# Computation Practical 9: Analysing phylogenetic trees

**Module Developers:** Dr. Pakorn Aiewsakun and Dr Francesc Coll I Cerezo

## Table Of Contents

Expected learning outcomes ........ 1
Introduction to phylogenetic analysis ........ 2
Creating a multiple sequence alignment ........ 3
Estimating a phylogeny ........ 7
Rooting your tree ........ 12
Phylogenetic uncertainty estimation with bootstrapping ........ 13
How to read a phylogenetic tree? ........ 15
Relatedness ........ 15
Trait evolution ........ 19
Types of phylogenetic groups ........ 21
Detecting potential conflicting evolutionary signals within the MSA ........ 22
Phylogenetic network reconstruction ........ 25
Take-home messages ........ 26
Answers to exercises on interpreting phylogenetic trees ........ 27
Bibliography and references ........ 28

# Expected learning outcomes

After this practical session, you should be able to:

- familiarise with phylogenetic concepts and nomenclature;
- reconstruct a phylogeny from whole-genome sequences using a maximum likelihood framework;
- interpret a phylogeny and assess phylogenetic uncertainty;
- identify and mask recombination from whole-genome sequence alignments;
- understand how to handle conflicting evolutionary signals with phylogenetic network analysis;

# Introduction to phylogenetic analysis

***Phylogenetic analysis*** is central to many areas of modern microbiological research, from pathogen classification to examination of pathogen-host co-evolution and tracking of transmission of infectious pathogens. The main result from a phylogenetic analysis is a ***phylogeny***, also known as ***phylogenetic tree***, or simply ***a tree***, depicting evolutionary relationships of a set of taxa – these can be different species or different strains of the same species, as it is often the case in genomic epidemiology studies. **Figure 1** to familiarise with commonly used phylogenetic terminology (e.g., internal nodes, branches, etc,).

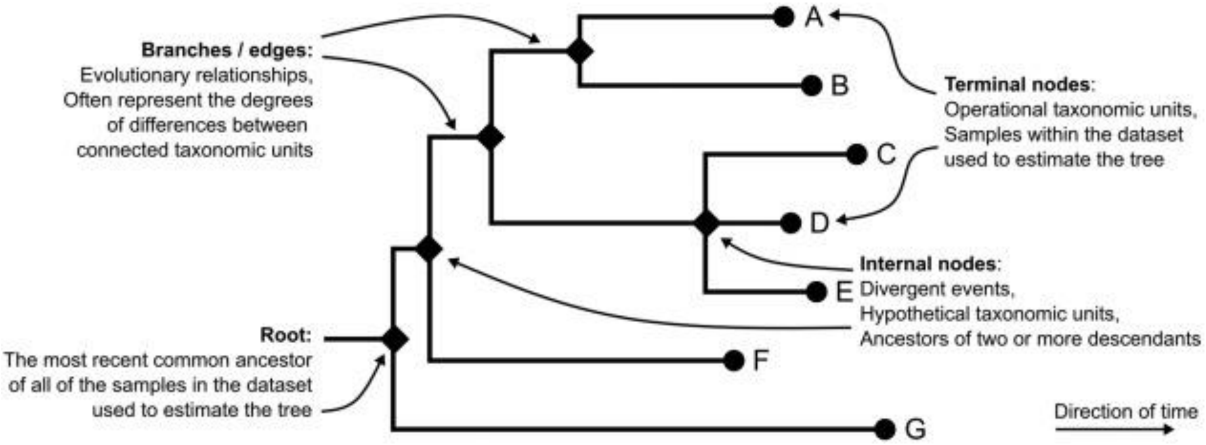

Branches / edges:
Evolutionary relationships,
Often represent the degrees
of differences between
connected taxonomic units
A
B
C
D
E
F
G
Terminal nodes:
Operational taxonomic units,
Samples within the dataset
used to estimate the tree
Internal nodes:
Divergent events,
Hypothetical taxonomic units,
Ancestors of two or more descendants
Root:
The most recent common ancestor
of all of the samples in the dataset
used to estimate the tree
Direction of time

**Figure 1 Basic phylogenetic terminology**. This figure has been reproduced from Aiewsakun 2024.

Ideally, we would like to have a complete knowledge of the entire genealogical history of the investigated taxa to draw their phylogeny, but such information is almost always impossible to ascertain. In practice, a phylogeny is commonly *estimated* from similarity of the investigated organisms' ***orthologous characters*** (i.e., biological features with their most recent diversification event coincides with that of the organisms bearing them), with the assumption that the more similar the characters, the more likely that the organisms bearing them would be more closely related (i.e., sharing a more recent common ancestor). Theoretically speaking, any orthologous characters can be used to reconstruct a phylogeny under this framework; however, researchers nowadays most commonly use DNA sequence data, such as genes or genome sequences, for this purpose, due to its relative ease of generation, curation, interpretation, transferability, and reproducibility. A phylogeny estimated from molecular sequences is commonly referred to as ***a molecular phylogeny***.

In the context of infectious disease epidemiology, a phylogenetic tree is commonly used to depict estimated evolutionary relationships between strains of the same bacterial species, which in turn can be used to track their transmission. Changes of population size can also affect how evolutionary changes accumulate on molecular sequences, and thus a phylogeny can be analysed to estimate epidemiology dynamics as well. While bacteria typically reproduce its genome with high fidelity (estimated at 1 in 10 million to 1 in a billion base substitutions per nucleotide per generation, very much comparable to that of us human really), random errors (i.e., ***mutations***) in DNA replication may still occur, and they can be passed down from one generation to the next. Some mutations are beneficial, while some can be neutral or deleterious.

As time goes by, the frequency of beneficial mutations that increase the bacterial ***fitness*** (i.e., increase the chance of survival and / or reproductive success of the bacteria in the environment that they find themselves in) will tend to increase in the population, while the frequency of deleterious mutations, i.e., those that decrease the fitness, will tend to decrease as bacteria carrying them will tend to be outcompeted, out-survived, and out-reproduced by others. This in turn will result in continuous change of genetic composition of the bacterial population over time. This process is known as ***evolution by natural selection***, first proposed by Charles Darwin in 1859. Although the mutation rate of bacteria is not high, given their typically short generation times and large population size, a bacterial population will nearly always have ***mutants*** around, allowing natural selection to operate rather efficiently. Combined with the usually strong selection pressure that bacteria face constantly, for example from the host immune response, antimicrobial drugs, and other competing organisms, the genetic composition of a bacterial population can and do change substantially within a relatively short amount of time. It is all of these features combined that make molecular phylogenetic analysis of bacteria populations feasible and meaningful.

A bacterial phylogeny is typically estimated from a (set of) orthologous gene(s), more commonly from ***polymorphic sites***, i.e., sites showing multiple forms of molecular variants within the population. The number and pattern of shared evolutionary changes between bacterial strains can be used to reconstruct their genealogical and evolutionary relationships. The advancement of sequencing technology nowadays also makes it possible to '***read***' the entire bacterial genome virtually within a day at an affordable price. This allows researchers to estimate a bacterial phylogeny from genes or polymorphic sites sampled across their whole genome, providing the ultimate level of resolution possible to discriminate between closely related strains, and in turn the finest disease transmission history.

# Creating a multiple sequence alignment

The very first and arguably the most critical step in reconstructing a phylogeny from molecular sequences, i.e., ***molecular phylogenetic reconstruction***, is to align molecular sequences (most often DNA sequences) of the studied organisms to create a ***multiple sequence alignment (MSA)***. With an MSA, we can then estimate the degree of organisms' (dis)similarity or infer the process of evolutionary changes to estimate their phylogenetic tree. It is very important to note that all positions within the MSA should be ***homologous positions*** to ensure that we compare *'like with like'*, otherwise the resultant tree will be meaningless.

In the context of genomic epidemiology investigations such like ours, a large number of homologous ***core genes***, i.e., genes that are shared among (almost) all strains, can often be readily identified and confidently aligned to generate an MSA, since the pathogens analysed tends to be very closely related. However, finding and aligning homologous sequences can be difficult when dealing with large and highly diverse organisms. When this is the case, this issue must be dealt with utmost care, making sure that there are as many correctly aligned positions (or conversely as few potentially misaligned positions) as possible, in order to obtain the most reliable and meaningful tree.

**Figure 2** illustrates the common workflow to generate an MSA from a collection of bacterial strains. Generally, bacterial DNA is extracted from a single colony picked from culture plates (therefore commonly referred to as '***isolate***'), followed by sequencing library preparation and whole-genome sequencing using rapid benchtop sequencers. Raw sequence data generated by sequencers are then processed using bioinformatic and genomic pipelines, which generally involve read cleaning, and mapping them to a reference genome to reconstruct the isolate's DNA sequence along the whole bacterial chromosome. *With good quality control*, mapping the reads of multiple sequenced isolates to the same reference genome is often the way to create an MSA while ensuring that all positions are homologous positions.

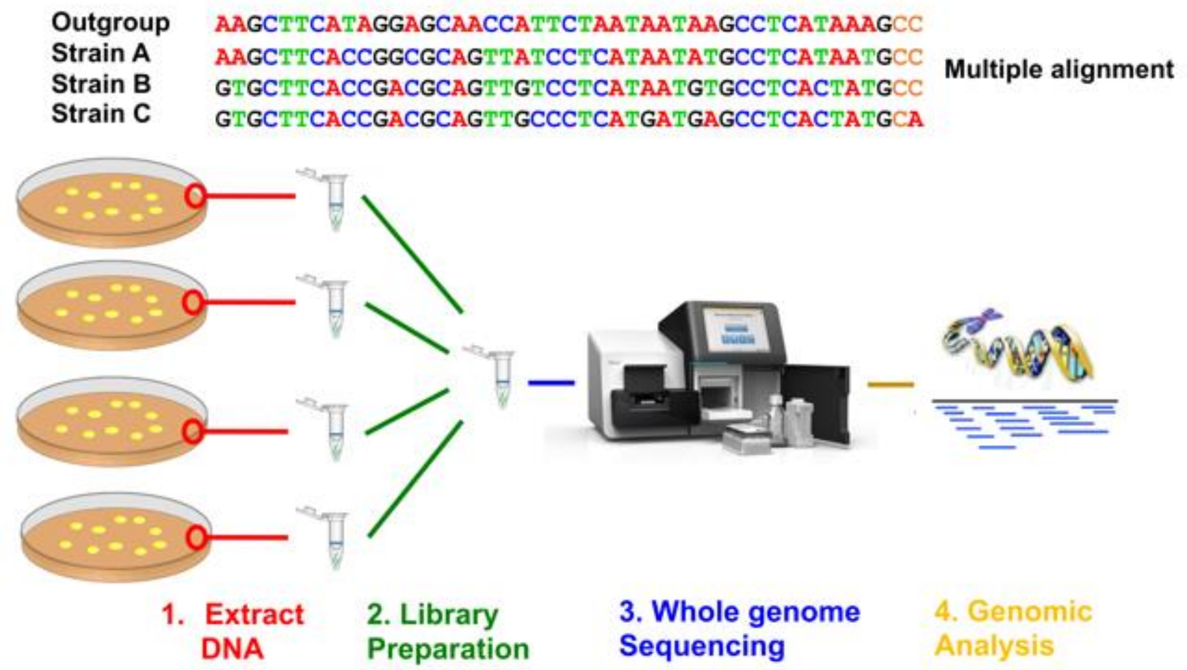

Outgroup
Strain A
Strain B
Strain C
AAGCTTCATAGGAGCAACCATTCTAATAATAAGCCTCATAAAGCC
AAGCTTCACCGGCGCAGTTATCCTCATAATATGCCTCATAATGCC
GTGCTTCACCGACGCAGTTGTCCTCATAATGTGCCTCACTATGCC
GTGCTTCACCGACGCAGTTGCCCTCATGATGAGCCTCACTATGCA
Multiple alignment
1. Extract DNA
2. Library Preparation
3. Whole genome Sequencing
4. Genomic Analysis

**Figure 2 Common workflow (bottom) to generate a multiple sequence alignment (top) from a collection of bacterial strains.**

Polymorphic sites in the MSA provide information useful for evolutionary relationship inference (i.e., the more the similar the sequences among '*orthologous*' polymorphic sites, the more likely they share a recent common ancestor), whereas monomorphic sites (nucleotide positions with all individuals showing the same molecular variants) are generally ignored (although it is a good practice to take these into account when specifying equilibrium base frequencies in the phylogenetic estimation). **Figure 3** shows how polymorphic sites may inform phylogenetic estimation.

