Title	A study on the population genetics of influenza A viruses
Author(s)	Kim, Kiyeon
Citation	北海道大学. 博士(獣医学) 甲第12395号
Issue Date	2016-09-26
DOI	10.14943/doctoral.k12395
Doc URL	http://hdl.handle.net/2115/67185
Type	theses (doctoral)
File Information	Kim_Kiyeon.pdf

A study on the population genetics of influenza A viruses

(A 型インフルエンザウイルスの集団遺伝学に関する研究)

Kiyeon KIM

CONTENTS

Chapter I

Host-specific and Segment-specific Evolutionary Dynamics of Avian and Human Influenza A Viruses: A Systematic Review

Chapter II

Estimating epidemiological parameters of pandemic influenza using approximate Bayesian computation and Tajima's D

Abbreviations

ABC	approximate Bayesian computation
BEAST	Bayesian evolutionary analysis by sampling trees
CI	credible interval
GARD	generic algorithm recombination detection
HA	hemagglutinin
HPD	highest posterior density
MCMC	Markov chain Monte Carlo
MC method	Monte Carlo method
MSE	mean square error
NA	neuraminidase
NCBI	National Center for Biotechnology Information
ODE	ordinary differential equation
R_0	basic reproduction number
R_E	effective reproduction number
SBP	single breakpoint recombination
SIR model	susceptible–infectious–removed model
TMRCA	time to the most recent common ancestor
WF-model	Wright-Fisher model
β	transmission rate
Υ	removal rate
μ	mutation rate

Preface

The influenza A virus is a zoonotic pathogen that infects a wide range of mammalian and avian species [1]. According to the antigenicity of hemagglutinin (HA) and neuraminidase (NA), influenza A viruses are divided into 18 HA subtypes and 11 NA subtypes [2]. The natural hosts of influenza A viruses are aquatic birds, such as ducks, geese, and gulls [3]. Sixteen HA subtypes and 9 NA subtypes of influenza A viruses are circulating among these aquatic bird species. So far, H1N1, H2N2, and H3N2 subtype viruses have caused pandemics in humans [4,5]. H5N1, H5N2, and H7N7 subtype viruses cause highly pathogenic avian influenza to chickens, and they have damaged poultry industry for long time [6,7]. Zoonotic transmissions of viruses from pigs and chickens to humans have been reported frequently [8–10].

All the influenza A viruses circulating in humans and poultry originated from their natural hosts. Kida *et al.* [11] showed ducks infected with influenza A viruses did not show clinical signs of diseases and they produced only low levels of serum antibodies. These results suggested that influenza A viruses have undergone neutral evolution in their natural host population.

Understanding the population genetics in the pathogen of infectious disease is important for controlling the outbreak. Genetic variation maintained in a microorganism population contains information about the evolutionary dynamics of the microorganism in the past.

Tajima's D is a statistic that can be used to test whether or not the population structures of target organisms follow the Wright-Fisher model (WF-model) [12–15]. The WF-model starts from three assumptions. First, the population of target organisms is selectively neutral that no mutation affects to fitness of viral population. Second, the population is constant in size and not subdivided. Using nucleotide sequence data from surveillance studies, Tajima's D can test whether or not these assumptions hold with the population. Tajima's D is often used to analyze genetic variation maintained in a population of organisms, including bacteria and viruses [16,17].

Estimating epidemiological parameters in early stage of epidemic/pandemic of infectious disease is important to establish control measure and intervention policy especially concerning vaccination strategy [18]. Basic reproduction number (R_0) is one of the most important parameter that defines an average number of successful transmission number per infectious person when the infectious was introduced into the totally susceptible population [18].

There are several sources of available information to estimate R_0 in real–time. Firstly, R_0 could be estimated using temporal changes of epidemiological incidence rate [19,20]. If the collected data could not represent the outbreak for several reasons—a rare disease, a biased sampling proportion, difficulty of sampling, or minor outbreak, then an accurate estimation would be difficult [21–23]. There was research to overcome such a shortage of data in minor outbreak, but there should be further studies to solve this problem [24]. Secondly, sequences information of pathogens could be used to estimate epidemiological parameters. Pybus *et al.* used coalescent theory to reconstruct genealogy and to estimate population changes by showing skyline–plot [25,26]. This method assumed that the population changed exponentially or logistically. Volz *et al.* estimated population size using mathematical model and viral sequence data collected from one time point [27]. This method combined coalesce model and compartment model, Susceptible–Infectious–

Removed (SIR) model, which is unsolvable ordinary differential equation (ODE). Stadler *et al*. developed add–on for BEAST2 program named BDSKY to estimate effective reproduction numbers using coalescent theory and Birth–Death model [28]. This method expanded the application of sequence information for estimating reproduction number.

This thesis consists of two chapters. The chapter I analyzed segment-specific and host-specific Tajima's D values using Influenza A virus sequences. The chapter II introduced new method to estimate epidemiological parameters using sequence data of pandemic influenza (2009) with Tajima's D.

Chapter I

Host-specific and Segment-specific Evolutionary Dynamics of Avian and Human Influenza A Viruses: A Systematic Review

Introduction

In this chapter, I analyzed host-specific and segment-specific Tajima's D trends of influenza A viruses. To minimize bias from viral population subdivision, I conducted a systematic review of surveillance studies on influenza A viruses of wild mallards, chickens, and humans using nucleotide sequences registered in the database of National Center for Biotechnology Information (NCBI). To my knowledge, this is the first comprehensive Tajima's D study that uses datasets obtained by stratifying NCBI database sequences according to their isolation hosts, sampling sites, and sampling year. To clarify theoretical detectability of influenza outbreaks by Tajima's D, I also conducted computer simulations of viral evolution with changing viral demography and confirmed a clear relationship between Tajima's D and the viral population changes.

Materials and methods

Tajima's D

Tajima's D [13] is the normalized difference between two statistics, Watterson's estimator and Tajima's estimator. Watterson's estimator θ_w, that is, the expected number of segregating sites between n sequences, is given by

$$\theta_W = \frac{S_n}{\sum_{k=2}^{n} \frac{1}{(k-1)}}. \tag{1}$$

The numerator of equation (1), S_n is the observed number of segregating sites, and the denominator of equation (1) is the expected total length of genealogy of n samples divided by 2 times total population N. Tajima's estimator θ_T, which is the average number of nucleotide differences, is given by

$$\theta_T = \frac{2}{n(n-1)} \sum_{i<j} \pi_{ij} . \tag{2}$$

Here π_{ij} denotes the pairwise difference between the i^{th} sequence and the j^{th} sequence in the samples, and $n(n–1)/2$ is the total number of pairs in the samples.

Tajima's D is derived by subtracting Watterson's estimator from Tajima's estimator and by normalizing its numerator as follows;

$$D = \frac{\theta_T - \theta_W}{Std(\theta_T - \theta_W)}. \tag{3}$$

From equation (3), the sample size for Tajima's D have to be larger than three because the denominator of Tajima's D becomes zero. Fig. 1 showed simple example of calculation of Tajima's D.

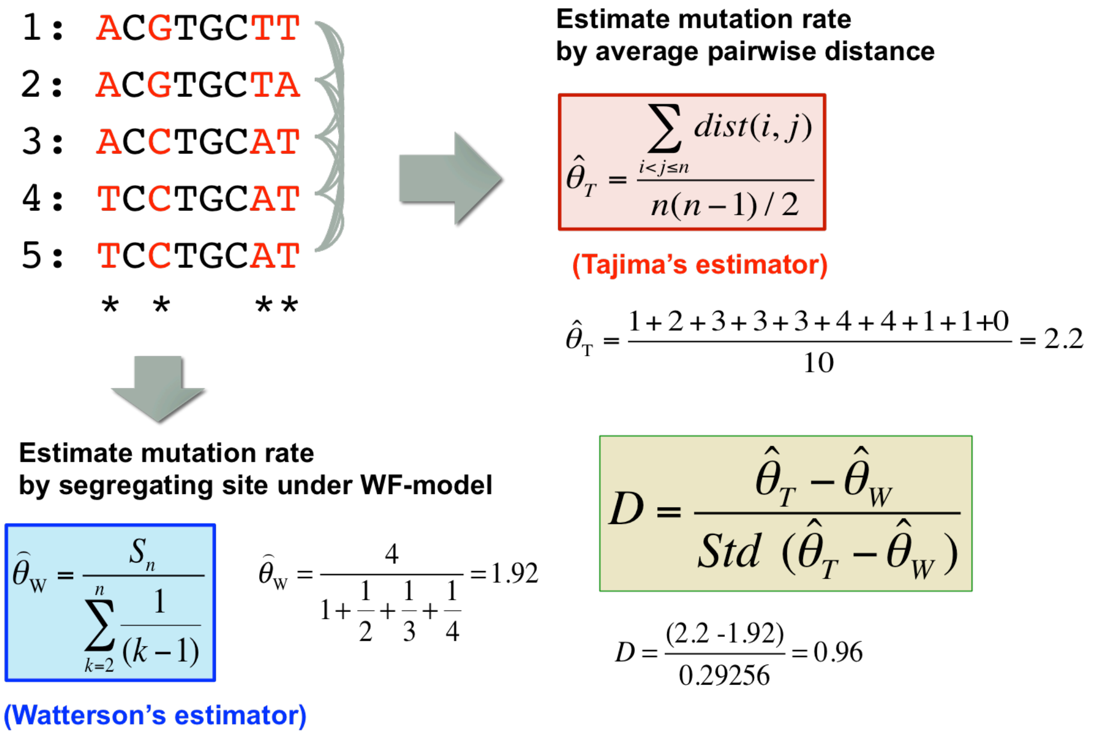

1: ACGTGCTT
2: ACGTGCTA
3: ACCTGCAT
4: TCCTGCAT
5: TCCTGCAT
* * **
Estimate mutation rate by average pairwise distance
$\hat{\theta}_T = \frac{\sum_{i<j\le n} dist(i,j)}{n(n-1)/2}$
(Tajima's estimator)
$\hat{\theta}_T = \frac{1+2+3+3+3+4+4+1+1+0}{10} = 2.2$
Estimate mutation rate by segregating site under WF-model
$\hat{\theta}_W = \frac{S_n}{\sum_{k=2}^{n} \frac{1}{(k-1)}}$
(Watterson's estimator)
$\hat{\theta}_W = \frac{4}{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} = 1.92$
$D = \frac{\hat{\theta}_T - \hat{\theta}_W}{Std\ (\hat{\theta}_T - \hat{\theta}_W)}$
$D = \frac{(2.2 - 1.92)}{0.29256} = 0.96$

Figure 1. Simple example of calculation of Tajima's estimator, Watterson's estimator, and Tajima's D

Given the five arbitrarily sequences, Tajima's estimator was 2.2, Watterson's estimator was 1.92, and Tajima's D was 0.96. The asterisk marks represent polymorphic site.

