Gene Reports 40 (2025) 102233

Contents lists available at ScienceDirect

Gene Reports

journal homepage: www.elsevier.com/locate/genrep

Detecting signatures of episodic positive selection based on observed amino acids in hemagglutinin of H3N2 human influenza A virus

Yoshiyuki Suzuki

Graduate School of Science, Nagoya City University, 1 Yamanohata, Nagoya-shi, Aichi-ken 467-8501, Japan

ARTICLE INFO

Edited by Jormay Lim

Keywords:
Antigenicity
Epistasis
Hemagglutinin
Influenza A virus
Positive selection

ABSTRACT

In the parsimony method for detecting natural selection at amino acid sites of proteins, the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions that have accumulated over the evolutionary history of observed sequences were computed assuming that any amino acid was compatible at each site. Positive selection was inferred to have operated recurrently when the null hypothesis of no selection was rejected with $d_S < d_N$. Here an attempt to detect episodic positive selection within the framework of parsimony method was demonstrated. The d_S and d_N values were computed assuming that only the observed amino acids were compatible at each site. Positive selection was inferred to have operated episodically when the null hypothesis of the same fitness effects among observed amino acids was rejected with $d_S > d_N$. In the analysis of 18,444 sequences for hemagglutinin of H3N2 human influenza A virus, recurrent and episodic positive selections were inferred mainly at the sites related to antigenicity. Episodic positive selection was detected particularly at the sites under epistasis. Although it may be necessary to eliminate slightly deleterious amino acids from the population genetic data, the analysis based on observed amino acids may be useful for screening the sites with signatures of episodic positive selection.

1. Introduction

The rate of molecular evolution, such as the rates of nucleotide and amino acid substitutions, is equal to the mutation rate without natural selection (Kimura, 1983). However, the rate is accelerated and decelerated when positive and negative selections operate, respectively. Thus, natural selection operating at the amino acid sequence level can be detected by comparing the rates of synonymous (r_S) and nonsynonymous (r_N) substitutions under the assumption that synonymous mutations are selectively neutral or nearly neutral; positive and negative selections are inferred when $r_S < r_N$ and $r_S > r_N$, respectively (Hughes and Nei, 1988). Parsimony (Suzuki and Gojobori, 1999), likelihood (Suzuki, 2004; Kosakovsky Pond and Frost, 2005; Massingham and Goldman, 2005), and Bayesian (Yang et al., 2000; Murrell et al., 2012) methods have been developed for detecting natural selection at amino acid sites of proteins.

In the ordinary parsimony method, the number of synonymous substitutions per synonymous site (d_S) and that of nonsynonymous substitutions per nonsynonymous site (d_N) are compared, where $d_S < d_N$ and $d_S > d_N$ correspond to $r_S < r_N$ and $r_S > r_N$, respectively (Suzuki and Gojobori, 1999). The d_S and d_N values are computed by dividing the numbers of synonymous (c_S) and nonsynonymous (c_N) changes with those of synonymous (s_S) and nonsynonymous (s_N) sites, respectively. Here c_S and c_N are the counts of synonymous and nonsynonymous differences that have accumulated over the evolutionary history of observed sequences, respectively, whereas s_S and s_N represent the relative probabilities that a mutation randomly occurring at the codon site would be synonymous and nonsynonymous, respectively.

In cellular organisms, however, ~ 36% of random amino acid mutations appeared to be deleterious, among which > 80% may affect the thermodynamic stability of proteins (Tokuriki and Tawfik, 2009). In addition, ~ 50% of random amino acid mutations appeared to be

