

# Sequence amplification via cell passaging creates spurious signals of positive adaptation in influenza virus H3N2 hemagglutinin

Claire D. McWhite,<sup>1,2</sup> Austin G. Meyer,<sup>1,3,4</sup> and Claus O. Wilke<sup>1,3,4,\*</sup>

<sup>1</sup>Center for Systems and Synthetic Biology and Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX 78712, USA, <sup>2</sup>Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX 78712, USA, <sup>3</sup>Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, TX 78712, USA and <sup>4</sup>Department of Integrative Biology, The University of Texas at Austin, Austin, TX 78712, USA

\*Corresponding author: E-mail: wilke@austin.utexas.edu

## Abstract

Clinical influenza A virus isolates are frequently not sequenced directly. Instead, a majority of these isolates (~70% in 2015) are first subjected to passaging for amplification, most commonly in non-human cell culture. Here, we find that this passaging leaves distinct signals of adaptation, which can confound evolutionary analyses of the viral sequences. We find distinct patterns of adaptation to Madin–Darby (MDCK) and monkey cell culture absent from unpassaged hemagglutinin sequences. These patterns also dominate pooled datasets not separated by passaging type, and they increase in proportion to the number of passages performed. By contrast, MDCK–SIAT1 passaged sequences seem mostly (but not entirely) free of passaging adaptations. Contrary to previous studies, we find that using only internal branches of influenza virus phylogenetic trees is insufficient to correct for passaging artifacts. These artifacts can only be safely avoided by excluding passaged sequences entirely from subsequent analysis. We conclude that future influenza virus evolutionary analyses should appropriately control for potentially confounding effects of passaging adaptations.

**Key words:** hemagglutinin; influenza virus; egg passaging; cell passaging; positive adaptation.

## 1. Introduction

The routine sequencing of clinical isolates has become a critical component of global seasonal influenza virus surveillance (World Health Organization Global Influenza Surveillance Network, 2011). Analysis of these viral sequences informs the selection of future vaccine strains (Stöhr et al., 2012; WHO Writing Group et al., 2012), and a wide variety of computational methods have been developed to identify sites under selection or immune-escape mutations (Koelle et al., 2006; Nelson et al., 2006; Wolf et al., 2006; Blackburne et al., 2008; Suzuki, 2008), or to predict the short-term evolutionary future of influenza virus (Łuksza and Lässig, 2014; Neher et al., 2014). However, sites that

appear positively selected in sequence analysis frequently do not agree with sites identified experimentally in hemagglutination inhibition assays (Tusche, Steinbrück, and McHardy, 2012; Meyer and Wilke, 2015; Kratsch et al., 2016), and the origin of this discrepancy is unclear. Here, we argue that a major cause of this discrepancy is widespread passaging of influenza virus before sequencing.

Clinical isolates are often passaged in culture one or more times to amplify viral copy number, as well as to introduce virus into a living system for testing strain features such as vaccine response, antiviral response, and replication efficiency (World Health Organization Global Influenza Surveillance Network, 2011; Kumar and Henrickson, 2012). A variety of culture systems

**Table 1.** Parsing of passage-annotated FASTA sequences into passage history groups. For each passage group, we defined a regular expression that could reliably identify sequences with that passage history. Regular expressions were applied through the built-in python library 're'. SIAT1 and non-SIAT1 cell culture regular expressions were applied to the subset of sequences identified as generic cell culture sequences. The three middle columns list the number of sequences we identified for each passage group, for years 2014 only, 2015 only, and 2005–2015. The two furthest right columns list the number of single and multiply passaged sequences from 2005 to 2015 for each condition.

Passage group	Regular expression	Number of sequences			Rounds of passage	
		2014	2015	2005–2015	Single	Multiple
Chicken egg amniotes	AM[1–9][E[1–7]] AMNIOTIC EGG EX AM_[1–9]	6	0	79	1	79
Monkey cell culture	TMK RMK RHMK RII PMK R[1–9] RX	366	290	917	904	13
Generic cell culture	S[1–9] SX SIAT MDCK C[1–9] CX C_[1–9] M[1–9] MX X[1–9]]^X_\$	794	787	3,041	867	2,158
SIAT1	^S[1–9]_\$ ^SX_\$ SIAT2_SIAT1 SIAT3_SIAT1	389	626	1,046	459	587
Non-SIAT1 cell culture	not SIAT SX S[1–9]	297	56	1,755	408	1,331
Unpassaged	LUNG P0 OR_ ORIGINAL CLINICAL DIRECT	249	506	1,703	N/A	N/A
Pooled		1,508	1,601	6,873	1,772	2,317

are used for virus amplification. Cell cultures derived from Madin–Darby canine kidney (MDCK) cells are by far the most widely used system, with the majority of sequences in influenza repositories deriving from virus that has been passaged through an MDCK or modified MDCK cell culture (Balish et al., 2013; Bogner et al., 2006). Influenza virus may also be passaged through monkey kidney (RhMK or TMK) cell culture or injected directly into egg amniotes. Alternatively, complete influenza virus genomes can be obtained from PCR-amplified influenza samples without intermediate passaging (Katz, Wang, and Webster, 1990; Lee et al., 2013a).

Several experiments have demonstrated that influenza virus hemagglutinin (HA) accumulates mutations following rounds of passaging in both cell (Wyde et al., 1977; Ilyushina et al., 2012; Lee et al., 2013b) and egg culture (Robertson et al., 1993). The decreased number of mutations in MDCK-based cell culture is the main argument for use of this system over egg amniotes in vaccine production (Katz and Webster, 1989), with MDCK cells expressing human SIAT1 having the highest fidelity to the original sequence and reduced host adaptation (Hamamoto et al., 2013). Viral adaptations to eggs were recently linked to reduced vaccine efficacy (Skowronski et al., 2014; Xie et al., 2015) and were implicated as potentially contributing to reduced efficacy of 2014–2015 seasonal H3N2 influenza vaccination in the World Health Organization's recommendations for 2015–2016 vaccine strains (The World Health Organization, 2015). As the majority of influenza vaccines worldwide are produced in eggs, vaccine strain selection is limited to virus with the ability to replicate rapidly in this system (World Health Organization Global Influenza Surveillance Network, 2011).

Although egg-passaged sequences are increasingly excluded from influenza virus phylogenetic analysis (see, e.g., the NextFlu tracker; Neher and Bedford, 2015), due to the known high host-specific substitution rates, cell culture is generally not thought to be sufficiently selective to produce a discernible evolutionary signal. One of few existing evolutionary analyses of passaging effects on influenza virus (Bush et al., 2000) found that passaging caused no major changes in clade structure between egg and cell passaged viruses. However, several studies have recommended the use of internal branches in the phylogenetic tree to reduce passaging effects in evolutionary analysis of influenza A virus (Bush et al., 2001; Suzuki, 2006). Another study discovered egg culture to be the cause of misidentification of several sites under positive selection in influenza B virus (Gatherer, 2010), but this study was limited to

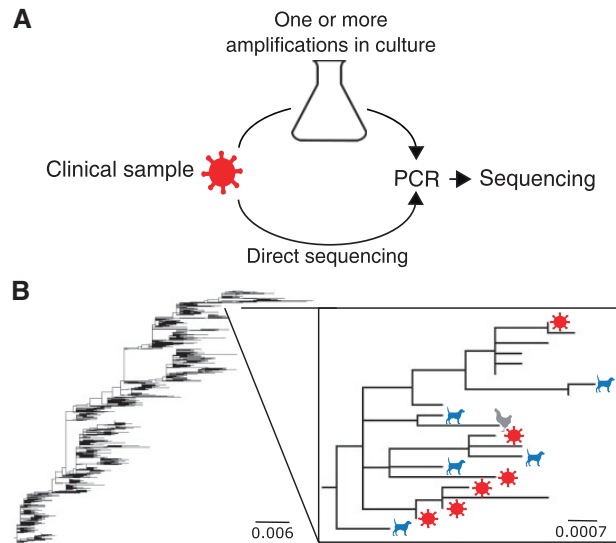
comparing egg-cultured to cell-cultured virus. As the availability of unpassaged influenza virus sequences has dramatically increased over the past 10 years, we can now perform a direct comparison of passaged to circulating virus.

Here, we compare patterns of adaptation in North American seasonal H3N2 influenza virus HA sequences derived from passaged and unpassaged virus. We divide viral sequences by their passaging history, distinguishing between unpassaged clinical samples, egg amniotes, RhMK (monkey) cell culture, and generic/MDCK-based cell culture. For the latter, we also distinguish between virus passaged in MDCK–SIAT1 cell culture (SIAT1) and in unmodified MDCK or unspecified cell culture (non-SIAT1). We find clear signals of adaptation to the various passaging conditions, and demonstrate that passaging artifacts become more severe with additional rounds of passaging. These signals are strongly present in the tip branches of the phylogenetic trees but can also be detected in internal and trunk branches. We demonstrate the accumulation of these passaging artifacts with additional rounds of serial passaging in non-SIAT1 cells. Finally, we demonstrate that the identification of antigenic escape sites from sequence data has been confounded by passaging adaptations, and that the exclusion of passaged sequences allows us to use sequence and structural data to highlight regions involved in antigenic escape.

## 2. Methods

### 2.1 Influenza sequence data

Non-laboratory strain H3N2 HA sequences collected in North America were downloaded from The Global Initiative for Sharing Avian Influenza Data (GISAID) (Bogner et al., 2006) for the 1968–2015 influenza seasons. We used exclusively North American sequences to reduce regional variation between influenza virus strains. Non-complete HA sequences were excluded. Sequences were trimmed to open reading frames, filtered to remove redundancies, and aligned by translation–alignment–back-translation using MAFFT (Katoh and Standley, 2013) for the alignment step. Sequence headers of FASTA files were standardized into an uppercase text format with non-alphanumeric characters replaced by underscores. As H3N2 strains have experienced no persistent insertion or deletion events, we deleted sequences that introduced gaps to the alignment. To ascertain overall data quality, we built a phylogenetic tree of the entire sequence set (using FastTree 2.0 compiled for short branch lengths; Price



**Figure 1.** Schematic of influenza A virus sequence collection and analysis. (A) Typical processing steps of influenza A virus clinical isolates. Virus collected from patients may be passaged a single time or multiple times prior to PCR amplification and sequencing in a variety of different environments (Ex. canine cell culture, monkey cell culture, egg amniotes). However, some clinical virus is not passaged and is sequenced directly. (B) Phylogenetic tree of H3N2 HA sequences from the 2005–2015 seasons. The inset shows a small clade of sequences from the 2006/2007 season, with colored dots representing sequences with passage annotations (red virion: unpassaged, blue dog: canine cell culture, gray hen: egg amniote, unlabeled: missing or unclear passage-history annotation).

et al., 2010) and checked for any abnormal clades or other unexpected tree features. We found one abnormal clade of approximately 20 sequences with an exceptionally long branch length ( $> 0.01$ ) and removed the sequences in that clade from further analysis. Our final dataset consisted of 6,873 sequences from 2005 to 2015 as well as one outgroup of 45 sequences from 1968 to 1977 (not considered for further analysis). We did not consider sequences collected from 1978 to 2004.

