Received: 5 June 2024 | Revised: 5 July 2024 | Accepted: 10 July 2024

DOI: 10.1111/tran.12703

THEMED INTERVENTION

To save lives: Lessons of a pandemic cartographer

John Hessler

The Biomap-Lab and Johns Hopkins University, Baltimore, Maryland, USA

Correspondence
John Hessler, The Biomap-Lab and Johns Hopkins University, Baltimore, MD, USA.
Email: jhessle1@jhu.edu

Abstract

For almost everyone in the world, the last few years have been unlike any experienced in their lifetimes. The public health crisis, spawned by the SARS-CoV-2 virus and the outbreak of COVID-19, presented a geospatial analysis challenge like none other as public health officials, emergency rooms, and the general public struggled to track the spread of the disease, allocate resources for testing and care, and understand the origin of this new zoonotic pathogen. During this period, the combination of rapid and open access genomic data combined with case counts and the mapping tools of geographic information systems allowed for near real-time tracking of the pandemic. This paper describes these tools, how they were used to advise policy-makers in the US at the height of the pandemic, and some of the lessons learned.

KEYWORDS

COVID 19, disease mapping, geographic information science, medical geography, SARS-CoV-2, spatial epidemiology

> Everybody knows that pestilences have a way of recurring in the world, yet somehow, we find it hard to believe in ones that crash down on our heads from a blue sky.
>
> Albert Camus, *The Plague*

1 | INTRODUCTION

For almost everyone in the world, the last few years have been unlike any experienced in their lifetimes. The public health crisis, spawned by the SARS-CoV-2 virus and the outbreak of COVID-19, has reminded us that viral pathogens pose an ever-present danger to global human health and economic stability.

For cartographers and epidemiologists, the rapid spread of the disease and the evolution and mutation of the virus presented a geospatial analysis challenge like none other as public health officials, emergency rooms, and the general public struggled to track the spread of the disease, allocate resources for testing and care, and understand the origin of this new zoonotic pathogen.

For much of the COVID-19 pandemic, beginning in March 2020, the author, who at the time was a specialist in computational geography and geographic information science at the Library of Congress in Washington DC and a lecturer at Johns Hopkins University, advised members of the US House of Representatives and Senate on the ongoing

accumulation of cases and gave confidential briefings to members on new ways to map and track the geographic spread of this previously unknown pathogen.

This kind of mapping and analysis, which combined new breakthroughs in genomics and the rapid accumulation of large amounts of bioinformatic data, was critical to the real-time public health response. This massive amount of geospatially structured bioinformatic data, gathered during what can be described as the first pandemic of the information age, also presents us now with a unique opportunity, as we begin to retrospectively examine the various social and political aspects of the pandemic response and its effects on communities, from both a global and local perspective (Sokhansanj & Rosen, 2022).

Although produced for scientific and public health response reasons, the maps and techniques discussed in this paper have also been used in wider geographical and political analysis. As is well known to geographers, case counts and disease transmission pathways are not only biological phenomena but are also influenced by the complex interplay of human behaviour and the historical economic inequities found across the globe. Many commentaries and papers focusing on the social aspects and the political uses and abuses of disease mapping have been published over the last few years, both during the height of the pandemic and in its aftermath (Marvin et al., 2023; Sparke & Anguelov, 2020). Several have also focused on the use of deep learning and natural language processing to map and examine the social media responses to the outbreak and to the enforcement of the ensuing public health policies (Hessler, 2020a).

In this paper, however, I will describe the real-time use of novel bioinformatic data analysis which, combined with the power of geographic information systems and mapping, was used to inform policy-makers and to help save lives during the pandemic (Hessler, 2020b).

At the beginning of the pandemic, case counts and genomic data about the virus were being generated at rapid pace and, when paired with modern GIS and other computational techniques, allowed for the near real-time spatial and temporal tracking of the disease's spread and the study of its mysterious and complex phylodynamics (van Dorp et al., 2021).

This data, and the need to present it in an understandable way to non-virologists and policy-makers, catalysed the development of new cartographic visualisation methodologies which helped us understand the movement of COVID-19 around the world. Resources like the Johns Hopkins COVID-19 dashboard (https://coronavirus.jhu.edu/map.html) became critical sources of information for the media, medical professionals, and the general public, as everyone struggled to comprehend the geospatial implications of the outbreak for the health of populations and economies (Peeples, 2022).

Much of the mapping and analysis that took place was made possible because of new data sources and providers who were capable of delivering up-to-date information on the genomics and spatial and temporal distribution of cases in real time. GISAID, the Global Initiative for Sharing All Influenza Data (www.gisaid.org), for example, brought together and still aggregates genomic data from labs around the world during serious disease flare-ups, like COVID-19 and influenza, and makes that data available to qualified and registered users (Sokhansanj & Rosen, 2022).

