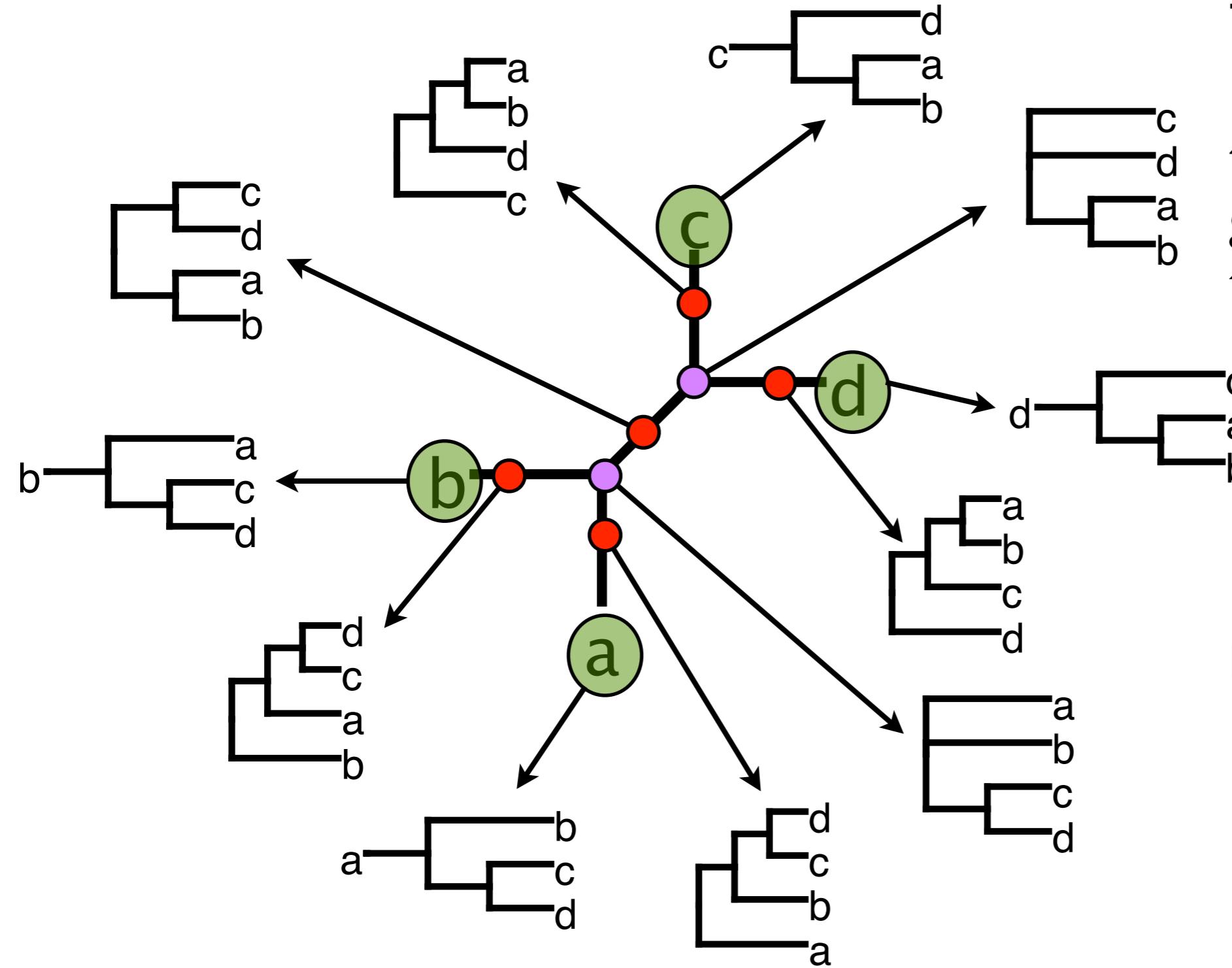
...which is usually interpreted as meaning that these taxa are related by a rooted tree but we don't know where the root is



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

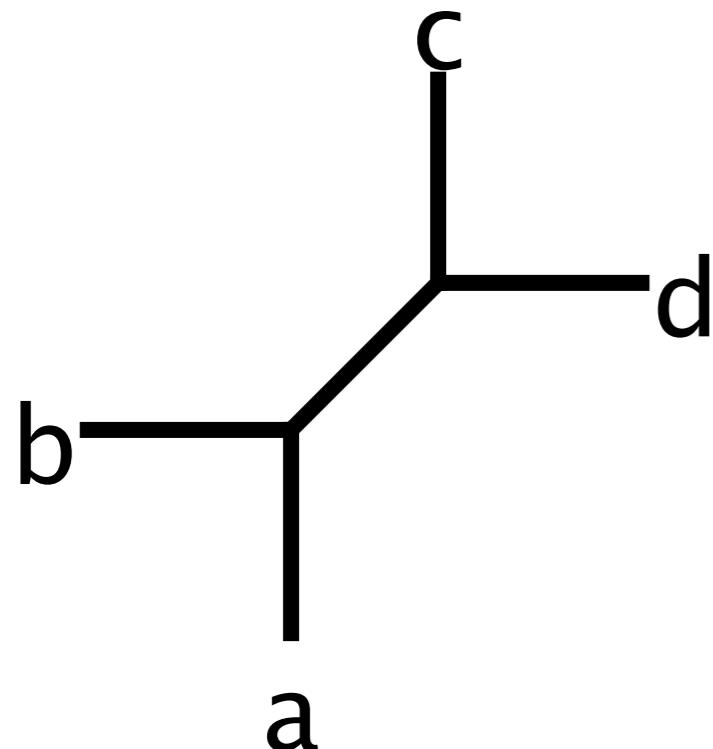
Unrooted trees can be rooted on their:

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

# Quiz

Assume we estimate an unrooted tree and have no additional data to infer the location of the root

In this case, which (if any) of the following statements would we be justified in making about the pattern of relatedness of the taxa shown in the tree?

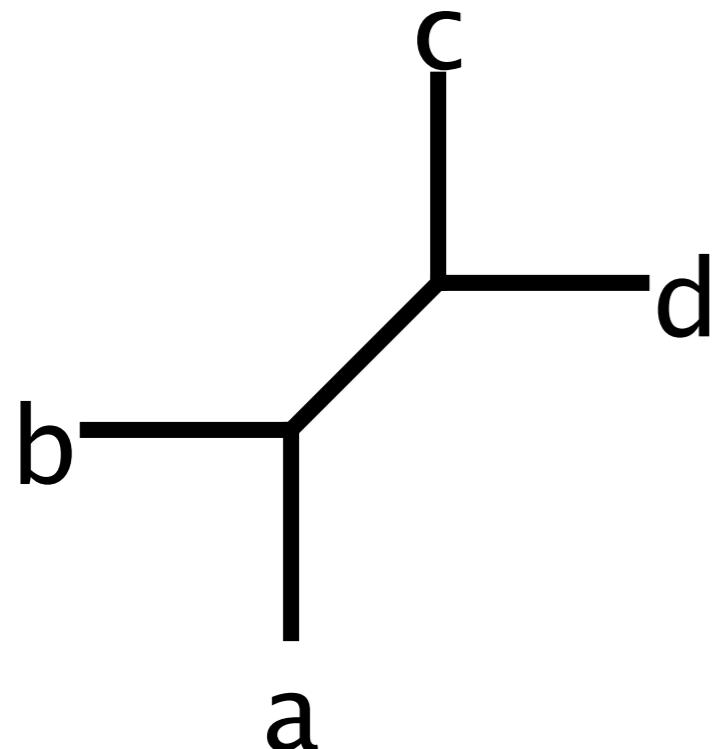


- d more closely related to:**
1. **a** than it is to **b** or **c**
  2. **b** than it is to **a** or **c**
  3. **c** than it is to **a** or **b**

# Quiz

We aren't justified in making any of these statements, as the tree is unrooted, and none of them is true under all possible footings 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 trues, in others it is not

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

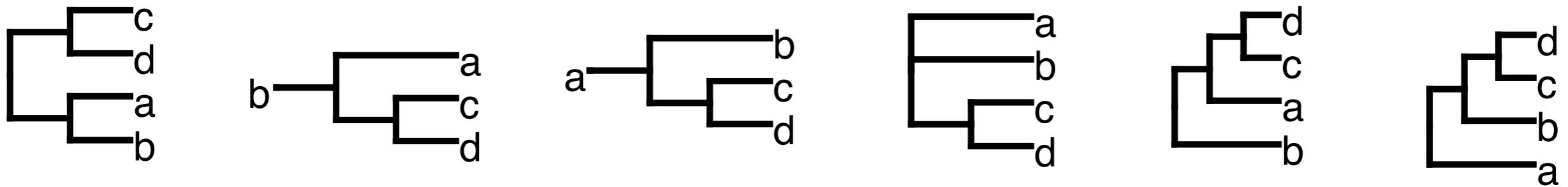
- true
- false

**d more closely related to:**

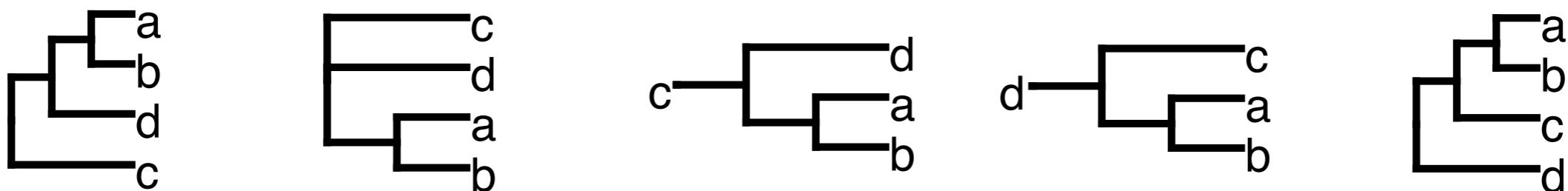
1. **a** than it is to **b** or **c**
2. **b** than it is to **a** or **c**
3. **c** than it is to **a** or **b**

# 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



# Example Phylogeny Estimation Workflow

# Phylogenetic Workflows

---

Every analysis is different - no “one size fits all” approach!

