Computational practical 9

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.

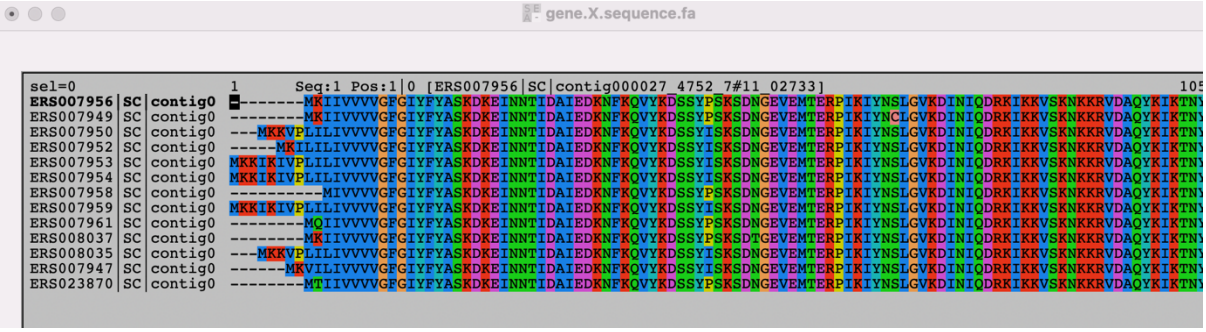
Multiple sequence alignment

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**"

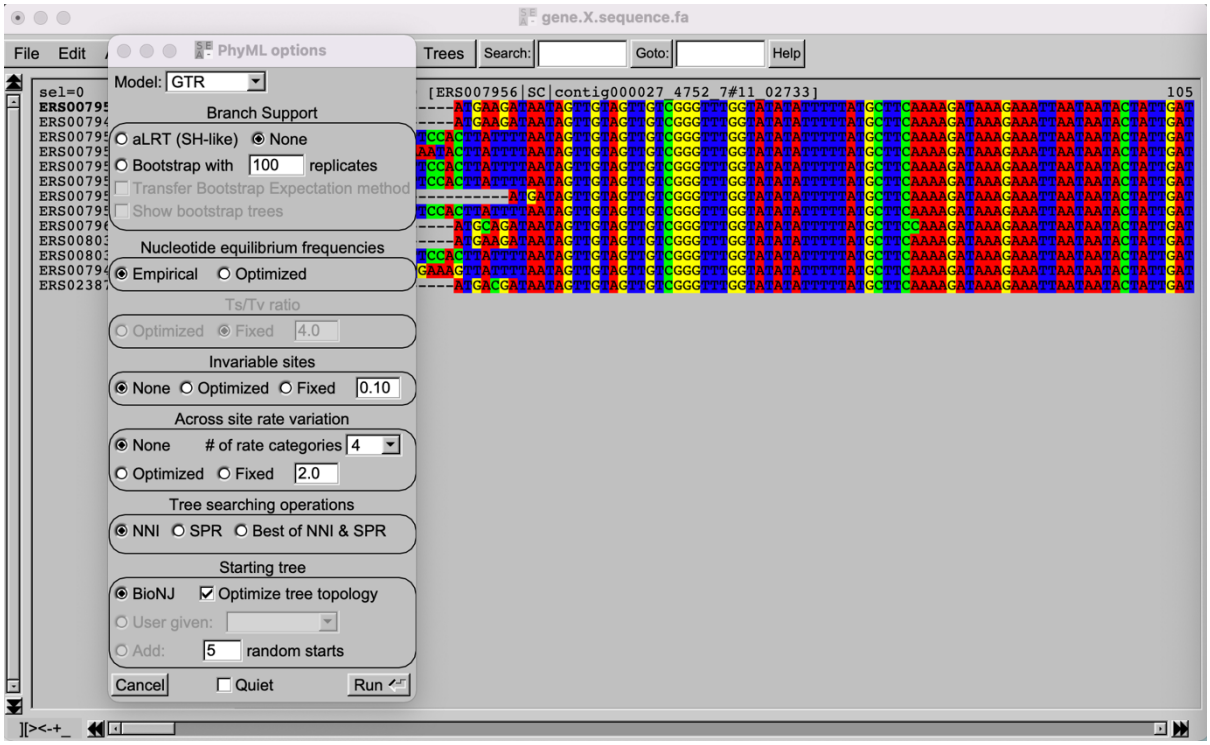
When the alignment process is complete, SeaView will have inserted gaps into the sequences so that homologous sites are lined up in colunms. Please inspect the alignment. You should be able to see which sequences are most closely related. If an alignment is still misaligned when observed by eye, you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



