medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Phylogeographic Inference of SARS-CoV-2 Delta Wave in Texas, USA using a Novel Spatial Transmission Count Statistic

Leke Lyu[a], Gabriella Veytsel[a], Guppy Stott[a], Spencer Fox[b], Cody Dailey[a], Lambodhar Damodaran[c], Kayo Fujimoto[d], Jacky Kuo[d], Pamela Brown[e], Roger Sealy[e], Armand Brown[e], Magdy Alabady[f], Justin Bahl[a]*

a. Institute of Bioinformatics, Department of Infectious Diseases, Department of Epidemiology and Biostatistics, Center for Ecology of Infectious Diseases, Center for Applied Pathogen Epidemiology and Outbreak Response, University of Georgia, Athens, GA, USA

b. Institute of Bioinformatics, Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA, USA

c. Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, USA

d. Department of Health Promotion and Behavioral Sciences, The University of Texas Health Science Center at Houston, Houston, TX, USA

e. Division of Disease Prevention and Control, Houston Health Department, Houston, TX, USA

f. Georgia Genomics and Bioinformatics Center, University of Georgia, Athens, GA, USA

Email: justin.bahl@uga.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Viral genomes contain records of geographic movements and cross-scale transmission dynamics. However, the impact of population heterogeneity, particularly among rural and urban areas, on viral spread and epidemic trajectory has been less explored due to limited data availability. Intensive and widespread efforts to collect and sequence SARS-CoV-2 viral samples have enabled the development of comparative genomic approaches to reconstruct spatial transmission history and understand viral transmission across different scales. Large genomic datasets with few mutations present challenges for traditional phylodynamic approaches. To address this issue, we propose a novel spatial transmission count statistic that efficiently summarizes the geographic transmission patterns imprinted on viral phylogenies. Our analysis pipeline reconstructs a time-scaled phylogeny with ancestral trait states and identifies spatial transmission linkages, categorized as imports, local transmission, and exports. These linkages are summarized to represent the epidemic profile of the focal area. We demonstrate the utility of this approach for near real-time outbreak analysis using over 12,000 full genomes and linked epidemiological data to investigate the spread of the SARS-CoV-2 Delta variant in Texas. Our goal is to trace the Delta variant's origin, timing and to understand the role of urban and rural areas in the spatial diffusion patterns observed in Texas. Our study shows (1) highly populated urban centers were the main sources of the epidemic in Texas; (2) the outbreaks in urban centers were connected to the global epidemic; and (3) outbreaks in urban centers were locally maintained, while epidemics in rural areas were driven by repeated introductions.

Significance Statement

We developed a novel phylogeographic approach that analyzes transmission patterns at low computational cost. This method not only facilitates the inference of spatial scales of transmission but also enables exploration of how specific demographic characteristics influence transmission patterns among heterogenous populations. The rural population in the US, comprising approximately 60 million individuals, has been significantly impacted by COVID-19. Applying our new method, we examined the variations in epidemic patterns between urban centers (e.g., Houston) and rural areas in Texas. We found that urban centers are the primary source for SARS-CoV-2 in rural areas. This analysis lays the groundwork for designing effective public health interventions specifically tailored to the needs of affected areas.

Main Text

Introduction

Genomic epidemiology is a field that utilizes pathogen genomes to study the spread of infectious diseases through populations (1). This approach has become increasingly popular due to the decreasing cost of genomic sequencing combined with increasing computational power. During the COVID-19 pandemic, increased number of countries started generating genomic data to inform public health responses (2). The Global Initiative on Sharing All Influenza Data (GISAID) (3) expanded to accommodate these novel data and now maintains the world's largest database of SARS-CoV-2 sequences. As of December 2023, over 16 million sequences, sampled from over 200 countries/regions, have been submitted and archived. Such a vast and diverse dataset enables researchers and public health officials to identify key mutations (4, 5) and track the emergence of variants of interest (VOIs) or variants of concern (VOCs). Additionally, this wealth of genomic information creates opportunities to uncover the hidden characteristics of the pandemic, such as the spatial scales of transmission and the demographic characteristics contributing to transmission patterns. However, effectively handling the complexity of the SARS-CoV-2 genomic dataset requires addressing key challenges, such as developing efficient computational algorithms/pipelines and establishing robust sampling frameworks to draw reliable conclusions.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Viral transmission happens at different spatial scales, encompassing international pandemics, domestic dispersal, and local outbreaks such as those in jails, nursing homes, hospitals, or schools. By mapping how pathogens spread through space and time, evidence-based interventions can be better developed and applied across various scales (6). Within the framework of phylogeographic analysis, sampling locations are assigned as sequence traits, and the ancestral states for traits are estimated on the phylogenetic tree. The well-established software package, Bayesian Evolutionary Analysis Sampling Trees (BEAST) (7), implements discrete (8) and continuous (9) phylogeographic models. Previous studies have used the discrete model to identify the transmission clusters of SARS-CoV-2 introduced in Europe (10), United States (11), Denmark (12) and England (13). Additionally, the continuous model has been applied to elucidate the spatial expansion of SARS-CoV-2 in Belgium (14) and New York City (15). Moreover, the BEAST module can accommodate individual travel history (16) to yield high-accuracy prediction regarding the location of ancestral nodes. Apart from Bayesian analysis, TreeTime (17) applies a maximum likelihood approach to infer the transitions between discrete characters. As a component of the Nextstrain (18) pipeline, this fast analysis enables real-time tracking of pathogens. With the rapid growth in SARS-CoV-2 data, we are now facing extensive phylogenies with thousands of tips. This raises the question: How can we translate the evolutionary changes of geographic traits from such expansive trees into clear epidemiological insights?

The transmission dynamics of SARS-CoV-2 are shaped by host immunity, host movement patterns, and other demographic characteristics (19). For instance, in Chile, people aged under 40 in municipalities with the lowest socioeconomic status had an infection fatality rate 3.1 times higher than those with the highest socioeconomic status (20). The severity of SARS-CoV-2 infection and the risk of mortality increased significantly with age (21). Accordingly, understanding the demographic composition of a population at risk is crucial in determining the potential burden on public health. In the US, rural populations have been particularly vulnerable to COVID-19 complications (22), experiencing higher incidences of disease, mortality, and unemployment. This vulnerability is largely attributed to limited access to healthcare and social services compared to their urban counterparts (23). There have been few studies that investigate the heterogeneity of transmission between rural and urban areas.

In genomic epidemiology, it is critical to analyze sampling biases and develop an appropriate sampling strategy (24). Recent studies have shown that differences in epidemiology and sampling can impact our ability to identify genomic clusters (25). For instance, decreased sampling fraction can lead to the identification of multiple, separate clusters. Moreover, sampling biases can also impact phylogeographic analyses. In the discrete model, if a specific area is overrepresented in the dataset, it may lead to an overrepresentation of the same area at inferred internal nodes (1). In the continuous model, extreme sampling bias might cause the posterior distribution to exclude the true origin location of the root (26).

In this study, we developed a novel phylogeographic pipeline to mitigate genome sampling bias, infer viral phylogenetic relationships, and summarize transmission patterns across multiple scales. To demonstrate the utility of this method, we focused on Texas, aiming to characterize the viral diffusion within the state and compare epidemic trends between urban and rural areas.