## 2.2 Identification of passage history and evolutionary-rate calculations

We divided sequences into groups by their passage-history annotation and collection year, determining passage history by parsing regular expressions using the built-in python package ‘re’ for keywords in FASTA headers (Table 1). We classified 1,133 sequences with indeterminate or missing passage histories, or passage through multiple categories of hosts (i.e. both egg and cell), as ‘other’. The final datasets for individual passage groups contained between 79 and 3,041 sequences (Table 1).

We additionally divided passage groups into singly and serially passaged subgroups. Sequences matching the regular expression ‘2|3|4|5|6|1\_C|X\_|X\_C|AND\_MV1|X\_S|1\_S|CX|MX|C1S1|EX|X\_E|MIX\_RHMK|RII’ were classified as having been passaged two or more times. All remaining sequences were passaged only once.

To determine the number of times a sequence was passaged, we used a different set of regular expressions, which we applied only to non-SIAT1 cell-passaged sequences. We first excluded sequences with an indeterminate number of passages by excluding sequences whose record IDs matched the regular expression ‘^X\_\$|DETAILS\_MDCK|MX\_C|^X\_C1|^MX\_\$|CX\_C1|X\_C1|DETAILS\_ND’. We then collected multiply passaged

sequences using the regular expression ‘3|1\_C2|2\_C1|M1M1\_C1|1\_MDCK2|4|2\_C2|3\_C1|1\_C3|5|2\_C3|3\_C2’ and doubly passaged sequences using the regular expression ‘2|1\_C1’. The remaining sequences were only passaged a single time.

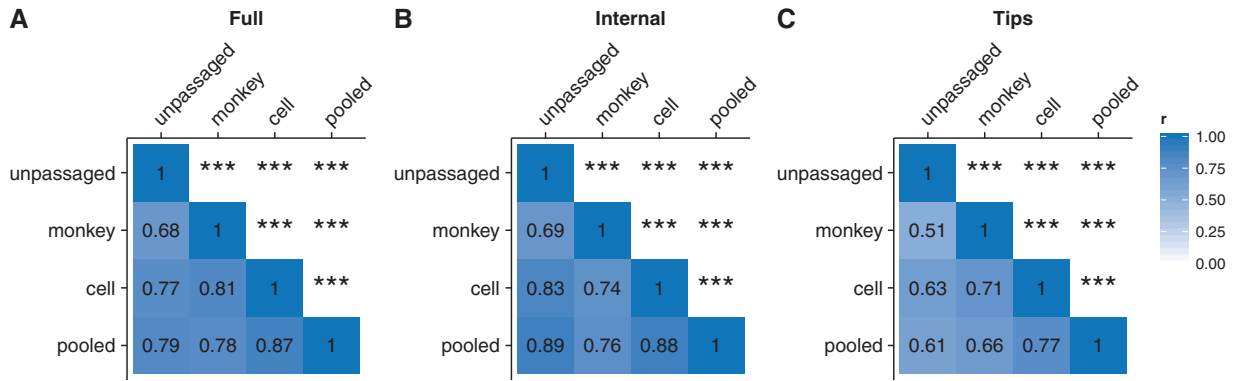
We next constructed phylogenetic trees for each passage group as well as one tree for a pooled dataset combining all individual passage groups and other sequences. All phylogenetic trees were constructed using FastTree 2.0 (Price et al., 2010). We calculated site-specific  $dN/dS$  values using a one-rate Single-Likelihood Ancestor Counting (SLAC) model implemented in HyPhy (Pond et al., 2005). One rate models fit a site-specific  $dN$  and a global  $dS$ , where the global  $dS$  is the mean site-wise  $dS$  for a given condition (Spielman et al., 2016). Among different one-rate, site-specific models, SLAC performs nearly identically to other approaches (Spielman et al., 2016), and it was chosen here due to its speed and ease of extracting  $dN/dS$  estimates along internal and tip branches. To obtain internal and tip branch-specific estimates, we extracted the  $dN/dS$  values calculated by the SLAC algorithm. We manually calculated  $dN/dS$  along trunk branches by counting the number of synonymous and non-synonymous substitutions at each trunk site. We defined the trunk as the sequence of branches from the root to the penultimate node before a randomly chosen terminal sequence from the most recent year represented in the tree.

We chose sequences from 2005 to 2015 as our sample set due to the low number of available sequences prior to this period. As  $dN/dS$  estimates can be confounded by sample size (Spielman et al., 2016), we sought to limit this effect by down-sampling each experimental set to match the number of sequences in the smallest group being considered in a particular analysis (Table 1). To reduce season-to-season variation in the comparison of unpassaged, SIAT1, and non-SIAT1 cell culture, we performed one analysis with sequences from only 2014, which is the year that maximizes sequences available from all three conditions ( $n = 249$  each).

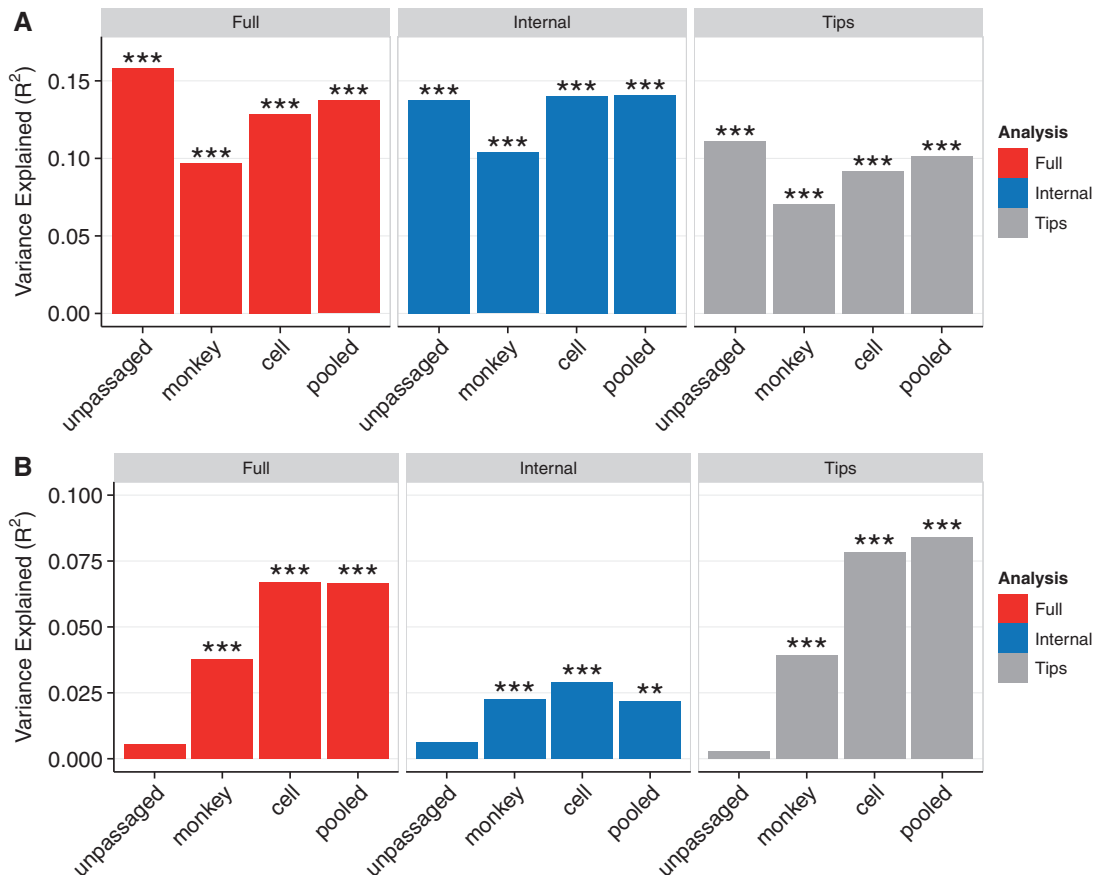
## 2.3 Geometric analysis of $dN/dS$ distributions

For each amino acid site  $i$  in HA, we computed the inverse Euclidean distance to each amino acid site  $j$  ( $j \neq i$ ) in the 3D crystal structure. For each site  $i$ , we then correlated the list of inverse distances to sites  $j$  with site-wise  $dN/dS$  values at sites  $j$ . This procedure resulted in a correlation coefficient for each site  $i$ , and we mapped these correlation coefficients onto the corresponding sites  $i$  in the 3D structure model of HA. In this analysis, sites spatially closest to positively selected regions in the protein yielded the highest correlation coefficients. Thus, this approach allowed us to visualize regions of increased positive selection. As the correlation coefficient for site  $i = 224$  is consistently highest for sets of sequences undivided by passage history, we chose this site as a reference to compare passage conditions. See Meyer and Wilke (2015) for additional discussion of this approach.

We processed the HA PDB structure to allow for easy alignment with site-wise measures as discussed previously (Meyer and Wilke, 2015). We provided a renumbered and formatted H3N2 HA structure derived from PDBID:2YP7 (2YP7clean.pdb) (Lin et al., 2012) with our data analysis code (see below). Noting that the HA protein and gene numbering is offset by 16, all site numbering in this article refers to the protein site position. The alignment of gene and protein numbering schemes to amino acid sequence is available in each Supplementary Data File.



