RESEARCH ARTICLE

National Science Review
8: nwaa246, 2021
doi: 10.1093/nsr/nwaa246
Advance access publication 24 September 2020

MOLECULAR BIOLOGY & GENETICS

On the founder effect in COVID-19 outbreaks: how many infected travelers may have started them all?

Yongsen Ruan[1], Zhida Luo[1], Xiaolu Tang[2], Guanghao Li[3], Haijun Wen[1], Xionglei He[1], Xuemei Lu[4,*], Jian Lu[2,*] and Chung-I Wu[1,*]

[1]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China; [2]State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, School of Life Sciences, Peking University, Beijing 100871, China; [3]CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China and [4]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology; Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

*Corresponding authors. E-mails: wzhongyi@mail.sysu.edu.cn; LUJ@pku.edu.cn; xuemeilu@mail.kiz.ac.cn

Received 10 June 2020; **Revised** 22 July 2020; **Accepted** 11 August 2020

ABSTRACT

How many incoming travelers (I_0 at time 0, equivalent to the 'founders' in evolutionary genetics) infected with SARS-CoV-2 who visit or return to a region could have started the epidemic of that region? I_0 would be informative about the initiation and progression of epidemics. To obtain I_0, we analyze the genetic divergence among viral populations of different regions. By applying the 'individual-output' model of genetic drift to the SARS-CoV-2 diversities, we obtain $I_0 < 10$, which could have been achieved by one infected traveler in a long-distance flight. The conclusion is robust regardless of the source population, the continuation of inputs (I_t for $t > 0$) or the fitness of the variants. With such a tiny trickle of human movement igniting many outbreaks, the crucial stage of repressing an epidemic in any region should, therefore, be the very first sign of local contagion when positive cases first become identifiable. The implications of the highly 'portable' epidemics, including their early evolution prior to any outbreak, are explored in the companion study (Ruan *et al.*, personal communication).

Keywords: SARS-CoV-2, COVID-19, population genetics, genetic drift, founder effect

INTRODUCTION

It is generally accepted that, in principle, a small number of infected people arriving in a new place could trigger an epidemic (if the basic reproduction number, R_0, is not too small [1–3]). The main issue is how many travelers *actually* started each epidemic. Here, 'a new place' may mean a country, or a bordered region, within which the bulk of human interactions happen. Relative to the within-region movement, a bordered region is lightly connected to the rest of the world. Since the epidemic in any bordered region could have been started by one single infected traveler, or by 1000 of them, we take the population genetic approach to analyzing the divergence among viral populations in relation to the 'founder effect' [4].

We shall let I_t be the amount of input at time t (i.e. the number of infected people coming into an uninfected region). The crucial number is I_0, i.e. the first batch of input. The magnitude of I_t is important in public health practice. If I_t has been large with continual input lasting for weeks, then a bordered region may be able to prevent the epidemic from being exported to (or being imported from) other regions, solely by restricting human movements out of (or into) its borders. On the other hand, if the epidemic in a region could be started with $I_0 < 10$ (and $I_{t>0} \sim 0$), then sealing off either emigration or immigration would not be effective in stopping a pandemic. Unless the bordered regions are maintaining zero infections, the danger would be coming mainly from within their borders. Here, we aim to infer I_t, in particular, in the early period of an epidemic.

In this study, we use a population genetic framework [5]. Because the focus is the stochastic differentiation among viral populations, epidemiological models generally do not cover such topics. The genetic drift formulation used here also permits the calculation of the extinction probability of the invasion of the virus ([5], Ruan *et al.*, personal communication). Epidemiological parameters, such as the number of uninfected individuals, the effects of quarantine and the development of immunity [6,7], are not considered here as population differentiation takes place in the

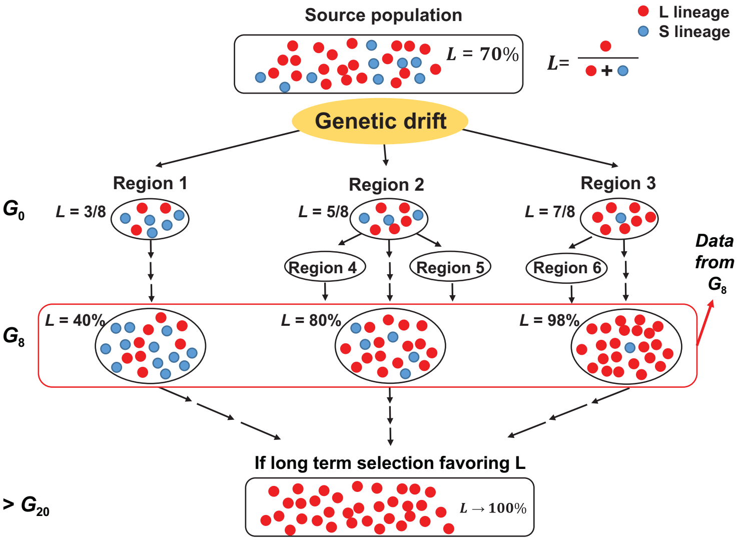

Figure 1. Schematic diagram of viral population divergence among regions. In G_0 (generation 0), $I_0 = 8$ and infected individuals arrive in regions 1–3 with 3, 5 and 7 of them, respectively, carrying the L-type virus. In the beginning, genetic drift is particularly strong and the frequency of L fluctuates as modeled by the Branching Process. G_8 is about 5 weeks after the first arrival when the data are collected. After G_8, the fluctuation is greatly dampened due to the large population size. In the later stage (after G_{20}), even weak selection could drive gene frequency toward the fixation of the more contagious genotype. Regions 4–6 are not independent samples and are not included in the analysis (see Data of the main text).

earliest stage of the invasion. During this stage, neither quarantine nor herd immunity has yet become a major factor in the outbreak.

