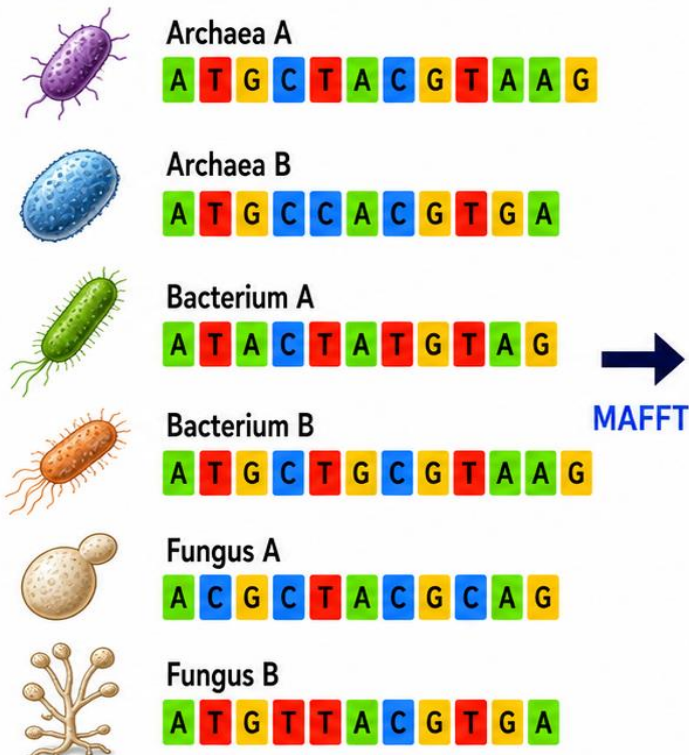


Lesson 3: Unknown Microbes and the Tree of Life: Revealing Genetic Relationships with MAFFT

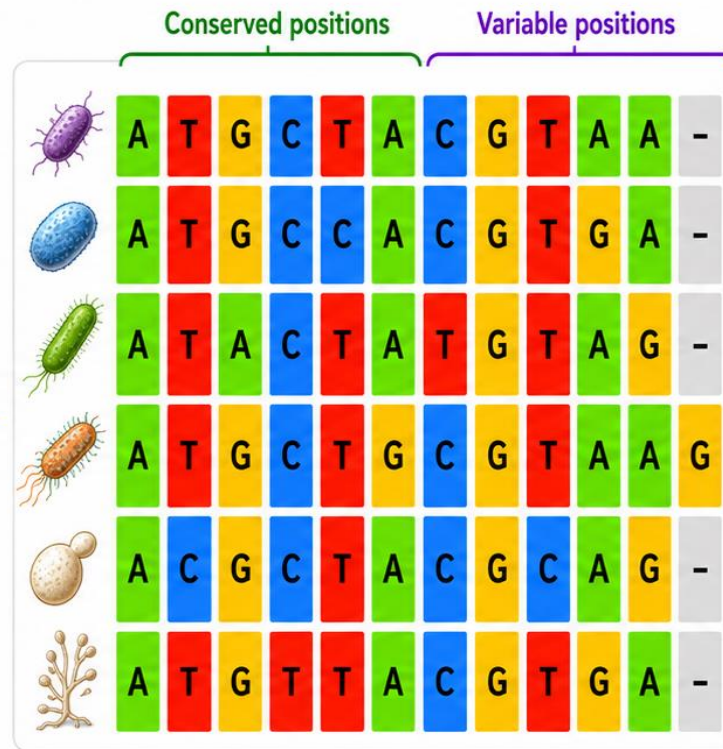
Lesson 3: Bioinformatics Activity 1

Revealing Genetic Relationships with MAFFT: Aligning small subunit RNA (ssRNA) Barcode Sequences (Bacteria, Archaea and Fungi).

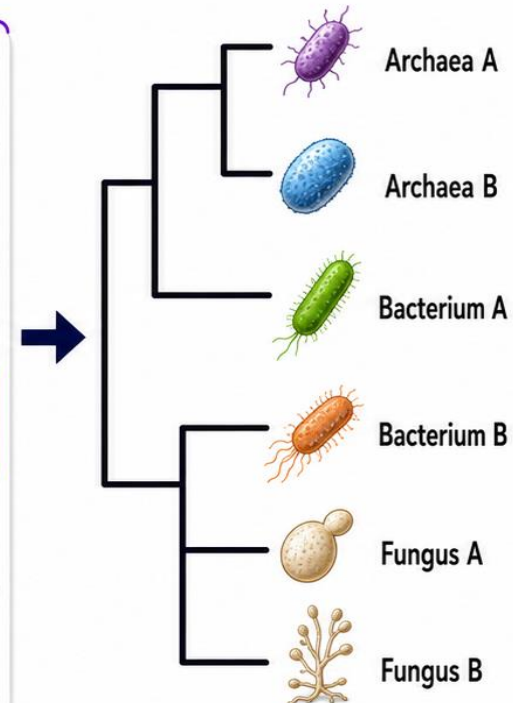
1 DNA barcode sequences



2 Multiple sequence alignment



3 Phylogenetic tree



Bioinformatics Activity 1

- **Use the provided ssRNA sequence* file** containing a curated dataset of bacteria, archaea, and fungi.
- **Upload the sequence file into MAFFT** to generate a multiple sequence alignment.
- **Compare conserved and variable regions** across the aligned sequences.
- **Use the alignment to construct a phylogenetic tree** representing the three domains of life.
- **Identify major clusters** corresponding to Bacteria, Archaea, and Eukarya.
- **Use the tree to discuss genetic relationships** among organisms based on barcode sequence similarity and the identification of an unknown bacterial isolate.

