Deep mutational scanning of hemagglutinin helps distinguish the evolutionary fate of human H3N2 influenza virus lineages

Juhye M. Lee[a,d,e,1], **John Huddleston**[b,f,1], **Michael B. Doud**[a,d,e], **Kathryn A. Hooper**[a,f], **Trevor Bedford**[b,c,2], **and Jesse D. Bloom**[a,c,d,2]

[a]Basic Sciences Division; [b]Vaccine and Infectious Diseases Division; [c]and Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [d]Department of Genome Sciences; [e]Medical Scientist Training Program; [f]and Molecular and Cellular Biology Program, University of Washington, Seattle, WA 98195, USA

This manuscript was compiled on January 8, 2018

The evolution of seasonal H3N2 influenza virus is mediated by the rapid accumulation of mutations. A better understanding of how mutations affect viral growth is critical for studying the pressures and constraints underlying the evolution of the influenza virus in nature. Here we experimentally measure the effect of every possible single amino-acid mutation to an H3 influenza hemagglutinin on viral growth in cell culture. We find the stalk domain to be fairly mutationally tolerant, and there is not a large disparity in tolerance between the head and stalk domains. Furthermore, our measurements reveal mutations in successful seasonal influenza viral lineages to be significantly more preferred than mutations in lineages that have died out, and this was true at both epitope and non-epitope sites. Epistasis in hemagglutinin is also seemingly driven by antigenic drift. The preferences measured in a distantly related protein homolog do not reveal such differences in viral lineages. A comparison of the preferences between the two hemagglutinin homologs suggests that mutational effects increasingly shift as proteins diverge. Overall, our work highlights how experimental measurements of mutational effects can be leveraged to gain insight into the evolutionary fates of viral strains and potentially be used to predict viral evolution.

influenza virus | hemagglutinin | deep mutational scanning | antigenic drift | epistasis

Seasonal H3N2 influenza virus evolves rapidly, readily accumulating mutations in the hemagglutinin (HA) surface protein that contribute to antigenic drift. The evolution of human H3N2 influenza virus is further characterized by strain competition and frequent population turnover (1–3), producing a distinct spindly shape in its phylogenetic tree with a persistent trunk lineage and short-lived side branches (4) (Figure 1). Given the rather distinctive pattern of H3N2 evolution and that vaccination remains an effective way of reducing disease burden, a key challenge in selecting a vaccine strain lies in being able to predict the dominant strain in the upcoming influenza season.

Vaccine strain selection decisions mainly rely on experimentally characterizing the antigenicity of circulating strains. To aid in the timeliness and throughput of making such decisions, previous work have endeavored to accurately predict the evolutionary outcomes of viral strains based on the antigenicity and/or tree structure of viral clades (5, 6). However, the accuracy of these predictions are subject to uneven geographical and temporal sampling, and for (6), the lack of information about a mutation's effect on antigenicity. Others have attempted to model the antigenic properties of viral strains using sequence information or phylogenetic relationships (7–9).

While mutations that contribute to antigenic evolution of the virus certainly determine strain success, the non-antigenic impacts of mutations play an important role in determining the persistence of viral strains (2, 10–13). Due to the high mutation rate of RNA viruses [*citation here*] and because the influenza virus does not appreciably recombine within a given segment, deleterious mutations become linked to beneficial ones. A predictive fitness model takes into account mutations at both antigenic and non-antigenic sites in HA to infer clade fitness (12). In this model, mutations at epitope sites are beneficial to viral fitness, while mutations at non-epitope sites incur a fitness cost. Although this assumption may generally be true, the effect of mutations on viral fitness, and particularly on protein function, are much more nuanced. We therefore need a greater understanding of the phenotypic effects of mutations on viral growth to comprehend the forces and constraints shaping influenza virus evolution in nature.

We can measure such mutational effects in the lab using deep mutational scanning approaches (14). Previously, we experimentally measured the effect of all possible single amino-acid mutations to an H1 HA from the A/WSN/1933 (H1N1) strain (15, 16). However, the WSN/1933 is a highly lab-adapted strain, and these measurements may not be particularly relevant for studying the mutational processes of more contemporaneous strains circulating in the human population. Here we measure the effects of mutations to the HA from the A/Perth/16/2009 (H3N2) strain on viral growth in cell culture. We apply these experimental measurements to examine differences in the trunk and side branch lineages of human H3N2 influenza virus. Trunk mutations were found to be more

Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the Significance Statement is to explain the relevance of the work in broad context to a broad readership. The Significance Statement appears in the paper itself and is required for all research papers.

Please provide details of author contributions here.

Please declare any conflict of interest here.

[1]J.M.L. and J.H. contributed equally to this work.

[2]To whom correspondence should be addressed: trevor@bedford.io, jbloom@fredhutch.org

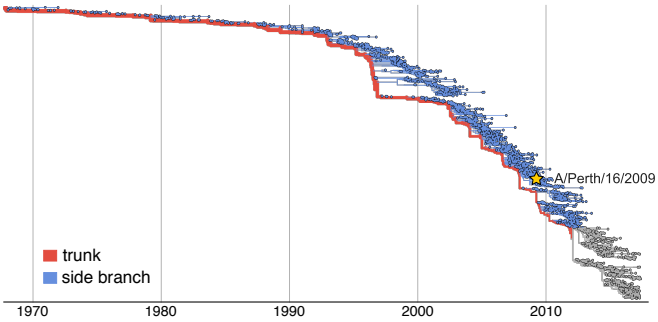

Fig. 1. Human H3N2 HA phylogeny from 1968-2017. The trunk is shown in red, and side branches are shown in blue. The gray branches represent the part of the tree for which we cannot yet distinguish the trunk from side branches. The Perth/2009 strain is indicated with a star.

favorable than side branch mutations, both at epitope and non-epitope sites. We also find evidence that epistasis in HA has primarily been driven by antigenic drift. The measurements also enabled a comparison of how the amino-acid preferences have shifted between two diverged HAs. The extent of shifts in amino-acid preferences increase as homologs diverge. Our work highlights the potential for using high-throughput experimental measurements of mutational effects to inform evolutionary forecasting of human seasonal influenza virus.

Results

Deep mutational scanning of HA from a recent strain of human H3N2 influenza virus. We performed a deep mutational scan to measure the effects of all amino-acid mutations to HA from the A/Perth/16/2009 (H3N2) strain on viral replication in cell culture. This strain was the H3N2 component of the influenza vaccine from 2010-2012 (17, 18). Relative to the consensus sequence for this HA in Genbank, we used a variant with two mutations that enhanced viral replication in cell culture, G78D and T212I (Figure S1 and Dataset S1). The G78D mutation occurs at low frequency in natural H3N2 sequences, and T212 is a site where a mutation to Ala rose to fixation in human influenza in ∼2011.

We mutagenized the entire HA coding sequence at the codon level to create mutant plasmid libraries harboring an average of ∼1.4 codon mutations per clone (Figure S2). We then generated mutant virus libraries from the mutant plasmids using a helper-virus system that enables the efficient generation of complex influenza virus libraries (16) (Figure 2A). These mutant viruses derived all of their non-HA genes from the lab-adapted WSN/1933 strain. Using WSN/1933 for the non-HA genes reduces biosafety concerns, and also helped increase viral titers. To further increase viral titers, we used MDCK-SIAT1 cells that we had engineered to constitutively express the TMPRSS2 protease, which facilitates HA cleavage and activation (19, 20).

After generating the mutant virus libraries, we passaged them at low MOI in cell culture to create a genotype-phenotype link and select for functional HA variants (Figure 2A). All of the experiments were completed in full biological triplicate (Figure 2B). We also passaged and deep sequenced library 3 in duplicate (denoted as library 3-1 and 3-2) to gauge to the amount of experimental noise occurring *within* a single biological replicate. As a control to measure sequencing and mutational errors, we used the unmutated HA gene to generate and passage viruses carrying wildtype HA.

