Updated: October 2021

The *Wolbachia* Project

Discover the Microbes Within!

Lab 5: Bioinformatics III

Wolbachia Phylogenetics

Project Guide

The Wolbachia Project

The *Wolbachia* Project: Discover the Microbes Within! was developed by a collaboration of scientists, educators, and outreach specialists. It is directed by the Bordenstein Lab at Vanderbilt University.
https://www.vanderbilt.edu/wolbachiaproject

Activity at a Glance

Goals

- To generate a phylogenetic tree of *Wolbachia*
- To determine the relatedness of an unknown sequence to those of known *Wolbachia* strains and identify Supergroup designation

Learning Objectives

Upon completion of this activity, students will build a phylogenetic tree to explore the relatedness of their sequence(s) to other *Wolbachia* strains within the NCBI database.

Prerequisite Skills

While no computer programming skills are necessary to complete this work, prior exposure to personal computers and the Internet is assumed.

Teaching Time: One to two class periods

Recommended Background Reading

This activity discusses *Wolbachia* Supergroups. For a quick refresher, review the Supergroup discussion in Lab 5: *Wolbachia* Identification and Naming.

The following online textbooks provide text, videos, and assessment materials for understanding the basics of phylogenetics.

CK-12: Biology for High School

- https://flexbooks.ck12.org/cbook/ck-12-biology-flexbook-2.0/section/5.11

Khan Academy: AP/College Biology

- https://www.khanacademy.org/science/ap-biology/natural-selection

OpenStax: Biology 2e

- https://openstax.org/books/biology-2e/pages/20-introduction

Required Resources

- Computer with internet browser, such as Firefox or Chrome
- Phylogenetic analysis software, NGPhylogeny.fr - https://ngphylogeny.fr
- DNA Sequence Files: https://www.vanderbilt.edu/wolbachiaproject/lab-5-dna-sequences/#moduleiii

Multiple software options are available for building phylogenetic trees. NGPhylogeny.fr is highlighted here due to its user-friendly interface and online, cross-platform accessibility. Another highly recommended tool is MEGA X (https://www.megasoftware.net/home).

Introduction to Phylogenetics: Reading a Tree

Phylogenetics is the study of evolutionary relatedness among biological organisms. Phylogenetic trees are generally based on molecular data, such as DNA or amino acid sequence, and use tree-like branching patterns to illustrate evolutionary histories (Fig 5.1). The tips of each branch represent a single sequence or organism, termed **taxon** (plural: taxa). Each **node** on the tree represents the common ancestor for all taxa branching out of that node. Clusters of taxa that originate from the same ancestral node are called **clades**.

