The evolutionary dynamics of influenza A virus within and between human hosts

Running Title: Influenza dynamics within and between hosts

John T. McCrone¹, Robert J. Woods², Emily T. Martin³, Ryan E. Malosh³, Arnold S. Monto³, and Adam S. Lauring¹,² *

¹ Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109
² Division of Infectious Diseases, Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109
³ Department of Epidemiology, University of Michigan, Ann Arbor, MI 48109

* Corresponding author

Adam S. Lauring
1150 W. Medical Center Dr.
MSRB1 Room 5510B
Ann Arbor, MI 48109-5680
(734) 764-7731
alaurin@med.umich.edu

Abstract Word Count:

Main Text Word Count:

Key Words – influenza virus, bottleneck, transmission, diversity, evolution
Abstract

A complete understanding of influenza virus evolution requires studies at all levels, as viral evolutionary dynamics may differ across spatial and temporal scales. The relative contribution of deterministic processes, such as selection, and stochastic processes, such as genetic drift, is influenced by the virus' effective population size. While the global evolution of influenza A virus (IAV) is dominated by the positive selection of novel antigenic variants that circulate in the tropics, much less is known about the virus' evolution within and between human hosts. With few exceptions, most of the available data derive from studies of chronically infected, immunocompromised hosts, experimental infections with attenuated viruses, or animal models. Here we define the evolutionary dynamics of IAV in human hosts through next generation sequencing of 249 upper respiratory specimens from 200 individuals collected over 6290 person-seasons of observation. Because these viruses were collected over 5 seasons from individuals in a prospective community-based cohort, they are broadly representative of natural human infections with seasonal viruses. Within host genetic diversity was low, and we found little evidence for positive selection of minority variants. We used viral sequence data from 35 serially sampled individuals to estimate a within host effective population size of 30-50. This estimate is consistent across several models and robust to the models' underlying assumptions. We also identified 43 epidemiologically linked and genetically validated transmission pairs. Maximum likelihood optimization of multiple transmission models estimates an effective transmission bottleneck of 1-2 distinct genomes. Our data suggest that positive selection of novel viral variants is inefficient at the level of the individual host and that genetic drift and other stochastic processes dominate the within and between host evolution of influenza A viruses.

Introduction

The rapid evolution of influenza viruses has led to reduced vaccine efficacy, widespread drug resistance, and the continuing emergence of novel strains. Broadly speaking, evolution is the
product of deterministic processes, such as selection, and stochastic processes, such as genetic drift (1). The relative contribution of each is greatly affected by the effective population size, or size of an idealized population whose dynamics are similar to that of the population in question (2). If the effective population size of a virus is large, as in quasispecies models, evolution is largely deterministic and the frequency of a mutation can be predicted based on its starting frequency and selection coefficient. In small populations, selection is inefficient, and changes in mutation frequency are strongly influenced by genetic drift.

Viral dynamics may differ across spatial and temporal scales, and a complete understanding of influenza evolution requires studies at all levels (3, 4). The global evolution of influenza A virus (IAV) is dominated by the positive selection of novel antigenic variants that circulate in the tropics and subsequently seed annual epidemics in the Northern and Southern hemisphere (5). Whole genome sequencing has also demonstrated the importance of intrasubtype reassortment to the emergence of diverse strains that differ in their antigenicity. While continual positive selection of antigenically drifted variants drives global patterns, whole genome sequencing of viruses on more local scales suggests the importance of stochastic processes such as strain migration and within-clade reassortment (6).

With the advent of next generation sequencing, it is now feasible to efficiently sequence patient-derived isolates at sufficient depth of coverage to define the diversity and dynamics of virus evolution within individual hosts (7). Studies of IAV populations in animal and human systems suggest that most intrahost single nucleotide variants (iSNV) are rare and that intrahost populations are subject to strong purifying selection (8-14). While positive selection of adaptive variants is commonly observed in cell culture (15-17), it has only been documented within human hosts in the extreme cases of drug resistance (8, 18, 19), long-term infection of immunocompromised hosts (20) or experimental infections with attenuated viruses (21). Indeed,
we and others have been unable to identify evidence for positive selection in natural human infections (13, 14), and its relevance to within host processes is unclear.

Despite limited evidence for positive selection, it is clear that novel mutations do arise within hosts. Their potential for subsequent spread through host populations is determined by the size of the transmission bottleneck (22, 23). If the transmission bottleneck is sufficiently wide, low frequency variants can plausibly be transmitted and spread through host populations (24). Because the transmission bottleneck is conceptually similar to the effective population size between hosts, its size will also inform the relative importance of selection and genetic drift in determining which variants are transmitted. While experimental infections of guinea pigs and ferrets suggest a very narrow transmission bottleneck (25, 26), studies of equine influenza support a bottleneck wide enough to allow transmission of rare iSNV (9, 27). The only available genetic study of influenza virus transmission in humans estimated a remarkably large transmission bottleneck, allowing for transmission of 100-200 genomes (11, 28).

Here, we use next generation sequencing of within host influenza virus populations to define the evolutionary dynamics of influenza A viruses (IAV) within and between human hosts. We apply a benchmarked analysis pipeline to identify iSNV and to characterize the genetic diversity of H3N2 and H1N1 populations collected over five post-pandemic seasons from individuals enrolled in a prospective household study of influenza. We use these data to estimate the in vivo mutation rate and the within and between host effective population size. We find that intrahost populations are characterized by purifying selection, a small effective population size, and limited positive selection. Contrary to what has been previously reported for human influenza transmission (11), but consistent with what has been observed in other viruses (23), we identify a very tight transmission bottleneck that limits the transmission of rare variants.
Results

We used next generation sequencing to characterize influenza virus populations collected from individuals enrolled in the Household Influenza Vaccine Effectiveness (HIVE) study (29-32), a community-based cohort that enrolls 213-340 households of 3 or more individuals in Southeastern Michigan each year (Table 1). These households are followed prospectively from October to April, with symptom-triggered collection of nasal and throat swab specimens for identification of respiratory viruses by RT-PCR (see Methods). In contrast to case-ascertained studies, which identify households based on an index case who seeks medical care, the HIVE study identifies individuals regardless of illness severity. In the first four seasons of the study (2010-2011 through 2013-2014), respiratory specimens were collected 0-7 days after illness onset. Beginning in the 2014-2015 season, each individual provided two samples, a self-collected specimen at the time of symptom onset and a clinic-collected specimen obtained 0-7 days later. Each year, 59-69% of individuals had self-reported or confirmed receipt of that season’s vaccine prior to local circulation of influenza virus.

Over five seasons and nearly 6,290 person-seasons of observation, we identified 77 cases of influenza A/H1N1pdm09 infection and 313 cases of influenza A/H3N2 infection (Table 1). Approximately half of the cases (n=166) were identified in the 2014-2015 season, in which there was an antigenic mismatch between the vaccine and circulating strains (33). All other seasons were antigenically matched. Individuals within a household were considered an epidemiologically linked transmission pair if they were both positive for the same subtype of influenza virus within 7 days of each other. Several households had 3 or 4 symptomatic cases within this one-week window, suggestive of possible transmission chains (Table 1).

Within host populations have low genetic diversity
We processed all specimens for viral load quantification and next generation sequencing. Viral load measurements (genome copies per µl) were used for quality control in variant calling, which we have shown is highly sensitive to input titer (34) (Figure 1A). Accordingly, we report data on 249 high quality specimens from 200 individuals, which had a viral load of $>10^3$ copies per microliter of transport media, adequate RT-PCR amplification of all eight genomic segments, and an average read coverage of $>10^3$ across the genome (Table 1, Supplemental Figure 1).

We identified intrahost single nucleotide variants (iSNV) using our empirically validated analysis pipeline (34). Our approach relies heavily on the variant caller DeepSNV, which uses a clonal plasmid control to distinguish between true iSNV and errors introduced during sample preparation and/or sequencing (35). Given the diversity of influenza viruses that circulate locally each season, there were a number of instances in which our patient-derived samples had mutations that were essentially fixed ($>0.95$ frequency) relative to the clonal control. DeepSNV is unable to estimate an error rate for the control or reference base at these positions. We therefore performed an additional benchmarking experiment to identify a threshold for majority iSNV at which we could correctly infer whether or not the corresponding minor allele was also present (see Methods). We found that we could correctly identify a minor allele at a frequency of $\geq 2\%$ when the frequency of the major allele was $\leq 98\%$. We therefore report data on iSNV present at frequencies between 2 and 98%. As expected, this threshold improved the specificity of our iSNV identification and decreased our sensitivity to detect variants below 5% compared to our initial validation experiment (34), which did not employ a frequency threshold (Supplemental Table 1).

Consistent with our previous studies and those of others, we found that the within host diversity of human influenza A virus (IAV) populations is low (11, 13, 14, 21, 34). Two hundred forty-three out of the 249 samples had fewer than 10 minority iSNV (median 2, IQR 1-3). There were 6
samples with greater than 10 minority iSNV. In 3 of these cases, the frequency of iSNVs were tightly distributed about a mean suggesting that the iSNV were linked and that the samples represented mixed infections. Consistent with this hypothesis, putative genomic haplotypes based on these minority iSNV clustered with distinct isolates on phylogenetic trees (Supplemental Figures 2 and 3). While viral shedding was well correlated with days post symptom onset (Figure 1A) the number of minority iSNV identified was not affected by the day of infection, viral load, subtype, or vaccination status (Figure 1B and Supplemental Figure 4).

