More Than Just Cases: Mapping the Spread of COVID-19 Using Geospatial Nucleotide Mutations and Pathogen Phylodynamics

By John Hessler

INTRODUCTION

For almost everyone in the world, the last few months have been unlike any experienced in their lifetimes. The current public health crisis, spawned by the outbreak of the Novel Coronavirus, COVID-19, has shown that viral pathogens pose an ever-present danger to global human health and economic stability. For cartographers and epidemiologists the rapid evolution and mutation of nucleotides and amino acids of the SARS-CoV-2 virus present a geospatial analysis challenge like none other, as public health officials, emergency rooms and the general public struggle to track the spread of the disease and allocate resources. Genomic data, paired with modern GIS and computational techniques, have allowed for the spatial and temporal tracking of the disease's spread, have catalyzed the development of new cartographic visualization methodologies, and have helped predict the movement of COVID-19 around the world.[1]

Much of this mapping and analysis has been possible because new data sources and providers have been organized in the last few years, who are capable of providing up-to-date information on the genomics and the spatial and temporal distribution of rapidly transmitted and dangerous diseases, in real time, during outbreaks. GISAID, the Global Initiative for Sharing All Influenza Data, for example, aggregates rapidly accumulating genomic data from labs around the world during serious disease pandemics and makes that data available to qualified and registered users.[2]

Analysis of this critically important data has also been streamlined and made easier by platforms like Next Strain,[3] an interactive software suite for phylodynamics that consists of data curation, analysis and visualization components, which use Python scripts to maintain and update a database of available sequences and related metadata, sourced from public repositories such as NCBI (www.ncbi.nlm.nih.gov), GISAID (www.gisaid.org) and ViPR (www.viprbrc.org), as well as GitHub repositories and other sources of genomic data. The software contains powerful tools used to perform phylodynamic modeling, geographic mapping of mutations, and transmission networks, and includes tools that allow the inference of the most likely past and future transmission events.[4]

This short paper will introduce readers to a few of these new mapping tools and explain how genomic data, along with advanced GIS applications, are being used to track COVID-19.

GEOSPATIAL MAPPING THE PHYLODYNAMIC DATA

Phylogenetic trees are complex graphs used to represent the genetic history of an organism and are constructed computationally using statistical algorithms. In the case of COVID-19 these algorithms look for common mutations in the strains of the virus as they evolve and jump through multiple human hosts. The structure of these trees give epidemiologists and cartographers a great deal of information regarding how fast a virus is changing and the path each mutation takes as it geographically moves from place to place. When these phylogenetic trees are inferred using testing and sequencing data from laboratories across the world, in real time, they give cartographers unique insights into where and how each new outbreak originated. **[Figure 1]**

The spatial-temporal propagation parameters of the new Corona virus are not well understood, with critical factors, such as the basic reproductive number and the effective reproductive number, not accurately known[5]. The attachment of geospatial metadata to the nucleotide mutation information however, allows us to use the phylogenetic tree and map the relationship between location, mutation and transmission along particular paths and begin to reconstruct the virus' transmission networks, such as those to and from the United States. The red lines in **Figure 2** represent locations and pathways of the disease in the United States and Canada that are all related genetically. In this case we can see the first outbreak that occurred in Washington State came from China and was collected by the lab on January 19th.

Here the purple represents transmission within China, with the red dots sprinkled in amongst them representing the genetic and geographic pathways to four locations in the United States. Mapping the locations shows the first outbreaks in the United States occurring in Washington, California, Illinois and Arizona happening at nearly the same time in mid-to-late January of 2020. All this

