Population genetics of variants in infectious diseases and its application to the prediction of variant replacement

（感染症における変異株の集団遺伝学と変異株の置き換わりの予測）

CHAYADA PIANTHAM

Table of Contents

List of figures

List of tables

Symbols and notation

Chapter 1

f	Relative frequency of amino acid
a	A statistic of patient age
e	Epitope flag indicating whether the amino acid is located on an antigenic site
C_a	Coefficient for patient age statistic
C_e	Coefficient for epitope flag
C_0	Intercept
$a.diff$	Difference in patient median age
$a.wilcox$	Difference in patient age distribution calculated using z-value of rank-sum
P_{fix}	Fixation probability of a new amino acid at a residue position on HA
N	Effective population size
s	Selective advantage of amino acid
r^2	Squared correlation coefficient
tp	The number of true positive predictions
tn	The number of true negative predictions
fp	The number of false positive predictions
fn	The number of false negative predictions

Chapter 2

q_X	Relative frequency of a variant X
t	Calendar time
t_0	Calendar time which a variant was introduced into the population
k	Relative effective reproduction number
f	Probability density function of generation time
g	Probability mass function of discretized generation time
R_X	Effective reproduction number of variant X
I	Total number of new infections
N	Number of sequences
M	The sum of the two shape parameters of beta distribution
s_p	Calendar time when variant frequency becomes p

Abbreviations

AIC	Akaike Information Criterion
CI	Confidence interval
D	Aspartic acid
E	Glutamic acid
GISAID	Global Initiative on Sharing All Influenza Data
HA	Hemagglutinin
I	Isoleucine
K	Lysine
N	Asparagine
P	Proline
Q	Glutamine
R	Arginine
ROC	Receiver Operating Characteristic
S	Serine
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
T	Threonine
V	Valine
VOC	Variant of concern
w.r.t.	with respect to

Notes

The contents in Chapter 1 have been published in *Virus Evolution.* The contents in Chapter 2 have been published in *Viruses*.

Piantham, C., & Ito, K. (2021). Modeling the selective advantage of new amino acids on the hemagglutinin of H1N1 influenza viruses using their patient age distributions. *Virus Evol*, *7*(1), veab049. doi:10.1093/ve/veab049

Piantham, C., & Ito, K. (2022). Predicting the trajectory of replacements of SARS-CoV-2 variants using relative reproduction numbers. *Viruses*, *14*(11), 2556. doi:10.1093/ve/veab049

Preface

Natural selection has been observed in many viruses circulating in the human population. For example, the antigenicity of H3N2 subtype of influenza A virus circulating in humans has been altering for more than 40 years since the virus caused a pandemic in 1968. The transmissibility of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has increased as the dominant virus changed from its original Wuhan strain to Alpha, Delta, and Omicron variants. When a new variant of a virus is selected by natural selection, phenotype such as antigenicity or transmissibility changes, because natural selection favors variants having phenotypes better adapted to the environment. Evolving viruses require public health sectors to update control measures in order to respond to the altered phenotypes of new variants.

Population genetics is a field of study that focuses on how the genetic compositions within populations change over time. Natural selection is a process where phenotypes of individuals in the population change depending on the difference in individuals' ability to generate offspring. My research aims to use population genetics to study the natural selection acting on variants of viruses. This dissertation develops population genetics models describing the process of variant selection in the viral population. The parameters of the models are estimated using observed data. The estimated parameters are used to explain epidemiological implications and to predict future. The results will serve as additional information for public health sectors to develop control measures against epidemics caused by the selected variants.

This dissertation is divided into two chapters. Chapter 1 describes the fixation of variants having new amino acids on the hemagglutinin (HA) protein of H1N1pdm2009 influenza viruses and how it is linked to patient age distribution. Chapter 2 describes the relative reproduction number of the Delta variant with respect to the Alpha variant of SARS-CoV-2 and how it can predict the time course of the Alpha-to-Delta replacement from early observations.

Influenza viruses alter their antigenicity via amino acid substitutions on antigenic sites on HA, which is the virus's major antigen. Natural selection drives the fixation and extinction of variants having amino acid substitutions on HA. In Chapter 1, associations between the fixation probability of variants having amino acid substitutions on HA of H1N1pdm2009 influenza and the attributes of the variants were investigated. Kimura's formula of fixation probability under advantageous selection was extended to deal with not only variant frequency but also other attributes of viruses. The importance of the distribution of patient ages in predicting adaptive evolution of seasonal influenza viruses was discussed in the chapter.

Variants of SARS-CoV-2 have undergone natural selection since the virus was introduced in the human population in China in 2019. The factor that drives the natural selection is the difference in

reproduction numbers among variants. Chapter 2 describes a method to estimate the difference in reproduction numbers among variants. I used the relative reproduction numbers, which is the ratio between the reproduction number of a variant and that of another variant. The relative reproduction numbers of variants with respect to another variant was estimated using temporal data on variant frequency. The predictability of the trajectory of variant replacement using the relative reproduction numbers estimated from early phases of variant replacement are discussed in the chapter.

Finally, findings from Chapter 1 and Chapter 2 are summarized in the Conclusion of this dissertation. Limitations and future direction of the research are also discussed in the Conclusion.

Chapter 1. Modeling the selective advantage of variants having new amino acids on the hemagglutinin of H1N1 influenza viruses using their patient age distributions

1.1. Summary

In 2009, a new strain of H1N1 influenza A virus caused a pandemic, and its descendant variants are causing seasonal epidemics worldwide. Given the high mutation rate of influenza viruses, variants having different amino acids on HA continuously emerge. To prepare vaccine strains for the next influenza seasons, it is an urgent task to predict which strains will be selected in the viral population. An analysis of 24,681 pairs of amino acid sequences of HA of H1N1 influenza viruses circulating from 2009 to 2020 and their patient age showed that the empirical fixation probability of variants having new amino acids on HA significantly differed depending on their frequencies in the population, patient age distributions, and whether or not it involves the antigenic sites. The selective advantage of a variant having an amino acid substitution on HA was modeled by linear combinations of patient age distributions and its involvement on antigenic sites. The fixation probability of the variants having a new amino acid was modeled using Kimura's formula for advantageous selection. Cross validation tests showed that the fixation and extinction of variants having a new amino acid on HA could be predicted with a sensitivity of 0.78, specificity of 0.86, and precision of 0.83 once the relative frequency of the amino acid exceeded 0.11. Estimated parameters showed that the fixation probability increased when variants having a new amino acid could infect patients in higher age groups better than those having the old amino acid. This result suggested that variants of H1N1 influenza viruses tend to be selected in the adult population and that patient ages of variants are useful for predicting fixation and extinction of variants of H1N1 influenza viruses.

1.2. Introduction

One billion seasonal influenza cases, with three to five million severe cases, and around 409 thousand influenza-related deaths are being reported annually (World Health Organization, 2019). The seasonal influenza is caused by two subtypes of influenza A viruses, H1N1 and H3N2, and two lineages of influenza B viruses circulating in the human population. Quadrivalent vaccines, containing two strains from type A viruses and two strains from type B, or trivalent vaccines, containing two from type A viruses and one type B strain, are used to reduce the risk of severe symptoms caused by seasonal influenza (World Health Organization, 2020).

The HA protein, the major antigen of influenza A viruses, undergoes adaptive evolution that alters their antigenicity in human population. This antigenic evolution is caused by a process where human immunity selects strains that are antigenically different from strains that have been circulating in the past (Smith et al., 2004). Amino acid substitutions on antigenic sites on HA are responsible for the difference in antigenicity (Koel et al., 2013). HA of subtype H3N2 has 5 antigenic sites (Wilson & Cox, 1990), while H1N1 HA has 4 antigenic sites (Igarashi et al., 2010). It is known that antigenic sites on HA show positive selection, under which non-synonymous mutations occurred more frequently than synonymous mutations (Bush, Bender, et al., 1999; Suzuki, 2008). The adaptive evolution of circulating influenza strains can be observed using genomic sequences stored in public databases in real-time (Neher & Bedford, 2015).

Different age groups have different adaptive immune profiles against influenza viruses. An individual's immunity is known to be mostly affected by the first infection in life. This phenomenon is called the original antigenic sin (Francis et al., 1947; Davenport et al., 1955; Francis, 1960; Lessler et al., 2012; Nachbagauer et al., 2017). Using a mathematical model, Kucharski and Gog demonstrated that the more influence the original antigenic sin has on current immunity against seasonal influenza, the more it alters the age distribution of immunity (Kucharski & Gog, 2012). Gostic et al. showed that the subtype with which an individual was infected first in life affected the severity of infections with H5N1 and H7N9 (Gostic et al., 2016). Using data on vaccine efficacy, Arevalo et al. showed that severity of H1N1 and H3N2 influenza infections was reduced depending on the subtype that the individual was first infected with (Arevalo et al., 2020).

The driving force of the adaptive evolution of seasonal influenza viruses is immunity in the human population, which are different depending on age groups. Several studies have developed computational models to predict influenza variants that would become dominant in subsequent seasons. Bush et al. predicted strains that became dominant in next seasons by using positively selected codons (Bush, Fitch, et al., 1999). Ito et al. used statistics on the number of different amino acids from past strains to predict future dominant strains of H3N2 viruses (Ito et al., 2011). Physicochemical properties of amino acids

on HA have also been used to predict the antigenic variations of H3N2 (Du et al., 2012; Suzuki, 2013; Cui et al., 2014; Suzuki, 2015). Steinbrück et al. combined serological data with the phylogenetic tree of HA to predict suitable vaccine strains (Steinbrück et al., 2014). Łuksza and Lässig developed a model to estimate the fitness of H3N2 strains using adaptive mutations on antigenic sites and deleterious mutations outside the antigenic sites on HA (Łuksza & Lässig, 2014). Neher et al. used the shape of genealogical tree to predict progenitor lineage of the upcoming season (Neher et al., 2014). Huddleston et al. developed a model to predict the frequency of an H3N2 strain in the future using its current frequency and fitness, determined by the novelty of antigenic sites and the mutational load in non-antigenic sites of HA (Huddleston et al., 2020). A comprehensive list of previous models attempting to predict the evolution of influenza viruses can be found in a review paper by Morris et al. (Morris et al., 2018). However, none of these previous studies considered the age distribution of patients to predict the evolution of influenza viruses.

Given the high mutation rate of influenza viruses, variants having different amino acids on HA continuously emerge during seasonal epidemics (Fitch et al., 1991). However, only a limited number of new amino acids become fixed in the viral population while most of them become extinct. The probability that a new allele becomes fixed is called fixation probability. The relationship between allele frequency and fixation probability was investigated in conditions under neutral evolution (Kimura, 1955), adaptive evolution (Kimura, 1962), nearly neutral evolution (Ohta, 1992), and various relaxed assumptions (Gerrish & Lenski, 1998; Gavrilets & Gibson, 2002; Wilke, 2003; Lambert, 2006; Patwa & Wahl, 2008). There are a few previous works studying the fixation probability of variants of influenza viruses. Steinbrück et al. analyzed the allele frequency of H3N2 viruses over time and showed that alleles that increases in frequency more rapidly were more likely to become fixed (Steinbrück & McHardy, 2011), and this phenomenon was later confirmed by computer simulations (Castro et al., 2020). Strelkowa et al. found that non-synonymous mutations on non-antigenic sites of HA reduced the fixation probability of strains (Strelkowa & Lässig, 2012). Illingworth et al. modeled the effect of linkage disequilibrium on the selection of alleles in adaptive evolution, and they estimated the influence of interference by other alleles in the evolution of H3N2 strains (Illingworth & Mustonen, 2012).

In this chapter, the relationships between fixation probabilities of amino acid substitutions on HA and variant frequencies, patient age distributions, and the involvement on antigenic sites are investigated. Mathematical models of the selective advantage of a variant having a new amino acid are constructed using patient age distributions and antigenic site involvement. The fixation probability is calculated using Kimura's formula for advantageous selection. The model parameters are estimated by maximizing the likelihood of fixation and extinction events observed in the HA sequence data of H1N1 influenza viruses circulating from 2009 to 2020. The predictability of models is evaluated using training-test cross-validations.

1.3. Materials and Methods

1.3.1. Sequence data

Complete HA sequences of H1N1 pandemic 2009 influenza A viruses isolated from humans during March 2009 to May 2020 were downloaded from the Global Initiative on Sharing All Influenza Data (GISAID) (Shu & McCauley, 2017). The HA sequences that had metadata about ages of patients were selected and used for subsequent analyses. As a result, a total of 24,681 unique pairs of an HA amino acid sequence of the virus variant and the age of its patient were obtained. To investigate temporal changes in the frequency of variants having different amino acids on HA, the sequences were grouped into four-month sliding windows. A total of 130 four-month sliding windows were obtained. The first sliding window contained HA sequences from March to June 2009, the second contained those from April to July 2009, and the last contained those from February to May 2020. The four-month sliding windows contained an average of 741.58 sequences with a minimum of 47 and a maximum of 4,347 sequences. The temporal changes in the number of sequences in each four-month window is shown in Figure 1.1A. The patient age distribution for each year from March 2009 to May 2020 is shown in Figure 1.1B.

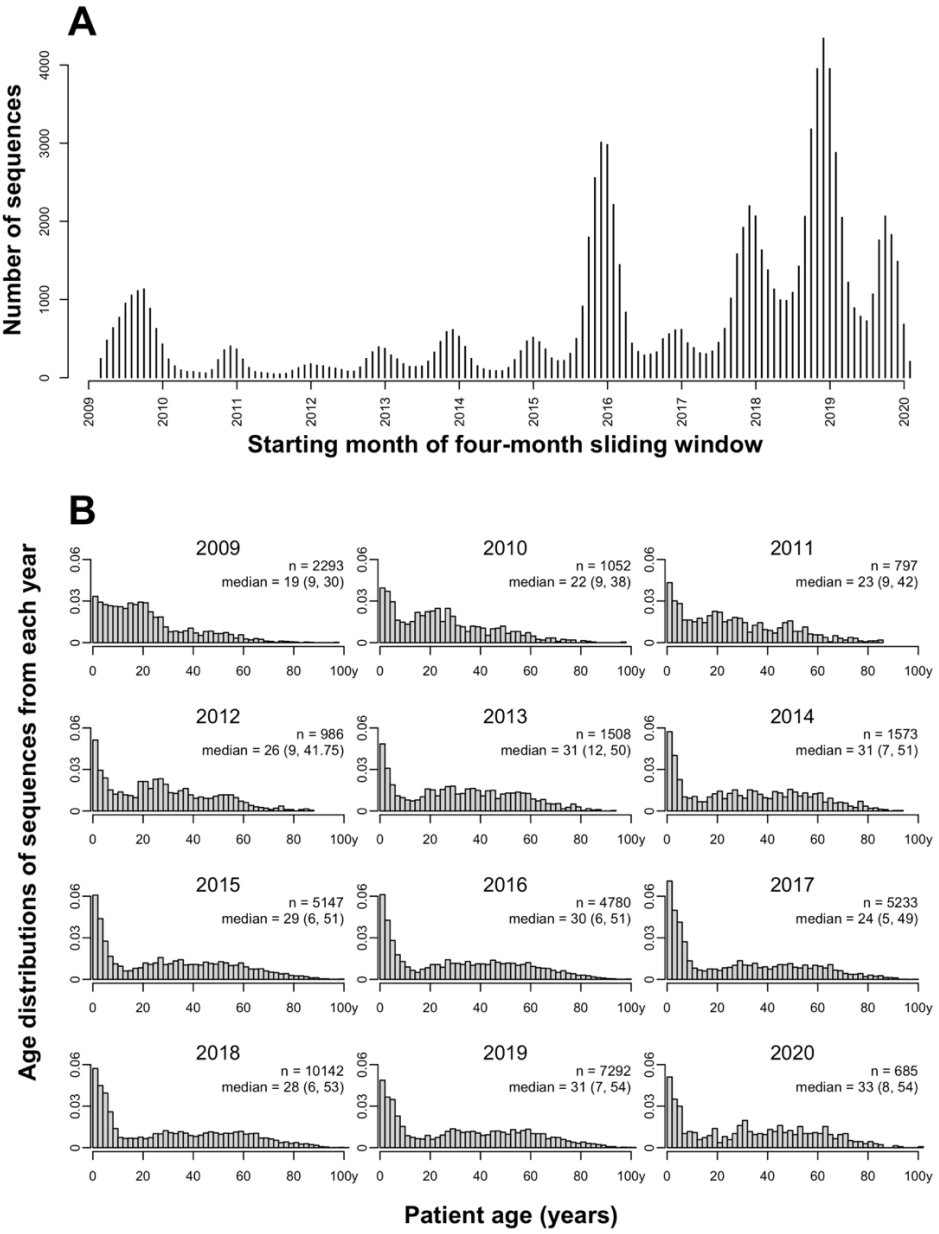

Figure 1.1 Temporal changes in the number of sequences and patient age distributions. (A) The temporal changes in the number of sequences in each four-month sliding window. Each bar represents the number of sequences in a four-month sliding window in which the metadata about patient ages of viruses were provided. The X-axis represents the starting month of the four-month windows. (B) Yearly distributions of patient ages obtained from the downloaded sequence data from March 2009 to May 2020. X-axis represents the patient age in years. Y-axis represents the frequency of variant sequences having patient age shown in the X-axis. Total number of sequences, median patient age, and second and fourth quartile of patient ages in each year are shown in each panel.

1.3.2. Tracking frequencies of newly emerged amino acids

Amino acid substitution is a process where an allele having a new amino acid at a residue position on a protein becomes fixed and those having the other amino acids at the position become extinct. To track the transition from an old amino acid to a new amino acid at a position on HA, frequencies of amino acids on each position for each sliding window were calculated. Historically, an allele is called fixed when its frequency becomes 1.0, and extinct when it becomes 0.0. In this study, however, the condition of fixation was relaxed and considered an amino acid as fixed when its frequency exceeds 0.95 in a sliding window. The reason for this relaxation is that variants with new amino acids other than the amino acid of interest emerge from time to time, and there is almost no chance for the frequency to become 1.0. The condition for extinction remains the same as the classical definition, where an allele becomes extinct at a frequency of 0.0. For each residue position on HA, an amino acid in a window is called old if the amino acid has just become fixed in the current window or it has been old in its preceding windows. After a fixation event, the other amino acids found at this position are considered as new amino acids. An old amino acid remains old, even though its frequency drops below 0.95, until another amino acid becomes fixed at its position. When an old amino acid has been substituted by another amino acid and appears after the substitution, it is considered as a new amino acid. In the first window, the old amino acid at each position is defined by the amino acid with the highest frequency.

1.3.3. Months from emergence and frequency of amino acids

For each new amino acid, a set of consecutive four-month sliding windows from its emergence to its evolutionary outcome, namely fixation or extinction, was identified. Frequencies of the new amino acid in these windows were recorded with its evolutionary outcome. Amino acids were stratified in the identified four-month sliding windows into strata of frequency ranges with a width of 0.1 starting with (0.0, 0.1] and ending with (0.9, 1.0] according to their frequencies. For each frequency range, the evolutionary outcomes of new amino acids of which frequencies at a time point were within the range were collected and used to calculate the empirical fixation probabilities. The empirical fixation probability of new amino acids within each frequency range was calculated as the number of new amino acids that later became fixed divided by the total number of new amino acids. Amino acids in which their outcomes have not yet been determined were excluded from the calculation. Amino acids with frequencies higher than 95% were excluded from the analysis because they were considered to have already been fixed. The 95% binomial confidence intervals of fixation probabilities were calculated by the method of Clopper and Pearson (Clopper & Pearson, 1934).

1.3.4. Comparison of patient age distributions between new and old amino acids

The transition phase of an amino acid substitution was defined as the period from its emergence to its fixation. For each new amino acid at a position on HA that later became fixed, sequences in each four-month sliding window during its transition phase were divided into three groups: those having the new amino acid that later became fixed at the position, those having new amino acids which later became extinct, and those having the old amino acid. The patient age information paired with the sequences in each group were collected. Patient ages of sequences with the new amino acids that later became fixed and those with old amino acids were compared using two-tailed Wilcoxon rank-sum test, with a null hypothesis that the distribution of patient ages of sequences with the new amino acid that later became fixed are the same as those of the old amino acid. The resulting p-values from the two-tailed Wilcoxon rank-sum test were adjusted by Bonferroni's correction. Cohen's d (Cohen, 1992) was used to estimate the effect size of having a new amino acid on the median patient ages for fixed amino acid substitution.

1.3.5. Relationship between empirical fixation probability and patient ages and epitope flags

For each four-month sliding window that contains at least one new amino acid, the evolutionary outcomes of all new amino acids were collected. To exclude four-month sliding windows that had extremely small numbers of sequences, four-month sliding windows containing less than 60 sequences, the 1st percentile of numbers of sequences of all windows, were excluded from the analyses. A threshold on the minimum frequency was set for a new amino acid in a four-month sliding window to be included in the calculation of the empirical fixation probability. The threshold was set to 0.11 in order to have a total empirical fixation probability of 0.5 (Figure 1.2A). This ensures that the number of four-month sliding windows consisting of new amino acids which became fixed is almost equal to the number of those which became extinct. However, new amino acids which became extinct would naturally appear in less numbers of windows compared to those that became fixed. Thus, the number of unique new amino acids may not be equal. See Discussions for the reason for setting a threshold.

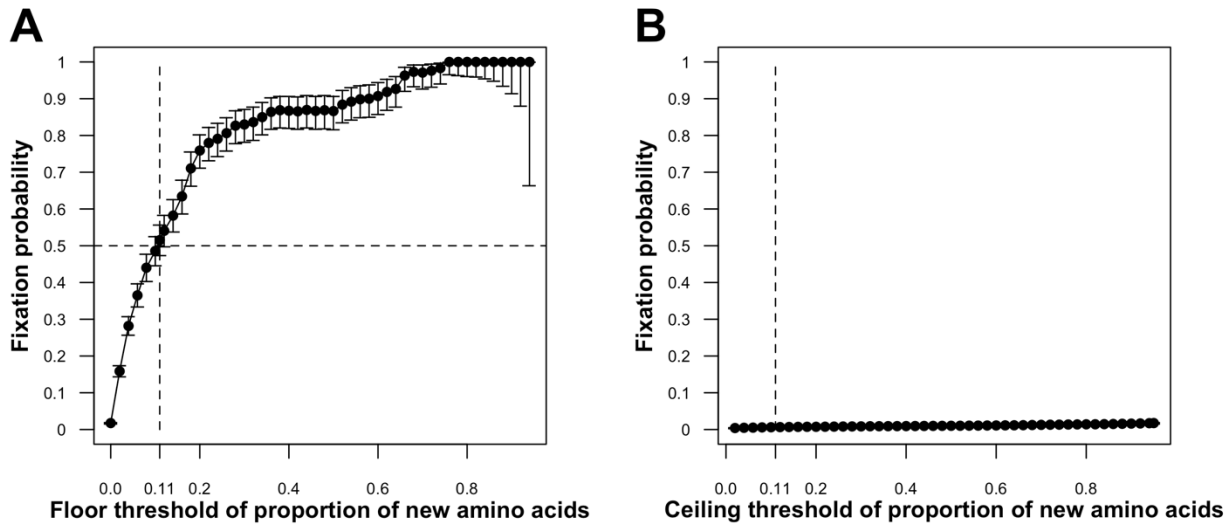

Figure 1.2 Effects of frequency thresholds on the empirical fixation probability of new amino acids. A point in panel (A) represents the empirical fixation probability of new amino acid having a frequency more than the threshold on the x-axis (floor threshold). A point in panel (B) represents the empirical fixation probability of new amino acid having a frequency less than the threshold on the X-axis (ceiling threshold). If no floor threshold for frequency was set during analysis, the fixation probability of all new amino acids would become no more than 0.02 as represented by the left most point in panel (A) and the right most point in panel (B). In this study, a floor threshold frequency was set to 0.11, where the overall empirical fixation probability was closest to 0.5 (A).

