Improving in silico prediction of influenza vaccine effectiveness by genome analysis incorporating epitope information

Authors:

Lirong Cao[1,2], Jingzhi Lou[1], Qi Li[1,2], Hong Zheng[1,2], Chris Ka Pun Mok[1,3], Zigui Chen[4], Renee Wan Yi Chan[5,6], Peter Pak Hang Cheung[7], Marc Ka Chun Chong[1,2], Eng Kiong Yeoh[1,8], Paul Kay Sheung Chan[5,9], 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] Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China.

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

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

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

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

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

[9] 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 87.1% 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[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

Vaccine effectiveness data and viral sequences

The VE data were extracted from published epidemiological studies. The VE data included in the analysis met the criteria below: (1) Studies applied test-negative design to evaluate VE; (2) RT-PCR

test was used to confirm influenza infections; (3) Target population is all-age group; (4) The primary outcome is influenza-like illness; (5) Studies included vaccine effectiveness against influenza A. The flu VE surveillance is continuous and exhaustive in Northern America. Compared to many studies included different country populations, the population cohort is relatively homogeneous in US and Canadian cohorts. Thus, Northern America is an ideal region to develop VE-GD models and perform validation. We included all VE observations published by the Centers for Disease Control and Prevention of US and Canada, and other VE reports met the including criteria. A complete summary of the VE data was described in the supplementary materials **Tables S1-2.**

Sequencing data of influenza A H1N1pdm09 and H3N2 were downloaded from the NCBI[12], and GISAID EpiFlu databases[13]. All available sequences that matched to the flu seasons and locations of the observational studies were downloaded. We thus obtained 2,408 hemagglutinin (HA) and neuraminidase (NA) protein sequences from United States between 2009/10 and 2018/19 flu seasons for model building. For validation, we retrieved 1,566 sequences from Canada during 2005/06 and 2018/19 flu seasons. Multiple sequence alignment was performed by MEGA (7.0.26). The influenza vaccine strains recommended by WHO for use in the northern hemisphere were summarized in supplement materials **Table S3**.

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 were listed in **Table S4** and EM sites in **Table S5**. The AS and EM locations on stereo view of HA and NA protein structures were plotted in **Figures S1-4**, obtained from the RSCB Protein Data Bank (PDB) (https://www.rcsb.org). The comparison of AS and EM sites was provided in **Figure 1**. 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 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.

Cross validation

Since VE data is scarce, we used leave-one-out cross-validation methods to test the VE-GD model performance. For the nth part, we fit the model to the other $N - 1$ parts of the data, and calculate the prediction error of the fitted model when predicting the nth part of the data. We do this for $n = 1, 2, \ldots, N$ and combine the N estimates of prediction error. Denote by $\hat{f}^{-n}(d)$ the fitted function, computed with the nth part of the data removed. Then the cross-validation estimate of prediction error is

$$CV(\hat{f}) = \frac{1}{N} \sum_{n=1}^{N} L(y_n, \hat{f}^{-n}(d_n)). \qquad \text{Equation 4}$$

For the nth observation the fit is computed using all the data except the nth. Herein, mean absolute error (MAE) was applied to measure regression model error. The concordance correlation coefficient (CCC) was used to measure precision and accuracy of prediction models.

Mediation analysis

With mediation analysis, we can gain insight and acquire deep understanding about the mechanism of mutations of different proteins affecting viral immune evasion. Such information provides an added dimension to understand how virus evolution causes low vaccine protection, which can stimulate the identification of more efficacious and predominant vaccine candidates. We postulated the causal model that is, HA mismatch led to increased NA mismatch and also led to some decrease in vaccine protection against disease from influenza virus infection. The effect caused by HA mutations was mediated by the co-mutations occurred on NA protein (referred to as an "indirect effect" in causal mediation terminology), and part of the drop in VE was introduced via HA-only mutations by other immune mechanisms ("direct effect" not mediated by NA mismatch). 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

We firstly investigated HA protein of two influenza A subtypes circulating among humans and found that the optimal predictor codon sites to quantify genetic mismatch between vaccine strain and circulating viruses are the intersection set of EM and AS (noted as EM-AS). The GD on EM-AS is significantly associated with VE against both H1N1pdm and H3N2. The same patterns were observed for NA protein that the genetic mismatch on EM-AS region showed a stronger association with VE for both subtypes. After aggregating the mismatch on HA and NA for each subtype, the integrated GD showed the best prediction power for VE (**Figure 2**).

