TEACHER GUIDE

Disease Detectives - Introduction to Sequence Analysis

Overview

In this lesson, students will analyze Ebola sequences that were obtained from patients in Sierra Leone during the 2014 outbreak in West Africa. Students are challenged to place sequences into groups based on similarities and create a story for transmission of the virus. They will then compare their results to those of scientists at the Broad Institute of MIT and Harvard, who followed a similar procedure in the beginning of the outbreak.

The lesson is structured to provide flexibility in the classroom. In a 60 minute class period, the lesson can be completed without the need for students to access a computer. In a 90 minute class period, the lesson can be taught using genome viewing software called Jalview if each student has access to a computer.

Learning Objectives

After completing the lesson, students will be able to analyze and interpret DNA sequence data. They will demonstrate their understanding by creating a visual representation of the Ebola virus transmission chain in Sierra Leone.

Grade Level

9-12

Suggested Time

60-90 minutes

Resources

Background Information

Ebola is a hemorrhagic fever virus that affects multiple organ systems in the body. Like all viruses, it uses receptor sites in host cells to infect healthy cells and replicate itself. However, the Ebola virus is extremely harmful because of how fast it can replicate and its ability to infect nearly every type of cell. Infected cells can attach themselves to blood vessels, causing uncontrollable internal bleeding in patients.

There have been several Ebola outbreaks in Central Africa, the first of which occurred in 1976 in the Democratic Republic of the Congo and infected about 300 people. This was the largest outbreak on record before 2014, when Guinea, Sierra Leone, and Liberia became the first countries in West Africa to become affected by the virus. As of October 2014 there were about 9,000 reported Ebola cases, nearly 30 times the 1976 outbreak. This drastic increase in cases could be attributed to regional differences. Central Africa is predominantly forested with limited access to roads, while West Africa has several large cities and better transportation infrastructure, making it easier for infected patients to travel between communities and across borders.

Ebola spreads by close contact with bodily fluids from an infected patient, such as blood, saliva, urine, and sweat. One way that scientists can track how Ebola is spreading from person to person is through the use of DNA sequencing. Each Ebola patient has the virus in their blood, and each virus has a genome made up of a sequence of letters (G, C, A, and T). Over time, random changes to the sequence of letters occur, referred to as mutations. When a new sequence is compared to an older one, a higher number of mutations are expected. By studying differences in the sequence of letters over time, scientists can reconstruct the history of how the virus is mutating.

Tracking mutations is important for several reasons. The test used to diagnose infected patients relies on identifying the virus's genome sequence, so it's important to track mutations over time to ensure the test will still be able to detect the virus. Similarly, there are several therapies and vaccines in development that work by attacking the protein sequences of the virus. If the virus mutates significantly, the antibodies in these therapies may not be effective. By continuing to sequence Ebola, scientists can stay on top of how the virus is changing to ensure that diagnostics and therapeutics will be successful. The sequencing data can also be used to track where the virus is moving geographically. As mutations accumulate and similar sequences are grouped together, scientists can better understand how the virus is being transmitted and may be able to predict where the virus is heading.

Scientists in Pardis Sabeti's lab at the Broad Institute played a vital role in the Ebola outbreak in Sierra Leone. Within the first few weeks of confirmed cases, they sequenced Ebola samples from 78 patients to try and gain insight into how the virus was mutating and what strains of the virus were present. They compared their data to a reference sequence from a patient in Guinea where the Ebola outbreak began. The group published their sequence results in a public database so that researchers around the world could do similar analyses, one example being to assess if mutations in the virus could effect the efficacy of the experimental drug ZMapp, which was used on the first few patients brought to the United States for treatment.

Note: In this lesson we have described Ebola as a DNA virus for simplicity purposes, when in actuality it is an RNA virus. All viruses contain tiny bits of nucleic acid that code for everything the virus needs to make copies of itself. Many viruses use DNA as their nucleic acid just as humans do. Other viruses like Ebola use RNA as their nucleic acid and require an additional step called reverse transcription to copy their RNA into DNA. These viruses then use transcription to make many RNA copies of the nucleic acid for packaging into new virus particles. Scientists use reverse transcription to copy the RNA to DNA prior to sequencing, so the data that students will be analyzing is actually DNA. These additional steps are not relevant to the lesson, which is about using changes in nucleic acid sequence to track how the Ebola virus moved through the population in Sierra Leone over time.

