

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

Juhye M. Lee<sup>a,d,e,1</sup>, John Huddleston<sup>b,f,1</sup>, Michael B. Doud<sup>a,d,e</sup>, Kathryn A. Hooper<sup>a,f</sup>, Trevor Bedford<sup>b,c,2</sup>, and Jesse D. Bloom<sup>a,c,d,2</sup>

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

This manuscript was compiled on January 5, 2018

Please provide an abstract of no more than 250 words in a single paragraph. Abstracts should explain to the general reader the major contributions of the article. References in the abstract must be cited in full within the abstract itself and cited in the text.

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

**V**ery rough outline:

- Mutations are rampant in the evolution of human influenza virus. Seasonal H3N2 influenza virus in particular rapidly accumulates mutations.
- The evolution of H3N2 is also characterized by clade competition and population turnover.
- There have been efforts to predict evolutionary success.
- Mutations that contribute to antigenic evolution largely determine strain success.
- RNA viruses can accumulate deleterious mutations, and because the influenza does not appreciably recombine, deleterious mutations are linked to beneficial ones.
- Mutations in HA that impact viral growth may influence strain success.
- We need to understand the functional impact of mutations in HA.
- Previously, we measured the effect of all possible single amino-acid mutations to an H1 hemagglutinin from the A/WSN/1933 (H1N1) strain (1, 2).
- However, this is a highly lab-adapted strain, and the measurements in this strain may not be particularly relevant for studying mutational processes of more contemporaneous strains circulating in the human population.
- We chose to study the Perth/2009 H3 HA.
- This also enabled a comparison of how the preferences have shifted for two diverged HA's.

## 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 (3, 4). 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 (2) (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 (5, 6).

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

## Significance Statement

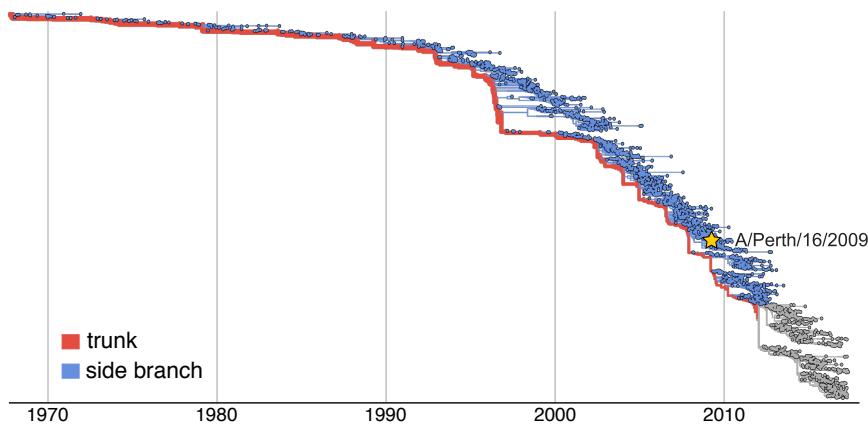
Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the Significance Statement is to explain the relevance of the work in broad context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

<sup>1</sup>J.M.L. and J.H. contributed equally to this work.

<sup>2</sup>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.

(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 [*let's cite some paper here, maybe Chris Brooke one?*] 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.

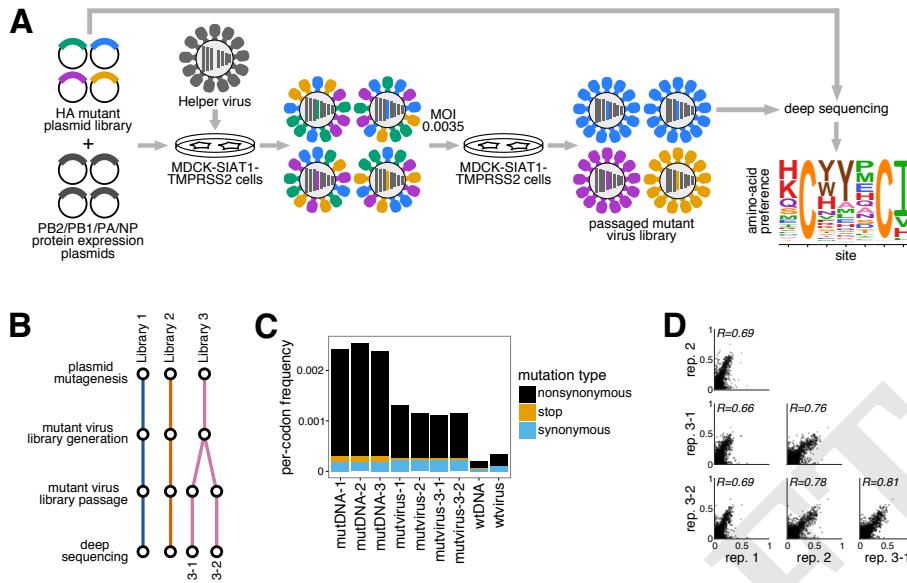
We next quantified the reproducibility of our deep mutational scanning measurements across biological and technical replicates. We first used the deep sequencing data for each replicate to independently estimate the preference of each site in HA for all 20 amino acids using the method described in (7). 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 (8). Table 3 shows that the ExpCM greatly outperforms conventional substitution models, showing that our experiments authentically capture some of the constraints on HA evolution. [*Jesse and Juhye are here in sentence-level editing*] The ExpCM also optimizes a stringency parameter for the preferences to more closely reflect the strength of selection in nature. The stringency parameter in the ExpCM is equal to 2.44, which indicates that although the same amino acids are preferred, the strength of selection is more stringent in nature than in our experiments. Figure 4 shows a logo plot of the Perth/2009 HA amino-acid preferences rescaled by this stringency parameter.

