Modeling viral mutations in the spread of epidemics

Vitor M. Marquioni[1], Marcus A. M. de Aguiar[1*]

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

* aguiar@ifi.unicamp.br

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.

Author summary

In this work we describe the genetic evolution of viruses in the course of an epidemic. The viruses are described by their RNA, modeled as a finite sequence of loci with four possible entries representing nucleotides. Viruses mutate at a fixed rate and we assume that genetic variations do not confer differential fitness, meaning that infected individuals acquire perfect cross immunity against all viral strains. Individuals in the population are represented by nodes of a network of contacts. We compute the diversity of viral population, measured by genetic distance between viral sequences, defined as number of loci bearing different nucleotides. We derive an equation for the evolution of the average genetic distance that depends only on epidemic variables, such as the number of infected and recovered individuals, number of nucleotides and mutation rate. We apply this equation to the beginning of SARS-CoV-2 epidemic in China and show that it agrees well with the available data. We also show how the genetic variability is affected when the virus spreads over connected communities, influencing the probabilities 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 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. As an application, we show that the connectivity among different communities (represented by modules of a larger 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 [42] 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 [43], 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 Appendix), after which it becomes infected. Every infected can recover with a

probability $p_R = 1/\tau_0$ per time step [43].

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.
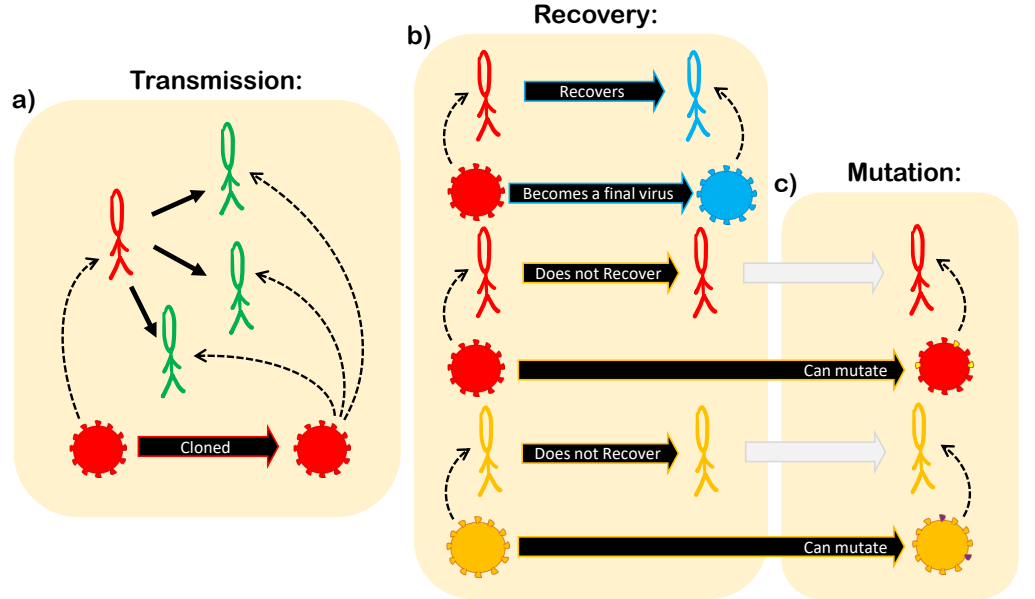

Fig 1. Model dynamics. *(a)* infected individuals (red) can transmit the virus to their susceptible first neighbors (green). When transmission is successful the virus is cloned to the new host, which is now an exposed individual (yellow) and will be able to mutate only in the next iteration. *(b)* infected 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.

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 [44, 45]. In our model the Hamming distance is given by

$$d^{\alpha\beta} = B - \sum_{i=1}^{B} \left(|b_{2i-1}^{\alpha} - b_{2i-1}^{\beta}| - 1 \right) \left(|b_{2i}^{\alpha} - b_{2i}^{\beta}| - 1 \right) \tag{1}$$

where $b_j^{\gamma} \in \{0, 1\}$ 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 $b_j^{\gamma} = 1$ 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 [42]. 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

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

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 are 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 equations (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,

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

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 [46]) 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 Eq.(2) or Eq.(3) and $d_0^{(ij)}$ must be a given matrix. The distance $d_{root,t}^{(i)}$ can be calculated similarly as Eq.(2),

