Updated: September 2023

The Wolbachia Project

*Discover the Microbes Within!*

# Lab 5: Introduction to Bioinformatics

## *Sanger Sequence Analysis, BLAST Searching, and Phylogenetics*

**Project Guide**

This activity was written by Sarah Bordenstein, Penn State University. 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. https://www.wolbachiaproject.org


*** Image credit for Activity 1:** (1) AJCI, CC BY-SA 2.0; (2) Public domain; (3) Public domain; (4) Wildreturn CC BY 2.0; (5) AJCI, CC BY-SA 2.0; (6) Ryan Hodnett, CC BY-SA 4.0; (7) April Nobile, CC BY-SA 4.0; (8) Mylène Durant, CC BY-SA 4.0; (9) Katja Schulz, CC BY 2.0; (10) Ryan Thumboor, CC BY 2.0; (11) Martin Cooper, CC BY 2.0

**Figure 5.5** created with BioRender.com.
**Word Search** & **Crossword Puzzle** created with TheTeachersCorner.net.

# Introduction to Bioinformatics

This Project Guide introduces participants to **bioinformatics**, the use of computational tools to understand biological data. Based on the insect barcoding gene, mitochondrial cytochrome c oxidase I (CO1), participants will analyze short sequences of DNA to determine taxonomic identity and evolutionary relatedness to other arthropods. Skills learned in this guide may be applied to individual Sanger sequencing results from The *Wolbachia* Project.

***Goals***

- To analyze and manipulate Sanger sequencing chromatograms
- To determine the taxonomic identification of unknown arthropods using NCBI BLAST
- To visualize the evolutionary relatedness of arthropods using phylogenetics

***Learning Objectives***

Upon completion of this activity, participants will (i) compare the identification of arthropods using morphological characterization vs. DNA sequencing; (ii) learn how to convert .ab1 chromatograms to fasta files; (iii) become familiar with NCBI BLAST; and (iv) build a phylogenetic tree to explore the evolutionary relatedness of arthropods.

***Prerequisite Skills***

No computer programming skills are necessary to complete this work; prior exposure to personal computers and the Internet is assumed. A review of **transcription** (DNA → RNA), **translation** (RNA → protein), and the **genetic code** (relationship between codons and amino acids, see *Appendix A: Codon Table*) is highly recommended.

***Teaching Time:*** Two class periods

***Recommended Background Reading & Tutorials***

**The Wolbachia Project: NCBI Taxonomy & BLAST Searching**

- https://wolbachiaproject.org/bioinformatics/

**LabXchange:**

- https://www.labxchange.org/
- Search the library for the "Central Dogma" Pathway to view videos and interactive simulations of transcription and translation.

***Required Resources***

- Computer with internet browser, such as Firefox or Chrome
- Download SnapGene Viewer (free software) - https://www.snapgene.com/snapgene-viewer
- Online access to NGPhylogeny.fr - https://ngphylogeny.fr

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. It provides a quick tree for educational purposes, while other software options (Geneious, MEGA, etc.) allow for higher scientific accuracy and are recommended for publication-quality trees. Like all web-based programs, users may experience issues with speed and responsiveness. Appendix B lists alternative options for building a phylogenetic tree in the classroom setting.

# Introduction to Bioinformatics: Phylogenetics

**Phylogenetics** is the study of evolutionary relatedness among biological organisms. Phylogenetic trees are generally based on molecular data (DNA or amino acid sequence) and use tree-like branching patterns to illustrate evolutionary histories (Fig 5.1). The tips of each branch, often referred to as leaves, represent the **operational taxonomic units** (OTUs) that are being compared. These might include genes, individuals, species, or populations. If the OTUs represent a formal taxonomic group, such as a species, they are termed **taxa** (singular: taxon). 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**.

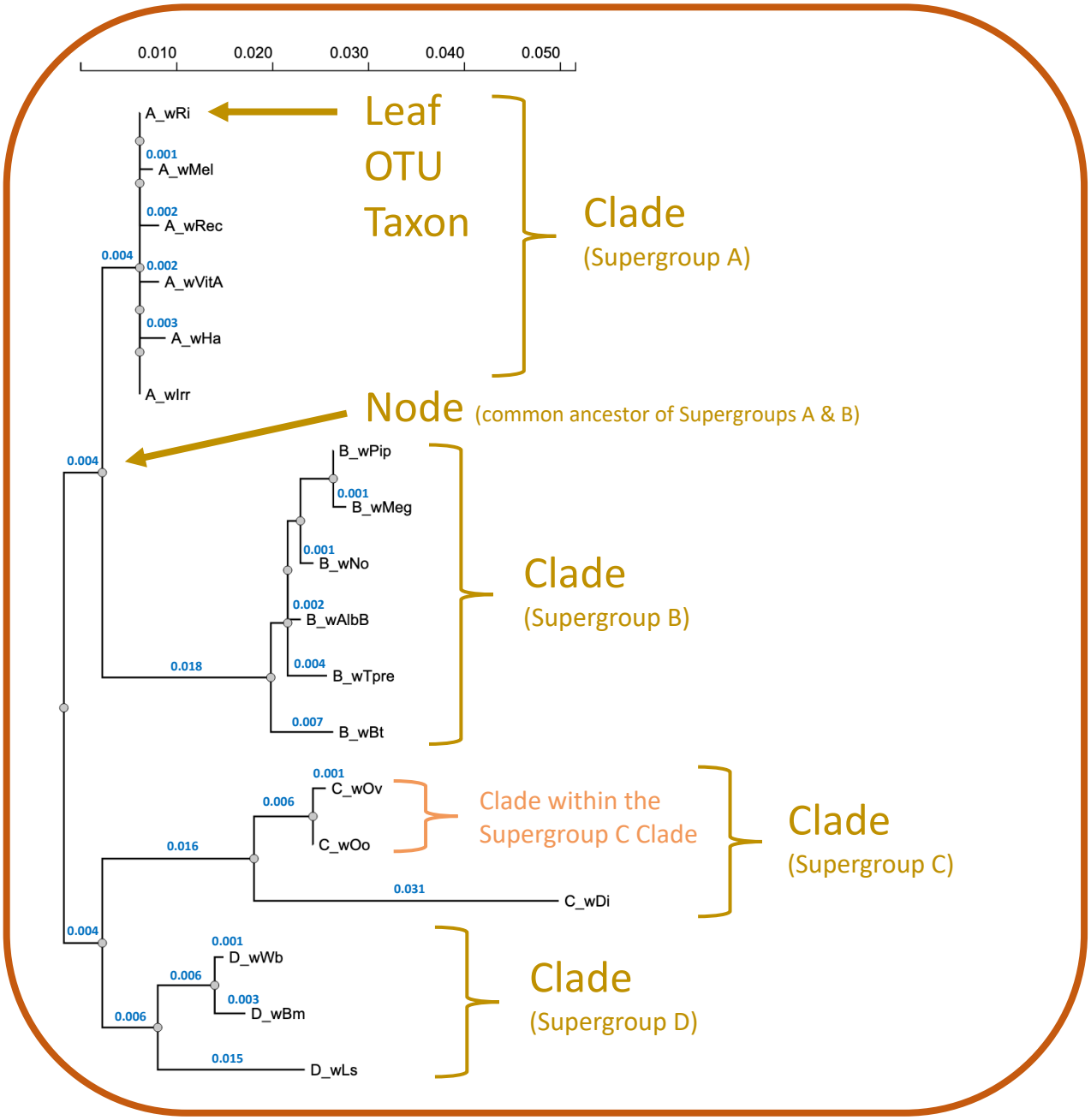



**Figure 5.1:** Unrooted phylogenetic tree of four *Wolbachia* supergroups.

# Introduction to Bioinformatics: 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* Supergroups, the ancestral strain is unknown so most trees will be unrooted. We can, however, include genera 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.

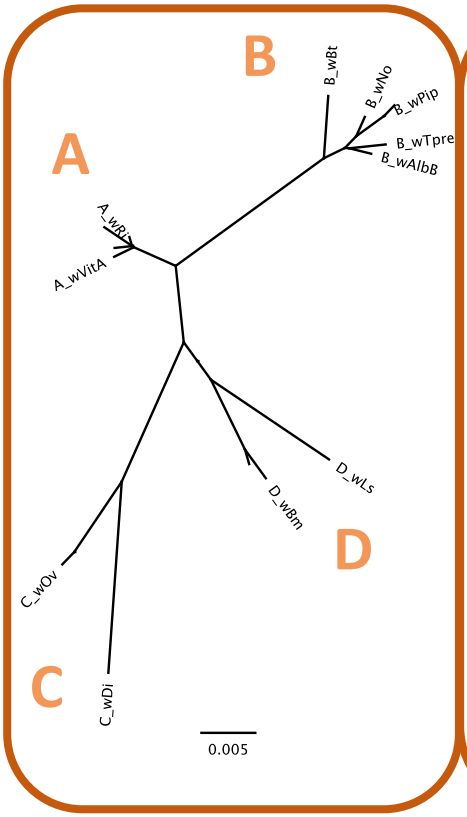



**Figure 5.2.** Unrooted *Wolbachia* tree illustrates the method of Supergroup (A-B-C-D) designation based on unique clades.

