PLOS ONE

RESEARCH ARTICLE

Modeling neutral viral mutations in the spread of SARS-CoV-2 epidemics

Vitor M. Marquioni, Marcus A. M. de Aguiar*

Instituto de Física "Gleb Wataghin", Universidade Estadual de Campinas - UNICAMP, Campinas, SP, Brazil

* aguiar@ifi.unicamp.br

Check for updates

OPEN ACCESS

Citation: Marquioni VM, de Aguiar MAM (2021) Modeling neutral viral mutations in the spread of SARS-CoV-2 epidemics. PLoS ONE 16(7): e0255438. https://doi.org/10.1371/journal.pone.0255438

Editor: Irene Sendiña-Nadal, Universidad Rey Juan Carlos, SPAIN

Received: May 5, 2021

Accepted: July 16, 2021

Published: July 29, 2021

Copyright: © 2021 Marquioni, de Aguiar. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its Supporting information files. Computer codes can be found at https://github.com/vitormmarquioni/Spread-and-Evolution.

Funding: This work was supported by the Sao Paulo Research Foundation (FAPESP - https://fapesp.br/), grants 2019/13341-7 (VMM), 2019/20271-5 and 2016/01343-7 (MAMA), and by Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPq - https://www.gov.br/cnpq/pt-br), grant 301082/2019-7 (MAMA). The funders

Abstract

Although traditional models of epidemic spreading focus on the number of infected, susceptible and recovered individuals, a lot of attention has been devoted to integrate epidemic models with population genetics. Here we develop an individual-based model for epidemic spreading on networks in which viruses are explicitly represented by finite chains of nucleotides that can mutate inside the host. Under the hypothesis of neutral evolution we compute analytically the average pairwise genetic distance between all infecting viruses over time. We also derive a mean-field version of this equation that can be added directly to compartmental models such as SIR or SEIR to estimate the genetic evolution. We compare our results with the inferred genetic evolution of SARS-CoV-2 at the beginning of the epidemic in China and found good agreement with the analytical solution of our model. Finally, using genetic distance as a proxy for different strains, we use numerical simulations to show that the lower the connectivity between communities, e.g., cities, the higher the probability of reinfection.

Introduction

In the late 2019, the world saw the emergence of a new disease, caused by a new type of coronavirus [1] which can cause severe injures to human respiratory system [2]. Since then, we witnessed an uninterrupted worldwide effort in the search for efficient treatments [2, 3], vaccines [4–6] and better understanding of the epidemic parameters and its pathways of spread [7–10].

A great number of SARS-CoV-2 genomes has been sequenced in different countries and regions, allowing scientists to study its genealogy and geographic origins [11]. Different strains have been characterized [12, 13], revealing cases of reinfection [14, 15]. Understanding the mechanisms of mutation and variability in viruses is of utmost importance to forecast forthcoming challenges, e.g. the appearance of other infectious strains or loss of acquired immunity. Mutation rates are usually high in RNA viruses [16] and are important mechanisms for spillover events [16–18]. Although mutations can have significant impact on the virus genetic machinery, leading to more or less infectious strains [19, 20], neutral mutations also occur in non-coding RNA regions or if they result in synonymous changes, that do not alter the corresponding protein. Counting the number of mutations and tracking their spread in the

had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

population is important for tracing pandemic routes through communities (neighborhoods, cities, or countries) and giving clues as to how the virus is moving [21].

Mathematical models of epidemic spreading are crucial to project how the disease will progress and plan intervention strategies, especially in the case of COVID-19 [22–25]. The great majority of epidemic models divide the population into categories, such as susceptible and infected individuals [26, 27]. Details concerning population structure and how different individuals respond to the infection are ignored, allowing the epidemic spreading to be described by differential equations that can be readily interpreted and solved numerically [28]. The SIR model, susceptible-infectious-recovered, is a classic example of this type of simplification and has set the foundations for the development of more detailed descriptions [26]. Important extensions include time dependent contact rates [29] and multiple infectious stages occurring in parallel [30].

One important drawback of the SIR and other related compartmental models is their inability to describe heterogeneity in individual behavior and response to the infection. Some of these features can be introduced with the help of network theory, which provides a framework for modeling explicit population structures [28]. A number of important results were demonstrated in this context, particularly in connection with the distribution of number of contacts among individuals [31]. The representation of individuals as nodes of a network can also be combined with stochastic infection and recovering processes, which might have important consequences for viral diversity [32].

More recently, efforts have been devoted to integrate models of epidemic spreading with population genetics through coalescent theory [33]. This allowed the study of pairwise genetic differences between viral haplotypes, estimation of the viral growth rate [33, 34] and times to most recent common ancestor [35, 36]. Genetic diversity has also been estimated by replacing birth-death models by deterministic epidemic equations [37] or introducing population structure [38]. Multi-strain models were also used to describe how epidemics shape pathogen diversity [39], considering different sources of heterogeneity, such as genotype networks [40] or, as we do here, the structure of the host' contact network [32, 41].

Here we consider an individual-based model for epidemic spreading where the population is represented by nodes of a network and viruses are modeled explicitly by a binary chain representing their RNA. This allows us to combine population structure using network theory, stochastic dynamics of epidemic spreading and population genetics into a single framework. One of the advantages of this formulation is that important epidemic features, such as the structure of social contacts through which contamination occurs, viral transmission rates, individual incubation and recover periods, virus's genome length and mutation rate can be readily included and analysed.

Although many studies have considered imperfect cross-immunity [32, 39–41], in the present model we consider only neutral mutations, which do not alter the immune escape or other viral parameters. This implies that, once the host has developed an immune response against a viral strain, it will have perfect cross immunity against all strains. We also assume that all viruses replicating inside the same host are identical, thus they can be modeled by a single RNA sequence. Viruses of two different hosts, however, can be different due to the mutations that happen randomly and independently at each nucleotide. These assumptions are justified if the periods of incubation and sickness are much shorter than the inverse of the mutation rate and the duration of the epidemic.

We track the spreading of the virus through the population network and compute its diversity by tracking the genetic distance between pairs of viruses along the epidemic propagation. Within this framework, it is possible to study the viral dynamics along different population structures, by changing only the contact network, which is suitable for computational

experiments. In the last decade, interconnected networks have been widely studied in the context of epidemiological models [42–44]. Here we show, as an application, that the connectivity among different communities (represented by modules of a larger interconnected network) changes significantly the viral pairwise distance distribution, suggesting how reinfections could arise if cross-immunity is lost.

Importantly, we derive a recurrence equation for computing the average genetic distance among viruses in the population in terms of the number of susceptible and infected individuals, length of the genome and mutation rate. We also derive a mean-field approximation for this equation that can be added to the usual SIR or SEIR models [45] to estimate the viral genetic evolution in homogeneous populations. Finally, we compare the genetic distance among viruses obtained theoretically from the recurrence equation to the SARS-CoV-2 genomic data, obtained from Chinese epidemic data during the period from 12/23/2019 to 03/24/2020.

The present work is a follow-up of a recently proposed SEIR model designed to study the effects of quarantine regimes [46], from which many parameters are obtained. The paper is organized as follows: in section *The Model*, we describe the SEIR model on networks and how the virus dynamics work. In *Analytical Description* we show how to analytically solve this dynamics for the average genetic distance among viruses. Our solution leads to a discrete equation, which we apply to the SARS-CoV-2 Chinese epidemic data. Taking the continuous time limit we argue that it can be included as a fourth equation to the classic SIR model, enabling one to infer genetic neutral evolution along an epidemic. The mathematical technique we have used can also be implemented in the case of more compartmentalized models. In *Communities and reinfection*, we simulate epidemic spreading along a chain of linearly connected communities and discuss how the risk of reinfection can be increased when the connectivity among them is decreased. This indicates that pandemics are more likely to yield early reinfections than epidemics. We discuss our conclusions in the Section *Conclusions*.

The model

We consider a SEIR individual based model where individuals are divided into four different compartments: *Susceptible*, individuals that can be infected; *Exposed*, individuals that are infected but not infectious; *Infected*, which can spread the virus by infecting others; and *Recovered*, who are recovered from the disease and can no longer be infected. We model the population as a network where nodes represent individuals and links indicate connections between them (linked nodes are also termed first neighbors). Time is discrete and at each step all infected individuals may transmit the disease to their susceptible first neighbors with probability p_I. The infection probability can be calculated as $p_I = R_0/(\tau_0 D)$, where τ_0 is the average time duration of symptoms, R_0, is the basic reproduction number and D the average network degree. Each exposed individual remains in this condition for a time τ distributed according to $\mathcal{P}(\tau)$ (see S1 Appendix), after which it becomes infected. Every infected can recover with a probability $p_R = 1/\tau_0$ per time step [46].