For each new amino acid of which frequency among all sequences at its position is more than 0.11 in a four-month sliding window, the profile of the new amino acid was defined as follows. The profile of new amino acid i in a four-month sliding window is represented by a combination of three variables (f_i, a_i, e_i), where f_i is the relative frequency of i in the four-month sliding window, a_i is its patient age statistic, and e_i is its epitope flag. The epitope information of HA of H1N1 viruses was acquired from Igarashi et al. (Igarashi et al., 2010). For epitope flags, $e_i = 1$ if i is a new amino acid in an antigenic site on HA and $e_i = 0$ otherwise.

Profiles of new amino acids in all four-month sliding windows were stratified according to patient age statistic a_i and epitope flag e_i. The patient age statistics included the median patient age of the new amino acids, median patient ages of old amino acids, differences of median patient ages between new and old amino acids, and differences of distribution of patient ages between new and old amino acids, which are defined as follows.

Let X and Y be sets of patient ages of amino acid sequences with a new amino acid and an old amino acid at a position on HA, respectively. The median age difference, $a.diff$, is defined by

$$a.diff = median(X) - median(Y). \tag{1.1}$$

As another statistic for the difference in distributions of patient ages between a new amino acid and an old amino acid at the same position, the z-value of the W statistic under its normal approximation, which is used in the Wilcoxon rank sum test with continuity correction, was used (Hollander et al., 2014). The z-value of rank-sum of a new amino acid, $a.wilcox$, is defined by

$$a.wilcox = \frac{\sum_{i=0}^{|X|} rank(x_i) - \mu_X}{\sigma_X} \tag{1.2}$$

Here, x_i represents an element of X and $rank(x)$ represents the rank of x in $X \cup Y$, and $|X|$ and $|Y|$ represent sizes of X and Y, respectively. The μ_X and σ_X are the expected mean and the standard deviation of the sum of ranks of x in X, which are obtained by

$$\mu_X = \frac{|X|(|X| + |Y| + 1)}{2} \tag{1.3}$$

and

$$\sigma_X = \sqrt{\frac{|X||Y|(|X| + |Y| + 1)}{12}}. \tag{1.4}$$

The empirical fixation probability for each stratum was calculated from the number of profiles of new amino acids that later became fixed and those that later became extinct.

1.3.6. Model of fixation probability

Kimura's formula for advantageous selection (Kimura, 1962) was used to represent the fixation probability of a new amino acid. Thus, the fixation probability of a new amino acid at a residue position on HA, $P_{fix}(f, Ns)$, is given by

$$Pfix(f, Ns) = \frac{1 - e^{-4Nsf}}{1 - e^{-4Ns}} \tag{1.5}$$

where N is the effective population size, s is the selective advantage of the amino acid, and f ($0 \leq f \leq 1$) is the frequency of viruses having the new amino acid at the position on HA in the viral population. N was assumed to be constant over time to use Equation 1.5 as the first approximation for its simplicity. This constant assumption of viral population is discussed in detail in the Discussion.

Let s_i be the selective advantage of viruses that have new amino acid i at a position on HA over those having the other amino acids at the same position. In this study, it is hypothesized that s_i can be represented as a linear combination of factors associated with survival in the human population. By assuming a constant effective population N, the product of N and selective advantage s_i are expressed as

$$Ns_i = C_a a_i + C_e e_i + C_0 , \tag{1.6}$$

where the parameter a_i is an age statistic representing how effectively the viruses with new amino acid i can infect adults compared to those with the old amino acid at the same position. The parameter e_i is the epitope flag of the position expressing whether the position is an antigenic site of HA. C_a, C_e, and C_0 represent coefficients for the age statistic, the epitope flag, and the intercept, respectively.

Combinations of age statistics $a.diff_i$, $a.wilcox_i$, and epitope flag e_i for a new amino acid i yield a total of six models.

(M1) $Ns_i = C_0$,

(M2) $Ns_i = C_a a.diff_i + C_0$,

(M3) $Ns_i = C_a a.wilcox_i + C_0$,

(M4) $Ns_i = C_e e_i + C_0$,

(M5) $Ns_i = C_a a.diff_i + C_e e_i + C_0$,

(M6) $Ns_i = C_a a.wilcox_i + C_e e_i + C_0$

Suppose $F = \{(f_1^F, a_1^F, e_1^F), (f_2^F, a_2^F, e_2^F), \ldots, (f_n^F, a_n^F, e_n^F)\}$ is a set of profiles of new amino acids that later became fixed and $E = \{(f_1^E, a_1^E, e_1^E), (f_2^E, a_2^E, e_2^E), \ldots, (f_m^E, a_m^E, e_m^E)\}$ is a set of those that later became extinct. The likelihood of coefficients $\theta = (C_a, C_e, C_0)$ is given by

$$L(\theta) = \prod_{i=1}^{n} \left(P_{fix}\left(f_i^F, Ns_i^F\right)\right) \prod_{j=1}^{m} \left(1 - P_{fix}\left(f_j^E, Ns_j^E\right)\right)$$

$$= \prod_{i=1}^{n} \frac{1 - exp(-4Ns^F{}_i f^F{}_i)}{1 - exp(-4Ns^F{}_i)} \prod_{j=1}^{m} (1 - \frac{(1 - exp(-4Ns^E{}_j f^E{}_j))}{1 - exp(-4Ns^E{}_j)})$$

$$= \prod_{i=1}^{n} \frac{1 - \exp\left(-4(C_a a^F{}_i + C_e e^F{}_i + C_0) f^F{}_i\right)}{1 - \exp\left(-4(C_a a^F{}_i + C_e e^F{}_i + C_0)\right)}$$

$$= \prod_{j=1}^{m} \left(1 - \frac{\left(1 - \exp\left(-4\left(C_a a^E{}_j + C_e e^E{}_j + C_0\right) f^E{}_j\right)\right)}{1 - \exp\left(-4\left(C_a a^E{}_j + C_e e^E{}_j + C_0\right)\right)}\right) \quad (1.7)$$

The maximum likelihood estimation of $\theta = (C_a, C_e, C_0)$ was performed by maximizing the logarithm of $L(\theta)$. The optim function in R software was used for the maximization of log likelihood (Bélisle, 1992). The 95% confidence intervals for each parameter were obtained by the profile likelihood methods (Pawitan, 2001).

1.3.7. Evaluation of models

The models of fixation probability were evaluated using four-fold cross-validation prediction tests. From March 2009 to May 2020, there were 62 new amino acids exceeding relative frequency of 0.11, of which 19 resulted in fixation and 43 resulted in extinction. The 19 fixed amino acids were considered as positive samples, $F = D_1^+ \cup D_2^+ \cup \cdots \cup D_{19}^+$, consisting of 304 profiles in total. The 43 extinct amino acids were considered as negative samples, $E = D_1^- \cup D_2^- \cup \cdots \cup D_{43}^-$, consisting of 286 profiles in total. Because it can take longer than four months for an amino acid to reach fixation or extinction, the same new amino acid appears in multiple profiles from different four-month windows during the course of

its evolutionary trajectory. For this reason, the number of profiles exceeds the number of fixed amino acids and extinct amino acids.

New amino acids at different positions may evolve in an almost perfect linkage disequilibrium. If profiles of an amino acid substitution in a test set is linked to another amino acid substitution in the training set, the result of the cross-validation test may be affected by the shared information of a single evolutionary event. To avoid sharing information of the linked amino acid substitution between the training set and the test set in cross-validation tests, groups of amino acid substitutions that are almost in perfect linkage disequilibrium were identified using correlation coefficient squared r^2 based on linkage disequilibrium coefficient (Sved & Hill, 2018).

Suppose we have a new amino acid, A, at a position of HA and a new amino acid, B, at another position. Let $p(A)$ and $p(B)$ denote the frequency of allele A and B in the population, respectively. The square of correlation coefficient between two alleles, r^2, which is commonly used to measure linkage disequilibrium of a pair of alleles at two loci, is defined by

$$r^2(AB) = \frac{\big(p(AB) - p(A)p(B)\big)^2}{p(A)\big(1-p(A)\big)p(B)\big(1-p(B)\big)} \tag{1.8}$$

An r^2 value of one means that the pair are in perfect linkage disequilibrium. To identify groups of linked new amino acids, pairwise r^2 between all new amino acids occurring during overlapping periods were calculated. Highly linked new amino acids, defined by having pairwise r^2 of more than 0.75, are grouped together using DBSCAN algorithm (Sander et al., 1998). The cutoff value for r^2 of 0.75 was selected so that all synchronized pairs of fixed new amino acids, visually identified from Figure 1.3, were grouped together and that the total number of groups remained as large as possible (Figure 1.4). Using the cutoff value, a total of 49 groups of amino acid substitutions, each of which consists of new amino acids that are almost in perfect linkage disequilibrium with another amino acid in the group, were identified.

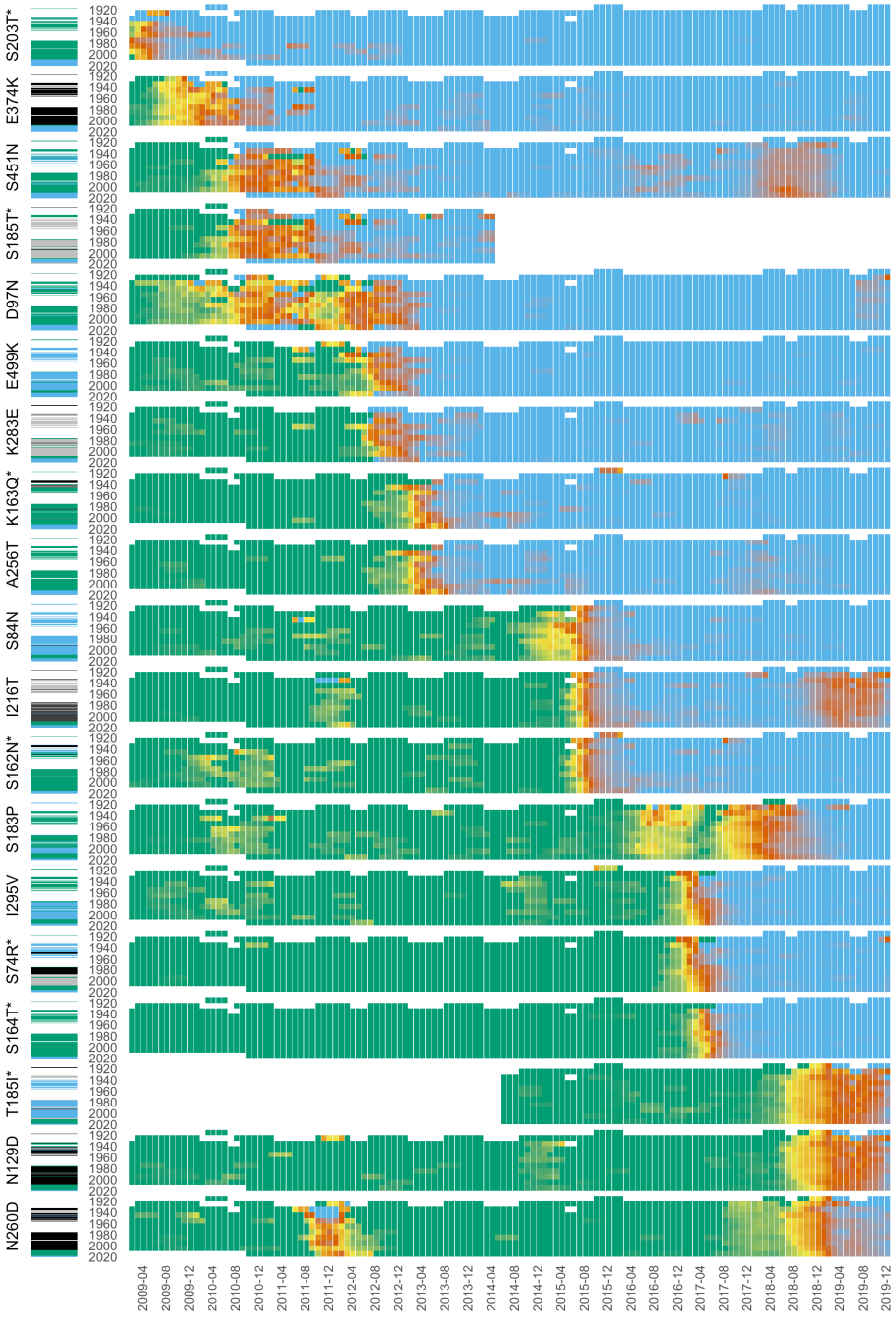

Figure 1.3 Timing chart of amino acid substitutions in different birth-year groups. Panels in the figure correspond to amino acid substitutions shown in panels in Figure 1.12. An asterisk indicates that the amino substitution occurred at a position on the antigenic sites. In order to see the synchronized transitions of amino acids in different positions, the frequency of viruses having the fixed new amino acid in the population of a birth-year group were depicted for four-month sliding windows started from March 2009 to February 2020. See the legend of Figure 1.12 for details of X-axis, Y-axis, heatmap, and bars.

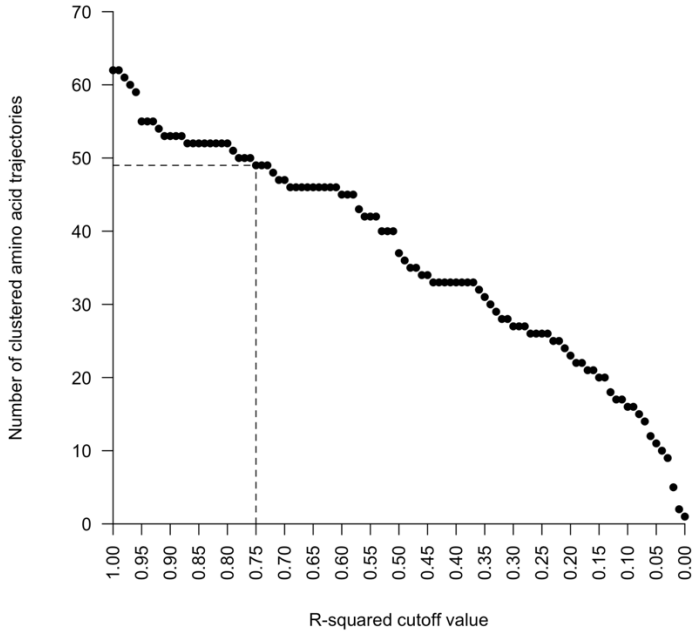

Figure 1.4 Number of clusters of new amino acid trajectories after grouping with r^2 cutoff values from 1 to 0. Grouping starts with a total of 62 singleton groups consisting of a single new amino acid trajectory when r^2 cutoff value equals to 1, where no two amino acid trajectories are grouped together. Grouping ends with a single group consisting of all the new amino acid trajectories when r^2 cutoff value equals to 0, where all amino acid trajectories are grouped together. Dotted line represents the r^2 cutoff value of 0.75, which is used in the analyses, as this is the maximum cutoff value where all visually identified synchronized pairs of fixed amino acids are grouped together.

Finally, in order to perform four-fold cross-validation, the 49 groups were randomly assigned to four datasets. Three of which consisted of 12 groups and the other consisted of 13 groups (Figure 1.5). For each random assignment, profiles in three of the datasets were used as training set to estimate parameters of each model by maximizing log likelihood to the observed evolutionary outcomes. The other dataset was used as test set to evaluate the predictability of the model. Four cross-validation tests were conducted in each random assignment and this process was repeated 100 times. The cross-validation was performed 400 times in total. The grouping of profiles prior to cross-validation ensured that no fixation or extinction events of the same amino acids or amino acids in linkage disequilibrium were shared between the training and test data during cross-validation. The models' Akaike information criteria (AICs) were calculated from the log likelihood estimation of the training set. Figure 1.5 shows the schematic diagram of cross-validation tests.

In each prediction test, the model predicts a new amino acid to become fixed if P_{fix} for its profile in its four-month sliding window is greater than 0.5, and extinct if otherwise. Sensitivity, specificity, precision, and Youden's index of each model were calculated from the number of true positive predictions (tp), true negative predictions (tn), false positive predictions (fp), and false negative predictions (fn) as follows:

$$Sensitivity = \frac{tp}{tp + fn}, \tag{1.9}$$

$$Specificity = \frac{tn}{tn + fp}, \tag{1.10}$$

$$Precision = \frac{tp}{tp + fp}, \tag{1.11}$$

$$Youden's\ index = Sensitivity + Specificity - 1 \tag{1.12}$$

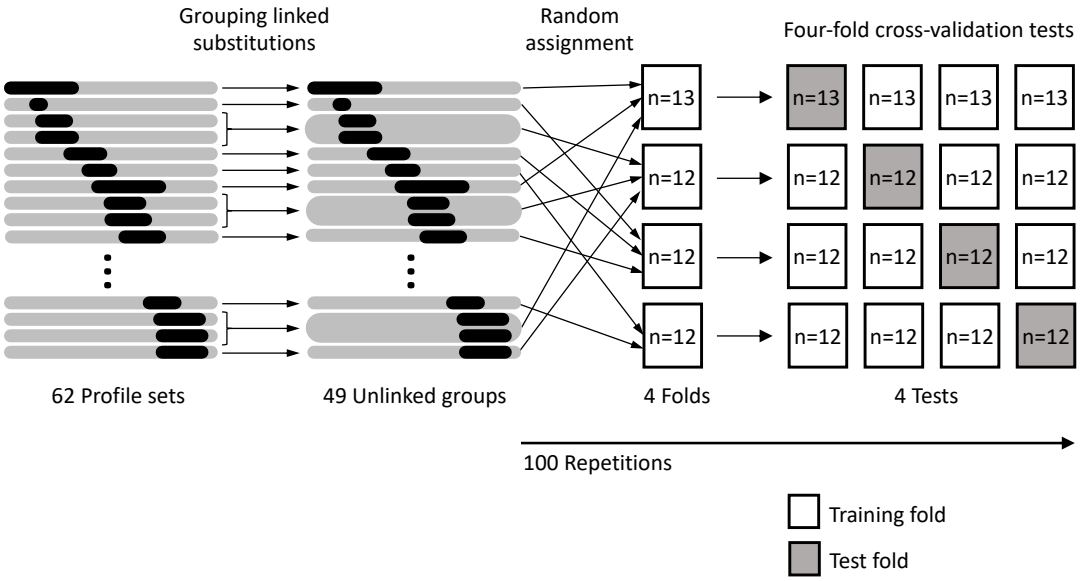

Figure 1.5 Schematic diagram of cross-validation tests

1.3.8. Timing of amino acid substitutions in different birth-year groups

The timing of amino acid substitutions in different birth-year groups were visualized as follows. For each new amino acid that later became fixed, amino acid sequences during its transition phases were divided into ten-year bins according to the year when patients were born. The relative frequency of sequences having the new amino acid at the position among those having new and old amino acids was calculated for each birth-year group in each four-month sliding window. The frequency of a new amino acid for a birth-year group equals zero when the new amino acid has not yet been found at its position on HA of viruses isolated from patients in the birth-year group. The frequency becomes one when viruses having the old amino acid at the position on HA was completely replaced by those having new amino acid in patients of the birth-year group.

The dominant amino acids on HA of the H1N1 strains circulating before the 2009 pandemic were determined from amino acid sequences obtained from GISAID database.

1.4. Results

1.4.1. Empirical fixation probability of new amino acids on HA

From March 2009 to May 2020, HA had a total of 4,580 new amino acids at 491 amino acid positions, which cover 89% of residues of the molecule. Figure 1.6A shows trajectories of frequencies in four-month sliding windows of all these amino acids. For some amino acids, it took twelve months to become fixed while some took fifty-four months. Most of the new amino acids became extinct shortly after their emergence, while a few of them remained for more than seventy months. Of 4,580 new amino acids, nineteen resulted in fixation (solid lines) while the others became extinct (dotted lines). The empirical fixation probability of all new amino acids was 0.004. However, the empirical fixation probability increased as the frequency of new amino acids increased (Figure 1.6B). Table 1.1 shows the number of fixed and extinct amino acid trajectories that reached a frequency of 0.10 and those that did not reach the frequency. The number of fixed and extinct amino acid trajectories is dependent on whether the frequency of viruses having new amino acids have exceeded 0.10 or not ($p < 10^{-16}$ with χ^2 test). The fixation probabilities exceeded the mid-point value of each frequency range of the new amino acids when the frequency is above 0.10. The lower 95% confidence intervals of fixation probabilities for the frequency ranges within 0.20 to 0.55 exceeded the mid-point values of those frequency ranges.

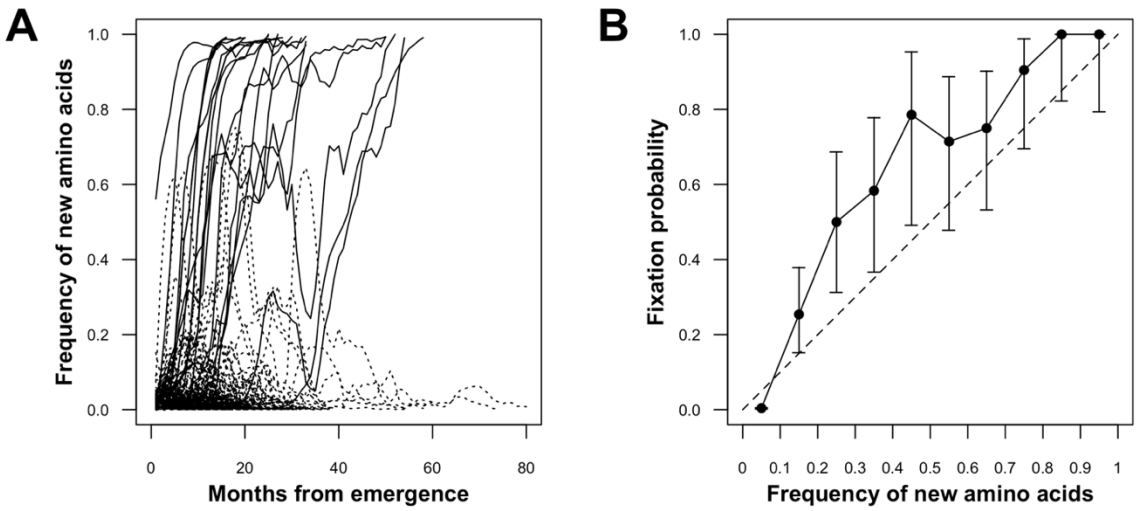

Figure 1.6 The frequency of new amino acids on HA and their empirical fixation probability. (A) Trajectories of the frequency of new amino acids in the population from their emergence to fixation or extinction. Solid lines represent the frequency of new amino acids that later became fixed while dotted lines represent those that later became extinct. (B) Empirical fixation probabilities of new amino acids stratified by their frequencies. X-axis represents the frequency of new amino acids at a position in the population. Y-axis represents the empirical fixation probability of new amino acid trajectories that reached the frequency on the X-axis. Error bars are the 95% binomial confidence intervals of the fixation probability.

Table 1.1 Number of fixed and extinct amino acid trajectories that reached a frequency of 0.10 and those that did not

Frequency	Number of new amino acids which became fixed	Number of new amino acids which became extinct
(0.00, 0.10]	18	4451
(0.10, 1.00]	143	89

Neutral evolution is an evolutionary process where every new allele becomes fixed with an equal chance. It is known that the fixation probability of a strain would be equal to its frequency under neutral evolution (Kimura, 1955). If the fixation of amino acid substitutions occurs under neutral evolution, the fixation probability will fall upon the neutral line (dashed line in Figure 1.6B). The excess of the empirical fixation probability indicates that the fixation of new amino acids on HA is under adaptive evolution where viral or environmental factors increase their chance of becoming fixed.

1.4.2. Fixation of new amino acids on HA

As of December 2020, nineteen new amino acids on HA have become fixed since the beginning of the pandemic in 2009 (Table 1.2). The nineteen fixations occurred on eighteen positions on HA as two fixations occurred at the same position, 185 on HA. These eighteen positions spread across the HA1 domain with three exceptions occurring on HA2 (positions 374, 451, and 499). Seven (36.84%) out of nineteen substitutions occurred on either of the four distinct antigenic sites, Sa, Sb, Ca, and Cb (Igarashi et al., 2010).

