©Copyright 2017
Michael B. Doud

Comprehensively mapping the effects of mutations to influenza virus

Michael B. Doud

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Jesse D. Bloom, Chair

Julie Overbaugh

Doug Fowler

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Comprehensively mapping the effects of mutations to influenza virus

Michael B. Doud

Chair of the Supervisory Committee:
Associate Member Jesse D. Bloom
Fred Hutchinson Cancer Research Center

Influenza virus is a rapidly evolving threat to public health. Influenza evolves through point mutations, so knowing the effects of all amino-acid mutations on the ability of the virus to withstand various selective pressures reveals the evolutionary paths accessible to the virus. There are approximately 10,000 different amino-acid mutations that can be made to the average influenza gene, but accurately predicting the effects of any one mutation is difficult. However, new methods leveraging the latest technologies in mutagenesis and high-throughput DNA sequencing have made it possible to measure the effects of all possible mutations to an influenza gene. This approach involves introducing all codon mutations to an influenza gene, reconstituting mutant virus libraries carrying these mutations and the corresponding protein variants, imposing selective pressure on the mutant virus libraries, and using accurate deep sequencing methods to quantify the frequencies of all mutations before and after selection. The effect of each mutation can then be computed from the change in mutation frequency during selection, where beneficial mutations will increase in frequency and deleterious mutations decrease in frequency. Here I describe several applications of this approach to comprehensively measure the effects of mutations within two influenza genes. First, I examine the extent that mutational effects shift during the course of protein evolution by measuring mutational effects to two homologs of influenza nucleoprotein separated by over thirty years of evolution. Although there are

a few protein sites with strong shifts in which amino acids are preferred, the effects of mutations at most sites are conserved across these homologs. The mutational effects measured in these two human influenza nucleoprotein homologs accurately describe the evolution of more distant influenza viruses infecting pigs, horses, and birds, demonstrating the feasibility of using measurements on one virus strain to model the evolution of more distantly related strains. Next, I describe technical improvements to the process of generating and selecting comprehensive mutant virus libraries of influenza hemagglutinin that yield more accurate and reproducible measurements of mutational effects than previously possible. Finally, I extend this approach to comprehensively map all mutations in hemagglutinin that enable the virus to escape from neutralizing antibodies. These results reveal the striking mutation-level idiosyncrasy of antibody escape: at most epitope sites only a subset of mutations confer escape to a given antibody, and similar antibodies targeting the same antigenic site elicit distinct profiles of escape mutations. Collectively, these studies enhance our understanding of influenza evolution and immune evasion and expand our ability to comprehensively map the evolutionary potential of viruses.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation would not have been possible without the guidance, support, and camaraderie of many. Here I will give my best attempt at acknowledging them, while admitting that with the words on these pages I will not be able to fully express my appreciation for their personal and professional support.

My advisor Jesse Bloom has been a fantastic mentor. From the very beginning of my time working with him, he has provided me the freedom to pursue the scientific questions that most interested me. His guidance was never overbearing; much of what I learned from him about experimental design, interpretation of results, computational analysis, and scientific writing and presentation was learned by him first letting me chart my own path (which undoubtedly involved making my own mistakes), and then through his thoroughly honest and constructive criticism. Jesse maintains a healthy and friendly lab culture where everyone's ideas are valued and where students receive ample recognition for their work. I never thought twice about joining his lab (as only his second graduate student), and upon completion of this dissertation I can say that I would certainly pick his lab again if I were to somehow magically have the pleasure of repeating graduate school. Thank you Jesse, especially for being a good sport when I revealed the design of your favorite pair of socks to everyone at my defense.

Within the past and present members of the Bloom Lab, Orr Ashenberg deserves a special acknowledgement. Orr is a wonderful collaborator, mentor, colleague, and friend. Collaborating with Orr on one of my earlier projects in the lab was incredibly fun (and I'm pretty sure Jesse felt we were having *too much fun* when he discovered that Orr and I chose to use `GitHub`'s repository name suggestion of "furry-octo-wookie" to host our

manuscript's work-in-progress. Jesse, being the good sport that he is, let us continue to use this repository for our project, although I'm pretty sure he felt a little uncomfortable every time he had to type "`git pull furry-octo-wookie`" just to see our latest drafts of the paper). Orr is a fellow coffee lover and I enjoyed countless conversations during our near-daily pour over coffee rituals about our work in the lab, interesting new papers we've encountered, and non-science topics alike. Orr's thought-provoking questions and enduring penchant for quirky humor has been central to the Bloom Lab's esprit de corps since he joined, and I was not the only one who benefited from the atmosphere he cultivated in the lab.

I also want to especially thank Bargavi Thyagarajan. Bargavi was my primary source for learning many techniques in the lab, and in many ways the work she did in the lab inspired two-thirds of the work in this dissertation. I really appreciate Bargavi's willingness to teach, her friendship, and her many random acts of kindness, such as bringing me dinner during my rotation when she discovered I was trapped in an hours-long experiment in the evening. Both Orr and I really felt awful when we were unable to keep her plant alive when she moved to New York.

Fellow graduate students and post-docs of the Bloom lab have always been supportive, both within and outside of the lab. I am grateful for many conversations with fellow lab members Heather Machkovech, Adam Dingens, Hugh Haddox, Juhye Lee, Sarah Hilton, Katie Hooper, Alistair Russel, Shirleen Soh, Katherine Xue, and Danny Lawrence, all of whom are helpful colleagues always willing to hear out my frustrations and suggest solutions. I am particularly indebted to Heather, Adam, Orr, and Juhye for tolerating my nerdiness and allowing me to talk about science with them while backpacking the Enchantment Lakes or climbing Mount St. Helens. Likewise, my colleges in the UW MSTP including Aaron Seo, BJ Valente, Ken Chen and Natalie Vandeven often put up with my musings during hiking and snowboarding trips. On that note, I'm really lucky to be surrounded by

many other really amazing colleagues in the UW MSTP and I've really treasured their friendship over the multiple phases of our training. Outside of the Bloom lab, I am deeply appreciative of the entire scientific community at the Fred Hutchinson Cancer Research Center and the University of Washington. I am especially grateful for feedback from my committee members Julie Overbaugh, Doug Fowler, and Marion Pepper. There are too many other graduate students, post-docs, faculty, and administrative staff to list, all of whom contribute to an inquisitive, supportive, collaborative, efficiently-run scientific environment. The weekly virology group meetings and the Basic Sciences Division seminar series always provided opportunities for valuable feedback throughout the course of these projects, and I appreciated many thought-provoking questions about my work over the years from Adam Geballe, Harmit Malik, Michael Emerman, Gerry Smith, Trevor Bedford, Erick Matsen, and many others. Furthermore, this work would not have been possible without the support of the UW MSTP and the UW Genome Sciences Department.

I'd also like to acknowledge my first scientific mentor, Tim Springer, and members of the Springer Lab including Chafen Lu (for instilling in me much of my practical knowledge about cellular and molecular biology lab techniques), and Drew Drabek and Alex Davies (for their friendship as the 'other two' amigos/musketeers). I'm ever grateful for the opportunity Tim gave me to work in his lab for two years. The experience was transformative in allowing me to discover my passion for science and inspiring me to pursue a PhD.

Finally, I want to thank the unwavering support of my family, especially my parents, and the ability of my partner Lydia to help me see awe in the world whenever I begin to lose sight of it. Her continuous support throughout my entire journey at the University of Washington thus far has been instrumental to my success and well-being.

Chapter 1
INTRODUCTION

In *The Library of Babel*, Jorge Luis Borges describes an imaginary universe as a library filled with every possible book containing a specified number of pages, with a specified number of lines of text per page, in which each line of text is comprised of a set number of characters from a defined alphabet [20]. Borges wrote this short story in 1941, but had he written it after the birth of molecular biology and our discovery of the basic workings of the genetic code, he may have explored the concept of *biological sequence space* [119, 38, 4] instead of the *literary sequence space* he contemplates. Indeed, Borges invokes a metaphor for the biological sequence space of point-mutants, informing the reader that for every book in the library, "there are always several hundred thousand imperfect facsimiles: works which differ only in a letter or a comma" [20]. This sub-library of imperfect facsimiles is reminiscent of the viral quasispecies [80] of rapidly-evolving RNA viruses, such as influenza virus. In this dissertation I will describe the construction and analysis of the virological equivalents of Borges' sub-libraries of "several hundred thousand imperfect facsimiles": libraries of mutant viruses sampling all amino-acid mutations in a viral gene, which can be used to measure inherent mutational tolerance, comprehensive antibody escape mutation repertoires, and in theory, a comprehensive sequence-function map for any viral phenotype that can be selected for in the laboratory.

Influenza virus is a rapidly evolving threat to public health. Influenza virus circulates globally, causing approximately three to five million cases of severe illness in each

annual epidemic, resulting in approximately 250,000 to 500,000 deaths [95]. The error-prone influenza RNA polymerase introduces up to three mutations per viral genome per replication cycle [98, 99], continually providing new mutants upon which complex selection pressures act [9]. One important consequence of this rapid evolution is that it allows the virus to escape from immune responses in a process termed “antigenic drift”, which can render seasonal influenza vaccines ineffective after several years. This necessitates the selection of new vaccine strains based on predictions of which viral strains will predominate in future seasons and how mutations in these strains might affect recognition by the immune system.

Various forces drive and constrain influenza evolution. New mutations can be neutral with respect to the fitness of the virus if they have little or no effect on the ability of the virus to replicate, or they can be beneficial or deleterious to varying degrees. Each of the eight influenza genes encode at least one viral protein, and collectively these proteins carry out the viral replication cycle. For example, the hemagglutinin protein on the surface of the virus is responsible for binding to host cell receptors and fusing the viral and host membranes, and mutations that interfere with these important functions can be detrimental to viral fitness. In addition to mutations disrupting specific functions of viral proteins, many mutations will be incompatible with viral replication because they will render the proteins misfolded or unstable. These various types of deleterious mutations, in the absence of compensatory mutations, are expected to be evolutionary “dead ends”, representing inherent constraints on which mutations are tolerated during evolution. On the other hand, some mutations will be beneficial for the virus, providing increased viral fitness in the face of some selective pressure: antigenic mutations that abrogate the ability of antibodies to neutralize the virus are one example. It is important to note that the genetic context of the virus in which a mutation appears can modulate the effect of the mutation on protein stability or function, and consequently viral fitness [56, 133], and so mutations that are “dead ends” in one viral strain might not necessarily be deleterious when they appear in

a different viral strain. The fact that mutational effects might differ between a particular strain assayed in the lab and another strain circulating in nature can complicate efforts to extrapolate experimental measurements to circulating viruses. There are many anecdotal examples of mutations that exhibit such epistasis, but it is unknown how frequently and to what magnitude the effects of mutations might change with sequence context throughout an entire protein - a question that can be answered only by comprehensively mapping the effects of mutations to two homologs of the same protein.

It is difficult to accurately predict the effects of the vast number of possible mutations. The average influenza gene has a length of approximately 500 residues, to which there are approximately 10,000 possible single amino-acid mutations. *A priori*, it is nearly impossible to accurately predict the consequences of these mutations in the face of the various drivers and constraints on influenza evolution, complicating efforts to predict which emerging strains will dominate in future seasons based on accumulated mutations. Even one of the most sophisticated predictive models for influenza evolution makes the unrealistic assumption that all mutations to influenza hemagglutinin are deleterious unless they occur in an area of the protein believed to be targeted by antibodies [86]. This is not a flaw in the model; instead, it is a reflection of the fact that the effects of any of the tens of thousands of possible mutations appearing during an influenza season are difficult to predict without the guidance of real experimental data. More accurate inferences of the evolutionary potential of circulating viral strains will greatly benefit from advances in experimental approaches capable of examining *all* possible mutations.

New experimental approaches can be used to comprehensively map the effects of viral mutations. The development of several new technologies in recent years has enabled a new style of high-throughput genetics in which the effects of tens of thousands of different mutations can be measured simultaneously in a single experiment, an approach I will refer to generally as “deep mutational scanning” throughout this dissertation [49, 52].

As applied to influenza virus, this experimental approach leverages the latest technologies in DNA synthesis and mutagenesis to create libraries of mutant genes, which are incorporated into viruses such that in the resulting mutant virus library each virus carries one (or several) of the many possible mutations in the viral gene of interest. The mutant virus libraries are then subjected to a selection step in which beneficial mutations lead to increased viral replication and deleterious mutations lead to decreased viral replication. Deep sequencing of the mutant pools before and after the selection step is then used to accurately measure changes in the frequency of all of the mutants assayed, and these measurements are used to infer the effects of each individual mutation in the context of the selective pressure used in the experiment (Figure 1.1). We interpret the results from these deep mutational scanning experiments as *site-specific amino-acid preferences*. At each residue in a protein, there are twenty possible amino acids. We define a *preference* for each of the twenty amino-acids at each residue based on the mutational scanning data, where deleterious amino-acid mutations receive smaller preferences and beneficial amino-acids receive larger preferences.

Faithfully recreating the selective pressures in nature is impossible in the lab, but as a first approximation, one can select for viral replication in cell culture. There are obvious caveats to this strategy. For instance, this type of experiment fails to consider important drivers of natural viral evolution such as adaptive immunity, so are the measurements of mutation effects made in cell culture actually concordant with the effects of mutations during natural influenza transmission?

Recent work has used the effects of mutations on viral replication measured by deep mutational scanning in cell culture to build models of protein evolution that reflect the unique site-specific amino-acid preferences for each site in the protein. This is in contrast with standard evolutionary models that are *not* site-specific, and assume that each site will evolve to reflect a uniform preference for all twenty of the amino acids. Site-specific models informed by deep mutational scanning of influenza virus outperform the standard models [123, 15], demonstrating that the evolutionary constraints measured with deep

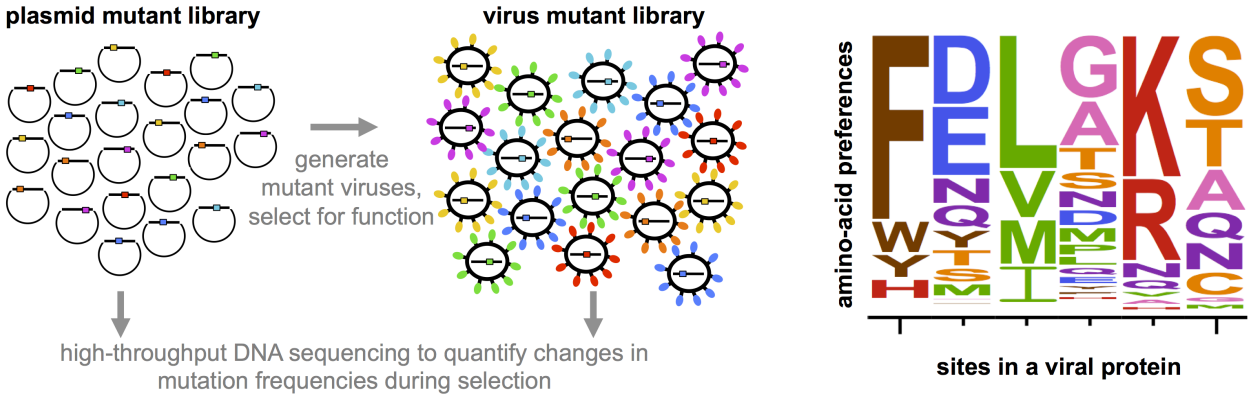

Figure 1.1: **Schematic of deep mutational scanning of influenza virus and site-specific amino-acid preferences.** Saturation codon mutagenesis introduces all possible codon mutations at all codon sites in an influenza gene on an influenza reverse genetics plasmid backbone to create a plasmid mutant library. Mutant viruses carrying these mutant genes and the corresponding protein variants are grown in cell culture and selected for viral replication. High-throughput DNA sequencing quantifies the frequencies of every mutation before and after selection. The relative enrichment or depletion of each amino-acid variant at each site in the protein is visualized as a stack of amino-acid preferences for the site, with the height of each letter proportional to the preference for that amino-acid. Mutations to highly preferred amino acids will increase in frequency during viral replication, and mutations to lowly preferred amino acids will decrease. Throughout this dissertation, amino acids are colored by physiochemical properties of the amino-acid side chain: hydrophobic (V, L, I, M, P) are green, nucleophilic (S, T, C) are orange, small (A, G) are pink, aromatic (F, Y, W) are brown, amide (N, Q) are purple, positively-charged (H, K, R) are red, and negatively-charged (D, E) are blue.

mutational scanning in the laboratory are largely reflective of the constraints during natural virus evolution. However, the degree to which these constraints are constant over evolutionary timescales is not known, and comprehensively measuring the effects of mutations in the context of immune pressure (e.g., neutralizing antibody) has not been feasible prior to my work.

The influenza virus nucleoprotein and hemagglutinin. In the following chapters I will describe several lines of work based on the application of deep mutational scanning to two influenza virus genes: nucleoprotein (NP) and hemagglutinin (HA), both of which perform multiple conserved functions essential to the virus lifecycle.

NP encapsidates the viral RNA, and along with the viral polymerase proteins PA, PB1,

and PB2, forms a viral ribonucleoprotein complex for each of the eight genome segments. During viral infection of a host cell, NP is expressed at high levels and is required for the transcription, replication, and trafficking of the viral ribonucleoprotein complexes [43], and the 3-D structure of NP is well conserved over divergent homologs [36, 138].

HA is the virus's most abundant surface protein, forming homotrimers on the viral surface responsible for both binding to host cell receptors and fusion of viral and endosomal membranes [130]. The structure of HA is also well conserved over divergent influenza virus types [114]; this consists of a fusion domain within the membrane-proximal region of HA (commonly referred to as the stalk or stem), and a membrane-distal receptor-binding domain (commonly referred to as the globular head) (Figure 1.2 A).

Most neutralizing antibodies against HA are directed against the globular head domain. These antigenic sites in HA were originally defined through repetitive selections of individual, de-novo escape mutants from wild-type stocks of virus with murine monoclonal antibodies (Figure 1.2 B) [53, 23]. These experiments uncovered five antigenic regions in the globular head domain of HA, each comprised of a handful of sites. However, for any given antibody used in the experiments, only a handful of escape mutations were identified, since the experiments were not exhaustive. To date there are no methods for completely mapping escape mutations from an antibody, so for any given antibody, it is difficult to accurately predict if a mutation in or near one of the classically defined antigenic sites actually confers escape from the antibody.

Layout of this dissertation. One potential concern about the applicability of deep mutational scanning data is that the effects of mutations measured in one viral strain in the laboratory might be different than the effects of the same mutations in other viral strains due to changes in the underlying genetic sequence. Many examples of this type of epistasis have been observed in the study of the evolution of various proteins, but prior to my work, no studies have measured how prevalent this phenomenon is within a protein by comprehensively measuring the effects of all mutations to two protein homologs. To what

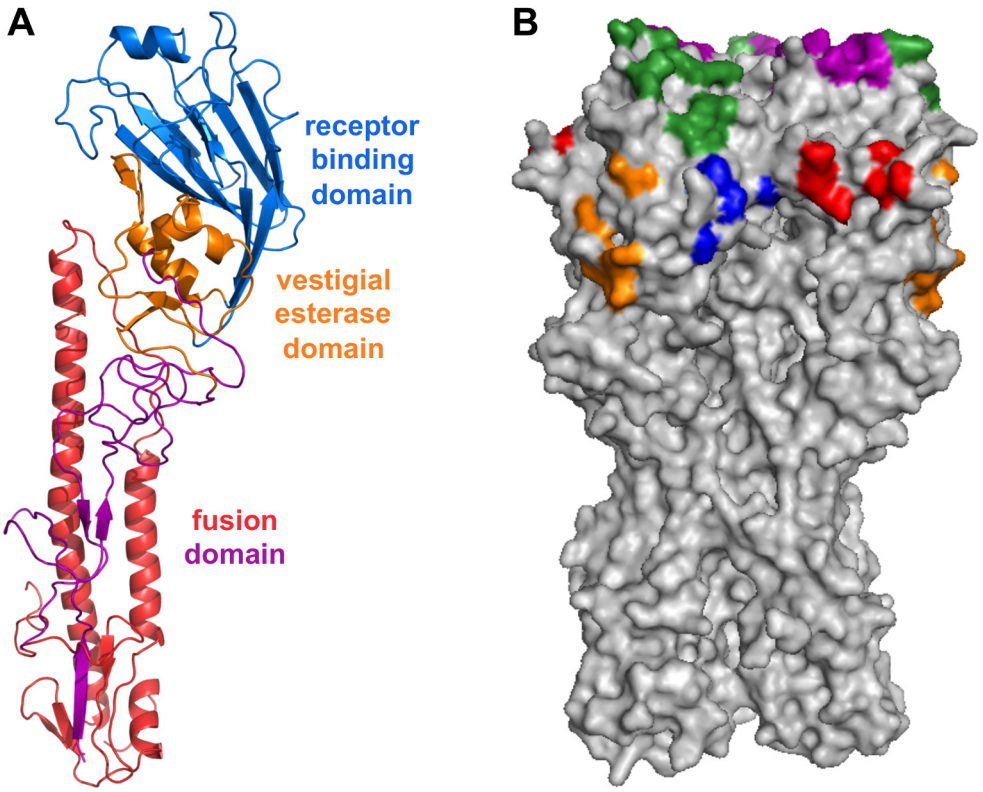

Figure 1.2: **Structure, function, and classic antigenic characterization of HA. (A)** Conserved domain structure of HA. The receptor-binding domain and vestigial esterase domain comprise the globular head domain; the fusion domain (also referred to as the stalk region) is comprised of portions of both the N-terminal (purple) and C-terminal (red) regions of HA; the C-terminal transmembrane and intracellular domains are not resolved in the structure. The receptor-binding domain is inserted into a surface loop of the vestigial esterase domain, and the vestigial esterase domain is inserted into a surface loop of the fusion domain. In distant homologs of hemagglutinin, such as the influenza C hemagglutinin-esterase-fusion protein, the esterase domain is functional and enzymatically destroys the receptor; in influenza A and B this function is served by the neuraminidase protein [114]. **(B)** Classical antigenic mapping of HA was performed by selecting escape mutant viruses with a panel of monoclonal antibodies. Sites where escape mutations were identified are colored by antigenic region (Sa: green, Sb: purple, Ca1: blue, Ca2: red, Cb: orange) [53, 23].

extent do site-specific amino-acid preferences change as a protein's sequence evolves? To what extent do these evolutionary constraints measured in one viral strain describe the evolution of more distantly-related viruses? This is the focus of **Chapter 2**, where, in a collaboration with Bloom Lab postdoctoral fellow Orr Ashenberg, I develop a statistical framework for comparing the site-specific amino-acid preferences measured by deep mu-

tational scans of the influenza NP from two human influenza strains: the PR/1934(H1N1) NP and the Aichi/1968(H3N2) NP. This work is the first comparative study of the effects of *all* amino-acid mutations to two homologous proteins.

We find that the effects of mutations to NP are conserved among most of the sites the protein, although we also identify a few sites with large shifts in their amino-acid preferences. Using the amino-acid preferences as substitution models for evolutionary analysis, we show that the data from either homolog can accurately describe the evolution of NP from a wide range of influenza viruses infecting humans, pigs, horses, and birds. Strikingly, when we average the amino-acid preferences from the two homologs in our experiments, we achieve the best evolutionary model for all of these influenza viruses, likely through a combination of averaging over experimental noise at some sites and capturing the changes in preferences at a few sites. While performing these analyses, the importance of accounting for the substantial amount of experimental noise between replicate deep mutational scanning experiments became clear, underscoring the need for an improved approach with less variability between identical experiments.

Motivated by this need to improve the reproducibility of deep mutational scanning of influenza, in **Chapter 3** I describe technical improvements to the generation of influenza mutant HA virus libraries, their selection for function, and accurate deep sequencing. The largest improvement comes from the use of a "helper virus" lacking the HA gene to more efficiently generate viruses carrying mutant HA genes. These improvements provide more accurate and reproducible measurements of the effects of mutations to HA, strengthening the conclusions from a previous deep mutational scan on HA made by previous Bloom Lab postdoctoral fellow Bargavi Thyagarajan [123] that antigenic sites in the head domain of HA are more tolerant of mutations than the rest of the protein. A separate question is whether the sites in the stalk region of HA, where rare broadly neutralizing antibodies bind, are conserved in natural evolution because they are intolerant of mutations, or because they aren't under strong immune pressure. I show that antibody epitopes in this stalk region of HA are inherently less tolerant to mutation than the rest of the protein.

The work in **Chapters 2 and 3** is solely based on measuring the effects of mutations during selection for viral replication in cell culture. Obviously, this fails to capture other relevant selective pressures on influenza, one of which is the presence of neutralizing antibodies targeting HA. In **Chapter 4**, I develop a new experimental approach that uses the mutant HA libraries described in **Chapter 3**, which carry all tolerated mutations to HA, to completely map the mutational potential for escape from neutralizing antibodies. The technique, which I refer to as "mutational antigenic profiling", identifies all tolerated mutations conferring strong escape from a neutralizing antibody. In contrast to existing high-throughput approaches to identifying antibody epitopes, mutational antigenic profiling identifies all amino-acid mutations that confer escape from neutralization in the context of authentically-displayed hemagglutinin on the surface of infectious influenza virus. I apply this to mapping escape mutations for four monoclonal antibodies and observe striking mutation-specific effects at sites targeted by each antibody. At many of these so-called 'antigenic sites' within the epitope of each antibody, only a handful of the possible mutations actually confer escape, and a common site targeted by two similar antibodies exhibits a unique mutational profile of escape for each, underscoring the complexities involved in predicting the antigenic consequences of mutations in antigenic sites.

Chapter 2

SITE-SPECIFIC AMINO-ACID PREFERENCES ARE MOSTLY CONSERVED IN TWO CLOSELY RELATED PROTEIN HOMOLOGS

A version of this chapter has been previously published as:

Michael B Doud, **Orr Ashenberg**, and Jesse D Bloom. Site-specific amino-acid preferences are mostly conserved in two closely related protein homologs. *Molecular Biology and Evolution,* 32 (11): 2944-2960 (2015).

Bold face indicates equal contributors.

Orr Ashenberg, Jesse Bloom, and I performed the experimental work. I developed statistical methods to compare deep mutational scanning datasets between homologs with contributions from Orr Ashenberg and Jesse Bloom. Orr Ashenberg performed phylogenetic analyses. Orr Ashenberg, Jesse Bloom, and I wrote the manuscript.

2.1 Abstract

Evolution drives changes in a protein's sequence over time. The extent to which these changes in sequence lead to shifts in the underlying preference for each amino acid at each site is an important question with implications for comparative sequence-analysis methods such as molecular phylogenetics. To quantify the extent that site-specific amino-acid preferences shift during evolution, we performed deep mutational scanning on two homologs of human influenza nucleoprotein with 94% amino-acid identity. We found that only a modest fraction of sites exhibited shifts in amino-acid preferences that exceeded the noise in our experiments. Furthermore, even among sites that did exhibit detectable shifts, the magnitude tended to be small relative to differences between non-homologous proteins. Given the limited change in amino-acid preferences between these close homologs, we tested whether our measurements could inform site-specific substitution models that describe the evolution of nucleoproteins from more diverse influenza viruses. We found that site-specific evolutionary models informed by our experiments greatly outperformed non-site-specific alternatives in fitting phylogenies of nucleoproteins from human, swine, equine, and avian influenza. Combining the experimental data from both homologs improved phylogenetic fit, partly because measurements in multiple genetic contexts better captured the evolutionary average of the amino-acid preferences for sites with shifting preferences. Our results show that site-specific amino-acid preferences are sufficiently conserved that measuring mutational effects in one protein provides information that can improve quantitative evolutionary modeling of nearby homologs.

2.2 Background

Since the first comparative analyses of homologous proteins by Zuckerkandl and Pauling [142] fifty years ago, it has been obvious that different sites in proteins evolve under different constraints, with some sites substituting to a wide range of amino acids, while others are constrained to one or a few identities. Zuckerkandl and Pauling [142] pro-

posed, and decades of subsequent work have confirmed [39, 61], that these constraints arise from the highly cooperative interactions among sites that shape important protein properties such as stability, folding kinetics, and biochemical function.

The complexity and among-sites cooperativity of these evolutionary constraints mean that a mutation at a single site can in principle shift the amino-acid preferences of any other site – and numerous experiments have demonstrated examples of such epistasis among sites [129, 96, 34, 87, 56, 92, 101]. However, experiments have also shown that despite such epistasis, the amino-acid preferences of many sites are similar across homologs [111, 5, 116]. For instance, protein structures themselves are highly conserved during evolution [27, 115], and sites in specific structural contexts often have strong propensities for certain amino acids [28, 110, 85]. Furthermore, many of the most successful methods for identifying distant homologs (e.g. `PSI-BLAST`) utilize site-specific scoring models [66, 2], implying that amino-acid preferences are at least somewhat conserved even among homologs with low sequence identity.

A half-century of work has therefore made it abundantly clear that site-specific amino-acid preferences can in principle shift arbitrarily during evolution, but nonetheless in practice remain somewhat conserved among homologs. The important remaining question is the *extent* to which site-specific amino-acid preferences are conserved versus shifted. This question is especially important for the development of quantitative evolutionary models for tasks such as phylogenetic inference. Initially, phylogenetic models unrealistically assumed that sites within proteins evolved both independently and under identical constraints. But more recent models have relaxed the second assumption that sites evolve identically. At first, this relaxation only allowed sites to evolve at different rates [136]. But newer models also accommodate variation in the amino-acid preferences among sites, either by treating these preferences as parameters of the substitution model [79, 82, 127, 113] or by leveraging their direct measurement by high-throughput experiments [15, 16]. Because these models retain the assumption of independence among sites, they will outperform traditional non-site-specific models only if site-specific

amino-acid preferences are substantially conserved among homologs.

Here we perform the first experimental quantification of the conservation of the amino-acid preferences at all sites in two homologous proteins. We do this by using deep mutational scanning [48, 49] to comprehensively measure the effects of all mutations to two homologs of influenza nucleoprotein (NP) with 94% sequence identity. We find that the amino-acid preferences are substantially conserved at most sites in the homologs, but some sites have significant shifts in preferences. We then test whether the experimentally measured site-specific amino-acid preferences can inform site-specific phylogenetic substitution models that describe the evolution of more diverged NP homologs. We find that the experimentally informed site-specific substitution models exhibit improved fit to NP phylogenies containing diverged sequences from human, swine, equine, and avian influenza lineages. Overall, our work shows that site-specific amino-acid preferences are sufficiently conserved that measurements on one homolog can be used to improve the quantitative evolutionary modeling of closely related homologs.

2.3 Results

2.3.1 Comparison of amino-acid preferences between two homologs

Deep mutational scanning of two influenza NP homologs. Our studies focused on NP from influenza A virus. NP performs several conserved functions that are essential for the viral life cycle, including encapsidation of viral RNA into ribonucleoprotein complexes for transcription, viral-genome replication, and viral-genome trafficking [43]. NP's structure is highly conserved in all characterized influenza strains [138, 36]. Our studies compared the site-specific amino-acid preferences of NP homologs from two human influenza strains, PR/1934 (H1N1) and Aichi/1968 (H3N2) (Figure 2.1). These NPs have diverged by over 30 years of evolution, and differ at 30 of their 498 residues (94% protein sequence identity).

We used our previously described approach for deep mutational scanning of influenza

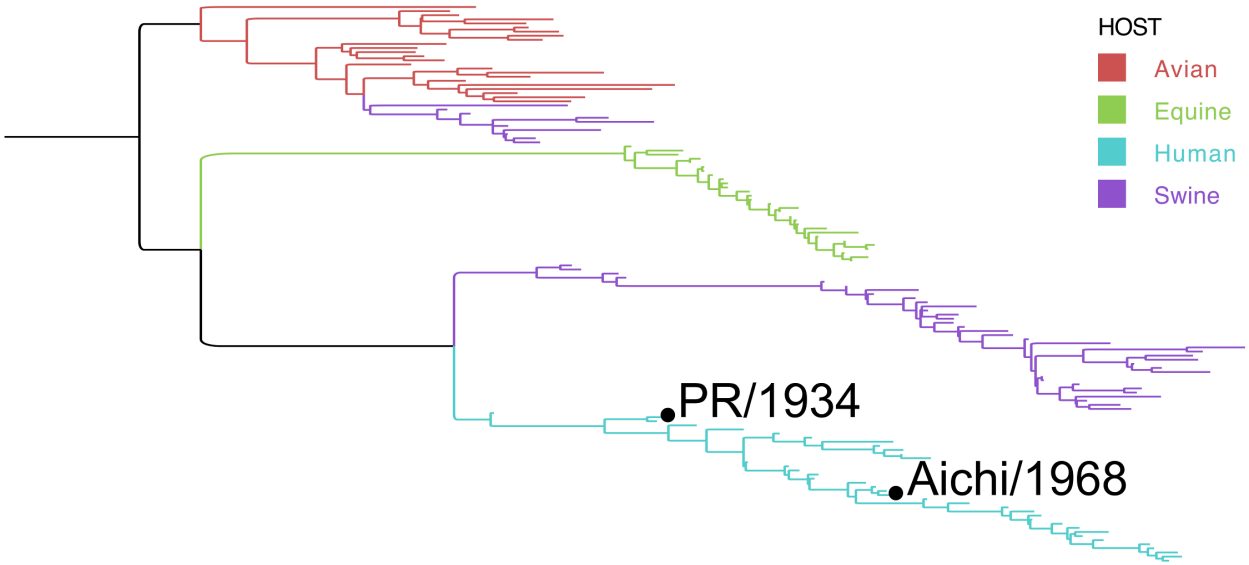

Figure 2.1: **Phylogenetic tree of influenza NPs.** The two homologs used in this work are labeled on the human influenza lineage. A diverse set of sequences was collected by sampling across years and hosts, and a maximum-likelihood tree was inferred using `CodonPhyML` [54] with the codon substitution model of Goldman and Yang [55]. The tree was rooted using the avian clade as an outgroup. The scale bar is in units of codon substitutions per site.

genes [15, 123] to measure the site-specific amino acid preferences of the PR/1934 and Aichi/1968 NPs. Briefly, this approach involved using a PCR-based technique to create mutant libraries of plasmids encoding NP genes with random codon mutations, using reverse genetics to incorporate these mutant genes into influenza viruses, and then passaging these viruses at low multiplicity of infection to select for viruses carrying functional NP variants. Deep sequencing was used to count the occurrences of each mutation before and after selection, and the amino-acid preferences for each site were inferred from these counts using `dms_tools` [17] (Figure A.1, Figure A.2). Our mutagenesis randomized 497 of the 498 codons in NP (the N-terminal methionine was not mutagenized), and so our libraries sampled all $497 \times 19 = 9,443$ amino-acid mutations at these sites. Our mutagenesis introduced an average of about two codon mutations per gene, with the number of mutations per gene following a roughly Poisson distribution (Figure A.4), and so the effect of each mutation was assayed both alone and in the background of variants that

contained one or more additional mutations.

Because deep mutational scanning is subject to substantial experimental noise, we performed several full biological replicates for each NP homolog, beginning with independent creation of the plasmid mutant library. In the current work, we performed three replicates of deep mutational scanning on the PR/1934 NP and two replicates on the Aichi/1968 NP. In a previous study [15] we performed eight replicates of deep mutational scanning on Aichi/1968 NP. We will refer to these previous replicates of the Aichi/1968 NP deep mutational scanning as the *previous study*, and the two new replicates as the *current study*. When not otherwise noted, we refer to the pooled data of all ten of these replicates simply as *Aichi/1968*.

Amino-acid preferences are well correlated between homologs. For each homolog we averaged the site-specific amino-acid preferences across all replicates and examined the correlations of the preferences for each of the 20 amino acids at each of the 497 sites we mutagenized (all sites can be unambiguously aligned between homologs). The mean preferences for the two NP homologs have a Pearson's correlation coefficient of 0.78 (Figure 2.2A). In comparison, the correlation between the preferences measured in the *previous study* and *current study* on the Aichi/1968 homolog is 0.83 (Figure 2.2B). Therefore, the amino-acid preferences correlate nearly as well between the two homologs as they do between different experiments on the same homolog. As expected, there is no correlation between the preferences of the PR/1934 NP and a non-homologous protein (hemagglutinin, HA) for which we have previously measured the site-specific amino-acid preferences using the same approach as in this work [123] (Figure 2.2C).

We also asked if the site-specific amino-acid preferences from each replicate showed the same pattern of correlation between homologs that we observed when comparing mean preferences. We again found that correlation coefficients are just as high between NP homologs as they are between replicate measurements on the same homolog, and that there is no correlation between the preferences for NP and the non-homologous

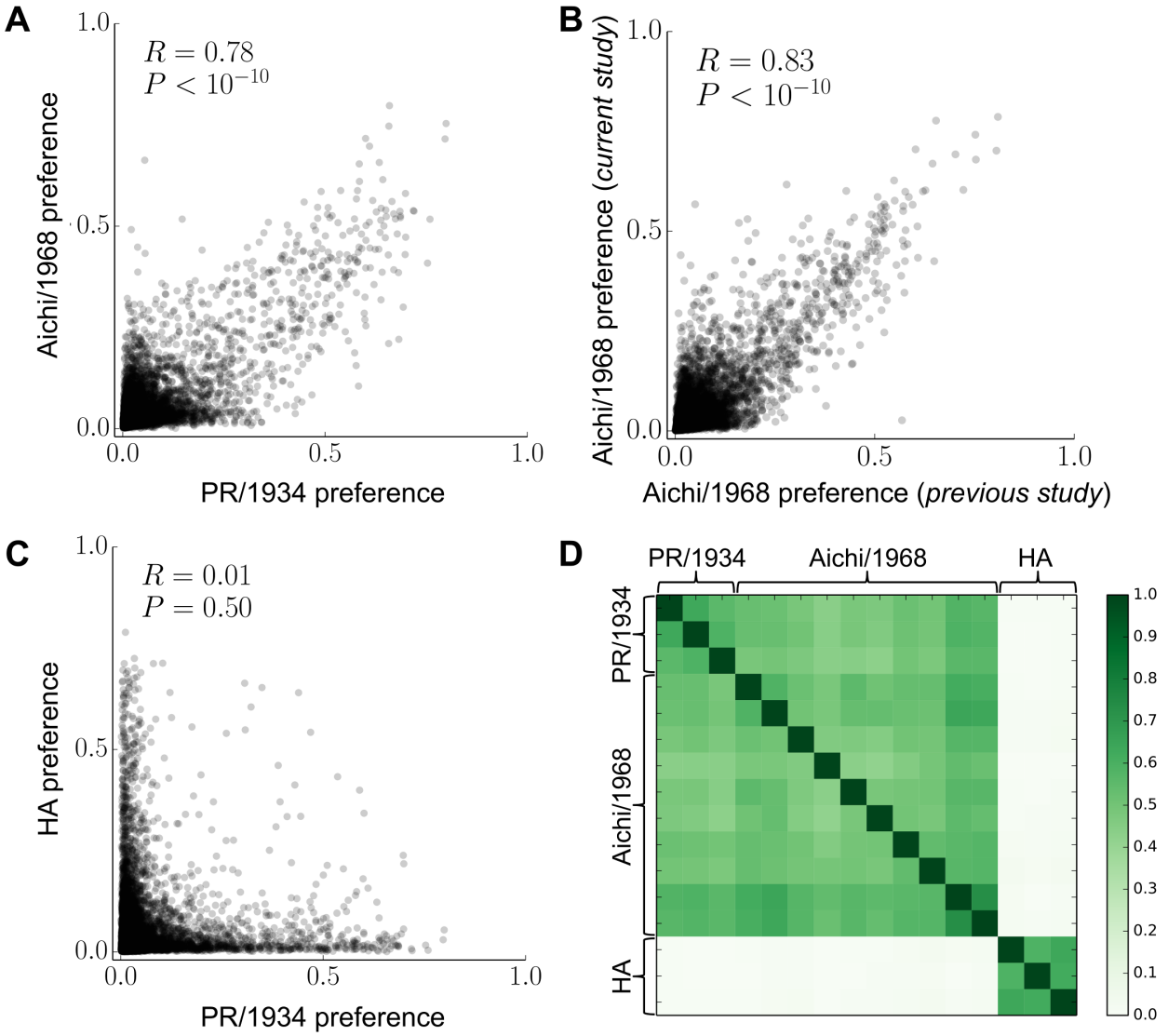

Figure 2.2: **Site-specific amino-acid preferences correlate nearly as well between NP homologs as between replicate measurements on the same homolog. (A), (B)** The correlation between the mean of the preferences taken over all replicates on each NP homolog is nearly as large as that between the preferences measured in the *current study* and *previous study* on the Aichi/1968 NP. **(C)** However, there is no correlation between the preferences measured for NP and the non-homologous protein HA. Each data point in (A)-(C) is the preference for one of the twenty amino acids at one of the 497 sites in NP. R is the Pearson correlation coefficient. **(D)** The Pearson correlations between the preferences measured in all pairs of individual replicates. Comparisons between NP and HA were made based on position in primary sequence for sites 2 through 498.

protein HA (Figure 2.2D). Overall, these results indicate that at the vast majority of sites, any differences in the amino-acid preferences between NP homologs are smaller than the noise in our experimental measurements, and vastly smaller than the differences between non-homologous proteins.

