Translation initiation during influenza virus infection and its role in cellular immunity

Heather M Machkovech

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

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Jesse D Bloom, Chair

Arvind R Subramaniam

Adam P Geballe

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Translation initiation during influenza virus infection and its role in cellular immunity

Heather M Machkovech

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

There are numerous factors that direct and constrain influenza evolution. On the one hand influenza virus is constrained to encode proteins that allow it to complete the viral life cycle, but the immune system also recognizes those same viral proteins to mount an immune response. The interplay between selection for viral propagation and escape from immunity shapes the rapid evolution of influenza virus.

In Chapter 2, I explore one of the immune pressures that drives influenza evolution. It is well known that antibodies are a driver of the rapid evolution of influenza. However, it was unknown whether CD8 T-cells, which are also a component of adaptive immunity to influenza, also shape influenza evolution. There is experimental support that CD8 T-cells help control influenza infection and there are even documented instances of mutations that confer escape from CD8 T-cells. However, conventional tests of selection consistently fail to find evidence that CD8 T-cells drive influenza evolution. Since CD8 T-cells tend to be found in regions that are fairly conserved in influenza, I hypothesized that CD8 T-cells are a selective pressure but that functional constraint masked our ability to detect such selection by conventional means. I therefore employed two novel statistical methods that have increased sensitivity to detect selection in influenza. I leveraged the existence of parallel lineages of influenza that infect different hosts with varying immune pressures to examine whether CD8 T-cells drive influenza evolution. I find statistical support that

that one influenza protein, nucleoprotein, is under selective pressure to evade the CD8 T-cell response. This means that the role of CD8 T-cells is sufficiently strong to shape the evolutionary trajectory of influenza virus. There is strong interest in generating vaccines with broad responses to influenza. Since T-cells are capable of eliciting broad responses, the finding that the biological role of T-cells is strong enough to shape influenza evolution validates the use of T-cells in the development of broad-acting vaccines.

In Chapter 3, I explore translation initiation in the context of influenza virus infection. Because CD8 T-cells shape influenza evolution and non-canonical translation initiation can generate novel CD8 T-cell epitopes, I examined whether there was an evolutionary signature consistent with selection against alternate translation initiation in influenza virus. I find evolutionary support that is consistent with selection against alternate translation initiation of CTG codons in mammalian lineages of influenza. I then performed ribosome profiling, a deep sequencing technique that enables capture of ribosome protected fragments, to annotate sites of translation initiation in the influenza genome. I did not find evidence of CTG initiation in the influenza viral genome, but I did find a small number of alternate translation initiation sites at ATG codons in the influenza viral genome. One of these alternate initiation sites generates an N-terminally truncated form of neuraminidase (NA), the influenza protein that mediates viral egress. This N-terminally truncated NA lacks the first fourteen amino acids, but is enzymatically active, supports viral growth, and is broadly conserved in the N1 lineage. In the Discussion, I speculate about the possible discordance of the evolutionary results and ribosome profiling data. Overall, this work increases our understanding of the range of viral proteins that are generated in the compact influenza genome and uncovers new evolutionary signatures of influenza virus.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my advisor Jesse Bloom for his guidance, knowledge, patience, and optimism. One goal I had in graduate school was to learn computational biology, and as is evident here, he has been extremely supportive of that goal. I am also particularly thankful for the kind and friendly lab culture that Jesse has created.

I thank members of the Bloom lab, past and present. Mike Doud is my 'big doc' in the MSTP program and was a major inspiration for rotating in the Bloom lab, and ever since he has continued to fulfill the role of a supportive colleague. I would especially like to thank Orr Ashenberg, Shirleen Soh, and Alistair Russell for their scientific advice, comedic relief, and moral support. I am thankful for the shared appreciation of the outdoors with fellow lab members - Mt Saint Helens with Orr, Juhye, Adam, and Mike, secret beaches with Sarah, tennis with Adam, and cross country skiing with Orr and Mike.

I thank the individuals I have been fortunate to collaborate with and be advised by. To Trevor Bedford, who was a collaborator on my first project in lab. Trevor provided much guidance for the analyses we developed and helped shape the arc of the story in a supportive and collaborative manner. I thank Rasi Subramaniam, an advisor on my project examining translation in the context of influenza infection. This collaboration gave to the opportunity to work on a project that was outside the general scope of the Bloom lab.

I thank my other Committee members: Adam Geballe, Julie Overbaugh, Phil Greenberg, and Mike MacCoss. I appreciate your advice on thinking critically about the big picture of research projects, and practical guidance for achieving those aims.

I thank my friends and colleagues. In particular, I am thankful for my friendship with

Rebecca Resnick and Sidney Bell. I appreciate the supportive community we have. I thank my fellow cohort of MSTP students, for the shared journeys and adventures of our training. I thank Madeleine Geisheker, who along with Rebecca forms the contingent of women MSTP trainees in the entering class of 2012. I also thank my friends Marilyn Sherris and Caitlin Lannigan.

I thank past advisors who have helped me grow as a scientist and see allow me to see how they model science as a career. I thank Nicole Dudley who took the time as a graduate student to mentor me as an undergraduate. To Professor Josh Maurer, who mentored me in my first 'independent' research project. To Professors Dewey Holten and Kevin Moeller who combined their passion for creativity and science with teaching. To Jean Schaffer and Dan Ory who model careers as physician-scientists. To Joel Schilling, whose enthusiasm for studying host-pathogen interactions motivated me to do the same. I deeply appreciate your passion for mentorship and how you conduct your career as physician-scientist.

I thank the greater community at the Fred Hutch and UW that foster a collaborative and efficiently run scientific community. This includes the Virology Group, Computational Biology, UW Genome Sciences, and UW MSTP. This also includes administrative staff and Fred Hutch Cores who help make research at the Fred Hutch run smoothly.

Most importantly, I thank my family, particularly my parents for their support, patience, kindness, and love through my training. In addition to my immediate family, I am thankful for my cousin Sam, who makes up the other half of the Seattle contingent of the Machkovech family.

Chapter 1
INTRODUCTION

In this Introduction, I begin by providing background on influenza virus. I then provide background relevant to Chapter 2 by introducing the experimental evidence for the role that CD8 T-cells have during influenza infection. Based on the results in Chapter 2 where I find that CD8 T-cells shape influenza evolution, I examine the extent and impact of protein products produced by influenza virus in Chapter 3. I provide more in depth background on the methods by which influenza increases the coding capacity of its genome and how alternately initiated protein products could be deleterious to influenza if they serve as T-cell epitopes. I also provide background on the experimental method that I use to annotate sites of translation initiation in influenza virus.

1.1 Introduction to influenza A virus

1.1.1 Influenza ecology and epidemiology

Influenza A virus (referred to here as influenza virus) is a member of the Orthomyxoviridae family [56]. The natural reservoir of influenza is wild waterfowl, but influenza infects a wide host range including humans, pigs, dogs, horses, marine mammals, cats, and bats [217, 204]. Influenza viruses cause annual epidemics and occasional pandemics in humans. Influenza epidemics infect up to 10 % of the population each year, causing approximately three to five million cases of severe illness and resulting in approximately 250,000 to 500,000 deaths (http://www.who.int/immunization/topics/influenza/en/). In contrast to epidemics which are characterized by viruses that have undergone slight anti-

genic evolution (antigenic drift), pandemics occur when a highly diverged virus enters the human population from an animal reservoir (antigenic shift). Influenza pandemics can result in a very high burden of morbidity and mortality, as pre-existing immunity to the antigenically shifted virus is low. Influenza virus remains a major issue in public health, underscoring the importance of influenza research.

1.1.2 Influenza genome structure

Influenza virus is a segmented single stranded negative-sense RNA virus. The influenza viral genome is approximately 12 kilobases in size and consists of eight segments [56]. The eight viral RNA segments encode at least ten functional viral proteins: polymerase basic protein 2 (PB2), polymerase basic protein 1 (PB1), polymerase acidic protein (PA), hemagglutinin (HA), nucleoprotein (NP), neuraminidase (NA), matrix 1 (M1), matrix 2 (M2), non-structural protein 1 (NS1), and non-structural protein 2 (NS2) [56]. Influenza NP encapsidates influenza gene segments, and the encapsidated segments also associate with the other components of the polymerase complex (PA, PB1, and PB2) to form ribonucleoprotein (RNP) complexes [233]. NP is also involved in viral genome packaging during virion formation [233]. The polymerase proteins carry out viral genome replication (production of vRNA) and transcription of mRNA [56, 57]. NS2 mediates nuclear export of viral RNP complexes [151]. NS1 has been ascribed numerous roles, mainly involving antagonism of the host immune response [77]. The viral envelope is comprised of an outer host derived lipid bilayer containing the viral surface proteins HA and NA [56]. HA mediates viral entry and NA mediates viral egress from host cells [56]. Under the viral envelope, M1 forms a matrix structure that forms the basis for the virion shape, and interacts with the vRNP complexes during viral packaging [141]. M2 is an ion channel that facilitates viral entry into the cytoplasm through acidification of the virion [156].

1.1.3 Influenza virus replication

Influenza virus replication begins with the virus attaching to the host cell by binding the host receptor sialic acid [185]. Following binding, the influenza virion is internalized into the host cell by receptor mediated endocytosis [185]. The low pH of the endosome triggers fusion of the viral membrane to the host endosomal membrane [185]. This occurs through a series of conformational changes in HA, which result in the fusion peptide of HA being inserted into the endosomal membrane [185]. Subsequently, M2 mediates an influx of H+ into the virion, thus lowering the pH of the virion [185, 156]. The lowered pH releases the RNPs from M1, allowing vRNPs to enter the cytoplasm.

Influenza virus transcription and replication occurs in the nucleus. Therefore, following entry of vRNPs into the cytoplasm, vRNPs enter the nucleus. The proteins of the vRNP complex contain nuclear localization signals, allowing nuclear entry [54].

The negative sense genome of influenza virus is complementary to viral mRNA, so genomic RNA has to be transcribed to generate viral mRNA. Viral transcription is initiated in cis by the polymerase that is a component of the vRNP [58, 57]. Viral transcription is primed by host mRNA derived 5' methylated caps. This occurs by PB2 binding host pre-mRNA and PA subsequently cleaving the 5' cap [45]. Transcription of vRNA proceeds until a short stretch of U residues which functions as a signal for polyadenylation is reached [170]. The viral polymerase stutters on the stretch of U residues, and generates the polyA tail [157]. Splicing of the viral M and NS transcripts is accomplished by cellular machinery. Viral transcripts are subsequently exported from the nucleus for translation. Unlike viral replication which proceeds late into viral infection, viral transcription peaks earlier in the viral replication cycle [16].

Viral replication has two steps: generation of complementary RNA (cRNA) by replication of genomic vRNA, and subsequent generation of vRNA using cRNA as a template [57]. Viral replication does not require a primer and is thought to occur in trans by a viral polymerase complex that is distinct from the polymerase present in the vRNP com-

plex [38]. Encapsidation of cRNA and vRNA with NP occurs co-transcriptionally, resulting in the formation of cRNP and vRNPs. Following viral replication, progeny vRNPs are selectively exported, while cRNPs are restricted to the nucleus [57]. NS2 facilitates nuclear export [154].

The last step of influenza virus replication is viral budding. Virus particles bud from the apical plasma membrane of polarized cells [172, 141]. HA, NA, and M2 are transported to the plasma membrane and localize to lipid raft domains [172, 141]. M2 facilitates formation of viral particles [141]. While packaging of viral segments into the virion is incompletely understood, portions of the 3' and 5' ends of vRNA serve as packaging signals [87]. M1 helps close the viral particle. The viral particle is then released from the cell by cleavage of host sialic acid by NA [172].

1.2 CD8 T-cells and influenza

1.2.1 Adaptive immunity to influenza

The specific nature of the adaptive immune response (comprised of humoral and cellular immunity) requires time to develop when encountering a pathogen for the first time. As a result, the adaptive immune response has an enhanced role in subsequent encounters with a pathogen.

Both arms of the adaptive immune system help control influenza virus replication. The humoral response, through influenza specific antibodies, neutralizes virus [104] and facilitates removal of infected cells [98]. Cellular immunity is comprised of CD4 and CD8 T-cells. T-cells recognize viral antigen on cells through interactions of their T-cell receptors with viral peptides presented on MHC molecules. CD4 cells help regulate the immune response [30], while CD8+ T cells kill infected cells that display viral peptides on their major histocompatibility complex (MHC) class I molecules [208, 169]. In contrast to antibodies which often target the rapidly evolving HA [211, 36, 39], T cells can provide broader protection against diverse strains since they tend to recognize epitopes in more

conserved internal viral proteins such as nucleoprotein (NP) and matrix protein (M1) [208, 169, 228, 79].

1.2.2 Evidence for protective role of CD8 T-cells

CD8 T-cells facilitate viral clearance during influenza infections. Naiive CD8 T-cells present in the draining lymph nodes are activated when they encounter viral epitopes presented on the MHC class I molecules of antigen presenting cells. Subsequently the activated CD8 T-cells differentiate into cytotoxic T-cells (CTL). CTL cells migrate to sites of infection and engage in a lytic response to promote clearance of infected cells and halt propagation of viral production [188]. Following resolution of infection, long-lived memory CD8 T-cells develop, enabling rapid responses to subsequent exposures to antigen [188].

Both experimental [75, 115, 111, 202, 136] and epidemiological studies [136, 79, 189]. provide evidence that pre-existing influenza specific CTL responses are protective. For example, in an experimental infection of human subjects, the magnitude of lytic activity of PBMCs was found to be inversely correlated with viral shedding [136]. A study from the 2009 H1N1 pandemic found that the presence of influenza specific CD8 T-cells correlated with decreased disease severity [189]. Furthermore, CD8 T-cell responses were correlated with recovery from severe H7N9 infection [216]. Similarly, T cells specific for NP were associated with a decreased incidence of symptomatic infection over a multiyear study of a large human cohort [79]. Overall, these studies implicate an important role for CTL responses in controlling influenza infection.

1.2.3 Examples of escape from T-cell immunity

Viruses utilize numerous strategies to evade recognition by T-cells. These immune evasion strategies can either globally affect antigen processing or they can alter just a particular epitope. Viruses can encode proteins that directly antagonize a step of of the antigen processing or presentation pathway [150]. This strategy is often employed by viruses with

large genomes such as HCMV [150]. However, influenza does not encode any direct antagonists of the antigen processing pathway.

The second way viruses can escape T-cells is through abrogation of the epitope itself. This can occur by accumulation of mutations outside of the epitope that altering processing of the protein so that an epitope is no longer produced, mutations in the anchor residues of the epitope that abrogate binding to the MHC molecule, or mutations in the TCR contact residues of epitope that ablate binding of a TCR to an MHC epitope complex. This type of escape has been documented for viruses that chronically infect humans such as HIV and HCMV [150].

There is evidence that influenza can accumulate T-cell escape mutations by mutation of anchor residues and TCR contact residues. These escape mutations occurred in and became fixed in the human H3N2 lineage of NP. An example of a mutation in an anchor residue of an epitope is the R384G substitution. This mutation fixed in 1993 [213]. There are three examples of escape by mutation of TCR contact residues: D421E/I425V mutations that fixed in 1979 [22, 23], K103R mutation that fixed in 1980 [15], and S259L mutation that fixed in 1990 [15]. T-cell escape has also been documented during a single infection in an immunocompromised setting. In a mouse study, viral mutations arose that conferred T-cell escape in RAG-1-deficient mice expressing an influenza virus NP-specific T-cell receptor (TCR) [160]. CD8 T-cell escape mutations also arose in a persistently influenza-infected infant [208]. These studies demonstrate that influenza virus accumulates substitutions that escape CD8 T cells.

1.2.4 Do CD8 T-cells exert selective pressure on human influenza?

Since influenza virus infects humans numerous times over their lifespan, influenza virus may be under pressure to escape from T-cell immunity. While it is well appreciated that the antibody response is a driver of influenza evolution [186, 26], the role of T-cells as a selective force is largely unknown. So even while there are documented instances of

T-cell escape, these instances do not indicate that T-cells are a selective force, as these mutations could also have arisen by chance. Therefore, in Chapter 2 I explore whether CD8 T-cells are a selective pressure on influenza. Addressing this question increases our understanding of the breadth of forces that act on influenza evolution and has implications for vaccine design.

1.3 Translation initiation in the setting of viral infection

Protein translation is a dynamic process that occurs in three stages: translation initiation, translation elongation, and translation termination. Viral infection and the resulting anti-viral response of infected cells are a cellular stress that alter all stages of translation.

1.3.1 Methods for increasing peptide diversity in viruses

Viruses are required to encode the necessary proteins to successfully complete viral replication in a compact genome. Viruses utilize a number of clever strategies to accomplish this. Influenza increases protein coding capacity of its genome through use of alternative splicing, ribosomal frameshifting, and alternate translation initiation.

Influenza contains two genes that can be alternatively spliced: NS and M. The NS segment generates unspliced NS1 and spliced NS2 transcripts. The M segment generates up to four transcripts: M1 is unspliced, and M2 M3, and M4 are spliced [182, 117, 89]. M1 and M2 generate canonical influenza proteins. M3 has no observed protein products. M4 is present in a small number of viral strains, as the splice donor site is poorly conserved [221]. Initiation of the M4 transcript can generate a protein called M42, which can functionally complement the ion channel M2 [221].

Influenza also utilizes ribosomal frameshifting to increase coding diversity. Specifically, influenza encodes PA-X by ribosomal frameshifting on the PA segment [96]. This occurs when the ribosome encounters a rare amino acid, causing the ribosome to sometimes shift into the +1 reading frame. PA-X contains the same endonuclease domain present

in PA . In PA, this enables snatching of host mRNA caps, however the unique c-terminus of PA-X enables differential localization such that PA-X nonspecifically cleaves host pol II generated mRNAs, thus blunting the host response to infection [96, 102].

Influenza also utilizes alternate translation initiation. The identified alternate initiation sites are all downstream ATG codons. The best described alternately initiated influenza protein is PB1-F2 [35]. PB1-F2 is an influenza protein initiated by a downstream ATG in the +1 reading frame of the PB1 segment [35]. PB1-F2 is thought to modulate the host immune response and be a virulence factor by inducing apoptosis [35, 230, 135].

Additional peptides generated by alternate start sites have been noted in the PB1, PA, and M segments of influenza [35, 230, 220, 1, 2, 139, 221]. Translation initiation at ATG codons downstream and in-frame of PB1-F2 have been detected [35]. An N-terminally truncated form of PB1 (PB1-N40) has also been detected [220]. PB1-N40 has been shown to interact with polymerase subunits. Whether this protein has a function is unknown, but it is speculated that it may balance expression of PB1 and PB1-F2 [220]. Two N-terminal truncations with no known function have been described in PA - PA155 and PA185 [139]. M42 is also generated by alternate translation initiation [221]. These examples of alternately initiated peptides provide insight into the complex coding capacity of the influenza genome, however there has yet to be a comprehensive survey of influenza translation initiation sites.

1.3.2 Alternate translation initiation products can serve as T-cell epitopes

There are many documented cases in which T-cell epitopes are derived from unexpected source [24, 227, 193]. These epitopes are described as 'cryptic' T-cell epitopes and can be derived from the 5' UTR, introns, alternate reading frames, and 3' UTRs.

Many cryptic T-cell epitopes are generated by alternate translation initiation. Though analysis of genome-wide data, CTG is the most commonly used alternate translation initiation codon following ATG [94, 121, 64, 62]. CTG initiation is of particular interest to the

anti-viral response and cryptic T-cell epitope production. There are examples of immune epitopes that are initiated from initiation at a CTG codon [24, 205, 34, 227, 179, 193]. There is even some evidence for CTG initiation in the non-coding region of M1 in the influenza genome [225]. Furthermore, initiation at CTG codons contained on reporter constructs can be induced during viral infection and activation of the immune response [191, 159]. It is hypothesized that upregulation of CTG initiation may be a deliberate host strategy that allows generation immune epitopes in the setting of reduced translation initiation. In support of this hypothesis [191, 159, 227], the mechanism of CTG initiation may not involve eIF2α which is limited during viral responses and instead may involve the alternate initiation factor eIF2A [191]. There has yet to a systematic analysis of CTG start sites in the influenza genome, and it is unknown whether start site usage differs at the global scale during viral infection is unknown.

1.3.3 Ribosome profiling

Ribosome profiling is a technique that enables a global insight into translation dynamics by high-throughput sequencing of ribosome-protected mRNA fragments [93, 94]. The principle of ribosome profiling is based on the classic technique of ribosome footprinting [194], which involves treating *in vitro* translated mRNA with a nuclease to digest mRNAs that are not protected by ribosomes. These ribosome footprints are of a characteristic size of approximately 30 nucleotides that can be sequenced to identify the location of the ribosome with high precision. Ribosome profiling adapts and extends the concept of ribosome footprints to a high-throughput assay that captures the entire genome in an *in vivo* setting. This technique can uncover global regulation of translation during an array of biological processes, provide insight into the mechanism of translation, the cellular location of translation, and provide a comprehensive method for annotating protein coding regions [27].

To perform ribosome profiling a sample is treated with translation inhibitors to stall initiating and elongating ribosomes on transcripts [93, 94]. The samples are then sub-

jected to nuclease digestion to generate ribosome footprints. The ribosome footprints are isolated and DNA libraries are generated and deep-sequenced. The resulting reads can be mapped back to the genome to determine which transcripts are being translated and where along transcripts ribosomes are located.

In addition, this technique can be applied to identify sites of translation initiation [94, 121, 64]. By treating samples with translation initiation inhibitors, one can specifically enrich for initiating ribosomes and deplete transcripts of elongating ribosomes. Subsequent generation and sequencing of DNA libraries reveals the location of initiating ribosomes, as revealed by peaks of sequencing reads in transcripts. Statistical methods can be used to call translation initiation sites. This method combined with traditional ribosome profiling has been employed to comprehensively annotate translation initiation sites and open reading frames in many experimental systems including in viral genomes such as HCMV, and during varying conditions of cellular stress [94, 121, 64, 198].

Importantly, there are limitations and caveats to this method. Each step of the library preparation can introduce biases or artifacts. During treatment with elongation inhibitors, there can be biases towards the 5' and 3' end of transcripts [67]. This can be minimized by flash freezing samples instead of pre-treating samples with inhibitors. During treatment with translation initiation inhibitors, there can be an inflation of ribosomes at the 5' of transcripts and a skew towards identification of upstream translation initiation sites [64]. Ribosome profiling libraries can also contain reads derived from non-ribosome protected RNA. Large ribonucleoprotein complexes can co-migrate with ribosome associated fragments during size exclusion or sucrose gradient purification of ribosome protected fragments [92, 13]. This can be mitigated in part by excluding reads that are outside the size distribution of ribosome protected fragments in downstream analyses. Despite these caveats, ribosome profiling is a powerful technique, particularly in combination with orthogonal techniques such as protein biochemistry and mass spectrometry that allows unparalleled insight into translation dynamics.

1.3.4 Comprehensive profiling of translation initiation in the setting of influenza virus infection

In Chapter 3, I examine how translation initiation is altered during influenza virus infection. I employ ribosome profiling [93, 90], which can provide a genome wide view of ribosomal occupancy, to comprehensively annotate sites of translation initiation. I find a small number of novel sites of alternate initiation in the influenza genome, including a translation initiation site that generates an N-terminally truncated but functional form of NA. My findings increase our understanding of the complex protein coding capacity of the influenza genome. I also employ evolutionary analyses to examine whether there may be selection against putative alternate initiation in the influenza genome, as viral protein products such as cryptic T-cell epitopes can also be deleterious to the virus. I find evidence of selection against putative alternate initiation in the influenza lineages of mammalian hosts. Overall this work provides insight into the evolutionary forces that shape influenza evolution.

Chapter 2

POSITIVE SELECTION IN CD8+ T-CELL EPITOPES OF INFLUENZA VIRUS NUCLEOPROTEIN REVEALED BY A COMPARATIVE ANALYSIS OF HUMAN AND SWINE VIRAL LINEAGES

A version of this chapter has been previously published as:

Heather M Machkovech, Marc A Suchard, Trevor Bedford, and Jesse D Bloom. Positive Selection in CD8+ T-Cell Epitopes of Influenza Virus Nucleoprotein Revealed by a Comparative Analysis of Human and Swine Viral Lineages. Journal of Virology, 89(22): 11275-11283 (2015).

Jesse Bloom and I developed and performed the analyses and wrote the manuscript. Trevor Bedford advised on the statistical analyses. Marc Suchard developed the software that we employed.

2.1 Abstract

Numerous experimental studies have demonstrated that CD8+ T cells contribute to immunity against influenza by limiting viral replication. It is therefore surprising that rigorous statistical tests have failed to find evidence of positive selection in the epitopes targeted by CD8+ T cells. Here we use a novel computational approach to test for selection in CD8+ T-cell epitopes. We define all epitopes in the nucleoprotein (NP) and matrix protein (M1) with experimentally identified human CD8+ T-cell responses and then compare the evolution of these epitopes in parallel lineages of human and swine influenza viruses that have been diverging since roughly 1918. We find a significant enrichment of substitutions that alter human CD8+ T-cell epitopes in NP of human versus swine influenza virus, consistent with the idea that these epitopes are under positive selection. Furthermore, we show that epitope-altering substitutions in human influenza virus NP are enriched on the trunk versus the branches of the phylogenetic tree, indicating that viruses that acquire these mutations have a selective advantage. However, even in human influenza virus NP, sites in T-cell epitopes evolve more slowly than do nonepitope sites, presumably because these epitopes are under stronger inherent functional constraint. Overall, our work demonstrates that there is clear selection from CD8+ T cells in human influenza virus NP and illustrates how comparative analyses of viral lineages from different hosts can identify positive selection that is otherwise obscured by strong functional constraint.

2.1.1 Importance

There is a strong interest in correlates of anti-influenza immunity that are protective against diverse virus strains. CD8+ T cells provide such broad immunity, since they target conserved viral proteins. An important question is whether T-cell immunity is sufficiently strong to drive influenza virus evolution. Although many studies have shown that T cells limit viral replication in animal models and are associated with decreased symptoms in humans, no studies have proven with statistical significance that influenza virus evolves

under positive selection to escape T cells. Here we use comparisons of human and swine influenza viruses to rigorously demonstrate that human influenza virus evolves under pressure to fix mutations in the nucleoprotein that promote escape from T cells. We further show that viruses with these mutations have a selective advantage since they are preferentially located on the 'trunk' of the phylogenetic tree. Overall, our results show that CD8+ T cells targeting nucleoprotein play an important role in shaping influenza virus evolution.

2.2 Background

Both arms of the adaptive immune system help control influenza virus replication: antibodies neutralize virus [104] and direct the clearance of infected cells [98], while CD8+ T cells kill infected cells that display viral peptides on their major histocompatibility complex (MHC) class I molecules [208, 169]. While antibodies against the viral surface protein hemagglutinin (HA) provide the most potent protection when they are well matched to the virus strain [211, 36, 39], T cells offer broader protection against diverse strains since they tend to recognize epitopes in more conserved internal viral proteins such as nucleoprotein (NP) and matrix protein (M1) [208, 169, 228, 79].

Studies in both mice [75, 115, 111, 202] and humans [79, 189, 216] have shown that preexisting influenza virus-specific CD8+ T cells reduce the severity of disease and enhance virus clearance. For instance, preexisting virus-specific CD8+ T cells were correlated with decreased symptoms in humans infected during the 2009 H1N1 pandemic [189]. Similarly, T cells specific for NP were associated with a decreased incidence of symptomatic infection over a multiyear study of a large human cohort [79], and CD8 T-cell responses were correlated with recovery from severe H7N9 infection [216] . Therefore, experimental and epidemiological work demonstrates that CD8+ T cells contribute to immunity against influenza.

Because humans are repeatedly infected with influenza over their lifetimes, one might

expect viruses to be under evolutionary pressure to accumulate substitutions in epitopes targeted by immune memory. Indeed, there are numerous examples of the fixation of antibody escape mutations in HA [186, 26], consistent with the notion that this protein evolves under strong selection from antibodies. Several studies have also described influenza virus mutations that escape recognition by CD8+ T cells [168]. In a mouse study, viral mutations arose that conferred T-cell escape in RAG-1-deficient mice expressing an influenza virus NP-specific T-cell receptor (TCR) [160]. Rimmelzwaan and coworkers identified the fixation of mutations in NP of human H3N2 virus that mediated escape from CD8+ T cells by altering the epitope recognized by the T-cell receptor [15, 22, 23] or by abrogating binding of the epitope to MHC class I molecules [213]. Valkenburg et al. described the emergence of CD8+ T-cell escape mutations in a persistently influenza-infected infant [208]. These elegant studies demonstrate that influenza virus accumulates substitutions that escape CD8+ T cells as well as antibody-mediated immunity.

However, these studies do not prove that positive selection for CD8+ T-cell escape is an important driving force in the evolution of influenza virus, since many sites in the virus genome will fix substitutions given enough time [161, 184, 145]. To rigorously establish the presence of positive selection, the field of molecular evolution has developed statistical tests to discern whether a subset of sites is evolving faster than expected. Most of these tests compute nonsynonymous and synonymous distances (referred to as dN and dS, respectively) and then test for sites with statistical evidence that the accumulation of nonsynonymous substitutions exceeds that of synonymous substitutions (dN/dS ratio of $>$1) [106, 140]. These tests consistently find overwhelming evidence for positive selection in the antigenic sites of influenza virus hemagglutinin [32, 181, 107] but little evidence for positive selection in CD8+ T-cell epitopes [107]. One study reported that CD8+ T-cell epitopes in NP have a higher dN/dS ratio than do other sites [14]; however, that study made a pairwise comparison of two sequences only and included no tests for statistical significance. Below, we describe the use of several state-of-the-art tests to verify that CD8+ T-cell epitopes have neither an elevated frequency of sites with a dN/dS ratio of

>1 nor an elevated rate of nonsynonymous substitutions. Therefore, by standard criteria, CD8+ T-cell epitopes are not under positive selection.

The results of these statistical tests for positive selection seem at odds with the extensive body of experimental work described above. We hypothesized that the discrepancy arises because known CD8+ T-cell epitopes are under strong functional constraint [14, 7, 167, 72]. If epitopes are highly constrained, then even strong positive selection might fail to elevate the rate of nonsynonymous substitutions in epitopes above that at less constrained nonepitope sites. To address this possibility, we developed new statistical tests that take advantage of the fact that some lineages of human influenza virus are paralleled by lineages of swine influenza virus that are not under selection from human CD8+ T cells. Using these tests, we show that CD8+ T-cell epitopes in NP evolve significantly faster in human influenza virus than in swine influenza virus. Furthermore, we show that substitutions in these epitopes are enriched on the trunk of the phylogenetic tree, indicating that viruses that acquire them have a selective advantage that promotes their evolutionary spread. Overall, our work provides clear statistical evidence that complements prior experimental studies showing that CD8+ T-cell epitopes are under selection in human influenza virus [21] and suggests that the failure of conventional tests to identify this selection is due to high levels of functional constraint in epitopes.

2.3 Materials and Methods

2.3.1 Inference of phylogenetic trees and mutation counts.

M1 and NP protein-coding sequences were downloaded from the Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) [8].

For human influenza virus, we assembled sequence sets by taking sequences for H1N1 (1918 to 1957), H2N2 (1957 to 1968), and H3N2 (1968 to 2013); if there were fewer than three sequences per year, we retained them all, and when there were more than three for a year, we randomly selected three to retain. For swine influenza virus, we

similarly assembled sequence sets containing up to 3 sequences per subtype per year for H1N1 (1918 to 2013), H1N2 (1999 to 2013), and H3N2 (1998 to 2013). For swine influenza virus, the first available sequence is from 1933. We excluded sequences that were previously classified as being misannotated [110] or that were strong outliers based on molecular clock analysis using RAxML [190] and Path-O-Gen (http://tree.bio.ed.ac.uk/software/pathogen/).

There are gaps in sequence availability in earlier years (most prominently, there are no sequences from between 1918 and the early 1930s). Therefore, we have a reduced power to identify substitutions in these early years. However, since our comparisons are between human and swine influenza viruses, and since both lineages have similarly sparse sequences in these early years, these gaps seem unlikely to systematically bias our study, although they may reduce its power.

We translated the sequences and inferred separate human and swine influenza virus phylogenies for each protein using Bayesian Evolutionary Analysis by Sampling Trees (BEAST) [51] with a strict molecular clock, a Jones-Taylor-Thornton (JTT) [99] model of substitution, and a constant population size demographic model. Figure 1 shows the maximum clade credibility trees rendered with FigTree (http://tree.bio.ed.ac.uk/software/figtree/). The trunk of each tree (dark lines in Figure 1) was defined by tracing from the most recent sequence back to the oldest sequence. We used a stochastic mapping technique [137, 149, 123] implemented via the 'MarkovJumps' feature in BEAST to estimate the posterior mean number of substitutions at each site for each phylogenetic tree and along the trunk of each tree. The times to the most recent common ancestor referred to in Figure 1 legend were estimated by a BEAST analysis of the joint swine and human influenza virus lineages. The dates of fixation of the CD8+ T-cell escape substitutions characterized by Rimmelzwaan and colleagues [15, 22, 23, 213] refer to estimates obtained previously (see Figure 2 supplement 2 in [72]).

2.3.2 Identification of CD8+ T-cell epitope sites.

We downloaded all epitopes with an experimentally identified CD8+ T-cell response (source organism Influenza A virus and host Homo sapiens) from the Immune Epitope Database (http://www.iedb.org/) [212]. We identified unique epitopes, as described in Results, using a previously described software package (https://github.com/jbloom/epitopefinder/) [71] and then determined the number of unique epitopes E_r to which each site r contributes (Figure 2).

2.3.3 Conventional dN/dS tests for positive selection.

We used DataMonkey (http://www.datamonkey.org/) [43] to perform two types of dN/dS analysis, the hierarchical Bayes method FUBAR (fast unconstrained Bayesian approximation) [140] and the maximum likelihood method FEL (fixed-effects likelihood) [106]. The maximum clade credibility tree from BEAST was used as the input phylogeny for FUBAR and FEL, and a REV (general reversible model) codon substitution model was specified for FEL. For both methods, we calculated the percentage of sites for which the estimated dN/dS ratio was $>$1 and the percentage of sites for which there was strong statistical support for this ratio being $>$1 (posterior probability of $>$0.95 for FUBAR; P value of $<$0.05 for FEL).

2.3.4 Statistics for substitutions at each site in human and swine influenza virus lineages.

The posterior mean estimate of the number of nonsynonymous substitutions, S_r, at each site r, was extracted from the BEAST trees. These estimates were used to compute the average substitution rates across all epitope sites (sites that fell into at least one epitope) and across all nonepitope sites, both for the entire tree and for the trunk alone. We also defined a statistic, F, which represents the average number of epitopes changed per substitution. This statistic is defined as:

$$F = \frac{\sum_r E_r \times S_r}{\sum_r S_r}, \tag{2.1}$$

where E_r is the number of unique epitopes to which site r contributes.

We performed statistical tests of whether we could reject the null hypothesis that there was no difference in the F statistics for human versus swine and for the trunk versus the tree. To do this, we calculated the ratios of these statistics for human versus swine or trunk versus tree and then created a null distribution by repeatedly recalculating the statistics after randomizing the epitope counts, E_r, among sites. The P values represent the fraction of time that the randomized statistic is greater than the actual statistic in 10^4 randomizations.

2.4 Results

2.4.1 Parallel human and swine influenza virus lineages reveal selection by CD8+ T cells.

Our goal is to determine whether epitopes targeted by human CD8+ T cells are under selection in influenza viruses that circulate in human hosts. The two most highly expressed influenza virus proteins are NP and M1 [113], and epitopes in these proteins are major targets of CD8+ T cells [208, 169, 228, 79]. The NP and M1 proteins in contemporary human H3N2 influenza viruses have circulated in humans since at least 1918 [49, 40]. For both genes, this unbroken lineage consists of H1N1 viruses from 1918 to 1957, H2N2 viruses from 1957 to 1968, and H3N2 viruses from 1968 to the present. The red lines in Figure 1 show phylogenetic trees of NP and M1 from this human influenza virus lineage.

This human influenza virus lineage is closely paralleled by a swine influenza virus lineage descended from the common ancestor of the virus that caused concurrent pandemics in humans and swine in 1918 [49, 28]. NP and M1 of this lineage have circulated exclusively in swine since 1918 [49, 28]. The blue lines in Figure 1 show phylogenetic trees of NP and M1 from this swine influenza virus lineage. The phylogenetic trees show

that both human and swine influenza viruses undergo substantial genetic evolution in NP and M1; however, this fact alone does not reveal what forces drive this evolution. Influenza virus genetic evolution can be driven by positive selection, but it can also be driven by stochastic forces such as genetic hitchhiking or drift [161, 184, 145].

