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

Claire D. McWhite[1,2], Austin G. Meyer[1,3,4], Claus O. Wilke[1,3,4*]

[1]Center for Systems and Synthetic Biology and Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX 78712

[2]Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX 78712

[3]Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, TX 78712

[4]Department of Integrative Biology, The University of Texas at Austin, Austin, TX 78712

Address correspondence to wilke@austin.utexas.edu

Clinical influenza A 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 in the viral sequences, 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 the influenza 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 evolutionary analyses should appropriately control for potentially confounding effects of passaging adaptations.

1. INTRODUCTION

The routine sequencing of clinical isolates has become a critical component of global seasonal influenza 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 (Blackburne et al., 2008; Koelle et al., 2006; Nelson et al., 2006; Suzuki, 2008; Wolf et al., 2006), 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 (Kratsch et al., 2016; Meyer and Wilke, 2015; Tusche et al., 2012), 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 (Kumar and Henrickson, 2012; World Health Organization Global influenza surveillance network, 2011). A variety of culture systems 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., 2005;

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 genomes can be obtained from PCR-amplified influenza samples without intermediate passaging (Katz et al., 1990; Lee et al., 2013a).

Several experiments have demonstrated that influenza virus hemagglutinin (HA) accumulates mutations following rounds of passaging in both cell (Ilyushina et al., 2012; Lee et al., 2013b; Wyde et al., 1977) 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 discernable 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 (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 (Gatherer, 2010), but this study was limited to comparing egg-cultured to cell-cultured virus. As the availability of unpassaged influenza sequences has dramatically increased over the past ten 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 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 hemagglutinin (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 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 6873 sequences from 2005-2015 as well as one outgroup of 45

sequences from 1968–1977 (not considered for further analysis). We did not consider sequences collected from 1978-2004.

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 with regular expressions for keywords in FASTA headers (Table 1). We classified 1133 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 3041 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 SLAC (Single-Likelihood Ancestor Counting) 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., 2015). Among different one-rate, site-specific models, SLAC performs nearly identically to other approaches (Spielman et al., 2015), 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 2005-2015 as our sample set due the low number of available sequences prior to this period. As *dN/dS* estimates can be confounded by sample size (Spielman et al., 2015), 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 hemagglutinin protein and gene numbering is offset by 16, all site numbering in this manuscript 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.

2.4 Local Branching Index analysis

We used the framework and code (https://github.com/rneher/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 fifty 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/rneher/FitnessInference) and calculated the hamming distance of the 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 fifty 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”, “1976 – 2015”. Acknowledgements for sequences used in this study are available in Supplementary File 11. 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 for the hemagglutinin trimer, and site-wise distance to protein site 224. Relative solvent accessibility of the hemagglutinin 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 script distances.py from: https://github.com/wilkelab/influenza_H3N2_passaging/tree/master/scripts. 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, 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/wilkelab/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 (Figure 1A). Samples may be passaged either once or serially (Table 1), even though a single passage is generally sufficient to obtain adequate amounts of viral DNA for sequencing. Reconstructed trees of influenza evolution contain a mixture of passage histories at their tips (Figure 1B). During passaging, influenza genomes accumulate adaptive mutations, and the effect of these mutations on evolutionary analyses of influenza 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 H3N2 hemagglutinin 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 Single Likelihood Ancestor Counting

(SLAC). Specifically, we calculated one-rate *dN/dS* estimates, i.e., site-specific *dN* values normalized by a global *dS* value (see Methods 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 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 (Figure 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 Figure 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, 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 analysis due to low sequence numbers ($n = 79$), however evolutionary rates from this condition correlated particularly poorly with those of random draws of 79 unpassaged sequences

(Supplementary Figure 1). This result is consistent with previous conclusions (Bush et al., 2000; Gatherer, 2010; Suzuki, 2006) that egg-derived sequences show specific adaptations not found otherwise in influenza 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 (Figure 2B), even though distinct differences between the passage groups remained. Conversely, when we only considered tip branches, correlations among most groups were relatively low (Figure 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 unpassaged sequences (Fig. 2A-C), with *dN/dS* values derived from monkey-cell-

passaged sequences being the least similar to *dN/dS* from unpassaged 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 unpassaged 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 hemagglutinin (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 H3N2 HA, this model explains over 30% of the variation in *dN/dS* using two simple physical measures, the relative solvent accessibility (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 hemagglutinin 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.

We first considered the correlation between *dN/dS* and RSA (Figure 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

342 structural measure of a persistent effect of *dN/dS* with stronger correlations in the full
343 tree and internal branches than in tip branches.
344
345 Next we considered the correlation between *dN/dS* and the inverse distance to site 224
346 to each site in the HA structure (Figure 3B). In contrast to RSA, correlations here were
347 systematically higher in tip branches, suggesting a recent adaptive signal. We found
348 virtually no correlation for unpassaged sequences, while a low correlation existed for
349 monkey-cell cultured sequences and a higher correlation for cell-passaged and pooled
350 sequences. To confirm that differences between pooled and unpassaged correlations
351 were not simply due to variation from random sampling, we created a null distribution of
352 correlations from 200 random draws of pooled sequences. The correlation for
353 unpassaged sequences was significantly lower than it was for pooled sequences ($z = -$
354 4.22, $p = 1.2 \times 10^{-5}$, Supplementary Figure 2). Correlations from pooled sequences
355 mirrored cell-culture correlations and persisted through internal branches. Thus, the
356 correlation of *dN/dS* with the inverse distance to site 224 seems to be primarily an
357 artifact of cell passage, even though its effect can be seen along internal branches as
358 well. This cell-specific signal dominates the pooled dataset. Further, this cell-specific
359 signal is partially attenuated along internal branches and amplified along tip branches,
360 as we would expect from a signal caused by recent host-specific adaptation. Even
361 though this signal is a true predictor of influenza evolutionary rates for virus grown in
362 cell culture, it does not transfer to unpassaged sequences and therefore has no
363 relevance for the circulating virus. This finding serves as a strong demonstration of

