

Standards,
Precautions &
Advances in
Ancient
Metagenomics

Practical 5B: Phylogenomics

Arthur Kocher
Aida Andrades Valtueña



Overview

- Basic concepts in phylogenomics
- The start: DNA sequence alignment
- Distance-based phylogenetic methods
 - Neighbour-Joining: Mega
- Character-based phylogenetic methods
 - Maximum Parsimony method
 - Probabilistic phylogenetic methods
 - Maximum Likelihood: raxML
 - Bayesian methods: BEAST2
- Outlook



Where would you find the data for this session?

- The data is in this path: “/vol/volume/5b-phylogenomics”
 - An alignment:.snpAlignment_session5.fasta
 - A txt file with the ages of the samples: sample.ages.txt
- Use this location to also store the trees and other results during this session
- Load the environment:

```
conda activate phylogenomics-functional
```

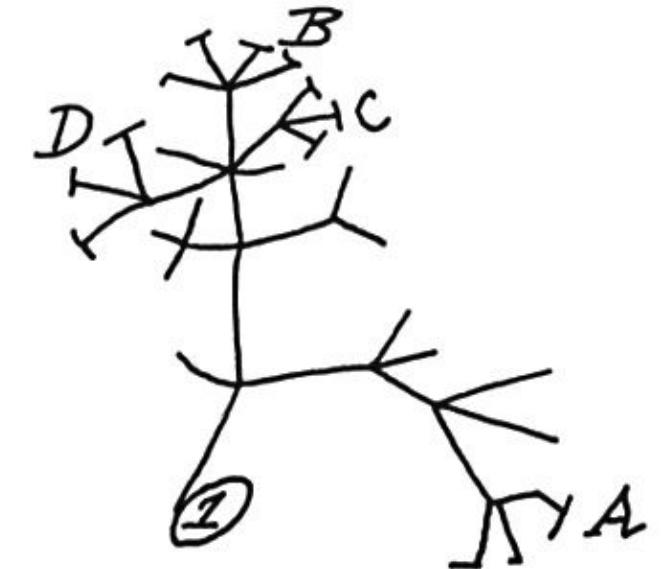


Basic concepts in phylogenomics



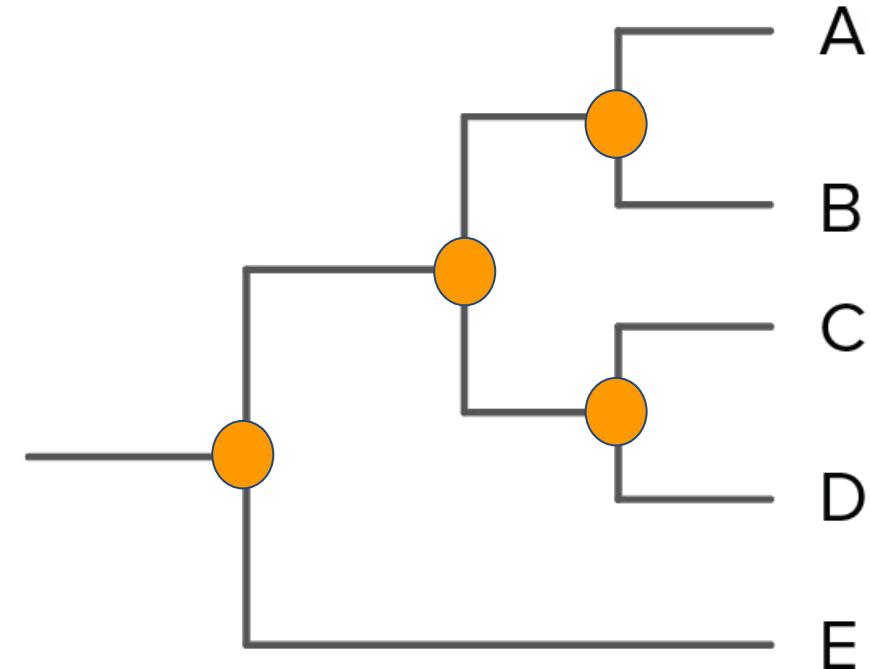
What is a phylogenetic tree?

- **Phylogenetics:** “The branch of biology that deals with phylogeny, especially with the deduction of the historical relationships between groups of organisms.” oxford dictionary
- The **relationships between organisms** are inferred based on sets of **homologous characters** (i.e. which were inherited from a common ancestor): these can be morphological or **molecular characters (like DNA sequences)**
- Evolutionary relationships are usually represented in a **phylogenetic tree**



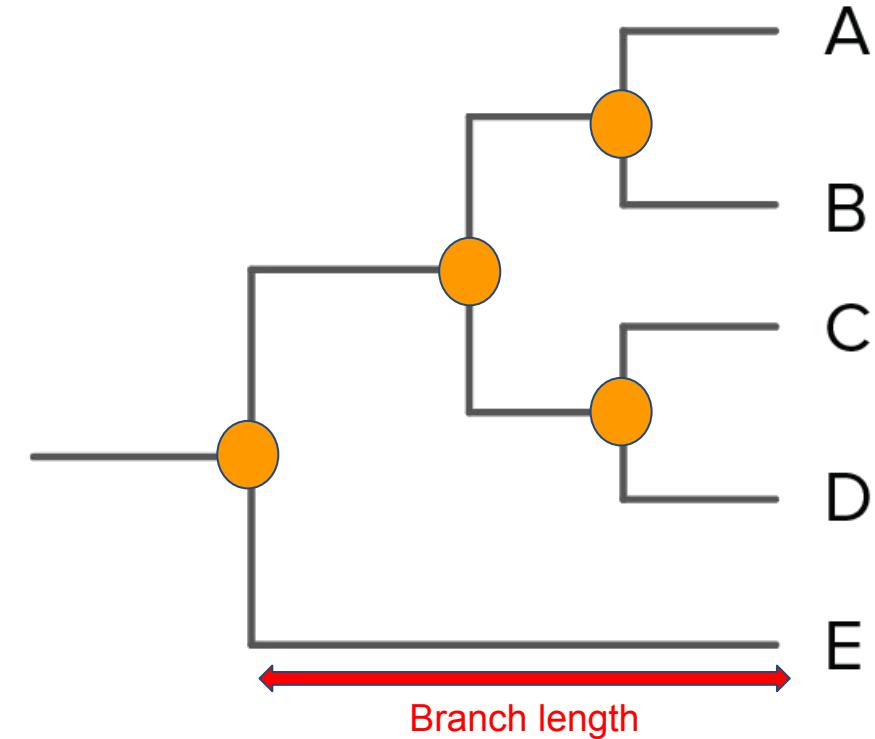
Parts of phylogenetic tree: Nodes and Branches

- **Tips/Leaves:** represent the sampled “individuals”
- **Node:** represent the ancestor shared by one or more tips
- **Branch:** it connects each node to each other and to the leaves (represent evolutionary path between nodes/leaves)

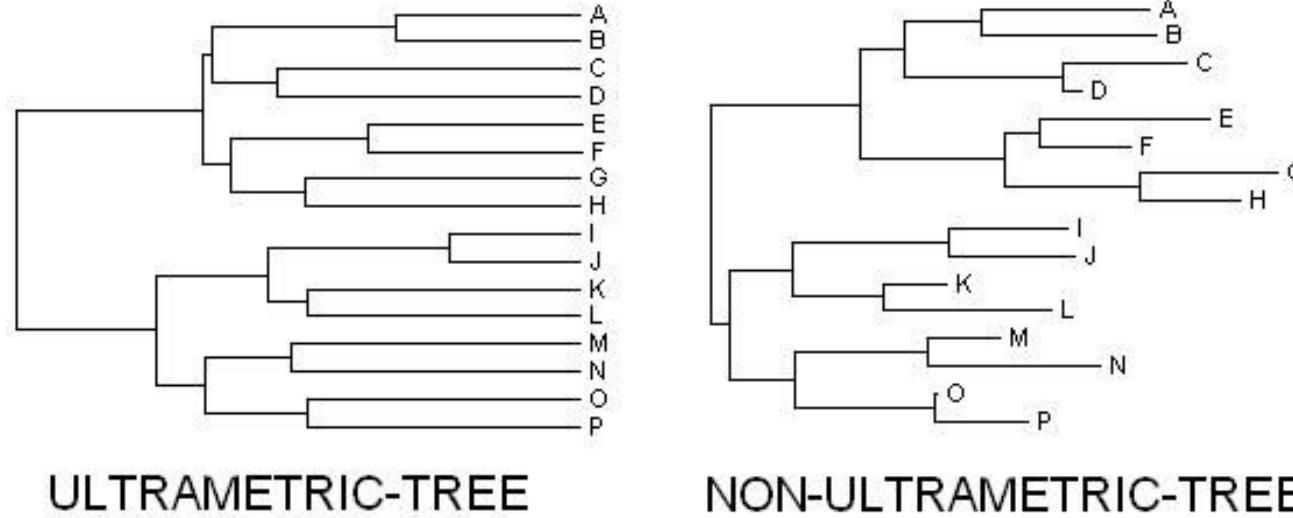


Parts of phylogenetic tree: Nodes and Branches

- **Tips/Leaves:** represent the sampled “individuals”
- **Node:** represent the ancestor shared by one or more tips
- **Branch:** it connects each node to each other and to the leaves (represent evolutionary path between nodes/leaves)
- **Branch length:** represents the number of changes, genetic/evolutionary distance, or time between two taxa or nodes



Types of trees: Ultrametric vs. non-ultrametric trees



- Ultrametric tree: the distance from the root to any leaf is the same (typically, tree with branch length in time, and all tips sampled at present)
- Non-ultrametric tree: the distance from the root to the leaves differs from leaf to leaf (typically, tree with branch lengths representing genetic distances)



The start: DNA sequence alignment



DNA sequence alignment

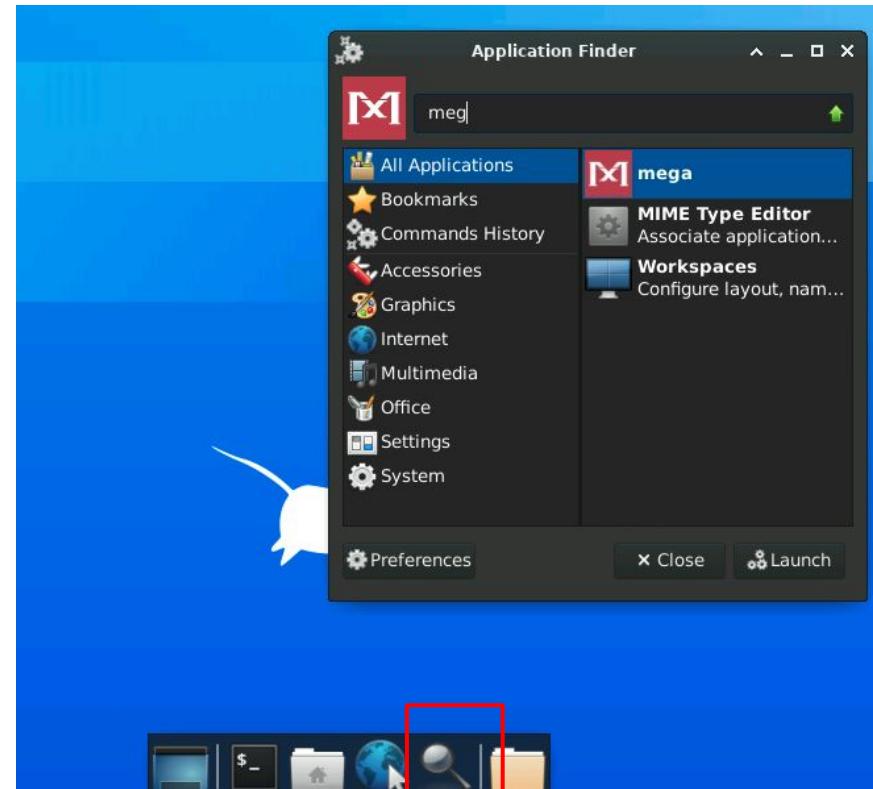
- For phylogenetic analyses, we should compare **homologous** genomic positions
- DNA sequences need to be “**aligned**”
- How can the sequence alignment be generated? This depends on your data:
 - Multi Sequence Alignment (MSA) of complete genomes
 - MAFFT
 - Clustel Omega
 - Reference based alignment: covered in the *Practical 4B: Genome mapping*

Note: for large genomic datasets, we often use single nucleotide polymorphism (SNP) alignments, i.e. alignments containing only variable genomic positions



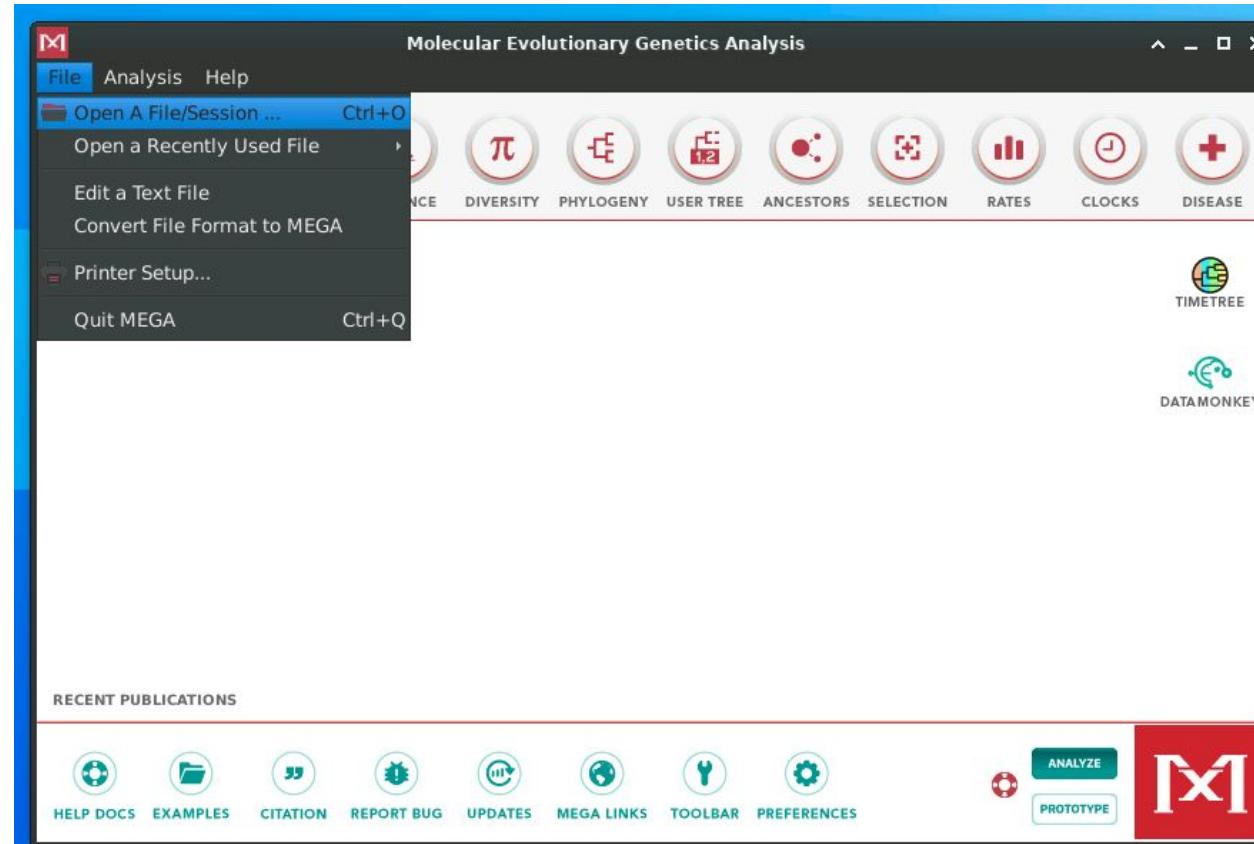
DNA sequence alignment

- Learned how to produce a SNP alignment in the Practical 4B: Genome mapping
- Let's explore the alignment in MEGA



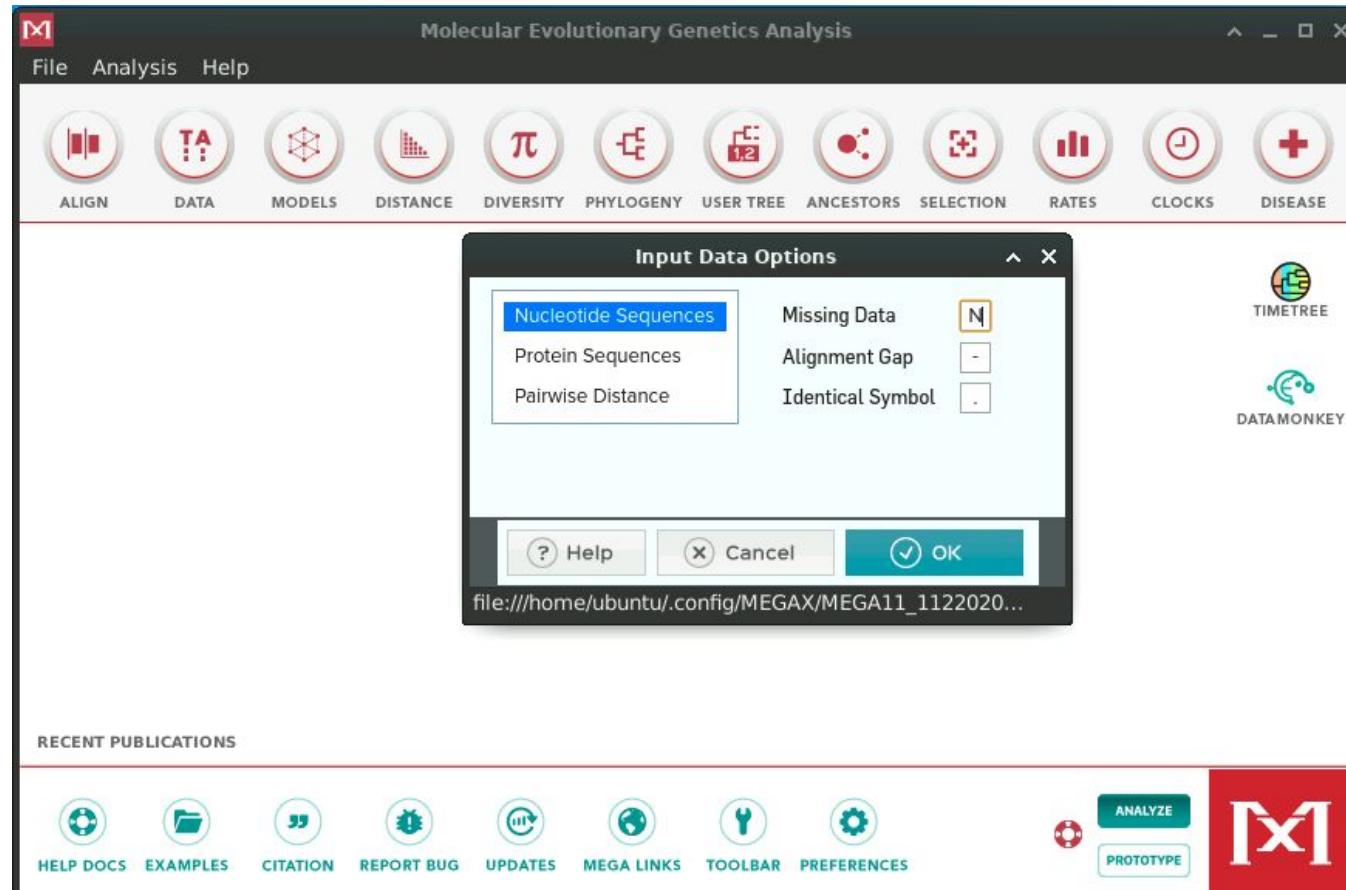
DNA sequence alignment

- File -> Open A File/Session -> Select
“/vol/volume/5b-phylogenomics/snpAlignment_session5.fasta”



