

LABORATORY ACTIVITY

Analyzing Coronavirus Genomes

Trace Jordan, Emery McKinstry, and Aisling Dunne

College Core Curriculum, New York University

Developed: May 1, 2020

Updated: January 13, 2021

Overview

In this laboratory activity, you will study the genomes of coronaviruses using the BLAST program to compare their base sequences. The activity begins by comparing the genomes of SARS-CoV-1, the virus that caused the original SARS outbreak in 2002, with the genome of SARS-CoV-2, which is responsible for COVID-19. Next, you will compare the reference genome of SARS-CoV-2 with the genomes of coronavirus samples from COVID-19 patients at various geographical locations and time points. Using the BLAST software, you will identify mutations that have occurred in the coronavirus genome and add your observations to a shared dataset. Finally, you will analyze and integrate all the genome data to propose a hypothesis for how SARS-CoV-2 mutated and spread geographically during the early months of the COVID pandemic. This laboratory activity provides an authentic experience of scientific investigation using real genomic data to address an important question for global public health.

Preparation

Prior to this lab, read the introduction and be familiar with the following concepts:

- The structure of a coronavirus
- Human diseases caused by coronaviruses
- The distinction between COVID-19 and SARS-CoV-2
- Composition of the SARS-CoV-2 genome
- Analyzing the base sequences of coronavirus genomes

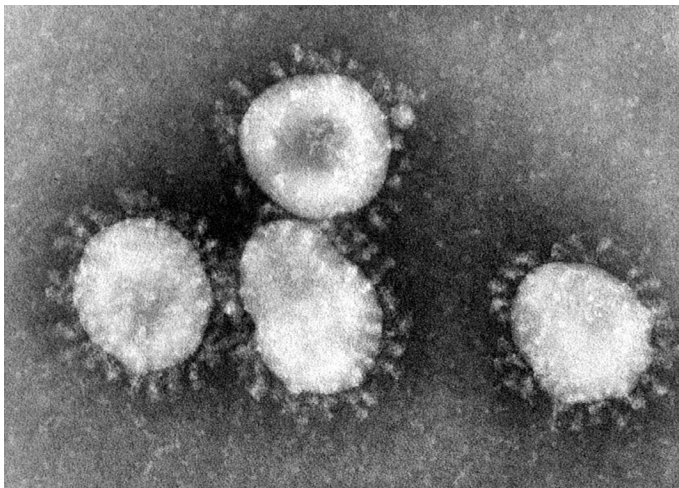
Table of Contents

1. Introduction
2. Procedures
3. Data Sheets

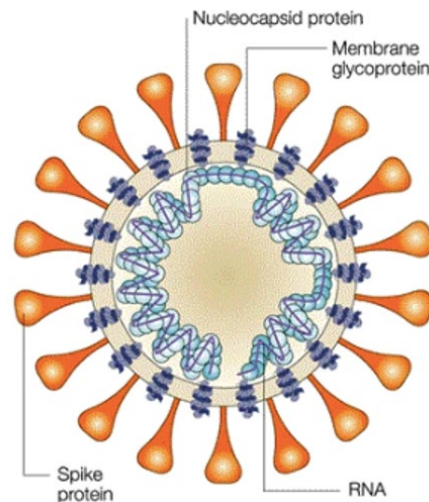
INTRODUCTION: Analyzing Coronavirus Genomes

What is a Coronavirus?

A **coronavirus** is a type of virus that has particular structural and genetic characteristics. The name “coronavirus” is derived from the word for crown (corona). When viewed using an electron microscope, a coronavirus has “spikes” protruding from its surface that look like a royal crown. Figure 1(a) shows an electron microscope image of a coronavirus, and Figure 1(b) illustrates the molecular composition of the virus. The genome of a coronavirus is composed of RNA, which is enclosed by a lipid envelope. Protruding from the surface of the virus are so-called “spike proteins,” which account for the characteristic appearance of a coronavirus. The virus uses these spike proteins to attach to a receptor protein on a host cell, which is the first step of cellular infection.



(a)



(b)

Figure 1 (a) An electron microscope image of a coronavirus, showing the characteristic spikes that resemble part of a crown. (b) The structure of a coronavirus and its molecular components.

Image sources:

- (a) https://commons.wikimedia.org/wiki/File:Coronaviruses_004_lores.jpg
- (b) https://www.researchgate.net/figure/Schematic-diagram-of-the-SARS-coronavirus-structure-reproduced-from-ref-20-The-viral_fig1_8953389

Laboratory Activity: Analyzing Coronavirus Genomes

Coronaviruses are capable of causing human diseases that usually affect the respiratory system. Examples of infectious coronaviruses include:

SARS	Severe Acute Respiratory Syndrome
MERS	Middle East Respiratory Syndrome

A new disease outbreak was reported in Wuhan, China, in late 2019. The cause of this disease was quickly identified as a new type of coronavirus. Initially, the virus was called the “novel coronavirus.” Subsequently, the World Health Organization (WHO) provided the disease and virus with their official names (note the distinction between the disease and the pathogen that causes it). The long title of the novel coronavirus indicates its close similarity with the virus that causes SARS.