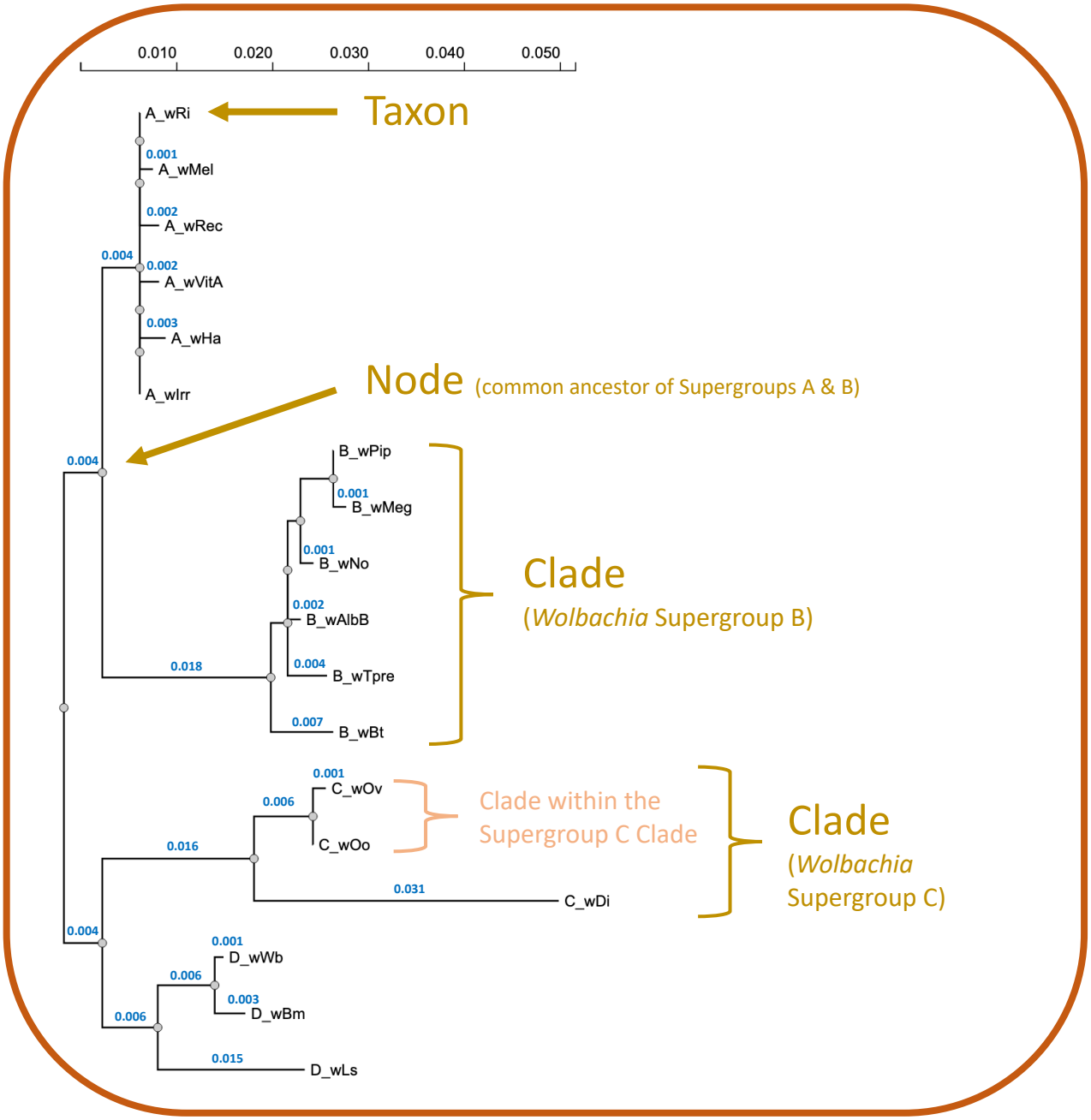
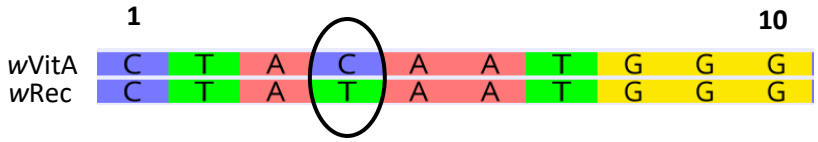