Analysis of this critically important data has also been streamlined and made easier by platforms like Nextstrain (Hadfield et al., 2018), an interactive software suite for phylodynamics that consists of data curation, analysis, and visualisation components, and that uses Python scripts to maintain and update a database of available sequences and related metadata, sourced from public repositories such as the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov), GISAID, and the Virus Pathogen Database and Analysis Resource (ViPR, www.viprbrc.org). The software contains powerful tools that help users perform phylodynamic modelling, the geographic mapping of mutations and creation of transmission networks, and also includes code that allows the inference of the most likely past and future transmission events.

The ease in the creation of phylogenetic trees of pathogens, which are complex graphs used to represent genetic history constructed computationally using statistical and Bayesian methods (Sagulenko et al., 2018), also made the mapping of changes and evolutionary dynamics of the virus more understandable to policy-makers.

In the case of SARS-CoV-2, these algorithms use the accumulating changes and deletions in the genome of the virus as it evolves and moves through multiple human hosts to infer the structure of these complex tree-like graphs. The structure of these trees give epidemiologists and cartographers critical information regarding how fast a virus is changing and infer the trajectory of each mutation as it geographically moves from place to place (Figure 1) (Volz et al., 2013).

2 | BIOINFORMATICS AND THE BIRTH OF PHYLOGEOGRAPHY

The microbial cause of the COVID-19 pandemic, SARS-CoV-2, is a single-stranded RNA virus whose origin and initial spread around the globe, from its spillover in Wuhan, China, has only recently begun to be understood in detail

FIGURE 1 Phylogenetic tree of SARS-CoV-2 based on sample genomes from December 2019 until March 2024. Colours indicate clades associated with the ancestral strain, variants, and subvariants. Data and visualisation platform from Nextstrain.

(Crits-Christoph et al., 2023; Hessler, 2023; https://arcg.is/CCfvP; Pekar et al., 2022; Worobey et al., 2020, 2022). The actual zoonotic spillover event, from a still to be identified animal host, remains, after four years, a continued locus of political controversy and scientific discussion.

What is certain is that this pathogen, the cause of the first large-scale global pandemic of the information age, generated more bioinformatic data on a shorter time scale, in the form of more than 16 million genetic sequences, than any that came before (GISAID current sequence count as of 3/10/2024). This unprecedented amount of data has allowed cartographers, epidemiologists, and geographers to combine the temporal and genomic data found in these sequences with geospatial information, and to analyse the movement and evolution of this virus in ways that could scarcely have been possible a decade ago.

While no spatial or cartographic model of viral pathogen transmission is perfect, the use of quickly accumulating and reliable data of the phylogenetic structure and dynamics of a disease has profound advantages over typical two-dimensional geographic case count data, especially when looking to map and spatially resolve accurate viral pathogen transmission pathways for large and global-scale pandemics.

Most epidemiological time-series, in the form of temporal cases and locations, are noisy, highly complex, and non-stationary, and are therefore difficult to analyse and map using traditional methods (Cazelles et al., 2007). Because of the rapidly changing nature of the spread of diseases like COVID-19, cartographers are trying to map far from equilibrium processes, whose global nature creates large uncertainty, especially in the modern era of globally connected and highly mobile populations.

The fact that international travel so quickly spread the various forms of the SARS-CoV-2 virus would have made tracing the geographic origins of COVID-19 difficult, if not impossible, if not for the availability of near real-time genomic sequence data from laboratories around the world. The combining of powerful and computationally efficient phylogenetic algorithms with accurate genome sequencing and geospatial data, while not perfect, went a long way in helping public health policy-makers and governments understand in detail the spread of COVID-19.

Mapping the movement of SARS-CoV-2 was, and still is, a complex undertaking. In order to trace the movement of COVID-19 around the world, it was necessary for geographers to look closely at a complex and large amount genomic data, and to spatially analyse the phylogeny of the nucleotide mutations and amino acid changes in individual cases, and then to trace those changes in other individuals as they caught and spread the disease. This exercise, in its simplest form, entails geographically mapping the process of mutation accumulation and examining at many scales the non-stationary movement of these changes as they bounced to new and sometimes distant locations.

One of the most difficult aspects of the enterprise was indeed the problem of scale. The SARS-CoV-2 virus had no geographic boundaries and no matter how strict the restrictions in specific regions, each of the variants of concern, like Delta and Omicron, eventually spread nearly everywhere. Just looking at the phylogenetic tree for the whole of the pandemic,

