Session 8: Phylogenetics

What are phylogenetic trees?

A phylogeny, also known as phylogenetic tree, depicts estimated evolutionary relationships between taxa - these can be species, strains or even genes.

In the context of infectious diseases epidemiology, phylogenetic trees are commonly used to define evolutionary relationships between strains of the same bacterial species. This is possible because bacteria reproduce clonally.

During clonal reproduction, bacterial progenitor cells replicate their DNA at high fidelity. Despite this, random errors in DNA replication may still occur, resulting in a clonal progeny that will inherit these genetic replication 'errors' (i.e. mutations) in their DNA and may not be strictly identical to their progenitor cells. Bacterial strains that have recently originated from the same progenitor cell are thus expected to share identical genomes, or have diverged at most by only a few genetic differences (mutations). The number and pattern of shared mutations between bacterial strains can be used to reconstruct their genealogical and evolutionary relationships.

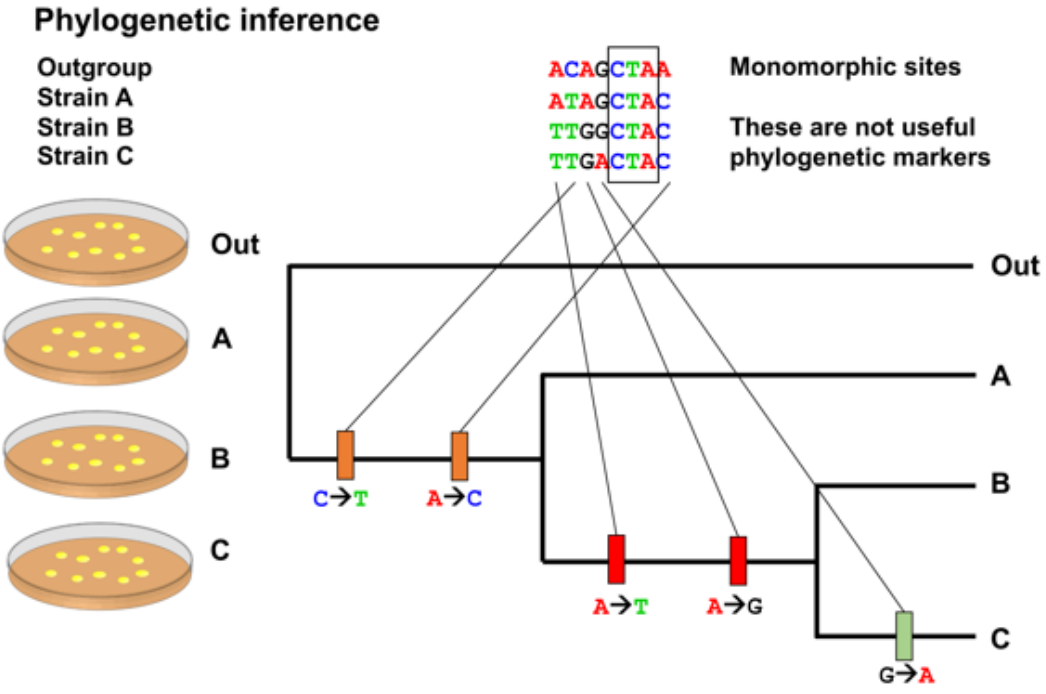
On a phylogenetic tree, isolated bacterial strains are depicted on the tips (or leaves) of the tree (i.e. taxa), whereas the internal nodes of the tree denote their hypothetical ancestors. Nodes and taxa are connected by branches, the length of which represent genetic distances between connected groups. Groups of bacterial strains (taxa) that share the same common ancestor form a monophyletic group (also known as clade). A group of strains that descends from a common ancestor, but does not include all descendants, is called paraphyletic.

How are phylogenetic trees reconstructed?

Today almost all phylogenetic trees are inferred from molecular sequence data, most often from DNA sequences. This is because DNA is an inherited material; it can easily, reliably and inexpensively be extracted and sequenced; and DNA sequences are highly specific to bacterial species and strains.

The application of whole-genome sequencing now makes it possible to 'read' the DNA sequence of the whole bacterial chromosome, which provides the ultimate level of resolution possible to discriminate between closely related strains. The figure below illustrates the common workflow to generate multiple DNA sequence alignments from a collection of bacterial strains. Generally, bacterial DNA is extracted from a single colony picked from culture plates (referred as to 'isolate'), followed by library preparation and whole-genome sequencing using rapid benchtop sequencers. Raw sequence data generated by sequencers are processed using bioinformatic and genomic pipelines, which generally involve mapping (aligning) the generated short reads to a reference genome to reconstruct the isolate's DNA sequence along the whole bacterial chromosome. Mapping the short reads of multiple sequenced isolates to the same reference genome allows the creation of multiple sequence alignments. Multiple sequence alignments are the first and critical point from which phylogenetic trees can be re-constructed.