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

with the continuum limit

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

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 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 [43]. 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 of 100 nodes. 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 Eq.(2) and Eq.(3). Each subfigure contains two different simulations and the mean-field solution for that respective set of parameters. We see that 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
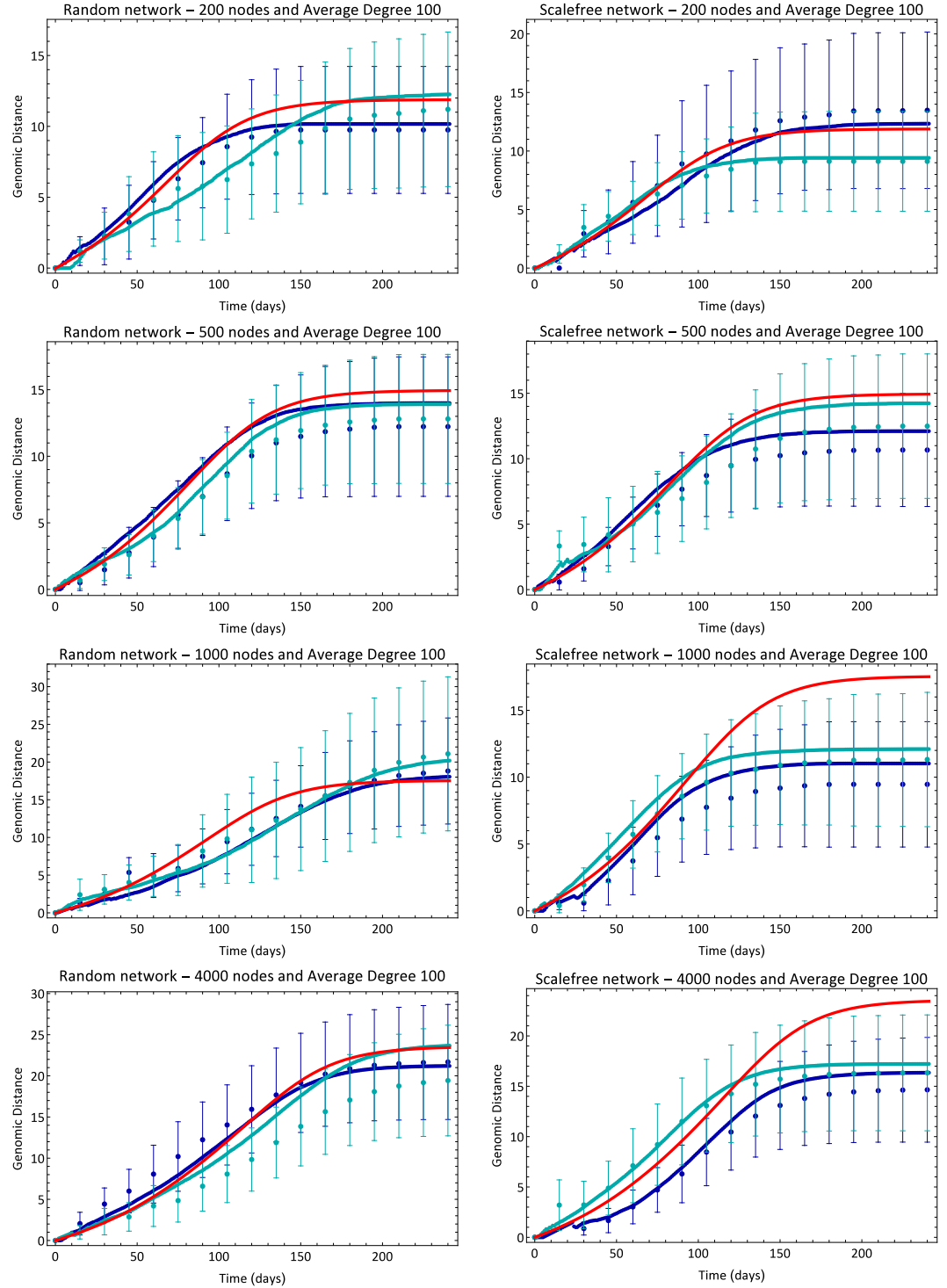
Fig 2. Evolution of average genetic distance. Blue lines and dots are, respectively, analytical (Eq.(2)) and simulation results for different simulations. Different shades of blue correspond to different simulations for the same set of parameters. The red line shows the result of mean-field Eq.(3). Error bars are standard deviation of the distance distribution in each simulation at each time.
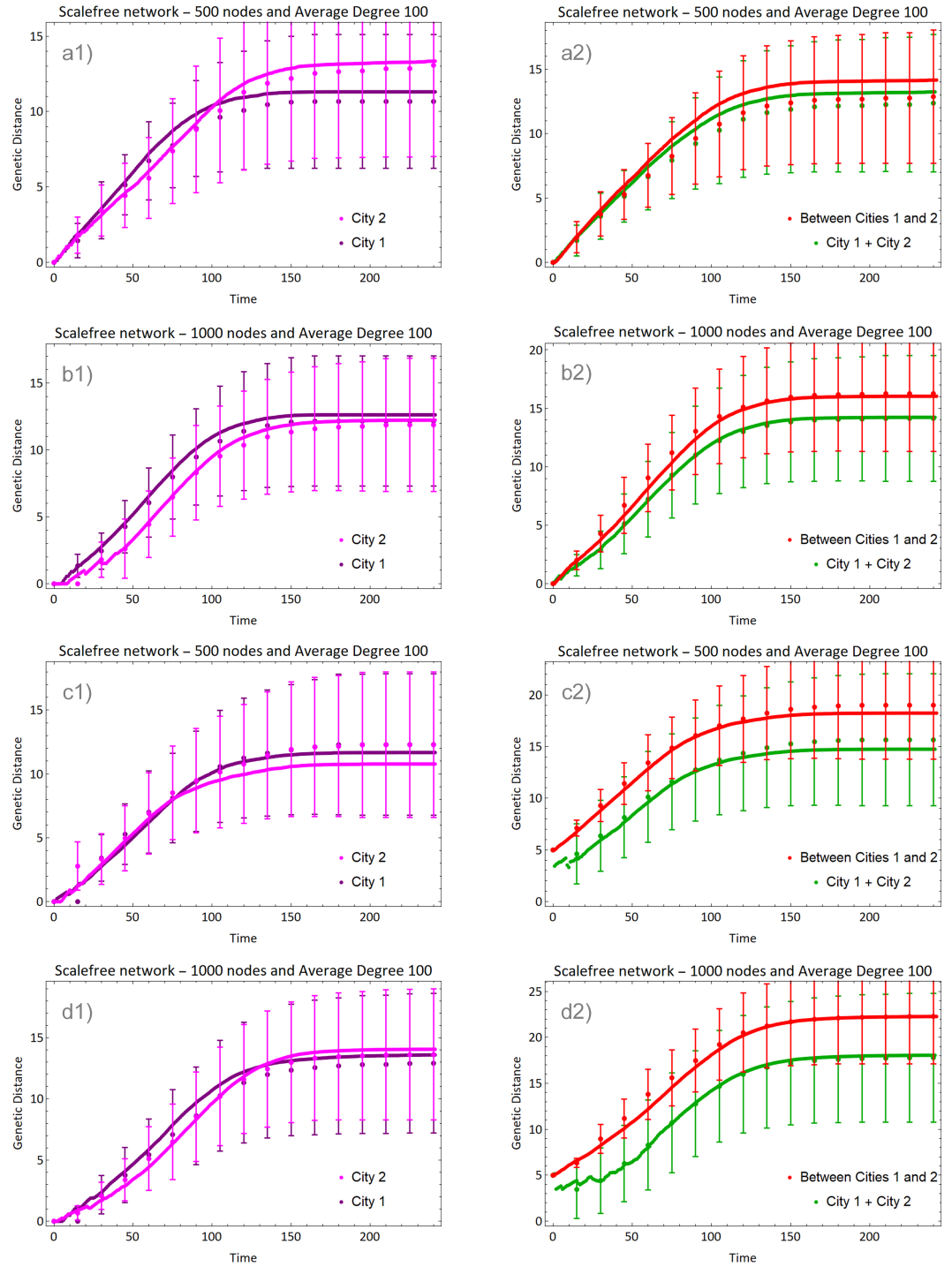
Fig 3. Evolution of average genetic distance in two isolated cities (sizes indicated in the panels). In (a) and (b) the initial viruses were identical and in (c) and (d) they differed by 5 nucleotides. Lines show the average distance within each city (pink), between cities (red) and total average distance (green).

estimate the 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 [47], and corrected it as in reference [48]. Because of the existence of undetected cases, we estimated the real number of cases considering references [48, 49]. 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.

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 [50]. The Hamming distance between each pair of genome was obtained by first aligning every two genomes with the Needleman-Wunsch algorithm with score matrix $+1$ for match and -1 for mismatch [45]. 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 $\pm10\%$ on the product μB, which is the only parameter in the equation (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.
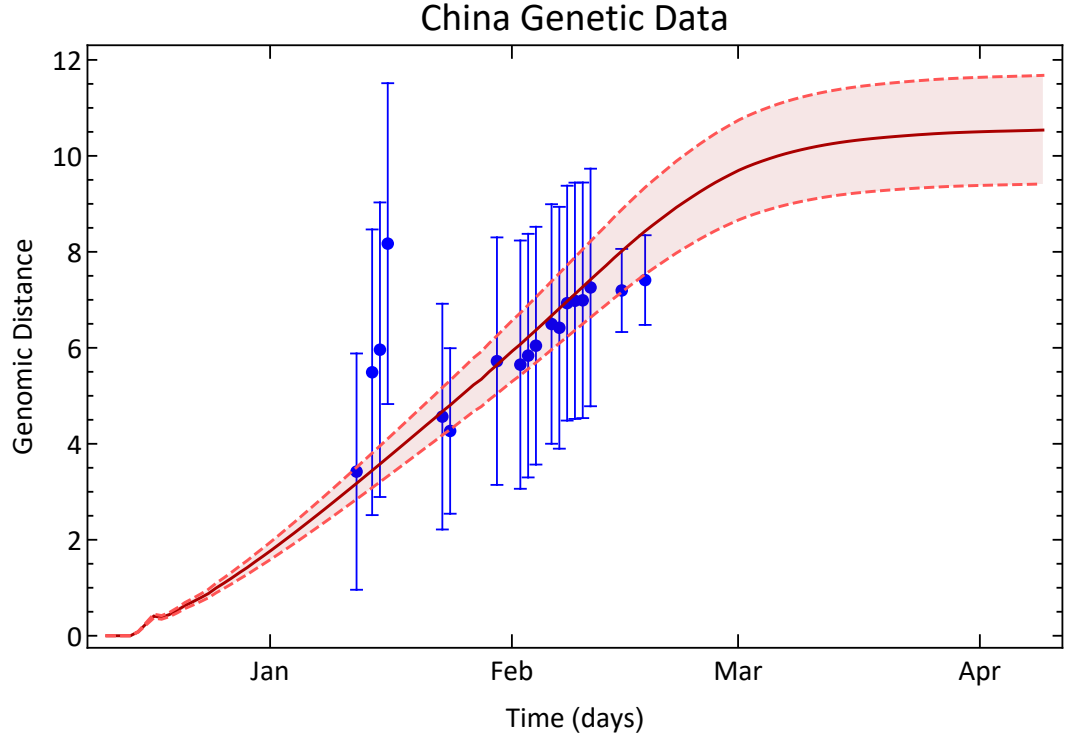
Fig 4. The genetic evolution of SARS-CoV-2 in China. Blue dots are the genetic distance among SARS-CoV-2 inferred from data collected in China between 12/23/2019 and 03/24/2020. The error bars are standard deviation of pairwise distance propagated through the equations. The brown line shows the genetic 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 $B\mu = 29900 \times 0.001/365$.

Communities and reinfection

In this section, we consider the spread of the epidemic through four communities, representing cities, connected linearly as in Fig.5. 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.

