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]Laboratory of Applied Evolutionary Biology, 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 within-host virus diversity of infections in otherwise 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/H3N2 or A/H1N1pdm09 viruses between 2007 and 2009. For A/H1N1pdm09 infections during the 2009 pandemic, nonsynonymous changes were common early in infection but decreased or remained constant throughout infection. For A/H3N2 viruses, early infection was dominated by purifying selection. However, as infections progressed, nonsynonymous variants increased in frequencies even though within-host virus titres decreased, leading to the maintenance of virus diversity via mutation-selection balance. Our findings suggest that this maintenance of genetic diversity in these children combined with their longer duration of infection may provide 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-

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

94

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

100

101 Results

102 *Study participants*

103 The A/H3N2 virus samples were collected from 51 unlinked individuals as part of an
104 oseltamivir dosage trial[20,21]. 48 of the 51 A/H3N2 virus infected individuals were young
105 children (median age=2 years; interquartile range (IQR)=2-3 years) at the time of sampling
106 and most had low or no detectable anti-influenza virus antibody titers on day 0 and 10 post-
107 symptom onset[21]. Given that young children are substantial contributors to influenza virus
108 transmission[22,23], the samples analysed here offer a valuable opportunity to investigate the
109 within-host IAV evolutionary dynamics in this key population. The A/H1N1pdm09 virus
110 specimens were collected from 32 individuals up to 12 days post-symptom onset. These
111 individuals include both children and adults (median age=10 years; IQR=4-20 years) infected
112 during the first wave of the pandemic in Vietnam (July-December 2009). 15 of the 32
113 individuals (including 6 index patients) were sampled in a household-based influenza cohort
114 study[24]. The remaining 16 unlinked individuals were hospitalised patients that were involved
115 in two different oseltamivir treatment studies[20,25] Details of all study participants are
116 described in the respective cited studies and Table S4.

117

118 *Genetic diversity of within-host virus populations*

119 We used the number of minority intra-host single nucleotide variants (iSNVs; ≥ 2% in
120 frequencies) to measure the levels of genetic diversity of within-host IAV populations.
121 Similar to previous studies[4,8,9,11], within-host genetic diversity of human A/H3N2 virus
122 populations was low (median = 11 iSNVs, interquartile-range (IQR) = 7-16; Figure 1A).
123 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).

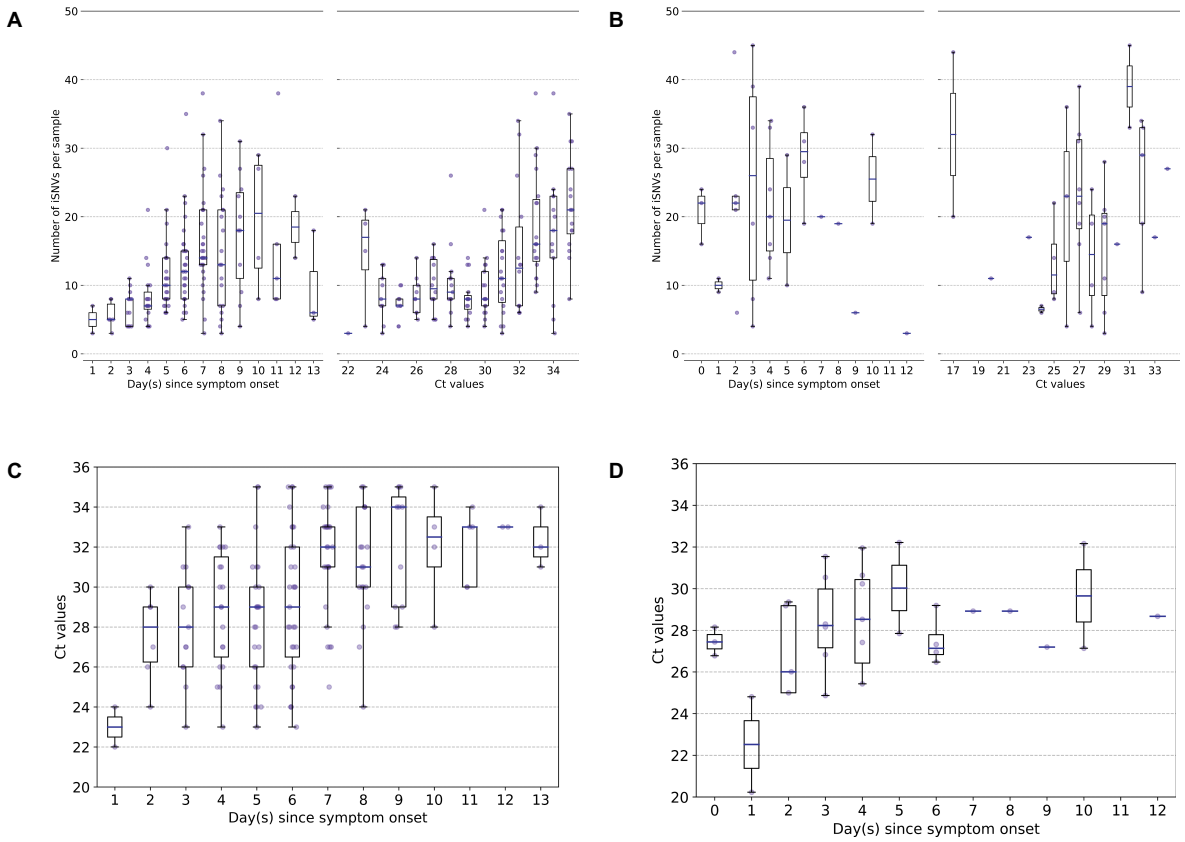


Figure 1: Genetic diversity of within-host influenza A virus populations. Box plots summarizing the number of intra-host single nucleotide variants (iSNVs; median, interquartile range (IQR), and whiskers extending within median $\pm 1.5 \times$ IQR) identified in samples with adequate breadth of coverage 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**).

144 *Within-host evolutionary rates of influenza A viruses*

145 To investigate within-host evolutionary dynamics, empirical rates of synonymous, non-
146 synonymous, and premature stop-codon (i.e. nonsense) iSNVs were calculated by
147 normalizing the summation of observed iSNV frequencies with the number of available sites
148 and time since symptom onset (see Methods). The overall within-host evolutionary rates of
149 A/H3N2 viruses observed here are in the same order of magnitude ($<\sim 10^{-5}$ divergence per
150 site per day) as those reported in previous within-host seasonal influenza virus evolution
151 studies (Figure 2A)[26]. Synonymous evolutionary rates were significantly higher than
152 nonsynonymous rates during the initial phase of A/H3N2 virus infections (Figure 2A),
153 primarily in the polymerase complex and HA genes (Figure 2A and S1-2). Importantly,
154 nonsynonymous variants gradually accumulated, increasing in rates around four days post-
155 symptom onset to similar levels relative to synonymous rates. Aggregating over all samples,
156 most nonsynonymous variants were found in the nucleoprotein (NP) and neuraminidase (NA)
157 gene segments (nonsynonymous to synonymous variant (NS/S) ratios = 1.69 (NP) and 1.32
158 (NA) whereas NS/S ratios were ≤ 1 for all other gene segments; Figure S1 and Table S1).
159 While nonsynonymous NA mutations associated with oseltamivir resistance were positively
160 selected for a subset of individuals in response to the antiviral treatment[21], nonsynonymous
161 changes to NP were likely mediated by protein stability, T-cell immune response and/or host
162 cellular factors (see next section).

