Zachary Bonnstetter
Claire Chen
Nadya Postolaki
CSCI 5481- Dr. Dan Knights

# Longitudinal Analysis of SARS-CoV-2 Genomes in Minnesota

## Abstract:

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is a member of the coronavirus group of RNA viruses known for causing respiratory tract infections of varying severity by strain. Since the end of 2019, SARS-CoV-2 has been responsible for a worldwide pandemic, during which time scientists have been collecting and sequencing sample viral genomes for upload to online repositories such as the National Center for Biotechnology Information's "NCBI Datasets", which currently contains almost 30,000 sequenced SARS-CoV-2 genomes from around the world, updated nearly daily. The benefit of such thorough cataloging of the SARS-CoV-2 genome is that as an RNA virus, transmission can be tracked and monitored by observing the acquisition of new mutations in the genome over time and measuring the genetic distance between given samples. Here we present a longitudinal analysis of SARS-CoV-2 complete genomes in the state of Minnesota, performing multiple sequence alignment of 758 SARS-CoV-2 genomes collected in Minnesota over the course of 9 months, assembling phylogenetic trees, and examining evolutionary time and genetic distance between genomes to explore SARS-CoV-2 transmission and mutation. We observe overall low genetic distances and evolutionary time between samples, and clear clustering by viral sample collection date in trees, demonstrating that SARS-CoV-2 is mutating slowly and proving phylogenetic trees to be an effective means of monitoring viral transmission and mutation.

## Summary of previous findings:

Previous studies have estimated that SARS-CoV-2 accumulates two single-letter mutations in its genome per month, relatively slow compared to other similar viruses such as influenza, which accumulates mutations roughly twice as fast as SARS-CoV-2 [1]. It has also previously been demonstrated that SARS-CoV-2 genome phylogenetic trees display clear clustering by geological region [1] however there is yet little analysis of further clustering within a region, such as by the date which the viral genome sample was collected. And although previous studies have explored predicting future SARS-CoV-2 mutation rate through the use of machine learning based on current global, continental, and national average mutation rate(s) [2], we are left wondering if these geologically large-scale mutation rate estimates are consistent with observed mutation rate(s) in smaller geological regions.

## Results:

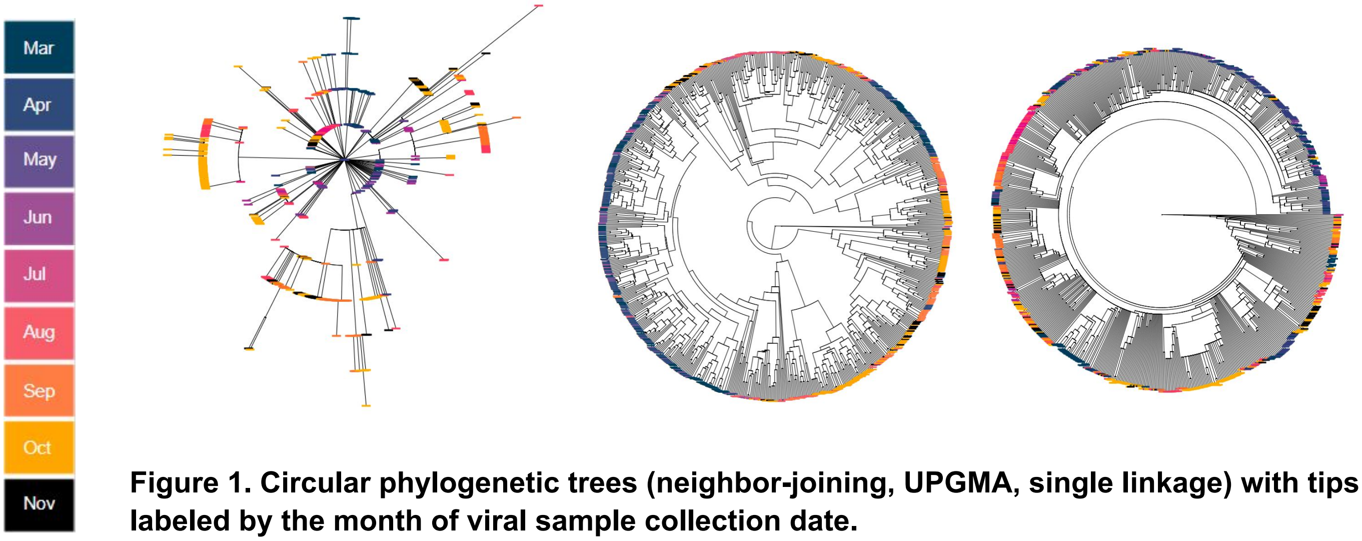

**Figure 1. Circular phylogenetic trees (neighbor-joining, UPGMA, single linkage) with tips labeled by the month of viral sample collection date.**

