January 2023

A Phylogeographic Analysis Of Sars-Cov-2 Variant Introduction

Kien Pham
duckien242@gmail.com

A Phylogeographic Analysis of SARS-CoV-2 Variant Introduction

Name: Kien Pham

Institution: Yale School of Public Health

Degree Awarded: Master of Public Health

Year Degree Awarded: 2023

Advisor 1/Committee Chair: Dr. Nathan Grubaugh

Advisor 2/Committee Member: Dr. Theodore Cohen

Abstract

The ongoing SARS-CoV-2 pandemic is marked by continuing emergence of novel Omicron variants. However, the mechanism with which these lineages establish themselves in new geographical areas remains an understudied subject. In this research, we utilized a discrete phylogeographical analysis framework on 19,608 SARS-CoV-2 sequences of the BA.5 Omicron lineages between February and June 2022, in order to better understand how the variant emerged in different regions of the United States. We found that the earliest BA.5 introductions came from Africa, befitting the origin of the lineage. Additionally, we discovered extensive domestic transmission between the Department of Health and Human Services (HHS) regions of America through both neighboring effects and cross-country transmission. We found further evidence that two key regions that include California and Florida drove most within-U.S. introduction events of BA.5. Our results suggest a pattern of foreign importation and domestic spread, through interaction of these hotspot areas, and form a framework for understanding novel variant emergence in the United States.

Table of Content

Figures and Tables

Introduction

As the COVID-19 pandemic enters its fourth year, SARS-CoV-2 virus has demonstrated the ability to evolve into novel variants to maintain the cycle of infection. Late 2021 saw the emergence of the Omicron (B.1.1.529) variant in southern Africa, which was deemed a variant of concern (VOC) by the World Health Organization (WHO) and soon became dominant in the United States and the rest of the world. Sequencing results of Omicron samples revealed 50 mutations across the viral genome, with over 32 in the spike protein, that granted it an advantage over co-circulating variants in terms of transmissibility and immune escape (Ye et al, 2021) (Viana et al, 2022) (Chaguza et al, 2022). Upon its discovery in November 2021, Omicron consisted of three original lineages: BA.1, BA.2, and BA.3; with the former two sequentially becoming more dominant around the world. By mid-2022, two additional lineages, BA.4 and BA.5, emerged in southern Africa and quickly co-circulated in the U.S. (Tegally et al, 2022) (U.S.-CDC). As 2023 rolled around, XBB.1.5, a recombinant variant of two BA.2 sub-lineages, overtook BA.5 in frequency in America (JHU Coronavirus Resource Center). This cycle of emergence is a natural consequence, as new variants are positively selected and carry mutational advantages enabling their dominance (Chaguza et al, 2022) (Cyupers et al, 2021). As the world began lifting social and travel restrictions, the landscape is rife for new variants to emerge. Understanding the process of lineage emergence and spatial dissemination is thus essential to scaling future surveillance and intervention of SARS-CoV-2, as well as those of other viral pathogens.

Phylogeography is the study of the processes underlying past and current spatial distribution of pathogen genealogical lineages. Recent advances in viral sequencing and phylogenetics have enabled more widespread and timely use of phylogeography to understand SARS-CoV-2 dynamics (Hill et al, 2023) (Dellicour et al, 2021) (Grubaugh et al, 2019). Early in the pandemic, Fauver et al (2020) combined sequencing data of 177 samples with air travel data to infer greater COVID-19 importation risk into

Connecticut from other U.S. states than from other countries. Alpert et al (2021) performed discrete phylogeographic analysis and ancestral trait reconstruction of 1,908 Alpha (B.1.1.7) sequences to infer 23 distinct B.1.1.7 introduction events leading to secondary transmission in the U.S... The emergence of Omicron variants coincided with scale-up of global immunization campaigns and eventually lifting of travel restrictions. Their effects were thus closely monitored by public health authorities and researchers. Performing a Bayesian phylogeographic analysis on a set of 10,403 genomes, Douglas et al (2022) estimated risk of Omicron importation into New Zealand at 1 per 5,000 arrivals into the island nation. No scholarship on a countrywide scale has been conducted to study Omicron variant introduction into the United States, however. We thus believed it was important to discern the dynamics on which these Omicron variants emerge and overtake one another, causing repeated infections and continuation of the SARS-CoV-2 pandemic.

This paper utilized the Bayesian discrete phylogeographic framework to understand the emergence of a novel SARS-CoV-2 lineage into the different regions of the United States. To this end, the Omicron BA.5, which first appeared in America in early 2022, was selected as the variant of study (U.S.-CDC). Unlike other Omicron lineages of BA.1 and BA.2, BA.5 established itself during times of relatively lower SARS-CoV-2 incidence and remained prominent in variant frequency for a longer period (John Hopkins University). While the most widely circulated variant in the U.S. for much of later 2022, BA.5 never achieved complete dominance, but instead co-circulated with other major lineages such as BA.2.12.1, BA.4, and XBB.1 (U.S.-CDC). As opposed to past dominating variants, BA.5 faced an American population having many vaccinated with the primary SARS-CoV-2 series plus booster, and who had largely been exposed to past Omicron infections of BA.1 and BA.2 (CDC COVID Data Tracker) (Chitwood et al, 2021). Moreover, its introduction approximately coincided with relaxing of restrictions such as masking and indoor gathering in most areas, as the U.S. attempted to return to pre-pandemic life (Our

World in Data). As such, BA.5 introduction dynamic would not only be unique from past lineages but would also largely be representative of that of future SARS-CoV-2 variants.

We define emergence to be the period in-between inferred introduction of the first BA.5 sequence and when the proportion of BA.5 samples in all regions reaches a minimum threshold in frequency. Upon examination of BA.5 frequency plot by U.S. region, we set this threshold to be 27 percent. While the first BA.5 sample in the U.S. was collected in late February 2022 (GISAID), its introduction likely predated this, given the limited proportion of SARS-CoV-2 cases sequenced and availability of home testing. Further, with the United States covering a large swarth of territory and having daily international arrivals disproportionately distributed throughout the country, importation rate and time into America would likely have varied by region (Alpert et al, 2021). Given the large volume of both international travel into and domestic travel around the country, we theorize that inter-region introduction events of BA.5 were prominent in the variant's early stages. Past scholarships by Fauver et al (2020) and McBroome et al (2022) have mentioned domestic introduction as a major mechanism of SARS-CoV-2 spread in the America. In the context of an Omicron variant, however, no research on a national scale has been conducted, much less focused on transmission between states. Through application of Bayesian discrete phylogeographic analysis, this paper shall explore the difference in timing, rate, and mechanism of early BA.5 emergence between the regions of the United States.

Methods

1. Data collection

In order to reconstruct the dynamics of variant introduction into the U.S., we performed analysis on a dataset of BA.5 genomic sequences submitted in the country and around the world during the inferred emergence period, estimated between February and June 2022. First, we downloaded all

sequences with assignment of a BA.5 lineage, as designated by *pango* nomenclature (Rambaut, 2020) from the GISAID database, with their corresponding metadata included. We performed data cleaning through the R programming language, first by filtering out sequences without proper Collection date and Location entry, before separating the data into two sets: *USA* and *Global*, for the latter subsampling step. The programs Geneious and Nextclade were also utilized to filter out sequences with a low quality control score (Nextclade) and large N gaps within their genomes (above 30 percent of genome) (Hadfield et al, 2018) (Geneious Prime 2021).