364 passage history as a confounder in analysis of hemagglutinin evolution, not just for egg
365 passage as previously demonstrated, but also for cell and monkey-cell passage.
366
367 Surprisingly, the correlation we found here between *dN/dS* and inverse distance to site
368 224 for pooled sequences (R^2 = 0.067) was less than half of the value previously
369 reported (Meyer and Wilke, 2015) (Fig. 3B). However, using a dataset of sequences
370 more temporally matched to the previously published analysis (2005–2014 instead of
371 2005–2015), we recovered the earlier higher correlation. This finding suggests that
372 there is some feature in the additional 2015 sequences that changes the pooled data's
373 relationship with inverse distance to site 224. In 2015, unpassaged and SIAT1
374 sequences each doubled in number compared to 2014, while the number of non-SIAT1
375 cell cultured sequences dropped dramatically (Table 1). SIAT1, an MDCK cell line which
376 overexpresses human-like 6-linked sialic acids over native 3-linked sialic acids
377 (Matrosovich et al., 2003) has higher sequence fidelity than unmodified MDCK
378 (Hamamoto et al., 2013). Therefore, we next investigated whether the drop in
379 correlation from 2014 to 2015 could be attributed to the recent reduction in cell culture
380 using non-SIAT1 cells.
381
382 **3.3 Adaptation to passage in SIAT1 cells is weak or absent**
383 In the preceding analyses, we lumped all cell cultures except monkey cells into the
384 same category. However, there are more subtle distinctions in cell passaging systems,
385 and they can exert differential selective pressures on human adapted virus (Hamamoto
386 et al., 2013; Oh et al., 2008). 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 Figure 3B.

SIAT1 is currently the dominant system for passaging of influenza virus in North America, with approximately half of the 2015 influenza 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 (Hamamoto et al., 2013; Oh et al., 2008). 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 = 1046), 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 (Figure 4A). By contrast, inverse distance to site 224 uniquely correlated with *dN/dS* from non-SIAT1-cultured virus (Figure 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

410 to site 224 represents a unique signal of adaptation to passaging in non-SIAT1 cells. In
411 other words, a non-SIAT1-specific signal can completely dominate all signals of positive
412 adaptation when a dataset contains a sufficiently high number of sequences passaged
413 in non-SIAT1 cells. In our analysis (Figure 3B), the high correlation of non-SIAT1 cell
414 *dN/dS* with inverse distance to site 224 is suppressed in the pooled condition because
415 the number of unpassaged and SIAT1-passaged sequences grew substantially in 2015.
416 This difference in sample composition explains the lower than expected correlations in
417 Figure 3B for pooled *dN/dS*.
418
419 As these three conditions were somewhat temporally separated (most non-SIAT1 cell
420 culture sequences were pre-2015, and most unpassaged and SIAT1 culture sequences
421 were post-2014), we controlled for season-to-season variation by drawing 249
422 sequences from each group from 2014. We again considered site-wise *dN/dS*
423 correlations among passaging groups, and we found that overall, unpassaged and
424 SIAT1-passaged sequences appeared the most similar (Supplementary Figure 3A–C).
425

426 **3.4 Signals of passaging adaptation accumulate with additional rounds of**
427 **passaging in non-SIAT1 cells**

428 Having identified non-SIAT1 cell culture as the source of the contaminating signal in
429 analyses of inverse distance to site 224, we next investigated the source of this signal at
430 the single amino acid level. We expected that a signal of adaptation to a passaging
431 system would strengthen with additional exposure to that system. Thus, we compared
432 the magnitude of the site-wise *dN/dS* values in sequences that had never been

433 passaged, had been passaged once, passaged twice, or passaged three to five times
434 (Figure 5A). For this analysis, we only considered passage in non-SIAT1 cells. This
435 analysis revealed distinct regions of increasing positive selection along the
436 hemagglutinin molecule (arrows in Figure 5A) and a strong relationship between the
437 magnitude of these signals and the number of times influenza viruses were passaged.
438 Further, we found an overall increase in *dN/dS* with increased numbers of passages in
439 non-SIAT1 cells (Figure 5B). The strongly selected sites 221 and 225 are adjacent to
440 site 224, explaining the specific relationship between *dN/dS* calculated from non-SIAT1
441 sequences and the inverse distance in 3D space to this site. The correlation between
442 *dN/dS* and inverse distance increased in strength with increasing numbers of passages
443 (Figure 5C), even though it was observable after a single passage in non-SIAT1 cells.
444 Mapping the raw *dN/dS* values onto the hemagglutinin structure showed how specific
445 sites light up as passage numbers increase (Figure 5D).
446
447 **3.5 Evolutionary variation in sequences from unpassaged virus predicts regions**
448 **involved in antigenic escape**
449 Our preceding analyses might suggest that the inverse distance metric for describing
450 regions of selection only captures effects of adaptation to non-SIAT1 cell culture.
451 However, this is not necessarily the case. Importantly, inverse distance needs to be
452 calculated relative to a specific reference point. Site 224 was previously used as the
453 reference point because it yielded the highest correlation for the dataset analyzed
454 (Meyer and Wilke, 2015). For a different dataset, one that doesn't carry the signal of

455 adaptation to non-SIAT1 cell culture, a different reference point may be more
456 appropriate.
457
458 We thus repeated the inverse distance analysis of Meyer and Wilke (2015) for a size-
459 matched sample of 1046 sequences from non-SIAT1, pooled, SIAT1, and unpassaged
460 virus collected between 2005 and 2015 (Figure 6). In brief, for each possible reference
461 site in the hemagglutinin structure, we measured the inverse distance in 3D space from
462 that site to every other site in the structure (see Methods for details). We then correlated
463 the inverse distances with the *dN/dS* values at each site, resulting in one correlation
464 coefficient per reference site. Finally, we mapped these correlation coefficients onto the
465 HA structure, coloring each reference site by its associated correlation coefficient. If
466 inverse distances measured from a particular reference amino acid have higher
467 correlation with the site-wise *dN/dS* values, then this reference site will appear
468 highlighted on the structure.
469
470 For non-SIAT1-passaged and pooled virus, this analysis recovered the finding of Meyer
471 and Wilke (2015) that the loop containing site 224 appeared strongly highlighted (Figure
472 7A). However, this signal was entirely absent in unpassaged and SIAT1 passaged virus
473 (Figure 7B), with no sites in that loop working well as a reference point. These results
474 suggested that this loop was specifically involved in adaptation of hemagglutinin to non-
475 SIAT1 cell culture, explaining the non-SIAT1-specific signal shown in Figure 4A.
476 Globally, the pattern of correlations from pooled sequences strongly resembled the non-
477 SIAT1 pattern, in contrast to the resemblance of SIAT1 to unpassaged. Thus, the

