

Phylogenetic inference of inter-population transmission rates for infectious diseases

Skylar A. Gay¹, Gregory Ellison², Jianing Xu², Jialin Yang², Yiliang Wei³, Shaoyuan Wu³, Lili Yu⁴, Christopher C. Whalen⁵, Jonathan Arnold^{1,6}, Liang Liu^{1,2,*}

¹Institute of Bioinformatics, University of Georgia, 120 Green Street, Athens, GA 30602, United States

²Department of Statistics, University of Georgia, 310 Herty Drive, Athens, GA 30602, United States

³Jiangsu Key Laboratory of Phylogenomics and Comparative Genomics, Jiangsu International Joint Center of Genomics, School of Life Sciences, Jiangsu Normal University, 101 Shanghai Road, Xuzhou, Jiangsu 221116, China

⁴Department of Biostatistics, Epidemiology and Environmental Health Sciences, College of Public Health, Georgia Southern University, 1332 Southern Drive, Statesboro, GA 30677, United States

⁵Global Health Institute, Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, 100 Foster Road, Athens, GA 30602, United States

⁶Department of Genetics, University of Georgia, 120 West Green Street, Athens, GA 30602, United States

*Corresponding author. Department of Statistics, Institute of Bioinformatics, University of Georgia, 310 Herty Drive, Athens, GA 30602, United States.
E-mail: lliu@uga.edu

Abstract

Estimating transmission rates is a challenging yet essential aspect of comprehending and controlling the spread of infectious diseases. Various methods exist for estimating transmission rates, each with distinct assumptions, data needs, and constraints. This study introduces a novel phylogenetic approach called transRate, which integrates genetic information with traditional epidemiological approaches to estimate inter-population transmission rates. The phylogenetic method is statistically consistent as the sample size (i.e. the number of pathogen genomes) approaches infinity under the multi-population susceptible-infected-recovered model. Simulation analyses indicate that transRate can accurately estimate the transmission rate with a sample size of 200 ~ 400 pathogen genomes. Using transRate, we analyzed 40,028 high-quality sequences of SARS-CoV-2 in human hosts during the early pandemic. Our analysis uncovered significant transmission between populations even before widespread travel restrictions were implemented. The development of transRate provides valuable insights for scientists and public health officials to enhance their understanding of the pandemic's progression and aiding in preparedness for future viral outbreaks. As public databases for genomic sequences continue to expand, transRate is increasingly vital for tracking and mitigating the spread of infectious diseases.

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

Introduction

Assessing transmission rates is essential for understanding the dynamics of infectious diseases [1, 2]. 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 [3, 4]. Various techniques and approaches have been devised to estimate transmission rates using epidemiological and genetic data in infectious diseases [5–9]. Conventional approaches, such as the computation of

the basic reproduction number R_0 , represent some of the most straightforward techniques for assessing transmission rates. Estimation of R_0 involves examining the growth rate of the epidemic curve under the susceptible-infectious-recovered (SIR) and susceptible-exposed-infectious-recovered (SEIR) models with the assumptions of a consistent transmission rate and a homogeneous population [10]. Several methods have been developed to account for temporal variations in estimating transmission rates [11–13]. The performance of these methods

Skylar Gay is an undergraduate student and ThinkSwiss Scholar conducting research in the Institute of Bioinformatics at the University of Georgia.

Gregory Ellison is a PhD candidate in the Department of Statistics at the University of Georgia.

Jianing Xu is a PhD candidate in the Department of Statistics at the University of Georgia.

Jialin Yang is a PhD candidate in the Department of Statistics at the University of Georgia.

Yiliang Wei is an Associate Professor of Jiangsu Key Laboratory of Phylogenomics and Comparative Genomics in School of Life Sciences at Jiangsu Normal University.

Shaoyuan Wu is a Professor of Jiangsu Key Laboratory of Phylogenomics and Comparative Genomics in School of Life Sciences at Jiangsu Normal University.

Lili Yu is a Professor in the Department of Biostatistics at Georgia Southern University.

Christopher Whalen is Karen and Jim Holbrook Professor of Global Health Institute in College of Public Health at the University of Georgia.

Jonathan Arnold is a Professor in the Department of Genetics at the University of Georgia.

Liang Liu is a Professor in the Department of Statistics at the University of Georgia.

Received: January 17, 2024. Revised: May 2, 2024. Accepted: June 14, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

