Phylogenetic inference of inter-population transmission rates for infectious diseases

Skylar Ann Gay[1], Greg Ellison[2], Jianing Xu[2], Jialin Yang[2], Yiliang Wei[3], Shaoyuan Wu[3], Lili Yu[4], Christopher C. Whalen[5], Jonathan Arnold[1,6], Liang Liu[1,2*]

[1]Institute of Bioinformatics, University of Georgia, Athens, GA 30602

[2]Department of Statistics, University of Georgia, Athens, GA 30602

[3]Jiangsu Key Laboratory of Phylogenomics & Comparative Genomics, Jiangsu International Joint Center of Genomics, School of Life Sciences, Jiangsu Normal University, Xuzhou, Jiangsu 221116, China

[4]Department of Biostatistics, College of Public Health, Georgia Southern University, Statesboro, GA 30677

[5]Global Health Institute, Department of Epidemiology & Biostatistics, College of Public Health, University of Georgia, Athens, GA 30602

[6]Department of Genetics, University of Georgia, Athens, GA 30602

*** Correspondence:**

Liang Liu

lliu@uga.edu

Keywords: "phylogenetic tree", "SIR model", "infectious disease", "COVID-19", "transmission rate"

Abstract

Estimating transmission rates is a challenging yet essential aspect of comprehending and controlling the spread of infectious diseases. There are various methods available for this purpose, each with its own assumptions, data requirements, and limitations. This paper introduces a phylogenetic approach called transRate, designed to estimate inter-population transmission rates. The phylogenetic method, which maintains statistical consistency under the multi-population Susceptible-Infected-Recovered (SIR) model, integrates genetic information with traditional epidemiological approaches. This integration improves the accuracy of transmission rate estimates, facilitating more effective disease control and prevention strategies. Simulation analyses validate the precision of transRate in estimating transmission rates. With the growing abundance of public databases for genomic sequences, transRate is becoming more prevalent in tracking and preventing the spread of such diseases.

Background

Assessing transmission rates is essential for understanding the dynamics of infectious diseases and developing effective control measures (ANDRAUD *et al.* 2009; ALSHAMMARI 2023). By gaining insights into transmission rates, public health officials can create strategies to mitigate the impact of a disease outbreak and be prepared for potential future outbreaks (AUDU *et al.* 2006; ARAVINDAKSHAN *et al.* 2022). Various techniques and approaches have been devised to estimate transmission rates using epidemiological and genetic data in infectious diseases (BECKER AND HASOFER 1998; STADLER *et al.* 2013; KIRKEBY *et al.* 2017; MUBAYI *et al.* 2021; CHANG AND DE JONG 2023). Conventional approaches, such as the computation of the basic reproduction number $\boldsymbol{R_0}$, represent some of the most straightforward techniques for assessing transmission rates. Estimation of $\boldsymbol{R_0}$ involves examining the growth rate of the epidemic curve under the Susceptible-Infected-Recovered (SIR) and Susceptible-Exposed-Infectious-Recovered (SEIR) models, assuming a consistent transmission rate and a homogeneous population (FRASSO AND LAMBERT 2016). This method fails to consider temporal fluctuations in the transmission rate (DERAKHSHAN *et al.* 2021), attributable to seasonal variations, interventions, or shifts in behavior. Several methods have been devised to account for temporal variations in estimating transmission rates (GANYANI *et al.* 2020; LIPPIELLO *et al.* 2022; BUCH *et al.* 2023). The performance of these methods relies on the assumption that symptom onset accurately reflects the date of infection, which might not hold true for cases of asymptomatic transmission (PARK *et al.* 2020).

Analyzing epidemiological data offers valuable insights into the transmission dynamics of a disease by fitting models to observed data and estimating relevant parameters (DABIS *et al.* 1993; KEELING *et al.* 2020; LARREMORE *et al.* 2021; MOON AND SCOGLIO 2021). Since transmission rates can exhibit spatial variability (KUO AND WEN 2022), analyzing the disease spread across different regions is instrumental in understanding the spatial distribution of transmission rates (OESTERHOLT *et*

al. 2006; LACHISH *et al.* 2011). Furthermore, advancements in genomic sequencing technologies have enabled researchers to trace the dissemination of pathogens at a molecular level (STOCKDALE *et al.* 2023). The analysis of genetic data, in particular, has emerged as a potent tool for estimating transmission rates in infectious diseases (MOHAMED *et al.* 2019). In the realm of infectious diseases, phylogenetic trees constructed from genetic data are the fundamental tools to elucidate the relatedness of different pathogen strains. By scrutinizing the branching patterns of the tree, researchers can deduce transmission dynamics, including the direction and frequency of transmission events (STADLER *et al.* 2013).

Traditional approaches for estimating transmission rates have primarily focused on understanding the spread of infectious diseases within a single population. In this paper, we introduce a multi-population susceptible-infectious-recovered (SIR) model to investigate transmission rates within and across populations (Figure 1). Based on the multi-population SIR model, a phylogenetic method is developed to accurately estimate the inter-population transmission rate. The phylogenetic approach aims to provide a comprehensive understanding of disease transmission dynamics across multiple populations.

Materials and Methods

Modelling transmissions for multiple populations

The multi-population SIR model is an expanded version of the conventional SIR model (KERNER AND MCKENDRICK 1927) that is used to simulate disease transmissions in a population (S1 in Supplementary data). In the multi-population SIR model, transmission events occur within and between K populations $\Omega_1, \dots, \Omega_K$ of size $N_1, \dots, N_K$ during a time interval $[0, d]$. Let $S_{t,k}, I_{t,k}, R_{t,k}$ be the number of susceptible, infectious, and recovered individuals at time t for the population $k =$

78 $1, \dots, K$. The variables $S_{t,k}, I_{t,k}, R_{t,k}$ in each population satisfy the differential equations of the SIR

79 model for a single population (S1 in Supplementary data), i.e.,

80
$$\begin{cases} \frac{dS_{t,k}}{dt} = -\beta_k I_{t,k} S_{t,k} \\ \frac{dI_{t,k}}{dt} = \beta_k I_{t,k} S_{t,k} - \gamma_k I_{t,k} \\ \frac{dR_{t,k}}{dt} = \gamma_k I_{t,k} \end{cases} \quad (1)$$

81 Moreover, transmissions occur between two populations i and j at a constant rate ω_{ij} for

82 transmissions from population Ω_j to population Ω_i and ω_{ji} for transmissions from population Ω_i to

83 population Ω_j (Figure 1). The inter-population transmission rate ω_{ij} represents the probability of an

84 individual from population Ω_j traveling to population Ω_i and contracting the infection in Ω_i, i.e.,

85
$$\omega_{ij} = wv \quad (2)$$

86 where w is the probability that an individual in population Ω_j travels to population Ω_i, and v is the

87 probability that an individual who has traveled to population Ω_i gets infected in Ω_i. If individuals in

88 population Ω_j independently have the same probability w of travelling to population Ω_i, then the

89 number $y_{t,j}$ of individuals in population Ω_j who travel to population Ω_i at time t follows the

90 binomial distribution, i.e.,

91
$$y_{t,j} \sim Binomial(N_j, w) \quad (3)$$

92 Given $y_{t,j}$, the number $x_{t,j}$ of individuals in population Ω_j who have travelled to and gotten infected

93 in population Ω_i at time t follows the binomial distribution, i.e.,

94
$$x_{t,j} | y_{t,j} \sim Binomial(y_{t,j} I_{t,i}, v) \quad (4)$$

95 The transmission events occurring in population Ω_i involve not only the infected individuals

96 in the population Ω_i, but also the $x_{t,j}$ individuals from population Ω_j who travel to and get infected in

