

# The mechanism shaping the logistic growth of mutation proportion in epidemics at population scale

Shi Zhao <sup>a, b</sup>, Inchi Hu <sup>c</sup>, Jingzhi Lou <sup>a</sup>, Marc K.C. Chong <sup>a, b</sup>, Lirong Cao <sup>a, b</sup>, Daihai He <sup>d</sup>, Benny C.Y. Zee <sup>a, b</sup>, Maggie H. Wang <sup>a, b, \*</sup>

<sup>a</sup> JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong, China

<sup>b</sup> CUHK Shenzhen Research Institute, Shenzhen, China

<sup>c</sup> Department of Information Systems, Business Statistics and Operations Management, Hong Kong University of Science and Technology, Hong Kong, China

<sup>d</sup> Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China

## ARTICLE INFO

### Article history:

Received 10 November 2022

Received in revised form 19 December 2022

Accepted 25 December 2022

Available online 26 December 2022

Handling Editor: Dr Lou Yijun

### Keywords:

Transmission advantage

Logistic growth

Population dynamics

Selective pressure

COVID-19

## ABSTRACT

Virus evolution is a common process of pathogen adaptation to host population and environment. Frequently, a small but important fraction of virus mutations are reported to contribute to higher risks of host infection, which is one of the major determinants of infectious diseases outbreaks at population scale. The key mutations contributing to transmission advantage of a genetic variant often grow and reach fixation rapidly. Based on classic epidemiology theories of disease transmission, we proposed a mechanistic explanation of the process that between-host transmission advantage may shape the observed logistic curve of the mutation proportion in population. The logistic growth of mutation is further generalized by incorporating time-varying selective pressure to account for impacts of external factors on pathogen adaptiveness. The proposed model is implemented in real-world data of COVID-19 to capture the emerging trends and changing dynamics of the B.1.1.7 strains of SARS-CoV-2 in England. The model characterizes and establishes the underlying theoretical mechanism that shapes the logistic growth of mutation in population.

© 2022 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Infectious diseases have been a long-term threat to human population globally (Kraemer et al., 2019; Troeger et al., 2018). The recurring epidemics of infectious pathogens are often driven by virus evolution to avoid extinction and to improve fitness. A small but important fraction of mutations that improve pathogen advantage are often observed to grow rapidly and reach fixation in population in a short period of time (Lowder et al., 2009; Steel, Lowen, Mubareka, & Palese, 2009; Zhu & Shu, 2015). For example, for influenza viruses, key mutations appeared in the initial stage of epidemic cycles and reached fixation as the epidemic progressed (Gog, Rimmelzwaan, Osterhaus, & Grenfell, 2003; Shih, Hsiao, Ho, & Li, 2007; Wang et al., 2021). In the

\* Corresponding author. JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong, China.

E-mail addresses: [zhaoshi.cmsa@gmail.com](mailto:zhaoshi.cmsa@gmail.com) (S. Zhao), [imichu@ust.hk](mailto:imichu@ust.hk) (I. Hu), [jzl@bethbio.com](mailto:jzl@bethbio.com) (J. Lou), [marc@cuhk.edu.hk](mailto:marc@cuhk.edu.hk) (M.K.C. Chong), [caolr@link.cuhk.edu.hk](mailto:caolr@link.cuhk.edu.hk) (L. Cao), [daihai.he@polyu.edu.hk](mailto:daihai.he@polyu.edu.hk) (D. He), [bzee@cuhk.edu.hk](mailto:bzee@cuhk.edu.hk) (B.C.Y. Zee), [maggiew@cuhk.edu.hk](mailto:maggiew@cuhk.edu.hk) (M.H. Wang).

Peer review under responsibility of KeAi Communications Co., Ltd.

recent COVID-19 pandemic, multiple new genetic variants of SARS-CoV-2 appeared in different locations and time, which is one of the major challenges against the control of COVID-19 (Baum et al., 2020; Tsetsarkin, Vanlandingham, McGee, & Higgs, 2007). These included the Alpha variant (B.1.1.7) that first arise in the United Kingdom (UK) around September 2020 (Tang, Tambyah, & Hui, 2021), the Gamma variant (P.1) in Brazil (Wise, 2021), and the Delta variant (B.1.617.2) in India in 2021 (Campbell et al., 2021). Characterizing and modeling the new genetic variants both biologically and epidemiologically, are crucial for advancing the control of infectious diseases.

At population (i.e., between-host) scale (Metcalfe et al., 2015), epidemiological studies assessed the relative fitness of a viral strain versus another using the ratio of reproduction numbers (Leung, Shum, Leung, Lam, & Wu, 2021; Zhao et al., 2021a, 2021b, 2021c), which was also used to infer the between-host transmission advantage of a viral strain relative to its competing strain. Transmission advantages are often reported among virus strains, such as the SARS-CoV-2 (Campbell et al., 2021; Davies et al., 2021a; Graham et al., 2021a, 2021b; Volz et al., 2021a; Zhao et al., 2021a, 2021b, 2021c) and influenza virus (Gog et al., 2003; Leung et al., 2021), which are usually accompanied with increase in both prevalence of mutations and incidences of infection (Davies et al., 2021a; Graham et al., 2021a, 2021b; Ito, Piantham, & Nishiura, 2021; Leung, Lipsitch, Yuen, & Wu, 2017). However, the increase of mutation prevalence was mainly reported descriptively, and the population dynamics underlying the growth pattern of mutation prevalence remain largely uncharacterized.

In this study, we derive that the growth of new mutants follows a logistic curve, and dissect the population dynamics shaping the pattern based on classical epidemiology theories of disease transmission. The relationship between the growth rate of mutations and transmission advantage is analyzed mathematically. We further discuss the effect of selective pressure on transmission advantage and the mechanism of dynamics. The theoretical framework is demonstrated using simulation data and applied on real world data of the SARS-CoV-2 Alpha (i.e., B.1.1.7) variants in England.

## 2. Conceptualization and parameterization: from epidemic model to logistic growth

The relationship of reproduction number and transmission advantage is first conceptualized and parameterized in the context of disease transmission. We formulate several widely-adopted epidemic models, and explore how transmission advantage shapes the growing dynamics of mutated strains. We characterize the changing patterns of mutated strains as logistic growth, see Eqn (3), and derive the relationship between transmission advantage and exponential growth rate mathematically.

### 2.1. Basic reproduction number and intrinsic transmission advantage

The basic reproduction number is the expected number of cases directly generated by one typical case during the infectious period in a completely susceptible population (Van den Driessche, 2017; Zhao et al., 2020). Epidemiologically, an outbreak may occur as the number of cases increases when basic reproduction number is larger than 1. As a well-studied metric that considers both reproducibility and survivability of the seed case, the reproduction number is typically adopted to measure the intrinsic fitness of a pathogen in maintaining its transmission in host population (Diekmann, Heesterbeek, & Roberts, 2010; Metz, Nisbet, & Geritz, 1992; Schreiber et al., 2021).

The multiplicative transmission advantage ( $\eta$ ) of one viral strain against another (e.g., its recent ancestor) is typically described by the relative fitness, calculated as the ratio of the two strains' reproduction numbers (Lythgoe, Gardner, Pybus, & Grove, 2017). Considering the context including two types of strains, the first type refers to the original strain, and the second type refers to the mutated strain. We denote the basic reproduction numbers of the original and mutated strains as  $\mathcal{R}_1$  and  $\mathcal{R}_2$ , respectively. The transmission advantage of the second type against the first type of strains is defined as

$$\eta = \frac{\mathcal{R}_2}{\mathcal{R}_1}, \quad (1)$$

which was adopted in studying the transmission dynamics of gonorrhoeae (Whittles, White, & Didelot, 2017), influenza (Leung et al., 2017), HIV (Kühnert et al., 2018), and COVID-19 (Campbell et al., 2021; Faria et al., 2021; Volz et al., 2021a; Zhao et al., 2021a, 2021b, 2021c).

Since the basic reproduction number is considered as an innate feature,  $\mathcal{R}_1$  and  $\mathcal{R}_2$  measure the intrinsic fitness where the biological viability of pathogens is the only contributing factor (Metz et al., 1992). Hence, the term  $\eta$  implies the intrinsic transmission advantage under a natural selection-free context (Coombs, Gilchrist, & Ball, 2007). When  $\eta > 1$ , the new strain is more transmissible than the original strain (Zhao et al., 2021a, 2021b, 2021c), and *vice versa*.

### 2.2. Simple power series

A closed population with size  $N$  is compartmentalized into four classes including the susceptible ( $S$ ), the infections of original strains ( $I_1$ ) and mutated strains ( $I_2$ ), and the removed ( $R$ ) individuals. We have  $S + I_1 + I_2 + R = N$ . To mimic the transmission process of disease, simple power series are modelled as a growing process that each primary case of original (or mutated) strains will generate  $\mathcal{R}_1$  (or  $\mathcal{R}_2$ ) secondary cases after a fixed generation time (GT)  $T_g$ . The GT is defined as the interval between the times of infection of the primary case and secondary case in a transmission chain (Fine, 2003). Here, we

fix the generation time as a constant  $T_g$  for simplicity. These quantities appear more clearly in the mathematical forms as follows.