Before the Lesson

1. For homework, ask students to read **The Virus Detectives: Sifting Through Genes in Search of Answers on Ebola**, a *New York Times* article highlighting the Sabeti Lab at the Broad Institute and their efforts to sequence the Ebola virus. Students should then answer the questions on the **Virus Detectives Homework** worksheet.
2. If you have 90 minutes and choose to use the genome viewing software for the main lesson, download the Jalview Desktop on all of the classroom computers ahead of time (http://www.jalview.org). Jalview is a free, open source program allowing you to edit, visualize, and analyze multiple sequence alignments. You may need to change the security settings on the computers to successfully open the program.

During the Lesson

1. Show the **Can DNA Help Us Fight Ebola?** video to introduce some key scientific concepts related to Ebola.
2. Use the **powerpoint slides** to give an overview of the 2014 Ebola outbreak and the importance of sequencing the virus. Discussion points are provided in the notes section of the slides.
3. Explain to students that they will be analyzing Ebola sequences that were obtained from patients in Sierra Leone during the 2014 outbreak. They will be following a procedure similar to that of the scientists who analyzed this data by grouping sequences together. Students' answers will vary based on the criteria they choose for grouping sequences. Researchers at the Broad Institute went through several iterations of data analysis before deciding how to group their sequences.
4. **If you have 60 minutes:**
    - Pass out the **Ebola Sequences** handout to students. Show the **Ebola Virus Analysis Tutorial - Visualizing the Data** video so students understand where the data in their handout comes from.

- Point out that the sequences in the handout are identical to the data shown in the video. The colored squares in sequences 1-15 represent mutations compared to the reference sequence. Remind students that the first sequence is a reference sequence from Guinea. It is from an early stage in the outbreak, and is several months older than the other sequences. Ask students to cut out the sequences into horizontal strips.

If you have 90 minutes:

- Have students download the sequence data from the NCBI website and analyze it using genome viewing software called Jalview. Directions to download the data can be found on the **Analysis Instructions - Downloading the Data** handout, or can be seen in the **Ebola Virus Analysis Tutorial - Downloading the Data** video.
- Show the **Ebola Virus Analysis Tutorial - Visualizing the Data** video so students understand how to use the software to analyze their sequences. After viewing the video, students can use the **Analysis Instructions - Visualizing the Data** handout to help guide them as they use the software on their own.

5. Ask students the following questions, giving them time between each question to move and analyze their sequences:
   - **Do any patterns in the sequences jump out at you?** *(*possible answers: *some sequences are identical, some only differ by one mutation, others differ by many mutations)*
   - **What attributes of the sequences could be used to divide them into groups?** *(*possible answers: *identical sequences, sequences that only differ by 1 mutation, location of the mutations, number of mutations)*
   - **Are there specific sequences that you think should be grouped together?** *(explain that grouping identical sequences is a good way to start, but students will have to decide how similar the sequences in each group should be. The number of groups students make will vary based on the criteria they use for grouping.)*
   - **How do you decide on the best number of groups?** *(explain that by grouping sequences, scientists can learn something about how they are related and how the virus is changing as it spreads between patients. It would not be helpful to put all the sequences in 1 group—or have each sequence in its own group—because scientists would not be able to compare changes in sequences to learn how the virus is transmitted between people.)*
6. Have students complete the **Disease Detectives - Introduction to Sequence Analysis** worksheet on their own, and review their responses as a class. Reference the **Worksheet Key** for class discussion tips.

After the Lesson

1. Hand out the **Scientist Visual** to students.
2. Explain that this visual, created by Broad scientists, shows how they chose to group sequences together to highlight the relationship between groups.
3. Ask:
   - **What can you infer by looking at this visual?** *(*possible answers: *group 3 represents patients that got sick after groups 1 and 2, the visual shows how the virus is changing over time so it's possible to figure out where the virus is moving throughout the population, it represents a timeline of the transmission of the virus)*
   - **Why is this type of analysis important to do during an outbreak?** (*knowing how the virus is mutating over time is important for diagnostics, therapeutics, and to understand the transmission of the virus as the outbreak is happening*)

Next Generation Science Standards Alignment

Science and Engineering Practices	
Asking questions and defining problems	✔
Developing and using models	
Planning and carrying out investigations	
Analyzing and interpreting data	✔
Using mathematics and computational thinking	
Constructing explanations and designing solutions	✔
Engaging in argument from evidence	✔
Obtaining, evaluating, and communicating information	✔
Crosscutting Concepts	
Patterns	✔
Cause and effect	✔
Scale, proportion, and quantity	
Systems and system models	
Energy and matter	
Structure and function	✔
Stability and change	✔
Interdependence of science, engineering, and technology	✔
Influence of engineering, technology, and science on society and the natural world	✔
Disciplinary Core Ideas	
PS1 - Matter and its interactions	
PS2 - Motion and stability: Forces and interactions	
PS3 - Energy	
PS4 - Waves and their applications in technologies for information transfer	
LS1 - From molecule to organisms: Structures and processes	
LS2 - Ecosystems: Interactions, energy, and dynamics	
LS3 - Heredity: Inheritance and variation of traits	
LS4 - Biological evolution: unity and diversity	✔
ETS1 - Engineering design	

