**Title: Understanding COVID – Nextstrain and variants**

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.
* Students are also introduced to the variants of concern and interest and how these mutations can impact the structure and binding of the SARS-CoV-2 virus to host cells.

**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
* Articulate the origin of variants and their potential impact

**Reading** (before proceeding):

Read [How to interpret the phylogenetic trees](https://nextstrain.org/help/general/how-to-read-a-tree)

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](https://nextstrain.org/) 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](https://nextstrain.org/) 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).

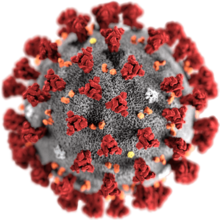
Listen to the podcast (before class):

[**Global network of scientists work to track COVID-19’s spread**](https://www.pri.org/file/2020-06-26/global-network-scientists-work-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.



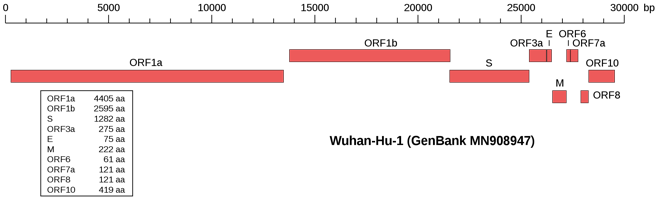


Figure 1. The figure above shows a schematic of the SARS-CoV-2 viral structure, a genome schematic (genome composition in bases) and the polypeptides that are made from the positive strand RNA genome (image from Wikipedia).

**Part 1. 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).

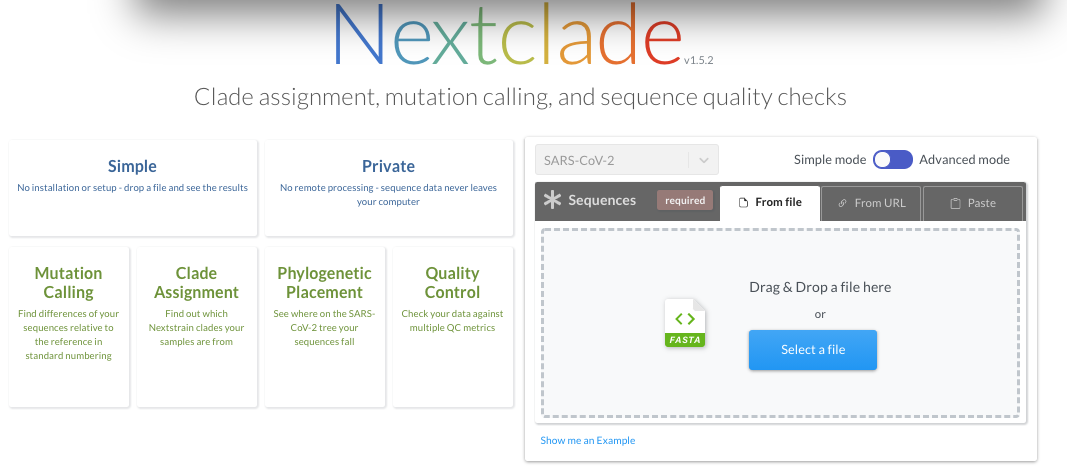


Figure 2. Nextclade (v1.5.2) main page.

1. 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.

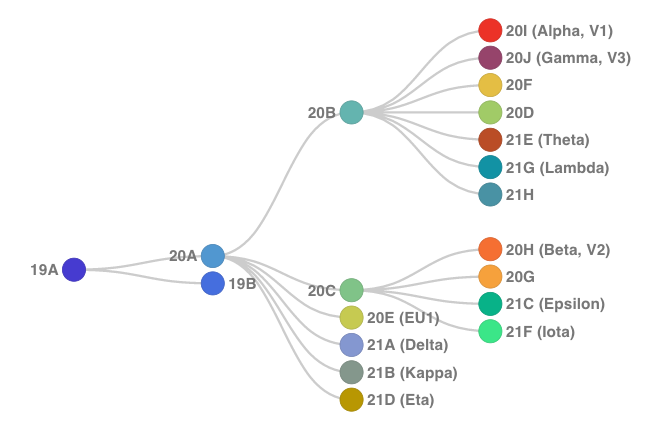


Figure 3. Illustration of phylogenetic relationship of clades, as defined by Nextstrain. (<https://clades.nextstrain.org/>). Image downloaded July 23, 2021.

