SIAM J. APPL. MATH.
Vol. 77, No. 6, pp. 2156–2171

TAJIMA'S D AND SITE-SPECIFIC NUCLEOTIDE FREQUENCY IN A POPULATION DURING AN INFECTIOUS DISEASE OUTBREAK*

RYOSUKE OMORI[†] AND JIANHONG WU[‡]

Abstract. Tajima's D measures the difference between two estimates of genetic diversity in a given set of nucleic acid sequences. Here we show how Tajima's D can be calculated/estimated by developing an inductive algorithm for calculating the site-specific nucleotide frequencies from a standard multistrain susceptible-infective-removed (SIR) model (both deterministic and stochastic). We show that these frequencies are fully determined by the mutation rate and the initial condition of the frequencies. We prove that the sign of Tajima's D is independent of the disease population dynamics and that the negative sign does not imply an expansion of the infected population in the deterministic model. Using individual-based simulations, we also show that dependence of Tajima's D on the disease transmission and evolution dynamics is a result of the stochasticity of those dynamics. The same is true for the dependence related to genetic diversity of a pathogen.

Key words. Tajima's D, SIR model, evolution, infectious disease

AMS subject classifications. 92B05, 92D30, 92D25

DOI. 10.1137/17M1114946

1. Introduction.

The rapid development of nucleic acid sequencing methodologies and technologies has contributed to the accumulation of enormous amounts of sequence data. This large quantity of data makes it easier to predict the impact of each mutation on the fitness of a pathogen via estimation of the selection pressure on the mutation [14]. For infectious disease epidemiology, the estimation of a mutation contributing to the fitness of a pathogen can be used to provide insight into mutations, e.g., discovering mutations determining pathogenicity, virulence, and host specificity. Rapid sequencing technologies can also capture detailed information about the time evolution of pathogen sequences. Such time series data are useful not only for the detection of pathogen evolution but also for capturing the pathogen population dynamics. Coalescent theory in neutral evolution, i.e., no natural selection and constant population size, can be used to estimate population size from the genetic diversity of sampled sequences [13]. The extension of basic coalescent theory to discrete changes in population over time allows us to estimate the time series of such changes [16, 17], which can then be applied to estimate the reproduction number of a pathogen [19].

Several methods have been proposed to measure the natural selection of a mutation at a specific site. Tajima's test is one of the most popular statistical tests of evolution neutrality at the sequence level. Tajima's D measures the difference

[*] Received by the editors February 6, 2017; accepted for publication (in revised form) September 29, 2017; published electronically December 6, 2017.
http://www.siam.org/journals/siap/77-6/M111494.html

Funding: The first author's work was supported by Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Agency (JST), and Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers, Japan Society for the Promotion of Science (JSPS). The second author's work was partially supported by the Canada Research Chair Program (230720) and by the Natural Sciences and Engineering Research Council of Canada (105588-2011).

[†] Division of Bioinformatics, Research Center for Zoonosis Control, Hokkaido University, Sapporo, Hokkaido, 001-0020, Japan, and JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama, 332-0012, Japan (omori@czc.hokudai.ac.jp).

[‡] Laboratory for Industrial and Applied Mathematics, York University, Toronto, Ontario, Canada, M3J1P3 (wujh@mathstat.yorku.ca).

between two estimates of genetic diversity [21]. These estimates are equal when evolution is neutral (i.e., there is no natural selection and a constant population). When this happens, Tajima's D = 0. Therefore, a nonzero Tajima's D implies that at least one condition for neutral evolution is violated. It is known that when Tajima's D is nonzero, the sign of Tajima's D gives us the interpretation of natural selection: balancing selection can result in positive Tajima's D, and positive selection can result in negative Tajima's D [1]. Not only natural selection but also population dynamics determines the sign of Tajima's D, and it is believed that negative Tajima's D results from an increase in population size and positive Tajima's D results from a decrease in population size [8, 18, 20, 11]. When it comes to the nonendemic situation of infectious disease transmission, if we assume that the number of pathogens is proportional to the number of infected hosts, i.e., the population dynamics of the pathogen is proportional to the disease dynamics, the population of the pathogen is always changing during an outbreak. Therefore, Tajima's D can change over time. To detect the natural selection of an infectious disease in a nonendemic situation, we need to understand the impact of disease dynamics on Tajima's D. This is the main objective of our study.

To focus on the effect of population dynamics alone, we assume that no mutation contributes to a change in the pathogen phenotypes. In section 2.1, we will introduce the concept of Tajima's D and discuss key components (strain-specific nucleotide frequencies) of Tajima's D. Then in section 2.2, we link these to the population infection dynamics of a pathogen through a simple multistrain SIR (susceptible-infected-removed) model to describe the population dynamics of the pathogen. We show in section 3.1 how Tajima's D can be calculated by developing an inductive algorithm for calculating the site-specific nucleotide frequencies from the multistrain SIR model. Then in section 3.2, we formulate the stochastic analogue of the multistrain SIR model and perform Monte Carlo simulations to compare the effects of disease dynamics described by both deterministic and stochastic SIR models on Tajima's D values. We observed the dependence of Tajima's D on the stochasticity of transmission dynamics during an outbreak. In the final section, we discuss the implication and the application of our results in evolutionary and epidemiological analyses.

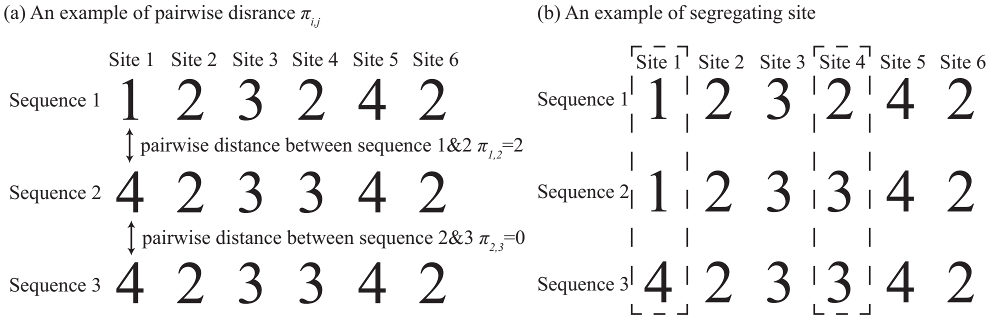
2. Site-specific frequency of nucleotide and models.

2.1. Tajima's D.

A nucleic acid sequence is a sequence of nucleotides consisting of four values. The nucleotides are A, T, G, or C for DNA sequences, and A, U, G, or C for RNA sequences. For the sake of simplicity, we refer to A, T (U), G, and C as 1, 2, 3, and 4. Here, we consider a set of sampled sequences with length L. The possible number of sequences is 4^L. We denote by x a sequence of nucleotides, with x_l representing the nucleotide at the lth site, that is,

$$x = \{x_1, x_2, \dots, x_L\}, x_l \in \{1, 2, 3, 4\}.$$

Figure 1 gives the sequence data of 3 sequences of length 6. To define Tajima's D, we define $\pi_{i,j}$ in a pair of sequences (ith sequence and jth sequence) as the number of unmatched sites. We call a site a segregating site when there are at least two different nucleotides among the sampled sequences. In Figure 1(a), we see $\pi_{1,2} = 2$ and $\pi_{2,3} = 0$, whereas in Figure 1(b), there are two segregating sites. For a given n sampled sequences, Tajima's D describes the variation in sequences among these samples. This is a function of the mean pairwise distance $\overline{\pi}$ and the number of


(a) An example of pairwise disrance $\pi_{i,j}$

Site 1 Site 2 Site 3 Site 4 Site 5 Site 6
Sequence 1 1 2 3 2 4 2
pairwise distance between sequence 1&2 $\pi_{1,2}$=2
Sequence 2 4 2 3 3 4 2
pairwise distance between sequence 2&3 $\pi_{2,3}$=0
Sequence 3 4 2 3 3 4 2

(b) An example of segregating site

Site 1 Site 2 Site 3 Site 4 Site 5 Site 6
Sequence 1 1 2 3 2 4 2
Sequence 2 1 2 3 3 4 2
Sequence 3 4 2 3 3 4 2

