Within-host evolution of emerging and re-emerging influenza A viruses

By

Jorge M Dinis Jr.

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Microbiology)

at the

University of Wisconsin-Madison

2015

Date of final oral examination: 09/10/2015
Pending approval by the following members of the Final Oral Committee:
Thomas Friedrich, Associate Professor, Pathobiological Sciences
Tony Goldberg, Professor, Pathobiological Sciences
Andrew Mehle, Assistant Professor, Medical Microbiology and Immunology
David O'Connor, Professor, Department of Pathology
Robert Striker, Associate Professor, Medical Microbiology and Immunology
Marulasiddappa Suresh, Professor, Pathobiological Sciences

Table of Contents

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Tables

Chapter 2

Chapter 3

Chapter 4

Acknowledgments

Thanks. No seriously, thank you. Graduate school is hard. And without the amazing support of my super friends none of this would be possible. I could list you all out, but why? You know who you are, and how much I appreciate you. You're now apart of my family, which is kind of a big deal because it comes with fancy benefits. You need help moving, I got you. Need help drinking, no worries, I'm there. Need some advice, for whatever it's worth, there you go. Thank you for being the wind beneath my wings. #BetteMidlerReference #hashtaginmydissertation

To my two amazing brothers, damn, 4 years was a long time to be away from my best friends. Unfortunately, we can't get that time back, which is lame, but what we can do is live the rest of our lives full of joy, togetherness and have some fun along the way. My brain hurts, and I know your backs hurt, but hopefully we can pull it together and save the world. I love and appreciate you both.

Mom and Dad, words can't express how grateful I am to have you in my life. You've made me into the person I am today (for better and for worse). Your encouragement gave me the strength to keep pushing through all of life's challenges and I can't wait to metaphorically come back home (I'm never moving back in, ever).

To my wonderful wife Mallory, you are the most amazing person I have ever met. Not just because of middle, but because your strength, courage and constant pursuit of excellence has transformed my life. People like to say that a PhD is the ultimate personal achievement. It's cool, sure. But, it's not a personal achievement. You put more into this degree than anyone could imagine. I'm grateful for your love, friendship and support. You helped me get through this "shit show" they call graduate school. Lets see what the future holds, but if the past is any indication of the future, it's going to be a hell of an adventure. Abby is great too!

On my first day you told me, "I will not let you fail." Thomas Friedrich, I want to thank you for fulfilling your promise, for all the time, tough love and support you have

given me over the years. I can't wait to show the world the scientist and person you shaped me into. Let's go!

Thank you to my mentors for your guidance and advice: Dr. Tony Goldberg, Dr. Andy Mehle, Dr. David O'Connor, Dr. Robert Striker and Dr. M. Suresh, Dr. Karen Singmaster, Dr. Herbert Sibler, and Dr. Cleber Ouverney.

Abstract

Occasionally, influenza viruses emerge unpredictably from their natural avian reservoirs, or through an intermediate species, in human populations. Once in a human, avian influenza viruses can rapidly adapt to their new environments, which can result in efficient replication in and transmission between susceptible hosts. Sustained human-to-human transmission of "mammalian" adapted influenza viruses can result in global pandemic outbreaks. As humans build up immunity against pandemic viruses, the virus can antigenically evolve allowing for seasonal "re-emergence." Consequently, seasonal vaccines must be reformulated annually and no current vaccine generates broad enough immunity to protect against all potentially emerging pandemic viruses. Moreover, the evolutionary mechanisms that govern the emergence or re-emergence of influenza viruses are not well defined. Due to error-prone genome replication, influenza viruses exist within infected hosts as a diverse collection of viral variants often referred to as a "quasispecies." Within-host genetic variability allows influenza viruses to rapidly adapt to changing selective pressures. As part of strong interdisciplinary collaborations, our goal was to elucidate the virological and evolutionary mechanisms of influenza emergence and re-emergence in humans. Throughout these collaborative studies, my role was to develop novel experimental and computational approaches to accurately assess influenza within-host genetic variability directly from infected hosts. This contribution resulted in the development of an amplicon-based shotgun deep sequencing approach capable of detecting influenza genetic variation below the detection limit of traditional surveillance approaches. Using the combined power of deep sequencing, bioinformatics and virological characterization, we revealed previously unappreciated features of influenza biology regarding the role of low-frequency influenza variants during host adaptation and transmission in mammals.

In Chapter 2, we describe the impact of within-host viral genetic diversity on replication and transmission of H5N1 viruses using a ferret model. In these experiments

we found that even very-low-frequency H5N1 variants could transmit between animals via respiratory droplets, providing evidence that minor viral variants can cause onward infections in mammals. To build on these initial animal experiments, in Chapter 3 we describe H5N1 within-host genetic and functional diversity from infected human patients. We showed that viral genotypic and phenotypic diversity remained limited during infection and even after continuous replication in humans H5N1 variants predominately retained “avian-like” phenotypes. Taken together, these data suggest that avian influenza adaptation did not occur via constant incremental fitness increases within infected individuals. In Chapter 4, we assessed within-host diversity in humans infected with seasonal influenza viruses. Deep sequencing revealed low-frequency mutations previously associated with escape from virus-specific antibodies in non-vaccinated and vaccinated humans. Intriguingly, these potential antigenic variants did not reach fixation during infection, suggesting that other evolutionary constraints may be hindering their re-emergence in nature.

Together, this dissertation describes interrelated research projects that uncover the contribution of within-host diversity on the emergence and re-emergence of influenza virus in mammals. Ultimately, these projects measured the extent of influenza within-host diversity during natural infections and have begun to reveal the biological significance of low-frequency variants within the influenza viral quasispecies. We revealed new opportunities to improve global surveillance by including the detection of low frequency influenza variants that can contribute to host adaptation, immune evasion and transmission in mammals. Implementing deep sequencing approaches in influenza surveillance may dramatically improve our ability to detect the early emergence of potentially worrisome variants in nature. Finally, understanding the evolutionary mechanism of viral host adaptation and immune evasion may inform future predictions of zoonotic and pandemic emergence.

Chapter 1

Introduction

1. Influenza is an emerging and re-emerging disease

As of May 2015, 657 confirmed human cases of H7N9 influenza virus infection have occurred in China, resulting in 261 deaths (39% case fatality rate). Currently, little is known about how these viruses evolved to cross species barriers and sporadically infect humans. Recent reports have demonstrated that H7N9 viruses likely emerged directly from birds; however, they possess several characteristic mutations in the attachment protein hemagglutinin (HA) and the polymerase subunit PB2 that facilitate efficient replication in mammals [1]. Fortunately, H7N9 viruses have not yet acquired the capacity for sustained human-to-human transmission. However, recent investigations of avian H5N1 and H7N9 viruses have shown that a limited number of mutations can enable droplet transmission among mammals, heightening concern that a future pandemic could arise from the acquisition of a transmissible phenotype among avian viruses currently circulating in nature [2-5].

Over the past two and a half centuries, 10 to 20 human influenza pandemics have swept the globe [6]. Influenza viruses have emerged in the past as a direct result of a process termed antigenic shift. Antigenic shift results from genetic reassortment, made possible because influenza viruses have a segmented genome. This allows for gene transfer when two or more viruses coinfect a single host cell, generating viral progeny with novel combinations of gene segments. Such “mixing” may occur in the avian reservoir itself, in intermediate hosts such as pigs or poultry, or perhaps even in humans [66]. When antigenically novel HA genes are acquired by viruses capable of human-to-human transmission, a new pandemic may result. The emergence of antigenic novelty occurs at the level of an individual host, and the degree to which genetic variation contributes to this process may determine the rate at which shifted viruses can emerge.

In 1918, mankind experienced the worst influenza pandemic in recorded history, which caused approximately 20-50 million deaths worldwide [7]. Reconstruction of the

viral genome from the tissues of several victims has demonstrated that the causative agent was related to contemporaneous H1N1 viruses found in birds [8]. Despite extensive analysis of multiple available 1918 virus sequences, the origin of the pandemic virus, including timing of its emergence in humans and whether an intermediate host was involved, remains unresolved [9]. We do not yet understand the mechanisms that would convert an H7N9-like or H5N1-like localized outbreak to a 1918-like pandemic, but adaptation of avian viruses to efficient replication and transmission in humans is critical.

2. Pandemic potential of avian origin viruses

Aquatic birds are the natural reservoirs of all HA and neuraminidase (NA) subtypes [6,10] with the exception of two influenza-like viruses recently detected in bats [11]. Occasionally these avian viruses cross species barriers and cause sporadic infections in humans. During such spillover infections, avian influenza viruses may acquire efficient replication in mammalian cells and the capacity to transmit via respiratory droplets between mammals [12]. The evolutionary dynamics that govern the generation and outgrowth of mammalian transmissible avian influenza viruses in nature are not completely understood. Historically, influenza pandemics are caused by viruses against which human populations have limited pre-existing immunity as defined by serum antibodies against specific HA subtypes [13]. Recent serological surveillance studies found that humans have little to no pre-existing immunity against potentially emerging H5N1 or H7N9 viruses [14,15]. Currently, H5N1 or H7N9 subtypes do not have the capacity for sustained human-to-human transmission. However, mathematical models indicate that virus mutation rate, strength of natural selection and the number of viral replication cycles in mammalian cells can dramatically impact the chance in which a transmissible avian influenza virus is generated during individual spillover infections in humans [13,16]. In addition to being generated, the transmissible variant needs to exist in the source population at sufficient frequencies to facilitate airborne transmission.

We recently found that selective forces acting on the HA segment can impose a strong population bottleneck after respiratory droplet transmission of H5 influenza viruses [17]. Interestingly, viral variants present in as little as 5.9% of viruses were transmitted through this bottleneck, demonstrating that viruses may not need to reach “high” frequency in one individual to be transmitted to another [17]. Surprisingly, current surveillance methods, based on decades-old gene sequencing technology can only detect mutations if present in more than 20% of viruses infecting a host [31], suggesting that transmissible variants could be missed by traditional surveillance methods. Therefore traditional surveillance methods may fail to detect the genetic markers associated with virulence or transmissibility that may currently exist at low frequency in wild and domestic birds [16,32].

3. Re-emergence of seasonal influenza

In a sense, human influenza viruses “re-emerge” each season through a process called genetic drift. Antigenic drift involves the accumulation of point mutations in the envelope proteins HA and NA, which allow circulating viruses to escape antibody recognition with variable efficacy. For this reason seasonal vaccines must be reformulated each year, and even so, they provide as little as 0% effectiveness and up to 75% effectiveness against infection and illness [18]. Despite widespread vaccine availability, seasonal influenza A epidemics continue to cause an estimated 3 to 5 million cases of severe respiratory illness each year, resulting in approximately 250,000 to 500,000 deaths worldwide [19].

The relationship between antigenic “match” of the vaccine strain to the circulating viruses on vaccine efficacy is not well defined. One reason for this confusion is that current antigenic characterization relies upon standard hemagglutinin-inhibition (HI) assays, in which influenza-specific antibodies are detected in blood serum. Although it is relatively cheap and easy to perform, the HI assay has several important drawbacks. First, not all circulating viruses can be easily propagated through in-vitro

culture, a prerequisite for HI characterization. Secondly, viral propagation in vitro often yields viruses with amino acid substitutions in HA that may affect antibody binding [20]. Thirdly, HI assays use sera isolated from ferrets exposed once to influenza. This "narrow" immunity may not accurately predict protection against infection or disease by antibody responses in humans with multiple past-exposures [21]. Finally, although HI may be a surrogate for neutralization, HI assays likely do not always accurately reflect the neutralizing ability of antibodies. For all these reasons, standard HI-based antigenic characterization may fail to detect the circulation of influenza viruses bearing antigenic changes that could affect vaccine protection. Low vaccine effectiveness was associated with circulation of H3N2 viruses bearing mutations in HA with respect to the vaccine strain as determined by consensus sequencing. However, traditional serology indicated that these variants viruses were "well matched" to the vaccine strain [22]. Importantly, this study did not show a direct link between the detected mutations in HA and reduced viral sensitivity to vaccine-induced antibodies; instead it demonstrated that genetic characterization might reveal important antigenic changes in circulating influenza viruses that traditional serology fails to detect.

The ways in which influenza viruses evolve antigenically (i.e., by drift of shift) at the population level have been the subject of intense study in the past decade [23-25]. However, new antigenic variants are initially generated and selected at the level of individual infected hosts. Unfortunately, current influenza surveillance cannot resolve the extent of viral genetic diversity within individual hosts. Therefore, little is known regarding the role of within-host viral variation and the contribution of low-frequency genetic variants during the adaptive processes leading up to seasonal antigenic changes.

4. Influenza virus within-host genetic diversity

Influenza viruses exist in the host as a diverse collection of genetically linked variants that arise due to the combined effects of error-prone genome replication, rapid replication kinetics, and large population size [26,27]. Within-host viral population

structure will ultimately reflect a balance between the generation of diversity through mutation and the loss of diversity through selection. The resulting “swarm” of genetic variants can rapidly adapt in response to selective pressures. The rate at which influenza genetic diversity is generated within hosts and the degree to which it is maintained upon transmission are therefore two main parameters determining the likelihood with which influenza viruses emerge or re-emerge in nature. The evolutionary advantage of maintaining a diverse quasispecies is that when selective pressures rapidly change, a variant possessing a fitness advantage may already exist in the population [28]. Therefore, higher levels of within-host genetic diversity may therefore enhance the potential of RNA viruses to emerge in human populations [29,30].

5. Deep sequencing influenza quasispecies

Traditional sequencing methods used to characterize within-host viral diversity relied on cloning and subsequent Sanger sequencing [33], but with the development of deep sequencing technologies, Sanger sequencing is now obsolete. Deep sequencing omits the need for plasmid-based cloning while simultaneously producing thousands to millions of short sequences (reads) in hours, thus allowing the high-throughput screening of viral populations. Due to the continual development of longer sequence lengths and decreases in operational costs, deep sequencing approaches have rapidly increased in popularity [11,34-39].

Several deep sequencing platforms exit, but the Illumina MiSeq has the highest throughput, lowest error rate and the lowest cost per base when compared to other benchtop platforms that are currently on the market [40,41]. A critical step in deep sequencing viral genomes is generating platform-compatible nucleic acid libraries. Influenza viruses are encoded by a RNA genome, and therefore must be reversed transcribed into cDNA. One potential strategy for reverse-transcription uses primers that are completely degenerate (random hexamers) to both randomly prime the cDNA synthesis as well as to prime multiple strand displacement amplification (MDA) carried

out by the highly processive phix29 DNA polymerase [42,43]. The second method uses a universal influenza primer 5'-AGC AAA AGC AGG-3' that specifically primes DNA synthesis of all eight genome segments, which is then followed by multiple rounds of PCR using segment-specific influenza primers [44-46]. Nevertheless, these approaches have their advantages and disadvantages, deciding on which to use depends upon a priori information regarding the source sample (e.g., virus genome structure, viral titers, amount of sequences needed).

6. Reserach Focus

Although previous work has identified key molecular determinants of influenza immune escape, pathogenicity and transmissibility, little work has been performed to understand how potentially pandemic influenza variants emerge in individual human infections. In this thesis, I described a novel deep sequencing approach that I developed to characterize influenza viruses directly from biological samples (i.e., without passaging viruses in tissue culture). Together with a custom designed analytical pipeline, I assessed the role of within-host genetic diversity on influenza emergence and re-emergence in mammals. With a team of outstanding collaborators, we were the first to demonstrate that low-frequency influenza variants can be transmitted from one host to another. Furthermore, we showed that selective pressures acting during influenza transmission among mammals imposed a significant bottleneck that impacts the trajectory of viral evolution. And finally, I show that minor viral variants, below the detection threshold of global influenza surveillance, have important roles in host adaptation, transmission, and potentially in antibody escape.

Chapter 2

Selection on hemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses

Published, in part as

Peter R. Wilker*, Jorge M. Dinis*, Gabriel Starrett, Masaki Imai, Masato Hatta, Chase W. Nelson, David H. O'Connor, Austin L. Hughes, Gabriele Neumann, Yoshihiro Kawaoka and Thomas C. Friedrich

*Peter R. Wilker and Jorge M. Dinis contributed equally

Nature Communications, October 2013; DOI: 10.1038/ncomms3636

Abstract

The emergence of human-transmissible H5N1 avian influenza viruses poses a major pandemic threat. H5N1 viruses are thought to be highly genetically diverse both among and within hosts, but the effects of this diversity on viral replication and transmission are poorly understood. Here we use deep sequencing to investigate the impact of within-host viral variation on adaptation and transmission of H5N1 viruses in ferrets. We show that although within-host genetic diversity in hemagglutinin (HA) increased during replication in inoculated ferrets, HA diversity was dramatically reduced upon respiratory droplet transmission, where infection was established by only 1-2 distinct HA segments from a diverse source virus population in transmitting animals. Moreover, minor HA variants present in as little as 5.9% of viruses within the source animal became dominant in ferrets infected via respiratory droplets. These findings demonstrate that selective pressures acting during influenza virus transmission among mammals impose a significant bottleneck.

Introduction

Avian H5N1 influenza viruses sporadically infect humans with a lethality rate approaching 60% among confirmed cases, but they have not yet acquired the capacity for sustained human-to-human transmission. Recent work has demonstrated that a limited number of mutations can enable droplet transmission of H5N1 influenza viruses among mammals, heightening concern that a future pandemic could arise from the acquisition of a transmissible phenotype by H5N1 viruses currently circulating in nature [47].

The dynamics governing the emergence of pandemic influenza viruses are not completely defined. Historically, most influenza pandemics have been caused by viruses with HAs against which the human population had limited pre-existing immunity [48-51]. The spread of H5N1 viruses in wild birds and poultry on several continents and human H5N1 infections have focused attention on H5N1 viruses as potential sources of future pandemics. Despite widespread circulation of H5N1 viruses and significant human contact with infected poultry and birds, the number of documented human cases of H5N1 infection is relatively low, and human-to-human transmission of H5N1 viruses is rare [15,52]. The likelihood of an H5N1 pandemic in humans is largely determined by factors affecting the emergence of viruses that efficiently replicate in humans and transmit person-to-person [13,53-58].

Influenza viruses exist in the host as a diverse collection of genetically linked variants that arise due to the combined effects of error-prone genome replication, rapid replication kinetics, and large population sizes [26,27]. The within-host viral population structure reflects a balance between the generation of diversity through mutation and its loss through selection. The resulting “swarm” of genetic variants can rapidly adapt in response to selective pressures. The rate at which influenza genetic diversity is generated within hosts and the degree to which it is maintained upon transmission are therefore two main parameters determining the likelihood with which mammalian-trans-

missible viruses might emerge in nature.

We recently examined the molecular features that enable mammalian transmission of H5N1 influenza viruses. HA has a major role in restricting influenza virus host range, in part due to receptor specificity [59]. HA proteins of avian H5N1 viruses preferentially recognize sialic acid linked to galactose by α2,3-linkages (Siaα2,3Gal), whereas human influenza isolates preferentially recognize α2,6-linked sialic acid (Siaα2,6Gal) [60]. In previous work [47], we identified two amino acid substitutions (N224K and Q226L) that cooperatively enabled Siaα2,6Gal human-type receptor binding by an HA protein derived from the pathogenic avian virus A/Vietnam/1203/2004 (VN1203; H5N1). We created reassortant viruses bearing VN1203 HA genes encoding N224K and Q226L substitutions, with the seven remaining segments derived from A/California/04/2009 (CA04; H1N1), and evaluated their replication in ferrets. An additional substitution at HA amino acid position 158 (N→D) was associated with increased virus titers in ferret nasal turbinates. To evaluate transmissibility in ferrets, we therefore first used a virus isolate bearing all three mutations in HA (N158D/N224K/Q226L; herein called VN1203-HA(3)-CA04). In these experiments, "index" ferrets were inoculated intranasally with 106 plaque-forming units (p.f.u.) of virus stock. One day later, each infected ferret was paired with an uninfected "contact" ferret placed in an adjacent cage that prevented direct contact between the animals but permitted airborne droplet transmission of influenza virus. VN1203-HA(3)-CA04 was transmitted between animals in 2 of 6 ferret pairs. During replication of this virus in a contact animal, an additional T318I amino acid substitution was detected. The N158D/N224K/Q226L/T318I virus (herein called VN1203-HA(4)-CA04) had an improved transmission efficiency, being transmitted between animals in 4 of 6 ferret pairs [47].

Here we evaluate the impact of within-host viral genetic diversity on the replication and transmission of H5N1 reassortant viruses described in our previous study [47]. Using deep sequencing, we assess viral genetic variation during infection of inoculated

ferrets and in contact ferrets infected via respiratory droplet transmission. We report that HA segment diversity increases rapidly following intranasal inoculation of ferrets, resulting in a genetically diverse population of viruses in each infected host. In contrast, we show that there is a transmission-associated bottleneck in which only a limited subset of the HA segments are transmitted to contact ferrets. Furthermore, we find that even variants present at frequencies of as little as 5.9% in one infected animal can be transmitted via respiratory droplets. This report establishes that selective forces acting on the HA segment impose a significant bottleneck during respiratory droplet transmission of reassortant H5 influenza viruses.

Materials and Methods

Infection and transmission in ferrets

Studies of H5N1 reassortant virus transmission in ferrets were conducted previously [47]. We summarize the approach for those experiments here to aid the reader in understanding the design of the experiments from which we obtained the samples used for analysis in the present study. Ten-month-old female ferrets (Mustela putorius furo) were obtained from Triple F Farms (Sayre, PA, U.S.A.). “Index” ferrets were intranasally inoculated with 106 plaque forming units (p.f.u.) of the indicated virus in 500µl phosphate buffered saline (PBS). Viruses used were 7:1 reassortant viruses composed of the HA gene segment from A/Vietnam/1203/2004 (VN1203) H5N1 with the indicated mutations and the remaining seven gene segments from A/California/04/2009 (CA04) H1N1 virus.

For transmission experiments, ferrets were housed in adjacent transmission cages that prevented direct and indirect contact between animals but allowed spread of influenza virus through the air (Showa Science, Chiyoda-ku, Japan). Twenty-four hours after infection, one naive “contact” ferret was placed in a transmission cage adjacent to the index ferret. Nasal washes were collected from index and contact ferrets on day 1 after inoculation or co-housing, respectively, and then every other day subsequently (Supplementary Figure S2). Viral loads in nasal washes were determined by standard plaque assay on MDCK cells using serially diluted nasal wash fluid. Animal studies were performed in accordance with Animal Care and Use Committee guidelines of the University of Wisconsin-Madison.

Biosafety and biosecurity

All biosafety protocols, including those for isolation and sequencing of viral nucleic acids, were approved by the University of Wisconsin-Madison’s (UW’s) Institutional Biosafety Committee after risk assessments conducted by the Office of Biological Safety. In addition, the UW Biosecurity Task Force regularly reviews the research

program and ongoing activities of the UW Influenza Research Institute (IRI). The task force has a diverse skill set and provides support in the areas of biosafety, facilities, compliance, security and health. Members of the Biosecurity Task Force are in frequent contact with the principal investigator and personnel of the IRI to provide oversight and assure biosecurity. Isolation of RNA from samples containing H5N1 reassortant viruses was performed in enhanced BSL3 containment laboratories approved for such use by the CDC and the USDA following procedures approved by the UW Office of Biological Safety. RNA was isolated using techniques documented to inactivate virus particles in samples before removal from the BSL3 laboratory space. DNA library preparation and sequencing were performed in a BSL2 laboratory space.

Amplification of genomic material for deep sequencing

Total RNA was purified from nasal wash fluid using the RNeasy Mini Kit (Qiagen, U.S.A.). Viral RNA encoding the HA, NA and M gene segments were reverse transcribed using the Superscript III reverse transcriptase (Invitrogen, U.S.A.) according to manufacturer's instructions and the primer 5'-AGC AAA AGC AGG-3'. The resultant cDNA was used as template in a PCR reaction to amplify the HA, NA and M gene segments using the following primer pairs (see Supplementary Table S9 for primers) and the high fidelity iProof polymerase and buffers (BioRad, U.S.A.). PCR was performed by incubating the reaction mixtures at 94°C for 2 minutes, followed by 35 cycles of 94°C for 30 seconds, 62°C for 30 seconds, and 72°C for 2 minutes, followed by a final extension step at 72°C for 10 minutes. PCR products were separated by electrophoresis on a 1% polyacrylamide gel. The band corresponding to the full-length amplified gene segment was excised and the DNA recovered using the QIAquick Gel Extraction Kit (Qiagen, U.S.A.).

Illumina MiSeq Sequencing

Amplified gel-purified PCR products were quantified using the Qubit dsDNA High Sensitivity Kit (Invitrogen, U.S.A.) and diluted in DEPC-treated water to a final con-

centration of 2.5 ng µL-1. Samples were prepared for sequencing on the Illumina MiSeq platform using the Nexteratm DNA Sample Preparation Kit according to the manufacturer's instructions with slight modifications. Individual sample preparation reactions were performed for each HA amplicon, while the NA and M amplicons generated for each individual animal at each timepoint were combined and processed together.

The prepared DNA was purified with Zymo DNA Clean & Concentrator Spin Columns and eluted in 25 µL resuspension buffer. Dual DNA bar-codes (Epicentre, U.S.A.) were added to each sample reaction by limited-cycle PCR to enable multiplexed sequencing of prepared DNA samples. Limited-cycle PCR products were purified using a 0.375X AMPure XP bead cleanup (Beckman Coulter, U.S.A.) and eluted in 32.5 µL of the AMPure XP resuspension buffer. Eluted DNA was quantified using the Qubit dsDNA High Sensitivity Kit (Invitrogen, U.S.A.) and average fragment length was determined using the Agilent High Sensitivity Bioanalyzer Kit.

Fragmented and indexed samples were pooled in equimolar amounts into two separate 2 nM libraries for sequencing on the Illumina MiSeq. Prepared samples were split into two sequencing runs to ensure an adequate depth of coverage to detect low-frequency viral variants. The sample libraries were denatured into single-strand DNA by mixing with an equal volume of 0.1 N NaOH for 5 minutes, and were then diluted to 20 pM using the supplied HT1 buffer (Illumina, U.S.A.). Denatured 20 pM libraries were then diluted to 6 pM with 2% phiX control library added to run quality assurances. Six hundred µL was loaded into a 300-cycle reagent cartridge. Illumina MiSeq run settings were entered into sample sheets by Illumina Experiment Manager software v1.3.66 as the following: Workflow: DenovoAssembly; Assay: Nextera; Chemistry: Amplicon; and Reads: 160 x 160 with automatic adapter removal. Sequence information was stored as fastq-formatted data and used for further analysis.

Quality trimming and assembly of Illumina MiSeq data

Illumina MiSeq sequences were imported into CLC Genomic Workbench, Ver-

sion 5.1 (CLC bio, Denmark). Sequence reads were deconvoluted using the DNA indices that were introduced during the limited cycle PCR stage of sample preparation as described above. Reads were trimmed using a quality limit threshold of 0.001 and reads, regardless of mate pair, greater than 100 base pairs in length were retained. Sequence "reads" for each sample were mapped to a full-length HA, NA or M reference sequence.

Single Nucleotide Polymorphism detection

SNPs were called using CLC Genomic Workbench, Version 5.1 using all available sequencing data with at least 100 sequence reads covering each nucleotide position and a central base quality score of Q30 or greater. The Geneious bioinformatic software suite, Version 5.6.3 (Biomatters, Ltd., New Zealand) was used as an independent method of calling SNPs to ensure call reproducibility. Geneious variant calling occurred only at sites inside coding regions with read coverage greater than 100. For each analysis, SNPs occurring in only one sequence read, i.e., "singleton SNPs," were discarded. No minimum variant frequency threshold was used and approximate variant p-values were calculated.

Single Nucleotide Polymorphism detection limit validation

The M gene segment of the seasonal H1N1 virus A/Kawasaki/173/2001 was amplified by PCR using Phusion DNA Polymerase (New England Biolabs, UK). The resulting cDNA was cloned using the Zero Blunt Cloning kit (Invirogen, U.S.A.). The presence of the K173 M gene insert in plasmid DNA was verified by sequencing. The K173 M gene-containing plasmid was used as template for an in vitro transcription reaction using the MEGAscript T7 Kit (Life technologies, U.S.A.). K173 M gene transcripts were purified using phenol/chloroform extraction.

To evaluate the error rate of our library preparation and deep sequencing, we used the K173 M gene transcripts and the K173 M gene-containing plasmid as starting templates. The K173 M gene transcript was reverse transcribed using the Superscript

III reverse transcriptase (Invitrogen, U.S.A.) according to manufacturer's instructions and the primer 5'-AGC AAA AGC AGG-3'. The resultant K173 M gene cDNA and the plasmid-encoded K173 M gene were used as input template (100,000 copies of each template total) to amplify a 430 base-pair product using the high fidelity iProof polymerase and buffers (BioRad, U.S.A.) with the following primer pairs (see Supplementary Table S9 for primers). PCR was performed with an initial step at 94°C for 2 minutes, followed by 35 cycles of 94°C for 30 seconds, 62°C for 30 seconds, and 72°C for 2 minutes, followed by a final extension step at 72°C for 10 minutes. The amplicon was gel purified using the QIAquick Gel Extraction Kit (Qiagen, U.S.A.). Extracted DNA was prepared for sequencing using the Nexteratm XT DNA Sample Preparation Kit and sequenced on the Illumina MiSeq as described above. Quality trimming, assembly, and SNP detection of Illumina MiSeq data were performed as described above. SNP and mapping statistics were calculated using R version 2.15.1 (http://www.R-project.org/).

Enumeration of Linked SNPs within HA gene segments

The PERL programming language was used to design a novel method of mining paired-end sequence data for linked nucleotide polymorphisms. Our script, LinkGE was used to enumerate linked polymorphisms within a predefined query-set. Query nucleotide positions used for LinkGE were included if they met the following three criteria: (a) SNPs fell within a defined window of approximately 500 base pairs to accommodate analysis of paired end reads within the physical constraints of library fragment distribution and average sequence read length; (b) SNPs were not fixed, i.e., less than 100% of sequence reads were identical to each other; (c) SNPs were detected at or above 1% of all reads at any time point in any animal. Parameter files were independently generated for both transmission groups. CLC Genomic Workbench-generated assemblies were used as input data for the LinkGE script and the frequencies of each constellation of linked polymorphisms were determined. Constellations of linked polymorphisms, which we call "haplotypes", produced by LinkGE were manually confirmed within the

original assemblies. To further validate the identity of haplotypes and their frequencies within the bulk viral population as determined using LinkGE, the HA gene segments from samples collected at a single time point of two pairs of ferrets used in the transmission experiment were independently reverse transcribed, PCR amplified, and cloned using the Perfectly Blunt Kit (EMD Millipore, U.S.A.). The HA gene segments contained within individual plasmids were sequenced using a conventional Sanger chain-termination sequencing approach and compared to the results produced through analysis of paired-end deep sequencing data by LinkGE. The LinkGE source code is freely available on http://dholk.primate.wisc.edu/project/dho/public/LinkGe/begin.view.

Calculation of nucleotide diversity estimates

To measure nucleotide sequence diversity in HA, NA and M1 genes, we first used deep sequencing "reads" from each animal to estimate the number of synonymous substitutions per synonymous site (dS) and the number of nonsynonymous substitutions per nonsynonymous site (dN). The numbers of synonymous and nonsynonymous sites in each coding sequence were estimated following the method of Nei and Gojobori [61]. Sequence heterogeneity across entire gene segments was then estimated by computing synonymous nucleotide diversity (πS), defined as the mean of dS for all pairwise comparisons among a set of sequences, and nonsynonymous nucleotide diversity (πN), which is the mean of dN for all pairwise comparisons among a set of sequences. Note that this method compares sequence "reads" from an individual sample to each other and does not use a consensus or other external reference as a basis for comparison. Since most random amino-acid-changing mutations are likely to be disadvantageous, we expect that πN will equal πS under strict neutrality. If πN exceeds πS for a gene segment, this indicates that selection is acting to favor nonsynonymous mutations. We therefore used paired t tests to evaluate the hypothesis that πN = πS within genes, or that, e.g., πN of one gene equals πN of another.

Results

Low sequence diversity in transmissible H5N1 virus stocks

Here we use deep sequencing to investigate H5N1 influenza virus variation during replication and transmission in mammals, using archived RNA samples that were collected from ferrets during the previous study [47] (see Supplementary Figure S1 for ferret experiment outline). Note that no new virus transmission experiments were conducted for this study. We emphasize that these studies use avian-human reassortant viruses that encode an avian H5 HA protein in the background of a human H1N1 virus isolate that is likely already well-adapted to growth and transmission in mammals. While the HA protein plays a central role in the adaptation of avian influenza viruses to mammalian hosts, proteins encoded by other gene segments can also influence avian influenza virus adaptation and replication in mammals [13,54,58,62,63]. Studies using fully avian viruses may therefore yield results that differ from those described here. We use the Illumina MiSeq instrument in these experiments due to its high throughput and low error rate [40]. Data generated by this instrument are largely free of the homopolymer-associated indel errors which are common to other sequencing platforms [41].

We first characterized the nucleotide sequence diversity in the HA gene of VN1203-HA(3)-CA04 and VN1203-HA(4)-CA04 virus stocks, which were harvested after a single passage in Madin-Darby canine kidney (MDCK) cells (Figure 1). As expected, the mutations previously reported in association with the transmissible phenotype of these H5N1 viruses are present at or near 100% fixation (Figure 1). The majority of HA nucleotide diversity in both viruses is present at less than 1% per site (that is, for a given nucleotide position, fewer than 1% of sequence “reads” varied from the consensus residue). However, multiple nonsynonymous single nucleotide polymorphisms (SNPs) are present at frequencies between 1% and 99% in both stock viruses. VN1203-HA(3)-CA04 stock virus SNPs are located at nucleotide 53 (1.5%, encoding an alanine-to-threonine substitution), 215 (1.1%, encoding a valine-to-leucine substitution),

788 (3.2%, encoding an alanine-to-serine substitution), and 1020 (1.9%, encoding a threonine-to-isoleucine substitution). VN1203-HA(4)-CA04 stock virus SNPs are located at nucleotide 788 (4.4%, encoding an alanine-to-threonine amino acid substitution) and 1580 (6.6%, encoding an glutamate-to-lysine amino acid substitution). We also note a synonymous substitution at HA nucleotide 1018 in VN1203-HA(3)-CA04 (4.1%). Deep sequencing of NA and M segments from both stock viruses reveals few SNPs above 1% frequency (Supplementary Figure S2A and 2B).

HA sequence variation increases with time in index animals

The presence of detectable low-level sequence diversity in HA, NA and M prompts us to consider the potential impact of within-host viral variation on adaptive processes in ferrets. To follow changes in the viral population in inoculated ferrets and after transmission to naïve contact ferrets, we analyze viruses isolated from nasal wash samples collected from only the ferret pairs in which H5N1 reassortant viruses were transmitted between animals in our previous experiments [47].

To test for evidence of selective pressures favoring viruses bearing specific mutations in the avian HA segment, we evaluate all individual SNPs occurring at a frequency of ≥1% in stock virus preparations or in any single sample from one or more of the infected index ferrets (Figure 2). To confirm that SNPs detectable above this threshold were not due to sequencing or PCR errors, we deep sequenced plasmid DNA and an in-vitro transcription product encoding a fragment of the influenza virus M gene (see Methods for details). In these samples, nucleotide variation did not exceed 0.4% at any position, suggesting that our threshold of ≥1% ensures that only bona fide SNPs are considered in our analyses (Supplementary Figure S2C and D).

During infection of ferrets, viruses encoding mutations away from the reference sequence at nucleotide 788, a site at which we observed low-frequency SNPs in both stock viruses, reach 30% frequency or greater in each of the six index ferrets examined. A G→T SNP at nucleotide 788 encoding an alanine-to-serine amino acid substi-

tution is present in the two ferrets infected with VN1203-HA(3)-CA04, and a G→A SNP at nucleotide 788 encoding an alanine-to-threonine substitution is present in the four ferrets infected with VN1203-HA(4)-CA04 (Figure 2). The respective SNPs are present in the VN1203-HA(3)-CA04 stock virus at 3.2% and in the VN1203-HA(4)-CA04 stock virus at 4.4% (Table 1). Nucleotide 788 is the first position of codon 238 of the mature H5 HA protein (amino acid 242 by H3 numbering), and is located within the molecule's globular head. Each of these mutations creates a potential signal for glycosylation of the upstream asparagine residue at amino acid position 236 (240 by H3 numbering). The frequency of variant nucleotides at position 788 in index animals is positively correlated with day post-infection ($r = 0.857$; $P < 0.001$, Pearson's correlation coefficient), showing a significant increase in the variant nucleotide over time (Supplementary Figure S3). These results indicate that amino acid substitutions away from the consensus alanine at HA position 238 are strongly favored during replication of H5N1 reassortant viruses in index ferrets that transmitted viruses to their contacts. Although both wild-type and variant sequences at nucleotide 788 are clearly replication-competent in vivo, our data do not allow us to draw further conclusions about the action of selection on this position in contact animals. Future work will be needed to fully characterize the effects of amino acid substitutions at HA position 238 on the replication fitness of H5N1 viruses in mammals.

A number of additional SNPs change markedly in frequency during infection of index ferrets with the either virus (Figure 2). In some cases, SNPs that change in frequency during infection of ferrets are detectable at frequencies between 1% and 5% in viral stocks (e.g., SNPs at nucleotide 788 of VN1203-HA(3)-CA04 and VN1203-HA(4)-CA04 virus stocks in Figure 2A and 2B). Multiple SNPs in both viruses rapidly increase to frequencies above 20% in index animals (e.g., SNPs at nucleotides 738, 788, and 1020 in virus VN1203-HA(3)-CA04, Figure 2A, and at positions 736 and 788 in VN1203-HA(4)-CA04, Figure 2B), although variant SNPs decline in frequency in some

index animals at very late timepoints. Overall, the reproducible increase in frequencies of nonsynonymous SNPs at sites such as nucleotide 788 in multiple index animals suggests that selection favors variant amino acids at these positions during virus replication in vivo. In contrast, the SNP located at nucleotide position 1580 in the VN1203-HA(4)-CA04 stock virus decreases in frequency following infection of each of four index ferrets (Figure 2B; ferrets 13, 15, 17, and 21), indicating that selection does not favor sequence variants at this position in vivo.

Droplet transmission of a small number of HA variants

To determine whether selective pressures might be acting during respiratory droplet transmission of H5N1 reassortant influenza viruses, we first compared the frequencies of HA SNPs in viruses replicating in index and contact ferrets. As shown in Figure 2, multiple HA SNPs arise and achieve frequencies up to 80% during infection of index ferrets with either virus. If infection of the contact results from transfer of a representative sample of virus from the transmitting animal, one would expect to see similar HA SNP frequencies in the contact animal upon establishment of infection. Strikingly, however, HA sequences of the viral population replicating at the earliest detectable timepoint in contact animals do not reflect the frequency of multiple SNPs present in the paired index ferret near the time of transmission (Figure 3 and Supplementary Figure S4; see Supplementary Figure S1B for information on approximate timing of transmission). Instead, SNPs are either present at nearly 100% or almost completely absent following transmission, suggesting that infection of contact animals is established by a virus population with a relatively homogeneous HA sequence. Indeed, even SNPs present as minor variants in index animals are found dominating the population replicating shortly after transmission in contact animals. For example, a lysine-to-asparagine substitution encoded by a SNP at nucleotide 643 (amino acid 193 by H3 numbering) is detected in 5.9% of viruses in the pair 1 index animal shortly before transmission, and this same substitution is present in virtually 100% of virus sequences detected shortly

after transmission in the paired contact animal (Figure 3, pair 1).

The difference in HA sequence composition in contact ferrets compared to index ferrets following respiratory droplet-mediated transmission suggests that there is a severe bottleneck associated with transmission of these H5N1 viruses. To more closely evaluate the genetic composition of HA segments in the mixed viral populations in index and contact animals, we identified linkage relationships among SNPs detected during infection and after transmission. In this analysis, we take advantage of the fact that information about the physical linkage among SNPs of interest is preserved in a subset of sequence reads for each HA gene segment. This is possible because, in preparation for deep sequencing, amplified HA segments are randomly sheared into fragments with an average length of 500 nucleotides, which are sequenced from both ends. Therefore, the linkage of SNPs located in a defined window of approximately 500 nucleotides of the HA segment can be assessed in a subset of high-quality sequencing reads that cover the region of interest. Using this approach, we define the identity and frequency of various SNP combinations in HA segments present at each timepoint in the index and contact animals. We term each distinct combination of SNPs an HA "haplotype." We confirmed the validity of this approach by cloning and sequencing a panel of full-length HA segments from representative samples (Supplementary Table S1). In general, the number of HA haplotypes present in index animals increases throughout infection (Figure 4). By day 5 post-infection, no single HA haplotype accounts for more than 50% of the HA segments detected in the viral population of any index animal, irrespective of the infecting virus stock, indicating that the HA genes of both viruses rapidly diversify in vivo (Figure 4). While many of the identified HA haplotypes are detected in all index animals infected with either virus, some haplotypes are unique to particular animals, suggesting that stochastic processes, host genetics, and/or other factors might impact within-host viral evolution. Collectively, these data demonstrate that the assemblage and frequency of HA haplotypes are dynamic during

infection of index animals with these H5N1 viruses, with a trend toward increasing HA haplotype diversity.