*Small subunit rRNA helps build ribosomes, so its genes are found across all life. They are useful barcodes because conserved regions allow broad comparison, while variable regions help distinguish organisms.

Note: The sequence file is already in FASTA format, a standard format where each sequence begins with an identifier line starting with ">", followed by the DNA sequence using the letters A, T, C, and G. Simply upload the file to MAFFT, no editing is required.



ssRNA Seq Euk Archaea Bact

File Edit View

```
>Escherichia coli
AGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGTAACAGGAAG
CAGCTTGCTGCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGGGAACTGCCTGATGGAGGGGGA
TAATACTGGAACGGTAGCTAATACCGCATAACGTCGCAAGCACAAGAGGGGGACCTTAGGGCCTCTT
GCCATCGGATGTGCCAGATGGGATTAGCTAGTAGGTGGGGTAACGGCTCACC TAGGCGACGATCCCTAG
CTGGTCTGAGAGGATGACCAGCAACACTGGAAC TGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTG
GGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCNGCGTGTATGAAGAAGGCCCTTCGGGTGT
AAAGTACTTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTACCCGCAAGAAG
GCACCGGCTAACTCCGTGCCAGCAGCCGCGTAATACGAGGGGTGCAAGCGTTAATCGGAATTACTGGGC
GTAAAGCGCACGCGAGGCGGTTTGTAAAGTCAGATGTGAAATCCCGGGCTCAACCTGGGAAGTGCATCTG
ATACTGGCAAGCTTGAGTCTCGTAGAGGGGGTAGAATCCAGGTGTAGCGGTGAAATGCGTAGAGATCT
GGAGGAATACCGGTGGCGAAGCGCGCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGC
AAACAGGATTAGATACCTCGTAGTCCACGCCGTAACGATGTGCACTTGGAGGTTGTGCCCTTGAGGCG
TGGCTTCGGANNTAACGCGTTAAGTCGACCGCTGGGGAGTACGGCCGCAAGGTTAACTCAAAATGAA
TTGACGGGGGCGCACAAGCGGTGGAGCATGTGGTTAATTGATGCAACGCGAAGAACCTTACCTGGTC
TTGACATCCACGGAAGTTTTCAGAGATGAGAATGTGCCCTCGGGAACCGTGAGACAGGTGCTGCATGGCT
GTCGTGAGTCTGTTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCTTATCCTTTGTTGCCA
CGGTTCGGCCGGGAAC TCAAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGACGTCAAGTC
ATCATGGCCCTTACGACCGAGGCTACACACGTGCTACAATGGCGCATACAAGAGAAGCGACCTCGCGAG
AGCAAGCGGACCTCATAAAGTGCCTGCTAGTCCGATTGGAGTCTGCAACTCGACTCCATGAAGTCGGAA
TCGCTAGTAATCGTGGATCAGAAATGCCACGGTGAATACGTTCCCGGGCTTGTACACACCGCCCGTCACA
CCATGGGAGTGGGTTGCAAAAGAAGTAGGTAGCTTAACCTTCGGGAGGGCG
```

```
>Staphylococcus aureus MULTIREsISTANT
ATCCTGGCTCAGGATGAACGCTGGCGCGTGCCTAATACATGCAAGTCGAGCGAACGGACGAGAAGCTTG
CTTCTCTGATGTTAGCGCGGACGGGTGAGTAACACGTGGATAACCTACCTATAAGACTGGGATAACCTC
GGAAACCGGAGCTAATACCGGATAATATTTTGAACGCATGGTTCAAAGTGAAAGACGGCTTGTGCTGT
CACTTATAGATGGATCCGCGCTGCATTAGCTAGTTGGTAAGGTAACGGCTTACCAAGGCAACGATGCATA
GCCGACCTGAGAGGGTGATCGGCCACACTGGAAC TGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGT
AGGGAATCTTCGCAATGGCGAAGGCTGACGGAGCAACGCCGCTGAGTGATGAAGGTTCTCGGATCG
TAAACTCTGTTATTAGGGAAGAATATGTGTAAGTAAC TGTCACATCTTGACGGTACCTAATCAGAA
AGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTATCCGGAATTATTGGG
CGTAAAGCGCGCTAGGCGGTTTTTAAAGTCTGATGTGAAAGCCACGGCTCAACCGTGGAGGGTCATTG
GAAACTGGAAAACCTTGATGCAGAAGAGGAAGTGGAATTCATGTGTAGCGGTGAAATGCGCAGAGATA
TGGAGGAACACCAAGTGGCGAAGGCGACTTCTGGTCTGTAAC TGACGCTGATGTGCAAGCGTGGGGAT
CAACAGGATTAGATACCTCGTAGTCCACGCCGTAACGATGAGTGCTAAGTGTTAGGGGGTTTCCGCC
CCTTAGTGCTGCAGCTAACGCATTAAGCACTCCGCCCTGGGGAGTACGACCGCAAGGTTGAAACTCAAAGG
AATTGACGGGACCCGCAACAAGCGGTGGAGCATGTGGTTAATTCGAAGCAACGCGAAGAACCTTACCAA
ATCTTGACATCCCTTGACAACCTTAGAGATAGAGCCTTCCCTTCGGGGGACAAAGTGACAGGTGGTGCA
TGGTTGTCGTGAGTCTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCTTAAAGCTAGT
TGCCATCATTAAGTTGGGCACTCTAAGTTGACTGCCGGTGACAAACCGGAGGAGGAGTGGGGATGACGTCA
AATCATCATGCCCTTATGATTTGGGCTACACACGTGCTACAATGGACAATACAAGGGCAGCGAAACCG
CGAGGTCAAGCAAATCCCATAAAGTTGTTCTCAGTTCGGATTGTAGTCTGCAACTCGACTACATGAAGCT
GGAATCGCTAGTAATCGTAGATCAGCATGTACGGTGAATACGTTCCCGGGCTTGTACACACCGCCCGT
CACACCAGAGAGTTTGAACACCCGAAGCGGTGGAGTAACCTTTTAGGAGCTAGCCGTGAAGGTGGG
ACAAATGATTGGGGTGAAGTCGTAACAGCT
```

