Examination of COVID-19 Dynamics through Phylogenetic Assembly and ISM Labeling

Anthony Knesis, ajk364@drexel.edu, Jacob Maier, jtm355@drexel.edu
John R Seitz, jrs488@drexel.edu, Dhairav Shah, dss322@drexel.edu
Wenhan Tan, wt99@drexel.edu, Alexander Tweed, adt55@drexel.edu

Advisor: Gail Rosen, Ph.D: glr26@drexel.edu

FINAL REPORT

Abstract

The unprecedented impact of the coronavirus pandemic on global society has accelerated genetic and immunological research into the SARS-CoV-2 virus that causes the COVID-19 respiratory illness. Rapid studies across every field of medicine are being conducted to assess the evolution and spread of this virus as well as its response to various treatments (Cascella et al. 2020). The growing number of cases and fatalities makes this research dynamic and time-sensitive; moreover, updated analyses are critical as more data is collected daily. Bioinformatics provides a unique lens to study this virus by examining its genetic composition and evolution. The objective of these inquiries is to understand the geographic and temporal spread of the SARS-CoV-2 virus from its origins in Wuhan, China. In this study, the goal is to generate a phylogenetic tree based on the GISAID database and label the tree nodes with Informative Subtype Markers (ISMs). These analyses can motivate policies on safety, provide predictions on future hotspots and outbreaks, and suggest avenues for treatment and control strategies. The results show the ability to track the virus temporally through its mutations alongside the geographical movement as well as through color-coded ISMs.

Literature Review

In their paper, *Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak*, Zhang et al. suggests that a type of coronavirus found in Malayan pangolins in 2019 (Pangolin-CoV) may be a potential origin of SARS-CoV-2. By performing sequence alignment between the pangolin coronavirus, the bat strain of CoV (RaTG13), and genomic data from the initial six reported cases at the seafood market, it was determined that Pangolin-CoV retains significant functional similarity to SARS-CoV-2 and RaTG13 to be considered its common phylogenetic ancestor. Although the bat strain of CoV has higher genetic similarity to SARS-CoV-2 (96.2%), history of interspecies disease transmission favors intermediate hosts such as civets or camels. This phylogenetic relationship was supported using nucleotide alignment as well as RNA polymerase similarity and protein expression. Evaluation of the S-protein functionality in Pangolin-CoV also provides insight into the possible mechanism that SARS-CoV-2 uses to facilitate entry into host cells as shown in Figure 1. Nonetheless, more sequence alignment must be performed to find intermediate SARS-CoV-2 hosts to reduce transmission (Zhang, Wu, and Zhang 2020).

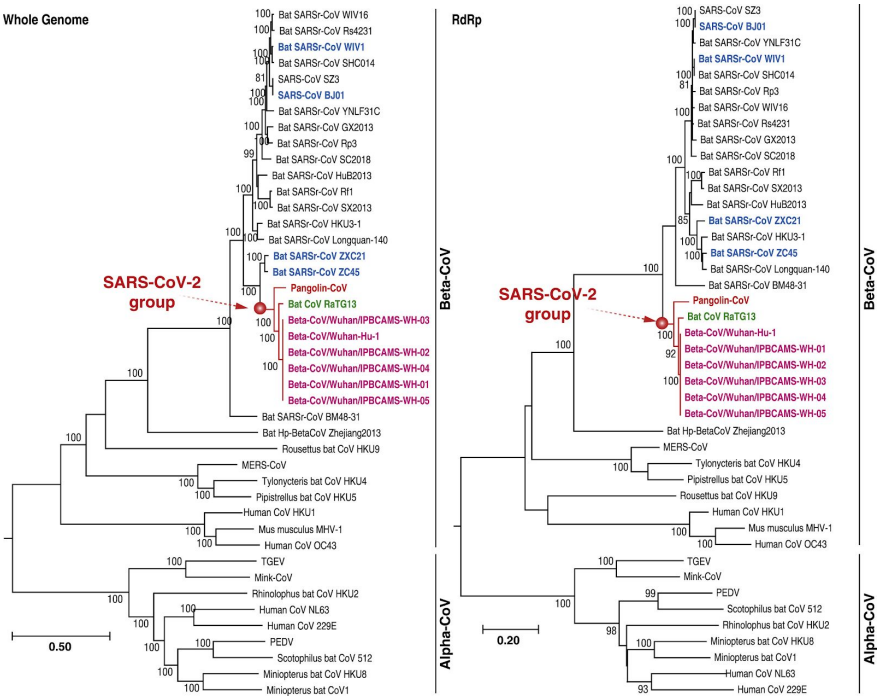

Figure 1. Phylogenetic Relationship of CoVs Based on the Whole Genome and RdRp Gene Nucleotide Sequences. (Zhang et al. 2020).