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 Fig.7 and Fig.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
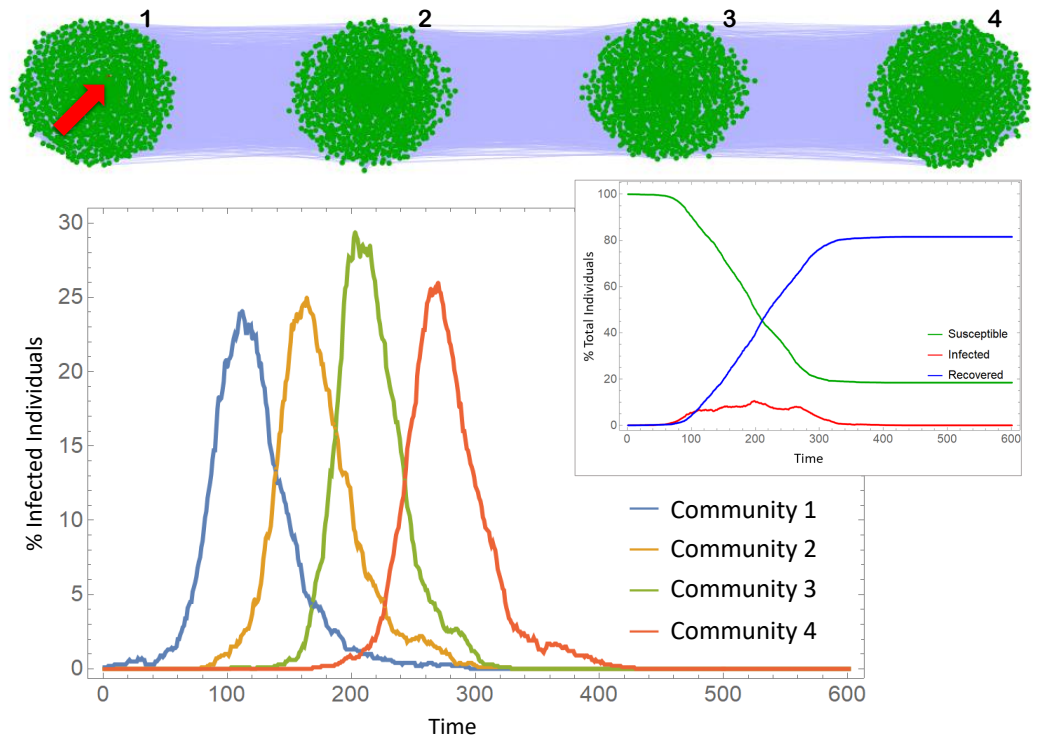
Fig 5. Contact network of four communities on a line and infection curves.
Communities are Barabasi-Albert networks with 1000 nodes. The infection starts with a
single infected individual in the first community (red node indicated with the red
arrow). The epidemic parameters are in S1 Appendix
.

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 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, 51]. 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
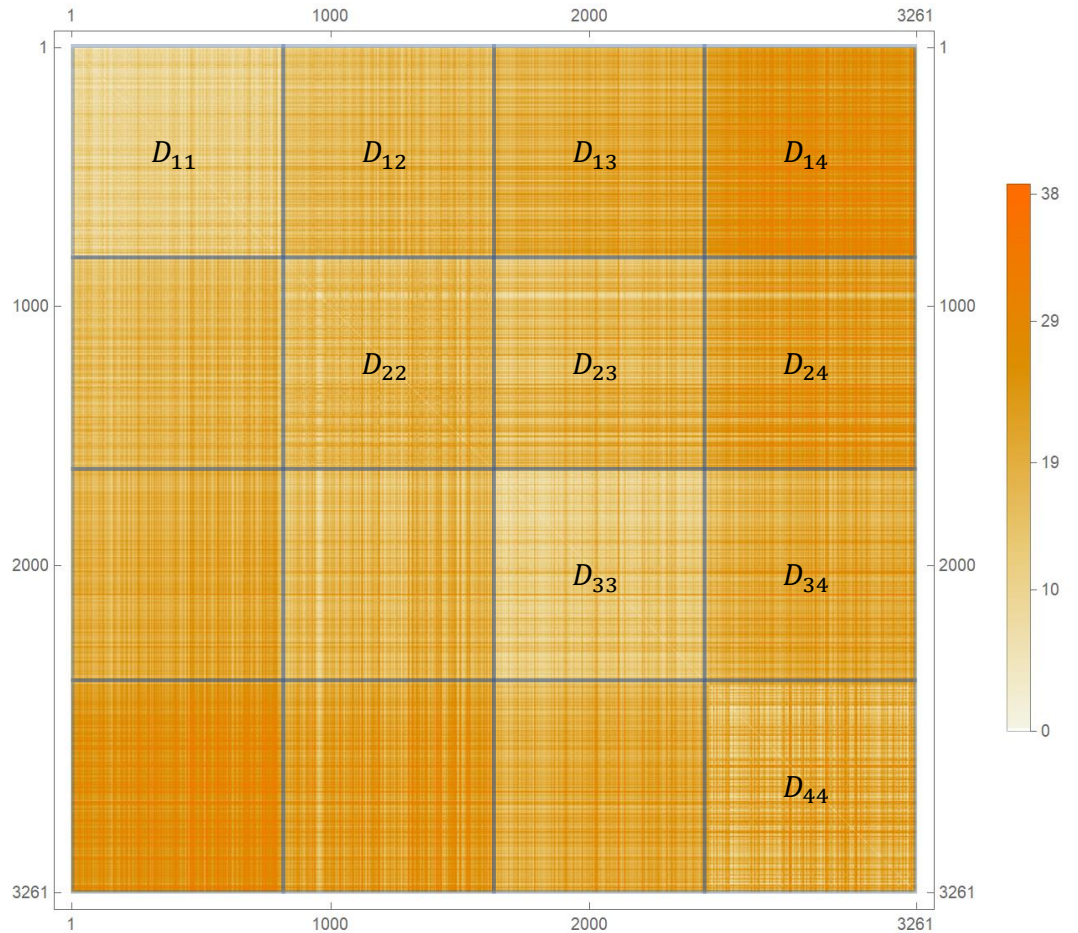
Fig 6. Hamming distance between pairs of viruses. The distance matrix is sorted by the city. Diagonal blocks show the distance between the viruses from a single city, while the non-diagonal blocks are the distances between the viruses from different cities.

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 [44,52]. 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 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
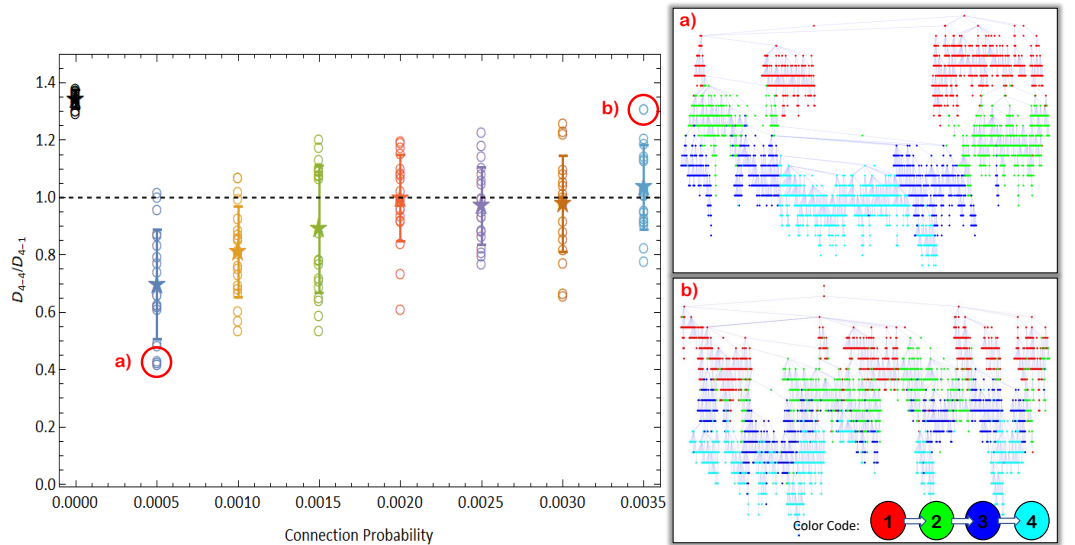
Fig 7. Ratio between the average distance in city 4 and the average distance between cities 1 and 4. Right panels show infection trees for the simulations highlighted with red circles. Open circles show results for individual simulations, the star is the average over 20 simulations and error bars are 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.