Systematic review

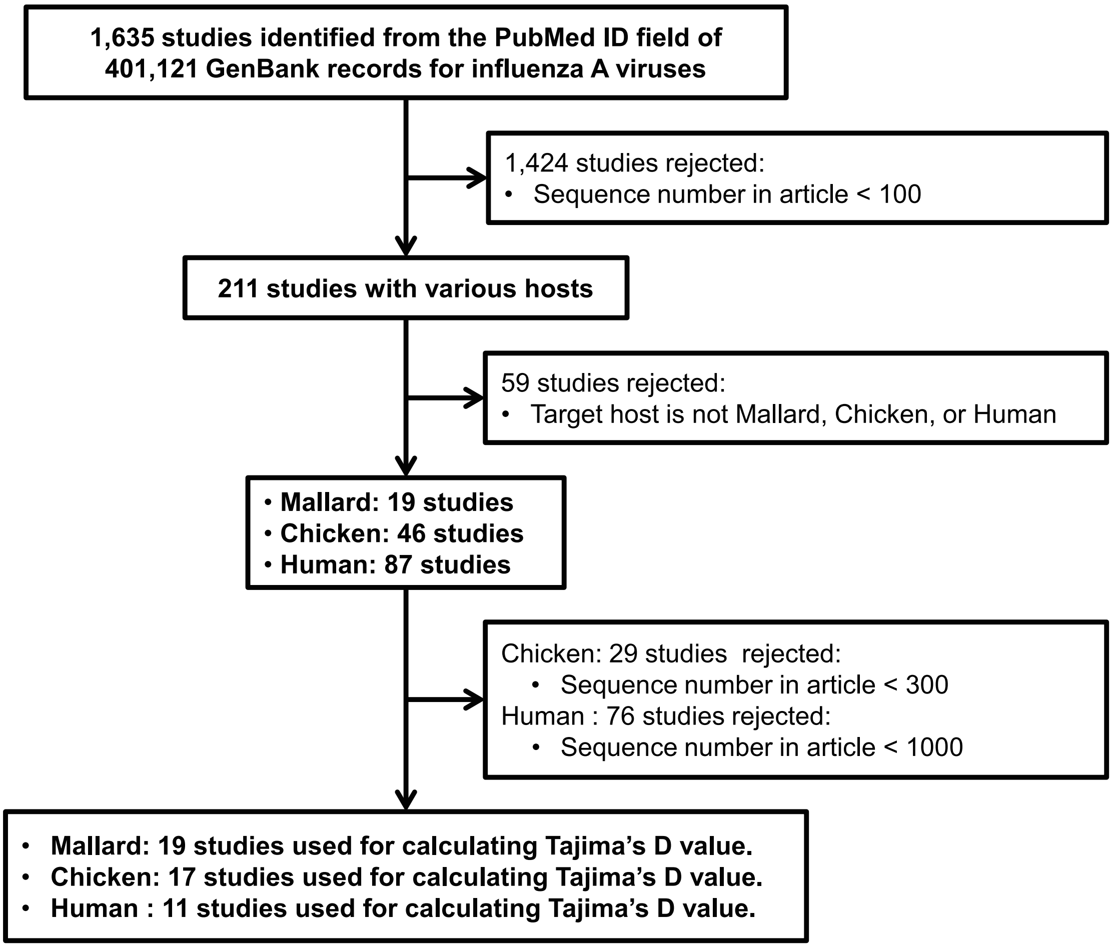
I downloaded all the database records of influenza A viruses from the GenBank on November 24th 2015 by using the Taxonomy ID of influenza viruses as a search condition, i.e. “txid=11320”. From the retrieved GenBank records, PubMed IDs were collected. Based on the PubMed ID, articles accompanied with more than 100, 300, and 1,000 GenBank sequence records respectively for mallard, chicken, and human viruses were collected. Influenza virus surveillance studies with wild mallards are conducted at a smaller scale than those for chickens and humans. To collect similar numbers of studies, I used these different thresholds on the minim sequence numbers for mallard, chicken, and human. To avoid bias from population subdivision [12,15], the abstract of articles were reviewed, and nucleotide sequences from surveillance studies conducted at a single sampling site from single host species were collected. Fig. 2 shows the selection process of the systematic review of surveillance studies.

Alignment of sequences and calculation of Tajima’s D

For each surveillance study selected by above criteria, nucleotide sequences of each gene segment were aligned using MAFFT, a multiple sequence alignment program (version 7) [29]. Sequences with a length less than 90% of complete gene were removed from the alignment. These aligned sequences were stratified according to their sampling years. Since Tajima’s D requires at least four sequences for its calculation, the datasets having less than four sequences were removed. For each dataset containing nucleotide sequences of the same gene segment of influenza A viruses isolated from the same sampling site in a single year, Tajima’s D was computed by a custom program implemented with Python3 (v.3.3.3).

Outbreak simulation

Viral sequence evolutions in a rapidly expanding population were simulated using Python3. I set the length of nucleotide sequences to 500 and mutation rate in the simulated evolution to 10^{-6} per base per generation. For each generation, viruses are randomly selected from the previous generation with replacement, and their nucleotide sequences were copied to the offspring in the current generation with mutations. I used equal mutation rates for all nucleotide bases (JC69 model) [30]. The simulation was started with 1,000 viruses with identical nucleotide sequences. During the first 5,000 generations, the population size was fixed to 1,000. In each generation from the 5,000th to 5,005th, the population size was doubled. From the 5,005th generation, the population size was fixed to 32,000 to the end of the simulation. For every 400 generations, 50 viruses were randomly sampled and Tajima's D was calculated from their nucleotide sequences. Totally, 100 simulations were conducted with the same setting, and averages of Tajima's D values were calculated.


1,635 studies identified from the PubMed ID field of 401,121 GenBank records for influenza A viruses

1,424 studies rejected:
- Sequence number in article < 100

211 studies with various hosts

59 studies rejected:
- Target host is not Mallard, Chicken, or Human

- Mallard: 19 studies
- Chicken: 46 studies
- Human: 87 studies

Chicken: 29 studies rejected:
- Sequence number in article < 300

Human : 76 studies rejected:
- Sequence number in article < 1000

- Mallard: 19 studies used for calculating Tajima's D value.
- Chicken: 17 studies used for calculating Tajima's D value.
- Human : 11 studies used for calculating Tajima's D value.

Figure 2. The selection process of systematic review of surveillance studies

Results

Data retrieval and sequence alignment

Using 401,212 GenBank records retrieved from the NCBI database, I identified 1,635 articles published with nucleotide sequences of the influenza A viruses. Among them 19, 17, and 11 articles satisfied the criteria for mallard, chicken and human, respectively (Fig. 2). A total of 42,664 nucleotide sequences accompanied with these 47 surveillance articles were used calculating Tajima's D. Table 1 shows the numbers of datasets for each segment and each host after removing dataset having less than four sequences in the alignment. The accession numbers and their nucleotide sequences used in this study can be found in the supplementary information.

Tajima's D in natural host species

Wild mallard

The mean of Tajima's D values of PB2, PB1, PA, NP, and M gene segments were 0.061, 0.028, 0.115, 0.077, and 0.048, respectively (Table 2). Medians of Tajima's D for the internal gene segments (PB2, PB1, PA, NP, and M) across datasets were close to zero, and the differences from zero were not significant (p>0.05, 1-sample Wilcoxon signed rank test) (Fig. 3(a)). The mean Tajima's D of the surface protein genes (HA and NA) and non–structural gene segment (NS) was 1.524, 1.769 and 0.657, respectively (Table 2). Medians of Tajima's D of these gene segments across datasets were significantly positive (p<0.05, 1-sample Wilcoxon signed rank test) (Fig. 3(a)).

Table 1. The number of datasets of nucleotide sequences

Host		PB2	PB1	PA	HA	NP	NA	M	NS
Mallard									
	Number of dataset	24	25	23	24	30	22	29	30
	Total number of sequences	237	240	228	244	367	206	388	315
Chicken									
	Number of dataset	14	19	19	18	18	13	31	23
	Total number of sequences	342	468	414	271	429	143	634	497
Human									
	Number of dataset	24	30	28	16	30	24	33	33
	Total number of sequences	1019	1110	1053	861	1169	880	1191	1191

Table 2. The mean and standard deviation of Tajima's D

Host		PB2	PB1	PA	HA	NP	NA	M	NS
Mallard									
	Mean	0.061	0.028	0.115	1.524	0.077	1.769	0.048	0.657
	SD	0.630	0.516	0.755	1.065	0.800	0.805	0.672	1.534
Chicken									
	Mean	–	–	–	–	–	–	–	–
		0.550	0.447	0.678	0.872	0.816	0.766	0.473	0.431
	SD	1.325	1.507	1.378	0.903	1.089	1.118	1.100	1.279
Human									
	Mean	–	–	–	–	–	–	–	–
		1.109	1.013	1.056	1.417	1.008	1.201	0.941	1.200
	SD	0.978	1.003	0.940	0.970	0.956	0.730	0.898	0.999