In the previous session, we run `snippy` to map short Illumina reads of *Klebsiella pneumoniae* isolates obtained from a suspected hospital outbreak to the reference ST78 cpe058 genome, and called '***single nucleotide polymorphisms***' ('***SNPs***'). Now, we will attempt to construct a molecular phylogeny from these sequence data.

In this practical, you are provided with Snippy consensus sequences of the 14 investigated *K. pneumoniae* ST78 isolates. Each of these sequences was created by replacing SNPs and missing alleles along the reference sequence (ST78 cpe058 isolate). The command line below concatenates the files (not the sequences) to create one FASTA file, containing a '***whole-genome alignment***'.

```
cat "./snippy_files/"*.consensus.fa > Kpn_ST78.cpe058.fas
```

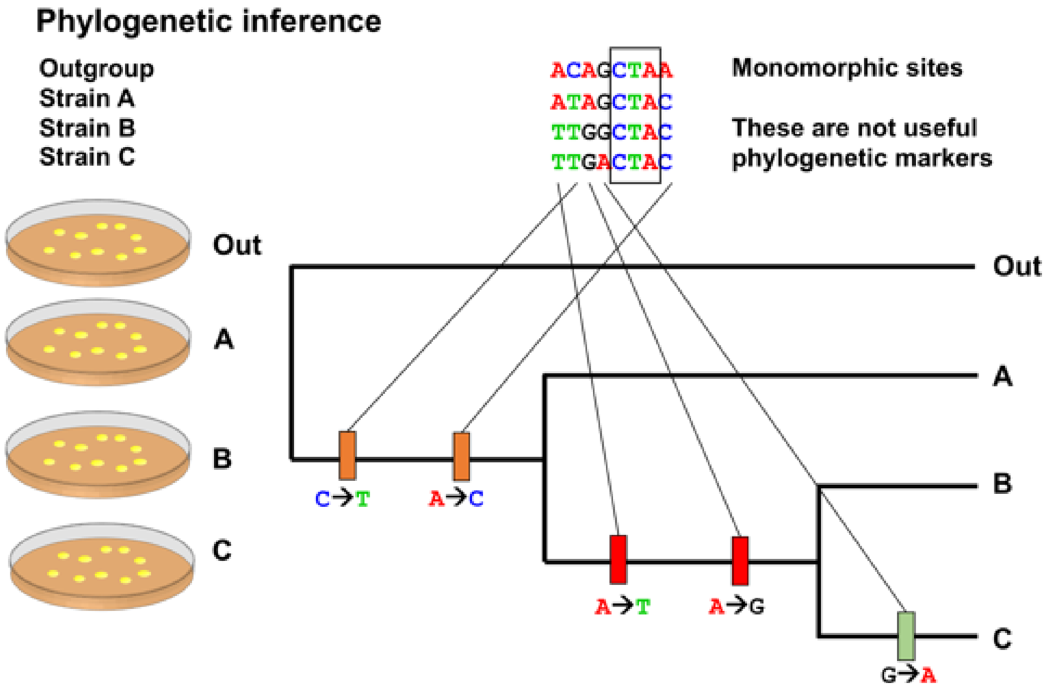


**Figure 3 How polymorphic sites may inform phylogenetic inference.** An example of a simple MSA of eight sites from four strains, which include monomorphic (squared) and polymorphic sites. Genetic changes in the phylogenetic tree are showed as coloured vertical rectangles on the branch where they originated. The identification of genetic changes (alleles) that are unique and common to multiple taxa (strains) are used to group them into phylogenetic clusters in a hierarchical manner with the goal of constructing the most plausible genealogical relationships between strains.

Next, it is good practice to run tools like `seqkit` to confirm the expected length of the MSA and the total number of sequences in it:

```
seqkit stats Kpn_ST78.cpe058.fas --all --gap-letters "- . N"
```

Next, replace Illumina run accessions in the MSA FASTA file with their corresponding strain ids:

```
python3 replace_fasta_ids.py -i Kpn_ST78.cpe058.fas \
-t Kpn_ST78.run_accessions.strain_ids.txt \
-o Kpn_ST78.cpe058.strain_ids.fas
```

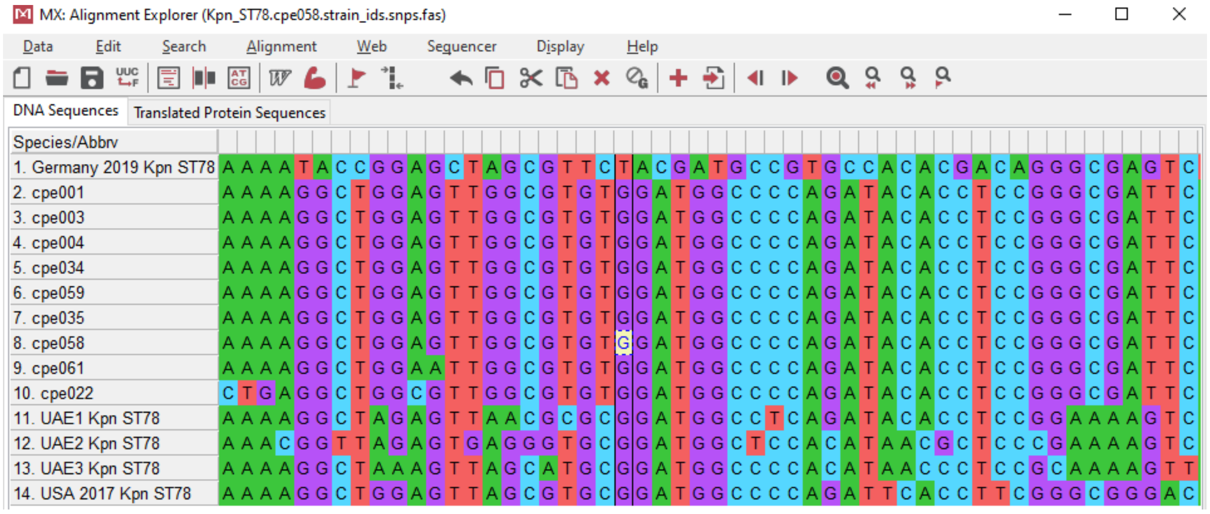
We can also manipulate the MSA to extract only SNP sites:

```
snp-sites   -c   -m   -o   Kpn_ST78.cpe058.strain_ids.snps.fas
Kpn_ST78.cpe058.strain_ids.fas
```

Finally, we will run *pairsnp*, a tool developed to quickly calculate SNP distances between all pairs of sequences within an MSA:

```
pairsnp      -c      Kpn_ST78.cpe058.strain_ids.snps.fas        >
Kpn_ST78.cpe058.pairsnp.csv
```

Now, let's have a look at your ***SNP alignment*** with `MEGA`. `MEGA` is a program with a graphical user interface, which is free and easy to use. You can examine your MSA in `MEGA` by dragging your MSA file to the program, and then the program will ask "*How would you like to open this fasta file*?". Click "`Align`", and you should see your alignment as shown in **Figure 4**.


MX: Alignment Explorer (Kpn_ST78.cpe058.strain_ids.snps.fas)
Data
Edit
Search
Alignment
Web
Sequencer
Display
Help
DNA Sequences
Translated Protein Sequences

| Species/Abbrv | |
|---|---|
| 1. Germany 2019 Kpn ST78 | AAAATACCGGAGCTAGCGTTCTACGATGCCGTGCCACACGACAGGGCGAGTC |
| 2. cpe001 | AAAAGGCTGGAGTTGGCGTGTGGATGGCCCCAGATACACCTCCGGGCGATTC |
| 3. cpe003 | AAAAGGCTGGAGTTGGCGTGTGGATGGCCCCAGATACACCTCCGGGCGATTC |
| 4. cpe004 | AAAAGGCTGGAGTTGGCGTGTGGATGGCCCCAGATACACCTCCGGGCGATTC |
| 5. cpe034 | AAAAGGCTGGAGTTGGCGTGTGGATGGCCCCAGATACACCTCCGGGCGATTC |
| 6. cpe059 | AAAAGGCTGGAGTTGGCGTGTGGATGGCCCCAGATACACCTCCGGGCGATTC |
| 7. cpe035 | AAAAGGCTGGAGTTGGCGTGTGGATGGCCCCAGATACACCTCCGGGCGATTC |
| 8. cpe058 | AAAAGGCTGGAGTTGGCGTGTGGATGGCCCCAGATACACCTCCGGGCGATTC |
| 9. cpe061 | AAAAGGCTGGAATTGGCGTGTGGATGGCCCCAGATACACCTCCGGGCGATTC |
| 10. cpe022 | CTGAGGCTGGCGTTGGCGTGTGGATGGCCCCAGATACACCTCCGGGCGATTC |
| 11. UAE1 Kpn ST78 | AAAAGGCTAGAGTTAACGCGCGGATGGCCTCAGATACACCTCCGGAAAAGTC |
| 12. UAE2 Kpn ST78 | AAACGGTTAGAGTGAGGGTGCGGATGGCTCCACATAACGCTCCCGAAAAGTC |
| 13. UAE3 Kpn ST78 | AAAAGGCTAAAGTTAGCATGCGGATGGCCCCACATAACCCTCCGCAAAAGTT |
| 14. USA 2017 Kpn ST78 | AAAAGGCTGGAGTTAGCGTGCGGATGGCCCCAGATTCACCTTCGGGCGGGAC |

**Figure 4 SNP alignment visualised in `MEGA`.**

***Q: How many positions are there in your MSA?***

***Q: Do you believe that all positions in your MSA are homologous positions? Why?***

***Q: Based on pairwise distances, can you already roughly group your organisms into distinct clusters? Is this consistent with what you already know about the bacterial isolates investigated?***

***Q: Let's have a closer look at your sequences, specifically their nucleotide composition. On the program's main page, click "MODELS" > "Compute Nucleotide Composition". Then a file navigation window will pop up. Locate and select your MSA file. The program will then ask you a couple more questions. Answer them based on your best understanding of the data. From the results, what is the average nucleotide composition of your SNP alignment? Compare your results with the actual nucleotide composition of the entire genome. How do they differ? Discuss.***

# Estimating a phylogeny

A large number of phylogenetic tree reconstruction methods have been developed, and based on their algorithmic natures and underlying philosophies, some of the most commonly used methods can be tentatively categorised into four large classes:

i) **The maximum parsimony methods**: this class of methods aims to identify the tree that requires the smallest number of evolutionary changes to explain the data.

ii) **The phenetic methods, also known as distance-based methods or distance-matrix methods**: methods of this class construct a tree diagram, or more precisely a ***phenogram***, by applying a hierarchical clustering algorithm to a dataset of overall organismal pairwise dissimilarity. Strictly speaking, a phenogram does not show evolutionary relatedness, and simply shows the degree of overall dissimilarity among a group of organisms. Nevertheless, in practice, it is often the case that a phenogram will closely resembles the true evolutionary tree as there tend to be a significant correlation between over similarity and evolutionary relatedness.

iii) **The maximum likelihood (ML) methods**: as the name suggests, this class of methods aims to identify an evolutionary model that gives the highest ***likelihood*** score, which numerically, is equal to the probability that the model will generate the observed MSA. In essence, these methods try to identify a set of evolutionary parameters that will most likely generate the observed MSA.

