Recreate the Evolutionary History of a Severe Acute Respiratory Syndrome Coronavirus 2 Outbreak in a US Major League Soccer Club Using Nextstrain and Ultrafast Sample placement on Existing tRee

Zijing (Carol) Zhou[1], Anne Piantadosi[2]

[1]Emory College of Arts and Sciences, Emory University, Atlanta, Georgia, USA
[2]Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, Georgia, USA

Author Contributions

Anne Piantadosi conceived of the research topic and provided the data. Zijing (Carol) Zhou conducted the data analysis and wrote the manuscript. All authors edited the manuscript.

Honor Code

By typing your name below, you agree that at least one of your mentors has read and provided comments on this research report. You agree that you have followed the honor code for the course, which includes doing none of the following violations:

- falsifying data.
- providing false information to the course instructor or any research mentor.
- using any text from a published research article, a grant or paper draft written by someone other than yourself in this report unless you have prior consent from the course instructor to collaborate on your research report.
- intentionally inappropriately referencing the research of others
- unintentionally plagiarizing or inappropriately referencing the writing or research of others because of use of AI.
- using AI to generate content. Note, you may use AI tools to edit content.

Your Name: Zijing (Carol) Zhou

Abstract

In recent years, many phylogenetic tools have been developed to solve the problem of constructing phylogenetic trees efficiently and precisely because of the overwhelming amount of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) data. However, considering the software is constantly updating, it is unknown whether using them and performing a phylogenetic analysis again would produce the same result as the one done several months ago. Many literatures have only focused on testing and proving the productivity and accuracy when the toolkit first came out but have yet to consider the importance of comparing the before- and after-update outcomes. In this report, we recreate the evolutionary history of a published SARS-CoV-2 outbreak in a US Major League Soccer (MLS) Club with two methods. We first construct a phylogenetic tree using the Nextstrain SARS-CoV-2 Workflow following the paper and then perform phylogenetic placement using Ultrafast Sample placement on Existing tRees (UShER). We expect the recreated and placement trees to imply a similar phylogenetic history as the published one and that this approach could give the public an idea of the reliability of the two toolkits.

1 Introduction

Ever since the Coronavirus Disease 2019 (COVID-19) pandemic, more and more genomic data of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) have been generated because many people were being tested. However, this makes the process of constructing a phylogenetic tree inefficient. In reality, to ensure reliability, researchers need to find and align different published sequences that are similar to their sequences and that are from different locations whenever performing phylogenetic analysis. The overwhelming amount of data makes it even more time-consuming to not only filter the intended sequences from online databases but also run tree construction algorithms. To solve this problem, many phylogenetic tools have been developed, helping to locate genetically alike sequences or place one or multiple sequences onto an existing tree consisting of all the publicized SARS-CoV-2 sequences instantly and precisely.

Nevertheless, one question emerges from this — would the same results be produced several months later using the same toolkits and methods given that they are updated so frequently? Several papers have reviewed the accuracy of phylogenetics software when they were first published but not after they had any major updates. Moreover, the results generated using the default settings of those tools should be considered with caution because they do not always represent the true evolution history (Morel et al 2021). Thus, it is necessary to test whether the results after the update would be the same as before.

To tackle this question, we performed phylogenetic analyses on existing data generated by a lab member for her paper using two methods (Carmola et al 2023). First of all, sequences were aligned to a reference sequence and down-sampled. Maximum likelihood trees were then constructed using Nextstrain. Next, sequences were uploaded and placed onto an existing phylogenetic tree using Ultrafast Sample placement on Existing tRees (UShER) (Turakhia et al 2021). All results were compared to the findings in Carmola's paper. In this report, we demonstrate that though different sequences are placed onto the phylogenetic tree several months later, the phylogeny conveys a similar evolution history overall.

