Phylogenetic Tree Practical Problems

Software Tools:

- MEGA – A software package for constructing phylogenetic trees using neighbor-joining, UPGMA, and maximum parsimony.
- ClustalW – A tool for constructing multiple sequence alignment. ClustalW alignment is integrated into MEGA.

Databases:

- GenBank – A database used for looking up DNA and protein sequences of organisms. GenBank can be queried directly from the MEGA tool.
- Ebola Genome Browser - An information hub about the Ebola virus and interactive viewer of the Ebola genome.
- NCBI Ebola Virus Variation – A database containing various DNA and protein sequences for the Ebolavirus obtained from different host species, different countries, and different dates.

Background: The 2014 Ebola outbreak was the largest Ebola epidemic in history, infecting over 26,000 people in West Africa and taken over 10,000 lives.

Fortunately, Ebola is not an airborne disease; it can only be spread through direct contact with body fluids of an infected individual. An infected individual becomes contagious after they begin to show symptoms of the disease, which can occur 2-21 days after exposure. Scientists are also investigating the possibility that the virus may be transmitted sexually in the semen of Ebola survivors. The Ebola virus has been found in semen 89 days after symptom onset, and RNA from the virus can be found up to 199 days after symptom onset, which is long after the virus can no longer be detected in the bloodstream.

Months of investigation resulted in the identification of patient zero for the 2014 outbreak as a 2-year old boy named Emile Ouamouno from Méliandou, Guinea. But how was he infected?

Objective 1: Constructing a Multiple Alignment of Ebola Genome Sequences

The 2014 Ebola epidemic was the most deadly one in history, but smaller outbreaks have occurred in sub-Saharan Africa at different times since 1976. In fact, there are five different species in the *Ebolavirus* genus: Zaire (EBOV), Sudan (SUDV), Bundibugyo (BDBV), Tai Forest (TAFV), and Reston (RESTV). The first four of these species cause disease in humans. Our first biological objective is to place the strain causing the 2014 outbreak within the *Ebolavirus* phylogeny.

The following map shows the locations for which the Ebolavirus species are named in addition to the origin of the 2014 outbreak. The pins are, in order of appearance from left to right, Guinea, Tai Forest, Zaire, Bundibugyo, and Sudan.

1. Based on the locations on the map above, which species of Ebolavirus do you think the 2014 outbreak in Guinea is most closely related to?

In the main text, we began with the goal of constructing an evolutionary tree for a distance matrix holding the "distances" between every two pairs of present-day species under consideration. We saw that given a multiple alignment, we can construct a distance matrix for which the distance between two species is the number of differing symbols between their rows of the alignment.

Fortunately, MEGA includes the multiple alignment program ClustalW. This allows us to create phylogenetic trees directly from a multiple alignment, removing the intermediate step of computing the distance matrix. Let's see how MEGA does this using ten different Ebola genomes isolated from humans in different Ebola outbreaks.

