Virus Evolution, 2022, **8(1)**, 1–8

DOI: https://doi.org/10.1093/ve/veac013
Advance access publication date: 21 February 2022
Research Article

Effects of host and pathogenicity on mutation rates in avian influenza A viruses

Gwanghun Kim,[1] Hyun Mu Shin,[1,2,3,4] Hang-Rae Kim,[1,5,2,3,4,†] and Yuseob Kim[6,*,‡]

[1]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul 03080, Republic of Korea, [2]BK21 FOUR Biomedical Science Project, Seoul National University College of Medicine, Seoul 03080, Republic of Korea, [3]Medical Research Institute, Seoul National University College of Medicine, Seoul 03080, Republic of Korea, [4]Wide River Institute of Immunology, Seoul National University, Hongcheon 25159, Republic of Korea, [5]Department of Anatomy & Cell Biology, Seoul National University College of Medicine, Seoul 03080, Republic of Korea and [6]Division of EcoScience and Department of Life Science, Ewha Womans University, Seoul 03760, Republic of Korea

[†]https://orcid.org/0000-0002-3983-6193

[‡]https://orcid.org/0000-0002-7975-6147

*Corresponding author: E-mail: yuseob@ewha.ac.kr

Abstract

Mutation is the primary determinant of genetic diversity in influenza viruses. The rate of mutation, measured in an absolute timescale, is likely to be dependent on the rate of errors in copying RNA sequences per replication and the number of replications per unit time. Conditions for viral replication are probably different among host taxa, potentially generating the host specificity of the viral mutation rate, and possibly between highly and low pathogenic (HP and LP) viruses. This study investigated whether mutation rates per year in avian influenza A viruses depend on host taxa and pathogenicity. We inferred mutation rates from the rates of synonymous substitutions, which are assumed to be neutral and thus equal to mutation rates, at four segments that code internal viral proteins (PB2, PB1, PA, NP). On the phylogeny of all avian viral sequences for each segment, multiple distinct subtrees (clades) were identified that represent viral subpopulations, which are likely to have evolved within particular host taxa. Using simple regression analysis, we found that mutation rates were significantly higher in viruses infecting chickens than domestic ducks and in those infecting wild shorebirds than wild ducks. Host dependency of the substitution rate was also confirmed by Bayesian phylogenetic analysis. However, we did not find evidence that the mutation rate is higher in HP than in LP viruses. We discuss these results considering viral replication rate as the major determinant of mutation rate per unit time.

Key words: avian influenza; mutation rate; host specificity; pathogenicity; neutral substitution.

1. Introduction

Mutation is a random error in copying the nucleotide sequence during DNA or RNA replication and generates the genetic and antigenic diversity of viruses for their evolutionary success (Elena and Sanjuan 2005; Duffy, Shackelton, and Holmes 2008). Influenza viruses, like other RNA viruses exhibiting error-prone replication (Steinhauer, Domingo, and Holland 1992), are characterized by frequent mutations that are essential for evading host immune responses by antigenic drift and expanding their host range (Webster et al. 1992; Woolhouse, Haydon, and Antia 2005; Nelson and Holmes 2007; Selman et al. 2012). Mutation rate is therefore a critical parameter for understanding the persistence and emergence of influenza viruses in a wide range of vertebrate hosts. For example, new antigenic variants might be more likely to emerge in a viral population with a higher mutation rate.

In most organisms, mutation rate is defined as the frequency of errors per generation or per replication. However, since how often viral genomes replicate is difficult to observe under non-laboratory environments, viral mutation rate is usually measured as the change of nucleotide sequence per unit time (year or day) rather than per replication. Note that mutation rate is different from substitution rate (or evolutionary rate, as frequently used in virus literature), as the latter depends on whether nucleotide changes produced by mutations are lost in the population of viruses from which sequences to be compared are sampled (Duffy, Shackelton, and Holmes 2008). Namely, substitution rate is critically affected by negative or positive natural selection on new variants, which is not the subject of this study. However, if neutrally evolving sites such as synonymous sites in the protein-coding sequence are examined, substitution rate should be equal to mutation rate (Charlesworth and Charlesworth 2010). This principle holds even when natural selection occurs at linked sites (i.e. at nonsynonymous sites on the same segment) (Birky and Walsh 1988). For example, based on synonymous site substitutions only, the mutation rate of subtype H3N2 in humans was estimated to be about 0.005 per year per nucleotide site (Croze and Kim 2021).

Viral mutation rate per unit time (e.g. one year) is determined by (1) the rate of errors per replication and (2) the number of replications per unit time. In the case of influenza virus, the former is given by the fidelity of the RNA polymerase complex composed of PA, PB1, and PB2 proteins. It is not clear whether the host cellular environment can affect the fidelity of replication performed by this virus-coded complex, although it has been suggested that the host cellular effect on replication fidelity exists in other RNA viruses (Pita et al. 2007; Combe and Sanjuan 2014). The second factor for yearly mutation rate, the replication rate, is expected to be influenced by numerous factors. Theories suggest that viruses evolve to attain the optimal rate of replication, thus viral load, which is high enough to ensure transmission success but low enough to avoid heavy damage to hosts (May and Anderson 1990; Frank 1996). Hosts should also evolve to prevent viral replication from reaching a harmful level. Therefore, the adaptive evolutionary history of both the virus and host is likely to determine the rate of viral replication. Different avian and mammalian hosts possess different cellular factors that either negatively or positively affect the infection and replication of influenza A viruses (Long et al. 2019), as they have distinct co-evolutionary histories with the virus. Therefore, we may consider the host cellular environment as a major factor leading to a difference in replication rates and thus in the estimates of mutation rate per unit time. Other characteristics of hosts that affect the transmission mode and frequency of infection cycles are also known to increase the range of substitution rates in various RNA viruses (Hanada, Suzuki, and Gojobori 2004; Streicker et al. 2012; Scholle et al. 2013; Hicks and Duffy 2014).

