Introductions

Monday 5th May 2014

EMBO Practical Course on Computational Molecular Evolution

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

Heraklion, Greece

Aidan Budd, EMBL Heidelberg

What You Get From a Course

Aidan Budd, EMBL Heidelberg

why participate in a course?

1. to learn

Learning

2. to build 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 getting to know each other…

speed dating

Aidan Budd, EMBL Heidelberg

Speed Dating: Aims

- make it easier to start chatting later in the course
- find people you have things in common with/get on with

Speed Dating: Format

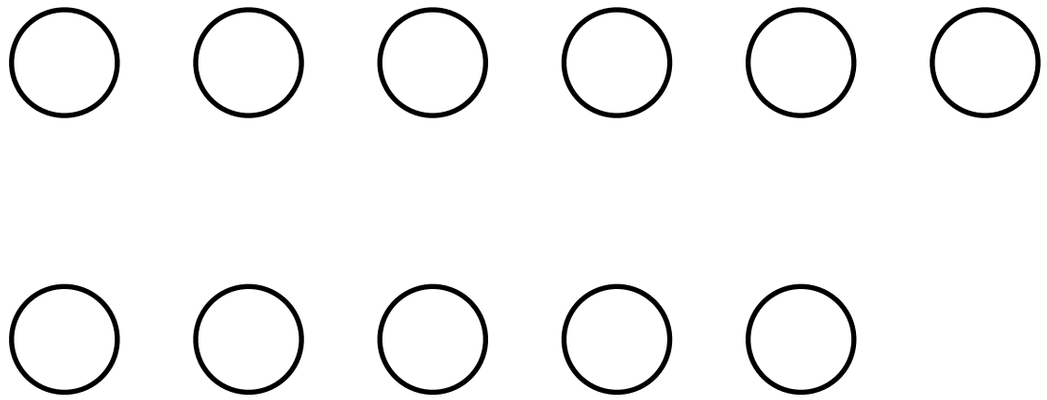
- meet other participants in many 1:1 chats
- tell each other
  - names
  - where you work
  - research topics
  - something people are often surprised to learn about you (e.g. I have three nationalities… am vain enough to choose clothing to fit my eye colour…)
  - try and find someone you know or somewhere you've been that you have in common

Speed Dating: Format

Stand, awkwardly, in two rows

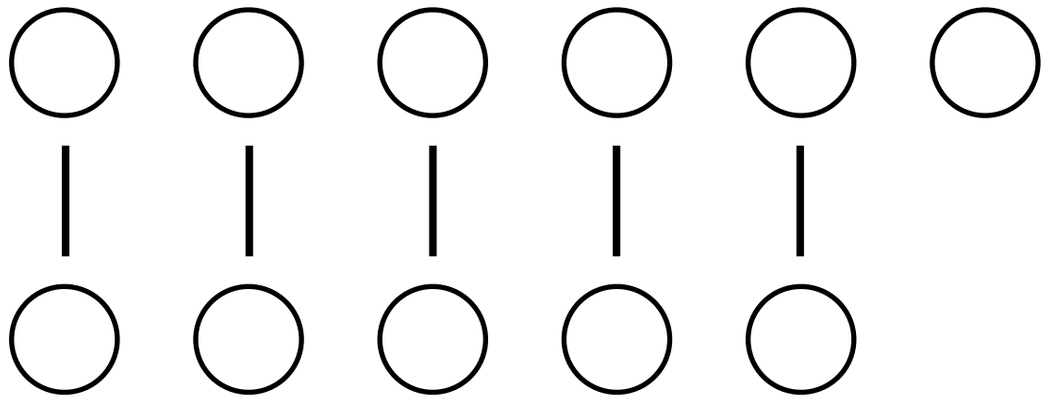
Face one person in the other row

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

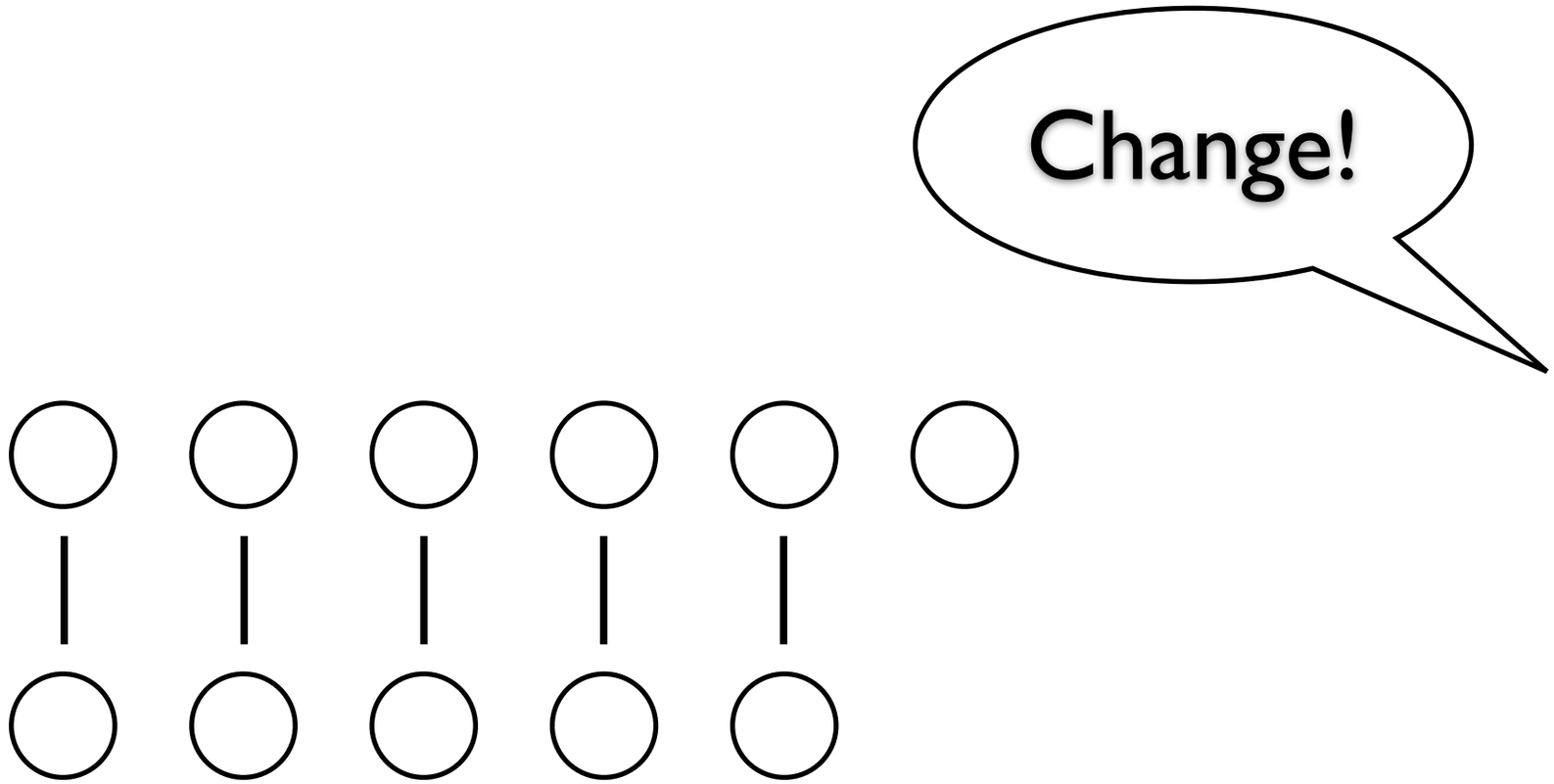


Speed Dating: Format

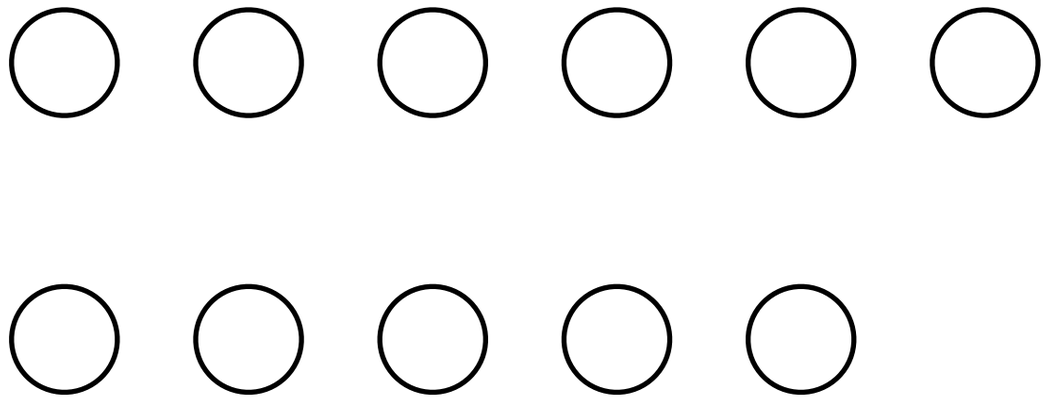
Chat!



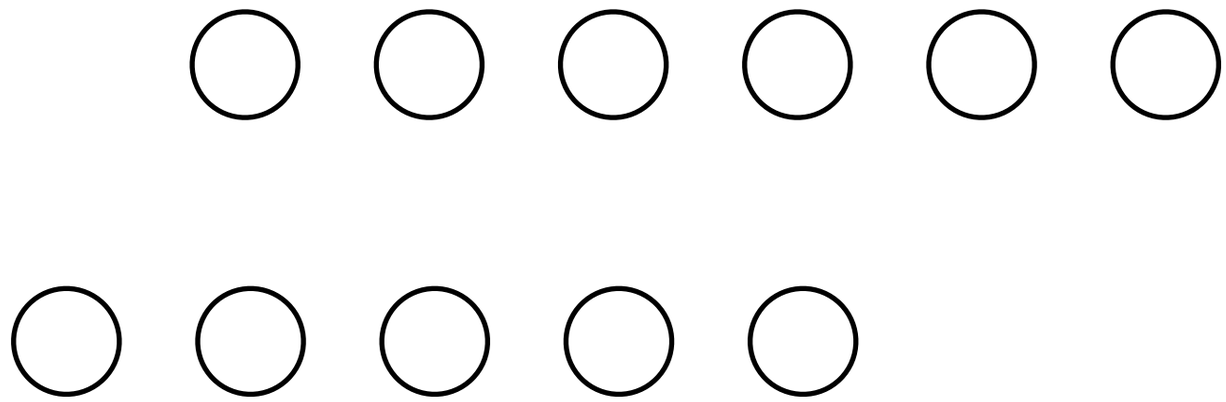
Speed Dating: Format



Speed Dating: Format

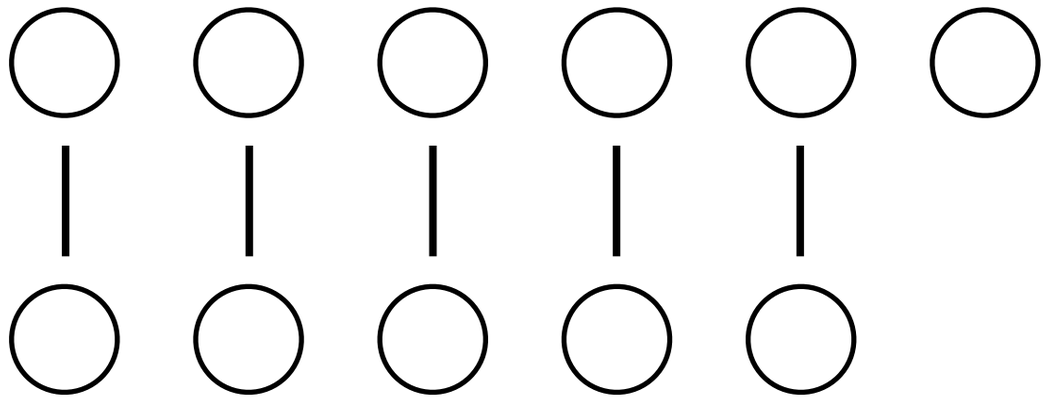


Speed Dating: Format



Speed Dating: Format

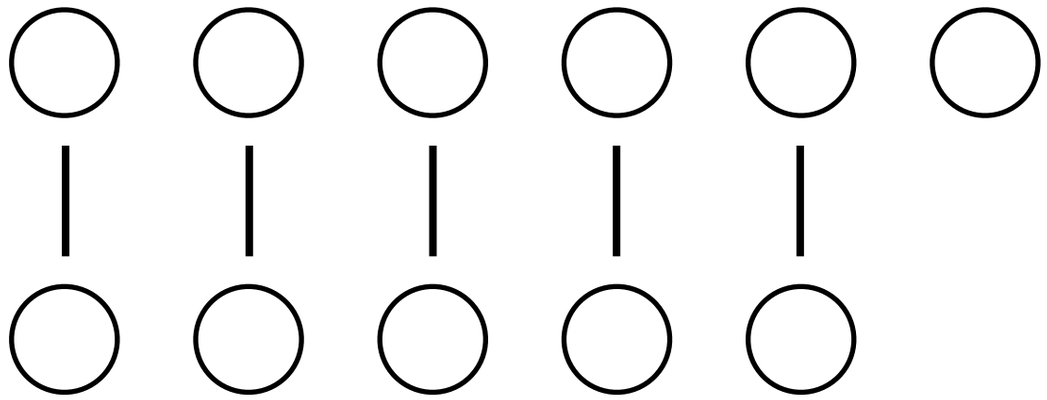
Chat!



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

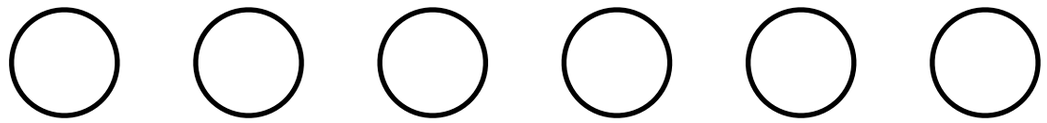
Speed Dating: Format

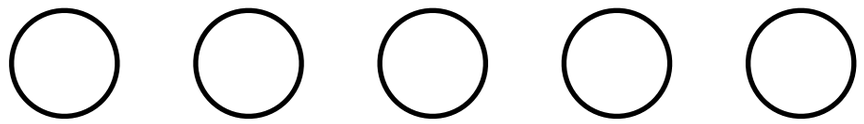
Chat!



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

Speed Dating: Format

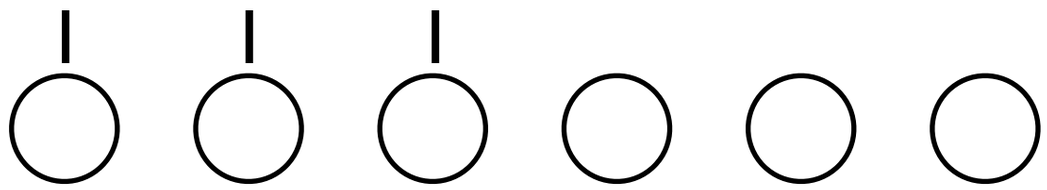


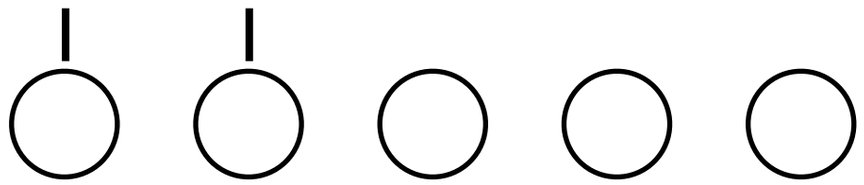


then split each row into two new rows

Speed Dating: Format

Chat!





make two new rows

and start again with the chat…

Session Homepage:

http://tinyurl.com/xxxxx

Interpreting Molecular Phylogenetic Trees

Monday 5th May 2014

EMBO Practical Course on Computational Molecular Evolution

Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Helenic 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

Aidan Budd, EMBL Heidelberg

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

Often non-trivial to interpret phylogenetic trees appropriately

Understanding better how to interpret such results helps us design better analyses

This session aims to provide:

An overview of terminology and concepts associated with phylogenetic trees

Experience with some commonly-used tools for examining phylogenies

Before we start

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

Why do people
care about phylogenetics?

and

How do people use
phylogenetics?

Aidan Budd, EMBL Heidelberg

Why Should We Care about Phylogenetics?

Useful to consider this question as...

... learning studies show: the more relevant/important a topic is to us...

... the more attention we pay when people talk about it...

... the more effectively we learn about the topic

You all already care, as you need to build/interpret them for your research

By presenting examples of specific applications of phylogenies, hopefully make it **even more relevant/important for you**

...and provides a context for considering **common ways in which phylogenies are used/interpreted**

Why Care About Phylogenies? 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

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

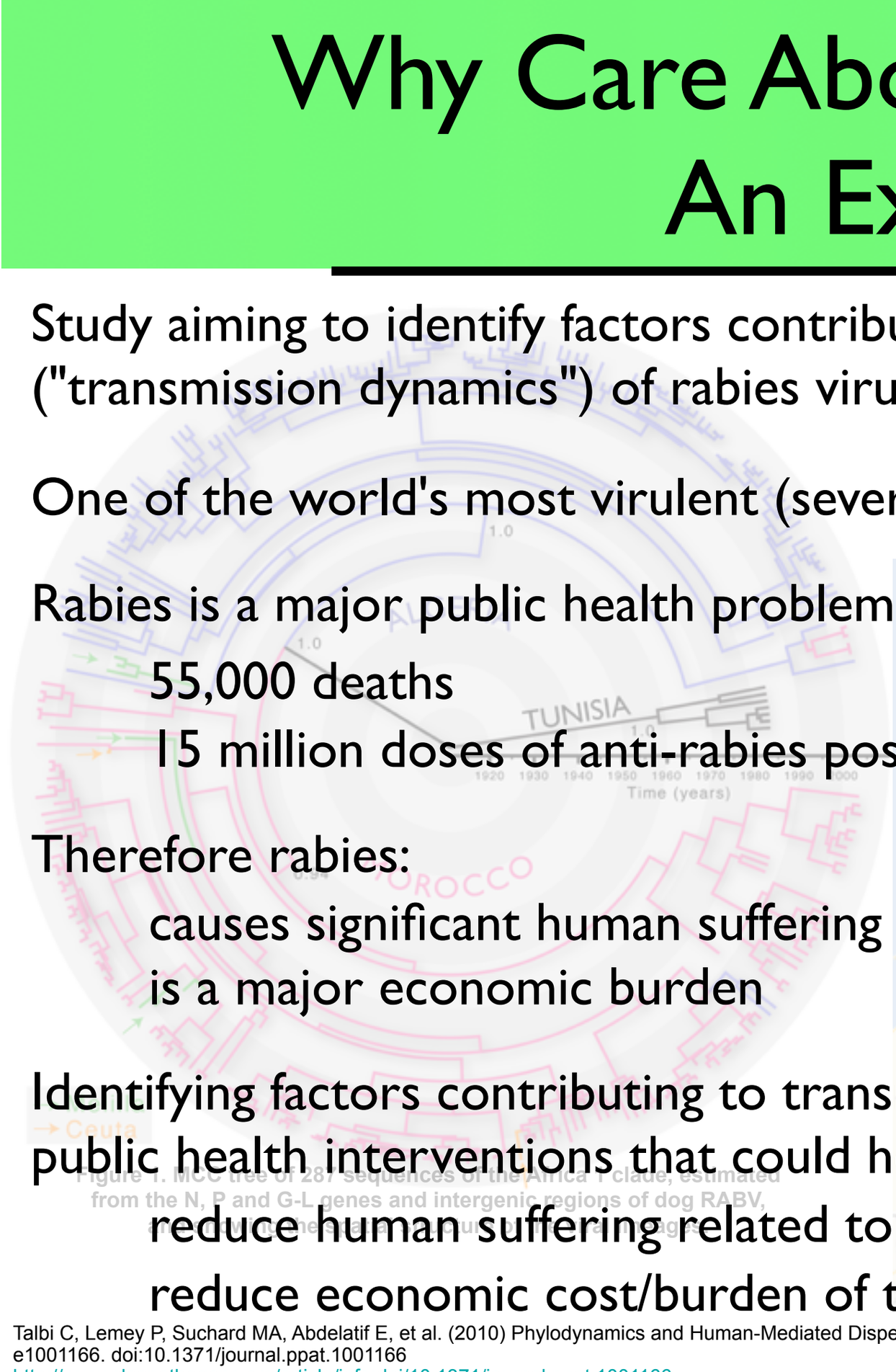
- 55,000 deaths
- 15 million doses of anti-rabies post-exposure prophylaxis administered

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



*autonomous cities of Spain

Talbi C, Lemey P, Suchard MA, Abdelatif E, et al. (2010) Phylodynamics and Human-Mediated Dispersal of a Zoonotic Virus. PLoS Pathog 6(10): e1001166. doi:10.1371/journal.ppat.1001166
http://www.plospathogens.org/article/info:doi/10.1371/journal.ppat.1001166

PLOS | PATHOGENS

Why Care About Phylogenies? An Example

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

Transmission via saliva, particularly via dog bites

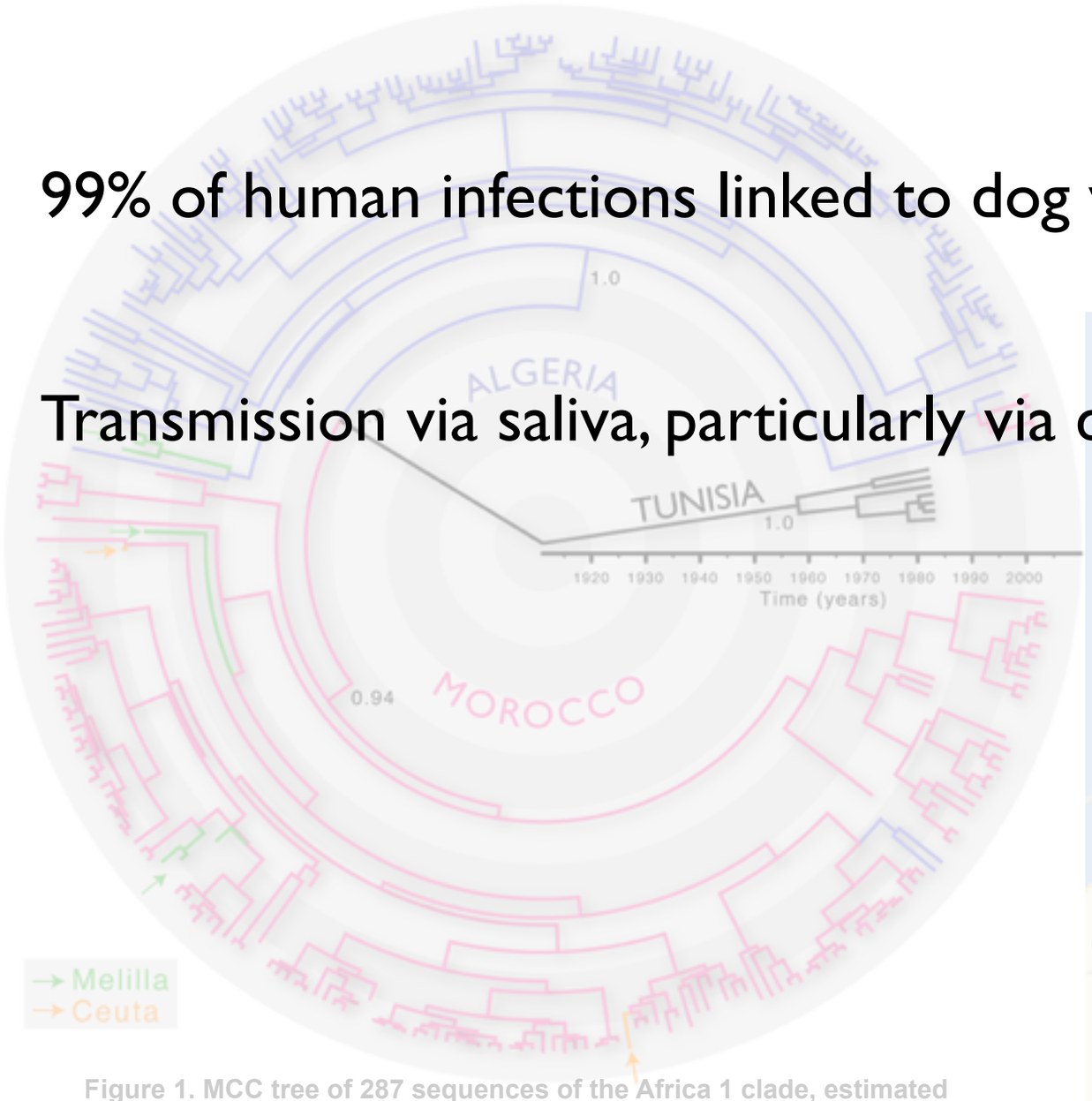


Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.

Talbi C, Lemey P, Suchard MA, Abdelatif E, et al. (2010) Phylodynamics and Human-Mediated Dispersal of a Zoonotic Virus. PLoS Pathog 6(10): e1001166. doi:10.1371/journal.ppat.1001166
http://www.plospathogens.org/article/info:doi/10.1371/journal.ppat.1001166

PLOS | PATHOGENS

Why Care About Phylogenies? An Example

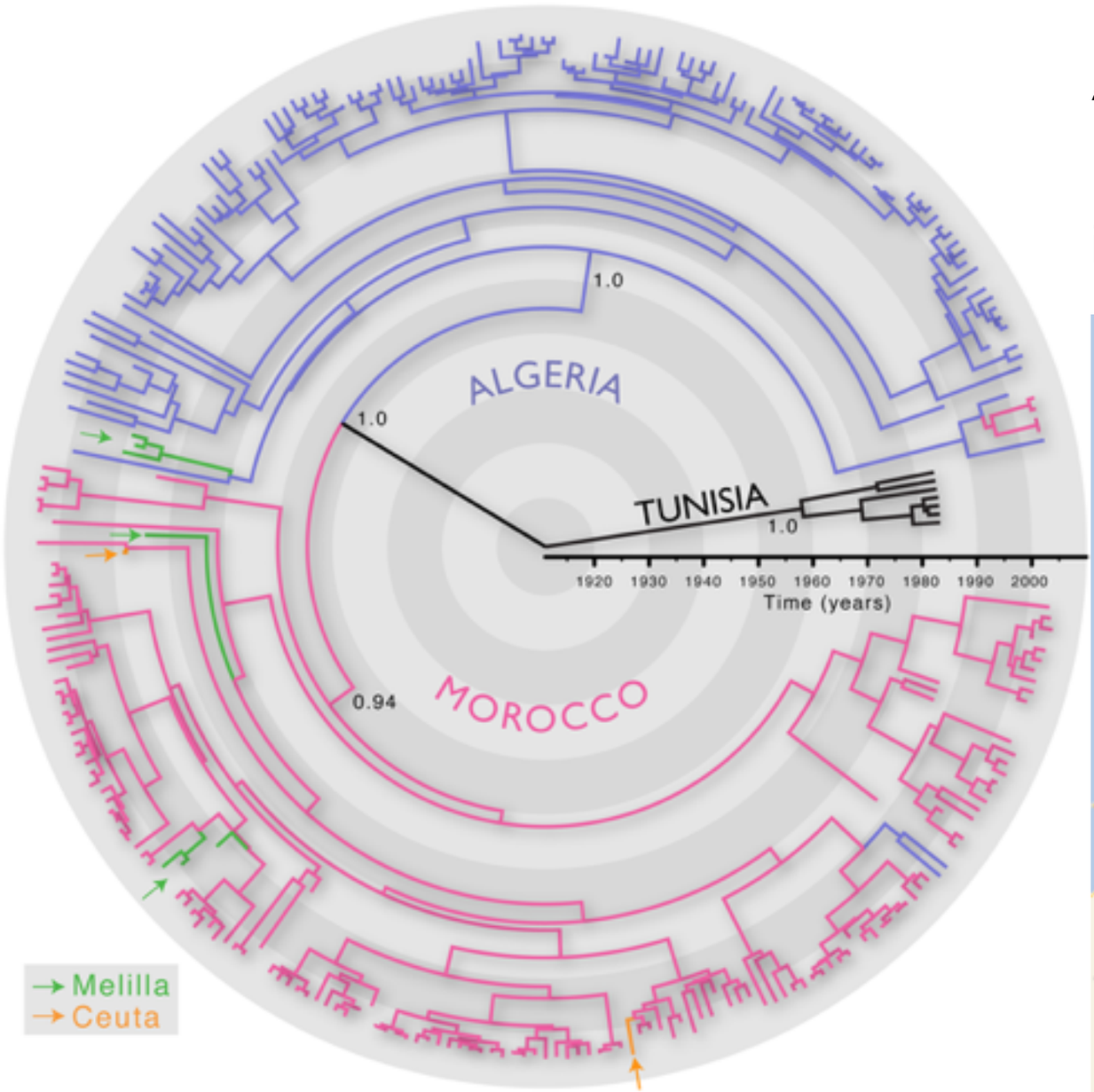
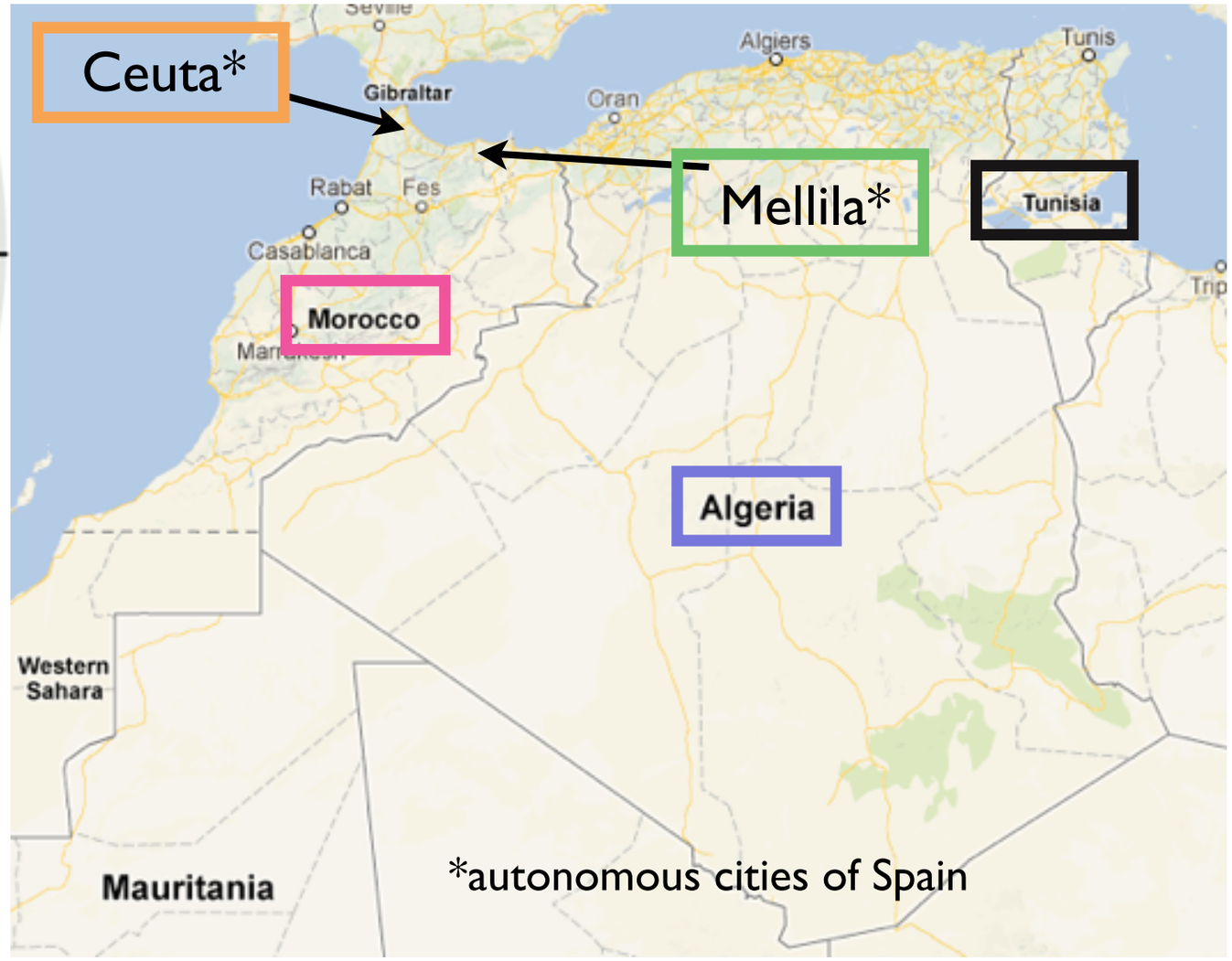


Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.