1. Scroll back to the top of the webpage and click on “show me an 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.

1. The colored vertical lines under Sequence view (right) show the positions of mutations relative to a SARS-CoV-2 reference genome. What does each color represent? *Hint: Scroll over the different colors to answer this question*.

|  |  |
| --- | --- |
| Color | Represents (exclude #s) |
| Green |  |
| Blue |  |
| Red |  |
| Gold |  |
| Gray |  |
| Black |  |

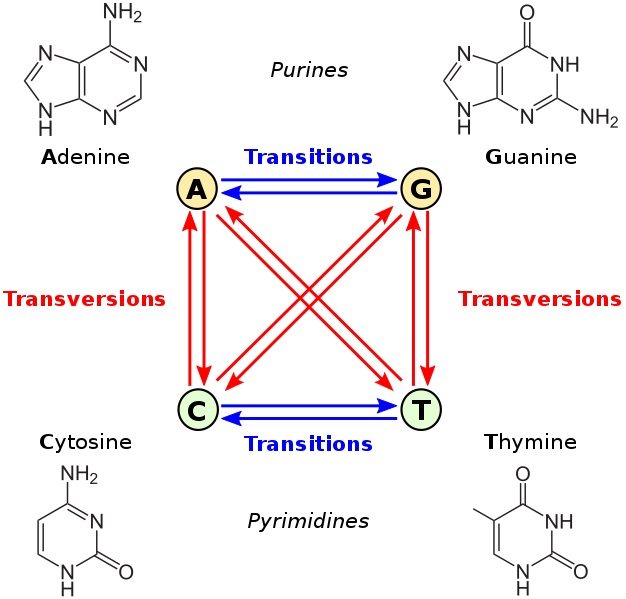


Figure 4. Transition and transversion mutations (image from Wikipedia)

These mutations each represent what is classified as transition mutations (see Figure 4).

1. Sort the example files by clicking on the down arrow in the Clade column.
2. Using your mouse or touchpad, scroll over the vertical lines in the Sequence view window.
   1. 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.

* 1. What does the number between the letters represent?

**Part 2. Looking at the accumulated data**

For the next analysis, you will look at the occurrence of any mutations that could impact the amino acids purportedly involved in ACE-2 interactions that we have been focused on in the previous worksheets from Understanding Covid or see figure below.

1. Go to [Nextstrain/ncov/global](https://nextstrain.org/ncov/global).
2. You should see an image similar to Figure 2.

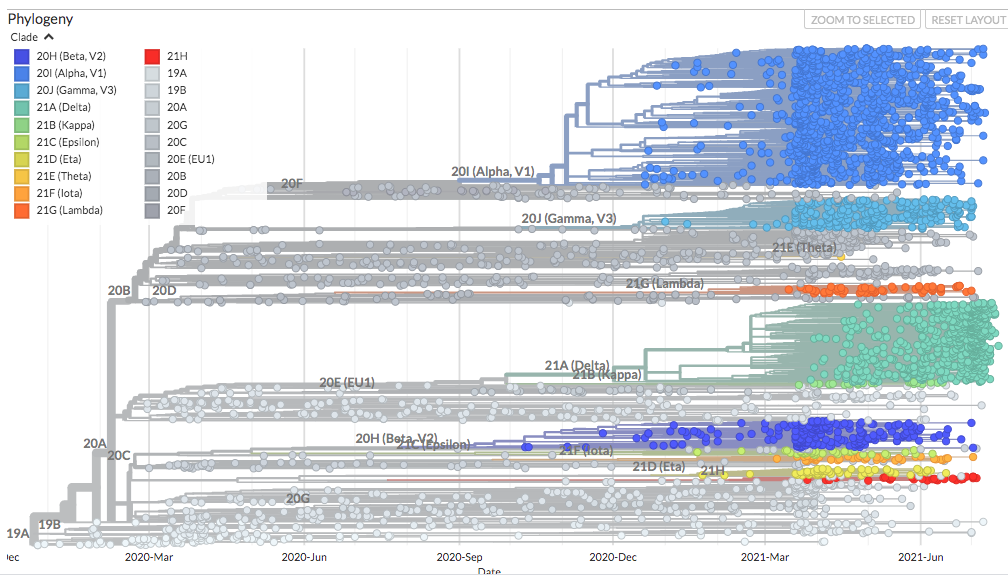


Figure 2. Nextclade showing SARS-CoV-2 clades looking at the approximately 30,000 base pair genome. (Image capture from <https://nextstrain.org/ncov/global> on July 26, 2021).

This ‘clade’ is looking at all of the sequences. Since we have focused on the S glycoprotein, we will look more at the RBD (receptor binding domain) and specific amino acids.

Figure 3A shows the SARS-CoV-2 Spike (S) glycoprotein receptor binding domain (RBD) interacting with the human ACE-2 protein, the site of viral attachment. Figure 3B shows specific chemical interactions between the protein domains. The next steps in the worksheet will focus on the interacting amino acids indicated in Figure 3C.