Acknowledgements

Christopher Angelli, Somerville High School
Harrison Dreves, Curious Minds
Rachel Gesserman, Broad Institute
Justine Lassar, Broad Institute
Aaron Lin, Broad Institute
Maria Maradianos, Somerville High School
Daniel Park, Broad Institute
Pardis Sabeti, Broad Institute
Rachel Sealfon, Broad Institute
Vivian Siegel, Broad Institute
Cisco Torres, Curious Minds
Joseph Vitti, Broad Institute
Shirlee Wohl, Broad Institute
Karen Woods, Somerville High School

References

1. Gire, Stephen K. *et al.* "Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak." *Science*. 345 (6202), 1369-1372 (2014).
2. Tam, Ruth. "This is how you get Ebola, as explained by science." *PBS Newshour*. (2014).
3. "Frequently Asked Questions on Ebola virus disease." *World Health Organization*. <http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/epr-highlights/3648-frequently-asked-questions-on-ebola-hemorrhagic-fever.html>.

The New York Times | http://nyti.ms/1yDlS5M

HEALTH

The Virus Detectives

Sifting Through Genes in Search of Answers on Ebola

By GINA KOLATA DEC. 1, 2014

CAMBRIDGE, Mass. — An old two-story brick building in a shabby part of town, formerly a distribution center for Budweiser beer, is now the world's most powerful factory for analyzing genes from people and viruses.

And it is a factory. At any given time, 10,000 tiny test tubes each holding a few drops of gene-containing fluid are being processed by six technicians, working 24 hours a day, 365 days a year — two on the night shift — using 50 dishwasher-sized machines in two large rooms.

The machines spit out sequence data onto a computer screen in the form of a long list, in order, of the letters that make up genetic material. That is three billion letters if the genes are from a person. Another 64 technicians do the more labor-intensive work of preparing the samples for analysis.

It is all in service of researchers who work for the Broad Institute, a gleaming, lavishly endowed genetics center a few blocks away. The sequencing center has worked on human DNA from an international effort, the 1,000 Genomes Project, that looks at the genes of thousands of people from around the world. It has gotten sequences of microbes, like dengue fever, malaria and West Nile virus. It has gotten genetic sequences from animals like chimpanzees.

And it is here that Broad scientists studying Ebola and a similar deadly disease, Lassa, send their samples, taking advantage of what the center's manager, Andrew J. Hollinger, referred to as superfast track sequencing in their urgent work on these diseases ravaging West Africa. Those scientists receive their sequence data in about 40 hours, compared with days for the usual work.

The Ebola and Lassa group, led by Pardis Sabeti, wants to know what the

viruses look like. Do they mutate while they are infecting people, possibly evading the immune system? Are some strains more deadly than others? And what about the genetics of the people who are infected? Are some people more resistant, perhaps even immune, to these viruses because of tweaks in their own genes?

The research is emblematic of a new direction in public health, which uses powerful genetic methods and applies them to entire populations.

The aim is to get a detailed picture of disease epidemiology, as the disease is happening. Armed with such data, doctors should be better able to stop epidemics and researchers can get clues to treating and preventing infections.

In one of their first investigations, the group traced the start of the Ebola epidemic in Sierra Leone from a single funeral in May that ended up infecting 14 women. One person who had been at that funeral showed up at Kenema Government Hospital a few hours' drive from the village where the funeral was held.

"That first case was manageable," Dr. Sabeti said. But several weeks after the funeral, there was a fear that an epidemic could have been sparked. The fear turned out to be true. "The virus was like a tidal wave coming into the country," Dr. Sabeti said.

Sierra Leone's department of health and safety sent epidemiologists to the remote village to trace the disease, asking who had been at the funeral and who had the funeral participants contacted. They found 14 ill with Ebola and an additional 35 people who tested negative for Ebola but had been exposed and had some symptoms.

Did they really have no virus in their bodies? That's where genetic sequencing provided an answer.