For the  $(j + 1)$ -th generation, the number of cases infected by original strains is  $I_1(j + 1) = \mathcal{R}_1 I_1(j) \frac{S(j)}{N}$ , and the number of cases infected by new strains is  $I_2(j + 1) = \mathcal{R}_2 I_2(j) \frac{S(j)}{N}$ . The term  $\frac{S(j)}{N}$  accounts for the probability of contracting an unprotected individual in the population under the homogeneously mixing mass-contacting assumption (Brauer, Driessche, & Wu, 2008), which mimics the depletion of susceptible individuals during an epidemic. Then, the terms  $\mathcal{R}_1 \frac{S(j)}{N}$  and  $\mathcal{R}_2 \frac{S(j)}{N}$  are the (effective) reproduction number for original and mutated strains, respectively. For the susceptible individuals,  $S(j + 1) = S(j) - [I_1(j + 1) + I_2(j + 1)]$ , and thus

$$S(j) = S(0) - \sum_{k=1}^j [I_1(k) + I_2(k)] = S(0) - \sum_{k=1}^j I(k),$$

where  $I(j) = I_1(j) + I_2(j)$ , for  $j > 0$ . For convenience, we denote the cumulative number of cases by

$$C(j) = I(0) + \sum_{k=1}^j I(k),$$

for the  $j$ -th generation. As such,  $S(j) = S(0) - [C(j) - I(0)]$ .

Extending the  $I_1(j)$  series iteratively, we obtain

$$I_1(j) = \mathcal{R}_1^j I_1(0) \prod_{k=0}^{j-1} \left[ \frac{S(k)}{N} \right] = I_1(0) \mathcal{R}_1^j \frac{\prod_{k=0}^{j-1} [S(0) + I(0) - C(k)]}{N^j},$$

and similarly,

$$I_2(j) = I_2(0) \mathcal{R}_2^j \frac{\prod_{k=0}^{j-1} [S(0) + I(0) - C(k)]}{N^j}$$

Then, for the  $j$ -th generation, the proportion of cases infected by new strains,  $\rho_2$ , is derived as  $\rho_2(j) = \frac{I_2(j)}{I_1(j) + I_2(j)} = \frac{I_2(0) \mathcal{R}_2^j}{I_1(0) \mathcal{R}_1^j + I_2(0) \mathcal{R}_2^j}$ . According to Eqn (1), we have

$$\rho_2(j) = \frac{I_2(0) \eta^j}{I_1(0) + I_2(0) \eta^j} = \frac{1}{1 + \frac{I_1(0)}{I_2(0)} \eta^{-j}}.$$

Since the GT is fixed at  $T_g$ , we transform the time scale from ‘per generation’ to ‘per day’, where 1 day is equivalent to  $1/T_g$  generation. We attain

$$\rho_2(t) = \frac{1}{1 + \frac{I_1(0)}{I_2(0)} \eta^{-t/T_g}},$$

on the  $t$ -th day after the initial stage when the seed cases are placed at  $t = 0$ . Let  $\ln(\eta)/T_g = r$ , we have

$$\eta = \exp(rT_g), \tag{2}$$

and we obtain  $\rho_2(t)$  as

$$\rho_2(t) = \frac{1}{1 + \left[ \frac{1 - \rho_2(0)}{\rho_2(0)} \right] \exp(-rt)}, \tag{3}$$

which is a logistic growth function depending on the calendar time  $t$ . The formulation in Eqns (2) and (3) are adopted previously (Campbell et al., 2021; Davies et al., 2021a; Metz et al., 1992; Volz et al., 2021b).

On the one hand, the term  $r$  is the exponential growth rate, or Malthusian parameter known to the demographics from the ‘law of geometric increase’ (Malthus, 1872), of the proportion of mutated strains in the population, which can be calculated empirically. On the other hand, the term  $r$  represents the difference between the variant-specific intrinsic growth rates, i.e., Malthusian parameters, attributed to the transmission advantage. Using the approximation from Taylor series, Eqn (2) becomes  $\eta - 1 = rT_g$ , where the value of  $\eta - 1$  is also known as the selection coefficient (Chevin, 2011; Schreiber et al., 2021; Travisano & Lenski, 1996; Volz et al., 2021b). Epidemiologically, this means the product of  $rT_g$  approaches the attributed risk of transmission of new strains, i.e.,  $(\eta - 1)$ , when the value of  $r$  is relatively small.

Importantly, from Eqns (2) and (3), we note that the logistic growth dynamics are not (directly) determined by the reproduction number of either type of strain (i.e.,  $\mathcal{R}_1$  or  $\mathcal{R}_2$ ), but determined by the transmission advantage ( $\eta$ ) between them. In other words, even if  $\mathcal{R}_1$  and  $\mathcal{R}_2$  change, the logistic growth in Eqn (3) holds as long as the ratio of  $\mathcal{R}_1$  and  $\mathcal{R}_2$

remains a constant. Additionally, the link between Eqns (1) and (2) implies the value of  $\eta$  can be determined by the knowledge of  $r$  and  $T_g$ .

### 2.3. Susceptible-infectious-removed compartmental model

We formulate the classic two-strain susceptible-infectious-removed (SIR) model as an ordinary differential equation (ODE) system expressed as

$$\begin{cases} \frac{dS}{dt} = -\beta \frac{S}{N} (I_1 + \eta_\beta I_2), \\ \frac{dI_1}{dt} = \beta \frac{S}{N} I_1 - \gamma I_1, \\ \frac{dI_2}{dt} = \eta_\beta \beta \frac{S}{N} I_2 - \frac{\gamma}{\eta_\gamma} I_2, \\ \frac{dR}{dt} = \gamma \left( I_1 + \frac{I_2}{\eta_\gamma} \right). \end{cases} \tag{4}$$

Here, the  $\beta$  is the transmission rate, or effective contact rate, and the  $\gamma$  is the removal rate, the reciprocal of which ( $\gamma^{-1}$ ) is the mean infectious period. The factor  $\eta_\beta$  indicates the multiplicative change in  $\beta$ , and the factor  $\eta_\gamma$  indicates the multiplicative change in  $\gamma^{-1}$  attributed to mutation. Similar formulations are adopted previously (Davies et al., 2021a; Gog et al., 2003). For the two types of strains, they are indicated by subscript ‘1’ for the original strain, and ‘2’ for the newly emerged (mutated) strain. Since  $\frac{dN}{dt} = 0$ , the total population ( $N$ ) is a constant.

Using the next generation matrix approach (van den Driessche & Watmough, 2002), the reproduction numbers of the cases infected by original and new strains are

$$\mathcal{R}_1 \frac{S}{N} = \frac{\beta}{\gamma} \bullet \frac{S}{N}, \text{ and } \mathcal{R}_2 \frac{S}{N} = \eta_\beta \eta_\gamma \frac{\beta}{\gamma} \bullet \frac{S}{N},$$

respectively, which are in line with (Anderson & May, 1985a, 1985b; Getz & Pickering, 1983). Hence, following the definition in Eqn (1), we have

$$\eta = \eta_\beta \eta_\gamma, \tag{5}$$

which is considered as a special case of Eqn (1) for compartmental model formulated in Eqn (4).

The incidence rate can be expressed as

$$c_1(t) = \beta \frac{S(t)}{N} I_1(t), \text{ and } c_2(t) = \eta_\beta \beta \frac{S(t)}{N} I_2(t),$$

respectively. For the proportion of cases infected by new strains,

$$\rho_2(t) = \frac{c_2(t)}{c_1(t) + c_2(t)} = \frac{1}{1 + \frac{I_1(t)}{I_2(t)} \eta_\beta^{-1}}.$$

By linearizing Eqn (4) at time  $\tau$ , we have

$$\frac{dI_1}{dt} = \left[ \beta \frac{S(\tau)}{N} - \gamma \right] I_1, \text{ and } \frac{dI_2}{dt} = \left[ \eta_\beta \beta \frac{S(\tau)}{N} - \frac{\gamma}{\eta_\gamma} \right] I_2$$

Then, we have the linear approximation at time  $\tau$  as

$$I_1(t) \approx I_1(\tau) \exp \left[ \left[ \beta \frac{S(\tau)}{N} - \gamma \right] (t - \tau) \right], \text{ and } I_2(t) \approx I_2(\tau) \exp \left[ \left[ \eta_\beta \beta \frac{S(\tau)}{N} - \frac{\gamma}{\eta_\gamma} \right] (t - \tau) \right]$$

Here, we remark that the linear approximation of  $\rho_2(t)$  applies only if the number of susceptible individuals ( $S$ ) is relatively stable.