[MAFFT alignment and NJ / UPGMA phylogeny](#)

Ctrl + click to follow link to MAFFT website

I. Having your sequences loaded and aligned.

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

[Download version](#)
[Mac OS X](#)
[Windows](#)
[Linux](#)
[Source](#)

[Online version](#)
[Alignment](#)
[mafft --add](#)
[Merge](#)
[Phylogeny](#)
[Rough tree](#)
[Merits / limitations](#)
[Algorithms](#)
[Tips](#)
[Benchmarks](#)
[Feedback](#)

To avoid overload, try [a light-weight option](#), for virus genomes (2024/Aug).

The function to email at the end of calculation was not working in August, 2024. Fixed on Aug 23.

For a large number of short sequences, try [an experimental service](#).

[Experimental service for aligning raw reads \(Updated, 2023/Nov\)](#)

If you need an MSA of only a specific region, then [try extracting the region first](#) (2022/Oct). **New!**

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:

Paste protein or DNA sequences in fasta format. [Example](#)

or upload a **plain text** file:

Choose File

No file chosen

ssRNA Seq Euk Archaea Bact

1

2 scroll down

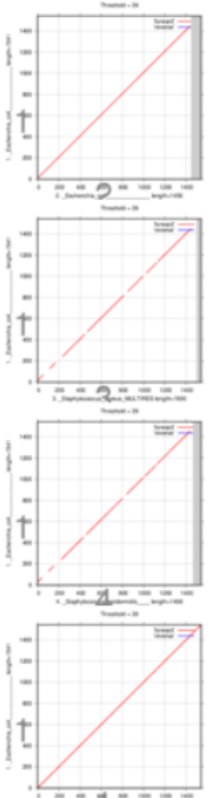
Submit

Reset

II. Accessing and visualizing the multiple sequence alignment.

[LAST](#) hits (score>39) between the top sequence and the others.

[Open all plots](#)



Be careful if there are **blue lines**. By default, MAFFT

[Clustal format](#) | [Fasta format](#) | MAFFT result | [View](#) | [Tree](#) | [Refine dataset](#) | [Return to home](#)

[View](#)

1. click here

[Reformat](#) to GCG, PHYLIP, MSF, NEXUS, uppercase/lowercase, etc. with Readseq

[GUIDANCE2](#) computes the residue-wise confidence scores and extracts well-aligned residues.

[Refine dataset](#)

[Phylogenetic tree](#)

MAFFT-L-INS-i Result

CLUSTAL format alignment by MAFFT (v7.511)

```
Escherichi aaattgaagagtttgatcatggctcagattgaacgctggcggcaggcctaacacatgca
Escherichi -----agtttgatcatggctcagattgaacgctggcggcaggcctaacacatgca
Pseudomona -----gagtttgatcatggctcagattgaacgctggcggcaggcctaacacatgca
Pseudomona -----agagtttgatcctggctcagattgaacgctggcggcaggcctaacacatgca
Pseudomona -----agtttgatcctggctcagattgaacgctggcggcaggcctaacacatgca
Staphyloco -----atcctggctcaggatgaacgctggcggcgctgcctaatacatgca
UNKNOWN -----atcctggctcaggatgaacgctggcggcgctgcctaatacatgca
Staphyloco -----gatgaacgctggcggcgctgcctaatacatgca
Staphyloco -----ctcaggatgaacgagcggcgctgcctaatacatgca
. ***** .*****.*****.*****
```

```
Escherichi gtcgaacggtaacaggaagaagcgttgct-ctttgctgacgagtggcggacgggtgagta
Escherichi gtcgaacggtaacaggaagcagcgttgctgctttgctgacgagtggcggacgggtgagta
Pseudomona gtcgagcggatg---aagggagccttgct---cctggattcagcggcgacgggtgagta
Pseudomona gtcgagcggtag---agagaagccttgct---tctcttgagagcggcgacgggtgagta
Pseudomona gtcgagcggatg---agaagagccttgct---cttcgattcagcggcgacgggtgagta
Staphyloco gtcgagcgaacg-gacgagaagccttgct-tctctgatgttagcggcgacgggtgagtaa
UNKNOWN gtcgagcgaacg-gacgagaagccttgct-tctctgatgttagcggcgacgggtgagtaa
Staphyloco gtcgagcgaaca-gacgagagccttgct-cctctgacgttagcggcgacgggtgagtaa
Staphyloco gtcgagcgaaca-gataaggagccttgct-cctttgacgttagcggcgacgggtgagtaa
*****.*****.*****.*****
```