"The government wanted to know if they were negative for real or was the diagnostic test just not picking Ebola up?" Dr. Sabeti said. The blood from those people was sent to the Broad Institute, where any viral genetic material in it was sequenced. Those 35 were not infected — they had no Ebola virus in their blood. But the test found the virus in the blood of the 14 who had the disease.

As the group examined the genetics of the Ebola viruses in different patients — 78 in the first few weeks of the outbreak in Sierra Leone — they noticed the virus

was continually mutating, which raises questions about whether it could become airborne or more deadly. Dr. Sabeti said the mutations were not a surprise because that was what viruses did. But, she added, "it is also always something we should be concerned about." It probably would take many major mutations for the virus to be able to spread through the air or become more virulent, she said. "But, again, any change is one change too many, and we should stop this thing as quickly as we can."

For their continuing work on why some who are exposed to Ebola become sick and die while others escape infection or become sick and recover, the investigators need to study the genes of the patients themselves. That can be difficult. In Sierra Leone, Dr. Sabeti said, people do not want researchers studying cells of people who died.

"We all want to work through these issues carefully, " Dr. Sabeti said. "We do not want to desecrate the memories of people who died, so for now, we will be studying survivors."

While Dr. Sabeti and others work on Ebola, they also are working on Lassa and asking the same questions.

Lassa virus is much more common than the Ebola virus, but Lassa and Ebola infections have many of the same symptoms: fever, vomiting, bleeding in some cases.

Lassa also can have dreadful consequences — only 16 percent of those admitted to hospitals in Sierra Leone with Lassa survive. Lassa, unlike Ebola, infects the brain, so survivors often end up with permanent neurological damage like deafness, dizziness or psychiatric symptoms.

Dr. Sabeti's interest in Lassa was piqued seven years ago, before there was an Ebola epidemic, and before sequencing reached today's low price and fast speed. She had decided to look at already-determined DNA sequences from people around the world with a simple question: Are there new gene mutations, ones that only recently emerged in a population, that might protect against disease? The idea was that if a disease entered a population and was deadly, those who carried a protective mutation would survive and reproduce and soon that good mutation would become common.

She saw one such mutation in Nigeria — it was a slight tweak in a gene and so common that 34 percent of the population there has it. The gene, called LARGE, is 10 to 50 times bigger than other genes. The gene still functioned, but why did so many people have this variation?

It turned out that the role of the LARGE gene was well known, studied by Dr. Michael B.A. Oldstone at Scripps and his colleagues. LARGE modifies a protein on the surface of cells that the virus uses as an entryway. Without LARGE, that group found, Lassa cannot get into cells.

Now that was interesting, Dr. Sabeti thought. Could the little tweak she had found in LARGE among so many Nigerians make it harder for the virus to infect them?

Lassa had been in Nigeria for about 1,000 years. If this gene mutation was protecting people, how would she know? Dr. Sabeti looked at DNA sequences from Sierra Leone, where Lassa entered about 150 years ago. Ten percent had the mutation in the LARGE gene.

Elsewhere in the world, the mutation was unheard-of. This told her that it was likely that the mutation was protective. To confirm her suspicion, she had to get data — cells from people who were exposed to Lassa and fell ill and those with similar exposure who resisted the virus. That way, she could test whether the LARGE mutation was linked to a better outcome. It is a difficult project and still underway, but so far, based on a small set of data, the mutation in LARGE does appear to be protective.

Dr. Sabeti was far from the first to investigate Lassa — a small contingent of researchers had been focusing on the illness for years but without the benefit of rapid genetic sequencing. The disease came to worldwide attention in the 1960s, said Dr. Joseph B. McCormick of the University of Texas School of Public Health in Brownsville, when some American missionaries became sick and died.

Lassa, Dr. Sabeti said, “likely kills tens to hundreds of thousands of people every year.” She is concerned — the virus is spread by mouse urine and outbreaks occur from winter until spring, when the mice enter homes. “Everyone is so myopically focused on Ebola,” she said, that they are not testing for Lassa and other infectious diseases in many places. The challenge with Lassa and Ebola,

though, was to follow the spread of the viruses in real time. And that meant finding a quick and accurate way to get the genetic sequences of Lassa and Ebola viruses from samples of blood.

It took the Broad group five years to develop such a viral blood test — there is very little virus in blood samples; the blood often has been stored under less-than-ideal conditions in tropical heat; and before a sample can be examined, the researchers have to kill any viruses in it so it does not infect laboratory workers. But the chemicals that kill the viruses make it even harder to fish out the virus.

