Title: Understanding COVID phylogeny using Nextstrain
Emily Jane McTavish

Modified from:
Baylee Blackburn, Andrea DePouw, Alexandra Beanblossom, Jennifer Ness, and Keith A. Johnson (Department of Biology, Bradley University, Peoria, IL 61625; Contact kajohnso@bradley.edu)

Adaptation to: **Understanding COVID-19 Biology to Design a Vaccine by Johnson, Vardar-Ulu and Dutta (2020), Doi: 10.25334/DNC7-8581**

Worksheet Nextstrain - Visualizing clade relationships between SARS-CoV-2 Viral sequences

Learning Objective:
- This worksheet introduces students to the Nextstrain website and its use as a tool for analyzing SARS-CoV-2 sequences, variants and their phylogeny.

Learning Goals: Students should be able to
- Summarize the role of clades in genomic analysis
- Examine a clade and identify the relationships between different branches
- Relate variants (mutants) to branches on a clade

Reading (before proceeding):
Read How to interpret the phylogenetic trees (https://docs.nextstrain.org/en/latest/learn/interpret/how-to-read-a-tree.html)
Read What are clades? (https://clades.nextstrain.org/)

Genetic Evolution of SARS-CoV-2 Virus

The evolution of viruses occurs as a result of changes in the genetic makeup of the virus. Many of the emerging viral diseases impacting humans are a result of a 'jump' from species that have been long-term hosts of the virus to new host species. The spread of the SARS-CoV-2 virus (causing the COVID-19 pandemic) across the world has led to intense research on the virus, the function of the various gene products, regulation of viral attachment and reproduction, and evolution of the RNA genome, as well as the enormous efforts being put forward to thwart the spread of the virus and treat the symptoms of COVID-19. The viral 'jump' is presumed to be from bats, with a likely intermediate host (possibly pangolin). *The actual origin of this virus is still under debate*. Sequence variation (see Box Phylogeny Terminology for highlighted terms) of the virus may influence virulence and may influence the effectiveness of different vaccines.

Nextstrain

Nextstrain is a website that gathers, tracks, and analyzes genome data for a variety of pathogens including West Nile virus, Mumps, Zika, seasonal influenza and SARS-CoV-2. The website provides current data on the evolution of pathogen populations. *There are other websites maintaining SARS-CoV-2 sequences, but you will be using Nextstrain.* Within the Nextstrain website, there is a bioinformatics tool called **Nextclade** which allows the users to upload a FASTA file and perform sequence analysis on SARS-CoV-2 genomes (you will not be doing this in this worksheet). The tool Nextclade performs a pairwise sequence alignment between a reference sequence and the uploaded sequences in the FASTA file. The sequences are assigned to clades based on differences in sequence mutations, and a phylogenetic tree is constructed. Nextclade also houses thousands of SARS-CoV-2 viral sequences from the earliest sequence (considered to be the reference sequence) to the most recent sequences (and is constantly updated).

Listen to the podcast (before class):
Global network of scientists work to track COVID-19's spread (about 7 min).

Box 1. Phylogeny Terminology (some Brooker, Genetics 7th edition)

Sequence variation - mutations that enter a sequence over time
Virulence - the severity of a disease
Mutation - a permanent change in the genetic material that can be passed from cell to cell or, if it occurs in reproductive cells, from parent to offspring.
Clade (monophyletic group) - a group of species consisting of all descendent of the group's most common ancestor
Phylogenetic tree - a diagram that describes a phylogeny and constitutes a hypothesis concerning the evolutionary relationships among different species
Reading frame - a series of codons determined by reading bases in groups of three beginning with the start codon as a frame of reference
Open reading frame (ORF) - a region in a genetic sequence that does not contain stop codons

SARS-CoV-2 Genome

The SARS-CoV-2 viral genome is a positive (+) strand RNA (similar to an mRNA). There are multiple coding sequences within the ~30,000 (actual 29,903) nucleotide genome. These coding sequences encode the spike glycoprotein, capsid proteins, proteases, the RNA-dependent RNA polymerase and other proteins (some of which are of currently unknown function - these are often identified as ORF with a number). Figure 1 shows a schematic RNA genome including some of the coding sequences.

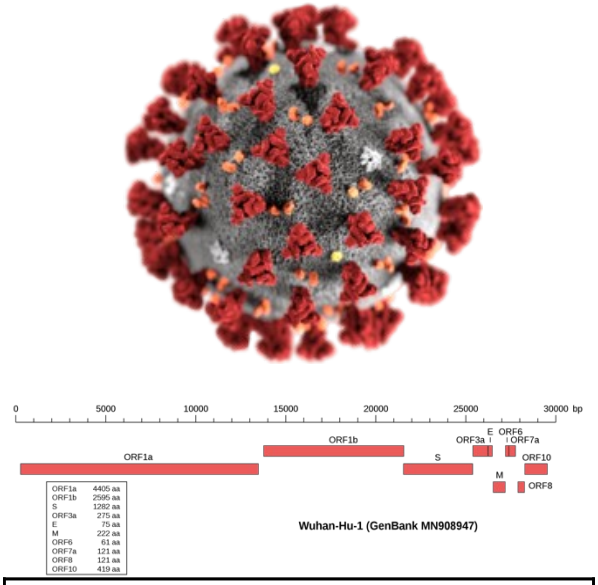

Figure 1. The figure above shows a schematic of the SARS-CoV-2 viral structure, a genome schematic (genome composition in bases) and the polypeptides that are made from the positive strand RNA genome (image from Wikipedia).

