NATIONAL CENTER FOR CASE STUDY TEACHING IN SCIENCE

Fatally Flawed? Early Genetic Testing for the COVID-19 Virus

by
Amy C. Groth
Department of Biology
Eastern Connecticut State University, Willimantic, CT

Part I – Viruses and Genetic Tests

The Problem

Stop touching your face! Dr. Brown silently admonished herself when she realized she was nervously biting her nails in her office at the state clinical laboratory. It was early February, 2020, and COVID-19, the novel coronavirus disease, had recently arrived in the United States. Experts had recommended that people wash their hands frequently and avoid touching their face to prevent the spread of the virus. Dr. Brown gazed out the window at the lights of the skyline and wondered how hard this virus would hit her city. How many cases would soon be flooding the hospital system? And more importantly, how many people out there were already infected, walking the streets, and unknowingly spreading the virus?

Two people who had recently traveled had just been hospitalized in her city with suspected COVID-19, and she was desperate to run a genetic test to determine whether they did have the SARS-CoV-2 virus and to begin contact tracing and community surveillance testing to stop the spread of this virus while it was still possible.

The problem was, there hadn't been a test available for labs like hers in the United States, until now. The Centers for Disease Control and Prevention (CDC) had conducted all the tests on site and had just sent a limited amount of testing kits out to a number of cities across the country. Her lab had received 50 tests, and the first step was to verify that the testing reagents worked the way they were supposed to. She had spent the day doing extractions and setting up the assays. She'd finally loaded the samples into the machine and forced herself to go back to her office to do paperwork, rather than stand by the machine waiting for the reactions to run.

Now it was 6:15 p.m. and the results should be ready. She hurried down the hallway, fighting the urge to run, calculating in her head how late into the evening she would have to stay to run the patient samples when she had verified that the test worked. As each result showed what she expected, she got more excited, until suddenly her heart dropped. One of the results was not what it was supposed to be. This is what she had been dreading; initial reports from some state labs indicated that the test did not work properly. The CDC had said those faulty tests could not be used and their samples would have to be sent to Atlanta for testing, but a delay of even a few days could be catastrophic.

Dr. Brown stared at the computer screen, lost in thought. She would run the test again, of course, but if the results came back the same, could she (and should she) use the test to at least determine whether those two patients were likely positive or likely negative?

Background

In late 2019, doctors reported a surge in an unidentified respiratory illness in the Wuhan region of China. This illness was attributed to a betacoronavirus, related to the viruses that caused the SARS outbreak in 2002–2004 and the MERS outbreak that began in 2012. (Zhu *et al.*, 2020) Eventually this virus was given the name SARS-CoV-2 and the disease that it caused, COVID-19, went on to rapidly spread across the globe, causing a long-feared global pandemic. In the United States, a lack of access to testing hampered early efforts to contain the spread of the virus. All testing was initially done at the CDC, which had limited capacity and caused a considerable delay in the ascertainment of results. Eventually, in early February, 2020, the CDC sent out limited amounts of test kits to state laboratories, but unfortunately a number of them did not function properly. One problem with many of the kits was a failure of the N3 reagent to yield a positive result (Shear, *et al.*, 2020; Baird, 2020; Whoriskey & Satija, 2020). Laboratories were initially instructed to not use the kits, but were eventually told to use them, without the N3 analysis. Around that same time, many labs were able to develop their own kits and get them approved through the Federal Drug Administration's "Emergency Use Authorization." (Emergency Use Authorization, 2020) However, the inability to perform tests in the early days prevented authorities from identifying community spread and containing the outbreak.

References

Baird, R.P. 2020. What went wrong with coronavirus testing in the U.S. *The New Yorker,* March 16, 2020.

Emergency Use Authorization. 2020. <https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/emergency-use-authorization>.

Shear, M.D., *et al.* 2020. The lost month: how a failure to test blinded the U.S. to COVID-19. *The New York Times* March 28, 2020.

Whoriskey, P., and N. Satija. 2020. How U.S. coronavirus testing stalled: flawed tests, red tape and resistance to using the millions of tests produced by the WHO. *The Washington Post,* March 16, 2020.

Zhu, N., D. Zhang, W. Wang, *et al.* 2020. A novel coronavirus from patients with pneumonia in China, 2019. *The New England Journal of Medicine* 382(8): 727–33. DOI: 10.1056/NEJMoa2001017.

