Passage adaptation correlates with the reduced efficacy of the influenza vaccine

Hui Chen[1,2], Jacob Josiah Santiago Alvarez[1,3], Sock Hoon Ng[4], Rasmus Nielsen[5,6], Weiwei Zhai[1,2,7,8,9,¶]

[1]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, P.R.China

[2]Human Genetics, Genome Institute of Singapore, A*STAR, Singapore

[3]Department of Biological Science, National University of Singapore, Singapore

[4]DSO National Laboratories, Singapore, Singapore

[5]Department of Integrative Biology, University of California, Berkeley, Berkeley, California 942720, USA

[6]Department of Statistics, University of California, Berkeley, Berkeley, California 942720, USA

[7]School of Biological Sciences, Nanyang Technological University, Singapore

[8]National Cancer Center, Singapore, Singapore

[9]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, P.R.China

¶ Correspondence should be addressed to weiweizhai@ioz.ac.cn or zhaiww1@gis.a-star.edu.sg

Summary: Using a powerful probabilistic method, we found the strength of passage adaptation (substitutions during vaccine production in embryonated eggs) correlates with influenza vaccine efficacy, suggesting passage adaptation is a strong contributing factor for the poor performance of the influenza vaccine.

Abstract

Background

As a dominant seasonal influenza virus, H3N2 virus rapidly evolves in human and is a constant threat to public health. Despite sustained research efforts, the efficacy of H3N2 vaccine has decreased rapidly. Even though antigenic drift and passage adaptation (substitutions accumulated during vaccine production in embryonated eggs) have been implicated in reduced vaccine efficacy, their respective contributions to the phenomenon remain controversial.

Methods

We utilized mutational mapping, a powerful probabilistic method of studying sequence evolution, to analyze patterns of substitutions in different passage conditions for an unprecedented amount of H3N2 hemagglutinin sequences (n=32278).

Results

We found that passage adaptation in embryonated eggs is driven by repeated convergent evolution over 12 codons. Based on substitution patterns at these sites, we developed a metric, Adaptive Distance (AD), to quantify the strength of passage adaptation and subsequently identified a strong negative correlation between AD and vaccine efficacy.

Conclusions

The high correlation between AD and vaccine efficacy implies that passage adaptation in embryonated eggs may be a strong contributor to the recent reduction in H3N2 vaccine efficacy. We developed a computational package called MADE to measure the strength of passage

adaptation and predict the efficacy of a candidate vaccine strain. Our findings hence shed light on strategies that reducing Darwinian evolution within the passaging medium can potentially restore an effective vaccine program in the coming future.

Key words: H3N2 influenza virus; vaccine efficacy; passage adaptation; mutational mapping

Introduction

Influenza viruses infect 5-15% of the total population, which result in an estimated half a million deaths annually [1, 2]. Vaccination against influenza viruses is the primary strategy for prevention and control of influenza outbreaks [3]. Despite decades of effort, vaccine efficacy against influenza viruses, particularly that of the H3N2 subtype has rapidly dropped to 10% or lower, calling into question the effectiveness of the current vaccination program [4, 5]. Consequently, understanding and controlling the sharp decline in vaccine efficacy is a pressing public health challenge.

Mismatches between circulating and vaccine strains are a critical factor affecting vaccine efficacy [6]. Two major sources of this mismatch are antigenic drift [7, 8] and passage adaptation [9]. While antigenic drift refers to the continued evolution of the influenza virus in host humans, passage adaptation relates to substitutions accumulated while propagating viral strains in the culturing medium during vaccine production. Even though the importance of antigenic drift and passage adaptation in vaccine efficacy have been proven in independent studies [9, 10], their relative contributions remain unknown and conclusions drawn between individual studies are discordant [11-13]. Since WHO enforces a consistent and rigorous vaccine program which matches the vaccine isolate to the circulating strains throughout the years, recent declines in VE may not be caused by antigenic drift, but might instead be driven by passage adaptation during vaccine production.