But there are concepts/features/tasks/tools common to many analyses

We present **example** analyses to highlight some of these

We'll look at similar workflows/analyses from two different perspectives

- considering a ‘standard’ phylogenetic analyses as a ‘standard’ statistical analyses
- highlighting tasks, tools, and concepts found in many molecular phylogenetic analyses

# Example Phylogeny Estimation Workflow

I. phylogenetic analysis as a standard  
statistical analysis

# 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

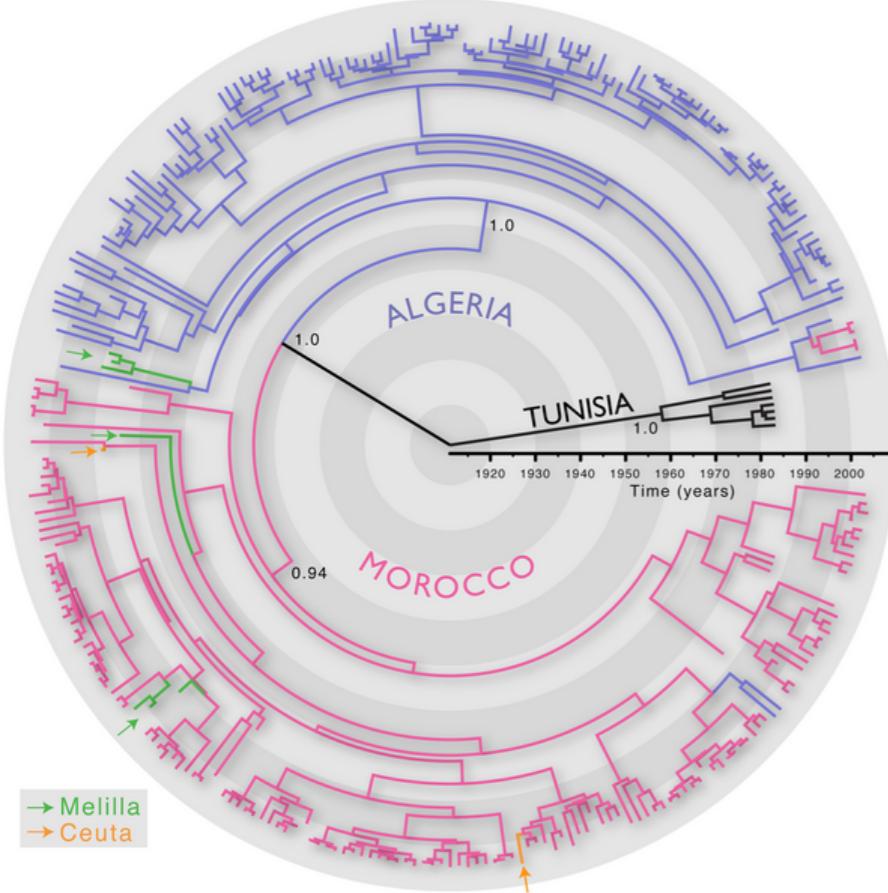
# Example Phylogeny Estimation Workflow

## 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



We saw earlier how our estimate of the **topology parameter** of this phylogenetic analysis of dog rabies viruses informs our answer to this question

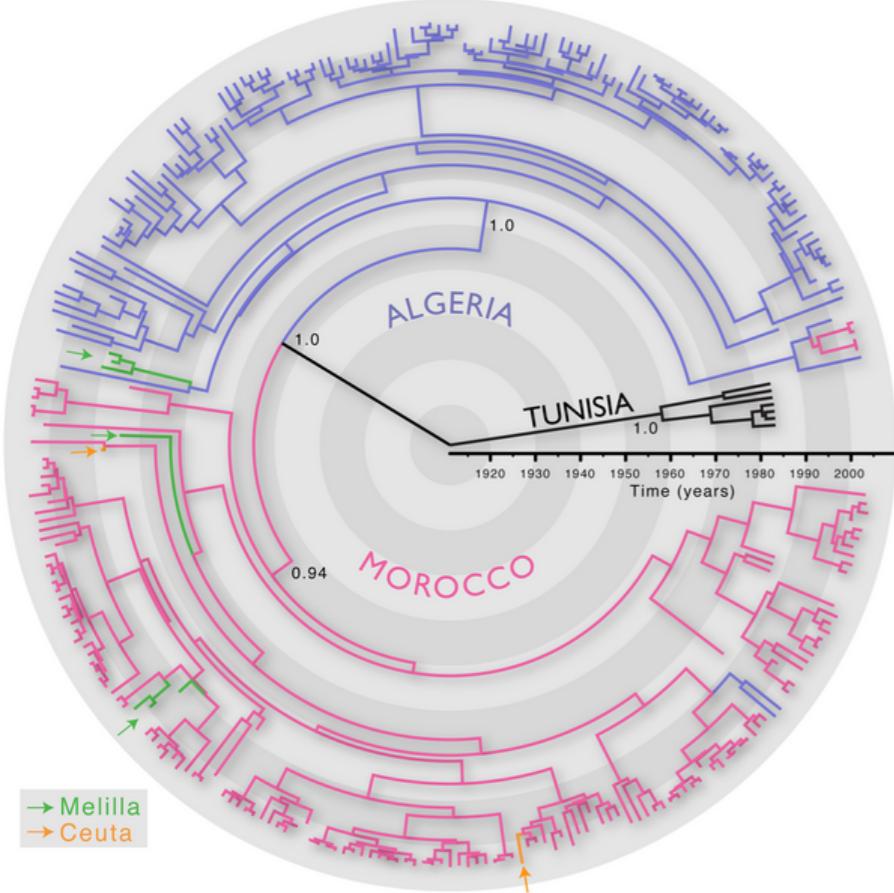
# Example Phylogeny Estimation Workflow

## 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

# Example Phylogeny Estimation Workflow

## 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

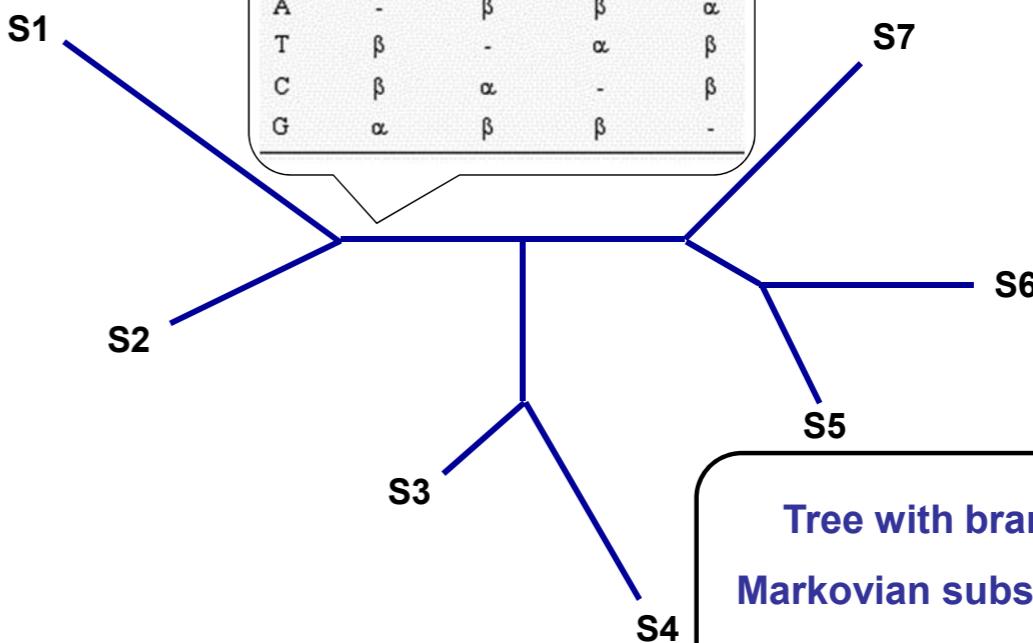
- A tree topology (to be estimated,  $n^n$ )
- Branch lengths (to be estimated,  $2n-3$ )
- A substitution model (to be (partly) estimated, 1, 3, 4, ...208 ...)
- A distribution of site rates (to be estimated, 1, 2, ...)

