Improving the prediction of influenza vaccine effectiveness by refined genetic distance measure

Authors:

Lirong Cao[1,2], Jingzhi Lou[3], Qi Li[1,2], Chris Ka Pun Mok[1,4], Zigui Chen[5], Renee Wan Yi Chan[6,7], Peter Pak Hang Cheung[8], Marc Ka Chun Chong[1,2], Eng Kiong Yeoh[1,9], Paul Kay Sheung Chan[5,10], Benny Chung-Ying Zee[1,2], and Maggie Haitian Wang[1,2*]

Affiliations:

[1] JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong SAR, China

[2] CUHK Shenzhen Research Institute, Shenzhen, China

[3] Beth Bioinformatics Co. Ltd., Hong Kong SAR, China

[4] Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China.

[5] Department of Microbiology, The Chinese University of Hong Kong, Hong Kong SAR, China

[6] Department of Paediatrics, The Chinese University of Hong Kong, Hong Kong SAR, China

[7] Hong Kong Hub of Paediatric Excellence, The Chinese University of Hong Kong, Hong Kong SAR, China

[8] Department of Chemical Pathology, CUHK; Hong Kong SAR, China.

[9] Centre for Health Systems and Policy Research, CUHK; Hong Kong SAR, China.

[10] Stanley Ho Centre for Emerging Infectious Diseases, CUHK; Hong Kong SAR, China.

Correspondence: Maggie Haitian Wang, email: maggiew@cuhk.edu.hk, JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong SAR, China.

Abstract

Our previous research demonstrated that genetic distance (GD) on effective mutation (EM) sites can be used to evaluate vaccine effectiveness (VE) *in silico* in real time. This study further investigates the relationship between VE and GD on antigenic sites (AS) and identifies key amino acid sites related to vaccine protection against influenza A/H1N1pdm09 and A/H3N2 between 2009 and 2019 flu seasons. We found that not any AS on hemagglutinin (HA) and neuraminidase (NA) may cause a decrease in VE, rather, GD on the intersection set of EM and AS is highly predictive of influenza VE. The integrated GD of HA and NA can explain up to 87.8% of VE variations for H3N2. Significant improvement is also found for VE prediction for pH1N1. Accurate prediction of influenza VE before vaccine deployment may facilitate reverse vaccinology to optimize vaccine antigen design and facilitate government preparedness of epidemics.

Keywords: Influenza; vaccine effectiveness; prediction; genetic distance; genetic mismatch; sequence analysis

Introduction

Influenza causes numerous morbidity and mortality annually worldwide, most of which occur among susceptible populations, including elderly adults, immunocompromised persons, and individuals with chronic cardiopulmonary disease[1]. Many countries have implemented influenza vaccination programmes to reduce disease burden due to influenza. However, the performance of influenza vaccines varies by vaccine match to circulating viruses each year. While the protective effect of vaccine can be assessed by vaccine efficacy through randomized control trials (RCTs) or vaccine effectiveness (VE) in observational studies[2-4], rapid evaluation of VE before mass vaccination can provide timely information for policy-makers to optimize public health resources allocation.

Previously, we developed a novel computational framework to predict VE for the influenza[5,6] and COVID-19[7] vaccine by virus sequencing data. We found that genetic distance (GD) between vaccine strains and circulating viruses is significantly correlated with vaccine effectiveness (VE). The key to model the relationship between VE and GD lies in the appropriate quantification of GD. Although it has been well recognized that mutations in the epitopes play a major role in virus antigenic drift and vaccine escape[8], GD computed on the epitope regions are not predictive to VE[5]. We found that GD measured on the set of effective mutations (EMs) shows good prediction performance of VE. The EMs were identified based on the prevalence of the substitution exceeding the dominance threshold in population[9]. Evident linear relationships were identified between GD at the EM sites and VE for both influenza A subtypes and influenza B, and were validated in independent data. Nevertheless, current prediction model is mainly constructed from a statistical point of view without utilizing functional information of the virus genome. Previous studies reported that mutations near the receptor binding domain is responsible for major antigenic drift of the influenza viruses[10,11]. In view of the biological significance for AS, complete survey of GD on antigenic epitopes is important for better understanding GD's relationship with VE and identifying optimal targets for GD quantification.

In this study, we investigate the incorporation of antigenic sites information in the VE-GD model for influenza. We explore the genetic predictors constructed for both the HA and NA proteins and their respective effect on VE. The outcome of this study is quantified measures of genetic escape influencing vaccine protection and improved VE prediction model, ready-to-use for VE assessment giving vaccine strain component against an arbitrary composition of circulating virus population.