In phylogenetic trees, two species are more related if they have a more recent common ancestor, and less related if they have a less recent common ancestor. According to the three phylogenetic trees from Fig.1, we observe clear clustering of tips by color, indicating that the genomes collected within the same month are more genetically similar. In addition, we observe that as genetic distance increases outward from the root in the neighbor-joining tree, tips are again clustered by color with genomes collected in later months (Aug-Nov) being more prevalent further from the root and genomes collected in earlier months (Mar-May) closer to the root. An additional visualization of genetic distance is shown in Fig.2. Using a custom R script, a matrix of pairwise genetic distances was computed from the aligned sequences. Two genomes were used as the “root” for this visualization: the first sample collected, and the sample with the most genetic similarity to other early samples. In both graphs, we observe a general linear trend of genetic distance increasing steadily but slowly over time. In each graph, there is an approximate increase of 0.000325 in genetic distance over the course of the 9 months. With 0.000325 new mutations per site in approximately 29,000 sites, this indicates a mutation rate of approximately 9.425 new mutations over the course of the 9 months, or 1.05 new mutations per month. This is very similar to estimated mutation rate in previous studies, as mentioned earlier, though almost twice as slow as other studies have estimated [1]. This could be a result of low sample size relative to these other large-scale studies, or it could be indicative of greater fluctuation in mutation rate of the virus by strain masked in large-scale studies; in order to determine which, if either, the case might be, additional study would be required.

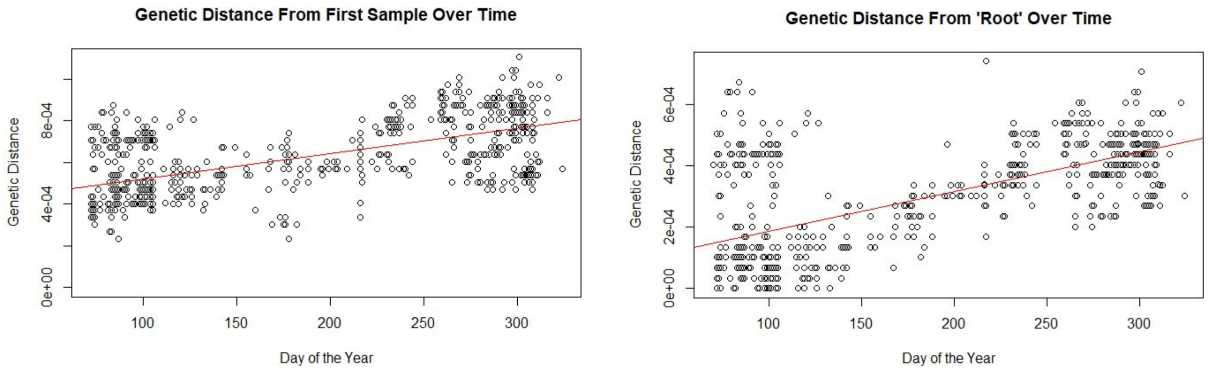

**Figure 2. Graphing genetic distance of genomes over time. Genetic distance clearly increases slightly but steadily over time, at a rate of approximately 1.05 new mutations per month.**

Left: genetic distance of samples as calculated compared to the first sample collected.

Right: genetic distance of samples as calculated compared to the sample with the most genetic similarity to other early samples.

**Methods:**

The source of assembled SARS-CoV-2 genomes were downloaded from the NCBI Datasets website using their datasets command-line tool. Four flags were used to filter the data by viral taxon (to specify only SARS-CoV-2 data), host organism (to specify only human hosts), geographic location (to specify only Minnesota), and genomic sequence completeness (to specify only complete genome sequences). The exact command-line prompt utilized to download the data is shown below.

```
        datasets download virus genome taxon 2697049 --geo-location
Minnesota --host human --complete-only --filename SARS2.zip
```

An additional NCBI Datasets table was downloaded manually from the NCBI Datasets website which contained the exact date on which each viral sample for sequencing was collected. This table was manually filtered down to only the data fitting the specifications outlined above, and an additional column was added to each genome ID's row corresponding to the color in which to label each genome's corresponding tip in the downstream phylogenetic trees based on the collection date for that viral sample.