Discrete phylogeographic analysis requires the specification of geographic locations for estimation of pathogen dissemination. Trait specification must be specific enough for meaningful phylogeographic results, but also not bias the analysis towards more locations with higher proportion of sequences. In a global context, there is a noticeable disparity in global capacity, where we see large numbers of published SARS-CoV-2 sequences belonging to developed countries (Brito et al, 2022). Considering this fact, the research categorized Global BA.5 data by continent. Within the U.S., genomic surveillance policy is largely decided by the individual state. Abbasi (2021) showed a difference between states with regards to variant surveillance efforts, resulting in a national genomic database heavily skewed towards states with more sequencing labs and more funding. In order to ameliorate this disparity in data, we decided against using U.S. states as traits. Rather, we divided the country into ten regions, according to the ten regional offices of the Department of Human and Human Services (HHS). All in all, the research shall examine sequences from six continents – Asia, Africa, Europe, North America (minus the U.S.), South America, and Oceania – and from the United States. A follow-up analysis would further split the U.S. into the ten HHS regions for domestic introduction analysis.

In order to reduce computational load and reduce sampling bias from regions with better sequencing capacity, we performed additional subsampling of BA.5 sequences. We defined emergence as the period in-between an initial variant introduction event, or events, leading to its spread and

achieving at minimum 27 percent in frequency within the region's SARS-CoV-2 population, according to GISAID data. We set the 27 percent threshold after plotting cumulative BA.5 progression by region, with a view to portray the period up until the end of exponential growth in every area of study, both globally and within the United States. By setting 27 percent as the minimum frequency, we ensured adequate representation of BA.5 genomes between all regions both in terms of size and of time.

As mentioned prior, genomic sequencing capacity differs noticeably between countries and between U.S. states (Brito et al, 2022) (Abbasi, 2021). To account for possible selection bias from this heterogeneity, we performed subsampling separately for each global and U.S. region, and elected to sample for 75 percent of all sequences. The final dataset selected for analysis consists of 19,608 sequences, 9,350 of which have "USA" in their country and the remaining 10,258 are from non-USA countries. For the emergence period, the earliest sample was collected on February 25, 2022, whereas the latest sample was collected on June 19, 2022. The following table shows the final count of sequences for all the respective regions of analysis:

Dataset				
Global			**USA**	
Africa	638		Region 1	373
Asia	1185		Region 10	341
Europe	6380		Region 2	1071
North America	770		Region 3	383
Oceania	466		Region 4	1879
South America	819		Region 5	959
			Region 6	822
			Region 7	212
			Region 8	508
			Region 9	2802

Table 1: Final sequence dataset for BA.5 discrete phylogeographical analysis

2. Data analysis

Once the data cleaning process for the 19,608 sequences was complete, we could begin the discrete phylogeographic analysis pipeline. First, multiple sequence realignment was generated using the Nextclade tools (Neher et al, 2021), with Wuhan-Hu-1/2019 as the reference genome. Afterwards, we constructed a maximum-likelihood phylogenetic tree using the IQ-Tree program (Nguyen et al, 2015). We assumed an HKY nucleotide substitution model, which assumes unequal stationary frequencies between nucleotide bases as well as different rates of transition and transversion (Hasegawa et al, 1985) (Dunn, 2023). Within the maximum likelihood tree, we pruned out any molecular clock outliers, as well as the outgroup used for rooting, using jclusterfunk toolset (Snake-flu Github) (Hill & Baele, 2019). Next, we generated an approximate tree-calibrated tree using TreeTime, which we used as the starting tree for BEAST to reduce the number of discarded states for burn-in (Sagulenko et al, 2018).

Given the large size of the genome set of analysis, we ran BEAST v.1.10.4 with the Thorney extension, which allowed for time calibration of larger phylogenetic dataset by using an alternative tree likelihood function (Suchard et al, 2018) (du Pleissis et al, 2019) (Didelot et al, 2020). With the starting time tree from TreeTime and the maximum likelihood tree obtained from IQ-Tree, the research generated an XML file for the Thorney BEAST run, with a fixed clock rate of 8x10E-4 substitutions/site/year, as the study period (February 2022 to June 2022) was too short to invoke any strong temporal signals. The Bayesian *skygrid coalescent model* (Gills et al, 2013) was used, which allows for estimation of effective population size at defined time points, with 23 grid points defined according to approximate equal intervals of the global emergence period shown above. We consequently ran the MCMC on BEAST v.1.10.4 for 1,000,000,000 iterations (Suchard et al, 2018). The program Tracer was then used to assess convergence upon run completion, with 10 percent of states discarded for burn-in

(Rambaut et al, 2018). After confirmation of model convergence, we selected the final tree of the resulting *.trees* file for the final phylogeographic analysis for computational efficiency purposes.

From the resulting tree generated with Thorney BEAST, we ran the final discrete traits analysis. Here, we added the discrete locations as traits for BEAST to reconstruct the locations of the internal nodes. The first run contained the seven states for the six global continents (with North America not including the U.S.) and USA as a country. The results of this run helped us assess the rate of BA.5 importation into the whole of America. The second run also contained the global continents plus the ten HHS regions representing the United States. Such subcategories would reveal the specific interactions between global regions and individual states, as well as between the U.S. regions. We continued with the Bayesian skygrid coalescent model (Gills et al, 2013) and ran on BEAST the MCMC chain of 2,000,000 iterations. We further wrote the complete history tree log to file at every 500 iterations, and otherwise maintained similar parameters as our previous Thorney BEAST run. Once this final BEAST run concluded and passed test of convergence on Tracer, we used Python scripts to estimate the average number of introductions across all posterior and then inferred a final tree that came closest to said figure. Tree #1024000 was selected as that most representative of average introductions. We then ran additional Python scripts to visualize the locations-annotated phylogenetic tree, map out proportion of introductions between locations, and construct a source-sink map for all locations.

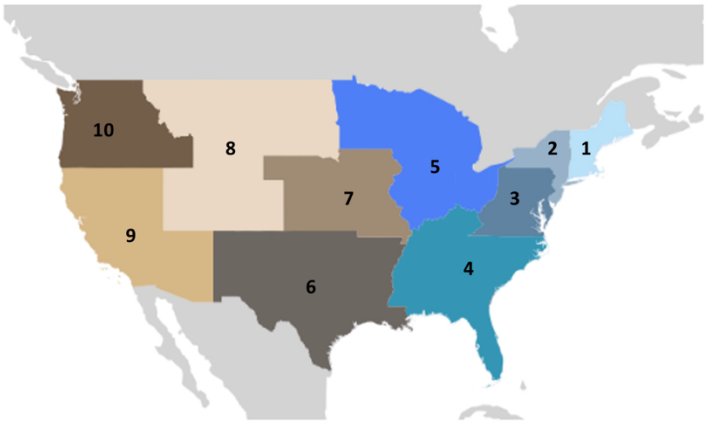
Results

BA.5 emerged at differing time periods between the U.S. regions.

In uncovering the patterns of variant emergence in the United States, we hypothesized that Omicron BA.5 emergence was different between regions over time. That SARS-CoV-2 general transmission within states differs by U.S. states and by presence of major travel centers (McBroome et al, 2022) suggests a novel lineage may be introduced into America through different introduction events

happening at different time points and in different regions. Any research into variant emergence would therefore need to investigate introduction in different areas of the United States. After filtering the GISAID sequence dataset for U.S. samples, we further subdivided these genomes by the regions of study, before calculating frequency of BA.5 over time. Analysis of this build-up phase showed a two-month gap between the region of earliest sequence submission (Region 9) and that of the latest (Region 10) (Fig. 2), as well as significant difference in variant frequency over time by HHS region.

Our preliminary analysis involved exploring GISAID database for BA.5 submission by HHS region and calculating variant frequency in data reporting. First, we downloaded from GISAID all SARS-CoV-2 sequences publicly available before Oct 25, 2022, for a total of 13,632,956 entries. These sequences were then limited to those after January 1, 2022, filtered for missing date and lineage data, and categorized into two datasets: USA genomes and Global (non-USA) genomes. To account for heterogeneity in states' sequencing capacity (Abbasi, 2021), we further separated U.S. data according to the ten Department of Health & Human Services (HHS) regions (Fig. 2A).