Methods

Viral sequences and vaccine effectiveness data

Sequencing data of influenza A H1N1pdm09 and H3N2 were downloaded from the NCBI[12], and GISAID EpiFlu databases[13]. The HA and NA protein sequences in the United States were collected for model building and genetic data in the United Kingdom, Germany and Hong Kong were collected for validation. The vaccine effectiveness data were extracted from published epidemiological studies and described in the supplementary materials of Cao et al 2021[5,6]. The influenza vaccine strains recommended by WHO for use in the northern hemisphere were summarized in supplement materials **Table S1**.

Genetic distance on antigenic epitopes

Detailed information on antigenic epitopes of HA and NA for H1N1 and H3N2 viruses were collected from literature[14-20]. Amino acid positions on AS are listed in **Table S2** and EM sites in **Table S3**. The AS and EM locations on stereo view of HA and NA protein structures is plotted in **Figures S1-4**, obtained from the RSCB Protein Data Bank (PDB) (https://www.rcsb.org). We compared predictive power of using GD calculated on AS sites with GD on EM-AS sites. A sample protein sequence *i* isolated in season *t* is denoted as,

$$X_i^t = \{x_{ij}^t\}, j = 1, \ldots, J, \text{ and } i = 1, \ldots, n_t,$$

where *j* is the *j*-th amino acid position, and n_t is sequence sample size in *t*-th flu season. Let $V^t = \{v_j^t\}$ denote the WHO recommended vaccine strain for the *t*-th flu season, and $W(t)$ denotes the predictor codon sites identified for season *t*. The genetic mismatch evaluated on the predictor codon sites between the vaccine strain and a given virus strain is calculated by the hamming distance, denoted by function d_i^t, between X_i^t and V^t as follows,

$$d_i^t = \sum_{j \in W(t)} I(v_j^t \neq x_j^t).$$

In the *t*-th flu season, the GD indicates the total mismatch d^t compared to a vaccine strain on $W(t)$, measured by the average distance of n_t sequence to V^t, as follows,

$$d^t = \sum_{i=1}^{n_t} d_i^t / n_t = \sum_{i=1}^{n_t} \sum_{j \in W(t)} I\left(v_j^t \neq x_j^t\right) / n_t. \qquad \text{Equation 1}$$

This metric measures the GD between a given vaccine strain to circulating viruses for the same influenza protein. The key to quantifying GD lies in the determination of $W(t)$, and in this study AS and EM sites are tested. Since the influenza vaccines incorporate multiple antigens, we need to take into account the total mismatch of these antigens, which is introduced in the following section.

Integrative analysis and mediation analysis for VE evaluation

Integrative analysis described in prior research[5] was used to aggregate genetic mismatch on HA and NA proteins. Let Y^s denote VE of a given subtype s, β_{gs} denote the effect size of genetic distance measured on gene g for influenza subtype s. A generalized linear model was used to model their relationship as follows,

$$E[Y^s \,|\, d^{s,g}] = \beta_0 + \beta_{gs} d(g,s), \qquad \text{Equation 2}$$

where $d(g, s)$ is a vector of genetic mismatch over the sampling periods, targeting gene g of an influenza type s. $\hat{\beta}_{gs}$ indicates sample estimate. Based on the method of the Polygenic Risk Score (PRS) to integrate risks of multiple genetic variants[21,22], the overall genetic distance can be calculated by,

$$d(s) = \sum_g \hat{\beta}_{gs} d(g,s) \,/ \sum_g \hat{\beta}_{gs}. \qquad \text{Equation 3}$$

We thus can use this equation to aggregate the mismatch of multiple subtypes to investigate their association with the subtype VE. This approach can be used to reasonably aggregate genetic mismatch for each protein and its effect size could be adjusted automatically to avoid underestimation or overestimation of the specific phenotype. Moreover, such a method can reduce the collinearity problems among the genetic markers compared to simultaneous estimation of the effect size.

Two regression models were required to perform mediation analysis: a model for the mediator (GD on NA) determined by the independent variable (GD on HA), and a model for the outcome (influenza subtype VE) determined jointly by the exposure and the mediator. The indirect effect is the effect of the HA mismatch on the VE through NA mismatch, and the direct effect is the effect of the HA mismatch on the outcome independent of the mediator. All analyses was carried out in **R** statistical software[23] (version 4.0.3) and the R package mediation was used[24]. The statistical significance is declared when p-value is less than 0.05.

