AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere

Check for updates

| Epidemiology | Research Article

Genomic epidemiology reveals the dominance of Hennepin County in the transmission of SARS-CoV-2 in Minnesota from 2020 to 2022

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]

AUTHOR AFFILIATIONS See affiliation list on p. 9.

ABSTRACT Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has had an unprecedented impact on human health and highlights the need for genomic epidemiology studies to increase our understanding of virus evolution and spread and to inform policy decisions. We sequenced viral genomes from over 22,000 patient samples tested at Mayo Clinic Laboratories between 2020 and 2022 and used Bayesian phylodynamics to describe county and regional spread in Minnesota. The earliest calculated introduction into Minnesota was to Hennepin County from a domestic source around 22 January 2020; 6 weeks before the first confirmed case in the state. This led to the virus spreading to Northern Minnesota and, eventually, the rest of the state. International introductions were most abundant in Hennepin (home to the Minneapolis/St. Paul International Airport) totaling 45 (out of 107) over the 2-year period. Southern Minnesota counties were most common for domestic introductions, with 19 (out of 64), potentially driven by bordering states such as Iowa and Wisconsin as well as Illinois, which is nearby. Hennepin also was, by far, the most dominant source of in-state transmissions to other Minnesota locations (n = 772) over the 2-year period. We also analyzed the diversity of the location source of SARS-CoV-2 viruses in each county and noted the timing of state-wide policies as well as trends in clinical cases. Neither the number of clinical cases nor the major policy decisions, such as the end of the lockdown period in 2020 or the end of all restrictions in 2021, appeared to have an impact on virus diversity across each individual county.

IMPORTANCE We analyzed over 22,000 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes of patient samples tested at Mayo Clinic Laboratories during a 2-year period in the COVID-19 pandemic, which included Alpha, Delta, and Omicron variants of concern to examine the roles and relationships of Minnesota virus transmission. We found that Hennepin County, the most populous county, drove the transmission of SARS-CoV-2 viruses in the state after including the formation of earlier clades including 20A, 20C, and 20G, as well as variants of concern Alpha and Delta. We also found that Hennepin County was the source for most of the county-to-county introductions after an initial predicted introduction with the virus in early 2020 from an international source, while other counties acted as transmission "sinks." In addition, major policies, such as the end of the lockdown period in 2020 or the end of all restrictions in 2021, did not appear to have an impact on virus diversity across individual counties.

KEYWORDS epidemiology, computational biology, SARS-CoV-2, Minnesota, high-throughput nucleotide sequencing

Editor Nicole M. Bouvier, Icahn School of Medicine at Mount Sinai, New York, USA

Address correspondence to Andrew P. Norgan, Norgan.Andrew@mayo.edu.

The authors declare no conflict of interest.

See the funding table on p. 9.

Received 27 April 2023
Accepted 20 September 2023
Published 26 October 2023

Genomic epidemiology has provided valuable insight into the 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, due to unprecedented viral genomic sequencing efforts across the globe. As of 29 July 2023, there are over 15.8 M virus sequences in GISAID (1) and over 8.1 M in NCBI Virus (2) and GenBank (3). Studies that focus on localized spread such as counties or regions 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 the 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 the 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), 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, and 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. (9) created county-level risk maps for international importation of early Alpha variants via air transport. More recently, Smith et al. (10) examined the transmission of the Omicron variant among four counties in Arizona. Other works such as Valesano et al. (11), Holland et al. (12), Currie et al. (13), and Srinivasa et al. (14) 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. Here, we leveraged amplicon-based high-throughput sequencing and Bayesian phylodynamics to analyze the evolution and spread of SARS-CoV-2 into, and within, the State of Minnesota to understand the roles of specific counties and regions in the transmission of the virus over a 2-year period and across different viral clades, variants of concern (VoC), and state-wide mandates and policies.

MATERIALS AND 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, USA). 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 × 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 (15). We used Illumina's Dynamic Read Analysis for GENomics (DRAGEN) COVID Lineage software and pipeline (16) (versions 3.5.1, 3.5.3, and 3.5.6) for reference-based alignment to Wuhan-1 (17), quality assessment, variant calling, and generation of consensus sequences. We excluded sequences from downstream analysis if they met any of the following criteria, including: (i) given an overall score of fail by the DRAGEN pipeline due to having an insufficient amount of detectable viral reads; (ii) given an overall quality score by Nextclade (18) as bad; (iii) potentially contaminated based on the presence of unusual allele frequencies (<0.9); (iv) duplicate runs; and (v) positive or negative controls.

Phylogenetic analysis of SARS-CoV-2

We assembled a representative data set (n = 6,188; Fig. S1) that included SARS-CoV-2 genome sequences from the 20 counties with the greatest number of reported COVID-19 cases as of 28 February 2022 as well as a global representation of sequences available via GenBank as part of an open access data set from Nextstrain (Table S1) (19). We used the list of accessions to download sequences from NCBI Virus (2). All but two of the genomes were ≥29,000 nucleotides in length. We also removed duplicates.

We included sequences from December 2019 (including Wuhan-1) to 28 February 2022, as well as their sampling location and collection date metadata. To partially address sampling bias, we sampled at a rate of five sequences per 1,000 county cases and used the filter module in augur (20) to distribute (as equally as possible) our heterochronous sequences by month (Fig. S1). For each county, we attempted to include SARS-CoV-2 genomes across the 2-year timeframe by supplementing our data set with sequences provided by the Minnesota Department of Health (MDH) (and available in GISAID). The MDH sequences were produced from randomly selected samples from clinics and community testing sites. Sample Ct values were equal to or below 30.