Currently, embryonated eggs [14] or mammalian cell lines (e.g. Madin Darby Canine Kidney Cell line, i.e. MDCK [15]) are mainly used for propagating influenza viruses. The receptors for viral entry into the host cell are sialic acids found on the surface of cells, which have two different forms of sialic acid linkages to galactose. The alpha-2,3 glycosidic-bond-linked sialic acids (SAα-2,3Gal) are found mostly in the intestinal tracts of birds, while the alpha-2,6 glycosidic-bond-linked residues (SAα-2,6Gal) predominate in the upper respiratory tract of humans [16]. Growing human influenza viruses which preferentially bind to the SAα-2,6Gal receptor can therefor cause strong adaptive evolution during virus culture in embryonated eggs which mainly contain the alternative SAα-2,3Gal form [17, 18]. One recent study revealed a steady increase in passage adaptation of the H3N2 virus in embryonated eggs [18] that seems to be correlated with the reduction in vaccine efficacy, further supporting the importance of passage adaptation in the recent reduction in vaccine efficacy [9, 19].

With rapid progress in viral sequencing, quantifying adaptive evolution in large datasets (>10000 sequences) is becoming a challenging task for the researchers due to the increased computational burden. To overcome this, a probabilistic approach known as ‘stochastic mutational mapping’ [20] has emerged as a powerful and efficient approach for sampling possible mutational histories of observed data [20, 21]. Using theories derived from the continuous time Markov Chain, mutational mapping first infers possible histories of sequence evolution for the observed data and subsequently many statistical tests can be constructed to test different hypotheses such as the existence of positive evolution. Simulation studies show that mutational mapping provides a very efficient and accurate method for sequence evolution [18, 20, 21]. In this work, we will employ

this approach to explore the correlation between passage adaptation and vaccine efficacy, and elucidate the genetic determinants of recent vaccine competence.

Materials and Methods

Sequence curation and passage histories

We analyzed a total of 38360 influenza H3N2 HA1 sequences from the Global Initiative on Sharing All influenza sequence data were summarized from Global Initiative on Sharing All Influenza Data (GISAID) database with collection dates from 1968/1/1 to 2016/07/08 [22]. After multiple quality control steps, 32278 sequences were used for subsequent analysis (Supplementary Material, Supplementary Table 1 and Supplementary Figure 1&2). Passage information of these sequences was extracted from both CDC/NCBI [17] and GISAID databases [22]. We categorized sequences according to their passage histories [18]. 661 egg-passaged sequences and 4544 MDCK-passaged sequences were identified (Figure 1A). Other passage information include CELL (3286), SIAT (3712), Clinical (4046), Monkey (2019), Mix (4088), others (1086) as well as unknown (8836).

Phylogenetic reconstruction and mutational mapping

Sequence alignment for 32278 sequences were performed using MUltiple Sequence Comparison by Log-Expectation (MUSCLE) [23]. After sequence alignment, the phylogenetic relationship among these sequences was reconstructed using Randomized Axelerated Maximum Likelihood

(RAxML) package under a General Time Reversible (GTR) model (Figure 1A) [24]. The probabilistic mutational mapping algorithm has been shown to be an efficient approach for sampling evolutionary histories according to their posterior probabilities [20]. Using the inferred maximum likelihood parameters and associated phylogenetic tree, mutational mapping was carried out as previous studies [18, 20, 21] (Supplementary Material). The mutational history inference was conducted at the base level and provided an efficient and accurate inference of the mutational events. After mutational mapping, summary statistics such as the number of substitutions that occurred along a given set of branches were extracted from the mutational mapping output.

The convergent and enrichment test

Changes at each of the three codon positions were conjugated together to examine the functional consequence of these substitutions. The convergent and enrichment tests survey two different aspects of the adaptive evolution. The enrichment test examines substitutions in terminal branches and investigates whether nonsynonymous substitutions at a given codon tend to happen specifically over terminal branches associated with certain passage histories (Figure 1B, Supplementary Material). The convergent test examines whether a given nonsynonymous transition between two codon states occurred more frequently than by chance. Using synonymous substitution rates as the baseline, we tested if a given nonsynonymous substitution takes place more often than by chance (Figure 1B, Supplementary Material). After calculating p-value from both the enrichment and convergent test, Fisher's method was used to combine these two p-values and multiple test correction was performed in R using the p.adjust() function. By taking the average q-value across 100 replicates from the mutational histories and using a false discovery rate cutoff of 0.01, we identified 12 codon positions driving egg passage adaptation.