The diversity of HA haplotypes in index ferrets is not reflected in the virus population found in contact animals following transmission (Figure 4, Supplementary Tables S2 and S3). Instead, a single predominant HA haplotype is detected in each of the six contact ferrets at the earliest timepoint at which we recovered virus (Figure 4, Supplementary Figure S1B). In VN1203-HA(3)-CA04-infected contacts, infection is established by a relatively monomorphic population of viruses possessing a single HA haplotype, although the specific transmitting HA haplotype differs between the two ferret pairs (Figure 4A). Similar patterns are found in ferret pairs infected with VN1203-HA(4)-CA04. In contrast to the heterogeneous mixture of HA haplotypes present in index ferrets, the first samples of virus collected from contact animals contains a single detectable HA haplotype (Figure 4B, pair 8 and pair 11), or a single predominant HA haplotype with a second detectable haplotype present at a low frequency (Figure 4B, pair 7 and pair 9, day 5). It is unclear if the minority HA haplotypes detected early in contact ferrets are transmitted from the animals' paired index ferrets or are derived from the more prevalent HA haplotype following the establishment of infection.

Using the same approach, we also identify NA and M1 gene haplotypes during infection and after transmission (see Supplementary Tables S4 and S5 for NA and Supplementary Tables S6 and S7 for M1). While multiple NA and M1 haplotypes are defined, the virus population in index ferrets is dominated by a single NA and M1 haplotype in each animal, with 2-5 minor variant haplotypes present at 1-5% (Supplementary Tables S4 and S6, respectively). In contrast to our observations for the HA segment, there is no narrowing of NA or M1 haplotype diversity after transmission. Together these data show that there is a severe bottleneck in HA segment diversity associated with the transmission of H5N1 influenza viruses via respiratory droplets. As a result, infection is established in the contact host by a population of viruses with a single

predominant HA haplotype that resembles only one of multiple haplotypes generated during viral replication in the infected transmitting animal.

Transmission bottleneck is associated with selection on HA

At least two non-mutually-exclusive processes could cause a severe bottleneck during transmission. First, a low infectious dose could account for the diminution in HA sequence diversity during transmission if only one or a few viral particles establish infection in the new host, a situation commonly referred to as the founder effect. In this situation, natural selection would not act to favor the transmission of particular viruses, and we would expect to observe low genetic diversity in all viral gene segments immediately after transmission. Second, strong selective pressures could favor transmission of viruses possessing particular HA haplotypes that confer improved transmission efficiency. In this situation, termed a "selective sweep," diversity is reduced as natural selection eliminates viruses that are poorly adapted for transmission [64]. In a selective sweep acting on HA, we would therefore expect transmitted viruses to show a greater reduction in HA genetic diversity as compared to diversity in other gene segments.

To determine the relative impact of these two processes, we measure genetic diversity across entire gene segments by computing the statistics πN and πS (for details, see Methods). πN, or nonsynonymous diversity, describes the frequency of mutations within a virus population that encode an amino acid change. Similarly, πS, or synonymous diversity, describes the frequency of silent mutations. Comparing these statistics for a given gene segment provides information about the "direction" of natural selection. Generally, a $\pi N/\pi S$ ratio >1 indicates that positive selection is favoring genetic diversification. By contrast, a $\pi N/\pi S$ ratio <1 indicates that purifying (or negative) selection is acting to maintain a "fit" consensus sequence by removing deleterious mutations. Performing pairwise comparisons of πN or πS across gene segments (e.g., comparing statistics for HA and NA within one virus sample, or for the HAs of two different viruses) provides information about the relative genetic diversity of any pair of

gene segments. It is important to reiterate here that our experiment uses a reassortant virus expressing an avian HA, but all other viral gene segments from a mammalian-adapted pandemic H1N1 virus; the results of these analyses may therefore differ when applied to fully avian viruses.

We compute overall πN and πS in each animal using combined sequence data from every available timepoint. In HA, overall means of πN and πS do not differ significantly either in index animals or in contact animals when all timepoints are considered together (Figure 5 A, B; compare dark blue and light blue or dark red and light red bars). This result suggests that, although positive selection is likely acting at specific sites, such as nucleotide 788 in index animals, it is not driving diversification throughout the HA gene. In NA and M1, overall mean πS is often higher than mean πN in both index and contact animals (Figure 5A, B; compare dark blue and light blue bars or dark red and light red bars). This result indicates that nonsynonymous mutations in NA and M1 are generally removed by purifying selection and therefore remain at low frequencies in the virus population in index animals. Together these data suggest that different selective pressures act on HA, as compared to NA and M1, during replication and transmission of reassortant H5N1 viruses in ferrets.

In our initial analyses, the average values for πN and πS in HA in index ferrets were higher than those in contact ferrets, but these differences did not attain statistical significance for every pairwise comparison (Figure 5A and B). To more closely examine the impact of transmission on HA genetic diversity, we therefore compare values for πN and πS in HA in index and contact animals, considering only the timepoints closest to the transmission event. Since only two ferret pairs were infected with VN1203-HA(3)-CA04, we cannot perform statistical analyses on that virus alone. We therefore first consider only animals infected with VN1203-HA(4)-CA04 (Figure 5C, left side of panel). In this group, mean πN for the transmitted virus in contact ferrets (0.00008 ± 0.00002) is significantly lower than that found in the virus replicating in index ferrets just before

transmission (0.00108 ± 0.00014; P = 0.008, Student t-test). Likewise, near the time of VN1203-HA(4)-CA04 transmission, mean πS is lower in contact than in index animals, although in this case the difference is not statistically significant. Considering all six ferret pairs regardless of infecting virus, mean πN for the transmitted virus in contact animals (0.00012 ± 0.00013) is significantly lower than mean πN in the index ferrets near the time of transmission (0.00124 ± 0.00014; P <0.001, Student t-test; Figure 5C, right side of panel, compare dark blue to dark red bars). Similarly, the mean πS value of contact animals is lower than that of index animals near the time of transmission, although this difference narrowly escapes statistical significance (P = 0.054, Student t-test; Figure 5C, right side, compare light blue to light red bars). Interestingly, when we considered the HA nucleotide diversity in index animals just prior to the estimated time of transmission, the mean of πN for the four index ferrets infected with VN1203-HA(4)-CA04 (0.00108 ± 0.00014) was significantly greater than mean of πS for the same four animals (0.00042 ± 0.00012; P = 0.004, Student t-test; Figure 5C, left side of panel, compare dark blue to light blue bars), suggesting that HA is undergoing positive selection during infection of index animals. Together, these results confirm that transmission of these H5N1 viruses to contact animals is associated with a significant reduction in overall HA nucleotide diversity.

We also compute an overall πN and πS for HA, NA and M1 in the stock viruses, finding that πN and πS is in general lower than the values seen in the viruses infecting all 6 index animals (Supplementary Table S8). In the case of HA, both πN and πS are significantly lower in stock viruses than in virus from index animals (P < 0.01 in each case, Student t-test; Supplementary Table S8). In the case of NA and M1, πN and πS are lower in stock viruses than in viruses from the index animals; but the differences are not statistically significant (Supplementary Table S8). Additionally, πN is lower than πS for the HA, NA, and M1 genes of each stock virus, providing evidence that positive selection on HA sequences did not occur during production of the virus stocks. To-

gether, our data therefore show that there is a dramatic reduction in genetic diversity in HA, but not NA or M, following droplet transmission of H5N1 reassortant viruses, while there is no evidence for a similar reduction in diversity following direct inoculation of index animals with stock viruses.

Discussion

Acquisition of a human-transmissible phenotype by H5N1 avian influenza viruses represents a major pandemic threat [47,65-68]. Despite the identification of specific mammal-adapting mutations in the HA genes of H5N1 viruses [47,65], the strength and nature of evolutionary barriers to such adaptation remain unclear [16]. Importantly, few models have measured the impact of within-host "quasispecies" diversity on influenza virus evolution. Here we use deep sequencing to evaluate the impact of viral variation on H5N1 influenza virus host adaptation and mammalian transmission dynamics. Our analyses show that selection on HA plays a main role in driving a bottleneck during transmission of these reassortant H5N1 influenza viruses among mammals. Together, our results suggest that selection could drive establishment of infection in mammals by viruses bearing "favorable" HA genes, even when such viruses are in the minority in the source population.

Restriction of influenza viral genetic diversity in contact animals likely reflects the various mechanical and immunological barriers to replication that viruses must overcome upon mucosal transmission. This bottleneck could result from the founder effect, i.e., a dramatic reduction in effective viral population size without the action of natural selection, and/or from a selective sweep, in which viruses with specific traits are best able to transmit. Influenza virus quasispecies dynamics during transmission have not been well characterized, but transmission of other diverse RNA viruses has been shown to dramatically reduce quasispecies diversity. In HIV transmission, infection via the cervicovaginal mucosa is initiated by only 1-2 virus clones [69-71]. This bottleneck may be due largely to the founder effect, although HIV variants bearing envelope proteins capable of using CCR5 as a co-receptor for cellular entry appear to be favored [71]. Transmission of hepatitis C virus (HCV) is also associated with a genetic bottleneck, which may result from a selective sweep acting on the viral envelope glycoprotein [72,73]. By controlling the viral dose administered to naïve ferrets via aero-

solized droplets, Gustin and colleagues demonstrated that influenza virus infection can be established by a dose as small as 4 p.f.u., suggesting that the founder effect could reduce viral diversity during influenza transmission by respiratory droplets [74]. Additionally, the anatomic location of replication of variant virus populations in the source host may affect the "availability" of specific virus variants for transmission, providing a further potential mechanism for founder effects to influence the diversity of viruses in individuals by respiratory droplets. In our experiment, however, HA diversity was reduced to a much greater extent than diversity in genes encoding NA or M1, suggesting that a selective sweep acted to favor transmission and/or replication of only a subset of HA sequences from index animals in contacts infected by respiratory droplets. While our results highlight a role for natural selection in determining the composition of HA sequences infecting contact animals, the potential impact of extremely low infecting doses or other factors in the bottleneck associated with transmission cannot be excluded.

Although we detect only 1-2 HA "haplotypes" early after infection in contact animals, no single haplotype was consistently associated with transmission (Figure 4). Perhaps this is because the main determinants of mammalian transmissibility of these reassortant viruses are the mutations identified in our previous study [47], which were fixed in the virus populations. The haplotypes we identified therefore existed in a background of fixed mutations that already facilitated mammalian transmission of these viruses. Our results suggest that additional amino acid substitutions may render individual viruses with this genetic background more or less fit for transmission. The specific HA haplotype that establishes infection in naïve contact ferrets may vary depending on a number of factors, including the unique constellation of viruses that co-exist in the inoculated ferret due to random mutation and intrahost selective pressures during infection, the titer of each variant virus in the inoculated ferret as the infection progresses, the anatomic location at which each variant virus is replicating, and the timing of

expulsion of respiratory droplets as it relates to the dynamically changing population of viruses in the inoculated ferret. Further work will therefore be required to understand the nature of selection pressures acting to restrict HA sequence diversity during H5N1 virus transmission among mammals.

Replication and transmission of purely avian H5N1 viruses in mammals may result in patterns of selection that differ from those we have observed here using a reassortant virus composed of altered versions of the HA gene segment from an H5N1 virus and seven remaining gene segments from a human H1N1 virus. Notably, adaptive mutations in gene segments other than HA can also affect the ability of purely avian viruses to productively infect mammalian cells [54, 55]. In particular, a lysine residue at PB2 amino acid 627 has consistently been associated with enhanced replication of avian-origin influenza viruses in mammals [58,62,63,65,75-77]. In our study, we observed very little diversification of the NA and M1 genes during infection of index or contact ferrets (Supplementary Tables S4 and S6) when compared to the H5 HA, perhaps because in our reassortant viruses these gene segments were already well adapted to replicate in mammals. The signatures of purifying selection we observed in these segments (i.e., levels of synonymous diversity that were higher than levels of nonsynonymous diversity) are consistent with this interpretation.

RNA viruses are characterized by high mutation rates, leading to the accumulation of deleterious mutations, while their short replication times and frequently large effective population sizes act to increase the efficacy of purifying selection; that is, as virus titers increase, so does the likelihood that purifying selection will act to remove deleterious mutations from the population [64,78,79]. Consistent with this model, in HA genes we detected only a small number of sites at which nonsynonymous substitutions accumulated to detectable levels. Positive selection may have acted on these particular sites to enhance viral fitness. For example, we found that low-frequency SNPs encoding alanine-to-serine or alanine-to-threonine substitutions at H5 amino acid 238

(residue 242 by H3 numbering) in the virus stocks rapidly increased in frequency in all 6 index animals we examined. Each substitution creates a potential site of N-linked glycosylation within the HA globular head, though the impact of glycosylation at this site (H5 amino acid 236; H3 amino acid 240) has not been characterized. Notably, we found that viruses encoding a serine or threonine at position 238 were not consistently transmitted to contact animals, despite their high frequency in index animals. We speculate that the biological function of serine/threonine 238 may enhance virus replication within mammals, but does not favor transmission between hosts. Together these observations suggest that different viral characteristics, such as transmissibility and replication, may have distinct effects on evolutionary fitness, so that, e.g., mutations providing optimal advantages for transmission may exact a cost to replicative capacity.

Finally, our findings also suggest that influenza surveillance efforts based on Sanger sequencing may fail to detect the early emergence of genetic markers associated with transmissibility or virulence in mammals, as others have recently speculated [16]. The use of Sanger sequencing for influenza surveillance typically defines consensus sequences, cannot resolve variants present below 20% of the viral population, and cannot provide information regarding the genetic linkage of variant nucleotides. As described above, our deep sequencing revealed substantial diversification of HA haplotypes during infection of index animals. In some instances, the transmitted HA variant was present at frequencies as low as 5.9% in the source animal near the time of transmission, a level below the ability of population-based Sanger sequencing to resolve SNPs [31]. Our results therefore demonstrate that low-level viral variants in the source viral population can nonetheless found infections in new hosts. This finding has important implications for surveillance activities aimed at detecting naturally occurring variants that may have the ability to replicate in and transmit among mammals. Importantly, Sanger sequencing may not only fail to detect biologically relevant viral species in a mixed population, but may define a viral consensus sequence that does not exist

in nature. Deploying deep sequencing approaches in surveillance may therefore dramatically enhance our understanding of influenza virus population diversity in reservoir hosts.

Acknowledgments

This work was supported by a supplement to National Institutes of Health grant RR000167 (now OD011106) awarded to TCF and the Wisconsin National Primate Research Center; grant PRJ29JN awarded to TCF and DHO and grant AI 077376 awarded to ALH and DHO. JMD was supported by National Science Foundation Graduate Research Fellowship DGE-0718123. YK gratefully acknowledges support from a grant-in-aid for Specially Promoted Research and from a contract research fund for the Program for Funding Research Centers for Emerging and Reemerging Infectious Diseases from the Ministries of Education, Culture, Sports, Science, and Technology, and from grants-in-aid of Health, Labor, and Welfare of Japan, by ERATO (Japan Science and Technology Agency), and from National Institute of Allergy and Infectious Diseases Public Health Service Research grants.

Figure Legends

Figure 1. Deep sequencing reveals low sequence variation in hemagglutinin of influenza virus stocks. We used deep sequencing to probe the nucleotide diversity of the (A) VN1203-HA(3)-CA04 and (B) VN1203-HA(4)-CA04 virus stocks used in mammalian transmission experiments. Individual HA sequence reads were mapped to a consensus HA sequence derived from the isolate A/Vietnam/1203/2004 (H5N1). SNPs were enumerated as described in Methods. No variants were detected below 0.01%. The frequency of variants at each site is presented as either closed circles (nonsynonymous substitutions) or open squares (synonymous substitutions). Mutations previously reported in association with mammalian transmission are highlighted in red: VN1203-HA(3)-CA04: N158D (nt 536), N224K (nt 736) and Q226L (nt 741) and VN1203-HA(4)-CA04: N158D (nt 536), N224K (nt 736), Q226L (nt 741) and T318I (nt 1020). A synonymous mutation at nucleotide position 1555 that was not present in the plasmid used to generate viruses by reverse genetics was detected in each virus stock after harvest and before infection of ferrets.

Figure 2. Within-host selection of HA segments harboring specific single nucleotide polymorphisms. Viral RNA recovered in nasal wash samples collected from ferrets at different timepoints following intranasal infection with the indicated viruses was used to measure HA segment variation by deep sequencing. Bar graphs depict changing frequencies of specific SNPs during infection of index ferrets with (A) VN1203-HA(3)-CA04 or (B) VN1203-HA(4)-CA04 viruses. Number of sequences used to calculate SNP frequencies ranged from n = 198 to 15206 for VN1203-HA(3)-CA04 and n = 111 to 11411 for VN1203-HA(4)-CA04. This analysis focused on SNPs detected in at least 1% of virus sequences in stock viruses or in one or more samples collected from any ferret at any time point. Each SNP was nonsynonymous, with the exception of a synonymous SNP at nucleotide position 1018. Inset line graphs depict virus titers

in nasal wash samples collected from each index ferret at the indicated timepoint. Virus titers were measured using a standard plaque assay on MDCK cells.

Figure 3. Detection of HA SNPs early after infection in contact animals. HA gene segments accumulated diversity over time during replication in index animals, as demonstrated by the increasing frequency of substitutions at positions 643, 788, 1018 and 1020 in index animals infected with VN1203-HA(3)-CA04 (left-hand portion of each panel). Following transmission, the founding virus population in contact animals displayed a shift in SNP frequencies, such that SNPs were either nearly fixed in, or were absent from, the replicating virus population. A SNP detected in 5.9% of viruses in the pair 1 index animal shortly before transmission was present in nearly 100% of virus sequences shortly after transmission in the paired contact animal. Number of sequences used to calculate SNP frequencies ranged from n = 1472 to 13324.

Figure 4. Enumeration of HA segment "haplotypes". To identify patterns of physically linked SNPs, we took advantage of the fact that paired-end deep sequencing provides "mate-paired" reads that are separated by intervening sequences of varying length, allowing us to identify reads containing sites of interest that are linked on the same viral RNA. By analyzing these reads, we identified linkage relationships among targeted SNPs, and use the term "haplotype" to denote a single unique combination of SNPs. SNPs used to define HA haplotypes are shown schematically above each panel. This analysis targeted nucleotides 728, 738, 744, 788, 1018 and 1020 in virus VN1203-HA(3)-CA04 (panel A) and nucleotides 494, 496, 557, 736, 754, 778 and 788 in virus VN1203-HA(4)-CA04 (panel B). We considered only "haplotypes" detected at or above a frequency of 1% of the total virus population. Within each panel, grey boxes indicate a transmission pair, each with an index animal (above) and a contact animal (below). The x-axis represents the sample collection timepoints for index or contact animals.

Grey bars denote the frequencies of non-transmitting haplotypes. The frequencies of HA haplotypes implicated in transmission are colored with the specific constellation of SNPs indicated in the schematic above each panel. Note that the minor haplotype detected at the first timepoint in which virus was recovered from the contact ferret of pairs 7 and 9 was also found in the paired index ferret. Number of sequences used to calculate haplotype frequencies ranged from n = 1418 to 3128 for VN1203-HA(3)-CA04 and n = 540 to 1050 for VN1203-HA(4)-CA04. Further details can be found in Supplementary Table S2 and S3.

Figure 5. Within-host nucleotide diversity in HA. We determine mean nonsynonymous (πN) and synonymous (πS) nucleotide diversity throughout the experiment in the HA, NA and M1 coding regions of viruses isolated from index and contact ferrets infected with (A) VN1203-HA(3)-CA04; n = 2 or (B) VN1203-HA(4)-CA04; n = 4. (C) To independently assess the impact of transmission on HA nucleotide diversity, we compared πN and πS at the single timepoint closest to transmission for each ferret pair. For this analysis we considered either the VN1203-HA(4)-CA04-infected group alone (panel C, left) or all 6 ferret pairs together (panel C, right). In each graph, vertical bars represent the mean nucleotide diversity for all index and contact samples; error bars represent standard error of the mean (s.e.m.). Dark and light blue bars indicate πN and πS, respectively, in index animals; dark and light red bars indicate πN and πS, respectively, in contact animals. We used paired t-tests to compare πN and/or πS values within and between gene segments. Horizontal bars highlight comparisons for which two values are significantly different. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

Tables, Figures and Supplementals

Table 1. Single nucleotide polymorphisms detected in HA segments of viral populations recovered from infected ferrets

Virus*	Nucleotide position (VN1203 Reference)	Nucleotide change	Amino acid position (H3 numbering)	Amino acid change	Frequency in HA(3)/HA(4) stock viruses	Potential function, if known
VN1203-HA(4)-CA04	204	A → G	53	D → G	0%	
Both	339	A → G	n/a	D → G	0% / 0%	
Both	496	G → T	144	K → N	0% / 0%	
Both	536	A → G	158	N → D	99.8 / 100%	Loss of glycosylation site
VN1203-HA(4)-CA04	557	A → G	165	K → E	0%	
Both	643	G → T	193	K → N	0% / 0%	
VN1203-HA(3)-CA04	728	A → C	222	K → Q	0%	Receptor recognition
Both	736	C → A	224	N → K	99.8% / 99.8%	Recognition of human-type receptors
Both	738	G → A	225	G → E	0% / 0%	Receptor recognition
Both	741	A → T	226	Q → L	100% / 100%	Recognition of human-type receptors
VN1203-HA(3)-CA04	788	G → T	242	A → S	3.20%	
VN1203-HA(4)-CA04	788	G → A	242	A → T	4.40%	
VN1203-HA(4)-CA04	956	A → G	297	I → V	0%	
VN1203-HA(3)-CA04	1018	G → A	317	syn	4.10%	
Both	1020	C → T	318	T → I	1.9% / 99.9%	
VN1203-HA(3)-CA04	1144	T → C	n/a	syn	0%	
VN1203-HA(4)-CA04	1375	A → G	n/a	syn	0%	
VN1203-HA(4)-CA04	1580	G → A	n/a	E → K	6.60%	

* Viruses were 7:1 reassortants with an HA gene segment derived from A/Vietnam/1203/2004 and the remaining segments derived from A/California/04/2009 (pH1N1; CA04).
n/a indicates H3 amino acid numbering not applicable, since residue is outside mature HA.
syn indicates synonymous nucleotide change.

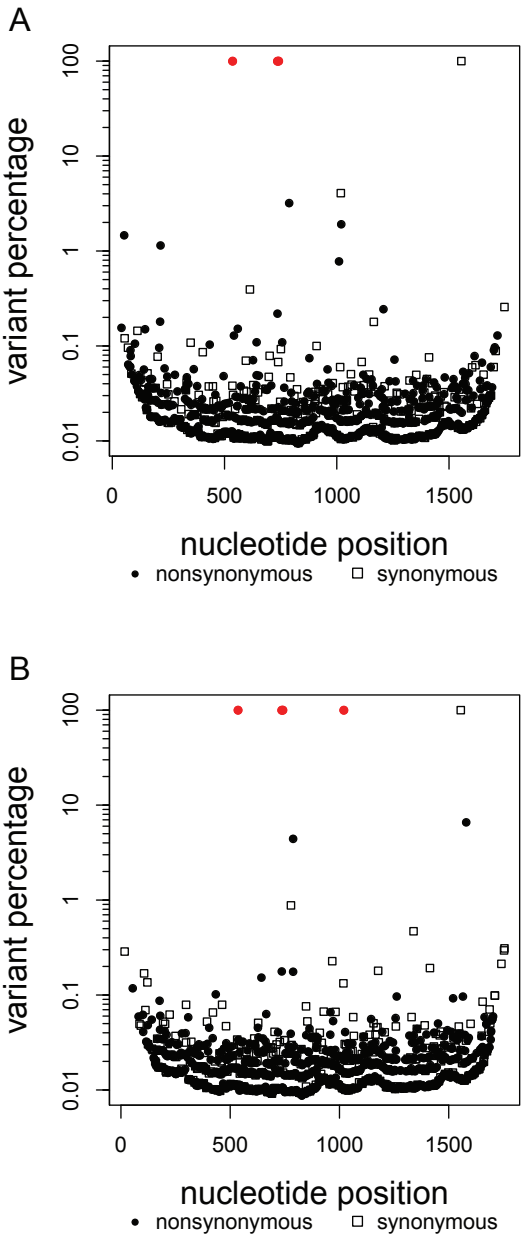


Figure 1. Deep sequencing reveals low sequence variation in hemagglutinin of influenza virus stocks.

We used deep sequencing to probe the nucleotide diversity of the (A) VN1203-HA(3)-CA04 and (B) VN1203-HA(4)-CA04 virus stocks used in mammalian transmission experiments. Individual HA sequence reads were mapped to a consensus HA sequence derived from the isolate A/Vietnam/1203/2004 (H5N1). SNPs were enumerated as described in Methods. No variants were detected below 0.01%. The frequency of variants at each site is presented as either closed circles (nonsynonymous substitutions) or open squares (synonymous substitutions). Mutations previously reported in association with mammalian transmission are highlighted in red: VN1203-HA(3)-CA04: N158D (nt 536), N224K (nt 736) and Q226L (nt 741) and VN1203-HA(4)-CA04: N158D (nt 536), N224K (nt 736), Q226L (nt 741) and T318I (nt 1020). A synonymous mutation at nucleotide position 1555 that was not present in the plasmid used to generate viruses by reverse genetics was detected in each virus stock after harvest and before infection of ferrets.

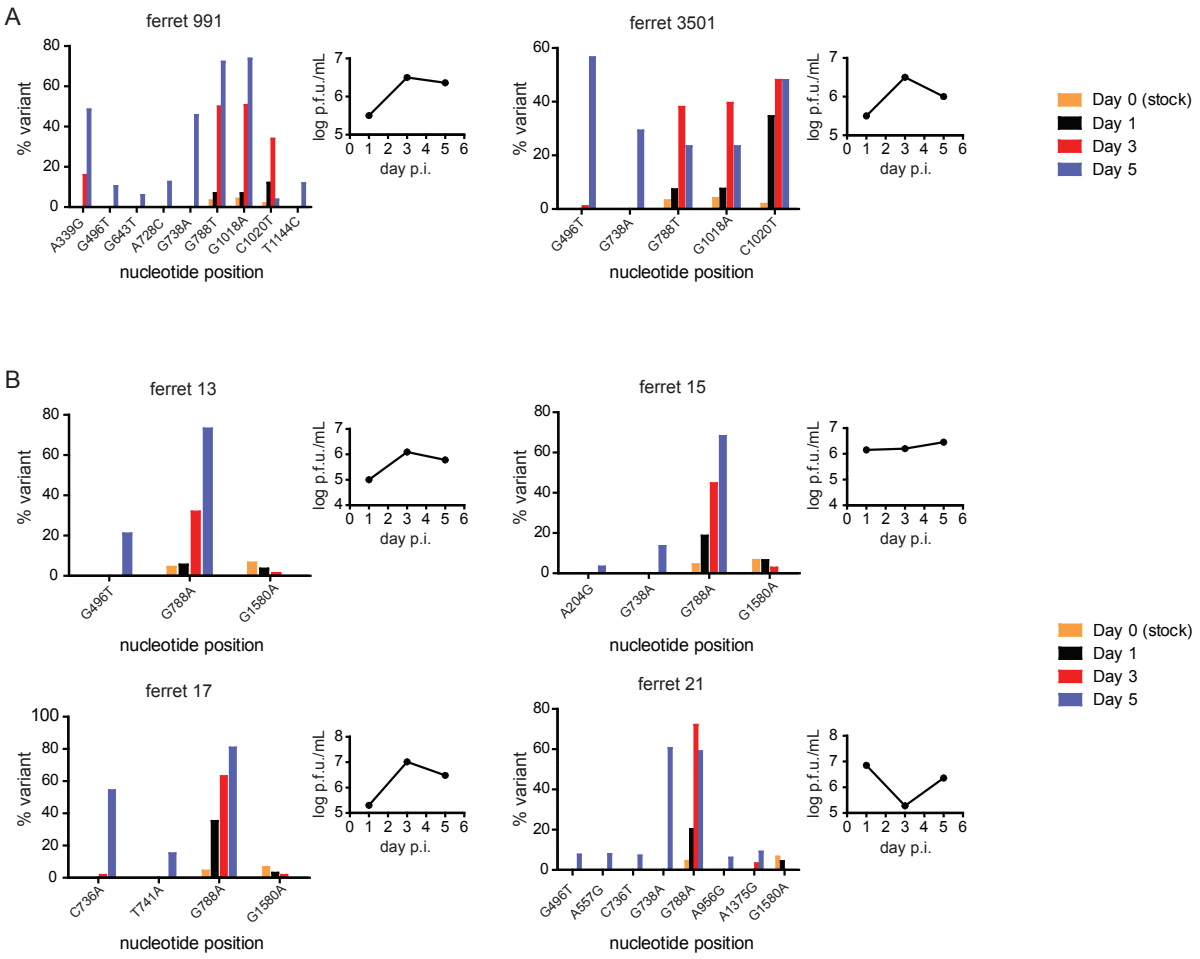


Figure 2. Within-host selection of HA segments harboring specific single nucleotide polymorphisms.

Viral RNA recovered in nasal wash samples collected from ferrets at different timepoints following intranasal infection with the indicated viruses was used to measure HA segment variation by deep sequencing. Bar graphs depict changing frequencies of specific SNPs during infection of index ferrets with (A) VN1203-HA(3)-CA04 or (B) VN1203-HA(4)-CA04 viruses. Number of sequences used to calculate SNP frequencies ranged from n = 198 to 15206 for VN1203-HA(3)-CA04 and n = 111 to 11411 for VN1203-HA(4)-CA04. This analysis focused on SNPs detected in at least 1% of virus sequences in stock viruses or in one or more samples collected from any ferret at any time point. Each SNP was nonsynonymous, with the exception of a synonymous SNP at nucleotide position 1018. Inset line graphs depict virus titers in nasal wash samples collected from each index ferret at the indicated timepoint. Virus titers were measured using a standard plaque assay on MDCK cells.

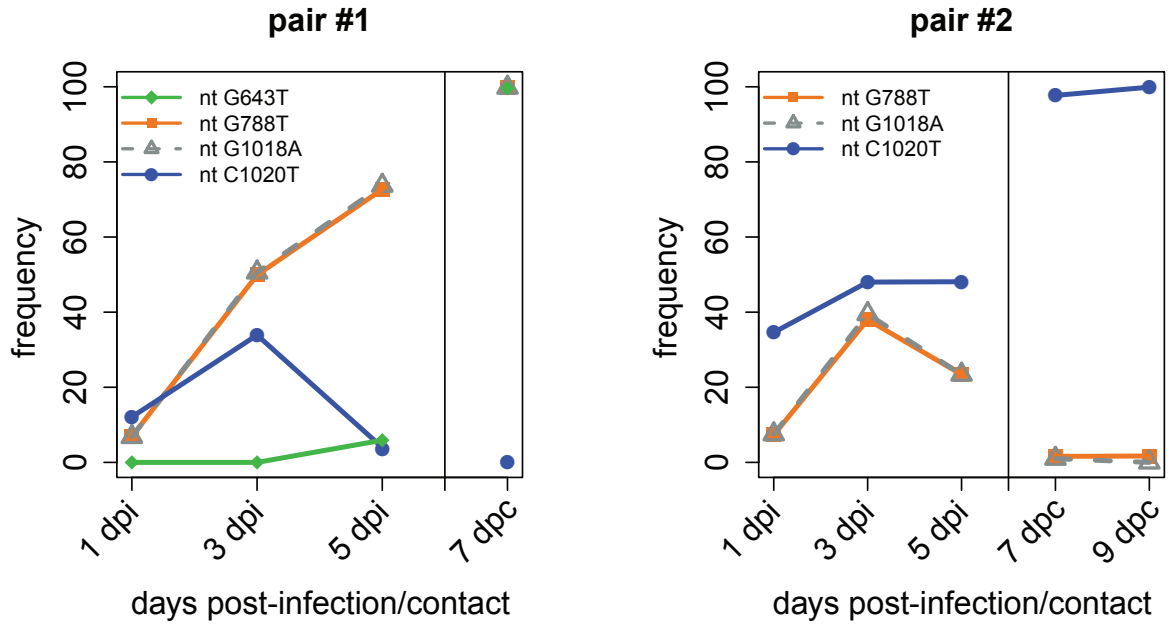


Figure 3. Detection of HA SNPs early after infection in contact animals. HA gene segments accumulated diversity over time during replication in index animals, as demonstrated by the increasing frequency of substitutions at positions 643, 788, 1018 and 1020 in index animals infected with VN1203-HA(3)-CA04 (left-hand portion of each panel). Following transmission, the founding virus population in contact animals displayed a shift in SNP frequencies, such that SNPs were either nearly fixed in, or were absent from, the replicating virus population. A SNP detected in 5.9% of viruses in the pair 1 index animal shortly before transmission was present in nearly 100% of virus sequences shortly after transmission in the paired contact animal. Number of sequences used to calculate SNP frequencies ranged from n = 1472 to 13324.

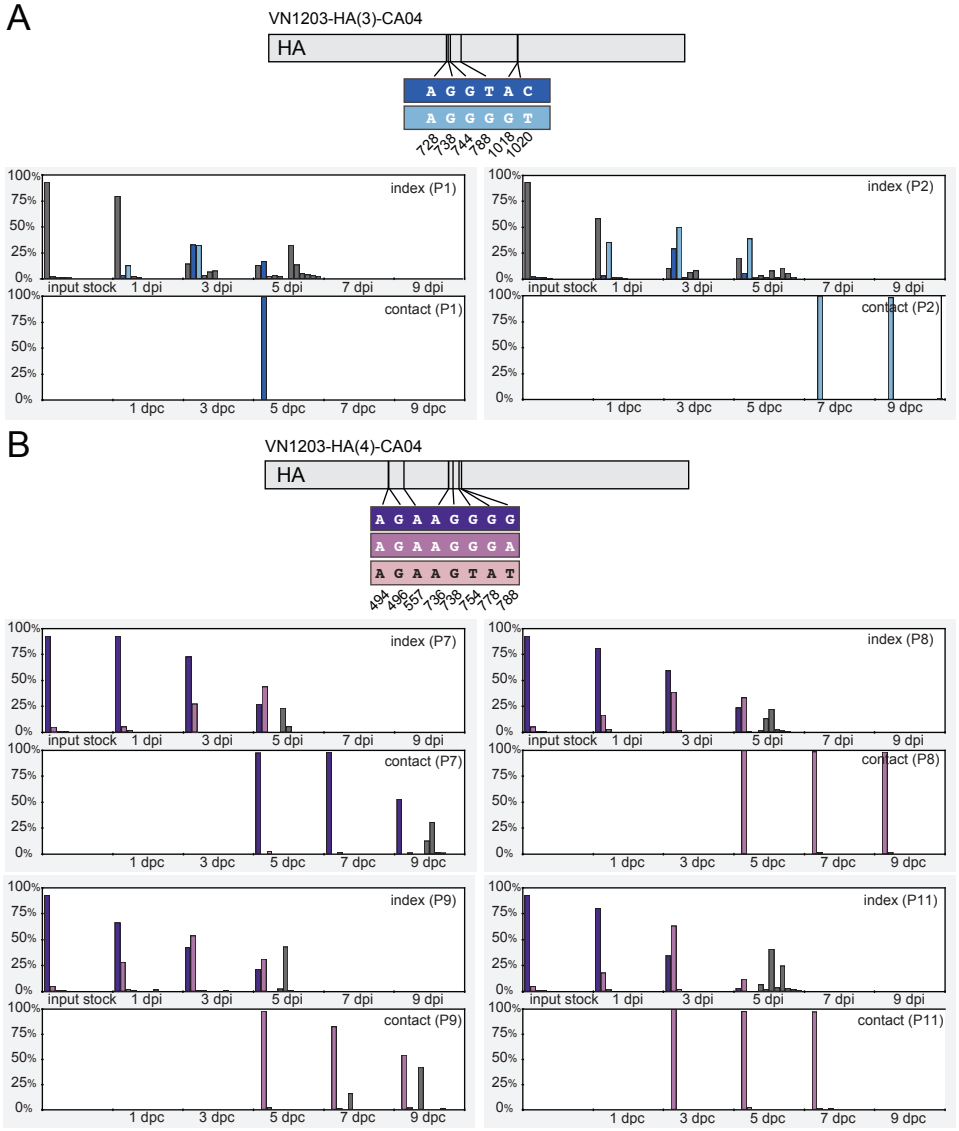


Figure 4. Enumeration of HA segment "haplotypes". To identify patterns of physically linked SNPs, we took advantage of the fact that paired-end deep sequencing provides "mate-paired" reads that are separated by intervening sequences of varying length, allowing us to identify reads containing sites of interest that are linked on the same viral RNA. By analyzing these reads, we identified linkage relationships among targeted SNPs, and use the term "haplotype" to denote a single unique combination of SNPs. SNPs used to define HA haplotypes are shown schematically above each panel. This analysis targeted nucleotides 728, 738, 744, 788, 1018 and 1020 in virus VN1203-HA(3)-CA04 (panel A) and nucleotides 494, 496, 557, 736, 754, 778 and 788 in virus VN1203-HA(4)-CA04 (panel B). We considered only "haplotypes" detected at or above a frequency of 1% of the total virus population. Within each panel, grey boxes indicate a transmission pair, each with an index animal (above) and a contact animal (below). The x-axis represents the sample collection timepoints for index or contact animals. Grey bars denote the frequencies of non-transmitting haplotypes. The frequencies of HA haplotypes implicated in transmission are colored with the specific constellation of SNPs indicated in the schematic above each panel. Note that the minor haplotype detected at the first timepoint in which virus was recovered from the contact ferret of pairs 7 and 9 was also found in the paired index ferret. Number of sequences used to calculate haplotype frequencies ranged from n = 1418 to 3128 for VN1203-HA(3)-CA04 and n = 540 to 1050 for VN1203-HA(4)-CA04. Further details can be found in Supplementary Table S2 and S3.

