Software Challenge: Constructing Evolutionary Trees (Answer Key)
02-604: Fundamentals of Bioinformatics
Spring 2018

Background

The 2014 Ebola outbreak was the deadliest Ebola epidemic in history, infecting over 26,000 people in West Africa and takin 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, where the virus has been found 89 days after symptom onset, 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

Smaller Ebola 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 (2014), Tai Forest, Zaire, Bundibugyo, and Sudan.

1. **Based on the locations on the map above, to which species of Ebolavirus would you expect the 2014 outbreak in Guinea to be most closely related? Why?**

Tai Forest, since it is closest geographically to Guinea.

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.

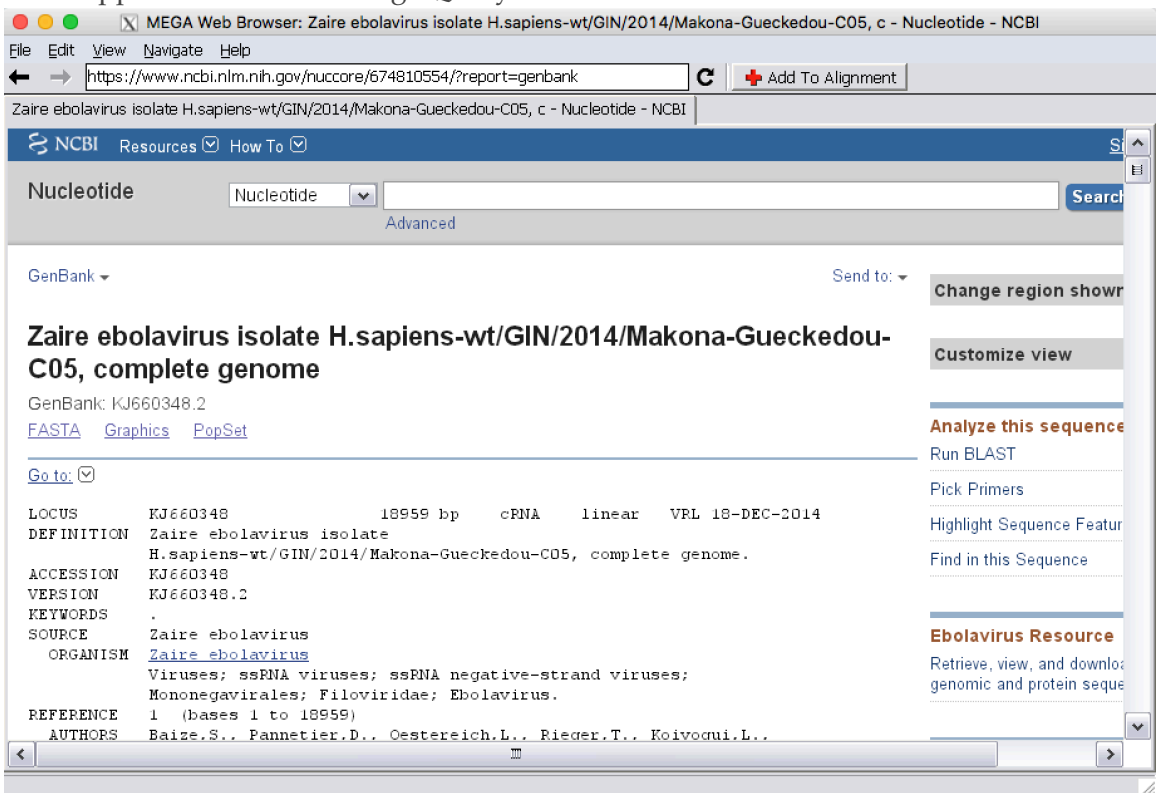
Fortunately, MEGA includes the multiple alignment program ClustalW, which we encountered in the alignment Software Challenge. This allows us to create phylogenetic trees directly from a multiple alignment, removing the intermediate step of computing the distance matrix. Let's see how MEGA does this using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

First, download and install MEGA v. 7.0 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 alignment. At this point, an alignment explorer should open and you can start selecting sequences to align.