**Figure 2.** Comparison of site-wise  $dN/dS$  values among sequences with differing passage histories. Pearson correlations between site-wise  $dN/dS$  values for HA sequences derived from passaged and unpassaged influenza virus collected between 2005 and 2015 (downsampled to  $n=917$  in all groups). Correlations were calculated separately for  $dN/dS$  estimated from complete trees (A), internal branches only (B), and tip branches only (C). Asterisks denote significance of correlations ( $0.01 \leq p < 0.05$ ,  $0.001 \leq p < 0.01$ ,  $***p < 0.001$ ). Data used to generate this figure are available in [Supplementary Data 1](#).

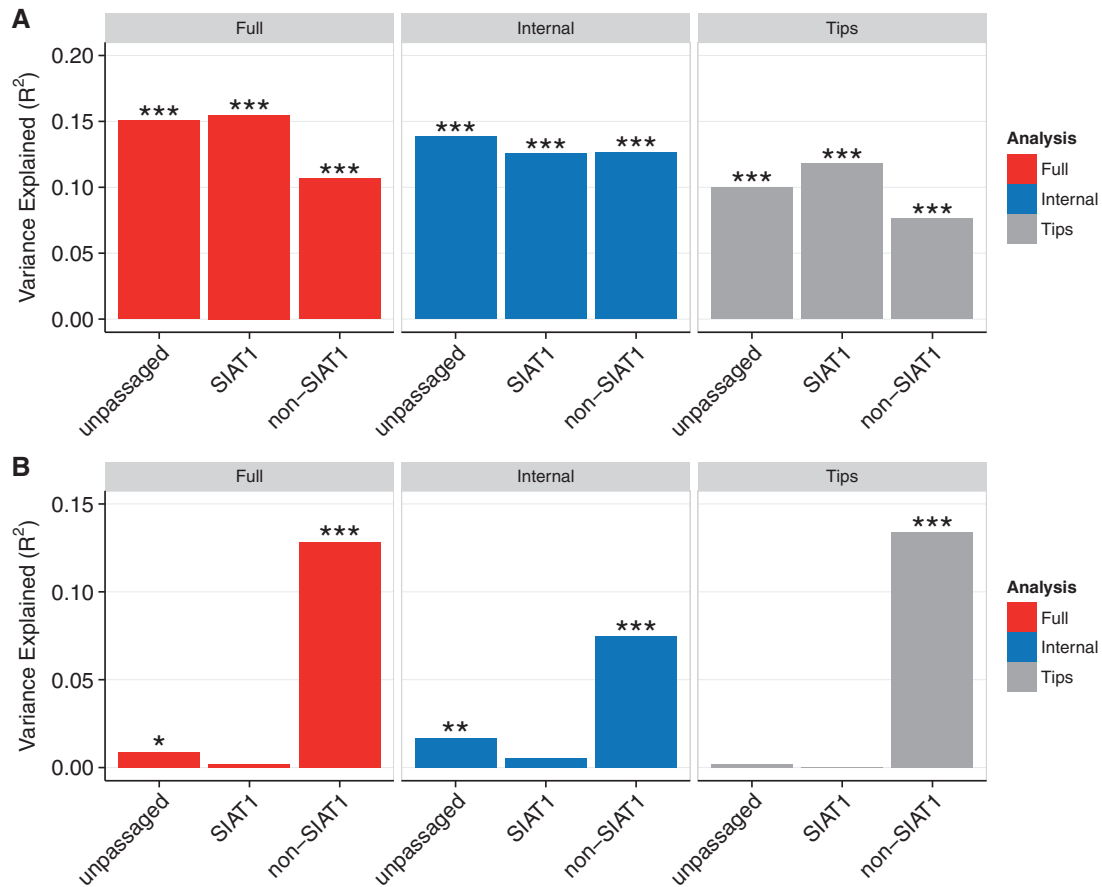


**Figure 3.** Percent variance in  $dN/dS$  explained by relative solvent accessibility (A) and by inverse distance to protein site 224 (B). (A) Relative solvent accessibility (RSA) explains ~10–16% of the variation in  $dN/dS$  for all sequences. (B) Inverse distance to site 224 explains ~7% of the variation in  $dN/dS$  for cell-passaged sequences and for all sequences (pooled), however it explains virtually no variation for unpassaged sequences. Asterisks denote significance of correlations ( $0.01 \leq p < 0.05$ ,  $0.001 \leq p < 0.01$ ,  $***p < 0.001$ ). Data used to generate this figure are available in [Supplementary Data 1](#).

#### 2.4 Local Branching Index analysis

We used the framework and code (<https://github.com/mneher/FitnessInference>) from [Neher et al. \(2014\)](#) to rank sequences according to their Local Branching Index (LBI), a metric that uses branching density to predict progenitor lineages. To build our sample set we divided sequences by year and passage history.

We then down-sampled alignments to 70% of the available sequences, up to a maximum of 100 sequences. We repeated each down-sampling 50 times for each condition. We then ranked sequences in each sample according to the LBI algorithm (script `rank_sequences.py` available from <https://github.com/mneher/FitnessInference>) and calculated the Hamming distance of the



**Figure 4.** Virus passaged in non-SIAT1 cells carries unique adaptations not present in unpassaged or SIAT1-passaged virus. (A) The correlation between  $dN/dS$  and RSA is weakened for virus passaged in non-SIAT1 cells. (B) The correlation between  $dN/dS$  and inverse distance to site 224, representing a positive-selection hotspot in the vicinity of that site, is only present in virus passaged in non-SIAT1 cells. Asterisks denote significance levels ( $0.01 \leq p < 0.05$ ,  $0.001 \leq p < 0.01$ ,  $p < 0.001$ ). Sequences analyzed were collected between 2005 and 2015. Alignments were randomly down-sampled to yield identical numbers of sequences in each alignment ( $n = 1046$ ). Data used to generate this figure are available in [Supplementary Data 4](#).

top ranked sequence from each condition to the ancestrally reconstructed root sequence of the following year's unpassaged and pooled trees. The Hamming distance derived from the top ranked sequence was divided by the Hamming distance of a randomly chosen sequence from the same condition, to assess if a predicted progenitor was better than a randomly chosen sequence. These ratios were averaged over all possible choices of the randomly chosen sequence and over the 50 trials, to yield the mean ratio score for a particular year and passage condition. A mean ratio score  $< 1$  indicates that the LBI algorithm performs better than random chance.

### 2.5 Statistical analysis and data availability

Raw influenza sequences used in this analysis are available for download from GISAID (<http://gisaid.org>) using the parameters 'North America', 'H3N2', and '1976–2015'. Acknowledgements for sequences used in this study are available in [Supplementary File 1](#). The complete, processed dataset used in our statistical analysis is available in [Supplementary Data 10](#), including protein and gene numbering, computed evolutionary rates, relative solvent accessibility (RSA) for the HA trimer, and site-wise distance to protein site 224. RSA of the HA trimer was taken from [Meyer and Wilke \(2015\)](#). Site-wise Euclidean distances between all amino acids in the HA structure PDBID:2YP7 were recalculated from structural coordinates using the python script `distances.py`

([Supplementary File 2](#)). Statistical analysis was performed using R ([Ihaka and Gentleman, 1996](#)), and all graph figures were drawn with the R package `ggplot2` ([Wickham, 2009](#)). Throughout this work, \* denotes a significance of  $0.01 \leq p < 0.05$ , \*\* denotes a significance of  $0.001 \leq p < 0.01$ , and \*\*\* denotes a significance of  $p < 0.001$ .

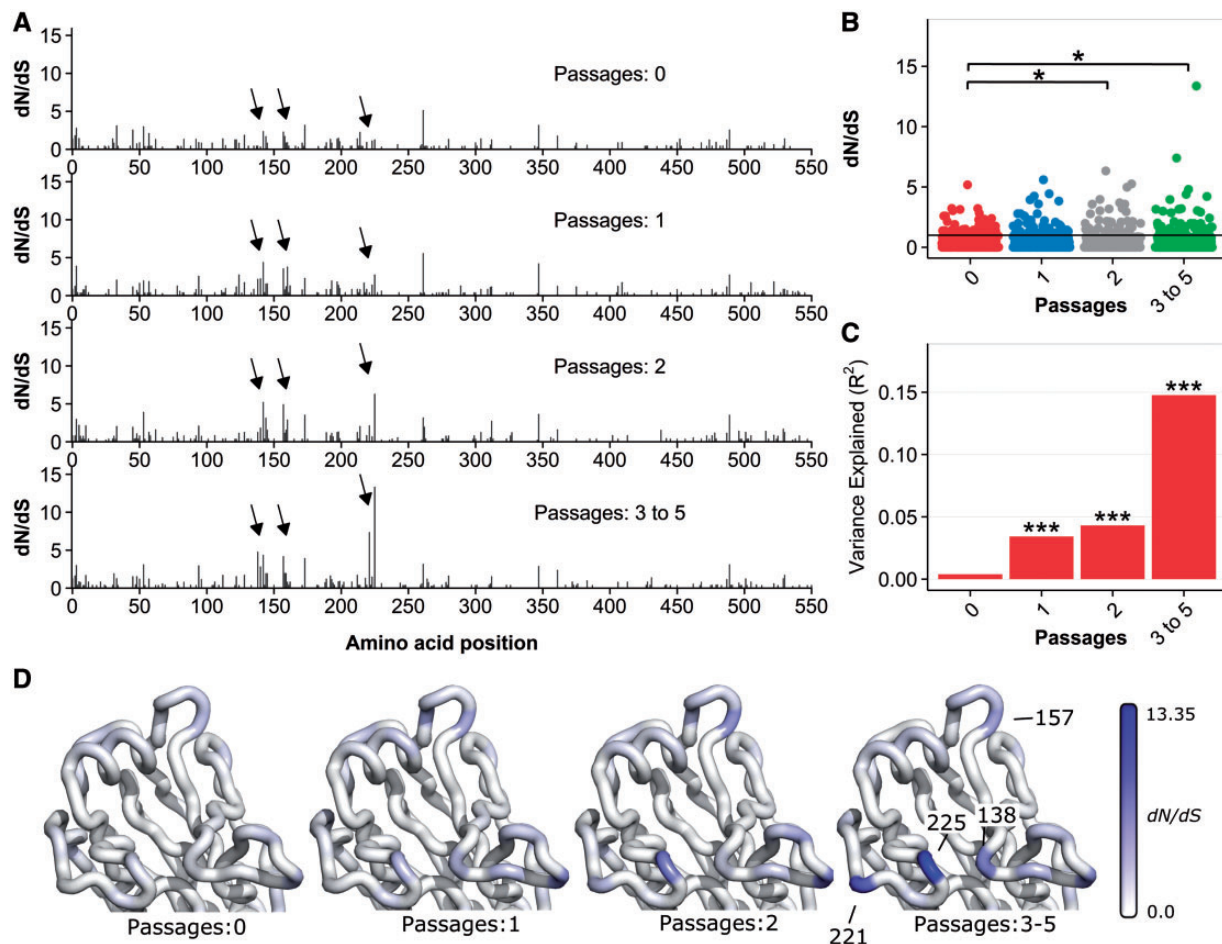
Linear models between site-wise  $dN/dS$  and RSA or inverse distance were fit using the `lm()` function in R. Correlations were calculated using the R function `cor()` and significance determined using `cor.test()`.

