Self-organized critical dynamics of RNA virus evolution

Xiaofei Ge
School of Pharmaceutical Sciences, Tsinghua University, Beijing, 100084, China.
E-mail: gxf20@mails.tsinghua.edu.cn

Kaichao You
School of Software, Tsinghua University, Beijing, 100084, China.
E-mail: ykc20@mails.tsinghua.edu.cn

Zeren Tan
Institute for Interdisciplinary Information Science, Tsinghua University, Beijing, 100084, China.
E-mail: tanzr20@mails.tsinghua.edu.cn

Hedong Hou
UFR de Mathématiques, Université de Paris, Paris, 75013, France.
E-mail: hedong.hou@etu.u-paris.fr

Yang Tian and Pei Sun
Department of Psychology & Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, 100084, China.
E-mail: tiany20@mails.tsinghua.edu.cn; peisun@tsinghua.edu.cn

February 2021

Abstract. RNA virus (e.g., SARS-CoV-2) evolves in a complex manner. Studying RNA virus evolution is vital for understanding molecular evolution and medicine development. Scientists lack, however, general frameworks to characterize the dynamics of RNA virus evolution directly from empirical data and identify potential physical laws. To fill this gap, we present a theory to characterize the RNA virus evolution as a physical system with absorbing states and avalanche behaviors. This approach maps accessible biological data (e.g., phylogenetic tree and infection) to a general stochastic process of RNA virus infection and evolution, enabling researchers to verify potential self-organized criticality underlying RNA virus evolution. We apply our framework to SARS-CoV-2, the virus accounting for the global epidemic of COVID-19. We find that SARS-CoV-2 exhibits scale-invariant avalanches as mean-field theory predictions. The observed scaling relation, universal collapse, and slowly decaying auto-correlation suggest a self-organized critical dynamics of SARS-CoV-2 evolution.
Interestingly, the lineages that emerge from critical evolution processes coincidentally match with threatening lineages of SARS-CoV-2 (e.g., the Delta virus). We anticipate our approach to be a general formalism to portray RNA virus evolution and help identify potential virus lineages to be concerned.

1. Introduction

With high mutation and replication capacities, RNA viruses exhibit intricate evolution [1, 2, 3]. Studying the complex evolution dynamics of RNA virus remains as an attractive challenge in biology, as RNA virus evolution is a frequent cause of new pathogens and may be a pivotal route in medicines [4, 5, 6]. Meanwhile, RNA virus interests physicians for its potential to serve as a representative model of long-term molecular evolution [7, 8].

Despite its apparent significance, the physics underlying RNA virus evolution remains elusive. While important progress has been accomplished in characterizing the dynamics of mutant virus population size (e.g., see Refs. [9, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]), there remains another non-negligible aspect of evolution for exploration. From the biological perspective, the dynamics accountable for phylogenetic tree formation is informative for the global evolution properties [24, 25, 26, 27]. Compared with the evolution previously considered in the context of population dynamics [9, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23], this unexplored aspect may convey key information about the macroscopic laws of RNA virus evolution [24, 25, 26, 27] but yet remains unclear. Moreover, existing models either non-trivially handle multiple time scales with extreme differences in magnitude orders or are simplified by phenomenological parameters [23]. Although phenomenological models are more solvable on empirical data than non-trivial analytic ones, these models are powerful in imitating phenomena but limited in explaining mechanisms. Interpretations of phenomenological parameters with experiments should be prudently considered [23]. In sum, much unknown remains about the general frameworks that are applicable to identify potential physical laws of RNA virus evolution from empirical data.