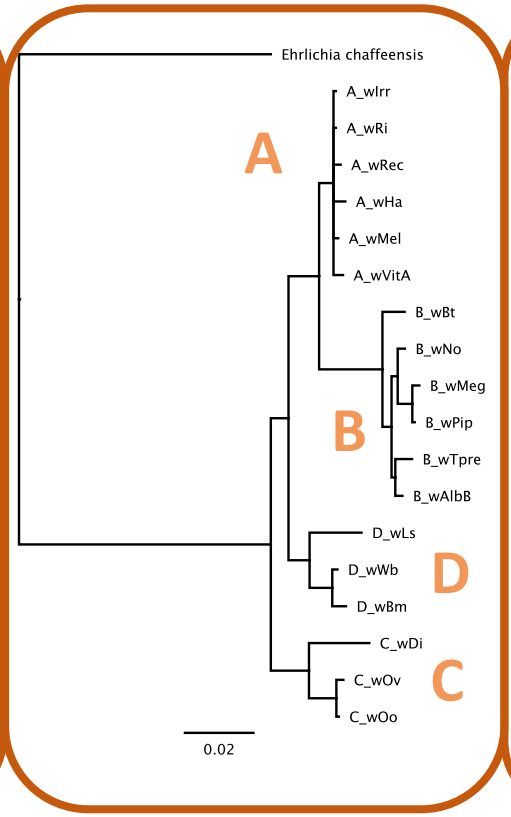



**Figure 5.3.** Outgroup rooting with *Ehrlichia chaffeensis* allows the unrooted *Wolbachia* tree to be ladderized.

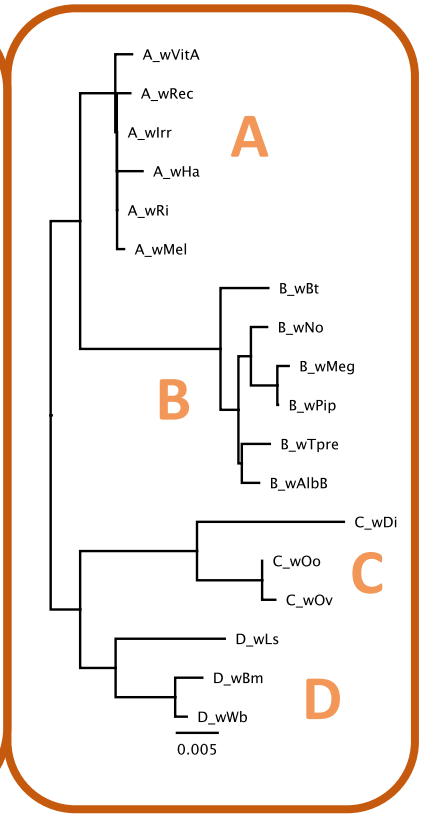



**Figure 5.4.** Midpoint rooting between clades A/B and C/D allows the unrooted *Wolbachia* tree to be ladderized.

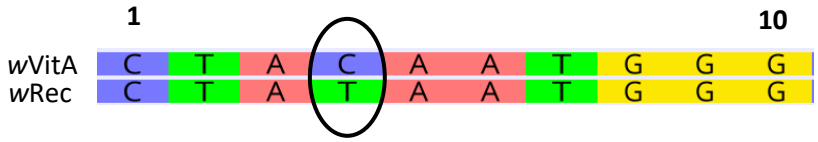
# Introduction to Bioinformatics: Alignments

The length of each branch on a phylogenetic tree represents the evolutionary distance between sequences. There are various approaches to constructing phylogenetic trees based on the dataset being analyzed; many methods begin with a distance estimation between nucleotide or amino acid sequences. The first step in distance-based phylogeny is the calculation of the pairwise distance, or **p-distance**, between two sequences. In simplest forms, this can be calculated by aligning two DNA sequences[1] and dividing the number of nucleotide differences ($n_d$) by sequence length, or the total number of nucleotide sites in the sequence ($n$).

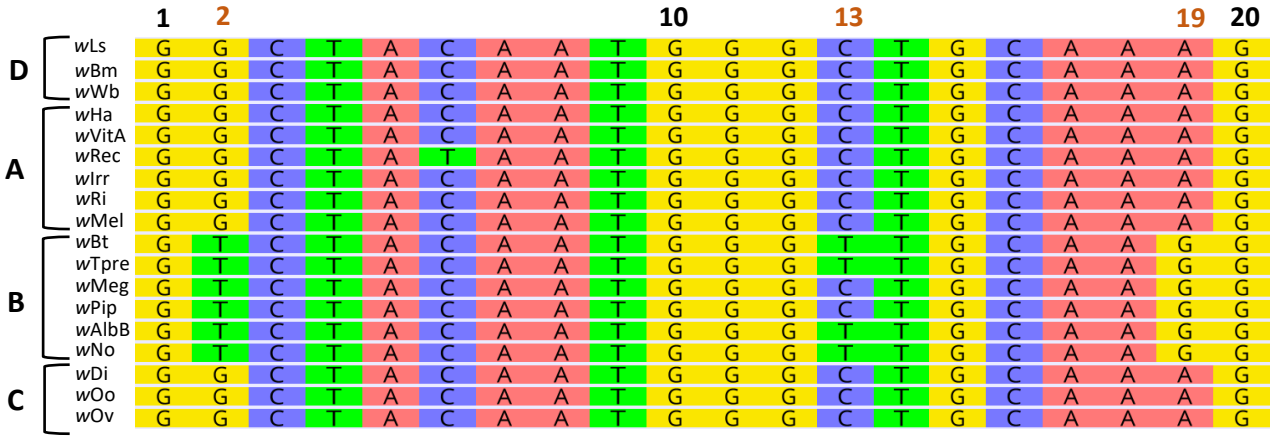
$$p = n_d/n$$

In the example below, there is one base pair substitution (C → T) across 10 nucleotide sites.

$$p = 1/10 = 0.10$$




Typically, an alignment would assess multiple sequences, each with hundreds to thousands of nucleotides. A higher p-distance (e.g., 0.30) would indicate that the sequences are more **divergent**, or different, relative to the above pairwise alignment. On the other hand, a p-distance of 0 would indicate that the sequences are identical with 100% nucleotide similarity.




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. These **mutations**, or changes in the DNA sequence, are specific to the B-Supergroup. Position 13, however, is variable within Supergroup B. Which two taxa are divergent from the four other B-*Wolbachia* at this site? Notice how this correlates with a smaller clade within the larger Supergroup B clade in Fig 5.1.

[1] Alignments compare a single strand of DNA. Bioinformaticians must confirm that the sequences are in the same orientation (such as "Forward", or 5' → 3') prior to the alignment. If a particular sequence does not align well, it may need to be reversed.

# Introduction to Bioinformatics: Point Mutations

Beyond the basic p-distance, many other variables in an alignment can inform evolutionary analyses. For example, **point mutations** are genetic mutations in which a single nucleotide in a genome is altered. Nucleotides can be substituted, inserted, or deleted in a genome.

***Nucleotide Substitutions***

**Transitions** involve the interchange of a purine with a purine (A ⇄ G) or a pyrimidine with a pyrimidine (C ⇄ T). **Transversions,** on the other hand, involve the interchange of a purine and a pyrimidine. As shown in Figure 5.5, transversions have a greater impact on genomic structure because a two-ring structure is exchanged with a one-ring structure, or vice versa.

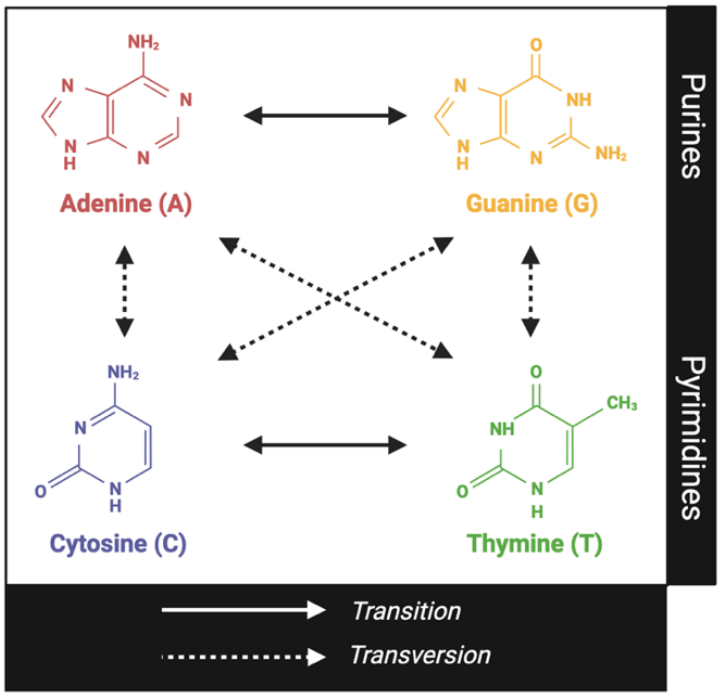



**Figure 5.5.** Transitions and transversions.

Both types of nucleotide substitutions - transitions and transversions – can influence the resulting protein sequence. **Nonsynonymous substitutions** are point mutations resulting in a different amino acid sequence. If the point mutation encodes the same amino acid, it is termed a **synonymous substitution.** Figure 5.6 illustrates the variation in amino acid sequences encoded by two different DNA sequences. Each **codon**, a sequence of three nucleotides that correlates to a specific amino acid, is marked #1-10 (see *Appendix A*: *Codon Table*). Codon #3 encodes a synonymous substitution because the correlating amino acid, threonine (Thr), remains the same even though the coding sequences are slightly different (ACA vs ACG) between the two nucleotide sequences. Codons #5, #7, and #10, however, encode nonsynonymous substitutions because the change in nucleotide sequence has an impact on the resulting amino acid products. A simple change in the amino acid sequence can impact both the structure and function of the resulting protein.