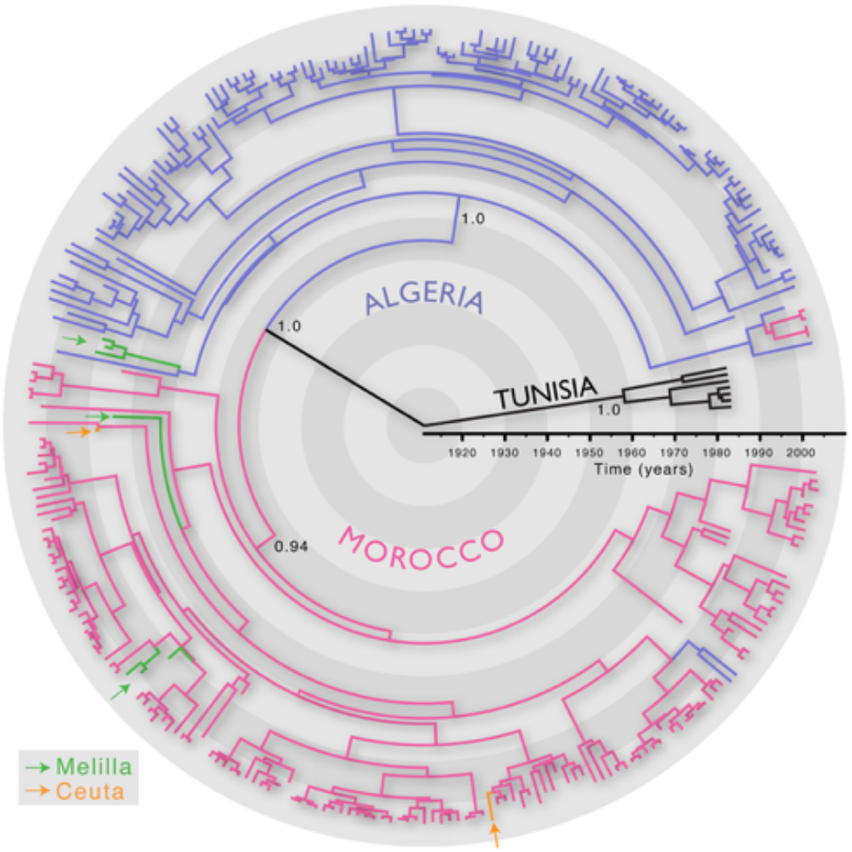
Phylogeny of rabies virus sampled from North African dogs

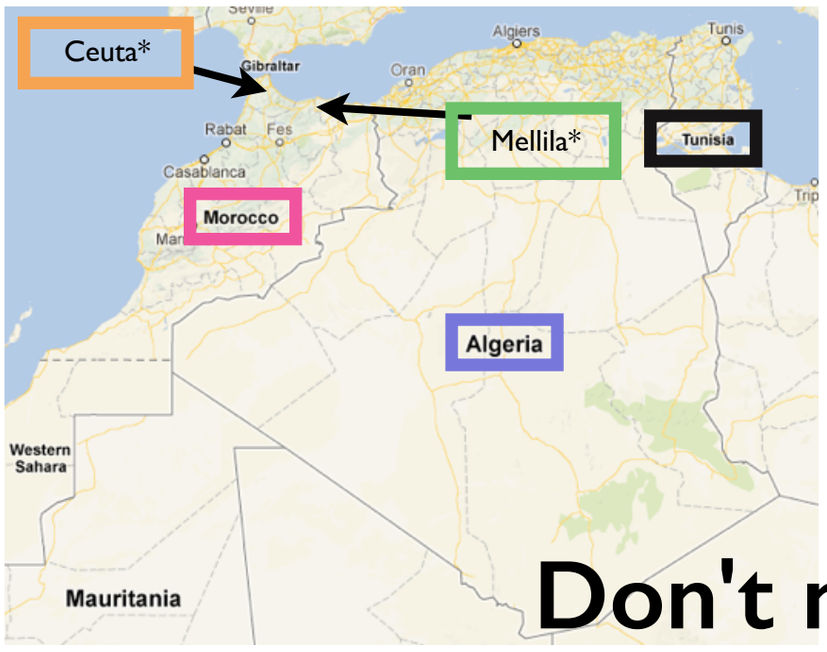
Branches coloured according to measured or inferred geographical location



Talbi C, Lemey P, Suchard MA, Abdelatif E, et al. (2010) Phylodynamics and Human-Mediated Dispersal of a Zoonotic Virus. PLoS Pathog 6(10): e1001166. doi:10.1371/journal.ppat.1001166
http://www.plospathogens.org/article/info:doi/10.1371/journal.ppat.1001166

Why Care About Phylogenies? 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

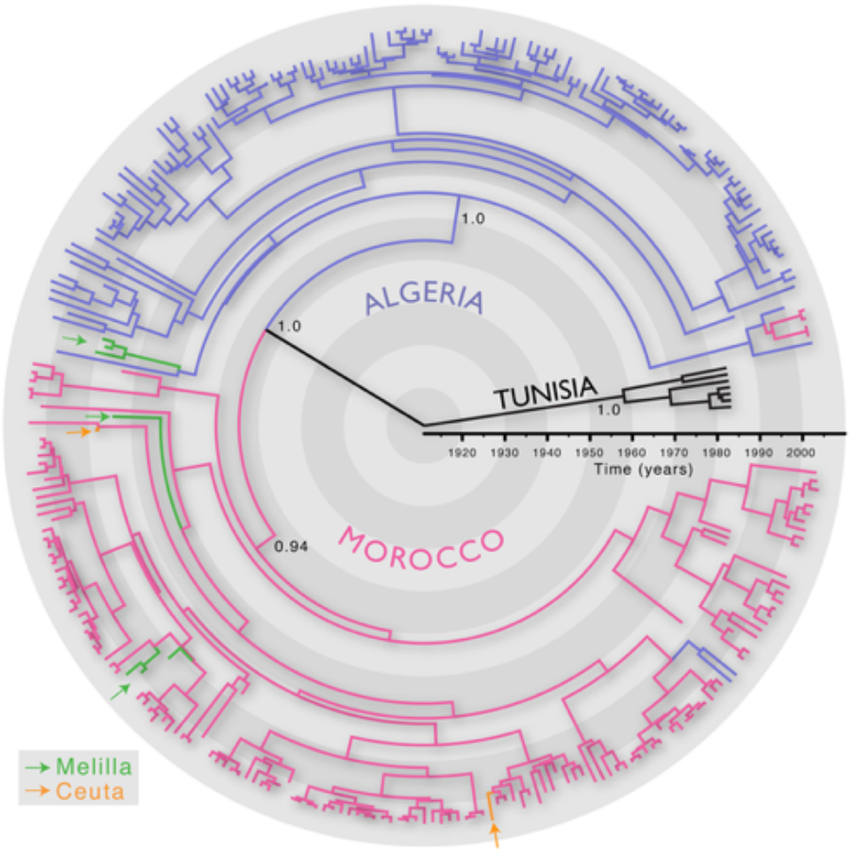
Then we'll take a vote, followed by discussing with your neigbours

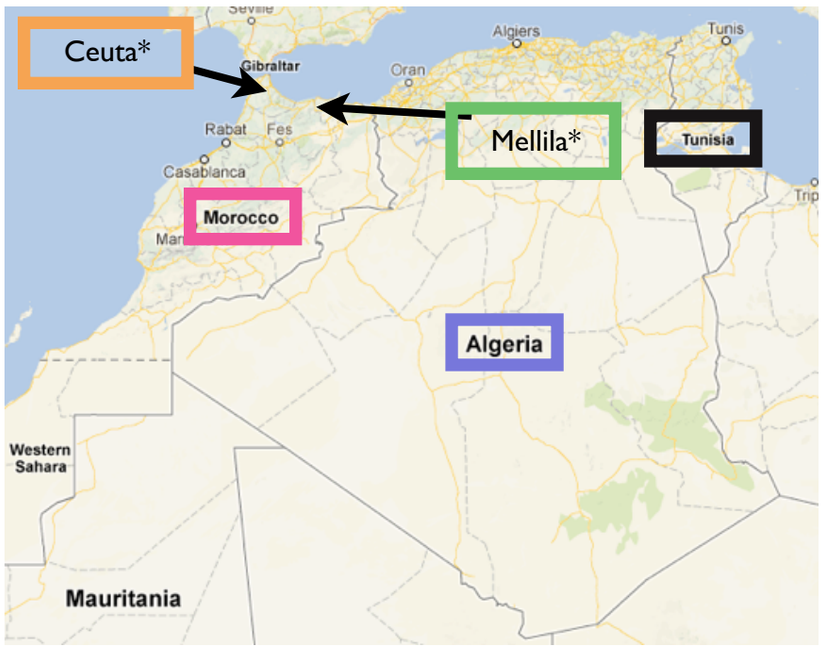
Feel uncomfortable that you don't know enough about the study/data to decide?

Make (and make 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!

Why Care About Phylogenies? 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

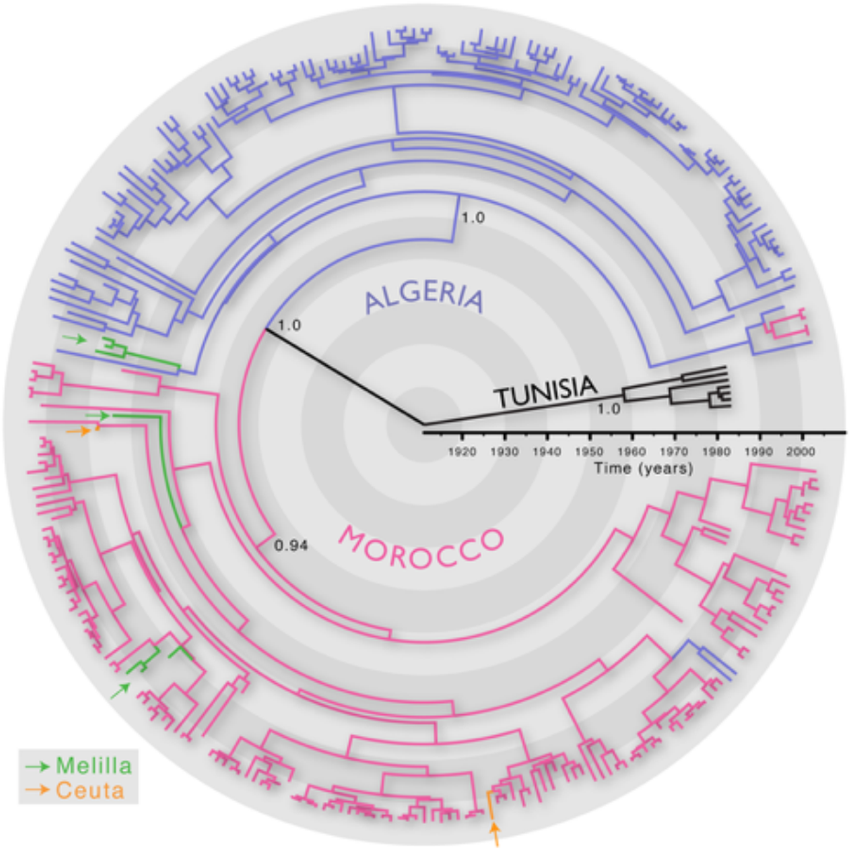
A. more probable

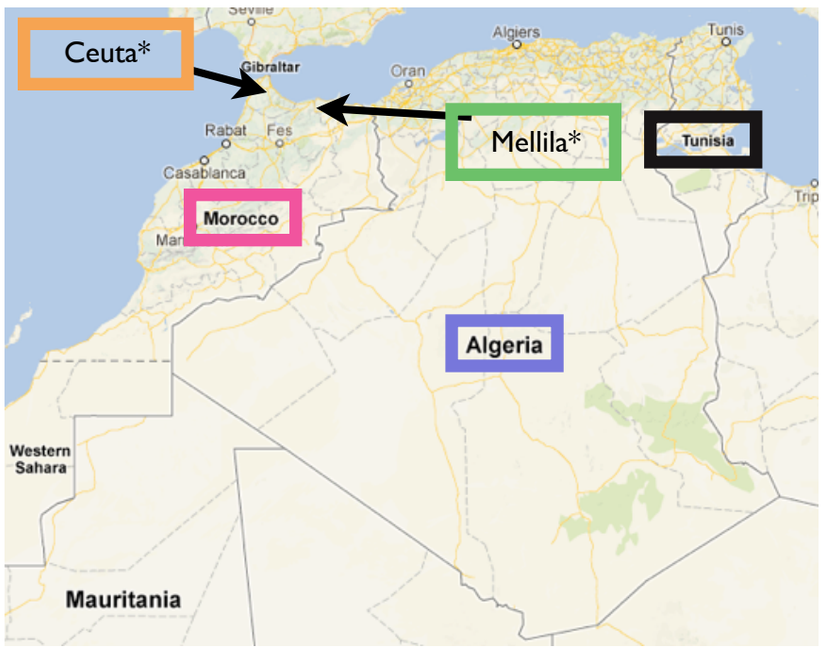
that human activity significantly influences the dynamics of rabies virus transmission between dogs

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

Why Care About Phylogenies? 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

add another slide about common comments/things we learn from the example, and make previous slide less fussy

Other Applications of Phylogenetics

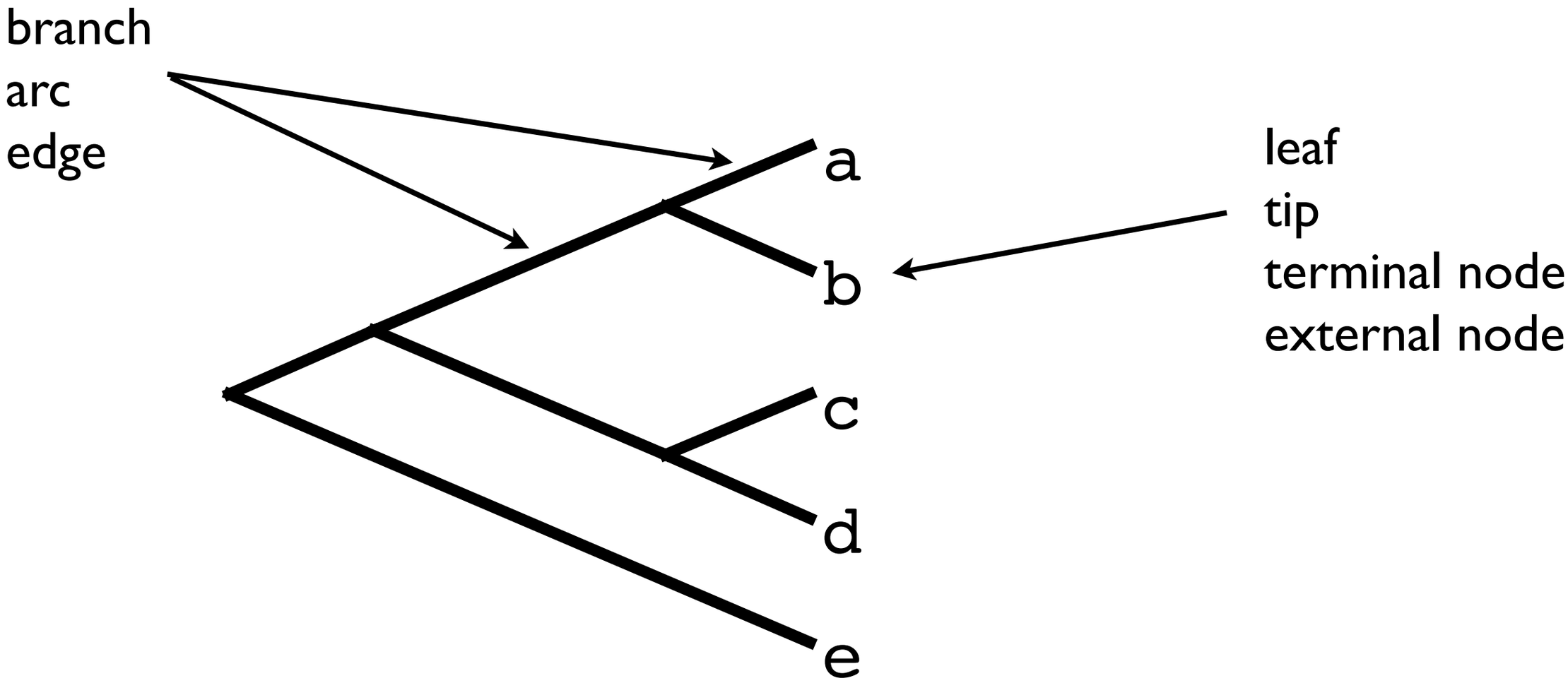
- Epidemiology
- Forensics
- Selecting conservation targets
- Monitoring trade in illegal organisms
- Bioinformatics tools - in particular:
  - building MSAs
  - predicting function
- Basic evolutionary research
  - characterising processes of evolutionary transformation
  - estimating patterns of transmission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

Rooted Phylogenies

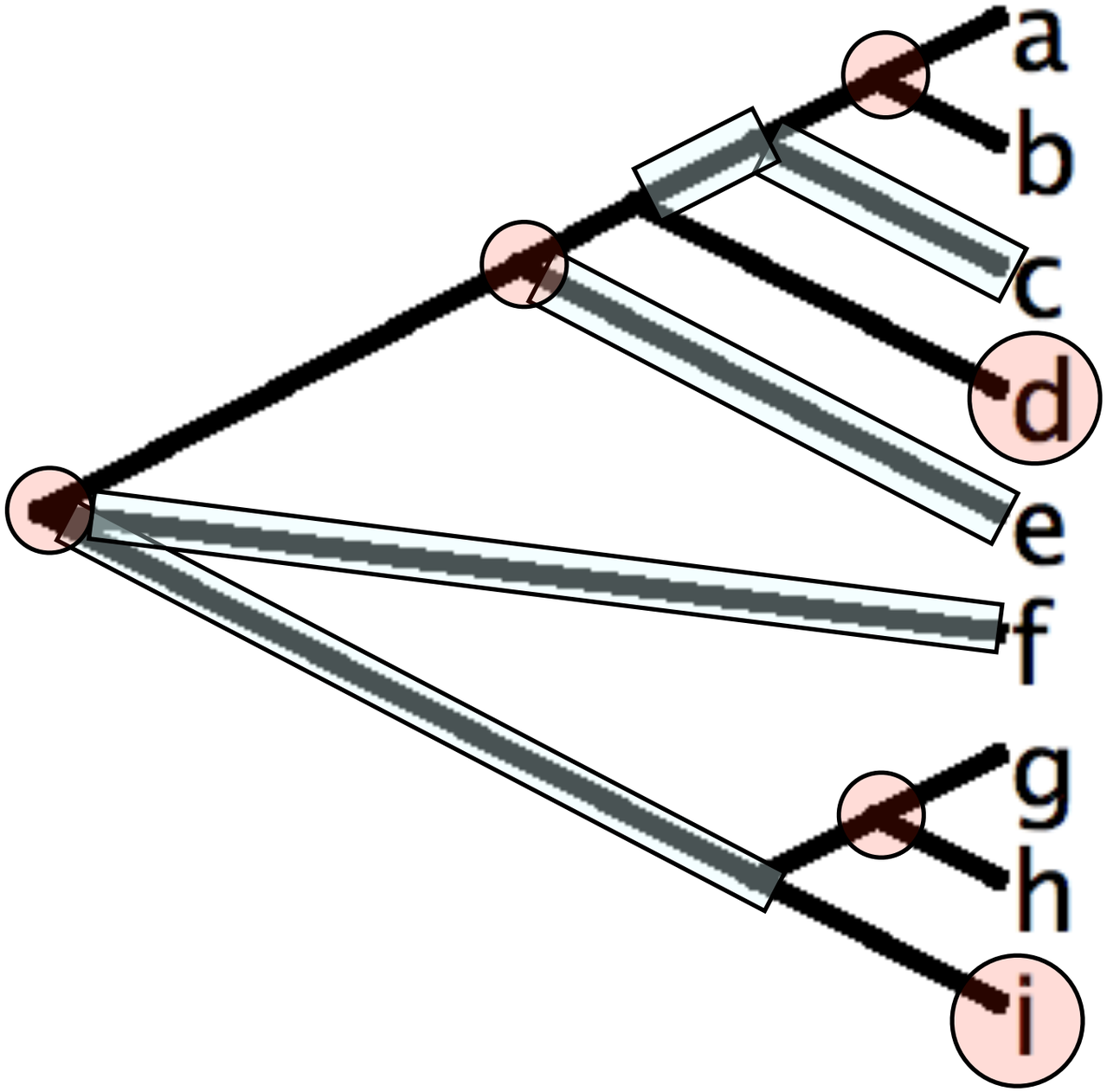
Terminology and Concepts

Aidan Budd, EMBL Heidelberg

Alternative Tree-Related Terminologies



Trees: Branches and Nodes



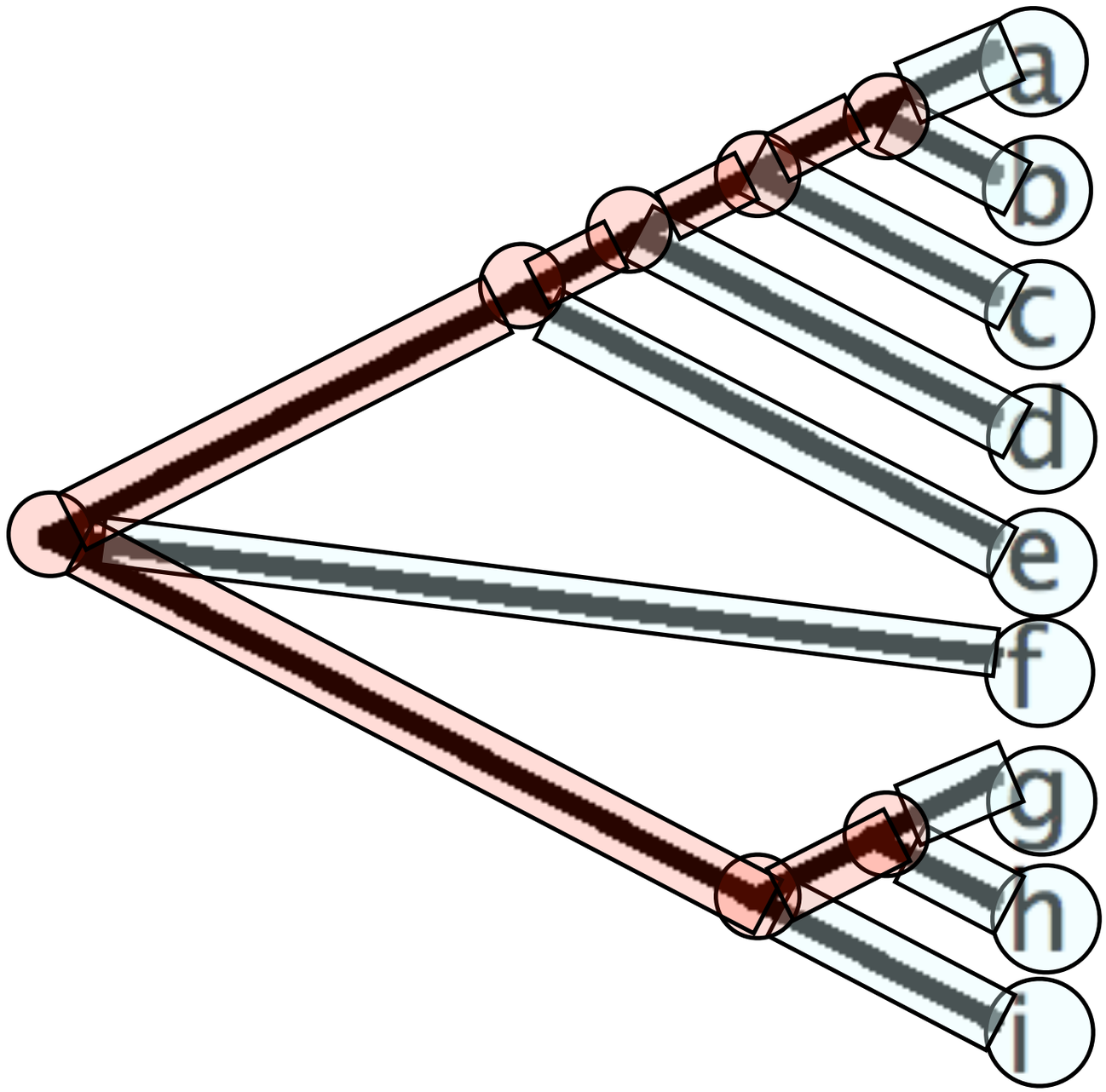
Trees consist of:

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

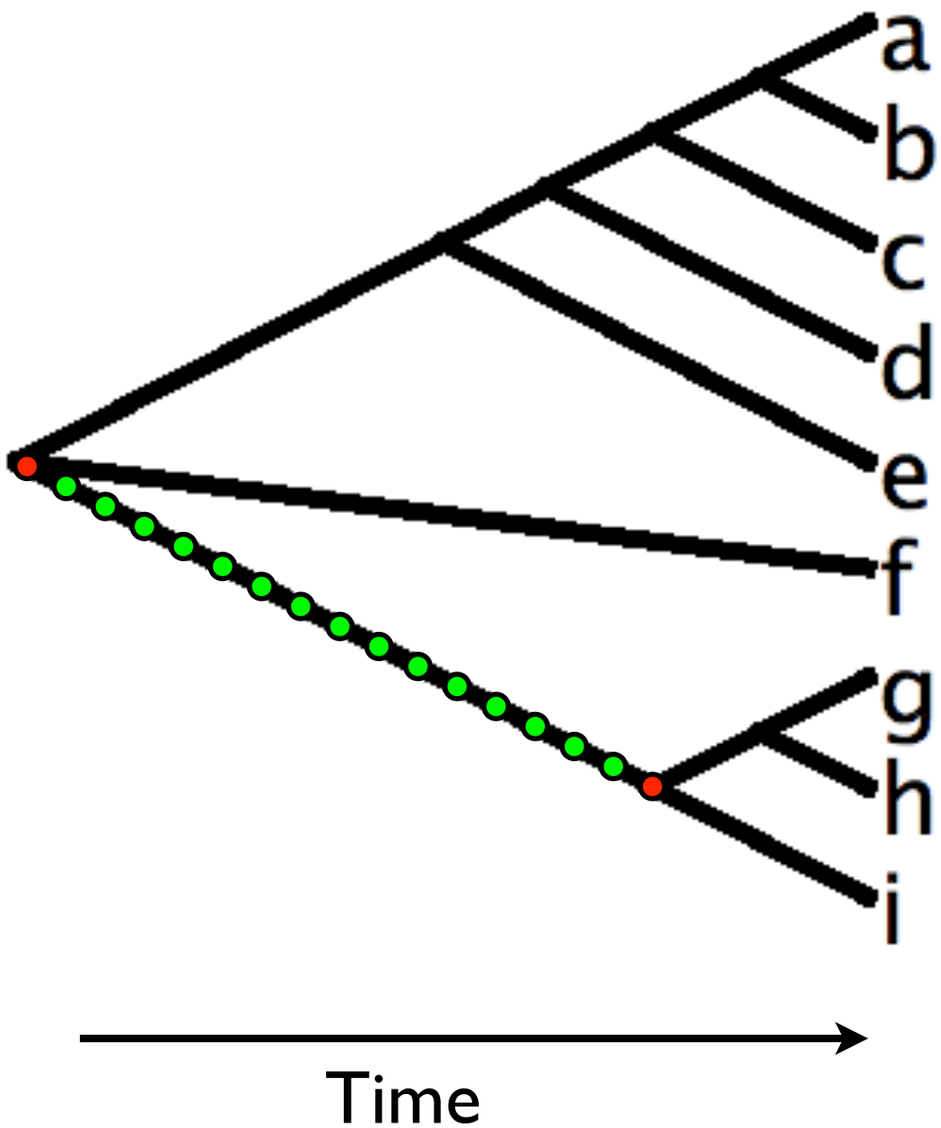
Branch - links two internal nodes

external/terminal

Node - associated with an extant sequence/**OTU** (operational taxonomic unit)

Branch - links an external and an internal node

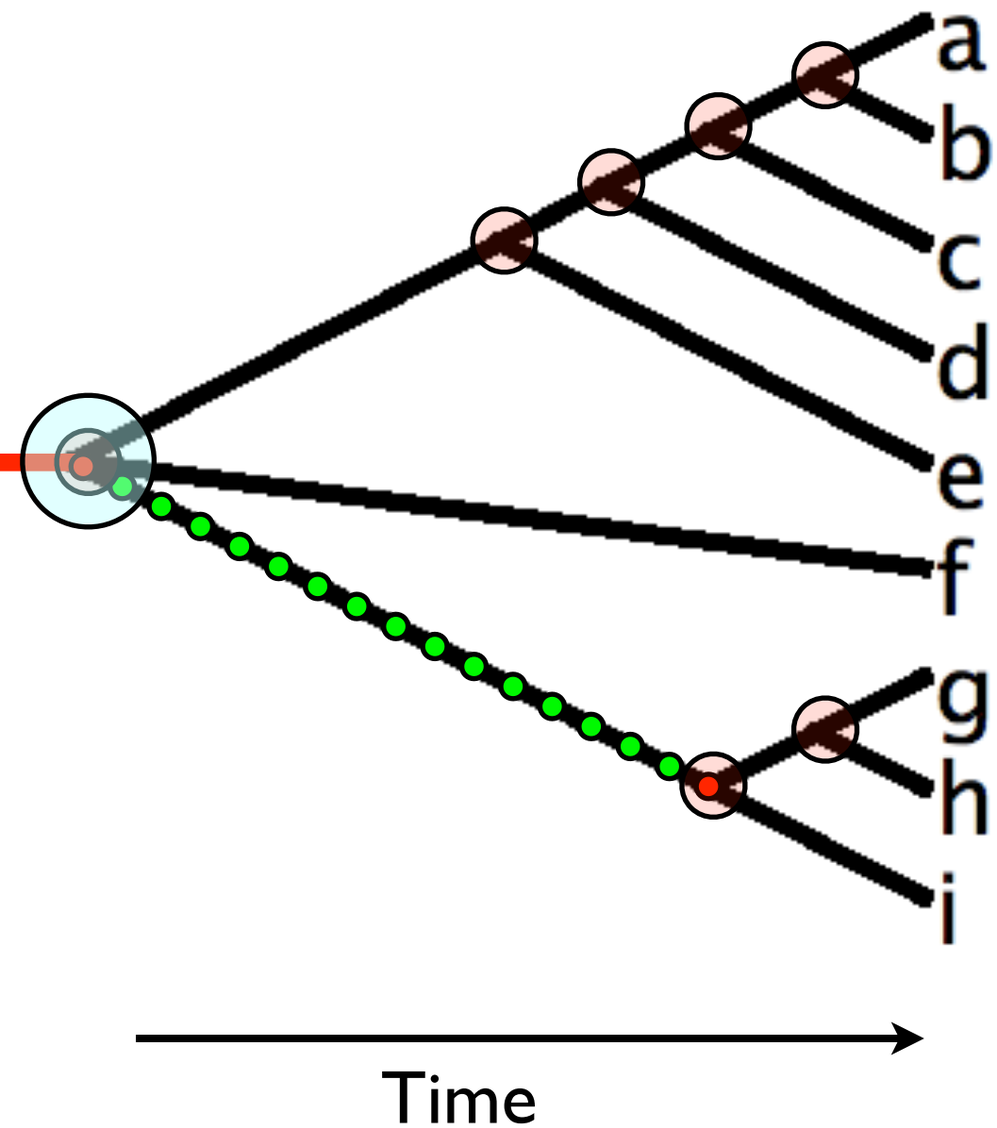
Branches



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

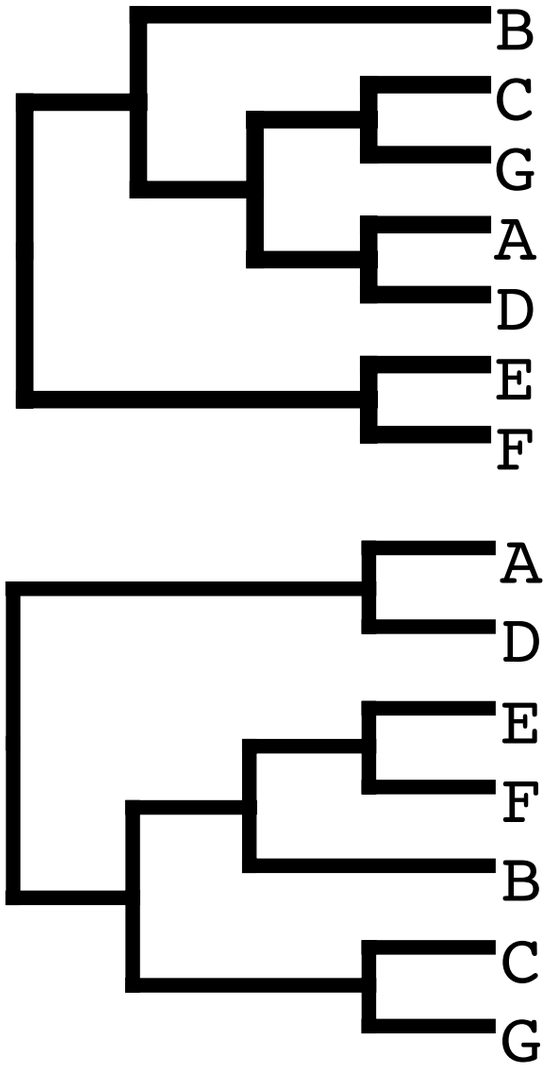
Branch Lengths

Aidan Budd, EMBL Heidelberg

Scaled and Unscaled Trees

Aidan Budd, EMBL Heidelberg

Unscaled Trees



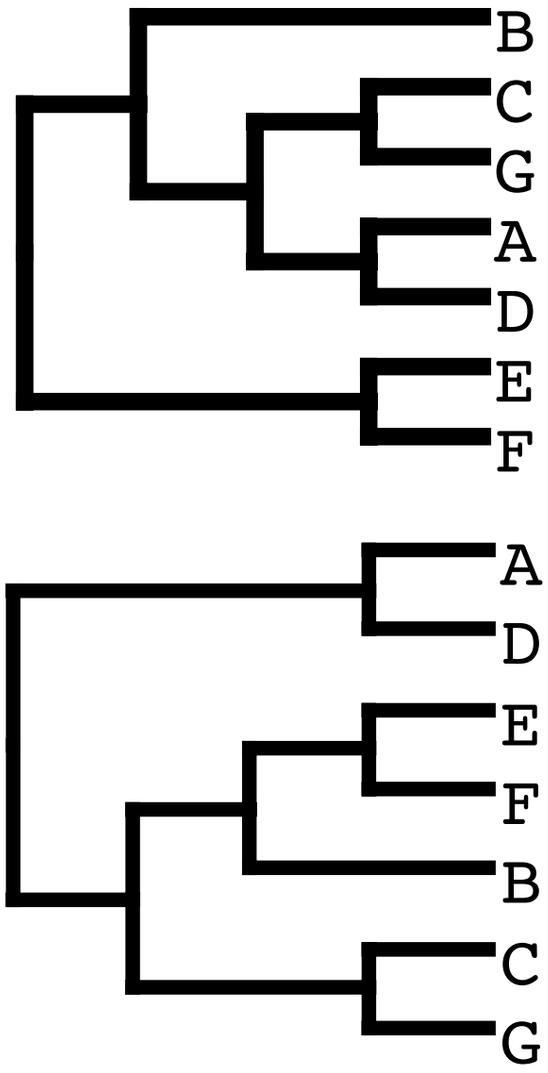
Same **unscaled** unrooted tree

Branch lengths provide no information

Branch lengths usually chosen to align OTU labels

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

Scaled Trees



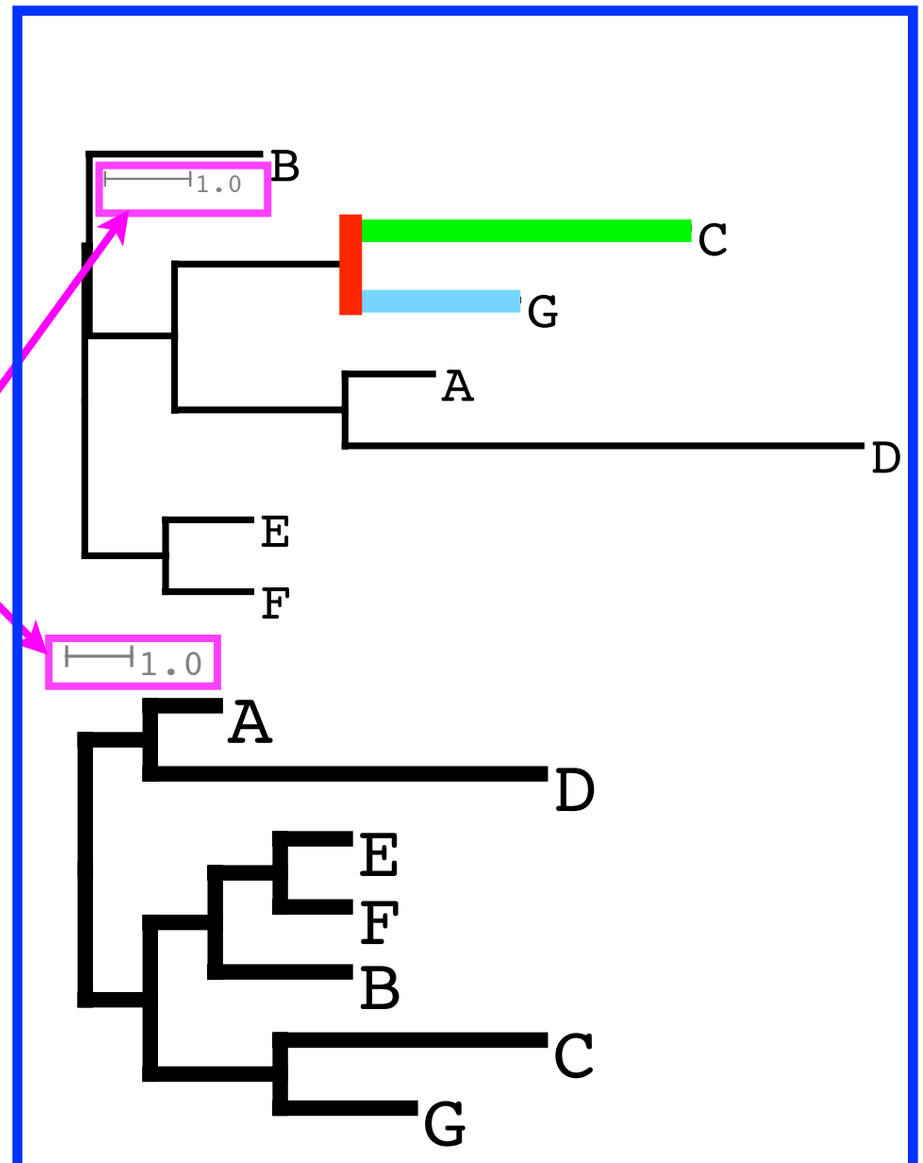
Same **unscaled** unrooted tree

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!)