Region 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont
Region 2: New Jersey, New York, Puerto Rico, Virgin Islands
Region 3: Delaware, District of Columbia, Washington DC, Maryland, Pennsylvania, Virginia, West Virginia
Region 4: Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, Tennessee
Region 5: Illinois, Indiana, Michigan, Minnesota, Ohio, Wisconsin
Region 6: Arkansas, Louisiana, New Mexico, Oklahoma, Texas
Region 7: Iowa, Kansas, Missouri, Nebraska
Region 8: Colorado, Montana, North Dakota, South Dakota, Utah, Wyoming
Region 9: Arizona, California, Stockton, Hawaii, Nevada, Northern Mariana Islands, Guam, Marshall Islands, Palau, American Samoa
Region 10: Alaska, Idaho, Oregon, Washington, Tribal

Figure 1: Department of Human Health & Services (HHS) regional designation of the United States.

For more fine-grained analysis, we restricted the data by the emergence period of BA.5 frequency, defined as time up to when all regions have had BA.5 reach a minimum 27 percent in proportion. Such time restriction ensured our dataset adequately captured the exponential growth taking place during the early phase of BA.5 entry in every region of study. In calculating for the proportion of BA.5 over time, we plotted the three-day moving average of BA.5 frequency to better adjust for temporal bias in sequencing (Fig. 2). We then constructed a mixed-effect negative binomial regression model on the moving average frequency, controlled for date and for effect modification via HHS regions, to see whether BA.5 frequency over time was significantly different by U.S. regions.

Our frequency graph displayed variation in period of BA.5 appearance into each HHS region (Fig. 2). The lineage first appeared in America in Region 9 on February 26, 2022; before surfacing in Region 6 one day later. Region 10 had the latest BA.5 discovery on April 27, 2022 – approximately two months since the lineage appeared in Region 9's surveillance radar (Fig. 2). The remaining eight regions had their first sequence reported in the one month's span between mid-March to mid-April. Further, there existed a submission lag after BA.5 discovery, as regions did not consistently report BA.5 sequences until the middle of April 2022. After mid-April, or once BA.5 have appeared in all ten HHS regions, variant frequency grew at an exponential rate in every region (Fig. 2). Moreover, our negative binomial regression model also showed a statistically significant difference in the interaction coefficient between time and four HHS regions, suggestive of variant frequency over time differing between the HHS regions (Supplementary Material).

From this initial data exploration, we may conclude that novel lineage got introduced at varying time points in different geographic regions of the United States. The timeline of lineage reporting to GISAID also suggested an initial period of slow BA.5 foreign importation into all regions, followed by faster growth in frequency in late May 2022. We also believed that heterogeneity in discovery time between regions suggested that regions receiving BA.5 introductions earlier played a part in inter-

regional transmission. We were cognizant of the fact, however, that the preliminary result was indicative of GISAID submission, a possible bias due to different surveillance policies, and not of actual variant dynamic, which would be better explained through phylogeography method.

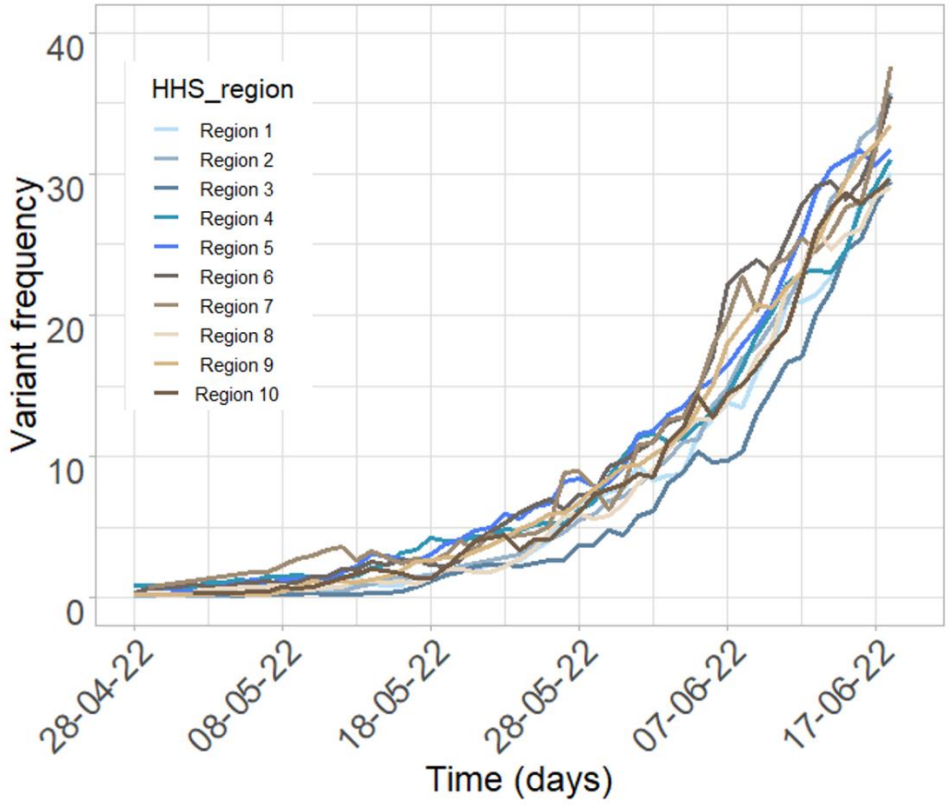


Figure 2: Frequency of BA.5 variant in United States regions during emergence period. *BA.5 3-day moving average variant frequency up to 06.19.22, by HHS regions.*

Phylogenetic evidence revealed early BA.5 introductions into the U.S. were from Africa, with later transmission characterized by inter-region spread.

Visualization of U.S. sequencing data suggested that the novel BA.5 variant emerged at various time points in different regions, with some areas of earlier introduction possibly playing a role in later countrywide spread (Fig. 2). To investigate the initial foreign importation of BA.5 and the possibility of inter-region transmission, we conducted Bayesian discrete phylogeographic analysis using 19,608 BA.5 sequences from the U.S. and other global regions. Our final location-annotated phylogenetic tree

revealed Africa and Europe playing a major role in exporting the novel variant to America, with the former continent responsible for the earliest introductions (Fig. 3A). Moreover, annotating with HHS regions showed early BA.5 inter-region spread as characterized by two major U.S. – only clades with origins in Region 4 of southeastern states and Region 9 of southwestern ones (Fig. 3B-C).

From the total of 13,632,956 SARS-CoV-2 entries publicly available on GISAID database, we filtered for and selected 1,400,088 USA sequences and 4,313,313 Global. The USA sequences were then subdivided according to the HHS regional categorization. Additional filters selected sequences according to the regional BA.5 emergence phase, defined as the period up to variant proportion in all regions reaching at least 27 percent. After subsampling by 75 percent, we finally reduced the data to a representative set of 19,608 BA.5 sequences, which consists of 9,350 sequences for USA and 10,258 for Global. From this dataset, we inferred a maximum-likelihood phylogenetic tree, scaled in time using Thorney BEAST, on which we performed Bayesian discrete phylogeographic analysis (Lerney et al, 2010) (Dellicour et al, 2020) (Anderson et al, 2021). From the resulting location-annotated MCC trees, we picked out an average tree based on the mean number of introductions, defined by the descending clades (Anderson et al, 2021) (Fig. 3A). We colored the final tree by global regions for inferring foreign introduction into the U.S. (Fig. 3A) and again by HHS regions for estimating domestic transmission between regions (Fig. 3B-C). For HHS regions, the color schemes were blue colors for East -Central regions and brown for West -Central regions (Fig. 3A-C).