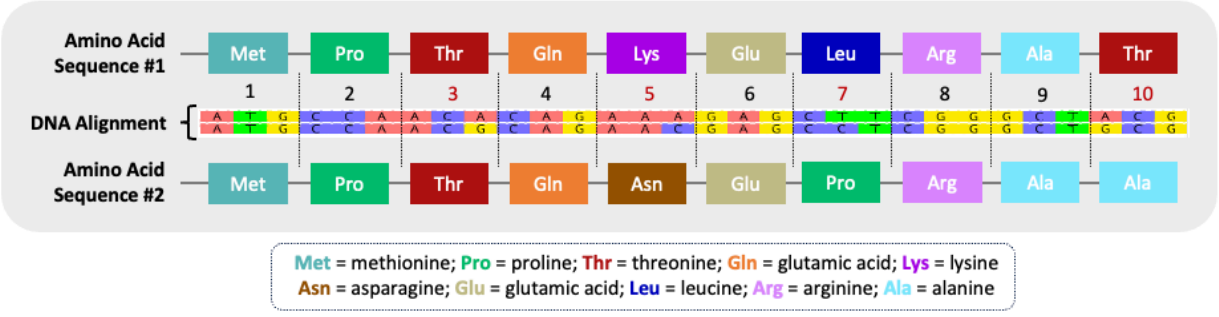



**Figure 5.6.** Synonymous and nonsynonymous substitutions.

***Nucleotide Insertions & Deletions***

Insertions and deletions in a nucleotide alignment are often referred to as **indels** (**in**sertions-**del**etions) because the term does not require information about the direction of mutation. When comparing two nucleotide sequences, for example, it is likely unknown if the indel is due to an insertion (gain of a nucleotide) in one sequence or a deletion (loss of a nucleotide) in the aligned sequenced. Indels are incredibly important to evolutionary genetics because they influence phenotypic traits, including human disease.

Indels will result in a **frameshift mutation** unless all three nucleotides in a codon are similarly inserted or deleted. As the term implies, the **reading frame** of the nucleotide sequence is shifted such that the amino acid sequence is altered. Figure 5.7 illustrates an indel at codon #6. Without comprehensive phylogenetic information, it is unknown whether this indel represents a deletion in the top sequence or an insertion in bottom sequence. In scenario (a), a single indel has drastic consequences for the downstream amino acid sequence and will result in a different protein product. In scenario (b), the insertion/deletion of an entire codon (3 nucleotides) results in only one amino acid change across the sequence. The phenotypic impact of this seemingly minor indel will depend on the specific role of the amino acid (glutamic acid) in overall structure and/or function of the protein.

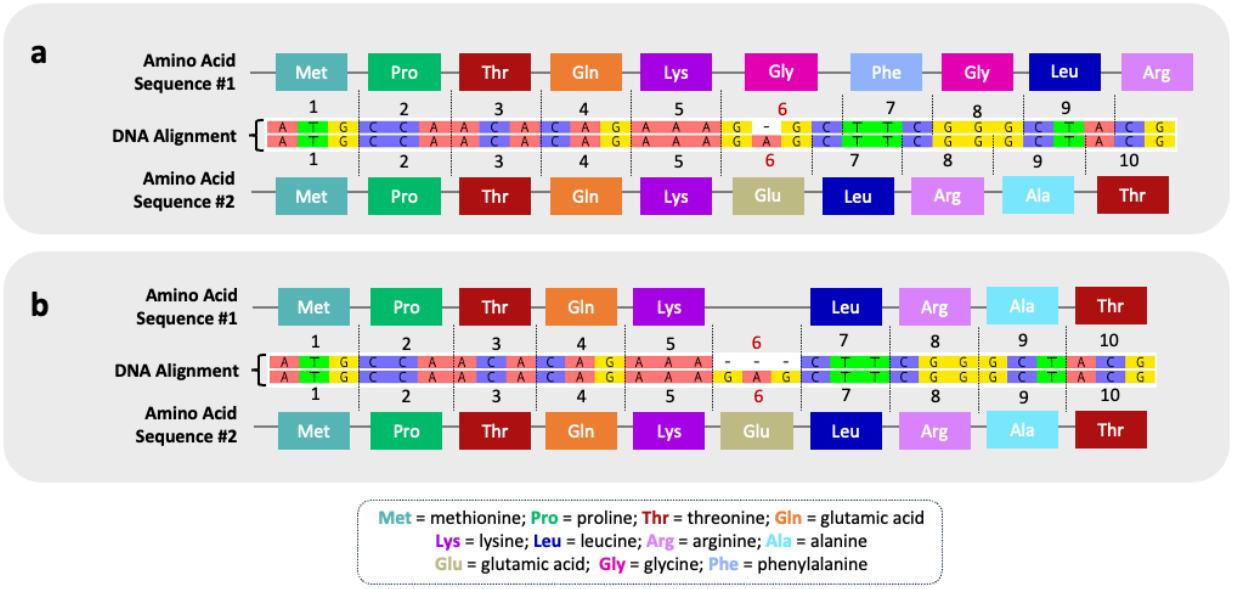



**Figure 5.7.** (a) Indels result in frameshift mutations unless (b) all three nucleotides of the codon are similarly inserted/deleted.

Together, point mutations occur at different evolutionary rates and have profoundly different impacts on the genome. When building phylogenetic trees, bioinformaticians must weigh the cost/benefit of using an advanced method with long computational time that better accounts for evolutionary variation vs. a simpler, less robust method that delivers fast results. For the purpose of this lab, we will focus on the latter.

# Technical Overview: Sanger Sequencing Files

The phylogenetic trees constructed in this lab activity are based on the insect barcoding gene, mitochondrial cytochrome c oxidase I (CO1). All sequences have been obtained by DNA extractions from individual arthropods followed by Sanger sequencing. A few simple steps, covered in Activity 2, are needed to convert the raw data file (.ab1) to a usable format (.fasta) for phylogenetic analysis. Before we begin, let's review some basic concepts:

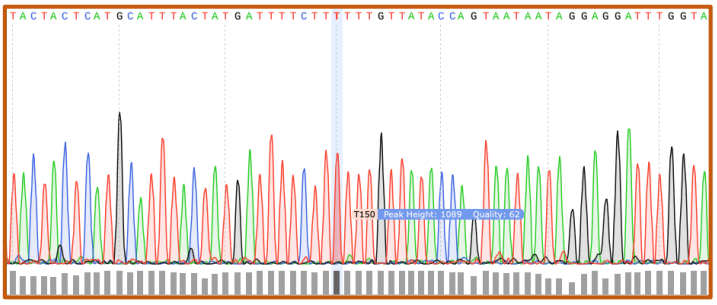
***File Extensions***

A **file extension** typically comes after the period in a file name and indicates the format and/or software program that generated the file. Below are a few common bioinformatics file extensions:

- **.ab1** (ABI sequencer data file): Known as the *trace file*, it includes raw data that has been output from Applied Biosystems' Sequencing Analysis Software. **.ab1** files include quality information about the base calls, the chromatogram (also called the electropherogram), and the DNA sequence.
- **.scf** (Standard Chromatogram Format): Like .**ab1** files, .**scf** files are also *trace files* that include quality information about the base calls, the chromatogram (also called the electropherogram), and the DNA sequence.
- **.fasta:** A text-based format for representing either nucleotide or amino acid sequences. The file often starts with a description or header line that begins with '>' and provides information about the sequence.

***Chromatograms***

Sanger sequencing (.ab1) files are visualized as **chromatograms**. Each peak corresponds with a unique base call, or nucleotide, and quality score. Chromatograms can be trimmed and modified to generate a text-based FASTA file.



***Quality Scores***

**Quality scores** indicate the probability that an individual base, or nucleotide, is called incorrectly during DNA sequencing. For this lab, we recommend a Q score ≥ 40.

| Q score | Probability of incorrect base call | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |

# Technical Overview: FASTA Files

***FASTA Format***

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

1. This file contains DNA sequence information for the red imported fire ant. The first line contains '>' followed by the name of the sequence. A line break separates the name from the actual DNA sequence.

```
>Red imported fire ant
AATTTTATATTTTATTCTTGCTATTTGATCAGGAATAATTGGATCCTCTATAAGAATGATTATTCGATTAGAAC
TAGGATCATGTAATTCACTTATCAATAATGATCAAATTTATAACACTTTAGTAACTAGTCATGCTTTTATTATA
ATCTTTTTTATAGTAATACCTTTTATAATTGGAGGGTTTGGCAACTTCTTAGTTCCTTTAATACTAGGTTCACC
AGATATAGCTTACCCCCGTATAAATAATATAAGATTTTGACTTTTACCCCCATCTCTCTCTCTTCTTATCATAA
```

2. This file contains three separate DNA sequences. The '>' symbol designates the beginning of a new sequence within a FASTA file. Names, or descriptors, are located on the top line, followed by a line break and the corresponding DNA sequence.

```
>Black widow spider
AACTTTGTATTTGATTTTTGGGGCTTGGGCTGCTATAGTGGGTACAGCTATAAGAGTTTTAATTCGTACTGA
ATTAGGGCAACCAGGAAGTTTATTGGGTGATGACCAATTATATAATGTTATTGTGACTGGTCATGCTTTTGT
TATAATTTTTTTTATGGTTATACCGATTTTAATTGGGGGATTTGGAAATTGGTTAGTTCCTTTGATATTAGGA
>German cockroach
AACTCTCTATTTTATTTTTGGAGCTTGATCTGGAATAGTAGGGACATCCTTAAGAATATTAATTCGAGCTGAA
TTAAATCAACCCGGCTCATTAATTGGAGATGATCAAATTTATAATGTTATTGTAACAGCACATGCCTTTGTAA
TAATTTTCTTTATAGTTATACCAATTTTAATTGGGGGATTCGGAAATTGGTTAGTACCTTTAATATTAGGAGC
>Stink bug
AACCTTATACTTCCTATTTGGAATATGAGCAGGAATAGTAGGATCAGCAATAAGATTAATTATCCGAATTGA
ATTAGGACAACCAGGAAGATTTATTGGAGATGATCAGATCTATAACGTAGTAGTTACAGCCCACGCATTCA
TTATAATTTTTTTTATAGTTATGCCTATTATAATTGGGGGATTTGGTAACTGACTTGTACCTTTAATAATTGGA
```