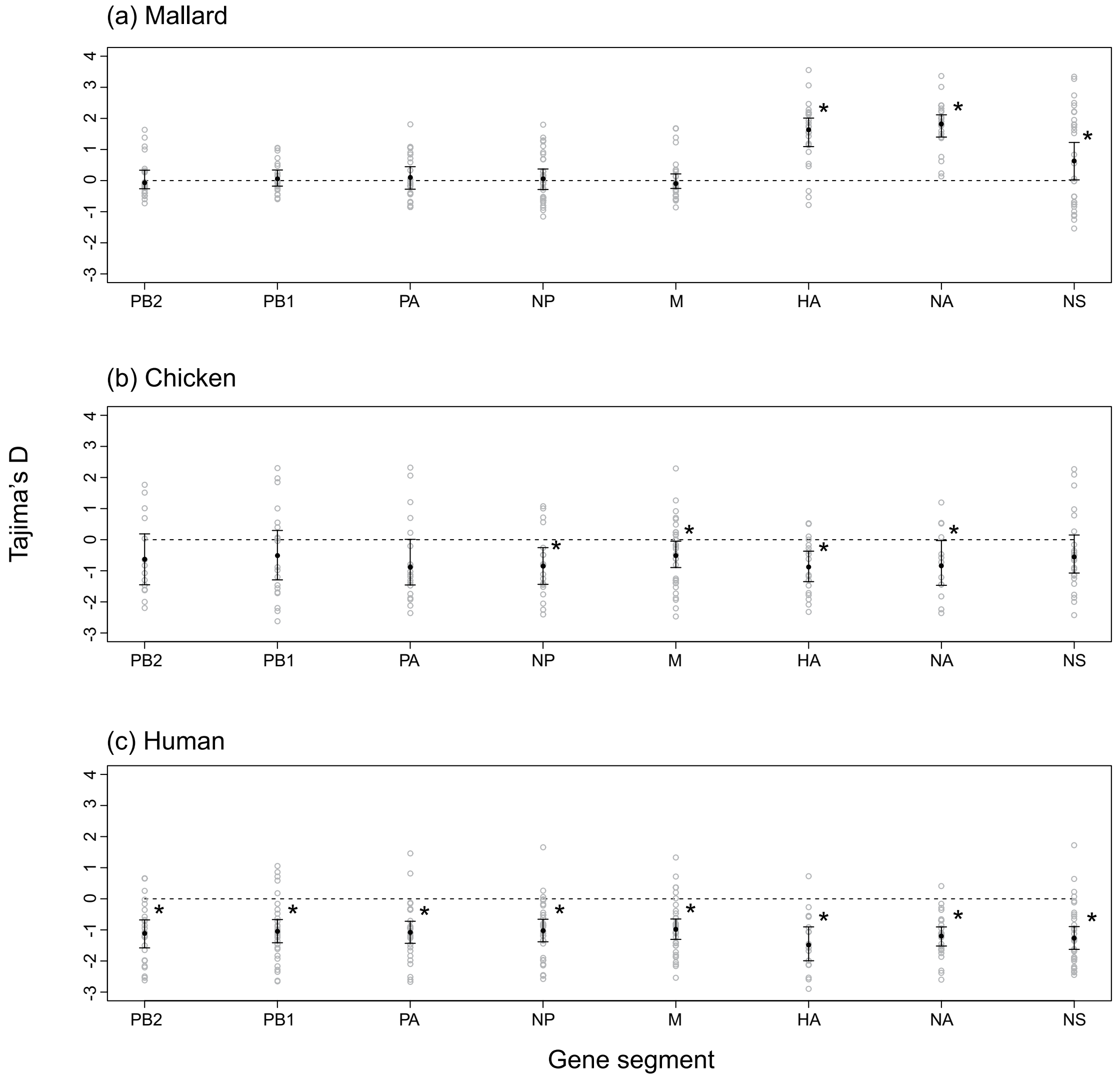


Figure 3. Tajima’s D values for gene segments sampled from the mallards, chickens and humans.

(a) shows Tajima's D values for the viruses isolated from wild mallards, (b) shows those from domestic chickens, and (c) shows those from humans. Black circles and error bars represent estimated medians and 95% confidence intervals for the median of Tajima's D across datasets using 1-sample Wilcoxon signed rank test. Gray circles represent Tajima's D values of each dataset. Asterisk denotes the significantly positive or negative Tajima's D based on the result of 1-sample Wilcoxon signed rank test.

Tajima’s D in non-natural host species

Chicken

Influenza A viruses that were circulating in chickens had an overall mean Tajima D of −0.629. The mean values of Tajima’s D in PB2, PB1, PA, HA, NP, NA, M, and NS gene segments were −0.550, −0.447, −0.678, −0.872, −0.816, −0.766, −0.473, and −0.431, respectively (Table 2). Medians of Tajima’s D values for HA, NP, NA, and MP gene segments across datasets were significantly negative (*p*<0.05, 1-sample Wilcoxon signed rank test) (Fig. 3(b)). There were significant differences in Tajima’s D between the wild mallard and chicken except NS gene segment (*p*<0.05; Two-sample Kolmogorov-Smirnov test).

Human

Influenza A viruses circulating in humans had a mean Tajima’s D of −1.118. The mean values of Tajima’s D in PB2, PB1, PA, HA, NP, NA, M, and NS gene segments were −1.109, −1.013, −1.056, −1.417, −1.008, −1.201, −0.941, and −1.200, respectively (Table 2). Medians of Tajima’s D for all gene segments were significantly negative (*p*<0.05, 1-sample Wilcoxon signed rank test) (Fig. 3(c)), and there were significant differences in Tajima’s D between the wild mallards and humans for all gene segments (*p*<0.05; Two-sample Kolmogorov-Smirnov test).

Outbreak simulation

At the first duration when the viral population size was constant over viral generations, the mean of Tajima’s D of 100 simulations was around zero and was within the range of 95% confidence interval for D=0 (the error distribution was assumed to be beta distribution [13], which agreed with the theory of Tajima's D. After a sudden increase of the viral population, the mean Tajima's D value decreased to –2.052, which is significantly negative (p<0.05; beta distribution) (Fig. 4). Consequently the mean Tajima's D value increased gradually and returned within the range of 95% confidence interval for D=0 (Fig. 4(b)).

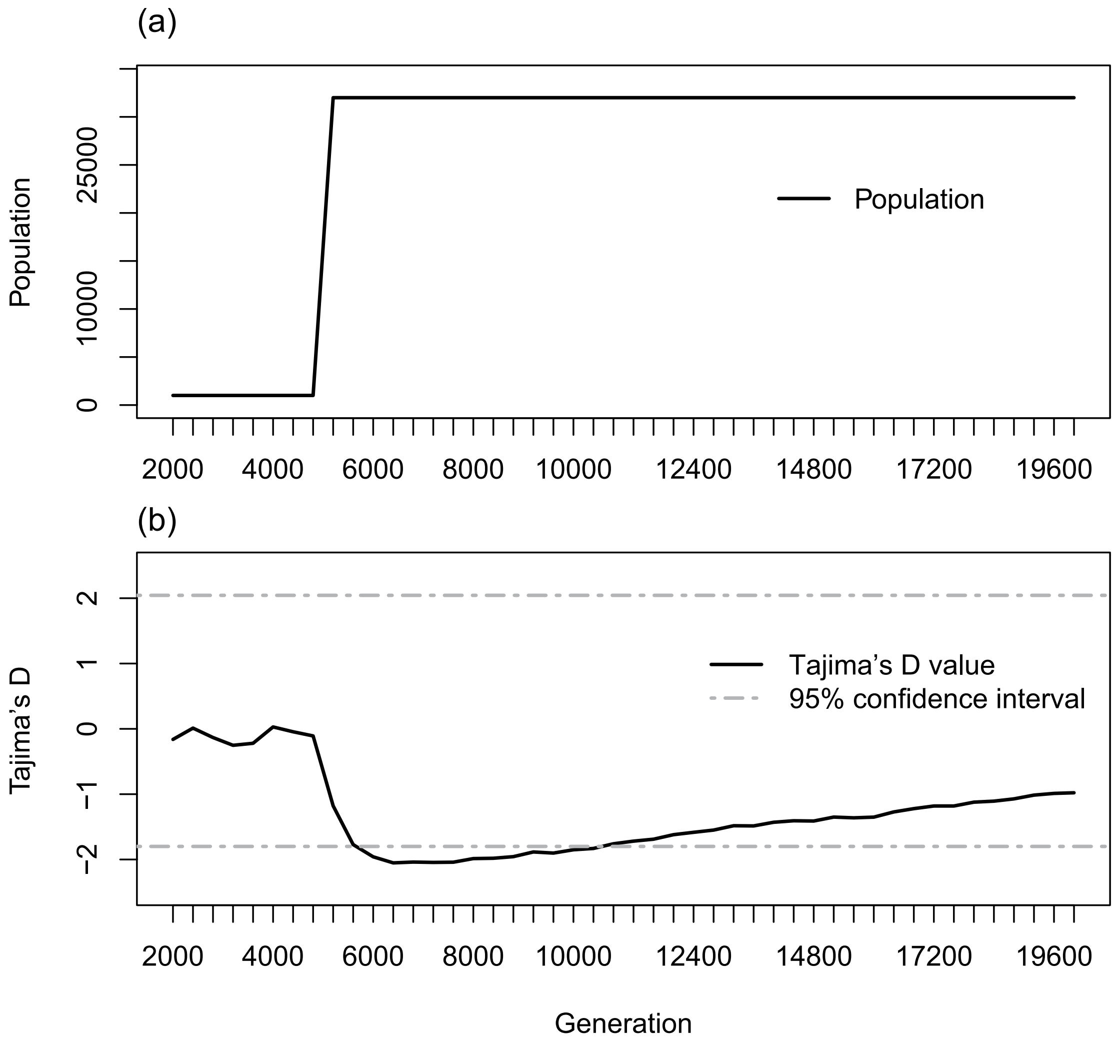


Figure 4. The change of Tajima's D with a sudden increase of population

(a) shows the setting of time evolution of viral population size and (b) shows the result of time series change of mean Tajima's D. Gray dot line represents 95% confidence interval of Tajima's d value for D=0.

