Within-host evolutionary dynamics of seasonal and pandemic human influenza A viruses in young children

Alvin X. Han[1,*], Zandra C. Felix Garza[1,*], Matthijs R. A. Welkers[1,*], René M. Vigeveno[1], Tran Nhu Duong[2], Le Thi Quynh Mai[2], Pham Quang Thai[2], Dang Dinh Thoang[3], Tran Thi Ngoc Anh[4], Ha Manh Tuan[4,] Nguyen Thanh Hung[5], Le Quoc Thinh[5], Le Thanh Hai[6], Hoang Thi Bich Ngoc[6], Kulkanya Chokephaibulkit[7], Pilaipan Puthavathana[7], Nguyen Van Vinh Chau[8], Nghiem My Ngoc[8], Nguyen Van Kinh[9], Dao Tuyet Trinh[9], Tran Tinh Hien[7,10], Heiman F. L. Wertheim[10,11,12], Peter Horby[12,13], Annette Fox[13,14,15], H. Rogier van Doorn[12,13], Dirk Eggink[1,16,†], Menno D. de Jong[1,†], Colin A. Russell[1,†]

*Contributed equally, †Contributed equally

[1]Department of Medical Microbiology & Infection Prevention, Amsterdam University Medical Center, Amsterdam, The Netherlands

[2]National Institute of Hygiene and Epidemiology, Hanoi, Vietnam

[3]Ha Nam Centre for Disease Control, Ha Nam, Vietnam

[4]Children's Hospital 2, Ho Chi Minh city, Vietnam

[5]Children's Hospital 1, Ho Chi Minh city, Vietnam

[6]Vietnam National Children's Hospital, Hanoi, Vietnam

[7]Siriraj Hospital, Mahidol University, Bangkok, Thailand

[8]Hospital for Tropical Diseases, Ho Chi Minh city, Vietnam

[9]National Hospital for Tropical Diseases Hanoi, Vietnam

[10]Oxford University Clinical Research Unit, Ho Chi Minh city, Vietnam

[11]Radboud Medical Centre, Radboud University, Nijmegen, The Netherlands

[12]Nuffield Department of Medicine, University of Oxford, Oxford, UK

[13]Oxford University Clinical Research Unit, Hanoi, Vietnam

[14]Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Australia

[15]WHO Collaborating Centre for Reference and Research on Influenza, Melbourne, Australia

[16]Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

Correspondence to A.X.H. (x.han@amsterdamumc.nl) & C.A.R. (c.a.russell@amsterdamumc.nl)

Abstract

The evolution of influenza viruses is fundamentally shaped by within-host processes. However, the within-host evolutionary dynamics of influenza viruses remain incompletely understood, in part because most studies have focused on infections in healthy adults based on single timepoint data. Here, we analysed the within-host evolution of 82 longitudinally-sampled individuals, mostly young children, infected with A/H1N1pdm09 or A/H3N2 viruses between 2007 and 2009. For A/H1N1pdm09 infections during the 2009 pandemic, nonsynonymous minority variants were more prevalent than synonymous ones. For A/H3N2 viruses in young children, early infection was dominated by purifying selection. As these infections progressed, nonsynonymous variants typically increased in frequency even when within-host virus titres decreased. Unlike the short-lived infections of adults where *de novo* within-host variants are rare, longer infections in young children allow for the maintenance of virus diversity via mutation-selection balance creating potentially important opportunities for within-host virus evolution.

Introduction

Influenza A viruses (IAV) are some of the most prevalent human respiratory pathogens, infecting hundreds of millions of people worldwide each year. Because of the high error rates of the viral RNA polymerase complex, *de novo* mutants are generated as the viruses replicate within infected hosts[1]. However, the emergence of these variants within host does not mean that they will become the majority variant within the infected host or be transmitted between hosts. The evolution of IAVs is the product of a complex mosaic of evolutionary processes that include genetic drift, positive selection[2] , transmission bottleneck effects[3,4] and global migration patterns[5,6]. Importantly, the resulting evolutionary dynamics can differ at the individual and population levels[7].

For seasonal IAVs at the global population level, antibody-mediated immune selection pressure from natural infection or vaccination positively selects for novel antigenic variants that facilitate immune escape resulting in antigenic drift[2]. However, at the within-host level, the role of positive selection exerted by immunity is less obvious. Several next generation sequencing studies of typical, short-lived seasonal IAV infections in adult humans showed that intra-host genetic diversity of influenza viruses is low and dominated by purifying selection[4,8–11]. Additionally, large scale comparative analyses of IAV haemagglutinin (HA) consensus sequences found limited evidence of positive selection on HA at the individual level regardless of the person's expected influenza virus infection history[12]. Importantly, these studies focused on virus samples from only one or two time points, mostly early in infection, limiting the opportunities to study how virus populations evolved over the course of infection.

Separate from seasonal IAVs, zoonotic IAVs constantly pose new pandemic threats. Prior to becoming human-adapted seasonal strains, IAVs are introduced into the human population from an animal reservoir through the acquisition of host adaptive mutations, sometimes via reassortment, resulting in global pandemics such as the 2009 swine influenza pandemic[13]. In the 2009 pandemic, global virus genetic diversity increased rapidly during the early phases of the pandemic as a result of rapid transmissions in the predominantly naïve human population[14]. Over subsequent waves of the pandemic, host adapting mutations that incrementally improved viral fitness and transmissibility in humans of A/H1N1pdm09 viruses emerged[15], eventually reaching fixation in the global virus population[16].

At the individual level, the within-host evolutionary dynamics of the pandemic A/H1N1pdm09 virus, particularly in the early stages of the 2009 pandemic, have been relatively underexplored. To date, the only within-host genetic diversity analysis of A/H1N1pdm09 viruses during the initial phase of the pandemic was based on mostly single-

timepoint samples collected within ~7 days post-symptom onset[17]. Despite initial findings of high within-host diversity and loose transmission bottlenecks[17], these results were later disputed due to technical anomalies and subsequent reanalyses of a smaller subset of the original data found that intra-host genetic diversity of the pandemic virus was low and comparable to levels observed in seasonal IAVs[18,19]. It remains unclear how frequently host adaptive mutations appear within hosts infected by a pandemic IAV and if these mutants are readily transmitted between individuals.

Here, we deep sequenced 275 longitudinal clinical specimens sampled from 82 individuals residing in Southeast Asia between 2007 and 2009 that were either infected with seasonal A/H3N2 or pandemic A/H1N1pdm09 viruses. By analysing minority variants found across the whole IAV genome, we characterised the evolutionary dynamics of within-host virus populations in these samples collected up to two weeks post-symptom onset.

Results

Study participants

The A/H3N2 virus samples were collected from 51 unlinked individuals as part of an oseltamivir dosage trial[20,21]. 48 of the 51 A/H3N2 virus infected individuals were young children (median age=2 years; interquartile range (IQR)=2-3 years) at the time of sampling and most had low or no detectable anti-influenza virus antibody titers on day 0 and 10 post-symptom onset[21]. Given that young children are substantial contributors to influenza virus transmission[22,23], the samples analysed here offer a valuable opportunity to investigate the within-host IAV evolutionary dynamics in this key population. The A/H1N1pdm09 virus specimens were collected from 32 individuals up to 12 days post-symptom onset. These individuals include both children and adults (median age=10 years; IQR=4-20 years) infected during the first wave of the pandemic in Vietnam (July-December 2009). 15 of the 32 individuals (including 6 index patients) were sampled in a household-based influenza cohort study[24]. The remaining 16 unlinked individuals were hospitalised patients that were involved in two different oseltamivir treatment studies[20,25] Details of all study participants are described in the respective cited studies and Supplemental File 4.

Genetic diversity of within-host virus populations

We used the number of minority intra-host single nucleotide variants (iSNVs; $\geq 2\%$ in frequencies) to measure the levels of genetic diversity of within-host IAV populations. Similar to previous studies[4,8,9,11], within-host genetic diversity of human A/H3N2 virus populations was low (median = 11 iSNVs, interquartile-range (IQR) = 7-16; Figure 1A). Within-host genetic diversity of pandemic A/H1N1pdm09 virus populations was also low,

with a median number of 21 iSNVs (IQR = 13.5-30.0; Figure 1B) identified. Cycle threshold (Ct) values, and thus likely virus shedding, correlated with the number of days post-symptom onset for both IAV subtypes (A/H3N2: Spearman's $\rho = 0.468$, $p = 1.38 \times 10^{-10}$; A/H1N1pdm09: $\rho = 0.341$, $p = 0.048$; Figure 1C and D). The number of iSNVs observed in within-host A/H3N2 virus populations weakly correlated with days since onset of symptoms in patients ($\rho = 0.463$, $p = 2.22 \times 10^{-10}$) and Ct values ($\rho = 0.508$, $p = 1.20 \times 10^{-12}$), suggesting that as infection progresses, genetic variants accumulate within-host even as virus population size decreases (Figure 1A). On the other hand, there was no significant correlation between the number of iSNVs observed in within-host A/H1N1pdm09 virus populations and Ct values ($\rho = 0.198$, $p = 0.21$) or days post-symptom onset ($\rho = -0.021$, $p = 0.91$) (Figure 1B).

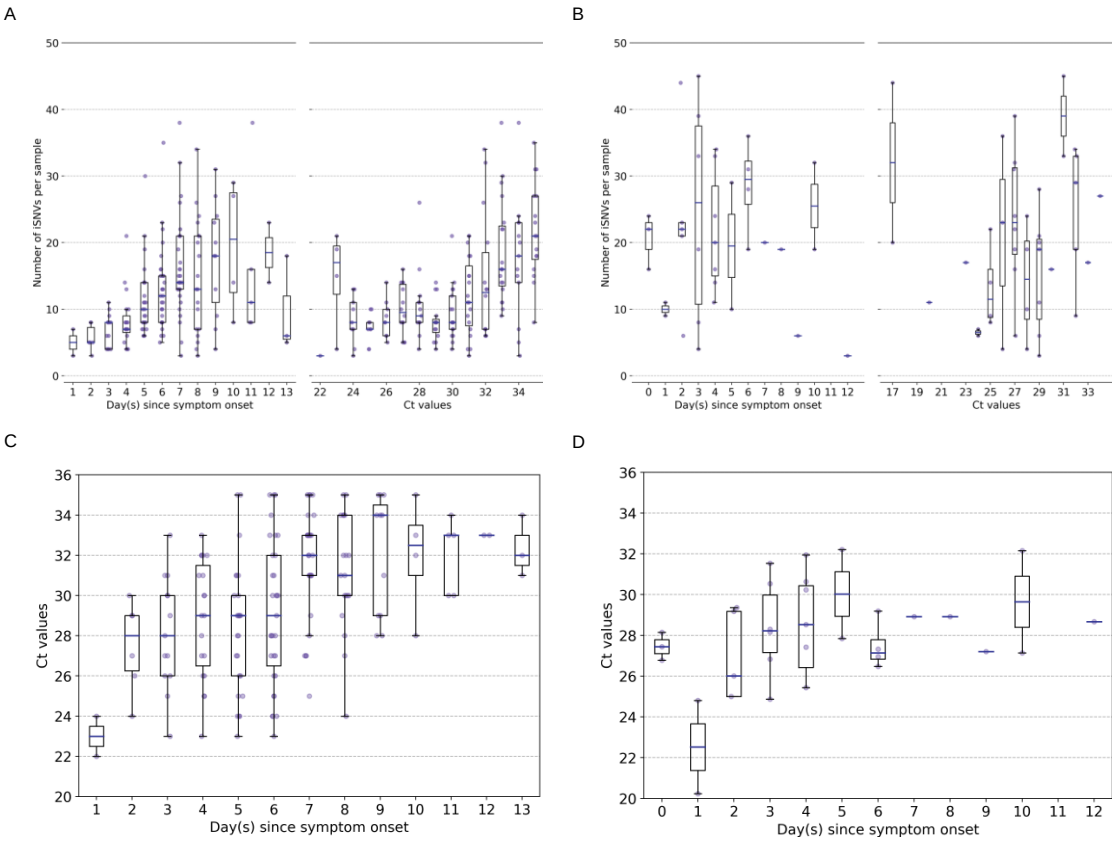


Figure 1: Genetic diversity of within-host influenza A virus populations. Box plots summarizing the number of intra-host single nucleotide variants (iSNVs; median, interquartile range (IQR), and whiskers extending within median $\pm 1.5 \times$ IQR) identified in samples with adequate breadth of coverage across the whole influenza virus genome in (**A**) seasonal A/H3N2 and (**B**) pandemic A/H1N1pdm09 virus samples, stratified by day(s) since symptom onset or qPCR cycle threshold (Ct) values. (**C**, **D**) Ct values as a function of day(s) since symptom onset for A/H3N2 viruses (**C**) and A/H1N1pdm09 viruses (**D**).

Within-host evolutionary rates of influenza A viruses

To investigate within-host evolutionary dynamics, empirical rates of synonymous, non-synonymous, and premature stop-codon (i.e. nonsense) iSNVs were calculated by normalizing the summation of observed iSNV frequencies with the number of available sites and time since symptom onset (see Methods). The overall within-host evolutionary rates of A/H3N2 viruses observed here are in the same order of magnitude ($<\sim 10^{-5}$ divergence per site per day) as those reported in previous within-host seasonal influenza virus evolution studies (Figure 2A)[26]. Synonymous evolutionary rates were significantly higher than nonsynonymous rates during the initial phase of A/H3N2 virus infections (Figure 2A), primarily in the polymerase complex and HA genes (Figure 2A, Figure 2 – figure supplement 1 and Figure 3 – figure supplement 1). Importantly, nonsynonymous variants gradually accumulated, increasing in rates around four days post-symptom onset to similar levels relative to synonymous rates. To ensure that this temporal trend was not due to aggregated effects across multiple individuals, we performed linear regression on the computed evolutionary rates for each A/H3N2 infected individual with at least three sampling timepoints (n=39). Nonsynonymous evolutionary rates were positively correlated against time for 25/39 individuals (64%; Figure 2 – figure supplement 3). In contrast, synonymous evolutionary rates were negatively correlated against time for 27 (69%) individuals.

Consolidating over all samples, most nonsynonymous variants were found in the nucleoprotein (NP) and neuraminidase (NA) gene segments (nonsynonymous to synonymous variant (NS/S) ratios = 1.69 (NP) and 1.32 (NA) whereas NS/S ratios were ≤ 1 for all other gene segments; Figure 2 – figure supplement 1 and Supplemental File 1). While nonsynonymous NA mutations associated with oseltamivir resistance were positively selected for a subset of individuals in response to the antiviral treatment[21], nonsynonymous changes to NP were likely mediated by protein stability, T-cell immune response and/or host cellular factors (see next section).

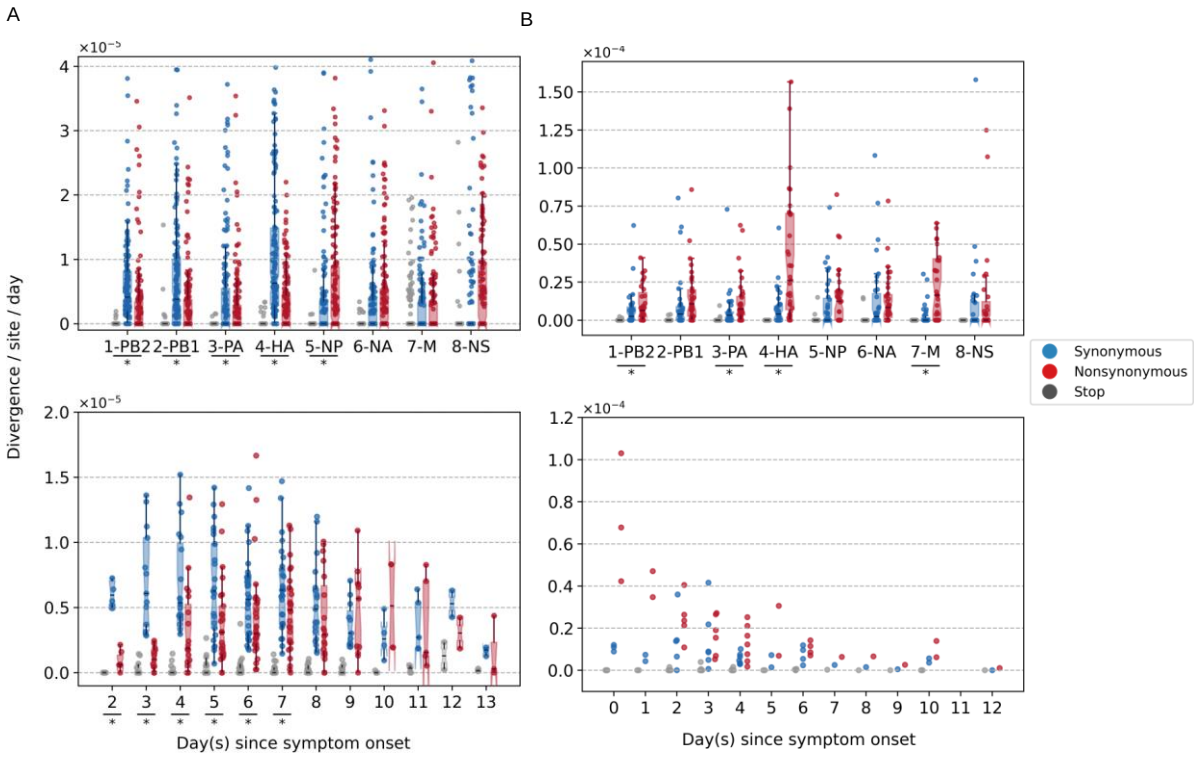


Figure 2: Box plots (median, interquartile range (IQR), and whiskers extending within median $\pm 1.5 \times$ IQR) summarizing the empirical within-host evolutionary rates of (**A**) seasonal A/H3N2 viruses and (**B**) pandemic A/H1N1pdm09 viruses. Top panel shows the evolutionary rate of individual gene segments over all timepoints (r_g) while the bottom panel depicts the genome-wide evolutionary rate (r_t) for each day since symptom onset. All rates are stratified by substitution type (synonymous – blue; nonsynonymous – red; grey – stop-codon). Wilcoxon signed-rank tests were performed to assess if the paired synonymous and nonsynonymous evolutionary rates are significantly distinct per individual gene segment or timepoint (annotated with "$*$" if $p < 0.05$). This was done for all sets of nonsynonymous and synonymous rate pairs except for those computed per day since symptom onset for A/H1N1pdm09 viruses due to the low number of data points available (median number of A/H1N1pdm09 virus samples collected per day since symptom onset = 2). Note that the scales of the y axes differ between A and B to better show rate trends.

For A/H1N1pdm09 viruses during the first wave of the pandemic, the overall within-host evolutionary rate was as high as $\sim 10^{-4}$ divergence per site per day in some samples on day 0 post-symptom onset (Figure 2B). We observed higher nonsynonymous evolutionary rates relative to synonymous ones initially after symptom onset but were unable to determine if they were significantly different due to the low number of samples (i.e. median = 2 samples per day post-symptom onset). In turn, we also could not meaningfully characterise the temporal trends of within-host evolution for the pandemic virus with this dataset. Nonetheless, consolidating over all samples across all time points, there was significantly higher rates of accumulation of nonsynonymous variants in the polymerase basic 2 (PB2), polymerase acidic (PA), HA and matrix (M) gene segments (Figure 2B, Figure 2 – figure supplement 2 and Figure 3 – figure supplement 2). All gene segments also yielded NS/S ratios > 1 (Supplemental File 1).

Intra-host minority variants

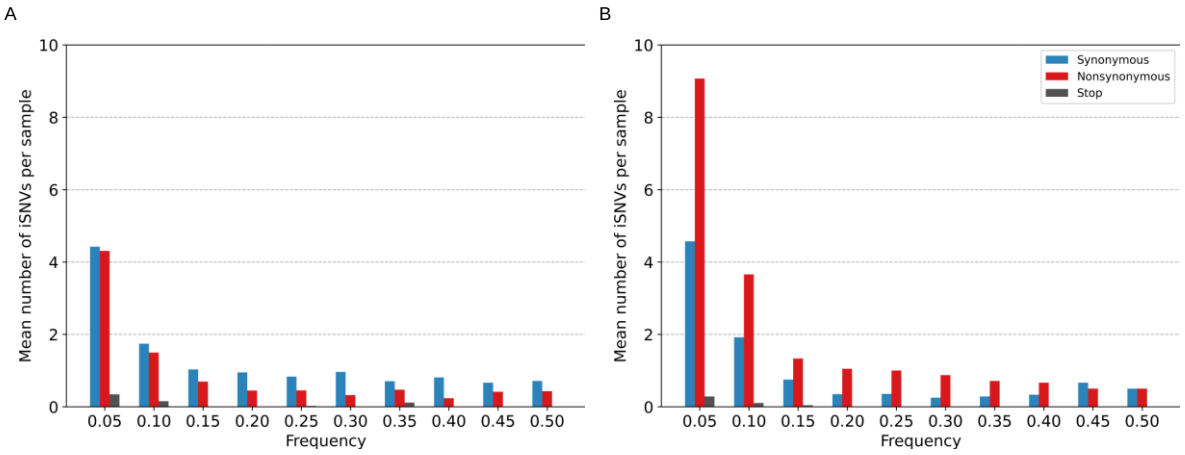


Figure 3. Histogram of the mean number of minority iSNVs identified per sample across all (**A**) A/H3N2 and (**B**) A/H1N1pdm09 virus specimens, sorted by frequency bins of 5% and substitution type (synonymous – blue; nonsynonymous – red; stop-codon – grey).

