Dynamic Convergent Evolution Drives the Passage Adaptation across 48 Years' History of H3N2 Influenza Evolution

Hui Chen,[1] Qiang Deng,[1,2,3] Sock Hoon Ng,[4] Raphael Tze Chuen Lee,[5] Sebastian Maurer-Stroh,[5,6,7,8] and Weiwei Zhai*[,1]

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

[2]Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[3]University of Chinese Academy of Sciences, Beijing, China

[4]DSO National Laboratories, Singapore, Singapore

[5]Bioinformatics Institute, A*STAR, Singapore

[6]School of Biological Sciences (SBS), Nanyang Technological University (NTU), Singapore

[7]National Public Health Laboratory (NPHL), Ministry of Health (MOH), Singapore

[8]Department of Biological Sciences, National University of Singapore (NUS), Singapore

***Corresponding author:** E-mail: zhaiww1@gis.a-star.edu.sg.

Associate editor: Thomas Leitner

Abstract

Influenza viruses are often propagated in a diverse set of culturing media and additional substitutions known as passage adaptation can cause extra evolution in the target strain, leading to ineffective vaccines. Using 25,482 H3N2 HA1 sequences curated from Global Initiative on Sharing All Influenza Data and National Center for Biotechnology Information databases, we found that passage adaptation is a very dynamic process that changes over time and evolves in a seesaw like pattern. After crossing the species boundary from bird to human in 1968, the influenza H3N2 virus evolves to be better adapted to the human environment and passaging them in embryonated eggs (i.e., an avian environment) leads to increasingly stronger positive selection. On the contrary, passage adaptation to the mammalian cell lines changes from positive selection to negative selection. Using two statistical tests, we identified 19 codon positions around the receptor binding domain strongly contributing to passage adaptation in the embryonated egg. These sites show strong convergent evolution and overlap extensively with positively selected sites identified in humans, suggesting that passage adaptation can confound many of the earlier studies on influenza evolution. Interestingly, passage adaptation in recent years seems to target a few codon positions in antigenic surface epitopes, which makes it difficult to produce antigenically unaltered vaccines using embryonic eggs. Our study outlines another interesting scenario whereby both convergent and adaptive evolution are working in synchrony driving viral adaptation. Future studies from sequence analysis to vaccine production need to take careful consideration of passage adaptation.

Key words: **passage adaptation, host-mediated changes, adaptive evolution, convergent evolution, influenza H3N2, embryonated egg, vaccine production, mutational mapping.**

Article

Introduction

The ability to detect and identify a virus is often the first step for regular surveillance of circulating pathogenic strains. Following viral identification, proper isolation and propagation of viral particles play a very crucial role for subsequent procedures including vaccine production (Hilleman 2000). For many years, researchers have noticed that additional adaptation might happen during this culturing process (e.g., in Influenza A virus, Burnet and Bull 1943). Host-mediated changes due to medium propagation is a widely observed phenomena and have been documented across a diverse set of viruses including HIV (Sawyer et al. 1994), influenza B (Gatherer 2010), Japanese encephalitis virus (Cao et al. 1995), hepatitis A (Graff et al. 1994), hepatitis C (Lohmann et al. 2001), hepatitis E (Lorenzo et al. 2008) and Sendai virus (Itoh et al. 1997).

Propagation of influenza virus is carried out via two main methods, through the use of embryonated eggs (often denoted as the egg culture) (Burnet 1940, 1941) and mammalian cell lines including Madin–Darby canine kidney cells (MDCK) (Madin and Darby 1958; Schepetiuk and Kok 1993), Vero cells (Govorkova et al. 1995) and many other cell types (Govorkova et al. 1999; Seo et al. 2001). Sialic acids, sugars terminally linked to different carbohydrates (e.g., galactose) on cell surfaces, are the host receptor for the influenza virus. The linkage between sialic acid and galactose takes two different forms, namely alpha-2,3 glycosidic bond (SAα-2,3Gal, found mostly in intestinal tracts of birds) or alpha-2,6

glycosidic bond (SAα-2,6Gal, mostly in human upper respiratory tract). Human and avian viruses favor different types of sialic linkage from their respective natural hosts (Rogers et al. 1983).

Passage adaptation can have an important impact on influenza research. First of all, earlier studies have found that many of the sites responsible for passage adaptation often overlap with sites that are adaptively evolving in humans and passage adaptation can potentially confound the inference of adaptive evolution in humans (Bush et al. 2000). Secondly, additional substitutions in the culturing medium can lead to differences between the original and passaged isolate, which can affect the efficacy of vaccines. Ineffective vaccines for the dominant strains of the year can cause serious medical and public health problems (Kodihalli et al. 1995; Robertson et al. 1995; Saito et al. 2004; Nakowitsch et al. 2014). Understanding the magnitude and features of passage adaptation is an important first step for proper vaccine production and evolutionary inference.

Even though many studies have been conducted on passage substitutions based on sequencing clinical strains before and after virus propagation (Katz et al. 1990), they each drew a snapshot of the landscape at a given time in history. Due to constant selection in human populations, the influenza virus will evolve to be more adapted to the human environment (Su et al. 2015). We hypothesize that passage adaptation in the early and late phase of viral evolution can be quite different. By jointly analyzing viral strains from multiple time points, we want to draw a systematic picture of passage adaptation across 48 years of influenza evolution (1968–2015).

Results

Sequence Collection, Phylogenetic Inference and the Passage History

We collected sequences and passage information for the influenza H3N2 HA1 sequence from two major databases (Materials and Methods). The first data set was gathered from the Genbank database at the National Center for Biotechnology Information (NCBI) and the second data set was retrieved from the EpiFlu database from the Global Initiative on Sharing All Influenza Data (GISAID) (supplementary tables S1 and S2, Supplementary Material online). After quality control (Materials and Methods), 25,482 sequences (Materials and Methods) were retained for the subsequent analysis. Maximum likelihood procedure implemented in the Randomized Axelerated Maximum Likelihood (RAxML) package (Stamatakis 2006) was employed to estimate the phylogenetic relationship for all the sequences (fig. 1*a*). The shape of the phylogenetic tree shows a typical stair or cactus like pattern where sequences from each year are consecutively connected by a truncal lineage (fig. 1*a* and supplementary fig. S1, Supplementary Material online). This unique evolutionary tree shape driven by sequential selective sweeps has been suggested to indicate strong positive selection in the history of influenza evolution (Fitch et al. 1991). Plotting the passage histories of all strains showed that influenza virus propagation using embryonated eggs was performed across the years with two major peaks before and after year 2000 (fig. 1*a* inset and supplementary fig. S2, Supplementary Material online). Virus propagation using MDCK cell lines increased rapidly around year 2000 (fig. 1*a* inset and supplementary fig. S2, Supplementary Material online).

Mutational Mapping, A Versatile Approach to Study Adaptive Evolution across Site and Branches

Since passage related adaptation tends to add extra evolution to the tips of the phylogeny, we want to focus on the substitutions along the terminal branches for strains with different culturing conditions. A rigorous probabilistic approach to infer the history of evolutionary changes along a phylogeny is the mutational mapping method (Nielsen 2002; Bollback 2006; Zhai et al. 2007). Using properties of the continuous time Markov Chain, we can sample possible evolutionary histories according to their posterior probabilities (Materials and Methods). The evolutionary history includes both the state of all the internal nodes as well as the changes along each branch. With the full evolutionary history, we can study mutations specific to terminal branches where passage adaptation is likely to occur (Bush et al. 2000; Zhai et al. 2007).