The vast majority of minority variants were rare (frequency 0.02-0.07), and iSNV were distributed evenly across the genome (Figure 1C and 1D). The ratio of nonsynonymous to synonymous variants was 0.64 and was never greater than 1 in any 5% bin, which suggests that within host populations were under purifying selection. We also found that minority variants were rarely shared among multiple individuals. Ninety-five percent of minority iSNV were only found once, 4.7% were found in 2 individuals, and no minority iSNV were found in more than 3 individuals. The low level of shared diversity suggests that within host populations were exploring distinct regions of sequence space with little evidence for parallel evolution. Of the 31 minority iSNV that were found in multiple individuals (triangles in Figure 1D), 4 were nonsynonymous.

Although the full range of the H3 antigenic sites have not been functionally defined, it is estimated that 131 of the 329 amino acids in HA1 lie in or near these sites (36). We identified 17 minority nonsynonymous iSNV in these regions (Supplemental Table 2). Six of these were in positions that differ among antigenically drifted viruses (37, 38), and two (193S and 189N) lie in the “antigenic ridge” that is a major contributor to drift (39). Three of these have been detected at the global level as consensus variants since the time of isolation (128A, 193S and 262N) with two (193S and 262N) seemingly increasing in global frequency (40) (Supplemental Figure 5).
Additionally, we identified 1 putative H1N1 antigenic variant (208K in Cα) (41, 42). In total, putative antigenic variants account for 1.0-2.5% of minority iSNV identified and were found in 3.5-7.5% of infections. None of these iSNV were shared among multiple individuals.

Estimation of effective population size

Given the above observations, we hypothesized that within host populations of IAV are under purifying selection and that variants that rise to detectable levels do so by a neutral process as opposed to positive selection. Consistent with this hypothesis, we found that nonsynonymous and synonymous iSNV exhibited similar changes in frequency over time in the 35 individuals who provided serial specimens that contained iSNV (Figure 2A and 2B). We used the diffusion approximation to the Wright-Fisher model in conjunction with maximum likelihood estimation to determine the within host effective population size (Nₑ) of IAV (43). This model assumes that changes in iSNV frequency are due solely to random genetic drift and not selection, that iSNV are independent of one another, and that the effective population is sufficiently large to justify a continuous approximation to changes in allele frequency. While it is impossible to predict with certainty the trajectory of allele frequencies under random genetic drift, the Wright-Fisher model, and the diffusion approximation in particular, assigns probabilities to frequency changes given an Nₑ and the number of generations between sample times. In our model we fixed the within host generation time as either 6 or 12 hours (24) and report the findings for the 6 hour generation time below. We then asked what population size makes the observed changes in frequency most likely (Figure 2B). We restricted this analysis to samples taken at least 1 day apart (n = 29), as there was very little change in iSNV frequency in populations sampled twice on the same day (R = 0.990, Figure 2B and Supplemental Figure 6). The concordance of same day samples suggests that our sampling procedure is reproducible and that less than a generation had passed between samplings. Maximum likelihood optimization of this diffusion model revealed a within host effective population size of 35 (95% CI 26-46, Table 2).
The diffusion approximation makes several simplifying assumptions, which if violated could influence our findings. In particular, the model assumes a large population. To ensure our results were robust to this assumption, we employed a discrete interpretation of the Wright-Fisher model which makes no assumptions about population size (44). In this case we found an effective population size of 32 (95% CI 28-41), very close to our original estimate (Table 2).

Both models assume complete independence of iSNV. To ensure this assumption did not affect our results, we fit the discrete model 1000 times, each time randomly subsetting our data such that only one iSNV per individual was included. This simulates a situation in which all modeled iSNV are independent and our assumption is met. Under these conditions we found a median effective population size of 33 (IQR 32-40), demonstrating negligible bias in the initial analysis due to correlation between iSNV.

As above, most iSNV in the longitudinal samples were rare (< 10%) and many became extinct between samplings. To ensure that our models were capable of accurately estimating the effective population size from such data, we simulated 1000 Wright-Fisher populations with iSNV present at approximately the same starting frequencies as in our data set an N_e of 30, 50, or 100. In these simulations, we found mean N_e of 34, 56 and 117 (Figure 2C), which suggests that our estimate is not an artifact of the underlying data structure.

To this point, we have assumed that neutral processes are responsible for the observed changes in iSNV frequency within hosts. Although this assumption seems justified at least in part by the analysis above, we tested the robustness of our models by fitting the nonsynonymous (n = 27) and synonymous iSNV (n = 36) separately. Here, we estimated an effective population size of 30 using the nonsynonymous iSNV and an effective population size of 37 using the synonymous iSNV (Table 2). These estimates are very close to that derived
from the whole dataset and suggest that nonsynonymous and synonymous mutations are
influenced by similar within host processes. To further ensure that our results were not driven by
a few outliers subject to strong selection, we ranked iSNV by their change in frequency over
time and consecutively removed iSNV with the most extreme changes. We estimated the
effective population size at each iteration and found we would have to remove the top 75% most
extreme iSNV to increase the effective population size by a factor of 10 (Figure 2D). Therefore,
our estimates are robust to a reasonable number of non-neutral sites. Finally, we also applied a
separate Approximate Bayesian Computational (ABC) method, which uses a non-biased
moment estimator in conjunction with ABC to estimate the effective population size of a
population as well as selection coefficients for the iSNV present (17, 45). This distinct approach
relaxes our assumption regarding neutrality. We applied this analysis to the 16 longitudinal pairs
that were sampled 1 day apart and estimated an effective population of 54. We were unable to
reject neutrality for just 4 of the 35 iSNV in this data set (Figure 2E). These four mutations were
distributed between 2 individuals. Each individual had one nonsynonymous iSNV and one
synonymous iSNV. Neither were putative antigenic variants.

Identification of forty-three transmission pairs

The amount of diversity that passes between individuals during transmission determines the
extent to which within host evolution can affect larger evolutionary trends. We analyzed virus
populations from 85 households with concurrent infections to quantify the level of shared viral
diversity and to estimate the size of the IAV transmission bottleneck (Table 1). Because
epidemiological linkage does not guarantee that concurrent cases constitute a transmission pair
(46), we used a stringent rubric to eliminate individuals in a household with co-incident
community acquisition of distinct viruses. We considered all individuals in a household with
symptom onset within a 7-day window to be epidemiologically linked. The donor in each putative
pair was defined as the individual with the earlier onset of symptoms. We discarded a
transmission event if there were multiple possible donors with the same day of symptom onset. Donor and recipients were not allowed to have symptom onset on the same day, unless the individuals were both index cases for the household. In these 6 instances, we analyzed the data for both possible donor-recipient directionalities. Based on these criteria, our cohort had 124 putative household transmission events over 5 seasons (Table 1). Of these, 52 pairs had samples of sufficient quality for reliable identification of iSNV from both individuals.

We next used sequence data to determine which of these 52 epidemiologically linked pairs represented true household transmission events as opposed to coincident community-acquired infections. We measured the genetic distance between influenza populations from each household pair by L1-norm and compared these distances to those of randomly assigned community pairs within each season (Figure 3A, see also trees in Supplemental Figures 2 and 3). While the L1-norm of a pair captures differences between the populations at all levels, in our cohort, it was largely driven by differences at the consensus level. We only considered individuals to be a true transmission pair if they had a genetic distance below the 5th percentile of the community distribution of randomly assigned pairs (Figure 3A). Forty-seven household transmission events met this criterion (Figure 3B). Among these 47 sequence-validated transmission pairs, 3 had no iSNV in the donor and 1 additional donor appeared to have a mixed infection. These four transmission events were removed from our bottleneck analysis as donors without iSNV are uninformative and mixed infections violate model assumptions of site independence (see Methods). We estimated the transmission bottleneck in the remaining 43 high-quality pairs (37 H3N2, 6 H1N1, Figure 3B).

A transmission bottleneck restricts the amount of genetic diversity that is shared by both members of a pair. We found that few minority iSNV where polymorphic in both the donor and recipient populations (Figure 3C). Minority iSNV in the donor were either absent or fixed in the
recipient (top and bottom of plot). The lack of shared polymorphic sites (which would lie in the middle of the plot in Figure 3C) suggests a stringent effective bottleneck in which only one allele is passed from donor to recipient.