Most of the iSNVs identified for both virus subtypes were observed at low frequencies (2-5%; Figure 3), and appear to be stochastically introduced across the virus genome (Figure 4). Purifying selection dominated within-host seasonal A/H3N2 virus populations as the ratio of nonsynonymous to synonymous variants was 0.72 across all samples and variant frequencies (Figures 3A and Figure 3 – figure supplement 1). Of note, the canonical antigenic sites of the HA gene segment[27] of the A/H3N2 virus populations experienced strong negative selection as evidenced by the occurrence of synonymous variants (median frequency = 0.14, IQR range = 0.09-0.27) at far greater frequencies relative to those at non-antigenic sites of HA (median frequency = 0.03, IQR range = 0.03-0.05; Mann-Whitney U test $p = 1.18 \times 10^{-24}$; Figure 4C). There were no significant differences in the frequencies of nonsynonymous iSNVs between the antigenic sites of H3 (median frequency = 0.04, IQR range = 0.03-0.06) and the rest of the HA gene segment (median frequency = 0.03, IQR range = 0.02-0.06; Mann-Whitney U test $p = 0.29$; Figure 4C). In contrast, there was 1.94 times as many nonsynonymous minority iSNVs relative to synonymous ones identified in the pandemic A/H1N1pdm09 virus samples (Figures 3B and Figure 3 – figure supplement 2). Variant frequencies of nonsynonymous iSNVs found in the antigenic epitopes of H1[28] (median frequency = 0.04, IQR range = 0.04-0.05) were, however, not significantly different from those of non-antigenic sites (median frequency = 0.05, IQR range = 0.03-0.16; Mann-Whitney U test $p = 0.34$; Figure 4D).

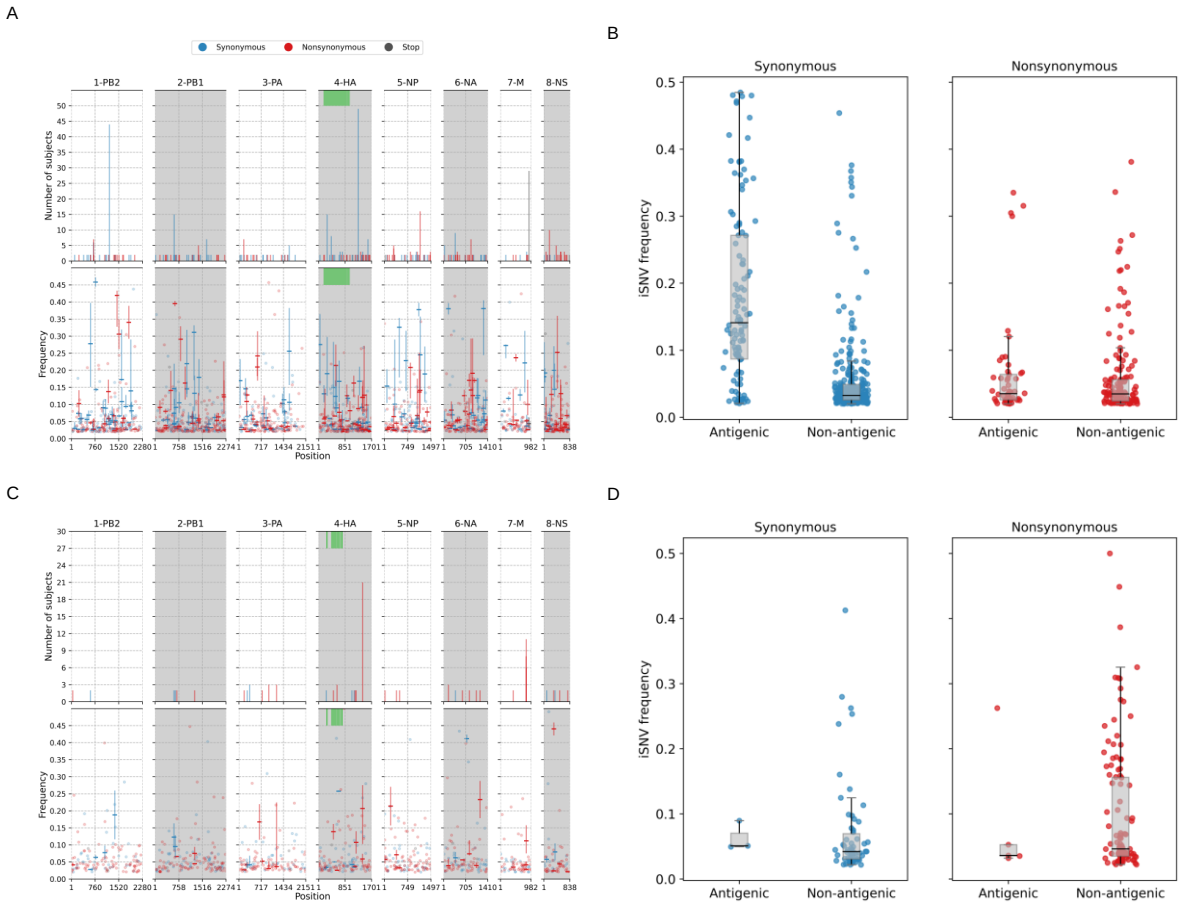


Figure 4: (**A**) Breakdown of iSNVs identified in seasonal A/H3N2 virus samples. The top panels plot the nucleotide positions where iSNVs were found in at least two subjects. The bottom panels shows the frequencies at which iSNVs were identified. For sites with iSNVs that were found in two or more subjects, the interquartile ranges of variant frequencies are plotted as vertical lines and the median frequencies are marked with a dash. If the iSNV was only found in one subject, its corresponding frequency is plotted as a circle. All iSNVs are stratified to either synonymous (blue), nonsynonymous (red) or stop-codon (grey) variants. Only the nonsynonymous variants are plotted if both types of variants are found in a site. Positions of antigenic sites of the haemagglutinin gene segment[29] are marked in green on the top panels. (**B**) Box plots of the frequencies of synonymous and nonsynonymous variants between antigenic and non-antigenic sites of seasonal A/H3N2 haemagglutinin gene segment. (**C**) Similar plots to (A) for iSNVs found in pandemic A/H1N1pdm09 virus samples. (**D**) Similar plots to (B) for HA iSNVs identified in the pandemic A/H1N1pdm09 virus samples.

As observed in a previous study using different data[26], premature stop-codon (nonsense) mutations accumulated within-host, though only at low rates. Here, we observed similarly low median nonsense rates, ranging between 0 and 1.29×10^{-6} divergence per site per day across the entire A/H3N2 virus genome over the course of infection (IQR limits range between 0 and at most, 1.82×10^{-6} divergence per site per day; Figure 2A). Premature stop-codons accumulated in the matrix (M) genes predominantly but also appeared in all other influenza gene segments within various individuals (Figures 2A and 4A). Nonsense mutations also accumulated within the A/H1N1pdm09 virus samples (Figure 2B). Similar to A/H3N2 viruses, nonsense mutation rates were much lower compared to the synonymous and nonsynonymous counterparts (median genome-wide rate across all samples between 0 and

1.43×10^{-6} divergence per site per day; IQR limits between 0 and 2.18×10^{-6} divergence per site per day).

The premature stop-codon mutations were mostly found at low frequencies for both influenza subtypes (<10%; Figure 3). The exception lies with one of the A/H3N2 virus samples where a premature stop codon was found in position 77 of the M2 ion channel with variant frequency as high as 34.6% (Patient 1843, day 6 since symptom onset; Figure 4A and Figure 4 – figure supplement 4). The premature stop codon in M2-77 was also found in 27 other individuals across multiple timepoints, albeit at a much lower frequency that never amounted more than 10% (Figures 4A and Figure 4 – figure supplement 4). This was unlikely to be a sequencing artefact resulting from a mistaken incorporation of the primer sequence as its carboxyl terminal falls outside the coding region of the M gene segment (Supplemental File 3) and the variant frequencies would have been much higher in all samples if this was the case.

Despite the dominance of purifying selection in seasonal A/H3N2 intra-host viral populations, we detected several nonsynonymous variants of interest. Amino acid variants emerging in the HA and NA proteins were discussed in a previous work[21] (Appendix A1). In the nucleoprotein, there were two notable nonsynonymous variants, D101N/G and G384R, that appeared in multiple individuals who were sampled independently between 2007 and 2009 (Figure 4A and Figure 4 – figure supplement 3). D101N/G was found in 7 different patients and at least for D101G, the mutation was previously linked to facilitating escape from MxA, a key human antiviral protein[30]. However, the nonsynonymous mutation was only found in low frequencies and remained invariant during the respective courses of infection for all seven patients (median variant frequency across all samples = 0.03; IQR = 0.02-0.07).

NP-G384R emerged in sixteen unlinked patients infected by A/H3N2 virus. Even though G384R did not become the majority variant in any of these individuals (median variant frequency across all samples = 0.14; IQR = 0.07-0.20), the variant emerged around day 4-5 post-symptom onset and mostly persisted within each individual for the rest of sampled timepoints. G384R is a stabilizing mutation in the A/Brisbane/10/2007 A/H3N2 virus NP background[31] that is similar to the viruses investigated here. Interestingly, position 384 is an anchor residue for several NP-specific epitopes recognised by specific cytotoxic T lymphocytes (CTLs) that are under continual selective pressure for CTL escape[32,33]. The wild-type glycine residue is known to be highly deleterious even though it was shown to confer CTL escape among HLA-B*2705-positive individuals[34–36].

Using a maximum likelihood approach to reconstruct and estimate the frequencies of the most parsimonious haplotypes of each gene segment, we computed linkage disequilibrium and found evidence of potential epistatic co-variants to NP-G384R in the A/H3N2 virus populations of multiple individuals (Figure 5 and Supplemental File 2). When analysing how these variants could alter protein stability using FoldX, the stabilizing effects of G384R (mean $\Delta\Delta G = -3.84$ kcal/mol (SD = 0.06 kcal/mol)) was found to alleviate the likely destabilizing phenotype of a functionally relevant linked variant in two of the three co-mutation pairs identified in separate individuals (i.e. G384R/M426I and G384R/G102R; Supplemental File 2). In the first individual (subject 1224), M426I was inferred to have emerged among the viral haplotypes encoding NP-G384R on the 10th day post-symptom onset (D10). M426I may be compensating for T-cell escape that was previously conferred by 384G even though the two amino acid sites are anchor residues of different NP-specific CTL epitopes[32]. M426I was found to be highly destabilizing (mean $\Delta\Delta G = 2.61$ kcal/mol (standard deviation (SD) = 0.05 kcal/mol); Table 1) but when co-mutated with G384R, stability changes to NP was predicted to be neutral (mean $\Delta\Delta G = -0.42$ kcal/mol (SD = 0.06 kcal/mol)). In the second individual (subject 1686), G102R was likely linked to G384R in the within-host virus populations found in the D10 sample. As a single mutant, G102R is also destabilizing to NP (mean $\Delta\Delta G = 4.87$ kcal/mol (SD = 0.00 kcal/mol)). However, when combined with G384R, NP protein stability was only weakly destabilizing (mean $\Delta\Delta G = 0.76$ kcal/mol (SD = 0.09 kcal/mol)). G102R was previously found to bypass the need for cellular factor importin-α7 which is crucial for viral replication and pathogenicity of IAVs in humans[37–39].

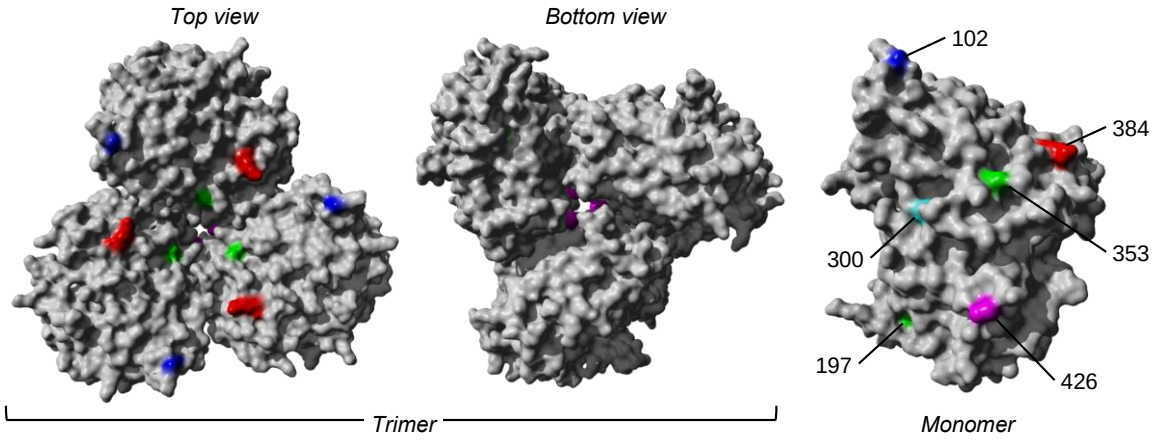


Figure 5: The trimeric and monomeric crystal structures of nucleoprotein (PDB: 3ZDP)[40] of influenza A viruses. Amino acid sites with potentially linked epistatic amino acid variants as tabulated in Table 1 are separately coloured, with their corresponding positions annotated on the monomeric structure.

	$\Delta\Delta G$ (kcal/mol)	
Variants	Mean	S.D.
G384R	-3.84	0.06
M426I	2.61	0.05
G384R,M426I	-0.42	0.06
G102R	4.87	0.00
G384R,G102R	0.76	0.09
A493T	11.96	0.30
G384R,A493T	5.56	0.19
V197I	-3.11	0.02
S353Y	-1.97	0.68
V197I,S353Y	-4.48	0.14

Table 1: FoldX stability predictions of likely linked nonsynonymous minority variants found in A/H3N2 nucleoprotein. The mean $\Delta\Delta G$ and standard deviation (S.D.) values reported are based on the results of five distinct simulations. Variants with mean $\Delta\Delta G < -0.46$ kcal/mol are deemed to be stabilizing while destabilizing mutants were estimated to yield $\Delta\Delta G > 0.46$ kcal/mol.

For the pandemic A/H1N1pdm09 viruses, most of the nonsynonymous variants were found singularly in individual patients (Figure 4B). Putative HA antigenic minority variants were found in four individuals in distinct amino acid sites (G143E, N159K, N197K and G225D; H3 numbering without signal peptide; Figure 4 – figure supplement 5). All of these variants were found at frequencies $\leq 5\%$ and the wild-type residues have been conserved in the corresponding positions globally to date, with the exception of position 225. Here, HA-225G was the majority variant (76%) in a hospitalised individual (subject 11-1022; Supplemental File 4) and D225G is linked to infections with severe disease outcomes[41]. Furthermore, one of the few nonsynonymous iSNVs that co-emerged in multiple unlinked patients was found in the usually conserved stem of the HA protein, L455F/I (H3 numbering without signal peptide), appearing in 17 separate individuals (Figures 4B and Figure 4 – figure supplement 5). The amino acid variant was found in patients from different time periods and geographical locations (Supplemental File 4), thus it is unlikely this was a unique variant shared among individuals in the same transmission cluster. It was observed as early as day 0 post-symptom onset for some patients and seemed to persist during the infection but only as a minority variant at varying frequencies (median frequency across all samples with mutation = 0.20; IQR = 0.08-0.28). However, this position has also been conserved with the wild-type Leucine residue in the global virus population to date. Hence, it is unclear if HA-L455F/I actually confers any selective benefit even though it was independently found in multiple patients.

We also found oseltamivir resistance mutation H275Y[42] in the NA proteins in two unlinked individuals who were infected with the A/H1N1pdm09 virus and treated with oseltamivir (Figure 4 – figure supplement 6 and Supplemental File 4). 275Y quickly became the majority variant in both patients within 3-4 days after the antiviral drug was first administered. Finally, there were two other amino acid variants in the M2 ion channel that appeared within multiple subjects in parallel across different geographical locations – L46P and F48S were identified in 8 and 16 patients respectively in a range of frequencies (L46P: median frequency = 0.04, IQR = 0.04-0.05; F48S: median frequency = 0.08, IQR = 0.03-0.13) but similarly, never becoming a majority variant in any of them (Figures 4 and Figure 4 – figure supplement 7). Again, the wild-type residues were mostly conserved in the global virus population since the pandemic.

Within-host simulations

To investigate the evolutionary pressures that likely underpin the observed within-host dynamics of A/H3N2 viruses in young children (Figure 2), we performed forward-time Monte Carlo simulations. Given that the median age of the children infected by A/H3N2 virus at the time of sample collection was 2 years of age (IQR=2-3 years), most of them were likely experiencing one of their first influenza virus infections. Furthermore, influenza vaccination for children is not part of the national vaccination programme in Vietnam. As such, most of the children analysed here lacked influenza virus specific antibodies based on haemagglutination inhibition assays[21]. Since seasonal A/H3N2 viruses have circulated within the human population since 1968, the virus is well adapted to human hosts at this point such that most nonsynonymous mutations are likely highly deleterious and would not reach detectable frequencies. We hypothesized that detected variants are mostly expected to be weakly deleterious, and thus not purged fast enough by selection such that mutation-selection balance was observed.

Our simulations used a simple within-host evolution model represented by a binary genome that distinguishes between synonymous and nonsynonymous loci. Given that the estimated transmission bottleneck sizes for seasonal A/H3N2 viruses[4,43] are narrow at 1-2 genomes, we modelled an expanding virus population size during the initial timepoints of the infection that started with one virion. If within-host virus populations were to evolve neutrally, we would observe similar synonymous and nonsynonymous evolutionary rates throughout the infection (Figure 6A). On the other hand, if negative selection is sufficiently strong, accumulation of deleterious nonsynonymous variants will decrease substantially with time (Figure 6B). Clearly, these patterns were not observed for A/H3N2 viruses (Figure 2A). However, if most *de novo* nonsynonymous mutations are only weakly deleterious, we would observe larger synonymous evolutionary rates initially before nonsynonymous variants accumulate to

similar levels (Figure 6C). By then, virion population size (N) would also be large enough relative to the virus mutation rate (μ) (i.e. $N\mu \gg 1$; Appendix A5) such that mutation-selection balance is expected and evolutionary rates remain fairly constant, similar to the patterns empirically observed for within-host A/H3N2 virus populations (Figure 2A).

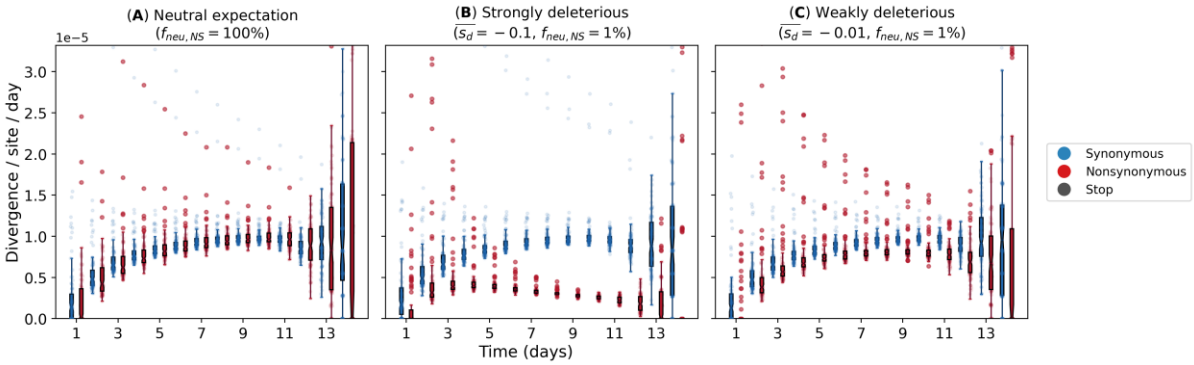


Figure 6: Evolutionary rates computed from forward-time Monte Carlo within-host simulations for different mean deleterious effects ($\overline{s_d}$) of nonsynonymous mutations . We assumed that synonymous mutations are neutral for all simulations. (**A**) Neutral expectation where all nonsynonymous mutations are neutral ($f_{neu,NS} = 100\%$). We tested our hypotheses where majority of nonsynonymous mutations are non-neutral ($f_{neu,NS} = 1\%$) and they are either (**B**) strongly ($\overline{s_d} = -0.1$) or (**C**) weakly ($\overline{s_d} = -0.01$) deleterious.

Discussion

Multiple next-generation sequencing studies have found little evidence of positive selection in seasonal influenza virus populations of acutely infected individuals[4,8–11,44]. Recent modelling work showed that the time required to initiate new antibody production and asynchrony with virus exponential growth limits the selection of *de novo* antigenic variants within host in acute seasonal influenza virus infections[45]. In contrast, phenotypically relevant variants that were positively selected in within-host virus populations of severely immunocompromised patients coincided with those selected by the global seasonal IAV population[46,47]. This implies that within-host evolutionary dynamics of seasonal IAVs in immunocompromised individuals are likely to be substantially different owing to the increased time for virus diversity to accumulate and for selection to act[48]. In other words, the duration of infection is likely to be critical for positive evolutionary selection to be effective within host.

Viral shedding duration is often longer in young children infected with seasonal influenza virus compared to otherwise healthy adults[49]. Children also play a critical role in "driving" influenza epidemics due to their higher contact and transmission rates[22,23]. As such, our seasonal A/H3N2 virus results fill an important gap in the current literature of within-host evolutionary studies of seasonal IAVs as most of the samples analysed were collected from

children under the age of six years up to two weeks post-symptom onset. Importantly, the absence of antibody-mediated immunity in young unvaccinated children, which would otherwise reduce the extended duration of infection, has the potential to facilitate other routes of virus evolution.

Similar to the aforementioned within-host studies, the A/H3N2 virus population within these children was characterised by low genetic diversity and dominated by purifying selection early in the infection. Due to a lack of antibody response against the antigenic regions of HA[21], it is unsurprising that we observed a lack of adaptive changes to the HA antigenic regions, similar to adults in previous studies[4]. We also found that the polymerase genes were subjected to purifying selection, indicating their critical role in virus replication as negative selection purges deleterious variation. However, while purifying selection is detectable, it is incomplete[26]. We observed that most nonsynonymous variants began to accumulate around 3-4 days post-symptom onset, with incrementally higher empirical rates as the infection progressed.