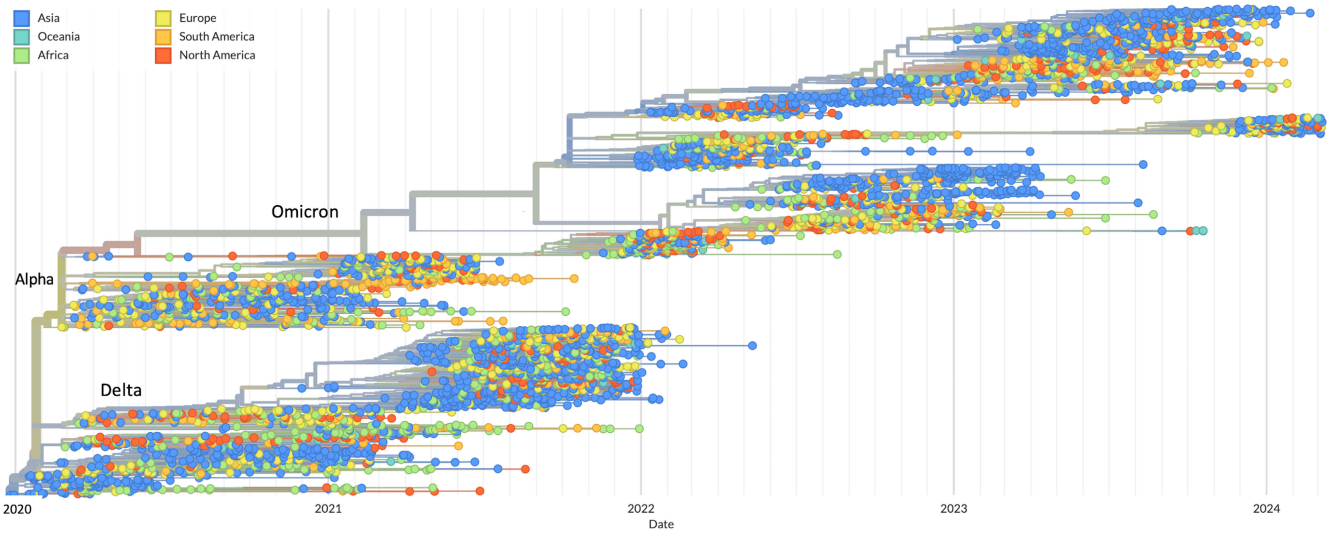


FIGURE 2 Phylogenetic tree of SARS-CoV-2 based on sample genomes from December 2019 until March 2024. Colours indicate the continental origin of the sample. Data from Nextstrain.

coloured by the continent where each case was sequenced, is enough to show the complexity and near intractability of understanding geographic spread in detail. (Figure 2).

To demonstrate the kind of spatial–temporal process-based cartography of genomic changes that developed during the pandemic, I will describe the mapping of a single mutation in the spike protein of SARS-CoV-2.

2.1 | The geography of mutations and amino acid changes

SARS-CoV-2 is an RNA virus with a little less than 30,000 nucleotides in its genome. These nucleotides make up the coding and non-coding regions of the virus and define the proteins that form its structure and how it functions. Many of the most critical changes and mutations that took place in SARS-CoV-2 over the course of the pandemic and that gave rise to new variants, like Omicron, occurred in the spike protein. This protein is the region responsible for giving the virus the tools to attach to and fuse with the membrane of a host cell. The spike has several parts, with the two called S1 and S2 the most important, as they participate either in catalysing the attachment of the virus to a human cell or help to fuse the virus to the cell wall. It was the region targeted by the vaccines.

Mapping the movement across the globe of changes that took place in the amino acid structure in this protein of SARS-CoV-2 was critical to forming public health policy, as changes in this region had the potential to alter the transmissibility of the virus and its ability to bind to human cells (Huang et al., 2020; Jackson et al., 2022). Because this was a dynamic and out of equilibrium process, the maps made had to consider time as an explicit mapping variable.

Examining the phylogenetic tree for SARS-CoV-2, and colouring its branches by changes in codon 681 of the spike protein, one can easily see the temporal evolution of the amino acids in this part of the genome (Figure 3). The fact that each of the dots on the tree also has a geospatial component means we can map the dynamics of these changes in any geographic region where we have sufficient data.

The phylogenetic tree indicates three important changes in the codon, with the initial amino acid Proline (P) changing to Histidine (H) and Arginine (R) at particular times and places. To truly appreciate the dynamics of both the changes happening in the genome, and its geographic movement into and out of human hosts, it was necessary to take a process-based approach to the map and to the visualisation of the data.

The whole process of codon change and geographic movement can be animated and compressed into shorter time scales that indicate where and when the amino acid changes in the spike protein took place (Figure 4, Video 1).

Although this simulation focuses on Europe, and is only one of the hundreds of mutations and amino changes in the SARS-CoV-2 genome that have been recognised through genomic sequencing over the course of the last few years, it can be thought of as a model for the innovations that took place during the pandemic (Hessler, 2020b).