Our entire analysis pipeline, scripts, instructions for running analyses, and raw data (except initial sequence data per the GISAID user agreement) are available at the following Github project repository:

[https://github.com/wikelab/influenza\\_H3N2\\_passaging](https://github.com/wikelab/influenza_H3N2_passaging).

### 3. Results

Many influenza virus samples collected from patients are first passaged through one or more culturing systems ([Table 1](#)) prior to PCR amplification and sequencing ([Fig. 1A](#)). Samples may be passaged either once or serially ([Table 1](#)), even though a single passage is generally sufficient to obtain adequate amounts of genetic material for sequencing. Reconstructed trees of influenza virus evolution contain a mixture of passage histories at



**Figure 5.** Accumulation of passing artifacts with increasing numbers of serial passages in non-SIAT1 cell culture. (A) Site-wise  $dN/dS$  values for virus which was not passaged, passaged once, passaged twice, and passaged 3-5 times in non-SIAT1 cell culture ( $n = 304$  for each group). Arrows highlight regions of increased  $dN/dS$  in passaged virus. Notably,  $dN/dS$  inflation is increased with increasing rounds of passaging. (B)  $dN/dS$  values vs. number of passages.  $dN/dS$  values are significantly elevated after two or more passages in non-SIAT1 cell culture, relative to unpassaged virus (paired t test). Asterisks denote significance levels ( $*0.01 \leq p < 0.05$ ,  $**0.001 \leq p < 0.01$ ,  $***p < 0.001$ ). (C) The correlation between  $dN/dS$  and inverse distance to site 224 increases with the number of passages. (D) Mapping  $dN/dS$  values onto the hemagglutinin head structure demonstrates the accumulation of passage adaptations with increasing rounds of passages. Labeled sites correspond to regions denoted with arrows in (A). Data used to generate this figure are available in [Supplementary Data 6](#).

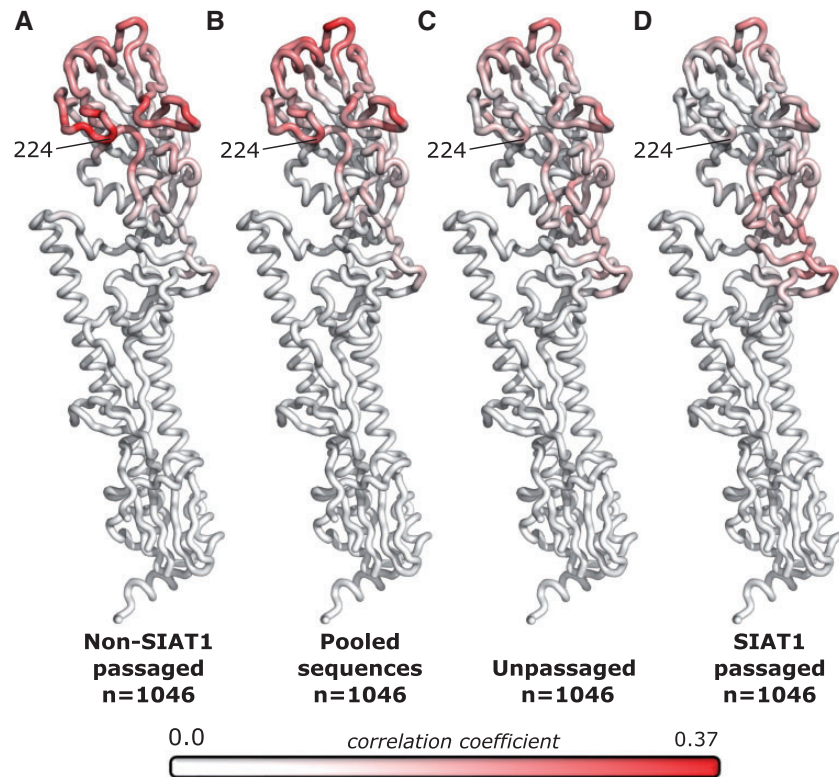
their tips (Fig. 1B). During passaging, influenza virus genomes accumulate adaptive mutations, and the effect of these mutations on evolutionary analyses of influenza virus sequences is not well understood.

### 3.1 Site-wise evolutionary rate patterns differ between passage groups

To quantify any evolutionary signal that may be introduced by passaging, we assembled, from the GISAID database (Bogner et al., 2006), a set of North American human influenza virus H3N2 HA sequences collected between 2005 and 2015. We initially sorted these sequences into groups by their passage history: (1) unpassaged, (2) egg-passaged, (3) generic cell-passaged, and (4) monkey cell-passaged (Table 1). To assess evolutionary variation at individual sites, we calculated site-specific  $dN/dS$  (Echave et al., 2016), using SLAC. Specifically, we calculated one-rate  $dN/dS$  estimates, i.e., site-specific  $dN$  values normalized by a global  $dS$  value (see 'Methods' section for details). In addition to considering groups of sequences with specific passage histories, we also calculated  $dN/dS$  values by pooling all sequences into one combined analysis. This pooled group corresponds to a

typical influenza virus evolutionary analysis for which passage history has not been accounted.

We first correlated the site-wise  $dN/dS$  values we obtained for virus sequences derived from different passage histories. If passage history did not matter, then the  $dN/dS$  values obtained from different sources should have correlated strongly with each other, with  $r$  approaching 1. Instead, we found that correlation coefficients ranged from 0.68 to 0.87, depending on which specific comparison we made (Fig. 2A). In this analysis, and throughout this work, we down-sampled alignments to the smallest number of sequences available for any of the conditions compared, to keep the samples as comparable as possible overall. The analysis of Fig. 2 used  $n = 917$  randomly drawn sequences for each condition. Unpassaged  $dN/dS$  correlated more strongly with cell and pooled  $dN/dS$  (correlations of 0.77 and 0.79, respectively) than with monkey-cell  $dN/dS$  (0.68). Note that the  $dN/dS$  values from the pooled group, which corresponds to a typical dataset used in a phylogenetic analysis of influenza virus, more closely correlated with the  $dN/dS$  values from the generic cell group ( $r = 0.87$ ) than from the unpassaged group ( $r = 0.79$ ). Egg-derived sequences were excluded from this



**Figure 6.** Correlations of  $dN/dS$  with inverse distances, mapped onto the hemagglutinin structure for non-SIAT1-passaged, pooled, unpassed, and SIAT1 passaged sequences. The correlation between  $dN/dS$  and inverse distance for each reference site was mapped onto the hemagglutinin structure for (A) non-SIAT1 sequences, (B) pooled sequences, (C) unpassed sequences, and (D) SIAT1 passaged sequences. Sequences analyzed were collected between 2005 and 2015. Alignments were randomly down-sampled to yield identical numbers of sequences in each alignment ( $n = 1046$ ). Red coloring represents positive correlations, while white represents zero or negative correlations. The four conditions group into two distinct correlation patterns, non-SIAT1/pooled and unpassed/SIAT1. In particular, the loop containing site 224 lights up strongly for non-SIAT1 and pooled sequences but not for unpassed and SIAT1 sequences. Data used to generate this figure are available in [Supplementary Data 7](#).

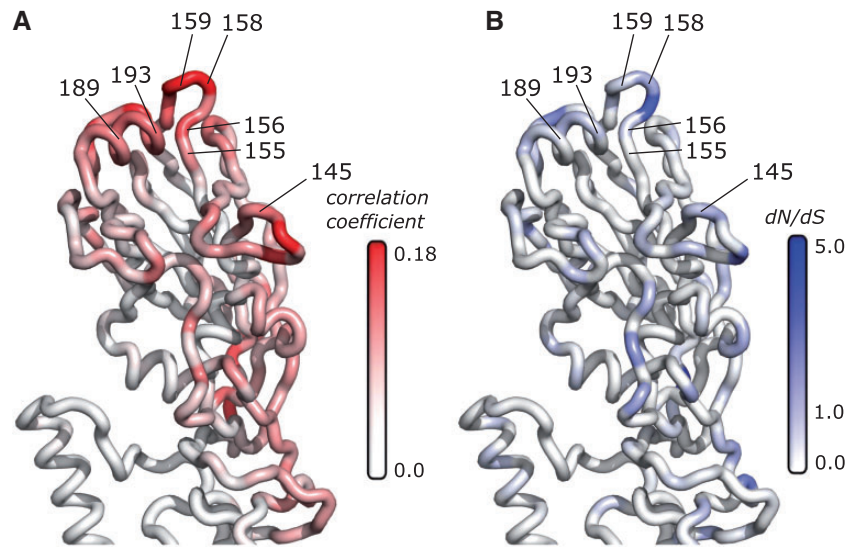
analysis due to low sequence numbers ( $n = 79$ ), however evolutionary rates from this condition correlated particularly poorly with those of random draws of 79 unpassed sequences ([Supplementary Fig. 1](#)). This result is consistent with previous conclusions ([Bush et al., 2000](#); [Suzuki, 2006](#); [Gatherer, 2010](#)) that egg-derived sequences show specific adaptations not found otherwise in influenza virus sequences.