The results of the phylogeographic analysis revealed on average 1306 separate introduction events into the 10 HHS regions of the United States during their emergence period in the first half of 2022. From the final tree, we could infer the mean Time to Most Recent Common Ancestor (tMRCA) and with it hypothesize the estimated time of variant introduction into the United States. As shown in Figure 3A, BA.5 most likely appeared in America around early February 2022, nearly three weeks prior to the collection date of the earliest USA sequence – February 26 (GISAID). The descending clades revealed

that Europe and Africa form almost all of BA.5 foreign introductions into the U.S (Fig. 3A). The three African clades, in particular, had earlier U.S. importation than the European ones, and the oldest and largest American clade also had its origin traced back to Africa (Fig. 3A-B). Specifically, Figure A showed the first American clade branching off from Africa in the first week of February, whereas . We thus inferred that BA.5 was first introduced into America from African countries. The assumption made sense on chronological and epidemiological levels, given that BA.5 was first discovered in southern Africa before being exported to Europe (Viana et al, 2022) (World Health Organization).

By categorizing the USA sequences by their HHS regions, we discovered heavy mixing of U.S. genomes, which suggested inter-region transmission as a key factor in BA.5 emergence. The earliest and largest USA clade originated in Region 4, after introduction from Africa, in early February 2022, approximately three weeks before the earliest BA.5 genome was reported to GISAID, and diverged to other regions in mid-March (Fig. 3A-B). Within this Region 4 clade, there were two notable subclades branching into Region 9 and Region 5 (Fig. 3B). We also noted another large clade that branched off from its Africa root in mid-March into Region 9 (Fig. 3B). This clade contained three subclades going into Region 9 (Fig. 3B). The remaining Europe clades saw introductions into the U.S. by late April to May 2022, one to two months later than the African clades. For every clade in the final tree, however, foreign introduction was followed by heavy domestic transmission, as we noticed a higher proportion of USA genomes later in the branches, by the last week of May (Fig. 3A-C). While the dichotomy of our two USA and Global datasets may have inflated this proportion, our result did show that inter-region transmission played as key a role as foreign importation during variant emergence. Close viewing of the USA clades (Fig. 3B-C) showed possible geographic factor involved in inter-region transmission. The Region 4 clade primarily saw transmission to other East Coast and East-Central regions (Fig. 3B), whereas the Region 9 clade to other West Coast and West-Central regions (Fig. 3C). We hypothesized from this observation that inter-region transmission of a novel lineage was influenced by geographic proximity,

which matched with McBroome et al (2022) assertion of the neighboring state impact on SARS-CoV-2 transmission. Nonetheless, interaction between Region 4 and Region 9 in the form of subclade branching (Fig. 3B-C) indicated the presence of coast-to-coast introduction as an important mechanism in BA.5 emergence.

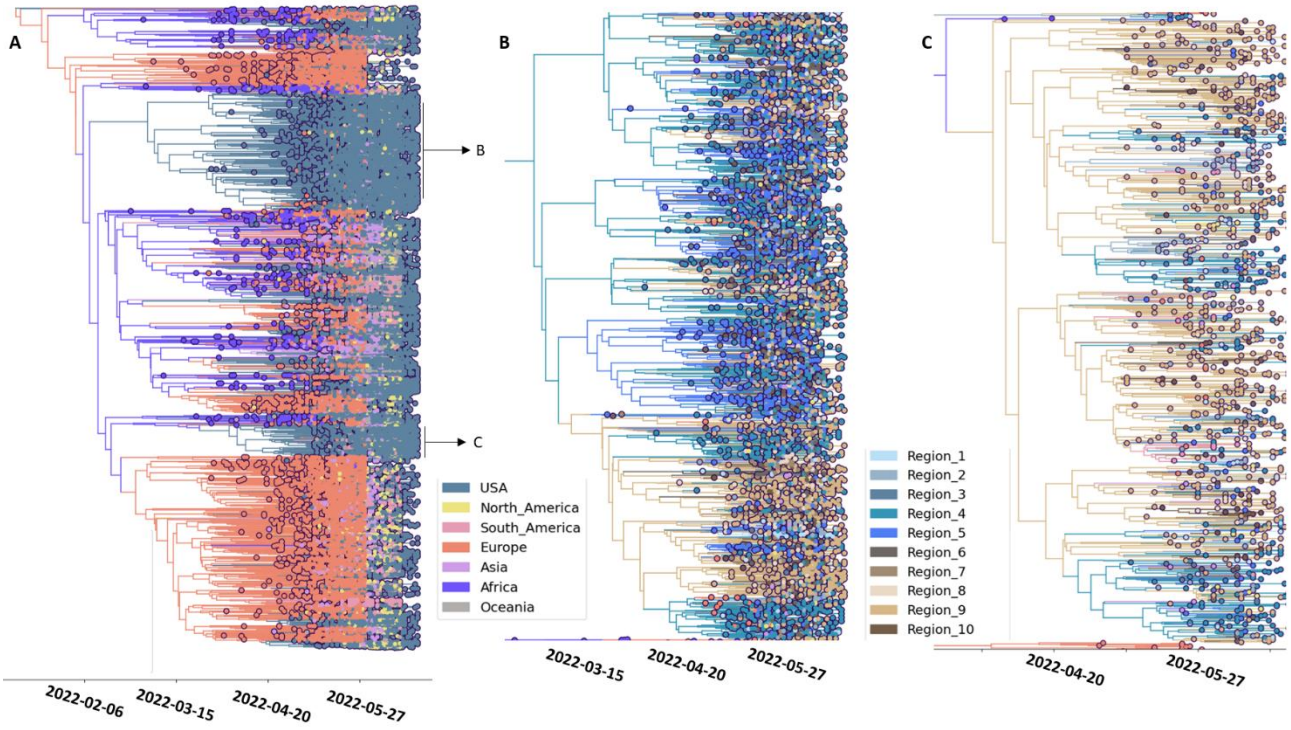


Figure 3: Phylogenetic representation of foreign introduction and domestic transmission of BA.5 during its emergence in the United States. *(A) Maximum likelihood phylogeny of BA.5 emergence using 19,608 USA and global sequences, rooted on Wuhan-Hu/1 genome. (B) Region 4 clade. (C) Region 9 clade.*

Earliest importation of BA.5 originated in Africa, while Europe formed the majority of introductions.

By parsing the Markov jumps in the discrete phylogeographic results, we plotted BA.5 lineage movements across the discrete regions to infer major sources of introduction. A timeline of variant introduction events by global region revealed that Africa was a major exporter of BA.5 to the U.S. prior to May 2022, before other regions overtook them at the latter half of the emergence period (Fig. 4A). Europe in particular was responsible for the majority of foreign introduction events, with numbers peaking in late May before slowing down by the end of the emergence period (Fig. 4A). Moreover,

examination of the global introduction rate by HHS regions revealed significantly higher Europe rates into Region 4, Region 9, and Region 2. Region 4 and Region 9 also had higher rates of introductions from Africa compared to the remaining eight, which suggests they might be the sites of initial BA.5 introduction into the United States.

Our discrete traits analysis recorded the entire history of our tree results, or every transition between geographic locations. From here, we filtered the history of our chosen tree to include only transitions from a global region into the United States, which we then defined as foreign introduction events into America. In order to understand how BA.5 emergence changed over time; we plotted the rates of introduction by global region for every week of the defined emergence period (Fig. 4A). Following observations of the time series from Figure 4A, we mapped out the global transmission of BA.5 into America by the initial and latter half of the emergence period, defined as before and after May 15, 2022 (Fig. 4C-D). Finally, in our HHS region-specific phylogeographic results, we graphed out foreign introduction rates into all ten regions (Fig. 4B). Such plot would discern variance in BA.5 importation patterns between U.S. regions, as well as demonstrate areas of high foreign introduction.