Abbreviations: A(H3N2)pdm68, IAV with subtype H3N2 circulating among humans after causing a pandemic in 1968; c_N, number of nonsynonymous changes; c_S, number of synonymous changes; d_N, number of nonsynonymous substitutions per nonsynonymous site; d_S, number of synonymous substitutions per synonymous site; GTR + Γ + I, general time reversible model with gamma-distributed rate heterogeneity among sites including invariable sites; H3HA, HA for A(H3N2)pdm68; HA, hemagglutinin; HA0, precursor of HA; IAV, influenza A virus; INSD, International Nucleotide Sequence Database; IVR, Influenza Virus Resource; ML, maximum likelihood; NA, neuraminidase; NJ, neighbor-joining; *P*, probability of Fisher's exact test; *q*, proportion of observed amino acid; r_N, rate of nonsynonymous substitution; r_S, rate of synonymous substitution; s_N, number of nonsynonymous sites; s_S, number of synonymous sites.

E-mail address: yossuzuk@nsc.nagoya-cu.ac.jp.

https://doi.org/10.1016/j.genrep.2025.102233
Received 11 February 2025; Received in revised form 3 April 2025; Accepted 17 April 2025
Available online 24 April 2025

deleterious in hemagglutinin (HA) of influenza A virus (IAV: *Alphainfluenzavirus influenzae*) (Nakajima et al., 2003, 2005). Although these mutations were included in the computation of s_N, they were not supposed to contribute to an increase in c_N, which may cause underestimation of d_N and reduction in the sensitivity for detecting positive selection. Thus, the ordinary method was modified to eliminate the fraction of s_N corresponding to deleterious mutations due to structural constraints, through predicting compatible and incompatible amino acids at each site of proteins based on the thermodynamic stability (Suzuki, 2013).

In both the ordinary and modified methods, d_S and d_N are computed as sums over the evolutionary history of observed sequences. Since negative selection always operates on amino acid sites as long as they are functional, positive selection may be detected as $d_S < d_N$ when it has operated recurrently, but may be masked as $d_S > d_N$ when it has operated episodically (Murrell et al., 2012). Notably, however, observed amino acids may have similar fitness effects at the amino acid sites evolving under the selective neutrality (Kimura, 1983). When the fitness effects are different, descendant amino acids are likely to have higher fitness effects than ancestral amino acids because evolution is not supposed to proceed toward decreasing the fitness. After the substitution increasing the fitness by positive selection, the reverse substitution, which is expected to decrease the fitness, may be suppressed by negative selection. Thus, positive selection operating on particular substitutions may be identified through detecting subsequent negative selection operating on reverse substitutions. This approach may be useful when the direction of evolution is known, and has been used to detect positive selection operating on the substitutions generating *N*-linked glycosylation sequons in HA of IAV (Suzuki, 2011).

However, even when the direction of evolution is unknown, positive selection may be identified at amino acid sites of proteins through detecting differences in the fitness effects among observed amino acids. The purpose of the present study was to demonstrate an attempt to detect signatures of episodic positive selection based on observed amino acids in HA of IAV.

2. Materials and methods

2.1. Sequence data

IAV is an etiological agent of influenza, with an enveloped virion of 80–120 nm in diameter containing a genome of eight-segmented, linear, negative-sense RNA (Krammer et al., 2018). The envelope glycoproteins, HA and neuraminidase (NA), have been classified into subtypes H1-H19 and N1-N11, respectively (Fereidouni et al., 2023). IAV with subtype H3N2 (A(H3N2)pdm68) has been circulating among humans after causing a pandemic in 1968 (Chang, 1969). HA is a homo-trimeric type 1 transmembrane glycoprotein, existing ~10 times more abundantly than NA (Mitnaul et al., 1996). The precursor of HA (HA0) for A (H3N2)pdm68 (H3HA), consisting of 566 amino acid sites, is cleaved into the signal peptide, HA1, and HA2, among which the latter two are linked by a bisulfide bond (Skehel and Wiley, 2000). Signal peptide directs co-translational transport of HA into the endoplasmic reticulum. HA1 is responsible for binding to the sialic acid receptor and is the major target of humoral immunity. HA2 anchors to the envelope and mediates its fusion with the endosomal membrane. Ectodomain of HA comprises distal globular head formed by a part of HA1 and proximal fibrous stem formed by the remaining part of HA1 and a part of HA2 (Lu et al., 2014). The 130-loop, 190-helix, 220-loop, and base in the globular head are involved in receptor binding. Antigenic regions A-E reside in the globular head as well as the fibrous stem (Wilson et al., 1981). Amino acid positions included in these regions are listed in supplementary Table S1.