Part 1. Looking at the accumulated data

1. Go to Nextstrain/ncov/global.
2. You should see an image similar to Figure 2.

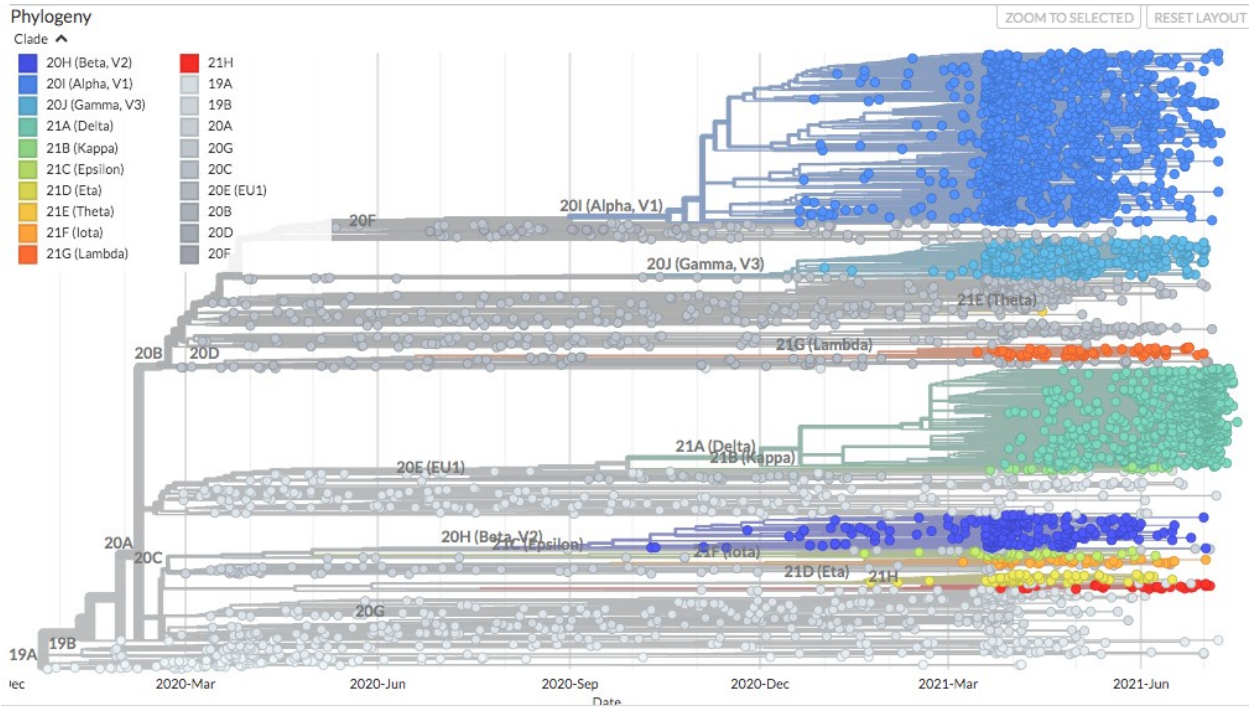

Figure 2. Nextclade showing SARS-CoV-2 clades looking at the approximately 30,000 base pair genome. (Image capture from https://nextstrain.org/ncov/global on July 26, 2021).

Tree thinking,:

1) What date was the first sampled SARS CoV lineage included in this tree sampled?

2) a)What is the approximate estimated time of the MRCA of the omicron lineages?

b) Delta?

c) all the sampled SARS CoV lineages?

3) Compare the trees with branch lengths in terms of 'time' and in terms of 'divergence'.

a) What are some lineages with especially high rates of evolution?

b) What is a lineage with an especially slow rate of evolution?

5) Are the omicron lineages more closely related to delta, to kappa, or equally closely related to both?

Part 2 Introduction to Nextclade

1. To get to the Nextclade application, go to: https://nextstrain.org/sars-cov-2/ and navigate (scroll down) to the Nextclade (sequence analysis webapp).