By sequencing clinical strains before and after medium propagation, researchers have empirically observed 22 codon sites as major targets of passage adaptation (often denoted as host-mediated sites or HM sites) (Nakajima et al. 1983; Robertson 1993; Rocha et al. 1993; Gubareva et al. 1994; Hardy et al. 1995; Bush et al. 1999). In order to study differences in passage adaptation across history, we partition the evolutionary histories into two time periods. Strains collected before year 2000 reflect the early phase of flu evolution and strains gathered after year 2000 correspond to recent adaptation. The number of egg passaged strains is relatively balanced between these two sets (supplementary table S3, Supplementary Material online). We hypothesize that, as the influenza virus adapts to the human environment, the passage adaptation to the culturing medium will also change accordingly.

Strong Adaptation in the Early History

When inspecting the evolutionary changes along the terminal branches for the HM sites (the 22 codons) as well as other sites (denoted as non-HM sites) during the early period (before year 2000), we noticed a much elevated ratio of nonsynonymous to synonymous changes (denoted as A/S ratio) at the HM sites as compared with the non-HM sites (348/124 vs. 811/887, $P < 0.001$, fig. 1*b*). Under neutral evolution, the expected A/S ratio is often between 2 and 3 depending on the codon structure of the sequences (Li et al. 1985). The A/S ratio in the HM sites suggests that, positive selection or relaxation of purifying selection is potentially acting on the HM sites along the terminal lineages.

Since passage adaptation tends to reside on the tip branches, internal branches represent the evolution of the influenza virus within human populations. When inspecting the nonsynonymous to synonymous changes along the internal branches for these 22 codons, we also noticed a much reduced A/S ratio for the HM sites along the internal branches (232/145 vs. 348/124, $P = 0.0002$). This suggests

that, passage adaptation can be different from the evolution of the influenza virus in human populations.

Given that terminal branches include all the different passage histories and that there might be differences in substitution patterns across different culturing mediums, we further classify terminal branches into those that were passaged in chicken embryos (denoted as the egg terminal) and those in MDCK cell lineages (denoted as the MDCK terminal, not including MDCK-SIAT, Materials and Methods). Within these two groups, we observed a much elevated A/S ratio (112/28, fig. 1*b*) along egg lineages as compared with those found in MDCK cell lines (66/19, fig. 1*b*). Using the binomial distribution to test for excess of nonsynonymous changes (Materials and Methods), the magnitude of nonsynonymous changes along the egg terminals are indeed more than expected by chance ($P = 0.001$). Applying the same test to the MDCK terminals, the calculated P value is also marginally significant at 0.043. In other words, there is statistical evidence that positive selection is acting along both the MDCK and egg terminals and the effect is stronger along the egg terminals.

Inspecting the A/S ratio at each of the 22 codons along the egg and MDCK terminals as well as the internal branches (fig. 1*c*), we observe that, whereas a large proportion (86%) of the sites have substitutions on the same set of sites, the substitution profiles at these sites are different between two culturing media as well as between terminal and internal branches (statistically significant for all pairwise test, fig. 1*c*). This suggests that, the adaptation to the two culturing media is fine-tuning different amino acid residues and passage adaptation is quite different from the evolution in the human population during the early phase of influenza evolution.

Dynamic Shift in Adaptation between the Early and Late Phase of H3N2 Evolution

When summarizing the mutation profile for the late time period (fig. 1*d*, post-year 2000), we observe that, the A/S ratio along the terminal branches is much reduced compared with the early period (699/487 vs. 348/124, $P = 0$). This suggests that the overall strength of positive selection at the HM sites was much attenuated in the later time period. Grouping the branches into the egg and MDCK terminal lineages, we found that the A/S ratio along the MDCK terminal is significantly reduced compared with the early phase (86/91 vs. 66/19, $P < 0.001$, fig. 1*b* and *d*), whereas the corresponding value along the egg terminals are much elevated (82/3 vs. 112/28, $P = 0.001$, fig. 1*b* and *d*).

It is very interesting to observe that the mutations at egg terminals at later time points (after 2000) are much more constrained at a few amino acid positions, rather than very scattered across many sites at earlier time points ($P < 0.001$, fig. 1*e* vs. *c*). Substitutions at codon position 186, 194 as well as 219 are the major contributing sites adapting to the avian environment in the later time period. It should be noted that these codon sites are prominently located in antigenic surface epitopes, which means that egg adaptation on these sites may alter antigenicity. Indeed, egg adaptation of these sites is suspected to be behind lower vaccine efficacy in recent seasonal H3N2 influenza seasons (Skowronski et al. 2014). Data from post-year 2000 indicate that the substitution profile for viruses propagated in MDCK cells is similar to the changes along the internal branches but is different from viruses passaged in embryonated eggs (fig. 1*e*). This is in contrast to the pattern observed from the data prior to year 2000, whereby substitutions along egg terminal, MDCK terminal and internal branches are all statistically different (fig. 1*c*).

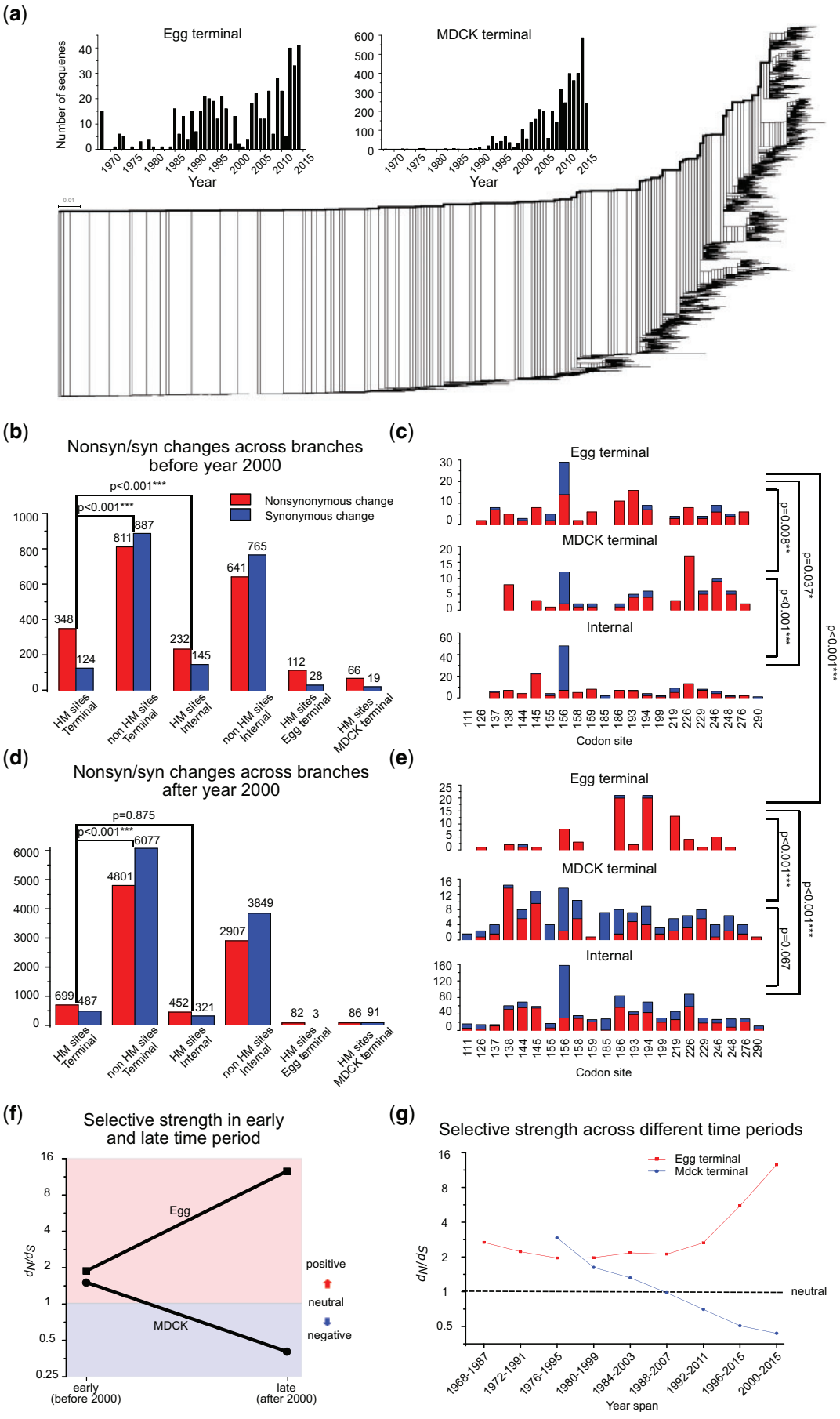
Comparing strains before and after year 2000, we revealed a dynamic history of passage adaptation in the two culturing conditions (fig. 1*f*). After crossing species boundary in 1968, the influenza H3N2 virus evolves to be fitter to the human environment and growing them in an avian medium will lead to stronger adaptation in these strains. On the other hand, the cellular environment in the MDCK cell (a mammalian cell line) resembles the conditions in the human population more than eggs and passage adaptation changes from positive selection to largely purifying selection (fig. 1*f*). When performing a sliding window analysis along the history by breaking the data set into discrete time windows, the pattern stays very similar (fig. 1*g*). Positive selection in the embryonated egg is found to be gradually increasing, whereas the opposite is true for the MDCK medium.