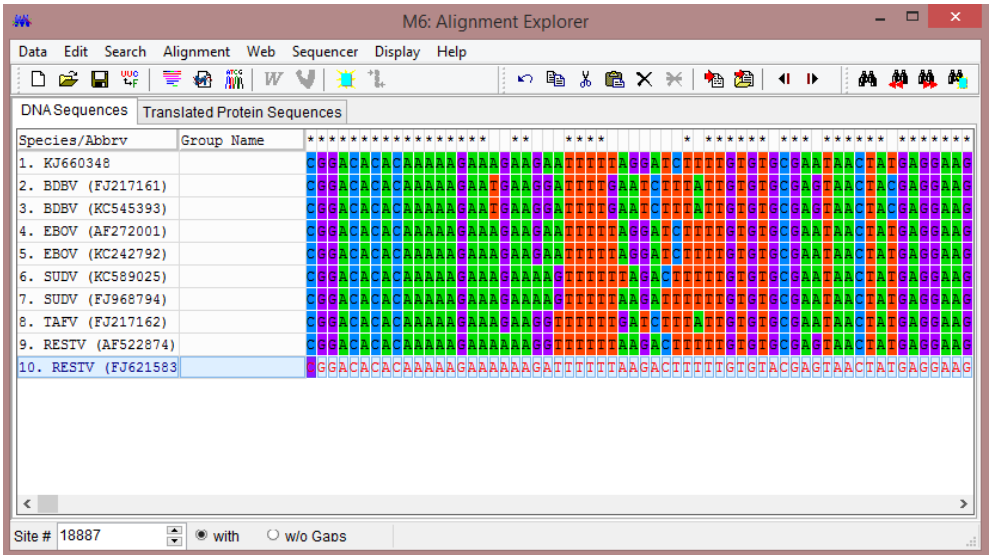
We are going to align the following ten DNA sequences. The length of these sequences fall in the range 18,875 – 18,959 base pairs.

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 as well as the virus species it derives from. Repeat for the other nine sequences until all ten sequences are populated. Your Alignment Explorer screen should resemble the following screenshot:



Now click on the "Alignment" menu and select "Align by ClustalW" (when prompted to select all, click "OK"). This will perform a multiple sequence alignment on your selected data. When prompted for parameters, use the default values.

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

Objective 2: Applying the Neighbor-Joining Algorithm

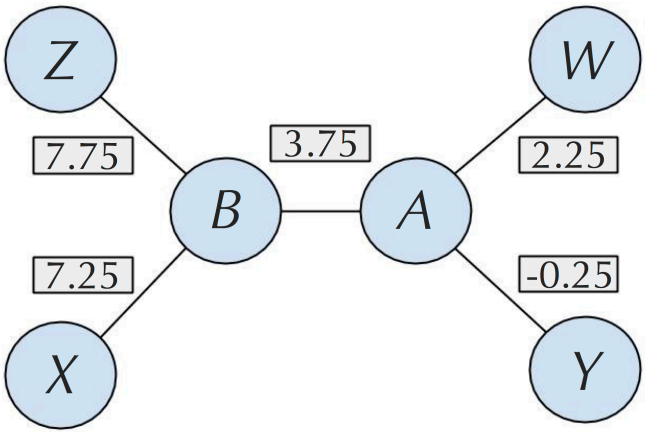
The neighbor-joining algorithm is one of the most popular methods for evolutionary tree reconstruction. We will first use the neighbor-joining algorithm to construct a small tree by hand.

Below is a distance matrix *D* containing distances between four different organisms

labeled W, X, Y, and Z.