Of nineteen fixed new amino acids, seventeen had higher median patient ages than old amino acids during their transition phases (Table 1.2). Exceptions were amino acid substitutions at positions 74 and 164. Arginine (R) at position 74 had the same median age as Serine (S). Threonine (T) at position 164 had lower median patient age than S. The median patient ages of viruses having the fixed new amino acids was higher than those of the old amino acids by an average of 4.4 years. Viruses having fourteen fixed new amino acids (73.68%) had significantly higher patient ages than those having the old amino acids at the same position during their transition phases. Cohen's d effect size based on the 19 pairs of median patient ages of new amino acid and old amino acid in Table 1.2 was estimated to be 1.11, with 95% CI from 0.45 to 1.77. The effect sizes are considered as negligible, small, medium, and large when $|d| < 0.2$, $|d| < 0.5$, $|d| < 0.8$, and $|d| \geq 0.8$, respectively (Cohen, 1992). Thus, we can reject a null hypothesis that the effect of having a new amino acid on median patient age is negligible in fixed amino acid substitutions. This result indicated that viruses that had been selected by human immunity had non-negligible excess infectivity to the adult population. The patient ages of viruses having new amino acids may be used as an indicator for viral fitness driving the amino acid substitutions.

Table 1.2 Amino acid substitutions on HA and median patient ages during their transition phases

Position	Antigenic site	Transition phase	Duration (month)	Old amino acids		New amino acids		Difference between median patient ages (year)	p-value
				Amino acid (n)	Median patient age (Q_1, Q_3)† (year)	Amino acid (n)	Median patient age (Q_1, Q_3)† (year)		
203	Ca	2009.03–2009.09	6	S (211)	18 (9, 31)	T (632)	20 (11, 33)	2	≈1
374	-	2009.03–2011.04	25	E (1747)	19 (10, 31)	K (984)	21 (8, 36.25)	2	0.6538
451	-	2009.04–2012.08	40	S (2468)	19 (9, 31)	N (691)	23 (10, 41)	4	0.0002***
185	Sb	2009.10–2012.08	34	S (1647)	19 (9, 31)	T (669)	23 (9, 41)	4	0.0004***
97	-	2009.04–2013.05	49	D (2654)	19 (9, 32)	N (1056)	24 (8, 41)	5	0.0002***
499	-	2011.07–2013.05	22	E (482)	22 (8, 37)	K (437)	28 (9, 43)	6	0.0262*
283	-	2012.03–2013.06	15	K (345)	21 (7, 37)	E (453)	28 (10, 44)	7	0.0141*
163	Sa	2012.07–2013.11	16	K (464)	26 (6, 40.25)	Q (303)	30 (16.5, 44)	4	0.0208*
256	-	2012.07–2013.11	16	A (588)	25 (6, 41)	T (308)	29 (16, 44)	4	0.0231*
84	-	2014.08–2016.06	22	S (1176)	23 (5, 47)	N (3900)	31 (7, 51)	8	5.13×10^{-6}***
216	-	2014.10–2016.06	20	I (1467)	23 (5, 47)	T (3579)	32 (7, 52)	9	6.33×10^{-9}***
162	Sa	2015.05–2016.06	13	S (852)	20 (4, 43)	N (3568)	32 (7, 52)	12	8.81×10^{-14}***
295	-	2016.09–2017.09	12	I (958)	25 (5, 47)	V (399)	29 (5, 51)	4	≈1
74	Cb	2016.09–2017.10	13	S (971)	24 (5, 47)	R (583)	24 (4, 48)	0	≈1
164	Sa	2016.10–2017.10	12	S (1082)	27 (5, 48)	T (385)	19 (3, 47)	–8	0.1883
183	-	2014.08–2019.02	54	S (8713)	27 (5, 49)	P (6392)	28 (6, 52)	1	4.51×10^{-5}***
260	-	2017.06–2020.01	31	N (7413)	22 (5, 48)	D (6020)	33 (8, 55)	11	6.50×10^{-48}***
185	Sb	2015.09–2020.02	53	T (13718)	28 (5, 51)	I (5088)	30 (6, 53)	2	0.0072**
129	-	2017.06–2020.02	32	N (8764)	26 (5, 50)	D (4718)	32 (7, 54)	6	1.58×10^{-18}***

†Q_1 and Q_3 represent the first and third quartiles of patient ages, respectively.
††The difference in median patient age is calculated by subtracting the median patient age of old amino acid from the median patient age of new amino acid.
*$p < 0.05$ by two-sided Wilcoxon rank-sum test adjusted by Bonferroni's correction with $n = 19$.
**$p < 0.01$ as above.
***$p < 0.001$ as above.

1.4.3. Factors associated with fixation probability

Figure 1.7 shows empirical fixation probabilities of new amino acids stratified by attributes in their profiles. Empirical fixation probabilities of new amino acids varied with median patient ages (Figure 1.7A and 1.7B). Table 1.3 shows the number of fixed and extinct new amino acid profiles having median patient ages between 25 and 35 and those of the others. The number of fixed and extinct new amino acid profiles is dependent on whether their median patient ages are between 25 and 35 or not ($p < 10^{-15}$ with χ^2 test). Table 1.4 shows the number of fixed and extinct new amino acid profiles in which median patient ages of old amino acids are less than or equal to 15 and those of the others. The number of fixed and extinct new amino acid profiles is dependent on whether the old amino acids have a median patient age less than or equal to 15 ($p < 10^{-9}$ with χ^2 test). These results indicated that new amino acids tended to become fixed when the viruses with the new amino acids infected the population with a median age between 25 to 35 or when the viruses having the old amino acids at the corresponding positions infected the population with a median age from 0 to 15.

The correlation between empirical fixation probabilities and the excess infectivity of strains with new amino acids to the adult population over strains with old amino acids was further investigated (Figure 1.7C and 1.7D). The excess infectivity of the new strains to the adult population was measured by comparing patient age distributions of amino acid sequences with new amino acids and old amino acids at the same positions on HA.

Empirical fixation probabilities of new amino acids were positively correlated with the excess of median patient ages of new amino acids with respect to those of old amino acids (Figure 1.7C). Pearson's correlation coefficient between fixation probability and excess in median patient ages was 0.94 ($p < 10^{-3}$). The empirical fixation probability was also positively correlated with the z-value of rank-sums of the patient ages of new amino acids (Figure 1.7D), with a correlation coefficient of 0.95 ($p < 10^{-2}$). Figure 1.8A shows a scatterplot of fixation probability versus excess in median patient ages with its regression line. Figure 1.8B shows a scatterplot of fixation probability versus z-value of rank-sums of the patient ages of new amino acids with its regression line. Both results indicated that fixation probabilities of new amino acids increased when viruses having the new amino acids on HA infected the population with a higher age than those infected with viruses with old amino acids at the corresponding positions.

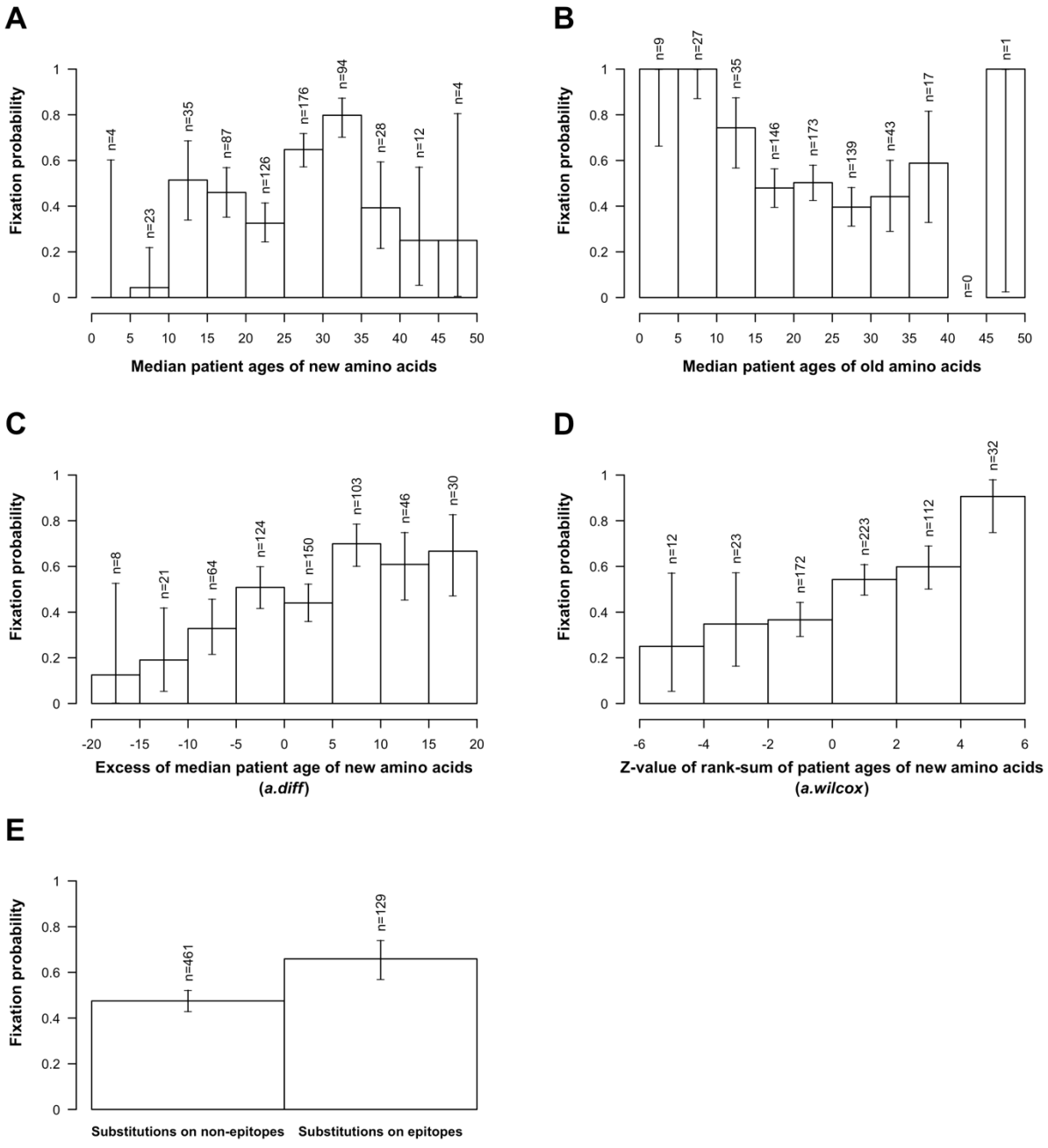

Figure 1.7 Distributions of new amino acids and their empirical fixation probabilities. Empirical fixation probabilities of new amino acids stratified with (A) the median patient ages of sequences having new amino acids, (B) median patient ages of sequences having old amino acids, (C) excesses of median patient age of sequences having new amino acids with respect to old ones, (D) z-values of rank-sum of patient ages of sequences having new amino acid, (E) epitope flags at the position of amino acids. The error bars indicate the 95% binomial confidence intervals of fixation probabilities.

Table 1.3 Number of fixed and extinct new amino acid profiles having median patient ages between 25 and 35 and those of the others

Median age of new amino acid	Number of new amino acids which became fixed	Number of new amino acids which became extinct
(25, 35]	189	81
Others	115	205

Table 1.4 Number of fixed and extinct new amino acid profiles in which median patient ages of old amino acids are less than or equal to 15 and those of the others

Median age of old amino acid	Number of new amino acids which became fixed	Number of new amino acids which became extinct
(0, 15]	62	9
Others	242	277

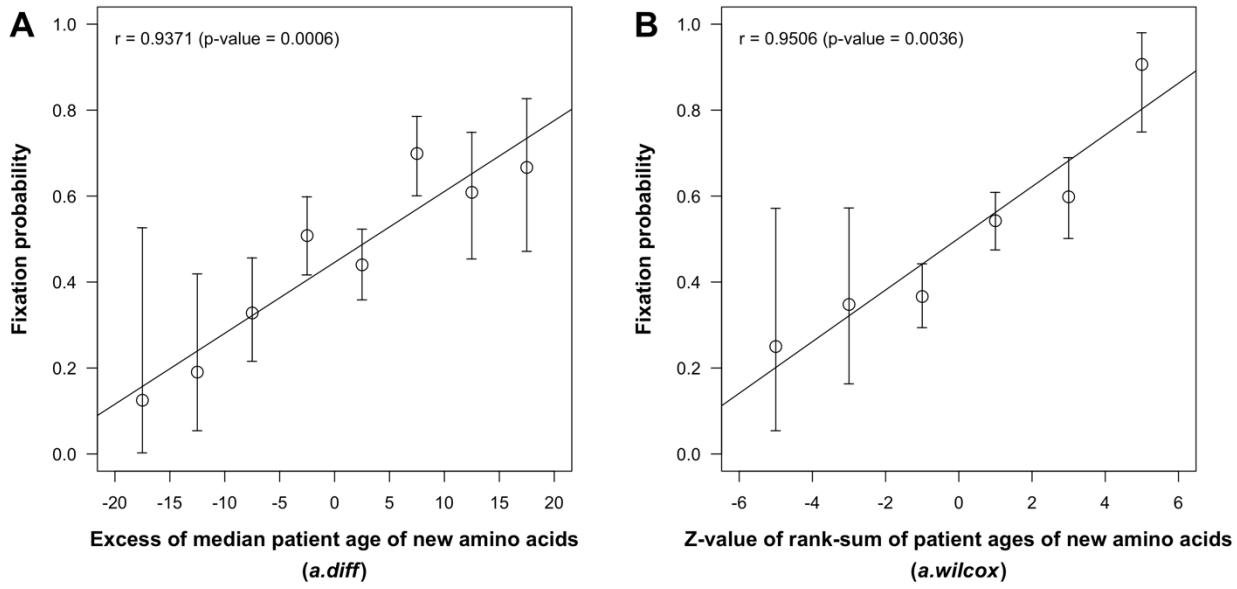

Figure 1.8 Scatterplots of fixation probability and difference between patient ages. (A) A scatterplot of fixation probability versus excess in median patient ages with its regression line. X-axis corresponds to the excess of median patient age of new amino acids ($a.diff$). (B) A scatterplot of fixation probability versus z-value of rank-sums of the patient ages of new amino acids with its regression line. X-axis corresponds to the z-value of rank-sums of the patient ages of new amino acids ($a.wilcox$). Y-axis of each panel represents fixation probability of new amino acids. Each circle represents the empirical fixation probability of new amino acids having $a.diff$ or $a.wilcox$ within ranges between the two adjacent tick marks on the x-axis. Error bars indicate the 95% binomial confidence intervals of fixation probabilities. Correlation coefficient is calculated using fixation probabilities plotted in circles.

Figure 1.7E shows empirical fixation probabilities of new amino acids stratified according to its involvement on antigenic sites. The empirical fixation probability of new amino acids at antigenic sites was 0.66 with a 95% binomial confidence interval of 0.57 to 0.74. On the other hand, the empirical fixation probability of new amino acids at non-antigenic sites was 0.48 with a 95% binomial confidence interval of 0.43 to 0.52. The fixation probability of new amino acids at antigenic sites was significantly higher than that of new amino acids at non-antigenic sites ($p < 10^{-3}$ with χ^2 test).

1.4.4. Models of fixation probability

Table 1.5 shows results of the maximum likelihood estimation of parameters of models using profiles having frequencies more than 0.11 in the dataset (See Discussion for details). Model M1, which assumes Ns is constant, has a maximum log likelihood of –260.441 and an AIC of 522.882. The maximum likelihood increased when included age statistics, $a.diff$ (M2) or $a.wilcox$ (M3), and the AIC decreased to 503.345 and 505.310, respectively. When the epitope flag e was added (M4), the maximum likelihood also increased, and the model's AIC decreased to 512.886. The increase in maximum likelihood from the constant advantage model (M1) when modeled with epitope flags (M4) was smaller than the increases when modeled with patient age statistics (M2 and M3).

When Ns was modeled using a combination of a patient age statistic and the epitope flag (M5 and M6), further increases in the maximum likelihood and decreases in AIC were observed. The AIC of M5 was 494.298, which was lower than those of its simpler models, M2 and M4. Similarly, the AIC of M6 was 493.761, which was lower than those of its simpler models M3 and M4. Models using a combination of a patient age statistic and epitope flag seemed to be better to represent the selective advantage of H1N1 influenza viruses than those using either parameter alone.

In the maximum likelihood estimation of M6 in Table 1.5, the 95% CI for C_0 contains zero, which means that the intercept may not necessarily be positive in M6. The epitope flag in the model takes the value of zero or one. The term $C_e e$ always takes a non-negative value if C_e is positive. The average of Ns can become positive even if C_0 takes a negative value. The lower bound of 95% CI for C_0 for M6 would be positive if epitope flags of –1 and +1 were used. The small positive value of the lower bound of 95% CI for M5 can be explained by the same reason.

Table 1.5 Maximum likelihood estimation of parameters of six models

Model	C_a (95% CI)	C_e (95% CI)	C_0 (95% CI)	Maximum log likelihood	AIC	ΔAIC†
(M6) $Ns = C_a a.wilcox + C_e e + C_0$	0.126 (0.073, 0.185)	0.572 (0.265, 0.922)	0.115 (–0.012, 0.243)	–243.881	493.761	0
(M5) $Ns = C_a a.diff + C_e e + C_0$	0.032 (0.018, 0.047)	0.521 (0.206, 0.870)	0.141 (0.016, 0.265)	–244.149	494.298	0.537
(M2) $Ns = C_a a.diff + C_0$	0.031 (0.018, 0.045)	-*	0.228 (0.115, 0.329)	–249.672	503.345	9.584
(M3) $Ns = C_a a.wilcox + C_0$	0.119 (0.065, 0.175)	-	0.217 (0.102, 0.333)	–250.655	505.310	11.549
(M4) $Ns = C_e e + C_0$	-	0.526 (0.219, 0.867)	0.225 (0.109, 0.343)	–254.443	512.886	19.125
(M1) $Ns = C_0$	-	-	0.313 (0.207, 0.421)	–260.441	522.882	29.121

*The hyphens indicate the model does not use that parameter.
†ΔAIC is calculated by subtracting AIC of M6 from the AIC of the model.

1.4.5. Linkage disequilibrium among amino acids

Table 1.6 shows groups of new amino acids that are almost in perfect linkage disequilibrium. Groups of linked amino acids were identified using correlation coefficient squared r^2 based on the frequency of the amino acids. Using a cutoff value of r^2 at 0.75, a total of 49 groups of amino acid substitutions, each of which consisted of new amino acids that are almost in perfect linkage disequilibrium with another amino acid in the group, were identified in the 62 sets of new amino acid profiles.

Table 1.6 New amino acids identified as being almost in perfect linkage disequilibrium

Group no.	Profile set	Substitution	Outcome	Emergence month	Outcome month	r^2
3	D^+_3	S185T	fixed	2009-10	2012-08	0.9592
	D^+_4	S451N	fixed	2009-04	2012-08	
5	D^+_6	E499K	fixed	2011-07	2013-05	0.7943
	D^+_7	K283E	fixed	2012-03	2013-06	
6	D^+_8	K163Q	fixed	2012-07	2013-11	0.9797
	D^+_9	A256T	fixed	2012-07	2013-11	
8	D^+_{11}	S162N	fixed	2015-05	2016-06	0.9689
	D^+_{12}	I216T	fixed	2014-10	2016-06	
9	D^+_{13}	I295V	fixed	2016-09	2017-09	0.9893
	D^+_{14}	S74R	fixed	2016-09	2017-10	
13	D^+_{18}	N129D	fixed	2017-06	2020-02	0.7528
	D^+_{19}	T185I	fixed	2015-09	2020-02	
24	D^-_{11}	I216V	extinct	2010-08	2012-08	0.8785
	D^-_{27}	R205K	extinct	2009-03	2014-04	
25	D^-_{12}	E356A	extinct	2010-10	2012-08	0.7829
	D^-_{21}	H138Q	extinct	2010-10	2013-06	
27	D^-_{14}	S69T	extinct	2011-09	2013-01	0.9506
	D^-_{15}	N260D	extinct	2011-07	2013-01	
34	D^-_{23}	A197T	extinct	2010-07	2013-07	0.9247
	D^-_{25}	S143G	extinct	2010-08	2013-11	
45	D^-_{36}	R45G	extinct	2017-06	2019-05	0.9504 with D^-_{37}
	D^-_{37}	P282A	extinct	2017-07	2019-05	0.9558 with D^-_{39}
	D^-_{39}	I298V	extinct	2017-07	2019-08	0.936 with D^-_{36}
47	D^-_{40}	E68D	extinct	2018-07	2020-02	0.9175
	D^-_{41}	S121N	extinct	2017-08	2020-02	

1.4.6. Evaluation by cross-validation

The predictability of models using four-fold cross-validation tests were evaluated. Table 1.7 shows the means and the standard deviations of AIC and maximum log likelihood for training sets and the means and the standard deviations of sensitivity, specificity, precision, and Youden's index for test sets in the cross-validations. The models were sorted in the descending order of Youden's indices, which is the sum of sensitivity and specificity minus one.

Consistent with Table 1.5, model M6 had the best AIC for training sets, followed by M5. However, model M3 had the highest mean Youden's index of 0.64 with a mean sensitivity of 0.78 and a mean specificity of 0.86. The model M2 had the second highest Youden's index of 0.63 with a mean sensitivity of 0.79 and a specificity of 0.84. The model M1, which assumed Ns is constant, had the third highest Youden's index of 0.62 with a mean sensitivity of 0.76 and mean specificity of 0.86. Models M6, M4, and M5 had lower Youden's indices than M1.

Youden's indices of M3 and M2 have a mean of 0.64 with a standard deviation of 0.11 and a mean of 0.63 with a standard deviation of 0.12, respectively. The difference between the Youden's indices of the models becomes clear when a result using a model for each test is compared to the result using M1 in a pair-wise manner. Figure 1.9 shows the distribution of excess Youden's indices of M2, M3, M4, M5, and M6 over M1 in 400 cross-validation tests. Panel A in Figure 1.9 clearly shows that the excess Youden's index of M3 over M1 was distributed more in the positive side than the negative side. Paired two-sided Wilcoxon rank-sum test adjusted by Bonferroni's correction shows that the Youden's indices of M3 and M2 are significantly larger than that of M1 with p-values of 5.76×10^{-20} and 0.003, respectively. There is no significant difference between the Youden's indices of M6 and M4, compared to M1 ($p \approx 1.000$ and $p = 0.218$, respectively). The Youden's indices of M5 is significantly lower than that of M1 ($p = 4.01 \times 10^{-5}$). These results indicate that the predictability of the fixation of new amino acids is significantly improved compared to the constant advantage model, M1, when Ns is modeled using $a.wilcox$ or $a.diff$ with an intercept.

Table 1.7 Results of cross-validation tests

Model	Training		Test				
	AIC	Maximum log likelihood	Sensitivity	Specificity	Precision	Youden's index	p-value ($n = 400$)
(M3) $Ns = C_a a.wilcox + C_0$	374.27 ± 45.02	–185.13 ± 22.51	0.78 ± 0.09	0.86 ± 0.11	0.83 ± 0.17	0.64 ± 0.11	5.76×10^{-20}***
(M2) $Ns = C_a a.diff + C_0$	373.00 ± 44.65	–184.50 ± 22.33	0.79 ± 0.10	0.84 ± 0.11	0.81 ± 0.17	0.63 ± 0.12	0.003**
(M1) $Ns = C_0$	388.27 ± 42.33	–193.13 ± 21.17	0.76 ± 0.08	0.86 ± 0.11	0.83 ± 0.18	0.62 ± 0.11	-
(M6) $Ns = C_a a.wilcox + C_e e + C_0$	366.35 ± 43.52	–179.17 ± 21.76	0.77 ± 0.11	0.84 ± 0.11	0.80 ± 0.18	0.61 ± 0.14	≈1.000
(M4) $Ns = C_e e + C_0$	381.72 ± 40.84	–187.86 ± 20.42	0.75 ± 0.10	0.85 ± 0.11	0.81 ± 0.18	0.6 ± 0.13	0.218
(M5) $Ns = C_a a.diff + C_e e + C_0$	366.92 ± 43.41	–179.46 ± 21.71	0.77 ± 0.11	0.82 ± 0.11	0.79 ± 0.18	0.59 ± 0.14	4.01×10^{-5}***

All values, except p-values, are presented as mean ± standard deviation in 400 cross-validation tests.
*$p < 0.05$ by two-sided paired Wilcoxon rank-sum test adjusted by Bonferroni's correction with $n = 5$ with a null hypothesis that the model's Youden's indices are the same as those of M1.
**$p < 0.01$ as above.
***$p < 0.001$ as above.