Infected and exposed individuals carry a strain of the virus, represented by a binary chain of size $2B$, where B is the number of nucleotides. Each pair of bits, b_{2i-1} and b_{2i} in the chain ($i = 1, \ldots, B$) represents a nucleotide, given, for instance, by 00 = A, 01 = U, 10 = C and 11 = G. As long as the virus remains hosted in the individual, it can mutate with probability of substitution μ per nucleotide at every iteration. When the virus is passed from one host to another, it is entirely copied to the new host. When the individual recovers, its virus' RNA stops mutating and its final configuration is saved for further analysis. We call this "a final virus". Fig 1 illustrates this dynamics.

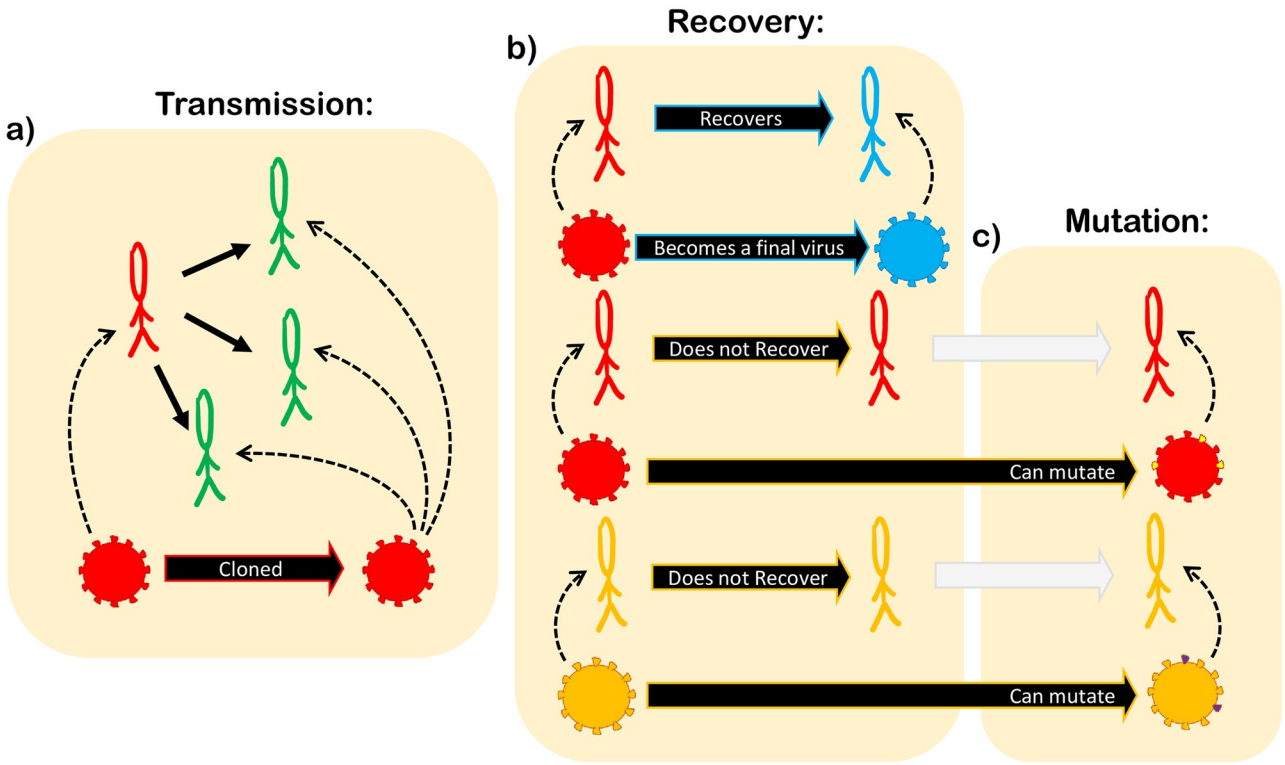

Fig 1. Model dynamics. *(a)* infected individuals (red) can transmit the virus to their susceptible first neighbors (green). When transmission is successful the virus is cloned to the new host, which is now an exposed individual (yellow) and will be able to mutate only in the next iteration. *(b)* infected individuals can recover with probability p_R. When an individual recovers (blue), its virus stops mutating and becomes a "final virus." *(c)* viruses on infected (red) or exposed (yellow) individuals can mutate.

https://doi.org/10.1371/journal.pone.0255438.g001

To compare the different viruses that appear during the simulation we use the Hamming Distance $d^{\alpha\beta}$, which counts the number of different nucleotides between two viruses α and β [47, 48]. In our model the Hamming distance is given by

$$d^{\alpha\beta} = B - \sum_{i=1}^{B} (|b^{\alpha}_{2i-1} - b^{\beta}_{2i-1}| - 1)(|b^{\alpha}_{2i} - b^{\beta}_{2i}| - 1) \quad (1)$$

where $b^{\gamma}_{j} \in \{0, 1\}$ is bit j of the virus γ.

We consider a neutral model for the virus evolution and do not include mechanisms of selection. The mutation probability is the same for all nucleotides, independent of its location in the genome or the nitrogenous base the nucleotide changes from or to. Additionally, once an individual recovers from infection by a strain it acquires perfect cross immunity against all strains.

We start the simulation with a single infected individual with genome $b^{\gamma}_{j} = 1$ for all j. All simulation parameters, can be found in S1 Appendix, and are scaled so that the time unit is one day.

Analytical description

The analysis presented here to calculate the average genetic distance between all viruses, living and final, is suitable for compartmental models in general [45]. Although we develop it to the

SEIR model, it can be applied to other models of this type. From now on we shall abbreviate *average genetic distance* by *average distance* for simplicity.

Single initial infection. Here we assume that the epidemic starts with a single infected individual. Our goal is to compute the average distance d_{t+1} at time $t + 1$ given the average distance d_t at time t. Notice that at the *beginning of iteration* $t + 1$, there are different *kinds* of viruses: those that are *already final* and have ceased to evolve (whose number is R_t); *viruses hosted* in exposed individuals (E_t), thus still evolving; and also those *hosted in infected* individuals (I_t). During the iteration, *new infections* appear (x_t) and some *infected individuals recover* (r_t), and thus do not evolve at this time step. Then, given d_t, we calculate the new average distance *between each kind of virus* which exists at the *end of iteration* $t + 1$, as well as the new average distance *within each kind of virus*.

Given that $\mu \ll 1$, we consider that the probability that two mutations happen in the same nucleotide in the course of the epidemic is negligible. This is a good approximation if the epidemic duration T remains sufficiently small, $\mu T \ll 1$. We also consider that each new infection in the same iteration comes from different hosts, which is valid for $R_0/\tau_0 < 1$, with τ_0 the average duration of symptoms. This means that we do not expect more than one new infection per infected individual in a single iteration. Highly connected nodes, however, can break this assumption, giving rise to super-spreaders. Network heterogeneity, therefore, can show deviations from our estimation. Under these assumptions, the new average distance (at the end of iteration $t + 1$) among the E_t is $d_t + 2B\mu$, once they distanced d_t at the begging of iteration $t + 1$ and evolved along the iteration, each virus getting $B\mu$ mutations. The new average distance between the E_t and the R_t is $d_t + B\mu$, since only the E_t evolved. We emphasize that the approximations used in this section are only for simplification of the analytical equations; the simulations in Section Results and discussion run as previously described.

Once all average pairwise distances have been calculated, d_{t+1} is given by a weighted average, where the weigths are the number of pairs sharing that distance. For instance, the number of pairs between exposed and recovered individuals is $E_t R_t$, while the number of pairs within exposed individuals is $E_t(E_t - 1)/2$.

All distances are calculated in S1 Appendix, and we find the recurrence equation

$$\begin{aligned} d_{t+1} = \frac{1}{Z_t} \quad & (d_t(R_t + E_t + I_t)(R_t + E_t + I_t - 1) \\ + \quad & x_t d_t \left(1 + 2B\mu \frac{R_t}{I_t + E_t + R_t}\right)(x_t - 3 + 2R_t + 2I_t + 2E_t) \\ + \quad & 2B\mu(E_t + I_t - r_t)(E_t + I_t + R_t + x_t - 1)) \end{aligned} \tag{2}$$

where $Z_t = (R_t + E_t + I_t + x_t)(R_t + E_t + I_t + x_t - 1)$, $r_t = R_{t+1} - R_t$ and $x_t = (E_{t+1} - E_t) + (I_{t+1} - I_t) + (R_{t+1} - R_t)$.

Therefore, given the epidemic curves S_t, E_t, I_t and R_t, respectively the Susceptible, Exposed, Infected and Recovered at time t, we can infer the evolution of average genetic distances. Taking the limit of continuous time between events we find the approximation,

$$\dot{d} = \frac{2\dot{S}d\left(1 - B\mu R\left(2 - \frac{3}{N - S}\right)\right)}{(N - S)(N - 1 - S)} + 2B\mu\left(1 - \frac{R}{N - S}\right) \tag{3}$$

where $N - S = I + R + E$ and $\dot{S} = -(\dot{E} + \dot{I} + \dot{R})$. The derivation of this limit is described in S1 Appendix. Since this equation depends only on the continuous curves $S(t)$ and $R(t)$, the initial and final compartment, it can be added to the classic SEIR model to infer the genetic evolution, or to the SIR model, if the exposed compartment is kept empty, meaning that all hosts are

infectious. This result holds if viral evolution occurs in the same way in every intermediate compartment and if every virus passes through all compartments. Adding more compartments with different dynamical behavior or changing the mutation mechanism through different compartments would change the Eqs (2) and (3) but the procedure described in the begging of this section to find d_{t+1} should remain the same.