478 inverse distance metric is useful for differentiating regions of selection particular to
479 different experimental groups.
480
481 Therefore, we next asked what residual patterns of positive selection remained once the
482 adaptation to non-SIAT1 cells was removed. Even though site-wise correlations are
483 relatively low for unpassaged virus compared to the ones observed for non-SIAT1-
484 passaged virus, we could still recover relevant patterns of HA adaptation after rescaling
485 our coloring. In particular, we found that sites opposite to the loop-containing site 224 lit
486 up in our analysis of unpassaged sequences (Figure 7A). Sites in this region are known
487 to be involved in antigenic escape. In fact, many of the highlighted regions contain
488 amino acid positions where substitution led to antigenic change (Table 2). We found a
489 similar pattern of concordance with antigenic sites when mapping *dN/dS* values directly
490 onto the structure (Figure 7B). The inverse-distance correlations, however, performed
491 better at identifying antigenic residues than did raw *dN/dS* values. When considering the
492 90th percentile (top 10% highest scored sites) by either metric, the inverse-distance
493 correlations recovered 5 of 7 sites while *dN/dS* alone recovered only 1 of 7 sites (Table
494 2). Additionally, while several sites involved in antigenic change had very low *dN/dS*, all
495 had inverse-distance correlations above the 86th percentile.
496
497 **3.6 Passaging artifacts extend deep into reconstructed trees**
498 Passaging adaptations could reasonably be expected to only affect peripheral clades of
499 influenza virus evolutionary trees, as they represent recent signals of adaptation that
500 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 (Figure 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 (Figure 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

524 the unpassaged tree trunk. This analysis demonstrates the ability of sequences
525 containing major passaging artifacts to confound both deep and peripheral analyses of
526 influenza virus evolution.
527
528 We next investigated the effect of passaging on a tree-topology based metric, Local
529 Branching Index (LBI) (Neher et al., 2014). Notably, this metric is entirely independent
530 of *dN/dS* values. The LBI algorithm uses the degree of local branching around a
531 terminal node to predict sequences similar to progenitors of the following season's
532 strain. As a read-out of the algorithm's performance, we calculated the mean ratio
533 score, as described (Neher et al., 2014) (see also Methods). Mean ratio scores below 1
534 indicate that the algorithm performs better than random chance. The lowest possible
535 mean ratio score possible corresponds to the mean ratio score of the sequence with the
536 lowest hamming distance to the following year's progenitor (i.e., the theoretical best
537 possible prediction from sequences in a condition).
538
539 We saw clear differences in accuracy of predictions made using trees composed of non-
540 SIAT1, pooled, or unpassaged sequences (Figure 8B, C). We could not examine SIAT1
541 patterns, as these sequences were not consistently available across seasons until
542 2013. As a general trend, predictions from unpassaged sequences seemed to be more
543 accurate (both less likely to exceed 1 and more likely to be closer to the best possible
544 prediction) than predictions from either passaged or pooled sequences.
545

546 **4. DISCUSSION**

We find that serial passaging of influenza virus introduces a measurable signal of adaptation into the evolutionary analysis of natural influenza 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 hemagglutinin 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 counterintuitive, 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.

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 hemagglutinin which contains site 224. This lack of selection concords with multiple experiments finding low levels of adaptation in this cell line (Hamamoto et al., 2013; Oh et al., 2008). 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 hemagglutinin sequences in the GISAID database from 2005-2015 were passaged more than once. Multiple passages cause increasing accumulation of passaging artifacts in non-SIAT1 cells, and we predict that any yet unknown passaging effects would accumulate similarly. However, 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 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 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 sequence databases in the near future. However, new human epithelial cell culture systems for influenza 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 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 manuscript, 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; Meyer and Wilke, 2013, 2015; Pan and Deem, 2011; Shih et al., 2007; Suzuki, 2006, 2008; Tusche et al., 2012) performing site-specific evolutionary analysis of hemagglutinin 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 Local Branching Index metric. Thus, future phylogenetic predictive models of influenza fitness and antigenicity, as in (Bedford et al., 2014; Łuksza and Lässig, 2014; Neher et al.,

638 2014), should also be checked for robustness to passage-related signals. Finally, while
639 it is beyond the scope of this work to investigate passage history effects in other
640 viruses, we suspect that passage-derived artifacts could be a factor in their phylogenetic
641 analyses as well. The use of datasets free of passage adaptations will likely bring
642 computational predictions of influenza positive selection more in line with corresponding
643 experimental results.
644
645 Sequences without passage annotations are inadequate for reliable evolutionary
646 analysis of influenza virus. Yet, passage annotations are often completely missing from
647 strain information, and, when present, are often inconsistent; there is currently no
648 standardized language to represent number and type of serial passage. We note,
649 however, that passage annotations from the 2015 season are greatly improved when
650 compared to previous seasons. Several major influenza repositories, including the
651 Influenza Research Database (Squires et al., 2012) and the NCBI Influenza Virus
652 Resource (Bao et al., 2008), do not provide any passaging annotations at all.
653 Additionally, passage history is not required for new sequence submissions to the NCBI
654 Genbank (Benson et al., 2012). The EpiFlu database maintained by the Global Initiative
655 for Sharing Avian Influenza Data (GISAID) (Bogner et al., 2006) and OpenFluDB
656 (Liechti et al., 2010), however, stand apart by providing passage history annotations for
657 the majority of their sequences. Of these, only the OpenFluDB repository allows filtering
658 of sequences by passage history during data download. Our results demonstrate the
659 strength of passaging artifacts in evolutionary analysis of influenza. The lack of a
660 universal standard for annotation of viral passage histories and a universal standard for

661 serial passage experimental conditions complicate the analysis and mitigation of
662 passaging effects.
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683

REFERENCES

Balish, A.L., Katz, J.M., and Klimov, A.I. (2005). Influenza: Propagation, Quantification, and Storage. In Current Protocols in Microbiology, (John Wiley & Sons, Inc.), p.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., and Lipman, D. (2008). The Influenza Virus Resource at the National Center for Biotechnology Information. J. Virol. *82*, 596–601.