The objectives of the current research are twofold. We aim at exploring the global properties of RNA virus evolution by developing a physics theory of the dynamics underlying phylogenetic tree formation. The theory is demanded to map fundamental biological factors to general characterizations of system dynamics so as to guarantee its applicability on real data and universality to arbitrary RNA viruses. Applying this theory, we present a systematic analysis of SARS-CoV-2, the virus accounting for COVID-19 [28]. The global infection of SARS-CoV-2 provides us a unique opportunity with rich data (e.g., a sufficiently large phylogenetic tree) to validate our theory. Moreover, confirming the macroscopic characteristics of SARS-CoV-2 evolution may benefit to developing coping strategies of COVID-19.
2. Backgrounds of RNA virus

To make our analysis comprehensible, we begin with a fundamental perspective of RNA virus. In RNA virus populations, the replication implies the vanishing of the original RNA virus and the birth of new RNA viruses that can also replicate themselves [29]. The proliferation of RNA virus is bounded by time-dependent host cell resource $\lambda(t)$ [1, 8, 30]. Specifically, the maximum cumulative mutation rate, $\sup_\omega \hat{\zeta}(\omega)$, of an RNA virus population of initial size $\omega \in \mathbb{N}^+$ during a duration of $[0, \kappa]$ is limited by $\lambda(t)$ in the following way

$$\sup_\omega \sup_\lambda \hat{\zeta}(\omega) = \omega \sigma \tau \left( \phi^{[\kappa/\tau]} - 1 - \frac{1}{\phi} \right),$$  \hspace{1cm} (1)

where $\sigma$ is the mutation rate of single RNA virus, parameter $\tau$ denotes entrance and eclipse periods (assume $\tau$ is divisible by $\kappa$), and $\phi \approx 10^3$ denotes burst size (the number of viruses produced by a host cell) [1, 8, 30]. Here $\sup_\omega \hat{\zeta}(\omega)$ can reach the maximum cumulative mutation rate of an RNA virus population under ideal conditions when host cell resources are ample $\lambda(t) \geq \phi^{[t/\tau]} - \phi^{[t/\tau]} - 1$. Derivations of Eq. (1) are presented in Appendix A.

In summary, host cell resource $\lambda(t)$ intrinsically determines the cumulative daily mutation rate of a RNA virus population. Instead of proposing detailed definitions of $\lambda(t)$ as previous studies (e.g., define life circles of cells) [13, 15, 17, 19, 20, 22, 23], we leave the characterization of $\lambda(t)$ as a task of epidemic and observe the evolution dynamics of RNA virus controlled by $\lambda(t)$ directly from empirical data.

3. Theory of RNA virus evolution

Certainly, one can not expect to count $\lambda(t)$ in a real epidemic. This parameter can be reasonably replaced by the number of human hosts. Assuming that a lineage $l_i$ of RNA virus emerges at moment $t'$ and one of its progeny lineage $l_j$ is born at moment $t''$, we can measure the duration length of evolution as $T = t'' - t'$. Meanwhile, we define $S(T)$ as the number of cumulative confirmed patients with lineage $l_i$ during $[t', t'']$ and denote the time-dependent number of confirmed patients with lineage $l_i$ by $S(t \mid T)$ (here $t \in [t', t'']$). The terminology “lineage” can be generally understood as a kind of definition of virus sub-type (e.g., see examples in Ref. [31]). Viruses of lineage $l_i$ need to accumulate sufficient mutations to make lineage $l_j$ emerge.

We can relate these concepts with a branching process concerning the number of copies of $l_i$ during $[t', t'']$, that is, a process where a patient with lineage $l_i$ infects multiple people before the virus mutates to lineage $l_j$ during replication. The realization of this branching process, from the initialization $t'$ to the termination $t''$ after which a new lineage $l_j$ emerges, is referred to as an avalanche. The size and life time of avalanche are exactly $S(T)$ and $T$. This branching process is slightly different from the standard one, where the termination is defined as a moment after which the population size (e.g., the number of patients with lineage $l_i$) remains 0 indefinitely [32, 33].
We can further relate the branching process of evolution with directed percolation, a universality class exhibiting a phase transition separating an absorbing state from an active state \[32,34,35\]. In directed percolation (e.g., the contact process \[36\]), sites (e.g., people) can be either active (infected) or inactive (healthy). With different balance between infection and recovery, the infection process may propagate over the system or vanish gradually. If infection vanishes, the system is trapped in a completely inactive state, the so-called absorbing state. Directed percolation is a non-equilibrium process since detailed balance breaks in the absorbing state (it can be reached but not be escaped) \[37\].