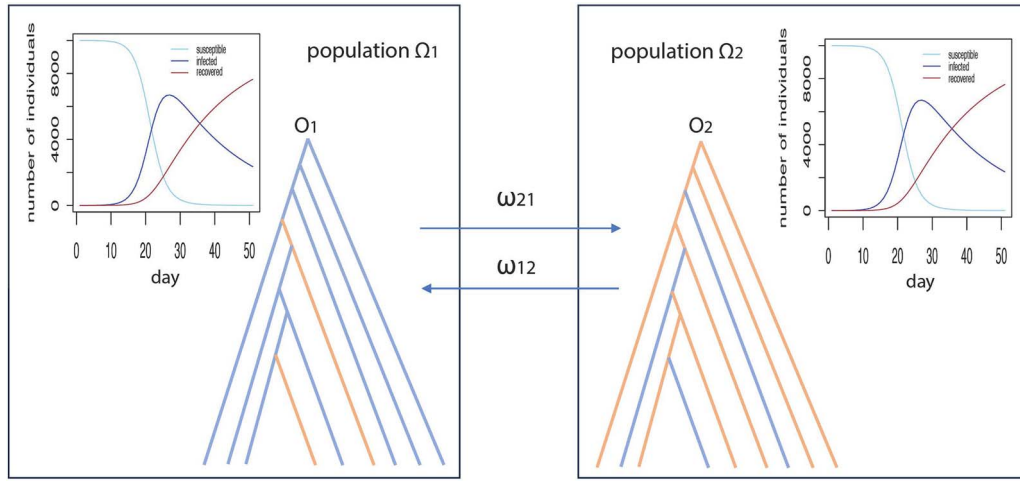


Figure 1. Transmission events generated from the two-population SIR model; the numbers of susceptible we consider the transmission events that occur within and between two populations Ω_1 (left panel) and Ω_2 (right panel); the numbers of susceptible, infected, and recovered individuals at time t (day) 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 the population Ω_2 to the population Ω_1 , and 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 most lineages are the transmissions within the population Ω_1 and two 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 most lineages are the transmissions within the population Ω_2 and three lineages are the inter-population transmissions from the population Ω_1 to the population Ω_2 .

relies on the assumption that symptom onset accurately reflects the date of infection, which might not hold true for cases of asymptomatic transmission [14].

Analyzing epidemiological data offers valuable insights into the transmission dynamics of a disease by fitting models to observed data and estimating relevant parameters [15–18]. Since transmission rates can exhibit spatial variability [19], analyzing the disease spread across different regions is instrumental in understanding the spatial distribution of transmission rates [20, 21]. Furthermore, advancements in genomic sequencing technologies have enabled researchers to trace the dissemination of pathogens at a molecular level [22]. Phylogenetic trees constructed from genetic data have emerged as fundamental tools to elucidate the relatedness of different pathogen strains [23]. By scrutinizing the branching patterns of the tree, researchers can deduce transmission dynamics, including the direction and frequency of transmission events [7].

Traditional methods for estimating transmission rates have predominantly centered on tracking the spread of infectious diseases within a single population. This paper introduces a novel approach by presenting a multi-population SIR model for analyzing transmission rates within and between populations (see Fig. 1). Leveraging this multi-population SIR model, we have developed a phylogenetic method to accurately estimate inter-population transmission rates. Our aim is to furnish a holistic comprehension of transmission dynamics across multiple populations for infectious diseases.

Materials and Methods

Modeling transmissions for multiple populations

The multi-population SIR model, an extension of the traditional SIR model [24], serves to simulate disease transmissions across multiple populations (S1 in Supplementary data). This model can be readily expanded into more intricate epidemiological models. For instance, it can adopt the SEIR model framework, which, compared to the SIR model, provides a more accurate

representation of infectious diseases with discernible incubation periods [25]. However, since infection and recovery rates are two pivotal parameters in the formula of estimating inter-population transmission rates, the multi-population SIR model is deemed sufficient for accurately estimating these rates. 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 $k = 1, \dots, K$. The variables $S_{t,k}, I_{t,k}, R_{t,k}$ in each population satisfy the differential equations of the SIR model for a single population (S1 in Supplementary data), i.e.

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

Moreover, transmissions occur between two populations i and j at a constant rate ω_{ij} for transmissions from population Ω_j to population Ω_i and ω_{ji} for transmissions from population Ω_i to population Ω_j (Fig. 1). The inter-population transmission rate ω_{ij} represents the probability of an individual from population Ω_j traveling to population Ω_i and contracting the infection in Ω_i , i.e.

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

where w is the probability that an individual in population Ω_j travels to population Ω_i , and v is the probability that an individual who has traveled to population Ω_i gets infected in Ω_i . If individuals in population Ω_j independently have the same probability w of traveling to population Ω_i , then the number $y_{t,j}$ of individuals in population Ω_j who travel to population Ω_i at time t follows the binomial distribution, i.e.

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

Given $y_{t,j}$, the number $x_{t,j}$ of individuals in population Ω_j who have traveled to and gotten infected in population Ω_i at time t follows the binomial distribution, i.e.

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

The transmission events occurring in population Ω_i involve not only the infected individuals in the population Ω_i but also the $x_{t,j}$ individuals from population Ω_j who travel to and get infected in Ω_i . Every newly infected individual $\mathcal{I}_{t,i}$ in Ω_i at time t , including the $x_{t,j}$ individuals who have 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 transmission event is indicated by a mapping τ defined as follows:

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

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

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

Moreover, the $x_{t,j}$ individuals from population Ω_j are infected by the $I_{t,i}$ infectious individuals of population Ω_i at time t . We assume that the $I_{t,i}$ infectious individuals at time t are equally likely to 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 represents one of the $I_{t,i}$ infectious individuals

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

All transmissions within an arbitrary time interval $[0, d_i]$ for $d_i > 2$ and $d_i \in \mathbb{N}$ in population Ω_i form a tree-like structure, which is the transmission tree T_i for population Ω_i (Fig. 1). We assume that the roots $O_1, \dots, O_K$ of K transmission trees $T_1, \dots, T_K$ share a common ancestor, denoted by O^* , 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 and 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 $E(x_{t,j}) = E(E(x_{t,j}|y_{t,j})) = N_j I_{t,i} w v = N_j I_{t,i} \omega_{ij}$, 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.

$$\hat{\omega}_{ij} = \frac{\sum_t x_{t,j}}{N_j \sum_t I_{t,i}}. \quad (8)$$

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 $\hat{\omega}_{ij}$ can be calculated by

$$\hat{\omega}_{ij} = \frac{\sum_t x_{t,j}}{N_j \sum_{m=1}^{I_i} (t_{m,i}^R - t_{m,i}^I)}. \quad (9)$$

The estimate $\hat{\omega}_{ij}$ 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 $\Omega_1, \dots, \Omega_K$. We assume that the infected individuals in the samples $\mathcal{S}_1, \dots, \mathcal{S}_K$ are randomly selected from populations $\Omega_1, \dots, \Omega_K$. 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 $\tilde{I}_i$ for $i = 1, \dots, K$ 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.

