Analysis of SARS-CoV-2 sequences reveals transmission path and emergence of $S^{D\,614G}$ mutation.

Mingjia Li[1], Nishita Prasad [2] , Dwight Hall[3], Huanmei Wu[3]
[1]Canterbury High School, Fort Wayne, Indiana. [2]Carmel High School, Carmel, Indiana. [3]Indiana University School of Informatics and Computing, Indianapolis, Indiana, USA

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313091

Abstract

The coronavirus disease 2019 (COVID-19) outbreak caused by SARS-CoV-2 virus began in Wuhan, China, and has spread quickly throughout the world. The development of vaccines for SARS-CoV-2 is difficult due to many obstacles, such as the lack of knowledge of important proteins, genes, and mutations of the viral genome. In this study, we selected and utilized 852 strains of COVID-19 from major countries in the National Center for Biotechnology Information (NCBI) global virus bank. The information of these strains was processed by using Nextstrain software, a program that provided a visual phylogenetic tree, transmission map, and diversity panel that explains entropy and number of mutations for each codon in the genome. The general data about the spread and evolution of COVID-19 supported the current knowledge that it began in China and spread throughout the country in an interrelated manner instead of a clear "patient zero" manner. A recent study reported that codon 614 on COVID-19 spike protein (S614) was an important codon for viral spread, specifically, a mutation from aspartic acid to glycine facilitated the spread of the virus. Therefore, we chose to geographically track this mutation during the spread of COVID-19 to investigate where it emerged and whether it can affect the spread COVID-19. Our results showed that the glycine mutation first emerged in France. Also, the transmission rates in France versus China, where the mutation was not prevalent, did reflect the hypothesized change in viral behavior.

Keywords: SARS-CoV-2, COVID-19 spike protein, Nextstrain

Introduction

The coronavirus disease 2019 (COVID-19) outbreak caused by SARS-CoV-2 virus began in Wuhan, China, and has spread quickly throughout the world. The development of vaccines for SARS-CoV-2 is difficult due to many obstacles, such as the lack of knowledge of important proteins, genes, and mutations of the viral genome. This study is aimed to investigate COVID-19's evolution and mutations through a timed phylogenetic tree built in Nextstrain, an open-source program that processes and visualizes public pathogen genome data for scientific and public health implication and gain a better understanding about the biology of COVID-19 and focused on specific codon mutations that have been tagged as important proteins to COVID-19 viral behavior.

Methods

Due to the political nature of COVID-19 test, it was important to ensure that the strains collected were accurate and credible. We defined this standard of strains as being from the NCBI database, having a complete or nearly complete sequence of over 29,000 bases, and being published by the NCBI consortium (NCBI Virus, 2020). Based on the standards, we selected and formatted 580 sequences from the countries in the NCBI database with the most strains available. All of the strains were SARS-CoV-2 strains from human hosts. We then added 580 sequences to the existing Nextstrain global database of human strains for a total of 997 sequences.

While running Nextstrain, we filtered and aligned the sequences. The filters removed any improperly formatted sequences as well as sequences with excessive absent bases. The remaining 852 strains were selected to create phylogenetic tree. After aligning the sequences, Nextstrain returned a list of base mutations, and these mutations are the basis of the phylogenetic tree. The remainder of the commands used to create the phylogenetic tree revolved around using Auspice and Augur, two subprograms of Nextstrain, to visually create the phylogenetic tree. Once this tree was exported, the visual phylogenetic tree, transmission map, and diversity index, were available (Figure 1).

Results

After inputting and processing the 852 SARS-CoV-2 strains, Nextstrain presents three panels with the phylogeny, transmissions, and diversity that visualizes the viral relationships. The Nextstrain features, supported by Auspice and Augur, are demonstrated in Figure 2.

The phylogeny feature represents the evolutionary relationships among different SARS-CoV-2 sequences on the molecular level by the observed mutation patterns. The x-axis depicts the degree of difference in time or genetic divergence, while the y-axis spreads the sequences for better visualization. To measure the differences over time, x-axis reflects each sampling date, with the tree's tips also representing the sampling date. Nextstrain infers the internal node dates, the "missing cases" that are intermediate and unsampled, through their descendants' sampling dates and viral mutation rates. Even though the sequences may have the similar mutations, the tree by date accounts for both the mutation and spread rates. The genetic divergence differences compare the number of mutations to the estimated start of the outbreak and organize the number of changes in the genome. The identical sequences are grouped together on the tree.