Shifts in amino-acid preferences are small for most sites. The previous section shows that any widespread shifts in site-specific amino-acid preferences are smaller than the noise in our experiments. However, it remains possible that a subset of sites show substantial shifts in their amino-acid preferences that are masked by examining all sites together. We therefore performed an analysis to identify specific sites with shifted amino-acid preferences between homologs.

This analysis needed to account for the fact that experimental noise induced variation in the preferences measured in each replicate. Figure 2.3 shows replicate measurements for both homologs at several sites in NP. At many sites, such as site 298, all replicate measurements yielded highly reproducible amino-acid preferences both between and within homologs. At many other sites, such as site 3, replicate measurements were quite variable both between and within homologs, probably due to fairly weak selection at that site. Some sites, like site 254, exhibited reproducible measurements within each homolog, and the most preferred amino acid was the same in both homologs, but the tolerance for mutations to other amino acids was distinct in each homolog. Finally, at a few sites, most prominently site 470, replicate measurements were highly reproducible within each homolog but clearly differed in which amino acid was most preferred between homologs. We therefore developed a quantitative measure of the shift in preferences between homologs that accounts for this site-specific experimental noise.

We used the Jensen-Shannon distance metric (the square root of the Jensen-Shannon divergence) to quantify the distance between the 20-dimensional vectors of amino-acid preferences for each pair of replicate measurements at each site. This distance ranges from zero (identical amino-acid preferences) to one (completely different amino-acid pref-

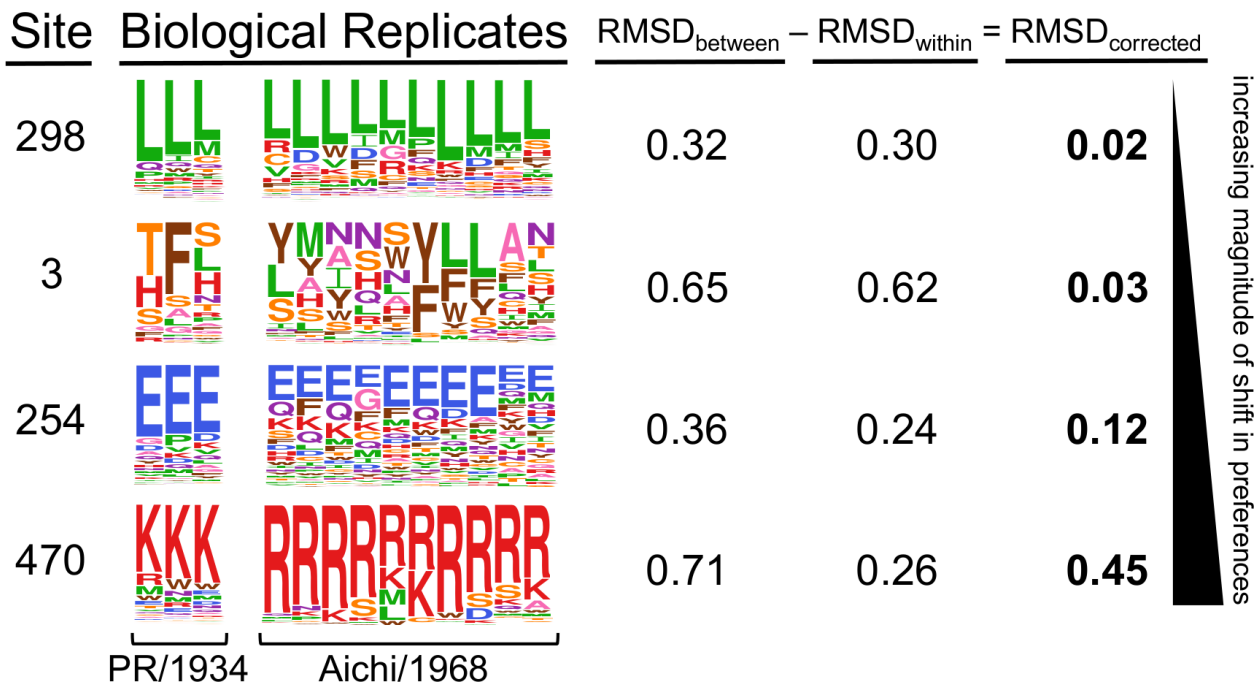

Figure 2.3: **Replicate measurements quantify the shift in amino-acid preferences between homologs after correcting for experimental noise.** The amino-acid preferences measured in multiple replicates of deep mutational scanning of both homologs are shown for selected sites ordered by the magnitude of preference change observed after correcting for site-specific noise. $RMSD_{between}$ (the average difference between the two homologs) and $RMSD_{within}$ (the average variation within replicates of each homolog) are shown to the right. $RMSD_{corrected}$ is calculated by subtracting $RMSD_{within}$ from $RMSD_{between}$.

erences). To quantify experimental noise at a site, we calculated the root-mean-square of the Jensen-Shannon distance for all pairwise comparisons among replicate measurements on the same homolog, and termed this quantity $RMSD_{within}$. Sites with large $RMSD_{within}$ have high experimental noise. We defined an analogous statistic, $RMSD_{between}$, to quantify the distance in preferences between homologs by calculating the root-mean-square of the Jensen-Shannon distance for all pairwise comparisons between replicates of PR/1934 and replicates of Aichi/1968. Figure 2.3 shows the values of these statistics for example sites.

The fact that we had data from two independent sets of experiments on the Aichi/1968 NP (the *current study* and *previous study*) enabled us to perform a control analysis by

calculating $RMSD_{between}$ and $RMSD_{within}$ for the replicates from these two experiments. As an additional control to gauge the extent of amino-acid preference differences between non-homologous proteins, we also calculated $RMSD_{between}$ and $RMSD_{within}$ for our experiments on Aichi/1968 NP and our previous study on HA (note that because NP and HA are non-homologous, they cannot be meaningfully aligned, so this control comparison simply pairs each site in NP with the corresponding residue number in HA).

The relationship between $RMSD_{between}$ (the observed difference between homologs) and $RMSD_{within}$ (the observed variation in repeated measurements on the same homolog) for all sites is shown for several different comparisons in Figure 2.4A-C. Sites with low $RMSD_{within}$ exhibit highly reproducible measurements between replicate experiments, whereas sites with higher values of $RMSD_{within}$ exhibit substantial experimental noise, probably due to weak selection at that site. Sites with large $RMSD_{between}$ exhibit amino-acid preference differences between homologs, but at each site some of this observed variation is due to the site-specific experimental noise (quantified by $RMSD_{within}$) rather than a true difference between the homologs.

When comparing two independent experiments on the same NP (Figure 2.4A) or comparing experiments on two homologs of NP (Figure 2.4B), the relationship between $RMSD_{between}$ and $RMSD_{within}$ is approximately linear, indicating that the difference in amino-acid preferences between homologs at a given site is usually comparable to the experimental noise. Deviations from this linear relationship are more frequent in the comparison between PR/1934 and Aichi/1968 (Figure 2.4B) than in the comparison between the two studies of Aichi/1968 (Figure 2.4A). These deviations mostly arise from sites that have larger $RMSD_{between}$ than $RMSD_{within}$, indicating that these sites have shifts in their amino-acid preferences between homologs that exceed the experimental noise. These results comparing NP homologs are in stark contrast with the $RMSD_{between}$ and $RMSD_{within}$ calculated when comparing NP to the non-homologous HA (Figure 2.4C), where the difference between proteins is almost always substantially greater than the experimental noise.

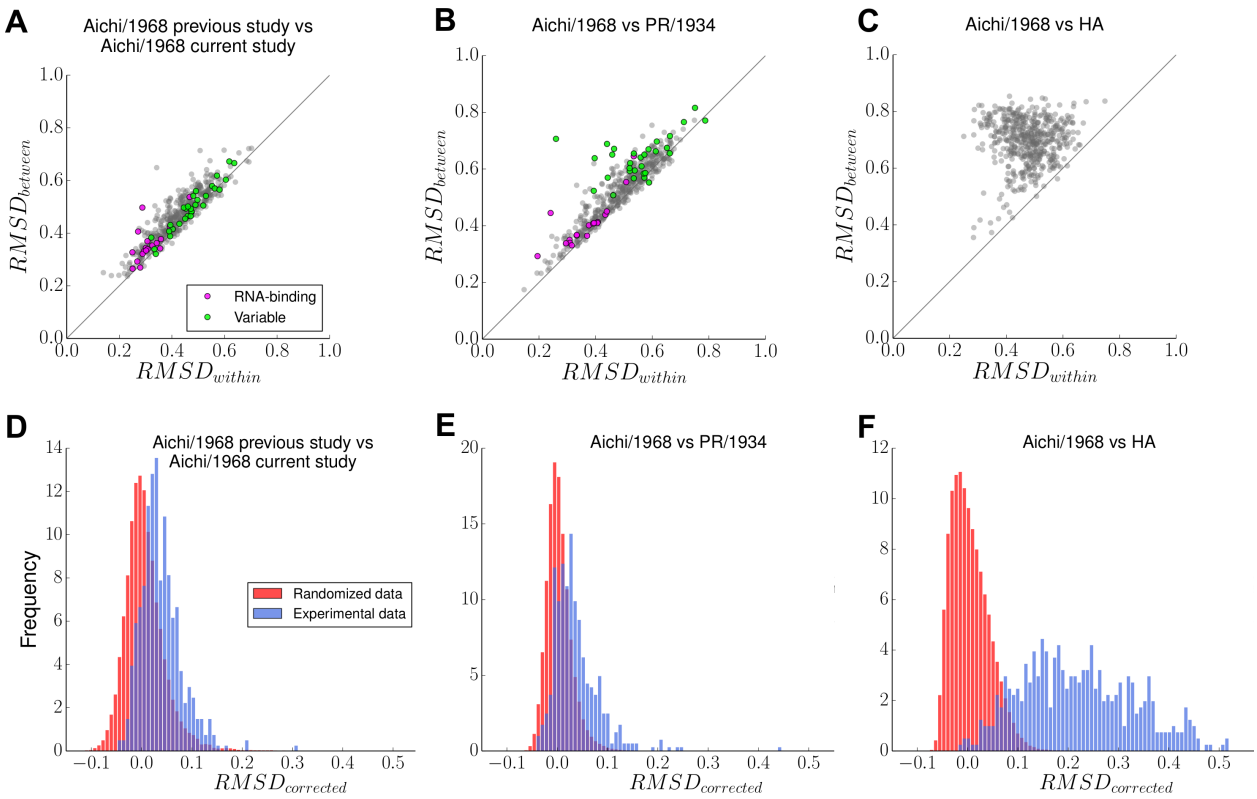

Figure 2.4: **Identification of sites with shifts in amino-acid preferences. (A)-(C)** Each plot shows statistics calculated for a comparison between two groups of replicate experiments. Each point represents a site in NP. $RMSD_{within}$ quantifies the average difference in amino-acid preferences within each of the two groups (experimental noise), and $RMSD_{between}$ quantifies the average difference in preferences between the two groups. Points above the $y = x$ diagonal represent sites with preference changes between homologs greater than experimental noise. Sites in the RNA-binding groove are in purple; sites that have different wild-type identities in PR/1934 and Aichi/1968 are in green. **(D)-(F)** The actual distribution of $RMSD_{corrected}$ values is shown in blue, and the distribution of $RMSD_{corrected}$ from data randomized between comparison groups is shown in red. Comparisons are made between the two studies on Aichi/1968 (A, D), between Aichi/1968 and PR/1934 (B, E), and between Aichi/1968 and the non-homologous HA (C, F).

To quantify the extent of amino-acid preference shifts between the two homologs in a way that corrects for the experimental noise, we defined another statistic, $RMSD_{corrected}$, by subtracting $RMSD_{within}$ from $RMSD_{between}$ (Figure 2.3). Sites with shifts in amino-acid preferences greater than the experimental noise have $RMSD_{corrected} > 0$. However, we also expect many sites to have positive $RMSD_{corrected}$ values due to statistical noise. To determine the distribution of $RMSD_{corrected}$ values expected due to such statistical

noise alone under the null hypothesis that the amino-acid preferences are the same in both groups being compared, we generated null distributions of $RMSD_{corrected}$ using exact randomization testing by shuffling which experimental replicates were assigned to which NP homolog. For every possible shuffling of replicates, we computed $RMSD_{corrected}$ at every site and combined the results across all shufflings.

The distribution of $RMSD_{corrected}$ obtained experimentally mostly overlaps the randomized distribution of $RMSD_{corrected}$ when comparing the two independent Aichi/1968 experiments (Figure 2.4D). This overlap is consistent with the hypothesis that the true amino-acid preferences are the same in both experiments on the Aichi/1968 NP. In contrast, when comparing PR/1934 to Aichi/1968, some $RMSD_{corrected}$ values are shifted in the positive direction substantially beyond the null distribution (Figure 2.4E), indicating larger differences in preferences at some sites than can be explained by experimental noise alone. This shift in preferences is particularly notable for site 470, which has a $RMSD_{corrected}$ of 0.45 as illustrated in Figure 2.3. However, most sites still fall within the null distribution when comparing the two NP homologs. In contrast, if NP is compared to the non-homologous HA, the vast majority of sites exhibit differences in preferences that vastly exceed the values expected under the null distribution (Figure 2.4F).

As an alternative approach to generating null distributions of $RMSD_{corrected}$, we performed simulations of observed amino-acid preferences in each replicate under a model where there are no differences in the underlying preferences between the two homologs, but varying levels of noise for each experiment. We simulated amino-acid preferences at each site by drawing from a Dirichlet distribution, which is well-suited for this purpose because its support is a normalized vector of values, in this case corresponding to the vector of amino-acid preferences at a site. Our null hypothesis is that the amino-acid preferences are the same for both homologs, so we performed simulations assuming that the *true* vector of amino-acid preferences at a site is equal to the average of our experimental measurements for both homologs. We simulated the amino-acid preferences for each replicate by drawing from a Dirichlet distribution centered on this vector of assumed true

preferences. The extent to which any given sample drawn from this Dirichlet distribution differs from the true vector can be tuned with a single scaling parameter (the concentration parameter). We identified a value for the concentration parameter for each experiment (Aichi/1968 *current study*, Aichi/1968 *previous study*, and PR/1934) that resulted in correlation coefficients between replicates that matched those in the actual experiment. We performed 1000 replicate simulations and combined the calculated $RMSD_{corrected}$ values from all simulations to build the null distribution. The distributions of $RMSD_{corrected}$ obtained by simulation are in Figure A.5 and are similar to those obtained using exact randomization testing.

Sites with clear shifts in amino-acid preferences. Using either of the two null distributions, we were able to identify specific sites with $RMSD_{corrected}$ values significantly larger than expected due to experimental noise alone. These are sites for which we can reject the null hypothesis that there is no shift in amino-acid preferences. To control for multiple hypothesis testing, we set a false discovery rate (proportion of rejected null hypotheses expected to be falsely rejected) of 5%.

Using exact randomization as a null distribution, we could reject the null hypothesis of no shift in amino-acid preference for 14 of the 497 sites. The simulated-data null distribution appeared to afford greater statistical power, and allowed us to reject the null hypothesis of no shift in preference for 76 sites (the 14 identified by the exact randomization plus an additional 62). Many of these additional sites, however, exhibit shifts that are small in magnitude; for instance, 30 of the additional 62 sites show a pattern similar to that of site 254 (Figure 2.3), where the most preferred amino acid is unchanged, but the tolerance for mutations to other residues is somewhat larger in one homolog than the other.

Figure 2.4 provides a more visual way to gauge the magnitude of the shifts in amino-acid preferences. If the preferences are completely conserved among homologs, the actual distribution in Figure 2.4E should look roughly like that in Figure 2.4D. In contrast, if

the preferences have completely shifted between homologs, the actual distribution should look more like that in Figure 2.4F. As is clear from visual inspection, only a handful of sites have amino-acid preferences that have shifted between the PR/1934 and Aichi/1968 homologs to be as different as is typical for pairs of sites from non-homologous proteins. The rest of the sites either exhibit a more modest shift in preference (this is the case for 14 or 76 sites depending on which null distribution is used) or no detectable shift in preference.

An important question is whether there are common characteristics of sites with shifted preferences. One reasonable hypothesis is that sites with wild-type amino-acid identities that differ between the homologs are more likely to have experienced shifts in their amino-acid preferences. Among the 14 sites identified as shifted by both null distributions, 5 have different wild-type amino-acid identities in PR/1934 and Aichi/1968 (Figure 2.5A). Therefore, of sites with variable amino-acid identity between the two homologs, 17% exhibit clear shifts in preference identified by both null distributions, while only 2% of conserved sites exhibit comparable shifts.

Having identified evolutionarily variable sites as enriched for the clearest shifts in amino-acid preferences, we next looked at sites with other special structural or functional properties. One group of functionally important sites are those that comprise the RNA-binding groove of NP. These RNA-binding sites have low $RMSD_{within}$ (Figure 2.4A and B), indicating below-average noise among replicates. RNA-binding sites also have low $RMSD_{corrected}$ (Figure 2.5B, Figure 2.6). These results are consistent with the expectation that RNA-binding sites in NP are under strong and conserved functional constraint, since RNA binding is essential for viral genome packing, transcription, and replication.

We next hypothesized that sites in structural proximity to evolutionarily variable sites may experience shifts in amino-acid preferences due to changes in the surrounding biochemical environment. We identified sites directly contacting the evolutionarily variable residues, and found that they do not have $RMSD_{corrected}$ values that differ from other sites (Figure 2.5B). Therefore, we are unable to identify any preferential tendency for substi-

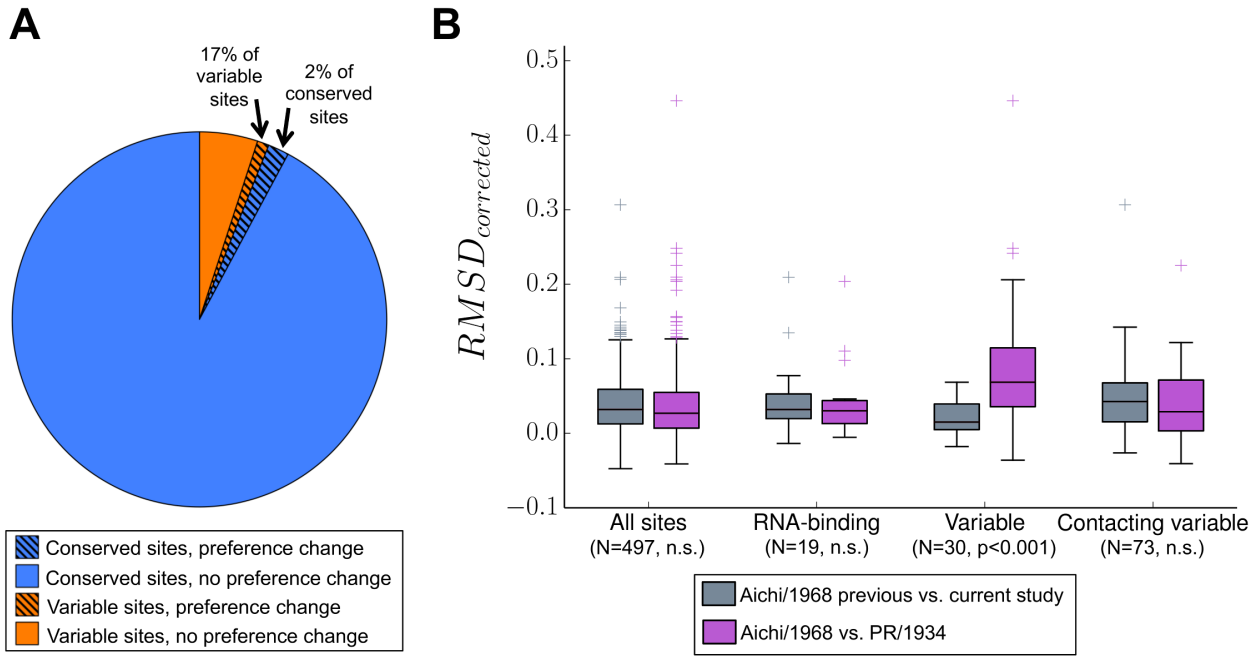

Figure 2.5: **Evolutionarily variable sites are enriched for changes in amino-acid preference.** **(A)** Sites with shifts in amino-acid preferences were identified by $RMSD_{corrected}$ values greater than expected under a null model assuming no difference between homologs (false discovery rate of 5% using a null model generated by exact randomization testing). *Variable* sites have different wild-type residues in the two NP homologs. **(B)** The distributions of $RMSD_{corrected}$ for various groups of sites. The median is marked by a horizontal line, boxes extend from 25th to 75th percentile, and whiskers extend to data points within 1.5 times the interquartile range. Outliers are marked with crosses. *Contacting variable* sites are conserved sites with side-chain atoms within 4.5 Ångströms of a variable side-chain atom. $RMSD_{corrected}$ distributions for each group of sites are shown for two comparisons: one comparing two independent experiments on Aichi/1968, and one comparing Aichi/1968 to PR/1934. P-values were determined using the Mann-Whitney U test and adjusted using the Bonferroni correction.

tutions to drive shifts in amino-acid preference at other sites in direct contact with the substituted residue.

The 14 sites with the clearest shifts in amino-acid preferences are distributed throughout the surface of NP in the body, head, and tail loop domains (Figure 2.6). Six of the 14 sites are located in the flexible tail loop, which inserts into a neighboring monomer during NP oligimerization. This suggestive clustering led us to test whether there was a significant tendency for the 14 sites with clearest shifts in preferences to be spatially

clustered in NP's structure. We calculated the distance between sites as the minimum distance between side chain atoms (using the alpha carbon for glycine). Eleven of the 14 sites with clearest shifts in preferences are resolved in the crystal structure, and of these 11 sites the median distance to the nearest neighbor among the 10 remaining sites is 5.8 Ångströms, which is significantly less than expected by chance for random selections of 11 sites (10.8 Ångströms, p=0.028). Thus, the clearest shifts in preferences between these two homologs occur in small clusters of proximal sites more often than in single isolated sites. This pattern also holds when considering the 76 sites identified by the simulation null distribution: among the 66 that are resolved in the crystal structure the median distance to the nearest neighbor is 4.5 Ångströms compared to a median distance to nearest neighbor among random selections of 66 sites of 5.0 Ångströms (p = 0.021). Therefore, sites with shifted preferences appear to cluster in NP's structure, even if they are not usually in direct physical contact with variable residues.

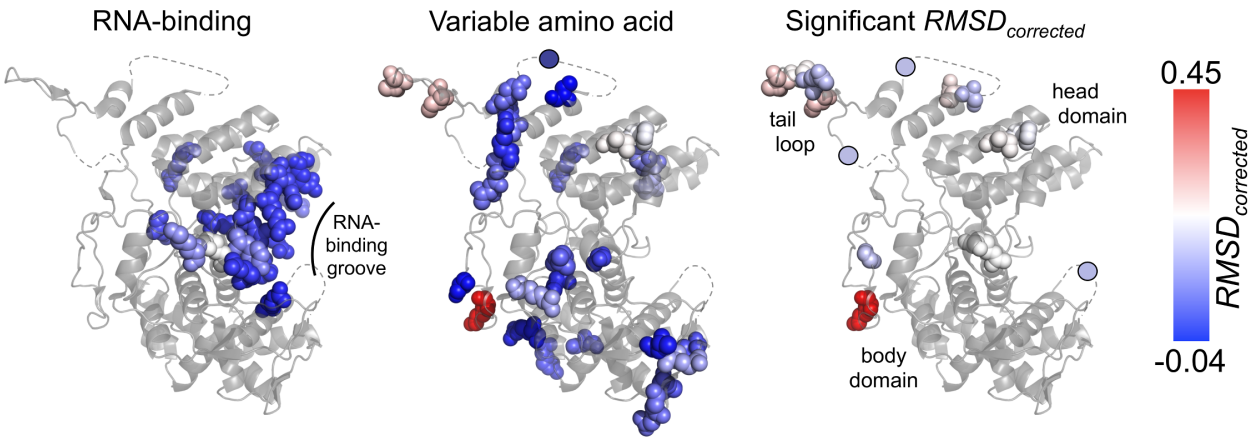

Figure 2.6: **Magnitude of the shift in amino-acid preferences mapped on the NP structure.** $RMSD_{corrected}$ values for each site are used to color space-filling models for the indicated sites in the NP crystal structure [PDB ID 2IQH, chain C; 138]). Sites are shown as circles when in regions that are not present in the crystal structure (dashed lines). Blue represents small shifts in amino-acid preferences between PR/1934 and Aichi/1968; red represents large shifts. *Variable amino acid* refers to sites where the wild-type residue differs between PR/1934 and Aichi/1968 NP. *Largest preference changes* refers to sites where the null hypothesis is rejected using exact randomization testing with a false discovery rate of 5%.

Overall, these results indicate that sites with evolutionarily variable amino-acid identity are more likely than conserved sites to exhibit shifts in amino-acid preferences, and that sites with shifted preferences tend to cluster in NP's structure. However, the majority of sites with variable identity do not exhibit large shifts in amino-acid preference, and overall, only between 3% and 15% (depending on the method used to generate the null distribution) of sites in NP undergo shifts in amino-acid preferences that are sufficiently large to justify rejecting the null hypothesis that the preferences are identical between homologs. Importantly, statistical significance does not necessarily imply a large magnitude in effect size – and indeed, with just a handful of exceptions (most prominently site 470), even the shifted sites are vastly more similar in their preferences than typical pairs of sites in non-homologous proteins.

2.3.2 Experimentally informed site-specific substitution models describe vast swaths of nucleoprotein evolution

We next quantitatively assessed how well our experimentally measured amino-acid preferences reflected the actual constraints on NP evolution. To do so, we used the amino-acid preferences to inform site-specific phylogenetic substitution models. We have previously shown that substitution models informed by experimentally measured site-specific amino-acid preferences greatly outperform common non-site-specific codon-substitution models [15, 16, 123].

In the prior work, site-specific amino-acid preferences were experimentally measured in a single sequence context. Here, we asked whether combining the preferences measured in the two different sequence contexts of Aichi/1968 and PR/1934 would more accurately describe NP sequence evolution. Any improvement could be due to two effects: First, a combined substitution model might better reflect the evolutionary average of the amino-acid preferences at sites with significant changes in preferences over time. Second, combining data from multiple experiments should reduce noise and yield more

accurate site-specific amino-acid preferences.

Combining deep mutational scanning datasets from nucleoprotein homologs improves phylogenetic fit. To compare the performance of different substitution models, we used a likelihood-based framework. We first built a maximum-likelihood tree for NP sequences from human influenza using `CodonPhyML` [54] with the codon-substitution model of Goldman and Yang [55] (GY94) (Figure 2.1). We fixed this tree topology and used `HyPhy` to optimize branch lengths and model parameters for each substitution model by maximum likelihood. The relative fits of the substitution models were evaluated using the Akaike information criterion (AIC) [107].

We tested experimentally informed substitution models derived from the Aichi/1968 and PR/1934 mutational scans either alone or in combination. The Aichi/1968 model used amino-acid preferences averaged across the *current study* and the *previous study*. To build a combined substitution model based on both NP homologs, we averaged the amino-acid preferences for the Aichi/1968 and PR/1934 homologs (Aichi/1968 + PR/1934). Each substitution model had five free parameters that were fit by maximum likelihood: four nucleotide mutation rates and a stringency parameter β that accounts for the possibility of a different strength of selection in natural sequence evolution compared to the mutational-scanning experiments [16]. Importantly, the amino-acid preferences themselves are not free parameters, as they are independently measured by experiments that do not utilize information from the naturally occurring NP sequences.

As a comparison to the experimentally informed substitution models, we also tested the non-site-specific GY94 model. Relative to the experimentally informed substitution models, the GY94 model includes more free parameters including equilibrium codon frequencies, a transition-transversion ratio, and parameters describing gamma distributions of the nonsynonymous-synonymous ratio and substitution rate across sites [136, 137].

The Aichi/1968 and PR/1934 experimentally informed models described the human NP phylogeny far better than the non-site-specific GY94 model (Table 2.1). Strikingly,

combining amino-acid preferences from both NP homologs (Aichi/1968 + PR/1934) resulted in a greatly improved substitution model (Table 2.1). For each experimentally informed model, the stringency parameter β fit with value greater than 1 (average $\beta = 2.5$), consistent with the idea that selection during natural evolution is more stringent than our laboratory selection.

Table 2.1: **Combining experimental data improves phylogenetic fit to NPs from human influenza.** Substitution models are sorted by ΔAIC, and the corresponding log likelihoods, number of free parameters, and values of optimized parameters are shown. Log likelihoods for each model were calculated through maximum-likelihood optimization of branch lengths and model parameters given the fixed tree topology of human NPs shown with blue lines in Figure 2.1. The only parameters in the experimentally informed models are the four nucleotide mutation rates and the stringency parameter β. The non-site-specific GY94 model [55] has nine empirical nucleotide equilibrium frequencies [104], and optimized parameters describing the transition-transversion ratio (κ), the gamma distribution of the nonsynonymous-synonymous ratio (ω) [137], and the gamma distribution of substitution rates [136]. In the Aichi/1968 model, the preferences from *current study* and *previous study* have been averaged.

model	ΔAIC	log likelihood	parameters (optimized + empirical)	optimized parameters
Aichi/1968 + PR/1934	0.0	-4395.8	5 (5 + 0)	$R_{A\to G}$ = 4.6, $R_{A\to T}$ = 0.8, $R_{C\to A}$ = 1.4, $R_{C\to G}$ = 0.1, β = 3.0
PR/1934	322.3	-4556.9	5 (5 + 0)	$R_{A\to G}$ = 4.9, $R_{A\to T}$ = 0.8, $R_{C\to A}$ = 1.4, $R_{C\to G}$ = 0.1, β = 2.1
Aichi/1968	485.7	-4638.6	5 (5 + 0)	$R_{A\to G}$ = 4.8, $R_{A\to T}$ = 0.7, $R_{C\to A}$ = 1.4, $R_{C\to G}$ = 0.1, β = 2.4
GY94, gamma ω, gamma rates	2582.3	-5678.9	13 (4 + 9)	κ = 6.2, ω shape = 0.1, mean ω = 0.1, rate shape = 2.4

Experimentally informed models also describe the evolution of more diverged non-human influenza strains. Given the success of the experimentally informed substitution models in describing the human NP phylogeny, we asked whether these models could be extended to more diverged NPs from non-human influenza strains. We expect these models to exhibit good fit if the NP site-specific amino-acid preferences are mostly con-

served across these viral strains. We examined NPs from influenza strains from three hosts: swine, equine, and avian. The average protein-sequence identity between human NPs and swine, equine, and avian NPs was 91%, 91%, and 93% respectively.

We built a phylogenetic tree of NPs of influenza viruses from human, swine, equine, and avian hosts (Figure 2.1). As previously reported, the avian sequences could be divided into western and eastern hemispheric clades, and the swine sequences consisted of the North American Classical H1N1 clade and the more recent Eurasian H1N1 clade [131]. Using this tree, we performed a phylogenetic analysis similar to that described above for human influenza NPs.

Again, the experimentally informed models greatly outperformed the non-site-specific GY94 model, and combining the Aichi/1968 and PR/1934 models resulted in a far superior model (Table 2.2). Since the amino-acid preferences were experimentally measured for human NP, we wanted to ensure that this superior performance was not driven solely by the human clade of the tree. We separately fit subtrees consisting only of swine, equine, or avian NP sequences (Table A.1, Table A.2, Table A.3). Each subtree showed the same trend as the full tree: the experimentally informed models were superior to the GY94 model, and combining data from the two NP homologs resulted in large improvements in likelihood. Therefore, site-specific amino-acid preferences of NP are sufficiently conserved across influenza A lineages that substitution models informed by deep mutational scanning of human influenza NP homologs can be extended to the NPs of influenza from other hosts.

Combining data from NP homologs improves phylogenetic fit to sites with shifted preferences. The results above show that the experimentally informed substitution models improved phylogenetic fit relative to the non-site-specific model, and that combining data from two NP homologs resulted in the best model. This increased performance when combining data may come from more accurate measurement of amino-acid preferences due to more replicates, or from averaging amino-acid preferences over multiple sequence

Table 2.2: **Combining experimental data improves phylogenetic fit to NPs from human, swine, equine, and avian influenza.** This table differs from Table 2.1 in that it fits the combined tree of human, swine, equine, and avian NPs in Figure 2.1.

model	ΔAIC	log like-lihood	parameters (optimized + empirical)	optimized parameters
Aichi/1968 + PR/1934	0.0	-17507.9	5 (5 + 0)	$R_{A\to G}$ = 6.0, $R_{A\to T}$ = 1.0, $R_{C\to A}$ = 1.4, $R_{C\to G}$ = 0.1, β = 2.7
PR/1934	700.2	-17858.0	5 (5 + 0)	$R_{A\to G}$ = 6.3, $R_{A\to T}$ = 1.0, $R_{C\to A}$ = 1.4, $R_{C\to G}$ = 0.1, β = 2.1
Aichi/1968	1030.2	-18023.0	5 (5 + 0)	$R_{A\to G}$ = 6.2, $R_{A\to T}$ = 0.9, $R_{C\to A}$ = 1.4, $R_{C\to G}$ = 0.1, β = 2.3
GY94, gamma ω, gamma rates	1784.7	-18392.2	13 (4 + 9)	κ = 6.9, ω shape = 0.3, mean ω = 0.1, rate shape = 3.1

contexts. To examine these possible explanations, we analyzed which sites in NP were more accurately modeled when the Aichi/1968 and PR/1934 experimental models were combined. This analysis was performed using the full phylogenetic tree of NP sequences (Figure 2.1).

While fixing the branch lengths and model parameters to their maximum-likelihood values for each model, we calculated for each site the difference in likelihoods (Δlog-likelihood) when the site was modeled using the combined Aichi/1968 + PR/1934 model compared to using the Aichi/1968 model. We binned sites into quintiles of Δlog-likelihood. Sites in the top quintile had the greatest increases in likelihood when the Aichi/1968 and PR/1934 models were combined. Overall 67% of sites in NP had increased likelihoods under the Aichi/1968 + PR/1934 model.

To determine whether these improved likelihoods came from lower noise in the combined experimental model, we used the $RMSD_{within}$ statistic. Sites with greater variance in amino-acid preferences across experimental replicates have higher $RMSD_{within}$ scores. We analyzed the distribution of the $RMSD_{within}$ scores for sites within each quintile (Figure 2.7). The top and bottom quintiles did not have significantly different

$RMSD_{within}$ distributions, indicating that sites prone to experimental noise contributed both positively and negatively to the tree likelihood when experimental datasets were combined. Thus, the improved modeling with the combined dataset was not chiefly due to reduced experimental noise.

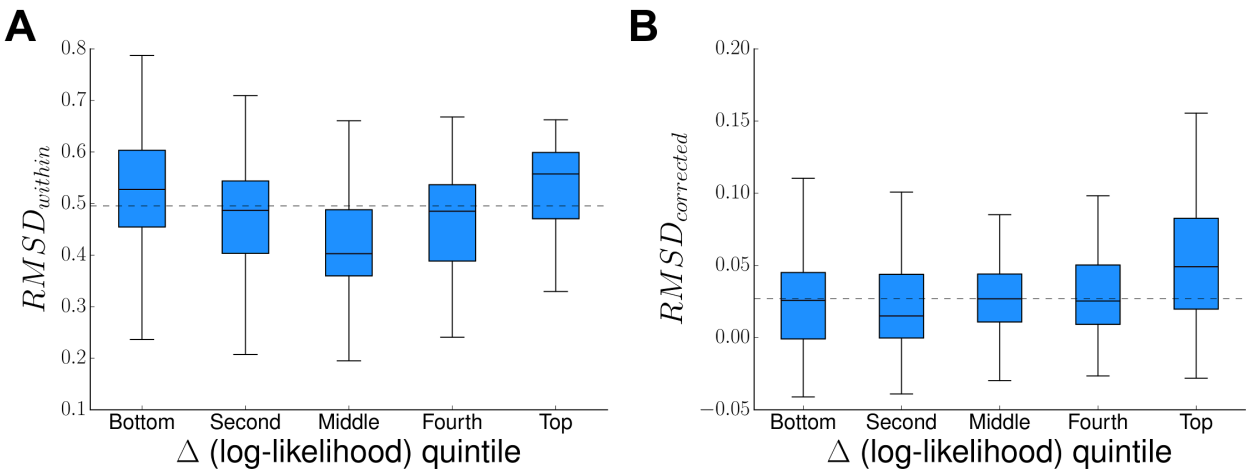

Figure 2.7: **NP sites that are better described by combining data from both homologs have shifted amino-acid preferences.** The change in per-site likelihood in going from the Aichi/1968 model to the Aichi/1968 + PR/1934 model was plotted against the per-site $RMSD_{within}$ (A) or per-site $RMSD_{corrected}$ (B). Sites were ranked by Δ(log-likelihood), divided into quintiles, and the per-site $RMSD_{within}$ or per-site $RMSD_{corrected}$ for sites in each quintile was displayed as a box and whisker plot. Outlier sites beyond the interquartile range are omitted. Quintiles are ordered left to right from least improved likelihoods to most improved likelihoods under the combined model. The median $RMSD_{within}$ or $RMSD_{corrected}$ is shown as a horizontal, dashed line. Sites with the most improved likelihoods did not have significantly higher variation in amino-acid preferences (high $RMSD_{within}$) across replicate measurements on the same homolog. However, these sites did have significantly higher differences in amino-acid preferences between Aichi/1968 and PR/1934 (high $RMSD_{corrected}$).

Next, to determine whether the improved likelihoods were driven by sites with different preferences between the two NP homologs, we used the $RMSD_{corrected}$ statistic (Figure 2.7). If the improvements under the combined model came from sites with different amino-acid preferences between Aichi/1968 and PR/1934, then we would expect that the sites with the greatest increases in likelihood would also have the greatest $RMSD_{corrected}$ values. This was indeed the case, as sites in the top quintile of log-likelihoods had the

highest median $RMSD_{corrected}$. The $RMSD_{corrected}$ scores in the top quintile were significantly different from those in the lower quintiles (Mann-Whitney U with Bonferroni correction $p<0.002$), whereas there were no significant differences in the $RMSD_{corrected}$ scores when comparing the lower quintiles. Therefore, improvements in the combined model were partly due to better describing those sites that had the largest shifts in amino-acid preferences over evolutionary time.

2.4 Discussion

Determining the extent to which site-specific amino-acid preferences shift during evolution is important for evaluating how well experimental measurements can be extrapolated across homologs, and for guiding the development of site-specific phylogenetic substitution models. We have performed the first comprehensive assessment of the conservation of site-specific amino-acid preferences by using deep mutational scanning to measure the effects of all mutations on two closely related homologs of influenza NP.

We found that for the majority of sites, any shift in amino-acid preferences between homologs was smaller than the noise in our experiments. We could reject the null hypothesis that the amino-acid preferences were identical among homologs for only between 3% and 15% of all sites, depending on the method used to generate the null distribution. Furthermore, even for those sites for which we could reject the null hypothesis of identical preferences between homologs, the magnitude of shifts tended to be small. Only a handful of the 497 sites exhibited shifts in preference between homologs with a magnitude comparable to the average difference between sites in non-homologous proteins. Sites that varied in amino-acid identity between the two homologs were more likely to have a detectable shift in amino-acid preferences – but even among variable sites, there was usually no shift. Admittedly, our experiments had substantial noise, so it is likely that other sites have undergone subtle shifts below our limit of detection. However, the fact that the preferences for the two NP variants are strongly correlated with each other

but completely uncorrelated with those for the non-homologous HA shows that the site-specific amino-acid preferences of homologs are tremendously more similar than those of unrelated proteins.