Polymorphic sites (that is, nucleotide positions that are variable across multiple strains) in multiple alignments are used to infer evolutionary relationships, whereas monomorphic sites (nucleotide positions with the same DNA base) are generally ignored. The figure below shows an example of a simple multiple alignment 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 monophyletic groups (clades) in a hierarchical manner (see example below) with the goal of constructing the most plausible genealogical relationships between strains and clades.



How are phylogenetic trees interpreted?

The preferred interpretation of a phylogenetic tree is as a depiction of lines of descent. That is, trees communicate the evolutionary relationships among strains and clades. Under this interpretation, internal nodes on a tree are taken to correspond to bacterial strains that existed in the past (ancestral) but could not be sampled.

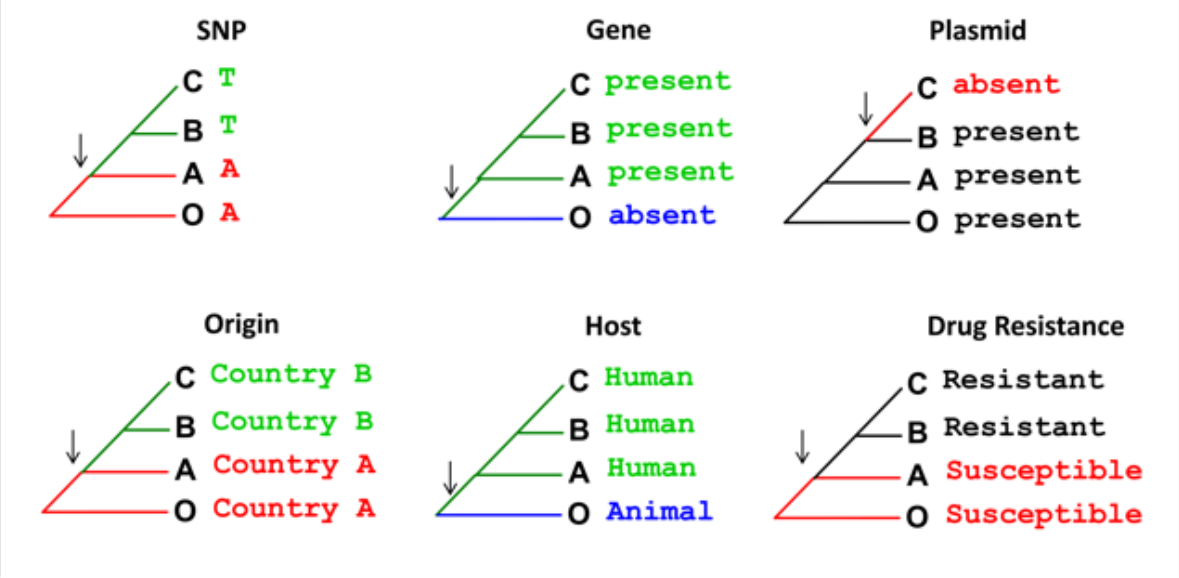
Phylogenies are commonly mis-interpreted when read along the tips. Instead, the correct way to read a tree is as a set of hierarchically nested groups (clades).

In the tree above, strain C is more closely related to strain B than it is to strain A. This is inferred by tracing the ancestor of strains (depicted as internal nodes) using the branch structure (i.e. topology) of the tree. Relatedness should be understood in terms of common ancestry— the more recently strains share a common ancestor, the more closely related they are.

How are phylogenetic trees used in infectious diseases epidemiology?

