Genomic epidemiology and phylodynamics for county-to-county transmission of SARS-CoV-2 in Minnesota, from 19A to Omicron

Matthew Scotch[1,2,3], Kimberly Lauer[4], Eric D. Wieben[5], Yesesri Cherukuri[6], Julie M Cunningham[7], Eric W Klee[4,8], Jonathan J. Harrington[8], Julie S Lau[8], Samantha J McDonough[8], Mark Mutawe[8], John C. O'Horo[9], Chad E. Rentmeester[7,10], Nicole R Schlicher[7], Valerie T White[7], Susan K Schneider[7], Peter T Vedell[4], Xiong Wang[11], Joseph D Yao[7], Bobbi S Pritt[7#], Andrew P Norgan[7#*]

1. Research Affiliate, Mayo Clinic Arizona, Phoenix, AZ USA
2. Biodesign Center for Environmental Health Engineering, Arizona State University, Tempe, AZ USA
3. College of Health Solutions, Arizona State University, Phoenix, Arizona USA
4. Department of Quantitative Health Sciences, Mayo Clinic Rochester, Rochester, MN, USA
5. Department of Biochemistry and Molecular Biology, Mayo Clinic Rochester, Rochester, MN, USA
6. Research Services, Mayo Clinic Florida, Jacksonville, FL, USA
7. Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA
8. Center for Individualized Medicine, Rochester, MN, USA
9. Division of Public Health, Infectious Diseases, and Occupational Medicine, Mayo Clinic, Rochester, Minnesota, USA
10. Saint Mary's University of Minnesota, Winona, MN, USA
11. Minnesota Department of Health, St. Paul, MN, USA

Joint Senior Authors

*Contact: Norgan.Andrew@mayo.edu

Keywords: Epidemiology; Computational Biology; SARS-CoV-2; Phylogeography; Minnesota; High-Throughput Nucleotide Sequencing

Abstract

SARS-CoV-2 has had an unprecedented impact on human health and highlights the need for genomic epidemiology studies to increase our understanding of the evolution and spread of pathogens and to inform policy decisions. Most efforts have focused on international or country-wide transmission, which are unable to highlight state-wide trends.

We sequenced virus genomes from over 22,000 patients tested at Mayo Clinic Laboratories between 2020-2022 and leveraged detailed patient metadata to describe county-to-county spread in Minnesota. Our findings indicate that spread in the state was mostly dominated by viruses from Hennepin County, which contains the largest metropolis. For many counties, we found that state government restrictions eventually led to a decrease in the diversity of circulating viruses from other counties and that their complete removal in May of 2021 saw a drastic revert to levels at or greater than those observed during the months before.

We also linked over 14,000 genomes with patient risk characteristics and infection-related phenotypes from the Mayo Clinic electronic health record. We found that the genetic relationship of Omicron viruses was structured by clinical outcomes when stratifying by patient risk factor and variant of concern. However, we were unable to identify nucleotide variants that drove this association.

MAIN TEXT

Genomic epidemiology has provided valuable insight into transmission, evolution, and public health surveillance of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the cause of coronavirus disease 2019 (COVID-19). This has been feasible, in large part, to unprecedented viral genomic sequencing efforts across the globe. As of the 24th of July 2022, there are over 12.1M virus sequences in GISAID [1] and over 5.9M in NCBI Virus [2] and GenBank [3]. Studies that focus on localized spread such as counties within a state or province can highlight and uncover transmission events that could inform statewide surveillance and prevention efforts. However, there have been limited SARS-CoV-2 genomic epidemiology studies at this geographic level in the United States. Work by Moreno *et al*. [4] examined evolution and spread of SARS-CoV-2 among two counties (Dane and Milwaukee) in Wisconsin, from the start of the pandemic until the end of April 2020, based on analysis of 247 new full-length SARS-CoV-2 genomes combined with sequences in GISAID. Using this data, they derived county data on synonymous and non-synonymous single nucleotide variants (SNVs), and performed a variety of phylogenetic analyses (using Nextstrain [5] and BEAST2 [6]), determined R0 for their region, and examined the number and timing of introductions to the two counties and how each introduction subsequently impacted the local transmission [4]. The authors were ultimately able to conclude that early transmission within Dane County was not due to its initial introduction followed by local spread, but rather multiple later introductions into the region [4]. In other work, Deng *et al.* [7] sequenced 36 clinical samples from different counties in Northern California, sourced from the California Department of Public Health, Santa Clara County Public Health Department, and the University of California San Francisco. Their phylogenetic analysis revealed multiple California clusters including Santa Clara County, Solano County, San Benito County, as well as lineages from Washington State and Europe [7]. They also identified several early notable SNVs including D614G in the Spike protein [7]. Work by Müller *et al.* [8] examined SARS-CoV-2 introduction and spread in the State of Washington including at the county level

early in the pandemic from February to July 2020 with a focus on the 614G variant and Google workplace mobility data. Alpert *et al*. created county-level risk maps for international importation of early Alpha variants via air transport [9]. Other work such as Valesano *et al*. [10], Holland *et al.* [11], and Currie *et al.* [12] examined spread on college campuses.

Although these studies have provided within-state snapshots into the genomic epidemiology of SARS-CoV-2, they did not consider how localized evolution and spread within a state changed over multiple years of the pandemic with the introduction and circulation of different variants of concern such as Alpha, Delta, and Omicron. As the virus continues to evolve, such studies are needed and can inform local and state public health response by highlighting the roles of various counties on state-wide transmission. In addition, they can elucidate the impact of out of state introductions on local spread which can inform policy on travel.

SARS-CoV-2 genome distribution by county and patient characteristics

We sequenced SARS-CoV-2 genomes from genomic material isolated from clinical samples of patients tested for SARS-CoV-2 infection at Mayo Clinic Laboratories (Extended Data Figs. 1-2) over a two-year period from March 2020 to March 2022. We combined these sequences with additional genomes generated for surveillance purposes by the Minnesota Department of Health (MDH) and performed Bayesian phylodynamics to understand county-to-county spread as well as the impact and timing of introductions into the State of Minnesota (see Methods). For subgroup analysis, we leveraged the Mayo Clinic's electronic health record (EHR) to link viral genomes to patient characteristics (demographics) and disease phenotypes (hospitalized ICU, etc.) to describe the genetic structure of variant-specific genomes in relation to clinical severity of disease.

Most of the patients from whom we collected a biological specimen and generated a SARS-CoV-2 genome resided in the State of Minnesota (96%) (Extended Data Table 1). The breakdown by gender was nearly 50/50 between males and females while 50 percent of the patients were between 18-45. Fifteen percent were under 18 while 11 percent were 65 or older.

Hennepin and Ramsey Counties drives in-state transmission

We down-sampled Minnesota genomes with county metadata proportional to their case rates for a final dataset of 4,586 genomes representing 20 counties (Extended Data Table 2). We implemented Bayesian phylodynamic models to examine transmissions in Minnesota from March 2020 to February 2022 (see Methods). We recorded Markov jumps [13] to estimate the timing of introductions and their directionality. Our analysis shows that Hennepin County, the largest population in Minnesota (which includes Minneapolis, the most populated city), drove the transmission of SARS-CoV-2 viruses in the state. This includes the formation of earlier clades for 20A, 20B, 20C, and 20G, as well as variants of concern Alpha, Delta, and Omicron. Ramsey County, the second most populated county (which includes St. Paul, the Capital and second most populated city) also contributed to early evolution and spread (such as 20G) however it had greater impact later in the pandemic including variants of concern Alpha and Delta (Fig. 1A).

Markov jump estimates (Fig. 1B) show that transmission of SARS-CoV-2 within the state largely originated from Hennepin and Ramsey counties (thick arcs and wider fragments at the outer circle). However, we also note existence of transmission back to these areas (white space between arc points and outer fragment) from bordering counties such as Anoka and Dakota. Analysis of the timing of introductions between counties shows the magnitude of Hennepin as a source for transmission throughout the pandemic (Extended Data Fig. 3, red dots).

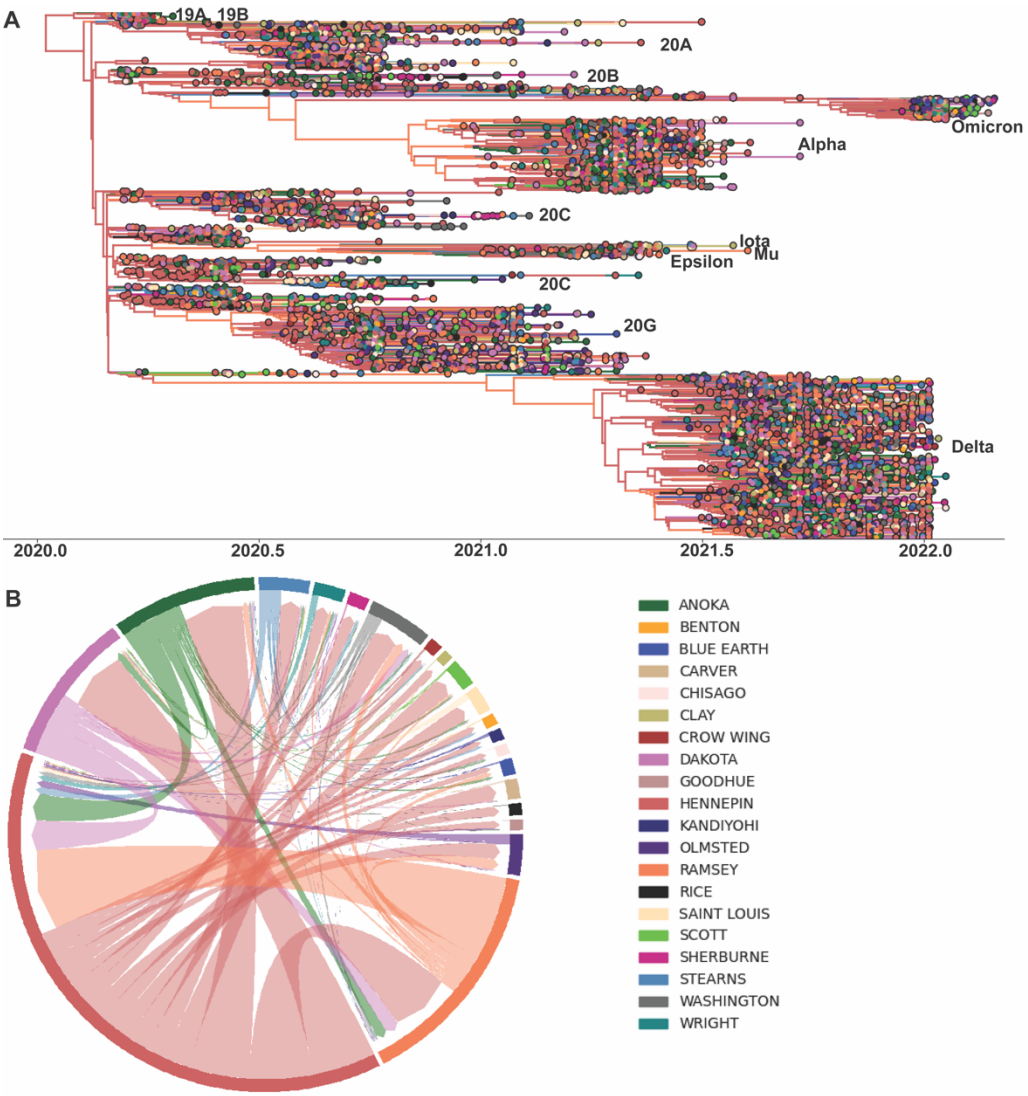


Fig. 1. County-to-county evolution and spread. A) Maximum clade credibility (MCC) tree of 4,586 SARS-CoV-2 genomes from twenty Minnesota counties. We see that Hennepin County (the root) dominated the early introductions into the other counties. We annotate the different clades by Nextclade-assigned names or VoCs [14]. B) Markov jumps between Minnesota Counties as shown via a Chord diagram. The colors for both diagrams represent the counties depicted in the legend.