Section A – Viruses and Tests to Detect Them

1. Describe how the three main types of viruses are classified.

2. How are the three different classes of viruses replicated? (Address the template, enzyme, and resulting type of genome in your answer.)

3. What are the average mutation rates for the three different classes of viruses? (*Note:* mutation rate is generally reported as substitutions per nucleotide per cell infection (s/n/c); essentially, how many changes are made at each position in the genome each time the virus is passed on to a new cell.)

4. List the various ways that one could test for the presence of a pathogen in a human clinical sample. (*Hint:* think about the central dogma of molecular biology.)

5. What type of virus is SARS-CoV-2?

6. What types of genetic tests would not work for SARS-CoV-2 and why?

7. One type of test is RT-qPCR. There are two main types of RT-qPCR tests (TaqMan and SYBR Green). Briefly describe how each is conducted (including the steps they have in common) and the advantages of each in the context of a test for a rapidly-spreading and evolving virus.

Section B – Genetic Tests and the Importance of Controls

1. The original TaqMan assay that was deployed by the CDC for SARS-CoV-2 had primer/probe sets to amplify regions specific to SARS-CoV-2 (N1 and N2), a primer/probe set to amplify a region from SARS-CoV-2 and other closely related viruses like the virus that causes SARS (N3), and a primer probe set to amplify a human specific gene (RP). Why was a human gene included?

2. In addition to the clinical sample being tested, researchers would also run the assays on three additional samples: uninfected human cells (human control), RNA provided by the CDC that matched parts of the SARS-CoV-2 genome (virus control) and "samples" containing just water. Why were those controls included (what would they show the researchers, and how would the results of the controls affect the interpretation of the outcome?). Was each considered a positive or a negative control?

3. Why do you think there are three primer/probe sets for SARS-CoV-2?

4. The dilemma that "Dr. Brown" and other researchers faced was a faulty N3 reagent in the CDC kit, meaning the N3 primer/probe set did not consistently yield a positive result with the positive control, although the N1 and N2 sets did, and the rest of the controls worked as expected. If you were Dr. Brown, would you use the kit to test patient samples, and why or why not?

5. Assuming all of the aspects of the test worked perfectly, and the patient did have COVID-19, what would be expected for each assay below (use a + to indicate a positive result, and a – to indicate a negative result).

Clinical Sample				*Human Control*				*Virus Control*				*Water control*			
N1	*N2*	*N3*	*RP*	*N1*	*N2*	*N3*	*RP*	*N1*	*N2*	*N3*	*RP*	*N1*	*N2*	*N3*	*RP*

6. For the following scenarios, designate each as positive, negative, or inconclusive (best, educated guess, with justification).

	Clinical Sample				*Human Control*				*Virus Control*				*Water control*				*Result?*
	N1	*N2*	*N3*	*RP*	*N1*	*N2*	*N3*	*RP*	*N1*	*N2*	*N3*	*RP*	*N1*	*N2*	*N3*	*RP*	
1	–	–	–	–	–	–	–	+	+	+	+	–	–	–	–	–	
2	+	+	+	+	+	+	+	+	+	+	+	–	+	+	+	–	
3	–	–	–	+	–	–	–	+	–	–	–	–	–	–	–	–	
4	+	–	+	+	–	–	–	+	+	+	+	–	–	–	–	–	
5	–	–	–	+	–	–	–	+	+	+	+	–	–	–	–	–	

Justifications:

Part II – Online Databases, Alignments and Mutation Rates

Introduction

During this online activity you will be utilizing a variety of publicly available genetic databases and analysis tools. First you will identify the regions of the SARS-CoV-2 virus that the original primer/probe sets were designed to amplify. Second you will compare the SARS-CoV-2 genome with the SARS-CoV genome (the virus that causes SARS) and determine which regions are most similar and dissimilar. You will then compare some of the sequenced versions of SARS-CoV-2 from patients in the first few months of the pandemic, to make a rough estimate of the mutation rate. Finally, you will determine whether any of those sequences have accrued mutations that might invalidate the established genetic test.

Note: Within the instructions, portions that should be turned in as part of the assignment are shown in **bold** (an answer sheet for you to fill out is provided at the end, including the table).