This general conservation of site-specific amino-acid preferences does not imply an absence of epistasis during NP's evolution. For instance, our results show that some (as yet mechanistically uncharacterized) epistatic interaction with other sites has driven a strong shift in the amino-acid preferences at site 470. At other sites, smaller shifts in amino-acid preferences are still certain to induce evolutionarily important epistasis, since natural selection is highly discerning. Indeed, we have previously demonstrated epistasis among mutations to NP [56], indicating that NP is no different than the many other proteins for which evolutionarily relevant epistasis has been identified [129, 96, 34, 87, 92, 101]. Our key result is not that epistasis is absent, but rather that its frequency and magnitude are sufficiently low that the amino-acid preferences for most sites are are still vastly more similar between homologs than between non-homologous proteins.

The implications of this finding are illustrated by the second part of our study, which shows that the experimentally measured site-specific amino-acid preferences can inform phylogenetic substitution models that greatly outperform non-site-specific models even for more diverged NP homologs. It is well known that the actual constraints on protein evolution involve cooperative interactions among sites [142, 39, 61], and so substitution models that treat sites either independently or identically are obviously imperfect. But computational biology must balance realism with tractability. Site-independent but site-specific substitution models are becoming feasible for real-world datasets [79, 82, 127, 113, 15, 16], but approaches that relax the assumption of independence among sites remain in their infancy [26, 19]. Are amino-acid preferences sufficiently conserved for site-independent but site-specific models to represent substantial improvements over existing non-site-specific alternatives? Both our experimental and computational results answer this question with a resounding yes.

Why are the site-specific amino-acid preferences mostly conserved? As is the case

for virtually all proteins [27, 115], the structure of NP is highly conserved among homologs [138, 36], and sites in specific structural contexts often have propensities for certain amino acids [28, 110, 85]. In addition, selection for protein stability is a major constraint on evolution [39, 13], and experiments on NP [5] and other proteins [111, 116] have shown that the effects of mutations on stability are similar among homologs. Therefore, conserved structural and stability constraints probably naturally lead to substantial conservation of site-specific amino-acid preferences. We refer the reader to an excellent recent study by Risso *et al.* [111] for a more biophysically nuanced discussion of these issues.

The extent to which site-specific amino-acid preferences will be conserved among more distant homologs remains an open question. Computational simulations of the divergence of distant homologs have been used to argue that preferences shift substantially [102], but the reliability of such simulations is unclear since computational predictions of the effects of even single amino-acid mutations are only modestly accurate [73, 108]. The only direct experimental data come from a study showing that the effects of a handful of mutations on stability are mostly conserved among homologs with about 50% protein-sequence identity [111]. More comprehensive determination of the relationship between sequence divergence and shifts in site-specific amino-acid preferences therefore remains an important topic for future work.

Chapter 3

ACCURATE MEASUREMENT OF THE EFFECTS OF ALL AMINO-ACID MUTATIONS TO INFLUENZA HEMAGGLUTININ

A version of this chapter has been previously published as:

Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses,* 8(6), 155 (2016).

Jesse Bloom and I designed the experiments. I performed the experiments and analyzed the data, with contributions from Jesse Bloom on preliminary work establishing the system to produce the HA-deficient helper virus. Jesse Bloom and I wrote the manuscript.

3.1 Abstract

Influenza genes evolve mostly via point mutations, and so knowing the effect of every amino-acid mutation provides information about evolutionary paths available to the virus. We and others have combined high-throughput mutagenesis with deep sequencing to estimate the effects of large numbers of mutations to influenza genes. However, these measurements have suffered from substantial experimental noise due to a variety of technical problems, the most prominent of which is bottlenecking during the generation of mutant viruses from plasmids. Here we describe advances that ameliorate these problems, enabling us to measure with greatly improved accuracy and reproducibility the effects of all amino-acid mutations to an H1 influenza hemagglutinin on viral replication in cell culture. The largest improvements come from using a helper virus to reduce bottlenecks when generating viruses from plasmids. Our measurements confirm at much higher resolution the results of previous studies suggesting that antigenic sites on the globular head of hemagglutinin are highly tolerant of mutations. We also show that other regions of hemagglutinin—including the stalk epitopes targeted by broadly neutralizing antibodies—have a much lower inherent capacity to tolerate point mutations. The ability to accurately measure the effects of all influenza mutations should enhance efforts to understand and predict viral evolution.

3.2 Background

Seasonal influenza is a recurrent threat to human health, largely because it rapidly accumulates amino-acid mutations in proteins targeted by the immune system [118]. Measuring the functional impact of every possible amino-acid mutation to influenza can therefore provide useful information about which evolutionary paths are accessible to the virus. Such measurements are now possible using deep mutational scanning [49, 21]. When applied to influenza, this technique involves creating all codon mutants of a viral gene, incorporating these mutant genes into viruses that are subjected to a functional selec-

tion, and estimating the functional impact of each mutation by using deep sequencing to quantify its frequency pre- and post-selection. We and others have used deep mutational scanning to estimate the effects of all amino-acid [123, 15, 40] or nucleotide [135, 133] mutations to several influenza genes, and Heaton and coworkers [63] have used a similar approach to examine influenza's tolerance to short insertions. However, these studies suffered from substantial noise that degrades the utility of their results. For instance, in every study that reported the results for independent experimental replicates, the replicate-to-replicate correlation was mediocre.

This experimental noise arises primarily from bottlenecking of mutant diversity during the generation of viruses from plasmids. The influenza genome consists of eight negative-sense RNA segments. During viral infection, gene expression from these segments is a highly regulated process [29, 112, 117]. Generating influenza from plasmids involves co-transfecting mammalian cells with multiple plasmids that must yield all eight viral gene segments and at least four viral proteins at a stoichiometry that leads to assembly of infectious virions [69, 94, 47]. This plasmid-driven process is understandably less efficient than viral infection. A small fraction of transfected cells probably yield most initial viruses, which are then amplified by secondary infection. This bottlenecking severely hampers experiments that require creating a diverse library of viruses from an initial library of plasmids.

Several strategies have been used to overcome problems associated with bottlenecks during the generation of influenza from plasmids. One strategy is to generate and titer each viral variant individually, and then mix them [124, 10]. A second strategy is to reduce the impact of bottlenecks by shrinking the complexity of the libraries, such as by only mutating a small portion of a viral gene [134, 71]. Neither of these strategies scale effectively to the deep mutational scanning of full-length proteins, since there are $\sim 10^4$ unique amino-acid mutants of a 500-residue protein.

To overcome these limitations, we have developed a novel approach that uses a "helper virus" to generate virus libraries without strong bottlenecking. We have combined

this approach with other technical improvements to perform deep mutational scanning of all amino-acid mutations to an H1 hemagglutinin (HA) with much higher accuracy and reproducibility than existing deep mutational scans of influenza genes. We use phylogenetic analyses to show that our measurements accurately reflect constraints on HA evolution in nature. We confirm that antigenic sites in the globular head of HA are highly tolerant of mutations, and identify other regions of the protein that are more constrained. These advances improve our understanding of HA's inherent evolutionary capacity and can help inform evolutionary modeling and guide the development of vaccines targeting sites with a limited capacity for mutational escape.

3.3 Results

3.3.1 A helper-virus enables efficient production of mutant virus libraries from plasmids

We reasoned that the process of generating viral libraries carrying HA mutants would be more efficient if transfected cells only needed to produce HA from plasmid, and the other gene segments and proteins were delivered by viral infection (Figure 3.1A). The Palese lab has previously shown that a seven-segmented HA-deficient virus can be propagated in cells that constitutively express HA protein [89]. We created HA-expressing cells and validated that we could propagate an HA-deficient A/WSN/1933 (H1N1) virus (Figure B.1).

We cloned triplicate plasmid libraries of random codon mutants of the A/WSN/1933 HA gene. These libraries contain multi-nucleotide (e.g., GGC$\rightarrow$CAT) as well as single-nucleotide (e.g., GGC$\rightarrow$GAC) codon mutations. There are $63 \times 565 \approx 3.5 \times 10^4$ different codon mutations that can be made to the 565-codon HA gene, corresponding to $19 \times 565 \approx 10^4$ amino-acid mutations. The deep sequencing described below found at least three occurrences of over 97% of these amino-acid mutations in each of the three replicate plasmid mutant libraries. These libraries have a somewhat lower mutation rate than our previous deep mutational scan of hemagglutinin [123], with the number of mutations per clone following a roughly Poisson distribution with a mean of about one (Figure B.2).

We cloned these HA libraries into both uni-directional and bi-directional reverse-genetics plasmids [94, 69].

We then transfected cells with one of the HA plasmid mutant libraries along with plasmids expressing the four viral polymerase-related proteins (PB2, PB1, PA, and NP) with the goal of generating pre-formed viral ribonucleoprotein complexes carrying the HA segment. These transfected cells were then infected with the HA-deficient helper virus, and 24 h later, we determined the titer of fully competent virus in the supernatant. The highest titers ($\sim$10^3 TCID$_{50}$ per μL) were obtained using the uni-directional reverse-genetics plasmid (Figure B.3). The reason that we co-transfected protein expression plasmids for the four polymerase-related proteins was to create pre-formed viral ribonucleoprotein complexes. Virus titers were $\sim$100-fold lower when the polymerase plasmids were not co-transfected (data not shown). Overall, these findings demonstrate the feasibility of the helper-virus strategy in Figure 3.1A.

We next used this helper-virus strategy to independently generate three mutant virus libraries, one from each of our triplicate plasmid mutant libraries. Each mutant virus library should sample most of the codon mutations to the A/WSN/1933 HA. We also generated a control virus library from a plasmid encoding the unmutated wild-type HA gene.

3.3.2 Low MOI passage combined with barcoded-subamplicon sequencing reveals strong selection against non-functional HA variants

To select for viruses carrying functional HA variants, we passaged the mutant virus libraries at a low multiplicity of infection (MOI) of 0.0075 TCID$_{50}$ per cell as outlined in Figure 3.1A. This MOI is substantially lower than the MOI of 0.1 that we used in our previous study to examine the effects of all mutations to HA [123], and was chosen with the goal of more effectively purging non-functional HA variants.

To quantify selection on HA, we needed our deep sequencing to be sufficiently accurate to determine the frequency of each mutation pre- and post-selection. Standard

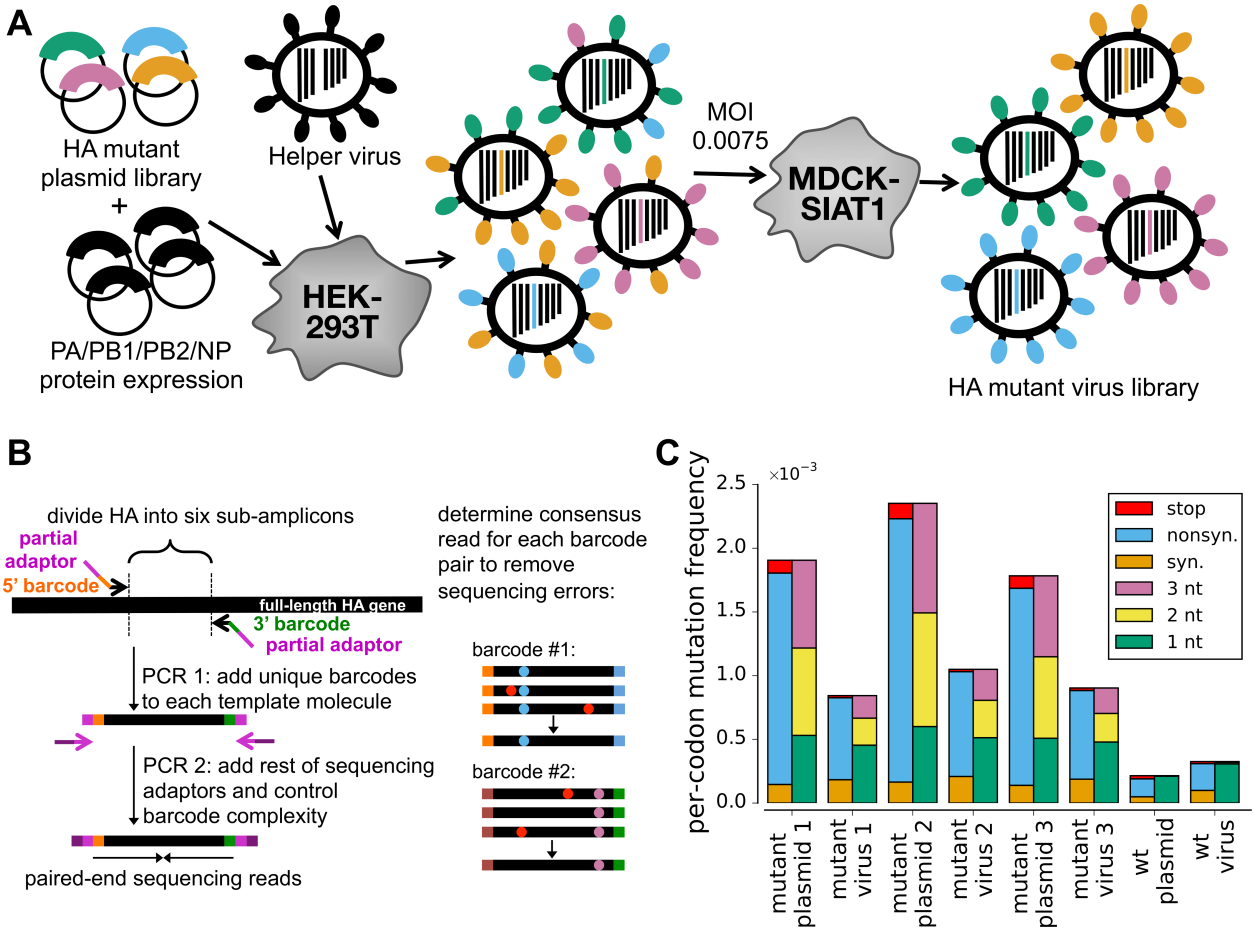

Figure 3.1: **Deep mutational scanning of HA**. (**A**) Cells transfected with a plasmid mutant library of HA are infected with an HA-deficient helper virus to yield a library of mutant viruses. This virus library is passaged at low MOI to select for functional variants and enforce genotype-phenotype linkage. The helper viruses themselves are propagated in cells constitutively expressing HA (Figure B.1). The variants in the plasmid mutant library contain an average of one codon mutation, with the number of mutations per clone following a roughly Poisson distribution (Figure B.2). The helper-virus works best when HA is provided on a plasmid that directs the synthesis of only viral RNA (Figure B.3). (**B**) Accurate Illumina sequencing using barcoded subamplicons. HA is divided into six sub-amplicons, and a first round of PCR appends random barcodes and part of the Illumina adaptor to each subamplicon. The complexity of these barcoded subamplicons is controlled to be less than the sequencing depth, and a second round of PCR adds the remaining adaptor. Sequencing reads are grouped by barcode, distinguishing sequencing errors that occur in only one read (red dots) from true mutations that occur in all reads (blue and purple dots). (**C**) The overall mutation frequencies reveal selection against stop codons and many nonsynonymous mutations in the mutant viruses relative to the plasmids from which they were generated (see also Figure B.4). Sequencing of unmutated plasmid and virus generated from this plasmid (denoted as "wt plasmid" and "wt virus" in panel C) indicates rates of sequencing, reverse-transcription, and viral replication errors are lower than the mutation rates in the libraries, enabling us to reliably distinguish the signal and noise.

Illumina sequencing has an error rate that is too high. In our previous deep mutational scanning of influenza [123, 15, 40], we reduced this error rate by using overlapping paired-end reads. Here, we used an alternative error-correction strategy that involves attaching random barcodes to PCR subamplicons and then clustering reads with the same barcode (Figure 3.1B). To our knowledge, this basic strategy was first described by Hiatt *et al.* [68] and first applied to influenza by Wu *et al.* [135]. Sequencing of the unmutated plasmid allows us to estimate that the error rate is $\sim 2 \times 10^{-4}$ per codon, corresponding to $< 10^{-4}$ per nucleotide (Figure 3.1C, sample referred to as "wt plasmid"). This error rate is substantially lower than we obtained previously using overlapping paired-end reads, consistent with the results of the sequencing-strategy comparison by Zhang *et al.* [141]. Sequencing of viruses generated from the unmutated plasmid shows that the error rates associated with reverse-transcription and viral replication are also tolerably low (below the mutation rate in the mutant libraries) (Figure 3.1C, sample referred to as "wt virus").

Figure 3.1C reveals strong selection against non-functional HA variants. The plasmid mutant libraries contain a mix of synonymous, nonsynonymous, and stop-codon mutations. However, stop-codon mutations are almost completely purged from the passaged mutant virus libraries, as are many nonsynonymous mutations. The selection against the stop codons is stronger than in our previous deep mutational scan [123] (Figure B.4). Overall, these results indicate strong selection on HA that can be quantified by accurate deep sequencing.

3.3.3 The mutant virus libraries have reduced bottlenecking and yield reproducible measurements of mutational effects

To evaluate whether the virus libraries were bottlenecked, we examined the distribution of synonymous mutation frequencies in each library. If bottlenecking causes a few mutants to stochastically dominate, we expect that in each library a few sites will have relatively high synonymous mutation frequencies and that these sites will differ among replicates.

Figure 3.2 shows normalized synonymous mutation frequencies across HA for each of the three replicate mutant virus libraries from both our previous deep mutational scan of HA that utilized reverse genetics [123], and the current study utilizing helper viruses. In the older study, each replicate had a different handful of sites with greatly elevated synonymous frequencies (green arrows), indicative of stochastic bottlenecking. In contrast, in our new virus libraries, the distribution of synonymous mutation frequencies is much more uniform across the HA gene. Specifically, the standard deviation of normalized synonymous frequencies was 1.63 $\pm$ 0.14 for the old libraries, but only 1.18 $\pm$ 0.05 for the new libraries, indicating less bottlenecking-induced variation in mutation frequencies in the new libraries.

We next evaluated the reproducibility of our measurements of the effects of each mutation. We estimated the effect of each mutation from its change in frequency in the mutant viruses relative to the original plasmid libraries, correcting for the site-specific error rates determined by sequencing unmutated virus and plasmid, and performing the analyses using the algorithms described in [17] and implemented in the `dms_tools` software (version 1.1.12, `http://jbloomlab.github.io/dms_tools/`). The results are quantified in terms of the *preference* of each site for each amino-acid; the set of all 20 preferences at a site can be thought of as representing the expected post-selection frequency of each amino acid at that site if all amino acids are initially present at equal frequencies.

Figure 3.3 shows the correlation between the amino-acid preferences from each experimental replicate. The replicate-to-replicate reproducibility is dramatically improved in our new experiments relative to our previous work utilizing reverse genetics [123], with the average Pearson's R^2 increasing from 0.34 to 0.61. The new experiments are also largely free of the most problematic type of noise that plagued the previous study, where an amino acid at a site is deemed highly preferred in one replicate but disfavored in another. Overall, these results demonstrate that our new strategies enable more reproducible measurement of the effects of all mutations to HA.

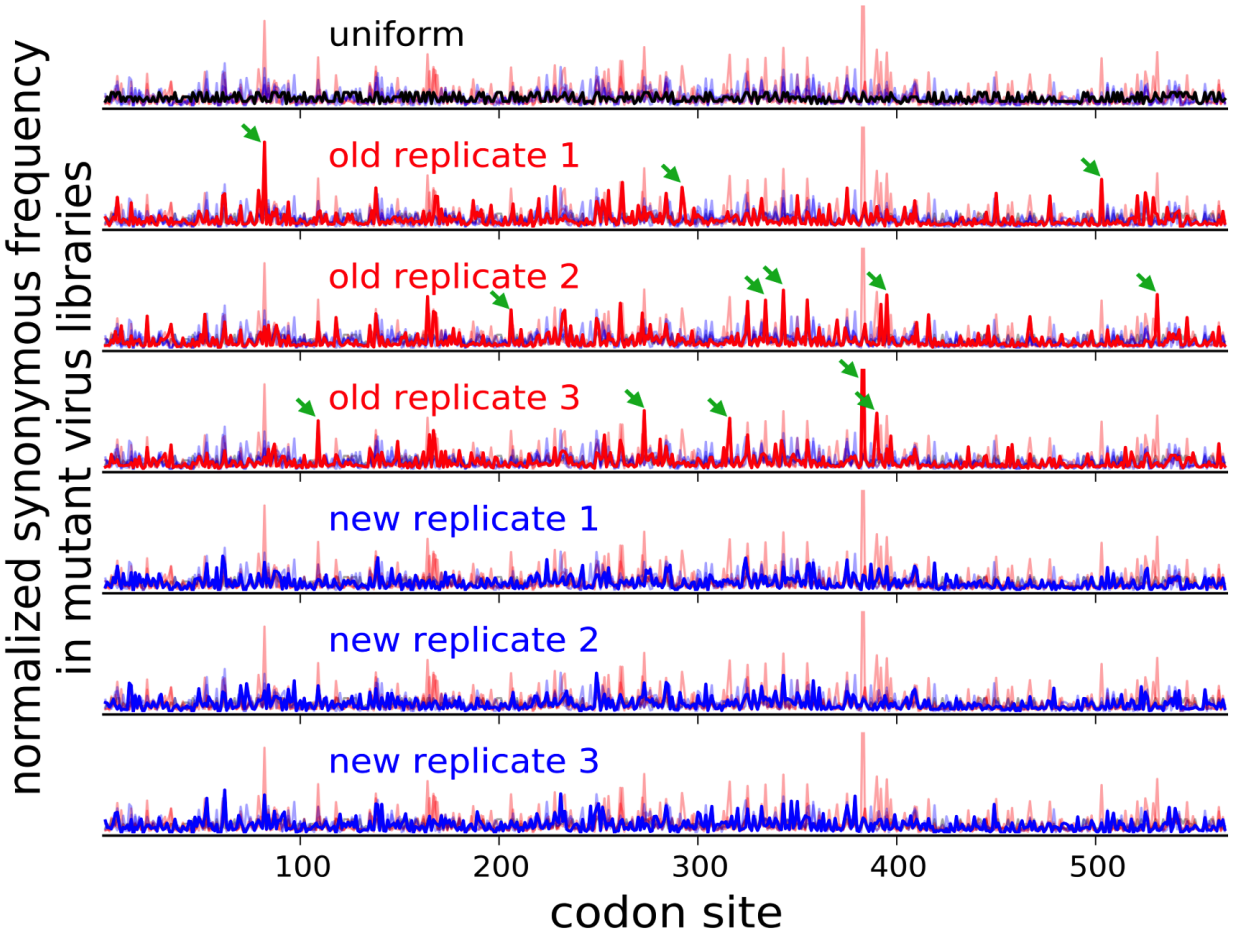

Figure 3.2: **The use of helper viruses reduces bottlenecking during the generation of the mutant virus libraries.** Each row shows the synonymous mutation frequency for every site normalized to the total synonymous frequency for that sample. If synonymous mutations are sampled uniformly, the data should resemble the black line in the top row (the line is not completely straight because different codons have different numbers of synonymous variants). The next six rows show the synonymous mutation frequencies for each replicate of the old (**red lines**) [123] and new (**blue lines**) experiments. To assist in comparing the locations and heights of peaks across all samples, the data for each replicate are shown as a thick line in front of thin lines representing the other five replicates. The old experiments have more bottlenecking as manifested by taller peaks indicating synonymous mutations that were stochastically enriched in each replicate (examples marked by **green arrows**). The differences between replicates are *not* due to differences in synonymous mutation frequencies in the plasmid libraries used to generate the viruses (Figure B.5).

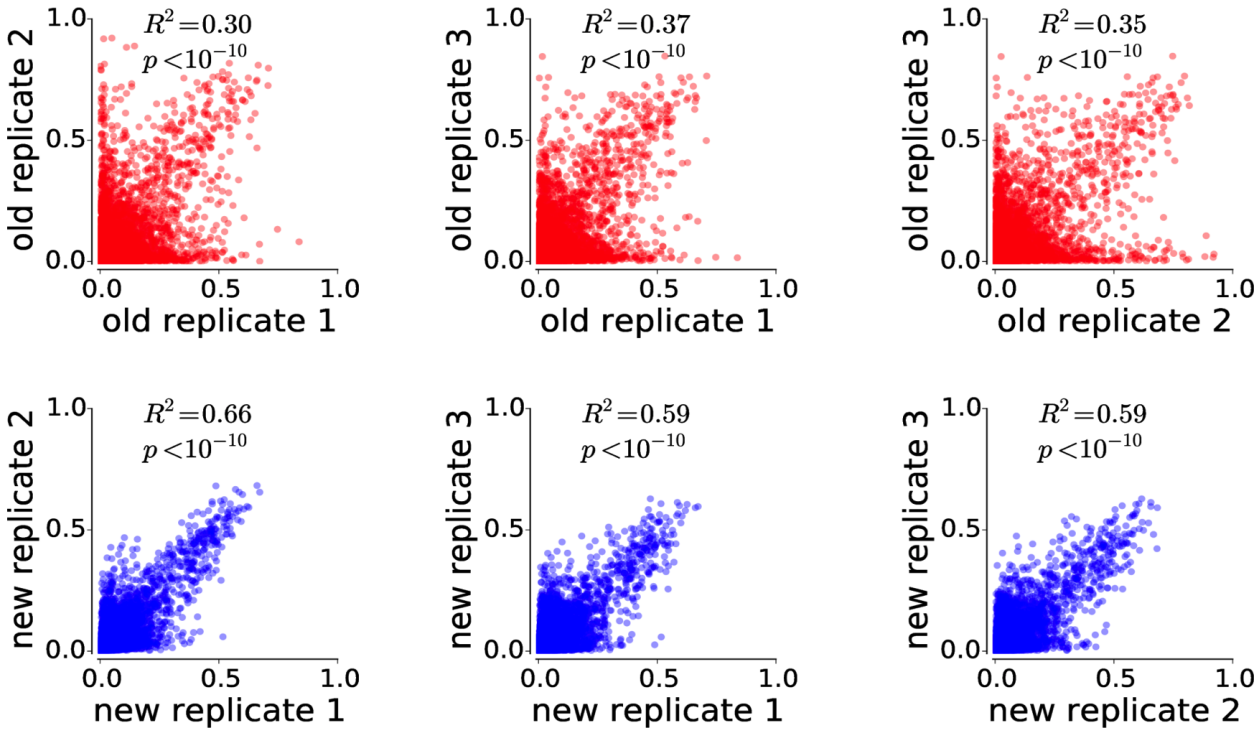

Figure 3.3: **The use of helper viruses increases reproducibility of measurements of mutational effects.** The mutational effects measured in the new experiments are much more reproducible across replicates. Each plot shows the squared Pearson correlation coefficient for all site-specific amino-acid preferences measured in a pair of independent experimental replicates. Each point represents the amino-acid preference for a specific amino acid at a specific site, as measured in the indicated replicate experiment.

3.3.4 The measurements better reflect the constraints on HA evolution in nature

We next tested whether our new measurements better describe the evolution of HA in nature. The accuracy with which experimental measurements of site-specific amino-acid preferences reflect the constraints shaping a protein's evolution in nature can be quantified by comparing the phylogenetic fit of experimentally informed substitution models [15]. We assembled a set of human and swine influenza HA sequences and fit substitution models using `phydms` [14] (version 1.1.1, `http://jbloomlab.github.io/phydms/`), which in turn uses `Bio++` [58] for the likelihood calculations.

A substitution model informed by our new measurements described the natural evo-

lution of HA better than a model informed by our older measurements from [123], and vastly better than conventional non-site-specific substitution models (Table 3.1). Averaging the measurements from both studies improved phylogenetic fit even further, a finding consistent with previous work reporting that combining data from multiple deep mutational scanning studies of the same protein tends to improve substitution models [40].

The phylogenetic model fitting optimizes a parameter that accounts for differences in the stringency of selection between the experiments and natural evolution [16]; a stringency parameter >1 indicates that natural selection prefers the same amino acids as the experimental selections but with greater strength. The best model in Table 3.1 has a stringency parameter of 1.8. The site-specific amino-acid preferences for this model scaled by this stringency parameter are displayed in Figure 3.4. Residues are numbered sequentially beginning with the initiating methionine; conversions to other numbering schemes are at `https://github.com/mbdoud/mutational_antigenic_profiling/blob/master/HA_numbering.txt`.

3.3.5 A handful of sites are under very different selection in our experiments than in nature

We next asked whether there are sites in HA that evolve in nature in a way that is highly discordant with our experimental measurements. To do this, we again used `phydms` [14] to identify selection in nature for amino acids that differ from the ones preferred in the deep mutational scanning, again using natural sequences from seasonal human H1N1 and classical swine H1N1 HAs. Briefly, this program uses a maximum-likelihood phylogenetics approach to estimate the difference in preference for each amino acid at each site between the experimental measurements and selection in nature (see [14] for details). Figure 3.5 shows the difference in amino-acid preferences between our experiments and natural evolution for each site in HA. At most sites, the magnitude of differential selection is small, indicating that the experimentally measured preferences mostly parallel constraints

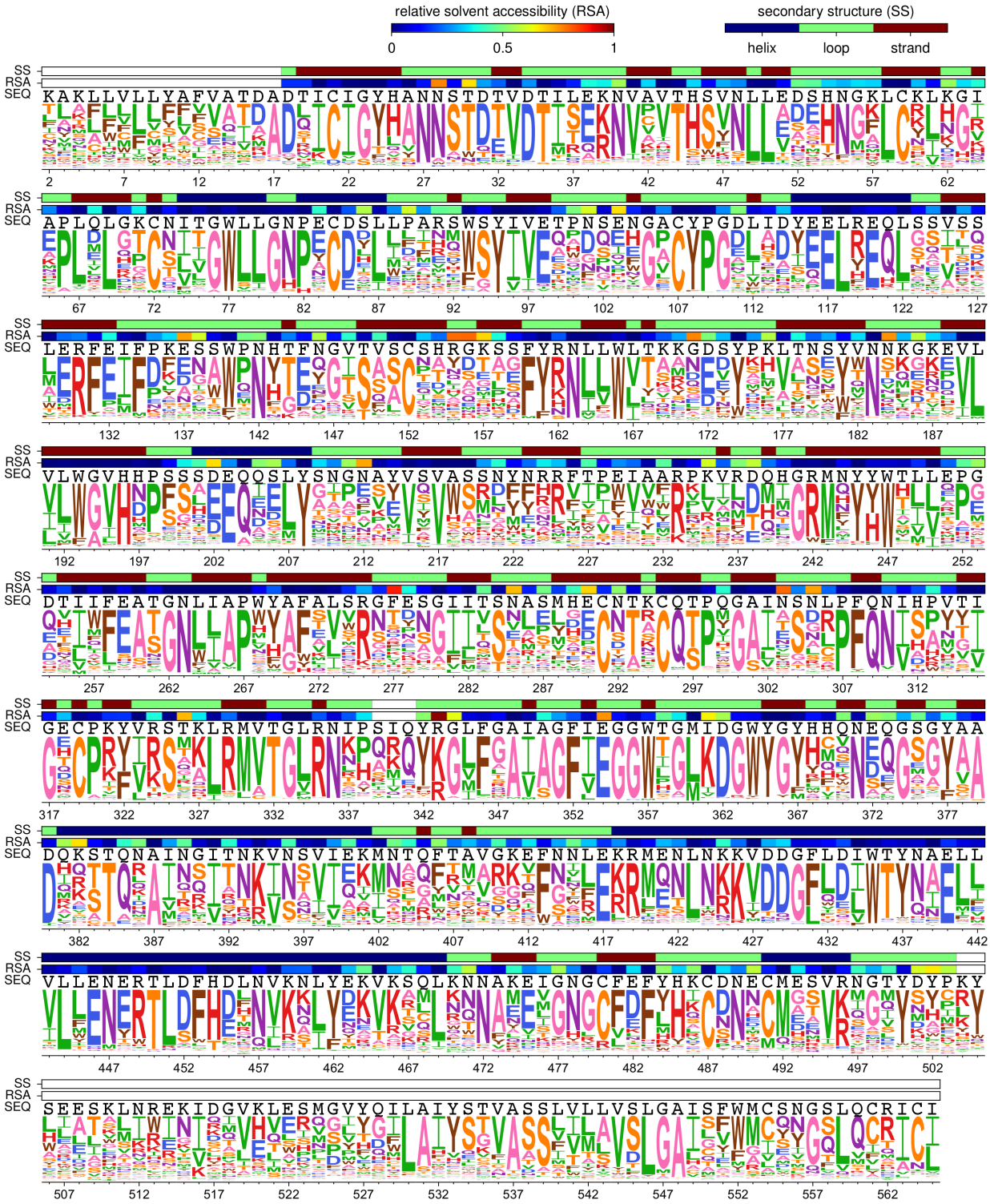

Figure 3.4: **HA's site-specific amino-acid preferences.** The preference of each site in HA for each of the 20 amino-acids as inferred by combining the new and old data and re-scaling by the stringency parameter inferred in Table 3.1. The height of each letter is proportional to the preference for that amino acid at that site. The overlay bars show each residue's secondary structure, relative solvent accessibility, and wildtype identity in the A/WSN/1933 HA.

Table 3.1: **The site-specific amino-acid preferences measured in the new experiments offer an improved description of HA evolution in nature.** Aikake information criterion (AIC) [107] was used compare the maximum likelihood phylogenetic fit of several models to an alignment of seasonal human H1N1 and classical swine H1N1 HAs. The experimentally informed substitution models are of the form described in [14] with the data from the average of all three replicates of the new or old experiments, or the average of the two. These models are compared to the variants of the substitution model of Goldman *et al.* [55] denoted as M0 and M8 in Yang *et al.* [137] with the equilibrium codon frequencies estimated empirically using the F3X4 method. The best model is the one that combines all experimental data, but a model informed by the new experiments alone is better than one informed by the old experiments alone. To confirm that the experimentally informed models are superior because they are site specific, we fit a control model in which the experimental data is averaged across sites. The tree topology was fixed to that inferred by maximum likelihood using the M0 version of the Goldman–Yang model. The free parameters for each model were then optimized along with the branch lengths; optimized parameters are in the last column.

model	ΔAIC	log likelihood	parameters (optimized + empirical): optimized values
new data + old data	0.0	-14933.5	6 (6 + 0): β = 1.82, ω = 0.51, κ = 4.95, ϕ_A = 0.40, ϕ_C = 0.18, ϕ_G = 0.20
new data	197.6	-15032.3	6 (6 + 0): β = 1.80, ω = 0.46, κ = 5.06, ϕ_A = 0.40, ϕ_C = 0.18, ϕ_G = 0.20
old data	341.2	-15104.1	6 (6 + 0): β = 1.40, ω = 0.46, κ = 4.90, ϕ_A = 0.39, ϕ_C = 0.18, ϕ_G = 0.20
Goldman–Yang M8	2156.8	-16003.9	14 (5 + 9): $p_{\omega>1}$ = 0.01, $\omega_{>1}$ = 1.91, p_β = 0.02, q_β = 0.76, κ = 4.94
new data + old data, averaged across sites	2971.6	-16419.3	6 (6 + 0): β = 0.50, ω = 0.20, κ = 5.38, ϕ_A = 0.38, ϕ_C = 0.18, ϕ_G = 0.21
Goldman–Yang M0	2980.8	-16418.9	11 (2 + 9): ω = 0.19, κ = 4.88

on natural evolution. Sites that are under strong differential selection usually show conservative changes; for example, site 78 prefers isoleucine in nature but leucine in our deep mutational scanning.

One of the most striking exceptions to this general concordance between natural selection and our experiments can be given a clear explanation. At site 342, the experimentally measured preference for tyrosine is at odds with nature's strong preference for serine (Figure 3.5). The lab-adapted A/WSN/1933 strain used in our experiments differs from naturally occurring influenza in that it uses plasmin to cleave and activate HA [81, 57]. Plasmin cleavage is enhanced by tyrosine at this site [122], so it is unsurprising that our experiments detected a preference at this site unique to the influenza strain we used. This

example illustrates how the occasional deviations from the general concordance between deep mutational scanning experiments and natural selection can point to interesting biological mechanisms.

3.3.6 Antigenic sites in HA's globular head are highly tolerant of mutations, but stalk epitopes targeted by broadly neutralizing antibodies are not

We computed the inherent mutational tolerance of each site using the stringency-scaled amino-acid preferences from the combined datasets (Figure 3.6A). The mutational tolerance is mapped onto the structure of HA in Figure 3.6B.

The H1 HA antigenic sites defined by Caton *et al.* [23] are significantly more mutationally tolerant than the average site (Figure 3.6C), even after accounting for relative solvent accessibility (Figure B.6A). This high mutational tolerance extends to other solvent-exposed residues in contact with the antigenic sites (Figure 3.6D, Figure B.6B), indicating that the HA molecular surfaces commonly targeted by antibodies have a high inherent capacity for evolutionary change. This high mutational tolerance does not extend to the receptor-binding pocket (Figure 3.6E, Figure B.6C,D) but may be a feature of the sites that make the greatest contributions to the punctuated antigenic evolution of H3N2 and seasonal H1N1 HA [75] (Figure 3.6F), albeit not at a level that is statistically significant after correcting for solvent accessibility (Figure B.6E). These results support the findings of our previous study [123] that the sites in HA that are the immunodominant targets of antibodies have a high inherent capacity to tolerate mutations.

Perhaps in part because of the high mutational tolerance of the antigenic sites in its globular head, HA is adept at escaping antibody-mediated immunity [118, 8]. New vaccines are being developed that aim to elicit immunity against other portions of HA [77], most commonly regions in the stalk that are relatively conserved among naturally occurring strains. An important question is whether these stalk regions are conserved because they are inherently intolerant of point mutations, or simply because they are not currently

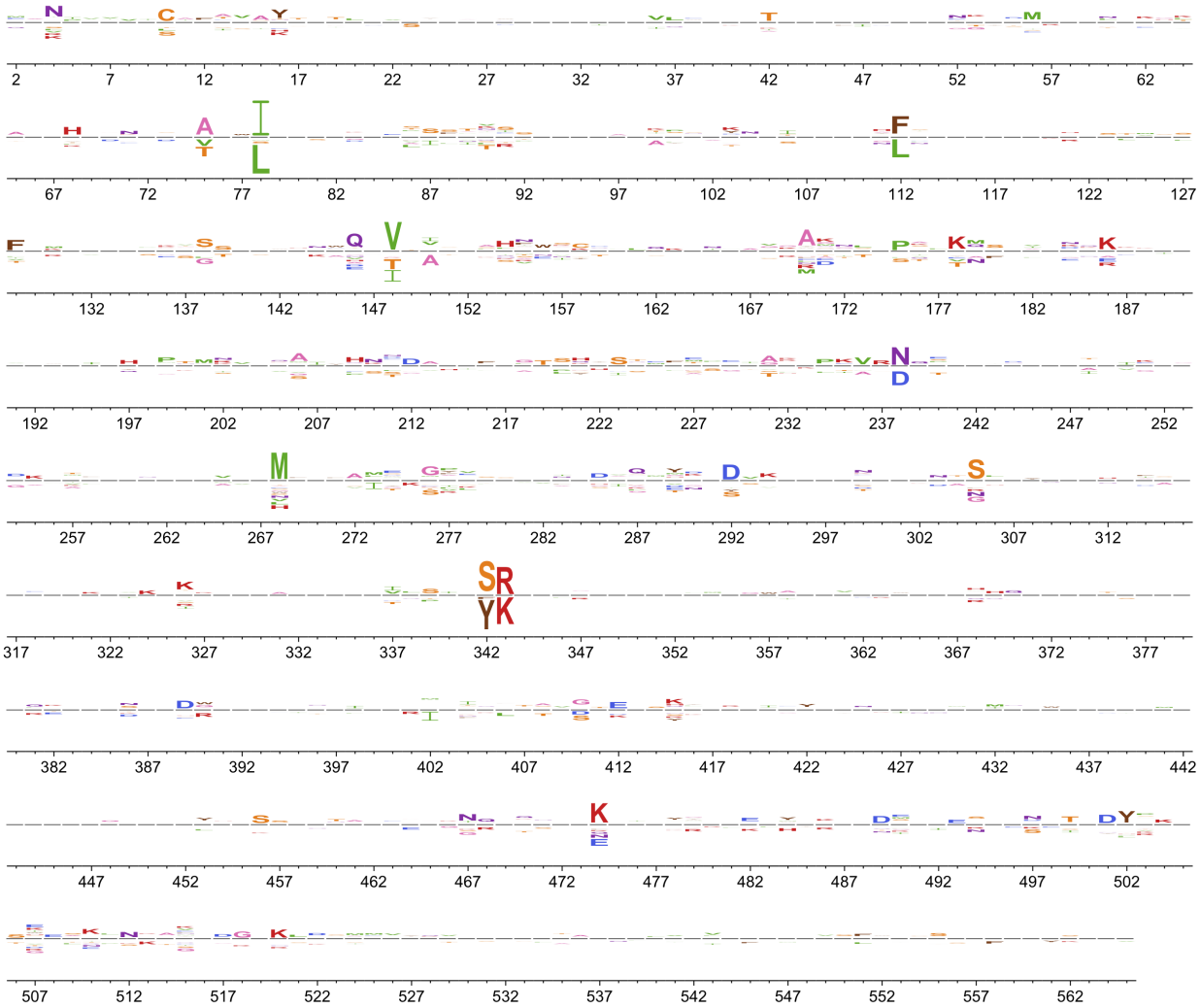

Figure 3.5: **Differential selection between our selection in the lab and HA's evolution in nature.** The approach in [14] was used to determine the magnitude of differential selection between the deep mutational scanning and the natural evolution of human seasonal H1N1 and classical swine H1N1 HAs. At each site, the height of a letter above or below the center line indicates that differential selection for or against that amino acid in nature as compared to our experiments. At most sites, the differential selection is very small, showing that the experimental measurements are mostly concordant with natural selection on HA. However, a few sites are under very different selection in nature as compared to our experiments.