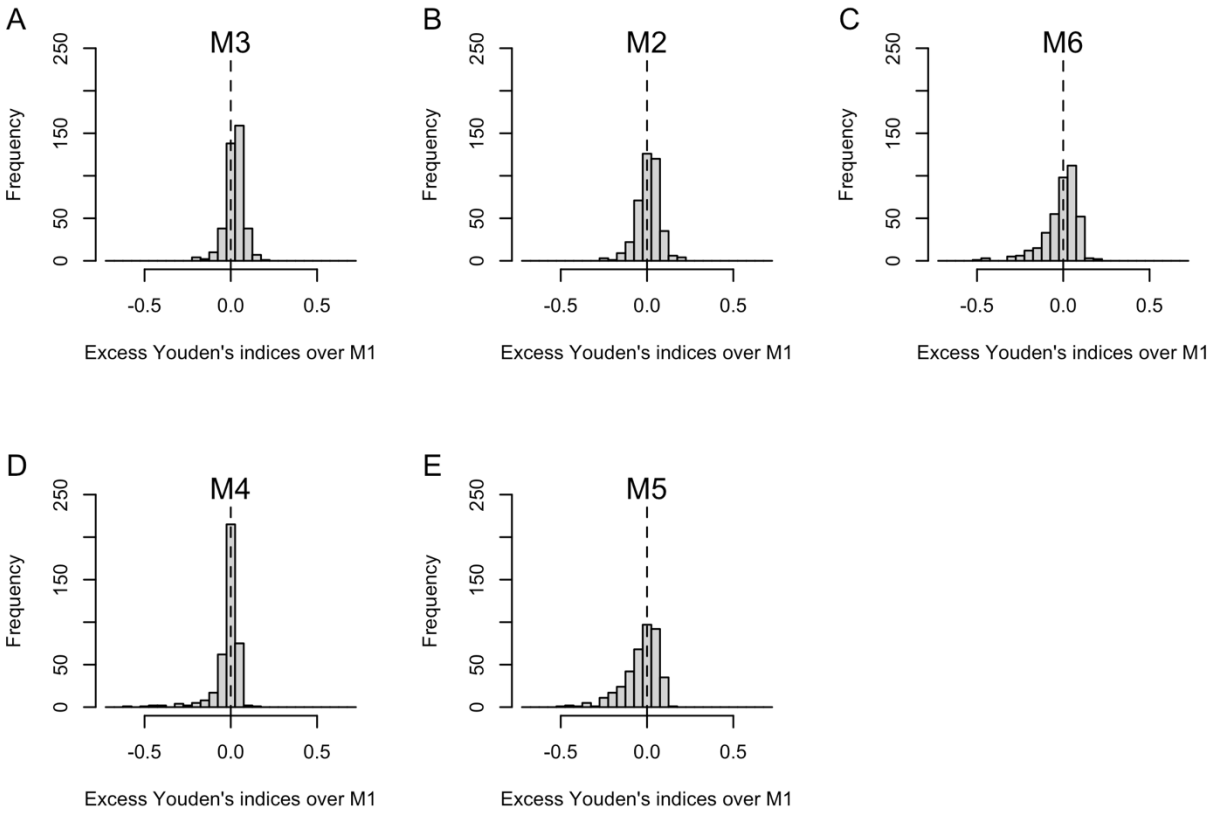

Figure 1.9 Distribution of excess Youden's indices of models over M1. The distribution of excess Youden's indices of M3 (A), M2 (B), M6 (C), M4 (D), and M5 (E) over M1 in 400 cross-validation tests. For each test, the excess of Youden's index of each model was calculated by subtracting the Youden's index of M1 from the Youden's index of the model.

In Table 1.7, a new amino acid was predicted to become fixed when P_{fix} is higher than a threshold of 0.5. The effect of the threshold of P_{fix}, τ, on the prediction of the fixation of new amino acids was further investigated. For each model, sensitivity and specificity for predicting the fixation of new amino acids were obtained by using different thresholds τ from zero to one in cross-validation tests. The sensitivities and specificities were averaged over 400 cross-validation tests for each threshold for each model. Figure 1.10 shows the receiver operating characteristic (ROC) curve created from the resulting sensitivities and specificities. Points around the lower left corner correspond to cross-validation tests using $\tau \cong 1$ and points around the upper right correspond to cross-validation tests using $\tau \cong 0$. The cross on the curve of M3 represents the sensitivity and specificity of M3 when τ equals 0.5, which is shown in Table 1.7. The ROC curve of M3 reached the maximum distance from the diagonal line when τ equals 0.43 (circle). This threshold increased Youden's index of M3 to 0.656 from 0.646 which is obtained when τ equals 0.5. The order of the furthest distances from the diagonal line was the same as the order of Youden's indices in Table 1.7. These results indicated that M3 had the highest predictive power when using a threshold of 0.43. The choice of threshold for P_{fix} faces the trade-off between the sensitivity and specificity, and the value should be determined by considering the purpose of prediction.

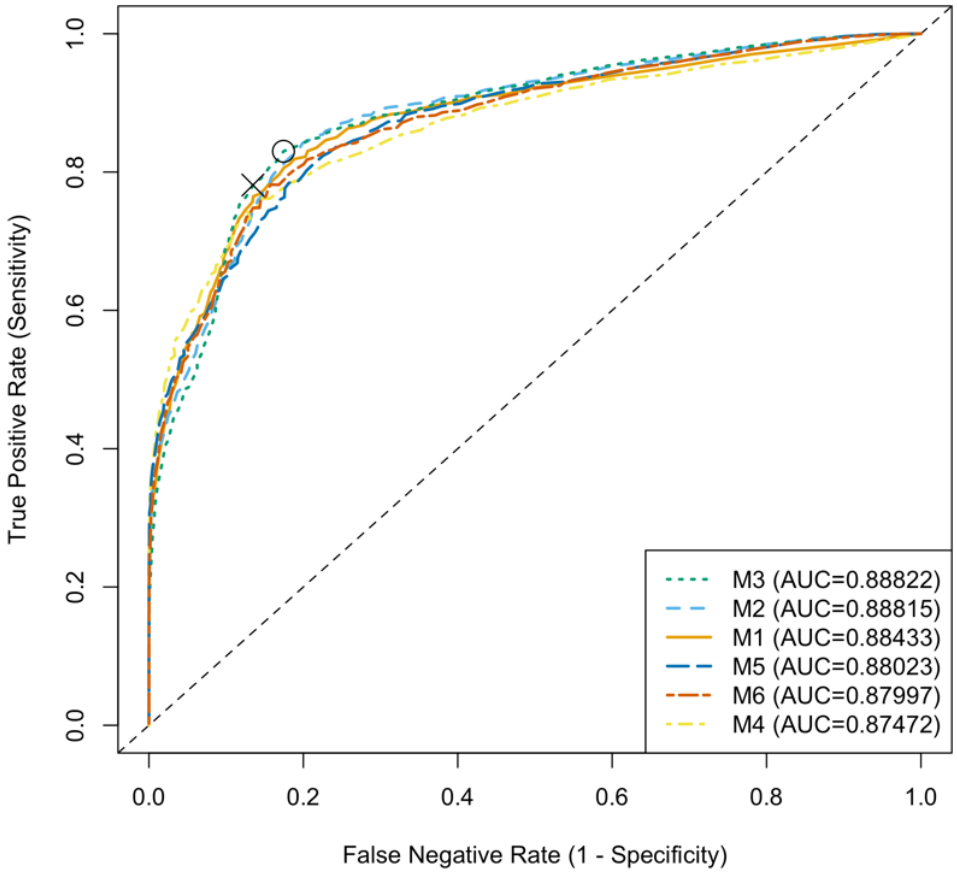

Figure 1.10 The receiver operating characteristic curve for predicting the fixation of new amino acids. A new amino acid was predicted to become fixed when P_{fix} is higher than threshold τ. The receiver operating characteristic curve for each model was obtained by varying the threshold τ. The cross represents the point where τ equals 0.5 for M3. The distance from the diagonal line to the curve of M3 reaches a maximum length when τ equals 0.43 (circle). Points around the lower left corner correspond to cross-validation tests using $\tau \approx 1$ and points around the upper right correspond to cross-validation tests using $\tau \approx 0$.

1.4.7. The fixation probability of a new amino acid on HA

Figure 1.11A shows the three-dimensional surface plot of the fixation probability of a new amino acid on HA based on model M3 with its parameters in Table 1.5. The fixation probability of a new amino acid increases as its frequency increases, as expected from the property of Equation 1.5. The fixation probability starts from zero when the frequency equals zero, as shown in green, and it approaches one when the frequency approaches one, shown in blue. The fixation probability also increases as the z-value of the rank-sum of patient ages of the new amino acid becomes larger, as one can observe an increase in height when looking at a band of the same color in the increasing direction of the z-value of the age rank-sum. This result indicates that viruses with a new amino acid on HA obtains additional chance to become fixed, when they can infect elderly patients more effectively than the viruses with the old amino acid at the same position. In other words, new variants' excess infectivity to adult population over old variants increases their chances to become fixed in addition to its chance of fixation gained from how large a fraction of the population they are currently infecting. Figure 1.11B shows the same information as Figure 1.11A in a two-dimensional presentation. The lines in Figure 1.11B represent the fixation probabilities at values of $a.wilcox$ from –5 to 5 with each increasing step of 1.

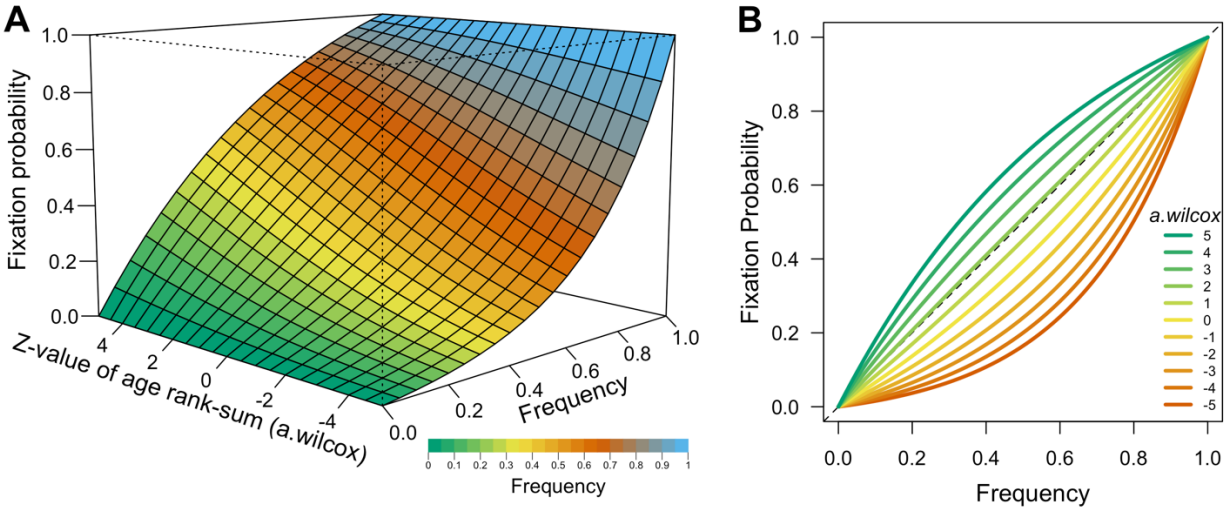

Figure 1.11 The fixation probability of viruses with a new amino acid on HA.(A) Three-dimensional surface plot of the fixation probability of viruses with a new amino acid on HA. X-axis and color represent the frequency of virus having the new amino acid in the viral population. Green represents frequencies close to zero, blue nearly reaching one, yellow and orange in-between. Y-axis represents the fixation probability of the new amino acid. Z-axis represents its $a.wilcox$, calculated by z-value of its patient age rank-sum compared to the old amino acid, representing the excess infectivity to adult population of viruses having the new amino acid compared to viruses with the old amino acid at the same position on HA. (B) The relationship between the frequency of a new amino acid on HA and fixation probability when the new virus infects different age groups compared to the old virus. X-axis represents the frequency of virus having the new amino acid in the viral population. Y-axis represents the fixation probability of the new amino acid. Color represents its $a.wilcox$, calculated by z-value of its patient age rank-sum compared to the old amino acid, representing the excess infectivity to adult population of viruses having the new amino acid compared to viruses with the old amino acid at the same position on HA.

1.4.8. Timing of selection of new amino acids in different birth-year groups

Figure 1.12 shows the time evolution of frequencies of amino acid sequences having the nineteen fixed new amino acids on HA in different birth-year groups during their transition phases. Panels A to S in the figure correspond to amino acid substitutions shown in Table 1.2 in the same order. The patients were assumed that they were firstly exposed to the dominant strain of the H1N1 viruses circulating in the year when they were born.

The transitions from old amino acids to new amino acids showed different timings of emergence and fixation depending on birth-year groups (Figure 1.12). Amino acid T at position 203 on HA have had frequencies of more than 0.30 in all birth-year groups since the first four-month sliding window starting from March 2009 (Figure 1.13A). The next three fixed new amino acids, lysine (K) at position 374, asparagine (N) at position 451, and T at position 185 exceeded a frequency of 0.10 firstly in the youngest birth-year groups followed by others (Figure 1.14B, 1.14C, and 1.14D). Amino acid N at position 97 exceeded a frequency of 0.10 and 0.30 firstly in the second oldest birth-year groups (Figure 1.14E and 1.14E). However, the tendency is not clear because of the drop in the frequency of new amino acid in the middle of its transition phase (Figure 1.12E).

After 2011, a general tendency that new amino acids exceeded a frequency of 0.30 earliest in the old and middle-aged birth-year groups was observed (Figure 1.13). These include K at position 499 (Figure 1.13F), glutamic acid (E) at position 283 (Figure 1.13G), glutamine (Q) at position 163 (Figure 1.13H), T at position 256 (Figure 1.13I), N at position 84 (Figure 1.13J), T at position 216 (Figure 1.13K), N at position 162 (Figure 1.13L), valine (V) at position 295 (Figure 1.13M), R at position 74 (Figure 1.13N), T at position 164 (Figure 1.13O), proline (P) at position 183 (Figure 1.13P), aspartic acid (D) at position 260 (Figure 1.13Q), isoleucine (I) at position 185 (Figure 1.13R), and D at position 129 (Figure 1.13S).

Figure 1.15 shows a clear trend that the fixation starts from old birth-year groups, followed by the middle-aged birth-year groups and ended with the young birth year groups for all the nineteen fixed amino acids (See Materials and Methods for the definition of fixation in this study). Despite this general tendency, three new amino acids became fixed quite early in the youngest birth-year group (Figure 1.15B, 1.15C, 1.15D). However, the timing of overturn, when the frequency of a new amino acid exceeds 0.50 in a birth-year group did not show a clear tendency (Figure 1.16).

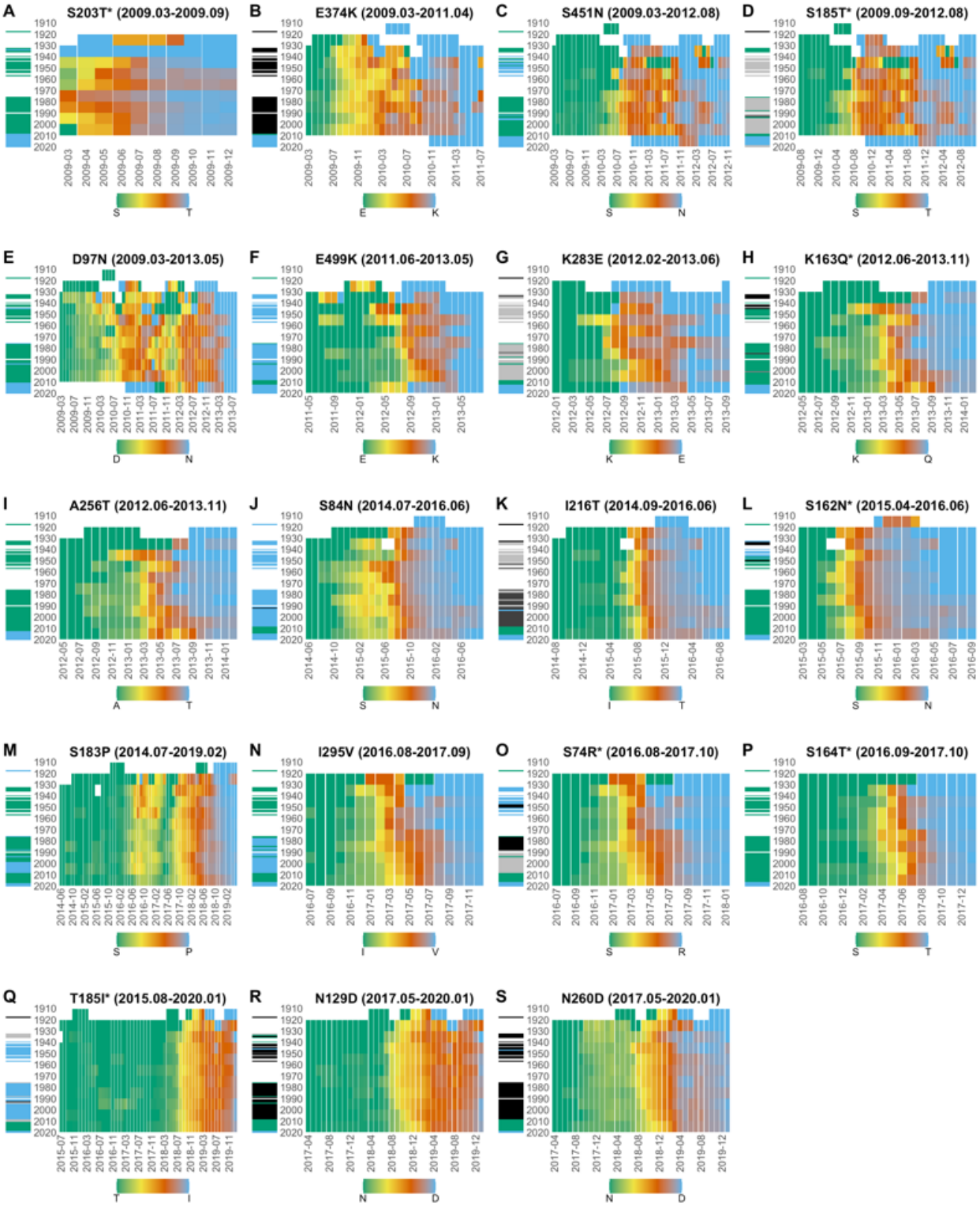

Figure 1.12 Timing of amino acid substitutions in different birth-year groups. (Figure caption continues next page)

Panels A to S in the figure correspond to amino acid substitutions shown in Table 1.2 in the same order. In each panel, X-axis represents the first month of a four-month sliding window, and Y-axis represents the birth-year of patients. The population of patients were grouped into 10-year birth-year groups. Each cell in a heatmap is color-coded according to the frequency of viruses having the fixed new amino acid in the population of a birth-year group at Y-axis in a four-month sliding window starting at the month on the X-axis. A cell is green if the frequency of new amino acid in the birth-year group is zero, and it is blue if the frequency in the birth-year group is one, as shown in the color key in the legend. Cells with no data are represented in white. The horizontal bars on the left of each heatmap represent the dominant amino acid at the corresponding position on HA of viruses circulating in the year on the Y-axis. The color of a bar on the left of each heatmap represents the dominant amino acid at the same position on HA of viruses circulating in the year when patients were born. A green bar indicates the circulation of viruses having the old amino acid at the substituted position on HA in the year when patients were born, and a blue bar indicates the circulation of viruses having the new amino acid at the same position. A bar with grey or black color indicates the circulation of viruses having a dominant amino acid different from both the old and new amino acids at the substituted position in the year when the patients were born. Amino acid substitutions with an asterisk represent substitutions which occurred at an antigenic site.

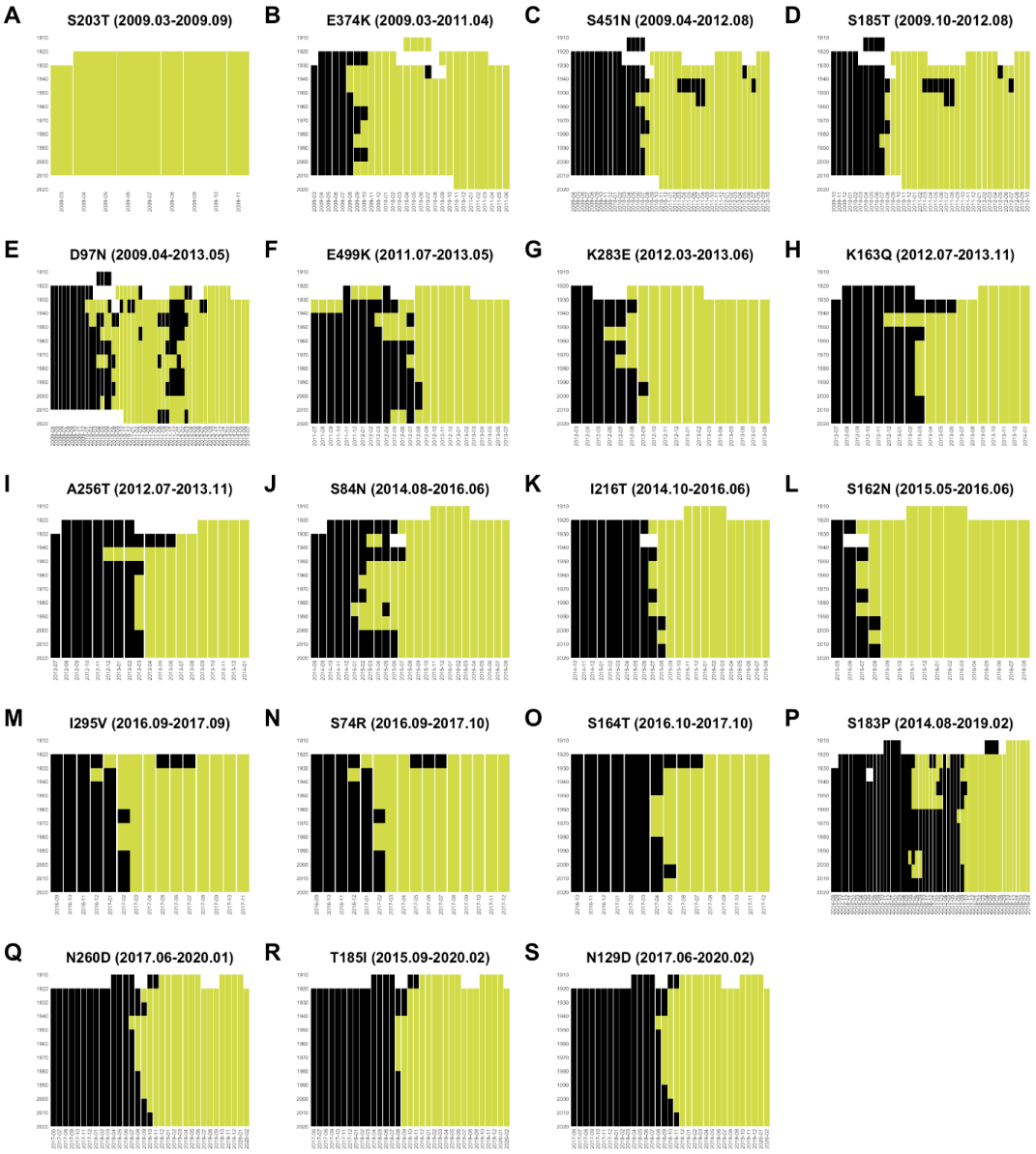

Figure 1.13 Timing when a fixed new amino acid has exceeded a frequency of 0.30 in different birth-year groups. Each cell in a panel corresponds to a cell representing the frequency of a fixed new amino acid shown in its corresponding heatmap in Figure 1.12. A cell is black if the frequency of the new amino acid in a birth-year group is less than or equal to 0.30, and it is lime-green if the frequency in the birth-year group is more than 0.30. Cells with no data are represented in white. X-axis represents the first month of a four-month sliding window, and Y-axis represents the birth-year of patients in the same manner as Figure 1.12. The population of patients were grouped into 10-year birth-year groups.

