Deep mutational scanning of an H3 hemagglutinin can inform evolutionary forecasting of human H3N2 influenza virus

Juhye M. Lee[1,4,5,†] John Huddleston[2,6,†] Michael B. Doud[1,4,5]
Kathryn A. Hooper[1,6] Trevor Bedford,[2,3]
Jesse D. Bloom[1,3,4*]

[1]Basic Sciences Division, [2]Vaccine and Infectious Diseases Division, and [3]Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[4]Department of Genome Sciences, [5]Medical Scientist Training Program, and [6]Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA

[†]These authors contributed equally

[*]Correspondence: jbloom@fredhutch.org

Abstract

Abstract text.

INTRODUCTION

RESULTS

Strategy for deep mutational scanning of an H3 hemagglutinin

Previously, we measured the effect of all possible single amino-acid mutations to a highly lab-adapted H1 hemagglutinin from the A/WSN/1933 (H1N1) strain (Thyagarajan and Bloom, 2014; Doud and Bloom, 2016). To examine the effects of mutations to a hemagglutinin of more direct relevance to influenza strains currently circulating in the human population, we chose to study H3 HA from the A/Perth/16/2009 (H3N2) strain. This strain was the H3N2 component of the influenza vaccine from 2010-2012. We first rescued virus carrying the wildtype Perth/2009 HA and NA and the WSN internal genes by reverse genetics, and serially passaged the virus in cell culture a total of six times. Two mutations, G78D and T212I, arose together from passaging, and the G78D-T212I Perth/2009 HA mutant supported viral growth to ~15- to 20-fold higher titers than the wildtype HA [*(supplementary data?)*]. We decided to perform all deep mutational scanning experiments in the background of the Perth/2009 HA carrying these two mutations to support

higher viral growth and to limit these cell culture adaptation mutations from swamping out the signal from other mutations.

We mutagenized the Perth/2009 HA gene to create mutant plasmid HA libraries encompassing all 567 codons and harboring an average of ~1.4 codon mutations per clone. We then used a helper-virus system previously established in Doud and Bloom (2016) to bypass the bottlenecks associated with reverse genetics to generate complex mutant virus libraries from the mutant plasmids (Figure 1A). In order to maximize viral titers and further avoid bottlenecking the diversity of the initially generated mutant viruses, we used an HA-deficient helper virus carrying WSN/1933 internal and NA genes to rescue the mutant viruses, thereby pairing the mutant HA's with the WSN NA. Additionally, we used MDCK-SIAT1 cells constitutively expressing the TMPRSS2 protease, which is found endogenously in the human airway and facilitates HA cleavage and activation [*cite Böttcher 2006 JVI, Böttcher-Friebertshäuser 2010 JVI*]. All of the experiments were completed in full biological triplicate (Figure 1B). We also passaged and deep sequenced library 3 in technical replicate (denoted as library 3-1 and 3-2) to gauge to the amount of experimental noise occurring *within* a single biological replicate.

Figure 1C shows the mutation frequencies of the mutant plasmids, mutant viruses, wildtype plasmids and wildtype viruses determined from deep sequencing. There is selection against non-functional HA variants as revealed by the reduced mutation frequencies of the mutant viruses compared to their starting frequencies in the mutant plasmids. Specifically, stop codons were purged to 20-45% of their original starting frequencies, after correcting for error rates from the wildtype controls. Although the majority of stop codons were purged, incomplete purging of stop codons is likely due to complementation of defective viral particles by infectious virions. We also observed purging of most nonsynonymous mutations to 30-40% of their initial starting frequencies after error correction, suggesting strong selection against deleterious HA variants.