Olivier Gascuel – Phylogenetic models – ISCB-ASBCB Casablanca 2013



### The full probabilistic model

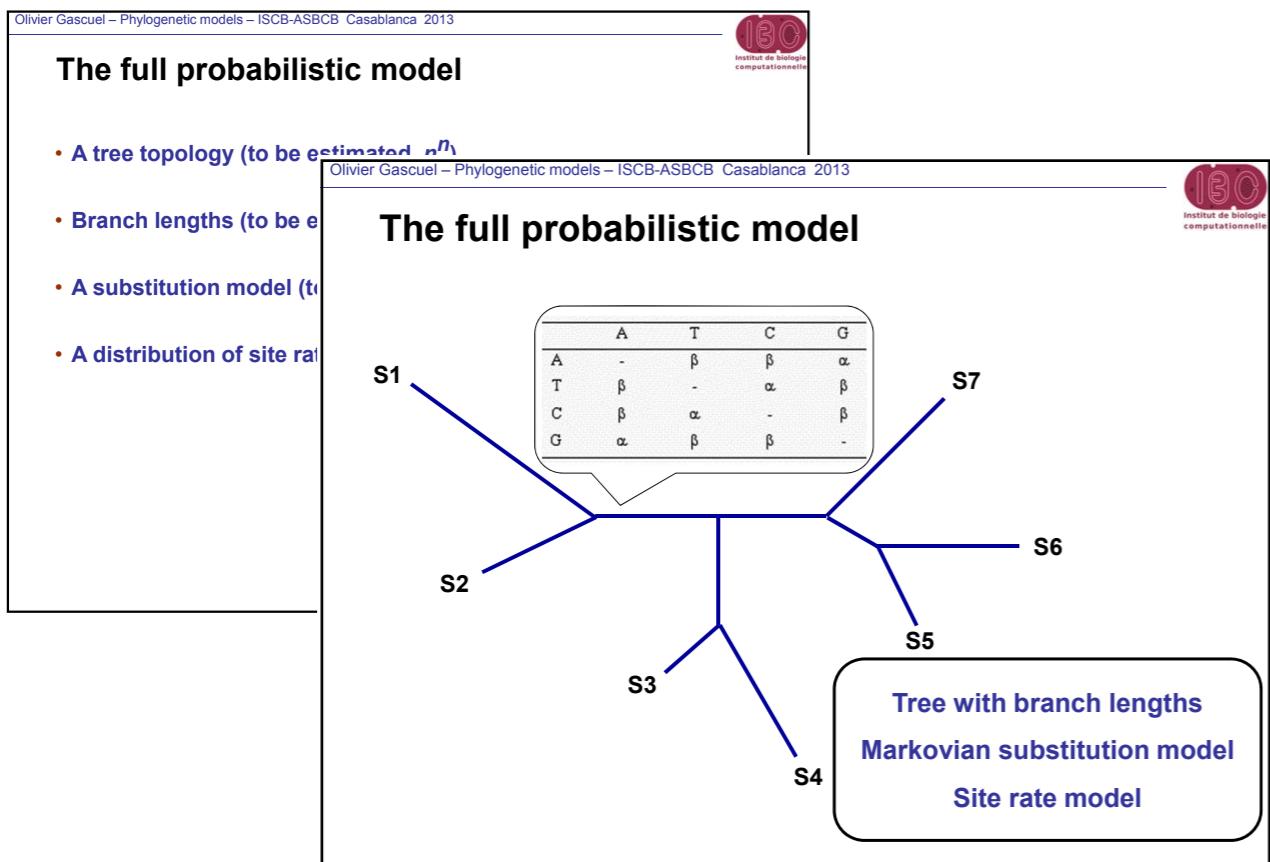
	A	T	C	G
A	-	$\beta$	$\beta$	$\alpha$
T	$\beta$	-	$\alpha$	$\beta$
C	$\beta$	$\alpha$	-	$\beta$
G	$\alpha$	$\beta$	$\beta$	-



Tree with branch lengths  
Markovian substitution model  
Site rate model

# Example Phylogeny Estimation Workflow

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



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

Sometimes, however, it's not the topology, but some 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

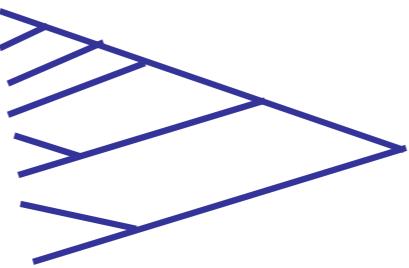
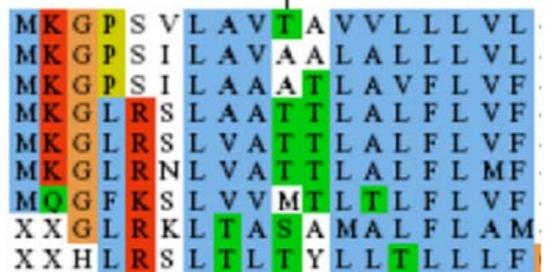
## 3. Collect observations informative about the model parameter(s)

Olivier Gascuel – Phylogenetic models – ISCB-ASCB Casablanca 2013

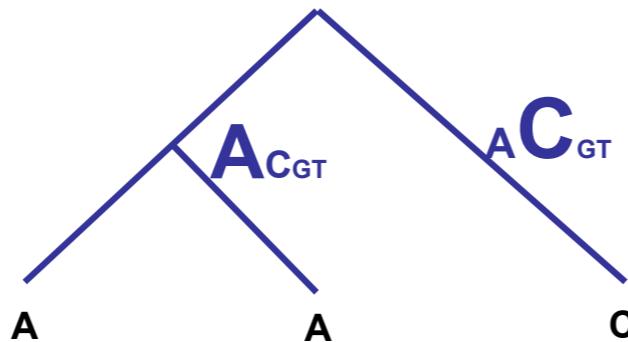


### Modeling sequence evolution: standard assumptions

MOUSE  
RAT  
RABBIT  
HUMAN  
DOG  
ELEPHANT  
COW  
CHICKEN  
FUGU



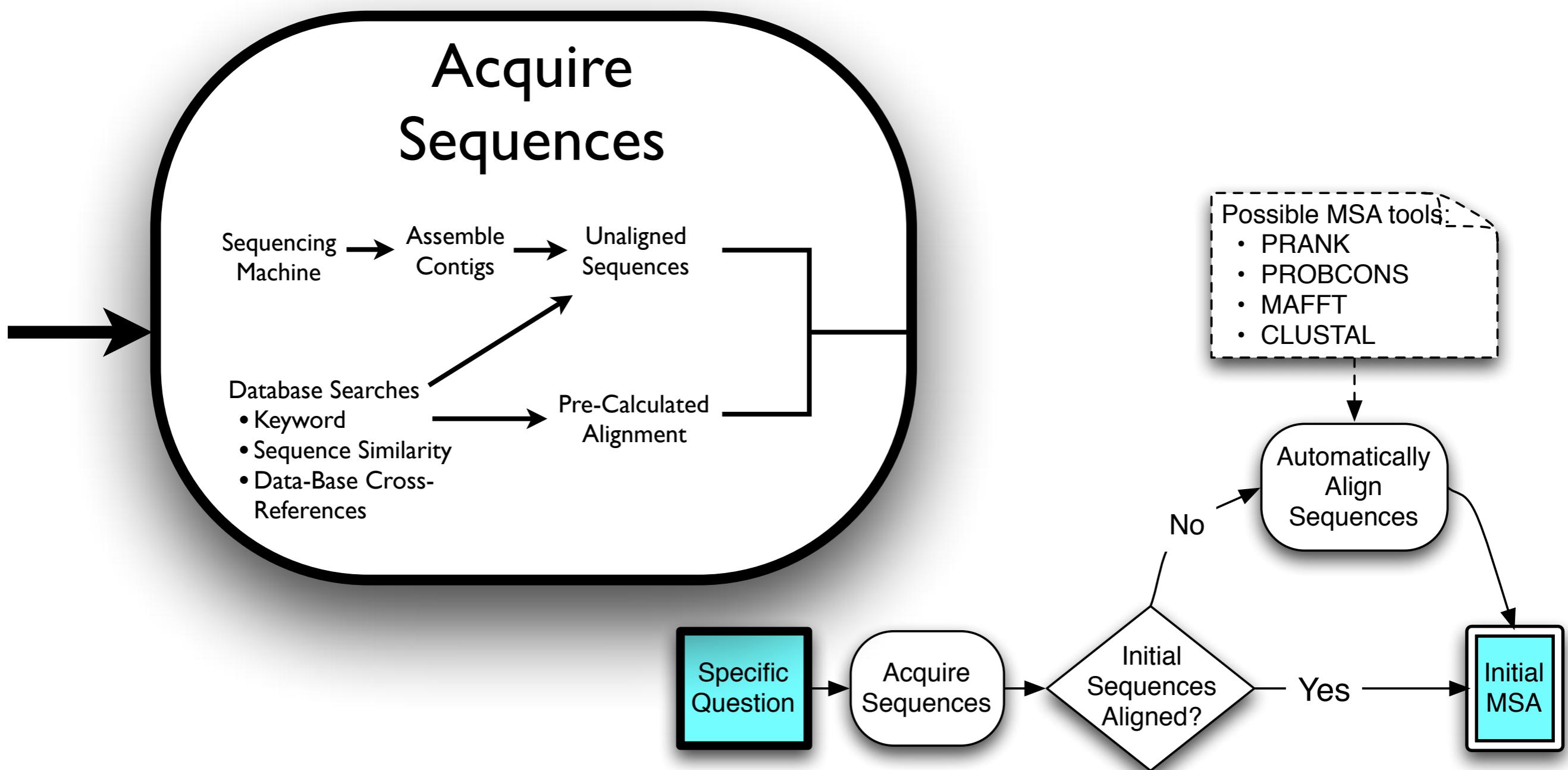
We aim at explaining the data (alignment) using a probabilistic scenario of the evolution of each of the sites along a phylogeny



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

# Example Phylogeny Estimation Workflow

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



# 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

---

4. Find best estimate(s) of the parameter(s), conditioned on these observations
5. Estimate sampling/random error associated with parameter estimate

In upcoming demonstration:

- Choose a substitution model
- Estimate a phylogeny using this model, with aLRT

# Example Phylogeny Estimation Workflow

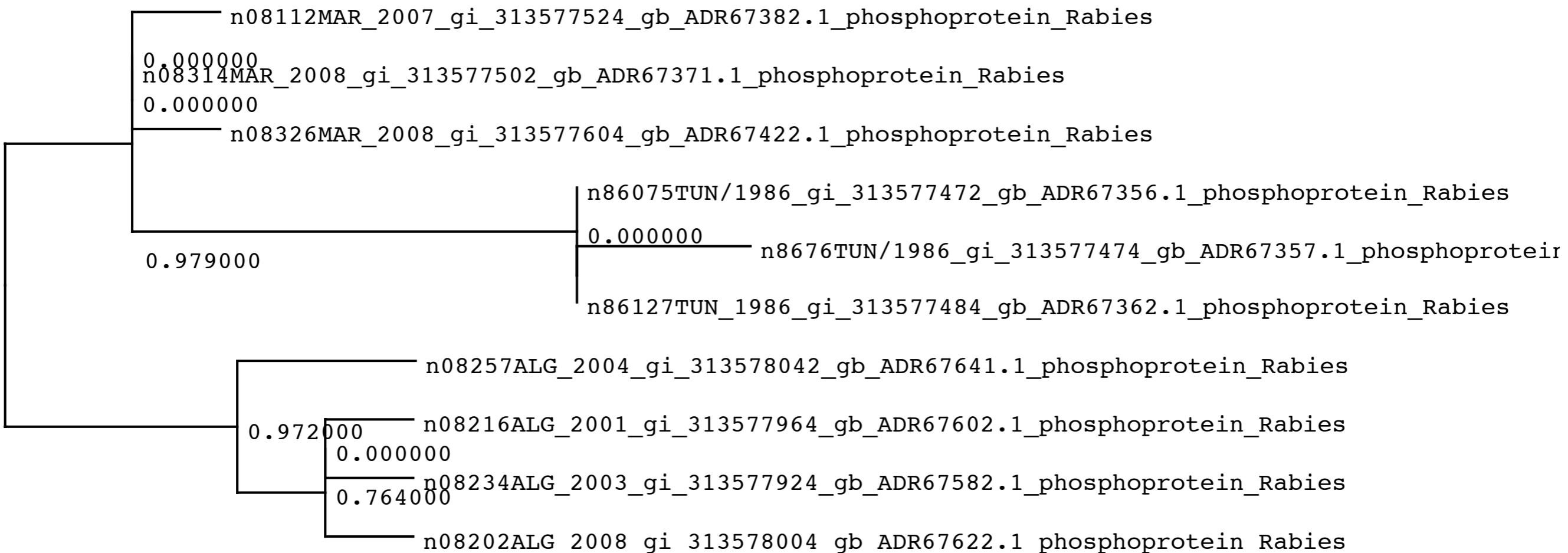
---

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 section:

“phylogenetic analysis as a standard statistical analysis”

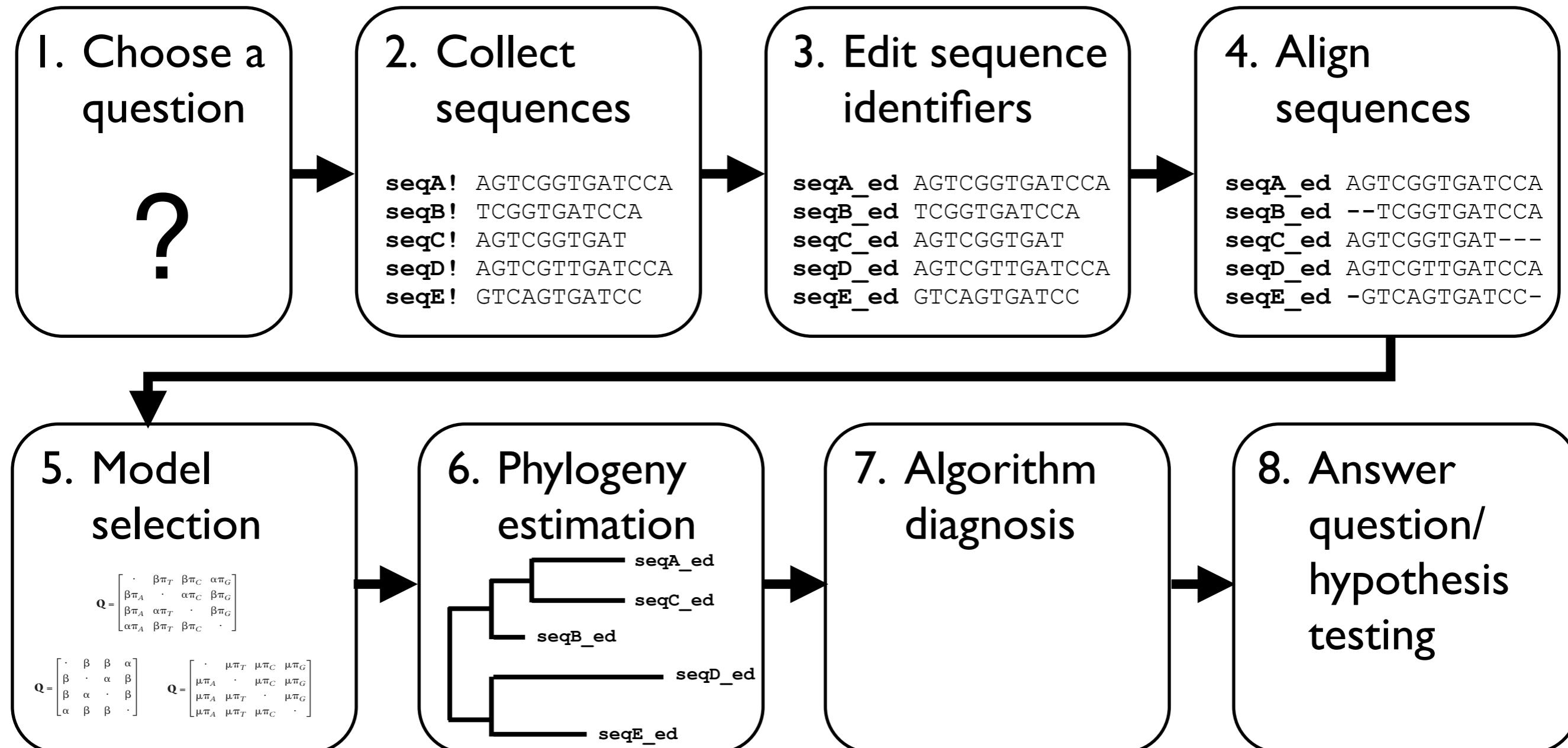
- **demonstration** with North African dog rabies viruses
- **exercise** with Louisiana gastroenterologist example

described in this HTML document [interpretingPhylogenies.html](#)