Deep sequencing of the initial plasmid mutant libraries and the passaged mutant viruses allowed us to observe selection for functional HA mutants. As expected, selection against non-functional HA variants as led to reduced mutation frequencies in the mutant viruses compared to the initial mutant plasmids (Figure 2C). Specifically, stop codons were purged to 20-45% of their initial frequencies after correcting for error rates estimated by sequencing the wildtype controls. The incomplete purging of stop codons is likely because genetic complementation due to co-infection (21) enabled the persistence of some virions with nonfunctional HAs. We also observed selection against many nonsynonymous mutations, with their frequencies falling to 30-40% of their initial values after error correction.

We next quantified the reproducibility of our deep mutational scanning measurements across biological and technical replicates. We first used the deep sequencing data for each replicate to independently estimate the preference of each site in HA for all 20 amino acids using the method described in (22). Because there are 567 residues in HA, there are $567 \times 20 = 11,340$ estimated amino-acid preferences. The correlations of the amino-acid preferences between pairs of replicates are shown in Figure 2D. The biological replicates are fairly well-correlated, with Pearson's R ranging from 0.69 to 0.78. Replicate 1 exhibited the lowest correlation with the other replicates, consistent with the observation that this replicate also showed the weakest selection against stop and nonsynonymous mutations perhaps indicating more experimental noise. The two technical replicates 3-1 and 3-2 were only slightly more correlated than pairs of biological replicates, suggesting that bottlenecking of library diversity after the reverse-genetics step contributes most of the experimental noise.

Our measurements are consistent with existing knowledge about HA's evolution and function. How do the HA amino-acid preferences measured in our experiments relate to the evolution of H3N2 influenza virus in nature? This question can be addressed by evaluating how well an experimentally informed codon substitution model (ExpCM) using our measurements describe H3N2 evolution compared to standard phylogenetic substitution models (23, 24). Table 1 shows that the ExpCM using the replicate-average measurements greatly outperforms conventional substitution models, showing that our experiments authentically capture some of the constraints on HA evolution. Note that while the relative ratio of nonsynonymous to synonymous substitutions (dN/dS or ω) is $\ll 1$ for conventional substitution models, this ratio is close to one for the ExpCM that accounts for the constraints measured in the experiments. ExpCM contain a stringency parameter that relates the selection in the experiments to that in nature (23, 24). We fit a stringency parameter of 2.44 (Table 1), indicating that natural selection favors the same amino acids as the experiments, but with greater stringency. Throughout the rest of this paper, we use experimental measurements re-scaled (23, 24) by this stringency parameter. These re-scaled preferences are shown in Figure 3.

A closer examination of Figure 3 reveals that the experimentally measured amino-acid preferences generally agree with existing knowledge about HA's structure and function.

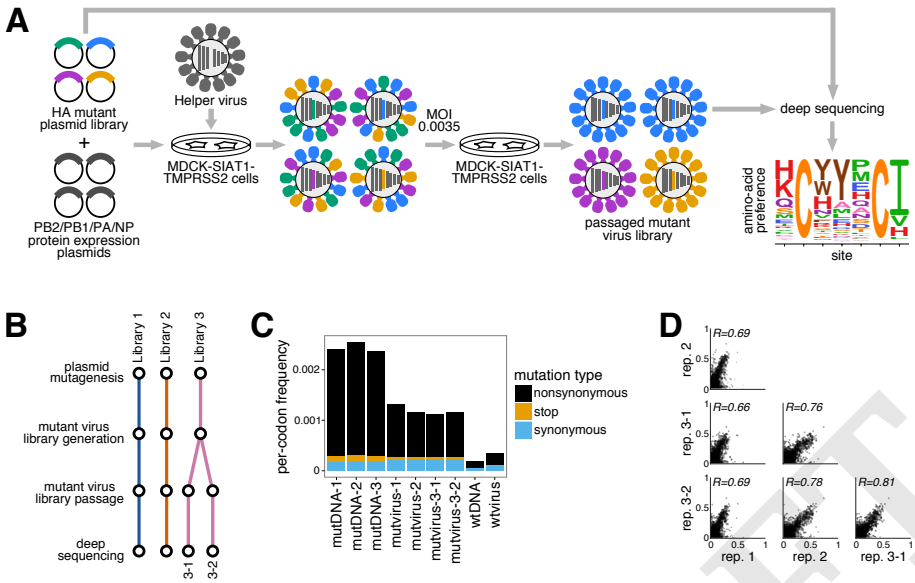

Fig. 2. Deep mutational scanning of the Perth/2009 H3 HA. (A) We generated mutant virus libraries using a helper-virus approach (16), and passaged them at low MOI to establish a genotype-phenotype linkage and to select for functional HA variants. Deep sequencing of the variants before and after selection allowed us to estimate each site's amino-acid preferences. (B) The experiments were performed in full biological triplicate. We also passaged and deep sequenced library 3 in duplicate. (C) Frequencies of nonsynonymous, stop, and synonymous mutations in the mutant plasmid DNA, the passaged mutant viruses, and wildtype DNA and virus controls. (D) The Pearson correlations among the amino-acid preferences estimated in each replicate.

Table 1. Substitution models informed by the experiments describe H3N2's natural evolution better than traditional substitution models.

Model	ΔAIC	Log Likelihood	Stringency	ω
ExpCM	0.0	-8439.33	2.44	0.91
Goldman-Yang M5	2166.06	-9516.36	–	0.30, 0.84, 0.36
ExpCM, averaged across sites	2504.18	-9691.42	0.68	0.32
Goldman-Yang M0	2607.92	-9738.29	–	0.31

Maximum likelihood phylogenetic fit to an alignment of human H3N2 influenza HAs using ExpCM (24), ExpCM in which the experimental measurements are averaged across sites, and the M0 and M5 versions of the Goldman-Yang model (25). Models are compared by AIC (26). The ω parameter is dN/dS for the Goldman-Yang models, and the relative dN/dS after accounting for the measurements for the ExpCM. For the M5 model, we give the shape parameter, rate parameter, and mean of the gamma distribution over ω.

For instance, sites that form structurally important disulfide bridges (sites 52 & 277, 64 & 76, 97 & 139, 281 & 305, 14 & 137-HA2, 144-HA2 & 148-HA2) (27) possess high preference for cysteine. At residues involved in receptor binding, there are strong preferences for the amino acids that are known to be involved in binding sialic acid, such Y98, D190, W153, and S228 (28–30)[*These old citations are for X-31 HA, you might also see if Paulson and/or Ian Wilson have newer structural papers on receptor-binding in recent H3 – if so, add them*]. A positively charged amino acid at site 329 is important for cleavage activation of the HA0 precursor, and indeed this site exhibits a high preference for arginine (31, 32). However, a notable exception occurs at the start codon at position -16, which does not show a strong preference for methionine. This codon is part of the signal peptide and is cleaved from the mature HA protein. One possible reason we do not see a strong preference for methionine at this site is due to alternative translation-initiation at a downstream or upstream start site, as has been described for other HAs (33).

There is less difference in mutational tolerance between the HA head and stalk domains for H3 than for H1. Our experiments measure which amino acids are tolerated at each HA site under selection for protein function. We can therefore use our experimentally measured amino-acid preferences to calculate the inherent mutational tolerance of each site, which we define as the Shannon entropy of the re-scaled preferences. In prior mutational studies of the WSN/1933, the stalk domain of this H1 HA was found to be substantially less mutationally tolerant than its globular head (15, 16).

We performed a similar analysis using the current data for the Perth/2009 H3 HA. Surprisingly, there was much less contrast in mutational tolerance between the stalk and head domains for the H3 HA than for the H1 (Figure 4). For instance, the in H3 HA, the short helix A in the stalk is very mutationally tolerant. Interestingly, more studies have reported selecting escape mutants from broadly neutralizing anti-stalk antibodies in H3 (35–38) than in H1 (39, 40) HAs.

We also see high mutational tolerance in many of the known antigenic regions of H3 HA (41). For instance, in recent H3N2 strains, antigenic region B is immunodominant and the location of most major antigenic drift mutations (42–44). We find that the most distal portion of the globular head near the 190-helix, which is part of antigenic region B, is highly tolerant of mutations (Figure 4). Antigenic region C is also notably mutationally tolerant.