Through simulations of a within-host evolution model, we investigated the hypothesis that in the absence of any positive selection, the accumulation of nonsynonymous iSNVs was a result of their neutral or only weakly deleterious effects and the expanding within-host virion population size during later timepoints in longer infections of naïve young children such that mutation-selection balance was reached. In contrast, this balance was not detected in otherwise healthy older children or adults with short-lived influenza virus infections lasting no more than a week where *de novo* nonsynonymous iSNVs are rarely found [4,8–11,44]. The maintenance of genetic diversity through mutation-selection balance within these young children may provide opportunities for the emergence of phenotypically relevant mutations which deleterious effects could be alleviated by the accumulation of a secondary compensatory mutations. For example, in one individual NP-G384R was accompanied by NP-M426I which is an anchor residue of a CTL epitope of NP, abrogating recognition by HLA-B*3501-positive CTLs[32] but is likely to be deleterious based on our computational protein stability predictions. G384R, which is located in a CTL epitope distinct from M426I[32], was previously shown to be a stabilizing substitution[31].

Interestingly, we also observed G384R in the minority virus population of 15 other unlinked individuals. Besides improving NP protein stability, G384R restores recognition by HLA-B*2705-positive NP-specific CTLs[36]. The NP gene segment in the global A/H3N2 virus population has an evolutionary history of fixating destabilizing amino acid mutations that promote CTL immune escape alongside stabilizing substitutions that compensate for the deleterious effects of the former[34]. The reversal R384G mutation confers CTL escape but is

known to be highly deleterious. This substitution was fixed in the global A/H3N2 virus population during the early 1990s as other substitutions such as S259L and E375G epistatically alleviated its destabilizing effects[34]. One possible explanation for the emergence of G384R as a minority variant within these unlinked individuals is that they are all HLA-B*2705 negative. However, we did not collect the necessary blood samples to investigate this possibility.

In contrast, we found a substantially higher fraction of nonsynonymous variants in the within-host virus populations of individuals infected A/H1N1pdm09 virus during the pandemic. Owing to the low number of A/H1N1pdm09 virus samples and different next-generation sequencing platforms used to sequence samples of the two virus subtypes and consequently differences in base calling error rates and depth of coverage (Figure 1 – figure supplement 1), we were unable to directly compare the observed levels of within-host genetic diversity and its temporal trends between the two influenza subtypes here. However, given that only iSNVs with frequencies $\geq 2\%$ were called, low-frequency minority variants arising from technical-related errors should be minimised[50]. Importantly, the relative number of nonsynonymous iSNVs identified were far greater than synonymous ones in the pandemic A/H1N1pdm09 virus infections, suggesting that there was room for further human host adaptation, particularly in the HA but also in the polymerase gene segments similar to those observed in other zoonotic influenza virus infections[51].

Given the tight estimated transmission bottleneck size (Appendix A4), the relatively large number of iSNVs identified at the start of symptom onset and simulations of within-host evolution (Figure 1B, 2B and 6D), it is unlikely that the initial within-host A/H1N1pdm09 virus populations sampled were the inoculating population that founded the infection. Instead, the inoculating viral population had already undergone substantial within-host replication during the incubation period before symptom-onset. In fact, four of the individuals analysed were asymptomatic (i.e. H058/S02, H089/S04, H186/S05 and H296/S04; Supplemental File 4). Additionally, pre-symptomatic virus shedding was observed in some of the secondary household cases[52] and presymptomatic transmission has been documented in other settings[53]. Nonetheless, this would not meaningfully impact our conclusions as most of the within-host viral populations sampled at the start of symptom onset should still constitute those found early in infection and the contrasting feature where nonsynonymous iSNVs outnumbered synonymous ones were not observed in the seasonal A/H3N2 virus samples.

For both A/H3N2 and A/H1N1pdm09 virus samples, nonsense iSNVs resulting in premature stop codons were found to accumulate within host, even though only at low proportions. The accumulation of premature stop-codon mutations further suggest that while purifying

selection dominates within-host influenza virus populations, it may not be acting strongly enough to completely purge these lethal nonsense mutations[26]. Additionally, it has been recently found that incomplete influenza virus genomes frequently occur at the cellular level and that efficient infection depends on the complementation between different incomplete genomes[54]. As such, nonsense mutations may not be as uncommon as previously thought. In particular, nonsense mutations in position 77 of the M2 ion channel were independently found in 27 unlinked individuals infected by A/H3N2 virus. While these nonsense mutations are generally considered to be lethal, ion channel activity is retained even if the M2 protein was prematurely truncated up to position 70 at its cytoplasmic tail[55].

Our study has several limitations. The number of iSNVs identified can potentially be biased by variations in sequencing coverage[56]. As such, the number of iSNVs observed in one intra-host virus populations may not be directly comparable to another with a distinct coverage profile (Figure 1 – figure supplement 1). As an alternative, the nucleotide diversity π statistic[57] may be a more robust measure of within-host diversity as it solely depends on the underlying variant frequencies[56]. Computing the corresponding π statistics for our data, we observed trends in genetic diversity that were similar to those inferred using iSNV counts (Appendix A2 and Appendix – Figure 1).

To ensure accurate measurements of virus diversity in intra-host populations, we would also need to be certain that the estimated variant frequencies precisely reflect the distributions of variants that comprise the sampled virus populations. The inferred variant frequencies can be significantly distorted if virus load is low[58,59]. As such, we limited our analyses for both virus subtypes to samples with Ct-values $\leq$35 which likely afford sufficient virus material for sequencing[59]. We were unable to estimate the amount of frequency estimation errors for the A/H1N1pdm09 virus samples as only one sequencing replicate was performed using the universal 8-segment PCR method[60]. However, for the A/H3N2 virus samples, independent PCR reactions were performed using three partly overlapping amplicons for all gene segments other than the non-structural and matrix genes. We compared the variant frequencies estimated for any overlapping sites generated by reads derived from distinct amplicons with sufficient coverage (>100x). Variant frequencies computed from independent amplicons agreed well with each other across the range of Ct values of the samples from which variants were identified (Figure 1 – figure supplement 2), affirming the precision of our iSNV frequency estimates for the A/H3N2 virus samples, including those with higher Ct values.

We also performed additional checks to ensure that our results were not driven by potential PCR and/or technical artefacts. First, we excluded all iSNVs found under the 75th percentile

of frequency range of A/H3N2 variants that were found in only one of the overlapping amplicons. We then recomputed the daily within-host evolutionary rates with the remaining iSNVs (Figure 2 – figure supplement 3) and found that the relative temporal trends in synonymous and nonsynonymous rates remain similar to those in Figure 2A. We also checked that the distributions of frequencies for iSNVs found in recurrent mutation sites (i.e. NP-384 and M2-77) that are below variant calling threshold are comparable to those found in their neighbouring sites (±10 nucleotide positions; Figure 4 – figure supplement 8). Furthermore, we remapped the sample reads to their respective consensus sequences to minimize mapping of technical artefacts. We were still able to detect the recurring NP-G384R and M2-R77* amino acid mutations in multiple individuals and timepoints at similar frequencies when mapped to the reference genome (Figure 4 – figure supplement 4-5 and 9). As such, these recurrent mutations are unlikely to have been resulted from erroneous variant calls of artefacts.

Finally, most study participants received oseltamivir during the course of their infections (Supplemental File 4). Although we were unable to identify any potential effects of enhanced viral clearance or any other evolutionary effects due to the treatment, besides oseltamivir-resistance associated mutations, it is unlikely that the antiviral treatment had a substantial impact on our results. First, the median timepoint in which the antiviral was initially administered was 4 days post-symptom onset (IQR = 3-6 days; Supplemental File 4). Previous studies showed that enhanced viral clearance of IAVs was mostly observed among patients who were treated with oseltamivir within 3 days of symptom onset[20,61,62]. Of note, late timepoint samples in this study (≥ 8 days since symptom onset) mostly came from individuals who started oseltamivir treatments ≥ 4 days post-symptom onset (Figure 1 – figure supplement 5). Second, at least *in vitro*, there were no differences in the levels of genetic diversity observed in influenza virus populations after multiple serial passages whether they were treated with oseltamivir or not[63].

To conclude, we presented how intra-host populations of seasonal and pandemic influenza viruses are subjected to contrasting evolutionary selection pressures. In particular, we showed that the evolutionary dynamics and ensuing genetic variation of these within-host virus populations changes during the course of infection, highlighting the importance for sequential sampling, particularly for longer-than-average infections such as those in the young children studied here.

Methods

Sample collection and viral sequencing

The A/H3N2 virus samples were collected from 52 patients between August 2007 and September 2009 as part of an oseltamivir dosage trial conducted by the South East Asia Infectious Disease Clinical Research Network (SEAICRN), which is detailed in a previous work[20]. Briefly, patients with laboratory confirmed influenza virus infection and duration of symptoms ≤ 10 days were swabbed for nose and throat samples daily between 0 and 10 days as well as day 14 upon enrolment for the study (Supplemental File 4). All PCR-confirmed A/H3N2 virus samples with cycle threshold (Ct) values ≤ 35 were included for sequencing.

Library preparation and viral sequencing protocols performed on these A/H3N2 virus samples are elaborated in detail in [21]. Here, we highlight key aspects of our preparation and sequencing procedures. Using segment specific primers (Supplemental File 3), we performed six independent PCR reactions, resulting in three partly-overlapping amplicons for each influenza virus gene segment other than the matrix (M) and non-structural (NS) genes where a single amplicon was produced to cover the entirety of the relatively shorter M and NS genes. The use of shorter but overlapping amplicons in the longer gene segments improve amplification efficiency, ensuring that these longer segments are sufficiently covered should there be any RNA degradation in the clinical specimen. These overlapping PCR products were pooled in equimolar concentrations for each sample and purified for subsequent library preparation. Sequencing libraries were prepared using the Nextera XT DNA Library Preparation kit (Illumina, FC-131-1096) as described in [21]. Library pools were sequenced using the Illumina MiSeq 600-cycle MiSeq Reagent Kit v3 (Illumina, MS-102-3003).

The A/H1N1pdm09 virus samples were obtained as part of a household-based influenza virus cohort study that was also performed by SEAICRN. The study was conducted between July and December 2009, involving a total of 270 households in Ha Nam province, Vietnam[64]. Similarly, combined nose and throat swabs were collected daily for 10-15 days from individuals with influenza-like-illness (i.e. presenting symptoms of fever >38°C and cough, or sore throat) and their household members, including asymptomatic individuals (Supplemental File 4). We also analysed additional samples collected from unlinked hospitalised patients who were infected by the A/H1N1pdm09 virus from two major Vietnamese cities (Hanoi and Ho Chih Minh) during the first wave of the pandemic[20,25]. A total of 32 PCR-confirmed A/H1N1pdm09-infected individuals originating from both households and hospitalised cases were selected for sequencing based on availability and Ct-values ≤ 33 (Supplemental File 4).

For the A/H1N1pdm09 virus samples, RNA extraction was performed manually using the High Pure RNA isolation kit (Roche) with an on-column DNase treatment according to the manufacturer's protocol. Total RNA was eluted in a volume of 50 µl. Universal influenza

virus full-genome amplification was performed using a universal 8-segment PCR method as described previously[65–67]. In short, two separate RT-PCRs were performed for each sample, using primers common-uni12R (5'-GCCGGAGCTCTGCAGAT ATCAGCRAAAGCAGG-3'), common-uni12G (5'-GCCGGAGCTCTG CAGATATCAGCGAAAGCAGG-3'), and common-uni13 (5'-CAGGAA ACAGCTATGACAGTAGAAACAAGG-3'). The first RT-PCR mixture contained the primers common-uni12R and common-uni13. The second RT-PCR mixture contained the primers common-uni12G and common-uni13, which greatly improved the amplification of the PB2, PB1, and PA segments. Reactions were performed using the One-Step RT-PCR kit High Fidelity (Invitrogen) in a volume of 50 µl containing 5.0 µl eluted RNA with final concentrations of 1xSuperScript III One-Step RT-PCR buffer, 0.2 µM of each primer, and 1.0 µl SuperScript III RT/Platinum Taq High Fidelity Enzyme Mix (Invitrogen). Thermal cycling conditions were as follows: reverse transcription at 42°C for 15 min, 55°C for 15 min, and 60°C for 5 min; initial denaturation/enzyme activation of 94°C for 2 min; 5 cycles of 94°C for 30 s, 45°C for 30 s, slow ramp (0.5°C/s) to 68°C, and 68°C for 3 min; 30 cycles of 94°C for 30 s, 57°C for 30 s, and 68°C for 3 min; and a final extension of 68°C for 5 min. After the PCR, equal volumes of the two reaction mixtures were combined to produce a well-distributed mixture of all 8 influenza virus segments. All RT-PCRs were performed in duplicate. Samples were diluted to a DNA concentration of 50 ng/µl followed by ligation of 454 sequencing adaptors and molecular identifier (MID) tags using the SPRIworks Fragment Library System II for Roche GS FLX+ DNA Sequencer (Beckman Coulter), excluding fragments smaller than 350 base pairs, according to the manufacturers protocol to allow for multiplex sequencing per region. The quantity of properly ligated fragments was determined based on the incorporation efficiency of the fluorescent primers using FLUOstar OPTIMA (BMG Labtech). Emulsion PCR, bead recovery and enrichment were performed manually according to the manufacturers protocol (Roche) and samples were sequenced in Roche FLX+ 454. Sequencing was performed at the Sanger Institute, Hinxton, Cambridge, England as part of the FP7 program EMPERIE. Standard flowgram format (sff) files containing the filter passed reads were demultiplexed based on the molecular identifier (MID) sequences using QUASR package version 7.0[50].

Read mapping

Trimmomatic (v0.39; Bolger et al. 2014) was used to discard reads with length <30 bases while trimming the ends of reads where base quality scores fall below 20. The MAXINFO option was used to perform adaptive quality trimming, balancing the trade-off between longer read length and tolerance of base calling errors (target length=40, strictness=0.4). For the A/H3N2 virus samples, the trimmed paired reads were merged using FLASH (v1.2.11)[69]. All remaining reads were then locally aligned to A/Brisbane/10/2007 genome (GISAID accession: EPI_ISL_103644) for A/H3N2 virus samples and A/California/4/2009 genome (EPI_ISL_376192) for A/H1N1pdm09 virus samples using Bowtie2 (v2.3.5.1)[70]. Aligned

reads with mapping scores falling below 20 alongside bases with quality score (*Q-score*) below 20 were discarded.

Variant calling and quality filters

Minority variants of each nucleotide site with a frequency of at least 2% were called if the nucleotide position was covered at least 50x (H1N1pdm09) or 100x (H3N2) and the probability that the variant was called as a result of base calling errors (p_{Err}) was less than 1%. p_{Err} was modelled by binomial trials[71]:

$$p_{Err} = \sum_{i=n}^{N} \binom{N}{i} p_e^i (1-p_e)^{N-i}$$

where $p_e = -10^{-\frac{Q-score}{10}}$, N is the coverage of the nucleotide site in question and n is the absolute count of the variant base tallied.

While lower coverage at both ends of individual gene segments was expected, there were also variable coverage results across gene segments for some samples that were mapped to A/H3N2 virus (Figure 1 – figure supplement 1). In order to retain as many samples deemed to have adequate coverage across whole genome, a list of polymorphic nucleotide sites found to have >2% minority variants in more than 1 sample was compiled. Each gene segment of a sample was determined to achieve satisfactory coverage if >70% of these polymorphic sites were covered at least 100x. For A/H1N1pdm09, the gene segment of a sample was deemed to be adequately covered if 80% of the gene was covered at least 50x.

The number of iSNVs observed in A/H3N2 virus samples collected from subject 1673 (39-94 iSNVs in three samples collected from three (D3) to five (D5) days post-symptom onset) and the D8 sample for subject 1878 (73 iSNVs) were substantially greater than numbers in all other samples. The putative majority and minority segment-concatenated sequences of these samples did not cluster as a monophyletic clade among themselves phylogenetically (Figure 1 – figure supplement 3), suggesting that these samples might be the product of mixed infections or cross-contamination. These samples were consequently excluded from further analyses.

Empirical within-host evolutionary rate

The empirical within-host evolutionary rate ($r_{g,t}$) of each gene segment (g) in a sample collected on t day(s) since symptom onset were estimated by:

$$r_{g,t} = \frac{\sum_i^{n_{g,t}} f_{g,t,i}}{n_{g,t} \cdot t}$$

where $f_{g,t,i}$ is the frequency of minority variants present in nucleotide site i for gene segment g and $n_{g,t}$ is the number of all available sites[26]. Distinct rates were calculated for synonymous and non-synonymous iSNVs. If a variant was found in overlapping reading frames and a nonsynonymous change was observed in any of those frames, it would be accounted for as a nonsynonymous mutation. The corresponding whole-genome evolutionary rate (r_t) on day t is computed by summing the rates across all gene segments:

$$r_t = \sum_g r_{g,t}$$

Haplotype reconstruction

The most parsimonious viral haplotypes of each gene segment were reconstructed by fitting the observed nucleotide variant count data to a Dirichlet multinomial model using a previously developed maximum likelihood approach to infer haplotype frequencies[43]. Assuming that the viral population is made up of a set of K haplotypes with frequencies $\boldsymbol{q_k}$, the observed partial haplotype frequencies $\boldsymbol{q_l}$ for a polymorphic site l can be computed by multiplying a projection matrix $\boldsymbol{T_l}$. For instance, if the set of hypothetical full haplotypes is assumed to be $\{AA, GA, AG\}$, the observed partial haplotype frequencies for site $l = 1$, q_{A-} and q_{G-} are computed as:

$$\boldsymbol{q_l} = \boldsymbol{T_l q_k} \Rightarrow \begin{bmatrix} q_{A-} \\ q_{G-} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} q_{AA} \\ q_{GA} \\ q_{AG} \end{bmatrix}$$

A list of potential full haplotypes was generated from all combinations of nucleotide variants observed in all polymorphic sites of the gene segment. Starting from $K = 1$ full haplotype, the optimal full haplotype frequency q_k is inferred by maximizing the likelihood function:

$$LL = \sum_l \log \mathcal{L}(\boldsymbol{x_l} | \boldsymbol{T_l q_k}, \varphi)$$

where $\mathcal{L}$ is Dirichlet multinomial likelihood, $\boldsymbol{x_l}$ is the observed variant count data for read type l and φ is the overdispersion parameter, assumed to be 1×10^{-3}. Simulated annealing was used to optimise the haplotype frequencies by running two independent searches for at least 5000 states (iterations) until convergence was reached. In each state, the distribution of $\boldsymbol{q_k}$ was drawn from a Gaussian distribution centered at the frequency distribution of the previous state with a standard deviation of 0.05. One additional haplotype was added to the set of K full haplotypes during each round of optimization.

The resulting *K* haplotypes reconstructed depend on the order in which the list of potential full haplotypes is considered. As mentioned above, paired-end reads were merged to produce longer reads (up to ~500-600 base pairs) for mapping in the case of the seasonal A/H3N2 virus samples. Additionally, the single-stranded A/H1N1pdm09 viral reads generated from 454 sequencing can be as long as ~500 base pairs. Consequently, there was a non-trivial number of reads where co-mutations were observed in multiple polymorphic sites. Since iSNV frequencies are generally low, haplotypes with co-mutating sites would inevitably be relegated to end of the list order if ranked by their expected joint probabilities. As such, the list of full potential haplotypes was ordered in descending order based on the score of each full haplotype set k (s_k):

$$s_k = f_{ss,k} \times f_{ms,k}$$

where $f_{ss,k}$ and $f_{ms,k}$ are both joint probabilities of the full haplotype k computed in different ways. $f_{ss,k}$ is the expected joint probability frequency calculated from the observed independent frequencies of each variant for each polymorphic site found in the full haplotype k. $f_{ms,k}$ is based on the observed frequencies of variants spanning across the sets of highest hierarchal combination of polymorphic sites ($f_{ms,k}$).

For example, given a segment where iSNVs were found in three sites, the following reads were mapped: (A, A, C), (T, A, C), (A, T, C), (A, C, –), (–, A, C) and (–, T, C). We can immediately see that the top hierarchal combination of polymorphic sites (i.e. possible haployptes) are (A, A, C), (T, A, C) and (A, T, C) (i.e. we would compute $f_{ms,(A,A,C)}$, $f_{ms,(T,A,C)}$ and $f_{ms,(A,T,C)}$ respectively). The observed number of reads with (–, A, C) will counted towards the computation of both $f_{ms,(A,A,C)}$ and $f_{ms,(T,A,C)}$ since they could be attributed to either haplotype. Similarly, reads with (–, T, C) will be absorbed towards the counts to compute $f_{ms,(A,T,C)}$. Finally, we see that reads with (A, C, –) are not a subset of any of the top hierarchal haplotypes considered. As such, they form the 4th possible top hierarchal haplotype on its own. As such, if we were to compute the ranking for haplotype (A, A, C):

$$\begin{aligned} s_{(A,A,C)} &= f_{ss,(A,A,C)} \times f_{ms,(A,A,C)} \\ &= \left\{f_{(A,-,-)} \times f_{(-,A,-)} \times f_{(-,-,C)}\right\} \times f_{ms,(A,A,C)} \end{aligned}$$

If any nucleotide variants in the observed partial haplotypes were unaccounted for in the current round of full haplotypes considered, they were assumed to be generated from a cloud of "noise" haplotypes that were present in no more than 1%. Bayesian information criterion (BIC) was computed for each set of full haplotypes considered and the most parsimonious set of *K* haplotypes was determined by the lowest BIC value.