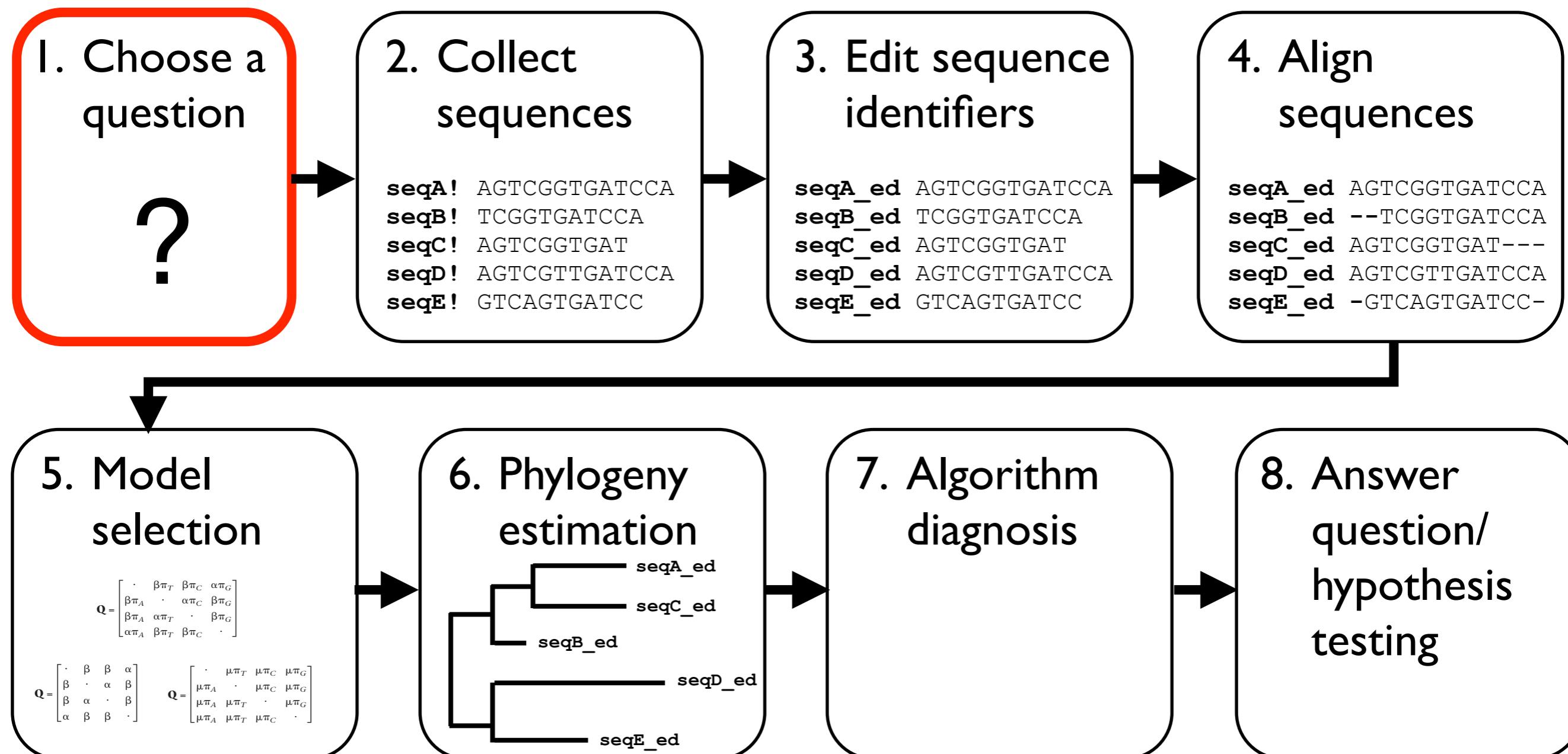
# Example Phylogeny Estimation Workflow

## 2. Common tasks and concepts found in many phylogenetic analyses

# Example Phylogeny Estimation Workflow

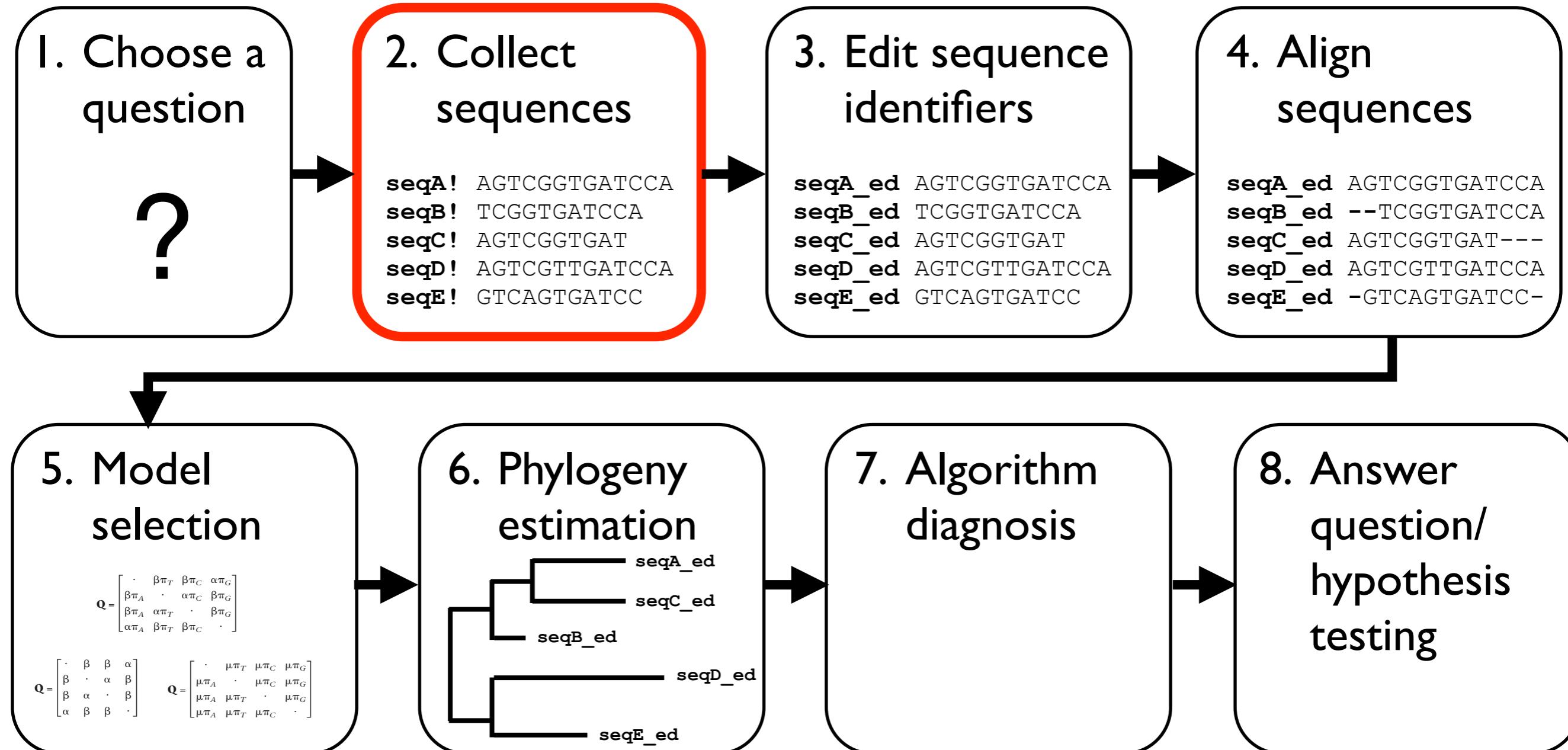


# Example Phylogeny Estimation Workflow



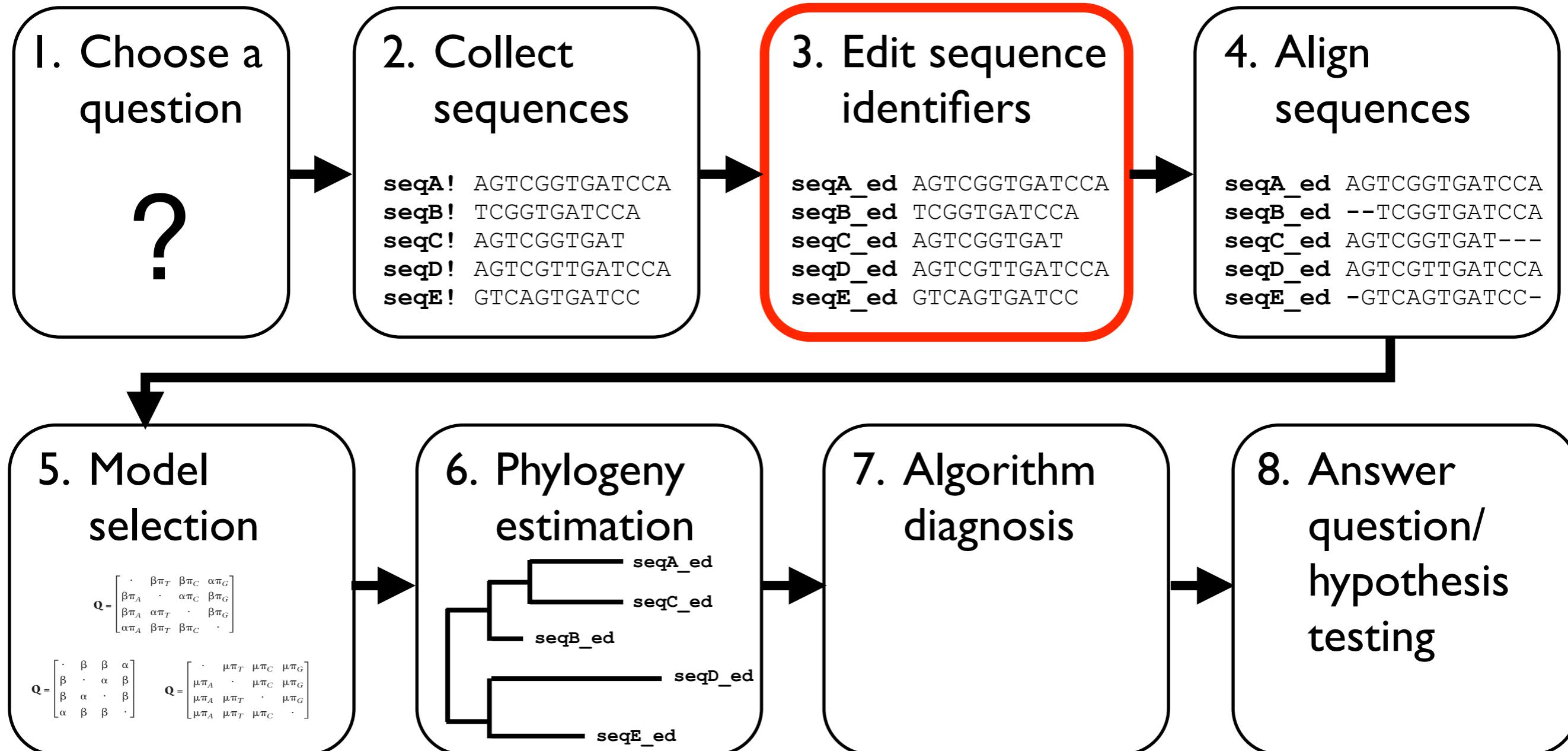
I. Choose a question: me, today, using your brain

# Example Phylogeny Estimation Workflow



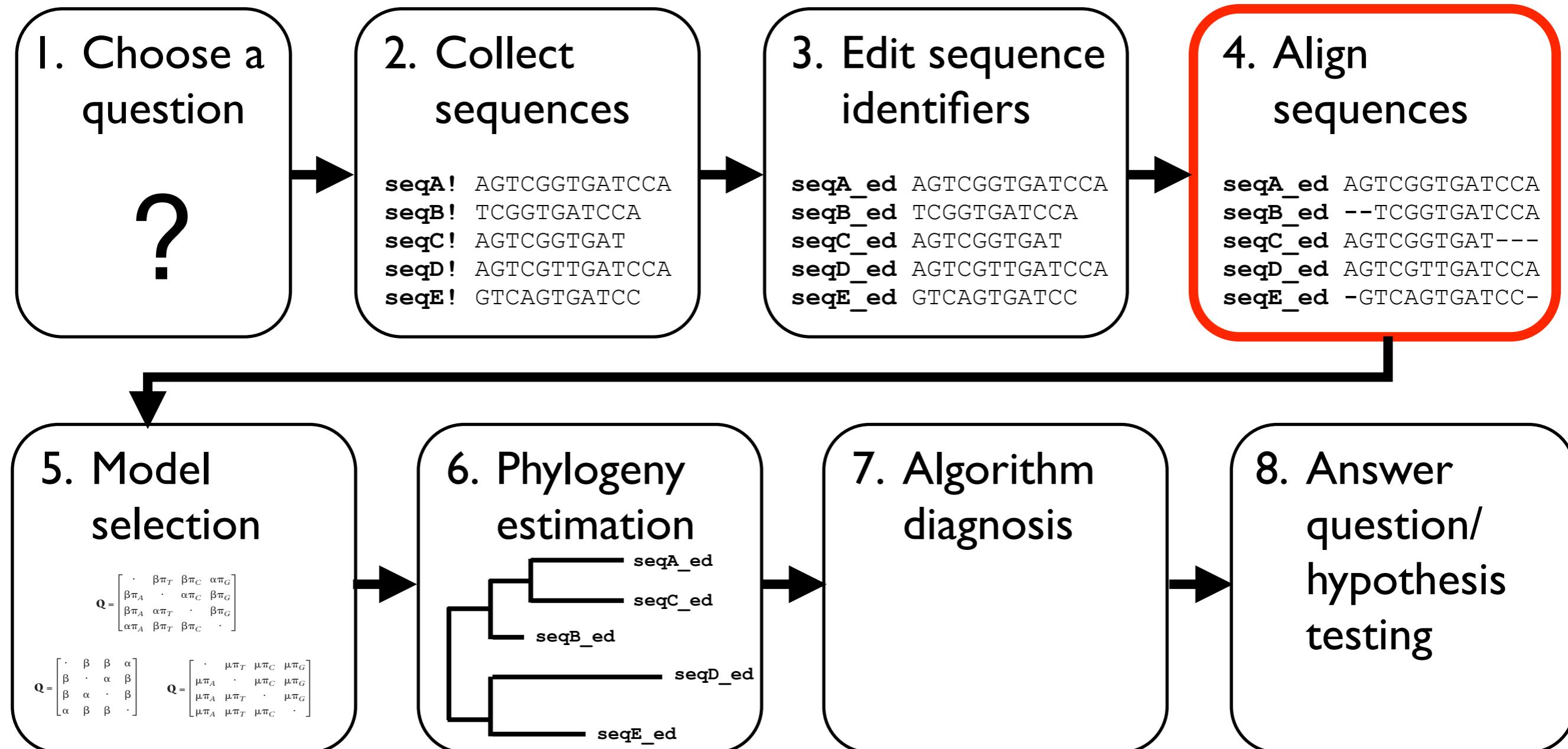
2. Collect sequences: Stephen, BLAST, MCL, etc.