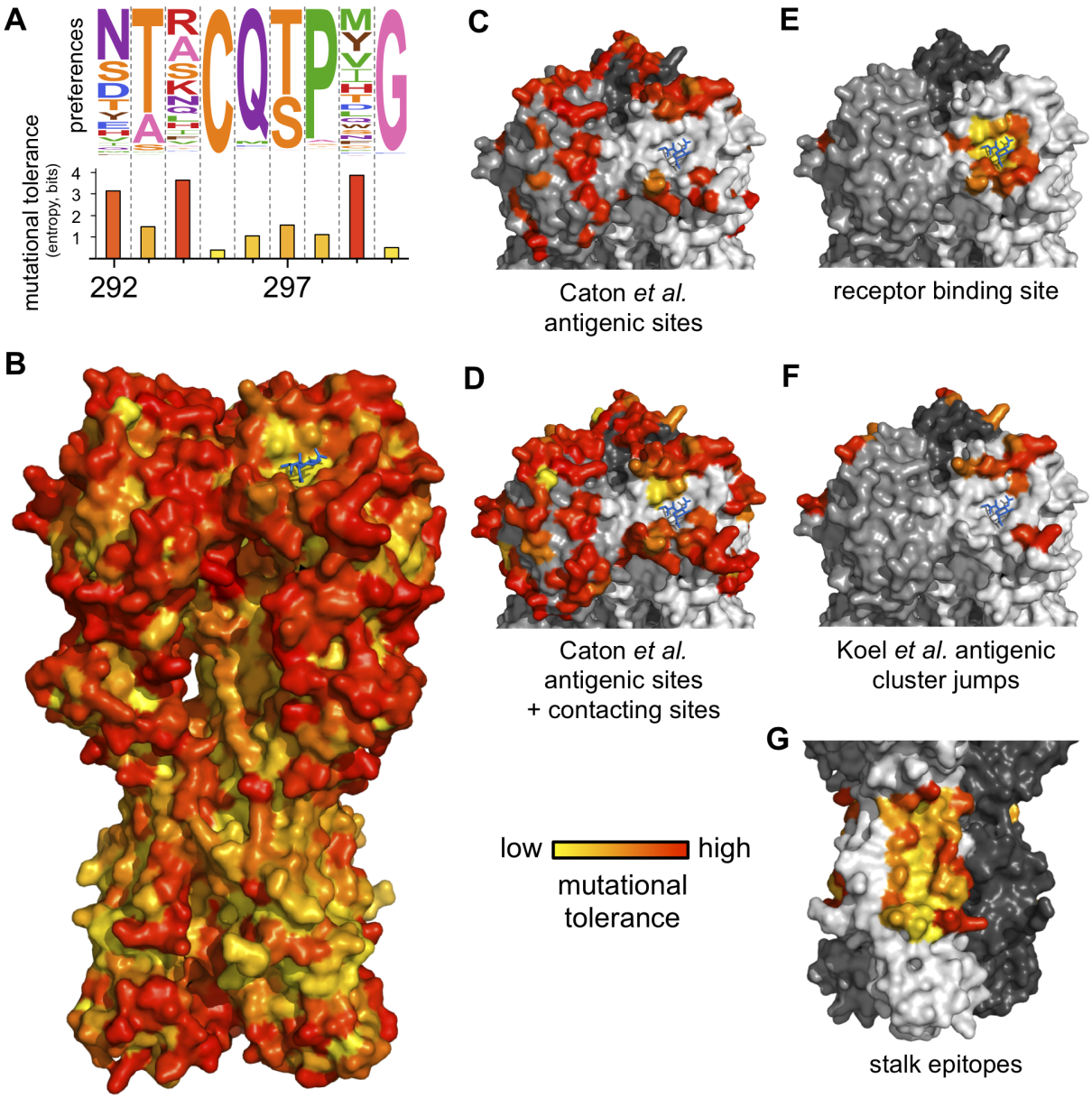

Figure 3.6: **Antigenic sites in HA's globular head have a high inherent tolerance for mutations, but HA's stalk is relatively intolerant of mutations.** (**A**) Mutational tolerance is calculated as the Shannon entropy of a site's amino-acid preferences. (**B**) Mutational tolerance mapped onto the HA trimer (**yellow** indicates low tolerance, **red** indicates high tolerance, **blue** sticks show the sialic-acid receptor). (**C**,**D**) The antigenic sites defined by Caton *et al.* [23] have high mutational tolerance, as do the residues contacting these sites. (**E**) Conserved receptor-binding residues have low mutational tolerance. (**F**) Sites that contribute to antigenic cluster jumps [75]. (**G**) Sites in the footprints of four broadly neutralizing antibodies have low mutational tolerance. Shown are footprints of F10, CR6261, FI6v3, and CR9114 [121, 44, 31, 42]. For panels B-G, tolerance is mapped onto PDB structure 1RVX [51]. For panels C-G, each monomer is shown in a different shade of gray. Figure B.6 reports statistical analyses of whether subsets of sites have higher or lower tolerance than expected given their solvent accessibility.

under immune pressure. To answer this question, we examined the inherent mutational tolerance of the largely overlapping epitopes of four broadly neutralizing anti-stalk antibodies: F10 [121], CR6261 [44], FI6v3 [31], and CR9114 [42]. Visual inspection of Figure 3.6G shows that these stalk epitopes have a low mutational tolerance, a result that is confirmed by statistical analysis (Figure B.6F). Therefore, the epitopes that next-generation vaccines aim to target indeed have a reduced capacity for immune escape by point mutations. This finding is also consistent with Heaton *et al.*'s report that HA's stalk is intolerant to insertions [63].

We wondered if some of HA's variation in mutational tolerance is explained by differences in the three ancient domains that compose the protein. HA is the product of a series of ancient insertions that merged a fusion domain, a receptor-binding domain (which contains the majority of the antigenic sites as well as the receptor-binding pocket itself), and a vestigial esterase domain [114]. We compared the inherent mutational tolerance of these three domains, again correcting for solvent accessibility. We found that sites in the receptor-binding domain on average have a significantly higher mutational tolerance than all sites in the protein, although sites in the receptor-binding pocket itself are often highly constrained (Figure 3.6, Figure 3.7). On the other hand, sites in the fusion domain have a significantly lower mutational tolerance than all sites (Figure 3.7). This enriched tolerance to point mutations throughout the receptor-binding domain is also concordant with the results of Heaton *et al.*, showing that the receptor-binding domain is uniquely tolerant to short insertions [63]. Therefore, HA's antigenic evolvability is not just a consequence of the immunodominant antigenic sites themselves having high mutational tolerance, but also because these sites are found within a protein domain that is intrinsically more mutable than the rest of HA.

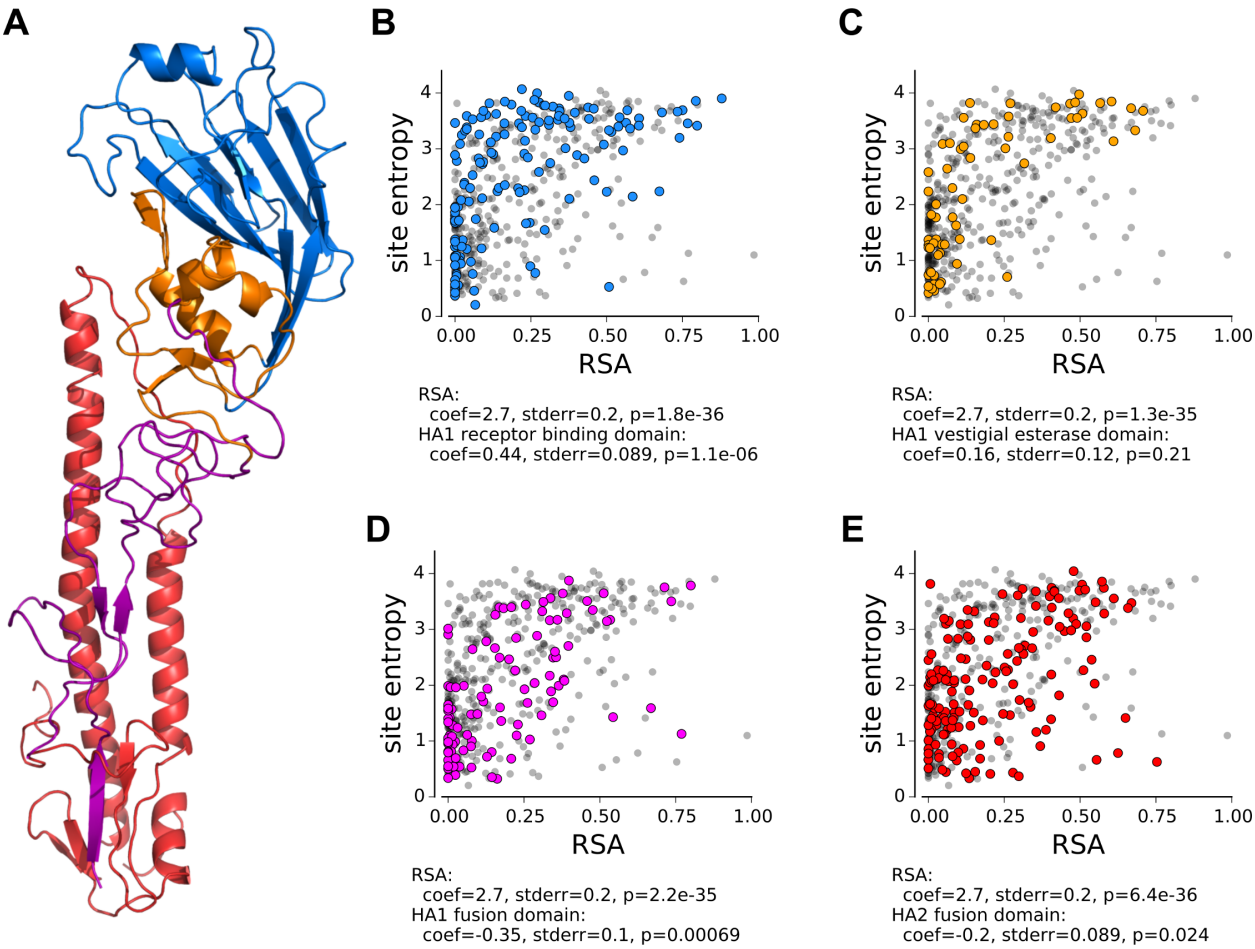

Figure 3.7: **The mutational tolerance of HA's three ancient protein domains. (A)** The domain architecture of HA. The receptor-binding domain is blue, the vestigial esterase domain is orange, and the fusion subdomains of HA1 and HA2 are purple and red, respectively. **(B)** The mutational tolerance of the receptor-binding domain is significantly higher than the rest of the protein. **(C)** The mutational tolerance of the vestigial esterase domain is not significantly different than the rest of the protein. **(D), (E)** The mutational tolerance of the fusion subdomains of HA1 and HA2 is significantly lower than the rest of the protein. Significance is assessed using multiple linear regression, correcting for solvent accessibility as in Figure B.6.

3.4 Discussion

We have described new techniques that greatly improve the reproducibility of deep mutational scanning of influenza. The largest improvement appears to result from using a helper virus to generate virus mutant libraries without the bottlenecks that plague the

creation of viruses purely from plasmids. We have used these techniques to more accurately measure the effects of all amino-acid mutations to HA. Our measurements confirm at greater precision and resolution the finding [123, 63] that HA's propensity for immune escape is underpinned by the high inherent mutational tolerance of the immunodominant receptor-binding domain. Our data also show that some regions of HA—including the stalk epitopes targeted by new broadly neutralizing antibodies— have a reduced capacity for evolutionary change.

In this study, we measured the effects of all mutations to the HA from a lab-adapted H1N1 strain. To what extent can these measurements be extrapolated to other HAs? Due to epistasis, the effects of mutations sometimes change as proteins evolve [56, 62]. However, many aspects of mutational effects are often roughly conserved during evolutionary divergence: for instance, experiments have shown that the effects of mutations on stability are often quite similar among homologs, both for HA [22] and proteins more generally [5, 111]. In a previous study, we used deep mutational scanning to estimate the effects of all mutations to two close homologs of influenza nucleoprotein, and found that only a few sites exhibited large qualitative changes in their amino-acid preferences [40]. Therefore, the limited existing experimental work on this topic suggests that site-specific amino-acid preferences will often be broadly similar among homologs of the same protein, but that there will also be some shifts that can have important implications for evolution. However, further systematic investigation of this question is needed to assess the extent that deep mutational scanning studies like the one reported here can be extrapolated across protein homologs.

Overall, our work demonstrates a method for making accurate large-scale measurements of the effects of mutations to influenza proteins. Our results offer insight into how protein-intrinsic mutational tolerance shapes influenza evolution, and provide a basis for using deep mutational scanning to improve quantitative models of viral evolution and understand virus-immune interactions.

Chapter 4

COMPLETE MAPPING OF VIRAL ESCAPE FROM NEUTRALIZING ANTIBODIES

A version of this chapter has been previously published as:

Michael B Doud, Scott Hensley, and Jesse D Bloom. Complete mapping of viral escape from neutralizing antibodies. *PLoS Pathogens*, 13(3): e1006271 (2017).

Jesse Bloom and I conceived of the idea. I designed and performed the experiments and analyzed the data. Scott Hensley provided crucial reagents and advice. Jesse Bloom and I wrote the paper.

4.1 Abstract

Identifying viral mutations that confer escape from antibodies is crucial for understanding the interplay between immunity and viral evolution. We describe a high-throughput approach to quantify the selection that monoclonal antibodies exert on all single amino-acid mutations to a viral protein. This approach, mutational antigenic profiling, involves creating all replication-competent protein variants of a virus, selecting with antibody, and using deep sequencing to identify enriched mutations. We use mutational antigenic profiling to comprehensively identify mutations that enable influenza virus to escape four monoclonal antibodies targeting hemagglutinin, and validate key findings with neutralization assays. We find remarkable mutation-level idiosyncrasy in antibody escape: for instance, at a single residue targeted by two antibodies, some mutations escape both antibodies while other mutations escape only one or the other. Because mutational antigenic profiling rapidly maps all mutations selected by an antibody, it is useful for elucidating immune specificities and interpreting the antigenic consequences of viral genetic variation.

4.2 Background

Host immunity drives the evolution of many viruses. For instance, potent immunity against influenza virus is provided by antibodies against hemagglutinin (HA), the virus's most abundant surface protein [130]. Unfortunately, these antibodies also select amino-acid substitutions in the HA of human seasonal influenza A virus at a rate of over two per year [118, 9]. This rapid evolution degrades the effectiveness of anti-influenza immunity, and is a major reason why humans are repeatedly re-infected over their lifetimes. Extensive antigenic variation is also a hallmark of several other medically relevant viruses, most prominently HIV. Efforts to induce immunity to such viruses must therefore account for antigenic variation, either by targeting vaccines against current circulating viral strains [86, 93] or developing methods to administer [6, 83] or elicit [64, 78] antibodies that recognize a broad range of strains. An important component of these efforts is identifying

which viral mutations escape neutralization by specific antibodies.

The classic approach for identifying such mutations is to select individual viral mutants that are resistant to neutralization by antibodies. For instance, escape-mutant selections with a panel of monoclonal antibodies were used to broadly define major antigenic regions of influenza HA [128, 53, 23]. However, each such selection typically only identifies one of potentially many mutations that escape an antibody, with a strong bias towards whichever mutations happen to be prevalent in the initial viral stock. Therefore, escape-mutant selections provide an incomplete picture of the ways that a virus can escape an antibody.

Another approach is to individually test antibody binding or neutralization for each member of a panel of viral variants. However, there are $\sim 10^4$ single amino-acid mutants to a 500-residue viral protein, so individually creating and testing all of them is a daunting task. Therefore, even the most ambitious such studies limit themselves to a small fraction of the possible point mutations, such as by only testing mutations to alanine [97, 126, 100]. But as the current work will underscore, the antigenic effect of mutating a residue to one amino acid can be poorly predictive of the effects of mutating the same residue to another amino acid. Furthermore, the difficulty in individually generating and testing large numbers of viral variants means that such studies often use simpler assays (e.g., hemagglutination-inhibition, pseudovirus neutralization, or protein binding) that can be imperfect surrogates for how well a mutation enables a replication-competent virus to escape antibody neutralization [125, 103, 25].

A complete structural definition of the interface between an antibody and antigen can be obtained using methods such as X-ray crystallography. However, obtaining such structures remains non-trivial, particularly since viral surface proteins are often heavily glycosylated [140] and sometimes conformationally heterogeneous [91]. In addition, structural definitions do not reveal which mutations actually escape antibody neutralization. Mutations at only a subset of the residues in the antibody-antigen interface actually disrupt binding [72, 90, 35, 106], a “hot spot” phenomenon observed in protein-protein interfaces

more generally [33, 30, 18].

Here we use massively parallel experiments to rapidly map all single amino-acid mutations to HA that enable influenza virus to escape from four neutralizing antibodies. Our approach involves imposing antibody selection on virus libraries generated from all amino-acid point mutants of HA, and using deep sequencing to quantify the selection on every mutation in the context of actual replication-competent virus. The resulting comprehensive map of antibody escape reveals remarkable mutation-level idiosyncrasy for each antibody: for instance, at many residues only some of the possible amino-acid mutations confer escape, and two antibodies targeting the same residue elicit unique profiles of escape mutations. Mutational antigenic profiling therefore enables complete and high-resolution mapping of viral antibody escape mutations.

4.3 Results

4.3.1 Reproducible measurement of antibody selection on all amino-acid point mutations to influenza HA

To quantify the selection that neutralizing antibodies exert on all single amino-acid mutations to a viral protein, we developed the mutational antigenic profiling strategy shown in Figure 4.1. A library of viruses is generated that contains all amino-acid point mutants of the protein that are compatible with viral replication. This library is incubated with or without a neutralizing antibody, and then used to infect cells. Deep sequencing of cellular RNA measures the frequency of each mutation among the viral variants that infect cells in the presence or absence of antibody, with molecular barcoding used to increase the sequencing accuracy. We quantify the *differential selection* for each mutation as the logarithm of its enrichment in the antibody-treated virus library relative to the no-antibody control, and display these data as in Figure 4.1B. In the analysis that follows, we only consider mutations with positive differential selection.

We applied mutational antigenic profiling to influenza HA. HA is a 565-residue gly-

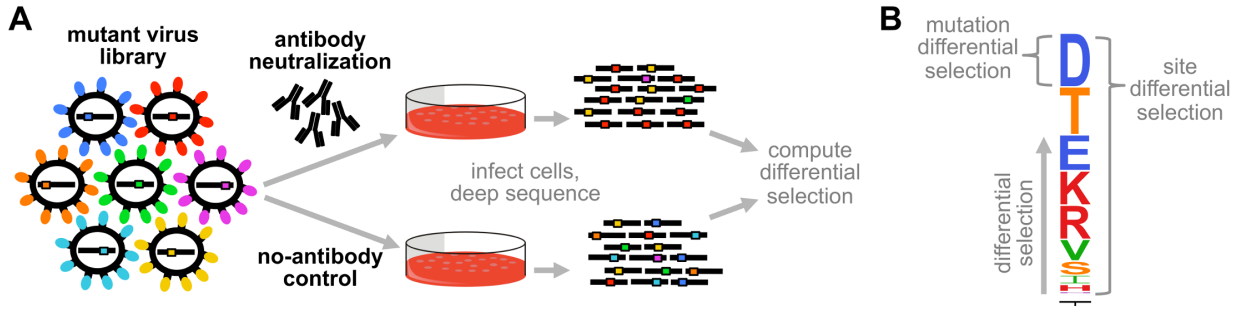

Figure 4.1: **Mutational antigenic profiling. (A)** Libraries of viruses carrying all amino-acid point mutants of a protein that support viral replication are incubated with or without antibody, and used to infect cells. Viral RNA is extracted from cells and accurately deep sequenced to quantify the frequency of each mutation in the antibody-selected and no-antibody control samples. **(B)** Differential selection is defined as the logarithm of the enrichment of each mutation in the antibody-selected sample versus the control. In the logo plots, the height of each letter is proportional to the differential selection for that amino-acid. The site differential selection is the total height of the logo stack at that site (the sum of mutation differential selection values). Only positive differential selection (corresponding to mutations enriched by selection) is shown. Logo plot letters are colored by physicochemical properties of amino-acids.

coprotein that forms homo-trimers on the virion surface that are responsible for both receptor-binding and membrane fusion [130]. Current influenza vaccines are designed to induce antibodies against HA, and the strains that compose these vaccines are chosen annually with the goal of matching the antigenicity of their HAs with those in circulating influenza variants [118, 9, 86, 93]. We chose to focus on the HA of an H1N1 strain (A/WSN/1933) that was isolated from humans early in the study of influenza and then serially passaged in the lab. Our reason for choosing this strain is that classic escape-mutant selections have extensively characterized the antigenicity of closely related HAs [53, 23], enabling us to compare our results to those obtained using more traditional methods.

The first step in mutational antigenic profiling is creating virus libraries (Figure 4.1A). A number of techniques have recently been described to create all amino-acid point mutants of a gene in the context of a plasmid [46, 74, 132]. The last few years have also seen the description of libraries of replication-competent virus mutants generated by adapting plasmid-based viral reverse-genetics systems to accommodate libraries of mutagenized

plasmids [123, 41, 135, 84, 59]. We utilized virus libraries created by melding these two techniques to create influenza viruses carrying all HA amino-acid point mutations compatible with viral replication [123, 41].

We initially selected these libraries with a monoclonal antibody (H17-L19) targeting the Ca2 antigenic region of HA [53]. We performed three biological replicates using independently generated virus libraries, as well as a technical replicate with one of the libraries (Figure 4.2A). The rationale for performing biological and technical replicates was to evaluate noise arising both from variability in the virus libraries and stochasticity in the antibody selections.

In each replicate, the antibody exerted strong selection for mutations at a handful of sites, and little selection on the rest of HA (Figure 4.2B). Figure 4.2C shows the selection for individual amino-acid mutations in a short region in HA containing most of the epitope. Visual inspection reveals consistent selection across technical and biological replicates. Statistical analysis confirms that the site differential selection is strongly correlated among replicates (Figure 4.2D).

We next asked how the differential selection depended on the concentration of antibody used. Figure 4.2 shows the results of mutational antigenic profiling at an antibody concentration where the virus libraries retained only 0.3% of their total infectivity. We performed additional experiments using dilutions of antibody that spanned a 20-fold range. Figure 4.3A shows the selection at each antibody concentration.

As expected, there is minimal selection when comparing replicate no-antibody controls. At progressively higher antibody concentrations, differential selection increases at most sites in the epitope, while noise at other sites remains similar to the no-antibody control. However, the increase in differential selection with antibody concentration is not entirely uniform across sites (Figure 4.3A). Figure 4.3B shows that despite these complexities, the sites of greatest differential selection are similar across concentrations, indicating that the identification of escape mutations does not strongly depend on antibody concentration within the 20-fold range tested here. Prior studies have shown that sub-

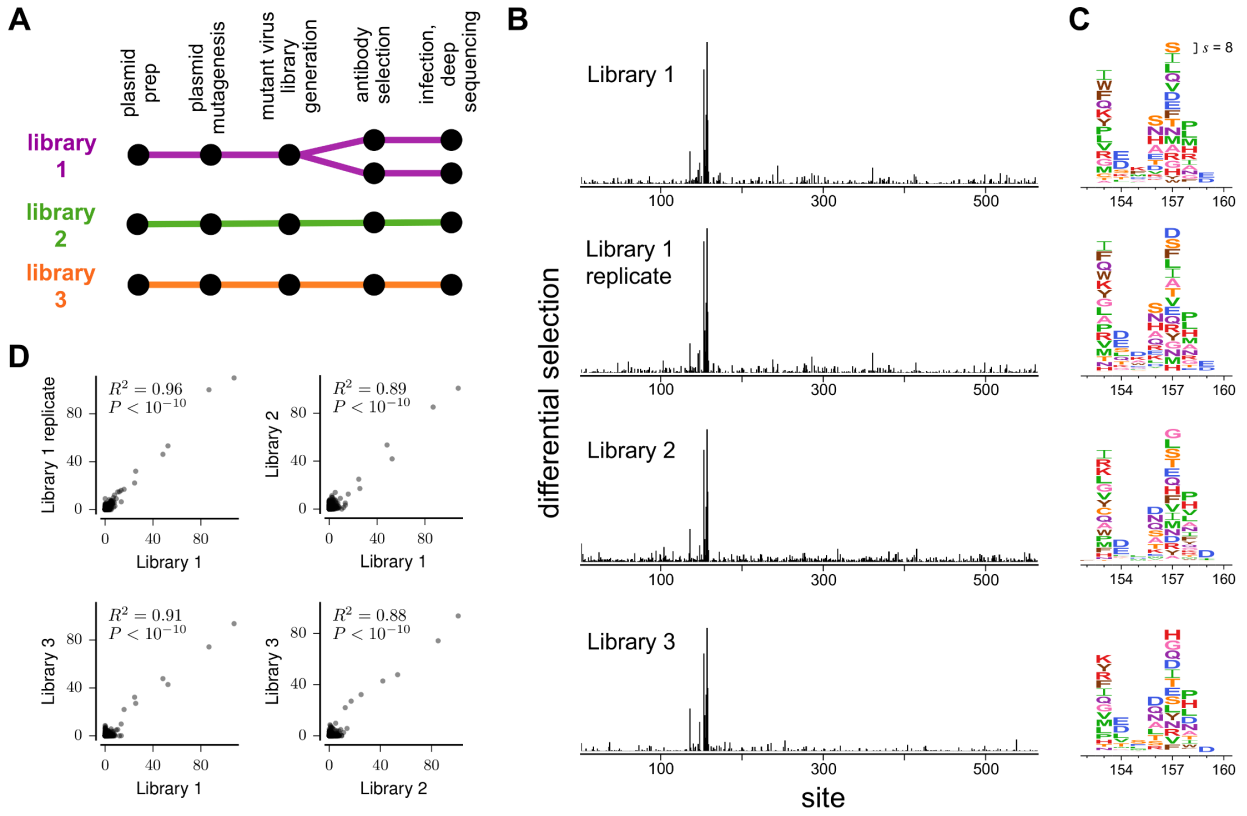

Figure 4.2: **Mutational antigenic profiling with antibody H17-L19 is highly reproducible. (A)** We performed three biological replicates and one technical replicate. **(B)** Site differential selection across HA is concentrated on the same subset of sites in all replicates. **(C)** Zoomed-in view of selection on the core of the epitope. The height of each letter is proportional to the differential selection for that amino-acid. The same scale is used in all panels of (B) and (C). The scale bar in the upper-right of (C) shows the letter height for a mutation with differential selection of 8, corresponding to $2^8 = 256$-fold enrichment by antibody selection. Residues are numbered sequentially beginning with the initiating methionine; conversions to other numbering schemes are at `https://github.com/mbdoud/mutational_antigenic_profiling/blob/master/HA_numbering.txt`. **(D)** Site differential selection across all sites is highly correlated among replicates. Each point represents selection at one site; correlation coefficients are Pearson's R. Data is shown for selections with antibody H17-L19 at 10 μg/ml.

neutralizing doses of mixtures of antibodies can select for mutations that increase the avidity of the virus for host cell receptors, as opposed to antigenic mutations within antibody epitopes [45, 139, 67]. The range of H17-L19 concentrations tested here is likely above the range of sub-neutralizing concentrations that have been used in the past to select for avidity-enhancing mutations, and it is also possible that mixtures of antibodies

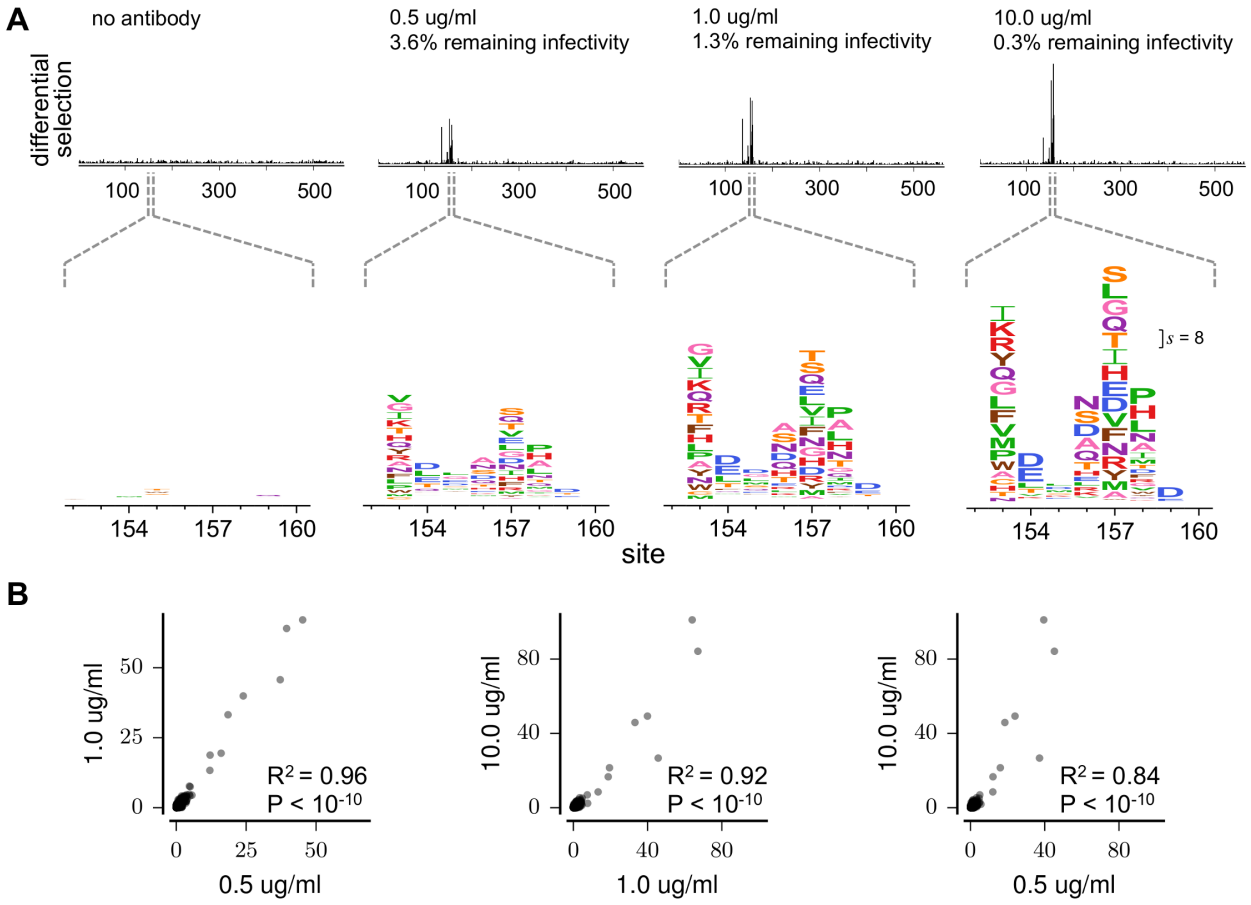

Figure 4.3: **Differential selection by H17-L19 at different antibody concentrations. (A)** Differential selection increases with antibody concentration. The top plots show site differential selection across HA; the bottom plots show the core of the epitope. All horizontally aligned plots use the same scale. The scale bar in the right-most plot shows the letter height for a mutation with differential selection of 8 (a 256-fold enrichment). The "no antibody" differential selection is computed between two replicate experiments on a single library. **(B)** Site differential selection is correlated between antibody concentrations, although the strength of selection increases at most sites with higher antibody concentration. Each point represents selection at one site in HA; correlation coefficients are Pearson's R. The data for each concentration is the average across the three biological replicates.

targeting different epitopes promote selection for avidity mutants. Understanding the determinants of how a mutation's differential selection depends on antibody concentration is an interesting area for future work.

Overall, these results confirm that mutational antigenic profiling reproducibly identifies

the HA mutations that confer escape from the monoclonal antibody H17-L19. The identified sites of selection are robust across replicate libraries and antibody concentrations.

4.3.2 Complete mapping of escape mutations for four monoclonal antibodies

We next extended the mutational antigenic profiling to three more antibodies. We performed selections with each antibody at concentrations at which the virus libraries retained 0.1 to 0.4% of their infectivity (Table C.1). Each antibody exerted strong selection at a small number of residues in HA. Figure 4.4A shows site differential selection across HA, while Figure 4.4B uses logo plots to show detailed mutation-level selection at some key positions in the antibody epitopes. We again performed three full biological replicates with each antibody, and the results were again highly reproducible among replicates (Figure C.1).

For each antibody, the sites of strongest differential selection were clustered in surface-exposed patches on HA's structure that are presumably within the antibody-binding footprint (Figure 4.4C and Figure C.2). The four antibodies target three antigenic regions: H17-L19 targets Ca2, H17-L10 targets Ca1, and H17-L7 and H18-S415 both target Cb [53, 23]. As expected, H17-L19 and H17-L10 exert strong selection on entirely distinct sets of residues, but H17-L7 and H18-S415 exert selection on similar sets of residues in the Cb antigenic region. For three of the antibodies, the strongly selected residues are within short contiguous stretches of primary amino-acid sequence, but for H17-L10 the strongly selected residues are distributed across 70 residues of HA's primary sequence. Overall, these results show that mutational antigenic profiling can comprehensively identify the selection imposed by diverse antibodies.

4.3.3 Comparison to traditional neutralization assays

The results above were obtained using experiments that examined tens of thousands of viral variants in parallel. How do these high-throughput measurements compare to the

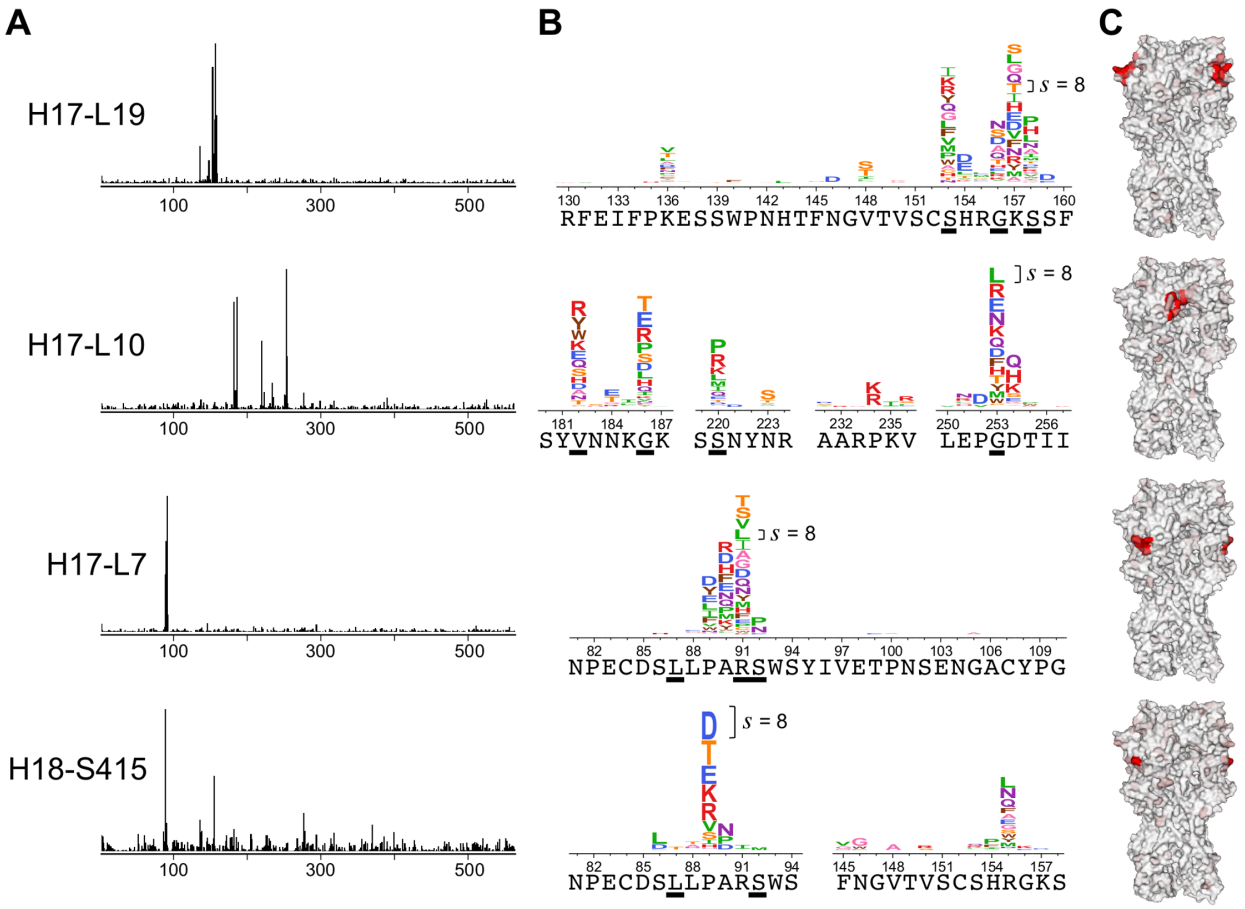

Figure 4.4: **Mutational antigenic profiling of four antibodies. (A)** Each antibody exerts a different profile of selection on HA. **(B)** Zoomed in view of some of the most strongly selected sites for each antibody. The wild-type amino acid is shown under the logoplots. Sites where mutations were selected in classical escape-mutant selections [53, 23] are underlined. Logoplots spanning all of HA are in Figure C.3, Figure C.4, Figure C.5, and Figure C.6. **(C)** The selection from each antibody visualized on HA's structure (PDB 1RVX [51]). Each site is colored from white to red based on the differential selection for the most strongly selected mutation at that site. Red indicates strong differential selection. All structures show trimeric HA in the same orientation (the epitope is visible for two of the three monomers for H17-L19, H17-L7, and H18-S415). See Figure C.2 for a zoomed-in structural view. The y-axis scale is set separately for each antibody; since the measured strength of differential selection depends on the concentration / potency of the antibody and the mutational tolerance of the viral epitope, it was impossible to precisely standardize selection strength across antibodies. The scale bar in each logo plot shows the letter height for a mutation with differential selection of 8 (a 256-fold enrichment). The data for each antibody is the average across three biological replicates.

antigenic effects of mutations measured by traditional low-throughput methods? To address this question, we tested some of our key findings with neutralization assays on individual viral mutants. To do this, we used site-directed mutagenesis to introduce single amino-acid mutations into the HA gene, generated viruses by reverse genetics, and performed GFP-based neutralization assays [70].

A clear observation from the mutational antigenic profiling is that at some residues, only a few of the possible amino-acid mutations are strongly selected by any given antibody, concordant with prior work showing that a limited number of mutations are sufficient for antigenic drift [75]. For instance, at HA residue 154, the H17-L19 antibody exerts strong selection only for mutations H154E and H154D, both of which introduce a negatively charged amino acid (Figure 4.4B, Figure 4.5A; residues are numbered sequentially beginning at the N-terminal methionine, other numbering schemes are in `https://github.com/mbdoud/mutational_antigenic_profiling/blob/master/HA_numbering.txt`). We generated viruses carrying the H154E mutation or a mutation to alanine (H154A), which mutational antigenic profiling did not find to be under differential selection. Neutralization assays confirmed that the H154E mutant completely escaped at all antibody concentrations tested, while the H154A mutant was as sensitive to antibody as wild-type (Figure 4.5A). Therefore, a more limited method such as alanine scanning would not have identified residue 154 as a site of escape mutations. This finding demonstrates the importance of assaying all amino-acid mutations if the goal is to comprehensively map sites of escape.