Same **scaled** unrooted tree

Branch Lengths

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

```
SeqA  IKTILKWWSP
SeqB  IKTIVKWDSP
```

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**

Branch Lengths

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

SeqA IKTILKWWSP
SeqB IKTIVKWDSP

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**

Branch Lengths and Visualising Trees

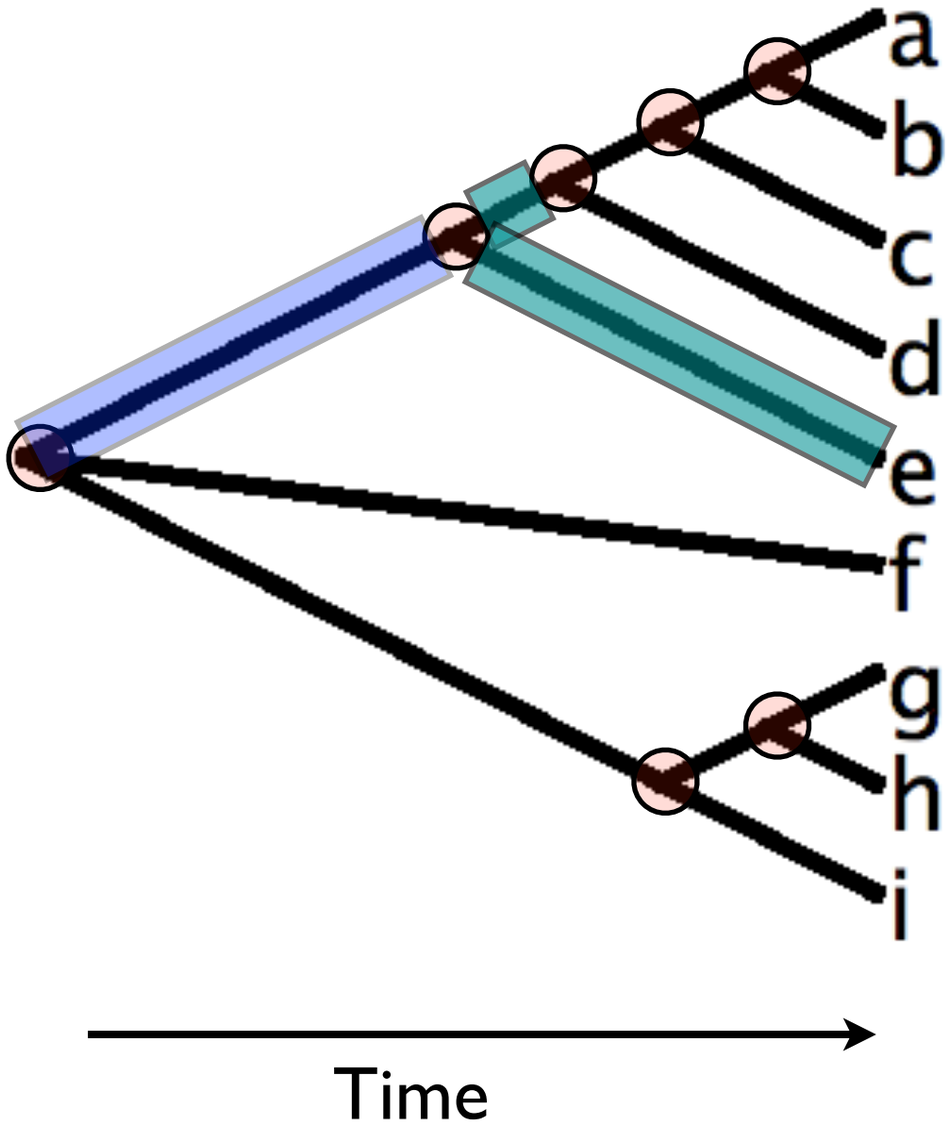
Demonstration and Exercise

- Working with scaled trees in Dendroscope
  - Formatting trees for figures
  - Working with large trees

More Rooted Tree Terminology

Aidan Budd, EMBL Heidelberg

Parent/Daughter Branches

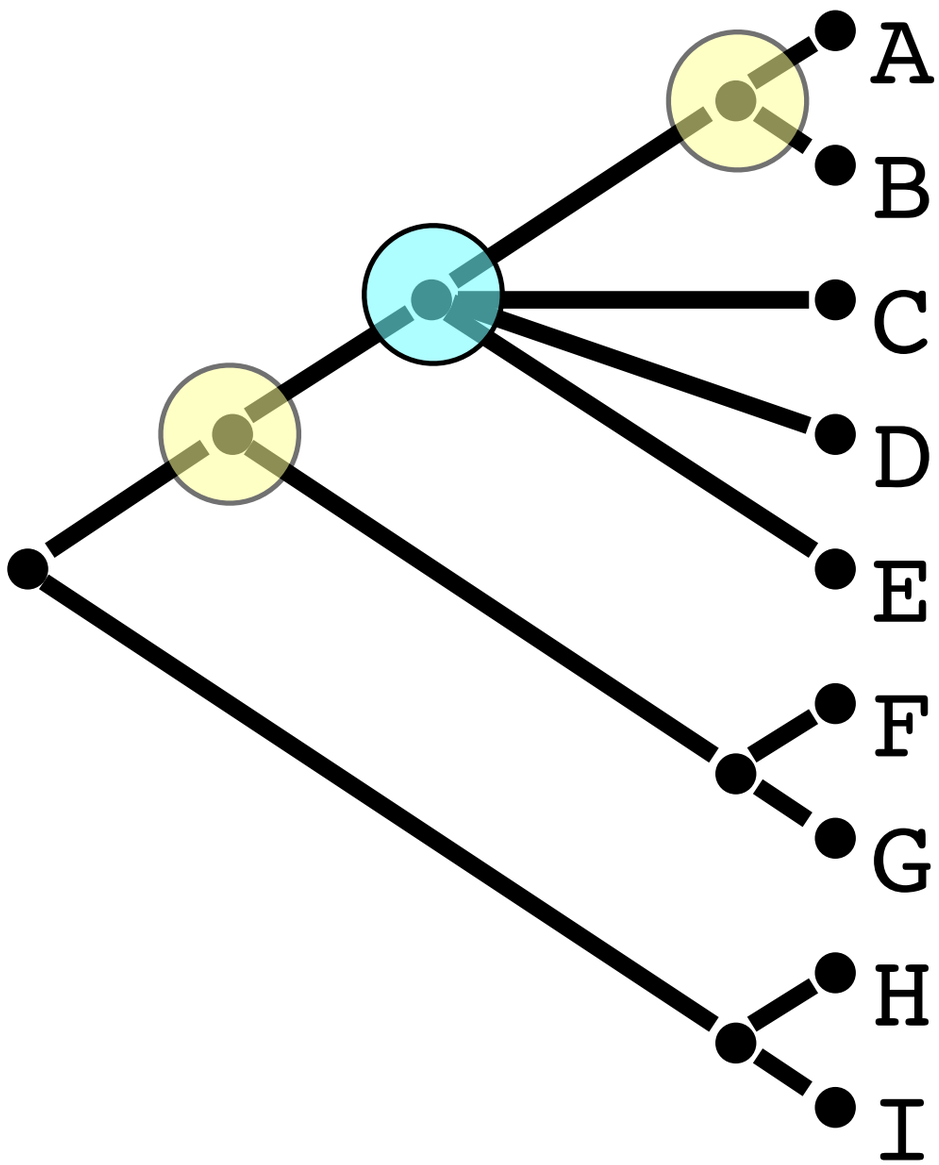


parental/ancestral lineages/branches

diverge into

multiple **daughter** lineages/branches

Polytomies



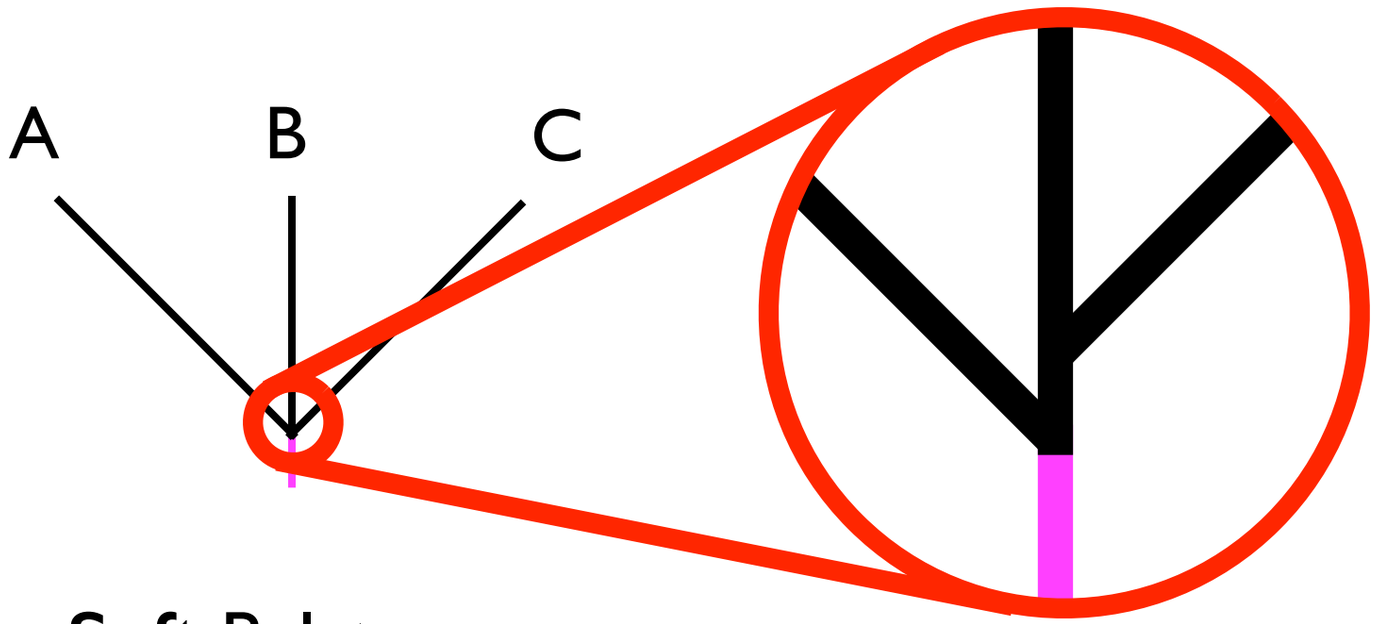
Polytomies

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**

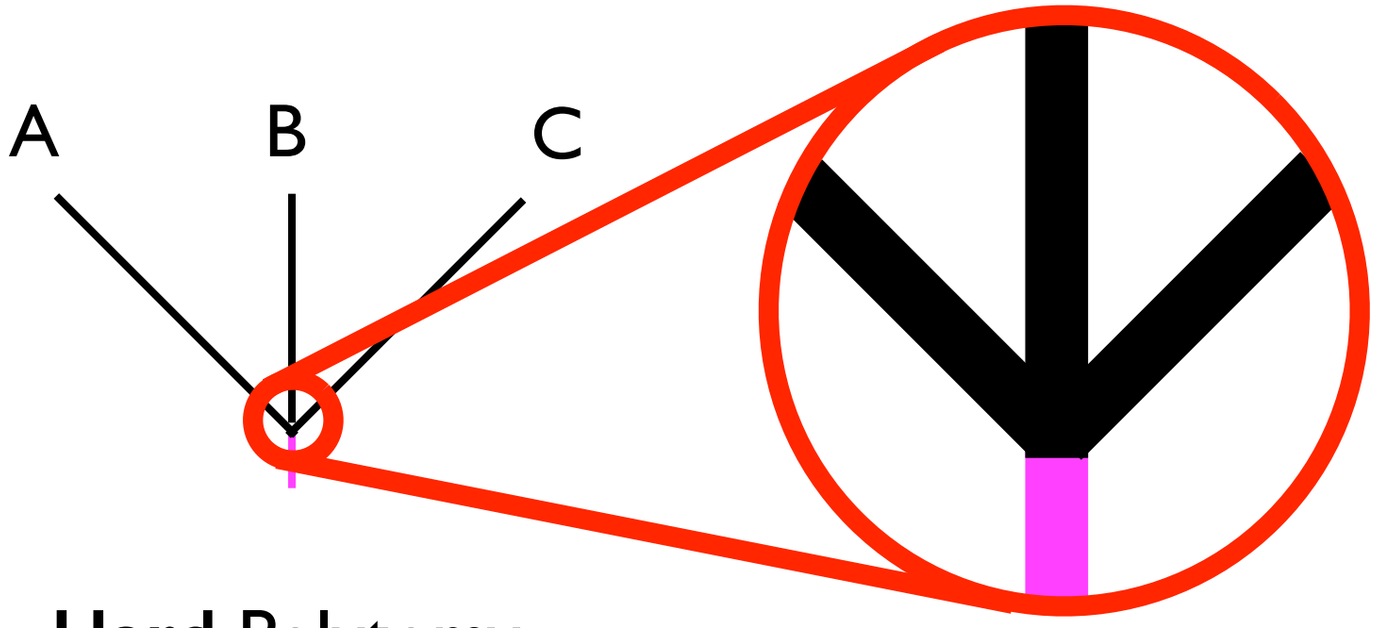
Interpreting Polytomies



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

Aidan Budd, EMBL Heidelberg

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

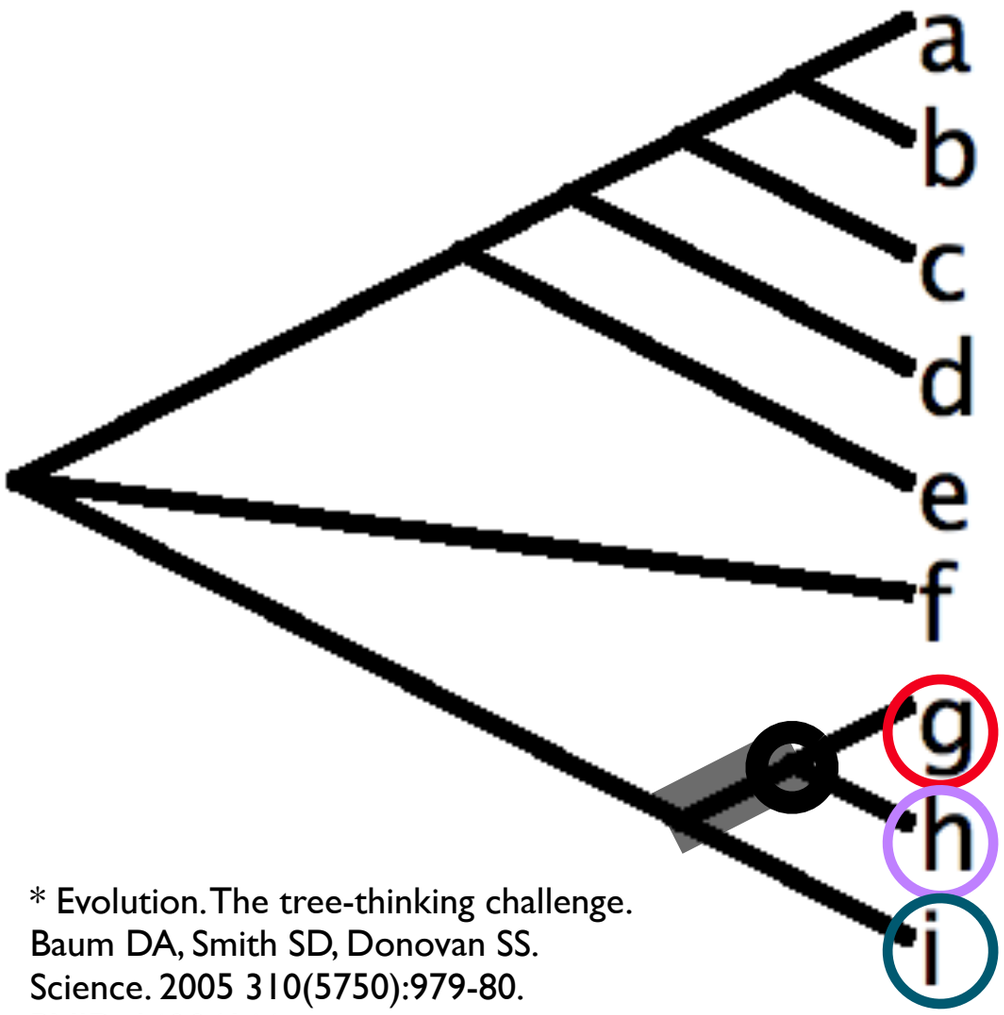
As many trees are estimated to inform ideas about patterns of relatedness, we need a clear understanding of how the term is used in this context

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

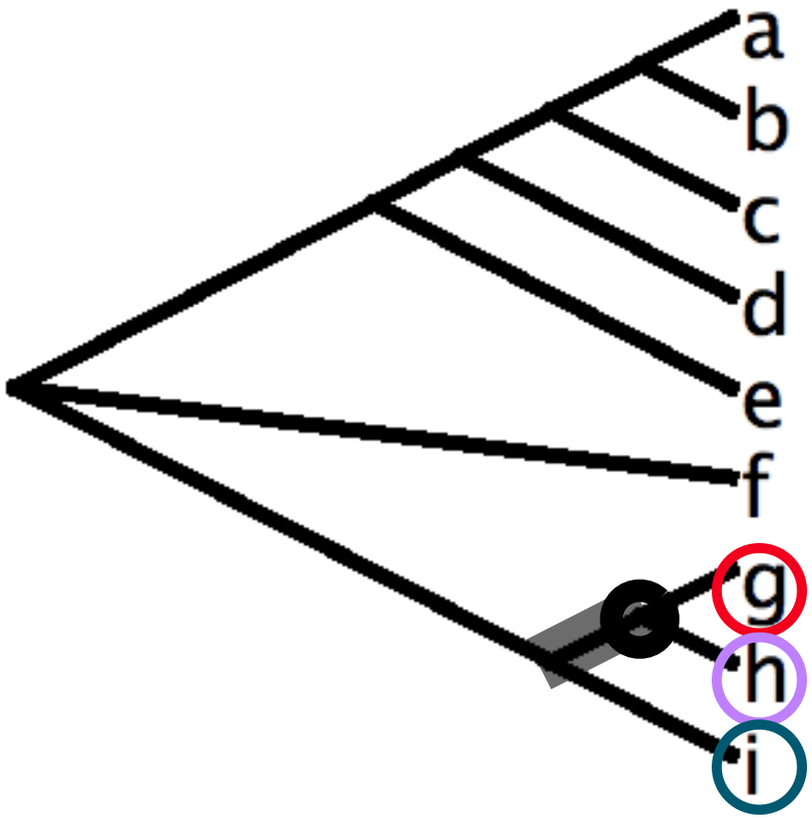
is associated with the extent of ancestry (i.e. the number of ancestors) taxa share with each other

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

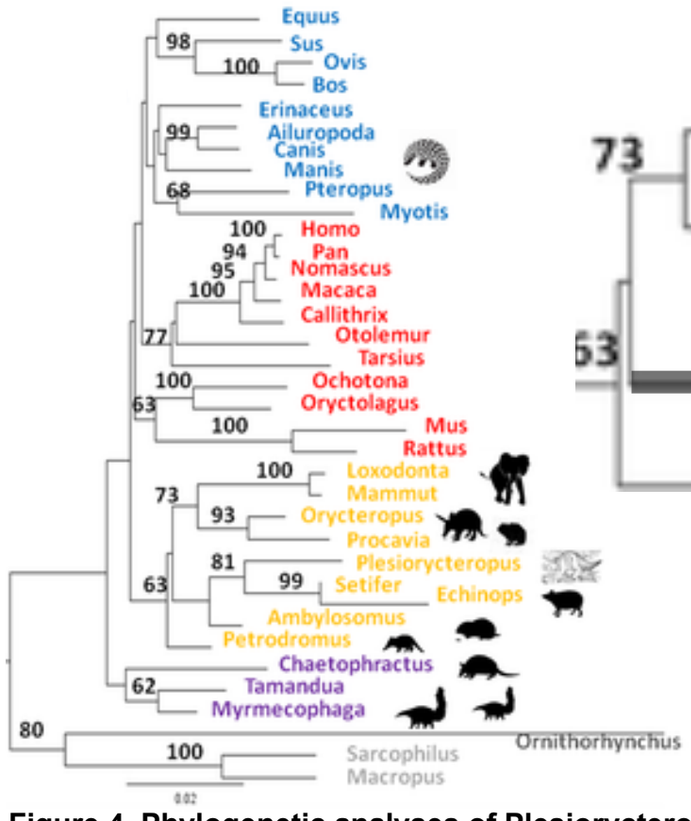
Relatedness
(in the context of phylogenetic trees)

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

g is **more closely related** to **h** than **g** is to **i**



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



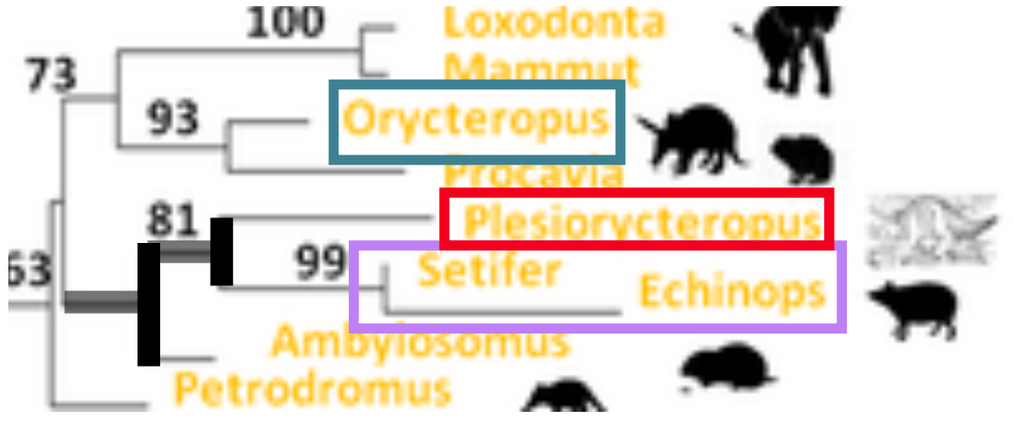


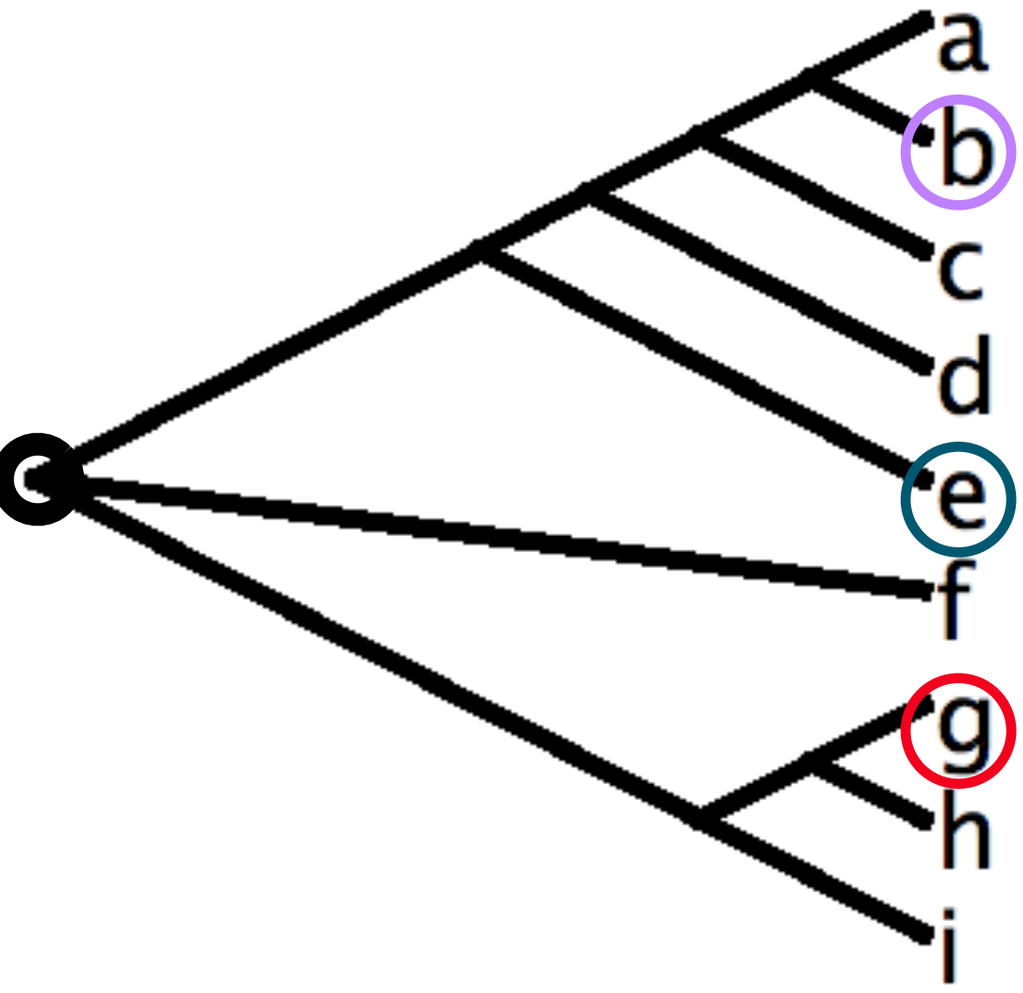
Figure 4. Phylogenetic analyses of Plesiorycteropus collagen (I) sequences obtained by LC-MS in comparison to previously postulated closest relatives.

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

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

Aidan Budd, EMBL Heidelberg

Relatedness
(in the context of phylogenetic trees)



of all the OTUs represented in this tree

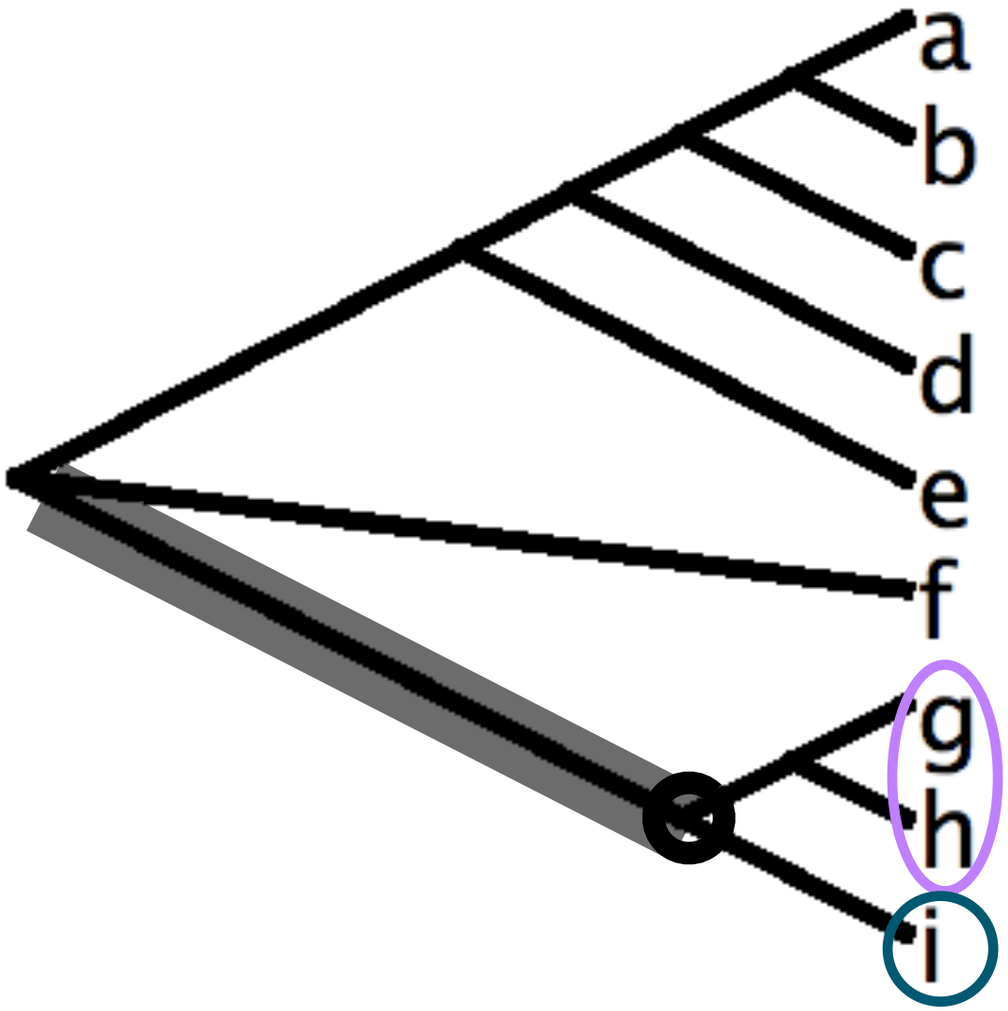
g is **equally distantly related** to **b** as it is to **e**

because **g** shares **the same** most recent common ancestor with **b** as it does with **e**

"the two species are equally distantly related to *S. cerevisiae*."

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)



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.

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**)

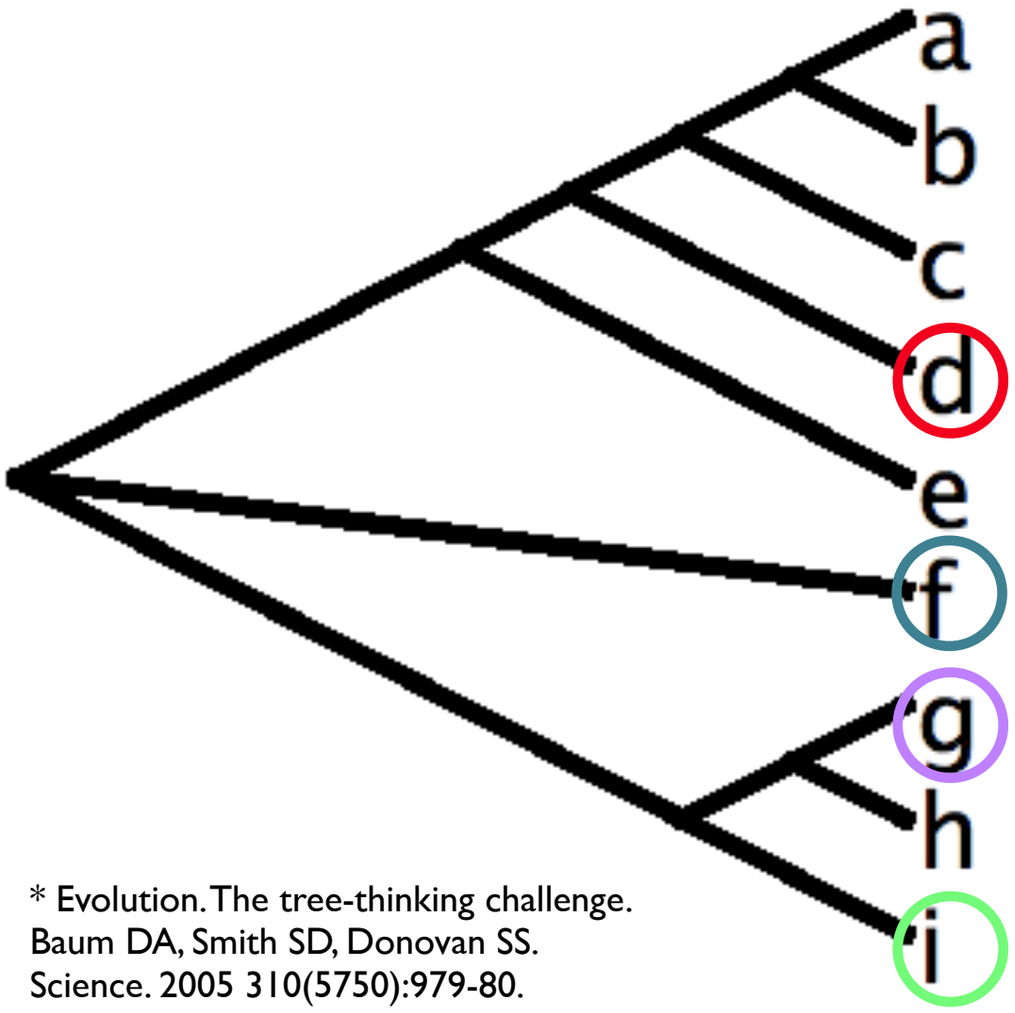
because **i** shares common ancestors with **g** and **h** that it does not share with any other OTUs in the tree

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

Aidan Budd, EMBL Heidelberg

Relatedness
(in the context of phylogenetic trees)

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



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

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

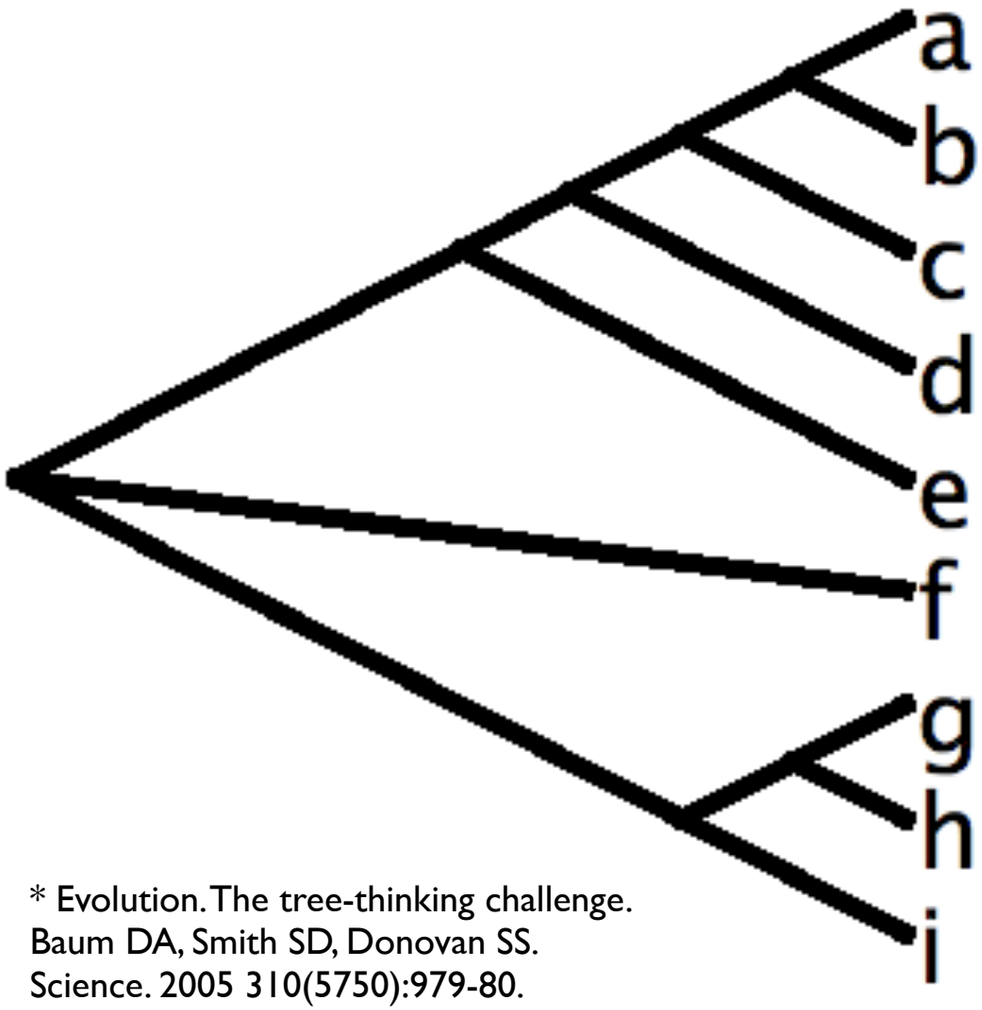
1. d is **more closely related** to g than to f or i

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

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

Relatedness
(in the context of phylogenetic trees)