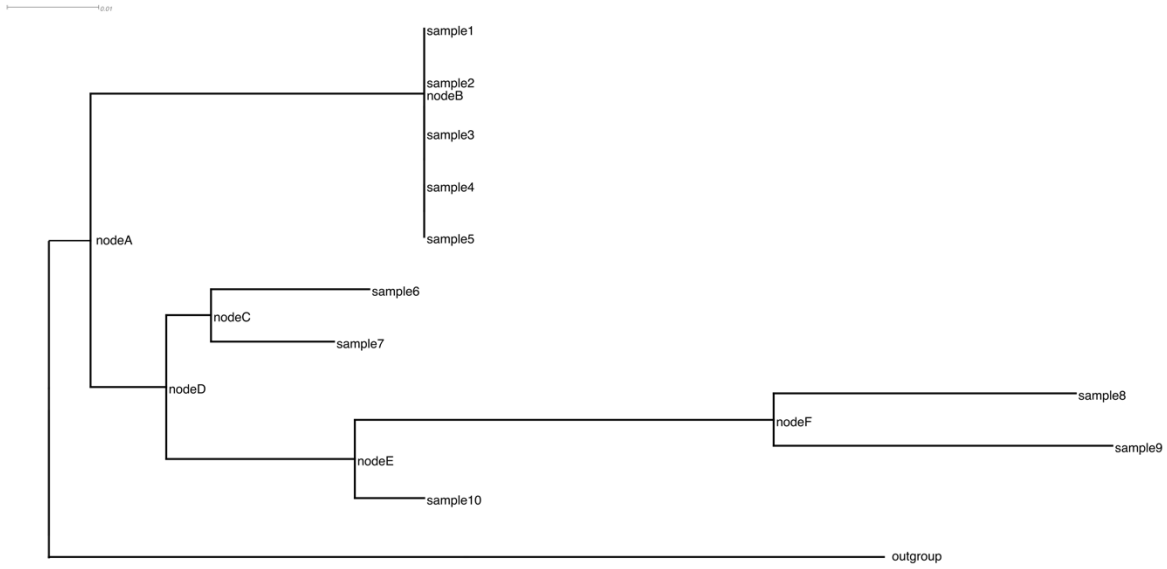
Phylogenetic trees are commonly used to identify where person-to-person transmission occurs; to identify the sources and study the transmission routes of outbreak and epidemic clones; and to determine whether bacterial clones are restricted to specific hosts and settings or, on the contrary, able to circulate among multiple ones.

A common phylogenetic method used to study how bacterial characteristics (traits) evolved is ancestral state reconstruction. In the example shown below, 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 AMR. The emergence and spread of individual mutations, genes and mobile genetic elements can also be reconstructed in a bacterial phylogeny using this method (see figure below).



Exercises on interpreting phylogenetic trees

This section includes questions on inferring genetic relatedness using phylogenetic trees.



Question 1. based on the tree above, what internal node corresponds to the most recent common ancestor 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

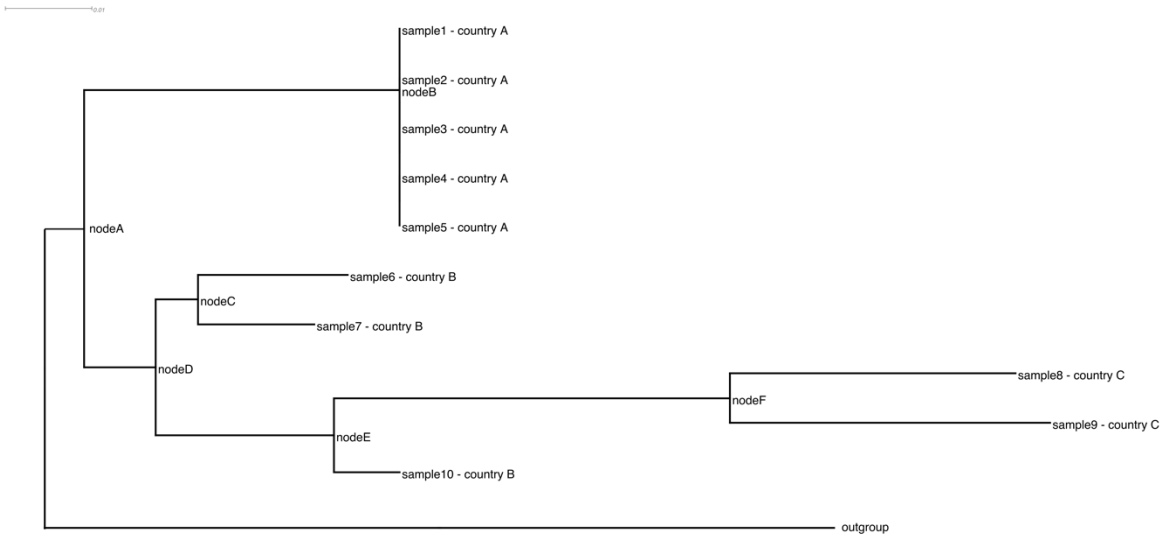
Question 3. Based on the tree above, which of the following statements referring to sample 10 is more accurate:

- 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 more 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



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

Exercises on constructing a phylogeny from gene sequences and whole genome SNPs

Having learned how to interpret the tree, we will next construct phylogenetic trees using the sequence from gene *mecX*, an imaginary hypothetical gene from *Staphylococcus aureus*. *MecX* has a high sequence similarity to *mecA*, a gene identified as responsible for methicillin resistance in *Staphylococcaceae* family. In this practical, you will compare and contrast the phylogenetic relationship of the gene *mecX* and the whole genome of *Staphylococcus aureus*. Both the *mecX* sequences, and the whole genome SNP alignment have been provided for you in the session folder. The whole genome SNP alignment was subsampled from the data you generated in the previous session. To ensure that the phylogenetic tree can be constructed in a timely manner, we will be working with 13 *S. aureus* genomes and their *mecX* genes.