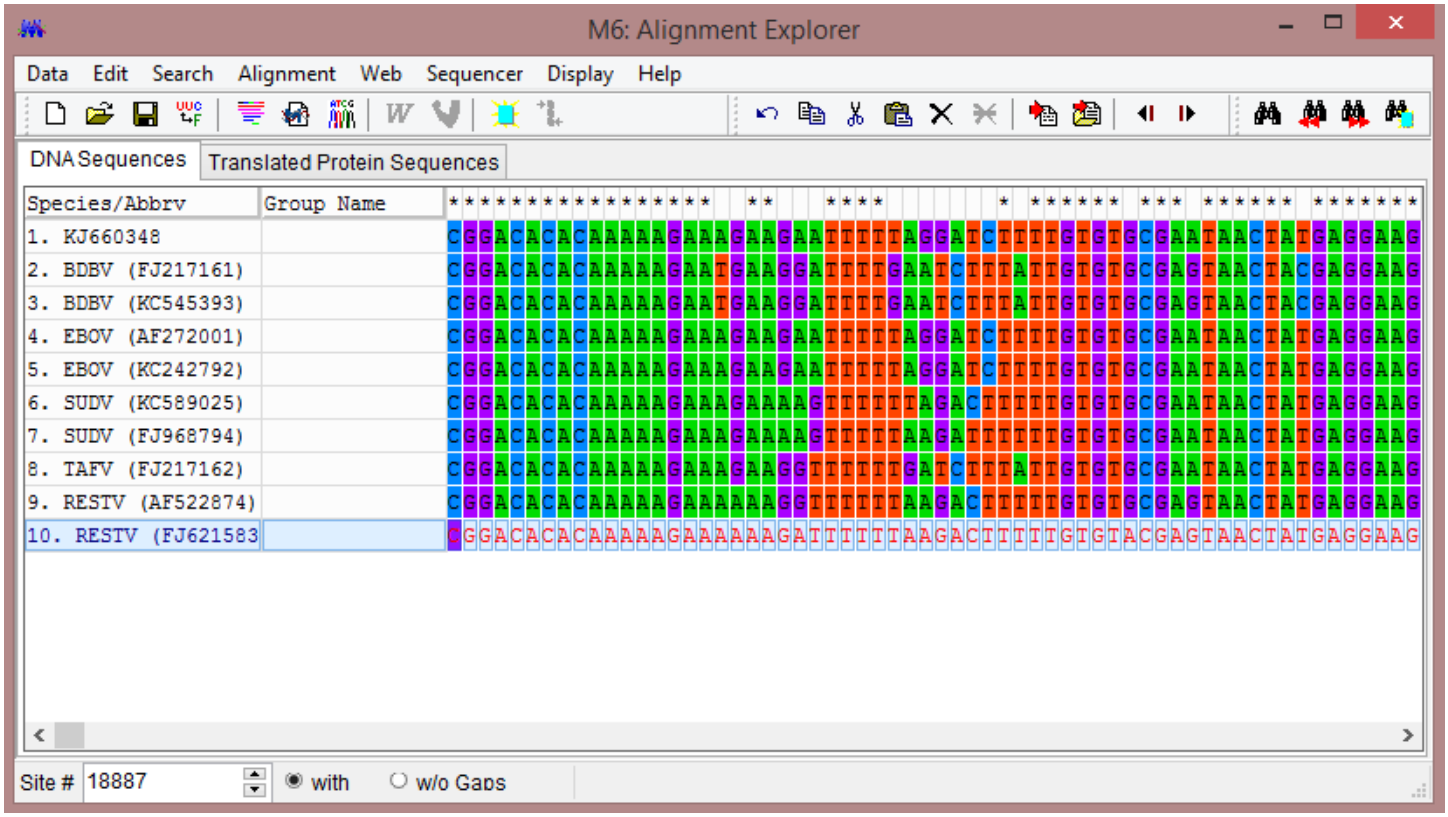
First, download and install MEGA v. 6.0 from the MEGA website. Then, open the program, click the "Align" icon near the top of the application, choose "Edit/Build Alignment" and create a new alignment. If asked, indicate that you are building a DNA sequence. At this point, an alignment explorer should open and you can start selecting sequences to align.

We are going to align the following ten DNA sequences. The length of these sequences fall in the range 18,875 – 18,959 base pairs.

Accession Number	Virus Species	Location	Date
KJ660348	????	Gueckedou, Guinea	2014

FJ217161	BDBV	Bundibugyo, Uganda	2007
KC545393	BDBV	Isiro, DRC	2012
AF272001	EBOV	Yambuku, DRC	1976
KC242792	EBOV	Mekouka, Gabon	1994
KC589025	SUDV	Luwero, Uganda	2012
FJ968794	SUDV	Sudan	1976
FJ217162	TAFV	Tai Forest, Ivory Coast	1994
AF522874	RESTV	Philippines	1990
FJ621583	RESTV	Philippines	2008