Linkage disequilibrium

Using the estimated frequencies of the most parsimonious reconstructed haplotypes, conventional Lewontin's metrics of linkage disequilibrium were computed to detect for potential epistatic pairs of nonsynonymous variants:

$$LD_{ij} = \hat{q}_{ij} - \hat{q}_i\hat{q}_j$$

where $\hat{q}_i$ and $\hat{q}_j$ are the estimated site-independent iSNV frequencies of sites i and j respective while $\hat{q}_{ij}$ is the frequency estimate of variants encoding co-variants in both i and j. Dividing LD by its theoretical maximum normalises the linkage disequilibrium measure:

$$LD' = \frac{LD}{LD_{max}}$$

$$LD_{max} = \begin{cases} \max\{-\hat{q}_i\hat{q}_j, -(1-\hat{q}_i)(1-\hat{q}_j)\} & if\ LD > 0 \\ \min\{\hat{q}_i(1-\hat{q}_j), (1-\hat{q}_i)\hat{q}_j\} & if\ LD < 0 \end{cases}$$

FoldX analyses

FoldX (https://foldxsuite.crg.eu/) was used to estimate structural stability effects of likely linked nonsynonymous minority variants found in the nucleoprotein (NP) of within-host A/H3N2 virus populations. At the time of writing of this paper, there was no A/H3N2-NP structure available. Although the eventual NP structure (PDB: 3ZDP) adopted for stability analyses was originally derived from H1N1 virus (A/WSN/33)[40], it was the most well resolved (2.69Å) crystal structure available, with 78.5% amino acid identity relative to the NP protein of A/Brisbane/10/2007. Previous work has shown that mutational effects predicted by FoldX using a NP structure belonging to A/WSN/33 (H1N1) was similar to those experimentally determined on a A/Brisbane/10/2007 nucleoprotein[31]. FoldX first removed any potential steric clashes to repair the NP structure. It then estimated differences in free energy changes as a result of the input amino acid mutation (i.e. $\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild-type}$) under default settings (298K, 0.05M ionic strength and pH 7.0). Five distinct simulations were made to estimate the mean and standard deviation $\Delta\Delta G$ values.

Within-host simulations

We implemented forward-time Monte Carlo simulations with varying population size using a simplified within-host evolution model to test if our hypotheses could explain the different evolutionary dynamics observed between A/H3N2 and A/H1N1 viral populations. We

assumed that a single virion leads to a productive influenza virus infection within an individual and computed changes in the virus population size (N) using a target cell-limited model. New virions are produced upon infection by existing virions at a rate of βCN where C is the existing number of target cells while β is the rate of per-cell per-virion infectious contact. Upon infection, a cell will produce r number of virions before it is rendered unproductive. We assume that infected individuals did not mount any antibody-mediated immune response, setting the virus' natural per-capita decay rate (d) such that virions continue to be present within host for 14 days (Figure 6 – figure supplement 1 and Table 2). β is then computed by fixing the within-host basic reproduction number (R_0):

$$R_0 = \frac{\beta C_0 r}{d}$$

where C_0 is the initial (maximum) target cell population size. We solve the following system of ordinary differential equations numerically to compute the number of virions per viral replicative generation ($N(t)$):

$$\frac{dC}{dt} = -\beta CN$$

$$\frac{dN}{dt} = \beta CN - dN$$

We assume a binary genome of length L, distinguishing between synonymous and nonsynonymous loci. For A/H3N2 viruses, we hypothesised that most *de novo* mutations are either weakly deleterious or neutral. To estimate the number of such sites, we aligned A/H3N2 virus sequences that were collected between 2007 and 2012 and identified all polymorphic sites with variants that did not fixate over time (i.e. <95% frequency over one-month intervals). We estimated $L = 1050$ with 838 and 212 synonymous and nonsynonymous loci respectively.

We tracked the frequency distribution of genotypes present for every generation t. We assumed that mutations occur at per-locus, per-generation rate μ. During each generation t, the number of virions incurring a single-locus mutation followed a Poisson distribution with mean $N(t)\mu L$. For each virion, the mutant locus was randomly selected across all loci. We assumed that all synonymous and a fraction of nonsynonymous sites ($f_{neu,NS}$) are neutral (i.e. (log) fitness effect $s = 0$). The remaining nonsynonymous sites either had an additive deleterious (s_d) or beneficial (s_b) fitness effect when mutated. The magnitude of s_d/s_b follow an exponential distribution with mean effect $|\bar{s}|$. Epistasis was neglected throughout. The distribution of genotypes in the next generation $t + 1$ was achieved by resampling

individuals according to Poisson distribution with mean $N(t+1)P_f(g,t)$ where $P_f(g,t)$ is the relative fitness distribution of genotype g during generation t.

To decrease the computational costs of the simulations, specifically when $N(t)$ reaches orders of $10^{10}-10^{11}$ virions (Figure 6 – figure supplement 1), we implemented an upper population size limit of 10^7 virions. Given the mutation rate assumed (Table 2), $N(t)\mu \gg 1$ for $N(t) \geq 10^7$ virions, mutation-selection balance is theoretically expected for a single-locus (deleterious) mutant model (Appendix A5). We ran 500 simulations for each variable set of $f_{neu,NS}$ and s_d/s_b values. All parameter values used in the model are given in Table 2.

Table 2: Parameter values used in within-host model

Parameter	Meaning	Value (units)	Source
-	Number of hours per replicative generation	6 hours	Assumption
r	Average number of virions produced by an infected cell	100 virions	[72]
C_0	Initial target cell population size	4×10^8 virions	[73]
d	Per-capita decay rate	2 per-generation	Assumption
R_0	Within-host basic reproduction number	5	[73]
μ	Per-site, per-generation mutation rate	3×10^{-5} per-site, per-generation	[44]

Phylogenetic inference

All maximum likelihood phylogenetic trees were reconstructed with IQTREE (v. 1.6.10)[74], using the GTR+I+G4 nucleotide substitution model.

Appendix

A1. Haemagglutinin and neuraminidase minority variants in A/H3N2 virus samples

The amino acid variants emerging in the haemagglutinin (HA) and neuraminidase (NA) proteins of A/H3N2 virus samples were discussed in a previous work[21]. Briefly, given the lack of antibody-mediated immune response in this cohort of mostly naïve children, HA amino acid variants emerging in putative antigenic sites were generally low in frequencies (median frequency = 0.04, IQR=0.03-0.06) and all only became detectable 3-4 days post

symptom onset (Figure 4 – figure supplement 1). Notably, two of these intra-host mutations were also found in the global A/H3N2 virus population in high frequencies: HA-S45N and -D53N (H3 numbering without signal peptide). Both mutations are part of the canonical antigenic site C of H3 and emerged within separate individuals, with D53N eventually becoming the majority variant (97%) as late as day 13 post-symptom onset in one of them. Oseltamivir-resistant amino acid mutations E119V, R292K and N329K arose in 10 patients that were treated with the antiviral drug and mostly rose in within-host frequencies only 4-7 days after administration of oseltamivir (Figure 4 – figure supplement 2).

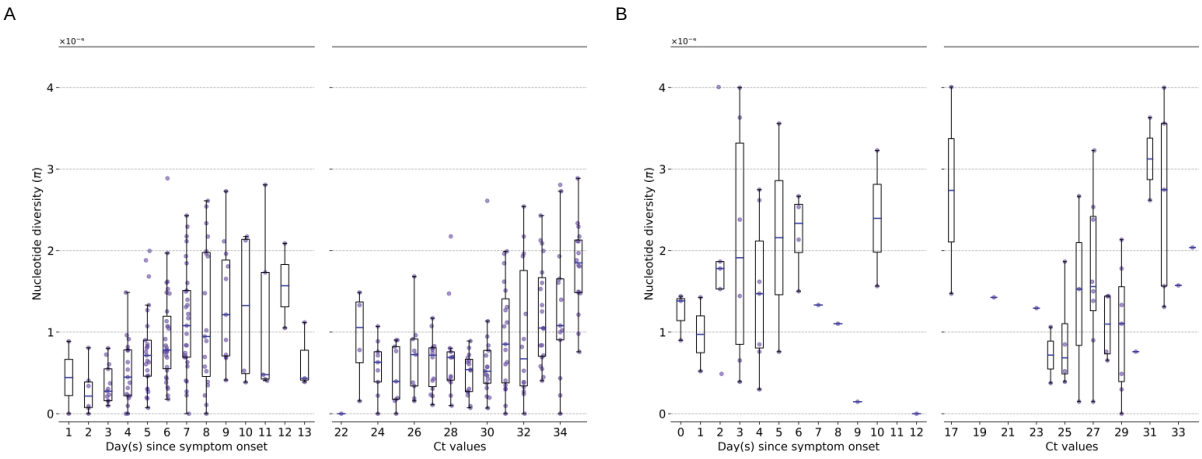
A2. Genetic diversity by π statistic

Given that the number of identified iSNVs can potentially be biased by variations in sequencing coverage, genetic diversity was also assessed using nucleotide diversity π statistic[57]. This approach constitutes a more robust measure of within-host diversity as it is solely dependent of the underlying variant frequencies. While the number of polymorphic sites provides an estimate of "richness" in the viral population, it may be incompatible to compare iSNV counts between samples with different read coverage profiles[56]. In contrast, π is a more robust metric that is unbiased by sequencing depth. For each site l:

$$\pi_l = \frac{N_l(N_l - 1) - \sum_j n_{j,l}\left(n_{j,l} - 1\right)}{N_l(N_l - 1)}$$

where N_l and $n_{j,l}$ are the coverage and number of reads encoding allele j in site l respectively[56]. To compute the π statistic for the entire genome of length L:

$$\pi = \sum_{l=1}^{L} \frac{\pi_l}{L}$$



Appendix – Figure 1: Genetic diversity of within-host influenza A virus populations as estimated by nucleotide diversity π statistic. Box plots summarizing the π statistic (iSNVs; median, interquartile range (IQR), and whiskers extending within median $\pm 1.5 \times$ IQR) computed for samples with adequate breadth of coverage

across the whole influenza genome. (**A**) seasonal A/H3N2 and (**B**) pandemic A/H1N1pdm09 viruses. All box plots are either stratified by day(s) since symptom onset or qPCR cycle threshold (Ct) values.

Here, we observed similar trends in genetic diversity when using π statistics compared to iSNVs counts (Appendix – Figure 1). π weakly increased with respect to time and CT values for A/H3N2 viruses (days since illness onset: Spearman $\rho = 0.388$, $p = 1.52 \times 10^{-8}$; CT: $\rho = 0.455$, $p = 3.66 \times 10^{-10}$) while remaining relatively invariant for A/H1N1pdm09 viruses (days since symptom onset: $\rho = 0.017$, $p = 0.92$; CT: $\rho = 0.240$, $p = 0.13$).

A3. Potential linked minority variants in within-host virus populations

For both within-host seasonal A/H3N2 and pandemic A/H1N1 virus populations, there were few instances of potentially linked nonsynonymous variants and if such co-variants were to exist, they were mainly found in the internal gene segments (Supplemental File 2). There was only one pair of HA amino acid mutations (E261G/L455F) that was encoded by a minority haplotype of A/H1N1pdm09 viruses infecting one individual but the normalized Lewontin's linkage disequilibrium measure (LD') was less than 0.5, suggesting a low likelihood that the mutation pair was linked non-randomly. These potentially linked variants tend to emerge late in the infection (6-7 days post illness onset) for both viral subtypes, in inferred haplotypes appearing at low frequencies within-host (median frequency = 0.08, IQR = 0.03-0.12) that were not shared between multiple individuals.

A4. Transmission bottleneck size estimation of pandemic A/H1N1pdm09 viral infections

Index cases were previously identified for six of the seven households where the pandemic A/H1N1pdm09 viral samples were collected[52]. Assuming that the non-index cases within the same household were secondarily infected by the index case, five transmission pairs were identified where samples with adequate breadth of coverage (>70% of genome covered with >50x coverage; Appendix – Figure 2) were collected from the index patient on an earlier date relative to the secondary case.

Virus transmission bottleneck sizes were then estimated using two binomial sampling models that were elaborated in detail by [75] and [4]. First, the presence/absence model computes transmission probability as the probability that a transmitted donor iSNV was found in at least one genome in the bottleneck population:

$$P_{d,i}(A|N_b) = p_{d,A}^{N_b}$$

where A refers to the transmitted iSNV in polymorphic site i, N_b is the bottleneck size and $p_{d,A}$ is the frequency of allele A in the sampled virus population within donor d.

The presence/absence model does not incorporate recipient frequencies of transmitted iSNVs in its probability calculations. It assumes that all transmitted iSNVs are detected in the recipient, and thus any donor iSNVs that are not present in the recipient are considered to have not been transmitted. It also does not account for any changes to p_d between the time of sampling and day of transmission. The maximum likelihood estimate of N_b would thus yield the largest log likelihood value given by:

$$LL(N_b) = \sum_{d} \sum_{i} \ln P_{d,i}$$

To incorporate information on recipient frequencies which can change between transmission and sampling, transmission bottleneck sizes were re-estimated using a second beta-binomial model formulated by Sobel Leonard et al. (2017). For each allele A observed in polymorphic site i that was transmitted from donor d to recipient r, the log-likelihood of N_b is given as:

$$LL(N_b)_{d,r}^{transmitted} = \sum_{A_i} \ln \left\{ \sum_{k=1}^{N_b} p_{beta}\left(p_{r,A_i} \middle| k, N_b - k\right) p_{bin}\left(k \middle| N_b, p_{d,A_i}\right) \right\}$$

where $p_{beta}\left(p_{r,A_i} \middle| k, N_b - k\right)$ is the conditional probability density, as modelled by the beta distribution, that the transmitted iSNV, A_i is found in the recipient at frequency p_{r,A_i} given that the variant is found present in k genomes out of the total transmission bottleneck of N_b genomes. $p_{bin}\left(k \middle| N_b, p_{d,A_i}\right)$ is the binomial probability of drawing k genomes with allele A_i in a sample of N_b genomes and variant frequency of p_{d,A_i} within the donor.

As some of the iSNVs in the donor may not be transmitted or were present below the 2% minimum variant frequency cut-off, the likelihood of these events was computed by:

$$LL(N_b)_{d,r}^{lost} = \sum_{A_i} \ln \left\{ \sum_{k=1}^{N_b} p_{beta,cdf}\left(p_{r,A_i} < 0.02 \middle| k, N_b - k\right) p_{bin}\left(k \middle| N_b, p_{d,A_i}\right) \right\}$$

where $p_{beta,cdf}$ is the cumulative distribution function of the beta distribution.

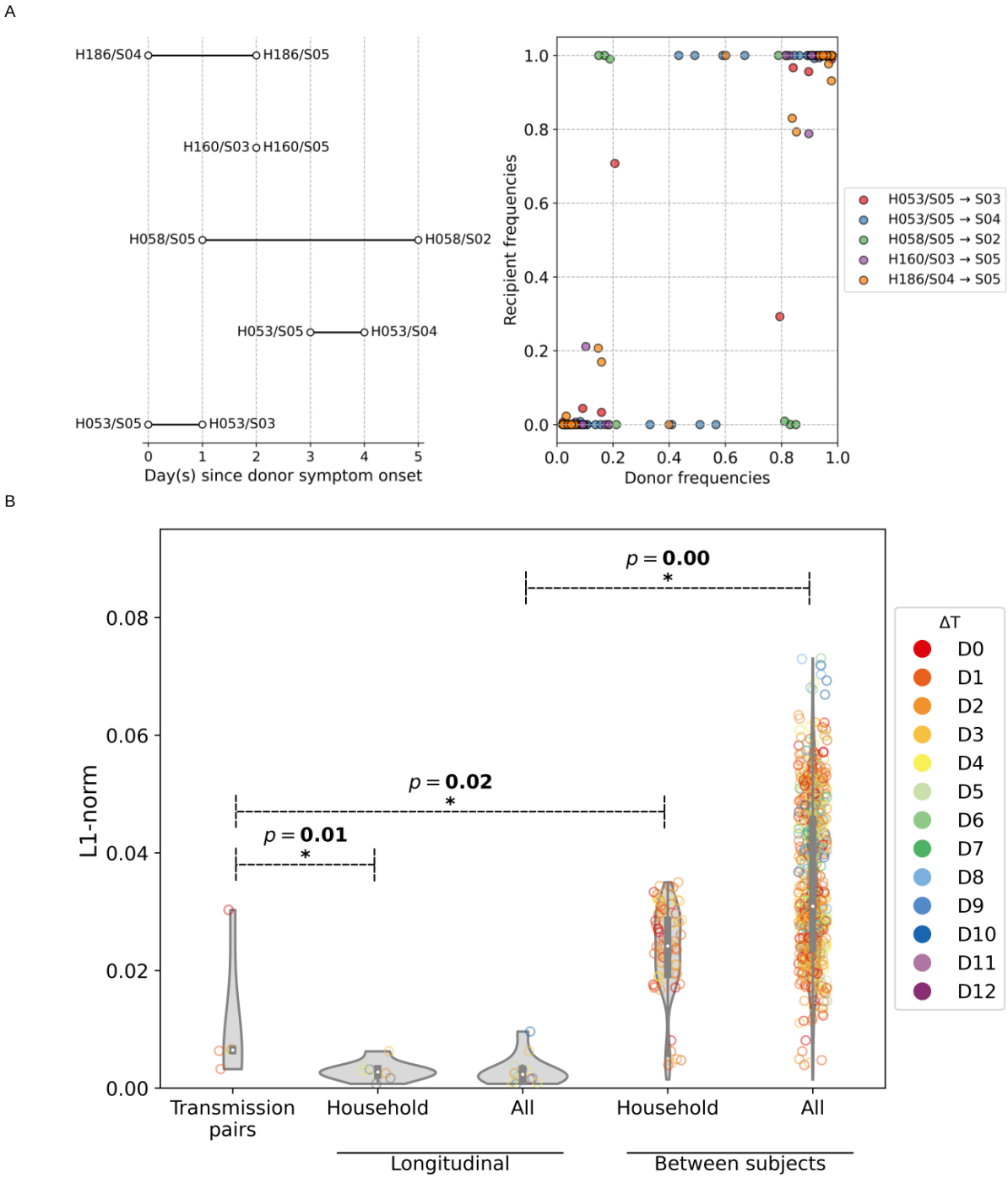
The maximum likelihood estimate of N_b as described by the beta-binomial was then computed by searching for the value of N_b that would give the largest value of:

$$LL(N_b) = \sum_{d,r} LL(N_b)_{d,r}^{transmitted} + LL(N_b)_{d,r}^{lost}$$

Log-likelihood values were computed for a range N_b between 1 to 1000 genomes for both models.

Transmission bottleneck size was then estimated by aggregating over all transmission pairs and applying two previously developed sampling based models[4,75]. We estimated the transmission bottleneck of pandemic A/H1N1pdm09 viruses to be 1-2 genomes (maximum likelihood (ML) presence-absence model estimate = 1 genome; ML beta-binomial model estimate = 2 genomes). The tight bottleneck was further reflected in the iSNV frequency plot where most of the donor variants were either present as the single majority allele or were not transmitted/undetected in the recipient, with few shared iSNVs between the two virus populations (Appendix – Figure 2A).

Furthermore, we observed that the virus populations between patients (median L1-norm distance = 0.031 divergence per site, IQR = 0.024-0.046 divergence per site) were significantly more different than longitudinal samples collected from the same individual (median L1-norm distance = 2.35×10^{-3} divergence per site, IQR = $1.54 \times 10^{-3} - 3.18 \times 10^{-3}$ divergence per site), suggesting that the haplotypes found within an individual remains relatively invariant throughout the course of infection (Appendix – Figure 2B). Combined with the fact that the per-site L1-norm genetic distance between samples attributed to the identified transmission pairs (median L1-norm distance = 6.49×10^{-3} divergence per site, IQR = $6.34 \times 10^{-3} - 6.63 \times 10^{-3}$ divergence per site) was greater than those computed for longitudinal samples pairs of each household individual (median L1-norm distance = 2.77×10^{-3} divergence per site, IQR = $2.18 \times 10^{-3} - 3.32 \times 10^{-3}$ divergence per site; Mann-Whitney U *p*-value = 0.01; Appendix – Figure 2B), it is likely only a limited number of haplotypes were shared between individuals in a transmission pair. Based on the most parsimonious reconstructed haplotypes for the five transmission pairs encoding shared iSNVs, we estimated the median number of haplotypes transmitted from donor to recipient to be between 1 and 2 haplotypes.



Appendix – Figure 2: (**A**) Schematic of A/H1N1pdm09 virus household transmission pairs identified by epidemiological linkage and plotted based on timing of sample collection. (**B**) iSNV frequencies found in the donors and recipients of the five transmission pairs. (**C**) Violin plots of L1-norm pairwise genetic distance per site between different A/H1N1pdm09 virus sample pairs (each circle = 1 pair of virus samples). Transmission pairs are those represented in Figure 5A. Longitudinal pairs are made up of sample pairs collected from the same individual on the first and any other later timepoints on which the patient was sampled. These pairs are stratified by whether they were collected from households located in the same community (i.e. household) or combined with the rest of the analyzed A/H1N1pdm09 virus samples collected from hospitals (i.e. all). For the same aforementioned categories, we also plotted the distribution of L1-norm distances for pairs of viruses collected from different individuals. All circles are coloured by the difference in days between which the sample pairs were collected (ΔT). All *p*-values reported are based on Mann-Whitney U tests which were used to

determine if the L1-norm genetic distance distributions of the two categories marked by the ends of the horizontal line above are statistically distinct.