Our weekly time series of global BA.5 movements into America (Fig. 4A) reconfirmed our assertion of early introductions coming from Africa. Specifically, we discovered that, prior to the first week of May, the vast majority of introductions were in the Europe or Africa categories (Fig. 4A). That Europe should dominate the other global regions in proportion is expected, given the bias in SARS-CoV-2 global sequencing (Brito et al, 2022) and in the distribution of passengers coming to U.S.. In our Global dataset, Europe comprised 7475 out of the 11216 BA.5 sequences, compared to only 683 from Africa. Phylogeographic results showed Africa only constituting 9 percent of total BA.5 foreign introductions, but, when restricted to the period up to mid-May 2022, that number increased to 23.1 percent. Such higher proportion of African introductions early in the BA.5 timeline, together with known origin of the lineage in southern Africa (World Health Organization) (Viana et al, 2022), would therefore indicate the

continent as the source of initial importation into America (Fig. 4A-C). Were our hypothesis true, it would mean that the BA.5 came to the United States from its southern Africa origin point through multiple introduction events, distinct from its transmission into Europe. The latter half of the time series (Fig. 4A-D) saw Europe continuing to dominate, with Asia and the remainder of the Americas rising in proportion, while Africa began to disappear. We interpreted this as: once BA.5 became established globally, we would see introduction rates more in line with the typical travel rate into the U.S. from regions of the world.

Through the HHS region-specific view, we saw a clear distinction in foreign introduction rates to different areas of the U.S. Three of these regions stood out for having significant higher rates than the rest: Region 9, Region 4, and Region 2 (Fig. 4B). Inclusion of domestic introduction rate in the proportional chart also showed Region 2 and Region 9 as the only two having foreign introductions of over 40 percent, while Region 4 stood at over 25 percent (Supplementary Material). As a result, the three regions were inferred to have been the key hubs for BA.5 introductions from other countries. Such assertion would be most probable, given all three include states of high population with large international airports (California, Florida, and New York), factors known to impact SARS-CoV-2 transmission (Anderson et al, 2021). Within all ten HHS regions, we saw European introductions forming the majority, although the disparity was especially noticeable in the three Region 9, Region 4, and Region 2 (Fig. 4B). Europe, therefore, was the main exporter of BA.5 sequences to the United States. Nevertheless, when viewed separately, rates of introduction from Africa were higher in Region 4 and Region 9 (Fig. 4B). Given what we theorized on the origins of BA.5 in the U.S. being from Africa, these higher proportions pointed to the first BA.5 introductions having come from these two regions. Finally, regions of low importation rates – such as Region 3 and Region 7 – most likely had introductions coming from domestic transmission (Fig. 4B).

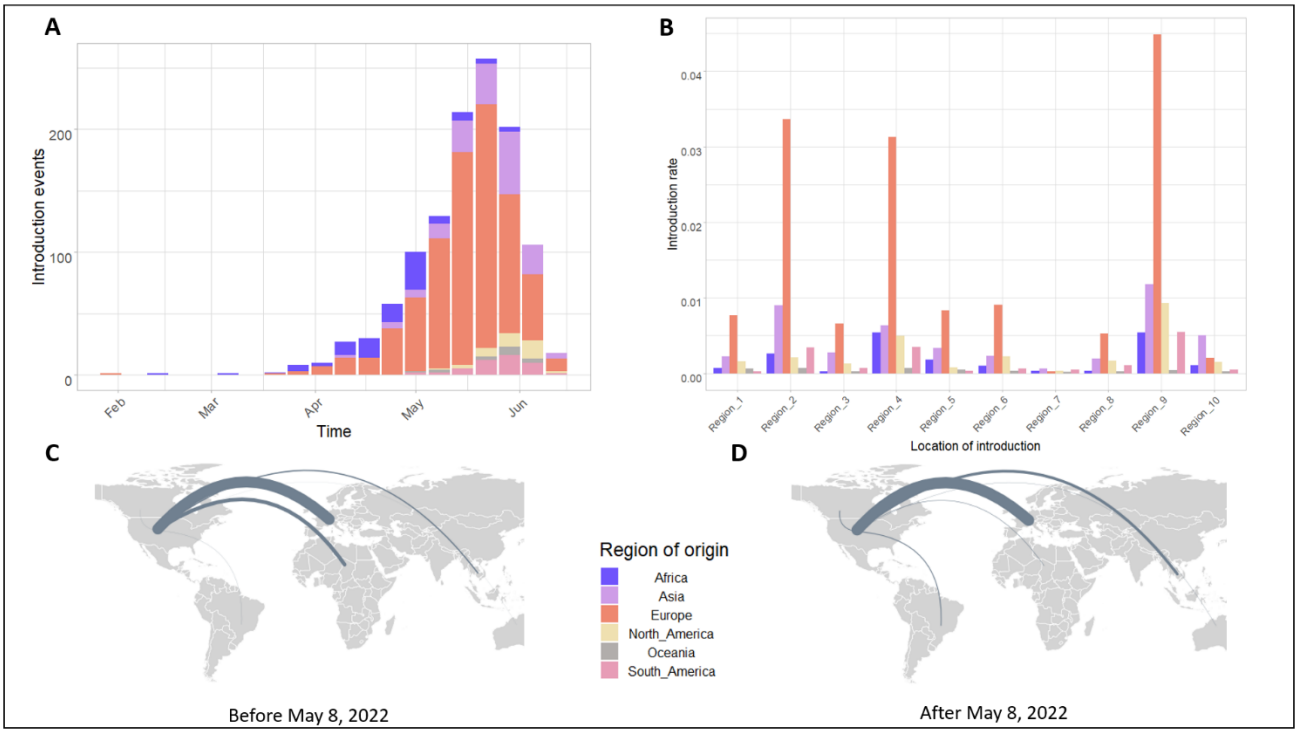


Figure 4: Spatiotemporal dynamics of BA.5 lineage foreign introductions to the United States. *(A) Timeline of BA.5 foreign introduction events to the United States, by continent. (B) Comparison of BA.5 foreign introduction events to ten U.S. HHS regions, by continent. (C) Map of cumulative BA.5 foreign introduction events to the United States.*

Key hotspots were responsible for facilitating domestic transmission of BA.5 during its emergence period.

Domestic transmission played a prominent role during BA.5 emergence in the United States. In fact, all 10 HHS regions had more than 50 percent of introduction events coming from within the U.S., with 8 having more than 80 percent (Supplementary Material). Therefore, we decided to investigate the characteristics of inter-region transmission of BA.5 through intra-U.S. movements in our phylogeographic results, which we indicated as domestic introduction. All HHS regions had considerably higher introduction rates from Region 4 (Florida and southeastern states) and Region 9 (California and other southwestern states) compared to others (Fig. 5C). We also witnessed a minor geographical proximity effect in the form of relatively higher rates between neighboring regions. Introduction rates were highest between Region 4 and Region 9, and to a lesser extent Region 2 (New York and New

Jersey) and Region 5 (Illinois and other midwestern states) (Fig. 5C). By plotting the source-sink map (Fig. 6A-B), we confirmed that Region 4 and Region 9 were the main distributors and recipients of BA.5 introductions. These results together demonstrated that domestic introduction, dominated by interactions of key hotspots, was the primary driver of inter-region introduction of BA.5.

From the phylogeographic results of our HHS region-specific analysis, we filtered for introduction events in-between U.S. regions, so as to investigate the role of domestic transmission in BA.5 emergence. We then plotted a map of BA.5 movements across regions of America (Fig. 5A), with line size representing the scale of interactions between the HHS regions. Similar to the U.S.-only analysis (Fig. 5B), we graphed the different domestic introduction rates for each HHS region (Fig. 5C). Finally, side-by-side mapping on Figure 6 showed the prominence of each HHS region as the "source" or "sink" regarding BA.5 emergence – or the degree to which they exported or received introduction during the study period. Overall, the visuals informed us on domestic transmission of BA.5, key interactions between U.S. regions, and potential hotspots for transmission during its emergence period.