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.

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.

Supporting information

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

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).

S2 Table   Included sequences sorted by Collection Date. All informations according to S1 Table.

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.

Acknowledgments

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;.

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.

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.

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.

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.

6. Graham BS. Rapid COVID-19 vaccine development. Science. 2020;368(6494):945–946.

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.

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.

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.

10. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PloS one. 2020;15(3):e0231236.

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.

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.

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.

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;.

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;.

16. Duffy S. Why are RNA virus mutation rates so damn high? PLoS biology. 2018;16(8):e3000003.

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.

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.

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.

21. Kupferschmidt K. Genome analyses help track coronavirus' moves; 2020.

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

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. G L Vasconcelos *et al* . Modelling fatality curves of COVID-19 and the effectiveness of intervention strategies. medRxiv. 2020;doi:10.1101/2020.04.02.20051557.

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;doi:10.25561/77735.

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.

28. Keeling MJ, Eames KT. Networks and epidemic models. Journal of the Royal Society Interface. 2005;2(4):295–307.

29. Kuznetsov YA, Piccardi C. Bifurcation analysis of periodic SEIR and SIR epidemic models. Journal of mathematical biology. 1994;32(2):109–121.

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.

31. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Physical review letters. 2001;86(14):3200.

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.

33. Slatkin M, Hudson RR. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics. 1991;129(2):555–562.

34. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SD. Phylodynamics of infectious disease epidemics. Genetics. 2009;183(4):1421–1430.

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.

37. Volz EM. Complex population dynamics and the coalescent under neutrality. Genetics. 2012;190(1):187–201.

38. Gordo I, Gomes MGM, Reis DG, Campos PR. Genetic diversity in the SIR model of pathogen evolution. PloS one. 2009;4(3):e4876.

39. Kucharski AJ, Andreasen V, Gog JR. Capturing the dynamics of pathogens with many strains. Journal of mathematical biology. 2016;72(1):1–24.

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.

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.

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

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

44. 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.

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

46. 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.

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

48. 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.

49. 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.

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

51. 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;.

52. 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.

Modeling viral mutations in the spread of epidemics - Appendixes

Vitor M. Marquioni[1], Marcus A. M. de Aguiar[1*]

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

* aguiar@ifi.unicamp.br

1 Appendix A: Simulation parameters

The network simulations follow the model proposed in reference [1], and the parameters are displayed in Table 1.

Parameter	Value
R_0	2.4 [2]
Average Symptoms Duration τ_0	14 days [3, 4]
Networks Average Degree D *	100 [1]
Incubation Time Distribution $\mathcal{P}(\tau)$	$\Gamma(6.25, 25/26)$ [5]
Mutation Rate μ	0.001 substitution per base, per year [6, 7]
Genome Size B	29900 bases [2]

Table 1. Simulation Parameters. The number of nodes in each simulation is described properly.
*This is the input average degree for the networks construction, but the actual value for each realization fluctuates. For the communities simulations, this is the parameter for constructing each isolated network, as also for the control case $p = 0$.

For the numerical solution of mean-field approaches, following the SEIR model

$$\begin{aligned}
\dot{S} &= -\beta SI/N \\
\dot{E} &= \beta SI/N - \sigma E \\
\dot{I} &= \sigma E - \gamma I \\
\dot{R} &= \gamma I
\end{aligned} \tag{A1}$$

we have used the following parameters: $R_0 = 2.4$, $\gamma = 1/14$ day^{-1}; $\beta = R_0 \gamma$ and $\sigma = 1/\langle t_i \rangle$, where $\langle t_i \rangle$ is the mean period of incubation, averaged over the distribution from Table 1 [1].

2 Appendix B: Analytical calculations

Our goal is to derive a recurrence equation for the average genetic distance, i.e., given the distance d_t at time t, we aim to calculate the distance d_{t+1} at time $t + 1$. The idea is to calculate d_{t+1} as a weighted average, where the weights are the number of pairs that are distanced by a certain amount. In a SEIR model, every iteration starts with a given number of recovered (R_t), infected (I_t) and exposed (E_t) individuals. When an individual recovers, its infecting virus stops to spread and to evolve, and we call it a

final virus. There are R_t final viruses at the beginning of a given iteration. Viruses infecting Exposed individuals can mutate during this iteration. However, viruses in Infected individuals can either evolve and mutate in this time step or not, since their hosts might recover. The latter become final and are counted as r_t. Infected individuals can also spread the virus, which replicate before evolving or becoming final. Such *offspring* (x_t) increase the number of viruses in Exposed individuals in the next iteration, when they will be allowed to evolve.

At the beginning of iteration $t+1$, there are $(R_t + E_t + I_t)(R_t + E_t + I_t - 1)/2$ pairs of viruses sharing an average distance equal to d_t, but along the iteration some of the distances may increase by a certain amount to be calculated, as also new viruses may arise. Therefore,

$$d_{t+1} = \frac{1}{Z'_t}\left(d_t \frac{(R_t + E_t + I_t)(R_t + E_t + I_t - 1)}{2} + \text{Increases} + \text{Offspring}\right), \quad (A2)$$

where Z'_t is a normalization factor, which counts the total number of pairs at the end of iteration $t+1$,

$$Z'_t = \frac{(R_t + E_t + I_t + x_t)(R_t + E_t + I_t + x_t - 1)}{2}. \quad (A3)$$

If the mutation rate is zero and no new infections occur ($x_t = 0$) the "Increases" term and the "Offspring" term are equal to zero, and $d_{t+1} = d_t$, as expected.

In the following two subsections, we shall calculate the "Increases" term and the "Offspring" term, which accounts for the evolution and for the spread, respectively.

2.1   Increases

Genetic distances between evolving viruses increase over time. In order to calculate how much these distances increase we first consider that mutations occurring in the same locus of different genomes are unlikely, as well as more than one mutation per locus on a single genome. This approximation holds as long as the epidemic duration T remains sufficiently small, $\mu T \ll 1$. Thus, after one time step, an evolving genome acquires, on average, $B\mu$ mutations. The distance between two evolving genomes will increase, on average, by $2B\mu$ nucleotides after one time step. The distance between viruses in exposed individuals, for example, increases by $2B\mu$ and because there are $E_t(E_t - 1)/2$ pairs of exposed individuals, their evolution along the iteration $t+1$ contributes $2B\mu E_t(E_t - 1)/2$ to the Increases term. On the other hand, the distance between viruses in an exposed and a recovered individual, or an infected individual that recovers, is only $B\mu$, because the latter two do not evolve. There are $E_t(R_t + r_t)$ pairs among these viruses, and thus their contribution to Increases is $E_t(R_t + r_t)B\mu$. We recall that the updates in our model occur in the order "Transmission", "Attempt to Recovery" and lastly, "Genome Evolution". Thus, if an infected individual recovers its virus does not have the chance to mutate.

Therefore, in order to compute the Increases term, we must calculate the average increase in distance between all pairs of viruses and how many pairs of these viruses exist. Table 2 summarizes this information. We obtain

$$\begin{aligned}
\text{Increases} =&E_t R_t B\mu + E_t r_t B\mu + (I_t - r_t)r_t B\mu + (I_t - r_t)R_t B\mu + (I_t - r_t)E_t 2B\mu \\
&+ \frac{E_t(E_t - 1)}{2}2B\mu + \frac{(I_t - r_t)(I_t - r_t - 1)}{2}2B\mu.
\end{aligned} \quad (A4)$$

Table 2. Increases in average distance and number of pairs of viruses.