Disease	Coronavirus Disease	COVID-19
Virus	Severe Acute Respiratory Syndrome Coronavirus 2	SARS-CoV-2

Composition of the SARS-CoV-2 Genome

The genome of a virus provides valuable information about its composition, origins, and evolution. One of the first scientific reports about the novel coronavirus was its genome sequence. The genome of SARS-CoV-2 contains almost **30,000 RNA bases**. The bases are divided into genes that provide the instructions to make **29 different proteins**. There are also regions of the genome—called noncoding regions—that do not make proteins.

Figure 2 shows the organization of the SARS-CoV-2 RNA genome, which is called a **genome map**. Like a road map, a genome map provides a set of “signs” that helps us know the location. The genome map begins with the 5'-end of the RNA genome on the left and ends with the 3'-end on the right. The figure highlights specific genes that correspond to particular proteins (not all proteins are shown). One example is the gene for the **spike protein**, which extends from base positions 21,563 to 25,384. Understanding the spike protein is particularly important because it is used by the virus to attach to human cells in the first stage of infection.

Laboratory Activity: Analyzing Coronavirus Genomes

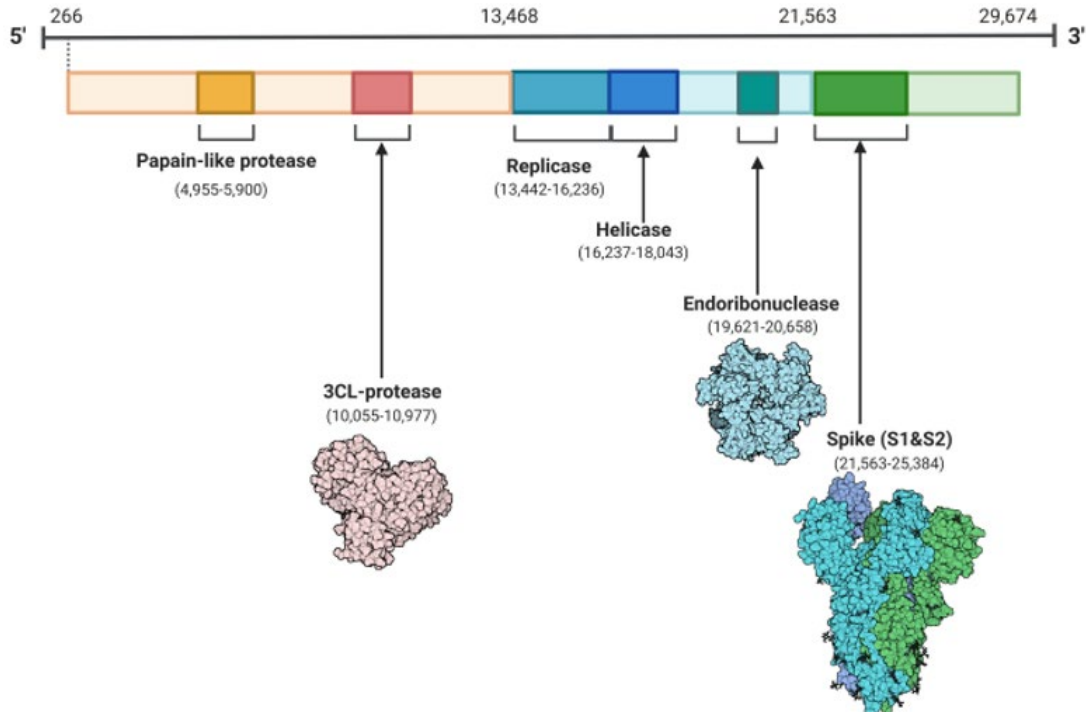


Figure 2 A map of the SARS-Cov-2 RNA genome. The map provides the location of genes that encode the information to make specific proteins.

Image source

<https://www.ncbi.nlm.nih.gov/books/NBK554776/figure/article-52171.image.f5/>

Analyzing the Base Sequences of Coronavirus Genomes

You will use a program called BLAST to analyze the base sequences in various coronavirus genomes. The RNA genome in a coronavirus is composed of four bases: A U G C. However, the tools of genome sequencing are based on working with DNA and not RNA. For this reason, the RNA genome is first converted to DNA in order to be sequenced. This process is called **reverse transcription** because it occurs in the opposite direction to the usual process of transcription within cells (Figure 3). To make this happen, scientists employ an enzyme called **reverse transcriptase** that converts single-stranded RNA into double-stranded DNA.

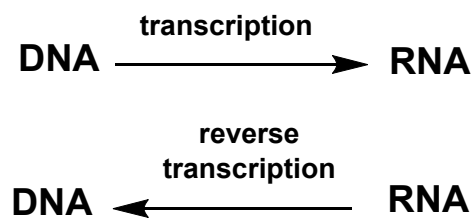


Figure 3 A comparison of transcription (DNA → RNA) and reverse transcription (RNA → DNA).

Laboratory Activity: Analyzing Coronavirus Genomes

After the viral RNA genome has been converted to DNA, scientists can use the rapid and powerful methods of DNA sequencing to analyze all the bases in the genome. The DNA sequences are stored as the **DNA coding strand**. As you will recall from Lab Project 6, the DNA coding strand has the same orientation ($5' \rightarrow 3'$) as the RNA strand, but the DNA coding strand contains thymine (T) instead of the RNA base uracil (U).