***FASTA File Name***

Just as PDF documents are identified with a .pdf file extension, FASTA files use .fasta or .fa 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 Hawaii
AGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGA
GTTATATTGTAGCCTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTA
CCTAGTAGTACGGAATAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAA

**FASTA 2**

>A_wHa Wolbachia endosymbiont of Drosophila simulans AGTTCTGGTCCATGATGACCC
AGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGA
GTTATATTGTAGCCTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTA
CCTAGTAGTACGGAATAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAA

**FASTA 3**

A_wHa
AGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGA
GTTATATTGTAGCCTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTA
CCTAGTAGTACGGAATAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAA

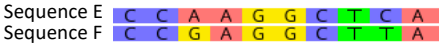
**FASTA 4**

>A_wHa
AGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGA
GTTATATTGTAGCCTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTA
CCTAGTAGTACGGAATAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAA
>A_wMel
AGAGTTTGATCCTAGCTCAGAATGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGA
GTTATATTGTAGCTTGCTATGGTATAACTTAGTGGCAGACGGGTGAGTAATGTATAGGAATCTA
CCTAGTAGTACGGAATAATTGTTGGAAACGGCAACTAATACCGTATACGCCCTACGGGGGAAA

2. For each short sequence alignment below, estimate the p-distance.


Alignment #1 = _____


Alignment #2 = _____

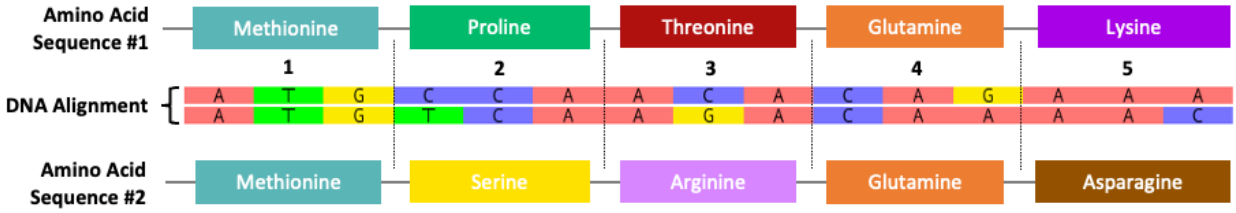



Alignment #3 = _____

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

**Use the following alignment to answer questions 5-8. Codons are labeled #1-5 and illustrated with their corresponding amino acids. Refer to Figures 5.5 and 5.6 for assistance.**




5. Calculate the p-distance (*p*) for the above nucleotide alignment.

$$p = n_d/n$$

($n_d$ = number of nucleotide differences; $n$ = sequence length)

$p$ = _____

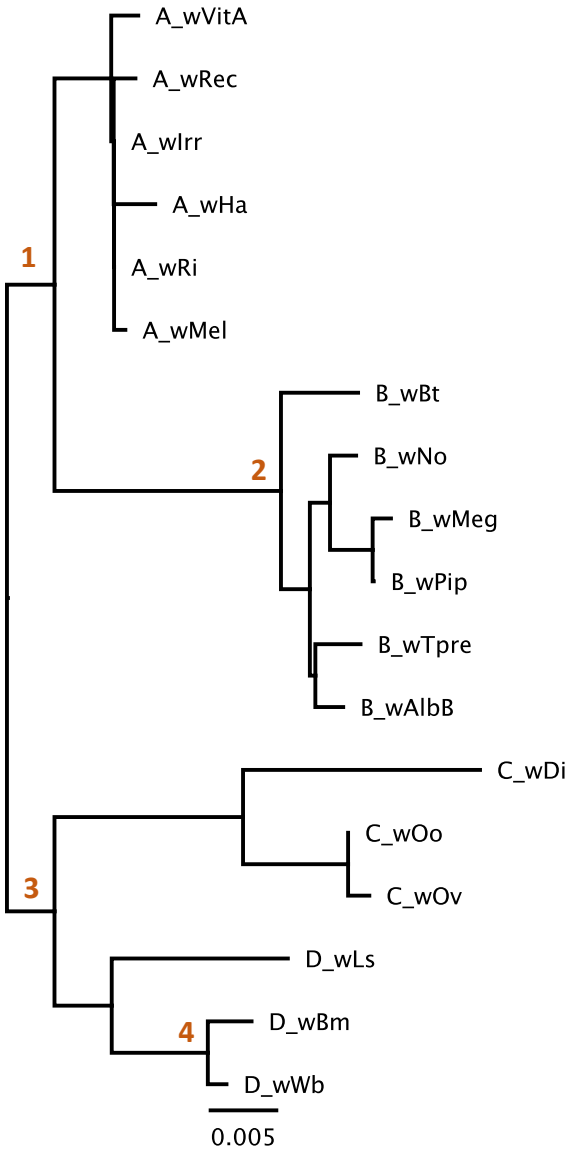
6. Fill in the chart to describe the type of substitution illustrated for each codon. Write N/A if the codons are identical.

| | Transition or Transversion? | Synonymous or Nonsynonymous? |
|---|---|---|
| Codon #1 | | |
| Codon #2 | | |
| Codon #3 | | |
| Codon #4 | | |
| Codon #5 | | |

7. Fill in each blank with either **PURINE** or **PYRIMIDINE**.
   a. Adenine is a _____.
   b. Cytosine is a _____.
   c. Thymine is a _____.
   d. Guanine is a _____.
   e. A _____ has a one-ring structure.
   f. A _____ has a two-ring structure.

8. Which type of substitution – transition or transversion - involves the exchange of one-ring and two-ring structures? _____

**Use the following phylogenetic tree to answer questions 9-11. Taxa are listed as "Supergroup_strain name". For example, A_wVitA is Supergroup A and strain wVitA.**




9. Label the four major *Wolbachia* clades (A, B, C, D).

10. According to the tree, B_*w*Pip and B_*w*Meg are most closely related to which other *Wolbachia* strain?

11. Which node (labeled in orange) represents the common ancestor of:

    a. B_*w*Bt and B_*w*No: _____
    b. Clade A and Clade B: _____
    c. D_*w*Bm and D_*w*Wb: _____
    d. C_*w*Ov and D_*w*Ls: _____

Name: _____
Date: _____

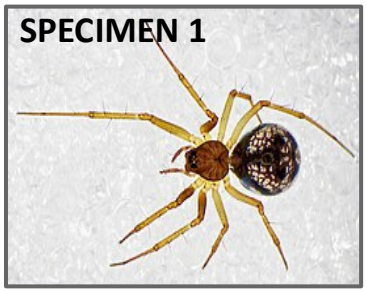
# Introduction to Bioinformatics: Data Sheet

Complete the following table for Lab Activities 1-3.

| Label | Morphology-based ID | BLAST – Taxonomic ID | Common Name |
|---|---|---|---|
| SPECIMEN_1 | | | |
| SPECIMEN_2 | | | |
| SPECIMEN_3 | | | |
| SPECIMEN_4 | | | |
| SPECIMEN_5 | | | |
| SPECIMEN_6 | | | |
| SPECIMEN_7 | | | |
| SPECIMEN_8 | | | |
| SPECIMEN_9 | | | |
| SPECIMEN_10 | | | |
| SPECIMEN_11 | | | |

# Activity 1: Arthropod Identification

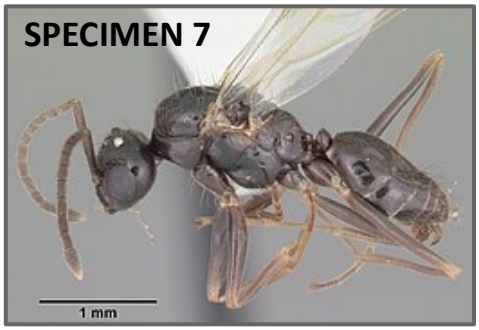
Eleven arthropods were collected from Central Pennsylvania for the *Wolbachia* Project. Based on the images below, putatively identify each arthropod to the best of your ability. You may use a field guide such as https://bugguide.net to identify to taxonomic order. Record answers on the Introduction to Bioinformatics Data Sheet.