We aligned all sequences using MAFFT (21) and used trimAl v1.4.rev22 (22) to remove columns that contained more than 70% gaps. We created an initial phylogenetic tree via Nextstrain's augur tree command (20) with IQTree and rooted the tree based on Wuhan-1 (17). We used TempEST (23) to examine the temporal signal of our heterochronous samples, which suggested the use of a strict molecular clock (correlation coefficient = 0.958) for our phylogenetic analysis. We also removed one sequence as an outlier. We used augur refine and the keep-root option to modify our tree with sequence metadata. We removed additional sequences that had potentially misassigned clades or produced inconsistencies with the phylogenetic structure as shown in Nextstrain's global all-time subsampled data set (24).

Phylodynamics of SARS-CoV-2 in Minnesota

We used R package ape (30) to confirm that our starting tree was rooted and non-bifurcating in order to comply with our downstream inferencing framework. 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 data sets (25, 26). We used our starting tree, a non-parametric Bayesian SkyGrid coalescent model for our tree prior (27), and a strict molecular clock. We ran two Markov-chain Monte Carlo (MCMC) simulations each for 5×10^8 steps and sampling every 5×10^4 steps. We combined these two runs via LogCombiner v.10.4 (28) after removing 10% burn-in. We checked for convergence of model parameters via Tracer v1.7.1 (29) with an ideal effective sample size threshold of 200. We generated log marginal likelihoods and evaluated population growth priors via a stepping stone and path sampling procedure (30). Our results suggested the use of the non-parametric Skygrid tree prior over a constant growth model (Table S2).

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 trait. We specified all non-U.S. sequences as "international" and non-Minnesota U.S. states as "USA." For computational efficiency, we kept the five counties with the greatest number of cases as independent locations and grouped the remaining 15 counties into three discrete regions including southern, central, and northern Minnesota (Table S1). In BEAUti (28), we specified an asymmetric transmission rate matrix of $K(K^*1)$, where K is equivalent to the number of discrete locations (n = 10 for our data set). We recorded Markov jumps (31) between locations to estimate the timing and source of introductions and specified an MCMC of 5×10^6 sampling every 5×10^2 steps. We used TreeAnnotator v.10.4 (28) to create a single maximum clade credibility (MCC) tree after a 10% burn-in. We used Baltic (32) for tree visualization and to extract the timing of discrete location transitions along the branches of the MCC for our estimates of introductions. For the latter, we estimated the time-point of the middle of the branch between the current node and its parent for any change in location state. We excluded transmission chains with low support of discrete origin and destination names by including only nodes with a posterior probability of ≥0.90 for the location state. We used SpreadD3 (33) to calculate the Bayes factors to identify the most parsimonious origin-destination scenarios (Table S3; Fig. S2). We used two programs of the BEAST library (28), introduced in reference (34), as part of our Bayesian phylodynamic analyses. TreeMarkovJumpHistoryAnalyzer samples from the posterior distribution of trees to collect the timing and location of each Markov jump (34). We used the output from this program to calculate the ratio of introductions to total viral flow into and out of each county [number of introductions/(number of introductions + number of exports)] as described in Lemey et al. (34) 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 (17). We used the output from this program to calculate the normalized Shannon diversity metric (34). 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 the NormShannon method in the R package QSutils (35) to calculate normalized monthly diversity metrics for each county and HDinterval (36) for the corresponding 95% highest posterior density region.

RESULTS

We sequenced SARS-CoV-2 genomes from genomic material collected from clinical samples of patients tested for SARS-CoV-2 infection at Mayo Clinic Laboratories (Fig. 1; Fig. S3) over a 2-year period from March 2020 to March 2022. We combined these sequences with additional genomes generated for surveillance purposes by the MDH and performed Bayesian phylodynamics to understand in-state spread as well as the impact and timing of introductions into the State of Minnesota (see Materials and Methods).

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%) (Table S4). The breakdown by gender was nearly 50/50 between males and females, while 50% of the patients was between 18 and 45. Fifteen percent was under 18, while 11% was 65 or older.

Hennepin County consistently drives in-state transmission

We down-sampled our genomes from Minnesota nearly proportional to the number of COVID-19 cases per county and then added additional genomes from NCBI GenBank (3) as part of an international data set used in Nextstrain (5) (see Materials and Methods). To address the computational burden of adding sequences to our already large data set, 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 (Fig. S4; Table S1).