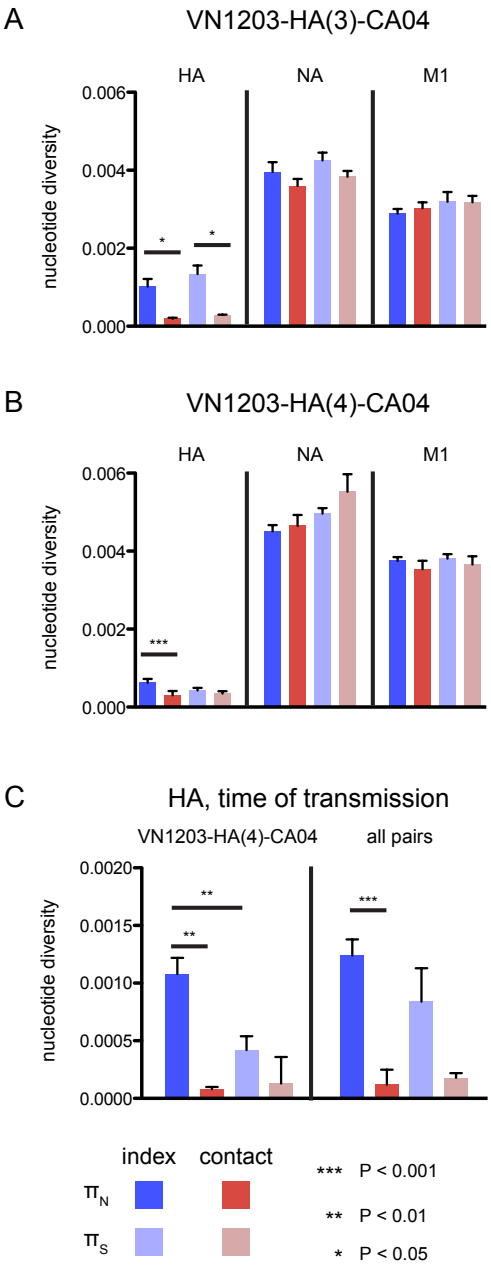
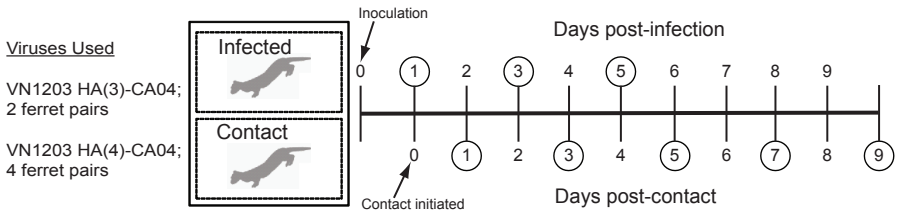


Figure 5. Within-host nucleotide diversity in HA. We determine mean nonsynonymous (πN) and synonymous (πS) nucleotide diversity throughout the experiment in the HA, NA and M1 coding regions of viruses isolated from index and contact ferrets infected with (A) VN1203-HA(3)-CA04; n = 2 or (B) VN1203-HA(4)-CA04; n = 4. (C) To independently assess the impact of transmission on HA nucleotide diversity, we compared πN and πS at the single timepoint closest to transmission for each ferret pair. For this analysis we considered either the VN1203-HA(4)-CA04-infected group alone (panel C, left) or all 6 ferret pairs together (panel C, right). In each graph, vertical bars represent the mean nucleotide diversity for all index and contact samples; error bars represent standard error of the mean (s.e.m.). Dark and light blue bars indicate πN and πS, respectively, in index animals; dark and light red bars indicate πN and πS, respectively, in contact animals. We used paired t-tests to compare πN and/or πS values within and between gene segments. Horizontal bars highlight comparisons for which two values are significantly different. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

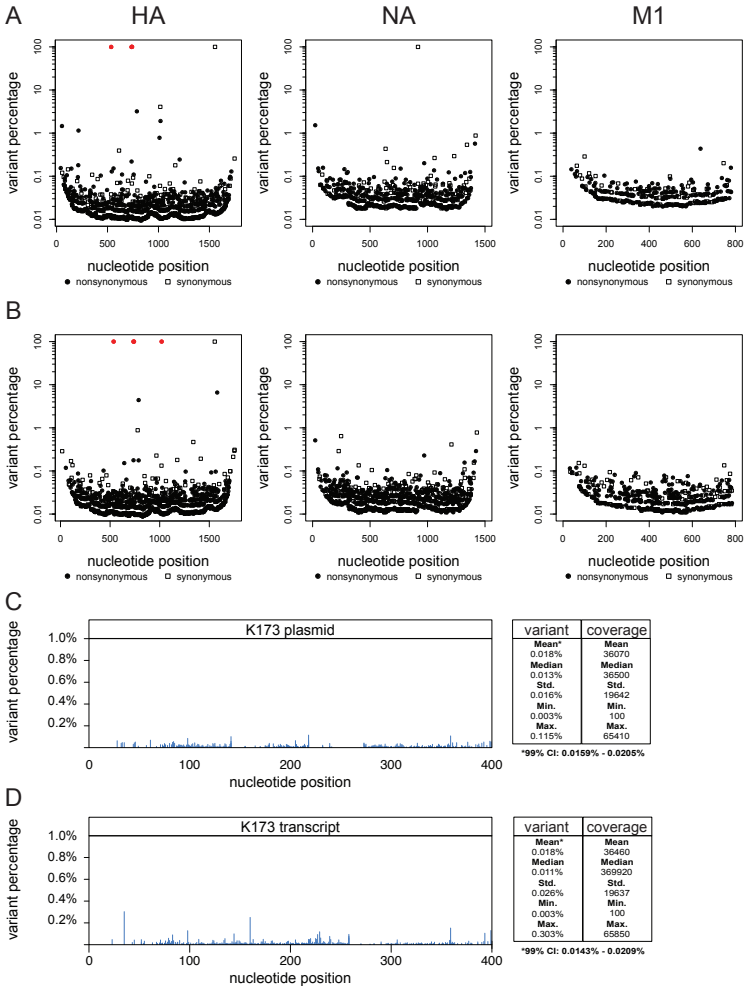
A



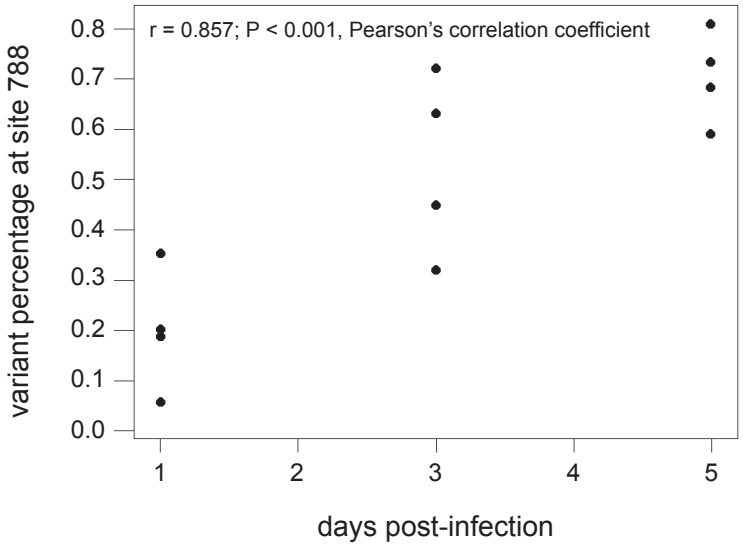
B

Earliest detectable timepoint with recovered virus from contact animals		
ferret pair number	**virus**	**day post-contact**
1	VN1203-HA(3)-CA04	5 DPC
2	VN1203-HA(3)-CA04	7 DPC
7	VN1203-HA(4)-CA04	5 DPC
8	VN1203-HA(4)-CA04	5 DPC
9	VN1203-HA(4)-CA04	5 DPC
11	VN1203-HA(4)-CA04	3 DPC

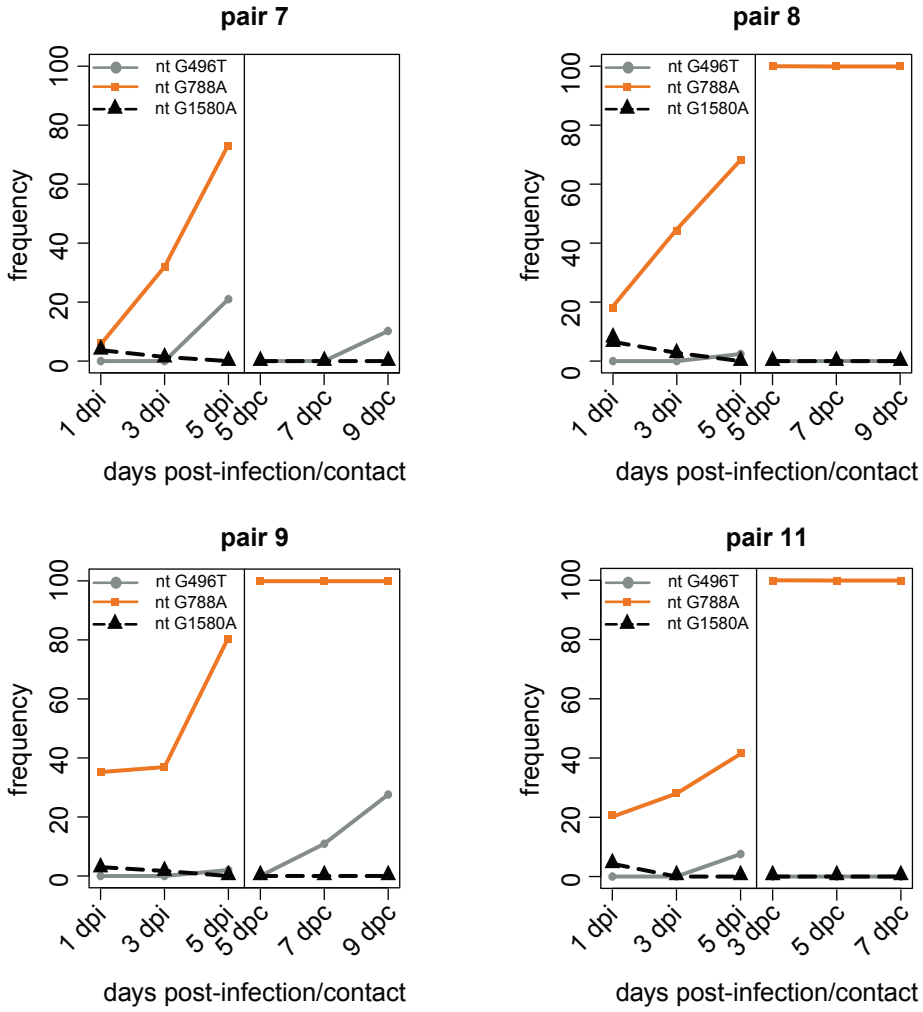
Supplementary Figure S1. Mammalian transmission experiment overview. This study uses samples taken from ferrets infected with reassortant H5N1 influenza viruses as part of a previously reported study. Here we present a schematic of the transmission experiments performed in that study; no new transmission experiments were performed for the study described here. (A) Transmission pairs consisted of an index (directly inoculated) and contact ferret: pair 1- ferrets 991/3498, pair 2 - ferrets 3501/3499, pair 7 - ferrets 13/14, pair 8- ferrets 15/16, pair 9- ferrets 17/18 and pair 11- ferrets 21/22. Index ferrets were intranasally inoculated with 10^6 plaque-forming units (500 µl) of either VN1203-HA(3)-CA04 virus (pairs 1 and 2) or VN1203-HA(4)-CA04 virus (pairs 7, 8, 9 and 11). One day after inoculation, a naïve contact ferret was placed in an adjacent isolator cage. Nasal washes were collected from the infected index ferrets on days 1, 3, and 5 after infection, and from contact ferrets on days 1, 3, 5, 7, and 9 after contact was initiated. Further details of the transmission experiment are presented in ref. 2. (B) The earliest timepoint in which virus was recovered from the contact animal nasal washes: pair 1 - 5 DPC, pair 2 - 7 DPC, pair 7 - 5 DPC, pair 8 - 5 DPC, pair 9 - 5 DPC, pair 11 - 3 DPC.



Supplementary Figure S2. Deep sequencing reveals low-level sequence variation in NA and M gene segments in stock H5N1 reassortant viruses. (A-B) Massively parallel Illumina sequencing revealed genetic variation present within the stock viruses VN1203-HA(3)-CA04 (panel A; HA, NA and M1) VN1203-HA(4)-CA04 (panel B; HA, NA and M1). Illumina MiSeq sequence reads for NA and M1 were mapped against reference sequences from A/California/04/2009. Results for HA are described in Figure 1 and are included here for comparison. HA mutations previously reported in association with mammalian transmission are highlighted in red: VN1203-HA(3)-CA04: N158D (nt 536), N224K (nt 736) and Q226L (nt 741) and VN1203-HA(4)-CA04: N158D (nt 536), N224K (nt 736), Q226L (nt 741) and T318I (nt 1020). The frequency of variants detected at each site is presented as closed circles for nonsynomyous mutations or open squares for synonymous mutations. A synonymous mutation at nucleotide position 1555 in HA that was not present in the plasmid used to generate viruses by reverse genetics arose during replication of the stock viruses in tissue culture prior to inoculation of ferrets. (C-D) Validation of the ≥1% SNP detection threshold was performed by deep sequencing a 430-bp M gene amplicon derived from either plasmid DNA or an in-vitro transcript. No SNPs were detected at frequencies above 0.115% from the plasmid-derived deep sequences (panel C) or above 0.303% from transcript-derived deep sequences (panel D). Thus, our 1% SNP detection threshold is approximately 3-fold higher than the highest frequency of spurious substitutions observed in our system. SNP and sequence coverage summary statistics for plasmid and transcript-derived deep sequences are found on the right side of each panel. Confidence intervals of the mean variant percentage were calculated based on a normal distribution (transcript, n = 360; plasmid, n = 543).

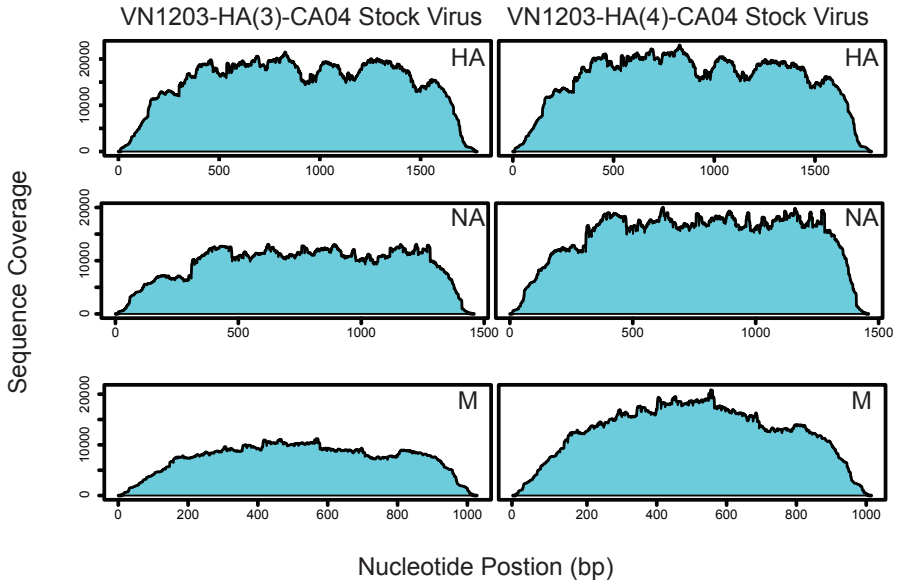


Supplementary Figure S3. Time-dependent increase in variant nucleotides at position 788 during infection of index ferrets. We found a positive correlation between the frequency of a variant nucleotide at position 788 (Ala → Thr substitution at amino acid position 242 by H3 numbering) and time post-infection for index animals infected with VN1203-HA(4)-CA04 stock virus (n = 4).



Supplementary Figure S4. SNPs in index virus HA segments were detected in nearly 100% or nearly 0% of viruses replicating early after infection of contact animals. HA gene segments diversified over time during replication in index animals, as demonstrated by the increasing frequency of substitutions at position 496, 788 and 1580 in the four index animals infected with VN1203-HA(4)-CA04 stock virus (left-hand portion of each panel). Virus populations replicating soon after demonstrated a shift in SNP frequencies. At the earliest timepoint in contact animals, SNPs were either nearly fixed in, or were absent from, the virus population. Number of sequences used to calculate SNP frequencies ranged from n = 175 to 29976.

A



B

ID	inoculating virus	ferret id number	ferret pair number	stock/index/ contact	sample date	HA average coverage ± standard deviation	NA average coverage ± standard deviation	M average coverage ± standard deviation
1	VN1203-HA(3)-CA04	n/a	n/a	stock	n/a	14638 ± 5504	9149 ± 3472	7254 ± 2853
2	VN1203-HA(4)-CA04	n/a	n/a	stock	n/a	15374 ± 5811	14148 ± 5153	12374 ± 5330
3	VN1203-HA(3)-CA04	991	1	index	1 DPI	17106 ± 6387	14060 ± 4986	14716 ± 5713
4	VN1203-HA(3)-CA04	991	1	index	3 DPI	18522 ± 6756	13440 ± 4834	13293 ± 5182
5	VN1203-HA(3)-CA04	991	1	index	5 DPI	29136 ± 10903	11610 ± 4343	11967 ± 4777
6	VN1203-HA(3)-CA04	3498	1	contact	7 DPC	14011 ± 5326	13700 ± 5022	13708 ± 5509
7	VN1203-HA(3)-CA04	3501	2	index	1 DPI	16033 ± 6131	16091 ± 5863	15962 ± 6676
8	VN1203-HA(3)-CA04	3501	2	index	3 DPI	16407 ± 6212	18040 ± 6799	15246 ± 6443
9	VN1203-HA(3)-CA04	3501	2	index	5 DPI	17517 ± 6667	15335 ± 5872	19071 ± 8286
10	VN1203-HA(3)-CA04	3499	2	contact	5 DPC	17613 ± 6669	16224 ± 6311	16314 ± 6869
11	VN1203-HA(3)-CA04	3499	2	contact	7 DPC	16245 ± 6101	15677 ± 5709	14800 ± 6003
12	VN1203-HA(4)-CA04	13	7	index	1 DPI	16277 ± 6058	8780 ± 3372	7411 ± 2965
13	VN1203-HA(4)-CA04	13	7	index	3 DPI	19246 ± 7047	13447 ± 5395	10362 ± 4261
14	VN1203-HA(4)-CA04	13	7	index	5 DPI	14249 ± 5298	15372 ± 5806	11245 ± 4485
15	VN1203-HA(4)-CA04	14	7	contact	5 DPC	11773 ± 4453	14843 ± 5669	10203 ± 3953
16	VN1203-HA(4)-CA04	14	7	contact	7 DPC	17307 ± 6331	12825 ± 5227	8956 ± 3707
17	VN1203-HA(4)-CA04	14	7	contact	9 DPC	15009 ± 5711	12852 ± 5052	10633 ± 4254
18	VN1203-HA(4)-CA04	15	8	index	1 DPI	15282 ± 5730	11144 ± 4517	8857 ± 3765
19	VN1203-HA(4)-CA04	15	8	index	3 DPI	13581 ± 5256	14042 ± 5598	11258 ± 4520
20	VN1203-HA(4)-CA04	15	8	index	5 DPI	13900 ± 5221	12602 ± 4901	9238 ± 3704
21	VN1203-HA(4)-CA04	16	8	contact	5 DPC	17082 ± 6516	11147 ± 4339	8370 ± 3315
22	VN1203-HA(4)-CA04	16	8	contact	9 DPC	14249 ± 5279	13067 ± 5466	10433 ± 4490
23	VN1203-HA(4)-CA04	16	8	contact	7 DPC	18182 ± 6694	19372 ± 7249	15548 ± 6528
24	VN1203-HA(4)-CA04	17	9	index	1 DPI	20506 ± 7453	12325 ± 4597	9463 ± 3688
25	VN1203-HA(4)-CA04	17	9	index	3 DPI	16121 ± 6116	11843 ± 4895	9805 ± 4209
26	VN1203-HA(4)-CA04	17	9	index	5 DPI	11567 ± 4554	10059 ± 4381	8599 ± 3663
27	VN1203-HA(4)-CA04	18	9	contact	5 DPC	20492 ± 7367	14616 ± 5361	12643 ± 4987
28	VN1203-HA(4)-CA04	18	9	contact	7 DPC	19018 ± 6967	15605 ± 5612	13228 ± 5066
29	VN1203-HA(4)-CA04	18	9	contact	9 DPC	21433 ± 8173	13128 ± 4509	13988 ± 5309
30	VN1203-HA(4)-CA04	21	11	index	1 DPI	18156 ± 6718	12657 ± 4884	11866 ± 4916
31	VN1203-HA(4)-CA04	21	11	index	3 DPI	21419 ± 7860	13878 ± 5082	13989 ± 5584
32	VN1203-HA(4)-CA04	21	11	index	5 DPI	18202 ± 6829	10858 ± 4193	10447 ± 4605
33	VN1203-HA(4)-CA04	22	11	contact	3 DPC	16739 ± 6388	11637 ± 4786	8417 ± 3737
34	VN1203-HA(4)-CA04	22	11	contact	5 DPC	15005 ± 5699	15949 ± 5553	15871 ± 6133
35	VN1203-HA(4)-CA04	22	11	contact	7 DPC	20032 ± 7554	11883 ± 4183	12376 ± 4950

Supplementary Figure S5. Reference-based assemblies of stock viruses HA, NA and M gene segments and mapping statistics for all sequencing samples. Sequences of viral populations were individually mapped against VN1203 HA, CA04 NA, or CA04 M reference sequences. **(A)** Graphical representation of the reference based assemblies for VN1203-HA(3)-CA04 and VN1203-HA(4)-CA04 stock viruses illustrates the "depth" of sequence coverage for each segment. The x-axis indicates the base pair positions relative to the reference sequence. The y-axis denotes the number of sequence reads that covered each base along the reference sequence. **(B)** Coverage statistics for HA, NA and M gene segments for each sample included in this study.

	Ferret # 991 (INDEX)				Ferret # 3498 (CONTACT)			
	LinkGE- 5 DPI		Sanger- 5 DPI		LinkGE- 7 DPC		Sanger- 7 DPC	
Pair #1	Read Total	Percentage	Read Total	Percentage	Read Total	Percentage	Read Total	Percentage
A G G G G C	408	13.0%	1.00	4.2%				
A G G T A C	530	16.9%	9.00	37.6%	1516	100%	24.00	100%
A G G G G T	77	2.5%						
A G G T G C	107	3.4%						
A G G G A C	75	2.4%						
A G G T G T								
A A G T A C	1019	32.6%	11.00	45.8%				
C G G G G C	427	13.7%	3.00	12.5%				
C A G T A C	164	5.2%						
A A G T G C	134	4.3%						
C G G T A C	106	3.4%						
C G G G A C	81	2.6%						
Total:	3,128		24		1,516		24	

	Ferret # 3501 (INDEX)				Ferret # 3499 (CONTACT)			
	LinkGE- 3 DPI		Sanger- 3 DPI		LinkGE- 5 DPC		Sanger- 5 DPC	
Pair #2	Read Total	Percentage	Read Total	Percentage	Read Total	Percentage	Read Total	Percentage
A G G G G C	196	9.8%	2.00	8.0%				
A G G T A C	554	27.6%	7.00	28.0%				
A G G G G T	945	47.0%	15.00	60.0%	2185	100%	10.00	100%
A G G T G C	33	1.6%						
A G G G A C	123	6.1%	1.00	4.0%				
A G G T G T	158	7.9%						
A A G T A C								
C G G G G C								
C A G T A C								
A A G T G C								
C G G T A C								
C G G G A C								
Total:	2,009		25		2,185		10	

Supplementary Table S1. LinkGE detected HA haplotypes with greater sensitivity than by conventional Sanger sequencing. To demonstrate that the assembly of HA haplotypes using deep sequencing data accurately reflected the assemblage of SNPs in the viral population, individual HA genes were cloned for Sanger sequencing from the viral population collected from representative ferrets infected with reassortant H5N1 virus and their paired contact ferret. The sequences of cloned HA segments were determined using conventional Sanger sequencing. The resulting haplotypes were then compared to haplotype frequencies determined using LinkGE haplotype assembly based on deep sequencing data. Clones were generated from RNA harvested from nasal washes collected from ferret pair 1 (left table) and pair 2 (right table) for the indicated timepoints.

	Ferret # 991 (INDEX)				Ferret # 3498 (CONTACT)			
Pair #1	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G G G G C	1323	1473	210	408				
A G G T A C	32	67	506	530			1516	
A G G G G T	25	240	470	77				
A G G T G C	22	44	51	107				
A G G G A C	16	26	98	75				
A G G T G T			113					
A A G T A C				1019				
C G G G G C				427				
C A G T A C				164				
A A G T G C				134				
C G G T A C				106				
C G G G A C				81				
A A G G G T								
A A G G G C								
A A G T G T								
A G G A G T								
Total:	1,418	1,850	1,448	3,128			1,516	

	Ferret # 3501 (INDEX)				Ferret # 3499 (CONTACT)			
Pair #2	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G G G G C	1323	1237	196	464				
A G G T A C	32	69	554	132				
A G G G G T	25	752	945	887		2185	1787	
A G G T G C	22	35	33	28				
A G G G A C	16	27	123	78				
A G G T G T		23	158	42				
A A G T A C				183				
C G G G G C								
C A G T A C								
A A G T G C				36				
C G G T A C								
C G G G A C								
A A G G G T				245				
A A G G G C				128				
A A G T G T				48				
A G G A G T							29	
Total:	1,418	2,143	2,009	2,271		2,185	1,816	

	Ferret # 13 (INDEX)				Ferret # 14 (CONTACT)			
Pair #7	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G A A G G G G	688	715	648	195		622	551	440
A G A A G G G A	38	41	244	315				
A G A A G T A T	8	15				16	10	10
A G A A G G A G	8							
A T A A G G G A				167				
A T A A G G G G				40				105
A G A A A G G G								254
A T A A A G G G								12
A G C A G G G G								9
A G A C G G G A								
A G A C G G G G								
A G A A A G G A								
G G A A G G G G								
A G A A G G A A								
A G A A A A A A								
A G G A G G G A								
A G A A G T G A								
A T A A A G G A								
Total:	742	771	892	717		638	561	830

	Ferret # 17 (INDEX)				Ferret # 18 (CONTACT)			
Pair #8	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G A A G G G G	688	630	542	131				
A G A A G G G A	38	129	353	186		724	744	672
A G A A G T A T	8	19	18	6			11	12
A G A A G G A G	8							
A T A A G G G A								
A T A A G G G G				12				
A G A A A G G G				74				
A T A A A G G G								
A G C A G G G G								
A G A C G G G A				124				
A G A C G G G G				13				
A G A A A G G A				7				
G G A A G G G G				6				
A G A A G G A A								
A G A A A A A A								
A G G A G G G A								
A G A A G T G A								
A T A A A G G A								
Total:	742	778	913	559		724	755	684

	Ferret # 15 (INDEX)				Ferret # 16 (CONTACT)			
Pair #9	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G A A G G G G	688	358	381	59				
A G A A G G G A	38	153	483	86		488	671	393
A G A A G T A T	8	11	10			10	11	14
A G A A G G A G	8	7	13					
A T A A G G G A				7			131	307
A T A A G G G G								
A G A A A G G G								
A T A A A G G G								
A G C A G G G G								
A G A C G G G A				118				
A G A C G G G G				4				
A G A A A G G A								
G G A A G G G G								
A G A A G G A A		11	11					
A G A A A A A A								10
A G G A G G G A								
A G A A G T G A								
A T A A A G G A								
Total:	742	540	898	274		498	813	724

	Ferret # 21 (INDEX)				Ferret # 22 (CONTACT)			
Pair #11	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G A A G G G G	688	674	281	28				
A G A A G G G A	38	152	515	100	577	766	1021	
A G A A G T A T	8	15	18			18	18	
A G A A G G A G	8							
A T A A G G G A				56			11	
A T A A G G G G				19				
A G A A A G G G				339				
A T A A A G G G				31				
A G C A G G G G								
A G A C G G G A								
A G A C G G G G								
A G A A A G G A				205				
G G A A G G G G								
A G A A G G A A								
A G A A A A A A								
A G G A G G G A				24				
A G A A G T G A				16				
A T A A A G G A				10				
Total:	742	841	814	828	577	784	1050	

Supplementary Table S2. LinkGE frequencies for HA gene segment haplotypes. LinkGE (see Methods for details) was used to mine paired-end sequencing data in order to link nucleotide polymorphisms and define 'HA haplotypes'. The genotype and frequency of each HA haplotype identified in samples harvested from paired index and contact ferrets are shown. Grey shaded regions indicate timepoints for which virus was not detected in nasal wash samples. Polymorphic nucleotide sites queried by LinkGE to define and enumerate HA haplotypes were as follows: pair 1 and pair 2 – nucleotide sites 728, 738, 744, 1018 and 1020; pair 7, pair 8, pair 9 and pair 11 – nucleotide sites 494, 496, 557, 736, 738, 754, 778 and 788.

	Ferret # 991 (INDEX)				Ferret # 3498 (CONTACT)			
Pair #1	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G G G G C	93.30%	79.62%	14.50%	13.04%				
A G G T A C	2.26%	3.62%	34.94%	16.94%			100.00%	
A G G G G T	1.76%	12.97%	32.46%	2.46%				
A G G T G C	1.55%	2.38%	3.52%	3.42%				
A G G G A C	1.13%	1.41%	6.77%	2.40%				
A G G T G T			7.80%					
A A G T A C				32.58%				
C G G G G C				13.65%				
C A G T A C				5.24%				
A A G T G C				4.28%				
C G G T A C				3.39%				
C G G G A C				2.59%				
A A G G G T								
A A G G G C								
A A G T G T								
A G G A G T								

	Ferret # 3501 (INDEX)				Ferret # 3499 (CONTACT)			
Pair #2	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G G G G C	93.30%	57.7%	9.8%	20.43%				
A G G T A C	2.26%	3.2%	27.6%	5.81%				
A G G G G T	1.76%	35.1%	47.0%	39.06%		100.00%	98.40%	
A G G T G C	1.55%	1.6%	1.6%	1.23%				
A G G G A C	1.13%	1.3%	6.1%	3.43%				
A G G T G T		1.09%	7.9%	1.85%				
A A G T A C				8.1%				
C G G G G C								
C A G T A C								
A A G T G C				1.6%				
C G G T A C								
C G G G A C								
A A G G G T				10.8%				
A A G G G C				5.6%				
A A G T G T				2.1%				
A G G A G T								1.6%

	Ferret # 13 (INDEX)				Ferret # 14 (CONTACT)			
Pair #7	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G A A G G G G	92.7%	92.7%	72.7%	27.2%		97.5%	98.2%	53.0%
A G A A G G G A	5.1%	5.3%	27.4%	43.9%				
A G A A G T A T	1.1%	2.0%				2.51%	1.78%	1.20%
A G A A G G A G	1.1%							
A T A A G G G A				23.3%				
A T A A G G G G				5.6%				12.7%
A G A A A G G G								30.6%
A T A A A G G G								1.5%
A G C A G G G G								1.1%
A G A C G G G A								
A G A C G G G G								
A G A A A G G A								
G G A A G G G G								
A G A A G G A A								
A G A A A A A A								
A G G A G G G A								
A G A A G T G A								
A T A A A G G A								

	Ferret # 17 (INDEX)				Ferret # 18 (CONTACT)			
Pair #8	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G A A G G G G	92.7%	81.0%	59.4%	23.4%				
A G A A G G G A	5.1%	16.6%	38.7%	33.3%		100%	98.5%	98.3%
A G A A G T A T	1.1%	2.4%	2.0%	1.1%			1.5%	1.8%
A G A A G G A G	1.1%							
A T A A G G G A								
A T A A G G G G				2.2%				
A G A A A G G G				13.2%				
A T A A A G G G								
A G C A G G G G								
A G A C G G G A				22.2%				
A G A C G G G G				2.3%				
A G A A A G G A				1.3%				
G G A A G G G G				1.1%				
A G A A G G A A								
A G A A A A A A								
A G G A G G G A								
A G A A G T G A								
A T A A A G G A								

	Ferret # 15 (INDEX)				Ferret # 16 (CONTACT)			
Pair #9	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G A A G G G G	92.7%	66.3%	42.4%	21.5%				
A G A A G G G A	5.1%	28.3%	53.8%	31.3%		98.0%	82.5%	54.3%
A G A A G T A T	1.1%	2.0%	1.1%			2.0%	1.4%	1.9%
A G A A G G A G	1.1%	1.3%	1.5%					
A T A A G G G A				2.6%			16.1%	42.4%
A T A A G G G G								
A G A A A G G G								
A T A A A G G G								
A G C A G G G G								
A G A C G G G A				43.1%				
A G A C G G G G				1.5%				
A G A A A G G A								
G G A A G G G G								
A G A A G G A A		2.0%	1.2%					
A G A A A A A A								1.4%

	Ferret # 21 (INDEX)				Ferret # 22 (CONTACT)			
Pair #11	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A G A A G G G G	92.7%	80.1%	34.5%	3.4%				
A G A A G G G A	5.1%	18.1%	63.3%	12.1%	100%	97.7%	97.2%	
A G A A G T A T	1.1%	1.8%	2.2%			2.3%	1.7%	
A G A A G G A G	1.1%							
A T A A G G G A				6.8%			1.1%	
A T A A G G G G				2.3%				
A G A A A G G G				40.9%				
A T A A A G G G				3.7%				
A G C A G G G G								
A G A C G G G A								
A G A C G G G G								
A G A A A G G A				24.8%				
G G A A G G G G								
A G A A G G A A								
A G A A A A A A								

Supplementary Table S3. LinkGE absolute counts for HA gene segment haplotypes. LinkGE (see Methods for details) was used to mine paired-end sequencing data in order to link nucleotide polymorphisms and define 'HA haplotypes'. The total number of recovered individual sequence reads that covered all of the queried polymorphic nucleotide sites are presented according to genotype. Grey shaded regions indicate timepoints for which virus was not detected in nasal wash samples. Polymorphic nucleotide sites queried by LinkGE to define and enumerate HA haplotypes were as follows: pair 1 and pair 2 – nucleotide sites 728, 738, 744, 1018 and 1020; pair 7, pair 8, pair 9 and pair 11 – nucleotide sites 494, 496, 557, 736, 738, 754, 778 and 788.

	Ferret # 991 (INDEX)				Ferret # 3498 (CONTACT)			
Pair #1	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
C G G C	1.20%	1.82%	1.21%				2.07%	
C G A A	1.08%							
C A C A				3.47%				
C G C A	97.72%	98.18%	97.58%	95.14%			97.93%	
C G G T			1.21%	1.39%				
C G G C								
C G A A								
C G C A								
C G G T								
T G C A								
C G C G								

	Ferret # 3501 (INDEX)				Ferret # 3499 (CONTACT)			
Pair #2	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
C G G C	1.20%	1.42%		1.27%		1.12%	1.33%	
C G A A	1.08%	1.14%				1.67%		
C A C A								
C G C A	97.72%	97.4%	100.0%	93.9%		97.2%	97.35%	
C G G T							1.3%	
C G G C								
C G A A								
C G C A								
C G G T								
T G C A				2.04%				
C G C G				2.80%				

	Ferret # 13 (INDEX)				Ferret # 14 (CONTACT)			
Pair #7	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A T G G C G T				1.41%				
A T G T A C T	2.21%	2.22%	2.02%			2.14%	1.92%	
A T G G C G C	97.79%	94.44%	95.68%	93.43%		97.86%	96.54%	98.77%
A T G C C C T		1.11%		3.76%			1.54%	1.23%
A T G C G A T		1.11%	2.31%	1.41%				
A T G A T G G		1.11%						
A T G C C G C								
A T G G T G C								
C T A G C G C								
A T G G T T T								
A T G T T A G								
C T A T A C T								
G T G G C G C								
A T T G C G C								
A T G A C G C								
T C T G C G C								
A T A G C G C								
G C G G C G C								
A A G G C G C								
G A T G C G C								
A T G G C A C								

	Ferret # 17 (INDEX)				Ferret # 18 (CONTACT)			
Pair #8	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A T G G C G T			1.14%					
A T G T A C T	2.21%	1.96%	2.27%	2.62%		1.12%	2.92%	1.54%
A T G G C G C	97.79%	96.08%	95.45%	95.88%		95.15%	93.86%	97.30%
A T G C C C T		1.96%	1.14%	1.50%		1.87%	3.22%	
A T G C G A T								
A T G A T G G								
A T G C C G C								1.16%
A T G G T G C						1.87%		
C T A G C G C								
A T G G T T T								
A T G T T A G								
C T A T A C T								
G T G G C G C								
A T T G C G C								
A T G A C G C								
T C T G C G C								
A T A G C G C								
G C G G C G C								
A A G G C G C								
G A T G C G C								
A T G G C A C								

	Ferret # 15 (INDEX)				Ferret # 16 (CONTACT)			
Pair #9	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A T G G C G T								
A T G T A C T	2.21%	1.29%		1.12%		1.26%	3.50%	1.43%
A T G G C G C	97.79%	96.77%	96.43%	97.77%		97.48%	96.50%	91.43%
A T G C C C T		1.94%	2.38%					1.43%
A T G C G A T			1.19%	1.12%				1.43%
A T G A T G G								
A T G C C G C						1.3%		
A T G G T G C								
C T A G C G C								
A T G G T T T								1.43%
A T G T T A G								1.43%
C T A T A C T								1.43%
G T G G C G C								
A T T G C G C								
A T G A C G C								
T C T G C G C								
A T A G C G C								
G C G G C G C								
A A G G C G C								
G A T G C G C								
A T G G C A C								

	Ferret # 21 (INDEX)				Ferret # 22 (CONTACT)			
Pair #11	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A T G G C G T					1.20%	1.19%	2.60%	
A T G T A C T	2.21%	2.12%	2.88%		1.20%	1.19%	2.60%	
A T G G C G C	97.79%	96.30%	90.38%	96.08%	94.58%	85.71%	76.90%	
A T G C C C T		1.59%	4.81%				5.10%	
A T G C G A T						1.19%		
A T G A T G G								
A T G C C G C								
A T G G T G C								
C T A G C G C					1.2%			
A T G G T T T								
A T G T T A G								
C T A T A C T								
G T G G C G C					1.8%			
A T T G C G C							2.6%	
A T G A C G C						2.4%	10.3%	
T C T G C G C						1.2%		
A T A G C G C				3.9%		1.2%		
G C G G C G C						1.2%		
A A G G C G C						2.4%		
G A T G C G C						2.4%		
A T G G C A C			1.9%					

Supplementary Table S4. LinkGE frequencies for NA gene segment haplotypes. LinkGE (see Methods for details) was used to mine paired-end sequencing data in order to link nucleotide polymorphisms and define 'NA haplotypes'. The genotype and frequency of each NA haplotype identified in samples harvested from paired index and contact ferrets are shown. Grey shaded regions indicate timepoints for which virus was not detected in nasal wash samples. Polymorphic nucleotide sites queried by LinkGE to define and enumerate NA haplotypes were as follows: pair 1 and pair 2 – nucleotide sites 647, 671, 1162 and 1185; pair 7, pair 8, pair 9 and pair 11 – nucleotide sites 447, 467, 516, 917, 925, 926 and 986.

	Ferret # 991 (INDEX)				Ferret # 3498 (CONTACT)			
Pair #1	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
C G G C	10	5	5				6	
C G A A	9							
C A C A				15				
C G C A	816	269	404	411			284	
C G G T			5	6				
C G G C								
C G A A								
C G C A								
C G G T								
T G C A								
C G C G								
Total:	835	274	414	432			290	

	Ferret # 3501 (INDEX)				Ferret # 3499 (CONTACT)			
Pair #2	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
C G G C	10	4		5		6	5	
C G A A	9	5				9		
C A C A								
C G C A	816	343	380	369		523	367	
C G G T							5	
C G G C								
C G A A								
C G C A								
C G G T								
T G C A				8				
C G C G				11				
Total:	835	352	380	393		538	377	

	Ferret # 13 (INDEX)				Ferret # 14 (CONTACT)			
Pair #7	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A T G G C G T				3				
A T G T A C T	3	4	7			6	5	
A T G G C G C	133	170	332	199		274	251	241
A T G C C C T		2		8			4	3
A T G C G A T		2	8	3				
A T G A T G G		2						
A T G C C G C								
A T G G T G C								
C T A G C G C								
A T G G T T T								
A T G T T A G								
C T A T A C T								
G T G G C G C								
A T T G C G C								
A T G A C G C								
T C T G C G C								
A T A G C G C								
G C G G C G C								
A A G G C G C								
G A T G C G C								
A T G G C A C								
Total:	136	180	347	213		280	260	244

	Ferret # 17 (INDEX)				Ferret # 18 (CONTACT)			
Pair #8	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A T G G C G T			3					
A T G T A C T	3	3	6	7		4	10	4
A T G G C G C	133	147	252	256		255	321	252
A T G C C C T		3	3	4		5	11	
A T G C G A T								
A T G A T G G								
A T G C C G C								3
A T G G T G C						5		
C T A G C G C								
A T G G T T T								
A T G T T A G								
C T A T A C T								
G T G G C G C								
A T T G C G C								
A T G A C G C								
T C T G C G C								
A T A G C G C								
G C G G C G C								
A A G G C G C								
G A T G C G C								
A T G G C A C								
Total:	136	153	264	267		269	342	259

	Ferret # 15 (INDEX)				Ferret # 16 (CONTACT)			
Pair #9	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A T G G C G T								
A T G T A C T	3	2		2		3	7	1
A T G G C G C	133	150	162	175		232	193	64
A T G C C C T		3	4					1
A T G C G A T			2	2				1
A T G A T G G								
A T G C C G C						3		
A T G G T G C								
C T A G C G C								
A T G G T T T								1
A T G T T A G								1
C T A T A C T								1
G T G G C G C								
A T T G C G C								
A T G A C G C								
T C T G C G C								
A T A G C G C								
G C G G C G C								
A A G G C G C								
G A T G C G C								
A T G G C A C								
Total:	136	155	168	179		238	200	70

	Ferret # 21 (INDEX)				Ferret # 22 (CONTACT)			
Pair #11	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
A T G G C G T					2	1	1	
A T G T A C T	3	4	3		2	1	1	
A T G G C G C	133	182	94	98	157	72	30	
A T G C C C T		3	5				2	
A T G C G A T						1		
A T G A T G G								
A T G C C G C								
A T G G T G C								
C T A G C G C					2			
A T G G T T T								
A T G T T A G								
C T A T A C T								
G T G G C G C					3			
A T T G C G C							1	
A T G A C G C						2	4	
T C T G C G C						1		
A T A G C G C				4		1		
G C G G C G C						1		
A A G G C G C						2		
G A T G C G C						2		
A T G G C A C			2					
Total:	136	189	104	102	166	84	39	

Supplementary Table S5. LinkGE absolute counts for NA gene segment haplotypes. LinkGE (see Methods for details) was used to mine paired-end sequencing data in order to link nucleotide polymorphisms and define 'NA haplotypes'. The total number of recovered individual sequence reads that covered all of the queried polymorphic nucleotide sites are presented according to genotype. Grey shaded regions indicate timepoints for which virus was not detected in nasal wash samples. Polymorphic nucleotide sites queried by LinkGE to define and enumerate NA haplotypes were as follows: pair 1 and pair 2 – nucleotide sites 647, 671, 1162 and 1185; pair 7, pair 8, pair 9 and pair 11 – nucleotide sites 447, 467, 516, 917, 925, 926 and 986.

	Ferret # 991 (INDEX)				Ferret # 3498 (CONTACT)			
Pair #1	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A							1.13%	
G G A G		1.12%	4.31%					
A G G G								
G G G A				1.23%				
G G G G	100.00%	98.88%	95.69%	98.77%			98.87%	

	Ferret # 3501 (INDEX)				Ferret # 3499 (CONTACT)			
Pair #2	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A								
G G A G								
A G G G							2.86%	
G G G A								
G G G G	100.00%	100.00%	100.00%	100.00%		100.00%	97.06%	

	Ferret # 13 (INDEX)				Ferret # 14 (CONTACT)			
Pair #7	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A		1.28%	1.24%					
A G G G	3.85%	1.50%	2.21%					
G G G G	96.15%	97.22%	96.55%	100.00%		100.00%	100.00%	100.00%

	Ferret # 17 (INDEX)				Ferret # 18 (CONTACT)			
Pair #8	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A						1.05%		
A G G G	3.85%	5.02%	2.77%					
G G G G	96.15%	94.98%	97.23%	100.00%		98.95%	100.00%	100.00%

	Ferret # 15 (INDEX)				Ferret # 16 (CONTACT)			
Pair #9	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A								
A G G G	3.85%	8.76%	4.60%					
G G G G	96.15%	91.24%	95.40%	100.00%		100.00%	100.00%	100.00%

	Ferret # 21 (INDEX)				Ferret # 22 (CONTACT)			
Pair #11	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A		1.12%						
A G G G	3.85%	8.70%	4.10%					
G G G G	96.15%	90.18%	95.90%	100.00%	100.00%	100.00%	100.00%	

Supplementary Table S6. LinkGE frequencies for M1 gene segment haplotypes. LinkGE (see Methods for details) was used to mine paired-end sequencing data in order to link nucleotide polymorphisms and define 'M1 haplotypes'. The genotype and frequency of each M1 haplotype identified in samples harvested from paired index and contact ferrets are shown. Grey shaded regions indicate timepoints for which virus was not detected in nasal wash samples. Polymorphic nucleotide sites queried by LinkGE to define and enumerate M1 haplotypes included nucleotides 77, 109, 295 and 364.