Next, we express the term  $\frac{I_1(t)}{I_2(t)}$  in the formulation of  $\rho_2(t)$  as

$$\frac{I_1(t)}{I_2(t)} = \frac{I_1(\tau) \exp\left[\left[\beta \frac{S(\tau)}{N} - \gamma\right](t - \tau)\right]}{I_2(\tau) \exp\left[\left[\eta_\beta \beta \frac{S(\tau)}{N} - \frac{\gamma}{\eta_\gamma}\right](t - \tau)\right]} = \eta_\beta \left[\frac{1 - \rho_2(\tau)}{\rho_2(\tau)}\right] \exp\left[-\left[(\eta_\beta - 1)\beta \frac{S(\tau)}{N} + \left(\frac{\eta_\gamma - 1}{\eta_\gamma}\right)\gamma\right](t - \tau)\right].$$

Substituting into  $\rho_2(t)$ , we have

$$\rho_2(t) = \frac{1}{1 + \left[\frac{1 - \rho_2(\tau)}{\rho_2(\tau)}\right] \exp[-r(t - \tau)]}, \tag{6}$$

where the relationship between transmission advantage ( $\eta_\beta$  and  $\eta_\gamma$ ) and exponential growth rate ( $r$ ) is given by

$$r = (\eta_\beta - 1)\beta \frac{S(\tau)}{N} + \left(\frac{\eta_\gamma - 1}{\eta_\gamma}\right)\gamma \tag{7}$$

Eqn (6) is also in the logistic form depending on the calendar time  $t$ . Specially, when we mark  $\tau = 0$ , the terms  $\rho_2(0)$  and  $S(0)$  represent the initial conditions of the compartmental model in Eqn (4), both of which can be learnt from the real-world settings. As such, Eqn (6) can be simplified as  $\rho_2(t) = \frac{1}{1 + \left[\frac{1 - \rho_2(0)}{\rho_2(0)}\right] \exp[-rt]}$  with  $\tau = 0$ , which is the same as Eqn (3).

Furthermore, we discuss the following four scenarios to illustrate the practical meaning of Eqn (7).

2.3.1. When the two types of variants have the same mean infectious periods

When the mean infectious periods ( $\gamma^{-1}$ ) of the two types of variants are the same, we have  $\eta_\gamma = 1$ , which indicates the multiplicative change in infectious period vanishes. Based on Eqn (5), the transmission advantage  $\eta = \eta_\beta$ , and thus Eqn (7) is reduced to  $r = (\eta - 1)\beta \frac{S(\tau)}{N}$ .

We remark that  $\gamma^{-1}$  is equivalent to the mean GT, denoted by  $T_g$ , for the classic SIR model (Svensson, 2007), and thus the term  $\eta$  can be expressed as

$$\eta = \frac{rT_g}{\mathcal{R}_1 \frac{S(\tau)}{N}} + 1 \tag{8}$$

Specially, if  $r$  is sufficiently small and further assume that the effective reproduction number of the original strains approaches 1, i.e.,  $\mathcal{R}_1 \frac{S(\tau)}{N} \rightarrow 1$ , which implies the original strains are circulating endemically, we have

$$\lim_{r \rightarrow 0, \mathcal{R}_1 \frac{S(\tau)}{N} \rightarrow 1} \eta = \exp(rT_g), \tag{9}$$

and the right-hand side of Eqn (9) is the same as the right-hand side of Eqn (2).

2.3.2. When the two types of variants have the same transmission rates

When the transmission rates ( $\beta$ ) of the two types of variants are the same, we have  $\eta_\beta = 1$ , which indicates that the multiplicative change in transmission rate vanishes. Then, based on Eqn (5), the transmission advantage  $\eta = \eta_\gamma$ , and thus Eqn (7) becomes  $r = \left(\frac{\eta - 1}{\eta}\right)\gamma$ .

Similarly, by replacing  $\gamma^{-1}$  with the mean GT,  $T_g$ , the term  $\eta$  can be expressed as

$$\eta = \frac{1}{1 - rT_g}$$

In addition, if  $r$  becomes sufficiently small, we have

$$\lim_{r \rightarrow 0} \eta = \frac{1}{\exp(-rT_g)} = \exp(rT_g),$$

the right-hand side of which is the same as the right-hand side of Eqn (2).

2.3.3. When both the two transmission rates and two mean infectious periods are close

We consider a special scenario that both transmission rates are close, and both mean infectious periods are also close, which indicates both  $\eta_\beta$  and  $\eta_\gamma$  approach 1 simultaneously. Then, from Eqn (7), we have

$$\lim_{\eta_\beta \rightarrow 1, \eta_\gamma \rightarrow 1} r = (\ln \eta_\beta) \beta \frac{S(\tau)}{N} + (\ln \eta_\gamma) \gamma$$

Similarly, by replacing  $\gamma^{-1}$  with the mean GT,  $T_g$ , we obtain

$$\lim_{\eta_\beta \rightarrow 1, \eta_\gamma \rightarrow 1} r = \frac{(\ln \eta_\beta) \mathcal{R}_1 \frac{S(\tau)}{N} + (\ln \eta_\gamma)}{T_g}$$

Specially, if we let the effective reproduction number of the original strains approach 1, i.e.,  $\mathcal{R}_1 \frac{S(\tau)}{N} \rightarrow 1$ , we have  $r = \frac{\ln \eta}{T_g}$  or equivalently,

$$\lim_{\eta_\beta \rightarrow 1, \eta_\gamma \rightarrow 1, \mathcal{R}_1 \frac{S(\tau)}{N} \rightarrow 1} \eta = \exp(rT_g),$$

the right-hand side of which is the same as the right-hand side of Eqn (2).

#### 2.4. Susceptible-exposed-infectious-removed compartmental model

Besides the classic SIR model, we also show that the logistic growth curve can be derived from the classic two-strain susceptible-exposed-infectious-removed (SEIR) model with a more generalized formulation of growth rate parameter. The technical frameworks are detailed in supplementary material S1.

#### 2.5. Renewable process

We consider the same context of susceptible ( $S$ ), infections of original ( $I_1$ ) and mutated ( $I_2$ ) strains, and removed ( $R$ ), but with a more realistic situation that the generation time (GT) follows a probability distribution. Thus, the transmission dynamics become a branching process, which was studied and adopted previously (Athreya, Ney, & Ney, 2004; Cori, Ferguson, Fraser, & Cauchemez, 2013; Volz et al., 2021a; Wallinga & Teunis, 2004; Zhao et al., 2021a, 2021b, 2021c).

The number of cases for original and new strains on day  $t$  can be expressed as

$$I_1(t) = I_1(0) \exp \left[ \int_0^t r_1(\tau) d\tau \right], \text{ and } I_2(t) = I_2(0) \exp \left[ \int_0^t r_2(\tau) d\tau \right].$$

Here, the terms  $r_1(t)$  and  $r_2(t)$  are the time-varying exponential growth rates, or Malthusian parameters, of the epidemic curves of original and new strains, respectively. The time-varying feature of the growth rate accounts for the depletion of susceptible population during an epidemic. Similar formulations are adopted in (Faria et al., 2021; Leung et al., 2017, 2021; Volz et al., 2021a). The proportion of cases infected by new strains equals

$$\rho_2(t) = \frac{I_2(t)}{I_1(t) + I_2(t)} = \frac{1}{1 + \left[ \frac{1 - \rho_2(0)}{\rho_2(0)} \right] \exp \left[ - \int_0^t [r_2(\tau) - r_1(\tau)] d\tau \right]}.$$

Specially, if we consider the number of susceptible individuals ( $S$ ) as a constant for a relatively short period of time, then both  $r_1$  and  $r_2$  are constants too. Let  $r = r_2 - r_1$ , which indicates the difference between the variant-specific growth rates attributed to transmission advantage, and we have

$$\int_0^t [r_2(\tau) - r_1(\tau)] d\tau = (r_2 - r_1)t = rt.$$

Thus,

$$\rho_2(t) = \frac{1}{1 + \left[ \frac{1 - \rho_2(0)}{\rho_2(0)} \right] \exp(-rt)},$$

which is the same as Eqn (3). We remark that this linear approximation of  $\rho_2(t)$  applies only if  $S$  holds relatively stable, which is generally acceptable because the proportion of susceptible population,  $\frac{S}{N}$ , is unlikely to have a large change during a relatively short period.

To explore the relation between growth rate ( $r$ ), GT distribution, and transmission advantage ( $\eta$ ), we consider stochastic GT with a probability density function denoted by  $w_1(\bullet)$  for the original strains, and  $w_2(\bullet)$  for the new strains. From the classic Euler-Lotka equation (Lotka, 1907), the equations

$$\mathcal{R}_1 \frac{S}{N} = \frac{1}{\int_0^\infty \exp[-r_1 \tau] w_1(\tau) d\tau}, \text{ and } \mathcal{R}_2 \frac{S}{N} = \frac{1}{\int_0^\infty \exp[-r_2 \tau] w_2(\tau) d\tau}$$

can be derived (Wallinga & Lipsitch, 2007), which describes the relationship among growth rate, reproduction number and GT. Note that  $\int_0^\infty \exp[-r_1 \tau] w(\tau) d\tau$  is the Laplace transform of  $w_1(\bullet)$  and the moment generating function (MGF)  $M(\bullet|w_1)$  well-known to statisticians. Following the definition in Eqn (1), we obtain a generalized version of  $\eta$ ,

$$\eta = \frac{M(-r_1|w_1)}{M(-r_2|w_2)}. \tag{10}$$

Since both  $r_1$  and  $r_2$  is constants, the value of  $\eta$  is also fixed. Therefore, each pair of  $r_1$  and  $\eta$  determines a unique  $(r_2 - r_1) = r$  as well as the changing profile  $\rho_2(t)$ .