Multiple initial infections. Eq (2) considers the epidemic starting with a single infected individual. To consider $m > 1$ initial infections, we must include the distance among the m different lineages. Let $\mathfrak{D}_t$ be the average distance among all viruses at time t, $d_t^{(i)}$ the average distance among the viruses of lineage i at time t, $d_0^{(ij)}$ the distance between the initial viruses i and j, and $d_{root,t}^{(i)}$ the average distance at time t of lineage i to the root of lineage i. Thus,

$$\begin{aligned}\mathfrak{D}_t = & \left[\sum_{i=1}^{m} d_t^{(i)}(R_t^{(i)}+E_t^{(i)}+I_t^{(i)})(R_t^{(i)}+E_t^{(i)}+I_t^{(i)}-1)/2\right. \\ & \left.+\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(d_0^{(ij)}+d_{root,t}^{(i)}+d_{root,t}^{(j)})(R_t^{(i)}+E_t^{(i)}+I_t^{(i)})(R_t^{(j)}+E_t^{(j)}+I_t^{(j)})\right] \\ \div & \left[\left(\sum_{i=1}^{m}(R_t^{(i)}+E_t^{(i)}+I_t^{(i)})\right)\left(\sum_{i=1}^{m}(R_t^{(i)}+E_t^{(i)}+I_t^{(i)})-1\right)/2\right]\end{aligned} \tag{4}$$

where $R_t^{(i)}$, $E_t^{(i)}$ and $I_t^{(i)}$ are, respectively, the number of recovered, exposed and infected individuals of lineage i at time t. The first sum represents the distances within each lineage i, while the double sum is due to the distance between each pair of lineages i and j. In this equation, we assume the $\mu \ll 1$ (for coronaviruses, μ lies in the range $\sim[10^{-5}, 10^{-2}]$ per site per year [49]) so that mutations for each virus are unlikely to occur twice at the same nucleotide.

For each lineage i, $d_t^{(i)}$ can be calculated from Eqs (2) or (3) and $d_0^{(ij)}$ must be a given matrix. The distance $d_{root,t}^{(i)}$ can be calculated similarly as Eq (2),

$$d_{root,t+1}^{(i)} = d_{root,t}^{(i)} + \frac{B\mu}{E_t^{(i)}+I_t^{(i)}+R_t^{(i)}+x_t^{(i)}}\left(E_t^{(i)}+I_t^{(i)}-r_t^{(i)}+\frac{4x_t^{(i)}R_t^{(i)}d_{root,t}^{(i)}}{E_t^{(i)}+I_t^{(i)}+R_t^{(i)}}\right) \tag{5}$$

with the continuum limit

$$\dot{d}_{root} = B\mu\left[1-\frac{R^{(i)}}{R^{(i)}+I^{(i)}+E^{(i)}}\left(1-\frac{4d_{root}(\dot{E}^{(i)}+\dot{I}^{(i)}+\dot{R}^{(i)})}{R^{(i)}+I^{(i)}+E^{(i)}}\right)\right] \tag{6}$$

where $R^{(i)}$, $I^{(i)}$ and $E^{(i)}$ are SEIR variables for lineage (i). The details behind these results are described in S1 Appendix.

Viral spread throughout communities

As an application of our model and computational framework, we studied the genetic evolution of a viral spread throughout four weakly and linearly connected communities, i.e., a network with four modules, representing different cities. The goal is to understand how the average genetic distance between viruses in distant communities change if the connectivity between the intermediary communities changes.

We start by generating four independent Barabasi-Albert networks, named 1, 2, 3 and 4. Then, we connect individuals from networks i and $i + 1$ with a *connection probability p in a way they form a line of communities. The Barabasi-Albert network is chosen in order to include heterogeneity in the contact network* [46]. *Finally, we analyse the average genetic distance*

between viruses from cities 1 and 4 for different values of p. The epidemic starts with a single infected individual in city 1 and spreads through the entire network.

Although in our model we always consider that individuals acquire perfect cross-immunity against all strains after being infected the cross-immunity could in principle be lost if a new infecting virus were too different from the original infection. Thus, if the distance between viruses from cities 1 and 4 is large, an infected individual from city 4 that travels to city 1 might reinfect an already recovered individual. Although our simulations do not include this possibility, this is an interesting way to investigate how the risk of reinfection changes due to changes in the network topology.

Results and discussion

Single initial infection

We ran our model for random (Erdos-Renyi) and scalefree (Barabasi-Albert) networks and calculated the average genetic distance. We used networks of 200, 500, 1000 and 4000 nodes, and average degree D of 100 nodes, which was the same for all simulations. In the range of parameters we have used, changing the average degree has two main consequences. First, for large values ($D \gg R_0$), the deviations around the mean of many simulations decreases; and secondly, once the probability of infection is proportional to $1/D$, increasing D delays the peak of infection. We note that the greater the number of connections, the greater the number of attempts to infect neighbors within a single iteration. Thus, we have chosen a value of D that produces reasonably small deviations around the mean and, at the same time, enables fast computation. Changing D in the interval 50 to 200 resulted in no qualitative changes. The infection starts with a single infected individual chosen at random and evolves according to the description in section 2. Fig 2 shows comparisons between the simulated distance and the average distance calculated from Eqs (2) and (3). Each subfigure contains two different simulations and the mean-field solution for that respective set of parameters. We see that Eq (3) approaches Eq (2) only for Erdos-Renyi networks, since only this topology mimics the well-mixed hypothesis considered in mean-field models. Because each genetic evolution curve is calculated from the corresponding epidemic curves, we cannot average over many simulations, thus the error bars are simply the standard deviation of the distribution of distances among all viruses that appeared at that specific simulation time step. Another important feature of this analytical formulation is that, once it is an average description, it does not capture the random appearance or extinction of viral lineages, which can introduce important deviations from our analytical description.

Multiple initial infections

Fig 3 shows the evolution of epidemic in two different cities (non-connected networks of random and scalefree types), each one starting its infection with a single infected individual chosen at random. The evolution in each city is calculated with Eq (2) (pink curves), while the distance between cities 1 and 2 is $d_t^{(1,2)} = d_0^{(1,2)} + d_{root,t}^{(1)} + d_{root,t}^{(2)}$, where $d_{root,t}^{(i)}$ is calculated with Eq (5) (red curve) and the total average distance $\mathfrak{D}_t$ (green curve) is given by Eq (4). The initial distance between the viruses that infected each city is $d_0^{(1,2)} = 0$ in panels (a) and (b), and $d_0^{(1,2)} = 5$ in panels (c) and (d).

The COVID-19 epidemic in China

Eq (2) describes the evolution of average genetic distance between viruses in a single community and depends only on the epidemic curves. It might, therefore, be used to estimate the