	Ferret # 991 (INDEX)				Ferret # 3498 (CONTACT)			
Pair #1	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A							7.00	
G G A G		6.00	25.00					
A G G G								
G G G A				7.00				
G G G G	310.00	529.00	555.00	560.00			611.00	
Total	310.00	535.00	580.00	567.00			618.00	

	Ferret # 3501 (INDEX)				Ferret # 3499 (CONTACT)			
Pair #2	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A								
G G A G								
A G G G							17.00	
G G G A								
G G G G	310.00	692.00	565.00	838.00		952.00	561.00	
Total	310.00	692.00	565.00	838.00		952.00	578.00	

	Ferret # 13 (INDEX)				Ferret # 14 (CONTACT)			
Pair #7	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A		6.00	9.00					
A G G G	19.00	7.00	16.00					
G G G G	474.00	454.00	700.00	546.00		667.00	688.00	683.00
Total	493.00	467.00	725.00	546.00		667.00	688.00	683.00

	Ferret # 17 (INDEX)				Ferret # 18 (CONTACT)			
Pair #8	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A						6.00		
A G G G	19.00	35.00	21.00					
G G G G	474.00	662.00	736.00	561.00		567.00	818.00	648.00
Total	493.00	697.00	757.00	561.00		573.00	818.00	648.00

	Ferret # 15 (INDEX)				Ferret # 16 (CONTACT)			
Pair #9	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A								
A G G G	19.00	38.00	33.00					
G G G G	474.00	396.00	684.00	882.00		515.00	435.00	338.00
Total	493.00	434.00	717.00	882.00		515.00	435.00	338.00

	Ferret # 21 (INDEX)				Ferret # 22 (CONTACT)			
Pair #11	Input Stock	1 DPI	3 DPI	5 DPI	3 DPC	5 DPC	7 DPC	9 DPC
G G C A		8.00						
A G G G	19.00	62.00	24.00					
G G G G	474.00	643.00	561.00	533.00	416.00	376.00	465.00	
Total	493.00	713.00	585.00	533.00	416.00	376.00	465.00	

Supplementary Table S7. LinkGE absolute counts for M1 gene segment haplotypes. Link GE (see Methods for details) was used to mine paired-end sequencing data in order to link nucleotide polymorphisms and define 'NA haplotypes'. The total number of recovered individual sequence reads that covered all of the queried polymorphic nucleotide sites are presented according to genotype. Grey shaded regions indicate timepoints for which virus was not detected in nasal wash samples. Polymorphic nucleotide sites queried by LinkGE to define and enumerate M1 haplotypes included nucleotides 77, 109, 295 and 364.

Gene	π_S ± S.E.	π_N ± S.E.
HA	0.00019 ± 0.00009**	0.00014 ± 0.00007**
NA	0.00351 ± 0.00048	0.00324 ± 0.00023
M1	0.00333 ± 0.00028	0.00277 ± 0.00038

t-tests of the hypothesis that π_S or π_N equals the corresponding value for virus from index animals: ** $P < 0.01$.

Supplementary Table S8. Mean synonymous (π_s) and nonsynonymous (π_N) nucleotide diversity in virus stocks. Synonymous and nonsynonymous nucleotide diversity for VN1203-HA(3)-CA04 and VN1203-HA(4)-CA04 virus stocks was determined (see Methods for details) for HA, NA and M1 gene segments.

Chapter 3

Natural selection limits human adaptation of H5N1 influenza viruses within individual hosts despite prolonged infection

Published, in part as

Hirotaka Imai*, Jorge M. Dinis*, Gongxun Zhong, Ryan McBride, Mai thi Q. Le, Julie Eggenberger, Anthony Hanson, Michael Lauck, Yuko Sakai-Tagawa, Shinya Yamada, Sarmila Basnet, Alexander I. Karasin, David H. O'Connor, Yasuo Suzuki, Masato Hatta, James C. Paulson, Gabriele Neumann, Thomas C. Friedrich and Yoshihiro Kawaoka

* Hirotaka Imai and Jorge M. Dinis contributed equally

This chapter represents a manuscript in preparation for publication.

Abstract

Highly pathogenic H5N1 avian influenza viruses occasionally cause “spillover” infections in humans, but so far have not evolved the capacity for sustained transmission among humans. Understanding the evolutionary pathways by which such viruses could acquire human transmissibility is key to assessing their potential to cause a new pandemic. Influenza viruses exist in each host as a collection of genetically diverse “quasispecies,” which is thought to enhance their ability to adapt to changing selective pressures. Here we used deep sequencing and functional analyses to probe the genotypic and phenotypic diversity of H5N1 quasispecies in infected humans. We characterized viral quasispecies replicating 2 to 10 days after symptom onset. At these timepoints viruses are estimated to have undergone 12 to 68 rounds of replication in infected individuals. Models of viral adaptation typically assume steady linear evolution and predict that natural selection will favor the outgrowth of multiple human-adapting quasispecies by this time. Instead, we find that, despite genetic diversification, H5N1 viruses replicating during late infections in humans retain strikingly avian phenotypes in receptor binding, hemagglutinin thermostability, polymerase activity and interferon antagonism. We posit that avian influenza virus adaptation to humans does not occur via constant incremental fitness increases in each infected host. Instead, once the virus reaches peak titers, positive selection may not be capable of displacing the existing avian consensus. These findings suggest that population bottlenecks during within-host infection or between-host transmission may be required to generate human-transmissible H5N1 viruses in nature.

Introduction

As of May 2015, highly pathogenic avian influenza H5N1 viruses have caused 826 confirmed human cases resulting in 440 deaths. Human H5N1 infections are primarily caused by direct or indirect contact with infected birds. However, in some cases, limited direct contact transmissions from infected persons have been reported [52,80,81]. Currently, H5N1 viruses do not readily transmit person-to-person, but laboratory experiments revealed a limited number of substitutions that render H5N1 viruses transmissible in ferrets [47,65,82]. The ease in which respiratory-droplet transmissibility is acquired by avian H5N1 viruses in spillover infections is essential for evaluating viral pandemic potential.

Experimental adaptation revealed the molecular basis for which avian H5N1 influenza viruses evolve airborne transmissibility in mammals [47,65]. In two independent studies, mutations in the avian hemagglutinin (HA) protein were required to modify receptor binding, stability and glycosylation to support efficient replication and transmission in mammals. A shift in receptor-binding preference from avian-type (α2,3-Sia) to human-type (α2,6-Sia) was essential for viral attachment to mammalian respiratory tracts [60,83,84]. Stabilization of the HA protein mediated optimal pH-depended membrane fusion and genome release in more acidic conditions of mammalian nasal mucosa [47,65]. Loss of a glycosylation signal on the HA globular head increased transmissibility of H5N1 viruses in mammals [47,65]. Additionally, the polymerase complex of each transmissible virus was able to efficiently replicate at lower temperatures of human upper airways (33 to 37°C) compared to avian gastrointestinal tracts (37 to 42°C) [65]. Experimental adaptation revealed a few molecular changes that enable H5N1 respiratory-droplet transmission in mammals, demonstrating that the mutational barrier for airborne transmissibility in mammals is remarkably low, raising the concern that a H5N1 virus with pandemic potential could arise in nature.

The evolutionary dynamics that govern the generation and outgrowth of mam-

malian transmissible avian influenza viruses in nature are not completely understood. Due to error-prone genome replication, influenza viruses exist within an infected host as a collection of genetically similar variants, often referred to as “quasispecies,” capable of rapidly adapting to changing selective pressures. Recent mathematical models showed that viral mutation rate, the number of replication cycles in a given host, and natural selection can impact the likelihood at which transmissible H5N1 viruses are generated in infection [85,86]. Humans infections caused H5N1 viruses are long lasting, with viral nucleic acids being detected up to 15 days post symptom onset [87]. Therefore, models predict that viral quasispecies from prolonged human infections are more likely to accumulate the minimum set of mutations needed to confer transmissibility in mammals [16]. In addition to being generated, the transmissible variant needs to exist in the source population at sufficient frequencies to facilitate airborne transmission. We recently found that selective forces acting on the HA segment can impose a strong population bottleneck after respiratory droplet transmission of reassortant H5 influenza viruses [17]. Interestingly, in one infected animal, a variant in as little as 5.9% of viruses was transmitted via respiratory droplets, demonstrating that variants may not need to reach high frequency in one individual to be transmitted to another [17]. Therefore, thorough investigation of major and minor variants within H5N1 quasispecies in humans is needed to accurately evaluate avian H5N1 pandemic potential.

Here we use deep sequencing and functional analyses to evaluate the impact of natural selection on H5N1 adaptation in prolonged human infections. Analysis of deep sequences revealed levels of quasispecies genetic diversity that varied considerably human-to-human. Despite robust within-host genetic diversity and rapid changes in frequency of individual amino acid substitutions, tested H5 variants predominantly retained avian phenotypes in receptor binding, HA protein stability and interferon antagonism. We found relatively few mutations that enhanced polymerase activity (PB2-627K, NP-77K, NP-454G) in mammalian cells at 33°C. Interestingly, we found that

several mutations that increased in frequency during infection did not encode a beneficial phenotype. This suggests that selection was not acting on these specific residues to enhance viral fitness, but instead, the frequency increase was likely due to random chance. These findings suggest that despite genomic plasticity of H5N1 quasispecies in humans, functional diversity is highly constrained in spillover infections. This report establishes that natural selection did not favor the rapid outgrowth of human-adapting variants after continued replication in these spillover infections.

Materials and Methods

Biosafety and biosecurity

Biosafety protocols for isolation and sequencing of viral nucleic acids was approved by the University of Wisconsin—Madison's (UW's) Institutional Biosafety Committee after risk assessments conducted by the UW Office of Biological Safety. The UW Biosecurity Task Force regularly reviews the research program and ongoing activities of the UW Influenza Research Institute (IRI). The task force has a diverse skill set and provides support in the areas of biosafety, facilities, compliance, security and health. Members of the Biosecurity Task Force are in frequent contact with the principal investigator and personnel of the IRI to provide oversight and assure biosecurity. Isolation of RNA from samples containing H5N1 reassortant viruses was performed in enhanced BSL3 containment laboratories approved for such use by the CDC and the USDA following procedures approved by the UW Office of Biological Safety. RNA was isolated using techniques documented to inactivate virus particles in samples before the removal from the BSL3 laboratory space. DNA library preparation and sequencing were performed in a BSL2 laboratory space.

RNA extraction, cDNA synthesis and PCR amplification

Total RNA was extracted directly from clinical specimens using the QIAamp MinElute Virus Spin Kit (Qiagen, Germany). Viral RNA segments 1, 2, 3, 4, 5 and 8, which respectively encode PB2, PB1, PA, HA, NP and NS genes were reverse transcribed using the SuperScript III reverse transcriptase (Invitrogen, California, USA). The resulting cDNA was PCR amplified using a high fidelity DNA polymerase with primers specific to the terminal or internal sequences of each gene segment (Table S1 for primers). Primers were designed against consensus sequences generated from viruses passaged in MDCK cells using the Hoffmann et al. universal primers [45]. PCR products were gel-purified and bands of the expected length were excised using the MinElute Gel Extraction Kit (Qiagen, Germany).

Illumina MiSeq Sequencing

Amplified gel-purified PCR products were quantified by using the Qubit dsDNA High Sensitivity Kit (Invitrogen, USA). Samples were prepared for sequencing on the Illumina MiSeq platform using the Nextera XT DNA Sample preparation Kit (Epicentre, USA) according to manufacturer's instructions. Limited-cycle PCR products were purified with two washes of 0.375X AMPure XT beads (Beckman Coulter, USA) and eluted in 10 μL of resuspension buffer. Eluted template was quantified using the Quibit dsDNA High Sensitivity Kit (Invitrogen, USA) and the average fragment length was determined using the Agilent High Sensitivity Kit (Agilent technologies, USA). Fragmented and indexed samples were pooled in equimolar amounts into separate 1 nM libraries and diluted to 8 pM for sequencing on the Illumina MiSeq. Illumina MiSeq run settings were Workflow: DenovoAssembly; Assay: Nextera; Chemistry: Amplicon; and Reads: 250 x 250 with automatic adapter removal.

Computational methods

Illumina MiSeq sequences were imported into CLC Genomic Workbench, Version 7.3 (CLC bio, Denmark). Reads were trimmed using a quality-limit threshold of 0.001 and reads ≥100 base pairs in length were retained. De novo assemblies were generated using from a single patient sample. From these de novo assemblies, a preliminary consensus sequence was extracted using majority rule (i.e., using the predominant nucleotide at each position). The preliminary consensus sequences was used as a scaffold for subsequent reference mappings in order to reduce indel-associated errors. SNPs were called from the reference mappings using CLC Genomic Workbench, Version 7.3 at nucleotide positions with at least 100 sequence reads covering each nucleotide position, a central base quality score of Q30 or greater, and if detected in two or more sequence reads. SNPs detected at or near the HA receptor binding domain in ≥ 5% of viral sequences were assessed for linkage to adjacently occurring mutations using a custom perl script called LinkGE. The LinkGE source code is freely available on

http://dholk.primate.wisc.edu/project/dho/public/LinkGe/begin.view. Phylogenetic tree substitution model was determined using Datamonkey webserver with the AIC-based model selection procedure (Generalized time reversible; GTR). Trees were generated using the ATGC online PhyML webserver. Trees were edited using the ggtree package in R. Mapping amino acid variations on a 3D HA structure was performed using Pymol (http://www.pymol.org) with the A/Vietnam/1203/2004 H5N1 HA (Protein Data Bank accession 2FKO). The nucleotide diversity statistics πN and πS were calculated in PoPoolation version 1.2.2 using subsampled sequence mappings containing 1000 randomly chosen sequences per nucleotide position to minimize potential coverage bias.

Cells

Human embryonic kidney 293T cells were maintained in Dulbecco's modified Eagle's MEM (DMEM) with 10% fetal calf serum. Madin-Darby canine kidney (MDCK) cells were maintained in minimal essential medium (MEM) with 5% newborn calf serum. Both cells were cultured at 37°C in 5% CO2.

Reverse genetics

HA gene consensus sequences were cloned into pPolI plasmids as previously described [88]. Individual HA amino-acid substitutions were incorporated into plasmid by site-directed mutagenesis. The sequence of HA gene plasmids were verified by Sanger-based sequencing. Plasmids encoding genes of A/California/04/2009 (H1N1; CA04) were used as a backbone for reassortant viruses. Reassortants possessing an avian H5 gene and the remaining genes derived from CA04 were generated in 293T cells by reverse genetics [88], and amplified in MDCK cells. To obtain large-scale stocks of H5-reassortants viruses for receptor binding assays, viruses were amplified again in MDCK cells unless stated otherwise. The sequence of HA gene segments after cellular passaging was confirmed by Sanger-based sequencing. H5-reassortants UT31312III-HA-203P virus and UT36282I-HA-138V respectively gained an unwanted mutation (203L) or reverted back to the consensus residue. In addition, UT31413II-HA-

486H reassortant virus was not successfully rescued and was therefore omitted. Reassortant virus possessing the HA genes of A/Kawasaki/173/2001 (H1N1; K173) virus [51] in the CA04 backbone was prepared by reverse genetics. All experiments with infectious viruses possessing the H5 HA with polybasic cleavage site were performed in enhanced biosafety level 3 (BSL3+) containment laboratories approved for such use by the Centers for Disease Control and Prevention (CDC) and the United States Department of Agriculture (USDA).

H5-Reassortant virus inactivation and purification

Viruses were amplified in MDCK cells, clarified by low-speed centrifugation, and inactivated with 0.1% β-propiolactone over night. After another low-speed centrifugation, inactivated viruses were purified by ultracentrifugation at 25,000 rpm for 2h at 4°C over 30% and 70% sucrose in PBS. Layer containing viruses was collected, diluted in PBS, laid over a cushion of 30% sucrose in PBS, and ultracentrifuged at 25,000 rpm for 2 h at 4 °C. The pellets were resuspended in PBS, aliquoted and stored at −80°C.

Solid-phase binding assay

Solid-phase binding assays were performed to assess the direct receptor-binding capacity of viruses as previously described [47] with slight modification. Nunc MaxiSorp flat-bottom 96 well plate (Thermo scientific, USA) were incubated with sodium salts of sialylglycopolymers (poly-l-glutamic acid backbones containing N-acetylneuraminic acid linked to galactose through either an α2,3 (Neu5Acα2,3Galβ1,4GlcNAcβ1-pAP) or an α2,6 (Neu5Acα2,6Galβ1,4GlcNAcβ1-pAP) bond) in PBS at 4°C overnight with constant rocking and irradiated under ultraviolet light at 254 nm for 2 min. After the sialylglycopolymer solution was removed, the plates were washed five times with 300 μl of PBS at room temperature and then blocked with 300 μl of 20% Blocking One (Nakarai Tesque Inc., USA) at room temperature for 1 h followed by five washes. Purified inactivated viruses diluted with PBS containing 2% bovine serum albumin (Sigma, USA) to 25 ng of M1 protein amount were incubated in the wells of the

sialylglycopolymer-precoated plates (50 μl/well) at 4°C overnight with constant rocking. After five washes, the plates were incubated for 2 h at 4°C with rabbit polyclonal antiserum to either K173 virus or VN1203 virus. The plates were washed and incubated with Goat Anti-Rabbit IgG (H+L), Horseradish Peroxidase Conjugate (Life Technologies, USA) for 2 h at 4°C. After washing at room temperature, the plates were incubated with 100 μl of TMB Substrate (Thermo scientific, USA) for 10 min at room temperature and the reaction was stopped with 100 μl of 0.18 M H2SO4. The absorbance at 450 nm and 570 nm was determined in a plate reader Infinite M1000 (Tecan, Switzerland). Washes were done using 300 μl of ice-cold PBS containing 0.05% Tween 20 (PBST) unless stated otherwise.

Tissue-virus binding assay

HA receptor binding specificity on human tissue sections was assessed as previously described [50] with modifications. For labeling, each inactivated virus was diluted with PBS(-) to 850 μl which contained the same M1 protein amount normalized by western blot and incubated with 1.0 M carbonate-bicarbonate buffer (pH 9.5) and 2 mg/ml FITC isomer I (Life Technologies, USA) at a volume ratio of 17:2:1 for 1 h at room temperature with constant rotating. Unbound FITC was cleared from the resulting labeled viruses by using PD MidiTrap G-25 (GE Healthcare Life Sciences, USA). The amount of M1 protein for each labeled virus was quantified as described below. Paraffin-embedded human adult normal trachea and lung sections were purchased from US Biomax, Inc. Sections were deparaffinized with xylenes and hydrated with ethanol followed by PBS. A hydrophobic barrier was drawn around each section with liquid blocker super pap pen (Cosmo Bio Company Ltd, Japan). Each tissue section was blocked with Carbo-Free Blocking Solution (Vector Labs, USA) and TNB Blocking Buffer (PerkinElmer, USA) at room temperature for 1 h, respectively. The tissue sections were incubated with each virus containing a normalized amount of M1 protein at 4°C for 16 h. The virus-treated sections were incubated in horseradish peroxidase conjugated

polyclonal rabbit anti-FITC antibody (Dako, USA). Signals were detected with AEC+ Substrate-Chromogen (Dako, USA) according to the manufacturer's instructions, followed by counterstaining with Mayer's hematoxylin (Sigma Aldrich) and mounting with Shandon Immu-Mount (Thermo Scientific, USA). The sections were observed under the upright microscope, Axio Imager.A2 (Zeiss, Germany) with a 100 X objective lens. To remove unbound residuals, sections were washed in cold PBS for 10 min 5 times each after virus incubation and antibody incubation. Signal specificity was confirmed by treating sections with an α(2->3,6,8,9) neuraminidase from Arthrobacter ureafaciens (Sigma Aldrich, USA). Tissue sections were kept in humidified containers for all incubations longer than 3 min to prevent drying.

Quantification of M1 protein

Protein amount of influenza virus matrix protein M1 was measured by western blotting. For deglycosylation of viral proteins, FITC-labeled viruses were treated with PNGase F (New England BioLabs, USA) according to the manufacture's protocol. Recombinant Influenza A Virus H1N1 M1 (A/California/04/2009), His-tagged (eEnzyme, USA) was deglycosylated and was used for protein standards. The deglycosylated samples were combined with NuPAGE LDS Sample Buffer (Life Technologies, USA) and NuPAGE Sample Reducing Reagent (Life Technologies, USA). Combined samples were heated at 70°C for 10 min and separated on NuPAGE Novex 10% Bis-Tris Midi Protein Gel (1.0mm 26 well, Life Technologies, USA) set in a box filled with NuPAGE MES SDS Running Buffer (Life Technologies, USA). Samples were transferred to a PVDF membrane (Life Technologies, SUA) by using iBlot Gel Transfer Device (Life Technologies, USA). The membrane was blocked with Blocking One reagent (Nacalai Tesque Inc., USA) and incubated in rabbit polyclonal antiserum to VN1203 followed by Goat Anti-Rabbit IgG (H+L), Horseradish Peroxidase Conjugate (Life Technologies, USA) and Super Signal West Dura Extended Duration Substrate (Thermo Scientific, USA). The prepared membrane was washed in PBS containing 0.05% Tween 20

(PBST) after blocking and antibody incubations. FluorChem HD2 system (Proteinsimple, USA) was used for chemiluminescent imaging. Bands corresponding to M1 protein were quantified by using AlphaView SA (Proteinsimple, USA).

Thermostability assay

Thermostability of the H5-reassortant viruses were assessed as previously described [47]. Briefly, viruses containing 64 HA units/50 µl in MEM supplemented with 0.3% BSA were heat-treated at 55°C for the indicated times. The haemagglutination activity of the heat-treated viruses was determined by HA assay using 0.5% TRBCs. The virus infectivity was determined by standard plaque assay in MDCK cells.

Mini-replicon assay

Viral polymerase activity of influenza polymerase complex was assessed as described previously [92] with slight modifications. The consensus sequence of PB2, PB1, PA and NP gene segments of each virus was cloned into pCAGGS plasmids [90]. If a segment was not successfully generated from the clinical specimen, plasmids were generated using genes from the respective H5N1 isolate passaged in MDCK cells. Site-directed mutagenesis was performed to obtain polymerases and NP segments with targeted amino acid substitutions found in human infection. Target sequences of plasmids were confirmed by Sanger-based sequencing. We seeded 293T cells in 24-well poly-D-lysine coated plates with 50000 cells per well. Plated cells were incubated for 24 h. The cells were transfected with 100 ng each pCAGGS plasmids encoding the polymerase protein PB2, PB1, PA, and NP, with 20 ng of plasmid encoding the firefly luciferase gene flanked by partial influenza virus genome sequences (pPolWSNNA F-Luc) and 10 ng of an internal control plasmid pGL4.74[hRluc/TK] (Promega, USA) by using TransIT293 (Mirus, USA). Cells were incubated at 33 or 37 °C. At 24 h post transfection the cells were lysed with 1×Passive Lysis Buffer (Promega, USA). Luciferase activity was determined using the Dual-Luciferase Reporter Assay System (Promega, USA) according to manufacturer's protocol. Firefly luciferase values were divided by

Renilla luciferase values to normalize transfection efficiency.

IFN Reporter Assays

To assess the effect of NS1 variations on virus-stimulated expression from IFNβ promoter, a reporter assay was performed as previously described [89] with some modifications. The consensus sequence of NS gene segment of UT36250I virus was cloned into pCAGGS plasmid [90]. Site-directed mutagenesis was performed to obtain the UT36250I-NS possessing methionine at position 124. The plasmid sequence was confirmed by Sanger sequencing. 293T cells were seeded in 24-well poly-D-lysine coated plates at 50000 cells per well and incubated for 24 h. The cells were transfected with 1000 ng of IFNβ promoter reporter plasmid (pGL-IFNβ luc) [91] by using TransIT293 (Mirus, USA). 100 ng of pCAGGS plasmid encoding NS1, pCAGGS plasmid encoding FLAG-GFP (pC-FLAG-GFP) or pCAGGS empty vector were co-transfected. After incubation for 24 h, cells were treated with 1000000 fluorescence forming units of SeV (Cantell strain). At 48 h post-transfection, cells were lysed with 1×Passive Lysis Buffer (Promega, USA). The cell lysate was mixed with Luciferase Assay Reagent II (Promega, USA) and the luciferase activity was determined.

Similarly, to assess the effect of NS1 variations on IFN-stimulated expression from an interferon-stimulated response element (ISRE), a reporter assay was performed as previously described [89] with some modifications. 293T cells were seeded in 24-well poly-D-lysine coated plates at 50000 cells per well and incubated for 24 h. The cells were transfected with 1000 ng of ISRE reporter plasmid (pISRE-Luc; Clontech, Japan) using TransIT293 (Mirus, USA). 100 ng of pCAGGS plasmid encoding NS1, pCAGGS plasmid encoding FLAG-GFP [92] or pCAGGS empty vector were co-transfected. Cells were incubated for 24 h and treated with 100 units of Human Interferon Beta 1a, mammalian (PBL Assay Science, USA). At 30 h post-transfection cells were lysed with 1×Passive Lysis Buffer (Promega, USA). The cell lysate was mixed with Luciferase Assay Reagent II (Promega, USA) and the luciferase activity was determined.

Results

Clinical characteristics of human H5N1 infection

To examine the genetic and functional diversity of H5N1 quasispecies in humans, we obtained clinical specimens from infected patients in northern Vietnam between 2004 and 2010 (n = 7; Table 1). Classical characterization of avian influenza viruses is performed using isolates passaged in eggs or MDCK cell lines. Passaging of human isolates in eggs and cells can select for variants with altered receptor binding [93-95]. To avoid generating and characterizing mutations that arise due to viral propagation, we examine quasispecies diversity directly from human respiratory specimens. Once hospitalized with suspected H5N1 infection, a throat swab was taken from each patient. A second throat swab, or if intubation was required, a tracheal aspirate was taken to confirm the initial diagnosis (Table 1). Previous lectin staining revealed that human pharynx and trachea tissues mainly express with α2,6-Sia [96] . Moderate differences in air temperature of human pharynx (33°C) and tracheal (35°C) has been reported [97]. Therefore, it is possible that anatomic location of viral replication in the human airways may affect the local composition of viral quasispecies during infection and cause sample-to-sample genetic variability. Humans normally develop symptoms less than 7 days since last exposure to H5N1-infected poultry [98] . In our study, human-patient samples were collected 2 to 10 days post-symptom onset for individuals with whom this information was available (Table 1). We therefore estimate that, in these human patients, avian H5N1 viruses had replicated for 3 to 17 days and undergone 12 to 68 rounds of replication (6 hours per replication cycle) [99,100]. Deep sequencing quasispecies late in human infection provides a unique opportunity to measure genetic and functional evolution of H5N1 viruses in single spillover infections after continued replication in mammalian cells.

Deep sequencing H5N1 segment diversity directly from clinical specimens

To examine H5N1 viral genetic diversity in late human infections, we amplified vi-

ral nucleic acids extracted directly from the throat swabs (TS) or tracheal aspirates (TA). Considering the limited quantity of viral nucleic acid extracted from these samples, we focused our efforts on viral segments encoding proteins associated with mammalian adaptation or transmissibility: polymerase basic protein 2 (PB2) [58,65,101,102], polymerase basic protein 1 (PB1) [102], polymerase acidic protein (PA) [102], nucleoprotein (NP) [102], HA [47,65] and the antagonist of interferon-mediated immune response by the nonstructural protein 1 (NS1) may contribute to host adaption during spillover infection [103]. Notably, polymerase complex and some HA segments were difficult to amplify; therefore, we divided each polymerase gene into two smaller amplicons and, for some samples, focused on a region around the HA receptor-binding domain (see methods for details; Table S1 for primers used). Amplified genes were prepared for deep sequencing using the Illumina MiSeq. High-quality sequences above a PHRED score 30 (i.e., 1 error per 1000 sequenced nucleotides) were used to generate a consensus sequence for each sample. Importantly, because the sequence of the initial infecting virus was not known, sequence reads were mapped against their own sample's consensus. On average, each assembly yielded approximately 5000-fold coverage per segment. From these mappings, we called genetic variation for each sample, resulting in the identification of synonymous (silent mutations) or nonsynonymous (amino-acid-changing mutations) detected at or above our quality cutoff (1%) and in less than 50% of sequences in a virus population.

Detection of genetic makers associated with H5N1 mammalian adaptation

We compared our H5N1 consensus sequences against publically available H5N1 sequences generated from birds, chickens, humans and environmental sites from Vietnam since 2004. First we generated individual maximum-likelihood phylogenies using near-full-length nucleotide sequences for PB2, PB1, PA, HA, NA and NS1 gene segments (Figure 1). We found that our segments clustered with previously circulated bird and chicken viruses. Based on nucleotide Basic Local Alignment Searced-

Tool (BLASTn), we found that our gene segments were ≥ 97% similar to bird, chicken and environmental viruses; Only PB2, PB1 and HA genes of UT3040 and the NP gene of UT31312 were identical reference isolates based on amino acid sequence, respectively (Table S2).

Past studies have identified molecular markers associated with H5N1 adaptation and transmissibility in mammals [47,58,77,82]. To identify genetic features associated with H5N1 mammalian adaptation and transmissibility, we queried our consensus segments against an inventory of publically residues associated to mammalian adaptation. Across all sequenced gene segments, we found 14 amino acid positions, which possess a residue associated with mammalian adaptation (Table S3). We found 3 PB2 segments with a canonical lysine at amino acid position 627 associated that increases polymerase activity at lower temperatures of the upper humans airways, and with additional mutations in HA, enhances respiratory-droplet transmissibility in mammals [58,65,77]. The PB2 K627 was detected at variable frequencies at 92.6%, 64.7% and 51.7% of viral sequences for UT3040I, UT31312III and UT31394II samples, respectively. All other residues associated with enhanced polymerase activity, human-type receptor binding specificity or mammalian virulence was fixed in their respective quasispecies. These data indicate H5N1 viruses sequenced from humans possess the genetic hallmarks mammalian-type receptor binding and increased polymerase activity, suggesting that some of these viruses likely adapted to humans.

Within-host evolution of H5N1 viruses in humans.

Error prone genome replication generates high rates of deleterious mutations that must be purified from the virus population to maintain fitness [29]. To estimate signatures of within-host natural selection we calculated nucleotide diversity using the statistic π (see methods for details). In brief, this statistic calculates the average pairwise substitution rate of nonsynonymous (πN) and synonymous (πS) mutations within a given sequence dataset. Comparing these values provides information about the

"direction" of natural selection. Generally, a πN/πS ratio > 1 indicates that positive selection is favoring genetic diversification. By contrast, a πN/πS ratio < 1 indicates that purifying (or negative) selection is acting to maintain a fit virus population by removing deleterious mutations. We estimated natural selection across each gene segment for all available samples by re-calculating πN and πS measurements using a sliding window of 30 nucleotides (10 codons) and a step size of 15 nucleotides (5 codon). Despite detecting localized peaks of greater nonsynonymous diversity, we found that gene-wide ratios of πN and πS were strongly indicative of purifying selection. This indicates that natural selection favors the removal of amino-acid changing mutations from the virus population at later time points in human infection (i.e., πN/πS ratio < 1; Figure 2). These data provide evidence that selection appears to favor the rapid removal of amino-acid changing mutations from the virus population in humans. The evolutionary advantage of maintaining a diverse quasispecies is that when selective pressures rapidly change, a variant possessing a fitness advantage may already exist in the population [28]. Therefore, we enumerated single nucleotide polymorphisms (SNPs) detected above our experimentally defined quality cutoff of ≥ 1% [17] and found 250 nonsynonymous and 145 synonymous mutations from all available gene segments (Figure S1). Interestingly, the number of SNPs detected within individual viral quasispecies varied considerably human-to-human (Table S4). That is, in some samples, quasispecies were highly diverse (e.g., UT36250I: 56 nonsynonymous and 20 synonymous mutations), while for others, we observed little genetic variability (e.g., UT31312I: 5 nonsynonymous and 6 synonymous). Taken together, these data reveal that in H5N1 quasispecies possess wide levels of within-host nucleotide diversity that could facilitate host adaptation to humans. However, these data provide conceptual support that strong purifying selection may be hindering the onward evolution of H5N1 viruses in humans by limiting the amout of genetic diversity maintained during infection.

Preparation of plasmid clones and viruses

To determine the functional diversity of H5N1 quasispecies, we generated plasmid clones and viruses by reverse genetics possessing single amino-acid-changing polymorphisms found in late human infection. We found 38 candidate amino nonsynonymous polymorphisms above this threshold. Also included in the candidate list are viral variants that rapidly changed in frequency (Figure 3 and Table 2), allowing us to determine the fitness advantage of variants putatively favored by positive selection. H5N1 plasmid clones or viruses bearing single point mutations are abbreviated based on sample metadata. For example, UT31394I-HA-T207 denotes variant that possess a threonine residue at amino acid position 207 in the HA segment of patient sample UT31394I. Mutations found in PB2, PB1, PA, NP and NS1 were introduced into their autologous consensus sequence to generate plasmid clones. Whereas HA polymorphisms were introduced into their autologous HA consensus sequence and viruses bearing all remaining genes from an A/California/04/2009 (H1N1; CA04) virus were generated by reverse genetics. Notably, for samples in which only a small region surrounding the HA receptor-binding domain (RBD) was sequenced, the consensus sequences from other available paired sample was used in replacement. Not all engineered viruses were rescued after cellular propagation. Only reverse-genetic viruses rescued as a pure population were used for functional characterization. Lastly, an isoleucine at amino acid position 511 of UT31413II was not generated because this mutation is outside the outside the 3D structure (Protein Data Bank accession 2FKO).

Receptor-binding specify of H5N1 quasispecies from prolonged human infection

To characterize the receptor-binding properties of mutations in the H5 HA globular head, we used solid-phase binding assays with synthetic sialylglycopolymers absorbed to plates that were then incubated with virus. We tested 5 consensus HA segments (herein called majority variants) and 8 nonsynonymous polymorphisms (herein called minority variants) located at or near the HA receptor-binding pocket (Figure 4A). A virus possessing the HA gene from seasonal human A/Kawasaki/173/2001 (H1N1;

K173) and the remaining genes from A/California/04/2009 (H1N1; CA04) (K173/CA04) served as a control virus for human-type receptor specificity. As expected the K173/CA04 virus preferentially bound to α2,6-Sia glycans (Figure 4B). In contrast each majority virus bound to α2,3-Sia glycans (Figure 4B), showing that consensus virus retained avian-like receptor binding despite numerous rounds of replication in human cells. As viruses continue to replicate, deleterious mutations are generate and subsequently removed from the population to maintain viral fitness. Therefore, we hypothesize that mutations detected in only one timepoint do not encode for human-type receptor-binding specificity. As expected, UT3040I-HA-138V, UT31312III-HA-186K, UT31312III-HA-206P variants bound to α2,3-Sia glycans. However, UT3040I-HA-186K was capable of binding to for α2,3-Sia and α2,6-Sia glycans, suggesting that dual avian- and human-type binding. The UT36250I/II-HA-138V and UT36282I/II-HA-138V variants were found to respectively decline in frequency from 28.0%→19.8% and 6.1%→5.3% in infection and bound α2,3-Sia glycans. We found 2 HA globular head polymorphisms that increased in frequency by more than 5% in infection. The UT31394II-HA-207T increased in frequency from 21.0%→47.2% between paired timepoints and is located in a region adjacent to the receptor-binding domain. The second variant, UT36250I/II-HA-226K increased 3.9%→14.0% of viral sequences, is located in the receptor-binding pocket and makes direct interactions host-cell receptors. Mutations that confer a fitness advantage will increase in a virus population by the action of positive selection. Surprisingly, UT31394II-HA-207T and UT36250I/II-HA-226K did not bind α2,6-Sia, but exclusively bound α2,3-Sia receptors, suggesting that these residues were not favored by positive selection.

Attachment of H5N1 quasispecies to human respiratory tissues

Human respiratory tracts are decorated with a wide spectrum of α2,3-Sia and α2,6-Sia glycans [105] . Lectin staining of human airways tissues show epithelial cells pharynx and trachea mainly express α2,6-Sia [96] . To evaluate the effects of H5 HA

globular head mutations on viral attachment to human respiratory airways, sections of tracheal tissues were exposed to K173/CA04 (human-type receptor binding) and H5-reassortant viruses bearing HA mutations found in the H5 HA globular head (Figure 4B). The K173/CA04 virus bound largely to α2,6-Sia expressing ciliated epithelial cells and non-ciliated goblet cells of trachea tissue cross-sections. The UT31312III-HA, UT31312III-HA-186K, UT31312-HA-206P, UT31394II-HA, UT31394II-HA-207T, UT36250I/II-HA, UT36250I/II-HA-138V, and UT36250I/II-HA-226K viruses displayed little to no binding to trachea epithelia. The UT3040I/II-HA majority virus weakly bind to goblet cells, but not to epithelial cells. Unlike the solid-phase binding assays, UT36282I-HA and UT36282I-HA-A138V did appreciably bind to α2,6-Sia expressing epithelial cells in trachea tissues (Figure 4B). When tracheal tissues were pre-treated with *Arthrobacter ureafaciens* neuraminidase, which cleaves non-reducing terminal sialic acids, virus binding was dramatically reduced and signals were caused by viral attachment to sialyl receptors (data not shown). These data indicate that H5N1 quasi-species do not appreciably bind to α2,6-Sia expressing epithelial cells in trachea tissues with the exception of UT36282I-HA and UT36282I-HA-A138V which may effect virus attachment to human airways.

HA protein stabilization

For efficient replication in and transmission between mammals, the H5 HA protein may require stabilization to overcome inactivation due to acidic environmental conditions of the human nasal mucosa [47,65,82,106]. Heat treatment at a neutral pH promotes a fusogeneic form of the HA protein and serves as a substitute assay to measure HA protein stability [47,82,107,108]. We tested HA stabilization for mutations detected external to the receptor-binding domain (Figure 5C). The K173/CA04, VN1203/CA04 and H5-reassortant viruses were incubated at 55°C at 15-min time intervals, after which the loss of infectivity and activity were respectively determined by plaque assay and hemagglutination assay using turkey red blood cells. We found

that the K173 HA was more stable than the avian VN1203 HA protein by heating for 120 min. Thermostability was evaluated for 3 majority and 3 minority variants, and each lost infectivity by heating for 120 min (8-8.5 log10 decrease in titre). But, we did find that UT31312II-HA and UT31312II-HA-92K were modestly more resistant to heat treatment (5 log10 decrease in titre by minute 60). In hemagglutination assays, viruses tested from patient UT31394 rapidly lost activity relative to those found within patient UT31312 demonstrating strain-to-strain variability of HA protein stability. Even though UT31394II-HA-54(+1)E (possessing glutamic acid at the position between 54 and 55 in H3 numbering) was found to increase in frequency from 33.5% to 45.5% in viral sequences, the glutamate substitution did not confer enhanced HA protein stabilization. Taken together, we show that the HA mutations tested did not promote enhanced HA thermostability in tissue culture.

Polymerase activity of H5N1 quasispecies

Avian polymerase activity is impaired in mammalian cells resulting in restricted transcription and replication of the viral genome in mammals [109,110]. Adaptive mutations in genes that encode the polymerase complex (PB2, PB1 and PA), and nucleoprotein (NP) have been shown to overcome restriction in mammalian cells [111] . A well-known mammalian host adaptation mutation, a lysine at amino acid position 627 of PB2, regulates polymerase activity in a species-specific fashion [109]. Viruses possessing PB2 627K replicate efficiently at lower temperatures characteristic of human upper airways (33 to 35°C) [77]. To identify mutations generated in prolonged human infections that enable efficient replication, we used a mini-replicon system to measure polymerase activity in mammalian cells. Polymerase activity of individual mutations were evaluated on human embryonic kidney 293T cells at 33°C and 37°C (Figure 6). We found the viral ribonucleoprotein complex (vRNP; PB2, PB1, PA and NP) of UT3040I, which encodes a PB2 627K residue, was significantly more active than other majority vRNPs tested ($P<0.02$, Student's t-test; Fig. 6), with the exception of UT31394II at

37°C. We found that the UT3040I polymerase activity was not significantly different than K173 vRNP (P=0.71, Student's t-test), suggesting that this polymerase complex is well suited for replication in mammalian cells. We then focused on the effect of minority amino acid polymorphisms detected within humans, we found that minority variants encoded a wide range of polymerase activities at 33°C and 37°C (Figure 6). Notably, the PA gene of UT3040I possessed a methionine at amino acid position 90 (UT3040I-PA-90M) and a glutamate at amino acid position 143 (UT3040I-PA-143E) that resulted in a respective 2.5- and 2-fold increase in polymerase activity relative to UT3040I vRNP. Interestingly, the UT3040I-PA-90M and UT3040I-PA-143E vRNPs were more active than the seasonal human K173 control; however, each of these putative human-adapting mutations was subsequently removed from the virus population (see Figure 3.). Furthermore, UT31394I/II-NP-K77 and UT31394I/II-NP-G454 resulted in 1.3- to 4-fold increase in polymerase activity relative to their autologous majority virus. Interesting, UT31394I/II-NP-K77 and UT31394I/II-NP-G454 increased in frequency from 7.7→21.0%% and 8.8%→21.0%, respectively, suggesting that these mutations were favored by positive selection in late human infection. As anticipated, vRNPs possessing a PB2-627K residue were more active in mammalian cells at lower temperature than those with an avian signature glutamic acid (PB2-627E). These data show that mutation detected late in human infection can significantly improve viral polymerase activity at lower temperatures in mammalian cells. Based on genetic and functional characterization, we show that UT31394I/II-NP-K77 and UT31394I/II-NP-G454 were favored by natural selection. Strikingly, mutations that improved polymerase activity did not always increase in frequency in the virus population. These data suggest variants possessing beneficial traits may not deterministically arise during infection, suggesting that other forces are hindering the emergence of fit viral variants in spillover infections.