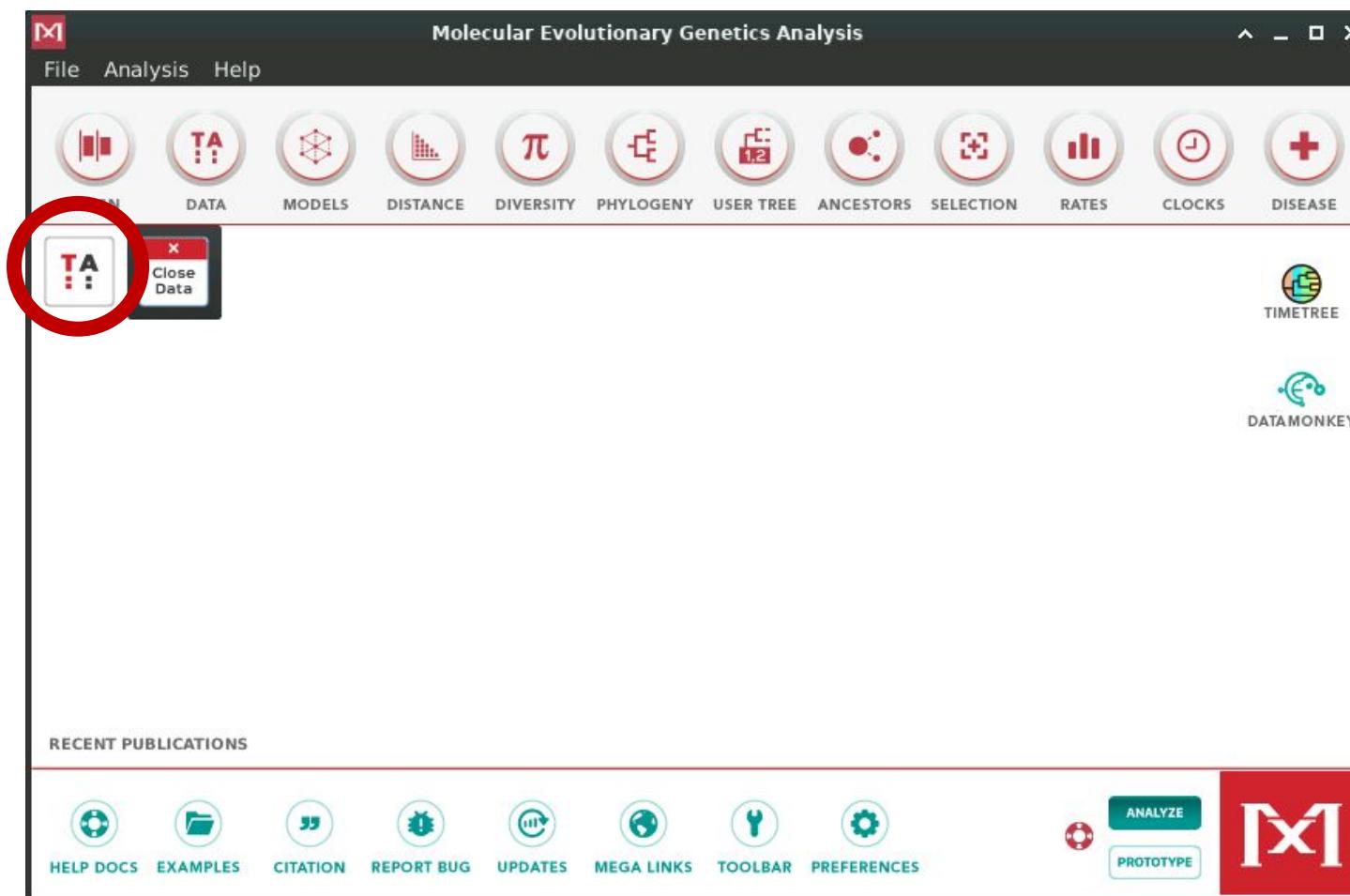
DNA sequence alignment

- Analyze -> Select “Nucleotide Sequences” change “Missing Data” by “N”
-> Click OK -> Select “No” since it is not a protein encoding sequence



DNA sequence alignment

- Open the Alignment by clicking on the TA box



DNA sequence alignment

- Alignment of *Y. pestis*
 - Our samples of interest are:
VLI092, CHC004 and KZL002 dating between 5000 and 2000 years Before Present
- What do the dots and letters represent?
- What are the Ns?
- How many sequences are we analysing?

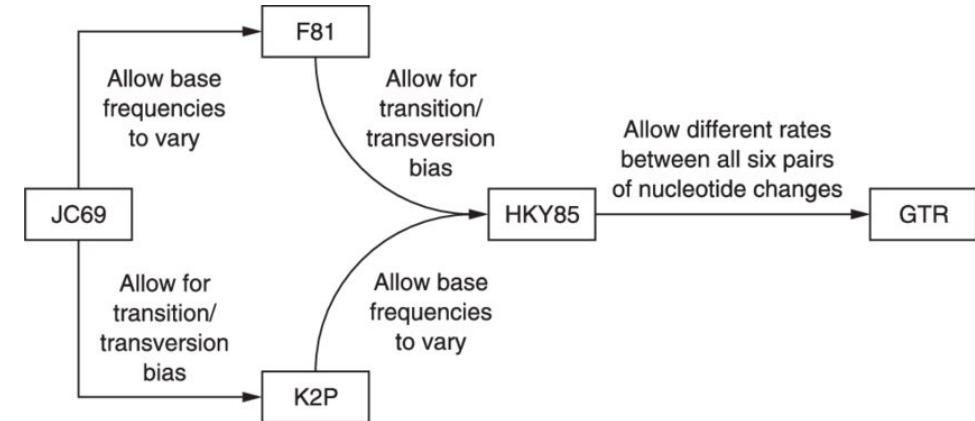


Distance-based phylogenetic methods

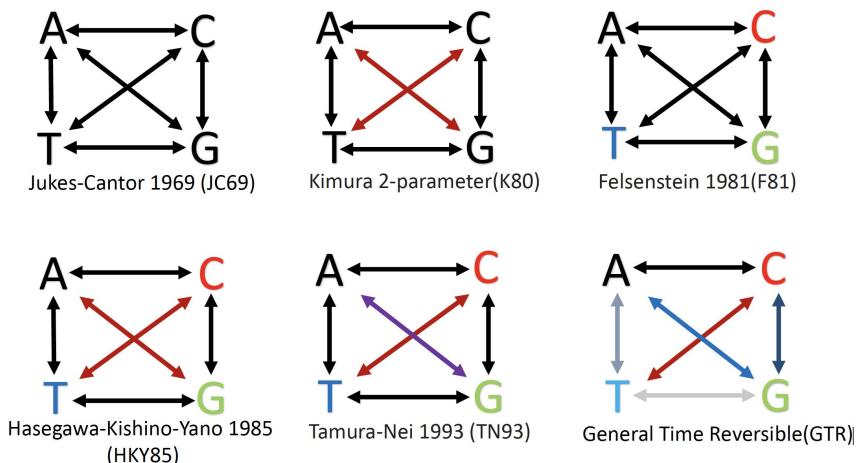


Distance-based phylogenetic methods

- Calculate the tree from a pairwise distance:
 - Number (#) of differences
 - p-distance: # differences/ total # of sites
- Substitution models:
 - evolutionary models allowing to account for multiple consecutive mutations
 - Different substitution models exists in which different types of mutations have different probabilities

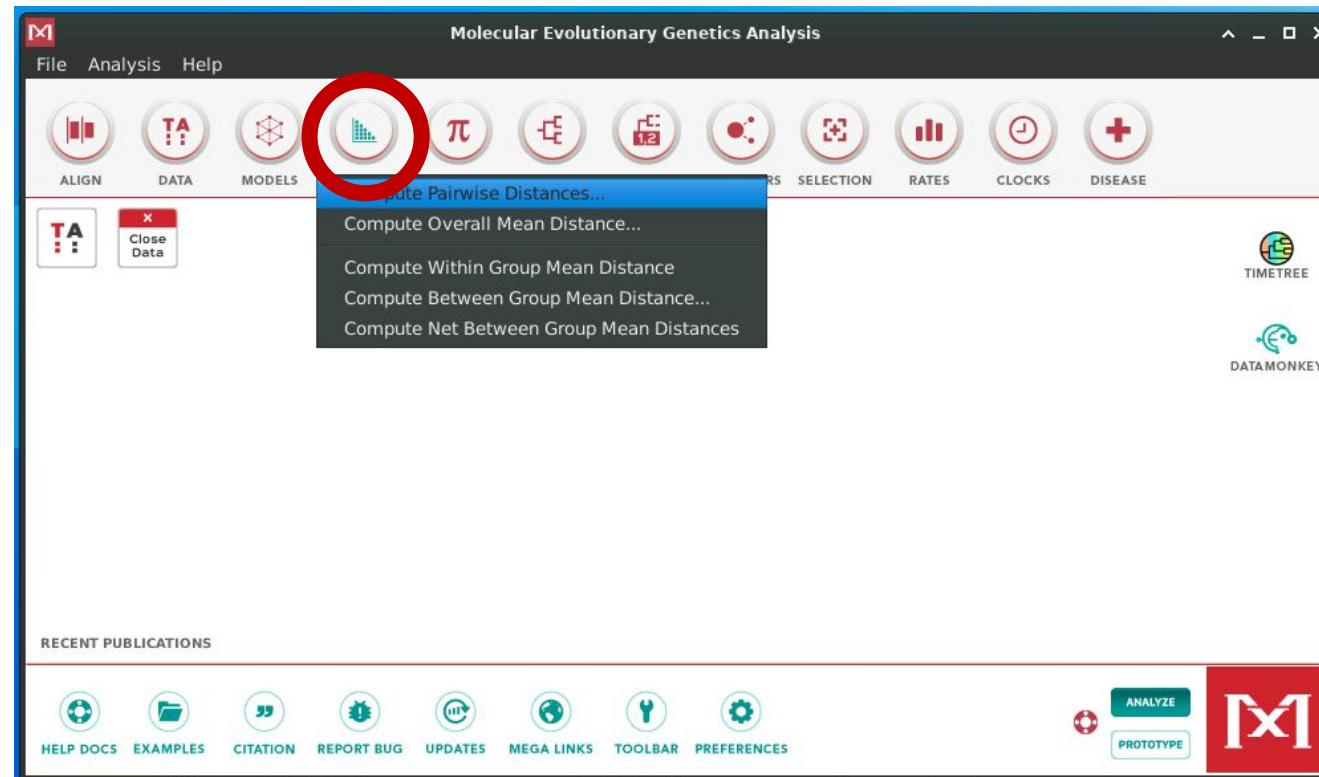


Egan A. N., Crandall K. A. (2006) Theory of Phylogenetic Estimation
 Evolutionary Genetics: Concepts and Case Studies (pp.426-443)
 Publisher: Oxford University Press
 Editors: Charles W Fox, Jason B Wolf



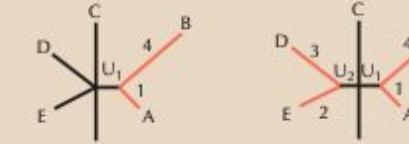
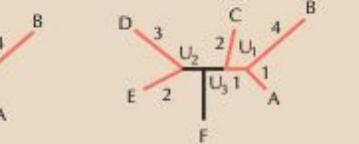
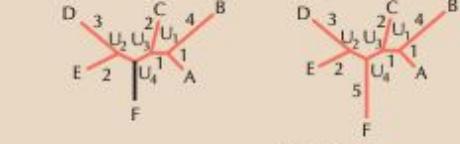
Calculate a pairwise distance matrix

- In MEGA:



Calculating a Neighbour-Joining tree

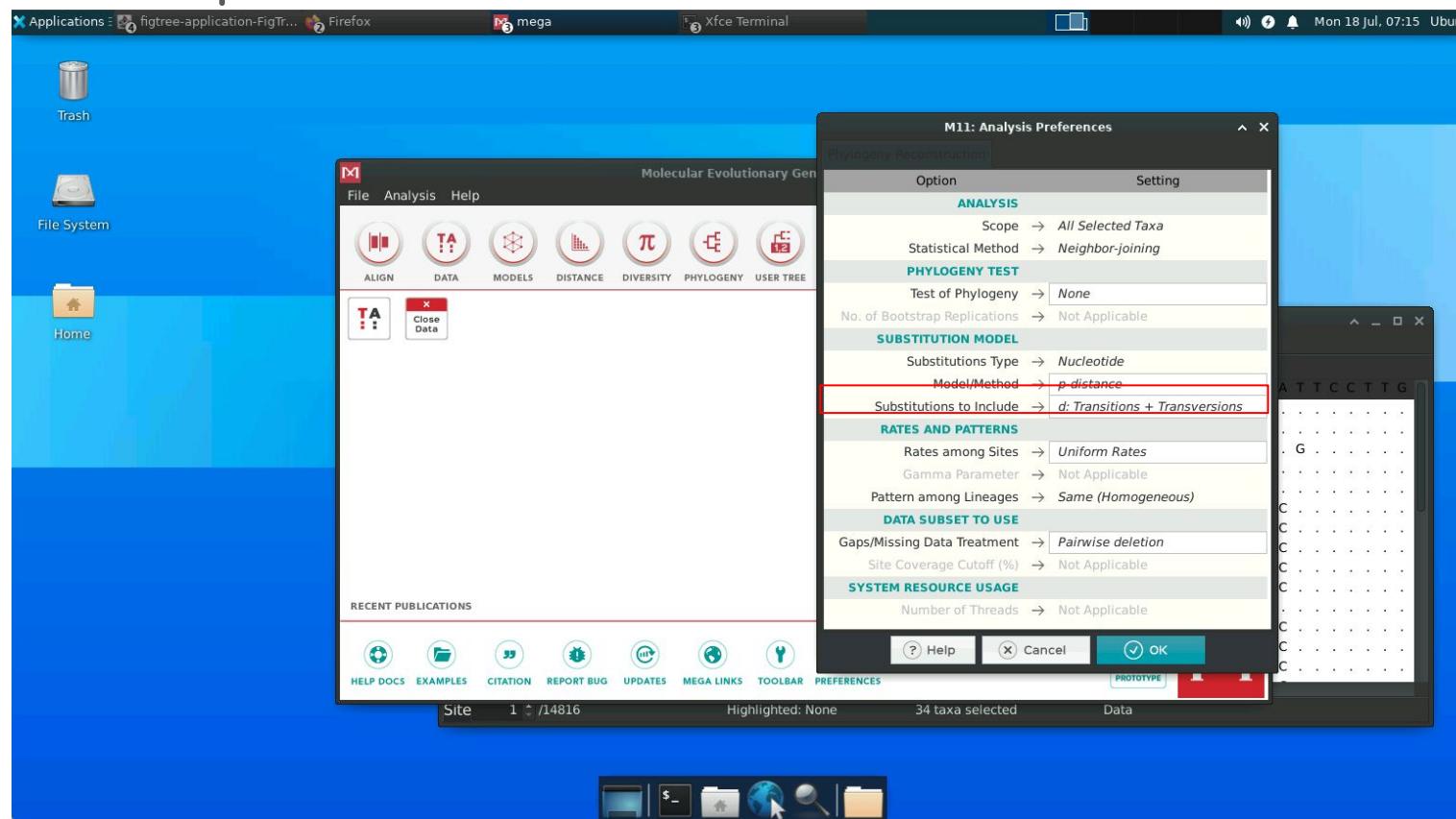
TABLE 27.11. Neighbor-joining example

	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5
Distance matrix	A B C D E B 5 C 4 D 7 10 7 E 6 9 6 5 F 8 11 8 9 8	U ₁ C D E C 3 D 6 7 E 5 6 5 F 7 8 9 8	U ₁ C U ₂ C 3 U ₂ 3 4 F 7 8 6	U ₂ U ₃ U ₃ 2 6 F 6 6	U ₄ F 5
Step 1	S _A = (5+4+7+6+8)/4 = 7.5 S _B = (5+7+10+9+11)/4 = 10.5 $S_k = \text{(sum all } D_{ij})/(N-2)$, where N is the # of OTUs in the set. S _C = (4+7+7+6+8)/4 = 8 S _D = (7+10+7+5+9)/4 = 9.5 S _E = (6+9+6+5+8)/4 = 8.5 S _F = (8+11+8+9+8)/4 = 11	S _{U₁} = (3+6+5+7)/3 = 7 S _C = (3+7+6+8)/3 = 8 S _D = (6+7+5+9)/3 = 9 S _E = (5+6+5+8)/3 = 8 S _F = (7+8+9+8)/3 = 10.6	S _{U₁} = (3+3+7)/2 = 6.5 S _C = (3+4+8)/2 = 7.5 S _{U₂} = (3+4+6)/2 = 6.5 S _E = (7+8+6)/2 = 10.5	S _{U₂} = (2+6)/1 = 8 S _{U₃} = (2+6)/1 = 8 S _E = (6+6)/1 = 12	Because N - 2 = 0, we cannot do this calculation.
Step 2	Calculate pair with smallest M_{ij} , where $M_{ij} = D_{ij} - S_i - S_j$. Smallest are M _{AB} = 5 - 7.5 - 10.5 = -13 M _{DE} = 5 - 9.5 - 8.5 = -13 Choose one of these (AB here).	Smallest is M _{CU₁} = 3 - 7 - 8 = -12 M _{DE} = 5 - 9 - 8 = -12 Choose one of these (DE here).	Smallest is M _{CU₁} = 3 - 6.5 - 7.5 = -11 Choose one of these (DE here).	Smallest is M _{U₂F} = 6 - 8 - 12 = -14 M _{U₃F} = 6 - 8 - 12 = -14 M _{U₂U₃} = 2 - 8 - 8 = -14 Choose one of these (M _{U₂U₃} here).	
Step 3	Create a node (U) that joins pair with lowest M_{ij} such that $S_{R,U} = D_{ij}/2 + (S_i - S_j)/2$. U ₁ joins A and B: S _{AU₁} = D _{AB} /2 + (S _A - S _B)/2 = 1 S _{BU₁} = D _{AB} /2 + (S _B - S _A)/2 = 4	U ₂ joins D and E: S _{DU₂} = D _{DE} /2 + (S _D - S _E)/2 = 3 S _{EU₂} = D _{DE} /2 + (S _E - S _D)/2 = 2	U ₃ joins C and U ₁ : S _{CU₃} = D _{CU₁} /2 + (S _C - S _{U₁})/2 = 2 S _{U₃U₁} = D _{CU₁} /2 + (S _{U₁} - S _C)/2 = 1	U ₄ joins U ₂ and U ₃ : S _{U₂U₄} = D _{U₂U₃} /2 + (S _{U₂} - S _{U₃})/2 = 1 S _{U₃U₄} = D _{U₂U₃} /2 + (S _{U₃} - S _{U₂})/2 = 1	For last pair, connect U ₄ and F with branch length = 5.
Step 4	Join i and j according to S above and make all other taxa in form of a star. Branches in black are of unknown length. Branches in red are of known length.				Comments Note this is the same tree we started with (drawn in unrooted form here).
Step 5	Calculate new distance matrix of all other taxa to U with $D_{iU} = D_{ix} + D_{jx} - D_{ij}$, where i and j are those selected from above.				
From http://www.icp.ucl.ac.be/~opperd/private/upgma.html .					