Viruses	Number of Pairs	Average Distance Increase
(E_t) and (R_t)	$E_t R_t$	$B\mu$
(E_t) and (r_t)	$E_t r_t$	$B\mu$
$(I_t - r_t)$ and (r_t)	$(I_t - r_t)r_t$	$B\mu$
$(I_t - r_t)$ and (R_t)	$(I_t - r_t)R_t$	$B\mu$
$(I_t - r_t)$ and (E_t)	$(I_t - r_t)E_t$	$2B\mu$
(E_t) and (E_t)	$E_t(E_t - 1)/2$	$2B\mu$
$(I_t - r_t)$ and $(I_t - r_t)$	$(I_t - r_t)(I_t - r_t - 1)/2$	$2B\mu$
(R_t) and (R_t)	$R_t(R_t - 1)/2$	0
(r_t) and (r_t)	$r_t(r_t - 1)/2$	0
(r_t) and (R_t)	$r_t R_t$	0

2.2 Offspring

The contribution of the new infections to the average distance d_{t+1}, the Offspring term, is more tricky. To simplify matters we will assume that an infected individual infects only one susceptible per time step, which is a good assumption if the basic reproduction number R_0 is small compared to the average duration of symptoms. Thus, x_t is also the number of individuals who infected a susceptible within the time step $t + 1$, which will be called *ancestors* from now on. Let D_1 be the average distance between ancestors and the other viruses at time t, and D_2, the distance between the exposed and the other viruses. Note that an ancestor may recover and, therefore, not mutate in this time step. The Offspring term is a sum of different contributions between offspring and the other viruses in the population, as explained in detail below.

1. *Genetic distance between offspring and recovered.* The number of pairs is $x_t R_t$. Because offspring do not evolve in the time step they appear, their average distance is D_1. Then, its contribution to the Offspring term is $\boldsymbol{x_t R_t D_1}$.

2. *Genetic distance between offspring and exposed.* The number of pairs is $x_t E_t$. Because the exposed evolve, these pairs contribute with $\boldsymbol{x_t E_t (D_2 + B\mu)}$ to the Offspring term.

3. *Genetic Distance between offspring of an infected (ancestor) that does not recover (there are $(I_t - r_t)$ of these individuals) and infected:*

   (a) The distance between an offspring and its ancestor is $B\mu$, since the ancestor evolves. There are $x_t(I_t - r_t)/I_t$ new infections of this type, contributing with $\boldsymbol{x_t((I_t - r_t)/I_t)B\mu}$ to the distance.

   (b) For each offspring there are $I_t - r_t - 1$ infected individuals that did not recover and are not its ancestral. The distance between the offspring and these individuals is $(D_1 + B\mu)$, adding $\boldsymbol{x_t((I_t - r_t)/I_t)(I_t - r_t - 1)(D_1 + B\mu)}$ to the Offspring term.

   (c) The distance between the offspring and individuals that recover is D_1, because neither of these viruses evolve in this time step. There are $x_t((I_t - r_t)/I_t)r_t$ pairs of these viruses, adding $\boldsymbol{x_t((I_t - r_t)/I_t)r_t D_1}$ to the Offspring term.

4. *Genetic distance between offspring of infected (ancestor) that recover in this iteration (there are r_t of these individuals) and infected:*

84
85
86
87
88
89
90
91
92
93

94
95
96
97
98
99
100
101
102
103
104
105
106

107

(a) The distance between offspring and its ancestor is zero, because none of them evolve.

(b) The distance between the offspring and the other viruses of type is D_1. There are $x_t r_t / I_t$ new infections of this type, contributing $(x_t r_t / I_t)(r_t - 1)D_1$ to the Offspring term.

(c) The distance between offspring and the other infected individuals is $(x_t r_t / I_t)(I_t - r_t)(D_1 + B\mu)$, since the other infected viruses evolve..

5. *Genetic distance between offspring.* Because each ancestor gives rise to only one new infection, this distance equals D_1, and once there are $x_t(x_t - 1)/2$ pairs of offspring, this contribution is $(x_t(x_t - 1)/2)D_1$.

6. By summing everything up, we get

$$\text{Offspring} = x_t R_t D_1 + x_t E_t (D_2 + B\mu)$$
$$+ x\frac{(I_t - r_t)}{I_t}B\mu + x_t\frac{(I_t - r_t)}{I_t}(I_t - r_t - 1)(D_1 + B\mu) + x_t\frac{(I_t - r_t)}{I_t}r_t D_1$$
$$+ x_t\frac{r_t}{I_t}0 + x_t\frac{r_t}{I_t}(r_t - 1)D_1 + x_t\frac{r_t}{I_t}(I_t - r_t)(D_1 + B\mu)$$
$$+ \frac{x_t(x_t - 1)}{2}D_1. \tag{A5}$$

Putting all these terms together and defining $Z_t \equiv 2Z_t'$ we obtain

$$d_{t+1} = \frac{1}{Z_t}\left(d_t(R_t + E_t + I_t)(R_t + E_t + I_t - 1)\right.$$
$$+ x_t D_1(x_t - 3 + 2R_t + 2I_t + 2E_t D_2/D_1)$$
$$\left. + 2B\mu(E_t + I_t - r_t)(E_t + I_t + R_t + x_t - 1)\right). \tag{A6}$$

The reason for assigning the distance D_1 between infected and other viruses, instead of d_t, is that infected individuals represent only a fraction of the viruses in the population, and the distance between them and other viruses grows over time, therefore being above the average d_t. The same holds for the exposed individuals.

Although we were not able to analytically find an expression for D_1 and D_2, we can approximate them as follows. First we assume that $D_2 \approx D_1$. When the epidemic begins, all viruses are infected, so that $D_1 = d_t$. However, the ratio between infected and recovered individuals decreases to zero along the epidemic, making D_1 larger than d_t. Thus, to first order, it is possible to approximate $D_1 \approx d_t(1 + \epsilon)$, with ϵ a function of the number of recovered individuals, $R_t/(I_t + E_t + R_t)$ and the average number of mutations $B\mu$. Our simulations showed that the linear function $D_1 = d_t(1 + 2B\mu R_t/(I_t + E_t + R_t))$ works well (considering the parameters in Appendix A), leading to the theoretical result expressed by Eq.(2) from the main text.

2.3 Continuum Limit

To achieve the continuum limit we start by substituting $r_t = R_{t+1} - R_t$ and $x_t = E_{t+1} - E_t + I_{t+1} - I_t + R_{t+1} - R_t$ in Eq.(2) from the main text and subtracting d_t from both sides of this equation:

$$d_{t+1} - d_t = \frac{1}{Z_t}\left\{2d_t\left(E_{t+1} - E_t + I_{t+1} - I_t + R_{t+1} - R_t\right) \times\right.$$
$$\times\left[-1 + B\mu\frac{R_t}{I_t + E_t + R_t}\left(R_{t+1} + R_t + I_{t+1} + I_t + E_{t+1} + E_t - 3\right)\right]$$
$$\left. + 2B\mu\left(E_t + I_t + R_t - R_{t+1}\right)\left(E_{t+1} + I_{t+1} + R_{t+1} - 1\right)\right\} \tag{A7}$$

with

$$Z_t = (E_{t+1} + I_{t+1} + R_{t+1})(E_{t+1} + I_{t+1} + R_{t+1} - 1). \tag{A8}$$

Then, we consider the first order approximations

$$f_t \approx f(t)$$
$$f_{t+1} \approx f(t) + \dot{f}(t)\Delta t,$$

and once $B\mu$ in the last line of Eq.(A7) is the number of mutations per time step, we replace it by $B\mu\Delta t$

$$\dot{d}(t)\Delta t = \tag{A9}$$

$$\frac{1}{Z_t} \left\{ 2d(t)\Delta t \left(\dot{E}(t) + \dot{I}(t) + \dot{R}(t) \right) \times \right.$$

$$\times \left[-1 + B\mu \frac{R(t)}{I(t) + E(t) + R(t)} \left(2R(t) + 2I(t) + 2E(t) + \Delta t(\dot{E}(t) + \dot{I}(t) + \dot{R}(t)) - 3 \right) \right]$$

$$\left. + 2B\Delta t\mu \left(E(t) + I(t) - \dot{R}(t)\Delta(t) \right) \left(R(t) + I(t) + E(t) + \Delta t(\dot{E}(t) + \dot{I}(t) + \dot{R}(t)) - 1 \right) \right\}$$
$$\tag{A10}$$

with

$$Z_t = (R(t) + I(t) + E(t) + \Delta t(\dot{E}(t) + \dot{I}(t) + \dot{R}(t))) \times$$
$$\times (R(t) + I(t) + E(t) + \Delta t(\dot{E}(t) + \dot{I}(t) + \dot{R}(t)) - 1). \tag{A11}$$

Finally, by taking the limit $\Delta t \to 0$ we obtain the continuous time equation.