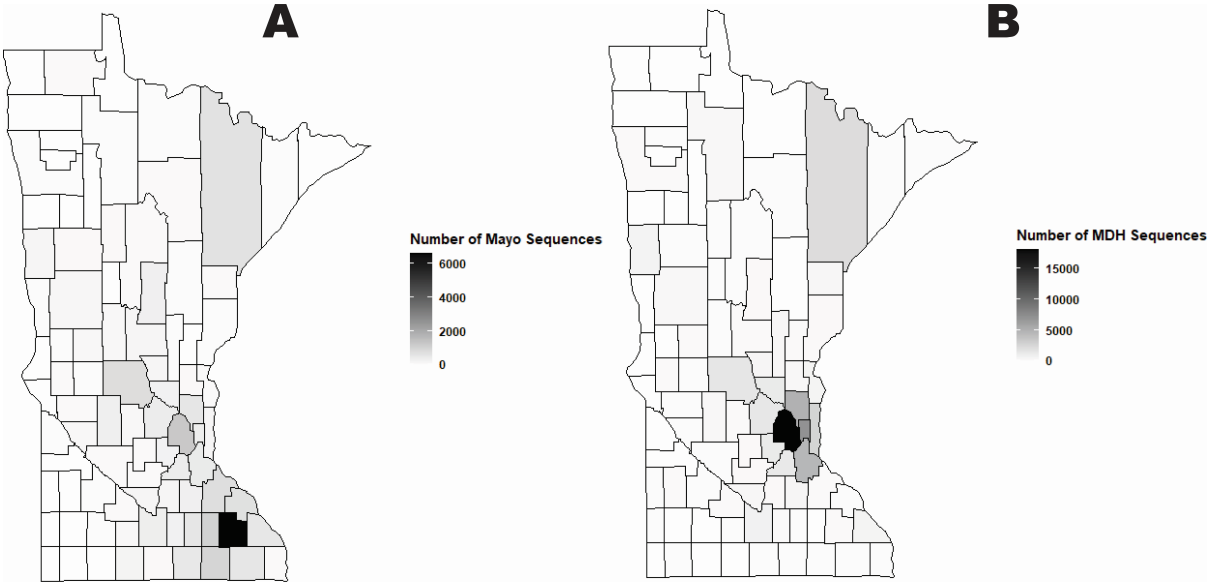

FIG 1 County map of Minnesota with number of sequences (*N* = 76,875) eligible for analysis by source. (A) New sequences (*N* = 21,669) generated from this study at Mayo Clinic Laboratories with a known sampling location in a Minnesota county. Olmsted County in Southeast Minnesota (where the Mayo Clinic Rochester campus is located) has the largest number of sequences. (B) Sequences (*N* = 55,206) available on GISAID with county metadata provided by the MDH. Hennepin County (the most populated county), north of Olmsted, has the largest number of sequences.

We implemented Bayesian phylodynamic models to examine the transmissions in Minnesota from early 2020 to early 2022 (see Materials and Methods). We recorded Markov jumps (31) to estimate the timing of introductions and their directionality. After introductions from domestic and international locations, our analysis shows that Hennepin County, the most populous county which includes Minneapolis, the most populated city, drove the transmission of SARS-CoV-2 viruses in the state (Fig. 2). This includes the formation of earlier clades including 20C and 20G, as well as variants of

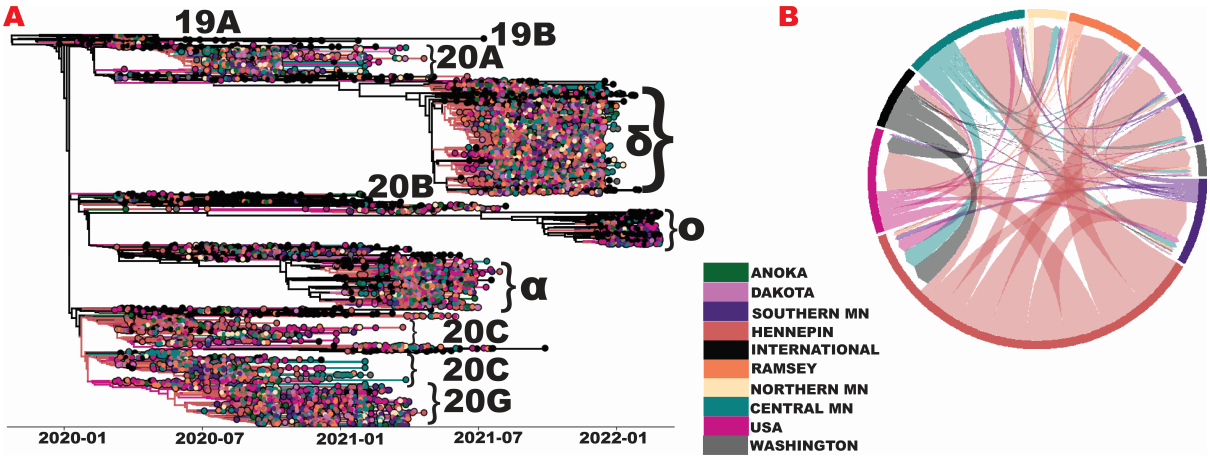

FIG 2 SARS-CoV-2 evolution and spread to and within the state of Minnesota. (A) MCC tree of 6,188 SARS-CoV-2 genomes from Minnesota counties and regions as well as international locations and other domestic locations in the United States, with manually annotated clades by Nextclade-assigned names or VoCs (18) using lowercase Greek letters. Less well represented VoCs in the tree (e.g., Gamma or Epsilon) are unlabeled. For clades (e.g., 20A) that are not monophyletic in the tree, the most populous clade is labeled. (B) Markov jumps between locations represented as a Chord diagram. Colors for both panels represent locations depicted in the legend. Central MN includes seven Minnesota counties: Benton, Carver, Chisago, Kandiyohi, Sherburne, Stearns, and Wright. Northern MN includes three counties: Clay, Crow Wing, and Saint Louis. Southern MN includes five counties: Blue Earth, Goodhue, Olmsted, Rice, and Scott. USA includes all states except for Minnesota. MN, Minnesota; *α*, alpha; *δ*, delta; and *O*, omicron. Panels A and B, legend, and text labels were recreated in Adobe Illustrator for visualization purposes.