FIG. 1. *A given set of three sampled sequences.* (a) *Pairwise distance. The pairwise distance is defined as the number of unmatched sites.* (b) *Segregating sites. A segregating site is a site where there is at least two different nucleotides among the sequence dataset.*

segregating sites σ. Namely,

$$D = \frac{\overline{\pi} - \sigma/a}{\sqrt{\mathrm{Var}\left(\pi_{i,j} - \sigma/a\right)}}, \tag{1}$$

where

$$\overline{\pi} = \left(\sum_i \sum_{j<i} \pi_{i,j}\right) \Big/ \binom{n}{2}, \tag{2}$$

$$a = \sum_{i}^{n-1} \frac{1}{i}.$$

In [21], the denominator of Tajima's D is given by the function of σ,

$$\sqrt{\mathrm{Var}\left(\pi_{i,j} - \sigma/a\right)} = \sqrt{g_1\sigma + g_2\sigma\left(\sigma - 1\right)},$$

where

$$g_1 = \frac{\frac{1+n}{3(n-1)} - \frac{1}{a}}{a},$$
$$g_2 = \frac{\frac{a_2}{a^2} - \frac{2+n}{an} + \frac{2\left(3+n+n^2\right)}{9n(n-1)}}{a^2 + a_2},$$
$$a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}.$$

To calculate Tajima's D, we require only (i) the site-specific frequency of each nucleotide among sampled sequences f, and (ii) the number of sampled sequences n. In the next subsection, we start with the calculation of Tajima's D for a specific site and then expand the calculation to include multiple sites.

2.1.1. Tajima's D for a specific site. In this subsection, we focus on a specific site, the lth site, in the sequences to calculate Tajima's D. Let f_l^k be the relative frequency of nucleotide k in the lth site among sampled sequences, where

$$k \in \{1,2,3,4\}, f_l^1 + f_l^2 + f_l^3 + f_l^4 = 1.$$

The numerator in the right-hand side of (2) is the sum of pairwise distances with respect to the lth site among all pairs of existing sequences and can be written as f_l^k. The sum of pairwise distances is the product of (i) the number of pairs classified by the nucleotide, e.g., 1 and 1, 1 and 2,..., 4, and 4, and (ii) the pairwise distance for the pair classified by the nucleotide. The number of pairs classified by the nucleotide except self-pairing is given by

$$\begin{pmatrix} n^2 f_l^1 f_l^1 - n f_l^1 & n^2 f_l^2 f_l^1 & n^2 f_l^3 f_l^1 & n^2 f_l^4 f_l^1 \\ n^2 f_l^1 f_l^2 & n^2 f_l^2 f_l^2 - n f_l^2 & n^2 f_l^3 f_l^2 & n^2 f_l^4 f_l^2 \\ n^2 f_l^1 f_l^3 & n^2 f_l^2 f_l^3 & n^2 f_l^3 f_l^3 - n f_l^3 & n^2 f_l^4 f_l^3 \\ n^2 f_l^1 f_l^4 & n^2 f_l^2 f_l^4 & n^2 f_l^3 f_l^4 & n^2 f_l^4 f_l^4 - n f_l^4 \end{pmatrix}.$$

The pairwise distance for the pair classified by the nucleotide is given by

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

Therefore, the total pairwise distance per each pair classified by the nucleotide is written as

$$\begin{pmatrix} 0 & n^2 f_l^2 f_l^1 & n^2 f_l^3 f_l^1 & n^2 f_l^4 f_l^1 \\ n^2 f_l^1 f_l^2 & 0 & n^2 f_l^3 f_l^2 & n^2 f_l^4 f_l^2 \\ n^2 f_l^1 f_l^3 & n^2 f_l^2 f_l^3 & 0 & n^2 f_l^4 f_l^3 \\ n^2 f_l^1 f_l^4 & n^2 f_l^2 f_l^4 & n^2 f_l^3 f_l^4 & 0 \end{pmatrix}.$$

Note that this matrix is a symmetry matrix, and the sum of all elements of this matrix counts each nucleotide pair twice. Then

$$\begin{aligned} \overline{\pi}_l &= \frac{n^2\left[f_l^1\left(1-f_l^1\right)+f_l^2\left(1-f_l^2\right)+f_l^3\left(1-f_l^3\right)+f_l^4\left(1-f_l^4\right)\right]}{2\binom{n}{2}} \\ &= \frac{n\left[f_l^1\left(1-f_l^1\right)+f_l^2\left(1-f_l^2\right)+f_l^3\left(1-f_l^3\right)+f_l^4\left(1-f_l^4\right)\right]}{n-1}. \end{aligned} \tag{3}$$

With respect to the number of segregating sites for the l_1th site, σ_l, we have from its definition the following:

$$\sigma_l = \begin{cases} 0 & \text{if } f_l^1 = 1 \text{ or } f_l^2 = 1 \text{ or } f_l^3 = 1 \text{ or } f_l^4 = 1, \\ 1 & \text{otherwise.} \end{cases}$$

Assuming that the sequence sampling process is random and the sampling process of the nucleotide in the lth site follows a multinomial process, the maximum likelihood estimate of the frequency of nucleotide k in the lth site among the population is equal to f_l^k. Therefore, the sampling probability of σ_l among n sampled sequences is given by

$$\begin{aligned} Pr\left(\sigma(t)=0\right) &= \left(f_l^1\right)^n + \left(f_l^2\right)^n + \left(f_l^3\right)^n + \left(f_l^4\right)^n, \\ Pr\left(\sigma(t)=1\right) &= 1 - \left(\left(f_l^1\right)^n + \left(f_l^2\right)^n + \left(f_l^3\right)^n + \left(f_l^4\right)^n\right). \end{aligned}$$

The expected value of σ_l is

$$\mathrm{E}\left(\sigma_l\right) = 1 - \left(\left(f_l^1\right)^n + \left(f_l^2\right)^n + \left(f_l^3\right)^n + \left(f_l^4\right)^n\right). \tag{4}$$

Meanwhile, Tajima's D at the lth site D_l is given by

$$D_l = \frac{\overline{\pi_l} - \sigma_l/a}{\sqrt{g_1\sigma_l + g_2\sigma_l\left(\sigma_l - 1\right)}}. \tag{5}$$

Assuming $\sigma_l = \mathrm{E}(\sigma_l)$, from (3), (4), and (5) we notice that D_l is determined only by f_l^1, f_l^2, f_l^3, f_l^4, and n.

2.1.2. Tajima's D for *L* sites. Assuming no linkage between sites, the mean pairwise distance among L sites of n sampled sequences is simply given by the sum of π_l,

$$\overline{\pi} = \sum_{l=1}^{L} \frac{n\left[f_l^1\left(1 - f_l^1\right) + f_l^2\left(1 - f_l^2\right) + f_l^3\left(1 - f_l^3\right) + f_l^4\left(1 - f_l^4\right)\right]}{n-1}. \tag{6}$$

σ is given by

$$\sigma = \sum_{l=1}^{L} \sigma_l,$$

where

$$\sigma_l = \begin{cases} 0 & \text{if } f_l^1 = 1 \text{ or } f_l^2 = 1 \text{ or } f_l^3 = 1 \text{ or } f_l^4 = 1, \\ 1 & \text{otherwise.} \end{cases} \tag{7}$$

If the site-specific frequency of nucleotides is assumed to be independent, i.e., the sampling process of site-specific nucleotides follows a multinomial process ($E(f_l^k) = f_l^k$), then the expected number of segregating sites σ among n sampled sequences is

$$\mathrm{E}\left(\sigma\right) = \sum_{l=1}^{L} 1 - \left(\left(f_l^1\right)^n + \left(f_l^2\right)^n + \left(f_l^3\right)^n + \left(f_l^4\right)^n\right). \tag{8}$$

Substituting (6) and (7) or (8) into (1) yields immediately that Tajima's D is determined only by f_l^k and n.