Results

Genome Sampling Bias and Subsampling Scheme Adjustments

With support from the Houston Health Department (HHD), we gained access to detailed metadata (zip codes) for 51,229 SARS-CoV-2 genomes sampled in Texas. Out of these, 24,593 were Delta samples (B.1.167.2 and AY*) with high-coverage complete genomes (> 29,000 bp). The metadata in our genome database spans over a thousand distinct zip code records. We subsequently translated these zip code records to their corresponding metropolitan areas to facilitate future phylogeographic analysis. Based on the USDA's Rural-Urban Continuum Code (RUCC), metropolitan areas can be categorized by their population size. Metropolitan areas with over 1 million residents are labeled RUCC-1; those housing between 250,000 and 1 million individuals are assigned RUCC-2; and areas with fewer than 250,000 residents are denoted as RUCC-3 (27). Four major urban centers in Texas - Dallas-Fort Worth, Houston, San Antonio, and Austin - are classified as RUCC-1. The detailed list of RUCC classifications for metropolitan areas can be found in Table S1.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

To explore sampling biases in Texas, we calculated the sampling ratio by dividing the count of available genomes by the count of reported cases. We noted significant heterogeneity in sampling ratio across different metropolitan areas from Epi-Week 14 to Epi-Week 43 (Figure S1A). During the Delta wave, Victoria, Wichita Falls, and Bryan-College Station emerged as the top three under-sampled metropolitan areas, while Houston, San Angelo, and Abilene were the most over-sampled. Subsequently, we introduced a subsampling scheme (Figure S1B) to reduce these sampling biases, thereby enhancing the accuracy of our phylogeographic analysis (1, 28). This proportional sampling method establishes a consistent sampling ratio to serve as a baseline for all regions. In cases where regions are under-sampled (sampling ratio below the baseline), all available genomes were retained. Conversely, over-sampled regions (with a sampling ratio exceeding the baseline) were down sampled to align with the baseline rate. By adopting a baseline value of 0.006, we selected 5,899 genomes. Consequently, the variance in sampling ratios (Table S2) across all metropolitan areas dropped substantially from 5.74e-05 to 7.56e-07.

The Transmission Dynamics of Delta Variant in Texas

To investigate the timing of virus importations into Texas and the dynamics of the resulting local transmission lineages, we conducted a comprehensive phylogeographic analysis of 12,285 SARS-CoV-2 Delta genomes sampled from March 27th 2021, to October 24th 2021. Of these genomes, 6,386 served as globally contextual sequences (Table S3), ensuring a balanced 1:1 ratio between Texas and non-Texas samples. We estimated the phylogenetic tree with IQ-TREE (29) and inferred the time-adjusted phylogeny using TreeTime (17). Through the time-calibrated phylogeny (Figure 1C), we inferred that the Delta variant was introduced into Texas as early as late March, preceding the first reported Delta case in Houston in mid-April 2021 (30).

The trait states of internal nodes were identified as either 'Texas' or 'non-Texas' using the 'mugration' model (17) implemented in TreeTime. The observed state changes at each node can be used to characterize three transmission categories: importations, local transmissions, and exportations (Figure 1A). These transmissions could be viewed as spatial transmission links (Figure 1B). Here, we defined the sum of these links as the spatial transmission counts, which, over time, offered a comprehensive summary of the epidemic's profile (Figure 1D). Given that the infectious periods for SARS-CoV-2 typically ranged from day 2 to day 15 post-infection (19), long branches on the phylogeny likely represented multiple transmission events. To reduce uncertainty, we excluded branches with durations exceeding 15 days (10,678 out of 19,841 were removed). The epidemic in Texas was characterized by at least 265 imports and 259 exports connecting statewide cases to the global pandemic. The Texas outbreak was driven by local transmission, with 4,750 local transmission branches. Our results indicated that numerous co-circulating transmission chains were introduced independently.

Characterizing spatial transmission heterogeneity

To understand the spatial transmission of the Delta SARS-CoV-2 wave in Texas, we estimated the ancestral location states (26 location traits: 25 metropolitan areas and 1 combined rural area) (Figure S2) on the phylogeny described above. The spatial transmission counts could be used to identify import and export hubs that determine the patterns of epidemic spread (Figures S3-S26).

To measure the magnitude of viral flow between metropolitan areas in Texas, we utilized the time series of spatial transmission counts derived from the phylogeny to construct a weighted network (Figure 2). In this network, edges carried weight corresponding to the spatial transmission counts connecting two metropolitan areas, without indicating directionality. The inferred network consisted of 25 nodes and 88 edges, with an average degree of 3.52. The centrality analysis, detailed in Table S4, underscored the top five nodes as identified consistently by betweenness and connectedness (31). These pivotal nodes were Dallas–Fort Worth, Houston, San Antonio, Austin, and Brownsville. The top 4 were classified as RUCC-1, which supported the idea that populated urban areas played a crucial role in the viral spread throughout Texas. The border city of Brownsville, classified as RUCC-2, was a well-connected border town, and this classification may not accurately capture its population density.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Community source-sink dynamics

We introduced the Source Sink Score to identify populations as either viral sources or sinks. The Source Sink Score ranged between -1 and 1. A score close to 1 indicates that the number of exports is significantly higher than the number of imports, classifying the population predominantly as a viral source, with very few or no imports. Conversely, a score closer to -1 indicates a predominance of imports over exports, classifying the population as a viral sink.

The metropolitan areas were categorized as sources or sinks based on the accumulated Source Sink Score, with the full list detailed in Table S5. We found that the RUCC-1 group, representing densely populated urban centers, had the highest Source Sink Score, highlighting its role as the major source during the Delta outbreak in Texas (Figure 3). Within the RUCC-1 group, Houston had the highest Source Sink Score at 0.147, followed by Dallas-Fort Worth, San Antonio, and Austin with values of 0.000, -0.101, and -0.363, respectively. In contrast, the rural areas, with a score of -0.711, suggested it primarily functioned as a viral sink.

Epidemic trends in populated urban center compared to rural areas

We introduced the Local Import Score to determine the characteristics of a source population. The Local Import Score can be used to evaluate when an outbreak is locally maintained rather than driven by continual introductions from other regions. The import score ranges from 0 to 1, where a value close to 1 indicates that the outbreak is driven by external introductions. Determining when most new cases are locally acquired is important to inform public health resource allocation, contact tracing efforts, and control strategies in emergency situations.

We calculated the Local Import Score for all metropolitan areas (Table S5). Selecting Houston as a representative city, we examined epidemic trends in populated urban centers and compared them to those in rural areas (Figure 4). The accumulated Local Import Score (0.168) over the entire Delta wave indicated that the outbreak in Houston was locally maintained. Conversely, with an accumulated Local Import Score of 0.634, the epidemic in rural areas relied on external introductions. Our results suggested that while an outbreak may initially rely on introductions from other sources, once the epidemic was established and locally maintained, the region may become a primary source of pathogen spread to other regions.

To assess the impact of global circulation on local-scale transmission patterns, we analyzed the viral flow between non-Texas and urban centers (e.g., Houston) (Figure 4A), as well as between non-Texas and the rural areas (Figure 4B). Introductions from non-Texas accounted for 56% of all imports to Houston and 19% of all exports from Houston went to locations outside of Texas. In comparison, Introductions from non-Texas accounted for 27% of all imports to rural areas, and 12% of rural exports were to locations outside of Texas. These statistics suggest that Houston, as a highly connected and large urban center, was an important hub connecting the outbreak in Texas to the global pandemic.

Discussion

In this study, we introduced a novel spatial transmission count statistic, which characterizes the weekly counts of local spread, viral inflow, and outflow, illustrating transmission trends over time. The Source Sink Score and Local Import Score allow for quantitative comparison of epidemic trends between regions. The Source Sink Score measures net viral exports, weighted by the outbreak size, while the Local Import Score compares the significance of external introductions versus local transmission in shaping the epidemic. To demonstrate the utility of this novel phylogeographic approach, we investigated the geographic diffusion pattern of the Delta variant of SARS-CoV-2. At the state level, our primary questions were when the outbreak in Texas began and the number of introductions that occurred. Within the state of Texas, we highlighted subregions that served as primary viral sources and contrasted the epidemic trends between urban centers and rural areas.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The extraordinary size of our genomic data offers valuable insights into the micro-epidemiological patterns that underlie the COVID-19 epidemic in Texas. Our analysis revealed that cryptic transmissions of the Delta variant began as early as late March, two weeks before the identification of the first Delta case in Houston (30). Additionally, we identified at least 265 imports and 259 exports, highlighting Texas's intensive connection to the global pandemic. Our results indicated that the Delta variant invaded Texas through multiple introductions. This pattern aligns with observations from Connecticut's initial COVID-19 wave (32), the UK's first wave (33), the emergence of B.1.1.7 in the United States (11), and the presence of Omicron BA.1 in England (13). These independent importations subsequently formed massive local transmission clusters in Texas.