Bedford, T., Suchard, M.A., Lemey, P., Dudas, G., Gregory, V., Hay, A.J., McCauley, J.W., Russell, C.A., Smith, D.J., and Rambaut, A. (2014). Integrating influenza antigenic dynamics with molecular evolution. eLife *3*, e01914.

Belanov, S.S., Bychkov, D., Benner, C., Ripatti, S., Ojala, T., Kankainen, M., Lee, H.K., Tang, J.W.-T., and Kainov, D.E. (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 Biol. Evol. evv240.

Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., and Sayers, E.W. (2012). GenBank. Nucleic Acids Res. *40*, D48–D53.

Blackburne, B.P., Hay, A.J., and Goldstein, R.A. (2008). Changing Selective Pressure during Antigenic Changes in Human Influenza H3. PLoS Pathog *4*, e1000058.

Bogner, P., Capua, I., Lipman, D.J., Cox, N.J., and others (2006). A global initiative on sharing avian flu data. Nature *442*, 981–981.

Bush, R.M., Fitch, W.M., Bender, C.A., and Cox, N.J. (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol. Biol. Evol. *16*, 1457–1465.

Bush, R.M., Smith, C.B., Cox, N.J., and Fitch, W.M. (2000). Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. Proc. Natl. Acad. Sci. *97*, 6974–6980.

Bush, R.M., Fitch, W.M., Smith, C.B., and Cox, N.J. (2001). Predicting influenza evolution: the impact of terminal and egg-adapted mutations. Int. Congr. Ser. *1219*, 147–153.

Echave, J., Spielman, S.J., and Wilke, C.O. (2016). Causes of evolutionary rate variation among protein sites. Nat. Rev. Genet.

Gatherer, D. (2010). Passage in egg culture is a major cause of apparent positive selection in influenza B hemagglutinin. J. Med. Virol. *82*, 123–127.

Hamamoto, I., Takaku, H., Tashiro, M., and Yamamoto, N. (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. J. Comput. Graph. Stat. *5*, 299–314.

716 Ilyushina, N.A., Ikizler, M.R., Kawaoka, Y., Rudenko, L.G., Treanor, J.J., Subbarao, K., and Wright, P.F.
717 (2012). Comparative study of influenza virus replication in MDCK cells and in primary cells derived from
718 adenoids and airway epithelium. J. Virol. *86*, 11725–11734.

719 Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7:
720 Improvements in Performance and Usability. Mol. Biol. Evol. *30*, 772–780.

721 Katz, J.M., and Webster, R.G. (1989). Efficacy of Inactivated Influenza A Virus (H3N2) Vaccines Grown in
722 Mammalian Cells or Embryonated Eggs. J. Infect. Dis. *160*, 191–198.

723 Katz, J.M., Wang, M., and Webster, R.G. (1990). Direct sequencing of the HA gene of influenza (H3N2)
724 virus in original clinical samples reveals sequence identity with mammalian cell-grown virus. J. Virol. *64*,
725 1808–1811.

726 Koelle, K., Cobey, S., Grenfell, B., and Pascual, M. (2006). Epochal evolution shapes the phylodynamics of
727 interpandemic influenza A (H3N2) in humans. Science *314*, 1898–1903.

728 Kratsch, C., Klingen, T.R., Mümken, L., Steinbrück, L., and McHardy, A.C. (2016). Determination of
729 antigenicity-altering patches on the major surface protein of human influenza A/H3N2 viruses. Virus
730 Evol. *2*, vev025.

731 Kryazhimskiy, S., and Plotkin, J.B. (2008). The Population Genetics of dN/dS. PLOS Genet *4*, e1000304.

732 Kumar, S., and Henrickson, K.J. (2012). Update on Influenza Diagnostics: Lessons from the Novel H1N1
733 Influenza A Pandemic. Clin. Microbiol. Rev. *25*, 344–361.

734 Lee, H.K., Tang, J.W.-T., Kong, D.H.-L., and Koay, E.S.-C. (2013a). Simplified Large-Scale Sanger Genome
735 Sequencing for Influenza A/H3N2 Virus. PLoS ONE *8*.

736 Lee, H.K., Tang, J.W.-T., Kong, D.H.-L., Loh, T.P., Chiang, D.K.-L., Lam, T.T.-Y., and Koay, E.S.-C. (2013b).
737 Comparison of Mutation Patterns in Full-Genome A/H3N2 Influenza Sequences Obtained Directly from
738 Clinical Samples and the Same Samples after a Single MDCK Passage. PLoS ONE *8*, e79252.

739 Liechti, R., Gleizes, A., Kuznetsov, D., Bougueleret, L., Mercier, P.L., Bairoch, A., and Xenarios, I. (2010).
740 OpenFluDB, a database for human and animal influenza virus. Database *2010*, baq004.

741 Lin, Y.P., Xiong, X., Wharton, S.A., Martin, S.R., Coombs, P.J., Vachieri, S.G., Christodoulou, E., Walker,
742 P.A., Liu, J., Skehel, J.J., et al. (2012). Evolution of the receptor binding properties of the influenza
743 A(H3N2) hemagglutinin. Proc. Natl. Acad. Sci. *109*, 21474–21479.