Many residues inside HA's receptor binding pocket are known to be highly functionally constrained (29, 45), and our data indicates that these sites are relatively mutationally intolerant in both H3 and H1 HAs. In contrast, the residues surrounding the receptor binding pocket are fairly mutationally tolerant, which may contribute to the rapidity of influenza's antigenic evolution, as these sites are under strong immune pressure. (41, 43).

The experimental measurements can help discriminate between successful and unsuccessful influenza virus lineages. A major goal in the study of rapidly evolving viruses such as influenza is to forecast evolutionary trajectories[*Michael Lassig has a Nature Ecology & Evolution review*]. For instance, forecasting which viral lineages will dominate the upcoming influenza season is an integral part of vaccine-strain selection[*add reference*]. Evolutionary forecasts must ultimately distinguish between successful and unsuccessful viral lineages, which in the case of human influenza virus means distinguishing between the trunk and side branches of the phylogenetic tree (Figure 1).

To investigate whether our experiments can aid in distinguishing trunk and side branch lineages, we calculated the experimentally measured effects of all mutations in a maximum-likelihood reconstruction of the phylogeny in Figure 1. The mutations that occurred on the trunk of the tree were consistently more beneficial for viral growth according to our experimental measurements (Figure 5A). This difference in the effects of mutations between the trunk and side branch lineages was statistically significant when taken across the entire phylogeny (Figure 5B). Some influenza sequences are determined using egg- or cell-passaged isolates that contain lab-adaptation mutations (48–50). These mutations will largely occur on the terminal side branches that lead to tip nodes on the phylogenetic tree. We therefore re-calculated the statistics separating side branches to tip and internal nodes, and again found that the difference between the trunk and both sets of side branches was statistically significant (Figure 5B). Therefore, strains with mutations that we experimentally measure to be favorable for viral replication tend to do better in nature than strains with mutations that we measure to be less favorable.

Our experiments were performed on the Perth/2009 HA, but we are scoring mutations on a phylogenetic tree that extends from 1968 to 2017. Because the effects of mutations can vary with genetic background[*cite some papers about epistasis, amino-acid shifts, etc*], it is possible that the effects of mutations that we measured in the Perth/2009 HA might have shifted somewhat in other related HAs. To explore this question, we scored the complete HA sequence of every node in the phylogenetic tree by quantifying its average per-site sequence preference for HA sequence $\mathbf{s}$, which is defined as

$$F\left(\mathbf{s}\right) = \frac{1}{L}\sum_{r=1}^{L}\ln \pi_{r,s_r}, \qquad [1]$$

where $\pi_{r,a}$ is the preference for amino-acid a at site r as measured in our experiments (e.g., Figure 3), s_r is the amino-acid at site r in HA sequence $\mathbf{s}$, and L is the length of the sequence. Figure 5C shows that the sequences of trunk nodes tend to be more highly preferred than those of side branch nodes across the entire timespan, consistent with the finding that trunk mutations are generally more favorable than side branch mutations. However, the average per-site sequence preference increases as the nodes approach the Perth/2009 strain. The Perth/2009 itself has among the highest per-site sequence preference of the entire tree, despite falling on a side branch. These results suggest that there is some HA background-dependence in the effects of mutations, but that despite this fact our experimental measurements consistently reveal the selective advantage of the trunk over a more than half-century of viral evolution.

Our experiments suggest different patterns of evolution at epitope and non-epitope sites. HA is under selection both to maintain its essential function in viral growth and to escape pre-existing immunity. Most of the immune selection is focused on a subset of so-called "epitope sites" that are targets of the

Fig. 3. **The site-specific amino-acid preferences of the Perth/2009 HA measure in our experiments.** The height of each letter is the preference for that amino acid, after taking the average of our experimental replicates and re-scaling (24) by the stringency parameter in Table 1. The sites are in H3 numbering. The top overlay bar indicates whether or not a site is part of the set of epitope residues delineated in (34) The bottom overlay bar indicates the HA domain (sig. pep. = signal peptide, HA1 ecto. = HA1 ectodomain, HA2 ecto. = HA2 ectodomain, TM = transmembrane domain, cyto. tail. = cytoplasmic tail). The letters directly above each logo stack indicate the wildtype amino acid at that site.

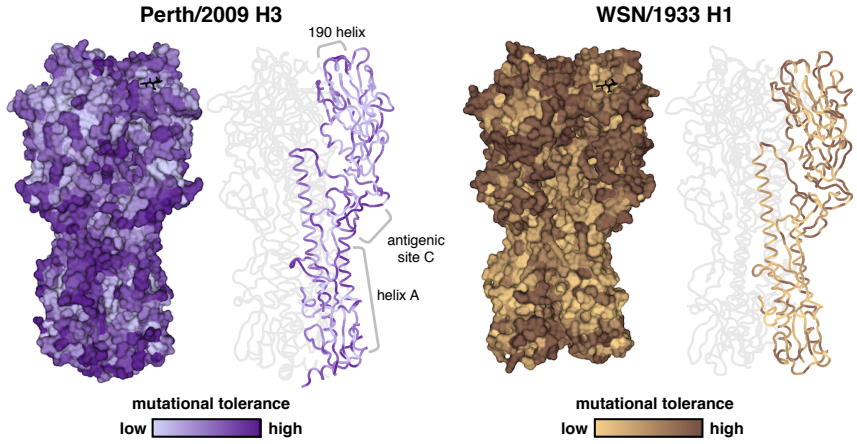

Fig. 4. **Mutational tolerance of each site in H3 and H1 HAs.** Mutational tolerance as measured in the current study is mapped onto the structure of the H3 trimer (PDB 4O5N; (46)). Mutational tolerance of the WSN/1933 H1 HA as measured in (16) is mapped on the structure of the H1 trimer (PDB 1RVX; (47)). Different color scales are used because measurements are comparable among sites within the same HA, but not necessarily across HAs. Both trimers are shown in approximately the same orientation. For each HA, the structure at left shows a surface representation of the full trimer, while the structure at right side shows a ribbon representation of just one monomer. The sialic acid receptor is shown in black sticks.

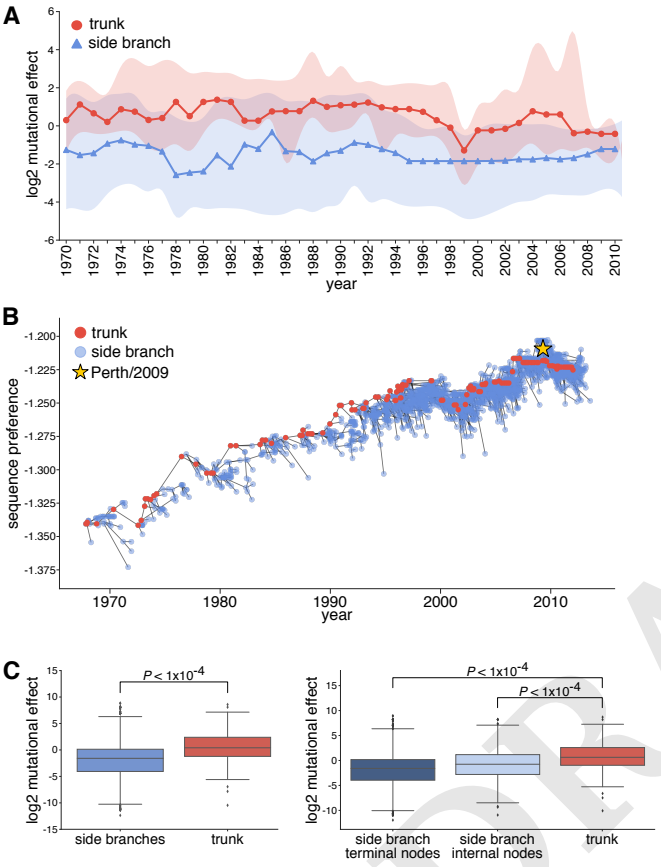

Fig. 5. Mutations on the trunk are more favorable than those on the side branches according to our experimental measurements. [*Legend will be written under assumption that panels (B) and (C) are swapped. Also add sequence preference* per site *to y-axis of the new (C).*] (A) We used our experiments to calculate the $\log_2$ mutational effect for all trunk and side branch mutations in five-year windows from 1968 to 2013. The central year in each window is denoted on the x-axis. Median mutational effects in each window are shown as circles for the trunk and triangles for side branches. The shaded regions demarcate the interquartile range. Negative numbers signify mutations towards less preferred amino acids, while positive numbers signify mutations to more preferred mutations. (B) The $\log_2$ mutational effect for all side branch and all trunk mutations (left panel), and the same data but separating side branches to terminal and internal nodes. P-values were computed by randomizing the experimental measurements among sites 10,000 times, and determining how often the difference in median mutational effect between the trunk and side branches for the randomized data exceeded the actual value. (C) The average per-site sequence preference of every node in the phylogenetic tree in Fig. 1. Trunk nodes are in red, side branch nodes in blue, and Perth/2009 is marked with a yellow star. More preferred sequences have larger values.