2 Materials and Methods

Data

MLS players' data was collected by GeoSentinel, and the sequence data and metadata of Delta variant lineages AY.36 and AY.4 were downloaded from GISAID as described in Carmola's paper.

Recreate the Phylogenetic Trees

According to Carmola's paper, two phylogenetic trees are constructed. For the first tree, one player's sequence was aligned with the reference strain Wuhan/Hu-1/2019 and 150 global AY.36 sequences that were down-sampled from a total of 2424 AY.36 sequences using the Nextstrain SARS-CoV-2 Workflow. For the second tree, seven players' sequences are aligned with the reference strain and 150 AY.4 sequences that were down-sampled from 64,083 AY.4 sequences using the same workflow. Two maximum likelihood trees are constructed using the default settings of the workflow with TreeTime.

Perform Phylogenetic Placement

The eight players' sequences were uploaded onto the web interface of UShER and placed onto an existing global tree containing 16,499,568 sequences. Two subtrees were created, having 150 samples per subtree — one containing the one sequence from the first tree while the other containing the seven sequences from the second tree.

Compare Results with the Original Trees

All trees were viewed on auspice.us and compared against the original trees produced in Carmola's paper. All the sequences were color-coded by world regions except for the players' samples, which were coded as "MLS" to distinguish them from other sequences.

3 Results and Discussion

A maximum likelihood phylogenetic tree was generated, and phylogenetic analysis revealed that the one player's sequence was most closely related to a sequence from Africa and a sequence from Europe (Fig. 1). Compared to Carmola's results, in which the one player's sequence was found to be closely related to a sequence from Africa, the recreated tree shows a similar phylogenetic history. However, it is interesting to note that though the data and methods were the same, the sequences in the two trees were different, and fewer sequences from Africa made it to this tree than the tree in Carmola's paper.