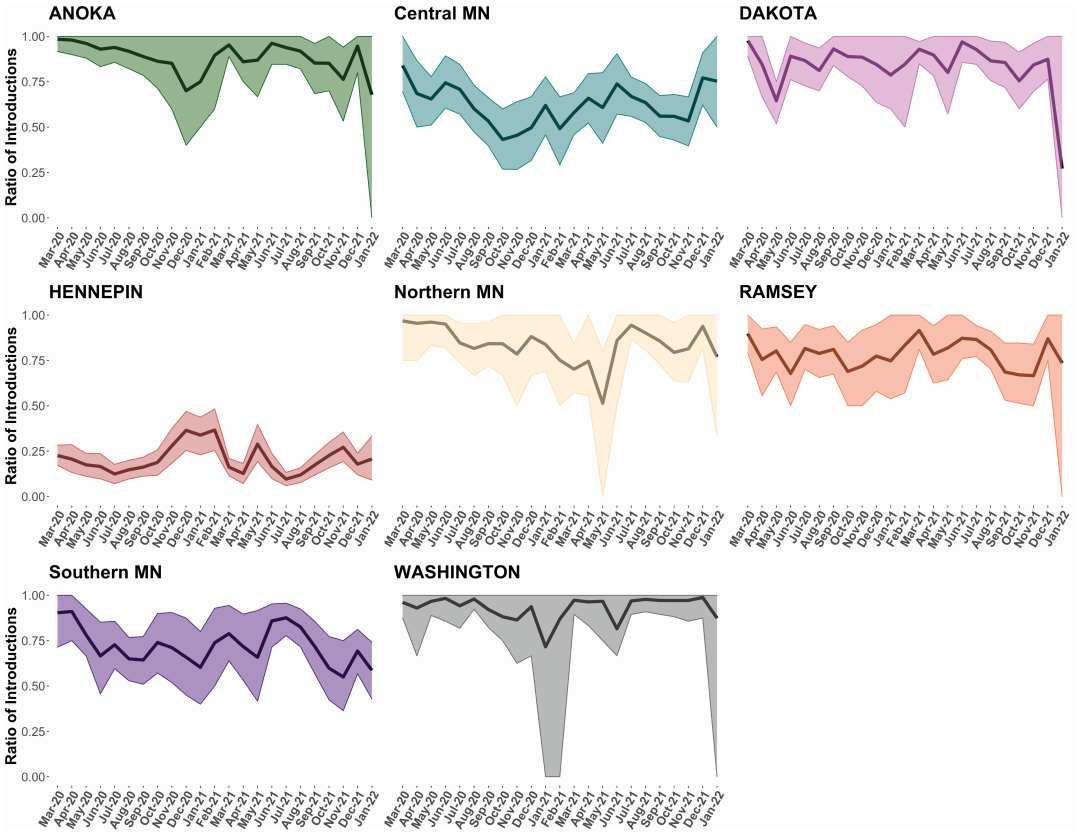

FIG 3 Ratio of introductions to total viral flow into and out of each discrete location by month from March 2020 to January 2022. The posterior mean ratio and 95% Bayesian highest posterior density interval are shown. Anoka, Dakota, Ramsey, and Washington have wider intervals during certain months, such as January 2022, due to a decrease in local sampling.

concern Alpha and Delta (Fig. 2A). The counties in central Minnesota contributed to spread including 20C, Alpha, and Delta, while southern Minnesota contributed mostly to 20G. Markov jump estimates (Fig. 2B) as shown via a Chord diagram suggest that transmission of SARS-CoV-2 within the state largely originated from Hennepin County (thick arcs and wider fragments at the outer circle). However, we also note the existence of transmission back to these areas (white space between arc points and outer fragment) from nearby counties in central Minnesota.

We measured the ratio of introductions to total viral flow into and out of each county by month from March 2020 to January 2022. A value of 1 suggests a county as solely being a "sink" (accepts SARS-CoV-2 lineages but never exports them to other counties), while a value of 0 indicates a county as solely being a "source." Anoka, Dakota, Ramsey, northern Minnesota, southern Minnesota, and Washington were fueled by introductions mostly throughout the pandemic (Fig. 3). Meanwhile, central Minnesota (outside of Hennepin and Ramsey) was dominated by introductions early in the pandemic but later in 2020 experienced brief trends of higher virus exportation. Hennepin County showed a drastically different trend than all others as it consistently acted as a source for other Minnesota counties over the nearly 2-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, potentially driven by the dominance of out-of-state introductions.

Low-to-intermediate spatial mixing within the state of Minnesota

We assessed county-specific virus diversity via a normalized Shannon diversity index (Fig. S5) that we computed based on the duration of time associated with continuous partitions of the phylogeographic tree as determined by Markov jumps (34). The index,