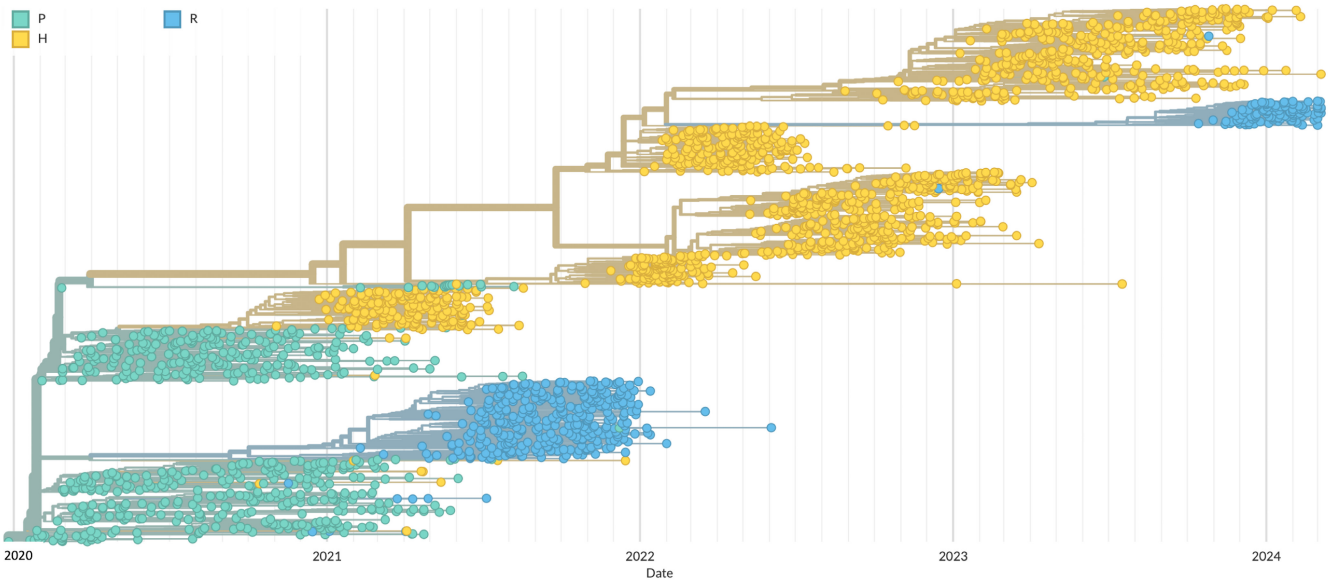


FIGURE 3 Phylogenetic tree of SARS-CoV-2, coloured for changes at codon 681 of the spike protein showing temporal evolution and amino acids in this part of the genome. Data from Nextstrain.

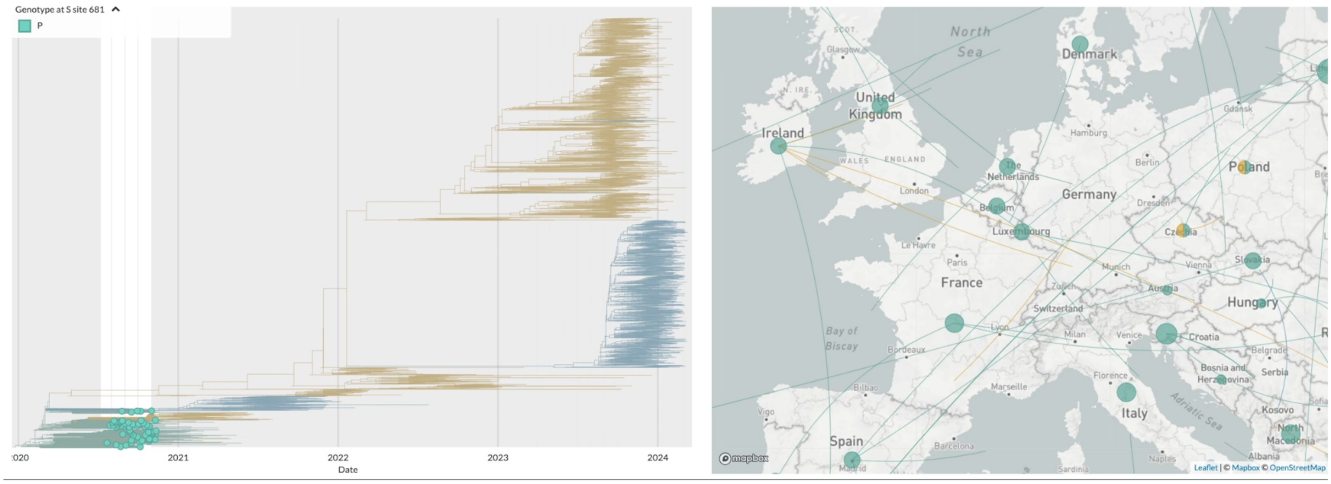
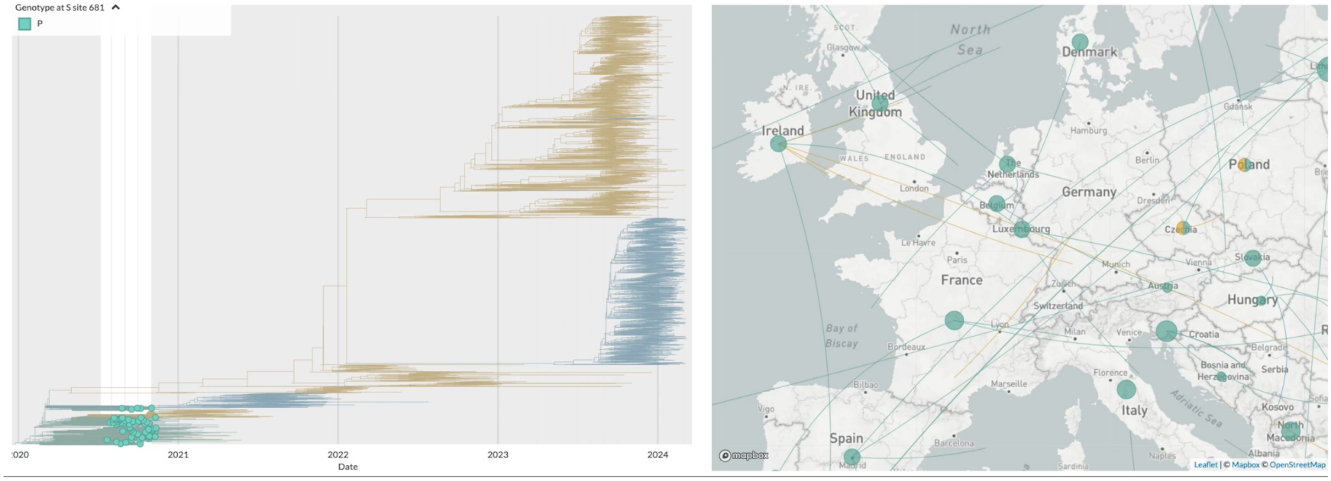