Results

Genetic distance on antigenic sites of HA

Firstly, we calculated genetic mismatch on five antigenic epitopes of HA for influenza H3N2 (Epitopes A-E) and H1N1pdm09 (Sites Sa, Sb, Ca1, Ca2 and Cb). For H3N2, the results showed that genetic distance on epitopes A and B were significantly associated with VE and approximately 50% of the variations in VE could be explained by these epitope mismatch, while no relationships were observed for genetic distance on epitopes D-E and all AS (**Figure 1**). The explained variations increased to 61.5% by aggregating epitopes A and B mismatch (**Figure 2b**). For H1N1pdm09, whereas genetic mismatch on any single AS seems not associated with VE, the integrated mismatch

on sites Sa and Sb showed negative linear relationship with VE (**Figure 3**). Unlike H3N2, the H1N1 VE can be better explained by the integrated mismatch of all AS compared to using combined Sa and Sb mismatch and the goodness of fit in terms of R^2 can reach 41.0% (**Figure 4b**).

Comparison of antigenic sites and effective mutation sites

Next, we compared the performance of using AS and EM sites to estimate VE and optimized the quantification of GD. For HA protein, the highest predictive power was observed for the GD on the sites that were the intersection set of EM and AS (EM-AS). The results were consistent for H1N1 and H3N2. Using this subset, the R-square could achieve 65.1% and 47.3%, for H3N2 and H1N1pdm09 VE, respectively (**Figure 2a, Figure 4a**). The EMAS set contains 11 codons for H3N2 and 5 codons for H1N1pdm09, as shown in **Figures S1-2**, in which per mismatch would generate an average reduction in VE of 11.0% (95% CI: 6.4 – 15.5) and 8.4% (95% CI: 3.2 – 13.7), respectively. Genetic distance calculated based on the non-intersection subsets of EM to AS showed no association with H3N2 VE but weaker association with H1N1 pdm09 VE (**Figure 2c, Figure 4c**). Interestingly, GD calculated based on only AS but not EM sites showed no association to VE for both influenza A subtypes.

For NA protein, genetic mismatch on 12 AS codons yielded better prediction performance compared to using 4 EM-AS codons, while other mismatch on EM-nonAS and nonEM-AS were not associated with H3N2 VE as shown in **Figure S5**. The expected VE would decrease by 19.8% (95% CI: 14.5 – 25.1) for per mismatch on NA's antigenic domain and the explained variations of VE by GD could attain 81.6% (**Figure S5**). The results of NA for H1N1 were similar that genetic mismatch on key EMAS codons showed the strongest predictive power compared to using other epitopes. The H1N1pdm09 VE would reduce by 10.6% (95% CI: 4.6 – 16.6) for per EMAS mismatch and the R^2 reach 52.4% (**Figure S6**). Similarly, NA's mismatch on epitopes other than EM sites appeared not associated with VE of two influenza A subtypes.

Integrative analysis for vaccine effectiveness prediction

An integrative analysis was conducted to explore the performance of integrated mismatch for HA and NA in VE evaluation. Except NA of H3N2, genetic distance on EM-AS showed the optimal performance in assessing VE. Hence, we aggregated each part mismatch and investigate their contribution to VE changes. The results showed that the relationship between combined GD and VE was stronger for both H1N1 and H3N2 than using any single protein's mismatch or other epitope mismatch (p-values < 0.01), see in **Figure 5**. Up to 87.8% variations of VE for H3N2 could be explained by GD and for H1N1 this quantity could reach 55.0%.

Mediation analysis of genetic factors on vaccine effectiveness

The HA protein, the target for influenza vaccines, plays a vital role in influenza pathogenesis and accumulated key mutations may induce immune escape and alter the efficacy of immunization. Considering that NA mismatch also contributes to VE changes, a mediation analysis was carried out to identify and explain NA's role in affecting VE. Notably, both direct and indirect effect of genetic mismatch on HA were significant (p-values < 0.05) (**Figure S7**). NA mismatch mediated HA mismatch effect on VE, accounting for 55.2% of the total effect.

Discussion

The use of GD to predict vaccine protection can help to evaluate the impact of influenza immunization programme before mass vaccination and inform vaccination policy details, such as determining whether to introduce a particular vaccine in a specific population or identifying preferred vaccine products for tropical or subtropical regions. In this study, we optimized GD measure for HA and NA by feature selection and found that GD on the intersection set of EM and AS is highly predictive of VE of influenza H1N1 and H3N2. However, the remaining AS but not EM sites showed few potentials in VE evaluation. The genetic mismatch on these AS but non-EM sites of HA and NA was consistently uncorrelated with VE of both influenza A subtypes. While literature reported amino acid substitutions emerged in HA or NA may weaken vaccine protection[25,26], few studies can quantify their effect on VE. This study bridges the molecular-level variations with population-level immunity and refines the quantification method of genetic distance in order to *in silico* predict the effectiveness of influenza vaccines in real time.