Because the common ancestor of any two passaged influenza viruses is a virus that replicated in humans, we expected that any adaptations introduced during passaging would not extend into the internal branches of a reconstructed tree. Therefore, we additionally subdivided phylogenetic trees into internal branches and tip branches, and calculated site-specific  $dN/dS$  values separately for these two sets of branches. In fact, [Bush et al. \(2000\)](#) recommended the use of internal branches to reduce variation seen between egg and cell culture-passaged virus. As expected, we found that when  $dN/dS$  calculations were restricted to the internal branches, the correlations between the passage groups increased overall ([Fig. 2B](#)), even though distinct differences between the passage groups remained. Conversely, when we only considered tip branches, correlations among most groups were relatively low ([Fig. 2C](#)), with the exception of cell-passaged sequences compared to the pooled sequences. This finding emphasizes once again that the pooled sample is most similar to the cell-passaged sample. We conclude that different passaging histories leave distinct, evolutionary signatures of adaptation to the passaging environment.

In aggregate, these results show that both generic-cell-passaged sequences and monkey-cell-passaged sequences yield different site-wise  $dN/dS$  patterns relative to unpassed sequences ([Fig. 2A–C](#)), with  $dN/dS$  values derived from monkey-cell-passaged sequences being the least similar to  $dN/dS$  from unpassed sequences ([Fig. 2A–C](#)). The pooled group of sequences, which corresponds to a typical dataset used in evolutionary analyses of influenza virus, describes evolutionary rates of specifically cell-passaged virus and poorly matches evolutionary rates of unpassed virus.

### 3.2 Adaptations to cell and monkey-cell passage display characteristic patterns of site variation

We next asked whether adaptations to passage history were located in specific regions of the HA protein. To address this question, we employed the geometric model of HA evolution we recently introduced ([Meyer and Wilke, 2015](#)), where structural measurements explain variation in  $dN/dS$ . For H3N2HA, this model explains over 30% of the variation in  $dN/dS$  using two simple physical measures, the RSA of individual residues in the structure ([Tien et al., 2013](#)) and the inverse linear distance in 3D space from each residue to protein site 224 in the HA monomer. Notably, the geometric model was previously applied to a pooled sequence set including sequences of various passaging histories. To what extent it carries over to sequences with specific passaging histories is not known.



**Figure 7.** Unpassed sequences allow recovery of antigenic regions from positive-selection analysis. For each site, the correlation between  $dN/dS$  and inverse distance (A) or  $dN/dS$  directly (B) were mapped onto the hemagglutinin structure, for  $dN/dS$  derived from unpassed sequences collected between 2005 and 2015 ( $n = 1703$ ). Red coloring represents higher correlation; blue coloring represents higher  $dN/dS$ . Highlighted regions contain residues (labeled with protein site number) which experimentally determined to cause antigenic change by Koel et al. (2013). Correlations and  $dN/dS$  for antigenic residues are given in Table 2. Data used to generate this figure are available in Supplementary Data 7.

**Table 2.** Evolutionary rates and inverse distance correlations of residues responsible for antigenic change. For each site, we determined  $dN/dS$  and the correlation between  $dN/dS$  and inverse distance for unpassed sequences collected between 2005 and 2015 ( $n = 1703$ ). 5/7 residues linked to antigenic changes have inverse-distance correlations above the 90th percentile, while only 1/7 have  $dN/dS$  values above the 90th percentile. Sites were experimentally determined by Koel et al. (2013).

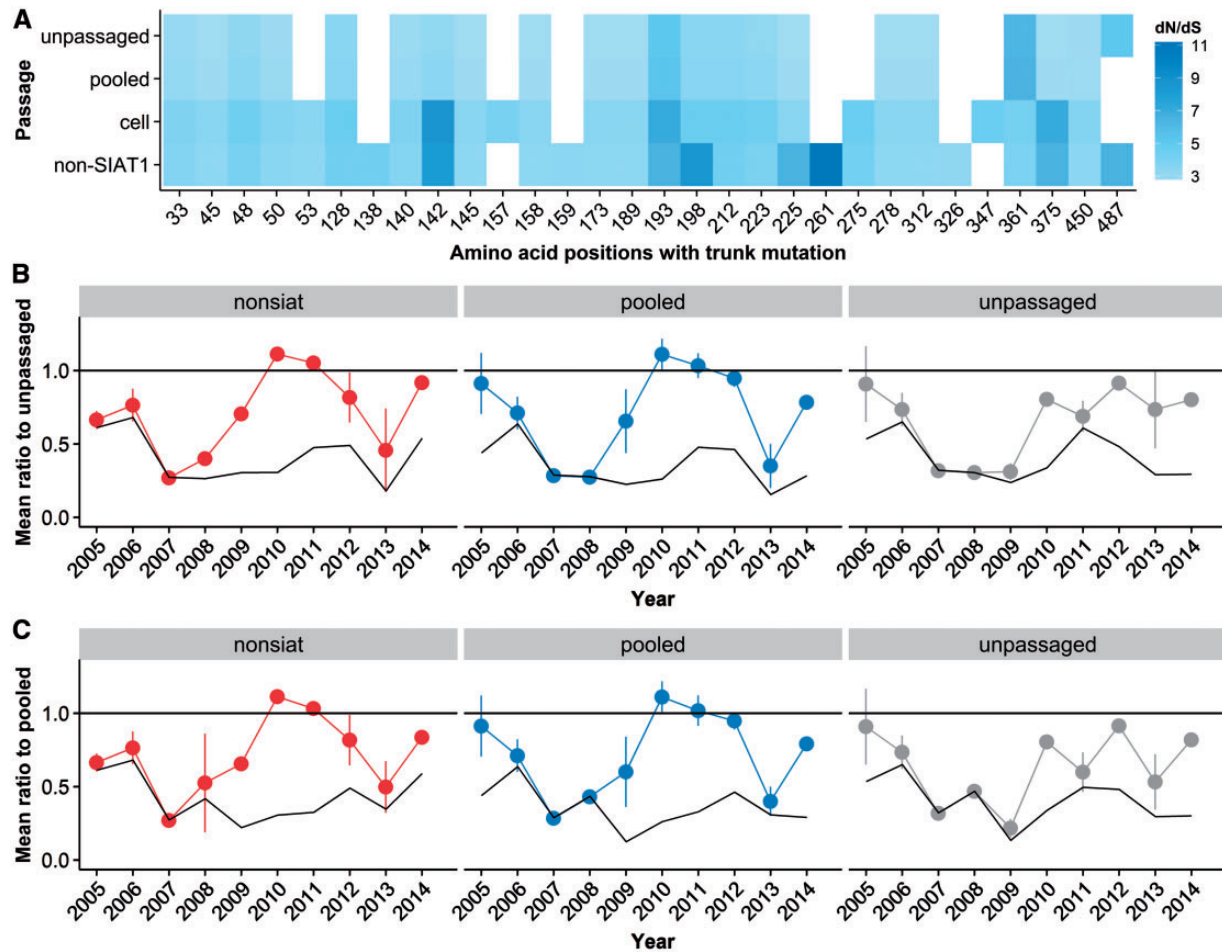
Site		Raw $dN/dS$		Inv.-dist. correlation	
Gene	Protein	$dN/dS$	Percentile	R	Percentile
161	145	0.672	0.823	0.082	0.883
171	155	0	0	0.077	0.867
172	156	0.672	0.832	0.1317	0.971
174	158	1.36	0.958	0.1797	0.996
175	159	0.49	0.75	0.1837	1
205	189	0.474	0.763	0.0887	0.905
209	193	0.672	0.845	0.098	0.936

We first considered the correlation between  $dN/dS$  and RSA (Fig. 3A). We found that for all passage groups,  $R^2$  values ranged from 0.10 to 0.16 in the full tree, consistent with our earlier work (Meyer and Wilke, 2015). The high congruence among  $R^2$  values for internal branches and all branches suggests that RSA imposes a pervasive selection pressure on HA, independent of passaging adaptations. Thus, RSA represents a useful structural measure of a persistent effect of  $dN/dS$  with stronger correlations in the full tree and internal branches than in tip branches.

Next we considered the correlation between  $dN/dS$  and the inverse distance to site 224 to each site in the HA structure (Fig. 3B). In contrast to RSA, correlations here were systematically higher in tip branches, suggesting a recent adaptive signal. We found virtually no correlation for unpassed sequences, while a low correlation existed for monkey-cell cultured sequences and a higher correlation for cell-passaged and pooled sequences. To confirm that differences between pooled and unpassed correlations were not simply due to variation from random sampling, we created a null distribution of correlations from 200 random

draws of pooled sequences. The correlation for unpassed sequences was significantly lower than it was for pooled sequences ( $z = -4.22$ ,  $p = 1.2 \times 10^{-5}$ , Supplementary Fig. 2). Correlations from pooled sequences mirrored cell-culture correlations and persisted through internal branches. Thus, the correlation of  $dN/dS$  with the inverse distance to site 224 seems to be primarily an artifact of cell passage, even though its effect can be seen along internal branches as well. This cell-specific signal dominates the pooled dataset. Further, this cell-specific signal is partially attenuated along internal branches and amplified along tip branches, as we would expect from a signal caused by recent host-specific adaptation. Even though this signal is a true predictor of influenza virus evolutionary rates for virus grown in cell culture, it does not transfer to unpassed sequences and therefore has no relevance for the circulating virus. This finding serves as a strong demonstration of passage history as a confounder in analysis of HA evolution, not just for egg passage as previously demonstrated, but also for cell and monkey-cell passage.





**Figure 8.** Passage artifacts affect trunk  $dN/dS$  and topology-based predictions. (A) Site-wise trunk  $dN/dS$  values for passage groups ( $n = 1703$ ). Only sites with at least one non-synonymous mutation along the trunk were included. Many sites appear under positive selection only in the trunk reconstructed from passed sequences, e.g. 159, 261, 275. (B and C) Prediction of future dominant clade by Local Branching Index (LBI) depends on passing history. LBI was calculated for trees derived from non-SIAT1, pooled, and unpassed sequences using a maximum of 100 sequences per condition. The mean ratio estimates the quality of the prediction (lower is better), and values  $< 1$  indicate the prediction performs better than random. Error bars indicate the standard deviation of the mean, estimated from resampling (see ‘Methods’ section). The solid black lines represent the best possible prediction in each year. Predictions were evaluated relative to the following year’s ancestrally reconstructed root sequence obtained from (B) unpassed and (C) pooled sequences. Non-SIAT1, pooled, and unpassed sequences have divergent prediction quality across years, and on an average, unpassed sequences seem to perform better than passed or pooled sequences. (The distances between dots and the solid black lines are, on an average, the smallest for predictions derived from unpassed sequences.) Data used to generate this figure are available in [Supplementary Datasets 8A and 9B and C](#).