97 Ω_i. Every newly infected individual $\mathcal{I}_{t,i}$ in Ω_i at time t, including the $x_{t,j}$ individuals who have

98 traveled from Ω_j to Ω_i, can be traced back to an infectious individual $\mathcal{A}_{t-1,i}$ in Ω_i at time $t-1$. This
99 transmission event is indicated by a mapping τ defined as follows

100 $$\tau: \mathcal{I}_{t,i} \mapsto \mathcal{A}_{t-1,i} \quad (5)$$

101 Since the number $x_{t,j}$ is negligible compared to the number $I_{t-1,i}$ of infectious individuals in the
102 population Ω_i at time $t-1$, we only consider the $I_{t-1,i}$ infectious individuals when we look
103 backward to find the infectious individual $\mathcal{A}_{t-1,i}$ at time $t-1$ who is the ancestor of a newly
104 infected individual $\mathcal{I}_{t,i}$ at time t. Furthermore, it is assumed that the $I_{t-1,i}$ infectious individuals at
105 time $t-1$ are equally likely to be the ancestor $\mathcal{A}_{t-1,i}$ of a newly infected individual $\mathcal{I}_{t,i}$, i.e., for $a =$
106 $1, \dots, I_{t-1,i}$, where a represents one of the $I_{t-1,i}$ infectious individuals in population i at time $t-1$,

107 $$P\left(\mathcal{A}_{t-1,i} = a\right) = \frac{1}{I_{t-1,i}} \quad (6)$$

108 Moreover, the $x_{t,j}$ individuals from population Ω_j are infected by the $I_{t,i}$ infectious individuals of
109 population Ω_i at time t. We assume that the $I_{t,i}$ infectious individuals at time t are equally likely to
110 be the ancestor $\mathcal{A}_{t,i}$ of one of the $x_{t,j}$ individuals from population Ω_j, i.e., for $b = 1, \dots, I_{t,i}$, where b
111 represents one of the $I_{t,i}$ infectious individuals,

112 $$P\left(\mathcal{A}_{t,i} = b\right) = \frac{1}{I_{t,i}} \quad (7)$$

113 All transmissions within an arbitrary time interval $[0, d_i]$ for $d_i > 2$ and $d_i \in \mathbb{N}$ in population Ω_i
114 form a tree-like structure, which is the transmission tree T_i for population Ω_i (Figure 1). We assume
115 that the roots $O_1, \dots, O_K$ of K transmission trees $T_1, \dots, T_K$ share a common ancestor, denoted by O^*,
116 i.e., the root of a super tree T^* in which the K transmission trees $T_1, \dots, T_K$ are the subtrees of T^*.

It follows from Equations 3-4 that the expected number $E(x_{t,j})$ of individuals who travel to and get infected in the population Ω_i is given by $E(x_{t,j}) = E\left(E(x_{t,j}|y_{t,j})\right) = N_j I_{t,i} wv = N_j I_{t,i} \omega_{ij}$, where $I_{t,i}$ is the number of infectious individuals in the population Ω_i at time t. The expectation of the total number $\sum_t x_{t,j}$ of individuals from the population Ω_j who get infected in the population Ω_i is equal to $E(\sum_t x_{t,j}) = N_j \omega_{ij} \sum_t I_{t,i}$, indicating that the transmission rate ω_{ij} can be estimated by the ratio of $\sum_t x_{t,j}$ and $N_j \sum_t I_{t,i}$, i.e.,

$$\widehat{\omega_{ij}} = \frac{\sum_t x_{t,j}}{N_j \sum_t I_{t,i}} \qquad (8)$$

The numerator $\sum_t x_{t,j}$ is the total number of individuals from the population Ω_j who travel to and get infected in the population Ω_i. The denominator $N_j \sum_t I_{t,i}$ can be calculated by $N_j \sum_t I_{t,i} = N_j \sum_{m=1}^{I_i} (t_{m,i}^R - t_{m,i}^I)$ where $t_{m,i}^R$ and $t_{m,i}^I$ are the recovery and infection time of the infected individual m in the population Ω_i, and I_i is the total number of infected individuals in the population Ω_i by time d_i. Thus, the estimate $\widehat{\omega_{ij}}$ can be calculated by

$$\widehat{\omega_{ij}} = \frac{\sum_t x_{t,j}}{N_j \sum_{m=1}^{I_i} (t_{m,i}^R - t_{m,i}^I)} \qquad (9)$$

The estimate $\widehat{\omega_{ij}}$ is unbiased and statistically consistent in estimating the inter-population transmission rate ω_{ij} (S2 and S3 in Supplementary data).

In real data analysis, however, we can only obtain a sample of infected individuals in populations $\Omega_1, \ldots, \Omega_K$. We assume that the infected individuals in the samples $\mathcal{S}_1,\ldots,\mathcal{S}_K$ are randomly selected from populations $\Omega_1, \ldots, \Omega_K$. Let n_i be the sample size of $\mathcal{S}_i$, and $\tilde{x}_{t,j}$ $(j \neq i)$ denotes the number of individuals in the sample $\mathcal{S}_j$ who travel to and get infected population Ω_i at time t. Let $\widetilde{I_i}$ for $i = 1, \ldots, K$ be the number of individuals in the samples $\mathcal{S}_i$ who get infected in population Ω_i. Let I_i be the total number of infected individuals by the time d_i in population Ω_i. The inter-population transmission rate ω_{ij} can be estimated by the samples $\mathcal{S}_i$ and $\mathcal{S}_j$, i.e.,

$$\widetilde{\omega_{ij}} = \frac{\frac{I_j}{\tilde{I}_j}\sum_t \tilde{x}_{t,j}}{N_j\left(\frac{I_i}{\tilde{I}_i}\sum_{m=1}^{\tilde{I}_i}\left(t^R_{m,i} - t^I_{m,i}\right)\right)} \quad (10)$$

The estimate $\widetilde{\omega_{ij}}$ converges to $\widehat{\omega_{ij}}$, i.e., $\widetilde{\omega_{ij}} \to \widehat{\omega_{ij}}$, as the sample sizes n_i and n_j approach to the total numbers N_i and N_j of the infected individuals in the populations Ω_i and Ω_j. We can show that $\widetilde{\omega_{ij}}$ is an asymptotically unbiased estimator of ω_{ij} and is statistically consistent in estimating the parameter ω_{ij} as the sample sizes n_1 and n_2 increase to infinity (S4 in Supplementary data).

We have developed a phylogenetic approach (transRate) to estimate the inter-population transmission rate ω_{ij} using the pathogen genomes labeled with their population origins $\Omega_1, \ldots, \Omega_K$. The phylogenetic method for transmission rate estimation consists of four steps: 1) Building a phylogenetic tree based on the pathogen genomes. 2) Identifying clades in the tree that have at least a certain percentage (default is 60%) of taxa with the same population origin. 3) Labeling each identified clade with the population origin of the majority of sequences in that clade. Any sequences labeled with a different population origin are inferred as inter-population transmission events. 4) Estimating the inter-population transmission rate ω_{ij} based on the identified clades. The variability inherent in the estimation of phylogenetic trees, including the formation of clades, could skew the accuracy of transmission rate estimation. However, given the constancy of the transmission rate over time, the percentage of transmission events remains the same throughout any given time frame. This indicates that, despite the inherent uncertainty in the phylogenetic tree and the identification of clades, the estimation of the transmission rate retains a certain degree of reliability.