Urban centers were the primary sources of the Delta epidemic in Texas. Utilizing the Source Sink Score, we ranked 26 subregions across Texas, categorizing them as viral sources or sinks. The analysis showed that Houston had the highest score, followed by Dallas-Fort Worth and San Antonio, which are the three most populous metropolitan areas in the state. The influence of these urban centers in spreading the Delta epidemic may be attributed to their critical positions in both road and air travel networks. Houston, Dallas-Fort Worth, and San Antonio, connected by Interstates 10, 45, and 35, form the vertices of the Texas Triangle (34), one of 11 megaregions in the US and home to the majority of the Texas's population. This complex connectivity, along with the presence of major airports such as George Bush Intercontinental Airport in Houston (a United Airlines hub), Dallas-Fort Worth International Airport (American Airlines' largest primary hub and headquarters), and San Antonio International Airport (a Southwest Airlines hub), highlights their pivotal role in airway travel. Our analysis underscored the crucial role of urban centers in driving the Delta outbreak. This insight provides valuable information that can guide public health decision-making. In particular, increased control efforts in highly connected urban centers may have a disproportionate impact on connected rural areas.

Rural areas exhibit a lower level of viral flow in relation to global contexts, with epidemics in these areas predominantly relying on external introductions, thus establishing them as viral sinks. Notably, urban centers and rural areas demonstrate distinct transmission patterns (35). In rural areas that are highly affected, implementing social distancing measures is crucial to reduce local spread. It is important to note that our analysis assumes virus transmission in each region is impacted only by population size and density and does not account for the influence of community behavior and beliefs, healthcare disparities, environmental factors, and other factors, on viral transmission. Future studies addressing these aspects will provide more comprehensive insights into the determinants of cross-scale transmission.

The proportional sampling scheme we employed ensured a consistent sampling ratio across all 25 metropolitan areas and 1 combined rural area in Texas. However, we did not have the specific count for Delta cases. Relying on overall SARS-CoV-2 case counts for approximation led to a lower projected sampling rate of the Delta variant, particularly at the beginning or end of the Delta wave. A study in Houston showed that 76.9% of the total sequences were identified as Delta during the study period from March 15, 2021, through September 20, 2021 (30). Unlike earlier variants, the Delta variant exhibited higher transmissibility and a higher rate of vaccine breakthrough cases (35–37), quickly becoming the dominant strain (39). This indicates that under-sampling of the Delta variant at the beginning and end of the outbreak would have a limited impact.

The spatial transmission count statistic represents the time-series of categorized transmission linkages related to the focal regions. Informed by the annotated viral phylogeny, it summarizes the trends of local spread and viral flow at a minimal computational cost. This efficiency allows for real-time surveillance of tens of thousands of viral genomes, which is essential in addressing the challenges posed by the current pandemic or potential future outbreaks. Adopting a simplified model, we assume that transmission events take place along all the branches of the viral phylogeny. However, phylogenetic trees are not equivalent to transmission trees; they do not directly reveal who infected whom (40, 41). As a result, our model may introduce bias in the estimation of local transmission counts. Nonetheless, it does provide insights into cross-scale transmission and epidemic trajectories that could be used to inform control efforts. The ongoing large-scale pathogen genomic surveillance of epidemic outbreaks will allow for the continued development of near real-time inferential methods to inform and improve public health practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Materials and Methods

Surveillance and genetic dataset

Texas comprises 25 metropolitan areas, as defined by the United States Office of Management and Budget (OMB). Any population, housing, and territory not included in a metropolitan area is classified as rural. The Rural-Urban Continuum Codes (RUCC) differentiates MAs based on the population size. Dallas–Fort Worth, Houston, San Antonio, and Austin, all characterized as RUCC-1, are the most populated metropolitan area in Texas.

Historical COVID-19 data of confirmed cases for Texas were accessed through the Texas Department of State Health Services (DSHS) website (42). These reported cases, counted by county, were then aggregated into metropolitan areas. The weekly tracking of new cases guided our proportional sampling strategy. All the scripts that facilitate the sampling scheme have been consolidated into an R package called Subsamplerr. The package takes case count tables and genome metadata as input, enabling the visual investigation of sampling heterogeneity and the implementation of the proportional sampling scheme. It is publicly available at https://github.com/leke-lyu/subsamplerr.

With the support of the Houston Health Department (HHD), we accessed a large dataset of SARS-CoV-2 genomes sampled in Texas: 51,229 genomes with linked metadata. Out of these, 24,593 were of the Delta variant, and 5,899 were sampled proportional to the case count. To investigate the introduction of the Delta variant into Texas, we also sampled worldwide Delta genomes from GISAID as global contextual. We randomly sampled 6,386 genomes from 49 counties. In total, our database consists of 12,285 whole genomes.

Phylogeographic analysis pipeline

The pipeline comprises two major components: (1) phylogenetic reconstruction and (2) spatial transmission linkages' characterization.

Phylogenetic reconstruction: This component aims to generate a time-labeled phylogeny with inferred ancestral trait states which utilizes the Nextstrain pipeline (18). Sequence alignment was conducted using Nextalign (18), while the maximum likelihood tree construction was achieved with IQ-TREE (29), applying a GTR substitution model. TreeTime (17) was employed to produce a time-scaled phylogeny and infer ancestral node states. The phylogeny was rooted using early samples from Wuhan (Wuhan-Hu-1/2019). Its temporal resolution was set based on an assumed nucleotide substitution rate of $8 * 10^{-4}$ substitutions per site per year (default setting of Nextstain build for SARS-CoV-2). Migration patterns between distinct geographic regions were inferred through time-reversible models, mirroring those characterizing genome sequence evolution (17). For a comprehensive understanding of the pipeline's setup and configurations, including Snakemake profiles, visit our GitHub repository at https://github.com/leke-lyu/surveillanceInTexas.

Spatial transmission linkages' characterization: This component utilized custom scripts to identify spatial transmission linkages from the phylogeny and summarize epidemic trends in the focal region. The tree file was imported using the 'treeio' package (43) in R. Following this, the tree was converted into a structured data frame for further analysis, aided by the 'tidytree' package (43). Branches with durations surpassing 15 days were excluded, and the shorter branches on the phylogeny were designated as spatial transmission linkages. By analyzing the trait states, we can determine whether the transmission occurred within the local area, involved an importation from another location, or resulted in exportation to another location. The time series of spatial transmission count, categorized by type, provides an overview of the focal area's epidemic trend. All scripts used in Texas case study are publicly accessible at https://github.com/leke-lyu/transmissionCount.

Metrics that describe transmission pattern

By employing an 'identify-and-count' approach for spatial transmission linkages, we were able to portray the epidemic profile of the area of interest. Different areas possess varying population sizes, levels of population mobility, and immunological characteristics, all of which can contribute to differences in the size and dynamics of the epidemic. To quantitatively compare the characteristics of epidemics in different areas, we introduced two metrics.