When you compare different coronavirus genomes, a variety of base mutations are possible. Some examples are illustrated in Figure 3, which shows the comparison between a reference sequence (our point of reference) and a sample sequence (from the sample we are investigating). A **deletion** occurs when a base in the reference sequence is missing (deleted) in the sample sequence. A **base substitution** occurs when a base in the reference sequence is substituted by a different base in the sample sequence (Figure 3 shows an $A \rightarrow T$ substitution). An **insertion** occurs when the sample sequence contains a new base that is not present in the reference.

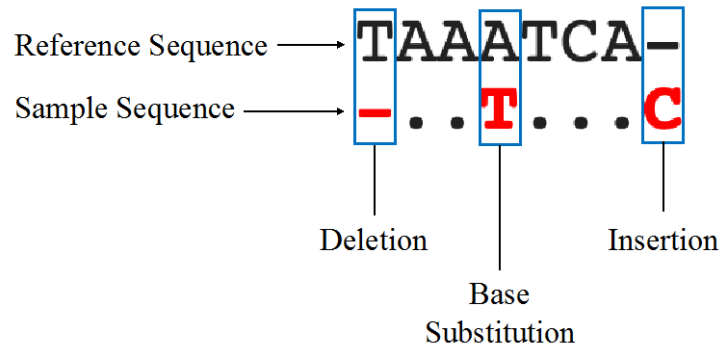


Figure 4 Examples of base mutations that can occur in a comparison of a sample sequence with a reference sequence.

During the lab activity, you will first compare the original SARS genome (SARS-CoV) with the genome of SARS-CoV-2. You will then compare various SARS-CoV-2 genomes from different geographical locations to determine the extent to which the viral genome has mutated. This type of research is currently underway in many research laboratories around the world, including NYU's Langone Medical Center. The map in Figure 5 shows how viral genome analysis has tracked the global spread of SARS-CoV-2. In this map, each color represents a particular variant of the virus as identified by its genome sequence. As an analogy, we can imagine the genome of each virus acting like the barcode on a shipping label. The unique barcode on the label enables us to monitor the distribution of a package around the world. Similarly, the unique base sequence of each viral variant enables us to track where it has traveled.

Laboratory Activity: Analyzing Coronavirus Genomes

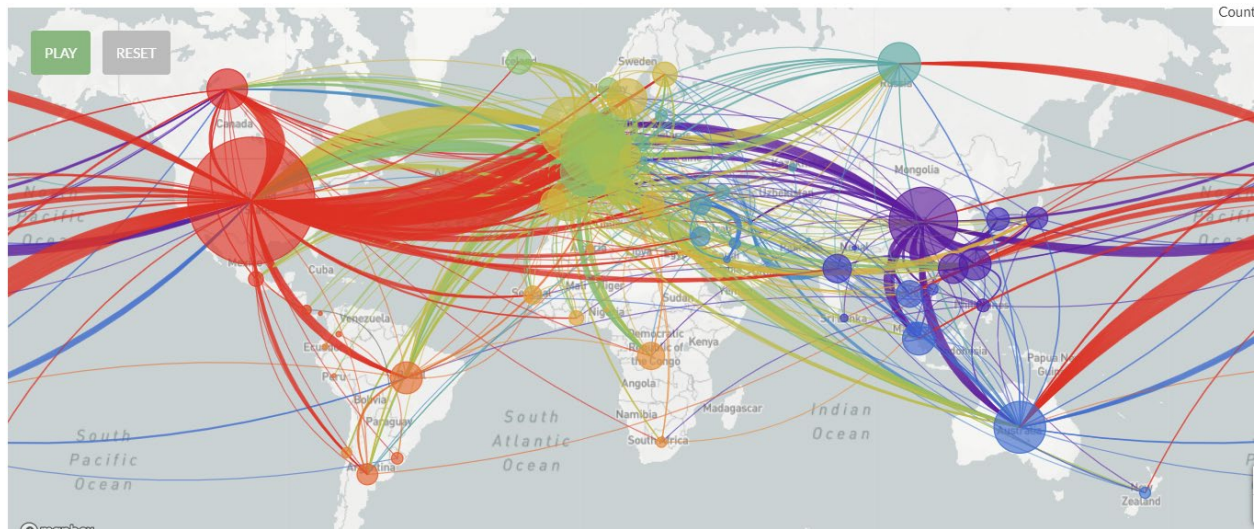


Figure 4 The global spread of SARS-CoV-2 as revealed by viral genome sequences. Each color represents a particular strain of the virus based on its specific genome sequence.

Image Source:

<https://wgicouncil.org/blogs/health-organizations-agencies-enabling-covid-19-spatial-dashboards/>

PROCEDURES: Analyzing Coronavirus Genomes

PART A: COMPARING THE GENOMES OF TWO SARS VIRUSES

For the first part of the lab activity, you will compare the genomes of two SARS viruses. The first genome sequence, called “SARS-COV-1,” is from the original SARS virus outbreak in 2003. The second sequence is the first genome that was analyzed for the recent SARS outbreak in Wuhan, China. This genome sequence, published in early 2020, is called “SARS-CoV-2 (Wuhan).”

1. Locate the virus genome files from NYU Classes. The virus genomes will be stored in text files for SARS-CoV-1 and SARS-CoV-2.
2. Go to the NCBI blast website (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>).
3. Select “nucleotide blast,” as shown in **Figure 1**. We select a nucleotide BLAST because we will be comparing virus genomes that are composed of nucleotides.