A closer examination of the logo plot reveals that the preferences generally agree with existing knowledge about HA's biochemistry. 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) (11) possess high preference for cysteine. At residues involved in receptor binding, there are strong preferences for the amino acids at sites Y98, D190, W153, and S228 (12, 13). 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 (14, 15). 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 explanation for why we do not see a strong preference for Met at this site is due to alternative translation initiation occurring at a downstream or upstream start site, as has been described for HA (16).

**The mutational tolerance of the head and the stalk shows less of a contrast in H3 than in H1.** We next sought to investigate the inherent mutational tolerance of the Perth/2009 HA. Figure 5 shows the mutational tolerance as calculated from the rescaled Perth/2009 H3 preferences and the rescaled WSN/1933 H1 preferences mapped onto the HA crystal structures. We found antigenic site C and the most distal portion of the globular head near the 190-helix in the Perth/2009 H3 to be tolerant of mutations. Interestingly, the H3 stalk including the shorter  $\alpha$ -helix (helix A) is relatively mutationally tolerant compared to the tolerance of the globular head domain. This observation suggests that the stalk may be prone to escape from antibodies, and agrees with previous work demonstrating

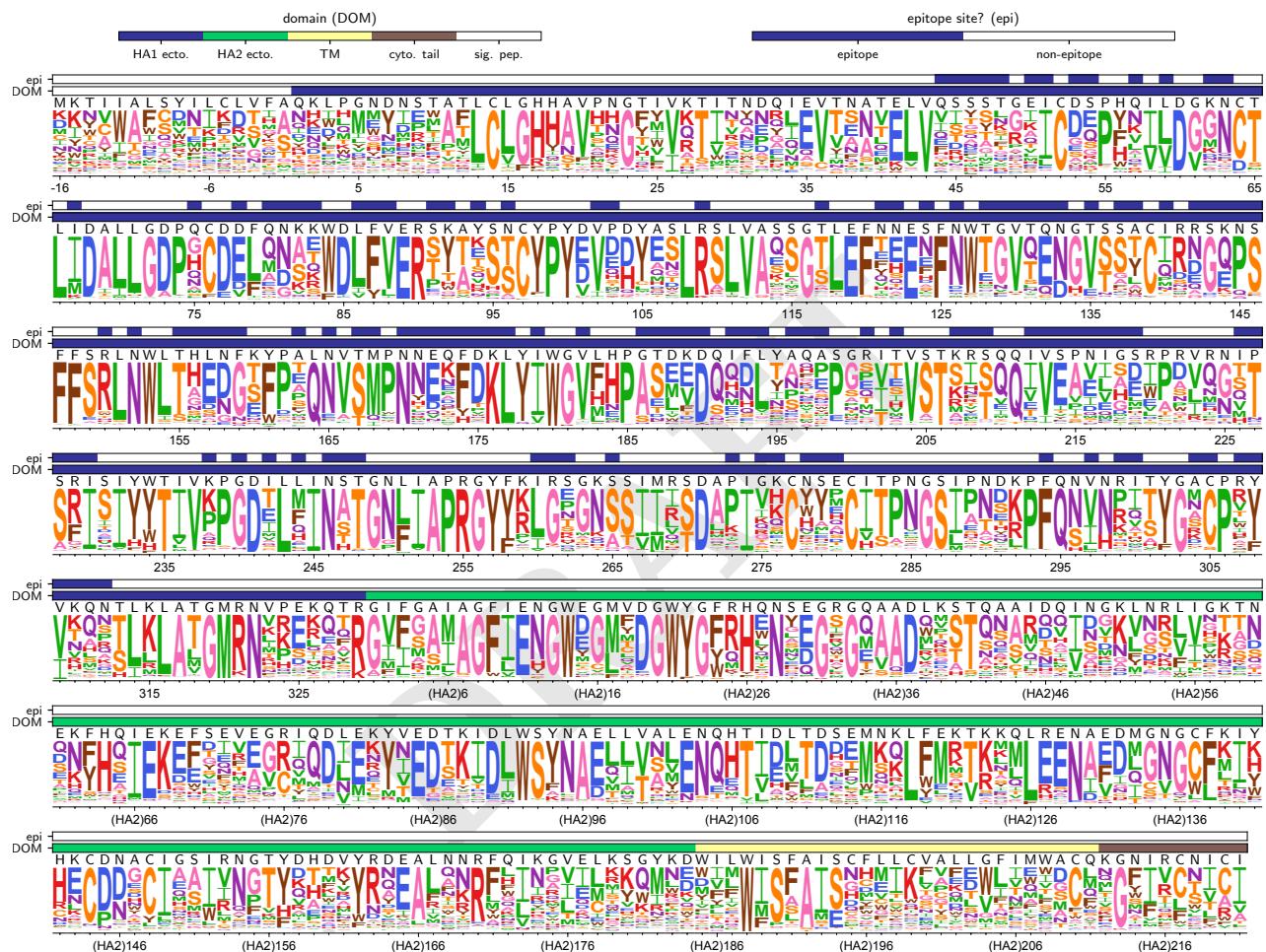


**Fig. 2. Deep mutational scanning of the Perth/2009 H3 HA.** (A) We generated mutant virus libraries using a helper-virus approach (2), and passaged them at low MOI to establish a genotype-phenotype linkage and to select for functional HA variants. Deep sequencing of the variants before and after selection 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.

**Fig. 3. The site-specific amino-acid preferences are informative for describing human H3N2 evolution in nature.**

Model	$\Delta \text{AIC}$	Log Likelihood	Stringency	$\omega$	$\bar{\omega}$	$\omega_\alpha$	$\omega_\beta$
ExpCM	0.0	-8439.33	2.44	0.91	—	—	—
Goldman-Yang M5	2166.06	-9516.36	—	—	0.36	0.30	0.84
ExpCM, averaged across sites	2504.18	-9691.42	0.68	0.32	—	—	—
Goldman-Yang M0	2607.92	-9738.29	—	0.31	—	—	—

We implemented several codon substitution models for phylogenetic fitting of an alignment of human H3N2 HA sequences. The maximum likelihood values for each model were compared using the Akaike information criteria ( $\Delta \text{AIC}$ ) (9). An experimentally-informed codon substitution model (ExpCM) built from the preferences averaged across all replicates performs better than conventional substitution models, specifically the M0 and M5 models in (10). A non-site-specific ExpCM informed by preferences averaged across all sites performs comparably to the GY94 class of models, indicating the importance of site-specificity in the ExpCM. The optimized parameters for each model are also shown.