We next quantified the reproducibility of our deep mutational scanning measurements across biological and technical replicates. We inferred the amino-acid preferences at each site in the Perth/2009 HA using the method described in Bloom (2015) and implemented in the `dms_tools2` software [`https://jbloomlab.github.io/dms_tools2/`]. These preferences represent an estimate of the 567 sites $\times$ 20 amino acids = 11340 experimental measurements and are normalized to sum to one at each site. The correlations of the amino-acid preferences between each pair of replicates is shown in Figure 1D. The biological replicates are fairly well-correlated, with a Pearson's R ranging from 0.69 to 0.78. Replicate 1 exhibits the least amount of correlation with the other biological replicates, consistent with the observation that this replicate showed the weakest selection against stop and nonsynonymous mutations and might therefore be subject to more experimental

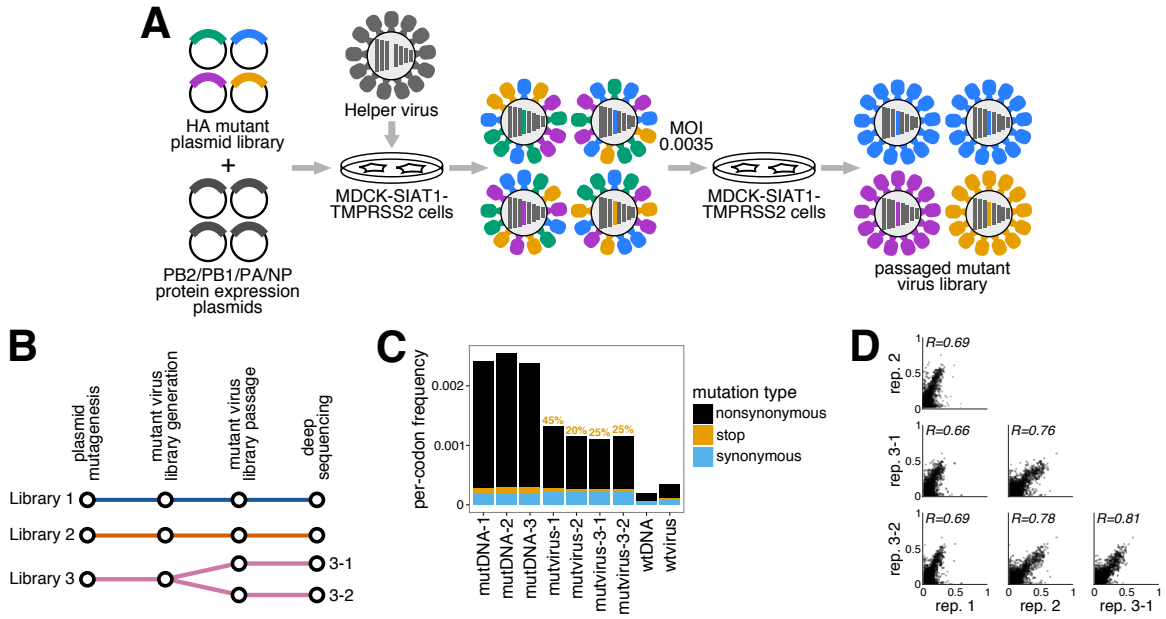

Figure 1: Overview of deep mutational scanning experiments of H3 hemagglutinin. (A) We used a helper virus approach previously described in Doud (2016) to generate the mutant virus libraries. We transfected MDCK-SIAT1-TMPRSS2 cells with the mutant plasmid library carrying all possible amino-acid mutations to the A/Perth/16/2009 (H3N2) HA, in addition to protein expression plasmids encoding the viral ribonucleoprotein complex. After transfection, we infected the cells with an HA-deficient helper virus carrying all of the WSN/1933 (H1N1) influenza genes. We then passaged the initially generated pool of mutant viruses at low MOI to establish a genotype-phenotype linkage and select for functional HA variants. (B) All of the experiments were completed in biological triplicate, starting from independent preps of the wildtype HA genes to create the mutant plasmids. In addition, we passaged and deep sequenced library 3 in technical replicate, denoted 3-1 and 3-2, to estimate the amount of experimental noise within a single biological replicate. (C) Mutation frequencies of nonsynonymous, stop, and synonymous mutations for the mutant DNA, mutant virus, wildtype DNA, and wildtype virus samples. There is selection against nonsynonymous and stop codons in the mutant viruses. The percentages signify the frequency of stop codons remaining in the passaged mutant viruses relative to their starting frequency in the mutant plasmid libraries after correcting for the stop codon frequencies in the wildtype DNA and viruses. (D) The correlations and the Pearson correlation coefficient for the amino-acid preferences between each pair of replicates are shown. The biological replicates are fairly well-correlated, as are the technical replicates, indicating some degree of bottlenecking of variants during the viral passage which can contribute to experimental noise.

noise. Of note, the two technical replicates 3-1 and 3-2 were only slightly more reproducible than that between biological replicates. This suggests that bottlenecking of the virus library during the low MOI passage contributes to much of the noise observed in our experiments, as we are only able to passage a finite number of viral particles.