Surprisingly, the correlation we found here between  $dN/dS$  and inverse distance to site 224 for pooled sequences ( $R^2 = 0.067$ ) was less than half of the value previously reported (Meyer and Wilke, 2015) (Fig. 3B). However, using a dataset of sequences more temporally matched to the previously published analysis (2005–2014 instead of 2005–2015), we recovered the earlier higher correlation. This finding suggests that there is some feature in the additional 2015 sequences that changes the pooled data’s relationship with inverse distance to site 224. In 2015, unpassed and SIAT1 sequences each doubled in number compared to 2014, while the number of non-SIAT1 cell cultured sequences dropped dramatically (Table 1). SIAT1, an MDCK cell line which overexpresses human-like 6-linked sialic acids over native 3-linked sialic acids (Matrosovich et al., 2003) has higher sequence fidelity than unmodified MDCK (Hamamoto et al., 2013). Therefore, we next investigated whether the drop in correlation from 2014 to 2015 could be attributed to the recent reduction in cell culture using non-SIAT1 cells.

### 3.3 Adaptation to passage in SIAT1 cells is weak or absent

In the preceding analyses, we lumped all cell cultures except monkey cells into the same category. However, there are more subtle distinctions in cell passing systems, and they can exert differential selective pressures on human adapted virus (Oh et al., 2008; Hamamoto et al., 2013). As our generic cell culture group was composed of a mixture of wild type MDCK, SIAT1, and unspecified cell cultures, we next investigated whether any one culture type was the source of the high cell-culture signal seen in Fig. 3B.

SIAT1 is currently the dominant system for passaging of influenza virus in North America, with approximately half of the 2015 influenza virus sequences currently available from GISAID deriving from serial passaging through SIAT1 cells. Experimental analysis of SIAT1 demonstrates improved sequence fidelity and reduced positive selection over unmodified MDCK cell culture (Oh et al., 2008; Hamamoto et al., 2013). We

sought to determine if the apparently cell-culture-specific correlation of site-wise evolutionary rates and inverse distance to site 224 extended to SIAT1 cell culture. To compare cell-culture varieties, we created sample-size matched groups of non-SIAT1 cell culture, SIAT1 cell culture, and unpassaged sequences collected between 2005 and 2015 ( $n=1,046$ ), excluding sequences that had been passaged through both a non-SIAT1 and a SIAT1 cell culture.

All groups showed similar correlations between  $dN/dS$  and RSA, regardless of whether  $dN/dS$  was calculated for the entire tree, for internal branches only, or for tip branches only (Fig. 4A). By contrast, inverse distance to site 224 uniquely correlated with  $dN/dS$  from non-SIAT1-cultured virus (Fig. 4B). This effect was strongest along tip branches ( $R^2=0.139$ ), but it was almost as strong along the entire tree ( $R^2=0.129$ ). The correlation was reduced, though still significant, among internal branches ( $R^2=0.075$ ). Thus, we conclude that the correlation between  $dN/dS$  and the inverse distance to site 224 represents a unique signal of adaptation to passaging in non-SIAT1 cells. In other words, a non-SIAT1-specific signal can completely dominate all signals of positive adaptation when a dataset contains a sufficiently high number of sequences passaged in non-SIAT1 cells. In our analysis (Fig. 3B), the high correlation of non-SIAT1 cell  $dN/dS$  with inverse distance to site 224 is suppressed in the pooled condition because the number of unpassaged and SIAT1-passaged sequences grew substantially in 2015. This difference in sample composition explains the lower than expected correlations in Fig. 3B for pooled  $dN/dS$ .

As these three conditions were somewhat temporally separated (most non-SIAT1 cell culture sequences were pre-2015, and most unpassaged and SIAT1 culture sequences were post-2014), we controlled for season-to-season variation by drawing 249 sequences from each group from 2014. We again considered site-wise  $dN/dS$  correlations among passaging groups, and we found that overall, unpassaged and SIAT1-passaged sequences appeared the most similar (Supplementary Fig. 3A–C).

### 3.4 Signals of passaging adaptation accumulate with additional rounds of passaging in non-SIAT1 cells

Having identified non-SIAT1 cell culture as the source of the contaminating signal in analyses of inverse distance to site 224, we next investigated the source of this signal at the single amino acid level. We expected that a signal of adaptation to a passaging system would strengthen with additional exposure to that system. Thus, we compared the magnitude of the site-wise  $dN/dS$  values in sequences that had never been passaged, had been passaged once, passaged twice, or passaged 3–5 times (Fig. 5A). For this analysis, we only considered passage in non-SIAT1 cells. This analysis revealed distinct regions of increasing positive selection along the HA molecule (arrows in Fig. 5A) and a strong relationship between the magnitude of these signals and the number of times influenza viruses were passaged. Further, we found an overall increase in  $dN/dS$  with increased numbers of passages in non-SIAT1 cells (Fig. 5B). The strongly selected sites 221 and 225 are adjacent to site 224, explaining the specific relationship between  $dN/dS$  calculated from non-SIAT1 sequences and the inverse distance in 3D space to this site. The correlation between  $dN/dS$  and inverse distance increased in strength with increasing numbers of passages (Fig. 5C), even though it was observable after a single passage in non-SIAT1 cells. Mapping the raw  $dN/dS$  values onto the HA structure showed how specific sites light up as passage numbers increase (Fig. 5D).

### 3.5 Evolutionary variation in sequences from unpassaged virus predicts regions involved in antigenic escape

Our preceding analyses might suggest that the inverse distance metric for describing regions of selection only captures effects of adaptation to non-SIAT1 cell culture. However, this is not necessarily the case. Importantly, inverse distance needs to be calculated relative to a specific reference point. Site 224 was previously used as the reference point because it yielded the highest correlation for the dataset analyzed (Meyer and Wilke, 2015). For a different dataset, one that doesn't carry the signal of adaptation to non-SIAT1 cell culture, a different reference point may be more appropriate.

We thus repeated the inverse distance analysis of Meyer and Wilke (2015) for a size-matched sample of 1,046 sequences from non-SIAT1, pooled, SIAT1, and unpassaged virus collected between 2005 and 2015 (Fig. 6). In brief, for each possible reference site in the HA structure, we measured the inverse distance in 3D space from that site to every other site in the structure (see 'Methods' section for details). We then correlated the inverse distances with the  $dN/dS$  values at each site, resulting in one correlation coefficient per reference site. Finally, we mapped these correlation coefficients onto the HA structure, coloring each reference site by its associated correlation coefficient. If inverse distances measured from a particular reference amino acid have higher correlation with the site-wise  $dN/dS$  values, then this reference site will appear highlighted on the structure.

For non-SIAT1-passaged and pooled virus, this analysis recovered the finding of Meyer and Wilke (2015) that the loop containing site 224 appeared strongly highlighted (Fig. 6A). However, this signal was entirely absent in unpassaged and SIAT1 passaged virus (Fig. 6B), with no sites in that loop working well as a reference point. These results suggested that this loop was specifically involved in adaptation of HA to non-SIAT1 cell culture, explaining the non-SIAT1-specific signal shown in Fig. 4A. Globally, the pattern of correlations from pooled sequences strongly resembled the non-SIAT1 pattern, in contrast to the resemblance of SIAT1 to unpassaged. Thus, the inverse distance metric is useful for differentiating regions of selection particular to different experimental groups.

Therefore, we next asked what residual patterns of positive selection remained once the adaptation to non-SIAT1 cells was removed. Even though site-wise correlations are relatively low for unpassaged virus compared to the ones observed for non-SIAT1-passaged virus, we could still recover relevant patterns of HA adaptation after rescaling our coloring. In particular, we found that sites opposite to the loop-containing site 224 lit up in our analysis of unpassaged sequences (Fig. 7A). Sites in this region are known to be involved in antigenic escape. In fact, many of the highlighted regions contain amino acid positions where substitution led to antigenic change (Table 2). We found a similar pattern of concordance with antigenic sites when mapping  $dN/dS$  values directly onto the structure (Fig. 7B). The inverse-distance correlations, however, performed better at identifying antigenic residues than did raw  $dN/dS$  values. When considering the 90th percentile (top 10% highest scored sites) by either metric, the inverse-distance correlations recovered 5 of 7 sites while  $dN/dS$  alone recovered only 1 of 7 sites (Table 2). Additionally, while several sites involved in antigenic change had very low  $dN/dS$ , all had inverse-distance correlations above the 86th percentile.

### 3.6 Passaging artifacts extend deep into reconstructed trees

Passaging adaptations could reasonably be expected to only affect peripheral clades of influenza virus evolutionary trees, as they represent recent signals of adaptation that should not penetrate far into the tree or significantly affect tree structure (Kryazhimskiy and Plotkin, 2008; Strelkowa and Lässig, 2012). Surprisingly, however, we found signals of passage adaptations in season-to-season fixed mutations and branch density.

To capture mutations that became fixed across seasons, we calculated  $dN/dS$  along the trunks of trees constructed from sequences of difference passage histories (Fig. 8A). To time-calibrate our trunk, we limited this analysis to passage types that had sequences at least in 2005 and in 2015, which excludes SIAT1 sequences. Trunk  $dN/dS$  measures season-to-season adaptation and might be expected to be robust to the effects of passaging. However, recurring adaptations to passaging conditions as samples were processed across seasons could falsely appear as trunk mutations.