744 Łuksza, M., and Lässig, M. (2014). A predictive fitness model for influenza. Nature *507*, 57–61.

745 Matrosovich, M., Matrosovich, T., Carr, J., Roberts, N.A., and Klenk, H.-D. (2003). Overexpression of the
746 alpha-2,6-sialyltransferase in MDCK cells increases influenza virus sensitivity to neuraminidase
747 inhibitors. J. Virol. *77*, 8418–8425.

748 Meyer, A.G., and Wilke, C.O. (2013). Integrating Sequence Variation and Protein Structure to Identify
749 Sites under Selection. Mol. Biol. Evol. *30*, 36–44.

Meyer, A.G., and Wilke, C.O. (2015). Geometric Constraints Dominate the Antigenic Evolution of Influenza H3N2 Hemagglutinin. PLoS Pathog. *11*.

Neher, R.A., and Bedford, T. (2015). nextflu: real-time tracking of seasonal influenza virus evolution in humans. Bioinformatics btv381.

Neher, R.A., Russell, C.A., and Shraiman, B.I. (2014). Predicting evolution from the shape of genealogical trees. eLife *3*, e03568.

Nelson, M.I., Simonsen, L., Viboud, C., Miller, M.A., Taylor, J., George, K.S., Griesemer, S.B., Ghedin, E., Sengamalay, N.A., Spiro, D.J., et al. (2006). Stochastic Processes Are Key Determinants of Short-Term Evolution in Influenza A Virus. PLoS Pathog *2*, e125.

Oh, D.Y., Barr, I.G., Mosse, J.A., and Laurie, K.L. (2008). MDCK-SIAT1 cells show improved isolation rates for recent human influenza viruses compared to conventional MDCK cells. J. Clin. Microbiol. *46*, 2189–2194.

Pan, K., and Deem, M.W. (2011). Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza. J. R. Soc. Interface *8*, 1644–1653.

Pond, S.L.K., Frost, S.D.W., and Muse, S.V. (2005). HyPhy: hypothesis testing using phylogenies. Bioinformatics *21*, 676–679.

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., Nicolson, C., Major, D., Robertson, E.W., and Wood, J.M. (1993). The role of amniotic passage in the egg-adaptation of human influenza virus is revealed by haemagglutinin sequence analyses. J. Gen. Virol. *74 (Pt 10)*, 2047–2051.

Shih, A.C.-C., Hsiao, T.-C., Ho, M.-S., and Li, W.-H. (2007). Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. Proc. Natl. Acad. Sci. *104*, 6283–6288.

Skowronski, D.M., Janjua, N.Z., De Serres, G., Sabaiduc, S., Eshaghi, A., Dickinson, J.A., Fonseca, K., Winter, A.-L., Gubbay, J.B., Krajden, 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. (2015). One-rate models outperform two-rate models in site-specific dN/dS estimation. bioRxiv 32805.

Squires, R.B., Noronha, J., Hunt, V., García-Sastre, A., Macken, C., Baumgarth, N., Suarez, D., Pickett, B.E., Zhang, Y., Larsen, C.N., et al. (2012). Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance. Influenza Other Respir. Viruses *6*, 404–416.

Stöhr, K., Bucher, D., Colgate, T., and Wood, J. (2012). Influenza Virus Surveillance, Vaccine Strain Selection, and Manufacture. In Influenza Virus, Y. Kawaoka, and G. Neumann, eds. (Humana Press), pp. 147–162.

785 Strelkowa, N., and Lässig, M. (2012). Clonal Interference in the Evolution of Influenza. Genetics *192*,
786 671–682.

787 Suzuki, Y. (2006). Natural selection on the influenza virus genome. Mol. Biol. Evol. *23*, 1902–1911.

788 Suzuki, Y. (2008). Positive selection operates continuously on hemagglutinin during evolution of H3N2
789 human influenza A virus. Gene *427*, 111–116.

790 The World Health Organization (2015). Recommended composition of influenza virus vaccines for use in
791 the 2015- 2016 northern hemisphere influenza season.

792 Tien, M.Z., Meyer, A.G., Sydykova, D.K., Spielman, S.J., and Wilke, C.O. (2013). Maximum allowed solvent
793 accessibilites of residues in proteins. PloS One *8*, e80635.

794 Tusche, C., Steinbrück, L., and McHardy, A.C. (2012). Detecting patches of protein sites of influenza A
795 viruses under positive selection. Mol. Biol. Evol. *29*, 2063–2071.

796 WHO Writing Group, Ampofo, W.K., Baylor, N., Cobey, S., Cox, N.J., Daves, S., Edwards, S., Ferguson, N.,
797 Grohmann, G., Hay, A., et al. (2012). Improving influenza vaccine virus selectionReport of a WHO
798 informal consultation held at WHO headquarters, Geneva, Switzerland, 14–16 June 2010. Influenza
799 Other Respir. Viruses *6*, 142–152.

800 Wickham, H. (2009). ggplot2: elegant graphics for data analysis (Springer New York).

801 Wolf, Y.I., Viboud, C., Holmes, E.C., Koonin, E.V., and Lipman, D.J. (2006). Long intervals of stasis
802 punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. Biol. Direct *1*,
803 34.

804 World Health Organization Global influenza surveillance network (2011). Manual for the laboratory
805 diagnosis and virological surveillance of influenza (Geneva, Switzerland).

806 Wyde, P.R., Couch, R.B., Mackler, B.F., Cate, T.R., and Levy, B.M. (1977). Effects of Low- and High-
807 Passage Influenza Virus Infection in Normal and Nude Mice. Infect. Immun. *15*, 221–229.

808 Xie, H., Wan, X.-F., Ye, Z., Plant, E.P., Zhao, Y., Xu, Y., Li, X., Finch, C., Zhao, N., Kawano, T., et al. (2015).
809 H3N2 Mismatch of 2014–15 Northern Hemisphere Influenza Vaccines and Head-to-head Comparison
810 between Human and Ferret Antisera derived Antigenic Maps. Sci. Rep. *5*.

