University of Massachusetts Medical School

eScholarship@UMMS

2017-12-18

Adapting the EMPIRIC Approach to Investigate Evolutionary Constraints in Influenza A Virus Surface Proteins

Aneth S. Canale
University of Massachusetts Medical School

Let us know how access to this document benefits you.

ADAPTING THE EMPIRIC APPROACH TO INVESTIGATE EVOLUTIONARY

CONSTRAINTS IN INFLUENZA A VIRUS SURFACE PROTEINS

A Dissertation Presented

By

Aneth Stephen Canale

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences, Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 18th, 2017

Interdisciplinary Graduate Program

ADAPTING THE EMPIRIC APPROACH TO INVESTIGATE EVOLUTIONARY

CONSTRAINTS IN INFLUENZA A VIRUS SURFACE PROTEINS

A Dissertation Presented

By

Aneth Stephen Canale

This work was undertaken in the Graduate School of Biomedical Sciences

Interdisciplinary Graduate Program

Under the mentorship of

Daniel Bolon, Ph.D., Thesis Advisor

Celia Schiffer, Ph.D., Member of Committee

Jennifer Wang, M.D., Member of Committee

William Royer, Ph.D., Member of Committee

Matthew Shoulders, Ph.D., External Member of Committee

Paul Clapham, Ph.D., Chair of Committee

Antony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

December 18th, 2017

DEDICATION

This work is dedicated to my brother Ibrahim 'Ibu' Steven Seme (07/08/1989 – 09/03/2016). Ibu, you were my biggest supporter through all my academic endeavors. I hope I have made you proud. You left us too soon. Mama, baba, Isaac and I miss you every day.

ACKNOWLEDGEMENT

I would like to thank my thesis advisor Dr. Daniel Bolon for giving me the opportunity to conduct my dissertation research in his lab. Dan's mentoring and guidance in the lab has helped me grow into the scientist I am today. After every EMPIRIC screen, Dan always encouraged me to follow up on observations that piqued my interest and made sure to acknowledge that when he presented my work. This small gesture was very encouraging to me as a young scientist and gave me even more reason to look forward to going to lab each morning. In addition to his great mentoring, Dan provided tremendous support for me as an international student. I cannot thank Dan enough for supporting me to visit my family in Tanzania for a few weeks every year.

I would also like to thank Dr. Jennifer Wang for her guidance and mentoring during my dissertation research. I am yet to meet someone that gives more detailed feedback on written work, and responds to email faster than Jennifer. I am grateful I got to work with a female scientist who showed me that it is possible to make time for family and still succeed as a scientist, and look really good doing it. I would like to also thank Dr. Celia Schiffer, Dr. Konstantin Zeldovich and Dr. Christopher Sassetti for their great mentoring and support as members of my thesis research advisory committee. Thanks to Dr. Paul Clapham, Dr. William Royer and Dr. Matthew Shoulders for serving as my dissertation examination committee. I would also like to thank all members of the AIRe (DoD) Project at UMass Med.

I would like to thank past and present members of the Bolon lab. Thanks to Li Jiang for sharing his virology protocols and for teaching me how to work with influenza A virus. Thanks to Parul Mishra, Pamela Cote-Hammarlof, Julia Flynn, Ammeret Russouw, Jeffrey Boucher, Gily Nachum, Carl Hollins, Neha Samant, Mohan Somasundaran and David Mavor. You all made the Bolon lab a great place work. I would also like to thank the Biochemistry and

Molecular Pharmacology Department for providing the great work atmosphere, and the department administration for always being willing to help.

I am also grateful to my family for their love and support throughout this process. Thanks to my dad Steven Seme for instilling in me the importance of education and hard work from an early age, and to my mum Joyce Mwakibwili for teaching me to not be afraid to fight for what I want. Thanks to my brothers Ibrahim and Isaac for keeping me sane and humble. Thanks to my in-laws Brian and Theresa Canale for their love and support, and for the glasses of wine shared at Sunday dinner. Thanks for listening to me complain about my research and for providing encouragement although you often had no idea what I was talking about.

Finally, I would like to extend my sincerest gratitude to my husband, David Canale. Thank you for your unconditional love, support, encouragement and patience. Thank you for coming to lab with me late at night and sometimes in the weekend. I always said I was making a quick stop by the lab and often that meant at least an hour, sometimes two. I cannot imagine going through this without you by my side. Here is to our next chapter in life.

ABSTRACT

Controlling influenza A virus (IAV) infections remains a challenge largely due to the high replication and mutation rates of the virus. IAV is a negative-sense RNA virus with two main surface proteins — hemagglutinin (HA) and neuraminidase (NA). HA recognizes and binds sialic acid on host cell receptors to initiate virus entry. NA also recognizes sialic acid on host cell receptors but functions by cleaving sialic acid interactions to release progeny virus. Because both HA and NA interact with sialic acid on the host cell surface with opposing effects, their balance is essential for optimal viral infectivity. However, the evolutionary constraints that maintain HA and NA function, while conserving a functional balance, are not fully understood.

I adapted the comprehensive and systematic mutational scanning technology, termed EMPIRIC (Exceedingly Meticulous and Parallel Investigation of Randomized Individual Codons), to investigate the local fitness landscape of regions of HA under standard conditions and under drug pressure. We observed that synonymous substitutions had a higher mean absolute fitness effect in the signal than a neighboring HA region used as a control. Folding ΔG calculations revealed a hairpin loop that appeared to be differentially enriched between human and swine IAV variants in sequences of circulating strains. However, the molecular mechanism resulting in the observed host species-specific constraints remains undefined.

Studying the fitness landscape of the receptor binding site of HA revealed the high sensitivity of this region to mutation. However, modulating the levels of NA activity by mutation and by using the NA inhibitor oseltamivir enabled the identification of HA mutations with adaptive potential under selection pressure by oseltamivir. These results highlight the importance of the HA-NA functional balance virus replication and in the development of resistance to oseltamivir inhibitors. These studies provide improved understanding of IAV biology, and can inform the development of improved antiviral agents with reduced likelihood for resistance.

Table of Contents

List of Tables

List of Figures

List of Abbreviations Used in This Work

PB2: Polymerase basic 2

PB1: Polymerase basic 1

PA: Polymerase acidic

NP: Nucleoprotein

bNAb: Broadly neutralizing antibodies

cRNA: Complementary RNA

DMS: Deep Mutational Scanning

DFE: Distribution of fitness effects

ELISA: Enzyme-linked immunosorbent assay

ER: Endoplasmic reticulum

EMPIRIC: Exceedingly Methodical and Parallel Investigation of Randomized Individual Codons

HA: Hemagglutinin

NA: Neuraminidase

IAV: Influenza A virus

IRD: Influenza Research Database

LAIV: Live attenuated influenza vaccine

M1: Matrix protein 1 M1 and

M2: Matrix protein 2

MD: Molecular dynamics

MDCK: Mardin-Darby canine kidney

NS1: Nonstructural protein 1

NS2: Nonstructural protein 2

PFU: Plaque forming unit

TIV: Trivalent influenza vaccine

SRP: signal recognition particle

vRNP: Viral ribonucleoprotein

vRNA: Viral RNA

Chapter I: Introduction

Influenza and global public health

Influenza viruses are important pathogens that cause upper respiratory tract infections in

humans and are a global public health concern. Inhaling virus-containing droplets is the most

common mode of transmission of influenza viruses [1]. Symptoms of influenza infections

typically last 3-5 days and include sore throat, runny nose, coughing, muscle ache, and fever.

Complications such as bronchitis, sinusitis, and pneumonia may occur in individuals such as the

elderly, pregnant women, and in children under 5 years of age, leading to increased morbidity

and mortality [2, 3]. The increasing cost of treatment combined with missed work days due to

illness make influenza infections a great burden to global economic development [4].

The World Health Organization estimates about 3-4 million cases of severe influenza

infections annually resulting in up to 500,000 deaths worldwide [5]. In temperate climates,

seasonal influenza outbreaks often happen in the colder months of October to February, while

influenza outbreaks may happen year-round in the tropics. In addition to seasonal influenza

outbreaks, influenza caused four pandemics in the 20th and 21st centuries [6, 7]. The 1918

Spanish flu alone caused 50-100 million deaths worldwide and remains the deadliest pandemic

in history [6, 8]. The Asian flu of 1957 and the Hong Kong flu of 1968 caused up to six million

deaths combined [7]. Even with this scale of mortality, effective control of influenza infections is

yet to be achieved. As such, improved pandemic preparedness will help reduce morbidity and

minimize mortality in preparation for the next pandemic.

The causative agent

Influenza infections are caused by viruses that belong to the Orthomyxoviridae family, a

family composed of three important genera [9]. Influenza B and C are known to only replicate in

human hosts whereas influenza A virus (IAV) replicates in human and non-human hosts such as chicken, ducks, swine, horses and bats [10-13]. IAV is the most virulent genera and causes the most severe infections in humans. Influenza B causes significantly fewer infections each year. However, a recent study by the Centers for Disease Control and Prevention reports that influenza B infections can be just as severe as infections caused by IAV [14], unlike previous theories.

IAV is an enveloped negative sense RNA virus that is roughly spherical in shape with a diameter of 100-120 nanometers [9]. The IAV genome is composed of eight single-stranded RNA segments, a genome organization with notable evolutionary plasticity by genetic reassortment. IAV genome segments are numbered by length from the longest to the shortest [15]. Each RNA segment is in complex with three polymerase subunits ((polymerase basic 2 (PB2), polymerase basic 1 (PB1) and polymerase acidic (PA)), and nucleoprotein (NP) encoded by RNA segments 1, 2, 3, and 5, respectively [16, 17]. Segments 4 and 6 encode the hemagglutinin (HA) and neuraminidase (NA) proteins that form the majority of surface glycoproteins of IAV. HA recognizes and binds host receptors to initiate infection [18], while NA cleaves host-virus interactions to allow the release of progeny virus [19]. Matrix proteins M1 and M2 are encoded by alternative reading frames on segment 7 [20]. M1 is the most abundant IAV protein that encapsulates viral ribonucleoprotein (vRNP) complexes [21]. M2 forms ion channels in the viral envelope allowing proton exchange between the virus and the cellular environment at low pH [22]. Segment 8 encodes the nonstructural proteins NS1 and NS2 in alternate open reading frames [23]. NS1 enhances the translation of viral mRNA and also inhibits host antiviral responses to allow viral propagation in the host [24]. NS2 mediates nuclear export of vRNP complexes after synthesis [25].

IAV is divided into subtypes based on the antigenic properties of HA and NA on the virus surface. Thus far 18 HA and 11 NA subtypes have been identified [13, 26]. Circulating viruses have different combinations of HA and NA subtypes - such as H1N1, H3N2 and H5N1. Most IAV strains are isolated from aquatic wild birds which serve as natural reservoirs for the virus [27]. Some variants such as H1N1, H3N2 and H7N2 are known to exclusively infect and circulate in humans, whereas H5N1 and H7N7 infect birds and are highly pathogenic. Pigs can be infected by both human and avian variants and therefore serve as mixing vessels for IAV. However, the role of viral and host factors in cross-species transmission and infectivity remain unclear. Understanding how IAV evolution influences virus function and infectivity in different host environments will help control the spread of influenza infections and inform improved prophylactic and therapeutic strategies.

New viral subtypes arise because the IAV genome is rapidly evolving. The process of genetic drift results from the error-prone RNA polymerase of IAV that leads to the accumulation of up to 10^{-4} nucleotide substitutions per site per year [28]. Additionally, the reassortment of viral RNA in hosts co-infected with different viral subtypes also contributes to genome changes by genetic shift. Transmission of antigenically novel zoonotic IAV strains to humans has been associated with all four influenza pandemics of the 20th and 21st centuries [29]. For instance, human IAV strains that had acquired RNA segments encoding avian-like variants were responsible for the 1957 and 1968 pandemics [30, 31]. Recently the H5 and H7 strains previously crossed the avian-human interspecies barrier and were responsible for the deadly avian flu outbreaks in Hong Kong [32]. Understanding the complex mechanism of cross-species transmission of IAV will help control cross-species transmission and provide better preparedness for pandemic IAV.

HA and NA are the main surface proteins in IAV

HA is a 75 kDa type I transmembrane protein that binds to host receptors and mediates virus entry into the host cell [18] (Fig. 1.1a). The 18 known subtypes of HA are classified into 2 groups. Group 1 contains H1, H2, H5, H6, H8, H9, H11, H12, H13, H16, H17, and H18, while group 2 is composed of H3, H4, H7, H10, H14, and H15 [13, 33]. During virus replication in the host cell, HA is synthesized as the precursor protein HA0 and undergoes co-translational trimerization in the ER-golgi complex [34]. Translocation to the ER is mediated by the signal sequence that is formed by the first 14 to 17 amino acids at the N-terminal [35]. The signal sequence is cleaved and does not form part of the surface expressed protein. Proteolytic cleavage of HA0 by host serine proteases results in the HA1 and HA2 subunits that remain linked by disulfide bonds. After cleavage, HA1 forms the globular head composed of the first 320 amino acids while HA2 is composed of 222 amino acids and forms the tail [36]. Previous studies have shown a correlation between HA0 cleavability by host proteases and virus pathogenicity [36]. Maintaining protease activation to form HA1 and HA2 while rapidly accumulating mutations highlights one the many functional constraints shaping IAV evolution.

The HA1 region of HA is distal to the virus membrane and carries both the antigenic sites and the receptor binding site. The amino acid sequence of the antigenic sites is unique for each HA subtype but evolves rapidly thus leading to evasion from neutralization by host antibodies [37]. On the other hand, the sequence of the receptor binding site is generally conserved [38]. The binding site is composed of the 190-a helix loop, the 130-loop, and the 220-loop arranged to form a receptor binding pocket [18, 38]. The interaction between the receptor binding site and the terminal sialic acid on host receptors determines host specificity. HA of human IAV is recognized by α-2,6-linked sialic acid terminating glycans on receptors while HA

of avian IAV is recognized by α-2,3-linked sialic acid [39]. The reason pigs can be infected by both human and avian strains of IAV and serve as mixing vessels for the virus is because epithelial cells in the porcine trachea contain both α-2,3- and α-2,6-linked receptors [40].

Mutations in the receptor binding site that lead to a switch in host specificity have been identified [41]. Specifically, the laboratory-selected L226Q mutant exhibits reduced binding affinity to α-2,6-linked receptors and improved binding to α-2,3 receptors. Switches in host specificity remain of great clinical importance due to the rich abundance of α-2,3-linked sialic acid in the mucins from human lungs [39]. The mucus lining is thought to minimize virus infections yet these receptors in the human lungs have been the source of high mortality pneumonia in influenza complications.

The HA2 region of HA mediates fusion between the viral and endosomal membranes that lead to the introduction of vRNP into the host cell. Membrane fusion is initiated by the fusion peptide region of HA2 [42]. This region is comprised of the first 23 amino acids of HA2 and is highly hydrophobic. In surface-expressed HA, the fusion peptide folds into a spring-loaded structure buried in the core of the trimer, giving HA a metastable conformation [43]. However, during membrane fusion, HA undergoes irreversible conformational change wherein the fusion peptide is released from the buried state and propelled to anchor onto the host membrane. Further structural rearrangement in HA2 ultimately leads to the fusion of viral and host membranes. As such, HA2 has highly conserved structural and functional requirements compared to HA1. Analysis of diverse HA sequences has revealed 30-70% sequence identity between different HA subtypes with higher conservation in HA2 than HA1 [44].

NA is a type II transmembrane protein (470 amino acids, 240 kDa). During infection, NA functions as a sialidase and cleaves sialic acid interactions between the virus and the host cell to

prevent virus aggregation and allow the release progeny virus [45]. NA sequences are classified into two phylogenetic groups. Group 1 consists of N1, N4, N5 and N8 subtypes, while group 2 consists of N2, N3, N6, N7 and N9; the newly identified N10 and N11 appear to be distantly related to both group 1 and group 2 [13]. Comparing amino acid sequences reveals 37.3–66.8% sequence identity between NA subtypes [46].

Structurally, NA is divided into a head region that is formed by six identical antiparallel β-sheets and a fibrous stalk region [19] (Fig. 1.1b). Surface NA functions as a homotetramer. The enzymatic active site of NA is located on the head region and is highly conserved. Eight catalytically important residues, R118, D151, R152, R224, E276, R292, R371, and Y406 (in N2 numbering), are conserved in all NA subtypes [47]. Due to the functional importance and high sequence conservation, the NA active site serves as an effective target for anti-IAV agents. NA inhibitors such as oseltamivir and zanamivir are used widely to treat influenza infections. However, the emergence of oseltamivir resistance mutations like H275Y and N295S undermine the dependability of NA inhibitors for long-term use against influenza infections [48]. Nonetheless, there is increased interest in developing universal influenza vaccines that target these functionally conserved residues on the active site of NA [46, 49]. While such antibodies have demonstrated neutralization potential in mouse models, their application in human subjects and the potential for neutralization evasion remains underexplored.

The IAV infection cycle

During infection, HA binds terminal sialic acid on host cell receptors to initiate viral entry (Fig. 1.1c). At this stage of infection NA enzyme activity must be appropriately balanced to allow
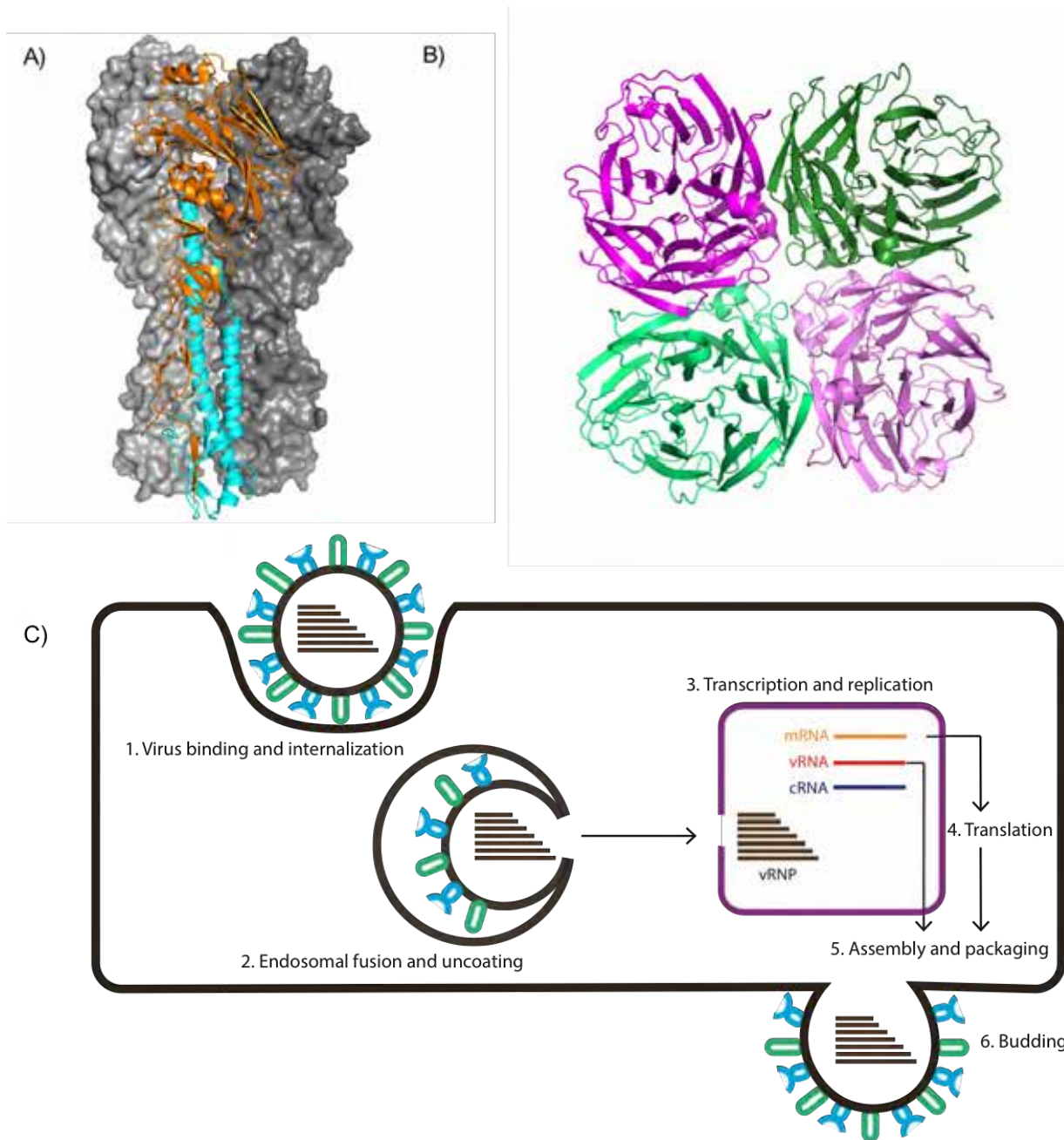
Figure 1.1. HA and NA form the major surface proteins of IAV. A) HA functions as a homotrimer. Each monomer is composed of a head region (orange) and a stem region (cyan) (PDB 1RUZ). B) NA functions as a tetramer. Each monomer is composed of a mushroom-shaped head region and a fibrous stalk (PDB 3BEQ). C) Illustration of the IAV infection cycle.

productive host-pathogen interaction culminating in receptor-mediated endocytosis [50]. A drop in the endosomal pH to between 5 – 6 triggers an irreversible conformational change in HA leading to membrane fusion and the release of vRNP into the host cell [51]. Studies have shown that net charge in the cavity surrounding the fusion peptide determines the pH at which membrane fusion occurs [42]. Mutating charged residues near the fusion peptide leads to the elevation of the fusogenic pH. Targeting the membrane fusion step as a therapeutic target proved futile as viruses quickly developed resistance against the proton pump inhibitors amantadine and rimantadine [52].

Upon entering the cytoplasm of the host cell, vRNP are trafficked to the nucleus for transcription and replication. Transcription of viral RNA (vRNA) is initiated by cis-acting viral RNA polymerase bound to the promoter region on the vRNA template [53]. The viral transcription mechanism relies on 5'-capped host mRNA primers acquired by cap-snatching [54]. IAV also depends on host splicing machinery to splice mRNA encoding matrix and non-structural proteins. Viral factors are proposed to regulate splicing of viral mRNA, although the mechanism remains unknown [55, 56]. Nuclear export factor encoded by NS2 mediates the export of polyadenylated mRNA to the cytoplasm for translation into viral proteins [25]. IAV uses host translation and secretory machinery to synthesize, fold, and modify viral proteins [34].

vRNA replication in the nucleus occurs in two steps. The first step involves copying the vRNA into complementary RNA (cRNA) and the second step entails using cRNA as a template to produce more vRNA [57]. Previous studies have shown that newly synthesized vRNA is selectively exported from the nucleus, yet the molecular basis of this selectivity remains unclear [58]. The transcription and replication machinery of IAV serve as potential targets for antiviral

agents. As such, understanding the molecular mechanism and the regulation of transcription and replication will help to facilitate the development of improved therapeutic agents against IAV.

In preparation for the release of newly synthesized virions, segment-specific packaging signals in the vRNA segments promote the packaging of a complete set of unique RNA segments into progeny virus [59]. Budding of the host cell membrane and NA cleavage of pre-existing interactions between HA and sialic acid promotes the release of newly synthesized viruses from infected cells. However, the molecular mechanism by which IAV hijacks cellular processes and promotes viral replication and release while evading complete neutralization by the host is not fully understood.

Functional balance in receptor binding by HA and receptor destroying by NA

HA and NA both interact with sialic acid on the surface carbohydrates of host cells resulting in virus entry and exit, respectively. Because HA and NA share a substrate, balance between the receptor binding avidity of HA and the enzymatic activity of NA is important for productive infection. However, the influence of this functional balance on IAV pathogenicity and cross-species transmission remains poorly understood. Studies have shown that different HA subtypes bind receptors with different affinity and that different NA subtypes have different levels of enzyme activity [60]. It is generally accepted that circulating viruses have matched receptor binding and receptor-destroying activity. The fact that there are 18 HA and 11 NA subtypes but only a limited number of HA-NA combinations have been isolated in circulating variants provides evidence that some combinations are not well-suited for virus infectivity [61]. Zhang et al. have shown that experimental reassortment studies of the IAV genome result in

strains with varying titers and infectivity [62], suggesting the likely unmatched levels of HA and NA function.

Studies by the Kawaoka group have shown that NA stalk length is important for optimal NA function. Shortening the NA stalk by introducing large deletions led to attenuated enzyme function [63]. However, this group showed that manipulating HA receptor binding strength by introducing mutations near the receptor binding site allowed the rescue of virus function. On the flip side, other studies have shown that mutations that alter HA receptor binding strength can also be rescued by mutations in NA [64-66]. These observations indicate that the HA-NA functional balance is often re-established following offsetting events such as mutations, thus highlighting its importance in virus function.

The significance of this functional balance has been well established. However, little is known on the role of this functional balance in interspecies transmissibility and pathogenicity of IAV. For example, it is now known that the 2009 pandemic originated from the reassortment of two IAV strains of swine origin neither of which was suited for human-human transmission [29, 65, 67]. Analysis of the HA and NA functional levels revealed that the 2009 pandemic strain had elevated levels of NA function to balance HA function. This functional balance was absent in the progenitor swine variants [65]. However, the molecular changes that must happen in IAV in the swine that enable the crossing of the human-swine interspecies barrier followed by sustained transmission in human hosts are not fully understood. Nevertheless, productive virus entry into host cells is only successful if the HA-NA functional balance at the onset of infection allows productive binding of HA to host receptors without interference by NA. The balance in receptor binding and receptor cleavage at the onset of infection remains understudied.

Cross-reactive broadly neutralizing antibodies against IAV

Broadly neutralizing antibodies (bNAb) neutralize IAV by targeting highly conserved functional regions of HA and either inhibit HA binding to host cell receptors, or inhibit the essential pH-induced conformational changes in HA required for membrane fusion. The first cross-reactive bNAb, C179, was isolated in 1993 from mice vaccinated with H2N2 viruses [68]. Since then numerous bNAb have been described. Antibodies targeting the receptor binding site such as (C05, S139 and F045-092) have been shown to neutralize several strains within each IAV subtype [69]. The HA stem has more stringent functional requirements than the head region. As such, stem-binding antibodies that lock the HA stem in the pre-fusion state to inhibit the pH-induced conformational changes in HA2 have shown heterosubtypic neutralization potential. CR6261, F10 and C179 are bNAbs that share an epitope that covers a conserved Trp-21 hydrophobic pocket on the HA stem and neutralize group 1 viruses. bNAbs that neutralize group 2 viruses, such as CR8043 and CR8020, bind a separate but overlapping epitope near the base of the HA stem. Heterosubtypic antibodies that neutralize both group 1 and group 2 viruses, such as 39.29, FI6 and CR9114, have also been identified. These antibodies bind an epitope that overlaps that of group 1-neutralizing antibodies and extends to the base of the HA stem while excluding subtype-specific glycans [69].

Due to the potential for use of antibody-based therapies as an additional strategy to control influenza infections, the field of bNAb discovery has grown exponentially in the last decade. Several heterosubtypic anti-HA antibodies have reached various stages of clinical development [69, 70]. These antibodies have shown significant reductions in virus titer in healthy adult volunteers infected with attenuated virus [71, 72]. The bNAb 39.29 advanced to use in naturally infected uncomplicated patients. However, laboratory evolution studies with

both 39.29 and F10 have identified neutralization resistant variants ([73], and in submitted work by Prachanronarong et al.). The molecular mechanism and evolutionary significance of neutralization evasion against stem-binding antibodies is not fully understood. Studying resistance to broad spectrum antibodies will help facilitate the design of antibody-based therapies, and structure-based vaccines that can elicit production of broad spectrum antibodies and minimize neutralization escape.

Controlling seasonal IAV outbreaks

The most effective way to control influenza infections to date is by annual vaccination. There are two vaccine formulations currently in use – the trivalent influenza vaccine (TIV) and the live attenuated influenza vaccine (LAIV) [74]. TIV consist of two influenza A strains and one influenza B strain inactivated viruses. The vaccine is reformulated annually based on virus strains projected to dominate the upcoming flu season, and is predominantly administered as an intramuscular injection. LAIV also contain 3 virus strains and is administered intranasally. Unlike TIV, LAIV consist of functionally attenuated TIV and are currently not recommended for use due to ineffectiveness. Due to antigenic drift, the influenza vaccine has to be reformulated annually. Mismatch in the vaccine strains of IAV and the circulating strains increases the difficulty of controlling influenza infections. A universal vaccine that can elicit the production of broadly neutralizing antibodies that protect against all influenza variants will provide a solution to the vaccine mismatch problem. As of yet, this holy grail of influenza vaccines still remains to be developed.

The most common method used to produce the influenza vaccine in the United States is the egg-based method [75]. This process begins when the WHO surveillance and response team

provides the candidate influenza strains predicted to dominate in the upcoming flu season. This often happens in February in the northern hemisphere and in September in the southern hemisphere [76]. The candidate viruses are injected into eggs and cultured to allow for virus replication. Virus-containing fluid is then harvested and purified. Recovered virus is inactivated and goes through several rounds of testing before packaging for use in administered vaccines. Predicting the influenza variants projected to dominate in the upcoming flu season follows the analysis of circulating clinical isolates. However, factors such as mismatch in the predicted viral strain and poor growth of certain strains in chicken eggs continue to undermine the effectiveness of the influenza vaccine.

The identification of cold-adapted IAV that exhibited dominant negative properties and suppressed the growth of wild type viruses in cell culture proved to be a promising strategy to control the virus [77]. These viruses carried a mutation in the matrix protein and showed evidence of reduced infection when studied as a live-virus in ferrets [78]. However, this new approach to control IAV infections was not developed for use in a clinical setting.

Antiviral drugs are currently used to reduce the viral load in the host helping to alleviate symptoms and shorten the duration of illness. Most individuals respond well to antiviral drugs. However, life-threatening complications may occur in immunocompromised individuals. M2 ion channel inhibitors, amantadine and rimantadine are some of the earliest drugs that were approved to treat influenza. Widespread resistance to M2 inhibitors has rendered these drugs unfit for use in treating influenza [52]. Oseltamivir, zanamivir and peramivir inhibit NA function therefore blocking viral shedding [79]. Even the relatively successful therapeutic agents such as oseltamivir have limitations. For instance, the initiation of treatment relative to the onset of

infection affects the outcome. Additionally, the emergence of drug resistant viruses following treatment with these drugs remains a challenge.

Approaches to investigate influenza evolution

Phylogenetic methods have been widely used to study IAV evolution [80, 81]. One limitation of using this method is the sparsity of sequencing data for strains that circulated in the early 20th century. Inadequate sampling of sequences may result in punctuated and inconclusive analyses of some phylogenies. Additionally, phylogenetic inferences may undermine important evolutionary constraints underlying the clustering of IAV variants. For instance, about 53% of diversity observed in HA1 of H3 variants results from the 18 codons near the antigenic sites that appear to be under positive selection, while the rest of the 311 residues are more conserved [82]. Additionally, in analyzing IAV evolution, the inability to distinguish genomic reassortment between related strains from genetic drift events may lead to inaccurate phylogenetic clustering [83]. Ultimately, phylogenetic inferences and the corresponding predictive power provide useful information but need to be backed by experimental data.

Efforts to reconstitute influenza virus for study in the laboratory began in the 1980s and ultimately led to the isolation and characterization of replication-competent vRNP [84, 85]. Parvin and colleagues were able to transcribe full-length vRNA by combining purified polymerase with vRNP. These advances led to research using recombinant influenza virus recovered from cloned cDNA and helper viruses [86]. The over-abundance of helper virus in the pool of recovered virus greatly limited the applications of this approach. Technological advances, by the Newmann and Webster groups, in modifying the polymerase I systems combined with cloning all 8 segments of the A/WSN/1933 H1N1 strain of IAV greatly improved

the reverse genetics approaches used to study IAV [15, 86]. The success of this technology has provided an excellent platform to study IAV biology.

The 8-plasmid system revolutionized the study of IAV evolution. Passaging viruses in cell culture and using Sanger sequencing to analyze mutations that emerge allowed the study of genetic drift and the analysis of reassortment between IAV subtypes [15]. Incorporating DNA manipulation technologies allowed the introduction of point mutations and sequence deletions to the IAV genome thus providing a better understanding of the structure-function relationships in IAV proteins, viral pathogenicity and evolution [87]. However, these approaches were extremely low-throughput allowing the analysis of a handful of mutations at a time. Additionally, these methods failed to sample all possible mutations similar to the mutational space available to circulating viruses.

Massively parallel mutational scanning approaches used to protein evolution

Alanine scans were the earliest methods used for mutational scanning [88]. Mutating each amino acid to alanine allowed the study of the structural and functional properties of each amino acid in a protein. While this approach is informative, alanine scans fail to capture the kind of mutations that happen in nature wherein random nucleotide mutations result in either synonymous or non-synonymous amino acid changes. The ability to randomize amino acids to all other possible amino acids would provide a more comprehensive picture of the effects of mutations on protein function.

Fitness landscapes represent the fitness effects of all possible combinations of mutations in a protein [89]. However, even for the smallest protein, the number of variants required to generate a complete fitness landscape remain experimentally intractable. Studying local fitness

landscapes can provide information on the evolutionary process of a protein. Recently developed methods that combine high-throughput site-directed mutagenesis and massively parallel sequencing, present a comprehensive study of sequence-structure-function relationships in a protein. These high-throughput mutational scans, such as the exceedingly methodical and parallel investigation of randomized individual codons (EMPIRIC) [90] and the deep mutational scanning (DMS) approach [91] use methods such as error-prone PCR, cassette-based mutagenesis and chemical synthesis to introduce all possible mutations to parts or the whole protein. Expression of plasmid libraries in a genetically tractable system such as phage display, bacteria, yeast, virus or mammalian cells followed by bulk competition allows for the enrichment of functional variants and the depletion of deleterious ones. Using next-generation deep sequencing and barcoding technology allows the analysis of the change in mutant frequency for up to hundreds of thousands of mutants [90-93].

High-throughput mutational scanning approaches have been used in directed evolution studies to study the effects of different selective pressures on the fitness landscape of proteins. Additionally, mutational data in combination with structural, biochemical and thermodynamic analyses has provided information on epistasis, protein evolvability and the development of drug resistance. However, analysis of additional viral proteins will be important in discerning whether or not the fitness constraints observed thus far apply to other proteins. Combining results of fitness scans and mathematical modeling provides improved understanding of protein evolution.

Limitations of massively parallel deep sequencing approaches

Patterns of mutational tolerance observed in experimental evolution studies occasionally diverge from sequence conservation patterns seen in nature [94]. Some mutations appear to have

minimal experimental fitness effects and yet appear to be absent in natural sequences. This may be due to factors such as the role of epistasis in protein evolution, differences in the environmental conditions tested [94, 95], differences in the time scale between natural evolution and experimental evolution studies, and the effective population size [96]. Despite this discontinuity, high-throughput mutagenesis studies enhance our understanding of protein function. With the rapid increase in the number of large-scale mutagenesis datasets, the future will see a rise in experimentally informed models that can improve protein engineering and enable us to predict functional consequences of mutations with confidence and therefore predict protein evolution. Predicting the evolution of viral proteins will allow the prediction of drug resistance mutations before they arise and become a public health concern.

Advances in high-throughput sequencing technologies have made high-throughput sequence-based methods a cost-effective approach to study protein evolution. However, there are some limitations that need to be addressed in order to ensure increased accuracy in measuring mutational fitness effects. First, quick and affordable methods that warrant unbiased representation of mutants in starting plasmid libraries are yet to be achieved. Second, some random mutagenesis approaches generate libraries carrying a mixture of single and multiple mutations. These methods allow the measurement of fitness effects as averages of effects in closely related but not identical genetic backgrounds, thus complicating the interpretation of sequence-function relationships. Another challenge is in the bottleneck effects encountered during expression of the plasmid libraries or in the bulk competition assays. Such bottlenecks can be avoided by improved population management to ensure sufficient mutational diversity throughout the experiment thus avoiding stochastic sampling of mutations [97]. Low sequencing coverage and high read errors leading to high sampling noise in sequencing results pose another

challenge [98, 99]. Sequencing samples to sufficient read depth, using paired-end sequencing when possible and the use of unique barcodes have greatly improved the quality of sequencing results. Applying this technology to re-emerging viral diseases, such as influenza, that pose great threat to global public health will help inform the control of seasonal flu outbreaks and minimize pandemics.