Which antigenic sites are the key to maintain high level of vaccine protection?

We calculated genetic mismatch on five antigenic sites of HA for influenza H1N1pdm09 (Sites Sa, Sb, Ca1, Ca2 and Cb) and H3N2 (Epitopes A-E). For H1N1pdm09, no obvious trend was observed between genetic mismatch on any single AS and VE (**Figure S5a-e**), but after aggregating genetic mismatch on Sa and Sb sites, a significant relationship with VE was detected (*p*-value = 0.029). We conducted feature selection and found that using all AS mismatch performs better than using Sa and Sb sites in VE estimation (**Figure S6c**), while no significant association with VE was observed for other combinations.

For H3N2 from single epitope level, we found that GD were significantly associated with VE for epitopes A and B, while no relationships were observed for genetic mismatch on epitopes C-E (**Figure S7a-e**). Moreover, all epitopes mismatch also showed no obvious relationship with VE (**Figure S7f**). The feature selection results illustrated that GD on the union of epitopes A and B were significantly associated with VE and the goodness of fit in terms of R-squared can reach 61.5% higher

than single epitope model (**Figure S8c**). As for other combinations of single epitope, no significant relationship between genetic mismatch and VE was detected. Thus, H1N1 VE could be better explained by the GD based on all AS, while the total GD on epitopes A and B was superior in predicting H3N2 VE.

Comparing prediction performances of antigenic sites and effective mutations sites for vaccine effectiveness

Previously, we demonstrated that genetic mismatch of HA based on EM sites was negatively associated with VE against two influenza A subtypes, as was NA mismatch for H1N1pdm09. The key to modeling relationships between VE and GD lies in how to appropriately quantify genetic factors. We compared prediction performances of AS and EM sites and explored how to optimize VE-GD model through integrating biological information. For HA protein of H1N1pdm09, we found that EM sites and AS performed similarly in estimating VE with an R-squared of approximately 41%, while the performance could increase to 47.3% by using EM-AS as predictor codon sites (**Figure S6**). The results were similar for HA protein of H3N2 virus that the highest predictive power was observed for the GD on EM-AS and up to 65.1% VE variations could be explained by a single factor EM-AS mismatch (**Figure S8**). The results were consistent for NA protein. While genetic mismatch on EM sites or AS was also associated with H1N1 VE, the model fitting result was better by using EM-AS with a goodness-of-fit of 52.4% (**Figure S9**). Similarly, for H3N2 VE estimation, EM-AS is highly correlated with vaccine protection and explained variations by GD could attain 81.8% (**Figure S10**).

We also checked the associations of GD on other genetic regions with VE. We found that GD based on the non-intersection subsets of EM to AS (noted as EM-nonAS) showed weaker association with H1N1 VE for both HA and NA (**Figure S11a,c)**. Interestingly, no significant relationship was found for GD calculated by only AS but not EM sites (noted as nonEM-AS) for two proteins (**Figure S11b,d)**. However, for H3N2, EM-nonAS and nonEM-AS mismatch both showed no obvious association with VE for HA and NA proteins (**Figure S12**). The model selection results were consistent for HA and NA and for both influenza A subtypes, which showed that the EM-AS performs the best in VE prediction.