[Clustal format](#) | [Fasta format](#) | [MAFFT result](#) | [View](#) | [Tree](#) | [Refine dataset](#) | [Return to home](#)

MAFFT MSA Viewer

[View this MSA in new window](#) (in alpha testing)

2. click here

MSAViewer

[Start MSAViewer in this window](#)

[Start MSAViewer in new window](#)

MSAViewer runs on any modern browser.

Reference:

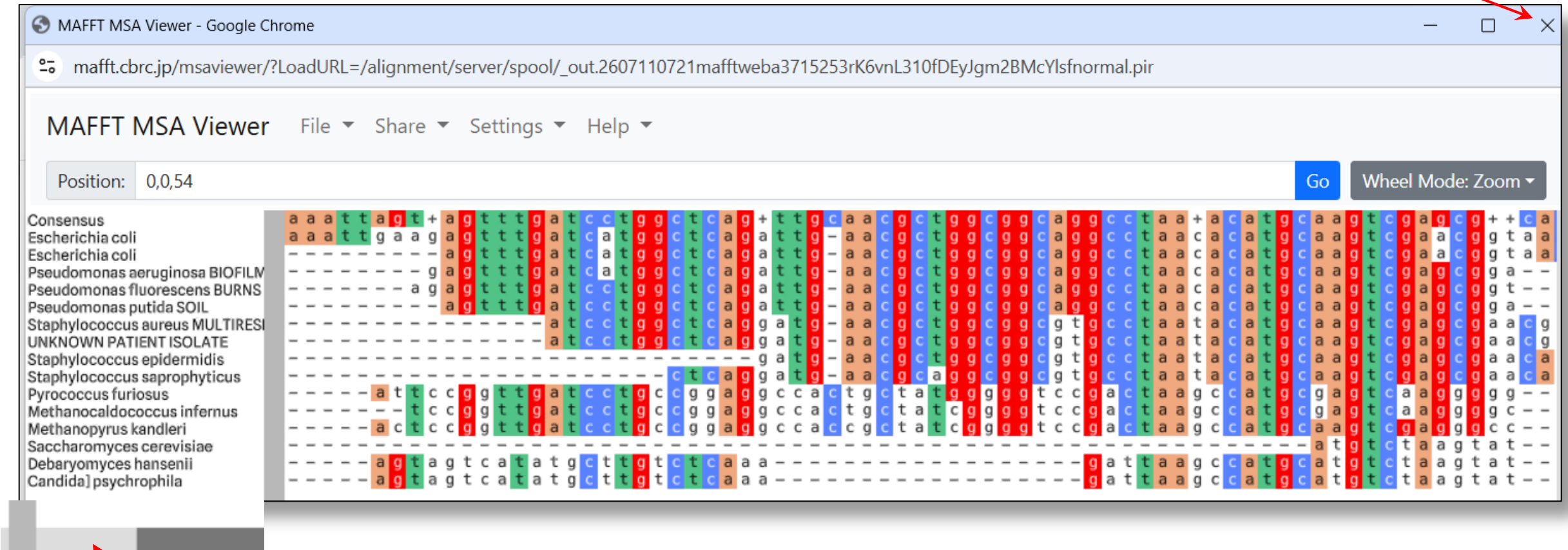
[Yachdav et al. \(2016\)](#)

MSAViewer Home:

<http://msa.biojs.net/>

II. Accessing and visualizing the multiple sequence alignment.

✓ 2. close alignment window



1. scroll bar

Multiple sequence alignments can be used for:

- building **phylogenetic trees**
- finding **conserved regions**
- designing **primers / probes for DNA analyses**
- infer function of unknown DNA, RNA or Protein Sequences

III. Accessing and visualizing the phylogenetic tree constructed from the multiple sequence alignment.

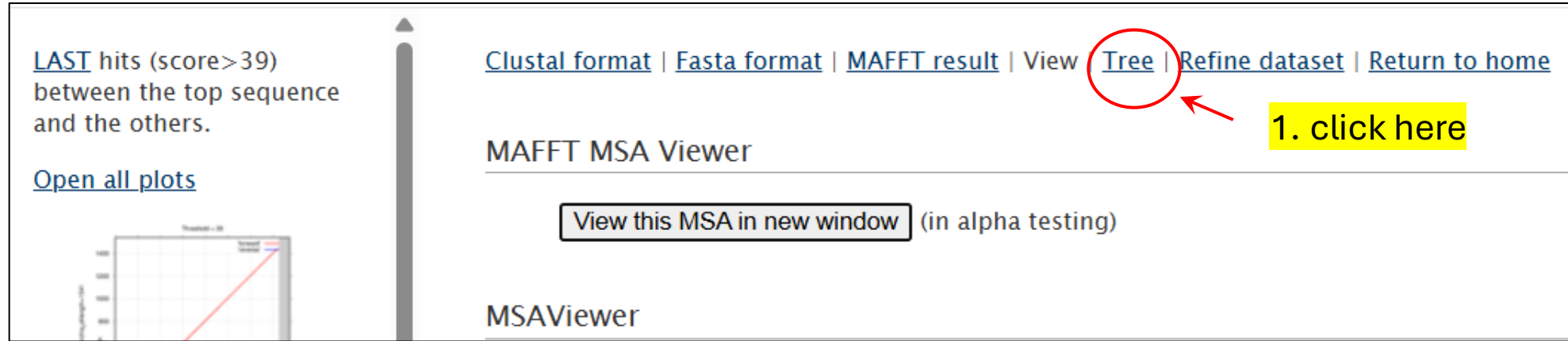
[LAST hits \(score>39\)](#) between the top sequence and the others.
[Open all plots](#)