We measured the ratio of introductions to total viral flow into and out of each county by month from March 2020 to January 2022. The top six sampled counties (Fig. 2A) contain considerably less uncertainty around the posterior mean ratio (depicted by the shaded regions of the 95% Bayesian highest posterior density in Fig2A and Fig2B) than the remaining counties

(Supplemental Fig. 1). While Anoka, Dakota, Ramsey, Stearns, and Washington were all fueled by introductions early in the pandemic, they experienced different trends over the remaining time of the pandemic. Anoka, Ramsey, and (to a lesser extent) Stearns experienced a relatively equal divide of both introductions into, and viral flow out of, their counties for the remaining duration of the study period, while Dakota and Washington consistently experienced more introductions than viral flow out of their counties. Meanwhile, Hennepin County showed a drastically different trend than all others as it consistently produced more viral flow into other Minnesota counties over the nearly two-year period. However, it did experience brief periods of fluctuation such as a spike in the ratio of introductions towards the end of 2020 and early 2021 (likely driven by out of state sources [Fig. 3]). This increase in Hennepin likely contributed to a decrease in viral flow into other counties (Fig. 2) as the Alpha variant began to circulate in the state.

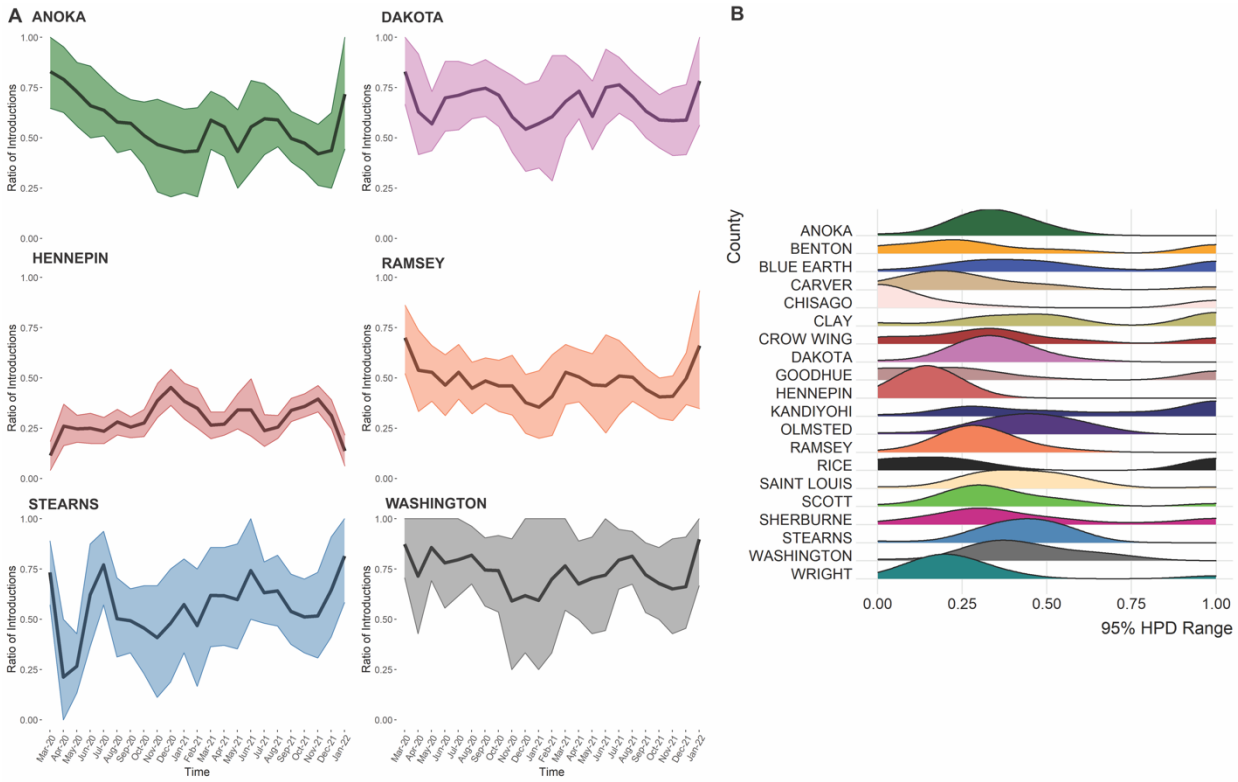


Fig. 2. Ratio of introductions to total viral flow into and out of the respective county by month for the six most sampled counties from March 2020 – January 2022. A) We show the posterior mean ratio and 95%

Bayesian highest posterior density interval. B) Many counties contain considerably more uncertainty around the median ratio given their limited number of sequences as depicted in the ridgeline plot of 95% HPD range. In Supplemental Fig. 1, we show the illustrations for all twenty counties.

While Hennepin County was also contributing to in-state transmission, it was also steadily dominating the harboring of SARS-CoV-2 viruses as illustrated by the Markov reward times (Extended Data Fig. 4). Ramsey county also consistently contributed to the proportion of viruses but to a much less extent.

Restrictions eventually reduced spatial mixing of SARS-CoV-2, but their elimination saw an immediate reverse of its effectiveness

We assessed county-specific virus diversity via a normalized Shannon diversity index (Extended Data Fig. 5) that we computed based on the duration of time associated with continuous partitions of the phylogeographic tree as determined by Markov jumps [15] (see Methods). The index, in this context, measures the degree of spatial structure (based on counties) during the evolution and spread of SARS-CoV-2 viruses in Minnesota. A value of *0*, indicates exclusive spatial structure such as an outbreak contained to only one county [15]. Conversely, a value of *1*, suggests significant spatial mixing of SARS-CoV-2 between counties [15]. Most counties show low to intermediate (0.25 to 0.5 Shannon) spatial mixing with brief periods of waxing and waning. The two dotted vertical lines indicate changes in state-wide policy. The second line on May 28, 2021 indicates the end of all COVID-19 restrictions in the state [16]. For many counties, this timepoint is preceded by a drop in diversity suggesting the restrictions were eventually effective in suppressing county-specific diversity. However, removal of state restrictions appears to have resulted in an immediate and sharp rise (to levels at or above before the drop) in diversity. This all precedes a spike in COVID-19 cases and the circulation of the Omicron variant into the state.

Hennepin County received the vast majority of out of state introductions

We analyzed the timing and impact of introductions into the Minnesota by adding additional genomes from NCBI GenBank [3] as part of an international dataset used in NextStrain [5] (see Methods). To address the computational burden of adding additional sequences to our already large dataset, we aggregated the additional samples into discrete traits *International* and *USA* and grouped counties with less sequences into areas in the state such as Southern, Central, and Northern Minnesota (Extended Data Table 3). We focused on the timing and source of introductions into the state during the pandemic (Fig. 3) as estimated from our maximum clade credibility tree (Extended data Fig. 6). The earliest introduction into Minnesota was to Hennepin County from a domestic source on around January 23, 2020 (depicted with an arrow in Fig. 3). This is about one month before the first patient in the state, a man from Ramsey County (which borders Hennepin), developed symptoms and around six weeks before (March 6, 2020) the Department of Health confirmed the infection [17]. The first international introduction occurred in Hennepin on around February 10 with another shortly thereafter on around February 13. The first two county-to-county introductions were estimated to originate from Hennepin to somewhere in Central Minnesota around February 28 and from Hennepin to Ramsey around March 1. International introductions were most abundant in Hennepin County (home to the Minneapolis/St. Paul International (MSP) airport) totaling 119 separate incidents over the two-year pandemic. The domestic (USA) introductions were most frequent to Hennepin as well (n=89) followed by Southern Minnesota (n=57), potentially from nearby states.

Fig. 3. International and out of state introductions into Minnesota. The colors correspond the location of origin. The circles represent a county or region in Minnesota. The triangles represent an outside location, USA or International. The x-axis of the scatter plot displays the posterior time of the introduction while the height is jittered for visualization only. We use an arrow to show the first introduction into the state which occurred in Hennepin County at

around January 23, 2020 (from a domestic location). We used Baltic to extract introductions (migration events) along the annotated branches of the phylogeographic tree.

Evolutionary relationship of Omicron viruses structured by clinical outcome

We linked a subset of our SARS-CoV-2 genomes to patient characteristics and COVID-19-related outcomes in the Mayo Clinic electronic health record (see Methods). We first stratified genomes by variant and then the patient's calculated risk of hospitalization based on the Monoclonal Antibody Screening Score (MASS) score [18] where a patient receiving a score of 1 or 2 was considered "mild" risk, a score of 3 was considered "medium", and, a score of 4 or more "severe". We labeled each taxa by outcome category (mild, moderate, severe, and death) based on a classification score from the Adaptive COVID-19 Treatment Trial (ACTT) [19] (Extended Data Table 4). For each *VoC-risk group* phylogeny, we measured the probability of matching clinical outcome traits to adjoining taxa in the tree[20]. For all Alpha and Delta scenarios, we found no relationship between the structure of the tree and COVID-19 clinical outcome. However, for our low risk (N=816) and high risk (N=390) Omicron genomes we found significant P-values for the association index (AI) metric (0.03 and 0.04 respectively) suggesting that the phylogeny (genetic relationship) of Omicron viruses, when stratifying for patient risk of COVID-19 hospitalization, is structured by clinical outcome.

While it has been demonstrated that the risk of severe clinical disease is much lower for Omicron viruses overall than previous versions of the virus [21], the possible relationship of genomes between mild and moderate clinical outcomes warrants further investigation. We studied single nucleotide variants (Fig. 4) and maximum clade credibility trees (Fig. 5) of our mild and moderate clinical cases of Omicron to identify any differences that could partially explain disease outcome among patients at the same risk of COVID-19 hospitalization. We note that both the mild (N = 1,259) and moderate (N = 84) clinical outcomes were similar in terms of known Omicron variants in the Spike region including receptor binding substitutions G339D, S373P, S375F, K417N, N440K, G446, S477N, E484A, Q493R, G496S, N501Y, and Y505H.

Both groups of outcomes also had S371F and S371P mutations. In addition, there was also evidence of the R346K mutation in both (47% of mild cases and 60% of moderate cases), a known BA.1 mutation [22] which has shown to be associated with a decrease in monoclonal antibody effectiveness [23]. In Fig. 5, we show the maximum clade credibility (MCC) trees of sequences from samples taken from patients with low and high risk of hospitalization and color the taxa by clinical outcome. Both show the dominance of mild cases (in red) with more moderate cases (blue) intermingled on the *high-risk* tree. While the test for trait association via BaTS relies on a posterior collection of trees, the MCC tree represents a single view of the genetic relationship between viruses that contributed to different clinical outcomes.

Table 1. COVID-19 clinical outcome trait association results of SARS-CoV-2 phylogenies by variants for 6,932 genomes. We show the results via BaTS for association index (AI). The four potential outcome traits were: mild, medium, severe, and death, however only *mild* and *moderate* cases were observed among these patients. We did not calculate AI for datasets of Alpha *medium* (N = 17) and *high* (N = 46) and Omicron *medium* (N=91) due to low sample sizes.

VoC	**Risk Group (N)**	**Observed Mean**	**Expected Mean**	***P-value***
Alpha				
	Low (N= 671)	1.38	1.35	0.57
Delta				
	Low (4,116)	13.74	14.54	0.2
	Medium (280)	2.27	1.75	0.95
	High (659)	8.85	9.02	0.39
Omicron				
	Low (816)	3.27	3.91	0.03*
	High (390)	12.66	14.37	0.04*

*Significant at the 0.05 P-value level.