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

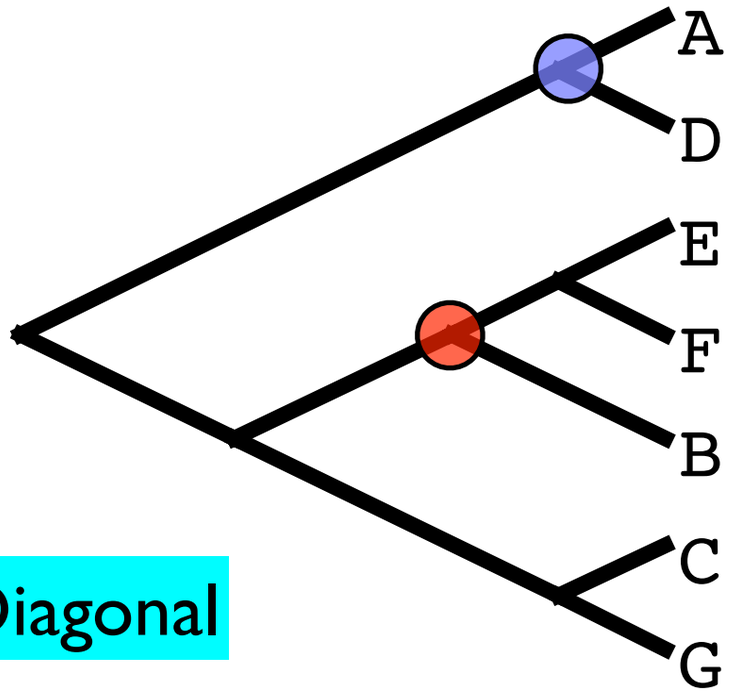


* Evolution. The tree-thinking challenge.
Baum DA, Smith SD, Donovan SS.
Science. 2005 310(5750):979-80.
PMID: 16284166

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

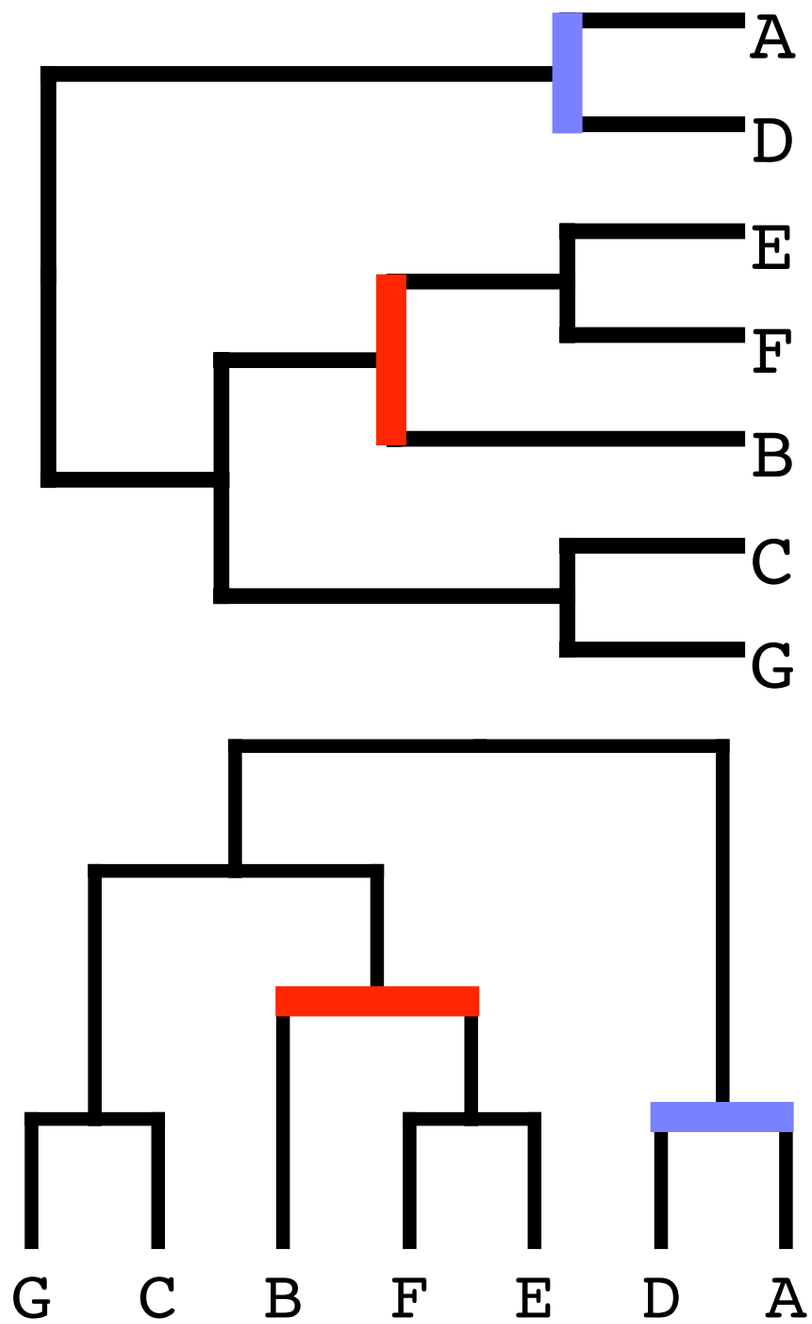
Tree Representations



Diagonal

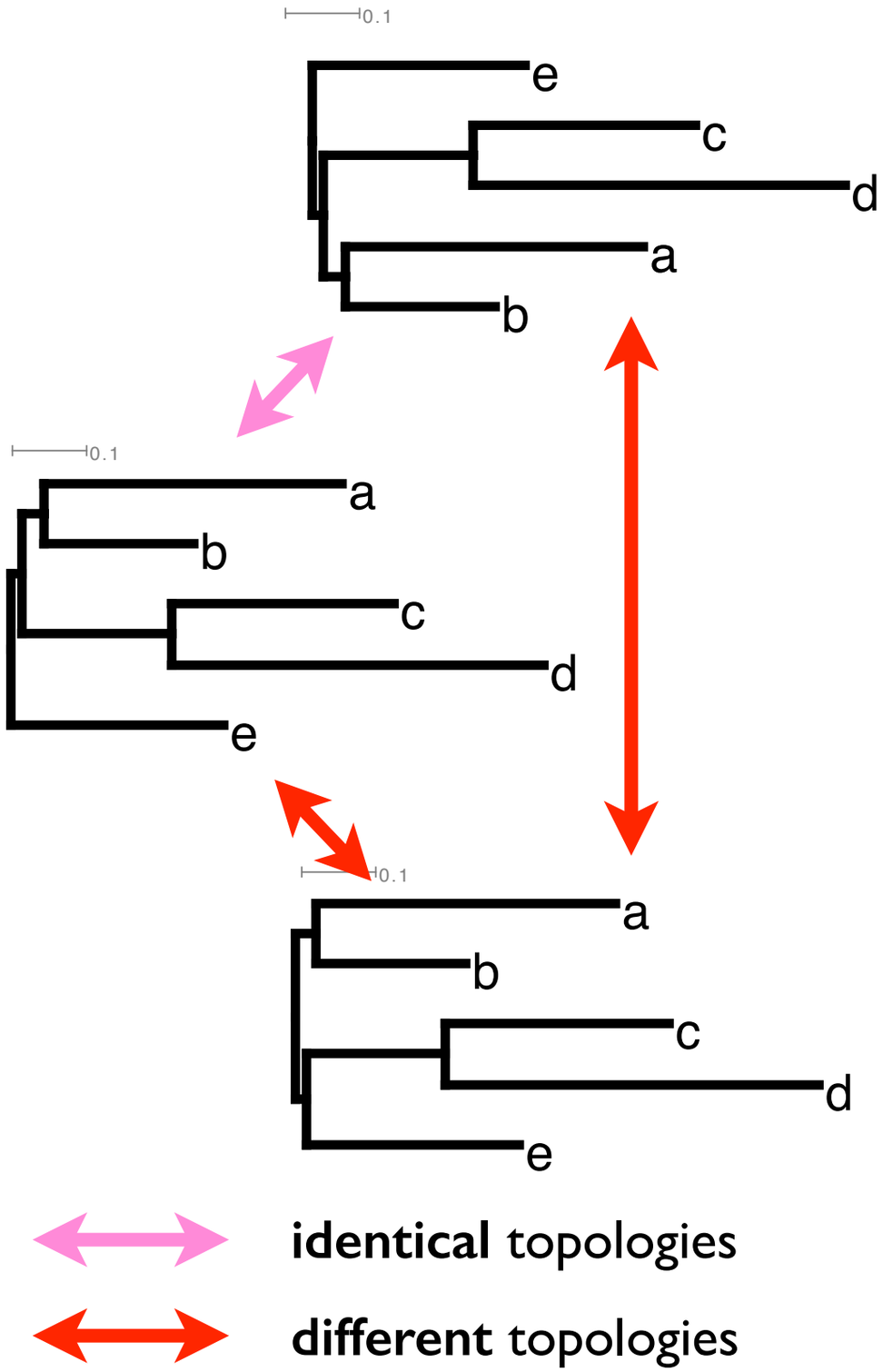
Most rooted tree figures use a "rectangular" rather than a "diagonal" representation

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



Rectangular

Tree Topology



identical topologies

different topologies

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

Trees with different topologies...

... describe **different sets** of "relatedness statements" between taxa

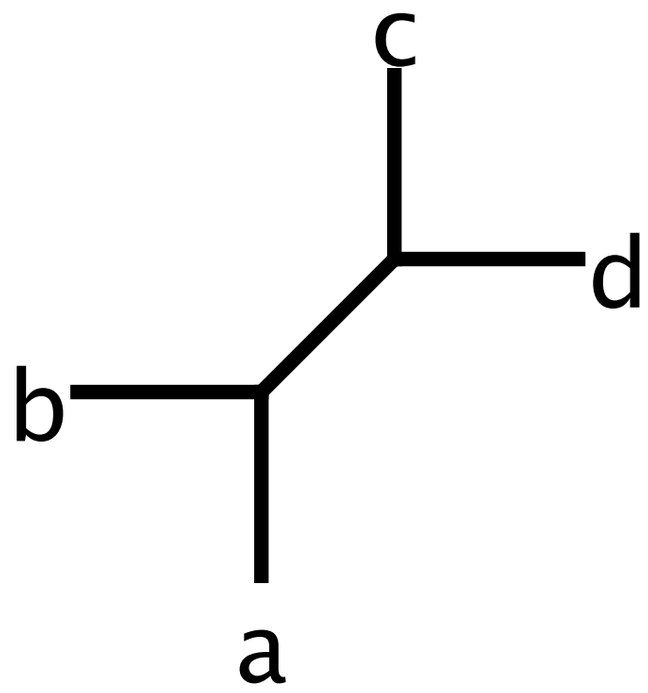
Unrooted Phylogenies

Aidan Budd, EMBL Heidelberg

Unrooted Trees

There's no root on the tree...

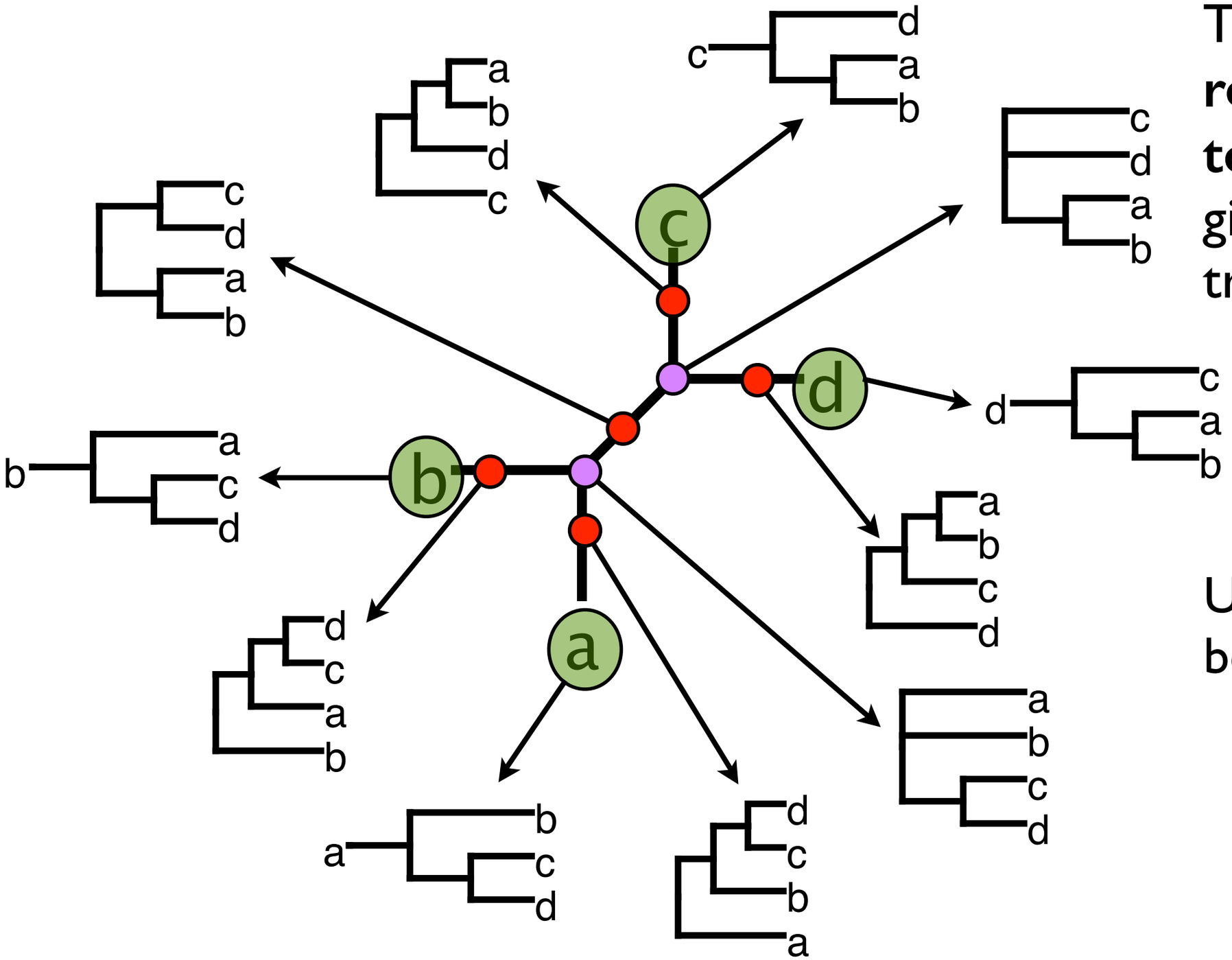
... which is usually interpreted as meaning that these taxa are related by a rooted tree **but** we don't know where the root is



Many applications of phylogenies require a rooted tree

But many tree estimation tools yield only unrooted trees!

Unrooted ⟶ Rooted



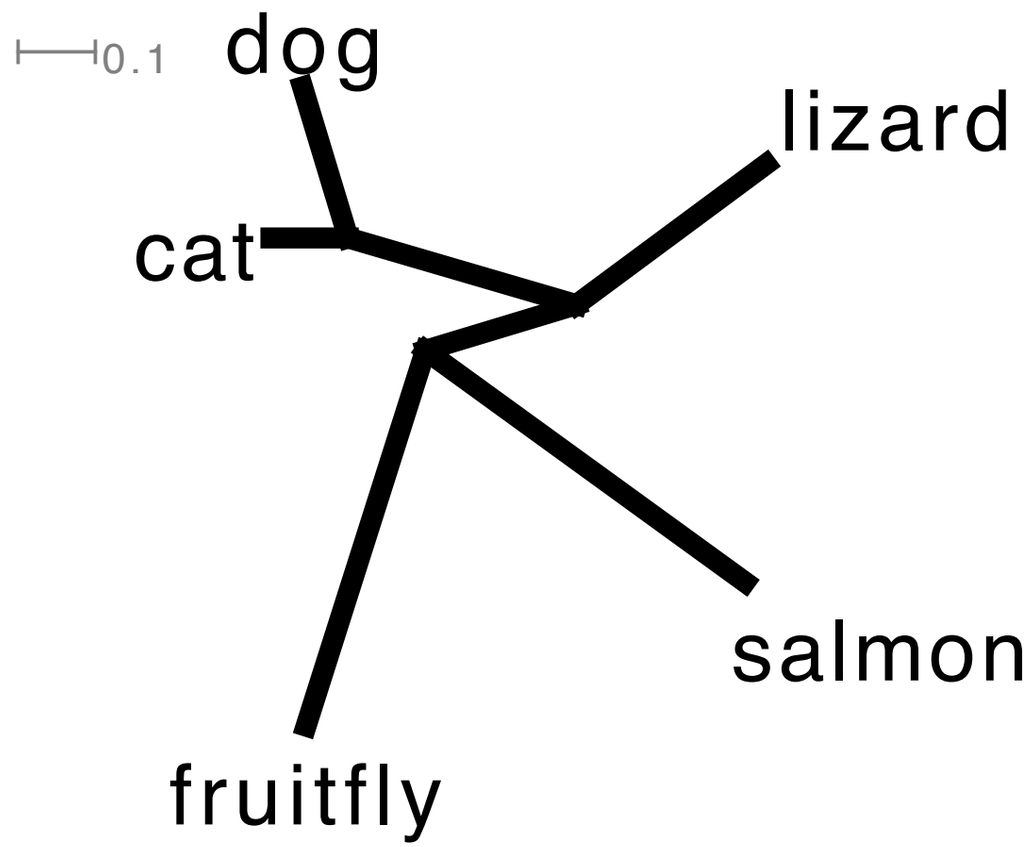
There are multiple **rooted tree topologies** for a given unrooted tree topology

Unrooted trees can be rooted on their:

- branches
- interior nodes
- terminal nodes

Unrooted ⟶ Rooted

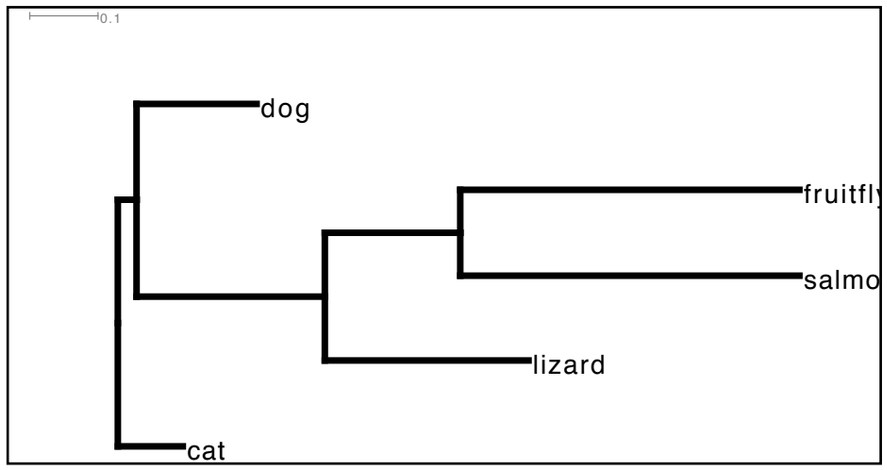
This **unrooted** tree can be rooted to yield several different **rooted topologies**.

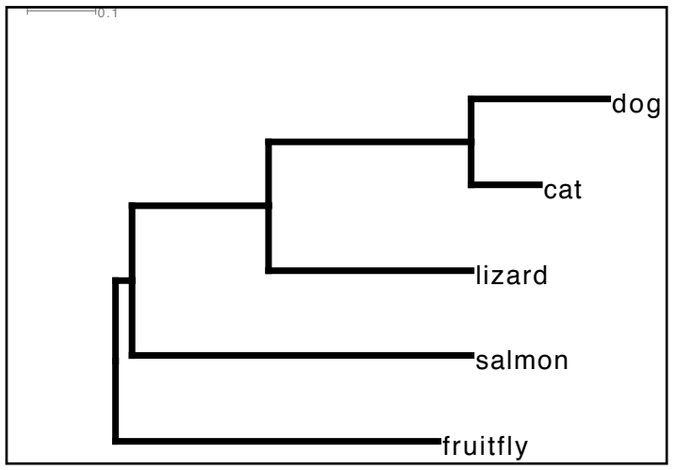


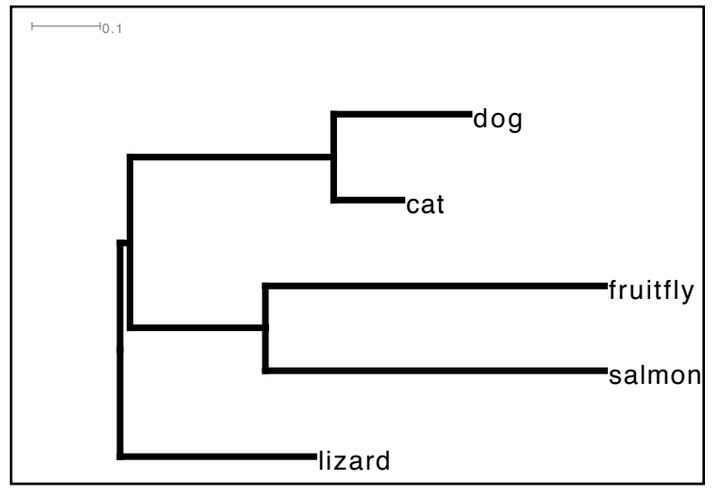
How many different rooted topologies exist by placing the root on a/an:

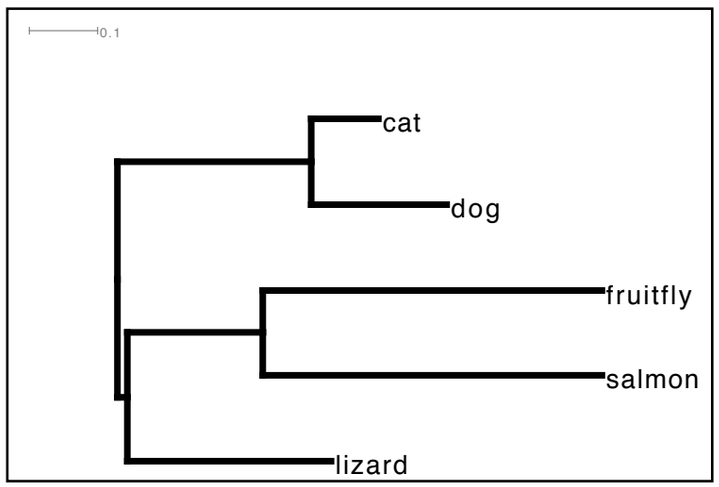
- branch? Draw all of these.
- terminal node? Draw one of these.
- internal node? Draw one of these.

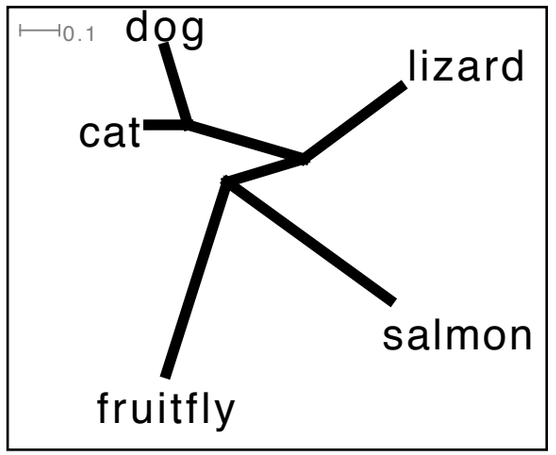
All Topologies Rooted on a Branch of the Unrooted Tree

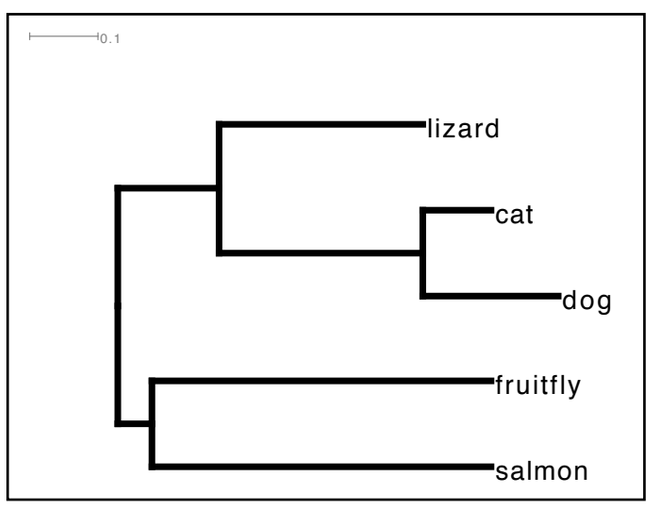


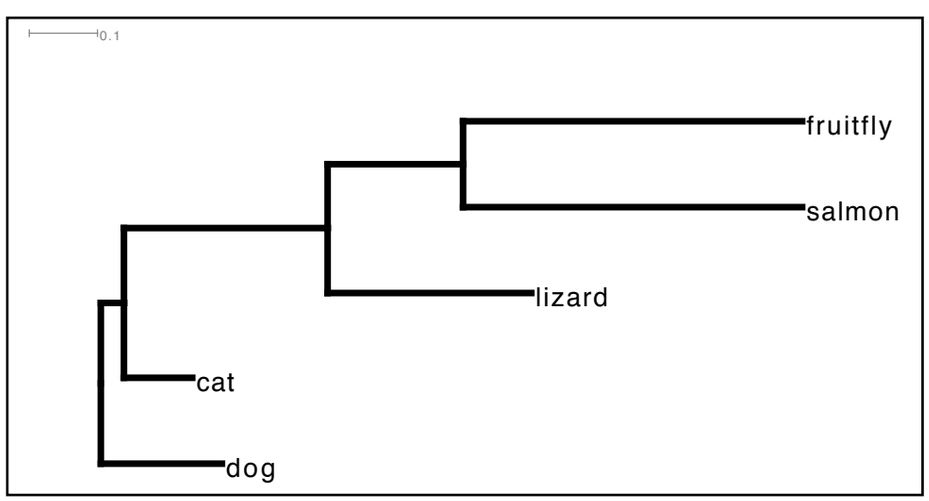


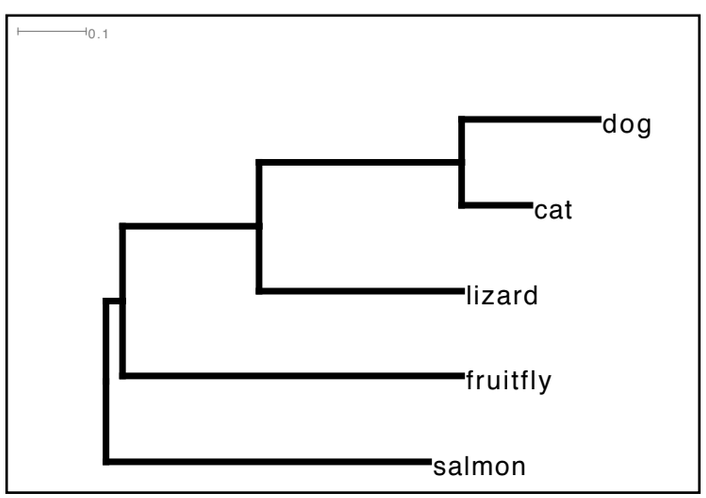






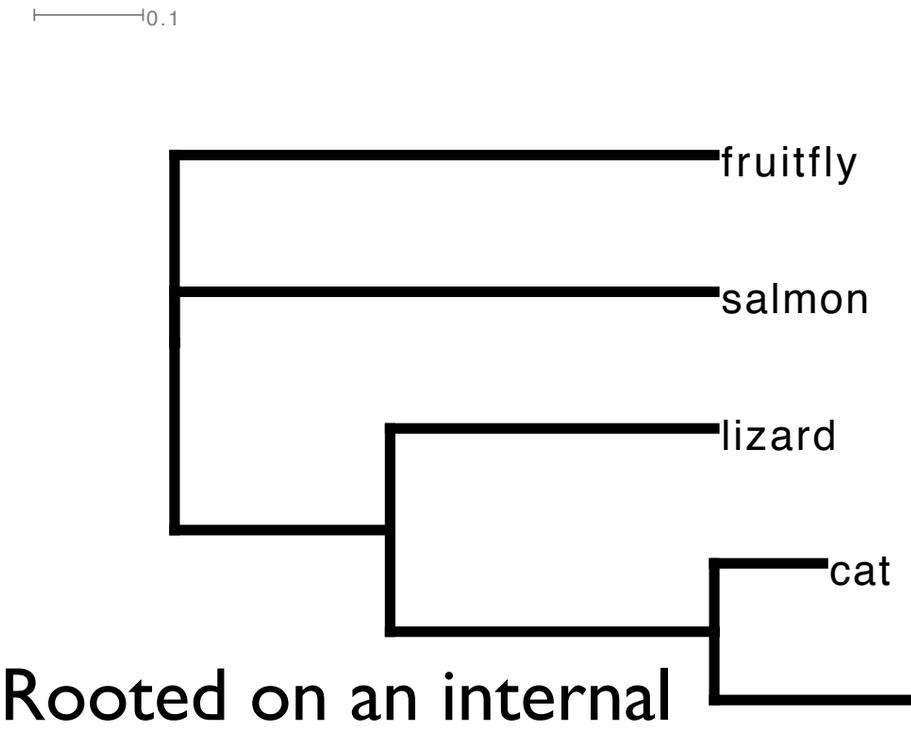




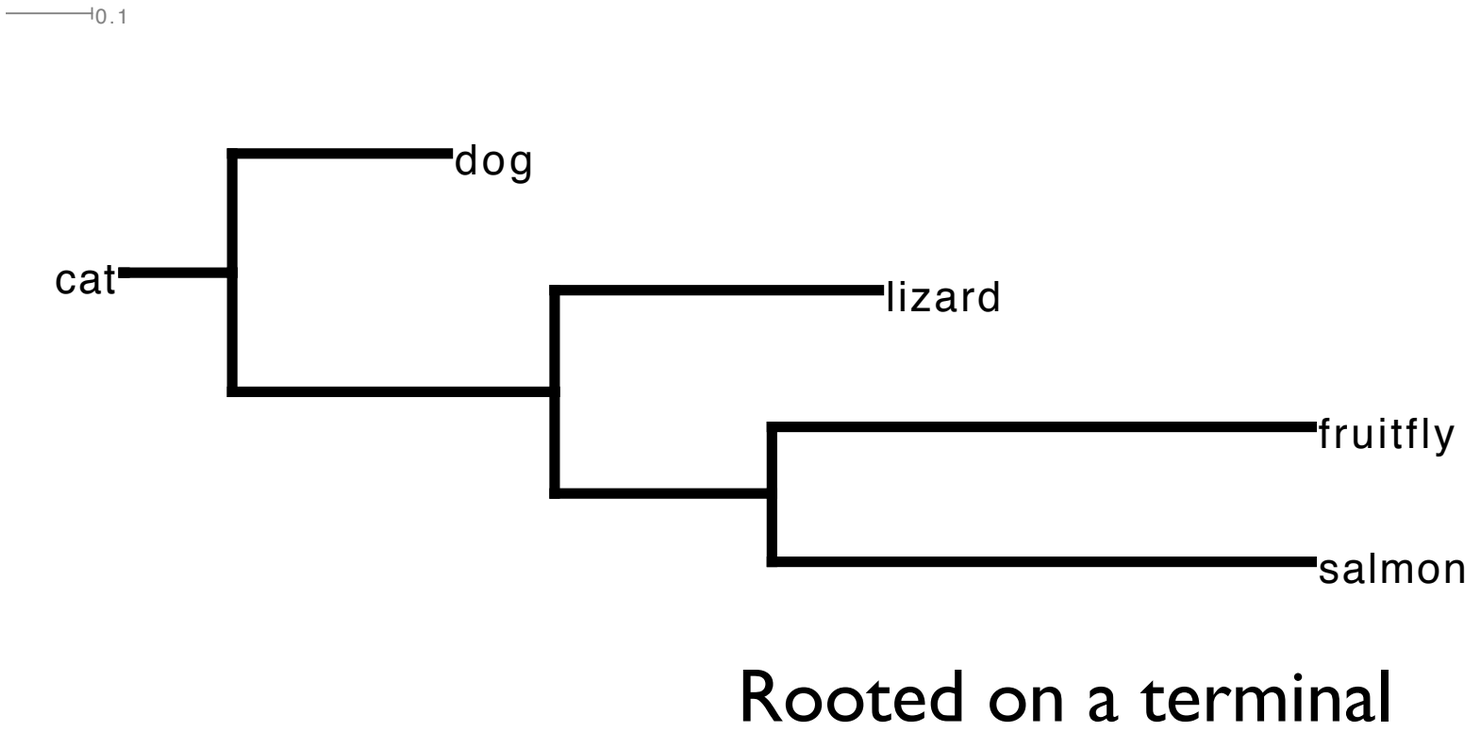


Aidan Budd, EMBL Heidelberg

Roots on Terminal and Internal Nodes

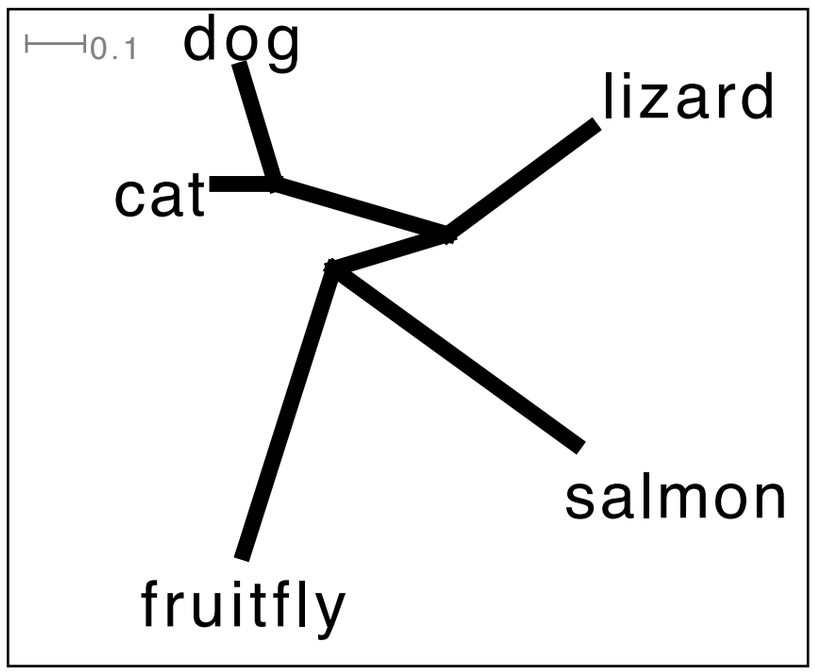


Rooted on an internal node of unrooted tree

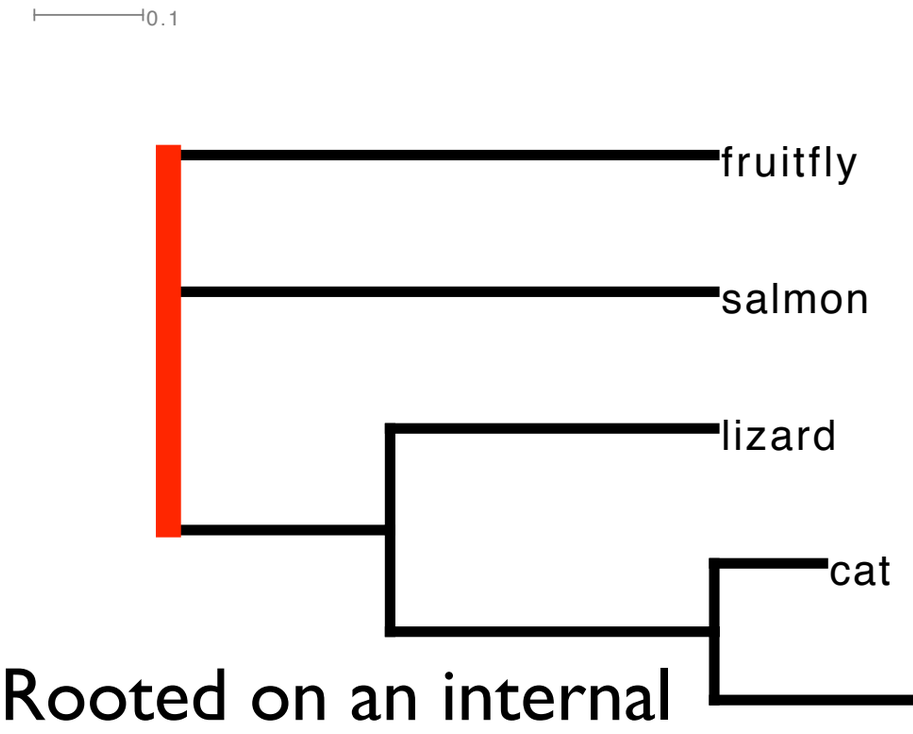


Rooted on a terminal node of unrooted tree

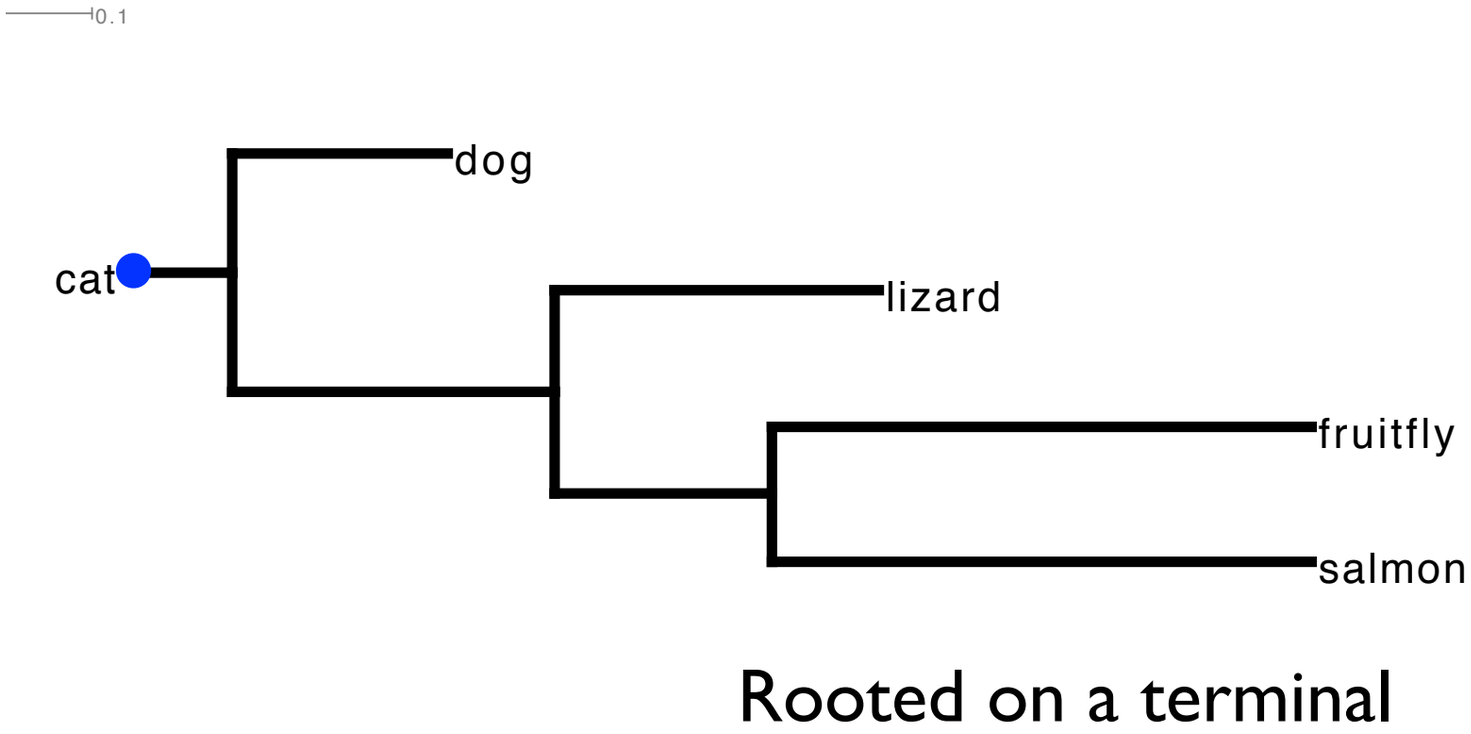
On the unrooted tree image to the right, label the two nodes on which the two above trees are rooted