2.4 Multiple Infections

The average distance $d_{root,t}^{(i)}$ between viruses from a lineage and its root is calculated using the same technique discussed above, however it is much simpler, once we only need to calculate the average distance from a kind of virus and the root (a single virus which does not evolve). Using the same notation, but now with a super-index to denote the lineage, we obtain

$$d_{root,t+1}^{(i)} = \frac{1}{Z_t} \left[\left(R_t^{(i)} + E_t^{(i)} + I_t^{(i)} \right) d_{root,t}^{(i)} + E_t^{(i)} B\mu + \left(I_t^{(i)} - r_t^{(i)} \right) B\mu + x_t^{(i)} D_{1,root}^{(i)} \right]$$
$$\tag{A12}$$

with $Z_t = (E_t^{(i)} + I_t^{(i)} + R_t^{(i)} + x_t^{(i)})$ and $D_{1,root}^{(i)}$ being the average distance between infected and the root, which is given (similarly to D_1) by

$$D_{1,root}^{(i)} = d_{root,t}^{(i)} \left(1 + 4B\mu \frac{R_t^{(i)}}{E_t^{(i)} + I_t^{(i)} + R_t^{(i)}} \right).$$

The factor 4 is a fit from numerical investigations. The continuum limit is obtained by subtracting $d_{root,t}^{(i)}$ from both sides of Eq.(6) from the main text, applying the continuous approximation for each epidemic curve and taking the limit $\Delta t \to 0$.

3 Appendix C: Real genetic evolution algorithm

In order to estimate the real (from genetic data) genetic evolution, we used 55 complete genome sequences collected in China [8]. First, these sequences were ordered and

numbered by its collection date and a matrix of genetic distances d_{ij} between genomes i and j has been constructed. Each pair of sequences were alligned with the Needleman-Wunsch algorithm, with score $+1$ for match and -1 for mismatch [9]. Then, the distance between two genomes was computed counting the number of substitutions between the sequences, neglecting *indels*.

We defined a time window $\tau_W = 14 = \tau_0$ days. Thus, every genome collected within τ_W are considered infected, and the genomes collected before this time window are considered recovered. Now, we calculate the average distance among the infected $d_{I,t}$, recovered $d_{R,t}$ and among infected and recovered $d_{IR,t}$ at the time t. Fig.1 shows an example of a distance matrix with a specific time window. Finally, the average distance at time t can be computed as

$$d_t = \frac{d_{I,t}I_t(I_t-1) + 2d_{IR,t}I_tR_t + d_{R,t}R_t(R_t-1)}{(R_t+I_t)(R_t+I_t-1)} \tag{A13}$$

where I_t and R_t are respectively given by $I(t)$ and $R(t)$ described evaluated in the supplemental material.

With this algorithm, we obtained 20 non-overlapping sets of infected genomes. One of these sets contained only one sequence and was not usable; a second set was too far from all other data and was also discarded. Thus, we were able to calculate 18 points (that appear in Fig.4 from the main text) with error bars given by the standard deviation of each set of distances (between infected, recovered and between infected and recovered) at each time t.

Fig 1. Example of distance matrix to illustrate the algorithm to infer the genetic evolution. Every genome collected within a time window τ_W is considered to belong to an infected individual. The red block shows distances between these viruses. The blue block shows viruses that appeared before the present time window, whose individuals are considered to have recovered. Green blocks are distances between infected and recovered individuals. The remaining entries are distances from viruses that have not appeared yet at that considered time, i.e., they appeared after the considered time window.

4 Appendix D: The COVID-19 data from China

We got the Chinese epidemic data from the dataset "Epidemic Data for Novel Coronavirus COVID-19" from Wolfram data repository [10]. Unfortunately, this dataset starts on 22 January (going up to 18 August by the date of our analysis), lacking the previous data. Another concern is about the change in the notification protocols adopted by the Chinese government. On 13 February, the Hubei province started to report not only the positive laboratory tests, but also the clinically diagnosed cases as infected too, appearing a sudden increase in infected curve [11]. We also need to correct the data by including undetected cases.

Firstly, in order to correct the notification problem, we smoothly distribute the sudden increase number of cases among the previous dates. Following reference [11], the

corrected accumulated number of cases $I_{a,c}(t)$ is given by

$$I_{a,c}(t) = I_a(t) + 15133 \frac{\sum_{i=22 \text{ Jan}}^{t} I_a(t)}{\sum_{i=22 \text{ Jan}}^{13 \text{ Feb}} I_a(t)} \tag{A14}$$

for $t \in \{22 \text{ Jan}, \ldots, 12 \text{ Feb}\}$, where $I_c(t)$ is the accumulated number of cases at date t, and $15133 = I_a(13 \text{ Feb}) - I_a(12 \text{ Feb})$ is the sudden increase due to the changes in the notification protocol.

Now, the undetected cases in China were estimated in reference [12], and also following reference [11], we get

$$I_{a,c'}(t) = \frac{I_{a,c}(t)}{1 - \theta(t)} \tag{A15}$$

for the estimated total number of cases at time t, where θ is the undetected fraction,

$$\theta(t) = \begin{cases} 0.86, \text{ for } t \leq 24 \text{ Jan} \\ linear\ decrease, \text{ for } 24 \text{ Jan} \leq t \leq 08 \text{ Feb} \\ 0.31, \text{ for } t \geq 08 \text{ Feb} \end{cases} \tag{A16}$$

This correction is also applied to the recovered curve. However, the Wolfram data distincts recovered $Rec(t)$ from deaths $Dea(t)$, while our theory does not differentiates these numbers. Thus, the number of recovered individuals we must consider is

$$R(t) = \frac{Rec(t) + Dead(t)}{1 - \theta(t)} \tag{A17}$$

and the infected curve is now given as

$$I(t) = I_{a,c'}(t) - R(t) \tag{A18}$$

Fig.2 shows the curves after these corrections. Once we do no have directly access to exposed data, we did not consider exposed individuals, meaning, at this point, that we are dealing with a SIR model without any prejudice to the present theory. However, bad data is an important source of error.
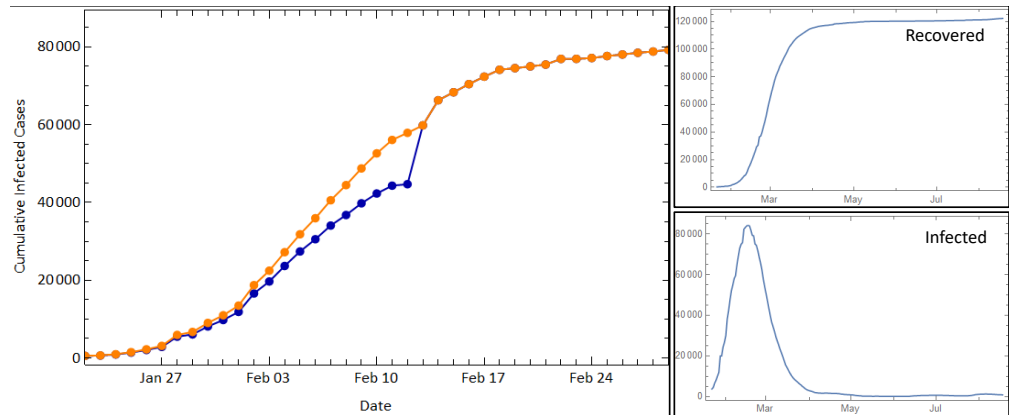

Fig 2. Chinese epidemic curves after corrections. The left chart shows the cumulative number of infections in China. The blue curve is the reported number of cases before the smoothness procedure of Eq.(14) and the orange curve is the result of this procedure. The right charts are the recovered and infected curves $R(t)$ and $I(t)$.