The integrative analysis of multiple proteins for vaccine effectiveness prediction

After establishing the optimal predictor codon sites of each protein for VE estimation, we applied integrative analysis to build the VE-GD models for two influenza A subtypes. As shown in Figure 1, the expected VE would decrease by 8.4% (95% CI: 3.2 – 13.7) per mismatch of HA protein, and 10.6% (95% CI: 4.6 – 16.6) per NA mismatch (**Figure 2a,b**). The integrated GD of HA and NA proteins was highly associated with VE and the goodness-of-fit could reach 55.0%, highest than all other H1N1 VE-GD models (**Figure 2c**). For H3N2 virus, per mismatch of HA would generate an average reduction in H3N2 VE of 11.0% (95% CI: 6.4 – 15.5), while 20.2% (95% CI: 14.8 – 25.6) for NA

protein (**Figure 2d,e**). We found that up to 87.1% variations of H3N2 VE could be explained by genetic mismatch of HA and NA proteins (**Figure 2f**).

Validation and retroactive predictions for 2019/20 to 2022/23 flu seasons

The leave-one-out cross validation was conducted for the integrative VE-GD model for two influenza A subtypes. The average distance between the observed values and the predicted VE was 7.1% for H1N1 and 6.3% for H3N2. The calibration plot of **Figure 3** depicts that the validation data points are surrounding the 45° line, meaning that the predictions are nearly identical to the observed VE. The test statistic CCC was 90.6% (95% CI: 73.8-96.9) for the H3N2 model and 60.2% (95% CI: 13.3-85.1) for the H1N1 model. In **Figure 2**, the independent data from Canada also showed an obvious negative relationship between GD and VE against H1N1 and H3N2. The Canadian VE data points were always around the fitted line of US models for single protein and integrated models. In addition, we conducted retroactive predictions for 2019/20 to 2022/23 flu seasons (**Figure 4**, supplementary materials **Table S6)** and found that the predicted H3N2 VE for US was 35.8% (95% prediction interval: 21.4-50.2), close to the observation of 36% (95% CI: 20-49) in 2021/22 flu seasons[25]. In the next flu season, the estimated H3N2 VE was 49.0% (95% prediction interval: 34.0-64.1), which is in very close agreement with the reported effectiveness of 45% (95% CI: 29-58) by US CDC[26], which indicates good predictive power of the model.

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 cape and alter the efficacy of immunization. Considering that NA mismatch could also lead to a decrease in VE, a mediation analysis was carried out to explore the relationship of two protein mismatch to VE. Notably, both direct and indirect effect of genetic mismatch on HA were significant (p-values < 0.05) (**Figure 5**). For H3N2 virus, the direct effect of GD on HA on VE, that is, the effect not related to NA substitutions, was estimated as a 4.7% (95% CI: 1.5 – 8.2) decrease in VE per mismatch. The indirect effect due to the HA-NA mismatch pathway resulted in a 6.3% (95% CI: 1.5 – 10.6) VE drop per unit, accounting for 57.0% of the total effect. Currently, no significant indirect effect was detected for H1N1 virus.

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 on HA or NA may weaken vaccine protection[27,28], 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[29-32]. 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[33], 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[34-36]. 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[37,38] and mutations on NA can also affect viral antigenicity[39]. Inclusion of HA and NA in influenza vaccines provides enhanced protection against influenza-associated illness compared to HA alone[40-42]. 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[43-48]. The antigenic distance (AD) hypothesis proposed that when prior and current vaccines are antigenically distinct, the influence of the former vaccination should be minimal[49]. 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 Strategic Seed Funding for Collaborative Research Scheme [3136001] and the Chinese University of Hong Kong Direct Grant [4054738]. 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., H.Z., 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. 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. Price, A.M., *et al.* Influenza Vaccine Effectiveness Against Influenza A(H3N2)-Related Illness in the United States During the 2021-2022 Influenza Season. *Clin Infect Dis* **76**, 1358-1363 (2023).

26. CDC. Preliminary Flu Vaccine Effectiveness (VE) Data for 2022-2023. Retrieved from https://www.cdc.gov/flu/vaccines-work/2022-2023.html. (2023).

27. 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).

28. 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).

29. 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).

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

31. 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).

32. 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).

33. 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).

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

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

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

37. 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).

38. 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).

39. 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).

40. 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).

41. 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).

42. 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).

43. 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).

44. 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).

45. 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).

46. 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).

47. 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).

48. 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).

49. 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).