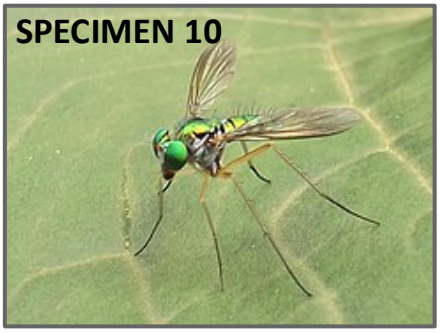











# Activity 2: Sanger Sequence Analysis

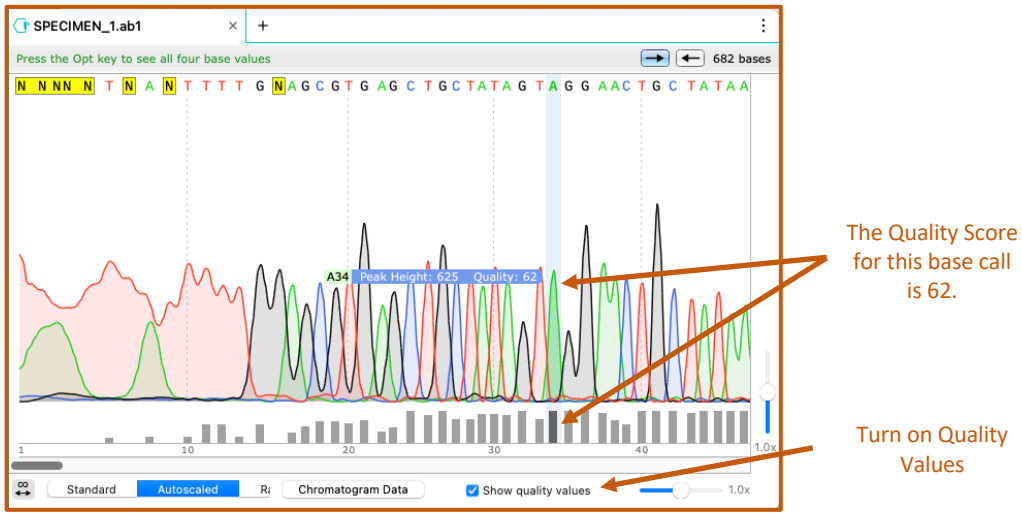
***Getting Started***

1. **Download Example DNA Chromatograms to your desktop.**
   https://wolbachiaproject.org/bioinformatics/
2. **Download SnapGene Viewer to your computer.**
   https://www.snapgene.com/snapgene-viewer/
3. **Open SnapGene Viewer.**

***Edit a Trace File***

4. **Select Open >> Open Files >> SPECIMEN_1.ab1**

5. **Select "Show quality values" in the lower right-hand corner.**
   The gray bars correlate to quality score. Hover the cursor over each bar to visualize the quality score.




6. **Use the bottom scroll bar to scan the sequence.**
   Confirm that most of the sequence contains unique peaks with quality values ≥ 40.

7. **Trim/delete poor base calls. Beginning at the 5'-end (left), identify the beginning of the "high quality sequence."**
   The ends of the sequence will likely be low-quality (as seen above). Therefore, it is necessary to remove poor base calls. Determining where to trim based on the chromatogram and quality scores requires some personal judgement. For example, scroll to base 39. It has a distinct 'C' peak, but the Quality Score is only 36. According to our ≥ 40 cutoff, there are a few possible options:
   - **OPTION 1:** Trim everything before base 40. You may do this by highlighting bases 1-39 along the top and hit "Delete" on the keyboard. Select "OK" from the popup menu.
   - **OPTION 2:** Include this base call but change the 'C' to 'N'. You may do this by highlighting the base at the top of the screen and entering "N" on the keyboard. Select "Insert" from the popup menu.

For the sake of this example, we will take the conservative approach and select **Option 1**. Most importantly, maintain consistency throughout your analysis. Define guidelines, record them in your lab notebook, and apply the same guidelines to all sequences.

8. **Using the cursor, highlight ALL bases from 1-39.**

9. **Hit 'Delete.'**
   ONLY trim from the ends; NEVER trim the interior portion of the sequence!

10. **Repeat for the 3'-end (right).**
    Applying the most conservative guideline, we will trim the last 66 bases. This results in a DNA sequence that contains 577 bases.

11. **Scroll through the sequence. Are all quality scores ≥ 40?**
    If there is a low-value base call in the middle of the sequence, DO NOT DELETE. Use your judgment here. Does the peak look unique? Is the value near 40 (i.e., 37-39)? If yes, you can leave as is. If you are not confident with this base call, highlight with your cursor and type 'N'. This will replace the base call with 'N', indicating that the exact base is unknown.

12. **Select File >> Export >> FASTA Format.**

13. **Check your desktop (or designated folder). You should now have 2 files for this sequence: the original trace file (.ab1) and the sequence file (.fasta; also written as .fa).**
    Note: If you want to save the modified chromatogram file, use the SCF format.

14. **Repeat for all remaining example sequences.**

**Special Note for SPECIMEN_6**

As you scroll through the DNA sequence, notice that many base calls are < 40 and feature a smaller, secondary peak. There are many possible explanations, including:

- NUMTS: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8625877/
- Contamination during the DNA extraction process
- Presence of a secondary arthropod, such as a mite
- Low-quality sequencing run

For the purpose of this activity, we will trim the low-quality ends and make a note of the low QS scores in our notebook. If this happened during an experiment, the researcher could troubleshoot by repeating the Sanger sequencing reaction, analyzing DNA extracted from additional individuals in the population, and/or identifying a more specific primer set.

# Activity 3: Determine Arthropod Identity with BLAST

1. **Download Phylogeny.fasta to your desktop.**
   https://wolbachiaproject.org/bioinformatics/

2. **Right click on Phylogeny.fasta and open the file with a Text Editor.**
   This file contains ALL taxa that will be used to generate the phylogenetic tree, including the 11 collected specimens as well representatives from common arthropod taxa.

3. **Locate the first eleven sequences "SPECIMEN_1 to SPECIMEN_11" at the top of the file. Highlight the DNA sequences with your cursor and right-click >> Copy.**

4. **Open a browser window and navigate to NCBI BLAST.**
   https://blast.ncbi.nlm.nih.gov/Blast.cgi

5. **Click on "Nucleotide BLAST".**

6. **Paste your sequences into the top box marked "Enter Query Sequence".**

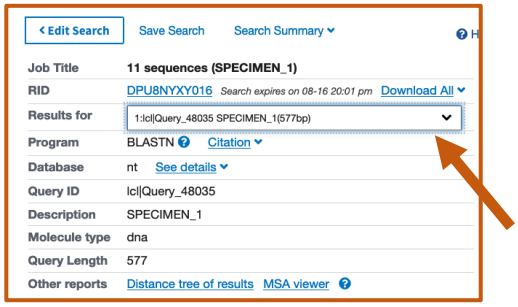
7. **Scroll to the bottom of the page and select the blue BLAST button.**
   The National Center for Biotechnology Information (NCBI) will reference the uploaded sequences against the entire database. This may take a few seconds to a few minutes depending on the level of usage. Each DNA sequence will serve as a unique query and NCBI will deliver eleven separate results.

8. **Scroll through the list of results. The most relevant columns are Query Coverage (the % length that is shared between the query and reference sequences) and Percent Identity (the % nucleotides that are shared between the query and reference sequences).**
   If you want to analyze the sequences, click on the description name. This will link to an alignment of your query sequence with the reference sequence. Click the "Descriptions" tab to return to the list of results.

9. **Based on the BLAST result, record the most likely taxonomic identity (genus and species, if applicable) of SPECIMEN_1 on your sheet.**

10. **In the top-left panel, select the "Results for" drop-down menu to select results for SPECIMEN_2. Record results.**

11. **Repeat Steps 8-10 for the remaining DNA sequences.**



# Activity 4:
# Build a Phylogenetic Tree with NGPhylogeny.fr

1. **Download Phylogeny.fasta to your desktop.**
   https://wolbachiaproject.org/bioinformatics/

2. **Right click on Phylogeny.fasta and open the file with a Text Editor.**

3. **Replace the SPECIMEN_1 through SPECIMEN_11 labels with customized names.**
   Use the Introduction to Bioinformatics Data Sheet as a guide. These can be the common names, scientific names, or your own labels. Make sure that (i) each name begins with a ">" and (ii) there is a line break between the name and the sequence.

4. **Save your changes and close the file.**

5. **Open a browser window and navigate to the online phylogenetics site.**
   https://ngphylogeny.fr/

6. **Click on the green button: Let's Go! With One Click Workflow.**

7. **Upload Phylogeny.fasta as the "Input File".**
   You may also copy the sequences and insert them into the "Pasted text" box.

8. **Click on the blue "Submit" button.**
   The program should complete in less than 10 minutes. If the pipeline stalls at a single step for longer than a minute, there may be an issue with the server. First try to refresh the webpage. If the program still doesn't progress to the next step, Appendix B lists alternative web-based options for building a phylogenetic tree.

9. **Once the program finishes running, select the green Tree Viewer button (Output Tree at step #12).**

10. **Click on Tardigrade and select "Reroot at this node". This will be the outgroup.**

11. ***Optional:* Use the visualization tools on the left to modify the tree.**

# Post-Lab Questions

1. Refer to the Introduction to Bioinformatics Data Sheet.
    a. Do the morphology-based IDs match the NCBI BLAST results? _____

    b. In your opinion, which is more reliable and why?

2. Based on this analysis, is SPECIMEN_9 most likely a bee (order Hymenoptera) or a fly (order Diptera)? _____

3. Copy/paste the phylogenetic tree from Activity 4 into a blank document such as Microsoft PowerPoint or Google Slides.
    a. Use an online browser to research the taxonomic order of each arthropod. Label the corresponding order next to each taxon. For example, yellow fever mosquito would be labeled Diptera.

    b. Do the clades roughly correlate with taxonomic order? Explain.