© 2020. John Hessler

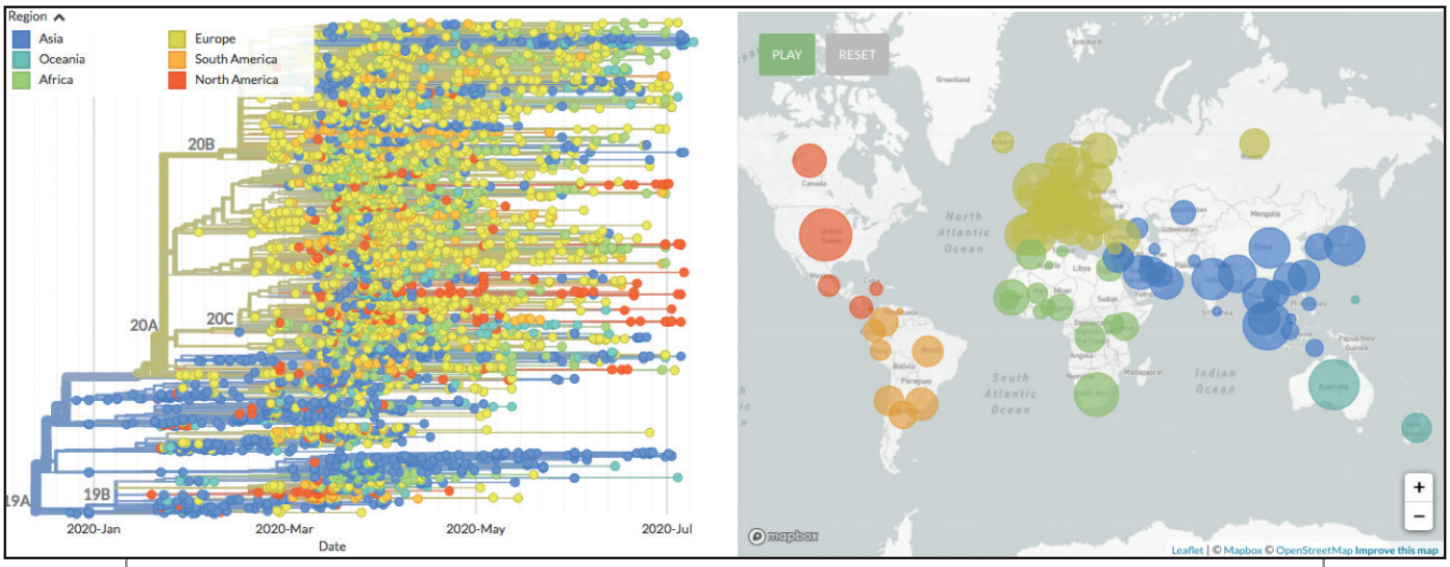

Figure 1. Geo-spatially resolved phylogenetic tree of COVID-19. The blue represents mutations and spread of the disease from China. Red represents North American mutations of the virus. Courtesy Next Strain.

