**Phylogenetic Epidemiology**

You are part of a team working for the Center for Disease Control in Atlanta, GA. There has been an outbreak of a disease in the Northeast. On June 12, 2010, 17 adolescent males at a boy scout camp in Milton, Massachusetts began suffering from headaches, fever, vomiting and drowsiness. Within days, 15 of the boys showed signs of confusion, respiratory distress and tremors. Encephalitis was found in the young men after a week. They began experiencing paralysis and five of the boys slipped into a coma. Two days after the onset of these symptoms, 13 of the boys died. The remaining four boys eventually recovered but still experience neurological damage.

An investigation into the region has revealed that this has happened before but never to the same severity. Cases dating back to May of 2009 show similar symptoms in individuals from Connecticut, Massachusetts, New Jersey, New York, Pennsylvania and Rhode Island.

The virus in question has had its genome sequenced from each of the victims from Milton, MA. You will find the complete sequence from one of the boys, identified as patient 0042, in the file marked “Unknown\_42.fsa” on ELC. The unknown virus is identified at Unknown\_42.

In this exercise, you will trace the infection history of the Unknown\_42 virus. On ELC you will find a file “Unknown\_42\_11cases.fsa.” This file has sequences for 1 individuals that have shown similar symptoms to the Milton, MA patients. There are also a few other viral genomes that will be helpful in understanding the phylogeny of Unknown\_42-like viruses. In each case, it was determined that a virus was responsible for the disease, and its genome was sequenced. Information about these patients can be found in the FASTA file ID line. This information will be helpful in understanding the spread of the disease in the United States.

You are tasked with the following:

1. Determine what the virus is or to what group of viruses it belongs.
2. Hypothesize how transmission of the virus occurs based on the transmission of related virus species.
3. Reconstruct the history of Unknown\_42 in the United States.
4. Identify individuals who have most likely contracted Unknown\_42 virus.
5. Identify patient zero.

THE PHYLOGENETIC ANALYSIS TAKES A WHILE TO COMPLETE. WE WILL BEGIN BY STARTING THIS ANALYSIS SO IT WILL BE READY WHEN WE NEED IT.

**Phylogenetic Reconstruction using Parsimony: dnapars from PHYLIP**

You will be using a web-based version of the program dnapars from PHYLIP 3.67 to conduct a parsimony analysis of individuals that have been symptomatic of Unknown\_42.

1. Go to <http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::dnapars>. Upload the alignment file “Unknown\_42\_11cases\_align.fsa”.
2. We will conduct a bootstrap analysis of the data. Bootstrap analyses give us some indication of reliability of our tree through randomly sampling the data many times and analyzing those samples. To start the bootstrap analysis, change randomize to “No”. Change bootstrap to “Yes”. Keep resampling methods on “Bootstrap” and number of replicates on “100”. Select an odd number to seed the process. Change “Compute a consensus tree” to “Yes”. Submit the job. We will come back to this analysis later.

**Determining Taxonomic Identity of Virus**

**Using BLAST and Genbank**

You will be using two tools that are vital to modern biology: BLAST and GenBank. BLAST is a program that takes your sequence of interest (in this case the viral genome) and uses it to search against a database of known sequences (GenBank). The results of the search are sequences that are very similar to the query sequence.

1. Go to <http://blast.ncbi.nlm.nih.gov/Blast.cgi>.
2. Click on “nucleotide blast”. Either upload the virus genome file (“Unknown\_42.fsa”) or copy and paste it into the box provided. Scroll down to “Database” and click “Others”. The database “Nucleotide Collection” should be selected. Change from “megablast” to “blastn”. Click “BLAST” to start the search.
3. Once the search is complete, answer the following questions:
   1. How do you determine which hits are the “best”?
   2. What are the top five hits for your sequence?
4. Click on “Taxonomy reports”.
   1. To what family of viruses does the unknown sample belong?
   2. What characteristics do members of this virus family share? (Use the Internet. Cite your source.)
   3. For which well-known virus is the family named? What disease does it cause?
   4. What virus species seems to be most closely related to Unknown\_42 given the distance-based tree provided by NCBI?
   5. What can you say about the identity of Unknown\_42?

NOW, WE WILL GO BACK TO THE BOOTSTRAP ANALYSIS YOU STARTED AT THE BEGINNING OF CLASS.

**Phylogenetic Reconstruction using Parsimony: dnapars from PHYLIP**

Today, we are only going to look at the bootstrap consensus tree to answer questions regarding the evolutionary history of Unknown\_42. Bootstrap trees provide a measurement of confidence for our phylogeny. We use a consensus tree of the bootstrap trees to give an idea of how many times a particular relationship is recovered from the resampling of data. This resampling gives us an estimate of confidence in our phylogeny.

1. View the consensus tree using archaeopteryx.
2. Sketch the consensus tree and include bootstrap values.
3. Use BLAST to check the sequences of ID\_39, ID\_91 and ID\_28 to determine if the major clades in the phylogeny are similar to Powassan virus or to another virus. You can find these sequences in the file Unknown\_42\_11cases.fsa. Report your findings.
4. Why were samples ID\_39 and ID\_11 included in the sampling of Unknown\_42-like viruses?
   1. Why do they have similar symptoms to Powassan virus and Unknown\_42-like viruses?

1. Which individuals most likely have contracted Unknown\_42 (You can use their ID number.)?
2. If we assumed that every instance of the Unknown\_42 virus in the United States has been accounted for in this phylogeny, who is most likely patient zero (the first infected)?
   1. How did you come to that conclusion?
3. Based on what you know about Powassan virus, postulate on how Unknown\_42 virus is transmitted.
4. Unknown\_42 virus is much more virulent than Powassan disease virus. Would you consider this a new virus species? Why or why not?