Model	ΔAIC	Log Likelihood	Parameters
ExpCM	0.0	-8439.33	$\beta = 2.44, \kappa = 5.78, \omega = 0.91$
Goldman-Yang M5	2166.06	-9516.36	$\omega_\alpha = 0.30, \omega_\beta = 0.84, \kappa = 5.10$
ExpCM, averaged across sites	2504.18	-9691.42	$\beta = 0.68, \kappa = 5.58, \omega = 0.32$
Goldman-Yang M0	2607.92	-9738.29	$\kappa = 5.05, \omega = 0.31$

Table 1: The site-specific amino-acid preferences are informative for describing human H3N2 evolution in nature. 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 (ΔAIC) (Posada and Buckley, 2004). 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 Yang et al. (2000). A non-site-specific ExpCM informed by preferences averaged across all sites does not perform as well as the M5 model, but still has a higher log likelihood than M0. The optimized parameters for each model are also shown.

H3 site-specific amino-acid preferences

How well do the Perth/2009 HA preferences inferred from experimental measurements describe the evolution of H3N2 influenza virus in nature? This question can be addressed by evaluating how well experimentally informed codon substitution models (ExpCM's) constructed from our laboratory measurements improve phylogenetic fit of H3N2 evolution (Hilton et al., 2017). The results in Table 1 show that the ExpCM outperforms conventional substitution models in describing the evolution of human H3N2 HA. The ExpCM also optimizes a stringency parameter (β) 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 2 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) (Waterfield et al., 1981) 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. 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 (Kido et al., 1992; Stech et al., 2005). 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 in [*need to list citations here, inc. Girard 2011 Infection Genetics and Evolution, Chen 2001 Nat Med*].

We next sought to investigate the inherent mutational tolerance of the Perth/2009 HA. Figure 3 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. Interestingly, the Perth/2009 H3 stalk is relatively mutationally tolerant compared to the tolerance of the globular head domain. In particular, the shorter α-helix (helix A) of the stalk domain, antigenic site C, and the most distal portion of the globular head near antigenic site B are all highly mutationally tolerant. On the other hand, the head domain of the WSN/1933 H1 is more mutationally tolerant relative to its stalk domain. The receptor binding pockets of both HA's are relatively mutationally intolerant, although the residues surrounding the receptor binding pocket are fairly tolerant of mutations. These findings are consistent with the high functional constraint of the receptor binding pocket, while the mutability of sites surrounding the pocket that are under strong immune pressure may contribute to antigenic evolution (Wilson et al., 1981; Wiley et al., 1981).

Estimating mutational effects from an H3N2 phylogeny

Comparing H1 and H3 preferences

DISCUSSION

Figure 2: The site-specific amino-acid preferences of H3 hemagglutinin. This logoplot 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.