Section A – Primer/Probe Sequences from the Original TaqMan Assay

One useful genomic database is the UCSC Genome Browser <https://genome.ucsc.edu>. This database contains the complete genomes of a number of different organisms. Although it has not historically focused on viral genomes, it now includes SARS-CoV-2 due to the global importance of this virus to the scientific community. The website also has a variety of tools that can be used to identify and compare sequences. You will utilize the "In-Silico PCR" tool to identify the regions of the viral genome that were targeted in the original test. With In-Silico PCR, the researcher provides the primer sequences and the program identifies the amplified sequence.

1. Navigate to <https://genome.ucsc.edu> and click on In-Silico PCR on the main page. (If the main page has changed, there should be a heading called "Tools" in a ribbon near the top; click there and select In-Silico PCR.) Use the Genome drop-down menu to select SARS-CoV-2 and the Assembly drop-down menu to select Jan2020.
2. For each of the primer probe sets below, paste the forward (F) primer sequence (just the sequence, not the header line) into the Forward Primer box, and the reverse primer (R) sequence into the Reverse Primer box, and click submit.

```
> N1F
GAC CCC AAA ATC AGC GAA AT
>N1R
TCT GGT TAC TGC CAG TTG AAT CTG
>N1Probe
ACC CCG CAT TAC GTT TGG TGG ACC

>N2F
TTA CAA ACA TTG GCC GCA AA
>N2R
GCG CGA CAT TCC GAA GAA
>N2Probe
ACA ATT TGC CCC CAG CGC TTC AG

>N3F
GGG AGC CTT GAA TAC ACC AAA A
>N3R
TGT AGC ACG ATT GCA GCA TTG
>N3Probe
AYC ACA TTG GCA CCC GCA ATC CTG
```

3. The output will have a first line that begins with >, and the sequence that is amplified will be the next two lines below that. **Copy this sequence, and paste it into your answer sheet. For each one, make a header line starting with a > followed by the name of the region (i.e., N1). Identify the probe sequence in bold (you can just compare by eye to identify the probe sequence).** Below is an example for N1. Determine the N2 and N3 sequences. (*Note:* the N3 probe has a "Y" in it, meaning pyrimidine; two different versions of the probe are included in the assay, one with a "C" at that position and one with a "T.")

   >N1
   GACCCCAAAATCAGCGAAATgc**accccgcattacgtttggtggacc**ctCA
   GATTCAACTGGCAGTAACCAGA

4. Once you have all three sequences, you will use another useful Genome Browser tool, a BLAT (BLAST-like alignment tool) search. This allows you to search for a given sequences or sequences within a genome. Under the Tools drop-down menu in the ribbon near the top, navigate to BLAT. Make sure the genome says SARS-CoV-2 and the assembly says Jan2020. Paste all three of your identified sequences into the box, each with >Header, followed by the sequence on the next line, followed by an empty line, etc., and click submit:

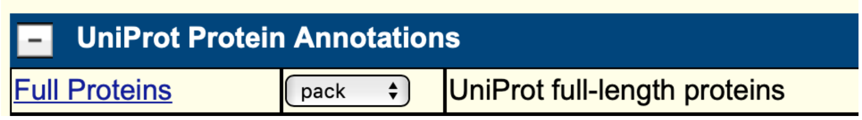
   >N1
   Sequence

   >N2
   Sequence

   >N3
   Sequence

5. The output page should have a line for each sequence, with hyperlinks to "browser" and "details." Other information on the output page refers to how similar the sequences are (% identity), the size of the match, whether it is on the + or – strand (reverse complement) and where in the viral genome it matches (START and END). The viral genome is 29,903 bp; **are these matches near the beginning or end of the virus?**

6. Details will show the exact base-by-base alignment of your sequence to the viral genome. Click on "Browser" for any of the three results. This will take you to the genome view page. The genome view page will show where your sequence matches the genome, in a viewer that allows you to see other features, such as the location of genes, etc. Your sequence is shown in black (it may also remember your PCR search and show that). In dark blue are gene sequences. (The Genome Browser remembers previous settings, so to make sure you are getting the proper view, click on configure below the genome browser window, scroll down, and under UniProt Protein Annotations, select Pack from the drop-down menu (Screenshot 1 below) and then click on submit.)