The transmission networks represent how the pathogens, SARS-CoV-2 in this instance, spread through rapid replication from one host to another. The genome sequences help inter sections of the transmission tree. Through replication and spread, the pathogenic genome experiences innumerable replication cycles, inevitably causing mutations, which are random copying mistakes. The sequences with similar mutations are more closely related than others, which allows the program to organize them into groups of associated viruses that are a part of the same transmission chains. Nextstrain colors the phylogenetic tree by the sample location and suggests viral spread throughout the outbreak. Therefore, these genome alterations accumulate, help track the spread, and establish the pathogen's routes and interactions.

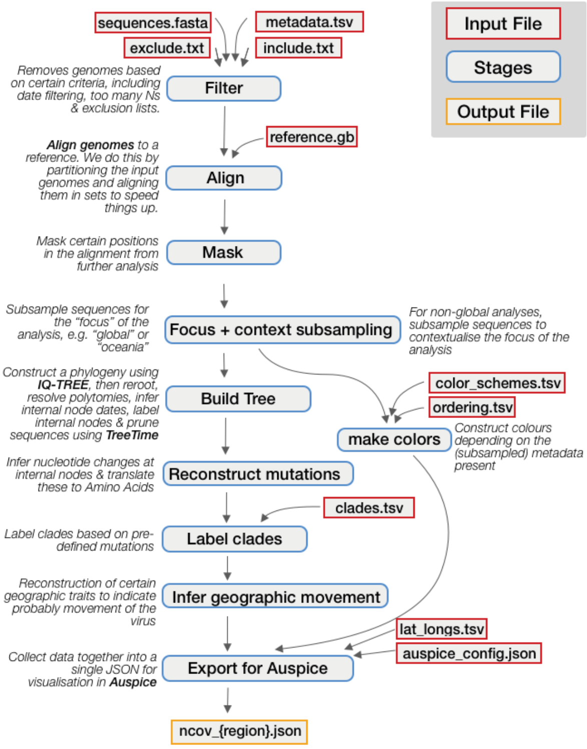

sequences.fasta
metadata.tsv
exclude.txt
include.txt
Input File
Stages
Output File
Removes genomes based on certain criteria, including date filtering, too many Ns & exclusion lists.
Filter
reference.gb
Align genomes to a reference. We do this by partitioning the input genomes and aligning them in sets to speed things up.
Align
Mask certain positions in the alignment from further analysis
Mask
Subsample sequences for the "focus" of the analysis, e.g. "global" or "oceania"
Focus + context subsampling
For non-global analyses, subsample sequences to contextualise the focus of the analysis
Construct a phylogeny using IQ-TREE, then reroot, resolve polytomies, infer internal node dates, label internal nodes & prune sequences using TreeTime
Build Tree
color_schemes.tsv
ordering.tsv
make colors
Construct colours depending on the (subsampled) metadata present
Infer nucleotide changes at internal nodes & translate these to Amino Acids
Reconstruct mutations
clades.tsv
Label clades based on pre-defined mutations
Label clades
Reconstruction of certain geographic traits to indicate probably movement of the virus
Infer geographic movement
lat_longs.tsv
auspice_config.json
Collect data together into a single JSON for visualisation in Auspice
Export for Auspice
ncov_{region}.json

Figure 1. *Steps to Process the SARS-CoV-2 data with Nextstrain.* This figure shows the steps in inputting, processing, and outputting the data with the Nextstrain program.

Demonstrated by the "diversity" panel, the program uses the variations in the genome, including mutations in nucleotides and amino acids, to construct the phylogenetic tree. The bar-chart has a horizontal axis that includes all the viral genome sites, which is approximately thirty thousand, and a vertical axis that reveals how much variability is at each site. Although there is no reason to believe that each alteration is a functional mutation, Nextstrain encompasses a feature that colors the tree by a mutation since the program uses the changes to organize and define the relationships of the sequences to build the tree.