Specially, the power series in Section 2.2 is equivalent to setting  $w_1(\bullet)$  and  $w_2(\bullet)$  to be the Dirac delta distribution function located at  $T_g$ .

Considering the practical implementation, we discuss the following two scenarios, which simplify Eqn (10).

2.5.1. When the reproduction number of original variants equals unity

If we let the effective reproduction number of the original strains be unity, i.e.,  $\mathcal{R}_1 \frac{S}{N} = 1$ , which implies the original strains are circulating endemically, we have  $r_1 = 0$ , and thus Eqn (4) becomes

$$\lim_{\mathcal{R}_1 \frac{S}{N} \rightarrow 1} \eta = \frac{1}{M(-r|w_2)},$$

where  $r = r_2$  when  $r_1 = 0$ . The density  $w_2(\bullet)$  can be inferred from the real-world contact tracing data as in (Ferretti et al., 2020; Ganyani et al., 2020; Ren et al., 2021; Tindale et al., 2020; Zhao et al., 2021a, 2021b, 2021c).

2.5.2. When the two types of variants have the same Gamma-distributed generation time

Since Gamma distribution is commonly adopted to model GT for infectious diseases (Ferretti et al., 2020; Ganyani et al., 2020; Tindale et al., 2020; Zhao et al., 2021a, 2021b, 2021c). We assume the GT distributions of the two types of variants follow the same Gamma distribution such that  $w_1(\bullet) = w_2(\bullet) = w(\bullet)$ . We denote the mean of  $w(\bullet)$  by  $T_g$ , and coefficient of variation (CV)  $1/\sqrt{k}$ , where  $k$  is also the shape parameter of the Gamma distribution. As such, the MGF of  $w(\bullet)$  is  $M(-x|w) = [1 + (\frac{T_g}{k})x]^{-k}$ . According to Eqn (10),  $\eta$  is expressed as

$$\eta = \frac{\left[1 + \left(\frac{T_g}{k}\right)r_1\right]^{-k}}{\left[1 + \left(\frac{T_g}{k}\right)r_2\right]^{-k}}.$$

Furthermore, when the growth rates  $r_1$  and  $r_2$  are small, we have

$$\lim_{r_1 \rightarrow 0, r_2 \rightarrow 0} \eta = \frac{\left[1 + \left(\frac{T_g}{k}\right)(r_2 - r_1)\right]^k}{\left[1 + \left(\frac{T_g}{k}\right)r\right]^{-k}}.$$

This simplifies the general relationship in Eqn (10). Note that the denominator of the right-hand-side of this equation is the same as  $M(-r|w)$  for Gamma-distributed  $w(\bullet)$ .

3. Time-varying selective pressure and structural change in logistic growth

From the formulation of the logistic growth of mutated strains in Section 2, we learn that the transmission advantage ( $\eta$ ), not the reproduction numbers (i.e.,  $\mathcal{R}_1$  or  $\mathcal{R}_2$ ), directly determines the growth rate. It is of interests to explore how the factors that affect the demonstration of transmission advantage may contribute to the growing pattern of mutations. In Section 3, we extend the framework by considering the impact of selective pressure (measured by  $\vartheta$ ) on the effective transmission advantage ( $\eta_{\text{eff}}$ ), see Eqns 11 and 12, as well as the growing patterns of the mutated strain's proportion ( $\rho_2$ ).

### 3.1. Effective transmission advantage and selective pressure

To complement the intrinsic concepts in Section 2.1, we consider a more realistic situation that selective pressures may affect the disease transmission by altering the fitness profiles, such that the mutated strain can only demonstrate a fraction of its transmission advantage. The effective transmission advantage, denoted by  $\eta_{\text{eff}}$ , is the realized relative fitness after accounting for the effects of selective pressures, which is an extension of the concept of intrinsic transmission advantage. Thus, similar to Eqn (1), the  $\eta_{\text{eff}}$  is defined as the ratio between the controlled reproduction numbers of the mutated ( $\rho_2$ ) and original ( $\rho_1$ ) strains

$$\eta_{\text{eff}} = \frac{\rho_2}{\rho_1}. \quad (11)$$

The controlled reproduction number represents the reproduction number under disease control measures. Since the control measures vary during the course of epidemic, both  $\rho_1$  and  $\rho_2$ , as well as  $\eta_{\text{eff}}$ , are time-varying. Like the intrinsic transmission advantage ( $\eta$ ), if  $\eta_{\text{eff}} > 1$ , the new strains are more transmissible than the original strains, and *vice versa*.

Under the controlled reproduction numbers ( $\rho_1$  and  $\rho_2$ ), the intrinsic (exponential) growth rates may also be changed, and thus become the controlled growth rates. Therefore, the relationship between  $\eta$  and  $r$  could be extended to the relationship between  $\eta_{\text{eff}}$  and controlled growth rate, and the mathematical relationships in Section 2 continue to hold.

For the control of infectious disease, the controlled reproduction numbers are typically less than their intrinsic values, i.e.,  $\rho_1 < \mathcal{R}_1$  and  $\rho_2 < \mathcal{R}_2$ , which implies reduction in the overall transmissibility. Although the reduction in reproduction numbers does not guarantee a reduction in their ratio ( $\eta_{\text{eff}}$ ), we expect a decrease in  $\eta_{\text{eff}}$ , i.e.,  $\eta_{\text{eff}} \leq \eta$ , under the selective pressure of nonpharmaceutical interventions (NPI). For instance, empirical evidence suggests that the decrease in the  $\eta_{\text{eff}}$  of SARS-CoV-2 B.1.1.7 (i.e., Alpha) variants is associated with the intensive social distancing in the United Kingdom (Graham et al., 2021a, 2021b; Volz et al., 2021a). Hence, we define the selective pressure ( $\vartheta$ ) as the attributed multiplicative change (or decrease) in the transmission advantage

$$\vartheta = 1 - \frac{\eta_{\text{eff}}}{\eta}, \quad (12)$$

where  $0 \leq \vartheta \leq 1$  if  $\eta_{\text{eff}} \leq \eta$ . Similar definition was also adopted previously (Leung et al., 2017). In the natural selection-free context as in Section 2.1,  $\vartheta = 0$ .

Since  $\eta_{\text{eff}}$  is time-varying,  $\vartheta$  is also time-varying. Hence, the selective pressure may lead to changes in the growth dynamics of mutated strains, i.e.,  $\rho_2(t)$  in Section 2.

### 3.2. Structural change in logistic growth attributed to selective pressure

A structural break is a considerable change over time in the values of parameters of statistical models, which could result in remarkable deviation from the original trends. The time-varying selection pressure may lead to changes in the exponential growth rate ( $r$ ), and thus cause structural changes in logistic growth.

Consider the situation that the selective pressure is absent in the first place, but then is placed consistently after time ( $t =$ )  $\tau_0$ . We model the structural changes using logistic regression with discontinuous design in the term  $r$

$$\mathbf{E} \left[ \ln \left[ \frac{\rho_2(t)}{1 - \rho_2(t)} \right] \right] = c + [r + \Delta r \mathbf{I}(t > \tau_0)]t, \quad (13)$$

where  $c$  is the intercept accounting for the initial settings of the original and mutated strains. The  $\mathbf{I}(\bullet)$  is an indicator function that takes the binary values (0 or 1); if  $t > \tau_0$ ,  $\mathbf{I}(t > \tau_0) = 1$ ; otherwise,  $\mathbf{I}(t > \tau_0) = 0$ . Thus, the growth rate is  $r$  when  $t \leq \tau_0$ , and  $(r + \Delta r)$  when  $t > \tau_0$ . Adopting the relation between transmission advantage and growth rate in Eqn (2), or other more complicated forms, we can further calculate the selective pressure using Eqn (12).

For the logistic growth curve in Eqn (3) and other similar forms, e.g., those derived in Sections 2.3 and 2.4, the discontinuous design in Eqn (13) for the term  $r$  can be expressed in differential forms,

$$\frac{d\rho_2}{dt} = [r + \Delta r \mathbf{I}(t > \tau_0)]\rho_2(1 - \rho_2), \text{ or equivalently } \frac{d}{dt} \log \left[ \frac{\rho_2(t)}{\rho_1(t)} \right] = r + \Delta r \mathbf{I}(t > \tau_0).$$

This formulation is consistent with the mathematical framework of natural selection originally proposed by Sir Ronald A Fisher (Fisher, 1958). Note that we expect  $\Delta r < 0$  such that the selective pressure leads to drops in growth rate as well as transmission advantage. As such, the growth with the selective pressure are slower than the situation without the selective pressure.

Furthermore, besides the discontinuous design in  $r$ , more flexible numerical approaches such as spline function are also applicable when the changes in selective pressures become more progressive.

