Virus Evolution, 2024, **10(1)**, veae034
DOI: https://doi.org/10.1093/ve/veae034
Advance Access Publication 25 April 2024
Research Article

Within-host influenza viral diversity in the pediatric population as a function of age, vaccine, and health status

Ashley Sobel Leonard,[1,2,*] Lydia Mendoza,[2] Alexander G. McFarland,[2] Andrew D. Marques,[2] John K. Everett,[2] Louise Moncla,[3] Frederic D. Bushman,[2] Audrey R. Odom John,[1,2,3,4] and Scott E. Hensley[2,*]

[1]Division of Infectious Diseases, Children's Hospital of Philadelphia, 3401 Civic Center Blvd., Philadelphia, PA 19104, USA, [2]Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, 3400 Civic Center Blvd., Philadelphia, PA 19104, USA, [3]Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, 3800 Spruce St., Philadelphia, PA 19104, USA and [4]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, 3400 Civic Center Blvd., Philadelphia, PA 19104, USA

*Corresponding author: E-mail: sobelleona@chop.edu; hensley@pennmedicine.upenn.edu

Abstract

Seasonal influenza virus predominantly evolves through antigenic drift, marked by the accumulation of mutations at antigenic sites. Because of antigenic drift, influenza vaccines are frequently updated, though their efficacy may still be limited due to strain mismatches. Despite the high levels of viral diversity observed across populations, most human studies reveal limited intrahost diversity, leaving the origin of population-level viral diversity unclear. Previous studies show host characteristics, such as immunity, might affect within-host viral evolution. Here we investigate influenza A viral diversity in children aged between 6 months and 18 years. Influenza virus evolution in children is less well characterized than in adults, yet may be associated with higher levels of viral diversity given the lower level of pre-existing immunity and longer durations of infection in children. We obtained influenza isolates from banked influenza A-positive nasopharyngeal swabs collected at the Children's Hospital of Philadelphia during the 2017–18 influenza season. Using next-generation sequencing, we evaluated the population of influenza viruses present in each sample. We characterized within-host viral diversity using the number and frequency of intrahost single-nucleotide variants (iSNVs) detected in each sample. We related viral diversity to clinical metadata, including subjects' age, vaccination status, and comorbid conditions, as well as sample metadata such as virus strain and cycle threshold. Consistent with previous studies, most samples contained low levels of diversity with no clear association between the subjects' age, vaccine status, or health status. Further, there was no enrichment of iSNVs near known antigenic sites. Taken together, these findings are consistent with previous observations that the majority of intrahost influenza virus infection is characterized by low viral diversity without evidence of diversifying selection.

Keywords: influenza; viral diversity; antigenic drift; next-generation sequencing; pediatrics.

Introduction

In most years, global influenza-associated respiratory mortality accounts for 291,000–645,000 deaths, with 9,200–105,000 deaths occurring in children under 5 years of age (Iuliano et al. 2018). Prior episodes of influenza disease do not protect against re-infection due to antigenic drift, whereby an accumulation of substitutions change the viral surface proteins to allow evasion of the host's antibody response. The influenza vaccine is updated annually to counter the effects of antigenic drift, providing boosted immunity against the influenza strains that are predicted to circulate. While mutations that facilitate antigenic drift are identifiable in retrospect (Li et al. 2013; Chambers et al. 2015), the inability to fully predict which mutations are most likely to emerge each season limits vaccine efficacy (D'Mello et al. 2015).

There has been increasing interest in characterizing intrahost diversity of rapidly evolving RNA viruses, including SARS-CoV-2, HIV, and influenza. The rise in interest has been driven by technological advancements and recognition of the link between the within-host and between-host scales (Xue et al. 2018). The development of deep-sequencing technologies facilitates quantifying intrahost viral population diversity beyond just the consensus-level genetic sequence. Yet, despite the advancements in next-generation sequencing (NGS), the identification of intrahost single-nucleotide variants (iSNVs) remains technically difficult due to the high background noise and sequencing artifacts. Studies of within-host viral evolution have demonstrated that variants identified during prolonged infections may be predecessors to variants later observed in the general

Table 1. Clinical metadata for subjects categorized by age. Clinical metadata categories are as follows. Cat: age range contained in category, N: number of subjects, age, cycle threshold (C_t), and symptoms are shown with mean ± standard deviation. Symptoms describe the # of days of symptoms prior to sample collection. Vaccine, PMCA, and comorbidities are shown with number of subjects (% of subjects in age category). PMCA: pediatric medical complexity algorithm, N-CD: no chronic disease, NC-CD: non-complex chronic disease, C-CD: complex chronic disease. CLD: chronic lung disease. IC: Immunocompromised.

Cat	N	Age	Ct	Symptoms	Vaccine		PMCA			Comorbidities	
					No	Yes	N-CD	NC-CD	C-CD	CLD	IC
0–4	59	1.6 ± 1.0	22.3 ± 3.4	3.2 ± 2.7	23 (39)	36 (61)	28 (47)	21 (36)	10 (17)	16 (27)	2 (3)
5–11	22	8.7 ± 2.0	21.9 ± 2.8	3.0 ± 1.8	8 (36)	14 (64)	2 (9)	9 (41)	11 (50)	16 (73)	1 (5)
12–18	30	15.8 ± 1.8	24.6 ± 3.3	2.8 ± 1.2	8 (27)	22 (73)	2 (7)	13 (43)	15 (50)	21 (70)	3 (10)

population (Xue et al. 2018). Moreover, prolonged infections may give rise to highly divergent viral lineages (Choi et al. 2020; Weigang et al. 2021; Ko et al. 2022; Gonzalez-Reiche et al. 2023), which in some cases may facilitate strain replacement, speculated to be the origin of the omicron lineage of SARS-CoV-2 (Shrestha et al. 2022).

While there are many examples of how prolonged viral infections give rise to divergent lineages, recent studies have shown that within-host viral populations generally show low-viral diversity (Debbink et al. 2017; McCrone et al. 2018; Moncla et al. 2020; Valesano et al. 2020). In the case of influenza viruses, cohort studies have shown that the within-host viral populations contain low numbers, <10 iSNVs, per sample (Debbink et al. 2017; Valesano et al. 2020). The paucity of within-host diversity is unsurprising given that most influenza infections are subject to narrow transmission bottlenecks and short infectious periods (Xue et al. 2018). Transmission bottlenecks describe the number of unique viral genomes that give rise to new infections and serve as essential determinants of within-host viral diversity for acute infections. Narrow transmission bottlenecks of 1–2 virions severely limit the starting viral diversity of the first infection (McCrone et al. 2018). Short-lived infections with few rounds of viral replication provide limited opportunity for viruses to accumulate sufficient mutations to substantially increase within-host diversity (Xue et al. 2018). Moreover, intrahost influenza virus evolution during acute infections is dominated by purifying selection to remove deleterious mutations. In contrast, diversifying or positive selection, such as antibody-mediated selection of antigenic variants, is rarely observed.

Observations suggest that a small subset of infections may be driving viral evolution at the population level (Lumby et al. 2020), akin to how rare individuals have an outsized effect on viral transmission, as seen with superspreaders (Lloyd-Smith et al. 2005). Identifying the proportion of the population with higher levels of intrahost influenza virus diversity could improve viral surveillance efforts and, thus, strain selection for the seasonal influenza vaccine. While high levels of viral diversity have been identified in severely immunocompromised individuals, these comprise a very small proportion of the general population and it is unclear if they are the primary source of pathogenic variants (Eden et al. 2017).

Some infection characteristics giving rise to prolonged, high-diversity infections in the immunocompromised can also be observed in children. Children shed virus for a longer duration than adults (Ng et al. 2016), which is likely related to limited prior immunity. The longer duration of viral replication and shedding, in turn, could potentially provide a greater opportunity for the accumulation of intrahost mutations. Delayed antibody induction can potentially provide a source of diversifying selection that gradually increases over the course of infection. Furthermore, children are already recognized as a key driver of influenza virus transmission (Worby et al. 2015), which could facilitate the transmission of antigenic variants. A recent study of influenza virus evolution of children in Vietnam has assessed within-host viral evolution in this population with longitudinal sampling (Han et al. 2021), and found that non-synonymous mutations tended to increase in frequency over the course of infection. However, given the key role of children in influenza virus propagation, further studies are needed of pediatric populations.

In this study, we evaluated 111 clinical influenza virus isolates from children collected during the 2017–18 influenza season using NGS. Our results show that the majority of these samples contain low numbers of iSNVs, suggesting low intrahost diversity. While we identified two clinical isolates with significantly higher levels of diversity, further analysis of the variants identified in many of the identified iSNVs were in phase each other. This observation suggests that the high diversity was not attributable to *de novo* evolution and, as such, they were excluded from subsequent analyses. For the remaining subjects, we found no association between the intrahost viral diversity and the age, vaccine status, or health status of the corresponding subject.

Results

Study participants

The viral isolates analyzed in this study were obtained from residual influenza virus-positive diagnostic nasopharyngeal swabs banked by the Children's Hospital of Philadelphia (CHOP) Infectious Diseases Diagnostics Laboratory (IDDL). We included samples from children between 6 months and 18 years of age. All banked samples from the 2017 to 2018 flu season were first stratified into distinct subgroups based on age and vaccination status. Vaccination status was defined relative to receipt of the 2017–18 seasonal flu vaccine only. From these defined subgroups, a directed random sampling strategy was implemented, resulting in the selection of 197 samples for further analysis. Of the identified samples, 118 met our quality control criteria. Further clinical and demographic metadata for these samples were obtained through medical record review. Previous studies of influenza and other viruses have shown that days since symptom onset and the amount of viral genetic material in a sample, often quantified using cycle threshold (C_t), may influence diversity metrics (Valesano et al. 2020; Han et al. 2021; Voloch 2021). The samples within our age-based categories did not differ based on the days since symptom onset or C_t (Table 1). We assessed the general health status of the subjects using the Pediatric Medical Complexity Algorithm, a tool used to stratify children based their level of medical need due to chronic health conditions (Simon et al. 2014). Children are stratified into three groups: children without chronic disease (N-CD), children with non-complex chronic disease (NC-CD), and

Table 2. Comparative distribution of iSNVs across influenza gene segments. The table delineates the iSNV counts across influenza gene segments. Synonymous (S) and non-synonymous (NS) mutations for each segment under different categories.