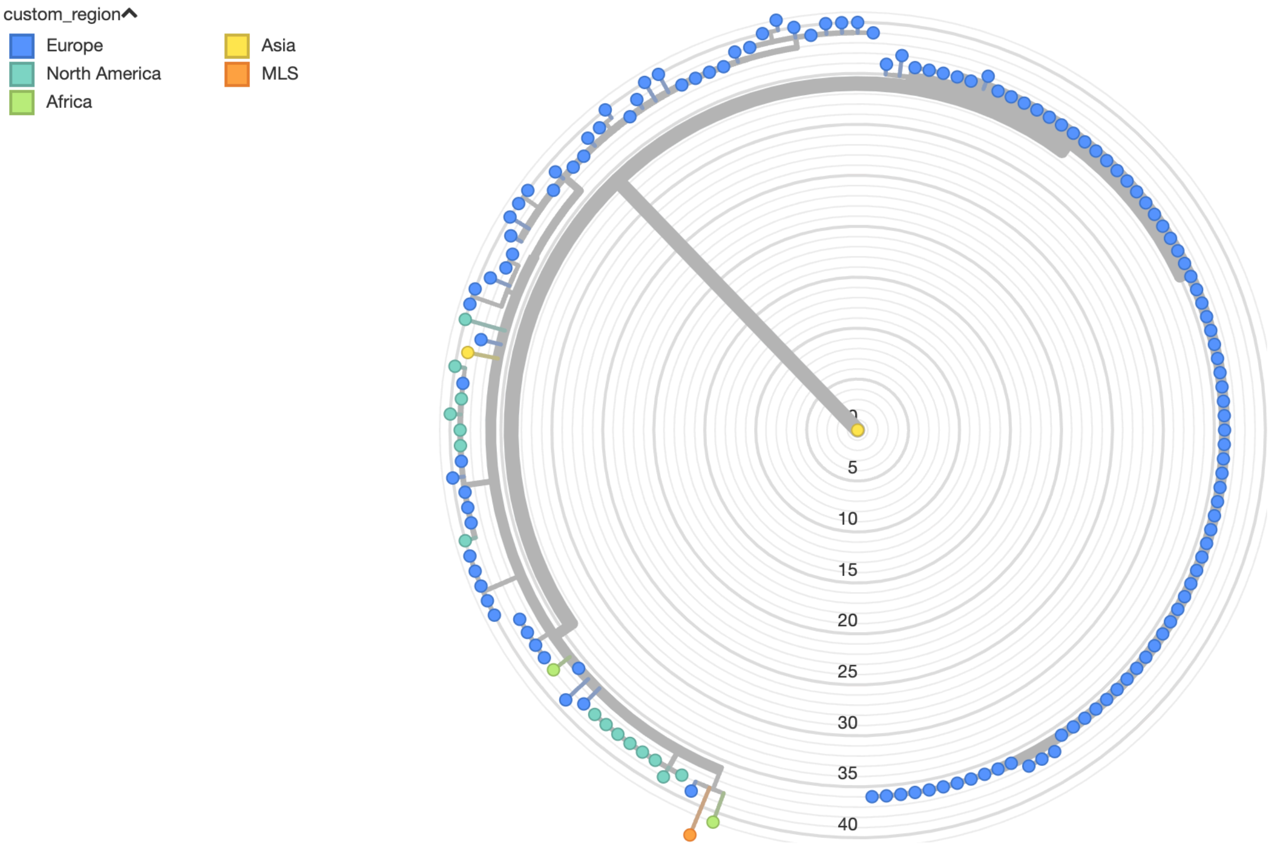

Figure 1. Phylogenetic analysis performed on Nextstrain indicates a similar evolutionary history discussed in Carmola's paper and that the one player's sequence is closely related to sequences from Europe and Africa. A maximum likelihood tree was constructed with the one player's sequence and 150 AY.36 sequences, rooted to reference Wuhan/Hu-1/2019. The player is colored orange.

A phylogenetic placement was performed, and a maximum likelihood phylogenetic tree was generated with sequences that were identical to the player's sequence. This tree hinted that the one player's sequence was most closely related to a sequence from North America (Fig. 2). A closer inspection of the tree reveals that the North American sequence is the player's sequence. Thus, in reality, the player's sequence clustered with different sequences from North America, Europe, and Africa. In contrast with Carmola's results, this tree conveys more uncertainty about the player's sequence but, for the most part, a similar genetic history.