2.2. Disease dynamics model. To explore the relationship between Tajima's D and disease dynamics, we construct a multistrain SIR model [3, 7, 15]. Here, the classification of strain is such that two sequences are considered the same strain if and only if they have identical sequences. We also assume that the number of sequences in the host population is proportional to the number of infected individuals corresponding to the strain. To focus only on the impact of disease dynamics, we assume that the evolution of the pathogen is neutral and has no effect on the phenotype. Then the transmission rate and recovery rate are identical among all sequences, and the established immunity against a strain carrying one sequence protects hosts against infections, with all strains carrying any other sequences. The population dynamics of the hosts infected with the strain carrying sequences x, I_x, can be described by

$$\begin{aligned} S'(t) &= -\beta S(t) \textstyle\sum_y I_y(t), \\ I_x'(t) &= \beta S(t) I_x - \gamma I_x(t) + \left(\textstyle\sum_y \mu_{y\to x} I_y(t)\right) - \left(\textstyle\sum_y \mu_{x\to y}\right) I_x(t), \\ R_x'(t) &= \gamma I_x(t), \end{aligned} \tag{9}$$

where β denotes the transmission rate, γ denotes the recovery rate, and $\mu_{x\to y}$ denotes the mutation rate from sequence x to sequence y. Here we assume that every mutation in a given host replaces the dominant genotype of the pathogen. Suppose the epidemic duration is short compared to the life span of the host; in this case the host population size is constant:

$$S(t) + \sum_x I_x(t) + \sum_x R_x(t) = N.$$

2.2.1. 1-site model. We start from the simplest 1-site model. We assume that only one site (the lth site) in the genetic sequences can mutate and that the nucleotides in other sites are the same among all sequences. Therefore, only four sequences can exist. Hereafter we refer to them as $x1$, $x2$, $x3$, and $x4$. The nucleotides in the lth site of sequences $x1$, $x2$, $x3$, and $x4$ are 1, 2, 3, and 4. Mutation can occur (I_x can be I_y by mutation, where I_x denotes the number of infected individuals with strain x). In population genetics, many models for mutation have been proposed. For example, Jukes and Cantor assumed that all site-specific mutation rates between nucleotides are the same [10], and Kimura assumed two mutation rates: a constant rate for the mutation between A and G and between C and T (U), and a different rate for all other mutations [12]. If we follow the simplest mutation model and use the Jukes–Cantor assumption, (9) can be written as

$$\begin{aligned} S'(t) &= -\beta S(t) \textstyle\sum_y I_y(t), \\ I'_x(t) &= \beta S(t) I_x(t) - \gamma I_x(t) + \Big(\textstyle\sum_y \mu I_y(t)\Big) - 4\mu I_x(t), \\ R'_x(t) &= \gamma I_x(t), \end{aligned} \tag{10}$$

where μ denotes the mutation rate among sequences $x1$, $x2$, $x3$, and $x4$ (i.e., a constant mutation rate among sequences $x1$, $x2$, $x3$, and $x4$ by the Jukes–Cantor assumption). Suppose that only one sequence, $x1$, exists at the beginning of an outbreak, the number of infected individuals is small, and all other individuals are susceptible; then we have (for sure $\epsilon \ll 1$)

$$\begin{aligned} S(0) &= N - \epsilon, \\ I_{x1}(0) &= \epsilon, \\ I_{x2}(0) &= I_{x3}(0) = I_{x4}(0) = R_{x1}(0) = R_{x2}(0) = R_{x3}(0) = R_{x4}(0) = 0. \end{aligned} \tag{11}$$

2.2.2. *L*-site model. We now expand the 1-site model into an L-site model, assuming that the number of mutable sites is L. In this model, 4^L sequences can exist. We denote the infected hosts and recovered hosts with the strain carrying sequence $x = \{x_1, x_2, \ldots, x_L\}$ by $I_{x_1,x_2,\ldots,x_L}$ and $R_{x_1,x_2,\ldots,x_L}$. The maximum number of mutations for each time step is assumed to be one for each sequence. Therefore, the number of unmatched sites between parent sequence and child sequence is always one. For the sake of simplicity, we use the Jukes–Cantor assumption for mutation, which is a constant rate for the mutations in a site; however, the mutation rates among different sites can vary, so μ_l denotes the mutation rate among nucleotides at the lth site. Following these assumptions, (9) can be written as

$$\begin{aligned} S'(t) &= -\beta S(t) \textstyle\sum_{y_1} \sum_{y_2} \cdots \sum_{y_L} I_{y_1,y_2,\dots,y_L}(t), \\ I'_{x_1,x_2,\dots,x_L}(t) &= (\beta S(t) - \gamma) I_{x_1,x_2,\dots,x_L}(t) + \Big(\textstyle\sum_{y_1=1}^{4} \mu_1 I_{y_1,y_2,\dots,y_L}(t) \\ &\quad + \textstyle\sum_{y_2=1}^{4} \mu_2 I_{y_1,y_2,\dots,y_L}(t) + \cdots + \sum_{y_L=1}^{4} \mu_L I_{y_1,y_2,\dots,y_L}(t) \Big) \\ &\quad -4 \left(\textstyle\sum_{l=1}^{L} \mu_l \right) I_{x_1,x_2,\dots,x_L}(t), \\ R'_{x_1,x_2,\dots,x_L}(t) &= \gamma I_{x_1,x_2,\dots,x_L}(t). \end{aligned} \tag{12}$$

Suppose that only one sequence, $\{1, 1, \dots, 1\}$, exists at the beginning of an outbreak, the number of infected individuals is small, and all other individuals are initially susceptible. Then

$$\begin{aligned} S(0) &= N - \epsilon, \\ I_{1,1,\dots,1}(0) &= \epsilon, \\ I_{x_1,x_2,\dots,x_L}(0) &= 0 \ \text{ for } \{x_1, x_2, \dots, x_L\} \neq \{1, 1, \dots, 1\}, \\ R_{x_1,x_2,\dots,x_L}(0) &= 0 \ \text{ for all } \{x_1, x_2, \dots, x_L\}. \end{aligned} \tag{13}$$

3. Main results.

3.1. Deterministic model. We start the analysis of the genetic variation of pathogens from the above deterministic SIR model. For a given site $l_1 \in \{1, \dots, L\}$ and a given nucleotide $k_1 \in \{1, 2, 3, 4\}$, we define

$$G_{l_1}^{k_1} = \{(x_1, \dots, x_L)\,;\ x_i \in \{1, 2, 3, 4\} \ \text{for } 1 \leq i \leq L, x_{l_1} = k_1\}$$

to be the set of all sequences with the nucleotide k_1 in the l_i site. Let $I_{l_1}^{k_1}(t)$ be the number of infectives whose sequences belong to $G_{l_1}^{k_1}$, and let $f_{l_1}^{k_1}(t) = I_{l_1}^{k_1}(t)/I(t)$ be the frequency of nucleotide k_1 in the l_1 site. Here $I(t)$ denotes the total number of infected individuals among all sequences:

$$I(t) = \sum_{k_1=1}^{4} I_{l_1}^{k_1}(t).$$

We start with the 1-site model as shown in (10). In the 1-site model, four different sequences can exist: $x1$, $x2$, $x3$, and $x4$. At only one site can a nucleotide mutate, and nucleotides in other sites are identical among sequences. For the purpose of induction, we introduce an inductive notation,

$$I_{x1} = I_{l_1}^{1_1}, I_{x2} = I_{l_1}^{2_1}, I_{x3} = I_{l_1}^{3_1}, I_{x4} = I_{l_1}^{4_1}.$$

From (10), we get

$$\left(I_{l_1}^{1_1}\right)' - \left(I_{l_1}^{2_1}\right)' = (\beta S - \gamma - 4\mu) \left(I_{l_1}^{1_1} - I_{l_1}^{2_1}\right)$$

and

$$I_{l_1}^{1_1}(t) - I_{l_1}^{2_1}(t) = \left(I_{l_1}^{1_1}(0) - I_{l_1}^{2_1}(0)\right) \left(\exp\left[\int_{\tau=0}^{t} \beta S - \gamma - 4\mu d\tau \right] \right). \tag{14}$$

It follows immediately that

$$I_{l_1}^{1_1}(t) - I_{l_1}^{2_1}(t) = I(0) \left(\exp\left[\int_{\tau=0}^{t} \beta S - \gamma - 4\mu d\tau \right] \right). \tag{15}$$

Substituting (11) and (15) into (14), we have

$$I_{l_1}^{1_1}(t) - I_{l_1}^{2_1}(t) = I(t)\exp\left[-4\mu t\right].$$

Recalling $f_{l_1}^{k_1}(t) = I_{l_1}^{k_1}(t)/I(t)$, we obtain the site-specific nucleotide frequency, $f_{l_1}^{k_1}(t)$ as follows:

$$\begin{aligned} &f_{l_1}^{1_1}(t) = \tfrac{I_{l_1}^{1_1}(t)}{I(t)} = \tfrac{1+3\exp[-4\mu t]}{4},\\ &f_{l_1}^{2_1}(t) = f_{l_1}^{3_1}(t) = f_{l_1}^{4_1}(t) = \tfrac{I_{l_1}^{2_1}(t)}{I(t)} = \tfrac{I_{l_1}^{3_1}(t)}{I(t)} = \tfrac{I_{l_1}^{4_1}(t)}{I(t)} = \tfrac{1-\exp[-4\mu t]}{4}. \end{aligned} \tag{16}$$

Therefore, $f_{l_1}^{k_1}(t)$ is completely determined by the mutation rate μ.