Theory

To estimate I_t, a conventional method is to inspect the changes in the population size of the viruses, N_t [8,9]. Viral population size corresponds to the number of infected individuals, assuming a viral clone in each person. Because N_t is only weakly dependent on I_t, the conventional approach does not offer the resolution we seek for. It would be more informative to examine multiple populations for their differences in genetic polymorphism. The differences would depend strongly on I_t at the very beginning of the epidemic (Fig. 1).

For studying population differentiation, the source population infecting the travelers needs to harbor genetic variants in non-trivial frequencies to yield informative data [10]. For example, Tang *et al.* [11] reported the existence of two lineages that are distinguishable by two Single Nucleotide Polymorphisms (SNPs), one being a Serine/Lysine (S/L) polymorphism. According to ref. 11, the S lineage accounts for ∼30% and the L lineage accounts for ∼70% among the 103 viral genomes they examined.

For the ease of estimating I_t's, the variants should ideally be neutral in fitness. Indeed, since variants under selection would have a short retention time in the population, SNPs are often neutral [12–14]. While the fitness differential between the L and S variants remains unclear, our simulations show that the estimation of I_t is only weakly dependent on selection (see below).

The estimation of I_0 as well as $I_{t>0}$ are conducted for multiple viral populations, which should ideally originate from independent samples. Among these populations, we model their differentiation as a function of I_t. I_t is not likely to be very large between regions (including the source) reachable only by air flight, each of which may carry at most a small number of infected passengers. As long as all extant populations are derived from the same source population, the estimates of I_t are only weakly dependent on the actual genetic make-up of the source. For that reason, the source population need not be known.

The hypothesis is that the viral populations seeded by the infected travelers have experienced strong fluctuation in gene frequency. This may happen at the beginning of the epidemic when N_t is small. Soon afterwards, the fluctuation in gene frequencies would be quickly dampened as N_t grows. The fluctuation in gene frequencies due to the random transmissions of genes is referred to as genetic drift [14,15] or the founder effect [4]. The standard formulation of genetic drift by the Wright-Fisher model (or the alternative model of Moran [16]) is not applicable for tracking the viral population. Instead, we use the 'individual output' model we previously proposed [5]. All models assume discrete generations. Based on the infection dynamics estimated in a recent study [17], we assume that each discrete generation is ∼4 days. If we use a longer or shorter generation time, the outcome would be similar as long as the progeny production is calibrated with the generation time.

From one generation to the next, each individual produces k 'descendants' (or infects k others) with the mean of $E(k)$ and the variance of $V(k)$. In the Wright-Fisher model, k follows the Poisson distribution and $V(k)$ is tied to $E(k)$ [5]. In the 'individual output' model, k may follow any distribution, which is often measurable but not in any common form. $E(k)$ dictates the population growth, N_t, while $V(k)$ determines the fluctuation in N_t and in gene frequency. We will attempt to obtain $E(k)$ and $V(k)$ from the empirical data and, for a comparison, will also allow $V(k) = E(k)$ to approximate the Wright-Fisher model. N_t is a function of $E(k)$, $V(k)$ and I_t. Here, we assume I_t to be a constant, hence, $I_t = I_0$

for all t's. At time T,

$$E(N_T) = \sum_{t=0}^{T} I_t E(k)^{T-t}$$

$$= I_0 \frac{E(k)^{T+1} - 1}{E(k) - 1} \quad \text{when } E(k) > 1 \quad (1)$$

If $E(k)$ is not too small, $E(N_T)$ would depend mainly on I_t of the first few generations. In fact, the results are often similar whether there is constant input or not (i.e. $I_t = I_0$, or $I_t = 0$ when $t \geq 1$). In other words, Equation $(1')$ below would yield similar results to Equation (1).

$$E(N_T) = I_0 E(k)^T \quad (1')$$

With reasonable accuracy, $E(k)$ can be obtained from the growth trajectory of N_t but I_0 has to be obtained by a different means. While many of the assumptions such as exponential growth and the constancy of I_t may hold for only a few generations, most of our results depend primarily on the dynamics of the first few generations. The actual trajectory of each population would also depend on $V(k)$. Using the simpler Equation $(1')$,

$$V(N_T) = I_0 V(k) E(k)^{T-1} \frac{E(k)^T - 1}{E(k) - 1}$$

$$\text{when } E(k) > 1 \quad (2)$$

To obtain I_0 for Equations (1) or $(1')$, we have to model gene frequencies. Using the example of the S/L polymorphism, we let X_T be the frequency of the L lineage at time T. If the fitness of the S and L type is the same, then

$$E(X_T) = E(X_{T-1}) = \cdots = E(X_0) \quad (3)$$

$$V(X_T) = \frac{V(k)}{E^2(k)} \frac{X_{T-1}(1 - X_{T-1})}{N_{T-1}} \quad (4)$$

where X_0 is the frequency in the source population. Equations (3) and (4) will need some modifications if we use Equation (1), or if we consider the fitness difference between L and S (see Supplement).