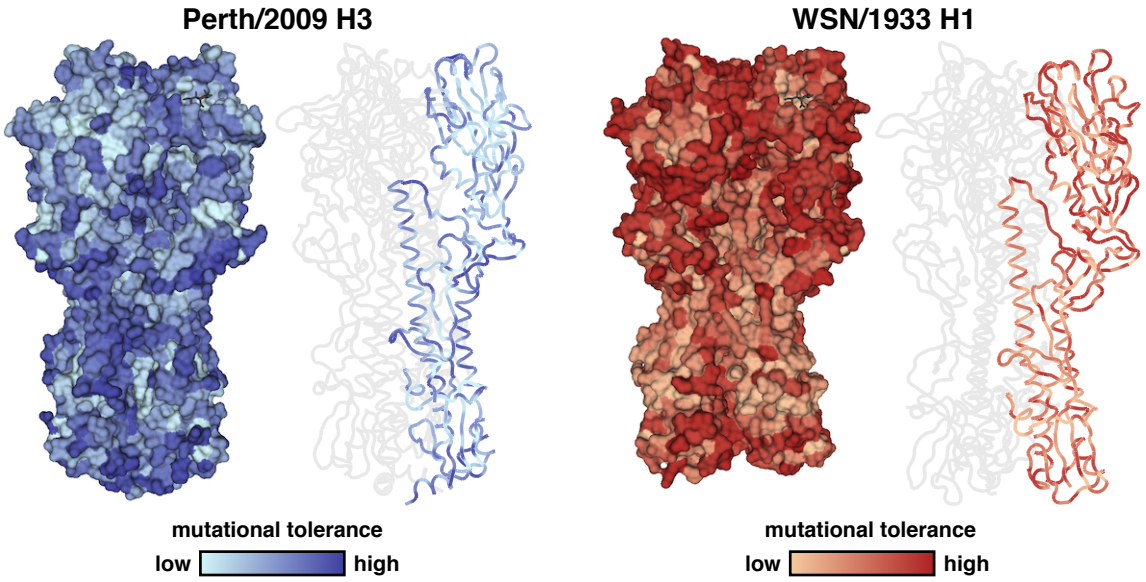

Figure 3: Different regions of HA exhibit varying degrees of mutational tolerance. Mutational tolerance as calculated by the Shannon entropy of a given site's amino-acid preferences are mapped onto the structure of the H3 trimer (PDB 4O5N; Lee et al. (2014)) and the H1 trimer (PDB 1RVX; Gamblin et al. (2004)), 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 Doud and Bloom (2016). Lighter shades of blue or red signify low mutational tolerance, while darker shades of blue or red 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 compared to the head region. The head region of the WSN/1933 H1 is mutationally tolerant compared to the relatively intolerant stalk region.