An evolutionary model can contain many parameters, including:

- tree topology (i.e., the order of diversification events),
- length of each branch in the tree (i.e., the number of molecular changes occurring along each branch),
- equilibrium molecular state frequencies,
- relative rates of molecular changes from one state to others (e.g., the probability that a base "A" will change to "T" within a certain among of time, etc.),
- rate variation among lineages / branches, and
- rate variation among sites within the MSA.

As you can imagine, given this large number of evolutionary parameters with each having its own many sub-parameters, and each sub-parameter can also have an enormous range of possible values, an evolutionary landscape that an ML method has to explore is vast, and it would be almost always impractical to explore the entire landscape of all possible models.

Heuristic search is almost always needed to find the ML solution. Conceptually, in a heuristic search, you start somewhere randomly within an evolutionary landscape with a presumably random set of evolutionary parameters and compute a likelihood score. Then, change the values of the investigated evolutionary parameters a little bit, and see if the score changes. If the score changes, you *'climb the hill'* by moving your model into the direction that increases the likelihood score the most. Repeat these steps until the score doesn't increase any further, at which point it means that you have reached the top of the hill, and there you get it, an ML solution! However, a naïve hill-climbing algorithm might get you stuck at a local optimal hill, but luckily several algorithms have been developed to overcome this issue (to some degree). With the continuous advancement of heuristic search algorithms and computer hardware, it is now possible to estimate a (close-to-)ML tree from an MSA of thousands of bacterial complete genomes within a practical and reasonable amount of time. Also, one may want to start at several random points within the evolutionary model landscape and see if all reach to similar models; if so that is great! It supports that your answer might really be the ML answer.

iv) **The Bayesian inference methods**: methods of this class are closely related to the ML methods, but instead of asking "*what is the evolutionary process that will most likely produce the observed MSA?*", these methods ask "*given a prior belief of how a set of organisms may evolve, represented by a set of possible (but may be unequally likely) phylogenies and evolutionary models, how is the presumed belief updated upon seeing the actual sequence data?*". This *'updated belief'* is more formally known as the ***posterior probability distribution*** of the models.

Philosophically, one may argue that these methods are the most attractive among all methods discussed so far; "*it computes what we most need, the probabilities of different hypotheses in the light of the data*" (Felsenstein 2004). Another very attractive feature of this class of methods is that, it is the only class of methods among the discussed four classes that intrinsically gives a distribution of models as an answer, and not just *'the best model'* like the other three, naturally accommodating result uncertainty assessment.

Nevertheless, there are some catches!

The main catch is that, as hinted earlier, one would need to have their own prior beliefs about all possible evolutionary models beforehand for the math to work — i.e., one would need to assign probabilities to all evolutionary models that one wants to explore *a priori* even before doing the analysis, which is very difficult to do sensibly, and different ***prior probability distributions*** of the models can lead to different outcomes. Many people opt to use the program's default settings for this, which is not the best thing to do. Similar to what suggested above for the ML method, several model prior probability distributions may also be explored to see if all inferences reach to the same / similar posterior probability distributions of the models. If so, then that's great. It means that your data contain sufficient information to update your beliefs in the same direction no matter where you start. Otherwise, it might indicate that it is your belief that drives the outcomes, and not the data.

Another catch is that methods of this class are also very computationally expensive — much more expensive than an ML method. In the ML method, you want just one '*best solution'*, but in a Bayesian method, you want to learn about the entire distribution of all possible models! As a result, what can be done within days by an ML method can take several weeks to complete with this approach even with a modern-day high-performance computer. So, for most common people, a Bayesian phylogenetic inference is only feasible with a relatively small dataset for now…

Here, we will use `IQ-TREE 2` to estimate an ML phylogeny from your whole-genome alignment. `IQ-TREE 2` is an open-source command line tool that can perform a variety of phylogenetic analyses.

To do this, open your computer '*terminal*', and type:

```
iqtree2 -s Kpn_ST78.cpe058.strain_ids.fas -T 4 --mem 4G --
ufboot 1000 --prefix Kpn_ST78.cpe058_iqtree -wbtl
```

That is it!

Spend some time exploring what each parameter mean—more information about `IQ-TREE 2` options can be found by typing `iqtree2 -h` on your terminal.

With this command, the program will automatically selects the '*best-fit nucleotide substitution model structure*' for you by using `ModelFinder` (see **Box 1**), infer an ML tree from your MSA file (as specified via the parameter `-s`) using the best-fit nucleotide substitution model, compute something called "clade support" based on 1,000 bootstrap MSAs (see below) using the ultrafast bootstrap approximation method (`--ufboot 1000`), and write the bootstrap trees into a file with branch length information (as specified by the flag `-wbtl`).

Behind the scene, this command also performs, under its default setting, tree searching with 100 initial parsimony trees (`--ninit 100`). 20 top initial parsimony trees are subsequently optimised with ML nearest neighbour interchange search to initialise the candidate set (`--ntop 20`), and the program then performs the ML tree search using the top 5 trees in the candidate set (`--nbest 5`).

***Q: Have a look at the output files. How many files are there? And what does each of them contain?***

Explore the output files obtained and identify the tree file we should take forward.

***Q: Given that it is SNP, and not constant, sites that are phylogenetically informative, why do you think we did not use the SNP alignment here, but the whole-genome alignment instead, which contains mostly monomorphic sites.***

**Box1: Nucleotide substitution model selection**

There are three major components to a nucleotide substitution model, including:

- **Equilibrium nucleotide frequencies**: the simplest model structure is to restrict all nucleotide bases to be the same (25%; +FQ), while the most relaxed structure is to allow all bases to have their own frequency (+F: empirical base frequencies or +FO: ML optimised base frequencies).
- **Relative transition rates among all 4 types of nucleotide bases**: most common nucleotide substitution models are time-reversible, meaning that the rate of base X changes to Y (X -> Y) is constrained to be equal to the rate of Y->X, reducing the number of parameters to just 6 (A-C, A-G, A-T, C-G, C-T and G-T) from 12 (A->C, A->G, A->T, C->A, C->G, C->T, G->A, G->C, G->T, T->A, T->C, and T->G). The Jukes and Cantor 1969 (JC69) model has the simplest model structure, assuming all

rates to be equal (and all four nucleotide frequencies to be equal as well), which is perhaps too simplistic / unrealistic for most cases. Other simple ones, but slightly more complicated, are the Kimura 1980 (K80) model and the Hasegawa, Kishino, and Yano 1985 (HKY85) model, which allow the rates of nucleotide transition (A-G, and C-T) and transversion changes (A-C, A-T, C-G, and G-T) to be different, with K80 assuming equal nucleotide frequencies while HKY85 allowing frequencies of different nucleotides to be different. The most general time-reversible model possible is called the Generalised Time-Reversible model of Tavaré 1986 or the GTR model, which allows nucleotide frequencies, and all of the six symmetrical rates of nucleotide changes to have different values. Note that `IQ-TREE 2` also has the unrestricted non-time reversible model, in which the rate X-> Y is allowed to be different from that of Y->X, and the frequencies of the four base types can also be different. This model structure is very flexible; however, in order for this model to be meaningfully estimated, the program must somehow know from the beginning the direction of time in your phylogeny (by including some outgroups, for example).

- **Distribution of rate variation among sites within the MSA**: Not all sites within the sequences evolve at the same rate. Some parts of a gene may evolve faster than others; for example, an active site of an enzyme, which is functionally very important, may be more conserved than other sites. When comparing several sequences to estimate a phylogeny, we should account for this rate heterogeneity to avoid errors. `IQ-TREE 2` supports site-wise rate heterogeneity modelling with various model structures. This includes:
    - "+I" model structure: allowing for a proportion within the MSA to be invariable sites.
    - "+G" model structure: modelling rate heterogeneity among sites by using the discrete Gamma model with default 4 rate categories (+G4).
    - "+I+G" model structure: invariable site plus discrete Gamma model.
    - "+R" model structure: FreeRate model that relaxes the assumption of Gamma-distributed rates, with default 4 rate categories (+R4).
    - "+I+R" model structure: invariable site plus FreeRate model.

Combined, a nucleotide substitution model may therefore be something like "GTR+F+I+R6", which means that:

- The relative transition rates among "A", "T", "C", and "G" base types are allowed to be different, but all are constrained to be time reversible (GTR), and…
- …their equilibrium state frequencies are taken to be equal to the empirical frequencies directly observed from the MSA (+F), and …
- … sites within the MSA are allowed have unequal overall rates of change, with 6 different rate categories (+R6), but …
- … some sites may not change at all (+I).

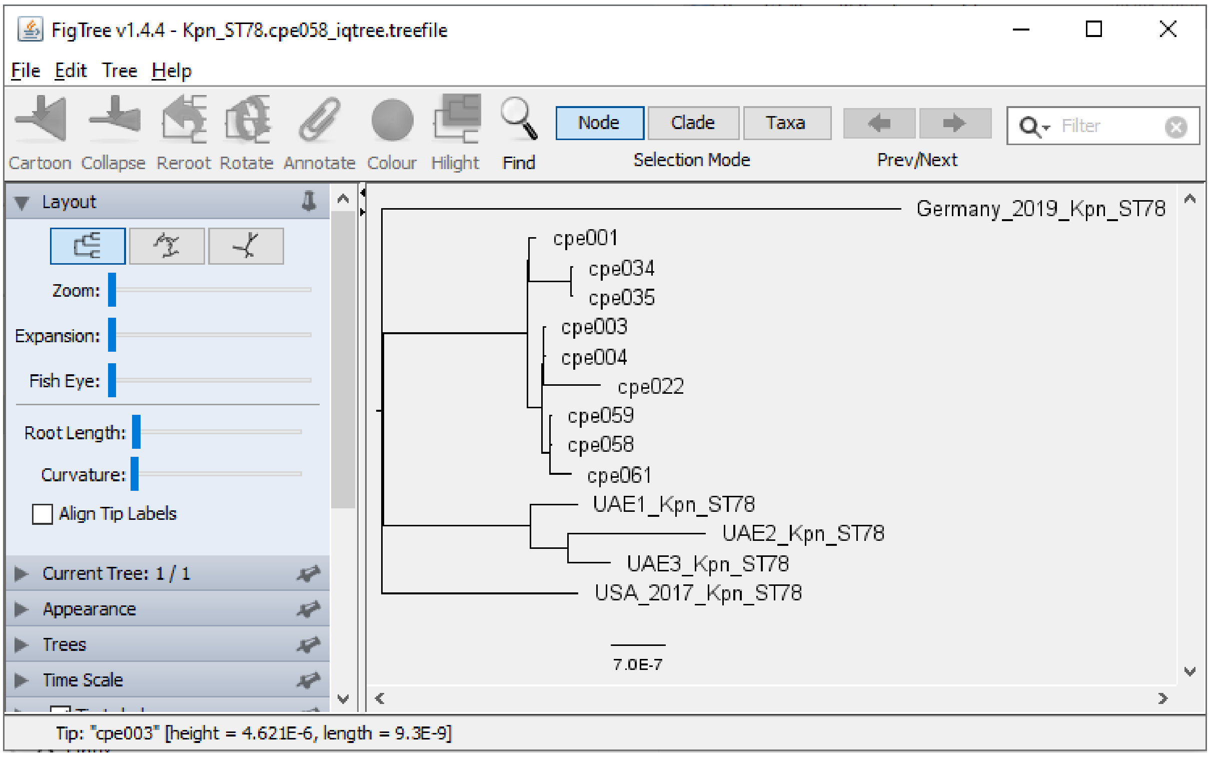
There are many more substitution models supported by `IQ-TREE 2`, such as Lie Markov models, codon models, or even protein models or binary and morphological models. You can learn more about them from http://www.iqtree.org/doc/Substitution-Models.