NS1 IFN antagonistic properties

Influenza virus nonstructural gene (NS) encodes for a multifunctional protein

(NS1) that acts to counteract host innate interferon (IFN) response [112]. Avian H5N1 viruses are relatively resistant to the antiviral effects of host interferon responses [113]. Point mutations in NS1 can modulate the virulence of highly pathogenic H5N1 strains [114]. A functional balance in influenza virulence is likely important for viral emergence, too virulent a strain will results in rapid host death that limits the potential for onward respiratory-droplet transmission. Furthermore, phylogenetic trees of NS genes for H5N1 viruses show that these segments evolve in a species-specific fashion, suggest that NS may play a role in host adaptation [103,115]. We only found single nonsynonymous polymorphism, a methionine at amino acid position 124 of NS1 in sample UT36250I that matched our functional characterization criteria. The UT36250I-NS1-124M was found to increase in infection from 14.2%→27.8% of viral sequences. We evaluated the impact of the UT36250I-NS1-124M variant on IFN antagonistic properties by using a reporter plasmid encoding firefly luciferase gene under the control of IFNβ promoter (pGL-IFNβ luc) [91] and the control of interferon-stimulated response element (ISRE) (pISRE-luc). We used an empty vector and GFP plasmids as negative control of IFN antagonism. We transfected 293T cells with UT36250I-NS1 and UT36250I-NS1-124M plasmids to determine their ability to antagonize mammalian INF responses. We found that UT36250I-NS1, UT36250I-NS1-124M, K173 possessed similar capacities to antagonize IFN responses under the control of IFNβ (Figure S2, panel A) and ISRE promoters (Figure S2, panel B).

Discussion

Mathematical models have recognized several factors that impact the probability in which respiratory droplet-transmissible H5N1 viruses are generated in spillover infections [13,16]. These models assume that a virus population grows exponentially, and once they reach peak titers, the population remains at a constant size resulting in the steady accumulation of substitutions during infection [13,16]. Polymerase fidelity, the number of replication cycles, and the strength of selection acting on individual mutations raise the probability that transmissible H5N1 viruses are generated in a host [16]. Here, we use deep sequencing and functional analyses to evaluate within-host evolution of H5N1 viruses in prolonged human infections and to provide experimental evidence to inform these models. In accordance with within-host evolution models, we found that viral quasispecies were genetically diverse after ~ 12 to 68 rounds of replication. Despite within-host genetic diversification, viral quasispecies largely retained avian-like phenotypes even after continuous replication in human cells. We found some mutations that increase polymerase activity at temperatures of human upper airways (e.g., PB2-K627, NP-77K and NP-G454). Strikingly, the vast majority of mutations detected were removed from virus population, even when the variant encoded for beneficial traits as determined by in vitro characterization (e.g., PA-M90 and PA-E143). We provide evidence that the principle evolutionary force shaping H5N1 quasispecies in late human infection was purifying selection, and rapid removal of genetic diversity may limit the onward evolution of H5N1 viruses in nature.

The evolutionary forces that govern mammalian adaptation and the emergence of viruses capable of transmission among humans remain unclear. RNA virus replication generates robust within-host genetic diversity that can rapidly change due to varying selective pressures. Error-prone replication leds to the accumulation of deleterious mutations that must be purified from the virus population to maintain viral fitness [64,79,116]. Pastore and colleagues found that deleterious mutations can serve as ge-

netic intermediates whereby subsequent mutations result in net fitness gain and virus survival [117]. Removal of deleterious mutations by purifying selection may limit viral adaptation to new hosts [118]. We predicted that positive selection would favor the rapid outgrowth of variants possessing mammalian-type traits in humans. Our results suggest contrary, with deleterious and beneficial mutations being removed from viral quasispecies. Recent models predict that within-host mutant emergence is less like to evolve after the wild-type-virus reaches peak titers in a host [86]. We hypothesize that the selective advantage of beneficial mutations were not strong enough to overcome the existing high-titer avian consensus. We speculate that reductions in population size during within-host infection or between-host transmission (i.e., bottlenecks) may be needed to re-organize the quasispecies to facilitate mammalian adaptation to humans

Determining which avian influenza virus will cause the next pandemic is impossible, but spillover infections provide an opportunity for influenza adaptation to mammals. Identifying and evaluating factors that impact the likelihood of viral emergence can inform preventative and control measures to decrease global morbidity and mortality. We recently showed that positive selection impacts the initial stages of H5 HA segment adaptation in ferrets, and thereafter purifying selection takes over to remove deleterious mutation late in infection to maintain fitness [17]. In our present study, we show that late in human infection, purifying selection limits the accumulation of within-host genetic diversity in the viral quasispecies. Additional studies are needed to evaluate the potential impact of natural selection on H5N1 adaptation early in human infection. Also, considering that the action of natural selection is dependent on the strain and host, other H5N1 strains may be differentially suited to evolve in mammals. Finally, as avian influenza viruses continue to evolve in nature, potentially closer to a mammalian transmissible phenotype, performing deep sequencing and functional assessments of viral quasispecies directly from biological samples may be important for early detection of potentially pandemic viruses.

Acknowledgments

The authors would like to acknowledge Yohei Watanabe at Osaka University for valuable technical advice. We thank Peter Halfmann for providing pGL-IFNβ luc, Seiya Yamayoshi for providing pC-FLAG-GFP, Peter Jester and Kelly Moore for technical support and Takeo Gorai and Eileen A. Maher for helpful discussions.

Figure Legends

Figure 1. Phylogenetic relationships of H5N1 influenza A genes sequenced from birds, humans and environmental sites in Vietnam. Maximum-likelihood phylogenies were created for H5N1 genes using consensus sequences generated from human clinical specimens. In some instances, clinical samples yielded insufficient sequence coverage for consensus generation (see Methods), for these genes we used consensus sequences generated from virus isolates passaged in Madin-Daryby canine kidney (MDCK) cells. Trees were built using partial length gene sequences with reference sequences from the Influenza Research Database project from Vietnam. The Generalized time reversible substitution model was determined using Datamonkey webserver with the AIC-based model selection procedure. Trees were generated using the ATGC online PhyML webserver. Host species or environmental site for which each sequence was generated is denoted as a colored circle: birds (dark blue), environmental sites (blue), chickens (light blue), humans (orange), and humans from this study (red).

Figure 2. Sliding window analysis reveals that selection acts to limit amino-acid changing diversity in H5N1 viruses. Individual values of πN and πS values were calculated for the entire length of PB2, PB1, PA, HA, NP, and NS gene segments using a window size of 10 codons and a step size of 5 codon. In this representation, individual valuates of πN and πS were summarized using a “loess” trend line (black) with the 99% confidence interval of the trend line is depicted in grey. The x-axis denotes the nucleotide position of each gene segment. The y-axis corresponds to the ratio of a πN/πS. Generally, a πN/πS ratio > 1 indicates that positive selection is favoring genetic diversification. By contrast, a πN/πS ratio < 1 indicates that purifying (or negative) selection is acting to maintain a fit virus population by removing deleterious mutations.

Figure 3. Frequencies of single nucleotide polymorphisms detected in ≥ 5% of

viral sequences within-infected humans. Nonsynonymous single nucleotide polymorphisms detected in ≥ 5% of viral sequences are candidates for in vitro functional characterization. This analysis focused on nonsynonymous SNPs detected in at least 5% of virus sequences in one or more samples collected from any human at any time point or anatomical site. Bar graphs depict changing frequencies of specific nonsynonymous SNPs during infection in humans from throat swabs (TS) or tracheal aspirates (TA) samples. Amino acid abbreviations: A; Alanine, C; Cysteine, D; Aspartic acid, E; Glutamic acid, F; Phenylalanine, G; Glycine, H; Histidine, I; Isoleucine, K; Lysine, L; Leucine, M; Methionine, N; Asparagine, P; Proline, Q; Glutamine, R; Arginine, S; Serine, T; Threonine, V; Valine, Y; Tyrosine.

Figure 4. Characterization of HA polymorphisms receptor binding properties. (A) Localization of amino acid variations on receptor binding domain identified at 5-50% frequency in our deep sequencing analysis. The variations are mapped on a receptor binding (sub)domain (amino acid positions 117-265 in H3 numbering) (Ha et al. 2002) of the three-dimensional structure of the monomer of VN1203 HA (Protein Data Bank accession 2FKO). (B) Receptor binding specificities of HA RBD variants. Direct binding of virus to α2,3-linked (blue) or α2,6-linked (red) sialylglycopolymers was assessed. The mean receptor binding activity (mean of triplicates of a single experiment) is shown. Binding of virus to human trachea sections was assessed. Signal was observed as red color. K173 (H1N1) virus was used as a control.

Figure 5. Characterization of HA polymorphisms on thermostability. (A) Amino acid substitutions on non-receptor binding domains are mapped on the three-dimensional structure of the monomer of VN1203 HA (Protein Data Bank accession 2FKO). (B) Thermostability of HA non-receptor binding domains variants. Viruses containing 64 HAU were incubated at 55 °C for 0-240 min. HA titers in heat-treated samples were

determined by performing HA assays with 0.5% TRBCs and virus titers in heat-treated samples were determined by plaque assays on MDCK cells. The mean HA titers or the mean virus titers (mean of triplicates of a single experiment) are shown. 20 PFU (dashed line) was the detection limit. All amino acid variations are designated in H3 numbering.

Figure 6. Characterization of vRNP polymorphisms on polymerase activity. 293T cells were transfected with plasmids encoding the polymerase proteins; PB2, PB1, PA and NP, with a plasmid for the expression of an influenza virus minigenome which encodes the firefly luciferase gene, and with a control plasmid encoding Renilla luciferase. The cells were incubated at 33°C (A) or at 37°C (B) for 24 h. Firefly and Renilla luciferase activities were measured by use of Dual-Luciferase Reporter Assay System. The firefly luciferase values were divided by Renilla luciferase values to normalize. The experiments (each in triplicates) were independently repeated twice. The mean relative viral polymerase activities with the standard deviations are shown. The viral polymerase activity of K173 was set to 100%.

Tables, Figures and Supplementals

Table 1. Sample history

Specimen IDs	HA clades	Sampling sites	Date of sympton onset	Date of hospitalization	Date of collection	Outcome
UT3040I	1	Throat swab	N.A.	N.A.	6-Jan-04	Died
UT3040II		Trachea aspirate			7-Jan-04	
UT31312I		Trachea aspirate			25-Jul-07	
UT31312II	2.3.4	Trachea aspirate	N.A.	N.A.	26-Jul-07	Died
UT31312III		Throat swab			25-Jul-07	
UT31394I	2.3.4	Throat swab	N.A.	N.A.	17-Jan-08	Died
UT31394II		Trachea aspirate			17-Jan-08	
UT31413I	2.3.4	Throat swab	3-Feb-08	N.A.	13-Feb-08	Died
UT31413II		Trachea aspirate			13-Feb-08	
UT36250I	2.3.4.2	Throat swab	5-Mar-10	10-Mar-10	10-Mar-10	Survived
UT36250II		Trachea aspirate			11-Mar-10	
UT36282I	2.3.4.1	Throat swab	27-Mar-10	2-Apr-10	1-Apr-10	Survived
UT36282II		Throat swab			3-Apr-10	
UT36285I	2.3.4.1	Throat swab	2-Apr-10	4-Apr-10	4-Apr-10	Survived
UT36285II		Throat swab			8-Apr-10	

Table 2. Within-host genetic variation for functional characterization

Protein (domain)	Specimens	Substitutions*
HA (RBD)	UT3040I	A138V
	UT3040II	N186K
	UT31312III	N186K
	UT31312III	S203P
	UT31394I/II	S207T
	UT36250I/II	A138V
	UT36250II	Q226K
	UT36282I/II	A138V
HA (non-RBD)	UT31312II	N92K
	UT31312III	M456R
	UT31394I	E54(+1)D
	UT31394II	D54(+1)E
	UT31394II	A67T
	UT31413I	T511I
	UT31413II	Y486H
PB2	UT3040I	K627E
	UT31312I/II	E627K
	UT31312III	K627E
	UT31394I	E627K
	UT31394II	K627E
	UT36250I	R15C
	UT36250I	E627K
PB1	UT31394I/II	F254L
	UT31394II	D538G
	UT36250I	T123A
	UT36250I	L598P
	UT36282II	N213D
PA	UT3040I	V90M
	UT3040I	K142E
	UT3040I	V387I
	UT3040I	L417V
	UT31394I/II	A651S
	UT31413I	P325S
	UT31413I	F260S

	UT31394I/II	R77K
NP	UT31394I/II	E454G
	UT36282II	A430V
NS	UT36250I/II	I124M

* The numbers refer to the amino acid positions in mature H3 HA.

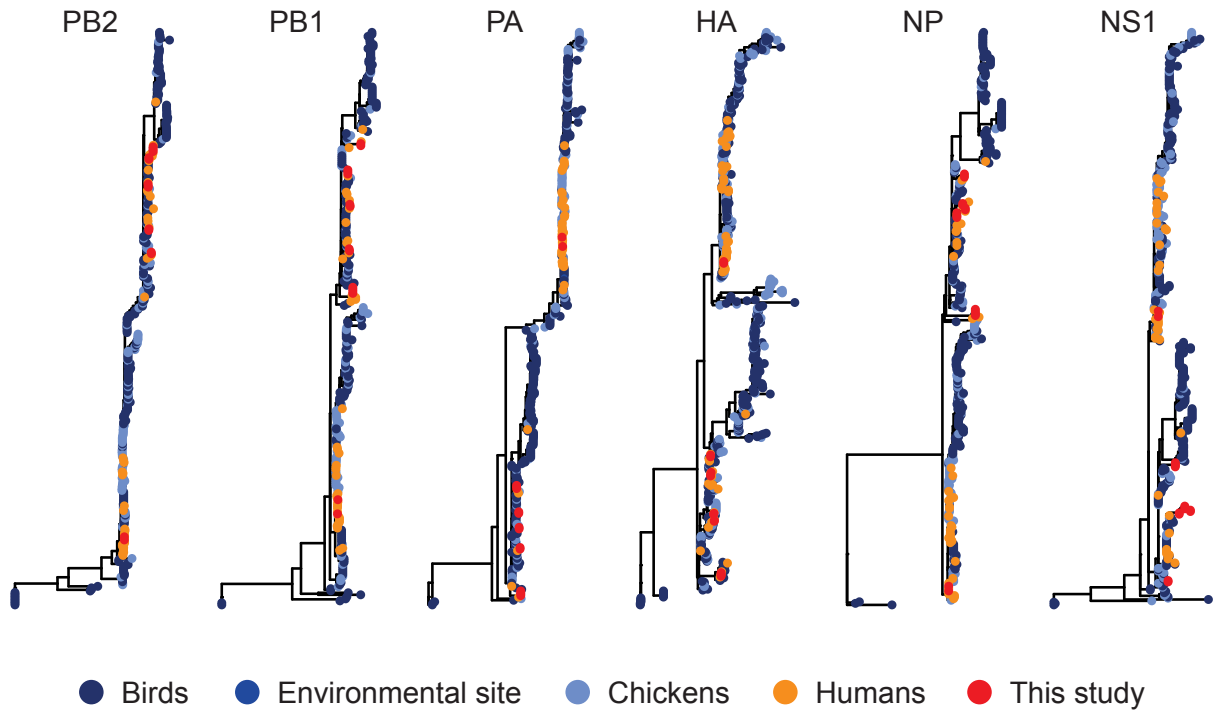


Figure 1. Phylogenetic relationships of H5N1 influenza A genes sequenced from birds, humans and environmental sites in Vietnam. Maximum-likelihood phylogenies were created for H5N1 genes using consensus sequences generated from human clinical specimens. In some instances, clinical samples yielded insufficient sequence coverage for consensus generation (see Methods), for these genes we used consensus sequences generated from virus isolates passaged in Madin-Daryby canine kidney (MDCK) cells. Trees were built using partial length gene sequences with reference sequences from the Influenza Research Database project from Vietnam. The Generalized time reversible substitution model was determined using Datamonkey webserver with the AIC-based model selection procedure. Trees were generated using the ATGC online PhyML webserver. Host species or environmental site for which each sequence was generated is denoted as a colored circle: birds (dark blue), environmental sites (blue), chickens (light blue), humans (orange), and humans from this study (red).

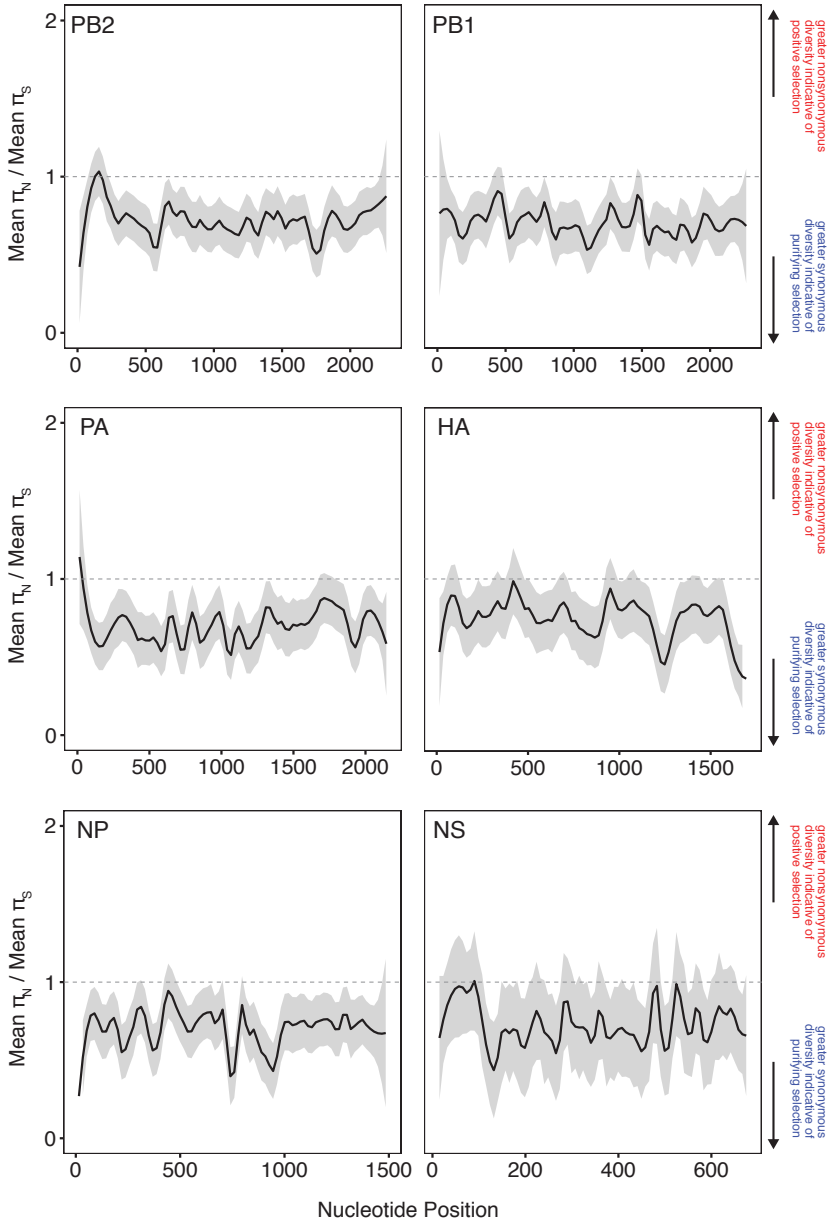


Figure 2. Sliding window analysis reveals that selection acts to limit amino-acid changing diversity in H5N1 viruses. Individual values of πN and πS values were calculated for the entire length of PB2, PB1, PA, HA, NP, and NS gene segments using a window size of 10 codons and a step size of 5 codon. In this representation, individual valuates of πN and πS were summarized using a "loess" trend line (black) with the 99% confidence interval of the trend line is depicted in grey. The x-axis denotes the nucleotide position of each gene segment. The y-axis corresponds to the ratio of a πN/πS. Generally, a πN/πS ratio > 1 indicates that positive selection is favoring genetic diversification. By contrast, a πN/πS ratio < 1 indicates that purifying (or negative) selection is acting to maintain a fit virus population by removing deleterious mutations.

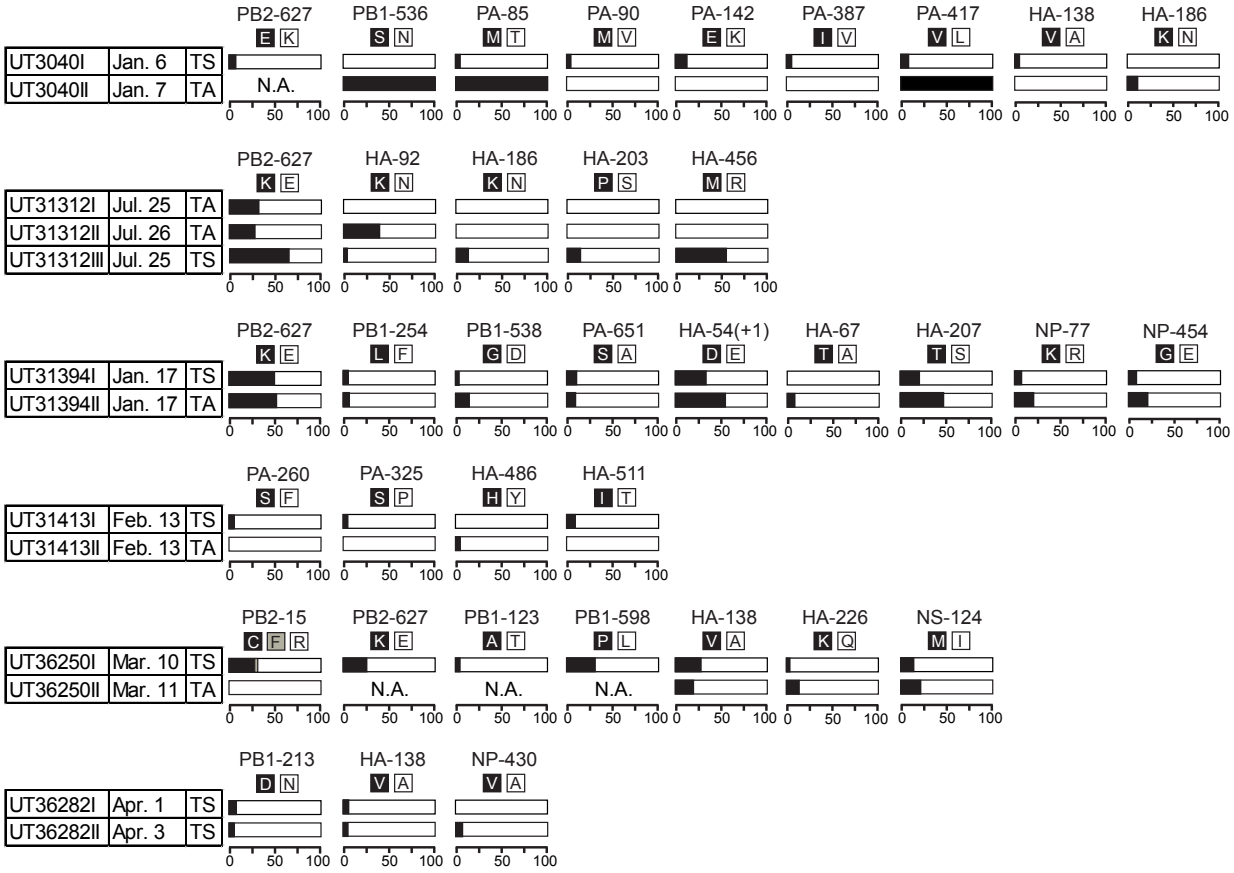


Figure 3. Frequencies of single nucleotide polymorphisms detected in ≥ 5% of viral sequences within-infected humans. Nonsynonymous single nucleotide polymorphisms detected in ≥ 5% of viral sequences are candidates for in vitro functional characterization. This analysis focused on nonsynonymous SNPs detected in at least 5% of virus sequences in one or more samples collected from any human at any time point or anatomical site. Bar graphs depict changing frequencies of specific nonsynonymous SNPs during infection in humans from throat swabs (TS) or tracheal aspirates (TA) samples. Amino acid abbreviations: A; Alanine, C; Cysteine, D; Aspartic acid, E; Glutamic acid, F; Phenylalanine, G; Glycine, H; Histidine, I; Isoleucine, K; Lysine, L; Leucine, M; Methionine, N; Asparagine, P; Proline, Q; Glutamine, R; Arginine, S; Serine, T; Threonine, V; Valine, Y; Tyrosine.

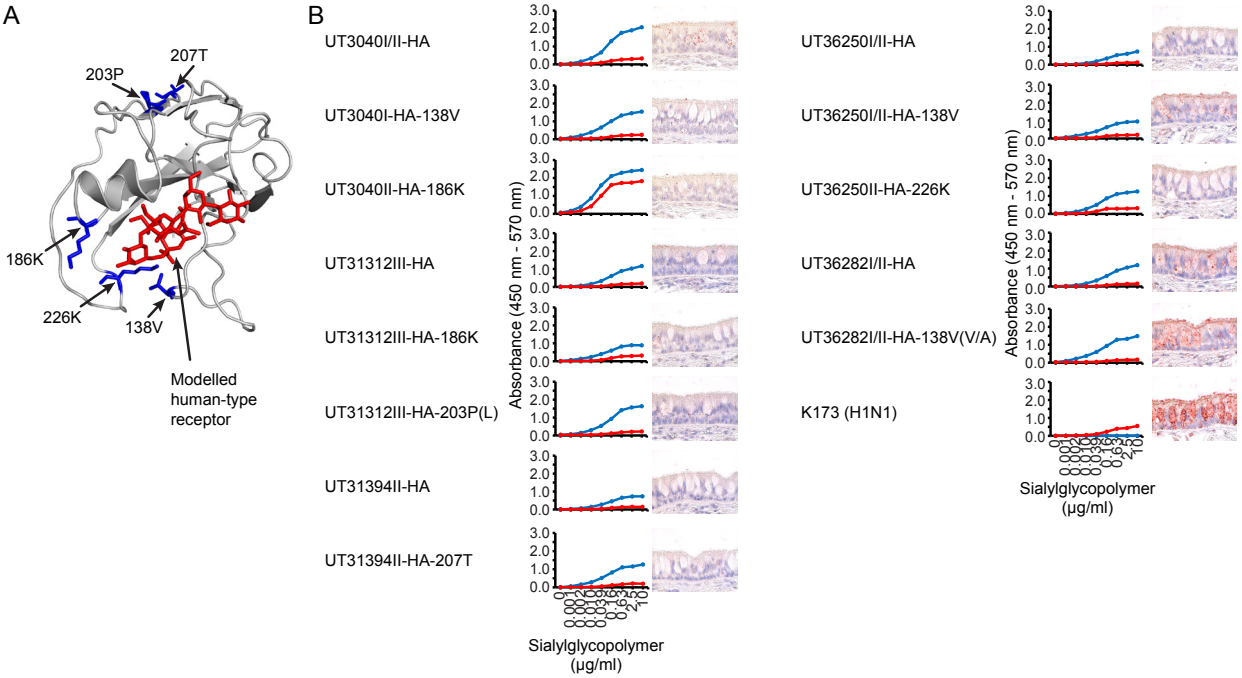


Figure 4. Characterization of HA polymorphisms receptor binding properties. (A) Localization of amino acid variations on receptor binding domain identified at 5-50% frequency in our deep sequencing analysis. The variations are mapped on a receptor binding (sub)domain (amino acid positions 117-265 in H3 numbering) (Ha et al. 2002) of the three-dimensional structure of the monomer of VN1203 HA (Protein Data Bank accession 2FKO). (B) Receptor binding specificities of HA RBD variants. Direct binding of virus to α2,3-linked (blue) or α2,6-linked (red) sialylglycopolymers was assessed. The mean receptor binding activity (mean of triplicates of a single experiment) is shown. Binding of virus to human trachea sections was assessed. Signal was observed as red color. K173 (H1N1) virus was used as a control.

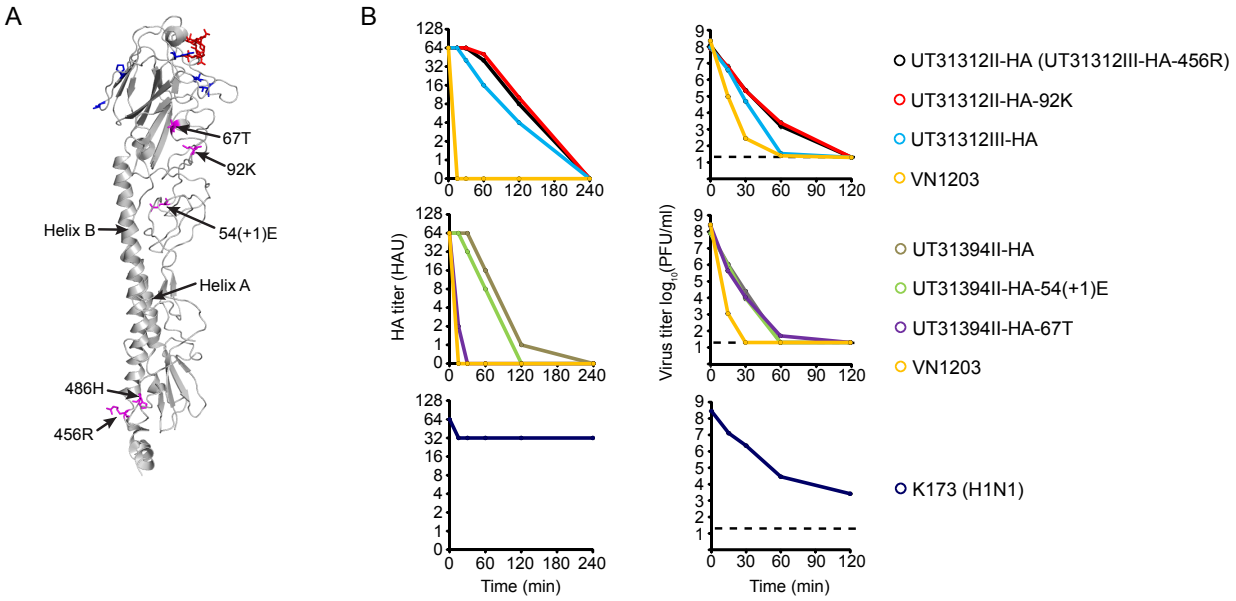


Figure 5. Characterization of HA polymorphisms on thermostability. (A) Amino acid substitutions on non-receptor binding domains are mapped on the three-dimensional structure of the monomer of VN1203 HA (Protein Data Bank accession 2FKO). (B) Thermostability of HA non-receptor binding domains variants. Viruses containing 64 HAU were incubated at 55 °C for 0-240 min. HA titers in heat-treated samples were determined by performing HA assays with 0.5% TRBCs and virus titers in heat-treated samples were determined by plaque assays on MDCK cells. The mean HA titers or the mean virus titers (mean of triplicates of a single experiment) are shown. 20 PFU (dashed line) was the detection limit. All amino acid variations are designated in H3 numbering.

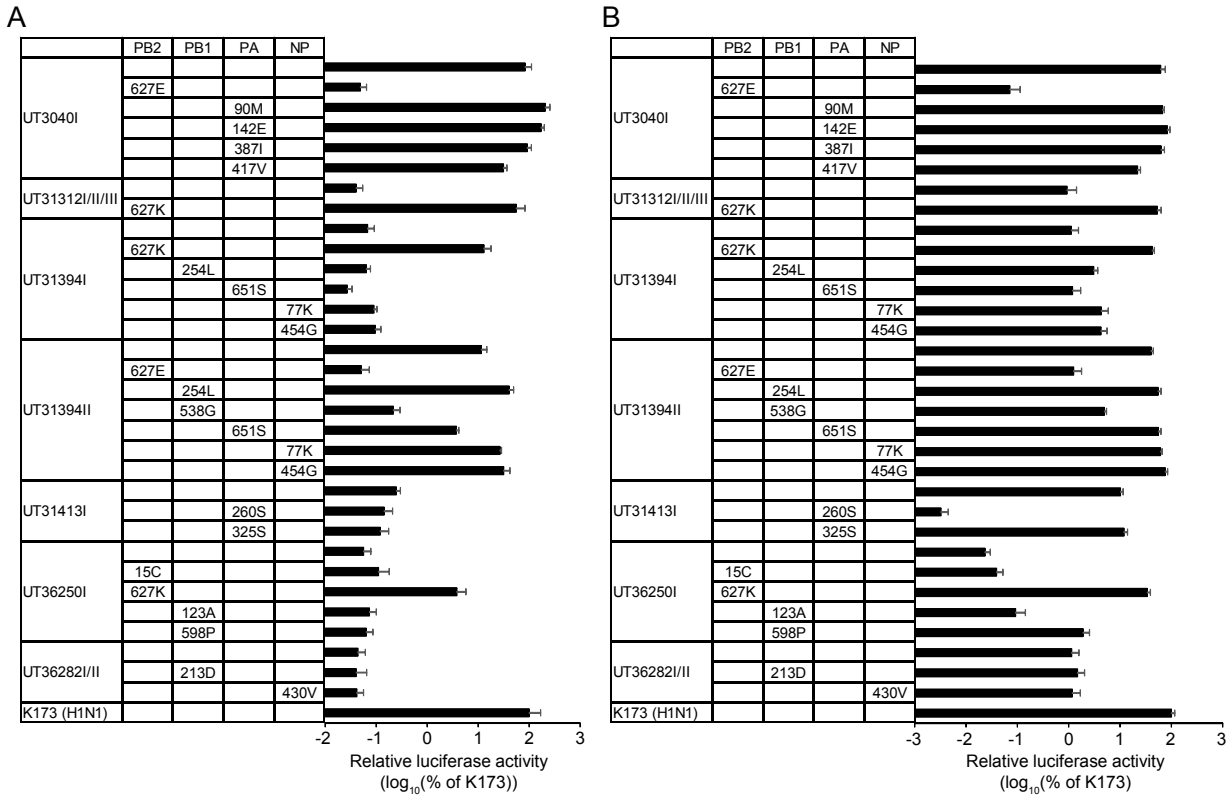


Figure 6. Characterization of vRNP polymorphisms on polymerase activity. 293T cells were transfected with plasmids encoding the polymerase proteins; PB2, PB1, PA and NP, with a plasmid for the expression of an influenza virus minigenome which encodes the firefly luciferase gene, and with a control plasmid encoding Renilla luciferase. The cells were incubated at 33°C (A) or at 37°C (B) for 24 h. Firefly and Renilla luciferase activities were measured by use of Dual-Luciferase Reporter Assay System. The firefly luciferase values were divided by Renilla luciferase values to normalize. The experiments (each in triplicates) were independently repeated twice. The mean relative viral polymerase activities with the standard deviations are shown. The viral polymerase activity of K173 was set to 100%.

Table S1: Primers used in this study

Gene segment	Nucleotide positions	Forward primer sequence	Reverse primer sequence
PB2	1-1784	AGCGAAAGCAGGTCAAATATATTC	TTGGGTACCAAGGATTGGAACGG
	532-2341	CCAAATGAAGTGG-GAGCTAGAATATTG	AGTAGAAACAAGGTCGTTTTTAAA
	167-2277	TGAAATGGATGATGGCAATG	CGCTGTCTGGCTGTCAGTAA
PB1	1-1829	AGCGAAAGCAGGCAAACC	CGGATATTGTATAGATTTGGTCCTCC
	501-2341	ATCGGGACGGCTAATAGATTTC	AGTAGAAACAAGGCATTTTTTCACG
	106-2214	CCTCCATACAGCCATGGAAC	GAAATCAATTCGTGCGTCAA
PA	1-1668	AGCGAAAGCAGGTACTGATCC	CATGTCTCCTATCTCGAGGACAC
	565-2233	GAAATGGCCAGTAGGGGTCTATG	AGTAGAAACAAGGTACTTTTTTGGAC
	74-2169	CGGAAAAGGCAATGAAAGAA	CAGTGCATGTGTGAGGAAGG
HA	1-1776*	AGCAAAAGCAGGGGT{TorC} CAATC	AGTAGAAACAAGGGTGTTTT{TorC} AACTA
	96-793	ACCATGCAAACAACTCGCAC	GATTGCATCGTCCGGTTTTA
NP	1-1565	AGCAAAAGCAGGGTAGATAATCAC	AGTAGAAACAAGGGTATTTTTCTTTA-ATTG
	51-1487	GTCTCAAGGCACCAAACGAT	ATGTCAAAGGAAGGCACGAT
NS	1-875	AGCAAAAGCAGGGTGACAAAAAC	AGTAGAAACAAGGGTGTTTTTTAT
	83-772	CAAACGATTTGCAGACCAAG	TCTGTTCGAAGCTGTTTTCTG

* Corresponding to 1-1779 of UT3040I and UT36250I viruses due to the three additional nucleotides at the multibasic cleavage site.

Table S2: Comparision of H5N1 consensus sequence with NCBI BLAST top hit from Vietnam.