Finally, we fit an exponential curve to a few initial data points of $I(t)$ and $R(t)$ and extrapolate it to previous dates. For the I-curve, we have adjusted the exponential $e^{a(t-t_0)}$, with fit parameters a and t_0, on the first $n_I = 10$ data points and extrapolated it up to the first case t_0 days before. With this approach, we found $t_0 = 11$ Dec, which is close to the first case reported by WHO, 08 Dec [13]. For the $R(t)$-curve, we have used the first $n_R = 13$ data points. The numbers n_I and n_R were chosen in order to make the exponential extrapolation makes sense according to WHO estimates of the first case, as also to make $R(t) < I(t)$ in a plausible way.

Now, the curves $R(t)$ and $I(t)$ can be implemented in the recurrence equation and the distance evolution can be estimated, with the first distance d_0 equalling zero.

References

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

2. Guo YR, Cao QD, Hong ZS, Tan YY, Chen SD, Jin HJ, et al. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. Military Medical Research. 2020;7(1):1–10.

3. Akhmetzhanov AR, Mizumoto K, Jung Sm, Linton NM, Omori R, Nishiura H. Epidemiological characteristics of novel coronavirus infection: A statistical analysis of publicly available case data. medRxiv. 2020;.

4. Morris DH, Rossine FW, Plotkin JB, Levin SA. Optimal, near-optimal, and robust epidemic control. arXiv preprint arXiv:200402209. 2020;.

5. 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.

6. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019. Clinical Infectious Diseases. 2020;71(15):713–720. doi:10.1093/cid/ciaa203.

7. 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.

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

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

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

11. 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.

12. 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.

13. World Health Organization. Coronavirus disease 2019 (COVID-19) Situation Report 37; 2020. `https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200226-sitrep-37-covid-19.pdf?sfvrsn=2146841e_2`.

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) [1].

Accession Number	Collection Date	Length	Geographic Location	Included?	Justification
MN908947	26 Dec 2019*	29903	Wuhan, Hubei**	Y	
MN938384	10 Jan 2020	29838	Shenzen, Guangdong	Y	
MN975262	11 Jan 2020	29891	Wuhan, Hubei**	Y	
MN988668	02 Jan 2020	29881	Wuhan, Hubei**	Y	
MN988669	02 Jan 2020	29881	Wuhan, Hubei**	Y	
MN996527	30 Dec 2019	29825	Wuhan, Hubei	Y	
MN996528	30 Dec 2019	29891	Wuhan, Hubei	Y	
MN996529	30 Dec 2019	29852	Wuhan, Hubei	Y	
MN996530	30 Dec 2019	29854	Wuhan, Hubei	Y	
MN996531	30 Dec 2019	29857	Wuhan, Hubei	Y	
MT019529	23 Dec 2019	29899	Wuhan, Hubei	Y	
MT019530	30 Dec 2019	29889	Wuhan, Hubei	N	MT19530 to MT19532: Might be biased data (sequences from the same researchers, collected at the same day and with quite the same length, with no other informations up to the date we have made the analysis).*
MT019531	30 Dec 2019	29899	Wuhan, Hubei	N	
MT019532	30 Dec 2019	29890	Wuhan, Hubei	N	
MT019533	01 Jan 2020	29883	Wuhan, Hubei	Y	
MT034054	03 Jan 2020	29885	Beijing	Y	
MT039873	20 Jan 2020	29833	Hangzhou, Zhejiang	Y	
MT039874	22 Jan 2020	29858	Hangzhou, Zhejiang**	Y	
MT049951	17 Jan 2020	29903	Kunming,† Yunnan	Y	
MT079843	22 Jan 2020	29915	Wuhan, Hubei**	Y	MT079843 to MT079854: Might be biased data (probable nosocomial transmission).** Then we have included only one genome.
MT079844	22 Jan 2020	29910	Wuhan, Hubei**	N	
MT079845	22 Jan 2020	29955	Wuhan, Hubei**	N	
MT079846	22 Jan 2020	29903	Wuhan, Hubei**	N	
MT079847	22 Jan 2020	29872	Wuhan, Hubei**	N	
MT079848	22 Jan 2020	29880	Wuhan, Hubei**	N	
MT079849	22 Jan 2020	29904	Wuhan, Hubei**	N	
MT079850	22 Jan 2020	29885	Wuhan, Hubei**	N	
MT079851	22 Jan 2020	30018	Wuhan, Hubei**	N	
MT079852	22 Jan 2020	29891	Wuhan, Hubei**	N	
MT079853	22 Jan 2020	29766	Wuhan, Hubei**	N	
MT079854	22 Jan 2020	29897	Wuhan, Hubei**	N	
MT093631	08 Jan 2020	29860	Beijing†	N	No detailed geographic information available.
MT121215	02 Feb 2020	29945	Shanghai	Y	

1

Accession	Date	Length	Location	Included	Notes
MT123290	05 Feb 2020	29891	Guangzhou, Guangdong	Y	
MT123291	29 Jan 2020	29882	Guangzhou, Guangdong	Y	
MT123292	27 Jan 2020	29923	Guangzhou, Guangdong	Y	
MT123293	29 Jan 2020	29871	Guangzhou, Guangdong	Y	
MT135041	26 Jan 2020	29903	Beijing	N	MT135041 to MT135044: Might be biased data (the lengths are all the same). Then we have included only one genome
MT135042	28 Jan 2020	29903	Beijing	N	
MT135043	28 Jan 2020	29903	Beijing	N	
MT135044	28 Jan 2020	29903	Beijing	Y	
MT226610	20 Jan 2020	29899	Kunming, Yunnan†	N	No detailed geographic information available.
MT253696	23 Jan 2020	29781	Hangzhou, Zhejiang	N	MT253696 to MT253710: Might be biased data (cluster of cases*; also they all have the same length). Then we have included only one genome.
MT253697	23 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253698	23 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253699	24 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253700	25 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253701	21 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253702	21 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253703	25 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253704	25 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253705	22 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253706	22 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253707	25 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253708	21 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253709	21 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253710	21 Jan 2020	29781	Hangzhou, Zhejiang	Y	
MT259226	10 Jan 2020	29868	Wuhan, Hubei	Y	
MT259227	26 Jan 2020	29863	Wuhan, Hubei	Y	
MT259228	26 Jan 2020	29861	Wuhan, Hubei	Y	
MT259229	26 Jan 2020	29864	Wuhan, Hubei	Y	
MT259230	25 Jan 2020	29866	Wuhan, Hubei	Y	
MT259231	25 Jan 2020	29865	Wuhan, Hubei	Y	
MT281577	10 Mar 2020	29903	Fujyang, Anhui	Y	
MT291826	30 Dec 2019	29807	Wuhan, Hubei	Y	
MT291827	30 Dec 2019	29858	Wuhan, Hubei	Y	
MT291828	30 Dec 2019	29858	Wuhan, Hubei	Y	
MT291829	30 Dec 2019	29774	Wuhan, Hubei	Y	
MT291830	30 Dec 2019	29807	Wuhan, Hubei	Y	
MT291831	24 Jan 2020	29872	Beijing	Y	
MT291832	25 Jan 2020	29828	Beijing	Y	