This study demonstrated that immunodominant antigenic epitopes are the optimal target to measure GD. Previous research reported that residues in epitopes A and B on HA of influenza H3N2 viruses are important in binding antibodies and genetic variants in these epitopes can help escape antibody response[27-30]. In line with previous findings, our results showed that genetic mismatch on epitopes A and B is associated with VE as amino acid changes on these epitopes could result in vaccine-mismatched influenza viruses predominant thereby causing a reduction in vaccine protection. Similar results are found for the H1N1pdm09 virus that the combined mismatch of sites Sa and Sb on HA can link to VE. The research on selective pressure also demonstrated that positively selected codons were associated with antibody combining epitopes A and B, while not all codons in these sites had predictive value for the evolution of human influenza A and only a few codons on epitopes C, D and E were under positive selection[31], which also support our findings.

Our previous research developed a novel bioinformatics approach to identify key residues relevant to escape from herd immunity, namely EM sites, and demonstrated the feasibility of using these sites to assess VE but failed for NA of H3N2. This work successfully determined genetic

measure for H3N2 NA. While genetic mismatch on AS of NA showed the best performance, the superiority compared to using EM-AS codons is relatively small with a 9.1% difference in terms of the goodness of fit, which also proves the importance of EM-AS in VE evaluation. Previous research also tried to use the dominant epitope as that which shows the most antigenic drift to estimate VE[32-34]. However, a precise determination of which epitopes are dominant was not completed. This study provides useful evidence for identifying key epitopes and measuring vaccine matching levels.

This work can also inform ongoing research efforts to understand the role of NA in the effect of HA mutations on VE. The existing evidence showed that NA-induced antibodies could also provide protection against influenza viruses[35,36] and mutations on NA can also affect viral antigenicity[37]. Inclusion of HA and NA in influenza vaccines provides enhanced protection against influenza-associated illness compared to HA alone[38-40]. From the perspective of genetic epidemiology, we uncovered the potential mechanisms underlying the association of HA mismatch with VE that a significant mediating effect of NA mismatch existed for H3N2.

The limitations of this study include that only a few countries annually investigate the effectiveness of influenza VE and thus it is still a challenge to customize a prediction model for each region. Moreover, numerous studies reported that pre-existing immunity from previous infection and vaccinations can also affect vaccine protective effects[41-46]. The antigenic distance (AD) hypothesis proposed that when prior and current vaccines are antigenically distinct, the influence of the former vaccination should be minimal[47]. When their AD is smaller, positive interference to the current vaccine protection is expected when the AD between prior vaccines and epidemic strains is small, while the interference becomes negative when the AD between pre-existing immunity and circulating viruses is large. In future studies, we will attempt to quantify the impact of pre-existing immunity on VE as data on immune history increases. Despite the limitations, the relationship between GD and VE is deterministic and the key mutation positions involved in this study are also well demonstrated in plenty of biology experiments and bioinformatics analyses. Improved quantification of GD based on EM-AS codons could assist real-time prediction of VE and candidate vaccine strain selection. The predictive models for VE developed herein may facilitate future vaccine deployment and resource coordination.

Acknowledgements

We thank the contributions of all the health care workers and scientists, the GISAID team, and the submitting and the originating laboratories. This work was partially supported by the Health and Medical Research Fund, the Food and Health Bureau, the Government of the Hong Kong Special Administrative Region [COVID190103] (MHW) and [INF-CUHK-1] (EKY), the National Natural Science Foundation of China [31871340] (MHW), and the Chinese University of Hong Kong Grant [PIEF/Ph2/COVID/06, 4054600] (MHW). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We appreciate the constructive comments from the editor and reviewers that improved this work.

Author Contributions

M.H.W conceived the study. L.C processed data, carried out the analysis and wrote the manuscript. J.L., Q.L., C.K.P.M, Z.C., R.W.Y.C., P.P.H.C., M.K.C.C., E.K.Y., P.K.S.C., and B.C.Y.Z. critically read and revised the manuscript and gave final approval for publication.

Competing interests

M.H.W and B.C.Y.Z are shareholders of Beth Bioinformatics Co., Ltd. B.C.Y.Z is a shareholder of Health View Bioanalytics Ltd. S.Y.C and J.L. are employees of Beth Bioinformatics Co., Ltd. All other authors declare no competing interests.