A5. Mutation-selection balance

Considering a single-locus mutant with deleterious fitness effect s (i.e. $s < 0$), the frequency of the mutant allele (f) can be modelled by the following stochastic differential equation, otherwise known as the Langevin equation[76]:

$$\frac{\partial f}{\partial t} = sf(1-f) + \sqrt{\frac{f-(1-f)}{N}}\eta(t) + \mu(1-f)$$

selection genetic drift mutation

where N is the population size, μ is the mutation rate and $\eta(t)$ is the stochastic noise term due to genetic drift. For any Langevin equation, we can find the time-dependent probability distribution of f (i.e. $\frac{\partial p(f,t)}{\partial t}$) by its corresponding Fokker-Planck equation. At stationarity (i.e. $\frac{\partial p(f,t)}{\partial t} = 0$), its solution is known to be:

$$p(f) \propto \frac{e^{-2N\Lambda(f)}}{f(1-f)}$$

where $-\frac{\partial \Lambda(f)}{\partial t} = s + \frac{\mu}{f}$

Integrating $-\frac{\partial \Lambda(f)}{\partial t}$, we will get:

$$-\Lambda(f) = sf + \mu \ln f$$

which can then be substituted to obtain:

$$p(f) \propto e^{2Nsf} \cdot f^{2N\mu-1}$$

In other words, $p(f)$ strongly depends on $N\mu$. If $N\mu \gg 1$, we can see that $p(f)$ will strongly peak at some characteristic value of f that minimizes $\Lambda(f)$. If N is large, we can assume drift effects is negligible and f is largely deterministic due to selection:

$$\frac{\partial f}{\partial t} = sf(1-f) + \mu(1-f) = \left\{-\frac{\partial \Lambda(f)}{\partial t}\right\}[f(1-f)]$$

$$\Rightarrow \frac{\partial \Lambda}{\partial t} = \frac{\partial f}{\partial t}\left(\frac{\partial \Lambda}{\partial f}\right) = \frac{\partial f}{\partial t}\left\{-\frac{\partial f}{\partial t}\left(\frac{1}{f(1-f)}\right)\right\} \leq 0$$

959

960 In other words, selection dynamics minimises the $\Lambda(f)$ term and as such, result in mutation-
961 selection balance:

$$-\frac{\partial \Lambda(f)}{\partial t} = s + \frac{\mu}{f} = 0$$

$$\Rightarrow f = -\frac{\mu}{s}$$

962

963

964

References

1. Andino, R. & Domingo, E. Viral quasispecies. *Virology* **479–480**, 46–51 (2015).
2. Smith, D. J. *et al.* Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science* **305**, 371–376 (2004).
3. Varble, A. *et al.* Influenza A Virus Transmission Bottlenecks Are Defined by Infection Route and Recipient Host. *Cell Host & Microbe* **16**, 691–700 (2014).
4. McCrone, J. T. *et al.* Stochastic processes constrain the within and between host evolution of influenza virus. *eLife* **7**, e35962 (2018).
5. Russell, C. A. *et al.* The global circulation of seasonal influenza A (H3N2) viruses. *Science (New York, N.Y.)* **320**, 340–6 (2008).
6. Rambaut, A. *et al.* The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453**, 615–619 (2008).
7. Nelson, M. I. & Holmes, E. C. The evolution of epidemic influenza. *Nature Reviews Genetics* **8**, 196–205 (2007).
8. Dinis, J. M. *et al.* Deep Sequencing Reveals Potential Antigenic Variants at Low Frequencies in Influenza A Virus-Infected Humans. *Journal of virology* **90**, 3355–65 (2016).
9. Debbink, K. *et al.* Vaccination has minimal impact on the intrahost diversity of H3N2 influenza viruses. *PLOS Pathogens* **13**, e1006194 (2017).
10. Valesano, A. L. *et al.* Influenza B Viruses Exhibit Lower Within-Host Diversity than Influenza A Viruses in Human Hosts. *Journal of Virology* **94**, e01710-19 (2020).
11. Sobel Leonard, A. *et al.* Deep Sequencing of Influenza A Virus from a Human Challenge Study Reveals a Selective Bottleneck and Only Limited Intrahost Genetic Diversification. *Journal of virology* **90**, 11247–11258 (2016).
12. Han, A. X., Maurer-Stroh, S. & Russell, C. A. Individual immune selection pressure has limited impact on seasonal influenza virus evolution. *Nature Ecology & Evolution* **3**, 302–311 (2019).
13. Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125 (2009).
14. Su, Y. C. F. *et al.* Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nature Communications* **6**, 7952 (2015).
15. Elderfield, R. A. *et al.* Accumulation of Human-Adapting Mutations during Circulation of A(H1N1)pdm09 Influenza Virus in Humans in the United Kingdom. *Journal of Virology* **88**, 13269 LP – 13283 (2014).

16. Nogales, A., Martinez-Sobrido, L., Chiem, K., Topham, D. J. & DeDiego, M. L. Functional Evolution of the 2009 Pandemic H1N1 Influenza Virus NS1 and PA in Humans. *Journal of Virology* **92**, e01206-18 (2018).

17. Poon, L. L. M. *et al.* Quantifying influenza virus diversity and transmission in humans. *Nature Genetics* **48**, 195–200 (2016).

18. Xue, K. S. & Bloom, J. D. Reconciling disparate estimates of viral genetic diversity during human influenza infections. *Nature Genetics* **51**, 1298–1301 (2019).

19. Poon, L. L. M. *et al.* Reply to 'Reconciling disparate estimates of viral genetic diversity during human influenza infections.' *Nature Genetics* **51**, 1301–1303 (2019).

20. South East Asia Infectious Disease Clinical Research Network. Effect of double dose oseltamivir on clinical and virological outcomes in children and adults admitted to hospital with severe influenza: Double blind randomised controlled trial. *BMJ (Clinical research ed.)* **346**, f3039 (2013).

21. Koel, B. F. *et al.* Longitudinal sampling is required to maximize detection of intrahost A/H3N2 virus variants. *Virus Evolution* **6**, veaa088 (2020).

22. Worby, C. J. *et al.* On the relative role of different age groups in influenza epidemics. *Epidemics* **13**, 10–16 (2015).

23. Viboud, C. *et al.* Risk factors of influenza transmission in households. *International Congress Series* **1263**, 291–294 (2004).

24. Horby, P. *et al.* The epidemiology of interpandemic and pandemic influenza in Vietnam, 2007-2010. *American Journal of Epidemiology* **175**, 1062–1074 (2012).

25. Hien, T. T. *et al.* Early Pandemic Influenza (2009 H1N1) in Ho Chi Minh City, Vietnam: A Clinical Virological and Epidemiological Analysis. *PLOS Medicine* **7**, e1000277 (2010).

26. Xue, K. S. & Bloom, J. D. Linking influenza virus evolution within and between human hosts. *Virus Evolution* **6**, 812016 (2020).

27. Wiley, D. C. C., Wilson, I. A. A. & Skehel, J. 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).

28. 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**, 417–27 (1982).

29. Igarashi, M. *et al.* Predicting the Antigenic Structure of the Pandemic (H1N1) 2009 Influenza Virus Hemagglutinin. *PLOS ONE* **5**, e8553 (2010).

30. Mänz, B. *et al.* Pandemic Influenza A Viruses Escape from Restriction by Human MxA through Adaptive Mutations in the Nucleoprotein. *PLOS Pathogens* **9**, e1003279 (2013).

31. Ashenberg, O., Gong, L. I. & Bloom, J. D. Mutational effects on stability are largely conserved during protein evolution. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 21071–6 (2013).

32. Berkhoff, E. G. M. *et al.* Functional Constraints of Influenza A Virus Epitopes Limit Escape from Cytotoxic T Lymphocytes. *Journal of Virology* **79**, 11239 LP – 11246 (2005).

33. Gog, J. R., Rimmelzwaan, G. F., Osterhaus, A. D. M. E. & Grenfell, B. T. Population dynamics of rapid fixation in cytotoxic T lymphocyte escape mutants of influenza A. *Proceedings of the National Academy of Sciences* **100**, 11143 LP – 11147 (2003).

34. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* **2**, e00631 (2013).

35. Rimmelzwaan, G. F., Berkhoff, E. G. M., Nieuwkoop, N. J., Fouchier, R. A. M. & Osterhaus, A. D. M. E. Functional Compensation of a Detrimental Amino Acid Substitution in a Cytotoxic-T-Lymphocyte Epitope of Influenza A Viruses by Comutations. *Journal of Virology* **78**, 8946 LP – 8949 (2004).

36. Berkhoff, E. G. M. *et al.* A Mutation in the HLA-B*2705-Restricted NP383-391 Epitope Affects the Human Influenza A Virus-Specific Cytotoxic T-Lymphocyte Response In Vitro. *Journal of Virology* **78**, 5216 LP – 5222 (2004).

37. Resa-Infante, P. *et al.* Targeting Importin-α7 as a Therapeutic Approach against Pandemic Influenza Viruses. *Journal of Virology* **89**, 9010 LP – 9020 (2015).

38. Resa-Infante, P. *et al.* Alternative interaction sites in the influenza A virus nucleoprotein mediate viral escape from the importin-α7 mediated nuclear import pathway. *The FEBS Journal* **286**, 3374–3388 (2019).

39. Gabriel, G. *et al.* Differential use of importin-α isoforms governs cell tropism and host adaptation of influenza virus. *Nature Communications* **2**, 156 (2011).

40. Chenavas, S. *et al.* Monomeric Nucleoprotein of Influenza A Virus. *PLOS Pathogens* **9**, e1003275 (2013).

41. Mak, G. C. *et al.* Association of D222G substitution in haemagglutinin of 2009 pandemic influenza A (H1N1) with severe disease. *Eurosurveillance* **15**, (2010).

42. Mai, L. Q. *et al.* A Community Cluster of Oseltamivir-Resistant Cases of 2009 H1N1 Influenza. *New England Journal of Medicine* **362**, 86–87 (2010).

43. Ghafari, M., Lumby, C. K., Weissman, D. B. & Illingworth, C. J. R. Inferring Transmission Bottleneck Size from Viral Sequence Data Using a Novel Haplotype Reconstruction Method. *Journal of Virology* **94**, (2020).

44. McCrone, J. T., Woods, R. J., Monto, A. S., Martin, E. T. & Lauring, A. S. The effective population size and mutation rate of influenza A virus in acutely infected individuals. *bioRxiv* 2020.10.24.353748 (2020) doi:10.1101/2020.10.24.353748.

45. Morris, D. H. *et al.* Asynchrony between virus diversity and antibody selection limits influenza virus evolution. *eLife* **9**, 1–62 (2020).

46. Xue, K. S. *et al.* Parallel evolution of influenza across multiple spatiotemporal scales. *eLife* **6**, e26875 (2017).

47. Lumby, C. K., Zhao, L., Breuer, J. & Illingworth, C. J. R. A large effective population size for established within-host influenza virus infection. *eLife* **9**, e56915 (2020).

48. Petrova, V. N. & Russell, C. A. The evolution of seasonal influenza viruses. *Nature Reviews Microbiology* **16**, 47–60 (2017).

49. Ng, S. *et al.* The Timeline of Influenza Virus Shedding in Children and Adults in a Household Transmission Study of Influenza in Managua, Nicaragua. *The Pediatric Infectious Disease Journal* **35**, 583–586 (2016).

50. Watson, S. J. *et al.* Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120205 (2013).

51. Welkers, M. R. A. *et al.* Genetic diversity and host adaptation of avian H5N1 influenza viruses during human infection. *Emerging Microbes & Infections* **8**, 262–271 (2019).

52. Thai, P. Q. *et al.* Pandemic H1N1 virus transmission and shedding dynamics in index case households of a prospective Vietnamese cohort. *Journal of Infection* **68**, 581–590 (2014).

53. Suess, T. *et al.* Comparison of Shedding Characteristics of Seasonal Influenza Virus (Sub)Types and Influenza A(H1N1)pdm09; Germany, 2007–2011. *PLOS ONE* **7**, e51653 (2012).

54. Jacobs, N. T. *et al.* Incomplete influenza A virus genomes occur frequently but are readily complemented during localized viral spread. *Nature Communications* **10**, 3526 (2019).

55. McCown, M. F. & Pekosz, A. The Influenza A Virus M₂ Cytoplasmic Tail Is Required for Infectious Virus Production and Efficient Genome Packaging. *Journal of Virology* **79**, 3595 LP – 3605 (2005).

56. Zhao, L. & Illingworth, C. J. R. Measurements of intrahost viral diversity require an unbiased diversity metric. *Virus Evolution* **5**, (2019).

57. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences* **76**, 5269 LP – 5273 (1979).

58. Illingworth, C. J. R. *et al.* On the effective depth of viral sequence data. *Virus Evolution* **3**, (2017).

59. Xue, K. S., Moncla, L. H., Bedford, T. & Bloom, J. D. Within-Host Evolution of Human Influenza Virus. *Trends in Microbiology* **26**, 781–793 (2018).

60. 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. *Archives of virology* **146**, 2275–2289 (2001).

61. Lee, N. *et al.* Viral Loads and Duration of Viral Shedding in Adult Patients Hospitalized with Influenza. *The Journal of Infectious Diseases* **200**, 492–500 (2009).

62. Ling, L. M. *et al.* Effects of early oseltamivir therapy on viral shedding in 2009 pandemic influenza A (H1N1) virus infection. *Clin Infect Dis* **50**, 963–969 (2010).

63. Renzette, N. *et al.* Evolution of the Influenza A Virus Genome during Development of Oseltamivir Resistance In Vitro *Journal of Virology* **88**, 272 LP – 281 (2014).

64. Horby, P. *et al.* The epidemiology of interpandemic and pandemic influenza in Vietnam, 2007-2010. *American Journal of Epidemiology* **175**, 1062–1074 (2012).

65. Watson, S. J. *et al.* Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120205 (2013).

66. Zhou, B. *et al.* Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza a viruses. *Journal of virology* **83**, 10309–13 (2009).

67. Jonges, M. *et al.* Emergence of the Virulence-Associated PB2 E627K Substitution in a Fatal Human Case of Highly Pathogenic Avian Influenza Virus A(H7N7) Infection as Determined by Illumina Ultra-Deep Sequencing. *Journal of virology* **88**, 1694–702 (2014).

68. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

69. Magoč, T., Magoč, M. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. **27**, 2957–2963 (2011).

70. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).

71. Illingworth, C. J. R. SAMFIRE: multi-locus variant calling for time-resolved sequence data. *Bioinformatics* **32**, 2208–2209 (2016).

72. Frensing, T. *et al.* Influenza virus intracellular replication dynamics, release kinetics, and particle morphology during propagation in MDCK cells. *Applied Microbiology and Biotechnology* **100**, 7181–7192 (2016).

73. Hadjichrysanthou, C. *et al.* Understanding the within-host dynamics of influenza A virus: from theory to clinical implications. *Journal of The Royal Society Interface* **13**, 20160289 (2016).

74. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32**, 268–74 (2015).

75. Sobel Leonard, A., Weissman, D. B., Greenbaum, B., Ghedin, E. & Koelle, K. Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. *Journal of Virology* **91**, e00171-17 (2017).

76. Good, B. H. & Desai, M. M. Fluctuations in fitness distributions and the effects of weak linked selection on sequence evolution. *Theoretical Population Biology* **85**, 86–102 (2013).

Data availability

All raw sequence data have been deposited at NCBI sequence read archive under BioProject Accession number PRJNA722099. All custom Python code and Jupyter notebooks to reproduce the analyses in this paper are available online: https://github.com/AMC-LAEB/Within_Host_H3vH1.

Acknowledgements

We thank Carolien van de Sandt for helpful discussions. We gratefully acknowledge the authors, originating and submitting laboratories (Supplemental File 5) for the reference sequences retrieved from GISAID's EpiFlu Database used in this study.

A.X.H., Z.C.F.G. and C.A.R. were supported by ERC NaviFlu (No. 818353). The South East Asia Infectious Disease Clinical Research Network (SEAICRN) was funded by National Institutes of Allergy and Infectious Diseases, National Institutes of Health (US), N01-A0-50042, HHSN272200500042C.

Competing interests

The authors declare no competing interests.

Supplemental Files

Supplemental File 1: Mean number of nonsynonymous (NS), synonymous (S) and stop codon (Stop) variants per sample for each gene segment as well as the corresponding NS/S ratio.

Supplemental File 2: Potentially linked nonsynonymous variants in within-host A/H1N1pdm09 and A/H3N2 virus samples. Sample names are given in the format of "Patient ID_Days since symptom onset". Both linkage disequilibrium (LD) and the normalized LD' measures are tabulated alongside the inferred maximum-likelihood haplotype frequencies (q_{10} and q_{01} are the haplotype frequencies with variant i or ii only while q_{11} is the frequency of haplotypes encoding both variants).

Supplemental File 3: A/H3N2 segment-specific primers

Supplemental File 4: Patients metadata (provided as an excel file).

Supplemental File 5: Acknowledgement table of reference sequences downloaded from GISAID.

Figure Supplements

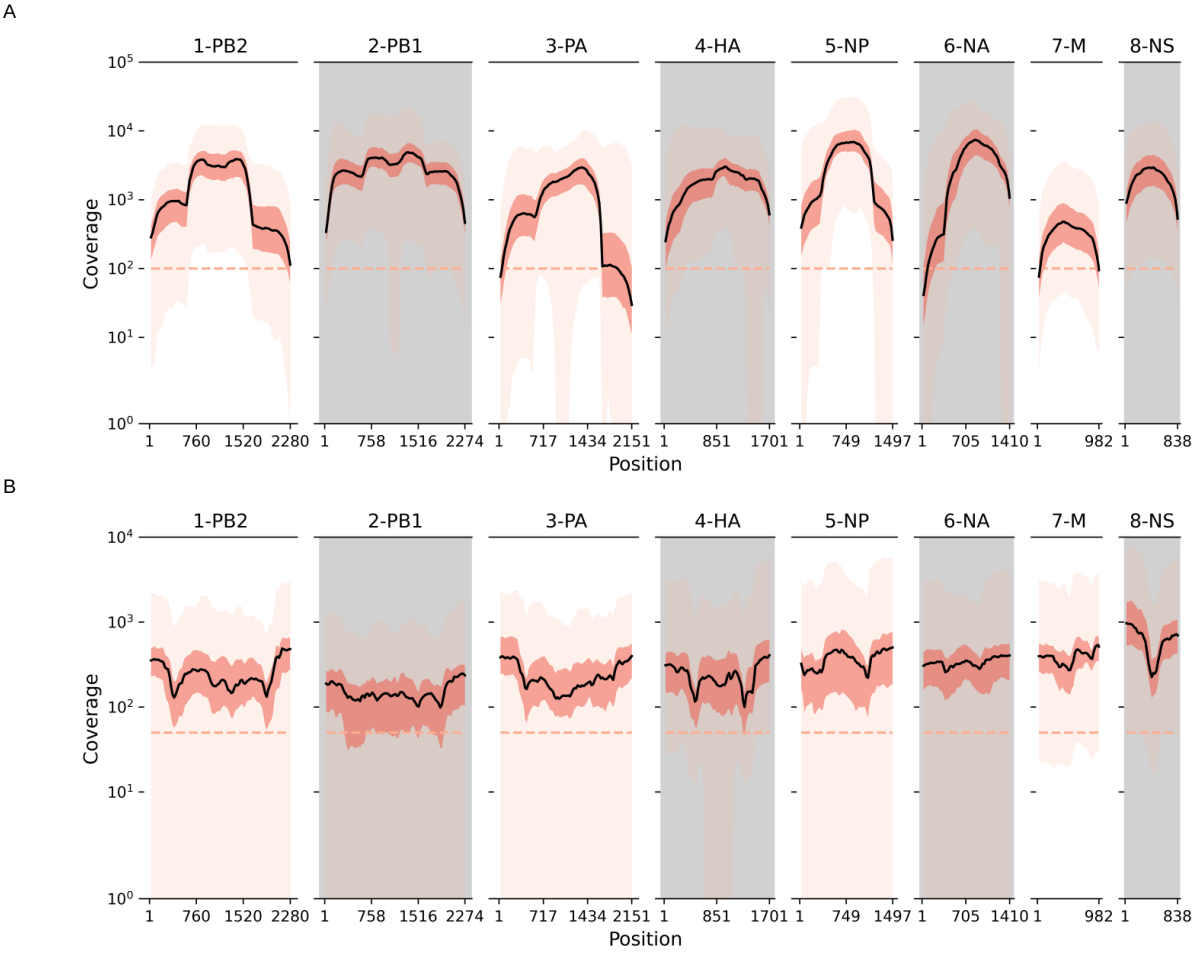


Figure 1 – figure supplement 1: Sequence coverage across all influenza gene segments and samples. Black line plots the mean coverage for a sliding window of 50 base pairs (stepsize = 25 base pairs). The interquartile range is shaded in dark pink while the full range is denoted in light pink. (**A**) H3N2. (**B**) H1N1pdm09

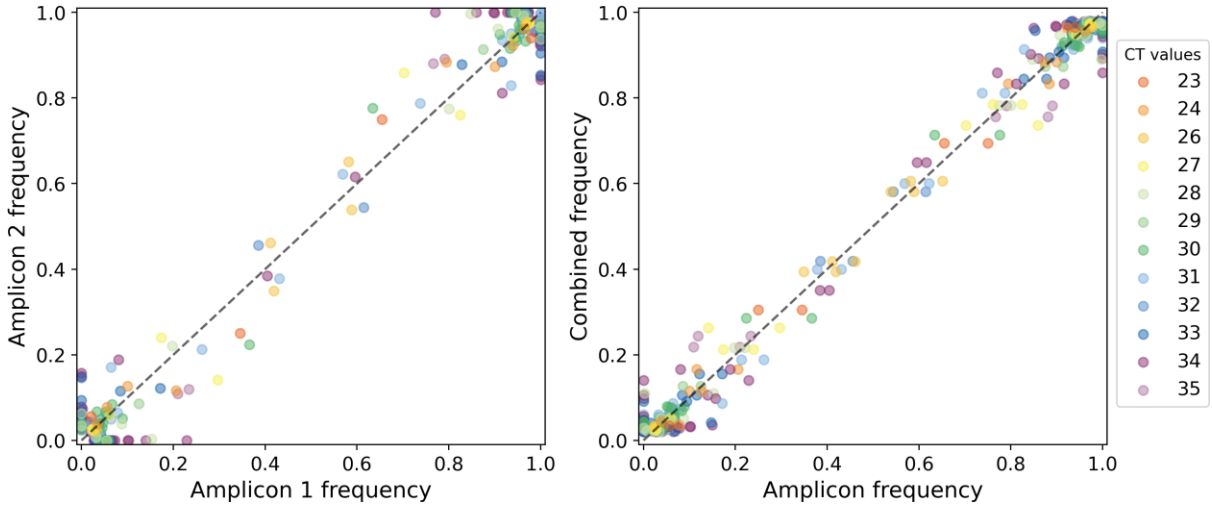


Figure 1 – figure supplement 2: Frequencies of nucleotide variants found in A/H3N2 viral reads sequenced from overlapping amplicons. Each circle represents a nucleotide variant site (with frequency estimated between 0.02 and 0.98) found in reads attributed to at least two different amplicons (at least 100x coverage for each amplicon), and is colored by the cycle threshold (CT) value of the sample from which the variant was found. Scatter plot on the left panel compares the variant frequencies between any two amplicons while the plot on the right panel compares the variant frequencies of each amplicon to that when combining across all overlapping amplicons (i.e. the frequencies used for main analyses). The dashed line is the one-to-one expected value.