To investigate the relative importance of repeated introductions versus continuing local spread, we define the Local Import Score:

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

$$Local\ Import\ Score = \frac{C_t(Import)}{C_t(Import) + C_t(Local\ Trans)}$$

where $C_t(Import)$ represents the count of importations over a specific time period, t, and $C_t(Local\ Trans)$ signifies the count of local transmissions during the same period. The choice of the time window for calculation is contingent on the research objective. It can encompass the entire duration of the epidemic wave to assess cumulative effects, or it might focus on shorter intervals, such as epidemiological weeks, for real-time surveillance. The Local Import Score ranges between 0 and 1. A Local Import Score value approaching 1 indicates a predominant role of importations, whereas a Local Import Score value nearing 0 suggests a dominance of local transmissions, implying that the epidemic is primarily sustained locally.
To identify whether a region acts primarily as a viral source or sink, we introduce the metric called Source Sink Score:

$$Source\ Sink\ Score = \frac{C_t(Export) - C_t(Import)}{C_t(Export) + C_t(Import)}$$

where $C_t(Export)$ represents the count of exportations over a specific time period t, and $C_t(Import)$ denotes the count of importations during that same period. The Source Sink Score ranges between -1 and 1. A Source Sink Score value approaching 1 suggests a dominant role of exportation, indicating that the research region mainly functions as a viral source. Conversely, a Source Sink Score value nearing -1 implies a dominant role of importation, suggesting that the research region predominantly serves as a viral sink.

Phylogenetic-based spatial network inference

To capture the viral flow between metropolitan areas in Texas, we constructed a weighted, undirected network. Each metropolitan area is represented as a node, and the edge carries weight corresponding to the spatial transmission counts. After establishing the network, we conducted the centrality analysis to rank the metropolitan areas based on their betweenness, closeness, and degree centrality. We processed the various network data objects using the 'igraph' package (44) in R. Visualizations were generated with the 'ggplot2' package (45). Additionally, we utilized the 'qgraph' package (46) to compute several node centrality statistics, including edge-betweenness centrality.

Acknowledgments

This work has been funded in part from the National Institute of Allergy and Infectious Diseases, a component of the NIH, Department of Health and Human Services, under contract no. 75N93021C00018 (NIAID Centers of Excellence for Influenza Research and Response, CEIRR) and Centers for Disease Control and Prevention, Department of Health and Human Services, under contracts 75D30121C10133 and NU50CK000626. We acknowledge the GISAID contributors (acknowledgment table of genomes used is provided on our GitHub repository) for sharing genomic data.

References

1. V. Hill, C. Ruis, S. Bajaj, O. G. Pybus, M. U. G. Kraemer, Progress and challenges in virus genomic epidemiology. *Trends in Parasitology* **37**, 1038–1049 (2021).

2. , Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health (July 25, 2023).

3. S. Elbe, G. Buckland□Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33–46 (2017).

4. E. B. Hodcroft, *et al.*, Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* **591**, 30–33 (2021).

5. N. D. Grubaugh, *et al.*, Tracking virus outbreaks in the twenty-first century. *Nat Microbiol* **4**, 10–19 (2019).

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

6. S. W. Attwood, S. C. Hill, D. M. Aanensen, T. R. Connor, O. G. Pybus, Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nat Rev Genet* **23**, 547–562 (2022).

7. M. A. Suchard, *et al.*, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, vey016 (2018).

8. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian Phylogeography Finds Its Roots. *PLOS Computational Biology* **5**, e1000520 (2009).

9. P. Lemey, A. Rambaut, J. J. Welch, M. A. Suchard, Phylogeography Takes a Relaxed Random Walk in Continuous Space and Time. *Mol Biol Evol* **27**, 1877–1885 (2010).

10. P. Lemey, *et al.*, Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature* **595**, 713–717 (2021).

11. T. Alpert, *et al.*, Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States. *Cell* **184**, 2595-2604.e13 (2021).

12. T. Y. Michaelsen, *et al.*, Introduction and transmission of SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark. *Genome Medicine* **14**, 47 (2022).

13. J. L.-H. Tsui, *et al.*, Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1. *Science* **381**, 336–343 (2023).

14. S. Dellicour, *et al.*, A Phylodynamic Workflow to Rapidly Gain Insights into the Dispersal History and Dynamics of SARS-CoV-2 Lineages. *Molecular Biology and Evolution* **38**, 1608–1613 (2021).

15. S. Dellicour, *et al.*, Variant-specific introduction and dispersal dynamics of SARS-CoV-2 in New York City – from Alpha to Omicron. *PLOS Pathogens* **19**, e1011348 (2023).

16. P. Lemey, *et al.*, Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nat Commun* **11**, 5110 (2020).

17. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution* **4**, vex042 (2018).

18. J. Hadfield, *et al.*, Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

19. P. V. Markov, *et al.*, The evolution of SARS-CoV-2. *Nat Rev Microbiol* **21**, 361–379 (2023).

20. G. E. Mena, *et al.*, Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile. *Science* **372**, eabg5298 (2021).

21. M. O'Driscoll, *et al.*, Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590**, 140–145 (2021).

22. , USDA ERS - Rural Residents Appear to be More Vulnerable to Serious Infection or Death From Coronavirus COVID-19 (August 2, 2023).

23. J. T. Mueller, *et al.*, Impacts of the COVID-19 pandemic on rural America. *Proc Natl Acad Sci U S A* **118**, 2019378118 (2021).

24. R. P. D. Inward, K. V. Parag, N. R. Faria, Using multiple sampling strategies to estimate SARS-CoV-2 epidemiological parameters from genomic sequencing data. *Nat Commun* **13**, 5587 (2022).

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

25. B. Sobkowiak, *et al.*, The utility of SARS-CoV-2 genomic data for informative clustering under different epidemiological scenarios and sampling. *Infection, Genetics and Evolution* **113**, 105484 (2023).

26. A. Kalkauskas, *et al.*, Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLOS Computational Biology* **17**, e1008561 (2021).

27. , USDA ERS - Rural-Urban Continuum Codes (September 7, 2023).

28. S. D. W. Frost, *et al.*, Eight challenges in phylodynamic inference. *Epidemics* **10**, 88–92 (2015).

29. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**, 268–274 (2015).

30. P. A. Christensen, *et al.*, Delta Variants of SARS-CoV-2 Cause Significantly Increased Vaccine Breakthrough COVID-19 Cases in Houston, Texas. *Am J Pathol* **192**, 320–331 (2022).

31. L. C. Freeman, Centrality in social networks conceptual clarification. *Social Networks* **1**, 215–239 (1978).

32. J. R. Fauver, *et al.*, Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181**, 990-996.e5 (2020).

33. , Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK | Science (September 12, 2023).

34. Y. Hagler, Defining U.S. Megaregions.

35. B. D. Dalziel, *et al.*, Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. *Science* **362**, 75–79 (2018).

36. V. Servellita, *et al.*, Neutralizing immunity in vaccine breakthrough infections from the SARS-CoV-2 Omicron and Delta variants. *Cell* **185**, 1539-1548.e5 (2022).

37. E. Hacisuleyman, *et al.*, Vaccine Breakthrough Infections with SARS-CoV-2 Variants. *N Engl J Med* **384**, 2212–2218 (2021).

38. T. Kustin, *et al.*, Evidence for increased breakthrough rates of SARS-CoV-2 variants of concern in BNT162b2-mRNA-vaccinated individuals. *Nat Med* **27**, 1379–1384 (2021).

39. R. Earnest, *et al.*, Comparative transmissibility of SARS-CoV-2 variants Delta and Alpha in New England, USA. *medRxiv*, 2021.10.06.21264641 (2021).

40. M. D. Hall, C. Colijn, Transmission Trees on a Known Pathogen Phylogeny: Enumeration and Sampling. *Mol Biol Evol* **36**, 1333–1343 (2019).

41. X. Didelot, C. Fraser, J. Gardy, C. Colijn, Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol Biol Evol*, msw075 (2017).