References

1. WHO. Influenza (Seasonal). (2023).
2. Smith, P.G., Rodrigues, L.C. & Fine, P.E. Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. *Int J Epidemiol* **13**, 87-93 (1984).
3. Halloran, M.E., Struchiner, C.J. & Longini, I.M., Jr. Study designs for evaluating different efficacy and effectiveness aspects of vaccines. *Am J Epidemiol* **146**, 789-803 (1997).
4. De Serres, G., Skowronski, D.M., Wu, X.W. & Ambrose, C.S. The test-negative design: validity, accuracy and precision of vaccine efficacy estimates compared to the gold standard of randomised placebo-controlled clinical trials. *Euro Surveill* **18**(2013).
5. Cao, L., *et al.* In silico prediction of influenza vaccine effectiveness by sequence analysis. *Vaccine* **39**, 1030-1034 (2021).
6. Cao, L., *et al.* Differential Influence of Age on the Relationship between Genetic Mismatch and A(H1N1)pdm09 Vaccine Effectiveness. *Viruses* **13**(2021).
7. Cao, L., *et al.* Rapid evaluation of COVID-19 vaccine effectiveness against symptomatic infection with SARS-CoV-2 variants by analysis of genetic distance. *Nat Med* **28**, 1715-1722 (2022).
8. Lee, M.S. & Chen, J.S. Predicting antigenic variants of influenza A/H3N2 viruses. *Emerg Infect Dis* **10**, 1385-1390 (2004).
9. Wang, M.H., *et al.* Characterization of key amino acid substitutions and dynamics of the influenza virus H3N2 hemagglutinin. *J Infect* **83**, 671-677 (2021).
10. Smith, D.J., *et al.* Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**, 371-376 (2004).
11. Koel, B.F., *et al.* Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* **342**, 976-979 (2013).
12. Bao, Y., *et al.* The influenza virus resource at the National Center for Biotechnology Information. *J Virol* **82**, 596-601 (2008).
13. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**(2017).
14. Wiley, D.C., Wilson, I.A. & Skehel, J.J. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* **289**, 373-378 (1981).
15. Gerhard, W., Yewdell, J., Frankel, M.E. & Webster, R. Antigenic structure of influenza virus haemagglutinin defined by hybridoma antibodies. *Nature* **290**, 713-717 (1981).
16. Wilson, I.A. & Cox, N.J. Structural basis of immune recognition of influenza virus hemagglutinin. *Annu Rev Immunol* **8**, 737-771 (1990).
17. Caton, A.J., Brownlee, G.G., Yewdell, J.W. & Gerhard, W. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell* **31**, 417-427 (1982).

18. Brownlee, G.G. & Fodor, E. The predicted antigenicity of the haemagglutinin of the 1918 Spanish influenza pandemic suggests an avian origin. *Philos Trans R Soc Lond B Biol Sci* **356**, 1871-1876 (2001).

19. Gulati, U., *et al.* Antibody epitopes on the neuraminidase of a recent H3N2 influenza virus (A/Memphis/31/98). *J Virol* **76**, 12274-12280 (2002).

20. Wan, H., *et al.* Structural characterization of a protective epitope spanning A(H1N1)pdm09 influenza virus neuraminidase monomers. *Nat Commun* **6**, 6114 (2015).

21. International Schizophrenia, C., *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).

22. Torkamani, A., Wineinger, N.E. & Topol, E.J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* **19**, 581-590 (2018).

23. Team, R.C. R: A language and environment for statistical computing. (2013).

24. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. Mediation: R package for causal mediation analysis (2014).

25. Tenforde, M.W., *et al.* Effect of Antigenic Drift on Influenza Vaccine Effectiveness in the United States-2019-2020. *Clin Infect Dis* **73**, e4244-e4250 (2021).

26. Gao, J., *et al.* Antigenic Drift of the Influenza A(H1N1)pdm09 Virus Neuraminidase Results in Reduced Effectiveness of A/California/7/2009 (H1N1pdm09)-Specific Antibodies. *mBio* **10** (2019).

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

28. Cox, N.J. & Bender, C.A. The molecular epidemiology of influenza viruses. *Seminars in Virology* **6**, 359-370 (1995).

29. Beer, K., *et al.* Characterization of neutralizing epitopes in antigenic site B of recently circulating influenza A(H3N2) viruses. *J Gen Virol* **99**, 1001-1011 (2018).

30. Ndifon, W., Wingreen, N.S. & Levin, S.A. Differential neutralization efficiency of hemagglutinin epitopes, antibody interference, and the design of influenza vaccines. *Proc Natl Acad Sci U S A* **106**, 8701-8706 (2009).