Enrichment scores and principal component analysis

In this study, we define an allele as a specific amino acid at a given codon position. For each allele observed at the 12 codons, we first calculated the proportion of strains carrying the specific allele in the focal set (e.g. egg passaged sequences, p_{egg}) and the background set (all sequences in this data set, p_{all}). The enrichment scores were then calculated using p_{egg}/p_{all} for all the alleles across the 12 codons. For each strain, the 12 alleles observed at the 12 positions were combined to define a 12-dimensional vector, indicating enrichment levels at these codons (i.e. similar to allelic barcodes). Each of these 32278 sequences is represented by a data point in the 12-dimensional space. Principal Component Analysis (PCA) was performed in R using the princomp() function to project the 12 dimensional space into the first two PCs.

Vaccine isolates and vaccine efficacy data

Given that seed viruses must be generated in eggs in the current WHO recommended vaccine production protocol [25], we collected the egg passaged strains in the public database with the same strain name as WHO recommended vaccine strains. The list of vaccine strains recommended by WHO were curated from WHO and Influenza Research Database (IRD) website [26]. When multiple egg-passaged sequences are available for a given strain, we included all of them to represent possible sequence variations during passage in embryonated eggs. The full list of vaccine strains is defined as the vaccine set and their GISAID isolate names are listed in the Supplementary Table 3. In total, 61 sequences from 13 vaccine strains were

extracted. In Figure 2H, the vaccine efficacy data was curated from Belongia et al. [5] where vaccine efficacies between 2010-2015 were calculated by pooling multiple test-negative studies using a random-effects model [5]. The adaptive distance for a given year was calculated as the mean of the adaptive distances for all the strains of that year in the vaccine set.

Results

Data curation and mutational mapping

32278 HA1 sequences of the H3N2 influenza virus were downloaded from GISAID database along with their passage histories. Of these sequences, 661 were from isolates cultured in embryonated eggs and more than 70% had passage annotations (Supplementary Figure 2A). After sequence alignment and the phylogenetic reconstruction (Figure 1A), maximum likelihood estimates of the mutational process (e.g. the GTR model) as well as the phylogenetic relationship were constructed. The probabilistic mutational mapping was subsequently employed to sample the nucleotide sequence of the internal nodes recursively from the root of the tree down to the tips, followed by simulating the mutational history along each branch according to the Markov Chain specified by the GTR model (Methods). Since passage adaptation will add extra changes to the original sequences and further extend the terminal branches [27], we merely focus on substitutions that occur on terminal branches.

Positively selected codons under egg passage adaptation and their functional roles

Similar to our previous work [18], in order to identify codon positions driving passage adaptation, we coupled the inferred substitutions at the three codon positions and used two statistical tests targeting different aspects of the mutational pattern across the phylogenetic tree. The first test (the enrichment test) investigates whether substitutions at a given codon tend to be enriched along terminal branches for the egg-passaged strains (n=661) relative to other terminal branches (n=31617) (Figure 1B and Methods). The second test (the convergent test) examines

whether certain nonsynonymous substitutions are observed more frequently than expected by chance along egg terminal branches (n=661, Figure 1B). We note that the tests for such patterns, using probabilistic mutational mapping, are specifically constructed to take uncertainty in mutational inference into account by incorporating mutation biases, multiple hits, and branch lengths of the phylogeny.

After adjusting for multiple testing and taking into consideration of uncertainty across multiple replicates (Methods), we identified 12 codon positions driven by positive selection in embryonated eggs (Figure 1C). Interestingly, most of these codons are found to be significant by both tests (p-value <= 0.01). This suggests that passage adaptation tends to be both highly convergent and enriched in embryonated eggs. Across all these 12 codons, the total ratio of nonsynonymous to synonymous changes along all the egg terminal branches is 200 to 7. The high nonsynonymous to synonymous ratio indicates exceptionally strong positive selection compared to many other instances of previously reported adaptive evolution detected in natural systems [28]. Of all substitutions across the 12 codons, amino acid positions 156, 186 and 194 contribute most of the nonsynonymous changes (Figure 1C). By correlating these 12 identified codon positions with functional domains (e.g. antigenic cluster A-E or receptor binding sites) on the hemagglutinin gene, we observe a strong overlap between antigenic sites (especially antigenic cluster B [9]) and the 12 codons. This suggests that receptor binding during viral infection or host cell immune response may drive passage adaptation in embryonated eggs.