In *Phylogenetic network analysis of SARS-CoV-2 genomes*, Forster et al. presents a phylogenetic network of 160 largely complete SARS-CoV-2 genomes to determine the virus' evolution before widespread migration. Their assessment of its animal origins confirmed that the Pangolin-CoV is poorly conserved with SARS-CoV-2, suggesting evolutionary mutation, while Bat-CoV retains the highest genomic similarity. In Figure 2, the network shows a mix of existing ancestral viral genomes alongside their mutated child genomes. Three core types of viral variants were identified: two of which were found mostly in Europeans and Americans, while the third was found mostly in East Asia. The predominant American strain was observed to derive from the ancestral (Bat-CoV related) node, linked to individuals with previous residence in Wuhan. The European strain is a mutation of the prominent East Asian subtype, but has little presence in mainland China. Significantly, the East Asian subtype itself appears to be environmentally adapted to the East Asian population, as evidenced by its delayed evolutionary mutation outside China. The reconstructed tree, which includes 100 different strains of SARS-CoV-2, illustrates the virus' alarming rate of mutation, and provides a quantitative record of early infection paths (Forster et al. 2020).

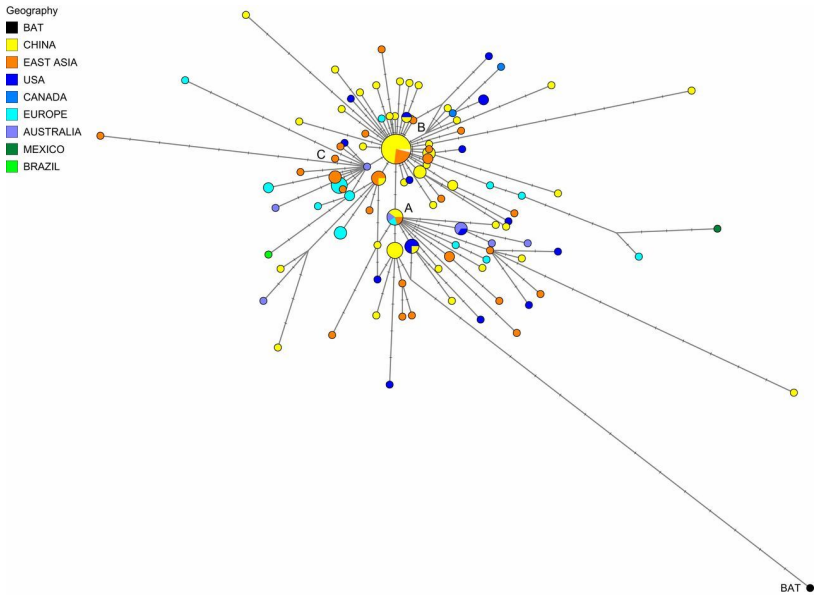

Figure 2. Phylogenetic network of 160 SARS-CoV-2 genomes. (Forster et al. 2020).

Further analysis of the distinguishing mechanisms of the spread of SARS-CoV-2 is performed in Brufsky (*Distinct Viral Clades of SARS-CoV-2: Implications for Modeling of Viral Spread*) using viral strains captured from Nextstrain. Like Zhang et al., Brufsky identifies mutations in the spike (S) protein as a driving factor in viral transmission. Moreover, he examines how different strains of CoV on the West and East coasts of the United States have resulted in differing mortality rates. The West Coast variant, he assesses, which derives from an ancestral subtype (China) with aspartic acid in this spike protein, has experienced lower rates of infection. The East Coast variant, however, which has spread to New York by way of Europe, contains a point mutation of aspartic acid to glycine. Pathogenetic theory and electron microscopy indicate that this glycine expression (which is highly conserved), leads to higher virulence and increased human-to-human spread. Due to biases in tree assembly, Brufsky also suggests a more rigorous analysis based on an organization of viral clades (Brufsky 2020) as shown in Figure 3.