Segment		H3N2		H1N1	
#	Name	S	NS	S	NS
1	PB2	9	4	5	5
2	PB1	9	4	2	2
3	PA	8	9	2	2
4	HA	9	4	1	2
5	NP	9	5	5	1
6	NA	5	7	0	5
7	MP	1	2	0	1
8	NS	4	1	1	5
	Avg	6.75	4.5	2	2.9

children with complex chronic disease (C-CD), based on diagnoses codes from the International Classification of Diseases coding system. Complex chronic conditions are associated with higher utilization of healthcare resources and poor health outcomes (Berry et al. 2015). Our study population had a higher proportion of children with complex chronic disease, particularly in the 5–11 and 12- to 18-year-old age groups, which can be explained by CHOP's role as a quaternary referral center. The most frequently identified medical comorbidity was chronic lung disease, which included asthma and chronic lung disease of prematurity, known risk factors for severe influenza (Coffin et al. 2007), which was identified in 48 per cent of our subjects overall. We included children with immunocompromising conditions, though they comprised a small proportion of the study participants in each age category.

Sample quality control

The samples were sequenced using the Illumina NextSeq platform. We applied rigorous quality control (QC) methods to ensure the robustness of our findings since sample contamination, technical errors, and sequencing artifacts can significantly affect the results of within-host viral diversity studies (McCrone, Lauring, and Dermody 2016; Xue and Bloom 2019; Roder et al. 2023). Previous studies have shown that low-starting cDNA concentrations can impact the accuracy of variant identification, thus we only considered samples with a $C_t < 30$ (McCrone, Lauring, and Dermody 2016). All samples were sequenced in duplicate from the beginning of our workflow at the viral RNA extraction step. We trimmed all short reads to the middle 50 per cent of the read prior to alignment with A/Michigan/45/2015 (H1N1) and A/Washington/17/2016 (H3N2) reference sequences. We masked nucleotides within short reads with phred scores <30 and/or mapping quality scores < 40. Within the alignment, we masked nucleotides with coverage of <100 reads. To pass QC, we required the sample to have at least 100× coverage for ≥95 per cent of the influenza genome. iSNV detection is highly sensitive to variant calling thresholds. We chose an iSNV threshold of 3 per cent to minimize the identification of artifactual iSNVs, a threshold that a prior study using amplicon-based sequencing validated as effective at removing false-positive iSNVs when combined with replicate sequencing (Grubaugh et al. 2019). Additional studies evaluating within-host viral diversity have also employed a minimum frequency threshold of 3 per cent (Braun et al. 2021; Lythgoe et al. 2021).

Even when stored at −80°C, viral RNA in clinical isolates can degrade over time (Cannon et al. 2019). Given that these samples were at least 5 years old at the time of sequencing, we expected

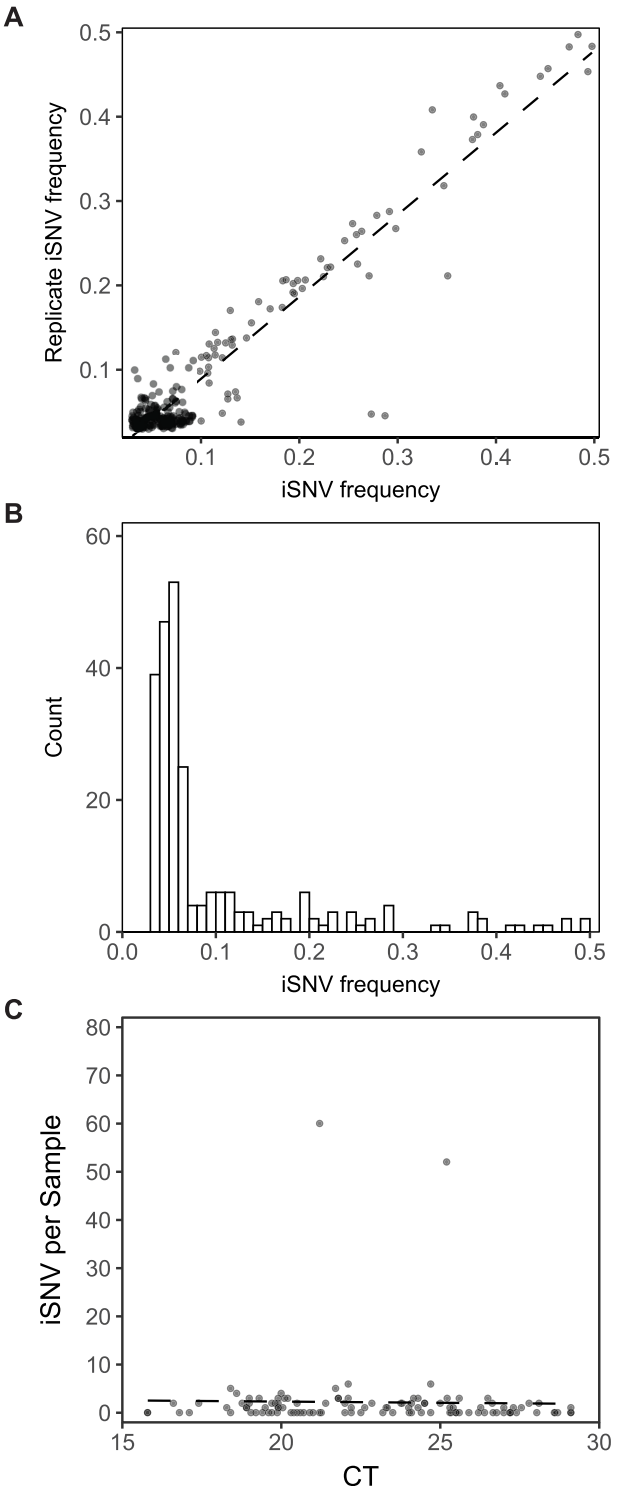

Figure 1. iSNV detection in technical replicates. (A) Concordance of iSNV frequency across between technical replicates. (B) Distribution of intersection iSNV frequency. (C) Association between the number of intersection iSNVs identified in a sample and that sample's C_t.

a subset of samples to yield low-quality sequences. To that end, we excluded all samples where <50 per cent of the detected iSNVs in each replicate were identified as intersection iSNVs, which we define as an iSNV found in both replicates with a frequency of at least 3 per cent. This criterion was applied only if there was more than one iSNV detected. As a result of this QC requirement, twenty

samples were excluded. For those samples passing our QC criteria, there was a strong linear correlation ($R^2 = 0.89$) between iSNV frequency across replicates (Fig. 1A).

The majority of identified iSNVs were generally observed at low frequency with a mean(υ_{var}) $\pm$ SD of 9.97 ± 10.09 per cent, where υ_{var} represents the average iSNV frequency across both replicates (Fig. 1B). While we excluded samples with high C_t values, we still had concerns about the potential lingering impact of a sample's C_t on the number of identified iSNVs. To investigate this possibility, we performed a linear regression analysis, examining the relationship between the samples' C_t values and the number of iSNVs detected in the corresponding sample. Our analysis failed to identify a significant association between C_t and the number of iSNVs ($R^2 = 3.03 \times 10^{-5}$) (Fig. 1C), indicating that, for the samples analyzed in this study, the within-host viral diversity was independent of the amount of viral material in the samples.

Phylogenetics

We reconstructed phylogenetic trees using the consensus sequence for hemagglutinin (HA) protein from each sample using Nextstrain for both influenza A subtypes (Hadfield, Megill, and Bell et al. 2018; Sagulenko, Puller, and Neher 2018). Of the 111 samples that passed our QC requirements, 84 sequences (76 per cent) were identified as H3N2 with the remaining 27 sequences identified as H1N1. This is consistent with the viral surveillance data, which showed that ~85 per cent of subtyped influenza A viruses circulating during the 2017–18 were H3N2 with a regional predominance ranging from 76 percent to 91 per cent (Garten et al. 2018). We aligned the hemagglutinin sequences obtained from our study samples with publicly available hemagglutinin sequences available from GISAID. We selected the sequences from GISAID using randomized, targeted, subsampling of all sequences submitted by the CDC for the United States between 2013 - 2018. We further enriched for sequences collected in Pennsylvania by incorporating additional randomly selected sequences obtained in Pennsylvania betwen 2017 and 2018, the year that the samples analyzed in this study were collected (Khare et al. 2021). Additional details describing the subsampling technique and phylogenetic reconstruction can be found in the 'Methods' section. The tree tips for both H3N2 and H1N1 sequences from our study were interspersed with the tips representing the general sequences from the USA and Pennsylvania (Fig. 2). The H3N2 samples fell within the 3C.2a2 and 3C.3a1 subclades and the H1N1 samples fell within the 6b1.A subclade, clades that were known to be circulating within the Northern Hemisphere during that time (Garten et al. 2018). Based on the categorization of our sequences into clades of the contemporaneous circulating influenza subtypes and the relative ratio of those subtypes in our data, we conclude that sequences from our study are representative of influenza viruses circulating in the USA during 2017–18.

Identification of mixed infections and possibility of cross-contamination of samples with high iSNVs

Mixed infections can lead to the detection of multiple mutations within a sample (Ghedin et al. 2011), which may be a potential explanation for higher numbers of iSNVs in some samples (McCrone et al. 2018). The iSNVs identified in CHOP-101 and CHOP-117 were predominately found on the HA segment. During inspection of the fastq files from the HA segment in these samples, we observed that many of the variants were in phase with one another. Furthermore, they occurred at residues that differed between the 3C.2a2, the clade of those samples' consensus sequences, and the 3C.3a1 subclade co-circulating at the time of sample collection (Supplementary Fig. S1). Thus, it is likely that these isolates represent mixed infections. We also cannot rule out the possibility of cross-contamination with other samples from our study as a subset of the other influenza isolates sequenced belonged to the 3C.3a1 clade. As the high diversity in these samples was not likely attributable to *de novo* evolution, we excluded these samples from subsequent analysis.

Variant analysis

We compared within-host diversity across the samples based on the number of identified synonymous and nonsynonymous iSNVs. We found low levels of within-host viral diversity with a mean and standard deviation of 0.64 ± 0.91 synonymous and 0.54 ± 0.84 nonsynonymous intersection iSNVs per sample (Supplementary Fig. S2A), consistent with the low within-host diversity described in previous studies (McCrone, Lauring, and Dermody 2016; Debbink et al. 2017; Han et al. 2021). All iSNVs identified in our study are shown in Supplementary Table S2. The distribution of the number of iSNVs per sample did not follow a normal distribution; instead, it exhibited a right-sided tail, indicative of overdispersion. To assess the association between the subtype and the number of iSNVs, we fitted a negative binomial regression model and compared the groups using a chi-squared (χ^2) test, applying a Bonferroni correction to account for multiple comparisons. We observed no association between subtype and the number of synonymous or non-synonymous iSNVs, $P = 1.00$ and 0.07, respectively (Supplementary Fig. S2B). Thus, we combined samples from H1N1 and H3N2 for the subsequent analyses, where appropriate, as has been done in other studies (McCrone et al. 2018).