Virus	Blast top hit	Nucleotide identity	Gene	Amino acid changes
UT3040I	Environment/2004 (GU052425)	100.0	PB2	n/a
UT3040II	Environment/2004 (GU052425)	100.0	PB2	n/a
UT31312I	Duck/2007 (JX241044)	99.8	PB2	A108T, V203I, E627K
UT31312II	Duck/2007 (JX241044)	99.8	PB2	A108T, V203I, E627K
UT31312III	Duck/2007 (JX241044)	99.6	PB2	A108T, V203I
UT31394I	Duck/2007 (GU050435)	99.0	PB2	A106T, R299K, M315I, K736R
UT31394II	Duck/2007 (GU050435)	99.0	PB2	A106T, R299K, M315I, E627K, K736R
UT31413I	Muscovy duck/2007 (GU050514)	99.9	PB2	T303I, P515S
UT31413II	Muscovy duck/2007 (GU050514)	99.9	PB2	T303I, P515S
UT36250I	Muscovy Duck/2007 (CY030725)	98.5	PB2	I451V, R630K
UT36250II	Muscovy Duck/2007 (CY030725)	98.4	PB2	I451V, E627K, R630K
UT36282I	Muscovy duck/2005 (EU931011)	97.9	PB2	E249G, T330M, R355K, I461V, M473I, V478I, V560L, Y658H
UT36282II	Muscovy duck/2005 (EU931011)	97.9	PB2	E249G, T330M, R355K, I461V, M473I, V478I, V560L, Y658H
UT36285I	Muscovy duck/2005 (EU931011)	97.9	PB2	E249G, T330M, R355K, I461V, M473I, V478I, V560L, Y658H
UT36285II	Muscovy duck/2005 (EU931011)	97.9	PB2	E249G, T330M, R355K, I461V, M473I, V478I, V560L, Y658H
UT3040I	Environment/2004 (GU052424)	100.0	PB1	n/a
UT3040II	Environment/2004 (GU052424)	100.0	PB1	n/a
UT31312I	Muscovy duck/2007 (GU050537)	99.7	PB1	K577R, D719E
UT31312II	Muscovy duck/2007 (GU050537)	99.7	PB1	K577R, D719E

UT31312III	Muscovy duck/2007 (GU050537)	99.7	PB1	K577R, D719E
UT31394I	Muscovy duck/2007 (GU050537)	99.0	PB1	M111I, T156V, S261N, M566T, D719E
UT31394II	Muscovy duck/2007 (GU050537)	99.0	PB1	M111I, T156V, S261N, M566T, D719Q
UT31413I	Duck/2007 (GU050521)	99.5	PB1	V200I, S361G, H634N, D719Q
UT31413II	Duck/2007 (GU050521)	99.6	PB1	V200I, S361G, H634N, D719E
UT36250I	Muscovy duck/2005 (EU931010)	97.9	PB1	M40L, E104D, M171I, S345A, I392V, E581D, K745R
UT36250II	Muscovy duck/2005 (EU931010)	97.9	PB1	M40L, E104D, M171I, S345A, I392V, E581D, K745R
UT36282I	Muscovy duck/2005 (EU931010)	98.1	PB1	V191L, M195I, K198R, N314S, I368V, D398E, R680K, M744T
UT36282II	Muscovy duck/2005 (EU931010)	98.1	PB1	V191L, M195I, K198R, N314S, I368V, D398E, R680K, M744T
UT36285I	Muscovy duck/2005 (EU931010)	98.1	PB1	V191L, M195I, K198R, N314S, I368V, D398E, R680K, M744T
UT36285II	Muscovy duck/2005 (EU931010)	98.1	PB1	V191L, M195I, K198R, N314S, I368V, D398E, R680K, M744T
UT3040I	Environment/2004 (GU052423)	100.0	PA	n/a
UT3040II	Environment/2004 (GU052423)	99.9	PA	E142K
UT31312I	Muscovy duck/2007 (CY029678)	99.7	PA	F35L, S409F
UT31312II	Muscovy duck/2007 (CY029678)	99.7	PA	F35L
UT31312III	Muscovy duck/2007 (CY029678)	99.9	PA	F35L
UT31394I	Muscovy duck/2007 (CY029678)	99.2	PA	S405I, A618T, A689S
UT31394II	Muscovy duck/2007 (CY029678)	99.1	PA	S405I, A618T, A689S
UT31413I	Duck/2007 (GU050520)	99.6	PA	M86V, A618T
UT31413II	Duck/2007 (GU050520)	99.6	PA	M86V, A618T
UT36250I	Muscovy duck/2005 (EU931009)	98.7	PA	G99E, I129V, S277F, N321K, L482M, K716E

UT36250II	Muscovy duck/2005 (EU931009)	98.7	PA	G99E, I129V, S277F, N321K, L482M, K716E
UT36282I	Muscovy duck/2005 (EU931009)	98.2	PA	N27S, V44I, T85A, V127A, A231T, C241Y, Y305F, A343T, K391R, R401K, I573V, T614N
UT36282II	Muscovy duck/2005 (EU931009)	98.2	PA	N27S, V44I, T85A, V127A, A231T, C241Y, Y305F, A343T, K391R, R401K, I573V, T614N
UT36285I	Muscovy duck/2005 (EU931009)	98.2	PA	N27S, V44I, T85A, V127A, A231T, C241Y, Y305F, A343T, K391R, R401K, I573V, T614N
UT36285II	Muscovy duck/2005 (EU931009)	98.2	PA	N27S, V44I, T85A, V127A, A231T, C241Y, Y305F, A343T, K391R, R401K, I573V, T614N
UT3040I	Environment/2004 (GU052418)	100.0	HA	n/a
UT3040II	Environment/2004 (GU052418)	100.0	HA	n/a
UT31312I	Duck/2007 (GU052533)	99.7	HA	F369Y
UT31312II	Duck/2007 (GU052533)	99.7	HA	F369Y
UT31312III	Duck/2007 (GU052533)	99.6	HA	F369Y, R472M
UT31394I	Quail/2008 (CY099963)	99.4	HA	M10I, D61E, I204T, I528T
UT31394II	Quail/2008 (CY099963)	99.5	HA	M10I, I204T, I528T
UT31413I	Muscovy duck/2008 (CY099958)	99.7	HA	A226V, G340E
UT31413II	Muscovy duck/2008 (CY099958)	99.7	HA	A226V, G340E
UT36250I	Chicken/2008 (CY099966)	98.4	HA	M10V, A149S, I214V, Del342R, T529I
UT36250II	Chicken/2008 (CY099966)	98.4	HA	M10V, A149S, I214V, Del342R, T529I
UT36282I	Duck/2010 (JN935004)	99.3	HA	F6L, T56K, R282K, G490D
UT36282II	Duck/2010 (JN935004)	99.3	HA	F6L, T56K, R282K, G490D
UT36285I	Duck/2010 (JN935004)	99.3	HA	F6L, T56K, R282K, G490D
UT36285II	Duck/2010 (JN935004)	99.3	HA	F6L, T56K, R282K, G490D
UT3040I	Environment/2004 (GU052421)	99.9	NP	n/a

UT3040II	Quail/2004 (DQ099763)	99.9	NP	n/a
UT31312I	Muscovy duck/2007 (CY029720)	99.9	NP	n/a
UT31312II	Muscovy duck/2007 (CY029720)	99.9	NP	n/a
UT31312III	Muscovy duck/2007 (CY029720)	99.9	NP	n/a
UT31394I	Muscovy duck/2007 (CY029720)	99.2	NP	T232I
UT31394II	Muscovy duck/2007 (CY029720)	99.2	NP	T232I
UT31413I	Muscovy duck/2007 (CY029720)	99.4	NP	V67A, N417S
UT31413II	Muscovy duck/2007 (CY029720)	99.4	NP	V67A, N417S
UT36250I	Duck/2005 (DQ366307)	97.6	NP	N23T, P390T, G401A, V428A
UT36250II	Muscovy duck/2007 (CY030545)	99.0	NP	N23T, P390T, G401A, V428A
UT36282I	Chicken/2004 (DQ099772)	97.1	NP	L47I, P318T, A373T, R391K, T430A, R446K, L466F
UT36282II	Duck/2005 (EU9309911)	97.3	NP	S22A, L47I, P318T, A373T, R391K, T430A, R446K, L466F
UT36285I	Duck/2005 (EU9309911)	97.4	NP	S22A, L47I, P318T, A373T, R391K, T430A, R446K, L466F
UT36285II	Chicken/2004 (DQ099772)	97.1	NP	S22A, L47I, P318T, A373T, R391K, T430A, R446K, L466F
UT3040I	Chicken/2004 (AY770611)	99.3	NS	E71G, T132I, G148E, S200N, K216Stop
UT3040II	Chicken/2004 (AY770611)	99.3	NS	E71G, T132I, G148E, S200N, K216Stop
UT31312I	Muscovy duck/2007 (CY029723)	99.3	NS	G108D, T122A
UT31312II	Muscovy duck/2007 (CY029723)	99.3	NS	G108D, T122A
UT31312III	Muscovy duck/2007 (CY029723)	99.3	NS	G108D, T122A
UT31394I	Muscovy duck/2007 (CY029723)	98.6	NS	A60T, K78R, I151V, I193V, D202N
UT31394II	Muscovy duck/2007 (CY029723)	98.6	NS	A60T, K78R, I151V, I193V, D202N

UT31413I	Muscovy duck/2007 (CY029723)	99.4	NS	E147V
UT31413II	Muscovy duck/2007 (CY029723)	99.4	NS	E147V
UT36250I	Muscovy duck/2005 (EU930984)	98.9	NS	P208L, K216Stop
UT36250II	Muscovy duck/2005 (EU930984)	98.9	NS	P208L, K216Stop
UT36282I	Muscovy duck/2005 (EU930984)	97.9	NS	L27M, P80S, T81A, F133C, T192A, D204G
UT36282II	Muscovy duck/2005 (EU930984)	97.9	NS	L27M, P80S, T81A, F133C, T192A, D204G
UT36285I	Muscovy duck/2005 (EU930984)	97.9	NS	L27M, P80S, Y81A, F133C, T192A, D204G
UT36285II	Muscovy duck/2005 (EU930984)	97.9	NS	L27M, P80S, Y81A, F133C, T192A, D204G

Table S3: Identification of amino acid residues associated to mammalian adaptation

	PB2				PB1		HA						NA	
	339	368	391	627	368	101	133(+1)	159	192	239	222	223	224	225
UT3040I	K	R	Q	K	I	D	L	S	T	P	E	S	E	V
UT3040IIS	K	R	Q	K	I	D	L	S	T	P	E	S	E	V
UT31312I	T	Q	E	E	I	N	S	N	T	P	G	S	E	V
UT31312II	T	Q	E	E	I	N	S	N	T	P	G	S	E	V
UT31312III	T	Q	E	K	I	N	S	N	T	P	G	S	E	V
UT31394I	T	Q	E	E	I	N	S	N	T	P	G	S	E	V
UT31394II	T	Q	E	K	I	N	S	N	T	P	G	S	E	V
UT31413I	T	Q	E	E	I	N	S	N	T	P	G	S	E	V
UT31413II	T	Q	E	E	I	N	S	N	T	P	G	S	E	V
UT36250I	T	Q	E	E	I	N	L	N	I	P	E	S	E	V
UT36250IIS	T	Q	E	K	I	N	L	N	I	P	E	S	E	V
UT36282I	M	Q	E	E	V	N	S	N	M	S	E	S	E	V
UT36282II	M	Q	E	E	V	N	S	N	M	S	E	S	E	V
UT36285IS	M	Q	E	E	V	N	S	N	M	S	E	S	E	V
UT36285II	M	Q	E	E	V	N	S	N	M	S	E	S	E	V

Consensus sequences generated from MDCK passaged virus stocks are labeled with an “S” following the sample ID in the left most column. We used the H5N1 Genetic Changes Inventory: A Tool for Influenza Surveillance and Preparedness (http://www.cdc.gov/flu/pdf/avianflu/h5n1-inventory.pdf) to identify mutations that have been reported to affect H5N1 biological properties. Substitutions demonstrated to alter viral property to a mammalian-like phenotype are colored red. Substitutions demonstrated to alter viral property to an avian-like phenotype are colored blue. The A/Vietnam/UT3040I/2004 sequence is used as reference. A dashed line in the sequence alignment denotes a position with missing amino acid information. Asterisks indicate a stop codon.

Table S4: Summary of nonsynonymous and synonymous mutations detected in our study.

	PB2		PB1		PA		HA		NP		NS	
	N	S	N	S	N	S	N	S	N	S	N	S
UT3040I	8	2	6	10	10	0	1	4	1	3	1	0
UT3040II			0	0	0	1	1	0			0	0
UT31312I	3	2	0	1	0	0	1	3	1	0	0	0
UT31312II	4	1	0	0	0	0	5	0	1	0	1	0
UT31312III	4	0	0	0	0	0	6	1	0	1	0	0
UT31394I	12	3	6	10	5	3	11	8	5	0	2	1
UT31394II	6	1	3	2	4	0	4	3	3	0	1	0
UT31413I	0	0	9	2	5	5	1	0	6	3	4	0
UT31413II	0	1	0	3	0	1	5	8	0	0	2	0
UT36250I	12	8	9	7	11	4	14	1	7	0	3	0
UT36250II	0	0					6	2			1	0
UT36282I	3	3	2	5	1	2	3	2	2	1	0	0
UT36282II	2	2	1	3	2	2	5	7	3	3	8	4
UT36285I	0	1			0	1	0	0	0	0	0	0
UT36285II	3	2			3	2			1	0	0	0

abbreviations: N; nonsynoymous, S; synonymous, n/a; not applicable due to low sequence coverage

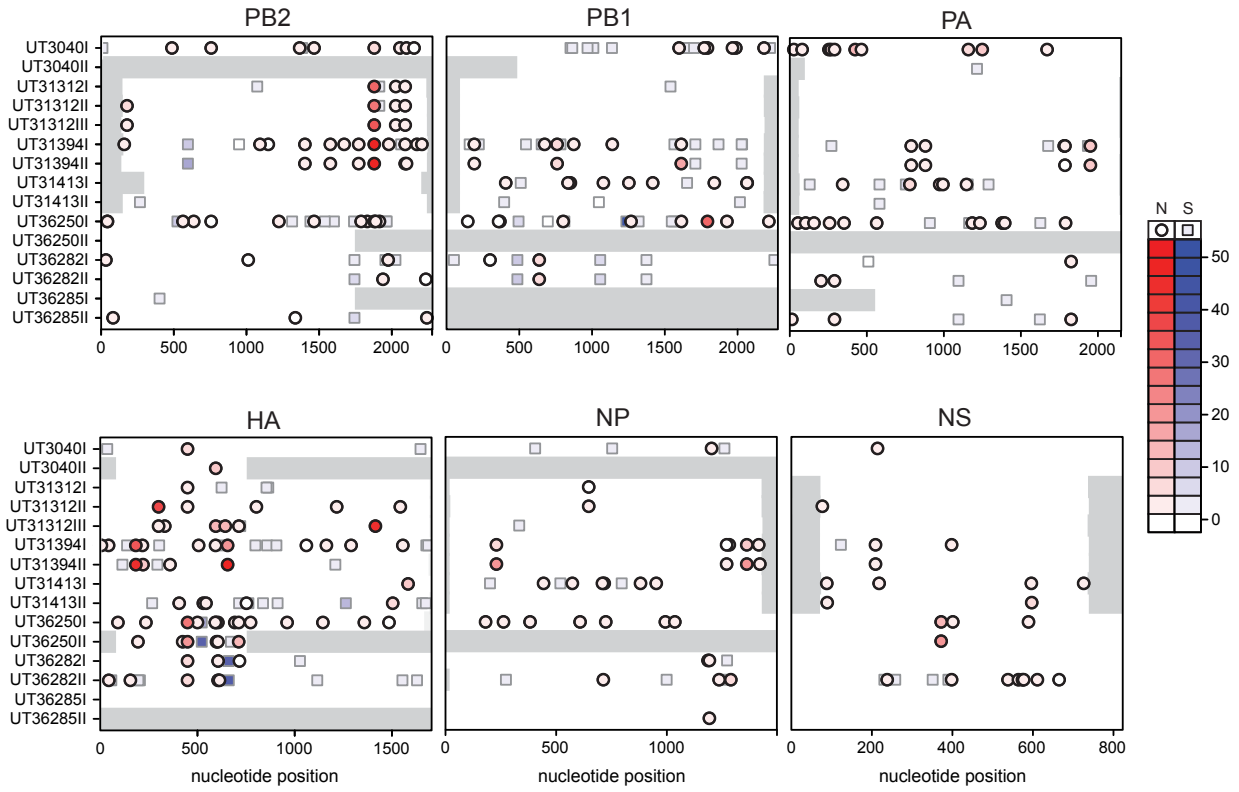


Figure S1. Deep sequencing reveals H5N1 influenza sequence variation within H5N1 infected humans. Using deep sequencing we characterized H5N1 within-host nucleotide diversity from viral nucleic acids extracted directly from human respiratory specimens. Single-nucleotide polymorphisms (SNPs) were called for each patient along the consensus gene sequence if detected in ≥ 1% of sequence reads (see Methods). Note, if an allele was detected in > 50% of sequences we consider it the "consensus" nucleotide. Every SNP was called relative to its consensus nucleotide; therefore, the maximum frequency of each SNP is > 50%. Grey bars indicate gene regions with insufficient sequence coverage (> 100 sequence reads) to call variants. Synonymous (squares) and nonsynonymous (circles). SNPs were mapped for each clinical specimen (y-axis) along their respective consensus gene sequence (x-axis). SNP frequencies are represented as a heatmap with synonymous and nonsynonymous mutations represented in blue and red, respectively.

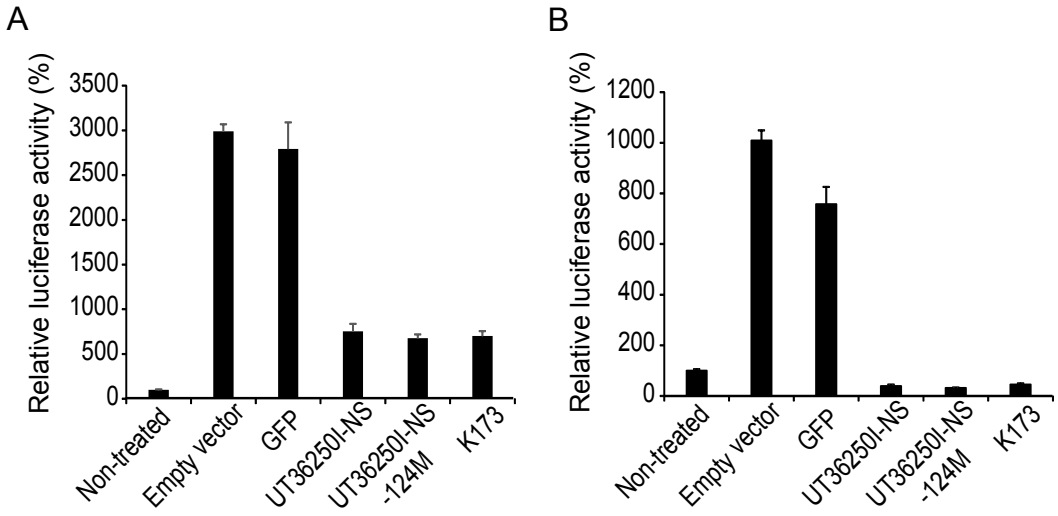


Figure S2. Characterization of NS1 polymorphisms on IFN antagonistic properties. 293T cells were transfected with a plasmid encoding the NS1 protein and with a plasmid encoding the firefly luciferase gene under the control of IFNβ promoter (A) or interferon-stimulated response element (ISRE) (B). The cells were treated with Sendai virus (SeV) or human IFNβ, respectively. Firefly luciferase activities was measured by use of Luciferase Assay Reagent II. The experiments (each in triplicates) were independently repeated twice. The mean relative IFN antagonistic activities with the standard deviations are shown. The activity of cells transfected with pCAGGS empty vector (non-treated with SeV) was set to 100%. Plasmid encoding GFP was used as a negative control. Plasmids encoding K173 (H1N1) virus NS1 was used for comparison.

Chapter 4

Deep sequencing reveals potential antigenic variants at low frequency in influenza A-infected humans

Published, in part as

Jorge M. Dinis*, Nicholas W. Florek*, Omolayo O. Fatola, Louise H. Moncla, James P. Mutschler, Olivia K. Charlier, Jennifer K. Meece, Edward A. Belongia, Thomas C. Friedrich

*Jorge M. Dinis and Nicholas W. Florek contributed equally

This chapter represents a manuscript in preparation for publication.

Abstract

Influenza vaccines must be frequently reformulated to account for antigenic changes in the viral envelope protein, hemagglutinin (HA). The rapid evolution of influenza virus under immune pressure is likely enhanced by the virus's genetic diversity within a host, though antigenic change has rarely been investigated on the level of individual infected humans. We used deep sequencing to characterize between- and within-host genetic diversity of influenza viruses in a cohort of patients that included individuals who were vaccinated and then infected in the same season. We characterized influenza HA segments from the predominant circulating influenza A subtypes during the 2012-2013 (H3N2) and 2013-2014 (pandemic H1N1; H1N1pdm) flu seasons. HA consensus sequences and the level of HA diversity were similar in non-vaccinated and vaccinated subjects. In both groups purifying selection was the dominant force shaping HA genetic diversity. Interestingly, viruses from multiple individuals harbored low-frequency mutations encoding amino acid substitutions in HA antigenic sites at or near the receptor-binding domain. These mutations included two substitutions in H1N1pdm viruses, G158K and N159K, which were recently found to confer escape from virus-specific antibodies. These findings raise the possibility that influenza quasispecies encoding antigenic diversity can be generated within individual human hosts, but may not become fixed in the viral population even when they would be expected to have a strong fitness advantage. Understanding constraints on influenza antigenic evolution within individual hosts may elucidate potential future pathways of antigenic evolution at the population level.

Introduction

Seasonal influenza A epidemics cause an estimated 3 to 5 million cases of severe respiratory illness each year, resulting in approximately 250,000 to 500,000 deaths worldwide. Available influenza vaccines only provide moderate protection against infection and illness. Vaccine effectiveness can vary greatly by season, but average effectiveness at preventing illness has been estimated at 59% [18]. Even when there is good antigenic match between circulating viruses and vaccine strains, people can become infected despite being vaccinated earlier in the same season, a situation known as vaccine failure [119,120]. In the most recent Northern Hemisphere influenza season (2014-2015), interim overall vaccine effectiveness (VE) was estimated at 23% in the United States, lower than in past flu seasons [121]. The low VE estimate coincided with a poor antigenic match between circulating H3N2 drifted variants and the H3N2 viruses included in 2014-15 vaccines [122].

Viral antigenicity is influenced by multiple factors, including the accumulation of point mutations in the gene encoding the viral attachment protein hemagglutinin (HA), the specificity and avidity of HA for its receptor, and epistatic interactions within HA and between gene segments [123,124]. The HA protein has five defined antigenic sites that are important for the recognition of neutralizing antibodies; amino acid substitutions in these sites can significantly change viral antigenicity [125,126]. Newly emerging antigenic variants of epidemiological importance have historically had four or more amino acid substitutions located in two or more antigenic sites [127]. “Antigenic cartography” using panels of reference sera to assess the degree to which antibodies of a given specificity cross-react with panels of reference viruses demonstrated that H3N2 strains circulating from 1968 to 2003 formed 11 distinct “antigenic clusters” [23]. Antisera against older antigenic clusters are poorly cross-reactive to new clusters. As a result, emerging antigenically distinct viruses have a selective advantage over older viruses and replace them in circulation [24]. More recently, Koel et al. found that single

amino acid substitutions in one of seven residues near the receptor-binding domain of HA were responsible for the majority of previously observed H3N2 antigenic cluster transitions [25]. Together these studies demonstrate that point mutations in crucial sites in HA can result in significant antigenic change.

In a host, influenza viruses exist as a diverse collection of genetically linked variants, sometimes termed "quasispecies," that arise due to low-fidelity genome replication, rapid replication kinetics and high viral loads [26,27]. The composition of the within-host viral population is shaped by the generation of diversity through random mutation, and the action of natural selection, which can either promote genetic diversification, or purify deleterious mutations from the virus population. Quasispecies diversity is therefore understood to enable influenza viruses to rapidly adapt to changing environments [116,128]. However, the degree to which quasispecies genetic diversity encodes antigenic diversity and contributes to the emergence of antigenic variants within individual hosts is not clear. Traditional influenza surveillance methods based on Sanger sequencing cannot resolve variants present below 20% of the viral population, making it difficult to characterize within-host viral genetic diversity [31]. Deep sequencing influenza viruses from immune-compromised humans has revealed the existence of low-frequency drug-resistant variants that rapidly dominate the viral population after oseltamivir treatment [36]. Moreover, we recently showed that HA variants present in a quasispecies below the detection limit of Sanger sequencing could be transmitted via respiratory droplets [17]. It is therefore conceivable that low-level antigenic variants generated in individual infections have the potential to rapidly emerge in nature.

Here we used deep sequencing to characterize HA segment genetic diversity in subjects acutely infected with influenza, with and without recent vaccination, to determine how viral genetic diversity potentially encoding antigenic change is generated and selected during human infection. We sequenced viruses directly from clinical samples without isolation in eggs or passage in tissue culture to avoid potential loss of diversity

and/or selection for specific variants within the natural quasispecies. There were no conserved mutations that distinguished HA consensus sequences in non-vaccinated subjects from those in vaccinated individuals. Within-host HA segment genetic diversity varied considerably among humans, although most viral polymorphisms were typically present either at or near fixation (90 to 100% of viral sequences) or at low frequency (1 to 10%) relative to the relevant reference sequence. Patterns of HA segment diversity were consistent with viruses undergoing purifying selection to maintain fit viral populations. Interestingly, we detected nonsynonymous mutations that had previously been shown to confer escape from neutralizing antibodies in both H3N2 and H1N1pdm viruses infecting both non-vaccinated and vaccinated subjects. These mutations were always detected at low frequencies in the viral population. Together, our data suggest that putative influenza antigenic escape variants are generated in influenza-infected humans, but often do not become fixed in the virus population.

Material and Methods

Study population

The source population for this study included residents of the Marshfield Epidemiologic Study Area (MESA), a population-based cohort of approximately 54,000 people in a 14-zip-code area surrounding the Marshfield Clinic campus in Marshfield, Wisconsin [129]. In the 2012-13 and 2013-14 seasons, most community-dwelling MESA residents aged ≥ 6 months were eligible to be recruited during an outpatient encounter for acute respiratory illness with cough. Patients were screened by study personnel, and those meeting symptom criteria were eligible if the duration of illness was ≤ 7 days. Participants completed an interview to assess illness onset and symptoms. Individuals were considered vaccinated if they had received an influenza vaccine for the current influenza season ≥ 14 days prior to illness onset. We obtained nasal and oropharyngeal swabs from non-vaccinated and vaccinated human subjects aged ≥ 9 years; swabs were combined in viral transport media for testing. Influenza infection was confirmed using quantitative RT-PCR (qRT-PCR) [130]. During the 2012-13 and 2013-14 seasons, subtypes H3N2 and 2009 pandemic H1N1 (H1N1pdm) respectively dominated circulation in the United States [120,131]; for this reason, we only had access to H3N2 during the 2012-13 season and H1N1pdm viruses during the 2013-14 season.

The Marshfield Clinic Institutional Review Board approved study procedures. Informed consent was obtained from all adults and parents/guardians of children. Additional details on the vaccine effectiveness studies from which these human samples were collected are published [120,131].

Template Preparation

Total nucleic acids were extracted from clinical specimens using the RNeasy Mini Kit (Qiagen, U.S.A). Influenza vRNAs were reverse transcribed using the Uni12-F AGCAAAAGCAGG primer [132] with Superscript III reverse transcriptase (Invitrogen, U.S.A.) according to the manufacturer's instructions. Single-stranded cDNA was used

as a template for PCR amplification to amplify the HA genes using either primers H3N2-F AGCAAAAGCAGGGGATAATTC and H3N2-R AGTAGAAACAAGGGTGTTTTTA or H1N1pdm-F ATATACGCGTAGCGAAAGCAGGGGAA and H1N1pdm-R ATATACGCGTAGTAGAAACAAGGTGT with high-fidelity Phusion 2X Master Mix (New England BioLabs Inc., U.S.A.). PCR was performed by incubating the reaction mixtures at 98°C for 20 seconds, followed by 35 cycles of 98°C for 10 seconds, 57°C for 10 seconds, and 72°C for 1 minute, followed by a final extension step at 72°C for 10 minutes. PCR products were separated by electrophoresis on a 1% agarose gel. The band corresponding to the full-length HA gene segment was excised and the DNA was recovered using the QIAquick Gel Extraction Kit (Qiagen, U.S.A.).

Deep Sequencing

Gel-purified PCR products were quantified using the Qubit dsDNA High Sensitivity Kit (Invitrogen, U.S.A.). Purified PCR products were made compatible for deep sequencing using the Nexteratm XT DNA Sample Preparation Kit (Illumina, U.S.A). Each sample was enzymatically fragmented and tagged with short oligonucleotide adapters followed by 14 cycles of PCR for template barcoding. Barcoded H3N2 and H1N1pdm libraries were pooled to a final concentration of 2 nM and 10 µL of the 2 nM pooled library was denatured in 10 µL of 0.1 N NaOH for 5 minutes. Denatured pooled libraries were diluted to a final concentration of 12 pM with a PhiX-derived control library accounting for 1% of total DNA. Next, 600 µl of diluted-denatured template was added to a 500-cycle reagent cartridge. Illumina MiSeq run settings were entered into the sample sheet by Illumina Experiment Manager Software v1.3.66 as Workflow - FASTQ, Assay - Nextera XT, Chemistry – Amplicon and Reads 250 x 250. Sequence reads were imported into CLC Genomics Workbench, Version 7.5.2 (CLC bio, Denmark). Sequence reads were trimmed using a quality score cutoff of Q30 (i.e., 1 error per 1000 sequenced bases). For each sample, we mapped trimmed sequence reads against a vaccine strain HA reference sequence: A/Victoria/361/2011 (H3N2; KC306165) or A/California/07/2009

(H1N1pdm; CY121680).

Between-host evolutionary analysis

For each reference-based mapping, we extracted a consensus sequence that consists of the majority nucleotide at each nucleotide site. Consensus sequences were imported into Geneious Version 5.6.3 (Biomatters Ltd., New Zealand). In Geneious, we created individual alignments using the MUSCLE alignment algorithm [133] for H3N2 and H1N1pdm HA gene sequences. Phylogenetic trees were constructed from alignments in PHYML version 3.0 [134,134] using the HKY85 substitution model with 1000 bootstrap replicates to assess statistical confidence. From these alignments, we assessed genetic differentiation between viral consensus sequences from non-vaccinated and vaccinated subjects by performing an analysis of molecular variance (AMOVA) [135] using the Pegas package [136] in R (http://www.R-project.org). The consensus HA gene segments has been deposited in NCBI GenBank under accession numbers KR611199 - KR611312.

Within-host evolutionary analysis

Single-nucleotide polymorphisms (SNPs) were called, at nucleotide positions with coverage of at least 100 sequence reads and a central base quality score of ≥Q30, relative to a vaccine strain reference sequence (A/Victoria/361/2011 H3N2 or A/California/07/2009 H1N1pdm) using CLC Genomics Workbench Version 7.5.2. SNPs were only considered for subsequent analysis if detected at a cutoff frequency of ≥ 1%, a validated threshold that reduces the carryover of errors introduced by reverse transcription, amplification, or basecalling [17]. The nucleotide diversity statistics πN and πS were calculated in PoPoolation version 1.2.2 [137] using subsampled sequence mappings containing 1000 randomly chosen sequences per nucleotide position to minimize potential coverage bias. HA protein structures were created with MacPymol (http://www.pymol.org/). Mature H3N2 numbering is used throughout this manuscript for H3N2 and H1N1pdm viruses.

Results

Participant Characteristics

Our study included a cross-sectional sampling of 114 human subjects aged ≥ 9 years with medically attended acute respiratory illness (MAARI) who tested positive for influenza A virus by qRT-PCR (Table 1). The mean (±SD) interval from illness onset to sample collection was 2.7 ± 1.5 and 2.7 ± 1.8 days respectively for subjects who tested positive for H3N2 or H1N1pdm infection. Of these cases, we obtained swabs from 68 individuals infected with H3N2 during the 2012-13 season and 46 individuals infected with H1N1pdm during the 2013-14 season. In total, we obtained samples from 69 non-vaccinated and 45 vaccinated individuals. As shown in Table 1, 34 of these subjects received an inactivated influenza vaccine (IIV), 7 received a live attenuated influenza vaccine (LAIV), and 2 received high-dose IIV (HD); vaccine type was not recorded for an additional 2 subjects (unknown or UN).

Deep sequencing HA segments

We used deep sequencing to assess HA segment diversity within infected non-vaccinated and vaccinated individuals. Considering that mutations near the receptor-binding pocket can be introduced during propagation of clinical specimens in eggs [138] [139,140], we sequenced PCR-amplified viral nucleic acids extracted directly from swabs. Amplified HA genes were prepared for deep sequencing using the Illumina MiSeq because of its high throughput and low error rate [40,41]. Reference mappings were performed separately for each individual, yielding reference maps with mean (±SD) sequence coverage of 16,370.91 ± 10,887.82 (H3N2) and 13,786.92 ± 5,711.98 (H1N1pdm) sequences per nucleotide site.

Vaccine failure is not associated with particular HA sequences

To evaluate the evolutionary relationships among viruses circulating in our study population during the 2012-13 and 2013-14 influenza seasons, we constructed maximum-likelihood phylogenetic trees using nearly full-length nucleotide HA consensus

sequences (Figure 1). The phylogenies yielded topologies consistent with established relationships among influenza strains circulating during the 2012-13 and 2013-14 seasons [141,142]. We detected the co-circulation of H3N2 genetic subgroups 3C (66/68) and 6 (2/68) during the 2012-13 season and H1N1pdm genetic subgroup 6 (46/46) during the 2013-14 season. It was previously established that H3N2 genetic subgroups 3C and 6 were poor antigenic matches to the A/Victoria/361/2011 vaccine (Vic361) strain propagated in eggs, whereas H1N1pdm subgroup 6 remained antigenically matched to the A/California/07/2009 (CA07) strain used in the inactivated vaccine [142].

Three causes of vaccine failure can be tested with HA consensus sequences. First, if infections were caused by the introduction and spread of a single antigenic variant in this human population, we would expect HA segments from non-vaccinated and vaccinated subjects to form one phylogenetic group. Second we might observe that HA sequences form two well-supported clusters, each consisting of sequences from only non-vaccinated or only vaccinated subjects. This outcome would imply that antibodies in vaccinated subjects consistently select for similar viruses, which could either be minor variants already present in the circulating quasispecies or mutants that arise independently in multiple hosts. Third, there may be multiple sets of mutations that cause escape from vaccine-induced antibodies within vaccinated subjects. In this case, we might observe that HA consensus sequences from vaccinated humans would be more variable person-to-person than those from non-vaccinated humans, signifying that natural selection is favoring mutations away from the vaccine strain sequence.

To address these possibilities, we examined our phylogenies for monophyletic clades consisting primarily of HA consensus sequences from vaccinated subjects. We found that sequences from vaccinated and non-vaccinated individuals were generally interspersed throughout the trees. The majority of H3N2 viruses detected belonged to subgroup 3C; within this subgroup viruses from vaccinated subjects did not cluster

into clades distinct from viruses isolated from non-vaccinated individuals (Figure 1). We did detect one monophyletic clade consisting of HA sequences entirely from vaccinated individuals (n = 2; H3N2 genetic subgroup 6), but no sequences of subgroup 6 viruses from non-vaccinated subjects were available for comparison. Complementary to our phylogenetic analysis, we performed an analysis of molecular variance (AMOVA) to estimate the degree to which HA sequences differentiated into two separate genetic groups based on vaccine status [135]. AMOVA analysis showed that only a small percentage of total observed genetic variation among consensus sequences was attributable to differences in host vaccine status (H3N2-2012/13; 3.0% and H1N1-1pdm-2013/14; 3.3%), and these modest differences were not statistically significant (H3N2-2012/13; $p = 0.06$ and H1N1-2013/14; $p = 0.20$; Table 2). Together, these data indicate that vaccine failure was not associated with mutations in the HA gene detectable at the consensus level.

Action of natural selection on HA genes in human infections

While antigenic drift and transmission bottlenecks affect viral evolution during global circulation, factors like mutation rate, replication time and within-host natural selection impact how influenza evolves in the confines of a single infection. Therefore, rates of molecular evolution might differ within and between hosts. Evaluating signatures of natural selection within individual hosts may provide unique evolutionary insights not observed in HA consensus sequences. To estimate within-host HA nucleotide diversity we used the statistic π. πN, or nonsynonymous diversity, describes the frequency of pairwise sequence differences within a virus population that encode an amino acid change. Similarly, πS, describes the frequency of silent mutations within a virus population. Note that these statistics describe the genetic diversity of viral populations within a single host and do not rely on an external reference sequence. Comparing the values of these statistics for a given gene provides information about the “direction” of natural selection. Generally, a $\pi N/\pi S$ ratio > 1 indicates that positive selection

is favoring genetic diversification. By contrast, a πN/πS ratio < 1 indicates that purifying (or negative) selection is acting to maintain a fit virus population by removing deleterious mutations. We computed πN and πS for the full-length HA coding region and also for HA1, HA2, and the receptor-binding domain (RBD) to account for the possibility that selection could act differently at these discrete functional domains of the HA gene.

We computed overall mean πN and πS values for H3N2 (n = 68) and H1N1pdm (n = 46) viruses in the RBD, HA1, and HA2 domains and in full-length HA genes. Mean πN and πS values for each region ranged widely (Table 3). For example, individual means of πN for full-length HA genes ranged from 0.00040 to 0.00281 substitutions per site for H3N2 viruses and from 0.00035 to 0.00108 substitutions per site for H1N1pdm. We found that overall mean πN was lower than πS at each HA domain for all viruses (Table 3; 2-tailed $P < 0.0001$ to 0.05; Student's t-tests), showing that HA genes are generally subject to purifying selection in these subjects. We did not detect significant differences in nucleotide diversity between non-vaccinated and vaccinated subjects. The one exception to this trend was in the RBD domain of H1N1pdm viruses, in which πN was slightly, but significantly, greater in vaccinated than in non-vaccinated subjects ($P = 0.016$; Table 4). Overall, our data indicate that purifying selection was strong on both full-length HA segments and individual HA functional domains in both non-vaccinated and vaccinated subjects.

Deep sequencing reveals HA genetic and potential antigenic diversity in humans

To further characterize HA diversity within infected humans we enumerated single nucleotide polymorphisms (SNPs) relative to Vic361 and CA07 HA vaccine strain reference sequences. This allows us to detect all genetic polymorphisms in the virus population as well as fixed mutations relative to the vaccine strain. We only considered SNPs if they were detected in above our experimentally defined cutoff of ≥ 1% of sequence reads in a sample [17]. In total, we detected 695 nonsynonymous and 875 synonymous SNPs from individuals infected with H3N2 viruses; in H1N1pdm-infected

individuals we detected 700 nonsynonymous SNPs and 735 synonymous SNPs (Table 5). The SNP frequency spectrum had a bimodal distribution, with most mutations detected either at low frequency (1%-10% of sequences) or at high frequency (90%-100% of sequences). Overall, the average number of SNPs relative to vaccine strains in each host was significantly greater for H1N1pdm viruses (nonsynonymous SNPs: 14.8 ± 3.2 and synonymous SNPs: 16.0 ± 3.1) than H3N2 viruses (nonsynonymous SNPs: 10.2 ± 2.4 and synonymous SNPs: 12.9 ± 3.8; mean ± SD; 2-tailed $P > 0.0001$; Student's t-test). This greater genetic distance between vaccine and circulating strains for H1N1pdm viruses may be due to the fact that ~5 years elapsed between emergence of the CA07 vaccine strain in 2009 and the 2013-14 influenza season evaluated here, while only ~1 year separated the emergence of Vic361 and the 2012-13 season.

As discussed above, the rate of molecular evolution can vary considerably among different regions of the same gene. Therefore, we plotted the frequency of SNPs detected in all H3N2 or H1N1pdm viruses along the HA segment (Figure 2). The density of SNPs was not even across HA. For example, synonymous SNPs occurred uniformly throughout the HA gene for H3N2 viruses, while nonsynonymous SNPs accumulated disproportionately in the receptor-binding domain (Figure 2a, top). Similarly for H1N1pdm viruses, we identified elevated frequencies of nonsynonymous SNPs near the receptor-binding domain, as well as an increased number of synonymous SNPs in the HA2 domain (Figure 2b, top). The distribution of SNPs across HA gene segments is consistent with previous reports demonstrating that influenza viruses experience increased rates of positive selection in the HA1 domain due to antibody selection pressure, whereas the HA2 domain remains relatively conserved [143,144].

To identify mutations with the potential to alter influenza antigenicity, we queried SNP data for mutations that lie in previously defined antigenic sites. Nonsynonymous mutations occurred in our subjects in 9 HA amino-acid positions in H3N2 and 8 in H1N1pdm located at or near the receptor binding pocket, which were previously found

to affect viral antigenicity (Table 6 and Figure 3). Notably, we found substitutions in HA amino acid position 156 in H3N2 viruses (Q156H/R) and positions 158 and 159 of H1N1pdm viruses (G158E and N159K), all of which lie in the “antigenic ridge” of the HA protein. The antigenic ridge is a span of residues on the periphery of the receptor-binding site previously shown to be responsible for antigenic change in H3N2 [25] and H1N1pdm viruses [145]. An A→G SNP at nucleotide 515 encoding a glutamine-to-histidine change at amino acid position 156 was detected at or near fixation all H3N2-infected subjects. This Q156H substitution was associated with vaccine mismatch during the 2003-2004 influenza season and is found in HA consensus sequences of circulating H3N2 viruses ever since [146]. The A→C SNP encoding the glutamine-to-arginine substitution at amino acid position 156 was detected in a single vaccinated subject (V130083) in 3.3% of viral sequences. The Q156R mutation was previously linked to egg adaptation of the 2012-2014 Vic361 vaccine virus [147]. In H1N1pdm viruses, we detected a G→A SNP at nucleotide 535 encoding a glycine-to-glutamate substitution at amino acid position 158 in 2.3% and 2.4% of viral sequences in U140225 and V140227 respectively. G158E is detected in consensus H1N1pdm HA sequences in Influenza Research Database [148] but only in 0.65% (n = 4001) of samples. Interestingly, viruses with the G158E substitution replicated in vitro in the presence of H1N1pdm-specific ferret antibodies [149]. A T→A SNP encoding an asparagine-to-lysine substitution at position 159 was detected in three H1N1pdm-infected subjects at low frequencies: V14022 at 1.4%, V140172 at 2.2%, and U14106 at 5.7% of viral sequences. In a previous study, H1N1pdm viruses bearing the N159K substitution (N156K in H1 numbering) outcompeted wild type viruses in tissue culture, replicated efficiently in H1N1pdm immunized ferrets, and were efficiently transmitted among animals [150]. Interestingly, despite causing escape from H1N1pdm-specific antibodies in vitro and in vivo in ferret studies, this mutation is found at the consensus level in a minority of viruses circulating in humans (0.1% of sequences; n = 4001). Taken together, these data show that muta-

tions with the potential to encode antigenic variation are generated in humans naturally infected with influenza viruses. These mutations exist well below the detection limit of Sanger sequencing in our subjects, and are only rarely detected at the consensus level in humans.

Discussion

Although influenza viruses are known to exist as genetically diverse quasispecies, the potential impact of this diversity on viral adaptation to immune selection within individual hosts is not well understood. Currently, the majority of human influenza isolates are propagated in eggs before genetic and antigenic characterization. Influenza growth in eggs selects for changes in the receptor-binding domain of HA, which can have a significant impact on the antigenicity [22]. Propagation in eggs can also impact the composition of viral quasispecies, reducing diversity by selecting for specific variants and eliminating others [95]. Here we characterized influenza HA genes directly from human clinical specimens and evaluated the contribution of HA within-host genetic variation to vaccine failure, as well as the potential ways in which vaccine-induced immunity shapes within- and between-host diversity. Our analyses showed that influenza viruses infecting non-vaccinated and vaccinated subjects were genetically similar and did not display patterns of sequence diversity consistent with the de-novo selection of antigenic variants within any individual host. Within-host HA genetic diversity in either non-vaccinated or vaccinated subjects consists of more synonymous diversity than nonsynonymous diversity. This pattern suggests that purifying selection, acting to remove deleterious mutations, is the dominant evolutionary force affecting HA in this cohort of humans infected with H3N2 or H1N1pdm viruses. Interestingly, we observed several nonsynonymous mutations below the detection threshold of Sanger sequencing in both H3N2 and H1N1pdm virus populations located in positions that can cause significant antigenic change. Moreover, several low-frequency nonsynonymous mutations were detected reproducibly in multiple subjects, in both non-vaccinated and vaccinated groups. This observation raises the possibility that antigenic variants are frequently generated during influenza infection in humans, but that they often do not become fixed in the viral population even when they would be expected to have a strong fitness advantage.

There are important caveats to the interpretation of our results. First, we initially compared HA gene sequences from non-vaccinated and vaccinated subjects to identify potential mutations associated with vaccine failure. Humans have highly variable pre-existing immune repertoires against influenza, which are shaped by individual infection and vaccination histories, as well as by age, comorbidities, and host genetics [149,151-153]. Grouping virus sequences purely by same-season vaccine status may not adequately capture important biological differences among these subjects. When possible, it would be desirable to use immunological characteristics (e.g., hemagglutination-inhibition titers) to stratify subjects into more biologically relevant groups; unfortunately that was not possible for our cohort. Secondly, because autologous serum was not available, we do not know whether putative antigenic variants specifically escape antibodies present in the subjects in which the variants were detected. Despite these caveats, past reports have shown that Q156H/R (H3N2) impacts antigenicity of vaccine strains, and G158E and N159K (H1N1pdm) confer escape from H1N1pdm-specific antibodies in vitro and in vivo.

The extent of within-host HA genetic diversity and HA consensus sequences were similar between non-vaccinated and vaccinated subjects. It is likely that virtually all of these subjects had some degree of pre-existing immunity to influenza due to previous infection and/or vaccination. This may be particularly relevant for H1N1pdm, as H1N1pdm viruses have not undergone antigenic change since their emergence in 2009 and subjects could have been exposed to antigenically similar viruses for several years prior to the 2013-14 season evaluated here. The evident lack of antibody escape mutations in HA consensus segments in subjects with vaccine failure is therefore surprising. We cannot explain why antibody escape mutations did not emerge in HA consensus sequences using sequence data alone. However, we speculate that for antigenic variants to reach fixation in the virus population, compensatory mutations maybe needed to enhance viral replicative fitness before the “less fit” variant is lost by purifying selec-

tion. Finally, we cannot exclude the possibility that antibody escape mutations identified in vitro or small animal models do not confer escape in humans. Despite these possibilities, we found sequence signatures demonstrating that nonsynonymous diversity was significantly greater in the receptor-binding domain of H1N1pdm viruses within infected vaccinated subjects. This region corresponds to an antigenic “hot spot” in which we identified putative antibody escape mutations (G158E and N159K) in human subjects. Considering that G158E and N159K caused escape from H1N1pdm-specific antibodies in tissue culture and in ferrets, we hypothesize that genetic diversification in the receptor-binding domain of H1N1pdm viruses in vaccinated humans was due to vaccine-induced antibody pressures. Conversely, we did not observe genetic diversification in H3N2 viruses likely because of poor antigenic match between circulating viruses and vaccine strain.