Figure 1 – figure supplement 3: Maximum-likelihood phylogeny of putative majority (consensus) and minority whole genome sequences (by concatenating all eight gene segments) of A/H3N2 virus samples. Tip names are given in the format: "Patient ID_Days since symptom onset_putative consensus or minority sequence". The tree is rooted to the A/Brisbane/10/2007 virus (H3N2_Bris07; EPI_ISL_103644). Subject 1673 (green tips) and the D8 sample of subject 1878 (pink tips) might have arose from mixed infections or were contaminated by other strains.

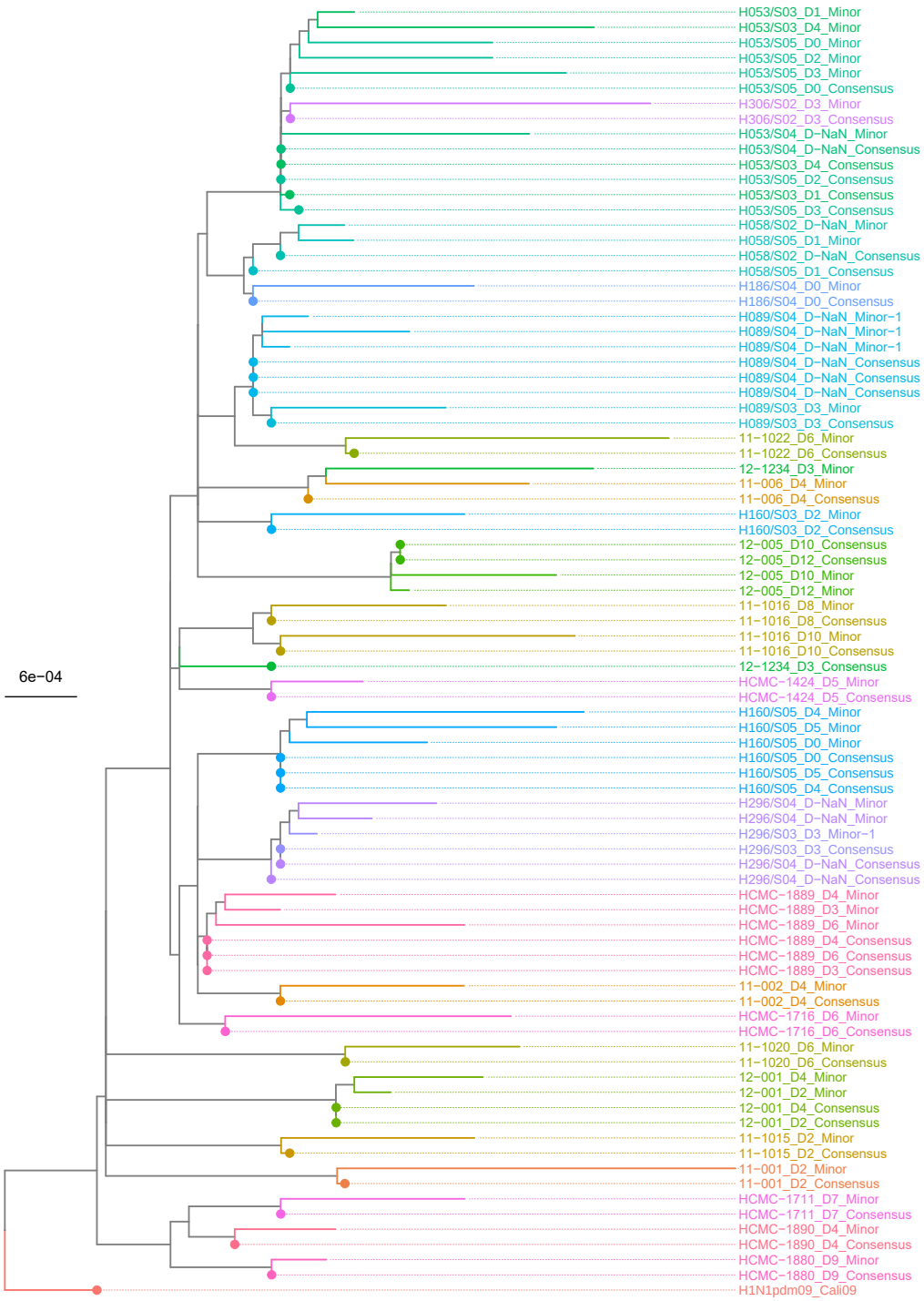


Figure 1 – figure supplement 4: Maximum-likelihood phylogeny of putative majority (consensus) and minority whole genome sequences (by concatenating all eight gene segments) of H1N1pdm09 virus samples. Tip names are given in the format: "Patient ID_Days since symptom onset_putative consensus or minority sequence". The tree is rooted to the A/California/04/2009 virus (H1N1pdm09_Cali09; EPI_ISL_376192). Encircled tips denote the consensus majority sequence of the sample.

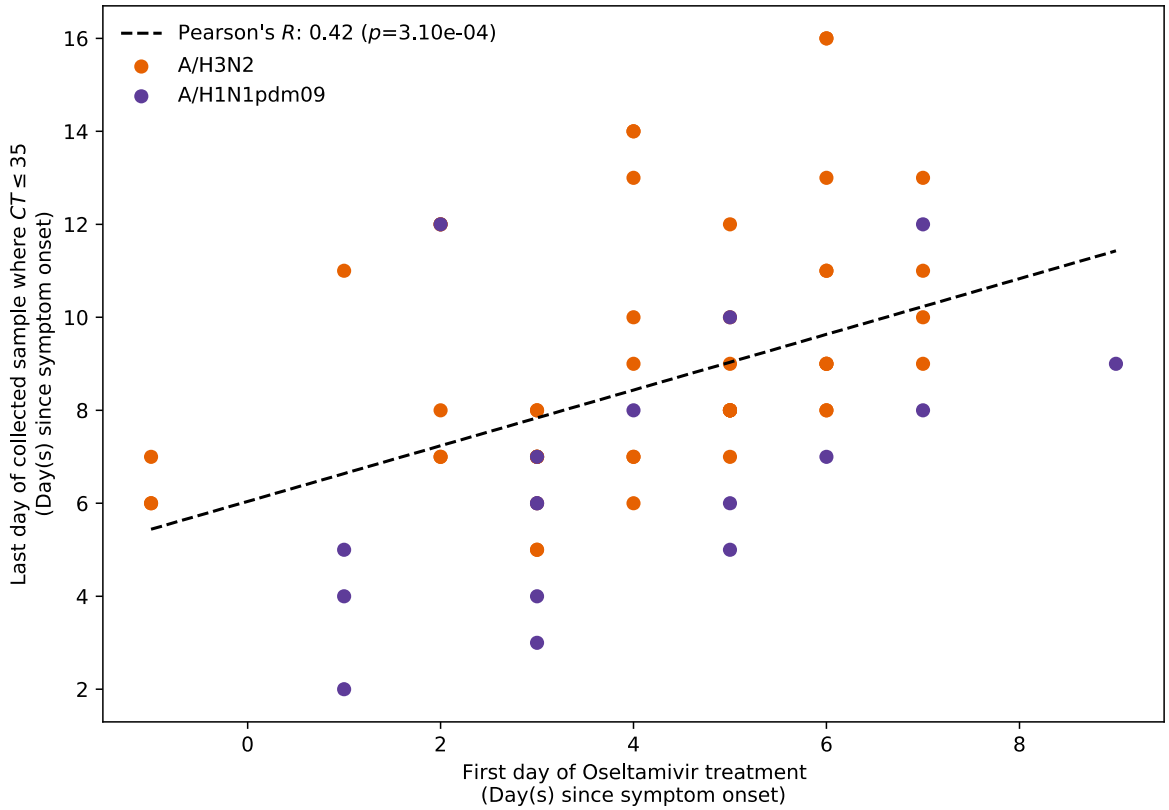


Figure 1 – figure supplement 5: Pearson's correlation between the first day of Oseltamivir treatment administered to patients and the last day on which viral samples with cycle threshold (CT) values ≤ 35 were collected. Time is measured by number of days since symptom onset. Each point represents a patient included in this study who was treated with Oseltamivir (Supplemental File 4).

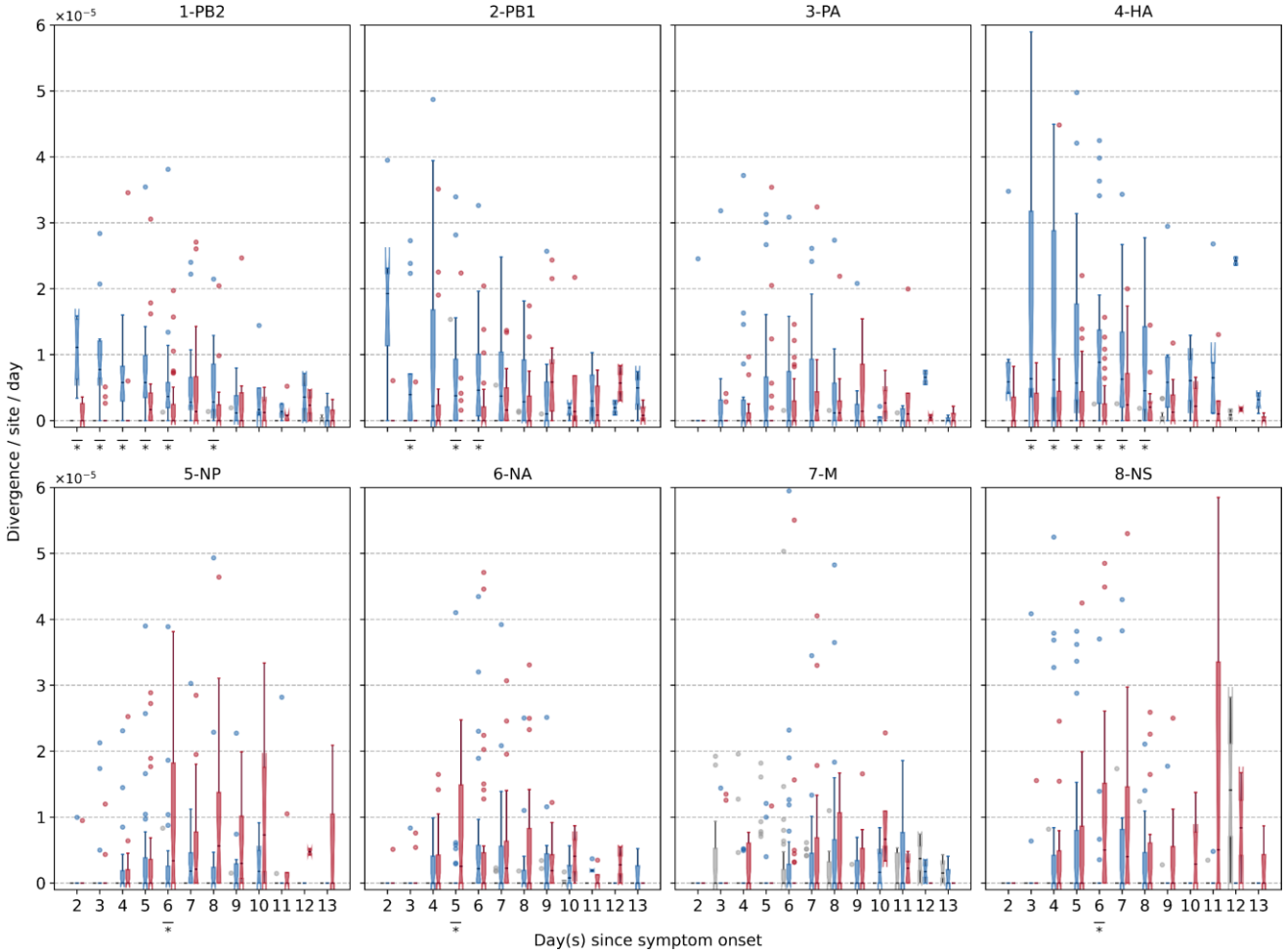


Figure 2 – figure supplement 1: Box plots (median, interquartile range (IQR), and whiskers extending within median $\pm 1.5 \times$ IQR) summarising the empirical within-host evolutionary rates ($r_{g,t}$) of different H3N2 viral gene segments. All rates are stratified by substitution type (synonymous – blue; nonsynonymous – red; stop codon – grey). Wilcoxon signed-rank tests were performed to assess if the paired synonymous and nonsynonymous evolutionary rates are significantly distinct per timepoint (annotated with "$*$" if $p < 0.05$).

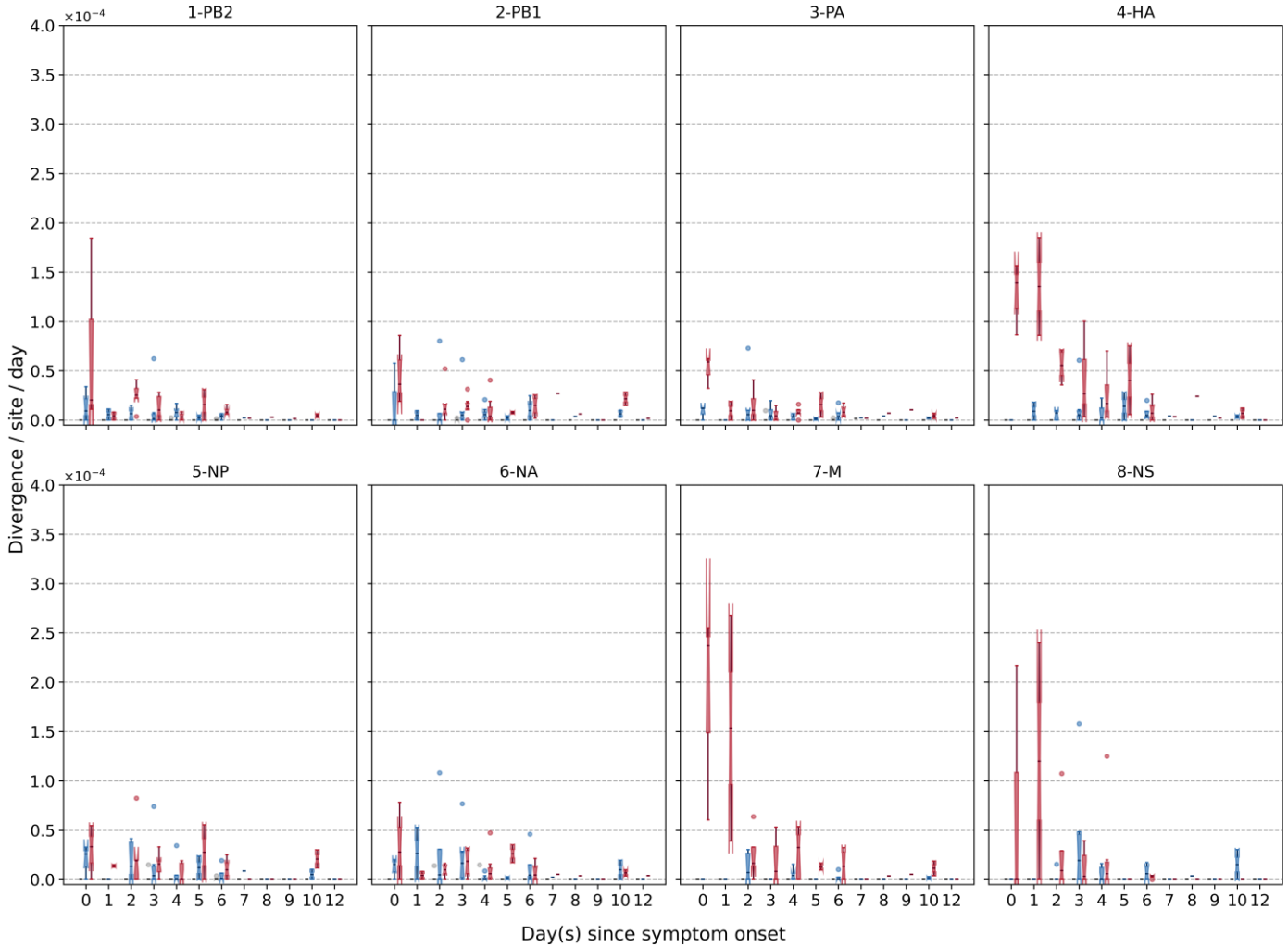


Figure 2 – figure supplement 2: Box plots (median, interquartile range (IQR), and whiskers extending within median $\pm 1.5 \times$ IQR) summarising the empirical within-host evolutionary rates ($r_{g,t}$) of different H1N1pdm09 viral gene segments. All rates are stratified by substitution type (synonymous – blue; nonsynonymous – red; stop codon – grey). Wilcoxon signed-rank test was *not* performed here due to low number of samples collected (i.e. median number of samples per day post illness onset = 2).

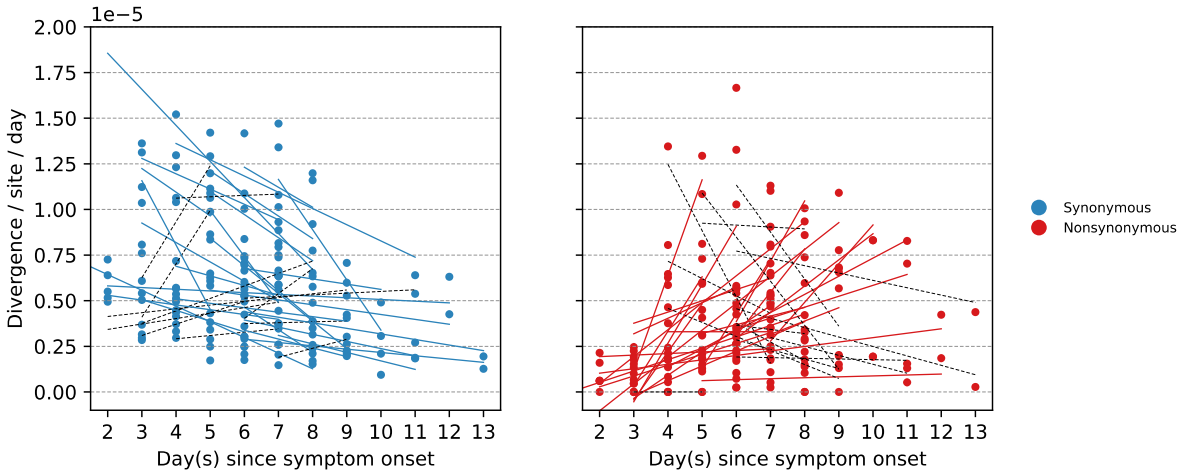


Figure 2 – figure supplement 3: Linear regression of within-host synonymous and nonsynonymous evolutionary rates of within-host A/H3N2 virus samples. Each plotted line is the linearly regressed line to the evolutionary rates computed for each A/H3N2 infected individual. Based on our findings, we expect that synonymous rates correlate negatively with time while nonsynonymous rates have a positive temporal correlation. Coloured lines represent those that fall within this expectation while dashed black lines represent those that did not.

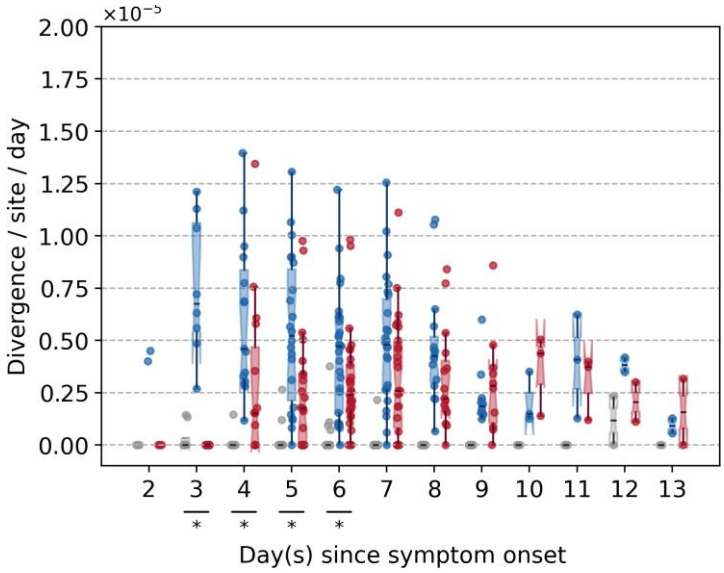


Figure 2 – figure supplement 4: Box plots (median, interquartile range (IQR), and whiskers extending within median $\pm 1.5 \times$ IQR) summarizing the empirical daily within-host evolutionary rates of seasonal A/H3N2 viruses. Variants that could potentially be PCR artefacts were removed (i.e. those found under the 75th percentile (6.3%) of frequency range of variants located in overlapping amplicons but were only detected in one amplicon, see Figure 1 – figure supplement 2). All rates are stratified by substitution type (synonymous – blue; nonsynonymous – red; grey – stop-codon). Wilcoxon signed-rank tests were performed to assess if the paired synonymous and nonsynonymous evolutionary rates are significantly distinct per individual gene segment or timepoint (annotated with "$*$" if $p < 0.05$). This was done for all sets of nonsynonymous and synonymous rate pairs computed between day 3 and day 9 since symptom onset.