For a more general L-site model, from (12) the dynamics of $I_{l_1}^{k_1}$ can be written as

$$\begin{aligned} \tfrac{d}{dt}I_{l_1}^{k_1}(t) &= \beta S(t)I_{l_1}^{k_1}(t) - \gamma I_{l_1}^{k_1}(t) + \left(\textstyle\sum_{m\neq l_1} 4\mu_m\right) I_{l_1}^{k_1}(t)\\ &\quad +\mu_{l_1}I(t) - \left(\textstyle\sum_m 4\mu_m\right) I_{l_1}^{k_1}(t)\\ &= \left(\beta S(t) - \gamma - 4\mu_{l_1}\right) I_{l_1}^{k_1}(t) + \mu_{l_1}I(t). \end{aligned} \tag{17}$$

Here the term $\left(\sum_m 4\mu_m\right) I_{l_1}^{k_1}(t)$ describes the new strains from sequences $\in G_{l_1}^{k_1}$, including the mutations from $G_{l_1}^{k_1}$ to $G_{l_1}^{k_1}$. The term $(\sum_{m\neq l_1} 4\mu_m) I_{l_1}^{k_1}(t)$ describes the new strain from $G_{l_1}^{k_1}$ to $G_{l_1}^{k_1}$, which should result from the mutation at sites other than the l_1 site. The term $\mu_{l_1}I(t)$ describes the new strains, which become elements of $G_{l_1}^{k_1}$ by the mutation at the l_1th site, including the mutation from $G_{l_1}^{k_1}$ to $G_{l_1}^{k_1}$. Therefore,

$$\begin{aligned} \tfrac{d}{dt}f_{l_1}^{k_1}(t) &= I(t)^{-2}\left[\left\{\left(\beta S(t) - \gamma - 4\mu_{l_1}\right) I_{l_1}^{k_1}(t) + \mu_{l_1}I(t)\right\} I(t)\right.\\ &\qquad \left. - \left\{\beta S(t) - \gamma\right\} I(t)I_{l_1}^{k_1}(t)\right]\\ &= -4\mu_{l_1}f_{l_1}^{k_1}(t) + \mu_{l_1}, \end{aligned} \tag{18}$$

from which we conclude that $f_{l_1}^{k_1}(t)$ is completely determined by the initial condition $f_{l_1}^{k_1}(0)$ and the mutation rate μ_{l_1}. Assuming that the sequence existing at the beginning of an outbreak has nucleotide 1 at the lth site as shown in (13), we obtain

$$\begin{aligned} &f_{l_1}^{1_1}(t) = \tfrac{1+3\exp\left[-4\mu_{l_1}t\right]}{4},\\ &f_{l_1}^{2_1}(t) = f_{l_1}^{3_1}(t) = f_{l_1}^{4_1}(t) = \tfrac{1-\exp\left[-4\mu_{l_1}t\right]}{4}. \end{aligned} \tag{19}$$

Recalling that Tajima's D is determined only by $f_{l_1}^{k_1}$ and n, we have the following.

LEMMA (INDUCTION LEMMA ON TAJIMA'S D). *Tajima's D is completely determined by the sample size n and site-specific mutation rate μ_{l_1}.*

In population genetics, the sign of Tajima's D is believed to be affected by the population dynamics [5]. However, we have observed here that from the deterministic disease dynamics, Tajima's D is independent of the disease dynamics (precisely, independent of the parameters characterizing the disease dynamics, i.e., β and γ). This is true even if the assumption of mutation follows Kimura's assumption [12]; see subsection 5.1. Figure 2 illustrates the dependence of the sign of Tajima's D with mutation rate μ and sample size n. Tajima's D tends to be negative at the beginning of an outbreak and becomes positive as time passes, except when n is small. Tajima's

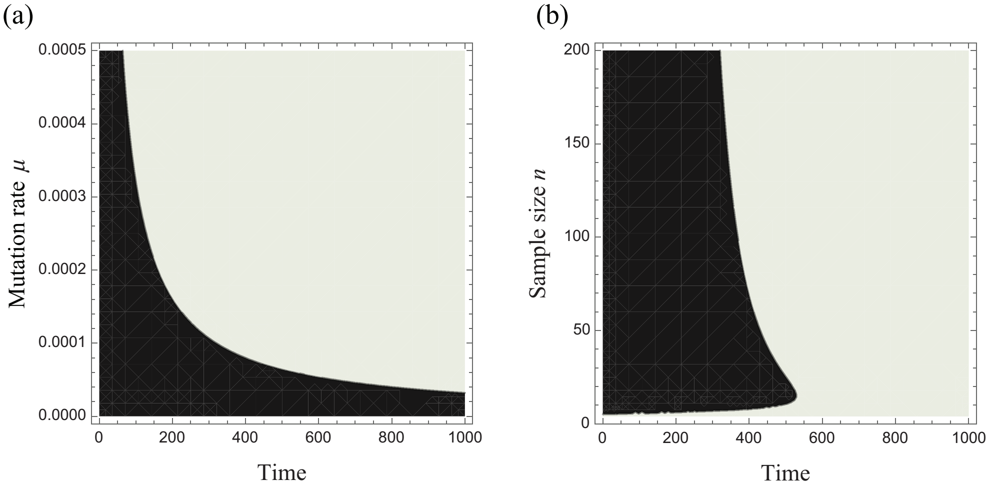


FIG. 2. *The dependence of the sign of Tajima's D with mutation rate μ* (a) *and sample size n* (b) *when the number of mutable sites $L = 500$ and the mutation rate is constant among sites. The black area denotes negative Tajima's D and the gray area denotes positive Tajima's D.*

D can be positive from the beginning to the end of an outbreak when n is small. The time point at which Tajima's D $= 0$ becomes earlier with increasing μ. Tajima's D becomes 0 at the latest time point when the sample size n is intermediate.

We continue to define G, I, and f by fixing two nucleotides k_1 and k_2 and the two different sites l_1 and l_2. That is,

$$\begin{aligned}
&G_{l_1,l_2}^{k_1,k_2} = \{(x_1,\ldots,x_L)\,;\, x_i \in \{1,2,3,4\} \text{ for } 1 \le i \le L, x_{l_1} = k_1, x_{l_2} = k_2\}\,,\\
&I_{l_1,l_2}^{k_1,k_2}(t) = \text{the number of infectives at time } t \text{ with the sequence belonging to } G_{l_1,l_2}^{k_1,k_2},\\
&f_{l_1,l_2}^{k_1,k_2}(t) = \tfrac{I_{l_1,l_2}^{k_1,k_2}(t)}{I(t)}.
\end{aligned}$$

Then we have

$$\begin{aligned}
\tfrac{d}{dt}\left[I_{l_1,l_2}^{k_1,k_2}\right] &= (\beta S - \gamma)\, I_{l_1,l_2}^{k_1,k_2} - 3\mu_{l_1} I_{l_1,l_2}^{k_1,k_2} - 3\mu_{l_2} I_{l_1,l_2}^{k_1,k_2}\\
&\quad + \mu_{l_1} \textstyle\sum_{j=1, j\neq k_2}^{4} I_{l_1,l_2}^{j,k_2} + \mu_{l_2} \textstyle\sum_{j=1, j\neq k_1}^{4} I_{l_1,l_2}^{k_1,j}\\
&= (\beta S - \gamma)\, I_{l_1,l_2}^{k_1,k_2} - 4\mu_{l_1} I_{l_1,l_2}^{k_1,k_2} - 4\mu_{l_2} I_{l_1,l_2}^{k_1,k_2} + \mu_{l_1} I_{l_2}^{k_2} + \mu_{l_2} I_{l_1}^{k_1},
\end{aligned}$$

where the term $\mu_{l_1} \sum_{j=1, j\neq k_2}^{4} I_{l_1,l_2}^{j,k_2}$ describes holding k_2 at the l_2 site with a mutation in the l_1 site, and the term $\mu_{l_2} \sum_{j=1, j\neq k_1}^{4} I_{l_1,l_2}^{k_1,j}$ describes holding k_1 at the l_1 site with mutation in the l_2 site, and

$$\frac{d}{dt}\left[f_{l_1,l_2}^{k_1,k_2}\right] = -4\,(\mu_{l_1} + \mu_{l_2})\, f_{l_1,l_2}^{k_1,k_2} + \mu_{l_1} f_{l_2}^{k_2} + \mu_{l_2} f_{l_1}^{k_1}. \tag{20}$$

Therefore, since we have proved in the induction lemma that $f_{l_i}^{k_i}(t)$ $(i = 1,2)$ are determined from $f_{l_i}^{k_i}(0)$ $(i = 1,2)$ and μ_{l_i} $(i = 1,2)$, we conclude that $f_{l_1,l_2}^{k_1,k_2}(t)$ is determined by μ_{l_1}, μ_{l_2}, and the initial frequencies $f_{l_1}^{k_1}(0)$, $f_{l_2}^{k_2}(0)$, and $f_{l_1,l_2}^{k_1,k_2}(0)$.