31. Bush, R.M., Bender, C.A., Subbarao, K., Cox, N.J. & Fitch, W.M. Predicting the evolution of human influenza A. *Science* **286**, 1921-1925 (1999).

32. Bonomo, M.E. & Deem, M.W. Predicting Influenza H3N2 Vaccine Efficacy From Evolution of the Dominant Epitope. *Clin Infect Dis* **67**, 1129-1131 (2018).

33. Xu, Y. & Ye, J. Predicting Influenza Vaccine Effectiveness With Hemagglutinin Sequences Is Not Enough. *Clin Infect Dis* **68**, 346 (2019).

34. Muñoz, E.T. & Deem, M.W. Epitope analysis for influenza vaccine design. *Vaccine* **23**, 1144-1148 (2005).

35. Giurgea, L.T., Morens, D.M., Taubenberger, J.K. & Memoli, M.J. Influenza Neuraminidase: A Neglected Protein and Its Potential for a Better Influenza Vaccine. *Vaccines (Basel)* **8** (2020).

36. Eichelberger, M.C. & Monto, A.S. Neuraminidase, the Forgotten Surface Antigen, Emerges as an Influenza Vaccine Target for Broadened Protection. *J Infect Dis* **219**, S75-S80 (2019).

37. Wan, H., *et al.* The neuraminidase of A(H3N2) influenza viruses circulating since 2016 is antigenically distinct from the A/Hong Kong/4801/2014 vaccine strain. *Nat Microbiol* **4**, 2216-2225 (2019).

38. Johansson, B.E., Matthews, J.T. & Kilbourne, E.D. Supplementation of conventional influenza A vaccine with purified viral neuraminidase results in a balanced and broadened immune response. *Vaccine* **16**, 1009-1015 (1998).

39. Bosch, B.J., *et al.* Recombinant soluble, multimeric HA and NA exhibit distinctive types of protection against pandemic swine-origin 2009 A(H1N1) influenza virus infection in ferrets. *J Virol* **84**, 10366-10374 (2010).

40. Johansson, B.E. Immunization with influenza A virus hemagglutinin and neuraminidase produced in recombinant baculovirus results in a balanced and broadened immune response superior to conventional vaccine. *Vaccine* **17**, 2073-2080 (1999).

41. Skowronski, D.M., *et al.* Serial Vaccination and the Antigenic Distance Hypothesis: Effects on Influenza Vaccine Effectiveness During A(H3N2) Epidemics in Canada, 2010-2011 to 2014-2015. *J Infect Dis* **215**, 1059-1099 (2017).

42. Keitel, W.A., Cate, T.R., Couch, R.B., Huggins, L.L. & Hess, K.R. Efficacy of repeated annual immunization with inactivated influenza virus vaccines over a five year period. *Vaccine* **15**, 1114-1122 (1997).

43. McLean, H.Q., *et al.* Impact of repeated vaccination on vaccine effectiveness against influenza A(H3N2) and B during 8 seasons. *Clin Infect Dis* **59**, 1375-1385 (2014).

44. Skowronski, D.M., *et al.* A Perfect Storm: Impact of Genomic Variation and Serial Vaccination on Low Influenza Vaccine Effectiveness During the 2014-2015 Season. *Clin Infect Dis* **63**, 21-32 (2016).

45. Ohmit, S.E., *et al.* Influenza vaccine effectiveness in households with children during the 2012-2013 season: assessments of prior vaccination and serologic susceptibility. *J Infect Dis* **211**, 1519-1528 (2015).

46. Kim, J.H., *et al.* Prior infection with influenza virus but not vaccination leaves a long-term immunological imprint that intensifies the protective efficacy of antigenically drifted vaccine strains. *Vaccine* **34**, 495-502 (2016).

47. Katz, J.M., Hancock, K. & Xu, X. Serologic assays for influenza surveillance, diagnosis and vaccine evaluation. *Expert Rev Anti Infect Ther* **9**, 669-683 (2011).