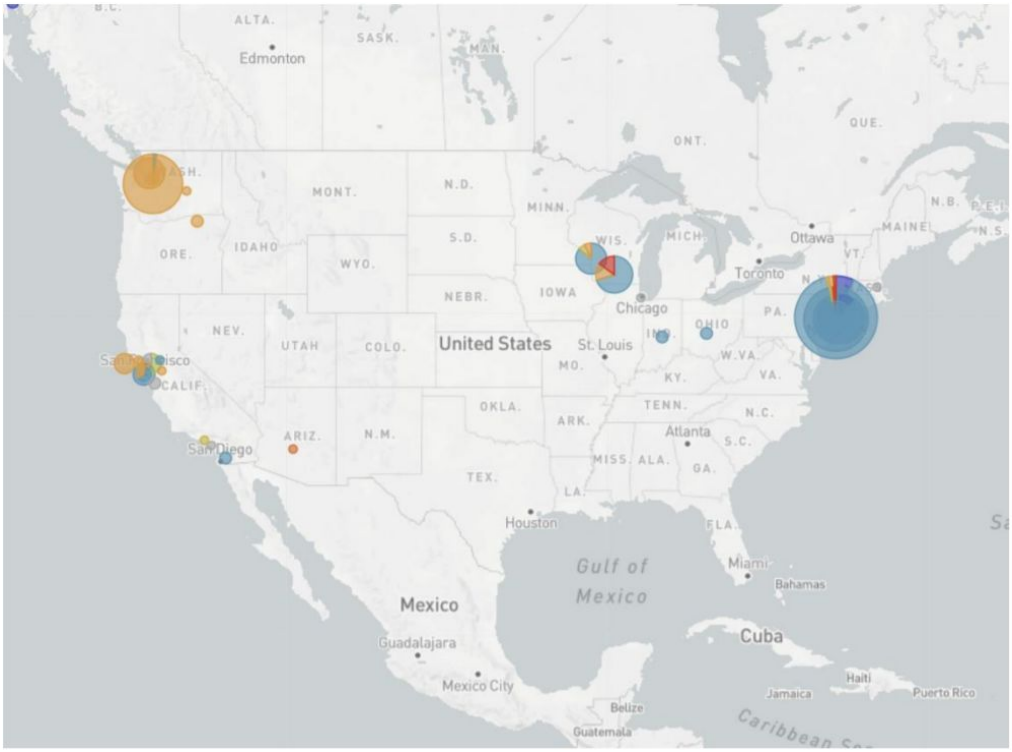

Figure 3. Domain Viral Clades in the United States. Clade B1 is in orange and Clade A2a is in blue. (Brufsky. 2020.)

Existing analyses of the distribution of SARS-CoV-2 subtypes have taken advantage of the influx of newly reported cases. The NextStrain project, which pairs a suite of bioinformatics tools with robust visualization abilities, is already assembling phylogenetic trees for the novel coronavirus based on geographical and temporal information about cases (Hadfield et al. 2018). Furthermore, research performed at the Ecological and Evolutionary Signal-Processing and Informatics Laboratory (EESI) has identified marker sequences in the SARS-CoV-2 genome that uniquely characterize viral subtypes using entropy-based measures (Zhao, Sokhansaj, and Rosen 2020). These informative subtype markers (ISMs) provide instructive barcode sequences for tracking the evolution of the coronavirus pandemic. This project is intended to integrate these complementary analyses by cross-referencing the ISMs against the branches of a phylogenetic tree. Of particular interest is determining the clustering of ISMs in phylogeny and the existence of clades for these markers. Such analysis is intended to facilitate a greater understanding of the spread of the SARS-CoV-2 virus and compare the observed mutations in its genetic sequence with the geotemporal distribution of subtypes as tracked through reported cases.

Methods

Complete genetic data of coronavirus cases is available from the GISAID database, an international accessible repository hosted by the Federal Republic of Germany which stores genetic information on SARS-CoV-2 and other viruses. The database also contains relevant sample metadata including the location of cases, the time of diagnosis, the originating lab, etc. As of June 12, the database contains over 45,000 complete sequences for the novel coronavirus. For the purposes of analysis, processing was performed on a static dataset dated May 17th 2020, and itself encompasses approximately 22,000 sequences. Due to the increasingly large number of genomic sequences that have been sampled, a serious obstacle for analysis was the computational complexity of performing sequence alignment and tree assembly. However,

certain computational steps have been drastically accelerated due to Proteus, the computing cluster available at Drexel University. Moderated by the University Research Computing Facility (URCF), the Proteus cluster is a robust computing platform, containing a combination of Intel and AMD CPUs, Nvidia GPUs, and several pre-installed programming libraries.