The parallel lineages of human and swine influenza viruses enable us to perform an internally controlled analysis of whether CD8+ T cells represent an important selective force in driving influenza virus evolution, since human CD8+ T cells target epitopes in human but not swine influenza viruses. There are two reasons why we can be confident that swine influenza virus is not under selection from human CD8+ T cells. First, the MHC class I molecules that restrict CD8+ T-cell epitopes are highly variable among species; therefore, epitopes displayed to human CD8+ T cells will differ from those displayed to swine CD8+ T cells [166, 127] (note that our approach does not require the human and swine epitopes to be completely nonoverlapping; it simply assumes that the MHC alleles are sufficiently diverged so that not all epitopes targeted by humans are also targeted by swine). Second, swine influenza virus is under weaker selection from immune memory than is human influenza virus because pigs are infected only once or a few times during their short lives [180, 147, 66, 218, 210]. Therefore, swine influenza virus is probably under less pressure from CD8+ T cells in general, and whatever pressure does exist will generally focus on different epitopes than those targeted by human T cells.

2.4.2 Experimentally identified human CD8+ T-cell epitopes.

We aimed to identify CD8+ T-cell epitopes targeted by individuals in the human population. There are two plausible ways to do this: by computationally predicting peptides that bind to MHC class I or by collating epitopes that have been experimentally identified as eliciting responses from CD8+ T cells isolated from humans. We chose to use experimentally identified epitopes since computational predictions are imperfect [171], and only a fraction of peptides that bind MHC class I molecules are targets of cytolytic CD8+ T cells [44, 226].

We extracted all influenza virus epitopes from the Immune Epitope Database [212] that are between 8 and 12 amino acids in length with an experimentally identified human CD8+ T-cell response. We retained all epitopes that aligned to at least one strain from our human and swine influenza virus lineages with no more than one amino acid mismatch. We classified epitopes as redundant if they shared 8 or more amino acids and were in the same MHC class I group (63) (or supertype [183] if the group was not specified). We identified 133 unique epitopes in the seven proteins that did not reassort during the human influenza pandemic of 1957 or 1968 (NS1, NS2, PB2, PA, M1, M2, and NP). Of 133 epitopes, 62 were in NP (47%), and 29 were in M1 (22%), consistent with reports that these two proteins are major targets of CD8+ T cells [79]. Figure 2 shows the number of epitopes to which each site in NP and M1 contributes; individual sites are involved in anywhere between zero and nine epitopes.

These experimentally identified epitopes probably do not represent an exhaustive list of all sites targeted by human T cells. In particular, some epitopes in historical strains may be overlooked since most studies use recent virus strains. However, since our analyses are internally controlled (we compare either human to swine influenza viruses or the trunk of the tree to side branches), missing some epitopes should not systematically bias our results.

Our approach identifies sites that contribute to epitopes in any of the influenza virus strains under consideration. Mutations of an epitope can mediate escape by abrogating peptide binding to MHC class I or by changing the sequence of the bound peptide such that it is no longer recognized by memory T cells. Both types of escape have been experimentally demonstrated for NP of human H3N2 virus. An example of escape by abrogation of MHC binding is the R384G mutation that fixed in 1993 [213]. Three examples of escape of T-cell recognition but not MHC binding are the D421E/I425V mutations that fixed in 1979 [22, 23] , the K103R mutation that fixed in 1980 [15], and the S259L mutation that fixed in 1990 [15]. Additionally, mutations outside an epitope can affect its processing [3], although we are unaware of documented examples of extraepitopic escape mutations that

have fixed in human influenza virus. Our analysis cannot distinguish among these types of escape, since most experiments identify epitopes without characterizing how prior or subsequent mutations affect their processing, MHC binding, and recognition by T cells.

We therefore classify sites according to the number of epitopes to which they contribute in any of our virus strains without attempting to determine whether the epitopes are present across all the virus strains. This approach is usually valid when mutations alter T-cell recognition without affecting processing or MHC binding, since epitopes that escape existing T cells via such mutations will often soon be targeted by new T cells [15]. However, our approach is imperfect for mutations that abrogate binding to MHC molecules and so eliminate the epitope from all subsequent strains. However, since the NP and M1 homologs are closely related (the maximal protein sequence divergence among the sequences in Figure 1 is just 12%), even if a site fixes just a single mutation that eliminates an epitope, this would represent a substantial elevation in its rate of evolution over the time frame of interest. Therefore, classification of sites by the number of epitopes as in Figure 2 should identify positions in NP and M1 that will exhibit an increased rate of substitution if there is T-cell selection.

2.4.3 Conventional dN/dS tests fail to find positive selection in CD8+ T-cell epitopes.

One might hypothesize that immune selection from CD8+ T cells would lead to a greater proportion of epitope than nonepitope sites with a dN/dS ratio of $>$1. We therefore used two state-of-the-art methods (a hierarchical Bayes approach [106] and a maximum likelihood approach [140]) to identify sites in human and swine influenza viruses with dN/dS ratios of $>$1 and partitioned the sites based on whether or not they were involved in at least one experimentally identified epitope. As shown in Figure 3A and B, the proportion of sites with a dN/dS ratio of $>$1 was actually lower for epitope than for nonepitope sites in human influenza virus. Additionally, we calculated the proportion of sites with strong statistical evidence of a dN/dS ratio of $>$1 (Figure 3C and D). In no instance is there a

greater proportion of epitope than nonepitope sites with a dN/dS ratio of $>$1. Our findings are consistent with data from previous work that failed to find evidence for positively selected sites in NP or M1 [107]. Thus, overall, state-of-the-art dN/dS tests fail to identify enhanced positive selection in T-cell epitopes in NP.

2.4.4 Epitope sites do not evolve faster than nonepitope sites, although the rate is higher on the trunk of the human influenza virus NP lineage.

We next estimated the number of nonsynonymous substitutions at each site in NP and M1 for both the full swine and influenza virus trees in Figure 1 and for the 'trunks' of these trees (dark lines in Figure 1). The rationale for examining the trunk separately is that we expect beneficial substitutions to be enriched on the trunk since they will confer a selective advantage that favors the propagation of sequences that contain them. Consistent with this idea, studies of human H3N2 influenza virus HA have found that presumably beneficial substitutions that alter antigenicity are enriched on the trunk versus the entire tree [12, 11].

Figure 4 shows the ratios of the substitution rates at epitope versus nonepitope sites. In none of the cases (NP or M1, swine or human influenza virus, or trunk or tree) is the ratio substantially greater than 1 when taken over the whole tree, so in no case are the epitope sites evolving faster than the nonepitope ones. This result helps explain why the dN/dS tests fail to find evidence for positive selection in the epitopes: for NP and M1 (unlike for HA), epitope sites simply do not evolve faster than their nonepitope counterparts.

However, further examination of the data in Figure 4 reveals an interesting trend. Although there are no instances where epitope sites are evolving substantially faster than nonepitope ones, the ratio of epitope to nonepitope substitution rates is higher for NP of human influenza virus than for NP of swine influenza virus. This trend is particularly pronounced along the trunk of the tree, which is exactly where we would expect to see the largest increase in rates if epitope substitutions confer a selective advantage. To test if this

trend was indicative of statistically significant CD8+ T-cell selection in NP, we undertook a more nuanced analysis, as described below.

2.4.5 Substitutions alter more NP epitopes in human than in swine influenza virus, especially on the trunk.

The above-described analyses simply subdivided sites as epitope or nonepitope sites based on whether they fall into at least one experimentally identified epitope. However, as shown in Figure 2, some sites are involved in far more unique epitopes than others. To account for this, we defined a new statistic (which we denote F) that gives the average number of unique epitopes that are altered per substitution. Figure 5A shows this F statistic for the trunk and entire tree for NP and M1 of both human and swine influenza viruses. There appears to be little difference in this F statistic for M1 between the trunk and the entire tree and between human and swine influenza viruses. However, for NP, F is greater for human influenza virus than for swine influenza virus, with the increase being much larger for the trunk than for the entire tree. This result indicates that the average substitution alters more NP epitopes in human influenza virus than in swine influenza virus and that this trend is especially pronounced on the trunk of the tree, exactly as we would expect under selection for epitope-altering substitutions.

To test if the trend is statistically significant, we computed the ratio of F for human influenza virus to that for swine influenza virus and then calculated a null distribution by repeatedly randomizing the epitopes among sites. For M1, the actual ratio falls near the center of the null distribution (Figure 5B), confirming that there is no enhancement of the rate of fixation of epitope-altering substitutions in M1 of human influenza virus. However, for NP, the actual ratio falls near the top the null distribution for both the trunk and the tree (Figure 5B). In particular, for the trunk, there is strong statistical support (P = 0.0002) for rejecting the null hypothesis that epitope-altering substitutions are equally likely to fix in human and swine influenza viruses. This result indicates that there is selection for

substitutions that alter CD8+ T-cell epitopes in NP of human influenza virus.

To corroborate this statistical finding of an enhanced rate of epitope-altering substitutions along the trunk of human versus swine influenza virus, we undertook a further subanalysis of the small subset of epitopes for which specific T-cell escape mutations have been identified. The vast majority of epitopes that comprise the analysis shown in Fig. 5 were identified by relatively high-throughput studies that characterized peptides eliciting T-cell responses without identifying escape mutations. However, a series of meticulous studies by Rimmelzwaan and coworkers [15, 22, 213] identified sites of specific escape mutations for a small number of epitopes. Table 1 shows that substitutions are fixed at all of these sites along the trunk of the human influenza virus tree but that swine influenza virus has no fixed substitutions at any of these sites. This finding lends further support to the idea that human T cells exert positive selection on human but not swine influenza virus.

We next tested whether selection for epitope-altering substitutions in human influenza virus NP was stronger on the trunk than on the rest of the tree, as would be expected if such substitutions confer a selective advantage. We computed the ratio of the F statistic for the trunk to that for the entire tree and generated null distributions for this statistic by randomizing the epitopes among sites (Figure 6). For M1, the actual ratios are near the center of the null distributions. However, for NP, the actual ratio is near the top of the null distribution (P = 0.003) for human influenza virus and near the bottom for swine influenza virus (P = 0.01). This result demonstrates that epitope-altering substitutions in NP are significantly enriched on the trunk of the human influenza virus lineage, suggesting that viruses that fix these mutations are more fit than other strains. The depletion of epitope-altering substitutions in the swine lineage can also be given a clear explanation: if the known CD8+ T-cell epitopes in NP are in functionally constrained regions of the protein (as has been suggested by a variety of experimental studies [14, 7, 167, 72], then substitutions in these epitopes will often be deleterious in swine influenza virus lineages that experience no human CD8+ T-cell selection and so will be relatively depleted on the

trunk. Thus, overall, the results described above provide strong statistical evidence of positive selection in the CD8+ T-cell epitopes of human influenza virus NP.

2.5 Discussion

We describe the first rigorous statistical evidence that CD8+ T-cell epitopes are under positive selection in human influenza virus. Our work adds to a growing body of evidence suggesting an important role for T-cell immunity in shaping influenza virus evolution. Previous studies showed that T cells help protect against human influenza virus [79, 189] and detailed specific instances of T-cell escape [15, 22, 23, 213, 208]. Our work shows that T-cell selection increases the rate at which mutations are fixed in epitopes of NP and indicates that viruses with these substitutions have a selective advantage that makes them more likely to fall along the trunk of the phylogenetic tree.

Our results also explain why conventional dN/dS tests fail to detect positive selection in CD8+ T-cell epitopes. Known human CD8+ T-cell epitopes tend to be under strong functional constraint [14, 7, 167, 72]. It is unclear whether this is because T cells inherently target conserved epitopes, because repeated infections preferentially boost T cells that target conserved epitopes, or because there is a bias toward experimentally identifying conserved epitopes. However, in any case, the fact that known epitopes are under strong constraint means that even fairly strong positive selection may not enhance the nonsynonymous substitution rate to a level detectable by conventional dN/dS tests. This contrasts with antibody epitopes in HA, where the ability of dN/dS tests to detect antibody-mediated positive selection is probably augmented by the fact that antigenic sites are disproportionately tolerant of point mutations [203].

The novel approach that we developed ameliorates this problem by comparing the evolution of epitopes of human and swine influenza viruses or the entire phylogenetic tree and only its trunk. These comparisons should better control for site-to-site variation in functional constraint, since comparisons are always made between homologous sites

that should be subject to similar functional constraints. Admittedly, there may also be other differences in functional constraints between human and swine influenza viruses beyond T cells, but unless these differential constraints are systematically biased toward occurring at T-cell epitopes, they should not alter the fundamental validity of our approach. By making comparisons in this way, we demonstrated clear positive selection in CD8+ T-cell epitopes in NP, both in human versus swine influenza viruses and in the trunk versus the entire phylogenetic tree.

One interesting aspect of our study is that we found positive selection in NP but not M1. This finding is consistent with a recent large-scale study that found that NP was the only protein for which the presence of preexisting memory T cells correlated with decreased rates of symptomatic infections [79]. However, our study does not preclude an important role of T cells targeting M1, which contains an immunodominant HLA-A2 epitope spanning residues 58 to 66 [21, 100]. One study argued that T cells targeting this epitope are ineffectual [100], although this interpretation is disputed [209, 101]. However, experiments have also shown that this epitope is under strong constraint [14]. If an epitope is completely intolerant of mutations, it will of course be unable to accumulate substitutions regardless of the strength of selection. It remains unclear if our failure to detect positive selection in M1 reflects a lack of effective immunity targeting this protein or strong constraints that simply prevent the fixation of escape mutations.

It is well established that antibodies are strong drivers of repeated selective sweeps in the evolution of human influenza virus [12, 163]. The fact that we can detect positive selection by CD8+ T cells even in the presence of these antibody-driven selective sweeps demonstrates the importance of T-cell immunity in driving viral evolution. The existence of such selection is consistent with modeling studies showing that T-cell immunity that reduces the infectious period can strongly favor viral escape [70].

There is considerable interest in developing vaccines that elicit stronger T-cell immunity to better protect against diverse influenza virus strains [208]. Our demonstration of the evolutionary importance of T-cell selection suggests that this interest is well founded.

In addition, our results suggest that attempts to forecast the seasonal evolution of influenza [130, 144] could benefit from examining changes in T-cell as well as antibody epitopes.

2.6 Acknowledgements

The research reported here was supported by the NIGMS of the National Institutes of Health under grants R01GM102198 (to J.D.B.) and MIDAS U01GM110721 (to T.B. and J.D.B). H.M.M. was supported in part by training grant T32AI083203 from the NIAID of the National Institutes of Health.

Chapter 3

COMPREHENSIVE PROFILING OF TRANSLATION INITIATION IN INFLUENZA-VIRUS INFECTED CELLS

Heather M. Machkovech, Jesse D. Bloom, and Arvind R. Subramaniam.

Jesse Bloom, Rasi Subramaniam and I devised experiments, developed and performed the analyses and wrote the manuscript. I performed the experiments.

3.1 Abstract

Translation can initiate at alternate, non-canonical start codons in response to stressful stimuli in mammalian cells. Recent studies suggest that viral infection and anti-viral responses alter sites of translation initiation, and in some cases, lead to production of novel immune epitopes. Here we systematically investigate the extent and impact of alternate translation initiation in cells infected with influenza virus. We perform evolutionary analyses that suggest selection against non-canonical initiation at CUG codons in influenza virus lineages that have adapted to mammalian hosts. We then use ribosome profiling with the initiation inhibitor lactidomycin to experimentally delineate translation initiation sites in a human lung epithelial cell line infected with influenza virus. We identify several candidate sites of alternate initiation in influenza mRNAs, all of which occur at AUG codons that are downstream of canonical initiation codons. One of these candidate downstream start sites truncates 14 amino acids from the N-terminus of the N1 neuraminidase protein, resulting in loss of its cytoplasmic tail and a portion of the transmembrane domain. This truncated neuraminidase protein is expressed on the cell surface during influenza virus infection, is enzymatically active, and is conserved in most N1 viral lineages. Host transcripts induced by the anti-viral response are enriched for translation initiation at non-canonical start sites and non-AUG start codons. Together, our results systematically map the landscape of translation initiation during influenza virus infection, and shed light on the evolutionary forces shaping this landscape.

3.2 Author summary

When viruses such as influenza infect cells, both host and viral mRNAs are translated into proteins. Here we investigate the sites in these mRNAs that initiate protein translation during influenza infection. In particular, we explore whether some of this translation initiates at codons other than the canonical ones used to produce the primary protein product of each gene. Using computational analyses, we find that mammalian influenza

viruses evolve to reduce the number of codons that can initiate such alternate translation initiation products. We next use the comprehensive experimental strategy of ribosome profiling to identify sites of translation initiation across all influenza and host mRNAs. We find a number of sites of alternate initiation on both influenza and host mRNAs. We study in detail one such alternate start site on an influenza mRNA, and show that it encodes a functional and previously uncharacterized variant of a viral protein. We also find evidence that the mRNAs that host cells express at higher levels during viral infection are enriched for translation initiation at non-canonical start codons. Overall, these results suggest that alternate translation initiation plays a role in shaping the repertoires of both viral and host proteins that are produced during influenza infection.

3.3 Introduction

Cellular stress can impact protein translation initiation. For example, viral infection can activate protein kinase R, which globally reduces translation initiation for most transcripts, while translationally upregulating a subset of stress-response transcripts [65]. Cellular stress can also alter the start codon on mRNAs at which ribosomes initiate translation [232, 198, 178]. A common alternate start codon, CUG, may be of particular relevance during immune stress. Previous work using sensitive reporter assays found that CUG initiation can be induced during viral infection and activation of the immune response [191, 159]. It has been hypothesized that upregulation of CUG initiation may be a deliberate host strategy to generate immune epitopes during cellular stress [191, 159, 229]. Indeed, there are numerous examples of immune epitopes that are initiated from alternate start sites, including CUG codons [24, 205, 34, 227, 179, 193]. However, it is unknown if start site usage is altered globally during viral infection, or skewed towards certain start codons such as CUG.

In addition, viruses often rely on alternate translation initiation to increase the diversity of proteins encoded in their compact genomes. Influenza is an eight segmented negative

stranded RNA virus that encodes ten canonical proteins, but additional influenza proteins can be generated through alternate initiation. For instance, the PB1-F2 protein is initiated at a downstream AUG in the +1 reading frame of the PB1 segment, and is thought to modulate the host immune response [35, 230, 135]. Additional peptides generated by alternate downstream AUG codons have been noted in the PB1, PA, and M influenza segments [35, 230, 220, 1, 2, 139, 221], but their abundance and functional relevance is unclear. There is some experimental support for CUG initiation in the non-coding region of M1 in the influenza genome [225], and very low levels of initiation on the negative sense RNA of the NS segment [234, 37, 80]. Even when such noncanonical peptides are generated at low levels and not functionally important for the virus, they can still be targeted by the host immune response [227]. However, despite the ad hoc identification of the aforementioned alternate protein products, there have been no systematic studies to map sites of translation initiation in all influenza transcripts.

Ribosome profiling, the high-throughput sequencing of ribosome-protected mRNA fragments, enables genome-wide interrogation of translation [93]. This method has been applied to influenza virus to reveal host shut-off [13] and translation of non-coding RNAs [164]. It has also been used to identify unannotated proteins in other viruses such as CMV [195]. Notably, ribosome profiling, in combination with inhibitors of translation initiation, can be used to identify candidate sites of translation initiation [94, 121].

Here we integrate computational and experimental approaches to systematically examine translation initiation in influenza-infected cells. We first perform computational analyses that suggest there is selection against putative CUG alternate translation initiation sites in mammalian influenza virus lineages. We then use ribosome profiling to experimentally map sites of translation initiation in viral and cellular mRNAs during influenza virus infection. Surprisingly, we find little signal for CUG initiation in viral genes, but we do identify a number of candidate AUG start sites that are downstream of the canonical start codon. We show that one of these candidate start sites leads to production of a truncated version of the viral neuraminidase protein that is expressed on the cell surface

and is enzymatically active. On the host side, we identify an excess of non-AUG start codons on transcripts that are upregulated during the anti-viral response. Together, our results highlight how alternate translation initiation can alter the protein-coding capacity of both viruses and the cells that they infect.

3.4 Results

3.4.1 Evolutionary analyses suggest selection against CTG start sites during influenza adaptation to mammalian hosts

We first examined whether we could identify any bioinformatic signature of evolutionary selection against alternate translation initiation. Such selection would occur if alternate translation initiation products are deleterious to the virus. One reason that such products might be deleterious is because they can serve as CD8 T-cell epitopes [24, 205, 34, 227, 179, 193]. Human influenza virus evolves under pressure to escape CD8 T-cell immunity [131, 160, 213], so we hypothesized that the virus might minimize the number of alternate translation initiation sites as it adapts to humans.

We focused our bioinformatic analysis on alternate translation initiation at CUG codons for multiple reasons. First, several reporter studies have shown that CUG can serve as an alternate translation initiation codon [155, 46, 191, 192, 225]. Second, genome-wide ribosome profiling studies suggest that CUG is the most common non-AUG start codon [94, 121, 62]. Third, CUG-initiated peptides can serve as CD8 T-cell epitopes [48, 177, 192, 219, 191]. Finally, recent evidence suggests that immune responses and viral infection cause specific upregulation of CUG initiation [191, 159].

One way to identify signatures of selection during influenza virus evolution is to compare changes in the genomes of viral lineages that have adapted to different host species [73, 197]. Such host-specific adaptation occurs frequently when the virus jumps from its natural reservoir—wild waterfowl—to other host species including humans and swine [200]. We speculated that these host jumps could lead to selection against CD8 T-cell epitopes.

Because humans are long lived and accumulate more adaptive immune memory, we further expected that immune selection on influenza will be generally stronger in humans than in birds and swine [147, 66, 218, 210, 103, 10]

For our analysis, we examined five sets of influenza virus sequences. The first set includes human viruses descended from the 1918 pandemic strain, which is believed to have originated from an avian virus that jumped to humans [201, 162, 187]. Five of the eight segments in contemporary human H3N2 influenza virus (PB2, PA, NP, M, and NS) are descended in an unbroken lineage from the 1918 virus, and so have been adapting to humans for a century. We analyzed the protein-coding genes on these five segments. The second set includes the same genes from the classical lineage of swine influenza, which is thought to have also arisen from an avian viral strain around 1918 [201, 162, 187]. Our third and fourth sets comprise these genes from viruses that were isolated from ducks and chickens, respectively. Since birds represent the virus's natural reservoir [153], we expect these viruses to be relatively stably adapted to their hosts. Our fifth set is sequences from human H5N1 infections, which are zoonotic infections by avian viruses that have had minimal time to adapt to humans [223].

To identify bioinformatic signatures of selection against alternate translation initiation, we began by examining whether there was depletion of CUG codons in reading frame 0 (in-frame with the canonical AUG) in human influenza. We found that the number of CUG codons in reading frame 0 of the human influenza virus lineage decreased over time (Fig. 7A, upper panel), with the recent human H3N2 viruses having about 40% fewer in-frame CUG codons than the 1918 virus. We also saw a similar trend in the swine lineage, suggesting that the loss of CUG codons results from a mammalian-specific rather than a human-specific evolutionary pressure. There is no notable decrease in the number of CUG codons in the avian viral sequences, which consistently have an in-frame CUG content similar to that of the 1918 virus. The in-frame CUG content of sequences from human H5N1 viruses (which mostly result from one-off infections with avian viruses) is similar to that of the avian and 1918 sequences. Notably, we did not see a comparable decrease in

frequency of CUH (H = A/C/U) leucine codons in any of the five lineages (Fig. 7A, lower panel), indicating that the depletion of in-frame CUG codons is not attributable to a general selection against CUN leucine codons. This depletion is also not due to selection against the amino acid leucine or biochemically similar amino acids (Fig. 13). The depletion against CUG codons is still observed if we correct for each sequence's nucleotide composition (Fig. 14).

Since the sequence context of a start codon plays an important role in determining how efficiently it can mediate translation initiation, we examined whether there is especially strong selection against CUG codons in sequence contexts that favor translation initiation. For AUG start codons, a purine at position -3 and a G at +4 can increase efficiency of initiation [109]. A recent reporter study has shown that CUG-mediated translation initiation is also promoted by a G at +4 [46]. Consistent with this study, we found that CUG initiation sites identified in mammalian transcripts using ribosome profiling are enriched for G at the +4 position (Fig. 15). Given the importance of the G at the +4 position, we examined whether there were differences in the depletion of in-frame CUGG and CUGH motifs in influenza virus. Indeed, we found that there is a more pronounced decrease in CUGG motifs than CUGH motifs in both human and swine viral lineages (Fig. 7B).

We next examined whether there was also selection against CUG codons that occur out-of-frame with the canonical AUG codon, as initiation in reading frames 1 and 2 would also generate novel peptides. We did not find evidence for selection against CUG codons in reading frames 1 and 2 (Fig. 14). However, in reading frames 1 and 2, eliminating the start codon is not the only way to eliminate long peptides initiated by alternate translation—these peptides can also be eliminated by introducing an early stop codon that shortens the peptide initiated by the alternate start codon. We therefore examined the length of putative CUG-initiated ORFs in reading frames 1 and 2. The human and swine viral lineages are depleted of long putative CUG-initiated ORFs (Fig. 7C, upper panel). Again, this selection is specific for CUG, since there is no host-specific depletion in long putative CUH-initiated ORFs (Fig. 7C, lower panel).

Since the viral sequences used in our analyses are phylogenetically related, we cannot easily test for the statistical significance of these evolutionary signals [55]. Nevertheless, the above analyses reveal bioinformatic signals consistent with selection against CUG-initiated alternate translation products in influenza virus lineages that have evolved in humans and swine.

3.4.2 Systematic experimental profiling of translation initiation during influenza virus infection

Having established that human-adapted lineages of influenza virus exhibit signatures consistent with selection against alternate initiation at CUG codons, we sought to experimentally delineate the codons in influenza transcripts that could serve as potential sites of translation initiation during infection of a human lung epithelial cell line (A549). To this end, we employed ribosome profiling [93], a deep sequencing method for genome-wide quantification of ribosome-protected mRNA fragments. We performed two different variations of the ribosome profiling assay (Fig. 8A). We used the standard ribosome profiling method (Ribo-seq) to capture both elongating and initiating ribosomes. We did not add the elongation inhibitor cycloheximide to cells, since there can be artifacts associated with cycloheximide treatment [68]. We also performed a variation of ribosome profiling that enriches for initiating ribosomes [121] by pre-treating cells for 30 minutes with lactidomycin, a drug that preferentially binds empty E-sites of initiating ribosomes [176] thus allowing elongating ribosomes to run off mRNAs. Finally, we performed RNA-seq to profile the genome-wide transcriptional changes that occur during influenza infection.

Anti-viral pathways have been shown to induce alternate translation initiation at CUG codons in reporter assays [159]. Therefore we also examined whether antiviral responses induced by interferon-β pretreatment alters the landscape of translation initiation in influenza virus. Altogether, we performed Ribo-seq and RNA-seq assays on four different samples of A549 cells (Fig. 8A): untreated cells (ctrl), interferon-β treated cells (+ifn), in-

fluenza virus infected cells (+vir), and cells that were pre-treated with interferon-β and subsequently infected with influenza virus (+ifn +vir). We utilized a reassortment virus with the polymerase components from A/PuertoRico/1934 and the remaining viral segments from A/WSN/1933. We used a high multiplicity of infection (MOI) such that most cells were productively infected by virus. We harvested cells 5 hours after viral infection, at which time the virus is expected to be amply transcribing and translating its mRNAs, but will not yet have undergone secondary rounds of infection [78, 174].

To detect potentially low levels of alternate translation initiation at CUG codons in influenza transcripts, we used a mix of two viruses that were synonymously recoded in their NP gene to either deplete the number of in-frame CUG codons (low CUG NP), or to increase the number of such codons (high CUG NP) (Fig. 16). Our rationale for recoding CUG codons in this gene is that NP is a major source of CD8 T-cell epitopes [228, 79], and previous work suggested that CUG codons may initiate peptides that contribute to the immune epitope pool [191, 159]. Using this virus mix allowed us to have an internally controlled experiment: if a CUG codon in the high CUG NP variant can initiate translation, we should be able to preferentially detect it against the low CUG NP variant background even in the presence of end-specific sequence biases in the ribosome profiling method [5].

Deep sequencing of ribosome protected fragments yielded 15–40 million reads, of which 2–3 million reads could be mapped to human or influenza transcripts after standard preprocessing steps (Fig. 17A). As observed previously [13], 15–25% of influenza-aligned reads mapped to the negative sense influenza genomic RNAs (Fig. 17A). In contrast to the reads mapping to the positive sense strand of influenza (Fig. 8C), the negative sense reads did not show a characteristic peak around the size of a ribosome footprint (Fig. 17D), and likely arise from co-purification of influenza genomic RNAs that are protected from nuclease digestion due to their association with the viral NP protein [52, 120]. Therefore, we only considered reads that mapped to the positive sense strand of influenza transcripts for the remaining analyses. Reads mapping to influenza transcripts accounted for between 29–39% of the mapped reads in the virus-infected samples (Fig. 8B), consistent with our

use of high MOI. As expected, interferon-β pretreatment reduced productive infection, and reads mapping to influenza transcripts only accounted for 6-13% of aligned reads in the +ifn +vir sample (Fig. 8B).

We performed additional checks to ensure adequate quality of our Ribo-seq data. Ribo-seq reads mapping to either human and influenza transcripts had read lengths between 29–34 nucleotides (Fig. 8C, Fig. 17B). This length distribution was on average a few nucleotides wider and longer than observed in other ribosome profiling studies [94, 91, 126], likely because of less stringent RNase I digestion. We assigned the ribosome P-site to either the 14th or 15th nucleotide from the 5′ end of footprint reads depending on their length (see Methods). Consistent with previous observations [121], mean P-site density in the Ribo-seq + LTM samples showed $>$100-fold enrichment at the annotated start codons of both human and influenza transcripts (Fig. 8D, Fig. 17C). P-site density in the Ribo-seq samples exhibited a peak at start and stop codons and was distributed within the coding region of human and influenza transcripts (Fig. 8E, Fig. 17C).

The mean P-site density in Ribo-seq samples displayed the expected 3-nucleotide periodicity along human transcripts (Fig. 17E), and was enriched in frame 0 of both human and influenza coding sequences (Fig. 8F, 17F). Notably, the known 191 nucleotide dual-coding region in the NS segment of the influenza genome [116] had a distinct P-site density distribution across the three reading frames compared to the single-coding region of the same segment (Fig. 8G, 17G). This dual-coding signature was present but less pronounced in the known dual-coding regions of M (45 nucleotides) and PB1 (261 nucleotides) segments (Fig. 8G, 17G), likely because M2 and PB1-F2 are expressed at lower levels during infection [112, 88, 207].

Finally, we used our sequencing data to evaluate the quality of the virus stocks used in our experiment. Virus stocks contain virions that range in biological activity, including virions defective in replication (defective particles) [214, 85, 29, 97, 42]. For influenza virus, defective particles often contain large internal deletions in the polymerase segments [214, 85, 29, 175, 97, 42]. The presence of defective particles with internal dele-

tions would diminish our ability to detect alternate initiation sites within the deleted regions. However, the burden of defective viral particles can be reduced by growing virus for a short amount of time and at a low MOI, as defective viral particles increase in frequency when viruses are grown at a high MOI where complementation of deleterious genotypes can occur [224]. The even read coverage across the polymerase segments in our data (Fig. 18B) indicates that we had a low burden of defective particles. This low burden was particularly evident when we compared our data to a previous ribosome profiling study of influenza virus infected cells (Fig. 18A) [13].

3.4.3 Translation initiation sites in the viral genome

We next used the Ribo-seq and Ribo-seq + LTM measurements to annotate candidate translation initiation sites (TIS) in the influenza genome. Our general strategy to identify candidate TIS was to find peaks in Ribo-seq + LTM coverage within each influenza transcript that was significantly higher than both the Ribo-seq coverage at the corresponding location and the background Ribo-seq + LTM coverage distribution for that transcript (Fig. 9A). Specifically, we used a zero-truncated negative binomial distribution (ZTNB) to statistically model the background distribution of Ribo-seq and Ribo-seq + LTM counts in each transcript [206, 64]. Candidate start sites were identified based on the following criteria: The ZTNB-based P-value for the Ribo-seq + LTM count at that location must be $<$0.01 and 1000-fold higher than the P-value of the Ribo-seq counts at the same location (Fig. 9A, left panel), or must have an absolute value less than 10^{-7}. In addition, the Ribo-seq + LTM counts must be a local maximum within a 30 nucleotide window (Fig. 9A, right panel), and the Ribo-seq counts must be non-zero. To account for the 1–2 nucleotide positional uncertainty of the P-site density in our Ribo-seq and Ribo-seq + LTM measurements, we pooled the read counts in 3 nucleotide windows before applying the above criteria.

We applied our candidate TIS identification strategy to each influenza transcript (Fig. 19)

separately for the +vir and +ifn +vir samples. We assigned the candidate TIS peaks to any near-cognate AUG codon (at most one mismatch from AUG) if that codon was within 1 nucleotide of either side of the TIS peak. We identified a total of 25 candidate TIS across both samples (Fig. 20). Fourteen of the 25 identified TIS overlapped between the two samples (Fig. 9B), and we used this overlapping subset as a high-confidence set of candidate TIS for downstream analyses. We did not detect a higher number of candidate TIS in our +ifn +vir sample compared to the +vir sample (4 vs. 7, Fig. 9B), suggesting that anti-viral response as mediated by interferon-β induction does not result in detectably higher number of alternate translation initiation sites in influenza transcripts [159]. An important caveat to this observation is that the Ribo-seq + LTM assay is likely to miss TIS that have a low frequency of initiation, but could still be detectable by sensitive immunological assays [191, 80].

The high-confidence set of 14 candidate TIS had multiple features consistent with being bona fide TIS. First, this set included 7 of the 8 annotated TIS (aTIS) for the 8 segments in the influenza genome (Fig. 9E). Only the annotated TIS of the PB2 segment was not identified, and this is due to the dense Ribo-seq + LTM coverage and lack of clear peaks in this segment (Fig. 9D, bottom right panel). Second, our set also correctly identified the start codon for two previously annotated protein products generated by initiation in an alternate reading frame of the PB1 and M genes. Initiation in the +1 reading frame at nucleotide 118 in the PB1 gene generates PB1-F2 (Fig. 9D, arrow). Initiation in the +1 reading frame at nucleotide 113 in the M segment is also a known to occur [221]. Third, 13 of the 14 candidate TIS had an AUG codon within 1 nucleotide even though less then 3% of trinucleotides in the influenza genome are AUG (Fig. 9F, top vs. bottom). Fourth, both the annotated and the novel candidate TIS in our set had an over-representation of A at –3 nucleotide and G at +4 nucleotide positions (Fig. 9C), which are known to be optimal contexts for translation initiation in vertebrate cells [109]. Finally, the candidate TIS are enriched towards the 5′ end of the transcripts, with 13 of the 14 candidate TIS located in the initial third of the influenza transcripts (Fig. 9G). This last observation is

consistent with the initiation of candidate TIS through scanning by the 43S pre-initiation complex from the 5′ end [108].

Six of the 7 alternate TIS in our candidate set are AUG codons that are downstream of the respective canonical aTIS (Fig. 9E). This downstream location of alternate TIS is expected given the short 5′ UTR (around 20 nucleotides) of influenza transcripts. While previous translation initiation profiling studies have identified a large number of putative non-AUG TIS [94, 121], these are predominantly located in the much longer 5′ UTRs of mammalian transcripts, upstream of aTIS. Even for mammalian transcripts, a majority of TIS identified downstream of annotated start codons are at AUG codons [215].

The alternate dTIS in our high-confidence set are distributed across multiple influenza segments: M, NA, NP, PB1, and PA (Fig. 9D, Fig 21). Ten of the 13 high-confidence TIS initiating at AUG are located in the canonical reading frame 0 (Fig. 9H). This set includes 7 annotated AUG starts, as well as three in-frame downstream AUGs in NA, NP, and PA segments that would result in N-terminally truncated forms of the annotated proteins. The three out-of-frame candidate dTIS are the known PB1-F2 ORF [35], the known start codon at nucleotide 113 of the M gene [221], and a short ORF of length 2 in NA (Fig. 9I, top panel). Excluding the well-characterized PB1-F2 ORF, the two candidate out-of-frame candidate ORFs have lengths that are typical of out-of-frame AUG-initiated sequences in the influenza genome (Fig. 9I, lower panel).