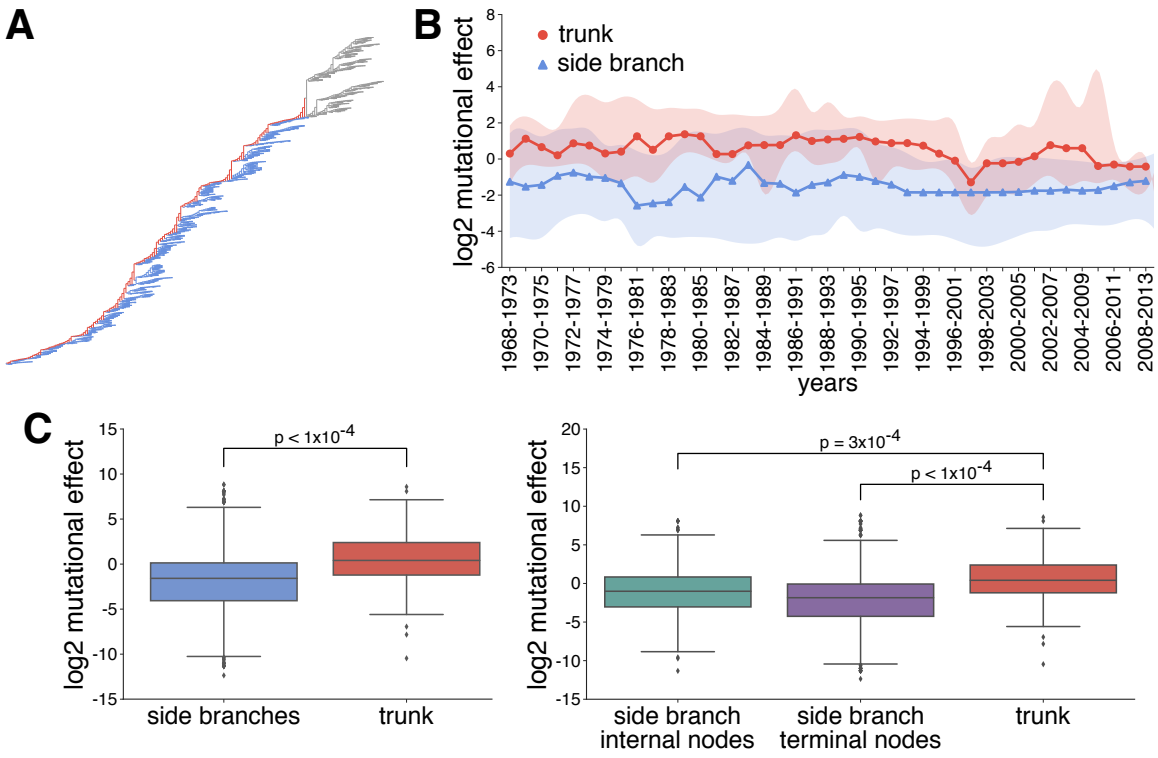

Figure 4: The trunk of a human H3N2 phylogeny has higher mutational effects than those of side branches. (A) Phylogenetic tree of human H3N2 influenza virus from 1968-present. [*We downloaded X sequences from the Influenza Virus Resource ?.... etc. inferred the tree, ancestral state reconstruction, visualized 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. All branches ancestral to the MRCA of the recent nodes were defined as the trunk (shown in red), and all other branches were defined as side branches (shown in blue). (B) 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. The median trunk mutational effects are consistently higher than the median side branch mutational effects for all windows. (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. To estimate significance, we performed 10,000 randomizations of the preferences to calculate the median difference in trunk vs side branch mutational effects, and counted how many of the randomizations exceeded the true difference in median trunk vs side branch mutational effects. The effects of trunk mutations are higher than side branch, internal and terminal node, mutations.

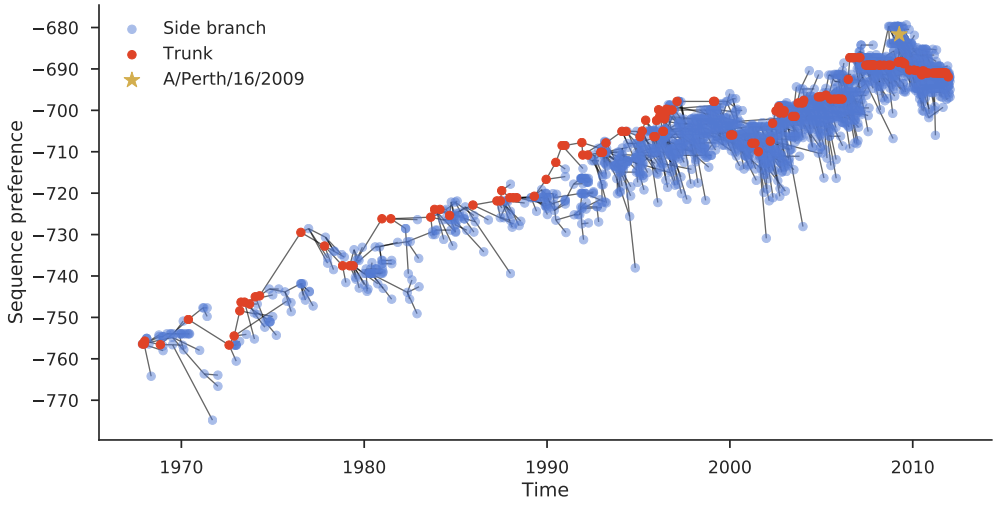

Figure 5: The trunk exhibits higher sequence preference than side branches. figure text