immunodominant antibody response. Although both epitope and non-epitope sites evolve rapidly, immune selection drives a higher rate of evolution at epitope sites. In the timeframe from 1968 to 2012, the trunk of the tree in Figure 1 fixed [*X*] epitope mutations per site, but only Y non-epitope mutations per site. Since our experiments only measure how mutations affect viral growth, they might be expected to describe the evolution of epitope and non-epitope sites differently.

Mutations that occur on the trunk are scored as significantly more favorable in our experiments than side-branch mutations at both epitope and non-epitope sites (Figure 6A). Therefore, our experiments can distinguish evolutionarily favorable mutations both at sites that are predominantly under functional constraint and ones that are also under immune pressure. The fact that our experiments can discriminate between the trunk and side branch at epitope sites despite only assaying for viral growth underscores the fact that most epitope sites are still important for HA function[*cite?*].

But interestingly, there are clear differences in how the average per-site sequence preference changes over time at the epitope and non-epitope sites (Figure 6B). The per-site sequence preference at epitope sites increases over time (Figure 6A [*and new associated supplement that zooms in*]), whereas at non-epitope sites it remains relatively constant. In addition, the per-site sequence preference is consistently higher at non-epitope than epitope sites, perhaps consistent with the fact that strong immune selection can lead to the fixation of less functionally favorable mutations at epitope sites. The fact that the epitope sites exhibit more of an increase in sequence preference over time may be due to the fact that they fix more mutations. It is known that as mutations accumulate at protein sites, they can lead to epistatic shifts in the effects of mutations[*general epistasis appears*]. Such epistasis has been experimentally demonstrated for HA[*Yewdell paper, maybe Scott Hensley has something else*], including at some specific epitope sites in H3 HA (51). We hypothesize that the rapid evolution of epitope sites in HA leads to epistatic interactions that shift the effects of mutations, leading to an apparent increase in per-site sequence preference over time when mutational effects are measured in the Perth/2009 HA.

Somehow merged into next subsection. Can we then distinguish lineage-specific mutational effects using the preferences measured in a distantly related HA homolog? We used the WSN/1933 H1 preferences measured in (16) to quantify the effects of H3 trunk and side branch mutations, shown in Figure 7. It is evident that we do not see trunk mutations significantly more favored than side branch mutations, suggesting that our ability to discriminate successful and unsuccessful strains degrades over sufficiently long evolutionary distances.

The H1 and H3 preferences have shifted at many sites. How shifted are the preferences between evolutionarily distant homologs such as H1 and H3? Although we have previously compared the experimentally measured preferences between related protein homologs of NP (52) and of Env (53), the HA homologs we have studied are considerably more diverged than either of these pairs. The WSN/1933 H1 and the Perth/2009 H3 fall into two phylogenetically distinct groups and share 42% amino-acid identity (Figure 8A), allowing us to examine what has shifted and what has remained conserved across such diverged homologs. Simply correlating the preferences

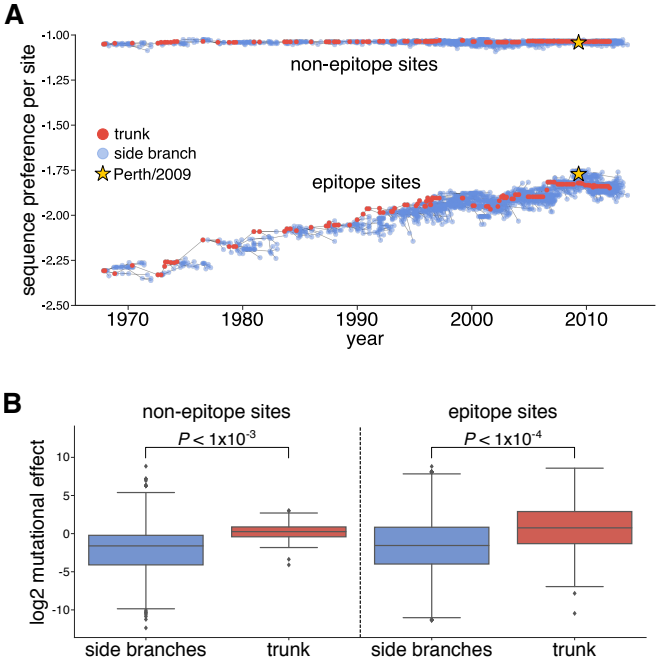

Fig. 6. [*Switch panels A and B, the legend is already written to reflect that.*] **Effects of mutations at epitope and non-epitope sites during HA evolution.** We partitioned HA into epitope and non-epitope sites using the definitions of Wolf *et al* (34). (A) The $\log_2$ mutational effects for side branch and trunk mutations at non-epitope (left) and epitope (right) sites. P-values were computed as in Figure 5 but only performing the randomizations over the appropriate set of sites (non-epitope or epitope). (B) The average per-site sequence preference for all nodes in the phylogenetic tree, calculated separately for each set of sites. [*Write the P-value as* $< 10^{-4}$ *if all* 10^4 *randomizations are less than the actual value. Otherwise write something like 0.0007 or (maybe if a lot of zeros* 7×10^{-4}*. Add supplementary figure for the zoom-in of non-epitope sites and then refer to it in the Figure legend here. You could even just add the old figure that split them and shows both epitope and non-epitope on their own y-axis as the supplement. I think the randomizations should be done just within the same set of sites.*]

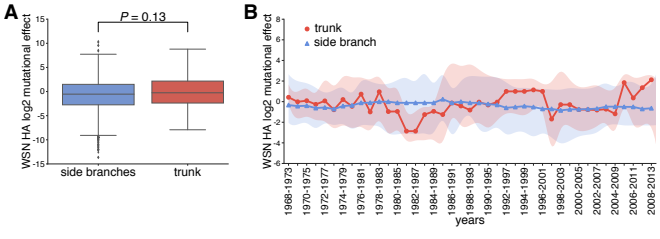

Fig. 7. Experimental measurements on an H1 HA do *not* provide evolutionarily relevant information for H3 HAs. (A), (B) This figure is analogous to that in Figure 5 except that it scores the H3 sequences using experimental measurements made on the lab-adapted WSN/1933 HA (16). [*Make into two vertical panels*]

between H1 and H3 reveals that the replicate measurements within a homolog are more correlated than between homologs (Figure 8B).

However, to quantify the extent of mutational shifts at a site-by-site level, we used an approach described in (52, 53). We aligned Perth/2009 H3 and WSN/1933 H1 and at each alignable site calculated the difference in preferences between homologs while correcting for experimental noise within homologs. The distribution of shifts is shown in Figure 8C. Although many sites have small shifts near zero, a considerable number of sites have large shifts in preference, reaching a difference of 0.86 out of a maximum possible of 1.0. When we compare HA to the preferences measured in the non-homologous protein HIV Env (53), nearly all sites are shifted. On the other hand, when we randomize the HA replicates to generate a null distribution, there is very little shift in preferences. For comparison, the preferences of the HIV Env homologs, which are 86% identical at the amino-acid level, are mostly similar. Although there are a small number of sites with larger shifts reaching a maximum of 0.52, we do not see a dramatic tail in the distribution as we see for the HA's. These observations suggest that as protein homologs diverge, their preferences increasingly shift.