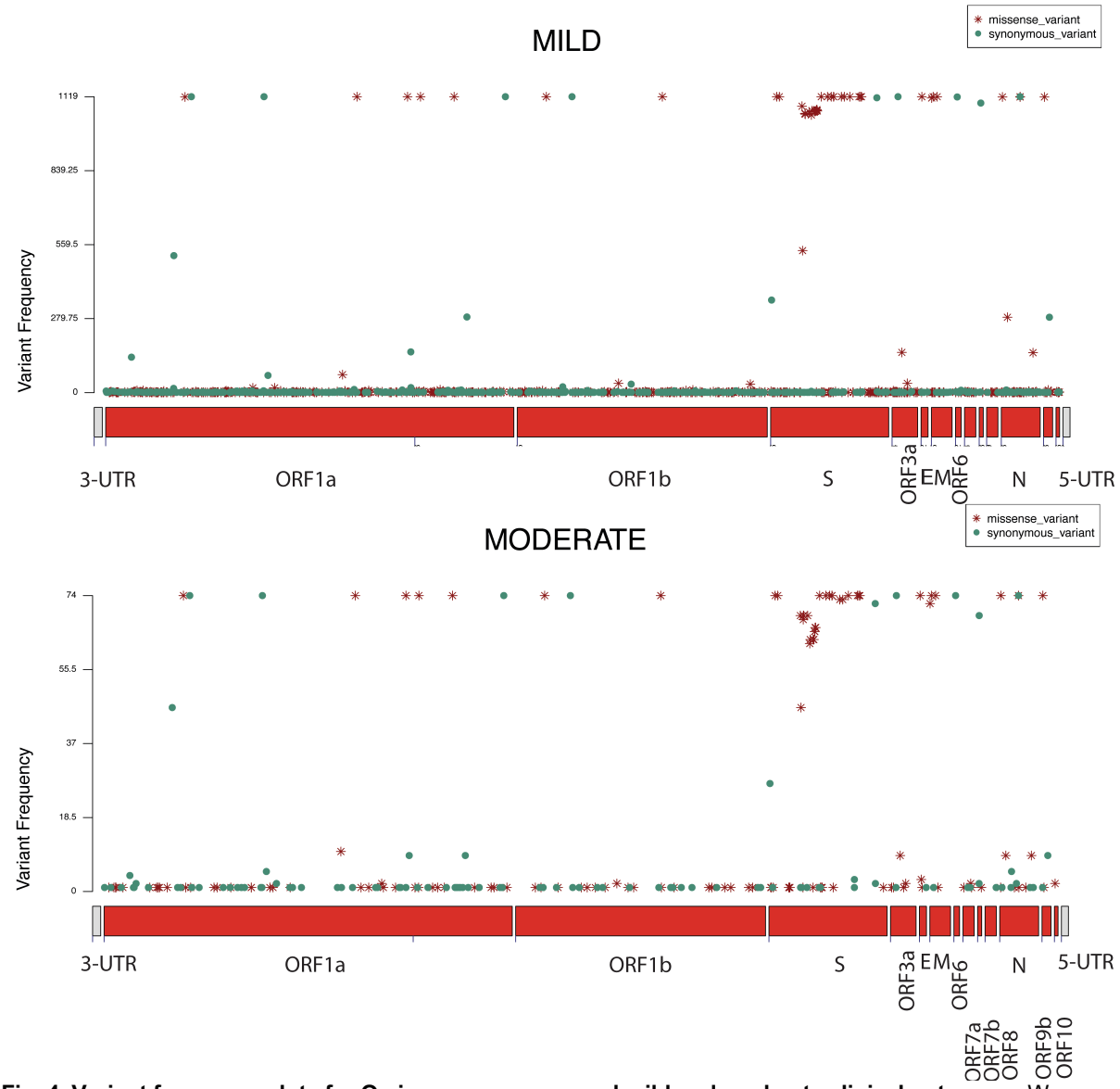


Fig. 4. Variant frequency plots for Omicron sequences and mild and moderate clinical outcomes. We combined sequences from patients of different risk levels for hospitalization (low, medium, etc.) that had either mild or moderate clinical cases.

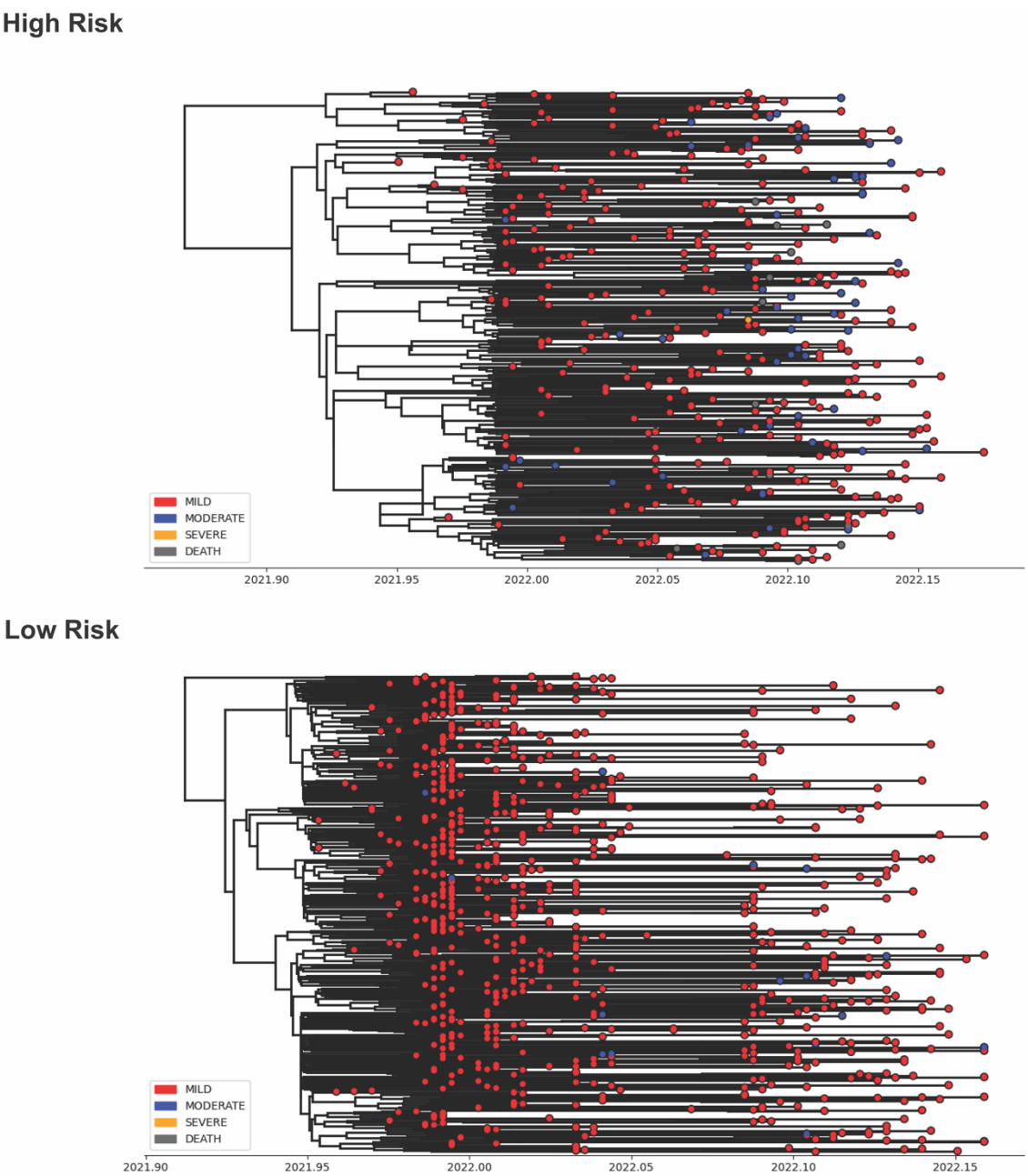


Fig. 5. Maximum clade credibility tree of sequences from patients infected with the Omicron variant that were at high (N = 390) and low (N = 816) risk of hospitalization. We color the taxa by clinical outcome (mild, moderate, severe, and death).

Discussion

We analyzed over 22,000 new genomes of patients tested at Mayo Clinic Laboratories during a two-year period in the COVID-19 pandemic. We focused our Minnesota analysis at the

county level (2nd administrative boundary) to examine the roles and relationships of virus transmission. Despite numerous efforts in genomic epidemiology, few studies have focused on county-to-county transmission in the US over most of the pandemic (including different VoCs). We expand on earlier efforts such Moreno *et al.* and [4] and Deng *et al.*[7] but include multiple variants and an extensive timeframe. We found that Hennepin County was the source for most of the county-to-county introductions after its initial introduction with the virus in early 2020 from an out of state sources. We found that state restrictions eventually led to a decrease in county-specific diversity but their complete removal in May of 2021 saw a sharp increase back to levels at or before their elimination.

We leveraged the Mayo Clinic's EHR to map and patient characteristics and clinical endpoints to over 14,000 SARS-CoV-2 genomes and found that the genetic evolution of sampled Omicron viruses in the state were structured by outcome, the majority of which were mild or moderate clinical cases. We were unable to find a specific, or a combination of, single nucleotide variants that explained the relationship of these viruses. However, the findings demonstrate the importance of linking virus sequences to patient metadata including risk factors (for confounding) and clinical endpoints arising from infection.

We note several limitations in the study including the likelihood of location-specific sampling bias. We attempted to supplement known locations of patients in our study (biased towards southeastern Minnesota) with existing sequences provided via the Minnesota Department of Health. We scaled our number of sequences to the rate of known COVID-19 cases, and, after doing so, omitted counties with a limited number of sequences. Thus, we are unable to account for virus spread from less populated areas of the state. In addition, our use of different versions of the DRAGEN pipeline over the course of our two-year study period, likely led to differences in variant frequencies across virus lineages/VoCs.

References

1 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22** (2017). https://doi.org:10.2807/1560-7917.ES.2017.22.13.30494

2 Hatcher, E. L. *et al.* Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic Acids Res* **45**, D482-D490 (2017). https://doi.org:10.1093/nar/gkw1065

3 Sayers, E. W. *et al.* GenBank. *Nucleic Acids Res* **49**, D92-D96 (2021). https://doi.org:10.1093/nar/gkaa1023

4 Moreno, G. K. *et al.* Revealing fine-scale spatiotemporal differences in SARS-CoV-2 introduction and spread. *Nat Commun* **11**, 5558 (2020). https://doi.org:10.1038/s41467-020-19346-z

5 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018). https://doi.org:10.1093/bioinformatics/bty407

6 Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **15**, e1006650 (2019). https://doi.org:10.1371/journal.pcbi.1006650

7 Deng, X. *et al.* Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* **369**, 582-587 (2020). https://doi.org:10.1126/science.abb9263

8 Muller, N. F. *et al.* Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State. *Sci Transl Med* **13** (2021). https://doi.org:10.1126/scitranslmed.abf0202

9 Alpert, T. *et al.* Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States. *Cell* **184**, 2595-2604 e2513 (2021). https://doi.org:10.1016/j.cell.2021.03.061

10 Valesano, A. L. *et al.* SARS-CoV-2 Genomic Surveillance Reveals Little Spread From a Large University Campus to the Surrounding Community. *Open Forum Infect Dis* **8**, ofab518 (2021). https://doi.org:10.1093/ofid/ofab518

11 Holland, L. A. *et al.* An 81-Nucleotide Deletion in SARS-CoV-2 ORF7a Identified from Sentinel Surveillance in Arizona (January to March 2020). *J Virol* **94** (2020). https://doi.org:10.1128/JVI.00711-20

12 Currie, D. W. *et al.* Interventions to Disrupt Coronavirus Disease Transmission at a University, Wisconsin, USA, August-October 2020. *Emerg Infect Dis* **27**, 2776-2785 (2021). https://doi.org:10.3201/eid2711.211306

13 Minin, V. N. & Suchard, M. A. Counting labeled transitions in continuous-time Markov models of evolution. *J Math Biol* **56**, 391-412 (2008). https://doi.org:10.1007/s00285-007-0120-8

14 Aksamentov, I., Roemer, C., Hodcroft, E. B. & Neher, R. A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. . *zenodo* (2021). https://doi.org:10.5281/zenodo.5607694

15 Lemey, P. *et al.* Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature* **595**, 713-717 (2021). https://doi.org:10.1038/s41586-021-03754-2

16 *Impact of Opening and Closing Decisions by state*, <https://coronavirus.jhu.edu/data/state-timeline/new-confirmed-cases/minnesota/99> (2022).

17 *Health officials confirm first case of novel coronavirus in Minnesota*, <https://www.health.state.mn.us/news/pressrel/2020/covid19030620.html> (2020).