Discussion

In this chapter, I analyzed host-specific and segment-specific Tajima’s D trends of influenza A viruses through a systematic review of viral sequences registered in the NCBI GenBank. To avoid bias from viral population subdivision, viral sequences were stratified according to their sampling locations and sampling years. Tajima’s D values for internal gene segments of influenza A viruses circulating in wild mallards were close to zero. On the other hand, interestingly, Tajima’s D for external gene segments of influenza A viruses circulating wild mallards showed positive. Tajima’s D values for both internal and external gene segments in non-natural hosts—chicken and human—were negative.

The trends of Tajima’s D are different between internal and external gene segments of influenza A viruses circulating in wild mallards. Wild mallard are considered as the natural host of influenza A viruses. Tajima’s D of influenza viruses in mallards is expected to be close to zero due to the low pathogenicity, or slightly negative due to the selective sweep by low immune response. However, Tajima’s D values for external genes showed positive value, suggesting balancing selection or population subdivision. Since all gene segments should show positive Tajima’s D if viral population were subdivided, balancing selection on external gene segments were more likely to be the cause of positive Tajima’s D values.

To analyze the selection on the external genes of influenza A viruses circulating in wild mallards, I compared Tajima’s D of the data containing only one subtype with those containing multiple subtypes using dataset from Bahl *et al.* [31]. Tajima’s D values of sequences containing two subtypes were positive: the values were 1.159 in 2006 and 1.032 in 2007, suggesting balancing selection. On the other hand, the Tajima’s D for sequences stratified by subtypes were not positive: −0.721

(−1.420) for the H3 HA in wild mallard in 2006 (2007), −1.222 (−0.535) for H4 HA in 2006 (2007), respectively, suggesting neutral or weak purifying selection (Table 3). A similar pattern was observed for NA (Table 4). These results suggested that selection within a subtype was neutral or weak purifying selection as observed in other non-natural hosts, on the other hand, selection across subtypes is balancing selection.

The diversity of influenza A viruses circulating wild mallard is much higher than other hosts. This high diversity is not able to be explained by relatively low pathogenicity or low immune response of wild mallard, which is one of the main reasons why wild mallard is considered to be the natural host of influenza A viruses. These factors can explain neutral selection on the viruses, but they cannot explain balancing selection.

Several studies have analyzed the evolutionary dynamics of avian influenza viruses using their nucleotide sequences. Time to the most recent common ancestor (TMRCA) of HA, NA and NS were much older than that of internal gene segments [32], and the result is consistent with my results. The phylogenetic analyses of HA and NA suggested high inter-subtype diversity and low intra-subtype diversity, which were not seen in internal gene segments [33]. The distinct divergence between two alleles of NS suggested balancing selection on NS [33], and this was consistent with this results. The dN/dS ratio—the ratio of the number of non-synonymous substitutions per site to the number of synonymous substitution per site—had suggested purifying selection on internal gene segments [31], while Tajima's D in this study supported neutral selection. This discrepancy between results from dN/dS and Tajima's D remains as an open question, and one possible explanation for this is that the discrepancy would be attributed to difference between the selection at the lineage level and the selection at the population level.

The negative Tajima's D values observed in the human and chicken viruses rejected the WF-model for these viral populations. These negative Tajima's D values should be attributed to the population increase due to recent outbreaks, purifying selection due to viral adaptation to new hosts, or combined effects of population change and selection. However, Tajima's D itself cannot be used to examine which of these factors are causes of negative Tajima's D values. This problem highlighted a need for the development of a new methodology that can be used to separate the composite signal into components of population change and selection.

Tajima proposed the use of beta-distribution to reject WF-model and to calculate the 95% confidence interval of Tajima's D under the WF-model [13]. Computer simulations showed that Tajima's D values fell outside 95% confidence interval right after the sudden increase of viral population. Although Simonsen *et al.* [34] showed that criteria using beta-distribution was too conservative to reject WF-model when neutrality assumption does not hold, computer simulations suggested that beta-distribution could be used to reject WF-model when population size is rapidly growing. When I have multiple samples independently collected form the population, an alternative approach to reject WF-model is to use 1-sample Wilcoxon signed rank test, as shown in the previous section.

It would be of particular interest to find connection between the Tajima's D of an infectious agent and the effective reproduction number of infectious disease caused by the agent. The effective reproduction number measures the continuance of an outbreak and the expected number of secondary infections. Recent studies have utilized coalescent theory to estimate the time evolution of population size of the ancestors of sampled sequences. By assuming constant-sized population between two coalescence events, Pybus *et al*. developed a method to estimate the time evolution of population size from their nucleotide sequences [25]. Mathematical models on population dynamics of infectious diseases have been also proposed to characterize infectious disease outbreaks from nucleotide sequences of infectious agents [27,28].

Table 3. Subtype specific Tajima’s D of HA in mallard.

Subtype		Year	
		2006	2007
H3			
	Sample size	11	17
	Tajima’s D	–0.721	–1.42
H4			
	Sample size	6	8
	Tajima’s D	–1.222	–0.535
H3, H4 and others (mixed)			
	Sample size	20	28
	Tajima’s D	1.519	1.032

Table 4. Subtype specific Tajima’s D of NA in mallard.

Subtype		Year	
		2006	2007
N6			
	Sample size	7	9
	Tajima’s D	–0.442	–1.315
N8			
	Sample size	11	14
	Tajima’s D	0.498	–0.011
N6, N8 and others (mixed)			
	Sample size	21	26
	Tajima’s D	2.125	2.052

Summary

Understanding the evolutionary dynamics of influenza viruses is essential to control both avian and human influenza. Here, we analyze host-specific and segment-specific Tajima's D trends of influenza A virus through a systematic review using viral sequences registered in the National Center for Biotechnology Information. To avoid bias from viral population subdivision, viral sequences were stratified according to their sampling locations and sampling years. As a result, we obtained a total of 580 datasets each of which consists of nucleotide sequences of influenza A viruses isolated from a single population of hosts at a single sampling site within a single year. By analyzing nucleotide sequences in the datasets, we found that Tajima's D values of viral sequences were different depending on hosts and gene segments. Tajima's D values of viruses isolated from chicken and human samples showed negative, suggesting purifying selection or a rapid population growth of the viruses. The negative Tajima's D values in rapidly growing viral population were also observed in computer simulations. Tajima's D values of PB2, PB1, PA, NP, and M genes of the viruses circulating in wild mallards were close to zero, suggesting that these genes have undergone neutral selection in constant-sized population. On the other hand, Tajima's D values of HA and NA genes of these viruses were positive, indicating HA and NA have undergone balancing selection in wild mallards. Taken together, these results indicated the existence of unknown factors that maintain viral subtypes in wild mallards.

Chapter II

Estimating epidemiological parameters of pandemic influenza using approximate Bayesian computation and Tajima's D

Introduction

In this chapter, I showed a new method to estimate epidemiological parameters using sequence data of infectious diseases. A Bayesian approach —approximate Bayesian computation (ABC) was used to estimate parameters. Based on Tajima's D of observed sequence data, I used sequence data randomly collected from simulation of individual based compartment model with mutations. By comparing Tajima's D values from observed sequences and those from simulated sequences, posterior distributions of epidemiological parameters were estimated [35]. Here, I applied this method to the viral sequence data sampled from pandemic influenza (2009) in the Buenos Aires, Argentina [36] and estimated epidemiological parameters of pandemic influenza (2009) and compared with precedent researches.

Materials and methods

Data selection

After downloading the influenza A virus database from GenBank on November 24th 2015, sequence data was subdivided to each assigned articles. To choose appropriate articles, of which sequence data was enough to get time–series Tajima's D values during the outbreak, I filtered articles, which had less than 200 sequences in one gene segment.

Alignment of sequences and Tajima's D values

The MAFFT, a multiple sequence alignment program (version 7) was used to align the sequences [29]. After alignment, sequences in each 3 consecutive days were stratified to make a new time unit. Tajima's D value was calculated using all sequence information in each time unit by a custom program implemented with Python3 (v.3.3.3).

SIR model simulation with evolving viral sequences

Viral sequence evolutions in the infectious population, which I assumed that they followed the SIR model, were simulated using Python3. Initially, total population (N) was set to constant to 300,000, the number of initial population of the infectious state was set to one, the length of nucleotide sequences was set to 500, and the number of sequences samples for Tajima's D value was calculated for each time unit to 10. One generation was assumed to be one day. The population of the infectious changed based on discrete SIR compartment model, which is given by,

$$\begin{aligned} S(t+\Delta t) &= S(t) - \Delta t \beta S(t) I(t) \\ I(t+\Delta t) &= I(t) + \Delta t \beta S(t) I(t) - \Delta t \gamma I(t) \, . \\ R(t+\Delta t) &= R(t) + \Delta t \gamma I(t) \end{aligned} \tag{4}$$

$S(t)$, $I(t)$ and $R(t)$ represent the population of the susceptible, the infectious, and the removed state at time t, respectively. At time $t+\Delta t$, the number of the infectious, $I(t+\Delta t)$, was the sum of the number of the infectious at time t and newly infected population and subtraction of newly removed population at time t. The newly infected population is a product of the number of susceptible at time t, the number of the infectious at time t, and transmission rate (β). The newly removed population is a product of the number of infectious at time t and removal rate (γ). Using equation (4), the basic reproduction number ($R_0 = \beta N/\gamma$) was calculated and the β was replaced to $R_0\gamma/N$. For stochastic simulation, parameter R_0 was assumed to follow uniform distribution from 1 to 6, parameter γ was assumed to follow uniform distribution from 0.05 to 1 per day, and parameter μ was assumed to follow uniform distribution from 10^{-6} to 10^{-4} per nucleotide per day. The range of each distribution of parameter was non–informative and was decided empirically to include all possible values.

During the simulation, the number of newly infected population (X) followed Poisson distribution of which rate was $S(t)I(t)R_0\gamma/N$, and the number of newly removed population (Y) followed Poisson distribution of which rate was $I(t)\ \gamma$. The viruses in each infectious individual at time t, were copied to the next individual at time $t+\Delta t$ with random mutations with equal mutation rates for all nucleotide bases

(JC69) [30]. To make a smooth population changes, Δt was set to be 0.1 day (Fig. 5). With this setting, I conducted 100,000 simulations of the SIR model.