Simulations

The actual realization of X_T in each population is obtained by iteration described here. We assume two types of viruses (L type and S type; [11]). The relative fitness of L type to S type is $1 + s$ ($s = 0$ represents no selection). In addition, there is a source population, in which the frequency of the L type is X_0. At generation t, there will be I_t immigrants from the source population. I_0 is the founder population size. A parameter T sets the time limit of immigration. Thus,

$$I_t = \begin{cases} 0, & \text{if } t > T \\ I_0, & \text{if } t \leq T \end{cases}$$

At generation $t - 1$, the numbers of the L type and S type are L_{t-1} and S_{t-1} respectively. Also, $N_{t-1} = L_{t-1} + S_{t-1}$ and $X_{t-1} = L_{t-1}/N_{t-1}$. After one generation, there will be I_t ($I_t = I_L + I_S$; I_L, I_S are the numbers of L and S type, respectively) immigrants from the source population. In addition, N_{t-1} will increase to $N_{t.}$. Thus, at generation t, the numbers of the L and S type are

$$L_t = I_L + \sum_{i=1}^{L_{t-1}} k_i$$

$$S_t = I_S + \sum_{j=1}^{S_{t-1}} k_j$$

where k_i is the progeny number of the i-th individual of either type. The distribution of k_i will be defined in the next section. If there is selection, the number of L type will change as follows:

$$L_t = (1 + s) L_t$$

At generation t, the population size and the frequency of L type are

$$N_t = L_t + S_t$$

$$X_t = L_t / N_t$$

Based on the definition above, we simulate the stochastic changes of the viral population until it reaches the 20th generation (i.e. $t = 20$). Since genetic drift is negligible when N_t is large (e.g. $> 10^5$), we simulate the trajectory by a deterministic model when $N_t > 10^5$.

To quantify the population differentiation, we calculate the pairwise Fst values [14,18], defined below. With X_t and Y_t for a pair of populations,

$$Fst = 1 - \frac{X_t(1 - X_t) + Y_t(1 - Y_t)}{2\bar{p}(1 - \bar{p})} \quad (5)$$

where $\bar{p} = (X_t + Y_t)/2$. If Fst $= 0$, $X_t = Y_t$ and the two populations are identical in gene frequency. If Fst $= 1$, the two populations are maximally differentiated with $X_t = 0$ and $Y_t = 1$, or vice versa.

Table 1. Two computationally generated datasets for the frequency of the L lineage.

Region	A1[a]	A2	A3	B1	C1	C2	C3	D1	D2[b]	D3
I (hypothetical)	0.73	0.70	0.71	0.64	0.75	0.61	0.70	0.74	0.82	0.69
II (realistic)	0.70	0.95	0.40	0.80	1.00	0.99	0.50	0.33	0.95	0.67

[a] Each letter in the region code indicates a separate continent and the number indicates a country or region; [b] D2 could have been derived from C1 or C2 due to the inter-continental travel pattern. Its exclusion from the analysis would increase the spread of Fst in Figs 2–4 and would thus lead to a smaller I_0 estimate.

Defining the parameter set (I_0, T, X_0, s) and the distribution of k

We set $I_0 = 1, 5, 10, 20, 50, 100$ and $T = 0, 1, 2, 3, 20$. Thus, $(I_0, T) = (5, 3)$ means five travelers each generation for four generations, counting $T = 0$ as the initial batch. We set T in this range but will show that $T = 0$, 1 or 20 hardly matters. The frequency of L lineage in the source population is $X_0 = 0.1, 0.3, 0.5, 0.7, 0.9$. Again, the polymorphism frequency in the source population has turned out to have little effect on the differentiation among populations. In addition, we also set $s = 0, 0.1$ to study the effect of selection. For each parameter set (I_0, T, X_0, s), we repeat the simulation 100 times.

As stated, the conventional Wright-Fisher Model requires k (progeny number of an individual) to follow a Poisson distribution with $V(k) = E(k)$. Here, we assume the spread of virus to be associated with the social network, which usually follows the power law [19,20]. Specifically, we let k follow Zipf's law (a discrete power-law distribution; [21]):

$$P(k = i; c, M) = \frac{1/(i+1)^c}{H_{M,c}}, \quad i = 0, 1, \ldots, M-1$$

where $H_{M,c} = \sum_{m=1}^{M} 1/m^c$. The mean and variance of k are

$$E(k) = \frac{H_{M,c-1}}{H_{M,c}} - 1 \quad (6)$$

$$V(k) = \frac{H_{M,c-2}}{H_{M,c}} - \frac{H_{M,c-1}^2}{H_{M,c}^2} \quad (7)$$

The estimate of the basic reproduction number (R_0) of SARS-CoV-2 ranges from 1.4 to 6.5 [22–24]. Here, we focus on the early phase of the viral population growth by using $R_0 = 6.5$. The relationship between $E(k)$ and R_0 is as follows:

$$N_t = N_0 R_0^{t/\tau} = N_0 E(k)^{t/G} \quad (8)$$

where τ is the serial interval and G is the generation time, and τ is estimated to be ~5 days [23,25] and G is 4 days [17]. Then, $E(k) = R_0^{G/\tau} = 4.5$. (Note that $E(k)$ would become smaller with smaller G, as stated above.) To have k follow the power law with $E(k) = 4.5$, we assume $M = 30$ and $c = 1.3$ in Equation (6), which yields $V(k) = 45$. As noted in Chen YX *et al.* [5], the strength of genetic drift depends mainly on $E(k)$ and $V(k)$ rather than the actual distribution. By setting $V(k)$ so large, we ensure that the I_0 estimate would be on the high side (see Discussion).