Among the other alternate translation initiation sites in the influenza genome that have been described previously, we did not find evidence for the initiation sites previously noted in PA and PB1 [35, 230, 220, 1, 2, 139]. This could be because of low initiation frequency at these sites under our infection conditions. Re-analysis of the harringtonine-treated ribosome profiling data from Ref. [13] revealed P-site count peaks that coincided with all the 7 annotated TIS as well as 3 of the 7 downstream TIS identified in our experiments (Fig. 22). Among the 4 dTIS that could not be clearly associated with a harringtonine peak, 3 are $>$200 nucleotide from the 5′ end of the transcripts, and could have been potentially affected by the high fraction of defective viral particles in [13] (Fig. 18A). Interestingly, the

dTIS in our dataset without a corresponding harringtonine peak is an in-frame AUG in the NA segment that is mutated to the near-cognate CUG codon in the PR8 influenza strain used in [13].

3.4.4 Translation initiation at CTG codons in influenza NP

As mentioned previously, our infections were performed with a mix of two otherwise isogenic viruses that were recoded to have different numbers of CUG codons in the NP gene (Fig. 16). This experimental design allows us to sensitively look for CUG initiation that cannot be detected by the start-site calling method in the previous section, but might still be detectable via an internally controlled comparison of the two NP variants.

Specifically, the high CUG NP variant has 20 leucine codons that were synonymously mutated to a non-CUG codon in the low CUG NP variant (Fig. 10A). We examined if there was evidence of enhanced initiation at any of these codon sites in the high CUG NP variant. Between 35% and 42% of reads mapping to NP could be uniquely assigned to either the high or low CUG NP sequences based on polymorphisms due to the synonymous mutations. These reads mapped to the two variants at a nearly equal (1:1.3 ratio) overall proportion (Fig. 10A). The ratio of P-site density between the two variants at individual NP sites obtained from the uniquely-mapping reads varied over a 500-fold range (Fig. 10B). Among the sites with an excess of high CUG NP P-site density, the CUG codon 322 nucleotides from the annotated TIS displayed the largest such excess with 16-fold more reads for the high CUG NP variant than the low CUG NP variant (Fig. 10B, red point). This excess P-site density was present in both the Ribo-seq and the Ribo-seq + LTM data (Fig. 10C). The length distribution of the reads generating this excess was consistent with them being derived from ribosome-protected fragments (Fig. 23). The excess P-site density at CUG322 did not arise from 3′ ends of reads mapping to the nearby recoded CUG328 (Fig. 24). The +ifn +vir sample displayed a similar excess of density at site 322 for the high CUG NP variant (Fig. 25).

To examine whether CUG322 could initiate translation in the high CUG NP variant, we used Western blots of heterologously expressed NP variants. However, we were unable to resolve any protein fragment of appropriate size that was present only in the high CUG NP variant (Fig. 26). It is possible that there is no initiation at this site, or that initiation occurs but at a very low level, consistent with the low overall Ribo-seq + LTM P-site density at CUG322 compared to the other locations on NP. In that case, more sensitive immunological assays [191, 80] might be necessary to identify initiation at CUG322 and other CUG codons in the high CUG NP variant. Another important caveat is that LTM treatment may not effectively arrest CUG-initiating ribosomes. Indeed, evidence for CUG-based initiation being refractory to several translation inhibitors has been noted in previous reporter-based studies [191].

3.4.5 Functional characterization of an in-frame alternate start site in NA

Among the three in-frame alternate TIS identified by our Ribo-seq experiments (Fig. 9D), the AUG at nucleotide 43 from the aTIS of the neuraminidase (NA) gene had the highest statistical significance in our start site calling method (Fig. 20). This candidate TIS (Fig. 11A) had several other features that prompted us to investigate it in further detail.

The AUG43 in NA has a favorable translation initiation context, with a G at +4 nucleotides (Fig. 11B). The function of NA is to mediate viral egress by cleaving the viral receptor sialic acid from the cell surface [152, 129]. NA is a type II membrane protein [25, 143], meaning that the cytoplasmic tail and the transmembrane domain are at the N-terminus and the ectodomain is at the C-terminus (Fig. 11B). The truncated NA protein resulting from translation initiation at AUG43 would lack the cytoplasmic tail and the first few amino acids of transmembrane domain. Interestingly, classical studies of type II membrane proteins characterized a series of artificial mutants of NA in an effort to determine the motifs required for membrane insertion of the protein [31, 83]. One of these NA mutants was an N-terminal deletion in which the first 42 nucleotides were removed,

effectively creating the NA43 protein that would be generated by translation initiation at AUG43. In the context of a protein expression plasmid, this NA N-terminal deletion was efficiently expressed and localized to the cell surface, indicating that the signal and anchor domains are reasonably intact [31, 83]. On the basis of this prior work, we hypothesized that any NA43 protein generated by alternate translation initiation at AUG43 of the wild-type NA gene would also create a cell-surface NA protein lacking the cytoplasmic tail and a portion of the transmembrane domain.

We analyzed the sequences of NA genes across different influenza lineages to examine if the AUG43 codon is evolutionarily conserved. Most N1 avian influenza and human pdmH1N1 NAs have an AUG at nucleotide 43, as do the majority of N1 swine influenza NAs (Fig. 11C). Some human seasonal H1N1 NAs and N1 swine influenza NAs lack an AUG at nucleotide 43, but these strains usually have another near-cognate start codon at the site instead (Fig. 11C). Therefore, the candidate downstream start site at NA nucleotide 43 is present in most N1 influenza strains.

We next sought direct experimental evidence for the NA43 protein initiated by the downstream start site. Towards this, we generated a series of mutants of the NA from the WSN strain of influenza. In these constructs, we mutated one or both of the canonical and candidate downstream start site to codons that are not near-cognate AUG codons. $1_{wt}43_{wt}$ contains the wildtype AUG codon at both the canonical and candidate downstream start sites, $1_{GUA}43_{wt}$ has the canonical start codon mutated to GUA, $1_{wt}43_{GUA}$ and $1_{wt}43_{UUA}$ have the candidate downstream start site mutated to GUA or UUA, and $1_{GUA}43_{GUA}$ and $1_{GUA}43_{UUA}$ have both codons mutated to the indicated identities. If both the canonical and downstream start sites initiate translation, then these constructs should encode both, just one, or neither of the full-length NA and NA43 proteins.

We first used these constructs to test whether NA43 is produced and localized on the cell surface. We did this by transfecting 293T cells with protein expression plasmids encoding the various NA mutants with a C-terminal V5 tag, which does not disrupt NA folding or function and can be detected by flow cytometry [84]. NA protein was detected

on the surface of cells transfected with plasmids encoding mutants that lacked an AUG at the canonical start codon (Fig. 11D, top left). Specifically, transfection of cells with the $1_{GUA}43_{wt}$ mutant led to ∼24% of the cell-surface NA found in cells transfected with wild-type NA. Production of this cell-surface NA in mutants lacking the canonical start codon is mostly dependent on the candidate downstream start site at nucleotide 43, since mutants that lack AUG at this position ($1_{GUA}43_{GUA}$ and $1_{GUA}43_{UUA}$) had much lower levels of cell-surface NA (Fig. 11D, top left). Interestingly, when we mutated the candidate downstream start site but not the canonical start site (mutants $1_{GUA}43_{wt}$ and $1_{wt}43_{GUA}$), we saw similar or even higher levels of cell-surface NA (Fig. 11D, top left). The slight increase could be because the amino acid mutations in these two constructs could increase NA expression or stabilize the protein (it is not possible to make a synonymous mutation to the downstream start site since AUG is the only codon for methionine). Such an increase in NA expression or stability is consistent with previous observations that a variety of amino acid mutations increase cell-surface NA levels [18, 19, 33, 41, 148].

We also quantified the levels of NA's enzymatic activity. As has been described previously [18], this can be done simply by transfecting cells with NA protein-expression plasmids and then measuring the total cell-surface enzymatic activity using the fluorogenic surrogate substrate MUNANA [158]. The results of the enzymatic activity assay mirror the surface expression (Fig. 11D, top left). Specifically, the mutant with only the downstream start site has ∼34% of the cell-surface activity of the wildtype NA, and this activity is largely dependent on having an AUG at position 43.

To confirm these results in another viral strain, we generated a similar set of mutants of the NA from the A/California/04/2009 (CA09) strain, which is a human isolate from the pdmH1N1 lineage. The results for this NA were broadly similar to those for the WSN strain: mutants without the canonical start codon still had detectable cell-surface protein and activity, which were largely dependent on having an AUG codon at position 43 (Fig. 11D, right panels). Therefore, the AUG at position 43 of NA is capable of initiating translation in multiple viral strains.

The results in Fig. 11D demonstrate that that the codon at position 43 can initiate translation in the absence of the canonical start site. However, because the assays used in those experiments cannot disambiguate between full-length NA and NA43, they do not fully validate the ribosome profiling data suggesting that the codon at position 43 initiates translation even in the presence of the canonical start site. We therefore performed Western blots to directly distinguish between full-length NA and NA43 on the basis of the size of the resulting protein.

NA43 is only 14 amino acids shorter than the 453 amino acid full-length NA. This size difference is difficult to resolve by Western blotting, so we created synthetic constructs by fusing the 5′ end of the NA gene (the 5′ UTR and the first 120 nucleotides of NA) to a short histone (H2B) coding sequence and a C-terminal V5 tag which results in a 180 amino acid reporter. Following the stop codon of the V5 tag, we retained the rest of the NA coding sequence. We again created variants of these synthetic constructs in which the individual AUGs were mutated to codons that are not near-cognate starts. Upon transfecting these constructs into HEK293T cells, the wildtype NA construct shows two bands (lane 2 of Fig. 11E) which run at the expected sizes for proteins initiated at the canonical AUG and the downstream AUG43. Constructs that lack AUG43 only show the band for full-length NA (lanes 3 and 4 of Fig. 11E), and a construct that lacks the canonical AUG start site only shows the band for NA43 (lane 5 of Fig. 11E). The mutants lacking an AUG at both the canonical start site and position 43 had an unexpected band intermediate in size between the full length NA and NA43 (lanes 7 and 8 of Fig. 11E). However, mutating a near cognate putative start codon at position 28 (lane 7 of Fig. 27A) largely eliminates this cryptic band.

As an additional check for initiation at AUG43 in the presence of upstream starts, we also generated constructs that introduce a frameshift just preceding position 43, such that any initiation that occurs 5′ to the frameshift will generate products that are not detected by our V5 antibody. Only the construct that contains an AUG at position 43 (and not GUA at position 43) produces a NA43 band (lane 5 versus lane 6 of Fig. 27B). Therefore, we conclude that the AUG at position 43 can initiate translation in both the presence and

absence of the canonical AUG start codon.

We next examined the extent to which NA43 could complement NA during viral replication. To do this, we used reverse genetics [82] to attempt to generate WSN influenza viruses with NAs that lacked one or both of the start sites. By far the highest viral titers were obtained by using the wildtype NA (Fig. 11F). However, we also obtained low but detectable viral titers when using NA that lacked the canonical start codon but had the AUG at position 43, but obtained no detectable viral titers when both start codons were mutated (Fig. 11F). The much lower viral titers when using the mutant that just produces NA43 could be due to a combination of reduced expression and the possible failure of NA protein lacking the cytoplasmic tail and a portion of the transmembrane domain to effectively localize to regions of viral budding [114, 128, 231, 9, 142, 124]. Notably, there was no detectable change in viral titer if we mutated the AUG at position 43 but maintained the canonical start codon. Therefore, although NA43 can weakly complement for full-length NA, it is not important for viral growth in the presence of full-length NA in cell culture.

The results in Fig. 11F suggest that NA43 does not substantially contribute to viral growth in cell culture, but simply titering viral supernatants generated by reverse genetics is a relatively insensitive way to quantify how mutations affect growth. We therefore performed competition assays between viruses with wildtype NA and NA lacking the start site at position 43, since such assays are a more sensitive way to detect small differences in viral fitness [6]. We performed these assays by mixing virus with wildtype NA with either the $1_{wt}43_{GUA}$ or $1_{wt}43_{UUA}$ mutant in roughly equal proportion, and then allowing the viruses to compete in low-MOI infections in cell culture. We then used deep sequencing to quantify the relative ratio of the wildtype and mutant NAs both prior to extensive secondary replication (10 hours post-infection) or after many rounds of replication (72 hours post-infection). If NA43 is important for viral fitness in cell culture, then we would expect virus with the wildtype NA (which has the start codon for NA43) to become enriched relative to either mutant. However, there was no enrichment of virus with the wildtype NA (Fig. 11G), indicating that NA43 is not detectably important for viral fitness in cell culture

in the presence of full-length NA.

We next considered the possibility that NA43 might contribute to viral fitness only *in vivo*. For instance, it is known that NA has additional roles that impact fitness *in vivo*, such as helping the virus access cells in the presence of mucins [134]. We therefore performed similar competition assays in mice, sequencing the viral population in the mouse lung 4 days after a low-dose infection with a mix of the wildtype and mutant viruses. In the mice, we did see enrichment of the wildtype virus relative to the $1_{wt}43_{GUA}$ NA mutant (Fig. 11H), but this was not statistically significant (Fig. 11H legend). Furthermore, the enrichment was present only for the $1_{wt}43_{GUA}$ mutant but not for the $1_{wt}43_{UUA}$ mutant, making it hard to disambiguate effects due to the lack of NA43 from the effects of introducing an amino acid substitution in NA that eliminates the downstream AUG. Therefore, our experiments did not reveal an unambiguous contribution of NA43 to viral fitness in the presence of full length NA either in cell culture or in a mouse model of infection.

Translation initiation on host transcripts during influenza infection and anti-viral response

Influenza virus infection and the host anti-viral response alter the expression of many host genes. For example, transcript levels of oxidative phosphorylation genes are refractory to host shutoff during influenza infection [13], and several hundred interferon-stimulated genes are induced during the host anti-viral response [173]. We sought to use our Ribo-seq and Ribo-seq + LTM data to examine translation initiation on host transcripts during influenza infection and the host anti-viral response.

Our strategy for calling translation initiation sites (TIS) on host transcripts was similar to the one we used for influenza transcripts (Fig. 9A). We used less stringent criteria to account for the lower read coverage of host transcripts compared to influenza transcripts (see Methods). We validated our TIS calling strategy for host transcripts using data from a previous study [121]. Our calling method resulted in slightly higher number of annotated TIS and lower number of downstream and upstream TIS than in the TISdb database [215]

created using the same data (Fig. 12A). This observation is expected from our conservative strategy of using Ribo-seq data in addition to Ribo-seq + LTM data to decrease false positives resulting from library preparation biases.

Application of our TIS calling strategy to the four samples in our study identified around 2000 candidate TIS in each sample (Fig. 12B). Across all four samples, over 75% of the called TIS were within 1 nucleotide of AUG codons, with CUG and GUG being the next most abundant codons at the called TIS (Fig. 12B). This proportion of AUG and near-cognate AUG codons was similar to those observed with data from HEK293T cells (Fig 28A) [121]. 875 of the called TIS were shared between all 4 of our samples (Fig. 12C), which we designate as high-confidence TIS for further analyses. Among these high-confidence TIS, over 60% corresponded to annotated canonical start codons (Fig. 12D), further validating our start site calling strategy. Less than half of the high confidence TIS identified in our study were shared with those identified from earlier work on HEK293T cells (Fig 28B). This modest overlap in called TIS likely reflects the distinct gene expression landscape between the A549 cells used in our study and HEK293T cells (Fig 28C). The uTIS were highly enriched for near-cognate AUG codons in comparison to dTIS among the high-confidence TIS (Fig. 12E), recapitulating previous observations (Fig. 28D) [94, 121].

Previous reporter studies found that non-AUG initiation could be specifically up-regulated under conditions of inflammation or infection leading to antigenic presentation of cryptic peptides [159]. We first sought to test the generality of this observation at the genome-wide level using our called TIS set. Comparison of number of called TIS between our four samples (Fig. 12B) did not reveal a globally higher proportion of non-AUG TIS upon influenza infection, interferon-β stimulation, or under both stimuli relative to the uninfected control sample ($P > 0.05$ for all three treatments, binomial proportion test). Since non-AUG TIS tend to be enriched among uTIS and dTIS in comparison to aTIS (Fig. 12E), consistently, we also did not find a higher proportion of uTIS or dTIS upon any of the stimuli relative to the uninfected control sample ($P > 0.05$ for all three treatments for both

uTIS and dTIS, binomial proportion test; Fig. 28E).

We then considered the possibility that the degree of increased non-AUG initiation during influenza infection or interferon-β stimulation might be too weak to be detected as a globally higher number of non-AUG start codons in our TIS set called using the Ribo-seq + LTM data. However, even a slightly higher degree of non-AUG initiation during these inflammatory stimuli might favor evolutionary selection for a higher proportion of non-AUG TIS in genes that are specifically up-regulated under these conditions. To test if there is a higher proportion of non-AUG TIS among induced genes, we first identified genes that were induced greater than 2-fold (average of Ribo-seq and RNA-Seq counts) under each of the three treatments in our study. Interferon-β treatment or interferon-β treatment followed by influenza infection resulted in up-regulation of around 300 genes ($>$2-fold, Fig. 12F). As expected, the most highly induced genes are well-characterized interferon-stimulated genes such as MxA or IFIT1. The genes induced upon either interferon-β treatment or interferon-β treatment followed by influenza infection show a high degree ($> 90\%$) of overlap (Fig 28F). By contrast, influenza infection on its own resulted in up-regulation of a small set of 16 genes ($>$2-fold, Fig. 12F) including only a few interferon-stimulated genes. This observation is consistent with recent work showing that activation of immune pathways by influenza virus is rare at the single cell level during infections with viruses that have relatively few defective particles [174].

We examined the TIS codon identity and type in induced genes, and compared it to a control set of all TIS that were called in any of our samples. This analysis revealed that genes that were induced upon either interferon-β treatment or interferon-β treatment followed by influenza infection had a higher proportion of near-cognate AUG TIS relative to AUG TIS [34% (+ifn induced genes) / 33% (+ifn +vir induced genes) vs 26% (all genes), $P = 0.02$, binomial proportion test, Fig. 12G]. uTIS and dTIS were also present in higher proportion than aTIS among the interferon-induced genes, as expected from the over-representation of non-AUG codons among these TIS types [59% (+ifn induced genes) / 58% (+ifn +vir induced genes) vs 47% (all genes), $P < 0.01$, binomial proportion test,

Fig. 12H]. We did not find the proportion of near-cognate AUG TIS or uTIS and dTIS in the influenza-induced genes to be significantly different from those in all genes ($P > 0.05$, binomial proportion test, Fig. 12G). This lack of statistical significance is likely due to the meager number of called TIS (N=5) in the set of influenza-induced genes. One important caveat that could bias these observations which are based off our Ribo-seq + LTM data is the runoff of elongating ribosomes caused by the lactidomycin treatment. The increased proportion of free ribosomal subunits under these conditions could lead to artifactual detection of initiating ribosomes at start sites that are not normally used in the absence of lactidomycin treatment. However, there is no a priori reason why such artifactual detection will favor an increased proportion of non-AUG initiation specifically on interferon-β induced transcripts over the remaining transcripts. Thus our observations are consistent with the hypothesis [159] that near-cognate AUG initiation at non-canonical start sites might be more common in an anti-viral cellular state induced by interferon-β treatment. However, in our data this increase arises from the presence of more non-canonical start sites in transcriptionally induced genes, rather than increased non-canonical initiation on existing transcripts.

Discussion

We have performed the first comprehensive characterization of translation initiation sites on viral and host transcripts during influenza virus infection. In viral transcripts, we identified a total of 14 high confidence translation initiation sites, including 7 of 8 canonical translation initiation sites and two previously characterized non-canonical translation initiation sites: PB1-F2 [35] and a start site at nucleotide 113 of M [221]. The seven alternate start sites that we identified were distributed across the PB1, M, NP, NA, and PA segments, and we found candidate novel N-terminal truncations in NP, PA, and NA.

We biochemically validated one of the new viral alternate start sites, a downstream AUG at codon 43 of NA. We showed that the NA43 protein initiated by this start site is

produced even in the context of the canonical start site, is expressed on the cell surface, and is enzymatically active. In addition, this NA43 protein is capable of supporting low levels of viral growth even in the absence of full length NA. The AUG codon at 43 is conserved in most N1 viral lineages. However, we were unable to find any evidence that NA43 impacts viral fitness in cell culture or mice, at least at the relatively low resolution with which viral fitness can be measured in laboratory settings. The N-terminal cytoplasmic tail of NA helps localize the protein to areas of viral budding on the cell surface [114, 128, 231, 9, 142, 124]. Therefore, the N-terminally truncated NA43 could localize to different regions of the cell surface than full-length NA. It is interesting to speculate whether such altered localization of the NA43 protein could have some phenotypic significance, such as reducing viral coinfection [86].

One of the motivations for our study was to search for evidence of alternate translation initiation at CUG codons in influenza virus. We hypothesized that initiation at such codons might lead to immune recognition of influenza virus, since CUG codons can initiate immune epitopes [24, 205, 34, 227, 179, 193], and initiation at CUG codons is thought to be upregulated during conditions of cellular stress including viral infection [191, 159, 229]. We found evolutionary signatures in the viral genome that were consistent with selection against CUG-mediated translation initiation in lineages of influenza that have adapted to mammalian hosts. However, we found minimal experiment support for CUG initiation on influenza transcripts in our ribosome profiling experiments. One possibility is that CUG initiation does generate evolutionarily relevant immune epitopes, but that the levels of CUG initiation are too low to be detected by our experimental design. For instance, CUG initiation could be partially refractory to capture by the translation initiation inhibitor lactidomycin used in our experiments [191], or perhaps only occurs in certain types of cells. In addition, even extremely low levels of translation initiation that are difficult to detect by most experimental methods can generate peptides that can still be recognized by T-cells [80]. A second possibility is that the selection against CUG codons is due to some pressure unrelated to translation initiation. There are other evolutionary signatures of in-

fluenza adaptation to mammalian hosts with uncertain origin, such as the decrease in GC content of influenza genome during viral adaptation to mammalian hosts [73].

We also did not find significant shifts in global start site usage towards non-AUG translation initiation during either viral infection of interferon treatment, despite the fact that CUG initiation has been shown to be enhanced in these conditions using biochemical and immunological assays [159]. However, the subset of transcriptionally induced genes upon interferon-β treatment did have a significantly higher number of non-AUG translation initiation sites than other genes. Whether these alternate start sites serve any biological function will require further study. One interesting possibility is that the peptides generated from these non-AUG start sites on anti-viral genes could harbor T-cell epitopes that are relevant during the host immune response.

Our ribosome profiling experiments used the translation initiation inhibitor lactidomycin to identify candidate start sites in influenza and host transcripts. One potential concern is that the extended lactidomycin arrest of initiating ribosomes could cause a traffic jam of scanning pre-initiation complexes [95] and lead to promiscuous initiation. Indeed more stringent initiation profiling methods have been developed to address this concern [64]. However, our goal was to detect start sites that might have very low levels of initiation. Hence we focused on achieving the highest possible sensitivity in our assay at the cost of possibly inflating the number of upstream start sites. Further, all the novel start sites identified on influenza transcripts occur downstream of the canonical initiation codon, indicating that traffic jams of pre-initiation complexes is not a major problem in analysis of the viral data. If anything, downstream start sites might be more prone to being missed by our method due to blocks caused by the initiating ribosome at the canonical start codon.

Overall, our work provides the first comprehensive analysis of translation initiation during influenza virus infection. We identified several new alternate translation initiation sites, one of which produces a functionally active viral protein that has not been previously described. However, we found little evidence for large-scale initiation of translation at non-canonical start codons such as CUGs on viral transcripts, and only a modest increase

in the proportion of non-canonical and non-AUG start codons on host transcripts upregulated during the anti-viral response. The relative paucity of evidence for virally induced alternate translation initiation in our ribosome profiling experiments is seemingly at odds with many studies [24, 205, 34, 227, 179, 193, 191, 159, 229] highlighting its potential role in generating immune epitopes during viral infection, as well as our own evolutionary analyses suggesting evolutionary selection against putative CUG start sites in the viral genome. Further work will be needed to resolve this apparent conundrum and define the extent and significance of alternate translation initiation during viral infection and the resulting immune response.

3.5 Materials and methods

3.5.1 Data and code availability

All high-throughput sequencing data is available from GEO under accession: GSE114636. Scripts for performing all analyses and generating figures in this manuscript are available at `https://github.com/rasilab/machkovech_2018`.

3.5.2 Evolutionary analysis of selection against initiation at CUG codons in influenza

Influenza sequence sets

To examine signatures of selection against translation initiation at CUG codons in influenza, we first assembled five influenza sequence sets: human, classical swine, duck, chicken, and human H5N1 influenza. For these sequence sets we only include the protein coding sequences from segments that do not reassort in the human influenza lineage descended from the 1918 lineage: PB2, PA, NP, NS1, M1, M2, and NS2. We retained sequences for a strain if that strain had sequences for all genes in the gene set. If there were multiple sequences for a given strain, we kept just one.

All human, H5N1, swine and avian protein coding sequences were downloaded from

the Influenza Virus Resource. (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html)

For human influenza, we kept coding sequences descended from the 1918 virus, which includes H1N1 from 1918 to 1957, H2N2 from 1957 to 1968, and H3N2 from 1968 to 2011 [49]. Viruses from the 2009 swine-origin H1N1 pandemic were excluded. We subsampled our sequences so that we kept at most 5 randomly chosen strains per year.

For human H5N1 influenza, we obtained sequences from 1997 to 2011, and did not subsample our sequences due to the low number of sequences.

For duck influenza, we obtained sequences from 1956 to 2011. We subsampled our duck influenza sequences so that we kept at most 5 randomly chosen strains per year.

For chicken influenza, we obtained sequences from 1934 to 2011. We subsampled our chicken sequences so that we kept at most 5 randomly chosen strains per year.

For swine influenza we selected classical swine H1N1 influenza viruses isolated in North America between 1934 and 1998.

For alignment and subsequent analysis, we appended the seven protein coding regions (PB2, PA, NP, NS1, M1, M2, and NS2) together for each strain. Sequence sets are aligned to A/Brevig Mission/1/1918 virus using MUSCLE [53], with gaps relative to the A/Brevig Mission/1/1918 strain stripped away. Alignments are provided as S1 File.

Number of sequence motifs and motif odds ratio

In Fig. 7A, Fig. 7B, Fig. 13, and Fig. 14B we counted the number of the indicated motif (eg. CUG codons) in the indicated reading frame for PB2, PA, NP, NS1, M1, M2, and NS2 protein coding sequences for the viral strains contained in the human, duck, chicken, swine, H5N1 sequence sets.

We also calculated the motif odds ratio (OR) as in [73]. The OR accounts for the individual nucleotide content of a sequence, and therefore removes the effect of any underlying changes in nucleotide content. The OR is defined as follows (shown for the codon CUG):

$$OR_{CUG} = \frac{frequency_{CUG}}{frequency_C \times frequency_U \times frequency_G}, \quad (3.1)$$

The OR is defined as the frequency of a given motif (CUG) in a sequence divided by the product of the frequency of each nucleotide that comprise the motif (C, U, and G) the sequence. Frequency of a motif such as CUG is defined as the number of codons in a sequence that are CUG divided by the total number of codons in a sequence. Frequency of a nucleotide is the number of a given nucleotide divided by the total number of nucleotides in a sequence. A value greater than 1 indicates there is an excess of the motif given the nucleotide content, and a value less than 1 indicates that there are fewer of the motif given the nucleotide content.

In Fig. 14A, we calculated the odds ratio for CUG for the indicated reading frame in each set of protein-coding sequences for the human, duck, chicken, swine, H5N1 sequence sets.

Length of CUG/CUH putative ORFs

For the non-canonical reading frames 1 and 2 of influenza we also examined selection against putative CUG initiation by examining whether there was a host-specific difference in putative ORF lengths (Fig. 7C). We selected a timeframe where the human and swine influenza strains should be reasonably host adapted (1970 to 2011), and took sequences for the human, duck, chicken, swine, H5N1 sequence sets from that time. We calculated the length of all putative ORFs initiated by CUG or CUH (where H is A, U or C) in reading frames 1 and 2 for each influenza sequence set. To plot the data, we took all of the ORF lengths and divided them into 5 ORF length bins. For each influenza host sequence set, we calculated the average number of ORFs for that bin by dividing the number of ORFs in that bin by the number of influenza strains for that host.

Consensus motif for CUG start sites

For Fig. 15, we generated a logoplot of the consensus context of CUG start sites called in the ribosome profiling data of Ingolia et al (Supplemental File 3 of [94]).

Plasmids

Influenza reverse genetics plasmids used to generate viral stocks for ribosome profiling

The virus used in the ribosome profiling experiments is a reassortant of the A/PuertoRico/1934 (PR8) and A/WSN/1933 (WSN) influenza strains. NA, HA, M, and NS gene segments are from the WSN strain, encoded by the bidirectional pHW plasmids [82]: pHW186-NA, pHW184-HA, pHW187-M, pHW188-NS. PB1, PB2, and PA segments are from the PR8 strain, encoded by the following plasmids: pHW192-PB1, pHW191-PB2, pHW193-PA. The NP segment was either one of two recoded variants of the PR8 NP, encoded by the following plasmids: pHW-lowCUG-PR8-NP or pHW-highCUG-PR8-NP.

To specifically examine putative alternate initiation at the CUG codons that are selected against in reading frame 0 of influenza, we recoded the PR8 NP to contain either few (low CUG PR8 NP) or many CUG codons (high CUG PR8 NP) (sequences are in S4 File). We specifically chose to recode the CUG content of NP because NP contains many CD8 T-cell epitopes[228, 79] and CUG initiation can lead to the generation of CD8 T-cell epitopes [191, 159]. Furthermore, we chose PR8 NP as the CD8 T-cell epitopes in PR8 NP for murine models of infection are well characterized [228, 20, 44]. To generate low CUG PR8 NP, we depleted PR8 NP of the most common alternate start codons AUG, CUG, and GUG as much as possible in all reading frames. We did this because we used low CUG PR8 to generate high CUG PR8, and we wanted to begin with a low background of possible alternate initiation sites. We generated high CUG PR8 NP by adding 20 CUG motifs that occur in reading frame 0 of natural influenza NP sequences to low CUG NP. The mutations we introduced to generate low CUG PR8 NP and high CUG PR8 NP are

made with the following constraints: changes must be synonymous with regards to reading frame 0, any synonymous codons that are introduced are chosen to be as frequent as possible in natural existing sequences, and codons must exist in at least 100 of the sequences. The sequences we used for this analysis are all full-length influenza A NP coding sequences from the Influenza Virus Resource. Sequences were aligned using MUSCLE [53], and alignments are included in S2 File.

NA43 plasmids for MUNANA and NA surface expression

To measure NA surface expression and NA MUNANA activity in Fig. 11D, we generated constructs by placing NA into an HDM plasmid in which expression of the insert is driven by the CMV promoter. Following the CMV promoter, we included the NA viral 5′ UTR, NA coding sequence, a C-terminal V-5 epitope tag (used for surface staining), and an internal ribosomal entry site (IRES) GFP (used for calculating transfection efficiency).

We made the following set of mutagenized constructs for WSN NA: pHDM-vUTR-WSN-NA_1ATG-43ATG_V5-IRES-GFP, pHDM-vUTR-WSN-NA_1ATG-43GTA_V5-IRES-GFP, pHDM-vUTR-WSN-NA_1ATG-43TTA_V5-IRES-GFP, pHDM-vUTR-WSN-NA_1GTA-43ATG_V5-IRES-GFP, pHDM-vUTR-WSN-NA_1GTA-43GTA_V5-IRES-GFP, and pHDM-vUTR-WSN-NA_1GTA-43TTA_V5-IRES-GFP.

Similar constructs were made for California/04/2009 NA (CA09). In these constructs, we mutated an additional in-frame AUG that was present at coding nucleotides 55-57 to UUA. We did this because the downstream AUG may function as a start site in the absence of either of the other AUG codons, and have the potential to complicate interpretation of our assays if the AUG starting at nucleotide 55 generates a functional NA.

For CA09 we made: pHDM-vUTR-CA09-NA_1ATG-43ATG-55TTA_V5-IRES-GFP, pHDM-vUTR-CA09-NA_1ATG-43GTA-55TTA_V5-IRES-GFP, pHDM-vUTR-CA09-NA_1GTA-43ATG-55TTA_V5-IRES-GFP, pHDM-vUTR-CA09-NA_1GTA-43GTA-55TTA_V5-IRES-GFP.

We also made a construct that lacks the V5 tag as a negative control for background

V5 staining (pHDM-WSN-NA-IRES-GFP).

NA43 plasmids for westerns

To examine whether there is initiation at site 43 of WSN NA (Fig. 11E), we generated constructs in an HDM plasmid in which expression of the insert is driven by the CMV promoter and the 5′ UTR of WSN NA. The protein coding sequence consists of codons 1-40 of WSN NA fused to Histone-2B, followed by a C-terminal V5 epitope tag (for immunoblot detection). The stop codon is followed by the remainder of NA and IRES GFP. We made the following constructs: pHDM-vUTR-WSN-NA-aa1-40_1ATG-43ATG_H2B-V5-IRES-GFP, pHDM-vUTR-WSN-NA-aa1-40_1ATG-43GTA_H2B-V5-IRES-GFP, pHDM-vUTR-WSN-NA-aa1-40_1ATG-43TTA_H2B-V5-IRES-GFP, pHDM-vUTR-WSN-NA-aa1-40_1GTA-43ATG_H2B-V5-IRES-GFP,pHDM-vUTR-WSN-NA-aa1-40_1GTA-43GTA_H2B-V5-IRES-GFP, and pHDM-vUTR-WSN-NA-aa1-40_1GTA-43TTA_H2B-V5-IRES-GFP. We made a construct that lacks the 5′ viral UTR and the first 42 coding nucleotides of NA, so begins at the AUG at coding nucleotide 43 which serves as a size control for NA43 (pHDM-WSN-NA-aa14-40-H2B-V5-IRES-GFP). We also made several constructs to eliminate cryptic products that appear in constructs lacking an AUG at both the canonical start and position 43. pHDM-vUTR-WSN-NA-aa1-40_1GTA-28GTT-43GTA_H2B-V5-IRES-GFP mutates a near cognate start codon at position 28. pHDM-vUTR-WSN-NA-aa1-40_1GTA-35Tinsert-43GTA_H2B-V5-IRES-GFP and pHDM-vUTR-WSN-NA-aa1-40_1GTA-35Tinsert-43ATG_H2B-V5-IRES-GFP contain a T at position 35 to introduce a frameshift.

NA43 plasmids for viral rescue

We used the bidirectional pHW plasmids [82] for all WSN segments (pHW181-PB2, pHW182-PB1, pHW183-PA,pHW184-HA, pHW185-NP, pHW187-M, pHW188-NS) except NA. For NA we used pHH [146], which includes only the Pol I promoter, so that NA mRNA would only be made from the native vRNA. We generated the following set of mutagenized con-

structs: pHH-WSN-NA_1ATG-43ATG, pHH-WSN-NA_1ATG-43GTA, pHH-WSN-NA_1ATG-43TTA, pHH-WSN-NA_1GTA-43ATG, pHH-WSN-NA_1GTA-43GTA, pHH-WSN-NA_1GTA-43TTA.

NP plasmids for westerns

To examine whether there is initiation at site 322 of NP (Fig. 11E), we generated constructs in an HDM plasmid in which expression of the insert is driven by the CMV promoter. Following the CMV promoter, we include the 5′ UTR of the indicated NP followed by a C-terminal 3x FLAG epitope tag (for immunoblot detection). We made the following constructs: pHDM-vUTR-lowCTP-PR8-NP-FLAG, pHDM-vUTR-highCUG-PR8-NP-FLAG, pHDM-vUTR-wt-PR8-NP-FLAG. We made a construct that lacks the 5′ viral UTR and begins at the AUG at coding nucleotide 322 which serves as a size control for NP322 (pHDM-322start-wt-PR8-FLAG).

3.5.3 Cells and cell culture media

We used the human lung epithelial carcinoma line A549 (ATCC CCL-185), the human embryonic kidney cell line 293T (ATCC CRL-3216), the MDCK-SIAT1 variant of the Madin Darby canine kidney cell line overexpressing human SIAT1 (Sigma-Aldrich 05071502), and MDCK-SIAT1-TMPRSS2 variant cells which express both SIAT1 and the protease TMPRSS2 [119].

All cell lines were maintained in D10 (DMEM supplemented with 10% heat-inactivated fetal bovine serum, 2 mM L-glutamine, 100 U of penicillin/ml, and 100 μg of streptomycin/ml). Experiments were performed with D10 unless otherwise indicated.

WSN growth media contains Opti-MEM (Gibco) supplemented with 0.05% heat inactivated FBS, 0.3% BSA, 100 U of penicillin/ml, 100 μg of streptomycin/ml, and 100 μg of calcium chloride/ml).