$$\tilde{\omega}_{ij} = \frac{\frac{1}{\tilde{I}_j} \sum_t \tilde{x}_{t,j}}{N_j \left(\frac{1}{\tilde{I}_i} \sum_{m=1}^{\tilde{I}_i} (t_{m,i}^R - t_{m,i}^I) \right)}. \quad (10)$$

The estimate $\tilde{\omega}_{ij}$ converges to $\hat{\omega}_{ij}$, i.e. $\tilde{\omega}_{ij} \rightarrow \hat{\omega}_{ij}$, 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 $\tilde{\omega}_{ij}$ 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 $\Omega_1, \dots, \Omega_K$. The phylogenetic method for transmission rate estimation consists of four steps: (i) building a phylogenetic tree based on the pathogen genomes. (ii) Identifying clades in the tree that have at least a certain percentage (α) of taxa with the same population origin. By default, α is set at 60%. Nonetheless, users have the flexibility to adjust α such that the count of strongly supported clades aligns with the number of populations. (iii) Labeling each identified clade with the population origin of the majority sequences in that clade. Sequences labeled with a distinct population origin are inferred as instances of inter-population transmission. (iv) 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 $[0, d]$. The population size was designated as $N_1 = N_2 = 10,000$ and $N_1 = N_2 = 1,000,000$. For the population size 10,000, we set the infection rate

$\beta = 5 \times 10^{-5}$ and the recovery rate $\gamma = 0.05$. For the population size 1,000,000, we set $\beta = 5 \times 10^{-7}$ and $\gamma = 0.05$. The values of the infection and recovery rates were based on the estimates obtained from the early stages of the COVID-19 pandemic. An infection rate (β) of 5×10^{-5} for a population of 10,000 suggest rapid infection with the entire population being infected within 30 days (Fig. 1). The recovery rate (γ) of 0.05 indicates a fast recovery with an average recovery time of 20 days. 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 [26]. 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 1000000. 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 10,000 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. $\omega_{12} = wv = 2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7}$ for the population size 10000 and $\omega_{12} = wv = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-9}$ for the population size 1000000, 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, i.e. $\omega_{12} = \omega_{21}$.

To understand the transmission dynamics of other pathogens with varying transmission modes, we simulated data for a population of 10,000 with slower infection and recovery rates, i.e. $\beta = 5 \times 10^{-6}$ and $\gamma = 0.005$. Compared with fast infection, the reduced infection rate of $\beta = 5 \times 10^{-6}$ would require 300 days for the entire population of 10,000 to become infected, while the slower recovery rate of $\gamma = 0.005$ indicated an average recovery period of 200 days (S5 in Supplementary data). 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. The number of individuals who traveled from the population Ω_2 to the population Ω_1 at time $t \in [0, 500]$ was simulated from the binomial distribution ($n = 10000, p = 0.0001$). The inter-population transmission rates ω_{12} and ω_{21} were designated as $\omega_{12} = \omega_{21} = 2 \times 10^{-8}, 4 \times 10^{-8}, 6 \times 10^{-8}, 8 \times 10^{-8}$.

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} included inter-population infections, and 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 coefficient 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. $MSE = \frac{1}{100} \sum_{i=1}^{100} (\hat{\omega}_{12}^i - \omega_{12})^2$ and $CV = \frac{sd(\hat{\omega}_{12})}{mean(\hat{\omega}_{12})}$, where $sd(\hat{\omega}_{12})$ is the standard deviation of $\hat{\omega}_{12}$.

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 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 deoxyribonucleic acid (DNA) sequences of 20,000 base pairs using the phylogenetic program Seq-Gen [27] with the mutation rate $\mu = 0.01, 0.001, 0.0001$. These sequences were utilized to reconstruct the maximum likelihood (ML) tree using FastTree [28]. 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 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 $i, j = 1, \dots, 5$) were set at values of $2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}$ and 8×10^{-7} , while for the larger population (1,000,000 individuals), transmission rates were configured at $2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}$, 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 labeled 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 the transmission rate estimates served as a measure of how well the phylogenetic approach performed in estimating transmission rates in the simulation.

Data analysis of SARS-CoV-2 genomes in the early pandemic

The proposed phylogenetic method transRate was applied to a genomic dataset consisting of 40,028 sequences of SARS-CoV-2 in human hosts during the early SARS-CoV-2 pandemic [29]. This dataset was formed from 41,910 coronavirus genomes retrieved