18 Razonable, R. R., Ganesh, R. & Bierle, D. M. Clinical Prioritization of Antispike Monoclonal Antibody Treatment of Mild to Moderate COVID-19. *Mayo Clin Proc* **97**, 26-30 (2022). https://doi.org:10.1016/j.mayocp.2021.11.017

19 *Adaptive COVID-19 Treatment Trial (ACTT)*, <https://clinicaltrials.gov/ct2/show/NCT04280705> (2022).
20 Parker, J., Rambaut, A. & Pybus, O. G. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect Genet Evol* **8**, 239-246 (2008). https://doi.org:10.1016/j.meegid.2007.08.001
21 Maslo, C. *et al.* Characteristics and Outcomes of Hospitalized Patients in South Africa During the COVID-19 Omicron Wave Compared With Previous Waves. *JAMA* **327**, 583-584 (2022). https://doi.org:10.1001/jama.2021.24868
22 Ou, J. *et al.* Tracking SARS-CoV-2 Omicron diverse spike gene mutations identifies multiple inter-variant recombination events. *Signal Transduct Target Ther* **7**, 138 (2022). https://doi.org:10.1038/s41392-022-00992-2
23 Uraki, R. *et al.* Therapeutic efficacy of monoclonal antibodies and antivirals against SARS-CoV-2 Omicron BA.1 in Syrian hamsters. *Nat Microbiol* (2022). https://doi.org:10.1038/s41564-022-01170-4
24 Illumina. *bcl2fastq2 Software Release Notes*, <https://support.illumina.com/content/dam/illumina-support/documents/downloads/software/bcl2fastq/bcl2fastq2-conversion-software-v2-19-release-notes.pdf> (2017).
25 Illumina. *DRAGEN COVID Lineage*, <https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/dragen-covid-lineage.html> (2021).
26 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016). https://doi.org:10.1186/s13059-016-0974-4
27 Wickham, H., François, R., Henry, L. & Müller, K. *dplyr: A Grammar of Data Manipulation*, <https://github.com/tidyverse/dplyr> (2022).
28 Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088-3090 (2017). https://doi.org:10.1093/bioinformatics/btx346
29 Huddleston, J. *et al.* Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. *J Open Source Softw* **6** (2021). https://doi.org:10.21105/joss.02906
30 *Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months*, <https://nextstrain.org/ncov/open/global/6m> (2022).
31 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066 (2002). https://doi.org:10.1093/nar/gkf436
32 Turakhia, Y. *et al.* Stability of SARS-CoV-2 phylogenies. *PLoS Genet* **16**, e1009175 (2020). https://doi.org:10.1371/journal.pgen.1009175
33 De Maio, N. *et al. Masking strategies for SARS-CoV-2 alignments*, <https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480/3> (2020).
34 Wu, F. *et al. Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome*, <https://www.ncbi.nlm.nih.gov/nuccore/MN908947> (2020).
35 Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007 (2016). https://doi.org:10.1093/ve/vew007
36 Gu, H. *et al.* Genomic epidemiology of SARS-CoV-2 under an elimination strategy in Hong Kong. *Nat Commun* **13**, 736 (2022). https://doi.org:10.1038/s41467-022-28420-7
37 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2019). https://doi.org:10.1093/bioinformatics/bty633
38 BEAST. *Approaches for analyzing large phylogenetic datasets*, <https://beast.community/thorney_beast> (2021).

39 McCrone, J. T. *BEAST v1.10.5 pre-release of ThorneyTreeLikelihood v0.1.1*, <https://github.com/beast-dev/beast-mcmc/releases/tag/v1.10.5pre_thorney_v0.1.1> (2021).
40 Hill, V. & Baele, G. Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model. *Mol Biol Evol* (2019). https://doi.org:10.1093/molbev/msz172
41 Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* **67**, 901-904 (2018). https://doi.org:10.1093/sysbio/syy032
42 Baele, G. *et al.* Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol* **29**, 2157-2167 (2012). https://doi.org:10.1093/molbev/mss084
43 Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, vey016 (2018). https://doi.org:10.1093/ve/vey016
44 Dudas, G. *baltic*, <https://github.com/evogytis/baltic> (2016).
45 Bielejec, F. *et al.* SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Mol Biol Evol* **33**, 2167-2169 (2016). https://doi.org:10.1093/molbev/msw082
46 Guerrero-Murillo, M. & Gregori, J. *QSutils*, <https://doi.org/doi:10.18129/B9.bioc.QSutils> (2022).
47 Meredith, M. & Kruschke, J. *HDInterval*, <https://CRAN.R-project.org/package=HDInterval > (2020).
48 *COVID-19 by County*, <https://www.cdc.gov/coronavirus/2019-ncov/your-health/covid-by-county.html> (2022).
49 *Governor Walz Extends Stay Home Order in Minnesota*, <https://mn.gov/governor/news/?id=1055-430501> (2020).
50 *County Population Totals: 2020-2021*, <https://www.census.gov/data/datasets/time-series/demo/popest/2020s-counties-total.html> (2022).

Methods

RNA extraction, library preparation and next generation sequencing

From March 2020 to March 2022, we analyzed patient nasopharyngeal or mid-nasal turbinate swabs that tested positive for COVID-19 via RT-qPCR at Mayo Clinic Laboratories and had a Ct value of 28 or lower. We extracted viral RNA on the Hamilton Microlab STAR Automated Liquid Handler system (Hamilton Company, Reno, NV USA) with the use of Promega Maxwell HT Viral TNA Kit (Fitchburg, WI). We generated libraries using the COVIDSeq Test reagent kit from Illumina (San Diego, CA USA) following the manufacturer's instructions. We sequenced the pooled libraries as 100 x 2 paired end reads using the NovaSeq SP sequencing kit and Xp 2-Lane kit with NovaSeq Control Software v1.6.0. We used the Illumina RTA version 3.4.4 for base-calling.

We de-multiplexed raw sequence data into individual sample fastq files using bcl2fastq2-v2.19.0 [24]. We used Illumina's Dynamic Read Analysis for GENomics (DRAGEN) COVID Lineage software and pipeline [25] (versions 3.5.1,3.5.3, and 3.5.6) for reference-based alignment to Wuhan-1 (NC_045512.2), quality assessment, variant calling, and generation of consensus sequences.

For downstream analysis and quality control, we used the following exclusion criteria including consensus sequences that were: 1) given an overall score of *fail* by the DRAGEN pipeline due to having an insufficient amount of detectable viral reads; 2) given an overall quality score by Nextclade [14] as *bad;* 3) potentially contaminated based on presence of unusual allele frequencies (< 0.9); 3) duplicate runs; 4) positive or negative controls.

Single Nucleotide Variant Annotation and Frequency Plotting

We used the Ensembl Variant Effect Predictor (VEP) [26] and a custom gff3 file containing matching gene coordinates used by Nextstrain [5], and Wuhan 1 (NC_045512.2) as the reference, to determine synonymous and non-synonymous variants. We then combined this information into a single file containing county and state information for each sample. We used

R and the package *dplyr* [27] to extract and import the annotated variant data and the package *Karyplot* [28] to generate plots of synonymous and missense variants across the SARS-CoV-2 genome.

Phylogenetic analysis of SARS-CoV-2

We created three different datasets for phylogenetic analyses including a:

1) *Minnesota counties* dataset (N=4,586) that included full genome SARS-CoV-2 sequences from the 20 counties with the greatest number of reported COVID-19 cases as of February 28, 2022. We included sequences from March 2020 through February 2022 as well as their sampling location and collection date metadata. To partially address sampling bias, we sampled at a rate of 5 sequences per 1,000 county cases (Extended Data Table 2 and Supplemental Fig. 2) and used the *filter* module in augur [29] to distribute (as equally as possible) our heterochronous sequences by month (Extended Data Fig. 7). For each county, we attempted to include SARS-CoV-2 genomes across the two-year timeframe by supplementing our dataset with sequences provided by the Minnesota Department of Health (MDH) (and available in GISAID).

2) *International* dataset (N=6,479) that contained our Minnesota county samples as well as a global representation of sequences available via GenBank as part of an open access dataset from Nextstrain (Extended Data Table 3) [30]. We used the list of accessions to download sequences from NCBI Virus[2]. We removed sequences less than 29K nucleotides in length as well as duplicates. We included sequences from December 2019 (including Wuhan-1, GenBank accession MN908947) to February 28, 2022.

3) *SARS-CoV-2 genomes-clinical phenotype* dataset (N=14,810) of sequences with linked Mayo Clinic EHR data including COVID-19 hospitalization risk and patient outcome. We separated these data to variant-specific sets of Alpha, Delta, and Omicron. To partially adjust for confounding, for each VoC, we further divided the data by *low*, *medium*, and *high* risk of COVID-

19 hospitalization as determined by the patient's Monoclonal Antibody Screening Score (MASS) score[18].

We aligned all sequences using Mafft [31] and soft masked problematic sites within coding and non-coding regions (UTRs) using a program described in [32,33]. We created initial phylogenetic trees via Nextstrain's augur pipeline [5] with IQTree and rooted based on Wuhan-1 (MN908947) [34]. We used TempEST [35] to examine the temporal signal of our heterochronous samples and removed outliers based on visualization of root-to-tip divergence and the residuals. We used least-square dating as done in [36] to refine the root. For our Minnesota-only phylogeny, we additionally pruned the Wuhan-1 outgroup.

Phylodynamics of SARS-CoV-2 in Minnesota

Both of our Minnesota-only (0.94) and international (0.93) datasets showed a positive correlation of sampling date and genetic divergence thus suggesting a strict molecular clock [35]. Since our downstream inferencing framework utilizes a rooted and non-bifurcating starting tree, we used the program *di2multi* in the R package *ape* [37] to manipulate our refined trees as necessary.

For Bayesian inference, we leveraged a pre-release of BEAST v1.10.5 (ThorneyTreeLikelihood v0.1.1) and BEASTGen v0.3 (pre-thorney) to specify a more efficient likelihood function intended for larger sequence datasets [38,39]. We used our starting trees and specified a GTR + Γ model of DNA substitution and a non-parametric Bayesian SkyGrid coalescent model for our tree prior [40]. We ran our Markov-chain Monte Carlo (MCMC) simulation for $5x10^8$ steps and sampling every $5x10^4$ steps for our Minnesota-only dataset and 10^9 steps with a sampling rate of 10^5 for our international dataset. We checked for convergence of model parameters via Tracer v1.7.1 [41] with an ideal effective sample size (ESS) threshold of 200. We also used Tracer to produce graphs of demographic reconstruction of our estimated virus population size via our Skygrid model [40] (Supplemental Fig. 3). We generated log marginal likelihoods and evaluated population growth priors via a stepping stone and path sampling

procedure [42]. For both datasets, our results suggested the use of the non-parametric Skygrid tree prior over a constant growth model (Supplemental Table 1).

We used LogCombiner to sample 1,000 trees from the posterior distribution and used this as empirical data for ancestral state reconstruction of our location traits. For our Minnesota-only phylogeny, we populated our *location trait* by the county name of each taxa's sampling location. For our *international* set, we specified all non-US sequences as "International" and non-Minnesota US states as "USA". For computational efficiency, we kept the five counties with the greatest number of cases as independent locations and grouped the remaining fifteen counties into three discrete regions including Southern, Central, and Northern Minnesota (Extended Data Table 3). In BEAUti [43], we specified an asymmetric transmission rate matrix of *K(K*1)* where *K* is equivalent to the number of discrete locations (20 for *Minnesota-only* and 10 for *international*). We recorded Markov jumps[13] between locations to estimate the timing and source of introductions and Markov rewards for the time maintained within a location. We combined posterior log data via LogCombiner v.10.4 [43] as needed and used TreeAnnotator v.10.4 [43] to create a single maximum clade credibility (MCC) tree after 10% burn-in. We used baltic [44] for tree visualization and to extract the timing of discrete location transitions along the branches of the MCC for our estimates of introductions. We used SpreadD3 [45] to calculate the Bayes factors to identify the most parsimonious origin-destination scenarios (Supplemental Table 2).

We used two new programs, introduced in [15], as part of our phylodynamic analyses. *TreeMarkovJumpHistoryAnalyzer* samples from the posterior distribution of trees to collect the timing and location of each Markov jump [15]. We used the output from this program to calculate the ratio of introductions to total viral flow into and out of each county as described in Lemey *et al.* [15] as well as the visualization of the weights of pairwise transmission between counties via a chord diagram. *TreeStateTimeSummarizer*, which also samples from the posterior distribution of trees, notes the contiguous partitions for a given discrete state[15]. We used the output from this program to calculate the normalized Shannon entropy metric as described in Lemey *et al.*[15]. We

used this measure to assess the level of location diversity for the viruses within each county during a specified time-period. For our analysis, we used *NormShannon method in* the R package *QSutils* [46] to calculate normalized monthly diversity metrics for each county and *HDinterval* [47] for the corresponding 95% highest posterior density region.