FIGURE 4 Simulation of the temporal changes at codon 681 in the spike protein for European cases. The simulation animates the geographic and temporal change from Proline (P) to Histidine (H) and Arginine (R) at particular times in the evolution of the virus. Data from Nextstrain.

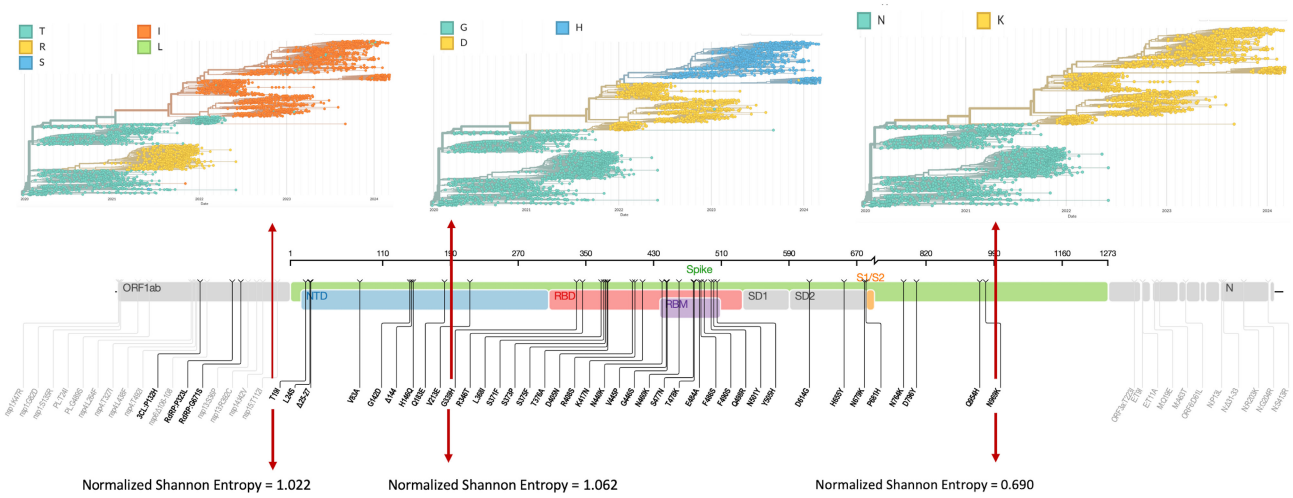
Identifying locations along the genome of the major sites of amino acid change and mapping their movement around the globe was not only a scientific problem, but became an important question for public health officials and legislators during the most intensive periods of COVID-19's spread. Several measures of diversity were developed during the pandemic that gave researchers the tools to narrow down those regions of the genome where large numbers of changes were occurring, and then to trace and map them geographically through space.

One of the most important of these measures was information entropy (Figure 5). Information entropy has applications in many fields and can be thought of as a measure of the diversity of change or mutations at the site of a particular amino acid or nucleotide (Leinster, 2021; Vopson & Robson, 2021).