For pooled sequences, trunk  $dN/dS$  appeared to be generally free of artifacts, resembling the trunk  $dN/dS$  of unpassaged sequences (Fig. 8A). In contrast, for datasets composed entirely of passaged sequences we found artifacts extending into the trunk. When trees were constructed from only cell-passaged sequences or only non-SIAT1 sequences, we observed a general inflation in  $dN/dS$  as well as several spurious sites of high  $dN/dS$  that do not occur in the unpassaged condition. Together, this result shows that the relative proportion of passaged to unpassaged sequences in a sample matters; when sequences with passaging artifacts are overrepresented compared to unpassaged sequences, there is a risk that a spurious signal will be found in the trunk. For example, a trunk  $dN/dS$  analysis of mainly non-SIAT1 sequences would direct attention to site 261, even though this site does not appear to be positively selected on the unpassaged tree trunk. This analysis demonstrates the ability of sequences containing major passaging artifacts to confound both deep and peripheral analyses of influenza virus evolution.

We next investigated the effect of passaging on a tree-topology based metric, LBI (Neher et al., 2014). Notably, this metric is entirely independent of  $dN/dS$  values. The LBI algorithm uses the degree of local branching around a terminal node to predict sequences similar to progenitors of the following season's strain. As a read-out of the algorithm's performance, we calculated the mean ratio score, as described (Neher et al., 2014) (see also 'Methods' section). Mean ratio scores  $<1$  indicate that the algorithm performs better than random chance. The lowest possible mean ratio score corresponds to the mean ratio score of the sequence with the lowest Hamming distance to the following year's progenitor (i.e., the theoretical best possible prediction from sequences in a condition).

We saw clear differences in accuracy of predictions made using trees composed of non-SIAT1, pooled, or unpassaged sequences (Fig. 8B and C). We could not examine SIAT1 patterns, as these sequences were not consistently available across seasons until 2013. As a general trend, predictions from unpassaged sequences seemed to be more accurate (both less likely to exceed 1 and more likely to be closer to the best possible prediction) than predictions from either passaged or pooled sequences.

## 4. Discussion

We find that serial passaging of influenza virus introduces a measurable signal of adaptation into the evolutionary analysis

of natural influenza virus sequences. There are unique, characteristic patterns of adaptation to egg passage, monkey cell passage, and non-SIAT1 cell passage. Monkey-cell-derived sequences show different molecule-wide evolutionary rate patterns. Non-SIAT1 cell-derived sequences instead display a hotspot of positive selection in a loop underneath the sialic-acid binding region. This hotspot has been previously noted (Meyer and Wilke, 2015) but no explanation for its origin was available. Additional passages in non-SIAT1 cell strengthen this artifact. Further, we find that virus passaged in SIAT1 cells seems to accumulate only minor passaging artifacts. Throughout our analyses, we find limited utility in subdividing phylogenetic trees to internal and terminal branches. While signals of passage adaptation are consistently elevated along terminal branches and attenuated along internal branches, evolutionary rates along internal branches remain confounded by passaging artifacts. Additionally, passage adaptation can resemble fixed season-to-season mutation along trunk branches and alter topology-based predictions of sequence fitness. Finally, we can accurately recover the experimentally determined antigenic regions of HA from evolutionary-rate analysis by using a dataset consisting of only unpassaged viral sequences.

Previous studies (Bush et al., 2001; Suzuki, 2006) suggest the use of internal branches to alleviate passage adaptations. However, we find here that this strategy is insufficient, because the evolutionary signal of passage adaptations can often be detected along internal branches. This finding may seem counter-intuitive, as internal nodes should exclusively represent human-adapted virus. We suggest that passaging adaptations in internal branches may be homoplasies caused by convergent evolution; if different clinical isolates converge onto the same adaptive mutations under passaging, then these mutations may incorrectly be placed along internal branches under phylogenetic tree reconstruction. Additionally, although the use of only internal branches removes some differences between the passage groups, the exclusion of terminal sequences can obscure recent natural adaptations and thus obscure actual sites under positive selection. Therefore, analysis of internal branches is not only insufficient for eliminating artifacts from passaging adaptations but also suboptimal for detecting positive selection in seasonal H3N2 influenza virus.

The safest route to avoid passaging artifacts is to limit sequence datasets to only unpassaged virus, although this approach limits sequence numbers. The human-like 6-linked sialic acids in SIAT1 (Matrosovich et al., 2003) greatly reduce observed cell culture-specific adaptations, particularly in the loop of HA which contains site 224. This lack of selection concurs with multiple experiments finding low levels of adaptation in this cell line (Oh et al., 2008; Hamamoto et al., 2013). As our analysis only detects minor differences between unpassaged and SIAT1 passaged virus, we posit that this passage condition is an acceptable substitute for unpassaged clinical samples. Even so, our findings do not preclude the existence of SIAT1-specific adaptations that may confound specific analyses.

Over half of the passaged HA sequences in the GISAID database from 2005 to 2015 were passaged more than once. Multiple passages in non-SIAT1 cells cause increasing accumulation of passaging artifacts, and we would expect that any yet unknown passaging effects might accumulate similarly. Nevertheless, even a single passage in non-SIAT1 cells introduces noticeable artifacts of adaptation. Influenza virus is often passaged multiple times to improve viral titers for hemagglutination inhibition assays, and thus, we expect that multiply passaged viruses

will continue to be deposited for the foreseeable future. We recommend that such viruses be used with care when studying the evolutionary dynamics of influenza virus strains circulating in the human population.

Although the majority of the sequences from the year 2015 are SIAT1-passaged or unpassaged, several hundred sequences from that year derive from monkey cell culture. The use of monkey cell culture surged in 2014 and 2015 compared to previous years. We recommend that these recently collected sequences be excluded from influenza virus evolutionary rate analysis, in favor of the majority of unpassaged and SIAT1-passaged sequences. As passaging is a useful and cost effective method for amplification of clinically collected virus, unpassaged viral sequences are unlikely to completely dominate influenza virus sequence databases in the near future. However, new human epithelial cell culture systems for influenza virus passaging (Ilyushina et al., 2012) could soon provide an ideal system that both amplifies virus and protects it from non-human selective pressures.

Passage history should routinely be considered as a potential confounding variable in future analyses of influenza virus evolutionary rates. Future studies should be checked against unpassaged samples to ensure that conclusions are not based on adaptation to non-human hosts. We recommend the exclusion of viral sequences that derive from serial passage in egg amniotes, monkey kidney cell culture, and any unspecified cell culture. Prior work that did not consider passaging history may be confounded by passaging adaptations, as occurred in our previous publication (Meyer and Wilke, 2015). In that article, we concluded that sites under positive selection differ from sites involved in immune escape. Here, we find that the origin of this positive selection is adaptation to the non-human passaging host, not immune escape in or adaptation to humans. In particular, we suggest that the evolutionary markers of influenza virus determined in (Belanov et al., 2015) be reevaluated to ensure these sites are not artifacts of viral passaging. Similarly, many of the earlier studies (Bush et al., 1999; Suzuki, 2006, 2008; Shih et al., 2007; Pan and Deem, 2011; Tusche, Steinbrück, and McHardy, 2012; Meyer and Wilke, 2013, 2015) performing site-specific evolutionary analysis of HA likely contain some conclusions that can be traced back to passaging artifacts. Additionally, even though passage artifacts do not appear to be sufficiently strong to affect clade-structure reconstruction (Bush et al., 2000), they do have the potential to cause artificially long branch lengths, due to  $dN/dS$  inflation, or misplaced branches, due to convergent evolution under passaging. We find samples composed of non-SIAT1 appear to behave differently than unpassaged samples under the LBI metric. Thus, future phylogenetic predictive models of influenza virus fitness and antigenicity, as in (Bedford et al., 2014; Łuksza and Lässig, 2014; Neher et al., 2014), should also be checked for robustness to passage-related signals. Finally, while it is beyond the scope of this work to investigate passage history effects in other viruses, we suspect that passage-derived artifacts could be a factor in their phylogenetic analyses as well. The use of datasets free of passage adaptations will likely bring computational predictions of influenza positive selection more in line with corresponding experimental results.

Sequences without passage annotations are inadequate for reliable evolutionary analysis of influenza virus. Yet, passage annotations are often completely missing from strain information, and, when present, are often inconsistent; there is currently no standardized language to represent number and type of serial passage. We note, however, that passage annotations

from the 2015 season are greatly improved when compared to previous seasons. Several major influenza repositories, including the Influenza Research Database (Squires et al., 2012) and the NCBI Influenza Virus Resource (Bao et al., 2008), do not provide any passaging annotations at all. Additionally, passage history is not required for new sequence submissions to the NCBI Genbank (Benson et al., 2012). The EpiFlu database maintained by the GISAID (Bogner et al., 2006) and OpenFluDB (Liechti et al., 2010), however, stand apart by providing passage history annotations for the majority of their sequences. Of these, only the OpenFluDB repository allows filtering of sequences by passage history during data download. Our results demonstrate the strength of passaging artifacts in evolutionary analysis of influenza virus. The lack of a universal standard for annotation of viral passage histories and a universal standard for serial passage experimental conditions complicate the analysis and mitigation of passaging effects.

## Supplementary data

Supplementary data are available at *Virus Evolution* online.

## Acknowledgements

We would like to thank Sebastian Maurer-Stroh for help with interpreting passaging annotations in GISAID. This work was supported in part by NIH grant no. R01 GM088344, DTRA grant no. HDTRA1-12-C-0007, and NSF Cooperative agreement no. DBI-0939454 (BEACON Center). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

## Data availability

Processed data are available as [supplementary material](#). Sequence data are available from GISAID as detailed in 'Methods' section. All analysis code used to generate the processed data is available at: [https://github.com/wilkelab/influenza\\_H3N2\\_passaging](https://github.com/wilkelab/influenza_H3N2_passaging).