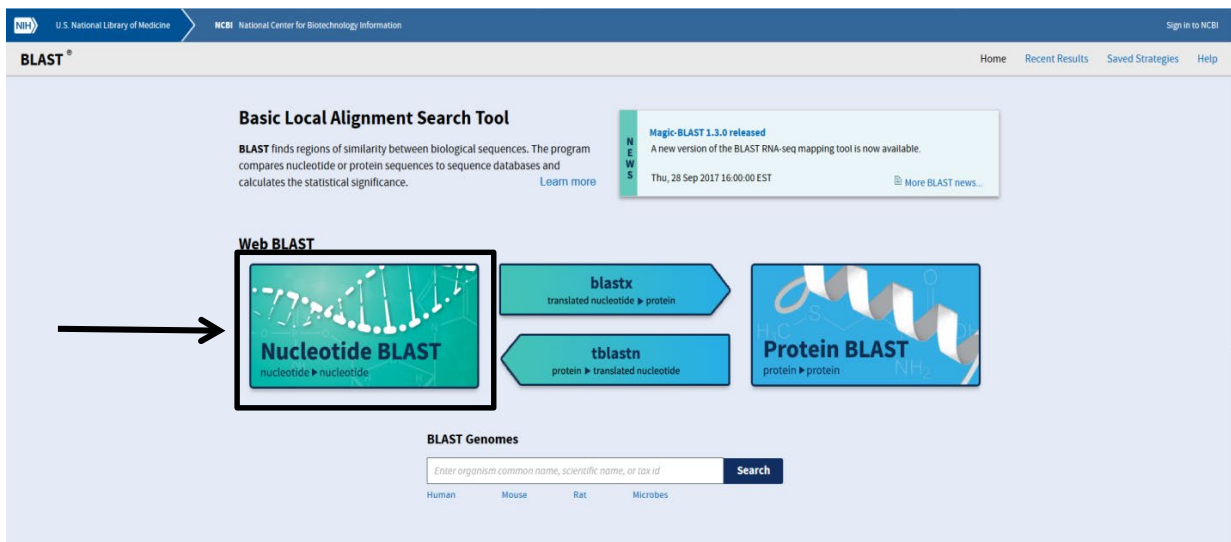


Figure 1: NCBI BLAST home screen.

Laboratory Activity: Analyzing Coronavirus Genomes

- From the nucleotide blast page, click the box “Align two or more sequences” (Figure 2).

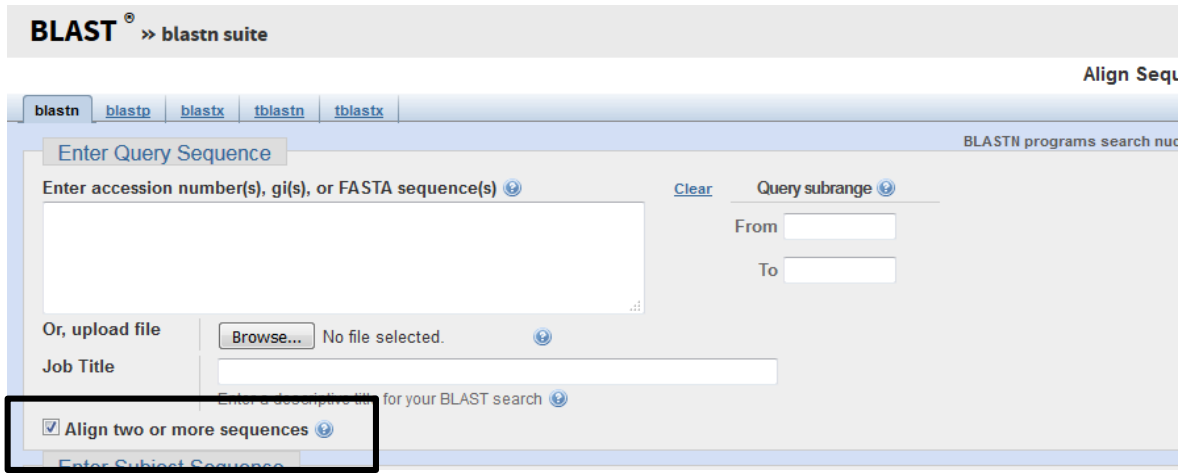


Figure 2: Nucleotide BLAST webpage.

- A second text box will appear (Figure 3).