811

812

813

814

815

816

817

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 manuscript.

DATA AVAILABILITY

Processed data are available as supplementary material. Sequence data are available from GISAID as detailed in Methods. All analysis code used to generate the processed data is available at: https://github.com/wilkelab/influenza_H3N2_passaging

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: CDM COW. Wrote scripts and analytic tools: CDM AGM. Performed the experiments: CDM. Analyzed the data: CDM COW. Wrote the paper: CDM AGM COW.

COMPETING FINANCIAL INTERESTS STATEMENT

The authors declare no competing financial interests.

838 **TABLES**

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

848

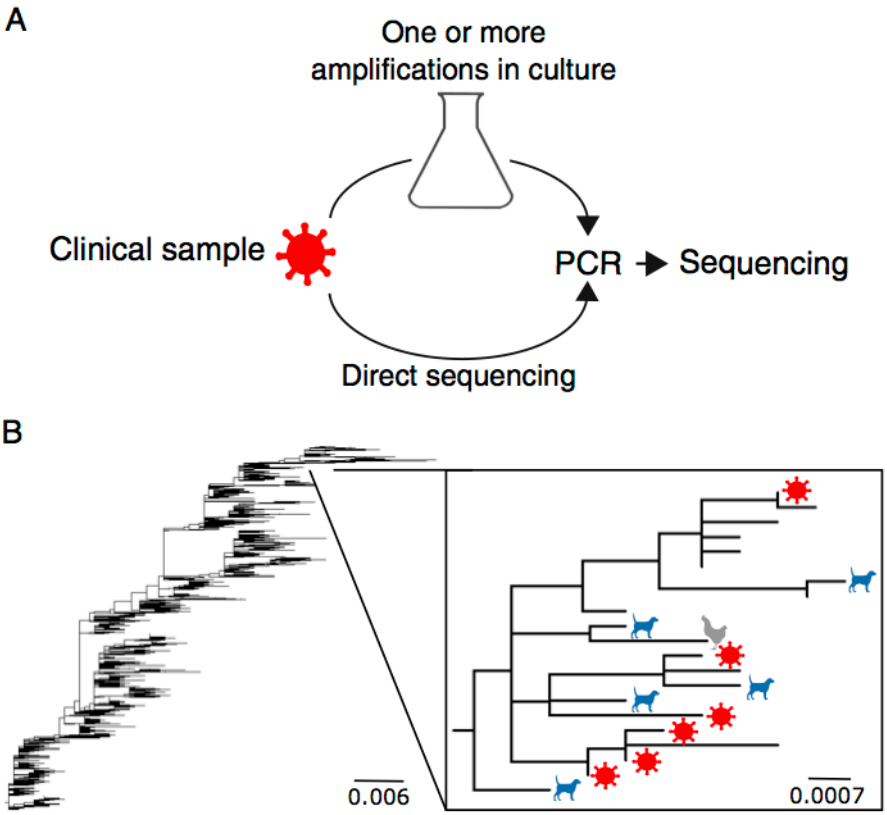
		Number of sequences			Rounds of passage	
Passage group	**Regular expression**	**2014**	**2015**	**2005–2015**	**Single**	**Multiple**
Chicken egg amniotes	AM[1-9]|E[1-7]|AMNIOTIC|EGG|EX|AM_[1-9]	6	0	79	1	79
Monkey cell culture	TMK|RMK|RHMK|RII|PMK|R[1-9]|RX	366	290	917	904	13
Generic cell culture	S[1-9]|SX|SIAT|MDCK|C[1-9]|CX|C_[1-9]|M[1-9]|MX|X[1-9]|^X_$	794	787	3041	867	2158
SIAT1	^S[1-9]_$\|^SX_$|SIAT2_SIAT1|SIAT3_SIAT1	389	626	1046	459	587
Non-SIAT1 cell culture	not SIAT|SX|S[1-9]	297	56	1755	408	1331
Unpassaged	LUNG|P0|OR_|ORIGINAL|CLINICAL|DIRECT	249	506	1703	N/A	N/A
Pooled		1508	1601	6873	1772	2317

849

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 unpassaged 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

859 **FIGURES**



860

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

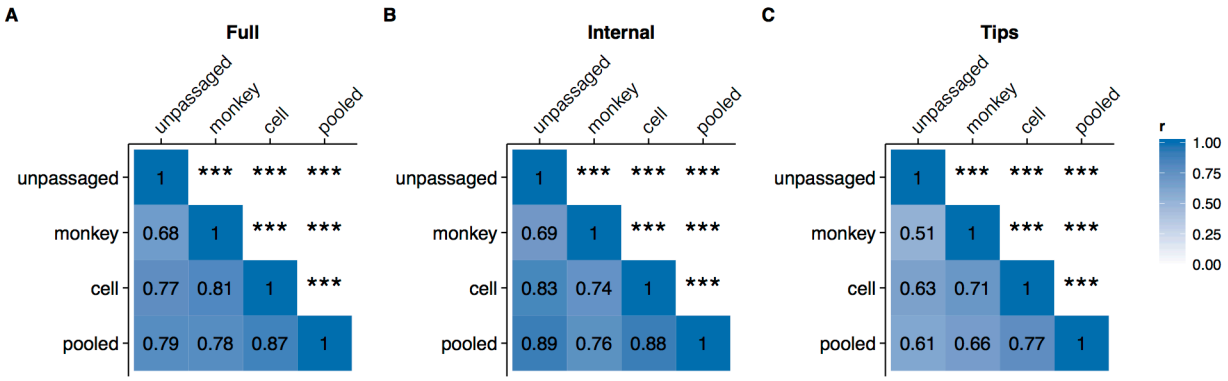


Figure 2. Comparison of sitewise *dN/dS* values among sequences with differing passage histories. Pearson correlations between sitewise *dN/dS* values for HA sequences derived from passaged and unpassaged influenza virus collected between 2005 and 2015 (downsampled to n = 917 in all groups). Correlations 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.