Estimation of the transmission bottleneck

We applied a simple presence-absence model to quantify the effective transmission bottleneck in our cohort. True to its name, the presence-absence model simply measures whether or not a donor allele is present or absent in the recipient sample. Under this model, transmission is a neutral, random sampling process, and the probability of transmission is simply the probability that the iSNV will be included at least once in the sample given its frequency in the donor and the sample size, or bottleneck. We estimated a distinct bottleneck for each transmission pair and assumed these bottlenecks followed a zero-truncated Poisson distribution. This model also assumes that the sensitivity for detection of transmitted iSNVs is perfect and that each genomic site is independent of all others. We then used maximum likelihood optimization to determine the distribution of bottleneck sizes that best fit the data. We found a zero-truncated Poisson distribution with a mean of 1.66 (\(\lambda = 1.12; 0.51-1.99, 95\% \text{ CI}\)) best described the data. This distribution indicates that the majority of bottlenecks are 1, and that very few are greater than 5 (probability 0.2\%). There were no apparent differences between H3N2 and H1N1 pairs. The model fit was evaluated by simulating each transmission event 1,000 times. The presence or absence of each iSNV in the recipient was noted and the probability of transmission given donor frequency determined. The range of simulated outcomes matched the data well, which suggests that transmission is a selectively neutral event characterized by a stringent bottleneck (Figure 3D).

The majority of transmitted iSNV were fixed in the recipients. Although this trend matches the expectation given a small bottleneck, these data could also be consistent with a model in which
the probability of transmission is determined by the frequency at which iSNV are found at the community level. To ensure our bottleneck estimates were an outcome of neutral transmission and not an artifact of the larger community population structure or selection for the community consensus, we created a null model by randomly assigning community “recipients” to each donor in our transmission pairings. Each community “recipient” was drawn from the pool of individuals that were infected after the donor but in the same season and with the same subtype as the donor. We then identified whether or not each donor iSNV was found in the community recipient and determined the relationship between donor frequency and probability of “transmission” for 1,000 such simulations. Given the low level of diversity in our cohort, we predicted that rare iSNV would be unlikely to be found in a random sample, while the major alleles should be fixed in most random samples. This trend is clearly demonstrated in Figure 3E. It is also clear that this null model fit the data much more poorly than the presence/absence model, suggesting that the observed data in our bona fide transmission pairs were not a product of community metapopulation structure, but rather an outcome of neutral sampling events.

Because our bottleneck estimates were much lower than what has previously been reported for human influenza (11), we investigated the impact that our simplifying assumptions could have on our results. In particular, the presence-absence model assumes perfect detection of variants in donor and recipient, and it can therefore underestimate the size of a bottleneck in the setting of donor-derived variants that are transmitted but not detected in the recipient. These “false negative” variants can occur when the frequency of an iSNV drifts below the level of detection (e.g. 2% frequency) or when the sensitivity of sequencing is less than perfect for variants at that threshold (e.g. 15% sensitivity for variants at a frequency 2-5%). Leonard et al. recently suggested that a beta binomial transmission model can account for the stochastic loss of transmitted variants, by allowing for a limited amount of time-independent genetic drift within the recipient (28). We modified this model to also account for our benchmarked sensitivity for rare
variants (Supplemental Table 1). For all donor-derived iSNV that were absent in the recipient, we estimated the likelihood that these variants were transmitted but either drifted below our level of detection or drifted below 10% and were missed by our variant identification. Despite the relaxed assumptions provided by this modified beta binomial model, maximum likelihood estimation only marginally increased the average bottleneck size (mean 1.71: lambda 1.19; 0.55-2.12, 95%CI) relative to the simpler presence-absence model. We simulated transmission and subsequent random drift using the beta binomial model and the estimated bottleneck distribution as above (Figure 3F). Although the model matched the data well, the fit was not substantially better than that of the presence-absence model (AIC 75.5 for beta-binomial compared to 76.7 for the presence-absence model).

The stringent influenza transmission bottleneck suggests that most infections are founded by one lineage and develop under essentially clonal processes. The diffusion approximation to the Wright-Fisher model (see above and Figure 2) can be used to predict the rate at which homogenous populations diversify from a clonal ancestor as a function of mutation rate and effective population size (2). By applying maximum likelihood optimization to the model and the frequency distribution of observed alleles (Figure 1C) we estimated an in vivo neutral mutation rate of $4 \times 10^{-6}$ mutations per nucleotide per replication cycle and a within host effective population size of 33 (given a generation time of 6 hours). This is consistent with the estimates above. As we have recently estimated that 13% of mutations in influenza A virus are neutral (47), we estimated that the true in vivo mutation rate would be approximately 8 fold higher than our neutral rate – on the order of $3-4 \times 10^{-5}$. This in vivo mutation rate is close to our recently published estimate of influenza A mutation rates in epithelial cells by fluctuation test (48).

Discussion
We find that seasonal influenza A viruses replicate within and spread among human hosts with very small effective population sizes. Because we used viruses collected over five influenza seasons from individuals enrolled in a prospective household cohort, these dynamics are likely to be broadly representative of many seasonal influenza infections. Other notable strengths of our study include a validated sequence analysis pipeline and the use of models that are robust to the underlying assumptions. The small effective size of intrahost populations and the tight transmission bottleneck suggest that stochastic processes, such as genetic drift, dominate influenza virus evolution at the level of individual hosts. This stands in contrast to prominent role of positive selection in the global evolution of seasonal influenza.

While influenza virus populations are subject to continuous natural selection, selection is an inefficient driver of evolution in small populations (2). Despite a large census, our findings indicate that intrahost populations of influenza virus behave like much smaller populations. We therefore expect random drift to be the major force driving the evolution of influenza virus within human hosts. This finding contradicts previous studies, which have found signatures of adaptive evolution in infected hosts (8, 19, 21, 49). However, these studies rely on data from infections in which selective pressures are likely to be particularly strong (e.g. due to drug treatment or infection with a poorly adapted virus), or in which the virus has been allowed to propagate for extended periods of time. Under these conditions, one can identify the action of positive selection on within host populations. We suggest that these are important exceptions to the drift regime defined here.

We used both a simple presence-absence model and a more complex beta binomial model to estimate an extremely tight transmission bottleneck. The small bottleneck size is driven by the fact that within host diversity was low, and there were very few minority iSNV shared among individuals in a transmission chain. While our methods for variant calling may be more
conservative than those used in similar studies, it is unlikely that our small bottleneck is an artifact of this stringency. The beta binomial model accounts for false negative iSNV (i.e. variants that are transmitted but not detected in the donor), which can lead to underestimated transmission bottlenecks (28). Our formulation of this model incorporates empirically determined sensitivity and specificity metrics to account for both false negative iSNV and false positive iSNV (34). Furthermore, if rare, undetected, iSNV were shared between linked individuals, we would expect to see transmission of more common iSNV (frequency 5-10%), which we can detect with high sensitivity. In our dataset, however, the majority of minority iSNV above 5% were not shared.

Although the size of our transmission bottleneck is consistent with estimates obtained for other viruses and in experimental animal models of influenza (23, 25), it differs substantially from the only other study of natural human infection (11, 28). While there are significant differences in the design and demographics of the cohorts, the influenza seasons under study, and sequencing methodology, the bottleneck size estimates are fundamentally driven by the amount of viral diversity shared among individuals in a household. Importantly, we used both epidemiologic linkage and the genetic relatedness of viruses in households to define transmission pairs and to exclude confounding from the observed background diversity in the community. Whereas we find that household transmission pairs and randomly assigned community pairs had distinct patterns of shared consensus and minority variant diversity, Poon et al. found that rare iSNV were often shared in both household pairs and randomly assigned community pairs (11).

Accurately modeling and predicting influenza virus evolution requires a thorough understanding of the virus’ population structure. Some models have assumed a large intrahost population and a relatively loose transmission bottleneck (24, 50, 51). Here, adaptive iSNV can rapidly rise in frequency and low frequency variants can have a high probability of transmission. In such a
model, it would be possible for the highly pathogenic H5N1 virus to develop the requisite 4-5 mutations to become transmissible through aerosols during a single acute infection of a human host (50, 52). Although the dynamics of emergent avian influenza and human adapted seasonal viruses likely differ, our work suggests that fixation of multiple mutations over the course of a single acute infection is unlikely.

While it seems counterintuitive that influenza evolution is dominated by drift on local scales and positive selection on global scales, these models are not necessarily in conflict. Within individuals we have shown that the effective population is quite small, which suggests that selection is inefficient. Indeed, we have deeply sequenced 332 intrahost populations from 283 individuals collected over more than 11,000 person-seasons of observation and only identified a handful of minority antigenic variants with little evidence for positive selection (this work and (14)). However, with several million infected individuals each year, even inefficient processes and rare events are likely to happen at a reasonable frequency on a global scale.

Methods

Description of the cohort

The HIVE cohort (30, 31), established at the UM School of Public Health in 2010, enrolled and followed households of at least 3 individuals with at least two children <18 years of age; households were then followed prospectively throughout the year for ascertainment of acute respiratory illnesses. Study participants were queried weekly about the onset of illnesses meeting our standard case definition (two or more of: cough, fever/feverishness, nasal congestion, sore throat, body aches, chills, headache if ≥3 yrs old; cough, fever/feverishness, nasal congestion/runny nose, trouble breathing, fussiness/irritability, decreased appetite, fatigue in <3 yrs old), and the symptomatic participants then attended a study visit at the research clinic.
on site at UM School of Public Health for sample collection. For the 2010-2011 through 2013-2014 seasons, a combined nasal and throat swab (or nasal swab only in children < 3 years of age) was collected at the onsite research clinic by the study team. Beginning with the 2014-2015 seasons, respiratory samples were collected at two time points in each participant meeting the case definition; the first collection was a self- or parent-collected nasal swab collected at illness onset. Subsequently, a combined nasal and throat swab (or nasal swab only in children < 3 years of age) was collected at the onsite research clinic by the study team. Families with very young children (< 3 years of age) were followed using home visits by a trained medical assistant.