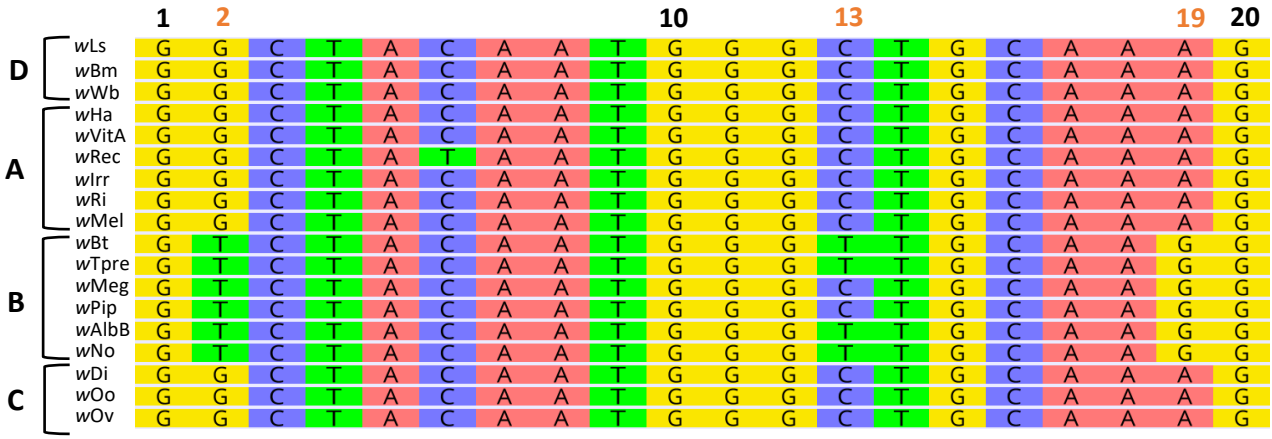
Figure 5.1

Introduction to Phylogenetics: Alignments

The length of each branch represents the evolutionary time between two nodes. This is often shown as **substitutions per site** (shown in blue in Fig 5.1). In simplest forms, this can be calculated by aligning the sequences and dividing the number of nucleotide differences by sequence length. In the example below, there is one base pair substitution (C → T) across 10 nucleotide sites. Therefore, there are 1/10 = 0.1 substitutions per site.



However, the reality of nucleotide alignments is much more complex. Substitutions may not be universal across all sites, substitutions occur at different rates (i.e., C → T vs G → T), some substitutions are synonymous (resulting in same amino acid product) whereas others are nonsynonymous (resulting in different amino acid product), etc. Therefore, software algorithms incorporate evolutionary models to better assess genetic change.



The above alignment features representatives from four *Wolbachia* Supergroups. Within the B-Supergroup, notice unique base pair substitutions at positions 2 and 19 relative to all other Supergroups. Position 13, however, is heterologous across Supergroup B. Which two taxa are **divergent** (or different) from the other B-*Wolbachia* at this site? Notice how this correlates with a smaller clade within the larger Supergroup B clade in Fig 5.1.

Introduction to Phylogenetics: Types of Trees

Rooted trees feature a distinct node, or root, that serves as the ancestral group for all taxa in the tree. The most common way to root a tree is by using an ancestral **outgroup**, a taxon that is known to be more distantly related than all other taxa in the tree. **Unrooted trees**, however, are necessary when ancestry is unknown (Fig 5.2). In the case of *Wolbachia*, the ancestral strain is unknown so most trees will be unrooted. We can, however, include taxa such as *Ehrlichia* or *Anaplasma* as outgroups because they are closely related yet outside the group of interest (*Wolbachia*). While this may not provide concise ancestral information (a true root), it will create a meaningful tree showing the relationship of all *Wolbachia* taxa relative to closely related taxa (Fig 5.3). Finally, unrooted trees are sometimes **midpoint rooted** (Fig 5.4). The hypothetical root can be placed midpoint in the tree if (i) the tree is balanced and closely related clades are separated by a long branch or (ii) taxa are evolving at the same rate.

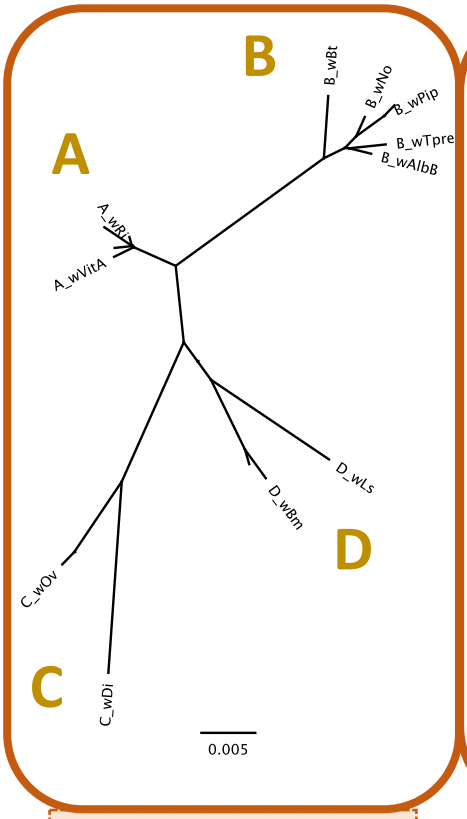


Figure 5.2. Unrooted *Wolbachia* tree illustrates the method of Supergroup (A-B-C-D) designation based on unique clades.