SARS-CoV-2 variant-specific genetic characterization and clinical severity

For our 14,810 linked clinical sequences that met our previously stated sequence quality inclusion criteria, we used Bayesian Tip-association Significance (BaTS) testing as introduced in [20] to examine whether our phylogenetic tree of complete SARS-CoV-2 genomes are structured by COVID-19. The null hypothesis of this test is that adjoining tips in the tree are no more likely to have the same trait of interest than by chance [20]. We first created a posterior distribution of Bayesian phylogenetic trees via BEAST v1.10.4 [43] and Tracer v.1.7.1 [41] to check for convergence and ESS values. We wrote a Python script to format our posterior set of trees and assign tips to COVID-19 classification scores (traits) as required by BaTS. We specified a single mode of analysis with 100 randomizations.

Human subjects and ethics approval

This research was conducted under approval of ethics by the Mayo Clinic Institutional Review Board and assigned a study ID IRB#: 20-005896 and title *Large Scale Whole Genome Sequencing of SARS-CoV-2.* All our published datasets contain randomly generated study IDs and no personal identifiers. All data analysis was performed behind the Mayo Clinic's firewall.

Data availability

We have deposited the SARS-CoV-2 genomes and metadata from this study in GISAID with a list available at doi.org/10.55876/gis8.220720me. The Minnesota Department of Health sequences used in this study are available on GISAID with acknowledgments available at doi.org/10.55876/gis8.220709mv. Our GenBank international sequences were identified via the Nextstrain [30] site and obtained from NCBI Virus[2].

Acknowledgements

This publication was supported by funding from the Center for Individualized Medicine-Mayo Clinic Research. Research reported in this publication was also supported by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under Award Number R01AI164481 (to MS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to thank Gytis Dudas for his assistance with the Baltic Python library. The authors would also like to thank Scott Lunt and Rob Timmer for their assistance with the mforge high-performance computing environment. The authors acknowledge the laboratories that submitted the SARS-CoV-2 genome data to GISAID that were used in this work.

Author contributions

MS designed the genomic epidemiology study, performed the data analysis, and wrote the manuscript. APN and BSP jointly supervised the study and providing funding, reviewed the manuscript, and approved the final manuscript. CER performed data analysis, assisted in reviewing and writing the manuscript. KL, YC, PTV and EWK implemented the DRAGEN and Nextstrain bioinformatics pipeline and variant analysis pipelines, performed variant analysis, and reviewed the manuscript. JMC, JSL, MM, SJM, VTW, NRS, SKS and EDW supervised and performed the sample extraction, viral genomic sequencing, primary analysis of the sequencing and variant data, and reviewed the manuscript. JJH managed the execution of the study and reviewed the manuscript. JCO extracted the clinical data, performed primary data analysis, and reviewed the manuscript. JDY coordinated the provision of samples for quality control and sequencing, assisted with variant interpretation, and reviewed the manuscript.
XW supervised the generation of the Minnesota Department of Health genomic sequences, performed primary analysis of the MDH sequence data, and reviewed the manuscript.

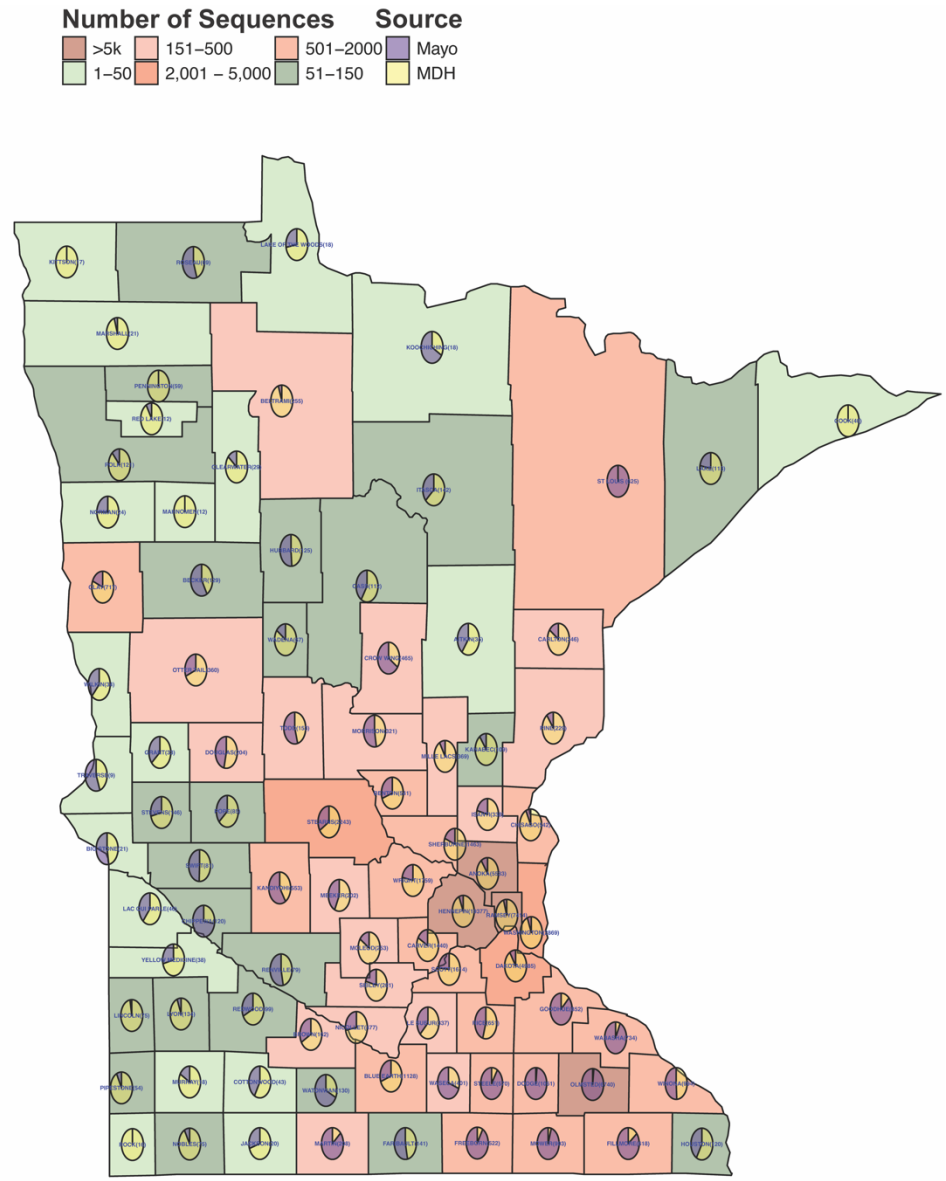
Competing interests

The authors declare no competing interests.

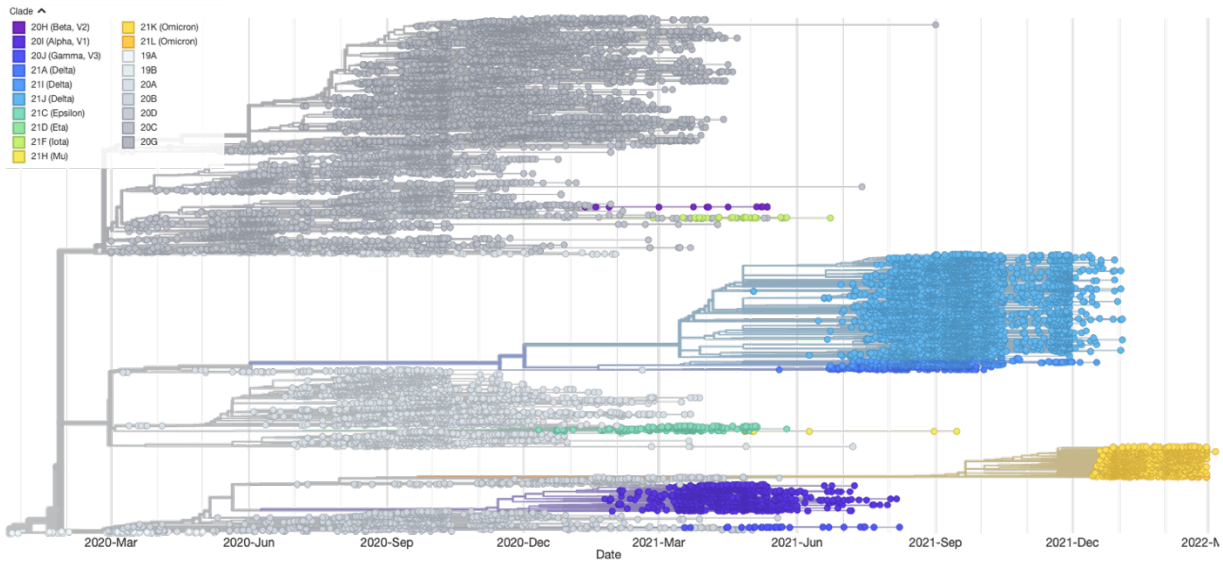
Materials & Correspondence

Please direct requests for materials or other correspondence to Andrew P. Norgan, MD, Mayo Clinic, 200 First Street SW, Rochester, MN USA 55905 (email: Norgan.Andrew@mayo.edu).

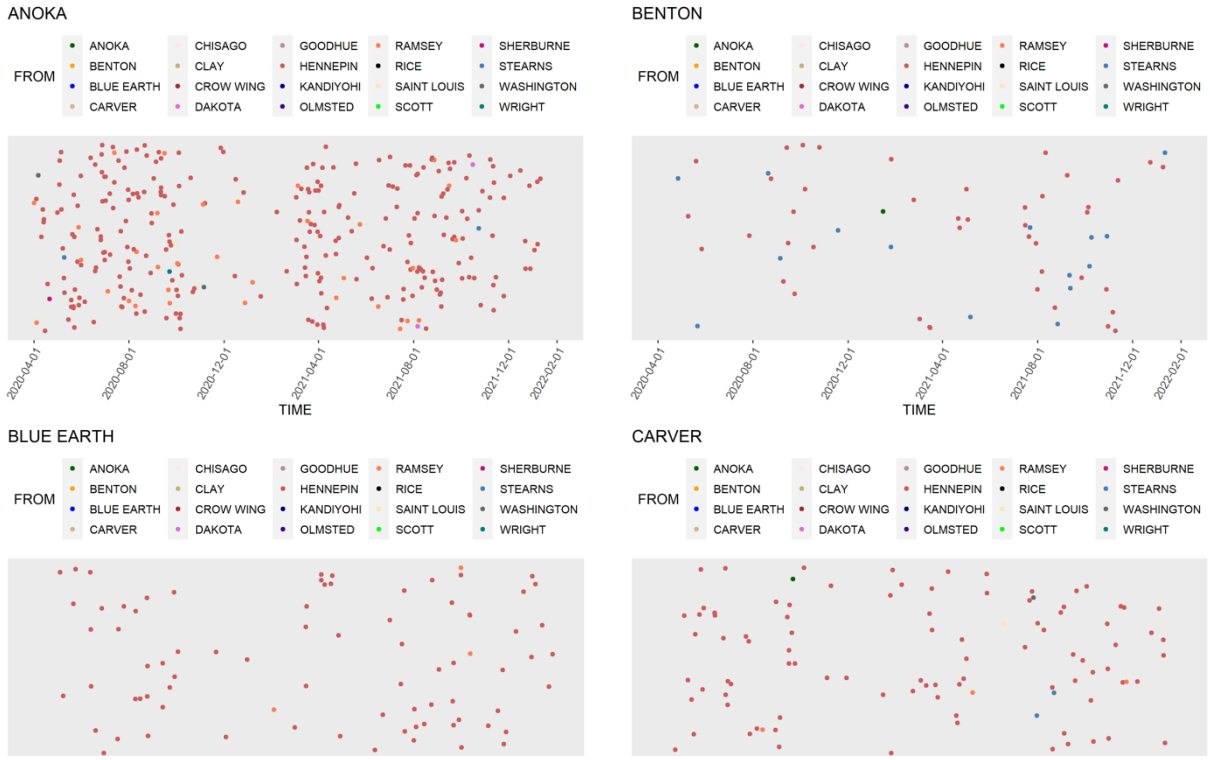
Extended Data



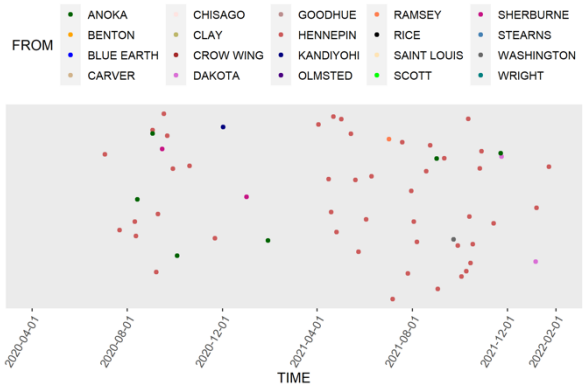
Extended Data Fig. 1. County map of Minnesota with number of sequences (N=76,875) eligible for analysis by source. Here, *Mayo* (N=21,669) represents new sequences generated from this study at Mayo Clinic Laboratories with a known sampling location in a Minnesota county. MDH (N=55,206), refers to a sequence available on GISAID with county metadata provided by the Minnesota Department of Health (MDH).