Active illness surveillance and sample collection for cases were conducted October through May and fully captured the influenza season in Southeast Michigan in each of the study years. Data on participant, family and household characteristics, and on high-risk conditions were additionally collected by annual interview and review of each participant’s electronic medical record. In the current cohort, serum specimens were also collected twice yearly during fall (November-December) and spring (May-June) for serologic testing for antibodies against influenza.

This study was approved by the Institutional Review Board of the University of Michigan Medical School, and all human subjects provided informed consent.

Identification of influenza virus
Respiratory specimens were processed daily to determine laboratory-confirmed influenza infection. Viral RNA was extracted (Qiagen QIAamp Viral RNA Mini Kit) and tested by RT-PCR for universal detection of influenza A and B. Samples with positive results by the universal assay were then subtyped to determine A(H3N2), A(H1N1), A(pH1N1) subtypes and B(Yamagata) and B(Victoria) lineages. We used primers, probes and amplification parameters
developed by the Centers for Disease Control and Prevention Influenza Division for use on the ABI 7500 Fast Real-Time PCR System platform. An RNAseP detection step was run for each specimen to confirm specimen quality and successful RNA extraction.

Quantification of viral load

Quantitative reverse transcription polymerase chain reaction (RT-qPCR) was performed on 5 μl RNA from each sample using CDC RT-PCR primers InfA Forward, InfA Reverse, and InfA probe, which bind to a portion of the influenza M gene (CDC protocol, 28 April 2009). Each reaction contained 5.4 μl nuclease-free water, 0.5 μl each primer/probe, 0.5 μl SuperScript III RT/Platinum Taq mix (Invitrogen 111732) 12.5 μl PCR Master Mix, 0.1 μl ROX, 5 μl RNA. The PCR master mix was thawed and stored at 4°C, 24 hours before reaction set-up. A standard curve relating copy number to Ct value was generated based on 10-fold dilutions of a control plasmid run in duplicate.

Illumina library preparation and sequencing

We amplified cDNA corresponding to all 8 genomic segments from 5 μl of viral RNA using the SuperScript III One-Step RT-PCR Platinum Taq HiFi Kit (Invitrogen 12574). Reactions consisted of 0.5 μl Superscript III Platinum Taq Mix, 12.5 μl 2x reaction buffer, 6 μl DEPC water, and 0.2 μl of 10 μM Uni12/Inf1, 0.3 μl of 10 μM Uni12/Inf3, and 0.5 μl of 10 μM Uni13/Inf1 universal influenza A primers (53). The thermocycler protocol was: 42°C for 60 min then 94°C for 2 min then 5 cycles of 94°C for 30 sec, 44°C for 30 sec, 68°C for 3 min, then 28 cycles of 94°C for 30 sec, 57°C for 30 sec, 68°C for 3 min. Amplification of all 8 segments was confirmed by gel electrophoresis, and 750 ng of each cDNA mixture were sheared to an average size of 300 to 400 bp using a Covaris S220 focused ultrasonicator. Sequencing libraries were prepared using the NEBNext Ultra DNA library prep kit (NEB E7370L), Agencourt AMPure XP beads (Beckman Coulter A63881), and NEBNext multiplex oligonucleotides for Illumina (NEB E7600S). The final
concentration of each barcoded library was determined by Quanti PicoGreen dsDNA quantification (ThermoFisher Scientific), and equal nanomolar concentrations were pooled. Residual primer dimers were removed by gel isolation of a 300-500bp band, which was purified using a GeneJet Gel Extraction Kit (ThermoFisher Scientific). Purified library pools were sequenced on an Illumina HiSeq 2500 with 2x125 nucleotide paired end reads. All raw sequence data have been deposited at the NCBI sequence read archive (BioProject submission ID: SUB2951236). PCR amplicons derived from an equimolar mixture of eight clonal plasmids, each containing a genomic segment of the circulating strain were processed in similar fashion and sequenced on the same HiSeq flow cell as the appropriate patient derived samples. These clonally derived samples served as internal controls to improve the accuracy of variant identification and control for batch effects that confound sequencing experiments.

Identification of iSNV
Intrahost single nucleotide variants were identified in samples that had greater than $10^3$ genomes/μl and an average coverage $>1000x$ across the genome. Variants were identified using DeepSNV and scripts available at https://github.com/lauringlab/variant_pipeline as described previously (34) with a few minor and necessary modifications. Briefly, reads were aligned to the reference sequence (H3N2 2010-2011 & 2011-2012 : GenBank CY121496-503, H3N2 2012-2013:GenBank KJ942680-8, H3N2 2014-2015 : Genbank CY207731-8, H1N1 GenBank : CY121680-8) using Bowtie2 (54). Duplicate reads were then marked and removed using Picard (http://broadinstitute.github.io/picard/). We identified putative iSNV using DeepSNV. Bases with phred <30 were masked. Minority iSNV (frequency <50%) were then filtered for quality using our empirically determined quality thresholds (p-value <0.01 DeepSNV, average mapping quality >30, average Phred >35, average read position between 31 and 94). To control for PCR errors in samples with lower input titers, all isolates with titers between $10^3$ and $10^5$ genomes/μl were processed and sequenced in duplicate. Only iSNV that were found in both
replicates were included in down stream analysis. The frequency of the variant in the replicate with higher coverage at the iSNV location was assigned as the frequency of the iSNV. Finally, any SNV with a frequency below 2% was discarded.

Given the diversity of the circulating strain in a given season, there were a number of cases in which isolates contained mutations that were essentially fixed (>95%) relative to the plasmid control. Often in these cases, the minor allele in the sample matched the major allele in the plasmid control. We were, therefore, unable to use DeepSNV in estimating the base specific error rate at this site for these minor alleles and required an alternative means of eliminating true and false minority iSNV. To this end we applied stringent quality thresholds to these putative iSNV and implemented a 2% frequency threshold. In order to ensure we were not introducing a large number of false positive iSNV into our analysis, we performed the following experiment. Perth (H3N2) samples were sequenced on the same flow cell as both the Perth and Victoria (H3N2) plasmid controls. Minority iSNV were identified using both plasmid controls. This allowed us to identify rare iSNV at positions in which the plasmid controls differed both with and without the error rates provided by DeepSNV. We found that at a frequency threshold of 2% the methods were nearly identical (NPV of 1, and PPV of 0.94 compared to DeepSNV).

Overview of models for effective population size

We estimated the effective population size using two separate interpretations of a Wright-Fisher population. At its base, the Wright-Fisher model describes the expected changes in allele frequency of an ideal population, which is characterized by non-overlapping generations, no migration, no novel mutation, and no population structure. We then asked what size effective population would make the changes in frequency observed in our dataset most likely. We calculated these values using two applications of the Wright-Fisher model (i) a diffusion
approximation (43) and (ii) a maximum likelihood approach based on the discrete interpretation (44).

For these estimates we restricted our analysis to longitudinal samples from a single individual that were separated by at least 1 day and only used sites that were polymorphic in the initial sample (29 of the 49 total serial sample pairs). We modeled only the iSNV that were the minor allele at the first time point, and we assumed a within host generation time of either 6 or 12 hours as proposed by Geoghegan et al (24).

Diffusion approximation

The diffusion approximation was first solved by Kimura in 1955 (43). This approximation to the discrete Wright-Fisher model has enjoyed widespread use in population genetics as it allows one to treat the random time dependent probability distribution of final allele frequencies as a continuous function (e.g. (55-60)). Here, we also included the limitations in our sensitivity to detect rare iSNV by integrating over regions of this probability density that were either below our limit of detection or within ranges where we expect less than perfect sensitivity. Our adaptation of Kimura’s original work is below.

Let \( P(p_0, p_t, t \mid N_e) \) be the time dependent probability of a variant drifting from an initial frequency of \( p_0 \) to \( p_t \) over the course of \( t \) generations given an effective population size of \( N_e \), where \( 0 < p_t < 1 \).

The time dependent derivative of this probability has been defined using the forward Kolmogorov equation and the solution is here adapted from Kimura, 1955 (43).
\begin{equation}
\begin{split}
P(p_0, p_t, t \mid N_e) = \sum_{i=1}^{\infty} p_0 q_0 (i+1)(2i+1)F(1-i, i+2, 2, p) \times F(1-i, i+2, 2, p_t)e^{-\left[\frac{i(i+1)}{2N_e}\right]^t}
\end{split}
\end{equation}

Where \( q = 1 - p \) and \( F \) is the hypergeometric function. We approximated the infinite sum by summing over the first 50 terms. When we added an additional 50 terms (100 in total) we found no appreciable change in the final log likelihoods.