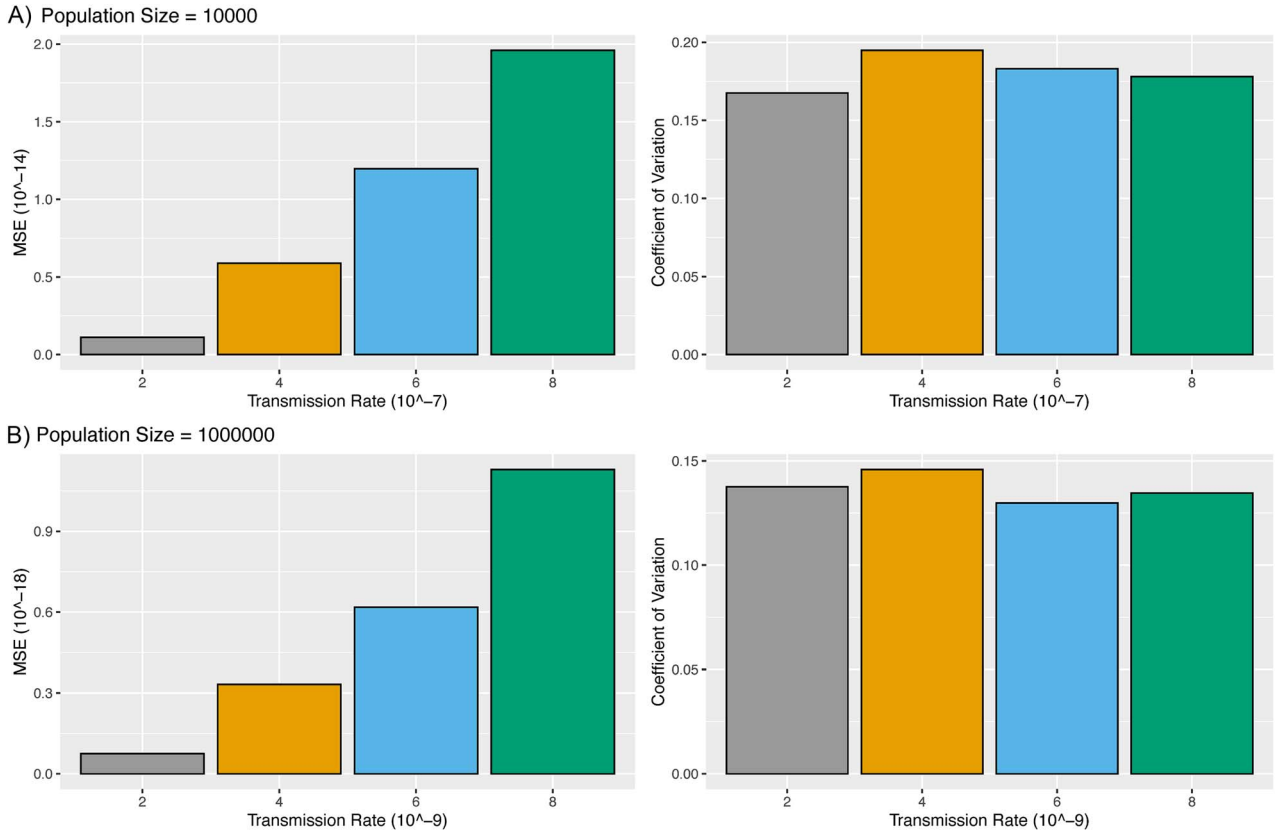


Figure 2. Estimation of the transmission rate from the phylogenetic tree of all infected individuals; the phylogenetic tree of all infected individuals was generated from the two-population SIR model; (a) for population size = 10,000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7}$, and (b) for population size = 1,000,000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-9}$; the phylogenetic tree was then utilized to estimate the transmission rate ω , and the simulation was repeated 100 times; the MSE and CV of the transmission rate estimates were calculated.

from NCBI GenBank nucleotide database on 26 August 2021, with all sequences collected from regions across Asia, Europe, Oceania, Africa, and the Americas between 31 December 2019, and 1 April 2020. The dataset was filtered to remove sequences with missingness, containing frame shift, and incomplete genomes. The filtration returned a dataset of 40,028 SARS-CoV-2 genomes isolated from human hosts. The genomes were annotated with the geographical locations corresponding to where the samples were collected. The sequence datasets for each of the 11 genes in SARS-CoV-2 and the full-genome sequence dataset were constructed and aligned, whereafter ML and bootstrap phylogenetic trees were built from the 11 gene alignments using RAxML [30] with the GTRCAT model. The species tree was estimated from ML and bootstrap gene trees using a coalescent method NJst [31]. The geographical distribution of the clades in the species tree are as follows: 11 clades geographically centered in the Americas, 5 clades geographically centered in Asia, 18 clades geographically centered in Europe, and 1 clade geographically centered in Oceania (S7 in Supplementary data). A number of these clades did not contain any geographical outliers. A group of datasets were formed for the following populations: “Africa” (includes any clades that were centered in countries within the continent of Africa), “Americas” (includes any clades that were centered in countries or localities that were in North, Central, and South America), “Asia” (includes any clades that were centered in countries located in Asia), “Europe” (includes any clades that were centered in countries located in Europe), and “Oceania” (includes any clades that were centered in localities in Oceania). The transmission rate estimation calculation was applied to the above populations. The recovery time is equal to 14 days as

initially issued by the World Health Organization in the early pandemic. Population data was collected based on the World Health Organization and United Nations published data (S8 in Supplementary data).

A transmission analysis airplane plot was constructed using the previously described dataset 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 illustrated in 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 in R.

Results

Simulation

Estimation of transmission rates from phylogenetic trees

In this simulation, we considered two scenarios: (i) the phylogenetic tree was constructed from all transmissions simulated by the two-population SIR model and (ii) the phylogenetic tree was constructed from a sample of transmissions simulated by