[Clustal format](#) | [Fasta format](#) | [MAFFT result](#) | View [Tree](#) | [Refine dataset](#) | [Return to home](#)

MAFFT MSA Viewer

[View this MSA in new window](#) (in alpha testing)

MSAViewer



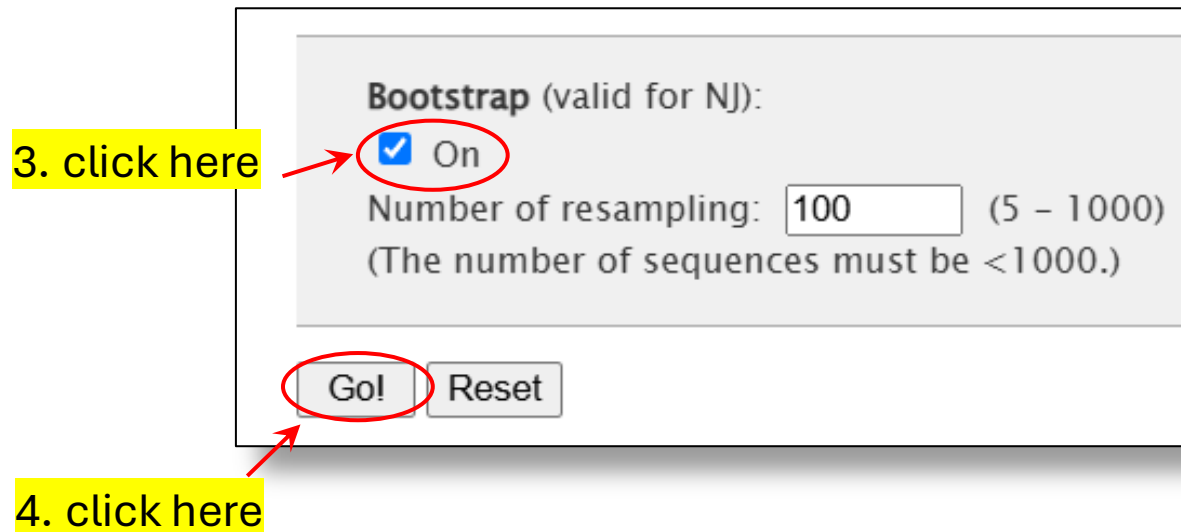
↓ 2. scroll down

3. click here → ☒ On

Bootstrap (valid for NJ):

Number of resampling: (5 – 1000)
(The number of sequences must be <1000.)

4. click here → [Go!](#) [Reset](#)



[Clustal format](#) | [Fasta format](#) | [MAFFT result](#) | [View](#) | [Tree](#) | [Refine dataset](#) | [Return to home](#)

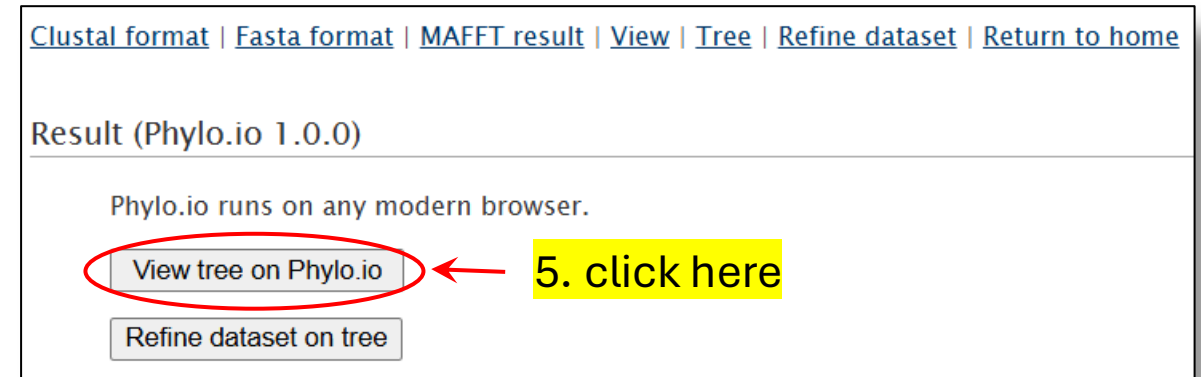
Result (Phylo.io 1.0.0)

Phylo.io runs on any modern browser.