50. Skowronski, D.M., *et al.* Interim estimates of 2019/20 vaccine effectiveness during early-season co-circulation of influenza A and B viruses, Canada, February 2020. *Euro Surveill* **25**(2020).

Figures

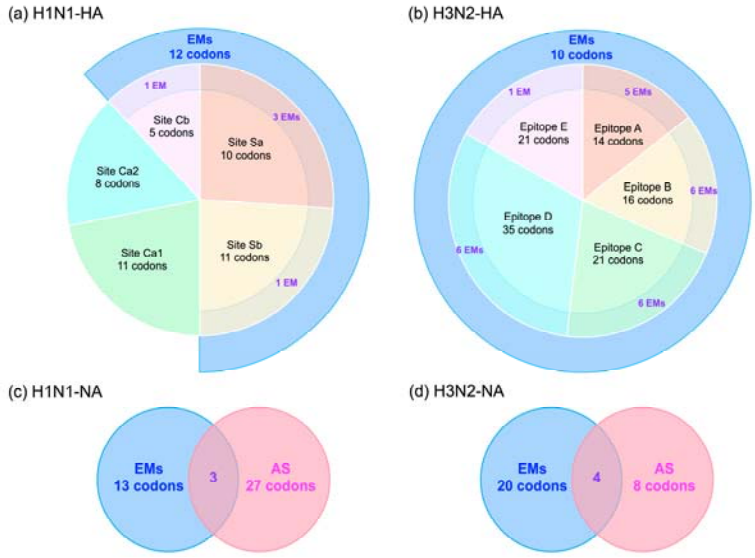

Figure 1. Venn diagram for comparing antigenic sites (AS) and effective mutation (EM) sites.

We compared EM sites and AS and identified the intersection codons of two genetic regions. The list of AS and EM sites are provided in supplementary materials Table S4 and Table S5, respectively.

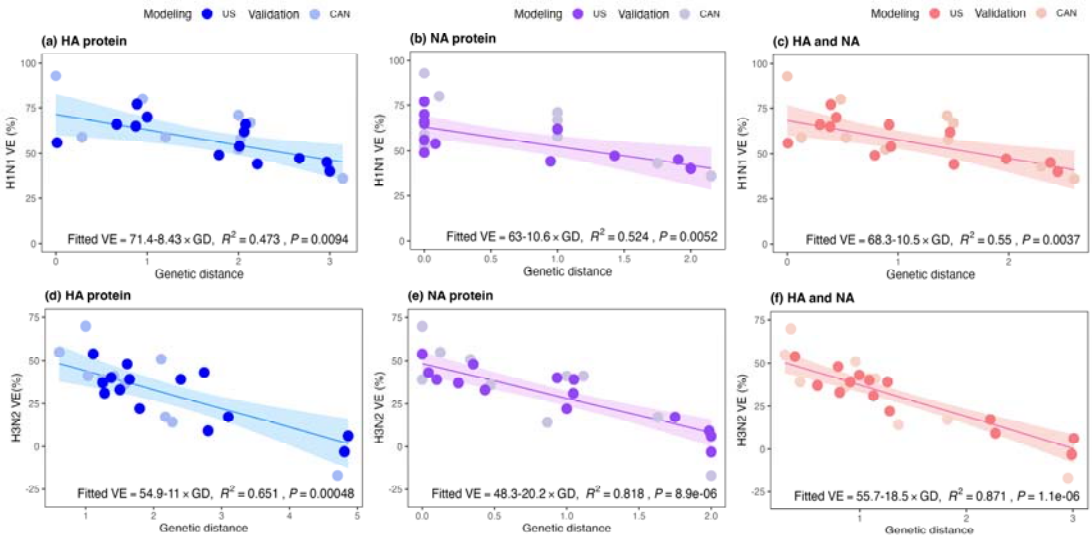

Figure 2. Prediction models of vaccine protection for two influenza A subtypes and validation.

Model selection was conducted for hemagglutinin (HA) and neuraminidase (NA) to determine the optimal predictor codon site for vaccine effectiveness (VE) prediction. After a host of experiments, the intersection of EM sites and AS (noted as EM-AS) was established as a best target to measure genetic measure. Panels (a-b) and (c-d) were calculated by genetic distance (GD) on EM-AS. Panels (c) and (d) were based on the integrated HA and NA mismatch.