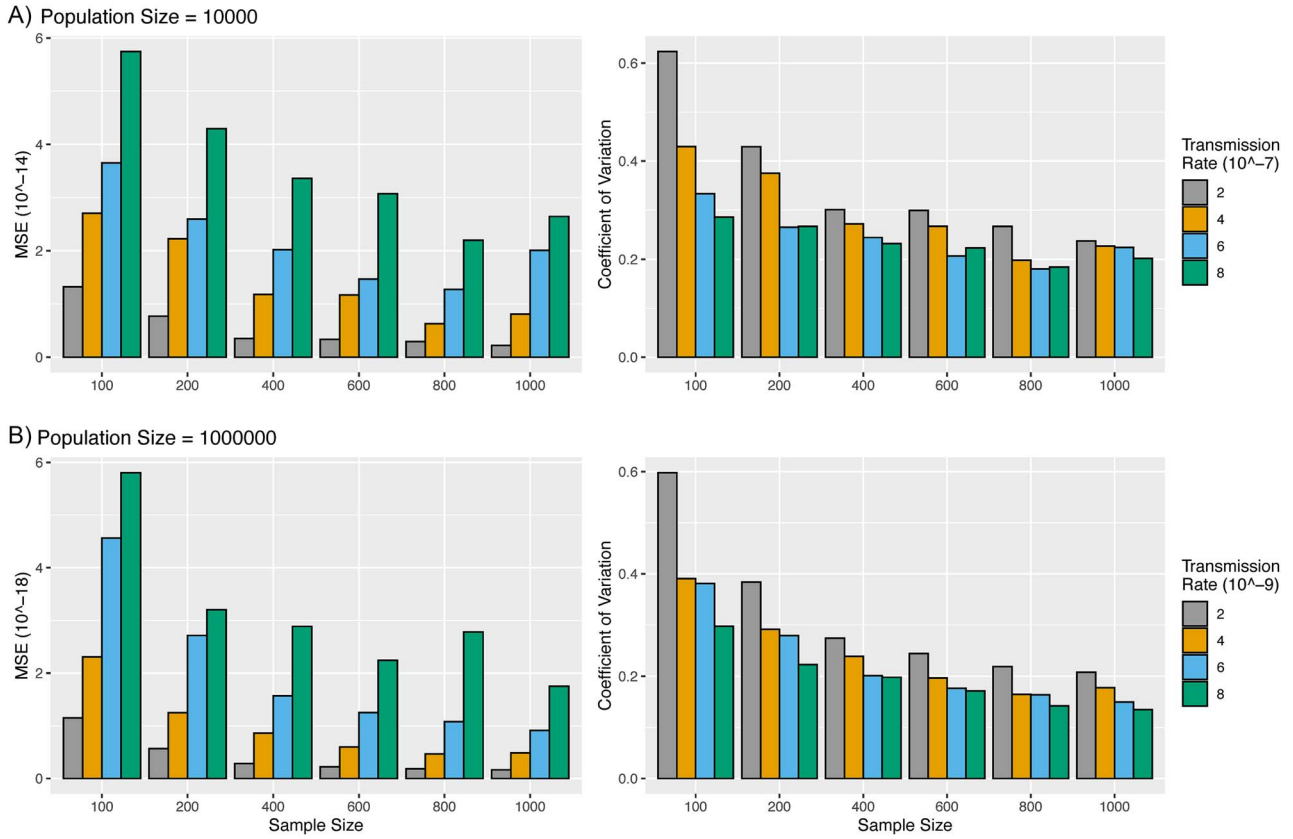


Figure 3. Estimation of the transmission rate from the phylogenetic tree of a sample of infected individuals; the phylogenetic tree of a sample of infected individuals (sample size = 100, 200, 400, 600, 800, 1000) was generated from the two-population SIR model; (a) for population size = 10,000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7}$, and (b) for population size = 1,000,000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-9}$, and the phylogenetic tree was then utilized to estimate the transmission rate ω , and the simulation was repeated 100 times, and the MSE and CV of the transmission rate estimates were calculated.

the two-population SIR model. For Scenario 1, the MSE of the transmission rate estimates increases as the true value of the transmission rate grows from 2×10^{-7} to 8×10^{-7} (Fig. 2). The MSEs for the population size of 10,000 are very small ($< 2 \times 10^{-14}$) (Fig. 2A). Similar results can be observed for the population size of 1,000,000 (Fig. 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 (Fig. 2B) is less than those for the population size of 10,000 (Fig. 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 (Fig. 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 Fig. 3A and 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 (Fig. 3A and b). The CV of the transmission rate estimates is less than 0.45 for the sample size ≥ 200 , suggesting that the sample size 200 is sufficient to accurately estimate the transmission rate when the phylogenetic tree of a sample of transmissions is given.

Similar results are obtained for the simulation with reduced infection and recovery rates. When the phylogenetic tree encompasses all transmissions, the MSEs of transmission rate estimates are below 3×10^{-17} (S6a in Supplementary data). When the phylogenetic tree is derived from a sample of transmissions, the MSEs of transmission rate estimates tend to decrease as the sample size increases (S6b in Supplementary data). Moreover, the MSEs decline more rapidly as the sample size increases from 100 to 400 (S6 in Supplementary data). The CV of transmission rate estimates is less than 0.4 for sample sizes ≥ 200 (S6 in Supplementary data), suggesting that a sample size of 200 is adequate for accurately estimating transmission rates when provided with a sample of transmissions.

Estimation of transmission rates from sequences

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 to reconstruct the 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 grows from 100 to 500 (Fig. 4). The rate of decrease appears to be positively correlated with the values of the transmission rate (Fig. 4). Moreover, the MSE of the transmission rate estimates decreases at a faster rate when the sample size grows from 100 to 200. In contrast, the MSE of