The pathogenicity or virulence of a virus is also expected to be correlated with replication rate. While it is not clear whether the positive relationship between these parameters is universal, at least for influenza virus subtype A/H5N1, the level of viral load is a major determinant of pathogenicity in mice and ducks (Hatta et al. 2010; Boon et al. 2011; Hu et al. 2013). Thus, we predict that the mutation rate estimated from highly pathogenic (HP) influenza viruses should be higher than that from low pathogenic (LP) influenza viruses.

Whether influenza A viruses infecting poultry evolve more rapidly than in the 'natural hosts'—wild ducks, gulls, and shorebirds—has been a question of particular interest, given the general hypothesis that the rate of virus-host co-evolutionary changes diminishes in time as they reach an evolutionary equilibrium (Suarez 2000; Simmonds, Aiewsakun, and Katzourakis 2019). Domesticated birds such as chicken are considered new hosts because their population density only recently (in the evolutionary time-scale) became high enough to sustain the infection cycle of influenza virus. However, Chen and Holmes (2006) found that the evolutionary (substitution) rate in wild birds is not much slower than that in domesticated birds and mammals, suggesting that the adaptive (e.g. antigenic) evolution of the virus has not ceased in natural reservoir hosts. Again, the authors measured the rates of substitutions at all sites that include nonsynonymous changes. Therefore, their results may not provide accurate information about host-dependent mutation or replication rate.

In this study, to investigate whether influenza viruses mutate at different rates, depending on hosts, we estimated yearly mutation rates for many subsets of serially sampled influenza virus, each of which is inferred to have evolved mainly within one of four different groups of avian hosts. We also tested whether the pathogenicity of a virus is associated with an elevated rate of mutation. We observed synonymous substitutions at four large segments coding 'internal proteins'—RNA polymerases (PB2, PB1, PA) and nucleoprotein (NP). As these genes affect the host range of influenza A viruses (Shu, Bean, and Webster 1993; Neumann and Kawaoka 2006; Cauldwell et al. 2014), we expect that the rate of host switching is minimal for these segments, making it easier to find a serial sample of sequences from a viral population that evolved while maintaining its association with one particular host taxon. We stress again that, although what we measure primarily from sequences is their evolutionary changes accumulated over time, we do so to estimate the viral mutation rate per unit time from them. Therefore, in the following, we will simply call our estimates mutation rates rather than synonymous substitution rates.

2. Data and methods

2.1 Sequence data

We searched public databases (NCBI and GISAID) for genomic nucleotide sequences of avian influenza A viruses ranging from 1956 to 2019, using the keyword 'influenza'. These publicly available sequences were subject to our own curation that used Vigor annotator (Wang, Sundaram, and Spiro 2010) for segment annotation and the identification of coding sequences. Only isolates with the sequences of all eight viral segments were retained. In addition, any isolate with an ambiguous nucleotide between the start and stop codons in any segment or unclear information about host species or the location (country) of isolation was not included for further analysis. The final sets of sequences are from 12,234 isolates and available from http://avian-flu.org (Avian Influenza Database at Seoul National University College of Medicine).

2.2 Sequence alignments, tree construction, and clade isolation

Multiple sequence alignment was conducted for the open reading frame of each segment using Multiple Alignment using Fast Fourier Transform (Kuraku et al. 2013; Katoh, Rozewicki, and Yamada 2019). Then, a maximum-likelihood phylogenetic tree was constructed using IQTree (Trifinopoulos et al. 2016) with default parameters. Then, to isolate discrete clades nested within the phylogeny that are associated with particular host taxa, we visually identified all subtrees (monophyletic clades) that satisfy the following criteria: (1) have ultra-bootstrapping support value of at least 0.95; (2) contain more than 30 sequences and (3) the range of sampling times (difference between the oldest and the latest sequences in the clade) is at least 10 years. Then viral sequences within a clade were classified, according to their avian hosts, into four groups: wild birds in order *Anseriformes*, wild birds in order *Charadriiformes*, domestic ducks and chickens (order *Galliformes*). Sequences found in domestic bird hosts other than ducks and chickens were excluded.

A slightly different approach was taken to identify clades that are made of sequences from either HP or LP viruses infecting chickens. For a given segment, we first constructed a tree using all available sequences from HP chicken viruses. Then we visually identified all subtrees (thus HP clades) that satisfy the following: (1) have ultra-bootstrapping support value of at least 0.95, (2) contain more than 10 sequences and (3) the range of sampling times (difference between the oldest and the latest sequences in the clade) is at least 5 years. The same procedure, using only sequences from LP chicken viruses, was applied to isolate LP clades. Then, constructing a tree using sequences that were included in both HP and LP clades above, whether these HP and LP clades are cleanly separated was examined. Then clades formed as the mixtures of HP and LP sequences were excluded for further analyses.

2.3 Estimation of yearly mutation rate by the regression method

Mutation rate (synonymous substitution rate) was estimated for all clades in which at least 80 per cent of sequences belong to one particular host group or all sequences are sampled from either HP or LP virus. Sequences belonging to minor host groups within a clade were excluded in the following procedures. We also excluded nucleotide positions from 575 to 760 within the PA coding sequence that corresponds to a region of alternative reading frame for PA-X protein. Synonymous sequence divergence (d_S) between the ith sequence (sampled at time τ_i) and the oldest sequence (sampled at time τ_0) of a clade was calculated by the Nei–Gojobori method (Nei and Gojobori 1986) implemented in codeml in the Phylogenetic Analysis by Maximum Likelihood package (Yang 2007). Sampling times were calculated in units of days. Then the linear regression of d_S on sampling time difference ($\tau_i - \tau_0$ for all i) was performed and the slope of the regression line was taken as the estimate of mutation rate ($\hat{\mu}$), assuming that synonymous mutations are neutral, under which the rate of synonymous base substitutions should be equal to the mutation rate.