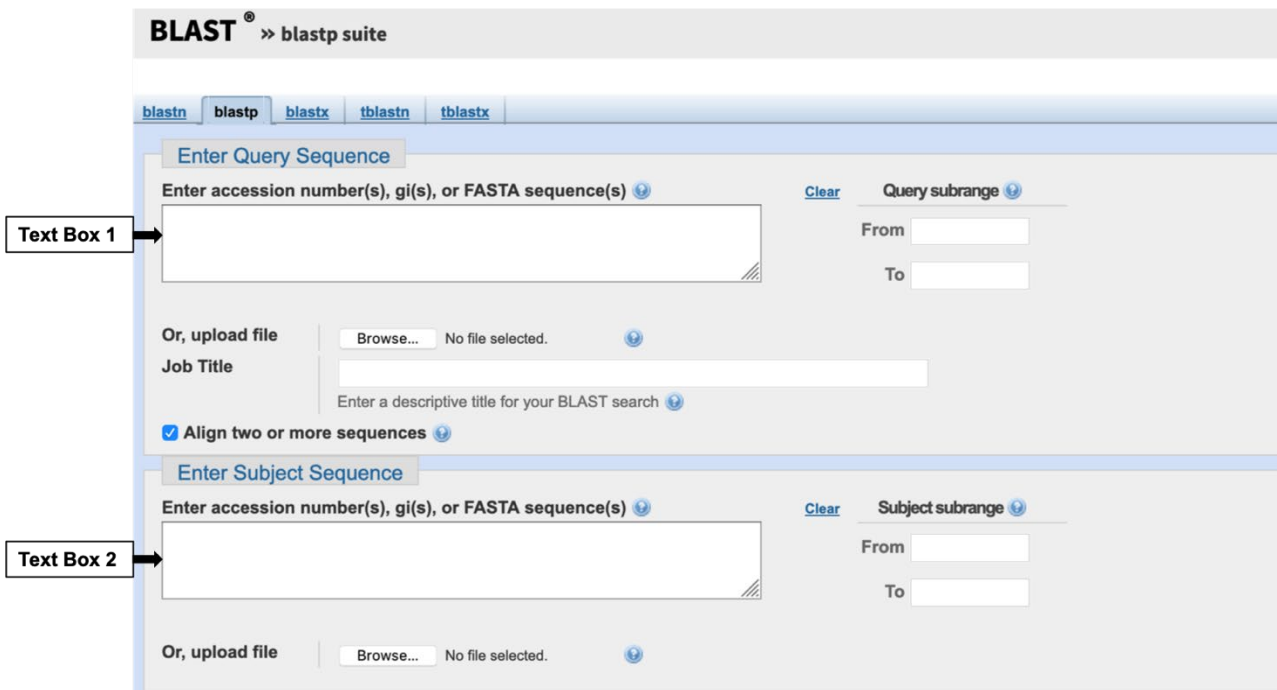


Figure 3: Alignment text boxes.

- Table 1 provides accession numbers for two different SARS viruses. An accession number identifies a specific DNA sequence. Instead of using text files to upload the viral genomes as you have done in a previous NCBI lab, you will copy and paste an accession number into the text box.

Table 1: Accession Numbers for Part A: Comparing the Genomes of Two SARS Viruses

SARS Virus	Accession Number
SARS-CoV-1	AY278488
SARS-CoV-2	NC_045512

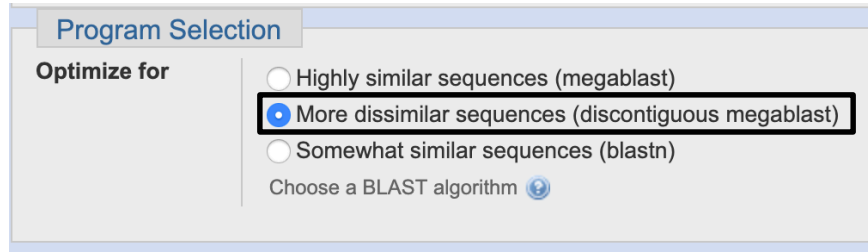
- Copy and paste the accession number for SARS-CoV-1 into Text Box 1 (**Figure 4**). This will be your query sequence. In general, a “query sequence” provides a point of reference for comparing another sequence, which is called the “subject sequence.” The query sequence is also called the “reference sequence.”
- Copy and paste the accession number for SARS-CoV-2 into Text Box 2 (**Figure 4**).

The screenshot shows the BLASTN web interface. At the top, there are tabs for 'blastn', 'blastp', 'blastx', 'tblastn', and 'tblastx'. Below the tabs, the text 'BLASTN programs search nucleotide subjects using a nucleotide q' is visible. The main interface is divided into two sections: 'Enter Query Sequence' and 'Enter Subject Sequence'. In the 'Enter Query Sequence' section, the text box contains 'AY278488'. To the right of this text box are 'Query subrange' fields for 'From' and 'To'. Below the text box, there is an 'Or, upload file' button and a 'Choose File' button with 'no file selected' text. The 'Job Title' field is populated with 'AY278488:SARS coronavirus BJ01, complete genome'. A checkbox labeled 'Align two or more sequences' is checked. The 'Enter Subject Sequence' section has a text box containing 'NC_045512', 'Subject subrange' fields for 'From' and 'To', and an 'Or, upload file' button with a 'Choose File' button and 'no file selected' text.

Figure 4: Inserting accession numbers of two SARS viral genomes into text boxes.

- Lastly, you will need to change the Program settings. Select “More dissimilar sequences (discontiguous megablast)” in the “Program Selection” box (**Figure 5**).

Laboratory Activity: Analyzing Coronavirus Genomes



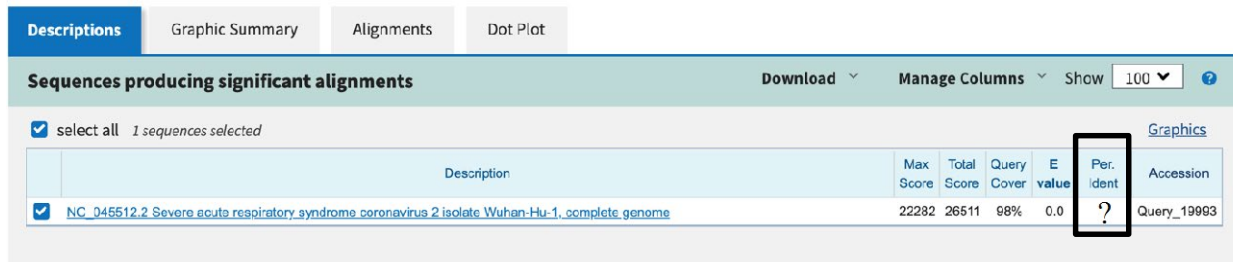
The image shows a 'Program Selection' box with the following options:

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Below the radio buttons is a link: 'Choose a BLAST algorithm' with a help icon.