Figures

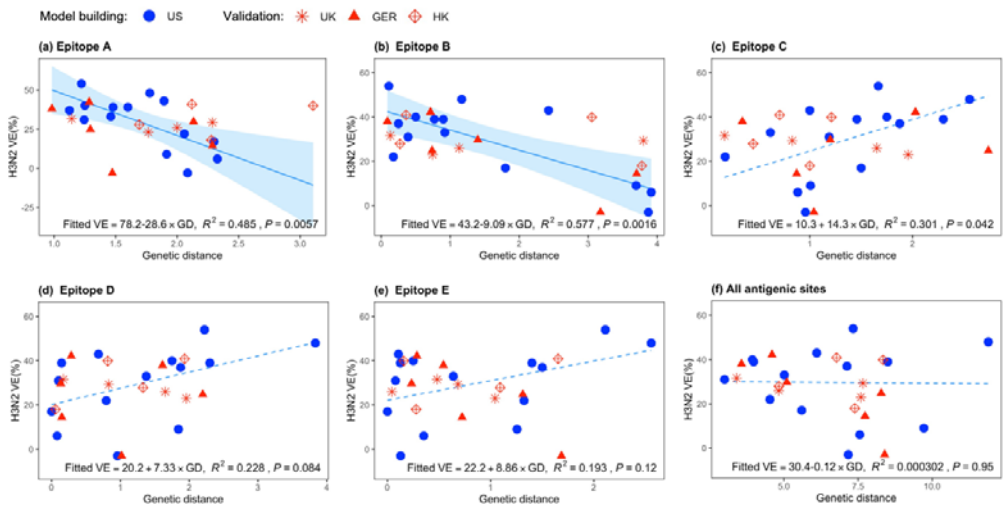

Figure 1. The relationship between VE and genetic distance on five epitopes of H3N2 hemagglutinin (HA) protein

Blue dots: data from the United States for model building; red dots: independent data from the United Kingdom, Germany and Hong Kong for validation. (a) H3N2 VE against genetic distance on epitope A of HA; (b) H3N2 VE against genetic distance on epitope B; (c) H3N2 VE against genetic distance on epitope C; (d) H3N2 VE against genetic distance on epitope D; (e) H3NA VE against genetic distance on epitope E; (f) H3N2 VE against genetic distance on all AS.

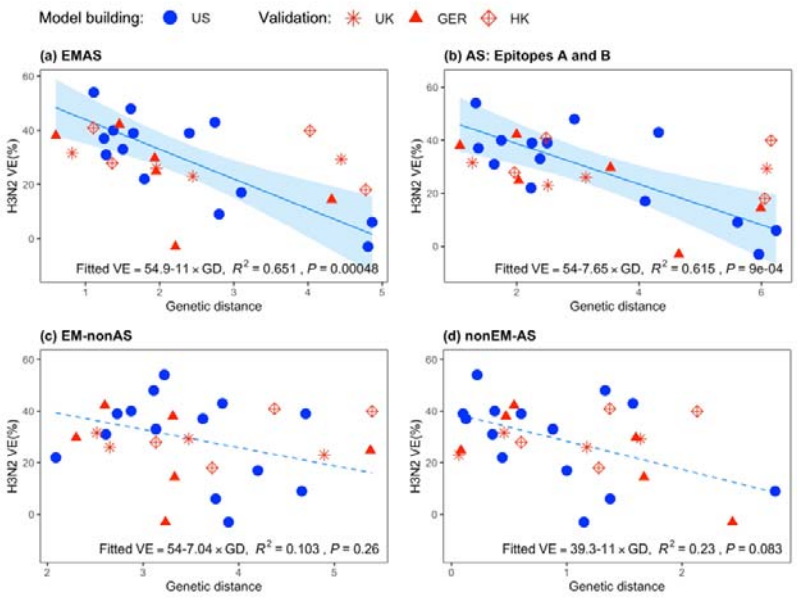

Figure 2. The best quantification of genetic distance for HA of the H3N2 virus in VE evaluation

(a) H3N2 VE against genetic distance on the intersection set of EM and AS for HA; (b) H3N2 VE against genetic distance on partial AS (epitopes A and B); (c) H3N2 VE against genetic distance on EM-nonAS; (d) H3N2 VE against genetic distance on nonEM-AS. AS sites here refers to codons on epitope A and B.