A total of 41,765 nucleotide sequences for H3HA, excluding those for laboratory and vaccine strains as well as strains with mixed subtype, were retrieved from the Influenza Virus Resource (IVR) as of September 25, 2024 (Bao et al., 2008). After eliminating the sequences containing ambiguous nucleotides or minor gaps, 18,444 unique sequences for the entire coding region (1698 nucleotide sites) of H3HA were retained for analyses of natural selection. Strain names and accession numbers in the International Nucleotide Sequence Database (INSD) for these sequences are presented in supplementary Figs. S1 and S2.

Multiple alignment of nucleotide sequences was made using MAFFT (version 7.427) (Katoh et al., 2002). Average numbers of transitional and transversional substitutions computed with the two-parameter model were 0.0234 and 0.00955, respectively, suggesting that the transition/transversion rate ratio was ~5 (Kimura, 1980). After translating the nucleotide sequences into amino acid sequences, repertoires of amino acids with the proportion (q) of $q > 0$ (all observed amino acids), $q > 0.01$, $q > 0.05$, and $q > 0.1$ were identified at each site.

2.2. Data analysis

Phylogenetic analyses were conducted using MEGA (version 10.2.6) (Kumar et al., 2018). HA sequence for A/duck/Hokkaido/10/1985 (H3N8) (INSD accession number: AB276113) was adopted as the outgroup. General time reversible model with gamma-distributed rate heterogeneity among sites including invariable sites (GTR + Γ + I) was selected as the best fit model of nucleotide substitution by corrected Akaike and Bayesian information criteria. Phylogenetic tree was constructed by the maximum likelihood (ML) method with the best fit model, as well as the neighbor-joining (NJ) method with the p distance, which has been reported to produce more reliable topologies than the ML method (Nei and Kumar, 2000; Yoshida and Nei, 2016). Phylogenetic trees were visualized using FigTree (version 1.4.4) (Rambaut, 2018).

Since the number of sequences analyzed in the present study was relatively large, it was not feasible to infer ancestral sequences at interior nodes of phylogenetic tree by the ML method. Thus, ancestral sequences were inferred as follows. First, a pair of exterior nodes constituting a neighbor was identified in the phylogenetic tree. Second, nucleotide sequences for the neighbor as well as the outgroup were analyzed to infer the ancestral sequence at the interior node connected to them by the maximum parsimony method using PAML (version 4.9j) (Yang, 2007). Third, the phylogenetic tree was updated such that the neighbor was eliminated and the interior node was regarded as an exterior node. This process was repeated until ancestral sequences were inferred at all interior nodes of phylogenetic tree.

Natural selection was detected at each amino acid site by the ordinary parsimony method using ADAPTSITE (version 1.6) (Suzuki et al., 2001). The d_S and d_N values were computed assuming that any amino acid was compatible at each site. The relationship $d_S = d_N$ was expected under the null hypothesis of no selection, and positive and negative selections were inferred when $d_S < d_N$ and $d_S > d_N$, respectively. Signature of positive selection was also detected as difference in the fitness effects among observed amino acids at each site. The d_S and d_N values were computed assuming that only the repertoire of observed amino acids with $q > 0$ were compatible at each site. Since some of unobserved amino acids may also be compatible, the relationship $d_S \leq d_N$ was expected under the null hypothesis of the same fitness effects among observed amino acids, and difference in the fitness effects was inferred when $d_S > d_N$. In the population genetic data, however, observed amino acids may contain slightly deleterious amino acids, which may create difference in the fitness effects without positive selection. Reportedly, the proportion of slightly deleterious amino acids may attain ~10% (Fay et al., 2001). Thus, to eliminate slightly deleterious amino acids, the above analysis was also conducted with $q > 0.01$, $q > 0.05$, and $q > 0.1$. In addition, since advantageous and slightly deleterious substitutions may be accumulated on the interior and exterior branches of phylogenetic tree, respectively (Pybus et al., 2007), the entire analysis was repeated using only the interior branches. In all cases, correction for multiple testing was conducted with the family-wise significance level of 0.05 (Suzuki, 2011).