One preliminary method of reducing the size of the dataset was to consider only sequences submitted in a specific geographic area. This method involved the removal of sequences sampled from countries outside the United States, or those which contain large numbers of gaps. This initial filtering step was originally performed in BioPython, and produced a dataset with 185 unique sequences exclusively from the state of Connecticut. These sequences were then aligned using MAFFT on the CIPRES Scientific Gateway, and assembled in a phylogenetic approximation using the FastTree algorithm, as will be discussed below. This tree structure was visualized using the Interactive Tree of Life viewer, a web-based platform for annotating and viewing phylogenetic trees. Figure 4, which shows the developed tree for only 185 coronavirus cases exclusively in Connecticut, highlights the difficulty in interpreting trees with large numbers of sequences.

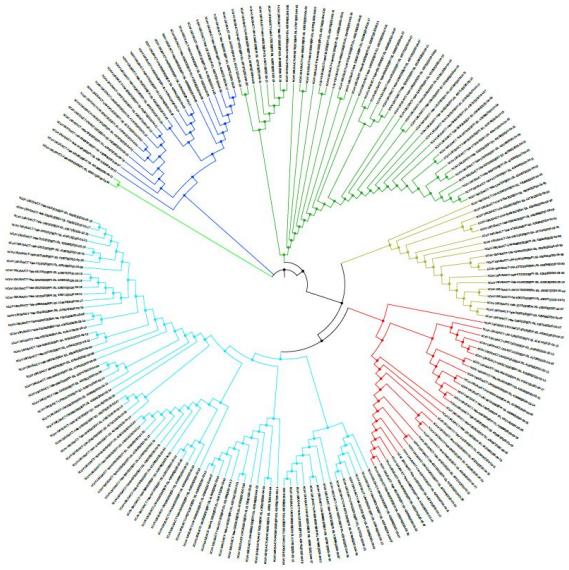

Figure 4: Phylogenetic tree for 185 samples in USA-CT, with false coloring.

Eventually, the filtering method was finalized to undersample the available sequences in each geographical and temporal region in order to retain a statistically significant amount of global diversity without excessive redundancy. Using a filtering process outlined by Nextstrain, only 100 sequences from each geotemporal region were preserved for alignment. Those with improper date annotations and less than 25,000 base pairs were also removed. The Nextstrain filtering procedure utilizes their custom bioinformatics toolkit known as Augur. Although Augur contains many functions applicable for large data pipelines, only the filtering step was leveraged in this analysis. This process (with 2000 sequences sampled per region), was utilized in Zhao et al. to calculate informative subtype markers. Our adjusted process truncated the dataset from 22,000 to 9,000 sequences. However, this size was still slightly cumbersome for purposes of visualization.

The alignment stage is the backbone of our project, the results of which are critical for producing credible analyses. Sequence alignment algorithms attempt to track and match similarities between nucleotide sequences, thereby identifying regions of variability and mutation. Most genetic research in microbiology leverages the genes of 16S ribosomal RNA, which mutates at a conservative rate suitable for alignment and phylogenetic comparison. Based on the size and scale of the genetic data available, this research utilized appropriate computational tools for multiple sequence alignment. Two efficient algorithms for sequence alignment include MUSCLE and MAFFT. MUSCLE (Multiple Sequence Comparison by Log-Expectation) uses distance matrices and iterative refinements to finalize alignment and can align up to 500 sequences (Edgar 2004). MAFFT, on the other hand, performs Multiple Alignment using Fast-Fourier Transforms by leveraging frequency-domain information about nucleotide arrangements. This method yields a much higher throughout, allowing alignment of up to 30,000 sequences (Katoh et al. 2002). Due to the volume of genetic coronavirus data and the importance of efficient data processing, our data pipeline used MAFFT for sequence alignment. The original implementation used the computing resources of the CIPRES Scientific Gateway to perform the multiple sequence alignment. The final method utilized the integrated MAFFT library in Proteus for convenience.