Another example of mutation-level sensitivity is HA residue 148, where antibody H17-L19 only selects for mutations to serine and threonine (Figure 4.4B). Both the V148T and V148S mutations introduce a motif (N-X-S/T) that potentially leads to glycosylation of the asparagine at site 146. To confirm that only some mutations at site 148 enable escape, we generated the V148T mutant as well as another mutant (V148R) that does not introduce a glycosylation motif. As expected, V148T dramatically reduced the virus's sensitivity to the antibody, whereas V148R only had a small effect (Figure 4.5A).

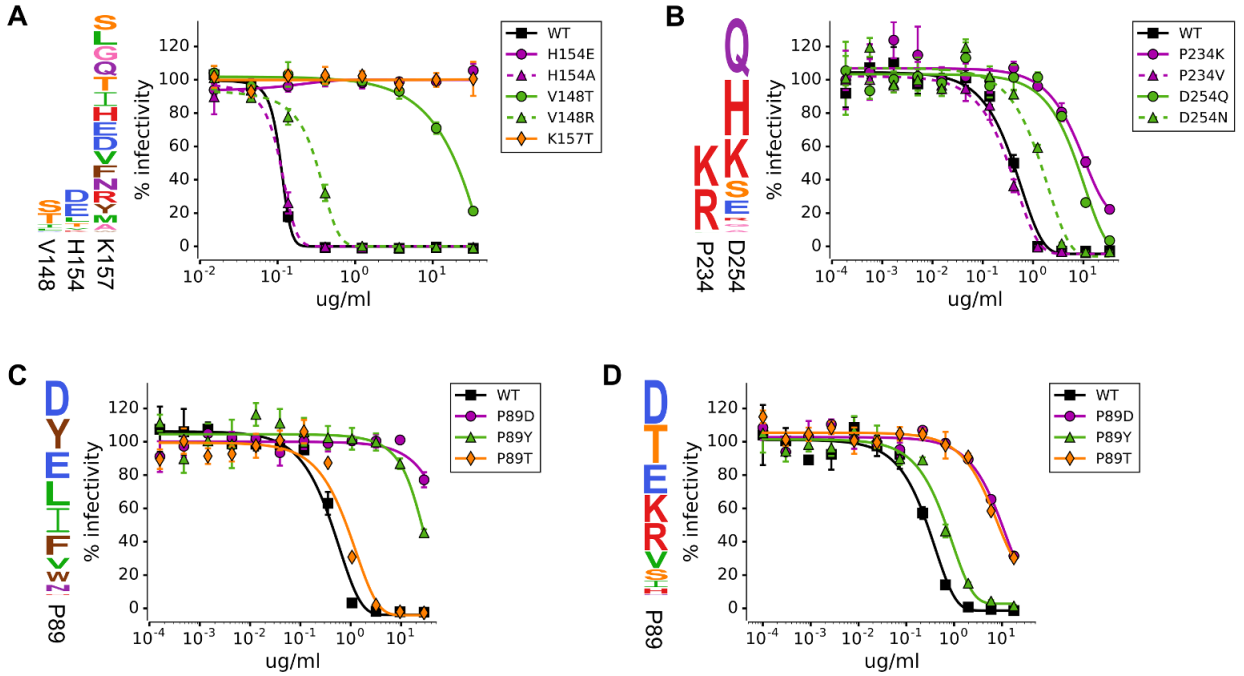

Figure 4.5: **Comparison of the selection measured by mutational antigenic profiling with the antigenic effects of mutations in traditional neutralization assays on individual viral mutants.** In each panel, the logo plot shows the results of the mutational antigenic profiling at the sites of mutations chosen for testing, and the graph shows the results of the neutralization assays. There is excellent concordance between whether a mutation is strongly selected in the mutational antigenic profiling and whether it has an effect in the neutralization assay. In many cases, only some of the amino-acid mutations at a site strongly affect neutralization by a given antibody – and the mutational antigenic profiling reliably distinguishes between mutations that do and do not have an effect. The antibodies in each panel are: **(A)** H17-L19, **(B)** H17-L10, **(C)** H17-L7, **(D)** H18-S415.

The mutational antigenic profiling suggests similar mutation-level sensitivity in escape from antibody H17-L10. At residue 234, there is strong differential selection only for mutations to the positively charged amino-acid residues lysine and arginine (Figure 4.4B). We generated a virus carrying one of these mutations (P234K) as well as a virus carrying another mutation at the same residue (P234V) that was not under differential selection. Neutralization assays confirmed that the P234K mutation escaped H17-L10, while the P234V mutation caused no change in antibody sensitivity (Figure 4.5B). Interestingly, in HA's structure, site 234 is on a neighboring protomer relative to all the other mutations strongly selected by H17-L10 (Figure C.2). Our finding that escape mutations from

H17-L10 cross the HA trimer interface is consistent with the fact that this antibody only recognizes trimeric HA [88]. Escape mutations at such epitopes are discernible because mutational antigenic profiling uses actual viruses that display intact HA; such conformational epitopes might not be properly displayed in the modified forms of viral glycoproteins often used in other high-throughput methods such as phage and yeast display.

Overall, these results indicate the power of mutational antigenic profiling to map residues where only a few specific amino-acid mutations lead to escape from antibody. Because this approach examines HA in its native context on influenza virions, it can comprehensively map escape mutations even in complex conformational epitopes.

4.3.4 Unique repertoires of escape mutations from two antibodies targeting the same site in HA

Two of the antibodies used in our study (H17-L7 and H18-S415) target the same antigenic region of HA, with residue 89 under strong selection from both antibodies (Figure 4.5C,D). Do these antibodies select the same or different amino-acid mutations at this residue? The mutational antigenic profiling suggests that both antibodies select mutations to negatively charged amino acids (P89D and P89E; Figure 4.5C,D). However, each antibody also selects a unique set of additional mutations, such as P89Y for H17-L7 and P89T for H18-S415.

We generated viruses containing the P89D, P89Y, or P89T mutations and tested their sensitivity to both antibodies using neutralization assays. In agreement with the mutational antigenic profiling, the P89D mutant escaped both antibodies, but P89Y only escaped from H17-L7 and P89T only escaped from H18-S415 (Figure 4.5C,D). Thus, when two antibodies target the same site, there can be both common and antibody-specific routes of escape. Characterizing antibody escape at the level of protein sites therefore only provides a partial picture of antigenicity. A complete understanding of escape requires consideration of every mutation at every site.

4.3.5 Comparison to classical escape-mutant selections

The antigenicity of H1 HA was originally characterized in classic experiments that selected individual viral escape mutants with a panel of mouse monoclonal antibodies [53, 23]. These experiments identified a handful of mutations that ablated binding by each antibody (Table C.2 and underlined residues in Figure 4.4B). All four antibodies used in our study are from the original panel used in the classic experiments. We expected that the sites of differential selection identified by mutational antigenic profiling would include the previously identified mutations.

Indeed, there is strong overlap between sites identified by mutational antigenic profiling and sites of mutations selected in the classic experiments (Figure 4.4B). However, we also identified numerous additional escape mutations at those and other sites. In some cases, the sites of strongest differential selection were not identified at all in the classic experiments. For instance, as shown in Figure 4.4, the classic escape-mutant selections failed to identify site 157 for H17-L19, site 89 for H17-L7, and site 89 for H18-S415. Differences in the virus strains used (as discussed below) may account for some of these discrepancies. Additionally, it is likely that mutations at these sites were not uncovered in escape-mutant selections because each such selection only finds one mutation, with a strong bias towards those that arise from single-nucleotide changes that are prevalent in the viral stock. In contrast, our approach simultaneously examines all amino-acid point mutations.

The exception to the concordance between mutational antigenic profiling and classic escape-mutant selections is antibody H18-S415 (Figure 4.4B). The classic selections failed to identify any mutations at site 89 despite the fact that mutational antigenic profiling finds by far the strongest differential selection at this residue. This discrepancy is not due to spurious signal in the mutational antigenic profiling, since Figure 4.5D validates that mutations at site 89 potently escape H18-S415. Perhaps the stochasticity of escape-mutant selections caused the classic experiments to fail to probe mutations at site

89.

It is worth noting that the differential selection exerted by H18-S415 in our experiments is substantially noisier than the differential selection for the other antibodies (Figure 4.4A, Figure C.1). In another recent study, H18-S415 selected an escape virus containing both a mutation in HA and a mutation in the neuraminidase (NA) gene that decreased NA protein expression, leading to increased virus avidity for host cell receptors [37]. The selection of avidity-enhancing mutations has been observed in selection of escape viruses using mixtures of antibodies [139, 67], and it is even possible that the H18-S415 hybridoma cell line is not completely monoclonal. Alternatively, it is possible that it is simply more difficult for the virus to escape H18-S415, and so there is more stochastic noise in which mutations appear in our selections.

The mutational antigenic profiling also failed to find strong selection from H18-S415 for some mutations reported in the classic experiments (L87P, S92P, and E132K; see Figure 4.4B, Table C.2, and Figure C.6). Why were these mutations selected in the classic experiments but not the mutational antigenic profiling? An important point is that the classic experiments used a different virus strain (A/Puerto Rico/8/1934) than the A/WSN/1933 strain used for our mutational antigenic profiling. In order for a mutation to be under differential selection, it must both support viral replication and affect antigenicity. The mutations L87P, S92P, and E132K are all strongly disfavored under simple selection for viral replication in the A/WSN/1933 strain [41], which likely explains why they are not under strong differential selection in our mutational antigenic profiling. This fact is an important reminder that while mutational antigenic profiling completely maps antibody selection on all single amino-acid mutations that support viral replication in a given viral strain, it does not reveal how the effects of mutations shift with changes in strain background. It remains an open question how well measurements of the effects of mutations on viral replication [40, 133] and antigenicity [37] can be extrapolated beyond the specific genetic backgrounds tested in the lab.

4.4 Discussion

We have used a new high-throughput approach to completely map the amino-acid mutations that enable influenza virus to escape from four neutralizing antibodies. Our approach is conceptually similar to recent methods that couple deep sequencing with phage or yeast display assays for antibody binding [76, 1, 50]. But whereas those methods select for binding to antigens expressed in bacteria or yeast, our approach selects for actual neutralization in the context of replication-competent virus. Our experiments therefore measure a phenotype directly relevant to virus evolution: whether a mutation enables a virus to escape neutralization by an antibody.

Our approach also bears similarities to the classic method of selecting individual viral escape mutants. However, escape-mutant selections rely on the occurrence of *de novo* mutations in a viral stock. Therefore, like evolution itself, such selections are stochastic, and only identify one of potentially many escape mutations. In contrast, our massively parallel experiments simultaneously examine all single amino-acid mutations, thereby minimizing stochasticity and allowing us to completely map antibody selection on all point mutations.

The most striking finding from our work is the exquisite mutation level-sensitivity of antibody escape. For each of the four antibodies, we identified residues in HA where only some of the possible amino-acid mutations conferred escape. In some cases, this mutation-level sensitivity is easy to rationalize: we found examples where escape required mutations that introduce glycosylation motifs or change the charge of the amino-acid sidechain. But in other cases, the effects are not only difficult to rationalize but depend on the antibody. For instance, we identified a residue targeted by two antibodies where a mutation that escaped the first antibody had no effect on the second and vice versa. Previous studies have distinguished between an antibody's "functional epitope" and physical footprint based on the observation that binding is disrupted by mutations at only some residues that contact the antibody [72, 90, 35, 106]. Our findings extend this concept by

showing that even within the functional epitope, only certain mutations mediate escape, consistent with the observation that a small number of amino-acid mutations in HA can cause extensive antigenic drift of H3N2 influenza virus [75].

These results underscore the shortcomings of thinking about viral antigenic evolution purely in terms of antigenic sites. For instance, many approaches to forecast and model influenza virus evolution are based on partitioning HA into antigenic and non-antigenic sites [86, 93]. However, our work shows that for any individual antibody, it is important to consider the exact amino-acid mutation as well as the site at which it occurs. Application of mutational antigenic profiling to contemporary viral strains and antibodies will enable the prospective mapping of immune-escape mutations on a vastly more comprehensive scale than previously possible.

Chapter 5

CONCLUSION

Deep mutational scanning can be used to comprehensively measure the effects of mutations to influenza genes. In this dissertation I presented applications of deep mutational scanning to the influenza nucleoprotein (NP) and hemagglutinin (HA) genes. In each case, I made plasmid mutant libraries with codon mutagenesis to introduce all possible amino-acid altering mutations, generated mutant virus libraries from these plasmid libraries, subjected mutant virus libraries to selection, and used deep sequencing to measure changes in mutation frequencies before and after selection. First, by selecting for viral replication in cell culture, I measured the inherent tolerance for mutation to all possible amino acids at every site in these two genes. The measurements of mutational tolerance in NP allowed me to make the first comparative study of the effects of *all* mutations to two homologs of the same protein. The HA mutant virus libraries subsequently allowed for the the first comprehensive profiling of escape mutations to neutralizing monoclonal antibodies. The experimental and analytical approaches developed in all three studies should be generally applicable to the generation of mutant influenza libraries for the study of specific phenotypes beyond replication and antibody escape, in the analysis of deep mutational scanning datasets on multiple homologs of a protein, and in the analysis of differential selection between defined selective pressures.

As expected, the mutational tolerance results were consistent with the known structure and biological function of both viral proteins. For example, one of HA's most important roles in the viral lifecycle is to bind the virus to host receptors, and the tolerance for mutations in the receptor-binding pocket is low, likely because most mutations at these sites interfere with receptor binding (Figure 3.6). Similarly, one of NP's biological functions is

to encapsidate the viral RNA genome along an RNA-binding groove on the surface of the protein, and these RNA-binding sites were among those with the most conserved functional constraint when I compared mutational effects across two NP homologs (Figure 2.6).

Despite the fact that there are obvious selective pressures that exist in nature that I could not recapitulate in a laboratory experiment, the mutational tolerance results are also reflective of evolutionary constraint on these genes. Site-specific substitution models informed by these experiments modeled the natural molecular evolution of HA and NP over the past century vastly more accurately than standard substitution models lacking experimental information on mutational effects (Tables 2.1, 2.2, 3.1).

Mutational effects are mostly conserved between two viral protein homologs. If mutational effects change significantly as a protein's sequence evolves, the utility of using these data to model evolution over large sequence divergences will degrade. There are many examples of experimentally validated epistasis, where a mutation at a single site can shift the amino-acid preferences at other sites [129, 96, 34, 87, 56, 92, 101]. Clearly this is a phenomenon that happens, but *how frequently, and to what extent*, do amino-acid preferences shift during the molecular evolution of a particular gene? I first approached this question by directly comparing mutational effects measured in deep mutational scanning on two NP homologs, and found that at most sites, mutational effects are conserved between these homologs (Figure 2.5).

Put another way, to what extent do the site-specific amino-acid preferences measured in one viral strain describe the evolution of more distantly-related viruses? This is the more practical question, since we are interested in applying the results of deep mutational scanning on specific strains of influenza virus to modeling and predicting the evolution of related, but diverged viral genotypes. The data from either NP homolog can accurately describe the evolution of NP from a wide range of influenza viruses infecting humans, pigs, horses, and birds (Figure 2.1, Tables A.1, A.2, A.3), demonstrating that mutational effects

measured in one protein homolog can model the evolution of more distant homologs.

How far across evolutionary time will this hold? As protein sequences become more and more diverged, we expect the effects of mutations to shift more as well. At what point will the frequency and magnitude of epistasis become so great that site-specific amino-acid preferences measured in one homolog fail to describe the evolution of distant homologs? Experimental evidence suggests that despite the fact that epistasis results in shifts at some sites, the amino-acid preferences of many, if not most sites remain similar across homologs with relatively low levels of sequence identity [111, 16]. Future work comparing comprehensive measurements of mutational effects directly between homologs of other proteins at greater levels of sequence divergence is needed to definitively answer this question. Of particular interest will be direct comparisons between homologs of rapidly-evolving viral proteins such as influenza hemagglutinin or HIV Env [59].

Mutational antigenic profiling completely maps viral mutations conferring escape from antibodies. Selection imposed from various components of the immune system, such as neutralizing antibodies, is thought to be a primary driver of influenza's rapid evolution. Some protein sites in HA have long been classified as antigenic sites, based on classic work selecting individual escape mutations with neutralizing monoclonal antibodies [53, 23], and binary classifications of antigenicity are often used to model and predict viral evolution [86, 93]. But do all mutations within these antigenic sites have equal effects on antigenicity? How complete is our understanding of the full spectrum of mutations that can confer escape from neutralizing antibodies?

Prior to the work presented here, many methods have been developed to map antibody epitopes, and to approach the related problem of identifying mutations within antibody epitopes that confer antibody escape. X-ray crystallography is often used to delineate the structural footprint of antibodies on antigens, but protein-protein interactions are driven by 'hot-spots' of binding energy, making it difficult, if not currently impossible, to accurately predict how mutations within the structural footprint of an antibody will affect molecular

recognition [72, 90, 35, 106, 33, 30, 18].

Since predictions are difficult, mutagenesis-based approaches to mapping antibody epitopes have the benefit of actually testing mutants for binding or neutralization by antibody. Alanine scanning mutagenesis has been a useful technique which, as opposed to testing all possible mutations, only tests alanine mutations, thereby reducing library complexity to match available experimental throughput. However, the effects of other mutations are difficult to predict. Physiochemically "conservative" mutations (eg., arginine to lysine) can have unexpectedly large effects on antibody binding [24], and mutations structurally predicted to interfere with binding can be accommodated by molecular rearrangements at the interface [35]. Truncating the amino-acid side chain at the site of interest to alanine will not necessarily result in the same effect as substituting to one of the other amino acids. Error-prone PCR libraries go a step beyond alanine scanning libraries by introducing random nucleotide point mutations instead of focusing solely on alanine mutations. However, the single-nucleotide mutations introduced by error-prone PCR are limited by the structure of the genetic code in which amino-acid substitutions are available for any given codon, and on average only half of the possible amino-acid sequence space can be accessed through error-prone mutagenesis. Codon mutagenesis, on the other hand, offers the ability to sample *all* possible amino-acid point mutations to a protein sequence. Recent methods have coupled deep sequencing with phage or yeast display assays for antibody binding [76, 1, 50]. However, these approaches select only for antibody binding to antigens expressed in bacteria or yeast, and some antibody epitopes, such as those crossing the interface between neighboring protomers in the influenza hemagglutinin trimer (Figure C.2, [88]), may not be appropriately assembled when displayed on the surface of phage or yeast.

As opposed to selection for binding to displayed antigens, the actual neutralization of replication-competent virus may be a more relevant phenotype to select for when identifying escape mutations is the goal. In the context of infectious viruses, escape mutant selections have classically been used to map antibody escape, but as discussed in **Chap-**

ter 4, this method relies on de novo mutations within wild-type stocks of virus, greatly limiting the mutation sampling and throughput.

I developed a new experimental approach utilizing codon mutagenesis, helper-virus based rescue of mutant virus libraries, selection by monoclonal antibody, and deep sequencing to completely map viral mutations conferring escape neutralizing antibodies. There are several strengths to this approach over existing approaches: it samples *all* amino-acid mutations that are compatible with viral replication, and completely and reproducibly identifies those mutations that confer antibody escape in the context of an infectious virus. Once the mutant virus libraries are made, it is relatively easy to profile multiple neutralizing antibodies in parallel. The results I obtained with four monoclonal antibodies targeting three antigenic regions on HA revealed that at most sites within each antibody epitope, only some of the possible amino-acid mutations confer escape, and even similar antibodies targeting the same site in HA elicit unique profiles of escape mutants (Figures 4.4, 4.5).

Previous work investigating the effects of mutations on antibody recognition have largely been limited to querying the effects of alanine substitutions in the epitope. This line of work established that an antibody's "functional epitope" is comprised of only a subset of the sites within the structural epitope [72, 90, 35, 106]. However, alanine substitutions are not necessarily indicative of the effects of all the other possible amino-acids. Here I extend this concept by showing that even within the "functional epitope", only certain mutations at each site mediate escape, and alanine scans are not suitable to detect all sites within the "functional epitope". This cautions against a simplified view of viral antigenic evolution defined by classifications of sites as antigenic or non-antigenic, since for individual antibodies, the specific mutations conferring escape vary both by site and by antibody.

There are several important questions raised by these results. This approach maps all mutations to one specific strain that confer antibody escape, but to what extent do these effects vary across sequence contexts? This remains an important question for

future work to address, since changes in either mutational tolerance or antigenicity between viral strains could in principle change the repertoire of escape mutations. To what extent does mutational tolerance constrain the repertoire of escape mutants? Are there many mutations that are normally deleterious, but would confer escape in the context of a compensating mutation that allows the virus to tolerate the escape mutation? Various approaches can be taken to begin to answer these questions. Mutational antigenic profiling can be performed on multiple diverged viral genotypes with a panel of cross-neutralizing antibodies in order to explore the extent that A) tolerated mutations can confer escape in one genotype context vs. another, and B) changes in mutational tolerance between genotypes can shape the available repertoire of escape mutations. Furthermore, alternative library construction strategies can prioritize a small set of sites (for instance, within the binding footprint of an antibody) to reduce the combinatorial complexity and allow for sufficient sampling of higher-order multiple mutants. It will be interesting to see the extent that epistasis among multiple sites within an antibody epitope can shape the effects of mutations on antibody escape.

Obviously, the selective pressures of the immune system on the natural evolution of influenza virus are much more complicated than the effects of monoclonal antibodies that can neutralize the virus *in vitro*. What is the best way to recapitulate the natural selective pressures in a mutational antigenic profiling experiment? How do multiple monoclonal antibody specificities combine to exert selection on the virus? Mutational antigenic profiling of mixtures of two or more monoclonal antibodies such as those used in these experiments can be used to begin to recapitulate the selective pressures of polyclonal serum. Such 'synthetic' polyclonal sera can be constructed with varying potencies of each monoclonal antibody to examine whether selection with multiple antibodies simultaneously is additive with respect to monoclonal selective pressures. Future work will need to evaluate various *in vitro* and *in vivo* methods for immune selection on mutant virus libraries with polyclonal sera. Applications of this approach with contemporary viruses and antibodies will be necessary to reveal the full potential for using these data to forecast viral evolution.

Appendix A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

Table A.1: **Combining experimentally informed substitution models for swine influenza NP.** This table differs from Table 2.1 in that the phylogenetic fit is for the tree of swine NPs shown in Figure 2.1.

model	ΔAIC	log like-lihood	parameters (optimized + empirical)	optimized parameters
Aichi/1968 + PR/1934	0.0	-6832.4	5 (5 + 0)	$R_{A\to G}$ = 4.9, $R_{A\to T}$ = 0.9, $R_{C\to A}$ = 1.4, $R_{C\to G}$ = 0.1, β = 2.8
PR/1934	390.4	-7027.6	5 (5 + 0)	$R_{A\to G}$ = 5.1, $R_{A\to T}$ = 0.9, $R_{C\to A}$ = 1.4, $R_{C\to G}$ = 0.1, β = 2.1
Aichi/1968	563.3	-7114.0	5 (5 + 0)	$R_{A\to G}$ = 5.0, $R_{A\to T}$ = 0.8, $R_{C\to A}$ = 1.4, $R_{C\to G}$ = 0.1, β = 2.4
GY94, gamma ω, gamma rates	2153.5	-7901.2	13 (4 + 9)	κ = 5.9, ω shape = 0.3, mean ω = 0.0, rate shape = 2.9

Table A.2: **Combining experimentally informed substitution models for equine influenza NP.** This table differs from Table 2.1 in that the phylogenetic fit is for the tree of equine NPs shown in Figure 2.1.

model	ΔAIC	log like-lihood	parameters (optimized + empirical)	optimized parameters
Aichi/1968 + PR/1934	0.0	-2458.1	5 (5 + 0)	$R_{A\to G}$ = 10.4, $R_{A\to T}$ = 0.7, $R_{C\to A}$ = 1.5, $R_{C\to G}$ = 0.4, β = 2.8
PR/1934	244.1	-2580.1	5 (5 + 0)	$R_{A\to G}$ = 10.6, $R_{A\to T}$ = 0.7, $R_{C\to A}$ = 1.5, $R_{C\to G}$ = 0.4, β = 2.0
Aichi/1968	337.5	-2626.8	5 (5 + 0)	$R_{A\to G}$ = 10.7, $R_{A\to T}$ = 0.6, $R_{C\to A}$ = 1.4, $R_{C\to G}$ = 0.3, β = 2.4
GY94, gamma ω, gamma rates	2270.1	-3585.2	13 (4 + 9)	κ = 13.0, ω shape = 0.0, mean ω = 0.1, rate shape = 1.7

Table A.3: **Combining experimentally informed substitution models for avian influenza NP.** This table differs from Table 2.1 in that the phylogenetic fit is for the tree of avian NPs shown in Figure 2.1.

model	ΔAIC	log like-lihood	parameters (optimized + empirical)	optimized parameters
Aichi/1968 + PR/1934	0.0	-5686.6	5 (5 + 0)	$R_{A\to G}$ = 8.7, $R_{A\to T}$ = 1.1, $R_{C\to A}$ = 1.3, $R_{C\to G}$ = 0.0, β = 3.2
PR/1934	334.3	-5853.7	5 (5 + 0)	$R_{A\to G}$ = 9.0, $R_{A\to T}$ = 1.1, $R_{C\to A}$ = 1.3, $R_{C\to G}$ = 0.0, β = 2.3
Aichi/1968	639.0	-6006.1	5 (5 + 0)	$R_{A\to G}$ = 8.9, $R_{A\to T}$ = 1.0, $R_{C\to A}$ = 1.3, $R_{C\to G}$ = 0.0, β = 2.5
GY94, gamma ω, gamma rates	1730.0	-6543.6	13 (4 + 9)	κ = 9.2, ω shape = 0.0, mean ω = 0.0, rate shape = 1.9

Figure A.1: **Logoplot of amino-acid preferences for PR/1934 NP.** The mean preferences for sites 2 through 498 of PR/1934 are represented in a sequence logo-like visualization created with the program `dms_logoplot`. The height of each letter is proportional to the preference for that amino-acid at that site.

Figure A.2: **Logoplot of amino-acid preferences for Aichi/1968 NP.** The mean preferences for sites 2 through 498 of Aichi/1968 are represented in a sequence logo-like visualization created with the program `dms_logoplot`. The height of each letter is proportional to the preference for that amino-acid at that site.

Figure A.3: **Logoplot of amino-acid preferences for combined PR/1934+Aichi/1968 NP.** The mean preferences for sites 2 through 498 of the combined Aichi/1968 + PR/1934 model are represented in a sequence logo-like visualization created with the program `dms_logoplot`. The height of each letter is proportional to the preference for that amino-acid at that site.

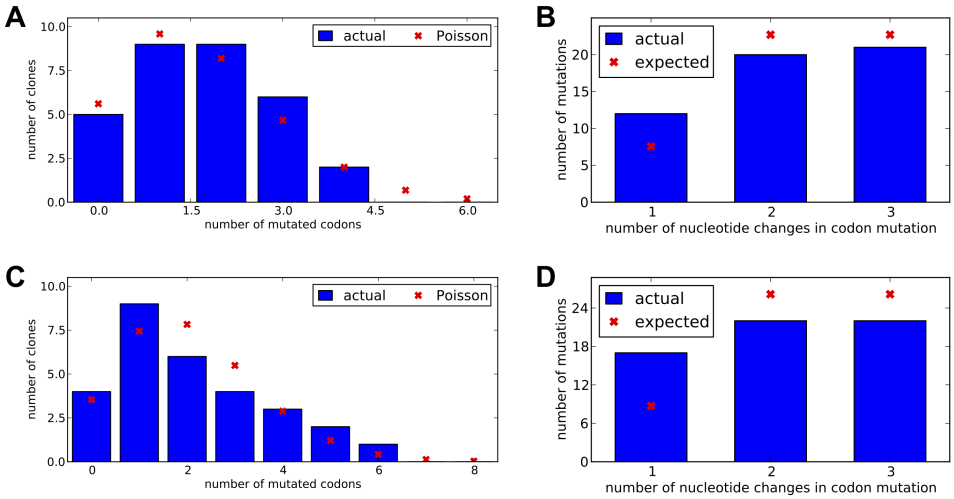

Figure A.4: **Characterization of NP plasmid mutant libraries generated by codon mutagenesis**. The distributions of number of mutated codons per clone (A, C) and number of nucleotide changes per codon mutation (B, D) were determined by full-length Sanger sequencing of individual clones. A-B: PR/1934 libraries, C-D: Aichi/1968 libraries.

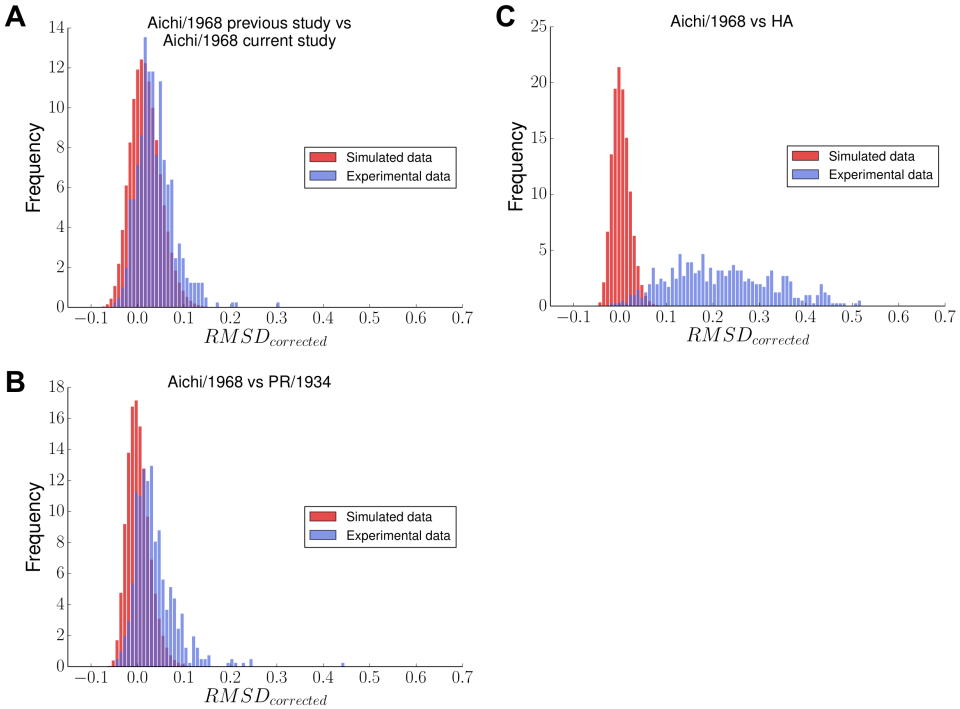

Figure A.5: **Null distributions of $RMSD_{corrected}$ generated by simulation.** The null distributions generated by simulation are shown in red; experimental distributions are shown in blue.

Materials and Methods

Availability of data and computer code

FASTQ files can be accessed at the Sequence Read Archive (SRA Accession SRP056028). The computer code necessary to reproduce all the analysis in this work is available at `https://github.com/mbdoud/Compare-NP-Preferences`.

Deep mutational scanning of two influenza nucleoprotein homologs

We performed deep mutational scanning of influenza nucleoprotein (NP) in three biological replicates for A/PR/1934 (H1N1) and two biological replicates for A/Aichi/1968 (H3N2) (termed here as Aichi/1968 *current study*). We broadly followed the methods used for mutagenesis, viral rescue, deep sequencing, and inference of amino-acid preferences from sequence data described in [15], with the following notable changes to the protocol.

Codon mutagenesis. For each replicate mutant library, we followed the mutagenesis protocol as previously described [15], but performed two rounds of mutagenesis instead of three to decrease the average number of mutations per clone. After ligation of mutagenized PCR products to the pHW2000 [69] plasmid backbone, multiple parallel transformations and platings were combined to ensure that each replicate library contained more than 10^6 unique transformants. Sanger sequencing of 30 clones from each homolog revealed that the number of mutations per clone was approximately Poisson distributed with an average of 1.7 mutations per clone for the PR/1934 libraries and 2.1 mutations per clone for the Aichi/1968 libraries, with mutations distributed uniformly across the length of the gene.

Growth of mutant virus libraries. We used reverse genetics [69] to rescue viruses carrying mutant NP genes. Co-cultures of 293T and MDCK-SIAT1 cells were plated 16

hours prior to transfection in D10 media (DMEM supplemented with 10% FBS, 100 U/mL of penicillin,100 μg/mL of streptomycin, and 2 mM L-glutamine) at cell densities of 3x10^5 293T/mL and 2.5x10^4 MDCK-SIAT1/mL. Co-cultures were transfected using BioT transfection reagent (Bioland Scientific) with a mixture of 250 ng of each of the eight reverse genetics plasmids per well in 6-well plates. In order to circumvent the possibility of rare mutants with exceptional replication fitness growing to high frequencies and limiting the growth of other mutants, we divided each transfection into multiple tissue-culture wells.

For the PR/1934 libraries, we rescued viruses containing the mutagenized PR/1934 NP with the seven remaining PR/1934 viral gene segments, and each replicate mutant library was transfected into the twelve wells of two 6-well plates. For the Aichi/1968 libraries, we used a viral rescue protocol that increases the number of parallel transfections and uses 293T cells that constitutively express protein V from hPIV2. This protein targets STATI for degradation, thereby inhibiting type I interferon signaling [3]. We rescued these Aichi/1968 virus libraries by transfecting the Aichi/1968 NP mutant library along with PB1/PB2/PA from Nanchang/933/1995 (using the plasmids in [56] and HA/NA/M/NS from WSN/1933 into 48 wells of eight 6-well plates. For both homologs, in parallel, we performed similar transfections using the corresponding unmutated NP genes to grow unmutated virus.

At 24 hours after transfection, co-culture media was aspirated, cells were rinsed with PBS, and the media was changed to influenza growth media (OptiMEM I media (Gibco) supplemented with 0.01% FBS, 0.3% BSA, 100 U/mL of penicillin,100 μg/mL of streptomycin, 100 μg/mL calcium chloride, and 3 μg/mL TPCK-trypsin). Co-culture supernatant was collected 72 hours after transfection, clarified by centrifugation at 2,000xg for 5 min, aliquoted and stored at -80° C.

Since many of the virions obtained from transfection with mutant NP library plasmids are likely to have originated in cells that contained more than one mutant NP gene and therefore might carry NP genes and NP proteins with different mutations, we passaged viruses in MDCK-SIAT1 cells at a low multiplicity of infection (MOI) to enforce genotype-

phenotype linkage. We titered viruses from thawed transfection supernatant aliquots for each replicate virus library using the TCID50 protocol described in [123]. We then passaged viral libraries in MDCK-SIAT1 cells. Cells were plated in D10 media at 2x10^5 cells/mL. After 16 hours, the media was changed to influenza growth media containing diluted transfection supernatant virus. PR/1934 libraries were each passaged in 20 wells of 6-well dishes at an MOI of 0.05 TCID50/cell, and Aichi/1968 libraries were each passaged in eight 10-cm dishes at an MOI of 0.1 TCID50/cell. After 48 hours, supernatant was clarified by centrifugation at 2,000xg for 5 min, aliquoted and stored at -80° C.

Sample preparation and deep sequencing. For each virus sample to be sequenced, 10 mL of clarified viral passage supernatant was centrifuged at 64,000xg for 1.5 hours to pellet viruses. RNA was extracted using the Qiagen RNEasy kit by lysing viral pellets in buffer RLT and following the manufacturer's recommended protocol. The NP gene was reverse transcribed using AccuScript High-Fidelity Reverse Transcriptase (Agilent Technologies) from both positive-sense and negative-sense viral RNA templates using the primers PR8-NP-RT-F (5'-agcaaaagcagggtagataatcactcactgagtgac-3') and PR8-NP-RT-R (5'-agtagaaacaagggtatttttcttta-3') for PR/1934 viruses or the primers 5'-BsmBI-Aichi68-NP (5'-catgatcgtctcagggagcaaaagcagggtagataatcactcacag-3') and 3'-BsmBI-Aichi68-NP (5'-catgatcgtctcgtattagtagaaacaagggtatttttcttta-3') for Aichi/1968 viruses.

To ensure a sufficiently large number of unique RNA molecules were reverse transcribed, we used qPCR (SYBR Green Real-Time PCR Master Mix, Life Technologies) using primers qWSN-NP-for (5'-ACGGCTGGTCTGACTCACAT-3') and qPR8-NP-rev (5'-TCCATTCCGGTGCGAACAAG-3') to quantify the concentration of first-strand cDNA molecules against a standard curve of linear NP amplicons quantified by Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies). We then made PCR amplicons with KOD DNA Polymerase (Merck Millipore) using at least 1x10^9 first-strand cDNA molecules as template in each reaction for viral gene sequencing. We also made PCR amplicons using 10 ng of the indicated plasmids for plasmid sequencing. For each biological replicate, we gener-

ated these PCR amplicons with 25 cycles of amplification using unmutated NP plasmid, mutated NP plasmid, NP cDNA from unmutated virus, and NP cDNA from mutated virus as template for the **DNA**, **mutDNA**, **virus**, and **mutvirus** samples, respectively.

To reduce the sequencing error rate, we developed a sequencing sample preparation protocol that results in sequencing libraries with inserts approximately 150 bp long. This allowed us to use paired-end 150 bp sequencing to achieve mostly overlapping reads so that sequencing errors resulting in mismatches between the two reads could be identified and ignored during data analysis. To make these sequencing libraries, we gel-purified the **DNA**, **mutDNA**, **virus**, and **mutvirus** PCR amplicons and sheared 1 μg of each amplicon using Covaris to a median size of approximately 150 bp. We followed the modified Illumina paired-end library preparation protocol provided in [65] for end repair, 3' A overhang, and adapter ligation steps, using Zymo DNA Clean & Concentrator columns (Zymo Research) or Ampure XP (Beckman Coulter) magnetic beads for DNA clean-up after shearing, end repair, and 3' A overhang steps. Barcoded Y-adapters were made by annealing 10 μL of 100 μM PAGE purified universal adapter (5'-AATGATACGGCGACCACCGAGATCTAC ACTCTTTCCCTACACGACGCTCTTCCGATC*T-3', where * indicates phosphorothioate bond) to 10 μL of 100 μM PAGE purified barcoded adapter (5'-PGATCGGAAGAGCACA CGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTT*G-3', where P indicates 5' phosphorylation, * indicates phosphorothioate bond, and NNNNNN indicates sample-specific barcode). Each 20 μL mixture (one mixture for each barcode sequence) was annealed by heating to 95° C for 5 minutes and cooling at 0.3° C/second to 4° C. The resulting Y-adapters were diluted to 25 μM by adding 20μL 10 mM Tris pH 7.5 and stored in 4 μL aliquots at -20° C. Y-adapters with unique barcodes (ATCACG, ACTTGA, TAGCTT, GGCTAC, TTAGGC, GATCAG, ACTGAT, CGTACG, CGATGT, TGACCA, CA-GATC, and CCGTCC) were ligated to samples derived from each biological replicate of each amplicon and ligation products were purified using 0.8X bead-to-sample ratio Ampure XP.

Purified adapter-ligated products for each sample were quantified by Quant-iT PicoGreen

dsDNA Assay Kit (Life Technologies) and 25 ng was used as template for a 4-cycle PCR using Phusion High-Fidelity Polymerase (Thermo Scientific) to amplify inserts with adapters properly ligated on both sides. This amplification step was performed with the following components: 25 ng template DNA, 5 μL 5X Phusion buffer, 2.5 μL mixture of each dNTP at 2.5 mM, 2 μL forward primer at 10 μM (5'-AATGATACGGCGACCACCGA GATCTACACTCTTTCCCTACACGA-3'), 2 μL reverse primer at 10 μM (5'-CAAGCAGAA GACGGCATACGAGAT-3'), and 0.25 μL Phusion polymerase in a final reaction volume of 25 μL. PCR products were purified using 1.0X bead-to-sample ratio Ampure XP and quantified using PicoGreen. Samples were pooled in equal amounts and size-selected on a 2.0% agarose gel for fragments between 240 bp and 300 bp, which contain sequencing inserts in the size range of 120-180 bp. The size-selected sample was then sequenced at the Fred Hutchinson Genomics Core on an Illumina HiSeq 2500 using a paired-end 150 bp sequencing strategy in rapid run mode.