#### 4. Numerical simulation

To illustrate how the transmission advantage may affect logistic growth of mutation proportion, we conduct numerical simulations using the settings and schemes introduced in this section. In this section, we adopted the classic SIR model, Eqn (4), for demonstration.

The technical frameworks are detailed in supplementary material S2.

##### 4.1. Simulation results

We show the key model settings and simulation results of the two scenarios in Fig. 1. Without the selective pressure, the proportion of mutated strains ( $\rho_2$ ) grows logistically, Fig. 1A. By contrast, an evident structural change in the growing trends can be found with the selective pressure places on Day 10, Fig. 1B.

As the consequence of the reduction in transmission rates, the numbers of cases as well as their growth rates decrease since Day 10, Fig. 1F, as well as the controlled reproduction numbers ( $r_1$  and  $r_2$ ), Fig. 1E. The intersection of the epidemic curves of two strains in Fig. 1F corresponds to the proportion of mutations at dominant threshold (i.e.,  $\rho_2 = 0.5$ ) in Fig. 1A or 1B, where we observed the timing of mutation reaching dominance is postponed due to the selective pressure. Regardless of the scenarios with or without selective pressure, the relatively stable levels of the susceptible population (almost constant) shown in Fig. 1G guarantee the logistic growth patterns of  $\rho_2$  discussed in Section 2.3. Additionally, we noticed a deviation of susceptible percentage from 100%, which reflected a consumption of susceptible population due to number of new cases in Fig. 1F. The exponential growth of cases was strictly relying on a constant susceptible population, and thus the growing dynamics of mutation frequency could no longer be solved analytically (but may be approximated numerically) if susceptible population has a major change. This minor decrease in susceptible population may lead to a slight change in the logistic growth of mutation frequency.

#### 5. Real-world example: SARS-CoV-2 B.1.1.7 variants in England

We further illustrate the setup in Section 3.2 by using the SARS-CoV-2 strains data in the United Kingdom as the real-world example.

The technical frameworks are detailed in supplementary material S3.

##### 5.1. Estimation results

We observe that the growth of B.1.1.7 roughly follows a logistic curve as shown in Fig. 2B. Assuming the absence of selective pressure under the scheme (I), the average transmission advantage is estimated at 1.38 (95%CI: 1.37, 1.39) or 1.38 (95%CI: 1.37, 1.39) by using the formula in Section 2.3.1 or Section 2.5.2, respectively, which is in line with many previous estimates (Campbell et al., 2021; Davies et al., 2021a; Graham et al., 2021a, 2021b; Leung et al., 2021; Volz et al., 2021a; Zhao et al., 2021a, 2021b, 2021c).

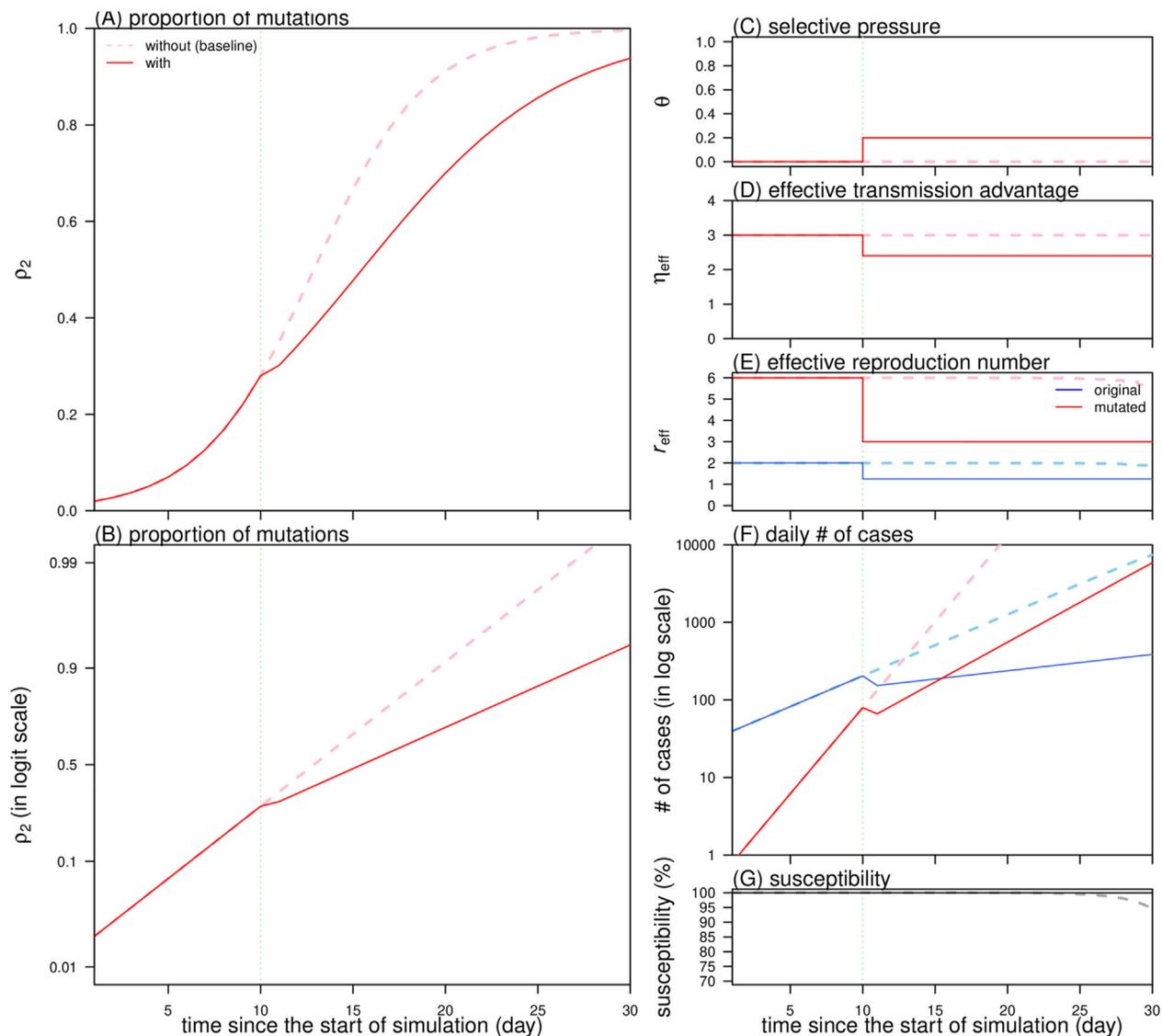
Using the fitting scheme (II), we detect a statistically significant structural change in the growth of B.1.1.7, Fig. 2C. Comparing the situations before and after the implementation of social distancing and local restrictions, we estimate the transmission advantage decreased from 1.65 (95%CI: 1.63, 1.67) to 1.40 (95%CI: 1.39, 1.41), Fig. 1D. Thus, the selective pressure was estimated increasing from 0 to 0.21 (95%CI: 0.19, 0.22), Fig. 1E. For validation, we also find decreasing trends in transmission advantage and increasing trends in selective pressure by using the fitting scheme (III) consistently.

For the goodness-of-fit, the values of McKelvey and Zavoina's pseudo-R-squared are 0.74, 0.80, and 0.81 for the fitting schemes (I)-(III), respectively. For the fitting performance, the AIC values of schemes (II)-(III) are 1150.1 and 1249.9 units less than that of fitting schemes (I), respectively. Regarding the decrease in transmission advantage in Fig. 2D, the superior performance of fitting schemes (II)-(III) matches the occurrence of the selective pressure on B.1.1.7 during the COVID-19 epidemic in England. Similar empirical findings are also reported (Graham et al., 2021a, 2021b), associating with the intensive social distancing and local restrictions (Volz et al., 2021a).

#### 6. Discussion

In this section, we summarize the key findings and conclusions, and discuss the limitations and extensions regarding the formulation and data analysis.