Adaptive distance and vaccine efficacy

As these 12 identified codons harbor mostly nonsynonymous substitutions, the classical d_N/d_S metric will give infinite estimates, hence making it ineffective in measuring the strength of passage adaptation. For highly adaptive substitutions that are strongly selected for in embryonated eggs, they will tend to be enriched in egg-passaged isolates while uncommon in the non-egg isolates. To measure the extent of adaptive convergent evolution for a specific amino acid at a given codon, we first calculate the proportion of strains carrying the amino acid at the focal codon in egg-passaged isolates (denoted as p_{egg} in 661 egg strains) and in all isolates (denoted as p_{total} in all 32278 sequences) respectively. We then define the enrichment score (ES) for that allele (e.g. amino acid V at position 186) as the ratio of p_{egg} to p_{total}. Higher ESs indicate strong adaptive convergent evolution driving the selection of certain alleles (i.e. amino acid) in embryonated eggs. By plotting allelic ESs across all the 12 identified codons, we found that a few alleles at certain codons are specific to egg passaged isolates and have very high enrichment scores (Figure 1D).

For each isolate in the database, the allelic states of the 12 codons are used to define a DNA barcode, with each allele having an associated ES that measures the strength of passage adaptation at that position. In order to understand the multi-dimensional distribution of the adaptive evolution, we projected the 12-dimensional enrichment scores for all existing sequences onto a two-dimensional space using PCA (Figure 2A). In the PCA map, each isolate is a data point projected from the 12-dimensional space. Interestingly, we observe a major cluster harboring most of the isolates at one corner of the PCA map. Scattered islands of sequences were

also found along the PC1 and PC2 axis. Since influenza strains driven by strong adaptive evolution will carry alleles with high ESs, they will tend to deviate from the major cluster in the PCA map. By plotting the distribution of the egg-passaged isolates, we clearly show that most of the distantly located data points in Figure 2A are indeed egg-passaged sequences (Figure 2B). Clinical strains or isolates passaged in cell lines (e.g. MDCK, Supplementary Figure 2) reside in the major cluster and bear little evidence of passage adaptation (Figure 2C).

In order to measure the extent of adaptive evolution, we defined the Adaptive Distance (AD) of a target isolate as its distance from the major cluster (Figure 2D). Using a well-studied meta-analysis for vaccine efficacies [5] (2010-2015), as well as a list of curated vaccine isolates (Supplementary Table 3), we found that a large proportion of these sequences were located far away from the major cluster and were possibly driven by egg passage adaptation (Figure 2E). When comparing strains with low and high ADs, we found that sequences with larger AD tend to have both a higher number of adaptive alleles (Figure 2F) as well as alleles with stronger ESs (Figure 2G). Thus, this shows that there is a very strong correlation between AD and the level of egg passage adaptation. Interestingly, when we plot the AD for each vaccine isolate as a function of time, there is a rapid increase in AD in the year 2013-2015 follow by a slight drop in the 2015/2016 season (Figure 2H). The increase in AD suggests that the effect of passage adaptation has manifested strongly in the vaccine isolates.

In order to further explore the connection between AD and vaccine efficacy (VE), we extracted VE data compiled by Belongia *et al.* [5] (Methods) and plotted them together with the AD of the vaccine strains. Interestingly, we observed a strong negative correlation between the AD and VE (p-value = 0.037). Linear regression between these two matrices yielded a significant R-square statistic of 75.0% (Figure 2I), indicating that passage adaptation at these 12 codon positions strongly correlates to the reduction in vaccine efficacy in the past few years.