De Novo Identification of Codon Positions Driven by Embryonated Egg Adaptation

The 22 HM sites were curated from a few empirical studies comparing the clinical strains before and after medium propagation (Nakajima et al. 1983; Robertson 1993; Rocha et al. 1993; Gubareva et al. 1994; Hardy et al. 1995; Bush et al. 1999). There are a few potential limitations of this earlier curation. First of all, the sites responsible for the medium adaptation might not be a set of fixed positions and can be changing along the history of influenza evolution (fig. 1*c* and *e*). There could be new sites that can contribute to the medium adaptation, but are not in the set collected using strains from the earlier time. Secondly, there are sites for which their substitutions are not responsible for medium adaptation, but happened simply due to genetic drift or hitchhiking in earlier experiments. Thus, we expect that not all of the sites from the earlier list are strongly contributing to the medium adaptation. For example, in figure 1*c* and *e*, several of the amino acid positions (e.g., site 199, 290) have no substitutions along the egg terminals.

Since substitution in mammalian cell lines is becoming similar to the natural evolution in human represented by the evolution along the internal branches (fig. 1*e*), we focus on the evolution taking place in embryonated eggs. We developed two statistical tests capturing different aspects of the mutational pattern in medium adaptation. The first test (denoted as the enrichment test) sets out to examine whether mutations at a given codon position are enriched in the egg terminals. Applying this test to all the egg terminal branches, we identified 16 codon positions for which associated nonsynonymous changes are enriched in the egg terminals ($P < 0.01$) (fig. 2*a*).

In addition to quantifying the amount of substitutions specific to the egg terminals in the first test, the second statistical test (denoted as the convergent test) focuses on the


(a)
Egg terminal
MDCK terminal
Number of sequenes
Year
(b)
Nonsyn/syn changes across branches before year 2000
p<0.001***
Nonsynonymous change
Synonymous change
348
124
811
887
232
145
641
765
112
28
66
19
HM sites Terminal
non HM sites Terminal
HM sites Internal
non HM sites Internal
HM sites Egg terminal
HM sites MDCK terminal
(c)
Egg terminal
MDCK terminal
Internal
p=0.008**
p=0.037*
p<0.001***
p<0.001***
Codon site
(d)
Nonsyn/syn changes across branches after year 2000
p=0.875
p<0.001***
699
487
4801
6077
452
321
2907
3849
82
3
86
91
(e)
p<0.001***
p=0.067
(f)
Selective strength in early and late time period
dN/dS
Egg
MDCK
positive
neutral
negative
early (before 2000)
late (after 2000)
(g)
Selective strength across different time periods
Egg terminal
Mdck terminal
neutral
Year span

specific changes along terminal branches. Since subjecting the influenza virus to the same egg environment can lead to convergent amino acid changes, the second test investigates whether there are multiple identical codon transitions across egg lineages. Using a binomial based test (Materials and Methods), we identified 10 codon sites which show significant evidence of convergent changes ($P < 0.01$). Combining results from these two tests, we identified 19 codons in total that are strongly driving the passage adaptation in embryonated eggs. Among these codons, 229 nonsynonymous changes and 25 synonymous changes occurred in 508 egg terminals, indicating extremely strong passage adaptation. Interestingly, 11 of the 19 identified HM codons overlap with 22 previously published sites. It is important to emphasize that the rest of the 11 HM sites not identified in this list have only 31 nonsynonymous and 7 synonymous changes in total, and are not strongly contributing to the host-mediated adaptation (fig. 1*c* and *e*).

The Functional and Structural Properties of Sites Responsible for Egg-Mediated Adaptation

When inspecting the functional annotation of these 19 sites, antigenic sites B and D overlap most with this set. Interestingly, the overlap between these 19 codons and the receptor binding sites (defined as those in contact with the sialic acid structurally, Wilson et al. 1981) is relatively mild. If most of the passage adaptation is due to a shift in receptor conformation, this mild overlap suggests that fine tuning of receptor binding might be achieved through sites that are not directly binding to sialic acid. When we map the 19 sites on the structure of the hemagglutinin (fig. 2*b*), we observe that all the 19 codon sites are indeed surrounding the receptor binding pocket, suggesting that functional fine tuning of receptor binding is a strong determining factor driving passage adaptation.

From five previously published studies looking for positively selected codons, we compiled a list of 39 positively selected codons for influenza during their evolution in human populations (Fitch et al. 1997; Bush et al. 1999; Suzuki and Gojobori 1999; Plotkin and Dushoff 2003; Pond et al. 2008) (supplementary table S4, Supplementary Material online). When checking these 39 sites against the 19 sites identified in this study, many codon sites overlap between these two sets (fig. 2*a*). This suggests that, many of the previously identified positively selected sites can be potentially confounded by the egg adaptation (Bush et al. 2000; Zhai et al. 2007).

In order to quantify the extent of substitutions due to egg passage, we calibrated the percentage of nonsynonymous changes contributed by the egg passage to the total number of nonsynonymous changes at each codon (Materials and Methods). Interestingly, 13 out of 19 sites have >10% of the total nonsynonymous changes contributed by egg terminals (fig. 2*c* and supplementary fig. S3, Supplementary Material online). If we calibrate convergent changes (defined as those codon transitions that appeared at least 3 times, Materials and Methods) at these 19 codons, convergent changes contribute a large proportion of the total nonsynonymous changes happened along egg terminals (fig. 2*d*). In summary, the same passage environment was driving very strong convergent adaptive evolution at these codons. Inference of signals of adaptation in influenza virus can be significantly confounded by passage adaptation.

The Proportion of Egg Passaged Isolates in the Public Databases

Since its creation in 2006 (Bogner et al. 2006), GISAID continually synchronizes the sequences from NCBI. In other words, most of the sequences before 2006 are shared between the two databases and GISAID has more unique sequences post-2006 (Materials and Methods). In GISAID, we observe a significant change in proportion of strains passaged in embryonated eggs. For example, of 1,826 sequences before year 2000 in GISAID, 234 (12.8%) sequences were passaged in the embryonated eggs. However, after year 2000, the number of available sequences increased dramatically and the proportion of sequences from egg passaged viruses dropped to be 1.1% (fig. 3*a*). The large proportion of "unknown" strains (sequences without any passage information) and the drop in the observed egg passaged influenza virus suggest a possibility that many of the egg passaged strains are not explicitly labeled in the database, and are in the unknown category.