[View tree on Phylo.io](#)

[Refine dataset on tree](#)

5. click here →

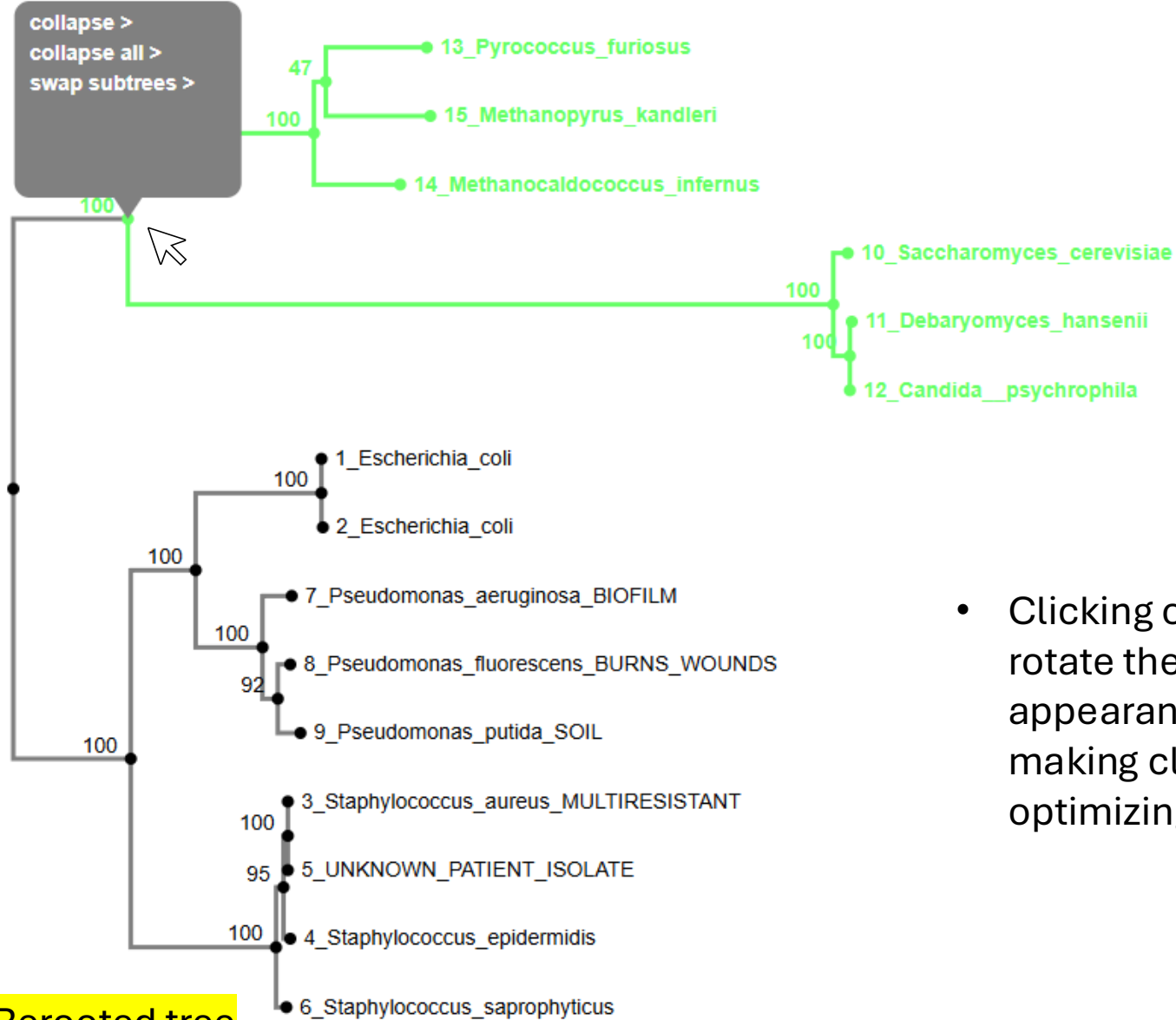


IV. Adjusting the Appearance of the Phylogenetic Tree to Improve Cluster Visualization

- The tree construction method used here is limited to inferring which taxa are more closely related, but it does not determine the position of the common ancestor. Therefore, the tree must be manually rooted using prior knowledge of well-established evolutionary relationships.
- To reroot the tree, hover the mouse pointer over the desired branch. A pop-up menu will appear with the Reroot option. Click Reroot, and the tree will be redrawn using that branch (or group) as the new root.