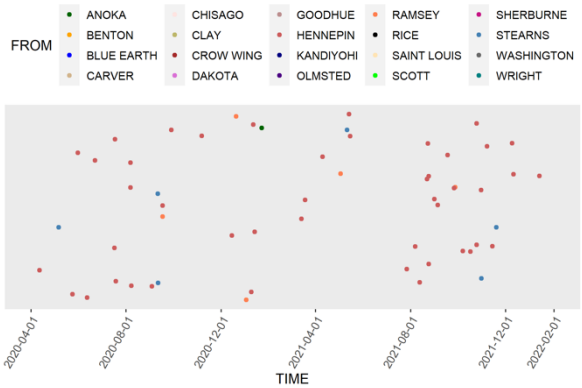
Extended Data Fig. 2. Phylogeny of 24,070 full genome SARS-CoV-2 sequences generated for this study from 2020-2022 via Nextstrain. Branch colors indicate the assigned clade.



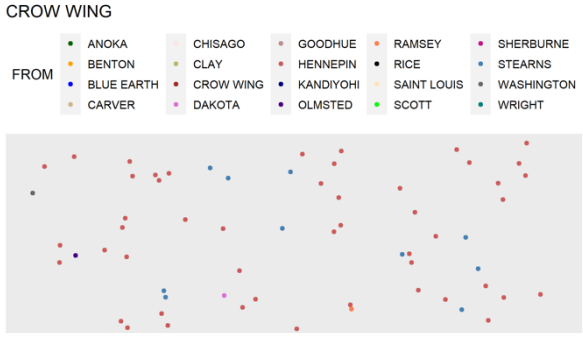
CHISAGO



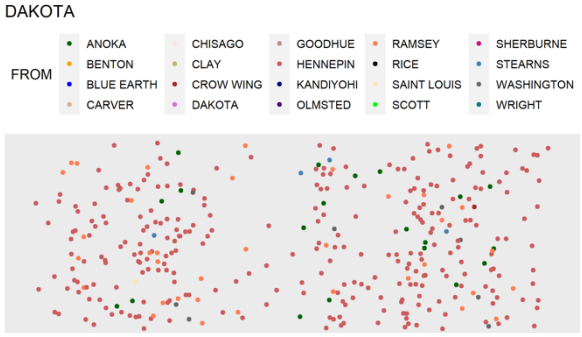
CLAY



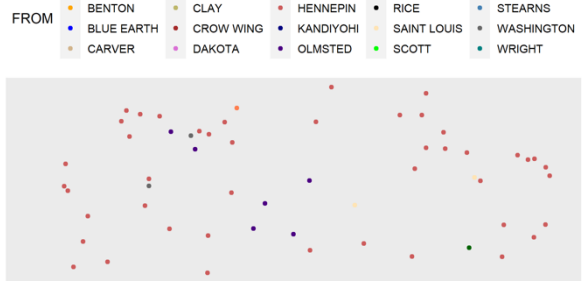
CROW WING



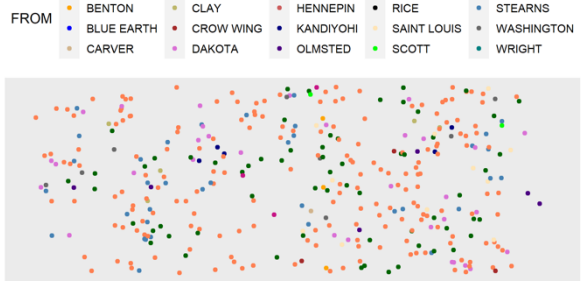
DAKOTA



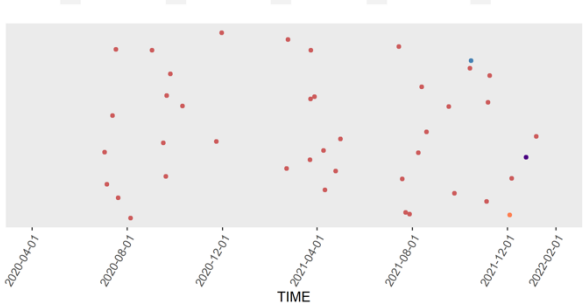
GOODHUE