Fig. 1. Nonsynonymous and synonymous changes across different branches and sites. (*a*) The maximum likelihood tree of all the sequences. The evolutionary relationship shows a typical stair/cactus like shape with sequences from same years clustered together (Fitch et al. 1991), and there is a truncal lineage connecting strains from different years (highlighted in bold, also see supplementary fig. S1, Supplementary Material online). The two inset figures show the year distribution of the isolates passaged in the embryonated egg as well as MDCK cell lines in the NCBI and GISAID combined data set. (*b*) Number of nonsynonymous and synonymous changes along different branches and sites for the early time period (before year 2000). The sites are classified as HM sites (previously identified 22 codon sites) and other sites (non-HM sites). The branches are categorized into terminal and internal branches. Within terminal branches, they were further categorized into those that were passaged in embryonated eggs (egg terminals) and in MDCK cells (MDCK terminals). (*c*) Number of nonsynonymous and synonymous changes along different branches for all the 22 sites during the early time period. Top panel is for egg terminals, middle panel is for MDCK terminals and bottom panel is for internal branches. Statistical test of mutational patterns between different panels are marked on the right hand-side of the figure. (*d*) Number of nonsynonymous and synonymous changes along different branches and sites for the later time period (after year 2000). The site and branch classes are the same as panel (*b*). (*e*) Number of nonsynonymous and synonymous changes along different branches for all the 22 sites during the later time period. The panels are the same as in (*c*). (*f*) Cartoon illustration of level of natural selection for the two culturing mediums along different time points. There is a seesaw like pattern between the egg and MDCK medium. (*g*) A sliding window of d_N/d_S ratios for the two types of culturing mediums along the history of the influenza evolution. For each time window, the d_N/d_S ratio is calculated for isolates within that time span.

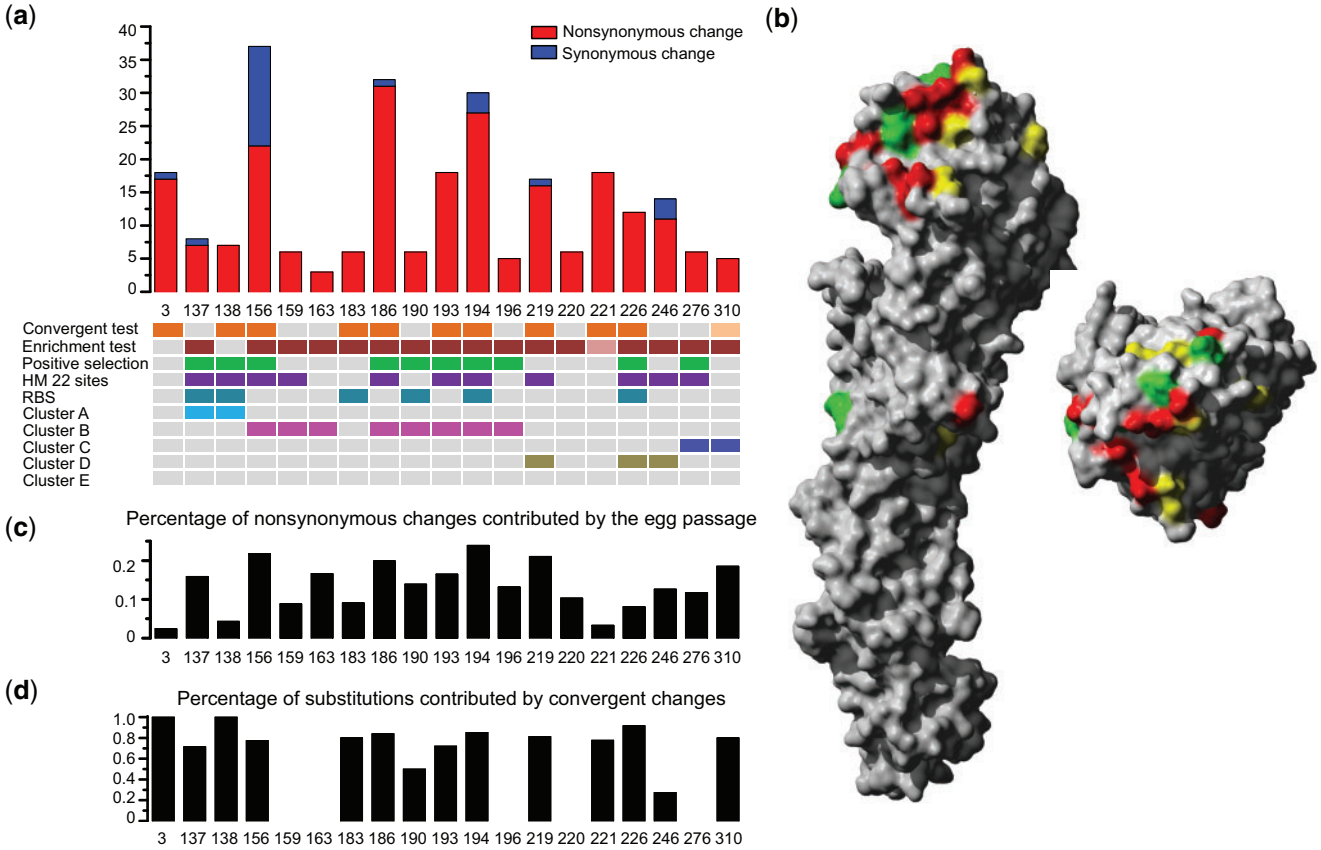


Fig. 2. Newly identified codons responsible for passage adaptation in the embryonated egg. (*a*) Number of nonsynonymous and synonymous changes at the newly identified 19 codons along the egg terminal branches in the GISAID/NCBI combined data set. The statistical evidences from the convergent and enrichment test are marked as a heatmap underneath the barplot. In the heatmap, dark colors label the codons whose *P* value is < 0.01. Light colors label the codons whose *P* value is between 0.01 and 0.05. Positively selected codon sites (curated from the literature), previously identified 22 codons, receptor binding sites (RBS) as well as antigenic sites A–E are also labeled. (*b*) Structural view of the codons. The red residues are those that overlap between the newly identified 19 codons and previously identified 22 codons. Yellow residues are those unique to the previously identified 22 codons. Green residues are those unique to the newly identified 19 codons. (*c*) The percentage of nonsynonymous substitutions at each codon contributed by egg passage. (*d*) The percentage of nonsynonymous substitutions contributed by convergent changes during egg passage at each codon position.

In order to test this hypothesis, we first tabulate the number of nonsynonymous to synonymous changes along strains passaged in different media from 2000 to 2015. In order to target the substitution pattern pertaining only to passage adaptation, we restrict the analysis to the 19 codon sites we identified earlier. As we see in figure 3*b*, the A/S ratio is very high in egg terminal branches at these 19 sites for the GISAID data set. Interestingly, the cell line lineages (CELL, MDCK) show similar A/S ratio and is in line with the earlier observation that mammalian cell lines impose weaker selection during viral passage.