Using high-throughput fitness scans to study IAV

High-throughput mutational scanning approaches have been used widely to study IAV function and evolution. Previous studies have used this approach to investigate effect of mutations on virus function [100-102] and in response to host innate immunity responses [103, 104]. Additionally, mutational scanning approaches have also been used to study the role of environmental factors, such as antiviral agents and neutralizing antibodies, on IAV evolution [100]. However, bottlenecks incurred during the virus recovery step using the eight-plasmid system greatly undermined the throughput and reproducibility of these experiments on IAV. Techniques, such as the use of helper virus in the viral recovery step, have helped overcome this hurdle[101]. This improvement will allow for improved studies on virus function and the development of drug resistance.

Scope of thesis

In this thesis, I use the EMPIRIC approach, developed in our lab for studying effects of mutations on genes, to perform a systematic and comprehensive study of the evolutionary constraints shaping the evolution of regions of IAV HA. Specifically, I first study effects of all possible single nucleotide mutations in the signal sequence of IAV HA and attempt to elucidate

the molecular mechanism of the observed fitness effects. I observe that, on average, synonymous mutations in the signal sequence had larger fitness effects compared to a neighboring region of HA. Additionally, the signal sequence appears to be more tolerant of non-synonymous mutations compared to the neighboring region used for comparison. This observation suggests that synonymous substitutions in HA may have greater impacts on IAV evolution than previously proposed. Next, I use the EMPIRIC approach to randomize a region near the HA receptor binding site and analyze effects of NA enzyme activity on the fitness landscape of this region of HA. Most mutations near the receptor binding site of HA were deleterious. However, modulating NA activity either by mutation or by adding the NA inhibitor oseltamivir appeared to influence the fitness landscape by rescuing defective HA mutations. Inhibiting NA function using oseltamivir has an effect great enough to allow the identification of potentially adaptive mutations. Modulating NA activity by mutation was not enough to discern HA mutations with adaptive potential. This observation underscores the importance of the HA-NA functional balance in shaping the evolution of IAV surface proteins. Moreover, these findings highlight the need to consider effects of NA inhibitors on HA fitness, and vice versa, on virus evolution and therefore drug resistance.

Chapter II: Synonymous mutations at the beginning of the influenza A virus hemagglutinin gene impact experimental fitness

This work has been submitted for publication as

Aneth S. Canale, Sergey V. Venev, Troy W. Whitfield, Daniel R. Caffrey, Wayne A. Marasco, Celia A. Schiffer, Timothy F. Kowalik, Jeffrey D. Jensen, Robert W. Finberg, Konstantin B. Zeldovich, Jennifer P. Wang[#], and Daniel N. A. Bolon[#]. "*Synonymous mutations at the beginning of the influenza A virus hemagglutinin gene impact experimental fitness*".

Abstract

The fitness effects of synonymous mutations can provide insights into biological and evolutionary mechanisms. We analyzed the experimental fitness effects of all single nucleotide mutations, including synonymous substitutions, at the beginning of the influenza A virus hemagglutinin (HA) gene. Many synonymous substitutions were deleterious in both bulk competition and for individually isolated clones. Investigating protein and RNA levels of a subset of individually expressed HA variants revealed that multiple biochemical properties contribute to the observed experimental fitness effects. Our results indicate that a structural element in the HA segment viral RNA (vRNA) may influence fitness. Examination of naturally evolved sequences in human hosts indicates a preference for the unfolded state of this structural element compared to that found in swine hosts. Our overall results reveal that synonymous mutations may have greater fitness consequences than indicated by simple models of sequence conservation, and we discuss the implications of this finding for commonly used evolutionary tests and analyses.

Introduction

Synonymous mutations change the nucleotide sequence of a gene without changing the

encoded protein sequence. According to a simple interpretation of the central dogma of

molecular biology, synonymous mutations do not impact fitness because they do not change

protein sequences. In this model, all synonymous mutations are neutral or "silent" and the

evolution of synonymous substitutions is thus determined by mutational processes and genetic

drift (e.g., the likelihood of different nucleotide substitutions made by replicative polymerases)

[105]. The availability of a greater number of gene sequences has made it clear that synonymous

mutations are subject to natural selection [106-109].

The signatures of selection acting on synonymous substitutions provide insights into

evolutionary and biochemical mechanisms. For example, highly expressed genes evolve at a

slow rate and are preferentially composed of codons that have cognate tRNA at high abundance

[110-112]. These observations have led to the hypothesis that high expression selects for

synonymous codons that can be translated efficiently and/or accurately in order to minimize

fitness costs that scale with expression level [113, 114].

Despite sequence conservation patterns indicating that natural selection commonly acts

on synonymous mutations [106], identification of synonymous mutations with measureable

changes in experimental fitness is rare (e.g., [115]). Many potential explanations could account

for this apparent discrepancy [116]. These include experimental conditions that do not

encompass the full breadth of environments encountered in nature [117], non-linear relationships

between protein function and fitness [118-120], and dramatically different timescales for natural

selection events to occur compared to the shorter timescales in measuring fitness experimentally

[116]. For these reasons, when synonymous mutations are made at sites that exhibit strong

nucleotide conservation during natural evolution they often do not cause measureable changes in experimental fitness [121].

Distinct patterns of sequence conservation of synonymous substitutions have been observed at the beginning of genes [122, 123]. Rare or non-preferred codons are disproportionately observed in the first 90-150 nucleotides of genes. Rare codons at the beginning of genes can impact the efficiency of translation initiation and appear to be important in regulating the spacing of ribosomes on mRNA [124]. Investigation of the impacts of synonymous mutations on GFP expression in bacteria indicated that secondary structure at the beginning of the gene has a large impact on the amount of GFP protein expressed [125]. These observations motivated us to search for potential experimental fitness effects at the beginning of the hemagglutinin (HA) gene of influenza A virus (IAV).

Because IAV has pandemic potential in the human population [126], a strong interest exists in understanding detailed mechanisms of how it can evolve. IAV is a negative-strand RNA virus with two main surface proteins, HA and neuraminidase (NA). HA and NA each bind to sialic acid but have opposing functions: HA mediates membrane fusion that facilitates viral entry [18] while NA cleaves sialic acid on host cells to release newly synthesized virions [127]. HA and NA both utilize host machinery for synthesis and trafficking.

Trafficking of proteins to appropriate cellular compartments is a highly conserved process that viruses utilize for propagation in host cells. Signal sequences, located at the N-terminus of proteins, often serve as molecular addresses for targeting secretory proteins to the endoplasmic reticulum (ER) [128-130]. During synthesis of these proteins, which include HA, the signal recognition particle (SRP) binds to the signal sequence and mediates co-translational trafficking of nascent polypeptide chains to the ER (Fig. 2.1a) [131, 132]. As for many secretory

proteins, the signal sequence of HA is cleaved by a host protease and does not form part of the mature protein [133-135]. The biochemical mechanisms of HA synthesis are clearly important for generating infectious IAV. Yet the sensitivity of IAV infectivity to quantitative changes in the efficiency of HA synthesis or the level of HA in viral particles has not been well defined. Because the signal sequence is ultimately cleaved off before HA exits the ER, mutations introduced into the signal sequence can in principle alter the amount of HA protein generated without changing the sequence of the mature HA protein.

Signal sequences of different proteins often share common biophysical properties. These generally include three key features: positively charged amino acids, followed by a hydrophobic region, followed by polar amino acids [136, 137]. The signal sequence of HA ranges from 14 to 17 amino acids in length and contains all three of these features (Fig. 2.1b), but the amino acid sequence is generally variable between different HA subtypes (Fig. 2.1c) [138, 139]. Previous studies have shown that large deletions in the signal sequence of HA affect protein maturation and cellular localization, underscoring this region's crucial role [140]. However, a systematic analysis of the impact of signal sequence mutations on HA synthesis, surface expression, and viral function has not been performed.

Several groups have used deep mutational scanning approaches to study effects of mutations on IAV genes, including HA and NA. The Bloom group studied mutational tolerance and antigenic evolvability in HA at the amino acid level by calculating the average effect of each mutation in a collection of mutants with closely related genetic backgrounds [101, 141, 142]. Wu et al. used deep mutational scanning to measure effects of HA single nucleotide mutations on viral experimental fitness in order to identify potential targets for vaccine and drug design [143]. Both groups used error-prone PCR to perform mutagenesis across the full-length HA protein.

(A)

(B) WSN signal sequence

```
  1    5       10       15      20
 MKAKLLVLLYAFVATDADTIC
```

Hydrophobic (h) Region   Peptidase cut site

(C)
```
A/WSN/33(H1N1) KAKLLVLLYAFVATDA
pH1            ..I.V....T.TTAN.
non-pH1        ..I.V....TFATAN.
H2            TITF.I..FTI.KGDQ
H3            .TIIALSYILCLVFAQ
H5            ERIVIA.AIISIVKGD
H7            NTQI..FALVAIVSDK
H9            ETVS.ITILLVATVSN
H10           YKIV.VFALLGAVNGL
```
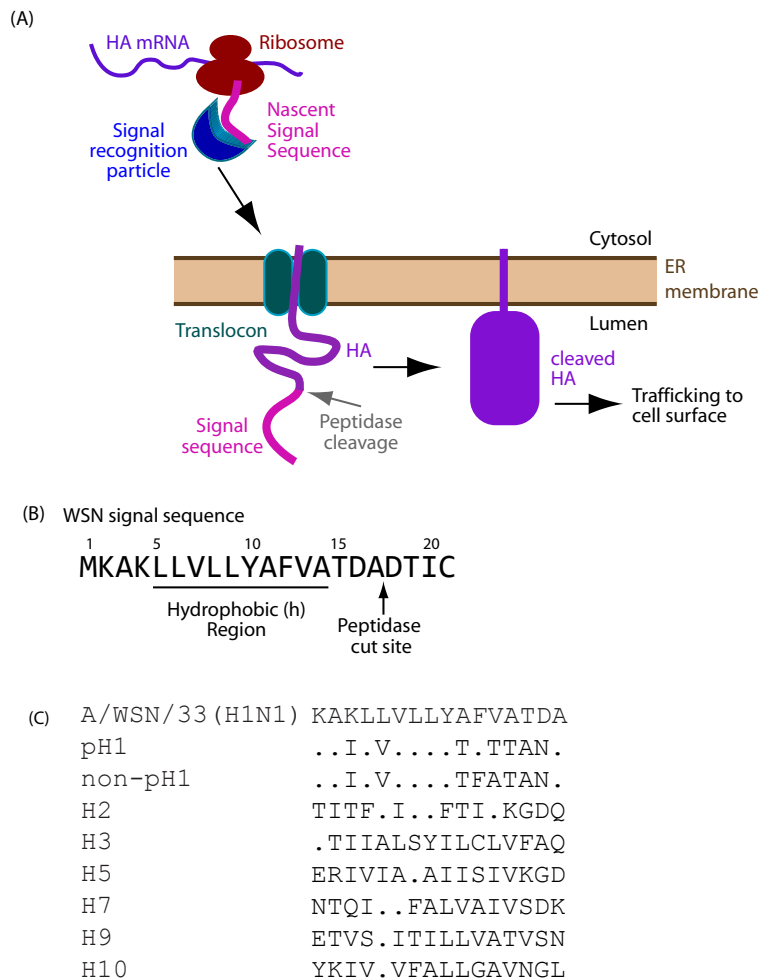
Figure 2.1. Function and properties of the signal sequence of HA. (A) Functional illustration. The signal recognition particle (SRP) recognizes the signal sequence of nascent HA, initiating the targeting of HA to the ER lumen. An ER peptidase cleaves the signal sequence, excluding it from the mature protein that is eventually trafficked to the cell surface. (B) Amino acid sequence of the HA signal sequence for A/WSN/33 IAV, highlighting the hydrophobic region important for SRP binding and the peptidase cut site. (C) Amino acid sequence of the HA signal peptide in different HA subtypes. Consensus sequences were generated from 1243 pandemic H1, 1315 non-pandemic H1, 598 H2, 1506 H3, 1585 H5, 1770 H7, 1168 H9 and 925 H10 sequences retrieved from the Influenza Research Database.

It is important to note here that experimental evolution studies investigate the effects of mutations in a limited genetic background in a controlled environmental condition. Such studies do not perfectly mimic natural evolution where more complex genotypes compete in varying environmental conditions and infect different hosts. For these reasons, the conclusions of experimental evolution studies should be carefully considered relative to potential distinctions with natural evolution.

In the present study, we used the EMPIRIC (Exceedingly Meticulous and Parallel Investigation of Randomized Individual Codons) approach that we previously developed to investigate the experimental fitness landscapes of genes [100, 118, 144, 145] to systematically investigate the effects of all possible single nucleotide mutations, including all possible synonymous mutations, in the signal sequence of HA (Fig. 2.2a). We measured experimental fitness which, in our study, represents the replication capacity of each mutant calculated as the change in relative mutant frequency before and after bulk competition. The results of our high throughput screen indicated that many synonymous mutations cause large experimental fitness effects. Using the screen as a guide, we independently analyzed a set of synonymous mutations and again observed many large experimental fitness effects. To probe the physical basis of the experimental fitness effects, we performed biochemical analyses, including estimates of vRNA and mRNA abundance, and efficiency of surface protein expression on a panel of point mutants.

Results

Quantification of experimental fitness effects of mutations in the HA signal sequence

We systematically generated and measured the relative effects of each single nucleotide mutation to the HA signal sequence in the A/WSN/33 H1N1 strain (Fig. 2.2). A number of

mutations were depleted during virus recovery from plasmid and/or during subsequent virus infection in the bulk competition assay to the point where the frequency change due to selection could not be accurately determined. While these mutations are likely deleterious, they were marked as "not determined" in the heatmap (Fig. 2.2b, Supplementary Fig. 2.1) as the severity of defects could not be quantified. Extra stop codons were included at elevated frequency in our mutant libraries to facilitate quantification of the depletion of null alleles. Using this approach, the experimental fitness effects of mutations were quantified on a scale from null ($s = -1$) to wild type ($s = 0$).

Experimental fitness measurements during viral infection were generally reproducible with an $R^2 = 0.76$ in replicates of the bulk competitions (Fig. 2.2c). This is the range of experimental reproducibility seen in similar studies ($R = 0.54$ to 0.78) using deep mutational scanning to study IAV HA [102, 141, 142]. For further validation, experimental fitness effects measured in bulk competition were compared to the experimental fitness effects of a panel of 22 HA variants, including 14 synonymous substitutions, that were randomly picked and individually cloned (Supplementary Table 2.1). For this panel of mutants, plaque size was measured as a surrogate of fitness, larger plaques being indicative of higher experimental fitness. A correlation ($R^2 = 0.69$) was observed between experimental fitness estimates from bulk competitions and plaque size measurements (Fig. 2.2d) indicating that our high-throughput experiments provide a generally reproducible readout of mutant effects.

Properties underlying selection on the signal sequence of HA

We examined potential similarities and distinctions in selection acting on the signal sequence in our experiments compared with circulating IAV. The experimental fitness effects
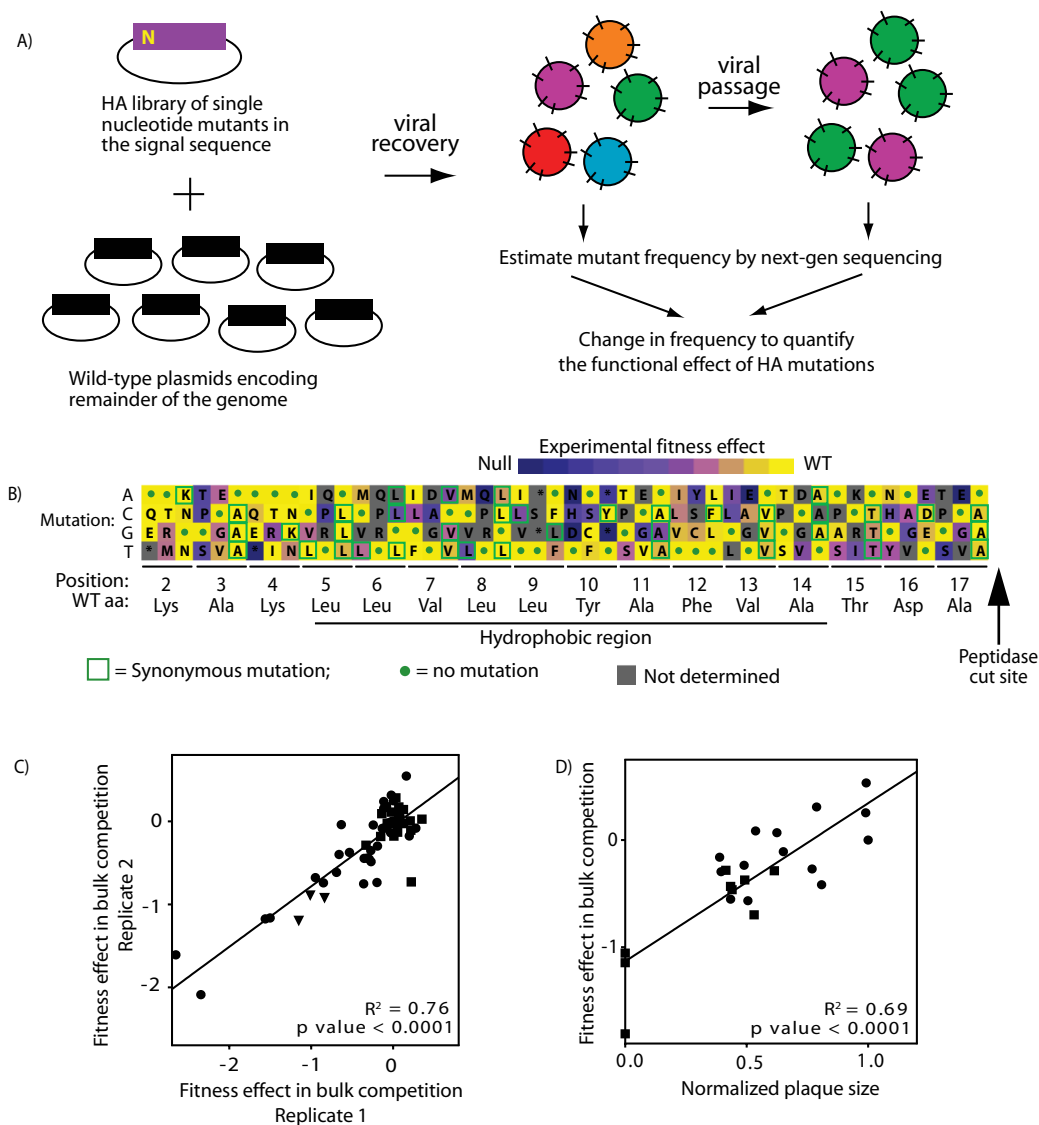
Figure 2.2. Experimental fitness effects of mutations in the signal sequence of HA. (A) Systematic approach to quantify the impact of all single base mutations in the signal sequence of HA on experimental fitness. The HA library of single nucleotide mutants was generated on plasmids that were utilized to recover viruses, recovered viruses were passaged in MDCK cells, and next-generation sequencing was used to estimate the impacts of each mutant on experimental fitness. (B) Heat map representation of the observed impacts of mutations in the HA signal sequence. Synonymous mutations are outlined in green and the wild type nucleotide at each position is shown by a green dot. (C) Reproducibility of the experimental fitness estimates from experimental repeats of the bulk viral expansion. Synonymous mutations are shown as circles, non-synonymous mutations as squares and nonsense mutations as triangles ($R^2 = 0.76$, $P < 0.0001$, from linear regression analysis). (D) Correlation between plaque size of a panel of mutants measured in isolation and average experimental fitness estimates in bulk competition

($n=3$, $R^2 = 0.69$, $P < 0.0001$, from linear regression analysis). Synonymous mutations are shown as circles and non-synonymous mutations are shown as squares.

measured in A/WSN/33 were compared to mutant frequency in 4374 sequenced, non-pandemic H1N1 isolates from the Influenza Research Database [146, 147] (see Methods for details of sequence alignment and filtering). Experimental fitness measurements show a rough trend with mutant frequency in sequenced circulating viral isolates (Fig. 2.3a & Supplementary Fig. 2.2). The experimental fitness effects of mutations observed at least twice in circulating isolates shows a narrower distribution compared to mutations that were not observed in circulating isolates (Fig. 2.3b). There were no mutations observed in circulating IAV that exhibited null fitness in our experiments with A/WSN/33, suggesting that similar biophysical properties may mediate selection on the HA signal sequence in our experiments and in circulating IAV. At a finer level of resolution, many mutations observed in circulating IAV exhibited either apparent fitness benefits or defects in our A/WSN/experiments (Fig. 2.3a), indicating that fitness effects of signal sequence mutations may depend on genetic background, or environmental conditions, and/or noise in our experimental fitness measurements.

We investigated how biophysical properties previously associated with signal sequence function correlate with observed experimental fitness effects. The hydrophobic region of the signal sequence serves as the binding site for the SRP and is essential for SRP-mediated targeting of nascent proteins to the ER. The relationship between residue hydrophobicity in the hydrophobic (h-) region of the signal sequence, quantified using the Kyte & Doolittle hydrophobicity score [148], and virus experimental fitness measured in bulk competition was analyzed (Fig. 2.3b). Hydrophobicity score in the h-region explains roughly 30% (R^2=0.3) of the variation we observed in experimental fitness. This correlation is lower than what we observed between experimental replicates, indicating that hydrophobicity alone does not fully explain the experimental fitness effects. The most hydrophobic amino acid at positions in the h-region did

not always provide the greatest experimental fitness, further suggesting selection on additional features. While mutations to slightly polar amino acids were compatible with robust experimental fitness at some positions, the most polar (least hydrophobic) amino acids were consistently strongly deleterious in our experiments.

Because net hydrophobicity is known to be important for signal sequence function, experimental fitness measurements of mutations that introduced charged residues were analyzed and compared to uncharged mutations. Charged residues were highly depleted in recovered virus despite their high abundance in the plasmid library (Fig. 2.3c). This observation indicates that charged amino acids in the hydrophobic region are strongly deleterious, consistent with the sparsity of charged amino acids in this region among sequenced isolates [137].

Previous studies have indicated that small amino acids are strongly preferred at the -1 and -3 sites relative to the peptidase cleavage site [134]. We assessed how amino acid size at these positions influenced experimental fitness effects. The molecular weight of residues at the -3 and -1 positions relative to the peptidase cleavage site had a significant correlation with experimental fitness in bulk competition (Fig. 2.3d, R^2=-0.36, P=0.02). We note that the strength of this correlation is driven by one large amino acid mutation, A17E, that is strongly deleterious. If A17E is removed, the correlation is not statistically significant (Supplementary Fig. 2.3). Two other large amino acid mutations, T15K and T15R, were severely depleted in recovered virus; although precise quantification of their experimental fitness effects is hindered, the depletion suggests that these substitutions are also strongly deleterious. While our observations are consistent with the -1, -3 rule of preference for small amino acids at these sites relative to the peptidase cut site [133, 134], the molecular weight of amino acids near the peptidase cut site only
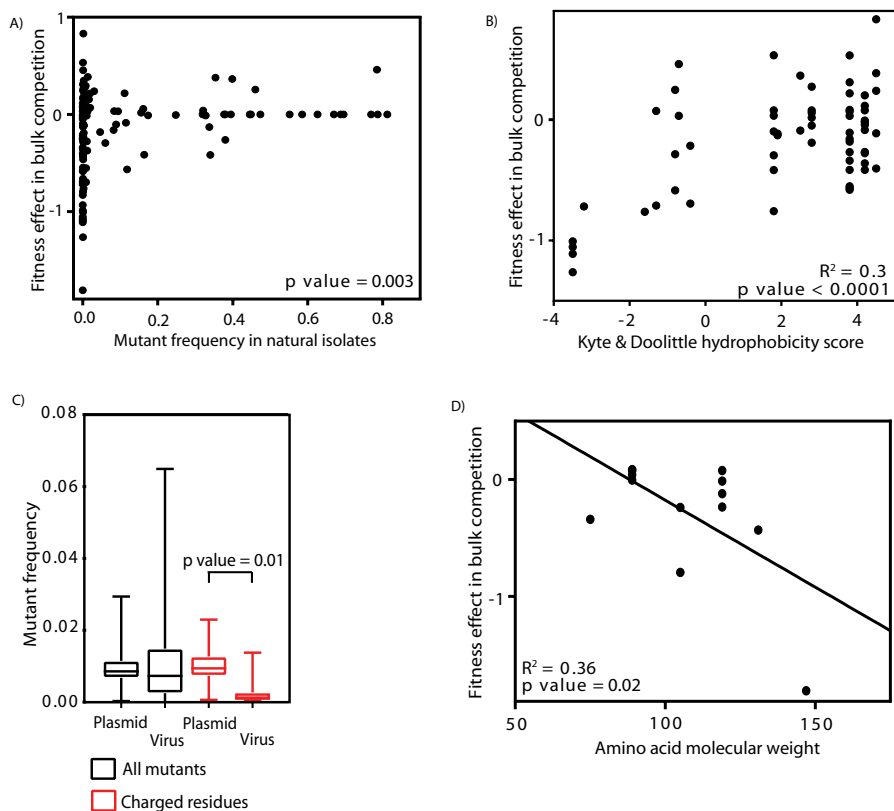
Figure 2.3. Relating experimental fitness with conservation patterns and biochemical properties. (A) Comparing experimental fitness measurements to mutant frequency in sequenced circulating viruses. Reported $P = 0.003$ from linear regression. (B) Correlation between experimental fitness measurements in bulk competition and the Kyte-Doolittle hydrophobicity scores for residues in the h-region ($P < 0.0001$ calculated by linear regression). (C) Box plot showing the frequency of 119 individual mutants in the mutant plasmid library before virus recovery and in recovered virus. Mutant frequency for all variants is shown in black and the frequency of charged residues is shown in red. $P = 0.01$ calculated using the Mann-Whitney test. (D) Correlation between experimental fitness effect and amino acid molecular weight for residues at position 15 and 17, which correspond to the -3 and -1 positions relative to the peptidase cleavage site ($R^2 = 0.36$, P value $= 0.02$ from linear regression).

explains part of the observed variance in experimental fitness effects. Other factors including structural context may also impact the experimental fitness effects of mutants in this region.

Experimental fitness effects of synonymous substitutions

Previous studies have found evidence of biased codon usage indicating that synonymous substitutions may be under selection for translation efficiency and accuracy [125, 149-153]. In our experiments, synonymous substitutions in the signal sequence exhibit a wide distribution of experimental fitness effects (Fig. 2.2), with fitness ranging from wild type-like to strongly deleterious. To investigate how selection on synonymous mutations may change with distance from the beginning of the gene, we used identical EMPIRIC procedures to quantify the experimental fitness effects of mutations at amino acids 32-41 of HA1. This 32-41 region is located on the HA stem proximal to the HA2 region involved in membrane fusion and is generally conserved in group 1 and 2 HA (Supplementary Fig. 2.4). HA is composed of different structural and functional domains that are under varying selection constraints. While the amino acid 32-41 region of HA does not fully represent the rest of HA, we use it here only as a comparison for measuring experimental fitness effects.

In the environmental condition of our experiment, synonymous substitutions exhibited an elevated mean absolute experimental fitness effect in the signal sequence compared to the 32–41 amino acid region (Fig. 2.4a), though this trend was not statistically significant *(P = 0.07)* at the 95% confidence level. However, non-synonymous mutations exhibited higher mean absolute experimental fitness effects in the amino acid 32-41 region than the signal sequence region *(P = 0.001)* (Fig. 2.4a). Using the ratio in mean absolute experimental fitness effect between non-synonymous and synonymous mutations as an estimate of the apparent ratio of selection on the

protein sequence to the nucleic acid sequence, the signal sequence's ratio was significantly lower than in the amino acid 32–41 region (Fig. 2.4b, $P = 0.003$). This difference in experimental fitness ratio indicates the presence of stronger selection on the protein sequence in the amino acid 32–41 region. Our observations are consistent with a combination of stronger selection on the nucleic acid sequence, but weaker selection on the amino acid sequence of the signal sequence relative to a nearby region of HA.

Because of variability in our bulk competitions, we pursued independent experiments on individual clones to further investigate the fitness effects of synonymous mutations in the signal sequence of HA. Consistent with our observations of fitness defects for many synonymous mutations from bulk competition, we also observed that many of the synonymous substitutions in our panel of individual mutants exhibited diminished plaque sizes relative to the parental strain (Fig. 2.4c&d). Both the bulk competitions and the analyses of individual clones indicate that synonymous mutations in the beginning of the HA gene caused measureable defects in experimental fitness.

To further investigate potential differences in the experimental fitness effects of synonymous mutations, infection kinetics experiments were performed on four synonymous substitutions, L5L^{CTC}, L6L^{CTC}, L6L^{TTG} and L8L^{CTA} picked from the panel of individually cloned mutants. Three of these mutants exhibited growth defects in bulk competition (L6L^{CTC}, L6L^{TTG} and L8L^{CTA}), while L5L^{CTC} did not exhibit a growth defect in bulk competition (Fig. 2.2). As individual mutants, each of these four mutants exhibited right-shifted infection growth kinetics relative to wild type virus (Fig. 2.4e&f).

It is challenging to directly compare the infection kinetics results with the bulk competition data and to reconcile differences in results because of distinctions in the inherent
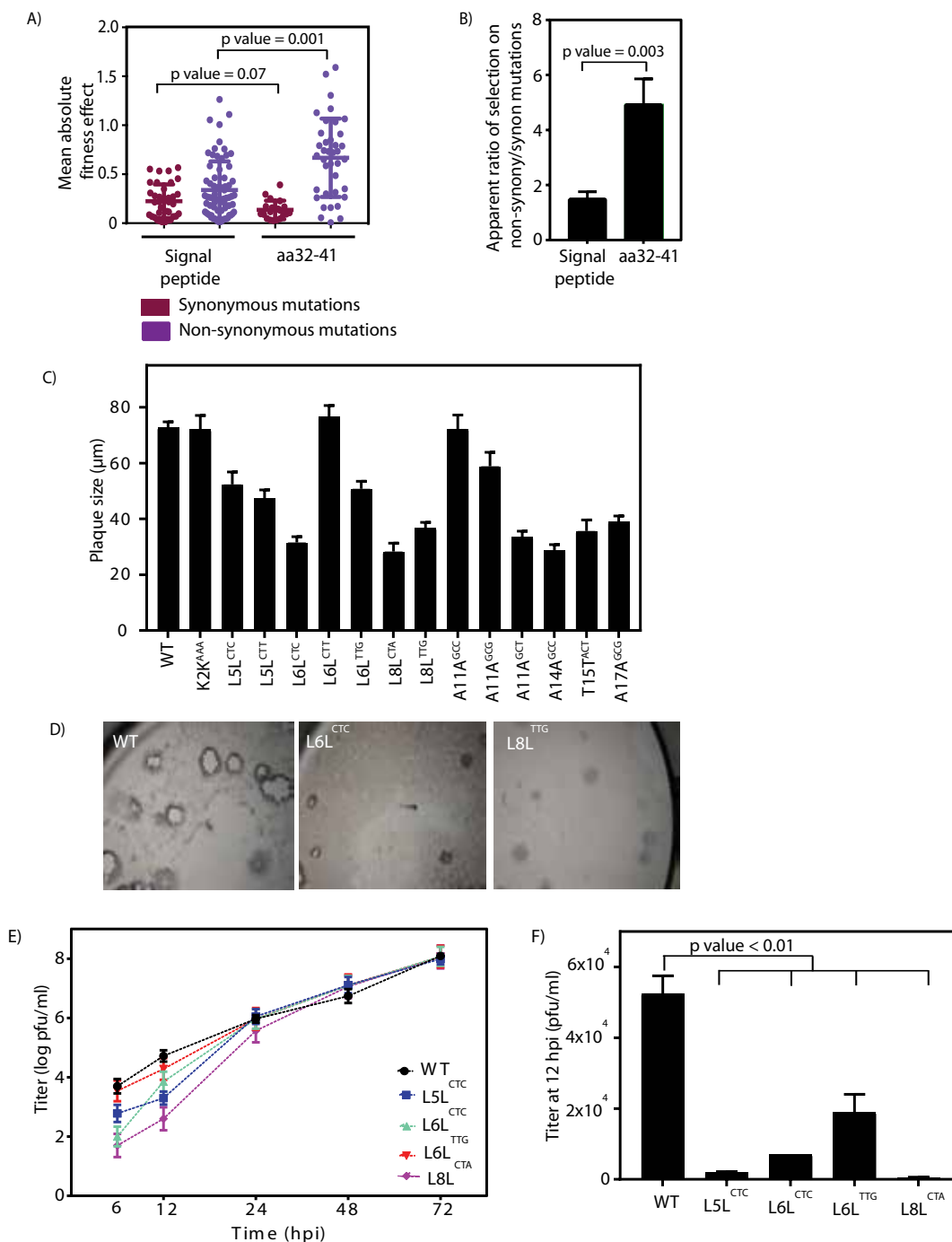
Figure 2.4. Synonymous mutations impact experimental fitness in the signal sequence of HA. (A) Beeswarm plot comparing the mean of absolute experimental fitness effects of synonymous and non-synonymous mutations in the signal sequence and in a region of HA encoding amino acids 32–41. The y-axis represents the magnitude of the experimental fitness effect from wild type (fitness = 0). Error bars represent the standard error of the mean and reported *P* value from multiple comparison analysis test. (B) The apparent ratio of selection acting on non-synonymous/synonymous mutations calculated as the ratio

between mean absolute experimental fitness effects of non-synonymous and synonymous mutations. Error bars are based on propagation of standard errors of the means and reported P value from Student's t-test. (C) Average plaque size measured for a panel of individually cloned synonymous mutations. Error bars represent standard error of the mean. (D) Images for plaques of wild type virus and 2 synonymous substitutions (L6L^{CTC} and L8L^{TTG}) that were found to have experimental fitness defects. (E) Infection kinetics of wild type IAV and four variants carrying synonymous mutations in the signal sequence of HA. Titer was determined based on replicate plaque assays ($n=2$) and the error bars represent the difference between the two assays. (F) Comparison of viral titer on a linear scale at $t = 12$ h during the infection time course. P values from multiple comparison analysis test.

experimental design. In the EMPIRIC experiments, multiple viral mutants are competing against each other, while the infection kinetics experiments follow the expansion of a single viral clone. The host cells available for infection over time likely differ in each of the two types of experiments. To try to account for these distinctions, we compared EMPIRIC selection coefficients with either the viral titer at 12 h or the change in titer between 6 and 24 h from the kinetic experiments and observed poor correlations in both cases (Supplementary Fig. 2.5a&b). While we were not able to develop a simple and effective comparison between the results of the bulk competitions and the kinetics analyses, we noted distinctions in the kinetic behavior of specific mutants. For example, the L5L^{CTC} mutant that did not show a growth defect in the bulk competition and had a low titer at 12 h but exhibited the most rapid expansion between 6 and 24 h in the kinetic experiments.

The discrepancy between the infection kinetics assay and bulk competition (Supplementary Fig. 2.5b) could also be due to inherent challenges in measuring experimental fitness effects with precision using these techniques. Consistent with our analyses of the reproducibility of experimental fitness measurements, it appears that our experiments can distinguish overall trends but are not precise enough to discern the effect of individual mutations with great confidence. For this reason, we have taken multiple approaches to investigate experimental fitness effects, as each approach has caveats. The slowed infection kinetics of three individual mutations that were deleterious in the bulk competitions lends evidence that synonymous mutations in the signal sequence can cause strong defects in viral expansion.

To further investigate the kinetic expansion results, a phenomenological 5-parameter model of IAV growth kinetics [154] was used to fit the experimental data (see Methods and Supplementary Information for details). Sensitivity analysis of the model showed that only two

parameters (β and τ_E) could be reliably estimated (Supplementary Fig. 2.6). The β parameter describes the rate of infection of target cells. The τ_E parameter represents the average duration of an eclipse phase, which is the time cells are infected before virus is produced (see Methods and Supplementary Fig. 2.6, 2.7 for details). Consistent with the infection kinetics data, L5L^{CTC}, L6L^{CTC}, and L8L^{CTA} demonstrated a delayed eclipse phase along with an increased infection rate. The L6L^{TTG} mutant demonstrated WT-like kinetics of viral replication in this model (Table 2.1).