Simulation

Estimation of transmission rates from phylogenetic trees

Given the transmission tree of two populations, we evaluated the performance of transRate for estimating the inter-population transmission rate. The transmission tree was generated from the two-population SIR model during a time interval $[0, d]$ (i.e., $d = 50$). The population size was set to be $N_1 = N_2 = 10{,}000$ and $N_1 = N_2 = 1{,}000{,}000$. For the population size $10{,}000$, we set the infection rate $\beta = 0.00005$ and the recovery rate $\gamma = 0.05$. For the population size $1{,}000{,}000$, we set $\beta = 0.0000005$ and $\gamma = 0.05$. The number of susceptible (S_t), infected (I_t), and recovered (R_t) at time t was obtained by solving the differential equations of the two-population SIR model using an R package deSolve (SOETAERT *et al.* 2010). The number $y_{t,2}$ of individuals who traveled from the population Ω_2 to the population Ω_1 at time $t \in [0, 50]$ was simulated from the binomial distribution with mean = $N_2 w$, where $w = 0.0001$ for the population size $10{,}000$ and $w = 0.000001$ for the population size $1{,}000{,}000$. The parameter w is the probability that an individual in the population Ω_2 travels to the population Ω_1, i.e., the average number of travelers from the population Ω_2 to the population Ω_1 of size 10000 is $10000 \times 0.0001 = 10$ individuals per day. Given $y_{t,2}$, the number $x_{t,2}$ of individuals who traveled to and were infected in the population Ω_1 at time t was simulated from the binomial distribution with the infection rate $v = 0.002, 0.004, 0.006, 0.008$. The inter-population transmission rate ω_{12} from the population Ω_2 to the population Ω_1 is equal to the product of two probabilities w and v, i.e., $\omega = wv = 2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7}$ for the population size 10,000 and $\omega = wv = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-9}$ for the population size 1000,000, respectively. Similarly, the transmissions from the population Ω_1 to the population Ω_2 were simulated with the transmission rate ω_{21}. Two inter-population transmission rates were assumed to be equal to each other, i.e., $\omega_{12} = \omega_{21}$.

A phylogenetic tree T_1 was subsequently constructed from the simulated transmissions in the population Ω_1. Let x_i be the number of new infections on day i. Let y_{i-1}be the number of infections on day $(i - 1)$. Note that x_i and y_{i-1} may include inter-population infections. It follows that the x_i

new infections on day i were infected by the y_{i-1} infectious individuals on day $(i-1)$. Since the infectious individuals on day $(i-1)$ were equally likely to infect the susceptible individuals on day i, the ancestor of each new infection on day i was found by randomly sampling an infectious individual on day $(i-1)$. The ancestors formed the ancestral history (i.e., the phylogenetic tree T_1) of the transmissions in the population Ω_1 generated from the two-population SIR model. Similarly, a phylogenetic tree T_2 was constructed from the transmissions in the population Ω_2. Two trees were combined into a super tree T. This super tree T was the input to estimate the transmission rates ω_{12} and ω_{21} using Equation 9. Moreover, we randomly selected $n = 100, 200, 300, 400, 500$ infected individuals (taxa) from each population in the super tree T. The phylogenetic tree of the sampled infections was utilized to estimate the inter-population transmission rates ω_{12} and ω_{21} using Equation 10. Each simulation was repeated 100 times and we calculated the mean squared error (MSE) and co-efficient variation (CV) of the estimates of the transmission rate. Since two transmission rates are equal to each other, we only present the MSE and CV of the transmission rate ω_{12}, i.e., $MSE = \frac{1}{100}\sum_{i=1}^{100}\left(\widehat{\omega_{12}}^{\,i} - \omega_{12}\right)^2$ and $CV = \frac{sd(\widehat{\omega_{12}})}{mean(\widehat{\omega_{12}})}$, where $sd(\widehat{\omega_{12}})$ is the standard deviation of $\widehat{\omega_{12}}$.

Estimation of transmission rates from molecular sequences

In the preceding simulation, the phylogenetic tree was derived from the transmissions generated by the two-population SIR model and was assumed to be a known input for estimating the inter-population transmission rate ω_{12}. However, in practice, the phylogenetic tree is typically inferred from the sequence alignments of pathogen genomes. Therefore, it becomes crucial to account for the uncertainty associated with the estimated phylogenetic tree when estimating the transmission rate ω_{12}. Once the phylogenetic tree was constructed based on the transmissions generated from the two-population SIR model, we proceeded to simulate DNA sequences of 20,000 base pairs using the phylogenetic program Seq-Gen (RAMBAUT AND GRASSLY 1997) with the mutation rate $\mu =$

$0.01, 0.001, 0.0001$. If the phylogenetic tree involved polytomies which were not readable by the program Seq-Gen, the polytomy nodes in the phylogenetic tree were replaced by bifurcating nodes with 0 internal branch length. These sequences were utilized to reconstruct the maximum likelihood (ML) tree using FastTree (PRICE *et al.* 2009), employing the following command line: *fasttree -nt -nosupport seqfile > outputfile*. The estimated phylogenetic tree served as the input for inferring the transmission rate ω_{12}. Each simulation was repeated 100 times and we calculated the mean squared error (MSE) and co-efficient variation (CV) of the estimates of the transmission rate ω_{12}.

Estimation of transmission rates for multiple populations

In this simulation, transmissions were generated from the multi-population SIR model for five populations $\Omega_1, \Omega_2, \Omega_3, \Omega_4, \Omega_5$. The populations were characterized by two different sizes, one with 10,000 individuals and another with 1,000,000 individuals. For the smaller population (10,000 individuals), transmission rates (ω_{ij} for $i, j = 1, \ldots, 5$) were set at values of 2×10^{-7}, $4 \times 10^{-7}, 6 \times 10^{-7}$ and 8×10^{-7}, while for the larger population (1,000,000 individuals), transmission rates were configured at 2×10^{-9}, $4 \times 10^{-9}, 6 \times 10^{-9}$, and 8×10^{-9}. A sample of infected individuals (i.e., $n_1 = n_2 = n_3 = n_4 = n_5 = 100, 200, 300$) was randomly selected from each of the five populations. Due to the limitation of the phylogenetic tree reconstruction method, we did not sample more than 300 individuals. The selected individuals were labelled with their population origins. Subsequently, a phylogenetic tree was constructed from the transmissions among the five samples of infected individuals. This phylogenetic tree featured five major clades, each corresponding to the sample selected from one of the five populations. DNA sequences of 20,000 base pairs were simulated from the phylogenetic tree using Seq-Gen. These simulated sequences were employed as input data to estimate ML trees using the FastTree algorithm. The ML tree, in turn, was utilized to infer the transmission rates (ω_{ij}). Finally, we evaluated the performance of transRate by calculating the MSE and CV of the estimates of the transmission rates (ω_{ij}). The MSE and CV of

231 the transmission rate estimates served as a measure of how well the phylogenetic approach performed
232 in estimating transmission rates in the simulation.

