

## **Exercise 4: Determine if the major capsid protein sequence can be used to cluster phages.**

Bacteriophage studies are of interest for a variety of reasons including phage therapy to treat antibiotic-resistant bacterial infections. The diversity of phages is enormous, and relationships between phages are predicted based upon their genomic sequences. Typically phages are clustered based upon complete genome sequences, and similar sequences will group together in a phylogenetic tree (see Figure 4). Clusters of phages that are evolutionarily related will emerge and relationships between different clusters can be predicted. However, sequencing entire genomes is more time intensive and expensive than sequencing a single gene. Analysis of conserved genes may be useful for identifying phage clusters prior to sequencing an entire genome, leading to identification of novel singleton phages that are not yet assigned to a cluster.

Several structural genes are conserved among phages, and you will use the major capsid protein for this analysis as it is conserved amongst viruses in general. [1]

### **Exercise 4:**

In this exercise, you will use your training from Sections 2 and 3 to generate multiple phylogenetic trees based upon the protein sequences of the major capsid proteins in a group of bacteriophages.

**Step 1:** Generate a fasta file with the major capsid protein sequences from the group of bacteriophages listed in Table 1 below. Note that 3 phages from each cluster are included.

Phage	Cluster	Phage	Cluster	Phage	Cluster	Phage	Cluster
Abrogate	A1	Wally	C1	Cerasum	F1	LilPharaoh	K1
BxB1	A1	BadAgartude	C1	Tweety	F1	BarrelRoll	K1
Seabiscuit	A1	Ewok	C1	Dante	F1	Anaya	K1
Cobra	B1	Sassay	E	Arial	J	Breezona	L2
Serendipity	B1	Dusk	E	Optimus	J	Faith1	L2
Orion	B1	Dumbo	E	Thibault	J	Crossroads	L2

**Table 1.** Bacteriophages whose major capsid protein sequence will be used to generate phylogenetic trees.

To generate your fasta file, do the following:

- Go to phagesdb.com and type the phage name in the top left corner to search for the phage.
- On the phage page, scroll down to Gene List and click to view a list of genes and gene products.
- Click on the 'major capsid protein' and it will take you to the gene page. Select amino acid sequence and copy that sequence into your fasta file that you will create in Notepad or the TextEdit program.
- Edit each sequence's descriptive name to identify only the phage and cluster (e.g. ">Abrogate\_A1"). No spaces allowed!
- No hard returns or spaces in each fasta sequence (see formatting example in exercise 2A)
- Save your capsid fasta file.

### **Question Set 6:**

- 4.1 Why include more than one phage per cluster?
- 4.2 Would you expect clearer results if all of phages in a cluster were included?
- 4.3 Since some of these clusters have over 100 phages, is it reasonable to choose a subset of phages to represent the cluster? (for phage students- go to phagesdb and find out what is special about Crossroads)

**Step 2:** Create a tree with phylogeny.fr (phyML program uses maximum likelihood) following the instructions that you used in Exercise 2 and using your capsid protein fasta file from Step 1.

### **Question Set 7:**

- 4.4 Submit your phylogenetic tree onto the worksheet.
- 4.5 How would you annotate this tree?
- 4.6 Are all members of a cluster next to each other in this tree?
- 4.7 Which clusters are predicted to be the closest in evolutionary time (aka. the most similar)?
- 4.8 Which are predicted to be the farthest apart in evolutionary time?
- 4.9 Which branch is the least supported based on bootstrapping results?

**Step 3:** Create a phylogenetic network with Splitstree4 using neighbor-joining and bootstrapping following the instructions in exercise 3 and using your capsid protein fasta file.

### **Question Set 8:**

- 4.10 Submit your phylogenetic network onto the worksheet.
- 4.11 How would you annotate this network?
- 4.12 Are all members of a cluster next to each other in this network?
- 4.13 Which are predicted to be the farthest apart in evolutionary time?
- 4.14 Describe similarities (if any) between the phylogeny.fr tree and the Splitstree network.
- 4.15 Describe differences (if any) between the phylogeny.fr tree and the Splitstree network..
- 4.16 What conclusions can be drawn from these analyzes using the capsid protein sequences?

### References:

1. Krupovic M, Koonin EV. 2017. Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A 114:E2401–E2410. doi:10.1073/pnas.1621061114.