Constructing a phylogeny from gene sequences using PhyML

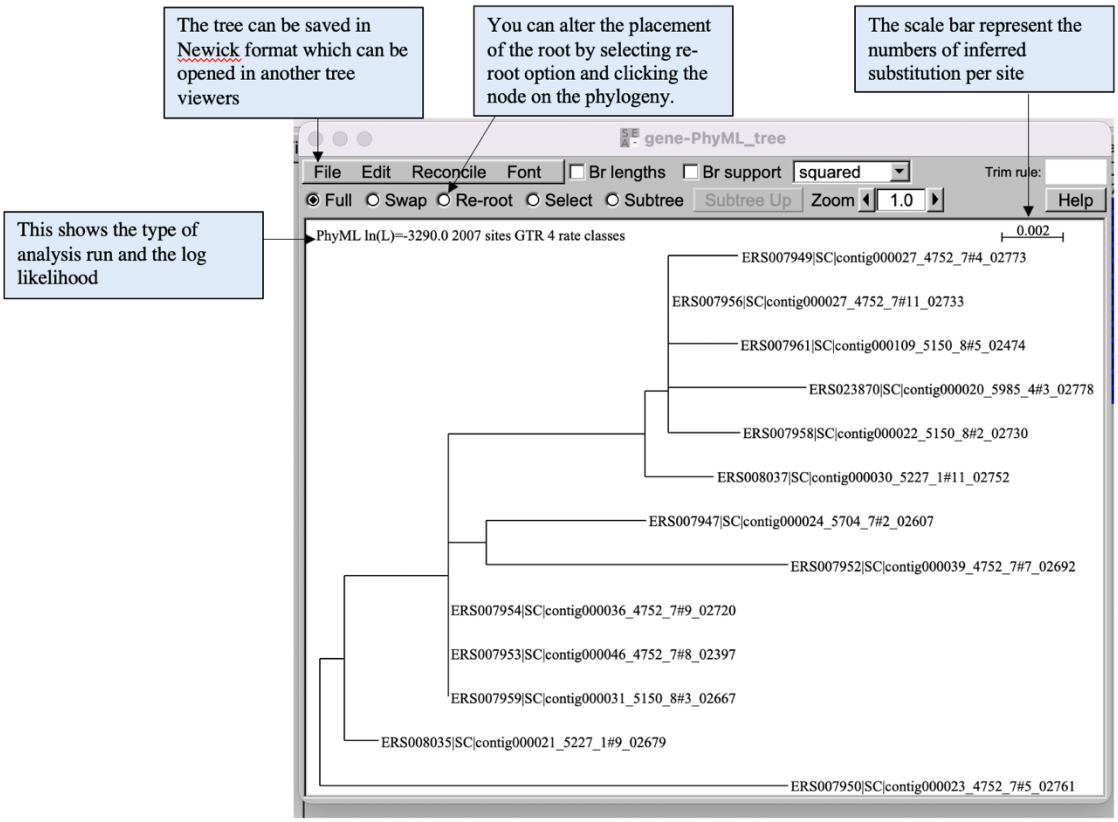
To estimate the phylogeny, we will use a programme called PhyML, which is already included in SeaView. PhyML includes a number of nucleotide substitution models ranging from the very simple (and could be unrealistic) to a complex one. PhyML uses maximum likelihood (ML) to estimate the tree. ML is more 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))