Signal sequence mutations may affect viral RNA structure

It has been reported previously that the folding free energy of a 42-nucleotide long region beginning four nucleotides before the start codon of both bacterial and eukaryotic mRNA affects translation initiation efficiency [155, 156]. Additionally, in the influenza virus, packaging signals in HA segment vRNA span the 5' UTR and extend for up to 45 nucleotides in the HA ORF [157, 158]. Therefore, synonymous mutations in the 5'-end region of the HA segment, which includes part of the signal sequence, can be expected to alter HA expression efficiency as well as vRNA packaging. To investigate the relationship between possible defects in viral RNA and experimental fitness, the folding free energy, ΔG, of secondary structures of both mRNA and vRNA in the 5' region (sense) for all synonymous HA variants was estimated using the ViennaRNA software [159] (Supplementary Fig. 2.8a). Local secondary structure of RNA was probed by restricting the sequence lengths for folding to fixed windows of several tested sizes: 24, 36, 42, 54 and 62 nucleotides. Analysis of the correlation between experimental fitness and ΔG for different window sizes and positions (Supplementary Fig. 2.8B) showed a statistically

significant correlation between experimental fitness and ΔG for a window of size 24 nucleotides beginning at nucleotide 6 in vRNA (permutation test, see Methods).

No significant correlation between the energy ΔG of mRNA structure in the HA signal sequence and experimental fitness was observed. Analysis of predicted vRNA structure at positions 6–30 revealed a small hairpin at positions 7–23 (Fig. 2.5a). The energy ΔG of mutant vRNA in this region has a bimodal distribution, with peaks around -4 kcal/mol (stable hairpin) and -1 kcal/mol (hairpin unfolded) (Supplementary Fig. 2.9). Mutant viruses predicted to be unstructured in this region exhibited a higher mean experimental fitness compared to viruses predicted to have an intact hairpin (Fig. 2.5b). While this trend was not statistically significant ($P = 0.29$), it suggests that less structured vRNA in this window may be beneficial for virus replication under our experimental conditions.

To check for similar findings in sequenced HA of circulating IAV, we evaluated HA sequences from the Influenza Research Database [146, 147]. Targeted RNA folding for the 30 nt region beginning at position 6 at the 5' end of HA segment in the vRNA was performed and the occurrence of different RNA secondary structures in the region was analyzed for strains isolated from either human or swine hosts (Fig. 2.5c). A two-sample Kolmogorov-Smirnov test indicates that the distributions of RNA folding energy for human-derived and swine-derived IAV are different ($P < 0.01$). While viruses from both human and swine are bimodal with peaks at ≈-0.5 kcal/mol and ≈-4 kcal/mol, the human viruses have a higher peak for the unfolded state compared to the swine viruses. These observations are consistent with selection favoring the folded RNA state in swine hosts relative to human hosts.

To investigate relationships between RNA structure and function, we analyzed the structures formed by the stable RNA state. Nearly 90% of the structures with the energy of ≈-4
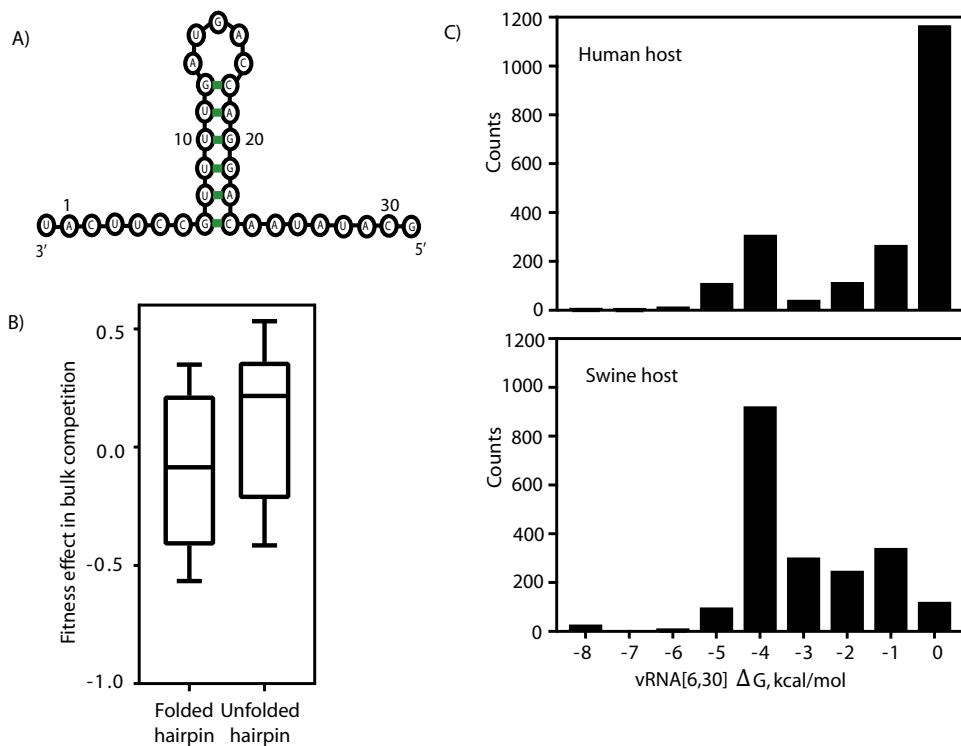
Figure 2.5. Probing effects of mutations on viral RNA. (A) Illustration of the hairpin at nucleotide position 7–23 in the HA vRNA. (B) Comparing experimental fitness of mutations that disrupt the hairpin at nucleotide 7–23 to mutations that leave the hairpin intact (*P*>0.62 calculated using the Mann-Whitney test). (C) Frequency distribution of calculated ΔG for sequenced IAV isolated from either human (top panel) [160] or swine (bottom panel) hosts.

kcal/mol have similar hairpin folds centered around a 5-nucleotide long loop at nucleotides 13 to 17 (or 3-nucleotide loop at positions 14 to 16), such as shown in Fig. 2.5a. We extracted the sequences of the 5-nucleotide loop from sequences of circulating strains that have the stable hairpin, and compared sequence occurrence with experimental fitness values in our bulk competition experiment. The sequence variants present in circulating HA all have WT-like experimental fitness, while assayed sequences that are not found in the database all have an experimental fitness defect (Supplementary Table 2.2). This observation indicates that EMPIRIC fitness measurements for this region are in agreement with trends seen in sequenced natural isolates. Additionally, it implies that the hairpin/loop may be important for viral fitness under specific conditions (e.g., when infecting a swine host), similar to the previously reported packaging interactions between vRNA hairpins of IAV segments 2 and 8 that encode the RNA polymerase subunit PB1 and non-structural protein (NS), respectively [161].

To assess the effects of signal sequence mutations on vRNA packaging, RT-qPCR was used to measure HA vRNA abundance for a panel of individual cell-free mutant viruses. HA vRNA levels were normalized to NA vRNA and used to quantify the packaging efficiency. No strong correlation was found between the efficiency of packaging HA vRNA into particles and EMPIRIC fitness (Supplementary Fig. 2.10). This lack of correlation reveals that viral expansion under our experimental conditions may not be tightly coupled to packaging efficiency, or that we are unable to capture that relationship at our current experimental precision. However, all of the signal sequence mutations analyzed reduced the efficiency of HA vRNA packaging into particles, indicating that the wild type signal sequences facilitate highly efficient packaging.

Signal sequence mutations influence HA protein abundance in viral particles

Because the HA signal sequence is cleaved and excluded from the mature protein, mutations in this region are likely to affect HA synthesis efficiency and expression rather than the ultimate function of the HA protein. ELISA was used to quantify HA protein expression for a panel of individual mutations to examine how signal sequence mutations impact HA protein abundance in cell-free viral particles (Fig. 2.6a, see Methods). All of the studied mutations precede the peptidase cleavage site located at amino acid 17, so mutations would result in identical mature HA protein. Optimal virus experimental fitness, measured in bulk EMPIRIC competition, was supported by a narrow range of HA expression – either too much or too little HA was associated with experimental fitness defects (Fig. 2.6b). The relationship between HA abundance and viral experimental fitness was nonlinear with $R^2 = 0.4$ and $P = 0.007$, and appears to be under stabilizing selection. Of note, we excluded L8L^{TTG} from the fit because its expression level is far removed from all other points. Nevertheless, the relationship is still statistically significant even if L8L^{TTG} is included ($P < 0.05$).

Changes in HA synthesis efficiency partly explain experimental fitness effects

To measure effects of signal sequence mutations on HA synthesis in the absence of other viral RNA segments, HA was transiently expressed in 293T cells and HA synthesis quantified by flow cytometry (Fig. 2.7a, see Methods for details). Non-synonymous substitutions had lower HA expression levels compared to the wild type, while synonymous mutations either reduced or increased HA expression. Additionally, a statistically significant correlation was found between HA abundance and experimental fitness effects measured in bulk (Fig. 2.7b, $P = 0.0002$).
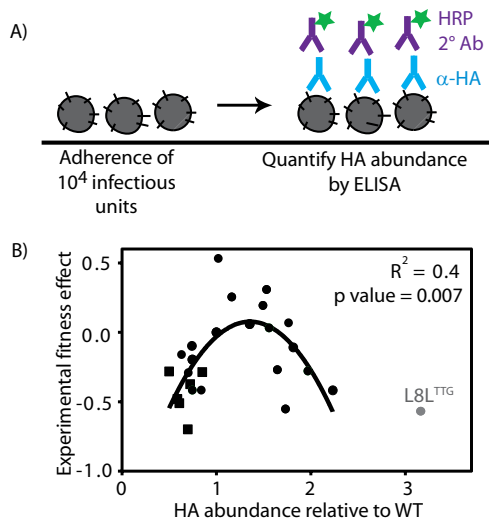
Figure 2.6. Relationship between experimental fitness and relative HA protein in cell-free viruses. (A) Experimental setup: viral particles were adhered to wells in plastic plates and HA abundance was quantified by ELISA. (B) Comparison of mutant impact on experimental fitness and HA abundance. Synonymous mutations are shown as circles and amino acid changing mutations as squares. The synonymous mutation of L8L (TTG), shown in grey, is excluded from the fit. $R^2 = 0.4$ and P value = 0.007 calculated by quadratic nonlinear fit.

To further discern whether experimental fitness effects arise from inherent defects in HA synthesis or from HA expression on virus particles, HA expression in 293T cells and HA abundance in viruses was compared. A significant correlation was found between HA synthesis and HA expression in virus particles (Fig. 2.7c, $P < 0.0001$). However, four synonymous mutations (L6L^{CTC}, V7V^{GTG} L8L^{TTG} and A11A^{GCG}) deviated far from the trend. The frequency of these variants in sequenced non-pandemic H1N1 isolates was analyzed and despite being synonymous substitutions, these variants are rare in circulating viruses. This observation suggests the importance of the nucleotide sequence in this region for virus function. When compared to wild type, these variants had wild type-like or reduced levels of HA expression in 293T cells, but had up to three times as much HA per plaque forming unit (PFU). These observations are consistent with an increased ratio of non-infectious to infectious particles, though further confirmatory studies would be required.

Multiple linear regression analysis of experimental fitness effects

To identify measurable factors affecting bulk experimental fitness of IAV mutants, multiple linear regression analysis was performed. As discussed above, the viral experimental fitness determined in the EMPIRIC experiments is correlated with various metrics of HA *in vitro*, using either a linear or quadratic model, see Figures 2.6b and 2.7b correspondingly. To assess whether additional factors explain the observed variance of experimental fitness, a linear regression model was used with HA synthesis in mammalian cells, HA protein abundance in cell-free viral particles, and mRNA folding energy ΔG as explanatory variables. The combination of HA synthesis in mammalian cells and HA protein abundance in cell-free viral particles improved the prediction of experimental fitness, $R^2 = 0.68$; ANOVA comparison of the

Figure 7
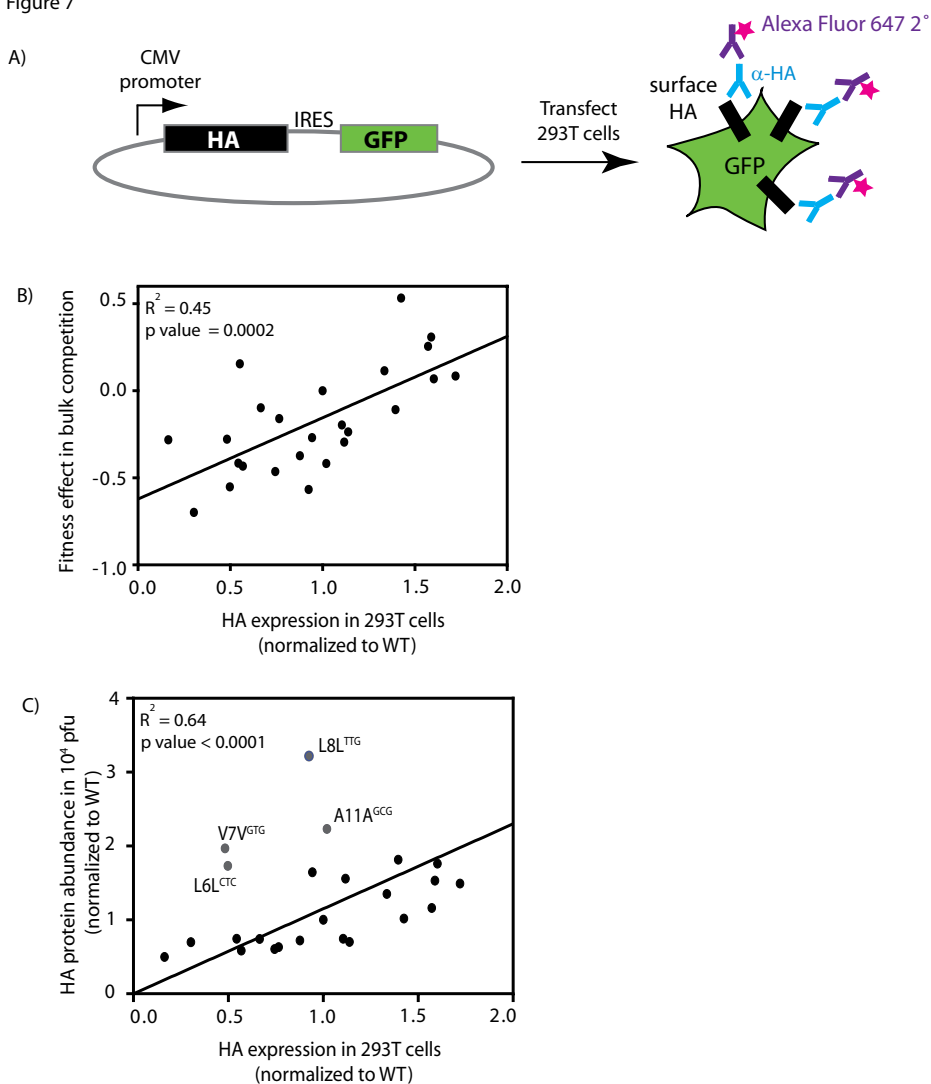
Figure 2.7. Relationship between experimental fitness and HA protein expression. (A) HA was expressed in 293T cells and HA protein expression was measured by flow cytometry. (B) Comparing experimental fitness to HA expression in 293T cells. *P* value calculated from linear regression. (C) Comparison of HA protein abundance in cell free viruses to HA expression in 293T cells (reported *P* value from linear regression). The points in grey exhibit the largest discrepancy between observed abundance in viruses and 293T cells and are omitted from the fit.

models yielding $P = 0.014$. The best prediction was achieved by a linear model combining all three factors, HA protein synthesis, HA abundance in released virus, and mRNA ΔG, yielding R^2 = 0.752, and marginally significant $P = 0.053$ for ANOVA comparison of the 2 vs. 3 factor models. The complete set of P-values of nested linear regression models is presented in Supplementary Table 2.3. The folding energy of mRNA was defined as the free energy ΔG of the viral mRNA segment consisting of nucleotides -4 to 37 relative to the HA start codon, as described in [140]. Adding a categorical variable representing DNA library comprising a particular mutant did not improve the fit, suggesting the measurements were not characterized by batch effect. Likewise, inclusion of additional metrics such as HA vRNA abundance normalized to NA viral RNA, vRNA ΔG, and position of the mutation did not improve the predictive power of the multiple linear regression model for experimental fitness (data not shown). However, the effectiveness of additional predictors in the multiple regression was also limited by the overall number of available data points.

These results suggest that the experimental fitness of synonymous mutants in IAV HA is affected by multiple factors including RNA secondary structure, protein synthesis and surface expression as well as functional interactions between the viral and host components.

Discussion

Our bulk competitions indicate increased selection on synonymous mutations in the signal sequence relative to a nearby region of the HA gene. Consistent with this idea, many individual synonymous mutations in this region caused defects in viral infectivity when analyzed in isolation. The finding that synonymous mutations in the signal sequence affect viral fitness agrees with previous studies that showed that synonymous mutations in the beginning of genes

can have large impacts on translation efficiency [125, 149, 156]. Biochemical analyses of a panel of synonymous mutations revealed that variants with reduced HA expression had lower experimental fitness, suggesting selection on the efficiency of protein synthesis. Signal sequence mutations also appeared to impact HA protein levels in viruses and during expression in the absence of other IAV genes in 293T cells. Our results are in agreement with selection against reduced HA expression that may in turn impact the ratio of infectious to non-infectious particles leading to apparent stabilizing selection on HA abundance per infectious particle. Yet the impacts of synonymous substitutions on translation efficiency appear to only partly explain the experimental fitness effects that we measured.

Changes in vRNA structure also explain some of the experimental fitness effects of synonymous substitutions. We found that change in RNA folding energy in a 24 nt window at position 6 is correlated with experimental fitness. Our analyses also indicate that the hairpin loop at position 7–23 may have structural significance as mutations that disrupted this hairpin loop had higher average experimental fitness in our experiments compared to mutations that left the hairpin loop intact. Intriguingly, clustering of HA sequences from circulating IAV based on the existence of the hairpin loop revealed correlations with the host (swine versus human), in line with earlier reports of lineage specificity of RNA structure in IAV segments [162-164]. This variability in structural features across different IAV strains and lineages may be responsible for the lack of conserved secondary structures in IAV vRNA, as proposed by previous genome wide computational analysis [165].

The panel of individual mutations analyzed experimentally caused a reduction in the levels of HA vRNA in virus particles compared to the wild type. This observation further supports the importance of the 5' UTR in packaging of functional IAV virus [157, 158]. Features

beyond hairpin or unfolded conformations in this region may also be important for experimental fitness. While future studies will be required to understand these details further, we note the presence of binding interactions similar to previously reported hairpin interactions between vRNA of segments PB1 and NS [161]. While the correlation between mRNA folding free energy and experimental fitness was not statistically significant, single nucleotide synonymous mutations in the signal sequence may introduce subtle structural changes that are poorly resolved by mRNA folding models and also poorly captured by a thermodynamic metric such as ΔG.

The effects of non-synonymous mutations on virus experimental fitness are generally consistent with the previously established biochemical properties of the signal sequence. Our results support the importance of hydrophobicity in the signal sequence [130, 137], and the importance of small amino acids near the protease cut site [133, 134]. Amino acid changes that introduce charged residues in the hydrophobic region or bulky residues near the protease cut site showed deleterious effects, indicative of structural and functional constraints on the amino acid sequence to ensure binding by the SRP. However, when compared to a region outside the signal sequence (amino acids 32–41), non-synonymous mutations have higher mean absolute experimental fitness effect suggesting that selection at the amino acid level is weaker in the signal sequence relative to this nearby region. This pattern agrees with previous reports that the amino acid sequence of the signal peptide of secretory proteins evolves up to five times faster than other regions [166, 167].

The biochemical underpinnings of selection on the signal sequence appear complex. Interestingly, we find measurable experimental fitness effects of synonymous mutations in the nucleotides encoding the HA signal peptide— a finding supported by the analysis of sequences from natural isolates as well. Our results demonstrate that selection acting on the signal sequence

is due to multiple biochemical pressures, including constraints on RNA secondary structure, RNA packaging efficiency, protein synthesis and trafficking. This type of selection produces a pattern of sequence evolution that is not readily identified by simple models of conservation. Specifically, our findings indicate that patterns of conservation in natural populations only reveal a fraction, perhaps only the tip of the iceberg, of the impacts that synonymous mutations can have on fitness. This result holds important implications for commonly used dN/dS (i.e., the ratio of non-synonymous to synonymous divergence) based statistics used for analyzing the selective pressures facing populations. Namely, the fundamental assumption of these approaches is that synonymous mutations are neutral with respect to fitness, and thus dS represents the rate of fixation owing to genetic drift alone (which is simply equal to the mutation rate). dS is then compared with dN, where a ratio larger than 1 is interpreted as evidence of positive selection on non-synonymous mutations. However, if synonymous mutations are indeed non-neutral as these results suggest, the rate of fixation of synonymous sites is no longer an appropriate proxy for neutrality, potentially resulting in dN/dS ratios that are greatly misleading. Thus, future experimental studies aimed at better defining the impacts of synonymous mutations on pathogen fitness and correlations with RNA structure will be of interest, in particular in pathogens with error-prone polymerases.

Materials and methods

Engineering plasmid libraries

Systematic single nucleotide mutant libraries of the signal sequence of HA were generated in reverse engineered IAV strain WSN/33 as previously described [100, 144]. Briefly, nucleotides encoding amino acid 1 to 1650 of HA were transferred to the pRNDM plasmid [144] and

mutations were introduced using a cassette ligation strategy. In addition to single nucleotide

mutants, we included additional stop codons at an elevated abundance as an internal control in

anticipation of potentially strong selection that might otherwise be difficult to detect. Mutant

libraries were transferred to the pHW2000 plasmid [168] using an engineered destination vector

and Gibson Assembly (New England Biolabs) as described [169].

Cell culture and viral recovery

MDCK and 293T cells were used for viral recovery and passaging as previously described [170].

Viruses were generated using the pHW2000-based 8 plasmid system using plasmids and protocol

developed by Webster and colleagues [168]. Briefly, 1 µg of HA plasmid library and 1 µg of

each plasmid encoding the rest of the A/WSN/33 genome, were transfected in a co-culture of

MDCK and 293T cells using TransIT LT1 reagent (Mirus, Madison, WI). Cell growth media

was changed to serum-free media 6 h post-transfection. TPCK trypsin (Sigma-Aldrich, St. Louis,

MO) was added to a final concentration of 0.5 µg/mL 30 h post-transfection. At 72 h post-

transfection virus-containing supernatant was harvested, centrifuged at 300g for 15 minutes and

stored at -80 °C. Viral titer of the recovered viruses was quantified by plaque assay as previously

described [170]. Briefly, 10-fold serial dilutions of virus-containing supernatant were added to

confluent MDCK cells and allowed to bind for 1 h at 37 °C. Unbound viruses were washed off

with PBS and infected cells were overlaid with 0.5% agar in DMEM and incubated at 37 °C.

Plaques were immunostained with antibody against IAV NP and visually counted 48 h after

infection.

Bulk competition experiments

Competition experiments were conducted as previously described [170]. Briefly, MDCK cells were plated at 10^6 cells per well in 6-well plates the day before infection. Cells were then washed with PBS followed by cDMEM Dulbecco's modified Eagle's medium, 100 U/mL penicillin, 100 µg/mL streptomycin, and 7.5% bovine serum albumin. Recovered viruses were diluted in influenza virus growth medium (cDMEM/ BSA with 1 µg/mL TPCK trypsin) and infections were performed in triplicates at a multiplicity of infection of 0.01. Cells were incubated with virus for 1 h at 37°C. Unbound virus was washed off twice with PBS. Infected cells were grown in influenza virus growth media at 37 °C. Virus-containing supernatant was collected 48 h post-infection and stored at -80 °C.

Quantification of mutant abundances

Mutant abundance was quantified as previously described [170]. Briefly, virus-containing supernatant was treated with DNase (New England BioLabs, Ipswich, MA) for 1 h at 37 °C for degradation of plasmid DNA. Influenza viral RNA was extracted using the QIAamp Viral RNA minikit (Qiagen, Germantown, MD). cDNA was generated using HA-specific RT primers and SuperscriptIII (Life Technologies, Beverly, MA). Sample processing for sequencing was performed as previously described [144]. Briefly, PCR was used to introduce cut sites for ligation of sample-specific adapters. Adapter-ligated samples were quantified, pooled and submitted for analysis of mutant abundance by Illumina single read 36-bp sequencing on a Genome Analyzer II. Reads with greater than 99.5% confidence were analyzed. Relative mutant abundance was calculated as $\log_2(\frac{mutant}{WT})$. Analyzing changes in mutant frequency between the plasmid library and recovered viral population provided an estimate of selection during viral

recovery. Experimental fitness effect in bulk competition was calculated as the difference in relative mutant abundance before and after competition.

Growth analyses of individual mutants

Twenty-two individual mutants (Supplementary Table 2.1) were cloned and recovered as described for mutant libraries. The viral titer for each variant was determined by plaque assay. Images of plaques were acquired using a Nikon SMZ1500 microscope. For each mutant, we used the NIS Elements-BR Analysis program to measure the diameter of 10–20 randomly selected plaques. The average plaque size for each mutant was calculated and utilized as an estimate of the growth rate. Viral infection kinetics for a subset of mutants were determined by infecting MDCK cells at a multiplicity of infection of 0.01, as described earlier, and collecting virus-containing supernatant at 6, 12, 48 and 72 h post infection. Viral titer analyzed by plaque assay was used to measure viral amplification over time in these experiments.

Investigating the relationship between vRNA/mRNA structural variation and observed experimental fitness effects of synonymous substitutions

Predicted Minimum Free Energy (MFE) of RNA secondary structure was used to test the connection between observed synonymous experimental fitness effects and local structural variation in the mRNA/vRNA of mutated HA segments. MFE was calculated for the synonymous HA mutations in a window of size W, sliding across the first 80 nt (30 nt UTR and 50 nt CDS) of the HA mRNA/vRNA sequence. The Pearson correlation coefficient R was used to quantify the relationship between calculated MFE values and measured experimental fitness effects. To assess the statistical significance of this correlation, a random permutation test was

used, permuting the energy ΔG and experimental fitness for each window, window size and position (Supplementary Fig. 2.8). Regions of mRNA/vRNA demonstrating significant ($P \leq$ 0.05) correlation between experimental fitness and MFE were subjected to additional criteria: (1) we set a lower limit of 10 for the number of measured experimental fitness values per folding window, as the number varied with size and starting position of the window (2) standard deviation of predicted MFE values per folding window was limited from below ($\sigma \geq 1$), to ensure sufficient structural variation in the window. Biological relevance of the regions with the significant correlation and both additional criteria satisfied was further investigated. The described analysis was performed independently for mRNA and vRNA sequences, with window sizes W (W=24,36,42,54,62 nt). As described by [125], we used a zero-based numbering of nucleotides in the mRNA and vRNA. The RNAfold program from the ViennaRNA 2 software package [159] was used for RNA folding with the default settings applied.

Alignment of naturally occurring HA sequences

We downloaded 4452 coding sequences (CDS) of HA segment from the NIAID Influenza Research Database (IRD) [147] using only H1N1 strains and excluding overrepresented pandemic H1N1 from the year 2009. HA sequences of sufficient length (> 50 nt) were aligned using TranslatorX [139], which aligns sequences of translated amino acids first and then translates back the amino acids using initial codons. The multiple sequence alignment was further filtered in order to eliminate HA sequences with insertions and deletions in the first 100 nt near the 5' end, resulting in 4374 HA sequences.

Analysis of structural variation of mRNA/vRNA in the naturally occurred HA variants

We focused on nucleotides 6 to 30 of the sequences retrieved from the IRD, a region that appears

biologically relevant according to our experimental and statistical analysis of WSN/33 H1N1

strain. To test if the predicted hairpin was still supported by the naturally evolved sequences we

folded a wider region of vRNA 0-36 nt, while restricting the size of secondary structure to 24 nt.

The RNAfold program from the ViennaRNA 2 software package was used for RNA folding with

the maxBPspan parameter set to 24 nt and otherwise default settings applied otherwise. All

statistical testing and visualization was performed using custom Python scripts, available at

https://github.com/sergpolly/EMPIRIC-HA and upon request.

Measuring HA protein abundance in viral particles

Indirect enzyme-linked immunosorbent assay (ELISA) was used to quantify HA abundance in

cell free viral samples. Mutations $L9L^{TTG}$, $L9L^{CTA}$, $A11A^{GCT}$ and $V13V^{GTC}$ were added to the

above panel for analysis. Recovered viruses were diluted in coating buffer (3.03 g Na_2CO_3; 6 g

$NaHCO_3$; 1 L distilled water; pH 9.6; sterile filtered) and 10^4 plaque-forming units per well were

incubated in 96-well MaxiSorp plates overnight at 4 °C (Sigma-Aldrich, St. Louis, MO).

Unbound virus was removed and wells were blocked with 5% BSA in PBS for 1 h. Primary

antibody, influenza A virus H1N1 HA antibody (Genetex, Irvine, CA. Cat.no GTX 127357), was

added and incubated on a rocker at room temperature for 2 h. Wells were washed and incubated

with sheep anti-mouse IgG HRP-conjugated secondary antibody (General Electric, UK Limited)

for 1 h at room temperature. Wells were washed again and HRP substrate (ThermoFisher

Scientific, Waltham, MA) was added to wells and incubated in the dark at room temperature for

30 minutes. 2 M phosphoric acid was added to stop the reaction. Absorbance was measured at 450 nm on a SpectraMax PLUS plate reader (Molecular Devices, Sunnyvale CA). For each viral stock, three ELISA replicates were performed.

Measuring HA expression in mammalian cells

In order to measure HA expression, the above panel of HA variants was cloned into the JB991 plasmid, kindly provided by J. Bloom (Fred Hutchinson Cancer Research Center, Seattle), that has a GFP reporter, an ampicillin resistance gene, and a CMV promoter. These constructs were transiently expressed in 293T. Cells were washed and re-suspended in isotonic buffer (15 mM MOPS, 145 mM NaCl, 2.7 mM KCl, 4 mM $CaCl_2$ pH 7.4. 2% heat-inactivated FBS added immediately before use). HA primary antibody (Genetex, Irvine, CA. Cat. no GTX 127357) was added to cells and incubated for 1 h at room temperature. Excess antibody was washed off twice with isotonic buffer. Alexa Fluor 647 goat anti-rabbit IgG secondary antibody (ThermoFisher Scientific, Waltham, MA) was added and incubated at room temperature for 1 h. Cells were washed, re-suspended in isotonic buffer and analyzed on a BD FACSCalibur. Data analysis was conducted using the FlowJo data analysis software package (Ashland, OR).

Quantification of vRNA levels

To measure effects of signal sequence mutations on viral RNA, virus-containing supernatant was treated with DNase for 1 h at 37 °C. Total viral RNA was extracted for each mutant variant using the QIAamp Viral RNA minikit (Qiagen, Germantown, MD). In separate reactions, cDNA was generated using HA- and NA-specific RT primers and Superscript III (Life Technologies, Beverly, MA). This cDNA was then used as template in qPCR performed using SYBR Green

and HA- and NA-specific PCR primers (ThermoFisher Scientific, Waltham, MA) on an

Eppendorf MasterCycler RealPlex machine (Eppendorf). Relative HA vRNA abundance was

calculated using the efficiency-corrected approach developed by Pfaffl [171] and normalized to

NA vRNA.

Modeling of the growth parameters of individual mutants

We used a multi-compartment ordinary differential equation (ODE) model [154, 172] to

compare viral yield kinetics of silent HA mutants. In the model, initial infection $V_{PFU}(0)$ spreads

in a population of N susceptible target cells (T), and infects cells at a rate $\beta V_{PFU}(t)$. Infected

cells spend on average τ_E time in an eclipse phase (E) and then, upon switching into infectious

mode (I), start producing virus at a rate p for average duration of τ_I. Newly generated virus

contributes to the total viral pool $V_{PFU}(t)$ that supports the spread of infection, and also degrades

at a rate c. We exclusively considered infectious particles (PFU) in this model, neglecting

$V_{RNA}(t)$ (for which we have no data) in the ODE system (Eq. 1 from [154]) and fit the

experimental data by minimizing the log-scaled sum of squared residuals (SSR) between

experimental measurements and simulated ones for a set of model parameters $\Theta =$

$(\beta, \tau_E, \tau_I, p, c)$:

$$SSR(\Theta) = \sum_{i=1}^{M_{exp}} (\log_{10} V_i - \log_{10} V_{PFU}(t_i, \Theta))^2,$$

where M_{exp} is the number of PFU measurements per mutant and V_i are values of PFU measured

at t_i =6, 12, 24, 48, 72 h. The number of compartments is fixed in our simulations $n_E = n_I =$

20. The size of initial target cell population was $N = 10^6$, with MOI= 10^{-3}. Supplementary

Figure 2.4 shows example of the viral growth measurements and the model fit with optimal parameters. Following Paradis et al. [154], the odeint function from scipy was used to solve the ODE system and the emcee python module was used to perform Markov Chain Monte Carlo (MCMC) simulations to extract the probability distribution of the model parameters. Initially we found an optimal set of parameters $\Theta = (\beta, \tau_E, \tau_I, p, c)$ for each mutant and then performed MCMC simulation using 100 walkers for 1000 steps. Upon convergence, data from step 600 and beyond was used to estimate the probability distribution, see Supplementary Figure 2.7. In order to assess the difference between estimated distributions, A and B, we used p-values derived from a two-sample Z-test,

$$Z = \frac{|\mu_A - \mu_B|}{\sqrt{\sigma_A^2 + \sigma_B^2}},$$

where for a given parameter and mutant, μ is the mean of the distribution and σ is its standard deviation. We used $\log_{10}$ of the β infection rate parameter for placing confidence limits on the estimated model parameters.

Acknowledgments

Tables

Table 2.1. Viral growth kinetics parameters estimated for wild type and 4 silent HA mutants, described by rate of infection of target cells β, and duration of eclipse phase τ_E. Median values and 95% bootstrap confidence intervals (BCI) are shown.