3.5.4 Virus generation by reverse genetics

Generation of virus for ribosome profiling

A co-culture of 293T and MDCK-SIAT1 cells seeded at 4×10^5 and 0.5×10^5 cells per well respectively in 6-well dishes were transfected using BioT and a mixture containing 250 ng of each of the eight plasmids encoding WSN NA, HA, M, NS (pHW186-NA, pHW184-HA, pHW187-M, pHW188-NS), PR8 PB1, PB2, PA (pHW192-PB1, pHW191-PB2, pHW193-PA), and either low or high CUG PR8 NP (pHW-lowCUG-PR8-NP or pHW-highCUG-PR8-NP). We refer to these viruses as low CUG NP WSN and high CUG NP WSN. At 20 hours post transfection, we changed the media from D10 to WSN growth media. At 48 hours post-transfection we collected the viral supernatant, centrifuged the sample at 2000g for 5 minutes to remove cellular debris, and stored aliquots of the clarified viral supernatant at -80 ℃. We titered the virus by TCID50 (using MDCK-SIAT1 cells). Viruses were expanded starting from an MOI of .05 for 55 hr. The resulting expanded virus was collected and stored in the same manner. The expanded virus was titered by HA staining (using A549 cells) and TCID50 (using MDCK-SIAT1) and used for ribosome profiling.

Generation of viruses for NA43 follow-up

A co-culture of 293T and MDCK-SIAT1-TMPRSS2 cells seeded at 4×10^5 and 0.5×10^5 cells per well in 6-well dishes were transfected using BioT and a mixture containing 250 ng of each of the eight plasmids encoding WT WSN PB2, PB1, PA, HA, NP, M, NS (pHW181-PB2, pHW182-PB1, pHW183-PA, pHW184-HA, pHW185-NP, pHW187-M, pHW188-NS) and the appropriate NA variant (pHH-WSN-NA_1ATG-43ATG, pHH-WSN-NA_1ATG-43GTA, pHH-WSN-NA_1ATG-43TTA, pHH-WSN-NA_1GTA-43ATG, pHH-WSN-NA_1GTA-43GTA, pHH-WSN-NA_1GTA-43TTA). MDCK-SIAT1-TMPRSS2 cells were used as it enables HA activation in the absence of NA activity, thus allowing us to minimize indirect effects of NA absence. Viruses were rescued in biological triplicate with three

separate minipreps of each NA construct. At 20 hours post transfection, we changed the media from D10 to WSN growth media. We collected 500 ul aliquots of virus 48 and 72 hours post transfection. We titered the virus by TCID50 (using MDCK-SIAT1 TMPRSS2 cells). Viruses were expanded starting from an MOI of .01 for 48 hours. We titered the expanded virus by TCID50 (using MDCK-SIAT1-TMPRSS2 cells).

3.5.5 Virus titering

By HA staining

We infected A549 cells in WSN growth media for 10 hours, collected cells, and stained for HA using H7-L19 antibody [69, 50] at 10 ug/ml, followed by goat anti-mouse IgG secondary antibody conjugated to APC, and analyzed by flow cytometry to determine the fraction of cells that were HA positive. We used the Poisson equation to calculate viral titer with respect to HA expressing units.

By TCID50

We titered virus by TCID50 on MDCK-SIAT1 or MDCK-SIAT1-TMPRSS2 expressing cells and calculated viral titer using the Reed-Muench formula [165] implemented here: `https://github.com/jbloomlab/reedmuenchcalculator`.

3.5.6 Ribosome profiling

Ribo-seq library generation

We performed ribosome profiling (Ribo-seq and Ribo-seq + LTM) with four different samples: untreated A549 cells, interferon-β treated A549 cells, influenza virus infected A549 cells, interferon-β pre-treated and influenza infected A549 cells (Fig. 8A). For each sample, we used two 15 cm plates of A549 cells, seeded with 7×10^6 cells 16 hours prior to start of treatment such that cells were 70-80% confluent at harvest. Five hours before

influenza infection, the medium in each plate was replaced with fresh D10 growth media either with or without interferon-β (at 20 U/ml). At the time of viral infection, the growth media was removed and replaced with four different WSN growth media corresponding to the four different samples in our experiment. For the +vir sample, we used concentrated virus stocks comprising approximately 2% of the growth media added to cells. Since virus was grown on MDCK-SIAT1 cells, we added the same volume of MDCK-SIAT1 supernatant as the viral stock to the uninfected sample (ctrl sample). For the +ifn sample, we used MDCK-SIAT1 supernatant supplemented with 20 U/ml of interferon-β (PBL 11415). For the +ifn +vir sample, we mixed the viral supernatant with interferon-β and added it to cells.

Influenza infections consisted of a 1:1 mixture of high CUG NP WSN and low CUG NP WSN viruses at a total dosage of 1 HA-expressing units / cell as calculated using A549 cells not treated with interferon-β. This corresponded to an MOI of approximately 30 TCID_{50} per cell as measured using MDCK-SIAT1 cells. These infection conditions resulted in approximately 70% HA high positive cells for the +vir sample. Samples were harvested 5 hours after influenza infection for ribosome profiling library preparation.

Ribosome profiling protocol was adapted from [90] with the following modifications. For our Ribo-seq + LTM samples, we removed the treatment medium from each plate and added media back containing LTM at 5 μM. These LTM-containing plates were then incubated for 30 min to allow runoff of elongating ribosomes [121]. For sample harvesting, we removed media from each plate and flash froze samples by placing the plate in liquid nitrogen and transferred to -80 ℃ until lysis. We took a portion of the cell lysate from our Ribo-seq samples for RNA-Seq library preparation. For ribosome profiling, we performed nuclease footprinting treatment by adding 600U RNase I (Invitrogen AM2294) to 18 A_{260} units of RNA [4]. We collected ribosome protected mRNA fragments using MicroSpin S-400 HR Columns (GE Healthcare 27-5140-01) following instructions from the Art-Seq ribosome profiling kit (RPHMR12126, Illumina), and purified RNA from the flow through for size selection. We gel-purified ribosome protected fragments with length between 26

and 34 nucleotides using RNA oligo size markers. We used polyA tailing instead of linker ligation following previous studies [93, 121]. Finally, we prepared RNA-seq libraries using the NEBNext Ultra Directional RNA Library Prep Kit (NEB E7420) after depleting ribosome RNA using the Ribozero Gold kit (Illumina MRZG126). Ribo-seq and RNA-seq libraries were sequenced on an Illumina HiSeq 2500 in 50bp single end mode resulting in between 15–40 million reads per sample (see Fig. 17A).

Pre-processing and alignment of sequencing data

For Ribo-seq libraries, we trimmed polyA tails using `cutadapt 1.12` [133] with parameters `--adapter=AAAAAAAAAA --length=40 --minimum-length=25` to retain trimmed reads that were between 25 and 40 nucleotides in length. For RNA-seq libraries, we first reverse complemented the read to obtain the sense orientation using `fastx_reverse_complement 0.0.13` (`http://hannonlab.cshl.edu/fastx_toolkit/`) with parameters `--Q33`, and then trimmed reads to 40 nucleotides. To remove rRNA, we discarded trimmed reads aligning to four human rRNA sequences (28S:NR_003287.2, 18S:NR_003286.2, 5.8S:NR_003285.2, and 5S:NR_023363.1) from the `hg38` genome. The alignment was done using `bowtie 1.1.1` [118] with parameters `--seedlen=23 --threads=8`. We aligned the remaining non-rRNA to human transcripts (`Gencode v24`) using `rsem 1.2.31` [125] with parameters `--output-genome- --sort-bam-by-coordinate`. We also aligned the non-rRNA reads to an influenza genome containing both low and high CUG NP using the above `rsem` parameters including the extra options `--seed-length 21 --bowtie-n 3`. The influenza genome file is provided in S4 File and the corresponding annotations in S5 File.

Calculation of influenza and host transcript coverage

Proportion of human versus influenza reads

The posterior probablilty score from `rsem` (`ZW` field in the `BAM` output) was used to calculate coverage for reads on transcripts. For influenza reads, we considered only reads that map

to the positive sense of influenza transcripts for all analyses except where explicitly indicated otherwise. To obtain higher resolution mapping of ribosome protected fragments, we performed variable trimming based on fragment length, reads that were between 25 and 32 nucleotides in length were assigned to a P-site at 14 nucleotides from the 5′ end of the read, and reads between 33 and 39 nucleotides were assigned to a P-site at 15 nucleotides from the 5′ end of the read.

Ribo-seq start or stop codon aligned read aggregation plots

For human transcripts, a set of non-redundant protein-coding transcripts was compiled by using the `gencode v24` annotations and applying the following criteria: The corresponding CDS must be part of the Consensus CDS (CCDS) project; It must be labeled as a `principal transcript` by APPRIS; For each gene, we selected the lowest numbered CCDS ID; For each CCDS, we selected the lowest numbered transcript ID.

For aggregation of sequencing reads, we further selected the set of transcripts requiring that each transcript has a minimum coverage of 0.33 reads per nucleotide in the coding region in a given sample. For each ribosome P-site in a transcript, the normalized ribosome density value was calculated by dividing the P-site read count by the average P-site read count across the coding region of that transcript. The normalized P-site density across all transcripts was calculated by averaging each P-site position (aligned relative to the start or stop codon).

Proportion of Ribo-seq reads aligning to each reading frame

Normalized P-site density across reading frames was calculated for the coding regions of the nonredundant set of transcripts included in the analysis used for read aggregation plots. The single and dual coding regions of NS and PB1 are defined in S5 File, and the normalized P-site density across the singly and dually coded regions was plotted.

Calling and analysis of candidate TIS

We used a zero-truncated negative binomial distribution (ZTNB) to statistically model the background distribution of Ribo-seq and Ribo-seq + LTM counts in transcripts with more than 50 positions with non-zero counts [206, 64]. We first added the Ribo-seq and the Ribo-seq + LTM read counts of the two neighboring positions to each position in the genome (referred to as pooled counts below). This was done in order to account for the +/–1 nucleotide uncertainty in the P-site assignment.

Candidate start sites were identified based on the following criteria: For influenza transcripts, the ZTNB-based P-value for the Ribo-seq + LTM pooled counts at that location must be $<$0.01 and 1000-fold higher than the P-value of the Ribo-seq pooled counts at the same location (Fig. 9A, left panel), or must have an absolute P-value less than 10^{-7}. For host transcripts, we required the Ribo-seq + LTM P-value to be only 100-fold higher than the P-value of the Ribo-seq pooled counts. Additionally, for host transcripts, we required that the read counts be greater than an absolute threshold across all transcripts. This threshold was estimated by requiring P$<$0.05 in a ZTNB model fit to the bottom 99% of all non-zero P-site Ribo-seq + LTM pooled counts across all transcripts. Only the highest pooled counts within each 30 nucleotide window was called as a candidate TIS. From the called TIS, we assigned the identity of the start codon by looking at a window -1 to +1 nucleotides from the TIS peak and assigning the start codon based on following hierarchy: AUG, CUG, GUG, UUG, AUA, AUC, AUU, AAG, ACG, AGG, and other. If there are multiple near cognate codons in the window, the codon was assigned based on the order in the above list.

For Fig. 9H and 9I, we consider the canonical frame defined by the aTIS as frame 0, and we designate any ORF out of frame with the aTIS as an out-of-frame ORF. For candidate host TIS, we designate the start codon of the canonical transcript of each gene (as defined above) as the annotated TIS. uTIS, dTIS, and the frame of their ORF are identified with respect to the annotated TIS. Host genes that have a median-normalized

fold-change in average Ribo-seq and RNA-seq counts greater than 2-fold upon +ifn, +vir, or +ifn +vir treatment are considered as induced genes under that condition.

For analysis of data from [121], we downloaded the raw sequencing data from the Sequence Read Archive, BioProject PRJNA171327, and analyzed it using the same pipeline that we used for our data. The same pipeline was also used for analysis of the raw sequencing data from [13] (NCBI Gene Expression Omnibus, GSE82232). The only difference from the analysis of our dataset is that we identified P-site as the 13th nucleotide from the 5′ end of the read for both these datasets.

Binomial proportion test for comparing proportions of different TIS was done using the `R` function `prop.test` with the alternate hypothesis set to `greater`.

Examining possible initiation at CUG codons in high CUG NP variants

In Fig. 10A, we considered reads that map with zero mismatches to either low or high CUG NP or both NP variants. We designated reads as belonging to one of the variants if it spans a SNP (recoded CUG codon), and thus could be identified uniquely. Coverage of non-unique, low CUG, and high CUG variants is shown as a stacked bar blot. For Fig. 10B, we plot the ratio of P-site count between the high and low CUG variants at each NP position to the mean value of the same quantity across the two variants. All P-sites have a pseudocount of 1 added to both the low and high CUG read counts.

NA surface expression and MUNANA NA activity assay

The transfected 293T cells were stained for cell-surface NA using an anti-V5 AF647-conjugated antibody (Invitrogen 45-1098) at a 1:200 dilution. Cells were analyzed by flow cytometry to determine the MFI of AF647 (APC channel) among GFP-positive (transfected) cells. Reported values were normalized to the $1_{wt}43_{wt}$ NA, with the background value from the pHDM-vUTR-NA-V5-IRES-GFP (lacking V5 tag) subtracted.

NA surface expression

The transfected 293T cells were stained for cell-surface NA using an anti-V5 AF647-conjugated antibody (Invitrogen 45-1098) at a 1:200 dilution. Cells were analyzed by flow cytometry to determine the MFI of AF647 (APC channel) among GFP-positive (transfected) cells. Reported values were normalized to the $1_{wt}43_{wt}$ NA, with the background value from the pHDM-vUTR-NA-V5-IRES-GFP (lacking V5 tag) subtracted.

MUNANA NA activity assay

The transfected 293T cells were used to assay for NA activity using MUNANA substrate (Sigma M8639). We performed the MUNANA assay in a non-lysing buffer (MOPS buffered saline supplemented with 2% FBS). We used 2 dilutions of our resuspended cell sample (25% and 2.5% of original total sample volume) to ensure that we are within the dynamic range of the assay. We performed the assay in black 96-well plates in a 200 ul volume. We added MUNANA to a concentration of 150 μM and incubated samples for 45 min at 37 ℃, and quenched by adding 50 μl of 142 mM NaOH in 100% ethanol. Fluorescence was measured at an excitation wavelength of 360 nm and an emission wavelength of 448 nm. Activity was calculated by subtracting the background value of untransfected cells, correcting for transfection efficiency (% GFP positive from NA surface expression assay), and normalizing to $1_{wt}43_{wt}$ NA.

3.5.7 Western blots

We transfected 293T cells seeded at 2×10^5 cells/well in a 12-well plate with the pHDM-vUTR-WSN-NA-aa1-40-H2B-V5-IRES-GFP constructs or pHDM-vUTR-NP-FLAG constructs. We collected cell lysates 20 hour post transfection, using RIPA buffer containing (1% NP40, 1% Sodium deoxycholate, 0.1% SDS, 150 mM NaCl, 50 mM Tris (pH 8), 0.1 mM EDTA, and Roche cOmplete ULTRA Tablet Protease Inhibitor).

For NA43, lysates were run on a 16.5% tris-tricine gel. For NP, lysates were run on a 4-20% tris-glycine gel. NA was detected using 1:2500 dilution of anti-V5 antibody (Invitrogen R960), followed by a 1:2500 dilution of DyLight 800 Rabbit anti-Mouse (Invitrogen SA5-10164). NP was detected using 1:2500 dilution of anti-FLAG antibody (Sigma F1804), followed by a 1:2500 dilution goat anti-mouse Alexa-Fluor 680 (Invitrogen A-21058). For loading controls, we used either GAPDH or H3. We used anti-GAPDH at a 1:2500 dilution (RD systems AF5718), followed by a 1:2500 Alexa-Fluor 680 donkey anti-goat secondary (Invitrogen A-21084). We used anti-H3 at a 1:10000 dilution (abcam 1791), followed by a 1:10000 Alexa-Fluor 680 goat anti-rabbit secondary (Invitrogen A-21109). Western blots were imaged using the LI-COR imaging system.

3.5.8 NA43 codon analysis

To examine the conservation of the AUG at coding site 43 in Fig. 11C, we downloaded all full-length N1 NA protein-coding sequences from the Influenza Virus Resource. We used `phydms` [81] to construct a codon-level alignment in reference to the WSN sequence used in the ribosome profiling experiment. The alignment is subsampled such that all sequences have at least 2 amino acid differences relative to other sequences. This is to avoid having the alignment dominated by many highly similar sequences that are heavily represented in the database. We parse sequences into the following four sets: human seasonal H1N1 sequences isolated before 2009, human pandemic H1N1 sequences isolated after 2009, all avian sequences, and all swine sequences. We further subset human seasonal H1N1 sequences so that we only keep 1 sequence per year. Sequences are in S3 File. We then count the number of each codon at coding nucleotides 43-45 for our set of sequences. We consider CUG, GUG, UUG, ACG, AGG, AUC, AUU, AAG, AUA as near cognate start codons.

3.5.9 Competition assays to examine impact of NA43 on viral fitness

For the competition assays (Fig. 11G and H) to examine if NA43 conferred a viral fitness advantage, we made 6 mastermixes of virus containing 1:1 mixes (as measured by TCID50) containing either $1_{wt}43_{wt}$ and $1_{wt}43_{GUA}$ (N=3) or $1_{wt}43_{wt}$ and $1_{wt}43_{UUA}$ (N=3) virus. For each virus, we used 3 biological replicates of expanded $1_{wt}43_{wt}$, $1_{wt}43_{GUA}$ and $1_{wt}43_{UUA}$ virus. The same viral mastermixes were used for both cell culture and mouse competition assays The viral mastermixes underwent 1 freeze thaw cycle prior to the mouse experiment.

Cell-culture NA43 competition assays

Competitions were done at a low total MOI of 0.001 (equivalent to an MOI of 0.0005 for each individual viral variant) in MDCK-SIAT1-TMPRSS2 cells. For each competition pair, we used one well of a 6-well plate (seeded at 5×10^5cells per well) and one 10 cm dish (seeded at 1.4×10^6cells per plate). We harvested the 6-well plate for cellular RNA at 10 hour post infection to get an early estimate of the viral ratios. We let the competitions proceed for 72 hours, and at 72 hours we collected 700 μl of virus supernatant from the 10 cm dish. Competitions were performed in WSN growth media. For the 10 hour timepoint, we extracted RNA using the Qiagen RNeasy mini plus kit, following the manufacturer's protocol. For the viral supernatant samples from the cell culture competition, we extracted viral RNA using the Qiagen QIAamp Viral RNA Mini Kit, following the manufacturer's protocol (using 140 μl virus containing supernatant).

Mouse NA43 competition assays

All animal experiments were conducted in accordance with National Institutes of Health guidelines for humane treatment of animals and were approved by the Institutional Animal Care and Use Committee of the Fred Hutchinson Cancer Research Center.

For each viral mastermix, we innoculated 3 female BALB/c mice (technical replicates) with 20 μl of virus mastermix (containing 2000 TCID50). BALB/c mice were 6-8 weeks old, from Jackson labs. For infections, mice were first anesthetized with 0.2 mg ketamine and 20 μg xylazine. Mice were weighed daily and lungs were harvested and flash frozen on day 4 post infection.

We homogenized whole lung in 2.4 ml buffer RLT using the gentleMACS dissociator. The homogenate was clarified by centrifugation, and 700 μl of supernatant was used for RNA extraction using the Qiagen RNeasy kit, following the manufacturer's protocol.

Sequencing to determine NA43 mutant frequency

We reverse transcribed NA vRNA gene from the extracted RNA using SuperScript III Reverse Transcriptase for first-strand cDNA synthesis. The reverse transcriptase reaction contained 500 ng cellular RNA template or 4 μl viral RNA and used the primer U12-NA-F (5-AGCGAAAGCAGGAGTUUAA-3) at 5 μM. We then carried out targeted deep sequencing of the mutated region in NA. We first amplified a region of the NA gene that encompassed position 43 using 1.5 μl cDNA template and the primers U12-NA-F and NA-249-R (5-GGACAAAGAGAUGAATTGCCGG-3). We used the following PCR program, 95 ℃ 2 min, followed by 27 cycles of: 95 ℃ 20s, 70 ℃ 1s, 50 ℃ 30s, 70 ℃ 20s.

We purified the PCR product using 1.5X AMPure XP beads (Beckman Coulter) We performed a second round of PCR to add a portion of the Illumina adapters using 3 ng DNA and the following PCR primers: Rnd1-F-WSN-NA (5-CTTTCCCTACACGACGCT CTTCCGATCTAGCGAAAGCAGGAGTUUAA-3) and Rnd1-R-WSN-NA (5-GGAGTTC AGACGTGTGCTCTTCCGATCUGGACAAAGAGAUGAATTGCCGG-3). We used the following PCR program: 95 ℃ 2 min, followed by 8 cycles of: 95 ℃ 20s, 70 ℃ 1s, 54 ℃ 20s, 70 ℃ 20s.

We purified the PCR product using 1.5X AMPure XP beads and used 1.5 ng of this product as template for a third round of PCR using the following pair of primers that added

the remaining part of the Illumina sequencing adaptors as well as a 6 or 8-mer sample barcode (xxxxxx): (5-AAUGATACGGCGACCACCGAGATCTACACTCTTTCCCTACAC GACGCTCTTCCGATCT-3) and Rnd2-R (5-CAAGCAGAAGACGGCATACGAGATxxxxx xGTGACUGGAGTTCAGACGTGTGCTCTTCCGATCT-3). We used the following PCR program: 95 ℃ 2 min, followed by 6 cycles of: 95 ℃ 20s, 70 ℃ 1s, 58 ℃ 10s, 70 ℃ 20s.

The resulting DNA libraries were sequenced on an Illumina Hi-Seq. We processed the sequencing reads to determine the frequency of the wildtype NA and mutant NA in each competition as follows. To determine the frequency of wildtype NA, we aligned our samples to the 1_{wt} 43_{wt} sequence, and counted the number of reads containing AUG at coding nucleotides 43-45. To determine the frequency of mutant NA, we aligned our samples to etiher the $1_{wt}43_{GUA}$ or $1_{wt}43_{UUA}$ sequence, and counted the number of reads containing GUA or UUA at coding nucleotides 43-45. We used `dms_tools` [17] for alignment of sequence data, and allowed only 2 nucleotide mismatches over coding nucleotides 1-48. ata, and allowed only 2nt mismatches over coding nucleotides 1-67.

Analysis of NA43 mutant frequency

To examine if there is an enrichment of wildtype viruses mutant over time, we first calculated the ratio of wildtype verses mutant at each timepoint using the counts of AUG, UUA, or GUA codons at postion 43-45. We then calculated the enrichment of wildtype relative to mutant over time. To do this we divided the ratio of wildtype to mutant at the endpoint (either 72 hour cell-culture or mouse value) by the ratio of wildtype to mutant at the 10 hour cell-culture timepoint. We compared the endpoint to the 10 hour timepoint instead of assuming that the initial ratio is 50:50 as the precision of TCID50 is less than that of deep sequencing, and the 10 hour timepoint allows us to measure relative levels of infectious virus for each variant. A value greater than 1 indicates an enrichment of wildtype over time. To assess significance, we performed two one-sided paired *t*-tests (mouse $1_{wt}43_{GUA}$ and mouse $1_{wt}43_{UUA}$). For the *t*-test, we used the log transformed ratios of wildtype to

mutant at each timepoint. For the mouse assay, we first took the mean of the 3 technical replicates (using the log transformed value), and paired the average mouse value for each biological replicate to the 10 hour cell culture timepoint. The sequencing counts of codons and the resulting ratios used in analysis are included in S1 Table.

3.6 Acknowledgments

We thank Adam Geballe, Ji Wan, & Shu-Bing Qian for helpful discussions. The work of ARS is supported by grant R35 GM119835 of the NIGMS of the NIH. The work of JDB is supported by grant R01 AI127893 of the NIAID of the NIH, and a Faculty Scholars Award from HHMI and the Simons Foundation. HMM was supported in part by grant T32 AI083203 of the NIAID of the NIH.

3.7 Supporting information

Supplemental material is described here and will be accessible upon publication.

Below is a description of supplemental files that will be available upon publication of the manuscript.

S1 Table. Deep sequencing from NA43 competition. Sequencing counts and ratios calculated for cell culture and mouse $1_{wt}43_{wt}$ verses $1_{wt}43_{GUA}$ and $1_{wt}43_{UUA}$ virus competitions.

S1 File. Influenza sequence alignments used for evolutionary analysis of CUG codons. Alignments of protein-coding sequences of influenza PB2, PA, NP, M and NS to the A/Brevig Mission/1/1918 virus. Alignments were performed by appending the seven protein coding sequences together for each viral strain. PB2 is from position 1 to 2280, PA is from position 2281 to 4431, NP from position 4432 to 5928, M1 from position 5929 to

6687, M2 from position 6688 to 6981, NS1 from position 6982 to 7674, NS2 from position 7675 to 8040.

S2 File. Influenza sequence alignments of NP used for generating low CUG PR8 NP and high CUG PR8 NP. Alignments of protein-coding sequences of influenza NP.

S3 File. Influenza sequence alignments of N1 NA . Alignments of protein-coding sequences of influenza NA used for analysis of codon identity at position 43.

S4 File. Influenza genome. This file contains the influenza genome used for our ribosome profiling analysis, including low and high CUG PR8 NP sequences.

S5 File. Influenza GTF. This file contains annotations for influenza used for our ribosome profiling analysis.

Chapter 4

CONCLUSION

There are numerous forces that shape influenza evolution. Immune pressure from antibodies is one of the well-described forces that drives influenza evolution. Surprisingly, while there is ample experimental support that CD8 T-cells control and limit influenza infection there is a lack of statistical support that CD-8 T-cells drive influenza evolution. I hypothesized that CD8 T-cells are a selective pressure but that conventional tests were not sensitive enough to detect selection in regions of functional constraint. In Chapter 2, I employed two novel statistical methods that have increased sensitivity to detect selection in influenza nucleoprotein. First, by comparing parallel lineages of influenza, I internally controlled for site-to-site variation in functional constraint. I compared epitope sites in the human lineage of influenza to homologous sites in the swine lineage of influenza (which served as a control as it should not be under pressure to evolve from human T-cells). I found that influenza nucleoprotein accumulates more substitutions in epitope regions in the human lineage. Secondly, my analysis of the phylogeny of both the human and swine lineage of influenza found that that mutations in epitope regions are more likely to persist than to die off for human influenza. This finding indicated that mutations in epitope regions targeted by human T-cells confer a fitness advantage in the human lineage of influenza. This is the first study that found statistical support that the role of CD8 T-cells is strong enough to drive influenza evolution. This knowledge deepens our understanding of how the immune system shapes influenza evolution and this information could be used to improve predictive modeling of influenza and inform vaccine development.

In Chapter 3, I examined the protein coding capacity of influenza as influenza has to generate enough protein products to successfully complete the viral life cycle yet limit

production of deleterious protein products such as T-cell epitopes. This work was motivated by the initial observation that there is selection against putative alternate initiation of CTG codons in mammalian lineages of influenza. I experimentally surveyed translation initiation sites in the influenza genome by performing ribosome profiling of influenza virus infected cells.

Surprisingly, I found little support for CTG initiation in the influenza genome. There are two explanations for this observation: either the experimental design did not capture initiation at these CTG codons, or the bioinformatic analysis is confounded by overlapping selective forces. With regards to experimental design, the cells, virus, or ribosome profiling method could fail to elicit or capture CTG initiation in the influenza genome. For example, CTG-initiation could be refractory to capture by the translation initiation inhibitor used in these experiments [191]. Furthermore, CTG initiation could occur at a level below the detection threshold of these assays, while still being functionally significant. The second explanation for the discrepancy between the evolutionary data and the experimental data on CTG initiation is that the evolutionary signal is due to a different evolutionary pressure. If the evolutionary signal is due to translation initiation, I would have expected the signal to be observed primarily near the beginning of transcripts, however I did not observe that to be the case - the selection occurred along the length of transcripts. Furthermore, initiation at CTG codons (as observed in ribosome profiling studies) occurs infrequently downstream of the annotated translation initiation site [94, 121, 198]. It is interesting to hypothesize what other factors could account for this selection. There are other evolutionary signatures with unclear origins in influenza as it adapts to mammalian hosts, such as the decrease in GC content of influenza genome decreases during adaptation to mammalian hosts [222]. The signature against CTG codons could be related to translation dynamics [105], as it is also noted that influenza evolves to a codon usage bias that is mismatched with the mammalian host [222]. The evolutionary signature could also be due to evasion from anti-viral factors [73], or other constraints to viral fitness in the mammalian host. Regardless, this analysis provides insight into the complexity of the

evolutionary forces influencing influenza evolution.

However our ribosome profiling experiments uncovered novel candidate initiation sites in the influenza genome. In addition to the canonical protein products, there are a small number of alternate translation initiation sites across multiple influenza segments. I found that the majority of these alternately initiated products are generated through use of ATG start codons. One of the alternate initiation sites generates a truncated form of influenza NA that is broadly conserved across N1 NAs. This N-terminally truncated NA is enzymatically active and capable of supporting viral growth. Together these results highlight the diversity of protein products produced by influenza.

Below I propose further experiments based on the research contained in this thesis. The first direction is related to enhancing our insight into the selective pressures of T-cells on influenza evolution by comprehensively mapping mutations in influenza that confer T-cell escape. The second direction is related to further understanding how immune pressures act to constrain both viral and host evolution at the level of protein translation.

4.1 Massively parallel interrogation of T-cell escape mutations in influenza

Having established that T-cells are a selective force that drives influenza evolution, it would be exciting to survey influenza for T-cell escape mutations in a high through-put manner. Traditional assays that discovered de novo mutations in influenza that confer T-cell escape were performed by infecting an immunocompromised mouse with wild-type virus [160], or by sampling an immunocompromised individual over time [208], and analyzing any resulting mutations that arose in the virus. In each case, one or a few escape mutations were discovered. However, it is now possible to perform experiments with massively parallel measurements of the effects of protein mutations using deep mutational scanning [59, 60]. For deep mutational scanning in the context of a virus, a large library of the viral gene carrying mutations is generated, mutant virus libraries are reconstituted, the viral libraries are subjected to selective pressures, and the frequencies of all mutations before and after se-

lection are quantified by deep sequencing [203, 50]. Deep mutational scanning has been successfully employed to interrogate viral mutations that confer escape from neutralizing antibodies [50, 47].

To apply deep mutational scanning to identify T-cell escape mutations in influenza, I will first map escape mutations to a single T-cell receptor (TCR) for an immunodominant epitope in influenza. I will generate a mutant virus library for influenza, infect the mutant virus library into cells capable of generating and presenting viral antigen, and then co-incubate the infected cells with a primary clonal T-cell line containing a TCR specific to the influenza epitope. The T-cell line will lytically kill cells containing viruses that generate epitopes that do not escape T-cell recognition. Cells containing viruses that escape T-cell recognition will not be targeted by T-cells, so those viruses will rise in frequency in comparison to the initial input library. Both initial and selected viral libraries will be sequenced and T-cell escape mutations will be identified by computing frequencies of all amino acids preceding and following selection, and subsetting on positions that rise substantially in frequency following selection.

As an initial validation of the experimental approach, I will perform this assay for an epitope that has been previously shown to T-cell escape in the literature. I will perform this assay with influenza nucleoprotein, which I demonstrated is under positive selection by CD8 T-cells and has previously described T-cell escape mutations such as R384G [213]. I will use the T-cell receptor that was employed to characterize escape from R384G [213]. This will ensure that there is a positive control built into the assay in that I should see the previously identified escape mutation R384G in the assay.

From these results, I can address whether there is high functional constraint in epitope regions, and whether escape mutations are generally deleterious to viral fitness. Standard metrics for selection have failed to find evidence of selection for CD8 T-cells on influenza, suggesting that the regions that CD8 T-cell epitopes reside may be under strong functional constraint. There is some experimental evidence that mutations that confer T-cell escape in influenza come at a fitness cost [14, 7, 167, 72]. I can address whether epitope sites

are generally constrained by analyzing the pre-selected frequencies of sites that allow escape and seeing if they are generally mutationally tolerant to allowing many amino acids (suggesting that there isn't functional constraint at that site) or if they only allow one or a few amino acids (suggesting strong functional constraint). The fitness of amino acids that confer escape can also be assessed - if the escape mutations are strongly disfavored compared to wild-type, this would provide additional support that these sites are under strong functional constraint and escape occurs at the cost of general viral fitness.

Once validated, this approach can be expanded to other T-cell receptors that target the same epitope to see if there are different routes accessible to escape. For example, one might expect that mutations in TCR contact residues will have varying impacts for different TCRs. This approach can also be extended to other epitopes. This research direction can provide insight into whether there are epitopes that have an easier time escaping T-cell immunity. This is relevant to vaccine development as there is much interest in generating a universal vaccine in part through eliciting a broadly protective T-cell response. This knowledge has the potential to direct efforts towards eliciting responses to epitopes which are more constrained.

4.2 Examine correlates of translation efficiency during activation of the anti-viral response

Translation of viral proteins is required for completion of the viral replication cycle. Therefore, host cells regulate translation during the anti-viral response that arise during viral infection. A classic example of the global regulation of translation that occurs during the anti-viral response is activation of protein kinase R (PKR) [65]. PKR becomes activated upon binding its ligand dsRNA and subsequently phosphorylates eIF2α [65]. eIF2α phosphorylation results in a global inhibition of translation initiation. Despite global decreases in translation initiation, some transcripts are selectively translationally upregulated [65]. A number of other anti-viral proteins have been shown to inhibit translation including: interferon-induced proteins with tetratricopeptide repeats 1(IFIT1) protein [122], stimula-

tor of IFN genes (STING) [61], and zinc-finger anti-viral protein (ZAP) [63, 16, 199].

ZAP inhibits the translation of a broad range of RNA viruses including HIV, influenza [199], Sindbis virus and Ross River virus [16], Ebola virus and Marburg virus [138], murine leukemia virus [63] and hepatitis B [132]. Each virus generally has a small number of transcripts that are sensitive to ZAP. ZAP inhibits the translation of the influenza proteins NA, PA, and PB2 [199]. The mechanism of ZAP activity includes binding to viral mRNA [76], inhibiting translation initiation [236] and promoting mRNA degradation [235]. Furthermore, the skew of ZAP activity towards modulating mRNA stability or translation can be context and transcript dependent.

There is evidence that dinucleotide usage is correlated with and impacts viral fitness. There is selection against CpG dinucleotides in influenza and other RNA viruses of mammalian hosts [73, 74]. It was recently shown that HIV viruses that have high CpG dinucleotide content are less fit than HIV viruses that have low CpG content in a ZAP dependent manner [196]. This fitness defect correlated with decreased mRNA levels of spliced HIV transcripts. Interrogation of ZAP binding sites using CLIP-Seq found that ZAP binds to regions of HIV transcripts that are enriched with CpG dinucleotides. It is unknown whether the selective pressure that results in depletion of CpG dinucleotides in influenza is due to ZAP.

In addition to selection against CpG dinucleotides in RNA viruses, there is also evidence of selection against CpG dinucleotides in host transcripts. Specifically there is depletion of CpG dinucleotides in human and mouse anti-viral transcripts [74]. It is reasonable to hypothesize that the selective pressure may be due to the unique conditions present during the anti-viral state. Furthermore, the majority of reads from the ZAP CLIP-Seq experiment in [196] were derived from host transcripts, suggesting that ZAP binds host transcripts. Given the signature of selection against CpG dinucleotides in anti-viral transcripts, the ability of ZAP to bind host transcripts, and the ability of ZAP to modulate translation, I hypothesize that ZAP can alter translation of host transcripts at a global level.

To examine whether ZAP alters translation of host transcripts, I will examine the IFN

treated and control samples of my ribosome profiling data to see if CpG dinucleotide content correlates with efficiency of translation during the anti-viral response at the global level. Specifically I hypothesize that transcripts enriched with CpG dinucleotides will be translated with low efficiency during activation of the anti-viral response.

To validate whether ZAP alters translation in a CpG dependent manner I propose making ZAP -/- cells and ZAP overexpressing cells. I will then use a fluorescent reporter constructs that differ only in CpG content (a high and low CpG fluorescent construct) to assay whether the high CpG construct is translated less well during IFN treatment in a ZAP dependent manner. This reporter based approach will provide orthogonal support that ZAP can alter translation in a CpG dependent manner.

In order to assess whether ZAP globally alters translation in a CpG dependent manner, I will perform ribosome profiling of ZAP -/- and ZAP overexpressing cells. If transcripts enriched with CpG motifs are poorly translated in the ZAP overexpressing cells, this finding could be a plausible explanation for why anti-viral genes are depleted of CpG motifs. Anti-viral transcripts need to be translated efficiently in the context of viral infection, during which ZAP is present.