Reportedly, positive selection may operate mainly at the amino acid sites within antigenic regions for H3HA (Murrell et al., 2012; Koel et al., 2013). Thus, results were evaluated with the efficiency for detecting positive and negative selections at the sites within and outside antigenic regions, respectively, using the probability (P) of Fisher's exact test.

3. Results

3.1. Positively selected amino acid sites identified by the ordinary method

When the ordinary method was applied to detecting natural selection operating at the amino acid sites of H3HA using all branches of the ML tree (supplementary Fig. S1), positive selection was detected mainly at the sites within antigenic regions ($P = 4.89 \times 10^{-17}$) (Tables 1 and 2). In contrast, negative selection was detected mainly at the sites outside antigenic regions ($P = 1.32 \times 10^{-8}$) (Table 1; supplementary Table S2). These results were consistent with those obtained in the previous studies (Murrell et al., 2012; Koel et al., 2013).

In the analysis using only the interior branches for eliminating slightly deleterious substitutions that may have accumulated on the exterior branches, the number of positively selected sites was not largely affected ($P = 7.45 \times 10^{-20}$) (Tables 1 and 2), which may reflect the tendency for advantageous substitutions to be located on the interior branches (Pybus et al., 2007). However, the number of negatively selected sites decreased significantly ($P = 0.0904$) (Table 1; supplementary Table S2), which was likely to be caused by elimination of synonymous substitutions on the exterior branches. Analyses using the NJ tree (supplementary Fig. S2) produced similar results (supplementary Tables S3 and S4).

3.2. Positively selected amino acid sites identified based on observed amino acids

Signature of positive selection was also identified as difference in the fitness effects among observed amino acids. In the analysis using all branches of the ML tree (supplementary Fig. S1), difference in the fitness effects among all observed amino acids (repertoire of observed amino acids with $q > 0$) was detected at many sites ($P = 3.76 \times 10^{-3}$) (Table 1; supplementary Table S5), which may reflect abundance of slightly deleterious amino acids (Nakajima et al., 2003, 2005). When slightly deleterious amino acids were eliminated by using repertoires of observed amino acids with $q > 0.01$, $q > 0.05$, and $q > 0.1$, the number of sites with difference in the fitness effects tended to be decreased along with the increase in the threshold value of q (Table 1; supplementary Table S5). However, similar results were obtained for $q > 0.05$ and $q > 0.1$, suggesting that slightly deleterious amino acids were mostly eliminated with $q > 0.05$. When $q > 0.05$, difference in the fitness effects was identified at 11 sites, which were mainly located within antigenic regions ($P = 5.98 \times 10^{-4}$) (Table 2). Out of these 11 sites, eight were inferred to be negatively selected by the ordinary method described above (supplementary Table S2).

In the analysis using only the interior branches, the number of sites with difference in the fitness effects was decreased when $q > 0$ and $q > 0.01$ (Table 1; supplementary Table S5), which was likely to be caused by existence of slightly deleterious amino acids with $q > 0$ and $q > 0.01$ and elimination of synonymous substitutions on the exterior branches. However, the number of sites with difference in the fitness effects was not largely affected when $q > 0.05$ and $q > 0.1$ (Table 1; supplementary Table S5), suggesting that slightly deleterious amino acids were mostly eliminated with $q > 0.05$ and advantageous substitutions tended to be located on the interior branches. Analyses using the NJ tree (supplementary Fig. S2) produced similar results (supplementary Tables S3 and S6).