Data

The estimation of I_0 as well as $I_{t>0}$ should be done on multiple viral populations that are independent samples from the same source. If they are not fully independent, then our estimates of I_t would be conservative (i.e. over-estimation) since any exchange between populations should reduce the divergence.

Here, we generate two sets of data as shown in Table 1. In the hypothetical Set I, the gene frequency is taken from a normal distribution with the mean of 0.7 and standard deviation of 0.06. The mean is close to the average frequency of the L type in this pandemic. Set I is generated to show the possible pattern of population divergence if I_0 is in the hundreds.

In Set II, we assign gene frequencies to 10 populations in Table 1 using the reported frequencies as a guide (GISAID (https://www.gisaid.org/); see Supplement). These 10 populations, distributed among four continents, are as likely to be independently derived as we could ascertain. The choice is based on three criteria: (i) the geography of the countries/regions and the distance between them; (ii) the timing of the documented onset of the epidemic; (iii) the abundance of DNA sequences. We consult the frequencies in samples collected before late March 2020, corresponding roughly to G_8 in Fig. 1. Due to the rapidly changing data reporting (GISAID [26]), the frequency profile of Set II is plausibly realistic as reported in mid-April (see Supplement). Readers with access to the more up-to-date data can compare the new observations with the theoretical results to improve the estimation of I_t.

Comparisons between simulations and data

In Fig. 2, the Fst distributions based on the datasets of Table 1 are presented. The two very different

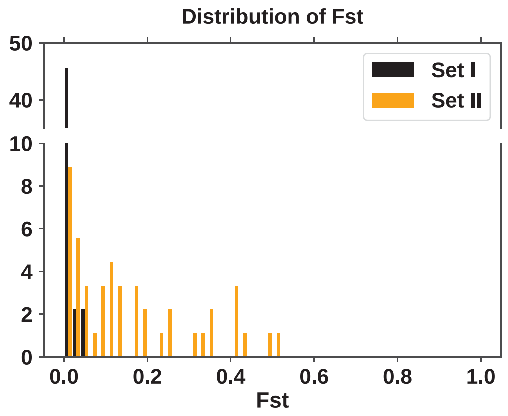

Figure 2. The pairwise Fst distribution from the two datasets of Table 1. Fst is calculated by Equation (5). Given 10 regions, there are 45 (= 10 × 9/2) pairwise comparisons.

distributions would be informative when compared with the simulated distributions.

Given the large number of possible combinations of parameters, (I_0, T, X_0, s), the task could have been daunting. Fortunately, all but one parameter would have little impact on the results. In Fig. 3A–D, only X_0 varies and the Fst distributions are very similar. The simplest explanation is that small populations would all deviate from the initial frequency, X_0, regardless of where it is. Figure 3E–H shows that the results do not depend strongly on T either, T being the time duration of traveler input. Since, at $T = 2$, the number of infected individuals would have swelled by 20-fold without any new input, the results are not significantly affected by later inputs.

In Fig. 4, the value of I_0 is varied from 10, 50 to 100. Here, we assume no travelers arriving after the first generation because their contributions, as shown in Fig. 3E–H, are insubstantial. The left panels (Fig. 4A, C and E) show the trajectories of 100 populations, which diverge in the first generation or two and then evolve steadily as the population size increases. On the right panels are the comparisons between the simulated and 'observed' distributions; the latter being those based on Dataset II (Fig. 4B and D) or Dataset I (Fig. 4F). It is clear that the expected distribution of Fst is very sensitive to I_t (I_0, in particular). The more realistic Dataset II can only be explained by $I_0 \leq 10$ whereas the artificial Dataset I would agree with $I_0 \geq 100$. We should note that the extensive survey of parameter space (see Supplement) shows the conclusion of $I_0 \leq 10$ to be robust.

In the results presented above, L and S lineages are assumed neutral. Intuitively, selection should drive the trajectories to converge and the left panels (Fig. 5A, C and E) do show that trend. If we let the L type enjoy a 10% selective advantage, the results of Fig. 5 still indicate $I_0 < 10$. Note that the range of I_0 spans a smaller range in Fig. 5 than that in Fig. 4. In other words, with selection, the estimated I_0 should be even smaller than indicated above.

The simulation results are based on the distribution of k that follows the power law (see Equations 6 and 7) with $V(k) \sim 10E(k)$. Such a large $V(k)$ means that the genetic drift would be very large, requiring large I_t's to reduce the drift. It is hence interesting that, even under such stringent conditions, I_0 is still $<$10. In the Supplement, we show that the estimated value of I_0 would be substantially lower if we use the Poisson distribution of k, associated with the conventional Wright-Fisher model. With $V(k) = E(K)$, I_0 would be 2–4. Hence, the conclusion presented in this section is robust.

Inference of parameters (I_0, T, X_0, s)

In the last section, we present a range of parameter values that yield the expected population divergence for a comparison with the data. To corroborate the visual comparisons, we also carry out formal inferences using the ABC (Approximate Bayesian Computation) procedure on Dataset II (see Supplement). The results of Fig. S3 and S4 indeed confirm the visual impression as the observed divergence is sensitive only to I_0, but not to the variation in T, X_0 and s. The formal inference of I_0 at 2–5 is even smaller than the visual impression would suggest. Finally, this analysis considers only the number of infections that can spread the virus further. A more extensive model that incorporates the development of symptoms, the border control and the local quarantine, all of which may contribute to the suppression of the epidemics, will be presented later (Ruan *et al.*, personal communication).