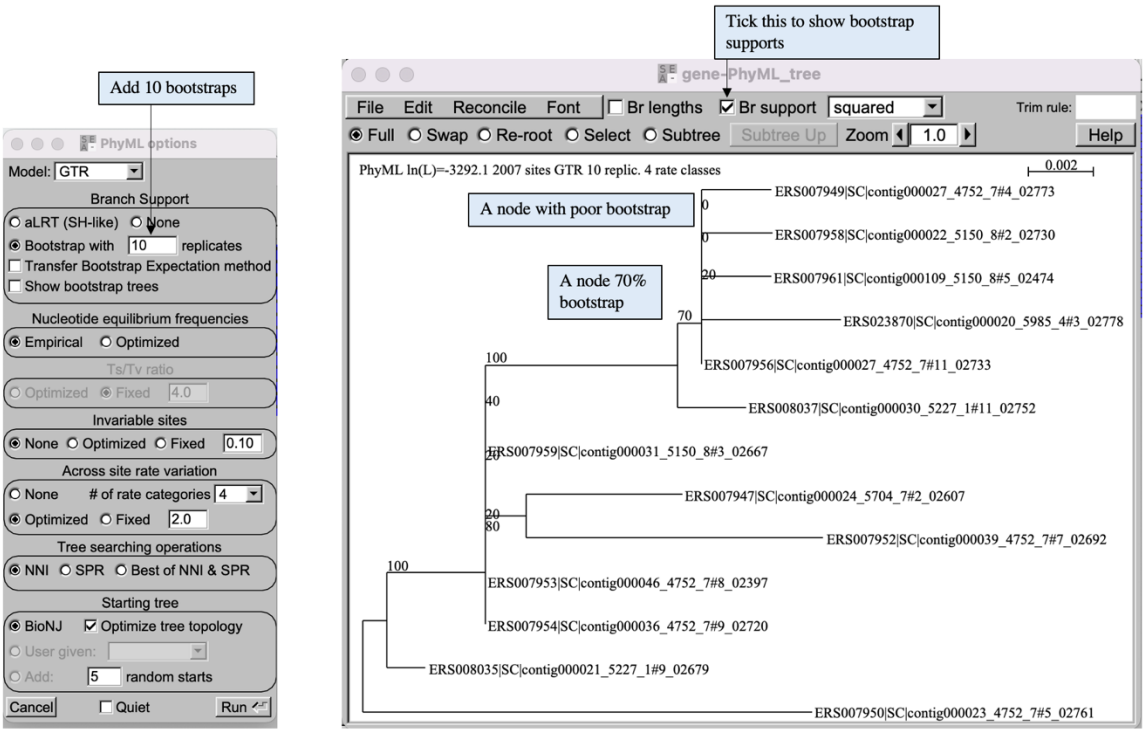
Where neg ln(L1) and neg ln(L2) present the negative log likelihood of the simpler and more complex model, respectively.

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.