163

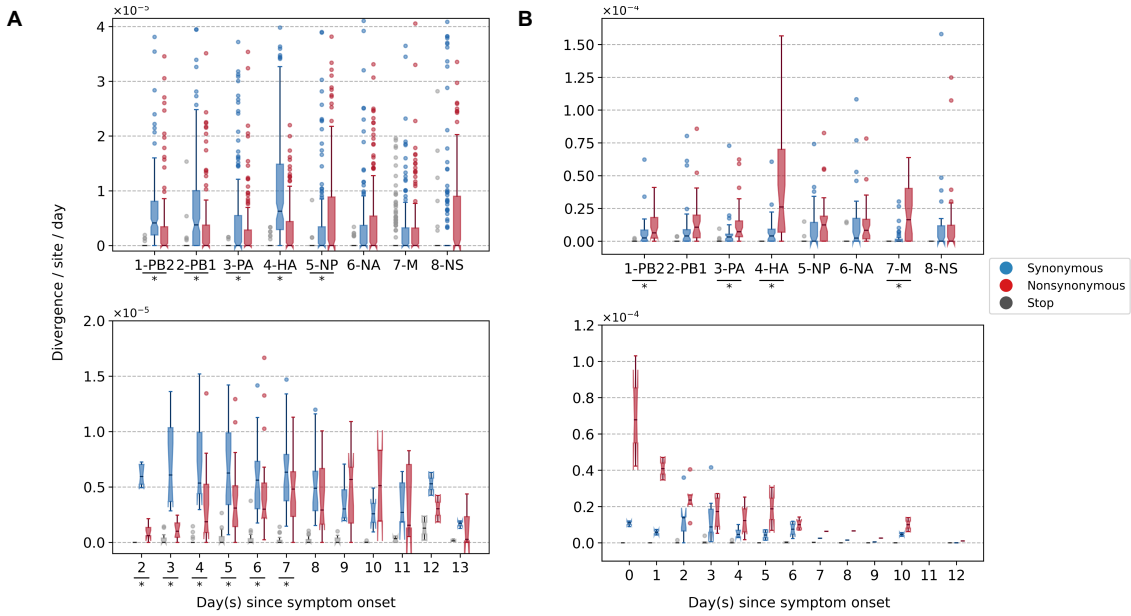
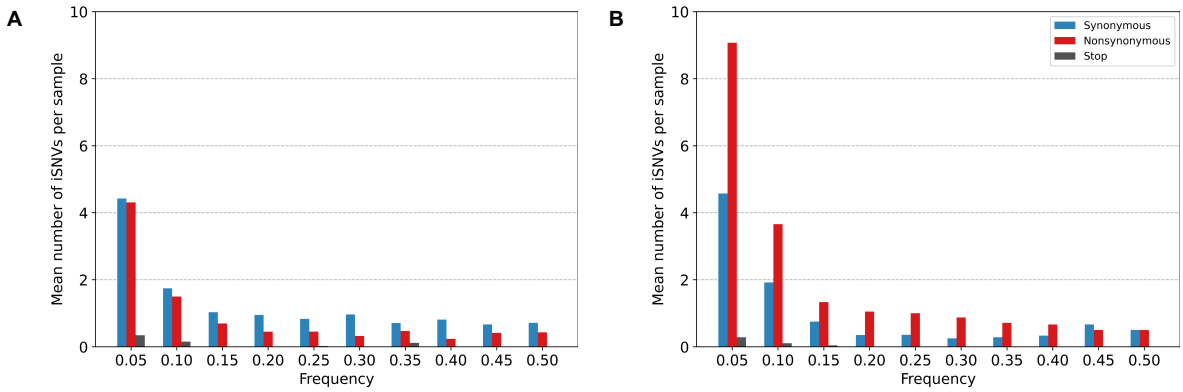


Figure 2: Box plots (median, interquartile range (IQR), and whiskers extending within median $\pm 1.5 \times$ IQR) summarizing the empirical within-host evolutionary rates of (**A**) seasonal A/H3N2 viruses and (**B**) pandemic A/H1N1pdm09 viruses. Top panel shows the evolutionary rate of individual 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). Nonsynonymous evolutionary rates were higher than synonymous rates from the start of symptom onset when overall evolutionary rates were also the highest. However, we were unable to determine if the per-day post-symptom onset nonsynonymous and synonymous rates were significantly different from each other due to the low number of samples (i.e. median = 2 samples per day post-symptom onset). Nonetheless, consolidating over all samples across all time points, the polymerase basic 2 (PB2), polymerase acidic (PA), HA and matrix (M) gene segments were the main contributors to the observed rate disparity (Figure 2B and S3-4) with nonsynonymous variants emerging at significantly higher rates relative to synonymous ones. All gene segments also yielded NS/S ratios > 1 (Table S1).

191 *Intra-host minority variants*

192



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

196

197 Most of the iSNVs identified for both virus subtypes were observed at low frequencies (2-
198 5%; Figure 3), and appear to be stochastically introduced across the virus genome (Figure 4).
199 Purifying selection dominated within-host seasonal A/H3N2 virus populations as the ratio of
200 nonsynonymous to synonymous variants was 0.72 across all samples and variant frequencies
201 (Figures 3A and S2). Of note, the canonical antigenic sites of the HA gene segment[27] of the
202 A/H3N2 virus populations experienced strong negative selection as evidenced by the
203 occurrence of synonymous variants (median frequency = 0.14, IQR range = 0.09-0.27) at far
204 greater frequencies relative to those at non-antigenic sites of HA (median frequency = 0.03,
205 IQR range = 0.03-0.05; Mann-Whitney U test $p = 1.18 \times 10^{-24}$; Figure 4C). There were no
206 significant differences in the frequencies of nonsynonymous iSNVs between the antigenic
207 sites of H3 (median frequency = 0.04, IQR range = 0.03-0.06) and the rest of the HA gene
208 segment (median frequency = 0.03, IQR range = 0.02-0.06; Mann-Whitney U test $p = 0.29$;
209 Figure 4C). In contrast, there was 1.94 times as many nonsynonymous minority iSNVs
210 relative to synonymous ones identified in the pandemic A/H1N1pdm09 virus samples
211 (Figures 3B and S4). Variant frequencies of nonsynonymous iSNVs found in the antigenic
212 epitopes of H1[28] (median frequency = 0.04, IQR range = 0.04-0.05) were, however, not
213 significantly different from those of non-antigenic sites (median frequency = 0.05, IQR range
214 = 0.03-0.16; Mann-Whitney U test $p = 0.34$; Figure 4D).

215

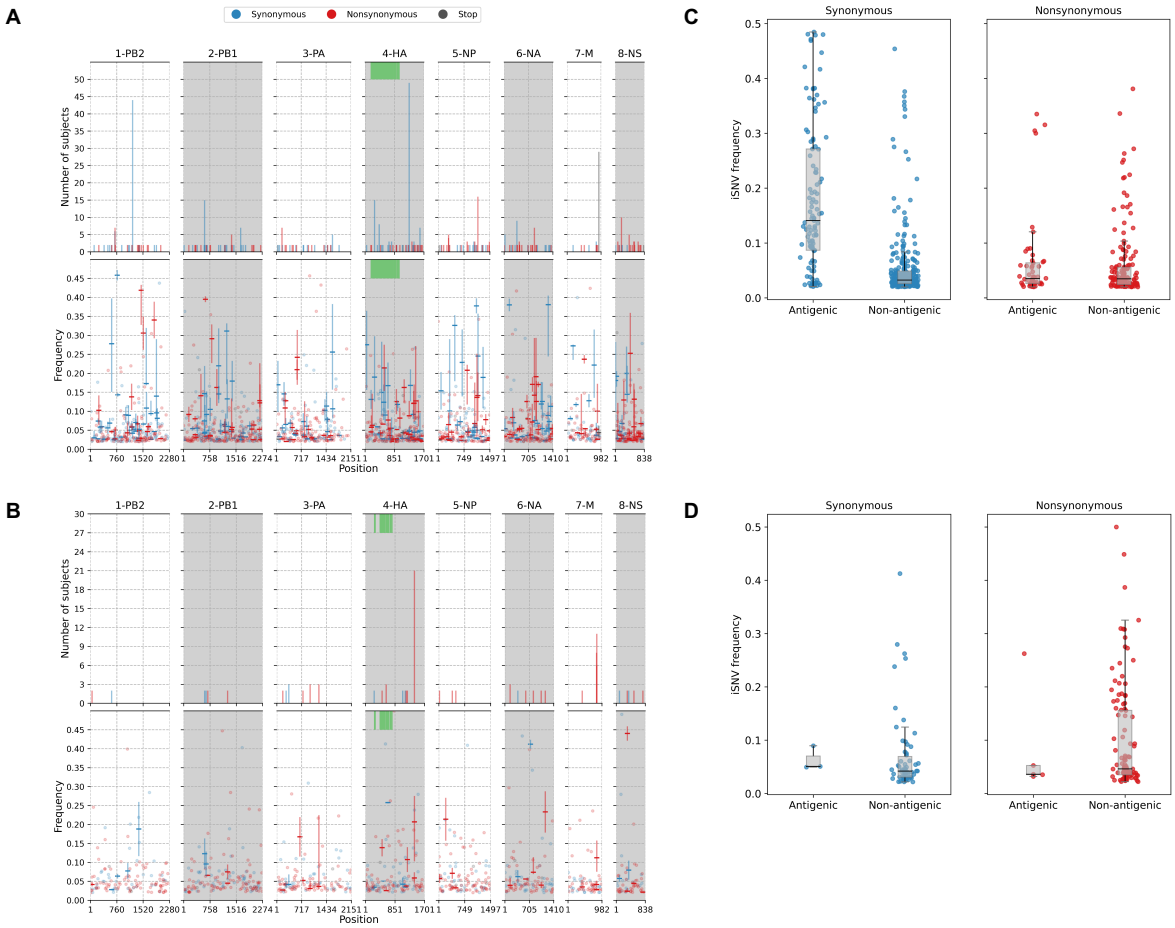


Figure 4: (**A**) Breakdown of iSNVs identified in seasonal A/H3N2 virus samples. The top panels plot the nucleotide positions where iSNVs were found in at least two subjects. The bottom panels shows the frequencies at which iSNVs were identified. For sites with iSNVs that were found in two or more subjects, 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**) Similar plots to (A) for iSNVs found in pandemic A/H1N1pdm09 virus samples. (**C**) Box plots of the frequencies of synonymous and nonsynonymous variants between antigenic and non-antigenic sites of seasonal A/H3N2 haemagglutinin gene segment. (**D**) Similar plots to (C) 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

237 A/H3N2 viruses, nonsense mutation rates were much lower compared to the synonymous and
238 nonsynonymous counterparts (median genome-wide rate across all samples between 0 and
239 1.43×10^{-6} divergence per site per day; IQR limits between 0 and 2.18×10^{-6} divergence
240 per site per day).