We are interested in potential self-organized criticality (SOC) due to its pervasiveness in biological systems (e.g., neural avalanches in the brain \[38,39\]). The self-organization to criticality distinguishes SOC (e.g., Refs. \[40,41\]) from ordinary critical phenomena \[37,42\]. At criticality, the system jumps between absorbing configurations by avalanches \[37,42\]. From the biological perspective, this property implies an unpredictable and unpreventable evolution of RNA virus. Numerous mean field theories of directed percolation have been proposed to determine avalanche exponents to study the scale-invariance underlying \(P_T\) and \(P_S\), the probability distributions of \(T\) and \(S\) (e.g., see Ref. \[43\]). Note that \(S\) is an abbreviation of \(S(T)\). In Appendix B, we re-calculate avalanche exponents following a similar idea of Refs. \[44,32,45\]. Our results show that

\[
P_T(t) \propto t^{-2},
\]

\[
P_S(s) \propto s^{-3/2},
\]

consistent with other mean field theory predictions of self-organized criticality \[43,44\].

To verify whether the system is at criticality, we can consider a scaling relation \[46,47,37,42\] in Eq. (4), where \(P_T(t) \propto t^{-\alpha}, P_S(s) \propto s^{-\beta}\), and \(\langle S(T) \rangle \propto T^\gamma\).

\[
\gamma = \frac{\alpha - 1}{\beta - 1}.
\] Inserting \(\alpha = 2\) and \(\beta = \frac{3}{2}\) into Eq. (4), we derive \(\gamma = 2\). Apart from that, more precise verification of criticality can be implemented according to the collapse shape \[47,39\] and the slow and exponential decay of auto-correlation \[39\]. Please see Appendix C for details of these approaches.

With host cell resource \(\lambda(t)\) as a clue, we have derived a directed percolation theory to characterize RNA virus evolution as a system with absorbing states and avalanche behaviors. A series of approaches are presented to verify the existence of self-organized criticality.

4. Evolution of SARS-CoV-2

Applying this theory, we aim at presenting a systematic analysis of SARS-CoV-2, the virus that causes COVID-19 \[28\]. We use an open-source database \[48\] to acquire the
necessary data. The criterion of lineage definition can be seen in [31]. Our data includes all lineages that emerge before 27 December 2021.

In the SARS-CoV-2 database [48], the probability for a patient to be infected with each lineage of virus is estimated with confidence intervals. By multiplying the midpoint of confidence interval with the total number of worldwide patients recorded by the World Health Organization (WHO) [49], we can estimate the accumulative and daily numbers of confirmed patients with each lineage of virus, corresponding to \( S(T) \) and \( S(t \mid T) \) of each lineage (see Fig. 1a). Similarly, we can estimate the upper and lower bounds of \( S(T) \) and \( S(t \mid T) \) based on the upper and lower bounds of confidence interval.

In Fig. 1b, we show the probability distributions of life time \( T \) and avalanche size \( S \) (for midpoint and bounds). In real cases, the power-law may only hold on specific tails \( \{ P(X) \mid X \geq X' \} \) of the empirical distribution of variable \( X \), where \( X' \) denotes a certain distribution cutoff [50]. We apply the approach in Ref. [51] to estimate cutoffs \( T' \), \( S' \), \( S'_{\text{L}} \) (lower bound), and \( S'_{\text{U}} \) (upper bound) as the points where the Kolmogorov-Smirnov (KS) statistic \( \eta \) reduces to a sufficiently small threshold at the first time (Fig. 1c). Details of cutoff estimation can be seen in Appendix D. Then, a maximum likelihood estimation of power-law exponent [51] is implemented on the samples above cutoffs to find ideal exponents that maximize the likelihood \( \mathcal{L} \) (Fig. 1c) [51]. In our results, exponents are estimated as \( \alpha = 1.69 \) (\( v = 7.13\% \)), \( \beta = 1.32 \) (\( v = -4.53\% \)), \( \beta_- = 1.32 \)