Roots on Terminal and Internal Nodes

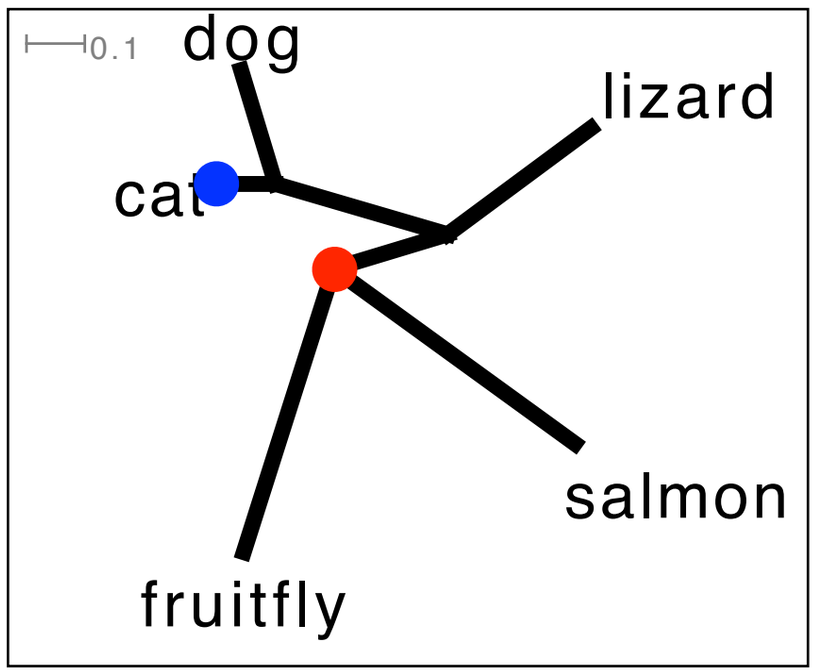


Rooted on an internal node of unrooted tree



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



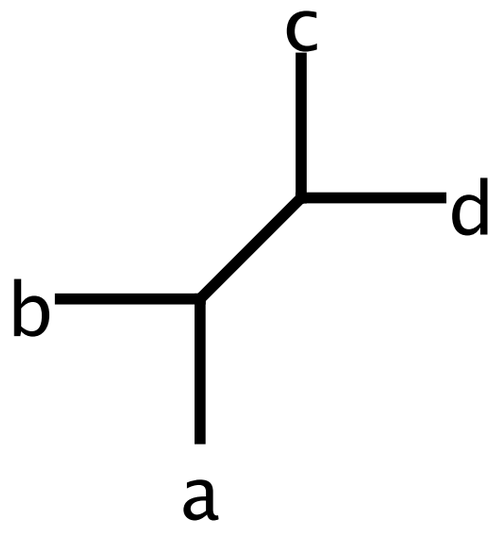
Aidan Budd, EMBL Heidelberg

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.

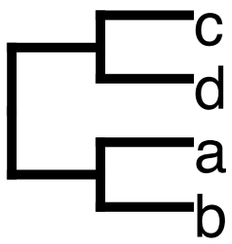
And, while 3. is true for some of the rooted trues, in others it is not

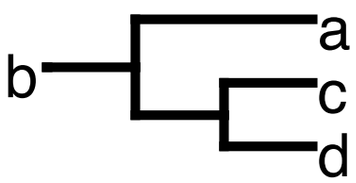
Draw the set of rooted tree topologies in which statement 3. is:

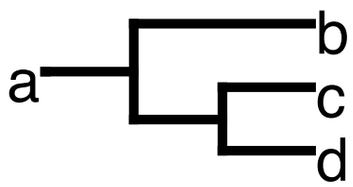
- true
- false

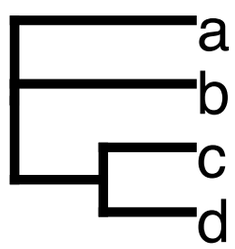
Quiz

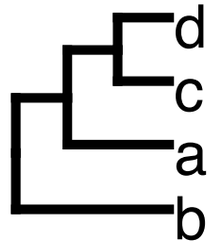
d more closely related to **c** than it is to **a** or **b**

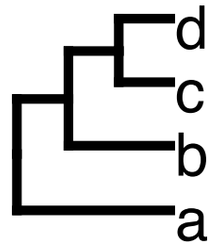




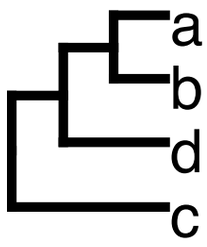


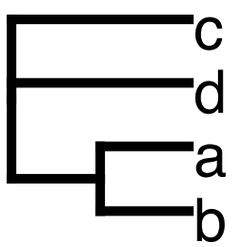


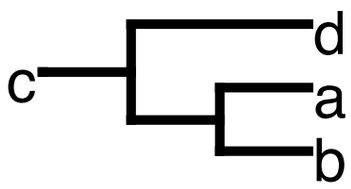


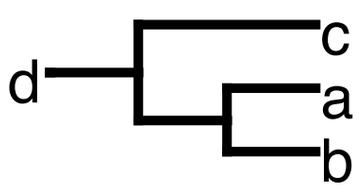


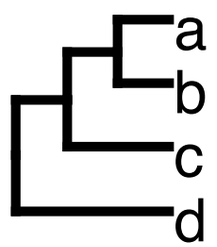
d not more closely related to **c** than it is to **a** or **b**





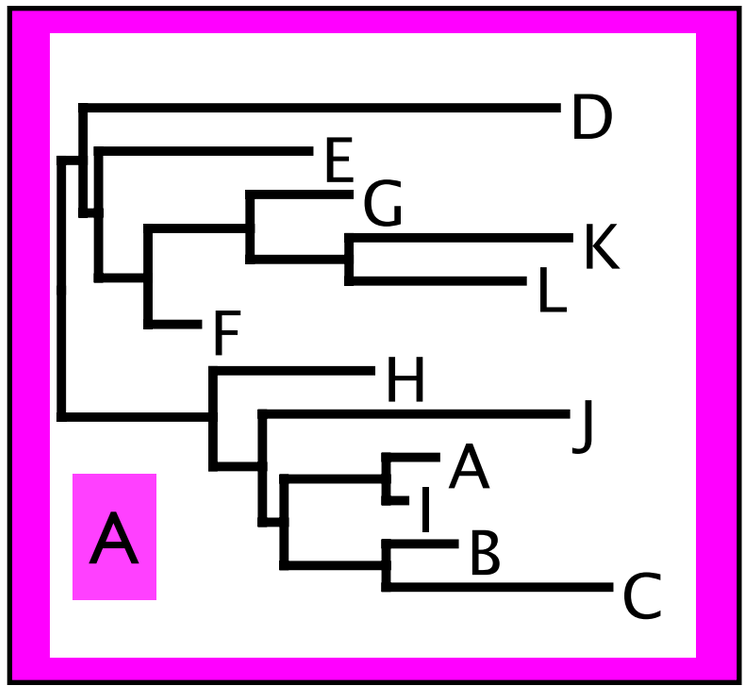


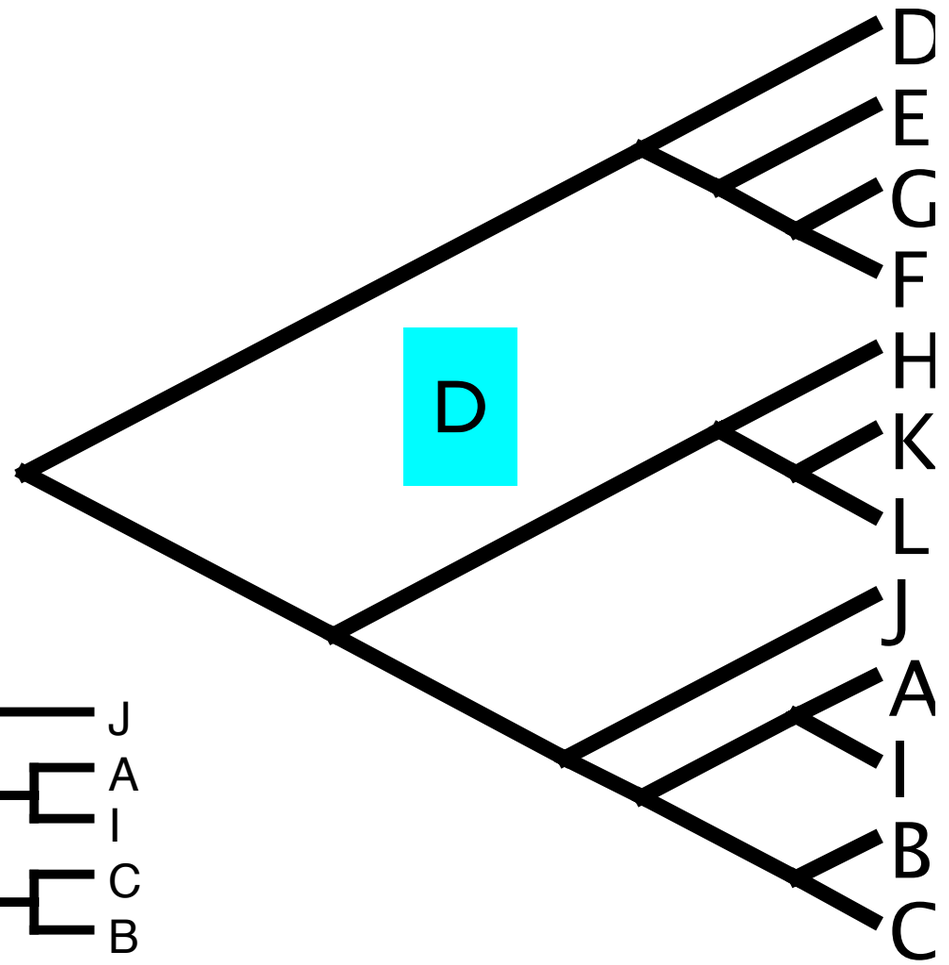


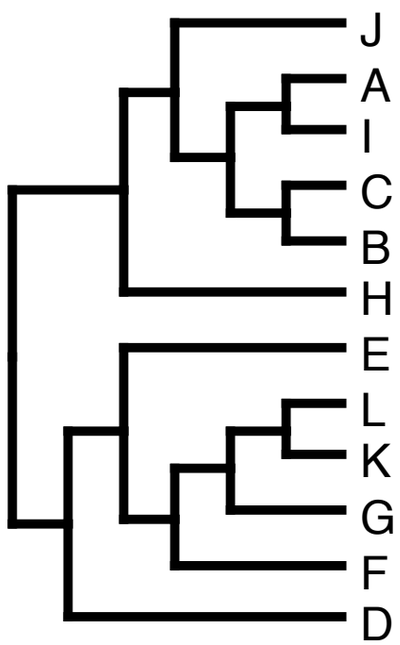


Quiz: recognise identical topologies

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







Aidan Budd, EMBL Heidelberg

Visualising Trees

Demonstration and Exercise

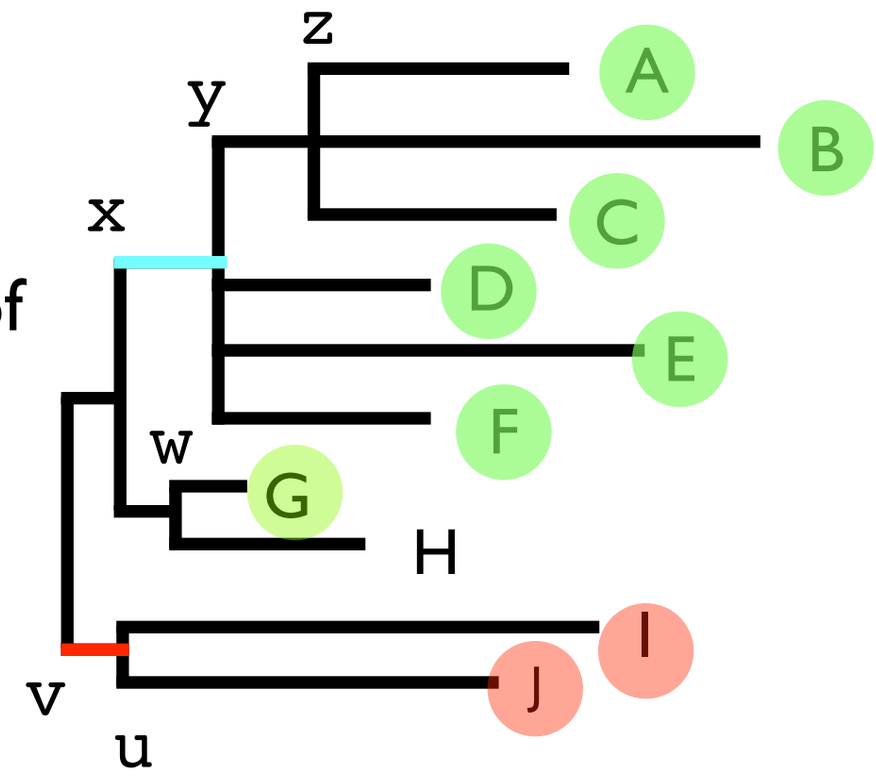
Viewing and manipulating unscaled trees with NJplot

- Rotating around internal branches
- Re-rooting

Clades

Clade:

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



Clades:

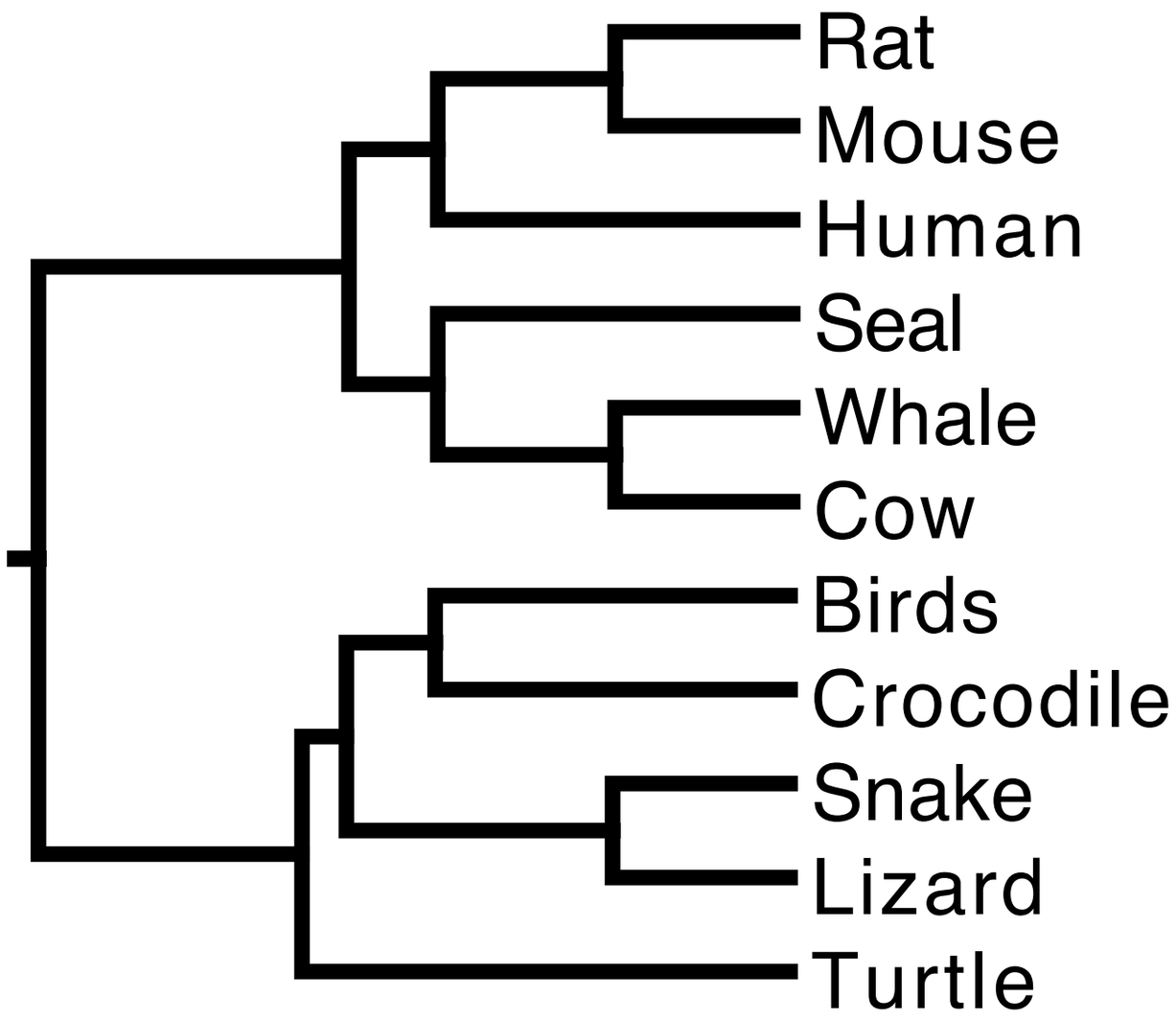
Branch xy specifies the clade ABCDEF

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

Clades



Cartoon phylogeny of selected amniotes

Identify groups of taxa, where the groups have common names, that are:

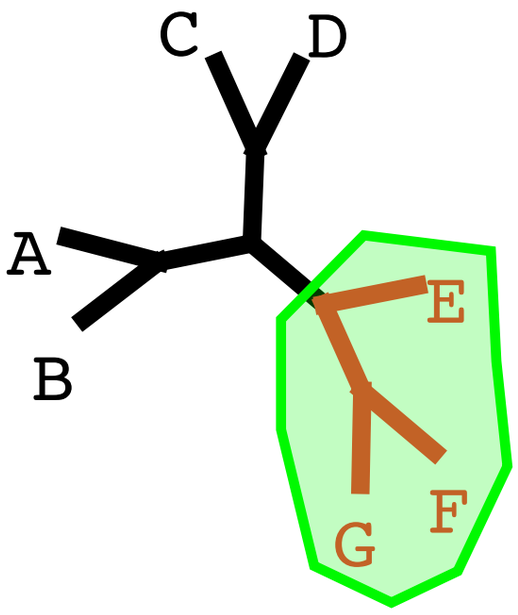
clades

not clades

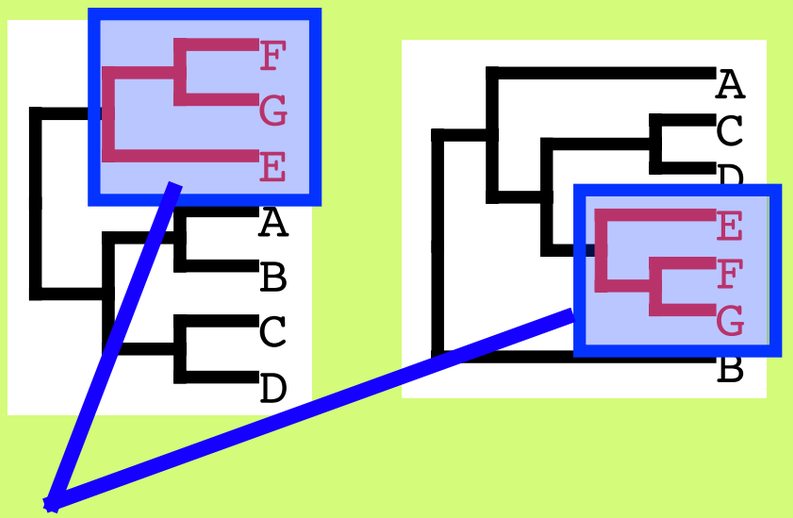
e.g. rodents, reptiles, mammals?

Clans

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



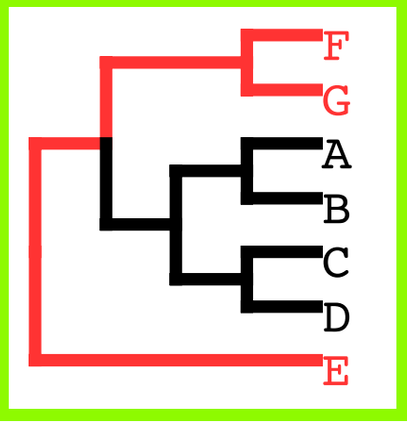
Of clades and clans: terms for phylogenetic relationships in unrooted trees.
Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM.
Trends Ecol Evol. 2007 Mar;22(3):114-5.
PMID: 17239486

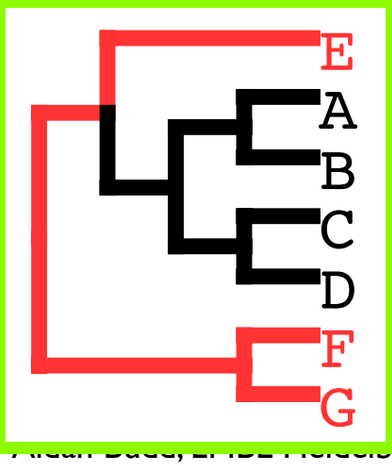


clades

therefore EFG form a **clan**

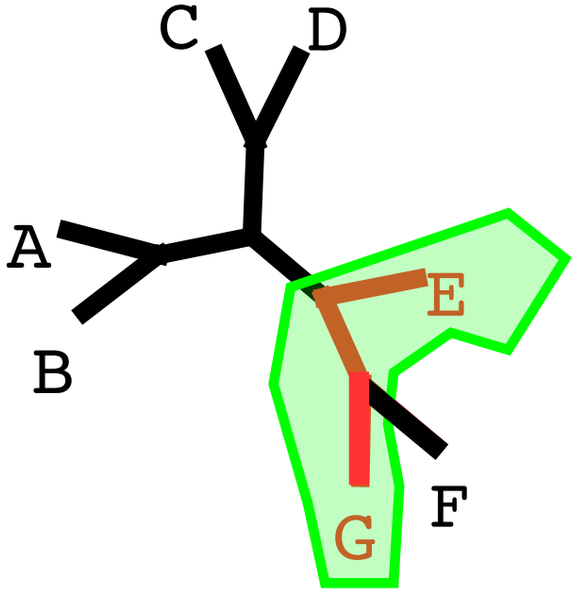
However! Under some rootings EFG **does not** form a clade

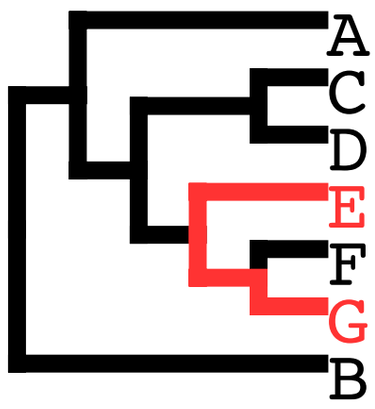


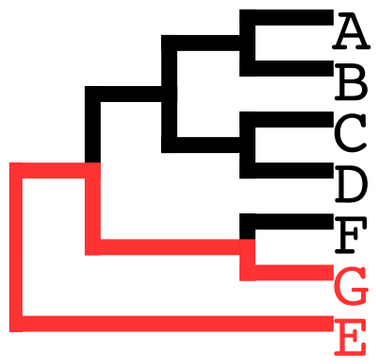


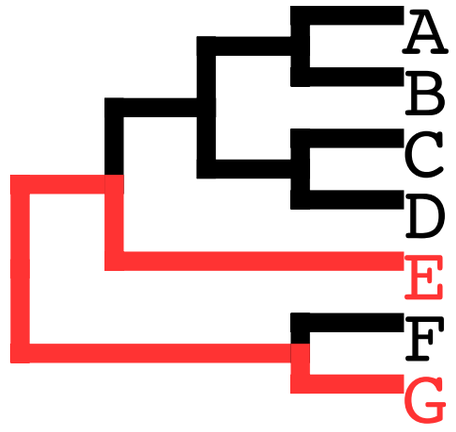
Clans

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









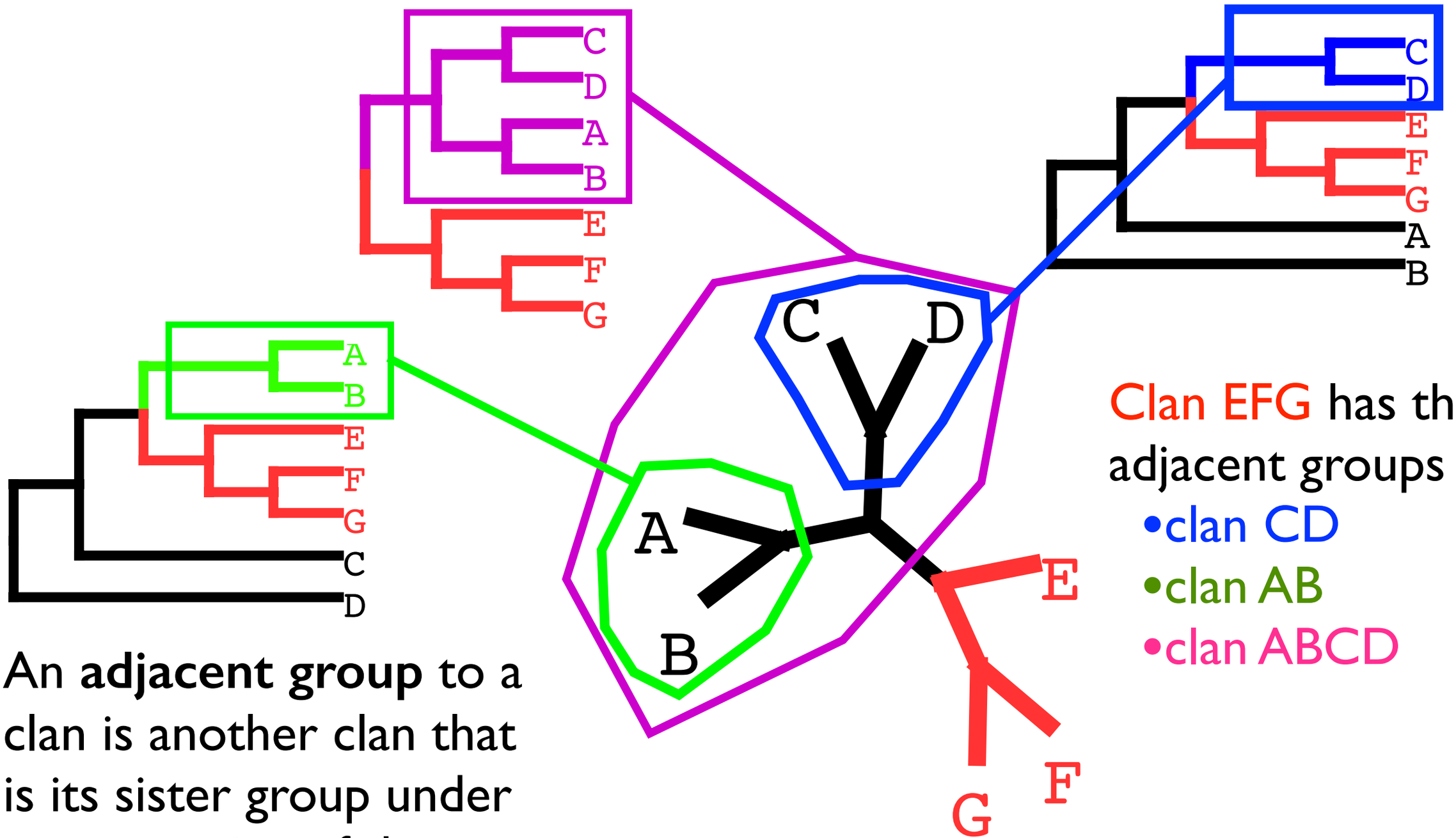
...