241

242 The premature stop-codon mutations were mostly found at low frequencies for both influenza
243 subtypes (<10%; Figure 3). The exception lies with one of the A/H3N2 virus samples where
244 a premature stop codon was found in position 77 of the M2 ion channel with variant
245 frequency as high as 34.6% (Patient 1843, day 6 since symptom onset; Figures 3A and S5D).
246 The premature stop codon in M2-77 was also found in 27 other individuals across multiple
247 timepoints, albeit at a much lower frequency that never amounted more than 10% (Figures
248 4A and S5D). This was unlikely to be a sequencing artefact resulting from a mistaken
249 incorporation of the primer sequence as its carboxyl terminal falls outside the coding region
250 of the M gene segment (Table S3) and the variant frequencies would have been much higher
251 in all samples if this was the case.

252

253 Despite the dominance of purifying selection in seasonal A/H3N2 intra-host viral
254 populations, we detected several nonsynonymous variants of interest. Amino acid variants
255 emerging in the HA and NA proteins were discussed in a previous work[21] (see
256 Supplementary Materials). In the nucleoprotein, there were two notable nonsynonymous
257 variants, D101N/G and G384R, that appeared in multiple individuals who were sampled
258 independently between 2007 and 2009 (Figure 4A and S5C). D101N/G was found in 7
259 different patients and at least for D101G, the mutation was previously linked to facilitating
260 escape from MxA, a key human antiviral protein[30]. However, the nonsynonymous mutation
261 was only found in low frequencies and remained invariant during the respective courses of
262 infection for all seven patients (median variant frequency across all samples = 0.03; IQR =
263 0.02-0.07).

264

265 NP-G384R emerged in sixteen unlinked patients infected by A/H3N2 virus. Even though
266 G384R did not become the majority variant in any of these individuals (median variant
267 frequency across all samples = 0.14; IQR = 0.07-0.20), the variant emerged around day 4-5
268 post-symptom onset and mostly persisted within each individual for the rest of sampled
269 timepoints. G384R is a stabilizing mutation in the A/Brisbane/10/2007 A/H3N2 virus NP
270 background[31] that is similar to the viruses investigated here. Interestingly, position 384 is an
271 anchor residue for several NP-specific epitopes recognised by specific cytotoxic T
272 lymphocytes (CTLs) that are under continual selective pressure for CTL escape[32,33]. The
273 wild-type glycine residue is known to be highly deleterious even though it was shown to
274 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 Table S2). 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; Table S2). 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].

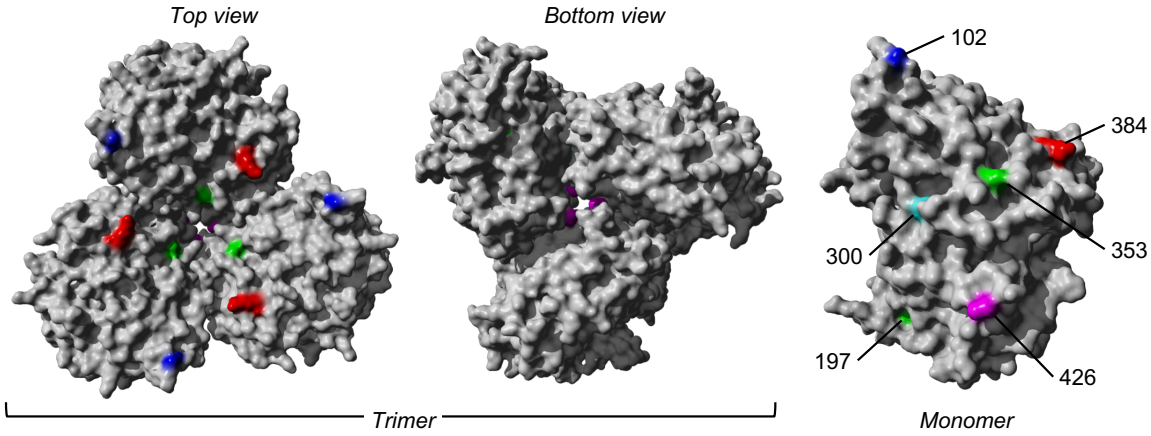


Figure 5: The trimeric and monomeric crystal structures of nucleoprotein (PDB: 3ZDP)[40] of influenza A viruses. Amino acid sites with potentially linked epistatic amino acid variants as tabulated in Table 1 are separately coloured, with their corresponding positions annotated on the monomeric structure.

Variants	$\Delta\Delta G$ (kcal/mol) 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 S5E). 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; Table S4) 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 S5E). The amino acid variant was found in patients from different time periods and geographical locations (Table S4), 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.

330 We also found oseltamivir resistance mutation H275Y[42] in the NA proteins in two unlinked
331 individuals who were infected with the A/H1N1pdm09 virus and treated with oseltamivir
332 (Figure S5F and Table S4). 275Y quickly became the majority variant in both patients within
333 3-4 days after the antiviral drug was first administered. Finally, there were two other amino
334 acid variants in the M2 ion channel that appeared within multiple subjects in parallel across
335 different geographical locations – L46P and F48S were identified in 8 and 16 patients
336 respectively in a range of frequencies (L46P: median frequency = 0.04, IQR = 0.04-0.05;
337 F48S: median frequency = 0.08, IQR = 0.03-0.13) but similarly, never becoming a majority
338 variant in any of them (Figures 4B and S5G). Again, the wild-type residues were mostly
339 conserved in the global virus population since the pandemic.

340

341 *Within-host simulations*

342 To investigate the evolutionary pressures that likely underpin the observed patterns of
343 synonymous and nonsynonymous substitutions (Figure 2), we performed forward-time
344 Monte Carlo simulations. Given that the median age of the children infected by A/H3N2
345 virus at the time of sample collection was 2 years of age (IQR=2-3 years), most of them were
346 likely experiencing one of their first influenza virus infections. Furthermore, influenza
347 vaccination for children is not part of the national vaccination programme in Vietnam. As
348 such, most of the children analysed here lacked influenza virus specific antibodies based on
349 haemagglutination inhibition assays[21]. For individuals infected by the pandemic
350 A/H1N1pdm09 virus, all but one patient was under 60 years of age and thus lacked immunity
351 to the virus as well. Furthermore, patients infected by either viruses mount little-to-no
352 humoral immune selection pressure during the first 7-10 days of infection[43]. As such, the
353 contrasting evolutionary dynamics between these viruses (Figure 2) are unlikely to have been
354 driven by antibody-mediated selection pressure.

355

356 Seasonal A/H3N2 viruses, having circulated within the human population since 1968, are
357 expected to be well adapted to human hosts at this point such that most nonsynonymous
358 mutations are likely highly deleterious and would not reach detectable frequencies. Those
359 that were detected are mostly expected to be weakly deleterious, and thus not purged fast
360 enough by selection such that mutation-selection balance was observed. In contrast, there was
361 evolutionary space for A/H1N1pdm09 virus to further adapt to its new found human hosts
362 during the initial waves of the pandemic. Since no mutation selected by rapid directed
363 positive selection was observed, most of the detected nonsynonymous mutations were
364 expected to be neutral and a small but non-trivial fraction are likely to be weakly beneficial.

365

366 Our simulations used a simple within-host evolution model represented by a binary genome
367 that distinguishes between synonymous and nonsynonymous loci. Given that the estimated

368 transmission bottleneck sizes for pandemic A/H1N1pdm09 (see Supplementary Materials)
369 and seasonal A/H3N2 viruses[4,44] are narrow at 1-2 genomes, we modelled an expanding virus
370 population size during the initial timepoints of the infection that started with one virion. If
371 within-host virus populations were to evolve neutrally, we would observe similar
372 synonymous and nonsynonymous evolutionary rates throughout the infection (Figure 6C). On
373 the other hand, if selection is sufficiently strong, accumulation of beneficial (or deleterious)
374 nonsynonymous variants will increase (or decrease) substantially with time (Figure 6B and
375 6E). Clearly, these patterns were not observed for both IAVs (Figure 2).

376

377 However, if most *de novo* nonsynonymous mutations are only weakly deleterious, we would
378 observe larger synonymous evolutionary rates initially before nonsynonymous variants
379 accumulate to similar levels (Figure 6A). By then, virion population size (N) would also be
380 large enough relative to the virus mutation rate (μ) (i.e. $N\mu \gg 1$; see Supplementary
381 Materials) such that mutation-selection balance is expected and evolutionary rates remain
382 fairly constant, similar to the patterns empirically observed for within-host A/H3N2 virus
383 populations (Figure 2A). Contrastingly, if the majority of nonsynonymous variants are
384 neutral and only a small subset confers weakly beneficial effects, nonsynonymous
385 evolutionary rates would consistently be larger than their synonymous counterpart but never
386 accumulate to levels akin to those observed for strong positive selection (Figure 6D).
387 Although the simulation results here does not entirely reflect the evolutionary dynamics
388 observed for A/H1N1pdm09 viruses in Figure 2B, we hypothesised that there was substantial
389 virus replication prior to symptom onset and that our samples better reflect the virus
390 populations present midway or nearing the end of the infection when compared to our
391 simulation results. This is further evidenced by the relatively large number of iSNVs detected
392 at the time of symptom onset (Figure 1B) and the tight transmission bottleneck sizes we
393 estimated for the pandemic virus (see Supplementary Materials).