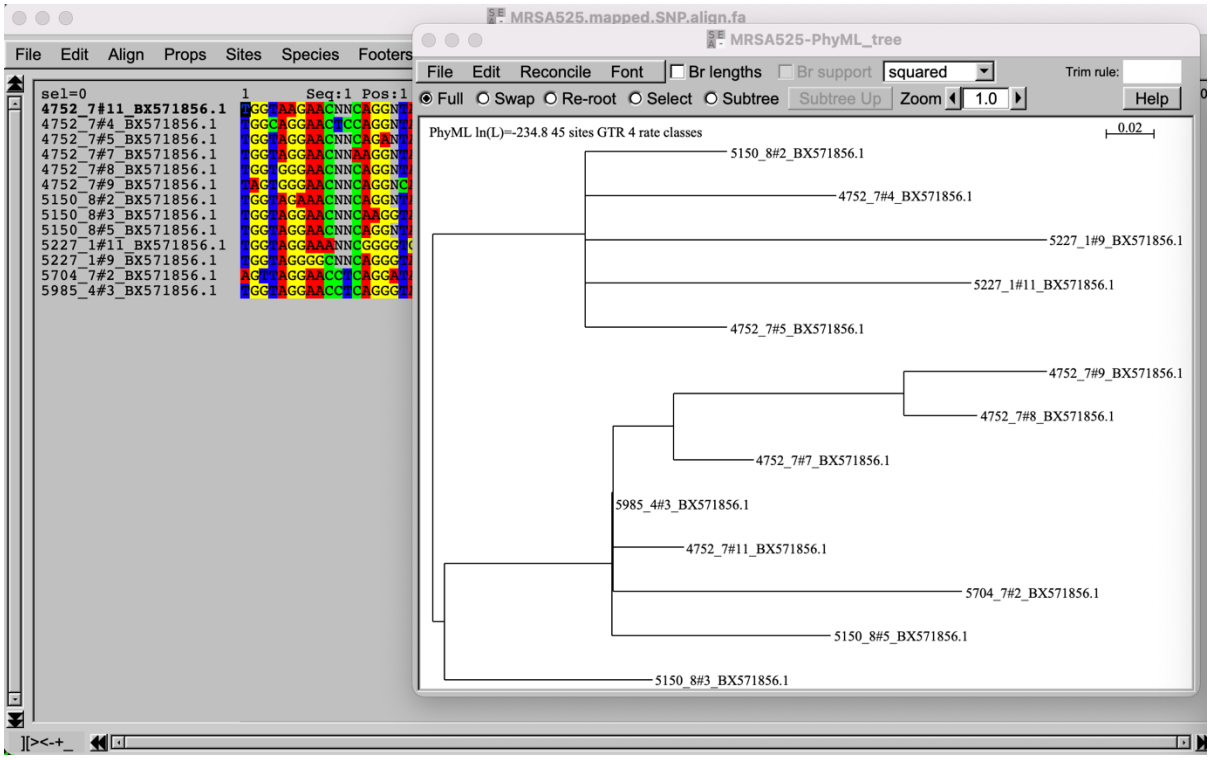
You can create a bootstrapped phylogeny for *mecX* dataset by creating a new phylogeny as before. This time click on "Bootstrap" in the "Branch support" box, and enter 10 in the replicates box. Ideally, you would need to run at least 100 or 1000 replicates. But due to speed and time limit, we will only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



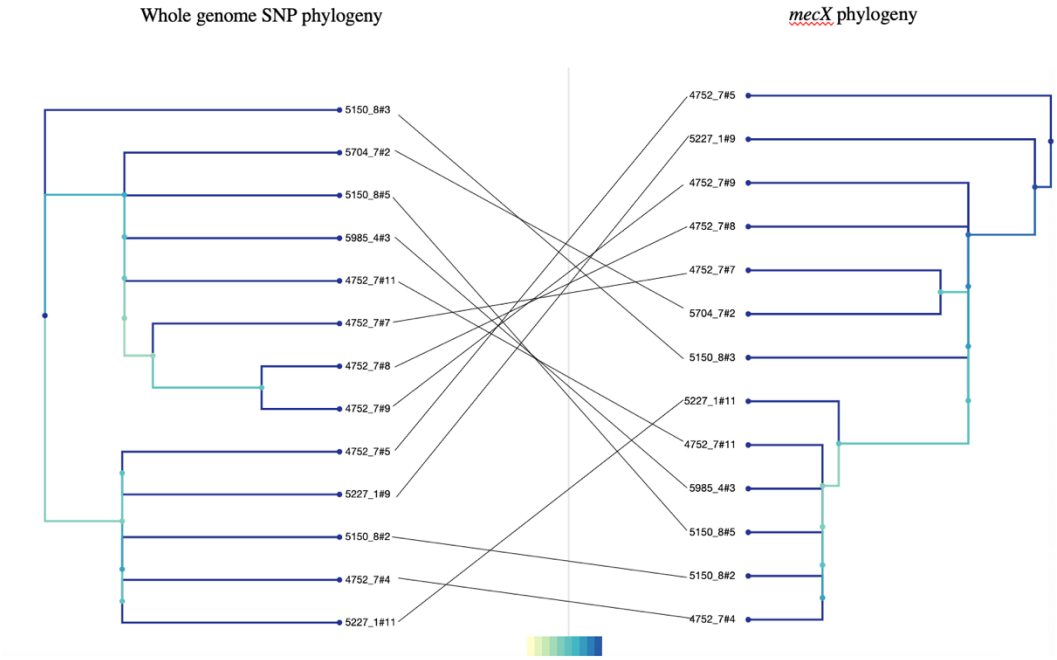
Constructing phylogeny from whole genome SNPs

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*?