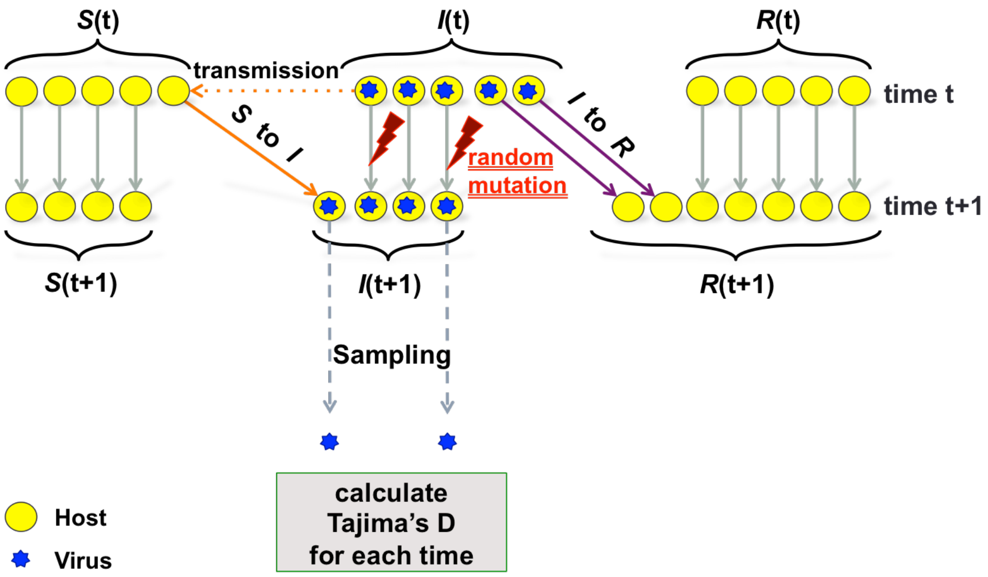


Figure 5. Simulation of SIR model with random mutations when $\Delta t = 1$: a conceptual overview

Calculation of Tajima's D values for simulated data

For each simulation, viral sequence data for three days were stratified to match the unit with selected Tajima's D values. Among accumulated sequences, which contained more than 40 sequences, 10% of them were randomly sampled to calculate Tajima's D values. For every calculation, the total number of accumulated sequences was recorded.

Approximate Bayesian computation (ABC)

The purpose of the ABC in this study was to estimate the three parameters—R_0, γ, and μ. At first, Tajima's D values were computed from selected sequences as an observed data. On the other hand, Tajima's D values calculated from simulated dataset as a simulated data and were used to calculate summary statistic for the reject algorithms as below,

$$\text{summary statistics} = \sum_{k=1}^{n} \frac{\left(D_k^{(obs)} - D_{j-i+k}^{(sim)}\right)^2}{n} . \qquad (5)$$

Here $D_t^{(obs)}$ and $D_t^{(sim)}$ are the observed and simulated Tajima's D value at time t, respectively. The i and j are the time point when $D_t^{(obs)}$ and $D_t^{(sim)}$ take the minimum value, respectively. If $D_{j-i+k}^{(sim)}$ are undefined the summary statistics is $+\infty$. The equation (5) is the mean square error (MSE) between the time evolution of Tajima's D values of observed sequences and those values of simulated sequences. If the MSE was less than 0.3, then the parameters used for the simulation were collected to estimate posterior distributions of the parameters.

Estimating effective reproduction number ($\boldsymbol{R}_{\mathrm{E}}$)

Estimating R_{E} was performed using the Bayesian evolutionary analysis by sampling trees (BEAST) 2 program (BDSKY add-on) [28,37] using JC69 [30] for substitution model with strict clock. "Birth Death Skyline Serial" was selected for tree prior. For other priors, except sampling proportion (beta distribution with default), uniform distributions were set as same as the priors that were used previously. For estimating serial R_{E} values, the number of dimension was selected to ten in prior setting of reproduction number. The analysis was performed with the Markov chain Monte Carlo (MCMC) algorithm, running 10 million chains with logging data for every 5,000 chains and MCMC result was analyzed by Tracer v1.6 program.

Results

Data selection and calculation of Tajima's D

A study conducted by Barrero *et al.* was qualified by conditions [36]. Sequences were collected from patients in children hospital located in Buenos Aires from 25th May to 24th August (Fig. 6(a)). A total of 265 NA gene segment sequence information—34 of full sequence data and 231 of fragment sequence data were used in this study. The number of sequences from 23rd June to 25th June was the peak, 58. After alignment, 212 sequences with 357 base pair were remained and 8 consecutive Tajima's D values were calculated from 15th June to 7th July. The minimum Tajima's D value, −1.926, was computed using sequences from 26th June to 28th June (Fig. 6(a)).

Estimation of epidemiological parameters

After 100,000 of stochastic simulations, 70,796 simulations were successful. Among them, 69,614 simulations were rejected after ABC algorithm. From accepted 1,182 simulations, three parameters from each simulation— R_0, γ, and μ—were stratified then plotted posterior distributions.

Simultaneously, each parameter was estimated with 95% credible interval (CI). The estimated mode value of R_0 was 1.47 (95% CI; 1.19– 4.93), and the estimated mode value of γ was 0.12 /day (95% CI; 0.1– 0.92), and the estimated mode value of μ was $1.63*10^{-4}$ substitutions/nucleotide/day (95% CI; $7.13*10^{-5}$ – $4.48*10^{-4}$) (Fig. 7). After collecting the date of epidemic start, epidemic peak, and epidemic end in simulated SIR data, the distribution of each date was plotted. The mode date of epidemic start was 7th June (95% CI: 14th May – 10th June), epidemic peak was 19th July (95% CI: 4th July – 15th August), and epidemic end was 23rd

October (95% CI: 3rd August – 28th November) (Fig. 8). Accumulated curve of the infectious of every accepted simulation with date showed overall distribution of epidemic dates (Fig. 9).

Effective reproduction number (*R*$_\mathrm{E}$)

After Bayesian MCMC approach with strict clock, the mean of R_E was estimated from ten divided time units. From time unit one to time unit seven, the mean values of R_E was estimated between 1 and 2. For the time unit 1, the estimated mean R_E was 1.42 (95% HPD; 0– 3.46). The value increased to 4.45 (95% HPD; 3.60– 5.00) sharply from time unit 7 to 8 then it decreased blow one from time unit 8 to 10 gradually (Fig. 10).

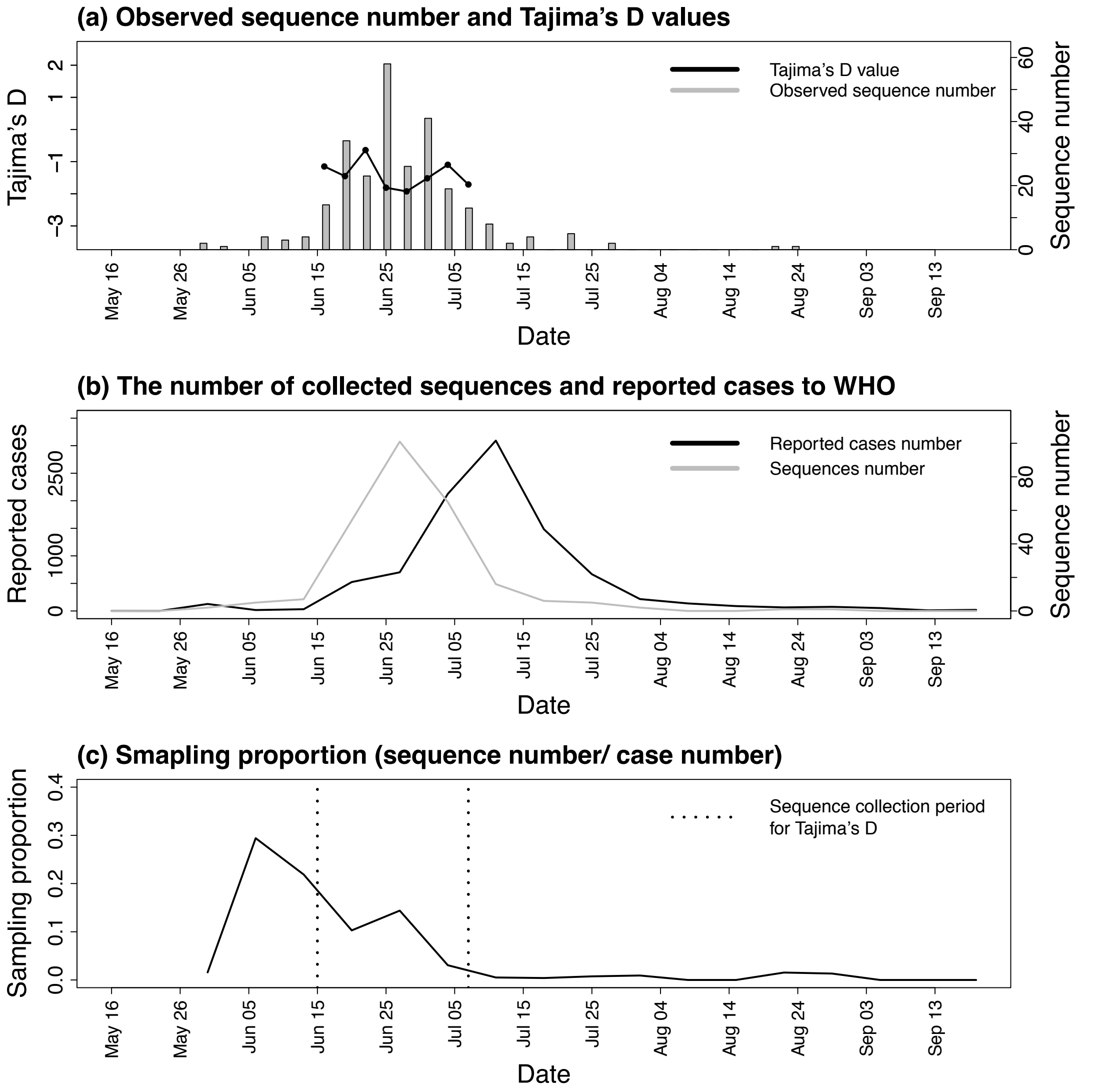


Figure 6. Sequence sampling information from pandemic influenza 2009 in Buenos Aires

(a) The number of collected sequence by three–days unit and the calculated Tajima's D value after alignment from selected study conducted by Barrero *et al.*. Black circles represent Tajima's D values using every sequence data in each time unit. Gray bar represent number of accumulated sequence data during each three–day unit. (b) The numbers of reported case number to WHO and sampled sequence number. Black line represents weekly–reported case report from FluNet and gray line represents weekly–accumulated sequence number. (c) Sampling proportion for each point. Gray dot–line represents period when Tajima's D values were calculated.