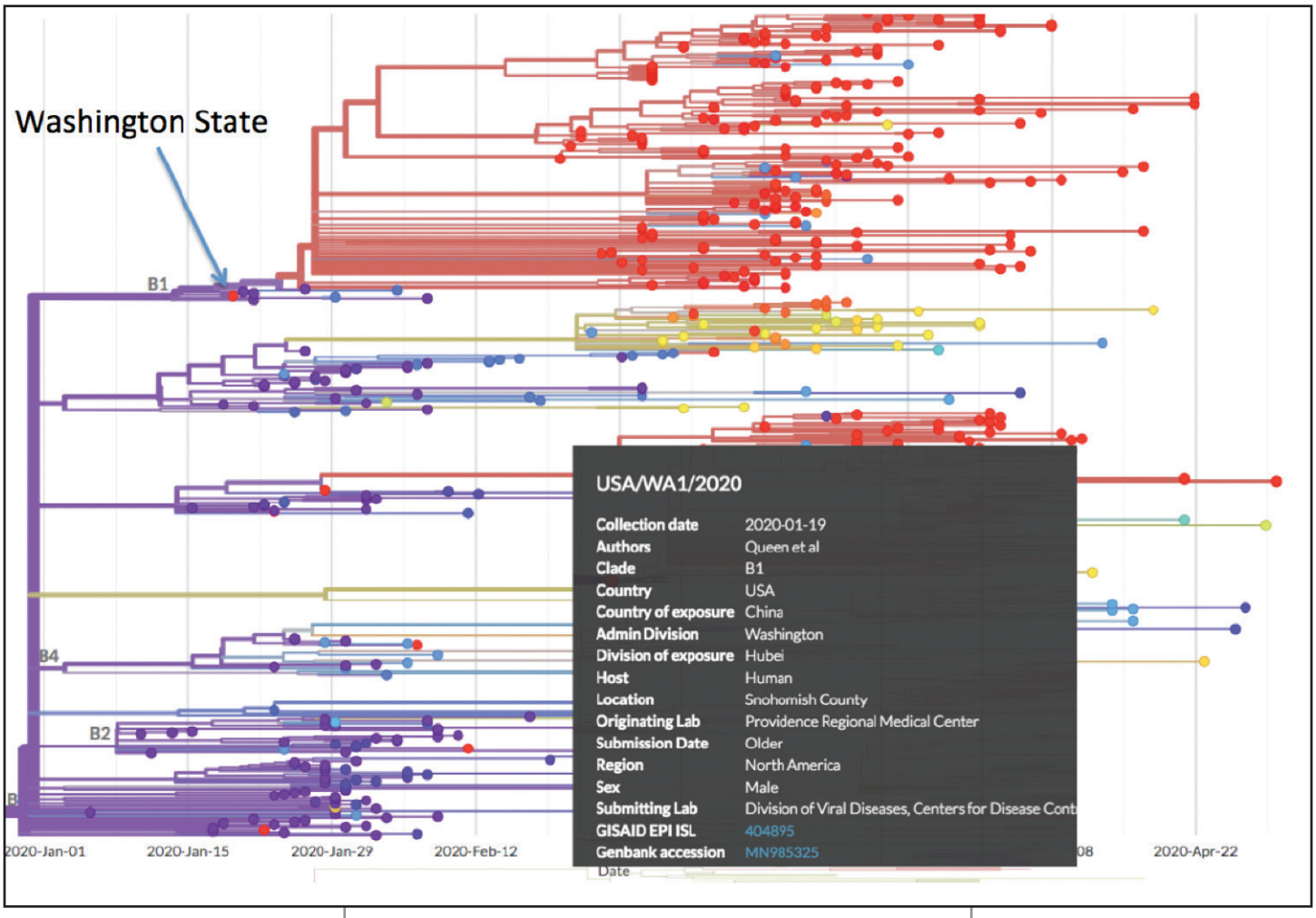

Figure 2. First case in the US resolved genomically and spatially

spatial resolution is possible because of the aggregation of GISAID nucleotide data from labs around the world, which also has geo-locations included in the metadata.

The key to all of this predictive and tracing power is the linkage between the genetic sequences of the virus and the map. For example, we can see in **Figure 3** the mutations found in cases around the United States and North America. Using the tree and the genomic data one can zoom into the various regions of the country and trace the spread geographically, not only from person to person, but also more importantly from place to place with great precision.

In this map (Figure 3) one can easily resolve, even at this scale, certain regional variations in the virus (shown by the different colors), but also observe, by looking for example at the red transmission events, how that particular mutation of COVID-19 moved from Washington State, to Alaska, to Chicago, to Hawaii, to the Northeastern US and to Canada.

For epidemiologists it is extremely important to understand the mutation and geospatial dynamics of any disease. As an RNA virus, as COVID-19 moves from human host to human host, the nucleotides that make up the virus's genetic material change. Understanding how each of the various forms of the virus travels geographically and whether one or another mutation becomes dominant is critically important, not only for those treating the disease, but also for scientists attempting to look for vaccines.

SARS-CoV-2 has several geographically important strains that have been mapped using these techniques. Looking at **Figure 4**, one can see the first two parts or clades, shown in blue and turquoise and labeled 19A and 19B, correspond to the split in the trees marked by mutations C8782T and T28144C.[6] These clades were both prevalent in Asia during the first months of the outbreak. The next clade that was named is 20A corresponding to the clade that dominated the large European outbreak in early 2020. It is distinguished from its parent 19A by the series of mutations C3037T, C14408T and A23403G. After this, there are two further clades that appeared later, 20B, another European clade separated clearly by three consecutive mutations G28881A, G28882A, and G28883C. And, finally there is 20C, a largely North American clade, distinguished by mutations C1059T and G25563T. This kind of mapping of the viral mutations has allowed cartographers to conclude, in the case of the United States, that two different paths of transmission led to the spread of the virus here. One coming from Asia in the first period of the outbreak, and another later event, which came from Europe and which was responsible for a considerable number of the cases in New York and the Northeast.