---

**Figure 1:** Avalanches of SARS-CoV-2 evolution. **a** Numbers of accumulatively and daily confirmed patients with viruses of different lineages (scatters with different colors) and the total accumulative number of patients (black dashed line). **b** Data distributions of \( P_T(\cdot) \) vs. \( T \), \( P_S(\cdot) \) vs. \( S \), and \( \langle S \rangle(T) \) vs. \( T \). **c** The KS statistic \( \eta \) for estimating distribution cutoffs and the likelihood \( \mathcal{L} \) for estimating avalanche exponents. Blue stars denote estimated values. **d** The power-law distributions of \( T \) and \( S \) (midpoint and bounds of confidence interval) above cutoffs.
(\(v = -5.06\%\)), and \(\beta_+ = 1.34\ (v = -14.88\%\)), where \(v < 10\%\) is a reasonable standard for the goodness of ideal estimations. Our estimated power-law models are shown in Fig. 1d. Please see Appendix D for details of power-law exponent estimation.

Meanwhile, we implement least square fitting on \(\langle S \rangle (T)\) vs. \(T\) to derive \(\gamma_+ = 1.8431 (R^2 = 0.9112)\), \(\gamma = 1.9363\ (R^2 = 0.8871)\), and \(\gamma_- = 1.9944\ (R^2 = 0.8652)\) in Fig. 1b. Together with our estimated avalanche exponents \(\alpha\) and \(\beta\), these results coincide with the scaling relation in Eq. (4), suggesting potential criticality.

To analyze criticality more precisely, we calculate collapse shape following Refs. [47, 39]. Please see details of collapse analysis in Appendix C. To ensure the robustness of our results, we further verify if the estimators of distribution cutoffs for \(T\) and \(S\) affects collapse shape. In Fig. 2a, we show \(\langle V(t \mid T)\rangle\) vs. \(\frac{t}{T}\) of avalanches with \(T \geq T'\) and \(S \geq \hat{S}\), where \(\hat{S} \in \{S', \ldots, 200S'\}\) (the main panel of Fig. 2a). By increasing \(\hat{S}\), we can simultaneously increase the distribution cutoff of \(T\). Our results suggest that the trend of \(\langle V(t \mid T)\rangle\) is principally independent of \(\hat{S}\). Meanwhile, we notice that \(\langle V(t \mid T)\rangle\) exhibits a mixture of global increasing trend and parabolic trend. The increasing trends may reflect the intrinsic property of the raw data. Therefore, we recalculate \(S\) by multiplying the midpoint of confidence interval [48] with the average total number of worldwide patients in the WHO database [49] for detrending. In Fig. 2a, \(\langle V(t \mid T)\rangle\)
Self-organized critical dynamics of RNA virus evolution

plausibly exhibits a parabolic trend in the detrended data. Meanwhile, we show several instances of $V(t | T)$ vs. $\frac{1}{T}$ in Fig. 2a. In Fig. 2b, the parabolic trend of $\langle V(t | T) \rangle$ is verified quantitatively by quadratic polynomial fitting. We show the average quadratic coefficient $\langle \mu \rangle$ (averaged across avalanches under each condition of $\widehat{S}$) as a function of $\widehat{S}$. Consistent with Fig. 2a, $\langle \mu \rangle < 0$ holds for every $\widehat{S}$ in both original and detrended data, suggesting plausible parabolic trends of $\langle V(t | T) \rangle$. The probability distributions of $\langle \mu \rangle$ and the goodness of fitting, $R^2$, are also shown.

In Fig. 2c, we show the auto-correlation (covariance) Cov ($S(t_i | T), S(t_j | T)$) under each condition of $\widehat{S}$, where we select $t_i \in \{0.3T, 0.5T\}$ as instances. Please see Appendix C for details. In Fig. 2d, the corresponding decays of auto-correlation are fitted to derive the decay rate $\xi$. In the raw data, we obtain $\langle \xi \rangle = -0.5407$ (averaged $R^2 = 0.6301$) for $t_i = 0.3T$ and $\langle \xi \rangle = -0.5159$ (averaged $R^2 = 0.6324$) for $t_i = 0.5T$. In the detrended data, we find that $\langle \xi \rangle = -1.6679$ (averaged $R^2 = 0.6301$) for $t_i = 0.3T$ and $\langle \xi \rangle = -1.4251$ (averaged $R^2 = 0.6332$) for $t_i = 0.5T$. These results suggest exponential decays of auto-correlation with reasonable errors, offering a practical verification of potential criticality.