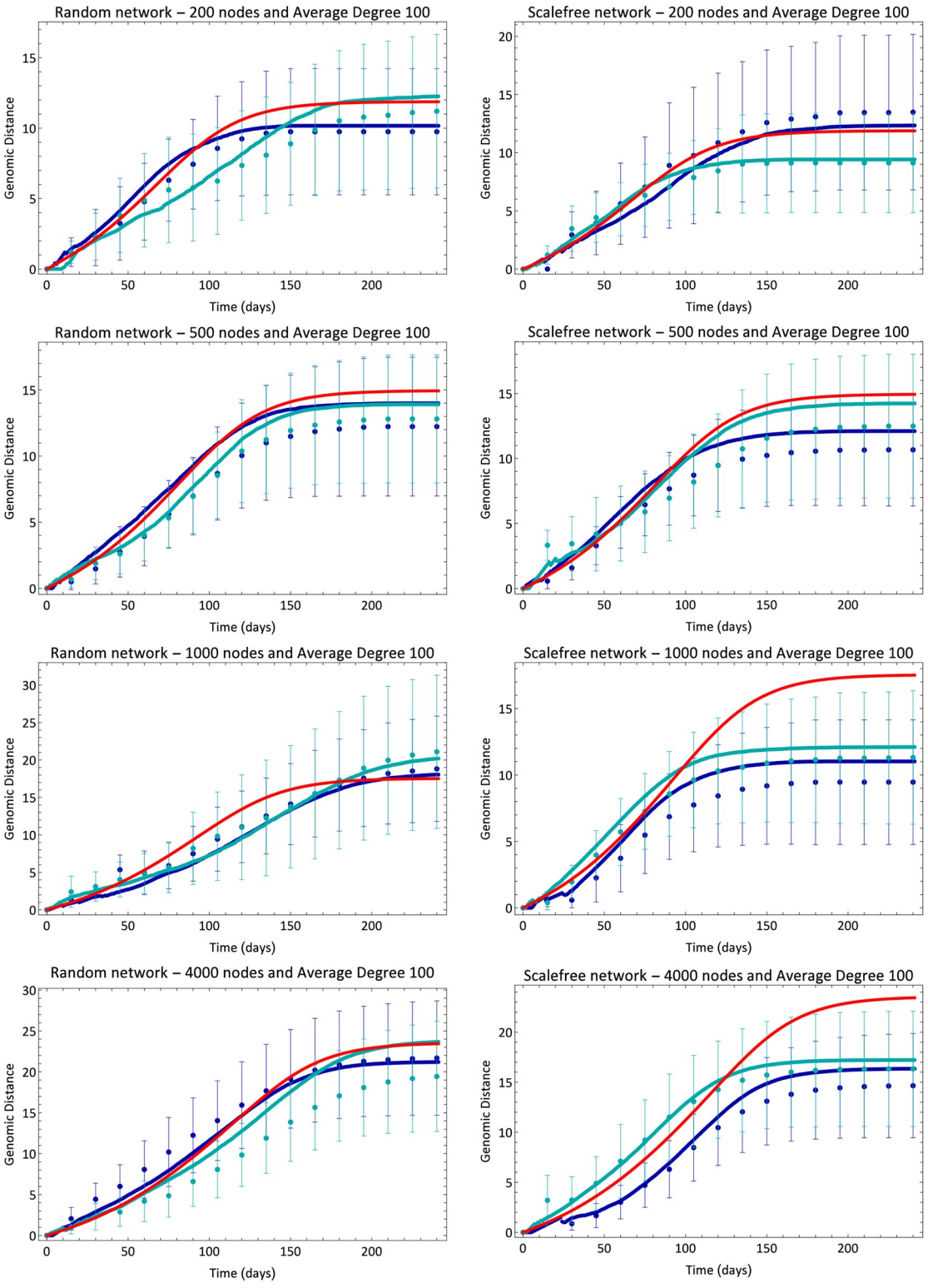

Fig 2. Evolution of average genetic distance. Blue lines and dots are, respectively, analytical (Eq (2)) and simulation results for different simulations. Different shades of blue correspond to different simulations for the same set of parameters. The red line shows the result of mean-field Eq (3). Error bars are standard deviation of the distance distribution in each simulation at each time.

https://doi.org/10.1371/journal.pone.0255438.g002

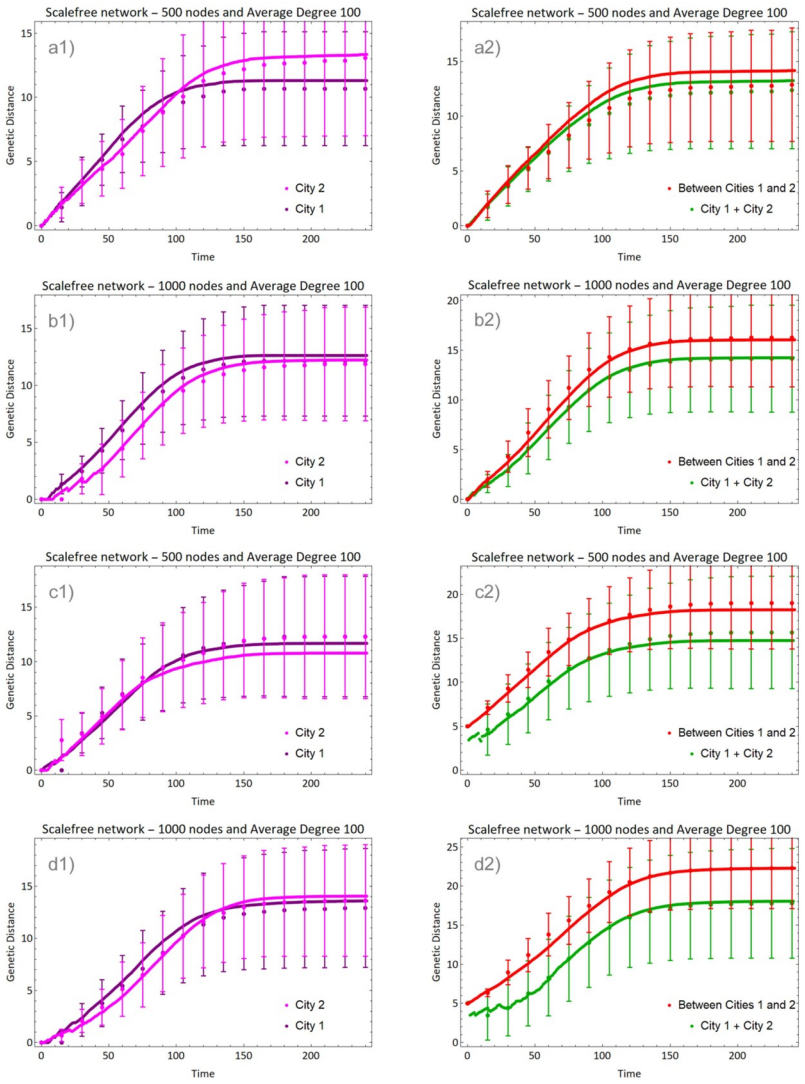

Fig 3. Evolution of average genetic distance in two isolated cities (sizes indicated in the panels). In (a) and (b) the initial viruses were identical and in (c) and (d) they differed by 5 nucleotides. Lines show the average distance within each city (pink), between cities (red) and total average distance (green).

https://doi.org/10.1371/journal.pone.0255438.g003

genetic evolution in real cases. The beginning of COVID-19 epidemic in China is a suitable example, considering the existence of a single patient zero. In any other country, the epidemic may have started with more than one individual, which would require the difficult task of tracking the lineages. The same applies to secondary waves of infection in China.

We obtained Chinese data from the Wolfram Data Repository [50], and corrected it as in reference [51]. Because of the existence of undetected cases, we estimated the real number of cases considering references [51, 52]. Because the number of exposed individuals is not directly available we choose to consider the simpler SIR model in this case. Notwithstanding, because the cases notification started only in January while the epidemic started in December, we extrapolated the data to previous dates, in order to calculate the genetic evolution since patient zero, as we have made in Fig 2. All these data corrections and considerations are described in the Supporting information.

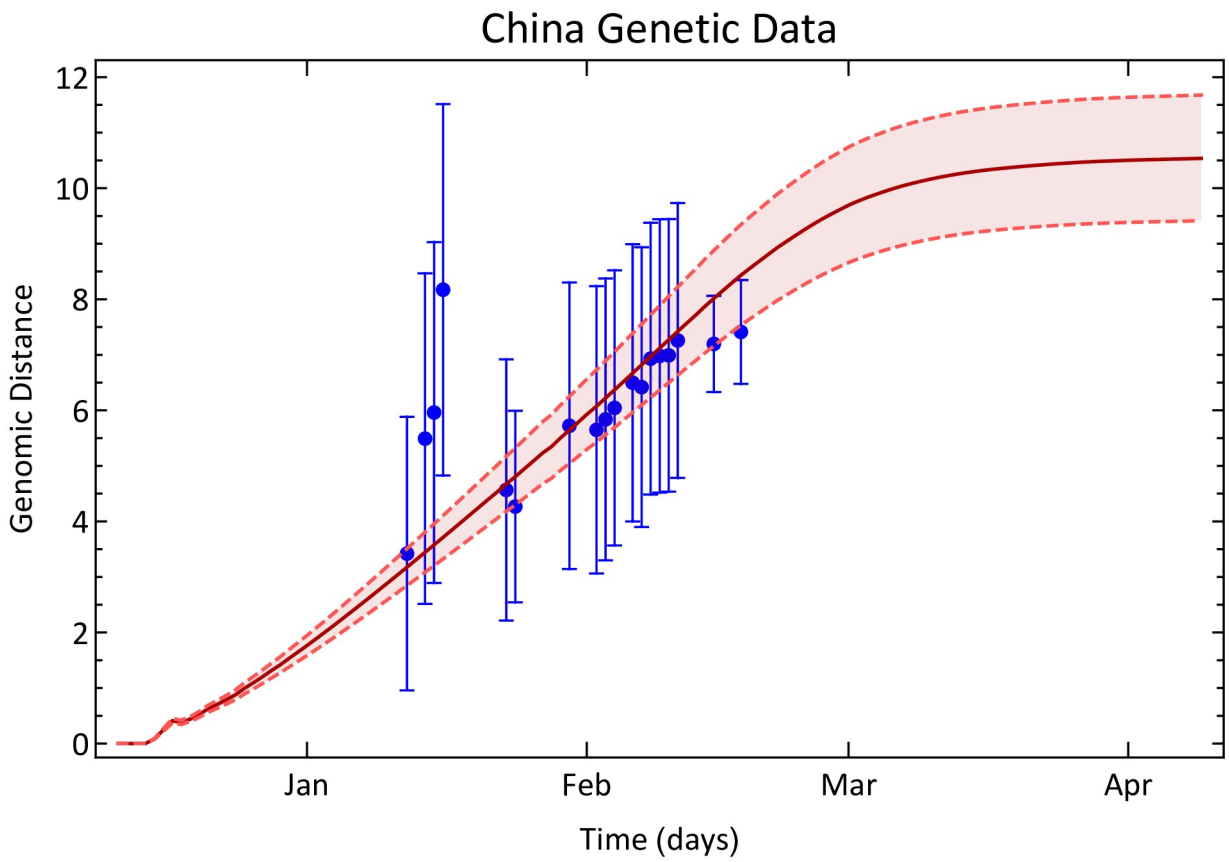

Fig 4. The genetic evolution of SARS-CoV-2 in China. Blue dots are the genetic distance among SARS-CoV-2 inferred from data collected in China between 12/23/2019 and 03/24/2020. The error bars are standard deviation of pairwise distance propagated through the equations. The brown line shows the genetic distance estimated with Eq (2) and the Chinese epidemic data. The interval around the brown curve is a ±10% error interval on the value $B\mu$, which we considered to be $B\mu = 29900 \times 0.001/365$.