`IQ-TREE 2` can automatically perform nucleotide substitution model selection for you, with its built-in program `ModelFinder`. For each model structure, `IQ-TREE 2` will try to estimate ML parameter values, and compute AIC (Akaike information criterion), AICc (small-sample-size-corrected AIC), and BIC (Bayesian information criterion) scores for each of them. All of these scores take into consideration how likely the model will produce your data (the likelihood score), and the number of parameters within the model, but they are slightly different in term of how the number of parameters is penalised. Generally, the lower these scores, the more preferred the model. Under the default setting, the BIC score is used to decide the best-fit model structure (`--merit BIC`), i.e., the one with the lowest BIC score.

***Q: What is the best fit nucleotide substitution model for your MSA? Can you make sense of it?***

Now, let's have a look at your tree in `Figtree`. `FigTree` is a program developed specifically for phylogenetic tree visualisation. And just like `MEGA` and `IQ-TREE 2`, it is free, and is quite easy to use!

To use `FigTree` to explore the tree you just made, first launch the program by simply double clicking the program icon. Click the menu "`File`" on the program's menu bar, and select "`Open…`", and a file navigation system should pop up. Search for your tree file, and then click "`Open`" to import the tree to the program. The program will alert you that "*The node/branches of the tree are labelled. Please select a name for these values.*" These are the bootstrap clade support values (see below). So, let us set the name of these labels to "`Clade support`" and click "`OK`". The program should then show you your tree (**Figure 5).**



Tip: "cpe003" [height = 4.621E-6, length = 9.3E-9]

**Figure 5 Our phylogenetic tree visualised in `FigTree`.**

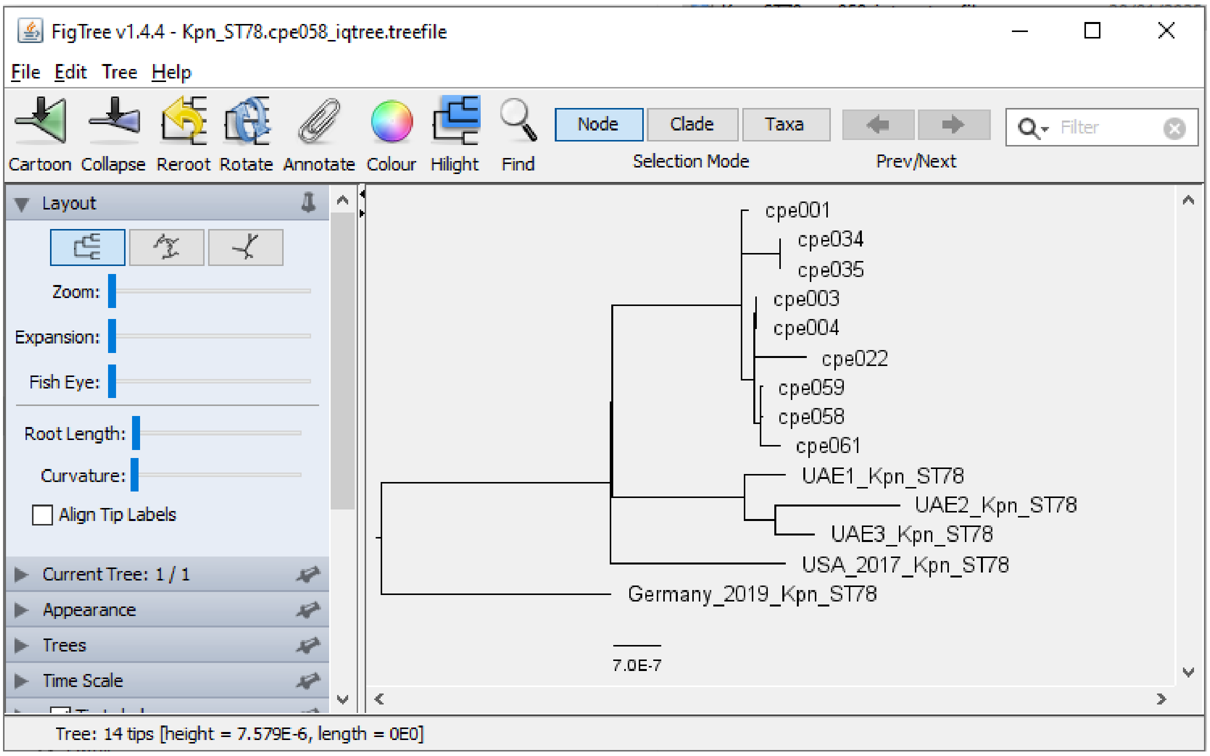
***Q: Inspect your ML tree using `Figtree`. Does the tree look alright in your opinion? How are strains clustered on the tree?***

# Rooting your tree

For a phylogeny to show genealogical relationships, which intrinsically have a time direction, you need to first ***root*** your tree (for further details, see "***How to read a phylogenetic tree?***" below). With just an MSA as input, `IQ-TREE 2` doesn't root the tree for you, and it actually cannot, as there is no temporal information to be found at all in your MSA—your MSA can only provide information about genetic dissimilarity among a set of organisms, and not at all about the direction of molecular changes or time. This is actually why most common nucleotide substation models are time-reversible, and why `IQ-TREE 2` can't give a rooted tree (without extra information) to you. An '***unrooted tree***', also known as an '*affinity tree*', simply indicates the overall degrees of dissimilarity among the depicted taxa.

There are several ways to root an unrooted tree. Here, we have included Germany_2019_Kpn_ST78 as our ***outgroup***, and we will use this information to specify the direction of time in our phylogeny. An outgroup is simply a group of organisms that you know for certain from some other sources of information that they are genetically related but not part of your ***in-group***, in our case, the outbreak strains investigated.

In `Figtree`, open our tree file, and to root our tree with strain Germany_2019_Kpn_ST78, simply click on its terminal branch and click the "`Reroot`" icon on the main menu bar. Now your tree should look like the one shown in **Figure 6**.



Tree: 14 tips [height = 7.579E-6, length = 0E0]

**Figure 6 Rooted phylogeny.**

***Q: Why do you think the strain 'Germany_2019_Kpn_ST78' was chosen as an outgroup? Hint: observe the SNP distances derived by pairsnp on your MSA.***

In fact, we could have told `IQ-TREE 2` that Germany_2019_Kpn_ST78 is our outgroup by adding '`-o Germany_2019_Kpn_ST78`' to our command. However, the program will still produce an unrooted tree, and simply drawing Germany_2019_Kpn_ST78 at the "root" (see the .iqtree file). Furthermore, without specifying an outgroup, the program will use the first sequence in the MSA file to 'soft-root' the tree default. So, in our case where Germany_2019_Kpn_ST78 is the first sequence in the file anyway, this option wouldn't make a difference. You can try it out!

# Phylogenetic uncertainty estimation with bootstrapping

It is important to appreciate that phylogenetic inference is inherently statistical, guided by only a limited number of organisms that we could get our hands on, and done under a large number of statistical and biological assumptions. Thus, uncertainty in the estimated tree should always be taken into consideration when describing inferred phylogenetic groups (and evolutionary relationships in general).

However, as previously mentioned, an ML method only provides a single *'best-fit evolutionary model'* that most likely generates the observed data (your MSA in this case). So, how can we estimate an uncertainty associated with this one ML solution?

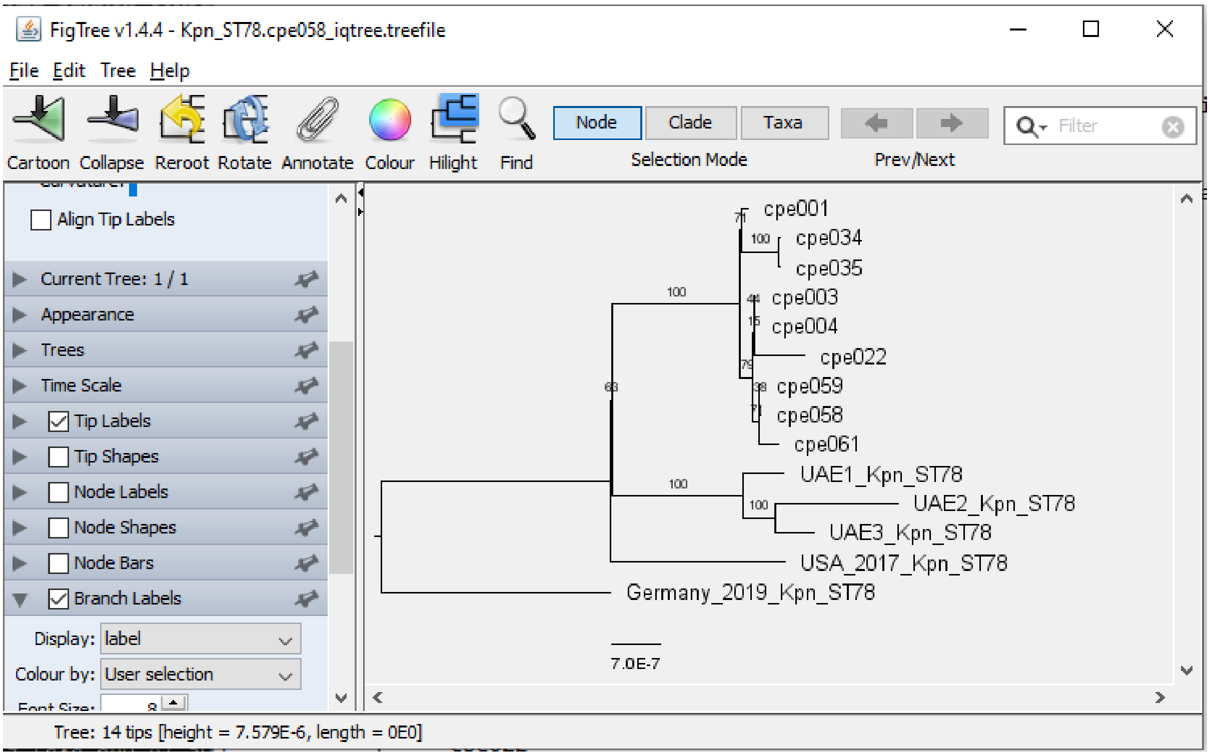
In this case, we can use the '***bootstrap***' method, invented by Bradley Efron in 1979. The term is derived from the phrase *'pull yourself up by your bootstraps'*, meaning *'to improve one's situation through hard work and self-determination, rather than getting assistance from someone else'*.