Accession	Date	Length	Location	Ref	Notes
MT291833	28 Jan 2020	29821	Beijing	Y	
MT291834	28 Jan 2020	29865	Beijing	Y	
MT291835	27 Jan 2020	29834	Beijing	Y	
MT291836	29 Jan 2020	29860	Beijing	Y	
MT407649	22 Jan 2020	29833	Hangzhou,† Zhejiang	Y	
MT407650	22 Jan 2020	29821	Hangzhou,† Zhejiang	Y	
MT407651	22 Jan 2020	29822	Hangzhou,† Zhejiang	Y	
MT407652	26 Jan 2020	29835	Hangzhou,† Zhejiang	Y	
MT407653	26 Jan 2020	29835	Hangzhou,† Zhejiang	Y	
MT407654	24 Mar 2020	29817	Hangzhou,† Zhejiang	Y	
MT407655	24 Mar 2020	29817	Hangzhou,† Zhejiang	Y	
MT407656	24 Mar 2020	29835	Hangzhou,† Zhejiang	Y	
MT407657	24 Mar 2020	29776	Hangzhou,† Zhejiang	Y	
MT407658	24 Mar 2020	29770	Hangzhou,† Zhejiang	Y	
MT407659	24 Mar 2020	29828	Hangzhou,† Zhejiang	Y	
MT412134	24 Feb 2020	29867	Zhengzhou, Henan†	N	No detailed geographic information available.
MT446312	05 Feb 2020	29879	Guangzhou, Guangdong	Y	
MT510727	15 Feb 2020	29903	Meizhou, Guangdong†	N	MT510727 and MT510728: Might be biased data (data from familial cluster*). There is also no detailed geographic information available.
MT510728	13 Feb 2020	29903	Meizhou, Guangdong†	N	
MT534630	26 Jan 2020	29845	Changzhou,Jiangsu	Y	
MT568634	25 Feb 2020	29861	Guangzhou, Guangdong	N	MT568634 to MT568641: data from a work presenting different approaches for genome sequencing.** Then, this data might have more errors than the others.
MT568635	25 Feb 2020	29854	Guangzhou, Guangdong	N	
MT568636	27 Feb 2020	29858	Guangzhou, Guangdong	N	
MT568637	25 Feb 2020	29860	Guangzhou, Guangdong	N	
MT568638	25 Feb 2020	29854	Guangzhou, Guangdong	N	
MT568639	25 Feb 2020	29861	Guangzhou, Guangdong	N	
MT568640	25 Feb 2020	29858	Guangzhou, Guangdong	N	
MT568641	25 Feb 2020	29868	Guangzhou, Guangdong	N	
MT622319	23 Jan 2020	29889	Shanghai†	N	No detailed geographic information available.
MT627325	28 Feb 2020	29859	Shanghai†	N	No detailed geographic information available.
NC045512	Dec 2019	29903	Wuhan, Hubei**	N	Identical to MN908947.*

[1]Wolfram Research. Genetic Sequences for the SARS-CoV-2 Coronavirus; 2020. Wolfram Data Repository https://doi.org/10.24097/wolfram.03304.data
GenBank information.
*Publication Information.
† Laboratory address.

S2 Table. Included sequences sorted by Collection Date. All informations according to S1 Table.

Number	Accession Number	Collection Date	Length	Geographic Location
#1	MT019529	23 Dec 2019	29899	Wuhan, Hubei
#2	MN908947	26 Dec 2019	29903	Wuhan, Hubei
#3	MT291829	30 Dec 2019	29774	Wuhan, Hubei
#4	MT291826	30 Dec 2019	29807	Wuhan, Hubei
#5	MT291830	30 Dec 2019	29807	Wuhan, Hubei
#6	MN996527	30 Dec 2019	29825	Wuhan, Hubei
#7	MN996529	30 Dec 2019	29852	Wuhan, Hubei
#8	MN996530	30 Dec 2019	29854	Wuhan, Hubei
#9	MN996531	30 Dec 2019	29857	Wuhan, Hubei
#10	MT291827	30 Dec 2019	29858	Wuhan, Hubei
#11	MT291828	30 Dec 2019	29858	Wuhan, Hubei
#12	MN996528	30 Dec 2019	29891	Wuhan, Hubei
#13	MT019533	01 Jan 2020	29883	Wuhan, Hubei
#14	MN988668	02 Jan 2020	29881	Wuhan, Hubei
#15	MN988669	02 Jan 2020	29881	Wuhan, Hubei
#16	MT034054	03 Jan 2020	29885	Beijing
#17	MN938384	10 Jan 2020	29838	Shenzhen, Guangdong
#18	MT259226	10 Jan 2020	29868	Wuhan, Hubei
#19	MN975262	11 Jan 2020	29891	Wuhan, Hubei
#20	MT049951	17 Jan 2020	29903	Yunnan
#21	MT039873	20 Jan 2020	29833	Hangzhou, Zhejiang
#22	MT253710	21 Jan 2020	29781	Hangzhou, Zhejiang
#23	MT407650	22 Jan 2020	29821	Zhejiang
#24	MT407651	22 Jan 2020	29822	Zhejiang
#25	MT407649	22 Jan 2020	29833	Zhejiang
#26	MT039874	22 Jan 2020	29858	Hangzhou, Zhejiang
#27	MT079843	22 Jan 2020	29915	Wuhan, Hubei
#28	MT291831	24 Jan 2020	29872	Beijing
#29	MT291832	25 Jan 2020	29828	Beijing
#30	MT259231	25 Jan 2020	29865	Wuhan, Hubei
#31	MT259230	25 Jan 2020	29866	Wuhan, Hubei
#32	MT407652	26 Jan 2020	29835	Zhejiang
#33	MT407653	26 Jan 2020	29835	Zhejiang
#34	MT534630	26 Jan 2020	29845	Changzhou, Jiangsu
#35	MT259228	26 Jan 2020	29861	Wuhan, Hubei
#36	MT259227	26 Jan 2020	29863	Wuhan, Hubei
#37	MT259229	26 Jan 2020	29864	Wuhan, Hubei
#38	MT291835	27 Jan 2020	29834	Beijing
#39	MT123292	27 Jan 2020	29923	Guangzhou, Guangdong
#40	MT291833	28 Jan 2020	29821	Beijing
#41	MT291834	28 Jan 2020	29865	Beijing
#42	MT135044	28 Jan 2020	29903	Beijing
#43	MT291836	29 Jan 2020	29860	Beijing
#44	MT123293	29 Jan 2020	29871	Guangzhou, Guangdong
#45	MT123291	29 Jan 2020	29882	Guangzhou, Guangdong
#46	MT121215	02 Feb 2020	29945	Shanghai
#47	MT446312	05 Feb 2020	29879	Guangzhou, Guangdong
#48	MT123290	05 Feb 2020	29891	Guangzhou, Guangdong
#49	MT281577	10 Mar 2020	29903	Fuyang, Anhui
#50	MT407658	24 Mar 2020	29770	Zhejiang
#51	MT407657	24 Mar 2020	29776	Zhejiang
#52	MT407654	24 Mar 2020	29817	Zhejiang
#53	MT407655	24 Mar 2020	29817	Zhejiang
#54	MT407659	24 Mar 2020	29828	Zhejiang
#55	MT407656	24 Mar 2020	29835	Zhejiang

S3 Table. Genome information used to calculate points in Fig.5. 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.

Point Number*	Infected Genomes	Recovered Genomes	Date Interval
#1	#03 → #19	#1 → #02	30 Dec 2019 → 12 Jan 2020
#2	#13 → #19	#1 → #12	01 Jan 2020 → 14 Jan 2020
#3	#14 → #19	#1 → #13	02 Jan 2019 → 15 Jan 2020
#4	#16 → #19	#1 → #15	03 Jan 2019 → 16 Jan 2020
#5	#17 → #27	#1 → #16	10 Jan 2019 → 23 Jan 2020
#6	#19 → #28	#1 → #18	11 Jan 2019 → 24 Jan 2020
#7	#20 → #45	#1 → #19	17 Jan 2019 → 30 Jan 2020
#8	#21 → #46	#1 → #20	20 Jan 2019 → 02 Feb 2020
#9	#22 → #46	#1 → #21	21 Jan 2019 → 03 Feb 2020
#10	#23 → #46	#1 → #22	22 Jan 2019 → 04 Feb 2020
#11	#28 → #48	#1 → #27	24 Jan 2019 → 06 Feb 2020
#12	#29 → #48	#1 → #28	25 Jan 2019 → 07 Feb 2020
#13	#32 → #48	#1 → #31	26 Jan 2019 → 08 Feb 2020
#14	#38 → #48	#1 → #37	27 Jan 2019 → 09 Feb 2020
#15	#40 → #48	#1 → #39	28 Jan 2019 → 10 Feb 2020
#16	#43 → #48	#1 → #42	29 Jan 2019 → 11 Feb 2020
#17	#46 → #48	#1 → #45	02 Feb 2019 → 15 Feb 2020
#18	#47 → #48	#1 → #46	05 Feb 2019 → 18 Feb 2020
#19**	#49 → #49	#1 → #48	
#20†	#50 → #55	#1 → #49	24 Mar 2019 → 06 Apr 2020

*In Fig.4 from the main text, points are numbered from left to right.

**Since there is only one genome in this time window, we cannot estimate a distance among the infected population, so genome #49 was not used.

† This point was not included in Fig.5 because it is lacking more than one month of genetic information between points #18 and #19, therefore the distance among the recovered population cannot be well inferred.