Analysis of deep sequencing data. Sequencing data processing was performed using the software package `mapmuts` [15]. Briefly, for each replicate sample of **DNA**, **mutDNA**, **virus**, and **mutvirus**, paired reads were stripped of any adapter sequence and aligned to each other. Read pairs were discarded if any of the following criteria were met: less than 100 bp of overlap between reads, average Q-score less than 25 across either read, more than 5 ambiguous nucleotides (N nucleotides) in either read, or more than 1 mismatch in the overlap between reads. Retained read pairs were then aligned to the appropriate reference NP gene sequence for PR/1934 or Aichi/1968 NP, and read pairs with more than 10 mismatches to the reference sequence or with any gaps or insertions were discarded. Once aligned to the reference sequence, codon identities at every position were called only if all three nucleotides in the codon matched unambiguously in both reads. The total number of codon identities at every codon position in the coding region were totaled for each sample (**DNA**, **mutDNA**, **virus**, and **mutvirus**), separately for each biological replicate.

Inference of amino-acid preferences. We specify that at every site r in the protein, there is an inherent preference $\pi_{r,a}$ for every amino acid a, and we specify that $\sum_a \pi_{r,a} = 1$. The preference $\pi_{r,a}$ can be considered to be the expected frequency of amino acid a at site r in a mutant virus library after viral growth from a starting plasmid mutant library that contains equal numbers of every amino acid encoded at site r. Thus, mutations to amino acids with high preferences are beneficial and will be selected for during viral growth, and mutations to amino acids with low preferences will inhibit viral growth and will be selected against. Since the plasmid mutant libraries we generated contain on average more than one mutation per clone, the amino-acid preferences we measure represent an average preference in a variety of genetic backgrounds very similar to the starting sequence.

Let $\mathcal{A}(x)$ represent the amino acid encoded by codon x and let $\mathcal{C}$ represent the set of all codons. The effect of the preference $\pi_{r,\mathcal{A}(x)}$ on the frequency f of observing codon x at site r in the mutant virus library sample **mutvirus** is given by:

$$f_{r,x}^{\textbf{mutvirus}} = \epsilon_{r,x} + \rho_{r,x} + \frac{\mu_{r,x} \times \pi_{r,\mathcal{A}(x)}}{\sum_{y \in \mathcal{C}} \mu_{r,y} \times \pi_{r,\mathcal{A}(y)}} \tag{A.1}$$

where $\epsilon_{r,x}$ is the rate of PCR and sequencing errors at site r resulting in codon x, $\rho_{r,x}$ is the rate of reverse transcription errors at site r resulting in codon x, and $\mu_{r,x}$ is the frequency of codon x at site r in the plasmid mutant library **mutDNA**.

We inferred the amino-acid preferences independently for each biological replicate using the Bayesian algorithm described in [17] as implemented in `dms_tools` where codon counts in the **DNA**, **virus**, and **mutDNA** samples are used to infer the unknown parameters ϵ, ρ, and μ at each site.

Amino-acid preferences for Aichi/1968 NP were previously published in [15], where 8 biological replicates of the entire experiment were performed. In this work we report two additional biological replicates of the deep mutational scanning experiment for Aichi/1968. We will distinguish the two data sets when they are used separately for comparison as Aichi/1968 *previous study* and Aichi/1968 *current study*, and we will call the combined

dataset of all 10 biological replicates for this homolog *Aichi/1968*.

Comparison of site-specific amino-acid preferences between homologs

Quantifying the magnitude of amino-acid preference difference between homologs. At every site in the protein, each replicate deep mutational scanning experiment allows for the inference of an amino-acid preference distribution $\vec{\pi}$ that provides the preference at that site for all 20 amino acids. We used the Jensen-Shannon distance metric (the square root of the Jensen-Shannon divergence) to quantify the distance d between two amino-acid preference distributions:

$$d(\vec{\pi}_1, \vec{\pi}_2) = \sqrt{H\left(\frac{\vec{\pi}_1 + \vec{\pi}_2}{2}\right) - \frac{H(\vec{\pi}_1) + H(\vec{\pi}_2)}{2}} \tag{A.2}$$

where $H(\vec{\pi})$ is the Shannon entropy of the amino-acid preference distribution $\vec{\pi}$. The Jensen-Shannon distance metric quantifies the similarity between two amino-acid preference distributions, ranging from 0 (identical distributions) to 1 (completely dissimilar distributions). The average distance d between amino-acid preferences inferred from replicate experiments in the same homolog varies across sites. In other words, at some sites in the protein $\vec{\pi}$ is measured with greater precision than others. We therefore sought to develop, for every site r, a quantitative measure of the magnitude of change in $\vec{\pi}$ between homologs that corrects for the variation in $\vec{\pi}$ within replicate experiments of the same homolog.

For two groups of replicate mutational-scanning experiments A and B done in different homologs, each containing several replicate inferences of $\vec{\pi}$ for every site, we calculate the root-mean-square distance at site r over all pairwise comparisons of $\vec{\pi}$ measured in replicate experiments i (from group A) and j (from group B):

$$RMSD_{r,between} = \sqrt{\frac{1}{N_{A,B}} \sum_{i \in A} \sum_{j \in B} d(\vec{\pi}_{r,i}, \vec{\pi}_{r,j})^2} \tag{A.3}$$

where $N_{A,B}$ is the total number of non-redundant pairwise comparisons between replicate preferences measured from groups A and B. At the same site, to estimate the amount of experimental noise within replicates of the same homolog, we calculate the root-mean-square distance over all pairwise comparisons of $\vec{\pi}$ *within* the same group of replicate experiments, and average this site-specific noise estimate across the two groups:

$$RMSD_{r,within} = \frac{1}{2}\sqrt{\frac{1}{N_{A,A}}\sum_{i,j\in A,i<j} d(\vec{\pi}_{r,i},\vec{\pi}_{r,j})^2} + \frac{1}{2}\sqrt{\frac{1}{N_{B,B}}\sum_{i,j\in B,i<j} d(\vec{\pi}_{r,i},\vec{\pi}_{r,j})^2} \quad \text{(A.4)}$$

where $N_{A,A}$ and $N_{B,B}$ are the number of non-redundant pairwise comparisons between replicates within groups A and B, respectively. We then subtract the magnitude of the noise at this site observed *within* groups from our measurement of the difference in amino-acid preferences seen *between* groups to obtain a corrected value for the change in $\vec{\pi}$ at site r between homologs:

$$RMSD_{r,corrected} = RMSD_{r,between} - RMSD_{r,within} \quad \text{(A.5)}$$

It is possible that the observed variation within groups is greater than the observed variation between groups, resulting in negative $RMSD_{corrected}$.

Identifying sites with statistically significant changes in amino-acid preference. To determine whether site-specific $RMSD_{corrected}$ values are significantly larger than expected if amino-acid preferences are unchanged between homologs, we applied two methods to generate null distributions of $RMSD_{corrected}$ values. First, we used exact randomization testing to make all possible shuffles of the replicate homolog datasets into the two groups A and B. For each permutation, we calculated the $RMSD_{corrected}$ at every site, and the results are combined for all permutations. If there are no differences in preferences between homologs, the distribution of scores generated through randomization should be similar to the distribution of scores from the actual experiment.

We next observed that the overall correlation of amino-acid preferences across all sites between replicates can vary between experiments. For instance, the average Pearson's correlation between PR/1934 replicates is 0.59, the correlation between Aichi/1968 replicates in the *previous study* is 0.50, and the correlation between Aichi/1968 replicates in the *current study* is 0.74. We considered whether the varying precision between homologs might lead to biases in the calculated $RMSD_{corrected}$.

To test this, we generated a second null distribution of $RMSD_{corrected}$ under the hypothesis that the "true" amino-acid preferences are the same for both homologs and can be approximated by averaging the mean observed preferences for each homolog:

$$\langle\langle\vec{\pi}_r\rangle\rangle = \frac{\langle\vec{\pi}_{r,homolog\ A}\rangle + \langle\vec{\pi}_{r,homolog\ B}\rangle}{2} \tag{A.6}$$

Under this hypothesis, the observed differences in amino-acid preferences between homologs is solely due to the different amounts of experimental noise between replicates of each homolog. To model the effects of this noise on our analysis, we drew replicate simulated amino-acid preferences at each site r from a Dirichlet distribution with mean centered on the "true" amino-acid preferences:

$$\vec{\pi}_{r,simulated\ A} = Dir(\langle\langle\vec{\pi}_r\rangle\rangle \times \sigma_A) \tag{A.7}$$

where σ_A is a scaling factor that is chosen to yield simulated replicate preferences across the entire protein that have an average Pearson's correlation between replicates equal to the correlation between experimental replicates. In other words, we simulate replicate amino-acid preference measurements with noise tuned to match the actual noise in each experiment. For each simulated experiment, we simulated the same number of replicates that were performed experimentally, and calculated $RMSD_{corrected}$ for all sites. We ran the entire simulation 1000 times, combining all $RMSD_{corrected}$ values to obtain a null distribution.

We then separately used the two null distributions (generated through randomization or simulation) to assign p-values to site-specific $RMSD_{corrected}$ at each site r:

$$p_r = \frac{\text{number of scores in null distribution} \geq RMSD_{r,corrected}}{\text{number of scores in null distribution}} \tag{A.8}$$

To control the false discovery rate across the 497 sites tested for significance, we used the procedure of Benjamini and Hochberg [11].

Structural analysis of sites with preference changes. We used the crystal structure of the influenza A H1N1 WSN/1933 NP [PDB ID 2IQH, chain C; 138] to calculate distances between sites. Distances between sites were defined as the minimum distance between any side chain atoms distal to the alpha carbons of each site (the alpha carbon was used for all glycine residues). A distance cutoff of 4.5 Ångströms was used to define sites that are in contact with evolutionarily variable sites. To test for spatial clustering of a group of N sites, the distribution of N distances to the nearest neighbor of the remaining $N-1$ sites was compared to a null distribution of distances calculated the same way for 1000 random selections of sites of size N. One-sided P-values were computed using the Mann-Whitney U test.

Phylogenetic analysis

Experimental substitution model overview. We used a previously described approach to build site-specific substitution models for influenza nucleoprotein [15, 16]. Briefly, this approach calculates the codon-substitution rate at each site in nucleoprotein based on the rate at which nucleotide mutations arise and the level of selection acting on these new mutations. The rate of codon substitution, $P_{r,xy}$, at site r of codon x to a different codon y is described as,

$$P_{r,xy} = Q_{xy} \times F_{r,xy} \tag{A.9}$$

where Q_{xy} is the rate of mutation from x to y, and $F_{r,xy}$ is the probability that a mutation from x to y at site r is selected and reaches fixation. In this equation, the mutation rates Q_{xy} are assumed to be identical across sites whereas the selection is modeled as site-specific and site-independent. The site-specific fixation probabilities $F_{r,xy}$ were calculated from the experimentally measured amino-acid preferences using the relationship proposed by Halpern and Bruno [60, 16]. The four mutation rate free parameters and the stringency parameter were defined as in [16].

We then calculated the phylogenetic likelihood of the observed nucleoprotein sequences given the resulting experimental substitution model $P_{r,xy}$, the nucleoprotein phylogenetic tree, and the model parameters. The tree consisted of influenza nucleoproteins from either human, swine, equine, or avian hosts. While holding the tree topology fixed, tree branch lengths, and any other model parameters (discussed below), were optimized by maximum likelihood.

To compare overall phylogenetic likelihoods calculated under various substitution models, we calculated the difference in the Akaike Information Criteria (ΔAIC) between models. We compared site-specific models derived from experimentally determined amino-acid preferences to a non-site-specific model. We tested separate site-specific models using the amino-acid preferences from PR/1934 and Aichi/1968. The Aichi/1968 preferences were an average of the amino-acid preferences from the *current study* and *previous study*. In addition, we tested a site-specific model where we combined data from the separate Aichi/1968 and PR/1934 mutational-scanning experiments, by averaging amino-acid preferences for each amino acid at each site across the two homologs, weighting each homolog equally.

The non-site-specific model used the Goldman-Yang (GY94) codon substitution model [55], with nucleotide equilibrium frequencies calculated by the CF3x4 method [104]. In this model, the transition-transversion ratio was optimized by maximum likelihood, along with the mean and shape parameters describing gamma distributions of the nonsynonymous-synonymous ratios [137] and the substitution rates [136] across sites. Each gamma distri-

bution was discretized with four categories. In previous comparisons of non-site-specific models, this non-site-specific model performed better than other variants of the GY94 model [15, 16]. All analyses were performed using the software packages `phyloExpCM` [15] and `HyPhy` [105], and the data, scripts, and descriptions to replicate the results in this article are available at `https://github.com/mbdoud/Compare-NP-Preferences`.

Phylogenetic trees for different influenza hosts. We built phylogenetic trees for nucleoprotein coding sequences from strains of human influenza, swine influenza, equine influenza, and avian influenza. Full-length nucleoprotein sequences were downloaded from the Influenza Virus Resource [7], and for each host, a small number of unique sequences per year per influenza subtype were retained. For human influenza, we retained one sequence every other year from each of the H1N1, H2N2, and H3N2 lineages. For swine influenza, we retained one sequence per year from either the North American Classical H1N1 lineage or the Eurasian H1N1 lineage. For equine influenza, we retained one sequence per year from the H3N8 lineage. For avian influenza, one sequence every other year per subtype was retained, and the examined hosts were further restricted to only duck species, to make a sequence set with a size manageable for phylogenetic modeling.

Sequences from each host were aligned by EMBOSS `needle` [109], and maximum-likelihood trees were built by `RAxML` [120]. Using these trees and the program `Path-O-Gen` (http://tree.bio.ed.ac.uk/software/pathogen/), we identified and removed any sequences that were noticeable outliers from the molecular clock. The final tree contained 37, 46, 29, and 24 sequences from human, swine, equine, and avian hosts respectively.

Maximum-likelihood phylogenetic trees were then built from the nucleoprotein sequence alignment using `codonPhyML` [54]. The GY94 model [55] was run using the CF3x4 nucleotide equilibrium frequencies [104] along with maximum-likelihood optimization of a transition-transversion ratio and of a mean and shape parameter describing a gamma distribution of nonsynonymous-synonymous ratios [137]. This gamma distribution was discretized with four categories. The final, unrooted tree was visualized with `FigTree` (htt

p://tree.bio.ed.ac.uk/software/figtree/) and rooted using the avian clade [131].

Acknowledgements

We thank Hugh Haddox, Alistair Russell, and Heather Machkovech for critical reading of the manuscript and Trevor Bedford for helpful discussions about statistical analysis. We thank the Summer Institute in Statistics and Modeling in Infectious Diseases at the University of Washington for helpful instruction and the Genomics Shared Resource at the Fred Hutchinson Cancer Research Center for performing high-throughput sequencing. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (grant number R01 GM102198). M.D. was supported by NIH Training Grant T32 AI083203 and a fellowship from the Seattle Chapter of the Achievement Rewards for College Scientists Foundation. O.A. was supported by a PhRMA Postdoctoral Fellowship in Informatics.

Appendix B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

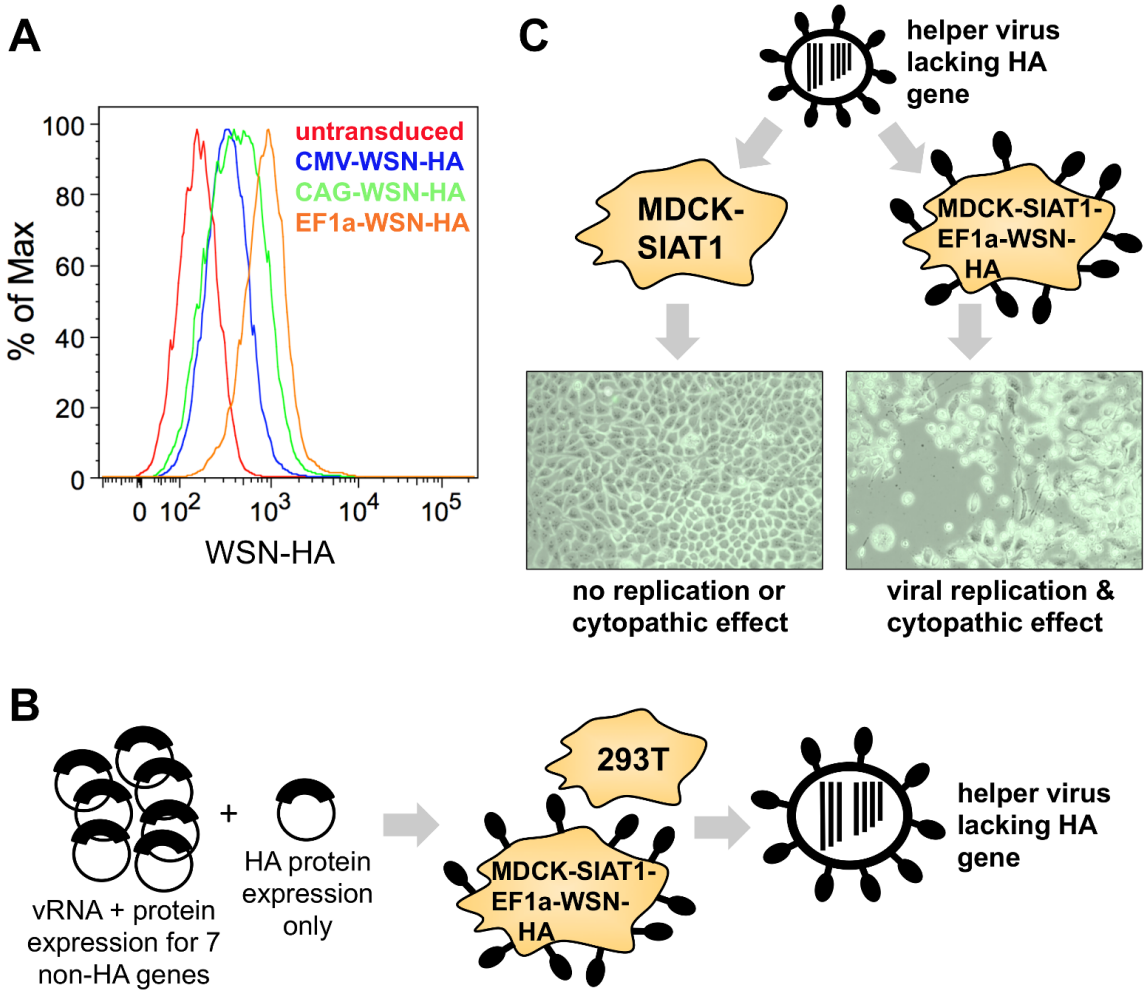

Figure B.1: **An HA-deficient helper virus can replicate in cells constitutively expressing HA protein. (A)** We engineered MDCK-SIAT1 cells by lentiviral transduction to constitutively express the HA protein of the A/WSN/1933 strain under control of the EF1a promoter (MDCK-SIAT1-EF1a-WSN-HA cells), which provided higher expression levels than the CMV or CAG promoters. Transduced and untransduced cells were stained with a 1:100 dilution of mouse polyclonal anti-WSN serum, followed by a 1:100 dilution of APC-conjugated anti-mouse IgG for secondary staining. **(B)** We transfected a co-culture of these MDCK-SIAT1-EF1a-WSN-HA and 293T cells with bidirectional reverse-genetics plasmids [69] for the seven non-HA segments of A/WSN/1933 plus a protein expression plasmid for HA. **(C)** The resulting transfection supernatant contained HA-deficient helper virus that could be propagated in MDCK-SIAT1-EF1a-WSN-HA cells but *not* in standard MDCK-SIAT1 cells. This virus typically reached titers of $\sim 10^3$ $TCID_{50}$ per μl when titered on the MDCK-SIAT1-EF1a-WSN-HA cells. This titer was about 3-fold lower than that obtained if we included an HA GFP segment similar to that described in Marsh *et al.* [89] that contains GFP flanked by the noncoding and 80 coding nucleotides. This difference in titer between an HA-deficient and HA-GFP virus is comparable to that reported by Marsh *et al.* [89].

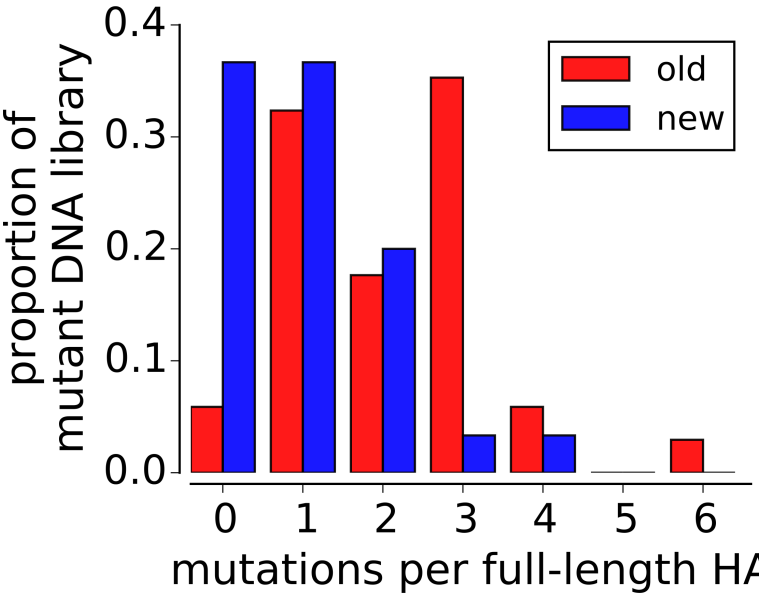

Figure B.2: **The mutant plasmid DNA library used in this study ("new") has a lower mutation rate than the library used by Thyagarajan and Bloom [123] ("old").** The old library was generated using two rounds of codon mutagenesis, leading to an average of two mutations per HA; the new library used only one round of mutagenesis, resulting in an average of one mutation per HA. At least 30 clones of each library were Sanger sequenced across the entire gene.

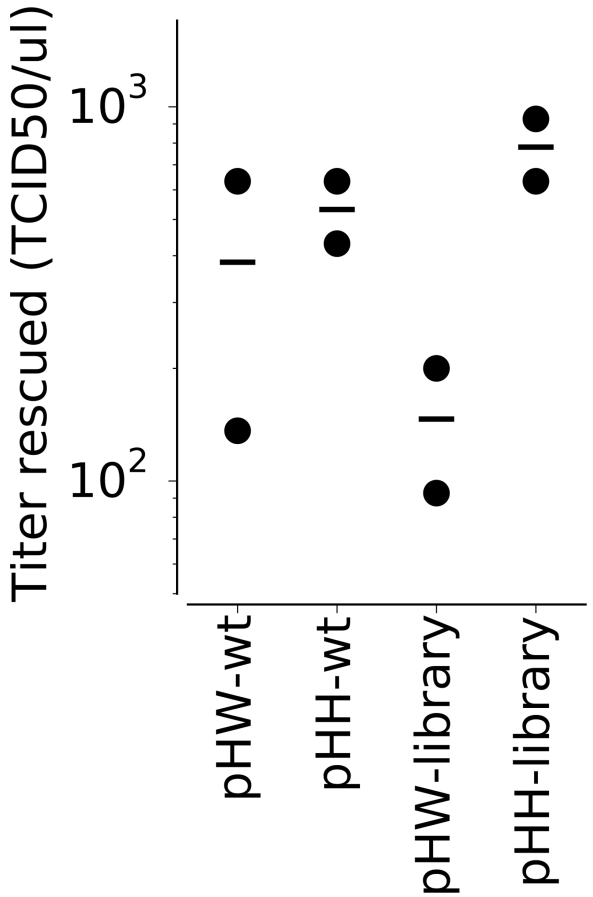

Figure B.3: **Mutant virus library generation is more efficient when HA is encoded on the pHH21 plasmid.** The pHH21 plasmid contains a single RNA polymerase I promoter for transcription of negative sense HA vRNA [94]; this vRNA molecule must then associate with the viral polymerase complex for mRNA transcription and protein expression for proper virus generation. The pHW2000 plasmid is similar to pHH21, but also contains an RNA polymerase II promoter [69] for the transcription of both negative sense HA vRNA and positive sense HA mRNA directly off the plasmid, so that HA protein expression is not limited by the transcription of mRNA by the viral polymerase. Cells were transfected with the indicated plasmid containing wild-type or mutant library HA along with protein expression plasmids for the viral polymerase-related proteins (PB2, PB1, PA, and NP). After 24 hours, cells were infected with helper virus, and 24 hours after infection, cell supernatants were titered in MDCK-SIAT1 cells to quantify the amount of virus generated. We hypothesize that the lower titer when using the pHW plasmid is due to expression of more HA mutants per cell, some of which might act as dominant negatives. Each virus generation was performed in duplicate, with a bar marking the mean of the two experiments.

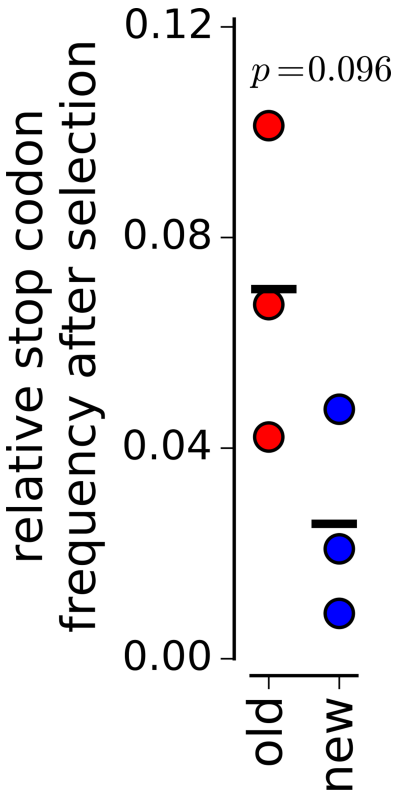

Figure B.4: **Purging of stop codons is more complete in our new experiments than in the previous experiments.** Each point shows the relative fraction of stop codons remaining after selection for one of the three replicates. The stop codon frequencies in the wild-type plasmid and virus samples are subtracted from the mutant plasmid and mutant virus samples to correct for errors arising during sample preparation and sequencing. "Old" refers to libraries from Thyagarajan and Bloom [123]; "new" refers to libraries in the current study. We hypothesize that the stronger selection against stop codons in the new experiments is the result of better genotype-phenotype linkage imposed by the the lower MOI used for viral passage, which leads to more effective selection on each mutation. Bars show the means for each set of libraries; the p-value was calculated with a two-sided T-test.

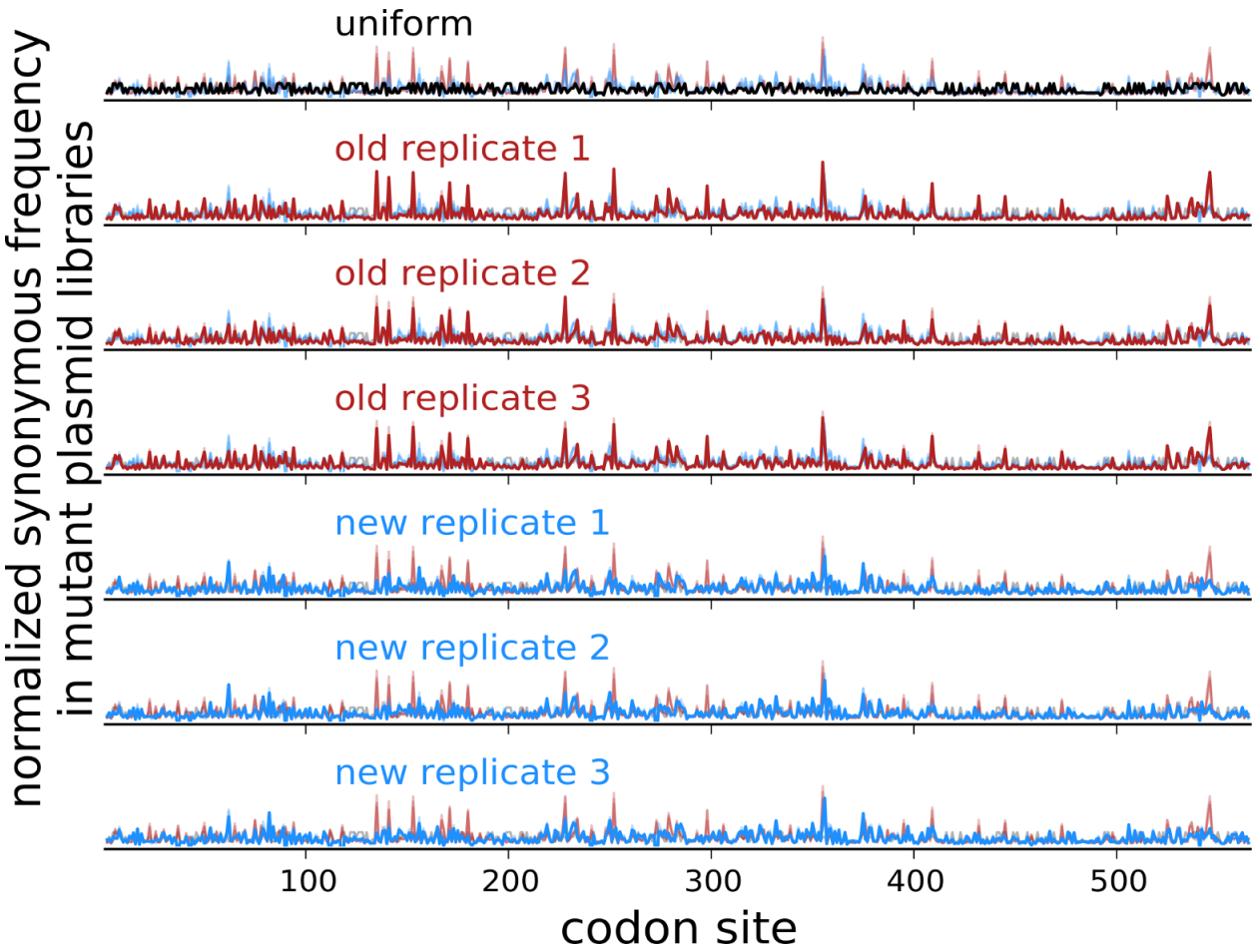

Figure B.5: **Synonymous frequency peaks observed in bottlenecked virus libraries are not due to the composition of plasmid mutant libraries.** Shown for each replicate is the normalized synonymous mutation frequency for plasmid mutant libraries in the same form as in Figure 3.2. The frequency of synonymous mutations in the plasmid mutant libraries is highly reproducible among replicates. Although there are some sites with peaked frequencies (likely due to PCR biases during mutagenesis), these sites are consistent across replicates and do not correspond to the peaks in the mutant virus libraries shown in Figure 3.2.

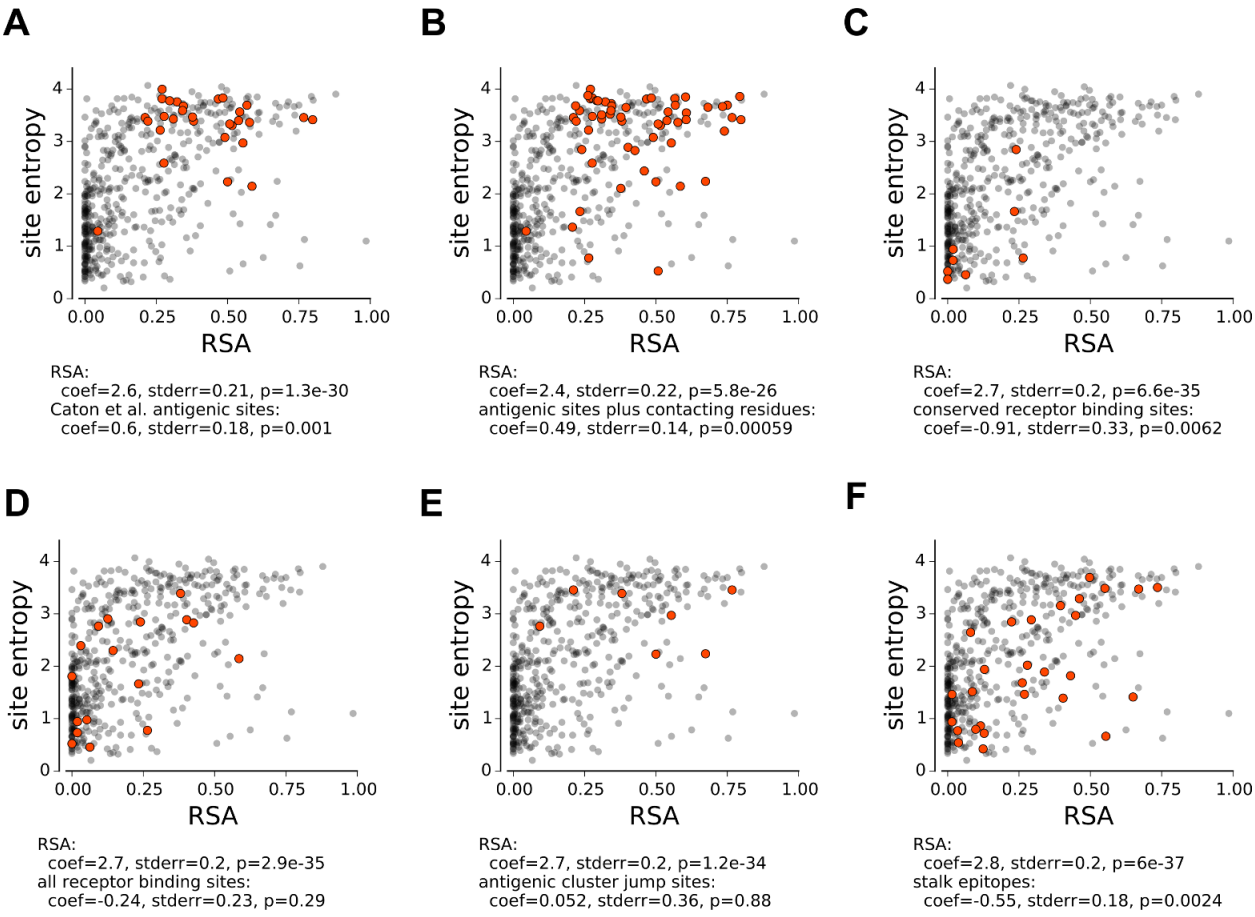

Figure B.6: **Statistical analyses of whether sets of sites have higher or lower mutational tolerance than expected given their solvent accessibility.** For each panel, a group of sites is selected and the results are shown for multiple linear regression of site entropy as a function of relative solvent accessibility (calculated from PDB structure 1RVX [51]) and whether or not the site belongs to that group. **(A)** Antigenic sites defined by Caton *et al.* [23] and **(B)** these sites plus their contacts have significantly higher mutational tolerance than expected from their solvent accessibility. **(C)** Conserved receptor binding sites have significantly lower mutational tolerance. **(D)** All sites contacting receptor have typical mutational tolerance. **(E)** Antigenic cluster jump sites defined by Koel *et al.* [75] have typical mutational tolerance. **(F)** Sites in the overlapping footprints of broadly neutralizing antibodies F10, CR6261, FI6v3, and CR9114 [121, 44, 31, 42] have significantly lower mutational tolerance.

Materials and Methods

Availability of data and computer code

Sequencing data are available from the Sequence Read Archive under accession numbers SRR3113656 (mutant DNA library 1), SRR3113657 (mutant DNA library 2), SRR3113658 (mutant DNA library 3), SRR3113660 (mutant virus library 1), SRR3113661 (mutant virus library 2), SRR3113662 (mutant virus library 3), SRR3113655 (wild-type DNA control), and SRR3113659 (wild-type virus control). The computer code necessary to reproduce all the analysis in this work is available at `https://github.com/mbdoud/2016_WSN_HA_analysis`.

Growth of HA-deficient helper virus in HA-expressing cells

MDCK-SIAT1 cells (Sigma, 05071502) were engineered to constitutively express the HA protein of A/WSN/1933 (H1N1) under control of the EF1a promoter by lentiviral transduction. These newly created cells will be referred to as MDCK-SIAT1-EF1a-WSN-HA cells since they are MDCK-SIAT1 cells that we have engineered to express the WSN HA under an EF1a promoter. HA surface expression was validated by flow cytometry (Figure B.1).

To generate HA-deficient helper viruses, we seeded co-cultures of 293T cells (obtained from the ATCC, number CRL-3216; seeded at 5×10^5 cells per well) and MDCK-SIAT1-EF1a-WSN-HA cells (5×10^4 cells cells per well) in 6-well dishes in D10 media (DMEM supplemented with 10% heat-inactivated FBS, 2 mM L-glutamine, 100 U of penicillin/mL, and 100 μg of streptomycin/mL). After 24 h, we transfected these co-cultures with bidirectional reverse-genetics plasmids for the seven non-HA segments of the A/WSN/1933 virus (pHW181-PB2, pHW182-PB1, pHW183-PA, pHW185-NP, pHW186-NA, pHW187-M, and pHW188-NS) [69] plus a protein expression plasmid for WSN HA (pHAGE2-CMV-WSNHA, which importantly does *not* contain non-coding regions of the HA segment or a promoter for the transcription of negative-sense viral RNA). Transfection was performed

with BioT transfection reagent (Bioland B01-02, Paramount, CA, USA) with each well receiving 250 ng of each plasmid. Twenty-two hours after transfection, we changed the media to WSN growth media (Opti-MEM supplemented with 0.5% heat-inactivated FBS, 0.3% BSA, 100 U of penicillin/mL, 100 μg of streptomycin/mL, and 100 μg of calcium chloride/mL). At 96 h post-transfection, we passed 400 μL of the transfection supernatant into 15-cm dishes containing 4×10^6 MDCK-SIAT1 cells (as a negative control) or MDCK-SIAT1-EF1a-WSN-HA cells in WSN growth media. HA-deficient helper virus could only be propagated in the HA-expressing cells as expected (Figure B.1). We collected the expanded helper virus from these cells after 68 h, aliquoted, and froze aliquots at -80 °C. We titered the helper virus in MDCK-SIAT1-EF1a-WSN-HA cells by $TCID_{50}$. We obtained titers between 10^3 and 10^4 $TCID_{50}$ per μL when titering in MDCK-SIAT1-EF1a-WSN-HA cells, and no cytopathic effect except with extremely concentrated helper virus in MDCK-SIAT1 cells (Figure B.1).

HA plasmid mutant libraries

Codon mutagenesis was performed as described in [123] except that we performed one overall round of the PCR mutagenesis to yield a lower mutation rate (Figure B.2). Ligation and eletroporation were also performed as in [123], except that we cloned the inserts into both pHW2000 [69] and pHH21 [94] plasmid backbones. All steps were performed in triplicate. For each replicate, we pooled over 3 million transformants, cultured in LB for 3 h in shaking flasks at 37 °C, and maxi-prepped plasmid libraries.

Generation of mutant HA virus libraries from mutant plasmids and helper viruses

To generate mutant virus libraries, we transfected 293T cells with a DNA mixture containing one of the three pHH21-MutantHA libraries (or the wild-type pHH21-WSN-HA control) and protein expression plasmids for the four proteins that compose the ribonucleoprotein complex, using plasmids HDM-Nan95-PA, HDM-Nan95-PB1, HDM-Nan95-PB2, and

HDM-Aichi68-NP [56]. Specifically, we plated 293T cells in D10 at a density of 8×10^5 per well in 6-well plates, changed the media to fresh D10 after 16 h, and then four hours later transfected cells with 500 ng of the HA reverse-genetics plasmid plus 375 ng of each of the PA, PB1, PB2, and NP plasmids using BioT. Twenty-four hours after transfection, we infected the cells with HA-deficient helper virus by making an inoculum of 1.3×10^3 TCID_{50} per μL in WSN growth media, aspirating the D10 media from the cells, and adding 2 mL of inoculum to each well. After 3 h, we removed the inoculum by aspiration and added 2 mL of WSN growth media supplemented with 5% D10. Twenty-four hours after helper virus infection, we collected the supernatants for each replicate, stored aliquots at -80 °C, and titered in MDCK-SIAT1 cells. Of note, we found that helper viruses that had been passaged more than once in MDCK-SIAT1-EF1a-WSN-HA cells tended to become less effective at rescuing fully replication competent viruses following infection of transfected cells, so we exclusively used single-passage helper virus in these experiments.