DISCUSSION

In the theory of genetic drift [5], even 100 infected travelers from a source viral population would give rise to a fairly uniform level of genetic polymorphism among bordered regions. In contrast, the reported data indicate substantial divergence among countries (GISAID (https://www.gisaid.org/); see Supplement). Dataset II of Table 1 is realistic in this respect. The divergent polymorphisms across countries depend mainly on a critical parameter—the size of the first cohort arriving in a country, I_0, which is estimated to be $<$10. The number may in fact be smaller than it seems since a long distance flight carrying one single infected but symptomless patient could infect this many people, all of whom would be without symptoms upon arrival [27].

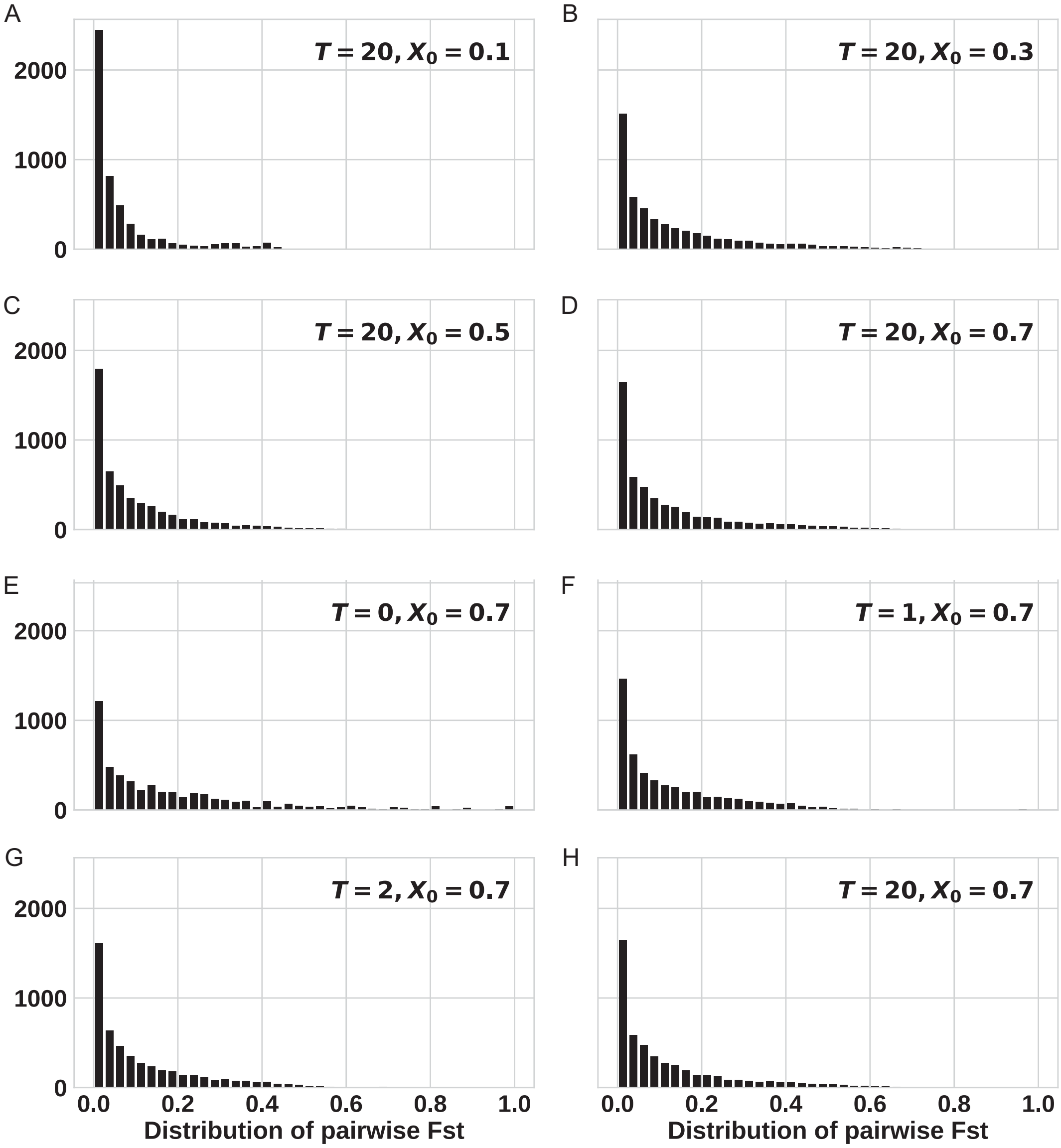

Figure 3. The Fst distribution at G20. For each parameter set of (I_0, T, X_0, s), we repeat the simulations 100 times. For all panels, $I_0 = 10$, $s = 0$. Panel A–D, $T = 20$ and X_0 ranges from 0.1 to 0.7. Panel E–H, $X_0 = 0.7$, T ranges from 0 to 20. These eight panels show that neither X_0 nor T would impact the Fst distribution much.

On the robustness of the estimation

In Figs 3 and 5, we show that, despite the complexity of the model with many parameters, none of them (T, X_0, s), except I_0, plays a significant role in the divergence among viral populations. As discussed in conjunction with Fig. 3, the distribution of k does not matter either. In fact, in the standard Wright-Fisher model, the estimate of I_0 would be <5. It is also noted that the $E(k)$ and $V(k)$ values used are for a generation time of 4 days. For a shorter generation time, the values would be correspondingly smaller and the results should be similar.