394

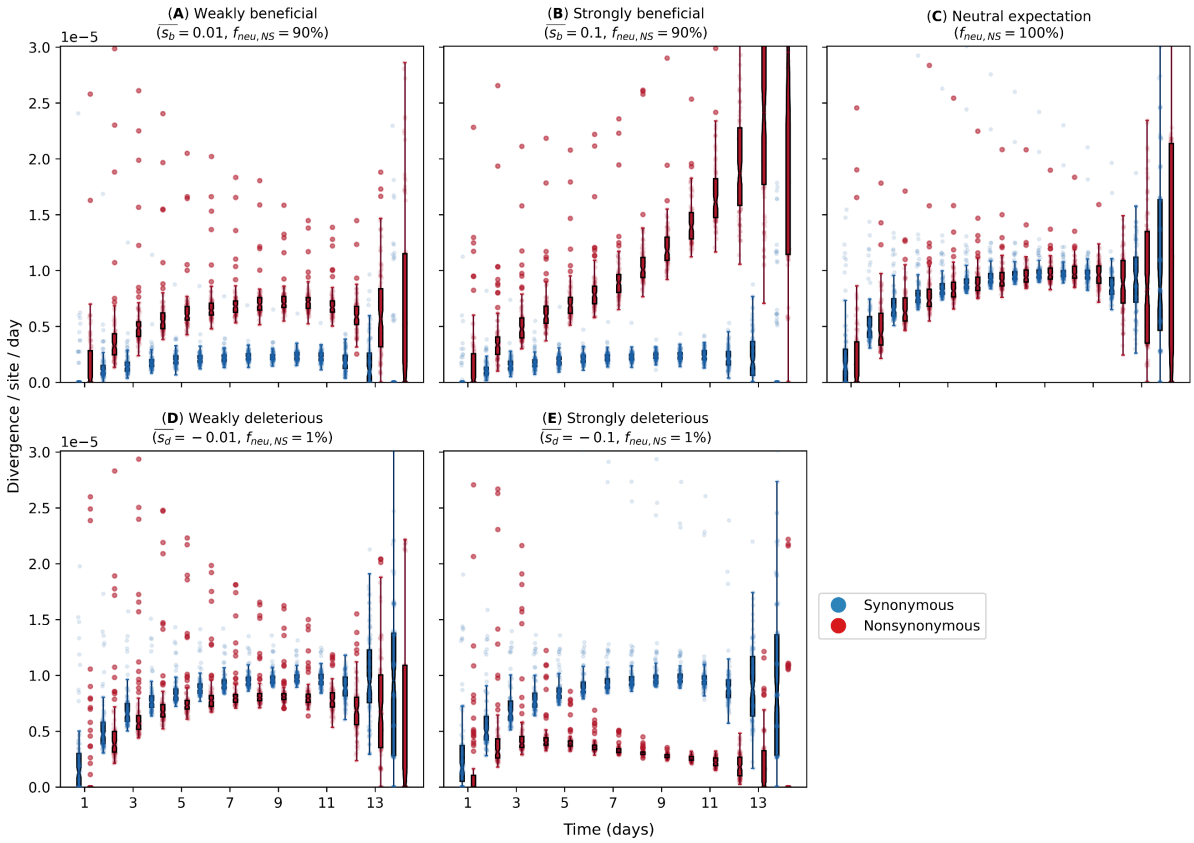


Figure 6: Evolutionary rates computed from forward-time Monte Carlo within-host simulations for different fitness effects of nonsynonymous mutations ($\overline{s_b}$ and $\overline{s_d}$ denote mean beneficial and deleterious effects respectively) and fraction of neutral nonsynonymous mutations ($f_{neu,NS}$). We assumed that synonymous mutations are neutral for all simulations. For A/H1N1pdm09 viruses, we assumed that only a small fraction of nonsynonymous mutations is neutral ($f_{neu,NS} = 1\%$) and performed simulations where the remaining nonsynonymous mutations are either (**A**) weakly ($\overline{s_b} = 0.01$) or (**B**) strongly ($\overline{s_b} = 0.1$) beneficial. For A/H3N2 viruses, we tested the hypotheses where majority of nonsynonymous mutations are neutral ($f_{neu,NS} = 90\%$) while the remaining ones are either (**D**) weakly ($\overline{s_d} = -0.01$) or (**E**) strongly ($\overline{s_d} = -0.1$) deleterious. (**C**) Neutral expectation where all nonsynonymous mutations are neutral ($f_{neu,NS} = 100\%$).

Discussion

Multiple next-generation sequencing studies have found little evidence of positive selection in seasonal influenza virus populations of acutely infected individuals[4,8–11,45]. 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[46]. 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[47,48]. 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[49]. In other words, the

417 duration of infection is likely to be critical for positive evolutionary selection to be effective
418 within host.

419

420 Viral shedding duration is often longer in young children infected with seasonal influenza
421 virus compared to otherwise healthy adults[50]. Children also play a critical role in "driving"
422 influenza epidemics due to their higher contact and transmission rates[22,23]. As such, our
423 seasonal A/H3N2 virus results fill an important gap in the current literature of within-host
424 evolutionary studies of seasonal IAVs as most of the samples analysed were collected from
425 children under the age of six years up to two weeks post-symptom onset. Importantly, the
426 absence of antibody-mediated immunity in young unvaccinated children, which would
427 otherwise reduce the extended duration of infection, has the potential to facilitate other routes
428 of virus evolution.

429

430 Similar to the aforementioned within-host studies, the A/H3N2 virus population within these
431 children was characterised by low genetic diversity and dominated by purifying selection
432 early in the infection. Due to a lack of antibody response against the antigenic regions of
433 HA[21], it is unsurprising that we observed a lack of adaptive changes to the HA antigenic
434 regions, similar to adults in previous studies[4]. We also found that the polymerase genes were
435 subjected to purifying selection, indicating their critical role in virus replication as negative
436 selection purges deleterious variation. However, while purifying selection is detectable, it is
437 incomplete[26]. We observed that most nonsynonymous variants began to accumulate around
438 3-4 days post-symptom onset, with incrementally higher empirical rates as the infection
439 progressed.

440

441 Through simulations of a within-host evolution model, we hypothesised that the
442 accumulation of nonsynonymous iSNVs was a result of their weakly deleterious effects and
443 expanding virion population size such that mutation-selection balance was reached. The
444 maintenance of genetic diversity through mutation-selection balance within these children
445 may provide opportunities for the emergence of phenotypically relevant mutations which
446 deleterious effects could be alleviated by the accumulation of a secondary compensatory
447 mutations. For example, in one individual NP-G384R was accompanied by NP-M426I which
448 is an anchor residue of a CTL epitope of NP, abrogating recognition by HLA-B*3501-
449 positive CTLs[32] but is likely to be deleterious based on our computational protein stability
450 predictions. G384R, which is located in a CTL epitope distinct from M426I[32], was previously
451 shown to be a stabilizing substitution[31].

452

453 Interestingly, we also observed G384R in the minority virus population of 15 other unlinked
454 individuals. Besides improving NP protein stability, G384R restores recognition by HLA-
455 B*2705-positive NP-specific CTLs[36]. The NP gene segment in the global A/H3N2 virus
456 population has an evolutionary history of fixating destabilizing amino acid mutations that
457 promote CTL immune escape alongside stabilizing substitutions that compensate for the
458 deleterious effects of the former[34]. The reversal R384G mutation confers CTL escape but is
459 known to be highly deleterious. This substitution was fixed in the global A/H3N2 virus
460 population during the early 1990s as other substitutions such as S259L and E375G
461 epistatically alleviated its destabilizing effects[34]. One possible explanation for the emergence
462 of G384R as a minority variant within these unlinked individuals is that they are all HLA-
463 B*2705 negative. However, we did not collect the necessary blood samples to investigate this
464 possibility.

465

466 In contrast, we found a substantially higher fraction of nonsynonymous variants in the
467 within-host virus populations of individuals infected A/H1N1pdm09 virus during the
468 pandemic. Owing to the different next-generation sequencing platforms used to sequence
469 samples of the two virus subtypes and consequently differences in base calling error rates and
470 depth of coverage (Figure S6), we did not directly compare the observed levels of within-host
471 genetic diversity between the two influenza subtypes here. However, given that only iSNVs
472 with frequencies ≥ 2% were called, low-frequency minority variants arising from technical-
473 related errors should be minimised[51]. Importantly, the relative number of nonsynonymous
474 iSNVs identified were far greater than synonymous ones early in the pandemic
475 A/H1N1pdm09 virus infections, suggesting that there was room for further human host
476 adaptation, particularly in the HA but also in the polymerase gene segments similar to those
477 observed in other zoonotic influenza virus infections[52].