Following the alignment step, high entropy regions of the aligned sequences were analyzed to produce the informative subtype markers (ISMs) as described in Zhang et al. The processing pipeline to identify informative subtype markers in aligned sequence data has been adapted for this research. For reproducibility, the BioPython implementation for this method has been provided in the research group's GitHub repository as a series of scripts. The final output of the Proteus scripts are the excised metadata which contain the dominant ISM of each sequence. A Jupyter notebook was subsequently created to label and filter the original Nextstrain metadata, such that each viral identifier in the dataset is paired with its ISM. Due to gaps and substitutions, there are over 250 unique ISM strings in the available data. Due to the frequency of only certain ISMs, it can be assumed that only the first 20 subtype markers will be of relevance for analysis, as shown in Zhang et al.

Concurrently with the ISM analysis, a phylogenetic tree of the multiple sequence alignment was produced using the FastTree algorithm. Due to its computational resources, the CIPRES gateway was the primary tool for this stage. Although the RaXML algorithm generally produces more accurate tree structures, the size and latency of the processing step precluded its use. However, bootstrapping steps incorporated into the FastTree algorithm may have contributed to a more stable and coherent structure. Visualization of the tree was conducted in the Interactive Tree of Life web viewer, which has its own format for node labelling. Using the labelled metadata, unique colors were assigned to the most prominent ISMs to highlight the concordance of certain viral strains by region and ISM.

The final step of analysis is a quantitative comparison of the produced tree and the global data produced by the Nextstrain team. Although the Nextstrain data is updated daily with every new case submitted to GISAID, status situation reports provide snapshots of the viral evolution across time. The closest available dataset for comparison summarized the coronavirus cases up to May 15, 2020. Using the "ape" and "phangorn" bioinformatics packages in R, the similarities of the two phylogenetic trees were assessed using the Robinson-Foulds distance (symmetric), path difference, and branch score.

Results

The results from Figure 5 show how ISMs are labeled on the phylogenetic tree. There are more than 250 ISMs extracted from the tree, which are ranked due to their number of occurrences. In Figure 7, the 20 most abundant ISMs within the total dataset are shown. Interestingly, the 20 ISMs are almost the same as the one in the paper (Zhao et al. 2020). The only difference is that more data samples are available for comparison, skewing the distribution toward the increased number of cases, especially in the United States. As for now, it seems like extracted ISMs are set at least for a few months since the first human case is reported in Wuhan, China. Figure 6 is a color-coded version of Figure 5. The 20 most abundant ISMs have been colored in order to provide a better and direct visualization.

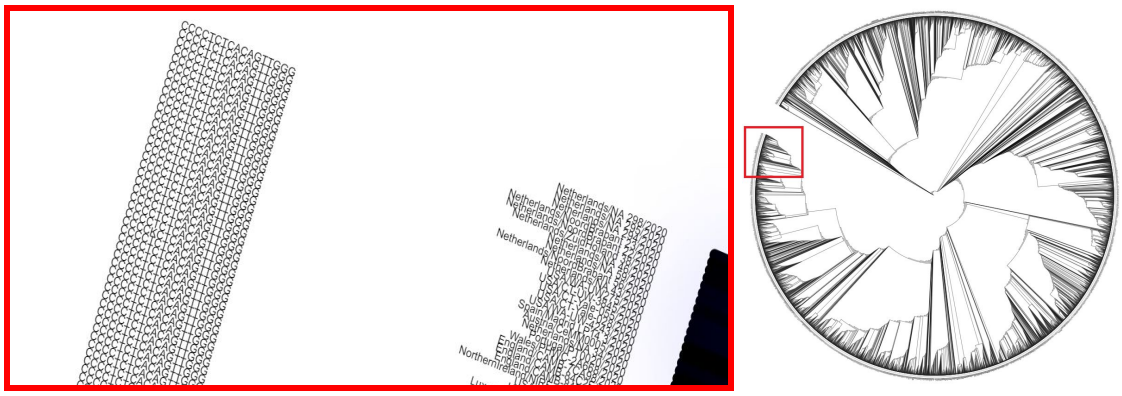

Figure 5: Phylogenetic tree containing all international sequences as of 5/17/20 w/ close up showing ISM.