	$Log_{10}\,\beta$, [95% *BCI*]	τ_E, [95% *BCI*]
WT	-6.28, [-6.50 -6.05]	6.16, [5.82 7.64]
L5L^{CTC}	-4.97, [-5.26 -4.69]	20.8, [19.0 22.3]
L6L^{CTC}	-4.85, [-5.10 -4.58]	18.8, [17.1 20.0]
L6L^{TTG}	-6.16, [-6.40 -5.87]	8.82, [7.52 10.9]
L8L^{CTA}	-5.11, [-5.30 -4.92]	23.1, [22.0 24.0]
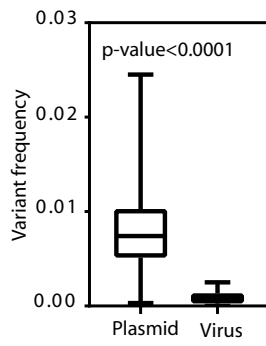
Supplementary Figure 1. Box plot showing the frequency of undetermined variants in the plasmid library and in recovered virus (n=35).
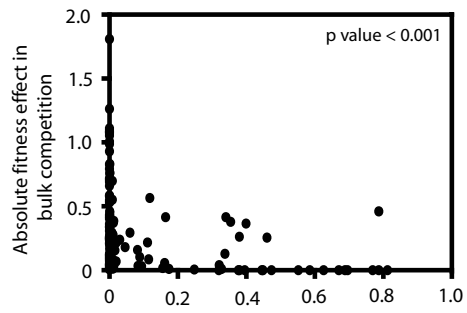
Supplementary Figure 2. Comparing absolute experimental fitness effects with mutant frequency in sequenced circulating viruses.
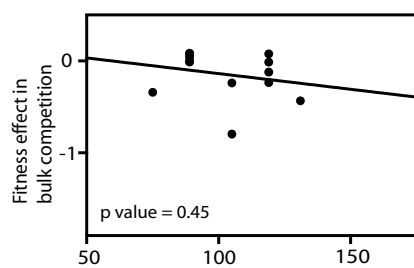
Supplementary Figure 3. Relationship between measured experimental fitness effects and amino acid molecular weight at positions -3 and -1 relative to the peptidase cleavage site excluding the mutation A17E.

```
            32          41
A/WSN/33  TVDTILEKNV
pH1       ....V.....
non-pH1   ....V.....
H2        ....V.....
H3        I....TNDQI
H5        Q....M....
H7        K...LT.RGI
H9        ....LT.TNV
H10       I.K.LTNEQE
```

Supplementary Figure 4. Sequence conservation for the amino acid 32–41 region in different HA subtypes.
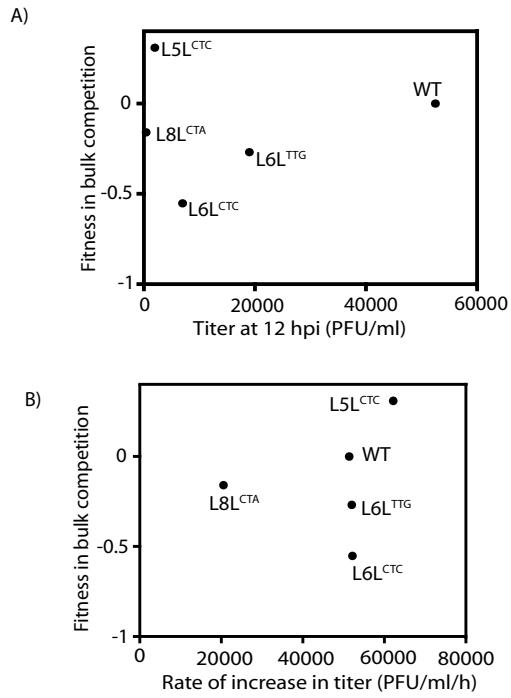
Supplementary Figure 5. Comparing mutant fitness in bulk competition and replication kinetics of individual mutants in isolation. (A) Comparing mutant fitness in bulk competition and viral titer at 12 hpi (pfu/ml) from replication kinetics assay of individual mutant viruses.
(B) Comparing mutant fitness in bulk competition and the rate of increase in titer (pfu/ml/h) of individual mutants.
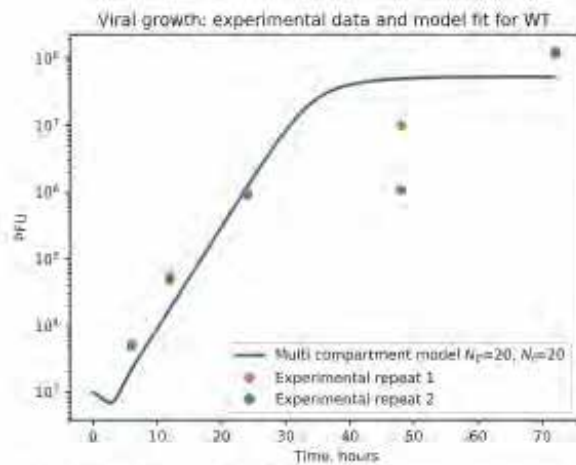
Supplementary Figure 6. Fitting growth measurements of IAV with the 5-parametric model.
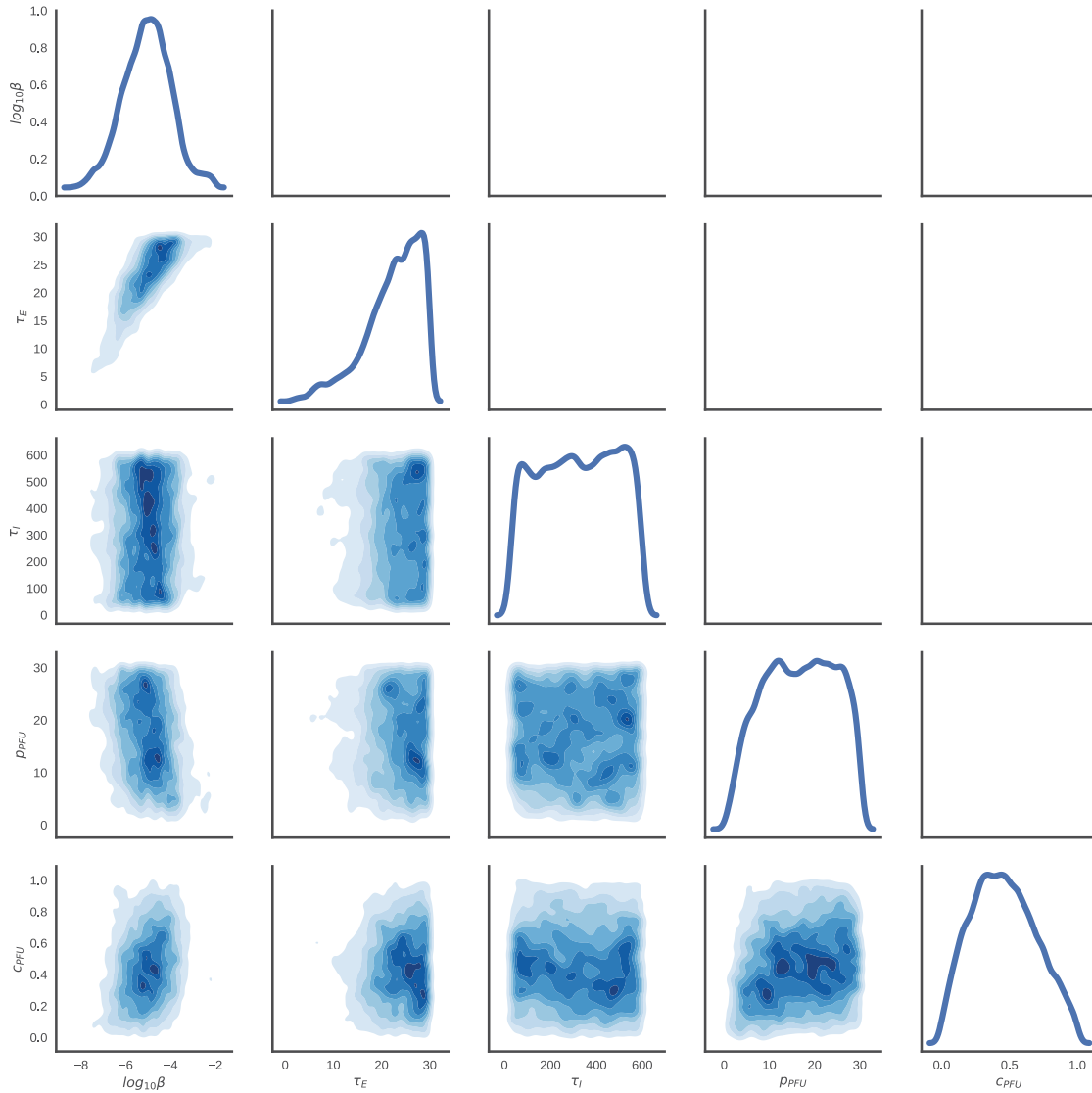
Supplementary Figure 7. Distributions of the kinetic model parameters β, τ_E, τ_I, p, c estimated for the L8L^{CTA} mutant using MCMC simulations of the ODE kinetic model of IAV infection. The MCMC simulation samples the parameter space, generating an estimate of the probability distribution for the parameters that minimize SSR($\Theta = (\beta, \tau_E, \tau_I, p, c)$). The plots on the diagonal represent the probability distributions of the five individual parameters ($\beta, \tau_E, \tau_I, p, c$). Qualitatively, the distributions of the parameters β and τ_E (rate of infection and duration of the eclipse phase, respectively) exhibit a sharp response in the model fit and were selected. The quality of model fit was much less sensitive to the changes of the other three parameters. The off-diagonal panels represent the joint parameter distributions for pairs of parameters, with darker shades of blue corresponding to a higher probability density.
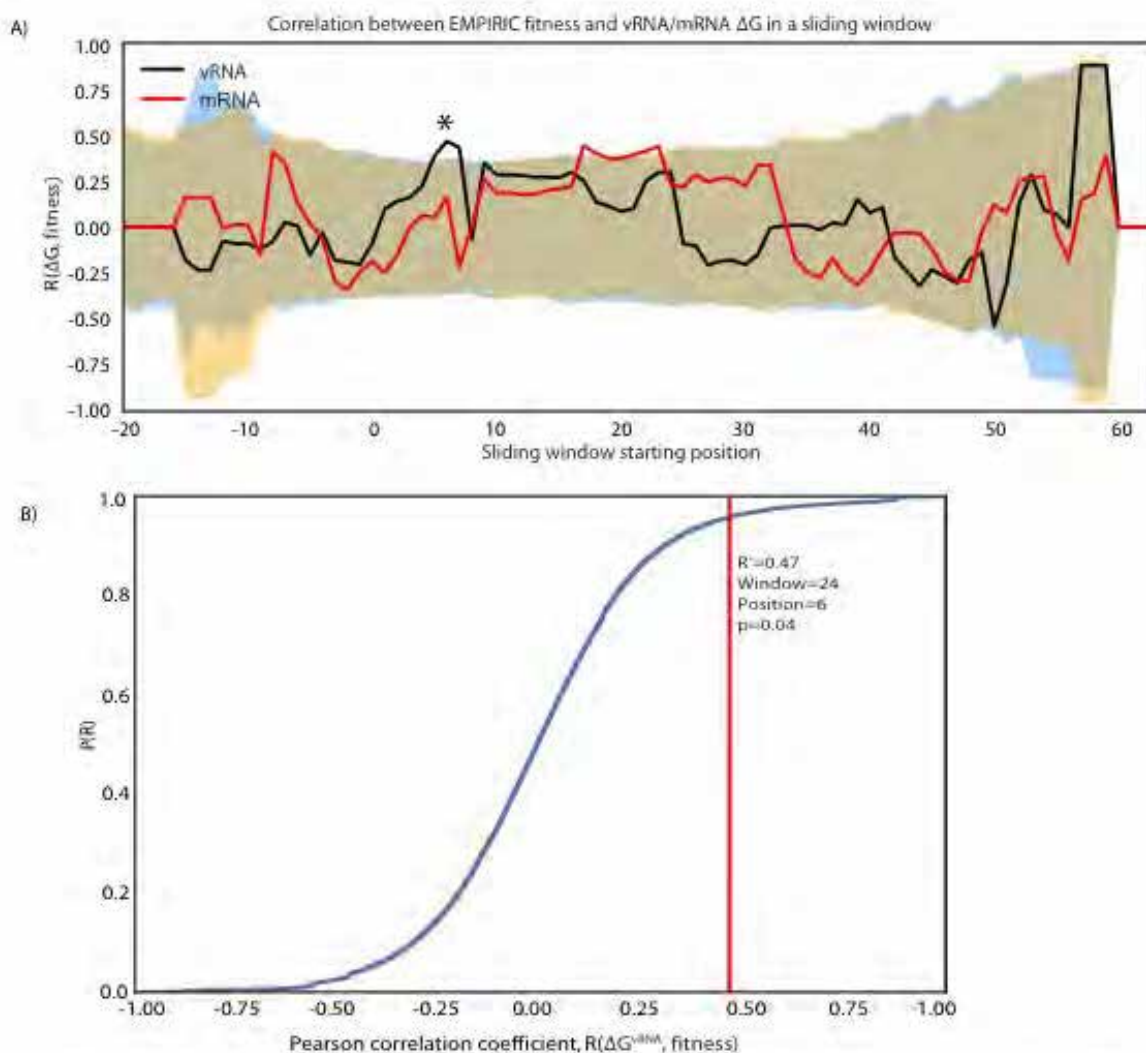
Supplementary Figure 8. Correlation between vRNA/mRNA ΔG and experimental fitness. (A) Folding free energy vRNA/mRNA ΔG of HA vRNA and mRNA in sliding window of size 24 nt (black and red, respectively). The shaded areas (blue for vRNA, tan for mRNA) represent the 98% confidence interval based on permutation tests (see Methods). The asterisk marks a statistically significant window, see below. (B) Cumulative distribution of Pearson correlation coefficient obtained in permutation tests for vRNA (blue). The red line corresponds to the window of size 24 nt beginning at position 6, indicating statistical significance (P=0.04).
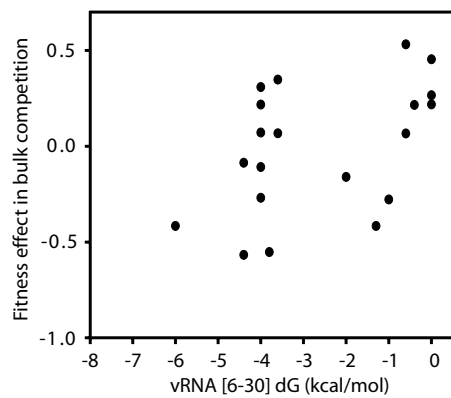
Supplementary Figure 9. Comparing experimental fitness and RNA folding free energy ΔG. vRNA was folded at positions 6-30.
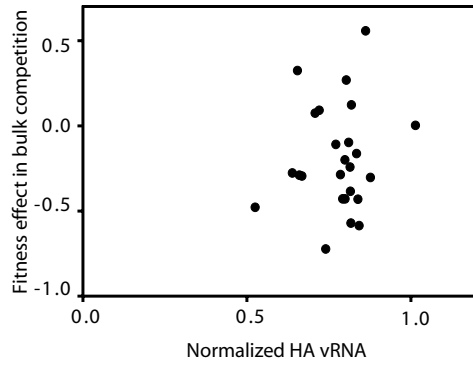
Supplementary Figure 10. Relationship between changes in HA vRNA levels normalized to NA (WT = 1) and experimental fitness measured in bulk competition.

Mutation	Codon
K2K	AAA
A3E	GAA
A3T	ACA
K4R	AGA
L5L	CTC
L5L	CTT
L6L	CTC
L6L	CTT
L6L	TTG
V7V	GTG
L8L	CTA
L8L	TTG
Y10N	AAT
A11A	GCC
A11A	GCG
A11S	TCA
A14A	GCC
T15I	ATA
T15T	ACT
D16H	CAT
A17A	GCG
A17E	GAA

Supplementary Table 1. Panel of mutations analyzed in isolation.

Loop	EMPIRIC fitness	Count in fludb
AUGAC	(wt)0.00	366
ACGAC	0.22	7
AUUAC	ND	3
AGGAC	0.31	1
AAGAC	-0.11	0
AUGAG	-0.55	0
AUGAA	0.07	1
AUAAC	-0.27	0

Supplementary Table 2. Comparison of the 5nt inner loops naturally occurring in HA segment of H1N1 strains with the mutants evaluated in EMPIRIC. ND implies that the loop sequence is absent/non-synonymous in the EMPIRIC scan.

Model	ANOVA *P*-value	R^2
s ~HAexpr		**0.528**
s~HAvir		0.136
s~vRNAratio		0.014
s~ΔGmRNA		0.061
s~ΔGvRNA		0.017
s~HAexpr + HAvirus	**0.014**	**0.680**
s~HAexpr + ΔGmRNA	0.104	0.602
s~HAexpr + ΔGvRNA	0.890	0.529
s~HAexpr + vRNAratio	0.574	0.537
s~HAexpr + HAvirus + ΔGmRNA	**0.053**	**0.752**
s~HAexpr + HAvirus + ΔGvRNA	0.700	0.683
s~HAexpr + HAvirus + vRNAratio	0.787	0.681

Supplementary Table 3. Two-way ANOVA comparison of multiple regression models linking fitness effects *s* to possible explanatory variables: HA protein abundance in viral particles (HAvir), HA expression in mammalian cells (HAexpr), vRNA HA/NA ratio (vRNAratio), and folding free energies ΔG of nucleotides -4 to 37 relative to start codon, of HA vRNA and mRNA (ΔGvRNA and ΔGmRNA, respectively). Among the linear regression models with a single explanatory variable, HA expression in mammalian cells was the best predictor of fitness (R^2=0.528). ANOVA P-values for nested, pairwise comparisons of models with two or three explanatory variables are shown, comparing the best model using *n* explanatory variables to all models including *n*-1 variables, *n*=2,3. Best models are highlighted in bold.

Chapter III: Modulating influenza neuraminidase function influences the fitness landscape of hemagglutinin

This manuscript is currently under preparation for submission as

Aneth S. Canale, Troy W. Whitfield, Daniel R. Caffrey, Wayne A. Marasco, Celia A. Schiffer, Timothy F. Kowalik, Jeffrey D. Jensen, Robert W. Finberg, Konstantin B. Zeldovich, Jennifer P. Wang[#], and Daniel N. A. Bolon[#]. "*Modulating influenza neuraminidase function influences the fitness landscape of hemagglutinin*".

Abstract

Influenza A virus is an important pathogen that causes hundreds of thousands of infections annually and continues to pose a great threat to global public health. The interaction of influenza A virus with the host receptors is mediated by the viral proteins hemagglutinin (HA) and neuraminidase (NA). HA recognizes and binds to sialic acid to initiate infection. NA recognizes and cleaves sialic acid interactions to prevent virus aggregation at the cell surface and release newly synthesized virus. Since HA and NA interact with the same molecule on the host cell surface but have opposing effects, a balance in receptor binding and receptor destroying activities of HA and NA is critical for efficient virus replication. This functional balance has been implicated in cross-species transmission of IAV, yet the molecular mechanism remains unclear. In the present study, we used the Extremely Methodological and Parallel Investigation of Randomized Individual Codons (EMPIRIC) approach to systematically investigate how changes in NA activity influence the fitness landscape of HA. We observed that inhibiting NA function by mutation did not rescue mutant HA. However, using the inhibitor oseltamivir enabled the rescue of HA mutations that were deleterious under wild type levels of NA function. Identifying oseltamivir-responsive HA mutations highlights the role of the HA-NA functional balance in the development of resistance to NA inhibitors.

Introduction

Influenza viruses are negative-sense, single-stranded RNA viruses that belong to the Orthomyxoviridae family. This family of viruses is composed of three main genera, influenza A, influenza B, and influenza C [9]. Most human influenza infections are caused by influenza A virus (IAV) [26]. IAV has two main surface glycoproteins – hemagglutinin (HA) and neuraminidase (NA) that have counteracting roles in the virus cycle [18, 19]. During infection, HA of human IAV preferentially recognizes and binds to alpha 2,6-linked sialic acid on host cell glycans, while avian IAV preferentially binds to alpha 2,3-linked sialic acid [173]. Following the replication of the viral genomic RNA and packaging of newly synthesized viral proteins, NA cleaves sialic acid interactions to release nascent virus. The differences in recognition and binding patterns between HA and sialic acid on host receptors and the corresponding receptor cleavage by NA form the basis of host-specificity in IAV infections. Because HA and NA recognize the same molecule but have opposing functions, balanced receptor-binding strength by HA and receptor-destroying activity by NA is essential for optimal viral replication (Fig. 3.1a). Eighteen HA and 11 NA subtypes have been identified to date [13, 26]. Without any constraints, 198 HA-NA combinations are expected to circulate in natural influenza. However, only a fraction of these subtypes has been isolated thus far. Reassortment studies conducted in the laboratory help to explain this observation [174, 175]. These studies showed that not all HA-NA combinations result in viable viruses. Kaverin et al. showed that combining a human H1N1 variant with different mammalian and avian viruses resulted in viruses with a wide array of infectivity. For instance, reassortants with avian HAs (H3, H4, H10, and H13) paired with human NA resulted in viruses that replicated poorly in embryonated chicken eggs, suggesting mismatch in HA and NA function [176]. These observations indicate that a balance in the levels

of HA and NA function is essential for viral infectivity, and is the determining factor in the emergence of new HA and NA combinations in circulating viruses. However, the functional constraints underlying genomic reassortment in IAV are poorly understood.

Other studies have shown that matching levels of receptor binding and receptor destroying activities of HA and NA are critical for productive infection by IAV. By manipulating NA properties, studies have shown that truncating the length of the NA stalk results in reduced NA function ultimately leading to viral attenuation in embryonated eggs [63, 177]. However, attenuated viruses appeared to be rescued by two mechanisms: the first was by insertions in NA that restored NA stalk length, and the second was by accumulating mutations in HA [63, 178]. Mapping HA mutations with rescue potential to structure showed that these mutations were mostly located near the receptor binding site, suggesting that reduced receptor binding established a new HA-NA functional balance. In another study, altering HA receptor binding strength by manipulating the N-linked glycans near the receptor site also resulted in attenuated viruses [179, 180]. In this study, optimal viral replication was only observed when truncated NA was matched with weak binding HA, or when full length NA was matched with stronger receptor binding. While complementary levels of HA and NA function allowed virus rescue, the limits of functional rescue under selection pressure remain understudied.
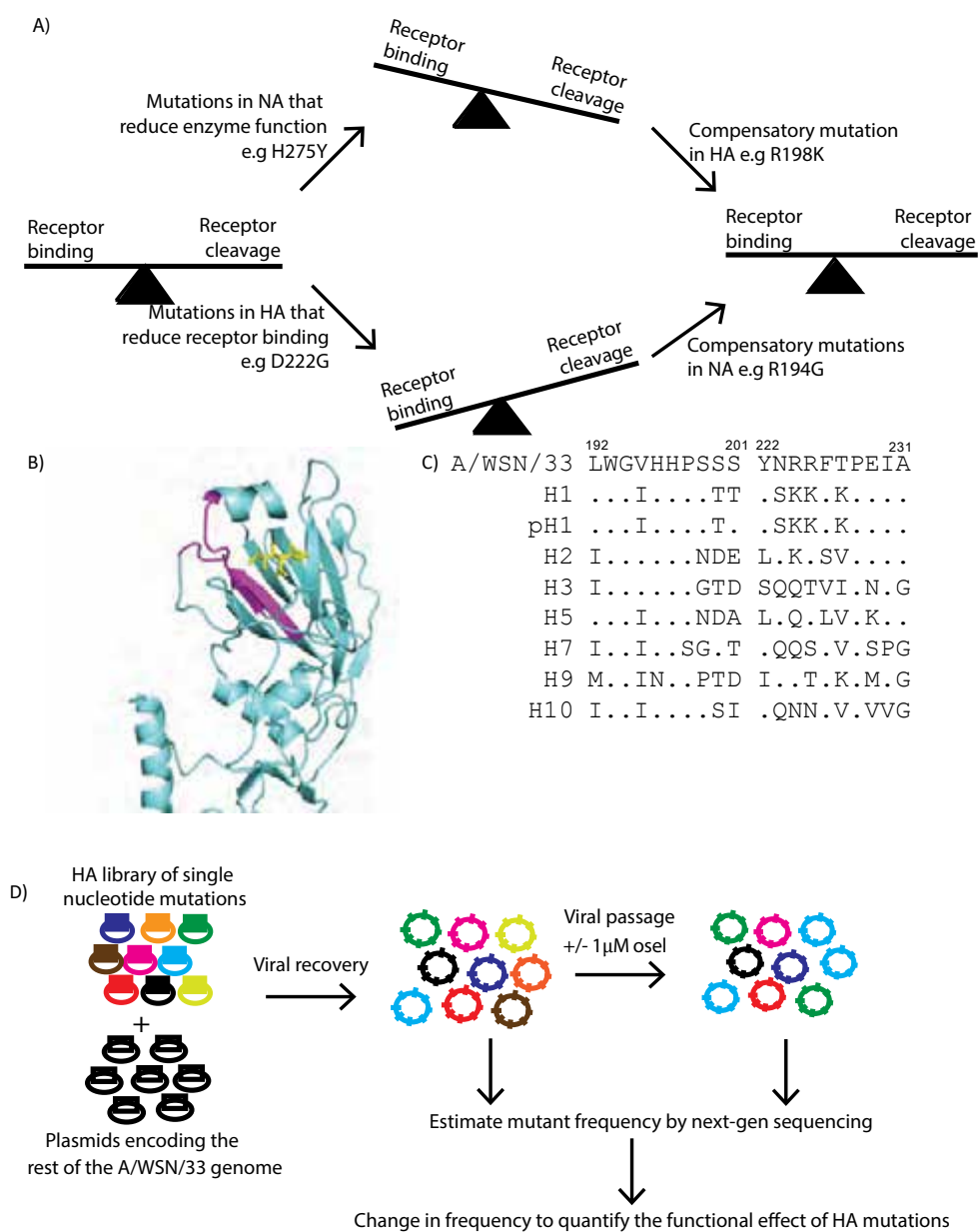
Figure 3.1. Using the EMPIRIC approach to study the role of HA-NA functional balance in IAV replication. A) Illustration of the match in receptor binding strength by HA and receptor cleavage by NA that is essential for optimal viral replication. B) Mapping the region covered in the current study (aa 192-210 and 222-231 in magenta) onto the HA structure (PDB 1HGG). Sialic acid is shown in yellow. C) Sequence conservation of the region under study over different HA subtypes. Consensus sequences were generated from 1243 pandemic H1, 1315 non-pandemic H1, 598 H2, 1506 H3, 1585 H5, 1770 H7, 1168 H9 and 925 H10 sequences retrieved from the Influenza Research Database D) Illustration of the experimental approach utilized to systematically study HA-NA functional balance.

Oseltamivir is a competitive NA inhibitor that inhibits the budding of progeny virus, and is widely used to treat influenza infections [181, 182]. However, oseltamivir-resistant viruses emerged in 2007 to 2008 and continue to undermined the reliability of oseltamivir in the treatment of influenza infections [48, 183]. Characterization of oseltamivir-resistant viruses revealed a high prevalence of the H275Y mutation (N1 numbering system) in NA [184, 185]. This mutation is located near the active site of NA and confers resistance by reducing oseltamivir binding affinity [186, 187]. HA mutations that confer resistance to NA inhibitors both in the presence and absence of the H275Y mutation have been identified [66, 188, 189]. These mutations exhibited reduced receptor binding strength, yet their role in the IAV evolutionary trajectory is not fully appreciated. Understanding the HA-NA functional balance will help define the evolutionary constraints shaping IAV evolution and provide better understanding of interspecies transmission and the development of drug resistance.

Taken together, these studies provide substantial evidence that the HA-NA functional balance is critical for productive infection by IAV and for viral adaptation to new environment. However, the degree to which IAV can tolerate a functional imbalance between receptor binding and receptor destroying while remaining infectious, and the molecular mechanism guiding the reinstatement of this functional balance are not well understood. Understanding the molecular factors governing the HA-NA functional balance under different environmental conditions will define what drives influenza evolution, potentially identify additional anti-influenza therapeutic targets, and help minimize the emergence of drug resistance.

In the present study, we adapted the Exceedingly Meticulous and Parallel Investigation of Randomized Individual Codons (EMPIRIC) approach that utilizes high-throughput mutational scanning in combination with bulk competition assays [90, 100, 118] to systematically

investigate how different NA perturbations affect the fitness landscape of HA. We modulated NA activity by mutation, by adding the NA inhibitor oseltamivir, and also used a combined mutation and drug pressure approach. We observed that inhibiting NA function by mutation alone was not enough to rescue HA mutations with reduced receptor binding strength. However, inhibition by oseltamivir alone or by NA mutation in combination with oseltamivir supported the rescue of HA mutations. Our results suggest that a threshold in NA function must be surpassed for the HA-NA tradeoff and rescue in virus function to occur.

Results

Measuring fitness effects of mutations in the HA receptor binding site

We studied the effects of all possible single-nucleotide mutations in two 10-amino acid regions covering the receptor binding site of HA (Fig. 3.1b). We chose the regions to include the highly conserved histidine, histidine and proline residues at positions 196-199 described to directly interact with the host receptor (Fig. 3.1c). Viruses that were recovered using a previously described approach [100] were used to infect MDCK cells and next-gen deep sequencing was used to measure mutant frequency in the viral pool before and after bulk competition (Fig. 1d). Experimental fitness effects were calculated relative to wild type ($s = 0$) and stop codons ($s = -1$). Fitness measurements were reproducible between bulk competition replicates ($R^2 = 0.89$, Fig. 3.2a). Stop codons were consistently depleted during the competition and wild type synonyms clustered around wild type-like fitness (Fig. 3.2b). Many mutations introduced in this region caused a fitness defect under wild type NA (Fig. 3.2c). This observation is consistent with the importance of this region in binding to host receptors to initiate infection.
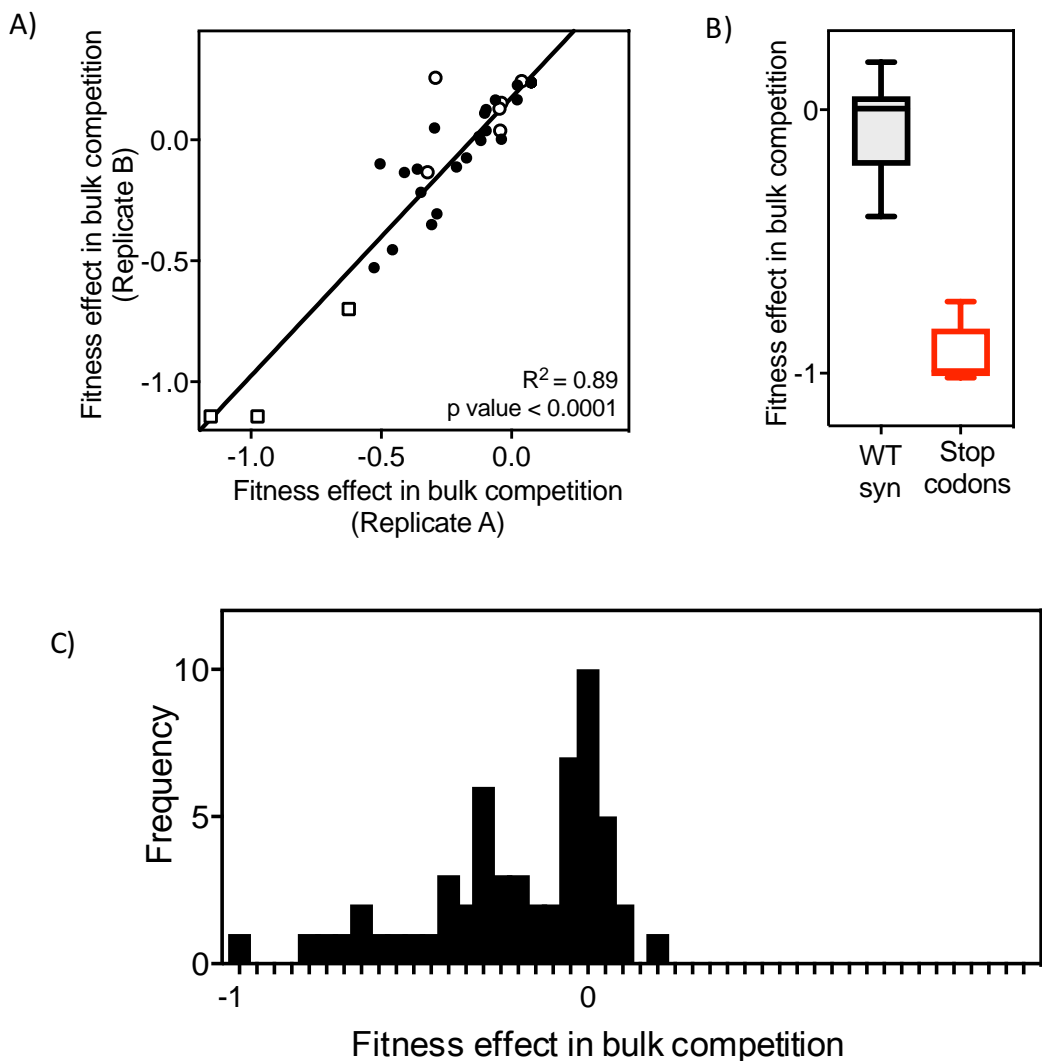
Figure 3.2. Mutations at the HA receptor binding site impact virus fitness. A) The EMPIRIC approach provides a reproducible approach to measure experimental fitness effects ($R^2 = 0.89$). Synonymous substitutions are shown as open dots, non-synonymous substitutions as black dots, and stop codons as open squares. B) Fitness effects of wild type synonyms cluster near wild type-like fitness ($s = 0$), while stop codons cluster near $s = -1$. C) Distribution of fitness effects (DFE) of mutations at positions 192-201 and 222-231.

Fitness effects of HA mutations in the presence of the NA inhibitor oseltamivir

Oseltamivir is a NA inhibitor that binds to the NA active site and reduces enzyme activity thus inhibiting the release of progeny of virus. Oseltamivir remains the most common targeted therapy against influenza infections, despite the described resistance. Even with the widespread use, the effect of oseltamivir on the HA-NA functional balance remains poorly understood. To systematically investigate the effects of oseltamivir on HA mutations, we measured the experimental fitness effects of HA mutations in the presence of drug pressure from the NA inhibitor oseltamivir (Fig. 3.1d).

We observed that even though oseltamivir does not bind directly with HA, this drug greatly influenced the fitness landscape of HA. Inhibiting NA function by oseltamivir allowed the identification of mutants that appeared to grow better than wild type virus ($s > 0$) under drug pressure (Fig. 3.3a). Fig. 3.3b shows the effects of oseltamivir on the distribution of fitness effects. Mutations that exhibited adaptive potential in the presence of oseltamivir had a wide distribution of fitness effects in the absence of drug pressure, ranging from wild type-like to highly deleterious (Fig. 3.3c).

Mapping mutations with the highest adaptive potential in the presence of oseltamivir to structure shows that the most adaptive mutations cluster at an unstructured loop at the neck of the receptor binding pocket (Fig. 3.3d).

van der Waals contacts between sialic acid and HA at the receptor binding site

In order to investigate the role of molecular interactions at the receptor binding site in the observed fitness effects, MD simulations were performed on a A/WSN/33 HA model bound to sialic acid. Mapping average vdW energy to structure showed that the strongest interactions were

with residues at the bottom of the receptor binding pocket (W166 at -5.6 kcal/mol; T149 at -5.4 kcal/mol) (Fig. 3.4a&b). Residues that bond the tightest to sialic acid were highly conserved between different HA subtypes (Fig. 3.4c). Of the 20-amino acid region randomized in this study, residues 196 – 200 were shown to interact with sialic acid (Fig. 3.4a). Mutations at these positions were generally deleterious under wild type levels but showed improved fitness and even adaptive potential under altered NA activity (Supplementary Fig. 3.1). This observation suggests that mutations that alter the optimized receptor binding strength were rescued under reduced levels of NA function.