https://doi.org/10.1371/journal.pone.0255438.g004

To compare the result of Eq (2) with the real genetic evolution, we used carefully selected 55 real genomes sequenced and collected in China, also available in the Wolfram Data Repository [53]. The Hamming distance between each pair of genomes was obtained by first aligning every two genomes with the Needleman-Wunsch algorithm with score matrix + 1 for match and −1 for mismatch [48]. Then, we considered the Hamming distance between a given pair of genomes as the number of mismatches that are not *indels*, i.e., we considered only nucleotide substitutions. The algorithm to estimate the distance evolution is explained in S1 Appendix, as we also detail the informations of the used genetic data.

Fig 4 shows the result obtained from Eq (2) (brown line) and the estimated genetic evolution (blue dots). The interval around the brown line is an error of ±10% on the product μB, which is the only parameter in the Eq (2). Despite all corrections to the epidemic data and the small number of real genomes we used to infer the real genetic evolution, except for a few points, all the inferred average genetic distances between RNA sequences lie in the predicted interval given by our theoretical model. Because the epidemic in China was readily contained, the average distance d_t saturated.

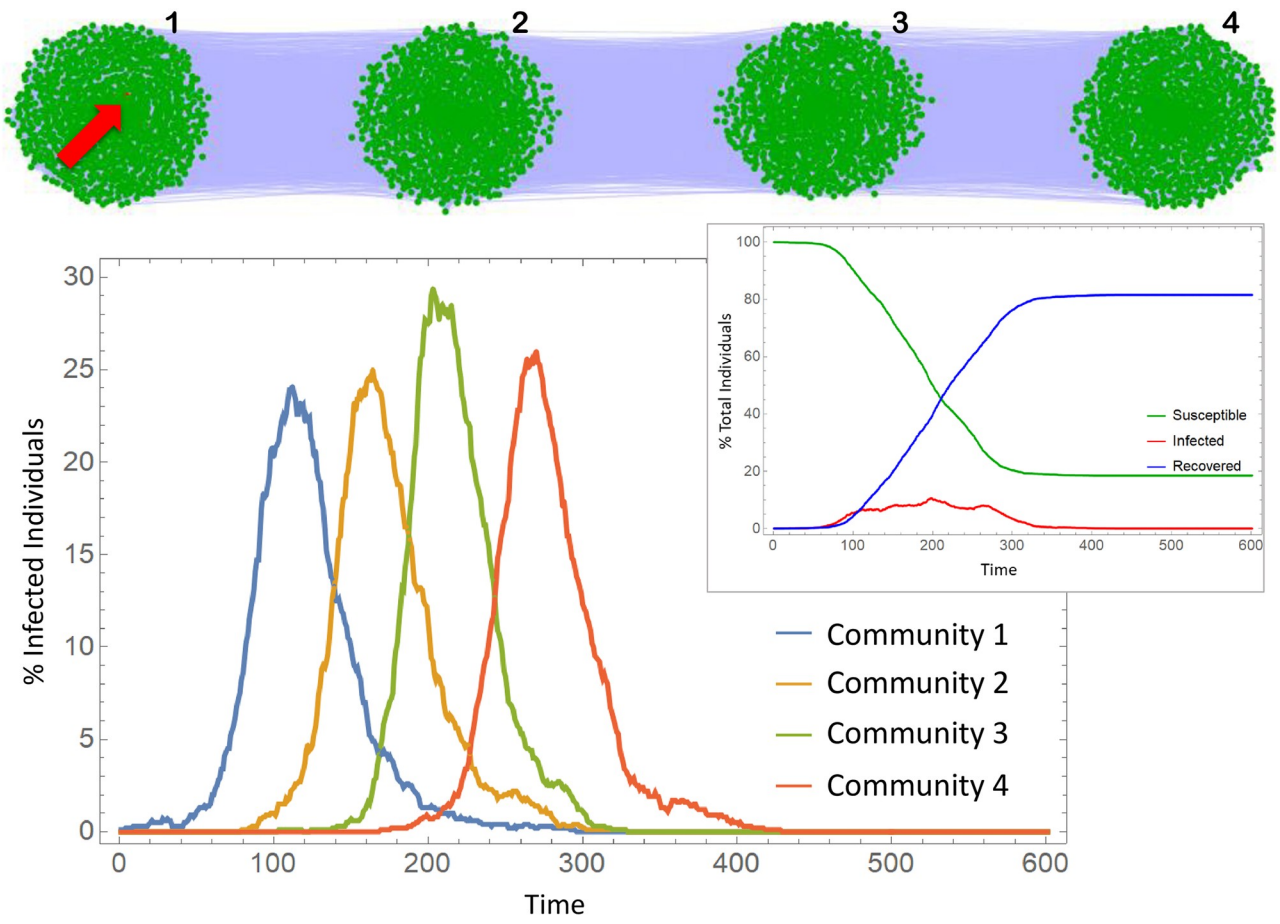

Fig 5. Contact network of four communities on a line and infection curves. Communities are Barabasi-Albert networks with 1000 nodes. We have kept the average degree constant and equal to 100 in all simulations. The infection starts with a single infected individual in the first community (red node indicated with the red arrow). The epidemic parameters are in S1 Appendix.

https://doi.org/10.1371/journal.pone.0255438.g005

Communities and reinfection

In this section, we consider the spread of the epidemic through four communities, representing cities, connected linearly as in Fig 5. The connections within each network are of Barabasi-Albert type, with 1000 nodes and average degree 100 (following the same considerations on average degree already mentioned). Every node from network i can be connected to a node in network $i + 1$ with connection probability p. Once p is small (ranging from 0.0005 to 0.0035) the degree distribution is not considerably distorted from a scale-free one. Fig 5 shows an example of the contact network. From left to right, we number the communities, or cities, from 1 to 4. The epidemic starts with a single infection in city 1 and spread through the entire network. Fig 5 also shows the Infection curves obtained from a simulation. The infection peak delay from one city to other is responsible for the plateau-type curve of total infections.

To analyse the genetic evolution in this system we simulated the dynamic until the epidemic was over and calculated the Hamming distance between every pair of final genomes α and β, constructing the distance matrix $d^{\alpha\beta}$ (Fig 6). Viruses are ordered according to their position in the line, i.e., first the genomes from city 1, then those from the city 2, and so on. We calculated the average distances D_{i-j} between the final genomes from cities i and j and compared with D_{i-i}, the average distance within city i.