Figure 5: Program Selection box.

10. Click “BLAST.” When your search is complete, a screen containing the BLAST results will be displayed.
11. In the Descriptions View, find the “Per. Ident.” Box that is highlighted in Figure 6. This stands by “Percent Identity” and gives the amount of similarity between the query sequence and the subject sequence.

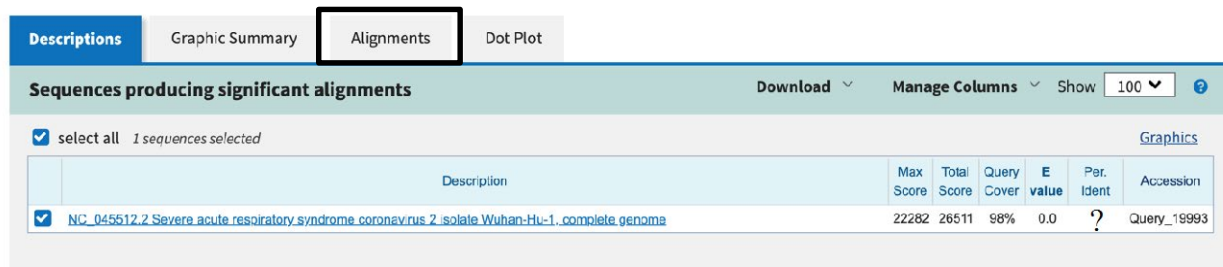


The image shows the 'Descriptions' view of BLAST results. The 'Per. Ident.' column is highlighted with a red box.

Description	Max Score	Total Score	Query Cover	E value	Per. Ident.	Accession
<input checked="" type="checkbox"/> NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome	22282	26511	98%	0.0	?	Query_19993

Figure 6: BLAST Results page featuring Percent Identity.

12. Answer **Question 1** in the Data Sheets.
13. Click the “Alignments” tab located near the middle of the page (**Figure 7**).



The image shows the 'Alignments' view of BLAST results. The 'Alignments' tab is highlighted with a red box.

Description	Max Score	Total Score	Query Cover	E value	Per. Ident.	Accession
<input checked="" type="checkbox"/> NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome	22282	26511	98%	0.0	?	Query_19993

Figure 7: BLAST Results page featuring the Alignment View.

14. Find the Alignment View and use the drop-down menu to choose “Pairwise with dots for identities” (**Figure 8**). The *query sequence* is the SARS-CoV-1 sequence. The pairwise with dots view shows the reference sequence at the top with the subject sequence aligned below (SARS-CoV-2_Wuhan). *Dots are used to show nucleotides that are identical, and letters are used to highlight nucleotides that differ.*

Laboratory Activity: Analyzing Coronavirus Genomes

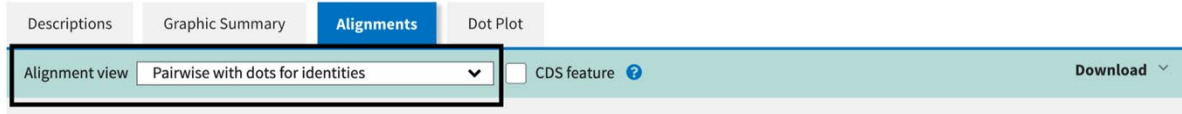


Figure 8: Changing the Alignment view on the BLAST Results page.

15. Find the “Sort by” drop-down menu in the light blue header and select “Subject start position” (Figure 9).

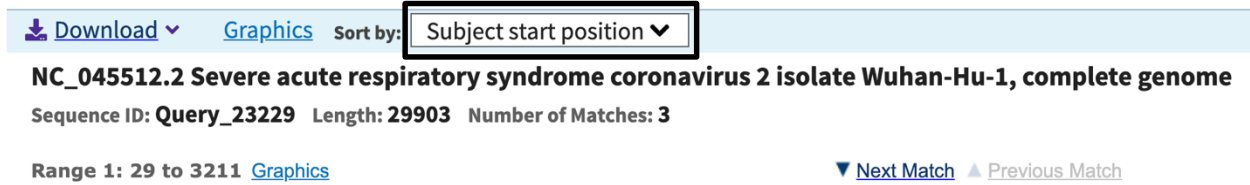


Figure 9: Changing how the sequences are sorted on the BLAST Results page.

16. Scroll down the page to see if there are positions where the *query* sequence (also called the *reference* sequence) differs from the subject (SARS-CoV-2) sequence. Note that “Sbjct” on the left side turns **red** whenever there is a difference between the subject sequence and the query/reference sequence. There are three ways the sequences could differ, which are illustrated in **Figure 10**.