	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.**



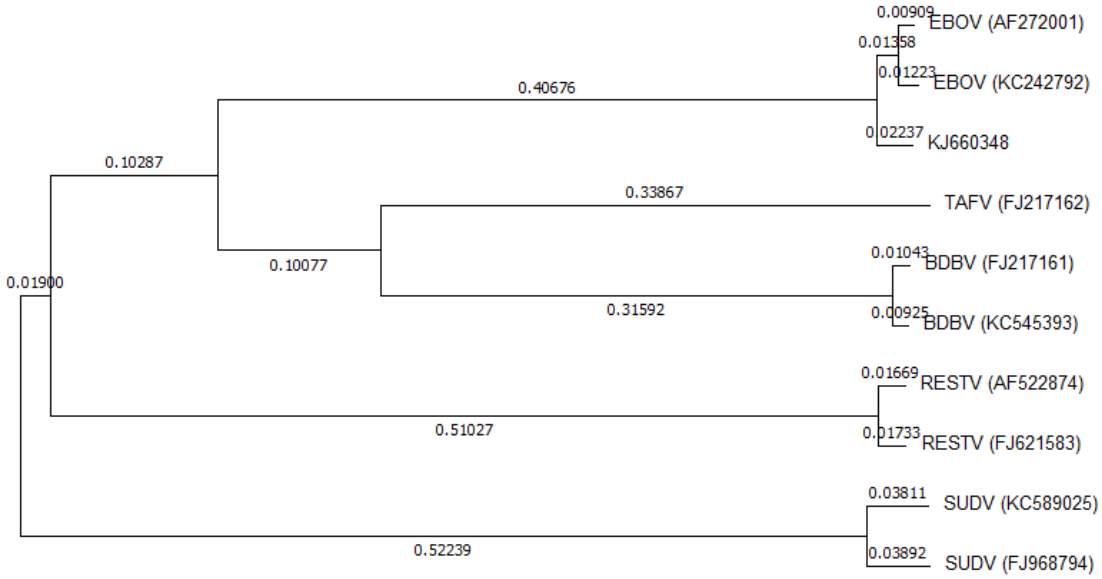
Objective 3: Using MEGA to Construct a Phylogenetic Tree

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 the program 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”. After clicking "Compute", you should now see an image of the evolutionary tree.

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”. 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”.) Click

"View" --> "Options" --> "Branch" and select 5 decimal places of accuracy.

3. **Provide an image of the phylogenetic tree that you created.**



Note to Instructor: The numbers might be off slightly in the 4th-5th decimal place.

4. **Based on the tree produced by Neighbor-Joining, which *Ebolavirus* species caused the 2014 outbreak? Why? Is it the same species that you predicted in question 1?**

The two EBOV strains are closest to KJ660348 in the tree, so it is likely that EBOV caused the 2014 outbreak. This is not the same species as the Tai Forest species predicted in Objective 1.

Objective 4: Constructing a phylogenetic tree with UPGMA.