**Fig. 4.** The site-specific amino-acid preferences of the Perth/2009 HA. This logplot shows the site-specific amino-acid preferences for the averaged replicates rescaled by the stringency parameter (Table 1) estimated by phydms. The height of each letter is proportional to its preference at that site, and the preferences for all sites are normalized to sum to 1. The sites are in H3 numbering. The top overlay bar shows the relative solvent accessibility. The bottom overlay bar is colored by 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 indicate the wildtype amino acid at that site.

that it is possible to select for antigenic mutants in H3 by broadly-neutralizing stalk-targeting antibodies (17–20).

The sites inside the receptor binding pockets are highly functionally constrained and were found to be relatively mutationally intolerant in both H3 and H1 (21). In contrast, the residues surrounding the receptor binding pocket are fairly mutationally tolerant, which may contribute to antigenic evolution as these sites are under strong immune pressure. (22).

**The experimental measurements can help discriminate successful influenza virus lineages.** A challenge in vaccine strain selection is predicting which strain will dominate the upcoming influenza season. An important question to thereby address is if our experimental measurements are useful in differentiating between strains of human H3N2 influenza virus that have succeeded and those that have died out. To investigate this, we used our preference dataset and an H3N2 phylogeny from 1968–2012 (Figure 6A) to calculate the effects of mutations for the evolutionarily successful trunk lineage and for side branches which have died out. We found that strains with mutations measured to be more beneficial to viral growth in our experiments tend to succeed in nature. Figure 6B shows the effects of trunk and side branch mutations in five-year intervals for every year from 1968–2012. On average, trunk mutations are towards more preferred amino acids compared to side branch mutations, and this was true for all intervals. Importantly, trunk mutations are significantly more favorable than side branch mutations when calculating the effects from the entirety of the phylogenetic tree (Figure 6C).

Because tip nodes can contain egg- or cell-passaged isolates (25–27) and our experiments were performed in cell culture, we examined the effects of mutations on terminal side branch nodes to see if these would rank more highly than internal node mutations. Instead, terminal node mutations are on average towards less preferred amino acids than internal node mutations (Figure 6C), and both internal and terminal node mutations are significantly less favorable than those on the trunk. Therefore, strains that have accumulated mutations that we experimentally measured to be unpreferred generally die out in nature, while more favored mutations provide a selective advantage to the trunk. These findings demonstrate the importance of the functional impacts of HA mutations in determining strain success.

How distantly can the preferences be extended to describe differences in successful and unsuccessful strains? To explore this question, we scored the complete HA sequence of every node in the tree using the preferences by quantifying a *sequence preference* metric. The sequence preference is defined as  $\sum_r \ln \pi_{r,a}$  where  $\pi_{r,a}$  is the preference for amino acid  $a$

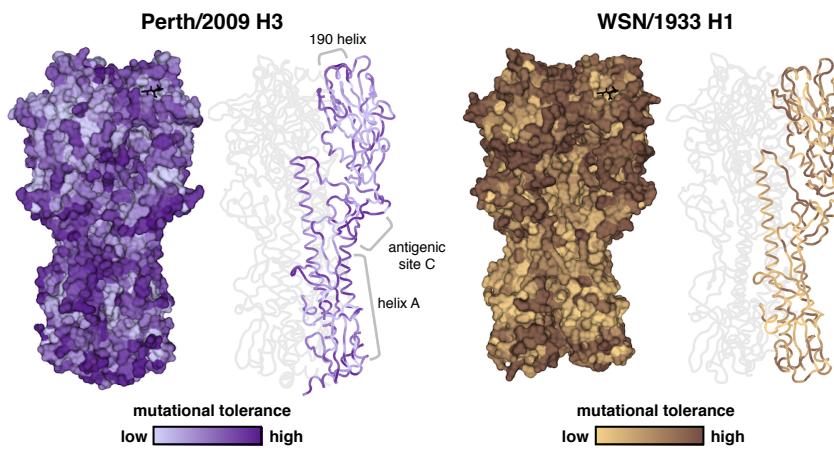
at site  $r$ . Consistent with the finding that trunk mutations are generally more favorable than side branch mutations, Figure 7 shows that the sequences of trunk nodes also tend to be more highly preferred than those of side branch nodes. Interestingly, the sequence preferences increase over time as the nodes approach the Perth/2009 strain and its closely related nodes, which all exhibit sequence preferences that are higher than that of the trunk. The observation that the strain in which we performed our deep mutational scan has one of the highest sequence preferences illustrates epistatic interactions among mutations such that an unpreferred mutation in one background may be preferred in another.

[this section will be fleshed out more once I figure out how to plot the substitutions from the root at epitope / non-epitope sites, as per Trevor's suggestion.] We reasoned that much of the increase in sequence preference toward Perth/2009 could be attributed to epitope sites, as most substitutions have occurred at sites under strong immune pressure. We therefore assessed each node's preference at epitope and non-epitope sites as defined by (28). Indeed, the preferences at epitope sites resemble those for all HA sites (Figure 7), with preferences increasing over time. At non-epitope sites, however, the trunk preferences remain relatively constant while side branch preferences tend to drop below the trunk. These observations highlight the extensive epistasis among epitope sites, as has even been noted for sites within the 220-loop surrounding the receptor binding pocket (29). Many successful epitope mutations are likely contingent upon the background in which they arise. Yet, we are able to distinguish trunk sequences as more favorable than side branch sequences, indicating that the preferences are still of utility over short evolutionary timescales.