2.4 Estimation of substitution rates by Bayesian phylogenetics

Model testing was performed in IQTree, to select an appropriate substitution model with the maximum Bayesian information criterion. (1) When entire protein coding sequences were used, the general time reversal (GTR) model with gamma distribution ratio heterogeneity in four ratio categories (G4) (Tavaré 1986; Yang 1994) was chosen for the PA and NP segments, and the model with the proportion of invariant site in addition to the GTR + G4 model (GTR + I + G4) (Gu, Fu, and Li 1995) for the PB2 and PB1 segments. (2) When composite sequences using only the third position of codons were used, GTR + I + G4 was chosen for all four internal segments. For a given clade, Bayesian Evolutionary Analysis Sampling Trees (BEAST) v1.10.4 (Suchard et al. 2018) was used to calculate the substitution rate based on the model of a constant population size and a relaxed uncorrelated log-normal molecular clock. Markov chain Monte Carlo was run for 100 million steps and sampled every 10,000 steps.

2.5 Estimation of host switching rates

Phylogenetic analysis was performed by BEAST in which the host taxon, one of four groups defined above, of a sequence was given as a ‘trait’. The rate of transition (the mean of posterior distribution) between traits (thus hosts) was estimated using the symmetric substitution model and Bayesian Stochastic Search Variable Selection procedure as options (Suchard et al. 2018).

2.6 Inference of viral pathogenicity

The pathogenicity of a virus isolate was indirectly determined by the number of polybasic amino acids in the cleavage site on the HA protein (Steinhauer 1999). The position of cleavage site residues was identified by an amino acid motif that starts with P, followed by either Q, L, or E and then by RGLF. The number of basic amino acids (H, R, K) in the subsequent positions was counted. If this number was equal to or more than 4, this virus was deemed HP.

2.7 Statistical analyses

Analysis of variance (ANOVA) with the Holm–Šídák test for multiple comparisons was used to compare mutation rates among host taxa. Two-way ANOVA was performed using the mutation rate estimated from clades as a dependent variable and the segment and the major host or pathogenicity of clades as independent variables. All statistical analyses were performed using GraphPad Prism 8 (GraphPad Software, La Jolla, CA, USA). All graphs were generated using Prism 8.

3. Results

3.1 The rate of host switching

We constructed phylogenetic trees for internal gene segments (PB2, PB1, PA, NP) using all available avian influenza genome sets. Unlike the HA segment, the phylogeny of which is characterized by discrete clades forming HA serotypes that are connected by long internal branches, sequences of an internal gene segment were aligned well regardless of serotypes or hosts and therefore formed phylogenies with relatively short internal branches (Supplementary Fig. S1). There was no clear, large-scale clustering of sequences according to host species, indicating that internal gene segments have switched hosts at non-negligible rates. To quantify the rate of host switching for each segment, we applied the method of estimating trait transition rate on the phylogenetic tree, implemented in BEAST, where the trait of interest is host (the same method was used for estimating the rate of reassortment; Lu, Lycett, and Brown 2014). Here, we classified hosts into four groups: wild birds in order *Anseriformes*, wild birds in order *Charadriiformes*, domestic ducks (livestock-*Anseriformes*) and chickens (livestock-*Galliformes*). We found that host switching occurred at similar rates (about 0.03 switching per year per lineage) for these four segments (Table 1). For comparison, the host switching rates were also estimated from HA segments, for eight subtypes (H1, 2, 4, 5, 6, 7, 9, 10) separately. Except for the subtype H5 (∼0.1; confirming its particularly wide host range; Kaplan and Webby 2013), the switching rates in HA subtypes were similar or lower than those of internal gene segments. Therefore, we found that genes coding polymerases and nucleoprotein were exchanged among different avian hosts as frequently as the antigenic gene although they confer host specificity (Cauldwell et al. 2014).

3.2 Effect of host on the mutation (synonymous substitution) rate

Despite frequent host switching in all internal gene segments, we attempted to identify subsets of sequences, which are likely to represent a viral subpopulation that evolved while remaining in one particular host. First, for each segment, we identified subtrees forming discrete monophyletic clades that contain more than 30 sequences spanning over at least 10 years. In total, 42, 36, 44, and 42 clades were selected for PB2, PB1, PA, and NP segments, respectively. Of these, we found that 20, 20, 24, and 24 clades were closely associated with particular hosts: each of them contained more than 80 per cent of sequences from one particular host group, when the avian hosts were classified into four groups as mentioned above (Fig. 1). By parsimony, we assumed that the within-clade diversity of these sequences was mainly the result of evolutionary processes that occurred in the corresponding host environment. For PB2 segment, 10, 4, 2, and 4 clades in which most sequences were from viruses infecting wild *Anseriformes*, *Charadriiformes*, domestic ducks, and chickens, respectively, were identified. The corresponding numbers for PB1, PA, and NP segments were (13, 3, 1, 3), (14, 3, 3, 4), and (15, 2, 4, 3),

Table 1. Host switching rate estimated by BEAST analysis.

Data set	PB2	PB1	PA	NP	Internal genes combined
Mean	0.029	0.031	0.031	0.026	0.029
95% HPD	[0.026, 0.033]	[0.028, 0.035]	[0.027, 0.034]	[0.023, 0.029]	[0.024, 0.034]
Data set	H1	H2	H4	H5	
Mean	0.018	0.011	0.011	0.104	
95% HPD	[0.006, 0.104]	[0.006, 0.020]	[0.0053, 0.026]	[0.080, 0.13]	
Data set	H6	H7	H9	H10	
Mean	0.031	0.043	0.023	0.032	
95% HPD	[0.025, 0.069]	[0.025, 0.069]	[0.014, 0.049]	[0.018, 0.097]	

Note: HPD = highest posterior density.