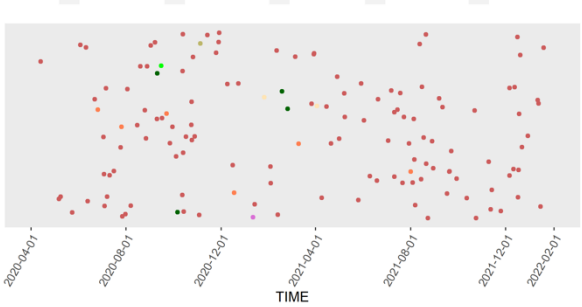
HENNEPIN

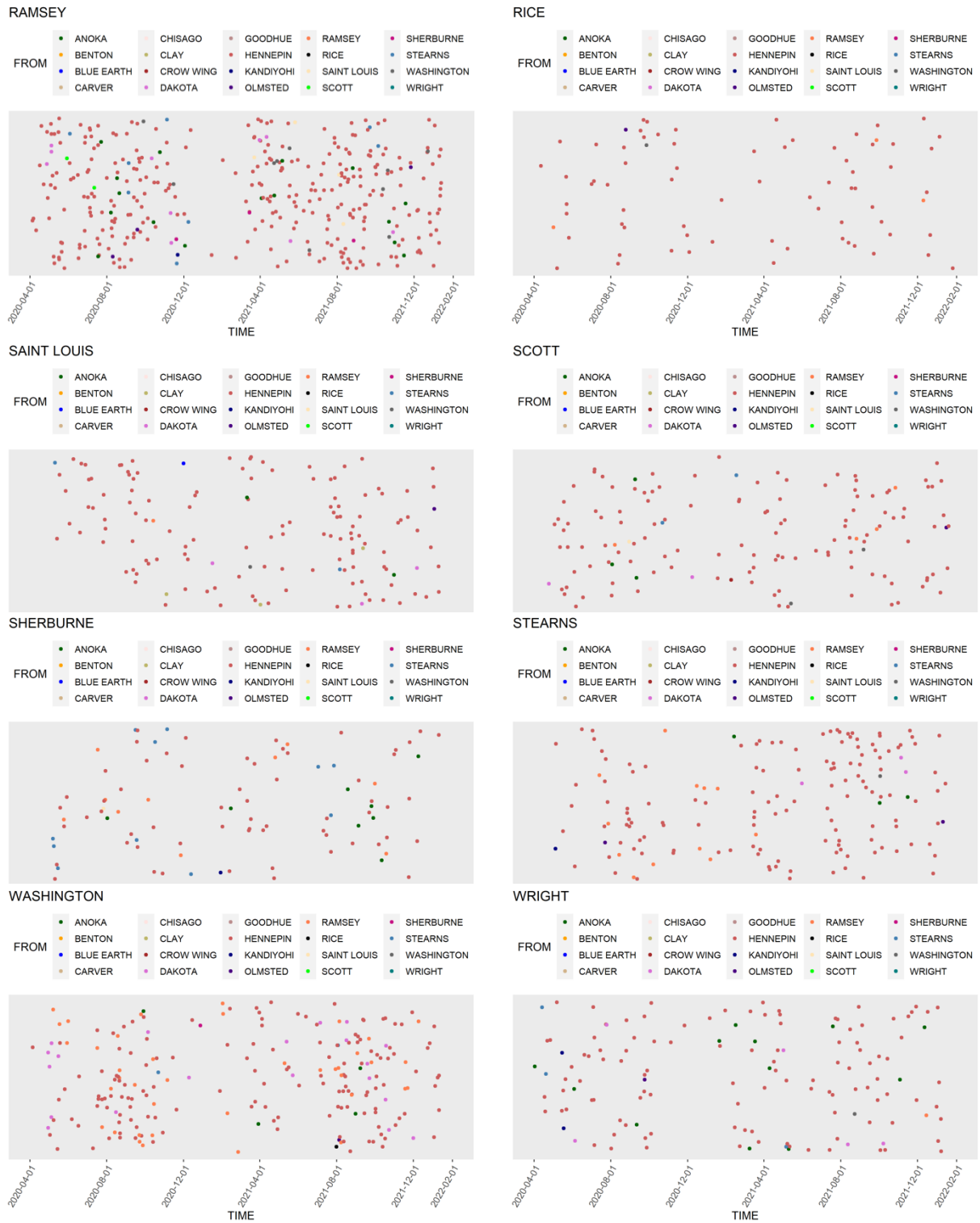


KANDIYOHI

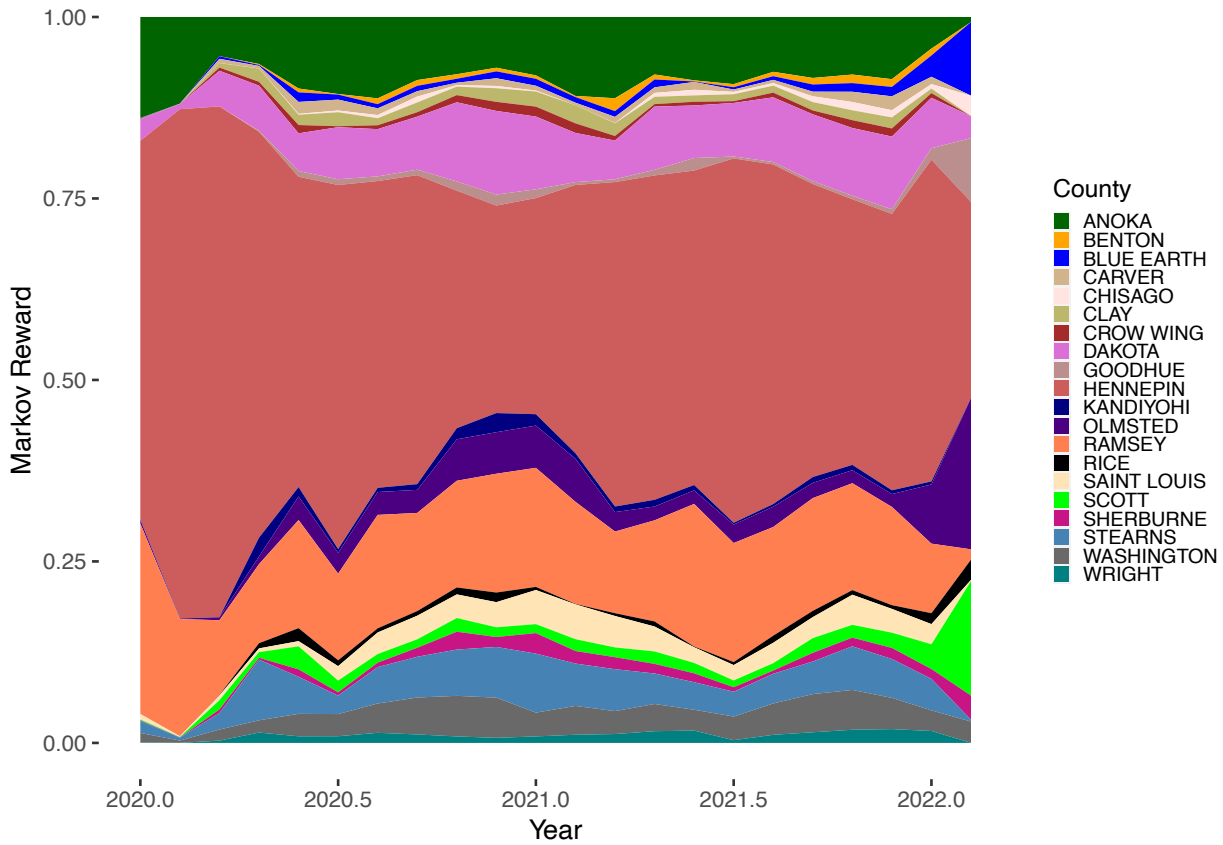


OLMSTED

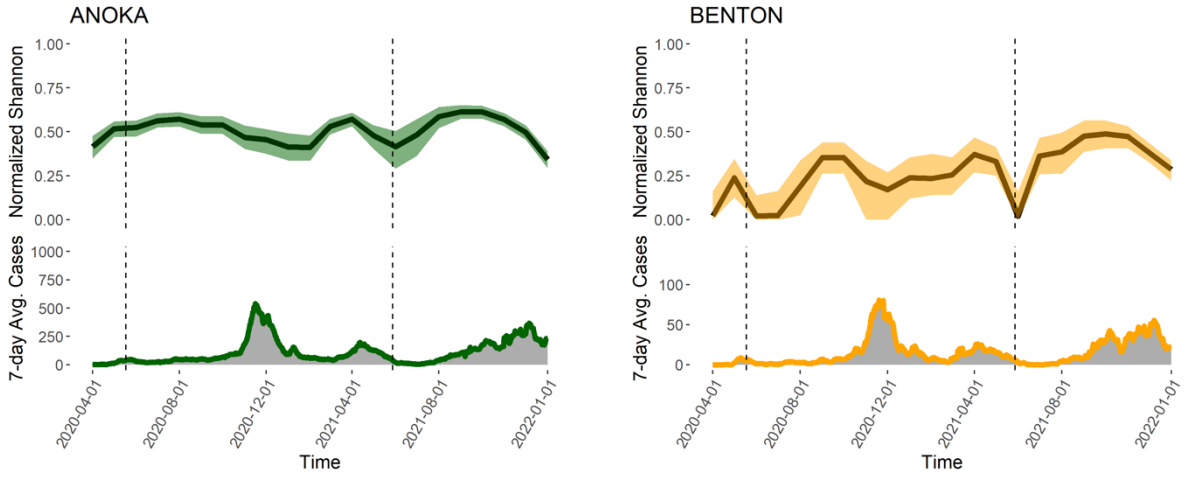


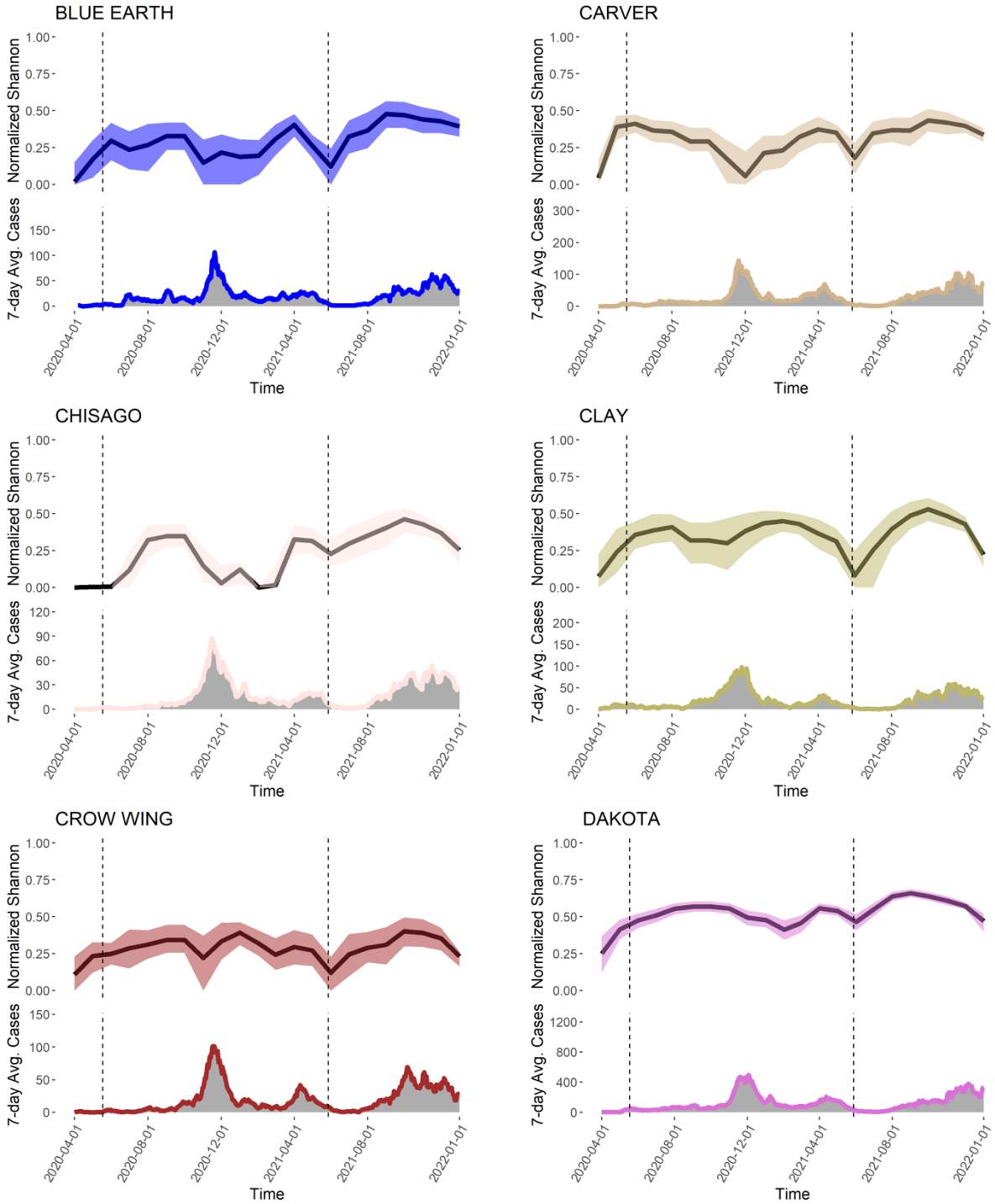


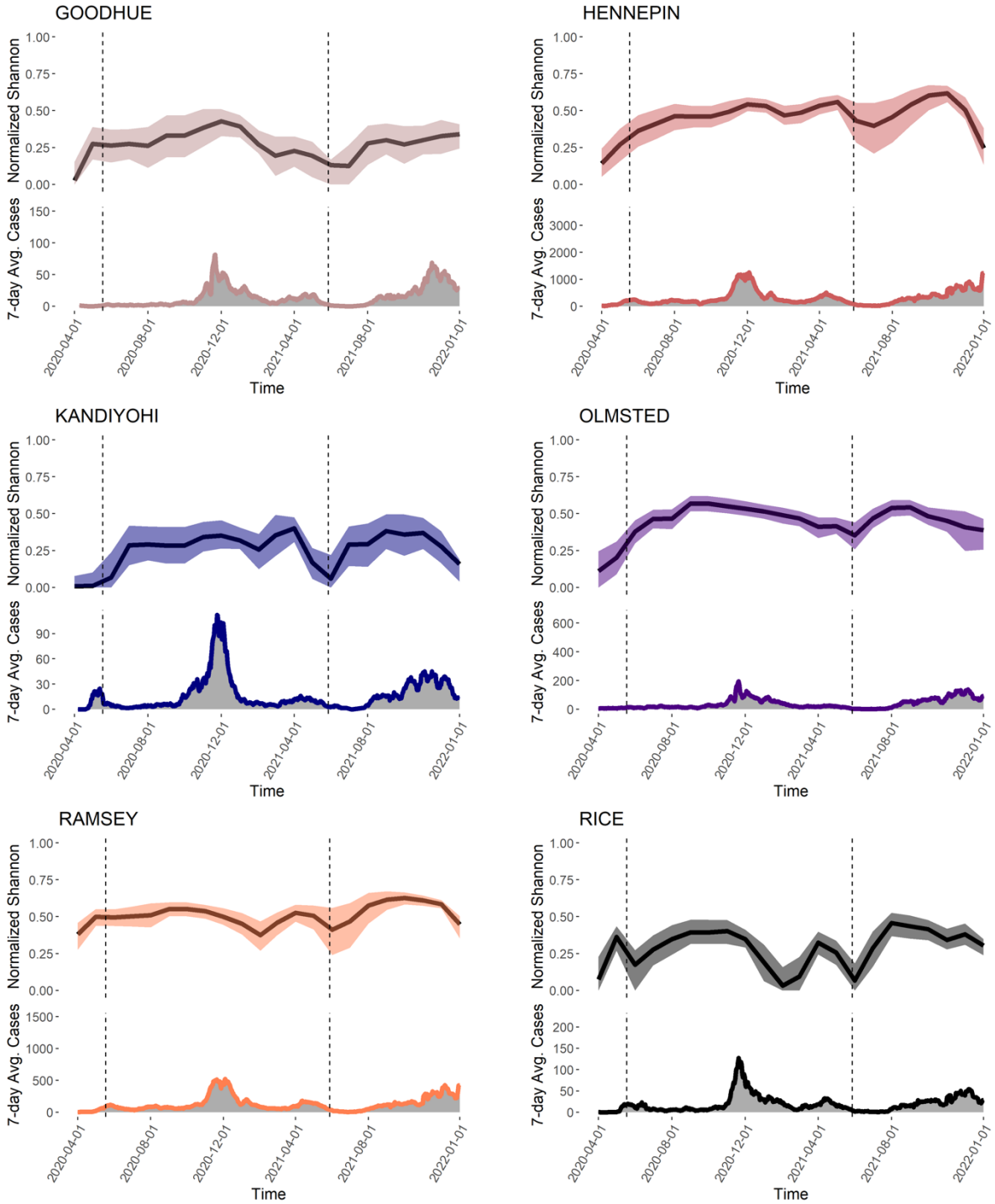
Extended Data Fig. 3. County-to-county introductions and timing extracted from the MCC tree (Fig. 1A). We limit this to the period of April 2020 – February 2022 for visualization purposes.