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



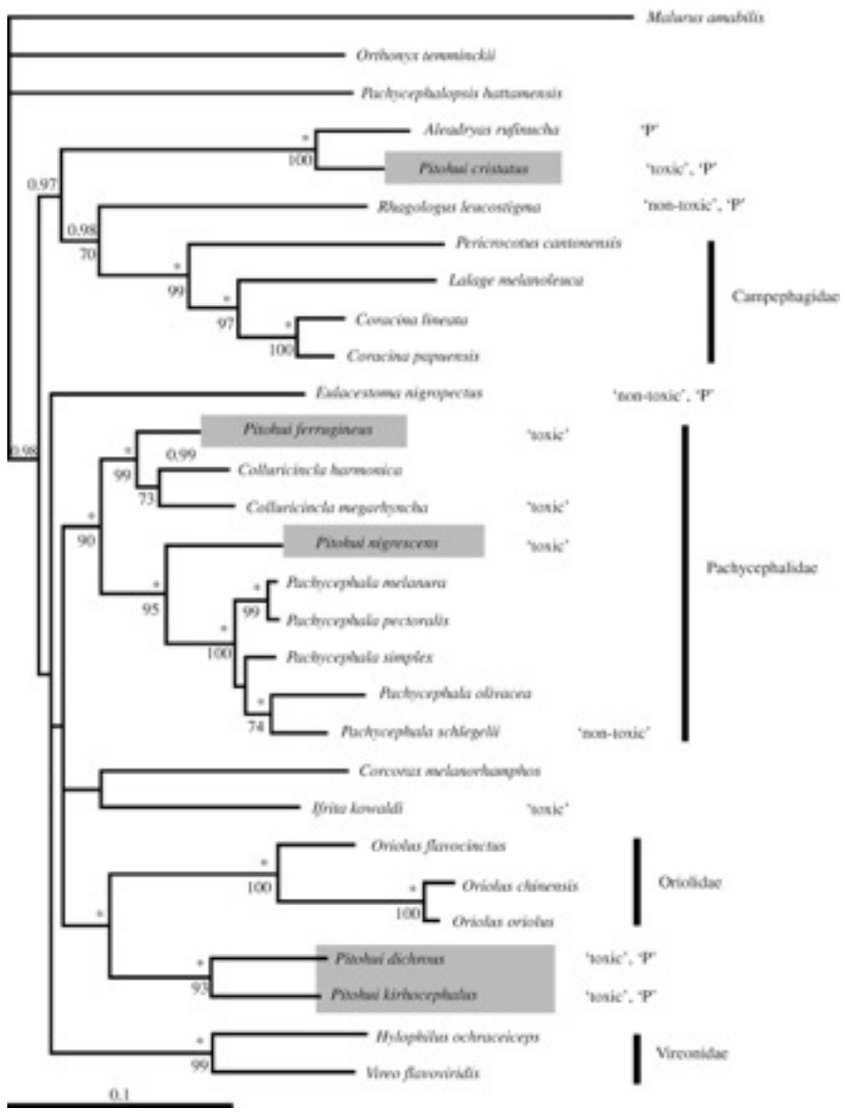
An **adjacent group** to a clan is another clan that is its sister group under some rooting of the tree

Clan EFG has three adjacent groups

- clan CD
- clan AB
- clan ABCD

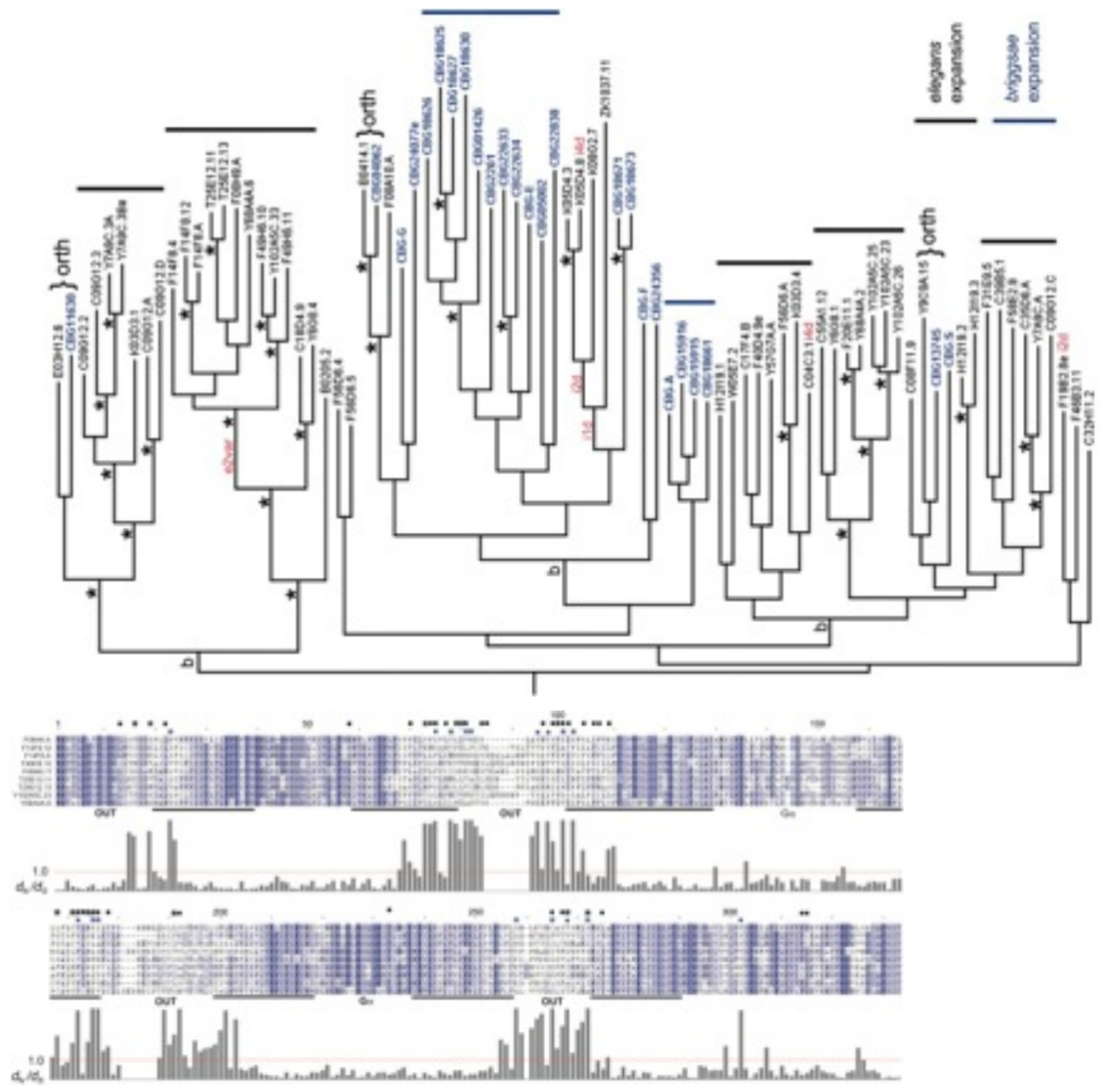
Of clades and clans: terms for phylogenetic relationships in unrooted trees.
Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM.
Trends Ecol Evol. 2007 Mar;22(3):114-5.
PMID: 17239486

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



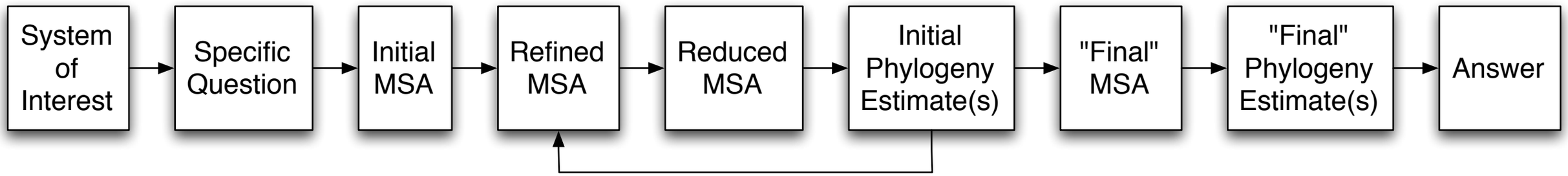
Sites of positive selection identified using reversible models of codon substitution

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

Example Phylogeny Estimation Workflow

Aidan Budd, EMBL Heidelberg

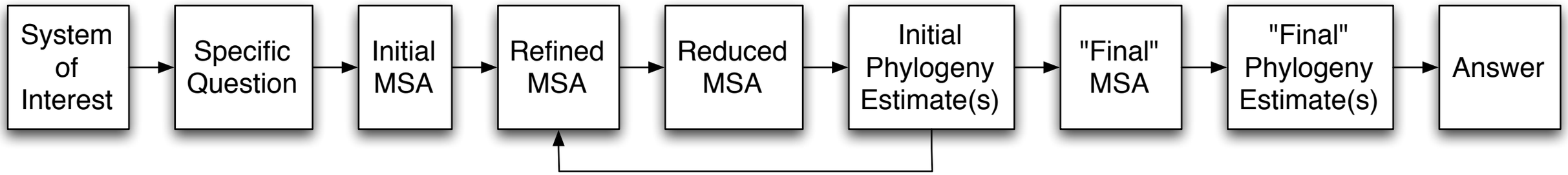
Example Phylogeny Estimation Workflow



Aims:

- Provide guidance/reference for planning your own analyses
- Show how different tools/stages in an analysis can link together
  - i.e. providing a context in which to place what you learn later on
- Provide a "target" to criticise
  - once you know more, what would you do differently?

Example Phylogeny Estimation Workflow



Aims:

- Introduce some common tools
- Highlight some common problems
- Highlight the importance of:
  - **Formulating an appropriate question**
  - **Examining the output from MSA and phylogenetic tools**

Example Phylogeny Estimation Workflow

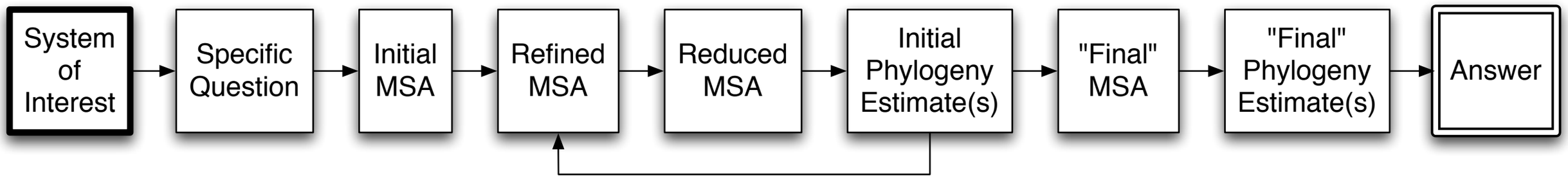
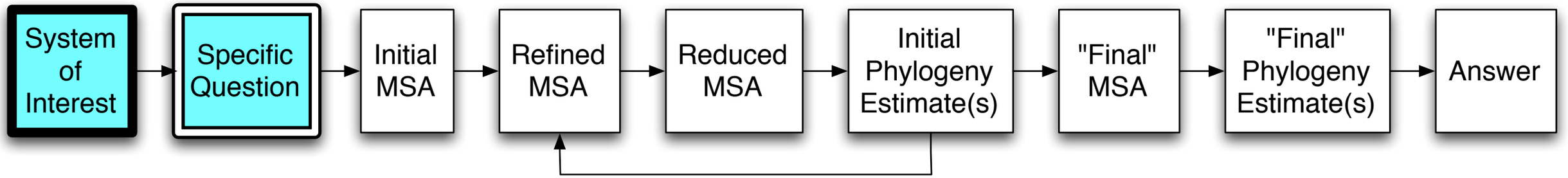
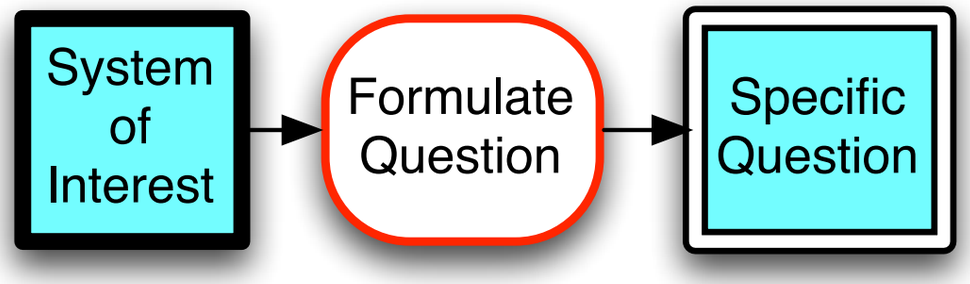


Figure 19

Formulating the Question





We do an analysis for a reason!

Usually because we are interested in the result...

... i.e. hope that it will help us choose between a set of alternative models for our observations of a biological system

Can you think of other possible reasons you might do a phylogenetic analysis?

Formulating the Question

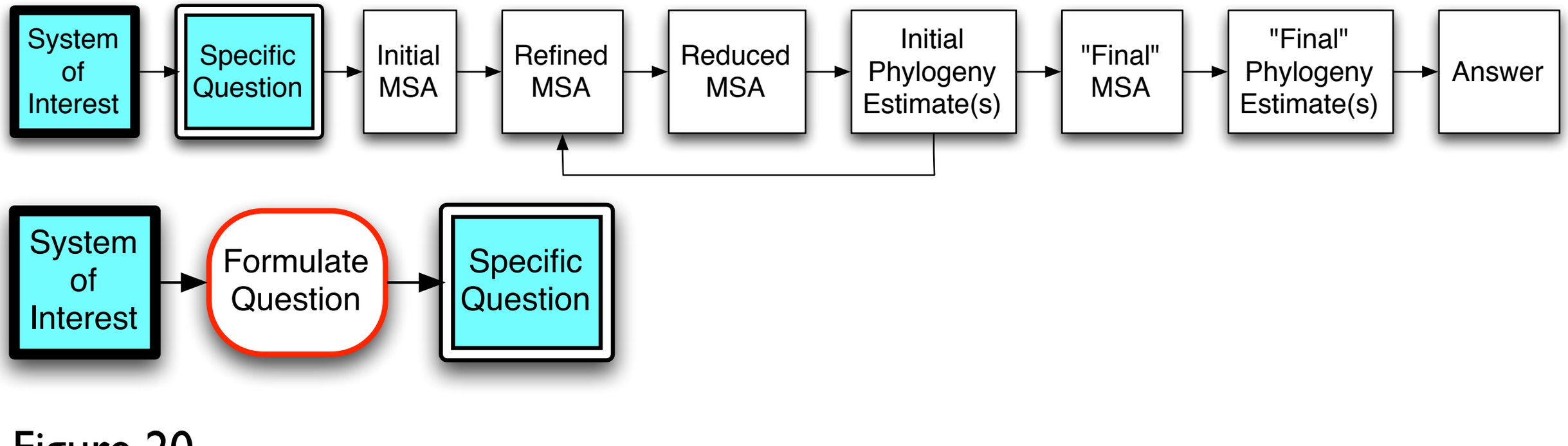
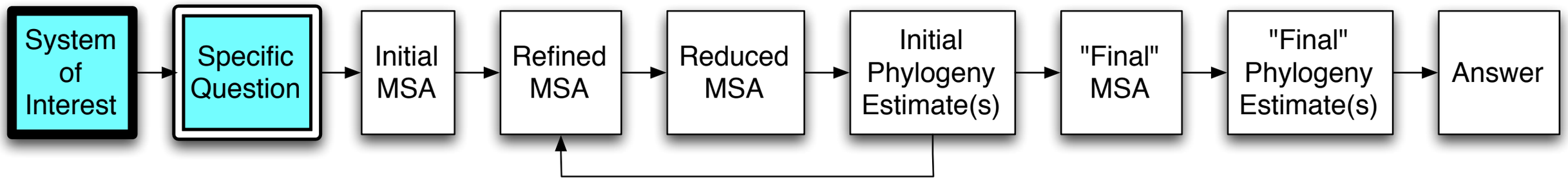
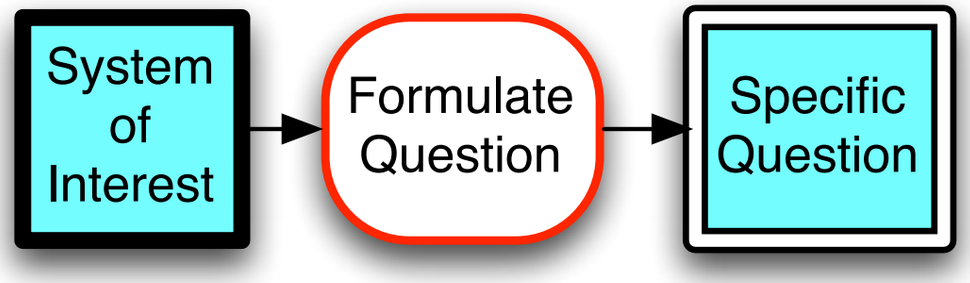


Figure 20

Formulating the Question



Having a clear awareness of our reason for doing the analysis (i.e. the specific question it aims to address) contributes to its success...



... because...

... most questions can be addressed in several (many!) different ways i.e. using different data/tools/tool parameters and...

... typically, a question is better addressed in some ways compared to others...

... so, our choice of data/tools/tool parameters influences the quality of our analysis

Thus... knowledge of:

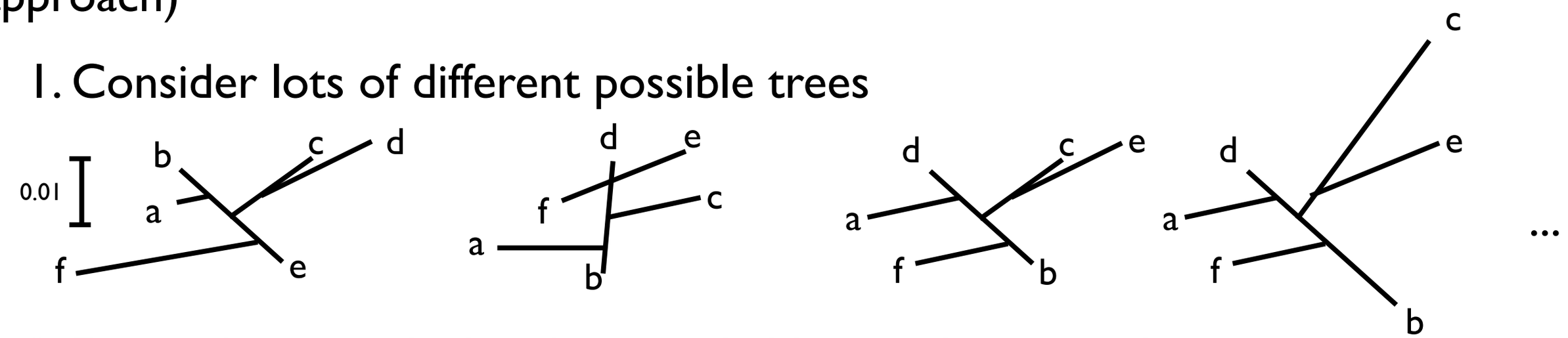
- how the tools work (and thus which questions they are good for addressing)
- the specific question being addressed

can help us choose an appropriate analysis

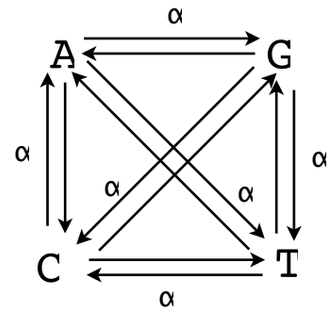
Phylogenetic Estimation in One Slide

One way you could try to estimate a phylogeny (the maximum likelihood approach)

1. Consider lots of different possible trees



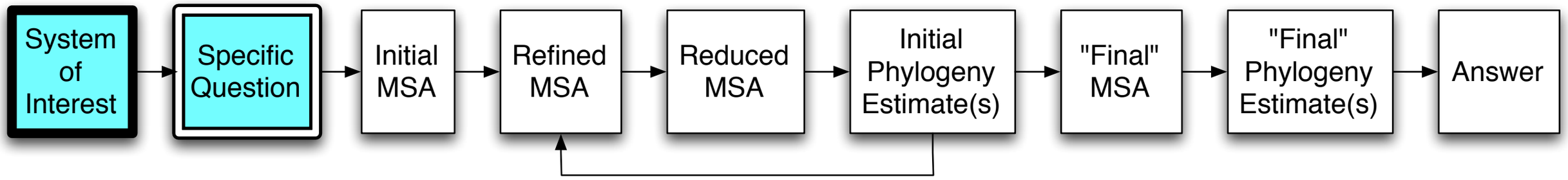
2. For each tree, calculate how probable it makes your data

(based on using a model of substitutions 

to calculate the probability of each branch, given the alignment)

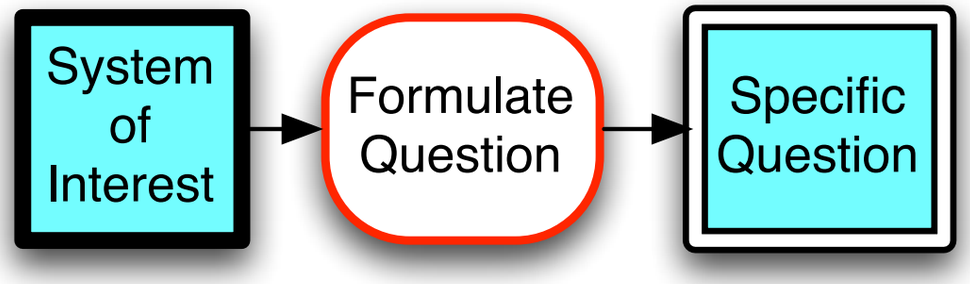
3. Choose the tree that makes the data most probable as your estimate of the phylogeny

Formulating the Question

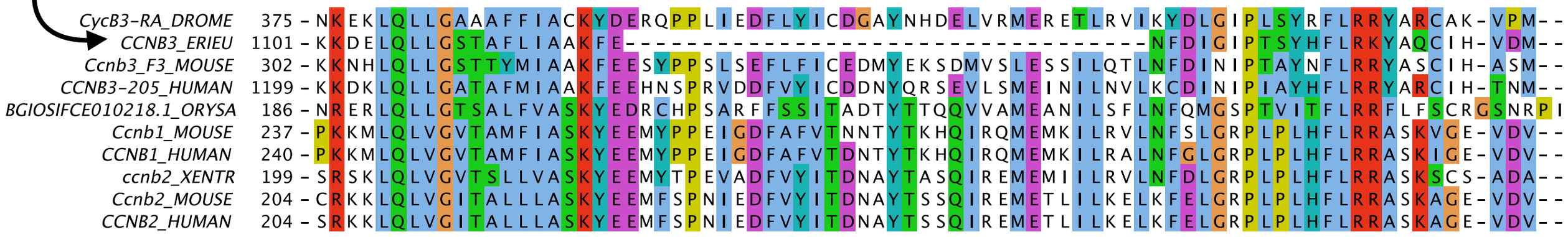


Example

Should the CCNB3_ERIEU sequence be removed from an analysis based on this alignment?

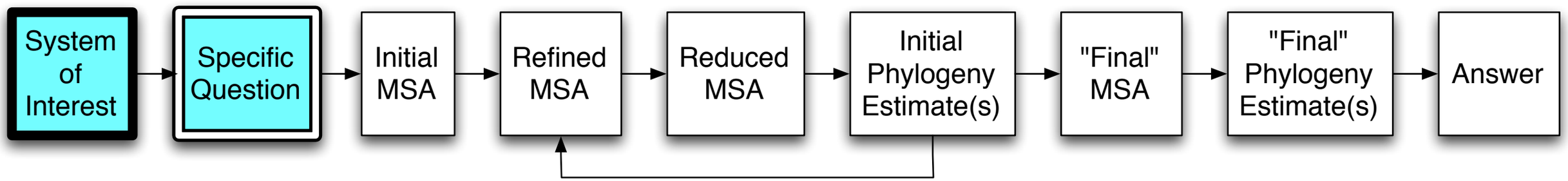


Who thinks it should be removed?



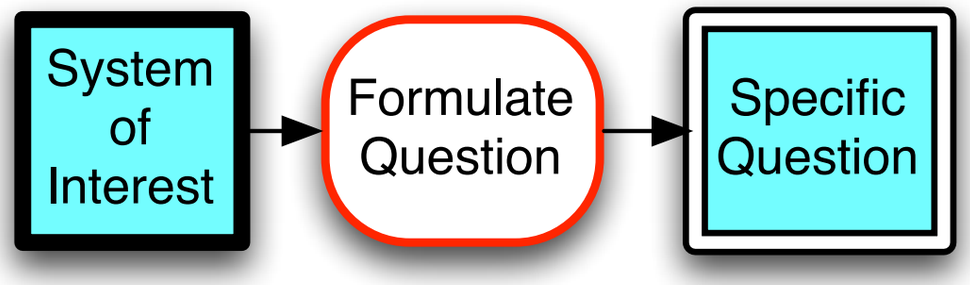
Cyclin B, TreeFam family TF101001

Formulating the Question



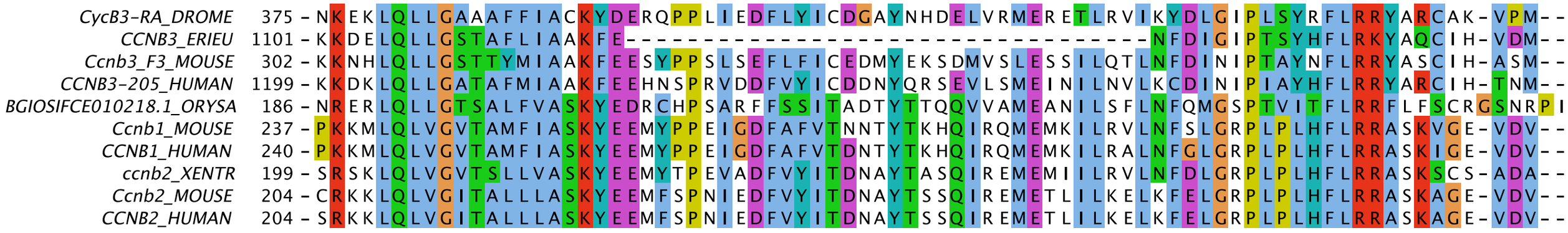
Example

Should the CCNB3_ERIEU sequence be removed from an analysis based on this alignment?

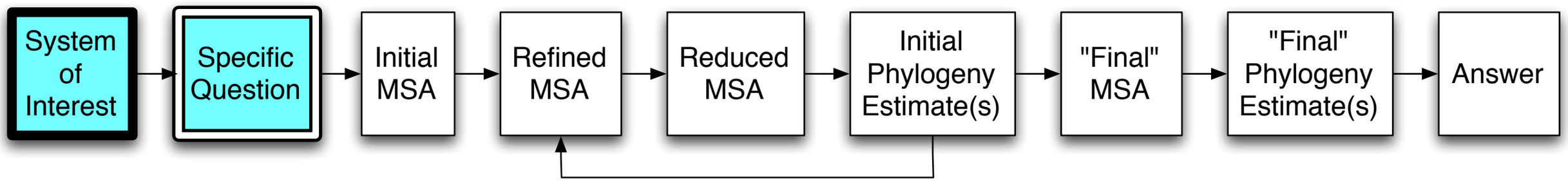


Decision is informed by our knowledge of:

- how the methods work
- the specific question(s) we want to address with the analysis

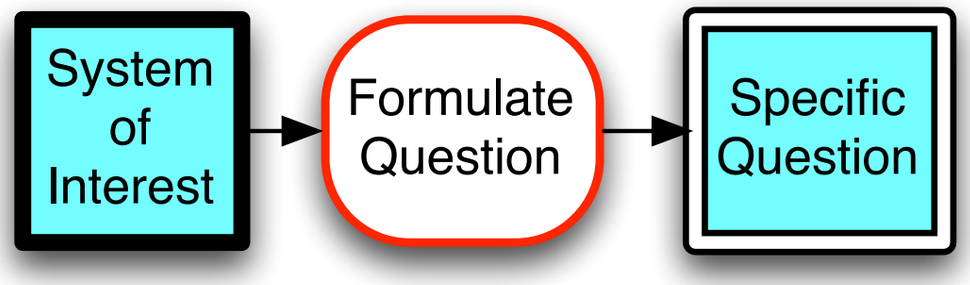


Formulating the Question



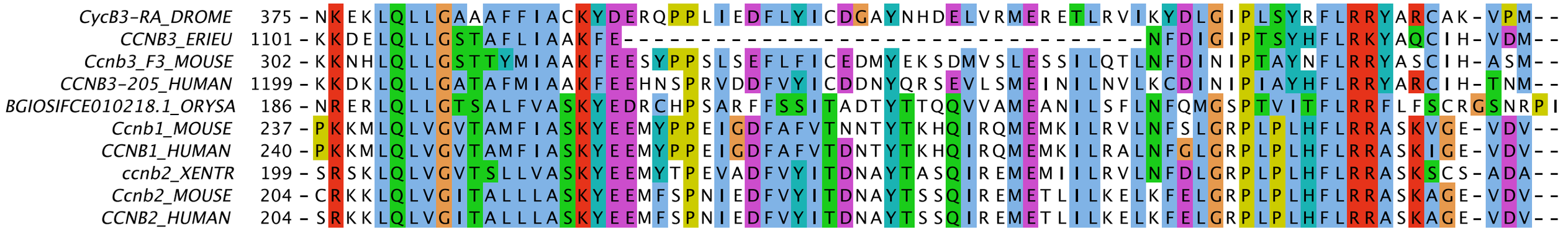
Example

Should the CCNB3_ERIEU sequence be removed from an analysis based on this alignment?



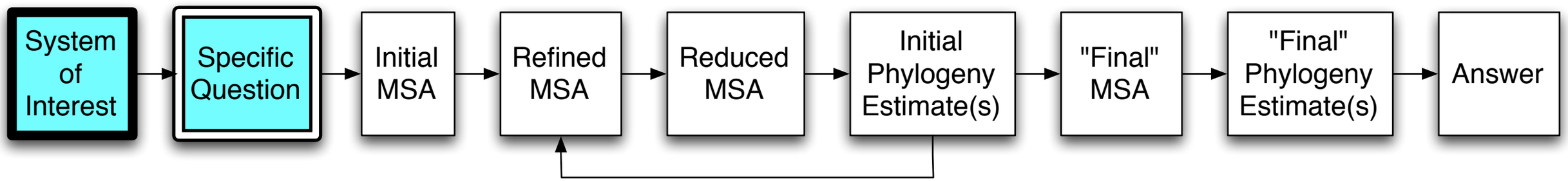
Decision is informed by our knowledge of:

- how the methods work
  - columns containing gaps are ignored by/cause problems for some phylogeny estimation software
  - removing CCNB3_ERIEU sequence may increase the number of columns analysed
- the specific question(s) we want to address with the analysis



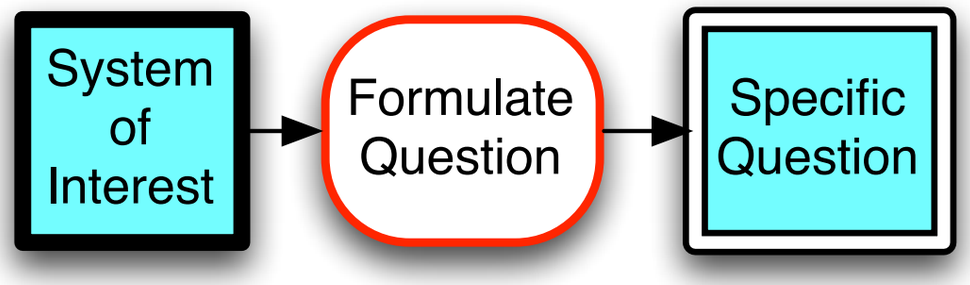
Cyclin B, TreeFam family TF101001

Formulating the Question



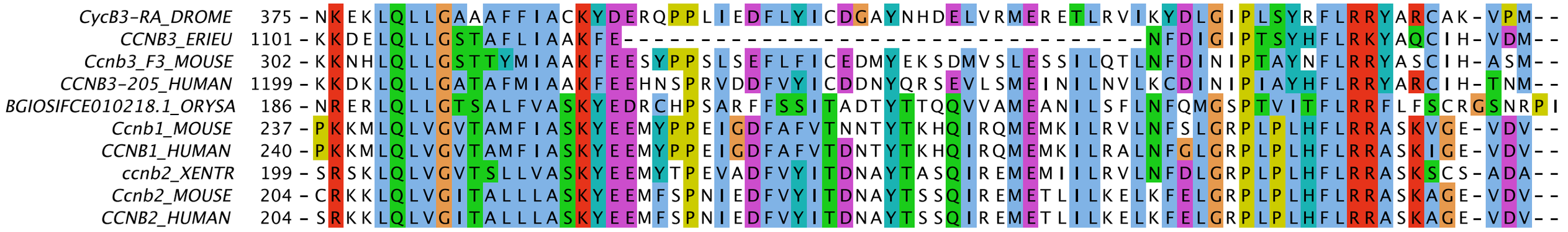
Example

Should the CCNB3_ERIEU sequence be removed from an analysis based on this alignment?



Decision is informed by our knowledge of:

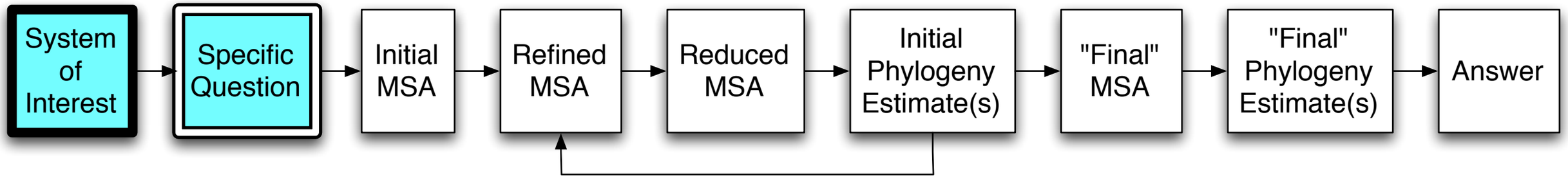
- how the methods work
- the specific question(s) we want to address with the analysis
  - **if** addressing the questions **requires** CCNB3_ERIEU? - then we **must** include it
  - **if** addressing the questions **does not require** CCNB3_ERIEU? - then we **could**



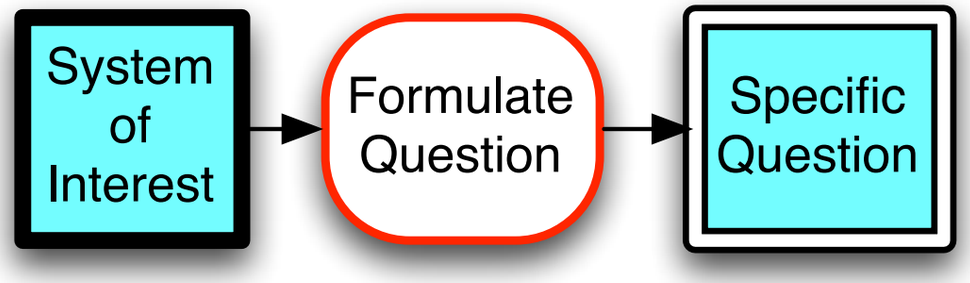
Cyclin B, TreeFam family TF101001

Aidan Budd, EMBL Heidelberg

Formulating the Question



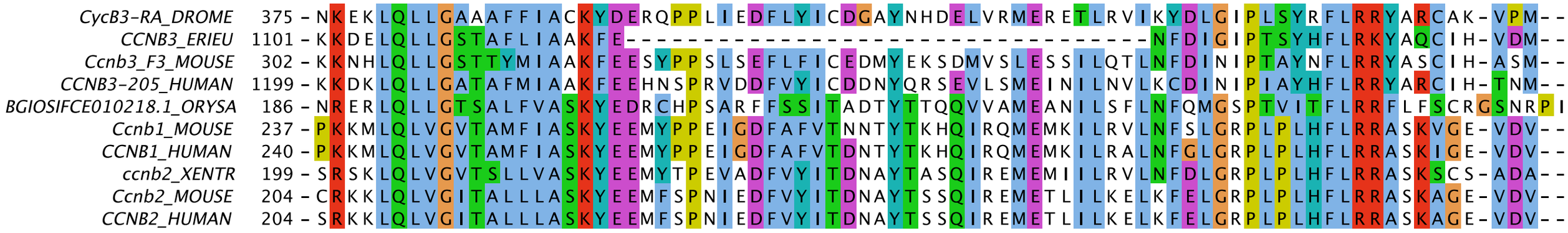
Usually, more focused/specific questions provide clearer guidelines for decision-making



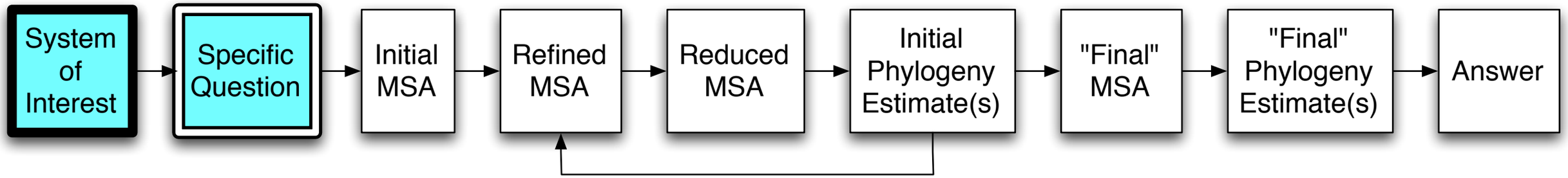
More Focused/Specific Example

Which mouse CCNB gene is most closely related to human CCNB1?

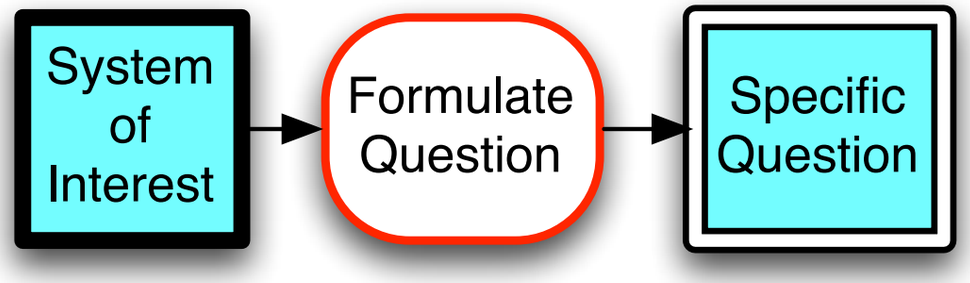
- Must include human CCNB1
- Should include as many mouse CCNB genes as possible
- Ideally include some other sequences - but could do analysis without them
- With this question, I **would remove CCNB3_ERIEU**



Formulating the Question



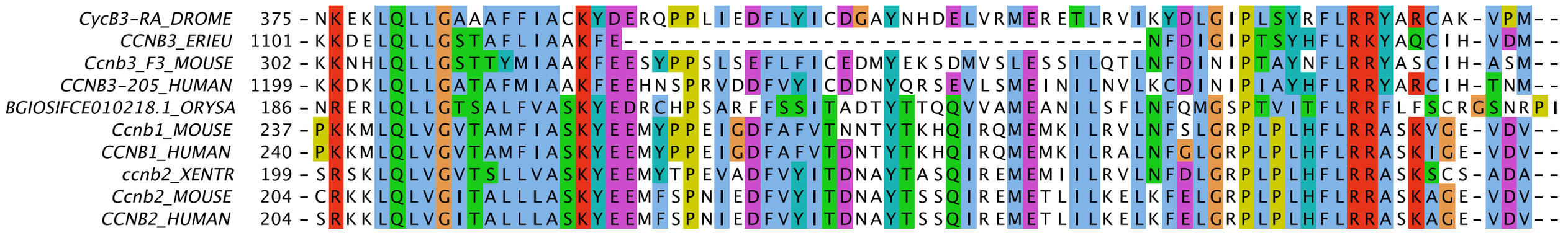
Usually, more focused/specific questions provide clearer guidelines for decision-making



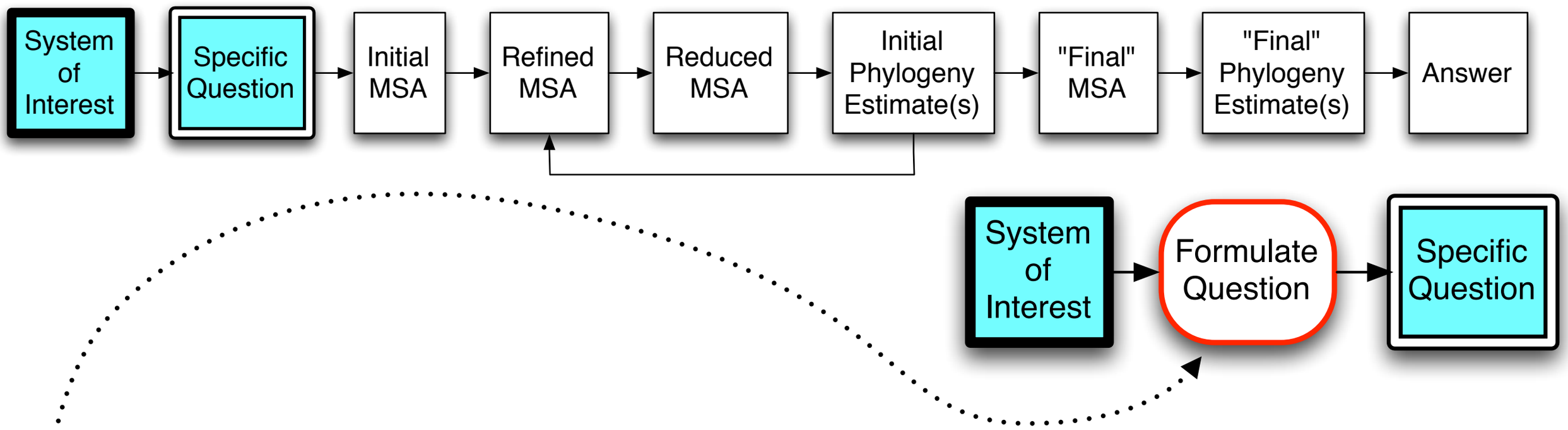
More General Example

What is the phylogeny of vertebrate CCNs?