Table 2
Amino acid positions where positive selection was identified by the ordinary method and based on the observed amino acids with $q > 0.05$ in H3HA using the ML tree.

Method	All branches[c]	Interior branches[c]
Ordinary[a]	**50, 53, 54, 57, 62, 63, 75, 81, 82, 83, 94, 121, 122, 124, 126, 135, 137, 140, 142, 143, 144, 146, 156, 157, 158, 159, 160, 172, 173, 188, 192, 193, 196, 198, 213, 227, 242, 260, 262, 276, 278, 299, 312**	**50, 53, 54, 57, 62, 63, 75, 81, 82, 83, 92, 94, 121, 122, 124, 126, 135, 137, 140, 142, 143, 144, 146, 155, 156, 157, 158, 159, 160, 172, 173, 188, 190, 192, 193, 196, 198, 213, 227, 242, 260, 262, 276, 278, 299, 312**
	[−8], [−3], 2, 3, 25, 31, 33, 202, 223, 225, 236, 406, 452, 453, 489, 522, 529, 530	[−8], 2, 3, 25, 31, 202, 223, 225, 236, 326, 406, 452, 453, 489, 529, 530
Observed[b]	**45, 128, 145, 164, 186, 190, 261, 276**	**45, 128, 145, 164, 186, 261**
	[−14], 195, 484	[−14], 484

[a] Ordinary parsimony method.
[b] Analysis based on observed amino acids.
[c] Positions within and outside antigenic regions are indicated in bold face and plain text, respectively, and those in the signal peptide are indicated in the brackets.

4. Discussion

4.1. Detecting signatures of episodic positive selection based on observed amino acids

In the ordinary parsimony method for detecting natural selection at amino acid sites of proteins, effects of positive and negative selections are averaged over the evolutionary history of observed sequences (Suzuki and Gojobori, 1999). Thus, this method may be suitable for identifying natural selection that has operated recurrently, but not for identifying fluctuating natural selection, including episodic positive selection. Positive selection may operate recurrently when the fitness

Table 1
Numbers of positively and negatively selected amino acid sites identified in H3HA using the ML tree.

				All branches		Interior branches	
Method	Compatible amino acid	Relationship	Selection	Antigenic regions [131][c]	Other regions [435][d]	Antigenic regions [131][c]	Other regions [435][d]
Ordinary[a]	Any	$d_S < d_N$	Positive	43	18	46	16
	Any	$d_S > d_N$	Negative	63	327	57	227
Observed[b]	$q > 0$	$d_S > d_N$	Positive	53	240	54	179
	$q > 0.01$	$d_S > d_N$	Positive	11	7	9	4
	$q > 0.05$	$d_S > d_N$	Positive	8	3	6	2
	$q > 0.1$	$d_S > d_N$	Positive	7	3	6	2

[a] Ordinary parsimony method.
[b] Analysis based on observed amino acids.
[c] Total number of amino acid sites within antigenic regions is indicated in the brackets.
[d] Total number of amino acid sites outside antigenic regions is indicated in the brackets.

landscape of amino acids changes continuously (Han et al., 2023). However, when the fitness landscape changes occasionally, adaptive evolution may occur through episodic positive selection. After the adaptation, evolution may proceed neutrally among amino acids with similar fitness effects (Kimura, 1983). Here fitness effects of the amino acids before and after the adaption should be different; descendant amino acids should have higher fitness effects than ancestral amino acids. Thus, signature of episodic positive selection was detected as difference in the fitness effects among observed amino acids in the present study.