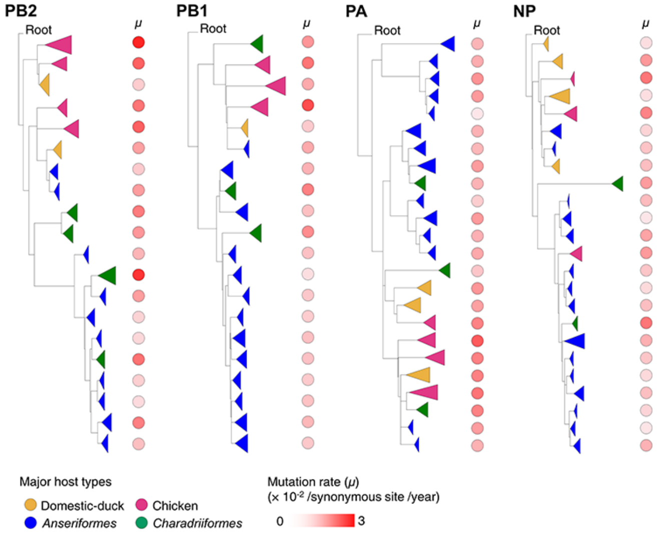

Figure 1. Phylogeny of PB2, PB1, PA, and NP segment sequences of avian influenza virus. Maximum likelihood trees were constructed from all available sequences and then only sequences forming host-specific monophyletic clades (shown in triangles) were retained. In each clade, more than 80 per cent sequences were sampled from one of four host groups: domestic duck, chicken, wild duck (*Anseriformes*), and shorebird (*Charadriiformes*). Mutation rates estimated by linear regression are shown next to the clades. The root of the trees is A/duck/Czech Republic/1/1956 (H4N6), which is the oldest isolate containing a full genome sequence in our avian influenza virus data.

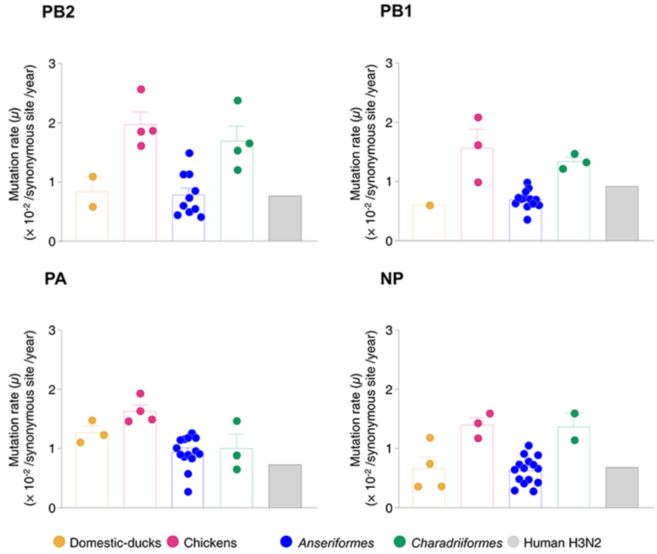

Figure 2. Comparison of the estimated mutation rates (synonymous substitution rates) of viruses infecting domestic duck, chicken, wild duck (*Anseriformes*), and shorebird (*Charadriiformes*). Mutation rate estimates from human H3N2 viruses (Croze and Kim 2021) are shown for comparison.

respectively. Detailed information for all clades chosen for the following analysis is presented in Supplementary Table S1.

Using a simple regression of synonymous sequence differences on sampling time differences, mutation rate ($\hat{\mu}$) was estimated for each of these 88 host specific clades. In the majority of clades, there was a clear linear increase of synonymous divergence with time; thus the pattern of the molecular clock expected for neutral evolution was observed (Supplementary Fig. S2). We found that, consistently over the segments, mutation rates were higher for viruses infecting wild-*Charadriiformes* and chickens than those infecting wild and domestic ducks (Fig. 2). When the results of all segments were combined, the difference in viral mutation rates by host group was highly significant (ANOVA, $P < 0.0001$; Table 2). The difference between segments was only marginally significant (highest for PB2 and lowest for NP). The difference in yearly mutation rate was particularly large between domestic duck versus chicken clades (Fig. 2), when these were two avian hosts between which viruses switch relatively frequently (Fig. 1). In addition, mutation rate was significantly higher in the *Charadriiformes* clades than in wild *Anseriformes* clades, while the former was generally nested within the clusters of the latter in phylogeny. It was also noted that similar mutation rates were estimated from wild and domestic duck clades.

On the phylogeny, clades with higher mutation rates do not appear to be located close to each other. To test whether divergence in clades' mutation rates increases with evolutionary genetic distance between them (corresponding to the length of the internal branches of phylogenies in Fig. 1), we used only wild duck clades and calculated the mean synonymous sequence difference between them minus within-clade diversity (Supplementary Fig. S3). A positive correlation was observed only in the PB1 segment. The overall pattern suggests that viral mutation rate is not a parameter that is determined by virus' evolutionary changes accumulating proportional to phylogenetic distance but one that changes readily, probably not due to viral evolution, upon switching host taxa.

3.3 Correlation with mutation rate estimates by a phylogenetic method

We attempted to validate the abovementioned results of host-specific mutation rates, estimated by simple regression, by a Bayesian phylogenetic method that jointly estimates tree topology and substitution rates on tree branches, which is implemented in

Table 2. Analysis of variance for mutation rates per year estimated from synonymous substitution rates in host-specific clades.

Two-way ANOVA table					
	SS (type III)	DF	MS	F (DFn, DFd)	P-value
Interaction	1.677	9	0.1863	F (9, 72) = 2.116	0.0389
Segment	0.813	3	0.2708	F (3, 72) = 3.077	0.0329
Host	10.06	3	3.3520	F (3, 72) = 38.08	< 0.0001
Residual	6.338	72	0.0880		
Ad hoc—by host types (Adjusted *P*-values using Holm–Šidák method)					
Group mean comparison				Mean difference	Adjusted *P*-value
Livestock/duck vs. livestock/chicken				−0.7981	<0.0001
Livestock/duck vs. *Charadriiformes*				−0.5075	0.0014
Livestock/duck vs. *Anseriformes*				0.0797	0.4899
Livestock/chicken vs. *Charadriiformes*				0.2906	0.0342
Livestock/chicken vs. *Anseriformes*				0.8778	<0.0001
Charadriiformes vs. *Anseriformes*				0.5872	<0.0001

the BEAST package. Substitution rates (the mean of posterior distribution) were estimated for all 88 clades above using the same sequences from which regression-based estimates were obtained. Since this method does not distinguish between synonymous versus nonsynonymous substitutions, the phylogeny was separately constructed using either the entire protein-coding sequence or the composite sequence made of third positions of codons. These two sets of sequences yielded near identical results (Supplementary Fig. S4), probably because substitutions on these internal genes are mostly synonymous (Bhatt, Holmes, and Pybus 2011). Thus, we may interpret the substitution rates estimated here to be approximately equal to the mutation rate.

