Look at the emergence of variants using NextStrain.org & identify mutations

NextStrain.org is an open-source research project that provides continually updated views of pathogen evolution and spread. A worldwide consortium of scientists and programmers collaborate on the NextStrain platform to create interactive visualizations for exploring changes in genome sequences over time.

This project provides a short introduction to NextStrain.org and instructions for obtaining a protein sequence from a viral sample.

1. Go to NextStrain.org.

You may want to scroll down and read about the philosophy of this website and why the scientists that developed NextStrain believe in the importance of monitoring changes in pathogen genomes.

1. Look for SARS-CoV-2 (COVID-19) and click “Latest global analysis - open data.”

Many if not all the samples are linked to sequence data at the National Center for Biotechnology Information (NCBI).

Part A. A quick tour of NextStrain.org

1. Phylogeny

You will see an interactive phylogenetic tree that represents the real time evolution of SARS-CoV-2. This tree is derived from genome sequences contributed by scientists around the world.

Each branch of the tree represents a new mutation. Time is shown on the X-axis. The tree is divided into groups of related sequences called “clades.” Clades are groups of sequences that share a common origin. When a group of viruses reaches a global frequency of at least 20%, they are designated a “major” clade. Major clades are named based on the next available letter in the Greek alphabet and the year they emerged. The panel in the top corner, above the phylogenetic tree, under “Phylogeny” shows the names of major clades and their corresponding colors.

If you hold your cursor over a branch of the tree, you will see the mutations that are found in that branch. Each circle represents a sample of virus that was collected from an individual and sequenced. Most of the samples were collected from humans, however it is possible to filter the data and find sequences from other animals, such as minks and hamsters. If you hold your cursor over a circle, you will see more details about that sample.

Timeline

Description automatically generated

1. World map

NextStrain.org also shows a map with pie charts representing the frequency of each clade in different parts of the globe. Genome sequencing is more common in some countries than others, which is why the size of each pie chart varies by country.

Diagram

Description automatically generated

1. Mutation frequency

Scroll downward and you’ll see under “Diversity” a map of the SARS-CoV-2 genome from the 5' end on the left to the 3' end on the right. The values on the Y-axis show the frequency of mutations at each position.

The gene that codes for the spike protein (S) contains the highest frequency of mutations compared to other regions of the virus genome. Some of these mutations make the virus more transmissible. Other changes help the virus escape the immune system. But for many of the mutations an impact on pathogenicity has not yet been discerned.

Timeline

Description automatically generated with medium confidence

1. Frequencies

This graph shows the proportion of the clades at different time points. Clicking one of the vertical lines in the Diversity plot will adjust the Frequency plot to display the proportion of samples that carry the different nucleotides.

A picture containing chart

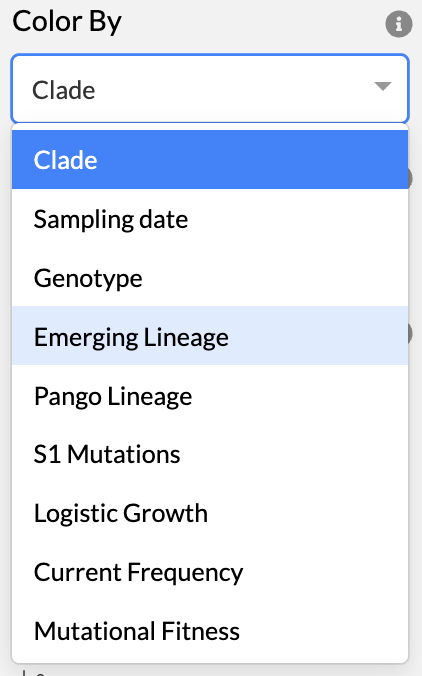
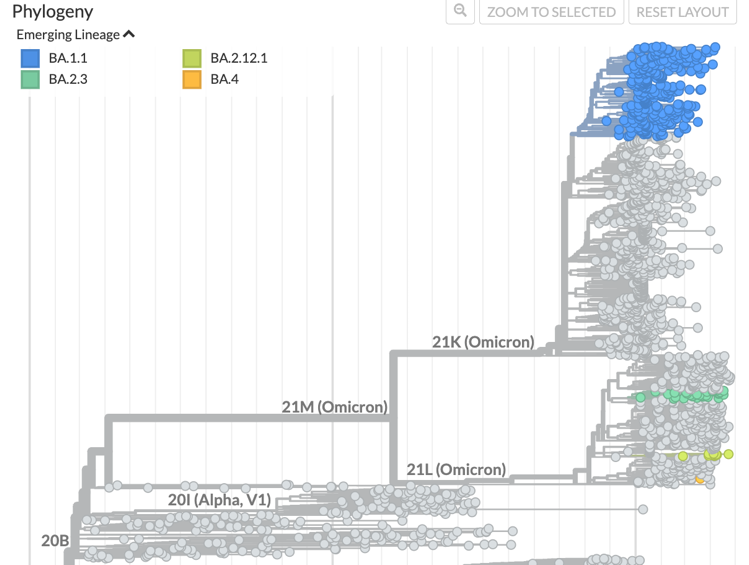
Description automatically generated

Part B. Obtaining a protein sequence from NextStrain.org and NCBI

1. Go to NextStrain.org
2. Find SARS-CoV-2 (COVID-19) and click “Latest global analysis - open data.”
3. Find the phylogenetic tree.
4. Choose a sample to investigate.