In the population genetic data, observed amino acids may contain slightly deleterious amino acids, which may create difference in the fitness effects among observed amino acids without positive selection. When difference in the fitness effects was examined among observed amino acids with $q > 0$, $q > 0.01$, $q > 0.05$, and $q > 0.1$, the number of sites with the difference tended to be decreased along with the increase in the threshold value of q. However, similar results were obtained for $q > 0.05$ and $q > 0.1$ using either all branches or interior branches of phylogenetic tree. Given that the proportion of slightly deleterious amino acids may attain $\sim 10\%$ (Fay et al., 2001), the above results suggested that these amino acids may be mostly eliminated with $q > 0.05$. Notably, however, slightly deleterious amino acids may not be contained in the species divergence data.

4.2. *Positively selected amino acid sites in HA of IAV*

Recurrent and episodic positive selections were identified mainly at the sites within antigenic regions, which were involved in immune escape (Murrell et al., 2012; Koel et al., 2013). Although positive selection was inferred at a few sites in the signal peptide, these results were likely to be false positives. This was based on the facts that functional constraint on the signal peptide is generally weak (Li et al., 2009) and synonymous substitution is suppressed at the ends of coding regions for IAV due to existence of bundling signals (Goto et al., 2013; Canale et al., 2018). Recurrent positive selection was also inferred at positions 223 and 225, which were outside antigenic regions. However, these positions were identified as determinants of antigenicity in H3HA by the machine learning model (Shah et al., 2024). Amino acid sites with recurrent and episodic positive selections within antigenic regions were often involved in *N*-linked glycosylation sequons (supplementary Table S7). Episodic positive selection may have operated on the gain of sequons for shielding antigenic regions from immune response (Suzuki, 2011; Kobayashi and Suzuki, 2012). Recurrent positive selection may also have operated on these sites before the gain of sequons for immune escape.

Amino acid sites with episodic positive selection largely (8/11) overlapped with those involved in an increase in the fitness identified from analyses of branching patterns in phylogenetic trees (Lefrancq et al., 2025). In addition, many of these sites (6/11) were also involved in epistasis (Mani et al., 2008; Lyons and Lauring, 2018). Positions 186 and 190 in antigenic region B did not appear to evolve independently because substitution at each of these positions was likely to reduce receptor binding avidity (Wu et al., 2018; Lei et al., 2024). However, these positions have co-evolved by changing receptor binding mode without reducing the avidity. At position 195 in the base of receptor binding region, tyrosine, phenylalanine, histidine, and asparagine were observed with $q > 0$, but only the former two were observed with $q > 0.05$. Although substitution from tyrosine to phenylalanine appeared to inhibit receptor binding (Skehel and Wiley, 2000), the predominant amino acid has been changed from tyrosine to phenylalanine during 2020–2021 (supplementary Fig. S3). This substitution has taken place apparently because it caused alteration in receptor binding mode together with substitutions at positions 159 and 160 (Liang et al., 2024). Positions 128, 276, and 484 have also been predicted to be under epistasis with other positions, although epistasis may not always induce adaptive evolution (Kryazhimskiy et al., 2011).

Supplementary data to this article can be found online at https://doi.org/10.1016/j.genrep.2025.102233.

CRediT authorship contribution statement

Yoshiyuki Suzuki: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The author declares no conflict of interest.

Acknowledgements

The author thanks anonymous reviewers for valuable comments. This work was supported by JSPS KAKENHI Grant Number JP19K12221 and AMED Grant Number JP23fk0108667 to Y.S.

Data availability

Data will be made available on request.