Overall, the BEAST results confirmed the significance of differences in viral mutation rates by the host group ($P < 0.0001$; Supplementary Table S2), although the rate of domestic duck-infecting viruses was not significantly lower than those of chicken- and shorebirds-infecting viruses. The correlation between BEAST and regression estimates was highly significant ($R^2 = 0.4716$, $P < 0.0001$; Fig. 3). Note that substitution rates by BEAST are about one third of the rates by regression method in Fig. 3 because the number of nucleotide differences between two sequences is divided by the length of entire sequence in the former method while a similar number (synonymous differences) is divided by the (effective) number of synonymous sites within the sequence in the latter (Nei and Gojobori 1986). How much two estimation methods disagree for a given clade may be quantified by the corresponding residual in regression (distance from the observed BEAST estimate to the expected value on the regression line in Fig. 3). As expected, this residual was observed to increase with uncertainties regarding BEAST estimates (i.e. the width of posterior distribution), quantified by the length of the 95 per cent highest posterior density divided by the mean ($P = 0.011$; Supplementary Fig. S5). However, the correlation of residuals with other features of clades—the total number of sequences, the total time span of sampling, the dispersion (standard deviation) of sampling times, or the rate of nonsynonymous substitutions—was not significant (Supplementary Fig. S5).

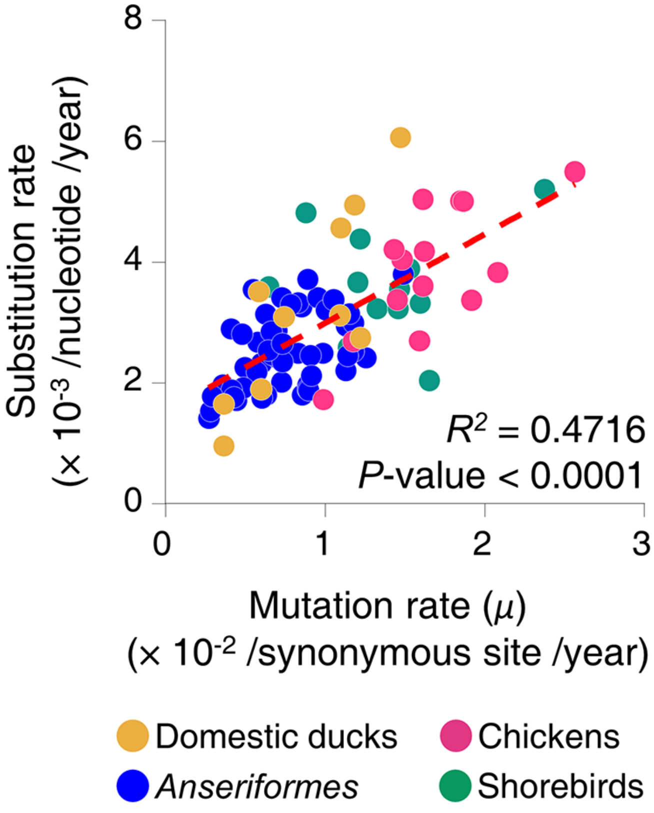

Figure 3. Correlation of mutation rate estimates (synonymous substitution rates) by the regression method (x-axis) and the BEAST estimates of substitution rate (y-axis), calculated for viruses infecting domestic ducks, chickens, wild ducks (*Anseriformes*), and shorebirds (*Charadriiformes*). R^2 and *P*-value were calculated using Pearson's correlation method.

3.4 Effect of pathogenicity on the yearly mutation rate

Next, we examined the effect of virus pathogenicity on the mutation rate. For most avian virus isolates from which we obtained genome sequences, clear information about whether they belong to HP strains was not given. For this reason, we indirectly inferred the pathogenicity of each isolate by the number of basic residues at the cleavage sites of the HA protein (see Section 2.6; Steinhauer 1999; Lopez-Martinez et al. 2013; Luczo et al. 2015). If this number was greater than 3, it was considered an HP virus. All others were defined as LP viruses. This method identifies the majority of HP viruses belonging to subtypes H5N1 or H5N2 infecting chickens. When the proportion of internal gene sequences from HP isolates (referred to as 'HP sequences' in the following) was counted for each of 88 minimum 10 year-span clades above, however, it was greater than 80 per cent in only one chicken clade (Supplementary Table S1). This is probably because the lineages of HP

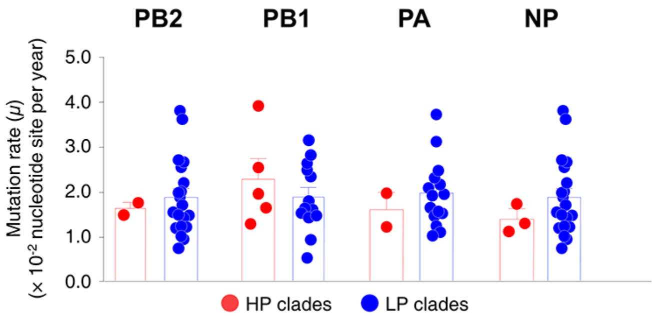

Figure 4. Comparison of mutation (synonymous substitution) rates on the PB2, PB1, PA, and PB segments estimated by linear regression between the highly pathogenic (HP) and low pathogenic (LP) clades of viral sequences infecting chickens.

viruses do not persist long enough. Since multiple clades made of HP sequences are needed for mutation rate comparison between HP and LP viruses, we searched the total phylogenies of four internal gene segments again to find new sets of distinct monophyletic clades containing at least 10 sequences that span at least 5 years (not 10 years). In addition, since we found that the host has a profound effect on mutation rate and that HP sequences are mostly observed in chicken host, we limited our search to compare mutation rates within the chicken host only. This led to the identification of 12 (2 PB2, 4 PB1, 3 PA, 3 NP) and 63 (19 PB2, 9 PB1, 18 PA, 17 NP) chicken clades that contain only HP and only LP sequences, respectively. Contrary to the expectation of a positive correlation between mutation rate and pathogenicity, estimated mutation rates by regression method from HP clades were slightly smaller than LP clades on average (1.808×10^{-2} vs. 1.855×10^{-2} per site per year) (Fig. 4). Therefore, we did not find the effect of pathogenicity on mutation rate (Supplementary Table S3).