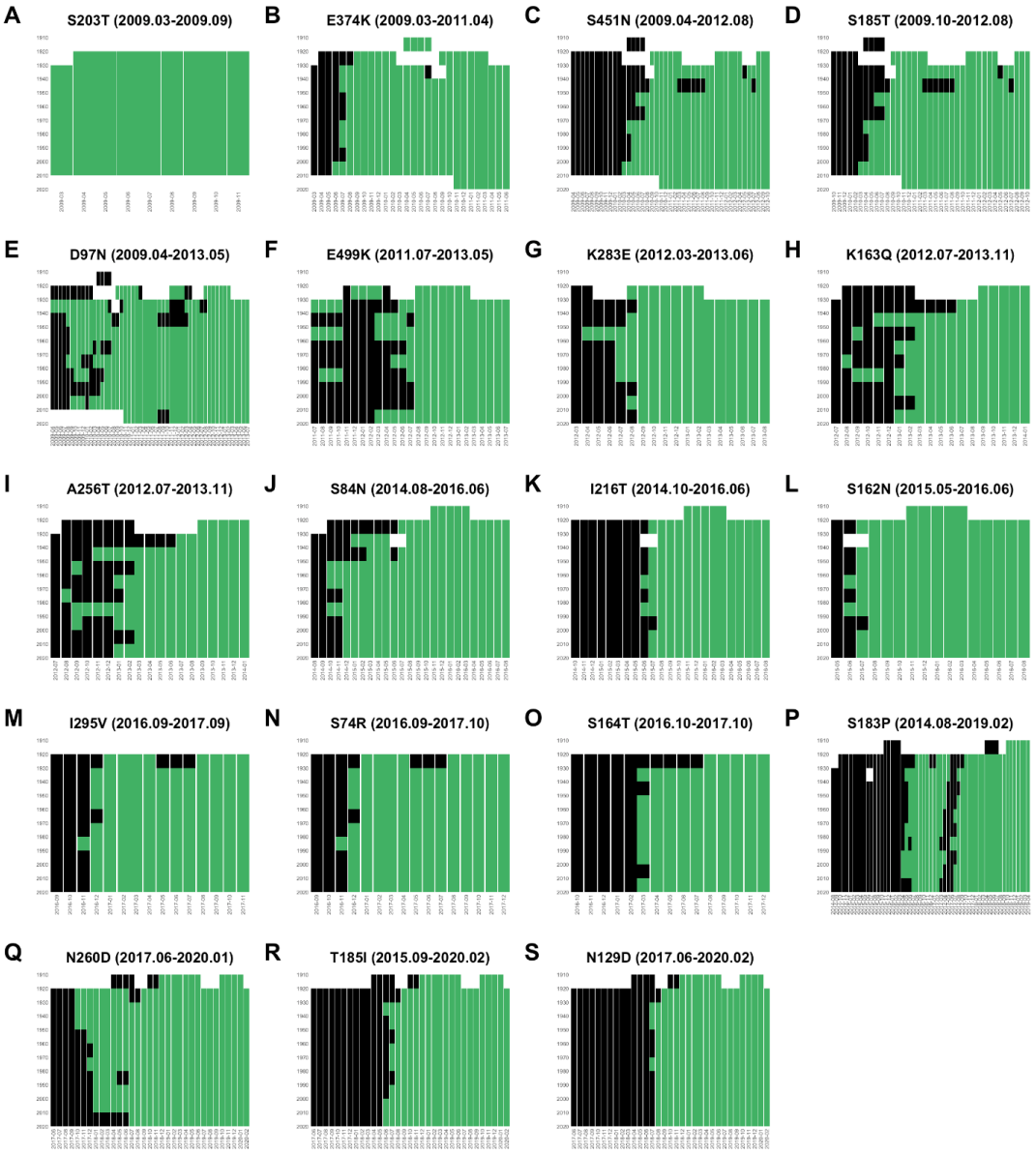

Figure 1.14 Timing when a fixed new amino acid has exceeded a frequency of 0.10 in different birth-year groups. Each cell in a panel corresponds to a cell representing the frequency of a fixed new amino acid shown in its corresponding heatmap in Figure 1.12. A cell is black if the frequency of the new amino acid in a birth-year group is less than or equal to 0.10, and it is green if the frequency in the birth-year group is more than 0.10. Cells with no data are represented in white. X-axis represents the first month of a four-month sliding window, and Y-axis represents the birth-year of patients in the same manner as Figure 1.12. The population of patients were grouped into 10-year birth-year groups.

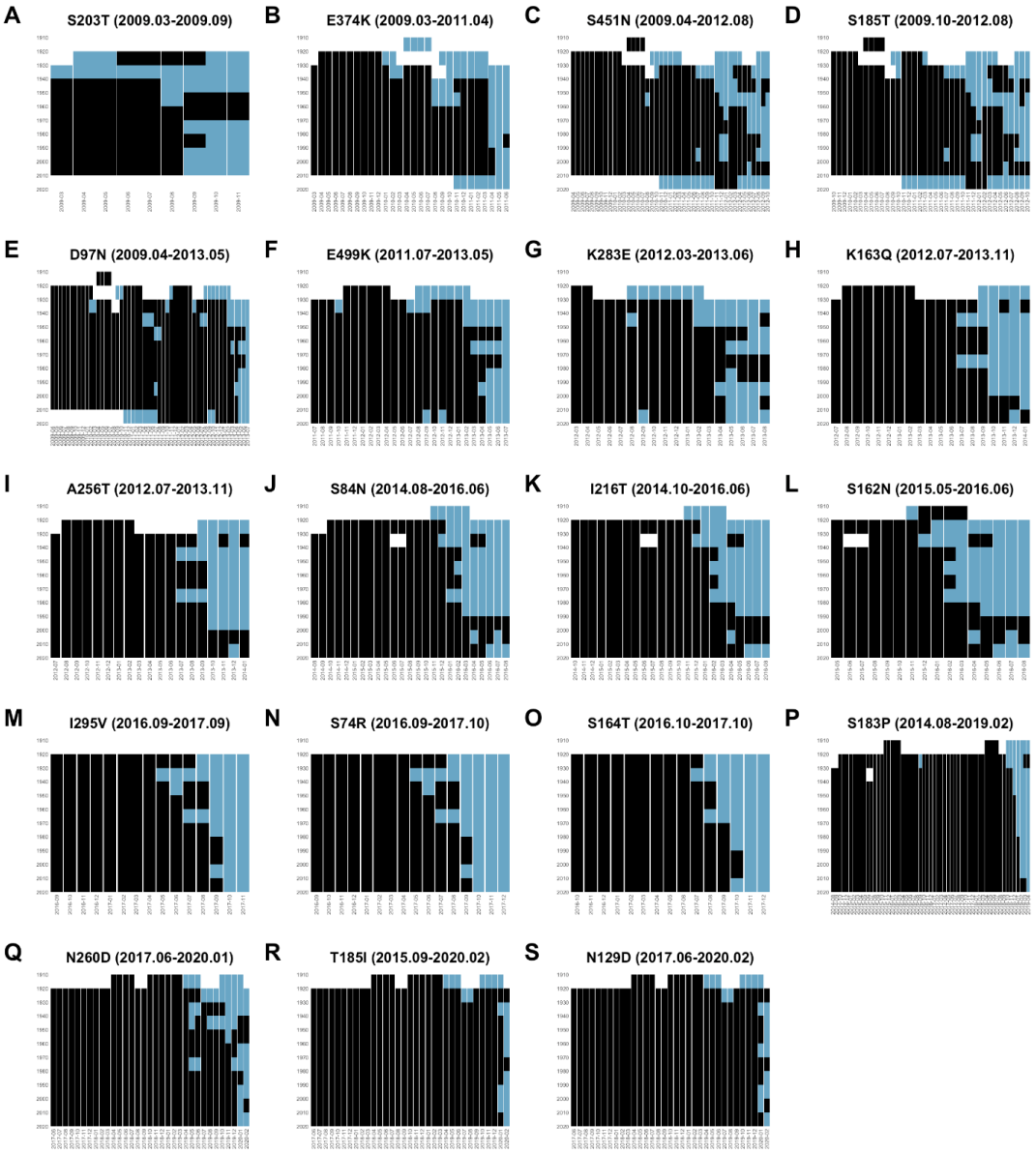

Figure 1.15 Timing when a fixed new amino acid has exceeded a frequency of 0.95 in different birth-year groups. Each cell in a panel corresponds to a cell representing the frequency of a fixed new amino acid shown in its corresponding heatmap in Figure 1.12. A cell is black if the frequency of the new amino acid in a birth-year group is less than or equal to 0.95, and it is blue if the frequency in the birth-year group is more than 0.95. Cells with no data are represented in white. X-axis represents the first month of a four-month sliding window, and Y-axis represents the birth-year of patients in the same manner as Figure 1.12. The population of patients were grouped into 10-year birth-year groups.

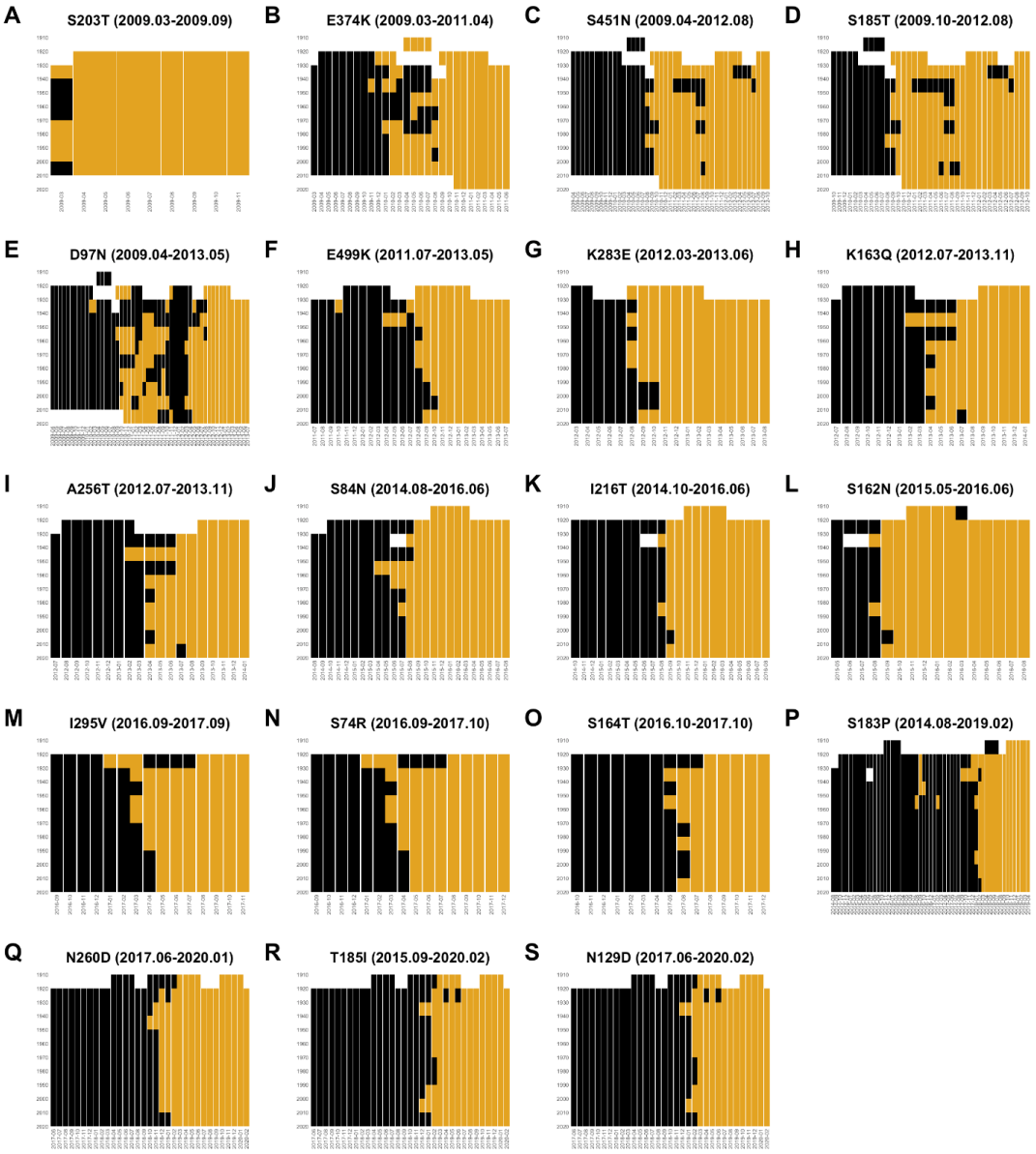

Figure 1.16 Timing when a fixed new amino acid has exceeded a frequency of 0.50 in different birth-year groups. Each cell in a panel corresponds to a cell representing the frequency of a fixed new amino acid shown in its corresponding heatmap in Figure 1.12. A cell is black if the frequency of the new amino acid in a birth-year group is less than or equal to 0.50, and it is orange if the frequency in the birth-year group is more than 0.50. Cells with no data are represented in white. X-axis represents the first month of a four-month sliding window, and Y-axis represents the birth-year of patients in the same manner as Figure 1.12. The population of patients were grouped into 10-year birth-year groups.

Some amino acid substitutions were associated with the dominant amino acids of the viruses circulating in the year when patients were born. For example, the transition from K to Q at position 163 on HA appeared earlier in patients born in 1940–1950 than those born in 1930–1940 (Figure 1.12H). More accurately, the timings when Q at position 163 on HA first exceeded a frequency of 0.30 in patients born in 1940–1950 preceded those born in 1930–1940 by seven months (Figure 1.13H). The dominant amino acid at position 163 on HA of viruses circulating during 1940–1950 was K (Figure 1.12H). In contrast, as shown in the black bars in Figure 1.12H, the dominant amino acid at position 163 on HA of viruses circulating during 1930–1940 was neither K nor Q. Patients born in 1940–1950 may be first exposed to viruses having K at position 163. The substitution from K to Q at position 163 may have selective advantage in patients born in 1940–1950. On the other hand, the birth-year group in 1930–1940 may be first exposed by viruses having a different amino acid other than Q or K at position 163. The viruses having Q at position 163 may not have large advantage compared to viruses having K at this position in birth-year group of 1930–1940. The difference in the timings of amino acid substitutions between the two birth-year groups may be attributed to the different amino acids at this position on HA of viruses that first infected to patients of the two birth-year groups.

Some amino acid substitutions may be associated with the disappearance of H1N1 viruses in the human population during 1957 to 1977, which is the period between the year of the Asian flu pandemic in 1957 and the year of the Russian flu pandemic in 1977. For example, the transition from S to N at position 84 on HA appeared earlier in patients born in 1950–1960 than those born in 1940–1950 (Figure 1.12J). Precisely, the first sliding window in which N at position 84 on HA exceeded a frequency of 0.30 in patients born in 1950–1960 was five months earlier than that in patients born in 1940–1950 (Figure 1.13J). The same tendency can be found for frequencies of 0.10 and 0.50 (Figure 1.14J and Figure 1.16J). The amino acid substitution from S to N at position 84 may have immunological disadvantage because most population have firstly exposed to viruses having N at position 84, as shown in the blue bars in Figure 1.12J. Due to the absence of H1N1 during 1957 to 1977, a considerable number of patients of birth-year groups 1950–1960 and 1960–1970 are likely to be infected first with H2N2 influenza viruses. Viruses having N at position 84 on HA may have less immunological disadvantage in these birth-year groups, resulting in an earlier transition from S to N at this position than the birth-year group of 1940–1950. The delay in transition from K to Q in 163 in patients of birth-year group 1950–1960 can also be explained by the absence of H1N1 strain during 1957 to 1977.

When the whole period of transition was considered, the age distributions of patients infected with viruses having the new amino acids were not significantly different from those of old amino acids at positions 203, 374, 295, 74, and 164 on HA (Table 1.2). However, when observing the timing of amino acid substitutions in different birth-year groups, amino acid substitutions at positions 295, 74, and 164 still followed a general tendency of beginning in the old and middle-aged birth-year groups and ending

in the youngest birth-year groups (Figure 1.12N, 1.12O and 1.12P). This suggests that, even though age distributions of patients may not significantly differ between viruses having new and old amino acids when considering the whole transition period, the new amino acid's distributions of patient ages in each window still tend to be skewed towards people in older birth-year groups for these amino acid substitutions.

1.5. Discussion

In this study, the patient age distributions and fixation probabilities of new amino acids on HA of 2009 pandemic strains of H1N1 influenza viruses were investigated. The empirical probability that a new amino acid on HA later became fixed in the viral population was only 0.004. The empirical fixation probability significantly increased when the frequency of viruses having new amino acids exceeded 0.1. The empirical fixation probability also significantly increased when the viruses having the new amino acids more effectively infected the adult age population from 25 to 35 years old than the viruses with old amino acids at the same position. Based on these observations, fixation probability of a new amino acid was modeled using Kimura's formula of advantageous selection. The selective advantage of a new amino acid was modeled by a linear combination of patient age distributions and the involvement on antigenic sites. The parameters of models were estimated by maximizing the likelihood of parameters from profiles of fixed and extinct new amino acids from 2009 to 2020. Four-fold cross-validation tests revealed that the model using the difference in patient age distribution and frequency of new amino acid predicted amino acid substitutions on HA with a sensitivity of 0.78, specificity of 0.86, and precision of 0.83.

When looking at trajectories of new amino acids that emerged on HA of H1N1 viruses from March 2009 to May 2020, the empirical fixation probability of a new amino acid was only 0.004 (Figure 1.6). It means that 99.6% of the new amino acids that appeared on HA went extinct. If we look at the relationship between the frequency of a new amino acid and its empirical fixation probability in a four-month sliding window, the empirical fixation probability became different, because the fixed new amino acids can be counted multiple times in different four-month sliding windows. A new amino acid having a frequency no more than 0.11 in a four-month sliding window had an empirical fixation probability of 0.02 (Figure 1.2B). This means that a new amino acid having a frequency below 0.11 in a four-month sliding window had almost no chance of becoming fixed later. In contrast, a new amino acid having frequencies more than 0.11 in a four-month sliding window had an empirical fixation probability close to 0.50 (Figure 1.2A). It means that a new amino acid having a frequency more than 0.11 in a four-month sliding window would have equal chance of becoming either fixed or extinct. Thus, the prediction of fixation of a new amino acid having a frequency more than 0.11 in a four-month sliding window, which have an empirical probability close to 0.50, is the most difficult setting to predict the

fixation of new amino acids. For this reason, the study focused on the prediction of the fixation of new amino acids which exceeded a frequency of 0.11 in a four-month sliding window.

Results found that the frequency of new amino acids alone can achieve high sensitivity, specificity, and precision in predicting the fixation of a new amino acid of which frequency is more than 0.11 in a four-month sliding window. Model M1, which modeled the fixation probability of a new amino acid using its current frequency under the assumption of a constant selective advantage, predicted the fixation of a new amino acid with an average sensitivity of 0.76, specificity of 0.86, and precision of 0.83 in four-fold cross-validations (Table 1.7). This result suggested that the fixation probability of a new amino acid is largely attributed to its frequency. The constant for the selective advantage, C_0, was estimated to be 0.313 with its confidence intervals of from 0.207 to 0.421 by the maximum likelihood method (Table 1.5). Since a positive coefficient for Kimura's formula indicates advantageous selection, it can be concluded that viruses with a new amino acid having a frequency higher than 0.11 in a four-month sliding window has a significantly higher selective advantage compared to viruses with the old amino acid at the same position.

The predictability of the fixation of a new amino acid was significantly improved by considering the Z-value of patient age rank-sums of new amino acids compared to the constant selective advantage model in cross-validation tests. Youden's index, which is the sum of sensitivity and specificity minus one, was significantly improved in model M3 from model M1 (Table 1.7). The coefficient for $a.wilcox$, C_a, was estimated to be 0.119 with its confidence intervals of from 0.065 to 0.175 by the maximum likelihood method (Table 1.5). Since $a.wilcox$ represents the excess infectivity to the adult population of viruses having new amino acids compared to those having old amino acids, this result suggests that inclusion of age statistics of viruses significantly improved the prediction of the fixation of a new amino acid. This result is consistent with the current understanding of the mechanism of evolution of influenza viruses, in which new strains are selected by the immunity of people who were infected with and recovered from strains circulating previously (Ferguson et al., 2003).

The proportion of adults who have been infected with influenza viruses is higher than that of children, because adults have more chance of being exposed to the viruses due to their longer time since birth compared to children. Therefore, viruses having different antigenicity from viruses which have been circulating before can have a higher advantage in adult population than in child population. In other words, the advantage of a new amino acid that alter the antigenicity of HA over its old amino acid in the child population can be lower compared to the advantage in adult population. It is likely that this is the main reason why a statistic of patient age distributions improved the accuracy of the prediction of amino acid substitutions. In addition to this straightforward interpretation, another explanation can be considered. Viruses infecting adults are more likely to be spread globally than children as adults are more likely to travel long distances (Bedford et al., 2015). This is an alternative explanation of the

phenomena, but a clear trend that the fixation starts from old birth-year groups, followed by the middle-aged birth-year groups and ended with the young birth year groups (Figure 1.15) supports the first interpretation that the higher selective advantage in adult population is attributed to the immunity from previous infections.

Since influenza viruses are transmitted among individuals of different age groups, the difference in the age distributions between new amino acids and old amino acids were not supposed to differ largely. As shown in Figure 1.1B, the distributions of patient ages of GISAID sequences are bimodal, with one mode in the children population younger than 15 years old, and another mode in the adult population older than 15, especially after 2012. The number of fixed and extinct new amino acid profiles in Table 1.3 is dependent on whether their median patient ages are between 25 and 35 or not ($p < 10^{-15}$ with χ^2 test). Furthermore, the number of fixed and extinct new amino acid profiles in Table 1.4 is dependent on whether the old amino acids have a median patient age less than or equal to 15 or not ($p < 10^{-9}$ with χ^2 test). The most probable hypothesis for the bimodal distribution is as the following. In the younger population who haven't experienced influenza infections, the variants with old amino acids can infect as effectively as variants with new amino acids. The number of infections in the younger population decreases as patient age increases because of acquisition of immunity by the first exposure to influenza viruses. In the adult population who has experienced previous exposures, on the other hand, variants having new amino acids is more infectious than those having old amino acids because of the original antigenic sin. The first mode in the patient age distribution is formed by the first influenza infection in life and the second mode was formed from the second or subsequent influenza infections. This is my best explanation for the results obtained in this study.

It is known that the 2009 pandemic strain shows cross-reactivity with the Spanish flu and Russian flu strains (Garten et al., 2009; Itoh et al., 2009). An individual's immunity profile against influenza is highly affected by their first infection in their childhood (Francis et al., 1947). Some strains having a new amino acid on HA seemed to have an advantage in infecting patients who were infected with the viruses having the old amino acid in their first infection. Examples of these amino acid substitutions include K163Q (Figure 1.12H). Some amino acid substitutions may be associated with the disappearance of H1N1 viruses in the human population during 1957 to 1977, which is the period between the year of the Asian flu pandemic in 1957, caused by a strain of H2N2 viruses, and the year of the Russian flu pandemic in 1977, caused by a strain of H1N1 viruses. A considerable number of patients of birth-year groups 1950–1960 and 1960–1970 are likely to be firstly infected with H2N2 influenza viruses. Viruses having the S84N substitution have less immunological disadvantage in these birth-year groups compared to the birth-year group of 1940–1950, resulting in the different timings of transition from S to N (Figure 1.12J). Selections of these strains can be explained as an effect of the original antigenic sin.