– UniProt Protein Annotations

Full Proteins	pack	UniProt full-length proteins

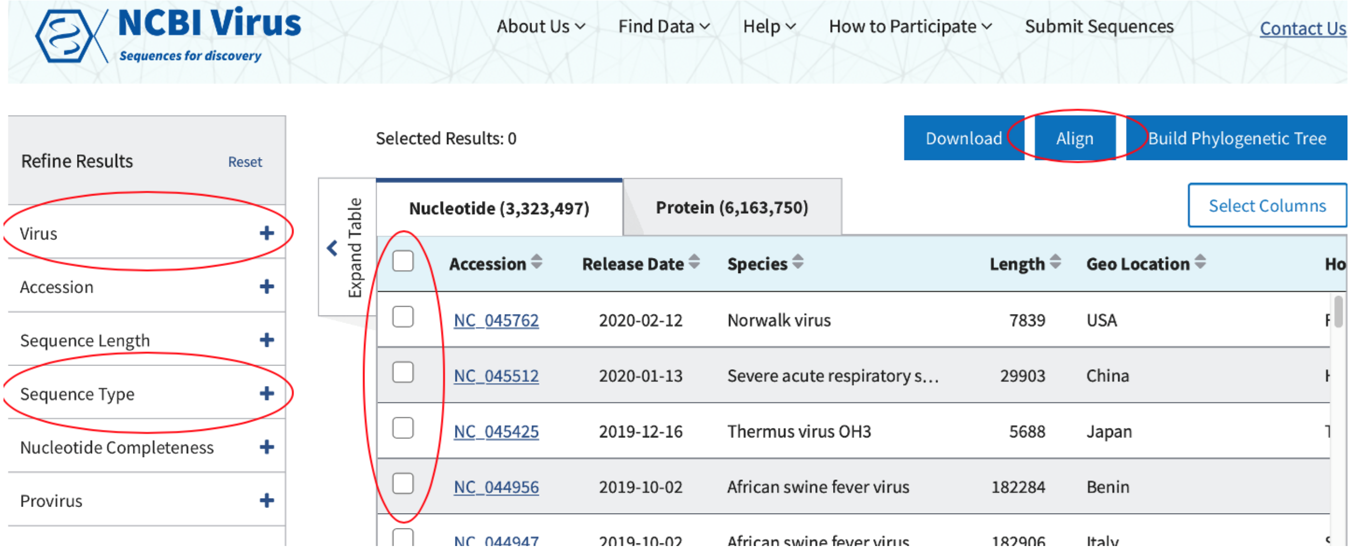
Screenshot 1.

7. Since your sequence is only ~70 bp (compared to ~30,000 bp for the viral genome), click Zoom out 100× (near the top right). The red box shows which part of the virus you are visualizing, and you should now be able to see black boxes in the genome view that are each of your three regions. All three of these are in the same gene. If you mouse over the blue box, the name of the gene should appear, and if you click on it, you should get a description of the gene. **What is the name and function of this gene?**

Section B – Comparing the Viruses that Cause SARS and COVID-19

When testing for a pathogen, a clinician might want a specific test for one pathogen, or a general test for a group of pathogens. For example, a doctor might want to test a patient for "the flu," rather than a specific strain of flu, like H1N1. You are going to determine approximately how closely related the viruses that cause SARS and COVID-19 are and determine regions that would be appropriate for general tests that would detect both, as well as specific tests that would detect SARS-CoV-2, but not SARS-CoV. You are now going to use another database, GenBank, and its associated alignment tool, BLAST (basic local alignment search tool). GenBank is housed at The National Center for Biotechnology Information (NCBI), which is a branch of the National Institutes of Health (NIH). Among many other functions, NCBI stores a wide variety of sequence data. In response to the COVID-19 crisis, they created a database housing sequence information from sequenced patient samples of SARS-CoV-2, from different times and locations of the outbreak. (*Note:* there are a variety of websites with this type of information; the Nextstrain website <https://nextstrain.org> is a great resource for graphics related to the spread and evolution of the SARS-CoV-2 genome.) First you will locate the two viral genomic sequences, and then you will use BLAST to compare them.