478

479 Given the tight estimated transmission bottleneck size (see Supplementary Materials), the
480 relatively large number of iSNVs identified at the start of symptom onset and simulations of
481 within-host evolution (see Supplementary Materials, Figure 1B, 2B and 6D), it is unlikely
482 that the initial within-host A/H1N1pdm09 virus populations sampled were the inoculating
483 population that founded the infection. Instead, the inoculating viral population had already
484 undergone substantial within-host replication during the incubation period before symptom-
485 onset. In fact, four of the individuals analysed were asymptomatic (i.e. H058/S02, H089/S04,
486 H186/S05 and H296/S04; Table S4). Additionally, pre-symptomatic virus shedding was
487 observed in some of the secondary household cases[53] and presymptomatic transmission has
488 been documented in other settings[54]. Nonetheless, this would not meaningfully impact our
489 conclusions as most of the within-host viral populations sampled at the start of symptom
490 onset should still constitute those found early in infection and the contrasting feature where

491 nonsynonymous iSNVs outnumbered synonymous ones were not observed in the seasonal
492 A/H3N2 virus samples.

493

494 For both A/H3N2 and A/H1N1pdm09 virus samples, nonsense iSNVs resulting in premature
495 stop codons were found to accumulate within host, even though only at low proportions. The
496 accumulation of premature stop-codon mutations further suggest that while purifying
497 selection dominates within-host influenza virus populations, it may not be acting strongly
498 enough to completely purge these lethal nonsense mutations[26]. Additionally, it has been
499 recently found that incomplete influenza virus genomes frequently occur at the cellular level
500 and that efficient infection depends on the complementation between different incomplete
501 genomes[55]. As such, nonsense mutations may not be as uncommon as previously thought. In
502 particular, nonsense mutations in position 77 of the M2 ion channel were independently
503 found in 27 unlinked individuals infected by A/H3N2 virus. While these nonsense mutations
504 are generally considered to be lethal, ion channel activity is retained even if the M2 protein
505 was prematurely truncated up to position 70 at its cytoplasmic tail[56].

506

507 Our study has several limitations. The number of iSNVs identified can potentially be biased
508 by variations in sequencing coverage[57]. As such, the number of iSNVs observed in one intra-
509 host virus populations may not be directly comparable to another with a distinct coverage
510 profile (Figure S6). As an alternative, the nucleotide diversity π statistic[58] may be a more
511 robust measure of within-host diversity as it solely depends on the underlying variant
512 frequencies[57]. Computing the corresponding π statistics for our data, we observed trends in
513 genetic diversity that were similar to those inferred using iSNV counts (see Supplementary
514 Materials and Figure S7).

515

516 To ensure accurate measurements of virus diversity in intra-host populations, we would also
517 need to be certain that the estimated variant frequencies precisely reflect the distributions of
518 variants that comprise the sampled virus populations. The inferred variant frequencies can be
519 significantly distorted if virus load is low[59,60]. As such, we limited our analyses for both virus
520 subtypes to samples with Ct-values $\leq$35 which likely afford sufficient virus material for
521 sequencing[60]. We were unable to estimate the amount of frequency estimation errors for the
522 A/H1N1pdm09 virus samples as only one sequencing replicate was performed using the
523 universal 8-segment PCR method[61]. However, for the A/H3N2 virus samples, independent
524 PCR reactions were performed using three partly overlapping amplicons for all gene
525 segments other than the non-structural and matrix genes. We compared the variant
526 frequencies estimated for any overlapping sites generated by reads derived from distinct
527 amplicons with sufficient coverage (>100x). Variant frequencies computed from independent
528 amplicons agreed well with each other across the range of Ct values of the samples from

529 which variants were identified (Figure S8), affirming the precision of our iSNV frequency
530 estimates for the A/H3N2 virus samples, including those with higher Ct values.

531

532 Finally, most study participants received oseltamivir during the course of their infections
533 (Table S4). Although we were unable to identify any potential effects of enhanced viral
534 clearance or any other evolutionary effects due to the treatment, besides oseltamivir-
535 resistance associated mutations, it is unlikely that the antiviral treatment had a substantial
536 impact on our results. First, the median timepoint in which the antiviral was initially
537 administered was 4 days post-symptom onset (IQR = 3-6 days; Table S4). Previous studies
538 showed that enhanced viral clearance of IAVs was mostly observed among patients who were
539 treated with oseltamivir within 3 days of symptom onset[20,62,63]. Of note, late timepoint
540 samples in this study (≥ 8 days since symptom onset) mostly came from individuals who
541 started oseltamivir treatments ≥ 4 days post-symptom onset (Figure S12). Second, at least *in*
542 *vitro*, there were no differences in the levels of genetic diversity observed in influenza virus
543 populations after multiple serial passages whether they were treated with oseltamivir or not[64].

544

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

551

552 **Methods**

553 *Sample collection and viral sequencing*

554 The A/H3N2 virus samples were collected from 52 patients between August 2007 and
555 September 2009 as part of an oseltamivir dosage trial conducted by the South East Asia
556 Infectious Disease Clinical Research Network (SEAICRN), which is detailed in a previous
557 work[20]. Briefly, patients with laboratory confirmed influenza virus infection and duration of
558 symptoms ≤ 10 days were swabbed for nose and throat samples daily between 0 and 10 days
559 as well as day 14 upon enrolment for the study (Table S4). All PCR-confirmed A/H3N2 virus
560 samples with cycle threshold (Ct) values ≤ 35 were included for sequencing.

561

562 Library preparation and viral sequencing protocols performed on these A/H3N2 virus
563 samples are elaborated in detail in [21]. Here, we highlight key aspects of our preparation and
564 sequencing procedures. Using segment specific primers (Table S3), we performed six

565 independent PCR reactions, resulting in three partly-overlapping amplicons for each
566 influenza virus gene segment other than the matrix (M) and non-structural (NS) genes where
567 a single amplicon was produced to cover the entirety of the relatively shorter M and NS
568 genes. The use of shorter but overlapping amplicons in the longer gene segments improve
569 amplification efficiency, ensuring that these longer segments are sufficiently covered should
570 there be any RNA degradation in the clinical specimen. These overlapping PCR products
571 were pooled in equimolar concentrations for each sample and purified for subsequent library
572 preparation. Sequencing libraries were prepared using the Nextera XT DNA Library
573 Preparation kit (Illumina, FC-131-1096) as described in [21]. Library pools were sequenced
574 using the Illumina MiSeq 600-cycle MiSeq Reagent Kit v3 (Illumina, MS-102-3003).

575

576 The A/H1N1pdm09 virus samples were obtained as part of a household-based influenza virus
577 cohort study that was also performed by SEAICRN. The study was conducted between July
578 and December 2009, involving a total of 270 households in Ha Nam province, Vietnam[24].
579 Similarly, combined nose and throat swabs were collected daily for 10-15 days from
580 individuals with influenza-like-illness (i.e. presenting symptoms of fever >38°C and cough,
581 or sore throat) and their household members, including asymptomatic individuals (Table S4).
582 We also analysed additional samples collected from unlinked hospitalised patients who were
583 infected by the A/H1N1pdm09 virus from two major Vietnamese cities (Hanoi and Ho Chih
584 Minh) during the first wave of the pandemic[20,25]. A total of 32 PCR-confirmed
585 A/H1N1pdm09-infected individuals originating from both households and hospitalised cases
586 were selected for sequencing based on availability and Ct-values ≤ 33 (Table S4).

587

588 For the A/H1N1pdm09 virus samples, RNA extraction was performed manually using the
589 High Pure RNA isolation kit (Roche) with an on-column DNase treatment according to the
590 manufacturer's protocol. Total RNA was eluted in a volume of 50 µl. Universal influenza
591 virus full-genome amplification was performed using a universal 8-segment PCR method as
592 described previously[51,65,66]. In short, two separate RT-PCRs were performed for each sample,
593 using primers common-uni12R (5'-GCCGGAGCTCTGCAGAT ATCAGCRAAAGCAGG-
594 3'), common-uni12G (5'-GCCGGAGCTCTG CAGATATCAGCGAAAGCAGG-3'), and
595 common-uni13 (5'-CAGGAA ACAGCTATGACAGTAGAAACAAGG-3'). The first RT-
596 PCR mixture contained the primers common-uni12R and common-uni13. The second RT-
597 PCR mixture contained the primers common-uni12G and common-uni13, which greatly
598 improved the amplification of the PB2, PB1, and PA segments. Reactions were performed
599 using the One-Step RT-PCR kit High Fidelity (Invitrogen) in a volume of 50 µl containing
600 5.0 µl eluted RNA with final concentrations of 1xSuperScript III One-Step RT-PCR buffer,
601 0.2 µM of each primer, and 1.0 µl SuperScript III RT/Platinum Taq High Fidelity Enzyme
602 Mix (Invitrogen). Thermal cycling conditions were as follows: reverse transcription at 42°C
603 for 15 min, 55°C for 15 min, and 60°C for 5 min; initial denaturation/enzyme activation of

604 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
605 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
606 extension of 68°C for 5 min. After the PCR, equal volumes of the two reaction mixtures were
607 combined to produce a well-distributed mixture of all 8 influenza virus segments. All RT-
608 PCRs were performed in duplicate. Samples were diluted to a DNA concentration of 50 ng/µl
609 followed by ligation of 454 sequencing adaptors and molecular identifier (MID) tags using
610 the SPRIworks Fragment Library System II for Roche GS FLX+ DNA Sequencer (Beckman
611 Coulter), excluding fragments smaller than 350 base pairs, according to the manufacturers
612 protocol to allow for multiplex sequencing per region. The quantity of properly ligated
613 fragments was determined based on the incorporation efficiency of the fluorescent primers
614 using FLUOstar OPTIMA (BMG Labtech). Emulsion PCR, bead recovery and enrichment
615 were performed manually according to the manufacturers protocol (Roche) and samples were
616 sequenced in Roche FLX+ 454. Sequencing was performed at the Sanger Institute, Hinxton,
617 Cambridge, England as part of the FP7 program EMPERIE. Standard flowgram format (sff)
618 files containing the filter passed reads were demultiplexed based on the molecular identifier
619 (MID) sequences using QUASR package version 7.0[51].