Inductively, for a fixed $m \in \{1, \ldots, L-1, L\}$, for fixed $l_1, \ldots, l_m \in \{1, \ldots, L-1\}$ with $l_i \neq l_j$ $(1 \leq i, j \leq m)$, and for fixed nucleotide, $k_1, \ldots, k_m \in \{1, 2, 3, 4\}$, we define

$$\begin{aligned} G^{k_1,\ldots,k_m}_{l_1,\ldots,l_m} &= \left\{(x_1, \ldots, x_L)\,;\, x_i \in \{1,2,3,4\}\,, 1 \leq i \leq L, x_{l_j} = k_j, 1 \leq j \leq m\right\}, \\ I^{k_1,\ldots,k_m}_{l_1,\ldots,l_m}(t) &= \text{the number of infectives with the sequence} \in G^{k_1,\ldots,k_m}_{l_1,\ldots,l_m}, \\ f^{k_1,\ldots,k_m}_{l_1,\ldots,l_m}(t) &= \tfrac{I^{k_1,\ldots,k_m}_{l_1,\ldots,l_m}(t)}{I(t)}. \end{aligned}$$

Then

$$\begin{aligned} &\tfrac{d}{dt}\left[I^{k_1,\ldots,k_m,k_{m+1}}_{l_1,\ldots,l_m,l_{m+1}}\right] \\ &\quad= (\beta S - \gamma)\, I^{k_1,\ldots,k_m,k_{m+1}}_{l_1,\ldots,l_m,l_{m+1}} - (3\mu_{l_1} + \cdots + 3\mu_{l_m} + 3\mu_{l_m+1})\, I^{k_1,\ldots,k_m,k_{m+1}}_{l_1,\ldots,l_m,l_{m+1}} \\ &\qquad + \mu_{l_1} \textstyle\sum_{j=1, j\neq k_1}^{4} I^{j,k_2,\ldots,k_m,k_{m+1}}_{l_1,\ldots,l_m,l_{m+1}} + \cdots + \mu_{l_{m+1}} \textstyle\sum_{j=1, j\neq k_{m+1}}^{4} I^{k_1,\ldots,k_m,j}_{l_1,\ldots,l_m,l_{m+1}} \\ &\quad= (\beta S - \gamma)\, I^{k_1,\ldots,k_m,k_{m+1}}_{l_1,\ldots,l_m,l_{m+1}} - 4\left(\mu_{l_1} + \cdots + \mu_{l_{m+1}}\right) I^{k_1,\ldots,k_m,k_{m+1}}_{l_1,\ldots,l_m,l_{m+1}} \\ &\qquad + \mu_{l_1} I^{k_2,\ldots,k_m,k_{m+1}}_{l_2,\ldots,l_m,l_{m+1}} + \mu_{l_2} I^{k_1,k_3,k_4,\ldots,k_m,k_{m+1}}_{l_1,l_3,l_4,\ldots,l_m,l_{m+1}} + \cdots + \mu_{l_{m+1}} I^{k_1,\ldots,k_m}_{l_1,\ldots,l_m}, \end{aligned}$$

from which we conclude that

$$\begin{aligned} \tfrac{d}{dt}\left[f^{k_1,\ldots,k_m,k_{m+1}}_{l_1,\cdots,l_m,l_{m+1}}(t)\right] = -4\left(\mu_{l_1} + \cdots + \mu_{l_m+1}\right) f^{k_1,\ldots,k_m,k_{m+1}}_{l_1,\ldots,l_m,l_{m+1}} \\ + \mu_{l_1} f^{k_2,\ldots,k_{m+1}}_{l_2,\ldots,l_{m+1}} + \cdots + \mu_{l_{m+1}} f^{k_1,\ldots,k_m}_{l_1,\ldots,l_m}. \end{aligned} \tag{21}$$

Thus, $f^{k_1,\ldots,k_m,k_{m+1}}_{l_1,\ldots,l_m,l_{m+1}}(t)$ is determined by $\mu_{l_1} \cdots \mu_{l_m+1}$ and the initial condition of $f^{k_i}_{l_i}(0)$ $(i=1, \ldots, m+1)$, $f^{k_i,k_j}_{l_i,l_j}(0)$ $(i, j=1, \ldots, m+1)$, $\ldots$, $f^{k_1,\ldots,k_m}_{l_1,\ldots,l_m}(0)$, $f^{k_1,\ldots,k_{m+1}}_{l_1,\ldots,l_{m+1}}(0)$. The induction ends when we are able to calculate $f^{k_1,\ldots,k_{m+1}}_{l_1,\ldots,l_{m+1}}(t)$. In summary, we have proved the following.

THEOREM. *The frequencies $f^{k_1,\ldots,k_L}_{l_1,\ldots,l_L}$ with $k_1, \ldots, k_L \in \{1,2,3,4\}$ and $l_1, \ldots, l_L \in \{1, 2, \ldots, L\}$ can be calculated inductively from* (18)*,* (20)*, and* (21)*, and these frequencies are independent of the disease population dynamics* $(\{\beta, \gamma\})$ *and are completely determined by the site-specific mutation rates and initial frequency values.*

The frequencies $f^{k_1,\ldots,k_L}_{l_1,\ldots,l_L}$ describe the genetic diversity of the pathogen. Thus, we observe from the deterministic model that the genetic diversity of the pathogen under neutral evolution (where mutation has no effect on the pathogens' phenotypes, i.e., β and γ) is independent of the disease dynamics.

3.2. Stochastic model. What happens if we expand the deterministic 1-site model into a stochastic 1-site model following a continuous time Markov process? From (21), the time differentiation of the expected number of infected hosts with the strain carrying the sequence whose nucleotide k_1 at the specific site l_1, $\mathrm{E}(I^{k_1}_{l_1})$, and the expected number of the total number of infected hosts, $\mathrm{E}\,(I)$, can be written as

$$\begin{aligned} \mathrm{E}\left(I^{k_1}_{l_1}(t)\right)' &= (\beta \mathrm{E}\,(S(t)) - \gamma - 4\mu)\, \mathrm{E}\left(I^{k_1}_{l_1}(t)\right) + \mu \mathrm{E}\,(I(t)) \\ &\quad + \beta \mathrm{Cov}\left(S(t), I^{k_1}_{l_1}(t)\right), \\ \mathrm{E}\,(I(t))' &= (\beta \mathrm{E}\,(S(t)) - \gamma)\, \mathrm{E}\,(I(t)) + \beta \mathrm{Cov}\,(S(t), I(t))\,. \end{aligned} \tag{22}$$