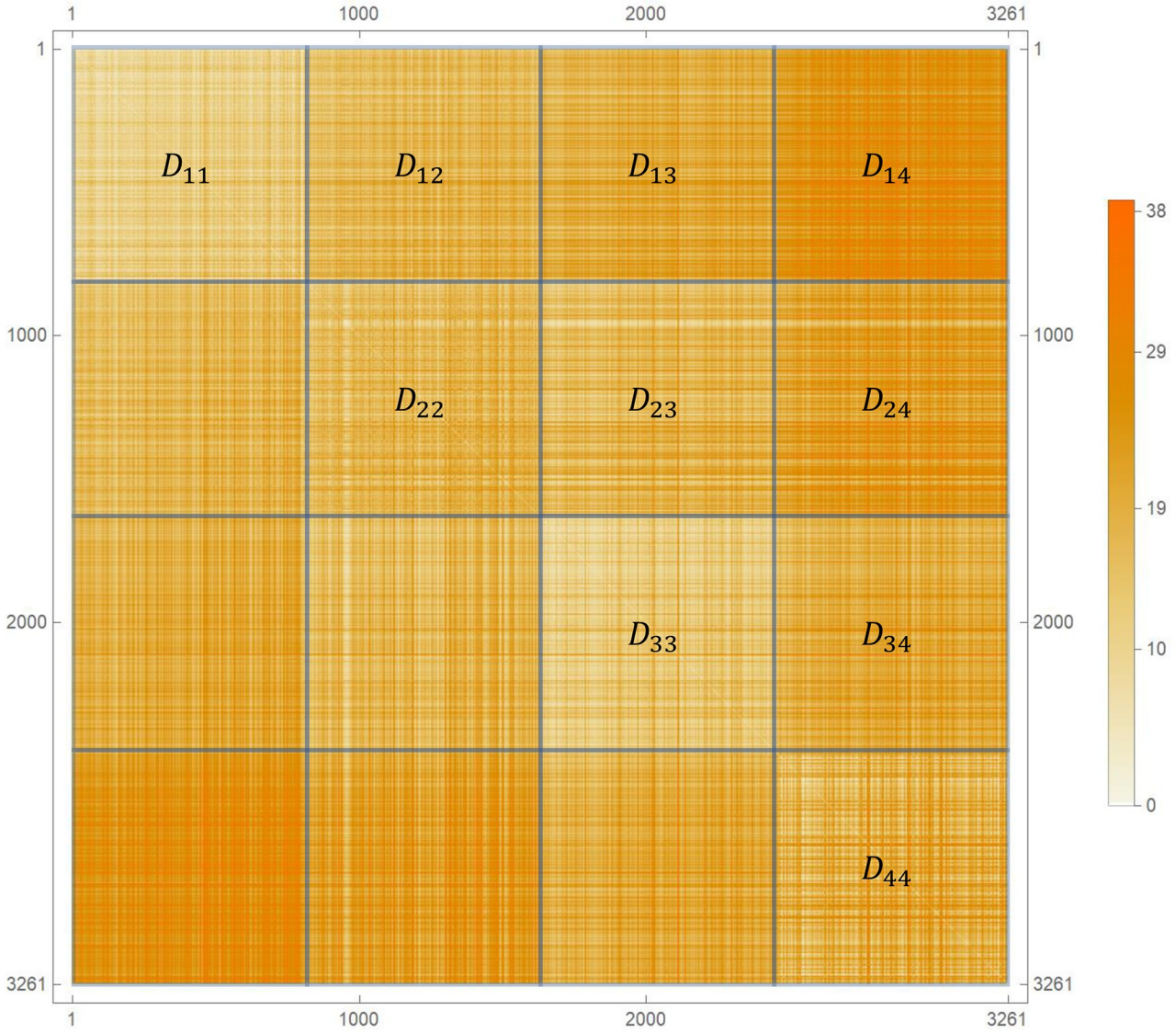

Fig 6. Hamming distance between pairs of viruses. The distance matrix is sorted by the city. Diagonal blocks show the distance between the viruses from a single city, while the non-diagonal blocks are the distances between the viruses from different cities.

https://doi.org/10.1371/journal.pone.0255438.g006

As a null model, we run the epidemic over a single Barabasi-Albert network wih the total size of the 4 cities. City *i*, in this case, means the i-th quarter of the infected nodes. We plot the results of the null model as $p = 0$ in Figs 7 and 8 for comparison. The single network behaves very differently from the four module network, not showing the same interesting results we find for the communities.

Fig 7 shows the ratio D_{4-4}/D_{4-1} as a function of the connection probability p. The results are averages over 20 different simulations for 7 different values of p. When p is small, $D_{4-4}/D_{4-1} < 1$, meaning that the viruses from city 4 are, in average, closer to each other than they are to the viruses from city 1. When p increases, the ratio D_{4-4}/D_{4-1} approaches 1, indicating that the viruses from city 4 are so close to each other as they are to viruses from city 1.

In order to understand the origin of this effect we analyse the infection trees in each case (Fig 7, left). Each node in the trees represents a recovered individual and is connected upwards

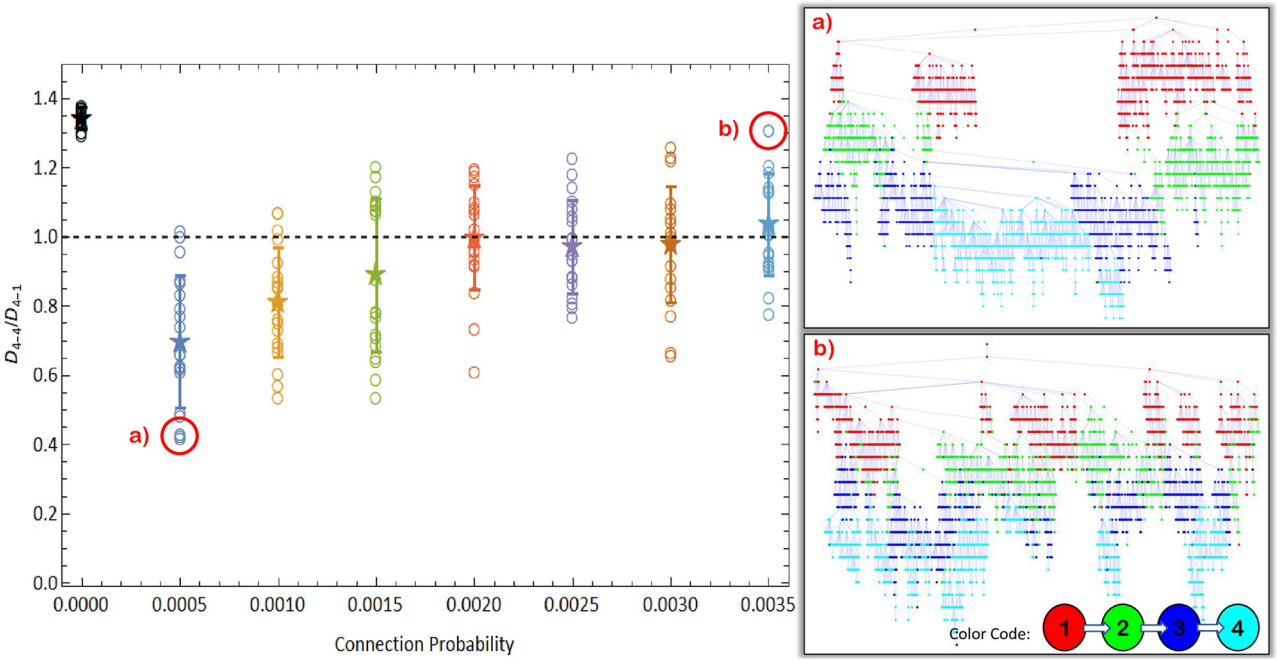

Fig 7. Ratio between the average distance in city 4 and the average distance between cities 1 and 4. Right panels show infection trees for the simulations highlighted with red circles. Open circles show results for individual simulations, the star is the average over 20 simulations and error bars are standard deviations. $p = 0$ represents a single Barabasi-Albert network with 4000 nodes (see text). Nodes in infection trees represent infected individuals, colored according to its city. City 4 (cyan) in panel (a), where $D_{4-4}/D_{4-1} < 1$, was almost entirely infected by a single viral lineage, while in panel (b) where $D_{4-4}/D_{4-1} > 1$, it was infected by many different viral lineages.

https://doi.org/10.1371/journal.pone.0255438.g007

with whoever infected it. Colors represent cities and it is possible to count how many initial infections each city had along the epidemic, i.e., how many lineages has infected each city. When p is small, very few lineages were responsible for infecting city 4 but for higher values of p, this number increases. This is expected, since more connected communities should have more infection gates. This result is a consequence of the founder effect, i.e., only a few individuals, "the founders", give rise to a new population in the new location [12, 54]. However, the system passes through a non-trivial bistable point. When $p = 0.0015$, the values of D_{4-4}/D_{4-1} accumulate around two different values, one above 1 and another below 1. In this case the average is not a good descriptor of the actual system behaviour and there is a competition between different lineages infecting city 4. In simulations where $D_{4-4}/D_{4-1} > 1$, many lineages were successful in infecting the city 4, whereas when $D_{4-4}/D_{4-1} < 1$, only a few did so successfully.