233 **Data Analysis of SARS-CoV-2 Genomes in the Early Pandemic**

234 The proposed phylogenetic method transRate was applied to a genomic dataset consisting of 40,028
235 sequences of SARS-CoV-2 in human hosts during the early SARS-CoV-2 pandemic (YANG *et al.*
236 2023). This dataset was formed from 41,910 coronavirus genomes downloaded from NCBI GenBank
237 "nucleotide" database on August 26, 2021, with all sequences collected between December 31, 2019,
238 and April 1, 2020. The dataset was filtered to remove sequences with missingness, containing frame
239 shift, and incomplete genomes. There was no geographical filtering. Samples originated from the
240 Americas, Europe, Oceania, and Asia. This filtration returned a dataset of 40,028 SARS-CoV-2
241 genomes isolated from human hosts.Thespecies tree was estimated from the genomic data using a
242 coalescent method NJst (LIU AND YU 2011). The geographical distribution of the clades in the species
243 tree are as follows: 11 clades geographically centered in the Americas, 5 clades geographically
244 centered in Asia, 18 clades geographically centered in Europe, and 1 clade geographically centered in
245 Oceania (S5 in Supplementary data). A number of these clades did not contain any geographical
246 outliers. A group of datasets were formed for the following populations: "Africa" (includes any
247 clades that were centered in countries within the continent of Africa), "Americas" (includes any
248 clades that were centered in countries or localities that were in North, Central, and South America),
249 "Asia" (includes any clades that were centered in countries located in Asia), "Europe" (includes any
250 clades that were centered in countries located in Europe), and "Oceania" (includes any clades that
251 were centered in localities in Oceania). The transmission rate estimation calculation was applied to
252 the above populations. The recovery time is equal to 14 days as initially issued by the World Health
253 Organization in the early pandemic. Population data was collected based on the World Health
254 Organization and United Nations published data (S6 in Supplementary data).

A transmission analysis airplane plot was constructed using the previously described data set of 40,028 whole genome sequences of SARS-CoV-2 in human hosts. These samples underwent the process of forming clades based on bootstrap support, location, and number of individuals per clade, as outlined above. The inferred transmission events were plotted as an airplane plot with the larger black dots representing the location of which the majority of clade members originated. The smaller black dots are geographical outliers (samples not from within the location of the majority) in the clade and the arcs connecting the clade center to the geographical outliers represent an inferred transmission event. The color of the arcs indicates the time point at which the inferred transmission event occurred. This plot was created using the maps (Becker et al, 2022) and geosphere (Hijmans et al, 2022) packages.

Results

Simulation

Estimation of transmission rates from phylogenetic trees

Given the transmission tree generated from the two-population SIR model, we evaluated the performance of transRate in estimating the inter-population transmission rate ω_{12}. We considered two scenarios: 1) the phylogenetic tree was constructed from all transmissions simulated by the two-population SIR model and 2) the phylogenetic tree was constructed from a sample of transmissions simulated by the two-population SIR model. For Scenario 1, the MSE of the transmission rate estimates increases as the true value of the transmission rate increases from 2×10^{-7} to 8×10^{-7} (Figure 2). The MSEs for the population size of 10,000 are very small ($< 2 \times 10^{-14}$) (Figure 2a). Similar results can be observed for the population size of 1,000,000 (Figure 2b), indicating that transRate can accurately estimate the transmission rate ω_{12} when the phylogenetic tree of all transmissions is given. The coefficient of variation (CV, i.e., the ratio of the standard deviation to the

mean) of the estimates of the transmission rate ω_{12} for the population size of 1,000,000 (Figure 2b) is less than those for the population size of 10,000 (Figure 2a). This result is consistent with our theory that increasing the population size leads to a more accurate estimate of the transmission rate.

For Scenario 2 where the phylogenetic tree was constructed from a sample of transmissions, the MSE of the transmission rate estimates appears to decrease as the sample size increases from 100 to 1000 (Figure 3a), indicating that transRate can accurately estimate the transmission rate when the sample size is large. This rate of decrease varies across different values $(2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7})$ of the transmission rate ω_{12}. Notably, the rate of decrease appears to be constant across different values of the transmission rate ω (see Figure 3a-b). Moreover, the MSE of the transmission rate estimates decreases at a faster rate when the sample size increases from 100 to 200, then it becomes stable after the sample size increases to 400 (Figure 3a-b). The CV of the transmission rate estimates is less than 0.45 for the sample size ≥ 200. This result indicates that the sample size 200 is sufficient to accurately estimate the transmission rate when the phylogenetic tree of a sample of transmissions is given.

Estimation of transmission rates from sequences

In real-world scenarios, the transmission tree is often inferred from the pathogen genomes. In this simulation, we assess the accuracy of transRate in the presence of uncertainty of the estimated transmission tree. We employed the two-population SIR model to generate the transmission tree for a sample of transmissions, followed by simulating DNA sequences based on this transmission tree to construct the Maximum Likelihood (ML) trees. These ML trees were then utilized to estimate the transmission rate. The simulation results, based on a population size of 10,000 individuals, indicate that the MSE of the transmission rate estimate decreases as the sample size increases from 100 to 500 (Figure 4). This rate of decrease varies across different values $(2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times$

10^{-7}) of the transmission rate. Notably, the rate of decrease appears to be positively correlated with the values of the transmission rate (Figure 4). Moreover, the MSE of the transmission rate estimate decreases at a faster rate when the sample size increases from 100 to 200, then it becomes stable after the sample size increases to 300. In contrast, the MSE of the transmission rate estimate remains consistent across various values $(0.0001, 0.001, 0.01)$ of the mutation rate. The CV of the transmission rate estimates is less than 0.3 when the sample size increases to 200, indicating that the sample size 200 is sufficient to accurately estimate the transmission rate. The MSE and CV of the transmission rate estimates for the population size 1,000,000 are less than those for the population size 10,000, which is consistent with the expectation of the multi-population SIR model that increasing the population size leads to a more accurate estimate of the transmission rate.

Estimation of transmission rates for five populations

In this simulation, transmission events were generated from the five-population SIR model. It was assumed that all of 20 inter-population transmission rates ω_{ij} for $i, j = 1, \ldots, 5$ and $i \neq j$ were equal to each other. For the population size of 10,000, $\omega_{ij} = 2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7}$. For the population size of 1,000,000, $\omega_{ij} = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-}$. To evaluate the performance of transRate for estimating the transmission rate, we calculated the MSE and CV of the average $\bar{\omega}$ of the estimates of 20 transmission rates. The MSE of the average estimate $\bar{\omega}$ appears to be constant as the sample size increases from 100 to 300 (Figure 5a-b). For the population size of 10,000 and 1,000,000, the CV of the transmission rate estimates is less than 0.15 across various values $(0.0001, 0.001, 0.01)$ of the mutation rate and the population size (10,000 and 1,000,000), indicating that the sample size of 100 is sufficient to accurately estimate the transmission rate.

Data Analysis of SARS-CoV-2 Genomes in the Early Pandemic

The transmission rate estimates for the early SARS-CoV-2 pandemic, between December 31, 2019 and March 31, 2020, reveal transmission between Europe, Africa, Americas, and Asia (Table 1). Population 1 is the population in which the majority of the clade is geographically centered and Population 2 is the population in which geographical outliers originate from. The analysis of the 40,028 whole genome sequences of SARS-CoV-2 in human hosts revealed the efficiency of certain protection measures enacted by public health officials. The findings suggest that, by the time many travel restrictions were in place, much transmission had already occurred between populations and with unstable availability in testing, inter-population transmission rapidly increased.