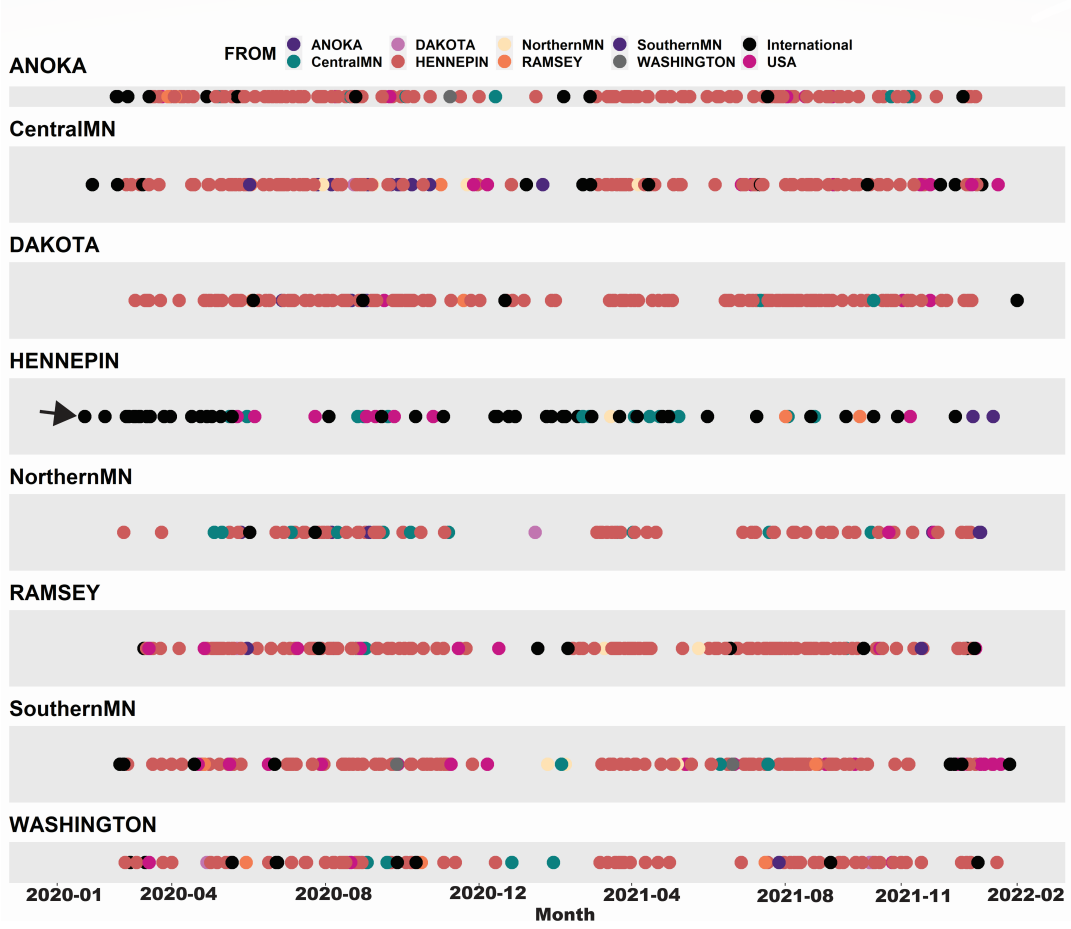

FIG 4 Timing and source of international, domestic, and within-state introductions for each discrete location. Colors correspond to the source location. An arrow shows the first predicted introduction into Minnesota, which is estimated to have occurred in Hennepin County at around 22 January 2020 (from an international location). We used Baltic to extract introductions (migration events) along the annotated branches of the phylogeographic tree for node states with a posterior probability of ≥0.90. Location panels were combined into one figure in R and original month labels (*x*-axis) were generated in Adobe Illustrator for visualization purposes.

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 an exclusive spatial structure such as an outbreak contained to only one county (34). Conversely, a value of 1, suggests significant spatial mixing of SARS-CoV-2 between counties (34). The counties and regions show low-to-intermediate (0.25–0.5 Shannon) spatial mixing with brief periods of waxing and waning. The two dotted vertical lines indicate changes in state-wide policy. The first vertical line indicates the end of lockdown in Minnesota on 18 May 2020 (37). The second line on 28 May 2021 indicates the end of all COVID-19 restrictions in the state (38). Neither of these policy decisions appeared to have a significant impact on virus diversity across each individual county. Anecdotally (looking at the trends of each graph), the changes in case counts over time do not appear to have a relationship with county-specific diversity.

Hennepin County received the vast majority of out-of-state introductions and was the dominant source for in-state transmission

We focused on the timing and source of introductions into the state during the pandemic (Fig. 4) as estimated from our maximum clade credibility tree (Fig. 2A). The earliest estimated introduction into Minnesota was to Hennepin County from an international source on around 22 January 2020 (depicted with an arrow in Fig. 4).

This is about 1 month before the first patient in the state, a man from Ramsey County (which borders Hennepin), developed symptoms and around 6 weeks before (6 March 2020) the Department of Health confirmed the infection (39). The first county-to-county introductions were estimated to originate from Hennepin to somewhere in northern Minnesota around 22 February and from Hennepin to Washington County (also in the northern part of the state) around 24 February. International introductions were most abundant in Hennepin (home to the Minneapolis/St. Paul International airport) totaling 45 (out of 107) over the 2-year period. Southern Minnesota counties were most common for domestic introductions, with 19 (out of 64), potentially driven by bordering states such as Iowa and Wisconsin as well as Illinois, which is nearby. Hennepin also was, by far, the most dominant source of in-state transmissions to other Minnesota locations (n = 772) over the 2-year period.

DISCUSSION

We analyzed over 22,000 new genomes of patients tested at Mayo Clinic Laboratories during a 2-year period in the COVID-19 pandemic. We focused our analysis on in-state transmission of SARS-CoV-2, mostly at the county (second administrative boundary) level, to describe the spread into and within Minnesota. Despite numerous efforts in genomic epidemiology, few studies have focused on county-to-county transmission in the United States over most of the pandemic (including different VoCs). We expand on earlier efforts such Moreno et al. (4) and Deng et al. (7) but include multiple variants and an extensive timeframe. We found that spread in the state was dominated by viruses from Hennepin County, which contains the largest metropolis, and that other regions including Northern and Southern Minnesota acted mainly as "sinks" for in-state transmission.