IV. Adjusting the Appearance of the Phylogenetic Tree to Improve Cluster Visualization



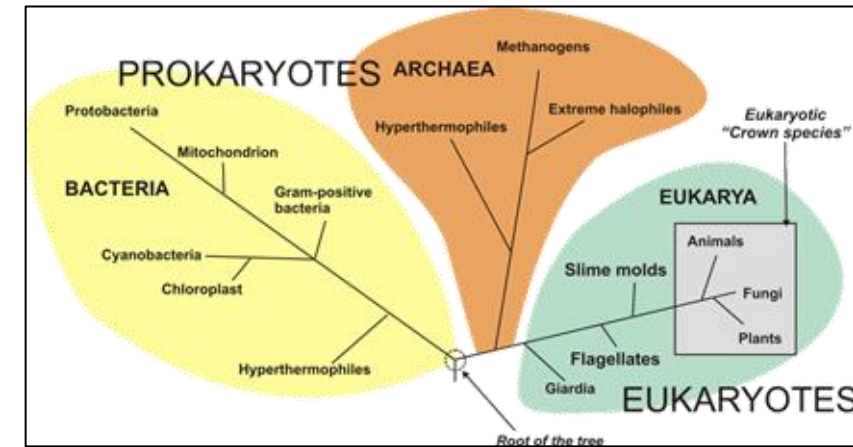
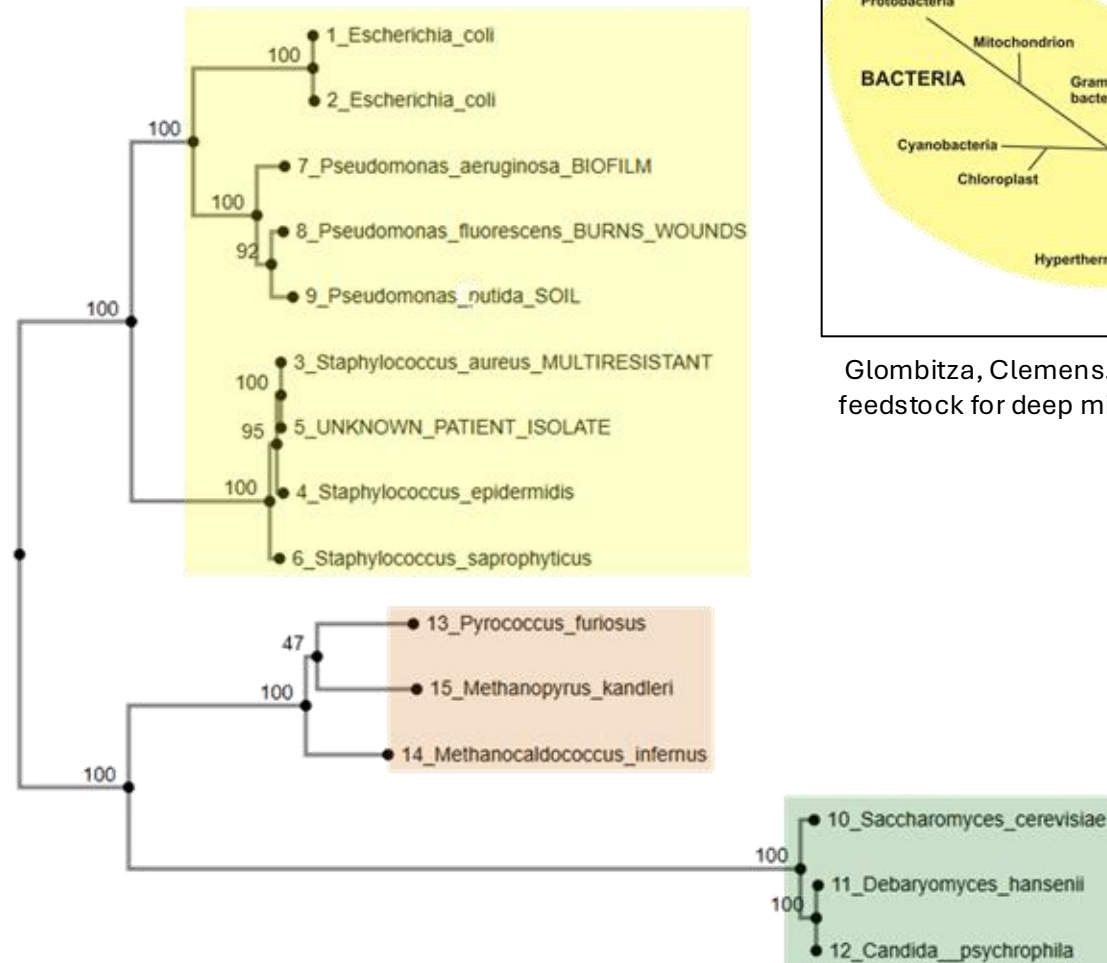
Rerooted tree

- Clicking on an internal node opens a menu that lets you rotate the connected branches. This changes the tree's appearance without altering the genetic relationships, making clustering patterns easier to visualize while optimizing spacing.

IV. Adjusting the Appearance of the Phylogenetic Tree to Improve Cluster Visualization

Adjusting the Appearance of the Phylogenetic Tree to Improve Cluster Visualization

1. **Rotate branches at internal nodes** to improve the visual arrangement of the tree **without changing the evolutionary relationships (tree topology).**
2. **Branch swapping groups closely related organisms together**, making major evolutionary clusters easier to recognize and interpret.
3. **Arrange the tree to resemble benchmark phylogenies**, such as **Carl Woese's Tree of Life**, which separates organisms into the three domains: **Bacteria, Archaea, and Eukarya.**
4. **Highlight major groups with transparent colored shapes** to emphasize taxonomic or evolutionary groups using transparent boxes (Select the shape box → Shape Format → Shape Fill → More Fill Colors → Transparency → set to 79% → OK)
5. **An optimized tree layout improves communication**, allowing genetic relationships and clustering patterns to be understood quickly by students and researchers alike.



Glombitza, Clemens. (2010). New Zealand coals - A potential feedstock for deep microbial life. 10.14279/depositonce-2755.

By organizing your sequences into evolutionary clusters, you construct your own simplified "Tree of Life," following the same principles Carl Woese used when he analyzed small-subunit ribosomal RNA sequences to uncover the three domains of life.