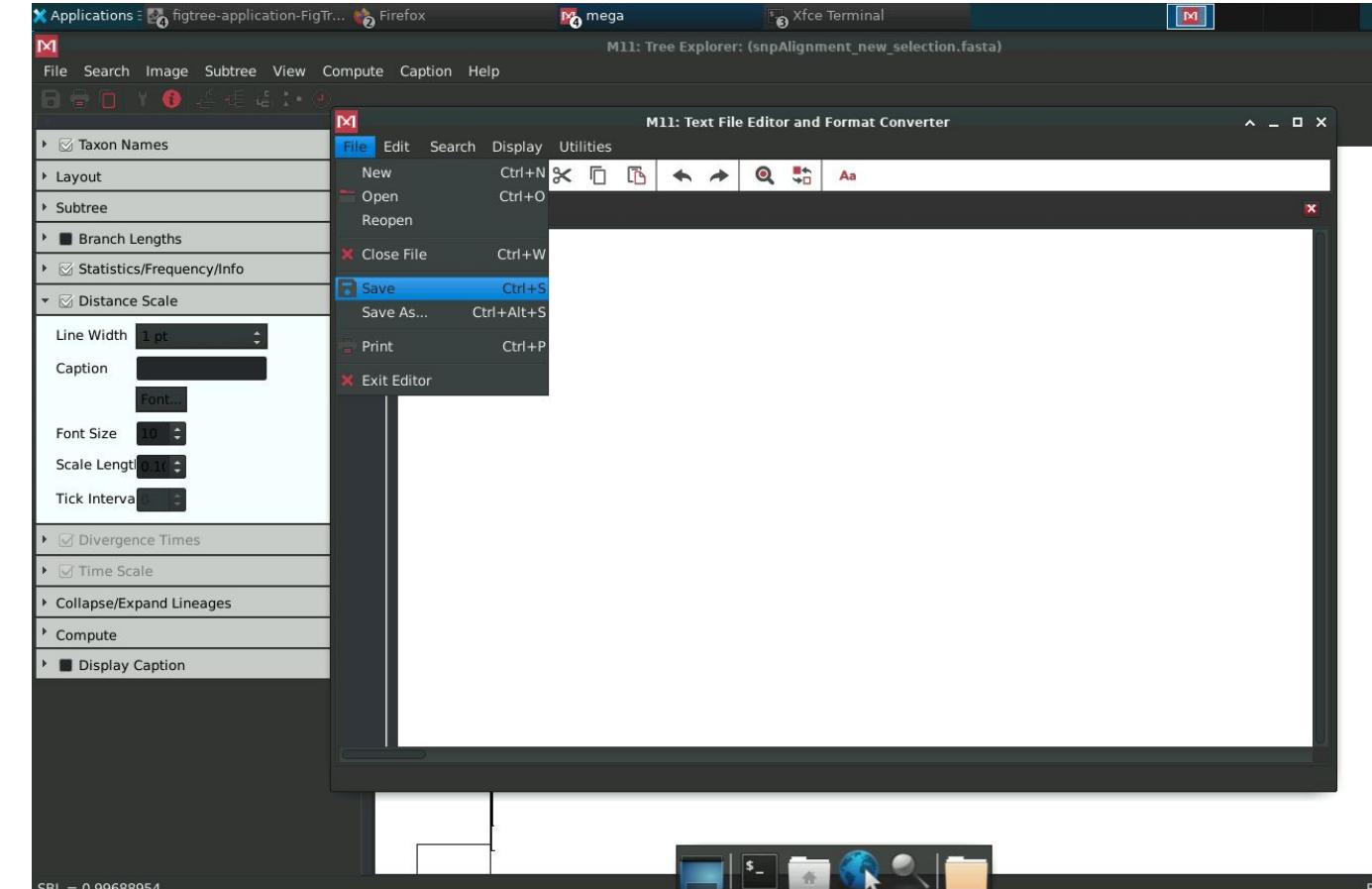
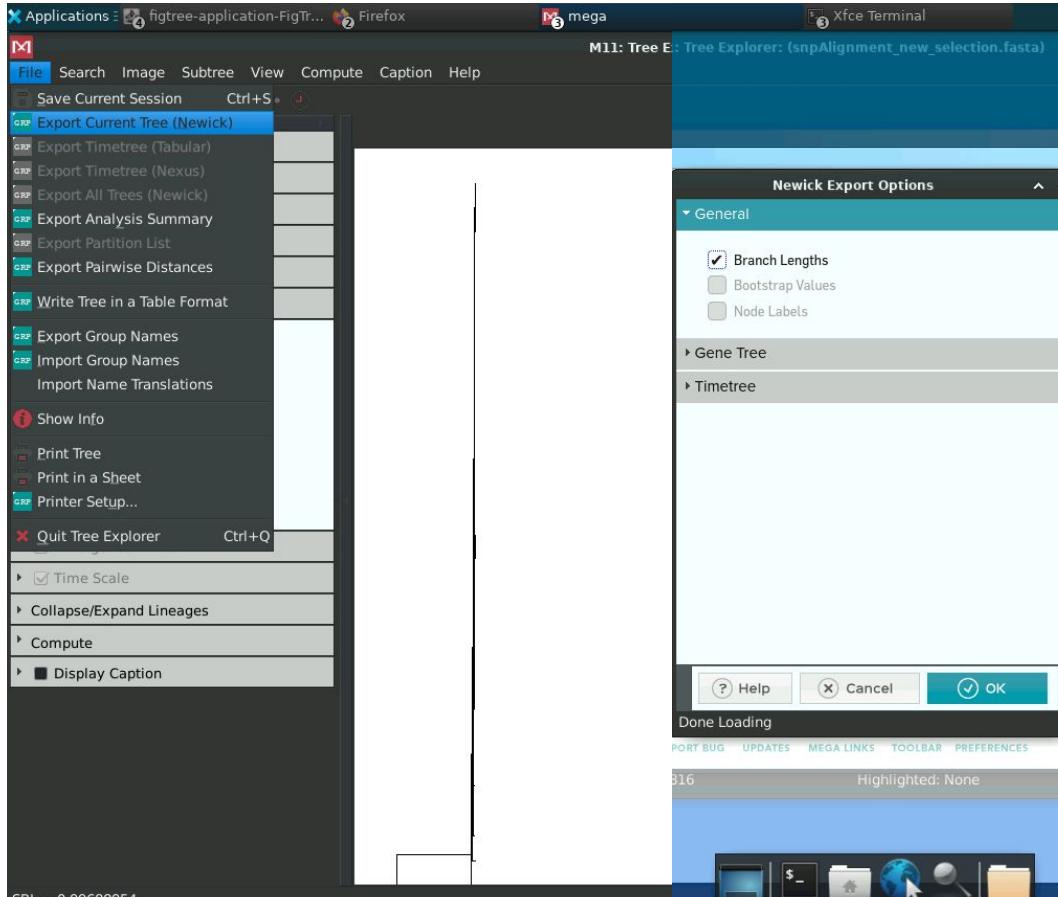
Let's make our own NJ tree!

- In MEGA: Phylogeny -> Construct Neighbour Joining Tree ->
Model/Method: “p-distance” -> OK!



Let's make our own NJ tree!

- In MEGA: File > Export current tree (Newick) -> click on "branch lengths" and save as NJ_tree.nwk



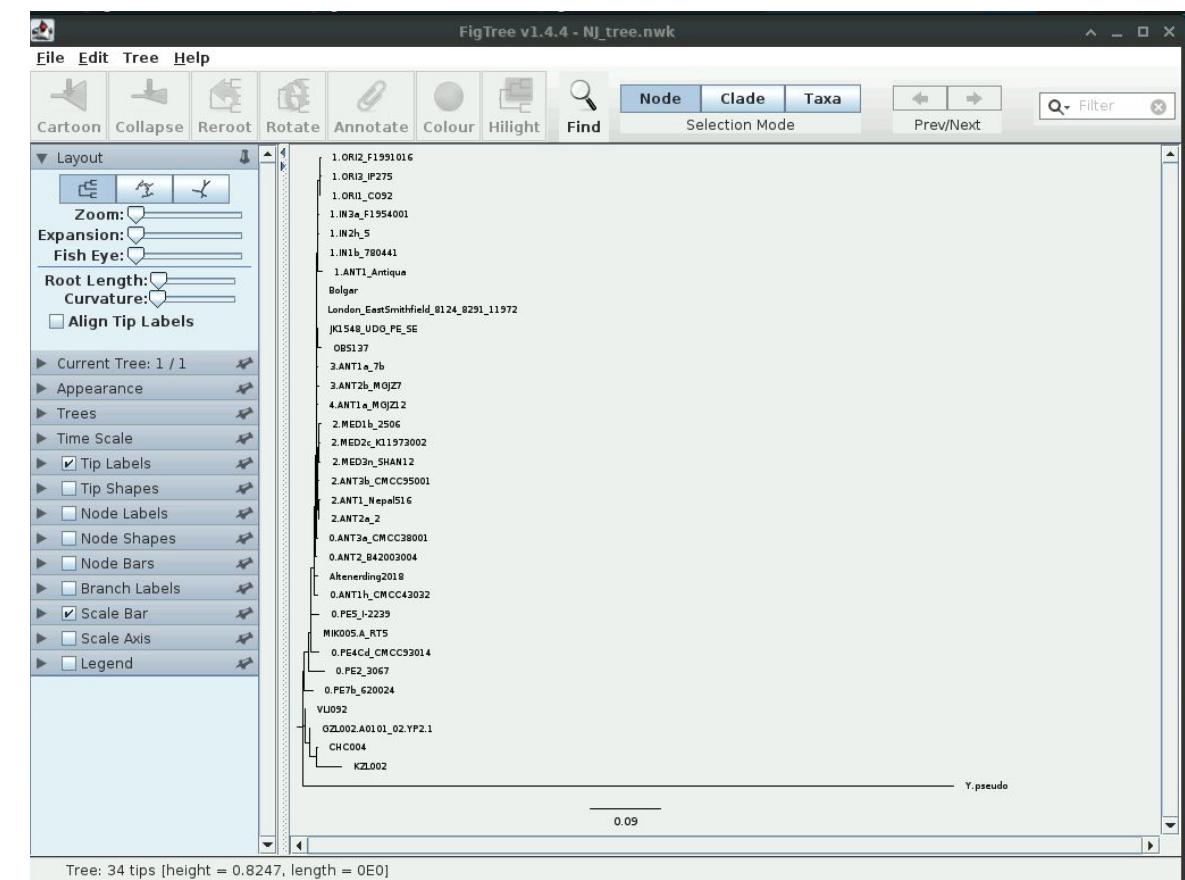
Let's make our own NJ tree!

- Explore in FigTree: type `figtree` in the terminal
- File -> Open -> select “NJ_tree.nwk”

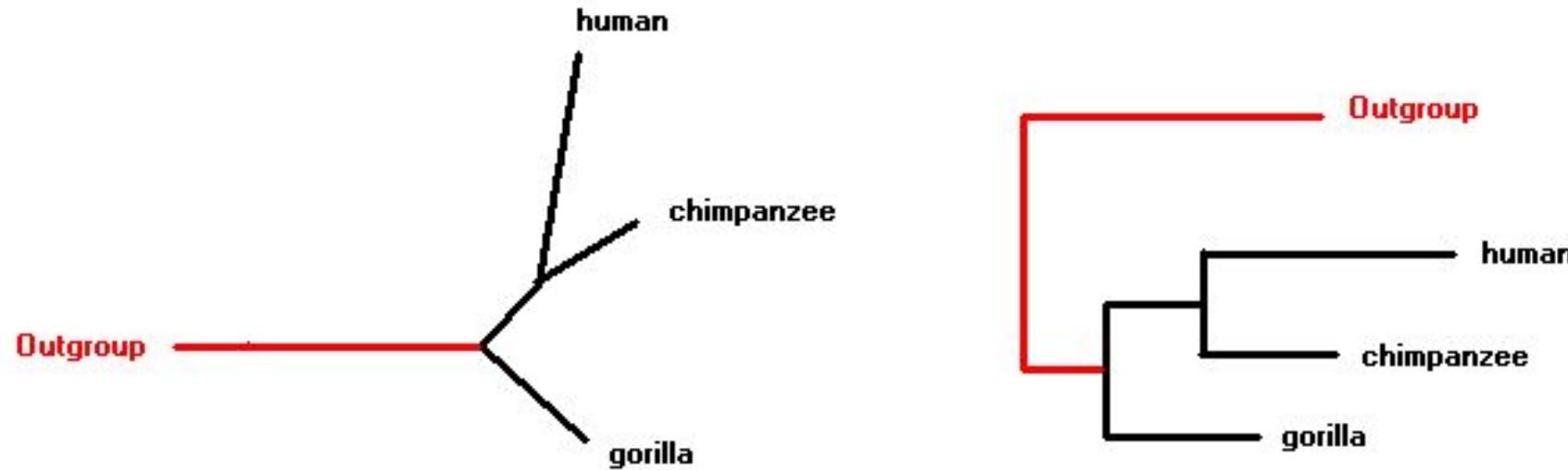


Let's make our own NJ tree!

- Explore in FigTree: type figtree in the terminal
- File -> Open -> select “NJ_tree.nwk”
- Which type of tree is this?



Types of trees: Rooted vs. Unrooted trees

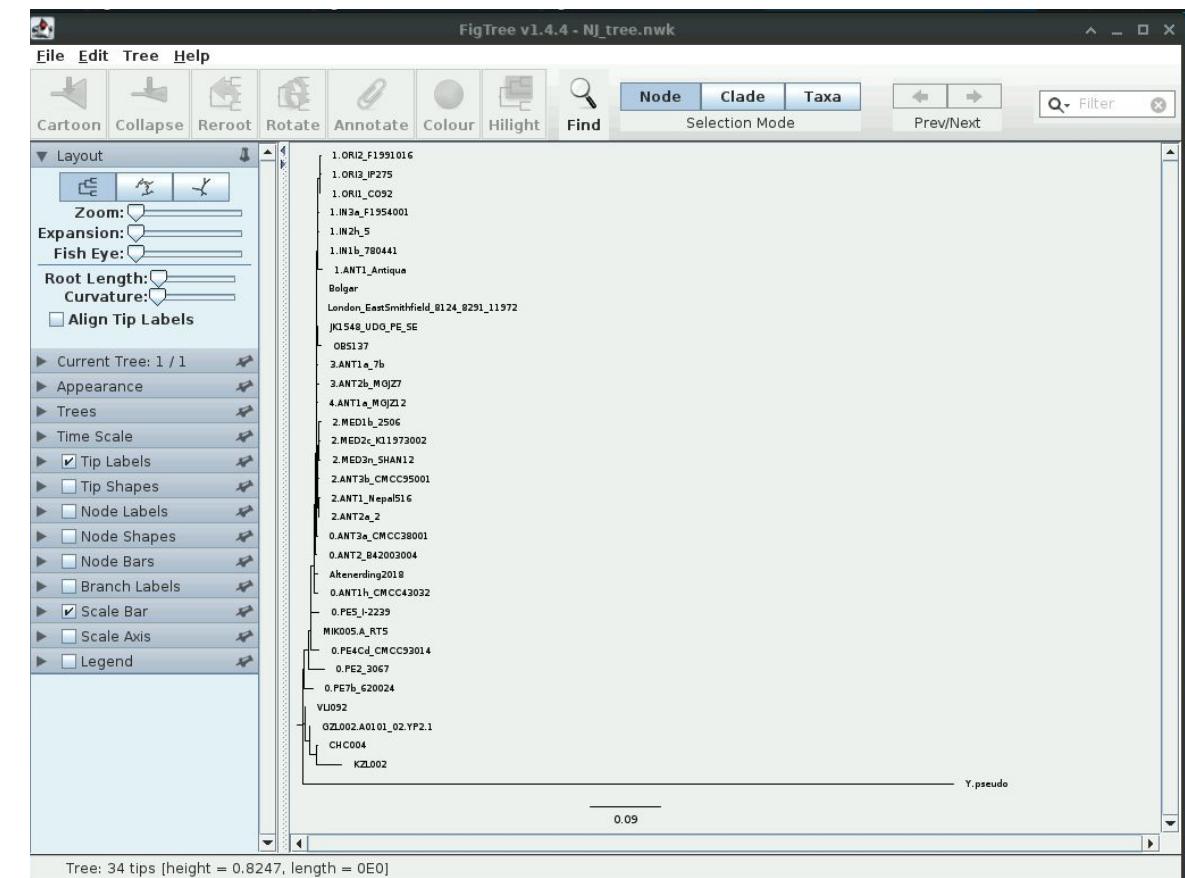


- Unrooted: represent the relationships but not the direction in time
- Rooted: represents the relationships and the direction in time



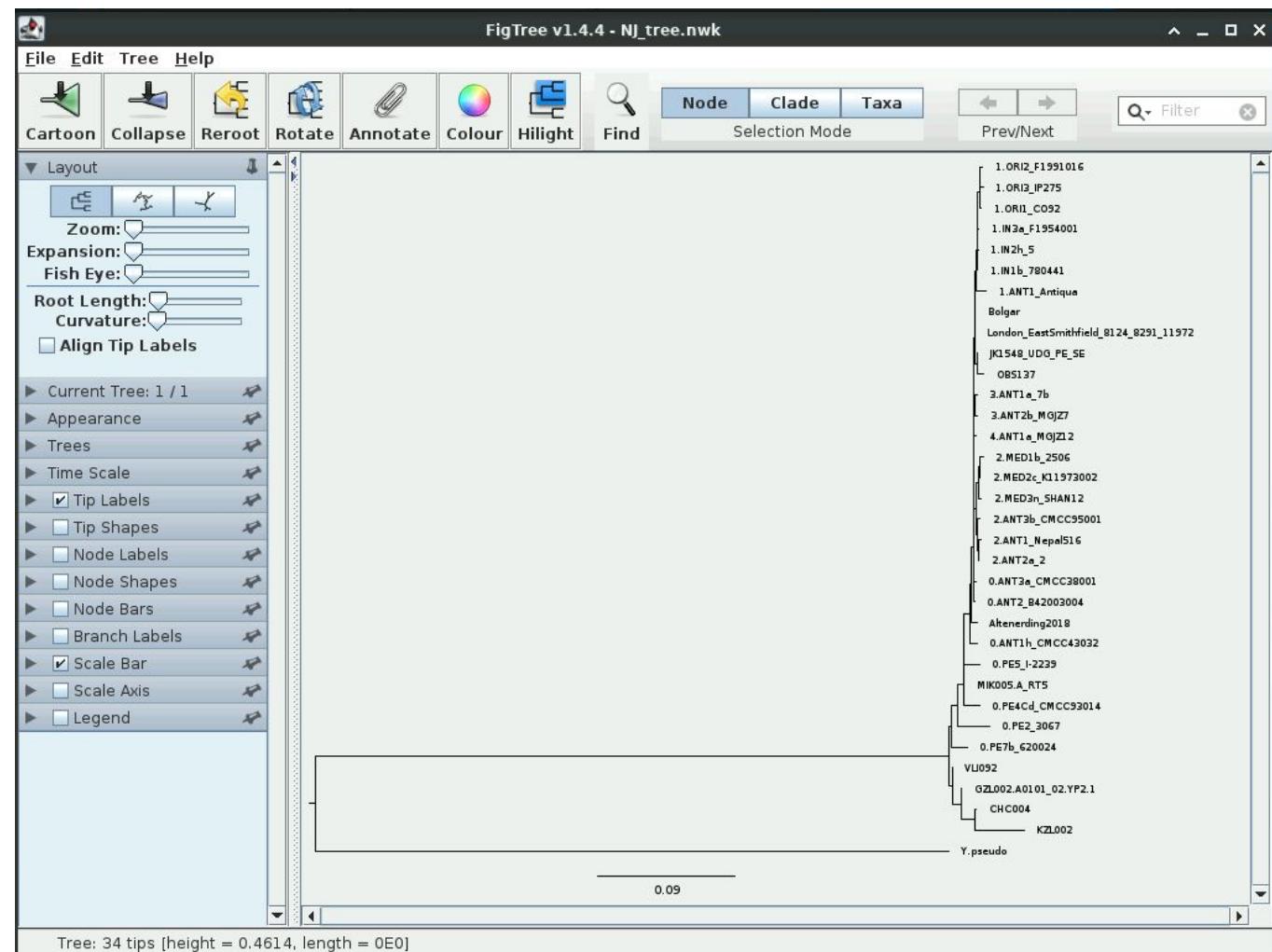
Let's make our own NJ tree!

- Explore in FigTree: type **figtree** in the terminal
- File -> Open -> select “NJ_tree.nwk”
- Which type of tree is this?
 - Unrooted tree
- We know *Y. pseudo* is the outgroup:
 - Root the tree by selecting *Y. pseudo* and pressing rerooot

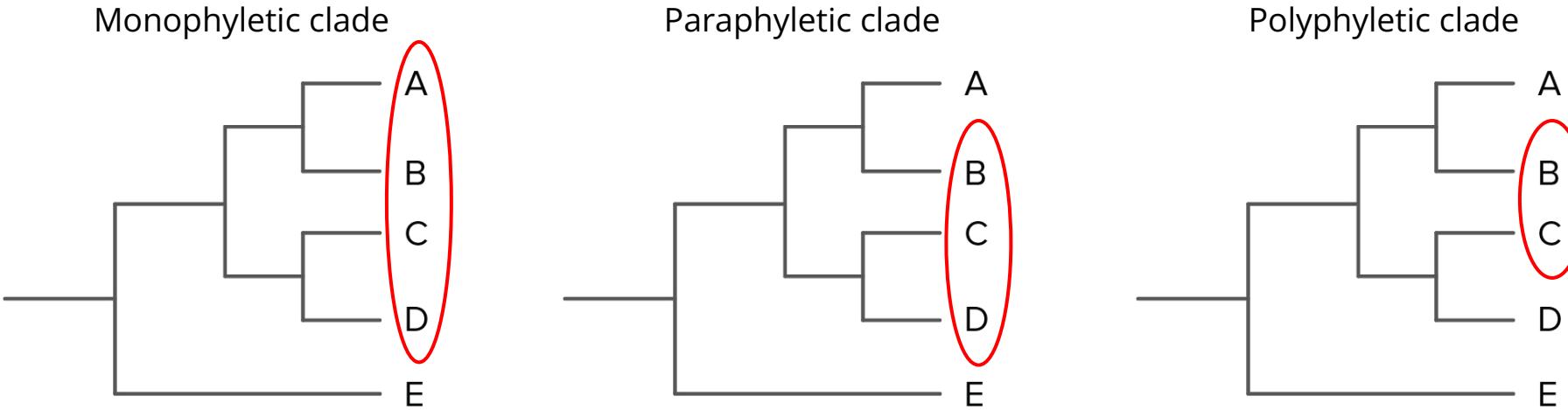


Let's make our own NJ tree!