We denote the event that an allele is not observed at the second time point as \( p_t \approx 0 \) and the probability of such an event as \( P(p_0, p_t \approx 0, t \mid N_e) \). This probability is given in equation 2 as the sum of the probability that the variant is lost by generation \( t \) (i.e. the other allele is fixed \( P(q_0, 1, t \mid N_e) \)), the probability that it is not detected due to the limit of detection (i.e. \( P(p, p_t \approx 0, t \mid 0 < p_t < 0.02, N_e) \)) and the probability the variant is not detected due to low sensitivity for rare variant detection (i.e. \( P(p_0, p_t \approx 0, t \mid 0.02 < p_t < 0.1, N_e) \)). The probability of not observing an allele at the second time is then

\begin{equation}
P(p_0, p_t \approx 0, t \mid N_e) = P(q_0, 1, t \mid N_e) + P(p, p_t \approx 0, t \mid 0 < p_t < 0.02, N_e) + P(p_0, p_t \approx 0, t \mid 0.02 < p_t < 0.1, N_e)
\end{equation}

The first term in equation 2 is adapted from Kimura, 1955 as

\begin{equation}
P(q_0, 1, t \mid N_e) = q_0 \sum_{i=1}^{\infty} (2i+1)p_0 q_0 (-1)^i F(1-i, i+2, 2, q_0)e^{-\left[\frac{i(i+1)}{2N_e}\right]^t}
\end{equation}
Where $q$ is defined as above. (Note that this is simply the probability of fixation for a variant at initial frequency $q$). As in equation 1 the infinite sum was approximated with a partial sum of 50 terms.

The probability of the allele drifting below our limit of detection can be found by integrating equation 1 between 0 and our limit of detection, 0.02. This was done numerically using the python package scipy (61).

\[
P(p, p_t \approx 0, t | 0 < p_t < 0.02, N_e) = \int_{0}^{0.02} P(p_0, p_t, t | N_e) dp_t
\]

Finally, the probability of an iSNV being present at the second time point, but escaping detection, is given by the integral of equation 1 between our benchmarked frequencies (0.02,0.05) times the false negative rate for that range. Here, we assumed the entire range had the same sensitivity as the benchmarked frequency at the lower bound and rounded recipient titers down to the nearest $\log_{10}$ titer (e.g. $10^3, 10^4, 10^5$). We also assumed perfect sensitivity above 10%.

\[
P(p_0, p_t \approx 0, t | 0.02 < p_t < 0.1, N_e) = \sum_{f_i}^{[0.02,0.05,0.10]} (\text{FNR} \mid \text{Titer}_r, f_i) \int_{f_i}^{f_i+1} P(p_0, p_t, t | N_e) dp_t
\]

Where (FNR $\mid$ Titer$_r, f_i$) is the false negative rate given the frequency and the sample titer (See Supplemental Table 1) and $P(p_0, p_t, t | N_e)$ is defined in equation 1.

The log likelihood of a given population size is then simply the sum of the log of $P(p_0, p_t, t | N_e)$ for each minor allele in the data set, where either the position is polymorphic at time $t$ (i.e. equation 1) or the allele is observed as lost at time $t$ (i.e. equation 2).
Discrete Wright-Fisher estimation of $N_e$

The diffusion approximation treats changes in frequency as a continuous process because it assumes sufficiently large $N_e$. That assumption can be relaxed, and the effective population size can be determined, by applying a maximum likelihood method developed by Williamsom and Slaktin 1999 (44). In this model, the true allele frequencies move between discrete states (i.e. the frequency must be of the form $i/N_e$ where $i$ is a whole number in the range $[0, N_e]$). In the original application, allele counts were used, and sampling error was added to the model as a binomial distribution with $n$ determined by the sample size. Here, we use the frequencies available from next generation sequencing and estimate sampling error as a normal distribution with mean equal to the observed frequency and a standard deviation equal to that observed in our benchmarking study for the $10^4$ genomes/μl samples ($\sigma = 0.014$) (34).

In this model, the probability of observing an allele frequency shift from $p_0$ to $p_t$ in $\tau$ generations provided an effective population of $N_e$ is the probability of observing $p_0$ given some initial state $q_0$ and the probability of the population having that state, times the probability of observing $p_t$ given some final state $q_\tau$ and the probability of moving from the initial to the final state summed across all possible states.

$$P(p_0, p_t \mid N_e) = \sum q_0 - q_\tau \cdot P(p_0 \mid q_0)P(q_0 \mid N_e)P(p_t \mid q_\tau)P(q_\tau \mid q_0, N_e)$$

(6)

Where $p$ are the observed probabilities and $q$ are the real ones (of the form $i/N_e$ discussed above). The likelihood of observing a given frequency $p_x$ given a defined state $q_x$ is given by the likelihood of drawing $p_x$ from a normal distribution with mean $q_x$ and standard deviation $0.014$. 

24
As in Williamson and Slatkin 1999, we assume a uniform prior on the initial state. Because we know that our specificity is near perfect (Supplemental Table 1) and we restrict our analysis to only polymorphic sites, the probability of any initial state is given by

\[ P(q_0 \mid N_e) = \frac{1}{N_e - 1} \]  

and finally the probability of moving from one state to another in \( t \) generations is given by

\[ P(q_t, q_0 \mid N_e) = v_0 M^t v_t \]  

Where \( M \) is a square transmission matrix with \( C = N_e + 1 \) rows and columns. Where \( m_{i,j} \) is the probability of going from the \( i \)th configuration to the \( j \)th or the probability of drawing \( j - 1 \) out of binomial distribution with mean \((i - 1)/N_e\) and a sample size \( N_e \). \( v_0 \) is a row vector of initial frequencies \( q_0 \) with 100% chance of initial state \( q_0 \), and \( v_t \) is column vector of the frequencies at time point \( t \) with 100% chance of the final state. In other words \( v_0 \) is a row vector of \( C \) states with 0 everywhere except in the \( i \)th position where \( \frac{i-1}{N_e} = q_0 \), and \( v_t \) is a column vector of \( C \) states with 0 everywhere except the \( j \)th position where \( \frac{i-1}{N_e} = q_t \).

Using the scalar and cumulative properties of matrix multiplication equation 6 reduces to

\[ P(p_0, p_t \mid N_e) = [0, P(p_0 \mid q_{0_1})P(q_{0_1} \mid N_e), \ldots, P(p_0 \mid q_{0_{N_e-1}})P(q_{0_{N_e-1}} \mid N_e), 0]M^t \begin{pmatrix} P(p_t \mid q_{t_1}) \\ \vdots \\ P(p_t \mid q_{t_{N_e}}) \end{pmatrix} \]  

\[ (10) \]
The first and last entries in $v_0$ are 0 because we assume all measured sites represent polymorphisms at the first time of sampling. As above, the log likelihood of a given population size is then simply the sum of the log of $P(p_0, p_t, t \mid N_e)$ for each minor allele in the data set.

Simulations

To simulate within host evolution we set $N_e$ in equation 10 to either 30, 50 or 100. For each minor allele we used the closest available non-zero state given the effective population size as the starting state. We then calculated the probability of moving to any other state and selected a final state from this distribution. We then drew a final measured frequency from the normal distribution accounting for measurement errors.

ABC model

We estimated both the effective population size and selection coefficients using the approximate Bayesian computation (ABC) described in (17, 45) with the scripts provided in (45). In its current implementation, this analysis requires the same time points for each sample, and we restricted this analysis to longitudinal samples taken 1 day apart. This subset constitutes 16 of the 29 modeled longitudinal samples. Briefly, we subsampled polymorphic sites to 1,000x coverage to estimate allele counts from frequency data as in (17). We then estimated the prior distribution of the effective population size using 10,000 bootstrap replicates. We selected a uniform distribution on the range [-0.5, 0.5] as the prior distribution for the selection coefficients. The posterior distributions were determined from accepting the top 0.01% of 100,000 simulations.

Overview of models used for estimating the transmission bottleneck

We model transmission as a simple binomial sampling process (28). In our first model, we assume any transmitted iSNV, no matter the frequency, will be detected in the recipient. In the second, we relax this assumption and account for false negative iSNV in the recipient. To
include the variance in the transmission bottlenecks between pairs we use maximum likelihood
optimization to fit the average bottleneck size assuming the distribution follows a zero-truncated
Poisson distribution.

Presence/Absence model

The presence/absence model makes many simplifying assumptions. We assume perfect
detection of all transmitted iSNV in the recipient. For each donor iSNV, we measure only
whether or not the variant is present in the recipient. Any iSNV that is not found in the recipient
is assumed to have not been transmitted. We also assume the probability of transmission is
determined only by the frequency of the iSNV in the donor at the time of sampling (regardless of
how much time passes between sampling and transmission). The probability of transmission is
simply the probability that the iSNV is included at least once in a sample size equal to the
bottleneck. Finally, we assume all genomic sites are independent of one another. For this
reason, we discarded the one case where the donor was likely infected by two strains as the
iSNV were certainly linked.

In our within host models, we only tracked minor alleles as in our data set we only ever find 2
alleles at each polymorphic site. In this case, the frequency of the major allele is simply one
minus the frequency of the minor allele. Because the presence/absence model is unaware of
the frequency of alleles in the recipient we must track both alleles at each donor polymorphic
site.