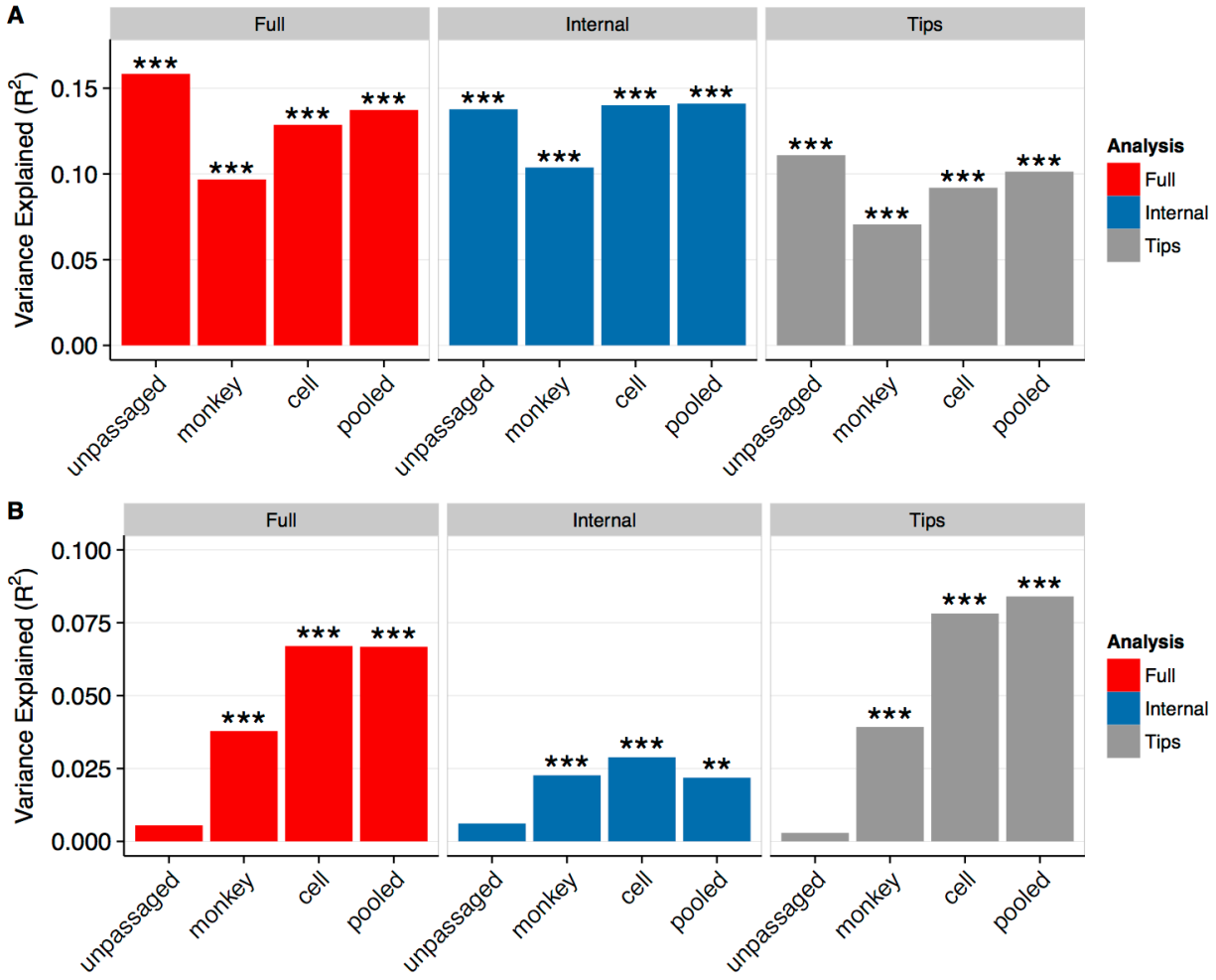


Figure 3. Percent variance in *dN/dS* explained by relative solvent accessibility (A) and by inverse distance to protein site 224 (B). (A) Relative solvent accessibility (RSA) explains ~10%–16% of the variation in *dN/dS* for all sequences. (B) Inverse distance to site 224 explains ~7% of the variation 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.