We next evaluated the association between the number and characteristics of iSNVs identified in the sample and elements of the clinical metadata abstracted from the electronic medical record (EMR). We fit a negative binomial linear regression to assess the relationship between a subject's age and the number of iSNVs identified in his/her viral isolate (Fig. 3A). The regression analysis demonstrated there was no relationship between age and either synonymous iSNVs ($R^2 = 0.0038$) or non-synonymous iSNVs ($R^2 = 0.019$). To assess the relationship between vaccination status and within-host diversity, we compared the mean number of iSNVs between vaccinated and unvaccinated subjects (Fig. 3B) using a negative binomial regression generalized linear model. The results of this analysis indicated that there was no statistically significant difference between vaccinated, n(Vaccinated) = 71, and unvaccinated individuals, n(Unvaccinated) = 38, for either synonymous iSNVs ($P = 1.00$) or non-synonymous iSNVs ($P = 0.35$). We also considered factors related to the severity of infections, such as days of symptoms prior to sample collection and the underlying health status of the subject, categorized by the pediatric medical complexity algorithm (PMCA), the setting where the sample was obtained, and whether the child was admitted. For this analysis, we considered the setting to be either outpatient or inpatient, where the inpatient category included children whose swabs were obtained following admission to CHOP or in the emergency department (ED). We again used a linear regression model to assess the relationship between iSNVs and age and we used a negative binomial regression model to assess the relationship between iSNVs and PMCA, setting and admission. We observed no significant association between the days of symptoms prior to sample collection for synonymous ($R^2 = 9.68 \times 10^{-7}$) or non-synonymous iSNVs ($R^2 = 1.83 \times 10^{-4}$) (Fig. 3C). Because we did not identify a significant relationship between days since symptom onset, we did not consider this a factor in our subsequent

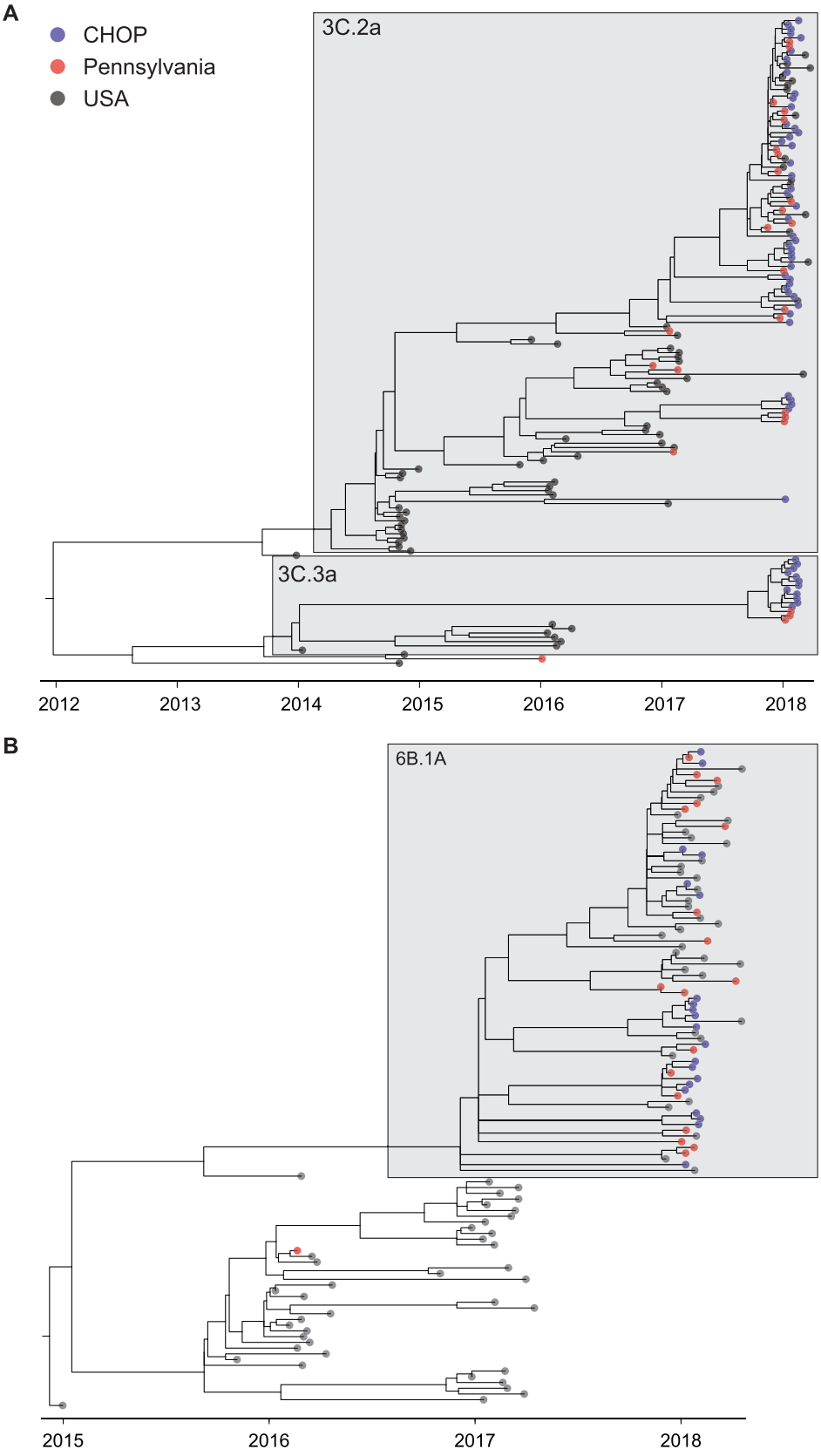

Figure 2. Time-resolved phylogenetic reconstruction of HA sequences from clinical influenza virus isolates for samples collected during the 2017–18 influenza season for: (A) H3N2 and (B) H1N1. The HA sequences from CHOP samples (blue) are shown in relation to sequences obtained from GISAID from the USA (black) and Pennsylvania (red). The clade designations for the CHOP sequences are identified at the clade-defining node.

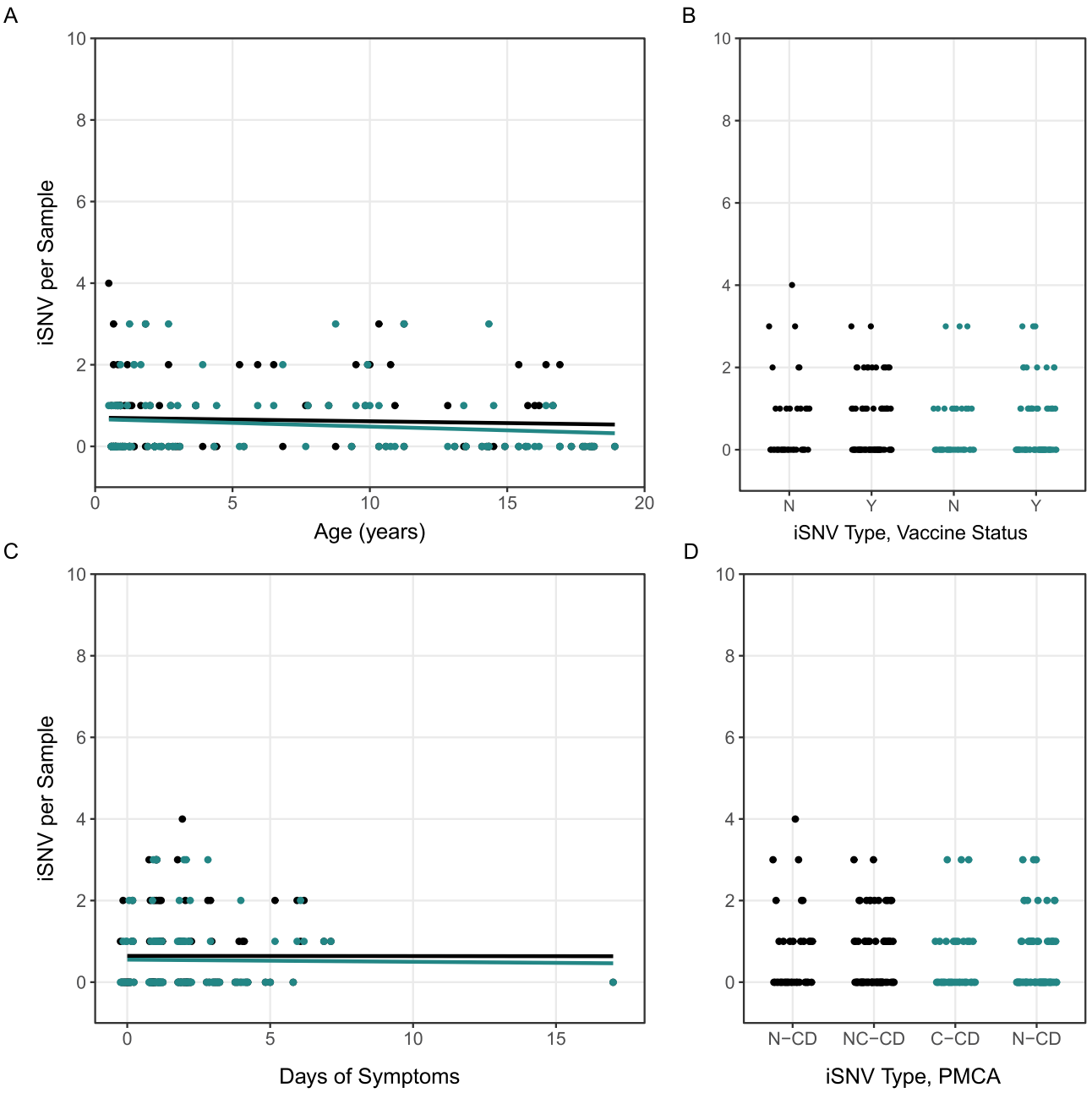

Figure 3. Relationship between the number of iSNVs and four different variables: (A) the subject's age, (B) vaccination status, (C) number of days of symptoms prior to sample collection, and (D) medical complexity as categorized by PMCA. Synonymous iSNVs are represented by black markers, while non-synonymous iSNVs are shown in teal. For numeric variables (age and days of symptoms), associations with the number of iSNVs were assessed using a linear regression model, with the best fitting regression line depicted in the respective plots. Associations between the number of iSNVs and categorical variables (vaccination status and medical complexity) were evaluated using a negative binomial regression model, though no signifciant associations were found at $\alpha = 0.05$. Plots B–D use minor spread along the x-axis to improve the distinctness of individual data points.

analyses. No significant association was observed between test setting and admission status, in relation to the occurrence of synonymous or non-synonymous iSNVs. Specifically, the comparison of inpatients (n(Inpatient) = 86) to outpatients (n(Outpatient) = 23) revealed no significant difference for synonymous ($P = 0.70$) or non-synonymous iSNVs ($P = 0.31$). Similarly, no significant difference was found when comparing admitted (n(Admitted) = 19) versus not admitted (n(Not Admitted) = 90) patients for synonymous ($P = 1.0$) and non-synonymous iSNVs ($P = 0.97$). In examining health status, children categorized as having complex chronic conditions (n(C-CD) = 34) were not found to have a higher number of iSNVs compared to children without chronic disease (n(N-CD) = 32) or with non-complex chronic disease (n(NC-CD) = 43) (Fig. 3D).