The model also assumes that each population is an independent sample of the source population. Since all populations are likely to exchange some individuals due to traveling, the actual divergence among populations would be even smaller than simulated. In other words, to attain the observed level of divergence, I_0 would have to be even smaller than estimated. Considering all these variables, we believe the conclusion of $I_0 < 10$ to be robust. In the analysis of regional divergence, the results would depend strongly on the smaller I_0 of the two regions being compared. Hence, the assumption

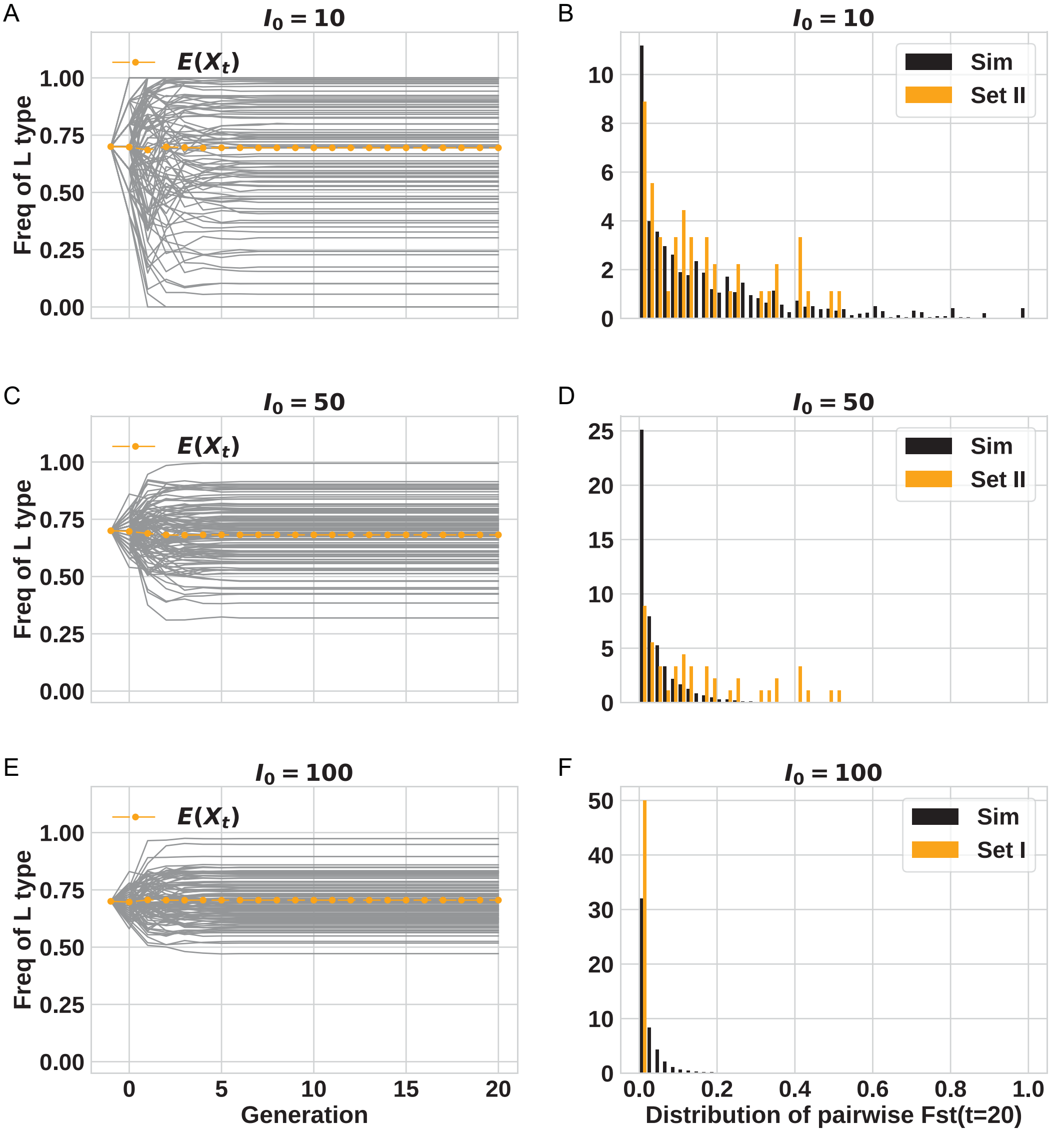

Figure 4. Comparisons between data and simulations with various I_0 values. For all panels, $s = 0$ (no selection), $X_0 = 0.7$ and $T = 0$. The value of I_0 is shown next to each panel. Panel (A, C and E): the frequency of L type over time (100 repeats), the average is showed by the orange dotted line. Panel (B and D): Fst distribution; the simulation results are in black and the distributions from Dataset II (realistic data) are in orange. Panel (F): like panels B and D but Dataset I is used.

of the same I_0 among all regions would be a reasonable one.

Subsequent viral evolution after arriving in a new continent

While our focus is on the divergence in the first few generations, we now briefly discuss the subsequent evolution after this initial critical period. The primary lineage delineation, the S/L polymorphism defined by two SNPs [11], has many subtypes (see Supplementary data and Table S1 for details). For example, western European countries including Italy, Switzerland, Germany and Belgium are predominantly of the L type with a similar abundance in the L2 subtype. In contrast, while Japan is also predominantly of the L type, it has mainly the L1 subtype. This contrast suggests that Japan may represent an independent sample from the western European samples, which have likely been spreading regionally after the initial seeding. Another example is the S1 and S2 subtypes, which differentiate between the samples from China and the west coast of the US.