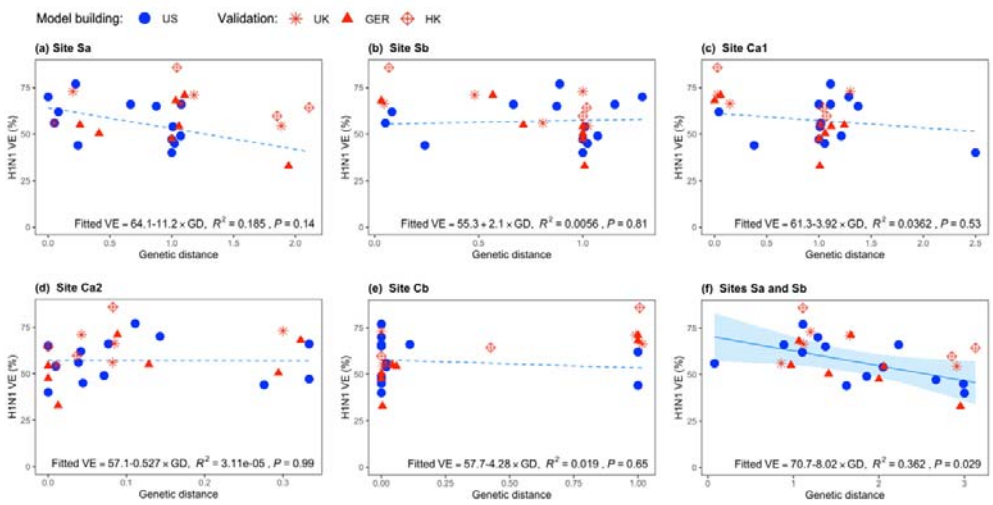

Figure 3. The relationship between VE and genetic distance on five AS of H1N1 HA protein

(a) H1N1 VE against genetic distance on site Sa of HA; (b) H1N1 VE against genetic distance on site Sb; (c) H1N1 VE against genetic distance on site Ca1; (d) H1N1 VE against genetic distance on site Ca2; (e) H1N1 VE against genetic distance on site Cb; (f) H1N1 VE against the integrated genetic distance of Sa and Sb.

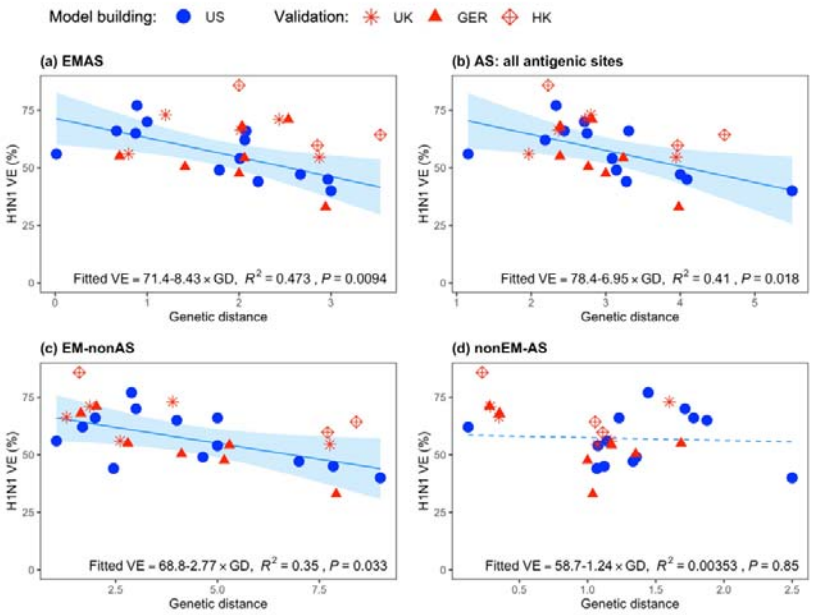

Figure 4. The best quantification of genetic distance for HA of the H1N1 virus in VE evaluation

(a) H1N1 VE against genetic distance on the intersection set of EM and AS for HA; (b) H1N1 VE against genetic distance on all AS; (c) H1N1 VE against genetic distance on EM-nonAS; (d) H1N1 VE against genetic distance on nonEM-AS. AS sites here refers to codons on all AS.

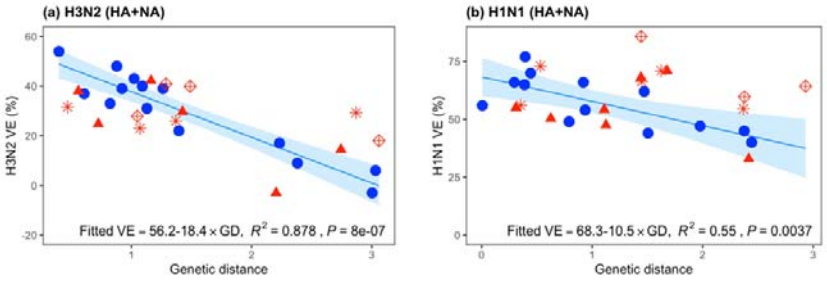

Figure 5. Prediction of VE for H3N2 and H1N1 by integrative analysis of HA and NA sequences

Panels (a-b): Relationship of integrated genetic distance on the HA and NA genes with VE for two influenza A subtypes.