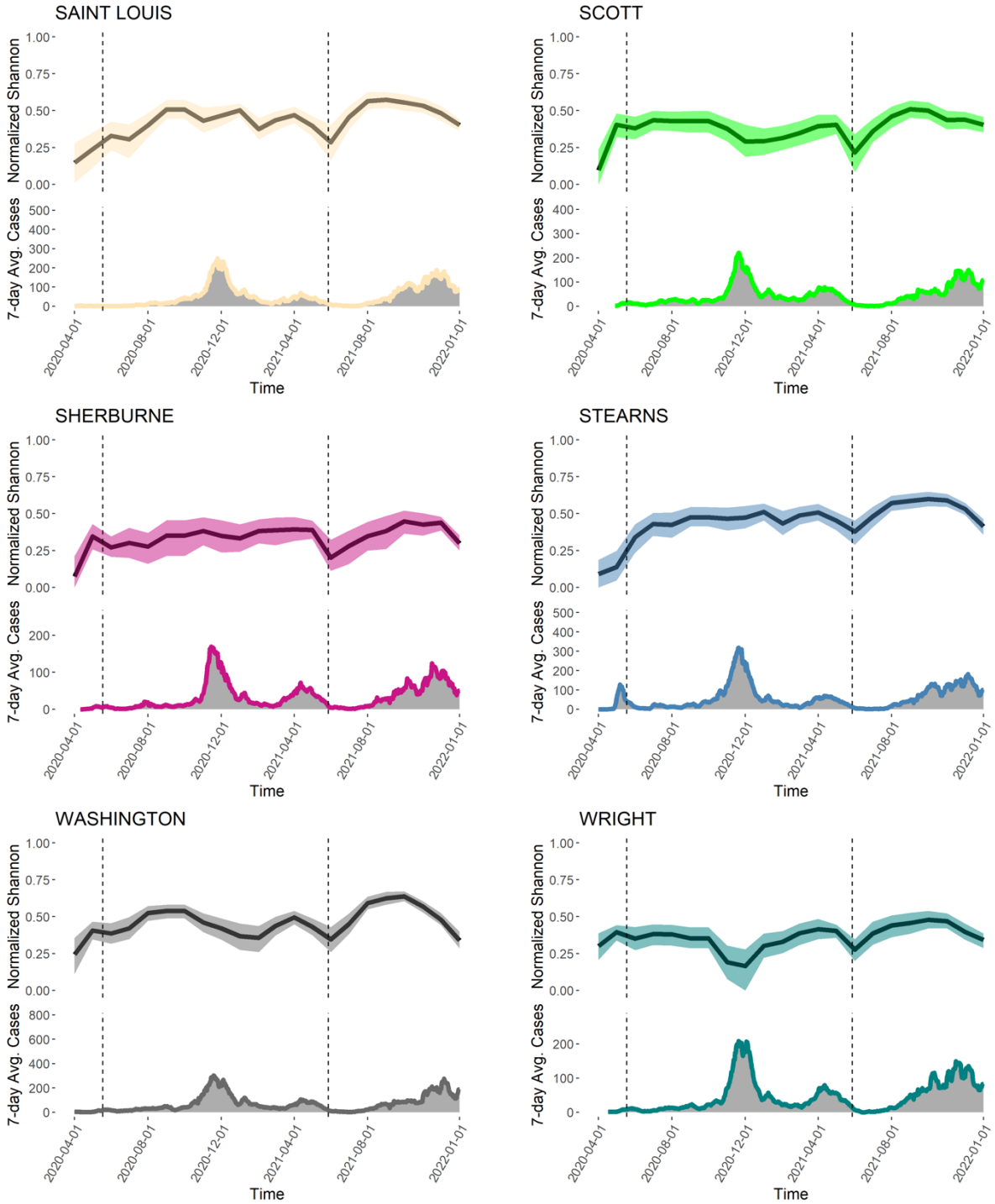


Extended Data Fig. 4. Markov rewards of SARS-CoV-2 throughout the pandemic for the 20 Minnesota counties included in the analysis.



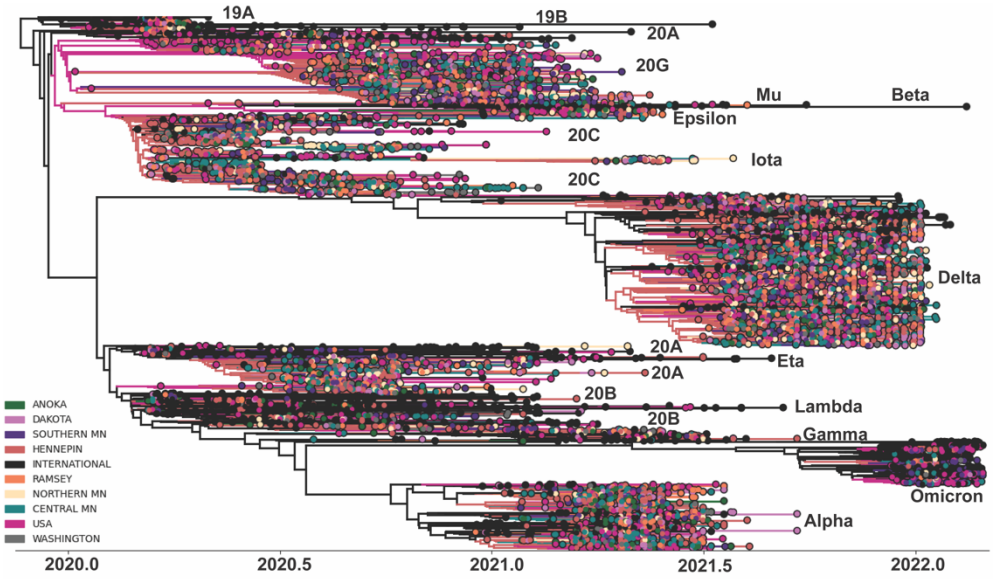

BLUE EARTH
CARVER
CHISAGO
CLAY
CROW WING
DAKOTA
Normalized Shannon
7-day Avg. Cases
Time
2020-04-01
2020-08-01
2020-12-01
2021-04-01
2021-08-01
2022-01-01


GOODHUE
HENNEPIN
KANDIYOHI
OLMSTED
RAMSEY
RICE
Normalized Shannon
7-day Avg. Cases
Time
2020-04-01
2020-08-01
2020-12-01
2021-04-01
2021-08-01
2022-01-01

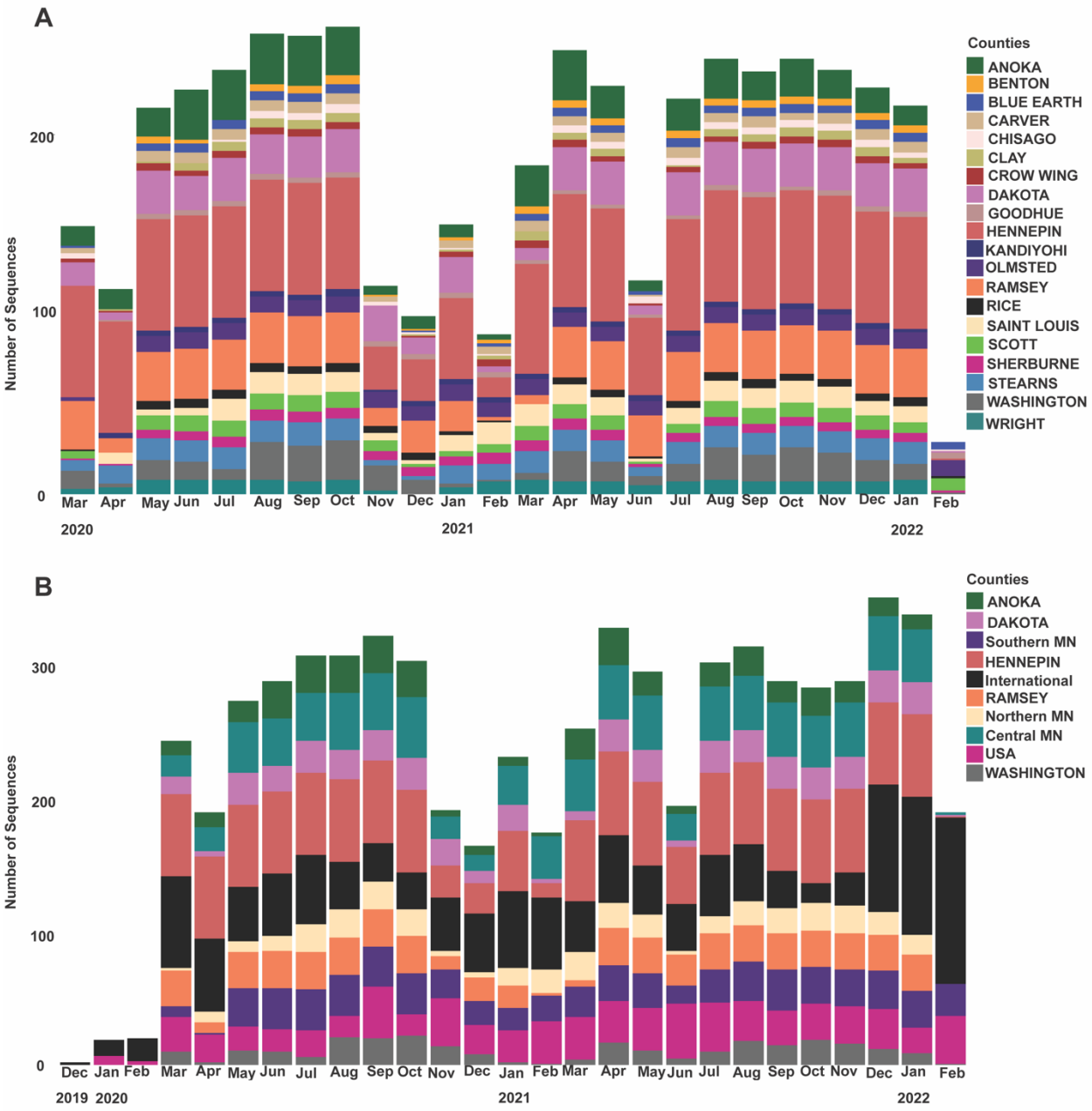


Extended Data Fig. 5. Virus diversity and cases per county. On the top of each panel, we show the normalized Shannon diversity index over time for each of the twenty Minnesota Counties. The shaded areas represent the 95% Bayesian highest posterior density (HPD). Below, we show the seven-day average cases over time of the 20 Minnesota Counties analyzed in this study and provided the CDC [48]. The first vertical line indicates the end of

lockdown in Minnesota on May 18, 2020 [49]. The second vertical line indicates the end of all COVID-19 restrictions in the State on May 28, 2021 [16].



Extended Data Fig. 6. Maximum clade credibility tree of our international dataset (N = 6,479).



Extended Data Fig 7. Sequence distribution by county and month for: A) Minnesota-only dataset and B) International dataset.

Extended Data Table 1. Demographics of N = 22,514 Mayo Clinic Laboratories COVID-19 patients included in the study.

Demographic	Category	Count (%)
Gender	Male	11,423 (49)
	Female	11,076 (51)
	Non-binary	15 (<1)
Age	< 18	3,416 (15)
	18-45	11,204 (50)
	46-64	5,350 (24)
	65+	2,544 (11)
State		
	Minnesota	21,669 (96)
	Wisconsin	413 (2)
	Iowa	184 (1)
	Other States	148 (1)

Extended Data Table 2. Number of sequences by county for the county-to-county phylodynamic analysis. We show the cumulative number of confirmed COVID-19 cases on February 28, 2022, used to guide the number of sequences (5 per 1,000 cases) for the analysis. We also include population estimates per county based on US Census data [50].

Counties	Number of Sequences (N = 4,586)	Confirmed COVID-19 Cases on February 28, 2022	2020 Population
Anoka	392	78,318	363,887
Benton	65	12,913	41,379
Blue Earth	85	16,900	69,112
Carver	106	21,100	106,922
Chisago	62	12,348	56,621
Clay	70	14,028	65,318
Crow Wing	71	14,130	66,123
Dakota	434	86,741	439,882
Goodhue	61	12,235	47,582
Hennepin	1266	253,289	1,281,565
Kandiyohi	64	12,817	43,732
Olmsted	195	38,908	162,847
Ramsey	521	104,283	552,352
Rice	82	16,310	67,097
Saint Louis	194	38,764	200,231
Scott	158	31,699	150,928
Sherburne	113	22,531	97,183
Stearns	238	47,520	158,292
Washington	264	52,881	267,568
Wright	145	29,096	141,337

Extended Data Table 3. Number of sequences by location for the international phylodynamic analysis.

Location	Number of Sequences (N = 6,479)
Anoka	392
Dakota	434
Hennepin	1266
Ramsey	521
Washington	264
Central Minnesota*	793
Northern Minnesota$	335
Southern Minnesota#	581
International	1227
USA	666

*Central Minnesota includes seven counties: Benton, Carver, Chisago, Kandiyohi, Sherburne, Stearns, and Wright.

$Northern Minnesota includes three counties: Clay, Crow Wing, and Saint Louis. #Southern Minnesota includes five counties: Blue Earth, Goodhue, Olmsted, Rice, and Scott

Extended Data Table 4. Adaptive COVID-19 Treatment Trial (ACTT) Severity Score[19]. We assigned each numerical score to one of four disease outcome variables.

Score	Definition	Categorical Variable
1	No limitation of activities / no hospitalization	Mild
2	Limitation of activities / no hospitalization	Mild
3	Hospitalization without oxygen therapy	Moderate
4	Hospitalization with oxygen therapy	Moderate
5	Hospitalization with high flow oxygen or non-invasive ventilation	Moderate
6	Hospitalization with mechanical and intubation ventilation	Severe
7	Hospitalization with ventilation and additional organ support	Severe
8	Death	Death