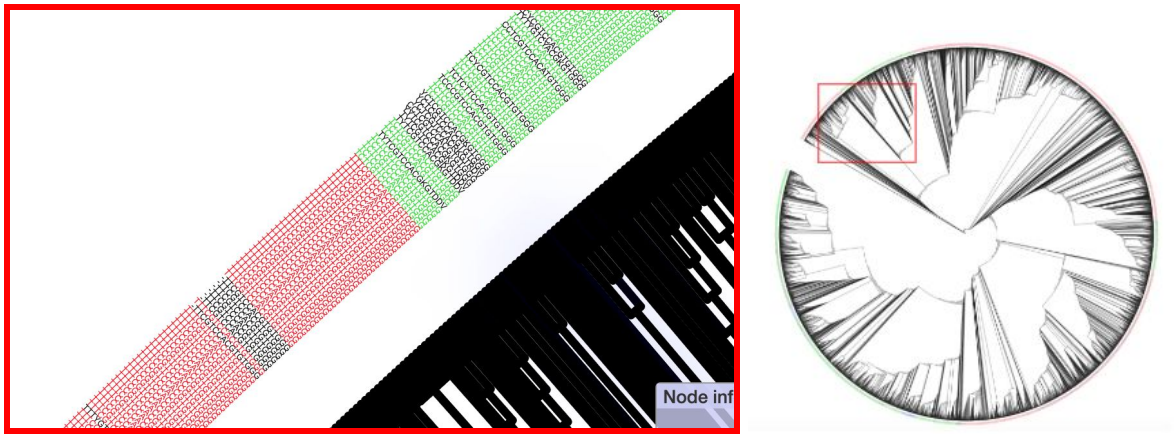

Figure 6: Phylogenetic tree containing all international sequences as of 5/17/20 w/ close up showing color coded ISM.

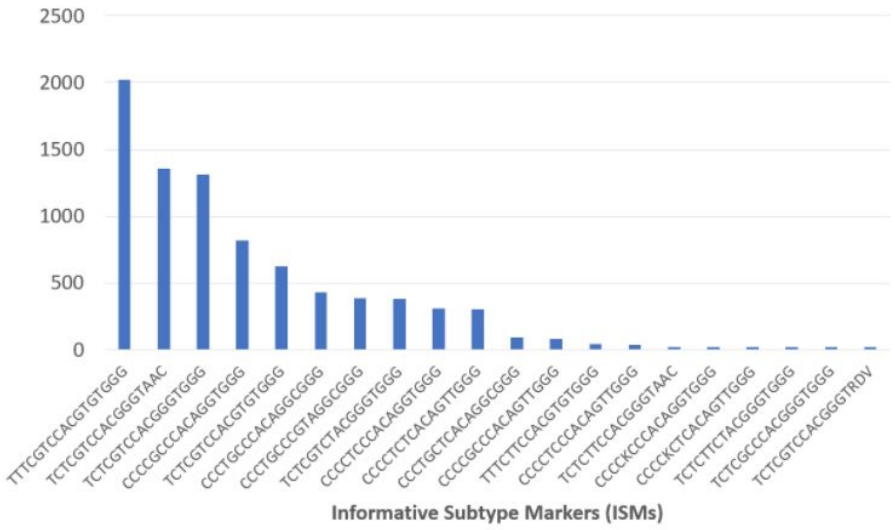

Figure 7: Number of sequences containing the 20 most abundant ISMs within the total data set.

ISM	Color Code	Geographic Region
TTTCGTCCACGTGTGGG	#FF0000	U.S.A., Canada, Europe, South America
TCTCGTCCACGGGTAAC	#A70000	Europe, Japan, Singapore, U.S.A.
TCTCGTCCACGGGTGGG	#FF5454	Europe, East Europe
CCCCGCCCACAGGTGGG	#19DC19	Europe, Middle East, U.S.A.

TCTCGTCCACGTGTGGG	#3ADF3A	U.S.A., Middle East
CCCTGCCCACAGGCGGG	#60DF60	U.S.A., Middle East, Mainland China
CCCTGCCCGTAGGCGGG	#47C847	North America
TCTCGTCTACGGGTGGG	#26AE26	Europe, Middle East
CCCCTCCCACAGGTGGG	#16C216	Mainland China, India, Singapore
CCCCTCTCACAGTTGGG	#13A413	Australia, Europe
CCCTGCTCACAGGCGGG	#0D62EE	U.S.A., Europe
CCCCGCCCACAGTTGGG	#2976F3	Spain, U.K.
TTTCTTCCACGTGTGGG	#4084F3	Hong Kong, Singapore, Mainland China
CCCCTCCCACAGTTGGG	#4084F3	U.S.A., Europe
TCTCTTCCACGGGTAAC	#4084F3	Hong Kong, Norway, Spain
CCCCKCCCACAGGTGGG	#4084F3	Russia, Europe, Australia
CCCCKCTCACAGTTGGG	#4084F3	India
TCTCTTCTACGGGTGGG	#4084F3	U.S.A., U.K.
TCTCGCCCACGGGTGGG	#4084F3	Congo
TCTCGTCCACGGGTRDV	#4084F3	Belgium, Mainland China, Spain

Table 1. The 20 most abundant ISMs with color code and geographic region.