These patterns suggest that, after the initial seeding, each major region or continent has been evolving along an independent path. Since the initial seeding may be extremely difficult to prevent, the onus is to suppress the regional spread. The analyses

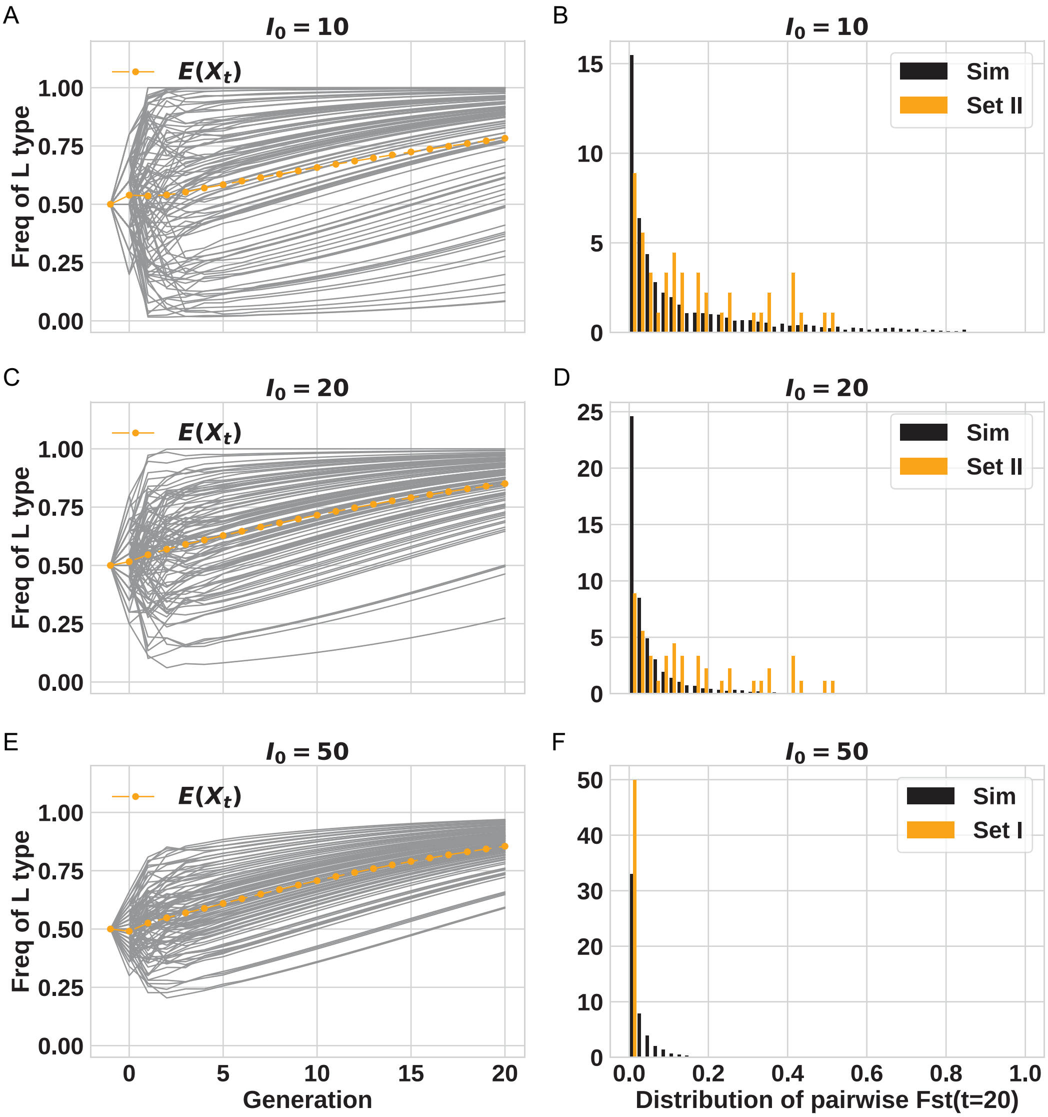

Figure 5. Same as Fig. 4 except that (i) $s = 0.1$ with selection favoring the L type; (ii) $X_0 = 0.5$ so the L type would not reach fixation so quickly.

of the subtypes in Asia, Australia and various parts of North America would offer additional details of the spread of the virus, as has been done recently [28–31]. These details are beyond the scope of this study, which focuses on the early stages of the viral spread.

Implications

The analysis suggests that the COVID-19 epidemic in each region surveyed was likely started by a very small number of travelers ($I_0 < 10$). With such a tiny trickle of human movement, it would have been very inefficient for any region to prevent infected individuals from exporting an epidemic to (or importing it from) other places. For that reason, the crucial stage of repressing an epidemic in any region should be the very first sign of local contagion.

Finally, due to the 'portability' of COVID-19, each epidemic, including the first one on record, could have easily been imported. Where then did all these epidemics begin? While the interest in the 'origin' is intense, we suggest the question be broadened as 'the origin and early evolution' of SARS-CoV-2. The latter implies a process whereas the former seems to mean a single time point. The process of early evolution may have stretched over different regions in a long time-span and involved multiple host species. Like many other evolutionary questions on origin, we suggest the question be phrased as the early evolution of SARS-CoV-2, rather than be about the 'origin'. The former implies a process whereas the latter seems to mean a single time point.

This distinction is important as seen in the debates on the 'origin' of dogs [32,33] and new species in novel environments [34]. By compressing a process into a simple 'origin', we may be asking a false question about, say, 'the first dog' or 'the first patient'. The possible early evolution of SARS-CoV-2 is addressed in the companion study (Ruan *et al.*, personal communication).