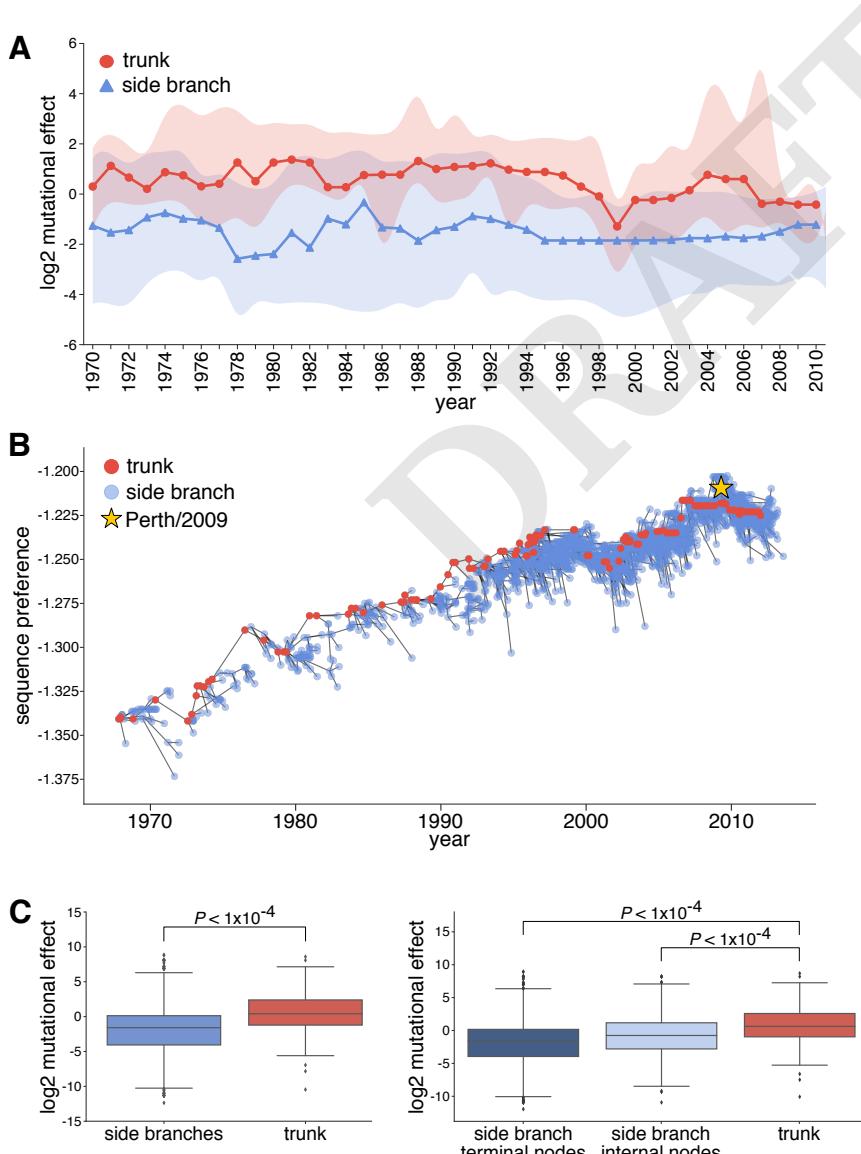
Can we then distinguish lineage-specific mutational effects using the preferences measured in a distantly related HA homolog? We used the preferences measured previously in the WSN/1933 H1 (2) to quantify the effects of H3 trunk and side branch mutations, shown in Figure 8. It is evident that we do not 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 preferences between related protein homologs of NP (30) and of Env [cite haddox2017], the HA homologs we have experimentally 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 9A), providing an opportunity to analyze what has shifted and what has remained conserved. Simply correlating the preferences between H1 and H3 reveals that the replicate measurements within a homolog are more correlated than between homologs (Figure 9B).

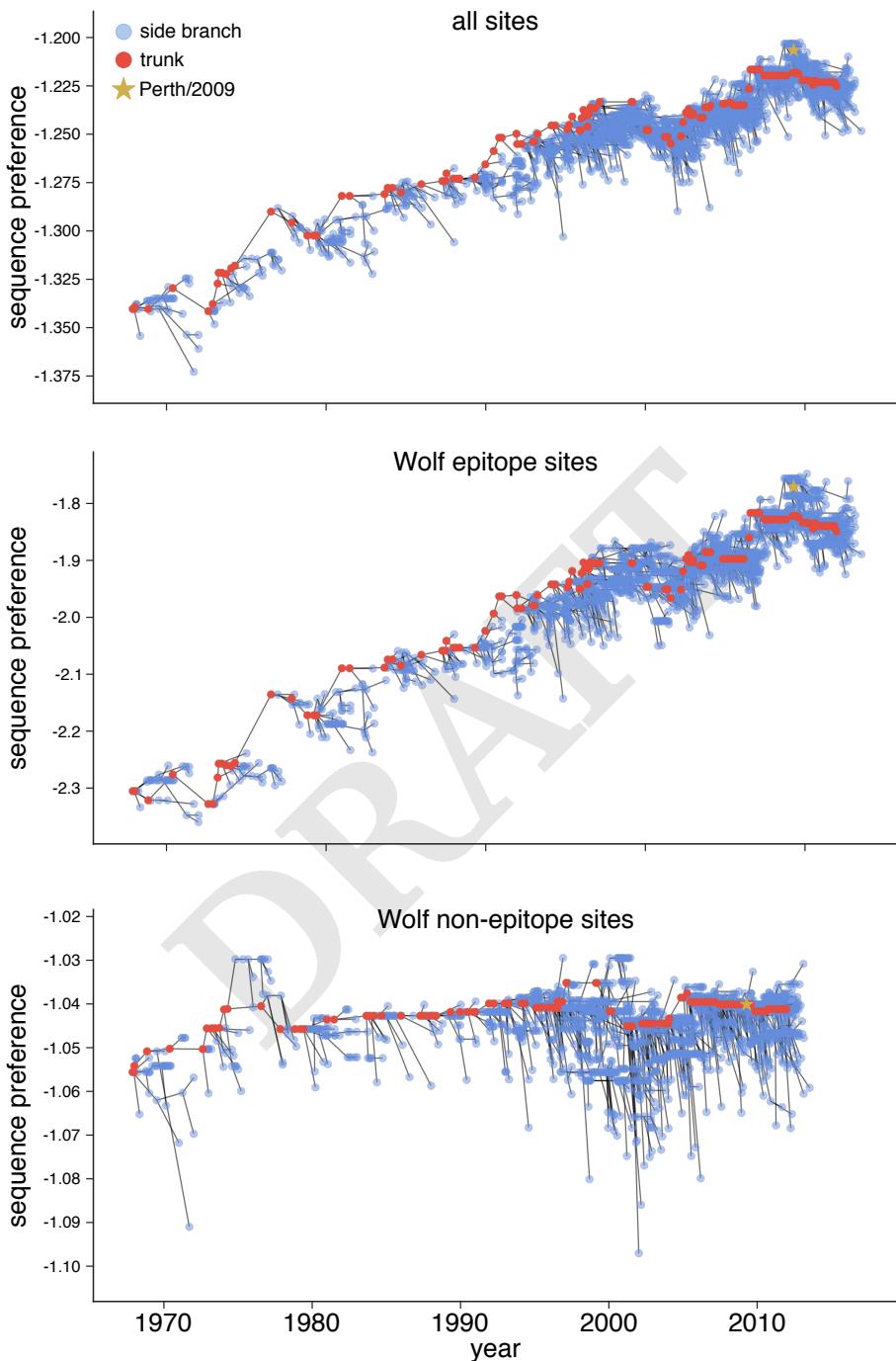
However, to quantify the extent of mutational shifts at a site-by-site level, we used an approach described in [haddox2017]. We aligned Perth/2009 H3 and WSN/1933 H1 and at each alignable site calculated the difference in preferences between homologs while correcting for experimental noise within homologs. The distribution of shifts is shown in Figure 9C. 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 [haddox2017], 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 homologs