We next asked if shifted sites tend to cluster together in the HA structure. However, the sites of large shifts were interspersed across HA and did not seem to obviously localize together (Figure 9A). Yet, at the domain level, we found the stalk domain to be significantly less shifted than the globular head domain (Figure 9B). This is unsurprising given that the stalk domain is more conserved both within and across HA subtypes (54–56). Sites absolutely conserved across all 18 HA subtypes were also found to be significantly less shifted than sites not conserved (Figure 9B), which is suggestive of the high functional importance of the sites that have remained unchanged throughout the divergence of HA.

Despite their large sequence divergence, H1 and H3 adopt nearly identical protein folds (57, 58). Previous work has found that although the subdomain structures are highly similar, there are differences in the rotation and upward translation of the globular head subdomains relative to the central stalk domain among the four HA clades represented by H1 & H5, H9, H7, and H3 (57, 58). One explanation for such clade-specific subdomain position differences could be local and/or long-range conformational interactions between sites that are variable between clades [*needs better wording*]. We would expect such sites to exhibit large shifts in preference. Indeed, clade-specific sites that are conserved within the H1 clade (including H1, H2, H5, and H6) or within the H3 clade (including H3, H4, and H14) are significantly shifted compared to non-clade-specific sites.

In particular, the clade-specific upward shift of the globular head relative to the stalk has largely been attributed to the interaction between sites 107 and 75(HA2) (57, 58). The H1 HA has a taller turn in the interhelical loop connecting helix A and helix B in the stalk domain, and this tall turn is stabilized by a hydrogen bond between Glu-107 and Lys-75(HA2) (Figure 9C). In the WSN/1933 HA, site 107 has a high preference for Glu and 75(HA2) strongly prefers positively charged Lys and Arg. In contrast, the interhelical loop in H3 HA makes a sharper and shorter turn which is facilitated by a Gly at 75(HA2). This site prefers Gly and to a lesser extent

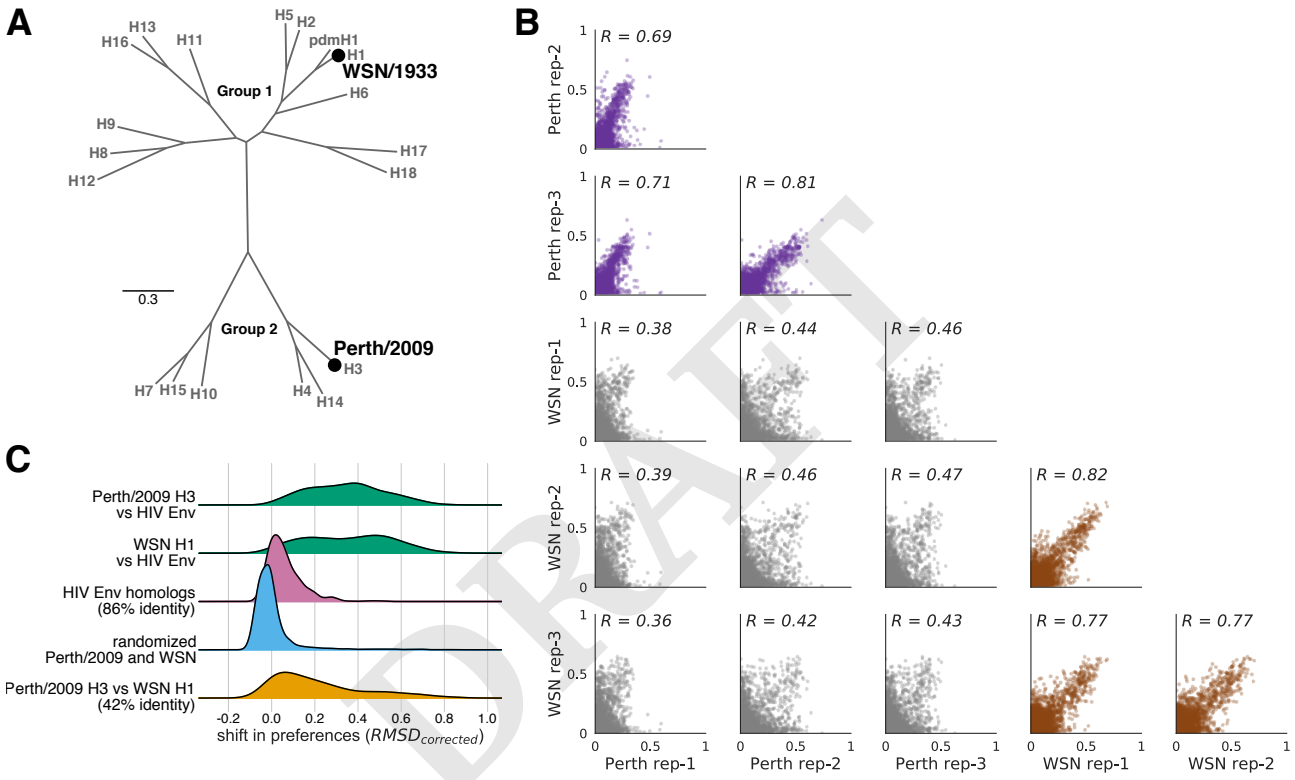

Fig. 8. **There are substantial differences in the effects of mutations between H1 and H3 HAs.** (A) Phylogenetic tree of HA subtypes, with the WSN/1933 H1 and Perth/2009 H3 HAs labeled. These HAs have ∼42% amino-acid identity[*get the exact number rounded the percent and remove the ∼ symbol*]. (B) All pairwise correlations of the amino-acid preferences measured in the three individual deep mutational scanning replicates in the current study and the three replicates in prior deep mutational scanning of an H1 HA (16). Comparisons between H3 replicates are in purple, those between H1 replicates are in brown, and those across H1 and H3 replicates are in gray. R indicates the Pearson correlation coefficient. (C) We calculated the shift in amino-acid preferences at each site between H3 and H1 HAs using the method in (53), and plotted the distribution of shifts for all sites. The shifts between H3 and H1 (yellow) are much larger than the null distribution (blue) expected if all differences to experimental noise. The shifts are also much larger than those previously observed between two variants of HIV envelope protein (Env) that share 86% amino-acid identity (pink). However, the shifts between H3 and H1 are still less than the differences between either HA and HIV Env (green). [*Let's remove RMSDcorrected from legend.*]

Val, while site 107 is fairly tolerant of mutations suggesting that the amino-acid identity at this site plays a smaller role in determining the conformation of the interhelical loop in H3.

Discussion

We have measured the effect of all possible single amino-acid mutations to Perth/2009 H3 on viral growth in cell culture.

Materials and Methods

Please describe your materials and methods here. This can be more than one paragraph, and may contain subsections and equations as required. Authors should include a statement in the methods section describing how readers will be able to access the data in the paper.

HA numbering. Unless otherwise indicated, all sites are in H3 numbering, with the signal peptide in negative numbers, the HA1 subunit in plain numbers, and the HA2 subunit denoted with "(HA2)". The conversion between sequential numbering of the A/Perth/16/2009 HA and H3 numbering was performed using an HA numbering Python script (available at https://github.com/jbloomlab/HA_numbering).

Creation of MDCK-SIAT1-TMPRSS2 cell line. The human TMPRSS2 cDNA ORF was ordered from OriGene (NM_005656), PCR amplified, and cloned into a pHAGE2 lentiviral vector under an EF1α-Int promoter and attached to mCherry through an IRES...etc etc [*Need to look at Katie's notebooks for this...*]

Generation of HA codon mutant plasmid libraries. Recombinant A/Perth/16/2009 (HA, NA) $\times$ A/Puerto Rico/8/1934 influenza virus, NIB-64, NR-41803 was ordered from BEI Resources, NIAID, NIH. Bulk RNA from the viral sample was extracted using the QIAamp Viral RNA Mini Kit (QIAGEN) according to manufacturer's instructions. The Perth/2009 HA and NA genes were then reverse transcribed, PCR amplified, and cloned into the pHW2000 (59) and pICR2 [*cite?*] plasmid backbones.