Discussion

The high correlation between VE and AD is concordant with a recent study showing that egg-adapted vaccine strains elicit specific antibodies during influenza immunization [29]. However, this also raises an interesting question - why does antigenic drift contribute so little to the reduction in vaccine efficacy? The R-square value (0.812) in the linear regression suggests that antigenic drift only contributes minimally to the total variability in vaccine efficacy. Several factors might explain the importance of passage adaptation in VE. First of all, antigenic evolution of circulating strains often follows a punctuated equilibrium model, whereby genetic changes only occasionally cause a disproportionally large shift in the antigenic space (averaging every three years) [7]. As most of the antigenic changes between years have little effect, they may not contribute strongly to the recent drop in vaccine efficacy. Secondly, the constant adaptation of the H3N2 influenza virus to the human population has enabled the H3N2 subtype

to become well adapted to the human cellular environment. This makes propagating them in an avian host difficult and often results in lower yields [30], as well as an increasing strength of passage adaptation in the embryonated eggs [18]. The resulting larger antigenic distances [12] between the egg adapted strains and original isolates may hence lead to large antigenic jumps and therefor reduced vaccine efficacy.

Since passage adaptation in embryonated eggs often results in the fixation of alleles highly specific to egg passaged sequences, the allelic status of the 12 codons identified by our study provide an important barcode for predicting the footprints of passage adaptation. Based on the principles illustrated in this work, we have developed a computational package MADE (Measuring Adaptive Distance and vaccine Efficacy based on allelic barcodes) that will enable vaccine developers to measure the strength of passage adaptation and hence predict the potential efficacy of a vaccine strain based on its nucleotide sequence [31].

It is of note that there are potential limitations in our current study. Firstly, the mutational mapping and the two statistical tests for enrichment and convergence implemented in this study are approximations of a full codon-based likelihood model. Nonetheless, it is worth pointing out that previous simulation studies have shown that statistical inference is quite accurate in cases of short term evolution such as in influenza viruses (see Supplementary Material for a discussion). Secondly, in addition to passage adaptation, some factors that may affect the measurement of VE were not able to be considered in this study, such as similarities between vaccine strains and

circulating strains[32], the time between vaccination and infection, the age group of the infected individuals, the immunogenicity variation across individuals [33], the dynamics of immune memory[34] as well as the study design for measuring VE. Moreover, although Test-Negative Design (TND) [35, 36] is a very popular approach for measuring VE, it may be affected by selection bias such as in seeking medical treatment and exposure misclassification [37]. Additionally, TND also relies on several important assumptions such as independence between vaccination and exposure or susceptibility to infection, as well as dichotomous protection (i.e. all-or-nothing) from vaccination. Lastly, the strong correlation between AD and VE is based on observations from a few years. Future studies with more extensive vaccine efficacy data will be able to further examine the relationship between passage adaptation and VE (Supplementary Figure 3).

The observations from our study suggest that strong passage adaptation in embryonated eggs will continue to impede the progress of any current H3N2 vaccine program. One solution may be to switch vaccine production to a cell based system. A natural question will be: will passage adaptation also occur in cell lines? Interestingly, when an analogous analysis on the MDCK medium was conducted, we observe that the allelic enrichment scores for MDCK-passaged isolates are much lower than those of egg-passaged isolates, and that the adaptive distances are negligible in most cases and are not correlated with vaccine efficacy. A cell-based vaccine production program and possibly other effective strategies are urgently needed to help increase the global protection against influenza [38].

Author contributions

W.Z conceived and supervised this work. H.C performed the analysis. J.J.S.A, S.H.N, R.N participated in the discussions and data interpretation. H.C, J.J.S.A, S.H.N, R.N, W.Z wrote the paper.

Acknowledgements

We would like to thank Sebastian Maurer-Stroh, Jianjun Liu, Chung-I Wu, Paola Florez DE SESSIONS, Swaine Chen, for helpful discussions. We want to thank Cheryl Zi Jin Phua and Chu Quan Tan for the help with the manuscript.

Funding

W.Z is supported in part by National Key R&D program of China grant 2018YFC1406902 and 2018YFC0910400.

Potential conflicts of interest

The authors declare no conflict of interests.

References

1. Nelson MI, Holmes EC. The evolution of epidemic influenza. Nat Rev Genet **2007**; 8(3): 196-205.

2. Stohr K. Influenza--WHO cares. Lancet Infect Dis **2002**; 2(9): 517.

3. Fiore AE, Uyeki TM, Broder K, et al. Prevention and control of influenza with vaccines: recommendations of the Advisory Committee on Immunization Practices (ACIP), 2010: Department of Health and Human Services, Centers for Disease Control and Prevention, **2010**.