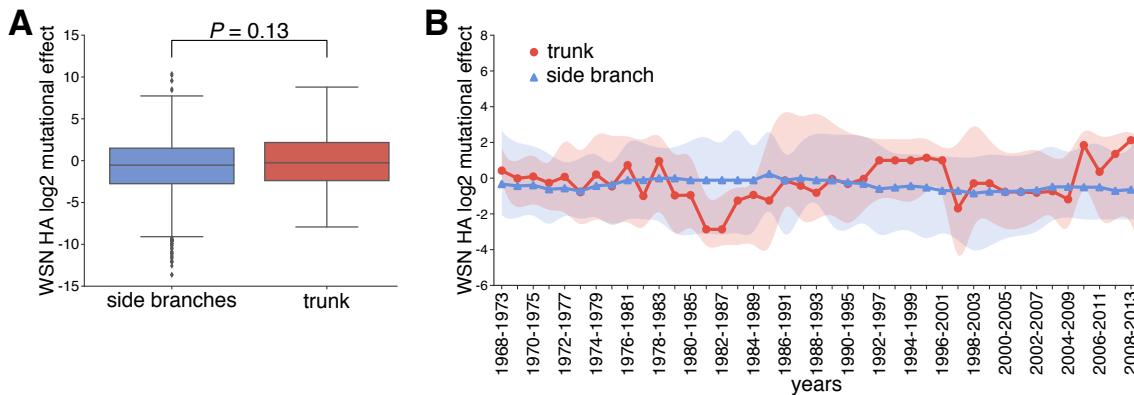
**Fig. 5. Mutational tolerance mapped onto each site of HA.** Mutational tolerance as calculated by the Shannon entropy of a given site's amino-acid preferences are mapped onto the structure of the H3 trimer (PDB 4O5N; (23)) and the H1 trimer (PDB 1RVX; (24)), with both trimers in approximately the same orientation. The site entropies were calculated from the preferences measured in the Perth/2009 H3 (left panel) from this study, or the preferences measured in the WSN/1933 H1 (right panel) from (2). Lighter shades of purple or brown signify low mutational tolerance, while darker shades signify high mutational tolerance. For each HA, the structure on the left side colors the full HA trimer, while the structure on the right side colors only one of the monomers. The sialic acid receptor is shown as black sticks. The Perth/2009 H3 shows relatively high mutational tolerance in the stalk region, particularly in helix A, compared to the head region. High mutational tolerance in H3 was also observed near antigenic site C and near the 190 helix. The head region of the WSN/1933 H1 is mutationally tolerant compared to the relatively intolerant stalk region.



**Fig. 6. Mutations in evolutionary successful strains tend to be more favorable than in strains that die out.** (A) Using the Perth/2009 H3 preferences, we calculated the  $\log_2$  mutational effect for trunk and side branch mutations in windows of 5 years for every year from 1968-2013. The median  $\log_2$  mutational effect in a given window is shown as circles for trunk mutations and triangles for side branch mutations. The shaded region demarcates the interquartile range of trunk and side branch mutational effects. Negative numbers signify mutations towards less preferred amino acids, while positive numbers signify more preferred mutations. The median trunk mutational effects are consistently higher than the median side branch mutational effects for all windows. (B) The sequence preference normalized by the number of sites of every node in the human H3N2 phylogenetic tree shown in Fig. 1. Trunk nodes are in red, side branch nodes in blue, and the Perth/2009 strain is marked with a star. Higher preferences are closer to zero while lower preferences are more negative. (C) The  $\log_2$  mutational effect for all side branch and all trunk mutations (left panel), in addition to all mutations in internal nodes and terminal nodes on the side branches (right panel) are shown. The preferences were randomized 10,000 times to estimate significance. The effects of trunk mutations are higher than side branch internal and terminal node mutations.