Our results showed distinct patterns in domestic BA.5 transmission within the regions of America. Specifically, Region 4 and Region 9 were the two areas with highest introduction rates into all remaining HHS zones (Fig. 5A-C). Region 2 and Region 5 also featured prominently in four other regions respectively (other East Coast regions for Region 2, neighboring south-central regions for Region 5) (Fig. 5C). Overall, we noted Region 4 and Region 9 as the top sources as well as top recipients of BA.5 introductions (Fig. 6). In fact, 19 out of the top 20 domestic introductions included one (or both) of these two regions, in either capacity (Fig. 5C) (Supplementary Material). Additionally, Region 4 and Region 9 were the other's top source of BA.5 introductions. Interactions between the two were the most numerous – comprising 16.8 percent of cumulative domestic introductions (Fig. 5A-B). These statistics indicated once more that Region 4 and Region 9 were primarily responsible for BA.5 inter-region transmission in the U.S. Given the presence of key population centers and airports within these regions

(Florida and California), these results are in line with past scholarship on SARS-CoV-2 transmission patterns being linked to urban areas and travel data (Alpert et al, 2021) (McBroome et al, 2022).

When observing domestic introduction rates of BA.5, we discovered a partial neighboring effect between the HHS regions (Abbasi, 2021). Specifically, a region would be more likely to receive introductions from its neighbor, regardless of overall introduction rates. An example of this would be Region 9 receiving 5.8 percent of its introductions from the adjacent Region 10, more than five times the rate of introductions from Region 10 to any other regions (Fig. 5C). Likewise, Region 1 and Region 3 received 25.8 percent and 22.1 percent of introductions from Region 2, respectively, much higher than Region 2 rates in other areas (Supplementary Material).

The exception to the neighboring effect lay in cross-country transmission taking place between Region 4 and Region 9, and to a lesser extent Region 5 and Region 2 (Fig. 5A-C). As mentioned, rates of introduction between Region 4 and Region 9 were the two highest in the U.S., together forming 16.8 percent of total domestic movements of BA.5 (Supplementary Material). Such extensive interaction and transmission between Region 4 and Region 9 formed the backbone of BA.5 emergence, given their high introduction rates into other regions. Region 5 and Region 2 also played supportive roles in this transmission network as the third and fourth most interactive areas. Together, the interactions between these 4 regions made up 91.1 percent of all inter-region introduction events in America. Overall, our research proved domestic introduction of novel BA.5 in the U.S. was characterized by neighboring effects and cross-country interactions between key hotspots.

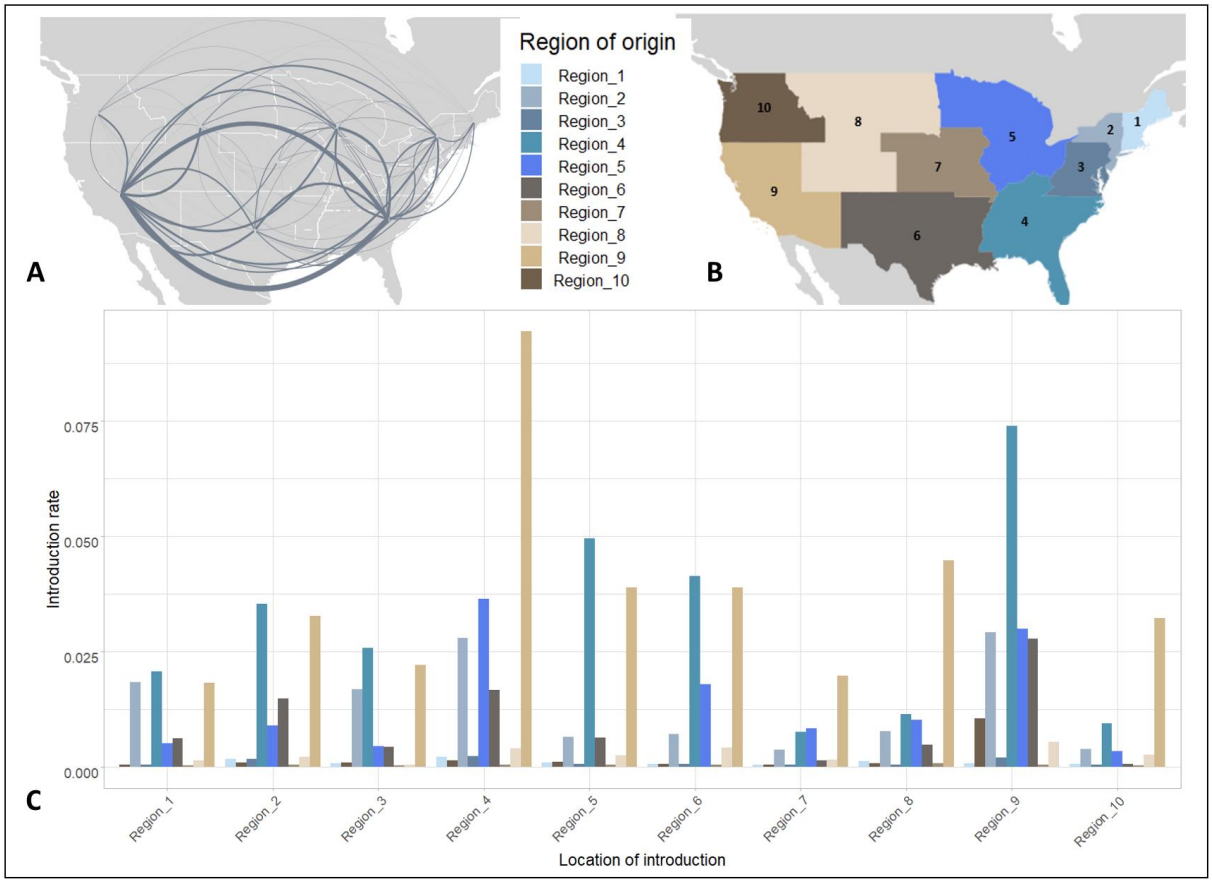


Figure 5: Spatiotemporal dynamics of BA.5 lineage domestic introductions within the United States HHS regions. *(A) Map of cumulative BA.5 inter-region introduction events within the United States (B) Department of Human Health & Services (HHS) regional designation of the United States. (C) Comparison of BA.5 inter-region introduction events between ten U.S. HHS regions, by HHS region.*

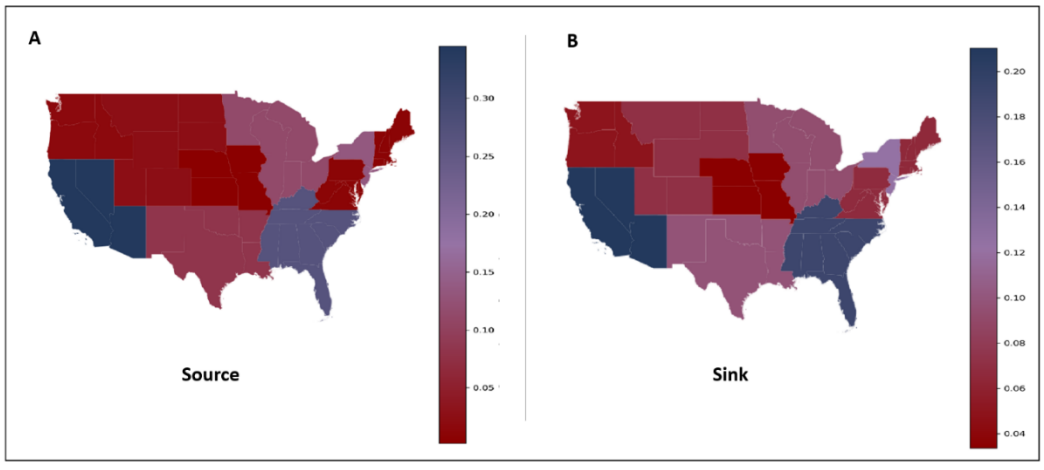


Figure 6: Source-Sink map of BA.5 lineage domestic introductions within the United States HHS regions.