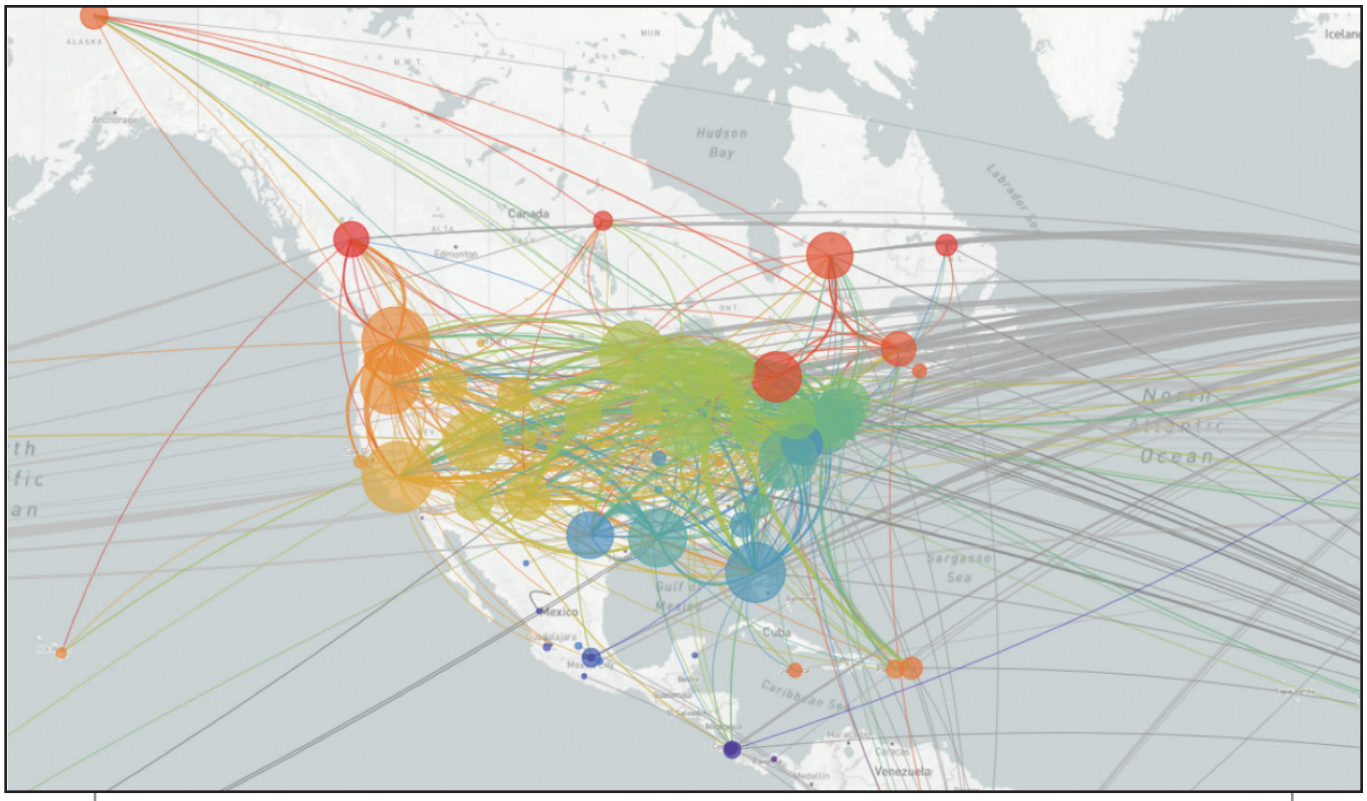

Figure 3. Network transmission map of COVID-19 as of July 15, 2020. The colors in the tree represent mutations of the virus and allow the geospatial mapping and analysis of the virus' spread. Built with Next Strain.