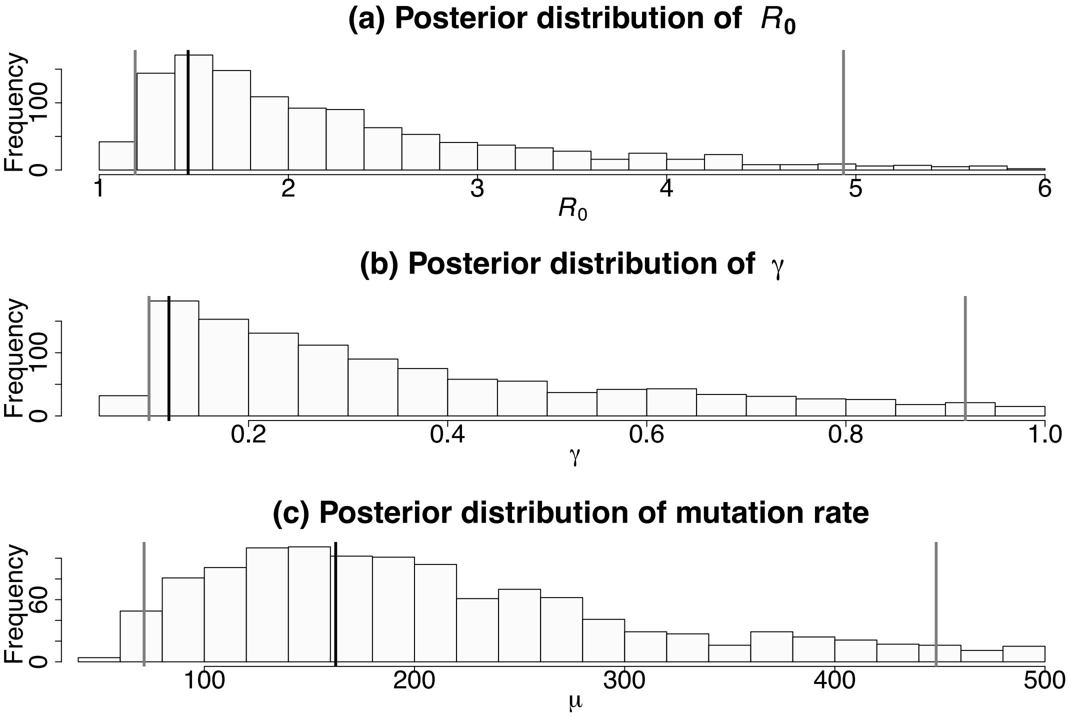


Figure 7. Posterior distribution of three parameters, basic reproduction number (R_0), removal rate (γ), and mutation rate (μ) after ABC rejection algorithms

(a), (b), (c) show posterior distribution of R_0, γ, and μ, respectively. Black vertical lines represent mode value and gray vertical lines represent 95% credible intervals.

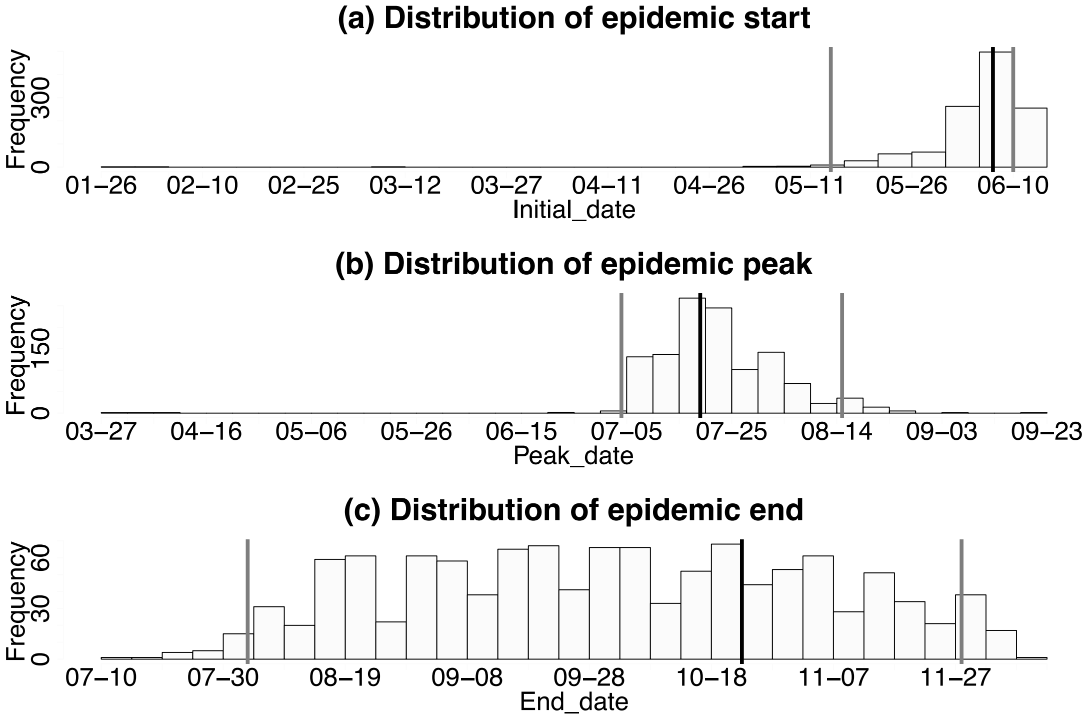


Figure 8. Posterior distribution of epidemic start, epidemic peak, and epidemic end

Black vertical line represents mode value of epidemic peak date and gray vertical lines represent 95% credible interval.

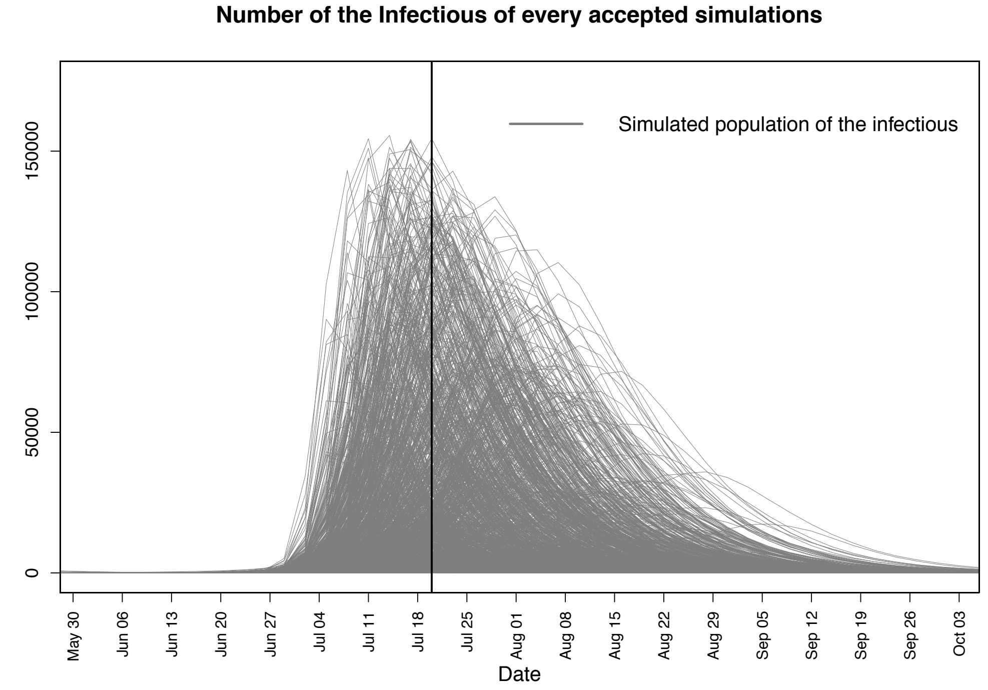


Figure 9. The changes of the population in the infectious in 1,182 accepted simulations

Each replicate is based on each parameter drawn randomly from each prior distribution as described in Materials and Methods. Black vertical line represents estimated mode value of epidemic peak date, 25th July.

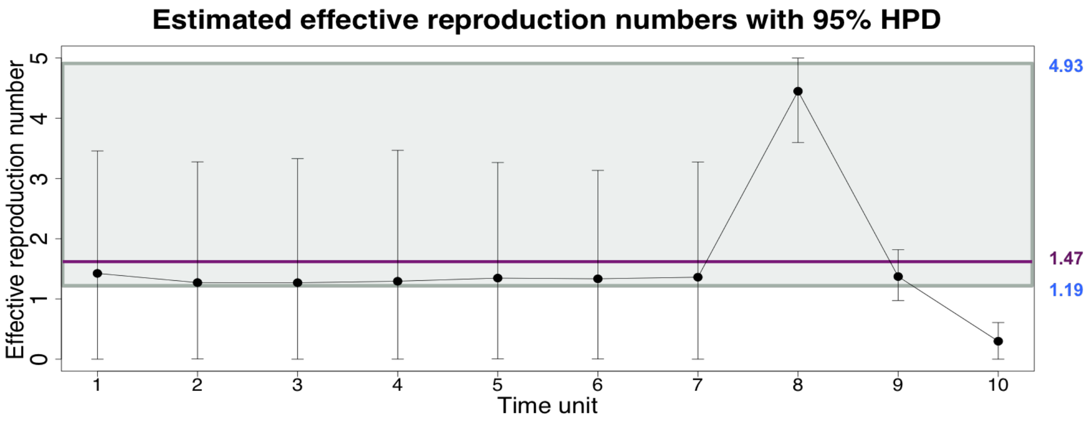


Figure 10. Time serial estimated mean R_E values using BEAST 2 program

Black circles represent estimated mean R_E and the vertical lines mean 95% HPDs. Purple horizontal line represents estimated mode R_0 using ABC and the gray box represents its 95% credible interval.