"The method really came together in the past year," Dr. Sabeti said. Now the group is starting to study how variations in the genetic sequences of the viruses affect the course of infection. And they are asking how quickly and easily the viruses spread by tracing the genetic footprints of the viral strains.

"There are hundreds of mutations evolving in individuals," she added. "We can see the new mutations emerging, and it helps us understand transmission." She said the work could also help with the development of methods of diagnosing the diseases as well as work on vaccines and treatments.

There should be practical payoffs, too. People who come to clinics ill with fevers, diarrhea or vomiting could receive an accurate diagnosis. Many clinics send blood samples to labs to test for Ebola, but those with Lassa have just been sent away, told that what they had was "not Ebola," which does not help much.

"I proposed several years ago to do a genetic study with Ebola," Dr. Pardis said. But it was infeasible: There were too few patients. The situation, unfortunately, has changed.

A version of this article appears in print on December 2, 2014, on page D1 of the New York edition with the headline: The Virus Detectives.

Name_____ **Date**_____

Directions

Read **The Virus Detectives: Sifting Through Genes in Search of Answers on Ebola**, a *New York Times* article highlighting the Sabeti Lab at the Broad Institute and their efforts to sequence the Ebola virus. Then answer the following questions.

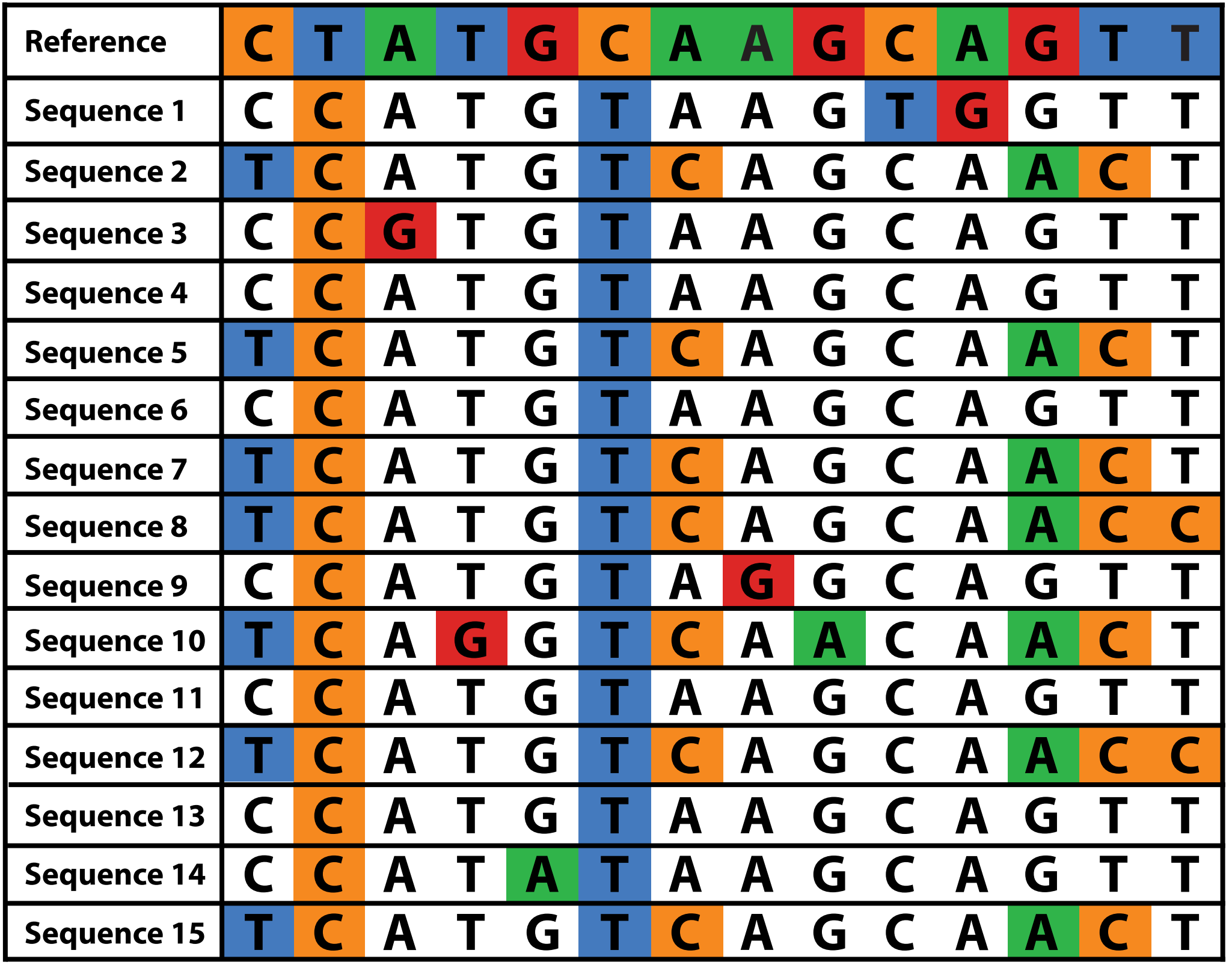
Questions

1. What are some of the questions the Sabeti Lab would like to answer by studying Ebola? Why would this information be useful during outbreaks in the future?

2. Why is it important to examine the genetics of the Ebola virus?

3. What did scientists learn about Lassa virus by studying a gene called LARGE in humans?

4. Describe one of the biggest challenges facing scientists who are trying to follow the spread of a virus in real time.


Reference C T A T G C A A G C A G T T
Sequence 1 C C A T G T A A G T G G T T
Sequence 2 T C A T G T C A G C A A C T
Sequence 3 C C G T G T A A G C A G T T
Sequence 4 C C A T G T A A G C A G T T
Sequence 5 T C A T G T C A G C A A C T
Sequence 6 C C A T G T A A G C A G T T
Sequence 7 T C A T G T C A G C A A C T
Sequence 8 T C A T G T C A G C A A C C
Sequence 9 C C A T G T A G G C A G T T
Sequence 10 T C A G G T C A A C A A C T
Sequence 11 C C A T G T A A G C A G T T
Sequence 12 T C A T G T C A G C A A C C
Sequence 13 C C A T G T A A G C A G T T
Sequence 14 C C A T A T A A G C A G T T
Sequence 15 T C A T G T C A G C A A C T