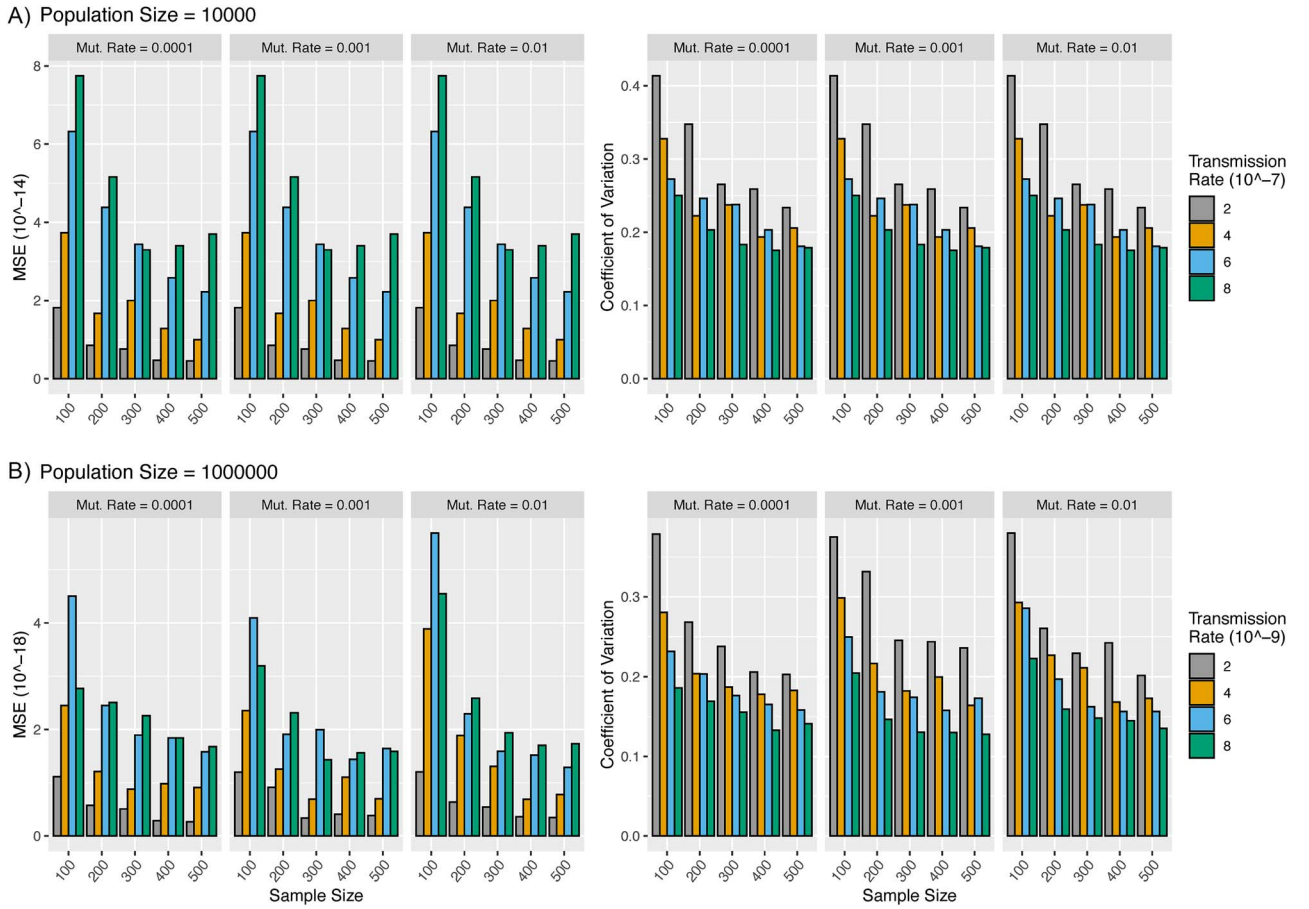


Figure 4. Estimation of the transmission rate from sequences; the phylogenetic tree of a sample of infected individuals (sample size = 100, 200, 300, 400, 500) was generated from the two-population SIR model; (a) for population size = 10,000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7}$, and (b) for population size = 1,000,000 transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-9}$, and 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 ML trees; finally, the transmission rate ω was estimated by transRate using ML trees; the simulation was repeated 100 times; the MSE and CV of the transmission rate estimates $\hat{\omega}$ were calculated.

the transmission rate estimates 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 for accurately estimating the transmission rate. The MSE and CV of 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 $i, j = 1, \dots, 5$ and $i \neq j$ were equal. 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, $\omega_{ij} = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-9}$. 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 (Fig. 5A and 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 for accurately estimating 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 reveal transmission between Europe, Africa, Americas, Oceania, and Asia (Table 1). Population 1 is the population in which the majority of the clade is geographically centered. 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 displays the inter-population transmissions of SARS-CoV-2 in human hosts between 31 December 2019 and 1 April 2020 (Fig. 6). The geographic coordinates are 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