The codon-mutant libraries were generated using a PCR-based approach described in (60, 61).

Generation and passaging of mutant viruses. The mutant virus libraries were generated and passaged using the approach described in (16) with several modifications.

Barcoded subamplicon sequencing.

Analysis of deep sequencing data.

Inference of phylogenetic trees. [*We downloaded X sequences from the Influenza Virus Resource ?.... etc. inferred the tree, ancestral state reconstruction, visualized the tree...*] To parse out trunk mutations from side branch mutations, we first defined a set of recent nodes sampled on or after Jan. 1, 2017, and traced these nodes back to their most recent common ancestor.

Quantification of mutational effects and sequence preferences from an H3N2 phylogeny.

Data availability and source code. Deep sequencing data are available from the Sequence Read Archive under BioSample accessions SAMN08102609 and SAMN08102610. Computer code used to analyze the data and produce the results in the paper are in...

ACKNOWLEDGMENTS. We thank Sarah Hilton, Hugh Haddox, Sidney Bell...the Fred Hutch Genomics Core... Funding...

References.

1. 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. USA* 94(15):7712–7718.
2. Strelkowa N, Lässig M (2012) Clonal interference in the evolution of influenza. *Genetics* 192(2):671–682.
3. Bedford T, et al. (2015) Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* 523(7559):217–220.
4. Fitch WM, Leiter J, Li X, Palese P (1991) Positive Darwinian evolution in human influenza A viruses. *Proc. Natl. Acad. Sci. USA* 88(10):4270–4274.
5. Steinbrück L, Klingen T, McHardy A (2014) Computational prediction of vaccine strains for human influenza A (H3N2) viruses. *Journal of Virology* 88(20):12123–12132.
6. Neher RA, Russell CA, Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. *Elife* 3:e03568.
7. Sun H, et al. (2013) Using sequence data to infer the antigenicity of influenza virus. *MBio* 4(4):e00230–13.
8. Harvey WT, et al. (2016) Identification of low-and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza A (H1N1) viruses. *PLoS Pathogens* 12(4):e1005526.
9. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI (2016) Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proceedings of the National Academy of Sciences* 113(12):E1701–E1709.
10. Pybus OG, et al. (2007) Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Molecular Biology and Evolution* 24(3):845–852.
11. Kucharski A, Gog JR (2011) Influenza emergence in the face of evolutionary constraints. *Proceedings of the Royal Society of London B: Biological Sciences* p. rspb20111168.
12. Łuksza M, Lässig M (2014) A predictive fitness model for influenza. *Nature* 507(7490):57–61.
13. Koelle K, Rasmussen DA (2015) The effects of a deleterious mutation load on patterns of influenza A/H3N2's antigenic evolution in humans. *Elife* 4:e07361.
14. Fowler DM, et al. (2010) High-resolution mapping of protein sequence-function relationships. *Nat. Methods* 7(9):741–746.
15. Thyagarajan B, Bloom JD (2014) The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife* 3:e03300.
16. Doud MB, Bloom JD (2016) Accurate measurement of the effects of all amino-acid mutations to influenza hemagglutinin. *Viruses* 8:155.
17. WHO (2010) Recommended viruses for influenza vaccines for use in the 2010-2011 northern hemisphere influenza season. *http://www.who.int/influenza/vaccines/virus/recommendations/201002_Recommendation.pdf?ua=1*.
18. WHO (2011) Recommended composition of influenza virus vaccines for use in the 2011-2012 northern hemisphere influenza season. *http://www.who.int/influenza/vaccines/2011_02_recommendation.pdf?ua=1*.
19. Böttcher E, et al. (2006) Proteolytic activation of influenza viruses by serine proteases TMPRSS2 and HAT from human airway epithelium. *Journal of Virology* 80:9896–9898.
20. Böttcher-Friebertshäuser, E, et al. (2010) Cleavage of influenza virus hemagglutinin by airway proteases TMPRSS2 and HAT differs in subcellular localization and susceptibility to protease inhibitors. *Journal of Virology* 11:5605–5614.
21. Marshall N, Priyamvada L, Ende Z, Steel J, Lowen AC (2013) Influenza virus reassortment occurs with high frequency in the absence of segment mismatch. *PLoS Pathogens* 9(6):e1003421.
22. Bloom JD (2015) Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics* 16(1):1.
23. Bloom JD (2017) Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology Direct* 12(1):1.
24. Hilton SK, Doud MB, Bloom JD (2017) phydms: Software for phylogenetic analyses informed by deep mutational scanning. *PeerJ* 5:e3657.
25. Yang Z, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155(1):431–449.
26. Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology* 53(5):793–808.
27. Waterfield M, Scrace G, Skehel J (1981) Disulphide bonds of haemagglutinin of Asian influenza virus. *Nature* 289:422–424.
28. Weis W, et al. (1988) Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature* 333:426–431.
29. Martin J, et al. (1998) Studies of the binding properties of influenza hemagglutinin receptor-site mutants. *Virology* 241(1):101–111.
30. Nobusawa E, Ishihara H, Morishita T, Sato K, Nakajima K (2000) Change in receptor-binding specificity of recent human influenza A viruses (H3N2): a single amino acid change in hemagglutinin altered its recognition of sialyloligosaccharides. *Virology* 278(2):587–596.
31. Kido H, et al. (1992) Isolation and characterization of a novel trypsin-like protease found in rat bronchiolar epithelial Clara cells. A possible activator of the viral fusion glycoprotein. *J Biol Chem* 267:13573–13579.
32. Stech J, Garn H, Wegmann M, Wagner R, Klenk H (2005) A new approach to an influenza live vaccine: modification of the cleavage site of hemagglutinin. *Nature Medicine* 11(6):683–689.
33. Girard G, Gultyaev A, Olsthoorn R (2011) Upstream start codon in segment 4 of North American H2 avian influenza A viruses. *Infect. Genet. Evol.* 11:489–495.
34. Wolf Y, Viboud C, Holmes E, Koonin E, Lipman D (2006) Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biology Direct* 1:34.
35. Ekiert DC, et al. (2011) A highly conserved neutralizing epitope on group 2 influenza A viruses. *Science* 333(6044):843–850.
36. Friesen R, et al. (2014) A common solution to group 2 influenza virus neutralization. *Proc. Natl. Acad. Sci. USA* 111:445–450.
37. Chai N, et al. (2016) Two escape mechanisms of influenza A virus to a broadly neutralizing stalk-binding antibody. *PLoS Pathogens* 12(6):e1005702.

