Inferring the population structure of Influenza A viruses with Tajima's D and its application to surveillance

KIM, Kiyeon
Research Center for Zoonosis Control, Hokkaido University

Ito, Kimihito
北海道大学人獣共通感染症リサーチセンター

https://hdl.handle.net/2324/1522094

Inferring the population structure of Influenza A viruses with Tajima's D and its application to surveillance

北海道大学　人獣共通感染症リサーチセンター　　　Kiyeon KIM
北海道大学　人獣共通感染症リサーチセンター　　　Kimihito ITO

概 要

Influenza A virus is a zoonotic pathogen which infects various hosts. Mallard is a natural host of influenza A viruses. If adapted to non-natural host, it could be highly pathogenic avian influenza virus in chicken or influenza pandemic in human. Thus, early detection a hidden outbreak in the new host species is important to control both avian and human influenza. So, we propose Tajima's D to detect viral demography in host. Influenza A viruses can be assumed to undergo neutral evolution in its natural host, and Tajima's D in this situation must be zero. But once the influenza A virus adapt to new host, Tajima's D values change to non-zero.　By monitoring the value of Tajima's D on viral nucleotide sequences, we could detect hidden outbreak. By analyzing 1,347 nucleotide sequences from the 6 published research papers, we show that the mean values of Tajima's D are different depending on segments. Tajima's D of internal genes of influenza viruses isolated from wild mallard sample is close to zero, meaning that the population of Influenza A virus in wild mallard is stable. When more than two subtypes of HA, NA and NS gene segment included in dataset, the Tajima's D showed positive tendency.

1. 序論

Influenza A virus is a zoonotic pathogen that is origin from mallard. Totally, 16 serotype of hemagglutinin (HA) and 9 subtypes of neuraminidase (NA) are isolated from mallard (Webster, Bean, Gorman, Chambers, & Kawaoka, 1992). Influenza A viruses infect a various range of hosts with different serotype. Early detection of the hidden circulation of influenza A viruses in non-natural host is imortant to control human and animal influenza. When viral infection to hosts is asymptomatic, the detection of circulating virus is difficult and it might remains unrecognized for a long time. This unrevealed outbreak would lead to further spread of viruses to other hosts including humans. The immunological pressure affects the evolution of viral sequences.　When hosts are exposesd to viral pahtogen for the first time, the hosts activate primary immune response and clear the viruses without selection pressure (Bean et al., 2013). And this results in neutral evolution. If the same virus infects the host more than one time, the host activates secondary immune response with immunological memory. However, if the primary immune response is too weak to activate adaptive immunity, then the same

viruses can infect to the hosts multiple times with neutral evolution of viruses (Kida, Yanagawa, & Matsuoka, 1980). Assuming that viral evolution is neutral, informs us viral demography. Based on coalescent theory, Tajima's D informs whether the observed variation is identical to the expected variation in given nucleotide sequences or not. The Tajima's D value (D value) should be zero when the viruses undergo neutral evolution in constant-size population. With the assumption of neutrality, if D value is negative, the population is inferred to be increasing or bottelneck situation with recent excessive mutations. If D value is positive, the population is declining or subdividing situation with of mutations in the past (Tajima, 1989b).

In this study, we introduce Tajima's D in detecting the viral demography in their hosts. Using public sequence data of influenza A viruses isolated from mallards, we analyze their D value. According to our hypothesis, Tajima's D vlaue of sequence samples of pathogens gives us information whether the pathogen population is increasing or not, possibly contributing the early detection of the hidden outbreak of zoonotic pathogens.

2. 本論

Tajima's D is a statistic that can be used to test whether or not the population structure of target organisms would follow the Wright-Fisher model. Wright-Fisher model starts from two assumptions. First, the population of target organisms is assumed to be selectively neutral. Second, the population does not have demographic structure—the population is constant-size and is not subdivided. Using nucleotide sequence data from surveillance study, Tajima's D can test whether these assumptions holds with the population or not. Mathematically, Tajima's D is normalized difference between two mutation parameters—Watterson's estimator and Tajima's estimator. Tajima's D is derived by subtracting Watterson's estimator from Tajima's estimator and by normalizing its numerator (eq. 1) (Tajima, 1989a).

$$D = (\theta\_T - \theta\_W)/Std(\theta\_T - \theta\_W)$$ ………………..(1)

If there are less than four nucleotide sequences in sample, the denominator of Tajima's D becomes zero. Samples should have at least four sequences to calculate Tajima's D.