The *mecX* is carried by the staphylococcal chromosome cassette *mec* (SCC*mec*), a mobile genetic element highlighted as recombination hotspot. The origins of *mecX* remained unclear with two competing hypotheses: (i) *mecX* was passed on vertically from generation to generation; and (ii) *mecX* was repeatedly horizontally acquired by independent recombination events. To address the hypothesis, you will construct phylogeny from *mecX* sequences and whole genome SNP of *S. aureus*. The topology of the gene tree versus the whole genome SNPs tree will give you ideas of how gene X dissemination in the population.

Viewing the alignment in SeaView

You will use SeaView to view and edit alignments, as well as to make a phylogeny. The program has a graphical user interface (GUI) that is easy to operate.

First you can navigate to Practical 8 and 9 Phylogeny

#change the working directory

cd Session_8_Phylogenetics

#start Seaview

seaview

#load the alignment of *mecX* sequences by selecting “Open” from the “File” menu.

Please spend some time to look at the diversity of *mecX* sequences in your samples. *MecX* is diverse with high SNP density and several indels, which makes mapping challenging and prone to errors. Therefore, we will use the assembly version of the gene in this part of the practical.

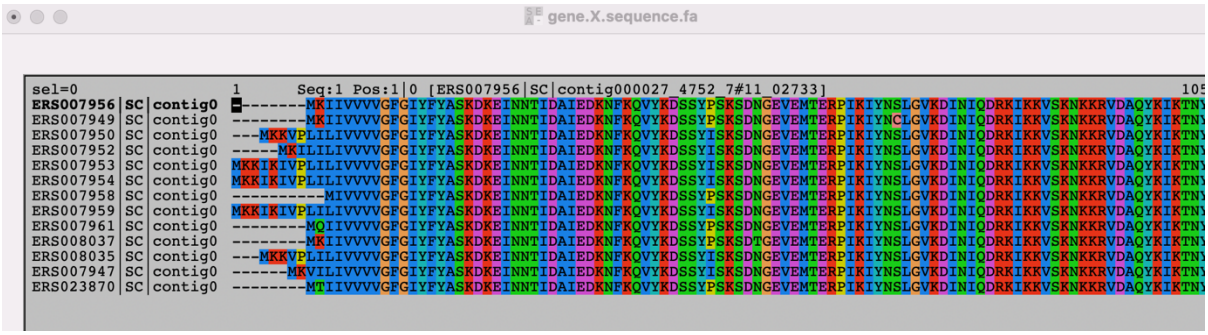
Multiple sequence alignment

It is important to make sure that the column in our data represent homologous bases by aligning the nucleotide or amino acid using a multiple alignment programme. For *mecX* sequences, length differences might complicate multiple sequence alignment as these require insertion or gaps into an alignment to ensure that homologous sites remain aligned. When possible, please check an alignment by eyes.

Seaview allows alignment using two programmes, clustal and muscle. Generally muscle is faster, and the protein alignments are of similar quality to clustal. In both cases, sequences are aligned by assigning costs to particular base changes and gaps insertions, with the optimal alignment having the lowest cost.

To start the alignment, select "**Align**" then "**Align all**"