# Example Phylogeny Estimation Workflow



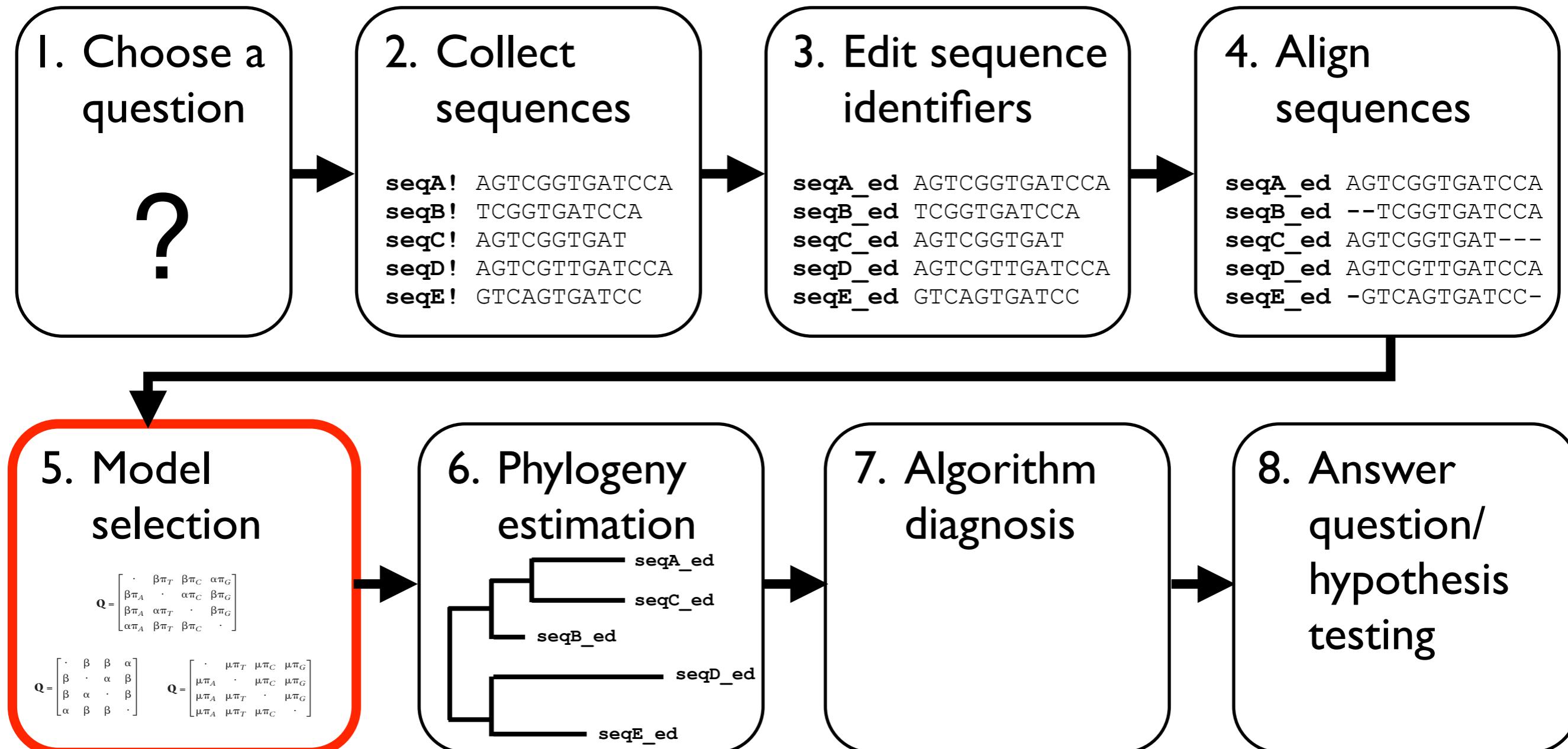
3. Edit sequence identifiers: me, today, text editors or scripting tools

# Example Phylogeny Estimation Workflow



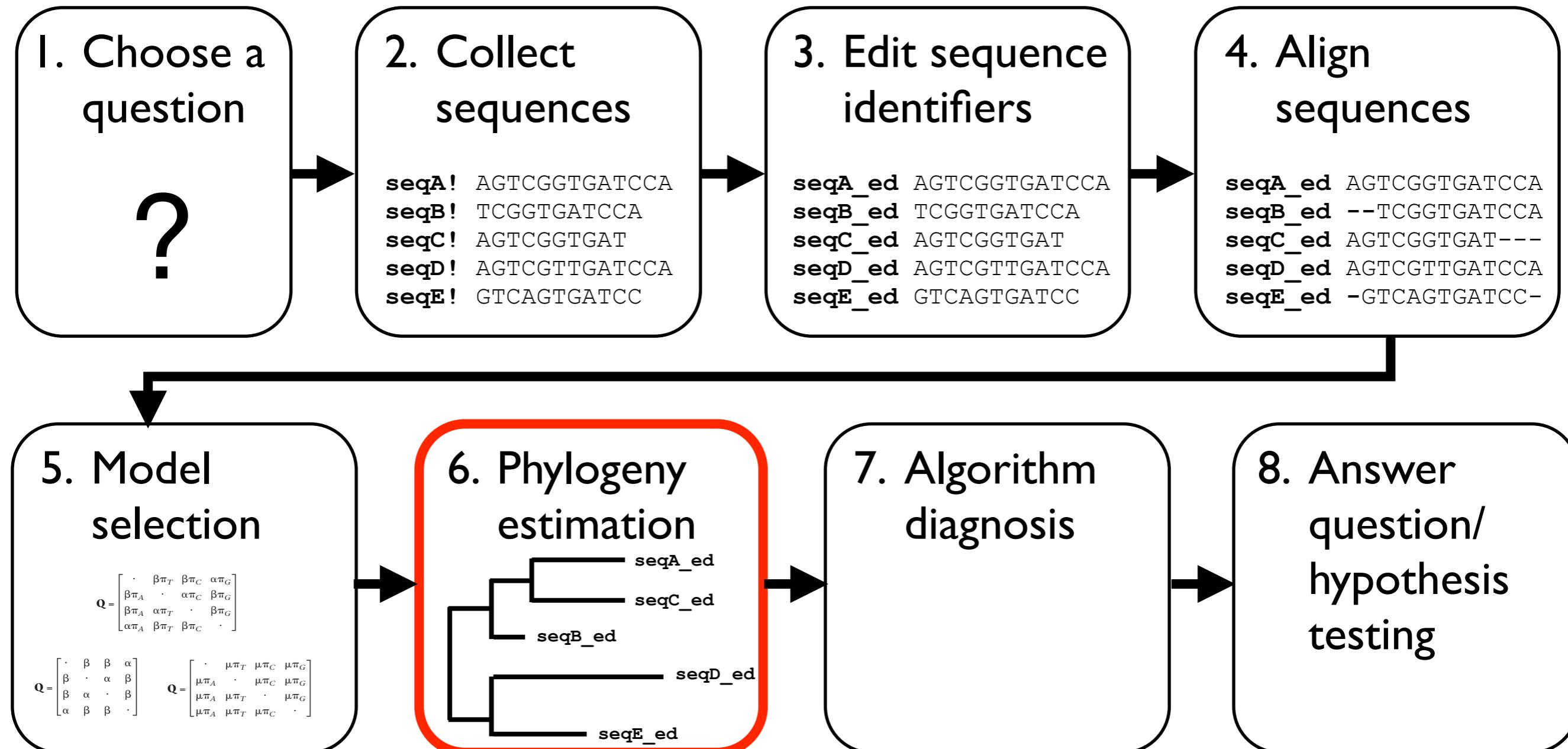
4. Align sequences: Ben, BaliPhy Prank, muscle, Probcons, mafft etc.

# Example Phylogeny Estimation Workflow



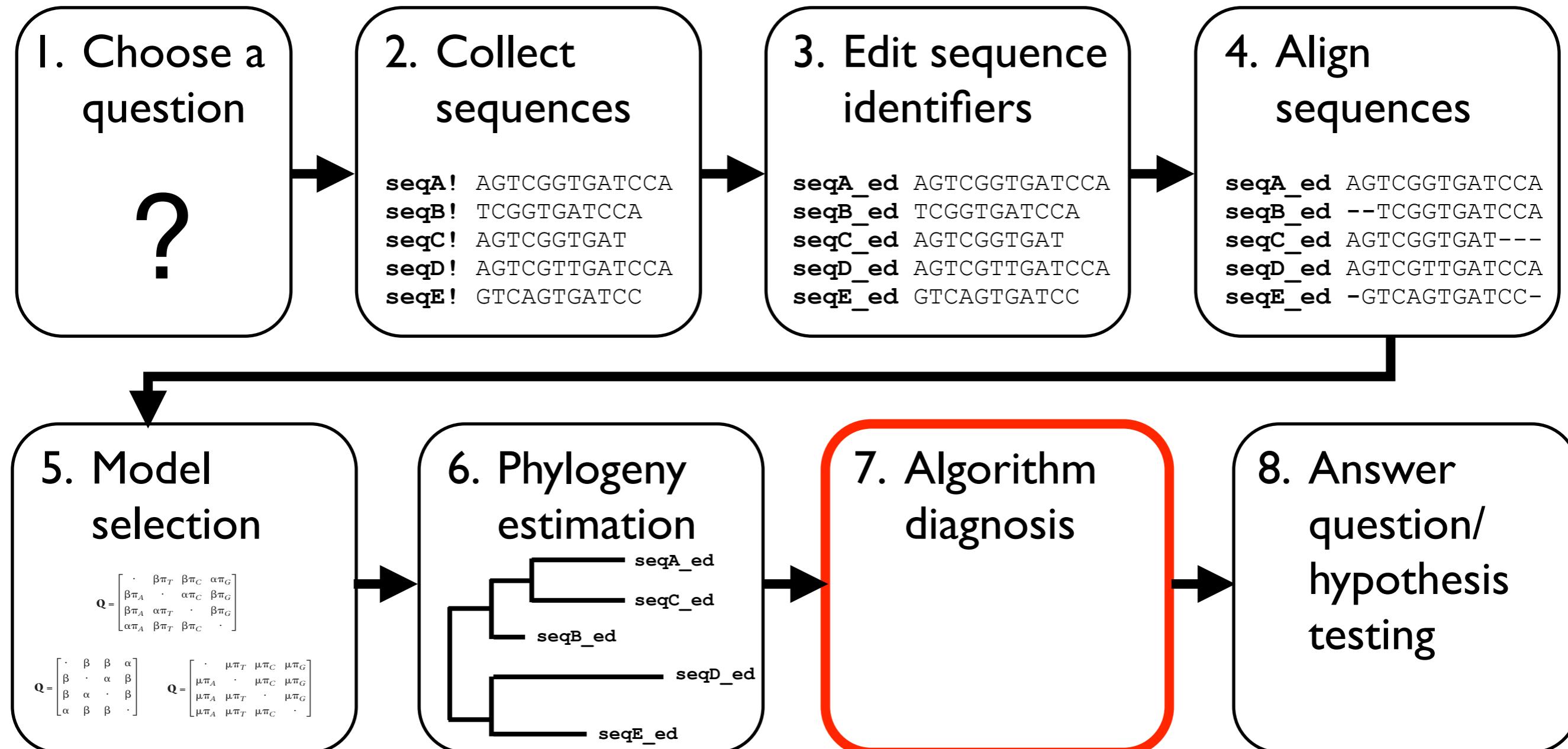
5. Model selection: Maria and Asif, PAML, jModelTest