620

621 *Read mapping*

622 Trimmomatic (v0.39; Bolger et al. 2014) was used to discard reads with length <30 bases
623 while trimming the ends of reads where base quality scores fall below 20. The MAXINFO
624 option was used to perform adaptive quality trimming, balancing the trade-off between longer
625 read length and tolerance of base calling errors (target length=40, strictness=0.4). For the
626 A/H3N2 virus samples, the trimmed paired reads were merged using FLASH (v1.2.11)[68]. All
627 remaining reads were then locally aligned to A/Brisbane/10/2007 genome (GISAID
628 accession: EPI_ISL_103644) for A/H3N2 virus samples and A/California/4/2009 genome
629 (EPI_ISL_376192) for A/H1N1pdm09 virus samples using Bowtie2 (v2.3.5.1)[69]. Aligned
630 reads with mapping scores falling below 20 alongside bases with quality score (*Q-score*)
631 below 20 were discarded.

632

633 *Variant calling and quality filters*

634 Minority variants of each nucleotide site with a frequency of at least 2% were called if the
635 nucleotide position was covered at least 50x (H1N1pdm09) or 100x (H3N2) and the
636 probability that the variant was called as a result of base calling errors (p_{Err}) was less than
637 1%. p_{Err} was modelled by binomial trials[70]:

638 $$p_{Err} = \sum_{i=n}^{N} \binom{N}{i} p_e^i (1 - p_e)^{N-i}$$

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

641

642 While lower coverage at both ends of individual gene segments was expected, there were also
643 variable coverage results across gene segments for some samples that were mapped to
644 A/H3N2 virus (Figure S6). In order to retain as many samples deemed to have adequate
645 coverage across whole genome, a list of polymorphic nucleotide sites found to have >2%
646 minority variants in more than 1 sample was compiled. Each gene segment of a sample was
647 determined to achieve satisfactory coverage if >70% of these polymorphic sites were covered
648 at least 100x. For A/H1N1pdm09, the gene segment of a sample was deemed to be
649 adequately covered if 80% of the gene was covered at least 50x.

650

651 The number of iSNVs observed in A/H3N2 virus samples collected from subject 1673 (39-94
652 iSNVs in three samples collected from three (D3) to five (D5) days post-symptom onset) and
653 the D8 sample for subject 1878 (73 iSNVs) were substantially greater than numbers in all
654 other samples. The putative majority and minority segment-concatenated sequences of these
655 samples did not cluster as a monophyletic clade among themselves phylogenetically (Figure
656 S9), suggesting that these samples might be the product of mixed infections or cross-
657 contamination. These samples were consequently excluded from further analyses.

658

659 *Empirical within-host evolutionary rate*

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

662 $$r_{g,t} = \frac{\sum_i^{n_{g,t}} f_{g,t,i}}{n_{g,t} \cdot t}$$

663 where $f_{g,t,i}$ is the frequency of minority variants present in nucleotide site i for gene segment
664 g and $n_{g,t}$ is the number of all available sites[26]. Distinct rates were calculated for
665 synonymous and non-synonymous iSNVs. The corresponding whole-genome evolutionary
666 rate (r_t) on day t is computed by summing the rates across all gene segments:

667 $$r_t = \sum_g r_{g,t}$$

668

669 *Within-host simulations*

670 We implemented forward-time Monte Carlo simulations with varying population size using a
671 simplified within-host evolution model to test if our hypotheses could explain the different
672 evolutionary dynamics observed between A/H3N2 and A/H1N1 viral populations. We
673 assumed that a single virion leads to a productive influenza virus infection within an
674 individual and computed changes in the virus population size (N) using a target cell-limited
675 model. New virions are produced upon infection by existing virions at a rate of βCN where C
676 is the existing number of target cells while β is the rate of per-cell per-virion infectious
677 contact. Upon infection, a cell will produce r number of virions before it is rendered
678 unproductive. We assume that infected individuals did not mount any antibody-mediated
679 immune response, setting the virus' natural per-capita decay rate (d) such that virions
680 continue to be present within host for 14 days (Figure S11 and Table 2). β is then computed
681 by fixing the within-host basic reproduction number (R_0):

682 $$R_0 = \frac{\beta C_0 r}{d}$$

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

686 $$\frac{dC}{dt} = -\beta CN$$

687 $$\frac{dN}{dt} = \beta CN - dN$$

688

689 We assume a binary genome of length L, distinguishing between synonymous and
690 nonsynonymous loci. For A/H3N2 viruses, we hypothesised that most *de novo* mutations are
691 either weakly deleterious or neutral. To estimate the number of such sites, we aligned
692 A/H3N2 virus sequences that were collected between 2007 and 2012 and identified all
693 polymorphic sites with variants that did not fixate over time (i.e. <95% frequency over one-
694 month intervals). We estimated $L = 1050$ with 838 and 212 synonymous and
695 nonsynonymous loci respectively. On the other hand, for A/H1N1pdm09 viruses, we
696 assumed that any variants that emerged are likely neutral or weakly beneficial. In the absence
697 of strong purifying selection, ~75% of mutations are expected to be nonsynonymous[26]. Here,
698 we assumed $L = 1000$ sites of which 750 of them are synonymous and the rest are
699 nonsynonymous.

700

701 We tracked the frequency distribution of genotypes present for every generation t. We
702 assumed that mutations occur at per-locus, per-generation rate μ. During each generation t,
703 the number of virions incurring a single-locus mutation followed a Poisson distribution with
704 mean $N(t)\mu L$. For each virion, the mutant locus was randomly selected across all loci. We

705 assumed that all synonymous and a fraction of nonsynonymous sites ($f_{neu,NS}$) are neutral (i.e.
706 (log) fitness effect $s = 0$). The remaining nonsynonymous sites either had an additive
707 deleterious (s_d) or beneficial (s_b) fitness effect when mutated. The magnitude of s_d/s_b follow
708 an exponential distribution with mean effect $|\bar{s}|$. Epistasis was neglected throughout. The
709 distribution of genotypes in the next generation $t + 1$ was achieved by resampling
710 individuals according to Poisson distribution with mean $N(t + 1)P_f(g, t)$ where $P_f(g, t)$ is
711 the relative fitness distribution of genotype g during generation t.

712

713 To decrease the computational costs of the simulations, specifically when $N(t)$ reaches
714 orders of $10^{10} - 10^{11}$ virions (Figure S11), we implemented an upper population size limit
715 of 10^7 virions. Given the mutation rate assumed (Table 2), $N(t)\mu \gg 1$ for $N(t) \geq 10^7$
716 virions, mutation-selection balance is theoretically expected for a single-locus (deleterious)
717 mutant model (see Supplementary Materials). We ran 500 simulations for each variable set of
718 $f_{neu,NS}$ and s_d/s_b values. All parameter values used in the model are given in Table 2.