42. , COVID-19 (Coronavirus Disease 2019) | Texas DSHS (September 10, 2023).

43. G. Yu, *Data Integration, Manipulation and Visualization of Phylogenetic Trees* (CRC Press, 2022).

44. G. Csardi, T. Nepusz, The Igraph Software Package for Complex Network Research. *InterJournal* **Complex Systems**, 1695 (2005).

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

45. P. M. Valero-Mora, ggplot2: Elegant Graphics for Data Analysis. *Journal of Statistical Software* **35**, 1–3 (2010).

46. S. Epskamp, A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, D. Borsboom, qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software* **48**, 1–18 (2012).

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figures and Tables

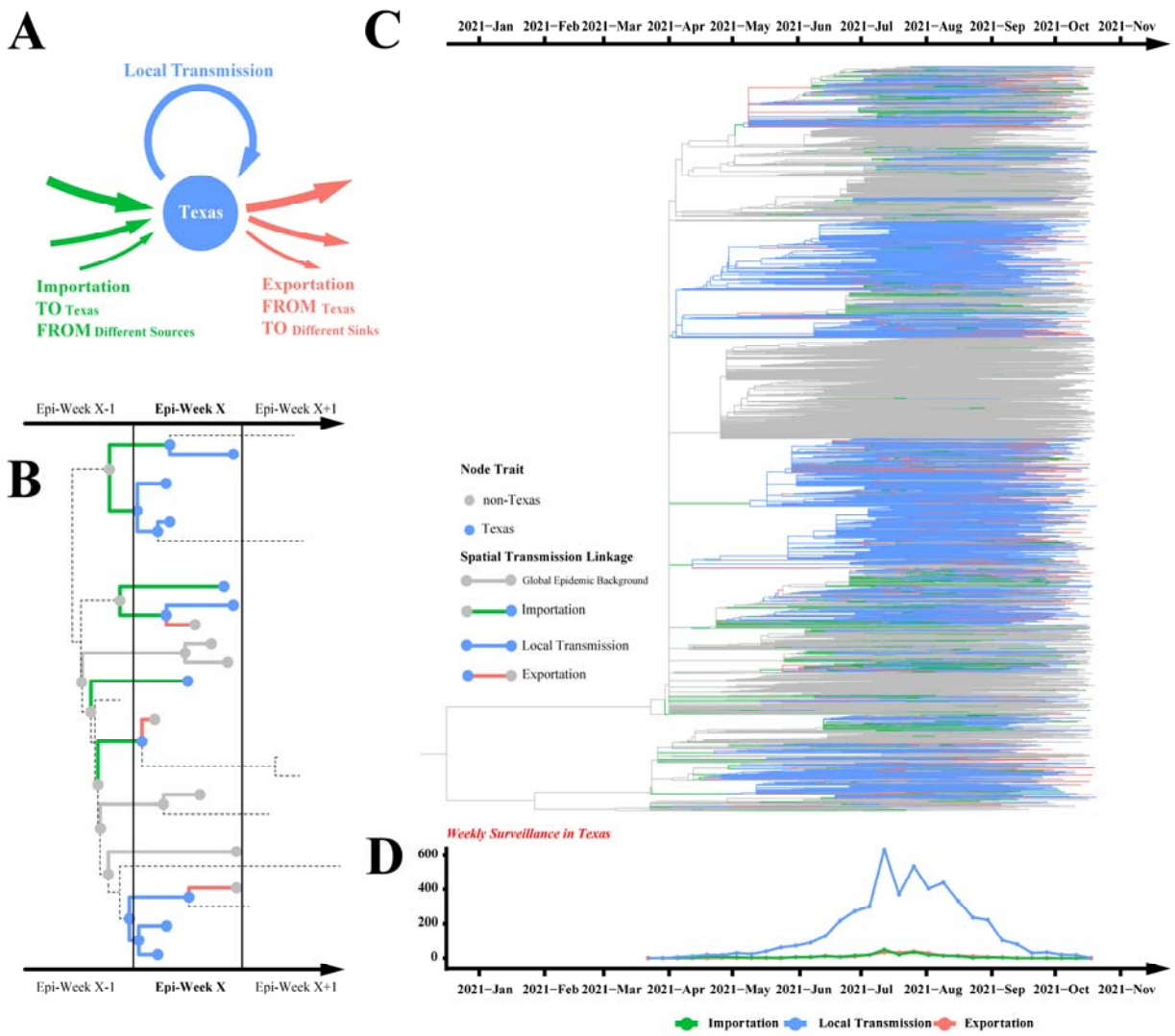


Figure 1. The spatial transmission count statistic investigates the transmission dynamics of SARS-CoV-2 Delta variant in Texas A. Texas-related transmissions can be classified into three categories: importation, local transmission, and exportation. Importation can have multiple sources, while exportation can have multiple sinks. **B.** The schematic tree depicts a total of 18 Texas-related spatial transmission linkages in Epi-Week X: 6 imports, 9 local transmissions, and 3 exports. **C.** In the time-adjusted phylogeny, branches are colored based on the categories of the corresponding spatial transmission linkages. **D.** The time series of spatial transmission counts summaries the epidemic trend in Texas.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