References

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., Lipman, D., 2008. The influenza virus resource at the national center for biotechnology information. J. Virol. 82, 596–601.
Canale, A.S., Venev, S.V., Whitfield, T.W., Caffrey, D.R., Marasco, W.A., Schiffer, C.A., Kowalik, T.F., Jensen, J.D., Finberg, R.W., Zeldovich, K.B., Wang, J.P., Bolon, D.N. A., 2018. Synonymous mutations at the beginning of the influenza A virus hemagglutinin gene impact experimental fitness. J. Mol. Biol. 430, 1098–1115.
Chang, W.K., 1969. National influenza experience in Hong Kong, 1968. Bull. World Health Organ. 41, 349–351.
Fay, J.C., Wyckoff, G.J., Wu, C.-I., 2001. Positive and negative selection on the human genome. Genetics 158, 1227–1234.
Fereidouni, S., Starick, E., Karamendin, K., Genova, C.D., Scott, S.D., Khan, Y., Harder, T., Kydyrmanov, A., 2023. Genetic characterization of a new candidate hemagglutinin subtype of influenza A viruses. Emerg. Microbes Infect. 12, 2225645.
Goto, H., Muramoto, Y., Noda, T., Kawaoka, Y., 2013. The genome-packaging signal of the influenza A virus genome comprises a genome incorporation signal and a genome bundling signal. J. Virol. 87, 11316–11322.
Han, A.X., de Jong, S.P.J., Russell, C.A., 2023. Co-evolution of immunity and seasonal influenza viruses. Nat. Rev. Microbiol. 21, 805–817.
Hughes, A.L., Nei, M., 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335, 167–170.
Katoh, K., Misawa, K., Kuma, K.-I., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066.
Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.
Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, New York, Melbourne.
Kobayashi, Y., Suzuki, Y., 2012. Evidence for N-glycan shielding of antigenic sites during evolution of human influenza A virus hemagglutinin. J. Virol. 86, 3446–3451.
Koel, B.F., Burke, D.F., Bestebroer, T.M., van der Vliet, S., Zondag, G.C., Vervaet, G., Skepner, E., Lewis, N.S., Spronken, M.I., Russell, C.A., Eropkin, M.Y., Hurt, A.C., Barr, I.G., de Jong, J.C., Rimmelzwaan, G.F., Osterhaus, A.D., Fouchier, R.A., Smith, D.J., 2013. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science 342, 976–979.
Kosakovsky Pond, S.L., Frost, S.D., 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22, 1208–1222.
Krammer, F., Smith, G.J.D., Fouchier, R.A.M., Peiris, M., Kedzierska, K., Doherty, P.C., Palese, P., Shaw, M.L., Treanor, J., Webster, R.G., García-Sastre, A., 2018. Influenza. Nat. Rev. Dis. Primers 4, 3.
Kryazhimskiy, S., Dushoff, J., Bazykin, G.A., Plotkin, J.B., 2011. Prevalence of epistasis in the evolution of influenza A surface proteins. PLoS Genet. 7, e1001301.
Kumar, S., Stecher, G., Li, M., Tamura, K., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35, 1547–1549.
Lefrancq, N., Duret, L., Bouchez, V., Brisse, S., Parkhill, J., Salje, H., 2025. Learning the fitness dynamics of pathogens from phylogenies. Nature 637, 683–690.
Lei, R., Liang, W., Ouyang, W.O., Hernandez Garcia, A., Kikuchi, C., Wang, S., McBride, R., Tan, T.J.C., Sun, Y., Chen, C., Graham, C.S., Rodriguez, L.A., Shen, I.R., Choi, D., Bruzzone, R., Paulson, J.C., Nair, S.K., Mok, C.K.P., Wu, N.C., 2024.