The airplane plot of 40,028 whole genome sequences of SARS-CoV-2 in human hosts between December 31, 2019-March 31, 2020 (Figure 6). The clades pictured in the airplane plot reflect the 35 clades in Table 1. The geographic coordinates were taken as the closest non-transmission taxon to the transmissions within a clade as an inference for "case 0" in a particular clade. The arcs indicate early transmission throughout Asia and spreading to Europe. Later transmission events are pictured from Europe the Americas. The latest transmission events shown in the airplane plot are within Oceania.

Discussion

Estimating transmission rate is a challenging but essential task for understanding and controlling the spread of infectious diseases. There are different methods for estimating transmission rate from data, each with its own assumptions, data requirements, and limitations. The choice of the best method depends on the availability and quality of the data, the characteristics of the disease, and the objectives of the analysis. In this paper, we develop a phylogenetic approach (transRate) for estimating inter-population transmission rates. TransRate is statistically consistent in estimating

inter-population transmission rates. The simulation and real data analyses indicate that transRate can accurately estimate inter-population transmission rates.

The accuracy of transRate is influenced by the number of pathogen genomes that have been sampled from different populations. This underscores the importance of conducting higher rates of whole genome sequencing during outbreak events. However, the availability of data can present significant gaps, as not all sequences may be publicly accessible. For instance, while the dataset for SARS-CoV-2 mentioned here is extensive, there can still be biases in the data, particularly in the early stages of a pandemic (YANG *et al.* 2023). The release policies for viral genomes can vary greatly between countries and change over time, as observed during the 2020 pandemic. Consequently, the sample is no longer random. Additionally, there are regions around the world where limited resources hinder the acquisition of viral genomes, as mentioned earlier. This can create limitations, especially when it comes to analyzing non-simulated data. Another important consideration is the presence of asymptomatic cases in viral infections. Asymptomatic individuals may not receive whole genome testing, which may introduce further challenges to the accuracy of the method.

Conclusion

Molecular epidemiology and genetic data play a crucial role in estimating transmission rates, providing a detailed understanding of the genetic diversity and dynamics of infectious agents. The phylogenetic approach developed in this paper integrates genetic information with traditional epidemiological approaches to improve the accuracy of transmission rate estimates. Simulation and analytic results indicate that transRate can accurately estimate transmission rates from genomic data, contributing to more effective strategies for disease control and prevention. This method is well-suited for estimating transmission rates on large multi-population datasets in both epidemic and

endemic states. With the increasing availability of public databases for genomic sequences, this methodology is expected to become more prevalent as a valuable policy tool.

Key Points

- We develop a novel phylogenetic approach for estimating transmission rates in infectious disease
- The phylogenetic approach integrates genetic information with traditional epidemiological approaches.
- The phylogenetic approach is statistically consistent in estimating transmission rates under the multi-population SIR model
- Simulation studies confirm the accuracy of the phylogenetic method in estimating transmission rates.
- The utilization of this phylogenetic approach enhances the efficacy of disease control and prevention strategies.

Data Availability

The datasets analyzed for this study can be found in "The species coalescent indicates possible bat and pangolin origins of the COVID-19 pandemic" (YANG *et al.* 2023). R code generated for the simulation study is available on Github at https://github.com/sagay2022/Phylogenetic-inference-of-inter-population-transmission-rates-for-infectious-diseases.

Author Contributions

LL and JA designed the research. LL, SAG, and JY wrote R code. SAG wrote and conducted SARS-CoV-2 data analysis. GE, JX, YW, SW, LY, CW, and LL conducted the simulation analysis. LL and SAG, and CW wrote the manuscript.

Funding

The National Science Foundation for support of this research: NSF DBI-2029595, NSF DBI-2243206, and NSF DBI-1946937. Thank you to the incredible collaborators for providing key data and information and for making this project possible.

Figures

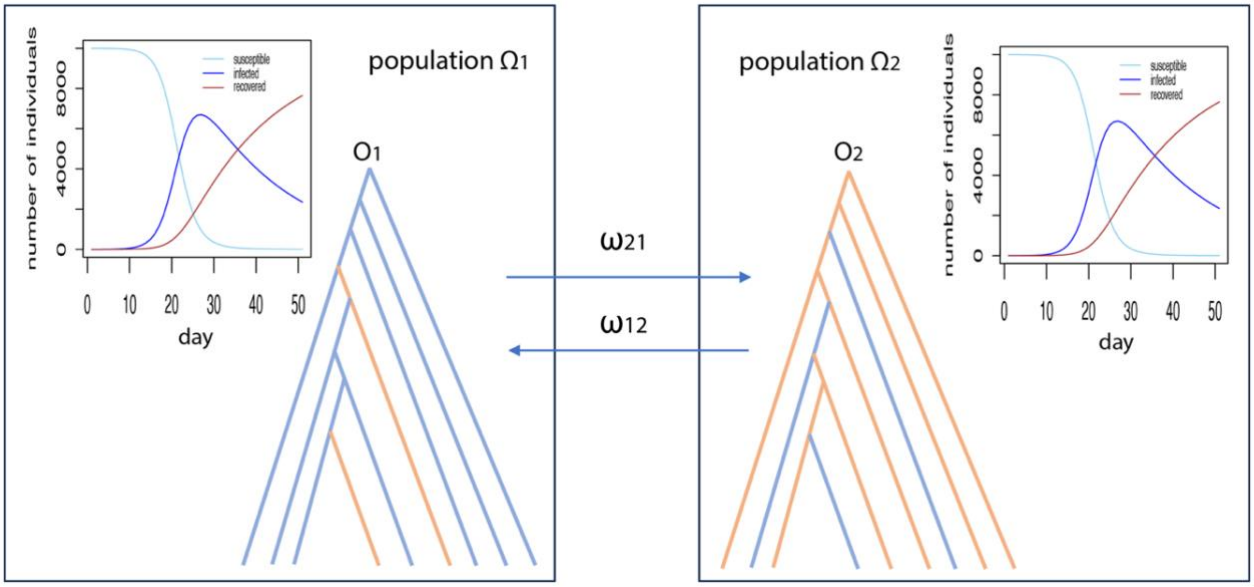

Figure 1: Transmission events generated from the two-population SIR model. The numbers of susceptible we consider the transmission events that occur within and between two populations Ω_1 (left panel) and Ω_2 (right panel). The numbers of susceptible (sky blue), infected (blue), and recovered (red) individuals at time t (day) were obtained by solving the differential equations for the two-population SIR model. Moreover, the model assumes that transmissions occur between two populations at a constant rate ω_{12} for transmissions from the population Ω_1 to the population Ω_2 and

ω_{21} for transmissions from population the Ω_2 to the population Ω_1. Every infected individual in the population Ω_1 can be traced back to an infector (i.e., ancestor) in the population Ω_1. The ancestral history of all transmissions in the population Ω_1 form a phylogenetic tree (left panel) with a root O_1 in which the blue lineages are the transmissions within the population Ω_1 and the orange lineages are the inter-population transmissions from the population Ω_2 to the population Ω_1. Similarly, a phylogenetic tree with a root O_2 (right panel) can be generated for the transmissions in the population Ω_2 where the orange lineages are the transmissions within the population Ω_2 and the blue lineages are the inter-population transmissions from the population Ω_1 to the population Ω_2.