Our results are consistent with a past report showing that prior immunity had little effect on the level and structure of genetic diversity of influenza viruses infecting vaccinated horses [154]. In that study, Murcia and colleagues speculated that purifying selection (i.e., more synonymous diversity relative to nonsynonymous diversity) is the dominant signal of influenza within individual hosts [154]; this is consistent with the idea that influenza viruses are well adapted to replication and transmission in their natural hosts, and therefore most mutations are likely to be deleterious. Indeed, mutations that alter HA antigenicity may frequently exact a fitness cost that requires compensatory mutations to restore viral replicative capacity [155]. This may explain why the putative antigenic variants we observed here remained at low frequency in infected humans. We speculate that compensatory mutations may be needed to overcome such fitness deficits, and the compensatory mutation(s) need to occur before the antigenic variant is removed from the virus population by purifying selection. Because each subject was sampled only once, we could not follow the changing frequencies of SNPs with time to detect potential positive selection at individual sites. Alternative approach-

es using time-series samples enable better detection of site-specific positive selection [156].

From 1968 to 2011 fourteen antigenic clusters of H3N2 viruses have emerged, each cluster being replaced by viruses with distinct antigenic characteristics [24]. Single amino acid substitutions, restricted to 7 positions (145, 155-159, 189, and 193) on the HA protein, were responsible for antigenic cluster transitions during the evolution of H3N2 viruses from 1968 to 2003 [25]. One mutation detected in H3N2 viruses in our study, encoding Q156H, fell within this defined "antigenic ridge." A Q156H substitution was responsible for the Sydney-1997 to Fujian-2002 antigenic cluster transition during the 2003-04 season resulting in vaccine mismatch [25,146]. Acquisition of glutamine (Q) at HA position 156 during egg passage of vaccine seed stocks was also associated with low vaccine effectiveness against H3N2 viruses in the 2012-13 season [22]. Contemporary H3N2 strains in circulation possess a histidine at HA position 156 in publically available consensus sequences, however we found that 3.3% of sequences within a single vaccinated case (V130083) possessed a Q156R substitution. Acquisition of arginine at position 156 has been previously linked to egg adaptation of the initial egg-grown A/Victoria/361/2011 wild-type virus for the production of the 2012-13 and 2013-14 vaccines [147]. However, the impact of Q156R on recognition of antibodies raised by vaccines encoding glutamine at position 156 has yet to be evaluated in vitro or in animal models.

The overwhelming majority of isolated H1N1pdm viruses remain antigenically similar to the prototypical A/California/07/2007 vaccine virus [157]. In the HA protein of H1N1pdm viruses, substitutions in amino acid positions 154-162 (H3 numbering) located adjacent to the receptor-binding site can cause escape from human infant and ferret antibodies after H1N1pdm virus infection (i.e., ≥ 4 fold reduction in HI titer) [145]. We detected a G158E substitution in a non-vaccinated case and a vaccinated case in 2.3 and 2.4% of sequences, respectively. The G158E substitution considerably alters

the electrostatic properties which may be important for surface interactions between the HA protein and antibodies [149], changing the antigenic properties of H1N1pdm viruses as mapped by human monoclonal antibodies [145,158], and polyclonal antibodies in ferrets [145] [149] [150]. In three humans, one non-vaccinated and two vaccinated, we detected an N159K substitution at low frequency. The N159K mutation has been described as affecting the local electrostatic charge potential of the HA protein [150]. In a ferret model, the N159K mutation was found to confer escape from H1N1pdm-specific antibodies elicited by suboptimal vaccination [150]. Viruses with the N159K mutation remained fit in a ferret transmission model and outcompeted the wild-type virus. During replication in ferrets, N159K variant viruses acquired additional mutations in HA and other gene segments. Interestingly, when the N159K mutant virus was generated by reverse genetics using backbones from A/Puerto Rico/8/1934 (H1N1) and A/Perth/261/2009 (H1N1pdm) it could not be rescued as a pure population demonstrating that additional mutations not in the reverse genetic virus must be acquired to restore fitness. Furthermore, because a pure population of N159K virus was not recovered, escape from adult human antisera could not be tested [150]. In another study, Koel and colleagues introduced different substitutions at position 159 (N→D, N→G, N→Y and N→S) and found that this position was important for antigenic change of H1N1pdm viruses, as measured against both human and ferret sera [145].

The ways in which influenza viruses evolve antigenically at the population level have been the subject of intense study in the past decade [24,25,159,160]. However, new antigenic variants are initially generated and selected at the level of individual infected hosts. Here we showed that influenza virus quasispecies populations infecting individual humans possess potential antigenic variation at low frequency. We previously showed that even very low-frequency influenza viral quasispecies can pass through selective bottlenecks and be transmitted to new hosts via the airborne route. It is therefore possible that low-frequency putative antigenic variants we observed here

could be transmitted via respiratory droplets. The use of Sanger sequencing for influenza surveillance typically defines consensus sequences and cannot resolve minority variants present below 20% of the viral population [31]. Traditional Sanger sequencing could therefore miss potentially antigenic variants that occur in nature. The evolutionary mechanism by which antigenic variants are generated and selected for within individual hosts remains unclear. Our findings suggest that generating antigenic variants through mutation alone may not be sufficient for the emergence of new antigenic variants within a host. Instead antigenic variants likely need compensatory mutations to restore replicative fitness, avoid loss by genetic drift, and be transmitted to new hosts to successfully emerge in nature. Additional studies are needed to more fully define the molecular basis of antigenic change, and perhaps more importantly, to determine the impact of antigenic changes on viral fitness and the evolutionary pathways by which fitness is restored in antigenic variants.

Acknowledgments

The authors would like to acknowledge Jennifer King, Huong McLean and Maria Sundaram for their outstanding support of this study and thoughtful scientific discussion. We gratefully acknowledge the authors, originating and submitting laboratories of the sequences used in this study. A full list of GISAID acknowledgements can be found in Table S1. This work was supported by the Centers for Disease Control and Prevention through a cooperative agreement (U01 IP000471) awarded to E.A.B and by a supplement to the WNPRC base grant (NIH P51 RR000167/OD011106), awarded to T.C.F. J.M.D. was supported by the National Science Foundation Graduate Research Fellowship DGE-0718123, and the National Academy of Sciences Ford Foundation Dissertation Fellowship. N.W.F. and L.H.M. were supported by NIH National Research Service Award T32 GM07215. O.O.F. was supported by the IBS-SRP summer research fellowship DBI-1063085.

Figure Legends

Figure 1. Phylogenetic relationships of influenza A viruses infecting non-vaccinated and vaccinated subjects in Marshfield, Wisconsin. (a) H3N2 phylogenetic tree constructed using 68 hemagglutinin genes from subjects infected during the 2012-13 season and 11 outgroup sequences representing WHO-recommended vaccine strains and representative H3N2 taxonomic clades (alignment length: 1641 nucleotides). (b) H1N1pdm phylogenetic tree constructed using 46 hemagglutinin genes from subjects infected during the 2013-14 season and 9 outgroup sequences representing WHO-recommended vaccine strains and representative H1N1pdm taxonomic clades (alignment length: 1698 nucleotides). GISAID and Genbank accession numbers for the included taxa are provided in Supplemental Table S1. Bootstrap values were determined from 1000 replicates and are indicated above the corresponding nodes when values are above 50%.

Figure 2. Deep sequencing reveals sequence variation in hemagglutinin genes during human infection. We used deep sequencing to assess within-host viral variation in non-vaccinated and vaccinated subjects directly from clinical specimens. Displayed are single-nucleotide polymorphisms (SNPs) detected in (a) H3N2 and (b) H1N1pdm viruses. Nonsynonymous (circles) and synonymous (squares) mutations were called relative to a Vic361 (H3N2) or CA07 (H1N1pdm) vaccine strain reference sequence. The x-axis represents the nucleotide position and the y-axis the frequency at which each mutation was detected. Above each SNP frequency plot is a cartoon depiction of the linear HA gene with shaded functional domains: HA1 (dark grey), HA2 (light grey), and the receptor-binding domain ("RBD"; black). Density plots indicate the likelihood for a nonsynonymous (black) or synonymous (grey) SNP to occur, regardless of frequency, in a given position across the HA gene.

Figure 3. Localization of amino acid substitutions identified in this study on the HA structure. Structure of an A/Aichi/2/1968 HA trimer (Protein Data Bank accession 5HMG) 66. The three monomers are shown in black and grey with the receptor-binding site in light brown and positions previously defined as responsible for antigenic cluster transitions shown in dark brown 14. Mutations detected in human subjects infected with H3N2 and H1N1pdm viruses are shown in red. All mutations are shown in H3 numbering. (a) H3N2 antigenic sites as defined by Wiley et al. 9; antigenic sites A-E are shown in blue, green, magenta, orange and yellow, respectively. (b and c) Mutations detected in subjects infected with H3N2 viruses. (d) H1N1 antigenic sites as defined by Caton et al. 10; antigenic sites Sa, Sb, Ca1, Ca2 and Cb are shown in blue, green, magenta, orange and yellow, respectively. (e and f) Mutations detected in subjects infected with H1N1pdm viruses. Images were created with MacPymol (http:// www.pymol.org/).

Tables, Figures and Supplementals

Table 1. Descriptive characteristics of influenza A cases by season

Characteristic	2012-13 N=68	2013-14 N=46	Total N=114
Age Group			
<9 years	18	9	27
9-17 years	5	5	10
8-49 years	20	17	37
≥50 years	25	15	40
Vaccination status			
Non-vaccinated	37	32	69
Vaccinated	31	14	45
IIV	24	10	34
LAIV	3	4	7
HD	2	0	2
Vaccinated, unknown type	2	0	2
Interval from onset to swab			
< 3 days	34	29	63
3-4 days	22	8	30
5-7 days	12	9	21
Swab month			
December	13	22	35
January	37	22	59
February	17	2	19
March	1	0	1
Influenza A subtype			
H3N2	68	0	68
H1N1pdm	0	46	46

Table 2. Analysis of molecular variance (AMOVA) of seasonal influenza viruses based on vaccination status.

Observed divergence for H3N2 viruses						
Source	df	SSD	MSD	σ	% variance	P value
Between populations	1	242.2	242.2	3.7	3.0	0.06
Within populations	66	7778.2	117.9	117.9	97.0	
Total	67	8020.3	119.7	121.5		
Observed divergence for H1N1pdm viruses						
Source	df	SSD	MSD	σ	% variance	P value
Among populations	1	37.1	37.1	0.6	3.3	0.20
Within populations	44	1098.2	25.0	25.0	96.7	
Total	45	1135.3	25.2	25.6		

Abbreviations: df; degrees of freedom, SSD; sum of squared deviations, MSD; mean square deviations, σ; variance component

Table 3. Estimates of nonsynonymous and synonymous nucleotide diversity for hemagglutinin genes of influenza A viruses.

Subtype	Domain	π_N	SEM	Range	π_S	SEM	Range	π_N vs. π_S, P value
H3N2	HA	0.00069	0.00004	0.00040-0.00281	0.00124	0.00019	0.00051-0.01348	0.0053
	HA1	0.00070	0.00004	0.00039-0.00261	0.00116	0.00015	0.00032-0.00968	0.0036
	HA2	0.00068	0.00005	0.00039-0.00342	0.00147	0.00038	0.00036-0.02642	0.0412
	RBD	0.00076	0.00006	0.00036-0.00444	0.00138	0.00021	0.00029-0.01333	0.0052
H1N1pdm	HA	0.00060	0.00002	0.00035-0.00108	0.00107	0.00011	0.00033-0.00382	0.0001
	HA1	0.00061	0.00003	0.00032-0.00135	0.00096	0.00010	0.00028-0.00315	0.0012
	HA2	0.00059	0.00003	0.00033-0.00106	0.00125	0.00017	0.00034-0.00499	0.0002
	RBD	0.00056	0.00003	0.00026-0.00100	0.00099	0.00012	0.00024-0.00392	0.0008

Abbreviations: π_N; nonsynonymous nucleotide diversity, π_S; synonymous nucleotide diversity; SEM; standard error of the mean P values correspond to comparisons of mean π_N and π_S values at each HA gene domain. Statistical analysis performed using student's t test.

Table 4. Estimates of nonsynonymous and synonymous nucleotide diversity in non-vaccinated and vaccinated subjects

		Non-vaccinated			Vaccinated			
Subtype	Domain	π_N	SEM	Range	π_N	SEM	Range	P value
H3N2	HA	0.00072	0.00006	0.00040 - 0.00281	0.00065	0.00003	0.00040 - 0.00114	0.329
	HA1	0.00072	0.00006	0.00041 - 0.00261	0.00068	0.00003	0.00039 - 0.00125	0.576
	HA2	0.00073	0.00008	0.00039 - 0.00342	0.00062	0.00003	0.00042 - 0.00144	0.235
	RBD	0.00081	0.00011	0.00043 - 0.00444	0.00070	0.00004	0.00036 - 0.00117	0.385
H1N1pdm	HA	0.00058	0.00003	0.00035 - 0.00108	0.00067	0.00004	0.00038 - 0.00090	0.094
	HA1	0.00059	0.00004	0.00032 - 0.00135	0.00066	0.00005	0.00036 - 0.00094	0.316
	HA2	0.00056	0.00003	0.00033 - 0.00096	0.00067	0.00005	0.00036 - 0.00106	0.056
	RBD	0.00052	0.00003	0.00026 - 0.00100	0.00066	0.00005	0.00042 - 0.00097	0.016
Subtype	Domain	π_S	SEM	Range	π_S	SEM	Range	P value
H3N2	HA	0.00139	0.00035	0.00051 - 0.01348	0.00106	0.00010	0.00065 - 0.00320	0.405
	HA1	0.00125	0.00025	0.00032 - 0.00968	0.00106	0.00011	0.00062 - 0.00305	0.517
	HA2	0.00180	0.00069	0.00036 - 0.02642	0.00108	0.00014	0.00064 - 0.00401	0.350
	RBD	0.00158	0.00037	0.00029 - 0.01333	0.00113	0.00015	0.00046 - 0.00518	0.296
H1N1pdm	HA	0.00110	0.00016	0.00033 - 0.00382	0.00099	0.00010	0.00047 - 0.00191	0.665
	HA1	0.00099	0.00014	0.00028 - 0.00315	0.00088	0.00006	0.00041 - 0.00128	0.614
	HA2	0.00128	0.00023	0.00034 - 0.00499	0.00117	0.00023	0.00043 - 0.00348	0.774
	RBD	0.00101	0.00017	0.00024 - 0.00392	0.00095	0.00013	0.00034 - 0.00228	0.827

Abbreviations: πN; nonsynonymous nucleotide diversity, πS; synonymous nucleotide diversity; SEM; standard error of the mean P values correspond to comparisons of mean πN and πN or πS and πS values for each HA gene domain (values were compared horizontally). Statistical analysis performed using student's t test.

Table 5. Summary of nonsynonymous and synonymous mutations detected in our study.

Mutation Frequency [%]	H3N2-2012/13 (n = 68)			H1N1pdm-2013/14 (n = 46)		
	N	S	N + S	N	S	N + S
1 - 10	97	84	181	100	64	164
10 - 20	6	6	12	9	6	15
20 - 30	2	6	8	1	2	3
30 - 40	3	0	3	1	0	1
40 - 50	0	0	0	0	0	0
50 - 60	0	0	0	0	0	0
60 - 70	2	6	8	0	0	0
70 - 80	2	1	3	0	2	2
80 - 90	0	0	0	1	7	8
90 - 100	583	772	1355	588	654	1242
Total	695	875	1570	700	735	1435

Table 6. Single nucleotide polymorphisms detected in antigenicity-associated HA positions from infected humans.

Subtype	Nt. pos.	Nucleotide change	H3 a.a. number	Amino acid change	Antigenic site	Number of subjects	Frequency range
H3N2	205	G → A	52	D → N	C	2	100%
	472	A → G	142	R → G	A	19	2.1 - 100%
	472	AG → GA	145	R → E	A	1	2.0%
	515	AA → GC	156	Q → R	B	1	3.3%
	516	A → C	156	Q → H	B	68	96.1 - 100%
	668	A → C	207	K → T	D	1	3.7%
	670	A → G	208	R → G	D	1	1.2%
	706	A → G	220	R → G	D	2	2.9 - 3.5%
	882	T → G	278	N → K	C	66	100%
H1N1pdm	283	C → A	80	S → Y	Cb	1	1.4%
	283	C → T	80	S → F	Cb	1	2.2%
	292	GC → AT	82b	S → N	Cb	1	100%
	490	G → A	143	G → E	Ca2	1	1.1%
	535	G → A	158	G → E	Sa	2	2.3 - 2.4%
	539	T → A	159	N → K	Sb	3	1.4 - 5.8%
	550	A → G	163	K → R	Sa	2	1.0 - 2.5%
	558	A → C	183	K → Q	Ca1	45	100%

Antigenic sites refer to defined HA protein sites important for the recognition of neutralizing antibodies [9,10]. Number of subjects signifies the total number of individuals (non-vaccinated and vaccinated) for which a particular nucleotide change was observed (H3N2; n = 68 and H1N1pdm; n = 46). Frequency range indicates the minimum and maximum frequency that a particular nucleotide change was observed in our cohort.

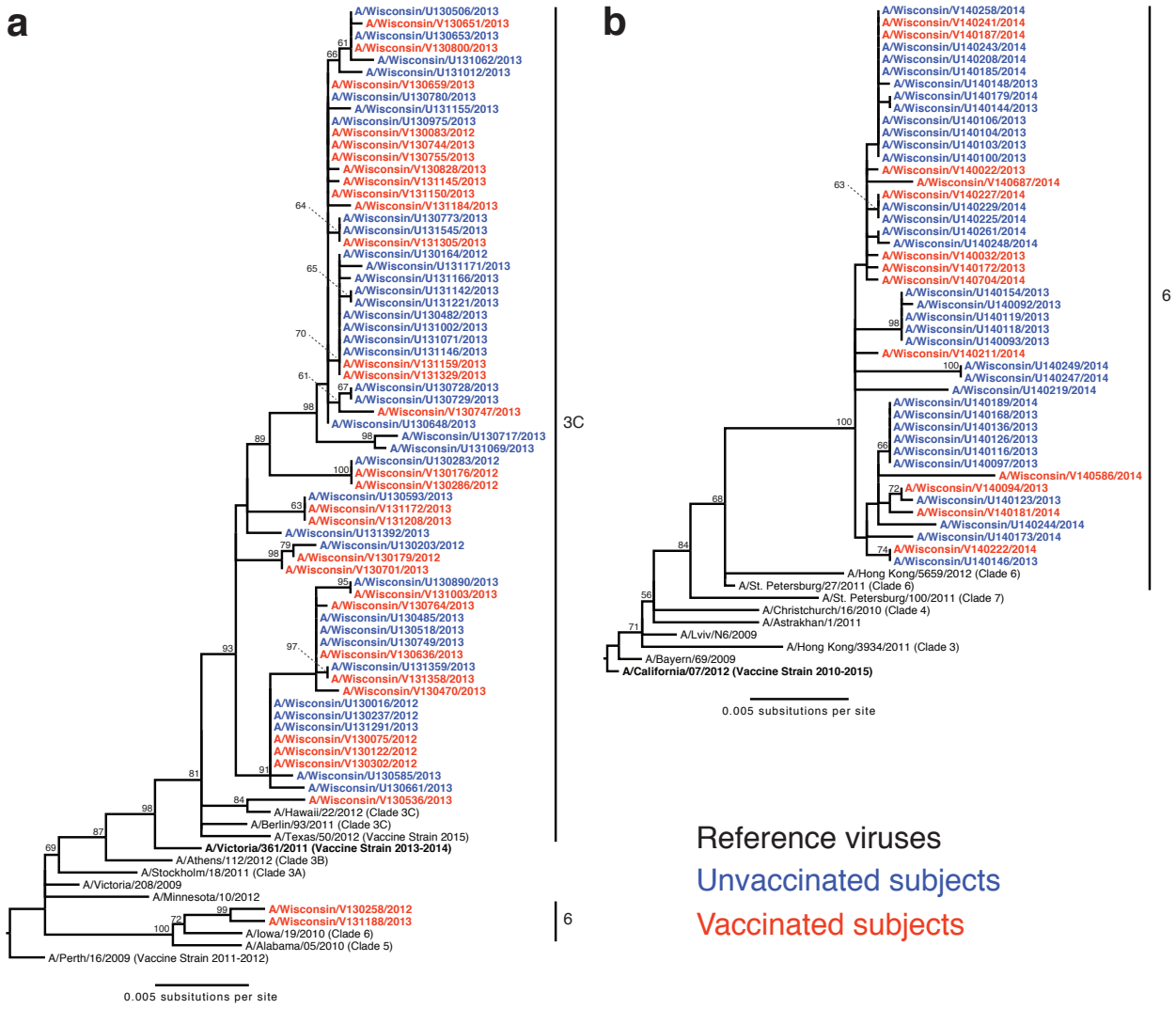


Figure 1. Phylogenetic relationships of influenza A viruses infecting non-vaccinated and vaccinated subjects in Marshfield, Wisconsin. (a) H3N2 phylogenetic tree constructed using 68 hemagglutinin genes from subjects infected during the 2012-13 season and 11 outgroup sequences representing WHO-recommended vaccine strains and representative H3N2 taxonomic clades (alignment length: 1641 nucleotides). (b) H1N1pdm phylogenetic tree constructed using 46 hemagglutinin genes from subjects infected during the 2013-14 season and 9 outgroup sequences representing WHO-recommended vaccine strains and representative H1N1pdm taxonomic clades (alignment length: 1698 nucleotides). GISAID and Genbank accession numbers for the included taxa are provided in Supplemental Table S1. Bootstrap values were determined from 1000 replicates and are indicated above the corresponding nodes when values are above 50%.

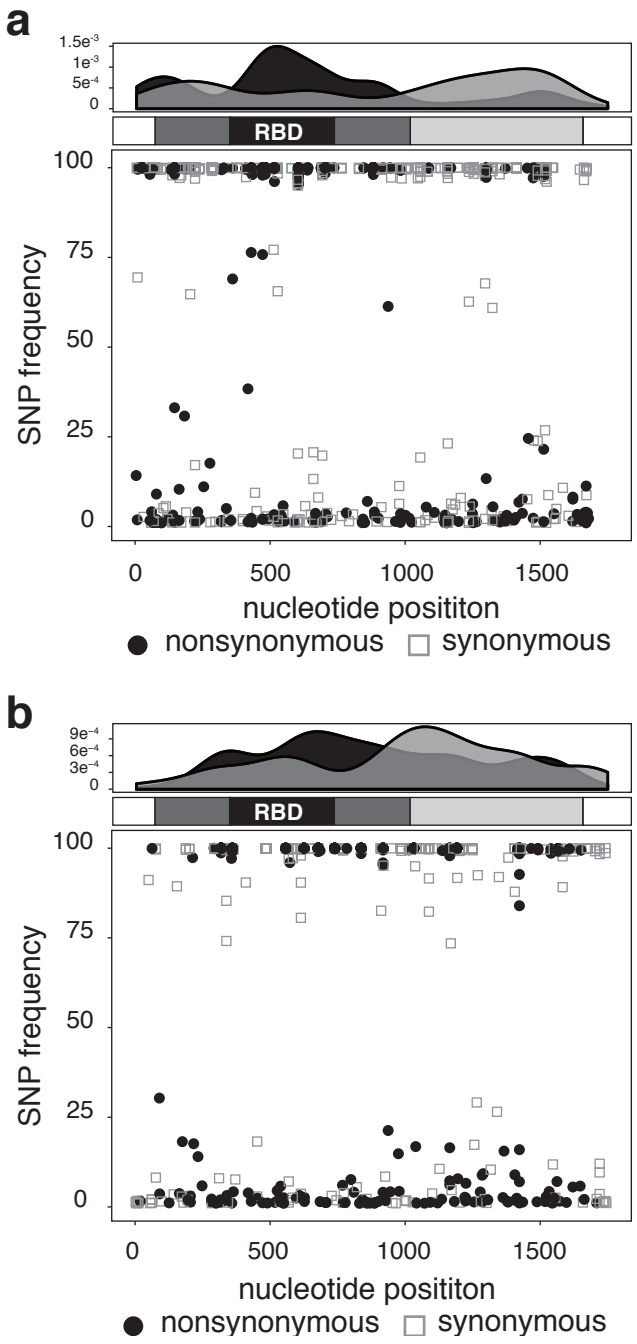


Figure 2. Deep sequencing reveals sequence variation in hemagglutinin genes during human infection. We used deep sequencing to assess within-host viral variation in non-vaccinated and vaccinated subjects directly from clinical specimens. Displayed are single-nucleotide polymorphisms (SNPs) detected in (a) H3N2 and (b) H1N1pdm viruses. Nonsynonymous (circles) and synonymous (squares) mutations were called relative to a Vic361 (H3N2) or CA07 (H1N1pdm) vaccine strain reference sequence. The x-axis represents the nucleotide position and the y-axis the frequency at which each mutation was detected. Above each SNP frequency plot is a cartoon depiction of the linear HA gene with shaded functional domains: HA1 (dark grey), HA2 (light grey), and the receptor-binding domain ("RBD"; black). Density plots indicate the likelihood for a nonsynonymous (black) or synonymous (grey) SNP to occur, regardless of frequency, in a given position across the HA gene.

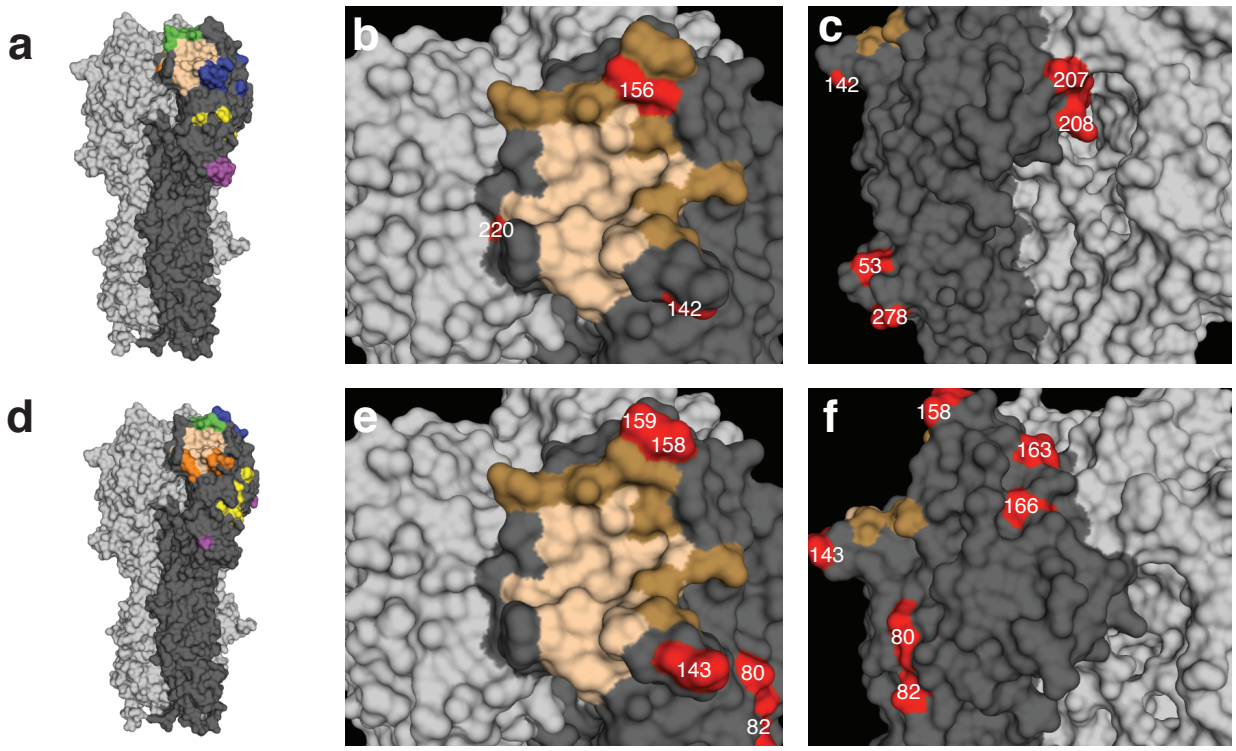


Figure 3. Localization of amino acid substitutions identified in this study on the HA structure. Structure of an A/Aichi/2/1968 HA trimer (Protein Data Bank accession 5HMG) 66. The three monomers are shown in black and grey with the receptor-binding site in light brown and positions previously defined as responsible for antigenic cluster transitions shown in dark brown 14. Mutations detected in human subjects infected with H3N2 and H1N1pdm viruses are shown in red. All mutations are shown in H3 numbering. (a) H3N2 antigenic sites as defined by Wiley et al. 9; antigenic sites A-E are shown in blue, green, magenta, orange and yellow, respectively. (b and c) Mutations detected in subjects infected with H3N2 viruses. (d) H1N1 antigenic sites as defined by Caton et al. 10; antigenic sites Sa, Sb, Ca1, Ca2 and Cb are shown in blue, green, magenta, orange and yellow, respectively. (e and f) Mutations detected in subjects infected with H1N1pdm viruses. Images were created with MacPymol (http:// www.pymol.org/).

Conclusions and Future Directions

Sequence-based genetic characterization of within-host viral diversity is central to all the findings presented in my thesis. The within-host spectrum of variant viruses reflects a balance between the generation of diversity through mutation and the loss of diversity through purifying selection. However capturing the genetic diversity of within-host virus populations requires “deep” genetic resolution to detect variants present at low frequencies. Traditional methods used to describe within-host viral diversity relied on cloning and Sanger sequencing [33]. Deep sequencing virus populations is a more rapid and effective strategy [11,39,162]. Therefore, as part of my graduate work, I developed novel methods to investigate viral genetic diversity from bench to the computer. As a result, I developed a novel next-generation sequencing approach for the Illumina MiSeq to characterize influenza virus populations from diverse animal and human samples provided by a talented investigative team. This approach circumvents the need for plasmid cloning while simultaneously producing thousands to millions of short sequences (reads), allowing the high-throughput screening of entire viral populations at unprecedented genetic resolution. Analytically, I took advantage of the large amount of sequence data generated in a single “run” to discover low-level single nucleotide polymorphisms (SNPs), i.e., genetic mutations relative to an external reference sequence or among viruses from the same population. Deep sequence “reads” are used as a relative measure of virus abundance, and, since each “read” represents a single molecule of template DNA, deep sequence data can be used to identify linkage relationships among individual SNPs. Our data strongly indicate that that current influenza surveillance methods, based on Sanger sequencing with its estimated SNP detection limit of ~20% [85], may not provide adequate genetic resolution to assess the true pandemic potential of influenza viruses circulating in nature. Implementing deep sequencing approaches in influenza genetic surveillance may provide critical information regarding the early emergence of potential troublesome variants currently circulating in nature.

Emergence of pandemic viruses requires efficient human-to-human transmission through direct contact, large droplets, or aerosols [163,164]. In chapter 2, we examined impact H5-reassortant influenza virus within-host genetic diversity of H5-reassortant influenza virus on host adaptation and transmission in ferrets. We identified a number of SNPs that changed markedly in frequency during infection of index ferrets with the reassortant viruses, suggesting that low-frequency variants with greater in-vivo replicative fitness than the consensus sequence can rapidly increase to fixation. Interestingly, we found that transmission did not result in the transfer of a representative sample of virus from the index to contact animal. Instead, we found that SNPs were either detected at nearly 100% or were almost completely absent following transmission. Strikingly, a minor variant detected at 5.9% of viruses in an index animal near the time of transmission could be found dominating the population replicating shortly after transmission in the paired contact animal. Upon additional computational analysis, we showed for the first time, that natural selection acting against HA imposed a bottleneck during transmission of influenza viruses. Since then, others have begun to focus on the ways in which transmission bottlenecks could impact cross-species transmission and the onward evolution of influenza viruses. We speculate that transmission bottlenecks driven by natural selection may favor the emergence of avian influenza viruses in mammals. Although, not all transmission bottlenecks are driven by natural selection, Varble and colleagues showed that mammalian H1N1pdm viruses randomly pass through bottlenecks without the influence of natural selection [95]. Together, these experiments demonstrate that differences in viral genetics (i.e., partially adapted H5N1 virus vs. mammalian pandemic virus) can dramatically affect the relative amount of selectivity and stochasticity observed during transmission in mammals. Therefore, it's conceivable that bottlenecks may be used as signatures of pandemic threat, such that partially-mammalian adapted or "pre-emerging" strains display selection-driven bottlenecks. Ideally, bottlenecks could be used as a new measure of viral pandemic potential are

needed to inform pandemic preventative and control measures.

Bottlenecks could be shaped by physical factors such as inoculum size; site of infection; and/or exchange between anatomical sites, and by virological factors, like receptor binding and polymerase activity, that are encoded by viral genes and subject to natural selection. We hypothesize that viral genetics determine the relative contribution of stochastic and selective processes on transmission bottlenecks in mammals. Moreover, bottlenecks may therefore have a profound impact on the "onward" evolution of influenza viruses in new hosts, and may form a critical barrier to pandemic emergence. The factors that influence their stringency are poorly understood. Deep sequencing could be used to probe viral quasispecies dynamics to disentangle the relative contribution of physical and virological factors that govern the stringency and composition of transmission bottlenecks. For example, mammal-adapted H1N1pdm reporter viruses can be used as a tool to elucidate the effects of inoculum size, site of infection, and viral "mixing" among anatomical sites on the transmission bottleneck. Engineering mutant H1N1pdm viruses with "avian-like" genetic motifs in the HA, polymerase complex, or other segments will determine the relative contribution for virological factors on transmission bottlenecks. We predict that these experiments will reveal how viral phenotypes can directly change the relative extent of stochasticity and/or selectivity that occurs during transmission bottlenecks in mammals. Understanding the role of transmission bottlenecks on avian influenza virus emergence in humans is critical for assessments of viral pandemic potential.

In chapter 2, we demonstrated that low-frequency influenza viruses are important for host adaptation and transmission in mammals, demonstrating that minor viral variants possess an underappreciated role in viral emergence. In chapter 3, we built on our ferret transmission experiments by examining the extent of H5N1 quasispecies adaptation in prolonged human infections. We used deep sequencing to characterize within-host genetic diversity, and for variants detected in more than 5% of sequences,

we generated plasmids and reassortant viruses for in vitro characterization (i.e., receptor-binding specificity, HA protein stability polymerase activity, and interferon antagonism). We found that H5N1 quasispecies were genetically diverse, but surprisingly, viruses within the quasispecies predominantly retained avian-like phenotypes. Interestingly, we found two low frequency mutations (PA-M90 at 5.1% and PA-E143 at 12.6%) that increased polymerase activity in vitro. Despite providing a replicative advantage in cell culture, these mutations were rapidly removed from the virus population in humans. We provide evidence that the principal evolutionary force shaping H5N1 quasispecies in late human infection was purifying selection. Interestingly, the rapid removal of genetic diversity during infection may limit the onward evolution of viruses in nature [118]. We posit that avian influenza virus adaptation to humans does not occur via constant incremental fitness increases in each infected host (Figure 1; Graphical abstract). Instead, once the virus reaches peak titers, positive selection may not be capable of outcompeting the existing avian consensus in a host and multiple transmission bottlenecks may be needed to reshape quasispecies composition and allow for adaptation.

Understanding the processes by which influenza viruses acquire new protein functions is important for public health. However, how to study influenza protein function, especially using experiments that generate viruses with enhanced pathogenicity and transmissibility in mammals, is fiercely debated [165,166]. In brief, research that improves the ability of a pathogen to cause disease is termed “gain-of-function” (GOF); for example, generating avian H5N1 viruses capable of mammalian transmissibility [47,65]. Due to the inherent biosafety and biosecurity risks, the United States government has instituted a mandatory moratorium on GOF studies for influenza, SARS, and MERS viruses (https://www.whitehouse.gov/blog/2014/10/17/doing-diligence-assess-risks-and-benefits-life-sciences-gain-function-research). Alternative experimental approaches have been proposed: mathematical modeling, in vitro studies using individual viral proteins, sequence database comparisons between avian and mammalian iso-

lates [167]. But, as I described in chapter 2 and 3, the mechanistic processes in which avian viruses acquire new biological functions is critical for understanding how influenza viruses emerge in nature. Assessing the role of minor variants during host adaptation could provide critical information about the ease with which circulating isolates adapt in and transmit between mammals and further elucidating the role of bottlenecks in viral emergence. Based on my work presented in this thesis, the next logical step is to serially passage avian influenza quasispecies with differing predicted phenotypes in ferrets. However, as the moratorium currently exists, these needed experiments cannot be performed. Uniquely, my past research experiences in GOF and "alternative" studies have shaped my opinion on GOF studies in three ways: (1) direct ferret inoculation with avian influenza viruses is the only biologically relevant way to study mammalian adaptation and other "alternative" approaches are only indirect measures with little biological relevance, (2) GOF studies are not without risks, and (3) to minimize the risk of accidental release, experiments should be performed in BSL-4 laboratories and all samples containing infectious virus should be inactivated at the end of the study. Overall, the mandatory moratorium of GOF experiments has forced a debate regarding the potential risk and benefits of GOF experiments. This debate needs to be resolved quickly because avian influenza viruses continue to evolve in nature, potentially closer to a transmissible phenotype, but our ability to directly study biological mechanisms that give rise to influenza pandemics remains in pause.

Finally, in chapter 4 we used deep sequencing to characterize between- and within-host HA segment diversity in a cohort of patients that included individuals who were vaccinated and then infected in the same season. Influenza between- and within-host genetic diversity was not significantly different in non-vaccinated and vaccinated humans, suggesting that vaccine-induced immunity does not exert a strong selective pressure on viruses replicating in individual people. Our results are consistent with a past report showing that prior immunity had little effect on the level and structure

of genetic diversity of influenza viruses infecting vaccinated horses and that purifying selection is dominant within individual hosts [154]. We found low frequency mutations, below the detection threshold of traditional surveillance methods, in non-vaccinated and vaccinated humans that were recently found to confer antibody escape. Interestingly, these potential antigenic variants did not reach fixation in infection, suggesting that other evolutionary factors may be hindering their emergence in individual humans. Determining the capacity of our putative escape variants to avoid antibody detection in vitro and in vivo is critical to understand how antigenic variants emerge in individual hosts. Our preliminary data identified 19 nonsynonymous (amino-acid-changing) substitutions occurring in HA putative antigenic sites and the receptor-binding domain in patients infected with H1N1pdm. Assessing whether these genetic variants encode phenotypic changes in antigenicity will be performed by hemagglutination-inhibition (HI) assays with ferret reference sera. Variant viruses that exhibit a phenotypic difference in antigenicity, defined by a ≥4-fold decrease in HI antibody titer with respect to the reference strain A/California/07/2009 (CA07), should be then evaluated for their capacity to grow in the presence of neutralizing antibodies from ferret reference sera and sera from vaccinated humans using plaque reduction assays. Antigenic variants that escape antibody responses without incurring a fitness cost should rapidly outcompete wild type viruses. We will compare the replicative fitness of variant and wild type viruses using competition assays in MDCK cell culture. Previously, mutations that alter HA antigenicity were found to cause a strong fitness cost that requires compensatory mutations to restore viral replicative capacity [155]. Therefore, we hypothesize that antigenic variants may require additional compensatory mutations to restore viral replicative fitness in humans. Importantly, these compensatory mutations must be acquired before the antigenic variant is purified from the virus population by natural selection. Determining the evolutionary processes that hinder the emergence of antigenic variants in humans may provide valuable insights on the molecular mechanisms that lead to

influenza re-emergence.

Through the completion of my graduate studies, I developed new experimental and analytical methodologies to rapidly characterize influenza quasispecies diversity directly from infected hosts. With strong collaborative research teams, I've adapted these sequence-based approaches to understand how other RNA viruses and parasites evolve in mammals; my minor contributions to published manuscripts not described as part of this thesis are outlined in Chapter 6, "Contributions to coauthored manuscripts." My thesis work uncovered underappreciated aspects influenza biology. Specifically, we were the first to describe influenza transmission bottlenecks and show that low-frequency variants can rapidly increase in frequency during acute infection to enhance viral replicative fitness. Building on our initial animal studies, we characterized H5N1 influenza viruses from human infections and detected viral variants with enhanced replicative capacity; interestingly, these variants did not increase in frequency during human infection or, at least to our knowledge, transmit between humans. Therefore, the generation of worrisome viruses in natures does not guarantee emergence, and other evolutionary processes may be hindering the emergence of pandemic viruses. Finally, this thesis serves as an early building block to better understand how within-host viral genetic diversity contributes to emergence and re-emergence of influenza viruses in nature.