Lastly, I will examine whether ZAP is the selective force that results in decreased CpG content of influenza that infects mammalian hosts. I will perform competition experiments between viruses that differ in CpG content in the presence and absence of ZAP. To do this, I will use a low CpG virus and a high CpG virus. I will co-infect the two viral variants into ZAP -/- and ZAP overexpressing cells and collect cellular RNA at several time points and perform targeted deep sequencing of the samples to identify relative amounts of each viral genotype. I expect the low CpG virus to increase in the ZAP overexpressing cells in comparison to the ZAP -/- cells.

Overall this will provide insight into how translation is altered during the anti-viral response and potentially further our mechanistic understanding of the selective forces that drive both human and viral evolution.

BIBLIOGRAPHY

[1] Ramesh K Akkina, James C Richardson, Marcela C Aguilera, and Yang Chi-Ming. Heterogeneous forms of polymerase proteins exist in influenza a virus-infected cells. *Virus research*, 19(1):17–30, 1991.

[2] RK Akkina. Antigenic reactivity and electrophoretic migrational heterogeneity of the three polymerase proteins of type a human and animal influenza viruses. *Archives of virology*, 111(3-4):187–197, 1990.

[3] Todd M Allen, Marcus Altfeld, G Yu Xu, Kristin M O'Sullivan, Mathias Lichterfeld, Sylvie Le Gall, Mina John, Bianca R Mothe, Paul K Lee, Elizabeth T Kalife, et al. Selection, transmission, and reversion of an antigen-processing cytotoxic t-lymphocyte escape mutation in human immunodeficiency virus type 1 infection. *Journal of virology*, 78(13):7069–7078, 2004.

[4] Dmitry E Andreev, Patrick BF O'Connor, Ciara Fahey, Elaine M Kenny, Ilya M Terenin, Sergey E Dmitriev, Paul Cormican, Derek W Morris, Ivan N Shatsky, and Pavel V Baranov. Translation of 5 leaders is pervasive in genes resistant to eif2 repression. *Elife*, 4, 2015.

[5] Carlo G Artieri and Hunter B Fraser. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome research*, 24(12):2011–2021, 2014.

[6] Orr Ashenberg, Jai Padmakumar, Michael B Doud, and Jesse D Bloom. Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by mxa. *PLoS pathogens*, 13(3):e1006288, 2017.

[7] Erika Assarsson, Huynh-Hoa Bui, John Sidney, Qing Zhang, Jean Glenn, Carla Oseroff, Innocent N Mbawuike, Jeff Alexander, Mark J Newman, Howard Grey, et al. Immunomic analysis of the repertoire of t-cell specificities for influenza a virus in humans. *Journal of virology*, 82(24):12241–12251, 2008.

[8] Yiming Bao, Pavel Bolotov, Dmitry Dernovoy, Boris Kiryutin, Leonid Zaslavsky, Tatiana Tatusova, Jim Ostell, and David Lipman. The influenza virus resource at the national center for biotechnology information. *Journal of virology*, 82(2):596–601, 2008.

[9] Subrata Barman, Lopa Adhikary, Alok K Chakrabarti, Carl Bernas, Yoshihiro Kawaoka, and Debi P Nayak. Role of transmembrane domain and cytoplasmic tail amino acid sequences of influenza a virus neuraminidase in raft association and virus budding. *Journal of virology*, 78(10):5258–5269, 2004.

[10] WJ Bean, Mo Schell, J Katz, Y Kawaoka, C Naeve, O Gorman, and RG Webster. Evolution of the h3 influenza virus hemagglutinin from human and nonhuman hosts. *Journal of virology*, 66(2):1129–1138, 1992.

[11] Trevor Bedford, Andrew Rambaut, and Mercedes Pascual. Canalization of the evolutionary trajectory of the human influenza virus. *BMC biology*, 10(1):38, 2012.

[12] Trevor Bedford, Marc A Suchard, Philippe Lemey, Gytis Dudas, Victoria Gregory, Alan J Hay, John W McCauley, Colin A Russell, Derek J Smith, and Andrew Rambaut. Integrating influenza antigenic dynamics with molecular evolution. *Elife*, 3, 2014.

[13] Adi Bercovich-Kinori, Julie Tai, Idit Anna Gelbart, Alina Shitrit, Shani Ben-Moshe, Yaron Drori, Shalev Itzkovitz, Michal Mandelboim, and Noam Stern-Ginossar. A systematic view on influenza induced host shutoff. *Elife*, 5:e18311, 2016.

[14] EGM Berkhoff, E De Wit, MM Geelhoed-Mieras, ACM Boon, J Symons, RAM Fouchier, ADME Osterhaus, and GF Rimmelzwaan. Functional constraints of influenza a virus epitopes limit escape from cytotoxic t lymphocytes. *Journal of virology*, 79(17):11239–11246, 2005.

[15] EGM Berkhoff, MM Geelhoed-Mieras, RAM Fouchier, ADME Osterhaus, and GF Rimmelzwaan. Assessment of the extent of variation in influenza a virus cytotoxic t-lymphocyte epitopes by using virus-specific cd8+ t-cell clones. *Journal of General Virology*, 88(2):530–535, 2007.

[16] Matthew J Bick, John-William N Carroll, Guangxia Gao, Stephen P Goff, Charles M Rice, and Margaret R MacDonald. Expression of the zinc-finger antiviral protein inhibits alphavirus replication. *Journal of virology*, 77(21):11555–11562, 2003.

[17] Jesse D Bloom. Software for the analysis and visualization of deep mutational scanning data. *BMC bioinformatics*, 16(1):168, 2015.

[18] Jesse D Bloom, Lizhi Ian Gong, and David Baltimore. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science*, 328(5983):1272–1275, 2010.

[19] Jesse D Bloom, Jagannath S Nayak, and David Baltimore. A computational-experimental approach identifies mutations that enhance surface expression of an oseltamivir-resistant influenza neuraminidase. *PLoS One*, 6(7):e22201, 2011.

[20] Helen C Bodmer, Roy M Pemberton, Jonathan B Rothbard, and Brigitte A Askonas. Enhanced recognition of a modified peptide antigen by cytotoxic t cells specific for influenza nucleoprotein. *Cell*, 52(2):253–258, 1988.

[21] ACM Boon, Gerrie de Mutsert, YMF Graus, RAM Fouchier, Krijn Sintnicolaas, ADME Osterhaus, and GF Rimmelzwaan. The magnitude and specificity of in-

fluenza a virus-specific cytotoxic t-lymphocyte responses in humans is related to hla-a and-b phenotype. *Journal of virology*, 76(2):582–590, 2002.

[22] ACM Boon, Gerrie de Mutsert, YMF Graus, RAM Fouchier, Krijn Sintnicolaas, ADME Osterhaus, and GF Rimmelzwaan. Sequence variation in a newly identified hla-b35-restricted epitope in the influenza a virus nucleoprotein associated with escape from cytotoxic t lymphocytes. *Journal of virology*, 76(5):2567–2572, 2002.

[23] Adrianus CM Boon, Gerrie de Mutsert, Debbie van Baarle, Derek J Smith, Alan S Lapedes, Ron AM Fouchier, Kees Sintnicolaas, Albert DME Osterhaus, and Guus F Rimmelzwaan. Recognition of homo-and heterosubtypic variants of influenza a viruses by human cd8+ t lymphocytes. *The Journal of Immunology*, 172(4):2453–2460, 2004.

[24] Thierry Boon and Aline Van Pel. T cell-recognized antigenic peptides derived from the cellular genome are not protein degradation products but can be generated directly by transcription and translation of short subgenic regions. a hypothesis. *Immunogenetics*, 29(2):75–79, 1989.

[25] Timothy J Bos, Alan R Davis, and Debi P Nayak. Nh2-terminal hydrophobic region of influenza virus neuraminidase provides the signal function in translocation. *Proceedings of the National Academy of Sciences*, 81(8):2327–2331, 1984.

[26] GERALD W Both, MJ Sleigh, NJ Cox, and AP Kendal. Antigenic drift in influenza virus h3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential amino acid changes at key antigenic sites. *Journal of virology*, 48(1):52–60, 1983.

[27] Gloria A Brar and Jonathan S Weissman. Ribosome profiling reveals the what,

when, where and how of protein synthesis. *Nature reviews Molecular cell biology*, 16(11):651, 2015.

[28] Christy Brockwell-Staats, Robert G Webster, and Richard J Webby. Diversity of influenza viruses in swine and the emergence of a novel human pandemic influenza a (h1n1). *Influenza and other respiratory viruses*, 3(5):207–213, 2009.

[29] Christopher B Brooke. Biological activities of'noninfectious' influenza a virus particles. *Future virology*, 9(1):41–51, 2014.

[30] Deborah M Brown, Eulogia Román, and Susan L Swain. Cd4 t cell responses to influenza infection. In *Seminars in immunology*, volume 16, pages 171–177. Elsevier, 2004.

[31] Donald J Brown, BRENDA G Hogue, and DEBI P Nayak. Redundancy of signal and anchor functions in the nh2-terminal uncharged region of influenza virus neuraminidase, a class ii membrane glycoprotein. *Journal of virology*, 62(10):3824–3831, 1988.

[32] Robin M Bush, Walter M Fitch, Catherine A Bender, and Nancy J Cox. Positive selection on the h3 hemagglutinin gene of human influenza virus a. *Molecular biology and evolution*, 16(11):1457–1465, 1999.

[33] Jeff Butler, Kathryn A Hooper, Stephen Petrie, Raphael Lee, Sebastian Maurer-Stroh, Lucia Reh, Teagan Guarnaccia, Chantal Baas, Lumin Xue, Sophie Vitesnik, et al. Estimating the fitness advantage conferred by permissive neuraminidase mutations in recent oseltamivir-resistant a (h1n1) pdm09 influenza viruses. *PLoS pathogens*, 10(4):e1004065, 2014.

[34] Sylvain Cardinaud, Arnaud Moris, Michèle Février, Pierre-Simon Rohrlich, Laurence Weiss, Pierre Langlade-Demoyen, François A Lemonnier, Olivier Schwartz,

and André Habel. Identification of cryptic mhc i–restricted epitopes encoded by hiv-1 alternative reading frames. *Journal of Experimental Medicine*, 199(8):1053–1063, 2004.

[35] Weisan Chen, Paul A Calvo, Daniela Malide, James Gibbs, Ulrich Schubert, Igor Bacik, Sameh Basta, Robert O'Neill, Jeanne Schickli, Peter Palese, et al. A novel influenza a virus mitochondrial protein that induces cell death. *Nature medicine*, 7(12):1306–1312, 2001.

[36] ML Clements, RF Betts, EL Tierney, and BR Murphy. Serum and nasal wash antibodies associated with resistance to experimental challenge with influenza a wild-type virus. *Journal of clinical microbiology*, 24(1):157–160, 1986.

[37] Monica Clifford, James Twigg, and Chris Upton. Evidence for a novel gene associated with human influenza a viruses. *Virology journal*, 6(1):198, 2009.

[38] Rocío Coloma, José M Valpuesta, Rocío Arranz, José L Carrascosa, Juan Ortín, and Jaime Martín-Benito. The structure of a biologically active influenza virus ribonucleoprotein complex. *PLoS pathogens*, 5(6):e1000491, 2009.

[39] Robert B Couch and Julius A Kasel. Immunity to influenza in man. *Annual Reviews in Microbiology*, 37(1):529–549, 1983.

[40] NJ Cox and K Subbarao. Global epidemiology of influenza: past and present. *Annual review of medicine*, 51(1):407–421, 2000.

[41] Diogo V da Silva, Johan Nordholm, Dan Dou, Hao Wang, Jeremy S Rossman, and Robert Daniels. The influenza virus neuraminidase protein transmembrane and head domains have coevolved. *Journal of virology*, 89(2):1094–1104, 2015.

[42] Alan R Davis, Alan L Hiti, and Debi P Nayak. Influenza defective interfering viral rna is formed by internal deletion of genomic rna. *Proceedings of the National Academy of Sciences*, 77(1):215–219, 1980.

[43] Wayne Delport, Art FY Poon, Simon DW Frost, and Sergei L Kosakovsky Pond. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics*, 26(19):2455–2457, 2010.

[44] Yuping Deng, Jonathan W Yewdell, Laurence C Eisenlohr, and Jack R Bennink. Mhc affinity, peptide liberation, t cell repertoire, and immunodominance all contribute to the paucity of mhc class i-restricted peptides recognized by antiviral ctl. *The Journal of Immunology*, 158(4):1507–1515, 1997.

[45] Alexandre Dias, Denis Bouvier, Thibaut Crépin, Andrew A McCarthy, Darren J Hart, Florence Baudin, Stephen Cusack, and Rob WH Ruigrok. The cap-snatching endonuclease of influenza virus polymerase resides in the pa subunit. *Nature*, 458(7240):914, 2009.

[46] Alexander J Diaz de Arce, William L Noderer, and Clifford L Wang. Complete motif analysis of sequence requirements for translation initiation at non-aug start codons. *Nucleic acids research*, 2017.

[47] Adam S Dingens, Hugh K Haddox, Julie Overbaugh, and Jesse D Bloom. Comprehensive mapping of hiv-1 escape from a broadly neutralizing antibody. *Cell host & microbe*, 21(6):777–787, 2017.

[48] Harry Dolstra, Hanny Fredrix, Frans Maas, Pierre G Coulie, Francis Brasseur, Ewald Mensink, Gosse J Adema, Theo M De Witte, Carl G Figdor, and Elly Van De Wiel-van Kemenade. A human minor histocompatibility antigen specific for b cell acute lymphoblastic leukemia. *Journal of Experimental Medicine*, 189(2):301–308, 1999.

[49] Mario Dos Reis, Alan J Hay, and Richard A Goldstein. Using non-homogeneous models of nucleotide substitution to identify host shift events: application to the origin of the 1918 “spanish” influenza pandemic virus. *Journal of molecular evolution*, 69(4):333, 2009.

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

[51] Alexei J Drummond, Marc A Suchard, Dong Xie, and Andrew Rambaut. Bayesian phylogenetics with beauti and the beast 1.7. *Molecular biology and evolution*, 29(8):1969–1973, 2012.

[52] PH Duesberg. Distinct subunits of the ribonucleoprotein of influenza virus. *Journal of molecular biology*, 42(3):485–499, 1969.

[53] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[54] Debra Elton, Martha Simpson-Holley, Kate Archer, Liz Medcalf, Roger Hallam, John McCauley, and Paul Digard. Interaction of the influenza virus nucleoprotein with the cellular crm1-mediated nuclear export pathway. *Journal of virology*, 75(1):408–419, 2001.

[55] Joseph Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15, 1985.

[56] BN Fields, DM Knipe, PM Howley, and DE Griffin. Fields virology. 5th, 2007.

[57] E Fodor. The rna polymerase of influenza a virus: mechanisms of viral transcription and replication. *Acta virologica*, 57(2):113–122, 2013.

[58] Ervin Fodor, David C Pritlove, and George G Brownlee. The influenza virus panhandle is involved in the initiation of transcription. *Journal of virology*, 68(6):4092–4096, 1994.

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

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

[61] Kate M Franz, William J Neidermyer, Yee-Joo Tan, Sean PJ Whelan, and Jonathan C Kagan. Sting-dependent translation inhibition restricts rna virus replication. *Proceedings of the National Academy of Sciences*, 115(9):E2058–E2067, 2018.

[62] Claudia Fritsch, Alexander Herrmann, Michael Nothnagel, Karol Szafranski, Klaus Huse, Frank Schumann, Stefan Schreiber, Matthias Platzer, Michael Krawczak, Jochen Hampe, et al. Genome-wide search for novel human uorfs and n-terminal protein extensions using ribosomal footprinting. *Genome research*, 22(11):2208–2218, 2012.

[63] Guangxia Gao, Xuemin Guo, and Stephen P Goff. Inhibition of retroviral rna production by zap, a ccch-type zinc finger protein. *Science*, 297(5587):1703–1706, 2002.

[64] Xiangwei Gao, Ji Wan, Botao Liu, Ming Ma, Ben Shen, and Shu-Bing Qian. Quantitative profiling of initiating ribosomes in vivo. *Nature methods*, 12(2):147–153, 2015.

[65] MA Garcia, J Gil, I Ventoso, S Guerra, E Domingo, C Rivas, and M Esteban. Impact of protein kinase pkr in cell biology: from antiviral to antiproliferative action. *Microbiology and Molecular Biology Reviews*, 70(4):1032–1060, 2006.

[66] Rebecca J Garten, C Todd Davis, Colin A Russell, Bo Shu, Stephen Lindstrom, Amanda Balish, Wendy M Sessions, Xiyan Xu, Eugene Skepner, Varough Deyde, et al. Antigenic and genetic characteristics of swine-origin 2009 a (h1n1) influenza viruses circulating in humans. *science*, 325(5937):197–201, 2009.

[67] Maxim V Gerashchenko and Vadim N Gladyshev. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic acids research*, 42(17):e134–e134, 2014.

[68] Maxim V. Gerashchenko and Vadim N. Gladyshev. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucl. Acids Res.*, 42(17):e134–e134, September 2014.

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

[70] Julia R Gog, Guus F Rimmelzwaan, Albert DME Osterhaus, and Bryan T Grenfell. Population dynamics of rapid fixation in cytotoxic t lymphocyte escape mutants of influenza a. *Proceedings of the National Academy of Sciences*, 100(19):11143–11147, 2003.

[71] Lizhi Ian Gong and Jesse D Bloom. Epistatically interacting substitutions are enriched during adaptive protein evolution. *PLoS genetics*, 10(5):e1004328, 2014.

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

[73] Benjamin D Greenbaum, Arnold J Levine, Gyan Bhanot, and Raul Rabadan. Patterns of evolution and host gene mimicry in influenza and other rna viruses. *PLoS pathogens*, 4(6):e1000079, 2008.

[74] Benjamin D Greenbaum, Raul Rabadan, and Arnold J Levine. Patterns of oligonucleotide sequences in viral and host cell rna identify mediators of the host innate immune system. *PloS one*, 4(6):e5969, 2009.

[75] Hailong Guo, Félix Santiago, Kris Lambert, Toru Takimoto, and David J Topham. T cell-mediated protection against lethal 2009 pandemic h1n1 influenza virus infection in a mouse model. *Journal of virology*, 85(1):448–455, 2011.

[76] Xuemin Guo, John-William N Carroll, Margaret R MacDonald, Stephen P Goff, and Guangxia Gao. The zinc finger antiviral protein directly binds to specific viral mrnas through the ccch zinc finger motifs. *Journal of virology*, 78(23):12781–12787, 2004.

[77] Benjamin G Hale, Richard E Randall, Juan Ortín, and David Jackson. The multifunctional ns1 protein of influenza a viruses. *Journal of general virology*, 89(10):2359–2376, 2008.

[78] Eriko Hatada, Masakazu Hasegawa, Jun Mukaigawa, Kazufumi Shimizu, and Ryuji Fukuda. Control of influenza virus gene expression: quantitative analysis of each viral rna species in infected cells. *The Journal of Biochemistry*, 105(4):537–546, 1989.

[79] Andrew C Hayward, Lili Wang, Nilu Goonetilleke, Ellen B Fragaszy, Alison Bermingham, Andrew Copas, Oliver Dukes, Elizabeth RC Millett, Irwin Nazareth, Jonathan S Nguyen-Van-Tam, et al. Natural t cell–mediated protection against seasonal and pandemic influenza. results of the flu watch cohort study. *American journal of respiratory and critical care medicine*, 191(12):1422–1431, 2015.

[80] Heather Hickman, Jacqueline Mays, James Gibbs, Ivan Kosik, Javier Magadan, Kazuyo Takeda, Suman Das, Glennys Reynoso, Barbara Ngudiankama, JiaJie Wei, et al. Influenza a virus negative strand rna is translated for cd8+ t cell immunosurveillance. *bioRxiv*, page 261966, 2018.

[81] Sarah K Hilton, Michael B Doud, and Jesse D Bloom. phydms: software for phylogenetic analyses informed by deep mutational scanning. *PeerJ*, 5:e3657, 2017.

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

[83] Brenda G Hogue and Debi P Nayak. Deletion mutation in the signal anchor domain activates cleavage of the influenza virus neuraminidase, a type ii transmembrane protein. *Journal of general virology*, 75(5):1015–1022, 1994.

[84] Kathryn A Hooper and Jesse D Bloom. A mutant influenza virus that uses an n1 neuraminidase as the receptor-binding protein. *Journal of virology*, 87(23):12531–12540, 2013.

[85] Alice S Huang and David Baltimore. Defective viral particles and viral disease processes. *Nature*, 226(5243):325, 1970.

[86] I-Chueh Huang, Wenhui Li, Jianhua Sui, Wayne Marasco, Hyeryun Choe, and Michael Farzan. Influenza a virus neuraminidase limits viral superinfection. *Journal of virology*, 82(10):4834–4843, 2008.

[87] Edward C Hutchinson, Johann C von Kirchbach, Julia R Gog, and Paul Digard. Genome packaging in influenza a virus. *Journal of general virology*, 91(2):313–328, 2010.

[88] SC Inglis and CM Brown. Differences in the control of virus mrna splicing during permissive or abortive infection with influenza a (fowl plague) virus. *Journal of general virology*, 65(1):153–164, 1984.

[89] Stephen C Inglis and Carol M Brown. Spliced and unspliced rnas encoded by virion rna segment 7 of influenza virus. *Nucleic acids research*, 9(12):2727–2740, 1981.

[90] Nicholas T Ingolia, Gloria A Brar, Silvia Rouskin, Anna M McGeachy, and Jonathan S Weissman. The ribosome profiling strategy for monitoring translation in

vivo by deep sequencing of ribosome-protected mrna fragments. *Nature protocols*, 7(8):1534, 2012.

[91] Nicholas T. Ingolia, Gloria A. Brar, Noam Stern-Ginossar, Michael S. Harris, Gaëlle J. S. Talhouarne, Sarah E. Jackson, Mark R. Wills, and Jonathan S. Weissman. Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.*, 8(5):1365–1379, September 2014.

[92] Nicholas T Ingolia, Gloria A Brar, Noam Stern-Ginossar, Michael S Harris, Gaëlle JS Talhouarne, Sarah E Jackson, Mark R Wills, and Jonathan S Weissman. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell reports*, 8(5):1365–1379, 2014.

[93] Nicholas T Ingolia, Sina Ghaemmaghami, John RS Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science*, 324(5924):218–223, 2009.

[94] Nicholas T Ingolia, Liana F Lareau, and Jonathan S Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802, 2011.

[95] Ivaylo P. Ivanov, Byung-Sik Shin, Gary Loughran, Ioanna Tzani, Sara K. Young-Baird, Chune Cao, John F. Atkins, and Thomas E. Dever. Polyamine Control of Translation Elongation Regulates Start Site Selection on Antizyme Inhibitor mRNA via Ribosome Queuing. *Molecular Cell*, 70(2):254–264.e6, April 2018.

[96] BW Jagger, HM Wise, JC Kash, K-A Walters, NM Wills, Y-L Xiao, RL Dunfee, LM Schwartzman, A Ozinsky, GL Bell, et al. An overlapping protein-coding region in influenza a virus segment 3 modulates the host response. *Science*, 337(6091):199–204, 2012.

[97] J Michael Janda, Alan R Davis, Debi P Nayak, and Barun K De. Diversity and generation of defective interfering influenza virus particles. *Virology*, 95(1):48–58, 1979.

[98] Sinthujan Jegaskanda, Emma R Job, Marit Kramski, Karen Laurie, Gamze Isitman, Robert de Rose, Wendy R Winnall, Ivan Stratov, Andrew G Brooks, Patrick C Reading, et al. Cross-reactive influenza-specific antibody-dependent cellular cytotoxicity antibodies in the absence of neutralizing antibodies. *The Journal of Immunology*, 190(4):1837–1848, 2013.

[99] David T Jones, William R Taylor, and Janet M Thornton. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3):275–282, 1992.

[100] Derin B Keskin, Bruce B Reinhold, Guang Lan Zhang, Alexander R Ivanov, Barry L Karger, and Ellis L Reinherz. Physical detection of influenza a epitopes identifies a stealth subset on human lung epithelium evading natural cd8 immunity. *Proceedings of the National Academy of Sciences*, 112(7):2151–2156, 2015.

[101] Derin B Keskin, Bruce R Reinhold, Guang Lan Zhang, Alexander R Ivanov, Barry L Karger, and Ellis L Reinherz. Reply to van de sandt and rimmelzwaan: Matching epitope display with functional avidity. *Proceedings of the National Academy of Sciences*, 112(19):E2418–E2418, 2015.

[102] Denys A Khaperskyy, Summer Schmaling, Jonah Larkins-Ford, Craig McCormick, and Marta M Gaglia. Selective degradation of host rna polymerase ii transcripts by influenza a virus pa-x host shutoff protein. *PLoS pathogens*, 12(2):e1005427, 2016.

[103] Hiroshi Kida, Yoshihiro Kawaoka, Clayton W Naeve, and Robert G Webster. Anti-

genic and genetic conservation of h3 influenza virus in wild ducks. *Virology*, 159(1):109–119, 1987.

[104] M Knossow, M Gaudier, A Douglas, B Barrere, T Bizebard, C Barbey, B Gigant, and JJ Skehel. Mechanism of neutralization of influenza virus infectivity by antibodies. *Virology*, 302(2):294–298, 2002.

[105] Anton A Komar. A pause for thought along the co-translational folding pathway. *Trends in biochemical sciences*, 34(1):16–24, 2009.

[106] Sergei L Kosakovsky Pond and Simon DW Frost. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution*, 22(5):1208–1222, 2005.

[107] Sergei L Kosakovsky Pond, Art FY Poon, Andrew J Leigh Brown, and Simon DW Frost. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza a virus. *Molecular biology and evolution*, 25(9):1809–1824, 2008.

[108] Marilyn Kozak. How do eucaryotic ribosomes select initiation regions in messenger rna? *Cell*, 15(4):1109–1123, 1978.

[109] Marilyn Kozak. Point mutations define a sequence flanking the aug initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, 44(2):283–292, 1986.

[110] Michael Krasnitz, Arnold J Levine, and Raul Rabadan. Anomalies in the influenza virus genome database: new biology or laboratory errors? *Journal of virology*, 82(17):8947–8950, 2008.

[111] JHCM Kreijtz, Rogier Bodewes, Geert van Amerongen, Thijs Kuiken, RAM Fouchier, ADME Osterhaus, and GF Rimmelzwaan. Primary influenza a virus infection induces cross-protective immunity against a lethal infection with a heterosubtypic virus strain in mice. *Vaccine*, 25(4):612–620, 2007.

[112] Andi Krumbholz, Anja Philipps, Hartmut Oehring, Katja Schwarzer, Annett Eitner, Peter Wutzler, and Roland Zell. Current knowledge on pb1-f2 of influenza a viruses. *Medical microbiology and immunology*, 200(2):69–75, 2011.

[113] Susann Kummer, Max Flöttmann, Björn Schwanhäusser, Christian Sieben, Michael Veit, Matthias Selbach, Edda Klipp, and Andreas Herrmann. Alteration of protein levels during influenza virus h1n1 infection in host cells: a proteomic survey of host and virus reveals differential dynamics. *PLoS One*, 9(4):e94257, 2014.

[114] A Kundu, RT Avalos, CM Sanderson, and DP Nayak. Transmembrane domain of influenza virus neuraminidase, a type ii protein, possesses an apical sorting signal in polarized mdck cells. *Journal of Virology*, 70(9):6508–6515, 1996.

[115] Brian J Laidlaw, Vilma Decman, Mohammed-Alkhatim A Ali, Michael C Abt, Amaya I Wolf, Laurel A Monticelli, Krystyna Mozdzanowska, Jill M Angelosanto, David Artis, Jan Erikson, et al. Cooperativity between cd8+ t cells, non-neutralizing antibodies, and alveolar macrophages is important for heterosubtypic influenza virus immunity. *PLoS pathogens*, 9(3):e1003207, 2013.

[116] Robert A Lamb and Ching-Juh Lai. Sequence of interrupted and uninterrupted mrnas and cloned dna coding for the two overlapping nonstructural proteins of influenza virus. *Cell*, 21(2):475–485, 1980.

[117] Robert A Lamb, Ching-Juh Lai, and Purnell W Choppin. Sequences of mrnas derived from genome rna segment 7 of influenza virus: colinear and interrupted mrnas code for overlapping proteins. *Proceedings of the National Academy of Sciences*, 78(7):4170–4174, 1981.

[118] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10:R25, March 2009.

[119] Juhye M Lee, John Huddleston, Michael B Doud, Kathryn A Hooper, Nicholas C Wu, Trevor Bedford, and Jesse D Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. *bioRxiv*, page 298364, 2018.

[120] Nara Lee, Valerie Le Sage, Adalena V Nanni, Dan J Snyder, Vaughn S Cooper, and Seema S Lakdawala. Genome-wide analysis of influenza viral rna and nucleoprotein association. *Nucleic acids research*, 45(15):8968–8977, 2017.

[121] Sooncheol Lee, Botao Liu, Soohyun Lee, Sheng-Xiong Huang, Ben Shen, and Shu-Bing Qian. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, 109(37):E2424–E2432, 2012.

[122] Wing Yiu Lee, Chen Liang, and Richard Sloan. Ifitm proteins inhibit hiv-1 protein synthesis. *bioRxiv*, page 201525, 2018.

[123] Philippe Lemey, Vladimir N Minin, Filip Bielejec, Sergei L Kosakovsky Pond, and Marc A Suchard. A counting renaissance: combining stochastic mapping and empirical bayes to quickly detect amino acid sites under positive selection. *Bioinformatics*, 28(24):3248–3256, 2012.

[124] George P Leser and Robert A Lamb. Influenza virus assembly and budding in raft-derived microdomains: a quantitative analysis of the surface distribution of HA, NA and M2 proteins. *Virology*, 342(2):215–227, 2005.

[125] Bo Li and Colin N. Dewey. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323, August 2011.

[126] Kifayathullah Liakath-Ali, Eric W. Mills, Inês Sequeira, Beate M. Lichtenberger, Angela Oliveira Pisco, Kalle H. Sipilä, Ajay Mishra, Harunori Yoshikawa, Colin Chih-

Chien Wu, Tony Ly, Angus I. Lamond, Ibrahim M. Adham, Rachel Green, and Fiona M. Watt. An evolutionarily conserved ribosome-rescue pathway maintains epidermal homeostasis. *Nature*, page 1, April 2018.

[127] Kristin Lienert and Peter Parham. Evolution of mhc class i genes in higher primates. *Immunology & Cell Biology*, 74(4):349–356, 1996.

[128] Sasa Lin, Hussein Y Naim, A Chapin Rodriguez, and Michael G Roth. Mutations in the middle of the transmembrane domain reverse the polarity of transport of the influenza virus hemagglutinin in mdck epithelial cells. *The Journal of cell biology*, 142(1):51–57, 1998.

[129] Chongguang Liu, Maryna C Eichelberger, Richard W Compans, and Gillian M Air. Influenza type a virus neuraminidase does not play a role in viral entry, replication, assembly, or budding. *Journal of virology*, 69(2):1099–1106, 1995.

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

[131] Heather M Machkovech, Trevor Bedford, Marc A Suchard, and Jesse D Bloom. Positive selection in cd8+ t-cell epitopes of influenza virus nucleoprotein revealed by a comparative analysis of human and swine viral lineages. *Journal of virology*, 89(22):11275–11283, 2015.

[132] Richeng Mao, Hui Nie, Dawei Cai, Jiming Zhang, Hongyan Liu, Ran Yan, Andrea Cuconati, Timothy M Block, Ju-Tao Guo, and Haitao Guo. Inhibition of hepatitis b virus replication by the host zinc finger antiviral protein. *PLoS pathogens*, 9(7):e1003494, 2013.

[133] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):pp. 10–12, May 2011.

[134] Mikhail N Matrosovich, Tatyana Y Matrosovich, Thomas Gray, Noel A Roberts, and Hans-Dieter Klenk. Neuraminidase is important for the initiation of influenza virus infection in human airway epithelium. *Journal of virology*, 78(22):12665–12667, 2004.

[135] Julie L McAuley, Felicita Hornung, Kelli L Boyd, Amber M Smith, Raelene McKeon, Jack Bennink, Jonathan W Yewdell, and Jonathan A McCullers. Expression of the 1918 influenza a virus pb1-f2 enhances the pathogenesis of viral and secondary bacterial pneumonia. *Cell host & microbe*, 2(4):240–249, 2007.

[136] Andrew J McMichael, Frances M Gotch, Gary R Noble, and Paul AS Beare. Cytotoxic t-cell immunity to influenza. *New England Journal of Medicine*, 309(1):13–17, 1983.

[137] Vladimir N Minin and Marc A Suchard. Counting labeled transitions in continuous-time markov models of evolution. *Journal of mathematical biology*, 56(3):391–412, 2008.

[138] Stefanie Müller, Peggy Möller, Matthew J Bick, Stephanie Wurr, Stephan Becker, Stephan Günther, and Beate M Kümmerer. Inhibition of filovirus replication by the zinc finger antiviral protein. *Journal of virology*, 81(5):2391–2400, 2007.

[139] Yukiko Muramoto, Takeshi Noda, Eiryo Kawakami, Ramesh Akkina, and Yoshihiro Kawaoka. Identification of novel influenza a virus proteins translated from pa mrna. *Journal of virology*, 87(5):2455–2462, 2013.

[140] Ben Murrell, Sasha Moola, Amandla Mabona, Thomas Weighill, Daniel Sheward, Sergei L Kosakovsky Pond, and Konrad Scheffler. Fubar: a fast, unconstrained bayesian approximation for inferring selection. *Molecular biology and evolution*, 30(5):1196–1205, 2013.

[141] Debi P Nayak, Rilwan A Balogun, Hiroshi Yamada, Z Hong Zhou, and Subrata Barman. Influenza virus morphogenesis and budding. *Virus research*, 143(2):147–161, 2009.

[142] Debi P Nayak, Eric Ka-Wai Hui, and Subrata Barman. Assembly and budding of influenza virus. *Virus research*, 106(2):147–165, 2004.

[143] DP Nayak and MA Jabbar. Structural domains and organizational conformation involved in the sorting and transport of influenza virus transmembrane proteins. *Annual Reviews in Microbiology*, 43(1):465–499, 1989.

[144] Richard A Neher, Colin A Russell, and Boris I Shraiman. Predicting evolution from the shape of genealogical trees. *Elife*, 3, 2014.

[145] Martha I Nelson, Lone Simonsen, Cecile Viboud, Mark A Miller, Jill Taylor, Kirsten St George, Sara B Griesemer, Elodie Ghedin, Naomi A Sengamalay, David J Spiro, et al. Stochastic processes are key determinants of short-term evolution in influenza a virus. *PLoS pathogens*, 2(12):e125, 2006.

[146] Gabriele Neumann, Tokiko Watanabe, Hiroshi Ito, Shinji Watanabe, Hideo Goto, Peng Gao, Mark Hughes, Daniel R Perez, Ruben Donis, Erich Hoffmann, et al. Generation of influenza a viruses entirely from cloned cdnas. *Proceedings of the National Academy of Sciences*, 96(16):9345–9350, 1999.

[147] Simon Noble, Martha S McGregor, David E Wentworth, and Virginia S Hinshaw. Antigenic and genetic conservation of the haemagglutinin in h1n1 swine influenza viruses. *Journal of general virology*, 74(6):1197–1200, 1993.

[148] J Nordholm, J Petitou, H Östbye, DV da Silva, D Dou, H Wang, and R Daniels. Translational regulation of viral secretory proteins by the 5'coding regions and a viral rna-binding protein. *The Journal of cell biology*, 216(8):2283, 2017.

[149] John D O'brien, Vladimir N Minin, and Marc A Suchard. Learning to count: robust estimates for labeled distances between molecular sequences. *Molecular biology and evolution*, 26(4):801–814, 2009.

[150] Michael BA Oldstone. How viruses escape from cytotoxic t lymphocytes: molecular parameters and players. *Virology*, 234(2):179–185, 1997.

[151] Robert E O'Neill, Julie Talon, and Peter Palese. The influenza virus nep (ns2 protein) mediates the nuclear export of viral ribonucleoproteins. *The EMBO journal*, 17(1):288–296, 1998.

[152] Peter Palese, Kiyotake Tobita, Masahiro Ueda, and Richard W Compans. Characterization of temperature sensitive influenza virus mutants defective in neuraminidase. *Virology*, 61(2):397–410, 1974.

[153] Colin R Parrish, Pablo R Murcia, and Edward C Holmes. Influenza virus reservoirs and intermediate hosts: dogs, horses, and new possibilities for influenza virus exposure of humans. *Journal of virology*, 89(6):2990–2994, 2015.