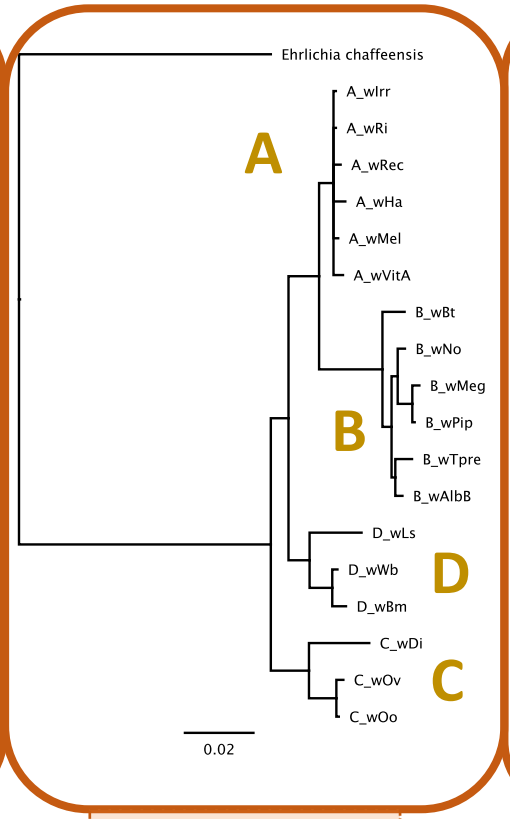


Figure 5.3. Outgroup rooting with *Ehrlichia chaffeensis* allows the unrooted *Wolbachia* tree to be ladderized.

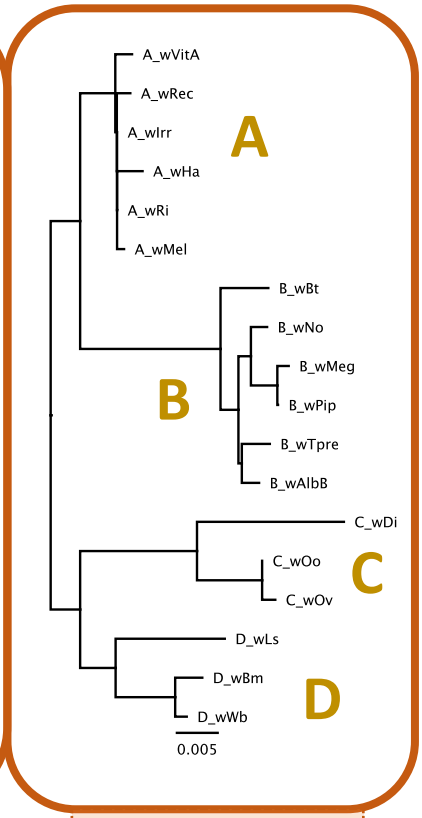


Figure 5.4. Midpoint rooting between clades A/B and C/D allows the unrooted *Wolbachia* tree to be ladderized.

Technical Overview: FASTA

FASTA Format

In bioinformatics, FASTA is a text-based format for representing either nucleotide (DNA/RNA) or peptide (amino acid) sequences. The file must have a top line that begins with '>' and includes a sequence name and/or short description. The actual sequence comprises the rest of the file. For example:

1. This file contains one sequence, *w*Mel, with only the name as a short descriptor on the top line.

```
>wMel
AGAGTTTGATCCTAGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGAGTTATATT
GTAGCTTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTACCTAGTAGTACGGAA
TAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAAAATTTATTGCTATTAGATGAGCCT
ATATTAGATTAGCTAGTTGGTGGAGTAATAGCCTACCAAGGCAATGATCTATAGCTGATCTGAGAGGATGA
TCAGCCACACTGGAACTGAGATACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACA
```

2. This file contains three separate sequences, each with a longer descriptor on the top line. The '>' line indicates it is the beginning of a new sequence.

```
>A_wHa Wolbachia endosymbiont of Drosophila simulans
AGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGAGTTATATT
GTAGCCTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTACCTAGTAGTACGGAA
TAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAAAATTTATTGCTATTAGATGAGC>
A_wIrr Wolbachia endosymbiont of Haematobia irritans
AAATTTGAGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGA
GTTATATTGTAGCTTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTACCTAGTA
GTACGGAATAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAAAATTTATTGCTATTA
>A_wRec Wolbachia endosymbiont of Drosophila recens
AGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGAAGTTATATT
GTAGCTTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTACCTAGTAGTACGGAA
TAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAAAATTTATTGCTATTAGATGAGC
```

For the purpose of this lab, the title of each FASTA sequence will be used as the corresponding taxon label on your phylogenetic tree.