In statistics, bootstrapping is a technique to assign variability to sample estimates, relying on ***random sampling with replacement***. In our context, it is our MSA that will be bootstrapped, that is, resampled by randomly selecting sites (i.e., columns) with replacement to construct a (pseudo)replicate alignment of the same length. This means that some columns may be sampled multiple times, while some columns would not be sampled at all since the sampling is with replacement. Thus, the resulting MSA would be slightly different from the original data. This sequence alignment is called a ***bootstrap replicate*** or sometimes a ***pseudoreplicate dataset***. We can then perform the same analysis on this bootstrap sample to get a ***bootstrap tree*** (and other associated parameters). Repeat this process a large number of times (typically 1,000 times, or more) until a population of bootstrap trees (and other evolutionary parameters) is obtained. Once we have a distribution of bootstrap trees, we can then compute ***bootstrap clade support values*** (also known as ***branch support*** values) for each clade in your main tree — classically, a bootstrap clade support value is defined as *'the proportion of the trees in the bootstrap distribution showing that clade'*, first proposed by Felsenstein in 1985. Robust relationship should be repeatable, and subsequently observed in a large proportion of the bootstrap data. Therefore, if you get 100 out of 100 times for a particular clade, you may therefore conclude that the inferred clade is unlikely due to chance, and a real one. A clade with a bootstrap support value of, say, 80 means that 80% of the trees in the bootstrap distribution contain that clade.

One of the most common misconceptions about clade support values is that they indicate the confidence in the '*split*' at the base of the clade. This is false — the number does not say anything about the branching structure of or within the clade it refers to. This confusion perhaps stems from the fact that support values are sometimes shown next to the nodes on the end of branches, which represent diversification events. Thus, if possible, it is best to display branch support values on branches, and not on nodes, to avoid propagating this confusion.

One important thing to note here is that, according to the `IQ-TREE 2` developer, the ultrafast bootstrap (UFBoot) method implemented in `IQ-TREE 2` is less biased compared to the classic (and more conservative) Felsenstein's bootstrap support value described above — while a clade with a Felsenstein's support value of 75–80% is typically already considered *'well-supported'*, it is recommended that a clade should be regarded as well-supported when its UFBoot support is ≥95%, corresponding roughly to a probability of 95% that a clade is true. Learn more about this here: http://www.iqtree.org/doc/Frequently-Asked-Questions#how-do-i-interpret-ultrafast-bootstrap-ufboot-support-values.

Now, let's have a look at the clade support in your tree. To display branch support on the tree in `Figtree`, check the "`Branch Labels`" box, and select "`Clade support`" under the "`Display`" option (**Figure 7**). If you find the numbers too small, you can increase the font size by adjusting the number in the "`Font Size`" box.



Display: label

Colour by: User selection

Font Size: 8

Tree: 14 tips [height = 7.579E-6, length = 0E0]

**Figure 7 Tree with clade support.**

***Q: What is the clade support value for each of the main distinct clades you see on your tree?***

# How to read a phylogenetic tree?

A ***phylogeny*** is a diagrammatic hypothesis that depicts relationships, or relatedness, of a set of biological entities. **Figure 1** above illustrates what a typical phylogeny might look like.

Mathematically speaking, a phylogeny is simply a '*tree'* graph that contain a set of ***nodes*** connected together by a set of ***branches*** or ***edges***. To illustrate genealogical relationships, which intrinsically has a time dimension, requires that the tree graph is directed, having all branches pointing away from the ***root node***. This in turn allows us to organise the genealogical relationships in a nested hierarchical fashion. Nodes at the tips of the tree are called ***terminal nodes***, or ***leaves***, or simply ***tips***, representing individual organisms used to reconstruct the tree. An ***internal node*** in a ***rooted tree***, in which the direction of time is well defined, can be thought of as a ***divergence event***, where a single taxon diversifies to give rise to two or more descendants, who may further diversify down the tree to give rise to more descendants, and so on. An internal node with descendants can thus be thought of as ***the most recent common ancestor (MRCA)*** of all of its subsequent descendants as well, with the root node representing the MRCA of all of the organisms in the tree. The root is the oldest node in the tree, defining the direction of time or the flow of genetic information in the tree, moving away from the root towards the tips. A branch connecting two organisms together indicates their evolutionary relatedness, representing the '*line of decent'* or the path of vertical transmission of genetic information from one organism (parent) to the next (descendent). One can also think of a branch as a continuous chain of organisms linking by (imperfect) reproduction. ***Branch length*** is *typically* drawn proportionally to the degree of differences between the two connected organisms, usually in the units of '*years'* or '*substitutions per site'* — the longer the branch, the larger the amount of changes or divergent time it represents. In such cases, a ***scale bar*** is often provided to give a scale for the branch length (instead of labelling each branch with its length). Note that when we talk about branches, only the lines along the time axis (**horizontal lines**, **Figure 1**) count. Lines perpendicular to the time axis (**vertical lines**, **Figure 1**) have no meaning — they do not count as parts of the branches, and are simply there to lay out the tree visually so that the labels and branches do not overlap on top of each other. Several terms can be used to refer to the general overall pattern of diversification process, including ***branching order***, ***branching pattern***, or simply ***tree topology***.

## Relatedness

One key concept in phylogenetics is ***relatedness.*** *Relatedness is determined by how recently organisms share a* ***common ancestor*** *— the more recently they share a common ancestor, the more closely related they are*. We will explore an example below from the family tree of the organisms A to F (**Figure 8**) to see how we can describe evolutionary relationships from a phylogeny.

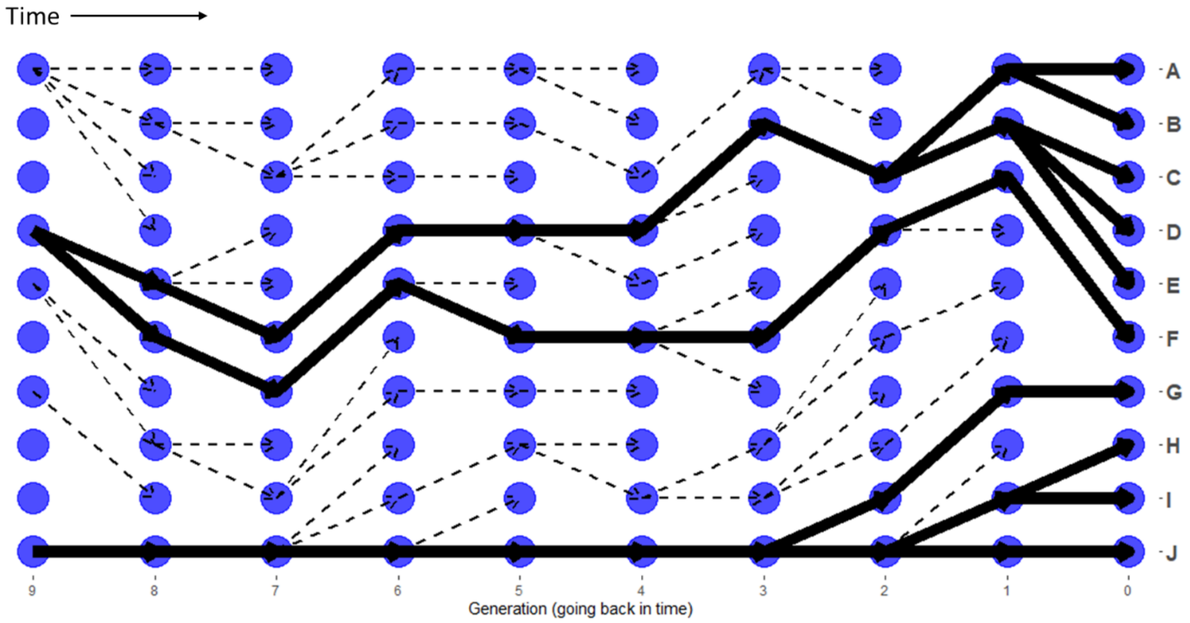


**Figure 8. Phylogenetic trees of hypothetical asexual organisms.** Each blue circle represents an individual organism. The direction of time is from left to right, and arrows indicate the direction of the relationship between individuals; parent is at the base and its descendant is at the tip. Lineages of organisms that manage to survive up until the 10th generation are drawn with thick solid arrows, otherwise drawn with thin dashed arrows.

Based on their family tree, we can say that:

- *"A and B share the MRCA one generation ago."*
- *"C, D, and E also share the MRCA one generation ago."*
- *"A, B, C, D and E share the MRCA two generations ago, and this ancestral organism exists prior to the MRCA of A, and B, and the MRCA of C, D, and E."*
- *"A is more closely related to B than C, since A shares a common ancestor more recently with B than it does with C."*
- *"In fact, since A and B are more closely related to each other than any other organisms on the tree, A and B are said to be* ***sister groups****."*
- *"Although B is depicted closer to C than A is to C on the family tree, both A and B are equally related to C since they both last share a common ancestor with C two generations ago."*
- *"The* ***lineage*** *of A, B, C, D and E could be traced back to the very first generation in the simulation (nine generations ago), when the ancestral organism diversified into two distinct lineages with the other being the ancestral lineage of F."*

Notice that we can make all these statements only because we know precisely the direction of time (i.e., that our tree is ***rooted***), making clear which events come before or after in the evolutionary history of the organisms.

One important thing to realise is that, in reality, unlike those shown in **Figure 8**, a phylogeny typically depicts a very much incomplete history of the organisms based on which it is reconstructed, showing only lineages of the organisms used to reconstruct the tree (**Figure 9**). Also bear in mind that ***a branch is a continuous chain of organisms that are linked together by the process of (imperfect) reproduction, and (multiple) evolutionary changes may occur***

***along the way (and not necessarily at a node, of which the main purpose is to depict a diversification event)***. Many people often forget this point as ancestral individuals along the branches, extinct lineages, and evolutionary changes are not usually shown on a phylogeny (**Figure 9**). Indeed, introductory students / researchers may mistakenly project the organisms from the tips backward in time to occupy internal nodes as if no changes have had occurred at all along the branches, which is inappropriate.

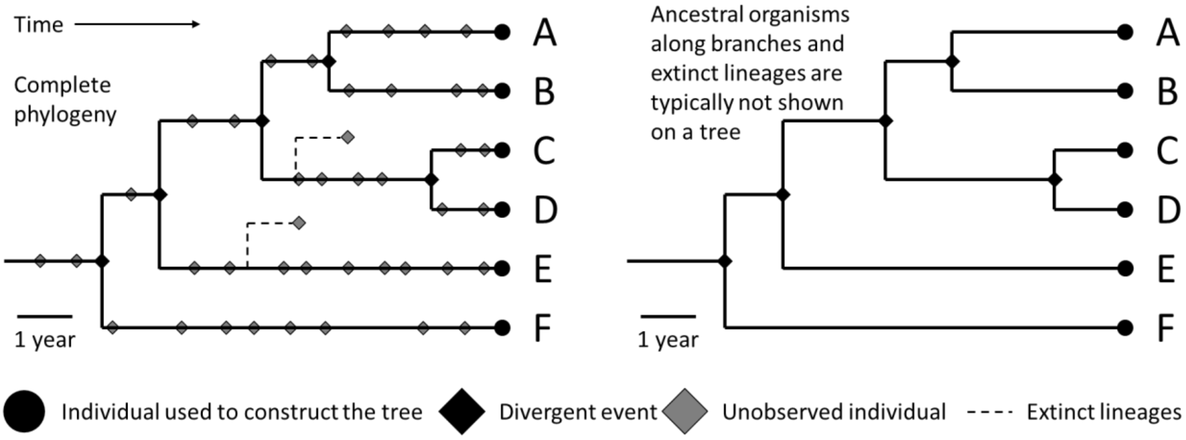

Time
Complete phylogeny
A
B
C
D
E
F
1 year
Ancestral organisms along branches and extinct lineages are typically not shown on a tree
A
B
C
D
E
F
1 year
● Individual used to construct the tree
◆ Divergent event
◇ Unobserved individual
---- Extinct lineages

**Figure 9. A phylogeny usually features an incomplete evolutionary history of the organisms used to reconstruct the tree, and ancestral organisms along branches, extinct lineages, and evolutionary changes, are usually not drawn on the tree.**

For example, based on the tree shown in **Figure 9**, it would be incorrect to say that the organisms A, B, C and D descended from E, when all the tree implies is that the organisms A, B, C and D are more closely related to each other than to E, and that the five organisms simply share a common ancestor at some point in the past.