Following our earlier intuition about strains with unknown passage history, the A/S ratio along the unknown lineages is much higher than in cell lines, suggesting that there is a significant proportion of strains that is passaged in embryonated eggs. Tracing back to the original publications and plotting the A/S ratio for the top 10 studies with the largest number of unknown strains, the A/S in the unknown terminals is largely driven by a few studies (supplementary fig. S4, Supplementary Material online). Using a well calibrated mixture model (Materials and Methods, supplementary tables S5 and S6 and fig. S5, Supplementary Material online), we estimated the proportion of egg passage in the unknown strains from the GISAID data set to be about 9% around year 2000 and drastically drop to be almost zero in year 2009 (fig. 3*c*). This suggests that, virus propagation using embryonated eggs is rapidly being phased out.

When we combine all the available information, we observe that the sequences from egg passaged viruses in GISAID peak around 2004 (fig. 3*d*). Overall, the percentage of egg passaged strains (actual and inferred from the unknown strains across all 48 years) in GISAID is found to be $\sim$3%. Given sequences with unknown passage information in the database, future analyses need to take into account that, many of the strains with unknown passage histories might still be grown in various culturing medium including embryonated eggs.

Discussion

Mutational mapping provides a very versatile method exploring mutations along different branches and sites. Compared with previous approaches such as parsimony based methods

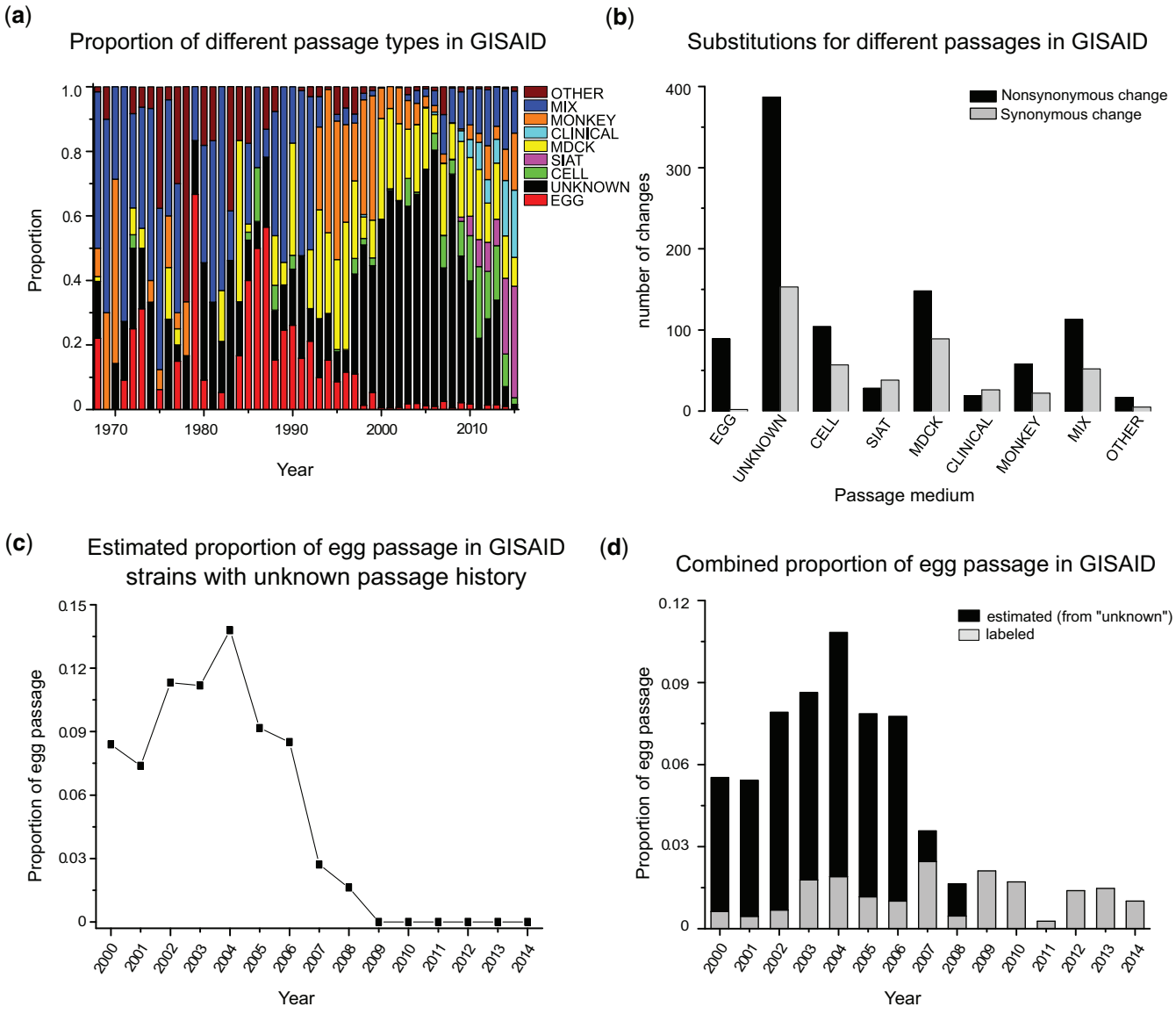


Fig. 3. Passage types of the GISAID data set. (*a*) The proportions of different passage types in the GISAID data set. (*b*) The nonsynonymous and synonymous substitutions along different terminal branches for the GISAID data set. (*c*) The estimated proportion of egg passage within the strains with unknown passage history from year 2000 to 2014 in GISAID. (*d*) Combined estimates of egg passage in the GISAID data set.

(Swofford and Maddison 1987), mutational mapping allows a better accommodation of uncertainties with the mutational history (Nielsen 2002; Zhai et al. 2007). The pattern we present here is based on one sampled history from the posterior distribution. Multiple other replicates are also checked and the results stay quantitatively very similar (supplementary fig. S6, Supplementary Material online). Computational simulations indicate that the inference procedure is very accurate and is robust to many confounding factors (Yang et al. 1995; Zhang and Nei 1997; Pond and Frost 2005) (supplementary note S1, Supplementary Material online).

Passage adaptation studied here represents one of the strongest forms of adaptation studied so far (Hughes 1999; Eyre-Walker 2006). Interestingly, the same passage environment also leads to strong convergent evolution (fig. 2*d*). This is in contrast with many scenarios happened in the natural world where divergent genetic mechanisms can evolve under convergent environmental pressure (e.g., the origin of flight across the tree of life) (Christin et al. 2010; Stern 2013). Since the probability of convergent changes depend on the mutation rate, the probability and magnitude of beneficial mutations (Chevin et al. 2010), convergent evolution in the same gene has often been observed, but much less often for identical substitutions in the same gene (i.e., so called hotspot gene theory, Martin and Orgogozo 2013). Understanding theoretical aspects of this observation in the influenza virus warrants another study going beyond the current work. Nevertheless, viral propagation conducted continually for the past 48 years at research centers across the world makes passage adaptation in influenza another textbook example for convergent adaptive evolution (Boucher et al. 1992; Kellam et al. 1994; Bull et al. 1997; Zhang 2006).

The differences in the number of positively selected codons in early stages versus in late stages are very interesting

(fig. 1c and e). We can imagine when the H3N2 virus had just crossed the species boundary in 1968, it was still a very "avian" like virus. Embryonated chicken medium will impose relatively weaker selective pressure on the influenza virus. In this case, possible adaptive trajectories can be quite abundant and positive selection is distributed across many codon positions. Similarly, mammalian cell lines will impose positive selection to the viral population for an avian like virus. As the virus evolves to adapt to the human environment, passage in the avian medium becomes much more challenging and positive selection on the influenza virus gets a lot stronger. Under strong selective pressure, only a few amino acid changes in a few codon positions will allow efficient passage in an avian medium (fig. 1e). As a corollary, passaging the viral population in mammalian cell lines will not apply strong selective pressure on the pathogen due to similar conditions between human and other mammalian cellular environment. This dynamic trajectory provides an important insight into the shape of the fitness landscape and variations in selective pressures can lead to drastically different behavior in adaptive trajectories.