**Conflict of interest:** None declared.

## References

- Balish, A. L., Katz, J. M., and Klimov, A. I. (2013) *Influenza: Propagation, Quantification, and Storage*. Current Protocols in Microbiology. 29:G:15G.1:15G.1.115G.1.24.
- Bao, Y. et al. (2008) 'The Influenza Virus Resource at the National Center for Biotechnology Information', *Journal of Virology*, 82: 596–601.
- Bedford, T. et al. (2014) 'Integrating Influenza Antigenic Dynamics with Molecular Evolution', *eLife*, 3: e01914.
- Belanov, S. S. et al. (2015) 'Genome-Wide Analysis of Evolutionary Markers of Human Influenza A(H1N1)pdm09 and A(H3N2) Viruses May Guide Selection of Vaccine Strain Candidates', *Genome Biology and Evolution*, 7: 3472–83.
- Benson, D. A. et al. (2012) 'GenBank', *Nucleic Acids Research*, 40: D48–53.
- Blackburne, B. P., Hay, A. J., and Goldstein, R. A. (2008) 'Changing Selective Pressure during Antigenic Changes in Human Influenza H3', *PLoS Pathogens*, 4: e1000058.
- Bogner, P. et al. (2006) 'A Global Initiative on Sharing Avian Flu Data', *Nature*, 442: 981.

- Bush, R. M. et al. (1999) 'Positive Selection on the H3 Hemagglutinin Gene of Human Influenza Virus A', *Molecular Biology and Evolution*, 16: 1457–65.
- , et al. (2001) 'Predicting Influenza Evolution: The Impact of Terminal and Egg-Adapted Mutations', *International Congress Series*, 1219: 147–53.
- , et al. (2000) 'Effects of Passage History and Sampling Bias on Phylogenetic Reconstruction of Human Influenza A Evolution', *Proceedings of the National Academy of Sciences*, 97: 6974–80.
- Echave, J., Spielman, S. J., and Wilke, C. O. (2016) 'Causes of Evolutionary Rate Variation Among Protein Sites', *Nature Reviews Genetics*, 17: 109–121.
- Gatherer, D. (2010) 'Passage in Egg Culture is a Major Cause of Apparent Positive Selection in Influenza B Hemagglutinin', *Journal of Medical Virology*, 82: 123–7.
- Hamamoto, I. et al. (2013) 'High Yield Production of Influenza Virus in Madin Darby Canine Kidney (MDCK) Cells with Stable Knockdown of IRF7', *PLoS One*, 8: e59892.
- Ihaka, R., and Gentleman, R. (1996) 'R: A Language for Data Analysis and Graphics', *Journal of Computational and Graphical Statistics*, 5: 299–314.
- Ilyushina, N. A. et al. (2012) 'Comparative Study of Influenza Virus Replication in MDCK Cells and in Primary Cells Derived from Adenoids and Airway Epithelium', *Journal of Virology*, 86: 11725–34.
- Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.
- Katz, J. M., and Webster, R. G. (1989) 'Efficacy of Inactivated Influenza A Virus (H3N2) Vaccines Grown in Mammalian Cells or Embryonated Eggs', *Journal of Infectious Diseases*, 160: 191–8.
- , Wang, M., and Webster, R. G. (1990) 'Direct Sequencing of the HA Gene of Influenza (H3N2) Virus in Original Clinical Samples Reveals Sequence Identity with Mammalian Cell-Grown Virus', *Journal of Virology*, 64: 1808–11.
- Koel, B.F. et al. (2013) 'Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution', *Science*, 342: 976–9.
- Koelle, K. et al. (2006) 'Epochal Evolution Shapes the Phylodynamics of Interpandemic Influenza A (H3N2) in Humans', *Science*, 314: 1898–903.
- Kratsch, C. et al. (2016) 'Determination of Antigenicity-Altering Patches on the Major Surface Protein of Human Influenza A/H3N2 Viruses', *Virus Evolution*, 2: vev025.
- Kryazhimskiy, S., and Plotkin, J. B. (2008) 'The Population Genetics of dN/dS', *PLoS Genetics*, 4: e1000304.
- Kumar, S., and Henrickson, K. J. (2012) 'Update on Influenza Diagnostics: Lessons from the Novel H1N1 Influenza A Pandemic', *Clinical Microbiology Reviews*, 25: 344–61.
- Lee, H. K. et al. (2013a) 'Simplified Large-Scale Sanger Genome Sequencing for Influenza A/H3N2 Virus', *PLoS One*, 8: e64785.
- , et al. (2013b) 'Comparison of Mutation Patterns in Full-Genome A/H3N2 Influenza Sequences Obtained Directly from Clinical Samples and the Same Samples after a Single MDCK Passage', *PLoS One*, 8: e79252.
- Liechti, R. et al. (2010) 'OpenFluDB, A Database for Human and Animal Influenza Virus', *Database*, 2010: baq004.
- Lin, Y. P. et al. (2012) 'Evolution of the Receptor Binding Properties of the Influenza A(H3N2) Hemagglutinin', *Proceedings of the National Academy of Sciences*, 109: 21474–9.
- Łuksza, M., and Lässig, M. (2014) 'A Predictive Fitness Model for Influenza', *Nature*, 507: 57–61.
- Matrosovich, M. et al. (2003) 'Overexpression of the Alpha-2,6-sialyltransferase in MDCK Cells Increases Influenza Virus Sensitivity to Neuraminidase Inhibitors', *Journal of Virology*, 77: 8418–25.
- Meyer, A. G., and Wilke, C. O. (2013) 'Integrating Sequence Variation and Protein Structure to Identify Sites under Selection', *Molecular Biology and Evolution*, 30: 36–44.
- , and —— (2015) 'Geometric Constraints Dominate the Antigenic Evolution of Influenza H3N2 Hemagglutinin', *PLoS Pathogens*, 11: e1004940.
- Neher, R. A., and Bedford, T. (2015) 'nextflu: Real-Time Tracking of Seasonal Influenza Virus Evolution in Humans', *Bioinformatics*, 31: 3546–48.
- , Russell, C. A., and Shraiman, B. I. (2014) 'Predicting Evolution From the Shape of Genealogical Trees', *eLife*, 3: e03568.
- Nelson, M. I. et al. (2006) 'Stochastic Processes Are Key Determinants of Short-Term Evolution in Influenza A Virus', *PLoS Pathogens*, 2: e125.
- Oh, D. Y. et al. (2008) 'MDCK-SIAT1 Cells Show Improved Isolation Rates for Recent Human Influenza Viruses Compared to Conventional MDCK Cells', *Journal of Clinical Microbiology*, 46: 2189–94.
- Pan, K., and Deem, M. W. (2011) 'Quantifying Selection and Diversity in Viruses by Entropy Methods, with Application to the Haemagglutinin of H3N2 Influenza', *Journal of the Royal Society, Interface*, 8: 1644–53.
- Pond, S. L. K., Frost, S. D. W., and Muse, S. V. (2005) 'HyPhy: Hypothesis Testing Using Phylogenies', *Bioinformatics*, 21: 676–9.
- Price, M. N., Dehal, P. S., and Arkin, A. P. (2010) 'FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments', *PLoS One*, 5: e9490.
- Robertson, J. S. et al. (1993) 'The Role of Amniotic Passage in the Egg-Adaptation of Human Influenza Virus is Revealed by Haemagglutinin Sequence Analyses', *Journal of General Virology*, 74(Pt 10): 2047–51.
- Shih, A. C. C. et al. (2007) 'Simultaneous Amino Acid Substitutions at Antigenic Sites Drive Influenza A Hemagglutinin Evolution', *Proceedings of the National Academy of Sciences*, 104: 6283–8.
- Skowronski, D. M. et al. (2014) 'Low 2012–13 Influenza Vaccine Effectiveness Associated with Mutation in the Egg-Adapted H3N2 Vaccine Strain Not Antigenic Drift in Circulating Viruses', *PLoS One*, 9: e92153.
- Spielman, S., Wan, S., and Wilke, C. O. (2016) 'A Comparison of One-Rate and Two-Rate Inference Frameworks for Site-Specific dN/dS Estimation', *Genetics*, Early online August 17, 2016; DOI: 10.1534/genetics.115.185264.
- Squires, R. B. et al. (2012) 'Influenza Research Database: An Integrated Bioinformatics Resource for Influenza Research and Surveillance', *Influenza and Other Respiratory Viruses*, 6: 404–16.
- Stöhr, K. et al. (2012) 'Influenza Virus Surveillance, Vaccine Strain Selection, and Manufacture', in Y., Kawaoka and G., Neumann (eds) *Influenza Virus*, pp. 147–62. Totowa, NJ: Humana Press.
- Strelkova, N., and Lässig, M. (2012) 'Clonal Interference in the Evolution of Influenza', *Genetics*, 192: 671–82.

- Suzuki, Y. (2006) 'Natural Selection on the Influenza Virus Genome', *Molecular Biology and Evolution*, 23: 1902–11.
- , et al. (2008) 'Positive Selection Operates Continuously on Hemagglutinin During Evolution of H3N2 Human Influenza A Virus', *Gene*, 427: 111–6.
- The World Health Organization (2015) *Recommended Composition of Influenza Virus Vaccines for Use in the 2015–2016 Northern Hemisphere Influenza Season*. Geneva: WHO Press.
- Tien, M. Z. et al. (2013) 'Maximum Allowed Solvent Accessibilities of Residues in Proteins', *PloS One*, 8: e80635.
- Tusche, C., Steinbrück, L., and McHardy, A. C. (2012) 'Detecting Patches of Protein Sites of Influenza A Viruses Under Positive Selection', *Molecular Biology and Evolution*, 29: 2063–71.
- WHO Writing Group et al. (2012) 'Improving Influenza Vaccine Virus Selection: Report of a WHO Informal Consultation Held at WHO Headquarters, Geneva, Switzerland, 14–16 June 2010', *Influenza and Other Respiratory Viruses*, 6: 142–52.
- Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.
- Wolf, Y. I. et al. (2006) 'Long Intervals of Stasis Punctuated by Bursts of Positive Selection in the Seasonal Evolution of Influenza A Virus', *Biology Direct*, 1: 34.
- World Health Organization Global Influenza Surveillance Network (2011) *Manual for the Laboratory Diagnosis and Virological Surveillance of Influenza*. Geneva: WHO Press.
- Wyde, P. R. et al. (1977) 'Effects of Low- and High-Passage Influenza Virus Infection in Normal and Nude Mice', *Infection and Immunity*, 15: 221–9.
- Xie, H. et al. (2015) 'H3N2 Mismatch of 2014–15 Northern Hemisphere Influenza Vaccines and Head-to-head Comparison between Human and Ferret Antisera derived Antigenic Maps', *Scientific Reports*, 5: 15279.