In addition to the number of iSNVs identified in the influenza isolates, we also characterized within-host viral diversity using the diversity statistic π (Nei and Li 1979; Nelson and Hughes 2015), calculated as:

$$\pi = \sum_{L}^{l=1} \frac{(D_l)}{L}$$

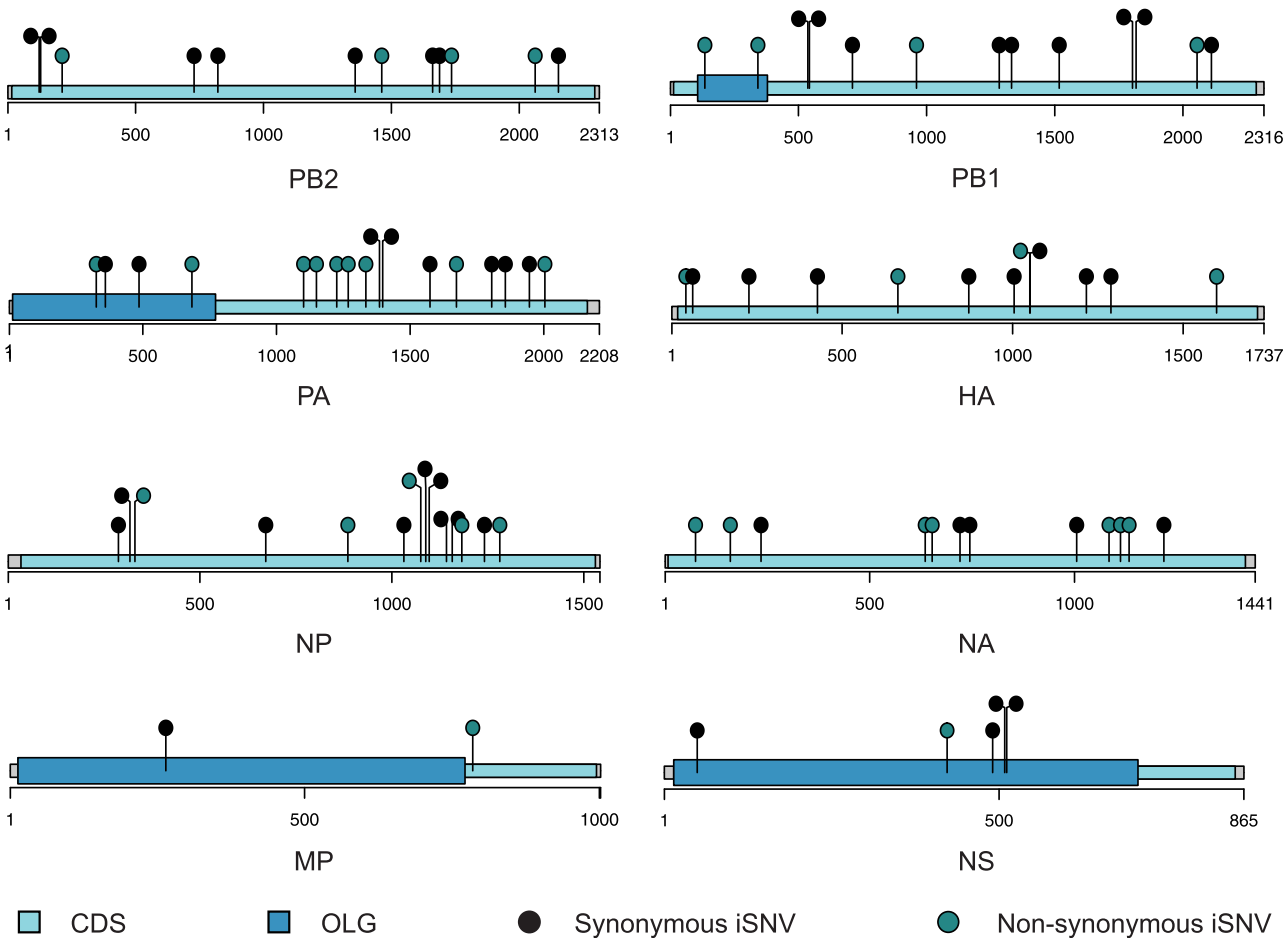

Figure 4. iSNV distribution across influenza gene segments for subjects infected with H3N2. The eight influenza gene segments are shown above with in grey with gene segment's complementary determining sequence (CDS) in light blue. The CDS for the overlapping genes (OLG), PB1-F2, PA-X, M1, and NS1, are shown in dark blue. iSNVs from all subjects plotted along the gene segment in which they were identified according to the figure legend.

where L represent the length of the viral genome and (D_l) represents the expectation value of pairwise diversity at locus l. The expression (D_l) represents the expected nucleotide diversity is calculated as:

$$(D_l) = 1 - \sum_{4}^{i=1} p_i^2$$

where p_i represents the frequency of allele at locus l (Zhao and Illingworth 2019). We chose this metric of estimating nucleotide diversity based on the evaluation by Zhao and Illingworth which showed this metric avoid bias introduced by differences in coverage. For our samples, we estimated the mean$(\pi) = 0.0012 \pm 0.00048$ (Supplementary Fig. S3A). As with the number of iSNVs, we failed to identify an association between π and either the subject's age, vaccine, or health (Supplementary Fig. S3B–D).

In addition to quantifying within-host viral diversity with iSNV counts, we evaluated the distribution of those iSNVs across influenza virus gene segments. Our analysis revealed an even distribution of iSNVs across the gene segments (Table 2), with no evidence of enrichment on or within any specific gene segment for either H3N2 (Fig. 4) or H1N1 subtypes (Fig. 5). This suggests a lack of pressure from diversifying selection in particular areas of the genome, particularly on the HA or NA gene segments.

Location of substitutions on influenza virus glycoproteins

Most neutralizing antibodies target influenza virus glycoproteins, HA, and neuraminidase (NA), embedded in virus membrane and, therefore, we next evaluated the non-synonymous substitutions identified in these gene segments. Overall, we identified few non-synonymous SNVs on either gene occurring at or near previously identified antigenic sites. There were two non-synonymous iSNVs identified on HA gene segment, one occurring on H3N2 and the other on H1N1 (Fig. 6A and B). The H3 iSNV G200V, identified in CHOP-080, was adjacent to antigenic site D, though it did not widely circulate during 2017–18 or subsequent influenza seasons (Neher and Bedford 2015). The H1 iSNV K311R, identified in CHOP-127, was not in close proximity to influenza's antigenic sites. We identified a greater number of mutations on the NA gene segment for both N2 and N1. For the N2 NA gene segment, we identified five unique mutations, each in a single subject (Fig. 6C–E). While we did not identify the same NA iSNVs across subjects, the mutations appeared to cluster around two regions. The first cluster contained R210K from CHOP-096 and V216I from CHOP-043. The numbering refers to the residue position in the mature N2 peptide (Zhu et al. 2012). Many of the iSNVs occurred within or near previously

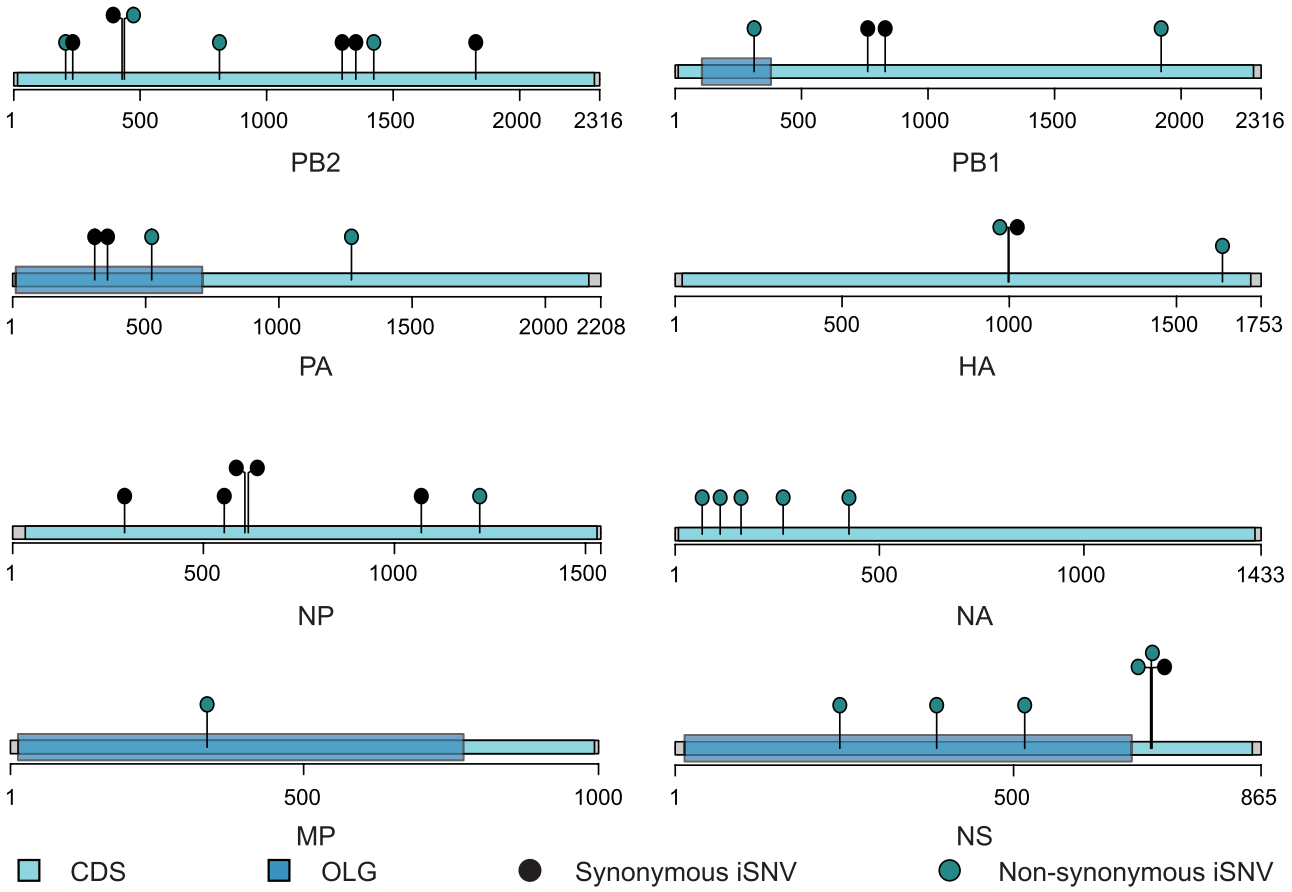

Figure 5. iSNV distribution across influenza gene segments for subjects infected with H1N1. The eight influenza gene segments are shown above with in grey with gene segment's complementary determining sequence (CDS) in light blue. The CDS for the overlapping genes (OLG), PB1-F2, PA-X, M1, and NS1, are shown in dark blue. iSNVs from all subjects plotted along the gene segment in which they were identified according to the figure legend.