4. Osterholm MT, Kelley NS, Sommer A, Belongia EA. Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. Lancet Infect Dis **2012**; 12(1): 36-44.

5. Belongia EA, Simpson MD, King JP, et al. Variable influenza vaccine effectiveness by subtype: a systematic review and meta-analysis of test-negative design studies. Lancet Infect Dis **2016**; 16(8): 942-51.

6. Belongia EA, Kieke BA, Donahue JG, et al. Effectiveness of inactivated influenza vaccines varied substantially with antigenic match from the 2004-2005 season to the 2006-2007 season. J Infect Dis **2009**; 199(2): 159-67.

7. Smith DJ, Lapedes AS, de Jong JC, et al. Mapping the antigenic and genetic evolution of influenza virus. Science **2004**; 305(5682): 371-6.

8. Chambers BS, Parkhouse K, Ross TM, Alby K, Hensley SE. Identification of Hemagglutinin Residues Responsible for H3N2 Antigenic Drift during the 2014-2015 Influenza Season. Cell Rep **2015**; 12(1): 1-6.

9. Skowronski DM, Janjua NZ, De Serres G, et al. Low 2012-13 influenza vaccine effectiveness associated with mutation in the egg-adapted H3N2 vaccine strain not antigenic drift in circulating viruses. PLoS One **2014**; 9(3): e92153.

10. Boni MF. Vaccination and antigenic drift in influenza. Vaccine **2008**; 26 Suppl 3: C8-14.

11. Carrat F, Flahault A. Influenza vaccine: the challenge of antigenic drift. Vaccine **2007**; 25(39): 6852-62.

12. Robertson J. Clinical influenza virus and the embryonated hen's egg. Reviews in Medical Virology **1993**; 3(2): 97-106.

13. Skowronski DM, Sabaiduc S, Chambers C, et al. Mutations acquired during cell culture isolation may affect antigenic characterisation of influenza A(H3N2) clade 3C.2a viruses. Euro Surveill **2016**; 21(3): 30112.

14. Burnet F. Growth of influenza virus in the allantoic cavity of the chick embryo. Aust J Exp Biol Med Sci **1941**; 19: 291-5.

15. Madin S, Darby N. Established kidney cell lines of normal adult bovine and ovine origin. Experimental Biology and Medicine **1958**; 98(3): 574-6.

16. Rogers GN, Pritchett TJ, Lane JL, Paulson JC. Differential sensitivity of human, avian, and equine influenza A viruses to a glycoprotein inhibitor of infection: selection of receptor specific variants. Virology **1983**; 131(2): 394-408.

17. Bush RM, Smith CB, Cox NJ, Fitch WM. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. Proc Natl Acad Sci U S A **2000**; 97(13): 6974-80.

18. Chen H, Deng Q, Ng SH, Lee RT, Maurer-Stroh S, Zhai W. Dynamic Convergent Evolution Drives the Passage Adaptation across 48 Years' History of H3N2 Influenza Evolution. Mol Biol Evol **2016**; 33(12): 3133-43.

19. Zost SJ PK, Gumina ME, Kim K, Diaz Perez S, Wilson PC, Treanor JJ, Sant AJ, Cobey S, Hensley SE. Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains. Proc Natl Acad Sci U S A **2017**.

20. Nielsen R. Mapping mutations on phylogenies. Syst Biol **2002**; 51(5): 729-39.

21. Zhai W, Slatkin M, Nielsen R. Exploring variation in the d(N)/d(S) ratio among sites and lineages using mutational mappings: applications to the influenza virus. J Mol Evol **2007**; 65(3): 340-8.

22. Bogner P, Capua I, Lipman DJ, Cox NJ. A global initiative on sharing avian flu data. Nature **2006**; 442(7106): 981-.

23. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research **2004**; 32(5): 1792-7.

24. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **2006**; 22(21): 2688-90.