Additional results confirmed that a higher mutation rate of viruses infecting chickens relative to that infecting domestic ducks was not due to a higher proportion of HP viruses in chickens than in ducks. First, the proportion of HP sequences within each of 14 minimum 10 year-span clades from chicken (data in Fig. 2) was not correlated with the estimated mutation rate (Supplementary Fig. S3). Second, the mutation rate difference was still highly significant between chicken and domestic duck hosts after removing clades that had more than 20 per cent sequences from HP isolates ($P < 0.0006$; Supplementary Fig. S4).

4. Discussion

Previous studies have measured the rate of evolution in influenza A virus over diverse vertebrate hosts (Webster et al. 1992; Chen and Holmes 2006; Nelson et al. 2006). However, most of them focused on the rapid molecular evolution at antigenic genes, quantified by substitution rates at both nonsynonymous and synonymous sites, where evolution at the former is mostly driven by selective pressure from host immunity. Here, we focused on the rate of synonymous substitutions, which are assumed to be neutral and thus provide the measure of mutation rate—how many errors in viral replication accumulate over a unit time. Our results indicate that this rate depends on host taxa; it is significantly higher when viruses infect chickens than domestic ducks and infect wild shorebirds than wild ducks. The rate estimates were similar between domestic and wild ducks, which belong to the same genus. It is possible that synonymous sites are under weak negative selection, for example, due to constraint for RNA secondary structure, in violation of our neutrality assumption. The rate of substitution at such sites can be different between two populations if their effective sizes are very different (stronger selection with larger effective size leading to lower rate) (McVean and Charlesworth 1999). Whether chickens and shorebirds have much smaller effective population sizes than ducks, compatible with their larger mutation rate estimates, needs to be investigated in the future. However, short external branches of phylogeny (coalescent tree, to be exact) observed in all avian hosts suggests that they all have small effective population sizes, as in the case of human H3N2 viral population (Croze and Kim 2021), leaving very small room for difference in the effective population size leading to substitution rate difference.

Since mutation (synonymous substitution) rate per unit time is determined by both the rate of mutation (copying errors) in one cycle of genome replication and the average number of replications per unit time, host dependence in mutation rate must be explained by inter-host heterogeneity in one or both of these parameters. For each parameter, heterogeneity can arise due to viral factors (evolutionary divergence of viral strains infecting different hosts) or host factors (host-specific cellular environment, immunity, and epidemiological dynamics). It is not clear if either viral or host factors create the host dependency of the first parameter, the fidelity of RNA replication in influenza virus, which depends on the performance of the polymerase complex of PA, PB1, and PB2 proteins. It is well known that several mutations of these genes are critically important in potentiating viral replication in mammalian host upon switching from avian hosts (Cauldwell et al. 2014; Long et al. 2019). It is also inferred that this polymerase complex interacts with host cellular factors, such as acidic leucine-rich nuclear phosphoprotein 32 family member A (ANP32A), that modulate the host-specific activity of replication (Moncorgé, Mura, and Barclay 2010; Long et al. 2016). However, it is unknown whether the activities of polymerase complex influenced by these viral and host factors include the fidelity of RNA replication.

To address the potential role of virus' functional variation in host-specific mutation rate, we examined if yearly mutation rate differences that we detected were associated with genetic divergence in PB2, PB1, and PA proteins. Amino acid variants, defined as the minor alleles when amino acid sequences of all clades were combined for a given gene, that are more than 50 per cent in frequency in at least two clades were found (Supplementary Table S4). Overall, we did not find a correlation between the number of amino acid variants observed in a clade and the mutation rate observed in the same clade. In PB2, sequences from chickens and shorebirds, hosts associated with higher mutation rates, have more variants than those from wild and domestic ducks. However, chicken clades yielding higher mutation rates are not distinguished from others by carrying a particular amino acid variant. Therefore, our own data do not provide evidence that differences in yearly mutation rate among avian hosts, either due to changes in fidelity or replication rate, result from genetic changes in virus-coded polymerases.

On the other hand, there is at least one difference among avian host cellular environments that is known to affect the reproductive cycle of influenza virus. Retinoic acid-inducible gene I protein (RIG-1), which is a cytoplasmic sensor of viral RNA and leads to the production of antiviral genes, is expressed in ducks and geese but not in chicken (Barber et al. 2010; Long et al. 2019). The replication of influenza virus was shown to be repressed in chicken cells after they were transfected to express duck RIG-I protein (Shao et al. 2014). With this kind of host-phyletic heterogeneity in cellular environments for viral reproduction, it seems reasonable that the replication rate of virus genome varies depending on the host. Therefore, we interpret differences in host-specific mutation rates as largely reflecting differences in viral replication rates; influenza virus probably replicates faster in chicken or shorebirds than in ducks. It is tempting to propose a hypothesis that chickens, which became a host to influenza virus only very recently, have not yet evolved to limit viral reproduction and thus allow faster replication. However, higher mutation rates in *Charadriiformes* than in wild *Anseriformes*, both of which are considered the natural reservoir of influenza A virus, may not be explained by hosts' disparate histories of adaptation to influenza virus.