Traditionally, embryonated eggs are often used for vaccine production (Lambert and Fauci 2010). In addition to being highly efficient in propagation, the egg environment can help screening out human pathogens and provide an important safety control for influenza vaccines (Feng et al. 2011). However, adaptation during the culturing process can often generate substitutions deviating from the original clinical specimen (Saito et al. 2004). Due to this constraint and many other factors, the generation of vaccines through embryonated eggs are gradually decreasing (Lambert and Fauci 2010). The patterns revealed from this work indicates that passage associated adaptive changes are more concentrated in a few codon positions in recent years (after year 2000). While one might think that this could make it a lot easier for the field to control passage adaptation compared with the earlier phase, the opposite may be true. The few sites dominating egg adaptation in recent strains are also prominently located in antigenic surface epitopes and their egg adaptation may be responsible for antigenic change associated with recent reduced vaccine efficacy (Skowronski et al. 2014). Since the genomic background of recent strains seems to strongly favor adaptation on these few sites, it is currently a great challenge to find strains that remain antigenically unaltered when passaged in eggs for vaccine production. On the contrary, growing influenza strains in MDCK cells is also not always straightforward and can induce changes that can affect HI titer measurements (Skowronski et al. 2016), drug development and usage (Okomo-Adhiambo et al. 2010).

Even though passage histories are very important for studying influenza evolution, they are not very well documented in the field. For example, most of the public databases do not require the passage annotation to be deposited when submitting sequences. In addition, there is no standard nomenclature when documenting passage histories, and some of the earlier recorded histories could be inaccurate (e.g., some of the strains that are passaged in MDCK cells could potentially have been grown in eggs, supplementary note S2, Supplementary Material online). Given the importance of passage history, the field needs to develop a systematic approach to document this information and this requirement might be generalizable to all viral fields.

In the GISAID and NCBI data set, the proportion of egg passage is low, but it can contribute >10% of the total nonsynonymous changes at a large set of codons (fig. 2c). Even though the percentage of egg passage is dropping for the two databases, the confounding effect from egg passage is likely to persist if we combine data across years. Future studies inferring positive selection and vaccine production using culturing conditions need to be careful in taking passage adaptation into consideration.

Materials and Methods

Data Curation

We combined two sources of information for influenza A (H3N2) HA1 sequences. The first data set was retrieved from the Genbank database at the NCBI in December 2015 (denoted as NCBI set). This NCBI data set included a special subset of strains from 1983 to 1997 generated by the Center for Disease Control and Prevention (denoted as CDC data set, NCBI accessions AF008656 to AF008909 and AF180564 to AF180666) for which passage adaptation was explored earlier (Bush et al. 1999, 2000). The second data set was retrieved from the EpiFlu database from the GISAID (denoted as GISAID set). After removal of low quality sequences (e.g., missing/ambiguous nucleotide bases or year information) and merging the two sets, the total combined data set used for subsequent analysis comprised of 25,482 strains. In detail, 16,638 sequences were found in both databases, 736 were unique from Genbank and 8,108 unique from GISAID. We acknowledge the originating and submitting laboratories of the GISAID sequences as summarized in supplementary table S1, Supplementary Material online.

The passage histories of all the strains were retrieved from the NCBI and GISAID databases. For the strains from the CDC data set included in the NCBI data set, we extracted the passage information from the original publication (Bush et al. 1999, 2000). Due to complex free-form text annotation for passage information (often from original publications), we classified the existing passage histories into nine groups. (1) egg (508 strains), (2) Madin–Darby Canine Kidney (MDCK) (4,148 strains), (3) cell (2,310 strains, being one of the cell lines, but are not documented with a specific type), (4) MDCK cells transfected with the cDNA of human 2,6-sialyltransferase (SIAT) (2,336 strains), (5) Primary Rhesus Monkey Kidney (RMK) cell culture (2,069 strains), (6) mixed (3,037 strains, propagated under at least two types of media, e.g., "C5/SIAT1", "C3/MDCK1", and "M1/S1"), (7) Clinical (1,939 strains, labeled as "CLINICAL" in the database), (8) unknown (8,795 strains, no passage information is available), and (9) other (340 strains, passage information is rather unclear, e.g., "1012_Daty_3" and "X2") (supplementary table S2, Supplementary Material online).

Sequence Alignment and Phylogenetic Inference

Using the computational algorithm implemented in the MUSCLE package (Edgar 2004), we performed multiple sequence alignment of all sequences with default parameter settings for each data set. Phylogenetic relationship among all the isolates was inferred using the RAxML program assuming a gamma distributed rate variation among sites and a general time reversible (GTR) model of sequence evolution (Stamatakis 2006). The inferred evolutionary relationship and associated mutational parameter estimates (e.g., base frequency and transition matrix) were used for the mutational mapping analysis.

Mutational History Inference Using the Mutational Mapping

With the evolutionary relationship and associated mutational parameters (e.g., GTR model and base frequencies), the mutational mapping algorithm provides an efficient way for sampling from all possible evolutionary histories according to their posterior probabilities (Nielsen 2002). The evolutionary histories include both the ancestral states of all the internal nodes and the mutational changes along each branch.

The inference procedure can be conducted in three major steps. First of all, the conditional probabilities of each ancestral node was calculated recursively from the tip of the tree down to the root using the pruning algorithm (Felsenstein 1981). The conditional probability is defined as the probability of observing the subclade information (all the descendants of the focal internal node) conditional on the node being A or T or C or G. Secondly, the state of the internal node is sampled recursively from the root of the tree down to the tips (Nielsen 2002). At each internal node, the probability of the four possible states (A/T/C/G) is a weighted product between the probability that flowed from the ancestor of the node and the conditional probability which flowed from the descendant of this node (Formula 10, Nielsen 2002). Lastly, given the ancestral state of all the internal nodes, a Markov chain can be set up to sample the evolutionary trajectories of that branch conditioning on the states at both ends of each branch.

On top of the mutational mapping algorithm (Nielsen 2002), we simultaneously couple the information across all three nucleotide positions for each codon, and restrict the sampling schemes (both ancestral state and the mutational history) to 61 possible codons (not allowing stop codons) (Zhai et al. 2007). This was done using rejection sampling and constraining the possible states to only 61 sensible codons.

The Binomial Test of Positive Selection

For a specific codon, we can calculate the expected number of nonsynonymous and synonymous sites using the continuous time Markov Chain. For example, the second codon position is always a nonsynonymous site. For the first and third codon position, the expected nonsynonymous site is calculated as the probability of a nonsynonymous substitution if a random mutation happens in that position. This can be easily calculated using the rate matrix from the continuous time Markov Chain. The expected synonymous site and the expected nonsynonymous site add up to 1 for any codon position.

The expected nonsynonymous/synonymous site for a codon along a given branch is calculated as the average of the corresponding values for the starting codon and end codon. The expected nonsynonymous/synonymous site for a codon along many branches can be calculated as the weighted mean of the values from each branch. The weights are simply the branch lengths of the individual branches. Using the expected number of nonsynonymous and synonymous sites, we can calculate the significance of the observed nonsynonymous and synonymous changes. The P value is the sum of the tail probabilities whose nonsynonymous changes are larger or equal to the observed value from the binomial distribution.