Results showed that epitope flags of substituted positions did not largely contribute to the prediction of amino acid substitutions in cross-validation tests. From nineteen fixed amino acid substitutions on HA observed in this study, only seven (36.84%) occurred on its antigenic sites (Table 1.2). It has been suggested that amino acid substitutions on non-antigenic sites compensate the fitness cost of substitutions on antigenic sites (Kryazhimskiy et al., 2011; Koel et al., 2013; Yokoyama et al., 2017). However, 90.5% of amino acid substitutions on the HA1 domain of H3N2 viruses were known to have occurred at its antigenic sites (Shih et al., 2007). A possible reason for the small contribution of epitope flags in prediction is that the positions of antigenic sites used in this study have been determined from H1N1 viruses before the 2009 pandemic (Igarashi et al., 2010). The antigenic sites for the 2009 pandemic strain may differ from the antigenic sites for previous seasonal strains circulating before 2009 pandemic. In fact, Ren et al. showed that antigenic regions cover a larger area than regions previously defined as the antigenic sites (Ren et al., 2015). The same was true for H3N2 (Lees et al., 2010). Further studies are required for a wider characterization of antigenic sites on HA of influenza viruses.

The human influenza shows seasonality, and the population of the viruses fluctuates depending on the time of year. Although the assumption of constant effective population size may not be valid for the population genetics of seasonal influenza viruses, this Kimura's formula was used under the assumption of constant effective population size for its simplicity. It is suggested that the fixation probability increases when the effective population size is growing (Fisher, 1930). This means the fixation probability predicted from the model would be underestimated during influenza seasons when the number of new cases is growing. Even so, the model has achieved high predictability in cross-validation tests, indicating that the error may be marginal and an acceptable trade-off for the model's simplicity. However, for more precise predictions, the method may adopt fixation probability models which take into account changing population sizes (Lambert, 2006).

Synchronized substitutions were observed at positions 451 and 185, positions 499 and 283, positions 163 and 256, positions 84, 216, and 162, positions 295 and 74, and positions 185 and 129 (Figure 1.3). Fixations of the synchronized amino acids may be hitchhiking substitutions, which do not contribute to the increase in viral fitness but became fixed due to the selective advantage gained from another substitution on HA of the same strain (Barton, 2000; Smith et al., 2004). For example, transitions from S to N at position 84, I to T at position 216, and S to N at position 162 occurred simultaneously (Figure 1.3). H1N1 strains circulating before the 2009 pandemic had S at position 162 on their HA (Figure 1.12L). Since position 162 is located in the antigenic site Sa (Table 1.2), variants having N at this position may have had selective advantage over variants having S at this position. In contrast, S at position 84 and I at position 216 of the 2009 pandemic strain were different from amino acids at these positions on HA of H1N1 strain circulating before the 2009 pandemic (Figure 1.12J and 1.12K). These two substitutions may not have a selective advantage in terms of antigenicity, and there

is a possibility of hitchhiking substitutions of S162N. Another explanation of the synchronized transitions of amino acids is that the fixations can occur through synergistic epistasis between several mutations (Neverov et al., 2015). Viruses with a new amino acid with slow transition may initially lack large advantage over viruses with an old amino acid at the same position. These viruses became fixed when they gained a synergistic advantage from another new amino acid on HA. For example, the slow transition from D to N at position 97 have become fixed when viruses have additional new amino acids at positions 499 and 283.

One of the limitations of this study's method is that the model can only predict the evolutionary outcome of new amino acids. Thus, the model cannot predict the time it takes before they became fixed in viral population. Each year, WHO makes recommendations for vaccine strains by reviewing the circulation and spread of new strains through their global influenza surveillance network (Russell et al., 2008). The recommendation of vaccine strains must be decided eight months before the season starts for the vaccine development and production process (World Health Organization, 2007). The method can predict the fixation of a new amino acid accurately once its frequency exceeds 0.11. According to the dataset used in this study, the time for a new amino acid to become fixed or extinct after exceeding a frequency of 0.11 had a mean of 18.8 months with a standard deviation of 13.6 months. Assuming that the time to fixation or extinction after exceeding a frequency of 0.11 follows a normal distribution with a mean of 18.8 months and a standard deviation of 13.6 months, predictions by the model can be obtained eight months earlier than its fixation or extinction for 79% of new amino acids that exceed a frequency of 0.11 in a four-month sliding window. Thus, the applicability of this study's method to the actual vaccine selection process is not largely restricted by the limitation due to the lack of a mechanism for predicting the timing of fixation.

The applicability of the method to H3N2 viruses should be tested in the future. Most studies to predict amino acid substitutions have targeted H3N2 viruses as sequence data are available from its emergence in 1968 (Agor & Ozaltin, 2018; Klingen et al., 2018). H1N1 viruses emerged in the Spanish flu pandemic in 1918 (Cohen, 2010), disappeared in 1957, and re-emerged in the Russian flu in 1977, and were replaced with a swine flu strain in the 2009 pandemic (Girard et al., 2010). In contrast, H3N2 viruses have been circulating in the human population since its pandemic in 1968. The structure of the population having immunity against H3N2 viruses may be simpler than that of H1N1 viruses. However, due to the limitation of patient age information of amino acid sequences of past H3N2 viruses, the method can be applicable only to the fixation of new amino acids that have been observed recently.

Chapter 2. Predicting the trajectory of replacements of SARS-CoV-2 variants using relative reproduction numbers

2.1. Summary

New variants of SARS-CoV-2 with high effective reproduction numbers are continuously being selected by natural selection. To establish effective control measures for new variants, it is crucial to know their transmissibility and timing of replacement in advance. In this chapter, retrospective prediction tests for the variant replacement from Alpha to Delta of SARS-CoV-2 in England were conducted using the relative reproduction numbers of Delta with respect to Alpha estimated from partial observations. Results showed that once Delta's relative frequency reached 0.15, the date when relative frequency of Delta would reach 0.90 was predicted with maximum absolute prediction errors of 3 days. This means that the time course of the variant replacement could be accurately predicted from early observations. Together with the estimated relative reproduction number of a new variant with respect to old variants, the predicted replacement timing will be crucial information for planning control strategies against the new variant.

2.2. Introduction

Since its first emergence in the human population in 2019, SARS-CoV-2 has been generating new variants. Natural selection favors new variants that has higher effective reproduction numbers than other circulating variants. As a result, the average transmissibility in the viral population increases over time (Tao et al., 2021). The emergence and replacement among variants of concern (VOCs), Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), and Omicron (B.1.1.529) (World Health Organization, 2022) are the observed process of natural selection.

It is important to know the transmissibility of new variants in comparison with previously circulating variants because the average reproduction number of the circulating virus changes when new variants become dominant. Several studies have analyzed the reproduction numbers of new variants that have replaced old ones. Volz et al. estimated the effective reproduction number of Alpha in England to be 1.5–2.0 times higher than that of non-VOCs using a logistic growth model for relative variant frequencies (Volz et al., 2021). Leung et al. estimated the basic reproduction number of Alpha to be 1.75 times higher than that of previously circulating variants in England using a renewal-equation-based model (Leung et al., 2021). Ito et al. estimated the effective reproduction number of Delta to be 1.35 times higher than that of Alpha from relative variant frequencies observed in Japan by using an approximated version of the renewal-equation-based model (Ito et al., 2021a). Using the same method, Ito et al. estimated the effective reproduction number of Omicron to be 3.15 times higher than that of Delta in Denmark (Ito et al., 2021b), and Nishiura et al. estimated the effective reproduction of Omicron to be 4.2 times higher than that of Delta in South Africa (Nishiura et al., 2021).

In order to prepare control measures against new variants, it is crucial to predict the trajectory of the variant replacements in advance. The prediction of variant selection has been widely studied in seasonal influenza viruses (Morris et al., 2018). Łuksza and Lässig developed a fitness model using mutations on antigenic sites and non-antigenic sites to predict selected variants (Łuksza & Lässig, 2014). Huddleston et al. predicted the future relative frequency of variants using its current relative frequency, the novelty of antigenic sites, and the mutational load in non-antigenic sites (Huddleston et al., 2020). Chapter 1 of this dissertation modeled the fixation probability of variants using relative variant frequency and statistics on patient ages (Piantham & Ito, 2021). In the case of seasonal influenza, the main driving force of natural selection was the population immunity acquired from previous infections. In contrast, most of the human population were considered naïve to SARS-CoV-2 at the beginning of the pandemic, and a method to predict the trajectory of variant replacements in the early stage of the pandemic can be simpler than those assuming pre-existing immunity from previous infections.

The transmissibility of an infectious agent can be measured by its reproduction number. The effective reproduction number at time t (R_t) is defined as the average number of infections that

someone infected at time t could expect to produce if conditions should remain unchanged (Fraser, 2007). When more than one variant of the infectious agent is circulating, the relative reproduction number can be used to measure the relative transmissibility of a variant compared to a baseline variant (Leung et al., 2017; Leung et al., 2021). However, the method requires the numbers of new cases in addition to the relative frequencies of variants, and it is not applicable for predicting variant replacement in the future. Using approximations, Ito et al. proposed a method to determine the relative reproduction number without knowing the number of new cases (Ito et al., 2021a). This method allows us to predict the future trajectory of variant replacements.

Nucleotide sequences of SARS-CoV-2 variants have been collected worldwide and accumulated in the GISAID database (Shu & McCauley, 2017). It is known that different geographical locations have different distributions of variants (Hadfield et al., 2018). As of 28 September 2022, a total of 13,283,666 sequences have been registered on the database worldwide. Of these, 2,286,890 (17.2%) were submitted from England, which has their population account for 0.71% of the world population. These numbers indicate that England is one of the locations having highest sequencing capacity. In England, the Alpha–Delta replacement was observed from March 2021 to June 2021. The sequence information during the Alpha–Delta replacement in England is one of the best datasets to evaluate the predictability of variant replacement in SARS-CoV-2.

In this Chapter, retrospective prediction tests were conducted using the nucleotide sequences collected in England during the Alpha–Delta replacement. For each given time point, partial sequence data observed only up to that time point were used to estimate the relative reproduction number of Delta with respect to (w.r.t.) Alpha. The estimated relative reproduction number is then used to predict the future trajectory of variant replacement. The estimated relative reproduction numbers and the predicted trajectories are evaluated by being compared to those estimated using the entire dataset.

2.2.1. Sequence data

Nucleotide sequences of SARS-CoV-2 viruses collected from England during 1 January 2021 to 31 July 2021 were downloaded from the GISAID database on 16 November 2021. Of these, 411,123 sequences had complete information about date of sample collection in the metadata. The PANGO lineage names (Rambaut et al., 2020) of these sequences were collected from metadata and recorded with their collection dates. Sequences that are labeled as "B.1.1.7" or sublineage names starting with "Q." were classified as the Alpha variant. Sequences that are labeled as "B.1.617.2" or sublineage names starting with "AY." were classified as the Delta variant. There were 11,773 sequences (2.9%) of lineages other than Alpha and Delta, and these were ignored in subsequent analyses. A total of 399,350 sequences of Alpha (192,250) and Delta (207,100) were used for counting the daily numbers of sequences belonging to Alpha and Delta (Figure 2.1).

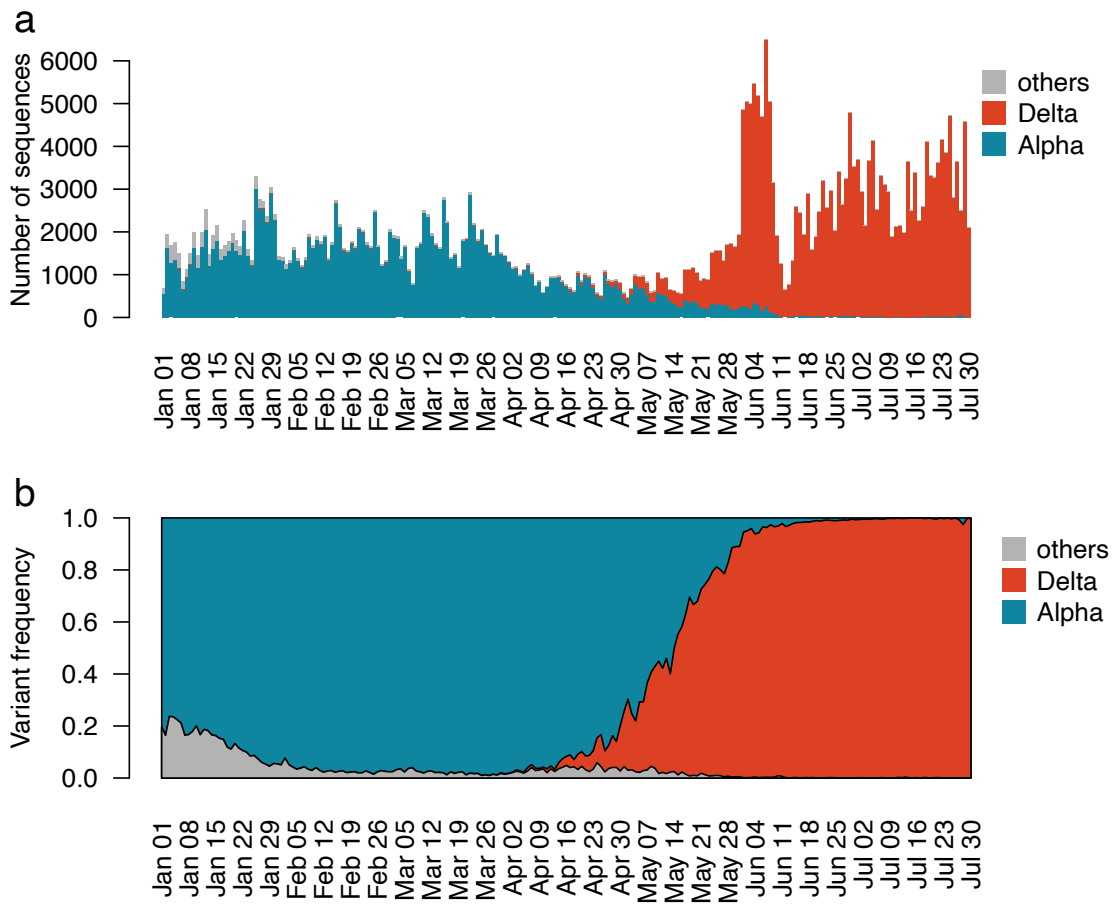

Figure 2.1 Daily variant frequencies in England from 1 January to 31 July 2021. Frequencies of Alpha (red), Delta (blue), and other variants (gray) were calculated from nucleotide sequences corresponding to those variants which were submitted in England on the GISAID database.

2.2.2. Model of advantageous selection

The relative reproduction number of a variant w.r.t. a baseline variant were estimated using an approximated version of the renewal-equation-based model (Ito et al., 2021a). Let X and Y represent variants circulating in the population and $q_X(t)$ and $q_Y(t)$ denote relative frequencies of X and Y at calendar time t, respectively. Suppose that variant X was dominant at t_0 and variant Y was introduced into the population at that time with an initial relative frequency of $q_Y(t_0)$. Assume that the effective reproduction number of variant Y was k times higher than that of variant X and that k is constant over time. Let $f(\tau)$ be the probability density function of generation time τ for SARS-CoV-2 infections. In this study, $f(\tau)$ was assumed to follow the gamma distribution with a shape parameter of 3.42 and a scale parameter of 1.36 (Nishiura et al., 2020). The gamma distribution $f(\tau)$ was discretized to $g(j) = \int_{j-1}^{j} f(\tau)d\tau$ for $1 \leq j \leq 19$. The generation time distributions were truncated at $\tau > 20$ and set $g(20) = \int_{19}^{\infty} f(\tau)d\tau$ so that $\sum_{j=1}^{20} g(j) = 1$. Let $I(t)$ be the total number of new infections by either X or Y at calendar time t. Based on Fraser's time-since-infection model (Fraser, 2007), the effective reproduction numbers of variant X and Y can be calculated as

$$R_X(t) = \frac{q_X(t)I(t)}{\sum_{j=1}^{20} g(j)q_X(t-j)I(t-j)} \tag{2.1}$$

and

$$R_Y(t) = \frac{q_Y(t)I(t)}{\sum_{j=1}^{20} g(j)q_Y(t-j)I(t-j)}. \tag{2.2}$$

Since the effective reproduction number of variant Y is k times higher than that of variant X, the effective reproduction number of variant Y at time t is given by

$$R_Y(t) = kR_X(t). \tag{2.3}$$

Assuming that the viral population at time t comprises of only variants X and Y, the relative frequency of variant Y at calendar time t, $q_Y(t)$, can be calculated as

$$q_Y(t) = \frac{q_Y(t)I(t)}{q_X(t)I(t) + q_Y(t)I(t)}. \tag{2.4}$$

The numbers of new infections were assumed to not vary greatly for 20 days, i.e.

$$I(t-1) \cong \cdots \cong I(t-20), \tag{2.5}$$

for $t > t_0$. In our previous publication using SARS-CoV-2 data from Denmark, models using Equation 2.5 were compared to models not using Equation 2.5. As a result, models using approximation with Equation 2.5 had lower AIC than their corresponding models without the approximation, suggesting that approximation using Equation 2.5 gives a better model than that without Equation 2.5 by eliminating noise in observed $I(t)$ (Ito et al., 2022). Using this approximation with Equations 2.1, 2.2, and 2.3, Equation 2.4 can be rewritten using $q_Y(t-j)$ for $1 \leq j \leq 20$ as

$$q_Y(t) = \frac{k\sum_{j=1}^{20} g(j)q_Y(t-j)}{\sum_{j=1}^{20} g(j)q_X(t-j) + k\sum_{j=1}^{20} g(j)q_Y(t-j)}. \tag{2.6}$$

The average reproduction number of circulating viruses can be determined by the expected value of reproduction numbers of circulating variants. Since relative reproduction number of X is 1, and that of Y is k, the average relative reproduction number of circulating viruses at time t w.r.t. variant X is given by

$$q_X(t) + kq_Y(t). \tag{2.7}$$

2.2.3. Parameter estimation from the number of sequences

Let $N_X(t)$ and $N_Y(t)$ be the number of sequences of variant X and Y observed at calendar time t, respectively. Suppose that variant Y is sampled and sequenced following a beta-binomial distribution having distribution parameters of $\alpha = q_Y(t)M$ and $\beta = (1-q_Y(t))M$, where $M = \alpha + \beta$. The parameter M represents the sum of the two shape parameters of the underlying beta distribution and it determines how proportions of variants vary during sampling. Note that this beta-binomial distribution has a mean of $\big(N_X(t) + N_Y(t)\big)q_Y(t)$ and a variance of $\frac{(N_X(t)+N_Y(t))q_X(t)q_Y(t)(N_X(t)+N_Y(t)+M)}{M+1}$. To reduce computational time, the upper limit of M is set to 2000. When $q_Y(t) = 0.5$ and $M = 2000$, the first and third quartiles of the beta distribution are 0.492 and 0.508, respectively. The beta-binomial distribution becomes the binomial distribution when $M = \infty$. The following equation gives the likelihood function of parameters k, $q_Y(t_0)$, and M for observing $N_X(t)$ and $N_Y(t)$ sequences of variants X and Y at calendar time t:

$$L\big(k, q_Y(t_0), M\ ;\ N_X(t), N_Y(t)\big) = \binom{N_X(t) + N_Y(t)}{N_Y(t)} \frac{B(N_Y(t) + \alpha, N_X(t) + \beta)}{B(\alpha, \beta)} \qquad (2.7)$$

where $\alpha = q_Y(t)M$, $\beta = (1 - q_Y(t))M$, and $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$. The likelihood for observing $N_Y(t)$ sequences of variant Y during the period on calendar times $t_1,\ldots,t_n$ is given by the product of the above formula for $1 \leq t \leq n$.

Alpha and Delta were considered as variants X and Y, respectively. $N_X(t)$ and $N_Y(t)$ are the numbers of sequences of Alpha and Delta in England at calendar time t, respectively. The date of first introduction of Delta, t_0, was set to the first date when $N_y(t) > 1$ (18 March 2022). The estimates of k, $q_Y(t_0)$, M, and $q_Y(t)$ were obtained by maximizing the likelihood function from $t = t_0$ until the latest t in which $q_Y(t) < 1$ (4 July 2021). The 95% confidence intervals (95% CI) of k, $q_Y(t_0)$, and M were determined using the profile likelihood method (Held & Bové, 2020). From the maximum likelihood estimates of k and $q_Y(t)$, the average relative reproduction number of circulating viruses w.r.t. Alpha at time t was estimated from Equation (7). The 95% CIs of $q_Y(t)$ and the average relative reproduction number of circulating viruses w.r.t. Alpha at time t were determined using combinations of parameters within 95% confidence region (Held & Bové, 2020).

2.2.4. Prediction of relative variant frequency and average relative reproduction number

Relative frequencies of Delta and average relative reproduction numbers of circulating viruses w.r.t. Alpha in future were predicted using the maximum likelihood estimates of parameters calculated from early observations. For each proportion $p = 0.05$, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, and 0.95, the calendar times s_p were determined as when the estimated relative frequency $q_Y(s_p)$ exceeded p using the maximum likelihood estimates calculated with the entire observations from 18 March to 4 July 2021. For each date s_p determined above, the maximum likelihood estimates of k, $q_Y(t_0)$, and M were calculated using observations no later than s_p. Relative frequencies of Delta and average relative reproduction numbers of circulating viruses w.r.t. Alpha in future were predicted by substituting k and $q_Y(t_0)$ in Equations 2.5 and 2.6, respectively. The 95% CIs of $q_Y(t)$ and the average relative reproduction number of circulating viruses w.r.t. Alpha at time $t > s_p$ were determined using combinations of parameters within 95% confidence region estimated from observations at time $t \leq s_p$.

2.3. Results

2.3.1. Estimation of relative reproduction number from entire observations

Table 2.1 shows maximum likelihood estimates and their 95% CIs of model parameters calculated from the entire observations in England from 18 March to 4 July 2021. The relative reproduction number (k) of Delta w.r.t. Alpha was estimated to be 1.88 (95% CI: 1.85, 1.91) with a beta-binomial distribution parameter (M) of 288.54 (95% CI: 202.96, 406.26). Each of these estimates are referred to as the 'final estimate' of each parameter.

Figure 2.2(a) shows the observed and estimated relative frequencies of Delta during the Alpha–Delta replacement in England. The blue curve and black curves around the blue curve represent the maximum likelihood estimates and 95% CI of relative frequencies of Delta. The gray area represents 95% equal-tailed intervals of the beta distribution with the parameters $q_Y(t)M$ and $(1 - q_Y(t))M$. Figure 2.2(b) shows the maximum likelihood estimates and 95% CI of the average relative reproduction number of circulating viruses w.r.t. Alpha during the same period. Dashed vertical lines in both panels indicate the dates when relative frequencies of Delta exceeded each 0.05 increment from 0.05 to 0.95 (Table 2.2). It took 47 days for Delta to reach from relative frequencies of 0.05 (21 April 2021) to 0.95 (7 June 2021).