Discussion

In this chapter, I introduced a new method to estimate epidemiological parameters— R_0, γ, and epidemic peak— using time evolution of Tajima's D values. I showed that distance between Tajima's D values of given sequence data and those calculated from SIR model simulations could be used for summary statistics in ABC algorithm.

BEAST is program developed for Bayesian analysis of molecular sequences using MCMC to reconstruct phylogenies in 2007 [38]. This program had been updated to BEAST2 with various extensions, recently [37]. Using BEAST2 program with BDSKY add–on and the same sequence information, I estimated R_E values and compared it with our estimation. The mean value of R_E at first time unit was 1.42 (95% HPD; 0– 3.46) and it was consistent with the mode value of our basic reproduction number, 1.47 (95% HPD; 1.19– 4.93). Though R_E is not identical to R_0, I can assume that these two value is theoretically similar because I used the dataset of pandemic influenza, which I can ignore initial immunity effect on the susceptible population. Also, R_0 itself could be calculated using cumulative incidence of outbreaks, which is increased exponentially constant growth rate, initially [20]. Serial R_E values, except time unit 8 and 10, showed similar values with the first estimated value (Fig. 10). The reason for inconsistence in time unit 8 and 10 would be sampling proportion. In this study, sampling proportion was calculated dividing weekly observed sequence numbers by weekly case report numbers downloaded from FluNet data ("FluNetLaboratorySurveillanceData," 2016.02.19.). Sampling proportions during 15th June to 7th July, Tajima's D value available period, showed decreasing tendency as times go by (Fig. 6(b) and 6(c)). This means that though Tajima's D value could be calculated using only more than four sequences data, if sampling proportion is too low, this value might not be enough to represent the population state.

I also compared my estimation of R_0 with previous studies, which used epidemic data of pandemic influenza H1N1 (2009) in countries in southern hemisphere. The mean value of estimated R_E of 18 values from 5 countries—Australia, Chile, New Zealand, Peru, and South Africa—was 1.50 (2.5–97.5 percentile: 1.19–2.21) [40]. Especially, the mean of R_E in Chile (bordering country with Argentina) was ranged from 1.19 to 1.8, which was overlapped with my estimation.

This method had several advantages. At first, my estimation used Tajima's D values as a statistics and used simple technique for summary statistics and did not need to reconstruct genealogy from sampled sequences. Second, this method had a good expandability. I could describe specific infectious disease with mutations explicitly by adding compartment, which could explain infectious disease precisely. Thirds, I could interpret the tendency of Tajima's D values calculated from sampled sequence data of accepted simulations. Changes of Tajima's D value could reflect population changes or selection pressure altogether [14]. The minimum Tajima's D value calculated from observed data was −1.92 and this meant that this viral population did not follow Wright–Fisher model with 95% confidence based on beta–distribution criteria [13]. This inferred that the viral population was growing or under purifying selection. Intuitively, it is easy to understand that the viral population is growing, because the sequence data was collected from the patients suffered from pandemic influenza. Barrero *et al.* also showed that overall dN/dS of this sequence data was 0.14 and there were 11 negatively selected codons [36]. The ratio below one could be interpreted as purifying selection. At last, this method could estimate epidemic start, peak, and end date. Estimated epidemic start date, 7th June (95% credible interval; 14th May – 10th June) (Fig. 8(a)), was consistent with the first isolation date in selected study and FluNet data, 27th May and 16th May, respectively. Estimation epidemic peak was available even the sequences data were collected before real epidemic peak had arrived. Though there was 8 days discrepancy between the estimated date, 19th July (95% credible interval; 4th July – 15th August), and observed data in FluNet, 11th Jul, but it was still located in 95% credible interval (Fig. 8(b)). This gap could be explained by the source of each data. The sequence data that used for estimation was collected only from Buenos Aires but FluNet data represent a whole country, Argentina. The estimation of epidemic end was 23rd October (95% credible interval; 3rd August – 28th November) (Fig. 8(b)) and observed last sequence was collected on 24th August, which was still in credible interval (Fig. 8(c)), but it did not match with FluNet data. Reason for this estimation would be its weak seasonality of H1N1 after pandemic in Argentina. To estimate accurate epidemic end date, I might need to modify compartment model to apply this seasonality.

There exist disadvantages also. At first, the more accurate estimation I have, the more computational expense I need. If I want to increase sensitivity of summary statistics—if I decrease the criteria of summary statistics from 0.3 to 0.1—, I could get more accurate estimation but this would take much more time than current estimation [41]. Though the time consumption is flexible depends on initial parameters and distribution of priors, it took seven days for 100,000 simulations in this study. Second, our method would be useful only to rapidly evolving viruses such as RNA virus, of which rapid mutation could explain it adjustment in new environment in short time [42]. Third, though, Tajima's D was a statistic that could infer population changes and selective pressure, it is difficult to quantify the extent of each effect on Tajima's D value [43]. Fourth, this method did not consider recombination. Fortunately, the sequences that I used in this study did not have evidence for recombination [43] but we should test whether or not the recombination was detected in the sequences data such as generic algorithm recombination detection (GARD) and single breakpoint recombination (SBP) to get accurate estimation.

This method succeeds to estimate the epidemiological parameters in Influenza pandemic 2009. The estimations were highly consistent with precedent researches and statistics. This method also showed some limitations. Current method could not tell how extent the population size or selection pressure affects on Tajima's D values. This method also did not consider possible recombination event that could affect evolutionary dynamic in pathogens. Further studies are needed to solve these limitations.

Summary

Estimating epidemiological parameters at initial stage of epidemic is important to establish effective control strategies of infectious diseases. Here, epidemiological parameters were estimated by approximate Bayesian computation (ABC) using Tajima's D. At first, NA gene sequence data during 2009 pandemic influenza in Buenos Aires were collected and stratified into 7 datasets according to their isolation dates. Then simulations of SIR model with mutations were conducted and randomly selected sequences after simulation were stratified in same manner. If the distance between Tajima's Ds of observed sequences and those of sequences evolved in the simulations were acceptable, then the parameters that were used for the simulation were assembled to make posterior distributions, respectively. After all, the mode value of R_0 was estimated to be 1.47 (95% CI; 1.19– 4.93). This estimation was consistent with other precedent researches. The mode of epidemic peak was estimated as 19th July (95% CI; 7th July – 30th August). Estimated epidemic peak was also consistent with WHO report. This analysis showed that epidemiological parameters of pandemic influenza (2009) could be estimated successfully using ABC with Tajima's D. I anticipate that this method could be applicable to another infectious diseases.

Conclusion

Tajima's D value of aligned sequence information informed us whether or not the sampled population follows WF-model. This value is helpful to estimate population changes or selective pressure given sequence information.

In chapter I, I calculated host-specific and segment specific Tajima's D values of influenza A viruses through a systematic review using viral sequences registered in the NCBI database. Sequences encoding external proteins of influenza A viruses showed positive Tajima's D in wild mallards, suggesting the existence of balancing selection, although zero or negative Tajima's D was expected. This result suggests the existence of missing factors other than low immune response or low pathogenicity to maintain the variation of the subtypes circulating in the natural hosts.

In chapter II, I extended application of Tajima's D to estimating epidemiological parameter with Bayesian approaches. I showed successful estimation of R_0, γ, epidemic peak, and μ of pandemic influenza H1N1(2009) in Buenos Aires, Argentina using viral sequence information. I applied ABC algorithm to bypass unsolvable ODE and to select accepted parameters from prior distributions. Our estimation was consistent with previous researches which using epidemic data in neighboring countries. I also confirmed our result is consistent with those obtained by BEAST2 program based estimation using same sequence. I expect this estimating method could be applied to other infectious disease if it is applied with appropriate compartment model with mutations.

Through the population genetic analyses of nucleotide sequences of influenza A viruses isolated from humans, chickens, and wild ducks, I showed that Tajima's D can be useful not only to understand population genetics of the pathogens but also to estimate epidemiological parameters of infectious diseases caused by the pathogens. I hope these knowledge obtained through this research would contribute to the control of zoonotic infectious diseases, of which pathogens are maintained in wild animal population and occasionally cause outbreaks in human population.

Acknowledgements

First of all, I greatly appreciate to my supervisor, Professor Kimihito Ito (Division of Bioinformatics, Hokkaido University Research Center for Zoonosis Control, Sapporo, Japan) whose kind and patient guidance and active discussions really helped me to conduct research and complete writing thesis. I am also grateful to Assistant Professor Ryosuke Omori (Division of Bioinformatics, Hokkaido University Research Center for Zoonosis Control, Sapporo, Japan) for his technical assistance and active discussions. Also I am grateful to the advisors who gave academic comments and advise to my research; Professor Ayato Takada (Division of Global Epidemiology, Hokkaido University Research Center for Zoonosis Control, Sapporo, Japan), Associate Professor Norikazu Isoda (Unit of Risk Analysis and Management, Hokkaido University Research Center for Zoonosis Control, Sapporo, Japan), and Associate Professor Kanako Koyanagi (Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan) .

I am thankful to Dr. Ueno Keisuke (Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia), who does help me to understand computer language. It was really helpful for me to develop coding to conduct researches. I also want to express my gratitude to Mr. Nipawit Karnbunchob , Mr. Jednipit Borthong, Ms. Heidi Tessmer, and Ms. Wessam Mohamed Ahmed, who are PhD students in Division of Bioinformatics, Hokkaido University Research Center for Zoonosis Control, Sapporo, Japan. Not only academic discussions with them but also splendid memories could make me continue all of researches in Japan. I feel really thankful to Dr. Wei Shi and Dr. Gabriel Gonzalez (Division of Bioinformatics, Hokkaido University Research Center for Zoonosis

Control, Sapporo, Japan), who gave me helpful advise from various point of view. Also I am greatly appreciated to Ms. Sayaka Iida for her sincere technical support.