The viral substitution rate also correlates with the host-dependent nature of transmission dynamics and the tropism of target cells (Hanada, Suzuki, and Gojobori 2004; Hicks and Duffy 2014). A large variation in substitution rate has been observed within a viral species that is under variable transmission modes, for example, in epidemic versus endemic cycles or with different host activities in tropical versus temperate regions (Salemi et al. 1999; Kurath et al. 2003; Streicker et al. 2012; Scholle et al. 2013). Heterogeneity in such transmission dynamics is translated to variation in the frequency of infection, thus replication, cycles per unit time. It will be therefore important to elucidate whether chickens, raised in farms in high density conditions, and shorebirds, with their population ecology potentially distinct from wild ducks, experience certain epidemiological dynamics that increase the replication rate of influenza virus.

We also investigated the correlation between pathogenicity of virus and mutation rate, as the pathogenicity is expected to increase with viral load, which in turn should increase with viral replication rate. However, a positive relationship between pathogenicity and mutation rate was not found. This negative result may be due to the short life span of pathogenic lineages that prevented us from obtaining substantially large HP clades from which mutation rates can be reliably measured. Alternatively, our assumption that replication rate should be positively correlated with viral load might not be true. A host might be able to limit the multiplication of virus up to a certain number while letting virus replicate at a constant rate. For example, the host defense might block virus from spreading beyond a small compartment in the host body or clear newly replicated viral ribonucleoproteins that are released into cytoplasm, both of which will lead to a lower viral load. In addition, significantly higher mutation rates in viruses infecting shorebirds than in those infecting wild ducks seem to suggest that the pathogenicity is not the primary determinant of mutation (or replication) rate, because both wild bird taxa are known to suffer little disease from influenza viruses, except H5N1.

Data availability

Data are available at http://avian-flu.org (Avian Influenza Database at Seoul National University College of Medicine).

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Funding

This work was supported in part by the Creative-Pioneering Researchers Program through Seoul National University (H.-R. Kim) and the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (2018M3A9H4055197 and 2021M3A9I2080497, H.-R. Kim).

Conflict of interest: No conflicts of interest declared.

Author contributions

G.K., H.M.S., and H.R.K. analyzed data, performed statistical tests, and participated in discussion. Y.K. conceived the study and wrote the manuscript.

References

Barber, M. R. W. et al. (2010) 'Association of RIG-I with Innate Immunity of Ducks to Influenza', *Proceedings of the National Academy of Sciences*, 107: 5913–8.

Bhatt, S., Holmes, E. C., and Pybus, O. G. (2011) 'The Genomic Rate of Molecular Adaptation of the Human Influenza A Virus', *Molecular Biology and Evolution*, 28: 2443–51.

Birky, C. W., and Walsh, J. B. (1988) 'Effects of Linkage on Rates of Molecular Evolution', *Proceedings of the National Academy of Sciences*, 85: 6414–8.

Boon, A. C. et al. (2011) 'H5N1 Influenza Virus Pathogenesis in Genetically Diverse Mice Is Mediated at the Level of Viral Load', *mBio*, 2: e00171–11.

Cauldwell, A. V. et al. (2014) 'Viral Determinants of Influenza A Virus Host Range', *Journal of General Virology*, 95: 1193–210.

Charlesworth, B., and Charlesworth, D. (2010) *Elements of Evolutionary Genetics*. Roberts and Company: Greenwood Village, Colorado.

Chen, R., and Holmes, E. C. (2006) 'Avian Influenza Virus Exhibits Rapid Evolutionary Dynamics', *Molecular Biology and Evolution*, 23: 2336–41.

Combe, M., and Sanjuan, R. (2014) 'Variation in RNA Virus Mutation Rates across Host Cells', *PLoS Pathogens*, 10: e1003855.

Croze, M., and Kim, Y. (2021) 'Inference of Population Genetic Parameters from an Irregular Time Series of Seasonal Influenza Virus Sequences', *Genetics*, 217: iyaa039.

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008) 'Rates of Evolutionary Change in Viruses: Patterns and Determinants', *Nature Reviews. Genetics*, 9: 267–76.

Elena, S. F., and Sanjuan, R. (2005) 'Adaptive Value of High Mutation Rates of RNA Viruses: Separating Causes from Consequences', *Journal of Virology*, 79: 11555–8.

Frank, S. A. (1996) 'Models of Parasite Virulence', *The Quarterly Review of Biology*, 71: 37–78.

Gu, X., Fu, Y.-X., and Li, W.-H. (1995) 'Maximum Likelihood Estimation of the Heterogeneity of Substitution Rate among Nucleotide Sites', *Molecular Biology and Evolution*, 12: 546–57.

Hanada, K., Suzuki, Y., and Gojobori, T. (2004) 'A Large Variation in the Rates of Synonymous Substitution for RNA Viruses and Its Relationship to A Diversity of Viral Infection and Transmission Modes', *Molecular Biology and Evolution*, 21: 1074–80.

Hatta, Y. et al. (2010) 'Viral Replication Rate Regulates Clinical Outcome and CD8 T Cell Responses during Highly Pathogenic H5N1 Influenza Virus Infection in Mice', *PLoS Pathogens*, 6: e1001139.

Hicks, A. L., and Duffy, S. (2014) 'Cell Tropism Predicts Long-term Nucleotide Substitution Rates of Mammalian RNA Viruses', *PLoS Pathogens*, 10: e1003838.

Hu, J. et al. (2013) 'The PA and HA Gene-mediated High Viral Load and Intense Innate Immune Response in the Brain Contribute to the High Pathogenicity of H5N1 Avian Influenza Virus in Mallard Ducks', *Journal of Virology*, 87: 11063–75.

Kaplan, B. S., and Webby, R. J. (2013) 'The Avian and Mammalian Host Range of Highly Pathogenic Avian H5N1 Influenza', *Virus Research*, 178: 3–11.

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019) 'MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization', *Briefings in Bioinformatics*, 20: 1160–6.

Kuraku, S. et al. (2013) 'aLeaves Facilitates On-demand Exploration of Metazoan Gene Family Trees on MAFFT Sequence Alignment Server with Enhanced Interactivity', *Nucleic Acids Research*, 41: W22–8.