719

720 **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	[71]
C_0	Initial target cell population size	4×10^8 virions	[72]
d	Per-capita decay rate	2 per-generation	Assumption
R_0	Within-host basic reproduction number	5	[72]
μ	Per-site, per-generation mutation rate	3×10^{-5} per-site, per-generation	[45]

721

722 *Haplotype reconstruction*

723 The most parsimonious viral haplotypes of each gene segment were reconstructed by fitting
724 the observed nucleotide variant count data to a Dirichlet multinomial model using a
725 previously developed maximum likelihood approach to infer haplotype frequencies[44].
726 Assuming that the viral population is made up of a set of K haplotypes with frequencies $\boldsymbol{q_k}$,
727 the observed partial haplotype frequencies $\boldsymbol{q_l}$ for a polymorphic site l can be computed by
728 multiplying a projection matrix $\boldsymbol{T_l}$. For instance, if the set of hypothetical full haplotypes is
729 assumed to be $\{AA, GA, AG\}$, the observed partial haplotype frequencies for site $l = 1$, q_{A-}
730 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

765 haploptes) are (A, A, C), (T, A, C) and (A, T, C) (i.e. we would compute $f_{ms,(A,A,C)}$,
766 $f_{ms,(T,A,C)}$ and $f_{ms,(A,T,C)}$ respectively). The observed number of reads with (–, A, C) will
767 counted towards the computation of both $f_{ms,(A,A,C)}$ and $f_{ms,(T,A,C)}$ since they could be
768 attributed to either haplotype. Similarly, reads with (–, T, C) will be absorbed towards the
769 counts to compute $f_{ms,(A,T,C)}$. Finally, we see that reads with (A, C, –) are not a subset of any
770 of the top hierarchal haplotypes considered. As such, they form the 4th possible top hierarchal
771 haplotype on its own. As such, if we were to compute the ranking for haplotype (A, A, C):

772 $$s_{(A,A,C)} = f_{ss,(A,A,C)} \times f_{ms,(A,A,C)}$$
773 $$= \{f_{(A,-,-)} \times f_{(-,A,-)} \times f_{(-,-,C)}\} \times f_{ms,(A,A,C)}$$

774

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

780

781 *Linkage disequilibrium*

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

785 $$LD_{ij} = \hat{q}_{ij} - \hat{q}_i\hat{q}_j$$

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

789 $$LD' = \frac{LD}{LD_{max}}$$

790 $$LD_{max} = \begin{cases} \max\{-\hat{q}_i\hat{q}_j, -(1-\hat{q}_i)(1-\hat{q}_j)\} & if\ LD > 0 \\ \min\{\hat{q}_i(1-\hat{q}_j), (1-\hat{q}_i)\hat{q}_j\} & if\ LD < 0 \end{cases}$$

791

792

793 *FoldX analyses*

794 FoldX (https://foldxsuite.crg.eu/) was used to estimate structural stability effects of likely
795 linked nonsynonymous minority variants found in the nucleoprotein (NP) of within-host
796 A/H3N2 virus populations. At the time of writing of this paper, there was no A/H3N2-NP

797 structure available. Although the eventual NP structure (PDB: 3ZDP) adopted for stability
798 analyses was originally derived from H1N1 virus (A/WSN/33)[40], it was the most well
799 resolved (2.69Å) crystal structure available, with 78.5% amino acid identity relative to the NP
800 protein of A/Brisbane/10/2007. Previous work has shown that mutational effects predicted by
801 FoldX using a NP structure belonging to A/WSN/33 (H1N1) was similar to those
802 experimentally determined on a A/Brisbane/10/2007 nucleoprotein[31]. FoldX first removed
803 any potential steric clashes to repair the NP structure. It then estimated differences in free
804 energy changes as a result of the input amino acid mutation (i.e. $\Delta\Delta G = \Delta G_{mutant} -$
805 $\Delta G_{wild-type}$) under default settings (298K, 0.05M ionic strength and pH 7.0). Five distinct
806 simulations were made to estimate the mean and standard deviation $\Delta\Delta G$ values.

807

808 *Phylogenetic inference*

809 All maximum likelihood phylogenetic trees were reconstructed with IQTREE (v. 1.6.10)[73],
810 using the GTR+I+G4 nucleotide substitution model.

811

812 **References**

813 1. Andino, R. & Domingo, E. Viral quasispecies. *Virology* **479–480**, 46–51 (2015).
814 2. Smith, D. J. *et al.* Mapping the Antigenic and Genetic Evolution of Influenza Virus.
815 *Science (80-.).* **305**, 371–376 (2004).
816 3. Varble, A. *et al.* Influenza A Virus Transmission Bottlenecks Are Defined by Infection
817 Route and Recipient Host. *Cell Host Microbe* **16**, 691–700 (2014).
818 4. McCrone, J. T. *et al.* Stochastic processes constrain the within and between host
819 evolution of influenza virus. *Elife* **7**, e35962 (2018).
820 5. Russell, C. A. *et al.* The global circulation of seasonal influenza A (H3N2) viruses.
821 *Science* **320**, 340–6 (2008).
822 6. Rambaut, A. *et al.* The genomic and epidemiological dynamics of human influenza A
823 virus. *Nature* **453**, 615–619 (2008).
824 7. Nelson, M. I. & Holmes, E. C. The evolution of epidemic influenza. *Nat. Rev. Genet.*
825 **8**, 196–205 (2007).
826 8. Dinis, J. M. *et al.* Deep Sequencing Reveals Potential Antigenic Variants at Low
827 Frequencies in Influenza A Virus-Infected Humans. *J. Virol.* **90**, 3355–65 (2016).
828 9. Debbink, K. *et al.* Vaccination has minimal impact on the intrahost diversity of H3N2
829 influenza viruses. *PLOS Pathog.* **13**, e1006194 (2017).
830 10. Valesano, A. L. *et al.* Influenza B Viruses Exhibit Lower Within-Host Diversity than
831 Influenza A Viruses in Human Hosts. *J. Virol.* **94**, e01710-19 (2020).
832 11. Sobel Leonard, A. *et al.* Deep Sequencing of Influenza A Virus from a Human
833 Challenge Study Reveals a Selective Bottleneck and Only Limited Intrahost Genetic
834 Diversification. *J. Virol.* **90**, 11247–11258 (2016).
835 12. Han, A. X., Maurer-Stroh, S. & Russell, C. A. Individual immune selection pressure
836 has limited impact on seasonal influenza virus evolution. *Nat. Ecol. Evol.* 1 (2018).
837 doi:10.1038/s41559-018-0741-x
838 13. Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin
839 H1N1 influenza A epidemic. *Nature* **459**, 1122–1125 (2009).
840 14. Su, Y. C. F. *et al.* Phylodynamics of H1N1/2009 influenza reveals the transition from
841 host adaptation to immune-driven selection. *Nat. Commun.* **6**, 7952 (2015).
842 15. Elderfield, R. A. *et al.* Accumulation of Human-Adapting Mutations during
843 Circulation of A(H1N1)pdm09 Influenza Virus in Humans in the United Kingdom. *J.*
844 *Virol.* **88**, 13269 LP – 13283 (2014).
845 16. Nogales, A., Martinez-Sobrido, L., Chiem, K., Topham, D. J. & DeDiego, M. L.
846 Functional Evolution of the 2009 Pandemic H1N1 Influenza Virus NS1 and PA in
847 Humans. *J. Virol.* **92**, e01206-18 (2018).
848 17. Poon, L. L. M. *et al.* Quantifying influenza virus diversity and transmission in humans.
849 *Nat. Genet.* **48**, 195–200 (2016).
850 18. Xue, K. S. & Bloom, J. D. Reconciling disparate estimates of viral genetic diversity
851 during human influenza infections. *Nat. Genet.* **51**, 1298–1301 (2019).
852 19. Poon, L. L. M. *et al.* Reply to 'Reconciling disparate estimates of viral genetic
853 diversity during human influenza infections'. *Nat. Genet.* **51**, 1301–1303 (2019).
854 20. South East Asia Infectious Disease Clinical Research Network. Effect of double dose
855 oseltamivir on clinical and virological outcomes in children and adults admitted to
856 hospital with severe influenza: Double blind randomised controlled trial. *BMJ* **346**,
857 f3039 (2013).
858 21. Koel, B. F. *et al.* Longitudinal sampling is required to maximize detection of intrahost
859 A/H3N2 virus variants. *Virus Evol.* **6**, veaa088 (2020).
860 22. Worby, C. J. *et al.* On the relative role of different age groups in influenza epidemics.