identified antibody binding sites. For the N1 protein, we identified a total of two iSNVs in the head domain (Fig. 6F-G)—both in subject CHOP-037, though neither occurred within or near previously described antibody-binding sites (Dai et al. 2021; Strohmeier et al. 2021).

Discussion

Our study shows that influenza viruses recovered from infected children generally display low levels of within-host viral diversity. This observation is in line with prior studies of within-host influenza evolution that have similarly shown that most infections are characterized by low-viral diversity (Debbink et al. 2017; McCrone et al. 2018; Han et al. 2021). While there were two individuals with significantly higher levels of diversity, further analysis indicated these were likely the result of mixed infection or cross-contamination. We also evaluated the relationship between metrics of within-host diversity, the number of iSNVs and the diversity statistics π. We showed that there was not a significant association between either of these metrics of diversity with elements of the clinical metadata including the child's age, vaccine, or health status. We did not explicitly consider dN/dS, a commonly used method for characterizing the type of selective pressure acting on a population, as this metric assumes that the population has had time to equilibrate, and may be neither sensitive nor specific for identifying diversifying selection in within-host viral populations (Kryazhimskiy, Plotkin, and Gojobori 2008; Lauring 2020). Finally, we show that a subset of the non-synonymous iSNVs detected on the influenza glycoproteins, HA, and NA, occur in proximity to previously described antigenic sites and known epitopes. Though, given the small number of these iSNVs and absence of paired serum samples, it is difficult draw broad conclusions related to immune escape from this observation.

The principal finding of this study is that there are low levels of within-host viral diversity in the pediatric population. The observation of low diversity, quantified by the number of iSNVs relative to the consensus sequence of the influenza virus in the corresponding sample, is consistent with other studies in the general (Debbink et al. 2017; McCrone et al. 2018) and pediatric populations (Han et al. 2021). Similar to previous reports, our findings also indicate no significant difference in viral diversity between vaccinated and unvaccinated individuals (Dinis et al. 2016; Debbink et al. 2017), or host age (Han et al. 2021). Furthermore, we did not identify a significant association between viral diversity and proxies for disease severity (test setting and whether the child was admitted) or health. Our study included six children who were immunocompromised. While previous work has shown higher levels of influenza viral diversity and diversifying selection can be observed in immunocompromised individuals, this observation was limited to prolonged infections (Xue et al. 2017). The viral isolates from children in our study were obtained between 1and 4 days following symptom onset, and it is possible that longer amounts of time is required to allow for the accumulation of viral mutations in immunocompromised children.

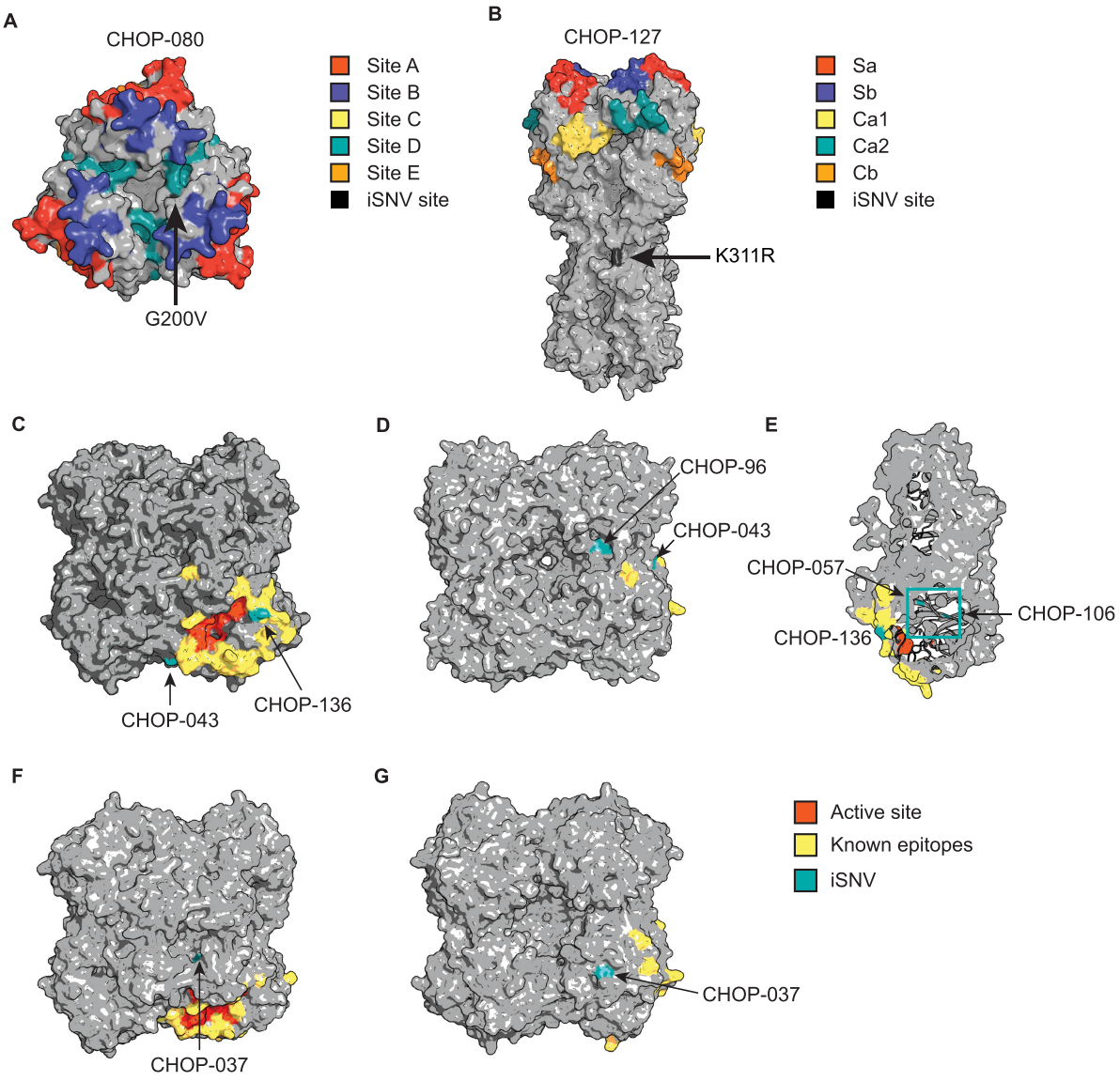

Figure 6. The structures of HA and NA featuring non-synonymous iSNVs identified in study subjects infected with H3N2 and H1N1 are illustrated. For HA, key antigenic sites on H3 (A, PDB 4O5N, (Burke, Smith, and Digard 2014)) and H1 (B, PDB 4LXV, (Yang et al. 2014; Liu et al. 2018)) are shown. The vaccine status and PMCA categorization as no chronic disease (N-CD), non-complex chronic disease (NC-CD), and chronic disease (C-CD) is listed for each of the subjects, where a mutation was identified. The most representative views of the HA structure are displayed for samples CHOP-080 (H3, unvaccinated, C-CD) and CHOP-127 (H1, unvaccinated, NC-CD), the subjects in which non-synonymous iSNVs were identified, with the iSNVs shown in black. For NA, the active site and previously identified epitope sites are depicted in varying colors as per the provided legend (Chen et al. 2018b; McAuley et al. 2019; Dai et al. 2021; Kirkpatrick Roubidoux et al. 2021), with iSNV locations shown in teal and labeled with the subject sample ID wherein they occurred. The subjects in which iSNVs affecting N2 were: CHOP-043 (vaccinated, N-CD), CHOP-057 (unvaccinated, NC-CD), CHOP-096 (vaccinated, C-CD), and CHOP-136 (vaccinated, C-CD). Three orientations, (C) top-down, (D) bottom-up, and (E) internal-side are shown for the N2 protein (PBD 4GZX, (Zhu et al. 2012). One iSNV was identified affecting N1 in CHOP-037 (N1, vaccinated, NC-CD). Two orientations, (F) top-down and (G) bottom-up for the N1 protein (PBD 7S0I (Strohmeier et al. 2021)). All illustrations were created with Pymol 2.1 (Schrödinger 2015).

The limited timing of viral replication prior to sample collection may also explain the lack of association between days since symptom onset and the number of detected iSNVs, as was observed by Han et al. (2021). The samples from our cohort were collected with a mean of 3 days of preceding symptom, whereas the sampis from the Han et al., cohort had a mean of 6.1 days of preceding symptoms, potentially allowing more time for viral mutations to accumulate. Additional studies of influenza virus within-host diversity have also failed to identify a link between days following symptom onset and the number of iSNVs, though the average number of days of symptoms at the time of sample collection was also <6 (Debbink et al. 2017; McCrone et al. 2018). Also, unlike Han et al., we did not observe an association between C_t and the number of detected iSNVs. This disparity can likely be attributed to differences in the C_t values of the samples used in our respective studies. For our study, we used a maximum C_t value of 30 as the upper threshold for samples we selected for sequencing, which contributed to the mean sample C_t value of 22.9 for our samples. In contrast, the mean C_t value for Han et al.'s study was 30.0. Finally, Han et al. (2021), showed that the number of iSNVs differed between influenza subtypes H1N1 and H3N2. While we did not observe a difference in the number of iSNVs identified in H3N2 vs H1N1 isolates, this reflect that our sample set included many fewer H1N1 samples than H3N2 samples, which is not surprising

given that the 2017–2018 flu season was H3N2 dominant (Garten et al. 2018). Therefore, it is possible that the distribution of synonymous versus non-synonymous iSNVs may have changed if more H1N1 samples had been identified. Thus, the seemingly discrepant findings between these two studies may be largely reconcilable, owing to differences in the characteristics of the samples studied.