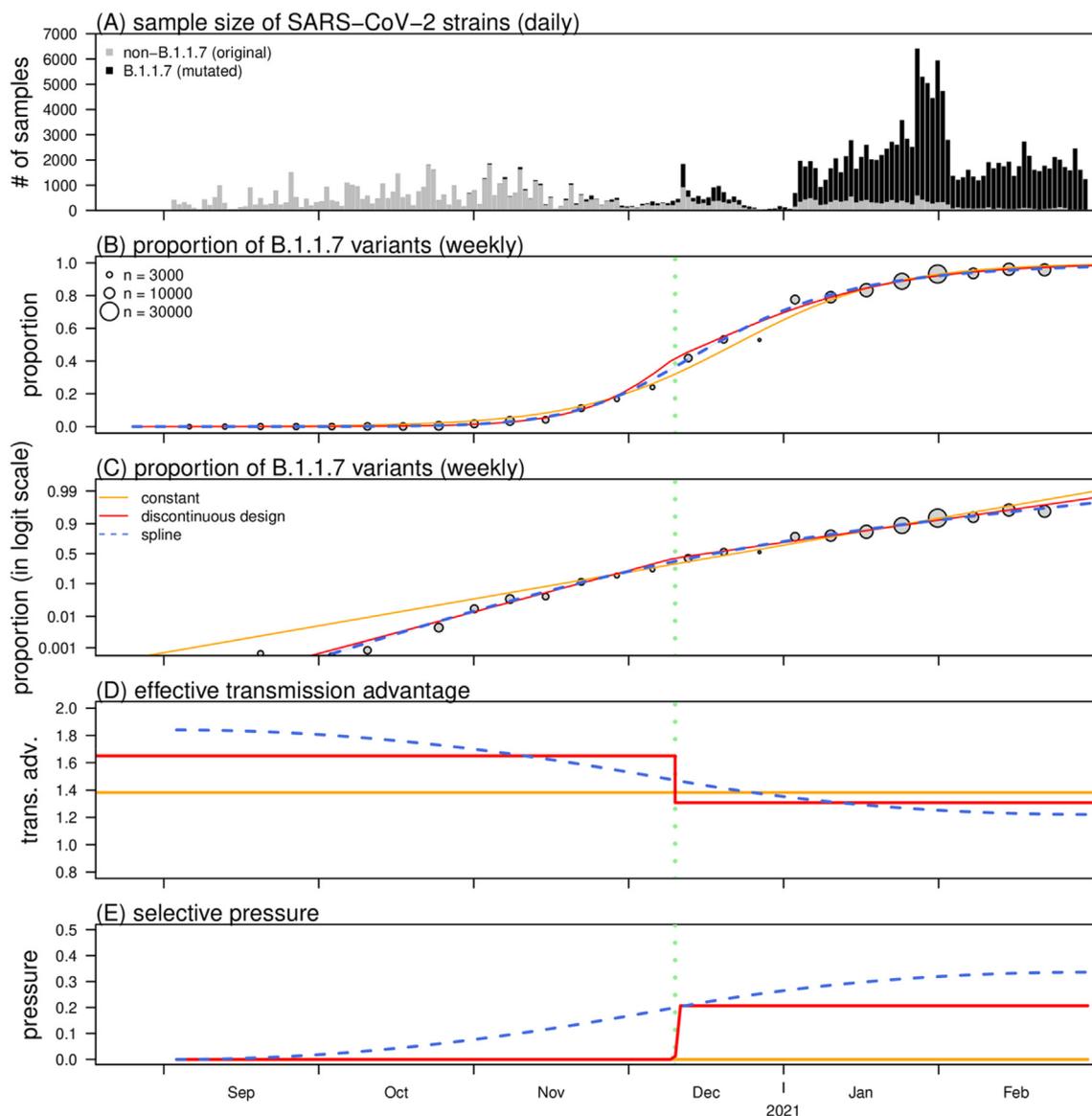
From several classic epidemic models, we derive the logistic growth of mutated strains in Section 2. When generalizing the epidemic model in settings and assumptions, the logistic growth requires more conditions, most of which appear to be biologically reasonable, and the relationship between the exponential growth rate ( $r$ ) and transmission advantage ( $\eta$ ) becomes more complex. The logistic growth is previously adopted as one of the parsimonious and classic models to study the competing process among different constitutions of lives under natural selection (Alizon et al., 2021; Fisher, 1958; Haim-Boukobza et al., 2021; MacArthur & Wilson, 2016). On the one hand, the logistic models in this study hold generally when the susceptibility (i.e.,  $S/N$ ) of the population is sufficiently stable, which means no dramatic changes occur during the growth of mutations. Notably, logistical regression models are frequently used to estimate the exponential growth rate of SARS-CoV-2



**Fig. 1.** The simulation outcomes of epidemic model based on Eqn (4) with and without the selective pressure. Panels (A) and (B) have the same contents, but show the simulated proportion of mutated strains in linear and logit scales, respectively. Panels (C), (D) and (E) show the time-varying selective pressure ( $\theta$ ), effective transmission advantage ( $\eta_{\text{eff}}$ ) and controlled reproduction numbers ( $r_1$  and  $r_2$ ), respectively. Panels (F) and (G) show the daily numbers of cases (in log scale) and susceptibility (i.e.,  $S/N$ ), respectively. In each panel, dashed curves are without the selective pressure, and normal curves are with the selective pressure; blue curves are for original strains, and red curves are for mutated strains. The vertical green dashed line indicates Day 10 of the simulation, when the selective pressure is placed.

mutations (Campbell et al., 2021; Davies et al., 2021a, 2021b; Volz et al., 2021b). On the other hand, when there is reduction in susceptibility, which may be caused by vaccine program (Anderson & May, 1985a, 1985b), the logistic growth no longer holds, and the growth rate will decrease. In the compartmental model, the vaccine effectiveness against susceptibility could be modelled as moving a portion of susceptible individuals to the “recovered” compartment, i.e., from class  $S$  to class  $R$ . However, such vaccination process is commonly depended on the external factors such as vaccine administration, vaccine seeking behaviors (Bauch, 2005), willingness of vaccine uptake, impacts of vaccine promotion programs, and sometimes, the projected intensity of outbreak situations (Bauch & Earn, 2004). Thus, it is difficult to find the analytical formula for the mutation growth, and thus numerical solutions are suggested alternatively. Although the intrinsic transmission advantage was unlikely to be changed by vaccines, genetic variants with a higher ability in immune escape would demonstrate an advantage in population with a substantial vaccine coverage.

One of the major benefits of our findings on logistic growth is to implement logistic regression models on the real-world strain samples, and infer the transmission advantage, see Section 5 and Eqn (15). Besides the selection-free context, we also



**Fig. 2.** Observed SARS-CoV-2 strains data in England, fitted, and estimated results. Panel (A) shows the daily sample size of SARS-CoV-2 strains collected in England, where the grey indicates the original strains (non-B.1.1.7), and the black indicates the mutated strains (B.1.1.7). Panels (B) and (C) have the same contents, but show the observed (dots) and fitted (curves) proportion of B.1.1.7 variants in linear and logit scales, respectively, where the size of dot indicates the sample size. Panels (D) and (E) show the estimated effective transmission advantage and the time-varying selective pressure ( $\vartheta$ , in Eqn (14)), respectively. The transmission advantage is calculated using Eqn (8), assuming the effective reproduction number of original strains equals 1. In panels (B)–(E), the vertical green dashed line indicates the day when the social distancing and (Tier 3 and Tier 4) local restrictions were enforced in the England. The orange, red, and blue dashed curves are from the results of fitting schemes (I)–(III) introduced in supplementary material S3, respectively.

investigated how the selective pressure may affect the demonstration of transmission advantage, and then change the growing dynamics of mutations, which is illustrated in Section 4. From the results of B.1.1.7 in England, we detect a statistically significant drop in the effective transmission advantage ( $\eta_{eff}$ ) in December 2020, which implies a change in selective pressure, see Fig. 2. Consistent with previous study (Volz et al., 2021a), we argue that this could be an outcome of intensive social distancing implemented during the same period. For clarification, we decompose the term  $\beta$ , i.e., the transmission rate in Eqns (4) and (10), into the contact rate (denoted by  $b$ ) and transmission probability per contact (denoted by  $\alpha$ , and  $0 \leq \alpha \leq 1$ ). Then, according to the classic epidemiological theories (Brauer et al., 2008; Held, Hens, D O'Neill, & Wallinga, 2019), we have  $\beta = ab$ . Considering the difference between the transmission rates of original and mutated strains in terms of  $\eta_\beta$ , the mutation is a biological factor, and thus will not affect the scale of  $b$ . Specifically, the factor  $\eta_\beta$  controls the multiplicative change in  $\alpha$  attributed to mutation. However, with social distancing, people are forced to stay at private locations, i.e., hotel and home, which implies a lower  $b$  but a higher  $\alpha$ . Social distancing increases the duration and proximity of each contact (Li et al., 2021),

and thus  $\alpha$  becomes higher (Luo et al., 2020). Note that the value of  $\alpha$  may become remarkably high, even close to 1, which means infectors are almost certain to transmit disease to their close contacts. Empirically, the proportion of household infections increases under intensive social distancing. As such, although the product of  $\alpha b$  decreases, the increase in  $\alpha$  may lead to decrease in the  $\eta_{\text{eff}}$ , which forms the selective pressure. Furthermore, as  $\alpha$  reaches 1, the transmission rates of the two types of strains approach each other and eventually converge to the same value. Thus,  $\eta_{\text{eff}}$  decreases from  $\eta_{\beta}\eta_{\gamma}$  to  $\eta_{\gamma}$ , which indicates that the transmission advantage controlled by the factor  $\eta_{\beta}$  vanishes.