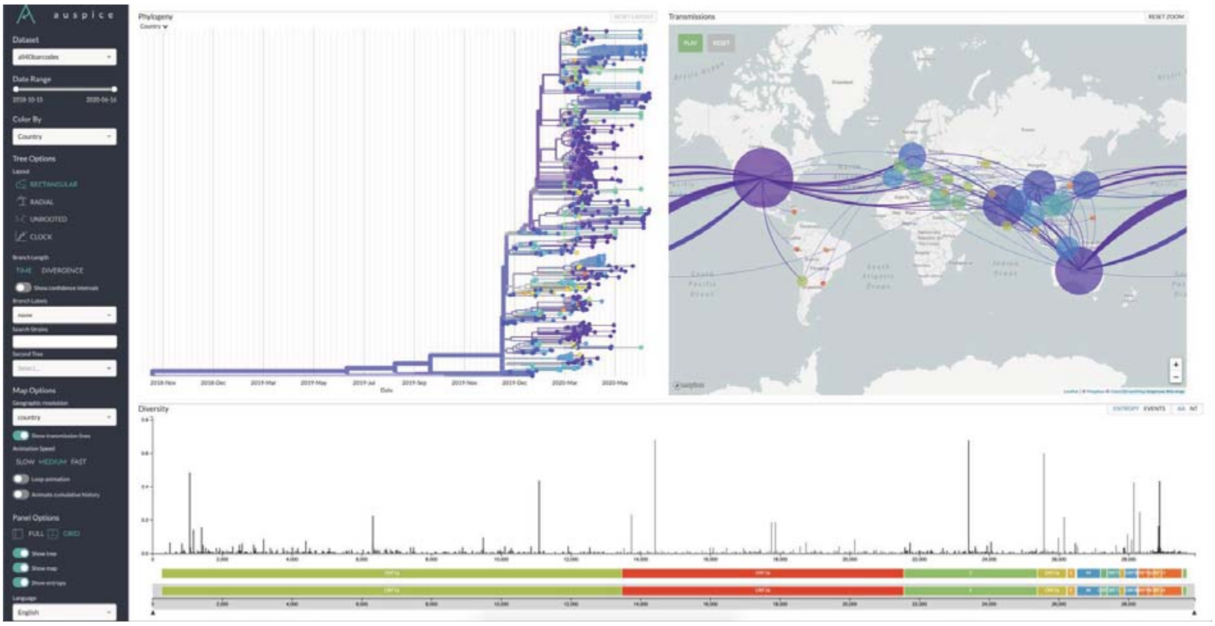
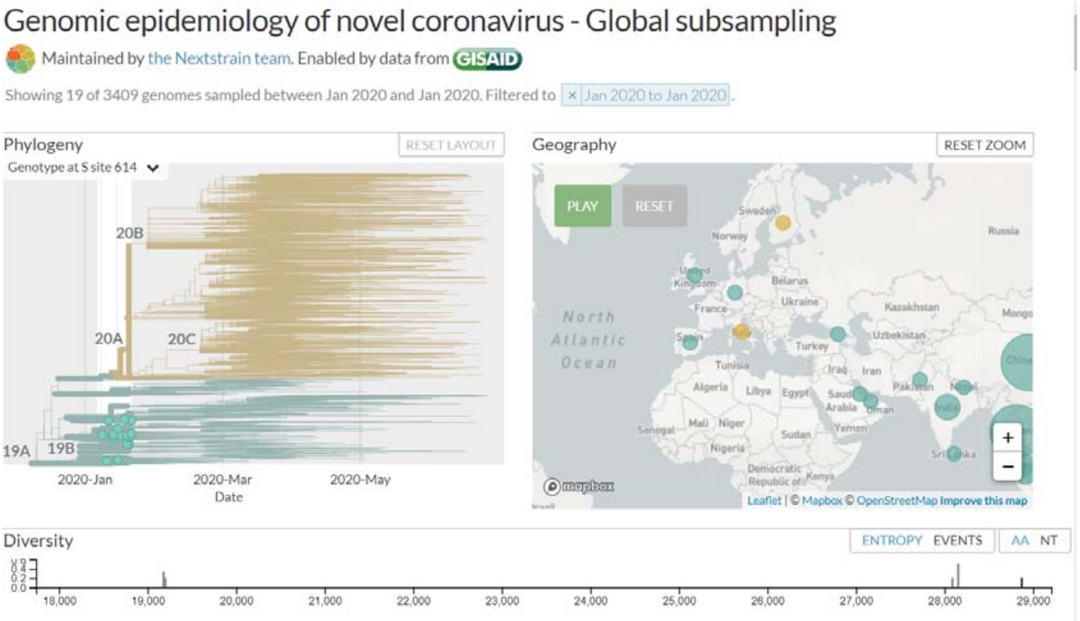


Figure 2. *Nextstrain panels after processing 852 SARS-CoV-2 strains.* This figure shows the panels of Nextstrain after the data of 852 strains was inputted and processed. In the first row, the phylogenetic tree and transmission map are pictured. In the second row, the diversity panel is pictured with "entropy" selected, so it is a measure of uncertainty of each of the codons on the genome. A visual representation of the genome is pictured below the diversity index as its x-axis.