Experimental fitness effects correspond with patterns in sequenced isolates

For these experiments, we used the 8-plasmid system to recover and study the H1N1 strain A/WSN/33. Due to differences in the environmental conditions and evolutionary constraints in natural infections compared to experimental conditions, we analyzed the relevance of our experimental results relative to natural evolution of IAV HA. We compared our experimental fitness measurements to the amino acid frequency of each mutation in a panel of 4452 non-redundant sequences extracted from the Influenza Research Database (see Methods). We observed a significant correlation between experimental fitness estimates and mutant frequency in sequences of circulating variants (Fig. 3.5a, $P = 0.02$). Mutations that were highly abundant in circulating strains had wild type-like fitness in our bulk competition assay, whereas mutations that had fitness defects in our assay were rarely observed in circulating viruses. Observed differences maybe be due to sequence differences between our model A/WSN/33 strain and circulating viruses (Fig. 3.1c), as well as other factors such as epistasis between HA residues outside the region covered in the current study.
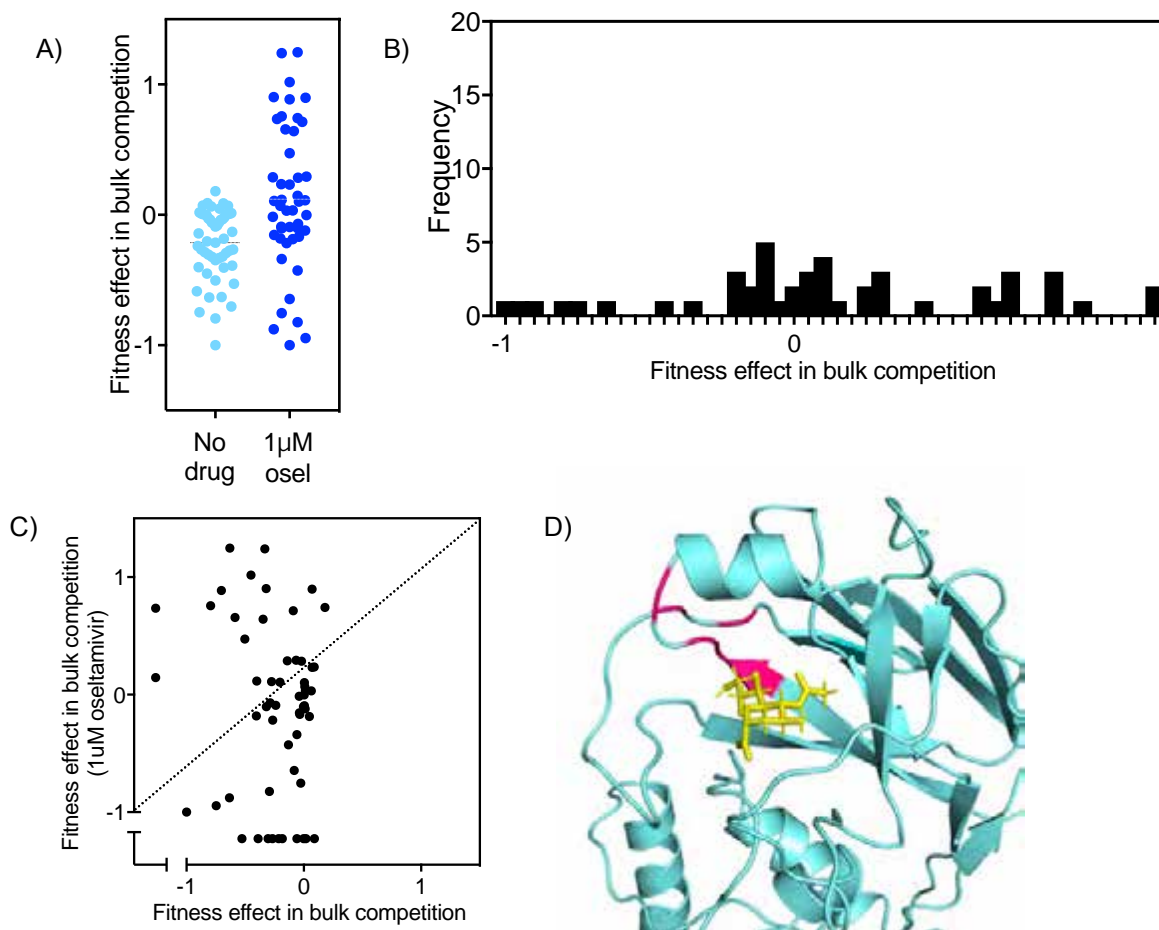
Figure 3.3. Modulating NA function using the inhibitor oseltamivir influences the fitness landscape of HA mutations. A) Scatter plot showing experimental fitness effects of HA mutations under wild type levels of NA function (light blue) and under inhibition by 1 µM oseltamivir. B) Oseltamivir alters the DFE of HA mutations and enables the rescue of mutant virus. C) Correlation between mutant fitness in the presence and absence of drug pressure. Oseltamivir responsive mutations deviate from the diagonal. D) Mapping HA mutations with the highest rescue potential to structure (PDB 1HGG). These mutations cluster on an unstructured loop region at the neck of the receptor binding pocket.
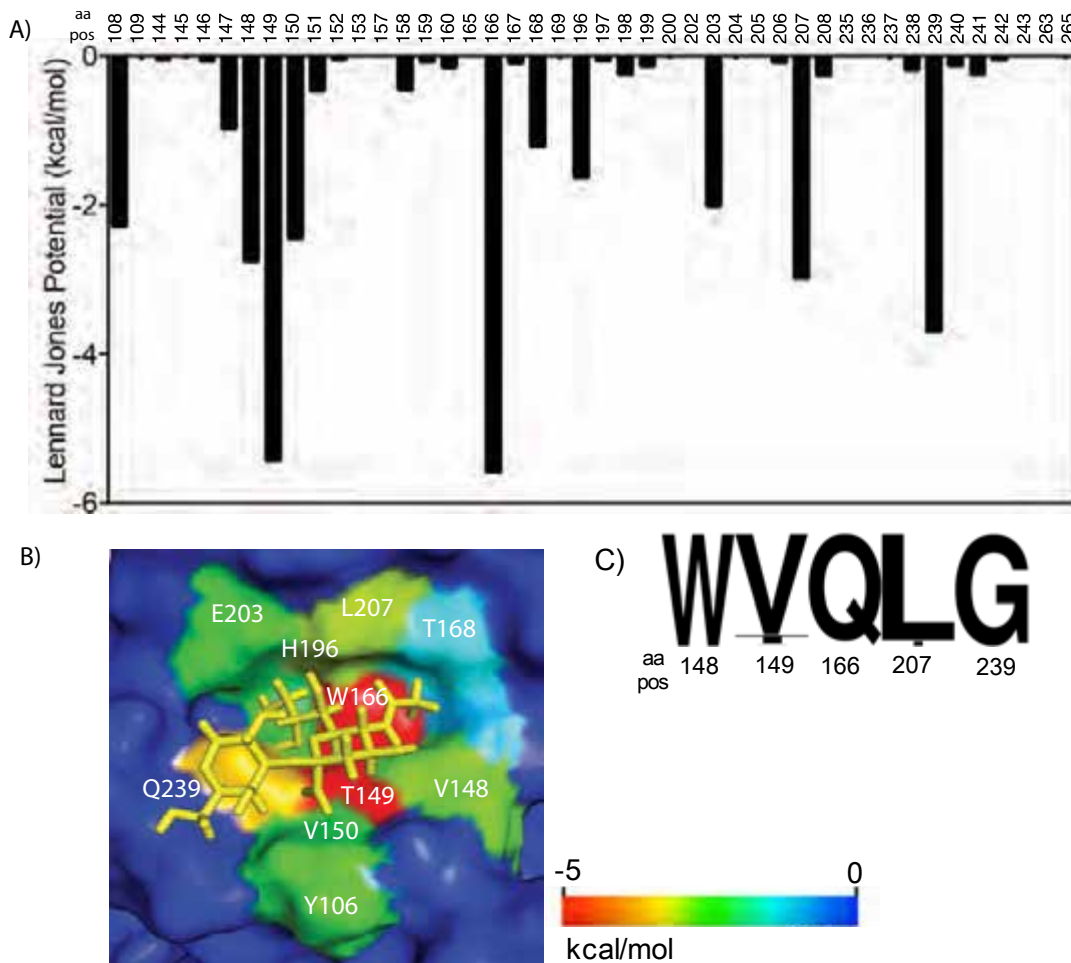
Figure 3.4. Using MD simulations to study van der Waals contacts between sialic acid on host receptor and HA residues at the receptor binding site. A) HA residues that interact with sialic acid (cut-off 0.05 kcal/mol). B) Mapping van der Waals energies to the model structure of A/WSN/33 HA. Red shows strongest interactions while blue shows weak interactions. Strongest interactions were observed at the bottom of the receptor binding pocket while the neck of the binding pocket, which is less structured, had weak interactions. C) Sequence conservation of HA residues that exhibited the strongest van der Waals interactions with sialic acid. Sequence logo was created from 2355 H1 sequences deposited from 2007 to 2009.

Measuring the fitness of HA mutants under selection pressure by oseltamivir showed that oseltamivir influences the fitness landscape of HA. In order to investigate the relevance of the effects of oseltamivir observed in our bulk competition study we analyzed 2355 H1 sequences retrieved from the Influenza Research Database from the years 2007 – 2009 when oseltamivir was widely used to treat influenza infections. We observed a switch in amino acid preference at positions 195, 199, 200, and 201 near the receptor binding sites, as well as at positions 223, 224, and 225 distal from the binding site (Fig. 3.5b&c) and that may coincide with the onset of widespread use of oseltamivir. Our experimental fitness measurements also show that lysines at positions 224 and 225 were deleterious under wild type levels of NA activity but became beneficial in the presence of oseltamivir.

Biochemical properties of amino acids partly explain observed fitness effects

We used the BLOSUM62 substitution matrix to investigate the role of amino acid similarity in the observed fitness effects (Supplementary Fig. 3.2a). Amino acid properties explained about 15% of the observed fitness effects under wild type NA ($P = 0.007$). Substitution to a biochemically similar amino acid generally resulted in minimal fitness defects, while substitution to a dissimilar amino acid resulted in a wide distribution of fitness effects. However, in the presence of selection pressure by oseltamivir, this pattern disappears (Supplementary Fig. 3.2b). This suggests that other factors, such as binding interactions and the local environment, may also influence fitness under stress conditions.
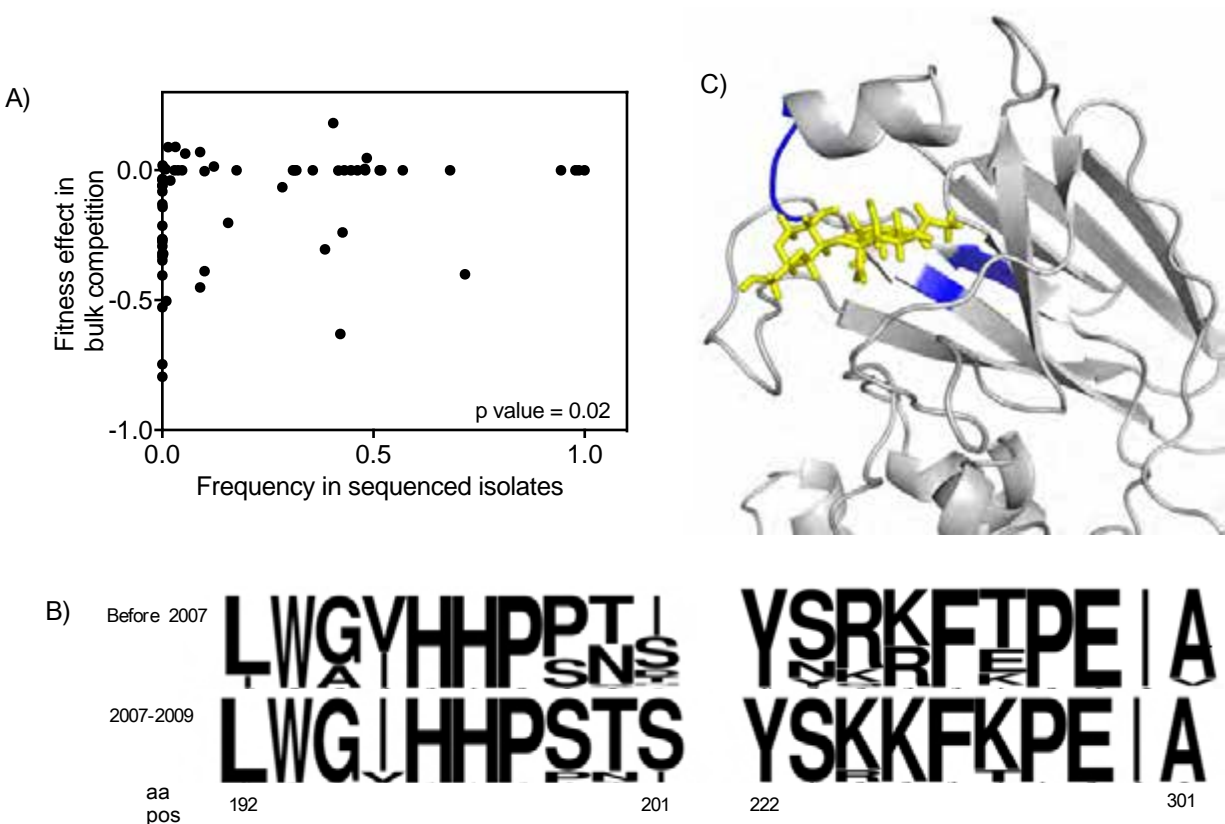
Figure 3.5. Experimental fitness effects measured in bulk competition correspond with patterns seen in circulating viruses. A) A significant correlation exists between experimental fitness measurements and mutant frequency observed in sequenced circulating isolates ($P = 0.02$). B) Comparing HA sequences before the widespread use of oseltamivir and sequences isolated in 2007-2009. Sequence variations are observed at positions 195, 199, 200, 201, 224, 225, and 227. C) Mapping regions with sequence variation to structure (PDB 1HGG). Some of these mutations cluster on the unstructured region of the receptor binding pocket, while residues 224, 225, and 227 are distal from the receptor binding pocket and do not interact with sialic acid.

Analyzing positional tolerance to mutation

We studied the effects of all possible single nucleotide mutations for a 20-amino acid region at the receptor binding pocket of HA. Calculating the average fitness effect of all mutations at each position revealed the sensitivity of each position to mutations (Supplementary Fig. 3.2c). We observed that most positions in this region were sensitive to substitutions. However, most positions had improved fitness in the presence of oseltamivir with positions 200 and 228 showing the greatest fitness advantage under drug pressure. Residues 194, 195, 196, 226 and 231 showed minimal response to oseltamivir. Analysis of sequenced isolates showed that residues 194, 195, and 196 are highly conserved in all HA subtypes (Fig. 3.1c) and have been shown to interact with sialic acid [190]. Positions 226 and 231 appear to only be conserved in H1 HA and have not been implicated in receptor binding (Fig. 3.1c).

Modulating NA function by mutation influences the HA fitness landscape

Previous studies have shown that a functional balance in receptor binding strength by HA and receptor cleavage by NA is important for productive infectivity by IAV [179, 191]. However, a systematic and comprehensive investigation of this functional balance has not yet been performed. In order to study the molecular mechanism of the HA-NA functional balance, we measured the experimental fitness effects of HA mutations under different levels of NA function (Fig. 3.6a). We studied effects of modulating NA activity by introducing one of the following mutations: I223M, H275Y, Y276F or N295S (Fig. 3.6b). We chose these mutations because of their relevance in oseltamivir resistance in circulating viruses [192, 193].

Perturbing NA activity by mutation had minimal effect on the fitness landscape of HA (Fig. 3.6c). We observed that the distribution of fitness effects had a bimodal distribution with a big peak near the fitness effect of the NA mutation and another peak at null fitness. Inhibiting NA function by mutation failed to identify HA mutations with adaptive potential, as seen in the presence of oseltamivir (Fig. 3.6c, d, e, f). Although present, beneficial mutations under mutant NA background have modest fitness benefits. Interestingly, there were no overlap between adaptive mutations in oseltamivir and adaptive mutations under mutant NA (Supplementary Figs. 3.3 & 3.4). This trend suggests that there may be two different mechanisms to adaptation when modulating NA function by drug compared to modulation by mutation. However, comparing HA fitness under different NA mutations showed some correlation (Supplementary Fig. 3.5), indicating that these mutations may have similar effects on the fitness landscape of HA.

Combined effects of mutation and drug pressure on NA has rescue potential

We analyzed the effects of different NA perturbations on the HA fitness landscape. Our results show that the means of modulating NA function does influence the resulting fitness landscape. The mutations we have used in our study impact virus growth by 25-30%. We used oseltamivir concentrations that inhibit virus growth by 50% under wild type conditions. As such, we were interested in studying the combined effects of these pressures on virus function (Fig. 3.6a). We previously observed that NA mutations alone had limited potential to identify adaptive HA mutations (Fig. 3.6c, d, e, f). However, under the influence of both mutation and drug pressure, we identified more mutations with even greater adaptive potential than in either condition alone (Fig. 3.7a). Analyzing the correlation between different NA backgrounds in the presence of oseltamivir revealed common adaptive mutations (Fig. 3.7b). For the NA mutations
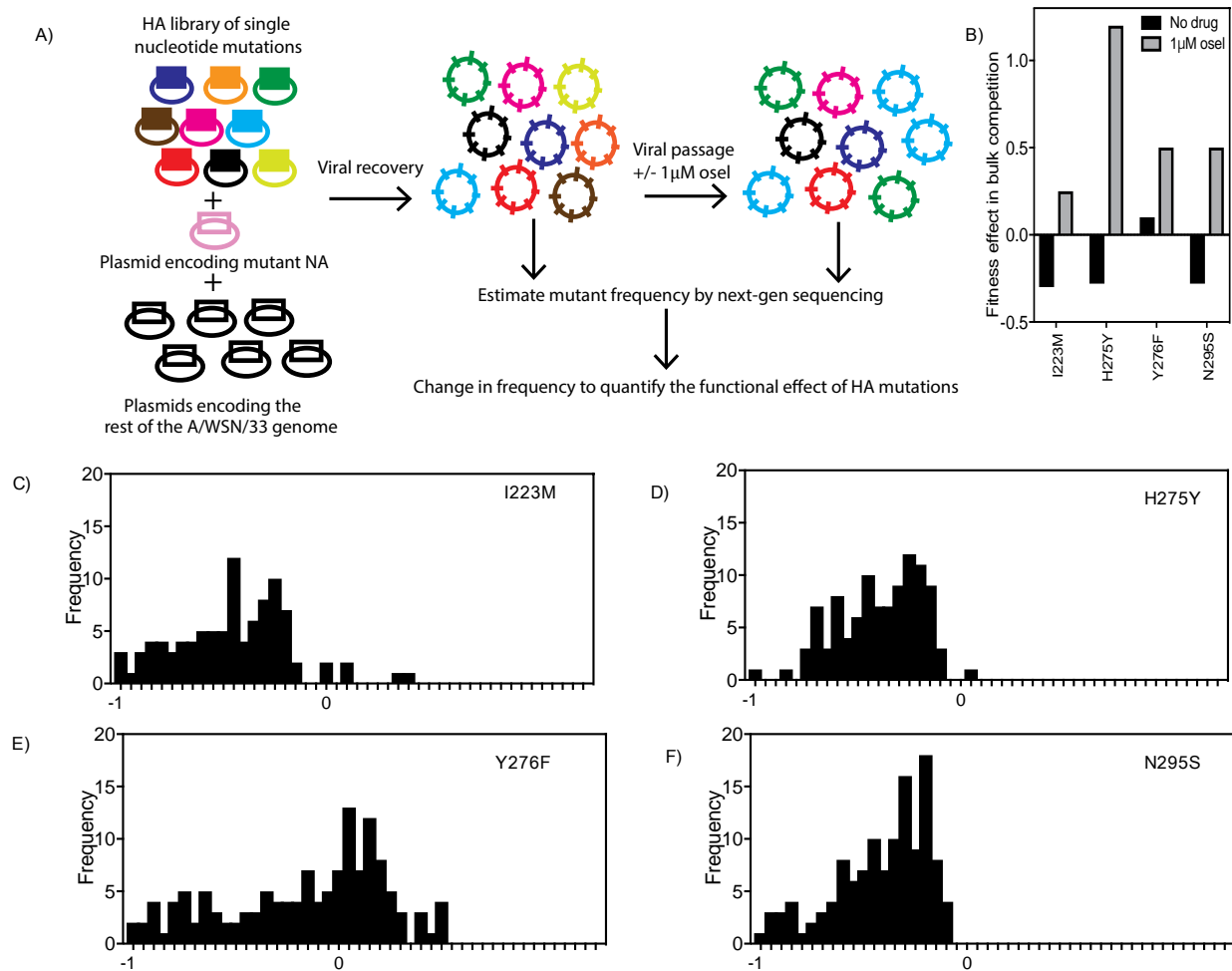
Figure 3.6. Modulating NA function by mutation shifts the DFE towards deleterious effects. A) Modifying the EMPIRIC approach to study effects of NA mutations on the fitness landscape of HA. B) Fitness effects of the NA mutations used in the study in the presence and absence of oseltamivir [100]. C-F) DFE of HA mutations under mutant NA backgrounds.

H275Y and N295S, oseltamivir appears to rescue mutations with wild type-like fitness, whereas under I223M and Y276F oseltamivir appears to mostly rescue deleterious mutations (Supplementary Fig. 3.6).

Positional sensitivity to oseltamivir under varying levels of NA function

In order to further investigate HA sensitivity to oseltamivir we compared the average fitness of all HA mutations for each position across different levels of NA function conferred by either mutation or by the combined effect of mutation and oseltamivir. We observed that under mutant NA, the oseltamivir insensitivity at the receptor binding site largely disappears (Supplementary Fig 3.7). This may indicate that inhibiting NA activity by both mutation and drug pressure may eliminate the need for the previously optimized strong binding to the receptor.

Validating bulk fitness measurements by studying individual clones in isolation

In order to validate experimental fitness effects measured in bulk competition, we measured the fitness of a panel of 22 individually cloned HA mutants (Table 3.1). For further analysis in isolation we chose mutations that were deleterious under wild type levels of NA function but were generally beneficial at non-wild type NA levels. Under unmodified levels of NA function, all of the mutants analyzed had smaller plaques compared to wild type (average radius = 123.66 $\pm$ 5.915 µm), except for the mutant N223I (Fig. 3.8a & b) (average radius = 129.37 $\pm$ 4.14 µm). In the presence of oseltamivir, all viruses had much smaller plaques (average radius = 16.3 µm), including the wild type (average radius = 11.48 $\pm$ 2.58 µm). Interestingly, in
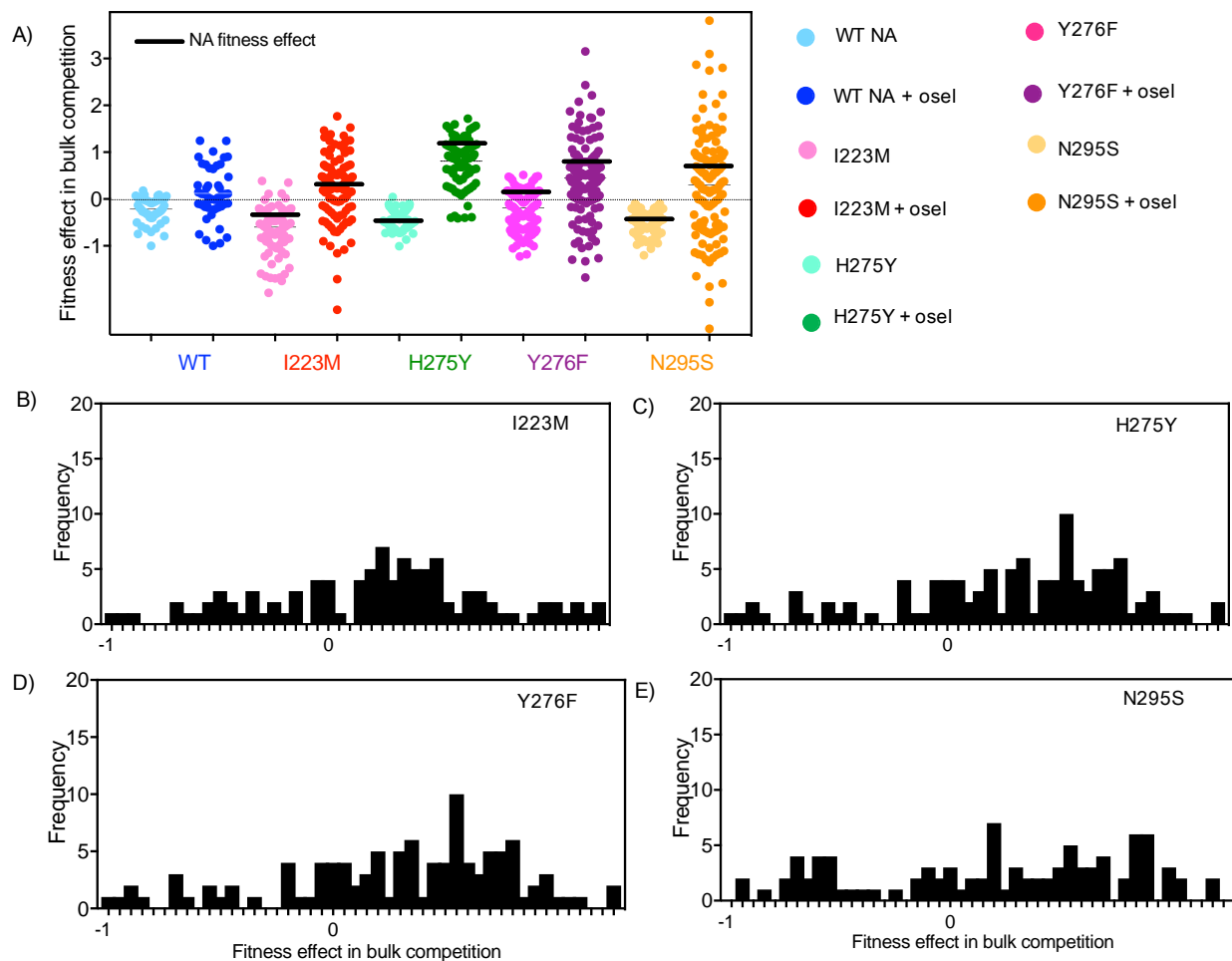
Figure 3.7. Combined effects of NA mutations and oseltamivir result in the rescue of HA mutants. A) Summary of the experimental fitness effects of HA mutants under wild type and mutant NA both in the absence (light hues) and presence (dark hues) of oseltamivir pressure. B-E) DFE of HA mutations under the combined pressure of NA mutations and oseltamivir.

the presence of oseltamivir, viruses carrying the N295S NA mutation had plaques that were more than double the size of wild type viruses under oseltamivir (average radius = 24.17 ± 3.2 μm) (Fig. 3.8c). We observed high reproducibility between our fitness measurements in bulk competition and mutant fitness measured in isolation ($R^2 = 0.82$ Fig 3.8d). Fitness effects were dependent on the environmental condition tested (see Fig 3.8e).

Rescued HA mutations exhibit reduced receptor binding

Functional balance in the receptor binding strength of HA and the sialidase activity of NA are important for the initiation and propagation of infection by IAV. Mutations that offset this balance have been shown to confer a fitness defect [63, 66]. However, previous studies have shown that over time new mutations in the virus will lead to the reinstatement of the HA-NA functional balance. As such, we speculate that inhibiting wild type NA function by using oseltamivir allows the identification and rescue of HA mutations with reduced receptor binding strength to match the reduced levels of NA function.

We used the hemagglutination assay to measure receptor binding normalized to virus titer for 5 mutations H197N, H197Y, S200N, N223Y and P228A under wild type and mutant NA (Fig 3.9a). We calculated receptor binding potential as the ratio between virus titer (pfu/ml) and hemagglutination units (HAU). Weak binding corresponds with a high pfu/HAU ratio, and vice versa. All mutations exhibited reduced receptor binding strength relative to wild type virus, except for the H197N mutation. There was no clear trend between receptor binding strength, calculated as pfu/HAU, and experimental fitness (Fig 3.9b, p value > 0.05). We compared
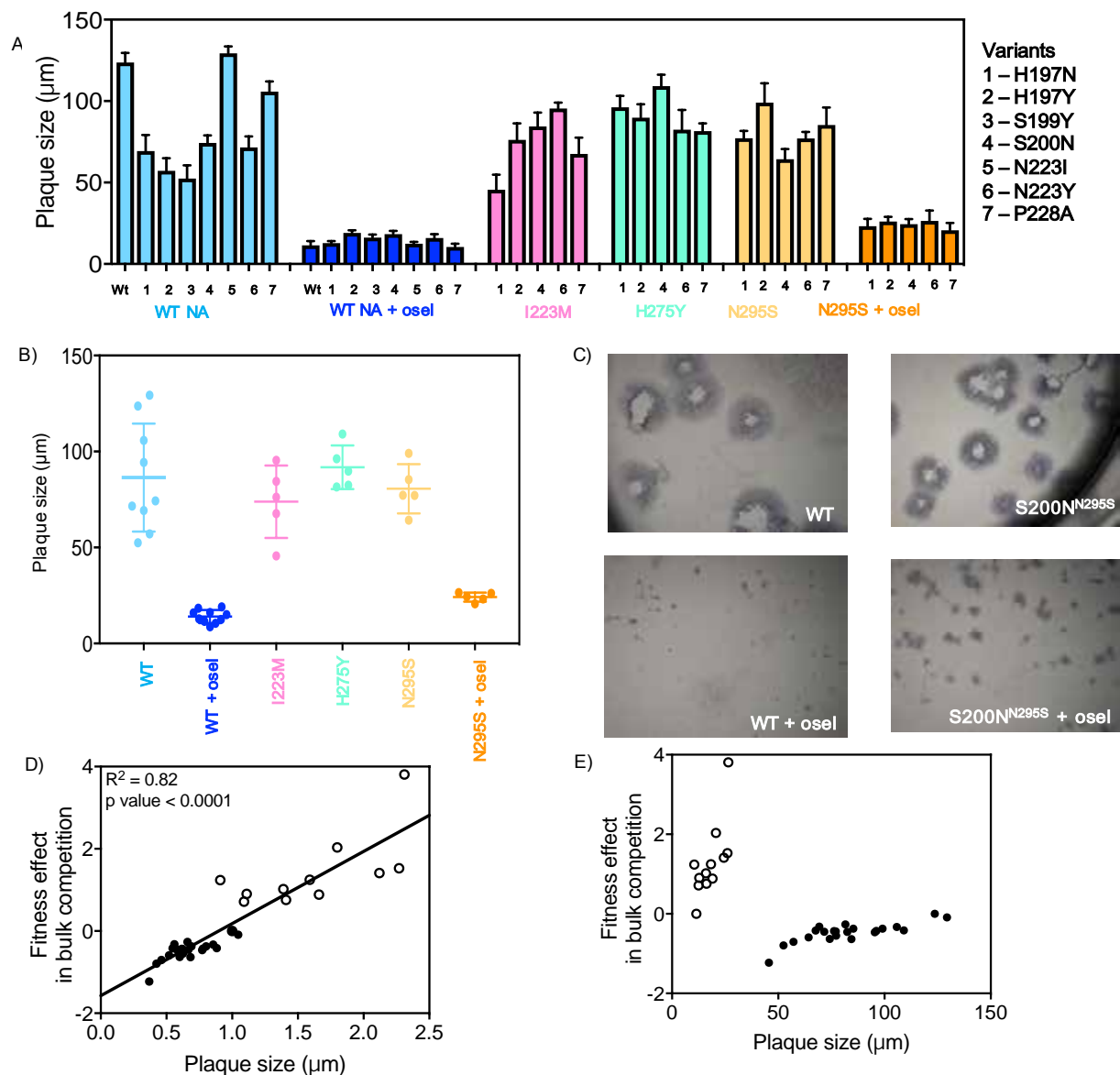
Figure 3.8. Validation of bulk fitness measurements by studies on individual IAV clones. A) Plaque size measurements of individual mutants in isolation. In blue are plaque size of HA mutants under wild type NA in the absence (light blue) and presence (dark blue) of oseltamivir. Plaque sizes were also measured under I223M, H275Y and N295S NA mutations. The panel of mutations carrying the N295S mutation was studied in the absence and presence of oseltamivir. B) Distribution of plaque sizes based on NA background. Color legend as in panel A. C) Images of plaques of wild type virus in the absence and presence of oseltamivir, and for the S200N mutation under N295S mutation in the presence and absence of oseltamivir. D) Results show high reproducibility between experimental fitness effects in bulk competition and plaque size normalized to wild type. Plaque size measurements in the absence of oseltamivir are shown as black dots and measurements in the presence of oseltamivir as shown as open dots. E) Plaque size measurements in the presence (open dots) and absence (black dots) of drug pressure.

receptor binding strength and NA activity previously measured by the enzymatic MUNANA assay (Fig 3.9c). Our results are inconclusive due to the number of data points available, but the trend suggests that HA mutations under wild type NA shift the HA-NA functional balance away from wild type, thus conferring a fitness defect. However, under the N295S NA mutation we see a modest increase in virus fitness at new levels of HA-NA ratios (Fig 3.9d). Further studies will help elucidate the functional constraints governing the observed balance in HA-NA function and the functional constraints governing this pattern of rescue.

Mutant plaque size corresponds with rescue potential

We speculated that when NA function is perturbed, the tradeoff in HA-NA function is condition-specific and only comes into play when a certain threshold in NA function is crossed. In order to test this hypothesis, we analyzed the number of adaptive HA mutations under each NA condition (Fig 3.10). Inhibiting NA function by mutation alone may not have reached the functional threshold to trigger the rescue of HA mutations. Inhibiting NA function by oseltamivir and by combining oseltamivir with mutation surpassed the threshold and enabled the rescue of HA mutations.
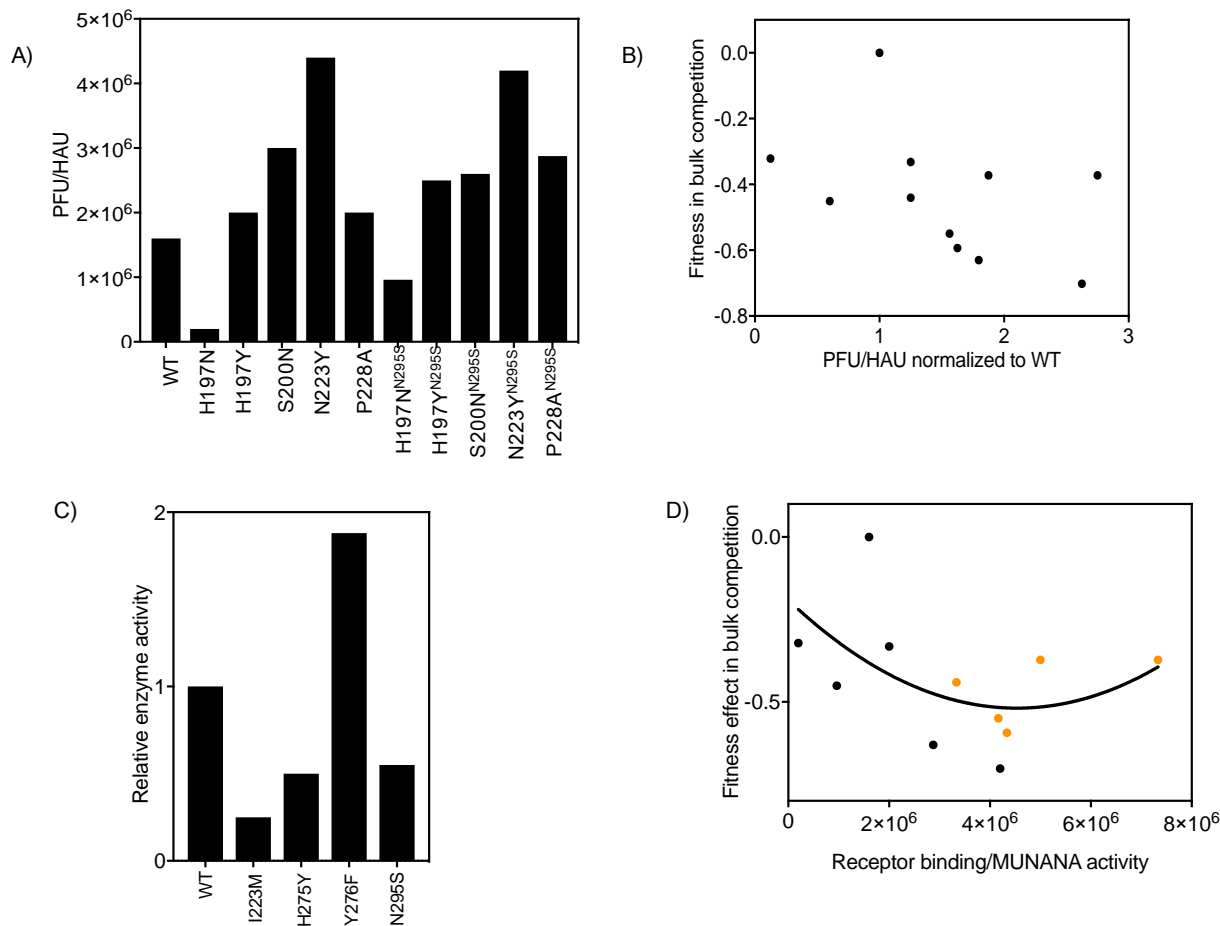
Figure 3.9. Mutations near the receptor binding site influence receptor binding. A) Receptor binding potential represented by the ratio between viral titer and hemagglutination units. HA mutations were analyzed under wild type and N295 NA backgrounds. B) The correlation between fitness measured in bulk competition and receptor binding potential was not significant (p value > 0.05). C) Previously quantified enzyme activity of NA mutations used in this study [100]. D) Relationship between mutant fitness measured in bulk competition and HA-NA activity represented by the ratio between receptor binding potential and NA enzyme activity measured by MUNANA.
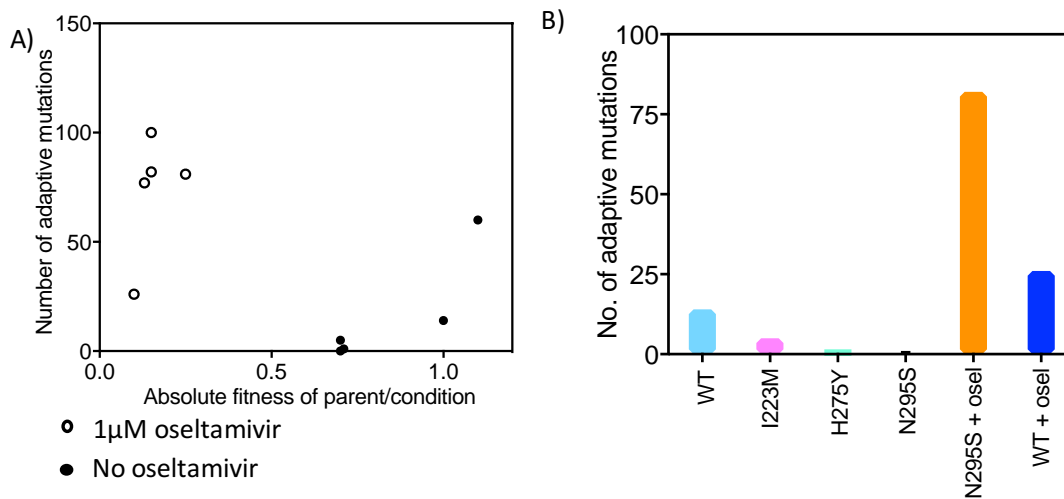
Figure 3.10. Relationship between NA activity and the rescue of HA mutations. A and B) Number of adaptive HA mutations correspond with the absolute effect of NA activity on virus function.