Values of the normalised information entropy near 1 are regions of high diversity (Mullick et al., 2021). Using this method, geographers and epidemiologists could compare regions of the genome and geographically map how particular amino acid changes had moved. It was therefore possible to trace the geographical distribution of variant introductions and important structure changes in the virus that might impact its transmissibility, for example, those at high-entropy codon position 19 in the spike protein (Figure 6).



VIDEO 1 Simulation of the temporal changes at codon 681 in the spike protein for European cases. The simulation animates the geographic and temporal change from Proline (P) to Histidine (H) and Arginine (R) at particular times in the evolution of the virus. Data from NextStrain.



$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

FIGURE 5 Diagram of the amino acid codons of the spike protein of SARS-CoV-2 showing the changes and locations of the changes to the XBB variant of Omicron. The diagram shows the Shannon entropy for three codons (19, 339, and 969) and their respective phylogenetic trees. Larger entropy values indicate that more genetic variation has been observed at those locations in the dataset. Data from Nextstrain. Amino change data and illustration for XBB from Stanford University Coronavirus Antiviral and Resistance Database (https://covdb.stanford.edu/).

Studies using this combination of geospatial metadata and the changes in the structure of the SARS-CoV-2 genome have informed theories of the virus's origins (Crits-Christoph et al., 2023), how it first moved into the US through multiple genetic strains in early 2020 (Worobey et al., 2020), and how it first entered into Europe (Worobey et al., 2020). By exploiting small differences and changes in the structure of viral genomes, epidemiologists and virologists began to understand how it spread geographically and could make informed hypotheses about the origins of the important variants of concern.

For example, Michael Worobey, from the University of Arizona, showed that there were actually two independent introductions to Washington state during the earliest weeks of the pandemic, based on the analysis of genomic sequences, the timing of infections, and most critically an accurate geographic map of cases from the state.