Previous studies that have shown evidence of diversifying selection in humans at the intrahost level for influenza virus have done so in the setting of prolonged infections (Ghedin et al. 2011; Xue et al. 2017). Recent studies have attempted to reconcile the disparate observations of levels of viral diversity at the intrahost level with frequent emergence of antigenic variants at the population level. Morris et al. have proposed that selection for antigenic variants occurs at the time of inoculation rather than over the course of infection (Morris et al. 2020). Experimental work using barcoded virus in guinea pigs shows high viral diversity for 1–2 days following inoculation before undergoing a steep reduction in viral diversity (Holmes et al. 2023). Importantly, both studies suggest the potential for immune-mediated selection to operate at or near the time of viral transmission. Thus, there is the potential to find evidence of selection by considering mutations detected at the consensus level relative to other co-circulating viruses within the same clade. Future studies should characterize the specificity of individual serum antibody responses to determine if these antibodies are able to bind and neutralize the infecting virus relative to the other co-circulating virus.

Despite the constraints of our study, notably the lack of host serum samples and availability of only a single viral isolate per subject, our findings provide key insights into the evolutionary dynamics of influenza virus in children. The findings of our study emphasize the need for additional research to better understand the intricate interplay of host immunity and influenza virus evolution, thereby shedding light on their collective implications for global viral evolutionary dynamics and influenza control.

METHODS

Subjects and specimens

This study was approved by the CHOP institutional review board (IRB) and we received a waiver of consent as this study was of minimal risk and, due to the retrospective nature and size of this study, it could not be practically carried out without this waiver. We collected banked samples from the CHOP Infectious Diseases Diagnostics Laboratory (IDDL) that were left over from nasopharyngeal swabs and nasal aspirates collected between October 2017 and March 2018. We selected samples for inclusion using random, targeted subsampling based on the subject's age and vaccination status. We considered all available samples for subjects between 6 months and 18 years of age at the time of sample collection but excluded all samples with a C_t ≥30. Cycle threshold was determined from the IDDL's laboratory derived RT-PCR targeting the matrix gene segment. Chart review was conducted manually with data stored in a secure, password-protected spreadsheet and REDCap (Harris et al. 2009, 2019). Data sources used for chart review include information contained in the EPIC, CHOP's EMR and Pennsylvania Statewide Immunization Information System (PA-SIIS) accessed via EPIC.

Virus sample isolation and NGS protocol

We extracted influenza viral RNA from the clinical isolates using QIAGEN QIAmp Viral RNA Mini Kits. The extracted RNA was reverse transcribed and amplified using the Superscript III One-Step RT-PCR Kit with Platinum™ Taq High Fidelity DNA Polymerase (Fisher #12574-035) and universal influenza primers Uni12/Inf1 (GGGGGGAGCAAAAGCAGG), Uni12/Inf3 (GGGGGGAGCGAAAGCAGG), and Uni13/Inf1 (CGGGTTATTAGTAGAAACAAGG). Reactions consisted of 12.5 µl 2× buffer, 0.2 µl Uni12/Inf1, 0.3 µl Uni12/Inf3, 0.5 µl Uni13/Inf1, 0.5 µl Taq HiFi DNA polymerase, 6 µl UltraPure DEPC-Treated Water (Thermofisher #750023), and 5 µl extracted viral RNA. The Thermocycler protocol was as follows: 60 min at 42°C, 2 min at 94°C, followed by 5 cycles of 30 s at 94°C, 30 s at 44°C, and 3 min at 68°C, followed by 28 cycles of 30 s at 94°C, 30 s at 57°C, and 3 min at 68°C, followed by storage at 4°C until the next step. We purified the resulting cDNA with Ampure XP magnetic beads (Beckman Coulter #A63881). Presence of cDNA was confirmed with Quant-IT™ PicoGreen (Thermo #PLL496) prior to proceeding with NGS library preparation. We completed library prep using the Illumina DNA Prep (M) Tagmentation (Illumina #20060059 and 20,018,705). Samples were barcoded with IDT for Illumina DNA/RNA UD (#20027213). Following library preparation, we repeated the PicoGreen protocol to normalize DNA concentrations for pooling. The pooled cDNA was again purified using Ampure magnetic beads. We quantified the concentration of the pooled library using the Qubit™ 1× dsDNA high-sensitivity assay kit (Thermo #Q33230) and diluted the library to 2 nM. The diluted library was loaded via either a NextSeq 500/550 High Output Kit 2.5 (300 cycles) (Illumina #20024908) or NextSeq 1000/2000 P2 (200 Cycles) v3 (Illumina #20046812) kits for sequencing on either the Illumina NextSeq 500 or Illumina NextSeq 2000 sequencer. Samples sequenced on 31 March 2022, 04 May 2022, 17 June 2022, and 12 September 2022 used the NextSeq 500 with 2x125bp paired-end reads. Samples sequenced on 23 February 2023, 10 March 2023 and 01 May 2023 used the NextSeq 2000 with 2x90bp paired-end reads. Each run included a positive control, either a virus with A/Cambodia/e0826360/2020 HA and NA segments and A/Puerto Rico/8/1934 (PR8) internals created by reverse genetics or a previously successfully sequenced clinical isolate, and a negative water control.

Sequence processing and variant identification

Samples were demultiplexed using custom bash scripts and then processed using FluPipeline. FluPipeline is a command line program that finds intra-host variants using short read data. It takes as input paired-end short reads in fastq format and a list of influenza genomes in Genbank format. In the pre-processing, step all reads are trimmed and quality filtered using fastp v0.22.01 (Chen et al. 2018a). Afterwards, a subset of reads from each sample is mapped to each influenza genome using BWA-MEM v0.7.172 (Li 2013). The genome with the highest read coverage and depth at the HA contig, or in the event of a tie, the highest coverage and depth over all contigs, is selected as the reference genome for first-pass variant detection. FluPipeline identifies variants in two passes. In the first pass, reads are aligned against the selected reference genome using BWA-MEM. Variants are called and major variants (default occurrence of ≥50 per cent for SNPs and ≥80 per cent for INDELs) are used to create a consensus sequence. In the second pass, reads are aligned using BWA-MEM against the consensus sequence to obtain intra-host variants (default ≥5 per cent for SNPs and INDELs). The default parameters use bcftools v1.15.1 (Li 2011) as the variant caller for the first pass and BBMap v38.14 (Bushnell 2014) for the second pass. The user can also specify whether a second pass is necessary depending on their needs. Users can also tune parameters that impact variant detection such as mapping quality, gap open and extension penalties, and

variant read depth, and score. FluPipeline is available on GitHub (https://github.com/agmcfarland/FluPipeline).

Our pre-processing steps with FluPipeline involved trimming either 22 bp (for samples sequenced on the NextSeq 500) or 31 bp (for samples sequenced on the NextSeq 2000) from the short read ends. We randomly selected 10,000 short reads for alignment to reference strains: B/Wisconsin/01/2010 Yamagata lineage (CY115183-CY115190), B/Brisbane/60/2008 Victoria lineage (FJ766840, FJ766841, KC866602-KC866607), influenza A virus ((A/turkey/Massachusetts/3740/1965(H6N2), CY087752), A/Michigan/45/2015(H1N1) (MK622934- MK622941), and A/Washington/17/2016(H3N2) (KX414254-KX414261). Following strain assignment, the short reads were assembled to the selected reference sequence with down-sampling to 100,000,000. We called major variants with bcftools (Li 2011) and minor variants with bbtools (Bushnell 2014) with a minimum frequency of 0.01. We used different variant callers as bcftools, while it has been used and validated extensive, has difficulty accounting for mutations, where multiple variant alleles may be present at single position within a sample, as could be expected for viral sequences. Short reads were filtered for a minimum average base quality score of 30 and a BWA-derived read mapping score of 40. Negative controls were considered successful if there was a coverage of ≤10 reads for ≤10 percent of the influenza genome. Our reads had high depth, with an average coverage between 1000 and 10,000 reads per position, Supplementary Figures S4–10.

Sequence post-processing was completed with custom R scripts available on GitHub (https://github.com/HensleyLab-UPENN/CHOP_Retrospective_GitHub). Portions of the code for the custom R scripts in for data processing and statistical analysis were reviewed and optimized with the assistance of the OpenAI GPT-4 language model (OpenAI). Packages employed in the analysis are identified in the Readme file. Packages used for statistical analysis and manipulation of genomic information used for the above analyses include: tidyverse v2.0.0 (Wickham et al. 2019), Rsamtools v2.14.0 (Morgan et al. 2022), Biostrings v2.66.0 (Pagès et al. 2022), Lawrence et al. 2013 dplyr 1.1.2 (Wickham et al. 2023a), seqinr 4.2.30 (Charif and Lobry 2007), stringr v1.5.0 (Wickham 2022), phylotools v0.2.2 (Zhang 2017), stringi v1.7.12 (Gagolewski 2022), ggplot2 v3.4.2 (Wickham 2016), tidysq v1.2.0 (Rafacz, Burdukiewicz, and Bakala 2022), patchwork v1.1.2 (Pedersen 2022), gridExtra v2.3 (Auguie 2017), ShortRead v1.56.1 (Morgan et al. 2009), Biostrings v2.66.0 (Pagès et al. 2022), and data.table v1.14.8 (Dowle and Srinivasan 2023). The fastq files for the influenza isolates analyzed in this study are available from the NCBI SRA database in BioProject PRJNA1066787.