Likewise, it would be incorrect to say that F is most similar to the root node, when the tree in fact shows that all of the six organisms are equally distantly related to the root (in the unit of time).

It is also incorrect to say that F evolved earlier than other taxa, when the tree implies that all organisms took the same amount of time to get to where they are on the tree.

Also remember that a phylogenetic tree is read along the time axis, and not across the tips of the tree. ***Relatedness is not about proximity of the names of the organisms on the tree, but how long ago they last share a common ancestor***. Therefore, a tree can be drawn in many ways — as long as they depict the same evolutionary relationships (i.e., showing the same tree topologies and branch lengths), they are the same trees (**Figure 10**).

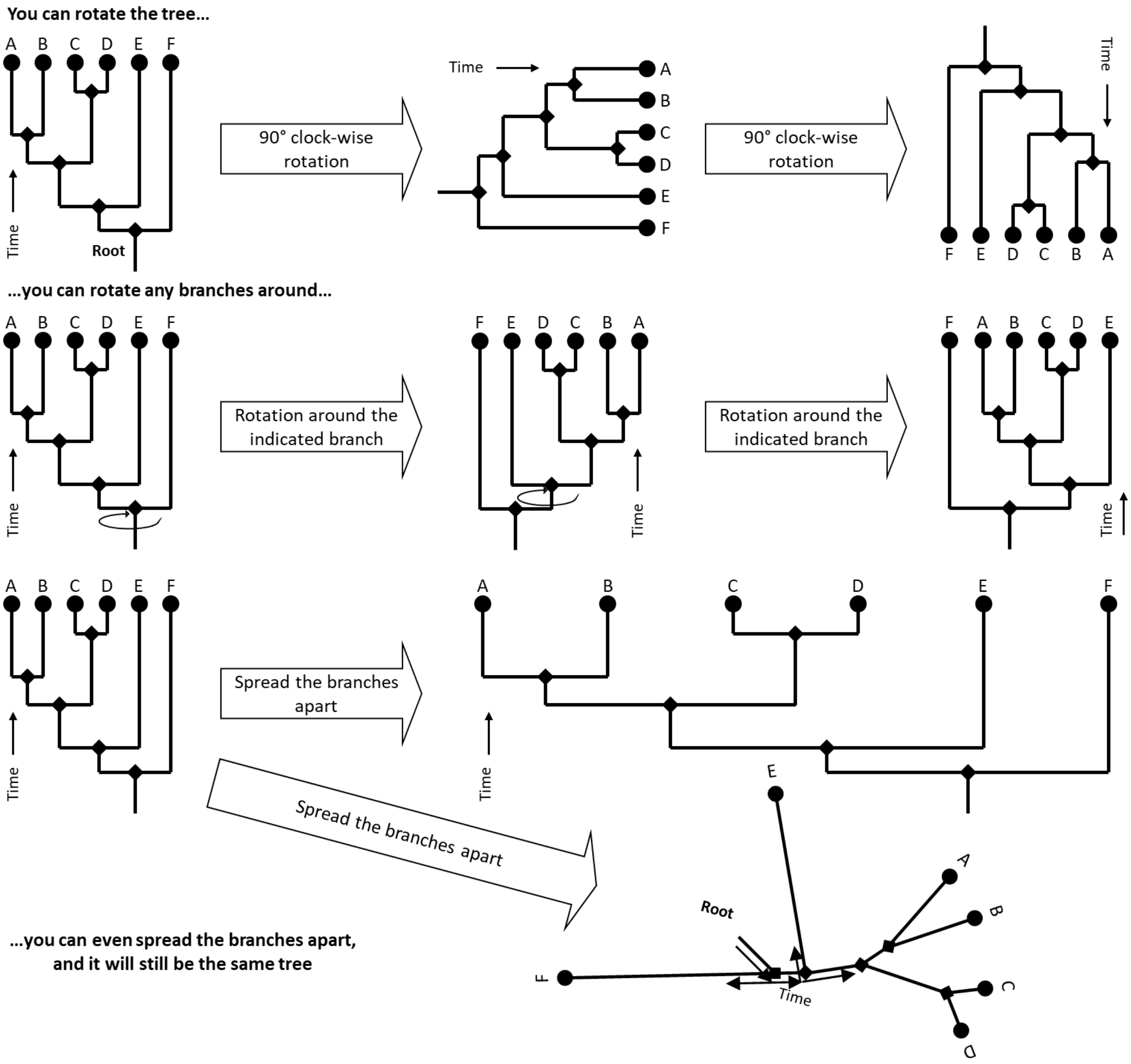

You can rotate the tree…
A B C D E F
Time
Root
90° clock-wise rotation
Time
A B C D E F
90° clock-wise rotation
Time
F E D C B A
…you can rotate any branches around…
A B C D E F
Time
Rotation around the indicated branch
F E D C B A
Time
Rotation around the indicated branch
F A B C D E
Time
A B C D E F
Time
Spread the branches apart
A B C D E F
Time
Spread the branches apart
E
A
B
C
D
F
Root
Time
…you can even spread the branches apart, and it will still be the same tree

**Figure 10. A tree can be drawn in many ways.** As long as they depict the same tree topologies and branch lengths, they are the same trees.

Now it is your turn to do some exercises! Answer the questions below based on this rooted tree shown in **Figure 11**.

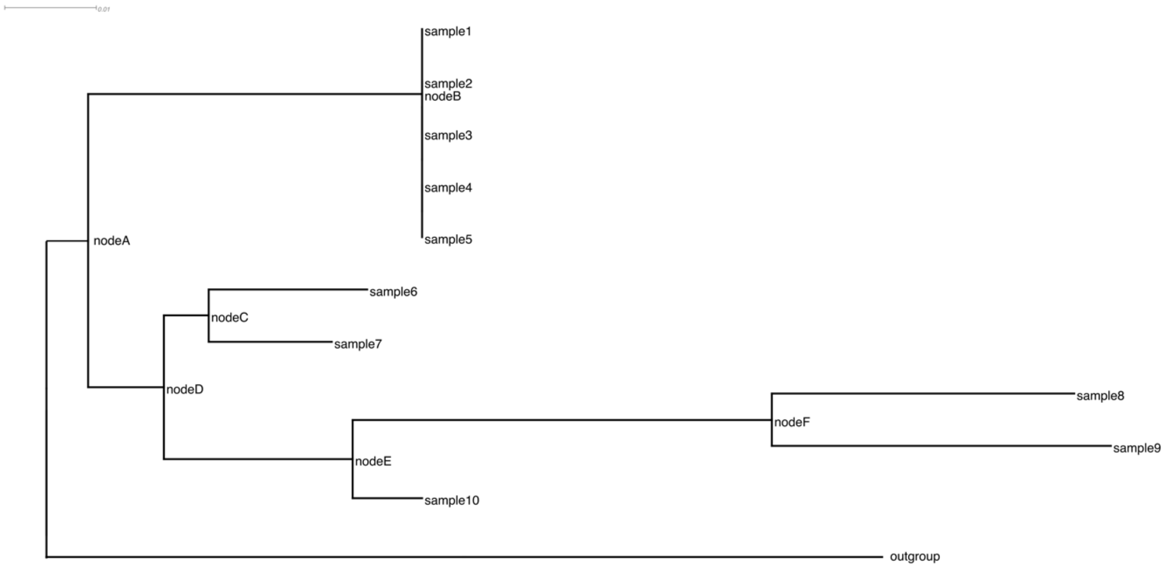


**Figure 11 A hypothetical tree of 10 samples and 1 outgroup**

Question 1. Based on the tree above, what internal node corresponds to the MRCA of samples 8 and 10:

- Node F
- Node D
- Sample 7
- Node E

Question 2. Based on the tree above, which group of samples are most closely related:

- Samples 1 to 5
- Samples 6 & 7
- Samples 6 to 10
- Samples 8 & 9

Question 3. Based on the tree above, which of the following statements referring to sample 10 is most accurate:

- Sample 10 is more closely related to sample 7 than to sample 8
- Sample 10 is more closely related to sample 8 than to sample 7
- Sample 10 is equally related to sample 7 and sample 8
- Sample 10 is related to sample 8, but it is not related to sample 7

Question 4. Based on the tree above, which of the following statements referring to sample 7 is most accurate:

- Sample 7 is more closely related to sample 8 than to sample 10
- Sample 7 is more closely related to sample 10 than to sample 8
- Sample 7 is equally related to sample 8 and sample 10
- Sample 7 is related to sample 8, but it is not related to sample 10

# Trait evolution

In molecular phylogenetic analysis, we often have information about our samples beyond their molecular sequences, such as their taxonomic group, sampling locations, biological features, etc. One thing we can do with a tree is to use it as a scaffold to unite all of this information under a single framework. Statistical methods can then be applied to infer how ancestral organisms might have looked like in the past. Phylogenetic structures of the investigated ***traits*** can then be examined to learn more about their past evolutionary history beyond the relatedness of the organisms. Such an analysis falls within the realm of ***phylogenetic comparative study*** — *a study of an evolutionary process from a combination of phylogenetic and phenotypic data.*

In the context of infectious disease epidemiology, a phylogenetic tree is commonly used to identify where person-to-person transmission occurs; to identify the sources and study the transmission routes of outbreak and epidemic clones; and to determine whether bacterial clones are restricted to specific hosts and settings or, on the contrary, able to circulate among multiple ones. A common phylogenetic method used to study how bacterial traits evolved is ***ancestral state reconstruction*** (**Figure 12**).