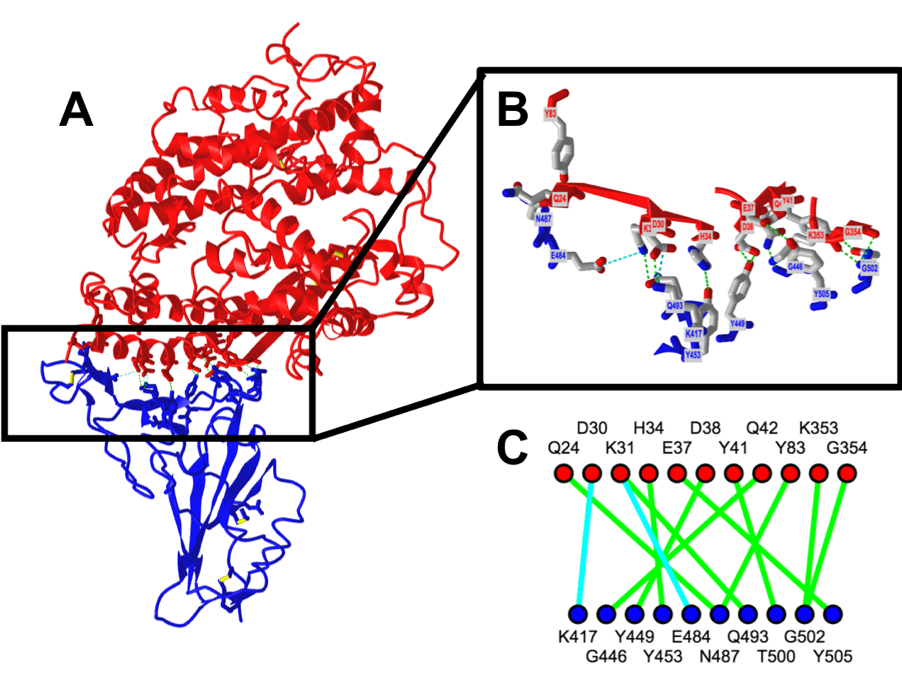
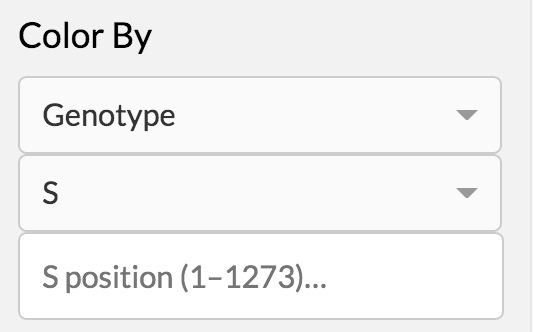


Figure 3. A. SARS-CoV-2 receptor binding domain (blue) interacting with ACE-2 protein (red) (PDB 6M0J). B. The interacting amino acids are shown and labeled in the box. C. Hydrogen (green) and ionic bonds (cyan) interactions between the ACE-2 protein (red) and SARS-CoV-2 receptor binding domain (blue) are shown in 2D. Image was generated using iCn3D protein visualization software.

1. You can focus on the specific Spike (S) glycoprotein amino acids involved in the chemical interaction with the ACE-2 protein (Figure 3C). On the left hand side of the [Nextclade window](https://nextstrain.org/ncov/gisaid/global), in “Color by”, scroll down and select “Genotype” and then select “S” for the gene.



1. In the box labeled “S position (1-1273)...” enter a number (not the amino acid letters) corresponding to one of the amino acids previously investigated. Hit return.
   1. Once you have entered the S protein amino acid position, hit return and look for a key in the upper left hand corner of the window. *Note: not all of the amino acids in the table may have been reported to be mutated.*
   2. With each S protein position, record in the table below if there have been any reported mutations that affect these amino acids.

|  |  |
| --- | --- |
| **Residue location within the full length SARS-CoV-2 S protein**  **(from Figure 3C)** | **Any changes (if changes, to what amino acid was the change?)**  ***See “Phylogeny Genotype at S site” # in the upper left corner of the image.*** |
| K417 |  |
| G446 |  |
| Y449 |  |
| Y453 |  |
| E484 |  |
| N487 |  |
| Q493 |  |
| T500 |  |
| G502 |  |
| Y505 |  |

You can also look at any of the other encoded genes in the SARS-CoV-2 virus to see where other mutations have occurred.

**Part 3. Understanding new mutations**

In the previous section, you looked at amino acids that were identified by the structure to view if they have been altered since the first sequence of SARS-CoV-2. Several of these amino acids have been altered and this may impact the binding of the S glycoprotein to the ACE2 protein. Scientists are looking more closely at these, but are also looking for the occurrence of other mutations.

Several mutations that have arisen have potential implications in the viral infectivity. It is unclear if these variants might have other impacts (e.g., are they still targets of the vaccines?).