You can open these sequences in the alignment explorer by clicking on “Web” at the top of the application and selecting “Query GenBank”. Use the Accession Numbers on the table above to look up ten sequences that we are using to build a phylogenetic tree. To search GenBank, simply type an Accession Number into the search box at the top of the page, keep “Nucleotide” in the dropdown box, and click “Search”. You will be taken to the GenBank page for this Accession Number. Next to the Address Bar of the web browser window, you will see a button that says “Add To Alignment” with a red plus sign next to it. Click this button to add the sequence to your dataset. For “Sequence Label”, for the sake of simplicity and consistency for this assignment, delete what is automatically entered and instead enter the Accession Number you are searching for. Repeat for the other nine sequences until all ten sequences are populated. Your Alignment Explorer screen should resemble the following screenshot:



Now select all ten sequences, then click on the "Alignment" menu and select "Align by ClustalW". This will perform a multiple sequence alignment on your selected data. If asked for parameters, use the default values.

The alignment takes quite a long time to complete (possibly up to an hour). Thus, while ClustalW is running, we will review how to construct a phylogeny from a distance matrix using neighbor-joining.

Objective 2: Applying the Neighbor-Joining Algorithm

The neighbor-joining algorithm is one of the most popular methods for evolutionary tree reconstruction. We will first use the neighbor-joining algorithm to construct a small tree by hand.

Below is a distance matrix *D* filled with distances between four different organisms.

	W	**X**	**Y**	**Z**
W	0	11	2	16
X	11	0	13	15
Y	2	13	0	9
Z	16	15	9	0

2. Use the Neighbor-Joining algorithm to construct a phylogeny for the above distance matrix.

Fill in the following distance matrices and tree reconstruction diagram:

Construct the neighbor-joining matrix D_1* from the distance matrix D given above.

	W	**X**	**Y**	**Z**
W				
X				
Y				
Z				

Construct a 3x3 distance matrix D_2 using the neighbor-joining matrix D_1*. "A" refers to 2 out 4 original leaves (W, X, Y, and Z) in the tree that were joined by the neighbor-joining algorithm. You will need to fill in the two branches that were not joined in addition to all distances in the resulting 3x3 distance matrix.

	A		
A			

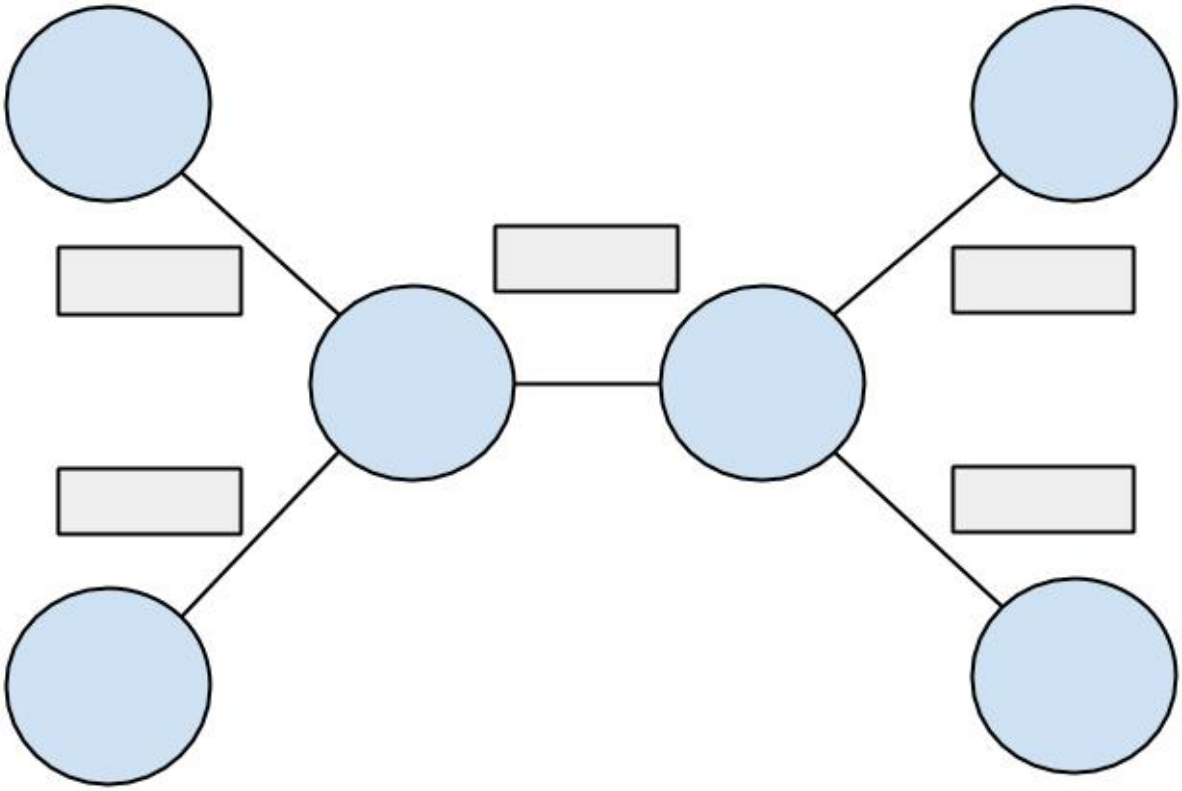
Construct the neighbor-joining matrix D_2* from the distance matrix D_2 given above.