FASTA File Name

Just as PDF documents are identified with a .pdf extension, FASTA files use .fasta at the end of the file name.

Creating and Modifying a FASTA File

Any bioinformatics program (such as MEGA, SnapGene, or Geneious) can create and modify FASTA files. Alternatively, a FASTA file may be manually edited using a basic text editing program (i.e., TextEdit for Mac or Notepad for PC). Text can be added and deleted as long as it retains the FASTA format (above).

Pre-Lab Questions

1. Which two FASTA files are correctly formatted?

FASTA 1

>A_wHa Wolbachia endosymbiont of Drosophila simulans
AGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGA
GTTATATTGTAGCCTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTA
CCTAGTAGTACGGAATAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAA

FASTA 2

>A_wHa Wolbachia endosymbiont of Drosophila simulans AGTTCTGGTCCATGATGACCC
AGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGA
GTTATATTGTAGCCTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTA
CCTAGTAGTACGGAATAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAA

FASTA 3

A_wHa
AGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGA
GTTATATTGTAGCCTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTA
CCTAGTAGTACGGAATAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAA

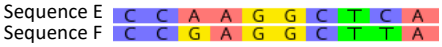
FASTA 4

>A_wHa
AGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGA
GTTATATTGTAGCCTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTA
CCTAGTAGTACGGAATAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAA
>A_wMel
AGAGTTTGATCCTAGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGA
GTTATATTGTAGCTTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTA
CCTAGTAGTACGGAATAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAA

2. For each short sequence alignment below, estimate the substitutions per site.

Alignment #1 = _____

Alignment #2 = _____



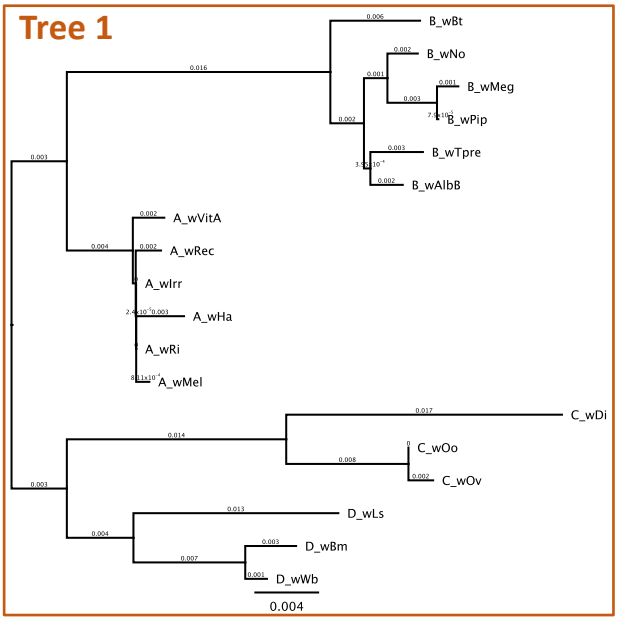
Alignment #3 = _____

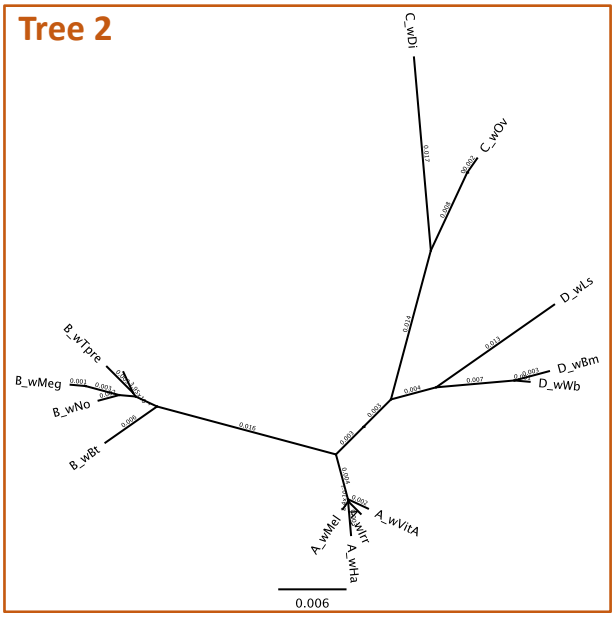
3. If the above sequence alignment is representative of the entire genome, which two genomes are most closely related?
   a. Alignment #1
   b. Alignment #2
   c. Alignment #3

4. If the above sequence alignment is representative of the entire genome, which two genomes are most divergent?
   a. Alignment #1
   b. Alignment #2
   c. Alignment #3

Pre-Lab Questions

5. In the unrooted trees below, each taxon label consists of 'Supergroup_strain name.' Label the four major clades corresponding to Supergroup A, B, C and D.