Interestingly, our results coincide with threatening lineages of SARS-CoV-2 if we filter directed edges in its original phylogenetic tree (edges correspond to avalanches and indicate evolution paths of lineages) as following: an edge is removed unless its avalanche satisfies $S \geq S'$ and $T \geq T'$ (Fig. 3). The filtered result, referred to as critical phylogenetic tree, is a sub-graph of the original one that includes only critical avalanches. Threatening lineages, such as B.1.167.2 (the Delta virus), B.1.1 (the parent of Alpha, Beta, and Omicron viruses), and B.1 (the parent of Epsilon, Eta, Iota, and Mu viruses), can be found by searching non-isolated lineages with high out-degree $\delta$ (number of offspring lineages) in the critical tree. Note that the out-degree in the original tree does not completely determine the possibility of being isolated in the critical tree. Numerous lineages with high out-degrees in the original tree still become isolated (e.g., 20.83%,

![Figure 3: Original and critical phylogenetic trees. Edges are colored according to $T$ of corresponding avalanches.](image-url)
8.33%, and 9.09% of lineages with $\delta \geq 5$, $\delta \geq 15$, and $\delta \geq 20$ in the original tree become isolated after filtering).

5. Conclusion

Following a clue that host cell (human host) resources shape RNA virus mutation and replication, we have naturally derived a theory to characterize RNA virus evolution as a system with absorbing states and avalanches. Our framework maps fundamental biological factors (e.g., phylogenetic tree and infection) to general theories of directed percolation and self-organized criticality, identifying potential physics laws of RNA virus evolution directly from accessible empirical data. Compared with the evolution characterized from the perspective of population dynamics [9, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23], our framework may be more applicable to analyzing the macroscopic characteristics of evolution dynamics underlying RNA virus phylogenetic tree formation [24, 25, 26, 27].

Applying this theory, we discover that SARS-CoV-2 evolution may proceed at self-organized criticality. Threatening lineages (e.g., the Delta virus) can be coincidentally found on the critical phylogenetic tree. These findings suggest the potential of our theory in analyzing RNA virus evolution and related phenomena. Although the present framework has been validated on SARS-CoV-2, its generalizability is suggested to be further verified on other RNA viruses. Moreover, future studies can also investigate the relation between the self-organized critical evolution process and the threatening degree of RNA virus. Hopefully, this direction may help identify potential virus lineages to be concerned from a physically-explainable perspective.

Acknowledgements

Correspondence should be addressed to Y.T. and P.S. Authors Z.R.T. and H.D.H. contribute equally to this work. This project is supported by the Artificial and General Intelligence Research Program of Guo Qiang Research Institute at Tsinghua University (2020GQG1017) as well as the Tsinghua University Initiative Scientific Research Program.

Appendix A. Host cell resources shape RNA virus evolution

With reasonable errors, the mutation rate $\sigma$ of a single RNA virus is typically

$$
\sigma \in \left[ 10^{-5} \chi \phi \frac{1}{\tau}, 10^{-4} \chi \phi \frac{1}{\tau} \right],
$$

(A.1)

where $\chi \in \mathbb{N}^+$ measures the number of nucleotides, parameter $\tau$ denotes entrance and eclipse periods, and $\phi \simeq 10^3$ denotes burst size (the number of viruses produced by a host cell) [1, 8, 30]. These parameters are principally constant for a given RNA virus. Note that the original RNA virus vanishes after replication [29].
If host cell resources are sufficient, RNA viruses may have an opportunity to replicate independently and freely. When these happen, a supremum of the cumulative mutation rate of an RNA virus population of initial size $\omega \in \mathbb{N}^+$ during a duration of $[0, \kappa]$ (assume $\tau$ is divisible by $\kappa$) can be reached

$$
\sup_\omega \zeta (\omega) = \omega \int_0^\kappa \left[ \phi^\lfloor t/\tau \rfloor - \phi^{\lfloor t/\tau \rfloor - 1} \right] dt,
$$

(A.2)

$$
= \omega \sigma \tau \left( \phi^{\lfloor \kappa/\tau \rfloor - 1} - \frac{1}{\phi} \right),
$$

(A.3)

where notion $\lfloor \cdot \rfloor$ denotes the floor function.