Let $A_1$ and $A_2$ be alleles in some donor $j$ at some genomic site $i$. Let $P(A_x)$ be the probability
that the $x$ allele is the only transmitted allele. There are then three possible outcomes for each
site. Either only $A_1$ is transmitted, only $A_2$ is transmitted, or both $A_1$ and $A_2$ are transmitted. The
probability of only one allele being transmitted given a bottleneck size of $N_b$ is
where \( p_x \) is the frequency of the \( x \) allele in the donor. In other words, this is simply the probability of only drawing \( A_x \) in \( N_b \) draws.

The probability of both alleles being transmitted is given by

\[
P_{i,j}(A_1, A_2 \mid N_b) = 1 - (p_1^{N_b} + p_2^{N_b})
\]

where \( p_1 \) and \( p_2 \) are the frequencies of the alleles respectively. This is simply the probability of not picking only \( A_1 \) or only \( A_2 \) in \( N_b \) draws.

This system could easily be extended to cases where there are more than 2 alleles present at a site; however, that never occurs in our data set.

For ease we will denote the likelihood of observing the data at a polymorphic site \( i \) in each donor \( j \) given the bottleneck size \( N_b \) as \( P_{i,j}(N_b) \) where \( P_{i,j}(N_b) = P_{i,j}(A_x \mid N_b) \) if only one allele is transmitted and \( P_{i,j}(N_b) = P_{i,j}(A_1, A_2 \mid N_b) \) if two alleles are transmitted.

The log likelihood of a bottleneck of size \( N_b \) is given by

\[
LL(N_b) = \sum_i \sum_j \ln(P_{i,j})
\]
where $\text{Ln}$ is the natural log, and $i, j$ refers to the $i$th polymorphic site in the $j$th donor. This is the log of the probability of observing the data summed over all polymorphic sites across all donors.

Because the bottleneck size is likely to vary between individuals, we used maximum likelihood to fit the bottleneck distribution as oppose to fitting a single bottleneck value. Under this model we assumed the bottlenecks were distributed according to a zero-truncated Poisson distribution parameterized by $\lambda$. The likelihood of observing the data given a polymorphic site $i$ in donor $j$ and $\lambda$ is

$$P_{i,j}(\lambda) = \sum_{N_b=1}^{\infty} P_{i,j}(N_b)P(N_b \mid \lambda)$$

(14)

where $P_{i,j}(N_b)$ is defined as above, $P(N_b \mid \lambda)$ is the probability of drawing a bottleneck of size $N_b$ from a zero-truncated Poisson distribution with a mean of $\frac{\lambda}{1-e^{-\lambda}}$. The sum is across all possible $N_b$ defined on $[1, \infty]$. For practical purposes, we only investigated bottleneck sizes up to 100, as $\lambda$ is quite small and the probability of drawing a bottleneck size of 100 from a zero-truncated Poisson distribution with $\lambda = 10$ is negligible. We follow this convention whenever this sum appears.

The log likelihood of $\lambda$ for the data set is given by

$$LL(\lambda) = \sum_j \sum_i \text{Ln} \left( \sum_{N_b=1}^{\infty} P_{i,j}(N_b)P(N_b \mid \lambda) \right)$$

(15)

**Beta Binomial model**
The Beta binomial model is explained in detail in Leonard et al. (28). It is similar to the presence/absence model in that transmission is modeled as a simple sampling process; however, it loosens a few restricting assumptions. In this model, the frequencies of transmitted variants are allowed to change between transmission and sampling according a beta distribution. The distribution is not dependent on the amount of time that passes between transmission and sampling, but rather depends on the size of the founding population (here assumed to equal to \( N_b \)) and the number of variant genomes present in founding population \( k \). Note the frequency in the donor is assumed to be the same between sampling and transmission.

The equations below are very similar to those presented by Leonard et al. with two exceptions. First, we fit a distribution to the bottleneck sizes in our cohort instead of fitting a single value, and second because we know the sensitivity of our method to detect rare variants based on the expected frequency and the titer, we can include the possibility that iSNV are transmitted but are missed due to poor sensitivity. Because the beta binomial model is aware of the frequency of the iSNV in the recipient, no information is added by tracking both alleles at a genomic site \( i \).

Let \( p_{i,j,d} \) represent the frequency of the minor allele frequency at position \( i \) in the donor of some transmission pair \( j \). Similarly, let \( p_{i,j,r} \) be the frequency of that same allele in the recipient of the \( j \)th transmission pair. Then, as in Leonard et al., the likelihood of some bottleneck \( N_b \) for the data at site \( i \) in pair \( j \) where the minor allele is transmitted is given by

\[
L(N_b)_{i,j} = \sum_{k=1}^{N_b} p_{\text{beta}}(p_{i,j,r} \mid k, N_b - k)p_{\text{bin}}(k \mid N_b, p_{i,j,d})
\]

Where \( p_{\text{beta}} \) is the probability density function for the beta distribution and \( p_{\text{bin}} \) is the probability mass function for the binomial distribution.
This is the probability density that the transmitted allele is found in the recipient at a frequency of $p_{i,jr}$ given that the variant was in $k$ genomes in a founding population of size $N_b$ times the probability $k$ variant genomes would be drawn in a sample size of $N_b$ from the donor where the variant frequency was $p_{i,jd}$. This is then summed for all possible $k$ where $1 < k \leq N_b$.

As in equation 14 the likelihood of a zero truncated Poisson with a mean of $\frac{\lambda}{1-e^{-\lambda}}$ given this transmitted variants is then given by

$$L(\lambda)_{i,j}^{\text{transmitted}} = \sum_{N_b=1}^{\infty} L(N_b)_{i,j} P(N_b \mid \lambda)$$

(17)

This is simply the likelihood of each $N_b$ weighted by the probability of drawing a bottleneck size of $N_b$ from bottleneck distribution.

In this model, there are three possible mechanisms for a donor iSNV to not be detected in the recipient. (i) The variant was not transmitted. (ii) The variant was transmitted but is present below our level of detection (2%). (iii) The variant was transmitted and present above our level of detection but represents a false negative in iSNV identification.

As in Leonard et al., the likelihood of scenarios (i) and (ii) for a given $N_b$ are expressed as

$$L(N_b)_{i,j}^{\text{lost}} = \sum_{k=0}^{N_b} p_{\text{beta_cdf}}(p_{i,jr} < 0.02 \mid k, N_b - k)p_{\text{bin}}(k \mid N_b, p_{i,jd})$$

(18)
Where $p_{\text{beta,cdf}}$ is the cumulative distribution function for the beta distribution. Note that if $k = 0$ (i.e. the iSNV was not transmitted) then the term reduces to the probability of not drawing the variant in $N_b$ draws.

The likelihood of the variant being transmitted but not detected in the recipient given a bottleneck of $N_b$ is described by

$$L(N_b)_{i,j}^{\text{missed}} = \sum_{k=0}^{N_b} \left\{ \sum_{f_e} p_{\text{beta,cdf}}(f_e < p_{i,i_r} < f_{e+1} | k, N_b - k) \times p_{\text{bin}}(k | N_b, p_{i,l_d})(\text{FNR} | \text{Titer_r}, f_e) \right\}$$

(19)

This is the likelihood of the variant existing in the ranges $[0.02,0.05]$ or $[0.05,0.1]$ given an initial frequency of $k/N_b$ and a bottleneck size of $N_b$, multiplied by the expected False Negative Rate (FNR) given the titer of the recipient and the lower frequency bound. As in our diffusion model, we assumed perfect sensitivity for detection of iSNV present above 10%, rounded recipient titers down to the nearest $\log_{10}$ titer (e.g. $10^3, 10^4, 10^5$) and assumed the entire range $[f_e, f_{e+1}]$ has the same sensitivity as the lower bound.

The likelihood of $\lambda$ for iSNV that are not observed in the recipient is then given by summing equations 18 and 19 across all possible $N_b$.

$$L(\lambda)_{i,j}^{\text{nontransmitted}} = \sum_{N_b=1}^{\infty} (L(N_b)_{i,j}^{\text{lost}} + L(N_b)_{i,j}^{\text{missed}})P(N_b | \lambda)$$

(20)
The log likelihood of the total dataset is then determined by summing log of equations 17 and 20 (as applicable) across all polymorphic sites in each donor. (As before here we sum of $N_b$ within the range $[1,100]$.)

**Simulation**

In order evaluate the fits of the two transmission models, we simulated whether or not each donor iSNV was transmitted or not. This involved converting each model to a presence absence model. In each simulation, we assigned a bottleneck from the bottleneck distribution for each transmission pair. We then determined the probability of only transmitting one allele ($A_x$ where $x \in [1,2]$ as in the presence/absence model above) and the probability of transmitted both alleles ($A_1, A_2$ above) for each polymorphic site.

For the presence/absence model, the probabilities for each possible outcome are given by equations 11 and 12. For the beta binomial model, the probability of only observing $A_x$ at site $i$ is given by

$$P(A_x \mid N_b) = L(N_b)_{i,j}^{\text{lost}} + L(N_b)_{i,j}^{\text{missed}}$$

(21)

where $L(N_b)_{i,j}^{\text{lost}}$ and $L(N_b)_{i,j}^{\text{missed}}$ are defined as in equations 18 and 19 respectively, but with $p_{i,j,i}$ replaced by $1 - p_{i,j}$. This is simply the probability of not observing the other allele in the recipient.

