nature portfolio

Corresponding author(s): Justin Bahl

Last updated by author(s): Feb 10, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Genomic data used in this study are publicly available from the GISAID database (Global Initiative on Sharing All Influenza Data). All SARS-CoV-2 genome sequences were obtained from GISAID's open access repository, which requires user registration and agreement to data-sharing terms.
Data analysis	Custom R scripts were developed for the exploration of sampling heterogeneity and the implementation of proportional sampling schemes. Phylogenetic analysis was performed using the Nextstrain pipeline. For the characterization of spatial transmission linkages, we used R code, utilizing the treeio and tidytree packages to read and manipulate tree structures. All scripts used to generate the results for the Texas case study are publicly available on GitHub at https://github.com/leke-lyu/transmissionCount.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The GISAID accession IDs of the genomes used in this study are provided on our GitHub repository (https://github.com/leke-lyu/transmissionCount). Additional data obtained during the study is available from the corresponding author upon reasonable request.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This study did not involve human participants, their personal data, or biological material. The research was conducted using publicly available SARS-CoV-2 genome sequences from the GISAID database. These genomic data do not include any identifiable information related to sex, race, ethnicity, or other personal attributes of the individuals from whom the samples were collected.
Reporting on race, ethnicity, or other socially relevant groupings	Same statement as above.
Population characteristics	Same statement as above.
Recruitment	Same statement as above.
Ethics oversight	No formal ethics approval was required for this study as it solely involved the use of publicly available SARS-CoV-2 genome sequences from the GISAID database. The data do not contain any personally identifiable information, and the study adhered to all relevant data-sharing policies outlined by GISAID.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study analyzed the spread of COVID-19 across urban and rural areas in Texas by examining over 50,000 viral genomes. We introduced a novel spatial transmission count statistic to efficiently summarize geographic transmission patterns reflected in viral phylogenies.
Research sample	The research was conducted using publicly available SARS-CoV-2 genome sequences from the GISAID database. The dataset was divided into two parts: the first focusing on Texas as the primary study area, and the second comprising a worldwide dataset used as a contextual background for comparative analysis.
Sampling strategy	The Texas dataset was proportionally sampled based on the case count in each region. The worldwide dataset was randomly sampled.
Data collection	The research was conducted using publicly available SARS-CoV-2 genome sequences from the GISAID database. The GISAID accession IDs of the genomes used in this study are provided on our GitHub repository (https://github.com/leke-lyu/transmissionCount).
Timing and spatial scale	Genome sequences were sampled from March 2021 to October 15, 2021, covering the full Delta variant outbreak. The dataset included samples from 49 countries across 6 continents globally.
Data exclusions	Genome sequencing, bioinformatic processing of raw data, and sequence alignment are steps where errors can occur, potentially distorting the phylogenetic analysis. The Nextstrain pipeline provides a list of problematic sequences, which we filter out before proceeding to ensure these errors do not affect the analysis.

Reproducibility	In this study, we introduced a novel spatial transmission count statistic to efficiently summarize geographic transmission patterns between rural and urban areas in Texas. The robustness of two key metrics—the Source Sink Score and the Local Import Score—was tested through a sensitivity analysis.
Randomization	In this study, we examined genome sampling bias and implemented subsampling scheme adjustments.
Blinding	Blinding was not employed during data acquisition or analysis in this study. Since the research focused on genomic sequencing and computational analysis of viral transmission patterns, all steps were based on objective data derived from viral genomes. The nature of bioinformatic workflows, such as sequence alignment and phylogenetic reconstruction, minimizes the need for subjective interpretation, making blinding less relevant to our study.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*