[154] Duncan Paterson and Ervin Fodor. Emerging roles for the influenza a virus nuclear export protein (nep). *PLoS pathogens*, 8(12):e1003019, 2012.

[155] David S Peabody. Translation initiation at non-aug triplets in mammalian cells. *Journal of Biological Chemistry*, 264(9):5031–5035, 1989.

[156] Lawrence H Pinto and Robert A Lamb. The m2 proton channels of influenza a and b viruses. *Journal of Biological Chemistry*, 281(14):8997–9000, 2006.

[157] Leo LM Poon, David C Pritlove, Jane Sharps, and George G Brownlee. The rna polymerase of influenza virus, bound to the 5? end of virion rna, acts in cis to polyadenylate mrna. *Journal of virology*, 72(10):8214–8219, 1998.

[158] M Potier, L Mameli, M Belisle, L Dallaire, and SB Melancon. Fluorometric assay of neuraminidase with a sodium (4-methylumbelliferyl-α-dn-acetylneuraminate) substrate. *Analytical biochemistry*, 94(2):287–296, 1979.

[159] Sharanya Prasad, Shelley R Starck, and Nilabh Shastri. Presentation of cryptic peptides by mhc class i is enhanced by inflammatory stimuli. *The Journal of Immunology*, 197(8):2981–2991, 2016.

[160] Graeme E Price, Rong Ou, Hong Jiang, Lei Huang, and Demetrius Moskophidis. Viral escape by selection of cytotoxic t cell–resistant variants in influenza a virus pneumonia. *Journal of Experimental Medicine*, 191(11):1853–1868, 2000.

[161] Oliver G Pybus, Andrew Rambaut, Robert Belshaw, Robert P Freckleton, Alexei J Drummond, and Edward C Holmes. Phylogenetic evidence for deleterious mutation load in rna viruses and its contribution to viral evolution. *Molecular biology and evolution*, 24(3):845–852, 2007.

[162] Raul Rabadan, Arnold J Levine, and Harlan Robins. Comparison of avian and human influenza a viruses reveals a mutational bias on the viral genomes. *Journal of virology*, 80(23):11887–11891, 2006.

[163] Andrew Rambaut, Oliver G Pybus, Martha I Nelson, Cecile Viboud, Jeffery K Taubenberger, and Edward C Holmes. The genomic and epidemiological dynamics of human influenza a virus. *Nature*, 453(7195):615, 2008.

[164] Brandon S Razooky, Benedikt Obermayer, Joshua Biggs OMay, and Alexander Tarakhovsky. Viral infection identifies micropeptides differentially regulated in smorf-containing lncrnas. *Genes*, 8(8):206, 2017.

[165] Lowell Jacob Reed and Hugo Muench. A simple method of estimating fifty per cent endpoints. *American journal of epidemiology*, 27(3):493–497, 1938.

[166] C Renard, E Hart, H Sehra, H Beasley, P Coggill, K Howe, J Harrow, J Gilbert, S Sims, J Rogers, et al. The genomic sequence and analysis of the swine major histocompatibility complex. *Genomics*, 88(1):96–110, 2006.

[167] GF Rimmelzwaan, EGM Berkhoff, NJ Nieuwkoop, RAM Fouchier, and ADME Osterhaus. Functional compensation of a detrimental amino acid substitution in a cytotoxic-t-lymphocyte epitope of influenza a viruses by comutations. *Journal of virology*, 78(16):8946–8949, 2004.

[168] GF Rimmelzwaan, ACM Boon, JTM Voeten, EGM Berkhoff, RAM Fouchier, and ADME Osterhaus. Sequence variation in the influenza a virus nucleoprotein associated with escape from cytotoxic t lymphocytes. *Virus research*, 103(1-2):97–100, 2004.

[169] Guus F Rimmelzwaan, Joost HCM Kreijtz, Rogier Bodewes, Ron AM Fouchier, and Albert DME Osterhaus. Influenza virus ctl epitopes, remarkably conserved and remarkably variable. *Vaccine*, 27(45):6363–6365, 2009.

[170] JAMES S Robertson, MANFRED Schubert, and ROBERT A Lazzarini. Polyadenylation sites for influenza virus mrna. *Journal of virology*, 38(1):157–163, 1981.

[171] Kirsten Roomp, Iris Antes, and Thomas Lengauer. Predicting mhc class i epitopes in large datasets. *Bmc Bioinformatics*, 11(1):90, 2010.

[172] Jeremy S Rossman and Robert A Lamb. Influenza virus assembly and budding. *Virology*, 411(2):229–236, 2011.

[173] Irina Rusinova, Sam Forster, Simon Yu, Anitha Kannan, Marion Masse, Helen Cumming, Ross Chapman, and Paul J. Hertzog. INTERFEROME v2.0: An updated database of annotated interferon-regulated genes. *Nucleic Acids Res*, 41(Database issue):D1040–D1046, January 2013.

[174] Alistair B Russell, Cole Trapnell, and Jesse D Bloom. Extreme heterogeneity of influenza virus infection in single cells. *eLife*, 7:e32303, 2018.

[175] Kazima Saira, Xudong Lin, Jay V DePasse, Rebecca Halpin, Alan Twaddle, Timothy Stockwell, Brian Angus, Alessandro Cozzi-Lepri, Marina Delfino, Vivien Dugan, et al. Sequence analysis of in vivo defective interfering-like rna of influenza a h1n1 pandemic virus. *Journal of virology*, 87(14):8064–8074, 2013.

[176] Tilman Schneider-Poetsch, Jianhua Ju, Daniel E Eyler, Yongjun Dang, Shridhar Bhat, William C Merrick, Rachel Green, Ben Shen, and Jun O Liu. Inhibition of Eukaryotic Translation Elongation by Cycloheximide and Lactimidomycin. *Nat Chem Biol*, 6(3):209–217, March 2010.

[177] Susan R Schwab, Katy C Li, Chulho Kang, and Nilabh Shastri. Constitutive display of cryptic translation products by mhc class i molecules. *Science*, 301(5638):1367–1371, 2003.

[178] Ataman Sendoel, Joshua G Dunn, Edwin H Rodriguez, Shruti Naik, Nicholas C Gomez, Brian Hurwitz, John Levorse, Brian D Dill, Daniel Schramek, Henrik Molina, et al. Translation from unconventional 5 start sites drives tumour initiation. *Nature*, 2017.

[179] Nilabh Shastri, Susan Schwab, and Thomas Serwold. Producing natures gene-chips: the generation of peptides for display by mhc class i molecules. *Annual review of immunology*, 20(1):463–493, 2002.

[180] MG Sheerar, BC Easterday, and VS Hinshaw. Antigenic conservation of h1n1 swine influenza viruses. *Journal of General Virology*, 70(12):3297–3303, 1989.

[181] Arthur Chun-Chieh Shih, Tzu-Chang Hsiao, Mei-Shang Ho, and Wen-Hsiung Li. Simultaneous amino acid substitutions at antigenic sites drive influenza a hemagglu-

tinin evolution. *Proceedings of the National Academy of Sciences*, 104(15):6283–6288, 2007.

[182] Shin-Ru Shih, Piin-Chau Suen, Yuan-Shou Chen, and Shih-Cheng Chang. A novel spliced transcript of influenza a/wsn/33 virus. *Virus genes*, 17(2):179–183, 1998.

[183] John Sidney, Bjoern Peters, Nicole Frahm, Christian Brander, and Alessandro Sette. Hla class i supertypes: a revised and updated classification. *BMC immunology*, 9(1):1, 2008.

[184] Lone Simonsen, Cécile Viboud, Bryan T Grenfell, Jonathan Dushoff, Lance Jennings, Marita Smit, Catherine Macken, Mami Hata, Julia Gog, Mark A Miller, et al. The genesis and spread of reassortment human influenza a/h3n2 viruses conferring adamantane resistance. *Molecular biology and evolution*, 24(8):1811–1820, 2007.

[185] John J Skehel and Don C Wiley. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annual review of biochemistry*, 69(1):531–569, 2000.

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

[187] Gavin JD Smith, Justin Bahl, Dhanasekaran Vijaykrishna, Jinxia Zhang, Leo LM Poon, Honglin Chen, Robert G Webster, JS Malik Peiris, and Yi Guan. Dating the emergence of pandemic influenza viruses. *Proceedings of the National Academy of Sciences*, 106(28):11709–11712, 2009.

[188] Saranya Sridhar. Heterosubtypic t-cell immunity to influenza in humans: challenges for universal t-cell influenza vaccines. *Frontiers in immunology*, 7:195, 2016.

[189] Saranya Sridhar, Shaima Begom, Alison Bermingham, Katja Hoschler, Walt Adamson, William Carman, Thomas Bean, Wendy Barclay, Jonathan J Deeks, and Ajit Lalvani. Cellular immune correlates of protection against symptomatic pandemic influenza. *Nature medicine*, 19(10):1305, 2013.

[190] Alexandros Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[191] Shelley R Starck, Vivian Jiang, Mariana Pavon-Eternod, Sharanya Prasad, Brian McCarthy, Tao Pan, and Nilabh Shastri. Leucine-trna initiates at cug start codons for protein synthesis and presentation by mhc class i. *Science*, 336(6089):1719–1723, 2012.

[192] Shelley R Starck, Yongkai Ow, Vivian Jiang, Maria Tokuyama, Mark Rivera, Xin Qi, Richard W Roberts, and Nilabh Shastri. A distinct translation initiation mechanism generates cryptic peptides for immune surveillance. *PloS one*, 3(10):e3460, 2008.

[193] Shelley R Starck and Nilabh Shastri. Non-conventional sources of peptides presented by mhc class i. *Cellular and molecular life sciences*, 68(9):1471–1479, 2011.

[194] Joan Argetsinger Steitz. Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage r17 rna. *Nature*, 224(5223):957, 1969.

[195] Noam Stern-Ginossar, Ben Weisburd, Annette Michalski, Vu Thuy Khanh Le, Marco Y Hein, Sheng-Xiong Huang, Ming Ma, Ben Shen, Shu-Bing Qian, Hartmut Hengel, et al. Decoding human cytomegalovirus. *Science*, 338(6110):1088–1093, 2012.

[196] Matthew A Takata, Daniel Gonçalves-Carneiro, Trinity M Zang, Steven J Soll, Ash-

ley York, Daniel Blanco-Melo, and Paul D Bieniasz. Cg dinucleotide suppression enables antiviral defence targeting non-self rna. *Nature*, 550(7674):124, 2017.

[197] Asif U Tamuri, Mario dos Reis, Alan J Hay, and Richard A Goldstein. Identifying changes in selective constraints: host shifts in influenza. *PLoS computational biology*, 5(11):e1000564, 2009.

[198] Leiming Tang, Jacob Morris, Ji Wan, Chelsea Moore, Yoshihiko Fujita, Sarah Gillaspie, Eric Aube, Jagpreet Nanda, Maud Marques, Maika Jangal, et al. Competition between translation initiation factor eif5 and its mimic protein 5mp determines non-aug initiation rate genome-wide. *Nucleic acids research*, 45(20):11941–11953, 2017.

[199] Qiannan Tang, Xinlu Wang, and Guangxia Gao. The short form of the zinc finger antiviral protein inhibits influenza a virus protein expression and is antagonized by the virus-encoded ns1. *Journal of virology*, 91(2):e01909–16, 2017.

[200] Jeffery K Taubenberger and John C Kash. Influenza virus evolution, host adaptation, and pandemic formation. *Cell host & microbe*, 7(6):440–451, 2010.

[201] Jeffery K Taubenberger, Ann H Reid, Raina M Lourens, Ruixue Wang, Guozhong Jin, and Thomas G Fanning. Characterization of the 1918 influenza virus polymerase genes. *Nature*, 437(7060):889, 2005.

[202] PM Taylor and BA Askonas. Influenza nucleoprotein-specific cytotoxic t-cell clones are protective in vivo. *Immunology*, 58(3):417, 1986.

[203] Bargavi Thyagarajan and Jesse D Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife*, 3, 2014.

[204] Suxiang Tong, Yan Li, Pierre Rivailler, Christina Conrardy, Danilo A Alvarez Castillo, Li-Mei Chen, Sergio Recuenco, James A Ellison, Charles T Davis, Ian A York,

et al. A distinct lineage of influenza a virus from bats. *Proceedings of the National Academy of Sciences*, 109(11):4269–4274, 2012.

[205] Akiko Uenaka, Toshiro Ono, Toshifumi Akisawa, Hisashi Wada, Tatsuji Yasuda, and Eiichi Nakayama. Identification of a unique antigen peptide prl1 on balb/c rl male 1 leukemia recognized by cytotoxic t lymphocytes and its relation to the akt oncogene. *Journal of Experimental Medicine*, 180(5):1599–1607, 1994.

[206] Philip J Uren, Emad Bahrami-Samani, Suzanne C Burns, Mei Qiao, Fedor V Karginov, Emily Hodges, Gregory J Hannon, Jeremy R Sanford, Luiz OF Penalva, and Andrew D Smith. Site identification in high-throughput rna–protein interaction data. *Bioinformatics*, 28(23):3013–3020, 2012.

[207] Juan Valcárcel, Agustín Portela, and Juan Ortín. Regulated m1 mrna splicing in influenza virus-infected cells. *Journal of general virology*, 72(6):1301–1308, 1991.

[208] Sophie A Valkenburg, John A Rutigliano, Ali H Ellebedy, Peter C Doherty, Paul G Thomas, and Katherine Kedzierska. Immunity to seasonal and pandemic influenza a viruses. *Microbes and infection*, 13(5):489–501, 2011.

[209] Carolien E van de Sandt and Guus F Rimmelzwaan. Immunodominant responses to the influenza virus m158–66 epitope: Stealth or protection? *Proceedings of the National Academy of Sciences*, 112(19):E2417–E2417, 2015.

[210] Amy L Vincent, Kelly M Lager, Wenjun Ma, Porntippa Lekcharoensuk, Marie R Gramer, Christina Loiacono, and Juergen A Richt. Evaluation of hemagglutinin subtype 1 swine influenza viruses from the united states. *Veterinary microbiology*, 118(3-4):212–222, 2006.

[211] Jean-Louis Virelizier. Host defenses against influenza virus: the role of anti-hemagglutinin antibody. *The Journal of Immunology*, 115(2):434–439, 1975.

[212] Randi Vita, James A Overton, Jason A Greenbaum, Julia Ponomarenko, Jason D Clark, Jason R Cantrell, Daniel K Wheeler, Joseph L Gabbard, Deborah Hix, Alessandro Sette, et al. The immune epitope database (iedb) 3.0. *Nucleic acids research*, 43(D1):D405–D412, 2014.

[213] JTM Voeten, TM Bestebroer, NJ Nieuwkoop, RAM Fouchier, ADME Osterhaus, and GF Rimmelzwaan. Antigenic drift in the influenza a virus (h3n2) nucleoprotein and escape from recognition by cytotoxic t lymphocytes. *Journal of virology*, 74(15):6800–6807, 2000.

[214] Preben von Magnus. Incomplete forms of influenza virus. In *Advances in virus research*, volume 2, pages 59–79. Elsevier, 1954.

[215] Ji Wan and Shu-Bing Qian. Tisdb: a database for alternative translation initiation in mammalian cells. *Nucleic acids research*, 42(D1):D845–D850, 2013.

[216] Zhongfang Wang, Yanmin Wan, Chenli Qiu, Sergio Quinones-Parra, Zhaoqin Zhu, Liyen Loh, Di Tian, Yanqin Ren, Yunwen Hu, Xiaoyan Zhang, et al. Recovery from severe h7n9 disease is associated with diverse response mechanisms dominated by cd8+ t cells. *Nature communications*, 6:6833, 2015.

[217] Robert G Webster, William J Bean, Owen T Gorman, Thomas M Chambers, and Yoshihiro Kawaoka. Evolution and ecology of influenza a viruses. *Microbiological reviews*, 56(1):152–179, 1992.

[218] Chih-Jen Wei, Jeffrey C Boyington, Kaifan Dai, Katherine V Houser, Melissa B Pearce, Wing-Pui Kong, Zhi-yong Yang, Terrence M Tumpey, and Gary J Nabel. Cross-neutralization of 1918 and 2009 influenza viruses: role of glycans in viral evolution and vaccine design. *Science translational medicine*, 2(24):24ra21–24ra21, 2010.

[219] Andreas O Weinzierl, Dominik Maurer, Florian Altenberend, Nicole Schneiderhan-Marra, Karin Klingel, Oliver Schoor, Dorothee Wernet, Thomas Joos, Hans-Georg Rammensee, and Stefan Stevanović. A cryptic vascular endothelial growth factor t-cell epitope: identification and characterization by mass spectrometry and t-cell assays. *Cancer research*, 68(7):2447–2454, 2008.

[220] Helen M Wise, Agnes Foeglein, Jiechao Sun, Rosa Maria Dalton, Sheetal Patel, Wendy Howard, Emma C Anderson, Wendy S Barclay, and Paul Digard. A complicated message: Identification of a novel pb1-related protein translated from influenza a virus segment 2 mrna. *Journal of virology*, 83(16):8021–8031, 2009.

[221] Helen M Wise, Edward C Hutchinson, Brett W Jagger, Amanda D Stuart, Zi H Kang, Nicole Robb, Louis M Schwartzman, John C Kash, Ervin Fodor, Andrew E Firth, et al. Identification of a novel splice variant form of the influenza a virus m2 ion channel with an antigenically distinct ectodomain. *PLoS pathogens*, 8(11):e1002998, 2012.

[222] Emily HM Wong, David K Smith, Raul Rabadan, Malik Peiris, and Leo LM Poon. Codon usage bias and the evolution of influenza a viruses. codon usage biases of influenza virus. *BMC evolutionary biology*, 10(1):253, 2010.

[223] Writing Committee of the World Health Organization (WHO) Consultation on Human Influenza A/H5. Avian influenza a (h5n1) infection in humans. *New England Journal of Medicine*, 353(13):1374–1385, 2005.

[224] Jia Xue, Benjamin S Chambers, Scott E Hensley, and Carolina B López. Propagation and characterization of influenza virus stocks that lack high levels of defective viral genomes and hemagglutinin mutations. *Frontiers in microbiology*, 7:326, 2016.

[225] Ning Yang, James S Gibbs, Heather D Hickman, Glennys V Reynoso, Arun K Ghosh, Jack R Bennink, and Jonathan W Yewdell. Defining viral defective ribosomal

products: standard and alternative translation initiation events generate a common peptide from influenza a virus m2 and m1 mrnas. *The Journal of Immunology*, 196(9):3608–3617, 2016.

[226] Jonathan W Yewdell. Confronting complexity: real-world immunodominance in antiviral cd8+ t cell responses. *Immunity*, 25(4):533–543, 2006.

[227] Jonathan W Yewdell, Luis C Antón, and Jack R Bennink. Defective ribosomal products (drips): a major source of antigenic peptides for mhc class i molecules? *The Journal of Immunology*, 157(5):1823–1826, 1996.

[228] Jonathan W Yewdell, Jack R Bennink, Geoffrey L Smith, and Bernard Moss. Influenza a virus nucleoprotein is a major target antigen for cross-reactive anti-influenza a virus cytotoxic t lymphocytes. *Proceedings of the National Academy of Sciences*, 82(6):1785–1789, 1985.

[229] Jonathan W Yewdell and Christopher V Nicchitta. The drip hypothesis decennial: support, controversy, refinement and extension. *Trends in immunology*, 27(8):368–373, 2006.

[230] Dmitriy Zamarin, Mila B Ortigoza, and Peter Palese. Influenza a virus pb1-f2 protein contributes to viral pathogenesis in mice. *Journal of virology*, 80(16):7976–7983, 2006.

[231] Jie Zhang, Andrew Pekosz, and Robert A Lamb. Influenza virus assembly and lipid raft microdomains: a role for the cytoplasmic tails of the spike glycoproteins. *Journal of Virology*, 74(10):4634–4644, 2000.

[232] Xingqian Zhang, Xiangwei Gao, Ryan Alex Coots, Crystal S. Conn, Botao Liu, and Shu-Bing Qian. Translational control of the cytosolic stress response by mitochondrial ribosomal protein L18. *Nat. Struct. Mol. Biol.*, 22(5):404–410, May 2015.

[233] Wenjie Zheng and Yizhi Jane Tao. Structure and assembly of the influenza a virus ribonucleoprotein complex. *FEBS letters*, 587(8):1206–1214, 2013.

[234] Weimin Zhong, Pedro A Reche, Char-Chang Lai, Bruce Reinhold, and Ellis L Reinherz. Genome-wide characterization of a viral cytotoxic t lymphocyte epitope repertoire. *Journal of Biological Chemistry*, 278(46):45135–45144, 2003.

[235] Yiping Zhu, Guifang Chen, Fengxiang Lv, Xinlu Wang, Xin Ji, Yihui Xu, Jing Sun, Li Wu, Yong-Tang Zheng, and Guangxia Gao. Zinc-finger antiviral protein inhibits hiv-1 infection by selectively targeting multiply spliced viral mrnas for degradation. *Proceedings of the National Academy of Sciences*, 108(38):15834–15839, 2011.

[236] Yiping Zhu, Xinlu Wang, Stephen P Goff, and Guangxia Gao. Translational repression precedes and is required for zap-mediated mrna decay. *The EMBO journal*, 31(21):4236–4246, 2012.

LIST OF FIGURES

Figure Number Page

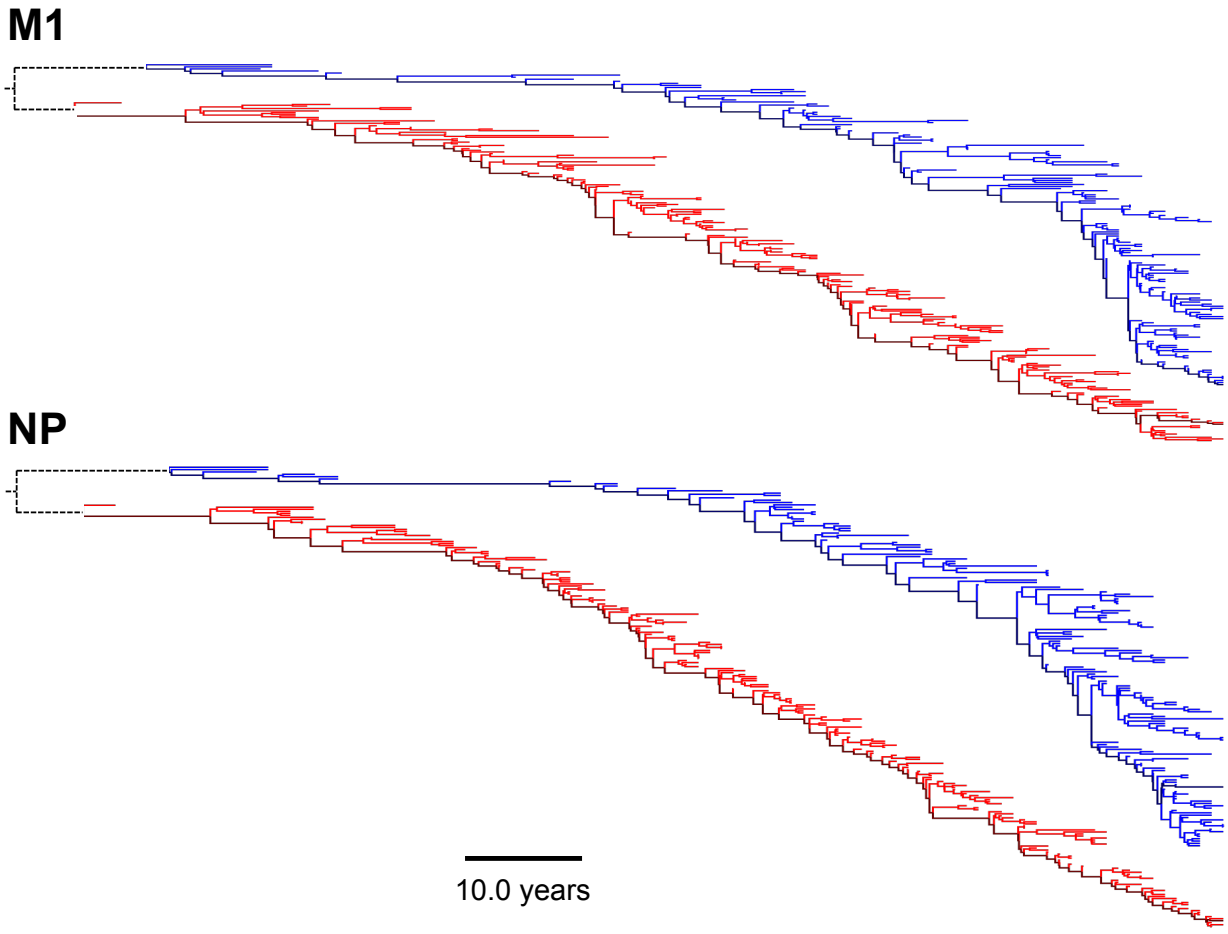


Figure 1: **Phylogenetic trees for human and swine influenza virus M1 and NP.** Shown are maximum clade credibility trees for NP and M1. The swine influenza virus lineage is in blue, and the human influenza virus lineage is in red. The dark blue and red lines represent the trunk. The dotted black lines indicate that human and swine influenza virus lineages share a recent common ancestor (estimated times to the most recent common ancestor are 13 and 6 years for M1 and NP, respectively).

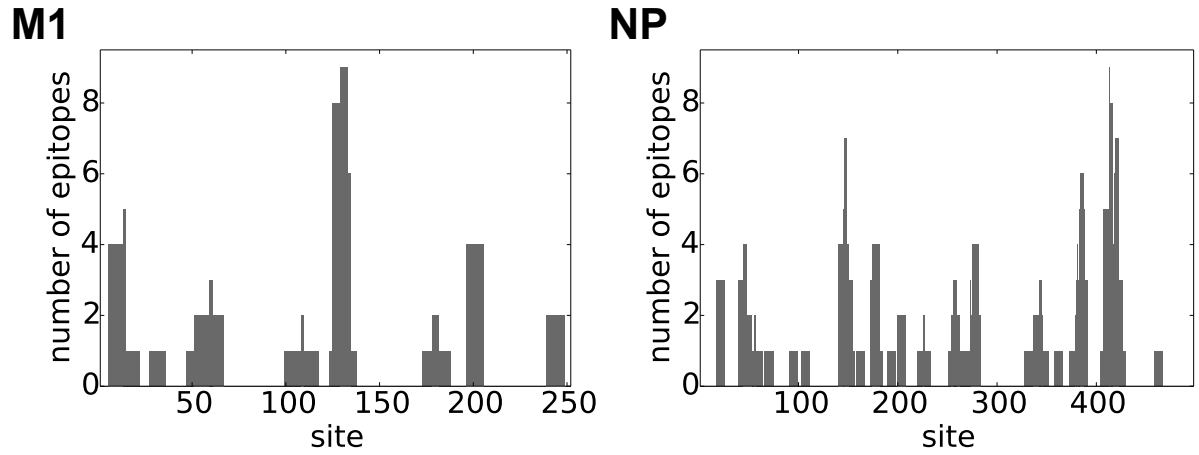


Figure 2: **Distribution of CD8 T-cell epitopes in M1 and NP.**
The numbers of unique experimentally identified CD8+ T-cell epitopes to which each site contributes for M1 and NP are shown.

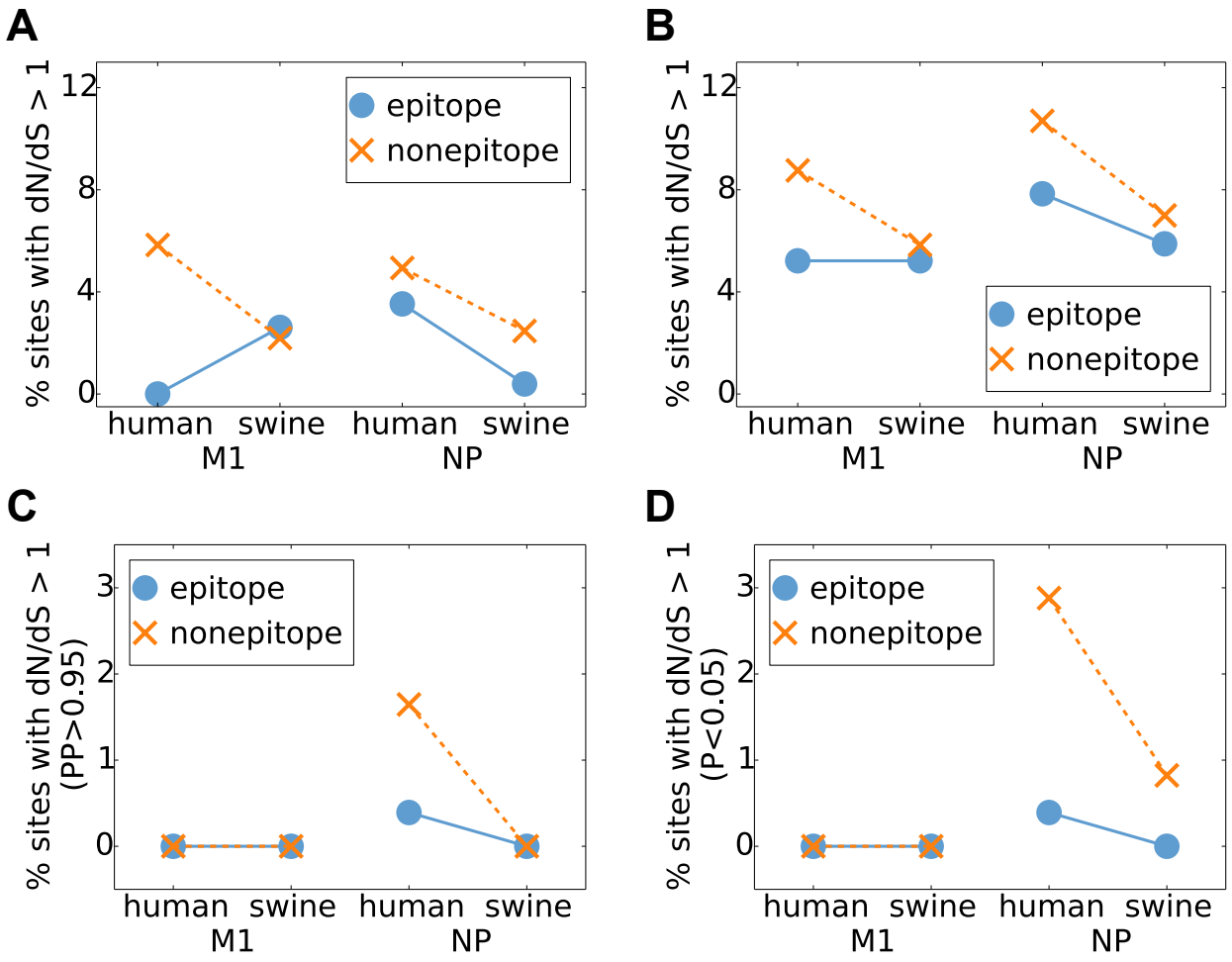


Figure 3: **Conventional dN/dS tests do not detect positive selection in T-cell epitopes.**
State-of-the-art methods for detecting positive selection fail to find any enrichment in sites with a dN/dS ratio of $>$1 at epitopes. **(A)** and **(B)** Percentages of sites estimated to have a dN/dS ratio of $>$1 by using the hierarchical Bayesian approach implemented in FUBAR **(A)** and the maximum likelihood approach implemented in FEL **(B)**. **(C)** Percentages of sites for which FUBAR reports a posterior probability of 0.95 that the dN/dS ratio is $>$1. **(D)** Percentages of sites for which FEL reports a dN/dS ratio of $>$1 with a P value of 0.05.

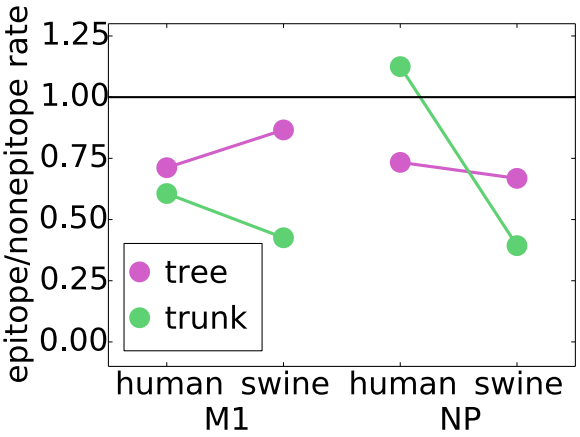


Figure 4: **Epitope sites do not evolve faster than nonepitope sites.**
The ratios of the average nonsynonymous substitution rate of epitope sites to that of nonepitope sites are shown. When the substitution rate is computed across the entire tree, the epitope sites always have a lower substitution rate than the nonepitope sites. However, along the trunk of the tree for NP from human influenza virus, the substitution rate in epitopes is slightly higher than the substitution rate at nonepitopes.

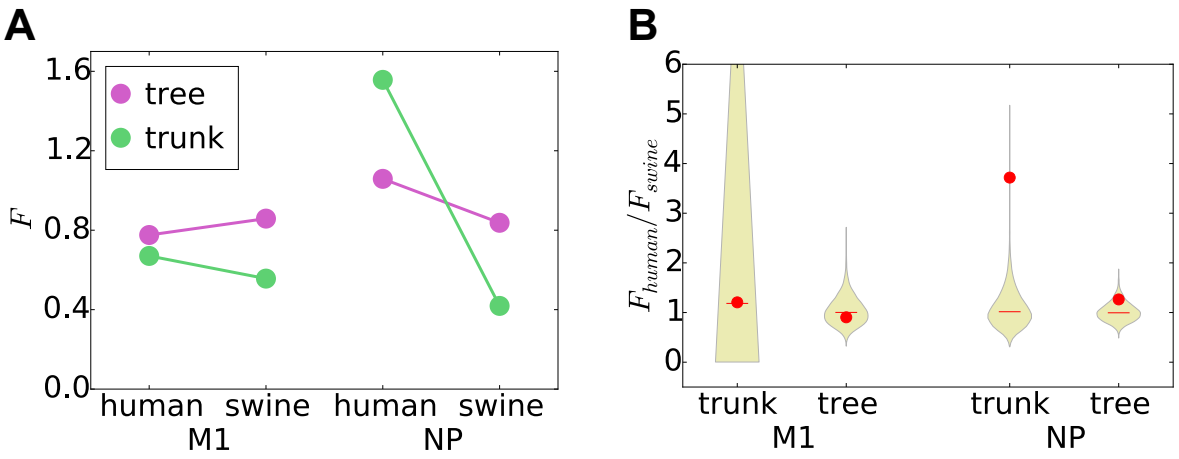


Figure 5: **The average substitution changes more NP epitopes in human influenza than swine influenza virus.**
The F statistic is the average number of epitopes altered per substitution. **(A)** The average substitution alters a similar number of epitopes in M1 proteins of both human and swine influenza viruses and for both the entire tree and the trunk alone. However, for NP, the average substitution alters substantially more epitopes in human than in swine influenza virus, particularly along the trunk of the tree. **(B)** The increased rate of epitope-altering substitutions in NP along the trunk of human versus swine influenza virus is statistically significant. The violin plots show the null distribution of the ratio of F for human influenza virus to that for swine influenza virus, with the median shown by the red lines and the actual value shown by red circles. The P values (fraction of the null distribution that is greater than or equal to the actual value) are 0.49 for the M1 trunk, 0.65 for the M1 tree, 0.0002 for the NP trunk, and 0.098 for the NP tree.

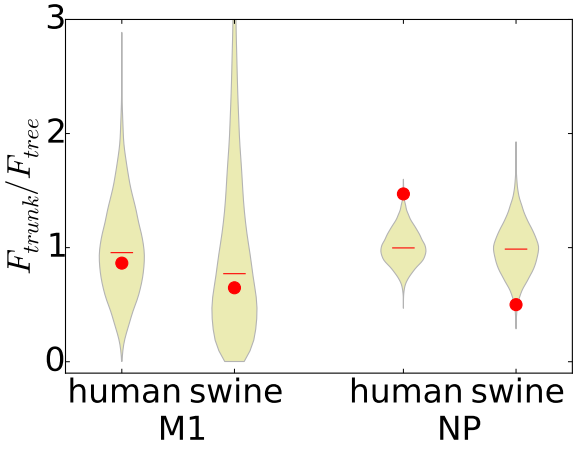


Figure 6: **Epitope-altering substitutions in human influenza virus NP are enriched on the trunk of the tree.**

Epitope-altering substitutions in human influenza virus NP are enriched on the trunk of the tree. Ratios of the average number of epitopes altered per substitution (F) for the trunk to that for the entire tree are shown. The violin plots show the actual values versus the null distributions, as described in the legend of Figure 5B. There is no difference between the trunk and the entire tree for M1 (P values are 0.59 for M1 of human influenza virus and 0.57 for M1 of swine influenza virus). However, for NP, epitope-altering substitutions are significantly enriched on the trunk of the human influenza virus tree relative to the rest of the tree (P = 0.003) and significantly depleted on the trunk of the swine influenza virus lineage (P = 0.01 for depletion) relative to the rest of the tree.