Table 2.1 Parameters estimated using the entire observations during the Alpha–Delta replacement in England

k (95% CI)	$q_Y(t_0)$ (95% CI)	M (95% CI)	Log Likelihood
1.88 (1.85, 1.91)	0.0005 (0.0004, 0.0006)	288.54 (202.96, 406.26)	–431.00

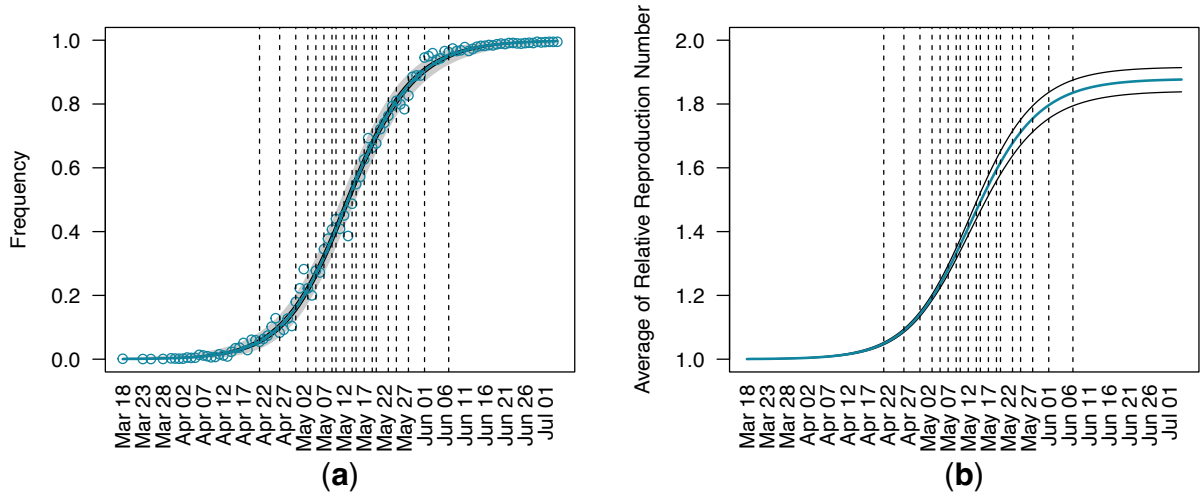

Figure 2.2 Estimated relative frequencies of the Delta variant and average relative reproduction number of circulating viruses with respect to Alpha using entire observations in England from 18 March to 4 July 2021. (a) The observed and estimated relative frequencies of Delta during the Alpha–Delta replacement. Circles represent relative frequencies of Delta sequences collected in England. The blue curve rep-resents the maximum likelihood estimates of relative frequencies of Delta. Black curves sur-rounding the blue curve represent 95% confidence intervals of the estimated relative frequencies of Delta. Gray area represents the 95% equal-tailed interval of beta distribution for the maximum likelihood estimations of parameters of the estimated beta-binomial distribution. (b) The maxi-mum likelihood estimates and 95% confidence intervals of the average relative reproduction number of circulating viruses with respect to Alpha. The blue curve and black curves represent the maximum likelihood estimates and 95% confidence intervals of the average relative reproduction number of circulating viruses with respect to Alpha. Vertical dashed lines in both panels indicate the dates when the estimated relative frequency of Delta reached 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, and 0.95.

Table 2.2 Maximum likelihood estimates for dates when Delta exceeded certain relative frequencies and the average relative reproduction numbers w.r.t. Alpha on those dates

Relative frequency	Date when Delta exceeded the relative frequency (95% CI)	Average relative reproduction number w.r.t. Alpha (95% CI)
0.05	2021-04-21 (2021-04-16, 2021-04-24)	1.049 (1.045, 1.053)
0.10	2021-04-26 (2021-04-23, 2021-04-29)	1.090 (1.084, 1.095)
0.15	2021-04-30 (2021-04-27, 2021-05-02)	1.141 (1.134, 1.148)
0.20	2021-05-03 (2021-04-30, 2021-05-05)	1.193 (1.185, 1.203)
0.25	2021-05-05 (2021-05-03, 2021-05-07)	1.235 (1.223, 1.246)
0.30	2021-05-07 (2021-05-05, 2021-05-09)	1.281 (1.267, 1.295)
0.35	2021-05-09 (2021-05-07, 2021-05-10)	1.332 (1.314, 1.349)
0.40	2021-05-10 (2021-05-08, 2021-05-12)	1.358 (1.339, 1.377)
0.45	2021-05-12 (2021-05-10, 2021-05-14)	1.412 (1.389, 1.434)
0.50	2021-05-14 (2021-05-12, 2021-05-15)	1.467 (1.440, 1.493)
0.55	2021-05-15 (2021-05-13, 2021-05-17)	1.493 (1.465, 1.521)
0.60	2021-05-17 (2021-05-15, 2021-05-19)	1.545 (1.513, 1.577)
0.65	2021-05-19 (2021-05-17, 2021-05-21)	1.594 (1.559, 1.628)
0.70	2021-05-20 (2021-05-18, 2021-05-23)	1.617 (1.580, 1.652)
0.75	2021-05-23 (2021-05-20, 2021-05-25)	1.677 (1.638, 1.716)
0.80	2021-05-25 (2021-05-23, 2021-05-28)	1.712 (1.671, 1.751)
0.85	2021-05-28 (2021-05-25, 2021-05-31)	1.754 (1.713, 1.794)
0.90	2021-06-01 (2021-05-29, 2021-06-05)	1.796 (1.754, 1.837)
0.95	2021-06-07 (2021-06-03, 2021-06-13)	1.835 (1.794, 1.875)

2.3.2. Relative reproduction number of Delta with respect to Alpha estimated from partial data

Table 2.3 shows the parameters of the model estimated using the partial data collected no later than each of dates in Table 2.2. The final estimate of k using observations of the entire period in the Alpha–Delta replacement was 1.88 (Table 2.1). The final estimate was within 95% CIs of estimations in seventeen out of nineteen estimations using the partial observations. Only the two early estimations, made at relative frequencies of 0.05 and 0.10, failed to cover the final estimate of k in their 95% CIs. All 95% CIs of k estimated at relative frequencies greater than or equal to 0.15 covered the final estimate of k. These results implied that it was possible to accurately estimate the relative reproduction number of Delta w.r.t. Alpha when relative frequencies of Delta became 0.15 or later. It took 38 days for Delta to reach a relative frequency of 0.95 (7 June 2021) from when it was 0.15 (30 April 2021) (Table 2.2). Therefore, it would be possible to estimate the relative reproduction number of Delta w.r.t. Alpha more than one month before its fixation.

Table 2.3 Parameters estimated using partial observations

Observed relative frequency	k (95% CI)	$q_Y(t_0)$ (95% CI)	M (95% CI)	Log Likelihood
0.05	2.15 (2.00, 2.45)	0.0002 (0.0001, 0.0003)	834.12 (346.81, 2000.00†)	–80.91
0.10	2.06 (1.92, 2.21)	0.0002 (0.0001, 0.0004)	581.40 (268.62, 1426.19)	–102.87
0.15	1.93 (1.83, 2.05)	0.0004 (0.0002, 0.0006)	399.87 (202.44, 832.82)	–123.23
0.20	1.97 (1.87, 2.08)	0.0003 (0.0002, 0.0005)	362.81 (188.91, 720.53)	–137.69
0.25	1.92 (1.83, 2.02)	0.0004 (0.0002, 0.0006)	307.07 (165.69, 575.00)	–148.88
0.30	1.91 (1.83, 2.00)	0.0004 (0.0003, 0.0006)	310.63 (170.09, 574.06)	–158.03
0.35	1.92 (1.85, 2.00)	0.0004 (0.0003, 0.0006)	328.39 (182.12, 607.05)	–166.02
0.40	1.93 (1.86, 2.00)	0.0004 (0.0003, 0.0006)	339.98 (189.18, 628.24)	–170.45
0.45	1.90 (1.84, 1.96)	0.0004 (0.0003, 0.0006)	315.00 (179.08, 565.97)	–180.84
0.50	1.86 (1.79, 1.92)	0.0005 (0.0004, 0.0008)	231.61 (136.25, 392.22)	–194.90
0.55	1.85 (1.79, 1.91)	0.0006 (0.0004, 0.0008)	234.52 (139.84, 401.81)	–198.95
0.60	1.86 (1.80, 1.91)	0.0005 (0.0004, 0.0007)	247.76 (147.88, 419.60)	–207.59
0.65	1.87 (1.81, 1.92)	0.0005 (0.0004, 0.0007)	248.77 (150.01, 415.23)	–217.85
0.70	1.86 (1.81, 1.91)	0.0005 (0.0004, 0.0007)	250.70 (152.17, 417.80)	–222.51
0.75	1.86 (1.82, 1.91)	0.0005 (0.0004, 0.0007)	271.02 (164.92, 448.18)	–235.19
0.80	1.87 (1.82, 1.91)	0.0005 (0.0004, 0.0007)	285.77 (174.67, 473.56)	–244.46
0.85	1.84 (1.81, 1.89)	0.0006 (0.0004, 0.0007)	250.12 (159.20, 426.80)	–261.43
0.90	1.86 (1.82, 1.90)	0.0005 (0.0004, 0.0007)	238.70 (154.56, 365.00)	–284.10
0.95	1.88 (1.85, 1.92)	0.0005 (0.0004, 0.0006)	209.79 (142.17, 313.90)	–321.85

†The upper bound of M in the maximum likelihood estimation was set to 2,000.

2.3.3. Prediction of relative variant frequency in future

Retrospective prediction tests were conducted on the future relative frequency of Delta and the average relative reproduction number of circulating viruses w.r.t. Alpha using model parameters in Table 2.3, which were estimated from partial observations. Figure 2.3 shows predicted trajectories of the Alpha–Delta replacement using partial observations up to different time points in Table 2.2. The maximum likelihood predictions made at relative frequencies of 0.05 and 0.10 overestimated the future relative frequencies of Delta (Figure 2.3(a) and 2.3(b)), while predictions made at relative frequencies greater than or equal to 0.15 fitted well with future observations (Figure 2.3(c–i)).

According to the final estimate using the entire observations, Delta exceeded relative frequencies of 0.50, 0.70, and 0.90 on 14 May, 20 May, and 1 June 2021, respectively (Table 2.2). The accuracy of predictions was evaluated by analyzing predictions targeted on these dates (Figure 2.4). When predictions were made before relative frequencies of Delta reached 0.15, the relative frequencies of Delta on the target dates were overestimated (Figure 2.4(a–c)) and the dates predicted to exceed target relative frequencies were earlier than the final estimates (Figure 2.4(d–f)). The reason for these was that these early predictions overestimated the relative reproduction numbers of Delta w.r.t. Alpha (Table 2.3). In contrast, predictions made when relative frequencies of Delta were greater than or equal to 0.15 were close to the final estimate of relative frequencies (Figure 2.4(a–c)) and dates (Figure 2.4(d–f)). When relative frequencies of Delta were greater than or equal to 0.15, the predicted relative frequencies targeted on 14 May, 20 May, and 1 June 2021 had median errors of 0.060 ($n = 7$), 0.023 ($n = 11$), and 0.004 ($n = 15$) with maximum absolute errors of 0.092 ($n = 7$), 0.060 ($n = 11$), and 0.034 ($n = 15$), respectively (Table 2.4). With the same setting, the predicted dates exceeding targeted relative frequencies of 0.50, 0.70, and 0.90 had median errors of 1 ($n = 7$), 1 ($n = 11$), and 1 ($n = 15$) days with maximum absolute errors of 2 ($n = 7$), 2 ($n = 11$), 3 ($n = 15$) days, respectively (Table 2.5).

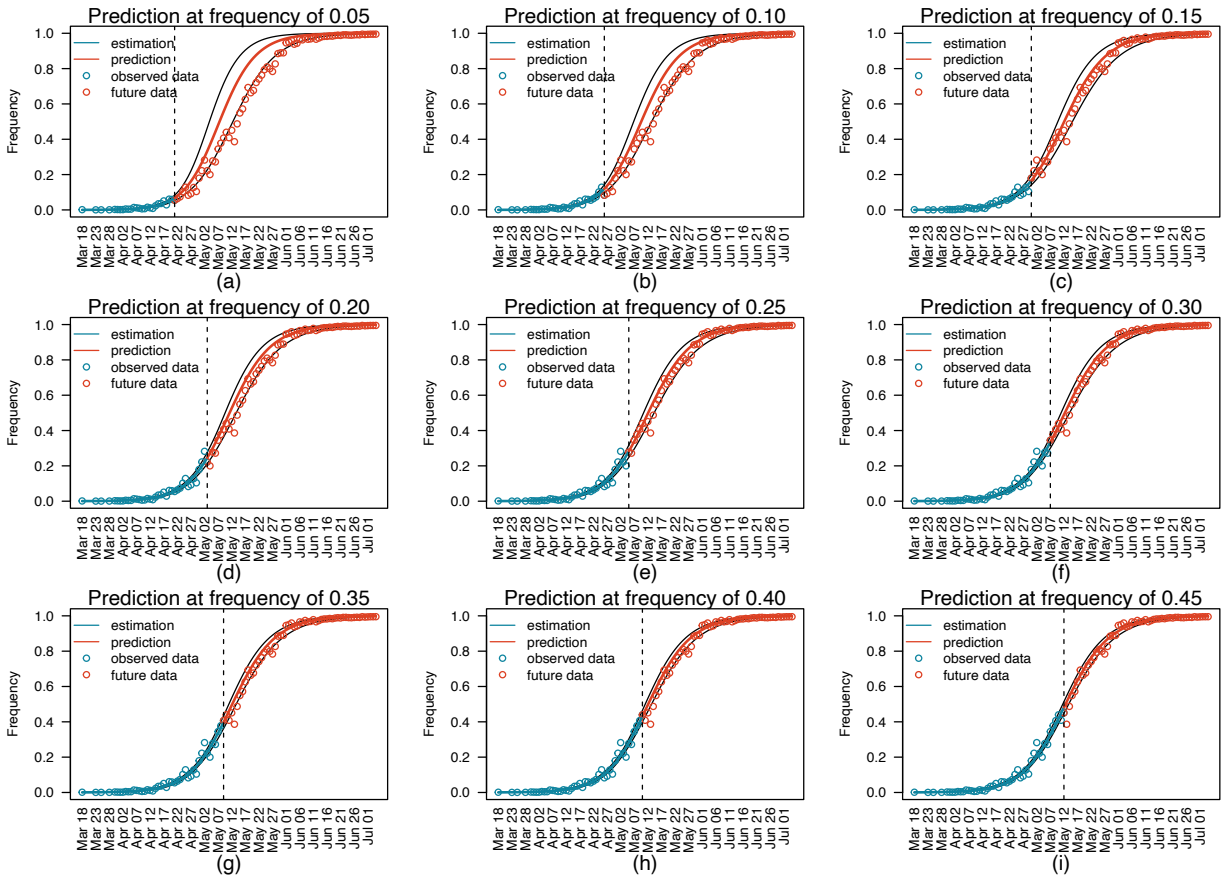

Figure 2.3 Prediction of future relative frequencies of the Delta variant using partial observations. Panels (a), (b), (c), (d), (e), (f), (g), (h), and (i) represent predictions estimated using observations until 21 April, 26 April, 30 April, 3 May, 5 May, 7 May, 9 May, 10 May, and 12 May 2021, re-spectively. Blue circles represent observed relative frequencies used for predictions. Red circles represent future observations that were not used for predictions. The vertical dashed line in each panel represents the date of the last observations used for prediction. The blue curve in each panel represents the maximum likelihood estimates of relative frequencies of Delta, and the red curve represents the relative frequencies predicted by the model using the estimated parameters. Black curves represent 95% confidence intervals of the relative frequencies of Delta.

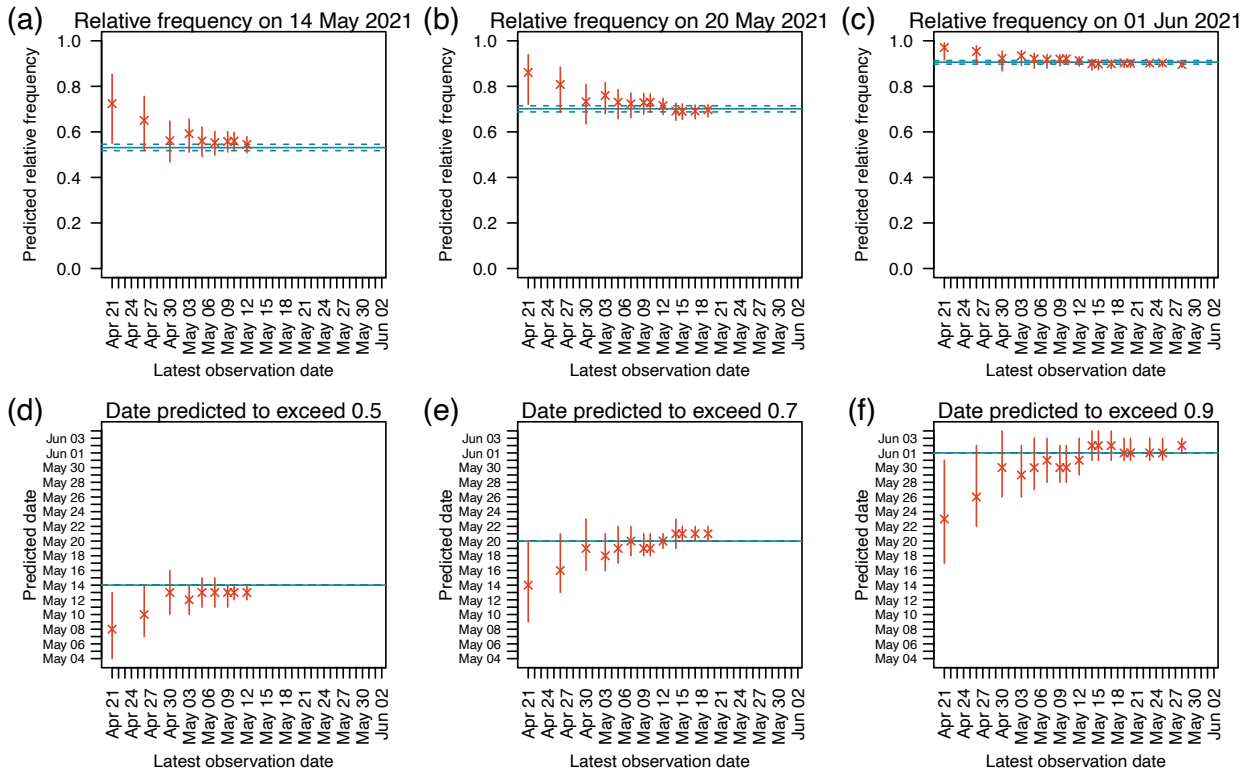

Figure 2.4 Predictions of relative frequencies of Delta on target dates and predictions of the dates when Delta would reach target relative frequencies. In each panel, x-axis represents dates until which observations were used in the prediction. Y-axes in panels (a), (b), and (c) represent the predicted relative frequencies on 14 May, 20 May, and 1 June, respectively. Y-axes in panels (d), (e), and (f) represent the predicted dates when Delta would reach relative frequencies of 0.50, 0.70, and 0.90, respectively. Cross marks represent predicted relative frequencies and dates with vertical bars showing their 95% confidence intervals. The blue horizontal solid lines represent the final estimates using the entire observations. The blue horizontal dashed lines represent 95% confidence intervals of the final estimates.

Table 2.4 Errors of predicted relative frequencies on target dates

Target date	Final estimate of relative frequency	Number of predictions†	Absolute errors in predicted relative frequency	
			Median	Maximum
14 May 2021	0.50	7	0.060	0.092
20 May 2021	0.70	11	0.023	0.060
1 June 2021	0.90	15	0.004	0.034

†The two earliest predictions, made when Delta was less than 0.15, were excluded.

Table 2.5 Errors of predicted dates exceeding target relative frequencies

Target relative frequency	Final estimate of date	Number of predictions†	Absolute errors of predicted dates	
			Median	Maximum
0.50	14 May 2021	7	1	2
0.70	20 May 2021	11	1	2
0.90	1 June 2021	15	1	3

†The two earliest predictions, made when Delta was less than 0.15, were excluded.

2.4. Discussion

The replacement from the Alpha variant to the Delta variant in England were analyzed using nucleotide sequences on the GISAID database collected from 18 March to 4 July 2021. The estimated relative reproduction number, k, of Delta w.r.t. Alpha was 1.88 (95% CI: 1.85–1.91) with a beta-binomial distribution parameter (M) of 288.54 (95% CI: 202.96–406.26) (Table 2.1). The relative reproduction number of Delta w.r.t. Alpha was accurately estimated from early observations once relative frequencies of Delta reached 0.15 (Table 2.3). Using these estimates of the relative reproduction number, the date when relative frequency of Delta would reach 0.90 was predicted with a maximum absolute prediction error of 3 days (Table 2.5).

Several studies have estimated the relative reproduction of Delta w.r.t. Alpha in different countries. Ito et al. estimated the relative reproduction number of Delta w.r.t. Alpha in Japan to be 1.35 (Ito et al., 2021a). Hansen estimated the relative reproduction number of Delta w.r.t. Alpha in Denmark to be 2.17 (Hansen, 2022). In this study, the relative reproduction number of Delta w.r.t. Alpha was estimated to be 1.88 (95% CI: 1.85–1.91) (Table 2.1). Figgins and Bedford found that the relative reproduction number of Delta and Alpha w.r.t. non-VOC variants in the United States were different depending on the states (Figgins & Bedford, 2021). The differences in relative reproduction numbers of Delta w.r.t. Alpha among countries or states may be attributed to the differences in the vaccine usage or the ethnicity of the target populations.

The model assumes that the sequences on GISAID database was sampled following a beta-binomial distribution. It is possible use the binomial distribution in the model instead of the beta-binomial distribution. The model using beta-binomial distribution resulted in lower Akaike information criterions (AIC) compared to the model using binomial distribution (Table 2.6). This means that the observed variance was larger than the variance of the binomial distribution. The additional variance to the binomial distribution may be attributed to the difference between relative variant frequencies among subpopulations, indicating that the target population was not well-mixed. For example, different regions may show different progresses in the variant replacement. The same may be true for different age groups.

The lockdown restrictions in the UK were relaxed from 17 May 2021. The relative reproduction number estimated using data up to 28 May 2021 was slightly lower than the final estimation using entire observations (Table 2.3). This is attributed to the decrease of relative frequency of Delta and the increase of the relative frequency of Alpha around 28 May 2021 (Figure 2.2). The underlying mechanism of the drop in the relative frequency of Delta after the relaxation of lockdown is unknown and needs to be further investigated.

Table 2.6 Parameters estimated from entire observation by the binomial distribution model and comparison of AIC values with that of the beta-binomial distribution model

Model	k (95% CI)	$q_Y(t_Y)$ (95% CI)	M (95% CI)	Log likelihood	AIC†
Beta-binomial distribution	1.88 (1.85, 1.91)	0.0005 (0.0004, 0.0006)	288.54 (202.96, 406.26)	–431.00	868.00
Binomial distribution	1.92 (1.91, 1.93)	0.0003 (0.0003, 0.0004)	–	–643.06	1292.12

†AIC refers to the Akaike information criterion of the model.

Prediction tests conducted in this study used the date of sample collection and did not consider delay in sequence submissions. Nucleotide sequences collected during the period from 18 March 2021, the introduction of Delta, to 13 May 2021, the date when Delta reached a relative frequency of 0.95, were submitted after 10.78 days on average with a standard deviation of 3.76 days. The 95% of sequences during this period were submitted to the GISAID database within 16 days after sample collection. This means real-time prediction may need additional 10–16 days to get prediction accuracy similar to results shown in this study.

The model assumes that there was no difference between the generation times of both variants with a mean value of 4.64 days (Nishiura et al., 2020). However, Hart et al. estimated the generation time of Delta (4.7 days) to be shorter than that of Alpha (5.5 days) (Hart et al., 2022). To allow differences between generation times of variants, it is necessary to extend the model to also estimate the relative generation times of the variant w.r.t. that of the baseline variant (Ito et al., 2022).

During the period of analysis, the percentage of people receiving two-dose vaccinations in England was increasing from 7% on 1 April to 39% on 1 June 2021. The model assumes that there was no difference between vaccination efficacy against Alpha and Delta. If there was difference between vaccine efficacies against Alpha and Delta, the relative reproduction number of Delta w.r.t Alpha would change as population vaccine coverage increases. Detection of the difference in vaccine efficacy among variants may be possible by analyzing temporal change of relative reproduction number using vaccination coverage data if there is a sufficient difference.

In this study, the relative reproduction number of Delta w.r.t. Alpha in England and its future trajectory of replacement could be predicted one month before it reached a relative frequency of 0.90. Public health policy makers have only one month to prepare control measures for the increase in viral transmissibility. This implies that a quick decision-making process is needed to take advantage of the prediction. This period can be extended if accurate predictions is available using data earlier than one month. Further research is needed to investigate how much additional information was required for obtaining accurate predictions using data earlier than one month.

Conclusion

This research developed two population genetics models of variant replacements of viruses. Chapter 1 developed models describing variant replacement under natural selection in an aging population. Chapter 2 developed a model describing variant replacement under natural selection among variants with different transmissibilities. In both chapters, parameters of the models were estimated using observed data of nucleotide sequences retrieved from a public database, and the predictability of variant replacements in future was evaluated using the developed models.