With manipulations similar to those for the deterministic 1-site model, we obtain

$$
\begin{aligned}
&\mathrm{E}\left(f_{l_1}^{1_1}(t)\right) = \mathrm{E}\left(\frac{I_{l_1}^{1_1}(t)}{I(t)}\right) = \tfrac{1}{4}\left(1+\exp\left[-4\mu t\right]\textstyle\sum_{k_1\neq 1} g_{l_1}^{k_1}(t)\right),\\
&\mathrm{E}\left(f_{l_1}^{2_1}(t)\right) = \tfrac{1}{4}\left(1+\exp\left[-4\mu t\right]\textstyle\sum_{k_1\neq 1} g_{l_1}^{k_1}(t) - 4\exp\left[-4\mu t\right] g_{l_1}^{2_1}(t)\right),\\
&\mathrm{E}\left(f_{l_1}^{3_1}(t)\right) = \tfrac{1}{4}\left(1+\exp\left[-4\mu t\right]\textstyle\sum_{k_1\neq 1} g_{l_1}^{k_1}(t) - 4\exp\left[-4\mu t\right] g_{l_1}^{3_1}(t)\right),\\
&\mathrm{E}\left(f_{l_1}^{4_1}(t)\right) = \tfrac{1}{4}\left(1+\exp\left[-4\mu t\right]\textstyle\sum_{k_1\neq 1} g_{l_1}^{k_1}(t) - 4\exp\left[-4\mu t\right] g_{l_1}^{4_1}(t)\right),
\end{aligned}
\tag{23}
$$

where $\mathrm{Cov}\,(x,y)$ denotes covariance of x and y and

$$
\begin{aligned}
g_{l_1}^{k_1}(t) &= \exp\left(-\beta\int_{\tau=0}^{t}\frac{\mathrm{Cov}(S(\tau),I(\tau))}{\mathrm{E}(I(\tau))} - \frac{\mathrm{Cov}\left(S(\tau),I_{l_1}^{1_1}(\tau)\right)-\mathrm{Cov}\left(S(\tau),I_{l_1}^{k_1}(\tau)\right)}{\mathrm{E}\left(I_{l_1}^{1_1}(\tau)\right)-\mathrm{E}\left(I_{l_1}^{k_1}(\tau)\right)}d\tau\right)\\
&= \exp\left(-\beta\int_{\tau=0}^{t}\frac{\mathrm{E}(S(\tau)I(\tau))}{\mathrm{E}(I(\tau))} - \frac{\mathrm{E}\left(S(\tau)\left(I_{l_1}^{1_1}(\tau)-I_{l_1}^{k_1}(\tau)\right)\right)}{\mathrm{E}\left(I_{l_1}^{1_1}(\tau)\right)-\mathrm{E}\left(I_{l_1}^{k_1}(\tau)\right)}d\tau\right).
\end{aligned}
$$

Unlike the deterministic model, the expected value of the site-specific nucleotide frequency in the stochastic model can be affected by disease dynamics. This is also true when we expand the stochastic 1-site model into a stochastic L-site model with varying mutation rates among sites as follows:

$$
\begin{aligned}
\mathrm{E}\left(I_{l_1}^{k_1}(t)\right)' &= \left(\beta\mathrm{E}\left(S(t)\right)-\gamma-4\mu_{l_1}\right)\mathrm{E}\left(I_{l_1}^{k_1}(t)\right)+\mu\mathrm{E}\left(I(t)\right)\\
&\quad+\beta\mathrm{Cov}\left(S(t),I_{l_1}^{k_1}(t)\right),\\
\mathrm{E}\left(I(t)\right)' &= \left(\beta\mathrm{E}\left(S(t)\right)-\gamma\right)\mathrm{E}\left(I(t)\right)+\beta\mathrm{Cov}\left(S(t),I(t)\right),
\end{aligned}
\tag{24}
$$

from which it follows that

$$
\begin{aligned}
&\mathrm{E}\left(f_{l_1}^{1_1}(t)\right) = \mathrm{E}\left(\frac{I_{l_1}^{1_1}(t)}{I(t)}\right) = \tfrac{1}{4}\left(1+\exp\left[-4\mu_{l_1}t\right]\textstyle\sum_{k_1\neq 1} g_{l_1}^{k_1}(t)\right),\\
&\mathrm{E}\left(f_{l_1}^{2_1}(t)\right) = \tfrac{1}{4}\left(1+\exp\left[-4\mu_{l_1}t\right]\textstyle\sum_{k_1\neq 1} g_{l_1}^{k_1}(t) - 4\exp\left[-4\mu_{l_1}t\right] g_{l_1}^{2_1}(t)\right),\\
&\mathrm{E}\left(f_{l_1}^{3_1}(t)\right) = \tfrac{1}{4}\left(1+\exp\left[-4\mu_{l_1}t\right]\textstyle\sum_{k_1\neq 1} g_{l_1}^{k_1}(t) - 4\exp\left[-4\mu_{l_1}t\right] g_{l_1}^{3_1}(t)\right),\\
&\mathrm{E}\left(f_{l_1}^{4_1}(t)\right) = \tfrac{1}{4}\left(1+\exp\left[-4\mu_{l_1}t\right]\textstyle\sum_{k_1\neq 1} g_{l_1}^{k_1}(t) - 4\exp\left[-4\mu_{l_1}t\right] g_{l_1}^{4_1}(t)\right),
\end{aligned}
\tag{25}
$$

where

$$
g_{l_1}^{k_1}(t) = \exp\left[-\beta\int_{\tau=0}^{t}\frac{\mathrm{E}\left(S(\tau)I(\tau)\right)}{\mathrm{E}\left(I(\tau)\right)} - \frac{\mathrm{E}\left(S(\tau)\left(I_{l_1}^{1_1}(\tau)-I_{l_1}^{k_1}(\tau)\right)\right)}{\mathrm{E}\left(I_{l_1}^{1_1}(\tau)\right)-\mathrm{E}\left(I_{l_1}^{k_1}(\tau)\right)}d\tau\right].
$$

3.2.1. Monte Carlo simulations. We now show that the impact of disease dynamics on Tajima's D is hidden in the stochasticity of the disease dynamics. The term in Tajima's D that includes the effect of stochasticity of disease dynamics, as shown in (24), cannot be expressed in a closed form [9]. To explore the behavior of Tajima's D in a stochastic model, we employ an individual-based Monte Carlo (IBM) simulation. The infection state of each host is recorded as 0, 1, or 2 in the IBM simulation. The infection states 0, 1, and 2 indicate that the host is susceptible (S), currently infected (I), or recovered (R), respectively, assuming each infected

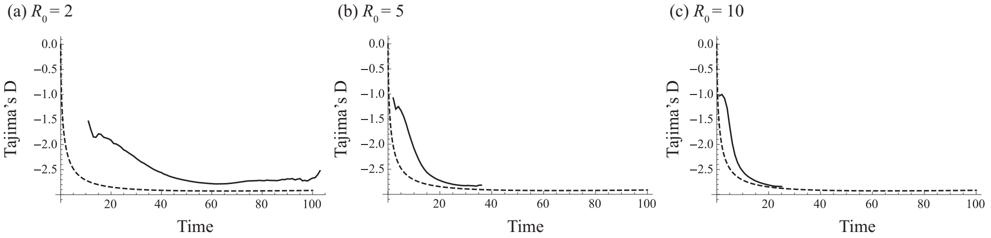


FIG. 3. *The impact of disease dynamics on the temporal evolution of Tajima's D with varied basic reproduction number $R_0 = \beta/\gamma$. The solid lines show the* 1000*-iteration average of the time evolution of Tajima's D when $R_0 = 2.0$* (a)*,* 5.0 (b)*, and* 10.0 (c)*, respectively. At each time step of the IBM simulation, Tajima's D was calculated when $I(t) \geq 200$. The dashed lines show Tajima's D in the deterministic model.*

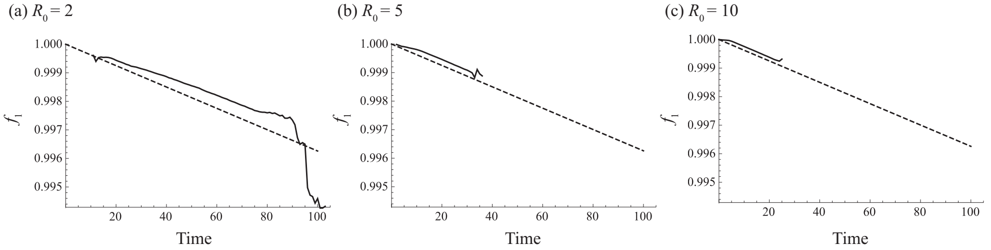


FIG. 4. *The impact of disease dynamics on the time evolution of f_1 with varied basic reproduction number R_0. The solid lines show the* 1000*-iteration average of the time evolution of f_1 when $R_0 = 2.0$* (a)*,* 5.0 (b)*, and* 10.0 (c)*, respectively. At each time step of the IBM simulation, f_1 was calculated when $I(t) \geq 200$. The dashed lines show f_1 in the deterministic model, which is given by* (19)*.*

individual has only one nucleic acid sequence of the pathogen, and each infected individual and nucleic acid sequence is recorded. Secondary cases have the same sequence as the primary case when transmission occurs. During infection the mutation replaces the dominant sequence of the pathogen by a constant rate μ.