(A) Source map for of top exporters of domestic BA.5 introductions. (B) Sink map for top importers of domestic BA.5 introductions.

Discussion

The continuing spread of SARS-CoV-2 in the U.S. and around the world after three pandemic years calls for the need to understand its mechanism of emergence and transmission. In this paper, we applied phylogeographic analysis of early BA.5 sequences as proxy for gauging the dynamics of Omicron emergence and establishment in different U.S. regions. Discrete traits analysis results revealed BA.5 introduction in America was characterized by both international importation events and domestic transmission, fueled by interactions between various key regions in the country. As the first nationwide, large-scale study of Omicron lineage establishment, our findings support the role of phylogenetic in surveillance of SARS-CoV-2 and contribute a phylogeographic framework for studying the emergence of other infectious pathogens in America.

Regarding importation dynamics, our research found the first BA.5 introductions into America to have come from Africa. Such discovery corroborates with the origins of the Omicron lineage in southern Africa (World Health Organization) (Viana et al, 2022), and suggests no intermediary was present in the initial transmission pathway of BA.5 into the U.S. Even after subsampling by global regions, our results saw the majority of foreign introductions hailing from Europe. Moreover, the latter half of the emergence period saw high proportion of European and Asian introductions of BA.5, and we theorized that this finding was proportional to the rates of travel into America. That our phylogeographic results still produced an African ancestry for initial BA.5, despite biases in sequencing and travel rates (Abbasi, 2021) (Anderson et al, 2022), reaffirms the hypothesis of original BA.5 in America hailing from Africa (Viana et al, 2022). The results also suggests that novel lineage introduction to the U.S. may be dependent on its place of origins, in addition to travel rates into America. Such theory is significant, given how BA.5 arose after relaxation of most global travel restrictions, and requires additional research that incorporates travel data into variant emergence analysis (Lemey et al, 2020).

Domestic introductions played a significant role in early BA.5 dissemination into the United States. In fact, we found rates of inter-region transmission to exceed those of global importation for all ten HHS regions. The result would imply widespread secondary transmission across the U.S. following the initial foreign introduction. It also corroborates past findings of SARS-CoV-2 transmission driven by domestic dynamics prior to and during federal travel restrictions (Fauver et al, 2020) (Alpert et al, 2021). For our research, we further managed to discern the different relative rates of inter-region introduction, giving us a more fine-grained look in the interactions between the individual U.S. regions. That early BA.5 spread within America was defined by neighboring effect is tangent to cross-border transmission dynamic between states during the pandemic, described by McBroome et al (2022). Moreover, deviation from the neighboring effect, seen in high cross-country interactions between Region 4 (California and other southwestern states) and Region 9 (Florida and other southeastern states), underlies the role of key hotspots in promoting BA.5 emergence. In discussing this long distance spread, Lerney et al (2014) applied phylogeographic analysis of influenza H3N2 and concluded global air travel as a key determinant global spread, with geographic proximity impacting local transmission. Our study of SARS-CoV-2, also a respiratory pathogen, appears to support such dynamics, although air travel data is again necessary for confirmation.

It is noteworthy that our results had Region 9 and Region 4 feature more prominently in BA.5 introduction than Region 2, which consists of New York City. In past research, Alpert et al (2021) found New York City to be the highest recipient of Alpha variant introductions from the United Kingdom into America, followed by California and Florida. Our results also revealed Region 2 as among the top three recipients of BA.5 foreign introductions (Fig. 4B), but the region fell behind Region 4 and Region 9 in terms of domestic introduction rate (Fig. 5C). That the region surrounding New York was more similar to that around Chicago and Texas, than regions with California or Florida, may suggest different mechanism underlying domestic variant introduction in comparison with foreign introduction. To a

lesser extent, the result may suggest changing emergence dynamics between SARS-CoV-2 variants over time, as the virus evolved and epidemiological landscape changed. In their methodology, Alpert et al (2021) obtained importation risk estimate through airport travel data. One possible alternative explanation is lack of travel data contributing to the under-representation of Region 2.

All in all, our research distinguished itself by proving heterogeneity in emergence, with multiple hotspots interacting to form most introduction and exportation events within America. We theorized that the four primary hotspots (Regions 4, 9, 2, and 5) were representative of the key urban centers within them: Miami, Los Angeles – San Francisco, New York City, and Chicago. These findings fit with Tsui et al (2022)'s description of early viral lineage movements between larger cities, followed by spatial expansion into nearby areas. We thus believe that future models that could identify these hubs of introduction, as well as their epidemiological interactions, would have important roles in predicting the dynamics of future variant emergence. Whether the prominence of such interaction is related to high population, presence of large urban centers, proxy to large airports, state policymaking, or differing regional behavior deserves additional investigation (McBroome et al, 2022) (Tsui et al, 2022) (Cuadros et al, 2021) (Alpert et al, 2021) (Zang et al, 2021).

Our research methods suffered from several limitations. First, while we ensured the 5 percent sequence-to-incidence ratio necessary for credible detection of emerging variants (Vavrek et al, 2021) (Alpert et al, 2021) through comparison with nowcasted SARS-CoV-2 case data from *covidestim* (Chitwood et al, 2021), our final dataset remain skewed (Supplementary Material). Specifically, 66.6 percent of our Global sequences were from Europe, causing our phylogenetic tree to have a European ancestral spine (Fig. 3A). Similar imbalance existed in the USA dataset, where four HHS regions (Regions 2, 4, 6, and 9) constituted the majority of sequences. These numbers reflected the broader inequality in genomic surveillance and GISAID submissions, both around the world and within America (Brito et al, 2022) (Abbasi, 2021). The research attempted to minimize these biases through subsampling and

categorization into broader continents and HHS regions. To this end, however, our categorization of the United States into ten areas was a coarse one, based on Department of Health and Human Services definition (DOH). As such, while the states within one region may share geographic proximity, the choice of states within one region was made seemingly without demographic, epidemiological, or other socio-political reasons. Furthermore, categorization into larger region may have covered for sequencing bias but introduced with it residual confounding, as we would be unable to gauge inter-state introduction events. Finally, our definition of the variant emergence phase was made based on the frequency growth curve with the aim to filter for early BA.5 sequences, which we deemed more important to our research purposes, and might not properly reflect true emergence time for the novel variant.

In conclusion, we have presented the first ever large-scaled phylogeographical analysis of the emergence of Omicron BA.5 variant in United States region in 2022. As countries lift pandemic restrictions and the general population largely immunized, the epidemiological landscape presents opportunities for positive selection of novel SARS-CoV-2 variants (Hou et al, 2022) (Otto et al, 2021). Understanding the different dynamics of introduction in U.S. regions is important for timely and cost-effective policymaking in part of health authorities. We are confident that our methods may extend beyond SARS-CoV-2 and form a framework for phylogeographic analysis of large datasets to discern the spatiotemporal spread of other novel pathogens.