In Chapter 1, the fixation probability of variants in an aging population was modeled based on Kimura's formula of advantageous selection. Fixation and extinction of variants of human H1N1 influenza viruses circulating from 2009 to 2020 were used to estimate the parameters of the models. Performances in predicting amino acid substitutions on HA were evaluated using cross-validation tests. Results showed that the best model could predict amino acid substitutions with a sensitivity of 0.78, specificity of 0.86, and precision of 0.83 once the amino acid relative frequency exceeds 0.11. Further investigations of model parameters revealed that the fixation probability of variants having a new amino acid increased as its frequency increased. The fixation probability also increased when variants having the new amino acid on HA could infect patients of higher age groups more effectively than those having the old amino acid on the same position. Possible explanation of this phenomenon is attributed to the different amino acids at this position on HA of the variants that first infected patients of the different age groups. However, an important limitation of the described method is that the model can only predict the evolutionary outcome of variants having new amino acids but cannot predict the time when they will become fixed in viral population.

In Chapter 2, the temporal change in variant frequencies was modeled using the relative reproduction number, which is the ratio of the reproduction number of a variant to that of another variant. Variant frequencies during the Alpha-to-Delta replacement in England were used to estimate the relative reproduction number of Delta w.r.t. Alpha, giving a value of 1.88 (95% CI: 1.85–1.91). The relative reproduction number of Delta w.r.t. Alpha was estimated using partial observations of variant frequencies during early phases and was accurately estimated once frequencies of Delta had reached 0.15, which was one month before its fixation. Public health policymakers would have had one month to prepare control measures for the increase in viral transmissibility and a quick decision-making process is needed to take advantage of the prediction. This period can be extended if accurate predictions are available using data earlier than one month. Further research is needed to investigate how much additional information was required for obtaining accurate predictions using data earlier than one month.

The methodologies in both studies in Chapter 1 and Chapter 2 may be applicable to the prediction of variant replacement of viruses other than the ones described here. The model described in Chapter 1

may be applied for antigenic evolution of other viruses such as H3N2 influenza A virus, influenza B virus, and SARS-CoV-2. The model introduced in Chapter 2 may be applied not only to other variants of SARS-CoV-2 but also to other newly emerged viruses. Evaluation of predictions using proposed models with other viruses must be investigated in future. The generalization of proposed models using additional parameters is another direction of future research. It is worth noting that this dissertation focuses on how natural selection act on new variants of viruses that have already emerged. It does not cover predictions of future variants that have not yet emerged in the viral population.

Acknowledgements

First of all, I would like to express my deepest sincere gratitude to Professor Kimihito Ito, my supervisor, for his support throughout my PhD journey. He has always been understanding and patient with his guidance. He is most generous with his time. He has always been caring and thoughtful of my well-being, especially during the COVID-19 pandemic when work had to be done remotely in isolation. I cannot express how fortunate I am to have him as my supervisor.

I also thank Professor Ayato Takada, Professor Yasuhiko Suzuki, Associate Professor Ryosuke Omori, and Professor Norikazu Isoda for their helpful suggestions during advisory sessions throughout my study.

I would like to express my gratitude to the Ministry of Education, Culture, Sports, Science, and Technology, Japan (MEXT) and the World-leading Innovative and Smart Education (WISE) Program for providing financial support without which this research would not have been possible.

I wish to thank Specially Appointed Lecturer Michael James Henshaw who has always reminded me that science is fascinating and research is fun.

I would also like to acknowledge the people who had unknowingly supported my research through all means. These include, but are not limited to, the team who was running GISAID, and people who asked and answered questions on Stack Overflow.

Lastly, I would like to extend my thanks towards my friends who had always kept me entertained throughout the journey.

References

Agor, J. K., & Ozaltin, O. Y. (2018). Models for predicting the evolution of influenza to inform vaccine strain selection. *Hum Vaccin Immunother, 14*(3), 678-683. doi:10.1080/21645515.2017.1423152

Arevalo, P., McLean, H. Q., Belongia, E. A., & Cobey, S. (2020). Earliest infections predict the age distribution of seasonal influenza A cases. *Elife, 9*, e50060. doi:10.7554/eLife.50060

Barton, N. H. (2000). Genetic hitchhiking. *Philos Trans R Soc Lond B Biol Sci, 355*(1403), 1553-1562. doi:10.1098/rstb.2000.0716

Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., Daniels, R. S., Gunasekaran, C. P., Hurt, A. C., Kelso, A., Klimov, A., Lewis, N. S., Li, X., McCauley, J. W., Odagiri, T., Potdar, V., Rambaut, A., Shu, Y., Skepner, E., Smith, D. J., Suchard, M. A., Tashiro, M., Wang, D., Xu, X., Lemey, P., & Russell, C. A. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature, 523*(7559), 217-220. doi:10.1038/nature14460

Bélisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}$ d. *J Appl Probab, 29*(4), 885-895. doi:10.2307/3214721

Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J., & Fitch, W. M. (1999). Predicting the evolution of human influenza A. *Science, 286*(5446), 1921-1925. doi:10.1126/science.286.5446.1921

Bush, R. M., Fitch, W. M., Bender, C. A., & Cox, N. J. (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol, 16*(11), 1457-1465. doi:10.1093/oxfordjournals.molbev.a026057

Castro, L. A., Bedford, T., & Ancel Meyers, L. (2020). Early prediction of antigenic transitions for influenza A/H3N2. *PLoS Comput Biol, 16*(2), e1007683. doi:10.1371/journal.pcbi.1007683

Clopper, C. J., & Pearson, E. S. (1934). The of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26*(4), 404-413. doi:10.1093/biomet/26.4.404

Cohen, J. (1992). A power primer. *Psychol Bull, 112*(1), 155-159. doi:10.1037//0033-2909.112.1.155

Cohen, J. (2010). Swine flu pandemic. What's old is new: 1918 virus matches 2009 H1N1 strain. *Science, 327*(5973), 1563-1564. doi:10.1126/science.327.5973.1563

Cui, H., Wei, X., Huang, Y., Hu, B., Fang, Y., & Wang, J. (2014). Using multiple linear regression and physicochemical changes of amino acid mutations to predict antigenic variants of influenza A/H3N2 viruses. *Biomed Mater Eng, 24*(6), 3729-3735. doi:10.3233/BME-141201

Davenport, F. M., Hennessy, A. V., Stuart-Harris, C. H., & Francis, T., Jr. (1955). Epidemiology of influenza; comparative serological observations in England and the United States. *Lancet, 269*(6888), 469-474. doi:10.1016/s0140-6736(55)93328-6

Du, X., Dong, L., Lan, Y., Peng, Y., Wu, A., Zhang, Y., Huang, W., Wang, D., Wang, M., Guo, Y., Shu, Y., & Jiang, T. (2012). Mapping of H3N2 influenza antigenic evolution in China reveals a strategy for vaccine strain recommendation. *Nat Commun, 3*, 709. doi:10.1038/ncomms1710

Ferguson, N. M., Galvani, A. P., & Bush, R. M. (2003). Ecological and immunological determinants of influenza evolution. *Nature, 422*(6930), 428-433. doi:10.1038/nature01509

Figgins, M. D., & Bedford, T. (2021). SARS-CoV-2 variant dynamics across US states show consistent differences in effective reproduction numbers. *medRxiv*, 2021.2012.2009.21267544. doi:10.1101/2021.12.09.21267544

Fisher, R. A. (1930). The evolution of dominance in certain polymorphic species. *Am Nat, 64*(694), 385-406. doi:10.1086/280325

Fitch, W. M., Leiter, J. M., Li, X. Q., & Palese, P. (1991). Positive Darwinian evolution in human influenza A viruses. *Proc Natl Acad Sci U S A, 88*(10), 4270-4274. doi:10.1073/pnas.88.10.4270

Francis, T. (1960). On the doctrine of original antigenic sin. *Proc Am Philos Soc, 104*(6), 572-578.

Francis, T., Salk, J. E., & Quilligan, J. J. (1947). Experience with vaccination against influenza in the spring of 1947: A preliminary report. *Am J Public Health Nations Health, 37*(8), 1013-1016. doi:10.2105/ajph.37.8.1013

Fraser, C. (2007). Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS One, 2*(8), e758. doi:10.1371/journal.pone.0000758

Garten, R. J., Davis, C. T., Russell, C. A., Shu, B., Lindstrom, S., Balish, A., Sessions, W. M., Xu, X., Skepner, E., Deyde, V., Okomo-Adhiambo, M., Gubareva, L., Barnes, J., Smith, C. B., Emery, S. L., Hillman, M. J., Rivailler, P., Smagala, J., de Graaf, M., Burke, D. F., Fouchier, R. A., Pappas, C., Alpuche-Aranda, C. M., Lopez-Gatell, H., Olivera, H., Lopez, I., Myers, C. A., Faix, D., Blair, P. J., Yu, C., Keene, K. M., Dotson, P. D., Jr., Boxrud, D., Sambol, A. R., Abid, S. H., St George, K., Bannerman, T., Moore, A. L., Stringer, D. J., Blevins, P., Demmler-Harrison, G. J., Ginsberg, M., Kriner, P., Waterman, S., Smole, S., Guevara, H. F., Belongia, E. A., Clark, P. A., Beatrice, S. T., Donis, R., Katz, J., Finelli, L., Bridges, C. B., Shaw, M., Jernigan, D. B., Uyeki, T. M., Smith, D. J., Klimov, A. I., & Cox, N. J. (2009). Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science, 325*(5937), 197-201. doi:10.1126/science.1176225

Gavrilets, S., & Gibson, N. (2002). Fixation probabilities in a spatially heterogeneous environment. *Popul Ecol, 44*(2), 51-58. doi:10.1007/s101440200007

Gerrish, P. J., & Lenski, R. E. (1998). The fate of competing beneficial mutations in an asexual population. *Genetica, 102*(0), 127. doi:10.1023/A:1017067816551

Girard, M. P., Tam, J. S., Assossou, O. M., & Kieny, M. P. (2010). The 2009 A (H1N1) influenza virus pandemic: A review. *Vaccine, 28*(31), 4895-4902. doi:10.1016/j.vaccine.2010.05.031

Gostic, K. M., Ambrose, M., Worobey, M., & Lloyd-Smith, J. O. (2016). Potent protection against H5N1 and H7N9 influenza via childhood hemagglutinin imprinting. *Science, 354*(6313), 722-726. doi:10.1126/science.aag1322

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics, 34*(23), 4121-4123. doi:10.1093/bioinformatics/bty407

Hansen, P. R. (2022). Relative contagiousness of emerging virus variants: An analysis of the Alpha, Delta, and Omicron SARS-CoV-2 variants. *Econom J, 25*(3), 739-761. doi:10.1093/ectj/utac011

Hart, W. S., Miller, E., Andrews, N. J., Waight, P., Maini, P. K., Funk, S., & Thompson, R. N. (2022). Generation time of the alpha and delta SARS-CoV-2 variants: an epidemiological analysis. *Lancet Infect Dis, 22*(5), 603-610. doi:10.1016/S1473-3099(22)00001-9

Held, L., & Bové, D. S. (2020). *Likelihood and Bayesian Inference* (2 ed.). Heidelberg: Springer Berlin.

Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric Statistical Methods* (3 ed.). New Jersey: Wiley & Sons.

Huddleston, J., Barnes, J. R., Rowe, T., Xu, X., Kondor, R., Wentworth, D. E., Whittaker, L., Ermetal, B., Daniels, R. S., McCauley, J. W., Fujisaki, S., Nakamura, K., Kishida, N., Watanabe, S., Hasegawa, H., Barr, I., Subbarao, K., Barrat-Charlaix, P., Neher, R. A., & Bedford, T. (2020). Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. *Elife, 9*, e60067. doi:10.7554/eLife.60067

Igarashi, M., Ito, K., Yoshida, R., Tomabechi, D., Kida, H., & Takada, A. (2010). Predicting the antigenic structure of the pandemic (H1N1) 2009 influenza virus hemagglutinin. *PLoS One, 5*(1), e8553. doi:10.1371/journal.pone.0008553

Illingworth, C. J., & Mustonen, V. (2012). Components of selection in the evolution of the influenza virus: linkage effects beat inherent selection. *PLoS Pathog, 8*(12), e1003091. doi:10.1371/journal.ppat.1003091

Ito, K., Igarashi, M., Miyazaki, Y., Murakami, T., Iida, S., Kida, H., & Takada, A. (2011). Gnarled-trunk evolutionary model of influenza A virus hemagglutinin. *PLoS One, 6*(10), e25953. doi:10.1371/journal.pone.0025953

Ito, K., Piantham, C., & Nishiura, H. (2021a). Predicted dominance of variant Delta of SARS-CoV-2 before Tokyo Olympic Games, Japan, July 2021. *Euro Surveill, 26*(27), 2100570. doi:10.2807/1560-7917.ES.2021.26.27.2100570

Ito, K., Piantham, C., & Nishiura, H. (2021b). Relative instantaneous reproduction number of Omicron SARS-CoV-2 variant with respect to the Delta variant in Denmark. *J Med Virol, 94*(5), 2265-2268. doi:10.1002/jmv.27560

Ito, K., Piantham, C., & Nishiura, H. (2022). Estimating relative generation times and reproduction numbers of Omicron BA.1 and BA.2 with respect to Delta variant in Denmark. *Math Biosci Eng, 19*(9), 9005-9017. doi:10.3934/mbe.2022418

Itoh, Y., Shinya, K., Kiso, M., Watanabe, T., Sakoda, Y., Hatta, M., Muramoto, Y., Tamura, D., Sakai-Tagawa, Y., Noda, T., Sakabe, S., Imai, M., Hatta, Y., Watanabe, S., Li, C., Yamada, S., Fujii, K., Murakami, S., Imai, H., Kakugawa, S., Ito, M., Takano, R., Iwatsuki-Horimoto, K., Shimojima, M., Horimoto, T., Goto, H., Takahashi, K., Makino, A., Ishigaki, H., Nakayama, M., Okamatsu, M., Takahashi, K., Warshauer, D., Shult, P. A., Saito, R., Suzuki, H., Furuta, Y., Yamashita, M., Mitamura, K., Nakano, K., Nakamura, M., Brockman-Schneider, R., Mitamura, H., Yamazaki, M., Sugaya, N., Suresh, M., Ozawa, M., Neumann, G., Gern, J., Kida, H., Ogasawara, K., & Kawaoka, Y. (2009). In vitro and in vivo characterization of new swine-origin H1N1 influenza viruses. *Nature, 460*(7258), 1021-1025. doi:10.1038/nature08260

Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc Natl Acad Sci U S A, 41*(3), 144-150. doi:10.1073/pnas.41.3.144

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics, 47*, 713-719.

Klingen, T. R., Reimering, S., Guzman, C. A., & McHardy, A. C. (2018). In silico vaccine strain prediction for human influenza viruses. *Trends Microbiol, 26*(2), 119-131. doi:10.1016/j.tim.2017.09.001

Koel, B. F., Burke, D. F., Bestebroer, T. M., van der Vliet, S., Zondag, G. C., Vervaet, G., Skepner, E., Lewis, N. S., Spronken, M. I., Russell, C. A., Eropkin, M. Y., Hurt, A. C., Barr, I. G., de Jong, J. C., Rimmelzwaan, G. F., Osterhaus, A. D., Fouchier, R. A., & Smith, D. J. (2013). Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science, 342*(6161), 976-979. doi:10.1126/science.1244730

Kryazhimskiy, S., Dushoff, J., Bazykin, G. A., & Plotkin, J. B. (2011). Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet, 7*(2), e1001301. doi:10.1371/journal.pgen.1001301

Kucharski, A. J., & Gog, J. R. (2012). Age profile of immunity to influenza: effect of original antigenic sin. *Theor Popul Biol, 81*(2), 102-112. doi:10.1016/j.tpb.2011.12.006

Lambert, A. (2006). Probability of fixation under weak selection: a branching process unifying approach. *Theor Popul Biol, 69*(4), 419-441. doi:10.1016/j.tpb.2006.01.002

Lees, W. D., Moss, D. S., & Shepherd, A. J. (2010). A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2. *Bioinformatics, 26*(11), 1403-1408. doi:10.1093/bioinformatics/btq160

Lessler, J., Riley, S., Read, J. M., Wang, S., Zhu, H., Smith, G. J., Guan, Y., Jiang, C. Q., & Cummings, D. A. (2012). Evidence for antigenic seniority in influenza A (H3N2) antibody responses in southern China. *PLoS Pathog, 8*(7), e1002802. doi:10.1371/journal.ppat.1002802

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. *Lancet Infect Dis, 17*(3), 339-347. doi:10.1016/S1473-3099(16)30465-0

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 Surveill, 26*(1), 2002106. doi:10.2807/1560-7917.ES.2020.26.1.2002106

Łuksza, M., & Lässig, M. (2014). A predictive fitness model for influenza. *Nature, 507*(7490), 57-61. doi:10.1038/nature13087

Morris, D. H., Gostic, K. M., Pompei, S., Bedford, T., Luksza, M., Neher, R. A., Grenfell, B. T., Lassig, M., & McCauley, J. W. (2018). Predictive modeling of influenza shows the promise of applied evolutionary biology. *Trends Microbiol, 26*(2), 102-118. doi:10.1016/j.tim.2017.09.004

Nachbagauer, R., Choi, A., Hirsh, A., Margine, I., Iida, S., Barrera, A., Ferres, M., Albrecht, R. A., Garcia-Sastre, A., Bouvier, N. M., Ito, K., Medina, R. A., Palese, P., & Krammer, F. (2017).

Defining the antibody cross-reactome directed against the influenza virus surface glycoproteins. *Nat Immunol, 18*(4), 464-473. doi:10.1038/ni.3684
Neher, R. A., & Bedford, T. (2015). nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics, 31*(21), 3546-3548. doi:10.1093/bioinformatics/btv381
Neher, R. A., Russell, C. A., & Shraiman, B. I. (2014). Predicting evolution from the shape of genealogical trees. *Elife, 3*, e03568. doi:10.7554/eLife.03568
Neverov, A. D., Kryazhimskiy, S., Plotkin, J. B., & Bazykin, G. A. (2015). Coordinated Evolution of Influenza A Surface Proteins. *PLoS Genet, 11*(8), e1005404. doi:10.1371/journal.pgen.1005404
Nishiura, H., Ito, K., Anzai, A., Kobayashi, T., Piantham, C., & Rodriguez-Morales, A. J. (2021). Relative reproduction number of SARS-CoV-2 Omicron (B.1.1.529) compared with Delta variant in South Africa. *J Clin Med, 11*(1), 30. doi:10.3390/jcm11010030
Nishiura, H., Linton, N. M., & Akhmetzhanov, A. R. (2020). Serial interval of novel coronavirus (COVID-19) infections. *Int J Infect Dis, 93*, 284-286. doi:10.1016/j.ijid.2020.02.060
Ohta, T. (1992). The Nearly Neutral Theory of Molecular Evolution. *Annu Rev Ecol Syst, 23*(1), 263-286. doi:10.1146/annurev.es.23.110192.001403
Patwa, Z., & Wahl, L. M. (2008). The fixation probability of beneficial mutations. *J R Soc Interface, 5*(28), 1279-1289. doi:10.1098/rsif.2008.0248
Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Croydon: Oxford University Press.
Piantham, C., & Ito, K. (2021). Modeling the selective advantage of new amino acids on the hemagglutinin of H1N1 influenza viruses using their patient age distributions. *Virus Evol, 7*(1), veab049. doi:10.1093/ve/veab049
Rambaut, A., Holmes, E. C., O'Toole, A., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol, 5*(11), 1403-1407. doi:10.1038/s41564-020-0770-5
Ren, X., Li, Y., Liu, X., Shen, X., Gao, W., & Li, J. (2015). Computational Identification of Antigenicity-Associated Sites in the Hemagglutinin Protein of A/H1N1 Seasonal Influenza Virus. *PLoS One, 10*(5), e0126742. doi:10.1371/journal.pone.0126742
Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., de Jong, J. C., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A. D., Rimmelzwaan, G. F., Shaw, M. W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R. A., & Smith, D. J. (2008). The global circulation of seasonal influenza A (H3N2) viruses. *Science, 320*(5874), 340-346. doi:10.1126/science.1154137
Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. *Data Min Knowl Discov, 2*, 169-194. doi:10.1023/A:1009745219419
Shih, A. C., Hsiao, T. C., Ho, M. S., & Li, W. H. (2007). Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc Natl Acad Sci U S A, 104*(15), 6283-6288. doi:10.1073/pnas.0701396104
Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill, 22*(13), 30494. doi:10.2807/1560-7917.ES.2017.22.13.30494
Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D., & Fouchier, R. A. (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science, 305*(5682), 371-376. doi:10.1126/science.1097211
Steinbrück, L., Klingen, T. R., & McHardy, A. C. (2014). Computational prediction of vaccine strains for human influenza A (H3N2) viruses. *J Virol, 88*(20), 12123-12132. doi:10.1128/JVI.01861-14
Steinbrück, L., & McHardy, A. C. (2011). Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic Acids Res, 39*(1), e4. doi:10.1093/nar/gkq909
Strelkowa, N., & Lässig, M. (2012). Clonal interference in the evolution of influenza. *Genetics, 192*(2), 671-682. doi:10.1534/genetics.112.143396
Suzuki, Y. (2008). Positive selection operates continuously on hemagglutinin during evolution of H3N2 human influenza A virus. *Gene, 427*(1-2), 111-116. doi:10.1016/j.gene.2008.09.012

Suzuki, Y. (2013). Predictability of antigenic evolution for H3N2 human influenza A virus. *Genes Genet Syst, 88*(4), 225-232. doi:10.1266/ggs.88.225
Suzuki, Y. (2015). Selecting vaccine strains for H3N2 human influenza A virus. *Meta Gene, 4*, 64-72. doi:10.1016/j.mgene.2015.03.003
Sved, J. A., & Hill, W. G. (2018). One Hundred Years of Linkage Disequilibrium. *Genetics, 209*(3), 629-636. doi:10.1534/genetics.118.300642
Tao, K., Tzou, P. L., Nouhin, J., Gupta, R. K., de Oliveira, T., Kosakovsky Pond, S. L., Fera, D., & Shafer, R. W. (2021). The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat Rev Genet, 22*(12), 757-773. doi:10.1038/s41576-021-00408-x
Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., O'Toole, A., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C. V., Boyd, O., Loman, N. J., McCrone, J. T., Goncalves, S., Jorgensen, D., Myers, R., Hill, V., Jackson, D. K., Gaythorpe, K., Groves, N., Sillitoe, J., Kwiatkowski, D. P., consortium, C.-G. U., Flaxman, S., Ratmann, O., Bhatt, S., Hopkins, S., Gandy, A., Rambaut, A., & Ferguson, N. M. (2021). Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature, 593*(7858), 266-269. doi:10.1038/s41586-021-03470-x
Wilke, C. O. (2003). Probability of fixation of an advantageous mutant in a viral quasispecies. *Genetics, 163*(2), 467-474. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12618386
Wilson, I. A., & Cox, N. J. (1990). Structural basis of immune recognition of influenza virus hemagglutinin. *Annu Rev Immunol, 8*, 737-771. doi:10.1146/annurev.iy.08.040190.003513
World Health Organization. (2007). *A description of the process of seasonal and H5N1 influenza vaccine virus selection and development*. Geneva: World Health Organization.
World Health Organization. (2019). *Global influenza strategy 2019-2030*. Geneva: World Health Organization.
World Health Organization. (2020). Recommended composition of influenza virus vaccines for use in the 2020 - 2021 northern hemisphere influenza season. Retrieved from https://www.who.int/influenza/vaccines/virus/recommendations/2020-21_north/en/
World Health Organization. (2022). Tracking SARS-CoV-2 variants. Retrieved from https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/
Yokoyama, M., Fujisaki, S., Shirakura, M., Watanabe, S., Odagiri, T., Ito, K., & Sato, H. (2017). Molecular Dynamics Simulation of the Influenza A(H3N2) Hemagglutinin Trimer Reveals the Structural Basis for Adaptive Evolution of the Recent Epidemic Clade 3C.2a. *Front Microbiol, 8*, 584. doi:10.3389/fmicb.2017.00584