4. Imagine that you are tasked with studying mosquito-borne diseases in your region. You first want to generate a list of DNA-sequenced mosquitoes.
    a. Which taxonomic level would be the best option for searching mosquitoes in NCBI?
    _____

        - **Kingdom** (Animalia)
        - **Phylum** (Arthropoda)
        - **Family** (Culicidae)
        - **Species** (*Aedes aegypti)*

    b. Open a browser window and navigate to NCBI Taxonomy: https://www.ncbi.nlm.nih.gov/taxonomy. Based on the above answer, enter a taxonomic name that encompasses all mosquitoes and click the Search button. Hover your mouse over each name to reveal taxonomic rank. From the list, record at least five mosquito genera:

        (1) _____

        (2) _____

        (3) _____

        (4) _____

        (5) _____

# Word Search

Name: _____

Created with TheTeachersCorner.net Word Search Maker

## Introduction to Bioinformatics

L V D T X Y N R J F A M I N O A C I D H F Z S X C N V N R O
C D U N R C X V M V U M L B C H U M P C O Y F D A O E F I T
T S L E G F L T M A T R N O X A T Z O U N P P Q A I D F R P
N K A G H I J Q R K O T H U T B E L I O L L A B Y T O A C X
T Z J R N C P Q X A F W S I J R R P N G Y B U D A I N M Q E
O E W E D Y V R N G N T B X N A O Y T K R Y G U E S E Q I T
H H T V F C L Y O Y B S O D P O M Y M L T Q X N C N H H G V
H E N I R U P Z N D O K V A A O I R U W P P D R L A U X I V
E Q Q D G Q T O S I Z E B E U O B T T W M D I O N R W X Y L
P U O R G T U O Y Y M M I S R E S W A R B P G O S T Y B U P
B Y W P F R O D N V A S K N J S X C T L T F I T S H I Q V E
O V W G E N O D O C R N I Q D E I W I I S R Z E S O R E T C
S S Y N R X B L N L G W L I I S M O O T L N R D I E A C I N
S K K R P G K X Y A O D W Z Q E L N N F E U A N A D E L C A
I O H C Y I W F M Q T Z V W C V A T M F L N F R E I S O M T
E S J S R I G R O E A P L S D X Y S F R A O E T T T B F N S
D S I P I M T R U Y M B Y X R T D M S I R S O G Q O T W G I
A F A Q M Z Y F S N O I T A T U M X M M H O T Q O E P V B D
L R B K I X V A N A R N Q S B Y S Z A N R S T A N L T J D P
C H F K D N D Z S R H I L P J J M T M B R J E N U C Y N T M
S T F B I H L U D Q C Z O L D X I I Y Z L K P M K U U H V J
Q J G E N E T I C C O D E P Y C U C Y Y Q D X M A N X Q P F
K G Q B E H Q U W M W D Y P S E N H U I A C P U H R O R G Q
M O C D O V D D D E Q X B S P F L Q F Q H F L M E V F M P A
Y I N D E L A G X Y J R J E G C Y M O J I D Z E V O B Y Z H

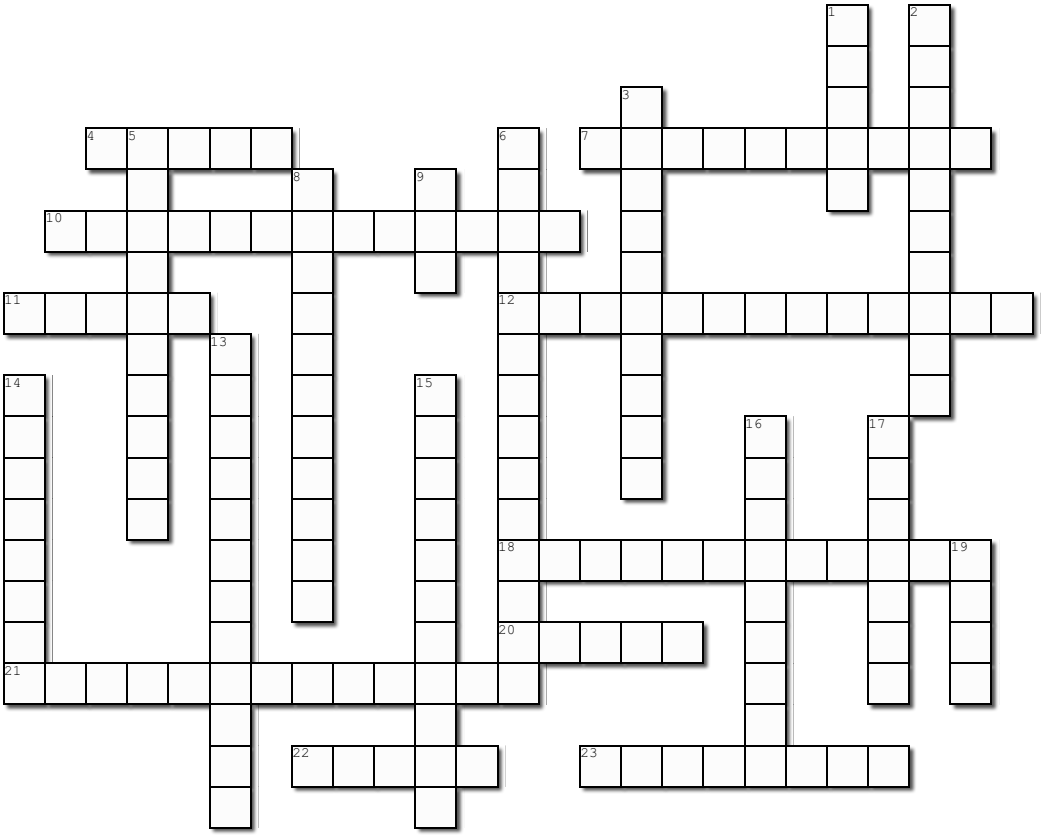
AMINO ACID
BIOINFORMATICS
CHROMATOGRAM
CLADE
CODON
DIVERGENT
FASTA
FRAMESHIFT
GENETIC CODE
INDEL
MUTATION
NONSYNONYMOUS
NODE
NUCLEOTIDE
OTU
OUTGROUP
P-DISTANCE
PHYLOGENETICS
POINT MUTATION
PURINE
PYRIMIDINE
ROOTED
SYNONYMOUS
TAXON
TRANSCRIPTION
TRANSLATION
TRANSVERSION
TRANSITION
UNROOTED

# Crossword Puzzle

Name:_____

## Introduction to Bioinformatics

Complete the crossword puzzle below.



Created using the Crossword Maker on TheTeachersCorner.net

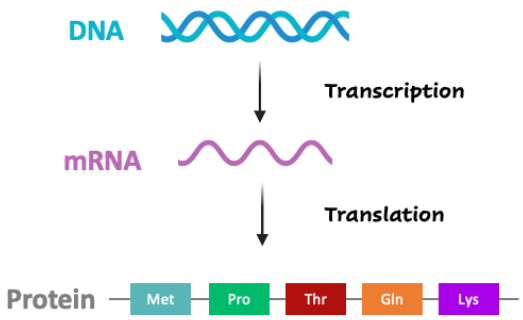
**Across**

**4.** A scientifically classified taxonomic unit, such as a genus or species.
**7.** The nucleotide interchange of a purine with a purine (A, G) or a pyrimidine with a pyrimidine (C, T).
**10.** Genetic mutation in which a single nucleotide in a genome is altered.
**11.** A sequence of three nucleotides that correlates to a specific amino acid.
**12.** A nucleotide substitution that results in a different amino acid product.
**18.** The nucleotide interchange of a purine with a pyrimidine or vice versa.
**20.** Cluster of taxa that originate from the same ancestral node.
**21.** The study of evolutionary relatedness among biological organisms.
**22.** An insertion or deletion in a nucleotide sequence.
**23.** A change in the DNA sequence of an organism.

**Down**

**1.** Text-based format for representing either nucleotide (DNA/RNA) or amino acid (protein) sequences.
**2.** A nucleotide substitution that results in the same amino acid product.
**3.** A genetic mutation in which an insertion or deletion in DNA shifts the coding region.
**5.** The building blocks of proteins.
**6.** The use of computational tools to understand biological data.
**8.** The building blocks of RNA and DNA.
**9.** The basic unit used in phylogenetics to compare closely related taxonomic groups or sequences.
**13.** The visual output of Sanger sequencing.
**14.** A taxon that is known to be more distantly related than all other taxa in the tree.
**15.** Cytosine (C) and thymine (T) nitrogenous bases.
**16.** Tending to be different.
**17.** Adenine (A) and guanine (G) nitrogenous bases.
**19.** The branching point of a phylogenetic tree that represents the common ancestor.

# Appendix A: Codon Table




The genetic code of DNA is transcribed into messenger RNA (mRNA) and then translated into proteins via codons, three consecutive nucleotides that correlate to specific amino acids. The table below summarizes the association between codons (from the coding strand of DNA) and amino acids.

1. ATG (methionine) is the most common start codon; therefore, methionine is often the first amino acid in a protein sequence.
2. TAA, TAG, and TGA encode stop codons. This terminates translation of the protein. If a frameshift mutation prematurely encodes a stop codon, the protein will be truncated.
3. Many codon tables display the genetic code from mRNA to amino acid (translation). In this case, thymine (T) would be replaced with uracil (U).
4. The genetic code is *nearly* universal across all forms of life. Some organisms, such as bacteria, feature slight deviations in the genetic code and may utilize alternative start codons.

**Second Base**

| First Base | T | C | A | G | Third Base |
|---|---|---|---|---|---|
| T | TTT, TTC: Phenylalanine (Phe)<br>TTA, TTG: Leucine (Leu) | TCT, TCC, TCA, TCG: Serine (Ser) | TAT, TAC: Tyrosine (Tyr)<br>**TAA: Stop**<br>**TAG: Stop** | TGT, TGC: Cysteine (Cys)<br>**TGA: Stop**<br>TGG: Tryptophan (Trp) | T C A G |
| C | CTT, CTC, CTA, CTG: Leucine (Leu) | CCT, CCC, CCA, CCG: Proline (Pro) | CAT, CAC: Histidine (His)<br>CAA, CAG: Glutamine (Gln) | CGT, CGC, CGA, CGG: Arginine (Arg) | T C A G |
| A | ATT, ATC, ATA: Isoleucine (Ile)<br>**ATG: Methionine (Met)** | ACT, ACC, ACA, ACG: Threonine (Thr) | AAT, AAC: Asparagine (Asn)<br>AAA, AAG: Lysine (Lys) | AGT, AGC: Serine (Ser)<br>AGA, AGG: Arginine (Arg) | T C A G |
| G | GTT, GTC, GTA, GTG: Valine (Val) | GCT, GCC, GCA, GCG: Alanine (Ala) | GAT, GAC: Aspartic acid (Asp)<br>GAA, GAG: Glutamic acid (Glu) | GGT, GGC, GGA, GGG: Glycine (Gly) | T C A G |

# Appendix B.1
# Build a Phylogenetic Tree with Phylogeny.fr

http://www.phylogeny.fr/

### Build the Phylogenetic Tree

1. **Download Phylogeny.fasta to your desktop.**
   https://wolbachiaproject.org/bioinformatics/

2. **Open a browser window and navigate to the online phylogenetics site.**
   http://www.phylogeny.fr/

3. **From the top navigation, click "Phylogeny Analysis" >> "One Click".**

4. **Click Browse… and upload the Phylogeny.fasta document.**
   You may also copy the sequences and insert them into the "Pasted text" box.