6. According to Tree 1, *w*Pip and *w*Meg are most closely related to which other *Wolbachia* strain?

7. In Tree 1, label the node representing the common ancestor of Supergroups A and B.

8. Which three taxa represent Supergroup D?

Lab Activity: *Wolbachia* Phylogenetics

MATERIALS

☐ ***Wolbachia* sequence(s):**
Use your FASTA sequence(s) from Module I or download Example *Wolbachia* Sequence(s):
https://www.vanderbilt.edu/wolbachiaproject/lab-5-dna-sequences/#moduleiii

☐ **FASTA Reference Sequences for *Wolbachia* Phylogenetics:**
https://www.vanderbilt.edu/wolbachiaproject/lab-5-dna-sequences/#moduleiii

☐ **Computer with Internet Access**

Step 1: Create a FASTA File

- Download the FASTA file of reference sequences for phylogenetics:
  - https://www.vanderbilt.edu/wolbachiaproject/lab-5-dna-sequences/#moduleiii
  - The majority of arthropod *Wolbachia* identified to-date fall within A and B Supergroups.
- Save the file to your Desktop or a specified folder on your Desktop.
- Open your file with a text editor (such as TextEdit for Mac, Notepad for Windows, Text for Chromebook).
- Scroll through the file. Each new sequence begins with > followed by the taxon name and description. A Supergroup prefix has been added to the beginning of each taxon name and complete names are listed in the Appendix.
  - *Note:* The text following '>' is the taxon name that will appear on your tree. If you make any changes to the FASTA file, make sure that the first line of each sequence begins with > and includes only the sequence name/description. The nucleotides sequences must begin on the second line.
- At the very end of the file, manually add your sequence. If you did not obtain a Wolbachia sequence for your arthropod, you may download an Example Wolbachia Sequence: https://www.vanderbilt.edu/wolbachiaproject/lab-5-dna-sequences/#moduleiii
  - The top line should simply be '>' followed by your Wolbachia strain name (for example, >wXXX) or sample ID
  - Hit enter to start a new line
  - Copy/paste your DNA sequence
- Select "File >> Save" to save your new FASTA sequence.
  - *Note:* The file extension must be .fasta. If the text editor changes to the extension to .txt, you will be unable to load the sequences into the bioinformatics pipeline. You must revert back to .fasta.

Step 2: Build a Phylogenetic Tree

- Open a browser and go to https://ngphylogeny.fr/
- Select One Click workflow
- From the left-hand column, you may select your preferred Tree Inference algorithm. For the purpose of this lab, use the default FastMe pathway.
- Under "Input file," use the "Choose File" button to load your FASTA sequence file.
- Select the blue Submit button. A new page will open.
- **Important:** note the URL for your results and/or enter your email address.
- The pipeline will show your tree being built in real-time.
    - **MAFFT** is a multiple sequence alignment tool that will create an alignment of all your uploaded sequences. This aligns each homologous nucleotide position. If you are comparing a 100-bp sequence with a 500-bp sequence, for example, you wouldn't start comparing them at position 1. Rather, you would find the overlapping region where the smaller sequence matches the larger sequence. Once this step is complete, you may click "MSAViewer" to visualize your alignment. Use the scroll bar immediately above the sequences to navigate around the alignment. A dash (–) means that there is no nucleotide at that position, either due to partial sequencing or an insertion/deletion event. Notice that your sequence is much shorter than the reference sequences. Does it align at the beginning (5') or end (3') of the reference sequences? Based on the alignment, does your sequence more closely resemble one Supergroup over the other? Select "Go back"
    - **BMGE** selects regions in the alignment that are suited for phylogenetic inference.
    - **FastME** provides the distance algorithms to build the phylogenetic tree. It will determine the relatedness of each sequence to other taxa and produce a tree.
    - **Newick Display** is the tree rendering software used to visualize the tree.