One of the mutations appears to outcompete the original virus (19A and 19B). This mutation (D614G – aspartic acid at S glycoprotein position 614 altered to glycine) first occurred in March of 2020 and has dominated the clade, suggesting that it outcompetes the in the real word (we will see this by the clade. Scientists have confirmed this in the laboratory.

* + - 1. Using the steps in Part 2 of this case, create the clade showing the D614G mutation.
         1. This mutation (D614G) does not occur in the region of interaction between the spike (S) glycoprotein and the ACE2 protein. The mutation appears to ‘open’ up the spike glycoprotein, which allows it to bind and infect more efficiently.
         2. Include an image of your clade.
         3. What effect has the D614G mutation had on the occurrence of the SARS-CoV-2 virus?

The WHO (World Health Organization) has tried in the summer of 2021 to consolidate the naming of SARS-CoV-2 variants and use Greek letters. The WHO has also created designations of Variants of Concern (Table 1) and Variants of Interest (Table 2). The WHO names have been added to the Nextclade analysis.

Table 1: WHO designated Variants of Concern in July 2021.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| WHO Label | Pango lineages | Nextstrain Clade | Additional amino acid changes | Earliest documented samples | Date of designation |
| Alpha | B.1.1.7 | 20I (V1) | +S: 484K  +S: 452R | United Kingdom (September 20) | December 2020 |
| Beta | B.1.351 | 20H (V2) | +S:L18F | South Africa (May 2020) | December 2020 |
| Gamma | P.1 | 20J (V3) | +S:681H | Brazil (November 2020) | January 2021 |
| Delta | B.1.617.2  AY.1 | 21A | +S: 417N | India (October 2020) | April-May 2021 |

Data was taken from <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>. GISAID clade was omitted but is unique from those identifiers in this table.

Table 2: WHO designated Variants of Interest in July 2021.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| WHO Label | Pango lineages | Nextstrain Clade | Earliest documented samples | Date of designation |
| Eta | B.1.525 | 21D | Multiple countries (December 2020) | March 2021 |
| Iota | B.1.526 | 21F | United States of America (November 2020) | March 2021 |
| Kappa | B.1.617.1 | 21B | India (October 2020) | April 2021 |
| Lambda | C.37 | 21G | Peru (December 2020) | June 2021 |

Data was taken from <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>. GISAID clade was omitted but is unique from those identifiers in this table.

While some of the amino acids identified as altered in the Variants of Concern are not located in the spike glycoprotein receptor binding domain for ACE-2 (shown in Figure 3), these amino acid changes are significant in making the virus either bind better/more efficiently or reproduce better or some other competitive advantage.

* + - 1. Using [Nextclade global analysis](https://nextstrain.org/ncov/gisaid/global), make some observations based on a clade of the current global data. Similar sequences are grouped and colored. Scroll over the dots (right hand side of clade) and note the origin of the sequence sample (should be part of the first name).
         1. Scroll over to the left hand side of the clade (older viral sequences) and determine the region that the sample was isolated from.
         2. How do those (older) samples compare with the origins of the newer (more right hand) samples in origin?

**Part 4. Exploring Uniprot**

The website [Uniprot.org](https://www.uniprot.org/) maintains a tremendous amount of information pertaining to proteins, including predicted cellular location and known mutations (and much more).

Go to the [Uniprot](https://www.uniprot.org/) website and search for ‘Spike glycoprotein’.

In the list that is returned, under ‘Organism’ identify the spike glycoprotein from SARS-CoV-2 and click on the link under ‘Entry’.

Looking at the resulting link, gather some information from the resource:

|  |  |
| --- | --- |
| UNIPROTKB identifier |  |
| GO\* Molecular Function (list first item) |  |
| GO\* Biological Process (list first item) |  |
| Subcellular location |  |
| Region – receptor binding domain of protein |  |

\*GO stands for Gene Ontology that provides information regarding the biological process, molecular function and cellular component (subcellular location) for the protein.

Scroll further down to the section indicating Natural variants. These are some of the variants that have been identified throughout the pandemic.

Using the information from Table 2 (Variants of Interest), determine the mutations that have been identified in the Eta and Iota virus clades. *Hint: Do a search for the Pango lineage from those clades – there may be more than one variation, list all).*

|  |  |  |
| --- | --- | --- |
| WHO Variant of Interest | Pango Lineage | Amino acid variations |
| Eta |  |  |
| Iota |  |  |

* + - * 1. Which amino acid variations are in the Spike glycoprotein receptor binding domain (circle your answer(s) from the table you created above).

**Assessment**

Based on the information that you have been able to gain from Nextstrain and Uniprot, what value do you see in being able to collect this data in nearly real time? How might this information be useful in further understanding the SARS-CoV-2 virus and the COVID-19 pandemic?