Document 4: Phylogenies- linking the cladograms presented in OneZoom to the genetic data.

Student Learning Outcome:To gain an understanding of genetic data used to generate cladograms as shown in OneZoom. Based on this exercise you should understand how genetic data submitted by thousands of researchers to GenBank, a curated warehouse of genetic sequences, is used by the developers of OneZoom to create the phylogenetic cladograms that are such effective visualizations of complex datasets.

Define the following terms:

Phylogeny

Cladogram

Genetic sequence

Part 1- OneZoom to NCBI: GenBank

* 1. Practicing database use and verification with a zombie snail parasite. Back to OneZoom. In the search all life box, enter zombie snail parasit*e*. The node should show the genus *Leucochloridium*. Zoom into the node and click on it to open the Wikipedia link. Notice, there are other options:



1. Click on Genetics to enter website for National Center for Biological Inventory (on mobile devices Genetics will appear as a double helix icon). You will enter the Taxonomy Browser home page. An example below is given for the amber snails which are part of the parasite’s life cycle. Check the box for nucleotide and select *Leucochloridium* from the list.



1. See the Entrez Records table and make sure only Nucleotide has a checkmark. Nucleotide records correspond to segments of gene regions that can be used to build phylogenetic trees.
2. Next, click on the term *Leucochloridium* in taxonomic list and another window will appear. Under Entrez Records you will see the Database Name. Next to nucleotide, the numbers of genetic sequences found on GenBank are listed here.





Question: How many nucleotide records are given?

1. Click on the number. You will enter the National Library of Medicine, GenBank. The number of items should match the number of nucleotides from the answer above.

Part 2: Using GenBank- the repository of gene sequences to build your phylogeny**.**

Based on the exercise in Part 1. You should be in GenBank- If not, paste this link into your browser. [GenBank Overview (nih.gov)](https://www.ncbi.nlm.nih.gov/genbank/)

One approach to adding new data to phylogenies of specific groups is to use *specific genes*. Genes from the large and small subunits of ribosomes (remember- ribosomes are protein/RNA combinations that create polypeptides needed for proteins) such as 12s rDNA, 18s rDNA, and 28s rDNA are great choices because they are often used in genetic studies.

In the example below I chose another Trematode parasite with a complex life cycle similar to the Zombie Snail Parasite. The example below uses the Asian Lung Fluke, *Paragonimus*, a major human pathogen in Southeast Asia that uses a snail host and a Human. When I typed in *Paragonimus* into the search bar it returned several thousand hits-way too many to download, organize, and create phylogenies. When I typed a specific gene into the search bar it narrowed it to 43 options (see screenshot below) that is a much more manageable dataset.



Back to the Zombie Snail Parasite.

* 1. Choose 28s and *Leucochloridium*.

Question: How many sequences for this gene are in GenBank?

* 1. Choose 18s and *Leucochloridium*.

Question: How many sequences for this gene are in GenBank?

* 1. Choose 12s and *Leucochloridium*.

Question: How many sequences for this gene are in GenBank?

* 1. Next, you will select only 28s sequences from GenBank to download to construct the phylogenetic tree. Make sure you select- sort by sequence length at the top of the screen.



* 1. Choose only sequences less than 1,000 bp in length as this will make creating a phylogenetic tree MUCH EASIER!

Question: How many sequences were longer than 1,000 basepairs (nucleotides)?

* 1. Select all 28s sequences using a checkmark as shown in the example below. Please note the example uses 16s. You are choosing only 28s sequences based on enough options available to build a phylogenetic tree. Do not include other gene regions (e.g. 18s, 16s, ITS) as they are very different than 28s.



Question: How many sequences were selected?

Question: What is the size range of sequences included?

* 1. Next, download your selected sequences by: Selecting **Send to** on the right side of the screen. Change the settings as shown below. In Choose Destination select choose file, in Format- choose FASTA and sort by organism name.



* 1. Click on Create File and GenBank will send the file to the downloads folder on your computer hard drive (On a mobile device, wherever downloads are stored this file should be found).
	2. Use the notepad app on your computer (or mobile device use the open in another app if the file does not immediately open) to view your sequences as your computer views the document as a text file. You should see as string of nucleotides for each record.
	3. *Save this file as your last name\_sequences.*

These sequences will be used to create a phylogenetic tree in an opensource phylogenetic software (MEGA). Given the complexity of this software, this step is not shown in this exercise. Instead, the phylogenetic tree generated in this software appears below. Four sequences were removed from phylogenetic analyses due to incompatible regions. Here, 12 28s sequences were used to generate the phylogenetic tree. It is much more difficult to read than the cladograms generated by OneZoom.



Question: Of the three species listed in the tree above, which two are most closely related?

Part 3: The relationship between GenBank and OneZoom**.**

Now, let’s compare this simple tree to the phylogenetic relationships shown in OneZoom to the MEGA generated tree above. To do this, use the advanced tracer tool option shown in previous exercises. Include the species shown in the tree above. Paste the tree below.

Question: Do the relationships presented in the tree match what is shown in OneZoom?

Question: How do you know?