	A		
A			

Construct a 2x2 distance matrix D_3 using the neighbor-joining matrix D_2*. "B" refers to the new leaf formed by joining two leaves in the neighbor-joining algorithm. You will need to fill in the branch that was not joined in addition to the distance in the resulting 2x2 distance matrix.

	B	
B		

Fill in the node labels and distances in the tree below. Specifically, fill in the circles with either the strain identifier (W, X, Y, Z), or labels A and B. Fill in the squares with the corresponding branch length.



Objective 3: Using MEGA to Construct a Phylogenetic Tree

Once ClustalW has finished constructing a multiple alignment of *Ebolavirus* sequences, you will need to load the alignment data to construct a tree. Click on the "Data" menu and select "Phylogenetic Analysis". If it asks you if the data is protein-coding nucleotide sequence data, select no (because we are aligning entire genomes, not just protein-coding regions). The data should now be loaded into the primary MEGA window. In the primary MEGA window, click on the "Phylogeny" icon and select "Construct/Test Neighbor-joining tree". It should then ask if you want to use the currently active data. Select "Yes". You should now see an image of the evolutionary tree.

If you do not see branch lengths, you should enable them. Click "View" at the top of the Tree Explorer. Hover over "Show/Hide", then click on "Branch Lengths". You should now be able to see all branch lengths of the tree. You can also go to "View" and then "Options" to adjust other properties of the tree. The "Tree" tab is especially of interest because you can adjust "Taxon Separation", "Branch Length", and "Tree Width". If you are having difficulty reading the branch lengths because they are too close together, you can click the "View" menu and select "Topology only".

3. Include an image of the phylogenetic tree that you created.
4. Which *Ebolavirus* species caused the 2014 outbreak? Is it the same species that you predicted in question 1?

Objective 4: Using MEGA to Find the Animal Reservoir of the Ebola Outbreak

We have now observed that the virus that caused the epidemics found in West Africa in 2014 should be classified as *Zaire ebolavirus* (EBOV). However, the 2014 outbreak began not in Zaire (present-day Democratic Republic of the Congo), but over a thousand miles away in Guinea! What has caused the virus to move so far without infecting a single patient along the way?