We passaged these transfection supernatants to create a genotype-phenotype link and impose functional selection on HA. We passaged over 9×10^5 TCID_{50} at an MOI of 0.0075 TCID_{50} per cell. Specifically, for each library, we plated ten 15-cm dishes with 6×10^6 MDCK-SIAT1 cells per dish and allowed cells to grow for 20 h, at which point they had reached a density $\sim 1.25 \times 10^7$ cells per dish. We then replaced the media in each dish with 25 mL of WSN growth media in each dish containing 3.7 TCID_{50} of virus per μL. We allowed virus replication to proceed for 40 h before collecting viruses from the supernatant for sequencing.

Barcoded subamplicon sequencing

For each of the three replicate HA virus libraries and the wild-type HA virus, we extracted viral RNA by ultracentrifuging 24 mL of supernatant at 22,000 rpm in a Beckman Coulter SW28 rotor. RNA was extracted using the Qiagen RNeasy kit by resuspending the viral pellet in 400 μL buffer of Qiagen RLT freshly supplemented with β-mercaptoethanol,

pipetting 30 times, transferring to an RNase-free microcentrifugefuge tube, adding 600 μL freshly-made 70% ethanol, and continuing with the manufacturer's recommended protocol, eluting the final RNA product in 40 μL of RNase-free water. HA was then reverse transcribed using AccuScript Reverse Transcriptase (Agilent 200820) with the primers WSNHA-For (5'-AGCAAAAGCAGGGGAAAATAAAAACAAC-3') and WSNHA-Rev (5'-AG TAGAAACAAGGGTGTTTTTCCTTATATTTCTG-3').

We generated PCR amplicons of HA for each of the eight samples (three replicate plasmid DNA libraries, three corresponding virus libraries, one wild-type plasmid DNA, and one wild-type virus) using KOD Hot Start Master Mix (71842, EMD Millipore) with the PCR reaction mixture and cycling conditions described in [15] and the primers WSNHA-For and WSNHA-Rev. The templates for these reactions were 2 μL of cDNA (for the virus-derived samples) or 2 μL of plasmid DNA at 10 ng/μL. To ensure that the number of molecules used as template did not bottleneck diversity, parallel PCR reactions were run with a standard curve of template molecules, and all products were analyzed by band intensity after agarose gel electrophoresis; all samples used $\geq 10^6$ molecules as a template for PCR. We purified these PCR amplicons using Agencourt AMPure XP beads (bead-to-sample ratio 0.9) (Beckman Coulter).

These PCR amplicons were quantified using Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies) and used as the templates for the barcoded-subamplicon sequencing in Figure 3.1B. We performed the first round of PCR ("PCR 1") in six parallel reactions (one for each of the six HA subamplicons) for each of the eight samples. Each reaction contained 12 μL 2X KOD Hot Start Master Mix, 2 μL forward primer diluted to 5 μM, 2 μL reverse primer diluted to 5 μM, and 8 μL purified amplicon diluted to 0.5 ng/μL (primer sequences for PCR 1 and PCR 2 are shown below). In addition to containing sequences targeting regions in HA, the forward and reverse primers for PCR 1 each contain an 8-base degenerate barcode and partial Illumina sequencing adaptors. To limit the generation of PCR artifacts, we performed only 9 cycles of PCR for PCR 1 using the following program: 1. 95 °C for 2:00; 2. 95 °C for 0:20; 3. 70 °C for 0:01; 4. 54 °C for 0:20; 5. 70 °C for

0:20; 6. Go to 2 (8 times); 7. 95 °C for 1:00; and 8. 4 °C hold. The denaturation step after cycling ensures that identical barcode pairs are not annealed at the end, so that most double-stranded molecules entering PCR 2 will contain two unique barcoded mutants. PCR 1 products were purified by Ampure XP (bead-to-sample ratio 1.0), quantified with Quant-iT PicoGreen, and diluted to 0.5 ng/μL.

We then mixed all six subamplicons from each experimental sample at equal concentrations and diluted these subamplicon pools such that the number of template molecules used in PCR 2 was less than the anticipated sequencing depth to ensure multiple reads per barcode. Specifically, we reduced the total amount of DNA for each experimental sample used as template in PCR 2 to 9.24×10^{-4} ng, which corresponds to 1.54×10^{-4} ng of each of the six subamplicons, corresponding to approximately 3.5×10^5 double-stranded DNA molecules (or 7×10^5 uniquely-barcoded single-stranded variants) per subamplicon per sample.

We performed PCR 2 for each sample with the following reaction conditions: 20 uL 2X KOD Hot Start Master Mix, 4 μL forward primer UniversalRnd2for diluted to 5 μM, 4 μL reverse primer indexXXRnd2rev diluted to 5 μM (a different index for each experimental sample), and 9.24×10^{-4} ng of the subamplicon pool of PCR 1 products described above, for a total volume of 40 μL. We used the following thermal cycling program: 1. 95 °C for 2:00; 2. 95 °C for 0:20; 3. 70 °C for 0:01; 4. 55 °C for 0:20; 5. 70 °C for 0:20; 6. Go to 2 (23 times); and 7. 4 °C hold. PCR 2 products were purified by Ampure XP (bead-to-sample ratio 1.0), quantified with Quant-iT PicoGreen, and equal amounts of each experimental sample were mixed and purified by agarose gel electrophoresis, excising the predominant DNA species at the expected size of approximately 470 bp. Sequencing was performed on one lane of a flow cell of an Illumina HiSeq 2500 using 2 $\times$ 250 bp paired-end reads in rapid-run mode.

Primer sequences for PCR 1 and PCR 2

PCR 1 primers: Each of 6 subamplicons is generated with a F/R primer pair. Lowercase sequence binds to HA, uppercase sequence is partial Illumina adaptor.

- amp1F (CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNaagcaggggaaaataaaaacaaccaaa)
- amp1R (GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNcattctcagagtttggtgtttctacaat)
- amp2F (CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNtccagcgagatcatggtcctac)
- amp2R (GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNgggtgatgaacaccccatagtac)
- amp3F (CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNtgtgaacaataaagggaaagaagtcctt)
- amp3R (GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNgtgttacactcatgcattgacgc)
- amp4F (CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNgtccggcatcatcacctcaaac)
- amp4R (GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNgttaatggcattttgtgtgctttttg)
- amp5F (CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNgatcaggctatgcagcggat)
- amp5R (GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNgaactcaaaacacccatttccgat)
- amp6F (CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNaaaaagccaattaaagaataatgccaaagaa)
- amp6R (GGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNgggtgtttttccttatatttctgaaatcctaatc)

PCR 2 primers: Each experimental sample is associated with one of the reverse primer index sequences (lowercase nucleotides) for multiplexing all experimental samples to a single flowcell.

- UniversalRnd2for (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC)
- index01Rnd2rev (CAAGCAGAAGACGGCATACGAGATacatcgGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT)
- index03Rnd2rev (CAAGCAGAAGACGGCATACGAGATcactgtGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT)
- index08Rnd2rev (CAAGCAGAAGACGGCATACGAGATgcctaaGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT)
- index09Rnd2rev (CAAGCAGAAGACGGCATACGAGATtcaagtGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT)
- index10Rnd2rev (CAAGCAGAAGACGGCATACGAGATctgatcGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT)
- index11Rnd2rev (CAAGCAGAAGACGGCATACGAGATaagctaGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT)
- index22Rnd2rev (CAAGCAGAAGACGGCATACGAGATcgtacgGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT)
- index25Rnd2rev (CAAGCAGAAGACGGCATACGAGATatcagtGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT)

Inference of amino-acid preferences from sequencing data

We used `dms_tools` (`http://jbloomlab.github.io/dms_tools/`), version 1.1.12, to align subamplicon reads to a reference HA sequence, group barcodes to build consensus sequences, quantify mutation counts at every site in the gene for each experimental sample,

and infer site-specific amino-acid preferences based on mutation frequencies pre- and post-selection using the algorithm described in [17].

Phylogenetic modeling using amino-acid preferences

We sub-sampled human and swine H1 sequences (1 sequence per host per year) from the set of sequences from [123], removed identical sequences, and built a sequence alignment. We then used `phydms` version 1.1.0 [14] (`http://jbloomlab.github.io/phydms/`), which in turn uses `Bio++` [58] for the likelihood calculations, to compare experimentally informed codon substitution models and other non-site-specific substitution models.

Statistical tests

Multiple linear regression of the continuous dependent variable of site entropy as a function of the continuous independent variable of relative solvent accessibility and a binary indicator of a site belonging to a specific classification (e.g., "antigenic sites") was performed with the same classifications as described in [123]. Additional classifications were obtained from [75] for sites responsible for antigenic cluster transitions in H3N2 and seasonal H1N1 (sites 158, 168, 169, 171, 172, 202, and 206 in sequential WSN H1 numbering starting with the initiating methionine), and the sites within antibody footprints of broadly-neutralizing antibodies F10, CR6261, FI6v3, and CR9114 (sites 25, 45, 46, 47, 48, 49, 305, 306, 307, 332, 361, 362, 363, 364, 379, 381, 382, 384, 385, 386, 388, 389, 391, 392, 395, 396, 399, and 400 in sequential WSN H1 numbering starting with the initiating methionine) [121, 44, 31, 42]. Definition of the protein domains within HA were from [51] (HA1 fusion domain: 18–72, 291–340; HA1 vestigial esterase domain: 73–125, 279–290; HA1 receptor binding domain: 126–278; HA2 fusion domain: 344–503; and all sites in sequential H1 numbering starting with the initiating methionine).

Acknowledgements

We thank Bargavi Thyagarajan for performing the PCR mutagenesis of the HA gene. We thank Anice Lowen for discussions that helped inspire the idea of using a helper virus to generate the mutant virus libraries. This work was supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) under grant R01 GM102198. M.B.D. was supported in part by a fellowship from the Seattle Chapter of the Achievement Rewards for College Scientists Foundation.

Appendix C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Table C.1: **Percentage of each mutant virus library remaining infectious after antibody neutralization in each replicate selection experiment.** Percent infectivity was measured by qRT-PCR of the influenza nucleoprotein gene and interpolated from a standard curve of infection prepared with serial dilutions of each virus library.

	library 1	library 1 replicate	library 2	library 3
H17-L19 0.5 μg/ml	2.1%	2.5%	4.9%	5.4%
H17-L19 1 μg/ml	0.7%	0.7%	1.9%	1.8%
H17-L19 10 μg/ml	0.3%	0.2%	0.4%	0.4%
H17-L10 3 μg/ml	0.2%		0.2%	0.2%
H17-L7 15 μg/ml	0.2%		0.1%	0.1%
H18-S415 3 μg/ml	0.2%		0.1%	0.2%

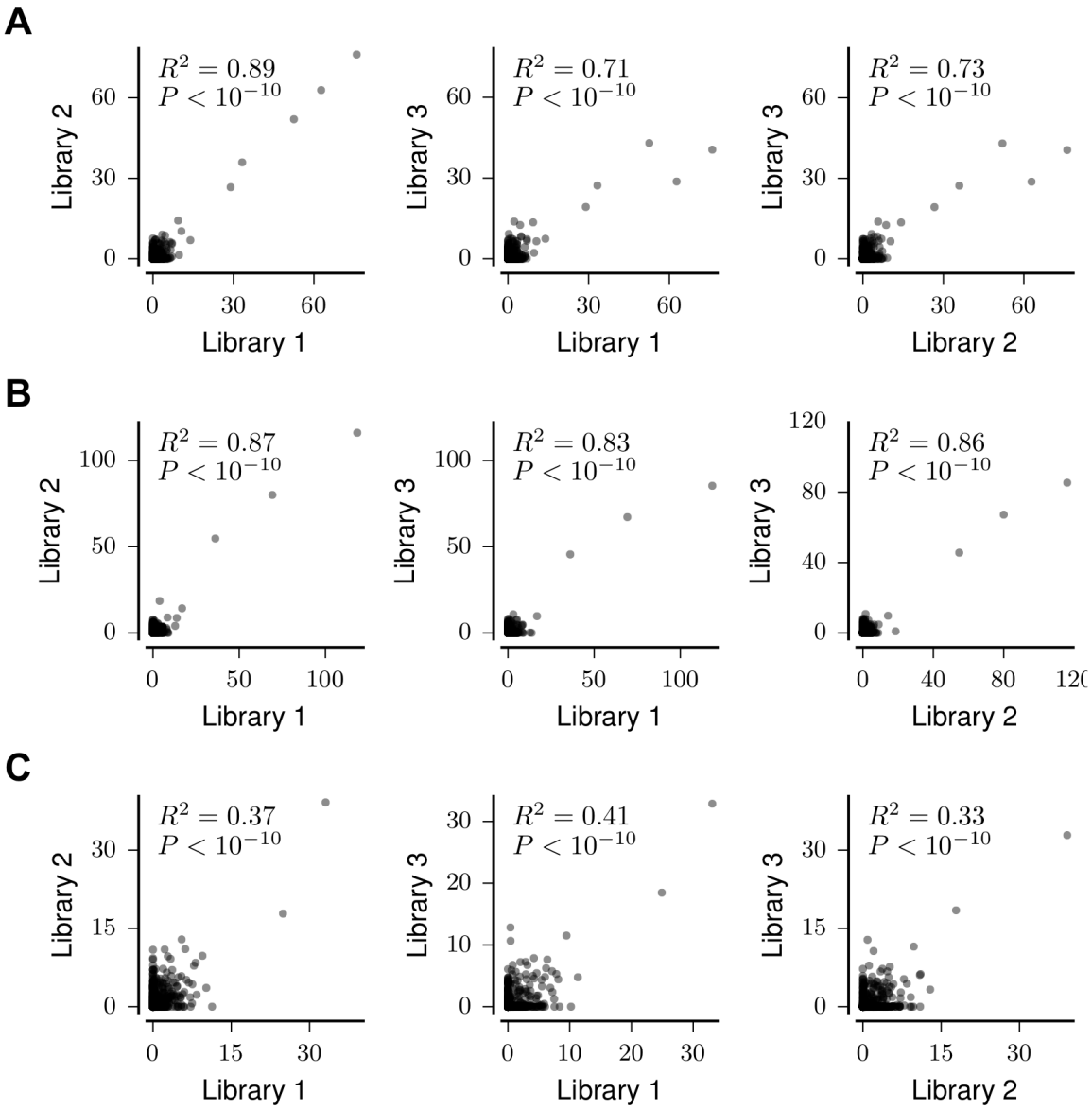

Figure C.1: **Positive site differential selection is highly correlated between full biological replicate measurements on independently generated mutant virus libraries.** Shown are correlations for antibodies **(A)** H17-L10, **(B)** H17-L7, and **(C)** H18-S415. Correlation coefficients are Pearson's R.

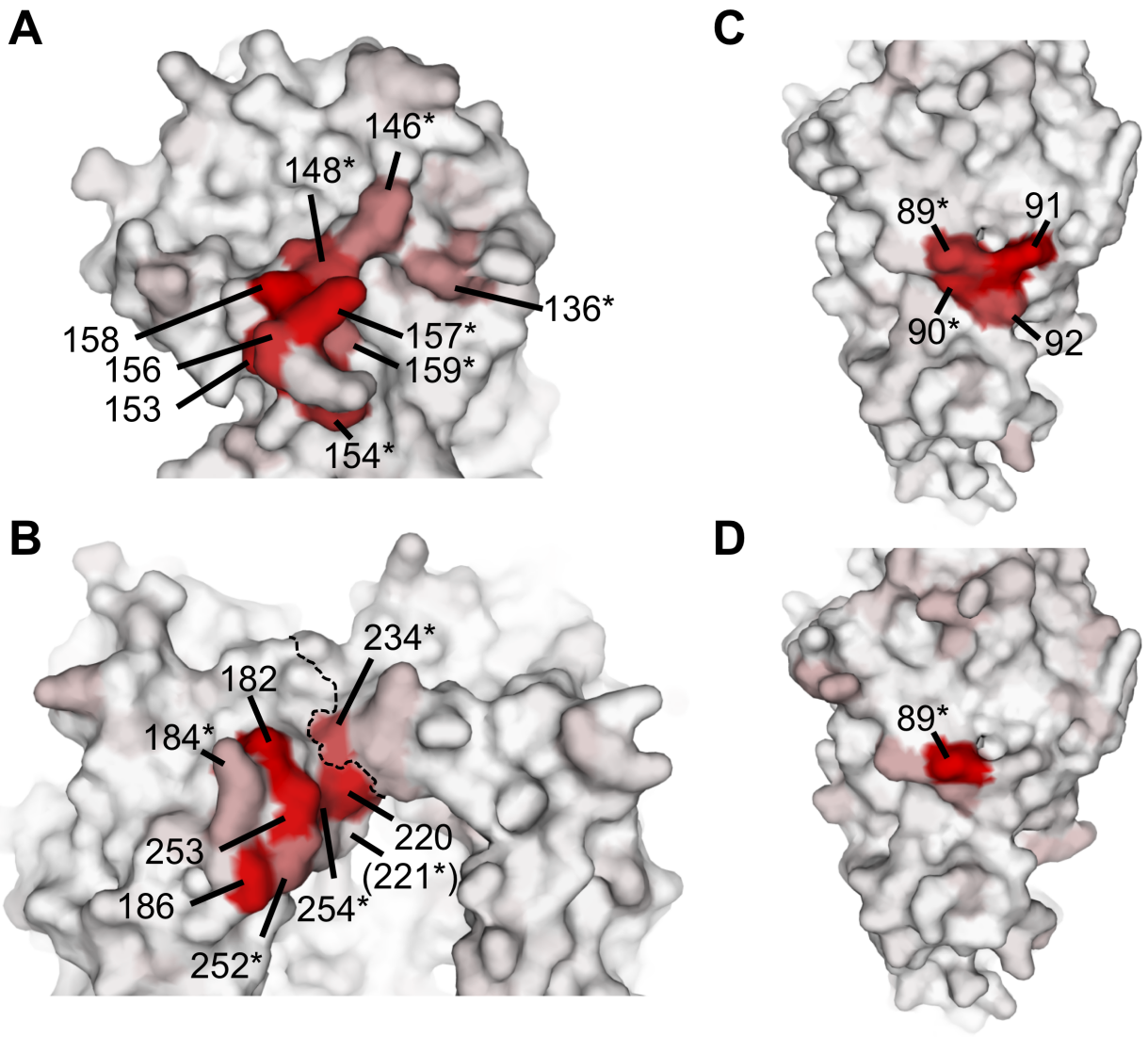

Figure C.2: **Detailed view of differential selection by each antibody projected onto HA's structure.** Each panel zooms into the relevant region of the structure shown in Figure 4.4C for that antibody. Residues are colored from white to red based on the differential selection for the most strongly selected mutation at that site for each antibody. Asterisks mark sites of strong differential selection which were not found in the original antigenic mapping of HA with that antibody [53, 23]. **(A)** H17-L19. **(B)** H17-L10. Strong differential selection at site 223 (not visible) results in putative glycosylation at site 221. The dashed line marks the boundary between two adjacent HA protomers. **(C)** H17-L7. **(D)** H18-S415.

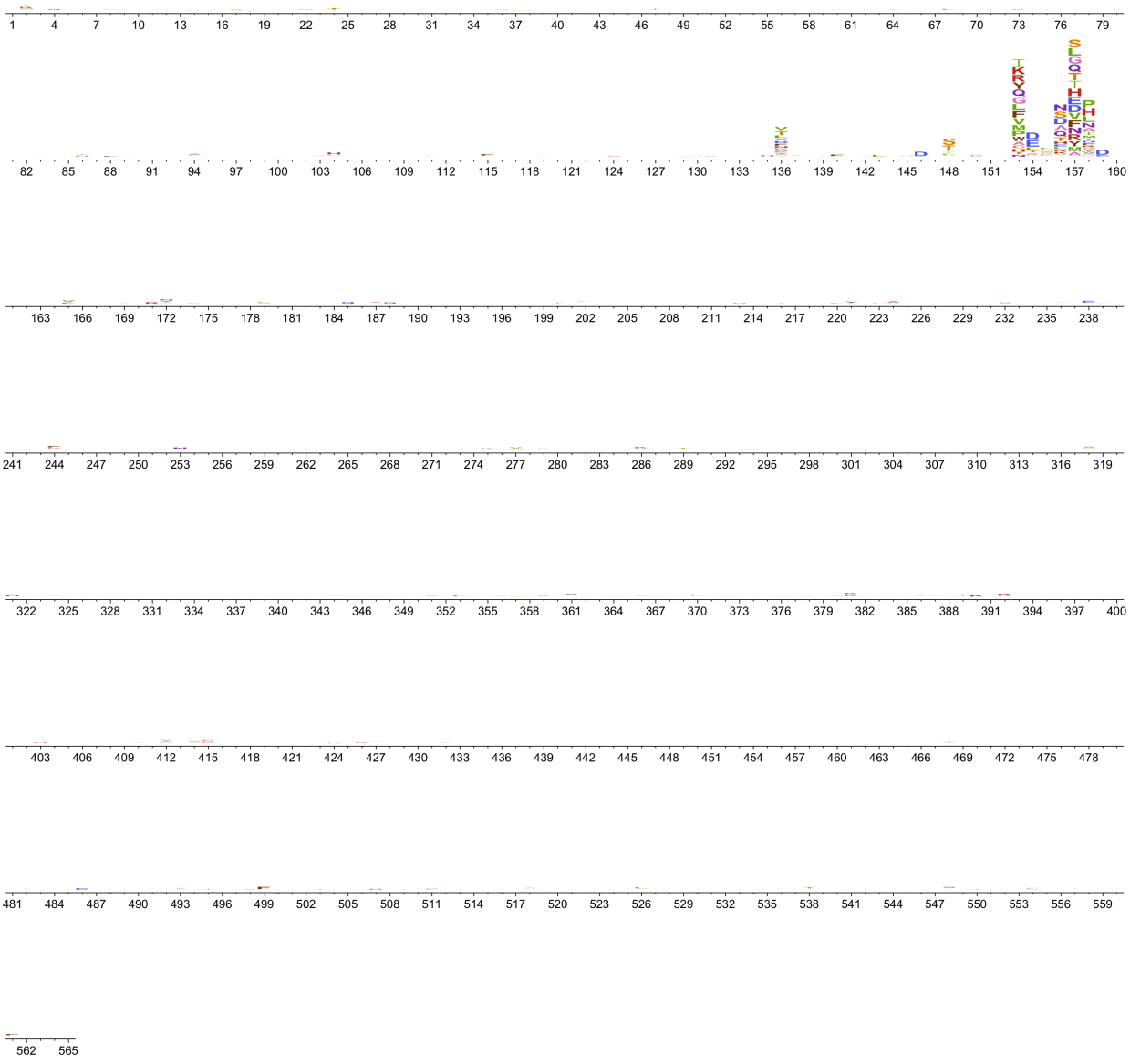

Figure C.3: **A logo plot showing the differential selection across all of HA from antibody H17-L19 at the concentration used in Figure 4.4.** These data are the average across the replicate libraries for each antibody.

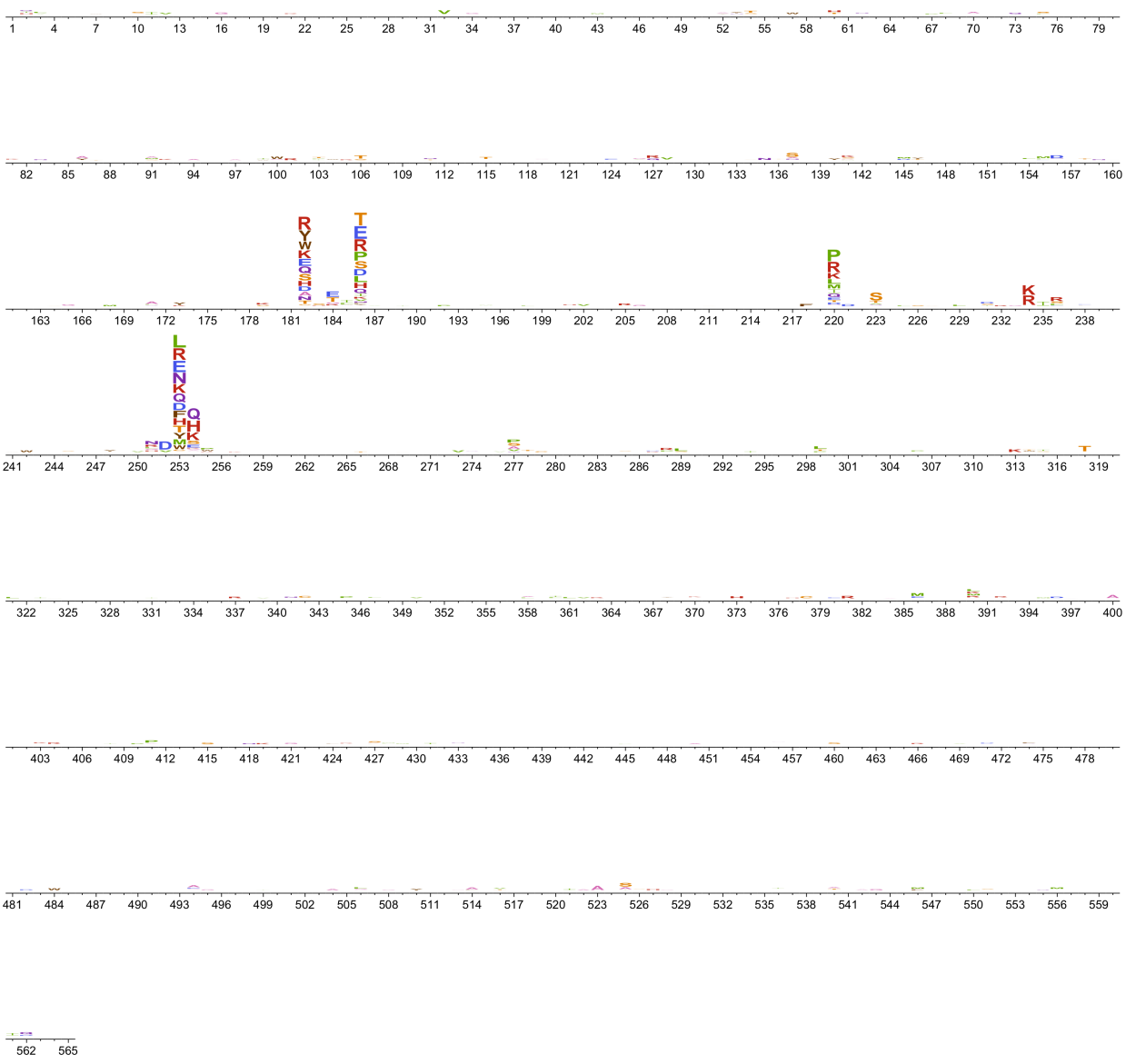

Figure C.4: **A logo plot showing the differential selection across all of HA from antibody H17-L10 at the concentration used in Figure 4.4.** These data are the average across the replicate libraries for each antibody.

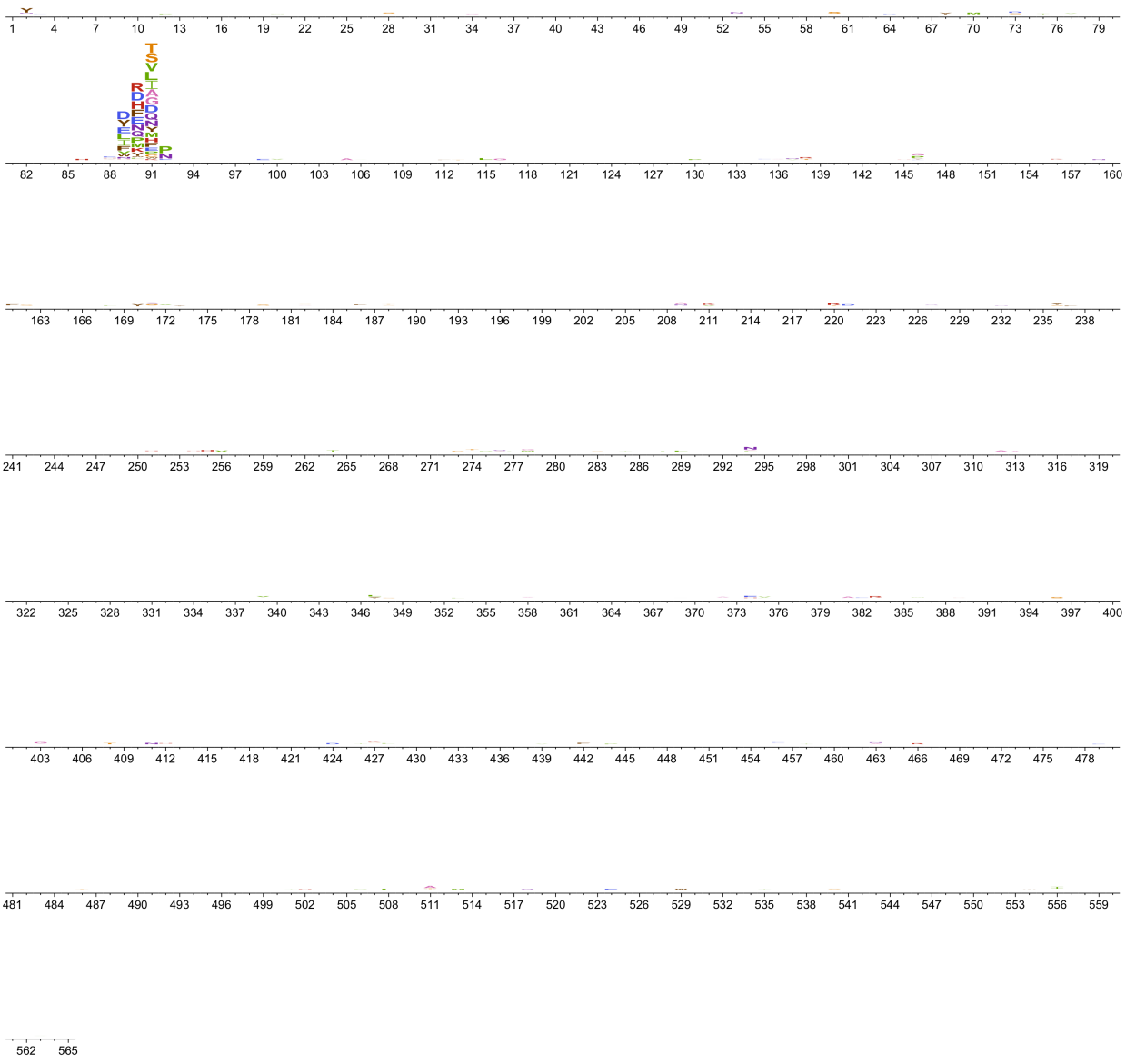

Figure C.5: **A logo plot showing the differential selection across all of HA from antibody H17-L7 at the concentration used in Figure 4.4.** These data are the average across the replicate libraries for each antibody.

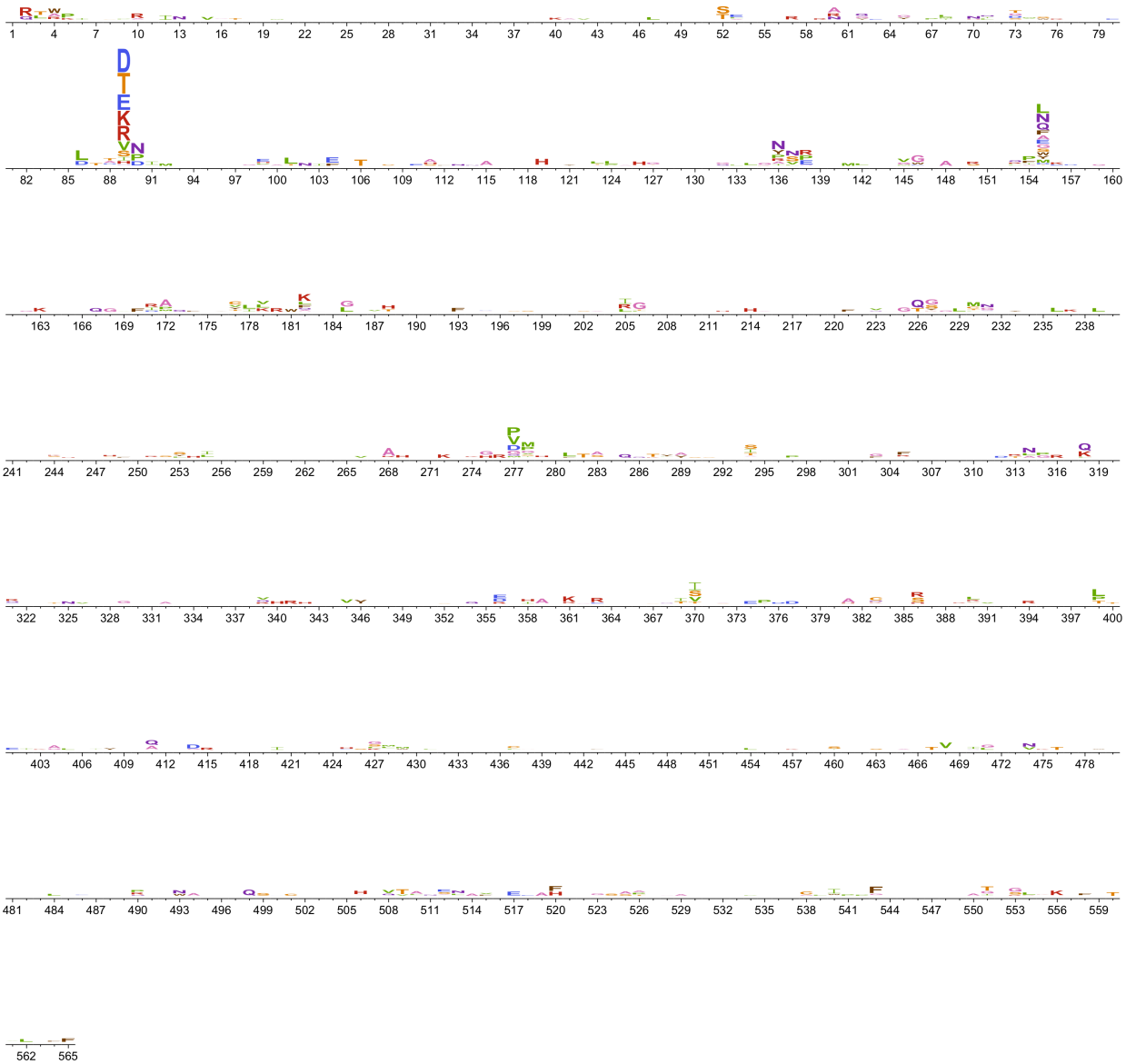

Figure C.6: **A logo plot showing the differential selection across all of HA from antibody H18-S415 at the concentration used in Figure 4.4.** These data are the average across the replicate libraries for each antibody.

Table C.2: **All mutations identified in the classic escape mutant selections with the four antibodies used in our study.** Note that the classic experiments used the A/Puerto Rico/8/1934 (H1N1) virus [53, 23], whereas our study used the A/WSN/1933 (H1N1) virus. In the older papers, multiple names were used to refer to the same antibody: H17-L19 was also called Ca3; H17-L10 was also called Ca6; H17-L7 was also called Cb15; H18-S415 was also called Cb5.

H17-L19

antigenic site	mutant virus name	mutation	WSN HA numbering
Ca2	DV4	S-P	153
Ca2	NV2	G-R	156
Ca2	NV7	S-G	158

H17-L10

antigenic site	mutant virus name	mutation	WSN HA numbering
Ca1	SV3	G-R	253
Ca1	WV8	S-L	220
Ca1	WV10	V-A	182
Ca1	WV11	G-R	186
Ca1	WV15	G-E	253
Ca1	ZV1	G-V	186

H17-L7

antigenic site	mutant virus name	mutation	WSN HA numbering
Cb	AV1	R-G	91
Cb	LV1	R-G	91
Cb	LV7	S-P	92
Cb	RV7	L-P	87

H18-S415

antigenic site	mutant virus name	mutation	WSN HA numbering
Cb	LV7	S-P	92
Cb	RV6	E-K	132
Cb	RV7	L-P	87

Materials and Methods

Availability of data and computer code

Deep sequencing data has been deposited at the Sequence Read Archive under BioSample accession SAMN05789126. The computer code necessary to reproduce all the analysis in this work is available at:

`https://github.com/mbdoud/mutational_antigenic_profiling`.

Mutant virus libraries

The influenza virus mutant libraries used have been described previously [41]. Briefly, reverse-genetics plasmids [94] encoding HA gene were mutagenized at the codon level using a previously described protocol [15]. These plasmid codon-mutant libraries were used to generate libraries of replication-competent influenza viruses using a helper-virus approach that reduced the bottlenecks associated with standard reverse genetics. The virus libraries were then passaged at low MOI to create a genotype-phenotype link between the HA protein on a virion's surface and the gene that it carries. The viral titers in these libraries were determined by TCID_{50} (50% tissue culture infectious dose) in MDCK-SIAT1 cells (obtained from Sigma Aldrich). Three fully independent virus libraries were generated beginning with independent plasmid mutant libraries as outlined in Figure 4.2A. It was these low-MOI passaged virus libraries [41] that formed the starting point for the antibody selections described in the current work.

Antibodies

The antibodies used in this study were originally isolated from mice [53, 23]. Note that in these older papers, two different naming schemes are used for the same antibodies: H17-L19 was also called Ca3; H17-L10 was also called Ca6; H17-L7 was also called Cb15; H18-S415 was also called Cb5. Antibodies secreted by H17-L19, H17-L10, H17-

L7, and H18-S415 hybridoma cell lines were purified using PureProteome A/G coated magnetic beads (Millipore). The hybridomas were originally derived from mice at the Wistar Institute [53], and were provided for this study by Scott Hensley.

Mutant virus selections with antibody

For the selections outlined in Figure 4.1, we began by diluting each virus library in influenza growth media (Opti-MEM supplemented with 0.01% heat-inactivated FBS, 0.3% BSA, 100 U of penicillin/ml, 100 μg of streptomycin/ml, and 100 μg of calcium chloride/ml) to a concentration of 1×10^6 $TCID_{50}$ per ml. Monoclonal antibody was also diluted in influenza growth media to a concentration twice that intended for use the selection. The virus library was then neutralized by mixing 1 ml of diluted virus with 1 ml of diluted antibody to give the final antibody concentrations listed in C.1. This virus-antibody mixture was then incubated at 37^{o}C for 1.5 hours. No-antibody controls were "mock-neutralized" in parallel by substituting influenza growth media for the diluted antibody. At the same time, serial ten-fold dilutions of mutant virus library were made from the 1×10^6 $TCID_{50}$ per ml virus stock to be used as a standard curve to measure infectivity. These dilutions represented 10%, 1%, 0.1%, 0.01%, and 0.001% of the 1×10^6 $TCID_{50}$ dose of library used in neutralizations.

The viral samples were then added to cells to allow infection by non-neutralized virions. We used MDCK-SIAT1 cells that had been plated four hours prior to infection in D10 media (DMEM supplemented with 10% heat-inactivated FBS, 2 mM L-glutamine, 100 U of penicillin/ml, and 100 μg of streptomycin/ml) at 2.5×10^5 cells per well in 6-well dishes. For the infections, we aspirated off the existing D10 media and added the 2 ml of viral sample. Duplicate infections were used for each point on the standard curve of serially diluted virus. After two hours, media in each well was then changed to 2 ml WSN growth media (Opti-MEM supplemented with 0.5% heat-inactivated FBS, 0.3% BSA, 100 U of penicillin/ml, 100 μg of streptomycin/ml, and 100 μg of calcium chloride/ml) after rinsing

cells once with PBS to remove residual virus in the supernatant.

Twelve hours later, RNA was isolated from the cells in each well using a Qiagen RNEasy Plus Mini kit by aspirating media, adding 350 μl buffer RLT freshly supplemented with β-mercaptoethanol, slowly pipetting several times to lyse cells, transferring the lysate to a RNase-free microfuge tube, vortexing for 20 seconds to homogenize, and proceeding with the manufacturer's suggested protocol, eluting in 35 μl of RNase-free water.