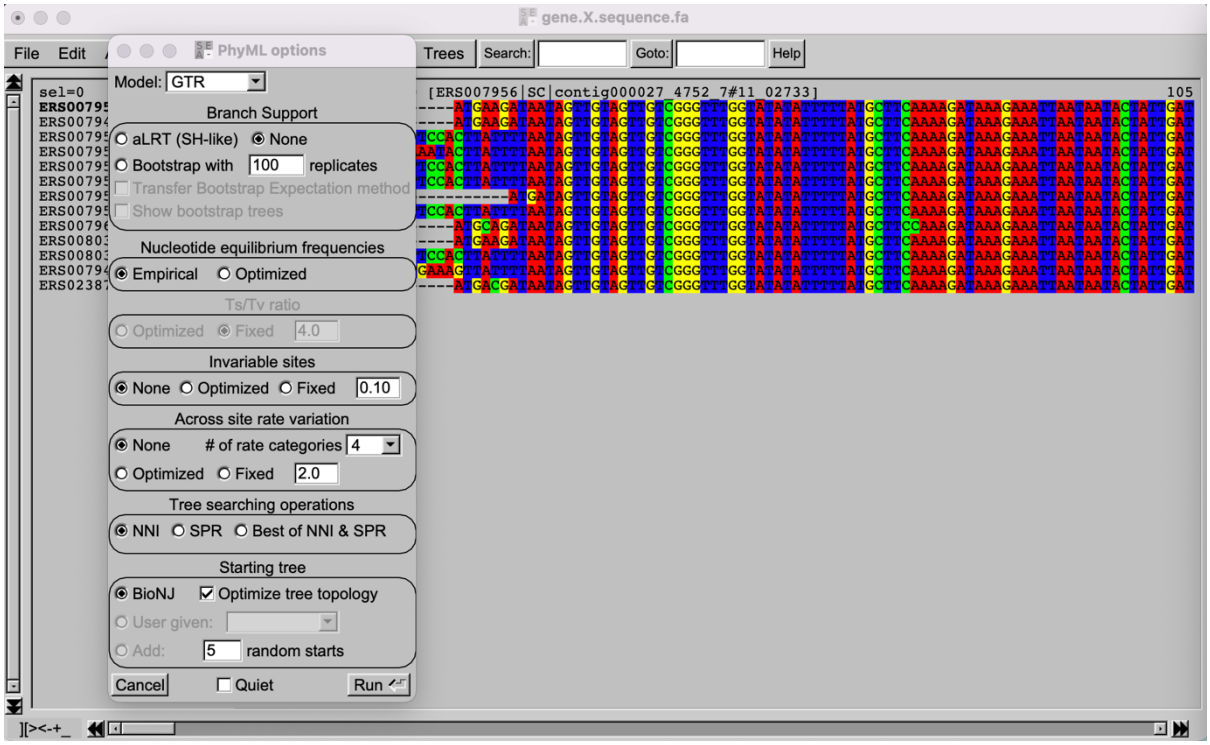
When the alignment process is complete, SeaView will have inserted gaps into the sequences so that homologous sites are lined up in colunms. Please inspect the alignment. You should be able to see which sequences are most closely related. If an alignment is still misaligned when observed by eye, you can edit the alignment as necessary.



Constructing a phylogeny from gene sequences using PhyML

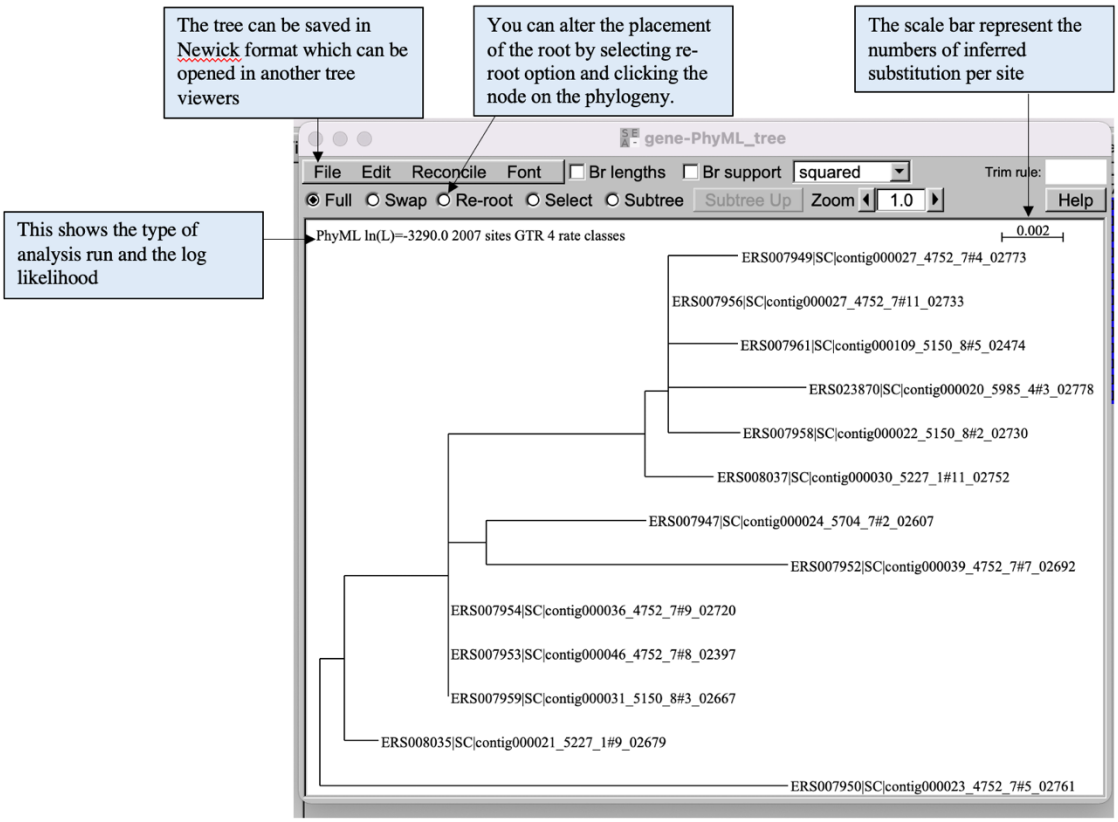
To estimate the phylogeny, we will use a programme called PhyML, which is already included in SeaView. PhyML includes a number of nucleotide substitution models ranging from the very simple (and could be unrealistic) to a complex one. PhyML uses maximum likelihood (ML) to estimate the tree. ML is more accurate than simpler approaches as it specifies an explicit evolutionary model to account for sources of homoplasy, while does not take too long to run. Moreover, it uses an optimality criterion (likelihood), which enable different trees to be compared. You can read more about the tree model from https://www.nature.com/articles/nrg3186