We set the parameters to their baseline values as follows: $N = 200{,}000$, $L = 500$, $\gamma = 0.3$, $\Delta t = 1$, $n = 200$. We assume initially that one host is infected with a strain carrying a sequence whose nucleotides at all sites are 1, and all other hosts are susceptible for all possible strains. The IBM simulations run until there are no infected individuals. Figure 3 shows the impact of disease dynamics on the temporal evolution of Tajima's D with varied basic reproduction number $R_0 = \beta/\gamma$. The difference between the deterministic model and the stochastic model increases as R_0 decreases. The duration of epidemics in the IBM simulation also increases when R_0 decreases.

As discussed in the previous sections, Tajima's D is a function of the site-specific nucleotide frequency $f_{l_1}^{k_1}$. The frequency of nucleotide 1, which is identical in all sequences at the beginning of an outbreak, $f_{l_1}^{1_1}$, is calculated. As we assume the site-specific mutation rate is identical among different sites, the expected value of $f_{l_1}^{k_1}$ is also identical among different nucleotides and different sites. We can then calculate the average of $f_{l_1}^{1_1}$ among all sites, f_1. Figure 4 shows the impact of disease dynamics on the time evolution of f_1 with varied basic reproduction number R_0. The deterministic model predicts f_1 in the corresponding stochastic model, and the impact

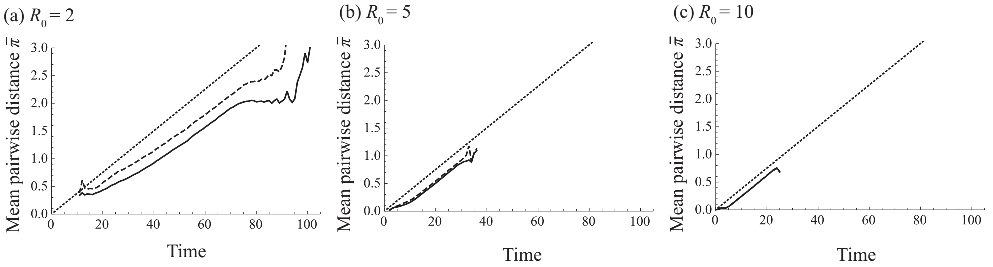


Fig. 5. *The time evolution of* $\overline{\pi}$ *with varied basic reproduction number* R_0*. The solid lines show the* 1000*-iteration average of the time evolution of* $\overline{\pi}$*. At each time step of the IBM simulation,* $\overline{\pi}$ *was calculated when* $I(t) \geq 200$*. The dotted lines show the theoretical* $\overline{\pi}$*, which is given by* (6) *using deterministic* f_1*,* f_2*,* f_3*, and* f_4 *given by* (19)*. The dashed lines show the theoretical* $\overline{\pi}$*, which is given by* (6) *using the* 1000 *iteration average of the time series of* f_1*,* f_2*,* f_3*, and* f_4 *with IBM.*

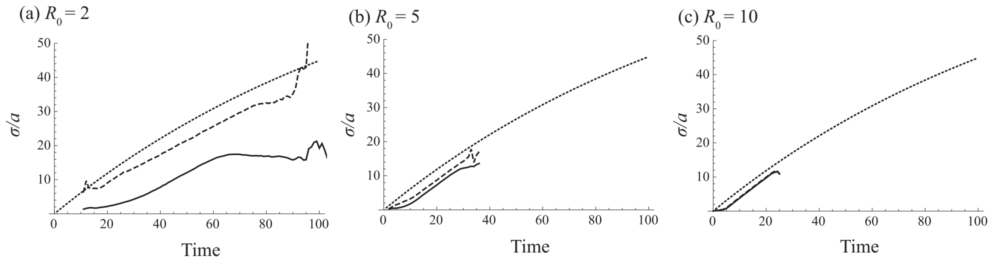


Fig. 6. *The time evolution of* σ/a *with varied basic reproduction number* R_0*.* σ *denotes the number of segregating site, and a denotes* $\sum_{i=1}^{n-1} 1/i$*. The solid lines show the* 1000*-iteration average of the time evolution of* σ *when* $R_0 = 2.0$ (a)*,* 5.0 (b)*, and* 10.0 (c)*. At each time step of the IBM simulation,* σ/a *was calculated when* $I(t) \geq 200$*. The dotted lines show the theoretical* σ*, which is given by* (8) *using the deterministic* f_1*,* f_2*,* f_3*, and* f_4 *given by* (19)*. The dashed lines show the theoretical* σ*, which is given by* (8) *using the* 1000 *iteration average of the time series of* f_1*,* f_2*,* f_3*, and* f_4 *with IBM.*

on R_0 is small for f_1.

As shown in (1), Tajima's D can be decomposed into two functions of f_1, the mean pairwise distance between sampled sequences $\overline{\pi}$ and the number of segregating sites among sampled sequences σ. Specifically, the relationship between $\overline{\pi}$ and σ/a determines the sign of Tajima's D. Therefore, we explore the difference between $\overline{\pi}$ and σ/a for the deterministic model and the stochastic model. Figure 5 shows the difference in the time evolution of $\overline{\pi}$ with varied basic reproduction number R_0. $\overline{\pi}$ is the function of f_1, f_2, f_3, and f_4, and there is a gap between theoretical and simulated $\overline{\pi}$ even if f_1, f_2, f_3, and f_4 are all fixed. This gap also depends on R_0: the smaller R_0, the larger the gap.

Figure 6 shows the difference in the time evolution of σ with the varied basic reproduction number R_0. This gap also depends on R_0: again, the smaller R_0, the larger the gap. The gap between theoretical and simulated Tajima's D is large at the early stage of an epidemic. Comparing $\overline{\pi}$ and σ/a, σ/a shows a large gap between the deterministic and stochastic models.

4. Discussion. In this paper we investigated the time evolution of Tajima's D in a nonendemic disease outbreak situation, and we conclude that (i) Tajima's D in a deterministic SIR model is completely determined by the mutation rate and sample size, and (ii) the time evolution of genetic diversity of an infectious disease pathogen in a deterministic SIR model is completely determined by the mutation rate. We observed that Tajima's D is independent of disease dynamics, and we found that the sign interpretation of Tajima's D (i.e., negative (positive) Tajima's D shows population expansion (contraction)) is not always true for disease transmission dynamics. Employing a stochastic model, we demonstrated that the dependence of Tajima's D on the disease transmission and evolution dynamics is a result of the stochasticity of those dynamics. The same is true for the dependence related to genetic diversity. Interestingly, the period when Tajima's D remains negative is longer than the period when the infected population increases, as shown in Figure 3.

Our observation that Tajima's D is dependent on stochasticity in disease transmission during an outbreak can be useful for the estimation of insightful epidemiological and evolutional parameters. If the sequence data reflect stochastic transmission and evolution dynamics, e.g., sequences of pathogens sampled from a small outbreak in a limited host population, then Tajima's D depends on both the mutation rate and R_0. Therefore, a joint estimation of the mutation rate and R_0 from Tajima's D is possible. If the disease dynamics can be approximated by a deterministic SIR model, then Tajima's D is not biased by disease dynamics and can be determined by the mutation rate alone.

Tajima's D depends largely on the sample size, as shown in Figure 2. In particular, with a small sample size, small changes in the sample size causes significant bias in Tajima's D. At a practical level, the sampling probability during an outbreak changes over time due to reporting bias, so careful study is required.

If we can assume that the sampling probability, p, is constant over time, and that the sample size is proportional to the number of infected individuals, $n = pI(t)$, then Tajima's D depends on the disease dynamics even in the deterministic model. Even in such a case, the sign of Tajima's D is not directly related to the sign of the time-derivative of the pathogen population $I'(t)$, as seen in the study with monotonic population growth [18, 20]. Tajima's D is determined not only by R_0, but also by the mutation rate and sampling probability. In this case we can estimate R_0 from deterministic Tajima's D. However, this estimate of R_0 is equivalent to the estimate of R_0 fitting an SIR model, with the time series of the sample size as the time series of the number of infective individuals.