1. Find the National Center for Biotechnology Information website (NCBI)

- Navigate to the NCBI website: http://www.ncbi.nlm.nih.gov/
  - NCBI is a massive, publicly available database containing many types of data, including sequence data for Ebolavirus and other viruses.

2. Find Ebola sequences

- Search *Zaire ebolavirus*. On the next screen, scroll down to the *Genomes* heading and click on *Nucleotide* on the left.
  - This should take you to a results page with all of the sequenced Ebolavirus genomes. For this lesson, we only want to analyze a subset of these sequences from the 2014 outbreak.
- On the left hand side under *Release Date,* click on *Custom range*.
- In the first three blank boxes, replace *YYYY* with *2014*, *MM* with *03* and *DD* with *01.* In the last three blank boxes, replace *YYYY* with *2014*, *MM* with *07* and *DD* with *20*.
- Click *Apply*. There should now be 18 results. These are the first 18 sequences from the 2014 Ebola outbreak that were published by scientists.
- 15 of these sequences have *SLE* in their title, which refers to Sierra Leone where the samples resulting in these sequences were collected. The remaining 3 sequences have the country code *GIN*, which refers to Guinea where the outbreak started.

3. Download 16 Ebola sequences

- Check the boxes next to the 15 sequences with *SLE* in the title.
- From of the remaining 3 sequences, check the box next to the one with *C15* in the title. This is the first sequence from Guinea published in 2014, and will be used as a reference sequence, which means you are going to compare all of the other sequences to it.
- At the top right hand corner of the page, click *Send to*.
- From the drop-down menu under *Choose Destination*, select *File.*
- Under the *Format* heading, change *Summary* to *FASTA*.
  - *FASTA* is a universally used format for storing sequence data.
- Click *Create File*. Your sequences should now be downloading into a single file called *sequence.fasta*. Save this file to your desktop or another location on your computer that you can easily navigate to.

1. Open Jalview

- Open Jalview on your computer.
- When the program opens, several windows will pop up showing examples of what you can do with various datasets. Close all of them.
- Drag your *sequence.fasta* file into the main Jalview window.
- You should now see 16 rows of Ebola sequences that you previously downloaded from the NCBI website. On the left you will see that 15 of the sequences are labeled starting with *gil66* and 1 sequence is labeled starting with *gil67.* The *gil67* sequence is your reference sequence from Guinea, while the rest are from patients in Sierra Leone.