- **Base substitution:** There is a **red letter** instead of a dot in the subject sequence and there is a letter above it in the query sequence.
- **Base deletion:** There is a **red dash** instead of a dot in the subject sequence.
- **Base insertion:** There is a **red letter** instead of a dot in the subject sequence and there is a dash above it in the query sequence.

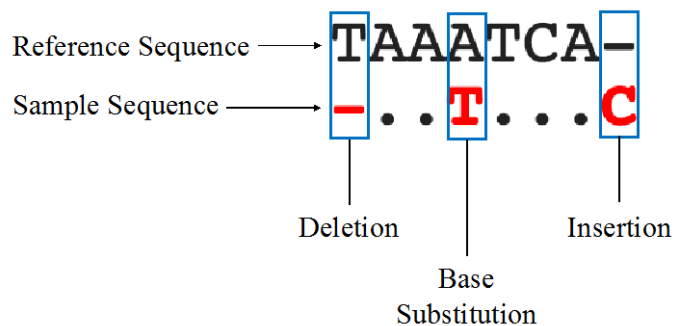


Figure 10: Three different types of mutation between a reference/query sequence and sample sequence on a BLAST Results page. First, a deletion can be seen when a “T” in the reference sequence changes to a “-” in the sample sequence. Second, a base substitution can be seen when an “A” in the reference sequence changes to a “T” in the sample sequence. Last, an insertion can be seen when a “-” in the reference sequence changes to a “C” in the sample sequence.

Laboratory Activity: Analyzing Coronavirus Genomes

17. Note that the numbers at the beginning and end of the lines refer to the position of the first and last nucleotide (**Figure 11**). The reference and subject sequence will start and end on different nucleotide numbers for each row of sequences. This happens because the BLAST program aligns the bases that are most similar, even though they may not be numbered the same in the sequence. *When reporting your data, use the subject sequence (Sbjct) to record the position of the base mutation.*

```
Reference Sequence → Query 8 AAGCCAACCAACCT-CGAT // GAACTTTAAAAATCTGTG 66
Subject Sequence → Sbjct 29 ..A.....T.T..... // ..... 88
```

Figure 11: Sequence alignments. The reference sequence refers to the top row (Query) and the subject sequence refers to the bottom row (Sbjct). The // in the middle of the sequence indicates a break in the sequence in order to show the whole row in the figure.

18. You will be placed into break-out rooms by your lab instructor and assigned two rows of the sequence for your analysis.
19. Answer **Question 2** by completing **Data Table 1** in the Data Sheets.
20. Add your data from Data Table 1 to the Google sheet called **Part A: Genome Comparison of SARS-CoV-1 and SARS-CoV-2**
21. Answer **Question 3** in the **Data Sheets**.

PART B: IDENTIFYING MUTATIONS IN SARS-CoV-2 SEQUENCES

1. In Part B, you will be using the Wuhan strain of SARS-CoV-2 as your query/reference sample and comparing different viral genomes to this strain. Your subject sequence will correspond to one of the following locations assigned by your lab instructor: California, New York, Spain, Washington State (Case 1), or Washington State (Case 5170).
2. The accession number for the Wuhan strain of SARS-Cov-2 (your reference) is NC_045512.
3. Table 2 provides the accession numbers for the genomes of SARS-CoV-2 samples that have been collected from patients in various locations. The first last row contains the accession number for Italy, which your lab instructor will use for a demonstration.

Table 2: Accession Numbers for Part B: Identifying Mutations in SARS-CoV-2 Sequences

Sample Location	Case Number	Accession Number	For Use By
Italy	1	MT077125	Lab Instructor
Washington State	1	MN985325	Students
Washington State	5170	MT375482	Students
New York City	Unknown	MT371038	Students
Spain	3	MT198652	Students
California	1	MN994467	Students

LAB INSTRUCTOR DEMONSTRATION

4. Start a new BLAST alignment. Scroll to the top of the page and click the “blastn suite-2sequences” button (**Figure 12**).

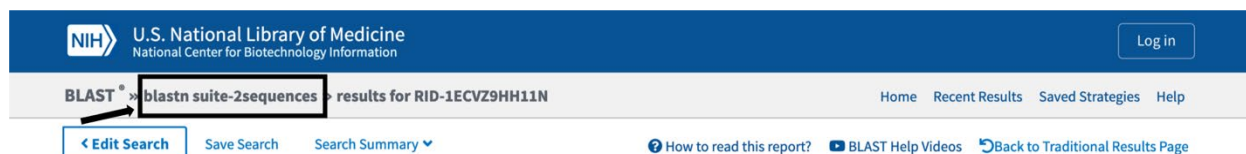


Figure 12: “blastn suite-2sequences” at top of results page.

5. Copy and paste the accession number for SARS-CoV-2 – Wuhan into Text Box 1. (**Figure 13**)
6. Copy and paste the accession number for SARS-CoV-2 – Italy into Text Box 2. (**Figure 13**)

Laboratory Activity: Analyzing Coronavirus Genomes

The screenshot shows the BLAST web interface. At the top, there are tabs for 'blastn', 'blastp', 'blastx', 'tblastn', and 'tblastx'. Below the tabs, the 'Enter Query Sequence' section has a text box containing 'NC_045512' and a 'Clear' button. To the right, there are 'Query subrange' fields for 'From' and 'To'. Below this, there is an 'Or, upload file' section with a 'Choose File' button and 'No file chosen' text. The 'Job Title' field contains 'NC_045512:Severe acute respiratory syndrome...' and a 'Clear' button. A checkbox labeled 'Align two or more sequences' is checked. The 'Enter Subject Sequence' section has a text box containing 'MT077125' and a 'Clear' button. To the right, there are 'Subject subrange' fields for 'From' and 'To'. Below this, there is another 'Or, upload file' section with a 'Choose File' button and 'No file chosen' text.