**Fig. 7. The HA sequences of trunk nodes tend to be more preferred than those of side branch nodes.** We used the preferences to calculate the HA sequence preference normalized by the number of sites of every node in a human H3N2 phylogenetic tree. To calculate the sequence preference, we used the entire HA sequence (top panel), only epitope sites (center panel) or non-epitope sites (bottom panel) as defined by (28). The preferences at all sites and at epitope sites increase as the tree approaches the Perth/2009 strain, indicating epistasis among epitope sites. The trunk nodes generally exhibit higher sequence preferences than do side branch nodes.



**Fig. 8.** The WSN/1933 H1 preferences do not reveal differences in trunk vs side branch mutational effects (A) We used the WSN/1933 H1 preferences from (2) to calculate the  $\log_2$  mutational effects of trunk and side branch mutations from the inferred H3N2 phylogeny in Figure 6. There is not a significant difference in trunk vs side branch mutational effects. (B) We also performed a sliding window analysis with the WSN/1933 H1 preferences. There is not a distinct difference in trunk and side branch mutational effects.

diverge, their preferences increasingly shift.

Upon mapping the shifts in preference onto the HA structure (Figure 10A), we did not see the shifted preferences obviously localize to specific regions in HA. However, we found the overall stalk domain to be significantly less shifted than the globular head domain (Figure 10B), which was expected given that the stalk domain is more conserved within and across HA subtypes. Sites absolutely conserved across all 18 HA subtypes were also found to be significantly less shifted than sites not conserved, which is suggestive of the high functional importance of the residues that have remained unchanged throughout the divergence of HA.

Despite their large sequence divergence, how is it that H1 and H3 adopt nearly identical protein folds [cite Ha, Russell]?

## 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](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 EF1 $\alpha$ -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 (?) and pICR2 [cite?] plasmid backbones.

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

**Generation and passaging of mutant viruses.** The mutant virus libraries were generated and passaged using the approach described in (2) with several modifications.

### Barcode 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. Mark Perth/2009 on 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.

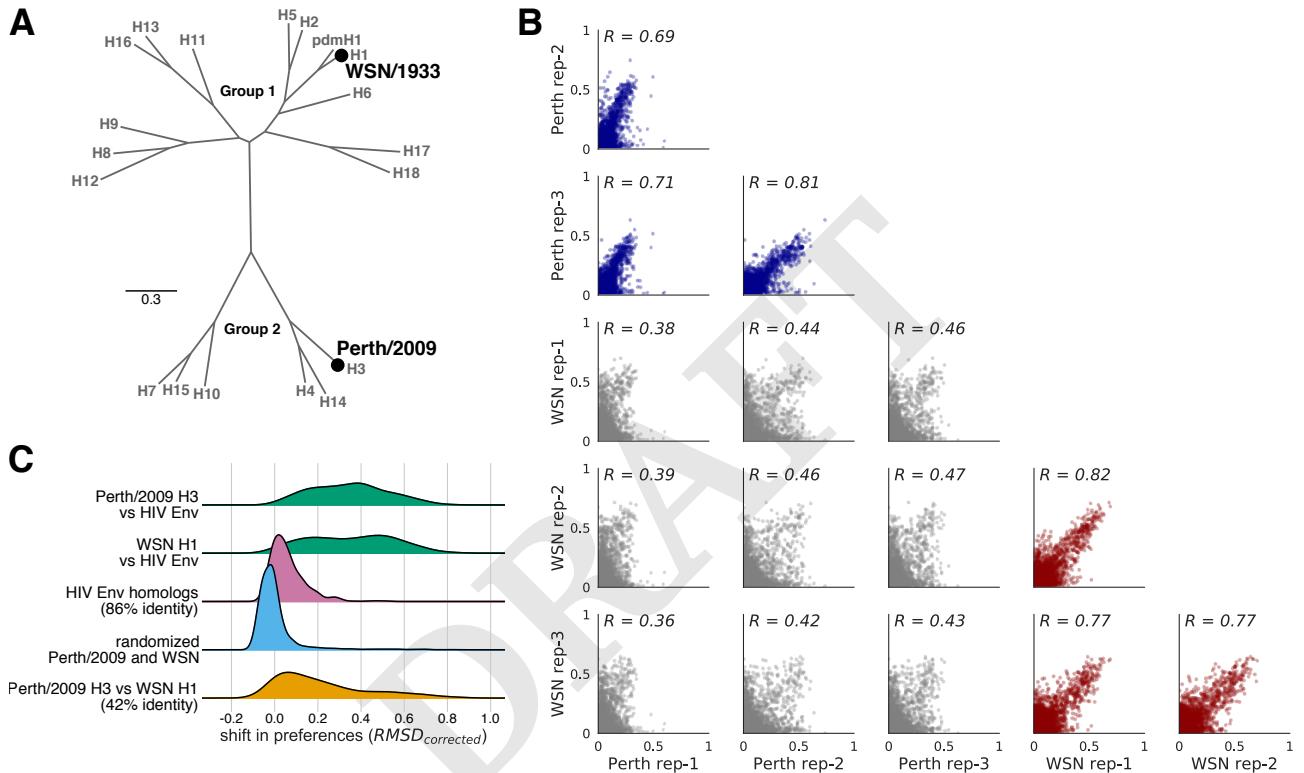
### Quantification of mutational effects and sequence preferences from an H3N2 phylogeny.