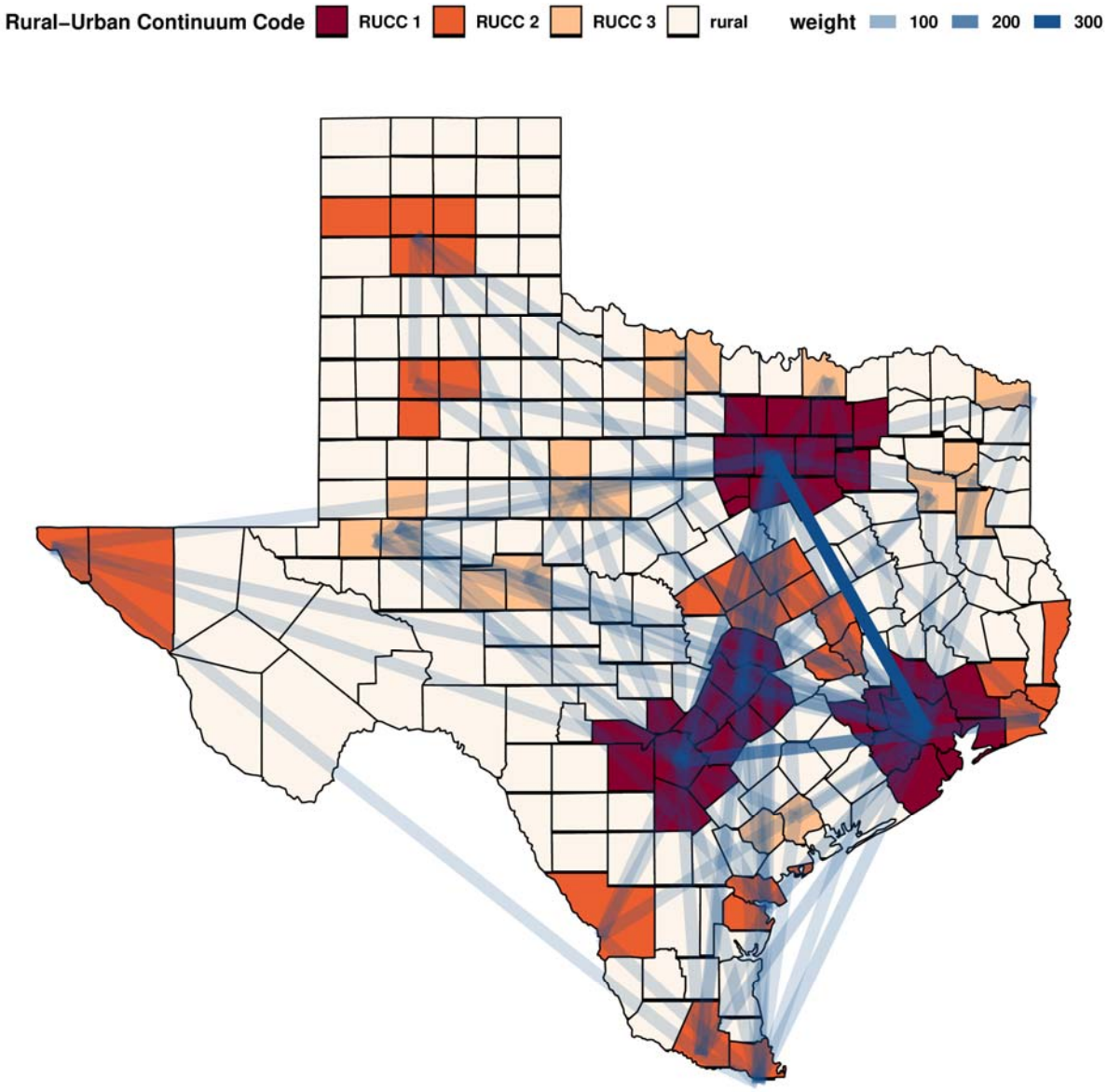


Figure 2. Spatial Network of metropolitan areas in Texas. This visualization overlays the network on a map of Texas. Each node represents a metropolitan area, and the transparency of the blue lines indicates the magnitude of viral flow between the respective metropolitan areas.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

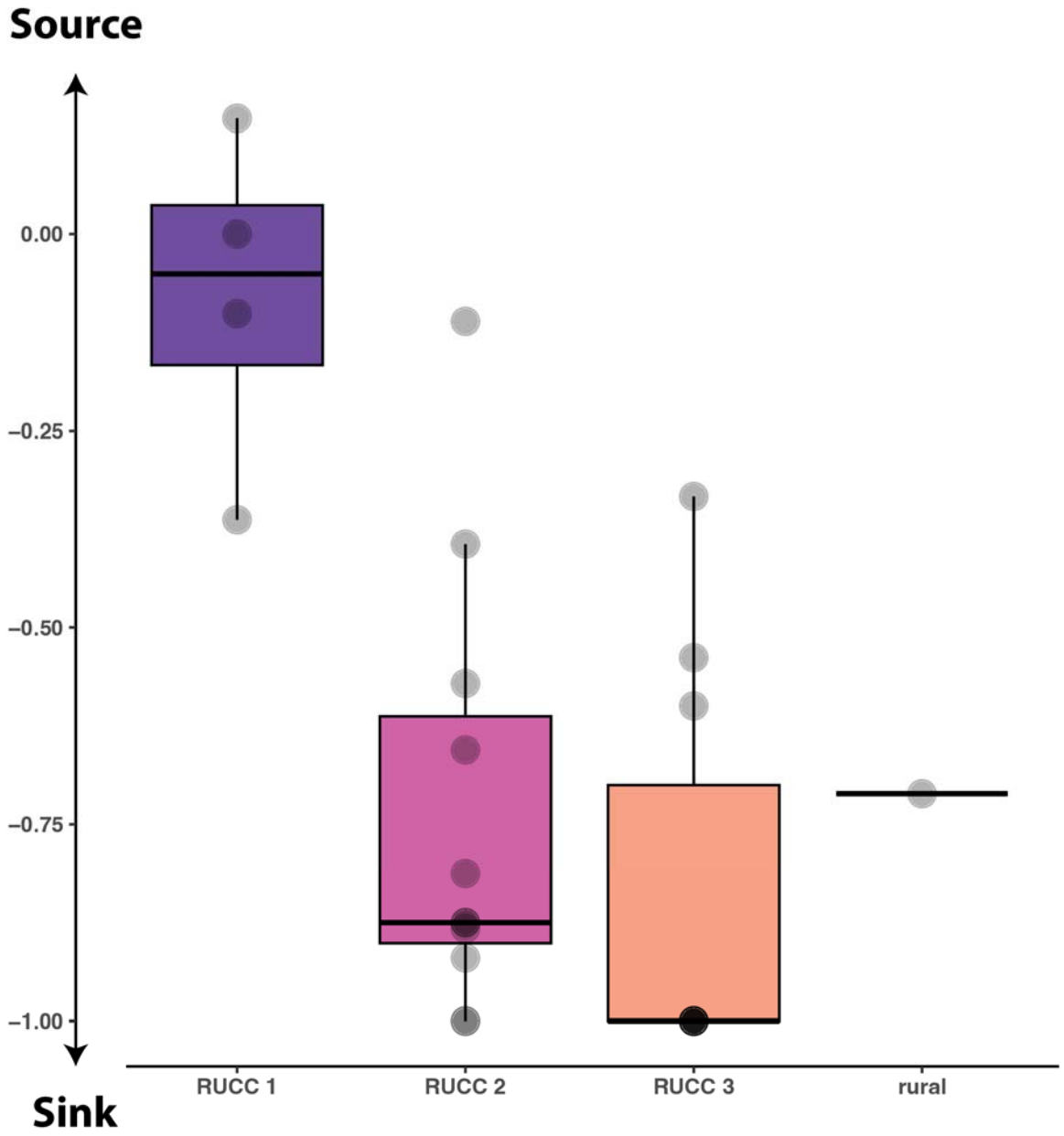


Figure 3. The Source Sink Score identifies the source hubs of Delta variant's spread in Texas. All location traits are categorized by RUCC code. The RUCC-1 group, representing the most populated urban centers, has the highest score.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