- Should use as many complete vertebrate CCN genes as possible/reasonable
- Decision "include or exclude" for each sequence depends on lengths/quality/ numbers of other sequences - i.e. guidelines are much less clear-cut
- With this question, I am unsure whether to include or exclude CCNB3_ERIEU - it would depend on considering also other sequences involved in the analysis



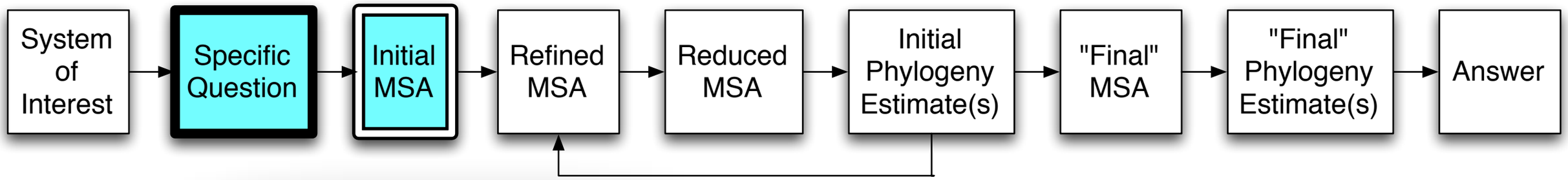
Formulating the Question

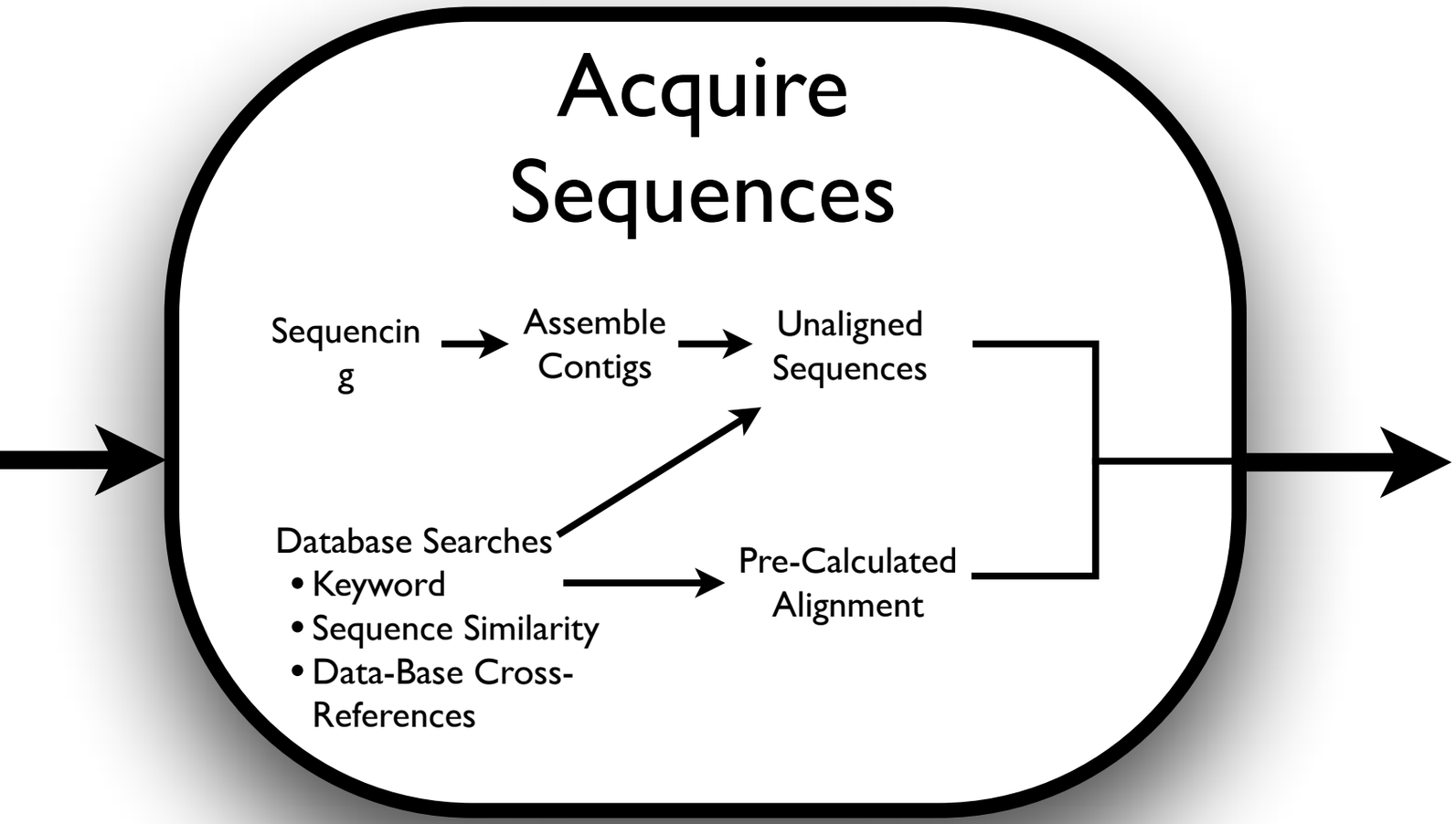


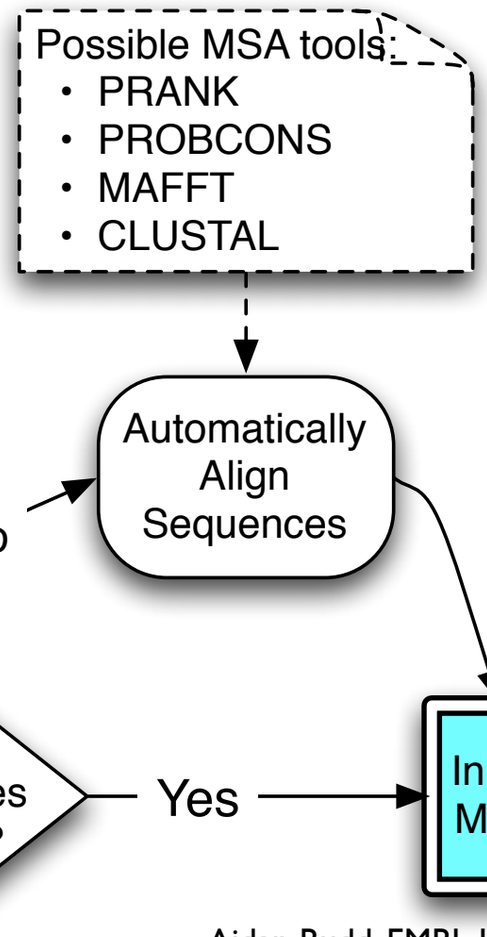
We emphasise the importance of question formulation (hence the red border in the diagram...)...

... as how we do this it influences many other decisions we make during the analysis

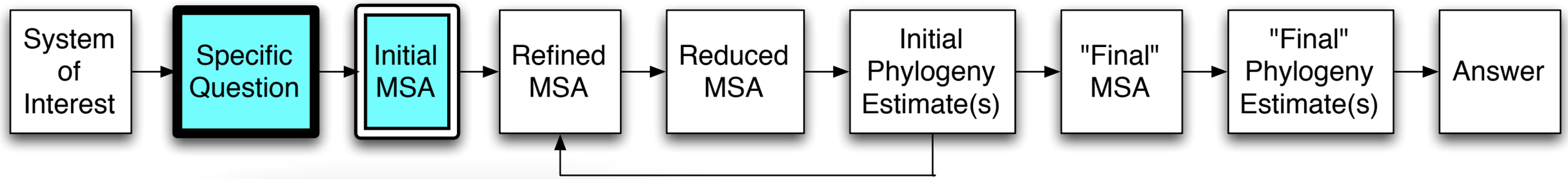
Building an Initial MSA: Acquiring Sequences

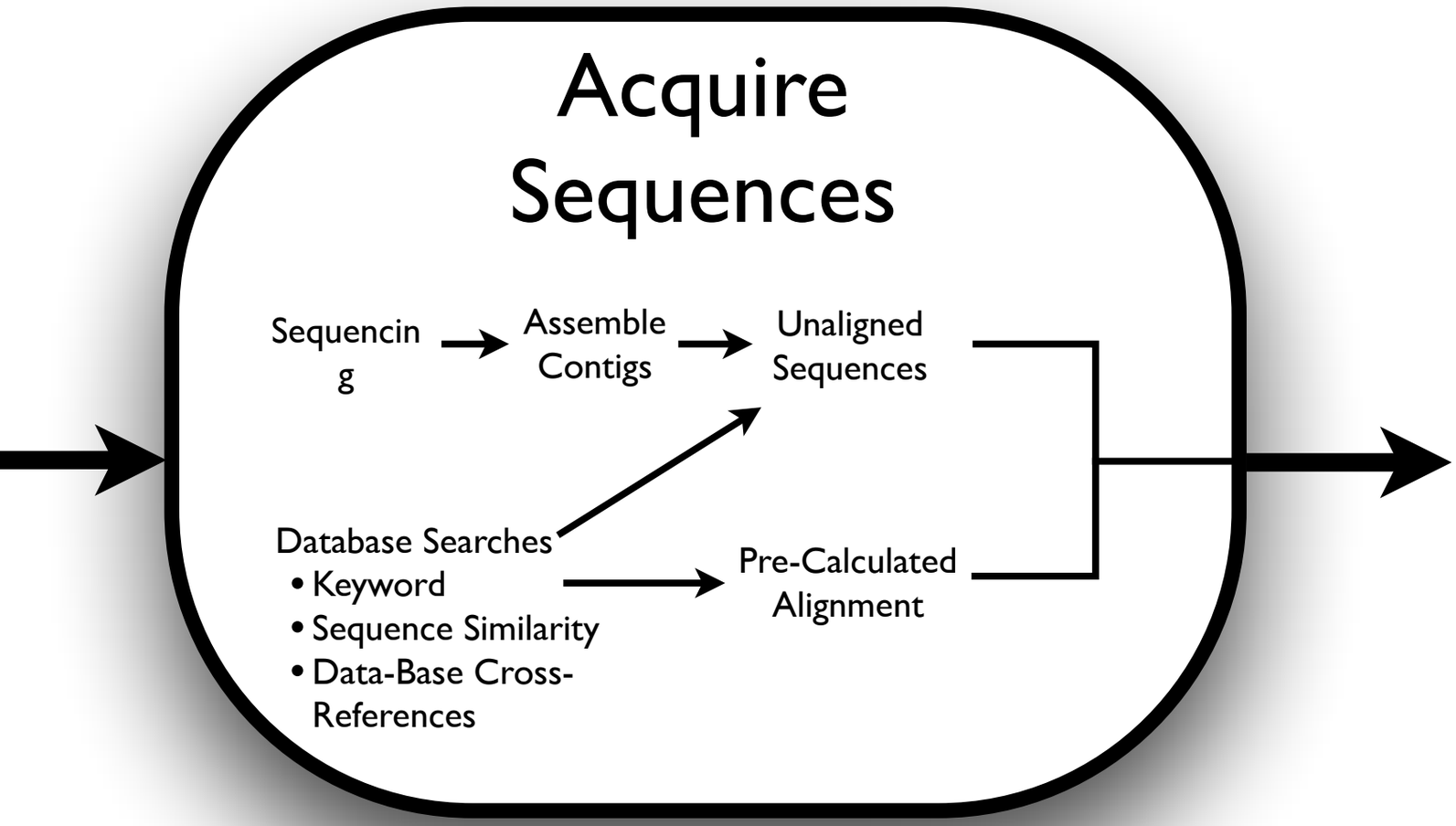






Building an Initial MSA: Acquiring Sequences





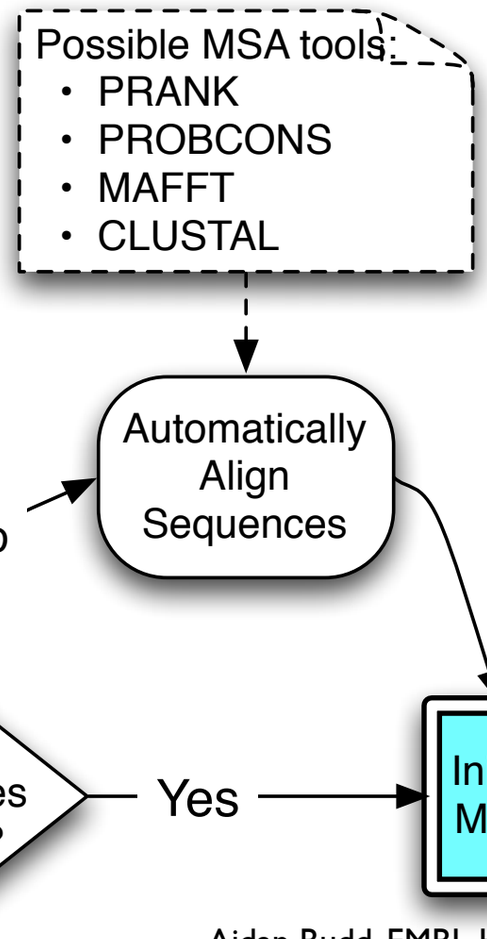
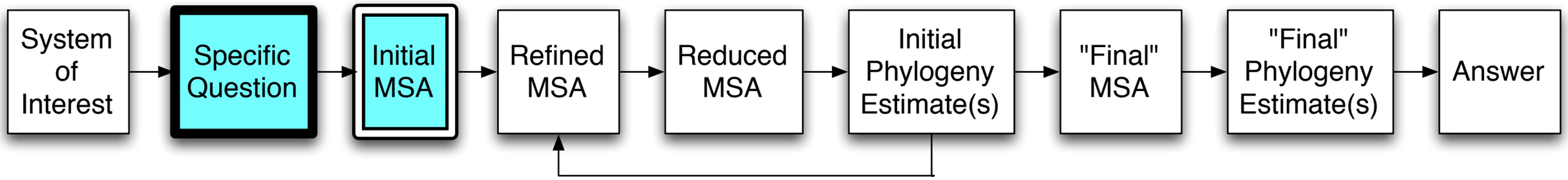


Figure 21

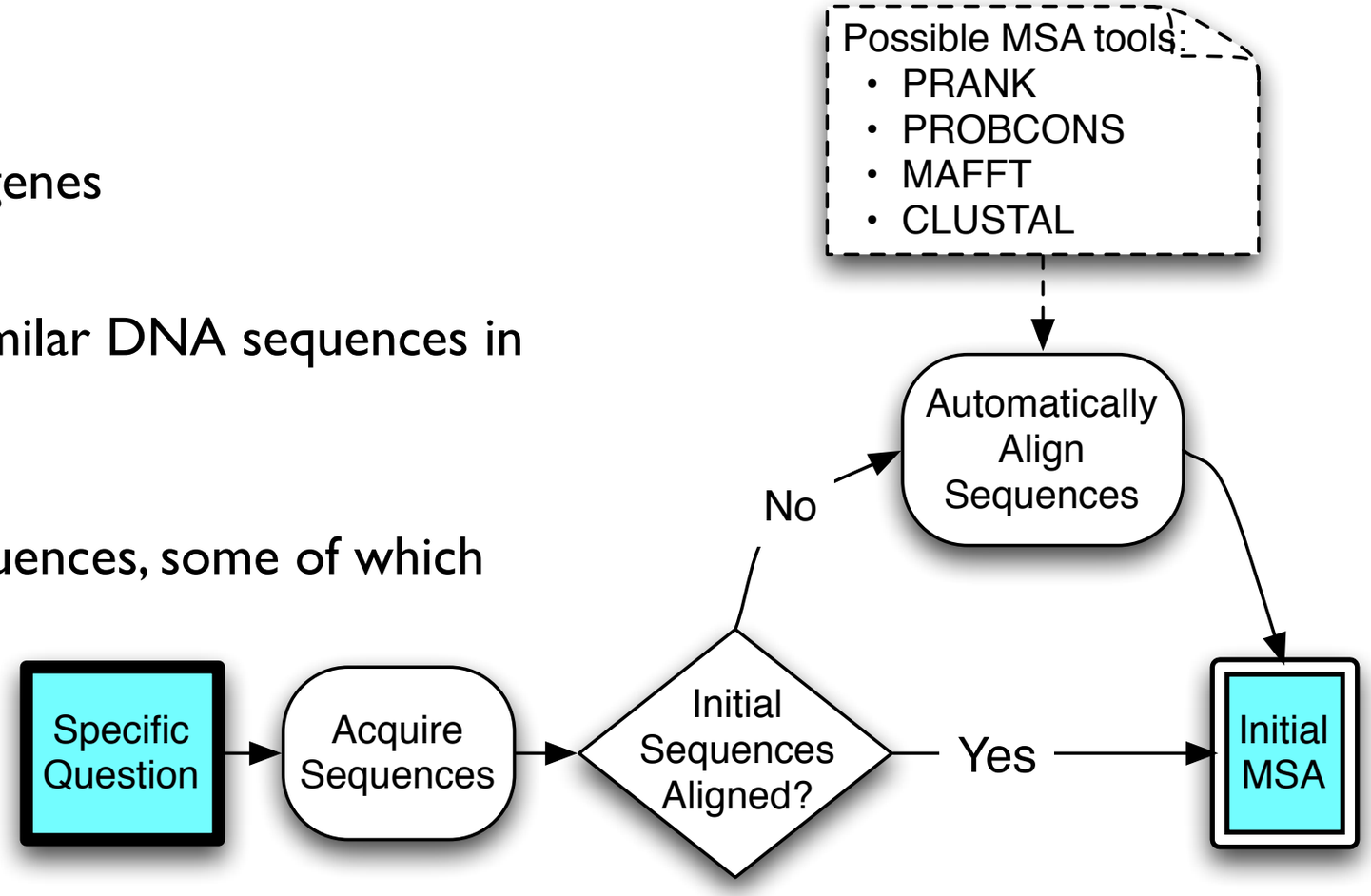
Aidan Budd, EMBL Heidelberg

Building an Initial MSA: Automatically Aligning Sequences

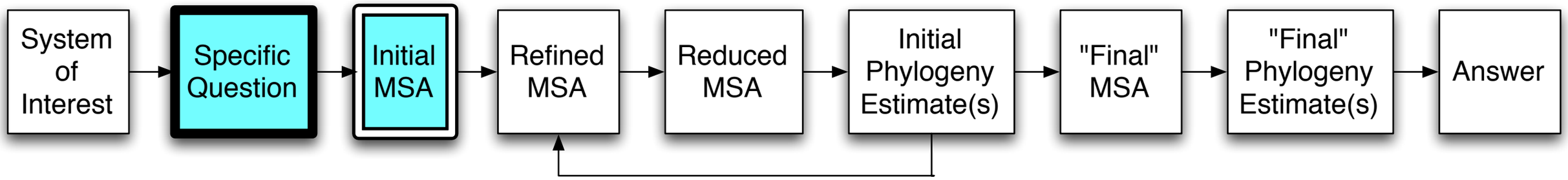


Different automatic MSA tools are designed for different tasks

- CLUSTALX, MUSCLE, PROBCONS
  divergent protein sequences
- NAST
  multiple alignment of 16S rRNA genes
- PRANK
  multiple alignment of relatively similar DNA sequences in an evolutionary context
- EXPRESSO(3DCoffee)
  multiple alignment of protein sequences, some of which have 3D structural information
- MAUVE, Enredo
  multiple alignment of genomes
- and many others...



Building an Initial MSA

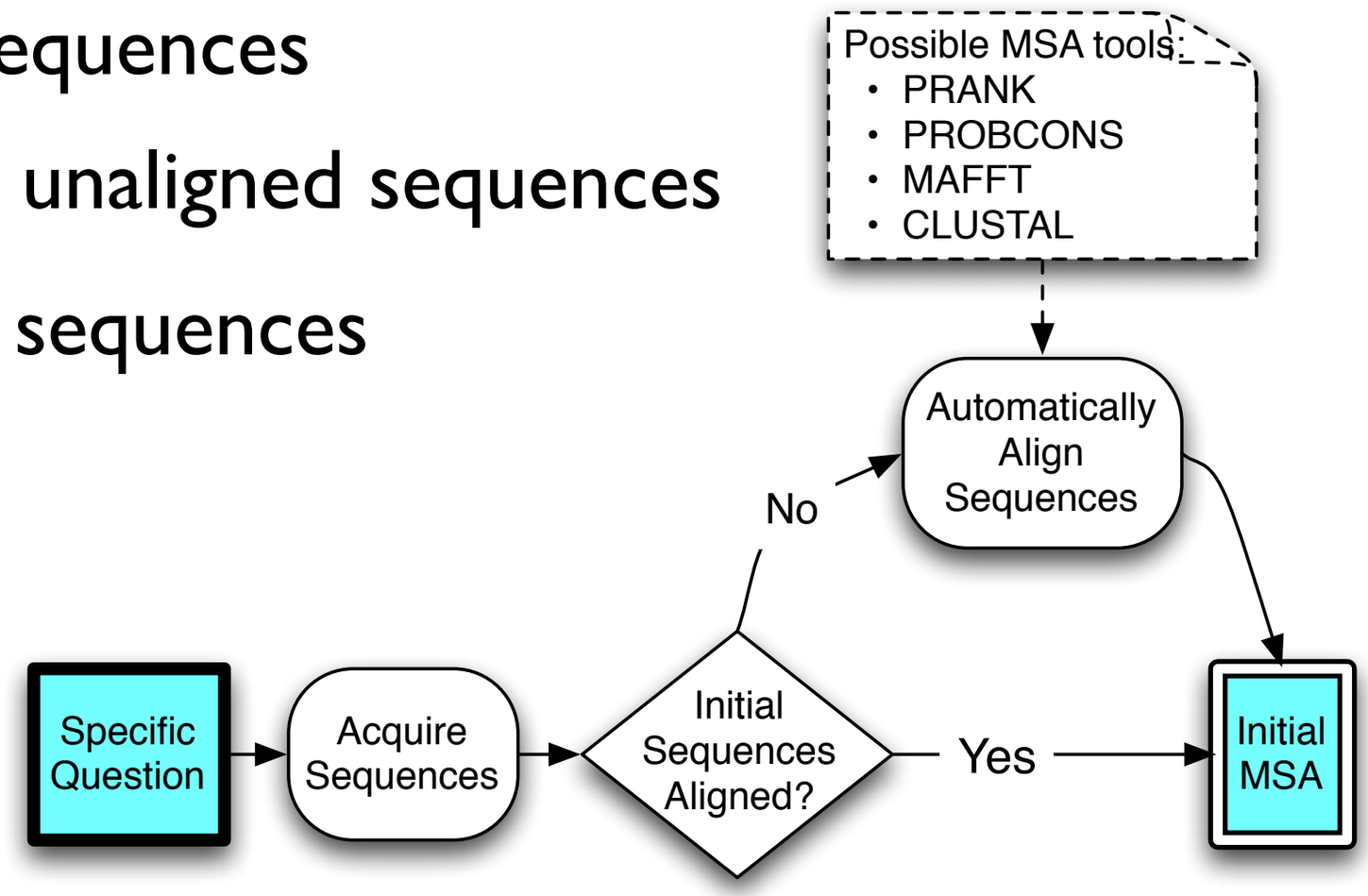


Demonstration and Exercise

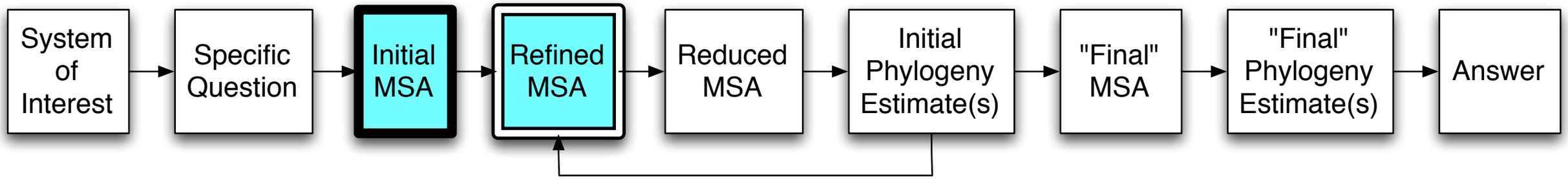
Obtaining unaligned sequences

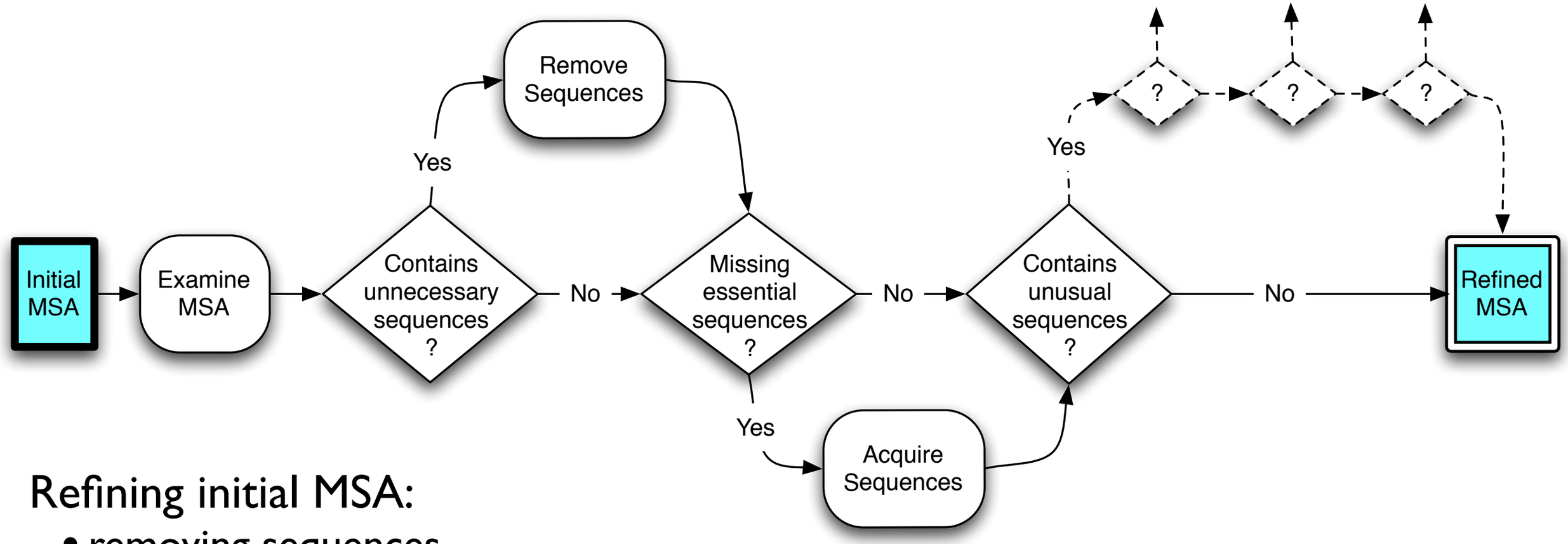
Building an MSA from unaligned sequences

Obtaining pre-aligned sequences



Refining Initial MSA: Unnecessary and Missing Sequences

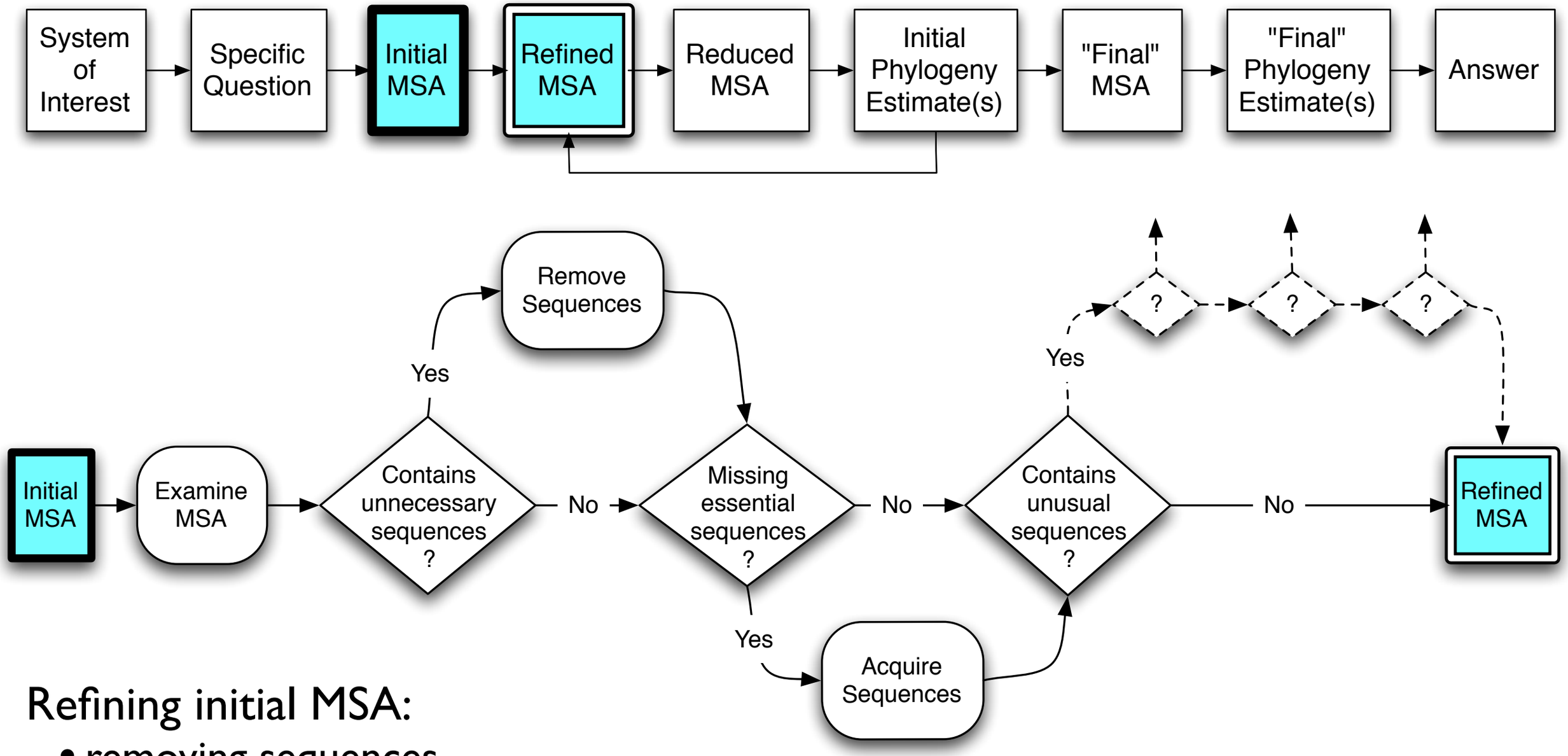




Refining initial MSA:

- removing sequences
- adding sequences
- correcting putatively mis-aligned regions

Refining Initial MSA: Unnecessary and Missing Sequences

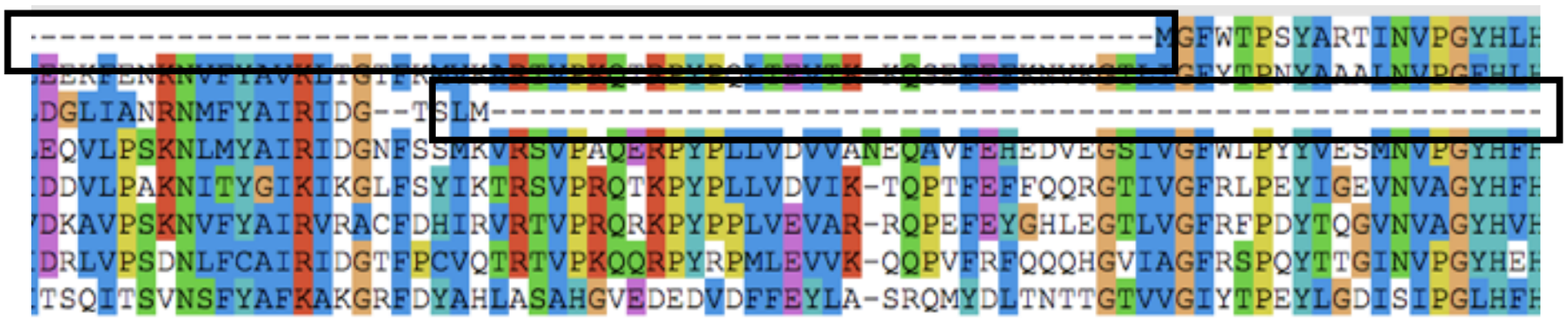


Refining initial MSA:

- removing sequences
- adding sequences
- correcting putatively mis-aligned regions

Figure 22

Unusual Sequences: Examples



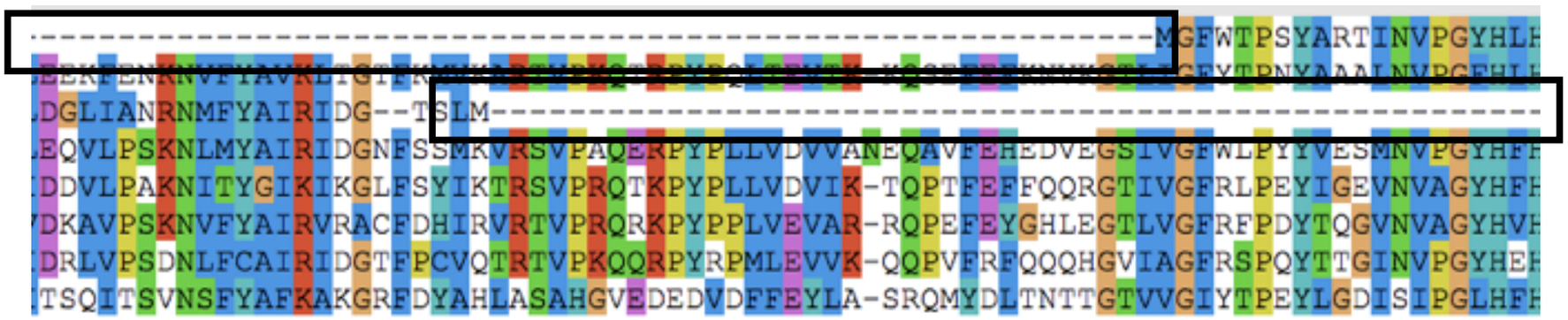
Short/fragmented sequences

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

Unusual pattern of "conservation"