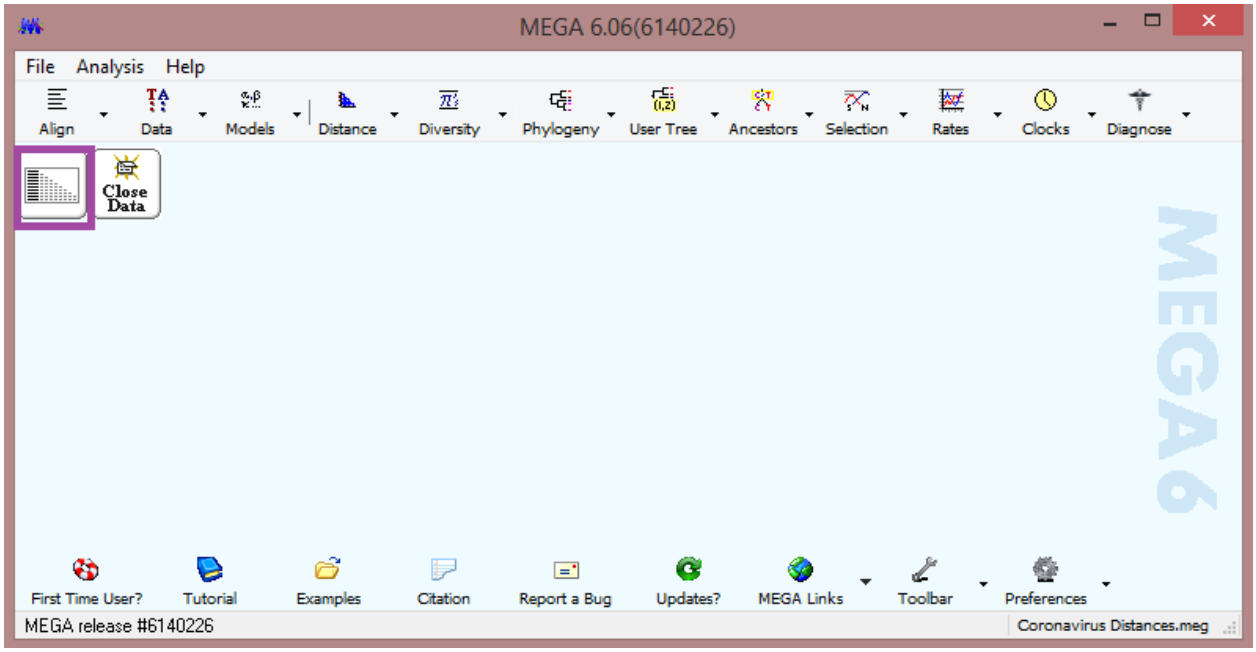
Animals are common reservoirs of disease, and viruses often live, multiply, and evolve in animal species before crossing over to humans. In the main text, we saw that the palm civet was the reservoir for SARS. Similarly, rats, chipmunks, and squirrels are reservoirs for bubonic plague; raccoons, skunks, and foxes are reservoirs for rabies; geese and ducks are reservoirs for bird flu; and ticks are the reservoir for Rocky Mountain spotted fever. In contrast, diseases like polio and smallpox have no animal reservoir. The lack of an animal reservoir makes it much easier to completely eradicate the disease, which is why smallpox was completely eradicated and polio has been limited down to just three countries (Nigeria, Pakistan, and Afghanistan).

As we saw in the main text, we can identify the animal reservoir of a virus by constructing an evolutionary tree of viruses taken from various animal species. Doing so will help us understand which animal brought *Zaire ebolavirus* to West Africa and initiated the 2014 outbreak.

Download the file *Species Distances.meg* from the class website. In MEGA, click on the "Data" option at the top of the application. Then select "Open a file/session…". Select the *Species Distances.meg* file that you just downloaded. Here is a summary of the virus sequences included in *Species Distances.meg*.