Discussion

We adapted the EMPIRIC approach to systematically investigate the functional balance between receptor binding by HA and receptor cleavage by NA. Results from bulk competition assays show that most mutations at the HA receptor binding site impact virus fitness. This observation agrees with previous findings that disturbing the HA-NA functional balance leads to reduced virus replication [179, 194, 195]. We validated the observed fitness effects by studying the fitness of a panel of individually cloned mutations in isolation. Mutant fitness effects were highly reproducible between bulk competition measurements and in studies on individual mutants in isolation. Experimental fitness effects measured by bulk competition were consistent with patterns of mutant frequency observed in sequenced circulating variants, suggesting that the EMPIRIC approach captures the constraints present in nature.

Previous studies have shown that compatible levels of HA and NA function are important for virus function. We speculated that modulating NA activity, either by using NA inhibitors or by introducing mutations, will lead to the rescue of HA mutants with reduced function. To test this hypothesis, we investigated how changes in the level of NA activity affect the fitness landscape of HA. In the presence of oseltamivir, we observed the rescue of HA mutations that had reduced fitness at wild type levels of NA function, indicating that HA-NA functional balance is more important than HA and NA activity levels for virus function [176, 195]. We identified oseltamivir-responsive HA mutations that may be involved in maintaining the HA-NA functional balance. These mutations appear to compensate for the reduced NA function thus allowing virus growth in the presence of oseltamivir. Some of these variants were observed in HA following the introduction of oseltamivir in 2007.

Investigating the rescue potential of mutant NA revealed that inhibiting NA function by mutation had minimal potential to rescue HA mutations under standard conditions. This observation indicates that the means of modulating NA activity does affect the resulting fitness landscape of HA. However, examining the combined effects of inhibiting NA function by mutation and by oseltamivir revealed greater rescue potential when compared to inhibition by oseltamivir alone. The improved fitness of the NA mutations used in this study in the presence of oseltamivir may have contributed to the increased rescue potential seen under drug pressure, but the mechanism is unclear.

In order to examine the limitations in HA rescue, we analyzed the relationship between the level of NA function and the number of HA mutations rescued. There was a sigmoidal non-linear relationship between NA activity, measured by mutant plaque size, and the potential to rescue HA mutations. This relationship suggests that reducing NA activity will only rescue HA mutations once a certain threshold level of NA activity is surpassed, after which the rescue potential plateaus. Further studies will be needed to further characterize this rescue pattern and to explain the role of this rescue pattern in cross-species transmission and infectivity.

HA-NA functional balance is often altered by genetic drift, antiviral agents that inhibit NA function, and during adaptation to a new host following cross-species transmission of IAV [60, 196]. Due to the high mutation rate inherent in IAV, the virus establishes a new balance that allows continued virus replication. Understanding the functional constraints shaping the re-establishment of the HA-NA functional balance can help inform the control of cross-species transmission of IAV, help to combat the ongoing problem of resistance to oseltamivir inhibitors, and help inform the design of more effective therapeutic agents.

Materials and methods

Engineering plasmid libraries

Single nucleotide mutant libraries were generated, as previously described, for a 20-amino acid region (aa 192-202 and 222-231) surrounding the receptor binding site of HA [90]. Briefly, nucleotides encoding amino acid 1 to 1650 of IAV strain WSN/33 HA were transferred to the pRNDM plasmid and mutations were introduced using a cassette ligation strategy. We included additional stop codons at an elevated abundance as an internal control for each 10-amino acid window to allow detection even in the presence of potentially strong selection. Mutant libraries were transferred to the pHW2000 plasmid using an engineered destination vector and Gibson Assembly (New England Biolabs) as described [100].

Virus recovery and quantification

The pHW2000-based 8-plasmid system developed by the Webster group was used to recover viruses [15]. Briefly, MDCK and 293T cells were cultured and passaged as previously described [100]. Virus recovery was performed using 1 µg of HA plasmid library and 1 µg of each plasmid encoding the rest of the A/WSN/33 genome. These plasmids were transfected in a co-culture of MDCK and 293T cells using TransIT LT1 reagent (Mirus, Madison, WI). Transfection media was changed to serum-free media 6 h post-transfection. At 30 h post-transfection TPCK

trypsin (Sigma-Aldrich, St. Louis, MO) was added to the serum-free media to a final concentration of 0.5 µg/mL. Virus-containing supernatant was harvested at 72 h post-transfection, centrifuged at 300g for 15 minutes, aliquoted and stored at -80ºC. Viruses were

recovered in two genetic conditions: with wild type NA or with either one NA mutation (I223M, H275Y, Y276F or N295S).

Plaque assays were performed to quantify viral titer of the recovered virus as previously described [197]. Briefly, virus-containing supernatant was diluted in a 10-fold serial dilution and added to confluent MDCK cells. Virus was allowed to bind to MDCK cells for 1 h at 37°C followed by washing off of unbound viruses with PBS. Infected cells were overlaid with 0.5% agar in DMEM and incubated at 37°C for 48 h. Plaques were stained with antibody against IAV NP and visually counted.

Growth competition experiments in cell culture

Bulk competition experiments in culture were conducted as previously described [100]. Briefly, MDCK cells were plated in 6-well plates at 10^6 cells per well about 24 h before infection. Cells were washed with PBS right before infection, followed by cDMEM Dulbecco's modified Eagle's medium, 100 U/mL penicillin, 100 µg/mL streptomycin, and 7.5% bovine serum albumin. Viruses were diluted in influenza virus growth medium (cDMEM/ BSA with 1 µg/mL TPCK trypsin) and used to infect MDCK cells at a multiplicity of infection of 0.01, in triplicates, either in the presence or absence of 1 µM oseltamivir. Cells were incubated with virus for 1 h at 37°C. Unbound virus was washed off with PBS. Infected cells were grown in influenza virus growth media in the presence or absence of oseltamivir at 37°C. At 48 h post-infection, virus-containing supernatant was harvested, centrifuged at 300g for 15 minutes and stored at -80°C.

Quantification of mutant frequency

Mutant frequency following bulk competition was quantified as previously described [100]. Briefly, harvested supernatant was treated with DNase (New England BioLabs, Ipswich, MA) for 1 h at 37ºC in order to degrade plasmid DNA. Viral RNA was extracted using the QIAamp Viral RNA minikit (Qiagen, Germantown, MD) and eluted in 40μl of elution buffer. HA cDNA was generated using Superscript III (Life Technologies, Beverly, MA) and HA-specific RT primers. Sample processing for sequencing was performed by using PCR to introduce restriction sites, followed by digestion and ligation of sample-specific adapters. Adapter-ligated samples were quantified, pooled and sequenced on the Illumina single read 36-bp sequencing on a Genome Analyzer II. Reads with greater than 99.5% confidence were analyzed. The relative abundance of each mutant was calculated as $\log_2$(mutant/WT). Fitness effect in bulk competition was calculated as the difference in relative mutant frequency before and after competition.

Validation of mutant fitness in clonal isolates

A panel of 22 individual mutations were cloned and virus recovery was performed as described above. The titer for each recovered viral variant was determined by plaque assay with minor modifications. Briefly, virus was diluted either in the presence or absence of 1 μM oseltamivir and incubated for 1 h at 37ºC. In the drug-treated condition, 1 μM oseltamivir was also added in the agar overlay. Plaques were visualized after 48 h and images were acquired using a Nikon SMZ1500 microscope. The NIS Elements-BR Analysis program was used to measure the radius of about 10-15 plaques from the same quadrant in each well. The average plaque size for each mutant was calculated and used as an estimate of viral growth.

Using molecular dynamics simulations to study molecular interactions

Molecular dynamics simulations were performed to calculate van der Waals contacts between

sialic acid and HA at the receptor binding site. A homology model of A/WSN/33 HA was

created using the SWISS-MODEL modelling server using the 1918 HA (PDB 1HGG) as the

template. Heteromultimer assembly and ligand docking was performed using the Prime Structure

Prediction panel on Schrodinger [198]. Molecular dynamics simulations were performed as

previously described [199]. Briefly, the modeled structures were prepared for simulation using

the octahedron solvent box. Desmond was used for all simulations using the OPLS2005 force

field. All simulations were performed for 10 ns.

Alignment of naturally occurring HA sequences

We retrieved 4452 HA sequences from the Influenza Research Database (IRD) [146] containing

only H1N1 strains and excluding overrepresented pandemic H1N1. These sequences were

aligned and filtered to eliminate HA sequences with insertions and deletions near the 5' end,

resulting in 4374 HA sequences. The frequency of each codon at each position was calculated

and compared to fitness data from our bulk competition experiment.

Hemagglutination assays to measure receptor binding

Human red blood cells were isolated by centrifuging serum at 875 rpm for 15 minutes. The

supernatant was carefully removed and the pellet containing isolated red blood cells was washed

three times with 10 ml PBS. A round-bottom 96-well plate was set up with duplicates of 2-fold

serially diluted virus to a total of 50 µl. A negative control containing PBS only was also

included. 50 µl of diluted red blood cells were added to each well and mixed carefully. Plates

were incubated at 4ºC for 45 minutes. Plates were visualized using a magnifying mirror. Wells

with uniform red color were considered positive while wells with a red button at the center of the

well were considered negative for hemagglutination. Resulting hemagglutination units were

normalized to viral titer.

Measuring NA activity *in vitro*

Measurements of NA enzyme activity were previously performed in [100] using the fluorogenic

2′-(4-methylumbelliferyl)-α-D-N-acetylneuraminic acid (MUNANA) substrate following

instructions provided by the manufacturer (Life Technologies, Carlsbad, CA). Briefly,

recombinant viruses were incubated with MUNANA substrate at a final concentration of 0.1 mM

for 1 h at 37 °C with shaking. Fluorescence was then measured using a Victor X5 plate reader

(PerkinElmer, Waltham, MA) with a 355 ± 10 nm excitation filter and a 460 ± 20 nm emission

filter. Relative fluorescence was normalized to virus titer determined by plaque assay.

Variant
H197N
H197Y
S199Y
S200N
N223I
N223Y
P228A
H197N^{I223M}
H197Y^{I223M}
S200N^{I223M}
N223Y^{I223M}
P228A^{I223M}
H197N^{H275Y}
H197Y^{H275Y}
S200N^{H275Y}
N223Y^{H275Y}
P228A^{H275Y}
H197N^{N295S}
H197Y^{N295S}
S200N^{N295S}
N223Y^{N295S}
P228A^{N295S}

Table 3.1. Mutants analyzed as individually cloned variants.
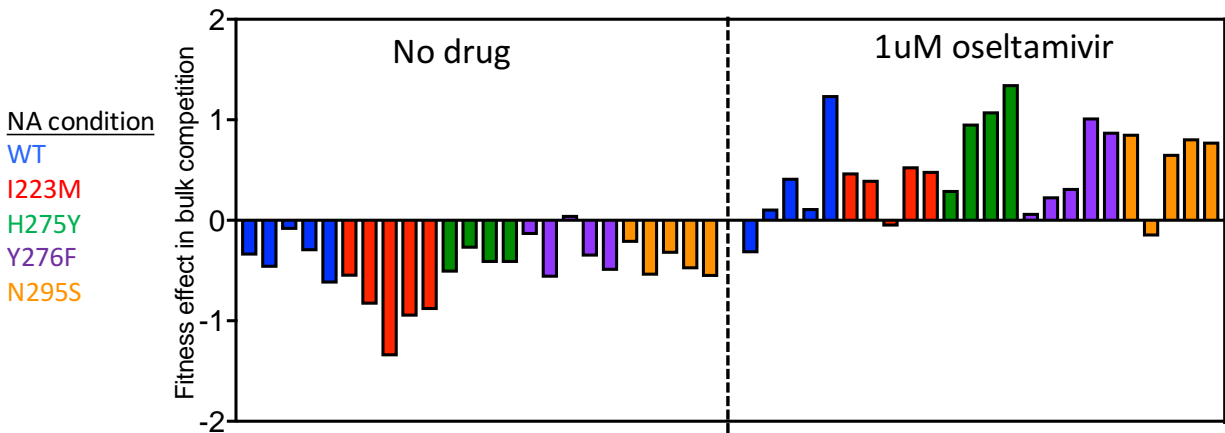
Supplementary Fig 3.1. Average experimental fitness effects of mutations at HA residues that interact with sialic acid that were covered in the present fitness scan. Mutations at these residues were deleterious under standard conditions (no drug) in wild type and in mutant NA backgrounds. In the presence of oseltamivir, mutations at these positions rescued virus function.
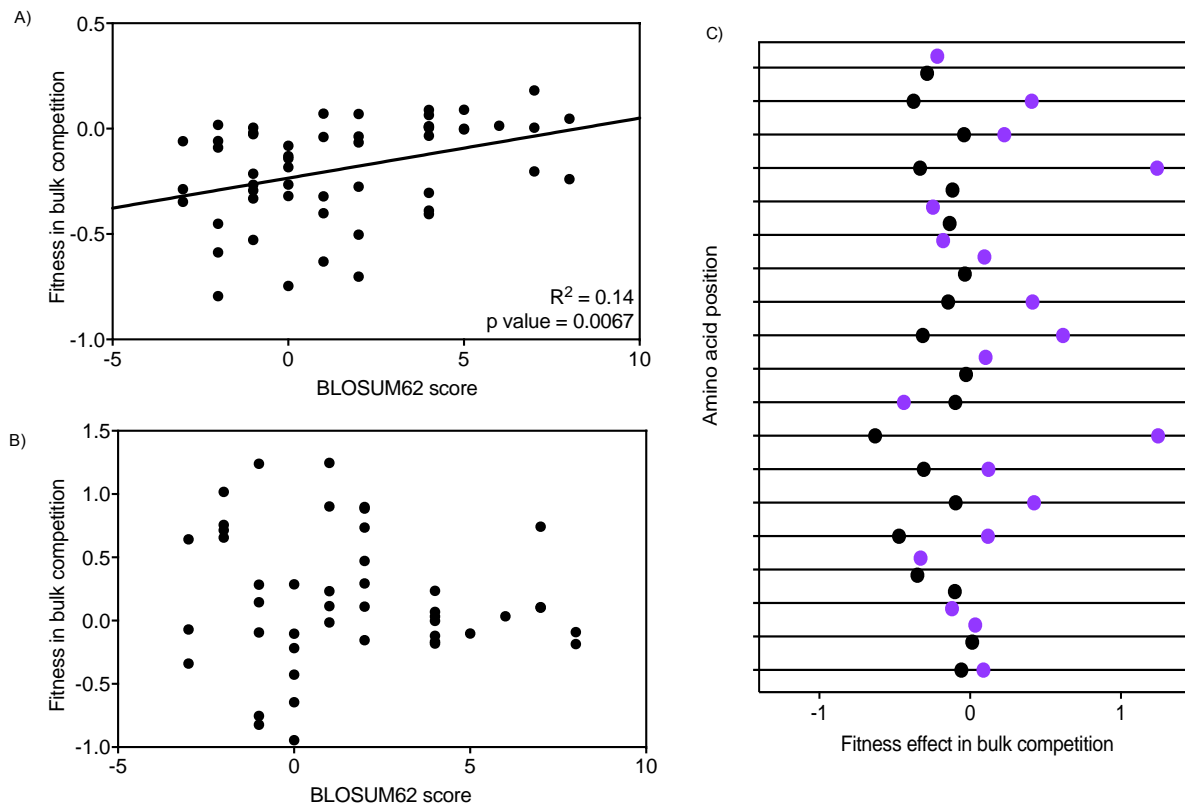
Supplementary Figure 3.2. Analyzing properties of HA mutations. A) Relationship between amino acid similarity and mutant fitness in bulk competition. B) Comparing amino acid similarity and mutant fitness under selection by oseltamivir. C) Analyzing positional sensitivity to mutation in the presence (purple dots) and absence (black dots) of drug pressure (wild type fitness = 0).
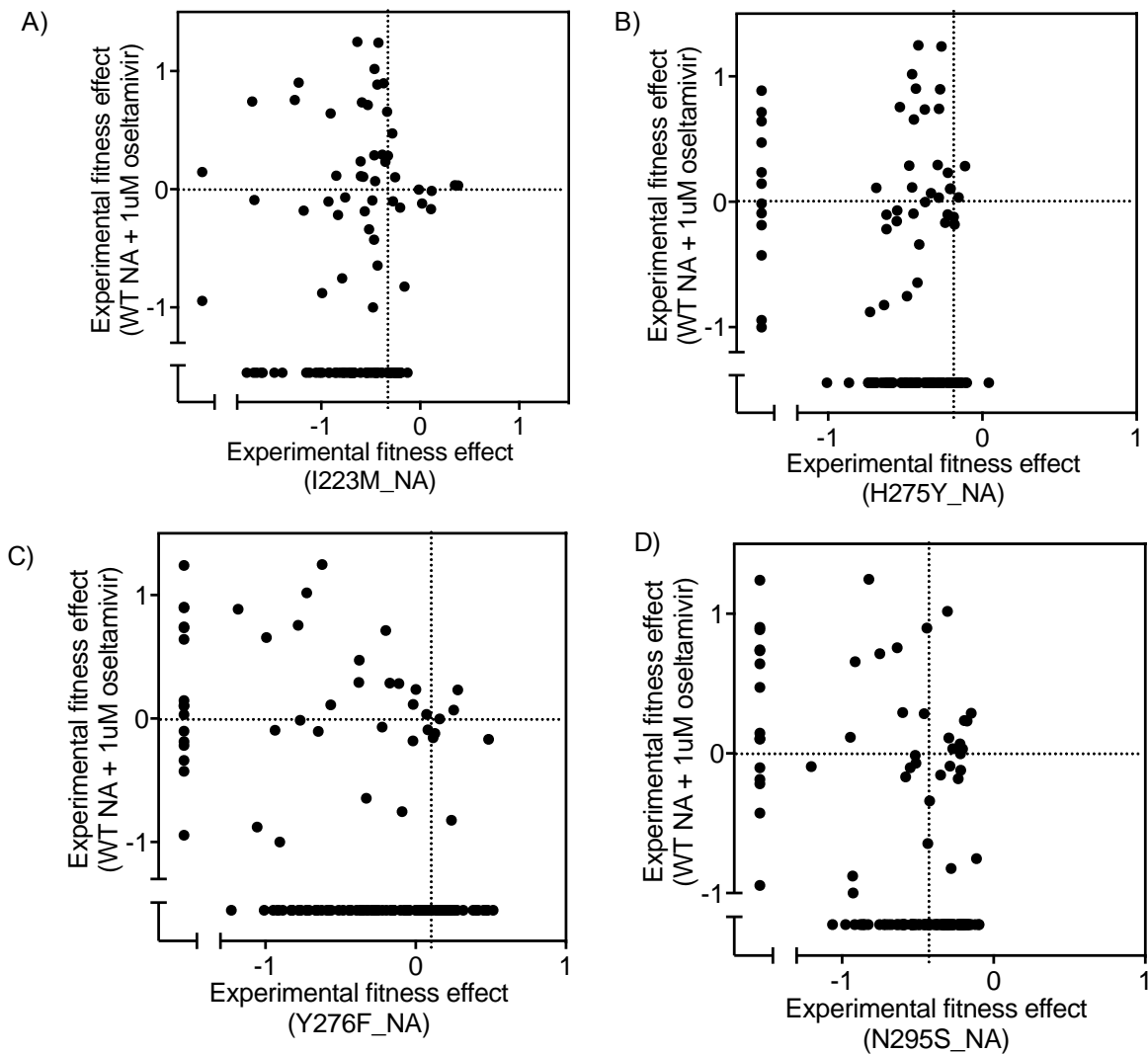
Supplementary Figure 3.3. Comparing the effects of inhibiting NA function by oseltamivir and by mutation A) I223M, B) H275Y, C) Y276F, and D) N295S.
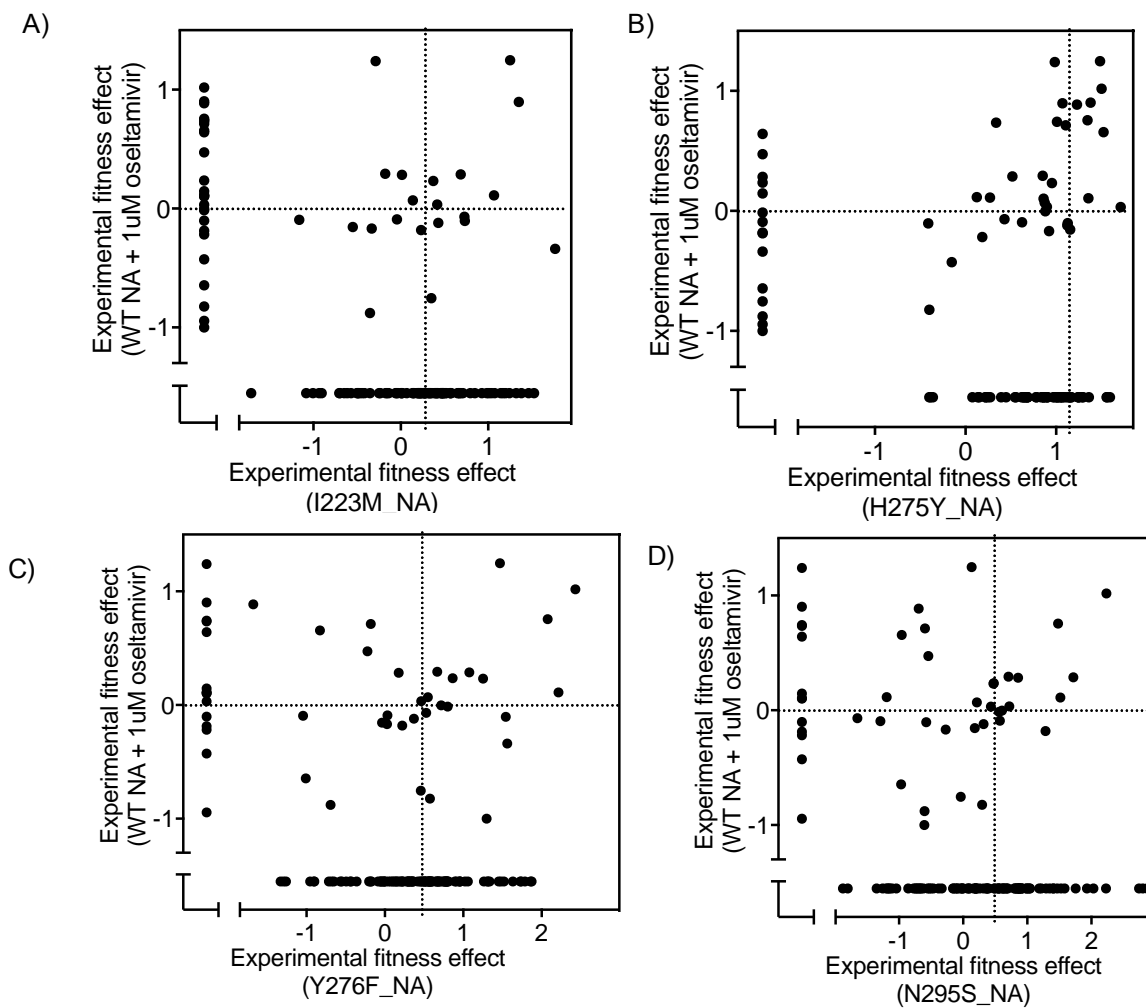
Supplementary Figure 3.4. Comparing effects of inhibiting NA function by oseltamivir and by the combined effect of oseltamivir and mutation A) I223M, B) H275Y, C) Y276F, and D) N295S in NA.
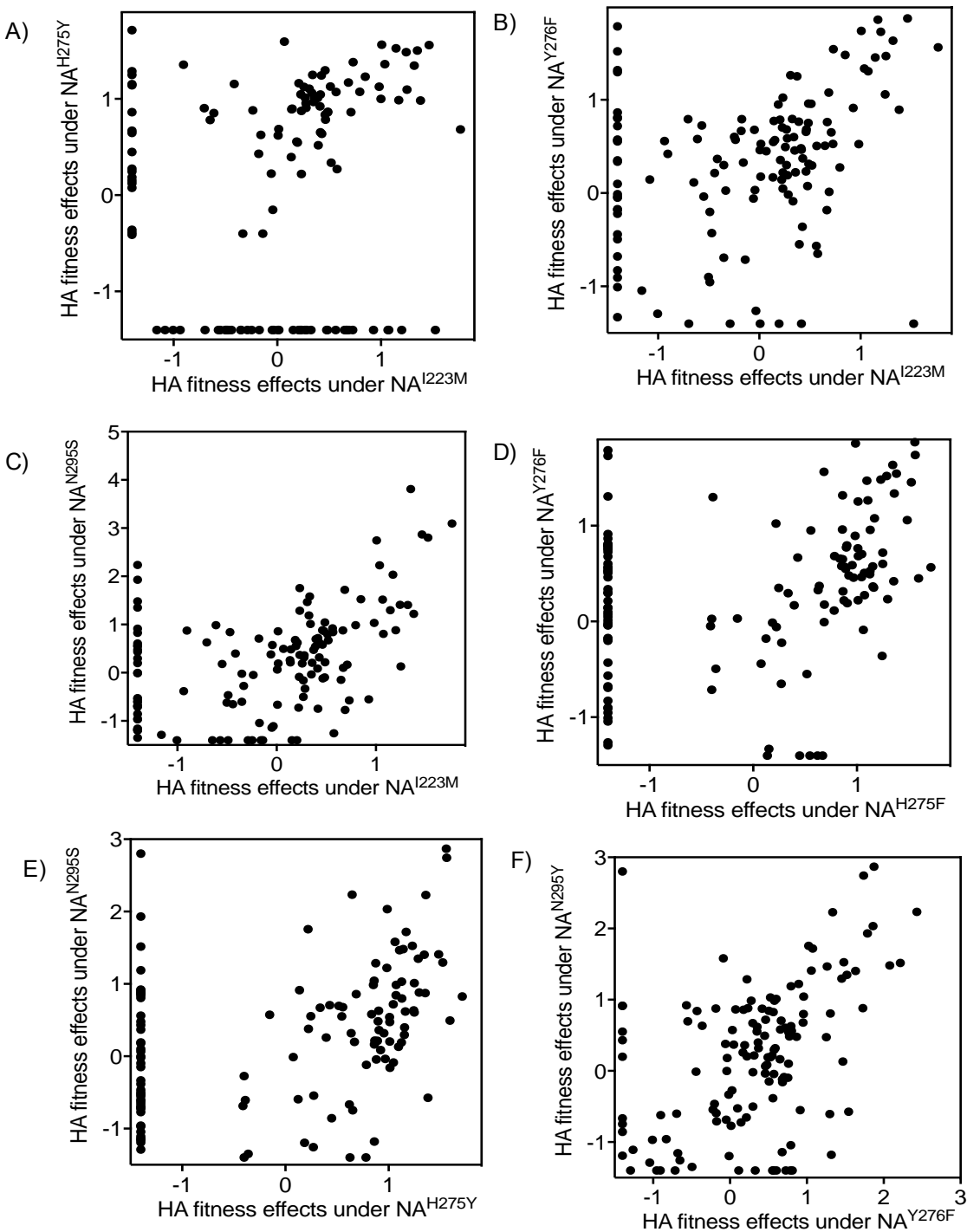
Supplementary Figure 3.5. Comparing effects of different NA mutations on the fitness of HA mutations. A) Comparing the effects of I223M and H275Y. B) Comparing the effects of I223M and Y276F. C) Comparing the effects of I223M and N295S, D) Comparing the effects of H275Y and Y276F. E) Comparing the effects of H275Y and N295S. E) Comparing the effects of N295S and Y276F.
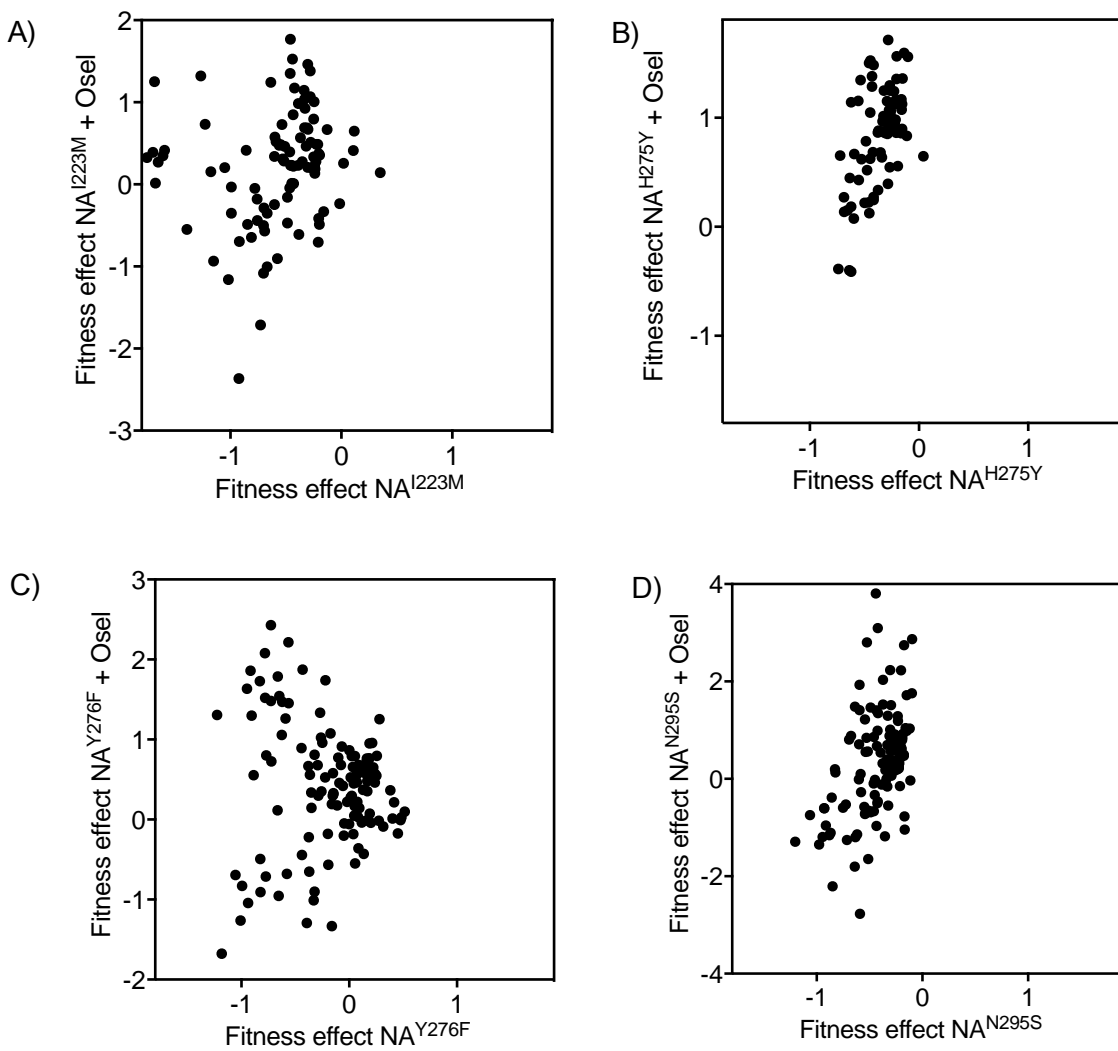
Supplementary Figure 3.6. Comparing the effects of oseltamivir under different NA backgrounds. A) I223M, B) H275Y, C) Y276F, and D) N295S.
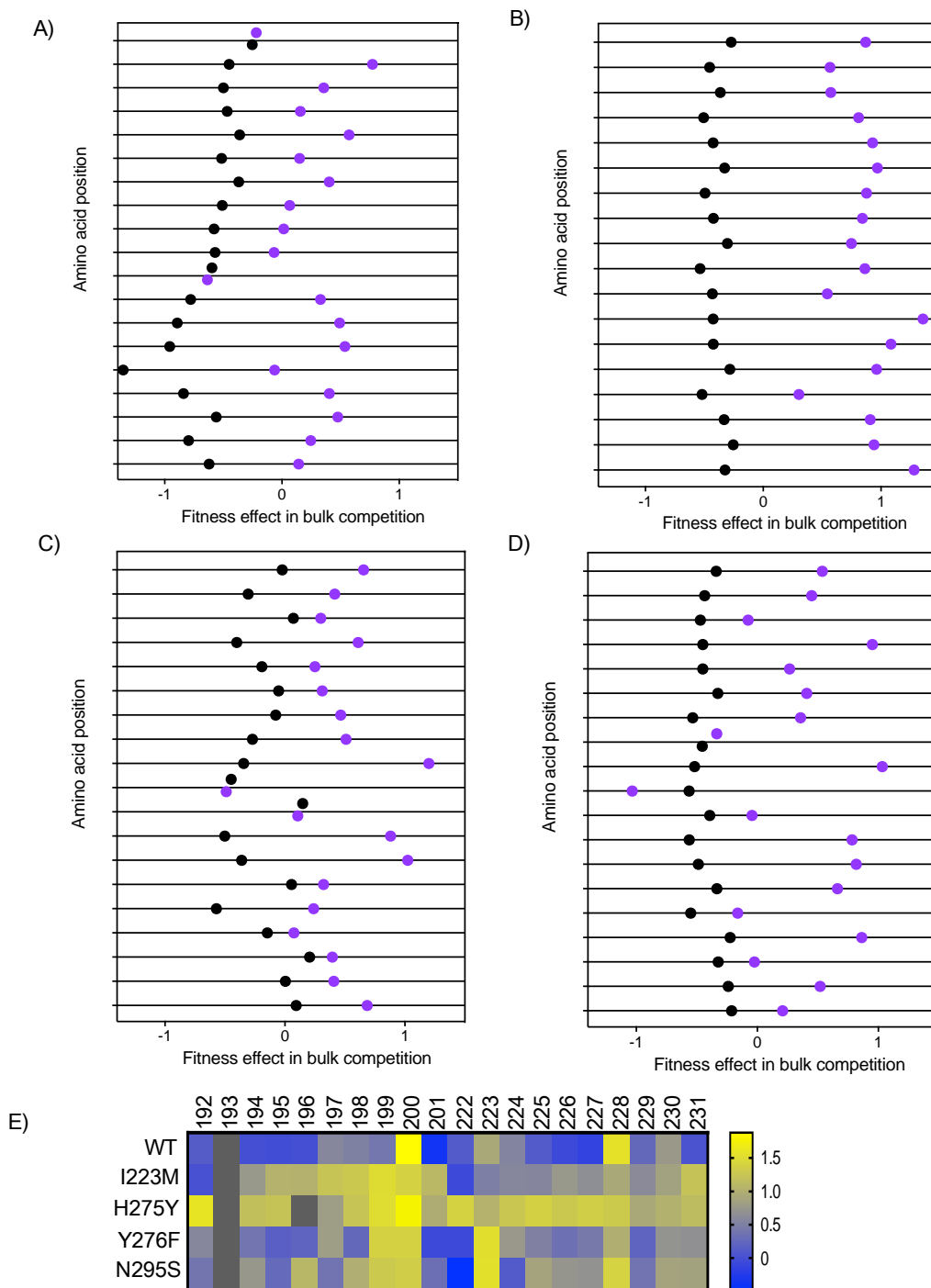
Supplementary Figure 3.7. Analyzing positional sensitivity of HA to mutation under different levels of NA activity mediated by mutation alone or by mutation and inhibition by oseltamivir. A) I223M, B) H275Y, C) Y276F, and D) N295S – black dots show fitness under the respective mutant NA and purple dots show fitness under mutant NA in the presence of oseltamivir (wild type fitness = 0). E) Summarizing the effects of oseltamivir on HA fitness (s_{drug} - $s_{no\ drug}$) under difference NA backgrounds.