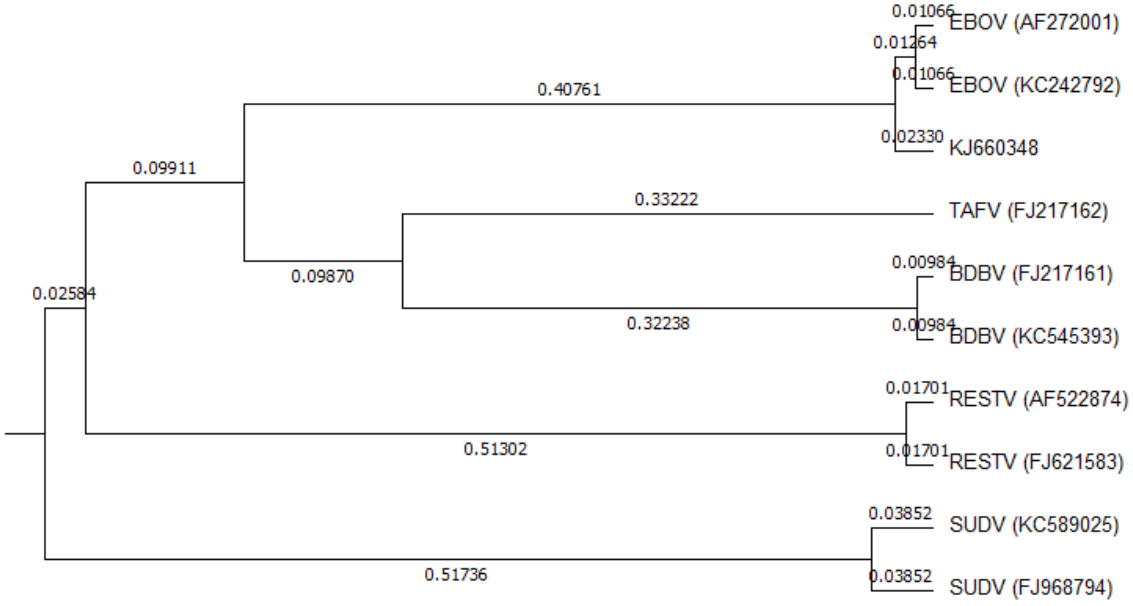
Previously, you generated a tree using the Neighbor-Joining algorithm. You will now use the alignment data generated in Objective 3 to construct a phylogenetic tree using two

other algorithms for phylogenetic tree reconstruction: the UPGMA and Maximum Parsimony algorithms.

Click on the "Phylogeny" button and click "Construct/Test UPGMA Tree". When asked to use the currently active data, click "Yes". 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.

5. **Include an image of the phylogenetic tree constructed by UPGMA.**



6. **Where is the root of the tree produced by UPGMA?**

The root of the tree is at the branch separating SUDV (KC589025 and FJ968794) from the rest of the Ebola sequences.

7. **What is the total distance D_{max} of the tree produced by UPGMA (i.e., the distance from any leaf to the root)?**

The total distance *Dmax* is **0.557** (because of rounding, answers may vary by about 0.005).

8. **What is the distance in the UPGMA tree between the leaf corresponding to KJ660348 and the internal node at the beginning of the branch leading to this leaf?**

$D_{2014} = 0.023$

9. **According to the UPGMA tree, how long did it take for the Ebola virus of 2014 to split from other Ebola viruses in the tree after their most recent common ancestor? How did you obtain this answer? Provide your answer in the distance units given in the tree (i.e. a decimal), not as actual time measurements.**

The distance from the root to the branch point where the 2014 Ebola virus branched away from its closest ancestor is **0.534.** This value was obtained by subtracting the two previous answers.