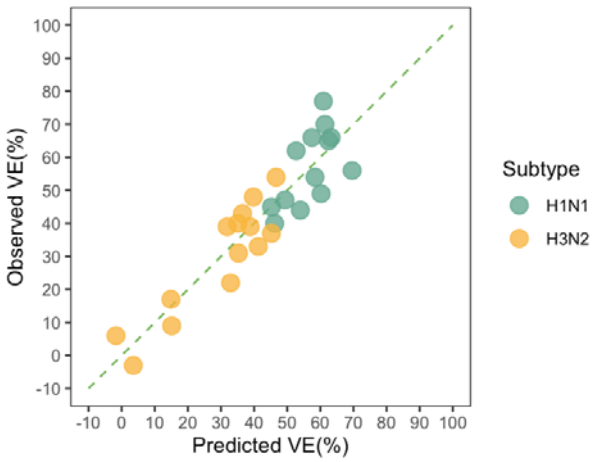

Figure 3. Calibration plot for the prediction outcome by cross validation.

The dotted line is the 45° line. Data points near the dashed line mean that the predicted value is close to the observed VE.

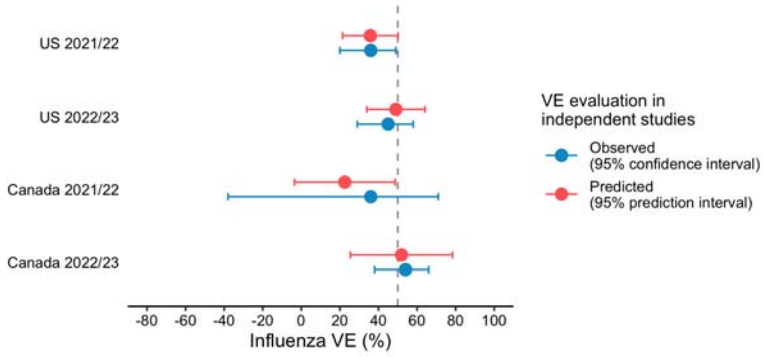

Figure 4. Retroactive prediction for H3N2 VE from 2019/20 to 2022/23 flu seasons.

To validate the VE-GD predictive power, we predicted vaccine protective effects from 2019/20 to 2022/23 flu seasons. Due to the COVID-19 pandemic, the VE observations of 2019/20 to 2020/21 were not available. The detailed prediction results were shown in **Table S6**. While H3N2 VE for 2019/20 in Canada was reported, the author emphasized "particularly noting the limited sample size of A(H3N2) cases, our interim estimate of 2019/20 A(H3N2) VE requires cautious interpretation pending further end-of-season evaluation our interim estimate of 2019/20 A(H3N2) VE requires cautious interpretation pending further end-of-season evaluation", thus this data was not included in the analysis[50].

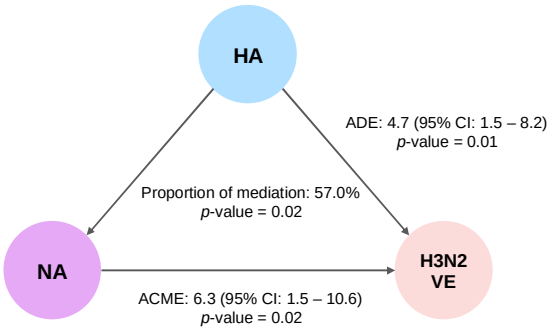

Figure 5. Mediation model showing the hypothesized relationship between genetic mismatch and influenza H3N2 VE.

The NA mismatch acts as a mediator of HA mismatch on vaccine protection. HA mismatch has an indirect effect on VE, which operates via a change on NA protein and thus causes additive effect. HA substitutions may also have a direct effect on VE that is not mediated by NA mismatch but operates through other immune mechanisms. Average direct effect (ADE), average causal mediation effects (ACME) and the proportion of mediation are reported.