Finally, I want to express my deep appreciation to my family, who all the time encourages me with deep patience.

和文要旨

感染症病原体の集団遺伝学的理解は，その病原体によって引き起こされる感染症を制御する上で重要である。病原微生物の集団にみられる遺伝子変異には，その微生物集団の過去の集団サイズと進化動態に関する情報が含まれている。Tajima の D は，解析の対象とする集団が，一定サイズのひとつ集団のもとで中立的に進化をしているか否かを検定する指標である。

本研究では，はじめに，米国国立生物工学情報センターに登録されている塩基配列を用いたシステマティックレビューにより，A 型インフルエンザウイルスの分節特異的および宿主特異的な Tajima の D の値の傾向を解析した。ウイルス集団の分断によるバイアスを避けるために，ウイルスの塩基配列をそれらの分離年と分離場所ごとに層化した。その結果，同じ年に同じ場所で同じ宿主動物から分離された A 型インフルエンザウイルスの塩基配列集合 580 セットを得た。これらの塩基配列集合を解析した結果，ウイルスの塩基配列の Tajima の D の値は，宿主および遺伝子分節によって異なることが判明した。ニワトリおよびヒトから分離したインフルエンザウイルスの Tajima の D の値は負であり，ウイルス株間での浄化選択または集団拡大が起きていることが示された。ウイルスの集団サイズの急激な増加により，Tajima の D が負の値をとることをコンピューターシミュレーションによっても確認した。野生のカモから分離されたインフルエンザの PB2，PB1，PA， NP および M 遺伝子においては，Tajima の D がおよそ 0 であり，これらの遺伝子は，一定サイズの集団のもとで中立的に進化していることが示唆された。一方， HA, NA および NS 遺伝子の Tajima の D は正であり，野生のカモにおいて，HA,

NA および NS が平衡選択を受けていることが示された。これらの結果は，野生のカモにおいてインフルエンザウイルスの亜型の多様性を保持する未知のメカニズムの存在が示唆された。

次に，本研究では，感染症流行時の病原体の塩基配列 Tajima の D の時系列変化から，感染症の流行を特徴付ける疫学的パラメータを推定する手法を開発した。米国国立生物工学情報センターのデータベースから，2009 年のインフルエンザパンデミック時にブエノスアイレスで分離された H1N1 亜型の A 型インフルエンザの NA 遺伝子の塩基配列 265 本を取得した。塩基配列の Tajima の D の時系列変化と，感染症流行における遺伝子変異のコンピューターシミュレーションで得られる Tajima の D の時系列変化とを比較した。近似ベイズ計算を用いることにより，ブエノスアイレスにおける 2009 年のパンデミックインフルエンザの基本再生産数の最頻値は， 1.47 (95% 信用区間: 1.19 – 4.93)であることが推定された。同様に，回復率の最頻値は 0.12／日(95%信用区間:0.1– 0.92)，流行のピークは 7 月 19 日 (95% 信用区間：7 月 7 日– 8 月 30 日)と推定された。これらの推定値は，これまでの他研究および世界保健機関の報告と無矛盾である。感染症流行初期における疫学的パラメータの推定は効果的な制御対策の策定に重要である。本研究で開発した手法は，病原体を限定しないため，インフルエンザ以外の感染症における疫学的パラメータの推定にも有用であると考える。

References

1. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of influenza A viruses. Microbiological reviews. 1992;56: 152–79.
2. Wu Y, Wu Y, Tefsen B, Shi Y, Gao GF. Bat-derived influenza-like viruses H17N10 and H18N11. Trends in microbiology. 2014;22: 183–91.
3. Olsen B, Munster VJ, Wallensten A, Waldenström J, Osterhaus ADME, Fouchier RAM. Global patterns of influenza a virus in wild birds. Science (New York, NY). 2006;312: 384–8.
4. Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, et al. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science (New York, NY). 2009;325: 197–201.
5. Kilbourne ED. Influenza Pandemics of the 20th Century. 2006;12: 9–14.
6. Alexander DJ. An overview of the epidemiology of avian influenza. Vaccine. 2007;25: 5637–44.
7. Capua I, Alexander DJ. Avian influenza: recent developments. Avian pathology : journal of the WVPA. 2004;33: 393–404.
8. Vincent AL, Lager KM, Anderson TK. A brief introduction to influenza A virus in swine. Methods in molecular biology (Clifton, NJ). 2014;1161: 243–58.
9. Webster RG, Govorkova EA. Continuing challenges in influenza. Annals of the New York Academy of Sciences. 2014; doi:10.1111/nyas.12462
10. Freidl GS, Meijer a, de Bruin E, de Nardi M, Munoz O, Capua I, et al. Influenza at the animal-human interface: a review of the literature for virological evidence of human infection with swine or avian influenza viruses other than A(H5N1). Euro surveillance : bulletin Européen sur les maladies

transmissibles = European communicable disease bulletin. 2014;19.

11. Kida H, Yanagawa R, Matsuoka Y. Duck influenza lacking evidence of disease signs and immune response. Infection and immunity. 1980;30: 547–53.

12. Robinson DA, Falush D, Feil EJ, editors. Bacterial Population Genetics in Infectious Disease [Internet]. 1st ed. WILEY-BLACKWEL. New Jersey: Wiley-Blackwell; 2010.

13. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123: 585–95.

14. Tajima F. The effect of change in population size on DNA polymorphism. Genetics. 1989;123: 597–601.

15. Innan H, Stephan W. Letter to the Editor The Coalescent in an Exponentially Growing Metapopulation and Its Application to Arabidopsis thaliana. 2000; 3–7.

16. Thomas JC, Godfrey P a., Feldgarden M, Robinson DA. Candidate targets of balancing selection in the genome of Staphylococcus aureus. Molecular Biology and Evolution. 2012;29: 1175–1186.

17. Frost SD, Dumaurier MJ, Wain-Hobson S, Brown AJ. Genetic drift and within-host metapopulation dynamics of HIV-1 infection. Proceedings of the National Academy of Sciences of the United States of America. 2001;98: 6975–6980.

18. Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. Princeton University Press; 2008.

19. Nishiura H, Chowell G, Heesterbeek H, Wallinga J. The ideal reporting interval for an epidemic to objectively interpret the epidemiological time course. Journal of the Royal Society, Interface / the Royal Society. 2010;7: 297–307.

20. Roberts MG, Heesterbeek JAP. Model-consistent estimation of the basic reproduction number from the incidence of an emerging infection. Journal of mathematical biology. 2007;55: 803–16.

21. Hsieh YH, Ma S. Intervention measures, turning point, and reproduction number for dengue, Singapore, 2005. American Journal of Tropical Medicine and Hygiene. 2009;80: 66–71.

22. Chiew M, Gidding HF, Dey A, Wood J, Martin N, Davis S, et al. Estimating the measles effective reproduction number in Australia from routine notification data. Bulletin of the World Health Organization. 2014;92: 171–7.

23. Morgan ER, Milner-Gulland EJ, Torgerson PR, Medley GF. Ruminating on complexity: Macroparasites of wildlife and livestock. Trends in Ecology and Evolution. 2004;19: 181–188.

24. Nishiura H, Yan P, Sleeman CK, Mode CJ. Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks. Journal of theoretical biology. 2012;294: 48–55.

25. Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. 2000;155: 1429–1437.

26. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nature reviews Genetics. 2009;10: 540–50.

27. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW. Phylodynamics of infectious disease epidemics. Genetics. 2009;183: 1421–30.

28. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth – death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Pnas. 2013;110: 228–233.

29. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 2013;30: 772–780.

30. Jukes TH, Cantor CR. Evolution of Protein Molecules. In: Munro HN, editor. Mammalian Protein Metabolism. New York: Academic Press; 1969. pp. 21–123.

31. Bahl J, Vijaykrishna D, Holmes EC, Smith GJD, Guan Y. Gene flow and competitive exclusion of avian influenza A virus in natural reservoir hosts. 2009;390: 289–297.

32. Chen R, Holmes EC. Hitchhiking and the Population Genetic Structure of Avian Influenza Virus. 2010; 98–105.

33. Dugan VG, Chen R, Spiro DJ, Sengamalay N, Zaborsky J, Nolting J, et al. The Evolutionary Genetics and Emergence of Avian Influenza Viruses in Wild Birds. 2008;4.

34. Simonsen KL, Churchill GA, Aquadro CF. Properties of Statistical Tests of Neutrality for DNA Polymorphism Data. 1995;429.

35. Ratmann O, Donker G, Meijer A, Fraser C, Koelle K. Phylodynamic inference and model assessment with approximate bayesian computation: influenza as a case study. PLoS computational biology. 2012;8: e1002835. doi:10.1371/journal.pcbi.1002835

36. Barrero PR, Viegas M, Valinotto LE, Mistchenko a S. Genetic and phylogenetic analyses of influenza A H1N1pdm virus in Buenos Aires, Argentina. Journal of virology. 2011;85: 1058–66.

37. Bouckaert R, Heled J, K??hnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS

Computational Biology. 2014;10: 1–6.

38. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology. 2007;7: 214.

39. FluNetLaboratorySurveillanceData [Internet]. [cited 19 Feb 2016].

40. Biggerstaff M, Cauchemez S, Reed C, Gambhir M, Finelli L. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. BMC infectious diseases. 2014;14: 480.

41. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. Approximate Bayesian Computation. PLoS Computational Biology. 2013;9.

42. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nature reviews Genetics. 2008;9: 267–76.

43. Kim K, Omori R, Ueno K, Iida S, Ito K. Host-Specific and Segment-Specific Evolutionary Dynamics of Avian and Human Influenza A Viruses: A Systematic Review. PloS one. Public Library of Science; 2016;11: e0147021.