2. Sequence Alignment

- Each row represents a single sequence, and each column is a DNA letter from the sequence. Because the sequences are from the same type of virus, they should look almost identical to each other. However, you'll notice that in each column most of the letters are not the same. This is because sequencing technologies have trouble finding the letters at the beginning and end of sequences. We will use a special program called Tcoffee to correctly align the columns.
- Navigate to *Web Service —> Alignment —> Run Tcoffee with preset —> Quick align*
  - It will take about 1 minute for the program to align the sequences. A new window will pop up when the alignment is complete. You'll notice the letters are now identical in most of the columns.
  - The dashes you see at the beginning of the sequences have been inserted by the alignment program.

3. Identify Mutations

- Your goal is to find mutations in the sequences and use this information to determine which sequences are most similar to each other.
- Make each DNA letter a specific color:
  - Navigate to *Colour —> Nucleotide*
- Uncolor the columns where all of the letters are the same in order to just highlight the columns where a mutation has occurred:
  - Navigate to *Colour —> by Annotation…*
  - Check the box that says *Use Original Colours*
  - Change *No Threshold* to *Below Threshold*
  - Change *50.0* to *99.9* and hit enter (the scroll bar should automatically slide to the right)
  - Click on *OK*

- Scroll to the right until you see a large block of colored columns, where one of the sequences has the letter *N* in multiple columns. The N means that this region of the genome was not sequenced properly.
- Delete all columns containing a dash, an N, or those that are not colored.
  - Scroll to the beginning of your sequences.
  - To select a block of adjacent columns to delete, click on the first column, hold shift, and click on the last column you want to delete in the block.
  - Hit *delete* on your keyboard and click *OK* when the warning message appears. Make sure to delete the **entire column**, otherwise you will need to re-align your sequences.
  - Repeat this process until you are left with just 14 colored columns that show the mutations.

4. Analyze Sequences

- On the left side of the window, click on the name of one of the sequences.
- Use your up and down arrow keys on your keyboard to move the sequence up or down.
- See if you can find patterns that might tell you about the spread of the virus. For example, in the first column all of the sequences have either a C or T. Grouping all of the sequences with C together will help you visualize which sequences are most similar to each other at this particular position in the genome.
- Keep re-arranging the sequences to find similarities and patterns.

Name_____ **Date**_____

Directions

- Move the Ebola sequences around to compare them to each other and find patterns in the sequences.
- Group similar sequences together, and use the patterns you find to answer the questions below.

Things to Remember

- The **reference sequence** is from Guinea. It is from the beginning of the Ebola outbreak and is several months older than the rest of the sequences which were taken from patients in Sierra Leone.
- Mutations are random and accumulate over time, so sequences with a larger number of mutations when compared to the reference sequence are from later in the outbreak.
- Every sequence should be in a group, even if there is no identical sequence. A sequence could also be placed in its own group.

Questions

1. How many groups of sequences did you create?

2. What attributes did you use to divide the sequences into groups?

3. How are the Ebola sequences (or groups of sequences) related to each other?

4. Create a visual that highlights the relationship between your groups.

5. Can you think of another way that tracking mutations over time might be useful?

Name_____ **Date**_____

Directions

- Move the Ebola sequences around to compare them to each other and find patterns in the sequences.
- Group similar sequences together, and use the patterns you find to answer the questions below.

Things to Remember

- The **reference sequence** is from Guinea. It is from the beginning of the Ebola outbreak and is several months older than the rest of the sequences which were taken from patients in Sierra Leone.
- Mutations are random and accumulate over time, so sequences with a larger number of mutations when compared to the reference sequence are from later in the outbreak.
- Every sequence should be placed in a group, even if does not have an identical sequence. A sequence could also be placed in its own group.

Questions

1. **How many groups of sequences did you create?**
   - Ask students to show with their fingers how many groups of sequences they created. Answers could vary from 2-8.
   - Prompt students with the following questions:
     - Why did you choose that number of groups?
     - Would it be useful to have just one group with all 15 sequences?
     - What about 15 groups that each have one sequence?
   - Explain to the students that there is no real correct answer to this question. Broad scientists went through several iterations of how many groups to make when analyzing this same data.