Step 3: Visualize and Modify the Tree

- Before modifying your tree, copy/paste the URL and save in a separate document.
- We recommend using the green "Viewer" button to make minor changes.
    - *Optional:* If you want more advanced options, select the yellow button to export your tree to iTol (Interactive Tree of Life, https://itol.embl.de/).
- Using the Viewer, hover the cursor over Ehrlichia chaffeensis and "Reroot at this node."
    - Do each of the major Subgroups form a clade?
    - You can "flip" the orientation of a clade by clicking the node (gray dot) and select "Swap subtree."
    - If you want to highlight your taxon, click on the branch (line) to turn it red.
- Visualization:
    - Phylogram (left menu) displays a tree where branch lengths (lines) are proportional to the amount of character change, or sequence divergence.

Dendrogram transforms the branches to a visually pleasing format that illustrates hierarchical cluster arrangement.
    - Click between the two. Which one is best illustrates the major clades? Which one best illustrates evolutionary relationships among the different sequences?
  - Linear, radial, and slanted are different formats to visualize the tree.
  - Display branch length will show the substitution rate.
  - Align text will align all taxon labels
  - Use the arrows and zoom features to the right of the tree to better visualize the phylogram.
  - Right-click >> Take a Screenshot

Advanced Option: Create a tree showing only *Wolbachia* Supergroups A and B

- Go back to Step 1 and open the original **FASTA Reference Sequences for *Wolbachia* Phylogenetics** using a Text Editor.
- Delete the *Ehrlichia* outgroup and all *Wolbachia* Supergroups *except* A and B. Make sure your sequences are still included.
- Repeat the activity. In Part 3, midpoint root the tree by selecting the node basal to Clades A/B and click "Reroot at this node."
- Are the results the same?

Did you determine a *putative* Supergroup classification for your Wolbachia strain(s)?

- We use the term "putative" because the 16S gene is just one indicator gene. Ideally, we would sequence multiple genes to confirm consensus for the Supergroup classification.

- Most arthropod sequences will likely fall within the A or B clades; most nematode sequences would likely fall within the C or D clades. If your sequence does not fall within a clade, refer to the initial chromatogram.
  - If the chromatogram was lower quality, this could represent a coinfection (your arthropod is infected with more than one Wolbachia strain), contamination, or a poor-quality sequencing run. In each of these cases, there may not be enough information to properly place your sequence.
  - If the chromatogram was high quality, your strain may belong to a less studied Supergroup. To properly classify, we would need to follow up by sequencing additional genes (or the entire genome).

Appendix:

Wolbachia strains included in this activity

Wolbachia Supergroup	Taxon Label	Complete Description
A	A_wHa	Wolbachia endosymbiont of Drosophila simulans from Hawaii
	A_wIrr	Wolbachia endosymbiont of Haematobia irritans
	A_wMel	Wolbachia endosymbiont of Drosophila melanogaster
	A_wRec	Wolbachia endosymbiont of Drosophila recens
	A_wRi	Wolbachia endosymbiont of Drosophila simulans from Riverside
	A_wVitA	Wolbachia endosymbiont of Nasonia vitripennis
B	B_wAlbB	Wolbachia endosymbiont of Aedes albopictus
	B_wBt	Wolbachia endosymbiont of Bemisia tabaci
	B_wMeg	Wolbachia endosymbiont of Chrysomya megacephala
	B_wNo	Wolbachia endosymbiont of Drosophila simulans from Nouméa
	B_wPip	Wolbachia endosymbiont of Culex pipiens
	B_wTpre	Wolbachia endosymbiont of Trichogramma pretiosum
C	C_wDi	Wolbachia endosymbiont of Dirofilaria immitis
	C_wOo	Wolbachia endosymbiont of Onchocerca ochengi
	C_wOv	Wolbachia endosymbiont of Onchocerca volvulus
D	D_wBm	Wolbachia endosymbiont of Brugia malayi
	D_wLs	Wolbachia endosymbiont of Litomosoides sigmodontis
	D_wWb	Wolbachia endosymbiont of Wuchereria bancrofti
E	E_wFol	Wolbachia endosymbiont of Folsomia candida
F	F_wCle	Wolbachia endosymbiont of Cimex lectularius
	F_wMo	Wolbachia endosymbiont of Mansonella ozzardi