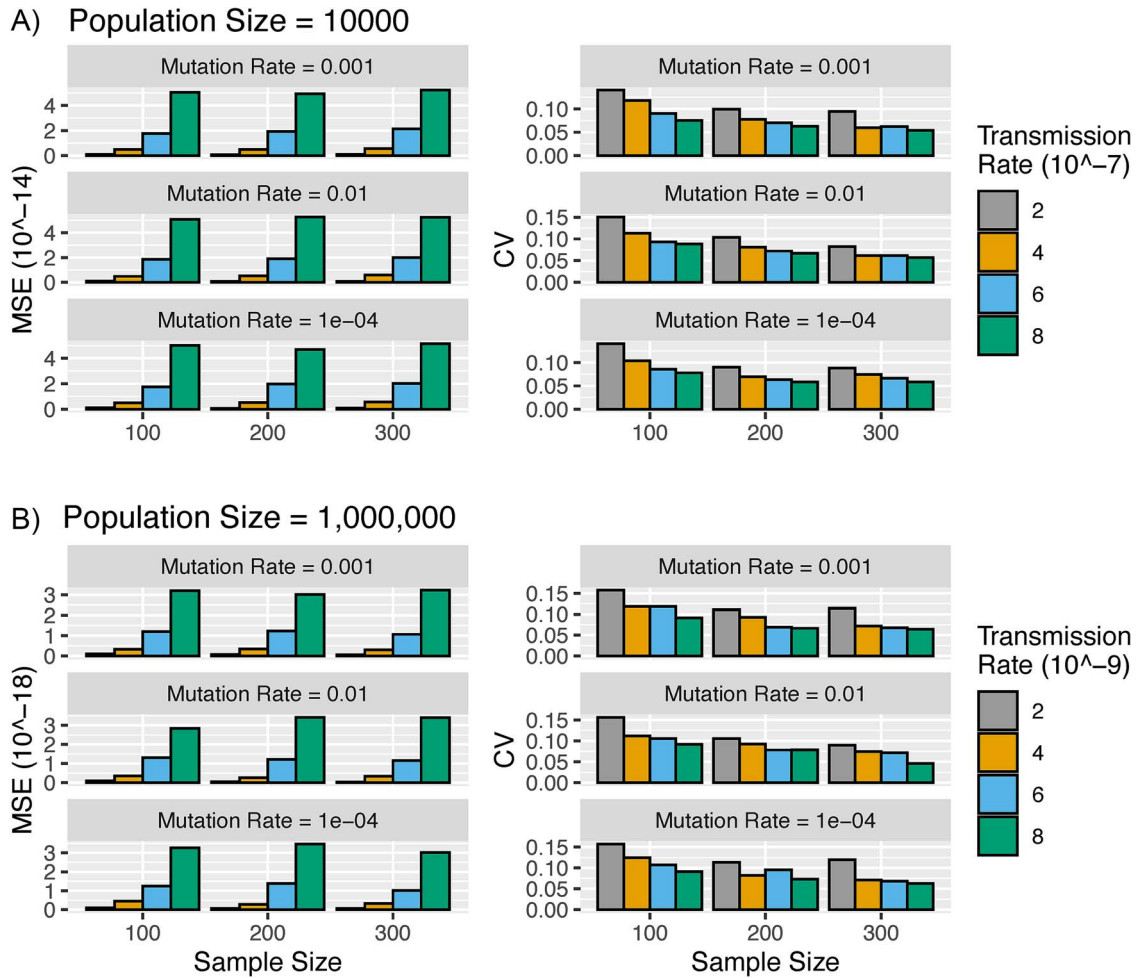


Figure 5. Estimation of the transmission rate for five populations; the phylogenetic tree of a sample of infected individuals (sample size = 100, 200, 300) was generated from the multi-population SIR model for five populations, and (a) for the population size = 10,000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-7}, 4 \times 10^{-7}, 6 \times 10^{-7}, 8 \times 10^{-7}$, and (b) for the population size = 1,000,000, transmission events were simulated with the transmission rate $\omega = 2 \times 10^{-9}, 4 \times 10^{-9}, 6 \times 10^{-9}, 8 \times 10^{-9}$, and 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 ML trees; finally, the transmission rate was estimated by transRate using ML trees; the simulation was repeated 100 times, and the MSE and CV of the transmission rate estimates $\hat{\omega}$ were calculated.

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

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

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. This paper introduces a phylogenetic approach (transRate) to estimate inter-population transmission rates using pathogen genomes. 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 [29]. 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.



Figure 6. An airplane plot of transmission analysis of 40,028 whole genome sequences of SRAS-CoV-2 in human hosts between 31 December 2019 and 1 April 2020; the larger black dots on the map represent the geographical location of clades formed based on bootstrap support values and 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, and 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.

Upon data analysis, there is an alarming lack of data availability during the COVID-19 pandemic. A collection of all publicly available sequence from December 2019 to March 2020 was pulled in late 2020. That totaled less than 6000 samples; however, when the same query was searched in late 2021 nearly 50,000 samples resulted. This slow release of publicly available data is a hindrance in the development of tools such as transRate to assess transmission data. Another critical factor is the presence of asymptomatic cases in viral infections. Asymptomatic individuals may not receive whole genome testing, which introduces further challenges to the accuracy of the phylogenetic method.

The performance of the phylogenetic approach relies on precisely identifying populations through well-supported clades in the inferred phylogenetic tree. Low-quality sequence data can introduce considerable uncertainty in tree estimation, compromising the accuracy of transRate in transmission rate estimation. Moreover, the multi-population SIR model assumes a consistent transmission rate, whereas real-world transmission rates can vary. Hence, adopting a more advanced model that accounts for time-varying transmission rates is essential to accurately capture the evolving dynamics of transmission events.

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

- Development of an innovative phylogenetic method to estimate transmission rates of infectious diseases.
- 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.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Acknowledgements

Thanks to the incredible collaborators for providing key data and information and for making this project possible.

Funding

This work was supported by the National Science Foundation (NSF DBI-2029595, NSF DBI-2243206, and NSF DBI-1946937).

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” [29]. 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

Christopher C. Whalen (Conducted the simulation analysis, Wrote the manuscript), Greg Ellison (Conducted the simulation analysis), Jialin Yang (Wrote R code), Jianing Xu (Conducted the simulation analysis), Jonathan Arnold (Designed the research), Liang Liu (Designed the research, Wrote R code, Conducted the simulation analysis, Wrote the manuscript), Lili Yu (Conducted the simulation analysis), Shaoyuan Wu (Conducted the simulation analysis), Skylar A. Gay (Wrote R code, Wrote and conducted SARS-CoV-2 data analysis, Wrote the manuscript), and Yiliang Wei (Conducted the simulation analysis)