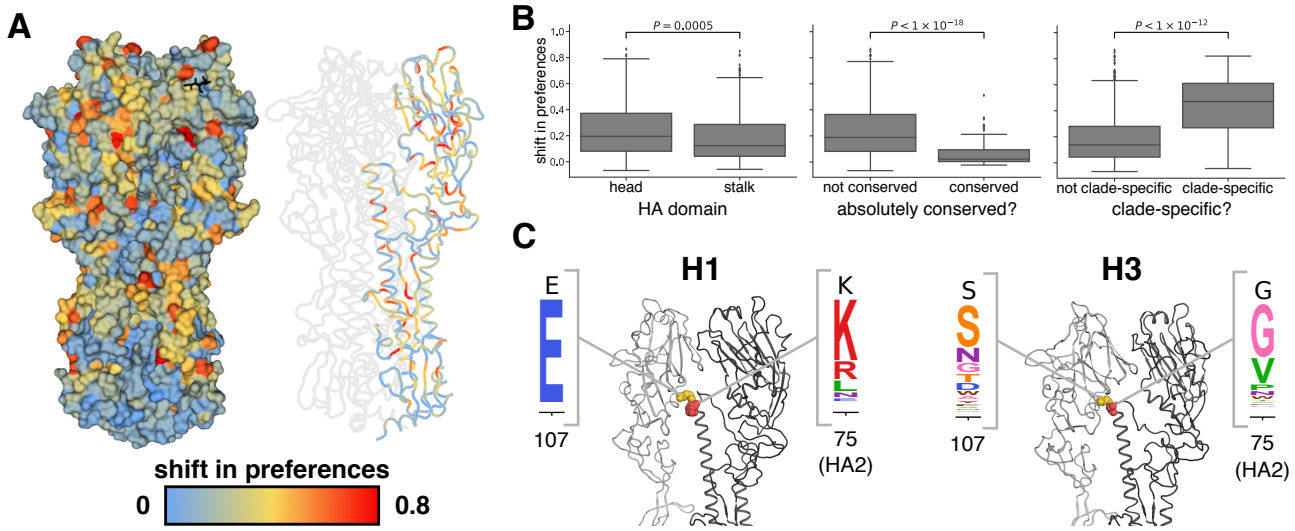

Fig. 9. Characterization of the preference shifts (A) The preference shifts as calculated by $RMSD_{corrected}$ between the two HA homologs is mapped onto the structure of HA (PDB 4O5N; (46)). The left structure shows the HA trimer, and the right structure colors one of the monomers. The sialic acid receptor is shown in black sticks. Blue indicates small shifts in preference near zero, while red indicates large shifts in preference. (B) The stalk domain was found to be significantly less shifted than the head domain (left plot). Sites absolutely conserved all 18 HA subtypes were also found to be significantly less shifted than the remaining non-conserved sites (middle plot). Sites specific to the clade containing H1, H2, H5, and H6 and specific to the H3, H4, and H14 clade are significantly more shifted than non-clade-specific sites (right plot). (C) HA group-specific interactions between sites 107 and 75(HA2) play a role in determining the upward translation of the globular head domain relative to the stalk domain. The amino acids at sites 107 and 75(HA2) are shown in spheres on the structure of H1 and H3. One monomer is shown in dark gray, while the HA1 domain of the neighboring monomer is in lighter gray. The sphere colors correspond to the shift in preference at that site. The preferences for each site in WSN/1933 and Perth/2009 are also shown.

38. Yamayoshi S, et al. (2017) Human protective monoclonal antibodies against the HA stem of group 2 HAs derived from an H3N2 virus-infected human. *Journal of Infection*.
39. Okuno Y, Isegawa Y, Sasao F, Ueda S (1993) A common neutralizing epitope conserved between the hemagglutinins of influenza A virus H1 and H2 strains. *Journal of Virology* 67(5):2552–2558.
40. Doud MB, Lee JM, Bloom JD (2017) Quantifying the ease of viral escape from broad and narrow antibodies to influenza hemagglutinin. *bioRxiv* p. 210468.
41. Wiley D, Wilson I, Skehel J, , et al. (1981) Structural identification of the antibody-binding sites of hong kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* 289(5796):373–378.
42. Chambers BS, Parkhouse K, Ross TM, Alby K, Hensley SE (2015) Identification of hemagglutinin residues responsible for H3N2 antigenic drift during the 2014–2015 influenza season. *Cell Reports* 12(1):1–6.
43. Koel BF, et al. (2013) Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* 342(6161):976–979.
44. Popova L, et al. (2012) Immunodominance of antigenic site B over site A of hemagglutinin of recent H3N2 influenza viruses. *PLoS One* 7(7):e41895.
45. Wilson I, Skehel J, Wiley D (1981) Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature* 289:366–373.
46. Lee PS, et al. (2014) Receptor mimicry by antibody F045-092 facilitates universal binding to the H3 subtype of influenza virus. *Nat Commun* 5:3614.
47. Gamblin S, et al. (2004) The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science* 303(5665):1838–1842.
48. Wu N, et al. (2017) A structural explanation for the low effectiveness of the seasonal influenza H3N2 vaccine. *PLoS Pathogens* 13(10):e1006682.
49. McWhite C, Meyer A, Wilke C (2016) Sequence amplification via cell passaging creates spurious signals of positive adaptation in influenza virus H3N2 hemagglutinin. *Virus Evolution* 2:vew026.
50. Skowronski D, 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.
51. Wu NC, et al. (2017) Diversity of functionally permissive sequences in the receptor-binding site of influenza hemagglutinin. *Cell Host & Microbe* 21(6):742–753.
52. Doud MB, Ashenberg O, Bloom JD (2015) Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol. Biol. Evol.* 32:2944–2960.
53. Haddox HK, Dingens AS, Hilton SK, Overbaugh J, Bloom JD (2017) Mapping mutational effects along the evolutionary landscape of HIV envelope. *bioRxiv* p. 235630.
54. Nobusawa E, et al. (1991) Comparison of complete amino acid sequences and receptor-binding properties among 13 serotypes of hemagglutinins of influenza A viruses. *Virology* 182(2):475–485.
55. Hai R, et al. (2012) Influenza viruses expressing chimeric hemagglutinins: globular head and stalk domains derived from different subtypes. *Journal of Virology* 86(10):5774–5781.
56. Mallajosyula VV, et al. (2014) Influenza hemagglutinin stem-fragment immunogen elicits broadly neutralizing antibodies and confers heterologous protection. *Proc. Natl. Acad. Sci. USA* 111(25):E2514–E2523.
57. Ha Y, Stevens DJ, Skehel JJ, Wiley DC (2002) H5 avian and H9 swine influenza virus haemagglutinin structures: possible origin of influenza subtypes. *The EMBO Journal* 21(5):865–875.
58. Russell R, et al. (2004) H1 and H7 influenza haemagglutinin structures extend a structural classification of haemagglutinin subtypes. *Virology* 325(2):287–296.
59. Hoffmann E, Neumann G, Kawaoka Y, Hobom G, Webster RG (2000) A DNA transfection system for generation of influenza A virus from eight plasmids. *Proc. Natl. Acad. Sci. USA* 97:6108–6113.
60. Bloom JD (2014) An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular Biology and Evolution* 31:1956–1978.
61. Dingens AS, Haddox HK, Overbaugh J, Bloom JD (2017) Comprehensive mapping of HIV-1 escape from a broadly neutralizing antibody. *Cell Host & Microbe* 21:777–787.

Supporting Information (SI). The main text of the paper must stand on its own without the SI. Refer to SI in the manuscript at an appropriate point in the text. Number supporting figures and tables starting with S1, S2, etc. Authors are limited to no more than 10 SI files, not including movie files. Authors who place detailed materials and methods in SI must provide sufficient detail in the main text methods to enable a reader to follow the logic of the procedures and results and also must reference the online methods. If a paper is fundamentally a study of a new method or technique, then the methods must be described completely in the main text. Because PNAS edits SI and composes it into a single PDF, authors must provide the following file formats only.

SI Text. Supply Word, RTF, or LaTeX files (LaTeX files must be accompanied by a PDF with the same file name for visual reference).

SI Figures.

SI Tables. Supply Word, RTF, or LaTeX files (LaTeX files must be accompanied by a PDF with the same file name for visual reference); include only one table per file. Do not use tabs or spaces to separate columns in Word tables.

SI Datasets.

Dataset S1. Genbank file giving the full sequence of the bidirectional reverse-genetics plasmid pHW-Perth2009-HA-G78D-T212I, which encodes the wildtype HA sequence used in this study.