Previous studies show that a sudden change in the sign of Tajima's D implies the replacement of the dominant strain under strong natural selection [6]. If R_0 and the mutation rate are given, the confidence interval of the site-specific Tajima's D, D_l, can be obtained from our stochastic model. This confidence interval gives the confidence interval for neutral mutation. Using this confidence interval, we can discover the significant mutations in terms of their impact on phenotypes. The estimation of R_0 using epidemiological data, e.g., the time series of the number of infected hosts [2] and the estimation of the mutation rate using sequence data [4] have already been developed, and the estimation of significant mutations is possible if both epidemiological and sequence data are available.

Sampling time series of sequences of pathogens during an outbreak has become a common practice for infectious disease surveillance. Methodologies analyzing evolution and disease dynamics from such sequence data are in great demand. Many sequence datasets of infectious disease are sampled from nonendemic situations, as described by the SIR model. In the SIR model, the fitness of a pathogen and effective reproduction number are changing over time, even if there is no mutation. In such nonequilibrium population dynamics, the interpretation of the existing methods for evolutionary analysis may be different from that in simple population dynamics, e.g., neutral evolution or exponential growth. Theoretical analysis of the impact of population dynamics on common evolutionary analysis can help interpret results of such analysis.

5. Appendix.

5.1. The disease dynamics with Kimura's assumption.

If we follow the mutation model of Kimura's assumption [12], (10) can be written as

$$\begin{aligned} S' &= -S\left(\textstyle\sum_x \beta I_x\right), \\ (I_{x1})' &= (\beta S - \gamma - (2\mu_1 + \mu_2))\, I_{x1} + \mu_1 I_{x2} + \mu_1 I_{x3} + \mu_2 I_{x4}, \\ (I_{x2})' &= (\beta S - \gamma - (2\mu_1 + \mu_2))\, I_{x2} + \mu_1 I_{x1} + \mu_2 I_{x3} + \mu_1 I_{x4}, \\ (I_{x3})' &= (\beta S - \gamma - (2\mu_1 + \mu_2))\, I_{x3} + \mu_1 I_{x1} + \mu_2 I_{x2} + \mu_2 I_{x4}, \\ (I_{x4})' &= (\beta S - \gamma - (2\mu_1 + \mu_2))\, I_{x4} + \mu_2 I_{x1} + \mu_1 I_{x2} + \mu_1 I_{x3}. \end{aligned} \tag{26}$$

The initial condition of the system shown in (26) is the same as that of (11). From (26),

$$(I_{x1})' - (I_{x4})' = (\beta S - \gamma - 2(\mu_1 + \mu_2))\,(I_{x1} - I_{x4}).$$

By a computation similar to that used in the 1-site model with the Jukes–Cantor assumption, we obtain

$$I_{x1} - I_{x4} = I(t)\exp(-2(\mu_1 + \mu_2)\,t).$$

We also compute the time derivative of $I_{x1}(t) - I_{x2}(t)$ using $I_{x2}(t) = I_{x3}(t)$,

$$(I_{x1})' - (I_{x2})' = (\beta S - \gamma - 4(\mu_1 - \mu_2))\,(I_{x1} - I_{x2}) - (\mu_1 - \mu_2)\,(I_{x1} - I_{x4}).$$

Therefore, we obtain the site-specific nucleotide frequency, $f_{l_1}^{k_1}(t)$, as follows;

$$\begin{aligned} f_{l_1}^{1_1}(t) &= \tfrac{1}{4}\left(1 + \exp(-4\mu_1 t) + 2\exp(-2(\mu_1 + \mu_2)\,t)\right), \\ f_{l_1}^{2_1}(t) &= f_{l_1}^{3_1}(t) = \tfrac{1}{4}\left(1 - \exp(-4\mu_1 t)\right), \\ f_{l_1}^{4_1}(t) &= \tfrac{1}{4}\left(1 + \exp(-4\mu_1 t) - 2\exp(-2(\mu_1 + \mu_2)\,t)\right). \end{aligned}$$

Regarding the L-site model, we can derive $f_{l_1}^{k_1}(t)$ similarly to the L-site model with the Jukes–Cantor assumption as follows;

$$\begin{aligned} f_{l_1}^{1}(t) &= \tfrac{1}{4}\left(1 + \exp(-4\mu_{l_1,1} t) + 2\exp(-2(\mu_{l_1,1} + \mu_{l_1,2})\,t)\right), \\ f_{l_1}^{2_1}(t) &= f_{l_1}^{3_1}(t) = \tfrac{1}{4}\left(1 - \exp(-4\mu_{l_1,1} t)\right), \\ f_{l_1}^{4_1}(t) &= \tfrac{1}{4}\left(1 + \exp(-4\mu_{l_1,1} t) - 2\exp(-2(\mu_{l_1,1} + \mu_{l_1,2})\,t)\right), \end{aligned}$$

where $\mu_{l_1,1}$ ($\mu_{l_1,2}$) denotes the mutation rate μ_1 (μ_2) in the 1-site model at the l_1th site.

Acknowledgments. The authors gratefully acknowledge the manuscript language review of Ms. Heidi L. Tessmer from Hokkaido University. The authors are grateful to the two anonymous referees for their careful reading and helpful suggestions which led to improvement of the original manuscript.

REFERENCES

[1] S. Biswas and J. M. Akey, *Genomic insights into positive selection*, TRENDS in Genetics, 22 (2006), pp. 437–446.

[2] G. Chowell and F. Brauer, *The basic reproduction number of infectious diseases: Computation and estimation using compartmental epidemic models*, in Mathematical and Statistical Estimation Approaches in Epidemiology, Springer, 2009, pp. 1–30.

[3] K. Dietz, *Epidemiologic interference of virus populations*, J. Math. Biol., 8 (1979), pp. 291–300.

[4] A. Drummond, O. G. Oliver, and A. Rambaut, *Inference of viral evolutionary rates from molecular sequences*, Adv. Parasitology, 54 (2003), pp. 331–358.

[5] J. C. Fay and C. Wu, *A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation*, Mol. Biol. Evol., 16 (2003), pp. 1003–1005.

[6] I. Gordo, M. G. M. Gomes, D. G. Reis, and P. R. Campos, *Genetic diversity in the SIR model of pathogen evolution*, PloS One, 4 (2009), e4876.

[7] S. Gupta, N. Ferguson, and R. Anderson, *Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents*, Science, 280 (1998), pp. 912–915.

[8] H. Innan and W. Stephan, *The coalescent in an exponentially growing metapopulation and its application to Arabidopsis thaliana*, Genetics, 155 (2000), pp. 2015–2019.

[9] V. Isham, *Assessing the variability of stochastic epidemics*, Math. Biosci., 107 (1991), pp. 209–224.

[10] T. H. Jukes and C. R. Cantor, *Evolution of protein molecules*, in Mammalian Protein Metabolism, vol. III, Academic Press, 1969, pp. 21–132.

[11] K. Kim, R. Omori, K. Ueno, S. Iida, and K. Ito, *Host-specific and segment-specific evolutionary dynamics of avian and human influenza A viruses: A systematic review*, PloS One, 11 (2016), e0147021.

[12] M. Kimura, *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences*, J. Mol. Evol., 16 (1980), pp. 111–120.

[13] J. F. Kingman, *Origins of the coalescent:* 1974-1982, Genetics, 156 (2000), pp. 1461–1463.

[14] R. Nielsen, *Molecular signatures of natural selection*, Annu. Rev. Genet., 39 (2005), pp. 197–218.

[15] R. Omori and A. Sasaki, *Timing of the emergence of new successful viral strains in seasonal influenza*, J. Theoret. Biol., 329 (2013), pp. 32–38.

[16] O. G. Pybus and A. Rambaut, *Evolutionary analysis of the dynamics of viral infectious disease*, Nature Reviews Genetics, 10 (2009), pp. 540–550.

[17] O. G. Pybus, A. Rambaut, and P. H. Harvey, *An integrated framework for the inference of viral population history from reconstructed genealogies*, Genetics, 155 (2000), pp. 1429–1437.

[18] A. Sano and H. Tachida, *Gene genealogy and properties of test statistics of neutrality under population growth*, Genetics, 169 (2005), pp. 1687–1697.

[19] T. Stadler, D. Kühnert, S. Bonhoeffer, and A. J. Drummond, *Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)*, Proc. Natl. Acad. Sci. USA, 110 (2013), pp. 228–233.

[20] F. Tajima, *The effect of change in population size on DNA polymorphism.*, Genetics, 123 (1989), pp. 597–601.

[21] F. Tajima, *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism*, Genetics, 123 (1989), pp. 585–595.