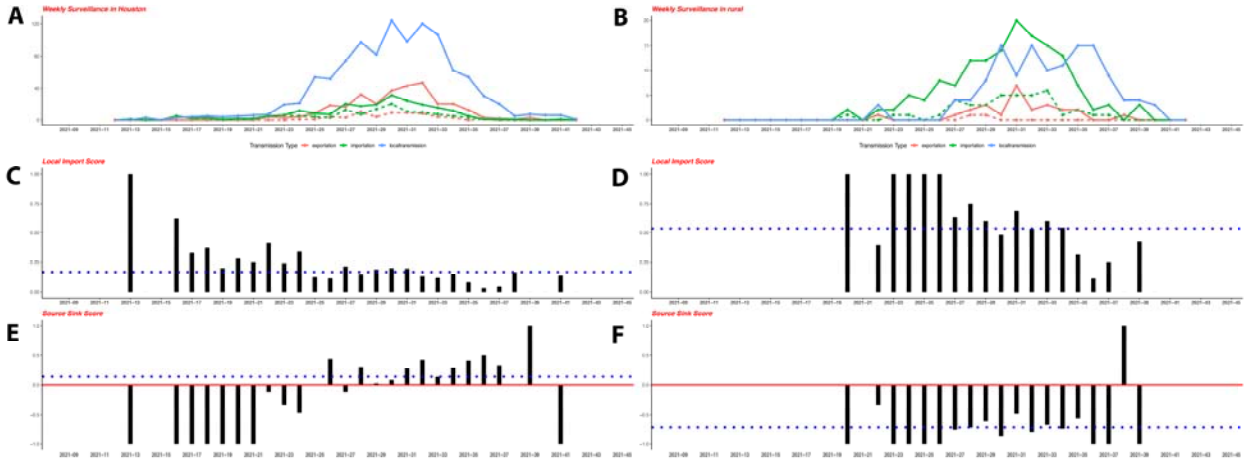
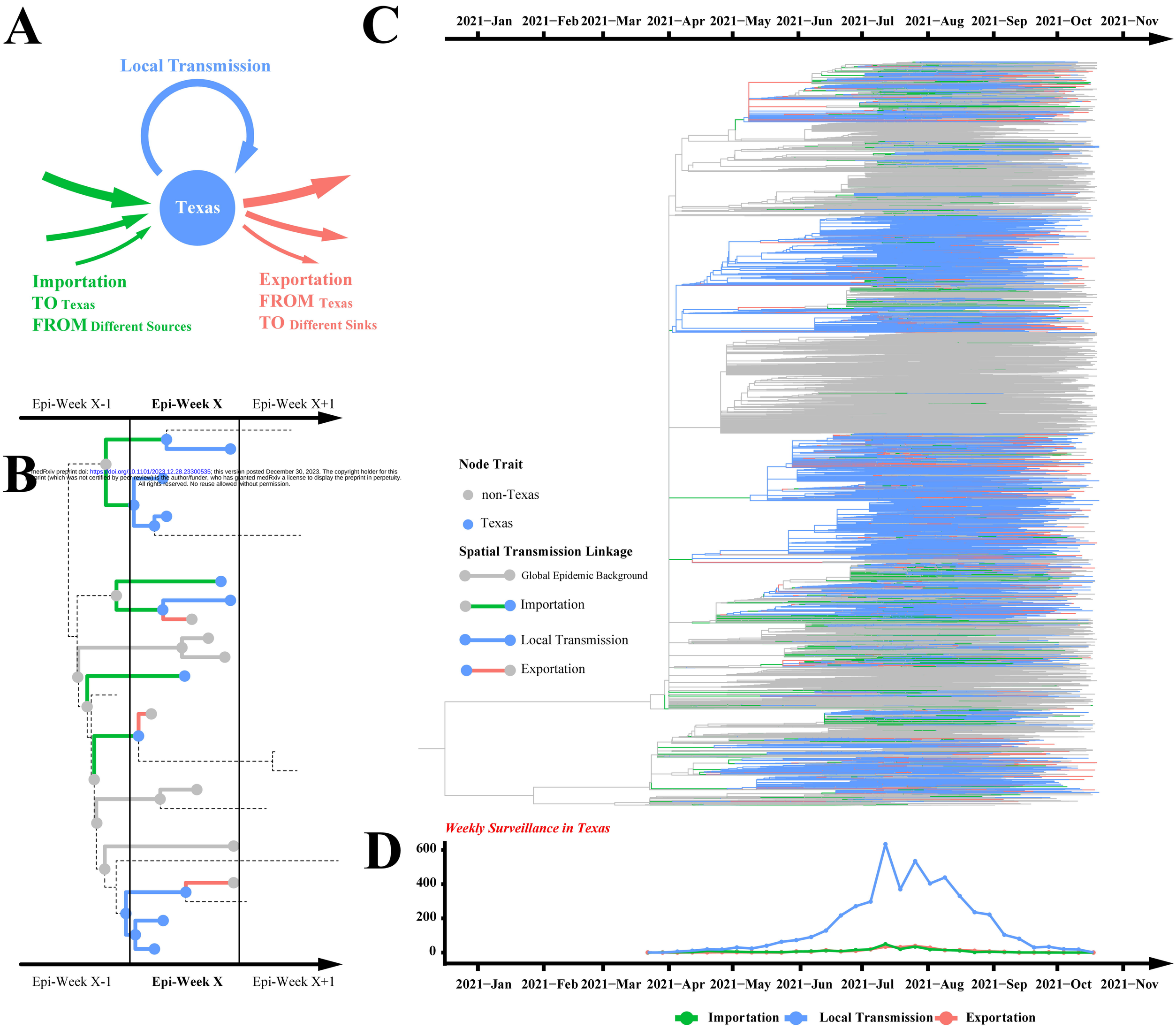
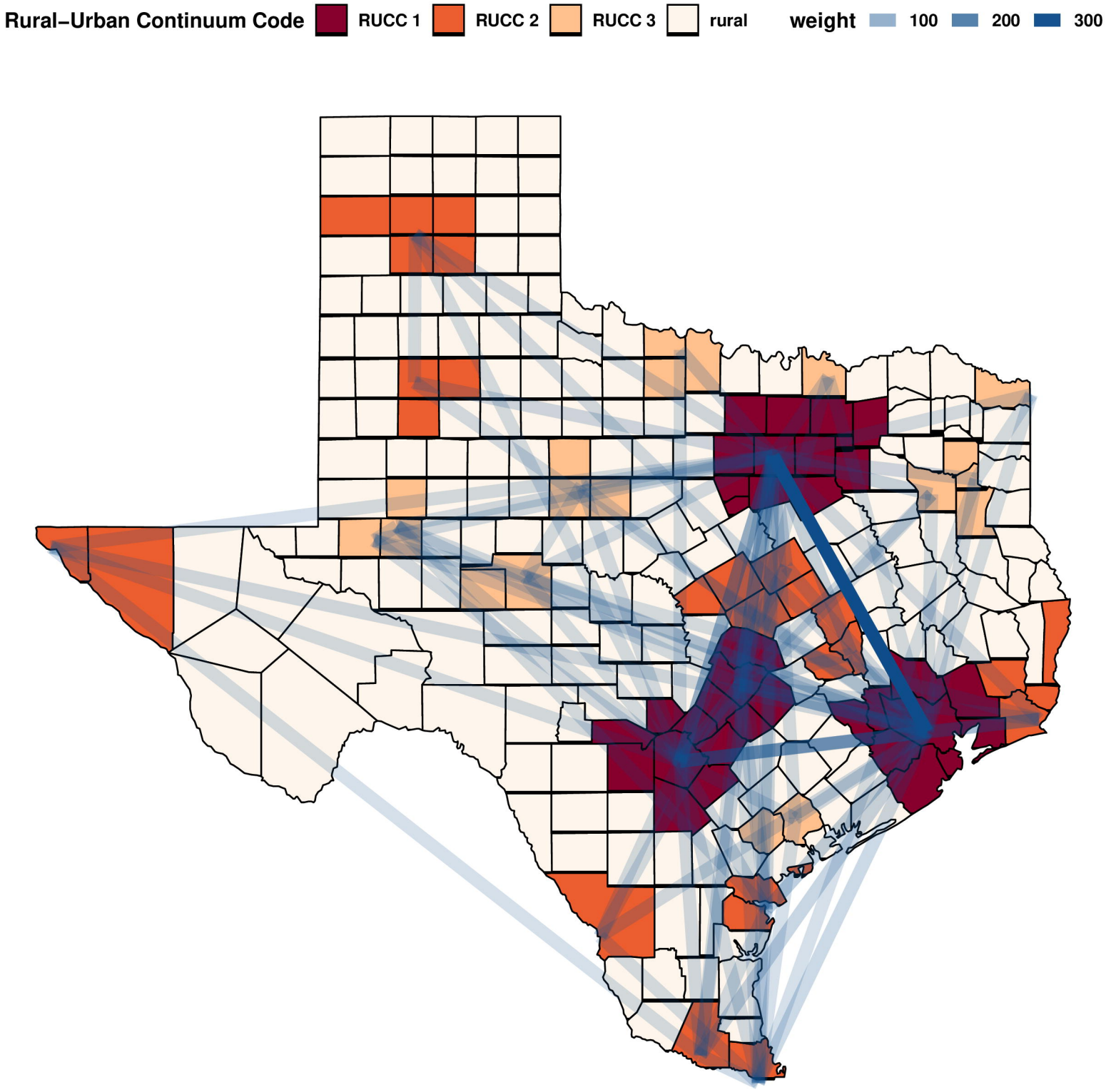
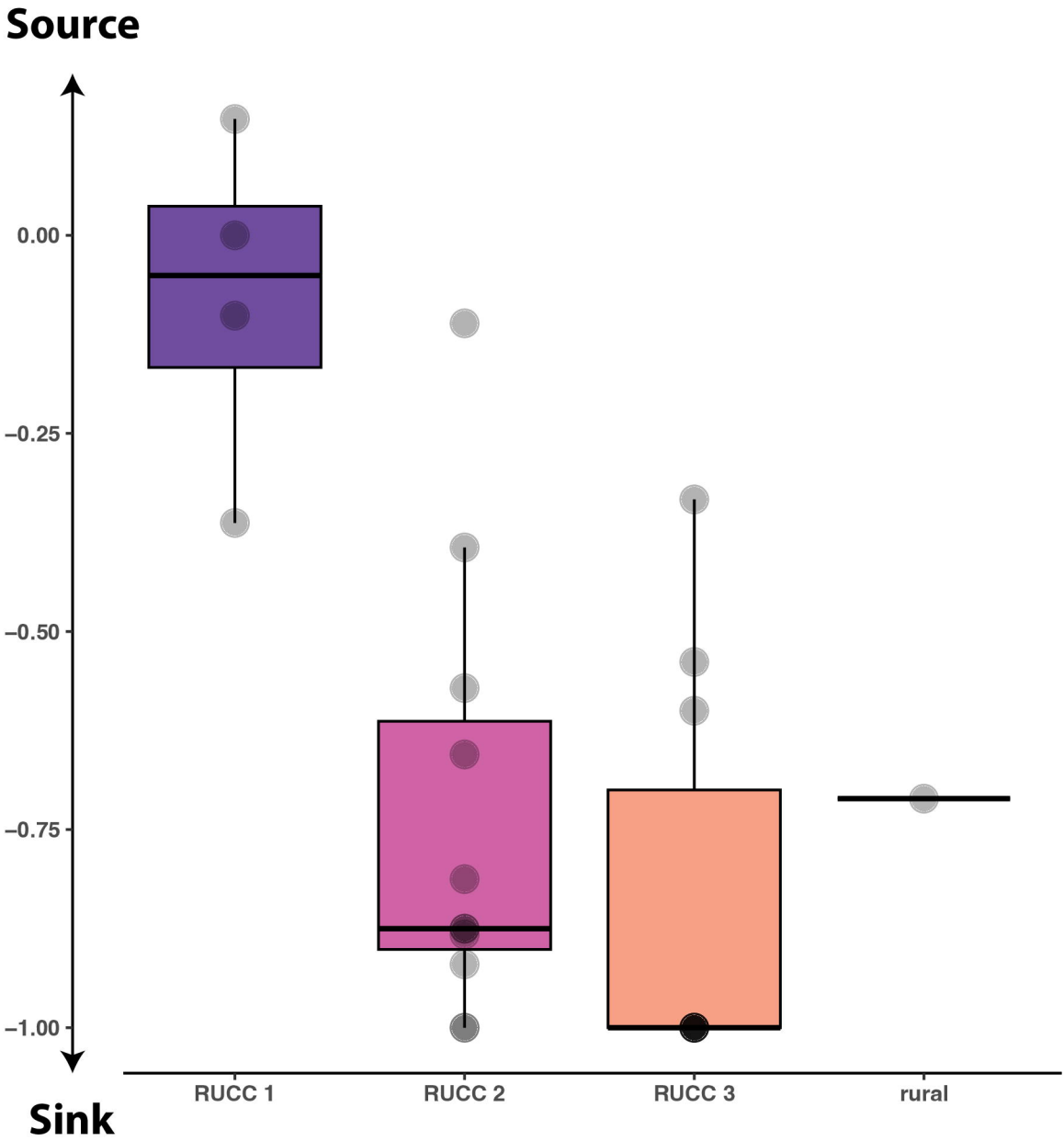