- How many leaves/tips has our tree?
- Is it an ultrametric tree?
- Where are our taxa of interest?
 - VLI092, GZL002, CHC004 and KZL002
- Do they form a clade?
- Which type of clade?



Types of Clades

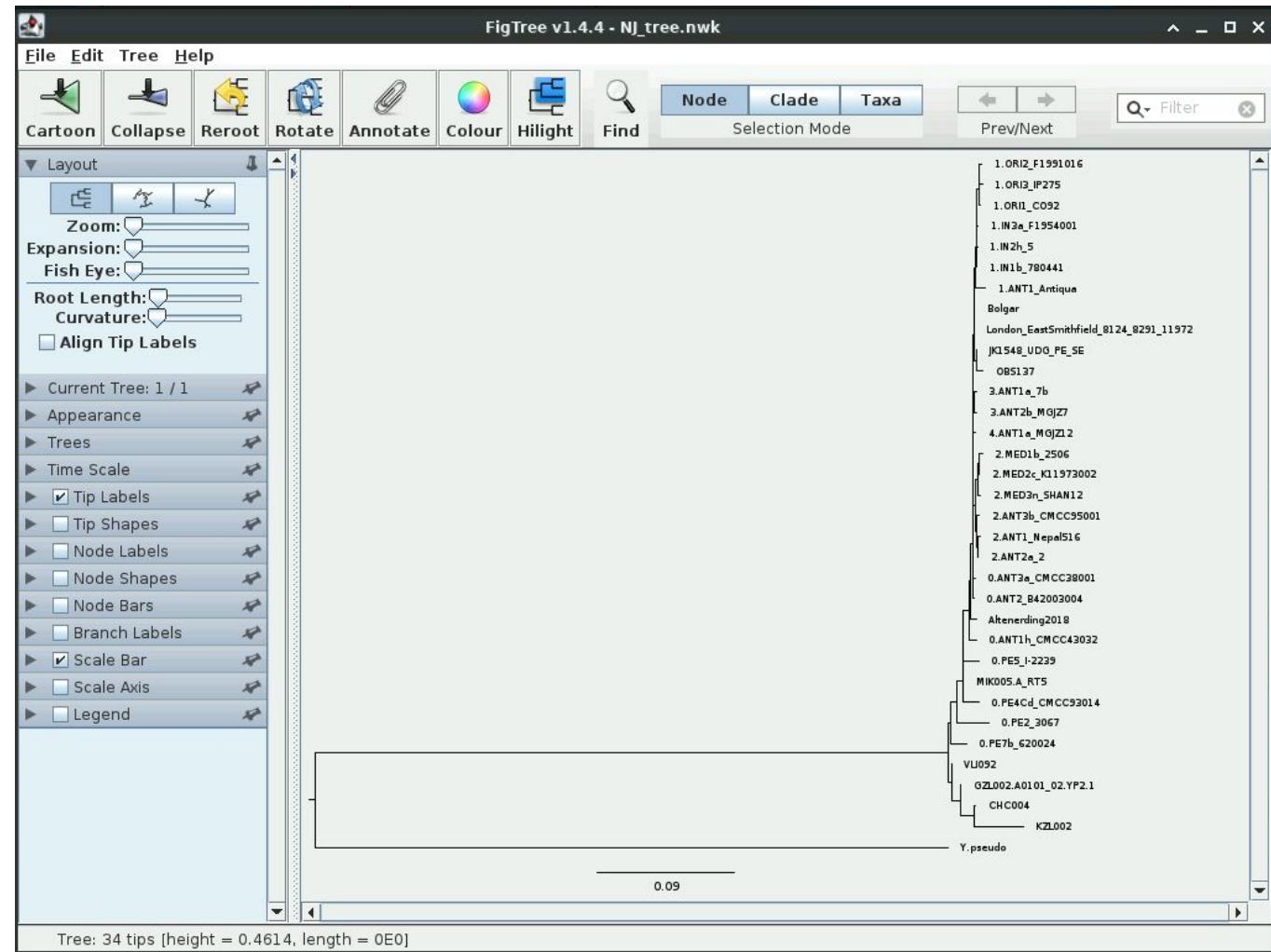


- Monophyletic clade: group of taxa that contain all the taxa that share a common ancestor
- Paraphyletic clade: group of taxa including all the taxa with a common recent ancestor except one or more taxa. In the example A is missing from the clade to be considered monophyletic
- Polyphyletic clade: group of taxa from different monophyletic clades



Let's make our own NJ tree!

- How many leaves/tips has our tree?
- Is it an ultrametric tree?
- Where are our taxa of interest?
 - VLI092, GZL002, CHC004 and KZL002
- Do they form a clade?
- Which type of clade?
 - Monophyletic



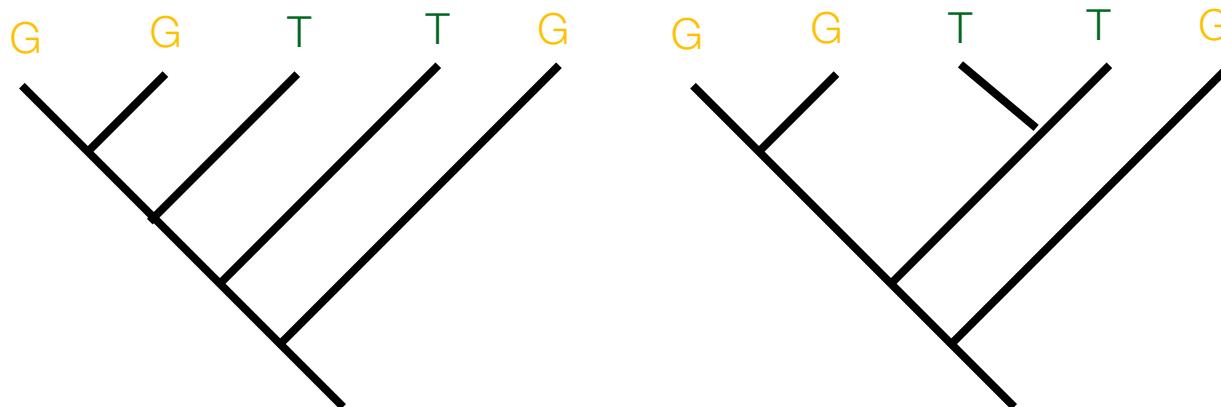
Character-based phylogenetic methods



Maximum parsimony method

1. look at all possible trees
2. reconstruct ancestral states and substitutions along the tree
3. select the tree that involves the least number of substitutions

Note: underlying assumption: the most likely tree is the one that involves the least evolutionary changes



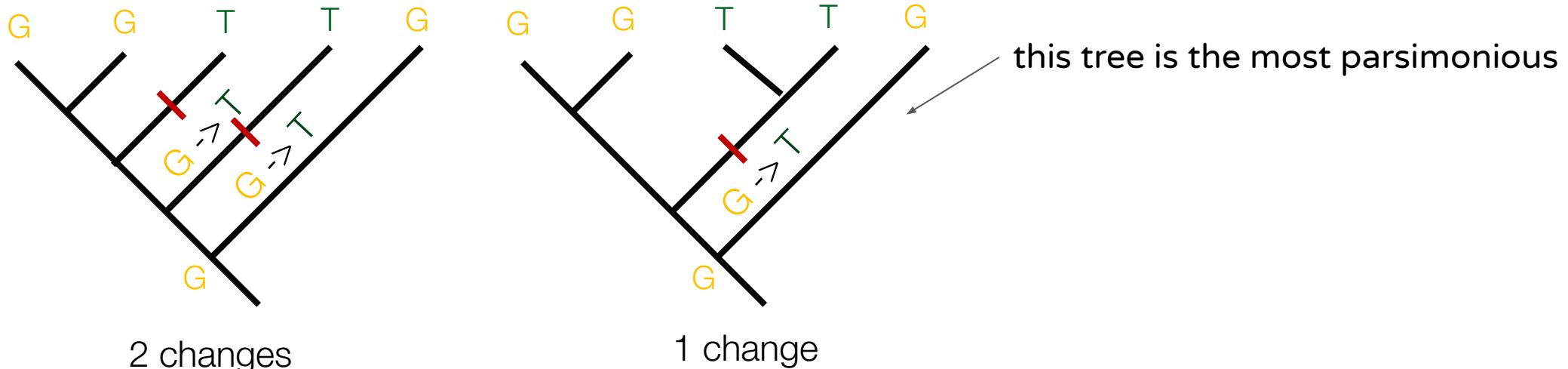
Which tree is the most parsimonious?



Maximum parsimony method

1. look at all possible trees
2. reconstruct ancestral states and substitutions along the tree
3. select the tree that involves the least number of substitutions

Note: underlying assumption: the most likely tree is the one that involves the least evolutionary changes



Maximum parsimony method

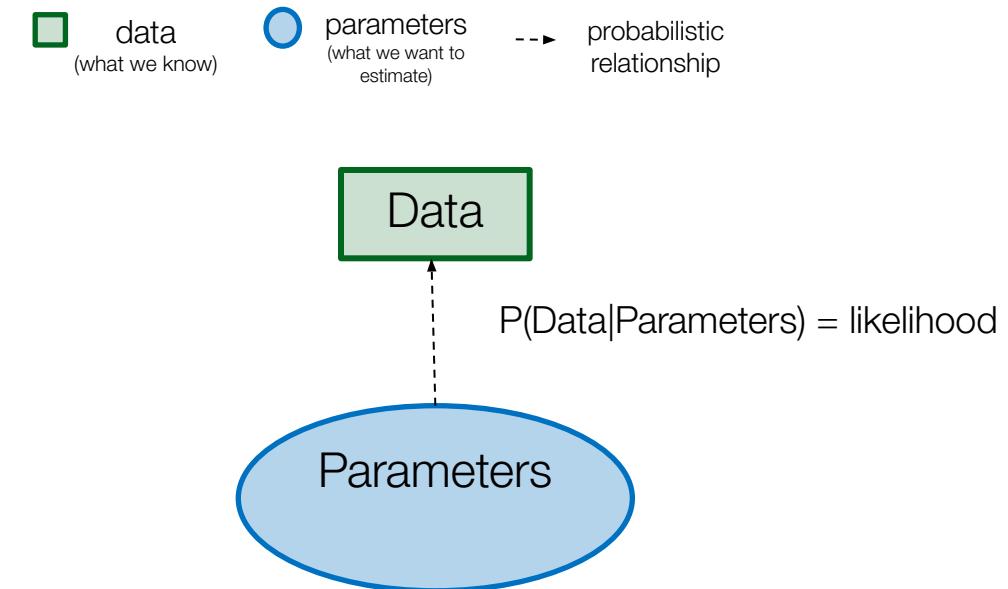
Maximum parsimony algorithms/softwares:

- MEGA
- PAUP
- RAxML
- ...



Probabilistic phylogenetic methods: Maximum likelihood

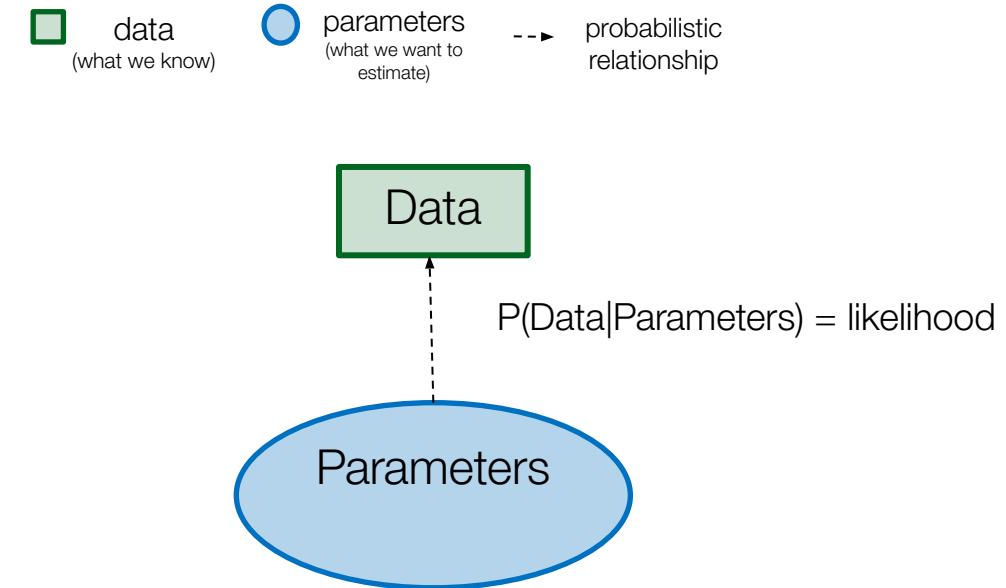
Probabilistic model: stochastic model under which data is generated following a probability distribution depending on a set of parameters



Probabilistic phylogenetic methods: Maximum likelihood

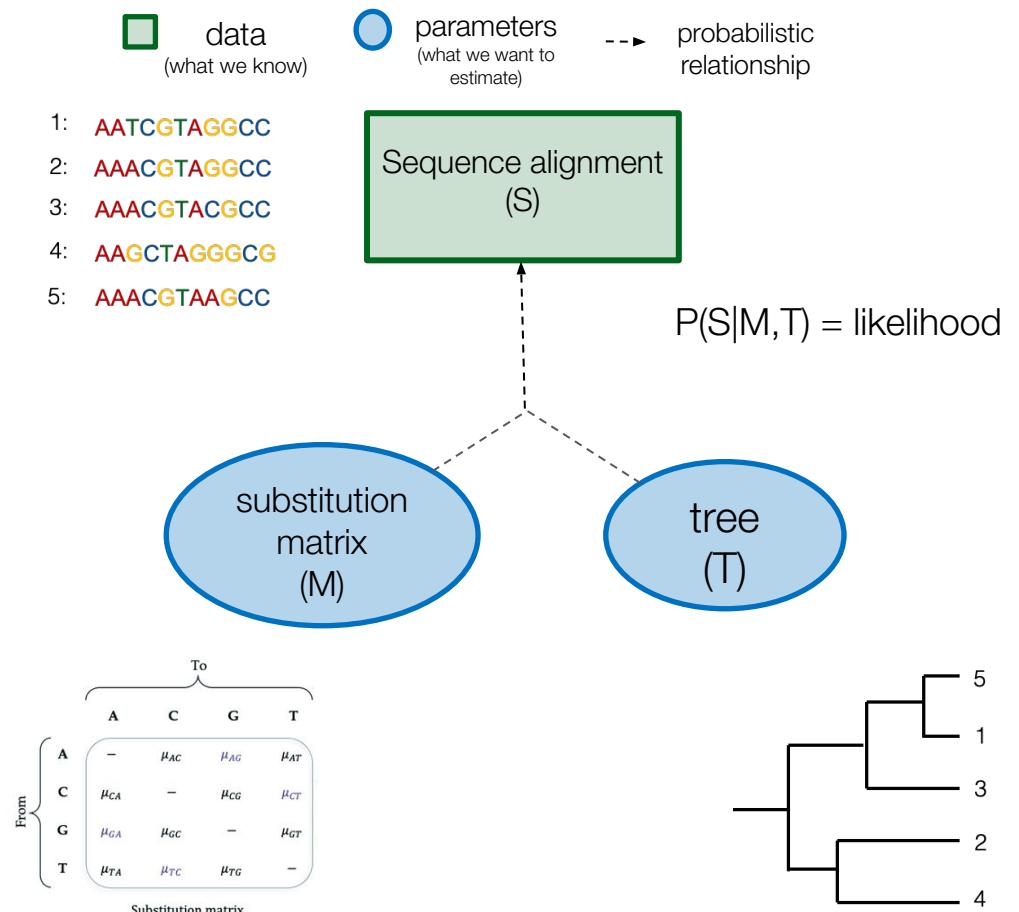
Probabilistic phylogenetic model:

- what is the data?
- what are the parameters?



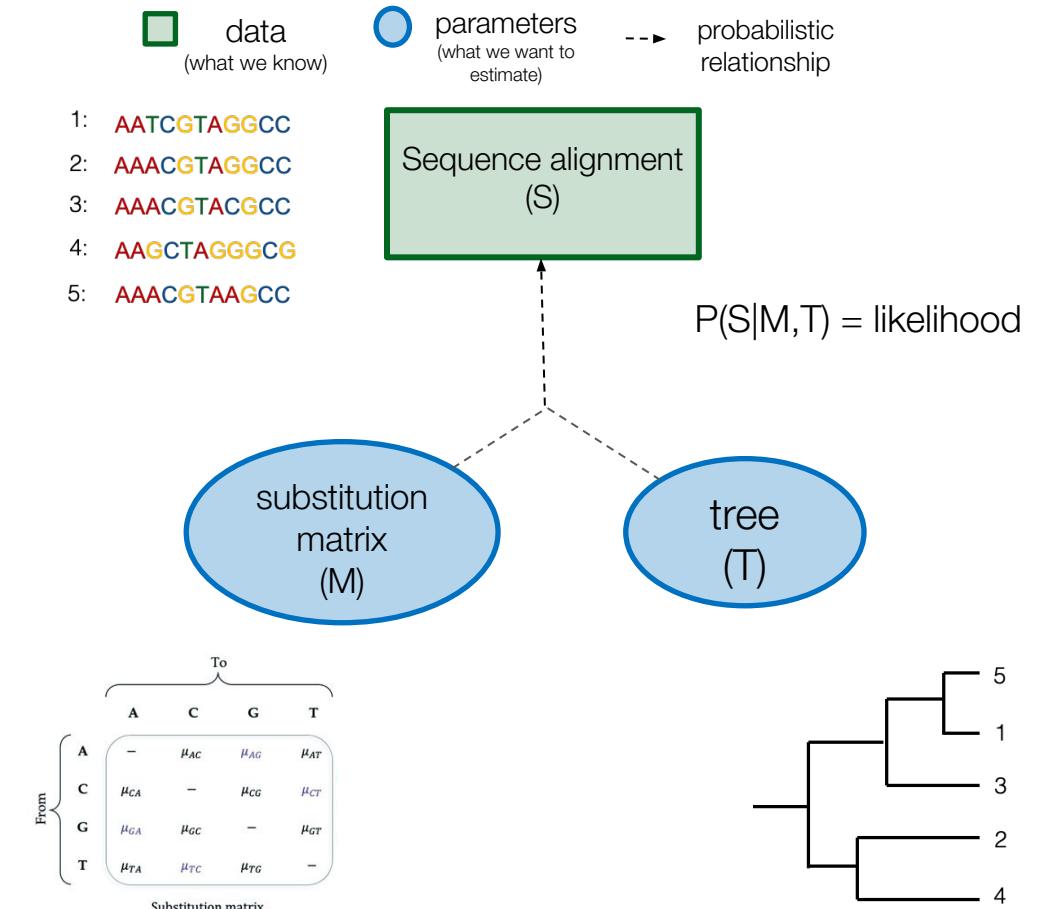
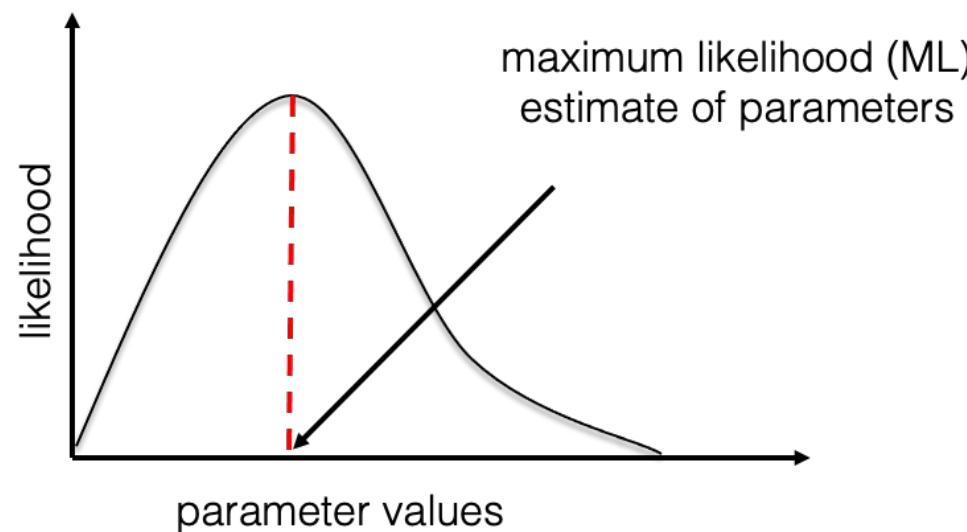
Probabilistic phylogenetic methods: Maximum likelihood