SUPPLEMENTARY DATA

Supplementary data are available at *NSR* online.

ACKNOWLEDGEMENTS

We thank all those who have contributed sequences to the GISAID (Global Initiative on Sharing All Influenza Data) database (https://www.gisaid.org/). Acknowledgements and detailed information of the genome sequences we consulted for this study are given in Table S2. The actual data fitted to our model were computationally generated (see the main text). We thank Drs Yaping Zhang, Suhua Shi, Weiwei Zhai, Jianzhi George Zhang, Bingjie Chen, Zheng Hu and Jindong Zhao for comments on and suggestions for this manuscript.

FUNDING

This work was supported by grants from the Ministry of Science and Technology of the People's Republic of China (No. 2020YFC0847000) to H.W., X.L. and J.L. This work was funded by the National Natural Science Foundation of China (31730046, 91731000 to C.I.W), Guangdong Basic and Applied Basic Research Foundation (2020B1515020030, 2019A1515010708), and CAS 'Light of West China' Program.

Conflict of interest statement. None declared.

REFERENCES

1. Smith DL, Battle KE and Hay SI *et al.* Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. *PLoS Pathog* 2012; **8**: e1002588.
2. Rothman KJ, Greenland S and Lash TL. *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins, 2008.
3. Woolhouse ME. Population biology of emerging and re-emerging pathogens. *Trends Microbiol* 2002; **10**: S3–7.
4. Rambaut A, Posada D and Crandall KA *et al.* The causes and consequences of HIV evolution. *Nat Rev Genet* 2004; **5**: 52–61.
5. Chen Y, Tong D and Wu CI. A new formulation of random genetic drift and its application to the evolution of cell populations. *Mol Biol Evol* 2017; **34**: 2057–64.
6. Hethcote HW. The mathematics of infectious diseases. *Siam Review* 2000; **42**: 599–653.
7. Zhao S and Chen H. Modeling the epidemic dynamics and control of COVID-19 outbreak in China. *Quant Biol* 2020; **8**: 11–9.
8. van den Driessche P. Reproduction numbers of infectious disease models. *Infect Dis Model* 2017; **2**: 288–303.
9. Heffernan JM, Smith RJ and Wahl LM. Perspectives on the basic reproductive ratio. *J R Soc Interface* 2005; **2**: 281–93.
10. Poo M-m and Wu C-I. Moral imperative for immediate release of 2019-nCoV sequence data. *Natl Sci Rev* 2020; **7**: 719–20.
11. Tang X, Wu C and Li X *et al.* On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev* 2020; **7**: 1012–23.
12. Kimura M and Ohta T. The average number of generations until fixation of a mutant gene in a finite population. *Genetics* 1969; **61**: 763–71.
13. Kimura M. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press, 1983.
14. Crow JF and Kimura M. *An Introduction to Population Genetics Theory*. New York, Evanston and London: Harper & Row Publishers, 1970.
15. Li W-H. *Molecular Evolution*. New York; Basingstoke: W.H. Freeman; Palgrave (distributor), 2007.
16. Moran PAP. Random processes in genetics. *Math Proc Cambridge Philos Soc* 1958; **54**: 60–71.
17. Wolfel R, Corman VM and Guggemos W *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature* 2020; **581**: 465–9.
18. Hartl DL and Clark AG. *Principles of Population Genetics*. Massachusetts: Sinauer Associates, 1997.
19. Muchnik L, Pei S and Parra LC *et al.* Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Sci Rep* 2013; **3**: 1783.
20. Newman M. *Networks: An Introduction*. Oxford: OUP, 2010.
21. Zipf GK. *Human Behavior and the Principle of Least Effort*. Oxford: Addison-Wesley Press, 1949.
22. Liu Y, Gayle AA and Wilder-Smith A *et al.* The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J Travel Med* 2020; **27**: taaa021.
23. Ferretti L, Wymant C and Kendall M *et al.* Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* 2020; **368**: eabb6936.
24. Kucharski AJ, Russell TW and Diamond C *et al.* Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infect Dis* 2020; **20**: 553–8.
25. Tindale L, Coombe M and Stockdale JE *et al.* Transmission interval estimates suggest pre-symptomatic spread of COVID-19. *eLife* 2020; **9**: e57149.
26. Shu Y and McCauley J. GISAID: global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* 2017; **22**: 30494.
27. Li R, Pei S and Chen B *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* 2020; **368**: 489–93.
28. Rambaut A, Holmes EC and O'Toole A *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020; **5**: 1403–7.
29. Bai Y, Jiang D and Lon JR *et al.* Evolution and molecular characteristics of SARS-CoV-2 genome. *bioRxiv* 2020; doi: 10.1101/2020.04.24.058933.

30. Fauver JR, Petrone ME and Hodcroft EB *et al*. Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. *Cell* 2020; **181**: 990–6.
31. Gonzalez-Reiche AS, Hernandez MM and Sullivan MJ *et al*. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* 2020; **369**: 297–301.
32. Wang GD, Shao XJ and Bai B *et al*. Structural variation during dog domestication: insights from gray wolf and dhole genomes. *Natl Sci Rev* 2019; **6**: 110–22.
33. Wei FW. Structural variation provides novel insights into dog domestication. *Natl Sci Rev* 2019; **6**: 123.
34. He Z, Li X and Yang M *et al*. Speciation with gene flow via cycles of isolation and migration: insights from multiple mangrove taxa. *Natl Sci Rev* 2019; **6**: 275–88.