Aidan Budd, EMBL Heidelberg

Unusual Sequences: Examples



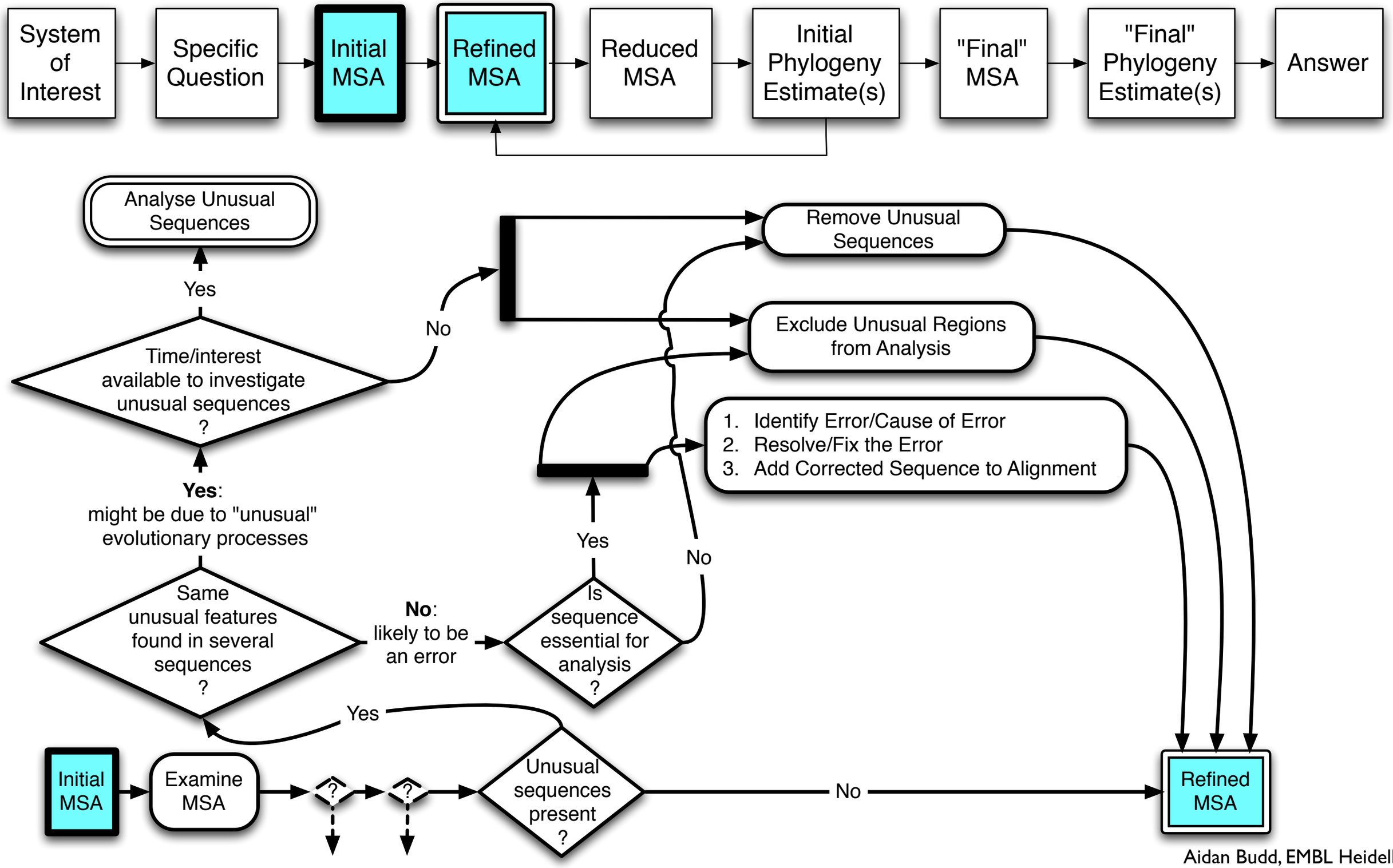
Short/fragmented sequences

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

Unusual pattern of "conservation"

Aidan Budd, EMBL Heidelberg

Refining Initial MSA: Unusual Sequences

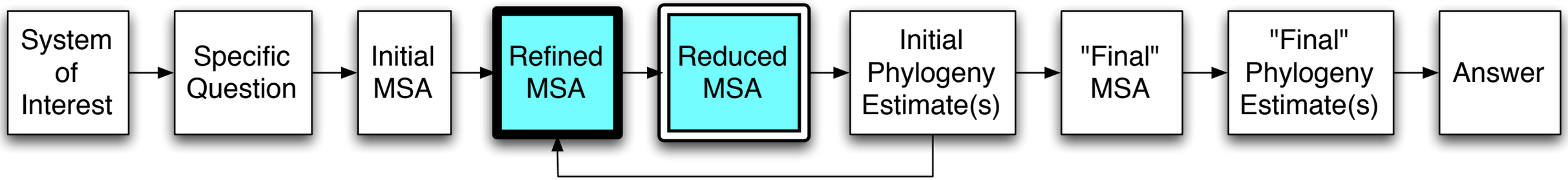


Refining Initial MSA: Unusual Sequences

Demonstration and Exercise

Refining an MSA

What is a "Reduced" MSA?



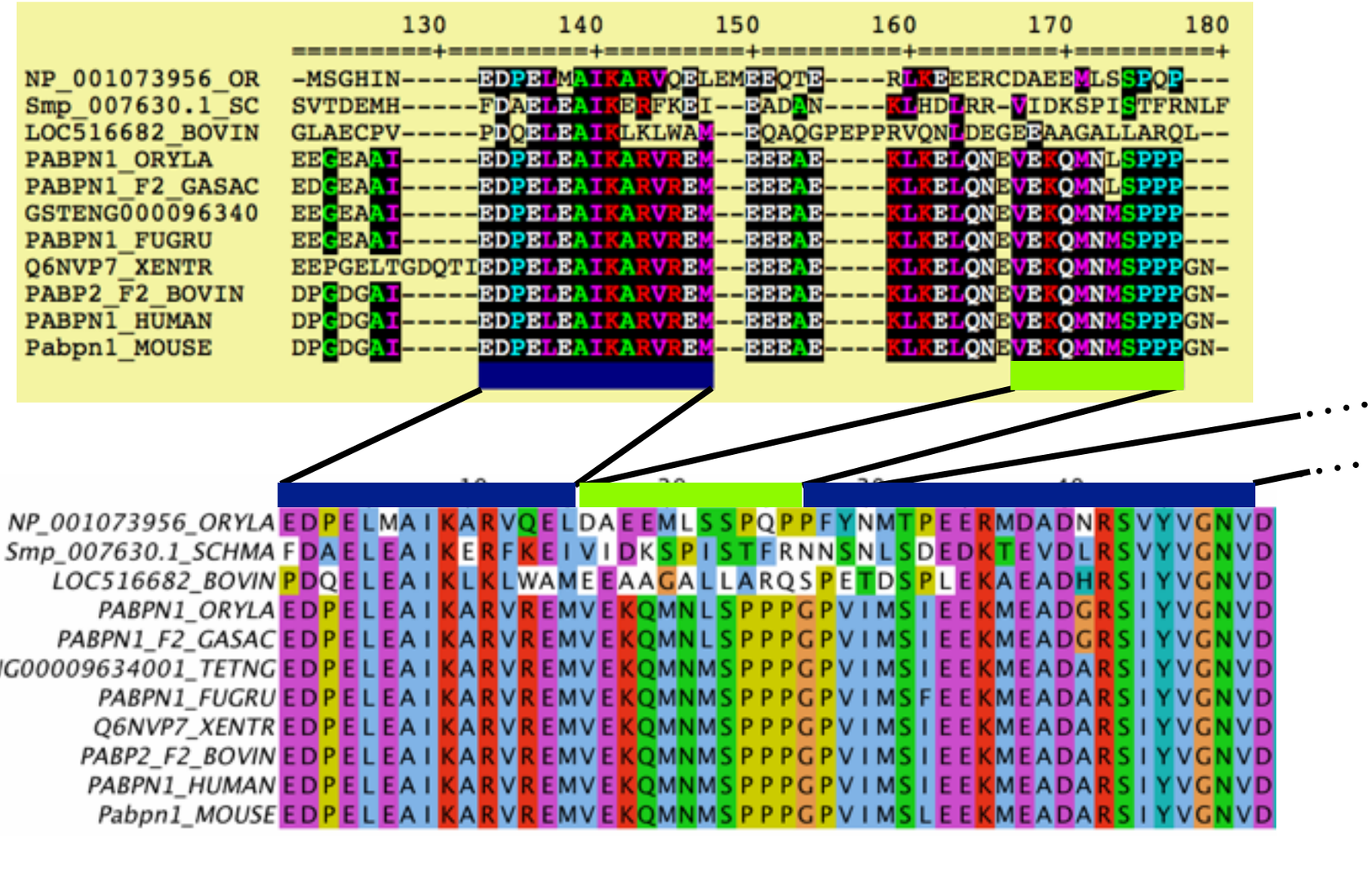
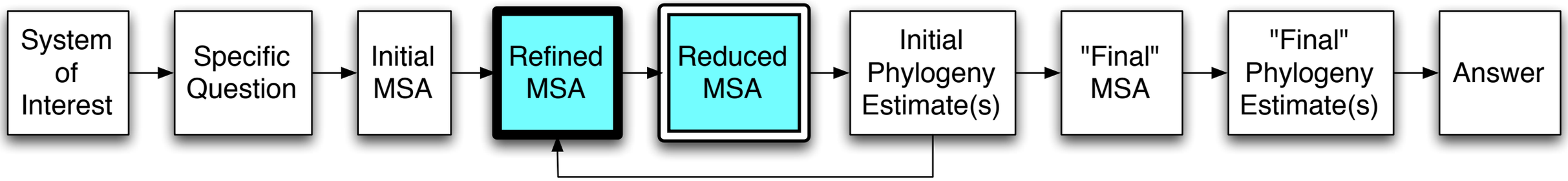
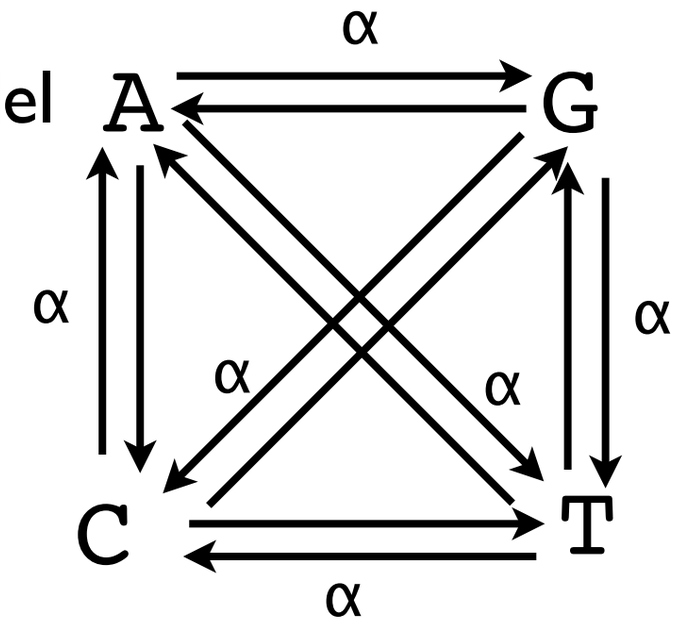


Figure 25

Why Use "Reduced" Alignments?



(Almost) all phylogeny estimation software ONLY model **point substitutions**



Analysing data (alignment columns) **related by any other process** introduces **systematic error** in the phylogeny estimate

Preparing "Reduced" Alignments

Demonstration and Exercise

Choose which columns to remove:

- "by eye" (using an alignment editor e.g. JalView)
- automatically (e.g. using GBLOCKS)

Building a "Final" MSA

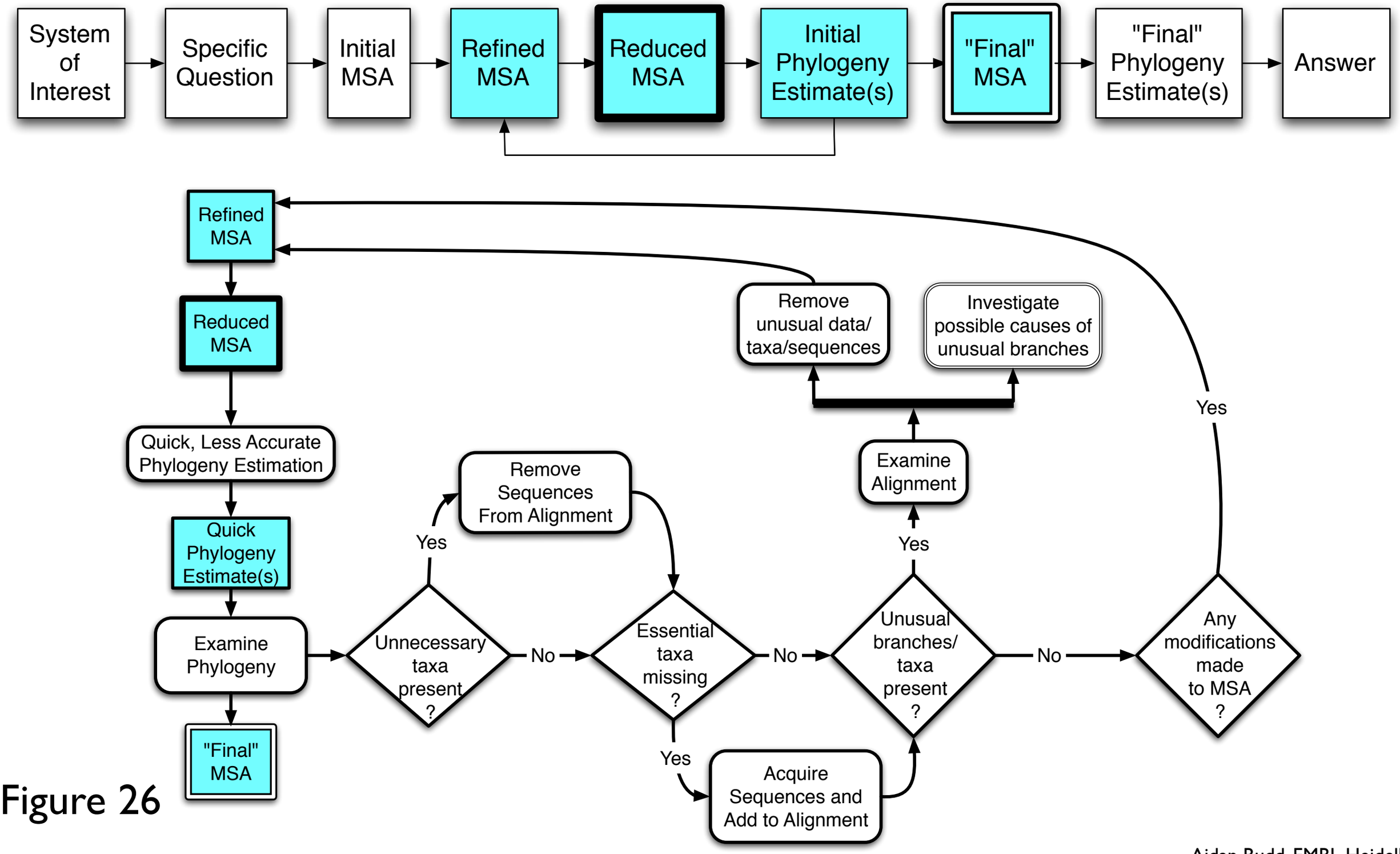
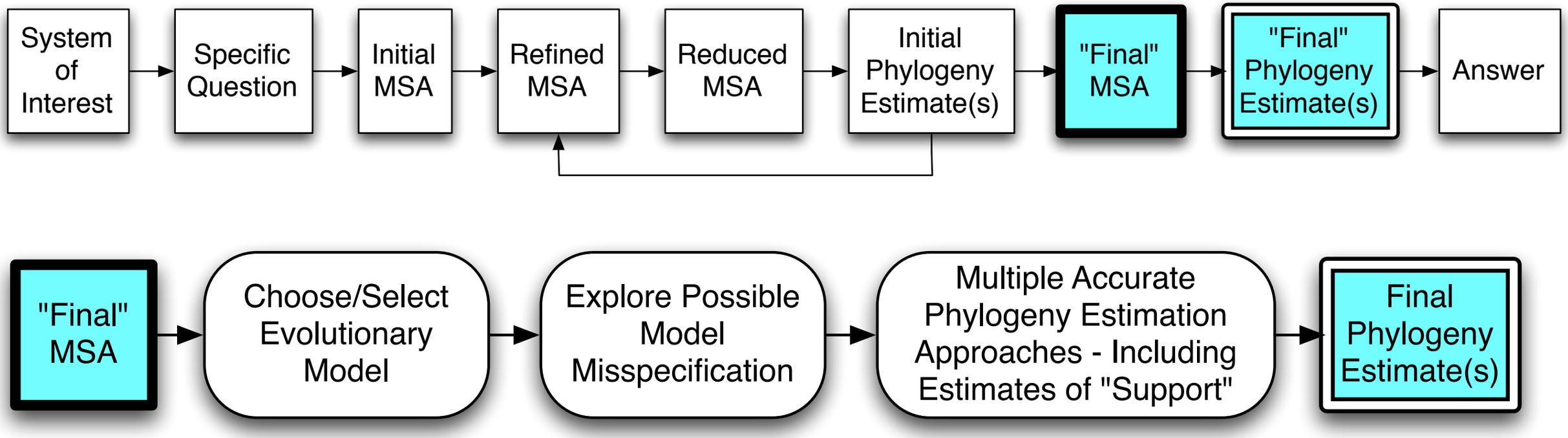


Figure 26

Final(ish) Phylogeny Estimate



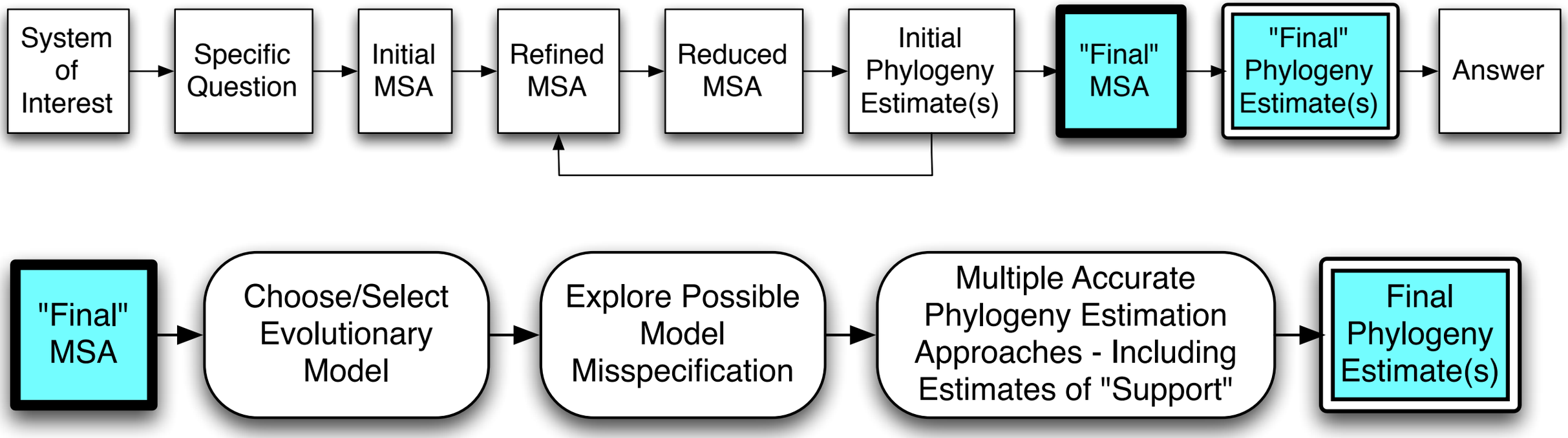
Explore different possible sources of model mis-specification

If not, either just acknowledge the diversity of possible answers, or try to understand why different analyses give different answers

Ideally they will tend to all give the same answer to the initial question

Estimate phylogeny using a wide range of different approaches, software, molecules, models, support-values

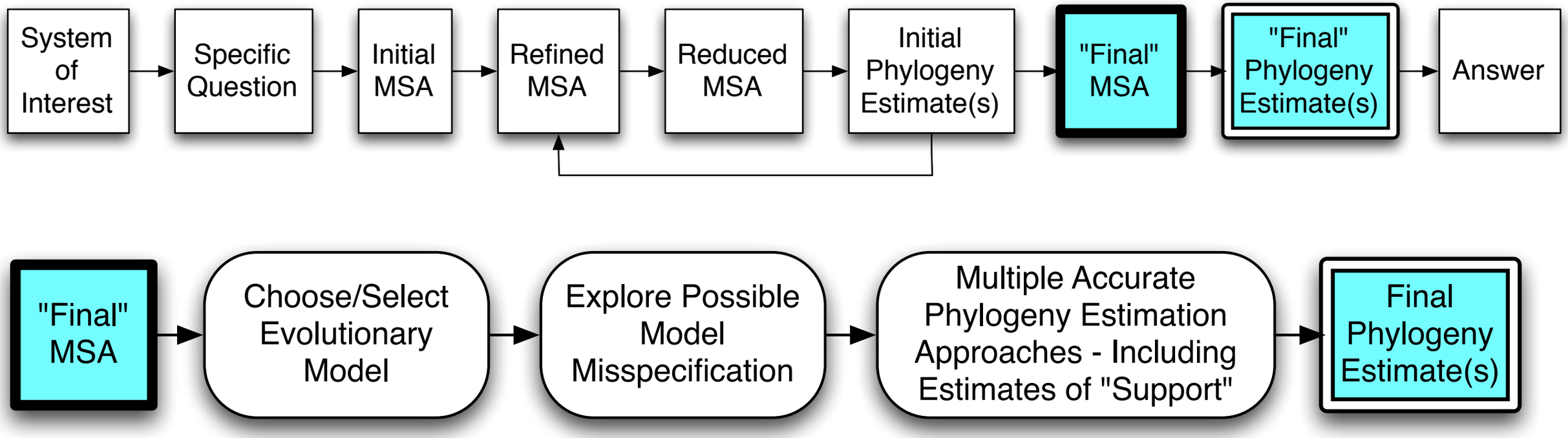
Final(ish) Phylogeny Estimate



Explore different possible sources of model mis-specification

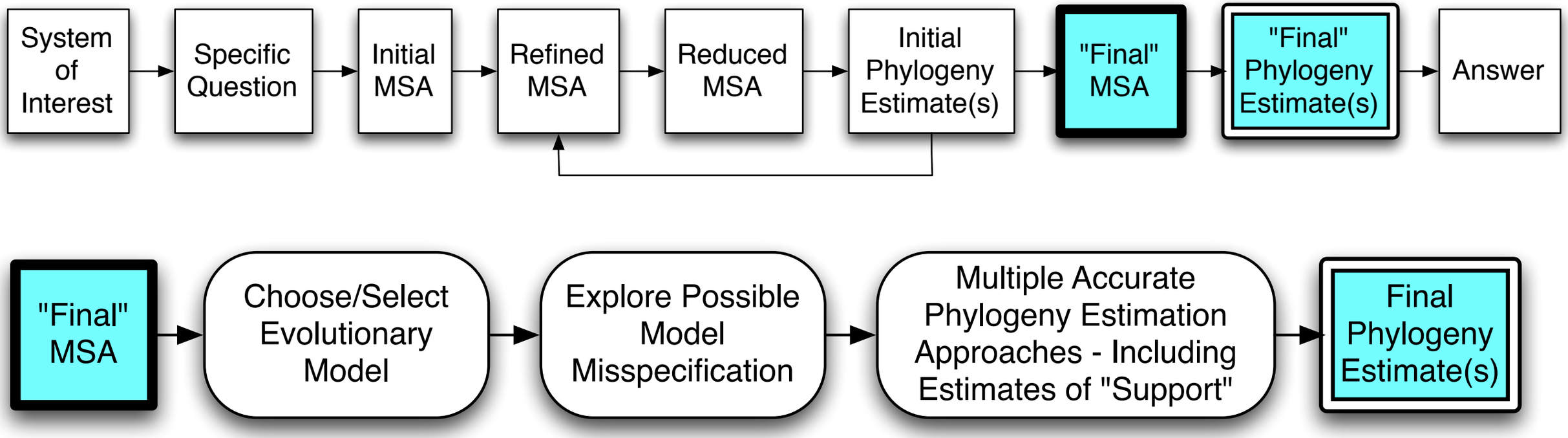
- Mis-alignment
- Recombination
- Heterotachy
- ...

Final(ish) Phylogeny Estimate



- Use several (more) accurate phylogeny estimation methods and implementations
  - Bayesian - MrBayes
  - ML - RAxML, PhyML
- Estimate using different parameter values within each implementation
  - Models
  - Specifics of tree search algorithm
  - Support values

Final(ish) Phylogeny Estimate



- Analyse phylogeny of datasets expected to have evolved under the same tree topology

  Paralogous groups of genes from the same family

  Different gene families from the same set of organisms

  Nucleotide dataset from the same gene

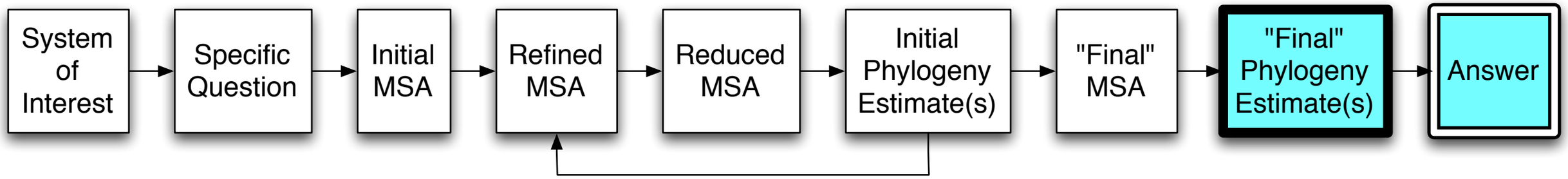
Final(ish) Phylogeny Estimate

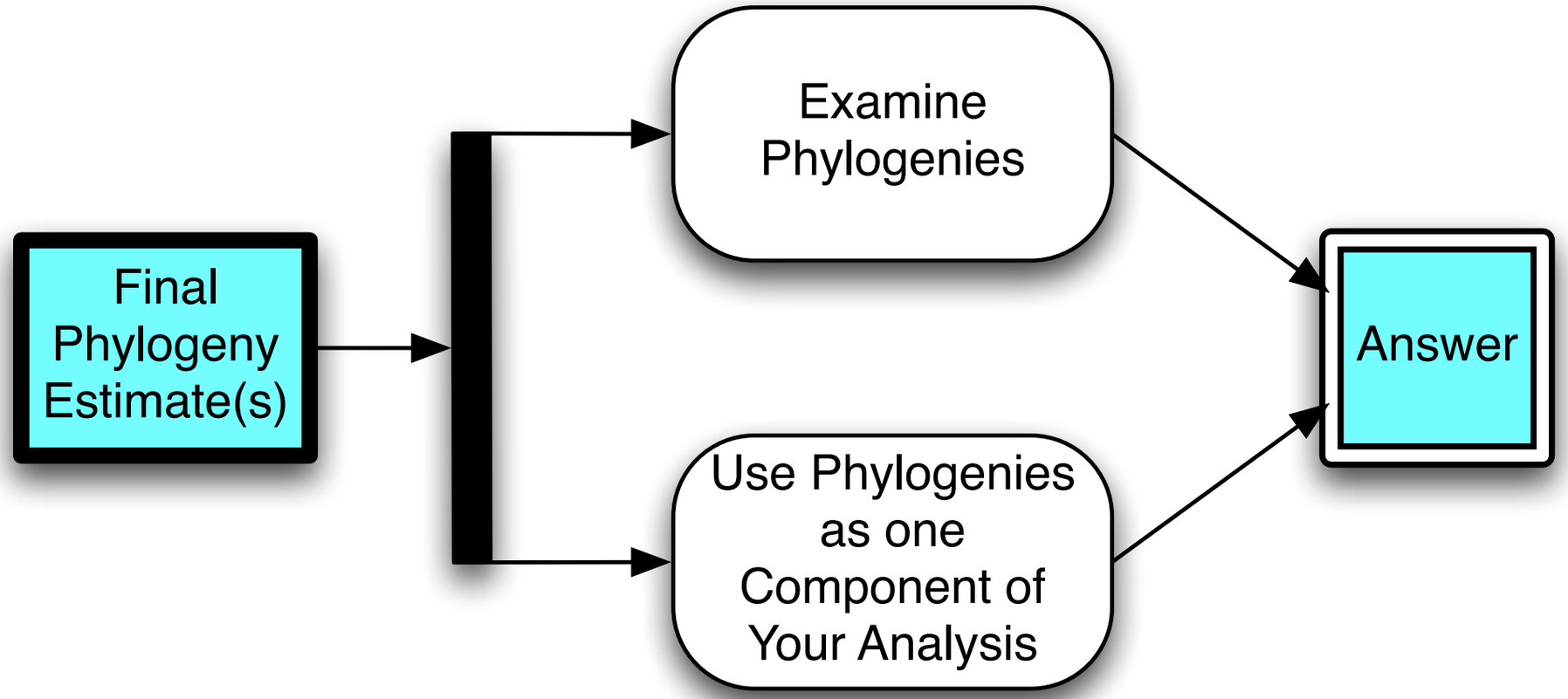
Demonstration and Exercise

Getting the final estimate:

- changing sequence names
- choose a model
- estimate phylogeny from boostrapped alignment
- maximum likelihood estimation of phylogeny
- project bootstraped trees onto ML phylogeny

Answer Your Question!





File Formats

Software only accepts data in particular format(s)

Format sometimes not very precisely specified

Common problems: **taxon labels**

- contain any characters other than A-Z and 0-9 (e.g. white space '" "', slashes '\' or'/', pipes '|' etc.)
- are not unique (e.g. two sequences labeled HumanA)
- are the wrong length (often a maximum of 10, sometimes exactly 10 is required)

File Formats

Software only accepts data in particular format(s)

Format sometimes not very precisely specified

Common problems: **sequence representation**

- if gaps allowed - wrong character used to represent them (e.g. '.' instead of '-')
- if gaps not allowed - the presence of gaps in the alignment
- sequence is of the wrong kind of molecule (DNA instead of protein etc.)
- sequence contains any characters other than the "alphabet" describing the sequences e.g. 'X' for protein alignments, 'N' for DNA alignments
- all sequences not the same length (check using JalView)

File Formats

Demonstration and Exercise

Re-formatting sequences for analysis by MrBayes

WHO FOUND THIS EASY/TRIVIAL?

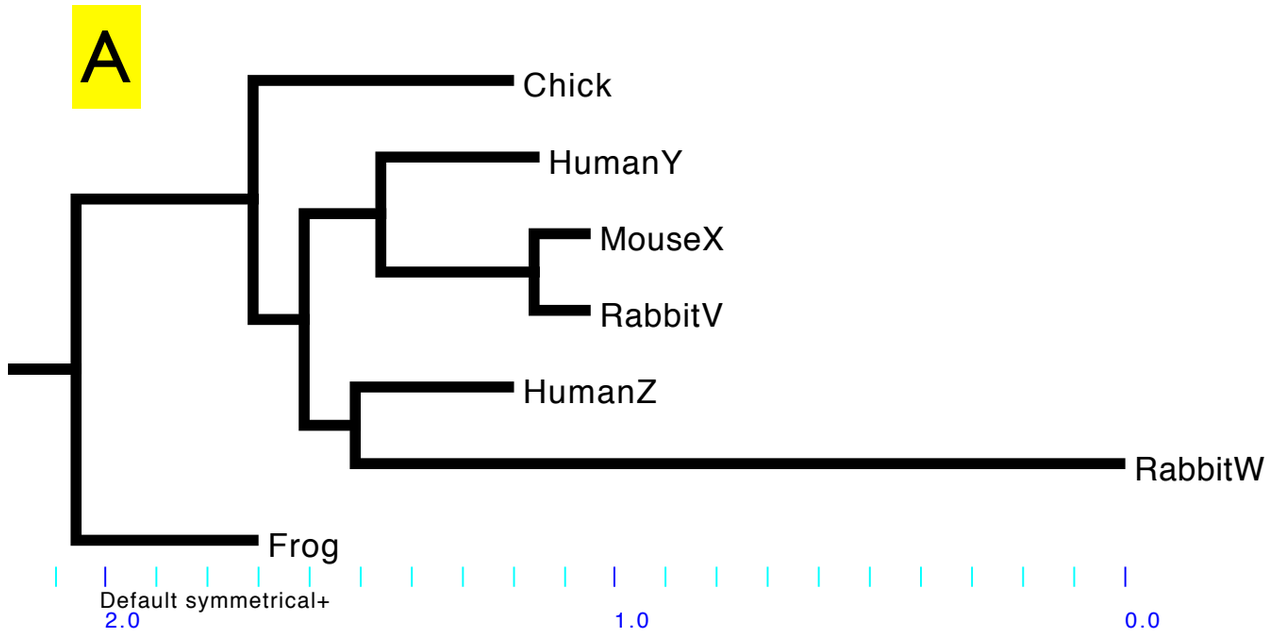
Software Execution

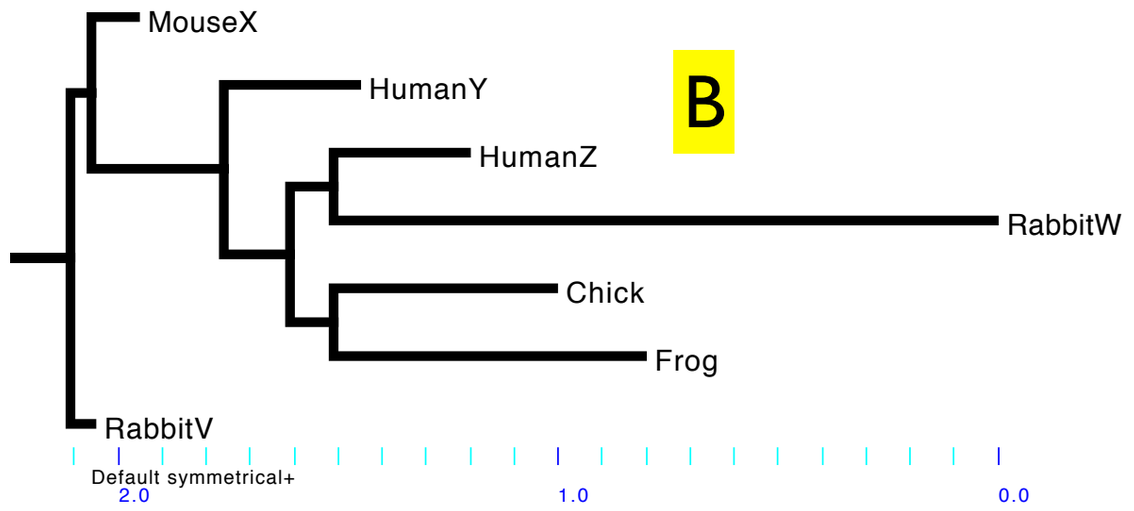
- correct order of command-line flags
- hyphen/no-hyphen
- space/no space

Rooting Trees/Reconciling Gene/ Species Trees

Aidan Budd, EMBL Heidelberg

Where does the root go?





A B and C are all from the same unrooted tree

Which rooted tree has the root in the most plausible position?