Epistasis mediates the evolution of the receptor binding mode in recent human H3N2 hemagglutinin. Nat. Commun. 15, 5175.
Li, Y.-D., Xie, Z.-Y., Du, Y.-L., Zhou, Z., Mao, X.-M., Lv, L.-X., Li, Y.-Q., 2009. The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. Gene 436, 8–11.
Liang, R., Peccati, F., Ponse, N.L.D., Uslu, E., Boons, G.-J., Unione, L., de Vries, R.P., 2024. Epistasis in the receptor binding domain of contemporary H3N2 viruses that reverted to bind sialylated diLacNAc repeats. bioRxiv 625384.
Lu, Y., Welsh, J.P., Swartz, J.R., 2014. Production and stabilization of the trimeric influenza hemagglutinin stem domain for potentially broadly protective influenza vaccines. Proc. Natl. Acad. Sci. U. S. A. 111, 125–130.
Lyons, D.M., Lauring, A.S., 2018. Mutation and epistasis in influenza virus evolution. Viruses 10, 407.
Mani, R., St. Onge, R.P., Hartman, J.L.I.V., Giaever, G., Roth, F.P., 2008. Defining genetic interaction. Proc. Natl. Acad. Sci. U. S. A. 105, 3461–3466.
Massingham, T., Goldman, N., 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics 169, 1753–1762.
Mitnaul, L.J., Castrucci, M.R., Murti, K.G., Kawaoka, Y., 1996. The cytoplasmic tail of influenza A virus neuraminidase (NA) affects NA incorporation into virions, virion morphology, and virulence in mice but is not essential for virus replication. J. Virol. 70, 873–879.
Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., Kosakovsky Pond, S.L., 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8, e1002764.
Nakajima, K., Nobusawa, E., Tonegawa, K., Nakajima, S., 2003. Restriction of amino acid change in influenza A virus H3HA: comparison of amino acid changes observed in nature and in vitro. J. Virol. 77, 10088–10098.
Nakajima, K., Nobusawa, E., Nagy, A., Nakajima, S., 2005. Accumulation of amino acid substitutions promotes irreversible structural changes in the hemagglutinin of human influenza AH3 virus during evolution. J. Virol. 79, 6472–6477.
Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Oxford, New York.
Pybus, O.G., Rambaut, A., Belshaw, R., Freckleton, R.P., Drummond, A.J., Holmes, E.C., 2007. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. Mol. Biol. Evol. 24, 845–852.
Rambaut, A., 2018. FigTree. http://tree.bio.ed.ac.uk/software/figtree.
Shah, S.A.W., Palomar, D.P., Barr, I., Poon, L.L.M., Quadeer, A.A., McKay, M.R., 2024. Seasonal antigenic prediction of influenza A H3N2 using machine learning. Nat. Commun. 15, 3833.
Skehel, J.J., Wiley, D.C., 2000. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu. Rev. Biochem. 69, 531–569.
Suzuki, Y., 2004. New methods for detecting positive selection at single amino acid sites. J. Mol. Evol. 59, 11–19.
Suzuki, Y., 2011. Positive selection for gains of N-linked glycosylation sites in hemagglutinin during evolution of H3N2 human influenza A virus. Genes Genet. Syst. 86, 287–294.
Suzuki, Y., 2013. Detection of positive selection eliminating effects of structural constraints in hemagglutinin of H3N2 human influenza A virus. Infect. Genet. Evol. 16, 93–98.
Suzuki, Y., Gojobori, T., 1999. A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 16, 1315–1328.
Suzuki, Y., Gojobori, T., Nei, M., 2001. ADAPTSITE: detecting natural selection at single amino acid sites. Bioinformatics 17, 660–661.
Tokuriki, N., Tawfik, D.S., 2009. Stability effects of mutations and protein evolvability. Curr. Opin. Struct. Biol. 19, 596–604.
Wilson, I.A., Skehel, J.J., Wiley, D.C., 1981. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. Nature 289, 366–373.
Wu, N.C., Thompson, A.J., Xie, J., Lin, C.-W., Nycholat, C.M., Zhu, X., Lerner, R.A., Paulson, J.C., Wilson, I.A., 2018. A complex epistatic network limits the mutational reversibility in the influenza hemagglutinin receptor-binding site. Nat. Commun. 9, 1264.
Yang, Z., 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.
Yang, Z., Nielsen, R., Goldman, N., Pedersen .K., A.-M., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.
Yoshida, R., Nei, M., 2016. Efficiencies of the NJp, maximum likelihood, and Bayesian methods of phylogenetic construction for compositional and noncompositional genes. Mol. Biol. Evol. 33, 1618–1624.