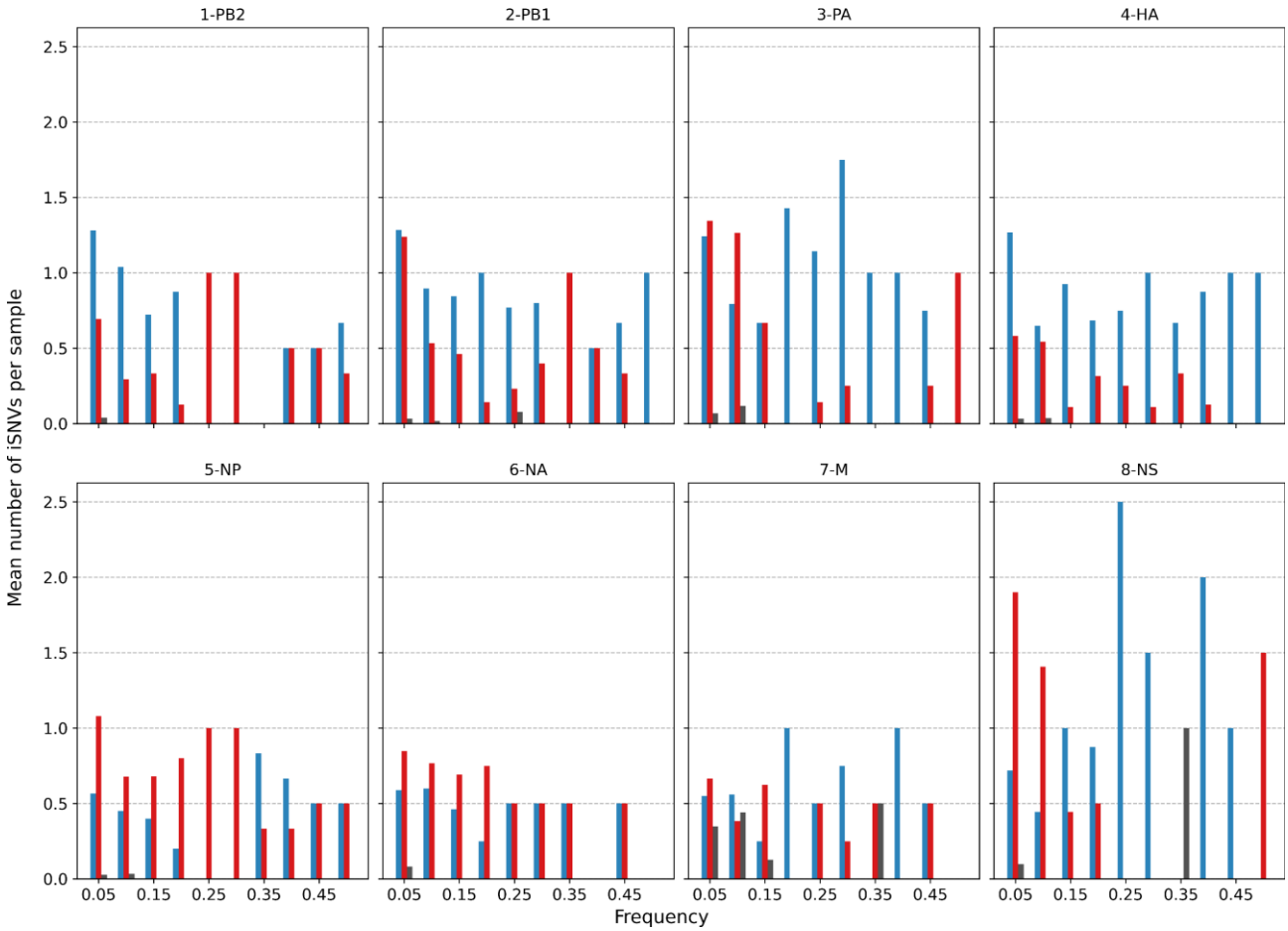


Figure 3 – figure supplement 1: Histogram of the mean number of minority iSNVs identified per sample across all H3N2 viral gene segments across all samples sorted by frequency bins of 5% and substitution type (synonymous – blue; nonsynonymous – red; grey – stop codon).

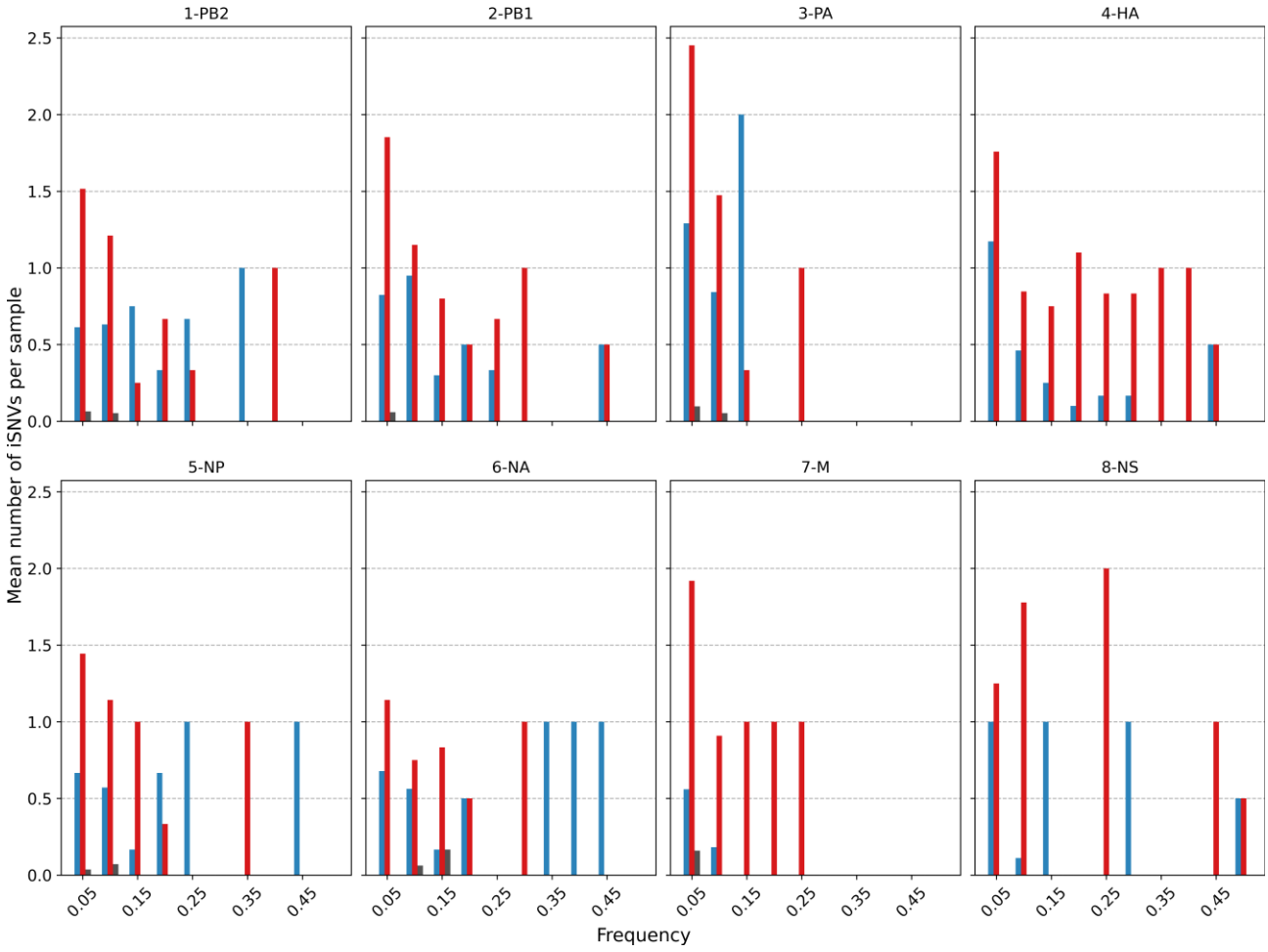


Figure 3 – figure supplement 2: Histogram of the mean number of minority iSNVs identified across all H1N1pdm09 viral gene segments across all samples sorted by frequency bins of 5% and substitution type (synonymous – blue; nonsynonymous – red; stop-codon – grey).

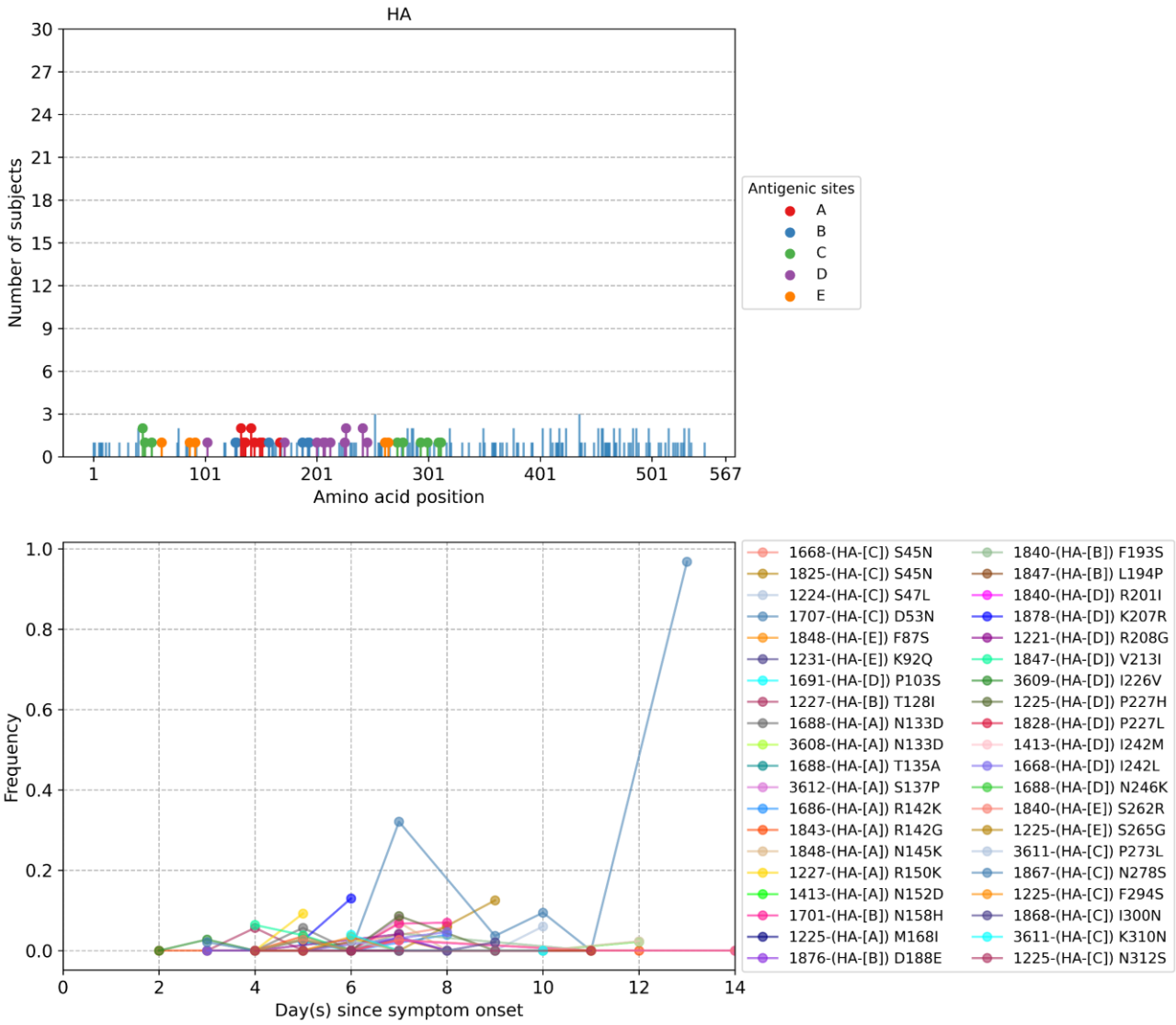


Figure 4 – figure supplement 1: Plots of intra-host haemagglutinin (HA) amino acid variants in A/H3N2 infected individuals. Top panel shows the number of subjects where nonsynonymous variants were found in the respective protein site. Different canonical antigenic sites of the HA protein are colored (HA numbering based on H3 numbering without signal peptide). Bottom panel plots selected as well as parallel amino acid mutations found in multiple patients against days since illness onset. Filled circles represent days on which samples were collected and sequenced. The variant frequencies of all putative antigenic sites are also plotted.

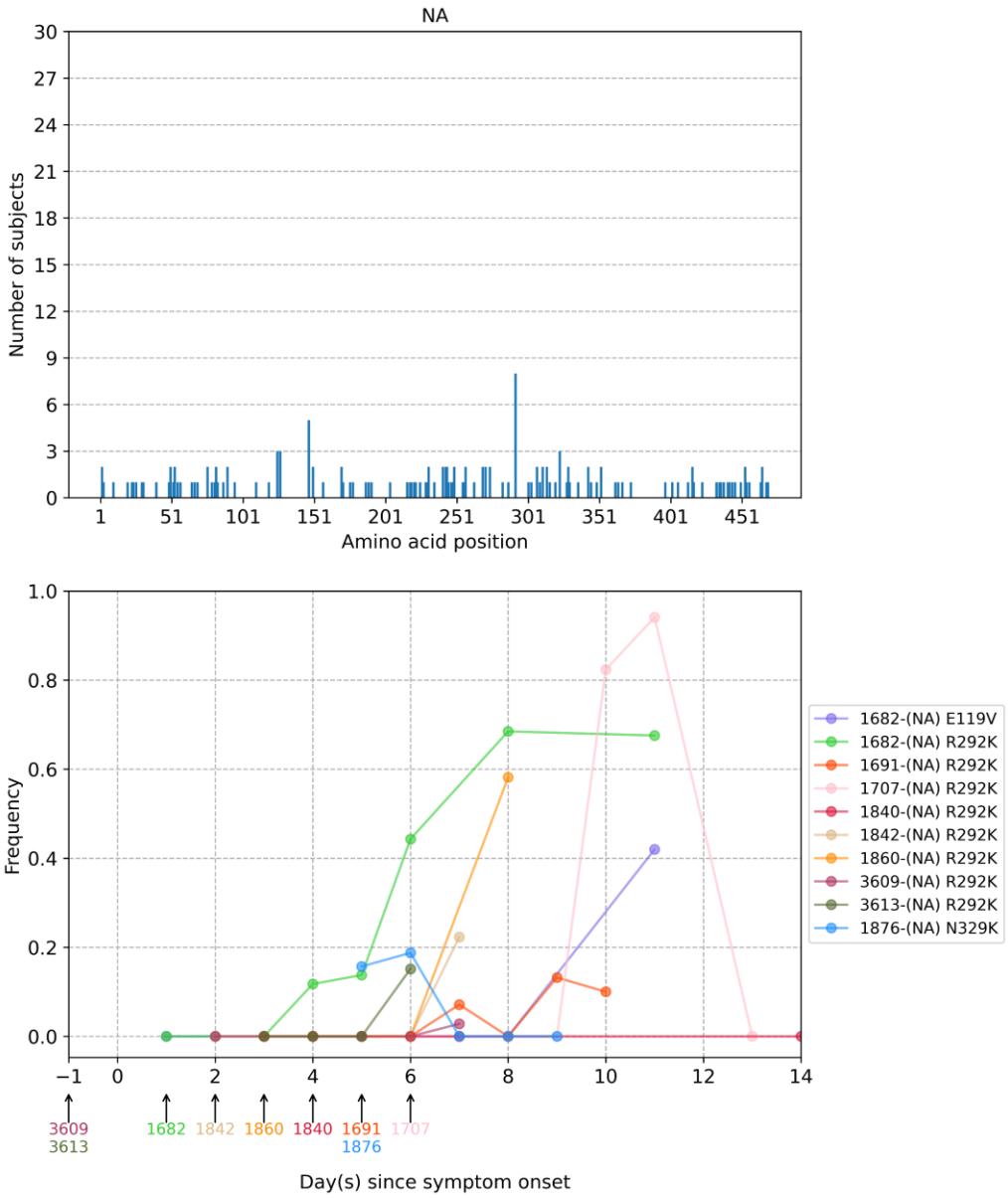


Figure 4 – figure supplement 2: Plots of intra-host neuraminidase (NA) amino acid variants in A/H3N2 infected individuals. Top panel shows the number of subjects where nonsynonymous variants were found in the respective protein site. Bottom panel plots selected as well as parallel amino acid mutations found in multiple patients against days since illness onset. Filled circles represent days on which samples were collected and sequenced. The first day of oseltamivir treatment for individuals with resistance mutations is annotated below the x-axis.

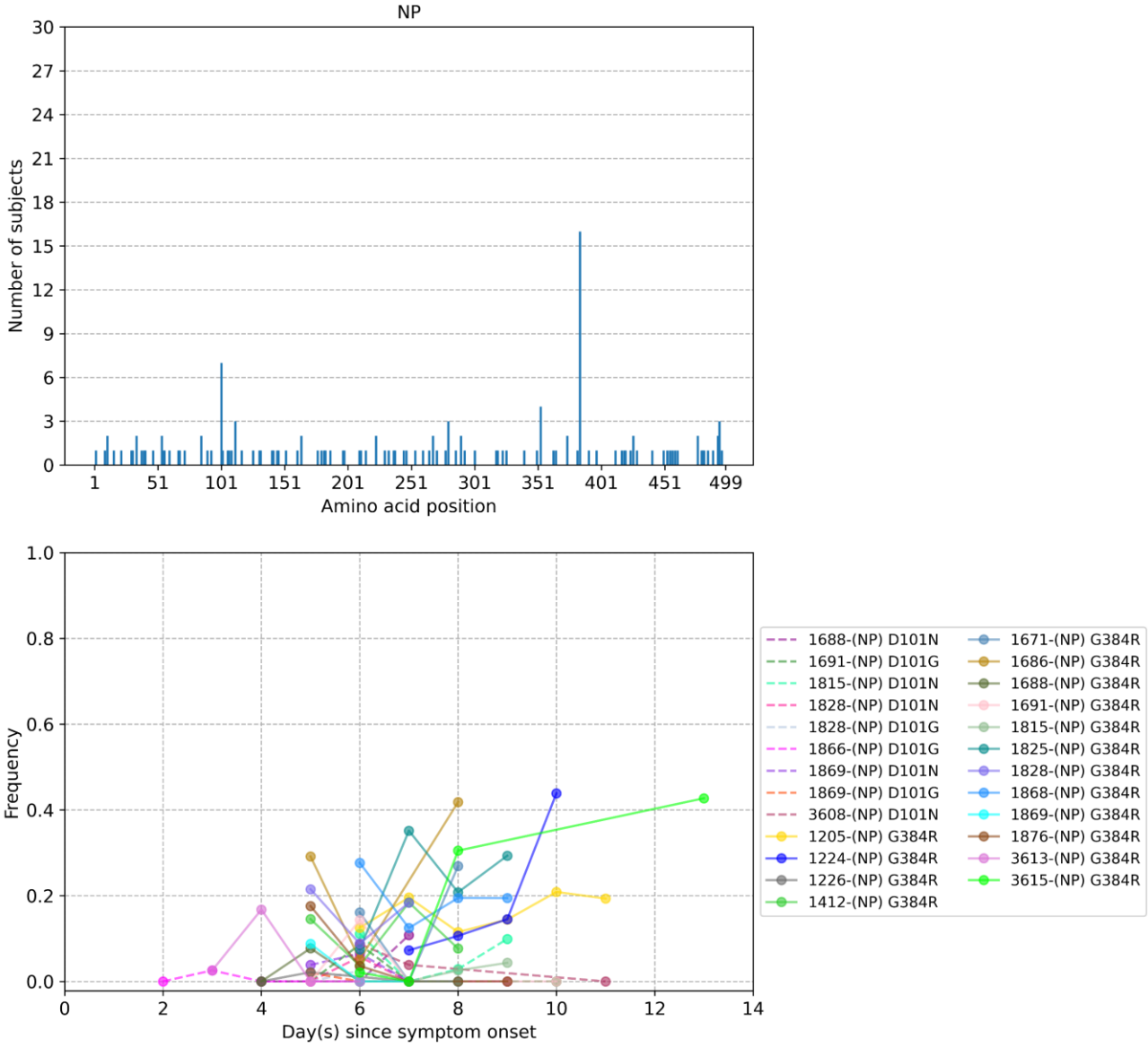


Figure 4 – figure supplement 3: Plots of intra-host nucleoprotein (NP) amino acid variants in A/H3N2 infected individuals. Top panel shows the number of subjects where nonsynonymous variants were found in the respective protein site. Bottom panel plots selected as well as parallel amino acid mutations found in multiple patients against days since illness onset. Filled circles represent days on which samples were collected and sequenced.

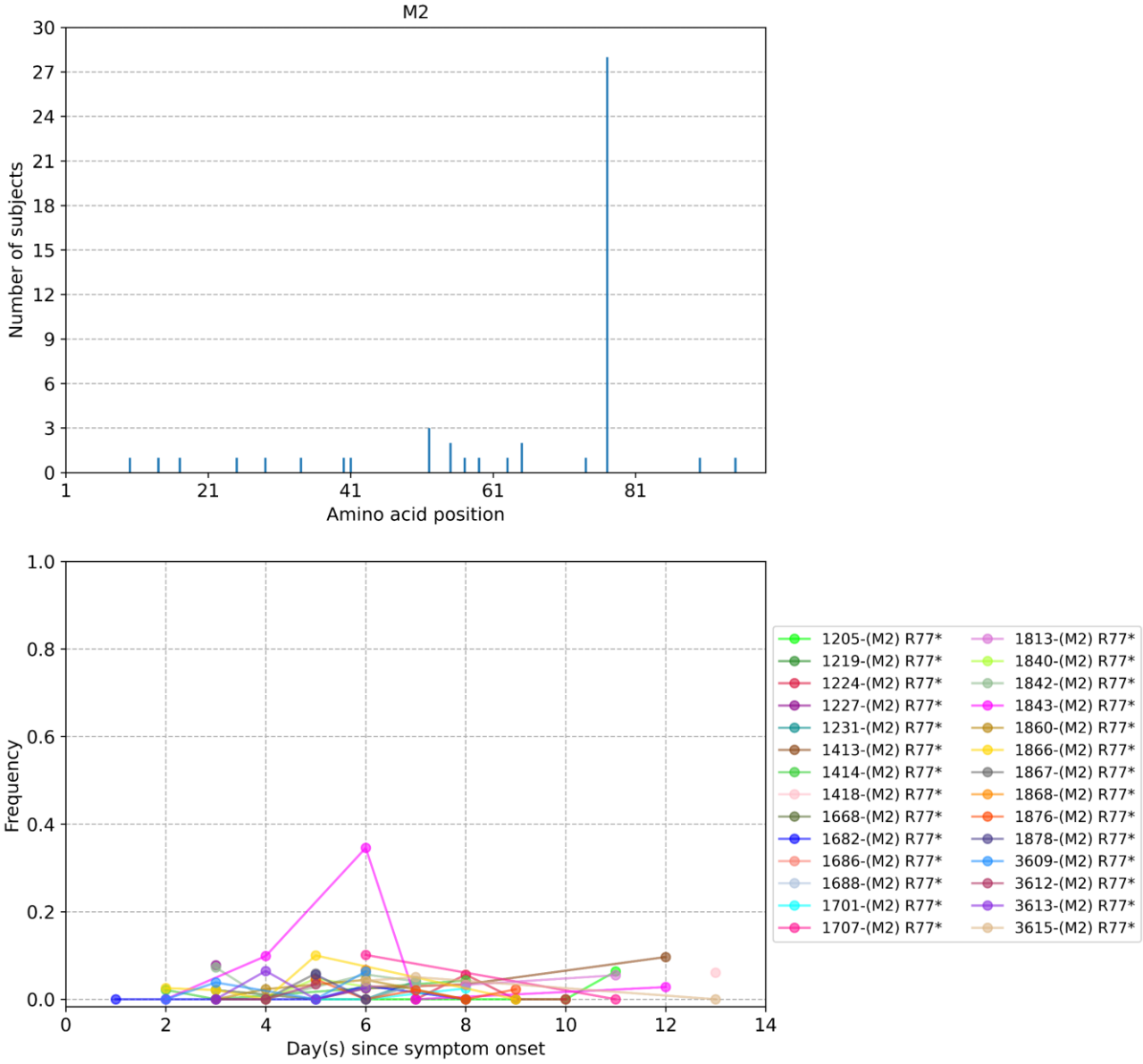


Figure 4 – figure supplement 4: Plots of M2 protein intra-host amino acid variants in A/H3N2 infected individuals. Top panel shows the number of subjects where nonsynonymous variants were found in the respective protein site. Bottom panel plots selected as well as parallel amino acid mutations found in multiple patients against days since illness onset. Filled circles represent days on which samples were collected and sequenced.