However, these ideal conditions are not realistic since host cell resources are always limited. Therefore, biologically sound models, including the discussed previous works [13, 15, 17, 19, 20, 22, 23], of RNA virus population dynamics bound the proliferation of RNA virus by time-dependent host cell resource $\lambda (t)$ (e.g., the number of susceptible cells)

$$
\sup_\omega \hat{\zeta} (\omega) = \omega \int_0^\kappa \min \left[ \lambda (t), \phi^{\lfloor t/\tau \rfloor} - \phi^{\lfloor t/\tau \rfloor - 1} \right] dt.
$$

(A.4)

Although Eq. (A.4) can not be solved until $\lambda (t)$ is given, we can know

$$
\sup_\lambda \sup_\omega \zeta (\omega) = \sup_\omega \zeta (\omega).
$$

(A.5)

The difference $\sup_\omega \zeta (\omega) - \sup_\omega \hat{\zeta} (\omega)$ will be minimized when $\lambda (t) \geq \phi^{\lfloor t/\tau \rfloor} - \phi^{\lfloor t/\tau \rfloor - 1}$.

**Appendix B. Avalanches of RNA virus evolution**

Consider a time-continuous infection process, where a random patient at moment $t \in [t', t'']$ implies three possibilities: becoming absorbed with probability $\rho$, creating another patient with probability $\theta$, or remaining effect-free with probability $1 - (\rho + \theta)$. In critical states, we have $\rho = \theta$ [14]. We define $A_n (t)$ as the probability for $n$ patients to exist at $t^* + t$ given that 1 patient exists at $t^*$. Assuming the independence of patient emergence, we have

$$
A_n (t) = \sum_{n_1 + \ldots + n_k = n} A_{n_1} (t) \ldots A_{n_k} (t).
$$

(B.1)

Assume that $A_n (t), n \in \mathbb{N}^+$ admits a Maclaurin expansion $A_n (t) = a_n t + o (t^2)$ (when $n \neq 1$) or $A_n (t) = a_n t + 1 + o (t^2)$ (when $n = 1$) where $a_n = dA_n (0) / dt$, we can readily derive $a_0 = a_2 = \rho$ and $a_1 = -2 \rho$ [14]. Meanwhile, we can know

$$
A_n (t + dt) - A_n (t) = \sum_{k=0}^{\infty} a_k A_{n-k} (t) dt.
$$

(B.2)
Eqs (B.1-B.2) readily lead to
\[
\frac{\partial}{\partial t} G(x,t) = \sum_{k=0}^{\infty} a_k \sum_{n=0}^{\infty} \left( \sum_{n_1+\ldots+n_k=n-k} \prod_{i=1}^{k} \mathcal{A}_{n_i}(t) \right) x^n,  
\]
where \( G(x,t) = \sum_{n=0}^{\infty} \mathcal{A}_n(t) x^n, \ x \in [0,1] \) denotes the generating function. Applying a trick introduced in Ref. [44], we define
\[
\mathcal{H}(x) = \frac{\partial}{\partial t} G(x,0),
\]
which naturally leads to
\[
\frac{\partial}{\partial t} G(x,t) = \mathcal{H}(G(x,t)), \quad (B.6)
\]
the backward Chapman-Kolmogorov equation. Meanwhile, \( \mathcal{H}(x) = \rho (1 - x)^2 \) can be derived based on \( a_0, a_1, \) and \( a_2 \) [44]. Taken together, we have
\[
\frac{\partial}{\partial t} G(x,t) = \rho (1 - G(x,t))^2. \quad (B.7)
\]
Note that the initial condition is \( G(x,0) = x \) since one patient emerges at \( t^* \). Solving Eq. (B.7), we derive an analytic expression
\[
G(x,t) = \frac{\rho (1 - x) t}{\rho (1 - x) t + 1}. \quad (B.8)
\]
Therefore, we have \( \mathcal{A}_0(t) = G(0,t) = \frac{\rho t}{\rho t + 1} \), supporting a calculation of lifetime distribution \( \mathcal{P}_T(t) \)
\[
\lim_{t \to \infty} \mathcal{P}_T(t) = \lim_{t \to \infty} \frac{d}{dt} G(0,t) \sim t^{-2}. \quad (B.9)
\]
Following Refs. [44, 32, 45], one can similarly calculate
\[
\lim_{s \to \infty} \mathcal{P}_S(s) \sim s^{-3/2}. \quad (B.10)
\]
These exponents are consistent as the predictions of other mean field theories [43, 44].