The earliest estimated introduction into Minnesota was to Hennepin from an international source about 6 weeks before the first confirmed case in the state. This suggests that earlier (and likely milder) infections of SARS-CoV-2 occurred before the first documented case. Interestingly, while Hennepin drove in-state transmission, it did not result in variations of location-specific spatial diversity. We found that all counties and regions had low-to-intermediate (0.25–0.5 Shannon) spatial mixing with brief periods of waxing and waning. The fluctuation in spatial diversity over time (that did exist) did not appear to be impacted by key state-mandated policies nor did it appear to have any relationship with reported clinical cases (Fig. S5).

As the virus continues to evolve, more within-state genomic epidemiology studies are needed to inform local and state public health response by highlighting the roles of various counties in state-wide transmission. In addition, they can elucidate the impact of out-of-state introductions on local spread which can inform policies such as travel.

We note several limitations in the study including the likelihood of location-specific sampling bias. We attempted to supplement the known locations of patients in our study (biased towards southeastern Minnesota) with existing sequences provided via the Minnesota Department of Health. We initially 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 (as well as outliers of sequences from included counties). Thus, we are unable to account for virus spread from less-populated areas of the state. We also attempted to include a representative sample of USA and international sequences. However, it is possible that additional sequences (context) might change the distribution of virus clades and the timing of introductions into the state, which could alter our interpretations of SARS-CoV-2 spread. We also only included early Omicron sequences and thus we are unable to describe an informed picture of its evolutionary diffusion in the state. In addition, our use of different versions of the DRAGEN pipeline over the course of our 2-year study period likely led to differences in variant frequencies across virus lineages/VoCs.

ACKNOWLEDGMENTS

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, which were used in this work.

This article is dedicated to Dr. Peter T. Vedell, who sadly passed away during the writing of this manuscript.

This publication was supported by funding from the Center for Individualized Medicine-Mayo Clinic Research. The 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 M.S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR AFFILIATIONS

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

PRESENT ADDRESS

Andrew P. Norgan, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA

AUTHOR ORCIDs

Matthew Scotch http://orcid.org/0000-0001-5100-9724
Bobbi S. Pritt http://orcid.org/0000-0003-0261-1326
Andrew P. Norgan http://orcid.org/0000-0002-2955-2066

FUNDING

Funder	Grant(s)	Author(s)
HHS | NIH | National Institute of Allergy and Infectious Diseases (NIAID)	R01AI164481	Matthew Scotch
Mayo Clinic | Center for Individualized Medicine		Bobbi S. Pritt Andrew P. Norgan

AUTHOR CONTRIBUTIONS

Matthew Scotch, Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing | Kimberly Lauer, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing | Eric D. Wieben, Data

curation, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review and editing | Yesesri Cherukuri, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – review and editing | Julie M. Cunningham, Data curation, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review and editing | Eric W. Klee, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – review and editing | Jonathan J. Harrington, Project administration, Writing – review and editing | Julie S. Lau, Data curation, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review and editing | Samantha J. McDonough, Data curation, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review and editing | Mark Mutawe, Data curation, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review and editing | John C. O'Horo, Formal analysis, Methodology, Writing – review and editing | Chad E. Rentmeester, Formal analysis, Methodology, Writing – review and editing | Nicole R. Schlicher, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – review and editing | Valerie T. White, Data curation, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review and editing | Susan K. Schneider, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – review and editing | Peter T. Vedell, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – review and editing | Xiong Wang, Formal analysis, Resources, Validation, Writing – review and editing | Joseph D. Yao, Supervision, Validation, Writing – review and editing | Bobbi S. Pritt, Funding acquisition, Resources, Supervision, Writing – review and editing | Andrew P. Norgan, Funding acquisition, Resources, Supervision, Writing – review and editing, Conceptualization, Data curation

DATA AVAILABILITY

We have deposited the SARS-CoV-2 genomes and metadata from this study in GISAID with a list available at 10.55876/gis8.220720me and in GenBank with a list available at 10.6084/m9.figshare.23802735. We randomly shifted the collection day by ±31 days for privacy protection. The Minnesota Department of Health sequences used in this study are available on GISAID with acknowledgments at 10.55876/gis8.220709mv. Our GenBank international sequences were identified via the Nextstrain (52) site and obtained from NCBI Virus (2). We have deposited our BEAST XML files, empirical set of posterior trees, and our introductions to figshare at 10.6084/m9.figshare.22679449, 10.6084/m9.figshare.21777995, 10.6084/m9.figshare.21777998, and 10.6084/m9.figshare.21778004.

ETHICS APPROVAL

This research was conducted under the approval of ethics by the Mayo Clinic Institutional Review Board and assigned a study ID IRB#: 20-005896 under the application title Large Scale Whole Genome Sequencing of SARS-CoV-2.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Figure Captions (mSphere00232-23-s0001.docx). Captions for supplemental figures.
Fig. S1 (mSphere00232-23-s0002.tif). Sequence distribution (n = 6,188) by Minnesota county/region by month for our phylodynamic analysis.
Fig. S2 (mSphere00232-23-s0003.tif). Supported routes of pairwise SARs-CoV-2 transmission as determined by the Bayes factor (BF) statistic.
Fig. S3 (mSphere00232-23-s0004.tif). Phylogeny of 24,070 full genome SARS-CoV-2 sequences generated for this study from 2020-2022 via Nextstrain (augur v15.0.2).

Fig. S4 (mSphere00232-23-s0005.tif). Map of Minnesota counties included in the phylodynamic analysis.
Fig. S5 (mSphere00232-23-s0006.tif). Virus diversity and cases per county/location.
Supplemental Tables (mSphere00232-23-s0007.docx). Tables S1 to S4.