**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...

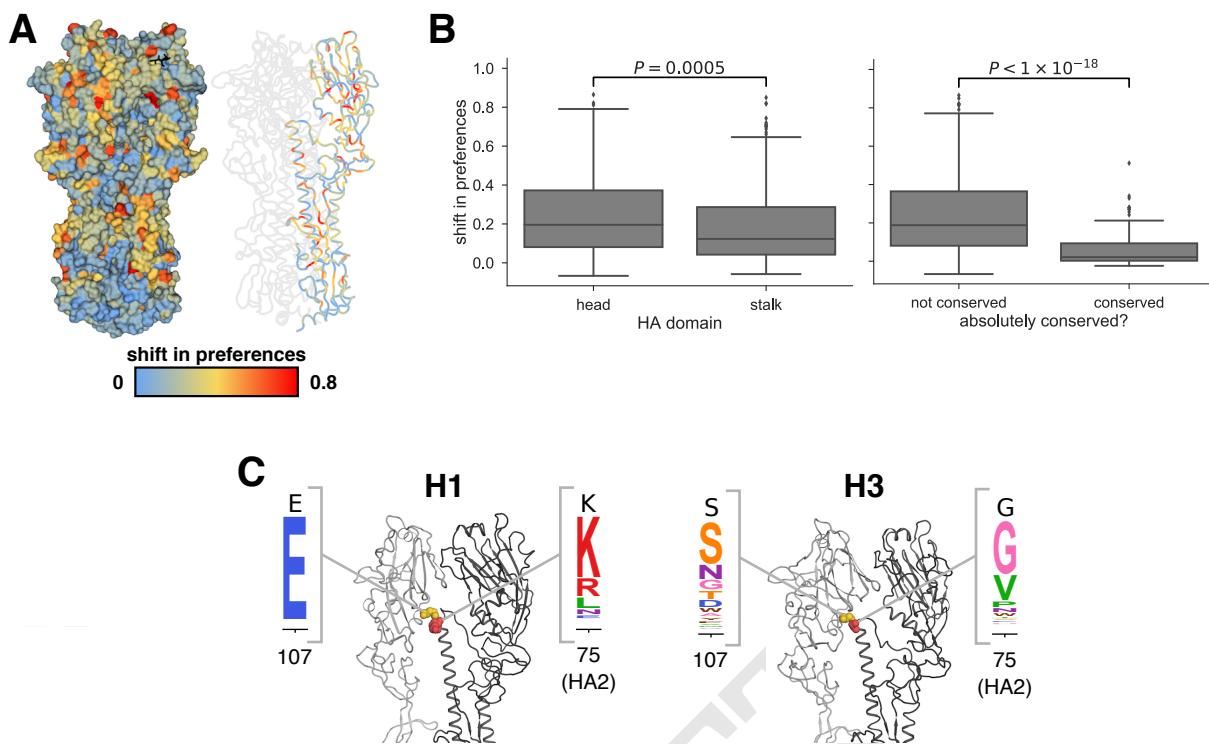
**ACKNOWLEDGMENTS.** We thank Sarah Hilton, Hugh Haddox, Sidney Bell...the Fred Hutch Genomics Core... Funding...

### References.

- Thyagarajan B, Bloom JD (2014) The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife* 3:e03300.
- Doud MB, Bloom JD (2016) Accurate measurement of the effects of all amino-acid mutations to influenza hemagglutinin. *Viruses* 8:155.
- 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](http://www.who.int/influenza/vaccines/virus/recommendations/201002_Recommendation.pdf?ua=1).
- 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](http://www.who.int/influenza/vaccines/2011_02_recommendation.pdf?ua=1).
- 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.
- 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* 81:5605–5614.
- Bloom JD (2015) Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics* 16(1):1.
- Hilton SK, Doud MB, Bloom JD (2017) phdms: Software for phylogenetic analyses informed by deep mutational scanning. *PeerJ* 5:e3657.



**Fig. 9. The HA homologs exhibit many large shifts in preference compared to shifts for other viral protein homologs** (A) A phylogenetic tree of the HA subtypes, with the two HA's, WSN/1933 H1 and Perth/2009 H3, for which we have measured amino-acid preferences denoted on the tree. The WSN/1933 H1 and the Perth/2009 H3 share  $\sim 42\%$  amino-acid identity. (B) The correlations of the amino-acid preferences for replicates both within and between the two HA homologs. The within-Perth/2009 and the within-WSN/1933 correlations are shown in blue and red, respectively. The between homolog correlations are in gray. The correlations for replicates within a homolog are higher than for replicates between homologs. (C) The distribution of shifts in preference for various homolog pairs are shown. The top two distributions show the distances between each of the HA homologs with the non-homologous HIV Env [cite haddox2017]. The center distribution shows the corrected distance between the two HIV Env homologs, which share 86% amino-acid identity. The fourth distribution from the top is a null generated by randomizing the HA replicates and computing the distances. The bottom distribution is the corrected distance between the Perth/2009 H3 and WSN/1933 H1 homologs at all sites that align.



**Fig. 10. Shifts in preferences mapped onto the structure of HA** (A) The preference shifts as calculated by  $RMSD_{corrected}$  between the two HA homologs is mapped onto the structure of HA (PDB 4O5N; (23)). 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. The top ten most shifted sites are shown in spheres on the monomer. (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 (right plot). (C)

9. 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.
10. 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.
11. Waterfield M, Scraff G, Skehel J (1981) Disulphide bonds of haemagglutinin of Asian influenza virus. *Nature* 289:422–424.
12. Weis W, et al. (1988) Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature* 333:426–431.
13. Martin J, et al. (1998) Studies of the binding properties of influenza hemagglutinin receptor-site mutants. *Virology* 241(1):101–111.
14. 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.
15. 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.
16. Girard G, Gulyaeva A, Olsfjoorn R (2011) Upstream start codon in segment 4 of North American H2 avian influenza A viruses. *Infect. Genet. Evol.* 11:489–495.
17. Ekert DC, et al. (2011) A highly conserved neutralizing epitope on group 2 influenza A viruses. *Science* 333(6044):843–850.
18. Friesen R, et al. (2014) A common solution to group 2 influenza virus neutralization. *Proc. Natl. Acad. Sci. USA* 111:445–450.
19. Chai N, et al. (2016) Two escape mechanisms of influenza A virus to a broadly neutralizing stalk-binding antibody. *PLoS Pathogens* 12(6):e1005702.
20. 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*.
21. Wilson I, Skehel J, Wiley D (1981) Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature* 289:366–373.
22. 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.
23. 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.
24. Gamblin S, et al. (2004) The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science* 303(5665):1838–1842.
25. Wu N, et al. (2017) A structural explanation for the low effectiveness of the seasonal influenza H3N2 vaccine. *PLoS Pathogens* 13(10):e1006682.
26. 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.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.
32. Bloom JD (2014) An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular Biology and Evolution* 31:1956–1978.
33. 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).