Tables, Figures and Supplementals

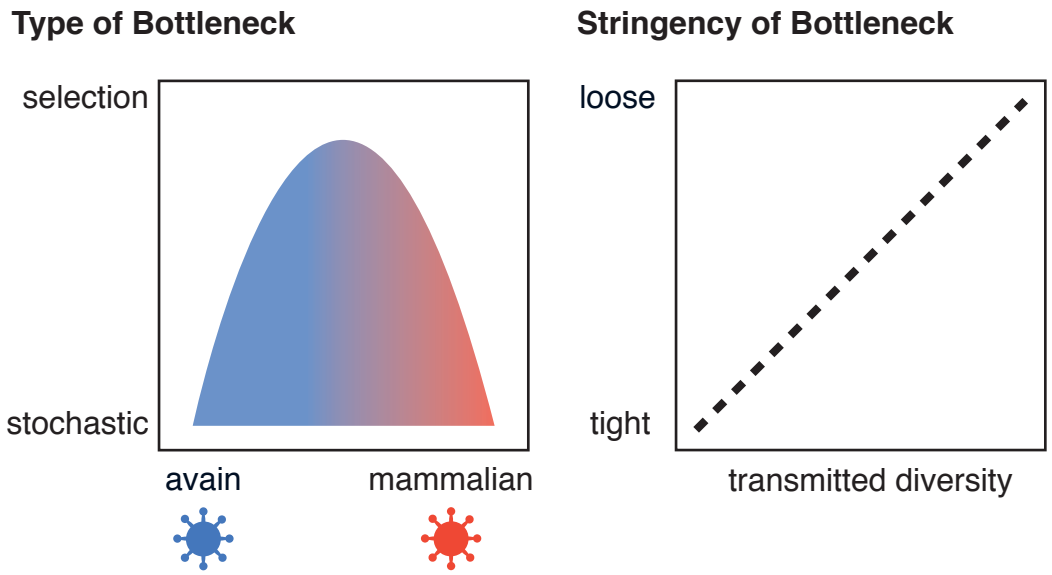


Figure 1. Conceptual overview. Influenza transmission bottlenecks are affected by both selection and stochastic factors that limit genetic diversity in recipient hosts. We predict that random processes dominate bottlenecks, but as avian influenza viruses evolve toward a “mammalian-like” phenotype, natural selection may favor the transmission of potentially pandemic viruses. Therefore, the type and stringency of transmission bottlenecks may impact influenza evolution and pandemic emergence.

Chapter 6

Appendix: Contributions to coauthored manuscripts

Influenza A Virus Polymerase Is a Site for Adaptive Changes during Experimental Evolution in Bat Cells.

Daniel S. Poole,[1] Shuǐqìng Yú,[2] Yíngyún Caì,[2] Jorge M. Dinis,[3] Marcel A. Müller,[4] Ingo Jordan,[5] Thomas C. Friedrich,[3,6] Jens H. Kuhn,[2] and Andrew Mehle[1].

[1]Department of Medical Microbiology and Immunology, University of Wisconsin—Madison, Madison, Wisconsin, USA. [2]NIH/NIAID Integrated Research Facility at Fort Detrick, Frederick, Maryland, USA. [3]Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin—Madison, Madison, Wisconsin, USA. [4]Institute of Virology, University of Bonn Medical Centre, Bonn, Germany. [5]ProBioGen AG, Berlin, Germany. [6]Wisconsin National Primate Research Center, Madison, Wisconsin, USA.

Journal of Virology, 88 (21) 12572-85, 2014 November.

Abstract: The recent identification of highly divergent influenza A viruses in bats revealed a new, geographically dispersed viral reservoir. To investigate the molecular mechanisms of host-restricted viral tropism and the potential for transmission of viruses between humans and bats, we exposed a panel of cell lines from bats of diverse species to a prototypical human-origin influenza A virus. All of the tested bat cell lines were susceptible to influenza A virus infection. Experimental evolution of human and avian-like viruses in bat cells resulted in efficient replication and created highly cytopathic variants. Deep sequencing of adapted human influenza A virus revealed a mutation in the PA polymerase subunit not previously described, M285K. Recombinant virus with the PA M285K mutation completely phenocopied the adapted virus. Adaptation of an avian virus-like virus resulted in the canonical PB2 E627K mutation that is required for efficient replication in other mammals. None of the adaptive mutations occurred in the gene for viral hemagglutinin, a gene that frequently acquires changes to recognize host-specific variations in sialic acid receptors. We showed that human influenza A virus uses canonical sialic acid receptors to infect bat cells, even though bat influenza A viruses do not appear to use these receptors for virus entry. Our results demonstrate that bats are unique hosts that select for both a novel mutation and a well-known adaptive mutation in the viral polymerase to support replication.

Key contributions: In this study, I applied deep sequencing to characterize human and avian-like viruses that were serially passaged in bat cells. My analysis of “deep” sequences revealed mutations that markedly changed in frequency during passage. Dr. Mehle’s group functionally characterized individual mutations and found that changes in human and avian polymerase complexes were required for efficient replication in bat cells.

High Genetic Diversity and Adaptive Potential of Two Simian Hemorrhagic Fever Viruses in a Wild Primate Population.

Adam L. Bailey,[1,2] Michael Lauck,[1,2] Andrea Weiler,[2,3] Samuel D. Sibley,[2,3] Jorge M. Dinis,[3] Zachary Bergman,[2,3] Chase W. Nelson,[4] Michael Correll,[5] Michael Gleicher,[5] David Hyeroba,[6] Alex Tumukunde,[6] Geoffrey Weny,[6] Colin Chapman,[6,7] Jens H. Kuhn,[8] Austin L. Hughes,[4] Thomas C. Friedrich,[2,3] Tony L. Goldberg,[2,3] and David H. O'Connor[1,2]

[1]Department of Pathology and Laboratory Medicine, University of Wisconsin–Madison, Madison, Wisconsin, United States of America. [2]Wisconsin National Primate Research Center, Madison, Wisconsin, United States of America. [3]Department of Pathobiological Sciences, University of Wisconsin–Madison, Madison, Wisconsin, United States of America. [4]Department of Biological Sciences, University of South Carolina, Columbia, South Carolina, United States of America. [5]Department of Computer Sciences, University of Wisconsin–Madison, Madison, Wisconsin, United States of America. [6]Makerere University, Kampala, Uganda. [7]Department of Anthropology and School of Environment, McGill University, Montreal, Quebec, Canada. [8]Integrated Research Facility at Fort Detrick, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Fort Detrick, Frederick, Maryland, United States of America.

PLoS ONE, 9(3) e90714, February 2014.

Abstract: Key biological properties such as high genetic diversity and high evolutionary rate enhance the potential of certain RNA viruses to adapt and emerge. Identifying viruses with these properties in their natural hosts could dramatically improve disease forecasting and surveillance. Recently, we discovered two novel members of the viral family Arteriviridae: simian hemorrhagic fever virus (SHFV)-krc1 and SHFV-krc2, infecting a single wild red colobus (*Procolobus rufomitratus tephrosceles*) in Kibale National Park, Uganda. Nearly nothing is known about the biological properties of SHFVs in nature, although the SHFV type strain, SHFV-LVR, has caused devastating outbreaks of viral hemorrhagic fever in captive macaques. Here we detected SHFV-krc1 and SHFV-krc2 in 40% and 47% of 60 wild red colobus tested, respectively. We found viral loads in excess of 106–107 RNA copies per milliliter of blood plasma for each of these viruses. SHFV-krc1 and SHFV-krc2 also showed high genetic diversity at both the inter- and intra-host levels. Analyses of synonymous and non-synonymous nucleotide diversity across viral genomes revealed patterns suggestive of positive selection in SHFV open reading frames (ORF) 5 (SHFV-krc2 only) and 7 (SHFV-krc1 and SHFV-krc2). Thus, these viruses share several important properties with some of the most rapidly evolving, emergent RNA viruses.

Key contributions: I computed and analyzed within-host genetic diversity from non-human primates infected with Simian Hemorrhagic Fever Viruses (SHFV). My contribution helped reveal that SHFV are highly diverse within individual primates and we speculate that this diversity may facilitate future cross-species transmission events.

A Novel Nonhuman Primate Model for Influenza Transmission.

Louise H. Moncla,[1,2,3] Ted M. Ross,[4] Jorge M. Dinis,[1,2,3] Jason T. Weinfurter,[1,2] Tatum D. Mortimer,[1,2,3] Nancy Schultz-Darken,[2] Kevin Brunner,[2] Saverio V. Capuano, III,[2] Carissa Boettcher,[2] Jennifer Post,[2] Michael Johnson,[2] Chalise E. Bloom,[5] Andrea M. Weiler,[2] and Thomas C. Friedrich[1,2,3]

[1]Department of Pathobiological Sciences, University of Wisconsin School of Veterinary Medicine, Madison, Wisconsin, United States of America. [2]Wisconsin National Primate Research Center, Madison, Wisconsin, United States of America. [3]University of Wisconsin Microbiology Doctoral Training Program, Madison, Wisconsin, United States of America. [4]Center for Vaccine Research, Dept. of Microbiology and Molecular Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America. [5]Vaccine and Gene Therapy Institute of Florida, Port St. Lucie, Florida, United States of America.
Johns Hopkins University - Bloomberg School of Public Health, United States of America.

PLoS One, 8(11) e78750, September 2013.

Abstract: Studies of influenza transmission are necessary to predict the pandemic potential of emerging influenza viruses. Currently, both ferrets and guinea pigs are used in such studies, but these species are distantly related to humans. Nonhuman primates (NHP) share a close phylogenetic relationship with humans and may provide an enhanced means to model the virological and immunological events in influenza virus transmission. Here, for the first time, it was demonstrated that a human influenza virus isolate can productively infect and be transmitted between common marmosets (*Callithrix jacchus*), a New World monkey species. We inoculated four marmosets with the 2009 pandemic virus A/California/07/2009 (H1N1pdm) and housed each together with a naïve cage mate. We collected bronchoalveolar lavage and nasal wash samples from all animals at regular intervals for three weeks post-inoculation to track virus replication and sequence evolution. The unadapted 2009 H1N1pdm virus replicated to high titers in all four index animals by 1 day post-infection. Infected animals seroconverted and presented human-like symptoms including sneezing, nasal discharge, labored breathing, and lung damage. Transmission occurred in one cohabitating pair. Deep sequencing detected relatively few genetic changes in H1N1pdm viruses replicating in any infected animal. Together our data suggest that human H1N1pdm viruses require little adaptation to replicate and cause disease in marmosets, and that these viruses can be transmitted between animals. Marmosets may therefore be a viable model for studying influenza virus transmission.

Key contributions: I optimized my deep sequencing approach to characterize influenza viruses from non-human primate respiratory samples (i.e., nasal swabs and bronchiolar lavages). With this improved protocol, I genetically characterized influenza quasispecies during infection and after transmission in marmosets. I wrote R scripts to identify signatures of natural selection and found that this human H1N1pdm isolate required little adaptation for replication in marmosets.

Deep sequencing identifies two genotypes and high viral genetic diversity of human pegivirus (GB virus C) in rural Ugandan patients.

Ria R. Ghai,[1] Samuel D. Sibley,[2] Michael Lauck,[3] Jorge M. Dinis,[2] Adam L. Bailey,[3] Colin A. Chapman,[4] Patrick Omeja,[5] Thomas C. Friedrich,[2,6] David H. O'Connor,[3,7] and Tony L. Goldberg[2,5,6]

[1]Department of Biology, McGill University, Montreal, QC, Canada. [2]Department of Pathobiological Sciences, University of Wisconsin–Madison, Madison, WI, USA. [3]Department of Pathology and Laboratory Medicine, University of Wisconsin–Madison, Madison, WI, USA. [4]Department of Anthropology and McGill School of Environment, Montreal, QC, Canada, and Wildlife Conservation Society, NY, USA. [5]Makerere University Biological Field Station, Fort Portal, Uganda. [6]Wisconsin National Primate Research Center, Madison, WI, USA. [7]School of Medicine and Public Health, University of Wisconsin–Madison, Madison, WI, USA.

Journal of General Virology, 94(12) 2670-2678, Sepember 2013.

Abstract: Human pegivirus (HPgV), formerly 'GB virus C' or 'hepatitis G virus', is a member of the genus Flavivirus (Flaviviridae) that has garnered significant attention due to its inhibition of HIV, including slowing disease progression and prolonging survival in HIV-infected patients. Currently, there are six proposed HPgV genotypes that have roughly distinct geographical distributions. Genotypes 2 and 3 are the most comprehensively characterized, whereas those genotypes occurring on the African continent, where HPgV prevalence is highest, are less well studied. Using deep sequencing methods, we identified complete coding HPgV sequences in four of 28 patients (14.3%) in rural Uganda, east Africa. One of these sequences corresponds to genotype 1 and is the first complete genome of this genotype from east Africa. The remaining three sequences correspond to genotype 5, a genotype that was previously considered exclusively South African. All four positive samples were collected within a geographical area of less than 25 km(2), showing that multiple HPgV genotypes co-circulate in this area. Analysis of intra-host viral genetic diversity revealed that total single-nucleotide polymorphism frequency was approximately tenfold lower in HPgV than in hepatitis C virus. Finally, one patient was co-infected with HPgV and HIV, which, in combination with the high prevalence of HIV, suggests that this region would be a useful locale to study the interactions and co-evolution of these viruses.

Key contribution: I was responsible for computing estimates of pegiviruses within-host genetic diversity from Ugandan patients.

Co-infection and cross-species transmission of divergent Hepatocystis lineages in a wild African primate community.

Mary I. Thurber,[1] Ria R. Ghai,[2] Hyeroba Hyeroba,[3] Geoffrey Weny,[3] Alex Tumukunde,[3] Colin A. Chapman,[3,4] Roger W. Wiseman,[5] Jorge Dinis,[5,6] James Steeil,[7] Ellis C. Greiner,[8] Thomas C. Friedrich,[1,5] David H. O'Connor,[5,6] and Tony L. Goldberg [1,3,5]

[1]Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison Wisconsin, USA 53706. [2]Department of Biology, McGill University, Montreal, Quebec, Canada, H3A 2T7. [3]Makerere University, P.O. Box 7062, Kampala, Uganda. [4]Department of Anthropology and School of Environment, McGill University, Montreal, Quebec, Canada, H3A 2T7. [5]Wisconsin National Primate Research Center, Madison Wisconsin, USA 53715. [6]Department of Pathology and Laboratory Medicine, School of Medicine and Public Health, University of Wisconsin-Madison, Madison Wisconsin, USA 53706. [7]University of Tennessee, College of Veterinary Medicine, Knoxville, Tennessee, USA 37996. [8]Department of Infectious Diseases and Pathology, College of Veterinary Medicine, University of Florida, Gainesville, Florida, USA 32611.

International Journal of Parasitology, 43(8) 613-9 2013, July 2013.

Abstract: Hemoparasites of the apicomplexan family Plasmodiidae include the etiological agents of malaria, as well as a suite of non-human primate parasites from which the human malaria agents evolved. Despite the significance of these parasites for global health, little information is available about their ecology in multi-host communities. Primates were investigated in Kibale National Park, Uganda, where ecological relationships among host species are well characterized. Blood samples were examined for parasites of the genera Plasmodium and Hepatocystis using microscopy and PCR targeting the parasite mitochondrial cytochrome b gene, followed by Sanger sequencing. To assess co-infection, "deep sequencing" of a variable region within cytochrome b was performed. Out of nine black-and-white colobus (*Colobus guereza*), one blue guenon (*Cercopithecus mitis*), five grey-cheeked mangabeys (*Lophocebus albigena*), 23 olive baboons (*Papio anubis*), 52 red colobus (*Procolobus rufomitratus*) and 12 red-tailed guenons (*Cercopithecus ascanius*), 79 infections (77.5%) were found, all of which were Hepatocystis spp. Sanger sequencing revealed 25 different parasite haplotypes that sorted phylogenetically into six species-specific but morphologically similar lineages. "Deep sequencing" revealed mixed-lineage co-infections in baboons and red colobus (41.7% and 64.7% of individuals, respectively) but not in other host species. One lineage infecting red colobus also infected baboons, but always as the minor variant, suggesting directional cross-species transmission. Hepatocystis parasites in this primate community are a diverse assemblage of cryptic lineages, some of which co-infect hosts and at least one of which can cross primate species barriers.

Key contribution: My role in this collaboration was to optimize de novo assembly and reference-based mappings to identify parasitic co-infections. My contribution helped identify parasitic co-infections in wild African primates.

Chapter 7

Bibliography

1. Kageyama, T. et al. Genetic analysis of novel avian A(H7N9) influenza viruses isolated from patients in China, February to April 2013. Euro Surveill 18, 20453 (2013).
2. Fraser, C. et al. Pandemic potential of a strain of influenza A (H1N1): early findings. Science 324, 1557-1561 (2009).
3. Richard, M. et al. Limited airborne transmission of H7N9 influenza A virus between ferrets. Nature 501, 560-563 (2013).
4. Belser, J. A. et al. Pathogenesis and transmission of avian influenza A (H7N9) virus in ferrets and mice. Nature 501, 556-559 (2013).
5. Watanabe, T. et al. Characterization of H7N9 influenza A viruses isolated from humans. Nature 501, 551-555 (2013).
6. Webster, R. G. Influenza: an emerging disease. Emerg Infect Dis 4, 436-441 (1998).
7. Johnson, N. P. & Mueller, J. Updating the accounts: global mortality of the 1918-1920 “Spanish” influenza pandemic. Bull Hist Med 76, 105-115 (2002).
8. Taubenberger, J. K. Initial Genetic Characterization of the 1918 “Spanish” Influenza Virus. Science 275, 1793-1796 (1997).
9. Smith, G. J. et al. Dating the emergence of pandemic influenza viruses. Proc Natl Acad Sci U S A 106, 11709-11712 (2009).
10. Krauss, S. et al. Influenza in migratory birds and evidence of limited intercontinental virus exchange. PLoS Pathog 3, e167 (2007).
11. Tong, S. et al. A distinct lineage of influenza A virus from bats. Proc Natl Acad Sci U S A 109, 4269-4274 (2012).
12. Schrauwen, E. J. & Fouchier, R. A. Host adaptation and transmission of influenza A viruses in mammals. Emerg Microbes Infect 3, e9 (2014).
13. Reperant, L. A., Kuiken, T. & Osterhaus, A. D. Adaptive pathways of zoonotic influenza viruses: from exposure to establishment in humans. Vaccine 30, 4419-4434 (2012).
14. Watanabe, T. et al. Characterization of H7N9 influenza A viruses isolated from humans. Nature 501, 551-555 (2013).
15. Wang, T. T., Parides, M. K. & Palese, P. Seroevidence for H5N1 influenza infections in humans: meta-analysis. Science 335, 1463 (2012).
16. Russell, C. A. et al. The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host. Science 336, 1541-1547 (2012).
17. Wilker, P. R. et al. Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. Nat Commun 4, 2636 (2013).
18. Osterholm, M. T., Kelley, N. S., Sommer, A. & Belongia, E. A. Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. Lancet Infect Dis 12, 36-44 (2012).
19. World Health Organization. Influenza. Fact sheet no. 211. [cited 2015 May 04] Available from: http://www.who.int/mediacentre/factsheets/fs211/en/
20. Stevens, J. et al. Receptor specificity of influenza A H3N2 viruses isolated in mammalian cells and embryonated chicken eggs. J Virol 84, 8287-8299 (2010).
21. Cox, R. J. Correlates of protection to influenza virus, where do we go from

here? vaccines 9, 405-408 (2013).
22. Skowronski, D. M. et al. Low 2012-13 influenza vaccine effectiveness associated with mutation in the egg-adapted H3N2 vaccine strain not antigenic drift in circulating viruses. PLoS One 9, e92153 (2014).
23. Smith, D. J. Mapping the Antigenic and Genetic Evolution of Influenza Virus. Science 305, 371-376 (2004).
24. Bedford, T. et al. Integrating influenza antigenic dynamics with molecular evolution. Elife 3, e01914 (2014).
25. Koel, B. F. et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science 342, 976-979 (2013).
26. Drake, J. W. & Holland, J. J. Mutation rates among RNA viruses. Proc Natl Acad Sci U S A 96, 13910-13913 (1999).
27. Sanjuan, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. J Virol 84, 9733-9748 (2010).
28. Schneider, W. L. & Roossinck, M. J. Genetic diversity in RNA virus quasispecies is controlled by host-virus interactions. J Virol 75, 6566-6571 (2001).
29. Holmes, E. C. The Evolutionary Genetics of Emerging Viruses. Annu. Rev. Ecol. Evol. Syst. 40, 353-372 (2009).
30. Parrish, C. R. et al. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol Mol Biol Rev 72, 457-470 (2008).
31. Simen, B. B. et al. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. J Infect Dis 199, 693-701 (2009).
32. Wilker, P. R. et al. Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. Nat Commun 4, 2636 (2013).
33. Breitbart, M. et al. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99, 14250-14255 (2002).
34. Bourret, V. et al. Whole-genome, deep pyrosequencing analysis of a duck influenza A virus evolution in swine cells. Infect Genet Evol 18, 31-41 (2013).
35. Croville, G. et al. Field monitoring of avian influenza viruses: whole-genome sequencing and tracking of neuraminidase evolution using 454 pyrosequencing. J Clin Microbiol 50, 2881-2887 (2012).
36. Ghedin, E. et al. Presence of oseltamivir-resistant pandemic A/H1N1 minor variants before drug therapy with subsequent selection and transmission. J Infect Dis 206, 1504-1511 (2012).
37. Ghedin, E. et al. Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. J Infect Dis 203, 168-174 (2011).
38. Kampmann, M. L. et al. A simple method for the parallel deep sequencing of full influenza A genomes. J Virol Methods 178, 243-248 (2011).
39. tale of three next generation sequencing platforms: comparison of Ion Torrent, P. B. A. I. M. S. Viral population analysis and minority-variant detection using short read next-generation sequencing. Philos Trans R Soc Lond B Biol Sci

368, 20120205 (2013).
40. Quail, M. A. et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13, 341 (2012).
41. Loman, N. J. et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30, 434-439 (2012).
42. Dean, F. B. et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A 99, 5261-5266 (2002).
43. Berthet, N. et al. Phi29 polymerase based random amplification of viral RNA as an alternative to random RT-PCR. BMC Mol Biol 9, 77 (2008).
44. Chan, C. H. et al. Amplification of the entire genome of influenza A virus H1N1 and H3N2 subtypes by reverse-transcription polymerase chain reaction. J Virol Methods 136, 38-43 (2006).
45. Hoffmann, E., Stech, J., Guan, Y., Webster, R. G. & Perez, D. R. Universal primer set for the full-length amplification of all influenza A viruses. Arch Virol 146, 2275-2289 (2001).
46. Stech, J. et al. Rapid and reliable universal cloning of influenza A virus genes by target-primed plasmid amplification. Nucleic Acids Res 36, e139 (2008).
47. Imai, M. et al. Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. Nature 486, 420-428 (2012).
48. Reperant, L. A., Kuiken, T. & Osterhaus, A. D. Influenza viruses: from birds to humans. Hum Vaccin Immunother 8, 7-16 (2012).
49. Neumann, G. & Kawaoka, Y. The first influenza pandemic of the new millennium. Influenza Other Respir Viruses 5, 157-166 (2011).
50. Watanabe, T. & Kawaoka, Y. Pathogenesis of the 1918 pandemic influenza virus. PLoS Pathog 7, e1001218 (2011).
51. Kobasa, D. & Kawaoka, Y. Emerging influenza viruses: past and present. Curr Mol Med 5, 791-803 (2005).
52. Ungchusak, K. et al. Probable person-to-person transmission of avian influenza A (H5N1). N Engl J Med 352, 333-340 (2005).
53. Zhang, Y. et al. Key molecular factors in hemagglutinin and PB2 contribute to efficient transmission of the 2009 H1N1 pandemic influenza virus. J Virol 86, 9666-9674 (2012).
54. Manz, B., Brunotte, L., Reuther, P. & Schwemmle, M. Adaptive mutations in NEP compensate for defective H5N1 RNA replication in cultured human cells. Nat Commun 3, 802 (2012).
55. Ilyushina, N. A., Bovin, N. V. & Webster, R. G. Decreased neuraminidase activity is important for the adaptation of H5N1 influenza virus to human airway epithelium. J Virol 86, 4724-4733 (2012).
56. Mehle, A., Dugan, V. G., Taubenberger, J. K. & Doudna, J. A. Reassortment and mutation of the avian influenza virus polymerase PA subunit overcome species barriers. J Virol 86, 1750-1757 (2012).
57. Sakabe, S., Ozawa, M., Takano, R., Iwastuki-Horimoto, K. & Kawaoka, Y. Mutations in PA, NP, and HA of a pandemic (H1N1) 2009 influenza virus contribute to

its adaptation to mice. Virus Res 158, 124-129 (2011).
58. Hatta, M., Gao, P., Halfmann, P. & Kawaoka, Y. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. Science 293, 1840-1842 (2001).
59. Imai, M. & Kawaoka, Y. The role of receptor binding specificity in interspecies transmission of influenza viruses. Curr Opin Virol 2, 160-167 (2012).
60. Rogers, G. N. & Paulson, J. C. Receptor determinants of human and animal influenza virus isolates: differences in receptor specificity of the H3 hemagglutinin based on species of origin. Virology 127, 361-373 (1983).
61. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3, 418-426 (1986).
62. Yamada, S. et al. Biological and structural characterization of a host-adapting amino acid in influenza virus. PLoS Pathog 6, e1001034 (2010).
63. Zhou, B. et al. PB2 residue 158 is a pathogenic determinant of pandemic H1N1 and H5 influenza a viruses in mice. J Virol 85, 357-365 (2011).
64. Nielsen, R. Molecular signatures of natural selection. Annu Rev Genet 39, 197-218 (2005).
65. Herfst, S. et al. Airborne transmission of influenza A/H5N1 virus between ferrets. Science 336, 1534-1541 (2012).
66. Horimoto, T. & Kawaoka, Y. Pandemic threat posed by avian influenza A viruses. Clin Microbiol Rev 14, 129-149 (2001).
67. Van Hoeven, N. et al. Human HA and polymerase subunit PB2 proteins confer transmission of an avian influenza virus through the air. Proc Natl Acad Sci U S A 106, 3366-3371 (2009).
68. Van Kerkhove, M. D. et al. Highly pathogenic avian influenza (H5N1): pathways of exposure at the animal-human interface, a systematic review. PLoS One 6, e14582 (2011).
69. Keele, B. F. et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A 105, 7552-7557 (2008).
70. Keele, B. F. et al. Low-dose rectal inoculation of rhesus macaques by SIVsmE660 or SIVmac251 recapitulates human mucosal infection by HIV-1. J Exp Med 206, 1117-1134 (2009).
71. Keele, B. F. & Estes, J. D. Barriers to mucosal transmission of immunodeficiency viruses. Blood 118, 839-846 (2011).
72. Kundu, S. et al. Tracking viral evolution during a disease outbreak: the rapid and complete selective sweep of a circovirus in the endangered Echo parakeet. J Virol 86, 5221-5229 (2012).
73. Bull, R. A. et al. Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. PLoS Pathog 7, e1002243 (2011).
74. Gustin, K. M. et al. Influenza virus aerosol exposure and analytical system for ferrets. Proc Natl Acad Sci U S A 108, 8432-8437 (2011).
75. Bussey, K. A., Bousse, T. L., Desmet, E. A., Kim, B. & Takimoto, T. PB2 residue 271 plays a key role in enhanced polymerase activity of influenza A viruses in mammalian host cells. J Virol 84, 4395-4406 (2010).

76. Shinya, K. et al. PB2 amino acid at position 627 affects replicative efficiency, but not cell tropism, of Hong Kong H5N1 influenza A viruses in mice. Virology 320, 258-266 (2004).
77. Hatta, M. et al. Growth of H5N1 influenza A viruses in the upper respiratory tracts of mice. PLoS Pathog 3, 1374-1379 (2007).
78. Domingo, E. & Holland, J. J. RNA virus mutations and fitness for survival. Annu Rev Microbiol 51, 151-178 (1997).
79. Moya, A., Holmes, E. C. & Gonzalez-Candelas, F. The population genetics and evolutionary epidemiology of RNA viruses. Nat Rev Microbiol 2, 279-288 (2004).
80. Nguyen, T. H., Farrar, J. & Horby, P. Person-to-person transmission of influenza A (H5N1). Lancet 371, 1392-1394 (2008).
81. Wang, H. et al. Probable limited person-to-person transmission of highly pathogenic avian influenza A (H5N1) virus in China. Lancet 371, 1427-1434 (2008).
82. Linster, M. et al. Identification, Characterization, and Natural Selection of Mutations Driving Airborne Transmission of A/H5N1 Virus. Cell 157, 329-339 (2014).
83. Connor, R. J., Kawaoka, Y., Webster, R. G. & Paulson, J. C. Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates. Virology 205, 17-23 (1994).
84. Matrosovich, M. et al. Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. J Virol 74, 8502-8512 (2000).
85. Russell, C. A. et al. The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host. Science 336, 1541-1547 (2012).
86. Reperant, L. A., Kuiken, T., Grenfell, B. T. & Osterhaus, A. D. The immune response and within-host emergence of pandemic influenza virus. Lancet 384, 2077-2081 (2014).
87. Beigel, J. H. et al. Avian influenza A (H5N1) infection in humans. N Engl J Med 353, 1374-1385 (2005).
88. Neumann, G. & Kawaoka, Y. Genetic engineering of influenza and other negative-strand RNA viruses containing segmented genomes. Adv Virus Res 53, 265-300 (1999).
89. Imai, H. et al. The HA and NS genes of human H5N1 influenza A virus contribute to high virulence in ferrets. PLoS Pathog 6, e1001106 (2010).
90. Niwa, H., Yamamura, K. & Miyazaki, J. Efficient selection for high-expression transfectants with a novel eukaryotic vector. Gene 108, 193-199 (1991).
91. Bale, S. et al. Marburg virus VP35 can both fully coat the backbone and cap the ends of dsRNA for interferon antagonism. PLoS Pathog 8, e1002916 (2012).
92. Yamayoshi, S. et al. Ebola virus matrix protein VP40 uses the COPII transport system for its intracellular transport. Cell Host Microbe 3, 168-177 (2008).
93. Ito, T. et al. Differences in sialic acid-galactose linkages in the chicken egg amnion and allantois influence human influenza virus receptor specificity and variant selection. J Virol 71, 3357-3362 (1997).
94. Crusat, M. et al. Changes in the hemagglutinin of H5N1 viruses during human infection--influence on receptor binding. Virology 447, 326-337 (2013).

95. Varble, A. et al. Influenza A virus transmission bottlenecks are defined by infection route and recipient host. Cell Host Microbe 16, 691-700 (2014).
96. Shinya, K. & Kawaoka, Y. [Influenza virus receptors in the human airway]. Uirusu 56, 85-89 (2006).
97. Thomachot, L. et al. Measurement of tracheal temperature is not a reliable index of total respiratory heat loss in mechanically ventilated patients. Crit Care 5, 24-30 (2001).
98. Huai, Y. et al. Incubation period for human cases of avian influenza A (H5N1) infection, China. Emerg Infect Dis 14(11), 1819-1821 (2008).
99. HENLE, W. & LIU, O. C. Studies on host-virus interactions in the chick embryo-influenza virus system. VI. Evidence for multiplicity reactivation of inactivated virus. J Exp Med 94, 305-322 (1951).
100. Abdoli, A., Soleimanjahi, H., Tavassoti Kheiri, M., Jamali, A. & Jamaati, A. Determining influenza virus shedding at different time points in madin-darby canine kidney cell line. Cell J 15, 130-135 (2013).
101. Steel, J., Lowen, A. C., Mubareka, S. & Palese, P. Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N. PLoS Pathog 5, e1000252 (2009).
102. Gabriel, G., Czudai-Matwich, V. & Klenk, H. D. Adaptive mutations in the H5N1 polymerase complex. Virus Res 178, 53-62 (2013).
103. Xu, J., Zhong, H. A., Madrahimov, A., Helikar, T. & Lu, G. Molecular phylogeny and evolutionary dynamics of influenza A nonstructural (NS) gene. Infect Genet Evol 22, 192-200 (2014).
104. Murcia, P. R. et al. Intra- and interhost evolutionary dynamics of equine influenza virus. J Virol 84, 6943-6954 (2010).
105. Walther, T. et al. Glycomic analysis of human respiratory tract tissues and correlation with influenza virus infection. PLoS Pathog 9, e1003223 (2013).
106. Krenn, B. M. et al. Single HA2 mutation increases the infectivity and immunogenicity of a live attenuated H5N1 intranasal influenza vaccine candidate lacking NS1. PLoS One 6, e18577 (2011).
107. Cui, L. et al. Dynamic reassortments and genetic heterogeneity of the human-infecting influenza A (H7N9) virus. Nat Commun 5, 3142 (2014).
108. Carr, C. M., Chaudhry, C. & Kim, P. S. Influenza hemagglutinin is spring-loaded by a metastable native conformation. Proceedings of the National Academy of Sciences 94, 14306-14313 (1997).
109. Subbarao, E. K., London, W. & Murphy, B. R. A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. Journal of virology 67, 1761-1764 (1993).
110. Clements, M. L. et al. Use of single-gene reassortant viruses to study the role of avian influenza A virus genes in attenuation of wild-type human influenza A virus for squirrel monkeys and adult human volunteers. Journal of clinical microbiology 30, 655-662 (1992).
111. Gabriel, G., Herwig, A. & Klenk, H. D. Interaction of polymerase subunit PB2 and NP with importin alpha1 is a determinant of host range of influenza A virus. PLoS Pathog 4, e11 (2008).

112. Hale, B. G., Randall, R. E., Ortin, J. & Jackson, D. The multifunctional NS1 protein of influenza A viruses. J Gen Virol 89, 2359-2376 (2008).
113. Peiris, J. S. M., Guan, Y. & Yuen, K. Y. Severe acute respiratory syndrome. Nature medicine 10, S88-S97 (2004).
114. Seo, S. H., Hoffmann, E. & Webster, R. G. Lethal H5N1 influenza viruses escape host anti-viral cytokine responses. Nature medicine 8, 950-954 (2002).
115. Finkelstein, D. B. et al. Persistent host markers in pandemic and H5N1 influenza viruses. J Virol 81, 10292-10299 (2007).
116. Domingo, E., Sheldon, J. & Perales, C. Viral quasispecies evolution. Microbiol Mol Biol Rev 76, 159-216 (2012).
117. Pastore, C. et al. Human immunodeficiency virus type 1 coreceptor switching: V1/V2 gain-of-fitness mutations compensate for V3 loss-of-fitness mutations. Journal of virology 80, 750-758 (2006).
118. Pybus, O. G. et al. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. Mol Biol Evol 24, 845-852 (2007).
119. McLean, H. Q. et al. Influenza Vaccine Effectiveness in the United States During 2012-2013: Variable Protection by Age and Virus Type. J Infect Dis (2014).
120. Flannery, B. et al. Interim estimates of 2013-14 seasonal influenza vaccine effectiveness - United States, February 2014. MMWR Morb Mortal Wkly Rep 63, 137-142 (2014).
121. Flannery, B. et al. Early estimates of seasonal influenza vaccine effectiveness - United States, January 2015. MMWR Morb Mortal Wkly Rep 64, 10-15 (2015).
122. CDC. 2014-2015 Influenza Season Week 53 ending January 3, 2015. Fluview
123. Hensley, S. E. et al. Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. Science 326, 734-736 (2009).
124. Kryazhimskiy, S., Dushoff, J., Bazykin, G. A. & Plotkin, J. B. Prevalence of epistasis in the evolution of influenza A surface proteins. PLoS Genet 7, e1001301 (2011).
125. Wiley, D. C., Wilson, I. A. & Skehel, J. J. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature 289, 373-378 (1981).
126. Caton, A. J., Brownlee, G. G., Yewdell, J. W. & Gerhard, W. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell 31, 417-427 (1982).
127. Wilson, I. A. & Cox, N. J. Structural basis of immune recognition of influenza virus hemagglutinin. Annu Rev Immunol 8, 737-771 (1990).
128. Lauring, A. S., Frydman, J. & Andino, R. The role of mutational robustness in RNA virus evolution. Nat Rev Microbiol 11, 327-336 (2013).
129. Greenlee, R. T. Measuring disease frequency in the Marshfield Epidemiologic Study Area (MESA). Clin Med Res 1, 273-280 (2003).
130. Ohmit, S. E. et al. Influenza vaccine effectiveness in the 2011-2012 season: protection against each circulating virus and the effect of prior vaccination on estimates. Clin Infect Dis 58, 319-327 (2014).
131. Early estimates of seasonal influenza vaccine effectiveness--United States, January 2013. MMWR Morb Mortal Wkly Rep 62, 32-35 (2013).

132. Skehel, J. J. & Hay, A. J. Nucleotide sequences at the 5' termini of influenza virus RNAs and their transcripts. Nucleic Acids Res 5, 1207-1219 (1978).
133. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113 (2004).
134. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59, 307-321 (2010).
135. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131, 479-491 (1992).
136. Paradis, E. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26, 419-420 (2010).
137. Kofler, R. et al. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 6, e15925 (2011).
138. Wang, M. L., Katz, J. M. & Webster, R. G. Extensive heterogeneity in the hemagglutinin of egg-grown influenza viruses from different patients. Virology 171, 275-279 (1989).
139. Robertson, J. S. et al. Structural changes in the haemagglutinin which accompany egg adaptation of an influenza A(H1N1) virus. Virology 160, 31-37 (1987).
140. Schild, G. C., Oxford, J. S., de Jong, J. C. & Webster, R. G. Evidence for host-cell selection of influenza virus antigenic variants. Nature 303, 706-709 (1983).
141. European Centre for Disease Prevention and Control. Influenza virus characterisation, S. E., July 2013. Stockholm: ECDC; 2013.
142. European Centre for Disease Prevention and Control. Influenza virus characterisation, S. E., July 2014. Stockholm: ECDC; 2014.
143. Zaraket, H., Bridges, O. A. & Russell, C. J. The pH of activation of the hemagglutinin protein regulates H5N1 influenza virus replication and pathogenesis in mice. J Virol 87, 4826-4834 (2013).
144. Bhatt, S., Holmes, E. C. & Pybus, O. G. The genomic rate of molecular adaptation of the human influenza A virus. Mol Biol Evol 28, 2443-2451 (2011).
145. Koel, B. F. et al. Identification of amino acid substitutions supporting antigenic change of A(H1N1)pdm09 viruses. J Virol (2015).
146. Jin, H. et al. Two residues in the hemagglutinin of A/Fujian/411/02-like influenza viruses are responsible for antigenic drift from A/Panama/2007/99. Virology 336, 113-119 (2005).
147. Barr, I. G. et al. WHO recommendations for the viruses used in the 2013-2014 Northern Hemisphere influenza vaccine: Epidemiology, antigenic and genetic characteristics of influenza A(H1N1)pdm09, A(H3N2) and B influenza viruses collected from October 2012 to January 2013. Vaccine 32, 4713-4725 (2014).
148. Squires, R. B. et al. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza Other Respir Viruses 6, 404-416 (2012).
149. Li, Y. et al. Immune history shapes specificity of pandemic H1N1 influenza antibody responses. J Exp Med 210, 1493-1500 (2013).

150. Guarnaccia, T. et al. Antigenic drift of the pandemic 2009 A(H1N1) influenza virus in A ferret model. PLoS Pathog 9, e1003354 (2013).
151. Miller, M. S. et al. Neutralizing antibodies against previously encountered influenza virus strains increase over time: a longitudinal analysis. Sci Transl Med 5, 198ra107 (2013).
152. Fazekas, D. S. G. & Webster, R. G. Disquisitions of Original Antigenic Sin. I. Evidence in man. J Exp Med 124, 331-345 (1966).
153. Fonville, J. M. et al. Antibody landscapes after influenza virus infection or vaccination. Science 346, 996-1000 (2014).
154. Murcia, P. R. et al. Evolution of equine influenza virus in vaccinated horses. J Virol 87, 4768-4771 (2013).
155. Das, S. R. et al. Defining influenza A virus hemagglutinin antigenic drift by sequential monoclonal antibody selection. Cell Host Microbe 13, 314-323 (2013).
156. Illingworth, C. J., Fischer, A. & Mustonen, V. Identifying selection in the within-host evolution of influenza using viral sequence data. PLoS Comput Biol 10, e1003755 (2014).
157. Recommended composition of influenza virus vaccines for use in the 2014-2015 northern hemisphere influenza season. Wkly Epidemiol Rec 89, 93-104 (2014).
158. O'Donnell, C. D. et al. Antibody pressure by a human monoclonal antibody targeting the 2009 pandemic H1N1 virus hemagglutinin drives the emergence of a virus with increased virulence in mice. MBio 3, (2012).
159. Ampofo, W. K. et al. Improving influenza vaccine virus selection: report of a WHO informal consultation held at WHO headquarters, Geneva, Switzerland, 14-16 June 2010. Influenza Other Respir Viruses 6, 142-52, e1 (2012).
160. Russell, C. A. et al. The global circulation of seasonal influenza A (H3N2) viruses. Science 320, 340-346 (2008).
161. Weis, W. I., Brunger, A. T., Skehel, J. J. & Wiley, D. C. Refinement of the influenza virus hemagglutinin by simulated annealing. J Mol Biol 212, 737-761 (1990).
162. Kampmann, M. L. et al. A simple method for the parallel deep sequencing of full influenza A genomes. J Virol Methods 178, 243-248 (2011).
163. Brankston, G., Gitterman, L., Hirji, Z., Lemieux, C. & Gardam, M. Transmission of influenza A in human beings. Lancet Infect Dis 7, 257-265 (2007).
164. Tellier, R. Aerosol transmission of influenza A virus: a review of new studies. J R Soc Interface 6 Suppl 6, S783-S790 (2009).
165. Casadevall, A. & Shenk, T. The H5N1 moratorium controversy and debate. MBio 3, e00379-e00312 (2012).
166. Duprex, W. P., Fouchier, R. A., Imperiale, M. J., Lipsitch, M. & Relman, D. A. Gain-of-function experiments: time for a real debate. Nat Rev Microbiol 13, 58-64 (2015).
167. Lipsitch, M. & Galvani, A. P. Ethical alternatives to experiments with novel potential pandemic pathogens. PLoS Med 11, e1001646 (2014).