The Enrichment Test and Convergent Test

In order to systematically look for the amino acid sites that are responsible for the passage adaptation, we devised two statistical tests targeting two different aspects of the substitution pattern. The enrichment test examines whether mutations that happened in a codon are enriched among the egg branches more often than expected. For example, at a given codon, if there are N changes along the terminal branches, a subset of them (denoted as x) will happen along egg terminals and N-x occur on other terminal branches. The enrichment test will employ a binomial distribution testing whether x is much larger than expected from random chances. The expectation can be calculated from the Binomial distribution (N, p) where p is ratio between the branch lengths of the egg terminals over all terminal branches.

The convergent test is meant to extract the pattern of excess changes from the same codon (e.g., U) to the other codon (e.g., V) along the tip branches. Here, we focus on the nonsynonymous changes only. Starting from a given codon U, we want to test whether the number of changes to V is more than expected. Using synonymous changes as the baseline, we can conduct this test by comparing the number of nonsynonymous changes from U to V to the number of synonymous changes at this codon. In other words, we are testing whether the number of specific nonsynonymous change is more than expected for a specific codon.

To be more precise, for a given codon W, there will be nine neighboring codons with one base difference. Let us denote these neighbors as $C_1, C_2, \ldots, C_9$ and the corresponding observed changes to these neighboring codons as $N_1\ N_2, \ldots, N_9$. Using the GTR model, we can calculate the probability of each of these changes (denoted as p_1 to p_9). The expected probability of synonymous changes will be the sum of the probabilities for the synonymous one-step neighbors. The convergent test will test whether the expected number of U to V changes is more than expected by chance, similar to the traditional d_N/d_S test.

Structural Modeling and the Proportion of Substitutions Due to Passage Adaptation and Convergent Changes

The structural view of the 19 codon positions were displayed using the YASARA package with protein databank (pdb) ID: 4FNK (Krieger and Vriend 2014). For a given codon, the

proportion of nonsynonymous changes contributed by egg passage was calculated as the nonsynonymous changes along the egg terminals divided by the total number of nonsynonymous changes at that codon. The proportion of convergent changes along egg terminals was defined as the contribution of those repeatedly occurring substitutions (happened at least 3 times) over all nonsynonymous changes at a given codon along egg terminals.

The Mixture Model and the Proportion of Egg Passage

For a given set of strains with unknown passage history, we estimate the proportion of egg passage by treating the unknown strains as a mixture of strains grown in eggs and cell medium. The expected nonsynonymous to synonymous substitution for the embryonated egg and cell culture at the 19 codons was first calculated. The proportion of substitutions contributed by egg passage (p_{sub}) can be estimated using a mixture model. Since the 19 codons tend to have more substitutions along the egg terminal branches (i.e., the enrichment test), the percentage of strains passaged in egg culture is then calculated as $(p_{sub}/L)/(p_{sub}/L + 1 - p_{sub})$, where L is the relative ratio between the average number of substitutions along egg terminal branches versus other terminal branches at these 19 codons.

In order to test the accuracy of the estimation procedure (the mixture model), we simulated a series of data sets by composing different proportions of isolates passaged in egg and cell culture. Applying the mixture model procedure to the simulated data and comparing estimated value to the true proportion from the simulated data, we can calibrate the accuracy of the method (supplementary table S5, Supplementary Material online) as well as the sensitivity and specificity of the estimation procedure (supplementary fig. S5, Supplementary Material online). Five-fold cross validation test is also performed (supplementary table S6, Supplementary Material online).

Supplementary Material

Supplementary figures S1–S6, tables S1–S6 and notes S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Author Contributions

W.Z. conceived and supervised the project. H.C., Q.D., and R.T.C.L. analyzed the data. S.H.N., R.T.C.L., and S.M.S. contributed materials and ideas. H.C., S.H.N., S.M.S., and W.Z. wrote the article.

Acknowledgments

We want to thank the editor and two reviewers for their constructive comments and suggestions. We acknowledge Catherine Smith (US CDC) for sharing her expertise on typical passage nomenclature and all originating and submitting laboratories of the sequences downloaded from GISAID. We would like to thank Prof Robin Bush for the communication on the passage history of the CDC data set. We also want to thank Rasmus Nielsen, Martin Lloyd Hibberd, Boon Huan Tan, Richard Sugrue, Hao Fan, Tong Zhang, Paola Florez DE SESSIONS for constructive comments and discussions. We would also like to thank Teck Por Lim for his help on the presentation of this work. Q.D. is supported by a visiting student support from the Genome Institute of Singapore as well as Chinese Academy of Sciences.

References

Bogner P, Capua I, Lipman DJ, Cox NJ. 2006. A global initiative on sharing avian flu data. *Nature* 442:981–981.

Bollback JP. 2006. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* 7:88.

Boucher CA, O'Sullivan E, Mulder JW, Ramautarsing C, Kellam P, Darby G, Lange JM, Goudsmit J, Larder BA. 1992. Ordered appearance of zidovudine resistance mutations during treatment of 18 human immunodeficiency virus-positive subjects. *J Infect Dis.* 165:105–110.

Bull J, Badgett M, Wichman HA, Huelsenbeck JP, Hillis DM, Gulati A, Ho C, Molineux I. 1997. Exceptional convergent evolution in a virus. *Genetics* 147:1497–1507.

Burnet F. 1940. Influenza virus infections of the chick embryo by the amniotic route. I. General character of the infections. *Aust J Exp Biol Med Sci.* 18:353–360.

Burnet F. 1941. Growth of influenza virus in the allantoic cavity of the chick embryo. *Aust J Exp Biol Med Sci.* 19:291–295.

Burnet F, Bull DR. 1943. Changes in influenza virus associated with adaptation to passage in chick embryos. *Aust J Exp Biol Med Sci.* 21:55–69.

Bush RM, Fitch WM, Bender CA, Cox NJ. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol.* 16:1457–1465.

Bush RM, Smith CB, Cox NJ, Fitch WM. 2000. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proc Natl Acad Sci U S A.* 97:6974–6980.

Cao JX, Ni H, Wills MR, Campbell GA, Sil BK, Ryman KD, Kitchen I, Barrett AD. 1995. Passage of Japanese encephalitis virus in HeLa cells results in attenuation of virulence in mice. *J Gen Virol.* 76 (Pt 11):2757–2764.

Chevin LM, Martin G, Lenormand T. 2010. Fisher's model and the genomics of adaptation: restricted pleiotropy, heterogenous mutation, and parallel evolution. *Evolution* 64:3213–3231.

Christin P-A, Weinreich DM, Besnard G. 2010. Causes and evolutionary significance of genetic convergence. *Trends Genet.* 26:400–405.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends Ecol Evol.* 21:569–575.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17:368–376.

Feng SZ, Jiao PR, Qi WB, Fan HY, Liao M. 2011. Development and strategies of cell-culture technology for influenza vaccine. *Appl Microbiol Biotechnol.* 89:893–902.

Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H (3) HA1 human influenza type A. *Proc Natl Acad Sci.* 94:7712–7718.

Fitch WM, Leiter J, Li X, Palese P. 1991. Positive Darwinian evolution in human influenza A viruses. *Proc Natl Acad Sci.* 88:4270–4274.

Gatherer D. 2010. Passage in egg culture is a major cause of apparent positive selection in influenza B hemagglutinin. *J Med Virol.* 82:123–127.

Govorkova E, Kaverin N, Gubareva L, Meignier B, Webster R. 1995. Replication of influenza A viruses in a green monkey kidney continuous cell line (Vero). *J Infect Dis.* 172:250–253.

Govorkova EA, Matrosovich MN, Tuzikov AB, Bovin NV, Gerdil C, Fanget B, Webster RG. 1999. Selection of receptor-binding variants of human influenza A and B viruses in baby hamster kidney cells. *Virology* 262:31–38.

Graff J, Normann A, Feinstone SM, Flehmig B. 1994. Nucleotide sequence of wild-type hepatitis A virus GBM in comparison with two cell culture-adapted variants. *J Virol.* 68:548–554.