Although the model frameworks in our study are formulated with two strains, they can be extended into multiple strains scenarios. Following the analytical procedures in Section 2.2, Eqn (3) is generalized as follows for the proportion of the  $j$ -th type of strains ( $\rho_j(t)$ ),

$$\rho_j(t) = \frac{\left[\frac{\rho_j(0)}{\rho_1(0)}\right] \exp(r_j t)}{1 + \sum_k \left[\frac{\rho_k(0)}{\rho_1(0)}\right] \exp(r_k t)}.$$

Here, the term  $r_j$  is the exponential growth rate of the  $j$ -th type of mutated strains, and thus the relationship in Eqn (2) is extended to

$$\eta_j = \exp(r_j T_g),$$

where  $\eta_j$  is the transmission advantage of the  $j$ -th against the first type (reference group) of strains. Therefore, the binary logistic framework can be generalized to a multinomial form (Zhao et al., 2021a, 2021b, 2021c), which is previously adopted in (Campbell et al., 2021). We note that likewise, all epidemic models in Section 2 can be extended to multi-strain forms, and similar multinomial logistic growth of each mutation can also be derived mathematically. Furthermore, the concepts of selective pressure and structural change in mutation growth are also applicable.

For the epidemic models in Section 2, one of the major assumptions is that once an individual is recovered from disease, the immunity is enduring and effective in preventing infection from different type of strains. In other words, re-infection by the same strains, and immune escape (or evasion) by different variants are ignored in the epidemic models. The assumption about re-infection might not hold for long-term epidemics where individuals can be infected more than once, e.g., measles (Sowers et al., 2016), or tuberculosis (Verver et al., 2005). However, if waning of immunity is sensitive to different types of variants, the logistic growth of mutation still holds despite the slight changes added to epidemic models. The detailed mathematical formulation is similar and is not covered in this study. For the assumption about immune escape, mutations may lead to an escape from the protection developed in previous infection of the original type of strains or vaccination, e.g., influenza (Park et al., 2009), and COVID-19 (Cao, 2020), or chronic virus infections, e.g., hepatitis (Lim et al., 2007), and HIV (Pennings, Kryazhimskiy, & Wakeley, 2014). As a result, the average susceptibility of the population to the mutated strains may be higher than that to the original strains, when there is partial or no cross-immunity. Hence, the transmission advantage of mutated against original strains becomes higher, and the growth rate of mutants increases. Since the depletion process of the susceptible becomes complex in this case, the growth dynamics of mutations may no longer be logistic. However, for the epidemics with short duration and relatively small outbreak size, the logistic frameworks in Section 2 are still applicable to govern the overall patterns of real-world data.

Several limitations for the statistical analysis in Section 5 regarding the data in the aspects of model assumptions and interpretations are discussed in supplementary material S4.

## 7. Conclusions

Based on classical epidemiology theories of disease transmission, we found that the between-host transmission advantage may shape the observed logistic growth of the viral mutation proportion at the population scale. The logistic growth model can be further generalized by incorporating time-varying selective pressure to account for impacts of external factors on pathogen adaptiveness. The model underlies theoretical mechanism that shapes the logistic growth of mutation in population.

## Ethics approval and consent to participate

The ethical approval or individual consent was not applicable. The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

## Data availability statement

The data of SARS-CoV-2 strains used in this work are publicly available, and their accessing identities are provided and acknowledged in the appendix.

## Funding

DH was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [HKU C7123-20G]. This study was supported by the National Natural Science Foundation of China (NSFC) [31871340, 71974165], Health and Medical Research Fund, the Food and Health Bureau, the Government of the Hong Kong Special Administrative Region [INF-CUHK-1, COVID190103], China, and partially supported by the CUHK grant [PIEF/Ph2/COVID/06, 4054600].

## Author contributions

Conceptualization: SZ, and MHW. Methodology: SZ, IH, and MHW. Software: SZ, and JL. Validation: SZ. Formal analysis: SZ, and JL. Investigation: SZ. Resources: SZ, and JL. Data Curation: SZ, and JL. Writing - Original Draft: SZ. Writing - Review and Editing: IH, MKCC, LC, DH, BCYZ, and MHW. Visualization: SZ. Supervision: MHW. Project Administration: SZ, and JL. Funding acquisition: MKCC, DH, and MHW. All authors critically read the manuscript, and gave final approval for publication.

## Declaration of competing interest

BCYZ is a shareholder of Health View Bio-analytics Ltd. MHW is a shareholder of Beth Bioinformatics Co., Ltd. Other authors declared no competing interests. The funding agencies had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

## Acknowledgements

None.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.idm.2022.12.006>.

## References

- Alizon, S., Haim-Boukobza, S., Foulongne, V., Verdurme, L., Trombert-Paolantoni, S., Lecorche, E., et al. (2021). Rapid spread of the SARS-CoV-2 Delta variant in some French regions, June 2021. *Euro Surveill*, 26(28), Article 2100573.
- Anderson, R. M., & May, R. M. (1985a). Helminth infections of humans: Mathematical models, population dynamics, and control. *Advances in Parasitology*, 24, 1–101.
- Anderson, R. M., & May, R. M. (1985b). Vaccination and herd immunity to infectious diseases. *Nature*, 318(6044), 323–329.
- Athreya, K. B., Ney, P. E., & Ney, P. E. (2004). *Branching processes*. Courier Corporation.
- Bauch, C. T. (2005). Imitation dynamics predict vaccinating behaviour. *Proceedings of the Royal Society B: Biological Sciences*, 272(1573), 1669–1675.
- Bauch, C. T., & Earn, D. J. (2004). Vaccination and the theory of games. *Proceedings of the National Academy of Sciences of the United States of America*, 101(36), 13391–13394.
- Baum, A., Fulton, B. O., Wloga, E., Copin, R., Pascal, K. E., Russo, V., et al. (2020). Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science*, 369(6506), 1014–1018.
- Brauer, F., Driessche, P. V., & Wu, J. (2008). *Lecture notes in mathematical epidemiology*, 75 pp. 3–22. Berlin: Germany Springer, 1.
- Campbell, F., Archer, B., Laurenson-Schafer, H., Jinnai, Y., Konings, F., Batra, N., et al. (2021). Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Euro Surveill*, 26(24), Article 2100509.
- Cao, X. (2020). COVID-19: Immunopathology and its implications for therapy. *Nature Reviews Immunology*, 20(5), 269–270.
- Chevin, L.-M. (2011). On measuring selection in experimental evolution. *Biological Letters*, 7(2), 210–213.
- Coombs, D., Gilchrist, M. A., & Ball, C. L. (2007). Evaluating the importance of within-and between-host selection pressures on the evolution of chronic pathogens. *Theoretical Population Biology*, 72(4), 576–591.
- Cori, A., Ferguson, N. M., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9), 1505–1512.
- Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., et al. (2021a). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, (6538), 372.
- Davies, N. G., Jarvis, C. I., van Zandvoort, K., Clifford, S., Sun, F. Y., Funk, S., et al. (2021b). Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature*, 593(7858), 270–274.
- Diekmann, O., Heesterbeek, J. A. P., & Roberts, M. G. (2010). The construction of next-generation matrices for compartmental epidemic models. *Journal of The Royal Society Interface*, 7(47), 873–885.
- van den Driessche, P., & Watmough, J. (2002). Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180, 29–48.
- Faria, N. R., Mellan, T. A., Whittaker, C., Claro, I. M., Candido, D. S., Mishra, S., et al. (2021). Genomics and epidemiology of the P. 1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*, 372(6544), 815(–+).
- Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dorner, L., et al. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491), Article eabb6936.
- Fine, P. E. (2003). The interval between successive cases of an infectious disease. *American Journal of Epidemiology*, 158(11), 1039–1047.
- Fisher, R. A. (1958). *The genetical theory of natural selection*. Рипол Классик.
- Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., et al. (2020). Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data. March 2020 *Euro Surveill : bulletin European sur les maladies transmissibles = European communicable disease bulletin*, 25(17), Article 2000257.
- Getz, W. M., & Pickering, J. (1983). Epidemic models: Thresholds and population regulation. *The American Naturalist*, 121(6), 892–898.