Click "**Trees"** option and selection "**PhyML**"



We will start by running a tree with choosing **GTR** model for substitution rates. In the **branch support**, select "**None**" as we will look at this parameter later. If you choose a model to include different transition and transversion substitution rates, you can set the ratio at Ts/Tv ratio. In the "**Across site rate variation**" box, select "**None**", we will look at the rate ratio later. The last two boxes help specify how the programme will choose the tree to start from (useful for large scale data). We will leave them as defaults.

Once the run has finished, click "OK". You will have a phylogenetic as in the figure.



The type of analysis run and the **loglikelihood** are shown above the tree.

You can alter the placement of the root by selecting **re-root** option and clicking the node on the phylogeny. You can tick **Br lengths** to add branch length values to the tree.

Please record likelihood score, that is ln(L) of your first tree.

Creating gene trees with different models

Not all sites within the sequence evolve at the same rate. Some parts of a gene evolve faster than others. For example, an active site of an enzyme which is functionally important may be more conserved than others. When comparing several sequences to estimate a phylogeny, we should account for rate heterogeneity to avoid errors.

Select “Optimised” from the “Invariable sites” box to allow the proportion of invariant sites to change but keep other parameters the same as the previous analysis. Press “Run” to generate a phylogeny as before. Please record the likelihood score of your second tree.

Comparing models with the likelihood ratio test (LRT)

ln(L) model 1 =

ln(L) model 2 =

You can compare the likelihood scores of the first (simpler model) and the second tree (more complex model) to work out which model fits better.

Alternatively, you can statistically test which model is better using the following formula:

LR = 2 x (neg ln(L1) – neg ln(L2))

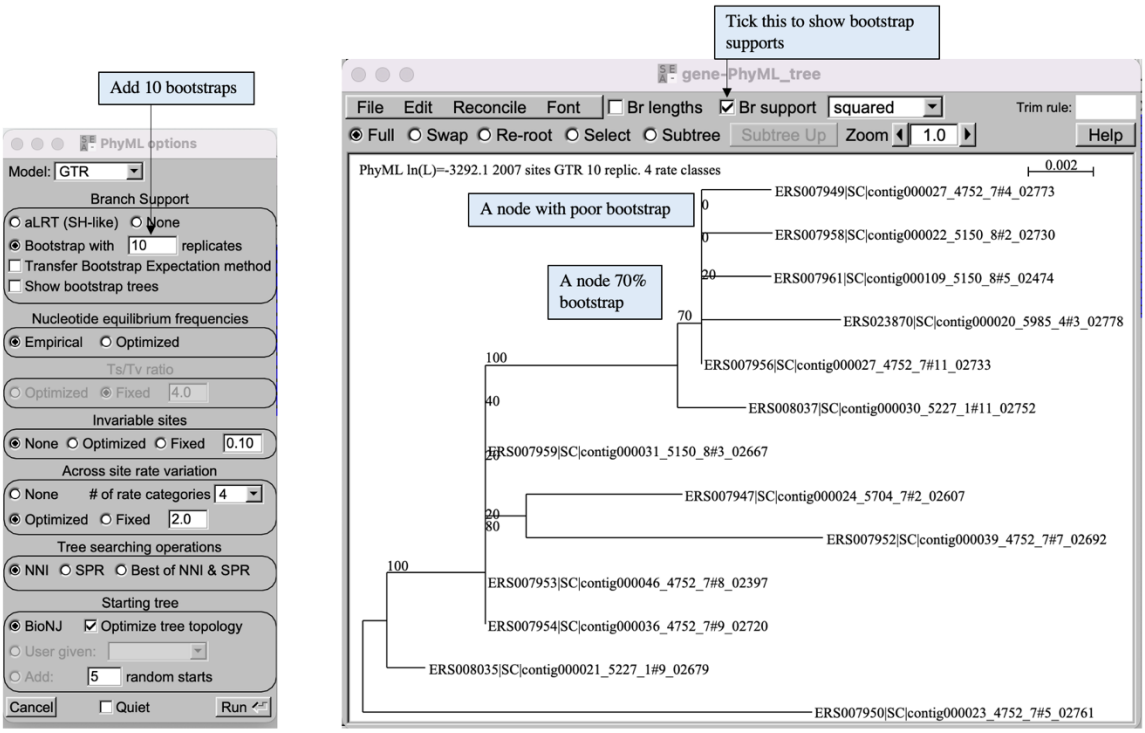
Where neg ln(L1) and neg ln(L2) present the negative log likelihood of the simpler and more complex model, respectively.