Appendix C. Verify the criticality of RNA virus evolution

Compared with the scaling relation [46, 47, 37, 42], a more precise verification of criticality can be implemented by the collapse shape [47, 39]
\[
\langle \mathcal{V} (t | T) \rangle = \mathcal{F} \left( \frac{t}{T} \right), \quad (C.1)
\]
in which the expectation is averaged across different $T$. Notion $F(\cdot)$ denotes a universal scaling function. Notion $V(t | T)$ is defined as the average collapse shape of all avalanches with the same life time $T$

$$V(t | T) = T^{1-\gamma} \langle S(t | T) \rangle. \quad (C.2)$$

where $S(t | T)$ denotes the time-dependent avalanche size during these avalanches. At criticality, all data of $V(t | T)$ should collapse onto $F(\cdot)$, a parabolic function, with reasonable errors [47, 39]. In some cases, function $F(\cdot)$ may contain both the parabolic component and a slight global trend (e.g., increasing or decreasing), making the shape not perfectly parabolic. De-trending can be used to deal with this issue during data pre-processing.

Moreover, a more practical verification concerns the slow and exponential decay of auto-correlation [39]

$$\ln \frac{\text{Cov}(S(t_i | T), S(t_j | T))}{\text{Cov}(S(t_i | T), S(t_i | T))} = -\xi \left( \frac{t_j - t_i}{T} \right) + c, \quad (C.3)$$

where $t_i \in [t', t'']$ and $t_j \in [t_i, t'']$.

**Appendix D. Power-law analysis**

The probability distributions of life time $T$ and avalanche size $S$ (for midpoint and bounds) are derived after a fine-grained data binning process with 1000 bins is applied, which is useful in de-noising while controlling information loss.

We discover that $P_T(\cdot)$ does not follow power-law properties for all values of $T$ (see Fig. 1b). Therefore, distribution cutoff estimation is necessary for $T$ and $S$. We apply the method introduced in Ref. [51] to estimate $T'$, $S'$, $S'_-$ (lower bound), and $S'_+$ (upper bound), which calculates the Kolmogorov-Smirnov (KS) statistic $\eta$ between the observed cumulative probability distribution above cutoff and the cumulative probability distribution of a standard power-law variable. An ideal cutoff is expected to minimize $\eta$ [51]. To control the reduction of sample size, we relatively relax the restriction and estimate cutoffs when $\langle \eta \rangle - 0.5 \text{ std}(\eta)$ is reached at the first time, where $\text{ std}(\cdot)$ denotes the standard deviation. Distribution cutoffs are estimated as $T' = 29$, $S' = 13320$, $S'_- = 12760$, and $S'_+ = 13980$. The distribution tails above cutoffs cover 93.879%, 76.38%, 70.526%, and 91.683% of original samples, supporting robust estimations of power-law exponents.

A maximum likelihood estimation of power-law exponent [51] is implemented on the samples above cutoffs. An ideal exponent can maximize the likelihood $L$ [51]. After estimating power-law exponents, we carry out a KS-statistic-based analysis to derive the goodness of estimation. To calculate the average KS statistic $\eta^*$ between sample distributions and estimated models, we first generate 1000 sample distributions (each sample distribution contains $n \in [500,5000]$ samples) of estimated power-law models. Then, we define $\nu = \frac{\hat{\eta} - \eta^*}{\eta^*}$ to reflect the goodness of estimation, where $\hat{\eta}$ is the KS
statistic between the cumulative probability distributions of estimated power-law models and empirical distributions above cutoffs. We suggest that $\nu < 10\%$ can be a reasonable standard for ideal estimations.