# Example Phylogeny Estimation Workflow



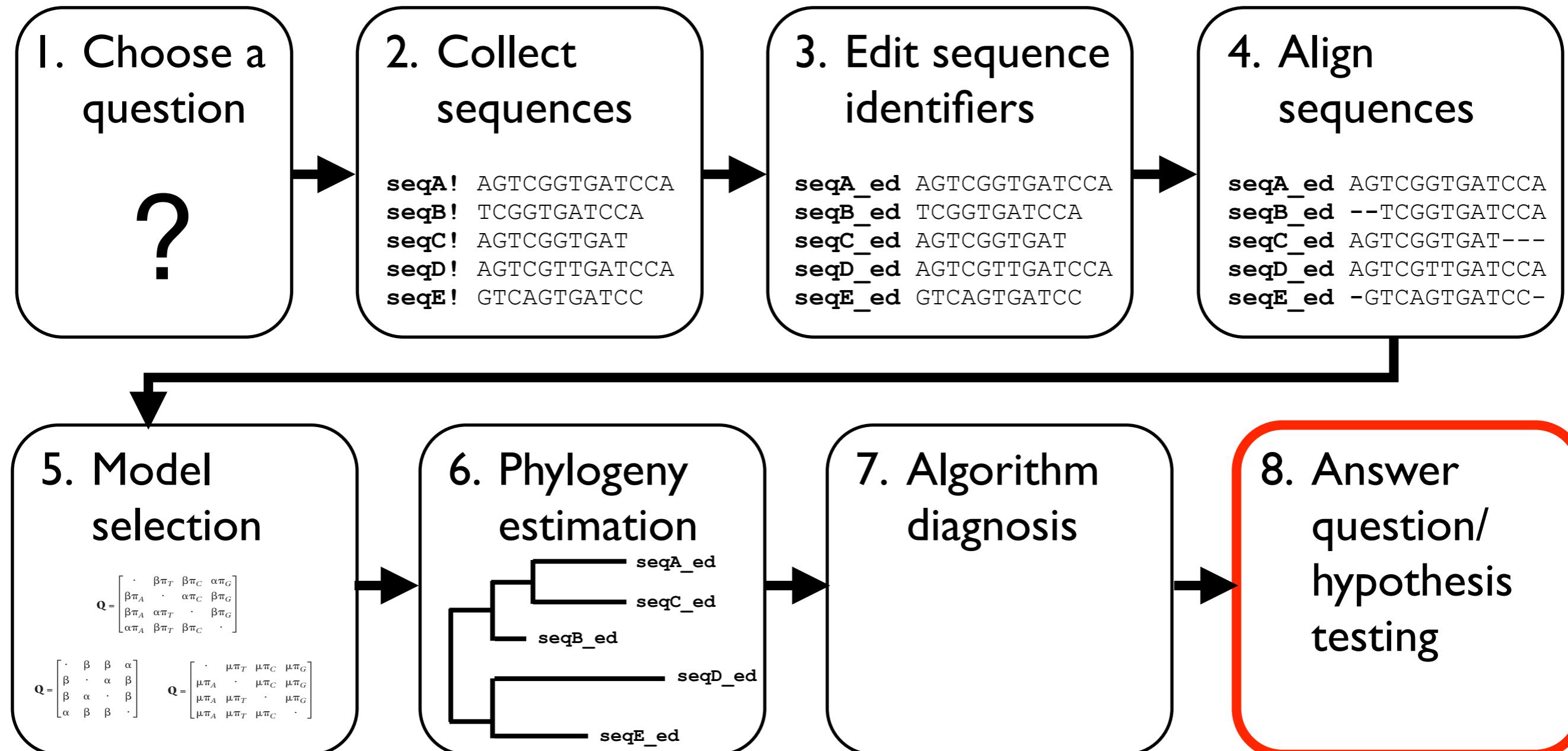
6. Phylogeny estimation: **Maria**, **Brian**, **Olivier**, **RAXML**, **PhyML**, etc.

# Example Phylogeny Estimation Workflow



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

# Example Phylogeny Estimation Workflow



8. Answer question/hypothesis testing: Asif, Nick, Tracy, Jeff, Maria, Brian, Olivier, aLRTs, AIC, Bayes factors

# Example Phylogeny Estimation Workflow

---

## Demo and Exercises

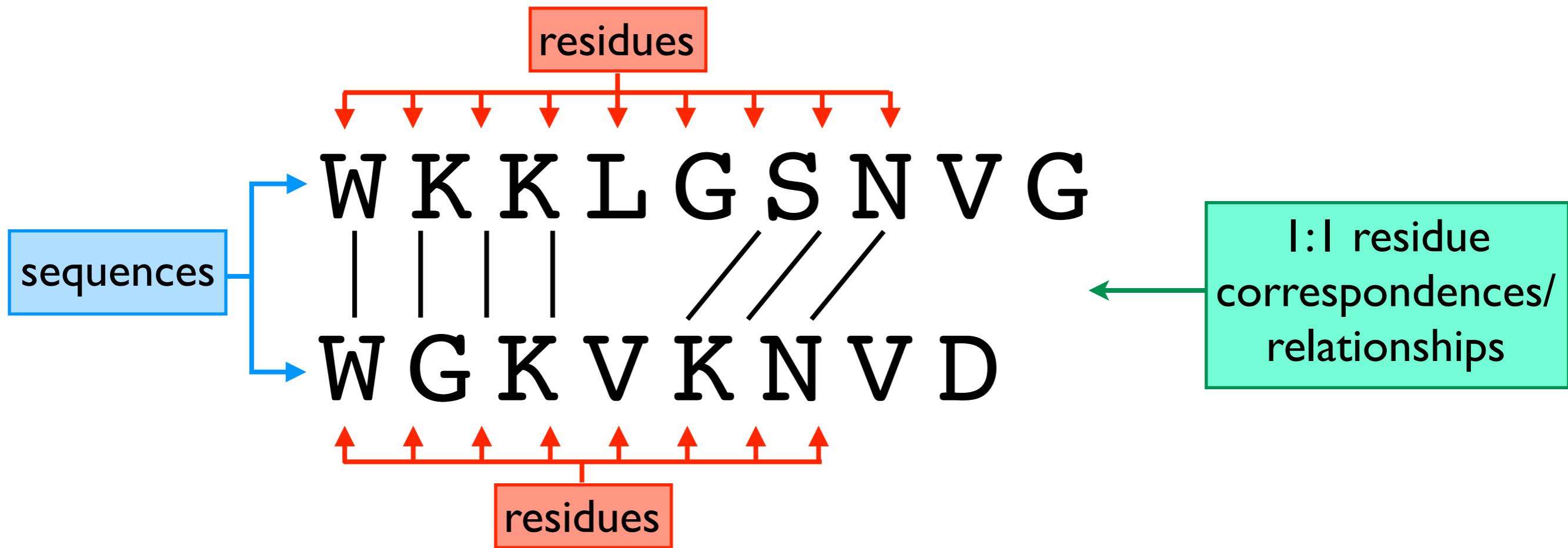
We'll follow a demonstration, and you'll have a chance to try this kind of phylogenetic workflow yourself, using the