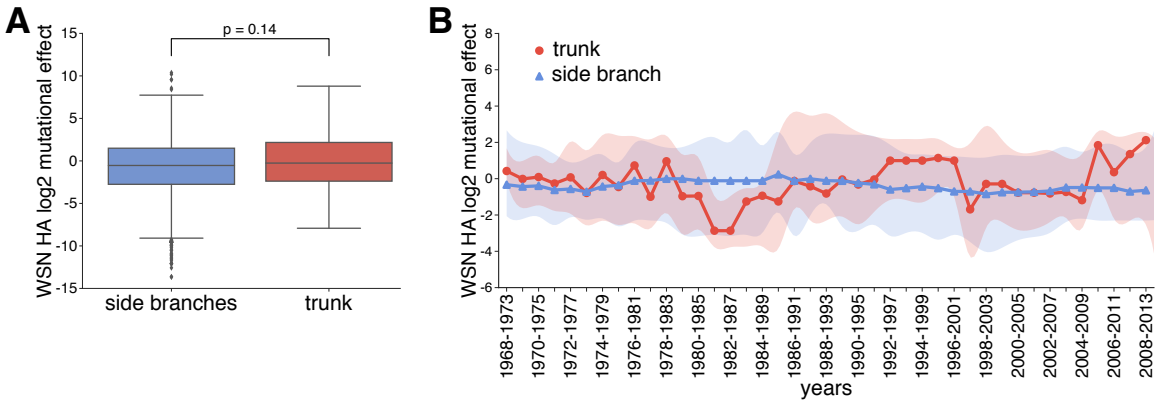

Figure 6: The WSN/1933 H1 preferences do not reveal differences in trunk vs side branch mutational effects (A) We calculated the $\log_2$ mutational effects of the same set of trunk and side branch mutations from the inferred H3N2 phylogeny in Figure 4 using the WSN/1933 H1 preferences. We again randomized the preferences for a total of 10,000 iterations and counted the number of randomizations that exceeded that true difference in median trunk vs side branch mutational effects to calculate a p-value. Using the WSN/1933 H1 preferences, there is not a significant difference in trunk vs side branch mutational effects. (B) We also performed the same sliding window analysis shown in Figure 4B, but using the WSN/1933 H1 preferences. There is not a distinct difference in trunk and side branch mutational effects.

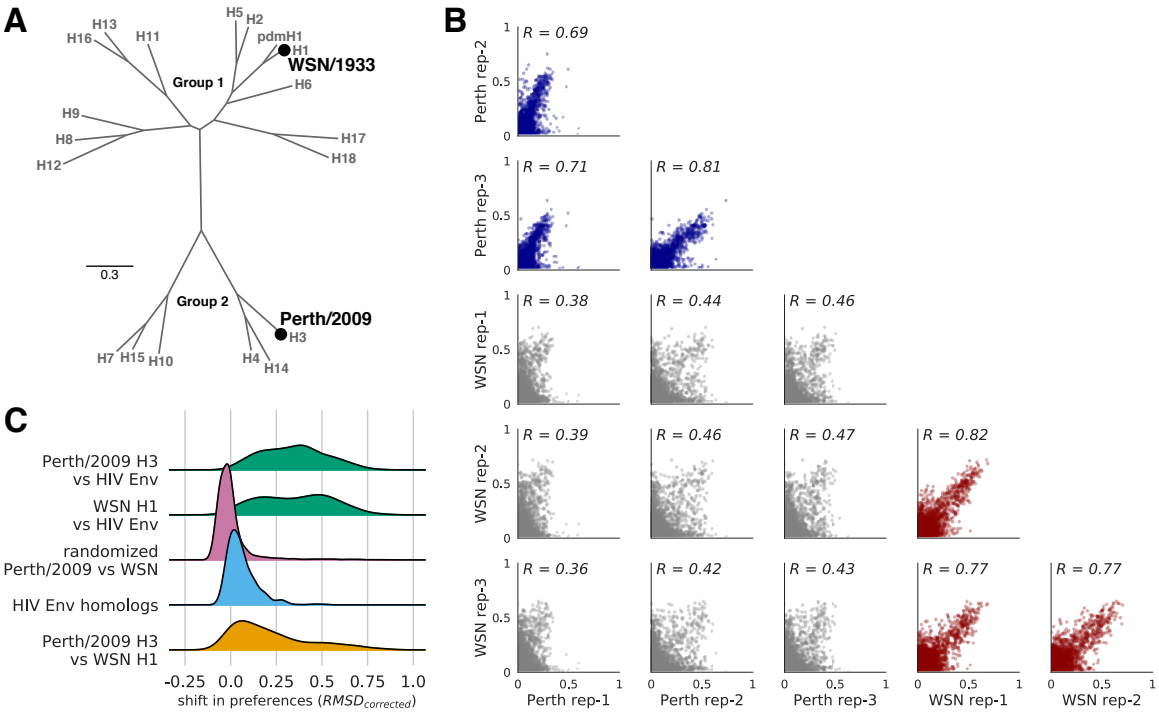

Figure 7: The HA homologs exhibit many large shifts in preference compared to shifts for other viral protein homologs (A) A phylogenetic tree of the HA subtypes, with the two HA's, WSN/1933 H1 and Perth/2009 H3, for which we have measured amino-acid preferences denoted on the tree. The WSN/1933 H1 and the Perth/2009 H3 share ~42% amino-acid identity. (B) The correlation for the amino-acid preferences for replicates both within and between the two HA homologs. (C) The distribution of shifts in preference for various homolog pairs are shown.