Again, the probability of observing both alleles is

$$P(A_1, A_2 \mid N_b) = 1 - (P(A_1) + P(A_2))$$
where \( P(A_1) \) and \( P(A_2) \) are defined as in equation 21.

**Fitting mutation rate and \( N_e \)**

The diffusion approximation to the Wright-Fischer model allows us to make predictions on the allele frequency spectrum of a population given a mutation rate and an effective population size. The probability of observing a mutation at frequency \( p_t \) given an initial frequency of 0 can be approximated as in (2)

\[
P(0, p_t, t \mid \mu, N_e) = \frac{2\mu N_e}{p_t} e^{-\frac{2N_e p_t}{t}}
\]

(23)

Where \( \mu \) is the mutation rate. In this model mutation increases an allele’s frequency from 0 but after that initial jump, drift is responsible for allowing the mutation to reach its observed frequency. Because the limit of equation 23 approaches infinity as \( p_t \) approaches 0 and for ease in numerical integration, we assumed that any variant present at less than 0.1% was essentially at 0%.

We then assumed each infection began as a clonal infection matching the consensus sequence observed at the time of sampling. The likelihood of observing minor alleles at the observed frequency is the given by equation 23.

As in the other within host models, we can account for nonpolymorphic sites by adding the likelihood that no mutation is present \( P(0, p_t \approx 0, t \mid p_t < 0.001, \mu, N_e) \), that a mutation is present but below our level of detection \( P(0, p_t \approx 0, t \mid p_t < 0.02, \mu, N_e) \), and that a mutation is present but missed due to low sensitivity at low frequencies \( P(0, p_t \approx 0, t \mid 0.02 < p_t < 0.1, \mu, N_e) \). In this
model we assumed 13133 mutagenic targets in each sample (the number of coding sites present in the reference strain from 2014-2015).

The probability of not observing a mutation is given by

\[ P(0, p_t \approx 0, t \mid p_t < 0.001, \mu, N_e) + P(0, p_t \approx 0, t \mid p_t < 0.02, \mu, N_e) + P(0, p_t \approx 0, t \mid 0.02 < p_t < 0.1, \mu, N_e) \]

(24)

Where

\[ P(0, p_t \approx 0, t \mid p_t < 0.001, \mu, N_e) = 1 - \int_{0.001}^{1} P(0, p_t, t \mid \mu, N_e)dp_t \]

(25)

and

\[ P(0, p_t \approx 0, t \mid p_t < 0.02, \mu, N_e) = \int_{0.001}^{0.02} P(0, p_t, t \mid \mu, N_e)dp_t \]

(26)

and

\[ P(0, p_t \approx 0, t \mid 0.02 < p_t < 0.1, \mu, N_e) = \sum_{f_i}^{[0.02,0.05,0.10]} \left( \text{FNR} \mid \text{Titer}_r, f_i \right) \int_{f}^{f_i+1} P(p_0, p_t, t \mid \mu, N_e)dp_t \]

(27)

Where we follow the same convention as in equation 5 for determining the false negative rate.

The log likelihood of a given \( \mu \) and \( N_e \) pair is then the sum of the log of equations 23 and 24 for all possible sites in the data set.

Annotated computer code for all analyses can be accessed at

https://github.com/lauringlab/Host_level_IAV_evolution
Acknowledgments

This work was supported by a Clinician Scientist Development Award from the Doris Duke Charitable Foundation (CSDA 2013105), a University of Michigan Discovery Grant, and R01 AI118886, all to ASL. The HIVE cohort was supported by NIH R01 AI097150 and CDC U01 IP00474 to ASM. JTM was supported by the Michigan Predoctoral Training Program in Genetics (T32GM007544). RJW was supported by K08AI119182. We thank Alexey Kondrashov and Aaron King for helpful discussion.

References


43. Kimura M (1955) Solution of a process of random genetic drift with a continuous model.


57. Nagasawa M, Maruyama T (1979) An application of time reversal of Markov processes to
a problem of population genetics. *Advances in Applied Probability*.


Figure 1. Within host diversity of IAV populations. (A) Boxplots (median, 25th and 75th percentiles, whiskers extend to most extreme point within median ± 1.5 x IQR) of the number of viral genomes per microliter transport media stratified by day post symptom onset. Notches represent the approximate 95% confidence interval of the median. (B) Boxplots (median, 25th and 75th percentiles, whiskers extend to most extreme point within median ± 1.5 x IQR) of the number of iSNV in 249 high quality samples stratified by day post symptom onset. (C) Histogram of within host iSNV frequency in 249 high quality samples. Bin width is 0.05 beginning at 0.02. Mutations are colored nonsynonymous (blue) and synonymous (gold) (D) Location of all identified iSNV in the influenza A genome. Mutations are colored nonsynonymous (blue) and synonymous (gold) relative to that sample’s consensus sequence. Triangles signify mutations that were found in more than one individual in a given season.

Figure 2. Within host dynamics of IAV. (A) Timing of sample collection for 35 paired longitudinal samples relative to day of symptom onset. Of the 49 total, 35 pairs had minor iSNV present in the first sample. (B) The change in frequency over time for minority nonsynonymous (blue) and synonymous (gold) iSNV identified for the paired samples in (A). (C) The distribution of effective population sizes estimated from 1,000 simulated populations. Simulations were run on populations with characteristics similar to the actual patient-derived populations and with the specified effective population size (x-axis). (D) The effect of iteratively removing iSNV with the most extreme change in frequency (fraction of iSNV removed, x-axis) on the estimated effective population size. The point represents the estimate when all iSNV are included (32). (E) The posterior distributions of selection coefficients estimated for the 35 iSNV present in isolates sampled one day apart. Distributions are colored according to class relative to the sample consensus sequence, nonsynonymous (blue) synonymous (gold). Variants for which the 95%
highest posterior density intervals exclude 0.0 are noted in the margin.

Figure 3. Between host dynamics of IAV. (A) The distribution of pairwise L1-norm distances for household (blue) and randomly-assigned community (gold) pairs. The bar heights are normalized to the height of the highest bar for each given subset (47 for household, 1,590 for community). The red line represents the 5th percentile of the community distribution. (B) Timing of symptom onset for 52 epidemiologically linked transmission pairs. Day of symptom onset for both donor and recipient individuals is indicated by black dots. Dashed lines represent pairs that were removed due to abnormally high genetic distance between isolates, see (A). (C) The frequency of donor iSNV in both donor and recipient samples. Frequencies below 2% and above 98% were set to 0% and 100% respectively. (D) The presence-absence model fit compared with the observed data. The x-axis represents the frequency of donor iSNV with transmitted iSNV plotted along the top and nontransmitted iSNV plotted along the bottom. The black line indicates the probability of transmission for a given iSNV frequency as determined by logistic regression. Similar fits were calculated for 1,000 simulations with a mean bottleneck size of 1.66. Fifty percent of simulated outcomes lie in the darkly shaded region and 95% lie in the lightly shaded regions. (E) The outcome from 1,000 simulated “transmission” events with randomly assigned recipients. The black line represents the observed data, as in (D) the shaded regions represent the middle 50% and 95% of simulated outcomes. The results from the simulated logit models were smoothed by plotting the predicted probability of transmission at 0.02 intervals. (F) The beta-binomial model fit. Similar to (D) except the simulated outcomes are based on a beta-binomial model using a mean bottleneck of 1.71.

Figure 4. Combined estimates of within host mutation rate and effective population size. Contour plot shows the log likelihood surface for estimates of the effective population size and neutral mutation rate. The point represents the peak ($\mu = 4 \times 10^{-6}$, $N_e = 33$, log likelihood = -
Supplemental Figure 1. Sequence coverage for all samples. For each sample, the sliding window mean coverage was calculated using a window size of 200 and a step of 100. The distributions of these means are plotted as box plots (median, 25th and 75th percentiles, whiskers extend to most extreme point within median ± 1.5 x IQR) where the y-axis represents the read depth and the x-axis indicates the position of the window in a concatenated IAV genome.

Supplemental Figure 2. Approximate maximum likelihood trees of the concatenated coding sequences for high quality H1N1 samples. The branches are colored by season; the tip identifiers are colored by household. Arrows with numbers indicate consensus and putative minor haplotypes for samples with greater than 10 iSNV.

Supplemental Figure 3. Approximate maximum likelihood trees of the concatenated coding sequences for high quality H3N2 samples. The branches are colored by season; the tip identifiers are colored by household. Arrows with numbers indicate consensus and putative minor haplotypes for samples with greater than 10 iSNV.

Supplemental Figure 4. The effect of titer and vaccination on the number of iSNV identified. (A) The number of iSNV identified in an isolate (y-axis) plotted against the titer (x-axis, genomes/μl transport media). (B) The number of iSNV identified in each isolate stratified by whether that individual was vaccinated or not. Red bars indicate the median of each distribution.

Supplemental Figure 5. Minority nonsynonymous iSNV in global circulation. The global frequencies of the amino acids that were found as minority variants in sample
isolates (x-axis) plotted overtime (y-axis). Each amino acid trace is labeled according to the H3 number scheme. All samples were isolated in December of 2014 (gray line).