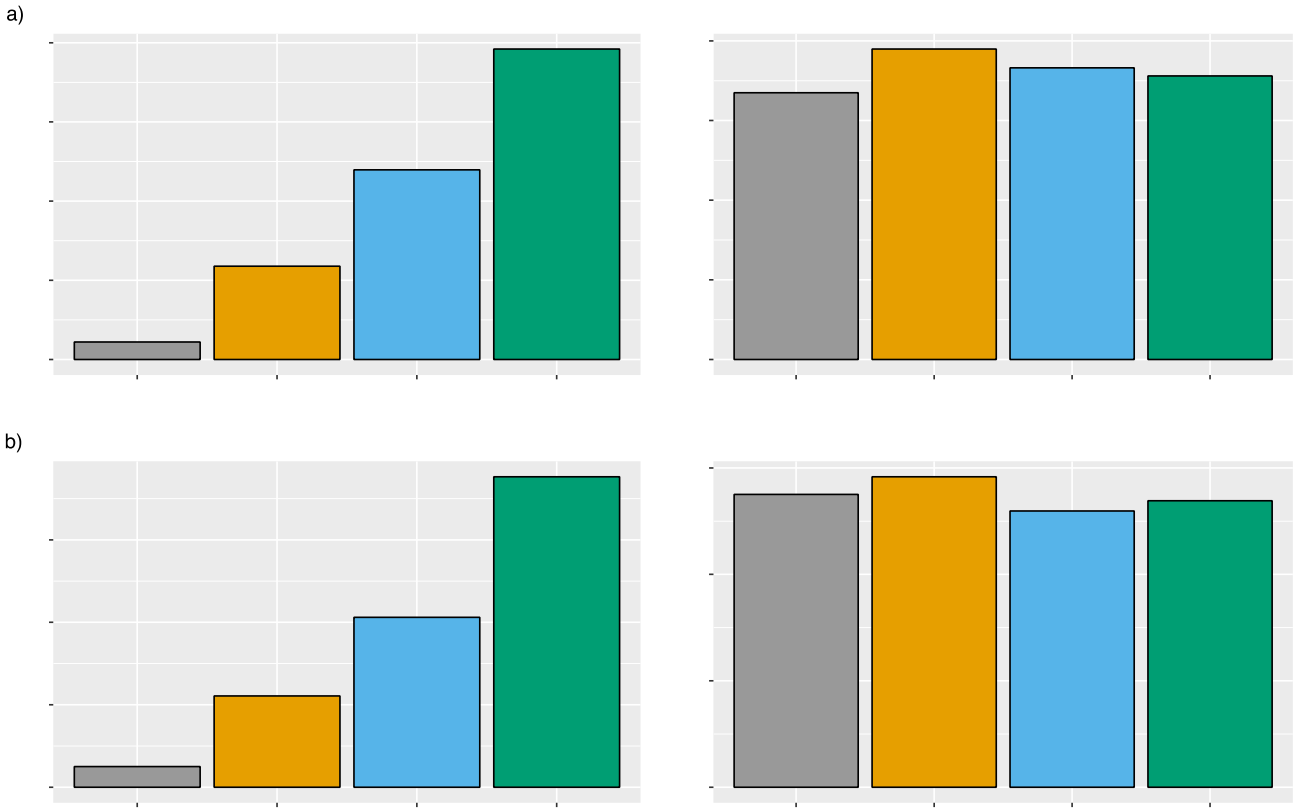

Figure 2: Estimation of the transmission rate from the phylogenetic tree of all infected individuals. The phylogenetic tree of all infected individuals was generated from the two-population SIR model. a) For population size = 10000, transmission events were simulated with the transmission rate $\omega =$

$2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7}$. b) For population size = 1000000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-9}$. The phylogenetic tree was then utilized to estimate the transmission rate ω. The simulation was repeated 100 times. The Mean Squared Error (MSE) and Coefficient Variation (CV) of the transmission rate estimates were calculated.

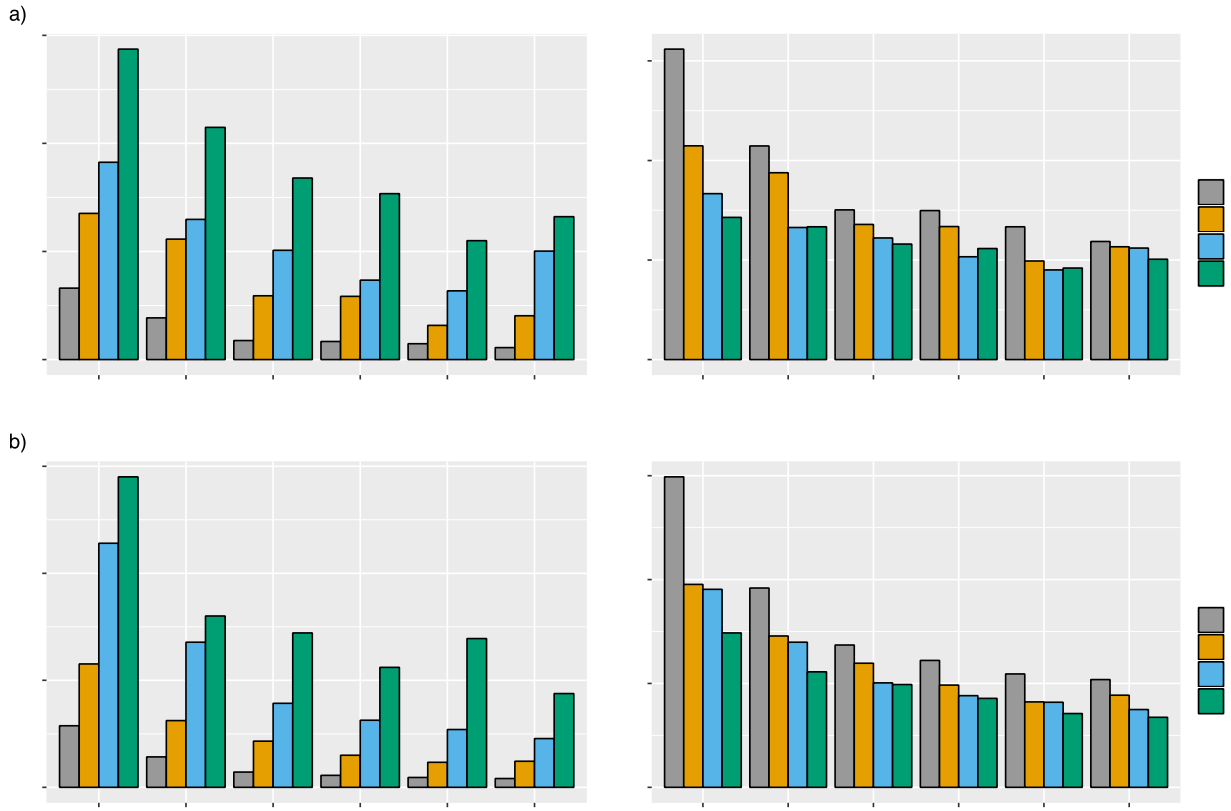

Figure 3: Estimation of the transmission rate from the phylogenetic tree of a sample of infected individuals. The phylogenetic tree of a sample of infected individuals (sample size = 100, 200, 400, 600, 800, 10000) was generated from the two-population SIR model. a) For population size = 10000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7}$. b) For population size = 1000000, transmission events were simulated with the

transmission rate $\omega = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-9}$. The phylogenetic tree was then utilized to estimate the transmission rate ω. The simulation was repeated 100 times. The Mean Squared Error (MSE) and Coefficient Variation (CV) of the transmission rate estimates were calculated.