The Nextstrain mapping reflects the common knowledge about the spread of COVID-19, in that it began its path from China with its inferred starting date in October 2019. When the date is unknown, Nextstrain assigns the outbreak start date by using the available dates from each sample and node in the phylogenetic tree. This "root" of the tree represents the "most recent common ancestor" for all of the SARS-CoV-2 sequences available thus far. With the virus beginning in China with 100% confidence, the first transmissions occurred in Germany and Australia, followed by Thailand, South Korea, Taiwan, Israel, and Nepal in December 2019. The first transmission to India was in In January 2020. By mid-January 2020, the virus travelled to France, Egypt, Hong Kong and Japan. Towards the end of the month, the United States, Italy, Pakistan, Chile, Kazakhstan, and Guam had SARS-CoV-2 cases. In February 2020, Spain, Peru, Greece, and Saudi Arabia had transmissions of the virus. In early March 2020, Brazil, Turkey, and Sri Lanka faced the virus. By mid-March 2020, Jamaica, Iran, Malaysia had COVID-19 cases. By the end of the month, there were SARS-CoV-2 strands in Uruguay, Bangladesh, and Timor-Leste. Between April and June, there were no transmissions, but the viral cases grew and diminished throughout the months, including transfers from countries other than China.

From the diversity panel, we also found that a significant majority of codons had only one mutation. Furthermore, there were only 3 codons with over 4 mutations. Although further studies are necessary to understand whether these codons are significant to COVID-19's function, the overarching summary is that COVID-19 is relatively stable and doesn't have a large number of mutations. Among the mutations, the majority are not prevalent and likely occur due to replication errors or sequencing errors.



Genomic epidemiology of novel coronavirus - Global subsampling

Maintained by the Nextstrain team. Enabled by data from GISAID

Showing 19 of 3409 genomes sampled between Jan 2020 and Jan 2020. Filtered to × Jan 2020 to Jan 2020.

Figure 3. *ncov phylogenetic tree and transmission map*. The above image is the phylogenetic tree and transmission map panels of the data produced by the Nextstrain ncov build. They show that the glycine mutation emerged in Italy and Finland in mid-late January.

Discussion

Using Nextstrain which provided us insight into both the general COVID-19 transmission pathway and the specific S614 gene mutation, we are able to map general data about the spread and evolution of COVID-19 supported the current knowledge that it began in China and spread throughout the country in an interrelated manner instead of a clear "patient zero" manner, we also tracked genes that have a significant impact on COVID-19's structure and function. For example, researchers of the Scripps Research Institute documented that S614 gene on the COVID-19 spike protein had a mutation of aspartic acid (D) to glycine (G) (Zhang et al., 2020). The glycine made COVID-19 more stable, and thus made it spread faster. We tracked the S614 gene spread to visualize where the mutation began. We found that the glycine mutation began in France and spread through Europe, then to the rest of the world. This is corroborated by much higher spreading rate in France than in China's first exponential growth (New York Times, 2020). This is significant because it shows that the mutation was not present in China during the virus's initial spread and that it presented itself in the viral behavior, or its spread.

SARS-CoV-2 virus has RNA as its genetic material that contains nucleotides in codons as its building blocks. Upon infecting a cell and making copies of its genetic instructions, viruses are constantly changing and generally do not have the required machinery to proofread their replicated RNA string for errors, resulting in various genetic accumulations (Garcia de Jesus, 2020). Unlike most RNA viruses, the *Nidovirales* order, which the Coronavirus genus belongs to, has the proof-reading capability, allowing them to have the largest RNA genomes. The order has a complex machinery for RNA synthesis that is operated by nonstructural proteins (nsps) to produce cleavage products of the ORF1a and ORF1b viral polyprotein to coordinate virus replication and transcription. With a high homology for SARS-CoV-2 RNA-dependent RNA polymerases (RdRps) used in replication and transcription, SARS-CoV-2 has conserved the machinery, demonstrating its importance (Pachetti et al., 2020). Despite these tools to prevent mutations, these changes can accumulate that may be slower than other RNA viruses without the enzymes such as influenza that results in mutations including S614 mutation. As stated before, the mutations are useful in tracing the virus throughout the world. For SARS-CoV-2 cases, researchers have been analyzing the viral path since the release of the first coronavirus genetic sequence in January 2020, allowing them to sequence the RNA changes as it spreads and infects more people even if they do not alter the protein (Garcia de Jesus, 2020).