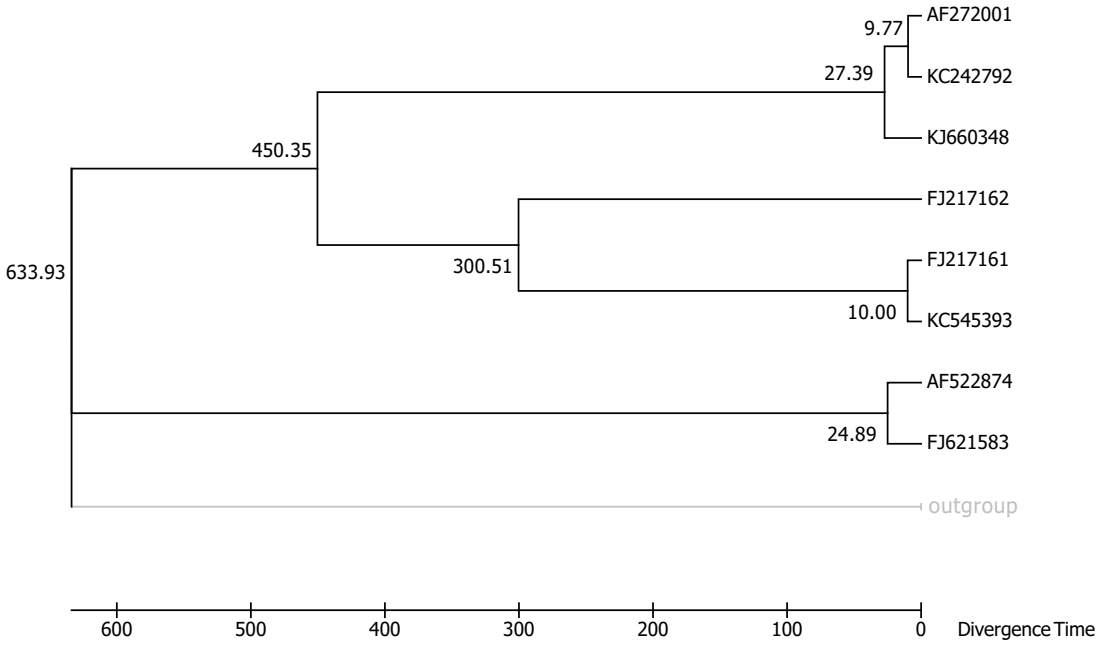
Export the tree (in Tree Explorer, go to File → Export Current Tree (Newick) → click "Export"). This exports the tree in a concise generalizable format called "Newick format": if you're interested, a short description of Newick format at http://rosalind.info/glossary/newick-format/. Then, save this text file in the same location you saved the .meg file.

In the MEGA home screen, go to "Clocks" → "Compute Timetree (RelTime-ML)". When prompted whether to use the currently active data, click yes (this refers to the alignment file). Then, under "Step 2: Load Tree File", click "Browse", and select the .nwk file you just saved. In Step 3, MEGA demands that we specify an "outgroup", or a group of species that we know are separate from the tree. For this tree, we want to separate KC589025 and FJ968794 – click "Select Branch" and click the branch connecting these two species. Then click "Finished", and "Yes" to confirm. Under step 4, click "add constraints". To calibrate the tree, we are going to assume that the two BDBV entries (FJ217161 and KC545393) diverged from a shared ancestor 10 years ago. To do this, click the parent of these two

nodes, and then select "Calibration" → "Calibrate Selected Node". Set both "Min Divergence Time" and "Max Divergence Time" to 10 (normally, this would allow us to specify a range for the divergence time of these two species. Click "Finished", then under Step 5, click "Set Options" and then "Save" to use default options. We are now ready for Step 6: Launch Analysis. Click "Execute".

10. **According to the resulting tree, how many years ago did the 2014 Ebola strain (KJ660348) split apart from its closest common ancestor? In other words, what is the number next to the rectangle at the branching point that is KJ660348's closest common ancestor with its nearest neighbors? Include an image of the tree.**

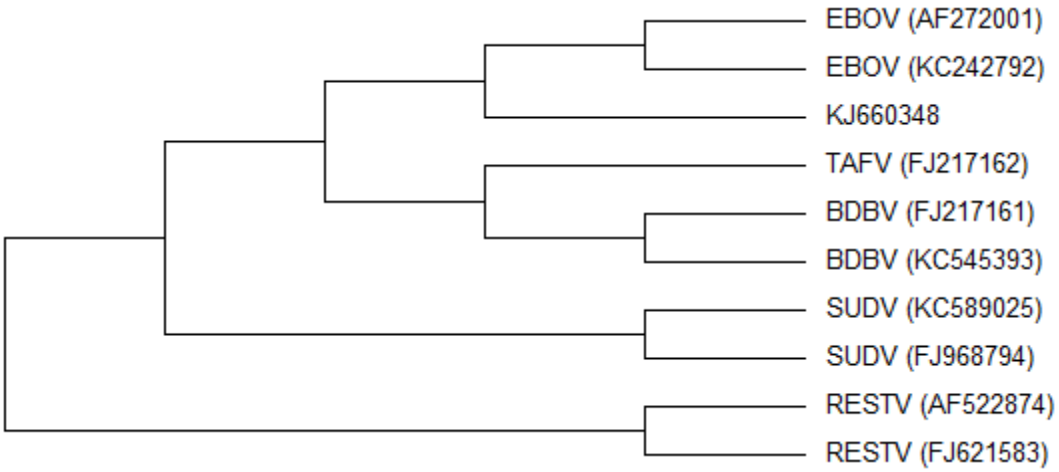
The answer is **27.39**; please see below for the tree.