Fig 8 shows the values D_{4-4} and D_{4-1} obtained in each simulation. The average over simulations of the average distance within the forth city D_{4-4} (highlighted blue circles) does not change considerably with p (around $D \approx 21$ nucleotides). Under a neutral evolutionary perspective, viruses will belong to different strains if they differ by more than G nucleotides, where G is a parameter whose value depends on the virus [47, 55]. If $D > G$, viruses in city 4 would belong, on average, to different strains when compared to city 1. As an example, if $G = 26$ new strains would arise, on average, in city 4 for $0 < p \leq 0.0010$, allowing a recovered individual from city 1 to be reinfected by an infected individual from city 4 if they are put in contact with each other (by travelling, for instance). Therefore, there is an increased risk of reinfection due to low connectivity among communities. In this sense, pandemics are more likely to originate new strains than epidemics, as they affect far more distant (therefore less

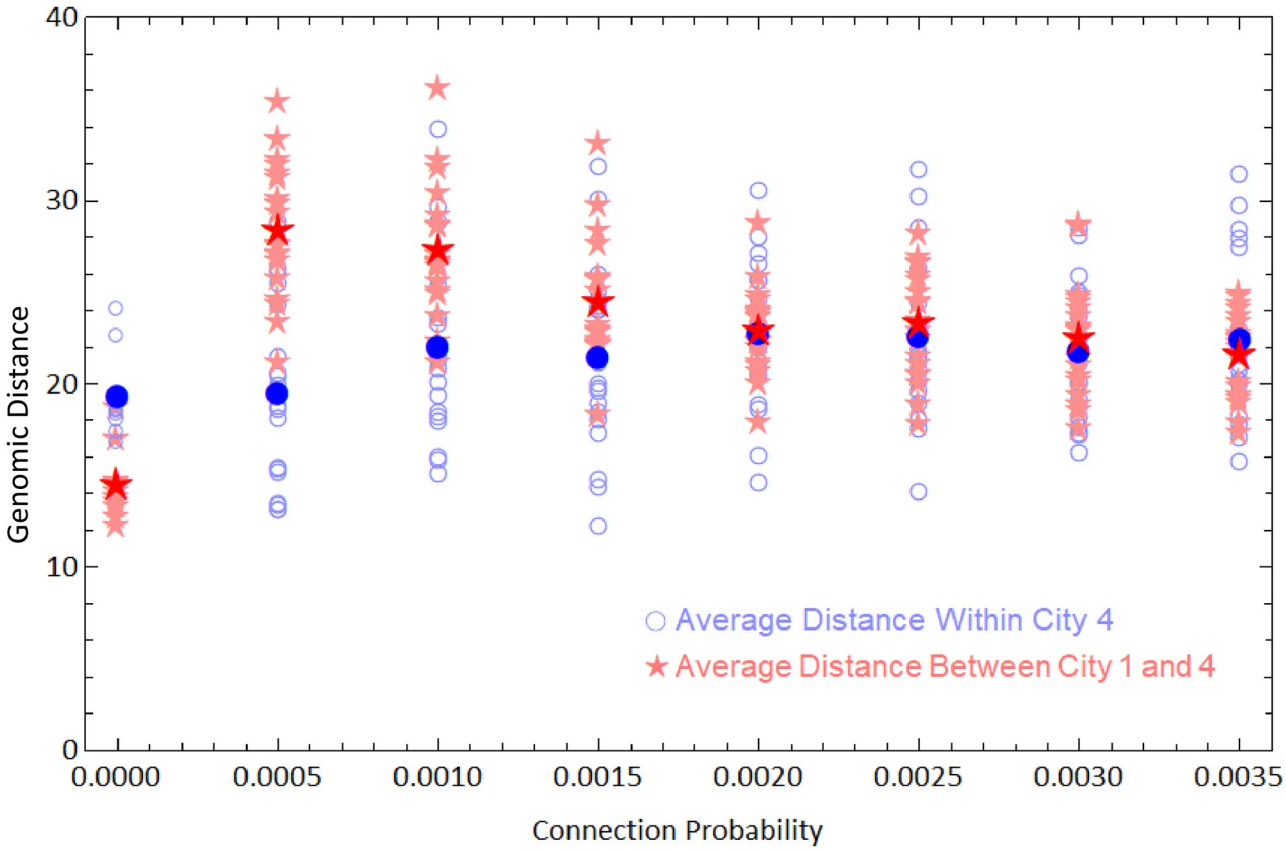

Fig 8. Average genetic distances within cities 1 and 4. Open blue circles are average distance between the viruses of city 4 from a single simulation, and the filled blue circle is average of these values. Light red stars are average distances between viruses from cities 1 and 4 and the dark red star is the average of these values. We ran 20 simulations for each value of connection probability.

https://doi.org/10.1371/journal.pone.0255438.g008

connected) communities. One confirmed case of reinfection by COVID-19 in Hong-Kong had the virus differing by 24 nucleotides from the first infecting virus [14]. This distance matches a value for G for which the network connectivity would strongly influence the rise of reinfections.

Conclusions

We have introduced an individual based model to describe the genetic evolution of a RNA-virus epidemic spreading. We used the SEIR model with four compartments on networks, but the evolutionary dynamics can be implemented in more compartmentalized epidemic models. We provided an analytical description that can be generalized for models with more compartments. An important result of this study is the mean-field approximation, Eq (3), for the evolution of the average genetic distance, which can be added directly to the mean-field SIR or SEIR models.

Our analytical description of the average genetic distance between viruses is neutral and depends only on the epidemic curves. This allows us to project the evolutionary scenario without using the actual genome sequences. Deviations from these predictions in genetic data could reveal the strength of selection or network effects. We compared our prediction using only fifty complete genomes sequenced and collected in China and found good agreement.

We have also analysed the genetic evolution of the epidemic when it spreads over different communities. By changing the connection probability p between 4 linearly arranged communities we investigated how different the viruses infecting city 4 would be from their ancestors in city 1. Our simulations showed that when p is sufficiently small, the genetic difference between these viruses can be quite large, spanning 30 loci. This could allow an infected individual from city 4 to reinfect a recovered individual from city 1. This is a consequence of the founder's effect, which is stronger if p is small as it decreases the number of infection gates of a community. Therefore, we expect increased risk of reinfection from contacts between travelling individuals living in distant territories.

Although the computational framework we described for the viral evolution is neutral, it can be adapted to including other evolutionary aspects, such as differential fitness for mutations in certain genome regions or loss of cross-immunity. These and other features are important topics to be added and studied in future works.

Supporting information

S1 Appendix. Simulation parameters, analytical calculations, real genetic evolution algorithm and Chinese epidemic data corrections.
(PDF)

S1 Table. All Chinese genome sequences. All genomes registered in Wolfram Repository "Genetic Sequences for the SARS-CoV-2 Coronavirus" with complete NucleotideStatus and human Host from China (data accessed 19/08/2020).
(PDF)

S2 Table. Included sequences sorted by collection date. All informations according to S1 Table.
(PDF)

S3 Table. Genome information used to calculate points in Fig 4. We have used a 14 days time window, i. e., every sequenced genome within an interval of 14 days were considered as infected ones, while the previous were considered to be recovered.
(PDF)

Acknowledgments

We are greatful to Dr. Débora Princepe, Dr. Flávia D. Marquitti and Luis F.P.P.F. Salles for critical readings and suggestions. We also thank the anonymous reviewers for the many valuable comments and suggestions.

Author Contributions

Conceptualization: Vitor M. Marquioni, Marcus A. M. de Aguiar.

Formal analysis: Vitor M. Marquioni.

Methodology: Vitor M. Marquioni, Marcus A. M. de Aguiar.

Software: Vitor M. Marquioni.

Supervision: Marcus A. M. de Aguiar.

Writing – original draft: Vitor M. Marquioni, Marcus A. M. de Aguiar.