Virus Species	***Host Species***	***GenBank ID***	***Protein Sequence Length (aa)***	***Protein Function***	***Location***
Ebolavirus	Human	AIN75249.1	184	L Gene	Democratic Republic of the Congo
Ebolavirus	Bat	ABB18310.1	88	L Gene	Gabon
Ebolavirus	Pig	ACT22790.1	251	vp24	Philippines
Ebolavirus	Pig	ABX75370.1	288	vp30	Russia
MERS Coronavirus	Camel	AHL18097.1	219	Glycoprotein	Egypt
Dengue Virus	Mosquito	AAD31720.1	501	Envelope protein	Kenya
Dengue Virus	Horse	ACS36231.1	106	Envelope protein	South Africa
Influenza	Chicken	AFC17983.1	171	hemagglutinin	Egypt
Influenza	Chicken	ADB25344.1	212	Nonstructural Protein	Nigeria
Influenza	Duck	ABJ53169.1	195	Matrix Protein 1	South Africa

In the middle section of MEGA, you should see two icons appear. The icon on the left, shaped like a staircase, will display the data selected (outlined in purple in the image below).

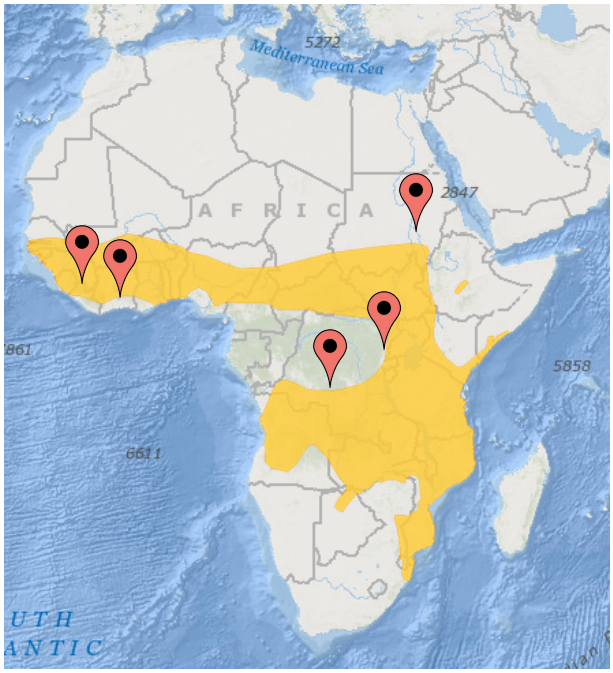


Next click on the "Phylogeny" option in the top row of MEGA. Then select "Construct/Test Neighbor-Joining Tree...". Use the default options and click "Compute".

You should now see an image of the phylogeny. If you do not see branch lengths, you should enable them. To do so, click "View" at the top of the Tree Explorer. Hover over "Show/Hide" then click on "Branch Lengths".

5. Include a screenshot of the generated phylogenetic tree with branch lengths displayed.

It turned out that Emile Ouamouno frequently played near a hollow tree where bats nested. The preceding exercise makes it clear that bats likely initiated the outbreak in Guinea after migrating from Central Africa. The following two images, which show the migration patterns of two species of African bats that have been confirmed as carriers of Ebola, make this theory even more compelling. Note that these migration patterns essentially cover every place name for *Ebolavirus* species (discussed at the beginning of objective 1).



Range of *Mops condylurus* (Angolan free-tailed bat)


Range of *Myonycteris torquata* (Little collared fruit bat)

Objective 5: Constructing phylogenetic tree with UPGMA.

Previously, you generated a tree using the Neighbor-Joining algorithm. You will now use the alignment data generated in objective 3 to construct a phylogenetic tree using two other algorithms for phylogenetic tree reconstruction: the UPGMA and Maximum Parsimony algorithms.

Click on the "Phylogeny" button and click "Construct/Test UPGMA Tree". Use the default options and click "Compute".

You should now see an image of the phylogenetic tree. If you do not see branch lengths, you should enable them. Click "View" at the top of the Tree Explorer. Hover over "Show/Hide" then click on "Branch Lengths". You should now be able to see all branch lengths of the tree. You can also go to "View" and then "Options" to adjust other properties of the tree. The "Tree" tab is especially of interest because you can adjust Taxon Separation, Branch Length, and Tree Width.