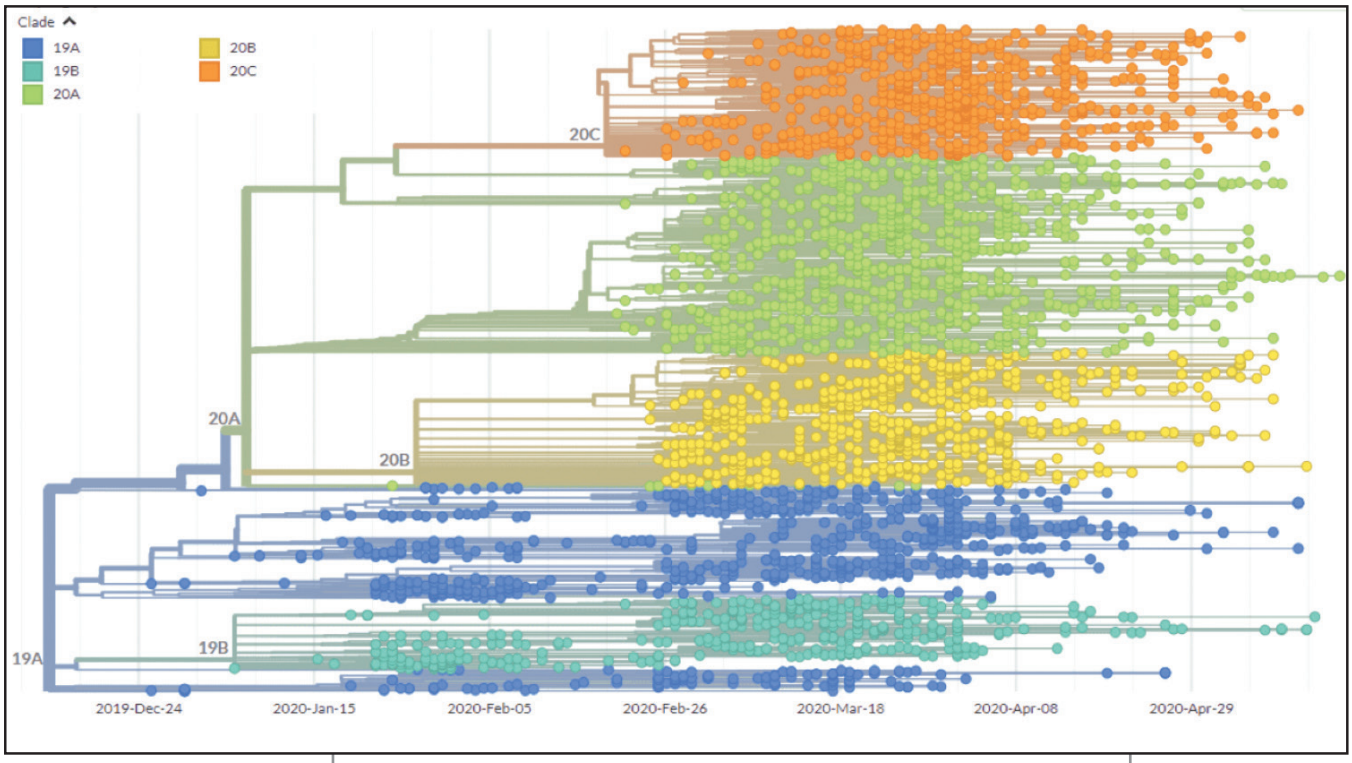

Figure 4. Major Clades and Mutations of COVID-19. Courtesy Next Strain.

CONCLUSIONS

While no spatial or cartographic model of viral pathogen transmission is perfect, the use of quickly accumulating and reliable data of phylogenetic structure and dynamics aggregated by a single source, such as GISAID, has profound advantages over typical geospatial data when looking to map and spatially resolve clear viral pathogen transmission pathways for large or global scale pandemics. This combining of powerful phylodynamic algorithms, accurate time genome sequencing and geospatial metadata, while not computationally perfect, has gone a long way in helping policy makers, medical geographers, GIS analysts and cartographers understand the spread of COVID-19.

Tools such as those talked about here are part of a major shift in mapping and cartography over the last few decades. Cartography has gone from a science that dealt with landforms and mostly static phenomena, to one whose main objective is to understand the spatial dynamics of quickly changing and non-equilibrium processes. The continuing development and increased sophistication of these mapping tools will help us to understanding future emerging pathogen threats and other problems effecting populations around the globe.

John Hessler is a Specialist in Computational Geography and Geographic Information Science at the Library of Congress. He is a Lecturer of Evolutionary Computation at the Johns Hopkins University and founder of the New Geometries Laboratory whose research uses geospatial data in conjunction with evolutionary computation to solve complex spatial analysis problems, like mapping COVID-19.

ENDNOTES

1 Pybus, O. et al. (2013) Evolutionary epidemiology: preparing for an age of genomic plenty. *Phil. Trans. R Soc. B*, 368, 20120193–20120193.

2 GISAID, the Global Initiative for Sharing All Influenza Data https://www.gisaid.org/.

3 NextStrain, https://nextstrain.org/.

4 Volz,E.M. et al. (2013) Viral phylodynamics. *PLoS Comput. Biol.*, 9, e1002947. https://doi.org/10.1371/journal.pcbi.1002947.

5 Bnaya Gross, et.al., "Spatio-temporal propagation of COVID-19 epidemics," (2020) https://arxiv.org/abs/2003.08382.

6 The notation for mutations of nucleotides uses initials for the four base pairs that make up DNA; adenine (A), cytosine (C), guanine (G), or thymine (T). C8782T means that at the position 8782, the nucleotide cytosine was substituted with thymine. COVID-19 has over 29,000 base pair positions in its genome.