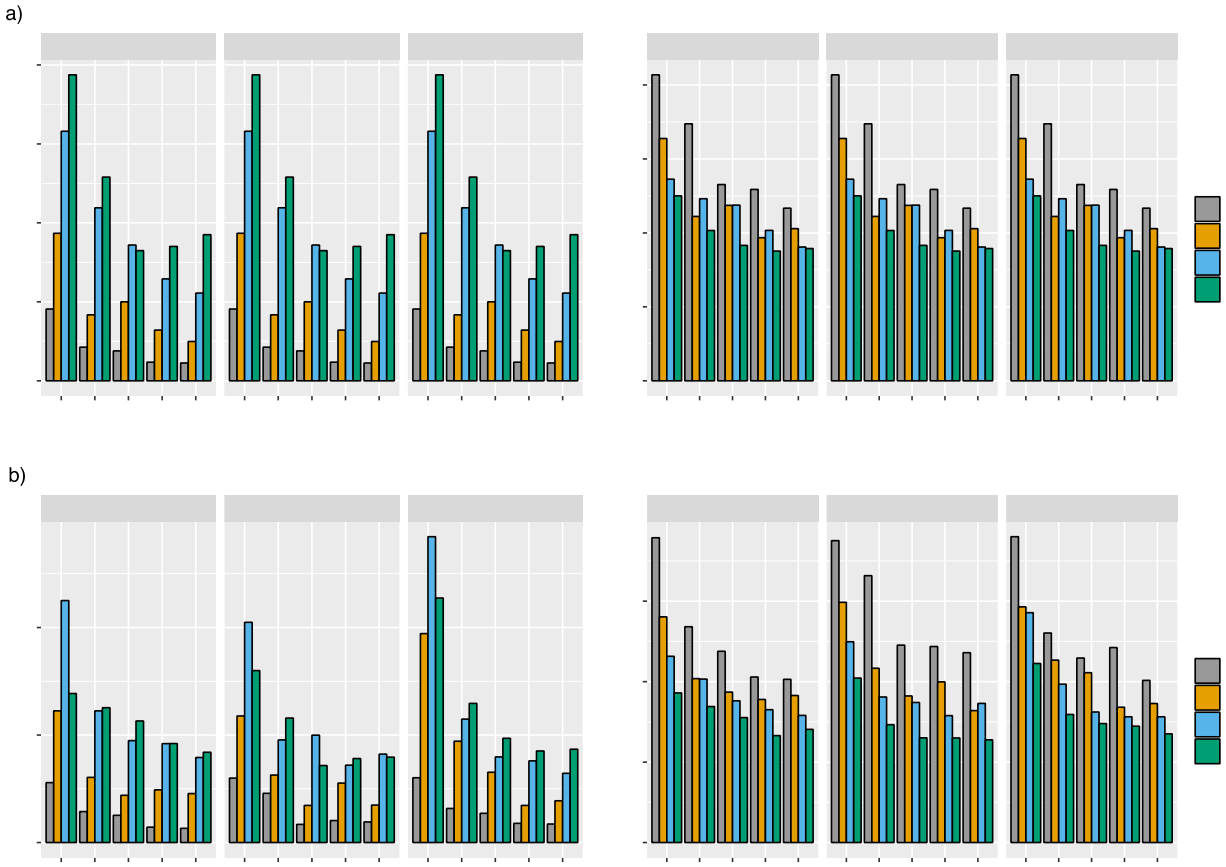

Figure 4: Estimation of the transmission rate from sequences. The phylogenetic tree of a sample of infected individuals (sample size = 100, 200, 300, 400, 500) was generated from the two-population SIR model. a) For population size = 10000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7}$. b) For population size = 1000000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-9}$. DNA sequences of 20,000 base pairs were simulated from the phylogenetic tree with the mutation rate = 0.0001, 0.001, 0.01 and then used to build the Maximum Likelihood (ML) trees. Finally, the

transmission rate ω was estimated by transRate using ML trees. The simulation was repeated 100 times. The Mean Squared Error (MSE) and Coefficient Variation (CV) of the transmission rate estimates $\hat{\omega}$ were calculated.

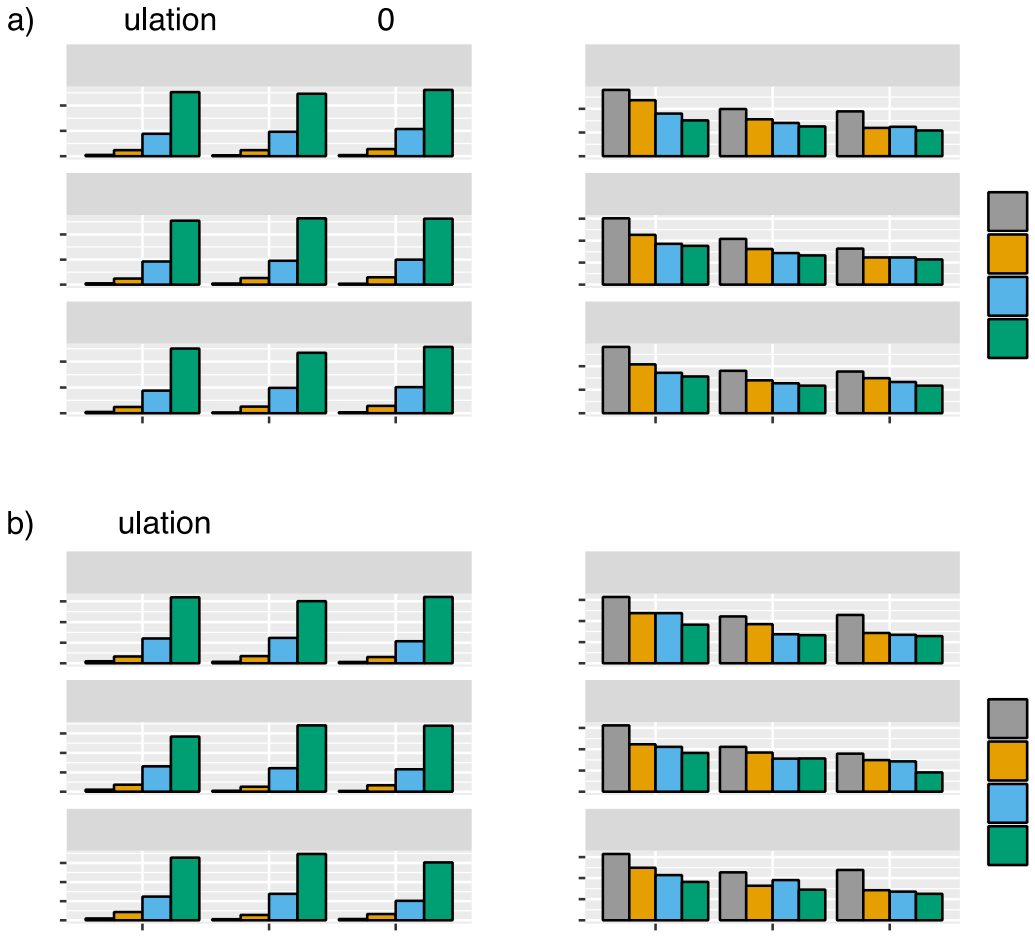

Figure 5: Estimation of the transmission rate for five populations. The phylogenetic tree of a sample of infected individuals (sample size = 100, 200, 300) was generated from the multi-population SIR model for five populations. a) For the population size = 10,000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7}$. b) For the population size = 1,000,000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-9}$. DNA sequences of 20,000 base pairs were simulated from the

phylogenetic tree with the mutation rate = 0.0001, 0.001, 0.01 and then used to build the Maximum Likelihood (ML) trees. Finally, the transmission rate was estimated by transRate using ML trees. The simulation was repeated 100 times. The Mean Squared Error (MSE) and Coefficient Variation (CV) of the transmission rate estimates $\hat{\omega}$ were calculated.