Objective 5: Constructing a phylogenetic tree with Maximum Parsimony

We will now apply MEGA to construct a phylogenetic tree using the 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.

11. **Include an image of the phylogenetic tree that you created.**



12. **Is the Maximum Parsimony tree rooted or unrooted?**

The Maximum Parsimony tree is **unrooted**.

13. **Under "Ancestors" (in the menu bar), click "Show All." Notice that nucleotides appeared at every node of the tree. What does the nucleotide at a given node represent?**

Recall that each node represents an individual sample: leaves are samples that we provide, and internal nodes are ancestors that are predicted by the algorithm. The nucleotide at a given node is the nucleotide predicted in the given sample at the specified position.

This step requires Microsoft Excel or OpenOffice, so if you have neither of these, please install OpenOffice (which is free). Click on the "Ancestors" button in the toolbar, then "Export Changes List" (only shown if "Show All" is selected in the "Ancestors" menu), then click "OK".

14. **How many sites have changes between KJ660348's ancestor and KJ660348? You can count (not by hand, as there are too many) how many entries are in the column that has an arrow pointing to KJ660348.**

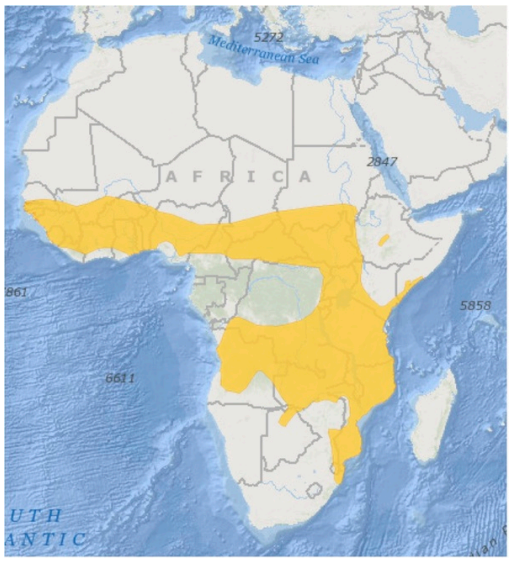
Using the COUNTA function of Excel, I count **10644** changes from KJ660348's closest ancestor to KJ660348.

Conclusion

We already 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 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).

15. **Take a look at the two figures below, which show the ranges of the Angolan free-tailed bat and little collared fruit bat, respectively. What can you conclude about these ranges?**



Note that the bat migration patterns essentially cover every place name for Ebolavirus species (discussed at the beginning of objective 1), which we have indicated below. Therefore, we could be led to hypothesize that bats are the animal reservoir of Ebola.

Emile Ouamouno, patient zero of the 2014 outbreak, frequently played near a hollow tree where bats nested. Researchers have therefore proposed that bats, as is the case for many other epidemics, may have caused the most recent SARS epidemic. In other words, the virus is still evolving within bats, and it crosses the species barrier to humans from time to time, causing an outbreak.

However, this theory is still just a hypothesis, and 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. Another cutting edge innovation that may help us track the spread of future epidemics is real-time genome sequencing as soon as patients are diagnosed. The result of this real-time, portable genome sequencing for the Ebola outbreak was described by a global team of researchers in January 2016:

- http://www.nature.com/nature/journal/vaop/ncurrent/full/nature16996.html

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