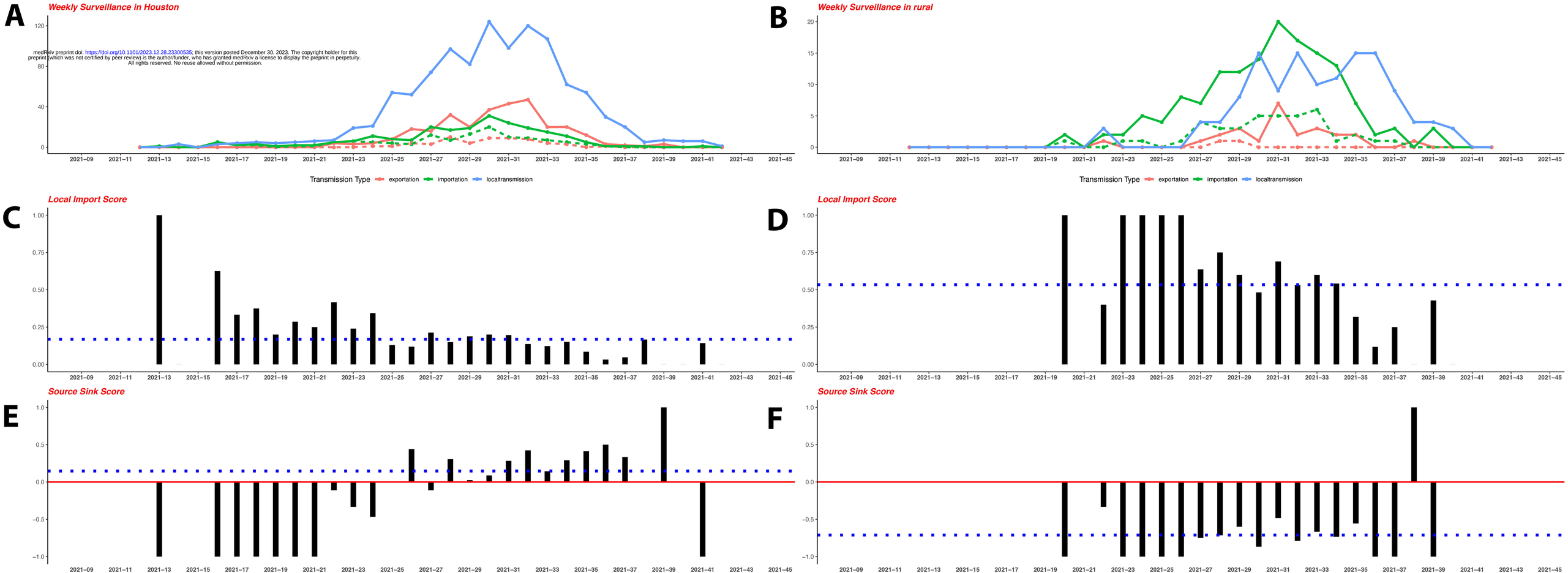
Figure 4. Epidemic trends in populated urban center vs the rural areas. A. The epidemic trend of Houston during the Delta wave. **B.** The epidemic trend of the rural areas during the Delta wave. The top of the panel shows the time series of spatial transmission counts by week. The dashed pink line represents exports from the analyzed regions to non-Texas. The dashed green line represents imports from non-Texas into the analyzed regions. **C.** The trend of Local Import Score in Houston. **D.** The trend of Local Import Score in rural areas. The black bars in the middle of the panel depict the weekly dynamics of Local Import Score. The dashed blue line indicates the accumulated Local Import Score during the Delta wave. **E.** The trend of Source Sink Score in Houston. **F**. The trend of Source Sink Score in rural areas. The solid red line represents the benchmark of 0, indicating a balance between imports and exports. The dashed blue line marks the accumulated Source Sink Score during the Delta wave.



medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.


Rural–Urban Continuum Code
RUCC 1
RUCC 2
RUCC 3
rural
weight
100
200
300


Source
0.00
−0.25
−0.50
−0.75
−1.00
RUCC 1
RUCC 2
RUCC 3
rural
Sink



medRxiv preprint doi: https://doi.org/10.1101/2023.12.28.23300535; this version posted December 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.