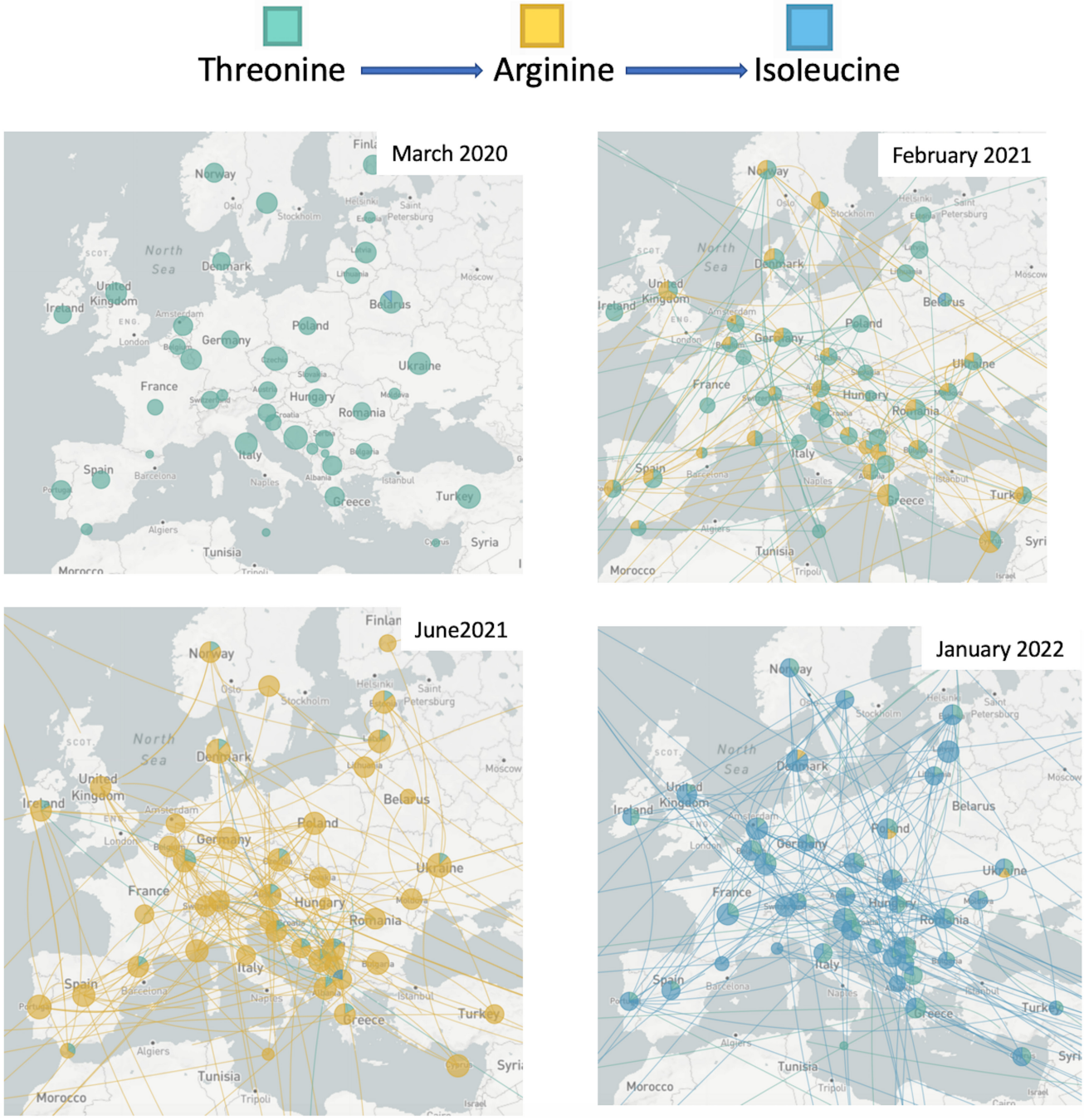


FIGURE 6 Static temporal map snapshots of the changes in codon 19 of the spike protein of SARS-CoV-2 in Europe. Data from Nextstrain.

Worobey and his collaborators used methods like those in our example and focused on just a few changes in the genome, and were able conclude that 'we must seriously consider the possibility that there were multiple introductions of genetically similar viruses into the United States' (Worobey et al., 2020, p. 567). Using geographic maps of the mutations of these slightly different genomes, they were also able to show that the genomic ancestors of one of these introductions of SARS-CoV-2 died out, while the other spread across the country.

Examples of this type of mapping and geographic analysis have revealed the power of the combination of large amounts of genomic, temporal, and geospatial information to aid policy-makers and epidemiologists in understanding the dynamics of a disease that changed the world. The combination of real-time genomic sequencing with geographic information created a new phylogeographic subdiscipline, which in the midst of a pandemic became critical to public policy and response.

As geographers, epidemiologists, and virologists continue to work through this massive amount of data and to investigate the complex phylodynamics of the pandemic and its aftermath, we can now only imagine what new combinations of advanced bioinformatic methods, machine learning algorithms, and GeoAI will yield in the near future (Cahuantzi et al., 2024).

3 | FINAL LESSONS

During the most difficult periods of the pandemic, I presented large amounts of data and the methods being used to map and analyse the spread of COVID-19 to policy-makers and legislators in the US. They all struggled to understand both the science and the impact the pandemic was having around a globe.

Speaking to policy-makers remotely, either in large groups over Zoom (sometimes more than 200 at a time) or individuals and staff in Senate offices, it was clear in the very beginning of the outbreak that much of the kind of geographically informed bioinformatic data that was being produced and that they needed was new to them. Most had no understanding of the power of geographic information systems (GIS), the methods of bioinformatics, and more importantly how geographic information could help in their decision-making. For most, this situation did not last long.

The lessons learned from the mapping tools developed and the geospatial analysis accomplished during the SARS-CoV-2 pandemic will be critical to our response to the emergence of a future new pathogen. One of the most lasting and important things accomplished during our struggle against COVID-19, at least from this cartographer's perspective, is the introduction of a whole generation of policy-makers to the power of geographic information.

Case mapping dashboards, the geography and network graphs of the transmission of variants, and the complexities of the geospatial spread of mutations all became common discussion points and part of the decision-making process. Many policy-makers and legislators came to appreciate how the discipline of cartography, broadly conceived at its multidisciplinary best, was critical to the world's response to the pandemic, and hopefully will remember how maps might save lives in the future.

ACKNOWLEDGEMENTS

The author would like to thank the Congressional Relations Office of the Library of Congress and the Library of the US House of Representatives, under whose direction many of the advising and policy discussions described in this paper were accomplished. I would also like to thank John Wertman, Program Manager for Public Policy at esri, and Paulette Hasier, Chief of the Geography and Map Division of the Library of Congress, for their encouragement of this work during the height of the COVID-19 pandemic. I would also like to acknowledge the data collection and visualisation platforms used in this paper, at the *biomap-lab*, Nextstrain and GISAID.

DATA AVAILABILITY STATEMENT

All genomic and geospatial data used in this paper are openly and publicly available at Nextstrain (www.netxstrain.org) or through GISAID (https://gisaid.org).

ORCID

John Hessler https://orcid.org/0009-0005-7049-0433

REFERENCES

Cahuantzi, R., Lythgoe, K., Hall, I., Pellis, L. & House, T. (2024) Unsupervised identification of significant lineages of SARS-CoV-2 through scalable machine learning methods. *Proceedings of the National Academy of Sciences of the United States of America*, 121, 1–8. Available from: https://doi.org/10.1073/pnas.231728412

Cazelles, B., Chavez, M., Magny, G.C., Guégan, J. & Hales, S. (2007) Time-dependent spectral analysis of epidemiological time-series with wavelets. *Journal of the Royal Society, Interface*, 4, 625–638. Available from: https://doi.org/10.1098/rsif.2007.0212

Crits-Christoph, A., Gangavarapu, K., Pekar, J., Moshiri, N., Singh Phd, R., Levy, J. et al. (2023) *Genetic evidence of susceptible wildlife in SARS-CoV-2 positive samples at the Huanan wholesale seafood market, Wuhan: Analysis and interpretation of data released by the Chinese Center for Disease Control*. Available from: https://zenodo.org/records/7754299

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C. et al. (2018) Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, 34, 4121–4123. Available from: https://doi.org/10.1093/bioinformatics/bty407

Hessler, J. (2020a) More than just cases II: Using artificial neural networks to map COVID-19 social media sentiment. *The Portolan: Journal of the Washington Map Society*, 109, 26–28.