5. **Click on the blue "Submit" button.**
   Use default settings. *Optional*: Enter your e-mail address to receive results by email.

### Assign the Outgroup

6. **Once the program finishes running, scroll down to the customizable options below the tree. Select the "Reroot (outgroup)" button.**

7. **Scroll back up the tree and click on Tardigrade. This will be the outgroup.**

### Change the SPECIMEN Names

8. **Scroll down to the customizable options below the tree. Select the "Change leaf name" button.**

9. **Scroll back up to the tree and click on SPECIMEN_1. Enter the new leaf name.**
   Use the Introduction to Bioinformatics Data Sheet as a guide. These can be the common names, scientific names, or your own labels.

10. **Repeat Steps 8 & 9 for remaining SPECIMEN labels. You may also adjust "Tardigrade" to "Water bear" for consistency (all names on the tree are common names).**

### Change Visualization Parameters and Download the Tree.

11. ***Optional:* Use the visualization tools to modify the tree.**
    Click on the action box, then return to the tree and select the taxon/node to perform the action. For example, the "Add leaf annotations" button can be used to add an annotation, or note, about each taxon. Annotations will appear to the right of the tree.

12. **Use the menu directly under the tree to download an image file.**

# Appendix B.2
# Build a Phylogenetic Tree with MAFFT & Phylo.io

https://mafft.cbrc.jp/alignment/server/

The MAFFT alignment program is highly customizable. Each parameter can be adjusted for specific needs. Phylo.io is an integrated, simple, and easy to use tree interface.

## Prepare the FASTA File

1. **Download Phylogeny.fasta to your desktop.**
   https://wolbachiaproject.org/bioinformatics/

2. **Right click on Phylogeny.fasta and open the file with a Text Editor.**

3. **Replace the SPECIMEN_1 through SPECIMEN_11 labels with customized names.**
   Use the Introduction to Bioinformatics Data Sheet as a guide. These can be the common names, scientific names, or your own labels. Make sure that (i) each name begins with a ">" and (ii) there is a line break between the name and the sequence.

4. **Save your changes and close the file.**

## Align the Sequences

5. **Open a browser window and navigate to the online MAFFT site.**
   https://mafft.cbrc.jp/alignment/server/

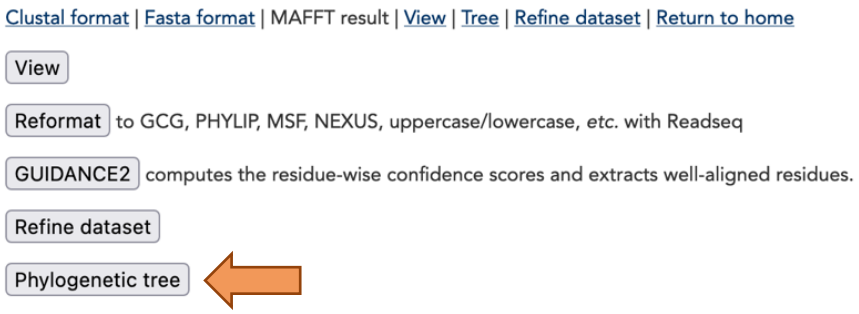
6. **From the top Input box, click "Browse…" and upload the Phylogeny.fasta document.**
   You may also copy the sequences and insert them into the white text box.

7. **In the UPPERCASE/lowercase box, select the "same as input" radio button.**
   Leave all other default settings the same. *Optional*: Enter a job name and your e-mail address to receive results by email.

8. **Click Submit.**

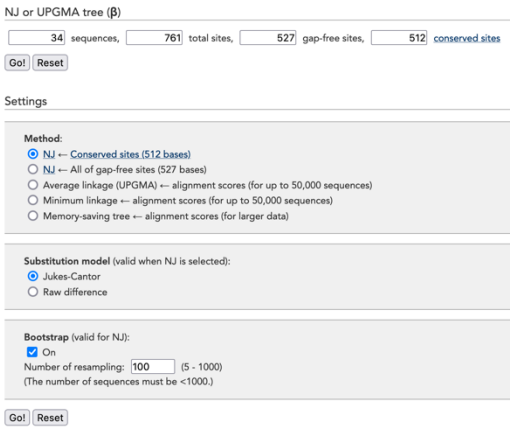
## Generate the Tree File

9. **Once the program finishes running and redirects to a new page, click on Phylogenetic tree.**



10. **Under the Settings panel, use the following parameters and click Go!**

    **Method:** NJ ← Conserved Sites
    **Substitution model:** Jukes-Cantor
    **Bootstrap:** On
    **Number of resampling:** 100

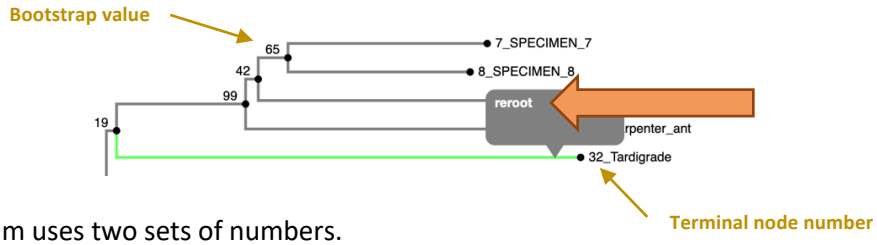



A bootstrap value indicates the number of times the same branch is observed when repeating the generation of the phylogenetic tree from the same dataset. Therefore, a resampling value of 100 means that the program generates 100 possible trees based on the same MAFFT alignment. The bootstrap value indicates how often the same branch is observed across the 100 trees. A bootstrap value of 100 indicates that it was always observed. A bootstrap value of 65 indicates that the branch was observed 65/100 times.

## Visualize the Phylogenetic Tree

11. **Select "View tree on Phylo.io" from the results page. The tree will appear in a pop-up window.**

12. **Click on the Tardigrade branch. When it is highlighted in green, click "reroot" from the popup.**




This program uses two sets of numbers.

The number at each leaf/taxon is the "terminal node number. It is a unique identifier for each taxon and correlates with the order that the sequence was listed in the FASTA file. The numbers located at each node are the bootstrap values.

## Download the Phylogenetic Tree

13. **From the left-hand menu, select the Share button. Copy the URL for future reference.**

    Alternatively, you may take a screenshot of the final tree.

# Appendix B.3
# Build a Phylogenetic Tree with MAFFT & Interactive Tree of Life (iTol)

MAFFT - https://mafft.cbrc.jp/alignment/server/
iTol - https://itol.embl.de/

The MAFFT alignment program is highly customizable. Each parameter can be adjusted for specific needs. The Interactive Tree of Life provides a comprehensive online environment for annotating and managing phylogenetic trees. It creates publication quality trees with an extensive set of customization options.

## Prepare the Alignment

1. **Download Phylogeny.fasta to your desktop.**
   https://wolbachiaproject.org/bioinformatics/

2. **Open a browser window and navigate to the online MAFFT site.**
   https://mafft.cbrc.jp/alignment/server/

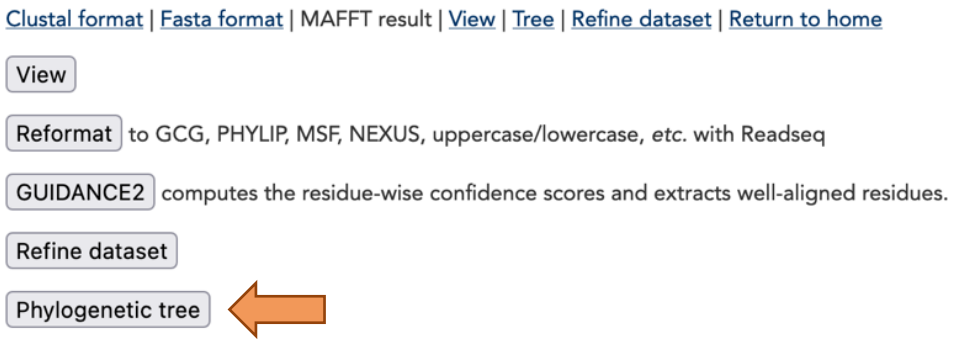
3. **From the top Input box, click "Browse…" and upload the Phylogeny.fasta document.**
   You may also copy the sequences and insert them into the white text box.

4. **In the UPPERCASE/lowercase box, select the "same as input" radio button.**
   Leave all other default settings the same. *Optional*: Enter a job name and your e-mail address to receive results by email.