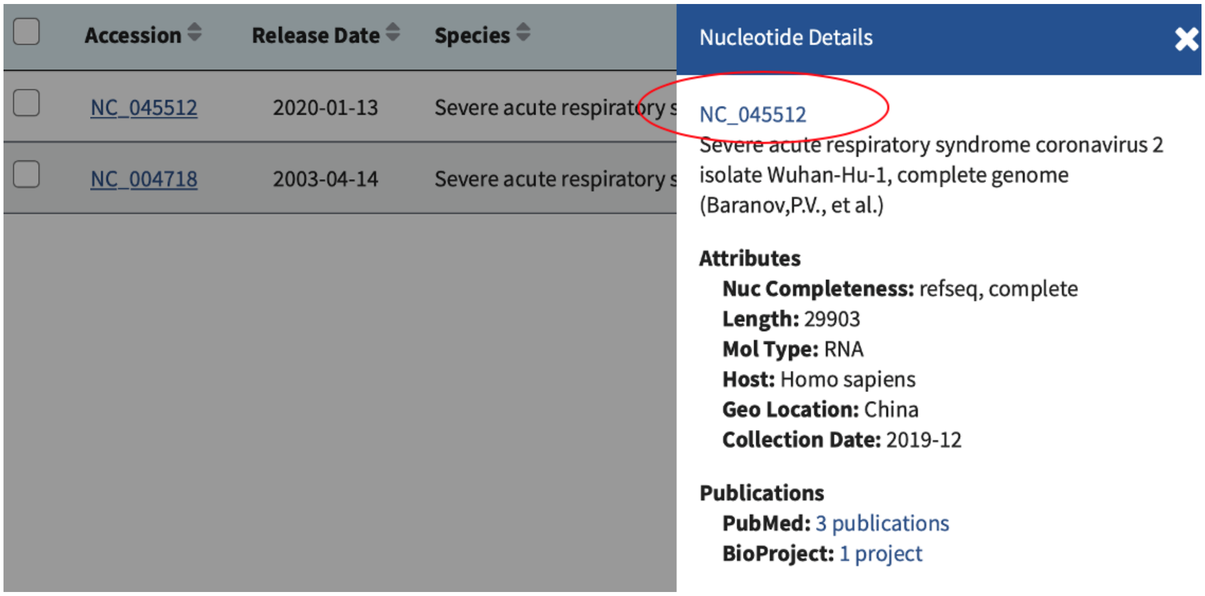
1. Navigate to <https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide>. In the "Refine Results" section on the left, click the plus sign (+) next to "Virus." (*Note:* the features you will use are indicated with red circles in Screenshot 2 below.) A box should appear that says Search Virus. In the box, type SARS. A box should appear with two choices; click on "Severe acute respiratory syndrome-related coronavirus" (not SARS-CoV-2). Under Sequence Type, click on the plus and check the box next to RefSeq. (RefSeq is the sequence that has been adopted as the standard; GenBank refers to all the other strains that have been sequenced and put in the database.) Now you should have two sequences showing in the main table. Clicking on the Accession Number will give you a description. The sequence with the January 2020 date is the SARS-CoV-2 genome (COVID-19) and the one with the April 2003 date is the SARS-CoV genome (SARS).



□	Accession	Release Date	Species	Length	Geo Location	Ho
□	NC_045762	2020-02-12	Norwalk virus	7839	USA	F
□	NC_045512	2020-01-13	Severe acute respiratory s…	29903	China	H
□	NC_045425	2019-12-16	Thermus virus OH3	5688	Japan	T
□	NC_044956	2019-10-02	African swine fever virus	182284	Benin	
□	NC_044947	2019-10-02	African swine fever virus	182906	Italy	S

Screenshot 2.

2. For each one in turn, after you click on the Accession number and the Genome Details panel appears (Screenshot 3, next page), click on the accession number within that panel.



□	Accession	Release Date	Species
□	NC_045512	2020-01-13	Severe acute respiratory s
□	NC_004718	2003-04-14	Severe acute respiratory s

Nucleotide Details ✖

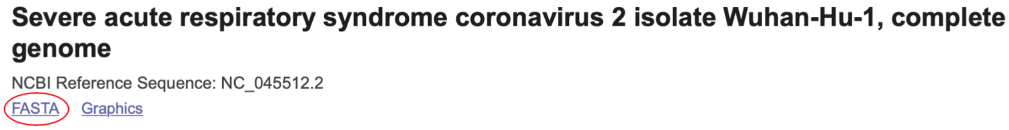
NC_045512
Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome (Baranov,P.V., et al.)

Attributes
Nuc Completeness: refseq, complete
Length: 29903
Mol Type: RNA
Host: Homo sapiens
Geo Location: China
Collection Date: 2019-12

Publications
PubMed: 3 publications
BioProject: 1 project

Screenshot 3.

This will take you to the GenBank file for that genome. There is a lot of information to be had in a GenBank file; if you scroll through you will see identifying information, the location of the various genes and the sequences of the proteins they code for, etc. You want the genome sequence in a particular (FASTA) format, so near the top of the page, under the title of the entry, click on "FASTA" (Screenshot 4).



Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

NCBI Reference Sequence: NC_045512.2
FASTA Graphics

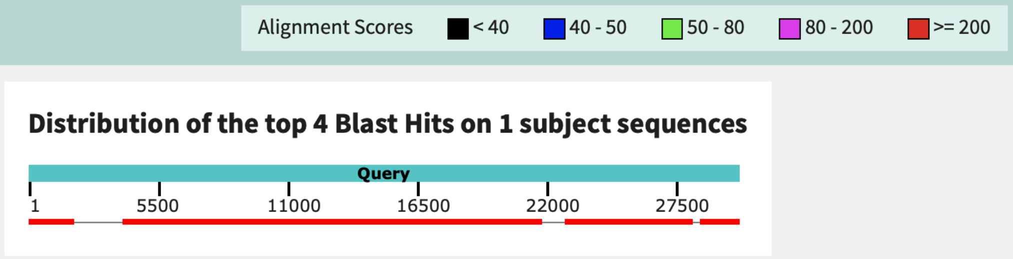
Screenshot 4.

Copy the entire entry from the ">" in the descriptor line through the last base of the sequence. It will take quite some time to highlight and drag all the way to the bottom; it is recommended that you highlight the first line or so, then hold the shift key down, scroll to the bottom of the page and click after the last base. Copy and paste this sequence into its own word document and save it with an appropriate title. Repeat for the other sequence.

3. Now you are going to use BLAST to compare the two sequences. Navigate to <https://blast.ncbi.nlm.nih.gov/Blast.cgi>. Click on Nucleotide BLAST. Below the top box, check the box next to "Align two or more sequences." Paste your SARS-CoV-2 sequence (including the identifier lines) in the top box, and your SARS-CoV sequence (including identifiers) in the bottom box. Click on the BLAST button near the bottom.

4. The results page has a number of items to look at, but it defaults to the description (you need to scroll down a little to see it). You used SARS-CoV-2 as the Query and attempted to align SARS-CoV to it. The "Query Cover" result tells you how much of the SARS-CoV genome could be aligned to the Query sequence. If this number was, say 70%, that does not mean that the two genomes are 70% identical, but rather that the algorithm could match up 70% of the SARS-CoV genome to similar pieces of the SARS-CoV-2 genome. "Per. Identity" gives a score that tells you how many bases matched exactly, however there is a caveat. If the percent identity was, say 60%, that doesn't mean that 60% of the bases in the SARS-CoV genome are identical to bases in the SARS-CoV-2 genome; it means that in those 70% of sequences that can be aligned, 60% of the bases are identical. To make these numbers make more sense, look at the other results tabs.

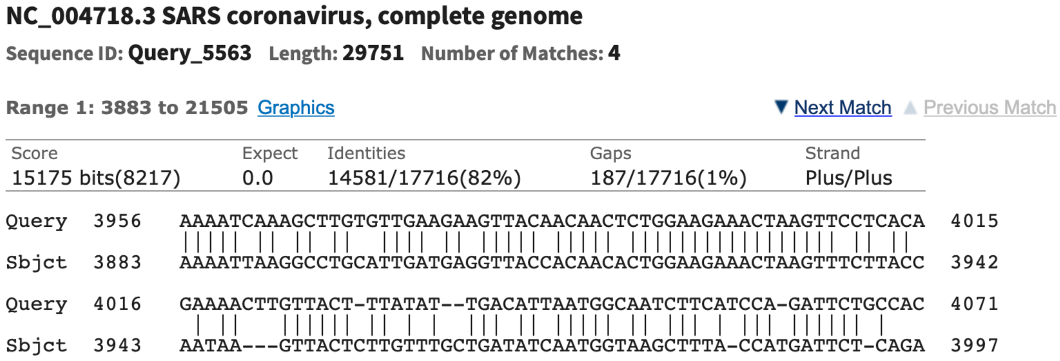
5. Click on the Graphic Summary tab. You should see something like Screenshot 5:



Screenshot 5.

This graphic is essentially saying that there were four sections of the SARS-CoV genome that aligned with high confidence to the SARS-CoV-2 genome (the red boxes). The lines in between the boxes were sections of the genome that did not align well. The total length of all of the aligned boxes compared to the entire genome is the number reported in the Query Cover score you saw previously. The percent of bases that aligned perfectly across those four regions is the percent identity score. Let's take a closer look at those alignments, by clicking on the "Alignments" tab.