Figure 13: Inserting accession numbers of SARS-COV-2 genome samples into text boxes.

7. Ensure the Program Settings are still set to “More dissimilar sequences (discontiguous megablast)” in the “Program Selection box (**Figure 5**).
8. Click “BLAST.”
9. Click the “Alignments” tab located near the middle of the page (**Figure 7**).
10. Find the Alignment View and use the drop-down menu to choose “Pairwise with dots for identities.”
11. Scroll down the page to look for base positions where the query (reference) sequence differs from the subject sequence. You will notice that most of the bases are identical. However, you can visually identify any mutations because the “sbjct” on the left side will turn **red** whenever there is a difference
12. The query sequence differs from the subject sequence where there is a **red letter** instead of a dot, showing a change in the base at that position. Note that the numbers at the beginning and end of the lines refer to the position of the first and last nucleotide as seen in **Figure 14**.

Reference Sequence →	Query	26097	AATGTTGATGAGCCTGAAGAACATGTCCAATTACACAATCGACGGTTCATCCGGAGT	26156
Subject Sequence →	Sbjct	26041 T	26100
			↑	
		Base #26,041		Base #26,144 Base #26,100

Figure 14: Sequence alignment when using “SARS-CoV-2 – Wuhan” as a reference and “SARS-CoV-2 – Italy.” You can see a base mutation at position 26,144 from a guanine (G) to a thymine (T).

IDENTIFYING GENOME MUTATIONS

13. You will be placed into breakout rooms by your lab instructor and assigned a specific SARS-CoV-2 genome sample to analyze.
14. Follow Steps 5 – 13 above, but this time use your assigned SARS-CoV-2 genome sample.
15. Answer Question 4 by completing Data Table 2 in the Data Sheets. Record the **type of base mutation** (e.g., C → A) and the **position of the mutation**.
22. Add your data from Data Table 1 to the Google sheet called **Part B: Genome Comparison of SARS-CoV-2 Strains**.
23. Answer **Question 5** in the **Data Sheets**.
24. To conclude the lab, answer **Question 6** in the **Data Sheets** about what you have learned from this lab activity.

DATA SHEETS: Analyzing Coronavirus Genomes

Name: _____ Date: _____

Laboratory Instructor: _____ Section: _____

PART A: COMPARING THE GENOMES OF TWO SARS VIRUSES

Question 1: What is the percent identity between the viral genomes of SARS-CoV-2 and SARS-COV-1?

Question 2: You have completed a BLAST using the SARS-COV-1 genome sequence that was published in 2003 and the SARS-CoV-2 sequence that was published in early 2020.

Your lab instructor will assign you two rows of bases in the genome sequence to analyze. Compare the subject sequence to the query sequence and complete the table below. You have been provided with an example below:

Row Number	Base Mutation to Adenine (A)	Base Mutation to Thymine (T)	Base Mutation to Guanine (G)	Base Mutation to Cytosine (C)	Insertion Mutation	Deletion Mutation
29	1	1	0	0	1	2

Data Table 1 Analysis of SARS-COV-1 vs SARS-CoV-2 Mutations

Row number	Base Mutation to Adenine (A)	Base Mutation to Thymine (T)	Base Mutation to Guanine (G)	Base Mutation to Cytosine (C)	Insertion Mutation	Deletion Mutation

Question 3: By pooling the observations by all the student groups, you now have a larger dataset to analyze. Look at all the base changes and answer the following questions.

Laboratory Activity: Analyzing Coronavirus Genomes

- (a) What type of base mutation occurs most frequently?

- (b) What type of base mutation occurs least frequently?

- (c) Do base insertions and deletions occur more or less frequently than base substitutions?

Continued on the next page.

PART B: IDENTIFYING MUTATIONS IN SARS-CoV-2 SEQUENCES

Question 4: You have performed a BLAST comparison between the Wuhan SARS-CoV-2 sequence and the coronavirus genome sequence from another patient. Complete Data Table 2 by recording information about the base mutations that have occurred.

NOTE: Record the position of the base mutation based on the **subject sequence** (Sbjct).

Data Table 2 Analysis of SARS-CoV-2 (Wuhan) vs. _____

Mutation: Base Difference (i.e. C → A)	Position of Mutation

Laboratory Activity: Analyzing Coronavirus Genomes

Question 5: You can now compare the original SARS-CoV-2 Wuhan sample to the genomes from other samples of SARS-CoV-2.

(a) What do you notice about the degree of similarity or difference between the original Wuhan SARS-CoV-2 sequence and the first recorded COVID case (Case 1) in Italy, Washington State, and California? Propose a hypothesis to explain these observations.

(b) What do you notice when comparing the BLAST results for Case 1 in Washington State and Case 5170 in the same state? Propose a hypothesis to explain these observations.

Question 6: What have you learned about coronavirus genomes by performing this lab activity? Provide a summary in 2-3 sentences.