- Probabilistic phylogenetic model:
 - what is the data?
 - what are the parameters?
- continuous-time Markov chain (CTMC) model/Felsenstein model
- good read: *A road map for phylogenetic models of species trees* (Cornuault & Sanmartín 2022)



Probabilistic phylogenetic methods: Maximum likelihood

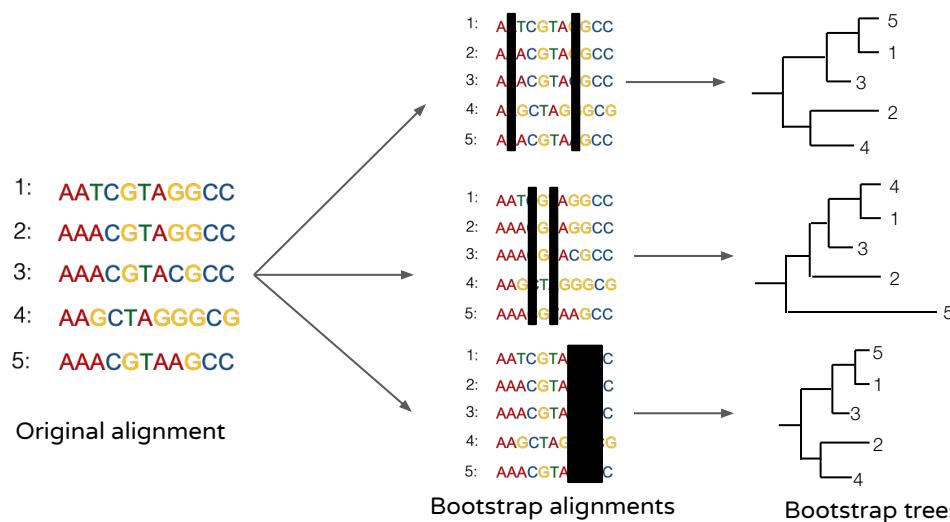
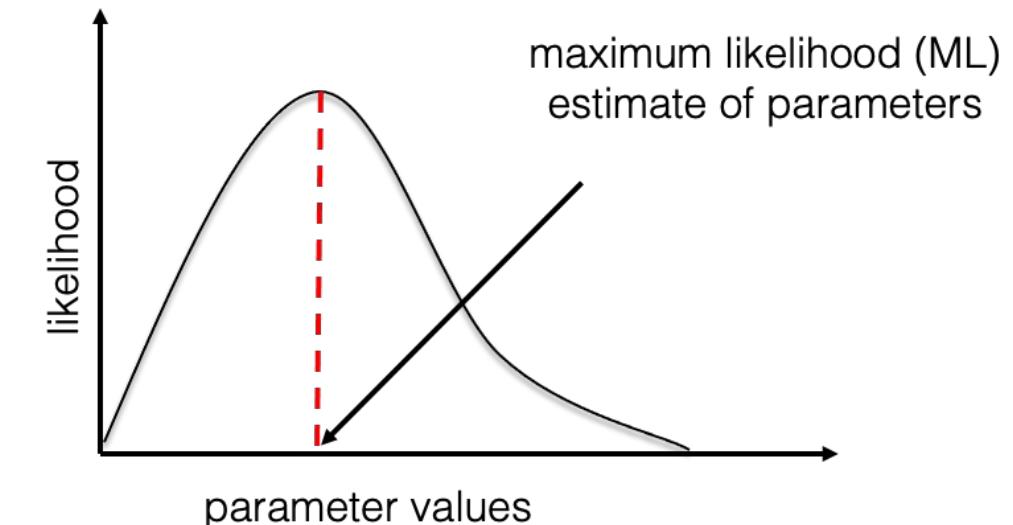
Maximum likelihood inference: most likely parameters are those that maximise the likelihood



Probabilistic phylogenetic methods: Maximum likelihood

A few words on bootstrapping:

- ML estimate is a point estimate: doesn't allow measuring uncertainty
- bootstrapping is a method to do so, by looking at trees estimated from subsampled sequence alignments
- bootstrap support of a clade in the ML tree: percentage of bootstrap trees that contain the clade

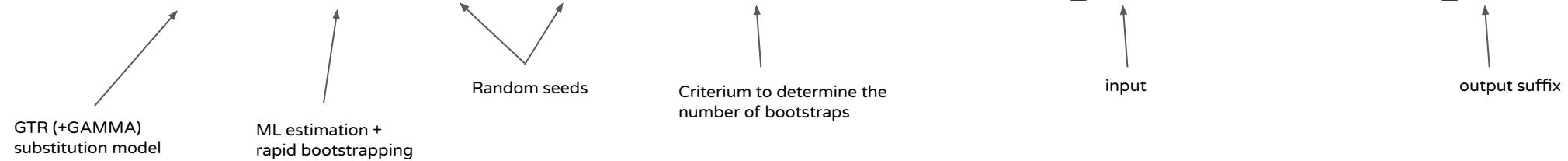


Probabilistic phylogenetic methods: Maximum likelihood

Practical:

- Estimate a maximum likelihood tree from the SNP alignment using RAxML:

```
raxmlHPC-PTHREADS -T 3 -m GTRGAMMA -f a -x 12345 -p 12345 -N autoMRE -s.snpAlignment_session5.fasta -n full_dataset.tre
```



Probabilistic phylogenetic methods: Maximum likelihood

Results:

- Open the tree using *figtree* (RAxML_bipartitions... file). Change “label” to “bootstrap support” at the prompt
- The tree estimated using this model is a substitution tree (branch lengths represent genetic distances in subst./site)
- Not oriented in time: unrooted tree (displayed with a random root in figtree)
- Reroot the tree in figtree using *Y. pseudotuberculosis* as an outgroup (click on the branch leading to *Y. pseudo* and then “Reroot”)
- Where do our genomes of interest from the LNBA period place compared to the rest of *Yersinia pestis* diversity:
VLI092, GZL002, CHC004, KZL002?
 - Do we observe the same topology than in the NJ tree?
- Is that placement well supported? (look at the bootstrap support value: click on the “Node Labels” box and open the drop-down menu, change “Node ages” to “bootstrap support”)



Probabilistic phylogenetic methods: Maximum likelihood

Practical: estimate an ML tree without the outgroup

- Go back to the alignment in mega
- Unclick Y.pseudo, and export in fasta format with the name snpAlignment_without_outgroup.fasta
- Run the same raxml analysis with this alignment

```
raxmlHPC-PTHREADS -T3 -m GTRGAMMA -f a -x 12345 -p 12345 -N autoMRE -s snpAlignment_without_outgroup.fasta -n without_outgroup.tre
```

- Visualize in figtree and re-root based on the previous results: **VLI092, GZL002, CHC004, KZL002 were placed basal to the Y. pestis clade**
- Export the rooted tree: file > export trees > "save as currently displayed" (you can name it ML_tree_rerooted.nexus)



Probabilistic phylogenetic methods: Assessing “temporal signal”

- Sample ages can be used to estimate a phylogeny which is calibrated in time: time tree (branch lengths in time units)
- Doing so requires to assume the “molecular clock hypothesis”: substitutions occur at a rate that is (relatively) constant in time
- It is good practice to assess whether the genetic sequences we analyse behave like molecular clocks before trying to estimate a time-tree
- Classic test: root-to-tip regression:
 - oldest samples should be closer to the root of the (substitution) tree because they had less time to accumulate substitutions
 - do the distance of the tips (sampled sequences) to the root (most recent common ancestor) correlate with sample age?



Probabilistic phylogenetic methods: Assessing “temporal signal”

Practical: root-to-tip regression with *tempest*:

- open *tempest* and load the re-rooted ML tree that we produced previously
- click on "import dates" in the "sample dates" tab, select the sample_age.txt file, and then change to "dates specified as years before the present"
- Look at the root-to-tip regression: is there a positive correlation?
- Which points correspond to the genomes of interest (VLI092, GZL002, CHC004, KZL002)?



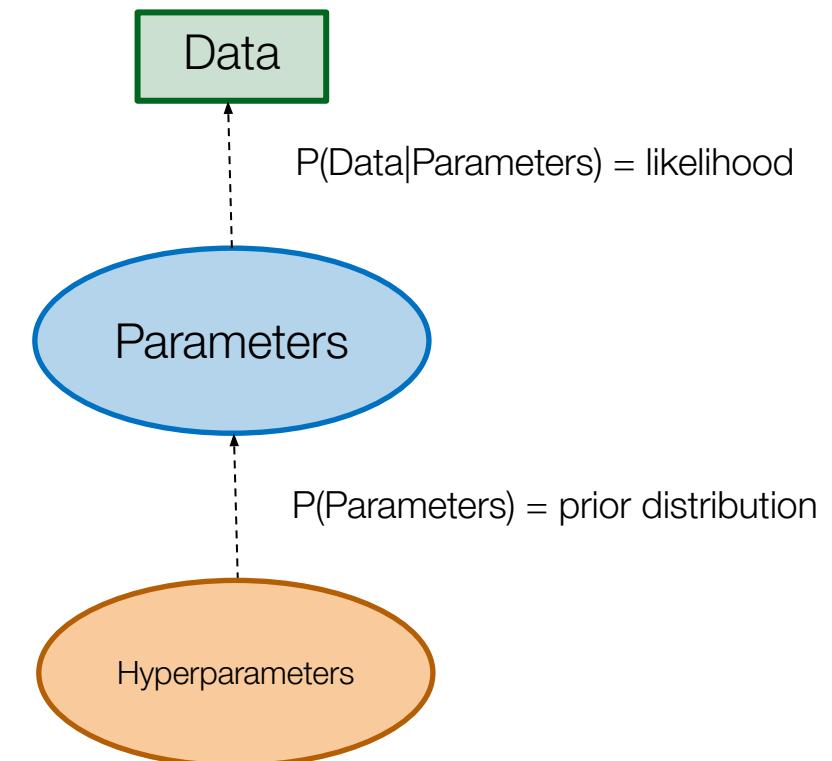
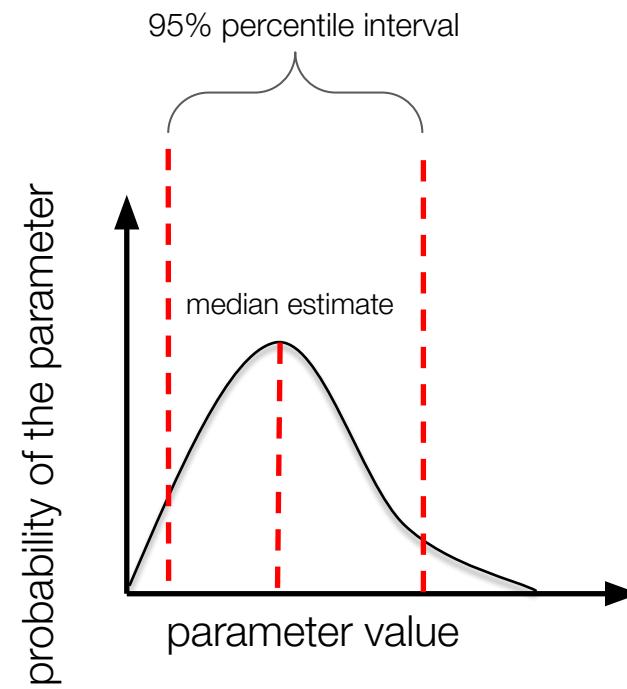
Probabilistic phylogenetic methods:

Estimating a time tree using Bayesian phylogenetics

Bayesian inference:

$$\begin{aligned} P(\text{Parameters}|\text{Data}) &\propto P(\text{Data}|\text{Parameters}) \times P(\text{Parameters}) \\ \text{Posterior} &\propto \text{Likelihood} \times \text{Prior} \end{aligned}$$

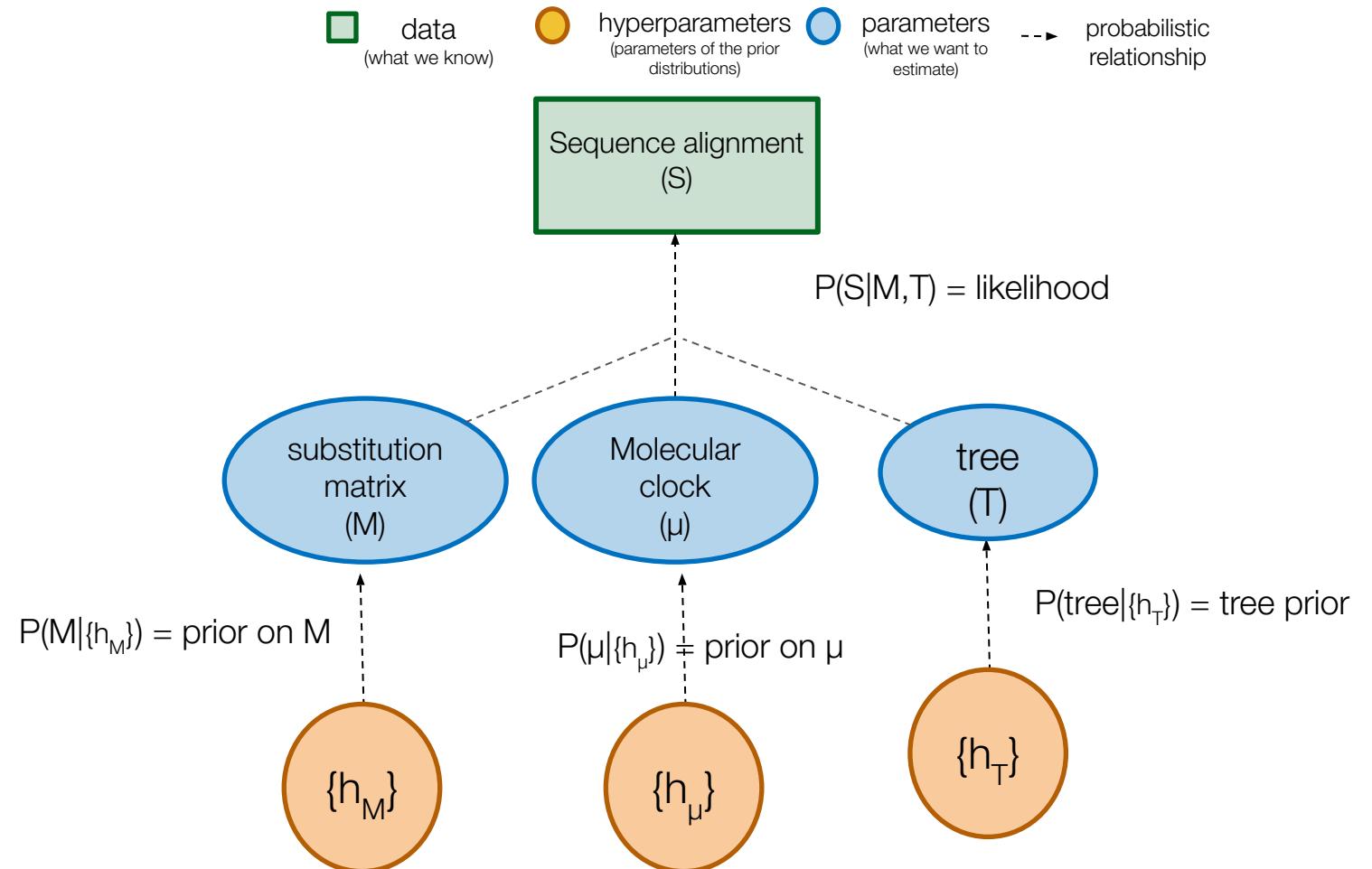
 data (what we know)  hyperparameters (parameters of the prior distributions)  parameters (what we want to estimate)  probabilistic relationship



Probabilistic phylogenetic methods:

Estimating a time tree using Bayesian phylogenetics

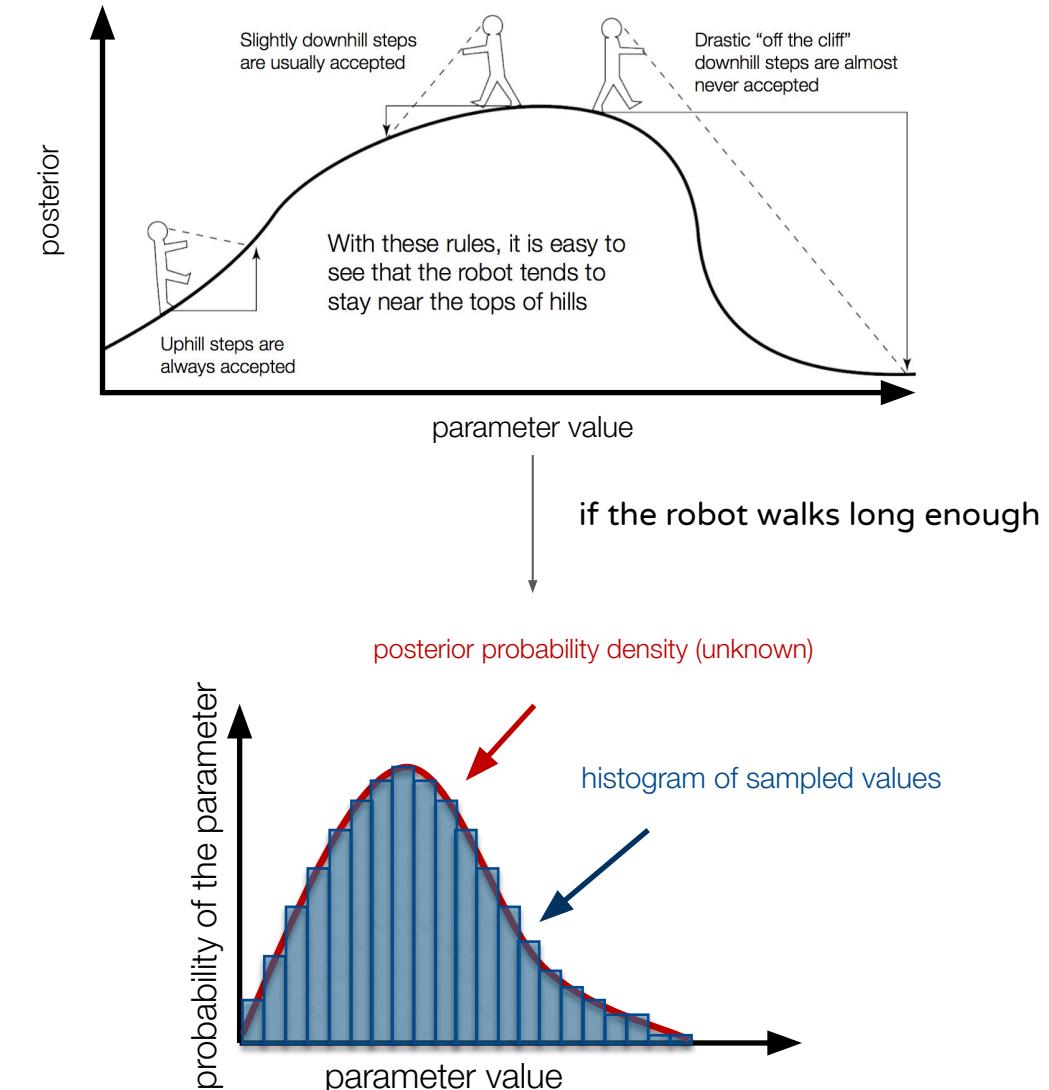
Bayesian phylogenetics:



Probabilistic phylogenetic methods: Estimating a time tree using Bayesian phylogenetics