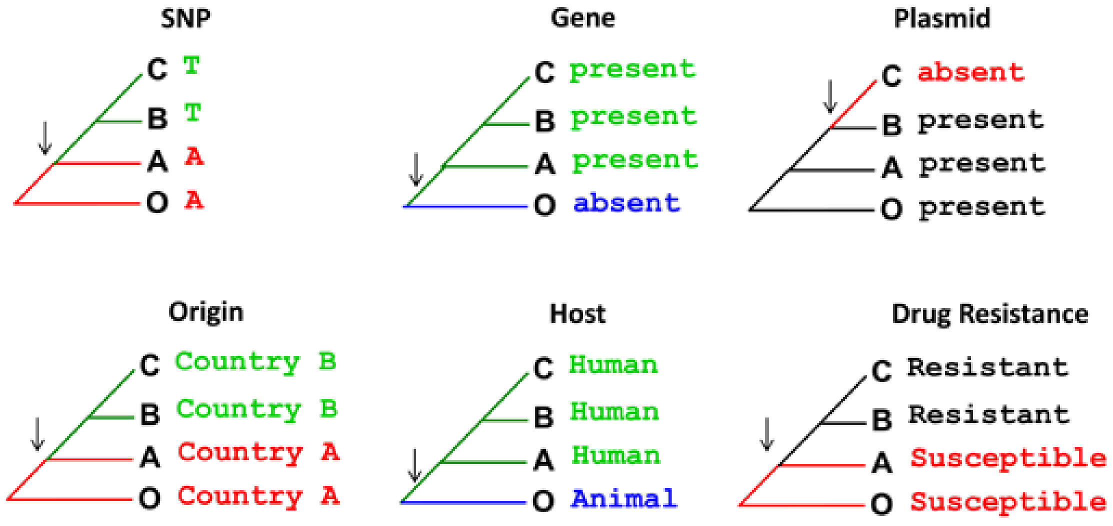


**Figure 12 Traits evolutionary study by ancestral state reconstruction.** Strains on the same tree are labelled based on the presence of different traits. Arrows indicate what internal node (ancestor) in the tree most likely changed (lost or gained) such a trait. Bacterial traits we may be interested in reconstructing include: geographical location — to then identify movement between regions / transmission events; colonising or infecting host — to enable us to identify host jumps; and antibiotic susceptibility — to enable us to identify evolution of antimicrobial resistance. The emergence and spread of individual mutations, genes and mobile genetic elements can also be reconstructed in a bacterial phylogeny using this method. © Wellcome Genome Campus Advanced Courses and Scientific Conferences

Ancestral state reconstruction is a difficult subject. One simple way to do this is to use ***the principle of parsimony*** to infer the most likely scenario, positing that *the most preferred trait evolutionary model is the simplest one, involving the smallest total number of trait changes required to explain the data.*

Now, apply this concept to the rooted tree shown in **Figure 13** with tips annotated with sampling locations to infer geographical locations at each internal node, and answer the following questions.

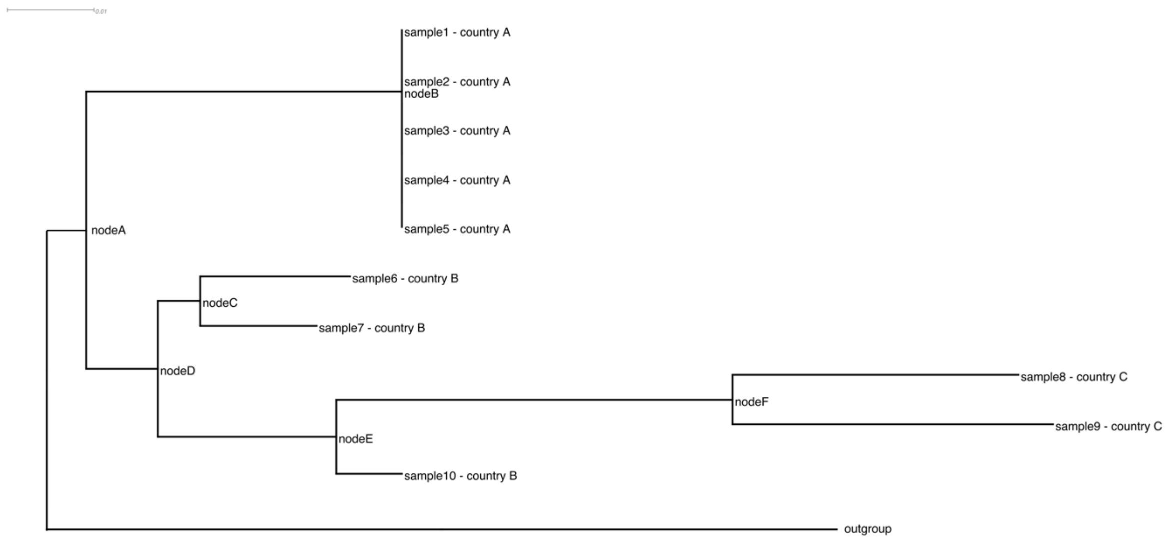


**Figure 13 A hypothetical tree of 10 samples and 1 outgroup with tips annotated with sampling location**

Question 5. Based on the country of origin of samples on the tree above, which of the following statements about transmission events is more certain:

- The common ancestor of samples 6 to 10 (node D) most likely circulated in country A first and later on transmitted to country B and C
- The common ancestor of samples 6 to 10 (node D) most likely circulated in country B first and later on transmitted to country C
- The common ancestor of samples 6 to 10 (node D) most likely circulated in country C first and later on transmitted to country B
- The common ancestor of samples 6 to 10 (node D) could have circulated in country A or B

Question 6. Based on the country of origin of samples on the tree above, which of the following statements about transmission events is more certain:

- The common ancestor of samples 1 to 10 (node A) most likely circulated in country A first and later on transmitted to country B and C
- The common ancestor of samples 1 to 10 (node A) most likely circulated in country B first and later on transmitted to country A and C
- The common ancestor of samples 1 to 10 (node A) most likely circulated in country C first and later on transmitted to country A and B
- The common ancestor of samples 1 to 10 (node A) could have circulated in country A or B

# Types of phylogenetic groups

In the case that you have a classification scheme, another thing you can do with your tree is to examine phylogenetic structures of your taxa. Under a phylogenetic framework, we can categorise organismal taxa into three classes (**Figure 14**):

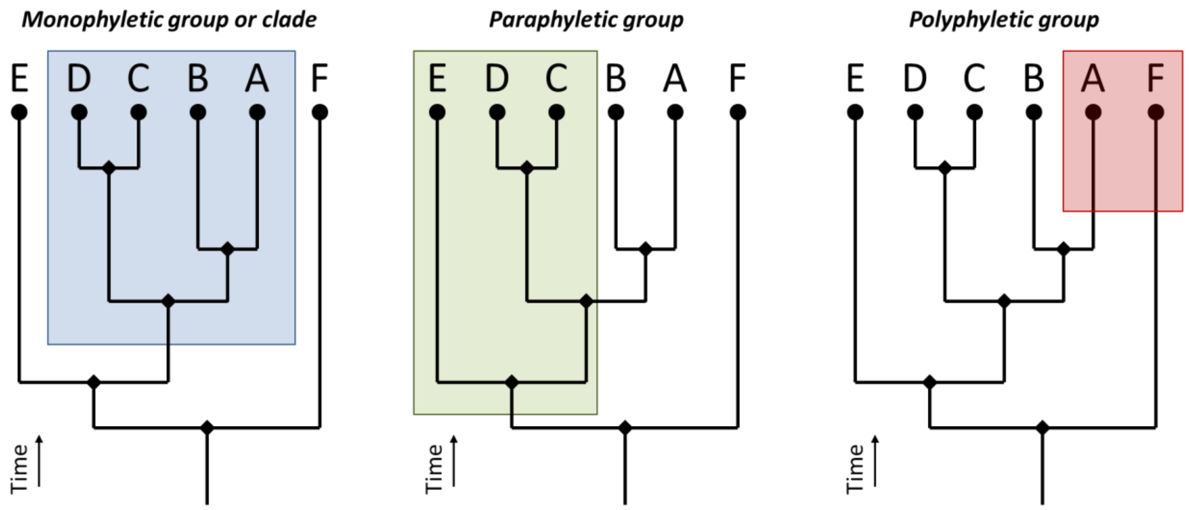


**Figure 14. Various types of phylogenetic groups.**

- ***Monophyletic group***: a group of an organism and all of its descendants. A monophyletic group is also known as a ***clade***.
- ***Paraphyletic group***: a group of an organism, and some of its descendants.
- ***Polyphyletic group***: a group of organisms that have multiple evolutionary origins, or a group that is not defined by a single common ancestor. This is equivalent to a group of organisms that do not contain their MRCA.

To test your understanding on this topic, let us apply this concept to the organisms on the tree shown in **Figure 13**.

***Q: If we are to classify organisms based on their country of origin (country A, B and C), and given your inferred country of origin of ancestral organisms, draw on the tree to indicate the 3 taxa, and are they monophyletic, paraphyletic, and polyphyletic?***

# Detecting potential conflicting evolutionary signals within the MSA

Owing to the now highly affordable and advanced sequencing technologies, phylogenetic analysis using entire whole genome sequences or gene sequences or SNP sites sampled across the entire genomes is now a common practice, especially in the field of bacteriology. This is actually precisely what we are doing here! It might not look like it as our SNP alignment is very short, but it contains all polymorphic sites from our bacteria's entire genomes already! While this might not be applicable to us, generally, especially in a large-scale ***phylogenomic study*** of bacteria, reconstructing a phylogeny from whole genome sequence data is very challenging, if not problematic. This is in part due to that many bacteria often have so-called **'*mosaic genomes*'**, in which different regions have different past histories.

***Horizontal gene transfer* (HGT)** — the process by which genetic material is transferred between (distinct lineages of) organisms — is very common among bacteria. Within a bacterial species, for example, it has been estimated that, on average, about 15.5% of accessory genes have undergone HGT (Oliveira *et al.* 2017). This phenomenon is in fact a well-known major driving force of the emergence of new virulent and drug-resistant bacterial strains, as well as the ability to evade host immunity (Kado 2009; Deng *et al.* 2019). Indeed, most antibiotic resistance genes are found in HGT hotspots (Oliveira *et al.* 2017), supporting this view.

In the context of molecular phylogenetic reconstruction, it is commonly assumed that all sites within the analysed MSA have the same underlying tree (although different sites may allow to have different evolutionary rates and probabilities of character substitutions). If this assumption is met, then a single tree diagram is sufficient for depicting the histories of all features used in the analysis. However, this assumption is violated more often than not, especially in large-scale phylogenetic analysis of bacteria due to the pervasive HGT outlined above. Thus, it is best to always check for potential presence of mosaic sequences in the dataset analysed; otherwise, naïve application of a standard method to a dataset containing mosaic sequences, which often assumes all molecular sites to share the same underlying phylogeny, could potentially produce a severely biased tree (reviewed in Aiewsakun 2024).

Since the datasets in this practical were derived from bacterial clonal populations, they are unlikely to be mosaic sequences, but we will use the '*pairwise homoplasy index*' test, also known as '*Phi*' test, to check if this is really the case. Phi test is one of the more recent tests for recombination (and homoplasy) detection within an MSA. The test computes the number of recombination events (and homoplastic mutations) required to explain the history of any two nearby sites on average, and sees if the observed value is significantly different from the number expected under the null hypothesis of no recombination (or homoplasy) or not (Bruen, Philippe and Bryant 2006). We will use `SplitsTree` to perform the test on our MSA.

`SplitsTree` is, again, a free program with a graphic user interface for computing unrooted '*phylogenetic'* networks from molecular sequence data (see below). To perform the Phi test on your MSA, open the program. Then click "`File`" > "`Open`". A file navigation window should pop up. Locate your MSA file and then open the file in the program. The program will automatically generate a phylogenetic network for you, but let's ignore it for now.

To perform phi test, click "`Analysis`" > "`Run Phi Test for Recombination`".

***Q: Does your MSA contain an overall significant signal of recombination?***

Now, we will run Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences), a commonly used bioinformatics tool to identify loci containing elevated densities of base substitutions, which are marked as recombinant regions, and to build a phylogeny based on the putative point mutations outside of these regions.