The significant of LR can be estimated using chi-square significant tables.

Phylogeny estimation with bootstrapping

Bootstrapping is a statistical technique to assess the confidence level around each phylogenetic node. The bootstrapping values indicates how many times out of 100, the same branch was observed when repeating the phylogenetic reconstruction on the re-sampled set of your data. Robust relationship should be repeatable, and subsequently observed in a large proportion of randomised data. Therefore, if you get 100 out of 100 times for a particular node, you can be certain that the observed branch is not due to chance, but likely to be real.

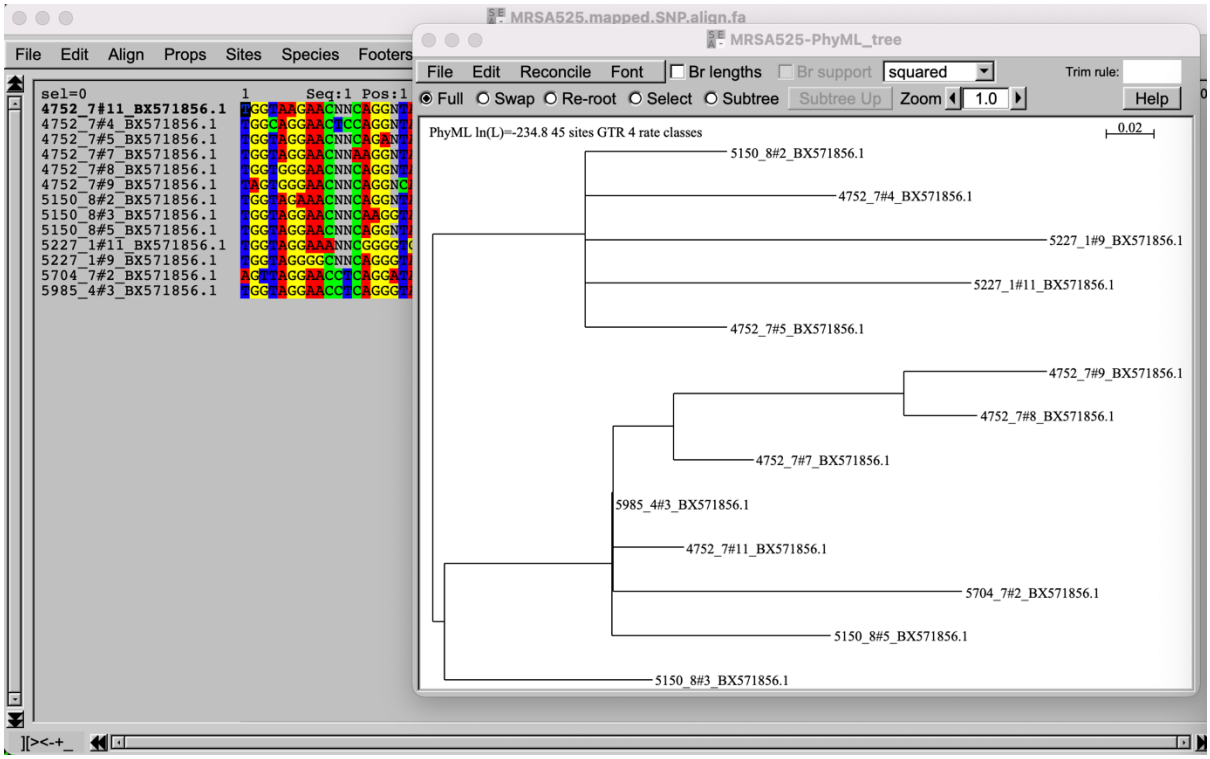
You can create a bootstrapped phylogeny for *mecX* dataset by creating a new phylogeny as before. This time click on "Bootstrap" in the "Branch support" box, and enter 10 in the replicates box. Ideally, you would need to run at least 100 or 1000 replicates. But due to speed and time limit, we will only work on 10 replicates in this exercise.