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

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

# Alignments

# "Anatomy" of a Sequence Alignment

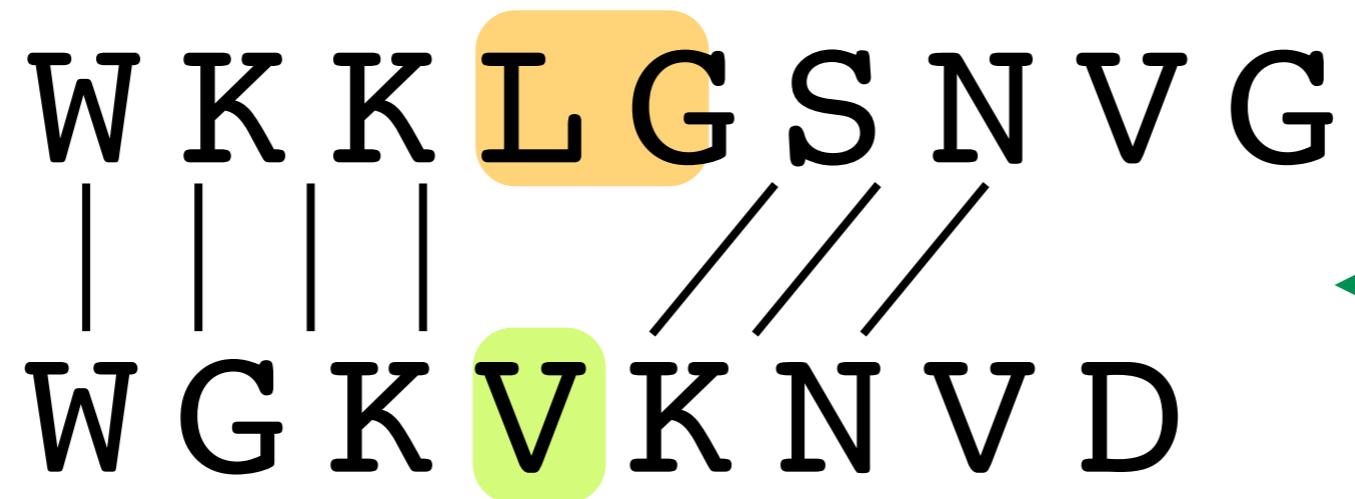


## I:I residue correspondences/relationships

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

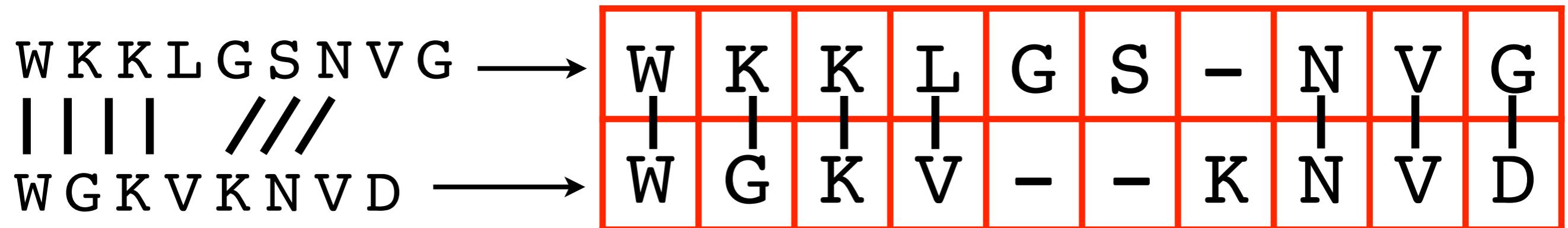
Residue has no link to a residue in the top sequence

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

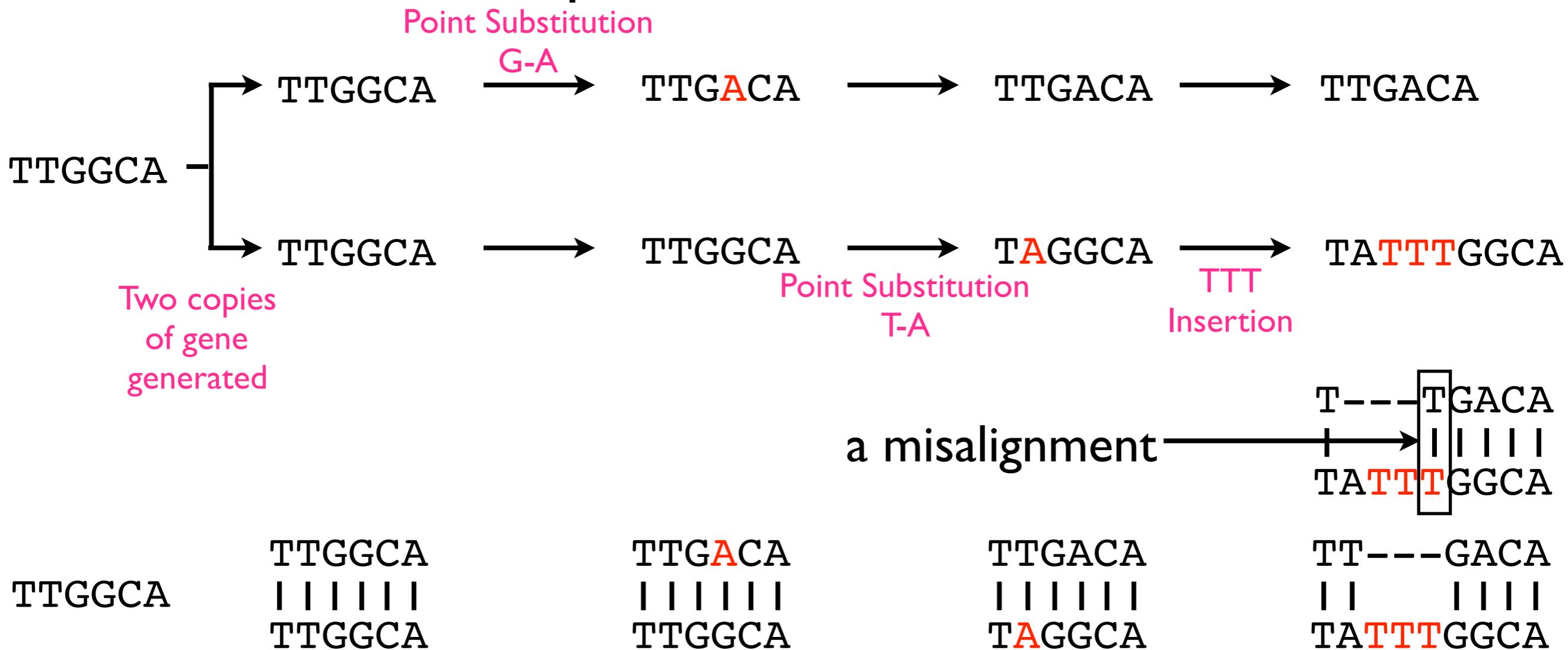
Residues in the same column are ‘linked’ (‘equivalent’? ‘homologous’?)

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

# Evolutionary "Equivalence"

Residues are "evolutionarily linked" when:

- you (believe) their evolutionary history can be traced back to the same residue in an (inferred) ancestral sequence...
- ... such that the only changes experienced during divergence from this ancestral residue were **point substitutions**



# Branch Lengths

# Branch Lengths

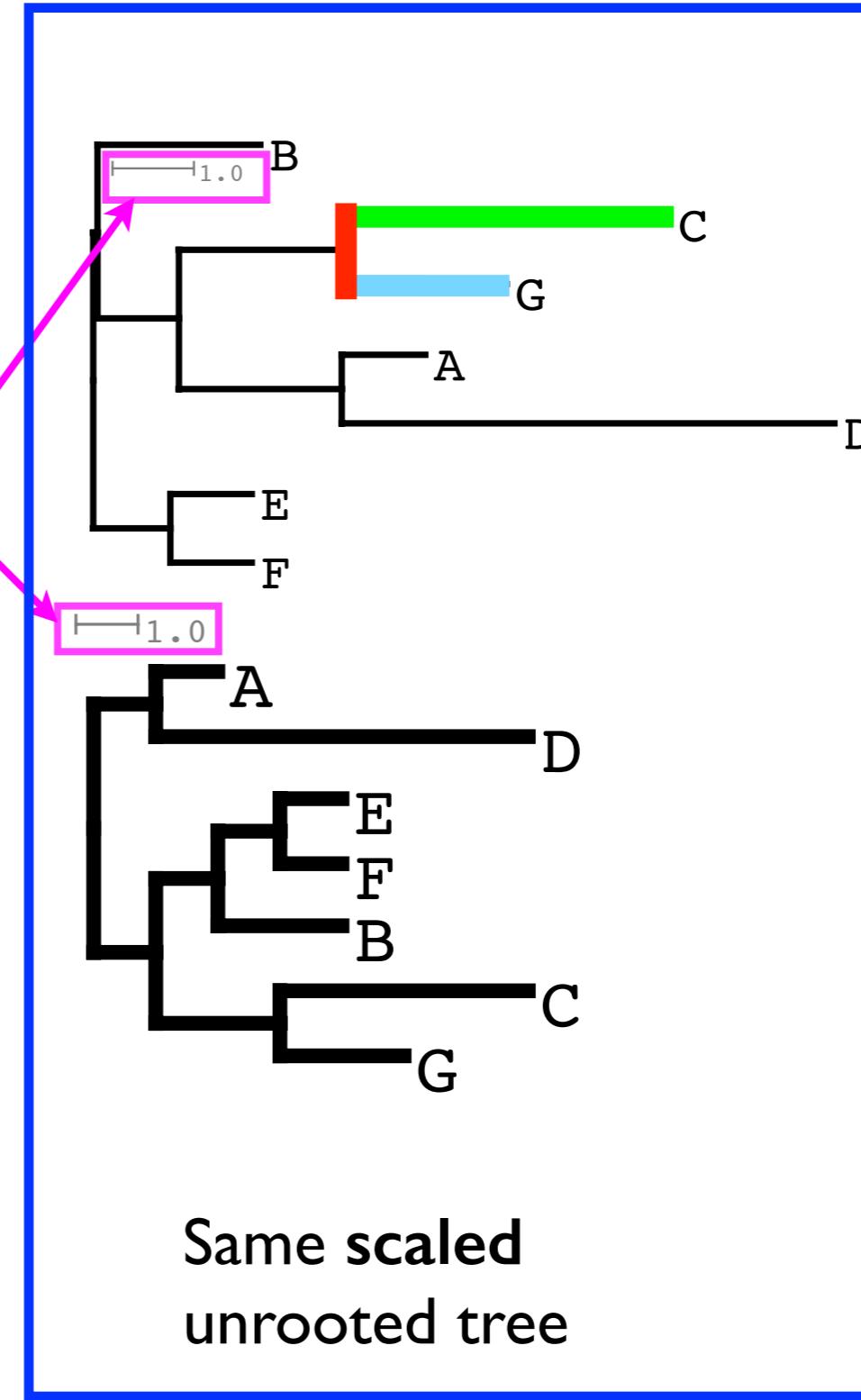
Branch length usually represents some measure of the difference/distance between TUs at ends of the branch

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



# Branch Lengths

Often 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

Often 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