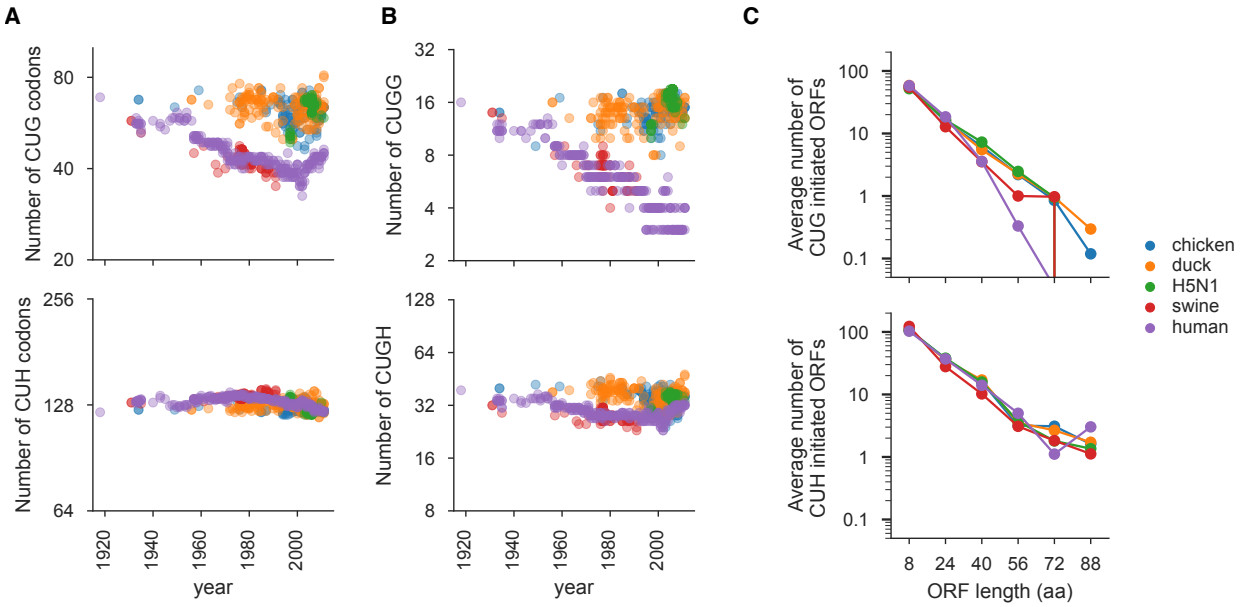


Figure 7: **Evolutionary selection against CTG codons in mammalian influenza. (A)** Number of CUG (upper panel) and CUH (where H is A, U or C, lower panel) codons in reading frame 0 of viral genes from human, avian, and swine influenza sequences. X-axis indicates the year that the viral sequence was isolated. **(B)** Number of in-frame CUGG (upper panel) and CUGH (lower panel) motifs in reading frame 0 of the viral genes. **(C)** Average length in codons of CUG- and CUH-initiated ORFs in reading frames 1 and 2 of the viral genes for sequences isolated between 1970 and 2011. All plots show data only for the genes on the PB2, PA, NP, NS and M segments, since these are the segments that have not re-assorted in human lineages derived from the 1918 virus.

Figure 8: **Translation initiation profiling of influenza infected cells. (A)** Schematic of assays performed on A549 cells. Infections contained a mixture of viruses with their NP gene recoded to have low and high numbers of CUG codons (see Fig. 16). **(B)** Proportion of reads that align to viral and human transcripts for RNA-seq (mRNA), ribosome profiling (Ribo-seq), or Ribo-seq with LTM treatment. **(C)** Length distribution of reads that aligned to viral and human transcripts for Ribo-seq and Ribo-seq + LTM samples. **(D)** Metagene alignment of average P-site density around annotated start codons in viral and human transcripts for Ribo-seq + LTM sample. **(E)** Metagene alignment of average P-site density around annotated start and stop codons in viral and human transcripts for Ribo-seq samples. **(F)** Normalized P-site density in each of the reading frames of viral and human transcripts for RNA-seq (gray) and Ribo-seq (orange) samples. Both single- and known dual-coded regions in influenza transcripts were considered. **(G)** Normalized P-site density in each of the reading frames of single- and dual-coded regions of M, NS, and PB1 segments of the influenza genome for RNA-seq and Ribo-seq samples. Panels C–G show the +vir sample; see Fig. 17 for comparable plots for other samples. Only reads mapping to + sense of influenza transcripts are considered for panels B–G.

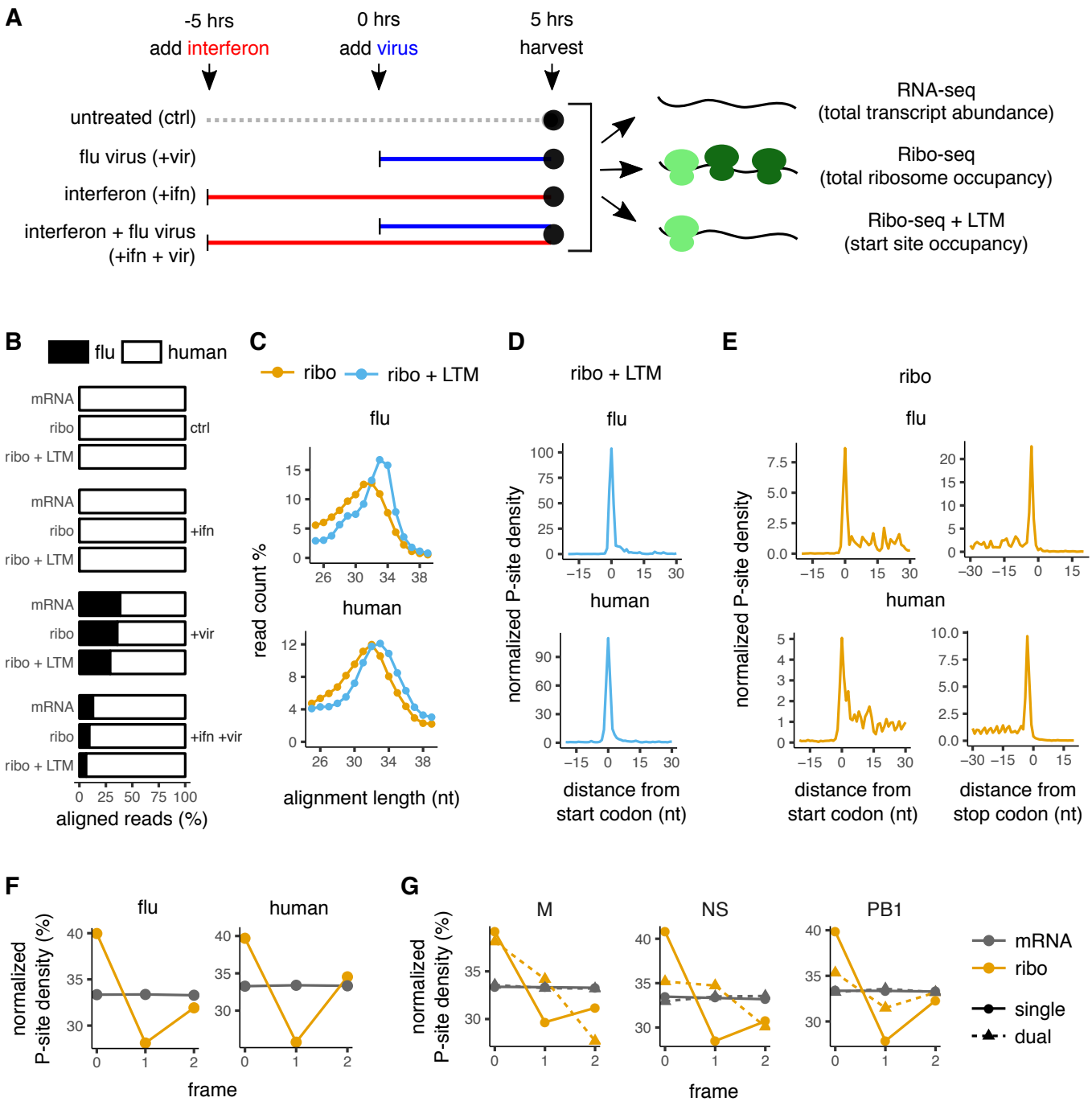

A
-5 hrs
add interferon
0 hrs
add virus
5 hrs
harvest
untreated (ctrl)
flu virus (+vir)
interferon (+ifn)
interferon + flu virus
(+ifn + vir)
RNA-seq
(total transcript abundance)
Ribo-seq
(total ribosome occupancy)
Ribo-seq + LTM
(start site occupancy)
B
flu
human
mRNA
ribo
ribo + LTM
ctrl
+ifn
+vir
+ifn +vir
aligned reads (%)
C
ribo
ribo + LTM
flu
human
read count %
alignment length (nt)
D
ribo + LTM
flu
human
normalized P-site density
distance from
start codon (nt)
E
ribo
flu
human
normalized P-site density
distance from
start codon (nt)
distance from
stop codon (nt)
F
flu
human
normalized
P-site density (%)
frame
G
M
NS
PB1
normalized
P-site density (%)
frame
mRNA
ribo
single
dual

Figure 8

Figure 9: **Candidate translation initiation sites on influenza transcripts. (A)** Method for identifying candidate translation initiation sites (TIS) in influenza. Left panel: LTM-specific outliers were identified over the background of Ribo-seq and Ribo-seq + LTM pooled counts at each site of each influenza transcript. Right-panel: Among LTM-specific outliers, only the highest LTM peak in P-site counts within each 30 nucleotide window was called as start site. The background Ribo-seq and Ribo-seq + LTM counts were fit to separate zero-truncated negative binomial distributions (lines in left panel). The final called TIS are indicated by grey triangles and arrows. The left panel shows the NA segment of our +vir sample and the right panel a window around the NA43 TIS. See Fig. 19 for other influenza segments. **(B)** Overlap in candidate TIS between the +vir and +ifn +vir samples. **(C)** Consensus motif of annotated translation initiation sites (aTIS, upper panel) and downstream translation initiation sites (dTIS, lower panel). **(D)** P-site counts from Ribo-seq and Ribo-seq + LTM assays are shown for all 8 influenza genome segments for our +vir sample. The counts from the two assays are shown as stacked bar graphs for ease of comparison. The candidate annotated TIS (circle) and downstream TIS (triangle) are indicated below the coverage plots. The known alternate initiation site PB1-F2 is highlighted with arrow. **(E)** Number of called annotated and downstream TIS. **(F)** Distribution of codon identity for candidate influenza TIS (top) and for all codons in influenza genome (bottom). Other indicates all non-AUG codons. **(G)** Distribution of AUG codons along transcripts for candidate influenza TIS (top) and for all AUG codons (bottom). The codon locations were binned into three fragments of equal length for each influenza CDS. **(H)** Distribution among reading frames for candidate downstream TIS (top) and for all AUG codons in the influenza genome (bottom). **(I)** Length of peptides initiated by called out-of-frame TIS (top) and all out-of-frame AUG codons in the influenza genome (bottom). Only the candidate TIS shared between the +vir and +ifn +vir samples were used for panels C–I.

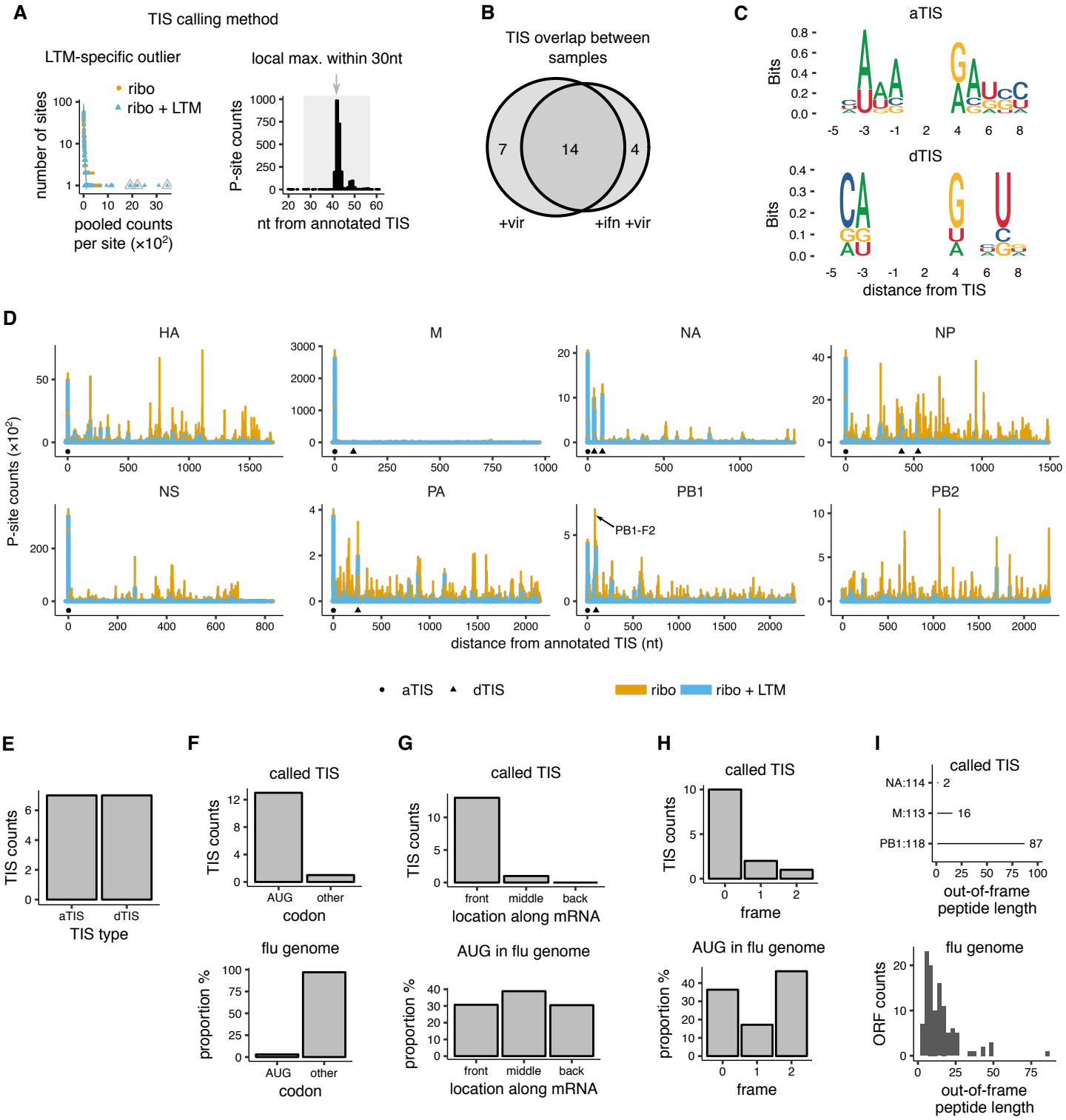

A
TIS calling method
LTM-specific outlier
ribo
ribo + LTM
number of sites
pooled counts per site (×10²)
local max. within 30nt
P-site counts
nt from annotated TIS
B
TIS overlap between samples
7
14
4
+vir
+ifn +vir
C
aTIS
dTIS
Bits
distance from TIS
D
HA
M
NA
NP
NS
PA
PB1
PB2
PB1-F2
P-site counts (×10²)
distance from annotated TIS (nt)
aTIS
dTIS
ribo
ribo + LTM
E
TIS counts
aTIS
dTIS
TIS type
F
called TIS
TIS counts
AUG
other
codon
flu genome
proportion %
G
called TIS
front
middle
back
location along mRNA
AUG in flu genome
H
called TIS
frame
AUG in flu genome
I
called TIS
NA:114
2
M:113
16
PB1:118
87
out-of-frame peptide length
flu genome
ORF counts

Figure 9

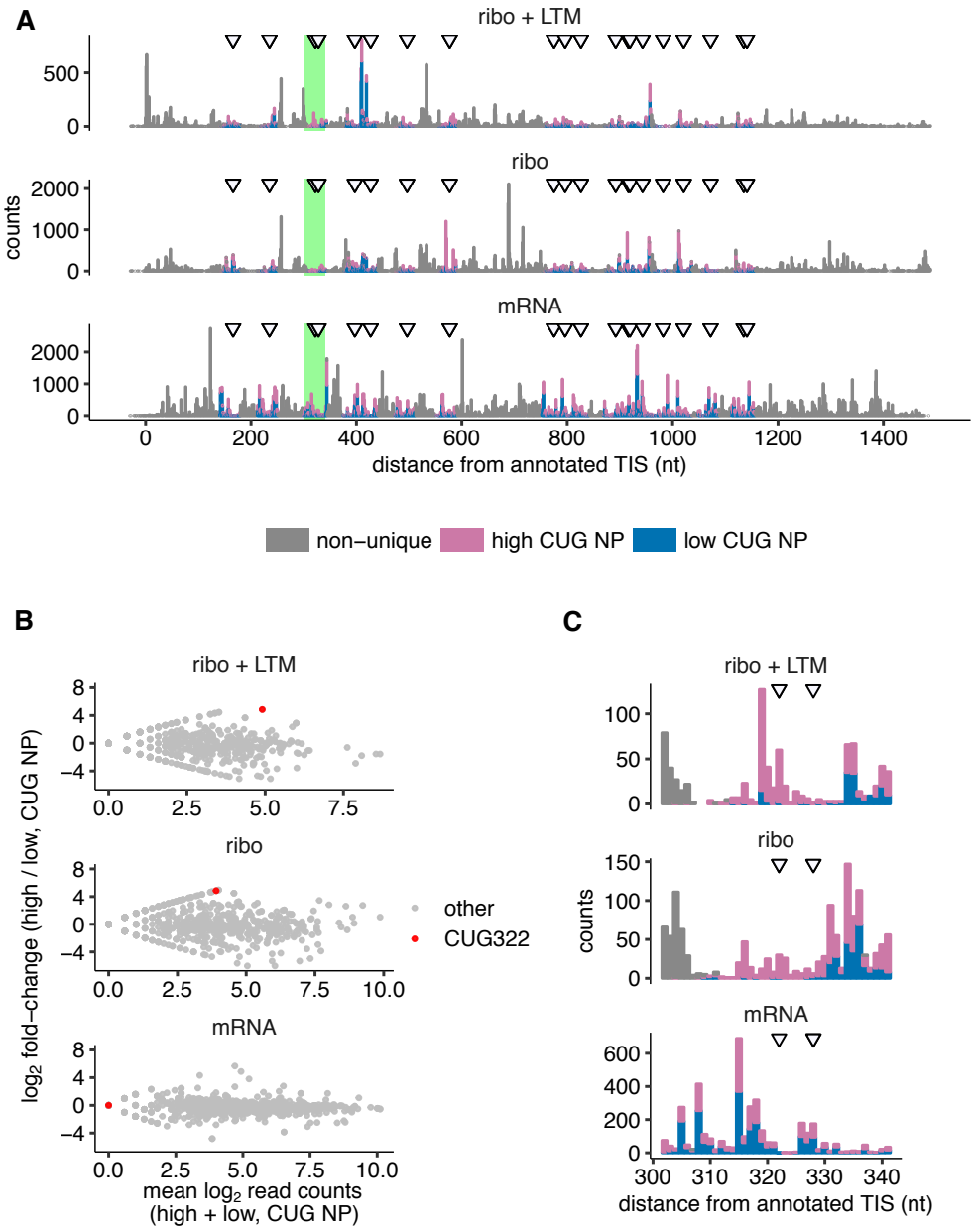


Figure 10: **(A)** Coverage of Ribo-seq + LTM, Ribo-seq, and RNA-seq reads that can be uniquely aligned to either the high CUG NP variant or the low CUG NP variant, along with reads that cannot be uniquely assigned. P-site counts are shown for Ribo-seq and Ribo-seq + LTM assays. 5'-end counts are shown for RNA-Seq. Data are plotted as a stacked bar graph. Locations of the 20 CUG codons that are present in high CUG NP and synonymously mutated in low CUG NP are indicated by arrows. **(B)** The ratio of high CUG NP to low CUG NP coverage from panel A is plotted against their sum along the horizontal axis. **(C)** The green-highlighted region in panel A around the CUG322 codon is shown at greater horizontal magnification. Data shown for +vir sample. See Fig. 25 for +ifn +vir sample.

Figure 11: **Characterization of the NA43 protein generated by alternate translation initiation in the NA gene. (A)** P-site counts from Ribo-seq and Ribo-seq + LTM assays are shown as stacked bar graphs for the 5′ end of the mRNA for our +vir sample. The candidate annotated TIS (circle) at coding nucleotide position 1 and the downstream TIS (triangle) at coding nucleotide position 43 are indicated below the coverage plots. **(B)** Schematic of NA's primary sequence. NA's cytoplasmic tail (CT) spans amino acids 1-6, the transmembrane domain (TMD) spans amino acids 7-35, and the ectodomain spans amino acids 36 to 453. The inset shows the AUG43 codon (in red) and surrounding sequence. **(C)** Most NAs in major lineages of N1 influenza have either an AUG or a near-cognate start codon at nucleotide 43. **(D)** Quantification of NA cell-surface protein levels and enzymatic activity in 293T cells transfected with plasmids encoding NA variants with mutations at the canonical or downstream start site. Results are shown for NA from the WSN and CA09 viral strains. Measurements are normalized relative to the mean for wild-type, and black lines indicate the mean of three measurements for all variants except for WSN $1_{wt}43_{GUA}$, which only had two measurements. ND indicates that the value was below the limit of detection. **(E)** Schematic of coding region of NA-H2B-V5 constructs used for Western blots. Blot of 293T cells transfected with NA-H2B-V5 constructs with mutations at the canonical or downstream start site. "43 start" is a size control construct that begins at site 43 of NA. Top panel: anti-V5 (blue arrow corresponds to full length NA and orange arrow corresponds to NA43); bottom panel: anti-GAPDH. **(F)** Viral titers in the supernatant at 48 and 72 hours post-transfection during reverse-genetics generation of WSN viruses carrying NA with the indicated mutations. Three independent replicates for each mutant are shown in gray, and the median is shown in black. Undetectable titers are plotted at the assay's limit of detection of 0.1 TCID50 / μl. **(G)** *In vitro* competition of wildtype virus versus virus carrying the indicated NA mutation to ablate the initiation codon at site 43. We mixed wildtype virus and mutant virus at a target ratio of 1:1 infectious particles, infected cells, and collected viral RNA at 10 hours and 72 hours post-infection to determine variant frequencies by deep sequencing. Shown is the enrichment of wildtype to mutant for the 72 hour timepoint relative to the 10 hour timepoint. The black bar indicates the mean of the three biological replicates. There was no consistent enrichment of the wildtype virus relative to the mutant lacking NA43. **(H)** Similar competition experiments performed *in vivo* in mice. We collected samples at 96 hours post-infection to determine viral frequencies. Shown is the enrichment of wildtype to mutant after 96 hours of viral growth in mice relative to the 10 hour cell culture timepoint. The black bar indicates the mean of the three biological replicates. We performed a paired *t*-test with the resulting P-values: 0.09 ($1_{wt}43_{GUA}$), 0.11 ($1_{wt}43_{UUA}$).

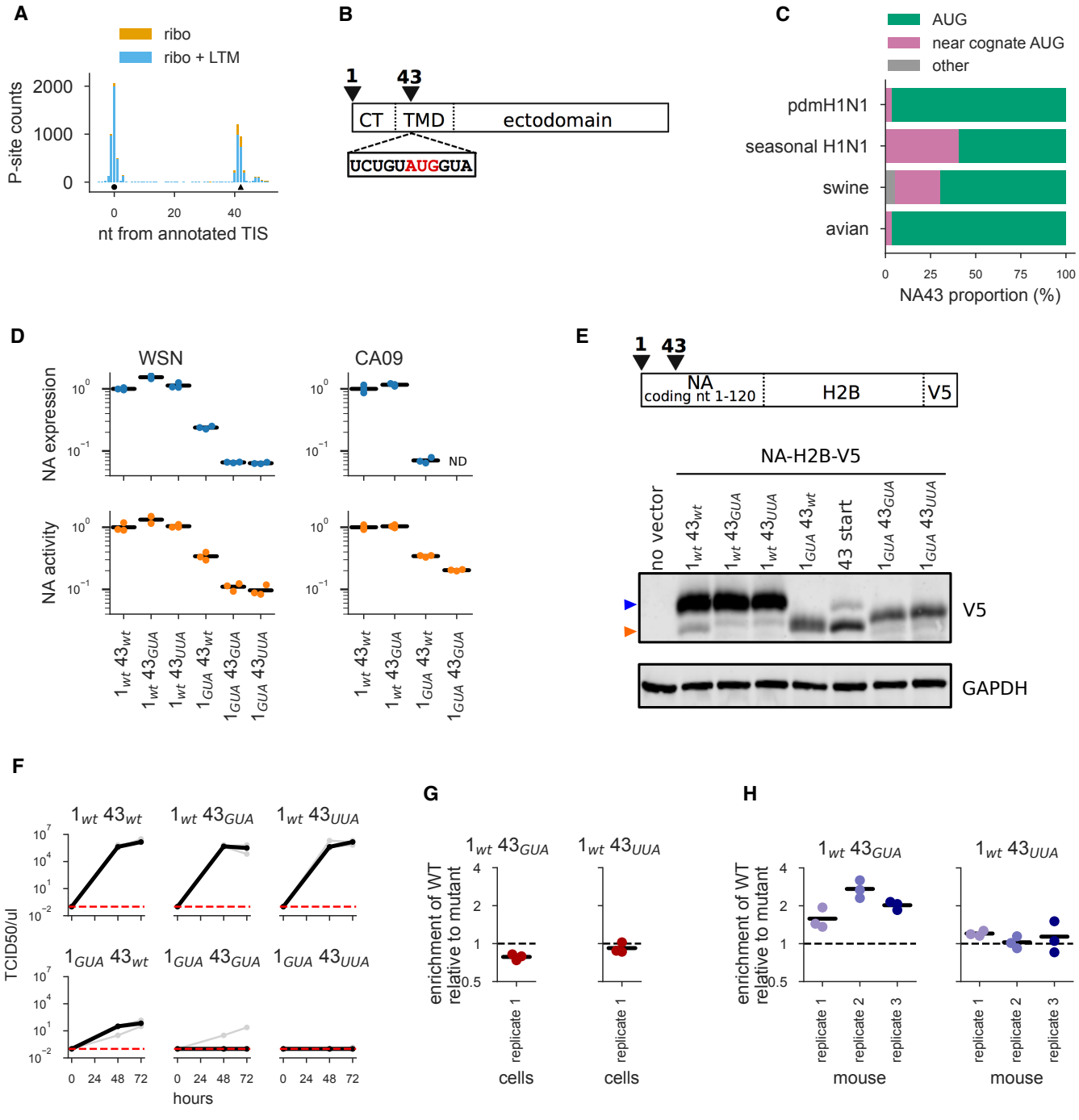

A
ribo
ribo + LTM
P-site counts
2000
1000
0
0
20
40
nt from annotated TIS
B
1
43
CT
TMD
ectodomain
UCUGUAUGGUA
C
AUG
near cognate AUG
other
pdmH1N1
seasonal H1N1
swine
avian
0
25
50
75
100
NA43 proportion (%)
D
WSN
CA09
NA expression
NA activity
ND
1wt 43wt
1wt 43GUA
1wt 43UUA
1GUA 43wt
1GUA 43GUA
1GUA 43UUA
E
1
43
NA
coding nt 1-120
H2B
V5
NA-H2B-V5
no vector
43 start
V5
GAPDH
F
TCID50/ul
hours
0
24
48
72
G
enrichment of WT
relative to mutant
replicate 1
cells
H
mouse
replicate 1
replicate 2
replicate 3

Figure 11

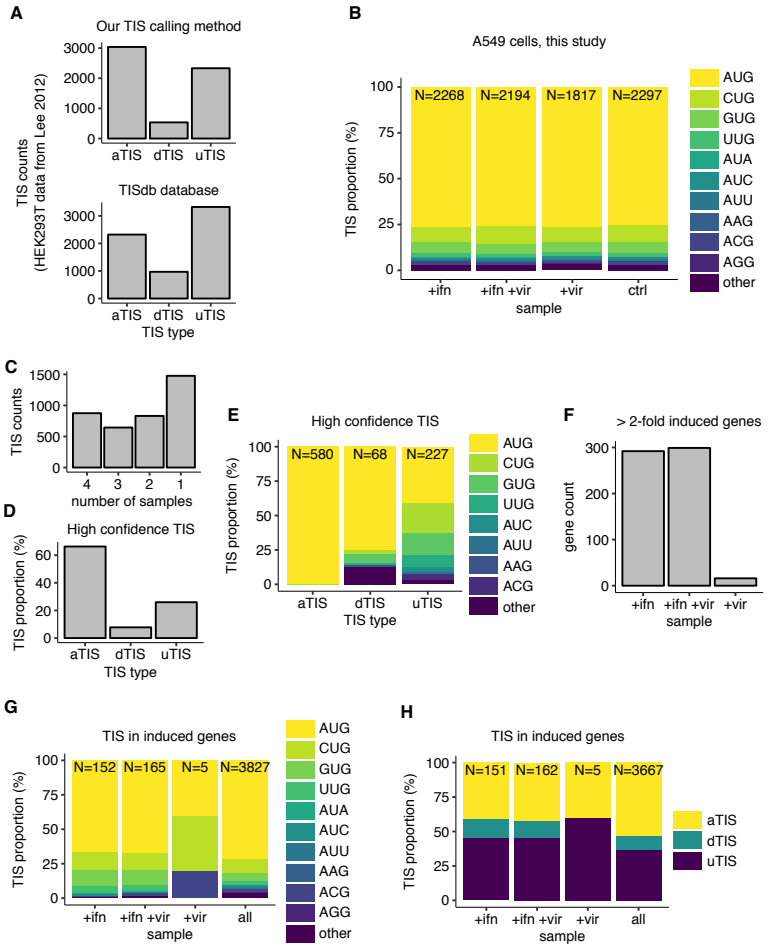


Figure 12: **(A)** Number of different TIS types—aTIS, dTIS and uTIS—called using our start site calling method (upper panel) or in the TISdb database [215] (lower panel). See main text for description of our calling method. **(B)** Proportion of different near-cognate AUG codons (or other codons) at the called TIS in each of the four samples in our study. N at the top of each bar indicates the total number of TIS called in each sample. **(C)** Overlap among TIS called in each of the 4 samples. Vertical axis indicates the number of TIS that are called in a given number of samples as indicated along the horizontal axis. **(D)** Proportion of each TIS type that are called in all 4 samples (designated as high-confidence TIS). **(E)** Proportion of different near-cognate AUG codons (or other codons) among the high-confidence TIS, stratified by TIS type. N at the top of each bar indicates the total number of high-confidence TIS of each type. **(F)** Number of genes that are induced greater than 2-fold (average of RNA-Seq and Ribo-seq counts, median-normalized) upon +ifn, +ifn +vir, or +vir treatment with respect to the control untreated sample. **(G)** Proportion of different near-cognate AUG codons (or other codons) among induced genes shown in F. The set of all TIS called in at least one of the four samples in B is shown as a control. N at the top of each bar indicates the total number of TIS that are called in each gene set. **(H)** Proportion of different TIS types among induced genes shown in F. The set of all TIS called in at least one of the four samples in B is shown as a control. N at the top of each bar indicates the total number of TIS that are called in each gene set.

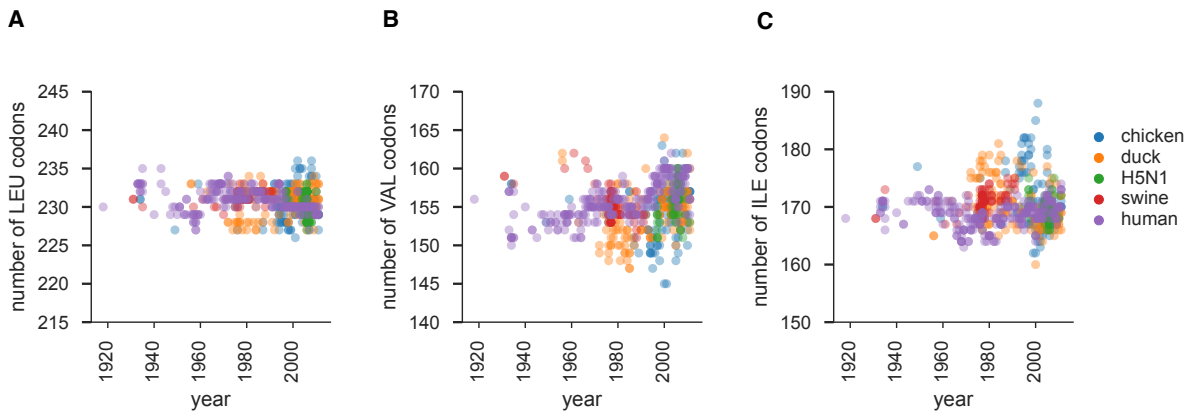


Figure 13: **Evolution of amino acids in influenza.** The number of LEU, VAL, ILE codons in reading frames 0 of the influenza genome over time in human, avian, and swine lineages. There is no systematic trend for enrichment or depletion of any of the amino acids.

Figure 14: **(A)** Evolution of the CUG odds ratio ($\frac{\mathrm{CUG}}{\mathrm{C}\times\mathrm{U}\times\mathrm{G}}$) in reading frames 0, 1, and 2 over time in human, avian, and swine lineages. The selection against CUG in reading frame 0 exists even when we use the odds ratio to correct for nucleotide usage. **(B)** Evolution of the number of CUG codons in reading frames 0, 1 and 2 of the influenza genome over time in human, avian, and swine lineages.

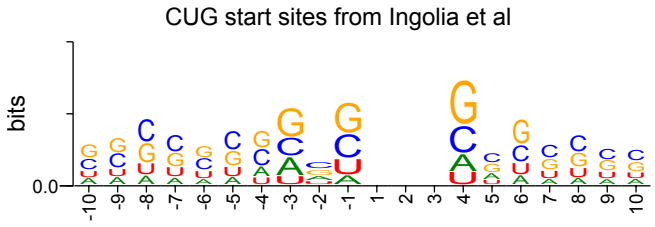
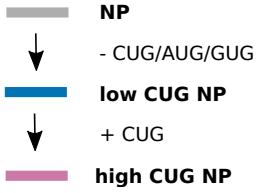


Figure 15: **Consensus motif for CUG translation initiation sites.** The consensus motif for CUG TIS was culled from the Ingolia ribosome profiling dataset [94].The most prominent feature of the consensus is a G at the +4 position.

A



B

Codon	Reading frame 0	Reading frame 1	Reading frame 2
GUG	9 → 3	11 → 0	5 → 5
AUG	25 → 25	19 → 4	1 → 0
CUG	6 → 0	16 → 0	6 → 0

Figure 16: **Summary of recoding influenza NP to generate low and high CUG NP. (A)** Low CUG NP was generated by depleting wild-type PR8 NP of the common alternate start codons AUG, CUG, and GUG in all reading frames under constraints explained in Methods. High CUG NP was generated by adding 20 CUG codons into the low CUG NP background. All changes were synonymous with respect to reading frame 0. **(B)** Summary of the differences between wild-type PR8 NP (first number) and low CUG NP (second number) at the indicated codons.

Figure 17: **Summary of quality control steps for Ribo-seq data.** **(A)** Number of input, trimmed, rRNA-aligned, and human/influenza transcript-aligned reads for assays and samples shown in Fig. 8. Number of reads aligning to + and – strands of influenza genome are shown separately. **(B)** Length distribution of viral (+ strand only) and human transcript aligned reads for Ribo-seq and Ribo-seq + LTM samples. **(C)** Metagene alignment of average P-site density around annotated start codons in viral and human transcripts for Ribo-seq + LTM and Ribo-seq samples. Right panels show the same for annotated stop codons for Ribo-seq samples. **(D)** Length distribution of viral (- strand only) genome aligned reads for Ribo-seq and Ribo-seq + LTM samples. **(E)** Metagene P-site density showing 3 nucleotide periodicity in a representative region of human transcripts for Ribo-seq samples. **(F)** Normalized P-site density in each of the reading frames of viral and human transcripts for RNA-seq and Ribo-seq samples. **(G)** Normalized P-site density in each of the reading frames of single- and dual-coded regions of M, NS, and PB1 segments of the influenza genome for RNA-seq and Ribo-seq samples for the +ifn +vir sample.