Figure 6: An airplane plot of transmission analysis of 40,028 whole genome sequences of SRAS-CoV-2 in human hosts between December 31, 2019-March 31, 2020. The larger black dots on the map represent the geographical location of clades formed based on 80% bootstrap support value and 80% locality identity. The smaller black dots are geographical outliers in the clade and the arcs connecting the clade center to the geographical outliers represent an inferred transmission event. The color of the arcs indicates the time point in which the inferred transmission event occurred. Transmission events are categorized into three time points: January 2020, February 2020, and March 2020. The geographic coordinates were taken as the closest non-transmission taxon to the transmissions within a clade as an inference for "case 0" in a particular clade.

Table 1: Inter-population transmission rate estimates of 40,028 whole genome SARS-CoV-2 sequences.

	Africa	Asia	Americas	Europe	Oceania
Africa	-				
Asia		-	2.205031e-09		6.178179e-07
Americas		1.562009e-09	-	1.01126e-08	7.398807e-08
Europe	1.123952e-07	6.297804e-10	1.414143e-09	-	1.337251e-07
Oceania			2.255848e-08	9.023393e-08	-

References

Alshammari FS. Analysis of SIRVI model with time dependent coefficients and the effect of vaccination on the transmission rate and COVID-19 epidemic waves. Infect Dis Model. 2023;8**:** 172-182.

Andraud M, Grasland B, Durand B, Cariolet R, Jestin A, Madec F *et al.* Modelling the time-dependent transmission rate for porcine circovirus type 2 (PCV2) in pigs using data from serial transmission experiments. J R Soc Interface. 2009;6**:** 39-50.

Aravindakshan A, Boehnke J, Gholami E, Nayak A. The impact of mask-wearing in mitigating the spread of COVID-19 during the early phases of the pandemic. PLOS Glob Public Health. 2022;2**:** e0000954.

Audu RA, Salu OB, Musa AZ, Onyewuche J, Funso-Adebayo EO, Iroha EO *et al.* Estimation of the rate of mother to child transmission of HIV in Nigeria. Afr J Med Med Sci. 2006;35**:** 121-124.

Becker NG, Hasofer AM. Estimating the transmission rate for a highly infectious disease. Biometrics. 1998;54**:** 730-738.

Buch DA, Johndrow JE, Dunson DB. Explaining transmission rate variations and forecasting epidemic spread in multiple regions with a semiparametric mixed effects SIR model. Biometrics. 2023.

Chang Y, de Jong MCM. A novel method to jointly estimate transmission rate and decay rate parameters in environmental transmission models. Epidemics. 2023;42**:** 100672.

Dabis F, Msellati P, Dunn D, Lepage P, Newell ML, Peckham C, Van de Perre P. Estimating the rate of mother-to-child transmission of HIV. Report of a workshop on methodological issues Ghent (Belgium), 17-20 February 1992. The Working Group on Mother-to-Child Transmission of HIV. AIDS. 1993;7**:** 1139-1148.

Derakhshan M, Ansarian HR, Ghomshei M. Temporal variations in COVID-19: an epidemiological discussion with a practical application. J Int Med Res. 2021;49**:** 3000605211033208.

Frasso G, Lambert P. Bayesian inference in an extended SEIR model with nonparametric disease transmission rate: an application to the Ebola epidemic in Sierra Leone. Biostatistics. 2016;17**:** 779-792.

Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, Hens N. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Euro Surveill. 2020;25.

514 Keeling MJ, Hollingsworth TD, Read JM. Efficacy of contact tracing for the containment of the 2019
515 novel coronavirus (COVID-19). J Epidemiol Community Health. 2020;74**:** 861-866.

516 Kerner WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proc. Math.
517 Phys. Eng. Sci. 1927;115**:** 700-721.

518 Kirkeby C, Halasa T, Gussmann M, Toft N, Graesboll K. Methods for estimating disease
519 transmission rates: Evaluating the precision of Poisson regression and two novel methods. Sci
520 Rep. 2017;7**:** 9496.

521 Kuo FY, Wen TH. Assessing the spatial variability of raising public risk awareness for the
522 intervention performance of COVID-19 voluntary screening: A spatial simulation approach.
523 Appl Geogr. 2022;148**:** 102804.

524 Lachish S, Knowles SC, Alves R, Wood MJ, Sheldon BC. Infection dynamics of endemic malaria in
525 a wild bird population: parasite species-dependent drivers of spatial and temporal variation in
526 transmission rates. J Anim Ecol. 2011;80**:** 1207-1216.

527 Larremore DB, Fosdick BK, Bubar KM, Zhang S, Kissler SM, Metcalf CJE *et al.* Estimating SARS-
528 CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological
529 surveys. Elife. 2021;10.

530 Lippiello E, Petrillo G, de Arcangelis L. Estimating the generation interval from the incidence rate,
531 the optimal quarantine duration and the efficiency of fast switching periodic protocols for
532 COVID-19. Sci Rep. 2022;12**:** 4623.

533 Liu L, Yu L. Estimating species trees from unrooted gene trees. Syst Biol. 2011;60**:** 661-667.

534 Mohamed W, Ito K, Omori R. Estimating Transmission Potential of H5N1 Viruses Among Humans
535 in Egypt Using Phylogeny, Genetic Distance and Sampling Time Interval. Front Microbiol.
536 2019;10**:** 2765.

537 Moon SA, Scoglio CM. Contact tracing evaluation for COVID-19 transmission in the different
538 movement levels of a rural college town in the USA. Sci Rep. 2021;11**:** 4891.

539 Mubayi A, Pandey A, Brasic C, Mubayi A, Ghosh P, Ghosh A. Analytical Estimation of Data-
540 Motivated Time-Dependent Disease Transmission Rate: An Application to Ebola and
541 Selected Public Health Problems. Trop Med Infect Dis. 2021;6.

542 Oesterholt MJ, Bousema JT, Mwerinde OK, Harris C, Lushino P, Masokoto A *et al.* Spatial and
543 temporal variation in malaria transmission in a low endemicity area in northern Tanzania.
544 Malar J. 2006;5**:** 98.

545 Park SW, Cornforth DM, Dushoff J, Weitz JS. The time scale of asymptomatic transmission affects
546 estimates of epidemic potential in the COVID-19 outbreak. Epidemics. 2020;31**:** 100392.

547 Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles
548 instead of a distance matrix. Mol Biol Evol. 2009;26**:** 1641-1650.

549 Rambaut A, Grassly NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence
550 evolution along phylogenetic trees. Comput Appl Biosci. 1997;13**:** 235-238.

551 Soetaert K, Petzoldt TR, Setzer W. Solving Differential Equations in R: Package deSolve. Journal of
552 Statistical Software. 2010;33**:** 1-25.

553 Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal
554 changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci U S A.
555 2013;110**:** 228-233.

Stockdale JE, Susvitasari K, Tupper P, Sobkowiak B, Mulberry N, Goncalves da Silva A *et al.* Genomic epidemiology offers high resolution estimates of serial intervals for COVID-19. Nat Commun. 2023;14**:** 4830.

Yang J, Skaro M, Chen J, Zhan D, Lyu L, Gay S *et al.* The species coalescent indicates possible bat and pangolin origins of the COVID-19 pandemic. Sci Rep. 2023;13**:** 5571.