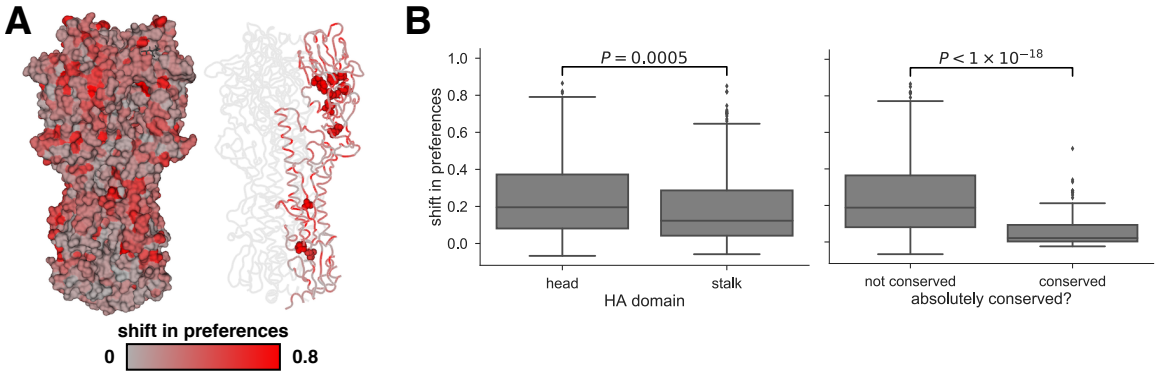

Figure 8: Shifts in preferences mapped onto the structure of HA (A) The preference shifts as calculated by $RMSD_{corrected}$ between the two HA homologs is mapped onto the structure of HA (PDB 4O5N, citation). The left structure shows the HA trimer, and the right structure colors one of the monomers. The sialic acid receptor is shown in black sticks. Gray indicates little shifts in preference, while red indicates large shifts in preference. The top ten most shifted sites are shown in spheres on the monomer. (B)

METHODS

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) × 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 (Hoffmann et al., 2000) and pICR2 [*cite?*] plasmid backbones.

The codon-mutant libraries were generated using a PCR-based approach described in Dingens et al. (2017).

Generation and passaging of mutant viruses

The mutant virus libraries were generated and passaged using the approach described in Doud and Bloom (2016) with several modifications.

Barcoded subamplicon sequencing

Analysis of deep sequencing data

Inference of phylogenetic trees

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

Bloom JD. 2015. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*. 16:1.

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.

Doud MB, Bloom JD. 2016. Accurate measurement of the effects of all amino-acid mutations to influenza hemagglutinin. *Viruses*. 8:155.

Gamblin S, Haire L, Russell R, et al. (11 co-authors). 2004. The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science*. 303:1838–1842.

Hilton SK, Doud MB, Bloom JD. 2017. phydms: Software for phylogenetic analyses informed by deep mutational scanning. *PeerJ*. 5:e3657.

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.

Kido H, Yokogoshi Y, Sakai K, Tashiro M, Kishino Y, Fukutomi A, Katunuma N. 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.

Lee PS, Ohshima N, Stanfield RL, Yu W, Iba Y, Okuno Y, Kurosawa Y, Wilson IA. 2014. Receptor mimicry by antibody F045-092 facilitates universal binding to the H3 subtype of influenza virus. *Nat Commun*. 5:3614.

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*. 53:793–808.

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*. 11:683–689.

Thyagarajan B, Bloom JD. 2014. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife*. 3:e03300.

Waterfield M, Scrace G, Skehel J. 1981. Disulphide bonds of haemagglutinin of Asian influenza virus. *Nature*. 289:422–424.

Wiley D, Wilson I, Skehel J, et al. (4 co-authors). 1981. Structural identification of the antibody-binding sites of hong kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*. 289:373–378.

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

Yang Z, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. 155:431–449.