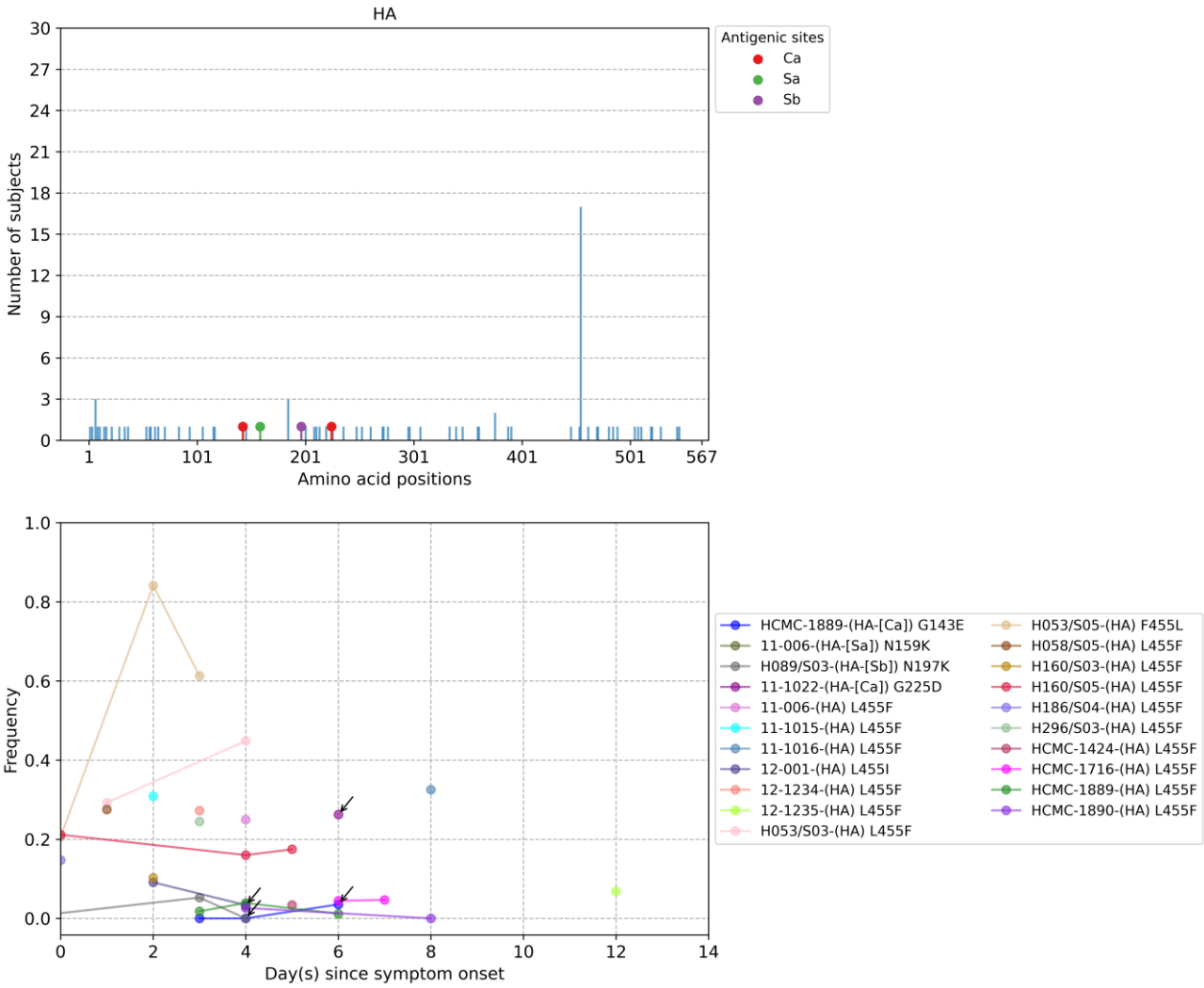


Figure 4 – figure supplement 5: Plots of haemagglutinin (HA) intra-host amino acid variants in A/H1N1pdm09 infected individuals. Top panel shows the number of subjects where nonsynonymous variants were found in the respective protein site. Different canonical antigenic sites of the haemagglutinin (HA) protein are colored (HA numbering based on H3 numbering without signal peptide). Bottom panel plots selected as well as parallel amino acid mutations found in multiple patients against days since illness onset. Filled circles represent days on which samples were collected and sequenced. The variant frequencies of all putative antigenic sites are also plotted. The frequencies of the five putative HA antigenic variants of A/H1N1pdm09 viruses are marked by arrows for better clarity.

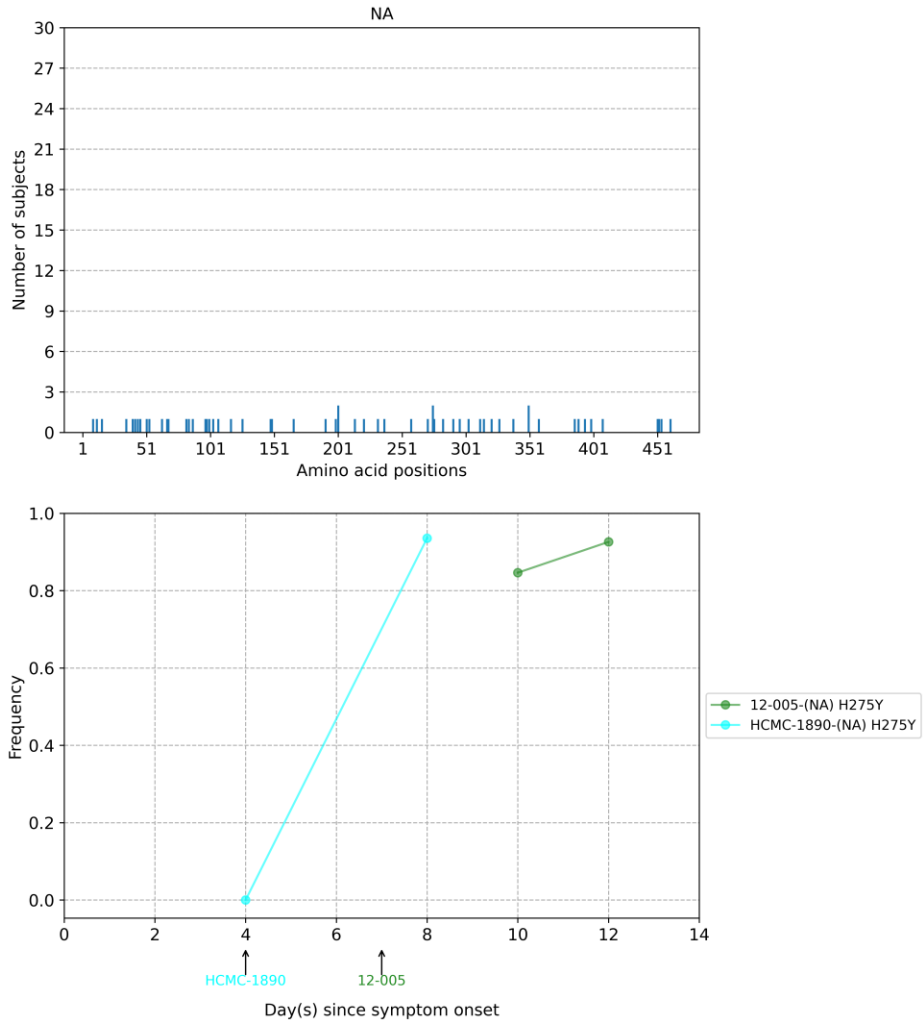


Figure 4 – figure supplement 6: Plots of neuraminidase (NA) intra-host amino acid variants in A/H1N1pdm09 infected individuals. Top panel shows the number of subjects where nonsynonymous variants were found in the respective protein site. Bottom panel plots selected as well as parallel amino acid mutations found in multiple patients against days since illness onset. Filled circles represent days on which samples were collected and sequenced. The first day of oseltamivir treatment for individuals with resistance mutations is annotated below the x-axis.

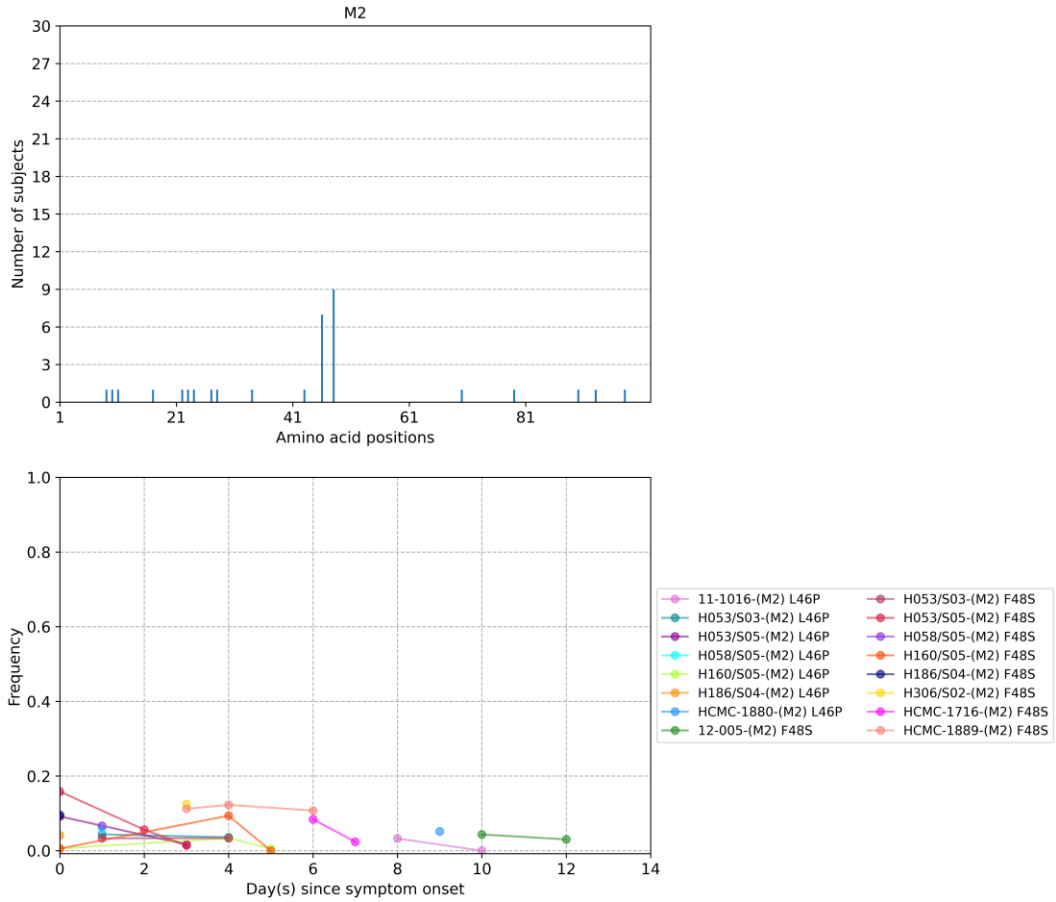


Figure 4 – figure supplement 7: Plots of M2 protein intra-host amino acid variants in A/H1N1pdm09 infected individuals. Top panel shows the number of subjects where nonsynonymous variants were found in the respective protein site. Bottom panel plots selected as well as parallel amino acid mutations found in multiple patients against days since illness onset. Filled circles represent days on which samples were collected and sequenced.

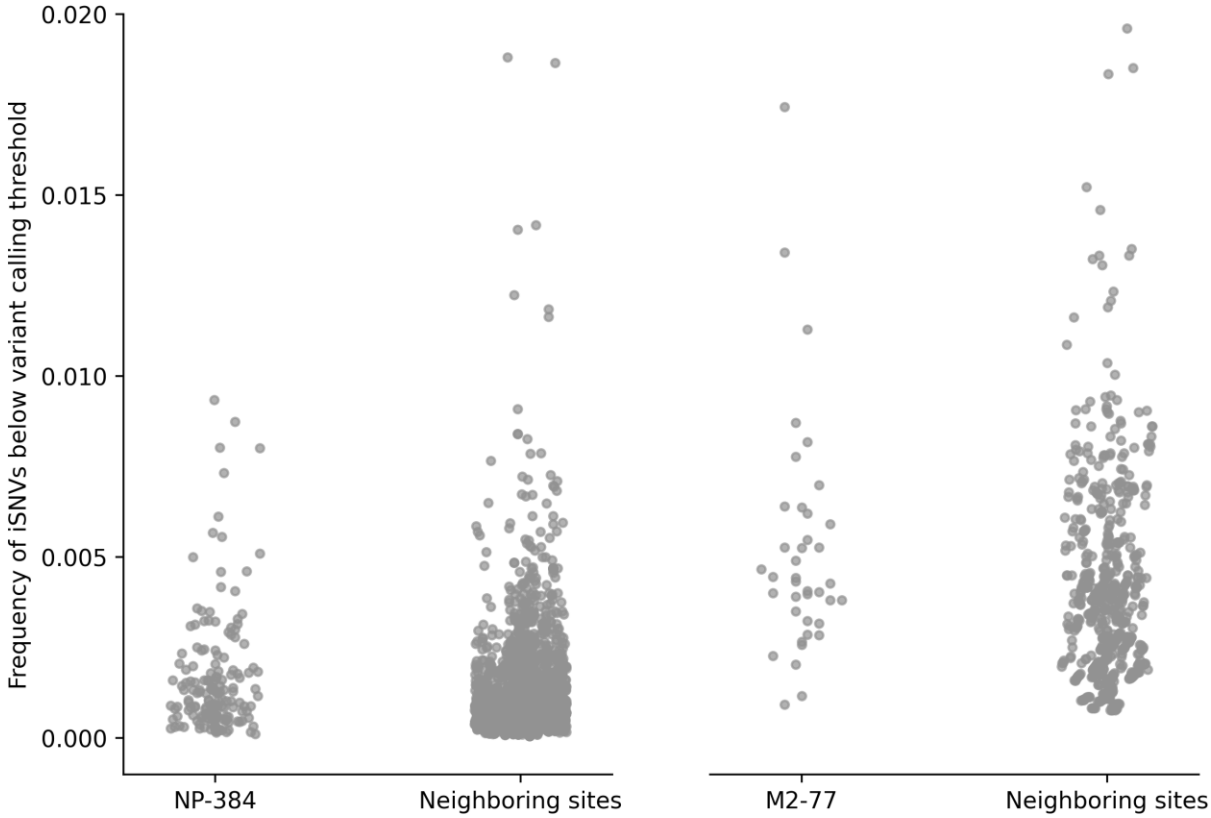


Figure 4 – figure supplement 8: Frequency distributions of iSNVs below the 2% variant calling threshold found in nucleotide positions NP-1150 and M-917 that encode for amino acid sites NP-384 and M2-77 respectively. All A/H3N2 virus samples collected from all patients with site coverage above the 100x are included. The distributions were compared to that of neighbouring sites, ±10 nucleotide positions adjacent to NP1150 and M-917.

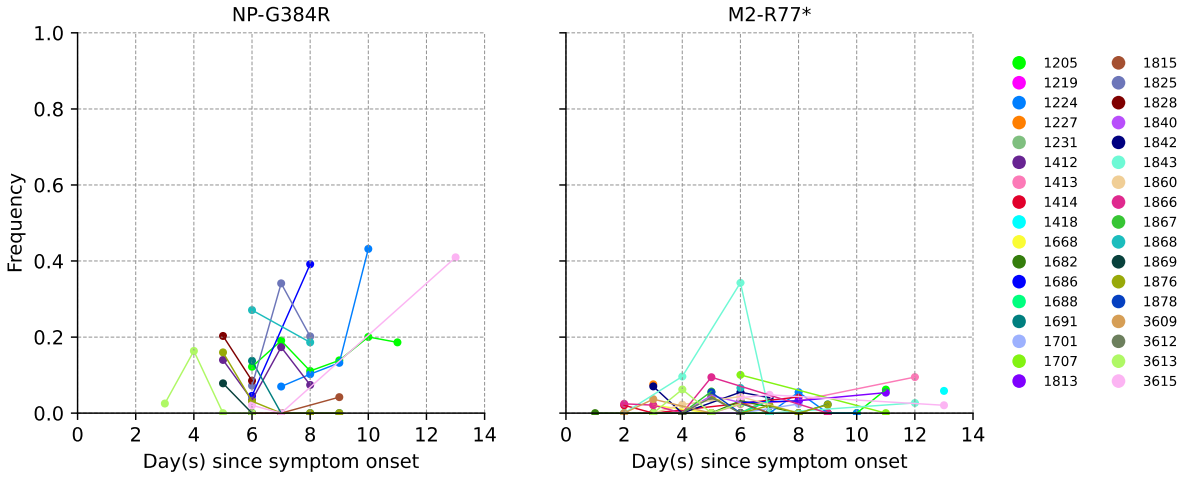


Figure 4 – figure supplement 9: Plots of within-host recurring A/H3N2 amino acid variants NP-G384R and M2-R77* based on variant calls and frequencies after remapping sample reads to their respective sample consensus sequence.

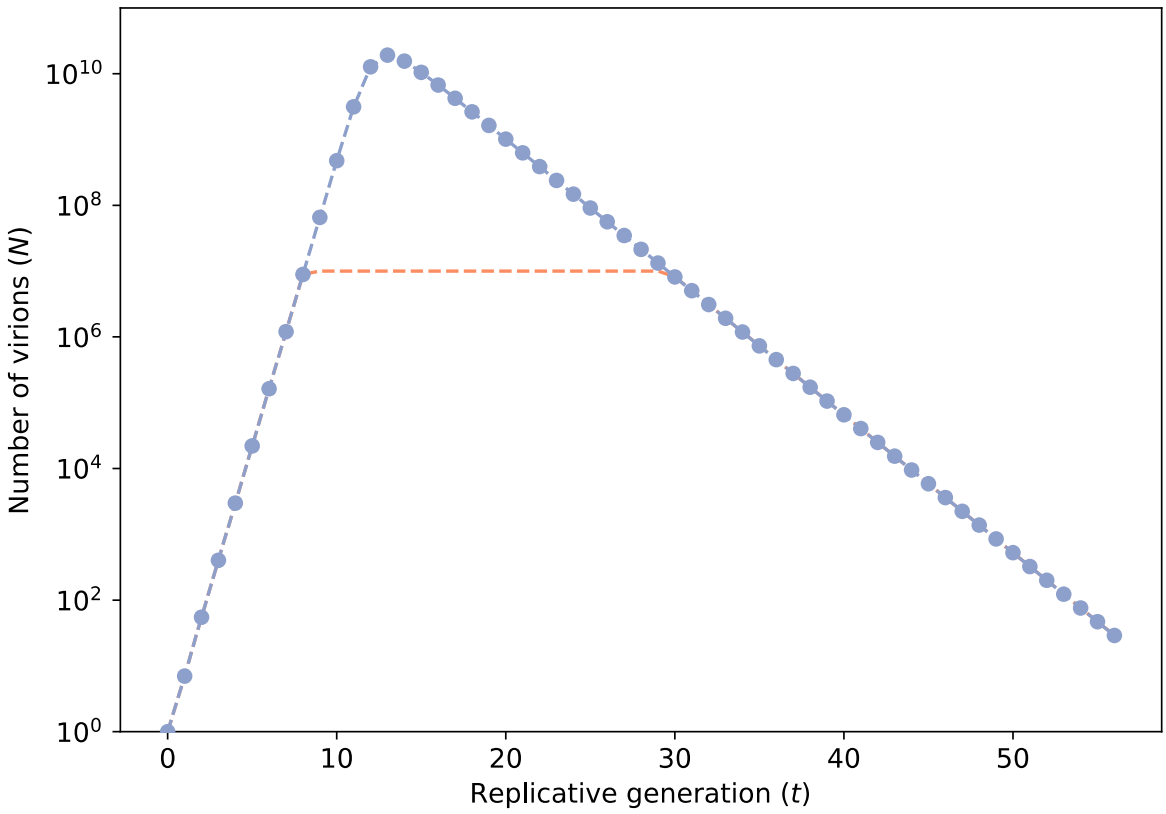


Figure 6 – figure supplement 1: Number of virions (N) against replicative generation (t) based on a target cell-limited within-host model. Blue line with markers denotes the population size computed from the model. When $N > 10^7$ virions, we assumed that N remained constant at 10^7 (pink dashed line) to reduce computational costs of forward-time simulations.