Multiple sequence alignment was performed using DECIPHER (Database of genomic variation and Phenotype in Humans using Ensembl Resources), a progressive-iterative alignment algorithm based tool available as a package in R. During alignment, default parameters were utilized as any modifications to parameters proved either unnecessary or suboptimal. The default parameters utilized, along with a justification for each, can be found below:

guideTree = Null

- Because this was a novel alignment, a guideTree had to be generated from scratch by clustering by genetic distance.

iterations = 2

- The alignment converged after 2 iterations, increasing the number of iterations would simply result in early termination of alignment after the second iteration.

refinements = 1

- Only one refinement step proved necessary per iteration as additional refinements resulted in negligible alignment alterations.

Tree building was performed through the use of the MAFFT Version 7 (Multiple alignment program for amino acid or nucleotide sequences) web server [3]. The three algorithms utilized were neighbor-joining, UPGMA, and single linkage. Output newick files were downloaded, modified with a custom script to contain the correct shortened tip labels, and then were visualized with colored tips corresponding to the month of genome sample collection by running a custom modification of hw3-plot-newick.r file, custom-plot-newick.r (Fig.1).

For genetic distance calculation and visualization, genetic distance was calculated using the dist.alingment() function from the seqinr package in R, visualized using the base R plot() function, and then a linear regression line was added using the stats package lm() function (Fig. 2).

**Conclusion:**

In this analysis we sought to present a longitudinal analysis of SARS-CoV-2 complete genomes in the state of Minnesota in order ascertain the rate at which the virus is mutating in Minnesota, as well as determine whether phylogenetic trees constructed from aligned viral genome sequences displays any form of clustering by date of viral sample collection that would enable the tree to be a tool for tracing viral transmission. After performing multiple sequence alignment on 758 Minnesota SARS-CoV-2 genomes collected over the course of 9 months, we do indeed observe clear clustering by date of viral sample collection, as well as a clear increase in genetic distance from early samples over time, both as expected. The observed mutation rate of 1.05 new mutations per month was slightly lower than expected compared to previous studies, but is a realistic and reasonable value. We hope that additional future studies can shed light on this discrepancy in mutation rate, and elucidate whether SARS-CoV-2 mutation rate does indeed vary widely by strain, or if this is the result of a relatively low sample size (758 genomes compared to thousands in global/national-scale studies). Additional exploration of alternative clustering methods for tree building and the incorporation of early Wuhan SARS-CoV-2 genome sequence data in the alignment to serve as a known root could prove beneficial.

**Acknowledgments:**

Zach:

- Downloading source data.
- Data preparation and manipulation.
- Performing multiple sequence alignment.
- Performing tree-building.
- Construction of figures.
- Preparing the presentation, editing as needed.
- Report abstract.

- Report results.
- Report methods.
- General report editing.

Nadya:

- Preparing the presentation, editing as needed.
- Editing resulting trees and images.
- Setting up and editing the report.
- Report Conclusion
- Report abstract
- Portion of the report results, and minor edits to methods.
- Assisting in finding what data to attain, algorithms to be used, utilizing some from HW3.
- Github to share files and discord meetings.

Claire:

- Finding the source for the SARS-CoV-2 genomes.
- Creating a spreadsheet for the downloaded data.
- Assisted in phylogenetic tree tip labeling.
- Preparing the presentation, editing as needed.
- Report results
- Report method

Citations:

1. B Korber, WM Fischer, S Gnanakaran, H Yoon, J Theiler, W Abfalterer, B Foley, EE Giorgi, T Bhattacharya, MD Parker, DG Partridge, CM Evans, TM Freeman, TI de Silva, on behalf of the Sheffield COVID-19 Genomics Group, CC LaBranche, DC Montefiori bioRxiv 2020.04.29.069054; doi: https://doi.org/10.1101/2020.04.29.069054
2. Pathan RK, Biswas M, Khandaker MU. Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model. *Chaos Solitons Fractals*. 2020;138:110018. doi:10.1016/j.chaos.2020.110018
3. “NJ / UPGMA Phylogeny.” *MAFFT Multiple Alignment Program for Amino Acid or Nucleotide Sequences*, mafft.cbrc.jp/alignment/server/phylogeny.html.