Bayesian inference:

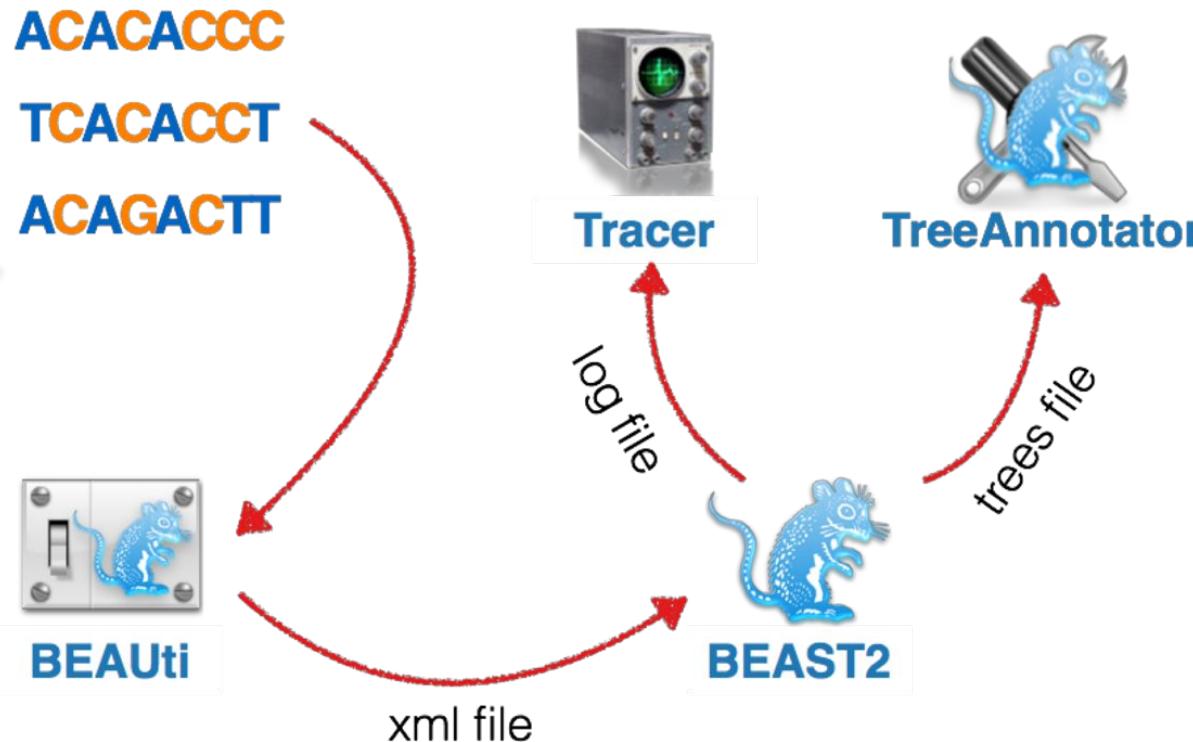
- characterizing the complete posterior distribution is impossible (computation for every possible combination of parameters)
- we “sample” from it using **MCMC** (Monte Carlo Markov Chain)
- the MCMC “robot” samples value along its walk with a frequency proportional to the posterior



Probabilistic phylogenetic methods:

Estimating a time tree using Bayesian phylogenetics

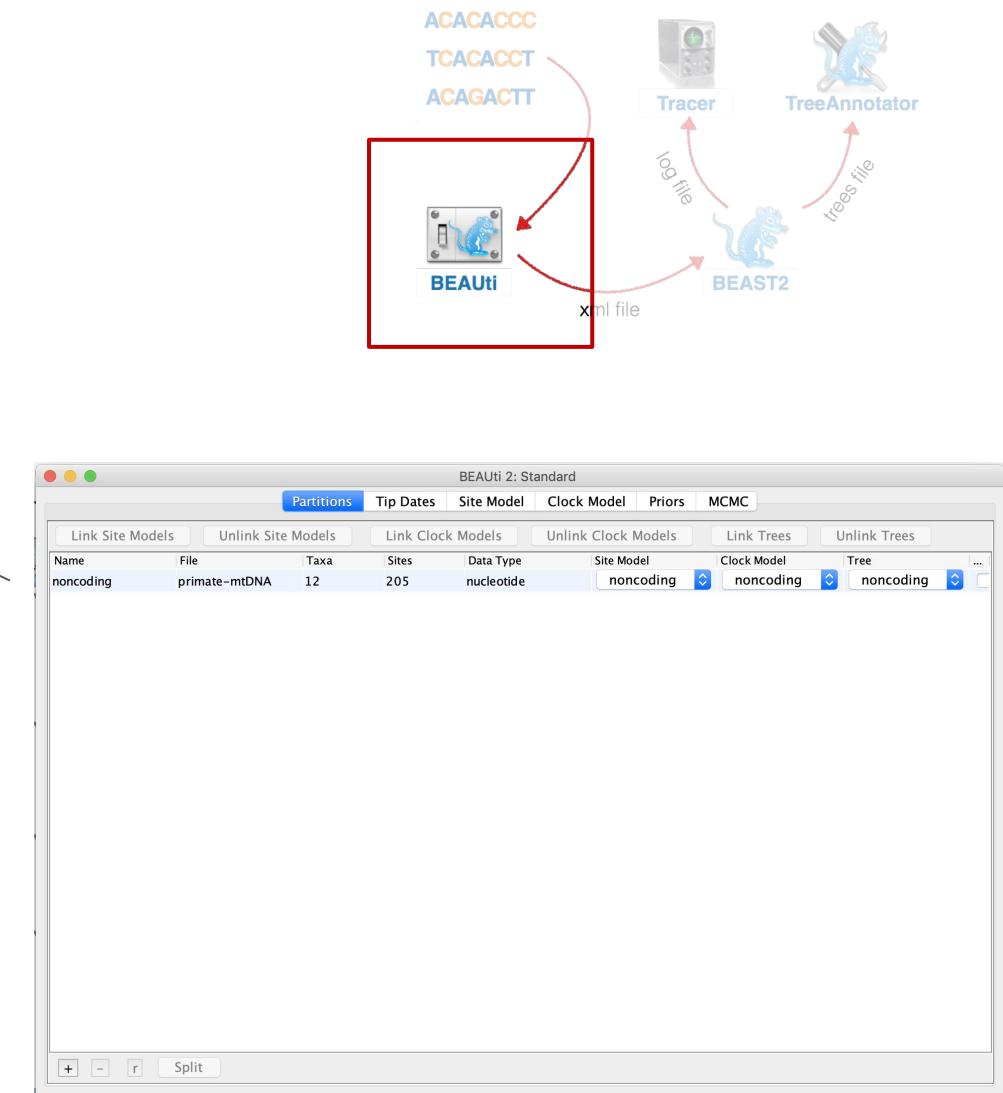
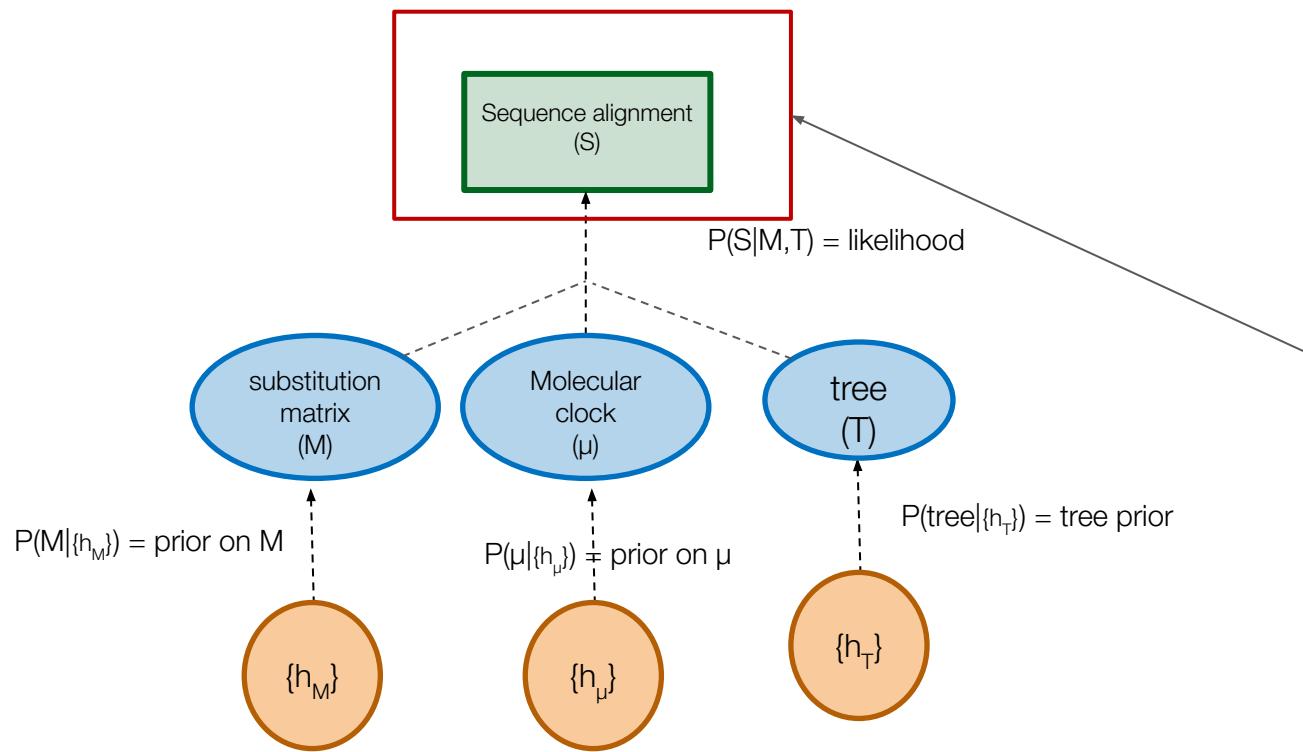
Practical: run a *Beast* analysis



Probabilistic phylogenetic methods:

Estimating a time tree using Bayesian phylogenetics

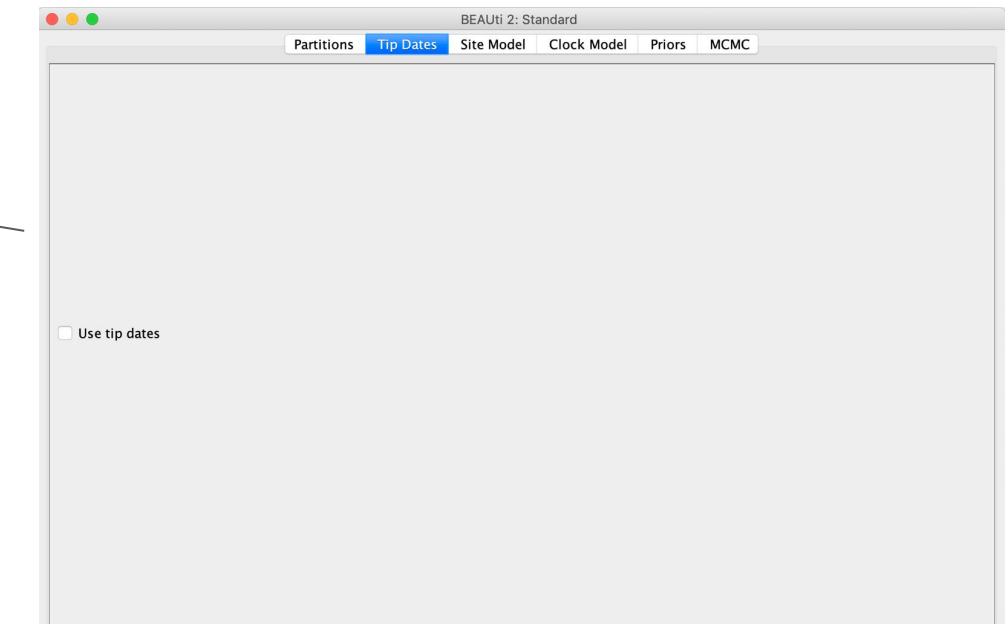
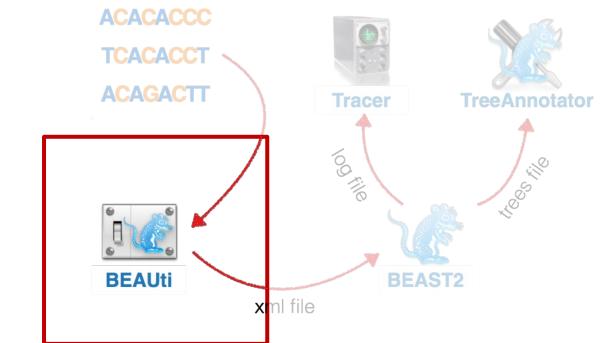
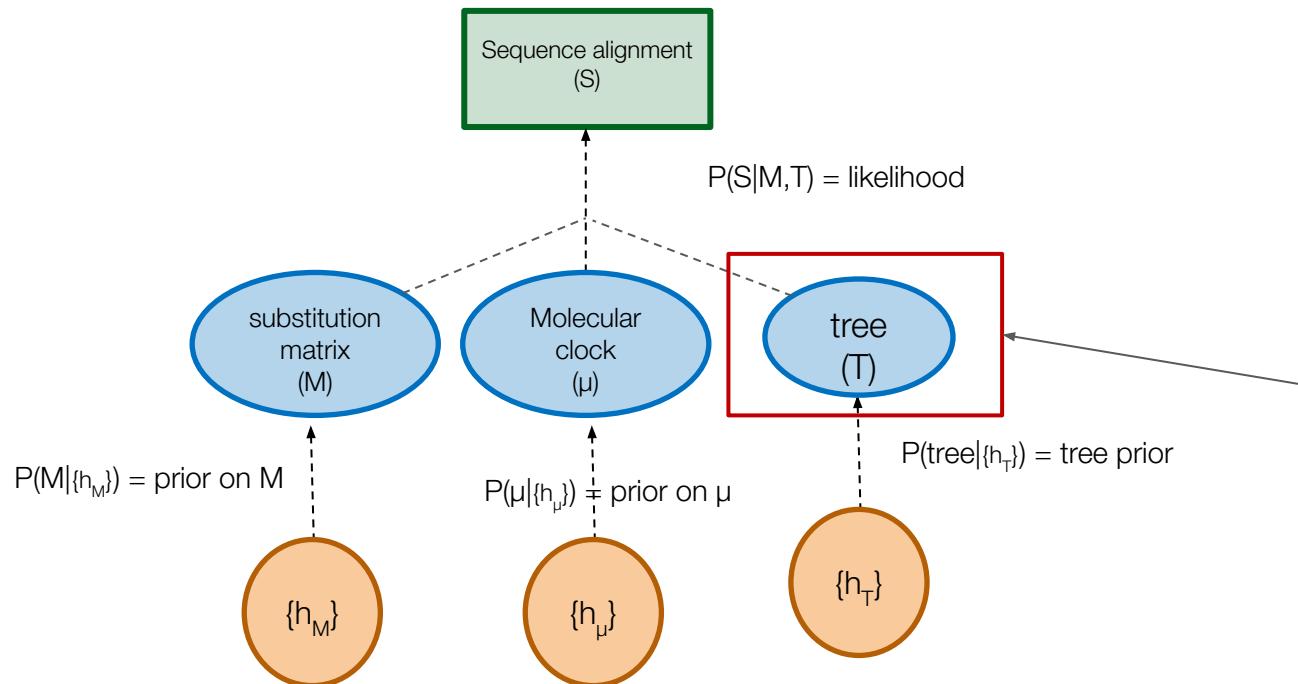
Practical: run a *Beast* analysis



Probabilistic phylogenetic methods:

Estimating a time tree using Bayesian phylogenetics

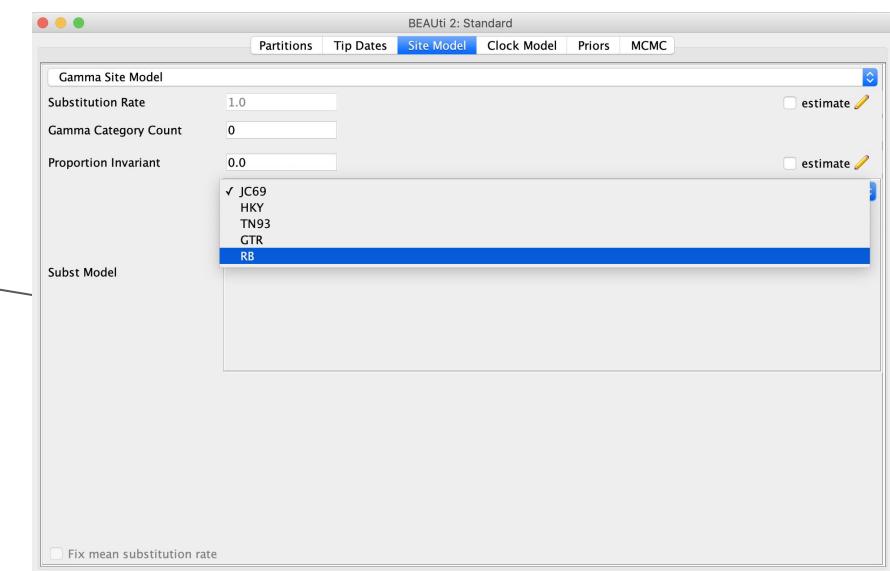
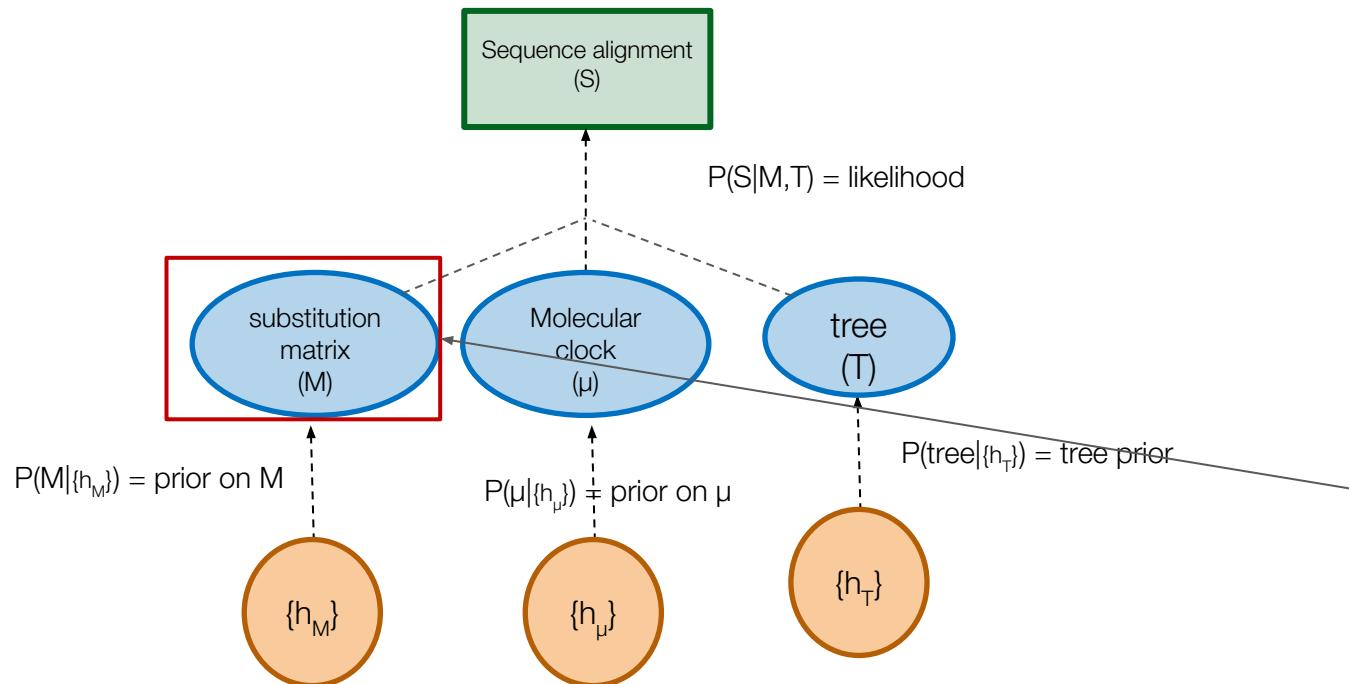
Practical: run a *Beast* analysis



Probabilistic phylogenetic methods:

Estimating a time tree using Bayesian phylogenetics

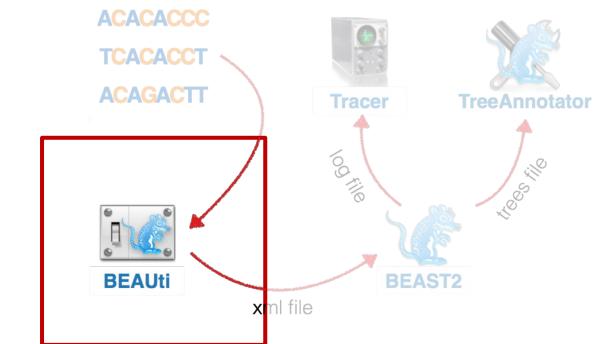
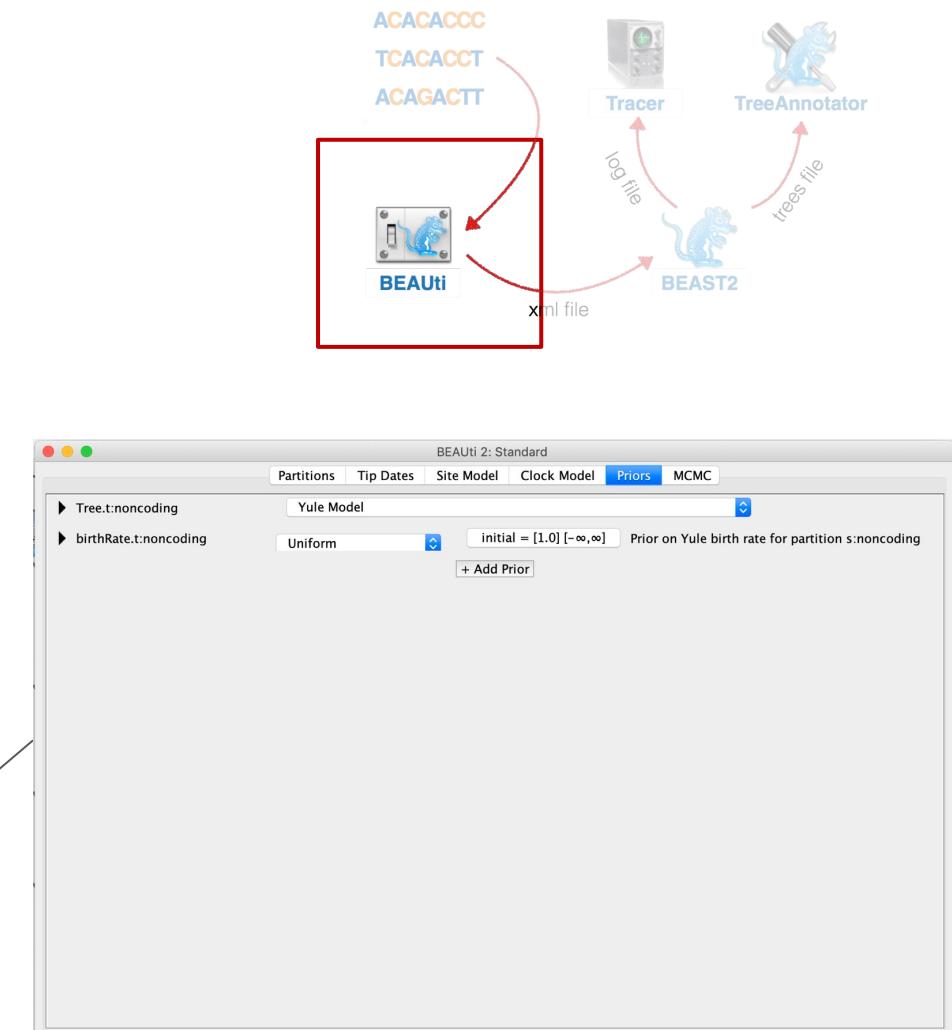
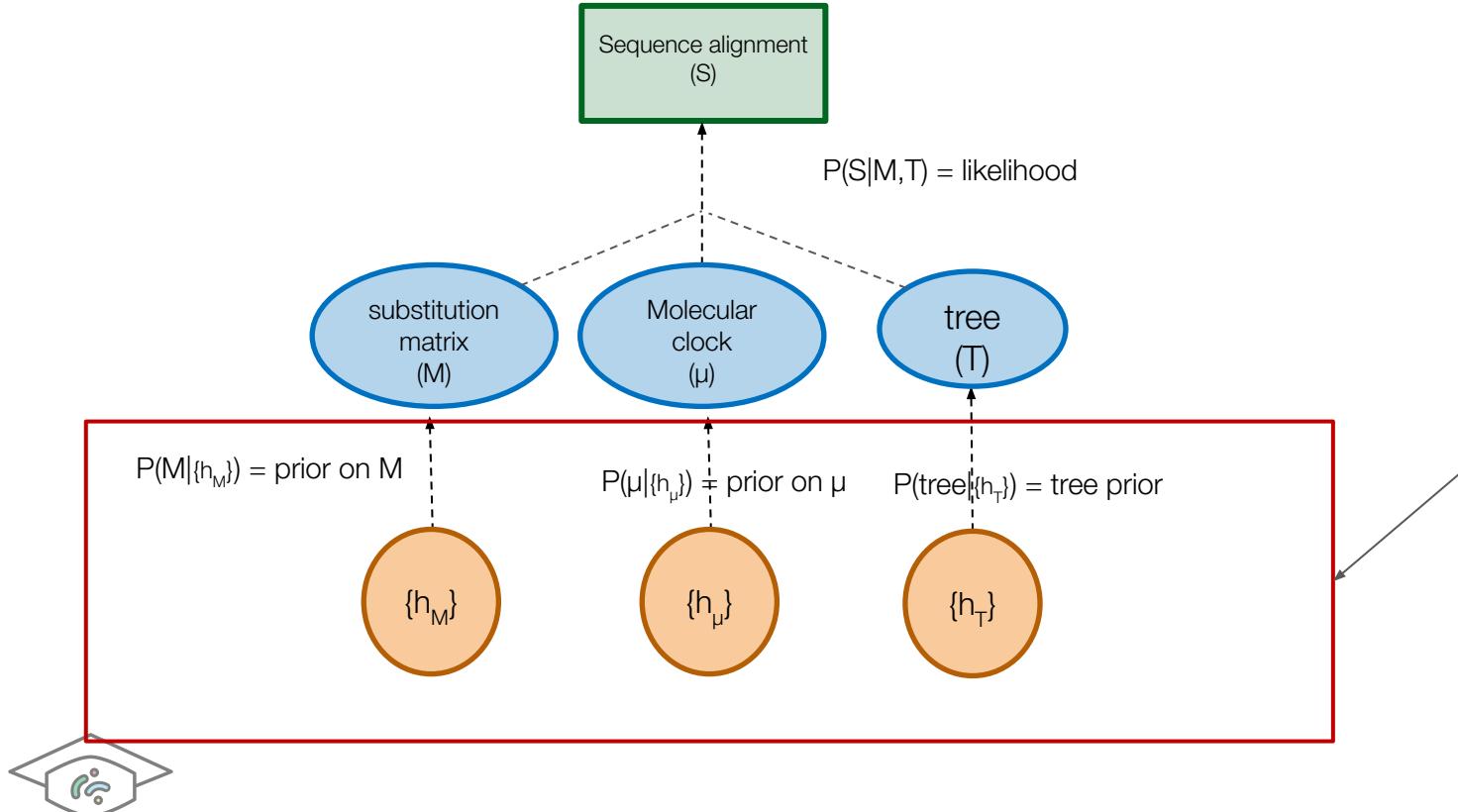
Practical: run a *Beast* analysis



Probabilistic phylogenetic methods:

Estimating a time tree using Bayesian phylogenetics

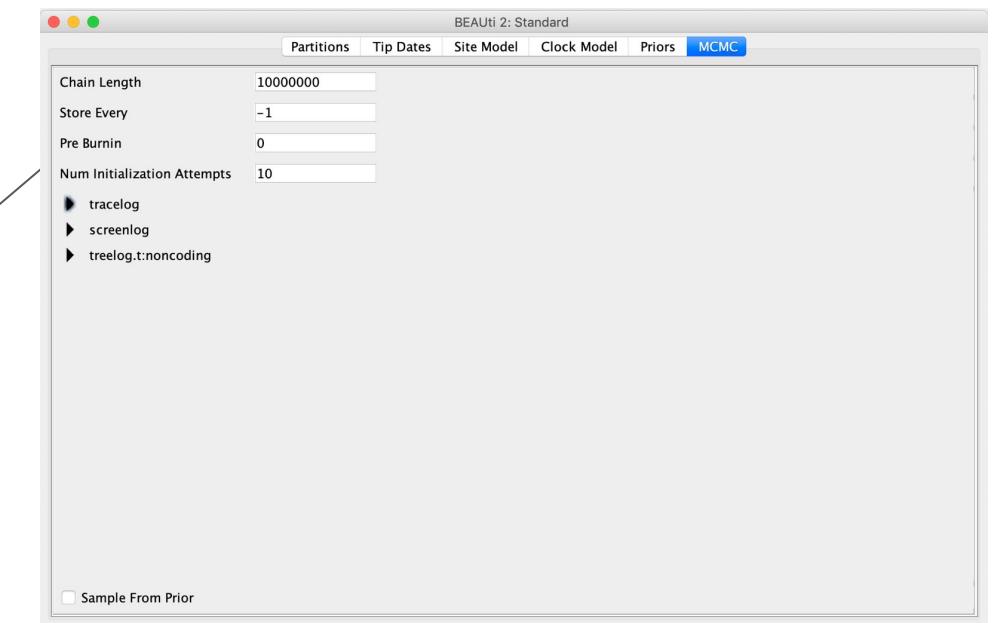
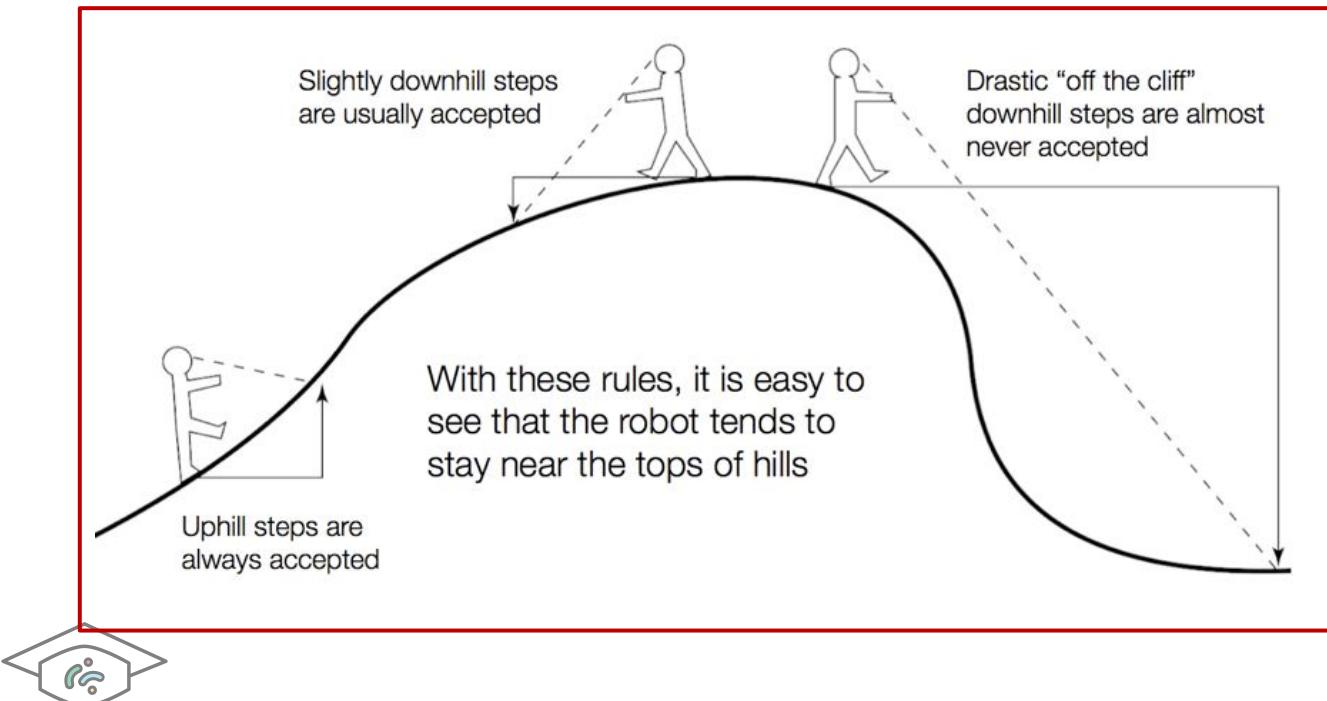
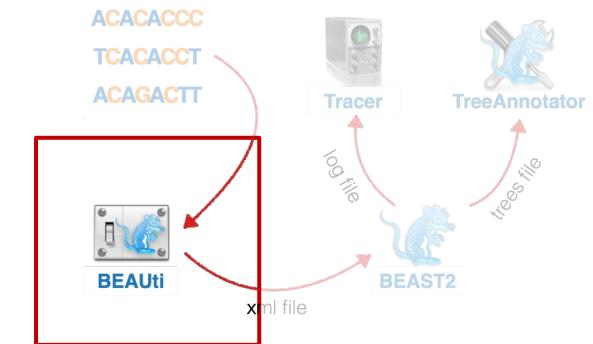
Practical: run a *Beast* analysis



Probabilistic phylogenetic methods:

Estimating a time tree using Bayesian phylogenetics

Practical: run a *Beast* analysis



Probabilistic phylogenetic methods:

Estimating a time tree using Bayesian phylogenetics

Practical: prepare a *BEAST* analysis in *beauti*

- load the alignment without outgroup (“Partitions” tab)
- load sample ages (“Tip Dates” tab; **don't forget to change to years before present**)
- Use a GTR substitution matrix with a Gamma site model and 4 gamma categories (“Site Model” tab)
- Use a relaxed clock lognormal model with an initial value of 1E-4 (“Clock Model” tab)
- Use a Bayesian Skyline Coalescent tree prior (“Priors” tab)
- Change the Mean clock prior to a uniform distribution between 1E-6 and 1E-3 subst/site/year (“Priors” tab)
- Use 300M iterations for the MCMC chain, and log the parameters and trees each 10,000 iterations (“MCMC” tab)
- Leave everything else as default and save the setup (an xml file will be created)
- Run the analysis:

```
beast input.xml
```

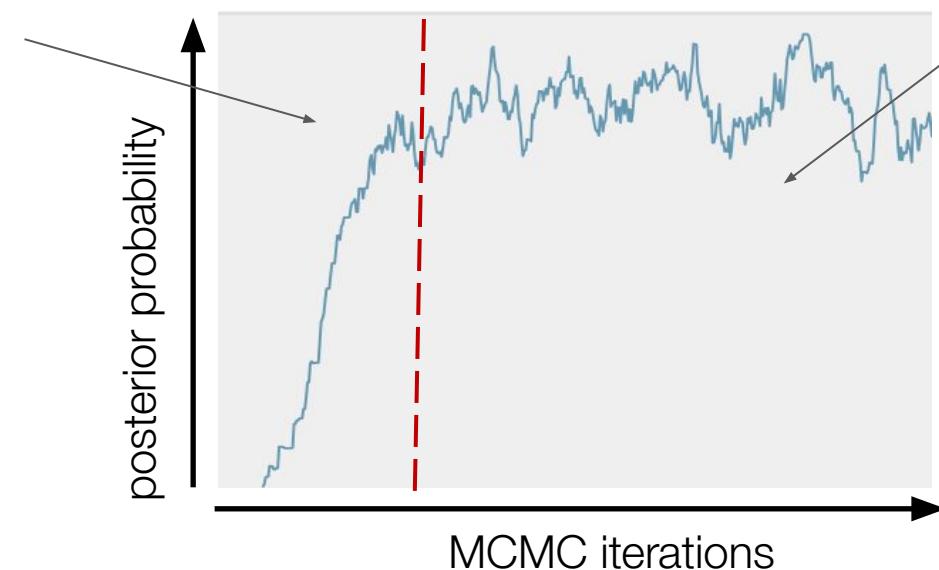


Probabilistic phylogenetic methods: Estimating a time tree using Bayesian phylogenetics

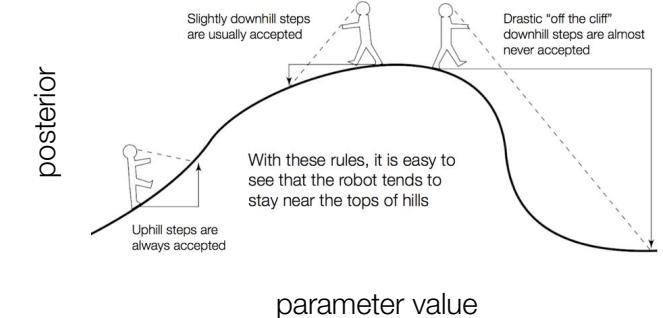
How do we know that the robot has walked long enough?

- trace plot: shows sampled parameter/posterior along the MCMC chain
- the MCMC robot starts at a random place in the parameter space. It is going to take some time before it finds the posterior “hill” and starts sampling values that make sense: this is the so called burn-in phase
- we need to check that we have passed the burn-in phase and remove the corresponding samples

burn in phase: the MCMC robot is climbing the posterior hill



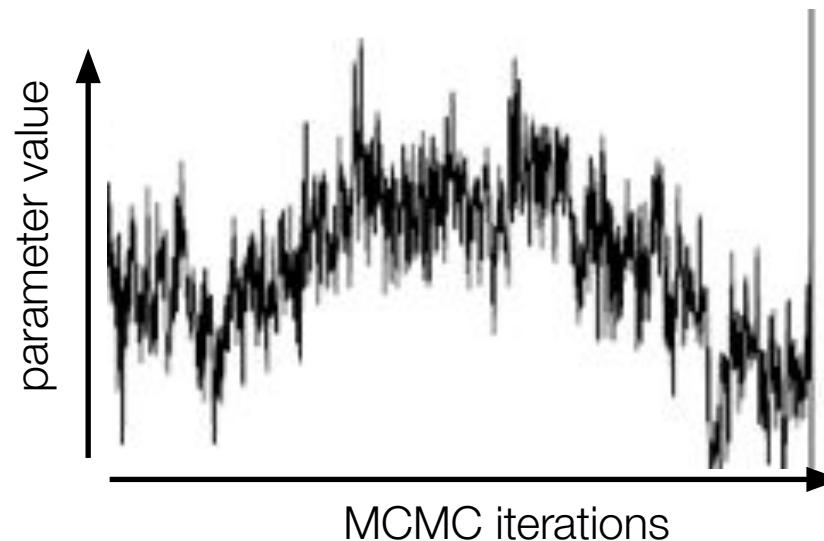
post-burn in phase: the MCMC robot is oscillating around the top of the posterior hill



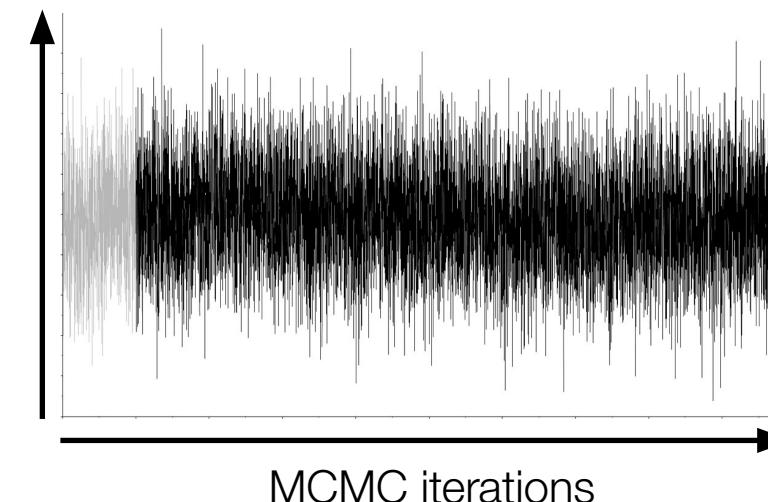
Probabilistic phylogenetic methods: Estimating a time tree using Bayesian phylogenetics

How do we know that the robot has walked long enough?