V. How to Read a Phylogenetic Tree

1. Each **branch** represents an evolutionary path connecting organisms that share a common ancestor.
2. **Branch length** shows the amount of genetic difference. Longer branches mean the DNA sequences are more different, while shorter branches mean they are more similar.
3. Very **short or no visible branch length** means the DNA sequences are identical or almost identical for the gene being analyzed.
4. Each **node** (branching point) represents a common ancestor. Organisms that meet at the same node are more closely related than organisms connected farther down the tree.
5. Sequences that cluster together are more closely related genetically than sequences found in different clusters.
6. **Numbers at the nodes (bootstrap values)** show how confident the computer is in each grouping. Values close to 100 indicate very strong support for that relationship.
7. The branches can be rotated to make clusters easier to see. Rotating branches does not change the evolutionary relationships—it only changes how the tree looks.

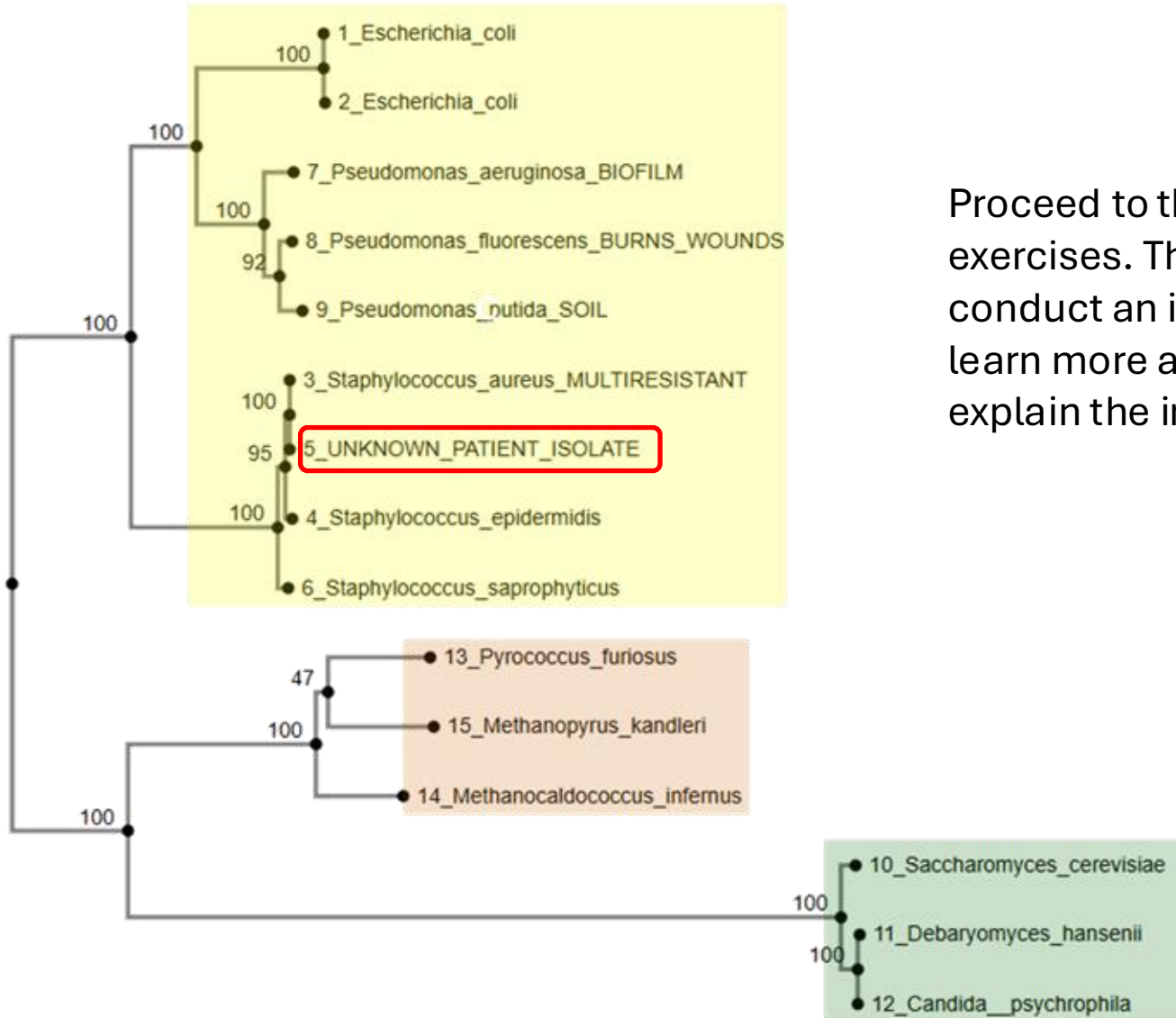


Take-Home Message

The branching pattern tells you who is most closely related, while the branch lengths show how much genetic change separates those organisms.

IV. Interpreting the clustering patterns depicted in the tree

Using the genetic relationships shown in the tree, infer the identity and likely biological characteristics of unknown patient isolate 5 (highlighted in red). Which known strain is it most closely related to?



Proceed to the worksheet and complete the background exercises. Then, complete the section that asks you to conduct an internet search using peer-reviewed articles to learn more about your unknown bacterial sequence and explain the implications of your results.