References

[1] Domingo E, Escarmís C, Sevilla N, Moya A, Elena S F, Quer J, Novella I S and Holland J J 1996 *The FASEB Journal* **10** 859–864
[2] Dolan P T, Whitfield Z J and Andino R 2018 *Annual Review of Virology* **5** 69–92
[3] Duffy S 2018 *PLoS biology* **16** e3000003
[4] Ojosnegros S and Beer twinkeln N 2010 *Immunome research* **6** 1–14
[5] Irwin K K, Renzette N, Kowalik T F and Jensen J D 2016 *Virus evolution* **2**
[6] Jakobsson E, Farimani A B, Taikhorshid E and Aluru N 2020 *Biophysical Journal* **118** 482a
[7] Sneppen K and Zocchi G 2005 *Physics in molecular biology* (Cambridge University Press)
[8] Tsimring L S, Levine H and Kessler D A 1996 *Physical review letters* **76** 4440
[9] Fontana W, Schnabl W and Schuster P 1989 *Physical Review A* **40** 3301
[10] Sasaki A 1994 *Journal of Theoretical Biology* **168** 291–308
[11] Haraguchi Y and Sasaki A 1997 *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **352** 11–20
[12] Sasaki A and Haraguchi Y 2000 *Journal of molecular evolution* **51** 245–255
[13] Nowak M and May R M 2000 *Virus dynamics: mathematical principles of immunology and virology: mathematical principles of immunology and virology* (Oxford University Press, UK)
[14] Kauffman S A et al. 1993 *The origins of order: Self-organization and selection in evolution* (Oxford University Press, USA)
[15] Korobeinikov A 2004 *Bulletin of Mathematical Biology* **66** 879–883
[16] Regoes R R, Wodarz D and Nowak M A 1998 *Journal of theoretical biology* **191** 451–462
[17] Adiwijaya B S, Herrmann E, Hare B, Kieffer T, Lin C, Kwong A D, Garg V, Randle J C, Sarrazin C, Zeuzem S et al. 2010 *PLoS computational biology* **6** e1000745
[18] Garcia V, Glassberg E C, Harpak A and Feldman M W 2018 *Journal of The Royal Society Interface* **15** 20170921
[19] Yan L, Neher R A and Shraiman B I 2019 *Elife* **8** e44205
[20] Held T, Klemmer D and Lässig M 2019 *Nature communications* **10** 1–11
[21] Rasmussen D A and Stadler T 2019 *Elife* **8** e45562
[22] Marchi J, Lässig M, Wälczak A M and Mora T 2021 *Proceedings of the National Academy of Sciences* **118** ISSN 0027-8424 URL https://www.pnas.org/content/118/27/e2103398118
[23] Korobeinikov A, Archibasov A and Sobolev V 2016 Multi-scale problem in the model of m1a virus evolution *Journal of Physics: Conference Series* vol 727 (IOP Publishing) p 012007
[24] Lässig M and Luksza M 2014 *Elife* **3** e05060
[25] Neher R A, Russell C A and Shraiman B I 2014 *Elife* **3** e03568
[26] Desai M M and Fisher D S 2007 *Genetics* **176** 1759–1798
[27] Rouzine I M, Wakeley J and Coffin J M 2003 *Proceedings of the National Academy of Sciences* **100** 587–592
[28] Velavan T P and Meyer C G 2020 *Tropical medicine & international health* **25** 278
[29] Vokovski P, Kratzel A, Steiner S, Stalder H and Thiel V 2021 *Nature Reviews Microbiology* **19** 155–170
[30] Bar-On Y M, Flamholz A, Phillips R and Milo R 2020 *elife* **9** e57309
[31] Network P 2021 Rules for the designation and naming of pango lineages URL https://www.pango.network/the-pango-nomenclature-system/statement-of-nomenclature-rules/
[32] Harris T E et al. 1963 *The theory of branching processes* vol 6 (Springer Berlin)
[33] Garcia-Millan R, Pausch J, Walter B and Pruessner G 2018 *Physical Review E* **98** 062107