REFERENCES

1. Shu Y, McCauley J. 2017. GISAID: global initiative on sharing all influenza data - from vision to reality. Euro Surveill 22:13. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494
2. Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, Schäffer AA, Brister JR. 2017. Virus variation resource - improved response to emergent viral outbreaks. Nucleic Acids Res 45:D482–D490. https://doi.org/10.1093/nar/gkw1065
3. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Sherry ST, Yankie L, Karsch-Mizrachi I. 2023. Genbank 2023 update. Nucleic Acids Res 51:D141–D144. https://doi.org/10.1093/nar/gkac1012
4. Moreno GK, Braun KM, Riemersma KK, Martin MA, Halfmann PJ, Crooks CM, Prall T, Baker D, Baczenas JJ, Heffron AS, Ramuta M, Khubbar M, Weiler AM, Accola MA, Rehrauer WM, O'Connor SL, Safdar N, Pepperell CS, Dasu T, Bhattacharyya S, Kawaoka Y, Koelle K, O'Connor DH, Friedrich TC. 2020. Revealing fine-scale spatiotemporal differences in SARS-CoV-2 introduction and spread. Nat Commun 11:5558. https://doi.org/10.1038/s41467-020-19346-z
5. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123. https://doi.org/10.1093/bioinformatics/bty407
6. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, Wu C-H, Xie D, Zhang C, Stadler T, Drummond AJ. 2019. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol 15:e1006650. https://doi.org/10.1371/journal.pcbi.1006650
7. Deng X, Gu W, Federman S, du Plessis L, Pybus OG, Faria NR, Wang C, Yu G, Bushnell B, Pan C-Y, Guevara H, Sotomayor-Gonzalez A, Zorn K, Gopez A, Servellita V, Hsu E, Miller S, Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Chu HY, Shendure J, Jerome KR, Anderson C, Gangavarapu K, Zeller M, Spencer E, Andersen KG, MacCannell D, Paden CR, Li Y, Zhang J, Tong S, Armstrong G, Morrow S, Willis M, Matyas BT, Mase S, Kasirye O, Park M, Masinde G, Chan C, Yu AT, Chai SJ, Villarino E, Bonin B, Wadford DA, Chiu CY. 2020. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into northern California. Science 369:582–587. https://doi.org/10.1126/science.abb9263
8. Müller NF, Wagner C, Frazar CD, Roychoudhury P, Lee J, Moncla LH, Pelle B, Richardson M, Ryke E, Xie H, Shrestha L, Addetia A, Rachleff VM, Lieberman NAP, Huang M-L, Gautom R, Melly G, Hiatt B, Dykema P, Adler A, Brandstetter E, Han PD, Fay K, Ilcisin M, Lacombe K, Sibley TR, Truong M, Wolf CR, Boeckh M, Englund JA, Famulare M, Lutz BR, Rieder MJ, Thompson M, Duchin JS, Starita LM, Chu HY, Shendure J, Jerome KR, Lindquist S, Greninger AL, Nickerson DA, Bedford T. 2021. Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington state. Sci Transl Med 13:595. https://doi.org/10.1126/scitranslmed.abf0202
9. Alpert T, Brito AF, Lasek-Nesselquist E, Rothman J, Valesano AL, MacKay MJ, Petrone ME, Breban MI, Watkins AE, Vogels CBF, Kalinich CC, Dellicour S, Russell A, Kelly JP, Shudt M, Plitnick J, Schneider E, Fitzsimmons WJ, Khullar G, Metti J, Dudley JT, Nash M, Beaubier N, Wang J, Liu C, Hui P, Muyombwe A, Downing R, Razeq J, Bart SM, Grills A, Morrison SM, Murphy S, Neal C, Laszlo E, Rennert H, Cushing M, Westblade L, Velu P, Craney A, Cong L, Peaper DR, Landry ML, Cook PW, Fauver JR, Mason CE, Lauring AS, St George K, MacCannell DR, Grubaugh ND. 2021. Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States. Cell 184:2595–2604. https://doi.org/10.1016/j.cell.2021.03.061
10. Smith MF, Holland SC, Lee MB, Hu JC, Pham NC, Sullins RA, Holland LA, Mu T, Thomas AW, Fitch R, Driver EM, Halden RU, Villegas-Gold M, Sanders S, Krauss JL, Nordstrom L, Mulrow M, White M, Murugan V, Lim ES. 2023. Baseline sequencing surveillance of public clinical testing, hospitals, and community wastewater reveals rapid emergence of SARS-CoV-2 Omicron variant of concern in. mBio 14:e0310122. https://doi.org/10.1128/mbio.03101-22
11. Valesano AL, Fitzsimmons WJ, Blair CN, Woods RJ, Gilbert J, Rudnik D, Mortenson L, Friedrich TC, O'Connor DH, MacCannell DR, Petrie JG, Martin ET, Lauring AS. 2021. SARS-CoV-2 genomic surveillance reveals little spread from a large University campus to the surrounding community. Open Forum Infect Dis 8:fab518. https://doi.org/10.1093/ofid/ofab518
12. Holland LA, Kaelin EA, Maqsood R, Estifanos B, Wu LI, Varsani A, Halden RU, Hogue BG, Scotch M, Lim ES. 2020. An 81-nucleotide deletion in SARS-CoV-2 ORF7a identified from sentinel surveillance in Arizona. J Virol 94:14. https://doi.org/10.1128/JVI.00711-20
13. Currie DW, Moreno GK, Delahoy MJ, Pray IW, Jovaag A, Braun KM, Cole D, Shechter T, Fajardo GC, Griggs C, Yandell BS, Goldstein S, Bushman D, Segaloff HE, Kelly GP, Pitts C, Lee C, Grande KM, Kita-Yarbro A, Grogan B, Mader S, Baggott J, Bateman AC, Westergaard RP, Tate JE, Friedrich TC, Kirking HL, O'Connor DH, Killerby ME. 2021. Interventions to disrupt Coronavirus disease transmission at a university, Wisconsin, USA, August-October 2020. Emerg Infect Dis 27:2776–2785. https://doi.org/10.3201/eid2711.211306
14. Srinivasa VR, Griffith MP, Waggle KD, Johnson M, Zhu L, Williams JV, Marsh JW, Van Tyne D, Harrison LH, Martin EM. 2023. Genomic epidemiology of severe acute respiratory syndrome coronavirus 2 transmission among university students in Western Pennsylvania. J Infect Dis 228:37–45. https://doi.org/10.1093/infdis/jiad041
15. Illumina. 2017. Illumina. Bcl2fastq2 Software Release Notes: bcl2fastq2 2.19.0
16. Illumina. 2023. 2023. DRAGEN COVID Lineage. *Illumina*. Illumina. Available from: https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/dragen-covid-lineage.html
17. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, Yuan M-L, Zhang Y-L, Dai F-H, Liu Y, Wang Q-M, Zheng J-J, Xu L, Holmes EC, Zhang Y-Z. 2020. A new Coronavirus associated with human respiratory disease in China. Nature 579:265–269. https://doi.org/10.1038/s41586-020-2008-3
18. Aksamentov I, Roemer C, Hodcroft E, Neher R. 2021. Nextclade: clade assignment, mutation calling and quality control for viral genomes. JOSS 6:3773. https://doi.org/10.21105/joss.03773
19. Nextstrain 2022. Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months. Nextstrain https://nextstrain.org/ncov/open/global/6m.
20. Huddleston J, Hadfield J, Sibley TR, Lee J, Fay K, Ilcisin M, Harkins E, Bedford T, Neher RA, Hodcroft EB. 2021. Augur: A bioinformatics toolkit for phylogenetic analyses of human pathogens. J Open Source Softw 6:2906. https://doi.org/10.21105/joss.02906
21. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res 30:3059–3066. https://doi.org/10.1093/nar/gkf436
22. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348
23. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using tempest (formerly path-O-Gen). Virus Evol 2:vew007. https://doi.org/10.1093/ve/vew007
24. BedfordT. 2023. *Genomic Epidemiology of SARS-CoV-2 with Subsampling Focused Globally since Pandemic Start*. Nextstrain. Available from: https://nextstrain.org/ncov/gisaid/global/all-time
25. BEAST 2023. Approaches for analyzing large phylogenetic datasets. BEAST https://beast.community/thorney_beast.