You can either select a sample from an older clade or a sample from an emerging lineage. In either case, the protein sequence will help us determine if a viral strain might be able to escape commercial antibodies or the immunity generated by current vaccines.

To investigate an emerging lineage, open the “Color By” menu and choose Emerging Lineage. Emerging lineages will appear in color and other clades will appear grey.

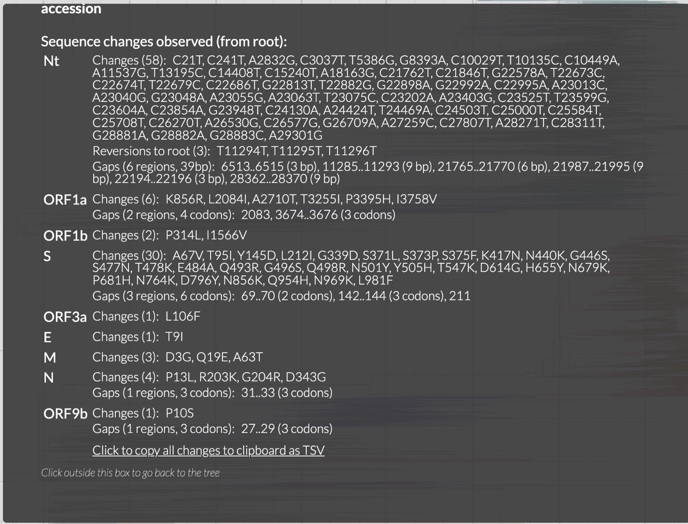


1. Pick a sample to investigate by clicking one of the circles in the tree.

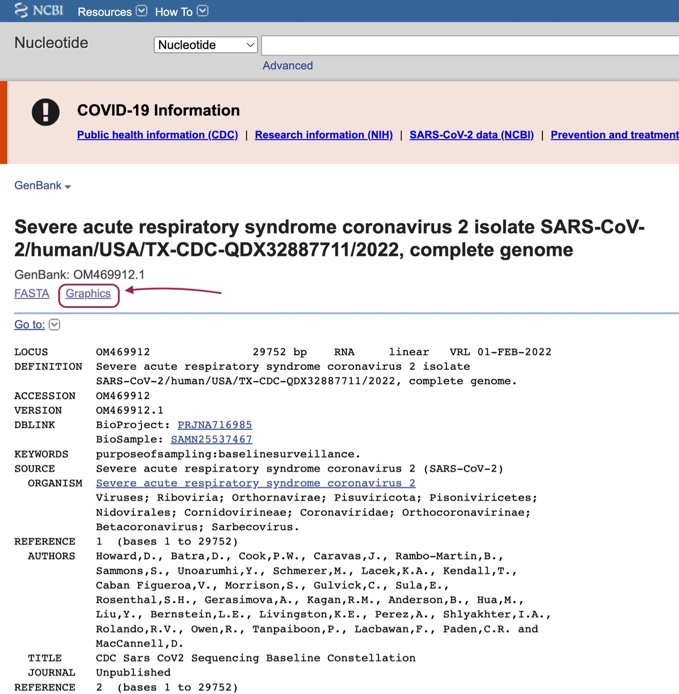
A report will appear with information about that sample such as the date and location where the sample was obtained. You will also see the name of the clade, the emerging lineage (if relevant), and all the mutations that have taken place since the first viral genome was obtained from the Wuhan fish market in 2020. Enter the respective information into the appropriate lines of the associated ‘Next Strain data’ document\*.

The report also contains a link to this sequence in NCBI’s Genbank (Genbank accession).

\*Note: When data are copied from the sample information and pasted into a document, the text will appear white. By changing the font color to black the pasted text becomes visible.