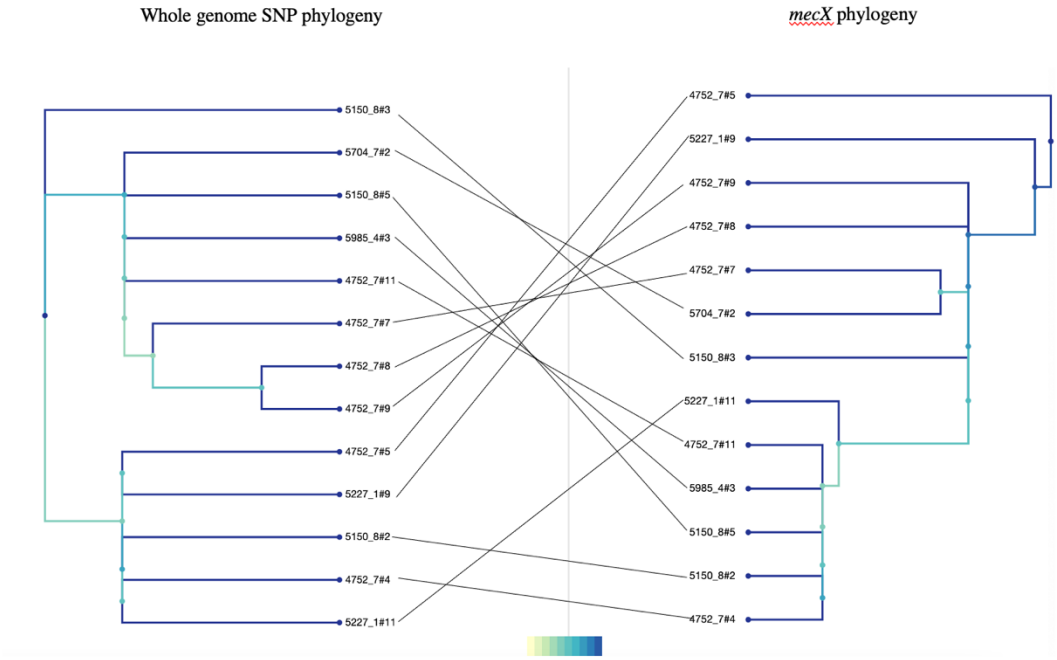
Constructing phylogeny from whole genome SNPs

You can either construct a phylogeny using a whole genome alignment produced from the previous practical, or using only sites that contain variations. The former will be slow to run. We therefore will only extract the sites which only contain SNPs using a C script called "snp_sites" (https://github.com/sanger-pathogens/snp-sites).

The SNP alignment of *S. aureus* has been prepared for you. Open the SNP alignment in Seaview and make a tree as before. Do not include the invariant sites parameter, as we just removed all invariant sites from the dataset.



Finally, you can compare the phylogenetic trees of *mecX* and *S. aureus* whole genome SNPs which have been produced for you to answer the following questions.



Questions:

1.How does the tree of whole genome SNPs compare to the *mecX* tree you have made?

2. Why the analyses of different part of the genome (whole genome vs *mecX*) lead to different phylogenies?

3. What does this tell you about the acquisition and dissemination of *mecX*?

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 common ancestor 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 most recent common ancestor of samples 10 and 8 is at node E, whereas the most common ancestor of samples 10 and 7 is at node D, which is a deeper (more ancestral) internal node in the tree.

Question 4: Sample 7 is equally related to sample 8 and sample 10. The most recent common ancestor of samples 7 and 8 is at node D, as is the most recent common ancestor 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, represented by ‘node F’, transmitted from country B to C.

Question 6: The common ancestor of samples 1 to 10 (node A) could have circulated in countries A or B. Unfortunately, information on the country of origin of contextual samples descendants from more ancestral nodes to ‘Node A’ are needed to draw more accurate conclusions.