26. McCroneJT. 2021. *BEAST v1.10.5 Pre-Release of ThorneyTreeLikelihood v0.1.1*. Github. Available from: https://github.com/beast-dev/beast-mcmc/releases/tag/v1.10.5pre_thorney_v0.1.1
27. Hill V, Baele G. 2019. Bayesian estimation of past population dynamics in BEAST 1.10 using the skygrid coalescent model. Mol Biol Evol 36:2620–2628. https://doi.org/10.1093/molbev/msz172
28. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4:vey016. https://doi.org/10.1093/ve/vey016
29. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using tracer 1.7. Syst Biol 67:901–904. https://doi.org/10.1093/sysbio/syy032
30. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol Biol Evol 29:2157–2167. https://doi.org/10.1093/molbev/mss084
31. Minin VN, Suchard MA. 2008. Counting labeled transitions in continuous-time markov models of evolution. J Math Biol 56:391–412. https://doi.org/10.1007/s00285-007-0120-8
32. Dudas G. 2023. Baltic. Available from: https://github.com/evogytis/baltic. Retrieved 3 Oct 2023.
33. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. 2016. SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes. Mol Biol Evol 33:2167–2169. https://doi.org/10.1093/molbev/msw082
34. Lemey P, Ruktanonchai N, Hong SL, Colizza V, Poletto C, Van den Broeck F, Gill MS, Ji X, Levasseur A, Oude Munnink BB, Koopmans M, Sadilek A, Lai S, Tatem AJ, Baele G, Suchard MA, Dellicour S. 2021. Untangling introductions and persistence in COVID-19 resurgence in Europe. Nature 595:713–717. https://doi.org/10.1038/s41586-021-03754-2
35. Guerrero-Murillo M. 2023. *QSutils: Quasispecies Diversity*. Available from: rdrr.io https://rdrr.io/bioc/QSutils/
36. Ngumbang J, Meredith M, Kruschke J. 2022. *HDInterval: Highest (Posterior) Density Intervals*. Available from: cran.r-project.org https://cran.r-project.org/web/packages/HDInterval/index.html
37. Unknown. 2020. *Governor Walz Extends Stay Home Order in Minnesota*. Available from: MN.gov https://mn.gov/governor/newsroom/press-releases/#/detail/appld/1/id/430501
38. Unknown 2022. Impact of Opening and Closing Decisions by State. Johns Hopkins University https://coronavirus.jhu.edu/data/state-timeline/new-confirmed-cases/minnesota/99.
39. SchultzD 2020. Health officials confirm first case of novel coronavirus in Minnesota. https://www.health.state.mn.us/news/pressrel/2020/covid19030620.html.