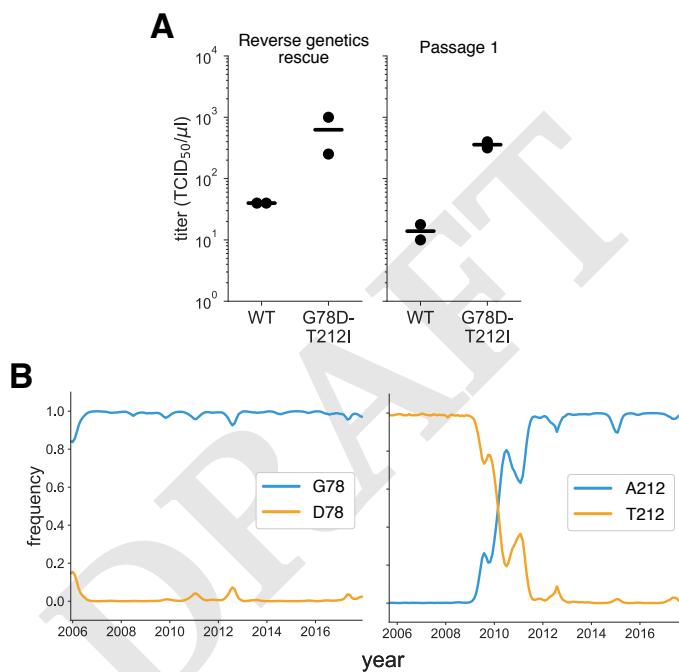
**SI Figures.**

**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.

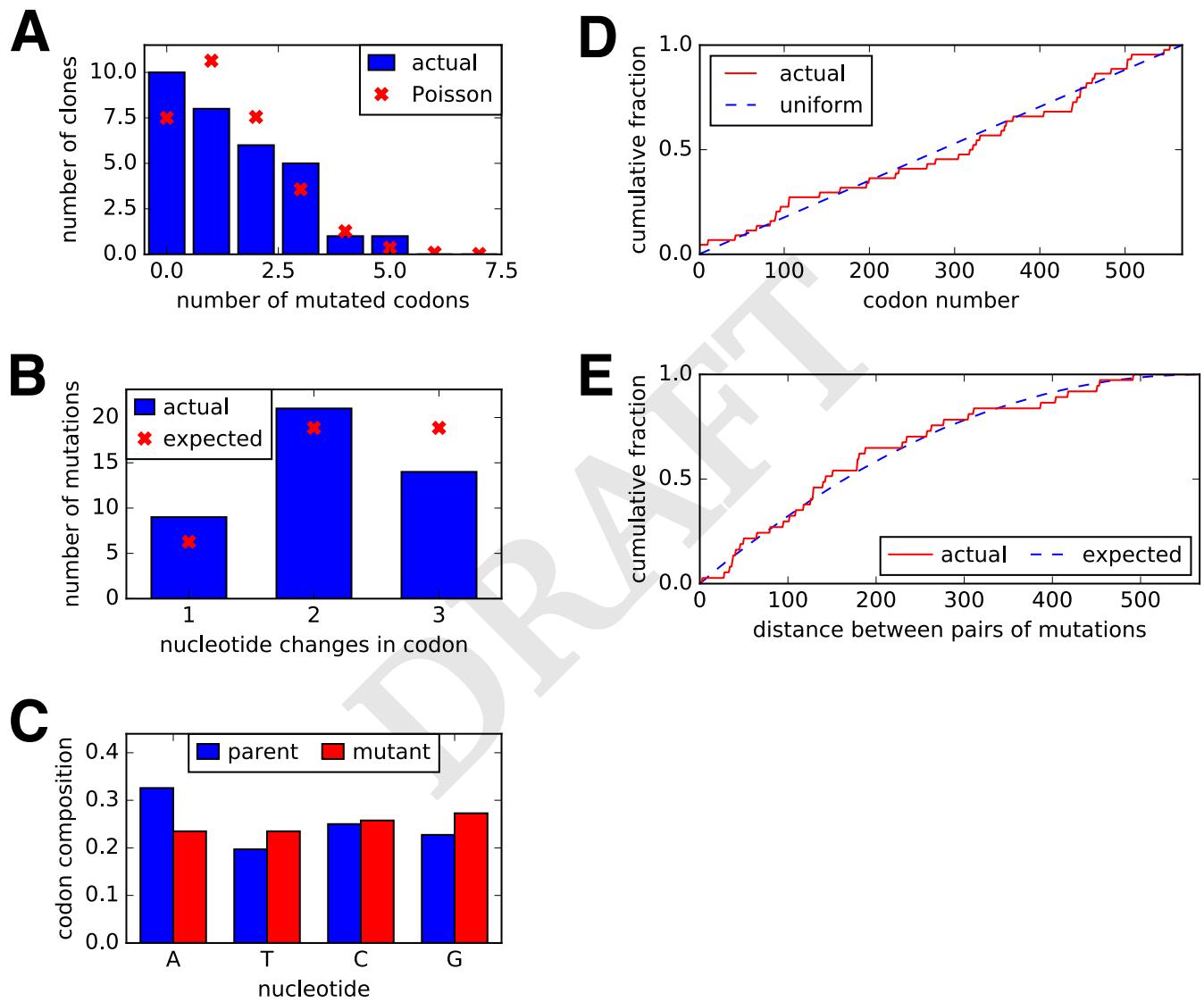
**SI Datasets.**

DRAFT

**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 30 clones from the mutant plasmid libraries. (A) (B) (C) (D) (E)