References

1. Center for Disease Control and Prevention. COVID Data Tracker. https://covid.cdc.gov/covid-data-tracker/#datatracker-home
2. World Health Organization. Coronavirus disease (COVID-19) Weekly Epidemiological Updates and Monthly Operational Updates. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
3. Our World in Data. COVID-19: International and Domestic Travel. https://ourworldindata.org/covid-international-domestic-travel
4. Abbasi J. How the US Failed to Prioritize SARS-CoV-2 Variant Surveillance. JAMA. 2021;325(14):1380-1382. doi:10.1001/jama.2021.3368
5. Alpert T, Brito AF, Lasek-Nesselquist E, et al. Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States. Cell. 2021;184(10):2595-2604.e13. doi:10.1016/j.cell.2021.03.061
6. Attwood SW, Hill SC, Aanensen DM, Connor TR, Pybus OG. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. Nat Rev Genet. 2022;23(9):547-562. doi:10.1038/s41576-022-00483-8
7. Brito AF, Semenova E, Dudas G, et al. Global disparities in SARS-CoV-2 genomic surveillance. Preprint. medRxiv. 2021;2021.08.21.21262393. Published 2021 Dec 9. doi:10.1101/2021.08.21.21262393
8. Chaguza C, Coppi A, Earnest R, et al. Rapid emergence of SARS-CoV-2 Omicron variant is associated with an infection advantage over Delta in vaccinated persons. Med (N Y). 2022;3(5):325-334.e4. doi:10.1016/j.medj.2022.03.010

9. Cuadros DF, Branscum AJ, Mukandavire Z, Miller FD, MacKinnon N. Dynamics of the COVID-19 epidemic in urban and rural areas in the United States. Ann Epidemiol. 2021;59:16-20. doi:10.1016/j.annepidem.2021.04.007
10. Dellicour S, Durkin K, Hong SL, Vanmechelen B, Martí-Carreras J, Gill MS, Meex C, Bontems S, André E, Gilbert M, Walker C, Maio N, Faria NR, Hadfield J, Hayette MP, Bours V, Wawina-Bokalanga T, Artesi M, Baele G, Maes P. A Phylodynamic Workflow to Rapidly Gain Insights into the Dispersal History and Dynamics of SARS-CoV-2 Lineages. Mol Biol Evol. 2021 Apr 13;38(4):1608-1613. doi: 10.1093/molbev/msaa284. PMID: 33316043; PMCID: PMC7665608.
11. Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. Bayesian inference of ancestral dates on bacterial phylogenetic trees. Nucleic Acids Res. 2018 Dec 14;46(22):e134. doi: 10.1093/nar/gky783. PMID: 30184106; PMCID: PMC6294524.
12. Douglas J, Winter D, McNeill A, et al. Tracing the international arrivals of SARS-CoV-2 Omicron variants after Aotearoa New Zealand reopened its border. Nat Commun. 2022;13(1):6484. Published 2022 Oct 29. doi:10.1038/s41467-022-34186-9
13. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics. 2002;161(3):1307-1320. doi:10.1093/genetics/161.3.1307
14. du Plessis L, McCrone JT, Zarebski AE, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. 2021;371(6530):708-712. doi:10.1126/science.abf2946
15. Fauver JR, Petrone ME, Hodcroft EB, et al. Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. Cell. 2020;181(5):990-996.e5. doi:10.1016/j.cell.2020.04.021
16. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-Based model for multiple loci. Mol Biol Evol. 2013

Mar;30(3):713-24. doi: 10.1093/molbev/mss265. Epub 2012 Nov 22. PMID: 23180580; PMCID: PMC3563973.

17. Hill V, Baele G. Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model [published online ahead of print, 2019 Jul 31]. Mol Biol Evol. 2019;36(11):2620-2628. doi:10.1093/molbev/msz172
18. Hill V, Githinji G, Vogels CBF, et al. Towards a global virus genomic surveillance network [published online ahead of print, 2023 Mar 6]. Cell Host Microbe. 2023;doi:10.1016/j.chom.2023.03.003
19. Hou Y, Zhao S, Liu Q, Zhang X, Sha T, Su Y, Zhao W, Bao Y, Xue Y, Chen H. Ongoing Positive Selection Drives the Evolution of SARS-CoV-2 Genomes. Genomics Proteomics Bioinformatics. 2022 Jun 26:S1672-0229(22)00079-1. doi: 10.1016/j.gpb.2022.05.009. Epub ahead of print. PMID: 35760317; PMCID: PMC9233880.
20. Lemey P, Hong SL, Hill V, et al. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. Nat Commun. 2020;11(1):5110. Published 2020 Oct 9. doi:10.1038/s41467-020-18877-9
21. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, Russell CA, Smith DJ, Pybus OG, Brockmann D, Suchard MA. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 2014 Feb 20;10(2):e1003932. doi: 10.1371/journal.ppat.1003932. PMID: 24586153; PMCID: PMC3930559.
22. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5(9):e1000520. doi:10.1371/journal.pcbi.1000520
23. L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies.. Mol. Biol. Evol., 32:268-274. https://doi.org/10.1093/molbev/msu300

24. McBroome J, Martin J, de Bernardi Schneider A, Turakhia Y, Corbett-Detig R. Identifying SARS-CoV-2 regional introductions and transmission clusters in real time. Virus Evol. 2022;8(1):veac048. Published 2022 Jun 16. doi:10.1093/ve/veac048
25. Otto SP, Day T, Arino J, Colijn C, Dushoff J, Li M, Mechai S, Van Domselaar G, Wu J, Earn DJD, Ogden NH. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. Curr Biol. 2021 Jul 26;31(14):R918-R929. doi: 10.1016/j.cub.2021.06.049. Epub 2021 Jun 23. PMID: 34314723; PMCID: PMC8220957.
26. Pipes L, Wang H, Huelsenbeck JP, Nielsen R. Assessing Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny. Mol Biol Evol. 2021;38(4):1537-1543. doi:10.1093/molbev/msaa316
27. Rambaut A, Drummond AJ, Xie D, Baele G and Suchard MA (2018) Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. Systematic Biology. syy032. doi:10.1093/sysbio/syy032
28. Rambaut, A., Holmes, E.C., O'Toole, Á. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 5, 1403–1407 (2020). https://doi.org/10.1038/s41564-020-0770-5
29. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018 Jan 8;4(1):vex042. doi: 10.1093/ve/vex042. PMID: 29340210; PMCID: PMC5758920.
30. Shrestha LB, Foster C, Rawlinson W, Tedla N, Bull RA. Evolution of the SARS-CoV-2 omicron variants BA.1 to BA.5: Implications for immune escape and transmission. Rev Med Virol. 2022;32(5):e2381. doi:10.1002/rmv.2381
31. Snake-flu Github, jclusterfunk. https://github.com/snake-flu/jclusterfunk
32. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ & Rambaut A (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10 Virus Evolution 4, vey016. DOI:10.1093/ve/vey016

33. Tegally H, Moir M, Everatt J, et al. Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. Nat Med. 2022;28(9):1785-1790. doi:10.1038/s41591-022-01911-2
34. Tian D, Sun Y, Xu H, Ye Q. The emergence and epidemic characteristics of the highly mutated SARS-CoV-2 Omicron variant. J Med Virol. 2022;94(6):2376-2383. doi:10.1002/jmv.27643
35. Tuekprakhon A, Nutalai R, Dijokaite-Guraliuc A, et al. Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1 serum. Cell. 2022;185(14):2422-2433.e13. doi:10.1016/j.cell.2022.06.005
36. Vavrek D, Speroni L, Curnow K, Oberholzer M, Moeder V, Febbo P. Genomic surveillance at scale is required to detect newly emerging strains at an early timepoint. medRxiv. 2021.01.12.21249613. doi: 10.1101/2021.01.12.21249613
37. Viana R, Moyo S, Amoako DG, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. Nature. 2022;603(7902):679-686. doi:10.1038/s41586-022-04411-y
38. Volz E, Mishra S, Chand M, et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature. 2021;593(7858):266-269. doi:10.1038/s41586-021-03470-x
39. Zang E, West J, Kim N, Pao C. U.S. regional differences in physical distancing: Evaluating racial and socioeconomic divides during the COVID-19 pandemic. PLoS One. 2021;16(11):e0259665. Published 2021 Nov 30. doi:10.1371/journal.pone.0259665