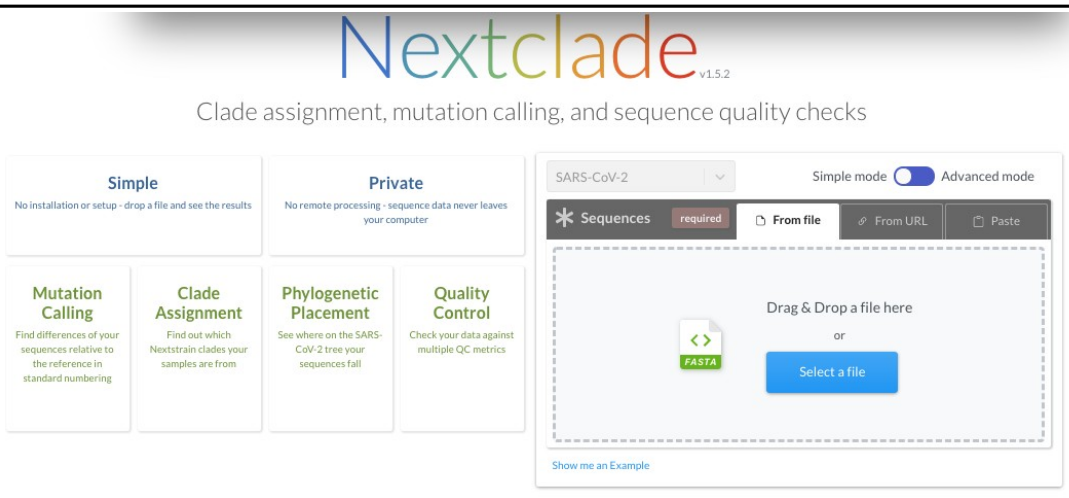

Figure 2. Nextclade (v1.5.2) main page.

2. Scroll down past the initial window shown above to read more about the site before moving on. You should be able to see a current phylogenetic tree of the viral genome, similar to Figure 3.

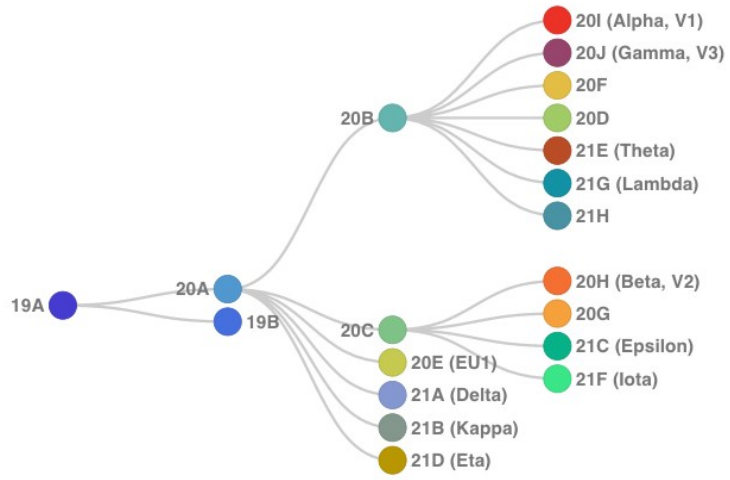

Figure 3. Illustration of phylogenetic relationship of clades, as defined by Nextstrain. (https://clades.nextstrain.org/). Image downloaded July 23, 2021.

3. Scroll back to the top of the webpage and click on "Load example" (Figure 2). You will see example sequences pasted into the box. This example represents many (but not all) of the SARS-CoV-2 viral sequences.

There are many SARS-CoV-2 sequences shown in this figure. An alignment of these sequences can be seen in the Sequence view window on the right. A schematic of the coding regions of the SARS-CoV-2 genome is shown on the bottom right.

Click on the box at the top to see the whole nucleotide sequence.

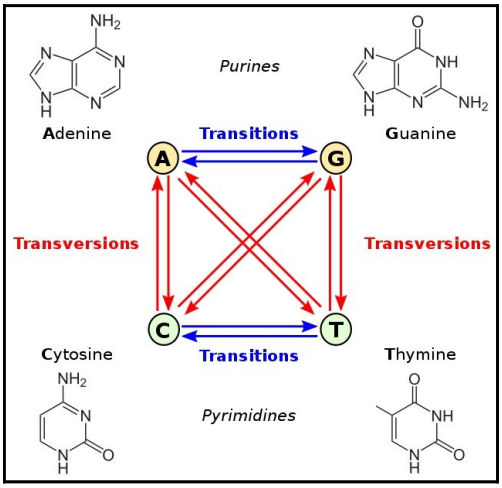

Figure 4. Transition and transversion mutations (image from Wikipedia)

4. Sort the example files by clicking on the down arrow in the Clade column.

5. Using your mouse or touchpad, scroll over the vertical lines in the Sequence view window.
    a. What are three mutations that are shared with the 20 clade that are not found in the 19 clade? Record the letter-number-letter of the mutation.

    b. What does the number between the letters represent?

6. The omicron lineage has many shared mutations in one region of the genome. What region is that? What does that gene do?

7. What sampled genome has the fewest differences from the reference? Is that strong evidence that it is closely related to the reference? Why or why not?

8. Find a mutation that appears to have arisen independently in more than one clade. Check to see if you are right going back to nextstrain and coloring the tree by the genotype at that nucleotide. Screenshot your tree including the repeated mutation and paste it here.