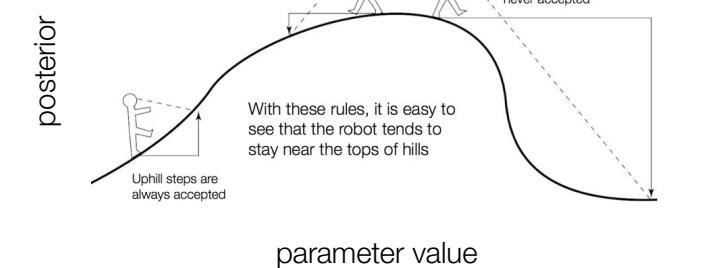
- trace plot
- number of uncorrelated samples = effective sampling size (ESS) > 200



low ESS



high ESS > 200

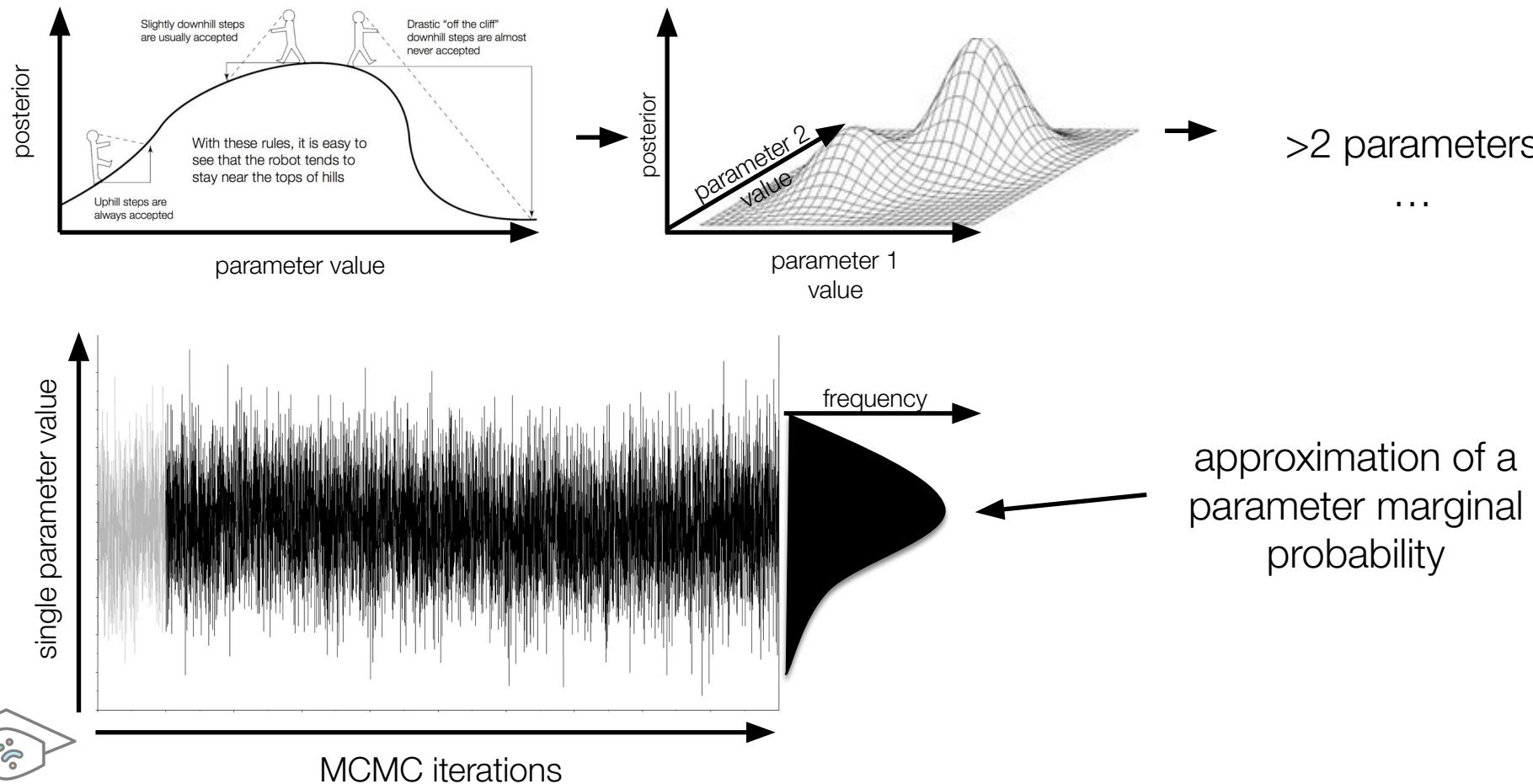


the hairy caterpillar!



Probabilistic phylogenetic methods: Estimating a time tree using Bayesian phylogenetics

MCMC and marginal posterior

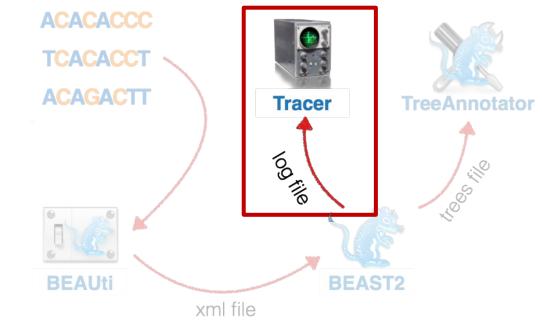


Probabilistic phylogenetic methods:

Estimating a time tree using Bayesian phylogenetics

Practical: look at the beast results with tracer

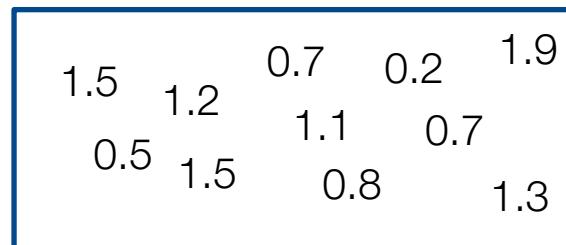
- open the log file (contains all the parameter values that have been sampled by the MCMC robot, except the trees)
- how many samples should we burn?
- do it, and look at the ESS values of the parameters
- has the MCMC robot run long enough?
- how much longer do you think we should run it?



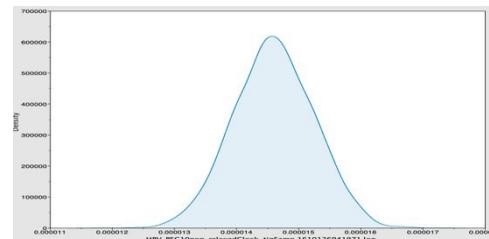
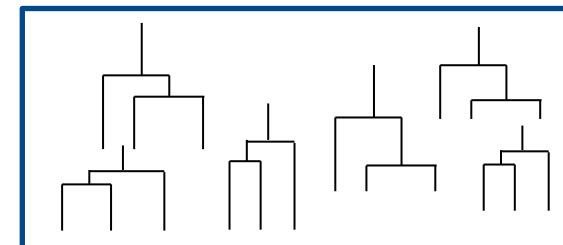
Probabilistic phylogenetic methods: Estimating a time tree using Bayesian phylogenetics

How to represent the posterior distribution of trees

Mutation rate



Tree



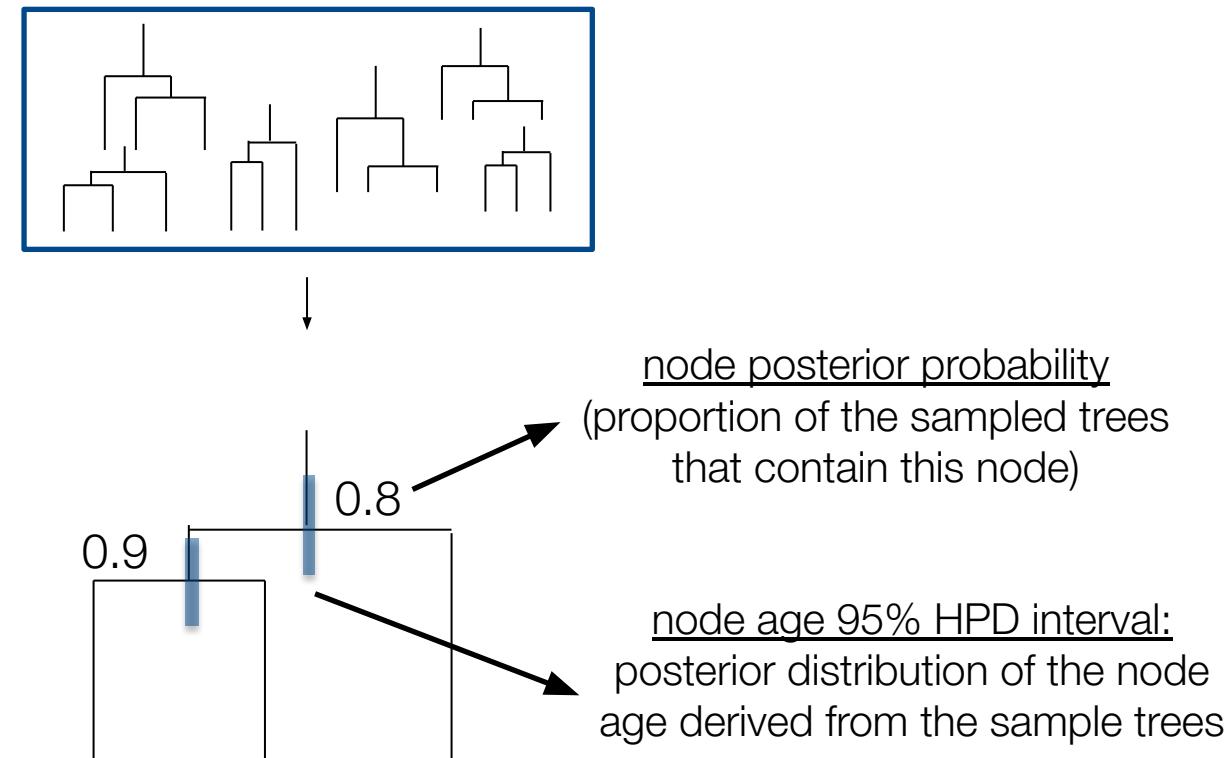
?



Probabilistic phylogenetic methods:

Estimating a time tree using Bayesian phylogenetics

How to represent the posterior distribution of trees



Maximum clade credibility (MCC) tree: sampled tree that
maximises the product of node posterior probability



Probabilistic phylogenetic methods: Estimating a time tree using Bayesian phylogenetics

Practical: construct an MCC tree from the tree sample using *treeannotator*

```
treeannotator -lowMem -burnin 10 input.trees output.MCC.tree
```

- open the tree in figtree
- is the root of the tree consistent with what we found previously?
- is this placement well supported?
- what is the age of the most recent common ancestor of all Y. pestis strains?



Outlook

How to deal with low coverage/noisy ancient DNA data?

- probabilistic methods should be able to handle missing data correctly
but
- potential biases with high levels of missing data difficult to predict: usual practice is to remove alignment positions that contain high levels of missing data (i.e., positions for which $>X\%$ of the sequences are not resolved; can be done in MEGA)
- highly incomplete sequences:
 - will introduce uncertainty/noise in the trees
 - are also usually low-coverage and therefore error-prone (sequencing errors, DNA damage)
 - good practice = exclude them when building the tree and trying to place them a posteriori (phylogenetic placement softwares: EPA-ng, MGPlacer, [pathPhynder](#))



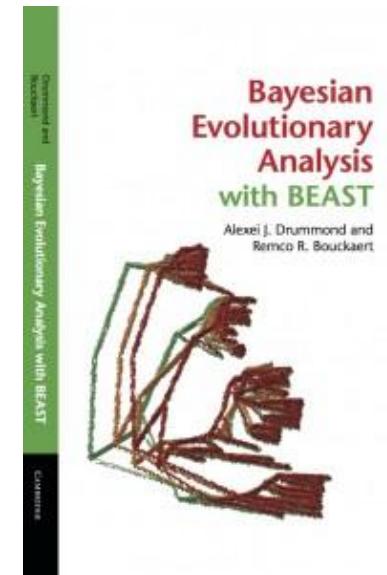
Outlook

- accounting for genetic recombination in phylogenetic inference?
 - identify recombinant genomes: *RDP4*
 - build a phylogeny excluding recombinant regions: *clonalFrame*
 - build an ancestral recombination graph: *bacter/recomb* packages (*BEAST2*)
- phylogenies can tell us more:
 - phylogeography
 - phylodynamics



Resources/contact

- aida_andrades@eva.mpg.de/arthur_kocher@eva.mpg.de
- taming the BEAST website (*BEAST2* tutorials): <https://taming-the-beast.org>
- the *BEAST2* book: Drummond A. J. & Bouckaert R. R. (2015) *Bayesian evolutionary analysis with BEAST*, Cambridge University Press.
- Cornuault, J., & Sanmartín, I. (2022). *A road map for phylogenetic models of species trees*. *Molecular Phylogenetics and Evolution*, 107483. doi: 10.1016/j.ympev.2022.107483
- Yang, Z. (2014). Molecular evolution: a statistical approach. Oxford University Press.

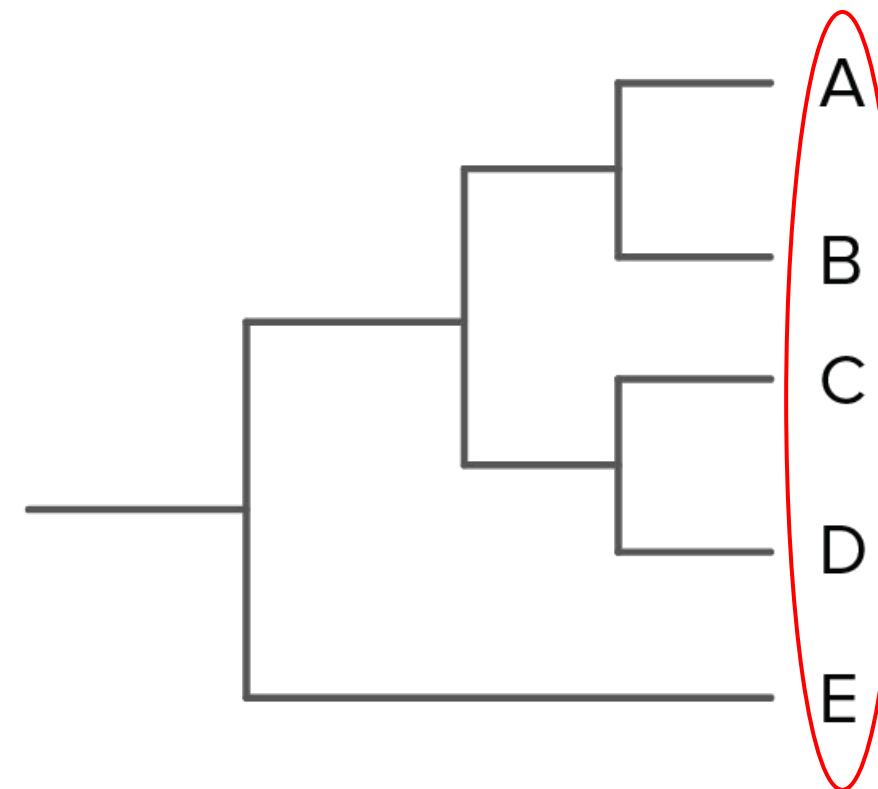


Extra slides



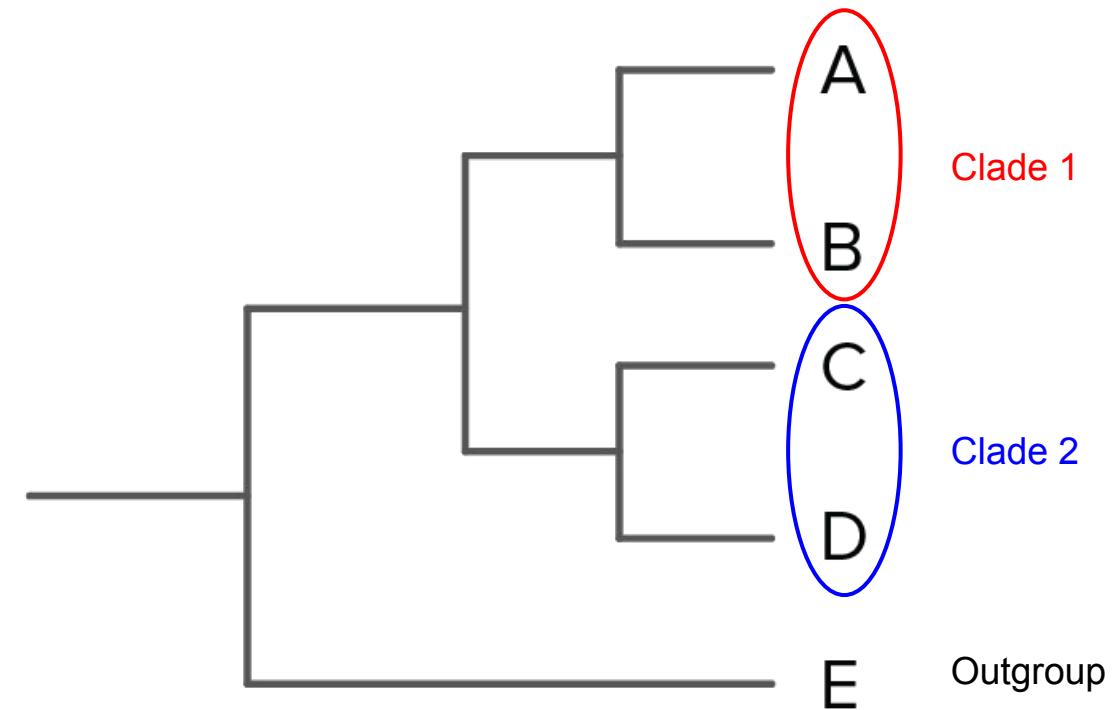
Parts of phylogenetic tree: Leaves

- A, B, C, D, E: the different taxa used to build the tree
- They are the leaves of our tree
- Leaves can also be referred as tips or terminal taxa



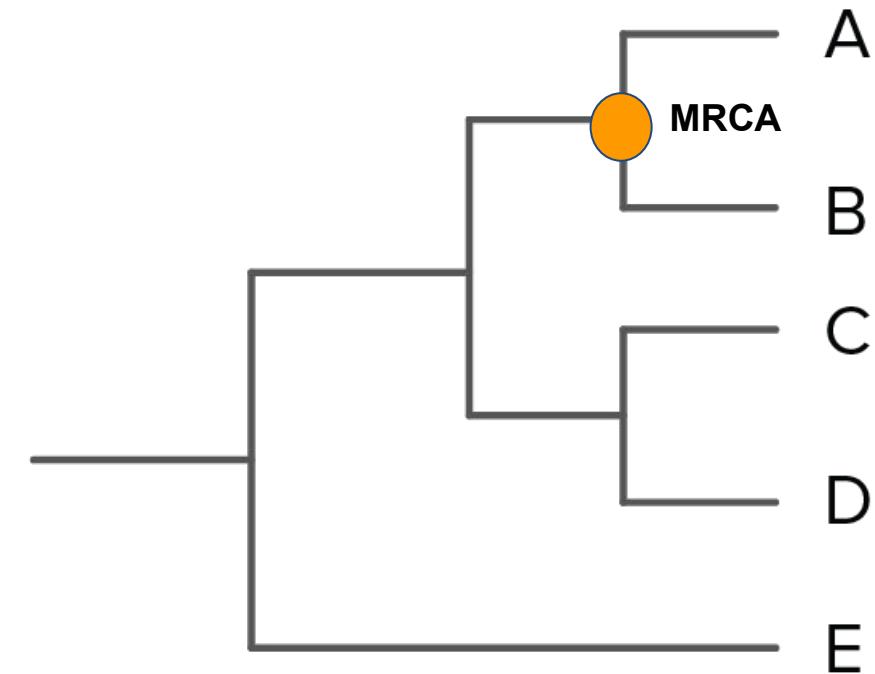
Parts of phylogenetic tree: Outgroup & Clades

- Outgroup: taxa that is outside the considered taxa to build the tree
- Clade: a group of leaves



Parts of phylogenetic tree: MRCA

- Node: represent the ancestor shared by one or more leaves
- It can be referred as: Most Recent Common Ancestor (MRCA)



Parts of phylogenetic tree: tMRCA

- Node: represent the ancestor shared by one or more leaves
- It can be referred as: Most Recent Common Ancestor (MRCA)
- The time to the MRCA is referred as tMRCA