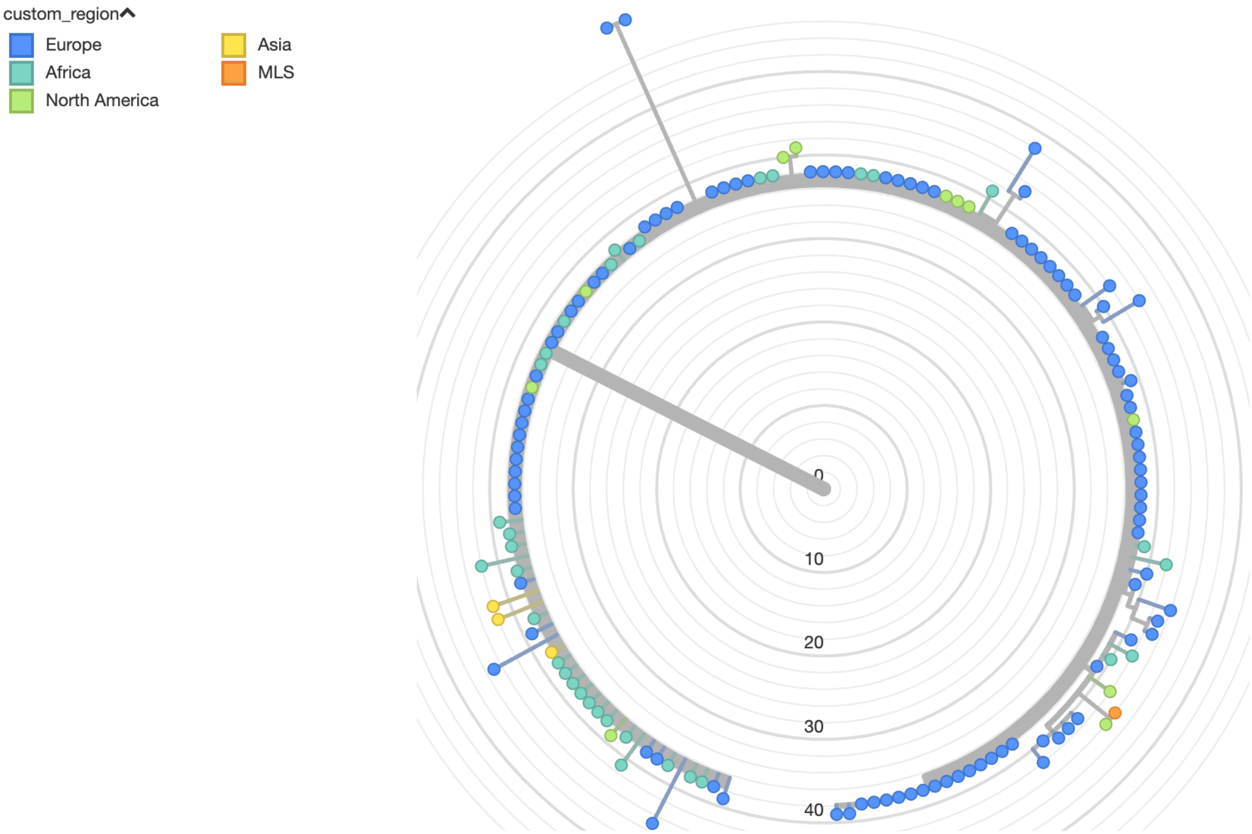

Figure 2. Phylogenetic analysis performed on UShER suggests that the one player's sequence is closely related to sequences from North America, Europe, and Africa, implying a similar evolutionary history discussed in Carmola's paper. A phylogenetic placement was performed on UShER using all 8 players' sequences, and two subtrees were made. This subtree contains 150 sequences: the 1 player's sequence and 149 published sequences from online databases. The player is colored orange.

To date, we are still in the phase of recreating the second tree due to the difficulties and challenges faced when using Nextstrain. However, we anticipate that the recreated phylogenetic tree will infer a similar evolution history as the one presented in Carmola's paper.

Another phylogenetic placement was performed, and a maximum likelihood phylogenetic tree was created with sequences that were identical to the sequences of the seven players. This tree conveyed that the players' sequences were most closely related to sequences from North America (Fig. 3). In comparison to Carmola's findings, in which the players' sequences were also closely related to sequences from North America, this tree implies the same similar evolutionary history.