References

1. Andraud M, Grasland B, Durand B. 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. <https://doi.org/10.1098/rsif.2008.0210>.
2. 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**: 172–82. <https://doi.org/10.1016/j.idm.2023.01.002>.
3. Audu RA, Salu OB, Musa AZ. et al. Estimation of the rate of mother to child transmission of HIV in Nigeria. *Afr J Med Med Sci* 2006;**35**:121–4.
4. Aravindakshan A, Boehnke J, Gholami E. et al. 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. <https://doi.org/10.1371/journal.pgph.0000954>.
5. Becker NG, Hasofer AM. Estimating the transmission rate for a highly infectious disease. *Biometrics* 1998;**54**:730–8. <https://doi.org/10.2307/3109779>.
6. 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. <https://doi.org/10.1016/j.epidem.2023.100672>.
7. Stadler T, Kühnert D, Bonhoeffer S. et al. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci U S A* 2013;**110**:228–33. <https://doi.org/10.1073/pnas.1207965110>.
8. Mubayi A. et al. Analytical estimation of data-motivated time-dependent disease transmission rate: an application to Ebola and selected public health problems. *Trop Med Infect Dis* 2021;**6**:141. <https://doi.org/10.3390/tropicalmed6030141>.
9. Kirkeby C, Halasa T, Gussmann M. et al. Methods for estimating disease transmission rates: evaluating the precision of Poisson regression and two novel methods. *Sci Rep* 2017;**7**:9496. <https://doi.org/10.1038/s41598-017-09209-x>.
10. 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**: 779–92. <https://doi.org/10.1093/biostatistics/kxw027>.
11. Ganyani T, Kremer C, Chen D. et al. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Euro Surveill* 2020;**25**:2000257. <https://doi.org/10.2807/1560-7917.ES.2020.25.17.2000257>.
12. 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;**79**:2987–97. <https://doi.org/10.1111/biom.13901>.
13. Lippiello E, Petrillo G, de Arcangelis L. Estimating the generation interval from the incidence rate, the optimal quarantine duration and the efficiency of fast switching periodic protocols for COVID-19. *Sci Rep* 2022;**12**:4623. <https://doi.org/10.1038/s41598-022-08197-x>.
14. Park SW, Cornforth DM, Dushoff J. et al. The time scale of asymptomatic transmission affects estimates of epidemic potential in the COVID-19 outbreak. *Epidemics* 2020;**31**:100392. <https://doi.org/10.1016/j.epidem.2020.100392>.
15. Dabis F, Msellati P, Dunn D. et al. 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* 1992;**7**:1139–48. <https://doi.org/10.1097/00002030-199308000-00027>.
16. Keeling MJ, Hollingsworth TD, Read JM. Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19). *J Epidemiol Community Health* 2020;**74**:861–6. <https://doi.org/10.1136/jech-2020-214051>.
17. Moon SA, Scoglio CM. Contact tracing evaluation for COVID-19 transmission in the different movement levels of a rural college town in the USA. *Sci Rep* 2021;**11**:4891. <https://doi.org/10.1038/s41598-021-83722-y>.
18. Larremore DB, Fosdick BK, Bubar KM. et al. Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys. *Elife* 2021;**10**:e64206. <https://doi.org/10.7554/eLife.64206>.
19. Kuo FY, Wen TH. Assessing the spatial variability of raising public risk awareness for the intervention performance of COVID-19 voluntary screening: a spatial simulation approach. *Appl Geogr* 2022;**148**:102804. <https://doi.org/10.1016/j.apgeog.2022.102804>.
20. Lachish S, Knowles SCL, Alves R. et al. Infection dynamics of endemic malaria in a wild bird population: parasite species-dependent drivers of spatial and temporal variation in transmission rates. *J Anim Ecol* 2011;**80**:1207–16. <https://doi.org/10.1111/j.1365-2656.2011.01893.x>.
21. Oesterholt MJ, Bousema JT, Mwerinde OK. et al. Spatial and temporal variation in malaria transmission in a low endemicity area in northern Tanzania. *Malar J* 2006;**5**:98. <https://doi.org/10.1186/1475-2875-5-98>.
22. Stockdale JE, Susvitasari K, Tupper P. et al. Genomic epidemiology offers high resolution estimates of serial intervals for COVID-19. *Nat Commun* 2023;**14**:4830. <https://doi.org/10.1038/s41467-023-40544-y>.
23. Mohamed W, Ito K, Omori R. Estimating transmission potential of H5N1 viruses among humans in Egypt using phylogeny, genetic distance and sampling time interval. *Front Microbiol* 2019;**10**:2765. <https://doi.org/10.3389/fmicb.2019.02765>.
24. Kerner WO, McKendrick AG. A contribution to the mathematical theory of epidemics. *Proc Math Phys Eng Sci* 1927;**115**: 700–21.
25. Fan Q, Li Q, Chen Y. et al. Modeling COVID-19 spread using multi-agent simulation with small-world network approach. *BMC Public Health* 2024;**24**:672. <https://doi.org/10.1186/s12889-024-18157-x>.
26. Soetaert K, Petzoldt TR, Setzer W. Solving differential equations in R: package deSolve. *J Stat Softw* 2010;**33**:1–25. <https://doi.org/10.18637/jss.v033.i09>.
27. Rambaut A, Grassly NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* 1997;**13**:235–8.
28. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 2009;**26**:1641–50. <https://doi.org/10.1093/molbev/msp077>.
29. Yang J, Skaro M, Chen J. et al. The species coalescent indicates possible bat and pangolin origins of the COVID-19 pandemic. *Sci Rep* 2023;**13**:5571. <https://doi.org/10.1038/s41598-023-32622-4>.
30. Stamatakis A. RAxML-VI-HP: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006;**22**:2688–90. <https://doi.org/10.1093/bioinformatics/btl446>.
31. Liu L, Yu L. Estimating species trees from unrooted gene trees. *Syst Biol* 2011;**60**:661–7. <https://doi.org/10.1093/sysbio/syr027>.