6. Include an image of the phylogenetic tree constructed by UPGMA.
7. How does this tree differ to the tree constructed by Neighbor-Joining?
8. Where is the root of the tree? What is the total distance D_{max} of the tree (i.e. the distance from any leaf to the root)?
9. What is the distance D_{2014} between the leaf corresponding to the Ebola virus of 2014 and the internal node at the beginning of the branch leading to this leaf?
10. How much time did it take for the Ebola virus of 2014 to split from other Ebola viruses after their most recent common ancestor?
11. Export the tree (in Tree Explorer, go to File → Export Current Tree (Newick) → choose a name and click OK). In MEGA, go to "Clocks" → "Compute Timetree (RelTime-ML)" and open the file you just saved. Click "Yes" in the popup about specifying time calibration constraints. Click "New". For "Taxon

A" and "Taxon B", choose the two BDBV entries. For "Calibration Name", type "BDBV". Set both "Min Divergence Time" and "Max Divergence Time" to 10 (we are estimating that the BDBV strains split apart from a common ancestor 10 years ago). Click "Save Changes" and click "Next Step", then click "Compute".

According to the resulting tree, how many years ago did the 2014 Ebola strain (KJ660348) split apart from its closest common ancestor? Include an image of the tree.

Objective 6: Constructing phylogenetic tree with Maximum Parsimony.

We will now apply MEGA to construct a phylogenetic tree using Maximum Parsimony algorithm. This time, click on the "Phylogeny" button and click "Construct/Test Maximum Parsimony Tree(s)". Note that Maximum Parsimony requires multiple sequence alignment data that we have generated (it cannot be applied to a distance matrix). Use the default options and click "Compute".

You should now see an image of the phylogenetic tree.

12. Is the Maximum Parsimony tree rooted or unrooted? Include an image of the phylogenetic tree you created.
13. What differences do you see between this tree and the trees generated using the previous two methods (Nearest-Neighbor and UPGMA)?
14. This step requires Microsoft Excel or OpenOffice, so if you have neither of these, please install OpenOffice (which is free). Recall that Maximum Parsimony can infer ancestral sequences. In the Tree Explorer, click "File", then "Export Current Tree (Newick)", and name your tree "parsimony.nwk". Return to the main MEGA screen, click on the "Ancestors" button, and select "Infer Ancestral Sequences (Parsimony)". Keep the default settings, and for "User Tree File" (the second yellow box), click the "…" symbol and browse to select the "parsimony.nwk" file you just exported. Click "Compute", and a new Tree Explorer window will open. Click on the "Ancestors" button in the toolbar, then "Export Changes List", then click "OK". How many sites have changes between KJ660348's ancestor and KJ660348? You can simply count (not by hand, as there are too many) how many entries are in the column that has an arrow pointing to KJ660348.

Conclusion

Throughout these practical challenges, we have explored three different methods for creating phylogenetic trees: Neighbor-Joining, UPGMA, and Maximum Parsimony. We used these methods to explore evolution and spread of the 2014 Ebola virus, and we pinpointed that bats are the most likely animal reservoir for Ebola.

The fight against Ebola is an ongoing one, as scientists hold out hope for a vaccine that would save thousands of lives in the future. For more information about *Ebolavirus* genomes, you can visit the Ebola Genome Browser or the NCBI Ebola Virus Variation website. You may also find interesting background information on Ebola at the following links.

http://www.who.int/mediacentre/factsheets/fs103/en/
http://apps.who.int/ebola/

http://www.who.int/reproductivehealth/topics/rtis/ebola-virus-semen/en/
http://www.npr.org/2014/10/23/358363535/why-do-ebola-mortality-rates-vary-so-widely
http://www.sciencealert.com/origin-of-2014-ebola-outbreak-traced-to-kids-favourite-hollowed-tree