6. At the top of the first alignment, you should see something like Screenshot 6:



NC_004718.3 SARS coronavirus, complete genome

Sequence ID: **Query_5563** Length: **29751** Number of Matches: **4**

Range 1: 3883 to 21505 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
15175 bits(8217)	0.0	14581/17716(82%)	187/17716(1%)	Plus/Plus

```
Query  3956   AAAATCAAAGCTTGTGTTGAAGAAGTTACAACAACTCTGGAAGAAACTAAGTTCCTCACA  4015
              ||||| || || ||  |||| || ||||| ||||| ||||||||||||||||| || ||
Sbjct  3883   AAAATTAAGGCCTGCATTGATGAGGTTACCACAACACTGGAAGAAACTAAGTTTCTTACC  3942

Query  4016   GAAAACTTGTTACT-TTATAT--TGACATTAATGGCAATCTTCATCCA-GATTCTGCCAC  4071
               | ||   |||||| || | |  ||| || ||||| || ||| | ||| |||||| |
Sbjct  3943   AATAA---GTTACTCTTGTTTGCTGATATCAATGGTAAGCTTTA-CCATGATTCT-CAGA  3997
```

Screenshot 6.

Number of matches shows how many different sections aligned (as we saw on the Graphic Summary page). "Range 1" is indicating it's the first displayed of those alignments (scrolling down through the page will show you the rest of them). 3883 to 21,505 are the bases of the SARS-CoV sequence that are represented in this alignment. Identities are the number of bases that matched exactly (indicated by a vertical line in the alignment), and gaps are the number of times there was a base (or multiple bases) in one sequence but not the other (small insertions and deletions, indicated by dashes). The rows in the alignment are 60 bp long.

7. Remembering that a primer/probe set is about 70 bp:

What are two regions that would be good targets for a test that would detect ONLY SARS-CoV-2, but not SARS-CoV?

What are two regions that would be good targets for a test that would detect BOTH SARS-CoV and SARS-CoV-2 (provide screenshots).

Section C – Mutation of SARS-CoV-2 Over Time

All viruses accumulate mutations over time, and different classes of viruses have different mutation rates based on the fidelity with which their genomes are replicated. Viral mutations can affect the accuracy of tests, the efficacy of vaccines, and virulence (the ability to spread, the seriousness of illness, etc.). In this section you will compare a number of sequenced genomes from clinical samples during the first several months of the COVID-19 outbreak to the RefSeq genome for SARS-CoV-2, to identify the number of mutations that have occurred and calculate an approximate mutation rate. (*Note:* In reality, mutation rates are not calculated this way; the math is much more complex. Usually viral rates are calculated after controlled experiments in tissue culture in the lab, with algorithms that account for many factors, including types of mutation, selection, etc. Calculations more similar to what you will be doing can be done based on extensive phylogenies, calculating time since divergence, etc. You are going to try to get a rough idea of how frequently the SARS-CoV-2 virus is mutating, based on simple comparisons and some simple assumptions, while continuing to work with GenBank and BLAST).

For this analysis, you will again use the NCBI virus database. The earliest samples in the database were collected from patients in late December, 2019 in China. As previously mentioned, one of these has been adopted as the “RefSeq” genome, the standard that others are compared to. Although an exact collection date is not listed, you will use 12/23/2019, the earliest date in the database. You will compare its sequence to others from different dates and locations in the first few months of the pandemic. You are going to assume that this is the virus from which all the others arose (in reality there is a more ancestral strain that was never sequenced, as evidenced by phylogeny, but this assumption will work for our rough estimate purposes).

There are some things to be aware of when considering the data. There are some whole genome sequences and some smaller partial gene sequences. We will be comparing whole genome sequences, but even within those, there are differences in the total number of bases reported due to differences on the ends outside of the genes (untranslated regions or artifacts of amplification and sequencing). We will therefore use the smaller of the sequences being compared to determine how many bases were (presumably) aligned between the two sequences. Also, the sequences are listed by release date (when the sequence was put in the database), but we are more interested in the collection date (when the sample was taken from the patient), which is a separate column and can also be accessed by clicking on the accession number.

Remember that the mutation rate for viruses is calculated as substitutions/nucleotide/cell infection. You are going to estimate the mutation rate by measuring two things and making some assumptions about mutations and transmission. The number of substitutions will be the number of differences between the two genomes. The number of nucleotides will be the number of bases that aligned between them (we will use the size of the shorter of the two genomes). For the cell infections, we are going to make a large assumption about viral transmission. We are going to estimate transmission events (how many times it jumped from one person to another).