861 *Epidemics* **13**, 10–16 (2015).
862 23. Viboud, C. *et al.* Risk factors of influenza transmission in households. *Int. Congr. Ser.*
863 **1263**, 291–294 (2004).
864 24. Horby, P. *et al.* The epidemiology of interpandemic and pandemic influenza in
865 Vietnam, 2007-2010. *Am. J. Epidemiol.* **175**, 1062–1074 (2012).
866 25. Hien, T. T. *et al.* Early Pandemic Influenza (2009 H1N1) in Ho Chi Minh City,
867 Vietnam: A Clinical Virological and Epidemiological Analysis. *PLOS Med.* **7**,
868 e1000277 (2010).
869 26. Xue, K. S. & Bloom, J. D. Linking influenza virus evolution within and between
870 human hosts. *Virus Evol.* **6**, 812016 (2020).
871 27. Wiley, D. C. C., Wilson, I. A. A. & Skehel, J. J. J. Structural identification of the
872 antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement
873 in antigenic variation. *Nature* **289**, 373–378 (1981).
874 28. Caton, A. J., Brownlee, G. G., Yewdell, J. W. & Gerhard, W. The antigenic structure
875 of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell* **31**, 417–27 (1982).
876 29. Igarashi, M. *et al.* Predicting the Antigenic Structure of the Pandemic (H1N1) 2009
877 Influenza Virus Hemagglutinin. *PLoS One* **5**, e8553 (2010).
878 30. Mänz, B. *et al.* Pandemic Influenza A Viruses Escape from Restriction by Human
879 MxA through Adaptive Mutations in the Nucleoprotein. *PLOS Pathog.* **9**, e1003279
880 (2013).
881 31. Ashenberg, O., Gong, L. I. & Bloom, J. D. Mutational effects on stability are largely
882 conserved during protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 21071–6
883 (2013).
884 32. Berkhoff, E. G. M. *et al.* Functional Constraints of Influenza A Virus Epitopes Limit
885 Escape from Cytotoxic T Lymphocytes. *J. Virol.* **79**, 11239 LP – 11246 (2005).
886 33. Gog, J. R., Rimmelzwaan, G. F., Osterhaus, A. D. M. E. & Grenfell, B. T. Population
887 dynamics of rapid fixation in cytotoxic T lymphocyte escape mutants of influenza A.
888 *Proc. Natl. Acad. Sci.* **100**, 11143 LP – 11147 (2003).
889 34. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the
890 evolution of an influenza protein. *Elife* **2**, e00631 (2013).
891 35. Rimmelzwaan, G. F., Berkhoff, E. G. M., Nieuwkoop, N. J., Fouchier, R. A. M. &
892 Osterhaus, A. D. M. E. Functional Compensation of a Detrimental Amino Acid
893 Substitution in a Cytotoxic-T-Lymphocyte Epitope of Influenza A Viruses by
894 Comutations. *J. Virol.* **78**, 8946 LP – 8949 (2004).
895 36. Berkhoff, E. G. M. *et al.* A Mutation in the HLA-B*2705-Restricted NP383-391
896 Epitope Affects the Human Influenza A Virus-Specific Cytotoxic T-Lymphocyte
897 Response In Vitro. *J. Virol.* **78**, 5216 LP – 5222 (2004).
898 37. Resa-Infante, P. *et al.* Targeting Importin-α7 as a Therapeutic Approach against
899 Pandemic Influenza Viruses. *J. Virol.* **89**, 9010 LP – 9020 (2015).
900 38. Resa-Infante, P. *et al.* Alternative interaction sites in the influenza A virus
901 nucleoprotein mediate viral escape from the importin-α7 mediated nuclear import
902 pathway. *FEBS J.* **286**, 3374–3388 (2019).
903 39. Gabriel, G. *et al.* Differential use of importin-α isoforms governs cell tropism and host
904 adaptation of influenza virus. *Nat. Commun.* **2**, 156 (2011).
905 40. Chenavas, S. *et al.* Monomeric Nucleoprotein of Influenza A Virus. *PLOS Pathog.* **9**,
906 e1003275 (2013).
907 41. Mak, G. C. *et al.* Association of D222G substitution in haemagglutinin of 2009
908 pandemic influenza A (H1N1) with severe disease. *Eurosurveillance* **15**, (2010).
909 42. Mai, L. Q. *et al.* A Community Cluster of Oseltamivir-Resistant Cases of 2009 H1N1
910 Influenza. *N. Engl. J. Med.* **362**, 86–87 (2010).

911 43. Lam, J. H. & Baumgarth, N. The Multifaceted B Cell Response to Influenza Virus. *J.*
912 *Immunol.* **202**, 351 LP – 359 (2019).
913 44. Ghafari, M., Lumby, C. K., Weissman, D. B. & Illingworth, C. J. R. Inferring
914 Transmission Bottleneck Size from Viral Sequence Data Using a Novel Haplotype
915 Reconstruction Method. *J. Virol.* **94**, (2020).
916 45. McCrone, J. T., Woods, R. J., Monto, A. S., Martin, E. T. & Lauring, A. S. The
917 effective population size and mutation rate of influenza A virus in acutely infected
918 individuals. *bioRxiv* 2020.10.24.353748 (2020). doi:10.1101/2020.10.24.353748
919 46. Morris, D. H. *et al.* Asynchrony between virus diversity and antibody selection limits
920 influenza virus evolution. *Elife* **9**, 1–62 (2020).
921 47. Xue, K. S. *et al.* Parallel evolution of influenza across multiple spatiotemporal scales.
922 *Elife* **6**, e26875 (2017).
923 48. Lumby, C. K., Zhao, L., Breuer, J. & Illingworth, C. J. R. A large effective population
924 size for established within-host influenza virus infection. *Elife* **9**, e56915 (2020).
925 49. Petrova, V. N. & Russell, C. A. The evolution of seasonal influenza viruses. *Nat. Rev.*
926 *Microbiol.* **16**, 47–60 (2017).
927 50. Ng, S. *et al.* The Timeline of Influenza Virus Shedding in Children and Adults in a
928 Household Transmission Study of Influenza in Managua, Nicaragua. *Pediatr. Infect.*
929 *Dis. J.* **35**, 583–586 (2016).
930 51. Watson, S. J. *et al.* Viral population analysis and minority-variant detection using short
931 read next-generation sequencing. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368**,
932 20120205 (2013).
933 52. Welkers, M. R. A. *et al.* Genetic diversity and host adaptation of avian H5N1 influenza
934 viruses during human infection. *Emerg. Microbes Infect.* **8**, 262–271 (2019).
935 53. Thai, P. Q. *et al.* Pandemic H1N1 virus transmission and shedding dynamics in index
936 case households of a prospective Vietnamese cohort. *J. Infect.* **68**, 581–590 (2014).
937 54. Suess, T. *et al.* Comparison of Shedding Characteristics of Seasonal Influenza Virus
938 (Sub)Types and Influenza A(H1N1)pdm09; Germany, 2007–2011. *PLoS One* **7**,
939 e51653 (2012).
940 55. Jacobs, N. T. *et al.* Incomplete influenza A virus genomes occur frequently but are
941 readily complemented during localized viral spread. *Nat. Commun.* **10**, 3526 (2019).
942 56. McCown, M. F. & Pekosz, A. The Influenza A Virus M₂
943 Cytoplasmic Tail Is Required for Infectious Virus Production and Efficient Genome
944 Packaging. *J. Virol.* **79**, 3595 LP – 3605 (2005).
945 57. Zhao, L. & Illingworth, C. J. R. Measurements of intrahost viral diversity require an
946 unbiased diversity metric. *Virus Evol.* **5**, (2019).
947 58. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of
948 restriction endonucleases. *Proc. Natl. Acad. Sci.* **76**, 5269 LP – 5273 (1979).
949 59. Illingworth, C. J. R. *et al.* On the effective depth of viral sequence data. *Virus Evol.* **3**,
950 (2017).
951 60. Xue, K. S., Moncla, L. H., Bedford, T. & Bloom, J. D. Within-Host Evolution of
952 Human Influenza Virus. *Trends Microbiol.* **26**, 781–793 (2018).
953 61. Hoffmann, E., Stech, J., Guan, Y., Webster, R. G. & Perez, D. R. Universal primer set
954 for the full-length amplification of all influenza A viruses. *Arch. Virol.* **146**, 2275–
955 2289 (2001).
956 62. Lee, N. *et al.* Viral Loads and Duration of Viral Shedding in Adult Patients
957 Hospitalized with Influenza. *J. Infect. Dis.* **200**, 492–500 (2009).
958 63. Ling, L. M. *et al.* Effects of early oseltamivir therapy on viral shedding in 2009
959 pandemic influenza A (H1N1) virus infection. *Clin Infect Dis* **50**, 963–969 (2010).
960 64. Renzette, N. *et al.* Evolution of the Influenza A Virus Genome during Development of

Oseltamivir Resistance In Vitro *J. Virol.* **88**, 272 LP – 281 (2014).
65. Zhou, B. *et al.* Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza a viruses. *J. Virol.* **83**, 10309–13 (2009).
66. 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. *J. Virol.* **88**, 1694–702 (2014).
67. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
68. Magoč, T., Magoč, M. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. **27**, 2957–2963 (2011).
69. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
70. Illingworth, C. J. R. SAMFIRE: multi-locus variant calling for time-resolved sequence data. *Bioinformatics* **32**, 2208–2209 (2016).
71. Frensing, T. *et al.* Influenza virus intracellular replication dynamics, release kinetics, and particle morphology during propagation in MDCK cells. *Appl. Microbiol. Biotechnol.* **100**, 7181–7192 (2016).
72. Hadjichrysanthou, C. *et al.* Understanding the within-host dynamics of influenza A virus: from theory to clinical implications. *J. R. Soc. Interface* **13**, 20160289 (2016).
73. 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. *Mol. Biol. Evol.* **32**, 268–74 (2015).

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 (Table S5) 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.

Author contributions

A.X.H., Z.C.F.G., M.R.A.W., D.E., M.D.d.J., and C.A.R. designed the research; A.X.H., Z.C.F.G. and M.R.A.W. performed the data analyses; M.R.A.W., R.M.V., T.N.D, L.T.Q.M., P.Q.T., T.T.N.A., H.M.T., N.T.H., L.Q.T., L.T.H., H.T.B.N., K.C., P.P., N.V.V.C., N.M.N., D.D.T., T.T.H., H.F.L.W., P.H., A.F., H.R.V.D., D.E. and M.D.d.J. collected the clinical samples and generated the sequencing data; A.X.H., Z.C.F.G. and C.A.R. wrote the first draft of the paper. All authors contributed to the critical review and revision of the paper.