Robinson-Foulds Distance	15584.00
Branch Score Distance	120.84
Path Difference	913079.09
Quadratic Path Difference	68300.67

Table 2: Metrics evaluating the similarity of the analyzed tree to the closest Nextstrain model.

Discussion

The goal of identifying and labeling all the ISMs on a phylogenetic tree has been achieved, as can be observed to some extent in Figures 5 and 6 containing two trees and ISM labeling with and without color coding. Identifying the ISM's on a phylogenetic tree provides the opportunity to track the virus temporally through its mutations alongside its geographical movement. Visual assessment of the color codes on the final tree illustrate that the most dominant ISMs are very strongly linked to the observed clades on the phylogenetic tree. Final results were achieved following the outlined four processing steps, namely, Data Collection, Alignment, Tree Building and ISM Labeling.

Assessment of the quantitative comparison of the Nextstrain and ISM trees, as shown in Figure 2, is ambiguous. The observed differences in the numeric scores (which are significant), can be justifiably explained by numerous factors, including the differences in filtered labels, mismatches in data set size, and the two-day latency between this research data and the Nextstrain data. Supplemental analysis, which reruns the entire pipeline using contemporary and exactly similar data for both sources, can fully account for the phylogenetic similarity.

Significant analytical bottlenecks were encountered as a result of improper visualization of the phylogenetic tree. Of the publicly available web viewers, none were sufficient for efficiently rendering the 9,000 sequences in the filtered dataset. Access to more powerful computational and graphic resources may have facilitated a more rigorous evaluation.

Future Goals

While tree visualization methods will continue to be investigated, an important corrective step for improving accessibility and interpretability will be to further condense the available data by using fewer sequences. Nextstrain's interactive SARS-CoV-2 phylogenetic tree, for instance, only uses ~4,000 sequences, but still includes global virus data for the last five months of reported cases. Optimal data filtering will help to accelerate the runtime of

computational alignment and tree-assembly processes. It may also permit the use of more accurate inference methods, such as RaXML for tree construction. Although FastTree is generally less accurate than RaXML, as the name suggests it can be performed substantially faster. As the size of current datasets make tree assembly computationally intensive even with FastTree, utilizing RaXML for more accurate trees is currently intractable.

Another area of future study concerns the temporal characteristics of the novel coronavirus as a function of ISM and geographic region. Certain socio-political conditions have facilitated the spread of the virus person-to-person; however, it remains to be seen if certain strains are inherently more contagious than others. A rigorous comparison of geographic data with the genetic identifiability of the ISM labelling may address particularly dangerous viral strains.

References

Brufsky, A. (2020), Distinct Viral Clades of SARS‐CoV‐2: Implications for Modeling of Viral Spread. J Med Virol. Accepted Author Manuscript. doi:10.1002/jmv.25902

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2020). Features, Evaluation and Treatment - Coronavirus (COVID-19). Treasure Island, FL: StatPearls Publishing.

Forster, P., Forster, L., Renfrew, C., & Forster, M. (2020). *Phylogenetic network analysis of SARS-CoV-2 genomes*. PNAS. https://www.pnas.org/content/117/17/9241

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. doi: 10.1093/nar/gkh340

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., … Neher, R. A. (2018). *Nextstrain: real-time tracking of pathogen evolution*. Bioinformatics. *34*(23), 4121–4123. doi: 10.1093/bioinformatics/bty407

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. doi: 10.1093/nar/gkf436

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, *26*(7), 1641–1650. doi: 10.1093/molbev/msp077

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi: 10.1093/bioinformatics/btu033

Zhang, T., Wu, Q., & Zhang Z. (2020). *Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak*. Current Biology, volume 30, issue 7, pages 1346-1351.e2.

Zhao, Z., Sokhansaj, B. A. & Rosen, G. L. (2020). Characterizing geographical and temporal dynamics of novel coronavirus SARS-CoV-2 using informative subtype markers. Manuscript submitted for publication. doi: 10.1101/2020.04.07.030759