Open a new terminal window and execute the commands below:

```
align="Kpn_ST78.cpe058.strain_ids.fas";
prefix="Kpn_ST78.cpe058.gubbins";
```

```
run_gubbins.py  $align  --prefix  $prefix  --use-time-stamp  --
threads  4  --first-tree-builder  fasttree  --tree-builder  raxml
--outgroup Germany_2019_Kpn_ST78
```

Spend some time familiarising yourself with the options and parameters available on Gubbins:

```
run_gubbins.py -h
```

Next, inspect the output files generated by Gubbins:

```
ls -l Kpn_ST78.cpe058.gubbins*
```

Pay particular attention to the files containing the detected recombination (i.e., those named with *recombination_predictions.*) and check if any recombination has been detected.

We will use a script made available by Gubbins to mask recombinant regions detected by Gubbins from the input alignment:

```
mask_gubbins_aln.py  --aln  Kpn_ST78.cpe058.strain_ids.fas  --
gff  Kpn_ST78.cpe058.gubbins.recombination_predictions.gff  --
out Kpn_ST78.cpe058.rmRCB.fas
```

You can inspect this new alignment with '`seqkit stats`' to confirm the alignment contains the same number of sequences and length of the alignment:

```
seqkit stats Kpn_ST78.cpe058.rmRCB.fas
```

`pairsnp` can also be run to extract the new pairwise SNP distances from this alignment:

```
pairsnp         -c         Kpn_ST78.cpe058.rmRCB.fas         >
Kpn_ST78.cpe058.rmRCB.pairsnp.csv
```

Finally, let's use `IQ-TREE 2` on this alignment with recombination regions masked:

```
align="Kpn_ST78.cpe058.rmRCB.fas";
pattern=`snp-sites -C $align`;
iqtree2 -fconst $pattern -s $align -T 4 --mem 4G --ufboot 1000
-m GTR --prefix Kpn_ST78.cpe058.rmRCB_iqtree -wbtl
```

Note that output files have been named with the suffix 'rmRCB' (removed recombination) not to overwrite the output files generated earlier.

***Q: Compare the phylogenetic trees before and after removing recombination. What differences can be observed in the topology, clustering, and branch length of the trees? Hint: inspect as well the pairwise SNP distances generated by pairsnp.***

# Phylogenetic network reconstruction

One defining feature of a tree graph is that any two nodes in the graph must be connected by just one path, and one path only. This structural constraint implies that, in a rooted phylogeny, a biological entity can have at most just one parent, and this has restricted its ability to depicting only vertical evolutionary relationships. However, HGT are fundamentally non-tree-like, giving rise to organisms that are amalgamations of multiple individuals of distinct evolutionary lineages, and thus a phylogenetic tree cannot handle this.

A '***phylogenetic network***', unlike a phylogenetic tree, allows nodes to have multiple parental nodes. A phylogenetic network can thus be thought of as a generalisation of a phylogenetic tree, accommodating both representation of vertical evolutionary processes and biological-fusion events between organisms of multiple evolutionary pasts. While internal nodes in a network still represent diversification events of the molecular sequences under investigation, they do not simply imply diversification events of the organisms bearing them anymore, but can refer to HGT events as well.

One major class of (implicit) phylogenetic networks is ***split networks***—networks computed based on a collection of weighted splits, i.e., a collection of weighted (bi)partitions of taxa divided into two nonempty sets. In the context of phylogenetic network, an edge, or a band of parallel edges when some of the splits in the collection are '*incompatible*', represents a separation between two taxon clusters, and their lengths are proportional to the split weights, such that the total length of the shortest path between a taxon pair in the graph best approximates their overall empirical dissimilarity. If one were to construct a network from a set of '*compatible*' splits, one would get an unrooted tree diagram, with each edge corresponding exactly to a split.

Open you SNP alignment file "Kpn_ST78.cpe058.strain_ids.snps.fas" in `SplitsTree`, and you should now be seeing a split network computed based on weighted splits derived from the P distance matrix (i.e., the normalised numbers of character-state differences between all taxon pairs) using the *neighbour-net* method (**Figure 15**).

On the "`SplitNetwork`" window, you should see the flow of commands executed by `SplitsTree`: "`Taxa Filter`" > "`P Distance`" > "`Neighbor Net`" > "`Show Splits`". Click on the "`P Distance`" area, and you should see a tiny window pop up on your lower left-hand side. Change from "`P Distance`" to "`GTR distance`", check the "`Use ML_Distance`" box, and then press the "play" button (the encircled triangle).

***Q: How does the network change? Is the results consistent with those obtained from `IQ-TREE 2`?***

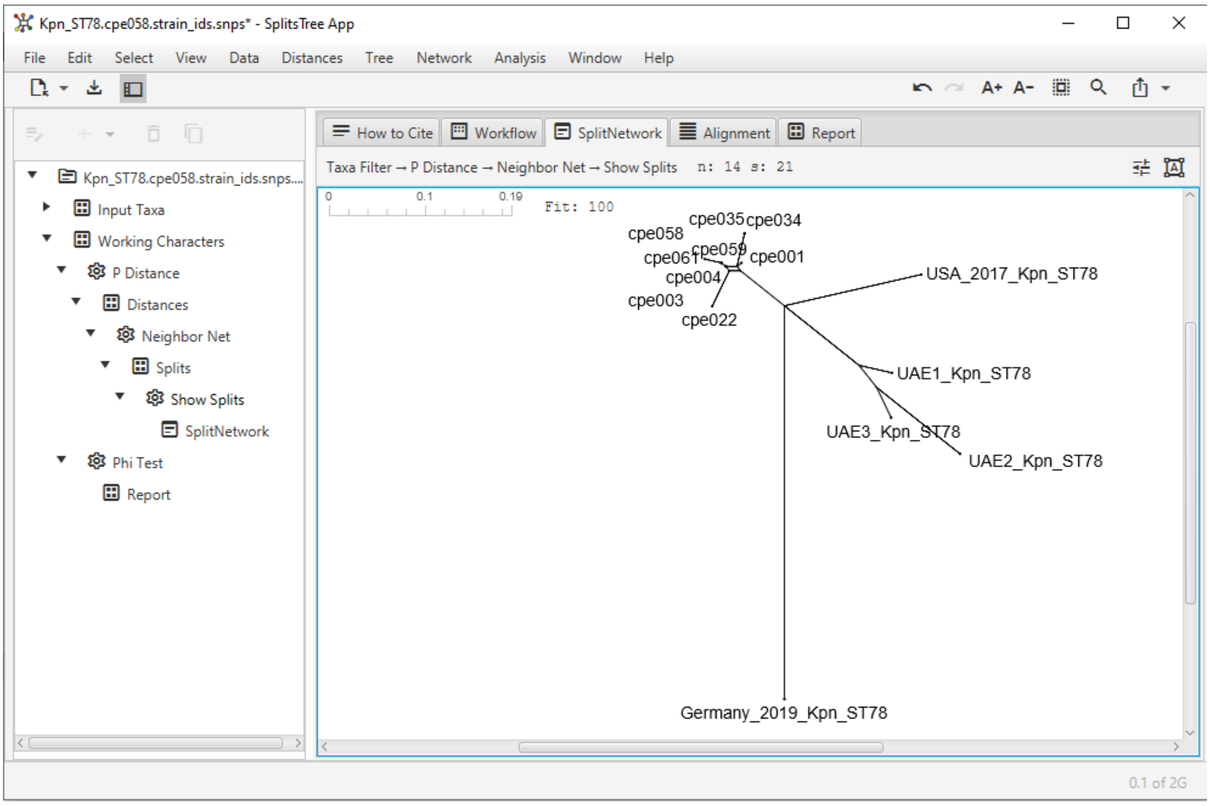


**Figure 15 A split network displayed on `SplitsTree`.**

# Take-home messages

Phylogenetic analysis is now a standard practice for microbiologists. Here, you have learnt how to estimate phylogenetic relationships from molecular data, and how to interpret phylogenies. We have also covered a few methods to detect conflicting evolutionary signals within molecular data, and how to deal with them using phylogenetic network analysis. Additionally, we briefly touched on examining the phylogenetic structure of trait evolution and grouping.

While this might seem like a lot already, this is just the beginning. Applications of phylogenetic analysis extend far beyond what we have covered here. To give you a few examples, phylogenetic analysis can be used to estimate effective population size dynamics, trace the detailed transmission history of pathogens, and establish correlations between genetic factors and phenotypic traits such as drug resistance. We hope that this practical has provided you a somewhat solid foundation and sparked your interest in the broader applications of phylogenetic analysis. 😊

Keep exploring!

# Answers to exercises on interpreting phylogenetic trees

Question 1: ‘Node E’ is the correct answer. ‘Node F’ is an ancestor of sample 8 but not of sample 10. ‘Node D’ is a common ancestor of samples 8 and 10, but it is more ancient than ‘node E’. ‘Sample 7’ is a living specimen and is not an ancestor.

Question 2: Samples 1 to 5 is the correct answer. Remember that in a tree represented as a rectangular layout, the length of horizontal lines (branches) represent genetic distances whereas vertical lines are only used to connect horizontal lines. In the tree above, samples 1 to 5 have the shortest branches connecting them to their common ancestor (node B).

Question 3: Sample 10 is more closely related to sample 8 than to sample 7. The MRCA of samples 10 and 8 is at node E, whereas the MRCA of samples 10 and 7 is at node D, which is deeper (more ancestral) in the tree.

Question 4: Sample 7 is equally related to sample 8 and sample 10. The MRCA of samples 7 and 8 is at node D, as is the MRCA of sample 7 and 10. All descendants of node E are equally related to sample 7.

Question 5: The common ancestor of samples 6 to 10 (node D) most likely circulated in country B first and later on transmitted to country C. Country B is the most likely origin of the common ancestor represented by ‘node D’ because its direct descendants (‘node C’ and ‘node E’) both contain samples collected on this country. Later on, one clone transmitted from country B to C before diversifying (represented by ‘node F’) to then give rise to sample 8 and 9.

Question 6: The common ancestor of samples 1 to 10 (node A) could have circulated either in countries A or B. Information on the countries of origin of samples that descended from more ancestral nodes to ‘Node A’ is needed to draw a more definitive conclusion.

# Bibliography and references

Aiewsakun P. Microbial evolutionary reconstruction in the presence of mosaic sequences. *Phylogenomics*. 1st ed. Elsevier, 2024, 177–217.

Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 2006;**172**:2665.

Deng Y, Xu H, Su Y *et al*. Horizontal gene transfer contributes to virulence and antibiotic resistance of Vibrio harveyi 345 based on complete genome sequence analysis. *BMC Genomics* 2019;**20**:761.

Felsenstein J. Bayesian inference of phylogenies. *Inferring Phylogenies*. Sunderland: Sinauer Associates, Inc., 2004, 288–304.

Kado CI. Horizontal gene transfer: sustaining pathogenicity and optimizing host–pathogen interactions. *Molecular Plant Pathology* 2009;**10**:143–50.

Oliveira PH, Touchon M, Cury J *et al*. The chromosomal organization of horizontal gene transfer in bacteria. *Nature Communications* 2017;**8**:841.

Parts of this handout were developed and modified from handouts used in the workshop “The ripple effect of COVID-19 pandmic or respiratory virus research and healthcare: genetic analysis of influenza virus and SARS-CoV-2: SARS-CoV-2 phylogenetic analysis”, originally developed by Pakorn Aiewsakun.