Kurath, G. et al. (2003) 'Phylogeography of Infectious Haematopoietic Necrosis Virus in North America', *Journal of General Virology*, 84: 803–14.

Long, J. S. et al. (2016) 'Species Difference in ANP32A Underlies Influenza A Virus Polymerase Host Restriction', *Nature*, 529: 101–4.

——— et al. (2019) 'Host and Viral Determinants of Influenza A Virus Species Specificity', *Nature Reviews. Microbiology*, 17: 67–81.

Lopez-Martinez, I. et al. (2013) 'Highly Pathogenic Avian Influenza A(H7N3) Virus in Poultry Workers, Mexico, 2012', *Emerging Infectious Diseases*, 19: 1531–4.

Lu, L., Lycett, S. J., and Brown, A. J. L. (2014) 'Reassortment Patterns of Avian Influenza Virus Internal Segments among Different Subtypes', *BMC Evolutionary Biology*, 14: 1–15.

Luczo, J. M. et al. (2015) 'Molecular Pathogenesis of H5 Highly Pathogenic Avian Influenza: The Role of the Haemagglutinin Cleavage Site Motif', *Reviews in Medical Virology*, 25: 406–30.

May, R. M., and Anderson, R. M. (1990) 'Parasite—Host Coevolution', *Parasitology*, 100: S89–101.

McVean, G. A., and Charlesworth, B. (1999) 'A Population Genetic Model for the Evolution of Synonymous Codon Usage: Patterns and Predictions', *Genetical Research*, 74: 145–58.

Moncorgé, O., Mura, M., and Barclay, W. S. (2010) 'Evidence for Avian and Human Host Cell Factors that Affect the Activity of Influenza Virus Polymerase', *Journal of Virology*, 84: 9978–86.

Nei, M., and Gojobori, T. (1986) 'Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions', *Molecular Biology and Evolution*, 3: 418–26.

Nelson, M. I., and Holmes, E. C. (2007) 'The Evolution of Epidemic Influenza', *Nature Reviews. Genetics*, 8: 196–205.

Nelson, M. I. et al. (2006) 'Stochastic Processes are Key Determinants of Short-term Evolution in Influenza a Virus', *PLoS Pathogens*, 2: e125.

Neumann, G., and Kawaoka, Y. (2006) 'Host Range Restriction and Pathogenicity in the Context of Influenza Pandemic', *Emerging Infectious Diseases*, 12: 881.

Pita, J. S. et al. (2007) 'Environment Determines Fidelity for an RNA Virus Replicase', *Journal of Virology*, 81: 9072–7.

Salemi, M. et al. (1999) 'Different Population Dynamics of Human T Cell Lymphotropic Virus Type II in Intravenous Drug Users Compared with Endemically Infected Tribes', *Proceedings of the National Academy of Sciences*, 96: 13253–8.

Scholle, S. O. et al. (2013) 'Viral Substitution Rate Variation Can Arise from the Interplay between Within-Host and Epidemiological Dynamics', *The American Naturalist*, 182: 494–513.

Selman, M. et al. (2012) 'Adaptive Mutation in Influenza A Virus Non-structural Gene Is Linked to Host Switching and Induces A Novel Protein by Alternative Splicing', *Emerging Microbes & Infections*, 1: 1–10.

Shao, Q. et al. (2014) 'Function of Duck RIG-I in Induction of Antiviral Response against IBDV and Avian Influenza Virus on Chicken Cells', *Virus Research*, 191: 184–91.

Shu, L., Bean, W., and Webster, R. (1993) 'Analysis of the Evolution and Variation of the Human Influenza A Virus Nucleoprotein Gene from 1933 to 1990', *Journal of Virology*, 67: 2723–9.

Simmonds, P., Aiewsakun, P., and Katzourakis, A. (2019) 'Prisoners of War - Host Adaptation and Its Constraints on Virus Evolution', *Nature Reviews. Microbiology*, 17: 321–8.

Steinhauer, D. A. (1999) 'Role of Hemagglutinin Cleavage for the Pathogenicity of Influenza Virus', *Virology*, 258: 1–20.

Steinhauer, D. A., Domingo, E., and Holland, J. J. (1992) 'Lack of Evidence for Proofreading Mechanisms Associated with an RNA Virus Polymerase', *Gene*, 122: 281–8.

Streicker, D. G. et al. (2012) 'Rates of Viral Evolution are Linked to Host Geography in Bat Rabies', *PLoS Pathogens*, 8: e1002720.

Suarez, D. L. (2000) 'Evolution of Avian Influenza Viruses', *Veterinary Microbiology*, 74: 15–27.

Suchard, M. A. et al. (2018) 'Bayesian Phylogenetic and Phylodynamic Data Integration Using BEAST 1.10', *Virus Evolution*, 4: vey016.

Tavaré, S. (1986) 'Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences', *Lectures on Mathematics in the Life Sciences*, 17: 57–86.

Trifinopoulos, J. et al. (2016) 'W-IQ-TREE: A Fast Online Phylogenetic Tool for Maximum Likelihood Analysis', *Nucleic Acids Research*, 44: W232–5.

Wang, S., Sundaram, J. P., and Spiro, D. (2010) 'VIGOR, an Annotation Program for Small Viral Genomes', *BMC Bioinformatics*, 11: 1–10.

Webster, R. G. et al. (1992) 'Evolution and Ecology of Influenza A Viruses', *Microbiological Reviews*, 56: 152–79.

Woolhouse, M. E., Haydon, D. T., and Antia, R. (2005) 'Emerging Pathogens: The Epidemiology and Evolution of Species Jumps', *Trends in Ecology & Evolution*, 20: 238–44.

Yang, Z. (1994) 'Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods', *Journal of Molecular Evolution*, 39: 306–14.

——— (2007) 'PAML 4: Phylogenetic Analysis by Maximum Likelihood', *Molecular Biology and Evolution*, 24: 1586–91.