Compared to Nextstrain novel coronavirus (ncov) built with 3,409 strains, our initial results regarding the spread of COVID-19 were very similar to the transmission map for Nextstrain ncov. For the S614 mutation, however, there was a slight difference. We found that France was the first location for glycine in S614, while the ncov build showed Italy and Finland as the primary location. This is likely due to smaller sample size of strains in our study, but in both situations the emergence was in Europe, which had a higher spread rate than in China (Figure 3).

Additionally, although there are no other publications about S614 transmissions, the information about the COVID-19 spread in France and China did correlate with the aspartic acid to glycine mutation. Specifically, it makes sense that France, with the glycine mutation arising, would have a higher rate of cases compared to China due to the nature of the S614 mutation.

There are several limitations of our work, first, the main limitation is the data collection. Although some countries have a large number of cases, they do not necessarily have a large number of strains in the NCBI database. For example, the United Kingdom, Brazil, and Russia are three countries that have a large number of cases but have very few strains: 0, 5, and 9 strains in the database, respectively. This occurrence does leave these countries out of analyses and interpretations, especially if they do not have any strains in the database. Although there was an attempt to select strains from nearby or within all countries with a significant number of COVID cases, the NCBI database limited the possibility. Second, in the countries with many strains available for download there were strains from the same date with consecutive accession numbers, potentially signifying that the data does not have significant differences and cannot demonstrate the evolution of the virus within the country. The date ranged from December 2019 to June 2020. To combat this, during sequence selection, we did try to ensure that the strains selected were not all from the same date. However, in the case where the only sequences from a country were from one date, we kept that in consideration when looking at our phylogenetic tree and analysis.Third, the spreading information we used is limited by the data provided by the countries, which is not necessarily accurate due to either lack of testing or lack of reporting accurate data.

In conclusion, Using the phylogenetic tree and transmission map created by the Nextstrain algorithm help understanding the overall spread of Covid-19. We found that D614G mutation on the COVID-19 spike protein emerged in France in early January and spread throughout the world. This is correlated to increased rate of spreading in Europe compared to China in the early stage of infection.

For future work, mutation emergence should be further investigated, especially in conjunction with epidemiology and microbiology. For example, once more codons and proteins are understood, they can be tracked to understand virus evolution. If these codon mutations are important to vaccine development, using Nextstrain builds can show how they affect the countries where the mutation was prevalent. Additionally, regarding D614G mutation, our methodology should be reproduced in the future with a different, larger set of data specifically from Europe and Asia to have a better understanding of where this mutation started.

Acknowledgement

This project was supported by Indianapolis Project STEM and sponsored by Indiana CTSI Project STEM for high school students. Thanks to Mr. Elmer Sanders, Dr. Huanmei Wu and Wu Lab group for all the support during the Summer 2020 session.

References

1. Centers for Disease Control and Prevention. (2020, May 28). *Using antibody tests for COVID-19*. Centers for Disease Control and Prevention. Retrieved July 9
2. Centers for Disease Control and Prevention. (2020, June 16). *How COVID-19 spreads*. Centers for Disease Control and Prevention. Retrieved July 12, 2020
3. Cucinotta, D., & Vanelli, M. (2020, March 19). WHO declares COVID-19 a pandemic. *Acta Biomedica*. Retrieved July 9, 2020
4. Garcia de Jesus, E. (2020, May 26). *Is the coronavirus mutating? Yes. But here's why you don't need to panic*. Science News. Retrieved July 12, 2020
5. Harvard Medical School. (2020, July 10). *Coronavirus resource center*. Harvard Health Publishing. Retrieved July 11, 2020
6. NCBI Virus. (2020). NCBI Virus. Retrieved July 15, 2020.
7. New York Times. (2020, July 13). *France coronavirus map and case count*. The New York Times. Retrieved July 13, 2020
8. Pachetti et al., M. (2020, April 22). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine*. Retrieved July 12, 2020.
9. Pathan et al., R. K. (2020, September). Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model. ScienceDirect. Retrieved July 6, 2020.
10. World Health Organization. (2020, June 30). Timeline of WHO's response to COVID-19. World Health Organization. Retrieved July 12, 2020.
11. Yale Medicine. (2020). COVID-19 (coronavirus disease 2019). Yale Medicine. Retrieved July 9, 2020.
12. Zhang et al., L. (2020). The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. Scripps Research. R2020.
13. Zhou et al., Y. (2020, February 29). [Analysis of variation and evolution of SARS-CoV-2 genome]. LitCovid