- Gog, J. R., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., & Grenfell, B. T. (2003). Population dynamics of rapid fixation in cytotoxic T lymphocyte escape mutants of influenza A. *Proceedings of the National Academy of Sciences*, *100*(19), 11143–11147.
- Graham, M. S., Sudre, C. H., May, A., Antonelli, M., Murray, B., Varsavsky, T., et al. (2021a). Changes in symptomatology, re-infection and transmissibility associated with SARS-CoV-2 variant B. 1.1. 7: An ecological study. *medRxiv*.
- Graham, M. S., Sudre, C. H., May, A., Antonelli, M., Murray, B., Varsavsky, T., et al. (2021b). Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B. 1.1. 7: An ecological study. *The Lancet Public Health*, *6*(5), e335–e345.
- Haim-Boukobza, S., Roquebert, B., Trombert-Paolantoni, S., Lecorche, E., Verdurme, L., Foulongne, V., et al. (2021). Detecting rapid spread of SARS-CoV-2 variants, France, January 26–February 16, 2021. *Emerging Infectious Diseases*, *27*(5), 1496–1499.
- Held, L., Hens, N., D’O’Neill, P., & Wallinga, J. (2019). *Handbook of infectious disease data analysis*. CRC Press.
- Ito, K., Piantam, C., & Nishiura, H. (2021). Predicted domination of variant delta of SARS-CoV-2 before Tokyo Olympic games, Japan. *Euro Surveillance*, *26*(27), Article 2100570.
- Kraemer, M. U. G., Cummings, D. A. T., Funk, S., Reiner, R. C., Faria, N. R., Pybus, O. G., et al. (2019). Reconstruction and prediction of viral disease epidemics. *Epidemiology and Infection*, *147*, Article e34.
- Kühnert, D., Kouyos, R., Shirreff, G., Pecerska, J., Scherrer, A. U., Böni, J., et al. (2018). Quantifying the fitness cost of HIV-1 drug resistance mutations through phylodynamics. *PLoS Pathogens*, *14*(2), Article e1006895.
- Leung, K., Lipsitch, M., Yuen, K. Y., & Wu, J. T. (2017). Monitoring the fitness of antiviral-resistant influenza strains during an epidemic: A mathematical modelling study. *The Lancet Infectious Diseases*, *17*(3), 339–347.
- Leung, K., Shum, M. H., Leung, G. M., Lam, T. T., & Wu, J. T. (2021). Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. *Euro Surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, *26*(1), Article 2002106.
- Li, Y., Liu, J., Yang, Z., Yu, J., Xu, C., Zhu, A., et al. (2021). Transmission of severe acute respiratory syndrome coronavirus 2 to close contacts, China, January–February 2020. *Emerging Infectious Disease Journal*, *27*(9), 2288–2293.
- Lim, S. G., Cheng, Y., Guindon, S., Seet, B. L., Lee, L. Y., Hu, P., et al. (2007). Viral quasi-species evolution during hepatitis Be antigen seroconversion. *Gastroenterology*, *133*(3), 951–958.
- Lotka, A. J. (1907). Relation between birth rates and death rates. *Science*, *26*(653), 21–22.
- Lowder, B. V., Guinane, C. M., Zakour, N. L. B., Weinert, L. A., Conway-Morris, A., Cartwright, R. A., et al. (2009). Recent human-to-poultry host jump, adaptation, and pandemic spread of *Staphylococcus aureus*. *Proceedings of the National Academy of Sciences*, *106*(46), 19545–19550.
- Luo, L., Liu, D., Liao, X., Wu, X., Jing, Q., Zheng, J., et al. (2020). Contact settings and risk for transmission in 3410 close contacts of patients with COVID-19 in Guangzhou, China: A prospective cohort study. *Annals of Internal Medicine*, *173*(11), 879–887.
- Lythgoe, K. A., Gardner, A., Pybus, O. G., & Grove, J. (2017). Short-sighted virus evolution and a germline hypothesis for chronic viral infections. *Trends in Microbiology*, *25*(5), 336–348.
- MacArthur, R. H., & Wilson, E. O. (2016). *The theory of island biogeography*. Princeton university press.
- Malthus, T. R. (1872). *An essay on the principle of population*.
- Metcalf, C. J. E., Birger, R., Funk, S., Kouyos, R., Lloyd-Smith, J., & Jansen, V. J. E. (2015). Five challenges in evolution and infectious diseases. *Epidemics*, *10*, 40–44.
- Metz, J. A. J., Nisbet, R. M., & Geritz, S. A. H. (1992). How should we define ‘fitness’ for general ecological scenarios? *Trends in Ecology & Evolution*, *7*(6), 198–202.
- Park, A. W., Daly, J. M., Lewis, N. S., Smith, D. J., Wood, J. L. N., & Grenfell, B. T. (2009). Quantifying the impact of immune escape on transmission dynamics of influenza. *Science*, *326*(5953), 726–728.
- Pennings, P. S., Kryazhimskiy, S., & Wakeley, J. (2014). Loss and recovery of genetic diversity in adapting populations of HIV. *PLoS Genetics*, *10*(1), Article e1004000.
- Ren, X., Li, Y., Yang, X., Li, Z., Cui, J., Zhu, A., et al. (2021). Evidence for pre-symptomatic transmission of coronavirus disease 2019 (COVID-19) in China. *Influenza Other Respir Viruses*, *15*(1), 19–26.
- Schreiber, S. J., Ke, R., Loverdo, C., Park, M., Ahsan, P., & Lloyd-Smith, J. O. (2021). Cross-scale dynamics and the evolutionary emergence of infectious diseases. *Virus Evolution*, *7*(1), Article veaa105.
- Shih, A. C.-C., Hsiao, T.-C., Ho, M.-S., & Li, W.-H. (2007). Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proceedings of the National Academy of Sciences*, *104*(15), 6283–6288.
- Sowers, S. B., Rota, J. S., Hickman, C. J., Mercader, S., Redd, S., McNall, R. J., et al. (2016). High concentrations of measles neutralizing antibodies and high-avidity measles IgG accurately identify measles reinfection cases. *Clinical and Vaccine Immunology*, *23*(8), 707–716.
- Steel, J., Lowen, A. C., Mubareka, S., & Palese, P. (2009). Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N. *PLoS Pathogens*, *5*(1), Article e1000252.
- Svensson, A. (2007). A note on generation times in epidemic models. *Mathematical Biosciences*, *208*(1), 300–311.
- Tang, J. W., Tambyah, P. A., & Hui, D. S. (2021). Emergence of a new SARS-CoV-2 variant in the UK. *Journal of Infection*, *82*(4), e27–e28.
- Tindale, L. C., Stockdale, J. E., Coombe, M., Garlock, E. S., Lau, W. Y. V., Saraswat, M., et al. (2020). Evidence for transmission of COVID-19 prior to symptom onset. *Elife*, *9*, Article e57149.
- Travisano, M., & Lenski, R. E. (1996). Long-term experimental evolution in *Escherichia coli*. IV. Targets of selection and the specificity of adaptation. *Genetics*, *143*(1), 15–26.
- Troeger, C., Blacker, B., Khalil, I. A., Rao, P. C., Cao, J., Zimsen, S. R. M., et al. (2018). Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: A systematic analysis for the global burden of disease study 2016. *The Lancet Infectious Diseases*, *18*(11), 1191–1210.
- Tsatsarkin, K. A., Vanlandingham, D. L., McGee, C. E., & Higgs, S. (2007). A single mutation in chikungunya virus affects vector specificity and epidemic potential. *PLoS Pathogens*, *3*(12), Article e201.
- Van den Driessche, P. (2017). Reproduction numbers of infectious disease models. *Infectious Disease Modelling*, *2*(3), 288–303.
- Verver, S., Warren, R. M., Beyers, N., Richardson, M., Van Der Spuy, G. D., Borgdorff, M. W., et al. (2005). Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, *171*(12), 1430–1435.
- Volz, E., Hill, V., McCrone, J. T., Price, A., Jorgensen, D., O’Toole, A., et al. (2021b). Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell*, *184*(1), 64–75 e11.
- Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., et al. (2021a). Assessing transmissibility of SARS-CoV-2 lineage B. 1.1. 7 in England. *Nature*, *593*(7858), 1–17.
- Wallinga, J., & Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1609), 599–604.
- Wallinga, J., & Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, *160*(6), 509–516.
- Wang, M. H., Lou, J., Cao, L., Zhao, S., Chan, R. W. Y., Chan, P. K. S., et al. (2021). Characterization of key amino acid substitutions and dynamics of the influenza virus H3N2 hemagglutinin. *Journal of Infection*, *83*(6), 671–677.
- Whittles, L. K., White, P. J., & Didelot, X. (2017). Estimating the fitness cost and benefit of cefixime resistance in *Neisseria gonorrhoeae* to inform prescription policy: A modelling study. *PLoS Medicine*, *14*(10), Article e1002416.
- Wise, J. (2021). *Covid-19: The E484K mutation and the risks it poses*. British Medical Journal Publishing Group.

- Zhao, S., Lou, J., Cao, L., Zheng, H., Chong, M. K. C., Chen, Z., et al. (2021a). Quantifying the transmission advantage associated with N501Y substitution of SARS-CoV-2 in the UK: An early data-driven analysis. *Journal of Travel Medicine*, 28(2), Article taab011.
- Zhao, S., Lou, J., Cao, L., Zheng, H., Chong, M. K. C., Chen, Z., et al. (2021b). Modelling the association between COVID-19 transmissibility and D614G substitution in SARS-CoV-2 spike protein: Using the surveillance data in California as an example. *Theoretical Biology and Medical Modelling*, 8(1), 10.
- Zhao, S., Lou, J., Cao, L., Zheng, H., Chong, M. K. C., Chen, Z., et al. (2021c). Real-time quantification of the transmission advantage associated with a single mutation in pathogen genomes: A case study on the D614G substitution of SARS-CoV-2. *BMC Infectious Diseases*, 21(1), 1039.
- Zhao, S., Musa, S. S., Hebert, J. T., Cao, P., Ran, J., Meng, J., et al. (2020). Modelling the effective reproduction number of vector-borne diseases: The yellow fever outbreak in luanda, Angola 2015-2016 as an example. *PeerJ*, 8, Article e8601.
- Zhu, W., & Shu, Y. (2015). Genetic tuning of avian influenza A (H7N9) virus promotes viral fitness within different species. *Microbes and Infection*, 17(2), 118–122.