25. Beer DG, Kardia SLR, Huang C, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med **2002**; 8: 816-24.

26. Squires RB, Noronha J, Hunt V, et al. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza and other respiratory viruses **2012**; 6(6): 404-16.

27. McWhite CD, Meyer AG, Wilke CO. Sequence amplification via cell passaging creates spurious signals of positive adaptation in influenza virus H3N2 hemagglutinin. Virus Evol **2016**; 2(2).

28. Eyre-Walker A. The genomic rate of adaptive evolution. Trends in Ecology & Evolution **2006**; 21(10): 569-75.

29. Raymond DD, Stewart SM, Lee J, et al. Influenza immunization elicits antibodies specific for an egg-adapted vaccine strain. Nat Med **2016**; 22(12): 1465-9.

30. Lu B, Zhou H, Chan W, Kemble G, Jin H. Single amino acid substitutions in the hemagglutinin of influenza A/Singapore/21/04 (H3N2) increase virus growth in embryonated chicken eggs. Vaccine **2006**; 24(44-46): 6691-3.

31. Chen Hui, Zhai W. MADE: Measuring Adaptive Distance and Vaccine Efficacy based on Allelic Barcodes (Manuscript in preparation).

32. Bonomo ME, Deem MW. Predicting Influenza H3N2 Vaccine Efficacy from Evolution of the Dominant Epitope. Clin Infect Dis **2018**.

33. Cobey S, Gouma S, Parkhouse K, et al. Poor Immunogenicity, Not Vaccine Strain Egg Adaptation, May Explain the Low H3N2 Influenza Vaccine Effectiveness in 2012-2013. Clin Infect Dis **2018**; 67(3): 327-33.

34. Lewnard JA, Cobey S. Immune History and Influenza Vaccine Effectiveness. Vaccines (Basel) **2018**; 6(2).

35. Skowronski D, Gilbert M, Tweed S, Petric M, Li Y, Mak A. Effectiveness of vaccine against medical consultation due to laboratory-confirmed influenza: results from a sentinel physician pilot project in British Columbia, 2004–2005. Can Commun Dis Rep **2005**; 31(18): 181-91.

36. Jackson ML, Nelson JC. The test-negative design for estimating influenza vaccine effectiveness. Vaccine **2013**; 31(17): 2165-8.

37. Joseph A Lewnard CAT, Benjamin J Cowling, Marc Lipsitch. Quantifying biases in test-negative studies of vaccine effectiveness. bioRxiv **2017**.

38. Hanon E, Van der Most R, Del Giudice G, Rappuoli R. Short-term and mid-term solutions for influenza vaccines. Lancet Infect Dis **2018**; 18(8): 832-3.

Figure legends

Figure 1. Mutational mapping and dynamic convergent evolution. A) The phylogenetic tree of the 32278 HA1 sequences retrieved from the GISAID database [22]. Egg and MDCK passage isolates are labeled on the phylogenetic tree. B) Principles of the convergent and enrichment test (also see Methods). The enrichment test asks whether substitutions at a given codon tend to be enriched along terminal branches for the egg-passaged strains (n=661) relative to other terminal branches (n=31617) while the convergent test examines whether certain nonsynonymous substitutions are observed more frequently than expected by chance along egg terminal branches.

C) Numbers of nonsynonymous and synonymous changes at the 12 positively selected codons. Significance of the two tests (p-value<=0.01 are labelled) for each codon as well as the functional category of the codons are indicated as a heatmap beneath the barplot. D) The enrichment scores across 12 codons. Enrichment scores higher than 15 are labeled.

Figure 2. Adaptive distance and vaccine efficacy. A) Principal component plot of the 12-dimensional enrichment scores across all isolates (n=32278) (PCA map). Each strain is a dot in the PCA map. The clusters of the strains are plotted as the heatmap in the PCA plot. B) PCA plot of the egg passaged strains (n=661). C) PCA plot of the clinical strains (n=4046). D) Adaptive Distance (AD) measures the distance between the center of the major cluster and the strain of interest. E) PCA plot of the vaccine strains between year 2010-2016 (n=61). F) The number of adaptive alleles (defined as enrichment scores >= 2) for strains with low and high adaptive distances. G) The sum of enrichment scores across 12 codon positions for strains with low and high adaptive distances. H) Adaptive distances for the vaccine strains between year 2010-2016. Vaccine efficacy (VE) data is plotted using the y axis on the right hand side. I) Linear regression between vaccine efficacy and adaptive distance. R-square is measured to be 0.750 (p-value=0.037).