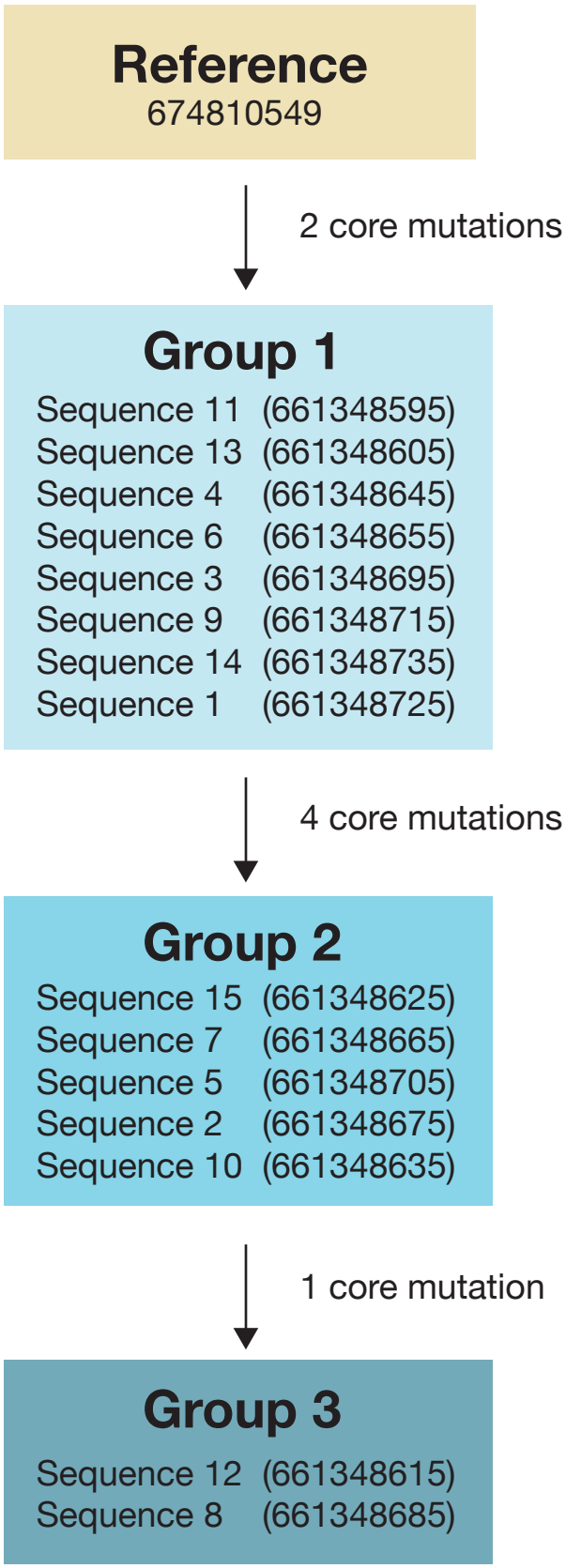
2. **What attributes did you use to divide the sequences into groups?**
   - Students could use one or more of the following attributes:
     - Identical sequences
     - Sequences that only differ by 1 mutation
     - Location of the mutations
     - Number of mutations

3. **How are the Ebola sequences (or groups of sequences) related to each other?**
    - The answer you are looking for is time. The groups give a clear picture of how the sequences are mutating and evolving over time.
    - Prompt students with the following question:
        - If a sequence has a larger number of mutations when compared to the reference sequence, does that mean it is from earlier or later in the outbreak? *(answer is later)*
    - Another implication of these grouped sequences is that they imply that viruses transmitted from one person to another are related by their sequence. In other words, one person gets sick, the virus replicates and mutates, and then several days later a second person is infected with a virus that has a similar sequence. This is a powerful way to track the spread of the virus geographically.

4. **Create a visual that highlights the relationship between your groups.**
    - An effective visual should show how the groups are related to each other over time. Some sort of flow chart or tree will serve this purpose. See the scientist visual for an example.

5. **Can you think of another way that tracking mutations over time might be useful?**
    - One possible answer is that scientists track how the influenza virus mutates in order to develop an effective flu vaccine each year.

Scientist Visual



Explanation. In this visual, sequence numbers 1-15 correspond to the labeling used on the paper strips. The numbers in parenthesis correspond to the sequence labeling used in Jalview.

Sequences are organized into three primary groups of nearly identical sequences. Core mutations refer to mutations that propogate into higher-level groups. For example, in Group 1 all sequences have the same two core mutations when compared to the reference sequence—C in column 2 and T in column 5. These mutations continue to be present in all Group 2 and Group 3 sequences. Although sequences 3, 9, 14, and 1 each have additional unique mutations, these do not propogate into other groups.

Sequences in Group 2 have four additional core mutations—T in colum 1, C in column 7, A in column 12, and C in column 13. These mutations continue to be present in Group 3 sequences. Although sequence 10 has two additional unique mutations, these do not propogate into Group 3.

Sequences in Group 3 have one additional core mutation—C in column 14. The sequences in this group are identical.

Grouping sequences in this manner clearly portrays that mutations are accumulating over time, with the arrows representing the evolution of the virus over time.