We used public sequence dataset of influenza A viruses of wild mallard by screening research articles. To restrict each dataset to single population, we selected papers, which conducted surveillance in a same area. Totally, 6 surveillance papers for wild mallard were found to be satisfied with the conditions. From these studies, 1,347 viral sequences from wild mallard were collected from Influenza Virus Resource in National Center for Biotechnology Information (NCBI).

Each dataset was divided into host and sampling area, then separated each group into 8 gene

segments. Then each segmented groups was aligned using MAFFT multiple sequence alignment program (version 7). Next, these aligned segment groups regrouped chronologically by year or season. If the gene segment group has subgroups or alleles, then it was subdivided into each subgroups or alleles for further analysis. Finally, D value was calculated for each time period in each research data. We also used bootstrap resampling analysis, 1500 replicates, to calculate 95% confidence interval of D value for each time period. Every classification of nucleotide sequences and computation of D value were programed by Perl 5 (v5.16.0).

D values were regrouped into each segment by hosts. Then, average and standard deviation value were computed for every segment in each host. To compare D value between hosts, box plot was drawn for each gene segments for each hosts. Additionally, mean value for the average of every segment in each host were calculated.

3. 結論

Influenza A viruses isolated from wild mallard had Tajima's D with a mean of 0.563 and a standard deviation of 1.144. The mean values of Tajima's D in PB2, PB1, PA, HA, NP, NA, MP, and NS gene segments were 0.037, 0.026, 0.203, 1.416, -0.081, 1.829, 0.002, and 0.637, respectively (Table 1).

D values of internal gene segments—PB2, PB1, PA, NP, MP, and NS—and gene segments of surface proteins—HA and NA—had different tendency (Fig 1A). D values of internal gene segment, except NS gene segment, had means and medians near zero with narrow interquartile range (IQR). The D values of PB2, PB1, PA, NP, MP genes had a mean of 0.036. These D values close to zero indicated that internal gene segments except NS had undergone neutral evolution in constant-size population in mallard. D value of NS gene segment showed similar mean value with other internal gene segments, but it had wide IQR with large standard deviaiton comparing to other seven gene segments (Fig 1A, Table 1). D value of HA and NA gene segment were clearly positive with narrow IQR. The mean of D value of HA and NA was significantly larger than that of internal gene segments ($p < 0.05$). Assuming neutral evolution, this positive D value indicated that HA and NA segments evolved in a decreasing population or subdividing population.

Watkins et al. classified the detection methods of outbreak into descriptive, derived, epidemiological and simulation methods (Watkins, Eagleson, Hall, Dailey, & Plant, 2006). These methods were based on number of disease cases. In contrast, our approach can be applied by only nucleotide sequence data of viral pathogens to detect viral demography with assumption that each sampled influenza A virus represent the viral population in each host and

whole sampled influenza A virus represent whole viral population in host population. Our results supported that influenza A viruses in their natural host experienced neutral selection without immune pressure(van de Sandt, Kreijtz, & Rimmelzwaan, 2012). For primary infection, influenza A viruses are cleared by primary immune response without any selection. During this process, natural host does not show symptoms and viral infection activates low levels of inflammation. These weak reactions do not stimulate adaptive immunity and natural host could be re-infected by same strain.   Thus the mutation is not prerequisite for virus to reinfect the same host and thus the viruses do not experience immune selection. Therefore we can assume that the viral population undergo neutral evolution in mallard. Based on this assumption, we could interpret the mean of D value in mallard. The mean of Tajima's D for PB2, PB1, PA, NP, MP were close to zero. Thus we could conclude that this genes have evolved in a constant-size viral population in mallard.

To figure out clearly about the host effect on Tajima's D, we need apply our method to various host, except present tatger host.

In conclusion, D value might be used to detect the change of pathogen population in the hosts   with assumption that viral population follow neutral evolution. Using nucleotide sequence data of influenza A viruses isolated from natural and non-natural host, we showed D values are depending on hosts. Thus, once we get sequence samples of influenza A virus in target host, we can test whether the viral population is constant-size or not by calculating D values. So, if there is persistent surveillance of concerned pathogen in some area, D value, will give information about viral demography in the host. Knowing demography of viral population will help us to prepare or detect hidden outbreaks. We anticipate that this technique would be applicable to various zoonotic pathogens other than influenza A viruses.