We estimated the percent remaining infectivity in the neutralized samples using qRT-PCR and a standard curve created using the infections with 10-fold serial dilutions of the virus libraries to give the estimates in C.1. For the qPCR, primers WSN-NP-qPCR-F (5'-GCAACGGCTGGTCTGACTCACA-3') and WSN-NP-qPCR-R (5'-TCCATTCCTG TGCGAACAAG-3') were used to amplify influenza nucleoprotein (NP) to quantify viral infectivity, and primers 5'-canineGAPDH (5'-AAGAAGGTGGTGAAGCAGGC-3') and 3'-canineGAPDH (5'-TCCACCACCCTGTTGCTGTA-3') were used to quantify canine GAPDH to correct for small differences in total RNA amounts. qRT-PCR was performed using Applied Biosystems PowerSYBR green RNA-to-Ct 1-step kit, with 40 ng of RNA in each 20 μl reaction, cycling conditions of 48 ^{o}C for 30 minutes, 95 ^{o}C for 10 minutes, and 40 cycles of: 95 ^{o}C for 15 sec, 58 ^{o}C for 1 min with data acquisition. All samples were measured in duplicate, and each assay included no-reverse-transcriptase controls. Linear regression of the relationship between the log(infectious dose) and the mean difference in Ct between NP and GAPDH was used to interpolate the remaining infectious dose of each antibody-neutralized sample, expressed as a percentage of the 1×10^6 $TCID_{50}$ used in each neutralization.

Deep sequencing and quantification of mutation frequencies

To prepare deep sequencing libraries, HA genes were amplified from the RNA isolated from infected cells by reverse transcription with AccuScript Reverse Transcriptase (Agilent 200820) using HA-specific primers WSN-HA-for (5'-AGCAAAAGCAGGGGAAAATA

AAAACAAC-3') and WSN-HA-rev (5'-AGTAGAAACAAGGGTGTTTTTCCTTATATTTCT G-3'). PCR amplification of HA cDNA and Illumina sequencing library preparation was then carried out using a previously described barcoded subamplicon sequencing protocol [41], which was in turn inspired by the approach of Wu and coworkers [135]. The only change made to the previous protocol [41] was that in order to more effectively spread sequencing depth across samples based on the expected diversity of mutations in each sample, the number of uniquely-barcoded single stranded variants used as template for round 2 PCR was 5×10^5 to 7×10^5 for the no-antibody control samples, and 1.5×10^5 for the antibody-neutralized samples. Sequencing libraries with unique indices for each experimental sample were pooled and sequenced on an Illumina HiSeq2500 using 2 x 250 bp paired-end reads in rapid-run mode.

The frequency of each mutation in each sample was determined by using `dms_tools` [17] (`http://jbloomlab.github.io/dms_tools/`), version 1.1.20, to align subamplicon reads to a reference HA sequence, group barcodes to build consensus sequences, and quantify mutation counts at every site in the gene for each experimental sample.

Computation of differential selection

We computed the extent that each mutation is enriched by each antibody selection by comparing mutation counts in each antibody-treated sample to mutation counts from the matching no-antibody control sample, also utilizing controls to account for PCR and sequencing errors. Specifically, we compute the *differential selection* on each mutation as follows. The error rate $\epsilon_{r,x}$ at each site r for codon x is estimated from the apparent frequency of that mutation in our previously described sequencing of HA from wild-type plasmid using barcoded-subamplicon Illumina sequencing [41]. Specifically, the error rate was calculated as:

$$\epsilon_{r,x} = \left(n_{r,x}^{\mathrm{err}}\right) / \left(\sum_y n_{r,y}^{\mathrm{err}}\right) \tag{C.1}$$

where $n_{r,x}^{err}$ is the number of counts of codon x at site r in the wild-type plasmid sequencing library. Note that for the wildtype codon $x = \mathrm{wt}\,(r)$, $\epsilon_{r,\mathrm{wt}(r)}$ does not represent the rate of "errors" to this codon, but rather the fraction of reads that give the wildtype codon as expected. We then adjusted the observed counts $n_{r,x}^{\mathrm{mock}}$ and $n_{r,x}^{\mathrm{selected}}$ for codon x at site r in the mock selected and antibody selected samples, respectively, to the error-corrected counts $\hat{n}_{r,x}$ for each sample:

$$\hat{n}_{r,x} = \begin{cases} \max\left[\left(\sum_y n_{r,y}\right)\left(\frac{n_{r,x}}{\sum_y n_{r,y}} - \epsilon_{r,x}\right), 0\right] & \text{if } x \neq \mathrm{wt}\,(r) \\ n_{r,x}/\epsilon_{r,x} & \text{if } x = \mathrm{wt}\,(r)\,. \end{cases} \tag{C.2}$$

This correction ignores second-order terms in which a mutant codon is incorrectly read as another mutant codon or wildtype due to sequencing errors; however, provided that both error rates and mutation rates are low (which is the case in our experiments), these second-order terms can be safely ignored.

To convert from codon counts to amino-acid counts, we summed the error-adjusted counts for all codons encoding each amino acid a at site r to give the error-adjusted amino-acid counts $\hat{n}_{r,a}^{\mathrm{mock}}$ and $\hat{n}_{r,a}^{\mathrm{selected}}$ for the mock selected and antibody selected samples, respectively. We then computed the relative enrichment $E_{r,a}$ of amino acid a at site r as

$$E_{r,a} = \frac{(\hat{n}_{r,a}^{\mathrm{selected}} + f_{r,\mathrm{selected}} \times P)/(\hat{n}_{r,\mathrm{wt}(r)}^{\mathrm{selected}} + f_{r,\mathrm{selected}} \times P)}{(\hat{n}_{r,a}^{\mathrm{mock}} + f_{r,\mathrm{mock}} \times P)/(\hat{n}_{r,\mathrm{wt}(r)}^{\mathrm{mock}} + f_{r,\mathrm{mock}} \times P)} \tag{C.3}$$

where $\mathrm{wt}\,(r)$ denotes the wildtype amino acid at site r, P is a pseudocount (set to 10 in our analyses), and $f_{r,\mathrm{selected}}$ and $f_{r,\mathrm{mock}}$ give the relative depths of the selected and mock samples at site r:

$$f_{r,\mathrm{selected}} = \max\left[1, \left(\sum_a n_{r,a}^{\mathrm{selected}}\right) / \left(\sum_a n_{r,a}^{\mathrm{mock}}\right)\right] \tag{C.4}$$

$$f_{r,\mathrm{mock}} = \max\left[1, \left(\sum_a n_{r,a}^{\mathrm{mock}}\right) / \left(\sum_a n_{r,a}^{\mathrm{selected}}\right)\right] \tag{C.5}$$

The reason for scaling the pseudocount by the library depth is that in the absence of such scaling, if the selected and mock samples are sequenced at different depths, the estimates of $E_{r,a}$ will tend to be systematically different from one even if the relative counts are the same in both conditions.

The mutation differential selection values are the logarithm of the enrichment values:

$$s_{r,a} = \log_2 E_{r,a}. \tag{C.6}$$

Mutations that confer escape from an antibody will have a larger relative frequency in the antibody-selected sample than the no-antibody control sample, and will thus have a large, positive differential selection. Therefore, we limited analysis to positive differential selection to identify antibody escape mutations. To summarize the differential selection at each site, we sum the mutation differential selection values $s_{r,a}$ over all amino-acids a with positive mutation differential selection and term this the positive site differential selection s_r for site r:

$$s_r = \sum_a \max\left(0, s_{r,a}\right). \tag{C.7}$$

Logoplots visualizing differential selection display each amino acid with a height proportional to the mutation differential selection $s_{r,a}$. Amino acid letter codes are colored based on the physiochemical properties of the amino-acid side chain: hydrophobic (V, L, I, M, P) are green, nucleophilic (S, T, C) are orange, small (A, G) are pink, aromatic (F, Y, W) are brown, amide (N, Q) are purple, positively-charged (H, K, R) are red, and negatively-charged (D, E) are blue.

The computer code to perform these differential selection analyses is incorporated in the `dms_tools` (`http://jbloomlab.github.io/dms_tools/`) software as the program `dms_diffselection`. The logoplots created by `dms_tools` are rendered with `WebLogo` [32].

GFP-based neutralization assays

We performed neutralization assays using viruses carrying GFP in the PB1 segment using a previously described protocol [70]. These GFP reporter viruses were generated using seven bidirectional reverse genetics plasmids [69] encoding the PB2, PA, HA, NP, NA, M, and NS segments of A/WSN/1933 (kindly provided by Robert Webster of St. Jude Children's Research Hospital), and a unidirectional reverse genetics plasmid pHH-PB1flank-GFP in which the coding sequence of PB1 is replaced by GFP [12]. Since these viruses carry GFP instead of PB1, they are grown in complementing 293T-CMV-PB1 (derived from cells purchased from the American Typed Culture Collection as described in [12]) and MDCK-SIAT1-CMV-PB1 cells (derived from cells purchased from Sigma Aldrich as described in [12]) that constitutively express the WSN PB1 protein.

For each HA mutation tested in the neutralization assay, the indicated amino-acid mutation was introduced into the WSN HA bidirectional reverse genetics plasmid by site-directed mutagenesis, and the HA sequence was verified by Sanger sequencing. To generate each mutant GFP-carrying virus, we transfected a co-culture of 293T-CMV-PB1 and MDCK-SIAT1-CMV-PB1 cells with the eight reverse genetics plasmids described above. For each transfection, 4×10^5 293T-CMV-PB1 and 4×10^4 MDCK-SIAT1-CMV-PB1 per well were plated in 6-well plates in D10 media four hours prior to transfection. Each well received a transfection mixture of 100 μl DMEM, 3 μl BioT transfection reagent, and 250 ng of each of the eight reverse genetics plasmids. At 20 hours post-transfection, the media was changed to WSN neutralization media, which has low autofluorescence in the GFP channel (Medium 199 supplemented with 0.3% BSA, 100 U of penicillin/ml, 100 g of streptomycin/ml, 100 g of calcium chloride/ml, 25 mM HEPES, 0.5% FBS). At 72 hours post-transfection, culture supernatants were clarified by centrifugation at 2,000$\times$g, aliquoted, and frozen at -80 ^{o}C.

The GFP-carrying viruses were titered by flow cytometry in MDCK-SIAT1-CMV-PB1 cells. For this titering, cells were plated in 12-well plates at 1×10^5 cells per well in

WSN neutralization media. Four hours after plating, cells were infected with dilutions of viral supernatant. At 16 hours after infection, wells with approximately 1% of cells GFP-positive were analyzed by flow cytometry, and the fraction of GFP-positive cells was used to calculate the titer of infectious particles in each viral supernatant.

For the neutralization assays, monoclonal antibody was diluted down columns of a 96-well plate in WSN neutralization media. Three replicate dilution columns were used for each virus-antibody combination. Columns without antibody were used to measure maximal fluorescence in the absence of neutralization, and columns without cells were used to measure background fluorescence in viral supernatants, which we found to contribute more background fluorescence than cells alone. The GFP reporter viruses were diluted in WSN neutralization media to 1×10^3 infectious particles per μl and 40 μl (4×10^4 infectious particles) was added to each well. Plates were incubated at 37 ^{o}C for 1.5 hours before adding 4×10^4 MDCK-SIAT1-CMV-PB1 cells to each well. After 16 hours incubation at 37 ^{o}C, GFP fluorescence intensity was measured on a Tecan plate reader using an excitation wavelength of 485 nm and an emission wavelength of 515 nm (12-nm slit widths). Percent of maximal infectivity was calculated by subtracting background fluorescence signal from all wells and dividing the signal from antibody-containing wells by the signal from corresponding wells without antibody.

Acknowledgements

We thank Susanne L. Linderman for assistance with preparing the antibodies used in this study. We thank Bargavi Thyagarajan for assistance in developing initial ideas related to this project. This work was supported by grants R01GM102198 and R01AI127893 from the NIGMS and NIAID of the NIH to J.D.B. The research of J.D.B. was also supported in part by a Faculty Scholar grant from the Howard Hughes Medical Institute and the Simons Foundation. M.B.D. was supported in part by training grant T32AI083203 from the NIAID of the NIH.

BIBLIOGRAPHY

[1] Rhys M Adams, Justin B Kinney, Thierry Mora, and Aleksandra M Walczak. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *eLife*, 5:e23156, 2016.

[2] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[3] J Andrejeva, DF Young, S Goodbourn, and RE Randall. Degradation of STAT1 and STAT2 by the V proteins of simian virus 5 and human parainfluenza virus type 2, respectively: consequences for virus replication in the presence of alpha/beta and gamma interferons. *Journal of virology*, 76(5):2159–2167, 2002.

[4] Frances H Arnold. The library of maynard-smith: my search for meaning in the protein universe. *Microbe*, 6(7):316, 2011.

[5] Orr Ashenberg, L Ian Gong, and Jesse D Bloom. Mutational effects on stability are largely conserved during protein evolution. *Proceedings of the National Academy of Sciences*, 110(52):21071–21076, 2013.

[6] Alejandro B Balazs, Joyce Chen, Christin M Hong, Dinesh S Rao, Lili Yang, and David Baltimore. Antibody-based protection against HIV infection by vectored immunoprophylaxis. *Nature*, 481(7379):81–84, 2012.

[7] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and

D. Lipman. The Influenza Virus Resource at the National Center for Biotechnology Information. *J. Virol.*, 82:596–601, 2008.

[8] T. Bedford, M. A. Suchard, P. Lemey, G. Dudas, V. Gregory, A. J. Hay, J. W. McCauley, C. A. Russell, D. J. Smith, and A. Rambaut. Integrating influenza antigenic dynamics with molecular evolution. *eLife*, 3:e01914, 2014.

[9] Trevor Bedford, Steven Riley, Ian G Barr, Shobha Broor, Mandeep Chadha, Nancy J Cox, Rodney S Daniels, C Palani Gunasekaran, Aeron C Hurt, Anne Kelso, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217–220, 2015.

[10] Asiel A Benitez, Maryline Panis, Jia Xue, Andrew Varble, Jaehee V Shim, Amy L Frick, Carolina B López, David Sachs, et al. In vivo RNAi screening identifies MDA5 as a significant contributor to the cellular defense against influenza A virus. *Cell reports*, 11(11):1714–1726, 2015.

[11] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[12] J. D. Bloom, L. I. Gong, and D. Baltimore. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science*, 328:1272–1275, 2010.

[13] J. D. Bloom, J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami, and F. H. Arnold. Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. USA*, 102:606–611, 2005.

[14] Jesse Bloom. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology Direct*, 12:1, 2017.

[15] Jesse D Bloom. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular Biology and Evolution*, 30:1956–1978, 2014. http://mbe.oxfordjournals.org/content/31/8/1956.

[16] Jesse D Bloom. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Molecular Biology and Evolution*, 31:2753–2769, 2014. http://mbe.oxfordjournals.org/content/31/10/2753.

[17] Jesse D Bloom. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*, 16(1):1, 2015.

[18] Andrew A Bogan and Kurt S Thorn. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, 280(1):1–9, 1998.

[19] Andrew J Bordner and Hans D Mittelmann. A new formulation of protein evolutionary models that account for structural constraints. *Molecular biology and evolution*, 31(3):736–749, 2014.

[20] Jorge Luis Borges. The library of babel. *Collected fictions*, 1998.

[21] Jeffrey I Boucher, Pamela Cote, Julia Flynn, Li Jiang, Aneth Laban, Parul Mishra, Benjamin P Roscoe, and Daniel NA Bolon. Viewing protein fitness landscapes through a next-gen lens. *Genetics*, 198(2):461–471, 2014.

[22] Lauren Byrd-Leotis, Summer E Galloway, Evangeline Agbogu, and David A Steinhauer. Influenza hemagglutinin (ha) stem region mutations that stabilize or destabilize the structure of multiple ha subtypes. *Journal of virology*, 89(8):4504–4516, 2015.

[23] Andrew J Caton, George G Brownlee, Jonathan W Yewdell, and Walter Gerhard. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell*, 31(2):417–427, 1982.

[24] Susan Chacko, Enid Silverton, Lauren Kam-Morgan, Sandra Smith-Gill, Gerson Cohen, and David Davis. Structure of an antibody–lysozyme complex unexpected effect of a conservative mutation. *Journal of molecular biology*, 245(3):261–274, 1995.

[25] Ning Chai, Lee R Swem, Mike Reichelt, Haiyin Chen-Harris, Elizabeth Luis, Summer Park, Ashley Fouts, Patrick Lupardus, Thomas D Wu, Olga Li, et al. Two escape mechanisms of influenza a virus to a broadly neutralizing stalk-binding antibody. *PLoS Pathogens*, 12(6):e1005702, 2016.

[26] Sang Chul Choi, Asger Hobolth, Douglas M Robinson, Hirohisa Kishino, and Jeffrey L Thorne. Quantifying the impact of protein tertiary structure on molecular evolution. *Molecular biology and evolution*, 24(8):1769–1782, 2007.

[27] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO J.*, 5:823–826, 1986.

[28] Peter Y Chou and Gerald D Fasman. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*, 13(2):211–222, 1974.

[29] Mark A Chua, Sonja Schmid, Jasmine T Perez, Ryan A Langlois, et al. Influenza A virus utilizes suboptimal splicing to coordinate the timing of infection. *Cell reports*, 3(1):23–29, 2013.

[30] Tim Clackson and James A Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383, 1995.

[31] Davide Corti, Jarrod Voss, Steven J Gamblin, Giosiana Codoni, Annalisa Macagno, David Jarrossay, Sebastien G Vachieri, Debora Pinna, Andrea Minola, Fabrizia Vanzetta, et al. A neutralizing antibody selected from plasma cells that binds to

group 1 and group 2 influenza A hemagglutinins. *Science*, 333(6044):850–856, 2011.

[32] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. WebLogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190, 2004.

[33] Brian C Cunningham and James A Wells. Comparison of a structural and a functional epitope. *Journal of Molecular Biology*, 234(3):554–563, 1993.

[34] J. da Silva, M. Coetzer, R. Nedellec, C. Pastore, and D. E. Mosier. Fitness epistasis and constraints in adaptation in a human immunodeficiency virus type 1 protein region. *Genetics*, 185:293–303, 2010.

[35] William Dall'Acqua, Ellen R Goldman, Wenhong Lin, Connie Teng, Daisuke Tsuchiya, Hongmin Li, Xavier Ysern, Bradford C Braden, Yili Li, Sandra J Smith-Gill, et al. A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. *Biochemistry*, 37(22):7981–7991, 1998.

[36] Kalyan Das, James M Aramini, Li-Chung Ma, Robert M Krug, and Eddy Arnold. Structures of influenza a proteins and insights into antiviral drug targets. *Nature structural & molecular biology*, 17(5):530–538, 2010.

[37] Suman R Das, Scott E Hensley, William L Ince, Christopher B Brooke, Anju Subba, Mark G Delboy, Gustav Russ, James S Gibbs, Jack R Bennink, and Jonathan W Yewdell. Defining influenza a virus hemagglutinin antigenic drift by sequential monoclonal antibody selection. *Cell host & microbe*, 13(3):314–323, 2013.

[38] Daniel C Dennett. Darwin's dangerous idea. *The Sciences*, 35(3):34–40, 1995.

[39] Mark A DePristo, Daniel M Weinreich, and Daniel L Hartl. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews Genetics*, 6(9):678–687, 2005.

[40] Michael B Doud, Orr Ashenberg, and Jesse D Bloom. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol. Biol. Evol.*, 32:2944–2960, 2015.

[41] Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all amino-acid mutations to influenza hemagglutinin. *Viruses*, 8:155, 2016.

[42] Cyrille Dreyfus, Nick S Laursen, Ted Kwaks, David Zuijdgeest, Reza Khayat, Damian C Ekiert, Jeong Hyun Lee, Zoltan Metlagel, Miriam V Bujny, Mandy Jongeneelen, et al. Highly conserved protective epitopes on influenza B viruses. *Science*, 337(6100):1343–1348, 2012.

[43] Amie J Eisfeld, Gabriele Neumann, and Yoshihiro Kawaoka. At the centre: influenza A virus ribonucleoproteins. *Nature Reviews Microbiology*, 13(1):28–41, 2015.

[44] Damian C Ekiert, Gira Bhabha, Marc-André Elsliger, Robert HE Friesen, Mandy Jongeneelen, Mark Throsby, Jaap Goudsmit, and Ian A Wilson. Antibody recognition of a highly conserved influenza virus epitope. *Science*, 324(5924):246–251, 2009.

[45] S Fazekas et al. Antigenic, adaptive and adsorptive variants of the influenza a hemagglutinin. In *The Influenza Virus Hemagglutinin*, pages 25–48. Springer, 1978.

[46] Elad Firnberg and Marc Ostermeier. Pfunkel: efficient, expansive, user-defined mutagenesis. *PLoS One*, 7(12):e52031, 2012.

[47] Ervin Fodor, Louise Devenish, Othmar G Engelhardt, Peter Palese, George G Brownlee, and Adolfo García-Sastre. Rescue of influenza A virus from recombinant DNA. *Journal of virology*, 73(11):9679–9682, 1999.

[48] Douglas M Fowler, Carlos L Araya, Sarel J Fleishman, Elizabeth H Kellogg, Jason J Stephany, David Baker, and Stanley Fields. High-resolution mapping of protein sequence-function relationships. *Nat. Methods*, 7(9):741–746, 2010.

[49] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, 2014.

[50] Tiziano Gaiotto and Simon E Hufton. Cross-neutralising nanobodies bind to a conserved pocket in the hemagglutinin stem region identified using yeast display and deep mutational scanning. *PloS One*, 11(10):e0164296, 2016.

[51] SJ Gamblin, LF Haire, RJ Russell, DJ Stevens, B Xiao, Y Ha, N Vasisht, DA Steinhauer, RS Daniels, A Elliot, et al. The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science*, 303(5665):1838–1842, 2004.

[52] Molly Gasperini, Lea Starita, and Jay Shendure. The power of multiplexed functional analysis of genetic variants. *Nature protocols*, 11(10):1782–1787, 2016.

[53] Walter Gerhard, Jonathan Yewdell, Mark E Frankel, and Robert Webster. Antigenic structure of influenza virus haemagglutinin defined by hybridoma antibodies. *Nature*, 290(5808):713–717, 1981.

[54] Manuel Gil, Marcelo Serrano Zanetti, Stefan Zoller, and Maria Anisimova. Codonphyml: Fast maximum likelihood phylogeny estimation under codon substitution models. *Mol. Biol. Evol.*, 30(6):1270–1280, 2013.

[55] Nick Goldman and Ziheng Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*, 11(5):725–736, 1994.

[56] Lizhi Ian Gong, Marc A Suchard, and Jesse D Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*, 2:e00631, 2013.

[57] Hideo Goto and Yoshihiro Kawaoka. A novel mechanism for the acquisition of virulence by a human influenza A virus. *Proceedings of the National Academy of Sciences*, 95(17):10224–10228, 1998.

[58] Laurent Guéguen, Sylvain Gaillard, Bastien Boussau, Manolo Gouy, Mathieu Groussin, Nicolas C Rochette, Thomas Bigot, David Fournier, Fanny Pouyet, Vincent Cahais, et al. Bio++: Efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol*, 30(8):1745–1750, 2013.

[59] Hugh K Haddox, Adam S Dingens, and Jesse D Bloom. Experimental estimation of the effects of all amino-acid mutations to HIV's envelope protein on viral replication in cell culture. *PLoS Pathogens*, 12(12):e1006114, 2016.

[60] Aaron L Halpern and William J Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 15(7):910–917, 1998.

[61] Michael J Harms and Joseph W Thornton. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nature Reviews Genetics*, 14(8):559–571, 2013.

[62] Michael J Harms and Joseph W Thornton. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*, 512(7513):203–207, 2014.

[63] Nicholas S Heaton, David Sachs, Chi-Jene Chen, Rong Hai, and Peter Palese. Genome-wide mutagenesis of influenza virus reveals unique plasticity of the hemagglutinin and ns1 proteins. *Proceedings of the National Academy of Sciences*, 110(50):20248–20253, 2013.

[64] Gunilla B Karlsson Hedestam, Ron AM Fouchier, Sanjay Phogat, Dennis R Burton, Joseph Sodroski, and Richard T Wyatt. The challenges of eliciting neutralizing antibodies to HIV-1 and to influenza virus. *Nature Reviews Microbiology*, 6(2):143–155, 2008.

[65] Jorja G Henikoff, Jason A Belsky, Kristina Krassovsky, David M MacAlpine, and

Steven Henikoff. Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences*, 108(45):18318–18323, 2011.

[66] Steven Henikoff and Jorja G Henikoff. Embedding strategies for effective use of information from multiple sequence alignments. *Protein Science*, 6(3):698–705, 1997.

[67] Scott E Hensley, Suman R Das, Adam L Bailey, Loren M Schmidt, Heather D Hickman, Akila Jayaraman, Karthik Viswanathan, Rahul Raman, Ram Sasisekharan, Jack R Bennink, et al. Hemagglutinin receptor binding avidity drives influenza a virus antigenic drift. *Science*, 326(5953):734–736, 2009.

[68] Joseph B Hiatt, Rupali P Patwardhan, Emily H Turner, Choli Lee, and Jay Shendure. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods*, 7(2):119–122, 2010.

[69] Erich Hoffmann, Gabriele Neumann, Yoshihiro Kawaoka, Gerd Hobom, and Robert G Webster. A DNA transfection system for generation of influenza A virus from eight plasmids. *Proceedings of the National Academy of Sciences*, 97(11):6108–6113, 2000.

[70] Kathryn A Hooper and Jesse D Bloom. A mutant influenza virus that uses an N1 neuraminidase as the receptor-binding protein. *Journal of Virology*, 87(23):12531–12540, 2013.

[71] Li Jiang, Ping Liu, Claudia Bank, Nicholas Renzette, Kristina Prachanronarong, Lutfu S Yilmaz, Daniel R Caffrey, Konstantin B Zeldovich, Celia A Schiffer, Timothy F Kowalik, et al. A balance between inhibitor binding and substrate processing confers influenza drug resistance. *Journal of molecular biology*, 428:538–553, 2015.

[72] Lei Jin, Brian M Fendly, and James A Wells. High resolution functional analysis of antibody-antigen interactions. *Journal of Molecular Biology*, 226(3):851–865, 1992.

[73] Elizabeth H Kellogg, Andrew Leaver-Fay, and David Baker. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Structure, Function, and Bioinformatics*, 79(3):830–838, 2011.

[74] Jacob O Kitzman, Lea M Starita, Russell S Lo, Stanley Fields, and Jay Shendure. Massively parallel single-amino-acid mutagenesis. *Nature Methods*, 12(3):203–206, 2015.

[75] Björn F Koel, David F Burke, Theo M Bestebroer, Stefan van der Vliet, Gerben CM Zondag, Gaby Vervaet, Eugene Skepner, Nicola S Lewis, Monique IJ Spronken, Colin A Russell, et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*, 342(6161):976–979, 2013.

[76] Caitlin A Kowalsky, Matthew S Faber, Aritro Nath, Hailey E Dann, Vince W Kelly, Li Liu, Purva Shanker, Ellen K Wagner, Jennifer A Maynard, Christina Chan, et al. Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep sequencing. *Journal of Biological Chemistry*, 290(44):26457–26470, 2015.

[77] Florian Krammer and Peter Palese. Influenza virus hemagglutinin stalk-based antibodies and vaccines. *Current opinion in virology*, 3(5):521–530, 2013.

[78] Florian Krammer and Peter Palese. Advances in the development of influenza virus vaccines. *Nature Reviews Drug Discovery*, 14(3):167–182, 2015.

[79] Nicolas Lartillot and Hervé Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6):1095–1109, 2004.

[80] Adam S Lauring and Raul Andino. Quasispecies theory and the behavior of rna viruses. *PLoS Pathog*, 6(7):e1001005, 2010.

[81] Sondra G Lazarowitz, Allan R Goldberg, and Purnell W Choppin. Proteolytic cleavage by plasmin of the HA polypeptide of influenza virus: host cell activation of serum plasminogen. *Virology*, 56(1):172–180, 1973.

[82] Si Quang Le, Nicolas Lartillot, and Olivier Gascuel. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. B*, 363(1512):3965–3976, 2008.

[83] JE Ledgerwood, EE Coates, G Yamshchikov, JG Saunders, L Holman, ME Enama, A DeZure, RM Lynch, I Gordon, S Plummer, et al. Safety, pharmacokinetics and neutralization of the broadly neutralizing HIV-1 human monoclonal antibody VRC01 in healthy adults. *Clinical & Experimental Immunology*, 182(3):289–301, 2015.

[84] Chengjun Li, Masato Hatta, David F Burke, Jihui Ping, Ying Zhang, Makoto Ozawa, Andrew S Taft, Subash C Das, Anthony P Hanson, Jiasheng Song, et al. Selection of antigenically advanced variants of seasonal influenza viruses. *Nature Microbiology*, 1:16058, 2016.

[85] Wendell A Lim and Robert T Sauer. The role of internal packing interactions in determining the structure and stability of a protein. *Journal of molecular biology*, 219(2):359–376, 1991.

[86] Marta Łuksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507(7490):57–61, 2014.

[87] Mark Lunzer, G Brian Golding, and Antony M Dean. Pervasive cryptic epistasis in molecular evolution. *PLoS Genetics*, 6(10):e1001162, 2010.

[88] Javier G Magadán, Surender Khurana, Suman R Das, Gregory M Frank, James Stevens, Hana Golding, Jack R Bennink, and Jonathan W Yewdell. Influenza A virus hemagglutinin trimerization completes monomer folding and antigenicity. *Journal of virology*, 87(17):9742–9753, 2013.

[89] Glenn A Marsh, Raheleh Hatami, and Peter Palese. Specific residues of the influenza A virus hemagglutinin viral RNA are important for efficient packaging into budding virions. *Journal of virology*, 81(18):9727–9736, 2007.

[90] Yves A Muller, Yvonne Chen, Hans W Christinger, Bing Li, Brian C Cunningham, Henry B Lowman, and Abraham M de Vos. VEGF and the Fab fragment of a humanized neutralizing antibody: crystal structure of the complex at 2.4 å resolution and mutational analysis of the interface. *Structure*, 6(9):1153–1167, 1998.

[91] James B Munro, Jason Gorman, Xiaochu Ma, Zhou Zhou, James Arthos, Dennis R Burton, Wayne C Koff, Joel R Courter, Amos B Smith, Peter D Kwong, et al. Conformational dynamics of single HIV-1 envelope trimers on the surface of native virions. *Science*, 346(6210):759–763, 2014.

[92] Chandrasekhar Natarajan, Noriko Inoguchi, Roy E Weber, Angela Fago, Hideaki Moriyama, and Jay F Storz. Epistasis among adaptive mutations in deer mouse hemoglobin. *Science*, 340(6138):1324–1327, 2013.

[93] Richard A Neher and Trevor Bedford. nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, 31(21):3546–3548, 2015.

[94] G. Neumann, T. Watanabe, H. Ito, S. Watanabe, H. Goto, P. Gao, M. Hughes, D. R. Perez, R. Donis, E. Hoffmann, G. Hobom, and Y. Kawaoka. Generation of influenza A viruses entirely from cloned cDNAs. *Proc. Natl. Acad. Sci. USA*, 96:9345–9350, 1999.

[95] World Health Organization. Influenza fact sheet, 2016. `http://www.who.int/mediacentre/factsheets/fs211/en/`, accessed February 12 2017.

[96] E. A. Ortlund, J. T. Bridgham, M. R. Redinbo, and J. W. Thornton. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, 317:1544–1548, 2007.

[97] Ralph Pantophlet, Erica Ollmann Saphire, Pascal Poignard, Paul WHI Parren, Ian A Wilson, and Dennis R Burton. Fine mapping of the interaction of neutralizing and nonneutralizing monoclonal antibodies with the CD4 binding site of human immunodeficiency virus type 1 gp120. *Journal of Virology*, 77(1):642–658, 2003.

[98] JD Parvin, A Moscona, WT Pan, JM Leider, and P Palese. Measurement of the mutation rates of animal viruses: influenza A virus and poliovirus type 1. *J. Virology*, 59(2):377–383, 1986.

[99] Matt D Pauly, Megan Procario, and Adam S Lauring. The mutation rates and mutational bias of influenza a virus. *bioRxiv*, page 110197, 2017.

[100] Brian G Pierce, Zhen-Yong Keck, Patrick Lau, Catherine Fauvelle, Ragul Gowthaman, Thomas F Baumert, Thomas R Fuerst, Roy A Mariuzza, and Steven KH Foung. Global mapping of antibody recognition of the hepatitis C virus E2 glycoprotein: Implications for vaccine design. *Proceedings of the National Academy of Sciences*, pages E6946–E6954, 2016.

[101] AI Podgornaia and MT Laub. Protein evolution. pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347(6222):673–677, 2015.

[102] David D Pollock, Grant Thiltgen, and Richard A Goldstein. Amino acid coevolution induces an evolutionary stokes shift. *Proceedings of the National Academy of Sciences*, 109(21):E1352–E1359, 2012.

[103] Victoria R Polonis, Bruce K Brown, Andrew Rosa Borges, Susan Zolla-Pazner, Dimiter S Dimitrov, Mei-Yun Zhang, Susan W Barnett, Ruth M Ruprecht, Gabriella Scarlatti, Eva-Maria Fenyö, et al. Recent advances in the characterization of HIV-1 neutralization assays for standardized evaluation of the antibody response to infection and vaccination. *Virology*, 375(2):315–320, 2008.

[104] Sergei Kosakovsky Pond, Wayne Delport, Spencer V Muse, and Konrad Scheffler. Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS One*, 5(7):e11230, 2010.

[105] Sergei Kosakovsky Pond, Simon DW Frost, and Spencer V Muse. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–679, 2005.

[106] Jaume Pons, Arvind Rajpal, and Jack F Kirsch. Energetic analysis of an antigen/antibody interface: alanine scanning mutagenesis and double mutant cycles on the HyHEL-10/lysozyme interaction. *Protein Science*, 8(05):958–968, 1999.

[107] David Posada and Thomas R Buckley. 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, 2004.

[108] Vladimir Potapov, Mati Cohen, and Gideon Schreiber. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Prot. Eng. Des. Sel.*, 22(9):553–560, 2009.

[109] Peter Rice, Ian Longden, and Alan Bleasby. EMBOSS: the european molecular biology open software suite. *Trends in Genetics*, 16(6):276–277, 2000.

[110] Jane S Richardson and David C Richardson. Amino acid preferences for specific locations at the ends of alpha helices. *Science*, 240(4859):1648–1652, 1988.

[111] Valeria A Risso, Fadia Manssour-Triedo, Asunción Delgado-Delgado, Rocio Arco, Alicia Barroso-delJesus, Alvaro Ingles-Prieto, Raquel Godoy-Ruiz, Jose A Gavira, Eric A Gaucher, Beatriz Ibarra-Molero, et al. Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Molecular Biology and Evolution*, 32(2):440–455, 2015.

[112] Nicole C Robb, Matt Smith, Frank T Vreede, and Ervin Fodor. NS2/NEP protein regulates transcription and replication of the influenza virus RNA genome. *Journal of general virology*, 90(6):1398–1407, 2009.

[113] Nicolas Rodrigue, Hervé Philippe, and Nicolas Lartillot. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, 107(10):4629–4634, 2010.

[114] Peter B Rosenthal, Xiaodong Zhang, Frank Formanowski, Wolfgang Fitz, Chi-Huey Wong, Herbert Meier-Ewert, John J Skehel, and Don C Wiley. Structure of the haemagglutinin-esterase-fusion glycoprotein of influenza C virus. *Nature*, 396(6706):92–96, 1998.

[115] Chris Sander and Reinhard Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, 9(1):56–68, 1991.

[116] L. Serrano, A. G. Day, and A. R. Fersht. Step-wise mutation of barnase to binase: a procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J. Mol. Biol.*, 233:305–312, 1993.

[117] GI Shapiro, T Gurney, and RM Krug. Influenza virus gene expression: control mechanisms at early and late times of infection and nuclear-cytoplasmic transport of virus-specific RNAs. *Journal of virology*, 61(3):764–773, 1987.

[118] Derek J Smith, Alan S Lapedes, Jan C de Jong, Theo M Bestebroer, Guus F Rimmelzwaan, Albert DME Osterhaus, and Ron AM Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, 2004.

[119] John Maynard Smith. Natural selection and the concept of a protein space. 1970.

[120] Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[121] Jianhua Sui, William C Hwang, Sandra Perez, Ge Wei, Daniel Aird, Li-mei Chen, Eugenio Santelli, Boguslaw Stec, Greg Cadwell, Maryam Ali, et al. Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nature structural & molecular biology*, 16(3):265–273, 2009.

[122] Xiangjie Sun, V Tse Longping, A Damon Ferguson, and Gary R Whittaker. Modifications to the hemagglutinin cleavage site control the virulence of a neurotropic H1N1 influenza virus. *Journal of virology*, 84(17):8683–8690, 2010.

[123] Bargavi Thyagarajan and Jesse D Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife*, 3:e03300, 2014.

[124] Andrew Varble, Randy A Albrecht, Simone Backes, Marshall Crumiller, Nicole M Bouvier, David Sachs, Adolfo García-Sastre, et al. Influenza a virus transmission bottlenecks are defined by infection route and recipient host. *Cell host & microbe*, 16(5):691–700, 2014.

[125] Chris P Verschoor, Pardeep Singh, Margaret L Russell, Dawn ME Bowdish, Angela Brewer, Louis Cyr, Brian J Ward, and Mark Loeb. Microneutralization assay titres correlate with protection against seasonal influenza H1N1 and H3N2 in children. *PloS One*, 10(6):e0131531, 2015.

[126] Laura M Walker, Sanjay K Phogat, Po-Ying Chan-Hui, Denise Wagner, Pham Phung, Julie L Goss, Terri Wrin, Melissa D Simek, Steven Fling, Jennifer L Mitcham, et al. Broad and potent neutralizing antibodies from an african donor reveal a new HIV-1 vaccine target. *Science*, 326(5950):285–289, 2009.

[127] Huai-Chun Wang, Karen Li, Edward Susko, and Andrew J Roger. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC evolutionary biology*, 8(1):331, 2008.

[128] RG Webster and WG Laver. Determination of the number of nonoverlapping antigenic areas on Hong Kong (H3N2) influenza virus hemagglutinin with monoclonal antibodies and the selection of variants with potential epidemiological significance. *Virology*, 104(1):139–148, 1980.

[129] Daniel M Weinreich, Nigel F Delaney, Mark A DePristo, and Daniel L Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770):111–114, 2006.

[130] Don C Wiley and John J Skehel. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annual review of biochemistry*, 56(1):365–394, 1987.

[131] Michael Worobey, Guan-Zhu Han, and Andrew Rambaut. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature*, 508(7495):254, 2014.

[132] Emily E Wrenbeck, Justin R Klesmith, James A Stapleton, Adebola Adeniran, Keith EJ Tyo, and Timothy A Whitehead. Plasmid-based one-pot saturation mutagenesis. *Nature Methods*, 13(11):928–930, 2016.

[133] Nicholas C Wu, Yushen Du, Shuai Le, Arthur P Young, Tian-Hao Zhang, Yuanyuan Wang, Jian Zhou, Janice M Yoshizawa, Ling Dong, Xinmin Li, et al. Coupling high-throughput genetics with phylogenetic information reveals an epistatic interaction on the influenza A virus M segment. *BMC genomics*, 17(1):1, 2016.

[134] Nicholas C Wu, C Anders Olson, Yushen Du, Shuai Le, Kevin Tran, Roland Remenyi, Danyang Gong, Laith Q Al-Mawsawi, Hangfei Qi, Ting-Ting Wu, et al. Func-

tional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS Genet*, 11(7):e1005310, 2015.

[135] Nicholas C Wu, Arthur P Young, Laith Q Al-Mawsawi, C Anders Olson, Jun Feng, Hangfei Qi, Shu-Hwa Chen, I-Hsuan Lu, Chung-Yen Lin, Robert G Chin, et al. High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. *Scientific reports*, 4, 2014.

[136] Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39(3):306–314, 1994.

[137] Ziheng Yang, Rasmus Nielsen, Nick Goldman, and Anne-Mette Krabbe Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–449, 2000.

[138] Qiaozhen Ye, Robert M Krug, and Yizhi Jane Tao. The mechanism by which influenza a virus nucleoprotein forms oligomers and binds rna. *Nature*, 444(7122):1078–1082, 2006.

[139] Jonathan W Yewdell, Andrew J Caton, and Walter Gerhard. Selection of influenza a virus adsorptive mutants by growth in the presence of a mixture of monoclonal antihemagglutinin antibodies. *Journal of virology*, 57(2):623–628, 1986.

[140] Ming Zhang, Brian Gaschen, Wendy Blay, Brian Foley, Nancy Haigwood, Carla Kuiken, and Bette Korber. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology*, 14(12):1229–1246, 2004.

[141] Tian-Hao Zhang, Nicholas C Wu, and Ren Sun. A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing. *BMC genomics*, 17(1):1, 2016.

[142] E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins*, pages 97–166, New York, NY, 1965. Academic Press.