5. **Click Submit.**

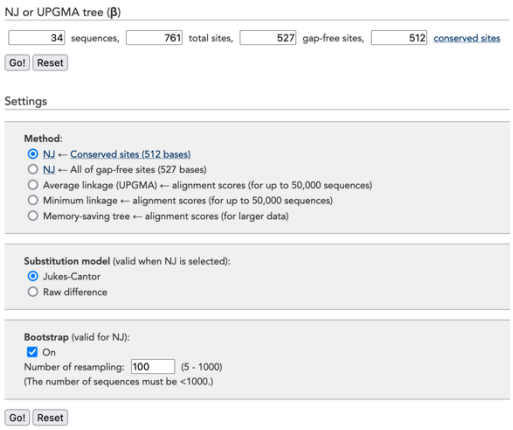
## Generate the Tree File

6. **Once the program finishes running and redirects to a new page, click on Phylogenetic tree.**



7. **Under the Settings panel, use the following parameters and click Go!**

   **Method:** NJ ← Conserved Sites
   **Substitution model:** Jukes-Cantor
   **Bootstrap:** On
   **Number of resampling:** 100




   A bootstrap value indicates the number of times the same branch is observed when repeating the generation of the phylogenetic tree from the same dataset. Therefore, a resampling value of 100 means that the program generates 100 possible trees based on the same MAFFT alignment. The bootstrap value indicates how often the same branch is observed across the 100 trees. A bootstrap value of 100 indicates that it was always observed. A bootstrap value of 65 indicates that the branch was observed 65/100 times.

## Visualize the Phylogenetic Tree

8. **Scroll to the bottom of the Results page. Under the "Result (for external tree viewers)" section, download one of the file types listed under "Tree file without terminal node number". Save to desktop.**

   Any of the file types will work.

9. **Open a second browser window and navigate to the Interactive Tree of Life site.**

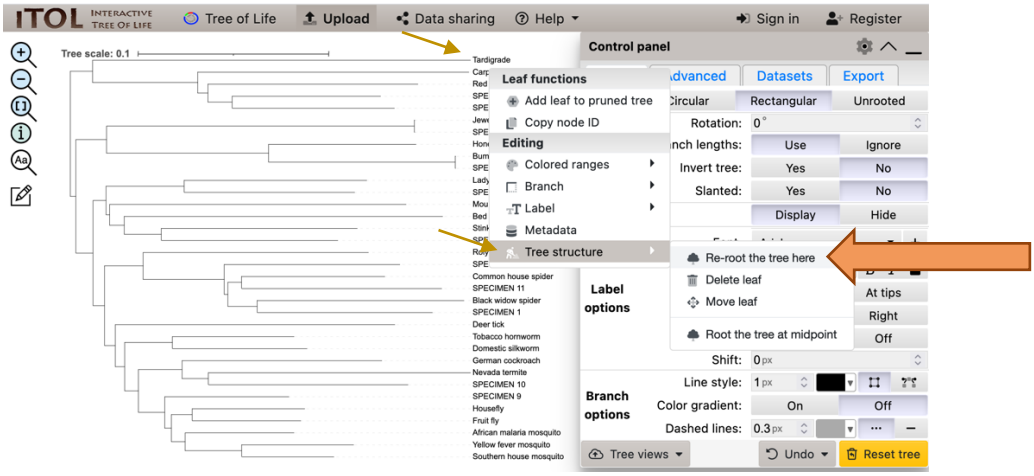
   https://itol.embl.de/

10. **In the top navigation menu, select the Upload button.**

11. **Click "Browse…" and upload the tree file from Step 13.**

    The tree file should be in one of the following formats, as indicated by the associated file extension: **NHX** (.nh); **phb** (.phb); **Newick** (.nwk); or **PhyloXML** (.xml).

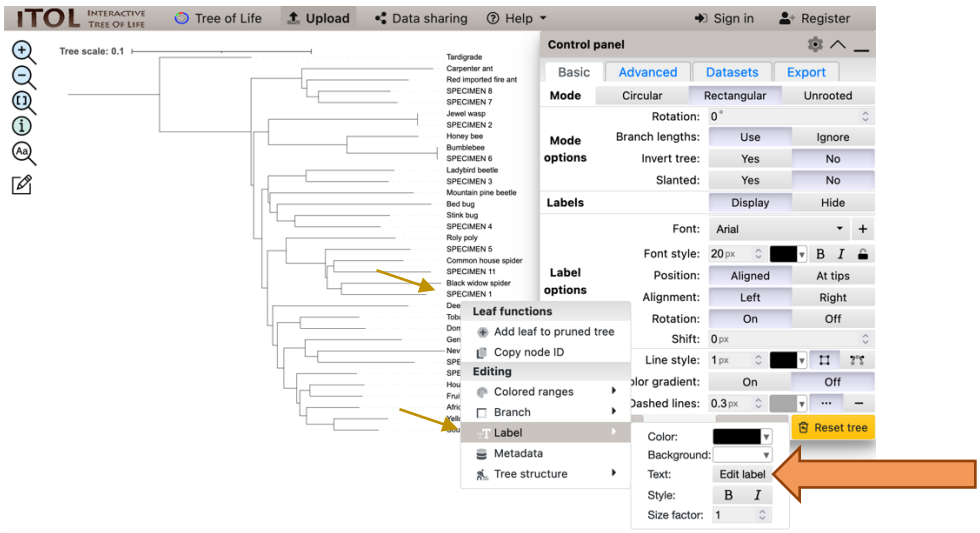
12. **Right-click on the Tardigrade label. Select Tree Structure >> Re-root the tree here.**



13. **To change label names, right click on SPECIMEN_1. Select Label >> Text >> Edit Label. Enter the new name in the text field.**

    Use the Introduction to Bioinformatics Data Sheet as a guide. These can be the common names, scientific names, or your own labels.

    In addition to changing the text, each branch and label can be color-coded. For example, you might color code all taxa from the order Diptera orange, all taxa from the order Coleoptera green, and so on.




14. **Use the Control Panel to adjust other parameters of the tree.**

    The tree can be visualized in three different modes: rectangular (default), circular, or unrooted. The font and font style can be adjusted under Label Options. Branches can be made lighter or darker by adjusting the Line Style.

15. **Use the Export tab in the Control Panel to save the tree image.**

# Glossary

**Amino acid:** An organic compound containing a carboxyl group, an amino group, and a variable side chain. Amino acids are the building blocks of proteins.

**Bioinformatics:** The use of computational tools to understand biological data.

**Chromatogram:** The visual output of Sanger sequencing. It consists of four colors where each peak corresponds with a unique base call, or nucleotide, and quality score.

**Clade**: Cluster of taxa that originate from the same ancestral node.

**Codon**: A sequence of three nucleotides that correlates to a specific amino acid.

**Divergent**: Tending to be different.

**FASTA:** Text-based format for representing either nucleotide (DNA/RNA) or amino acid (protein) sequences.

**File extension:** A suffix that typically comes after the period in a file name and indicates the format and/or software program that generated the file. Examples include .doc, .pdf, .abi, and .fasta.

**Frameshift mutation:** A genetic mutation in which an insertion or deletion in DNA shifts the coding region, resulting in an altered amino acid sequence.

**Genetic code:** A set of instructions, organized as three-nucleotide combinations (codons), that directs the translation of DNA into specific amino acids.

**Indel:** An insertion or deletion in a nucleotide sequence.

**Midpoint rooted tree**: An unrooted phylogenetic tree where the hypothetical root is placed midpoint in the tree.

**Mutation:** A change in the DNA sequence of an organism.

**Nonsynonymous substitution**: A nucleotide substitution that results in a different amino acid product.

**Node:** The branching point of a phylogenetic tree that represents the common ancestor for all descendants (taxa) branching out of that node.

**Nucleotide:** A compound consisting of a five-carbon sugar (either ribose in RNA or deoxyribose in DNA) attached to a phosphate group and nitrogenous base. Nucleotides are the building blocks of RNA and DNA.

**Operational taxonomic unit (OTU)**: The basic unit used in phylogenetics to compare closely related taxonomic groups or sequences. OTUs are found at the tips of phylogenetic trees.

**Outgroup**: A taxon that is known to be more distantly related than all other taxa in the tree.

**p-distance:** Pairwise distance ($p$) between two sequences that is calculated by aligning the two sequences and dividing the number of nucleotide differences ($n_d$) by the sequence length ($n$).

**Phylogenetics**: The study of evolutionary relatedness among biological organisms.

**Point mutations**: Genetic mutations in which a single nucleotide in a genome is altered.

**Purine:** A type of nitrogenous base that makes up DNA. Purine bases are adenine (A) and guanine (G), and feature two-ring structures.

**Pyrimidine:** A type of nitrogenous base that makes up DNA. Pyrimidine bases are cytosine (C) and thymine (T), and feature one-ring structures.

**Quality score:** A numerical value that indicates the probability that an individual base, or nucleotide, is called correctly during DNA sequencing.

**Reading frame:** A sequence of non-overlapping nucleotide triplets in DNA that is translated to amino acids. There are potentially three ways to read a DNA sequence depending on whether the frame starts at the first, second, or third nucleotide.

**Rooted tree:** A phylogenetic tree featuring a distinct node, or root, that serves as the ancestral group for all taxa in the tree.

**Synonymous substitution:** A nucleotide substitution that results in the same amino acid product.
**Taxon**: (plural *taxa*) A scientifically classified taxonomic unit, such as a genus or species.
**Transcription**: The process by which DNA is transcribed, or copied, to make RNA.
**Translation:** The process by which messenger RNA (mRNA) is translated into a specific series of amino acids to produce a protein.
T**ransversion substitution**: The nucleotide interchange of a purine with a pyrimidine or vice versa.
**Transition substitution**: The nucleotide interchange of a purine with a purine (A, G) or a pyrimidine with a pyrimidine (C, T).
**Unrooted tree**: A phylogenetic tree in which ancestry is unknown and the ancestral outgroup, or root, is not identified.