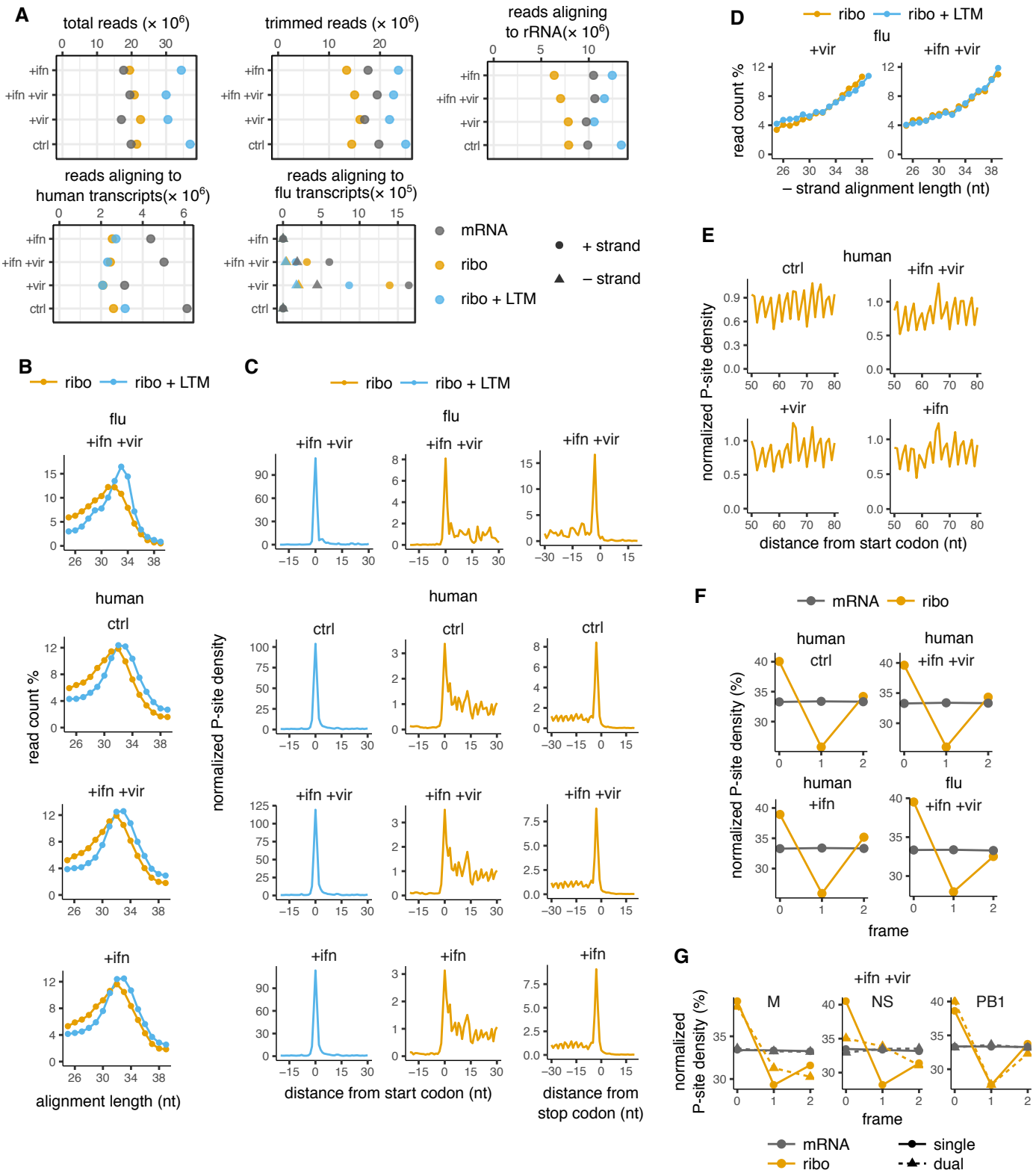

A
total reads (× 10^6)
trimmed reads (× 10^6)
reads aligning to rRNA(× 10^6)
reads aligning to human transcripts(× 10^6)
reads aligning to flu transcripts(× 10^5)
+ifn
+ifn +vir
+vir
ctrl
mRNA
ribo
ribo + LTM
+ strand
– strand
B
ribo
ribo + LTM
flu
+ifn +vir
human
ctrl
+ifn
read count %
alignment length (nt)
C
normalized P-site density
distance from start codon (nt)
distance from stop codon (nt)
D
flu
+vir
read count %
– strand alignment length (nt)
E
human
normalized P-site density
distance from start codon (nt)
F
mRNA
ribo
normalized P-site density (%)
frame
G
M
NS
PB1
single
dual

Figure 17

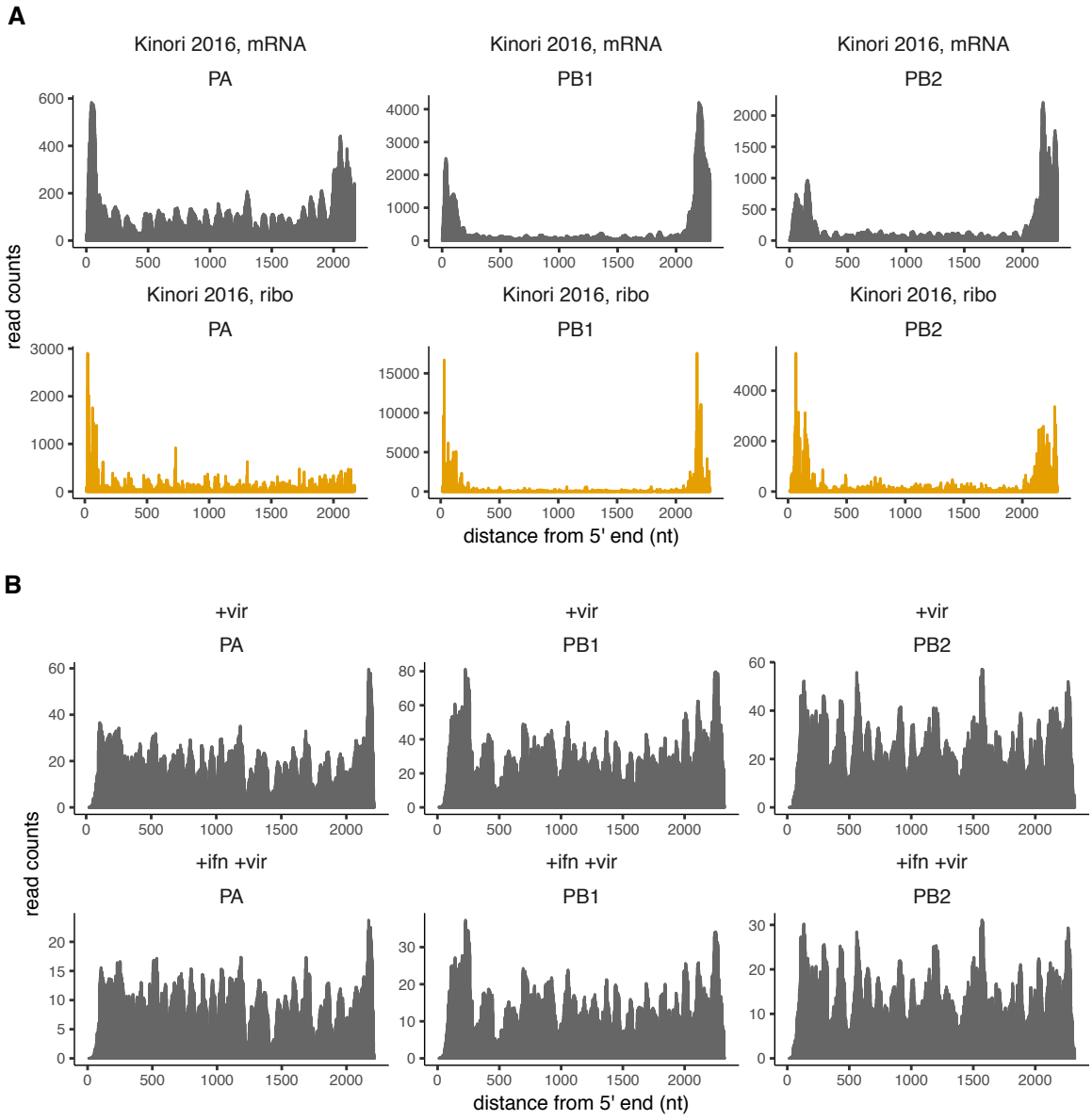


Figure 18: **Read coverage of influenza polymerase segments reflects defective viral particles. (A)** Distribution of RNA-seq and Ribo-seq read density along the polymerase segments of influenza. Raw data from [13]. **(B)** Distribution of RNA-seq read density along the polymerase segments of influenza in our data. See Fig. 9 for corresponding Ribo-seq density distribution. Defective viral particles are often characterized by the accumulation of large internal deletions in the polymerase segments. The large drop in coverage in panel A is consistent with the virus used in [13] containing a high burden of defective viral particles.

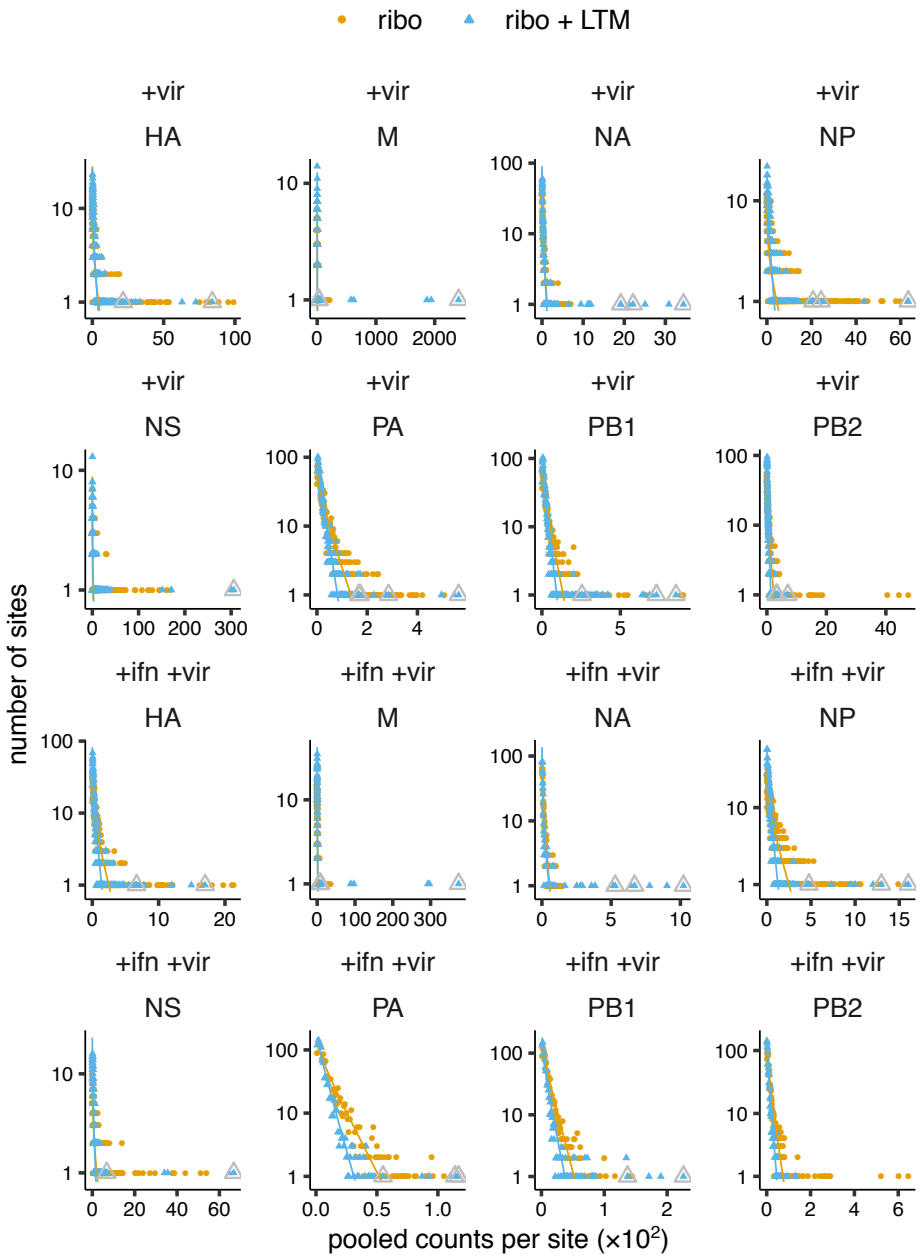


Figure 19: **Background distribution fits for each influenza genome segment.** The background Ribo-seq and Ribo-seq + LTM counts for each influenza genome segment were fit to separate zero-truncated negative binomial distributions (shown as lines). The final called TIS are indicated by grey triangles.

sample	codon	segment	frame	peptide length (aa)	distance to aTIS (nt)	ribo pval	LTM pval	called in
+ifn+vir	AUG	HA	0	565	0	2e-01	1e-14	both
+vir	AUG	HA	0	565	0	8e-02	0e+00	both
+ifn+vir	AUG	M	1	16	88	4e-01	8e-06	both
+vir	AUG	M	1	16	88	6e-01	7e-06	both
+ifn+vir	AUG	M	0	252	0	8e-14	0e+00	both
+vir	AUG	M	0	252	0	3e-14	0e+00	both
+ifn+vir	AUG	NA	2	2	95	1e-03	9e-09	both
+vir	AUG	NA	2	2	95	5e-05	0e+00	both
+ifn+vir	AUG	NA	0	453	0	5e-02	4e-15	both
+vir	AUG	NA	0	453	0	5e-02	0e+00	both
+ifn+vir	AUG	NA	0	439	42	1e-03	1e-10	both
+vir	AUG	NA	0	439	42	1e-04	0e+00	both
+ifn+vir	AUG	NP	0	498	0	3e-01	0e+00	both
+vir	AUG	NP	0	498	0	2e-01	0e+00	both
+ifn+vir	AUG	NP	0	362	408	7e-02	3e-14	both
+vir	AUG	NP	0	362	408	7e-02	3e-07	both
+ifn+vir	AUG	NS	0	230	0	1e-01	0e+00	both
+vir	AUG	NS	0	230	0	7e-02	0e+00	both
+ifn+vir	AUG	PA	0	716	0	3e-01	2e-08	both
+vir	AUG	PA	0	716	0	2e-01	2e-13	both
+ifn+vir	AUG	PA	0	631	255	1e-02	1e-08	both
+vir	AUG	PA	0	631	255	3e-03	3e-07	both
+ifn+vir	AUG	PB1	1	87	94	1e-01	2e-11	both
+vir	AUG	PB1	1	87	94	6e-02	0e+00	both
+ifn+vir	AUG	PB1	0	757	0	3e-01	2e-07	both
+vir	AUG	PB1	0	757	0	3e-01	2e-14	both
+vir	AUG	PA	1	1	1156	2e-01	1e-04	+vir
+vir	AUG	PB2	0	759	0	6e-01	2e-04	+vir
+ifn+vir	AUG	HA	2	7	800	2e-01	2e-06	+ifn+vir
+ifn+vir	AUG	M	0	5	741	5e-04	4e-08	+ifn+vir
+ifn+vir	AUG	NS	0	150	240	4e-01	3e-04	+ifn+vir
+ifn+vir	AUG	PA	2	17	113	4e-01	2e-04	+ifn+vir
+ifn+vir	other	NP	n.a.	n.a.	n.a.	9e-03	7e-06	both
+vir	other	NP	n.a.	n.a.	n.a.	5e-03	3e-06	both
+vir	other	HA	n.a.	n.a.	n.a.	5e-02	4e-05	+vir
+vir	other	M	n.a.	n.a.	n.a.	2e-01	1e-04	+vir
+vir	other	PA	n.a.	n.a.	n.a.	4e-01	1e-04	+vir
+vir	other	PB1	n.a.	n.a.	n.a.	4e-02	1e-05	+vir
+vir	other	PB2	n.a.	n.a.	n.a.	2e-01	6e-05	+vir

Figure 20: **List of candidate start sites in the influenza genome.**

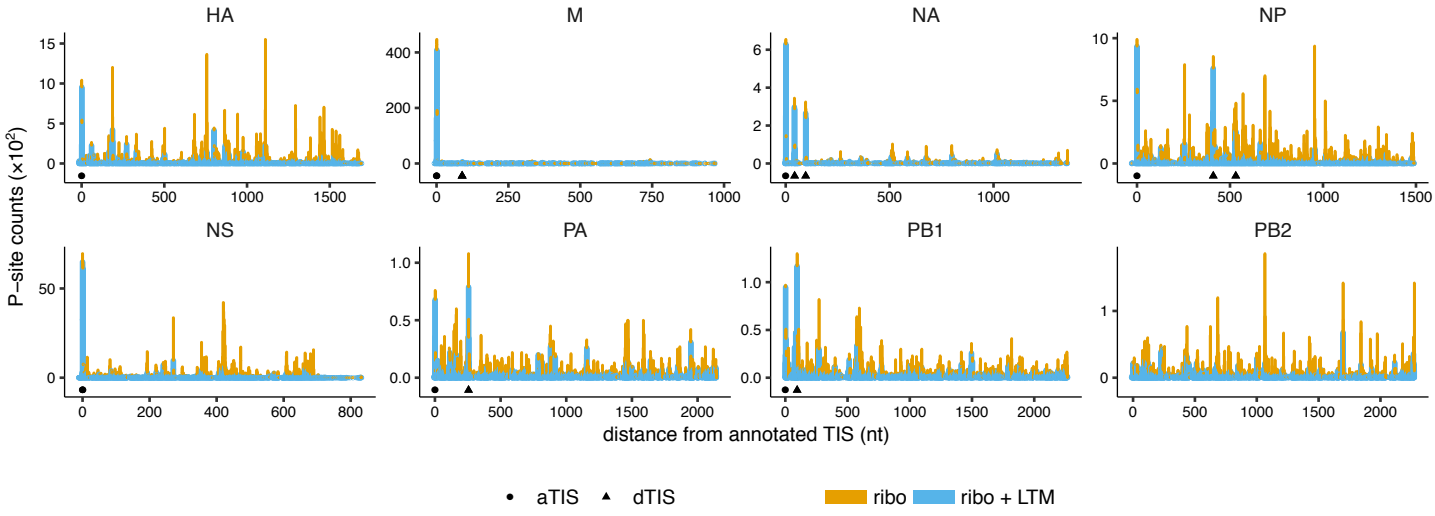


Figure 21: P-site counts from Ribo-seq and Ribo-seq + LTM assays are shown for all 8 influenza genome segments for our +ifn +vir sample. The counts from the two assays are shown as stacked bar graphs for ease of comparison. The candidate annotated TIS (circle) and downstream TIS (triangle) shared between the +vir and +ifn +vir samples are indicated below the coverage plots.

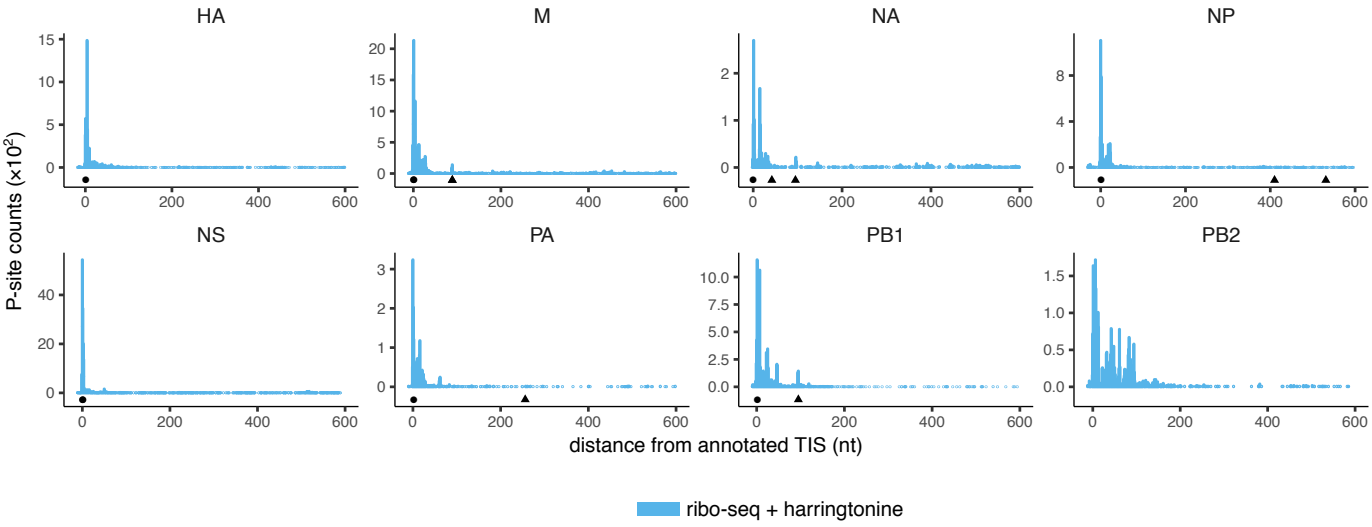


Figure 22: **Peaks in Harringtonine-treated influenza samples from [13]** The P-site count coverage from [13] is overlaid with the 14 putative TIS identified in both of our +vir and +ifn +vir samples. Only the first 600nt of all coding sequences which contains the putative TIS is shown.

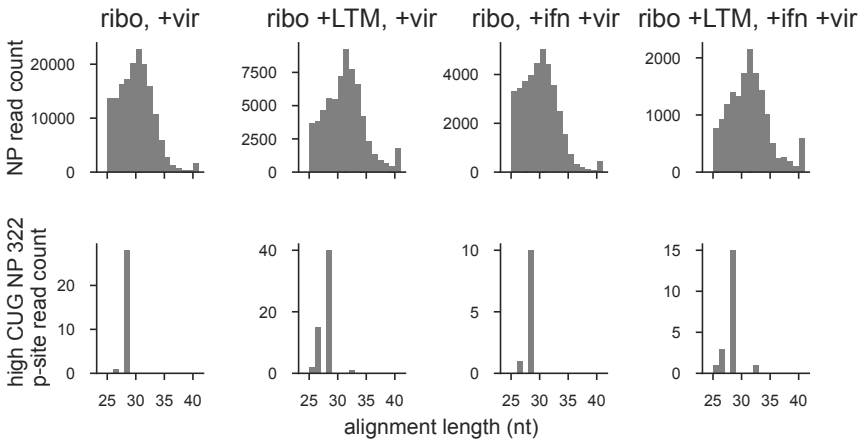


Figure 23: **Ribo-seq alignment length for position 322 of high CUG NP.** Distribution of alignment lengths for all NP reads and those with reads with P-site at position 322 of high CUG NP.

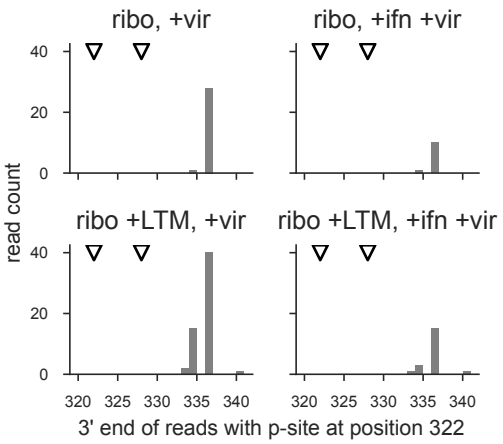


Figure 24: **3' end for reads mapping to postion 322 of high CUG NP.** For high CUG NP reads with p-site position at nucleotide 322 the distribution of 3' end of alignment is shown. Recoded CUG codons at position 322 and 328 are indicated with triangles.

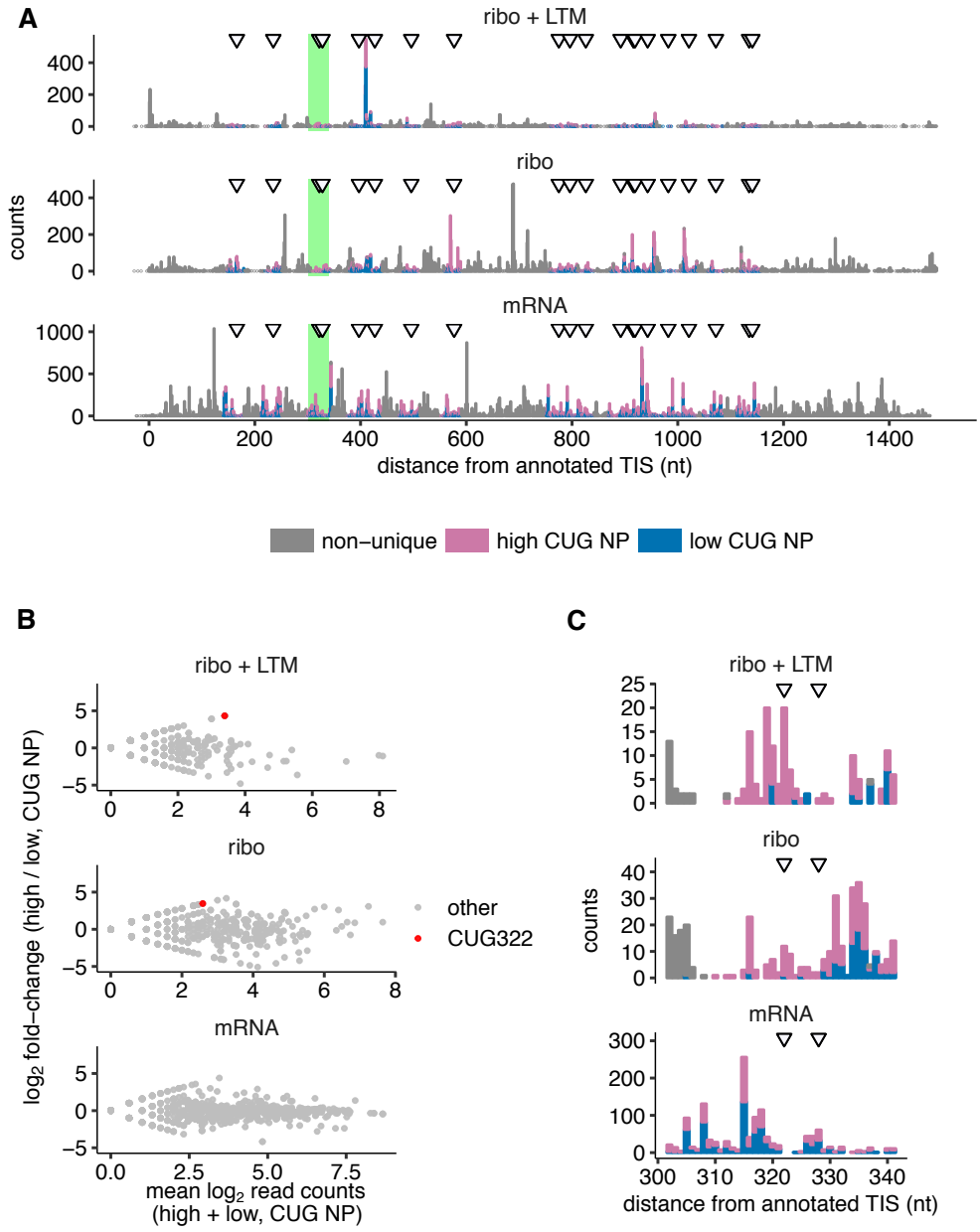


Figure 25: **Translation initiation at CUG codons in influenza nucleoprotein in +ifn +vir sample. (A)** Coverage of Ribo-seq + LTM, Ribo-seq, and RNA-seq reads that can be uniquely aligned to either the high CUG NP variant or the low CUG NP variant, and remaining non-unique reads. P-site counts are shown for Ribo-seq and Ribo-seq + LTM assays. 5end counts are shown for RNA-Seq. Data are plotted as a stacked bar graph. Locations of the 20 CUG codons that are present in high CUG NP and synonymously mutated in low CUG NP are indicated by arrows. **(B)** The ratio of high CUG NP to low CUG NP coverage from A is plotted against their sum along the horizontal axis. There were no RNA-seq reads with counts at 322, so this point is not highlighted. **(C)** The green-highlighted region in A around the CUG322 codon is shown at greater horizontal magnification. See Fig. 10 for +vir sample.

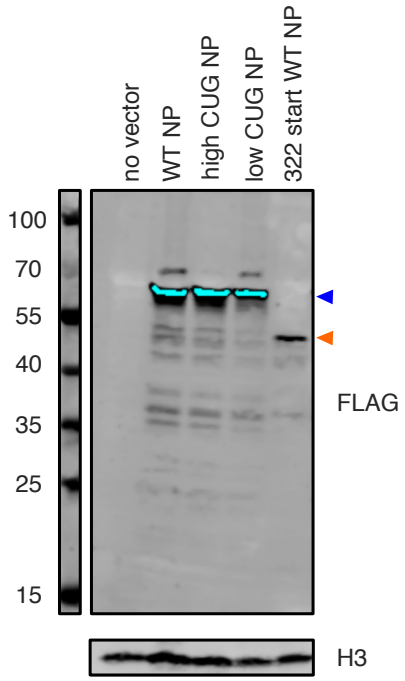


Figure 26: **Western blot to examine possible initiation at site 322 of high CUG NP.** Western blot of 293T cells transfected with the indicated NP protein expression constructs. "322 start WT NP" is a size control construct that begins at nucleotide 322 of WT PR8 NP. Top panel: anti-Flag; bottom panel: anti-H3. Blue arrow corresponds to full length NP and orange arrow corresponds to expected size of NP fragment due to initiation at nucleotide 322. The blot was overexposed to sensitively detect truncated peptides, leading to saturation of the the full length NP band (shown in cyan).

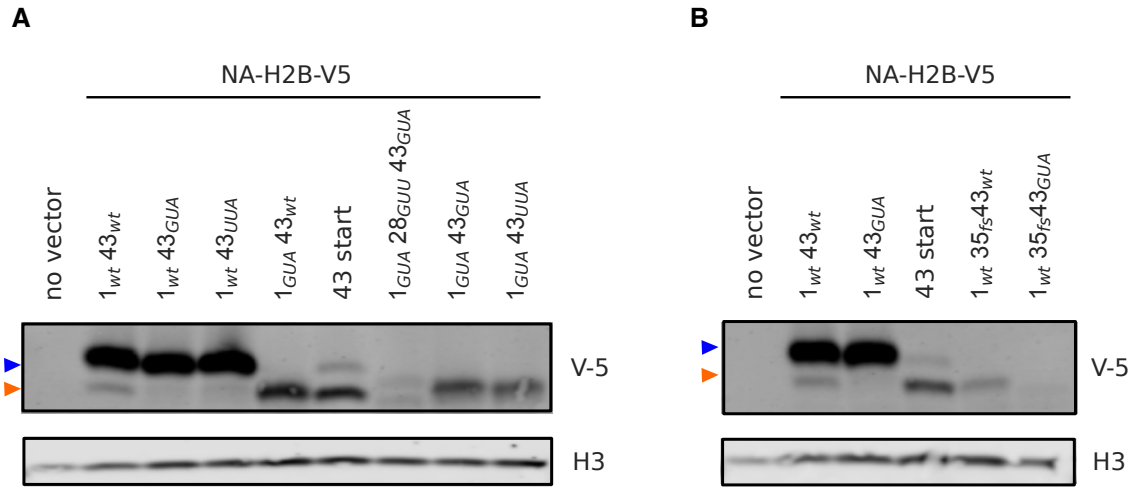


Figure 27: **Western blot with additional constructs to examine initiation at NA43. (A)** and **(B)** Western blot of 293T cells transfected with NA-H2B-V5 constructs with mutations at the canonical or downstream start site. “43 start” is a size control construct that begins at site 43 of NA. $1_{GTA}28_{GTT}43_{GTA}$ construct has a possible TIS (ATT codon) at position 28 mutated to GTT. $1_{GTA}35_{fs}43_{GTA}$ construct has a T inserted at coding nucleotide 35, such that any initiation 5’ to the insert should not be detectable with the V5 antibody. Top panel: anti-V5 (blue arrow corresponds to full length NA and orange arrow corresponds to NA43); bottom panel: anti-H3.

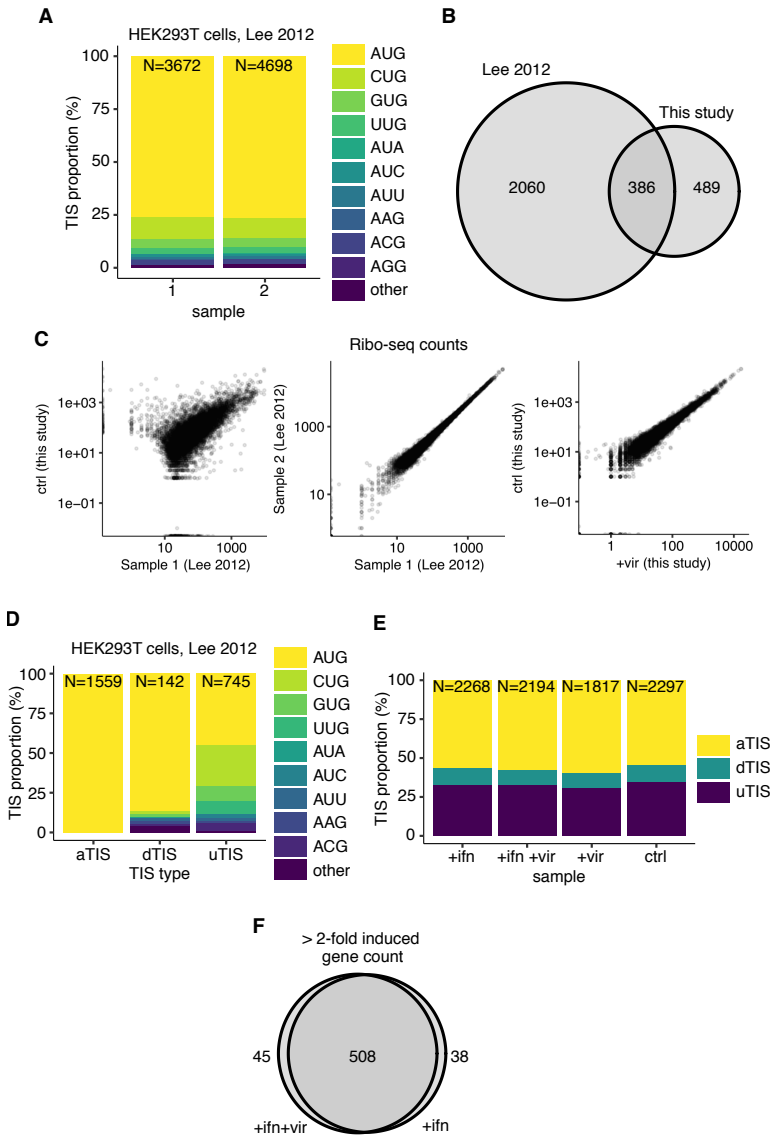


Figure 28: **(A)** Proportion of different near-cognate AUG codons (or other codons) overlapping with the called TIS in each of the two samples in [121]. N at the top of each bar indicates the total number of TIS called in each sample. **(B)** Overlap in high-confidence TIS between this study and Lee *et al.* [121]. High confidence TIS are the subset of TIS that are called across all samples in each study (2 in [121] and 4 in this study). **(C)** Comparison of Ribo-seq counts between samples in this study and from Lee 2012 [121]. Protein coding genes with at least 100 counts in one of the samples are plotted. **(D)** Proportion of different near-cognate AUG codons (or other codons) among the high-confidence TIS called in [121], stratified by TIS type. N at the top of each bar indicates the total number of high-confidence TIS of each type. **(E)** Proportion of different TIS types in each of the four samples used in this study. N at the top of each bar indicates the total number of TIS called in each sample. TIS not assigned to AUG or near-cognate AUG were excluded from this plot. **(F)** Overlap among the genes that are induced $>$2-fold upon either +ifn or +ifn +vir treatment with respect to the untreated sample. See Fig. 12 for definition of induced genes.

LIST OF TABLES

Table 1: Trunk substitutions for the small subset of NP epitopes where specific escape substitutions have been experimentally validated

Position in NP	Experimental evidence (reference)	No. of trunk substitutions for human influenza virus (substitution[s] and yr of isolation)
103	K103R escapes recognition by a HLA-B*1503-restricted T cell [15]	2 (K103R in 1981, R103K in 1997)
259	S259L escapes recognition by a HLA-B*4002-restricted T cell [15]	2 (L259S in 1972, S259L in 1990)
384	R384G abrogates MHC binding in HLA-B*2705 [213]	1 (R384G in 1990)
421	D421E escapes recognition by an HLA-B*3501-restricted T cell [22]	1 (D421E in 1977)
425	I425V escapes recognition by an HLA-B*3501-restricted T cell [22]	2 (I425V in 1975, V425I in 1999)