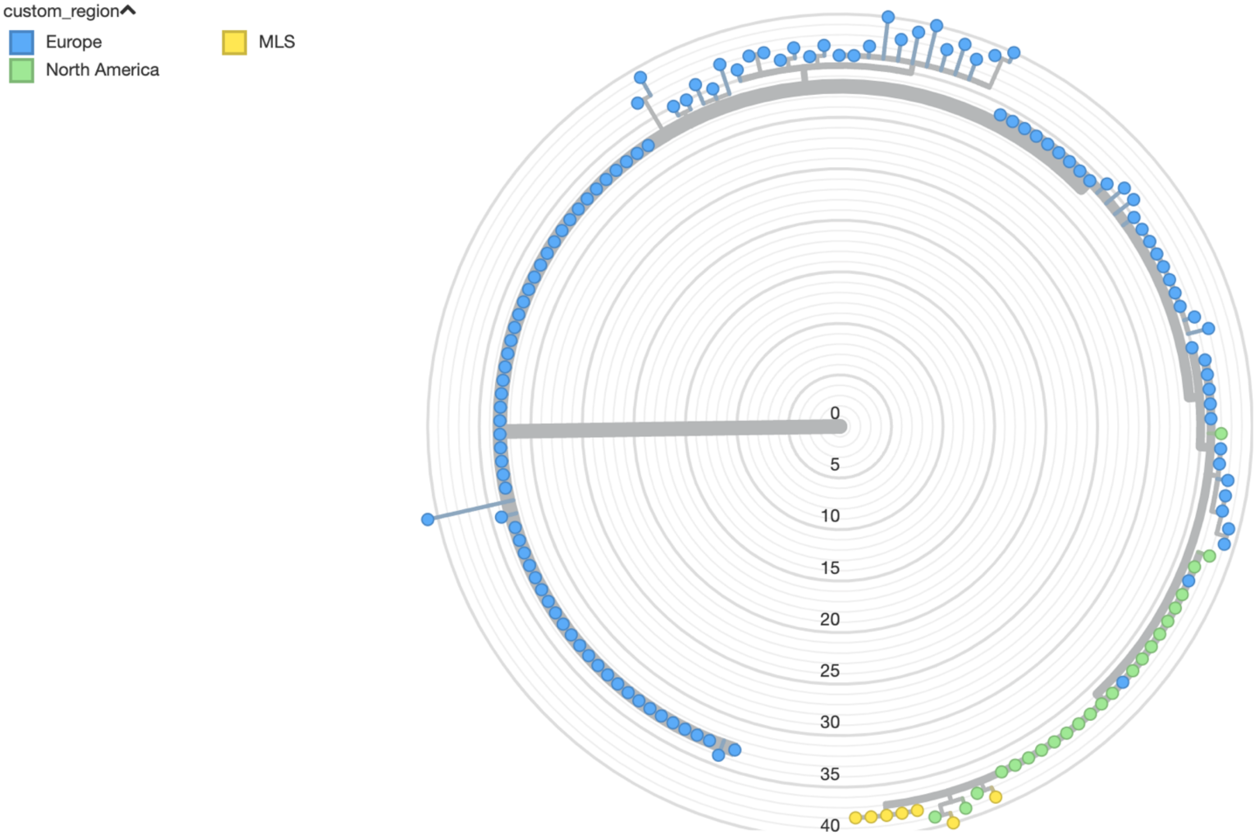

Figure 3. Phylogenetic analysis performed on UShER suggests that the seven players' sequences are closely related to each other and sequences from North America, aligning with Carmola's finding. A phylogenetic placement was performed on UShER using all 8 players' sequences, and two subtrees were made. This subtree contains 150 sequences: the 7 players' sequences and 143 published sequences from online databases. players are colored yellow.

To sum up, this report highlights the reliability of phylogenetic tools such as Nextstrian (before and after updating) and UShER in terms of inferring the true evolutionary history of SARS-CoV-2 as demonstrated using published data from Carmola's paper.

4 Funding

This research project is not funded.

5 Acknowledgments

We thank Tyshawn Ferrell for helping with using Nextstrain.

6 References

Morel, B., Barbera, P., Czech, L., Bettisworth, B., Hübner, L., Lutteropp, S., et al. (2021). Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. Molecular Biology and Evolution 38, 1777–1791. doi: 10.1093/molbev/msaa314.

Carmola, L. R., Turcinovic, J., Draper, G., Webner, D., Putukian, M., Silvers-Granelli, H., et al. (2023). Genomic Epidemiology of a Severe Acute Respiratory Syndrome Coronavirus 2

Outbreak in a US Major League Soccer Club: Was It Travel Related? Open Forum Infectious Diseases 10, ofad235. doi: 10.1093/ofid/ofad235.

Turakhia, Y., Thornlow, B., Hinrichs, A. S., De Maio, N., Gozashti, L., Lanfear, R., et al. (2021). Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet 53, 809–816. doi: 10.1038/s41588-021-00862-7.