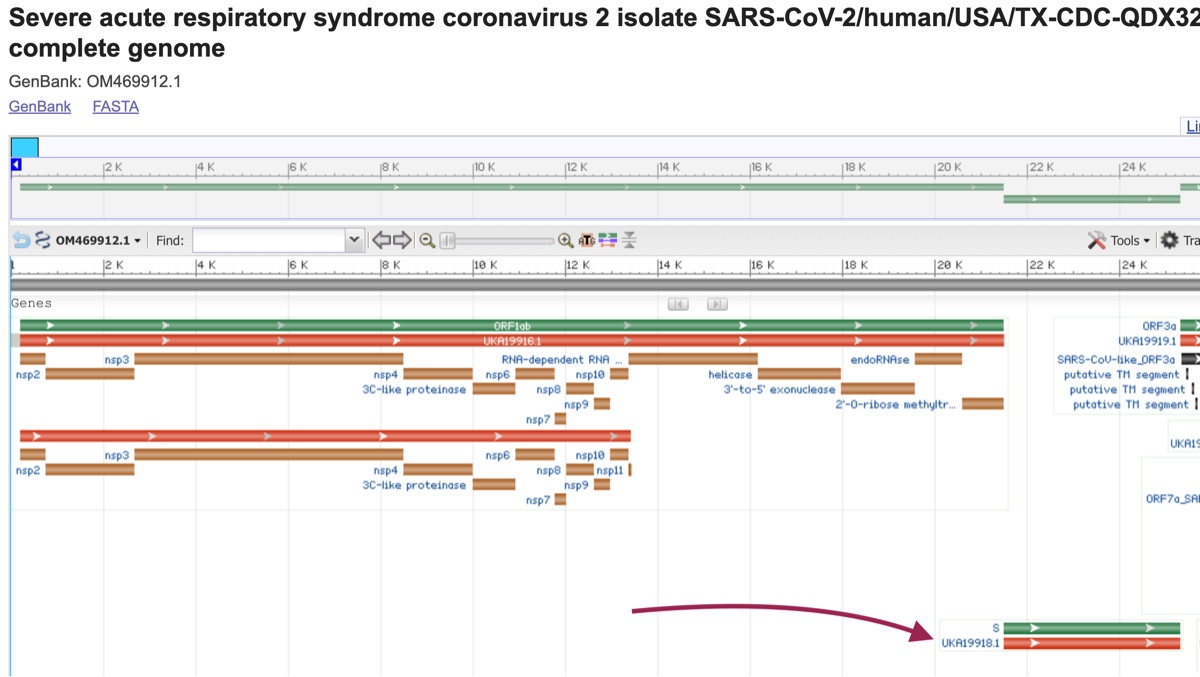


1. Click the Genbank accession number to open the nucleotide sequence record at NCBI.
   1. Find a link to “Graphics” below the title and the accession number.



* 1. Click the Graphics link to view a map of the viral genome.

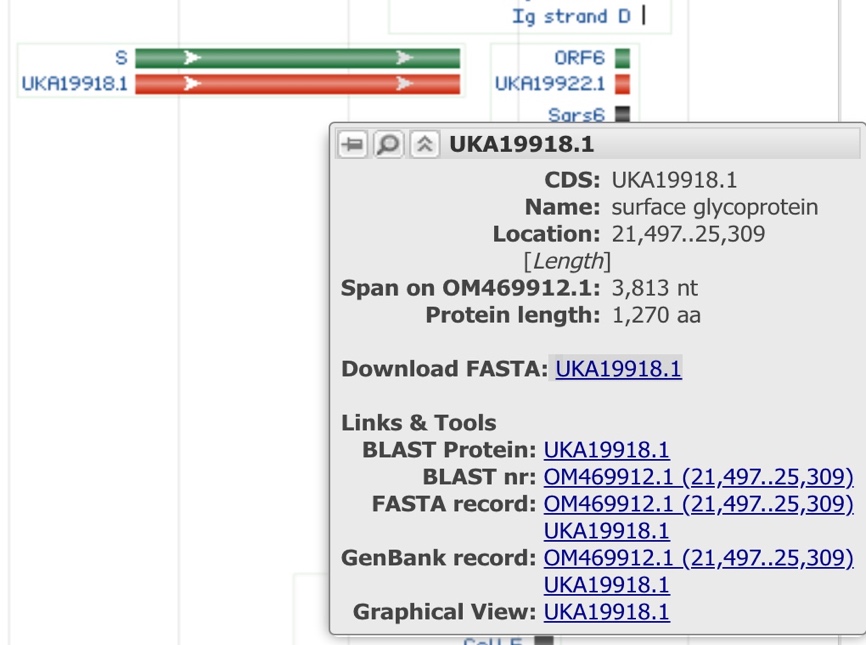
The genome map uses green rectangles to represent genes, red rectangles for proteins, and brown rectangles for proteins after they have been processed during virus maturation.



The arrow in the image above points to the accession number UKA19918.1 for the spike protein (a.k.a. surface glycoprotein) of this sample.

* 1. Find the spike protein in your genome map. (It may have a different accession number.)
  2. Hold your cursor over the accession number for what you believe is the spike protein for your sample. Your spike protein sequence will most likely appear in the same position on your map.

A grey box will appear below the protein.



* 1. Confirm that it is named surface glycoprotein
  2. Drag your cursor over the accession number (CDS) and copy it from this box. (CDS stands for Coding Sequence.)

1. Check the quality of your protein sequence.
   1. Paste the accession number in the search bar below the NIH > National Library of Medicine banner on the top of the page.
   2. From the dropdown to the left of the search bar select “Protein”, then click Search.



1. Examine the amino acid sequence. (You may need to scroll down to see it.)

If the amino acid sequence shows lots of X’s, go back and select another clade to investigate. X’s are used as placeholders for sequence parts that can’t be identified.