Gubareva LV, Wood JM, Meyer WJ, Katz JM, Robertson JS, Major D, Webster RG. 1994. Codominant mixtures of viruses in reference strains of influenza virus due to host cell variation. *Virology* 199:89–97.

Hardy CT, Young SA, Webster RG, Naeve CW, Owens RJ. 1995. Egg fluids and cells of the chorioallantoic membrane of embryonated chicken eggs can select different variants of influenza A (H3N2) viruses. *Virology* 211:302–306.

Hilleman MR. 2000. Vaccines in historic evolution and perspective: a narrative of vaccine discoveries. *Vaccine* 18:1436–1447.

Hughes AL. 1999. Adaptive evolution of genes and genomes. Oxford University Press.

Itoh M, Isegawa Y, Hotta H, Homma M. 1997. Isolation of an avirulent mutant of Sendai virus with two amino acid mutations from a highly virulent field strain through adaptation to LLC-MK2 cells. *J Gen Virol.* 78:3207–3215.

Katz JM, Wang M, Webster RG. 1990. Direct sequencing of the HA gene of influenza (H3N2) virus in original clinical samples reveals sequence identity with mammalian cell-grown virus. *J Virol.* 64:1808–1811.

Kellam P, Boucher CA, Tijnagel JM, Larder BA. 1994. Zidovudine treatment results in the selection of human immunodeficiency virus type 1 variants whose genotypes confer increasing levels of drug resistance. *J Gen Virol.* 75:341–351.

Kodihalli S, Justewicz DM, Gubareva LV, Webster RG. 1995. Selection of a single amino acid substitution in the hemagglutinin molecule by chicken eggs can render influenza A virus (H3) candidate vaccine ineffective. *J Virol.* 69:4888–4897.

Krieger E, Vriend G. 2014. YASARA View – molecular graphics for all devices – from smartphones to workstations. *Bioinformatics* 30:2981–2982.

Lambert LC, Fauci AS. 2010. Influenza vaccines for the future. *N Engl J Med.* 363:2036–2044.

Li W-H, Wu C-I, Luo C-C. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol.* 2:150–174.

Lohmann V, Korner F, Dobierzewska A, Bartenschlager R. 2001. Mutations in hepatitis C virus RNAs conferring cell culture adaptation. *J Virol.* 75:1437–1449.

Lorenzo FR, Tanaka T, Takahashi H, Ichiyama K, Hoshino Y, Yamada K, Inoue J, Takahashi M, Okamoto H. 2008. Mutational events during the primary propagation and consecutive passages of hepatitis E virus strain JE03-1760F in cell culture. *Virus Res.* 137:86–96.

Madin S, Darby N. 1958. Established kidney cell lines of normal adult bovine and ovine origin. *Exp Biol Med.* 98:574–576.

Martin A, Orgogozo V. 2013. The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution* 67:1235–1250.

Nakajima S, Nakajima K, Kendal AP. 1983. Identification of the binding sites to monoclonal antibodies on A/USSR/90/77 (H1N1) hemagglutinin and their involvement in antigenic drift in H1N1 influenza viruses. *Virology* 131:116–127.

Nakowitsch S, Waltenberger AM, Wressnigg N, Ferstl N, Triendl A, Kiefmann B, Montomoli E, Lapini G, Sergeeva M, Muster T, Romanova JR. 2014. Egg- or cell culture-derived hemagglutinin mutations impair virus stability and antigen content of inactivated influenza vaccines. *Biotechnol J.* 9:405–414.

Nielsen R. 2002. Mapping mutations on phylogenies. *Syst Biol.* 51:729–739.

Okomo-Adhiambo M, Nguyen HT, Sleeman K, Sheu TG, Deyde VM, Garten RJ, Xu X, Shaw MW, Klimov AI, Gubareva LV. 2010. Host cell selection of influenza neuraminidase variants: implications for drug resistance monitoring in A(H1N1) viruses. *Antiviral Res.* 85:381–388.

Plotkin JB, Dushoff J. 2003. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proc Natl Acad Sci.* 100:7152–7157.

Pond SLK, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol.* 22:1208–1222.

Pond SLK, Poon AF, Brown AJL, Frost SD. 2008. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol Biol Evol.* 25:1809–1824.

Robertson J. 1993. Clinical influenza virus and the embryonated hen's egg. *Rev Med Virol.* 3:97–106.

Robertson JS, Cook P, Attwell AM, Williams SP. 1995. Replicative advantage in tissue culture of egg-adapted influenza virus over tissue-culture derived virus: implications for vaccine manufacture. *Vaccine* 13:1583–1588.

Rocha EP, Xu X, Hall HE, Allen JR, Regnery HL, Cox NJ. 1993. Comparison of 10 influenza A (H1N1 and H3N2) haemagglutinin sequences obtained directly from clinical specimens to those of MDCK cell-and egg-grown viruses. *J Gen Virol.* 74:2513–2518.

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

Saito T, Nakaya Y, Suzuki T, Ito R, Saito T, Saito H, Takao S, Sahara K, Odagiri T, Murata T. 2004. Antigenic alteration of influenza B virus associated with loss of a glycosylation site due to host-cell adaptation. *J Med Virol.* 74:336–343.

Sawyer L, Wrin MT, Crawford-Miksza L, Potts B, Wu Y, Weber PA, Alfonso RD, Hanson CV. 1994. Neutralization sensitivity of human immunodeficiency virus type 1 is determined in part by the cell in which the virus is propagated. *J Virol.* 68:1342–1349.

Schepetiuk SK, Kok T. 1993. The use of MDCK, MEK and LLC-MK2 cell lines with enzyme immunoassay for the isolation of influenza and parainfluenza viruses from clinical specimens. *J Virol Methods* 42:241–250.

Seo SH, Goloubeva O, Webby R, Webster RG. 2001. Characterization of a porcine lung epithelial cell line suitable for influenza virus studies. *J Virol.* 75:9517–9525.

Skowronski DM, Janjua NZ, De Serres G, Sabaiduc S, Eshaghi A, Dickinson JA, Fonseca K, Winter AL, Gubbay JB, Krajden M, et al. 2014. Low 2012-13 influenza vaccine effectiveness associated with mutation in the egg-adapted H3N2 vaccine strain not antigenic drift in circulating viruses. *PLoS One* 9:e92153.

Skowronski DM, Sabaiduc S, Chambers C, Eshaghi A, Gubbay JB, Krajden M, Drews SJ, Martineau C, De Serres G, Dickinson JA, et al. 2016. Mutations acquired during cell culture isolation may affect antigenic characterisation of influenza A(H3N2) clade 3C.2a viruses. *Euro Surveill.* 21:30112.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Stern DL. 2013. The genetic causes of convergent evolution. *Nat Rev Genet.* 14:751–764.

Su YC, Bahl J, Joseph U, Butt KM, Peck HA, Koay ES, Oon LL, Barr IG, Vijaykrishna D, Smith GJ. 2015. Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nat Commun.* 6:7952.

Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol.* 16:1315–1328.

Swofford DL, Maddison WP. 1987. Reconstructing ancestral character states under Wagner parsimony. *Math Biosci.* 87: 199–229.

Wilson I, Skehel J, Wiley D. 1981. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature* 289:366–373.

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–1650.

Zhai W, Slatkin M, Nielsen R. 2007. Exploring variation in the d(N)/d(S) ratio among sites and lineages using mutational mappings: applications to the influenza virus. *J Mol Evol.* 65:340–348.

Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat Genet.* 38:819–823.

Zhang J, Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol.* 44:S139–S146.