Supplemental Figure 6. Reproducibility of iSNV identification for paired samples acquired on the same day. The x-axis represents iSNV frequencies found in the home-acquired nasal swab. The y-axis represents iSNV frequencies found the clinic-acquired combined throat and nasal swab.
### Table 1. Influenza viruses over five seasons in a household cohort

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Households</td>
<td>328</td>
<td>213</td>
<td>321</td>
<td>232</td>
<td>340</td>
</tr>
<tr>
<td>Participants</td>
<td>1441</td>
<td>943</td>
<td>1426</td>
<td>1049</td>
<td>1431</td>
</tr>
<tr>
<td>Vaccinated, n (%)</td>
<td>934 (65)</td>
<td>554 (59)</td>
<td>942 (66)</td>
<td>722 (69)</td>
<td>992 (69)</td>
</tr>
<tr>
<td>IAV Positive Individuals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1N1</td>
<td>26</td>
<td>1</td>
<td>3</td>
<td>47</td>
<td>0</td>
</tr>
<tr>
<td>H3N2</td>
<td>58</td>
<td>22</td>
<td>66</td>
<td>1</td>
<td>166</td>
</tr>
<tr>
<td>IAV Positive Households</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two individuals</td>
<td>13</td>
<td>2</td>
<td>9</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>Three individuals</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Four individuals</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>High Quality NGS Pairs</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>39</td>
</tr>
</tbody>
</table>

- **a** Self reported or confirmed receipt of vaccine prior to the specified season.
- **b** RT-PCR confirmed infection.
- **c** Households in which two individuals were positive within 7 days of each other. In cases of trios and quartets, the putative chains could have no pair with onset >7 days apart.
- **d** Samples with >$10^3$ genome copies per µl of transport medium, adequate amplification of all 8 genomic segments, and average sequencing coverage >$10^3$ per nucleotide.
Table 2. Within host effective population size of IAV

<table>
<thead>
<tr>
<th>Model</th>
<th>SNV Used</th>
<th>Generation Time (h)</th>
<th>Effective Population Size (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diffusion approximation</td>
<td>All</td>
<td>6</td>
<td>35 (26-46)</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>12</td>
<td>17 (13-23)</td>
</tr>
<tr>
<td>Discrete model</td>
<td>All</td>
<td>6</td>
<td>32 (28-41)</td>
</tr>
<tr>
<td></td>
<td>Nonsynonymous</td>
<td>6</td>
<td>30 (21-40)</td>
</tr>
<tr>
<td></td>
<td>Synonymous</td>
<td>6</td>
<td>37 (27-54)</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>12</td>
<td>23 (23-29)</td>
</tr>
<tr>
<td></td>
<td>Nonsynonymous</td>
<td>12</td>
<td>19 (19-21)</td>
</tr>
<tr>
<td></td>
<td>Synonymous</td>
<td>12</td>
<td>27 (22-33)</td>
</tr>
</tbody>
</table>
**Supplemental Table 1. Sensitivity and specificity of variant detection**

<table>
<thead>
<tr>
<th>Copy Number&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Variant Frequency</th>
<th>Original Pipeline&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Current Pipeline&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>&gt;10&lt;sup&gt;5&lt;/sup&gt;</td>
<td>0.05</td>
<td>1</td>
<td>&gt;0.9999</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>0.85</td>
<td>0.9999</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.95</td>
<td>0.9995</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.35</td>
<td>0.9999</td>
</tr>
<tr>
<td>10&lt;sup&gt;4&lt;/sup&gt;-10&lt;sup&gt;5&lt;/sup&gt;</td>
<td>0.05</td>
<td>0.95</td>
<td>0.9999</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>0.9</td>
<td>0.9999</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.8</td>
<td>0.9998</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.4</td>
<td>0.9999</td>
</tr>
<tr>
<td>10&lt;sup&gt;3&lt;/sup&gt;-10&lt;sup&gt;4&lt;/sup&gt;</td>
<td>0.05</td>
<td>0.8</td>
<td>&gt;0.9999</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>0.45</td>
<td>0.9999</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.2</td>
<td>0.9997</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.1</td>
<td>0.9999</td>
</tr>
</tbody>
</table>

<sup>a</sup> Per µl transport media

<sup>b</sup> As described in McCrone JT and Lauring AS, J. Virol. 90(15):6884, 2016.

<sup>c</sup> As described in Methods, benchmarked for frequencies 0.02-0.98 only
<table>
<thead>
<tr>
<th>House ID</th>
<th>Enrolment ID</th>
<th>Symptom Onset</th>
<th>Subtype</th>
<th>Frequency</th>
<th>Amino Acid Change</th>
<th>Antigenic Site</th>
<th>Vaccinated</th>
<th>Day of Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1111</td>
<td>300481</td>
<td>3-30-2011</td>
<td>H3N2</td>
<td>0.071</td>
<td>E62G</td>
<td>E*</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>2166</td>
<td>320661</td>
<td>2-13-2012</td>
<td>H3N2</td>
<td>0.071</td>
<td>V297A</td>
<td>C</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>1302</td>
<td>301355</td>
<td>3-20-2011</td>
<td>H3N2</td>
<td>0.088</td>
<td>L86i</td>
<td>E</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>3075</td>
<td>331045</td>
<td>12-10-2012</td>
<td>H3N2</td>
<td>0.066</td>
<td>I214T</td>
<td>D</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>5219</td>
<td>50935</td>
<td>12-5-2014</td>
<td>H3N2</td>
<td>0.175</td>
<td>F193S</td>
<td>B*†</td>
<td>No</td>
<td>3</td>
</tr>
<tr>
<td>5263</td>
<td>51106</td>
<td>12-6-2014</td>
<td>H3N2</td>
<td>0.111</td>
<td>T128A</td>
<td>B</td>
<td>Yes</td>
<td>3</td>
</tr>
<tr>
<td>5290</td>
<td>51225</td>
<td>12-15-2014</td>
<td>H3N2</td>
<td>0.405</td>
<td>I260V</td>
<td>E*</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>5302</td>
<td>51273</td>
<td>12-13-2014</td>
<td>H3N2</td>
<td>0.030</td>
<td>S262N</td>
<td>E*</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>5098</td>
<td>50419</td>
<td>12-22-2014</td>
<td>H3N2</td>
<td>0.364</td>
<td>G208R</td>
<td>D</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td>5033</td>
<td>50141</td>
<td>12-3-2014</td>
<td>H3N2</td>
<td>0.032</td>
<td>A163T</td>
<td>B</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>5034</td>
<td>50143</td>
<td>1-11-2015</td>
<td>H3N2</td>
<td>0.119</td>
<td>I307R</td>
<td>C</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>5289</td>
<td>51220</td>
<td>12-13-2015</td>
<td>H3N2</td>
<td>0.038</td>
<td>K189N</td>
<td>B*†</td>
<td>Yes</td>
<td>-1</td>
</tr>
<tr>
<td>5033</td>
<td>50141</td>
<td>12-3-2014</td>
<td>H3N2</td>
<td>0.025</td>
<td>D53E</td>
<td>C*</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>5033</td>
<td>50141</td>
<td>12-3-2014</td>
<td>H3N2</td>
<td>0.023</td>
<td>S312G</td>
<td>C</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>5269</td>
<td>51132</td>
<td>12-6-2014</td>
<td>H3N2</td>
<td>0.028</td>
<td>I242T</td>
<td>D</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>5147</td>
<td>50630</td>
<td>11-18-2014</td>
<td>H3N2</td>
<td>0.164</td>
<td>I242L</td>
<td>D</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>5034</td>
<td>50143</td>
<td>1-11-2015</td>
<td>H3N2</td>
<td>0.161</td>
<td>I307R</td>
<td>C</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>4185</td>
<td>UM40738</td>
<td>12-14-2013</td>
<td>H1N1</td>
<td>0.021</td>
<td>R208K</td>
<td>Ca</td>
<td>No</td>
<td>2</td>
</tr>
</tbody>
</table>

* Sites observed to vary between antigenically distinct strains in Wiley et al., 1981 and Smith DJ et al., 2004.
† Sites located in the “antigenic ridge” identified in Koel et al., 2013.
Figure 1

A

B

C

D

iSNV

Day post symptom onset

Day post symptom onset

Frequency

Nonsynonymous

Synonymous

PB2

PB1

PA

HA

NP

NA

M

NS
Figure 2

A

B

C

D

E

Days post symptom onset

Time within host (days)

Input Ne in Simulations

Fraction iSNV Removed from Analysis

Effective Population Size (Ne)

Effective Population Size (Ne)

PB1 G594 syn
NA N161S
NS1 I171T ; NS2 L14 syn
HA S165 syn

selection coefficient

-0.50 -0.25 0.00 0.25 0.50

−2 −1 0 1 2 3 4 5 6

−0.2 0.0 0.2 0.4

0 1 2 3 4 6

−0.5 0 0.5

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

240 200 160 120 80 40 0

30 50 100

0 100 200 300 400 500

0 100 200 300 400

PB1 G594 syn
NA N161S
NS1 I171T ; NS2 L14 syn
HA S165 syn
Figure 4

The diagram shows the relationship between effective population size and neutral mutation rate. The contour lines represent different values, with labels indicating specific values for the effective population size and neutral mutation rate.