Figure 1

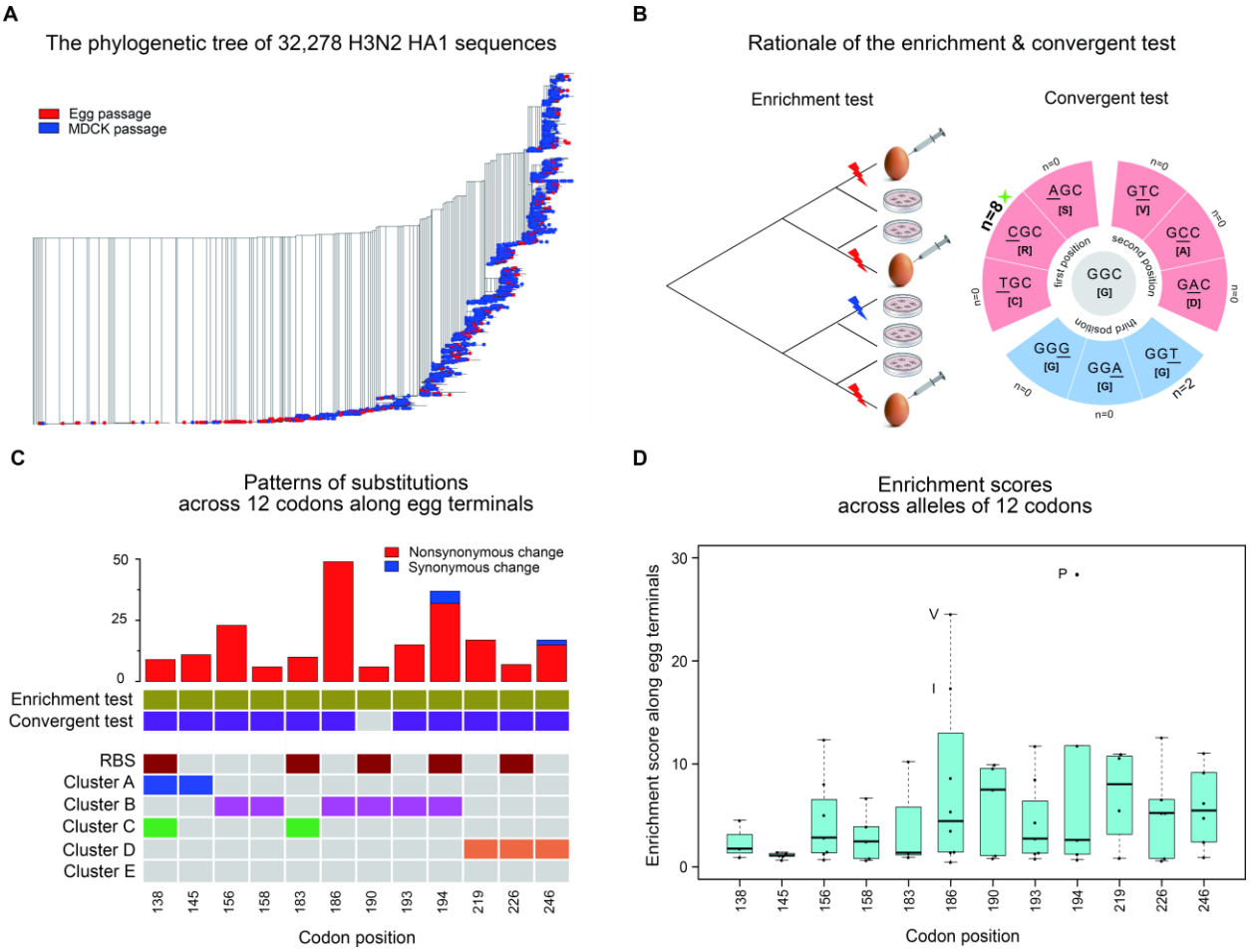

Figure 2