Chapter IV: Discussion

Overview

      IAV is a rapidly evolving virus that continues to threaten global public health, yet the

functional constraints shaping the evolutionary trajectory of the virus are not fully understood.

My dissertation research focused on adapting EMPIRIC, an approach that utilizes high

throughput mutagenesis, bulk competition assays, and next-generation deep sequencing, to study

the evolutionary constraints shaping IAV evolution and adaptation under different environmental

conditions. In chapter II, I used EMPIRIC to study the effects of all possible single nucleotide

mutations in the signal sequence of the IAV HA protein. We observed that both synonymous and

non-synonymous substitutions in this region impacted viral fitness. Comparing effects of

mutations in the signal sequence to a region outside the signal sequence showed that

synonymous mutations had larger effects in the signal sequence than in the control region. The

signal sequence also appeared to tolerate non-synonymous substitutions better than the control

region. Analyzing synonymous substitutions revealed that RNA structural elements in this

region, that appear to be species-specific, may be important for virus function. Other factors,

such as protein synthesis efficiency and protein abundance in viral particles, were also correlated

with fitness. We observed evidence of stabilizing selection on HA abundance in virus particles.

Both too much or too little HA appeared to negatively impact virus fitness. In chapter III, I used

the EMPIRIC approach to investigate how changes in NA activity influence the HA-NA

functional balance, and to investigate the effect of these changes on virus fitness. I used two

different methods to modulate NA function – by mutation, and by using the NA inhibitor

oseltamivir. Treatment with oseltamivir yielded much smaller plaques than viruses carrying

mutant NA. This suggests that oseltamivir has larger effects on virus replication than the

analyzed NA mutations. In agreement with this observation, analyzing the effects of reduced NA

activity revealed that inhibiting NA function by mutation had minimal effect on the fitness

landscape of HA. We observed a shift in the distribution of fitness effects of mutations, but

observed that inhibiting NA function by mutation failed to rescue HA mutations. However,

inhibiting NA function either by oseltamivir alone or by mutation combined with oseltamivir led

to the rescue of HA mutations. The combined effects of mutation and drug pressure by

oseltamivir exhibited improved rescue potential. The plaques of the rescued mutations were up to

three times as big as wild type plaques in the presence of oseltamivir. We observed a non-linear

relationship between NA activity and the potential to rescue HA function in virus. These results

suggest that the HA-NA tradeoff is observed when a certain threshold in NA function is crossed.

In summary, we observed that majority mutations introduced in IAV HA conferred a fitness cost

under standard conditions. Analyzing the patterns of mutational sensitivity combined with

biochemical assays on a panel of individually cloned mutations began to explain the molecular

mechanism underlying the observed fitness effects. Under drug pressure by oseltamivir, we

identified HA residues that respond to changes in NA function leading to the rescue of virus

replication.

Synonymous substitutions exhibit greater fitness effects in the signal sequence

Secretory proteins typically have a signal sequence that serves as a molecular address

targeting these proteins to the ER for trafficking and secretion [130, 132, 200]. The first 14 to 16

amino acids of the influenza A HA protein form the signal sequence. Previous studies have

shown that large deletions in this region result in cytoplasmic accumulation of HA, suggesting

the importance of the signal sequence in trafficking and surface expression [34, 140]. Analyzing

the amino acid sequences of the signal compared to the mature HA revealed higher sequence diversity in the signal sequence compared to the rest of HA. However, the evolutionary constraints shaping the evolution of the HA signal sequence remain unclear.

We used the EMPIRIC approach to study effects of mutations in the signal sequence of HA. Focusing on a small region of HA led to improved sensitivity and resolution of the fitness measurements. The improved resolution of this experimental setup allowed the measurement of the effects of both synonymous and non-synonymous substitutions on virus fitness. The observed fitness effects were also validated by studying individual mutations in isolation. When analyzed in isolation, synonymous mutations that exhibited deleterious effects in bulk competition also had smaller plaques compared to wild type. Effects of synonymous substitutions were further validated by studying their effects on the growth kinetics of viruses. Deleterious synonymous substitutions analyzed in isolation either exhibited a longer eclipse duration or reduced replication rate, leading to slower infection kinetics compared to wild type virus.

In order to quantify selection acting on the HA signal sequence relative to a neighboring region of HA, we compared the distributions of fitness effects of synonymous and non-synonymous mutations. Synonymous substitutions had larger effects in the signal sequence than in the neighboring region used for comparison. Non-synonymous mutations were more tolerated in the signal than the neighboring region. Calculating the apparent ratio of selection between the two regions indicated that amino acid sequence of the signal may be evolving faster than that of the control region. This observation is consistent with the high sequence diversity observed in the signal peptides of sequenced circulating IAV variants.

Studying the molecular basis of the observed fitness effects revealed that majority of the synonymous mutations analyzed impacted protein synthesis efficiency in cell culture. This

observation was consistent with the previous observation that rare codons at the N-termini of proteins regulate translation [107, 124, 153]. However, analyzing protein abundance in released viruses showed apparent stabilizing selection on HA abundance per plaque forming unit. Further studies will be necessary to elucidate how changes in HA synthesis affect the ratio of infectious to non-infectious viruses per plaque forming unit, and to explain how this change shaped IAV evolution.

Results of the fitness effects of mutations in the HA signal sequence demonstrated that RNA structure may also be shaping the evolution of this region. Moss and colleagues studied the structural components of the IAV genome by performing RNA structure predictions based on suppression of synonymous codon usage and unusual thermodynamic stabilities [201]. The HA segment appeared to be less structured than the rest of the genome. However, this study combined IAV genome sequences from human, avian and swine hosts for their analyses, and may have missed species-specific structural components of the IAV genome. Our fitness scan using the EMPIRIC approach combined with analysis of sequenced circulating isolates from different hosts revealed species-specific differential preferences for a structured or unstructured loop at nucleotide positions 7 – 23. This region is part of a larger region that has been implicated in virus packaging [158, 161]. However, species-specific differences in IAV packaging are not fully understood. Studying the role of host factors in constraining this region will help our understanding of influenza function and cross-species transmission.

Overall, this study highlights that synonymous mutations are not evolutionarily neutral as proposed by previous theories. Synonymous mutations often impact fitness, thus necessitating the need to re-valuate the expression dN/dS that is widely used to measure the strength of natural

selection acting on protein-coding genes. As such, it is important to consider effects of synonymous mutations when calculating and predicting the evolution rate of different proteins.

Adapting the EMPIRIC approach to study HA-NA functional balance

Many studies have shown the importance of the functional balance between receptor binding by HA and receptor cleavage by NA. Altogether, these studies show that productive viral infection is supported by matching levels of HA and NA activity. Viral evolution studies in the laboratory, as well sequence analysis of circulating viruses have shown that when this balance is perturbed, mutations in the virus that re-establish this balance will fix in the virus population [66, 195]. However, the molecular mechanism, and the role of this balance in shaping IAV evolution remains unclear.

We adapted the EMPIRIC approach to investigate how changes in the level of receptor cleavage activity of NA influences the fitness landscape of HA mutations near the receptor binding site. We also sought to study whether the means of modulating NA function – by mutation or using the NA inhibitor oseltamivir - impacted the resulting fitness landscape. We observed that inhibiting NA function using oseltamivir had a larger effect than inhibition by mutation. Oseltamivir allowed the identification of HA mutations with adaptive potential in the presence of drug pressure. The fitness trends seen in our bulk competition results in the presence of oseltamivir were consistent with mutant frequencies observed in circulating viruses following the introduction of oseltamivir in the clinic.

Our results show that there was a non-linear relationship between NA fitness effect, as measured by plaque size, and the potential to rescue HA mutations. Inhibiting NA function by mutation alone was not sufficient to rescue HA function, however it appears that oseltamivir

lowered NA function enough to establish a functional match with HA mutations. Previous studies have shown that IAV infection is initiated by many weak interactions between HA and sialic acid at the host cell receptor [202, 203]. However, the limitations in the range of HA binding strength to ensure successful virus binding and internalization remain unknown. Investigating the effect of HA and NA mutations on the HA-NA functional balance combined with phylogenetic analyses can elucidate the role of this balance on IAV evolutionary trajectories.

The HA-NA functional balance has been implicated in the development of resistance against oseltamivir inhibitors [66]. In the current study, we use a systematic approach to identify HA mutations that respond to oseltamivir. Expanding this mutational scan to study the whole HA protein, and to study the effects of different drug pressures, such as zanamivir, will further explain the role of the HA-NA functional balance on the development of resistance against oseltamivir inhibitors. Understanding the molecular underpinnings of the HA-NA functional balance in IAV may help explain the functional balance between virus entry and exit in other viruses in the paramyxoviridae family, such as measles, sendai, and newcastle viruses.

Using EMPIRIC to study protein function and evolution

EMPIRIC is a high throughput mutational scanning approach that has been used widely to study sequence-structure-function relationships of proteins in the yeast system [90, 117-119], to study disease-causing genes in cell culture [204, 205], and to study viral proteins [100, 206]. This approach uncovers positional sensitivity of a protein to mutation, both under normal conditions and under different selective pressures. Combining results from EMPIRIC fitness

scans and structural and biochemical analyses began to explain the molecular mechanisms underlying protein function [119] and the development of drug resistance [170].

Compared to other mutational scanning approaches, EMPIRIC provides a high throughput approach that is also systematic and comprehensive. Cassette-based mutagenesis provides more control on the frequency and coverage of mutations in mutant plasmid libraries compared to error-prone PCR. Combining this method of library preparation and straightforward functional assays, EMPIRIC results in highly reproducible fitness measurements. However, due to current technological advances and reduced cost, plasmid library preparation in mutational scanning studies is shifting towards the use of parallel oligonucleotide synthesis [207]. This, and other similar approaches, will enable the production of high diversity barcoded libraries that will in turn increase throughput.

Like all mutational scanning approaches, the EMPIRIC approach has shortcomings. The translation of EMPIRIC results to fitness effects of mutations in nature is dependent on the functional assay used. If the experimental conditions deviate from those of the natural function of the protein, then the fitness results obtained may be misleading. Additionally, the nature and strength of selective pressures that form part of the environmental condition under study must be in line with the pressures that will shape the evolution of a protein in nature. The EMPIRIC approach has additional limitations when used to study virus evolution. Bottlenecks such as during virus recovery from plasmid and during bulk competition assays can lead to inaccurate results. The presence of these bottlenecks often leads to limited throughput. Despite these drawbacks, the EMPIRIC approach provides a useful tool to study protein function and evolution under different environmental conditions.

On extrapolating EMPIRIC findings to circulating IAV

Results from EMPIRIC studies on HA generally agree with patterns seen in sequences

circulating viruses ($P < 0.05$). In these comparisons mutations with large fitness defects were

present at a low frequency or were absent in sequences of circulating viruses. On the other hand,

mutations that show minimal effects of virus fitness were abundant in circulating viruses. We

acknowledge that the EMPIRIC approach provides information on the general trends of the

fitness effects of mutations, and not the absolute fitness of each mutation. Studying effects of

mutations in individual virus clones in isolation helps to validate fitness measurements from bulk

competition experiments, indicating that this approach generally captures trends of virus

evolution present in nature.

IAV is a rapidly evolving virus. As such, sequence differences between the laboratory

strain A/WSN/33, that is commonly used for experimental evolution studies, and circulating

strains may limit the dependability of results from mutational scanning studies. Ideally,

performing IAV evolution experiments using viruses derived from the currently circulating

viruses will provide the most accurate information on the evolutionary trajectory and potential

for drug resistance in IAV. However, with the high rate of influenza virus evolution and failure

to accurately predict predominant circulating variants has deemed this approach unfeasible.

Nevertheless, information garnered from the current experimental setup provides insight into

virus function and evolution.

Future directions

In this dissertation project, I used the EMPIRIC approach to investigate the evolutionary

constraints shaping IAV evolution. Specifically, I studied how mutations in the signal sequence

of IAV HA affects virus replication, and examined how the HA-NA functional balance influences virus function and evolution.

Studying effects of signal sequence mutations on IAV replication

Studying effects of mutations in the signal sequence of HA showed that there is a complex relationship between viral factors, at the RNA and protein levels, and host factors that influence virus fitness. Future studies investigating host factors such as receptor specificity, tissue specificity, and antiviral responses may further explain the observed host-specific differences. Expanding these studies to the whole HA protein may reveal other host-specific structural constraints that as of yet remain unknown.

These studies also show that IAV may be under stabilizing selection for HA abundance in infectious particles, however, effects of changes in HA expression have not been explicitly investigated. Future studies analyzing the effects of changes in HA expression, by using different expression promoters for instance, in different IAV strains may explain virus function. Alternatively, similar studies investigating the effects of changes in NA abundance on virus function will be of interest. Since HA and NA are the main surface proteins of IAV, these studies may also elucidate virus antigenicity and host responses to IAV infection.

Investigating HA-NA functional balance

The balance in HA and NA function is critical for virus replication, yet remains underexplored. Our results show that modulating NA function by oseltamivir enables the rescue of HA mutations, and that there is a non-linear relationship between NA activity and rescue potential. However, the molecular mechanism underlying this balance remains unknown. Future studies using modern strains such as A/Brisbane/59/2007 will show whether the observed rescue

pattern is a general phenomenon or whether the pattern is specific for the A/WSN/33 strain used for these studies. Additionally, using viruses with a different receptor specificity – a2,3-linked sialic acid instead of a2,6-linked sialic acid – may explain the role of this functional balance across different IAV host species.

Conclusion

IAV remains a public health threat largely because we do not fully understand its evolution. Learning the evolutionary constraints governing the function and evolution of IAV proteins is important for the efficient control of influenza infections. To date, the most effective way to control influenza infection is by vaccination. However, mismatch between IAV variants contained in the vaccine and circulating viruses continually undermines the effectiveness of the vaccine. Antiviral agents that inhibit NA function are the common place therapies used to treat influenza infections. Resistance to oseltamivir inhibitors that are widely used to treat influenza infections makes the control of IAV a challenge. In this dissertation, I used the EMPIRIC approach to study the evolutionary constraints shaping the evolution of the HA signal sequence. Results indicate that there is a complex interaction between viral, environmental, and host factors that facilitate virus function. I also used the EMPIRIC approach to study the role of the HA-NA functional balance in virus function, and in adaptation to drug pressure in the environment. This study highlighted the need to focus on both NA and HA in investigating resistance to oseltamivir inhibitors. Understanding the role of HA in drug resistance can inform the design of antiviral agents with reduced chance for resistance, and help identify new targets for IAV therapies.

Appendix - Mutations proximal to the signal peptidase cleavage site of influenza A virus hemagglutinin exhibit dominant negative properties

This work was performed as a preliminary analysis to investigate the potential dominant negative properties of a list of HA mutations identified in a previous study. Follow-up experiments are required to elucidate the observed effects of HA mutations on the function of wild type protein.

Abstract

Dominant negative mutations are mutations that disrupt protein function when co-expressed with wild type protein. Cold-adapted influenza A virus that exhibited dominant negative properties were identified in the early 1990s. Exploiting these dominant negative properties provided a new strategy for producing an attenuated live-virus vaccine for use in immunocompromised individuals, and provided an alternative approach for the design of antiviral agents. This approach showed promising results in ferrets that were co-infected with a virulent strain of influenza A virus. The development of these constructs for use in the clinic has not been a success. In this study, we identified mutations in the signal sequence of IAV HA that exhibited dominant negative properties. These mutations caused up to 300-fold reduction in virus titer in mixed transfections with wild type. Studying their effects on wild type HA expression revealed up to 50% reduced HA expression in mammalian cells. Dominant negative mutations in HA, the most antigenic influenza protein, can help inform vaccine development.

Introduction

The N-terminus of the hemagglutinin protein of influenza A virus consists of a signal sequence that targets nascent HA to the endoplasmic reticulum (ER) for trafficking [135, 140]. Like many signal sequences, the HA signal sequence is composed of a positive charged amino terminal followed by a hydrophobic region and a polar region [130, 134]. Targeting of nascent HA to the ER is initiated by the signal recognition particle (SRP) binding to the hydrophobic region of the signal sequence which anchors at the SRP receptor on the ER membrane allowing insertion of the emerging protein into the translocon [132, 208]. A trypsin-like signal peptidase cleaves the signal peptide before HA exits the ER [134, 209]. HA trafficking mediated by the signal sequence has been shown to be critical for proper virus function as deletion mutations in the signal sequence show impaired function [140]. The amino acid sequence of the HA signal peptide is variable between different HA subtypes, yet the structural and biochemical properties are highly conserved. However, the role of the signal sequence in virus function and evolution remains understudied.

Mutations with dominant negative properties have the ability to disrupt wild type protein function when overexpressed in a polypeptide [210]. Since dominant negative mutants result in attenuated wild type phenotype, they provide a useful tool to study protein function [211, 212] and have exhibited therapeutic potential [77, 213]. For instance, influenza A viruses carrying mutations in the matrix gene were shown to harbor dominant negative properties and present a potential novel class of antivirals against influenza [214]. Exploring the molecular properties of dominant negative IAV mutations for therapeutic purposes is of high importance especially with the increasing prevalence of resistance against the currently used drug oseltamivir. However,

identifying and characterizing IAV mutations with dominant negative properties for candidates as therapies, and how host factors respond to these properties remains underexplored.

In the present study, we aim to identify signal sequence mutations that may have dominant negative effects on HA trafficking to the ER for surface expression. We used the dataset generated previously using the EMPIRIC (Exceedingly Meticulous and Parallel Investigation of Randomized Codons) approach to study effects of mutations in the signal sequence of IAV HA. We focused on mutations that were present in the plasmid library but were strongly depleted in the recovered virus population. In order to distinguish lethal from dominant negative mutations, we hypothesized that when studying these mutations in isolation only dominant negative mutations will disrupt wild type function. Furthermore, we speculated that wild type trafficking and surface expression will be greatly impacted by dominant negative mutations that inhibit signal peptide cleavage. We identified signal sequence mutations that caused up to 300-fold titer reduction in mixed transfections with wild type HA. These mutations also disrupted the surface expression and cellular localization of wild type HA.

Results and discussion

Mutations in the HA signal sequence disrupt wild type function

The signal sequence is generally hydrophobic and has low molecular weight residues near the peptidase cut site [215] (Fig. 5.1a). While studying effects of mutations in the HA signal sequence, we observed that some mutations were abundant in the plasmid library but were highly depleted in recovered viruses (Fig. 5.1b). In order to study the dominant negative potential of these mutants, we performed mixed transfection assays with wild type plasmid and compared virus titer to true wild type. Mutations L6R, V7D, V7G and L8R had no impact on wild type titer
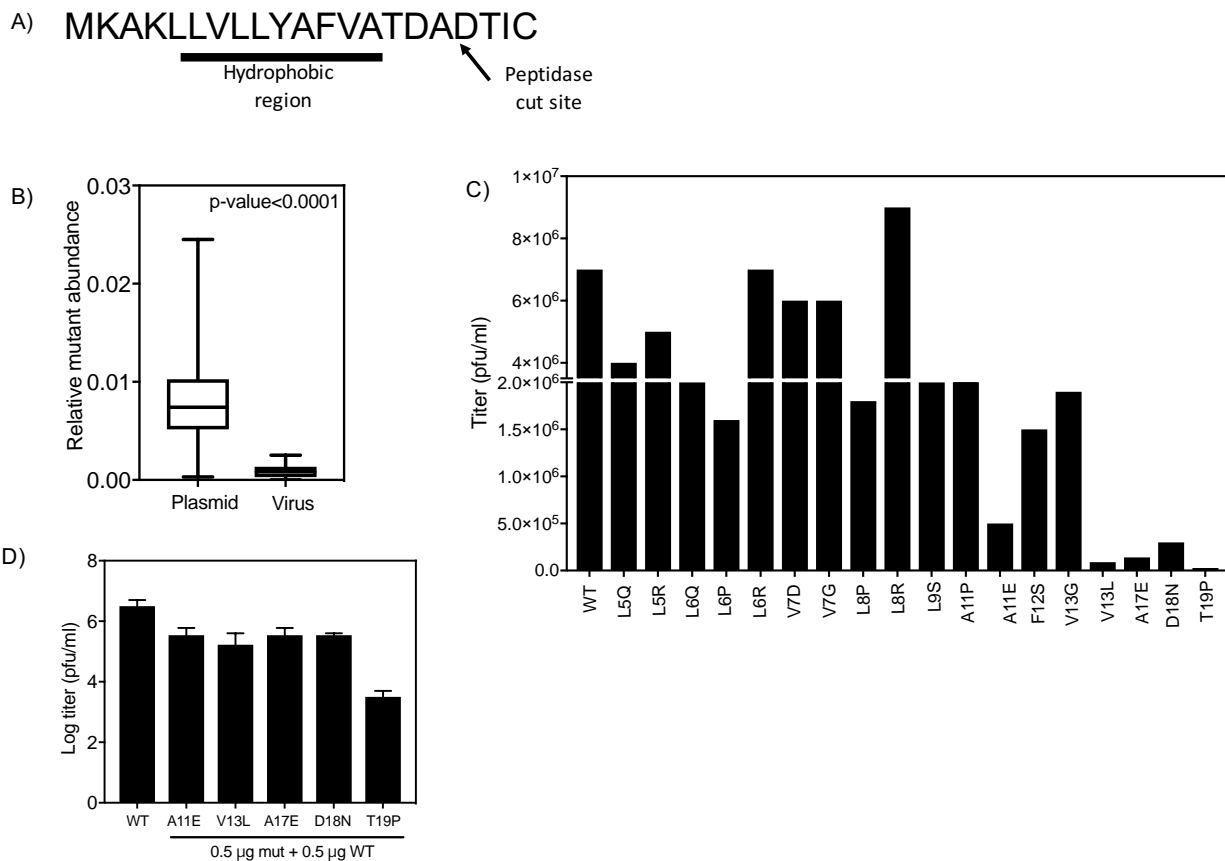
Figure 5.1. Mutations in the HA signal sequence can have dominant negative effects. A) Amino acid sequence of the A/WSN/33 HA signal sequence. The SRP binds at the hydrophobic region. The signal sequence is cleaved by host proteases and does not form part of the mature protein. B) Relative mutant abundance of highly depleted mutants in plasmid library and in recovered virus. C) Effects of a panel of individually cloned variants on wild type titer. D) Of the mutants in panel C, 5 variants were chosen for further analysis.

suggesting that these mutations may be lethal and not dominant negative. Mutations with the largest effect in mixed transfections were A11E, V13L, A17E, D18N and T19P (Fig. 5.1c). Our results show that T19P reduced wild type titer by about 300-fold. From this panel, we chose these 5 mutants for further analysis (Fig. 5.1d).

Candidate mutations interfere with wild type HA expression in mammalian cells

In order to study the effects of this panel of mutations on HA expression, we used a surface-display plasmid and co-expressed wild type and mutant plasmids in 293T cells (Fig. 5.2a). When expressed alone, each of these mutations exhibited reduced surface expression, with the A11E mutant expressing about 25% of wild type HA levels. When co-expressed with wild type, all mutations appeared to disrupt wild type expression. V13L reduced wild type expression to about 50% (Fig. 5.2b). A17E and T19P mutations had the least effect on surface expression (Fig. 5.2b). Since these mutations had the largest effects on wild type titer but exhibited minimal effect on HA expression, we speculate that A17E and T19P mutations may impact the function of surface expressed protein. Further studies will help elucidate the molecular mechanism of the observed effects. However,

We performed western blots to further investigate effects of mutations on HA expression. Consistent with our FACS results, all the mutants analyzed showed reduced HA expression (Fig. 5.2c). A11E had the lowest levels of surface expression and appeared to affect the oligomerization of wild type HA.

Previous studies have shown that the evolutionary trajectories of the IAV genome are significantly impacted by host proteostasis [216]. Although the molecular mechanism of the observed effects remains unknown, we speculated that abnormal signal sequence cleavage of
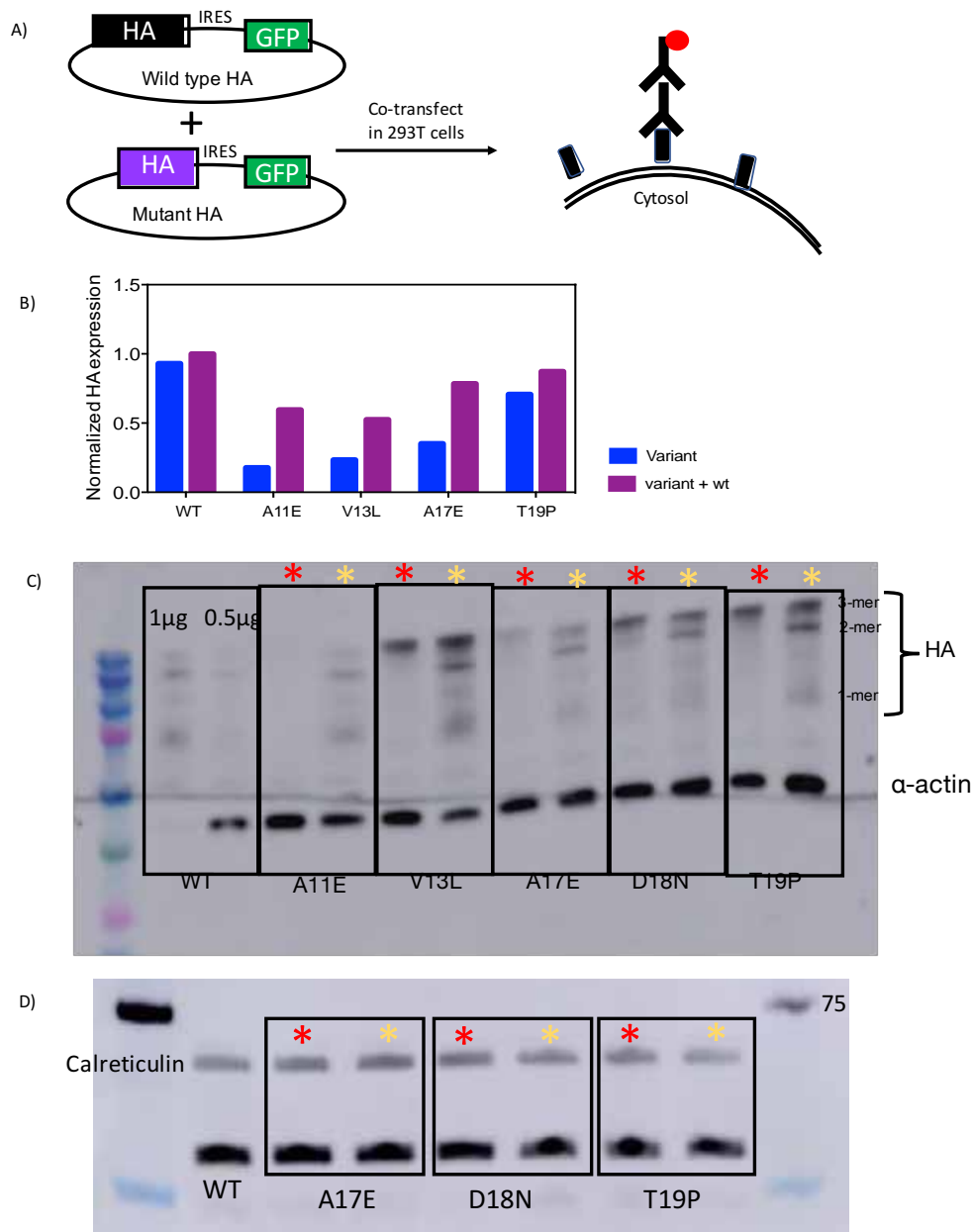
Figure 5.2. Candidate mutations disrupt HA expression in mammalian cells. A) Experimental approach to study effects of mutations on HA expression. B & C) Effects of HA mutants on wild type HA expression measured by flow cytometry (B) and by western blot (C). In (B) Blue bars show single-variant expression and magenta bars show co-expression of wild type and mutant variants. D) Analyzing effects of HA mutations on calreticulin expression. In C & D, a red star denotes mutant HA alone and a yellow star denotes co-transfected wild type and mutant HA.

these variants may trigger the protein folding stress response in mixed transfections. There were no changes in calreticulin expression in single and mixed transfections when compared to wild type (Fig. 5.2d). Studying other chaperones in the ER, specifically calnexin and BiP, may confirm whether or not these signal sequence mutations trigger the protein misfolding response.

Using immunostaining to study cellular localization of mutant HA

We used immunostaining and fluorescence microscopy to analyze effects of candidate mutations on expression levels and cellular localization of wild type HA. Wild type HA was highly expressed, and was localized along the cell membrane (Fig. 5.3a). Mixed transfections of wild type and mutant HA showed reduced HA expression on the cell surface with A11E showing the least expression (Fig. 5.3b – f). The mutation D18N appeared to change cellular localization to a punctate distribution. Further studies using quantitative confocal microscopy will provide further information on the changes in expression and cellular localization observed.
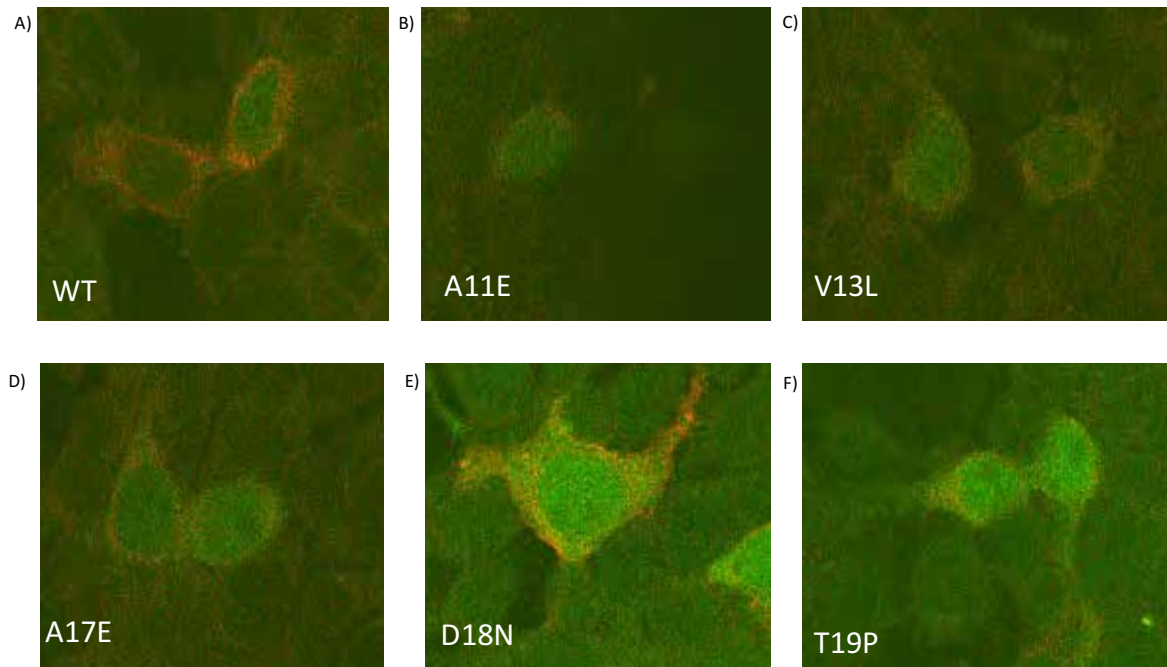
Figure 5.3. Comparing cellular localization and expression levels of wild type HA and wild type co-transfected with mutant HA. A) Wild type HA appears to localize along the plasma membrane. B – F) Analyzing expression of wild type HA co-expressed with mutant HA, A11E – B, V13L – C, A17E – D, D18N – E and T19P – F.

Materials and methods

Cell culture

MDCK cells were maintained in cell culture media consisting of Eagle's Minimal Essential Media supplemented with 10% fetal bovine serum, 100U/mL penicillin and 100µg/mL streptomycin grown at 37°C and 5% carbon dioxide. 293T cells were maintained in Opti-MEM I media supplemented with 5% fetal bovine serum, 100U/mL penicillin and 100µg/mL streptomycin grown at 37°C and 5% carbon dioxide.

Virus recovery

Mutations of interest were identified from a previous high-throughput mutational scan (in publication) that used next-gen deep sequencing to measures effects of all possible single nucleotide mutations in the HA signal sequence. A panel of 19 variants was generated using the pHW2000-based 8-plasmid system developed by Webster and colleagues [15]. Briefly, 1 µg of wild type HA plasmid library and 1 µg of each plasmid encoding the rest of the A/WSN/33 genome, were transfected in a co-culture of MDCK and 293T cells using TransIT LT1 reagent. For the mutants of interest 0.5 µg of wild type HA was mixed with 0.5 µg of mutant HA and transfected with plasmids encoding the rest of the genome. Media was changed to serum-free media 6 h post-transfection. At 30 h post-transfection, TPCK trypsin was added to a final concentration of 0.5 µg/mL. Virus-containing supernatant was harvested at 72 h post-transfection, centrifuged at 300g for 15 minutes and stored at -80°C.

Titer reduction assay

Viral titer of recovered viruses was quantified by plaque assay as previously described [100]. Briefly, $5x10^5$ MDCK cells per well were seeded 24 h before infection. Harvested virus was diluted in a 10-fold serial dilution in wash buffer and added to cells. Cells were incubated with virus for 1 h. Virus was removed from cells and replaced with agar overlay. Infected cells were incubated for 48 h at 37°C and 10% CO_2. Immunostaining of plaques was performed with anti-NP antibody, and plaques were visually counted and plaque size measured using a Nikon SMZ1500 microscope.

HA expression in mammalian cells

HA variants were cloned into the JB991 plasmid, kindly provided by J. Bloom (University of Washington, Seattle), that has a CMV promoter, a GFP reporter, and an ampicillin selection marker. HA clones were transiently expressed in 293T. Cells were harvested 24 h post-transfection, washed and resuspended in isotonic buffer (15 mM MOPS, 145 mM NaCl, 2.7 mM KCl, 4 mM $CaCl_2$ pH 7.4. 2% heat-inactivated FBS). Cells were stained with rabbit HA antibody for 1 h at room temperature followed by Alexa Fluor 647 goat anti-rabbit IgG secondary antibody. Stained cells were washed, resuspended in isotonic buffer and analyzed on a BD FACSCalibur. Data analysis was conducted using the FlowJo data analysis software package.

Studying cellular localization using immunofluorescence microscopy

HA variants were cloned into the JB991 plasmid and transiently expressed in 293T. At 24 h post-transfection, cells were fixed using 4% paraformaldehyde in PBS and stained with influenza A virus H1N1 HA primary antibody, followed by Alexa Fluor 647 secondary antibody, anti-

calreticulin antibody and Alexa Fluor 549 secondary antibody. Images were acquired using a Nikon Eclipse Ni-U microscope at 40x magnification using NIS-Elements imaging software (version 4.13).

Western blot analysis

HA variants were transiently expressed in 293T cells. Cells were washed and lysis buffer supplemented with PMSF was added to cells and incubated for 15 mins at 25°C. Cells were then scraped from wells and centrifuged at 1200 rpm for 2 mins. Cell lysates were collected and stored at -80°C. 20μg of total protein was loaded per well and blotted using HA antibody, calreticulin antibody and alpha-actin antibody.

Bibliography

[1] CDC. How Flu Spreads. Atlanta, Georgia: Center for Disease Control and Prevention; 2017.

[2] Kash JC, Tumpey TM, Proll SC, Carter V, Perwitasari O, Thomas MJ, et al. Genomic analysis of increased host immune and cell death responses induced by 1918 influenza virus. Nature. 2006;443:578-81.

[3] Kobasa D, Takada A, Shinya K, Hatta M, Halfmann P, Theriault S, et al. Enhanced virulence of influenza A viruses with the haemagglutinin of the 1918 pandemic virus. Nature. 2004;431:703-7.

[4] Molinari NA, Ortega-Sanchez IR, Messonnier ML, Thompson WW, Wortley PM, Weintraub E, et al. The annual impact of seasonal influenza in the US: measuring disease burden and costs. Vaccine. 2007;25:5086-96.

[5] WHO. Influenza (Seasonal). Geneva, Swtizerland2016.

[6] Taubenberger JK, Morens DM. 1918 Influenza: the mother of all pandemics. Emerg Infect Dis. 2006;12:15-22.

[7] Kilbourne ED. Influenza pandemics of the 20th century. Emerg Infect Dis. 2006;12:9-14.

[8] Taubenberger JK, Morens DM. The pathology of influenza virus infections. Annu Rev Pathol. 2008;3:499-522.