[Table 1] The mean and standard deviation value of calculated Tajima's D value for representative host.
a: compiled with 8 gene segments (PB2, PB1, PA, HA, NP, NA, MP, and NS)
b: compiled with 5 gene segments (PB2, PB1, PA, , NP, and MP)
a: compiled with 3 gene segments HA, ,NA, and NS)

		All[a]	PB2	PB1	PA	HA	NP	NA	MP	NS	mean of 5[b]	mean of 3[c]
Mallard	Number	1347	148	152	143	211	155	219	153	166	751	596
	Mean D	0.536	0.037	0.026	0.203	1.416	-0.081	1.829	0.002	0.637	0.036	1.302
	SD	1.144	0.655	0.591	0.776	1.188	0.711	0.857	0.628	1.539	0.664	1.291

(A) Tajima's D value of Influenza A virus in wild mallard
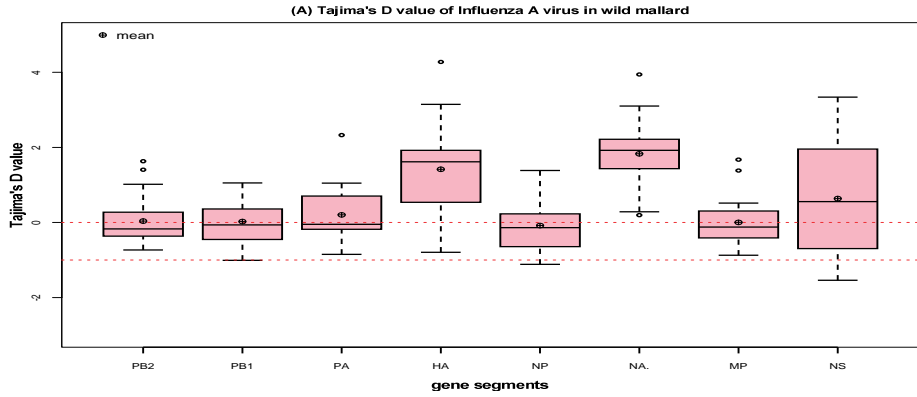
[Figure 1]   Box plot of Tajima's D value for each gene segments sampled from natural host, mallard, and human. The median and mean of Tajima's D value for each gene segments in wild mallard is close to 0, except HA, NA and partially NS gene segment. In case of HA and Ns gene segment both showed low accuracy and low

References

Bean, A. G. D., Baker, M. L., Stewart, C. R., Cowled, C., Deffrasnes, C., Wang, L.-F., & Lowenthal, J. W. (2013). Studying immunity to zoonotic diseases in the natural host - keeping it real. *Nature Reviews. Immunology*, *13*(12), 851–61. doi:10.1038/nri3551

Kida, H., Yanagawa, R., & Matsuoka, Y. (1980). Duck influenza lacking evidence of disease signs and immune response. *Infection and Immunity*, *30*(2), 547–53. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=551346&tool=pmcentrez&rendertype=abstract

Tajima, F. (1989a). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–95. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1203831&tool=pmcentrez&rendertype=abstract

Tajima, F. (1989b). The effect of change in population size on DNA polymorphism. *Genetics*, *123*(3), 597–601. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1203832&tool=pmcentrez&rendertype=abstract

Van de Sandt, C. E., Kreijtz, J. H. C. M., & Rimmelzwaan, G. F. (2012). Evasion of influenza A viruses from innate and adaptive immune responses. *Viruses*, *4*(9), 1438–76. doi:10.3390/v4091438

Watkins, R. E., Eagleson, S., Hall, R. G., Dailey, L., & Plant, A. J. (2006). Approaches to the evaluation of outbreak detection methods. *BMC Public Health*, *6*(1), 263. doi:10.1186/1471-2458-6-263

Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M., & Kawaoka, Y. (1992). Evolution and ecology of influenza A viruses. *Microbiological Reviews*, *56*(1), 152–79. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=372859&tool=pmcentrez&rendertype=abstract