Hessler, J. (2020b) More than just cases: Mapping the spread of COVID-19 using geospatial nucleotide mutations and pathogen Phylodynamics. *The Portolan: Journal of the Washington Map Society*, 108, 47–50.

Hessler, J. (2023) *Earliest Cases of COVID-19 in China* [Web mapping application]. Available from: https://arcg.is/CCfvP

Huang, Y., Yang, C., Xu, X.F., Xu, W. & Liu, S.W. (2020) Structural and functional properties of SARS-CoV-2 spike protein: Potential antivirus drug development for COVID-19. *Acta Pharmacologica Sinica*, 41, 1141–1149. Available from: https://doi.org/10.1038/s41401-020-0485-4

Jackson, C., Farzan, M., Chen, B. & Choe, H. (2022) Mechanisms of SARS-CoV-2 entry into cells. *Nature Reviews: Molecular Cell Biology*, 23, 3–20. Available from: https://doi.org/10.1038/s41580-021-00418-x

Leinster, T. (2021) *Entropy and diversity*. Cambridge, UK: Cambridge University Press.

Marvin, S., McFarlane, C., Guma, P., Hodson, M., Lockhart, A., McGuirk, P. et al. (2023) Post-pandemic cities: An urban lexicon of accelerations/decelerations. *Transactions of the Institute of British Geographers*, 48, 452–473. Available from: https://doi.org/10.1111/tran.12607

Mullick, B., Magar, R., Jhunjhunwala, A. & Barati Farimani, A. (2021) Understanding mutation hotspots for the SARS-CoV-2 spike protein using Shannon entropy and K-means clustering. *Computers in Biology and Medicine*, 138, 1–8. Available from: https://doi.org/10.1016/j.compbiomed.2021.104915

Peeples, L. (2022) Lessons from the COVID data wizards. *Nature*, 603, 564–567. Available from: https://doi.org/10.1038/d41586-022-00792-2

Pekar, J., Magee, A., Parker, E., Moshiri, N., Izhikevich, K., Havens, J.L. et al. (2022) The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. *Science*, 377, 960–966. Available from: https://doi.org/10.1126/science.abp8337

Sagulenko, P., Puller, V. & Neher, R. (2018) TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, 4, 1–9. Available from: https://doi.org/10.1093/ve/vex042

Sokhansanj, B. & Rosen, G. (2022) Mapping data to deep understanding: Making the most of the deluge of SARS-CoV-2 genome sequences. *mSystems*, 7, e0003522. Available from: https://doi.org/10.1128/msystems.00035-22

Sparke, M. & Anguelov, A. (2020) Contextualizing coronavirus geographically. *Transactions of the Institute of British Geographers*, 45, 498–508. Available from: https://doi.org/10.1111/tran.12389

van Dorp, L., Houldcroft, C., Richard, D. & Balloux, F. (2021) COVID-19, the first pandemic of the post-genomic era. *Current Opinion in Virology*, 50, 40–48. Available from: https://doi.org/10.1016/j.coviro.2021.07.002

Volz, E., Koelle, K. & Bedford, T. (2013) Viral Phylodynamics. *PLoS Computational Biology*, 9, e1002947. Available from: https://doi.org/10.1371/journal.pcbi.1002947

Vopson, M. & Robson, S. (2021) A new method to study genome mutations using the information entropy. *Physica A: Statistical Mechanics and its Applications*, 584, 1–9. Available from: https://doi.org/10.1016/j.physa.2021.126383

Worobey, M., Levy, J., Malpica Serrano, L., Crits-Christoph, A., Pekar, J.E., Goldstein, S.A. et al. (2022) The Huanan seafood wholesale market in Wuhan was the early epicenter of the COVID-19 pandemic. *Science*, 377, 951–959. Available from: https://doi.org/10.1126/science.abp8715

Worobey, M., Pekar, J., Larsen, B., Nelson, M., Hill, V., Joy, J. et al. (2020) The emergence of SARS-CoV-2 in Europe and North America. *Science*, 370, 564–570. Available from: https://doi.org/10.1126/science.abc8169

How to cite this article: Hessler, J. (2024) To save lives: Lessons of a pandemic cartographer. *Transactions of the Institute of British Geographers*, 00, e12703. Available from: https://doi.org/10.1111/tran.12703