[9] Fields BN, Knipe DM, Howley PM. Fields virology. 3rd ed. Philadelphia: Lippincott-Raven Publishers; 1996.

[10] Shope RE. Swine Influenza : Iii. Filtration Experiments and Etiology. J Exp Med. 1931;54:373-85.

[11] Geraci JR, St Aubin DJ, Barker IK, Webster RG, Hinshaw VS, Bean WJ, et al. Mass mortality of harbor seals: pneumonia associated with influenza A virus. Science. 1982;215:1129-31.

[12] Sovinova O, Tumova B, Pouska F, Nemec J. Isolation of a virus causing respiratory disease in horses. Acta Virol. 1958;2:52-61.

[13] Tong S, Zhu X, Li Y, Shi M, Zhang J, Bourgeois M, et al. New world bats harbor diverse influenza A viruses. PLoS Pathog. 2013;9:e1003657.

[14] CDC. Epidemiology and Prevention of Vaccine-Preventable Diseases. Atlanta, Georgi: CDC; 2015.

[15] Hoffmann E, Neumann G, Kawaoka Y, Hobom G, Webster RG. A DNA transfection system for generation of influenza A virus from eight plasmids. Proc Natl Acad Sci U S A. 2000;97:6108-13.

[16] Arranz R, Coloma R, Chichon FJ, Conesa JJ, Carrascosa JL, Valpuesta JM, et al. The structure of native influenza virion ribonucleoproteins. Science. 2012;338:1634-7.

[17] Moeller A, Kirchdoerfer RN, Potter CS, Carragher B, Wilson IA. Organization of the influenza virus replication machinery. Science. 2012;338:1631-4.

[18] Wilson IA, Skehel JJ, Wiley DC. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution. Nature. 1981;289:366-73.

[19] Varghese JN, McKimm-Breschkin JL, Caldwell JB, Kortt AA, Colman PM. The structure of the complex between influenza virus neuraminidase and sialic acid, the viral receptor. Proteins. 1992;14:327-32.

[20] Lamb RA, Choppin PW. Identification of a second protein (M2) encoded by RNA segment 7 of influenza virus. Virology. 1981;112:729-37.

[21] Safo MK, Musayev FN, Mosier PD, Zhou Q, Xie H, Desai UR. Crystal structures of influenza A virus matrix protein M1: variations on a theme. PLoS One. 2014;9:e109510.

[22] McCown MF, Pekosz A. The influenza A virus M2 cytoplasmic tail is required for infectious virus production and efficient genome packaging. J Virol. 2005;79:3595-605.

[23] Lamb RA, Choppin PW. Segment 8 of the influenza virus genome is unique in coding for two polypeptides. Proc Natl Acad Sci U S A. 1979;76:4908-12.

[24] Hale BG, Randall RE, Ortin J, Jackson D. The multifunctional NS1 protein of influenza A viruses. J Gen Virol. 2008;89:2359-76.

[25] O'Neill RE, Talon J, Palese P. The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins. EMBO J. 1998;17:288-96.

[26] CDC. Influenza Type A Viruses. Atlanta, Georgia: CDC; 2017.

[27] Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of influenza A viruses. Microbiol Rev. 1992;56:152-79.

[28] Ferguson NM, Galvani AP, Bush RM. Ecological and immunological determinants of influenza evolution. Nature. 2003;422:428-33.

[29] Taubenberger JK, Kash JC. Influenza virus evolution, host adaptation, and pandemic formation. Cell Host Microbe. 2010;7:440-51.

[30] Kawaoka Y, Krauss S, Webster RG. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. J Virol. 1989;63:4603-8.

[31] Scholtissek C, Koennecke I, Rott R. Host range recombinants of fowl plague (influenza A) virus. Virology. 1978;91:79-85.

[32] Yuen KY, Chan PK, Peiris M, Tsang DN, Que TL, Shortridge KF, et al. Clinical features and rapid viral diagnosis of human disease associated with avian influenza A H5N1 virus. Lancet. 1998;351:467-71.

[33] Russell RJ, Gamblin SJ, Haire LF, Stevens DJ, Xiao B, Ha Y, et al. H1 and H7 influenza haemagglutinin structures extend a structural classification of haemagglutinin subtypes. Virology. 2004;325:287-96.

[34] Braakman I, Hoover-Litty H, Wagner KR, Helenius A. Folding of influenza hemagglutinin in the endoplasmic reticulum. J Cell Biol. 1991;114:401-11.

[35] Porter AG, Barber C, Carey NH, Hallewell RA, Threlfall G, Emtage JS. Complete nucleotide sequence of an influenza virus haemagglutinin gene from cloned DNA. Nature. 1979;282:471-7.

[36] Klenk HD, Garten W. Host cell proteases controlling virus pathogenicity. Trends Microbiol. 1994;2:39-43.

[37] Caton AJ, Brownlee GG, Yewdell JW, Gerhard W. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell. 1982;31:417-27.

[38] Weis W, Brown JH, Cusack S, Paulson JC, Skehel JJ, Wiley DC. Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. Nature. 1988;333:426-31.

[39] Couceiro JN, Paulson JC, Baum LG. Influenza virus strains selectively recognize sialyloligosaccharides on human respiratory epithelium; the role of the host cell in selection of hemagglutinin receptor specificity. Virus Res. 1993;29:155-65.

[40] Ito T, Couceiro JN, Kelm S, Baum LG, Krauss S, Castrucci MR, et al. Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. J Virol. 1998;72:7367-73.

[41] Rogers GN, Paulson JC, Daniels RS, Skehel JJ, Wilson IA, Wiley DC. Single amino acid substitutions in influenza haemagglutinin change receptor binding specificity. Nature. 1983;304:76-8.

[42] Daniels RS, Downie JC, Hay AJ, Knossow M, Skehel JJ, Wang ML, et al. Fusion mutants of the influenza virus hemagglutinin glycoprotein. Cell. 1985;40:431-9.

[43] Thoennes S, Li ZN, Lee BJ, Langley WA, Skehel JJ, Russell RJ, et al. Analysis of residues near the fusion peptide in the influenza hemagglutinin structure for roles in triggering membrane fusion. Virology. 2008;370:403-14.

[44] Air GM. Sequence relationships among the hemagglutinin genes of 12 subtypes of influenza A virus. Proc Natl Acad Sci U S A. 1981;78:7639-43.

[45] Hirst GK. Adsorption of Influenza Hemagglutinins and Virus by Red Blood Cells. J Exp Med. 1942;76:195-209.

[46] Sylte MJ, Suarez DL. Influenza neuraminidase as a vaccine antigen. Curr Top Microbiol Immunol. 2009;333:227-41.

[47] Colman PM, Varghese JN, Laver WG. Structure of the catalytic and antigenic sites in influenza virus neuraminidase. Nature. 1983;303:41-4.

[48] Mehta T, McGrath E, Bheemreddy S, Salimnia H, Abdel-Haq N, Ang JY, et al. Detection of oseltamivir resistance during treatment of 2009 H1N1 influenza virus infection in immunocompromised patients: utility of cycle threshold values of qualitative real-time reverse transcriptase PCR. J Clin Microbiol. 2010;48:4326-8.

[49] Marcelin G, Sandbulte MR, Webby RJ. Contribution of antibody production against neuraminidase to the protection afforded by influenza vaccines. Rev Med Virol. 2012;22:267-79.

[50] Matlin KS, Reggio H, Helenius A, Simons K. Infectious entry pathway of influenza virus in a canine kidney cell line. J Cell Biol. 1981;91:601-13.

[51] Maeda T, Ohnishi S. Activation of influenza virus by acidic media causes hemolysis and fusion of erythrocytes. FEBS Lett. 1980;122:283-7.

[52] Mast EE, Harmon MW, Gravenstein S, Wu SP, Arden NH, Circo R, et al. Emergence and possible transmission of amantadine-resistant viruses during nursing home outbreaks of influenza A (H3N2). Am J Epidemiol. 1991;134:988-97.

[53] Fodor E, Seong BL, Brownlee GG. Photochemical cross-linking of influenza A polymerase to its virion RNA promoter defines a polymerase binding site at residues 9 to 12 of the promoter. J Gen Virol. 1993;74 (Pt 7):1327-33.

[54] Engelhardt OG, Smith M, Fodor E. Association of the influenza A virus RNA-dependent RNA polymerase with cellular RNA polymerase II. J Virol. 2005;79:5812-8.

[55] Garaigorta U, Ortin J. Mutation analysis of a recombinant NS replicon shows that influenza virus NS1 protein blocks the splicing and nucleo-cytoplasmic transport of its own viral mRNA. Nucleic Acids Res. 2007;35:4573-82.

[56] Robb NC, Fodor E. The accumulation of influenza A virus segment 7 spliced mRNAs is regulated by the NS1 protein. J Gen Virol. 2012;93:113-8.

[57] Jorba N, Coloma R, Ortin J. Genetic trans-complementation establishes a new model for influenza virus RNA transcription and replication. PLoS Pathog. 2009;5:e1000462.

[58] Tchatalbachev S, Flick R, Hobom G. The packaging signal of influenza viral RNA molecules. RNA. 2001;7:979-89.

[59] Liang Y, Hong Y, Parslow TG. cis-Acting packaging signals in the influenza virus PB1, PB2, and PA genomic RNA segments. J Virol. 2005;79:10348-55.

[60] Xu R, Zhu X, McBride R, Nycholat CM, Yu W, Paulson JC, et al. Functional balance of the hemagglutinin and neuraminidase activities accompanies the emergence of the 2009 H1N1 influenza pandemic. J Virol. 2012;86:9221-32.

[61] Parks CL, Latham T, Cahill A, O'Neill R E, Passarotti CJ, Buonagurio DA, et al. Phenotypic properties resulting from directed gene segment reassortment between wild-type A/Sydney/5/97 influenza virus and the live attenuated vaccine strain. Virology. 2007;367:275-87.

[62] Zhang Y, Lin X, Wang G, Zhou J, Lu J, Zhao H, et al. Neuraminidase and hemagglutinin matching patterns of a highly pathogenic avian and two pandemic H1N1 influenza A viruses. PLoS One. 2010;5:e9167.

[63] Castrucci MR, Kawaoka Y. Biologic importance of neuraminidase stalk length in influenza A virus. J Virol. 1993;67:759-64.

[64] Baum LG, Paulson JC. The N2 neuraminidase of human influenza virus has acquired a substrate specificity complementary to the hemagglutinin receptor specificity. Virology. 1991;180:10-5.

[65] Xu R, McBride R, Nycholat CM, Paulson JC, Wilson IA. Structural characterization of the hemagglutinin receptor specificity from the 2009 H1N1 influenza pandemic. J Virol. 2012;86:982-90.

[66] Ginting TE, Shinya K, Kyan Y, Makino A, Matsumoto N, Kaneda S, et al. Amino acid changes in hemagglutinin contribute to the replication of oseltamivir-resistant H1N1 influenza viruses. J Virol. 2012;86:121-7.

[67] Russier M, Yang G, Rehg JE, Wong SS, Mostafa HH, Fabrizio TP, et al. Molecular requirements for a pandemic influenza virus: An acid-stable hemagglutinin protein. Proc Natl Acad Sci U S A. 2016;113:1636-41.

[68] Okuno Y, Isegawa Y, Sasao F, Ueda S. A common neutralizing epitope conserved between the hemagglutinins of influenza A virus H1 and H2 strains. J Virol. 1993;67:2552-8.

[69] Corti D, Cameroni E, Guarino B, Kallewaard NL, Zhu Q, Lanzavecchia A. Tackling influenza with broadly neutralizing antibodies. Curr Opin Virol. 2017;24:60-9.

[70] Laursen NS, Wilson IA. Broadly neutralizing antibodies against influenza viruses. Antiviral Res. 2013;98:476-83.

[71] Wollacott AM, Boni MF, Szretter KJ, Sloan SE, Yousofshahi M, Viswanathan K, et al. Safety and Upper Respiratory Pharmacokinetics of the Hemagglutinin Stalk-Binding Antibody VIS410 Support Treatment and Prophylaxis Based on Population Modeling of Seasonal Influenza A Outbreaks. EBioMedicine. 2016;5:147-55.

[72] Nakamura G, Chai N, Park S, Chiang N, Lin Z, Chiu H, et al. An in vivo human-plasmablast enrichment technique allows rapid identification of therapeutic influenza A antibodies. Cell Host Microbe. 2013;14:93-103.

[73] Chai N, Swem LR, Reichelt M, Chen-Harris H, Luis E, Park S, et al. Two Escape Mechanisms of Influenza A Virus to a Broadly Neutralizing Stalk-Binding Antibody. PLoS Pathog. 2016;12:e1005702.

[74] Pica N, Palese P. Toward a universal influenza virus vaccine: prospects and challenges. Annu Rev Med. 2013;64:189-202.

[75] CDC. How Influenza (Flu) Vaccines Are Made. Atlanta, Georgia: CDC; 2014.

[76] CDC. Selecting Viruses for the Seasonal Influenza Vaccine. Atlanta, Georgia: CDC; 2016.

[77] Whitaker-Dowling P, Lucas W, Youngner JS. Cold-adapted vaccine strains of influenza A virus act as dominant negative mutants in mixed infections with wild-type influenza A virus. Virology. 1990;175:358-64.

[78] Whitaker-Dowling P, Maassab HF, Youngner JS. Dominant-negative mutants as antiviral agents: simultaneous infection with the cold-adapted live-virus vaccine for influenza A protects ferrets from disease produced by wild-type influenza A. J Infect Dis. 1991;164:1200-2.

[79] Gubareva LV. Molecular mechanisms of influenza virus resistance to neuraminidase inhibitors. Virus Res. 2004;103:199-203.

[80] Lam TT, Chong YL, Shi M, Hon CC, Li J, Martin DP, et al. Systematic phylogenetic analysis of influenza A virus reveals many novel mosaic genome segments. Infect Genet Evol. 2013;18:367-78.

[81] Suzuki Y, Nei M. Origin and evolution of influenza virus hemagglutinin genes. Mol Biol Evol. 2002;19:501-9.

[82] Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. Predicting the evolution of human influenza A. Science. 1999;286:1921-5.

[83] Chan JM, Carlsson G, Rabadan R. Topology of viral evolution. Proc Natl Acad Sci U S A. 2013;110:18566-71.

[84] Plotch SJ, Bouloy M, Ulmanen I, Krug RM. A unique cap(m7GpppXm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription. Cell. 1981;23:847-58.

[85] Parvin JD, Palese P, Honda A, Ishihama A, Krystal M. Promoter analysis of influenza virus RNA polymerase. J Virol. 1989;63:5142-52.

[86] Luytjes W, Krystal M, Enami M, Parvin JD, Palese P. Amplification, expression, and packaging of foreign gene by influenza virus. Cell. 1989;59:1107-13.

[87] Liu HL, Wei JC, Zhou B, Cao RB, Chen PY. [PCR site-directed mutagenesis of avian influenza virus hemagglutinin gene in vitro and expression in 293T cell]. Wei Sheng Wu Xue Bao. 2005;45:614-6.

[88] Morrison KL, Weiss GA. Combinatorial alanine-scanning. Curr Opin Chem Biol. 2001;5:302-7.

[89] Wright S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution Proceedings of the Sixth International Congress on Genetics. 1932:355-66.

[90] Hietpas R, Roscoe B, Jiang L, Bolon DN. Fitness analyses of all possible point mutations for regions of genes in yeast. Nat Protoc. 2012;7:1382-96.

[91] Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, et al. High-resolution mapping of protein sequence-function relationships. Nat Methods. 2010;7:741-6.

[92] Melamed D, Young DL, Gamble CE, Miller CR, Fields S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. RNA. 2013;19:1537-51.

[93] Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, et al. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol. 2012;30:543-8.

[94] Wu NC, Olson CA, Du Y, Le S, Tran K, Remenyi R, et al. Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality. PLoS Genet. 2015;11:e1005310.

[95] McLaughlin RN, Jr., Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature. 2012;491:138-42.

[96] Ohta T. Slightly deleterious mutant substitutions in evolution. Nature. 1973;246:96-8.

[97] Bank C, Hietpas RT, Wong A, Bolon DN, Jensen JD. A bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: uncovering the potential for adaptive walks in challenging environments. Genetics. 2014;196:841-52.

[98] Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods. 2008;5:1005-10.

[99] Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11:31-46.

[100] Jiang L, Liu P, Bank C, Renzette N, Prachanronarong K, Yilmaz LS, et al. A Balance between Inhibitor Binding and Substrate Processing Confers Influenza Drug Resistance. J Mol Biol. 2016;428:538-53.

[101] Bloom MBDJD. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. Viruses. 2016;8:155.

[102] Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, Qi H, et al. High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. Sci Rep. 2014;4:4942.

[103] Ashenberg O, Padmakumar J, Doud MB, Bloom JD. Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by MxA. PLoS Pathog. 2017;13:e1006288.

[104] Doud MB, Hensley SE, Bloom JD. Complete mapping of viral escape from neutralizing antibodies. PLoS Pathog. 2017;13:e1006271.

[105] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111-20.

[106] Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 2006;7:98-108.

[107] Gu W, Zhou T, Wilke CO. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol. 2010;6:e1000664.

[108] Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12:32-42.

[109] Andersson SG, Kurland CG. Codon preferences in free-living microorganisms. Microbiol Rev. 1990;54:198-210.

[110] Duret L. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. 2000;16:287-9.

[111] Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol. 1981;151:389-409.

[112] Moriyama EN, Powell JR. Codon usage bias and tRNA abundance in Drosophila. J Mol Evol. 1997;45:514-23.

[113] Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet. 2008;42:287-99.

[114] Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 2008;134:341-52.

[115] Acevedo A, Brodsky L, Andino R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature. 2014;505:686-90.

[116] Boucher JI, Bolon DN, Tawfik DS. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. Protein Sci. 2016;25:1219-26.

[117] Hietpas RT, Bank C, Jensen JD, Bolon DN. Shifting fitness landscapes in response to altered environments. Evolution. 2013;67:3512-22.

[118] Roscoe BP, Thayer KM, Zeldovich KB, Fushman D, Bolon DN. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. J Mol Biol. 2013;425:1363-77.

[119] Jiang L, Mishra P, Hietpas RT, Zeldovich KB, Bolon DN. Latent effects of Hsp90 mutants revealed at reduced expression levels. PLoS Genet. 2013;9:e1003600.

[120] Bershtein S, Choi JM, Bhattacharyya S, Budnik B, Shakhnovich E. Systems-level response to point mutations in a core metabolic enzyme modulates genotype-phenotype relationship. Cell Rep. 2015;11:645-56.

[121] Mayrose I, Stern A, Burdelova EO, Sabo Y, Laham-Karam N, Zamostiano R, et al. Synonymous site conservation in the HIV-1 genome. BMC Evol Biol. 2013;13:164.

[122] Qin H, Wu WB, Comeron JM, Kreitman M, Li WH. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics. 2004;168:2245-60.

[123] Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A comprehensive, high-resolution map of a gene's fitness landscape. Mol Biol Evol. 2014;31:1581-92.

[124] Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, et al. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell. 2010;141:344-54.

[125] Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in Escherichia coli. Science. 2009;324:255-8.

[126] Tautenberger JK, Morens DM. 1918 Influenza: The mother of all pandemics. . CDC Emerging infectious diseases. 2006;12:15-22.

[127] Varghese JN, Colman PM. Three-dimensional structure of the neuraminidase of influenza virus A/Tokyo/3/67 at 2.2 A resolution. J Mol Biol. 1991;221:473-86.

[128] Blobel G, Sabatini D. Dissociation of mammalian polyribosomes into subunits by puromycin. Proc Natl Acad Sci U S A. 1971;68:390-4.

[129] Blobel G, Dobberstein B. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. Journal of Cell Biology. 1975;67:835-51.

[130] von Heijne G. Signal sequences: The limits of variation Journal of Molecular Biology. 1985;184:99-105.

[131] Walter P, Blobel G. Purification of a membrane-associated protein complex required for protein translocation across the endoplasmic reticulum. Proc Natl Acad Sci U S A. 1980;77:7112-6.

[132] Gilmore R, Walter P, Blobel G. Protein translocation across the endoplasmic reticulum. II. Isolation and characterization of the signal recognition particle receptor. J Cell Biol. 1982;95:470-7.

[133] Dalbey RE, Von Heijne G. Signal peptidases in prokaryotes and eukaryotes--a new protease family. Trends Biochem Sci. 1992;17:474-8.

[134] von Heijne G. Patterns of amino acids near signal-sequence cleavage sites. Eur J Biochem. 1983;133:17-21.

[135] Air GM. Nucleotide sequence coding for the "signal peptide" and N terminus of the hemagglutinin from an asian (H2N2) strain of influenza virus. Virology. 1979;97:468-72.

[136] Inouye S, Soberon X, Franceschini T, Nakamura K, Itakura K, Inouye M. Role of positive charge on the amino-terminal region of the signal peptide in protein secretion across the membrane. Proc Natl Acad Sci U S A. 1982;79:3438-41.

[137] Izard JW, Rusch SL, Kendall DA. The amino-terminal charge and core region hydrophobicity interdependently contribute to the function of signal sequences. J Biol Chem. 1996;271:21579-82.

[138] Weltman JK, Skowron G, Loriot GB. Influenza A H5N1 hemagglutinin cleavable signal sequence substitutions. Biochem Biophys Res Commun. 2007;352:177-80.

[139] Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 2010;38:W7-13.

[140] Sekikawa K, Lai C. Defects in functional expression of an influenza virus hemagglutinin lacking the signal peptide sequences. Proceedings of the National Academy of Sciences of the United States of America. 1983;80:3563-7.

[141] Thyagarajan B, Bloom JD. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. eLIFE. 2014;3.

[142] Doud MB, Bloom JD. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. Viruses. 2016;8.

[143] Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, Qi H, et al. High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. Sientific Reports. 2014;4:4942.

[144] Hietpas RT, Roscoe BP, Jiang L, Bolon DNA. Fitness analyses of all possible point mutations for regions of genes in yeast. Nature Protocols. 2012;7:1382-96.

[145] Hietpas RT, Jensen JD, Bolon DNA. Experimental illumination of a fitness landscape. Proceedings of the National Academy of Sciences of the United States of America. 2011;108:7896-901.

[146] Squires RB, Noronha J, Hunt V, Garcia-Sastre A, Macken C, Baumgarth N, et al. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza Other Respir Viruses. 2012;6:404-16.

[147] Zhang Y, Aevermann BD, Anderson TK, Burke DF, Dauphin G, Gu Z, et al. Influenza Research Database: An integrated bioinformatics resource for influenza virus research. Nucleic Acids Res. 2017;45:D466-D74.

[148] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157:105-32.

[149] Sharp PM, Devine KM. Codon usage and gene expression level in Dictyostelium discoideum: highly expressed genes do 'prefer' optimal codons. Nucleic Acids Res. 1989;17:5029-39.

[150] Stoletzki N, Eyre-Walker A. Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol Biol Evol. 2007;24:374-81.

[151] Choi JY, Aquadro CF. Recent and Long-Term Selection Across Synonymous Sites in Drosophila ananassae. J Mol Evol. 2016;83:50-60.

[152] Singh ND, Bauer DuMont VL, Hubisz MJ, Nielsen R, Aquadro CF. Patterns of mutation and selection at synonymous sites in Drosophila. Mol Biol Evol. 2007;24:2687-97.

[153] Goodman DB, Church GM, Kosuri S. Causes and effects of N-terminal codon bias in bacterial genes. Science. 2013;342:475-9.

[154] Paradis EG, Pinilla LT, Holder BP, Abed Y, Boivin G, Beauchemin CA. Impact of the H275Y and I223V Mutations in the Neuraminidase of the 2009 Pandemic Influenza Virus In Vitro and Evaluating Experimental Reproducibility. PLoS One. 2015;10:e0126115.

[155] Bedouelle H. Mutations in the promoter regions of the malEFG and malK-lamB operons of Escherichia coli K12. J Mol Biol. 1983;170:861-82.

[156] Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. Rate-limiting steps in yeast protein translation. Cell. 2013;153:1589-601.

[157] Watanabe T, Watanabe S, Noda T, Fujii Y, Kawaoka Y. Exploitation of nucleic acid packaging signals to generate a novel influenza virus-based vector stably expressing two foreign genes. J Virol. 2003;77:10575-83.

[158] Marsh GA, Hatami R, Palese P. Specific residues of the influenza A virus hemagglutinin viral RNA are important for efficient packaging into budding virions. J Virol. 2007;81:9727-36.
[159] Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26.
[160] Nogales A, Baker SF, Ortiz-Riaño E, Dewhurst S, Topham DJ, Martínez-Sobrido L. Influenza A virus attenuation by codon deoptimization of the NS gene for vaccine development. Journal of Virology. 2014;88:10525-40.
[161] Gavazzi C, Yver M, Isel C, Smyth RP, Rosa-Calatrava M, Lina B, et al. A functional sequence-specific interaction between influenza A virus genomic RNA segments. Proc Natl Acad Sci U S A. 2013;110:16604-9.
[162] Vasin AV, Petrova AV, Egorov VV, Plotnikova MA, Klotchenko SA, Karpenko MN, et al. The influenza A virus NS genome segment displays lineage-specific patterns in predicted RNA secondary structure. BMC Res Notes. 2016;9:279.
[163] Gultyaev AP, Spronken MI, Richard M, Schrauwen EJ, Olsthoorn RC, Fouchier RA. Subtype-specific structural constraints in the evolution of influenza A virus hemagglutinin genes. Sci Rep. 2016;6:38892.
[164] Gultyaev AP, Tsyganov-Bodounov A, Spronken MI, van der Kooij S, Fouchier RA, Olsthoorn RC. RNA structural constraints in the evolution of the influenza A virus genome NP segment. RNA Biol. 2014;11:942-52.
[165] Moss WN, Priore SF, Turner DH. Identification of potential conserved RNA secondary structure throughout influenza A coding regions. RNA. 2011;17:991-1011.
[166] Williams EJ, Pal C, Hurst LD. The molecular evolution of signal peptides. Gene. 2000;253:313-22.
[167] Li YD, Xie ZY, Du YL, Zhou Z, Mao XM, Lv LX, et al. The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. Gene. 2009;436:8-11.
[168] Hoffmann E, Krauss S, Perez D, Webby R, Webster RG. Eight-plasmid system for rapid generation of influenza virus vaccines. Vaccine. 2002;20:3165-70.
[169] Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, 3rd, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods. 2009;6:343-5.
[170] Jiang L, Liu P, Bank C, Renzette N, Prachanronarong K, Yilmaz LS, et al. A Balance between Inhibitor Binding and Substrate Processing Confers Influenza Drug Resistance. Journal of Molecular Biology. 2016;428:538-53.
[171] Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res. 2001;29:e45.
[172] Holder BP, Beauchemin CA. Exploring the effect of biological delays in kinetic models of influenza within a host or cell culture. BMC Public Health. 2011;11 Suppl 1:S10.
[173] Connor RJ, Kawaoka Y, Webster RG, Paulson JC. Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates. Virology. 1994;205:17-23.
[174] Rudneva IA, Kovaleva VP, Varich NL, Farashyan VR, Gubareva LV, Yamnikova SS, et al. Influenza A virus reassortants with surface glycoprotein genes of the avian parent viruses: effects of HA and NA gene combinations on virus aggregation. Arch Virol. 1993;133:437-50.
[175] Rudneva IA, Sklyanskaya EI, Barulina OS, Yamnikova SS, Kovaleva VP, Tsvetkova IV, et al. Phenotypic expression of HA-NA combinations in human-avian influenza A virus reassortants. Arch Virol. 1996;141:1091-9.

[176] Kaverin NV, Gambaryan AS, Bovin NV, Rudneva IA, Shilov AA, Khodova OM, et al. Postreassortment changes in influenza A virus hemagglutinin restoring HA-NA functional match. Virology. 1998;244:315-21.

[177] Els MC, Air GM, Murti KG, Webster RG, Laver WG. An 18-amino acid deletion in an influenza neuraminidase. Virology. 1985;142:241-7.

[178] Mitnaul LJ, Matrosovich MN, Castrucci MR, Tuzikov AB, Bovin NV, Kobasa D, et al. Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza A virus. J Virol. 2000;74:6015-20.

[179] Wagner R, Wolff T, Herwig A, Pleschka S, Klenk HD. Interdependence of hemagglutinin glycosylation and neuraminidase as regulators of influenza virus growth: a study by reverse genetics. J Virol. 2000;74:6316-23.

[180] Ohuchi M, Ohuchi R, Feldmann A, Klenk HD. Regulation of receptor binding affinity of influenza virus hemagglutinin by its carbohydrate moiety. J Virol. 1997;71:8377-84.

[181] Kim CU, Lew W, Williams MA, Wu H, Zhang L, Chen X, et al. Structure-activity relationship studies of novel carbocyclic influenza neuraminidase inhibitors. J Med Chem. 1998;41:2451-60.

[182] CDC. Influenza Antiviral Medications: Summary for Clinicians. Atlanta, Georgia: CDC; 2017.

[183] Tai CY, Escarpe PA, Sidwell RW, Williams MA, Lew W, Wu H, et al. Characterization of human influenza virus variants selected in vitro in the presence of the neuraminidase inhibitor GS 4071. Antimicrob Agents Chemother. 1998;42:3234-41.

[184] Hurt AC, Holien JK, Barr IG. In vitro generation of neuraminidase inhibitor resistance in A(H5N1) influenza viruses. Antimicrob Agents Chemother. 2009;53:4433-40.

[185] Sheu TG, Deyde VM, Okomo-Adhiambo M, Garten RJ, Xu X, Bright RA, et al. Surveillance for neuraminidase inhibitor resistance among human influenza A and B viruses circulating worldwide from 2004 to 2008. Antimicrob Agents Chemother. 2008;52:3284-92.

[186] Malaisree M, Rungrotmongkol T, Nunthaboot N, Aruksakunwong O, Intharathep P, Decha P, et al. Source of oseltamivir resistance in avian influenza H5N1 virus with the H274Y mutation. Amino Acids. 2009;37:725-32.

[187] Karthick V, Shanthi V, Rajasekaran R, Ramanathan K. Exploring the cause of oseltamivir resistance against mutant H274Y neuraminidase by molecular simulation approach. Appl Biochem Biotechnol. 2012;167:237-49.

[188] Baigent SJ, Bethell RC, McCauley JW. Genetic analysis reveals that both haemagglutinin and neuraminidase determine the sensitivity of naturally occurring avian influenza viruses to zanamivir in vitro. Virology. 1999;263:323-38.

[189] Blick TJ, Sahasrabudhe A, McDonald M, Owens IJ, Morley PJ, Fenton RJ, et al. The interaction of neuraminidase and hemagglutinin mutations in influenza virus in resistance to 4-guanidino-Neu5Ac2en. Virology. 1998;246:95-103.

[190] Bradley KC, Galloway SE, Lasanajak Y, Song X, Heimburg-Molinaro J, Yu H, et al. Analysis of influenza virus hemagglutinin receptor binding mutants with limited receptor recognition properties and conditional replication characteristics. J Virol. 2011;85:12387-98.

[191] Wagner R, Matrosovich M, Klenk HD. Functional balance between haemagglutinin and neuraminidase in influenza virus infections. Rev Med Virol. 2002;12:159-66.

[192] Kiso M, Mitamura K, Sakai-Tagawa Y, Shiraishi K, Kawakami C, Kimura K, et al. Resistant influenza A viruses in children treated with oseltamivir: descriptive study. Lancet. 2004;364:759-65.

[193] Marjuki H, Mishin VP, Chesnokov AP, De La Cruz JA, Davis CT, Villanueva JM, et al. Neuraminidase Mutations Conferring Resistance to Oseltamivir in Influenza A(H7N9) Viruses. J Virol. 2015;89:5419-26.

[194] Gen F, Yamada S, Kato K, Akashi H, Kawaoka Y, Horimoto T. Attenuation of an influenza A virus due to alteration of its hemagglutinin-neuraminidase functional balance in mice. Arch Virol. 2013;158:1003-11.

[195] Baigent SJ, McCauley JW. Glycosylation of haemagglutinin and stalk-length of neuraminidase combine to regulate the growth of avian influenza viruses in tissue culture. Virus Res. 2001;79:177-85.

[196] Yen HL, Liang CH, Wu CY, Forrest HL, Ferguson A, Choy KT, et al. Hemagglutinin-neuraminidase balance confers respiratory-droplet transmissibility of the pandemic H1N1 influenza virus in ferrets. Proc Natl Acad Sci U S A. 2011;108:14264-9.

[197] Renzette N, Caffrey DR, Zeldovich KB, Liu P, Gallagher GR, Aiello D, et al. Evolution of the influenza A virus genome during development of oseltamivir resistance in vitro. J Virol. 2014;88:272-81.

[198] Schrodinger L. New York2015.

[199] Ozen A, Lin KH, Kurt Yilmaz N, Schiffer CA. Structural basis and distal effects of Gag substrate coevolution in drug resistance to HIV-1 protease. Proc Natl Acad Sci U S A. 2014;111:15993-8.

[200] Walter P, Blobel G. Translocation of proteins across the endoplasmic reticulum. II. Signal recognition protein (SRP) mediates the selective binding to microsomal membranes of in-vitro-assembled polysomes synthesizing secretory protein. J Cell Biol. 1981;91:551-6.

[201] Moss WN, Priore SF, Turner DH. Identification of potential conserved RNA secondary structure throughout influenza A coding regions. Bionformatics. 2011;17:991-1011.

[202] Benton DJ, Martin SR, Wharton SA, McCauley JW. Biophysical measurement of the balance of influenza a hemagglutinin and neuraminidase activities. J Biol Chem. 2015;290:6516-21.

[203] Sieben C, Kappel C, Zhu R, Wozniak A, Rankl C, Hinterdorfer P, et al. Influenza virus binds its host cell using multiple dynamic interactions. Proc Natl Acad Sci U S A. 2012;109:13626-31.

[204] Wagenaar TR, Ma L, Roscoe B, Park SM, Bolon DN, Green MR. Resistance to vemurafenib resulting from a novel mutation in the BRAFV600E kinase domain. Pigment Cell Melanoma Res. 2014;27:124-33.

[205] Ma L, Boucher JI, Paulsen J, Matuszewski S, Eide CA, Ou J, et al. CRISPR-Cas9-mediated saturated mutagenesis screen predicts clinical drug resistance with improved accuracy. Proc Natl Acad Sci U S A. 2017;114:11751-6.

[206] Duenas-Decamp M, Jiang L, Bolon D, Clapham PR. Saturation Mutagenesis of the HIV-1 Envelope CD4 Binding Loop Reveals Residues Controlling Distinct Trimer Conformations. PLoS Pathog. 2016;12:e1005988.

[207] Davidsson M, Diaz-Fernandez P, Schwich OD, Torroba M, Wang G, Bjorklund T. A novel process of viral vector barcoding and library preparation enables high-diversity library generation and recombination-free paired-end sequencing. Sci Rep. 2016;6:37563.

[208] Belin D, Bost S, Vassalli JD, Strub K. A two-step recognition of signal sequences determines the translocation efficiency of proteins. EMBO J. 1996;15:468-78.

[209] von Heijne G. A new method for predicting signal sequence cleavage sites. Nucleic Acids Res. 1986;14:4683-90.

[210] Herskowitz I. Functional inactivation of genes by dominant negative mutations. Nature. 1987;329:219-22.

[211] Burz DS, Hanes SD. Isolation of mutations that disrupt cooperative DNA binding by the Drosophila bicoid protein. J Mol Biol. 2001;305:219-30.

[212] Fukazawa J, Sakai T, Ishida S, Yamaguchi I, Kamiya Y, Takahashi Y. Repression of shoot growth, a bZIP transcriptional activator, regulates cell elongation by controlling the level of gibberellins. Plant Cell. 2000;12:901-15.

[213] Wilson JH, Wensel TG. The nature of dominant mutations of rhodopsin and implications for gene therapy. Mol Neurobiol. 2003;28:149-58.

[214] Maloy ML, Whitaker-Dowling P, Youngner JS. Dominance of cold-adapted influenza A virus over wild-type viruses is at the level of RNA synthesis. Virology. 1994;205:44-50.

[215] von Heijne G. Signal sequences. The limits of variation. J Mol Biol. 1985;184:99-105.

[216] Phillips AM, Gonzalez LO, Nekongo EE, Ponomarenko AI, McHugh SM, Butty VL, et al. Host proteostasis modulates influenza evolution. Elife. 2017;6.