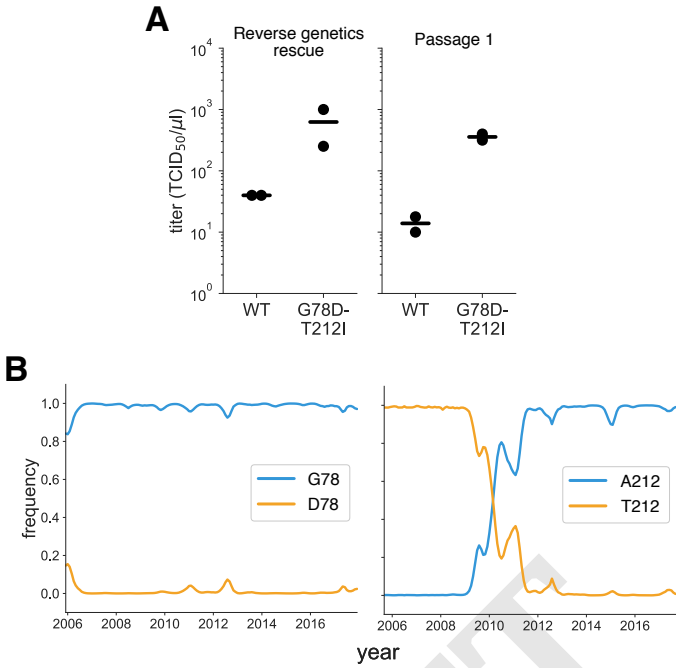

Figure S1. **Characterization of the G78D-T212I Perth/2009 HA variant.** (A) The G78D-T212I Perth/2009 HA variant supports better viral growth than the wildtype Perth/2009 HA. Viruses were generated in duplicate by reverse genetics with the Perth/2009 NA and WSN internal genes, and passaged once at MOI = 0.01 in MDCK-SIAT1-TMPRSS2 cells. The rescue and passage viral supernatants were collected at 72 hours post-transfection and 44 hours post-infection, respectively, and titered in MDCK-SIAT1-TMPRSS2 cells. The points mark each duplicate and the bar marks the mean. (B) The D78 variant remained at a low frequency in natural human H3N2 sequences over the past ∼10 years. The A212 variant rose to fixation in ∼2011, replacing the T212 variant.

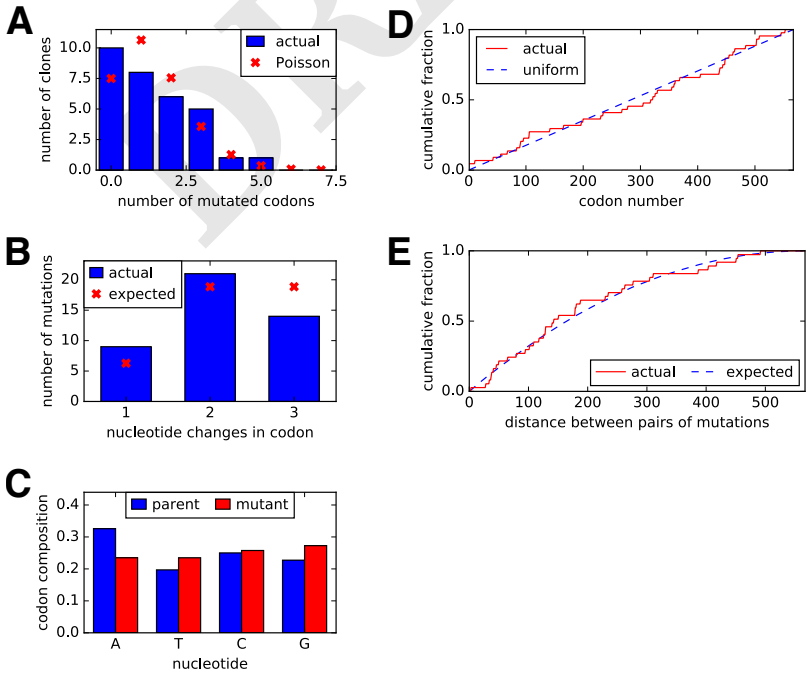

Figure S2. **Sanger sequencing of 31 clones from the mutant plasmid libraries.** (A) There were an average of ∼1.4 codon mutations per clone across the three plasmid mutant libraries. (B) A mixture of one-, two-, and three-nucleotide mutations were present, with slightly fewer triple-nucleotide changes than expected. (C) Nucleotide frequencies were uniform in the codon mutations. (D) The mutations were distributed relatively evenly across the length of the HA coding sequence. (E) We calculated the pairwise distances between mutations for clones carrying more than one mutation. The cumulative distribution of the distances is shown in the red line.

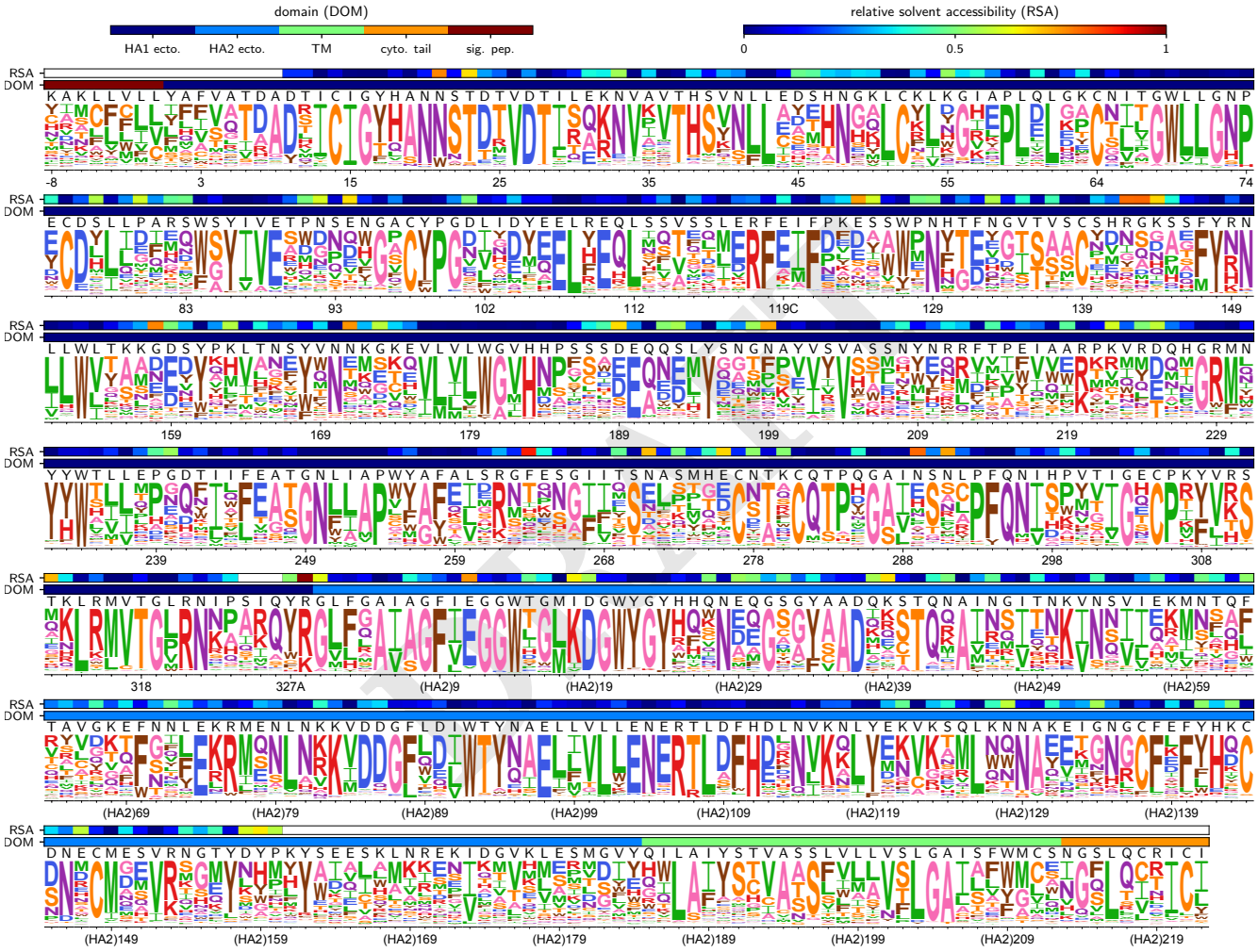

Figure S3. The site-specific amino-acid preferences of the WSN/1933 H1 HA.