Writing – review & editing: Vitor M. Marquioni, Marcus A. M. de Aguiar.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine. 2020;. https://doi.org/10.1056/NEJMoa2001017
2. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. Jama. 2020; 324(8):782–793. https://doi.org/10.1001/jama.2020.12839 PMID: 32648899
3. Sanders JM, Monogue ML, Jodlowski TZ, Cutrell JB. Pharmacologic treatments for coronavirus disease 2019 (COVID-19): a review. Jama. 2020; 323(18):1824–1836. PMID: 32282022
4. Lurie N, Saville M, Hatchett R, Halton J. Developing Covid-19 vaccines at pandemic speed. New England Journal of Medicine. 2020; 382(21):1969–1973. https://doi.org/10.1056/NEJMp2005630 PMID: 32227757
5. Le TT, Andreadakis Z, Kumar A, Roman RG, Tollefsen S, Saville M, et al. The COVID-19 vaccine development landscape. Nat Rev Drug Discov. 2020; 19(5):305–306. https://doi.org/10.1038/d41573-020-00073-5 PMID: 32887942
6. Graham BS. Rapid COVID-19 vaccine development. Science. 2020; 368(6494):945–946. https://doi.org/10.1126/science.abb8923 PMID: 32385100
7. Backer JA, Klinkenberg D, Wallinga J. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. Eurosurveillance. 2020; 25 (5):2000062. https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062 PMID: 32046819
8. Wu JT, Leung K, Bushman M, Kishore N, Niehus R, de Salazar PM, et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nature Medicine. 2020; 26(4):506–510. https://doi.org/10.1038/s41591-020-0822-7 PMID: 32284616
9. Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung Sm, et al. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. Journal of clinical medicine. 2020; 9(2):538. https://doi.org/10.3390/jcm9020538 PMID: 32079150
10. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PloS one. 2020; 15(3): e0231236. https://doi.org/10.1371/journal.pone.0231236 PMID: 32231392
11. of the International CSG, et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nature Microbiology. 2020; 5(4):536. https://doi.org/10.1038/s41564-020-0695-z
12. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences. 2020; 117(17):9241–9243. https://doi.org/10.1073/pnas.2004999117
13. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution. 2020; p. 104351. https://doi.org/10.1016/j.meegid.2020.104351 PMID: 32387564
14. To KKW, Hung IFN, Ip JD, Chu AWH, Chan WM, Tam AR, et al. COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 strain confirmed by whole genome sequencing. Clinical infectious diseases. 2020;. https://doi.org/10.1093/cid/ciaa1275 PMID: 32840608
15. Tillett RL, Sevinsky JR, Hartley PD, Kerwin H, Crawford N, Gorzalski A, et al. Genomic evidence for reinfection with SARS-CoV-2: a case study. The Lancet infectious diseases. 2020;. https://doi.org/10.1016/S1473-3099(20)30764-7 PMID: 33058797
16. Duffy S. Why are RNA virus mutation rates so damn high? PLoS biology. 2018; 16(8):e3000003. https://doi.org/10.1371/journal.pbio.3000003 PMID: 30102691
17. Froissart R, Doumayrou J, Vuillaume F, Alizon S, Michalakis Y. The virulence–transmission trade-off in vector-borne plant viruses: a review of (non-) existing studies. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010; 365(1548):1907–1918. https://doi.org/10.1098/rstb.2010.0068 PMID: 20478886
18. Hawley DM, Osnas EE, Dobson AP, Hochachka WM, Ley DH, Dhondt AA. Parallel patterns of increased virulence in a recently emerged wildlife pathogen. PLoS Biol. 2013; 11(5):e1001570. https://doi.org/10.1371/journal.pbio.1001570 PMID: 23723736
19. Stacey BC, Gros A, Bar-Yam Y. Eco-Evolutionary Feedback in Host–Pathogen Spatial Dynamics. arXiv preprint arXiv:11103845. 2013;.
20. de Aguiar MAM, Rauch E, Bar-Yam Y. Mean-field approximation to a spatial host-pathogen model. Physical Review E. 2003; 67(4):047102. https://doi.org/10.1103/PhysRevE.67.047102 PMID: 12786532
21. Kupferschmidt K. Genome analyses help track coronavirus' moves; 2020.

22. Cobey S. Modeling infectious disease dynamics. Science. 2020; 368:713–714. https://doi.org/10.1126/science.abb5659 PMID: 32332062

23. L F S Scabini *et al.* Social Interaction Layers in Complex Networks for the Dynamical Epidemic Modeling of COVID-19 in Brazil; 2020.

24. Vasconcelos G L et al. Modelling fatality curves of COVID-19 and the effectiveness of intervention strategies. medRxiv. 2020;. https://doi.org/10.7717/peerj.9421 PMID: 32612894

25. S Flaxman, S Mishra, A Gandy et al. Report 12: The Global Impact of COVID-19 and Strategies for Mitigation and Suppression. Imperial College London. 2020;.

26. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london Series A, Containing papers of a mathematical and physical character. 1927; 115(772):700–721.

27. Anderson RM. The epidemiology of HIV infection: variable incubation plus infectious periods and heterogeneity in sexual activity. Journal of the Royal Statistical Society: Series A (Statistics in Society). 1988; 151(1):66–93. https://doi.org/10.2307/2982185

28. Keeling MJ, Eames KT. Networks and epidemic models. Journal of the Royal Society Interface. 2005; 2 (4):295–307. https://doi.org/10.1098/rsif.2005.0051 PMID: 16849187

29. Kuznetsov YA, Piccardi C. Bifurcation analysis of periodic SEIR and SIR epidemic models. Journal of mathematical biology. 1994; 32(2):109–121. https://doi.org/10.1007/BF00163027 PMID: 8145028

30. Korobeinikov A. Global properties of SIR and SEIR epidemic models with multiple parallel infectious stages. Bulletin of mathematical biology. 2009; 71(1):75–83. https://doi.org/10.1007/s11538-008-9352-z PMID: 18769976

31. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Physical review letters. 2001; 86(14):3200. https://doi.org/10.1103/PhysRevLett.86.3200 PMID: 11290142

32. Buckee CO, Koelle K, Mustard MJ, Gupta S. The effects of host contact network structure on pathogen diversity and strain structure. Proceedings of the National Academy of Sciences. 2004; 101(29):10839–10844. https://doi.org/10.1073/pnas.0402000101 PMID: 15247422

33. Slatkin M, Hudson RR. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics. 1991; 129(2):555–562. https://doi.org/10.1093/genetics/129.2.555 PMID: 1743491

34. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SD. Phylodynamics of infectious disease epidemics. Genetics. 2009; 183(4):1421–1430. https://doi.org/10.1534/genetics.109.106021 PMID: 19797047

35. Griffiths RC, Tavaré S. Ancestral inference in population genetics. Statistical science. 1994; p. 307–319.

36. De Maio N, Wu CH, Wilson DJ. SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. PLoS computational biology. 2016; 12(9):e1005130. https://doi.org/10.1371/journal.pcbi.1005130 PMID: 27681228

37. Volz EM. Complex population dynamics and the coalescent under neutrality. Genetics. 2012; 190 (1):187–201. https://doi.org/10.1534/genetics.111.134627 PMID: 22042576

38. Gordo I, Gomes MGM, Reis DG, Campos PR. Genetic diversity in the SIR model of pathogen evolution. PloS one. 2009; 4(3):e4876. https://doi.org/10.1371/journal.pone.0004876 PMID: 19287490

39. Kucharski AJ, Andreasen V, Gog JR. Capturing the dynamics of pathogens with many strains. Journal of mathematical biology. 2016; 72(1):1–24. https://doi.org/10.1007/s00285-015-0873-4 PMID: 25800537

40. Williams BJ, St-Onge G, Hébert-Dufresne L. Localization, epidemic transitions, and unpredictability of multistrain epidemics with an underlying genotype network. PLoS Computational Biology. 2021; 17(2): e1008606. https://doi.org/10.1371/journal.pcbi.1008606 PMID: 33566810

41. Buckee C, Danon L, Gupta S. Host community structure and the maintenance of pathogen diversity. Proceedings of the Royal Society B: Biological Sciences. 2007; 274(1619):1715–1721. https://doi.org/10.1098/rspb.2007.0415 PMID: 17504739

42. Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S. Catastrophic cascade of failures in interdependent networks. Nature. 2010; 464(7291):1025–1028. https://doi.org/10.1038/nature08932 PMID: 20393559

43. Dickison M, Havlin S, Stanley HE. Epidemics on interconnected networks. Physical Review E. 2012; 85 (6):066109. https://doi.org/10.1103/PhysRevE.85.066109 PMID: 23005164

44. Saumell-Mendiola A, Serrano MÁ, Boguná M. Epidemic spreading on interconnected networks. Physical Review E. 2012; 86(2):026106. https://doi.org/10.1103/PhysRevE.86.026106 PMID: 23005824

45. Murray JD. Mathematical biology: I. An introduction. vol. 17. Springer Science & Business Media; 2007.

46. Marquioni VM, de Aguiar MAM. Quantifying the effects of quarantine using an IBM SEIR model on scalefree networks. Chaos, Solitons & Fractals. 2020; 138:109999. https://doi.org/10.1016/j.chaos.2020.109999 PMID: 32834581

47. De Aguiar MA. Speciation in the Derrida–Higgs model with finite genomes and spatial populations. Journal of Physics A: Mathematical and Theoretical. 2017; 50(8):085602. https://doi.org/10.1088/1751-8121/aa5701

48. Sung WK. Algorithms in bioinformatics: A practical introduction. CRC Press; 2009.

49. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang YP, et al. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC evolutionary biology. 2004; 4(1):21. https://doi.org/10.1186/1471-2148-4-21 PMID: 15222897

50. Wolfram Research. Epidemic Data for Novel Coronavirus COVID-19; 2020. Wolfram Data Repository https://doi.org/10.24097/wolfram.04123.data.

51. Ivorra B, Ferrández MR, Vela-Pérez M, Ramos A. Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. Communications in nonlinear science and numerical simulation. 2020; 88:105303. https://doi.org/10.1016/j.cnsns.2020.105303 PMID: 32355435

52. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science. 2020; 368(6490):489–493. https://doi.org/10.1126/science.abb3221 PMID: 32179701

53. Research W. Genetic Sequences for the SARS-CoV-2 Coronavirus; 2020. Wolfram Data Repository https://doi.org/10.24097/wolfram.03304.data.

54. Ruan Y, Luo Z, Tang X, Li G, Wen H, He X, et al. On the founder effect in COVID-19 outbreaks–How many infected travelers may have started them all? National Science Review. 2020;.

55. Costa CL, Marquitti FM, Perez SI, Schneider DM, Ramos MF, de Aguiar MA. Registering the evolutionary history in individual-based models of speciation. Physica A: Statistical Mechanics and its Applications. 2018; 510:1–14. https://doi.org/10.1016/j.physa.2018.05.150