Why are transmission events not the same as cell infections? Which number is expected to be higher, transmission events or cell infections? Will this lead to an over- or an underestimation of mutation rate in our calculations?

COVID-19 is a variable disease, with a wide range of incubation periods (the time from exposure to the first symptoms) and reports exist of asymptomatic people who were unknowingly infectious. Incubation periods range from 2–14 days, with early reported median incubation periods of around five days. People were potentially infectious before, and definitely after, they started showing symptoms, and many that did not show symptoms did not realize they were infected. For these purposes, you will estimate the time to transmission as five days from infection.

1. As you did previously in Section B, navigate to: <https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide>.

   This time, in the set of filters on the left, click the + next to virus and type SARS. Select the SARS-CoV-2 genome. You will compare a number of genomes from clinical samples to the RefSeq SARS-CoV-2 sequence. In the Sequence Length Filter, put a range of 28,000 to 30,000 bp. Within the table of sequences themselves, sort from earliest collection date (by clicking the arrow at the top; the RefSeq sequence should be the first listed), so only the first two pages will be required.
2. **Fill out Table 1 below** for each of the provided Accession Numbers. (*Hint:* they are provided in order from earliest collection date.) The basic information can be found in the GenBank table (accession number, collection date, country, genome size). To determine the number of mutations from the "original" sequence, the sequences will need to be aligned. For each sequence, you will align it to the Ref Seq sequence. To align two sequences, click the box next to each one, and then click on the blue box that says "Align" at the top right of the table (look back at Figure 2). Once you have aligned them, count the red lines; that is the number of differences between the two. (*Note:* you can also get this information from a BLAST alignment, but this is quicker.) Darker red lines indicate more than one change, but that should not be an issue with these sequences.
3. As you are looking up each sequence, go to the GenBank file, click on FASTA, and copy the genomic sequence with header as before. Make one Word file with all of your mutants (remember to skip a line in between each). You will use this file in Section D further below.
4. After you have aligned the first sequence to the RefSeq, go back to the previous tab, deselect that sequence (by clicking on the box next to the Accession number), and select the next one in the provided table. Repeat for each sequence.
5. To complete the table, calculate the number of days since the "original" RefSeq sequence (remember that January always has 31 days, and February in 2020 had 29 days). Divide that number by five (our estimate of the time to transmission). Finally, calculate the "substitutions/nucleotide/cell infections," in this case, mutations/bp of the smaller genome/# of transmissions for each genome. Complete the table by calculating the average mutation rate and standard deviation for the six different genomes.

   Table 1. Number of mutations.

Accession number	*Collection date*	*Country*	*Genome size*	*# of mutations compared to RefSeq*	*Days from RefSeq*	*# of transmissions since RefSeq*	*s/n/c*
NC_045512	12/23/19	China	29903	NA	NA	NA	NA
MT093631							
MN997409							
MT093571.1							
MT152824							
MT263074							
MT263430							
Average							
Std. Dev.							

How does this number compare to the known mutation rates for coronaviruses?

In addition to assumptions about transmission time and equating transmissions to cell infections, what other factors could contribute to differences between the mutation rate you calculated and the actual mutation rate?

Section D – Effect of Mutation Rate on Genetic Tests

For the final section of this activity you will determine whether any of the six variants you analyzed have accumulated mutations in the three primer/probe regions from the original CDC test (N1, N2 and N3).

1. Navigate back to nucleotide BLAST as before, and click the box to align multiple sequences together. Paste the N1 sequence that you identified in Part II, Section A into the top box. Paste the SARS-CoV-2 sequences from Section C into the bottom box and hit Blast. Look at your results. **Have any of the viruses obtained a mutation in the N1 region that would potentially prevent the test from working? Show evidence for or against using a screenshot.**

2. Repeat for N2 and N3. Note that if you navigate back using back arrows, when you get to the BLAST window, you'll need to refresh the page and paste your sequences in again.

   Have any of the viruses obtained a mutation in the any of the test regions that would potentially prevent the test from working? If so, show evidence using a screenshot.

   What are the implications of the viral mutation rate for the ability of the original test to positively identify new strains of the virus as the pandemic continues? What are the chances that a virus would accumulate a mutation in both the N1 and N2 regions, according to your calculations?