During post-processing, we applied additional QC criteria to ensure adequate coverage and sequence reproducibility. For coverage, we required that samples have >100 reads for ≥95 per cent of the genome. Individual sites with coverage <100 reads were masked. To ensure reproducibility, we required that all samples have technical replicates, i.e. replicates from separate RNA extractions, and that the consensus sequences for the replicates match. We did allow exceptions for disagreements between the consensus sequences if: (1) the mismatch occurred within in either the first or last 20 bp of the gene segment, (2) the mismatch occurred because one of the replicate positions had been masked, or (3) if there was a high frequency variant (>10 per cent) at the site of the mismatch. If there were more than two technical replicates, we selected the replicates with the lowest proportion of masked positions, i.e. positions with coverage <100 reads. In addition, excluded those samples with >1 iSNV detected and <50 per cent concordance between the replicates. Variants were considered nonsynonymous if they resulted in an amino acid change in the coding section of the gene segment. In cases where overlapping reading frames were present, we considered a mutation to be nonsynomous if it resulted in an amino acid change in either reading frame. We required that all variants called be present in both technical replicates above 3 per cent, our minimum frequency threshold, and have an average quality score of ≥20. We did consider variants called in a single replicate but present in the second replicate with an allele frequency of greater than or equal to the minimum variant frequency even if it was not called by bbtools. This criterion applied to 3 out of 241 iSNVs in our dataset: G1052A, identified in subjects CHOP-086 and CHOP-115, A541T identified in CHOP-117, and T548C identified in subjects CHOP-101 ad CHOP-117. While the discrepancy in whether or not the iSNV was called by bbtools in the replicates could indicate these iSNVs represent sequencing artifacts, we have chosen not to exclude these iSNVs as they represent a small proportion of the total iSNVs. Further, the decision of whether or not to exclude them would not affect the primary findings of this paper.

Statistical evaluations

For the assessment of the relationship between the number of iSNVs and elements of the clinical metadata, we used both linear regression models for numerical variables and nonlinear generalized mixed linear models for categorical variables. We fit the independent variables, the number of synonymous and nonsynonymous iSNVs, separately. We considered the following numerical variables: the number of days following symptom onset at the time of sample collection, age, and C_t. There was clearly no association between any of the numerical-dependent variables and the number of iSNVs detected in each sample. Therefore, we did not assess for covariance between these variables. In addition, we felt it was more statistically appropriate to exclude these variables from further analysis involving the categorical variables, rather than transforming the categorical variables to assess to variance. The categorical variables considered included the subject's vaccination status, categorization based on medical complexity, the setting in which the test was sent, and whether the subject was admitted to the hospital. We chose to simultaneously fit these variables using a negative binomial regression and a Bonferroni correction to account for multiple comparisons.

All statistical analyses were completed with custom R scripts available on GitHub (https://github.com/HensleyLab-UPENN/CHOP_Retrospective_GitHub). Statistical and genomic packages used for the above analyses include: tidyverse v2.0.0 (Wickham et al. 2019), phylotools v0.2.2 (Zhang 2017), dplyer v1.1.2 (Wickham et al. 2023a), stringi v1.7.12 (Gagolewski 2022), ggplot2 v3.4.2 (Wickham 2016), seqinr v4.2.30 (Charif and Lobry 2007), patchwork v1.1.2 (Pedersen 2022), data.table v1.14.8 (Dowle and Srinivasan 2023), cowplot v1.1.1 (Wilke 2020), reshape2 v1.4.4 (Wickham 2007), MASS v7.3.60 (Venables and Ripley 2002), trackViewer v1.34.0 (Ou and Zhu 2019), GenomicRanges v1.50.2 (Lawrence et al. 2013), phylotools v0.2.2 (Zhang 2017), DECIPHER v2.26.0 (Wright 2016), tidysq v1.2.0, (Rafacz, Burdukiewicz, and Bakala 2022), genbankr v1.26.0 (Becker and Lawrence, 2022), Biostrings v2.66.0 (Pagès et al. 2022).

Phylogenetic analysis

The findings of this study are based on metadata associated with sequences available on GISAID up to 4 October 2023, full details of which are recorded in Supplementary Table S2. For the

GISAID sequences, we obtained all original full-length, unpassaged, HA gene segment sequences for influenza strains H3N2 and H1N1pdm submitted by the Centers for Disease Control (CDC) collected during influenza season in the USA, between October 2013 and May 2018 (Khare et al. 2021). For each strain, we randomly subsampled these sequences to obtain up to fifteen sequences per flu season, pending availability, from influenza seasons between October 2013—May 2017 and 30 sequences from 2017 to 2018 influenza season. We also subsampled the October 2017–18 sequences to obtain 50 sequences collected in Pennsylvania (PA). Sequences and metadata were cleaned using custom R scripts. The sequences obtained from GISAID were combined with the sequences generated from the CHOP samples of the corresponding strain and duplicates were removed. We generated phylogenies with standard NextStrain tools and scripts (Hadfield et al. 2018; Sagulenko, Puller, and Neher 2018), with specific options as follows. The sequences were aligned in NextStrain to the reference strains A/Wisconsin/67/2004 for H3N2 (CY163680) and A/California/07/2009 (CY121680) for H1N1pdm using MAFFT (Katoh et al. 2002). Phylogenetic trees were constructed using IQ-TREE (Nguyen et al. 2015) with the GTR substitution model. Branch lengths for the time-resolved trees were inferred with TreeTime (Sagulenko, Puller, and Neher 2018) and we allowed up to ten iterations for convergence. The phylogenies were visualized and annotated in FigTree (Rambaut 2018). All custom scripts are available on Github (https://github.com/HensleyLab-UPENN/CHOP_Retrospective_GitHub).

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

We gratefully acknowledge all data contributors, i.e. the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. The authors would like to acknowledge the assistance of the OpenAI GPT-4 language model in proofreading the manuscript. We also acknowledge the sample curation efforts of the CHOP Infectious Diseases Diagnostics Laboratory (IDDL), particularly Michael Elkan and Jeffery Fink, without which this work would not have been possible. This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N93021C00015 and Grant No. 1R01AI108686 (to S.E.H.) and R33HD105594 (to A.O.J.). S.E.H. and A.O.J. are Investigators in the Pathogenesis of Infectious Disease Awards from the Burroughs Wellcome Fund. A.S.L. received funding support from the CHOP Department of Pediatrics T32 Research Training Grant No. 5T32HD043021-19 and the Pediatric Infectious Diseases Society (PIDS) – St. Jude Children's Research Hospital Fellowship Program in Basic and Translational Research. L. M. received funding support through the University of Pennsylvania Training Grant in Virology No. 5T32AI007324.

Conflict of interest: S.E.H. is a co-inventor on patents that describe the use of nucleoside-modified mRNA as a vaccine platform. S.E.H reports receiving consulting fees from Sanofi, Pfizer, Lumen, Novavax, and Merck.

References

Auguie, B. (2017) '*gridExtra: Miscellaneous Functions for "Grid" Graphics*'.

Becker G. and Lawrence M. (2022) '*genbankr: Parsing GenBank files into semantically useful objects*'.

Berry J. G., Hall M., Cohen E., O'Neill M. and Feudtner C. (2015) 'Ways to Identify Children with Medical Complexity and the Importance of Why', *The Journal of Pediatrics*, 167: 229–37.

Braun, K. M. et al. (2021) 'Acute SARS-CoV-2 Infections Harbor Limited Within-host Diversity and Transmit via Tight Transmission Bottlenecks', *PLoS Pathogens*, 17: e1009849.

Burke, D. F., Smith, D. J. and Digard, P. (2014) 'A Recommended Numbering Scheme for Influenza A HA Subtypes', *PLoS ONE*, 9: e112302.

Bushnell, B. (2014) 'BBMap: A Fast, Accurate, Splice-Aware Aligner'.

Cannon, J. L. et al. (2019) 'Impact of Long-term Storage of Clinical Samples Collected from 1996 to 2017 on RT-PCR Detection of Norovirus', *Journal of Virological Methods*, 267: 35–41.

Chambers, B. S. et al. (2015) 'Identification of Hemagglutinin Residues Responsible for H3N2 Antigenic Drift during the 2014-2015 Influenza Season', *Cell Reports*, 12: 1–6.

Charif, D. and Lobry, J. R. (2007) 'SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis', in Bastolla, U. et al. (eds) *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*, pp. 207–32. Springer: New York.

Chen, S. et al. (2018a) 'Fastp: An Ultra-fast All-in-one FASTQ Preprocessor', *Bioinformatics*, 34: i884–90.

Chen, Y. Q. et al. (2018b) 'Influenza Infection in Humans Induces Broadly Cross-Reactive and Protective Neuraminidase-Reactive Antibodies', *Cell*, 173: 417–429.e10.

Choi, B. et al. (2020) 'Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host', *New England Journal of Medicine*, 383: 2291–3.

Coffin, S. E. et al. (2007) 'Incidence, Complications, and Risk Factors for Prolonged Stay in Children Hospitalized with Community-acquired Influenza', *Pediatrics*, 119: 740–8.

Dai, M. et al. (2021) 'Analysis of the Evolution of Pandemic Influenza A(H1N1) Virus Neuraminidase Reveals Entanglement of Different Phenotypic Characteristics', *MBio*, 12: 10–128.

Debbink, K. et al. (2017) 'Vaccination Has Minimal Impact on the Intrahost Diversity of H3N2 Influenza Viruses', *PLoS Pathogens*, 13: 1–18.

Dinis, J. M. et al. (2016) 'Deep Sequencing Reveals Potential Antigenic Variants at Low Frequencies in Influenza A Virus-Infected Humans', *Journal of Virology*, 90: 3355–65.

D'Mello, T. et al. (2015) 'Morbidity and Mortality Weekly Report', *Morbidity and Mortality Weekly Report*, 64: 206–12.

Dowle, M. and Srinivasan, A. (2023) '*Data.Table: Extension of 'data.frame''*, <https://CRAN.R-project.org/package=data.table>

Eden, J.-S. et al. (2017) 'Persistent Infections in Immunocompromised Hosts are Rarely Sources of New Pathogen Variants', *Virus Evolution*, 3: 1–10.

Gagolewski, M. (2022) 'Stringi: Fast and Portable Character String Processing in R', *Journal of Statistical Software*, 103: 1–59.

Garten, R. et al. (2018) 'Update: Influenza Activity in the United States during the 2017–18 Season and Composition of the 2018–19 Influenza Vaccine', *MMWR Morbidity and Mortality Weekly Report*, 67: 634–42.

Ghedin, E. et al. (2011) 'Deep Sequencing Reveals Mixed Infection with 2009 Pandemic Influenza A (H1N1) Virus Strains and the

Emergence of Oseltamivir Resistance', *Journal of Infectious Diseases*, 203: 168–74.
Gonzalez-Reiche, A. S. et al. (2023) 'Sequential Intrahost Evolution and Onward Transmission of SARS-CoV-2 Variants', *Nature Communications*, 14: 3235.
Grubaugh, N. D. et al. (2019) 'An Amplicon-based Sequencing Framework for Accurately Measuring Intrahost Virus Diversity Using PrimalSeq and iVar', *Genome Biology*, 20: 8.
Hadfield, J. et al. (2018) 'NextStrain: Real-time Tracking of Pathogen Evolution', *Bioinformatics*, 34: 4121–3.
Han, A. X. et al. (2021) 'Within-host Evolutionary Dynamics of Seasonal and Pandemic Human Influenza a Viruses in Young Children', *eLife*, 10: e68917.
Harris, P. A. et al. (2009) 'Research Electronic Data Capture (Redcap)—a Metadata-driven Methodology and Workflow Process for Providing Translational Research Informatics Support', *Journal of Biomedical Informatics*, 42: 377–81.
Harris, P. A. et al. (2019) 'The REDCap Consortium: Building an International Community of Software Platform Partners', *Journal of Biomedical Informatics*, 95: 103208.
Holmes, K. E. et al. (2023) 'Viral Expansion after Transfer Is A Primary Driver of Influenza A Virus Transmission Bottlenecks', *bioRxiv: The Preprint Server for Biology*: 2023.11.19.567585.
Iuliano, A. D. et al. (2018) 'Estimates of Global Seasonal Influenza-associated Respiratory Mortality: A Modelling Study', *The Lancet*, 391: 1285–300.
Katoh, K. et al. (2002) 'MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform', *Nucleic Acids Research*, 30: 3059–66.
Khare, S. et al. (2021) 'GISAID's Role in Pandemic Response', *China CDC Weekly*, 3: 1049–51.
Kirkpatrick Roubidoux, E. et al. (2021) 'Identification and Characterization of Novel Antibody Epitopes on the N2 Neuraminidase', *mSphere*, 6: 10-12.
Ko, KKK. et al. (2022) 'Emergence of SARS-CoV-2 Spike Mutations during Prolonged Infection in Immunocompromised Hosts', *Microbiology Spectrum*, 10: e0079122.
Kryazhimskiy, S., Plotkin, J. B. and Gojobori, T. (2008) 'The Population Genetics of dN/dS', *PLoS Genetics*, 4: e1000304.
Lauring, A. S. (2020) 'Within-Host Viral Diversity: A Window into Viral Evolution', *Annual Review of Virology*, 7: 63–81.
Lawrence, M. et al. (2013) 'Software for Computing and Annotating Genomic Ranges', *PLoS Computational Biology*, 9: e1003118.
Li, H. (2011) 'A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data', *Bioinformatics*, 27: 2987–93.
——— (2013) 'Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM', arXiv:1303.3997, arXiv preprint arXiv.
Li, Y. et al. (2013) 'Single Hemagglutinin Mutations that Alter Both Antigenicity and Receptor Binding Avidity Influence Influenza Virus Antigenic Clustering', *Journal of Virology*, 87: 9904–10.
Liu, S. T. H. et al. (2018) 'Antigenic Sites in Influenza H1 Hemagglutinin Display Species-specific Immunodominance', *Journal of Clinical Investigation*, 128: 4992–6.
Lloyd-Smith, J. O. et al. (2005) 'Superspreading and the Effect of Individual Variation on Disease Emergence', *Nature*, 438: 355–9.
Lumby, C. K. et al. (2020) 'A Large Effective Population Size for Established Within-host Influenza Virus Infection', *eLife*, 9: 1–17.
Lythgoe, K. A. et al. (2021) 'SARS-CoV-2 Within-host Diversity and Transmission', *Science*, 372: eabg0821.
McAuley, J. L. et al. (2019) 'Influenza Virus Neuraminidase Structure and Functions', *Frontiers in Microbiology*, 10: 432609.
McCrone, J. T. et al. (2018) 'Stochastic Processes Constrain the within and between Host Evolution of Influenza Virus', *eLife*, 7: 1–19.
McCrone, J. T., Lauring, A. S. and Dermody, T. S. (2016) 'Measurements of Intrahost Viral Diversity are Extremely Sensitive to Systematic Errors in Variant Calling', *Journal of Virology*, 90: 6884–95.
Moncla, L. H. et al. (2020) 'Quantifying Within-host Diversity of H5N1 Influenza Viruses in Humans and Poultry in Cambodia', *PLoS Pathogens*, 16: e1008191.
Morgan, M. et al. (2009) 'ShortRead: A Bioconductor Package for Input, Quality Assessment and Exploration of High-throughput Sequence Data', *Bioinformatics*, 25: 2607–8.
Morgan, M. et al. (2022) '*Rsamtools: Binary Alignment (BAM), FASTA, Variant Call (BCF), and Tabix File Import*', <https://bioconductor.org/packages/Rsamtools>.
Morris, D. H. et al. (2020) 'Asynchrony between Virus Diversity and Antibody Selection Limits Influenza Virus Evolution', *eLife*, 9: e62105.
Neher, R. A. and Bedford, T. (2015) 'Nextflu: Real-time Tracking of Seasonal Influenza Virus Evolution in Humans', *Bioinformatics*, 31: 3546–8.
Nei, M. and Li, W. H. (1979) 'Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases', *Proceedings of the National Academy of Sciences of the United States of America*, 76: 5269–73.
Nelson, C. W. and Hughes, A. L. (2015) 'Within-host Nucleotide Diversity of Virus Populations: Insights from Next-generation Sequencing', *Infection Genetics & Evolution*, 30: 1–7.
Ng, S. et al. (2016) 'The Timeline of Influenza Virus Shedding in Children and Adults in a Household Transmission Study of Influenza in Managua, Nicaragua', *Pediatric Infectious Disease Journal*, 35: 583–6.
Nguyen, L.-T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.
OpenAI. (2023) '*ChatGPT (Sept 25 version) [Large language model]*', <https://chat.openai.com/chat>.
Ou, J. and Zhu, L. J. (2019) 'trackViewer: A Bioconductor Package for Interactive and Integrative Visualization of Multi-omics Data', *Nature Methods*, 16: 453–4.
Pagès, H. et al. (2022) '*Biostrings: Efficient Manipulation of Biological Strings*', <https://bioconductor.org/packages/Biostrings>.
Pedersen, T. L. (2022) '*Patchwork: The Composer of Plots*', <https://CRAN.R-project.org/package=patchwork>.
Rafacz, D., Burdukiewicz, M. and Bakala, L. (2022) '*Tidysq: Tidy Processing and Analysis of Biological Sequences*', <https://CRAN.R-project.org/package=tidysq>.
Rambaut A. (2018) '*Figtree v1.4.4. Institute of Evolutionary Biology*', University of Edinburgh: Edinburgh.
Roder, A. et al. (2023) 'Optimized Quantification of Intrahost Viral Diversity in SARS-CoV-2 and Influenza Virus Sequence Data', *mBio* 14: e01046–23.
Sagulenko, P., Puller, V. and Neher, R. A. (2018) 'TreeTime: Maximum-likelihood Phylodynamic Analysis', *Virus Evolution*, 4: vex042.
Schrödinger, L. L. C. (2015) '*The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version~1.8*', <http://www.pymol.org/pymol>.
Shrestha, L. B. et al. (2022) 'Evolution of the SARS-CoV-2 Omicron Variants BA.1 To BA.5: Implications for Immune Escape and Transmission', *Reviews in Medical Virology*, 32: e2381.
Simon, T. D. et al. (2014) 'Pediatric Medical Complexity Algorithm: A New Method to Stratify Children by Medical Complexity', *Pediatrics*, 133: e1647.

Strohmeier, S. et al. (2021) 'A Novel Recombinant Influenza Virus Neuraminidase Vaccine Candidate Stabilized by A Measles Virus Phosphoprotein Tetramerization Domain Provides Robust Protection from Virus Challenge in the Mouse Model', *mBio*, 12: e02241–21.

Valesano, A. L. et al. (2020) 'Influenza B Viruses Exhibit Lower Within-Host Diversity than Influenza A Viruses in Human Hosts', *Journal of Virology*, 94: 10–128.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*, 4th edn. Springer: New York.

Voloch C. M. et al. (2021) 'Intra-host evolution during SARS-CoV-2 prolonged infection', *Virus Evolution*, 7.

Weigang, S. et al. (2021) 'Within-host Evolution of SARS-CoV-2 in an Immunosuppressed COVID-19 Patient as a Source of Immune Escape Variants', *Nature Communications*, 12: 6405.

Wickham, H. (2007) 'Reshaping Data with the Reshape Package', *Journal of Statistical Software*, 21: 1–20.

——— (2016) *Ggplot2: Elegant Graphics for Data Analysis*. Springer: New York.

——— (2022) '*Stringr: Simple, Consistent Wrappers for Common String Operations*', <https://CRAN.R-project.org/package=stringr>.

Wickham, H. et al. (2019) 'Welcome to the Tidyverse', *Journal of Open Source Software*, 4: 1686.

Wickham, H. et al. (2023a) '*Dplyr: A Grammar of Data Manipulation*', <https://CRAN.R-project.org/package=dplyr>.

Wilke, C. O. (2020) '*Cowplot: Streamlined Plot Theme and Plot Annotations for "Ggplot2."*', <https://CRAN.R-project.org/package=cowplot>.

Worby C. J., Chaves S. S., Wallinga J., Lipsitch M., Finelli L. and Goldstein E. (2015) 'On the relative role of different age groups in influenza epidemics', *Epidemics*, 13: 10–16.

Wright, E. S. (2016) 'Using DECIPHER V2.0 To Analyze Big Biological Sequence Data in R', *The R Journal*, 8: 352–9.

Xue, K. S. et al. (2017) 'Parallel Evolution of Influenza across Multiple Spatiotemporal Scales', *eLife*, 6: 1–16.

Xue, K. S. et al. (2018) 'Within-Host Evolution of Human Influenza Virus', *Trends in Microbiology*, 26: 781–93.

Xue, K. S. and Bloom, J. D. (2019) 'Reconciling Disparate Estimates of Viral Genetic Diversity during Human Influenza Infections', *Nature Genetics*, 51: 1298–301.

Yang, H. et al. (2014) 'Structural Stability of Influenza A(H1N1)pdm09 Virus Hemagglutinins', *Journal of Virology*, 88: 4828–38.

Zhang, J. (2017) '*Phylotools: Phylogenetic Tools for Eco-Phylogenetics*', <https://CRAN.R-project.org/package=phylotools>.

Zhao, L. and Illingworth, C. J. R. (2019) 'Measurements of Intra-host Viral Diversity Require an Unbiased Diversity Metric', *Virus Evolution*, 5: vey041.

Zhu, X. et al. (2012) 'Influenza Virus Neuraminidases with Reduced Enzymatic Activity that Avidly Bind Sialic Acid Receptors', *Journal of Virology*, 86: 13371–83.

Virus Evolution, 2024, 10(1), veae034 , DOI: https://doi.org/10.1093/ve/veae034, Advance Access Publication 25 April 2024Research Article