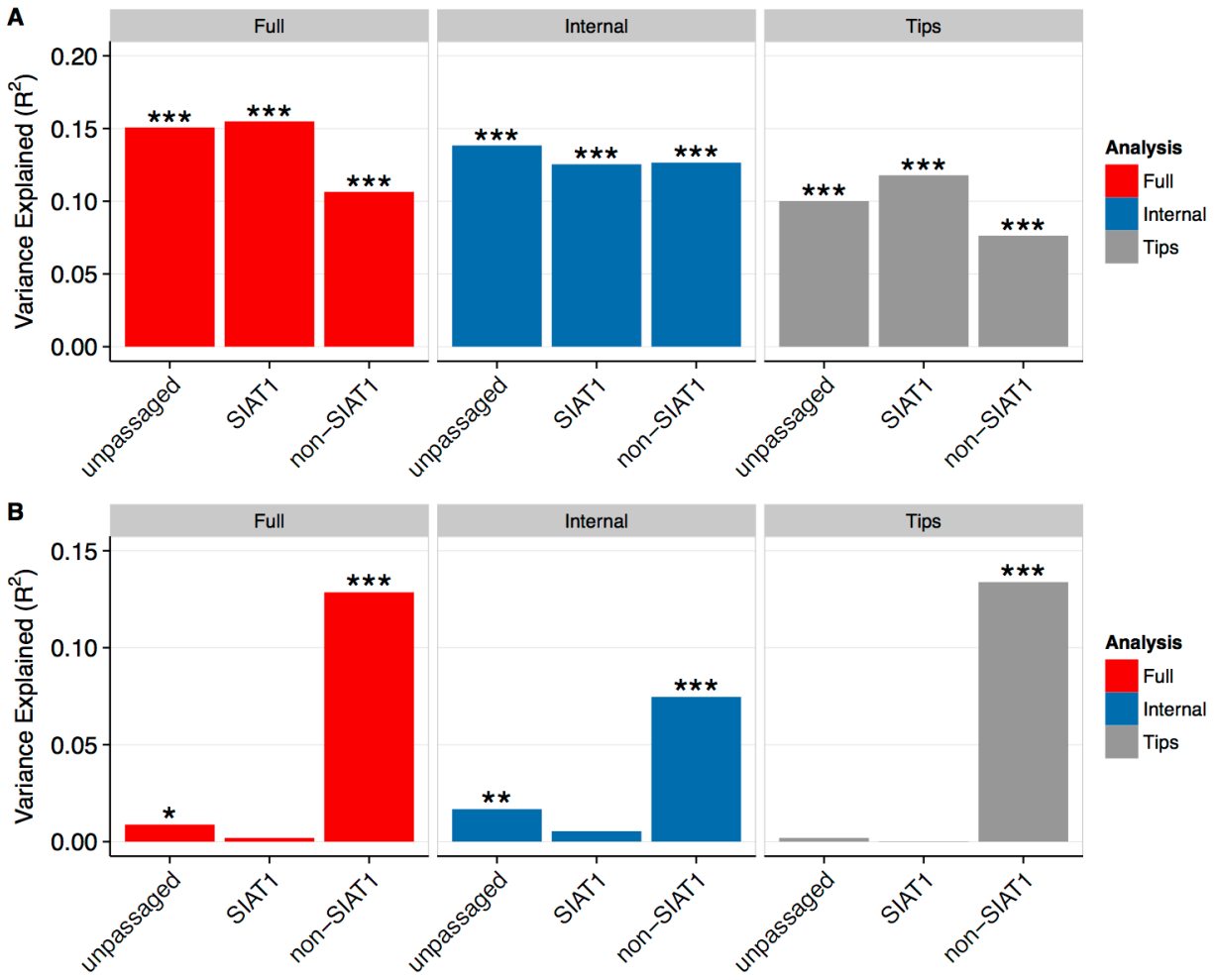


Figure 4. **Virus passaged in non-SIAT1 cells carries unique adaptations not present in unpassaged or SIAT1-passaged virus.** (A) The correlation between *dN/dS* and RSA is weakened for virus passaged in non-SIAT1 cells. (B) The correlation between *dN/dS* and inverse distance to site 224, representing 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.

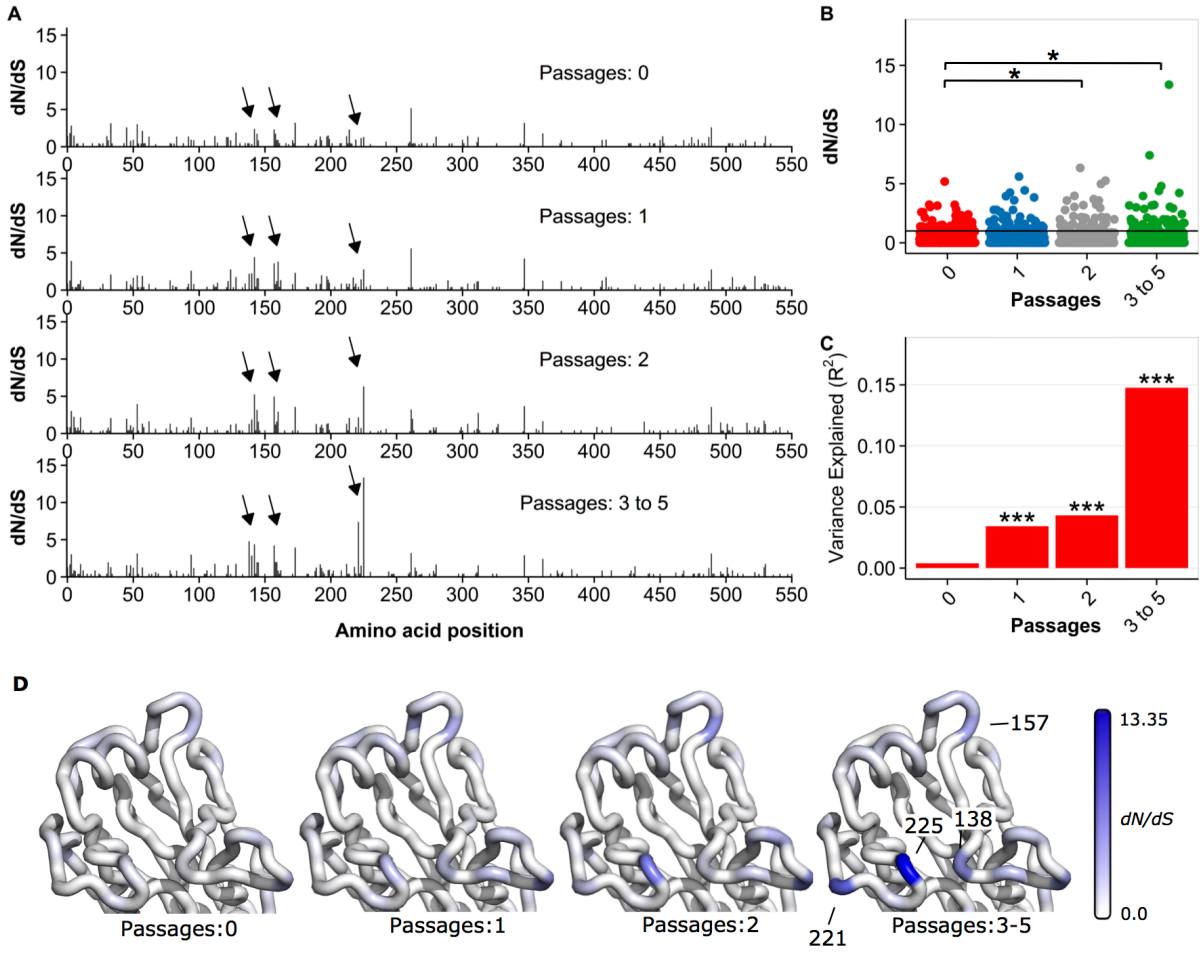
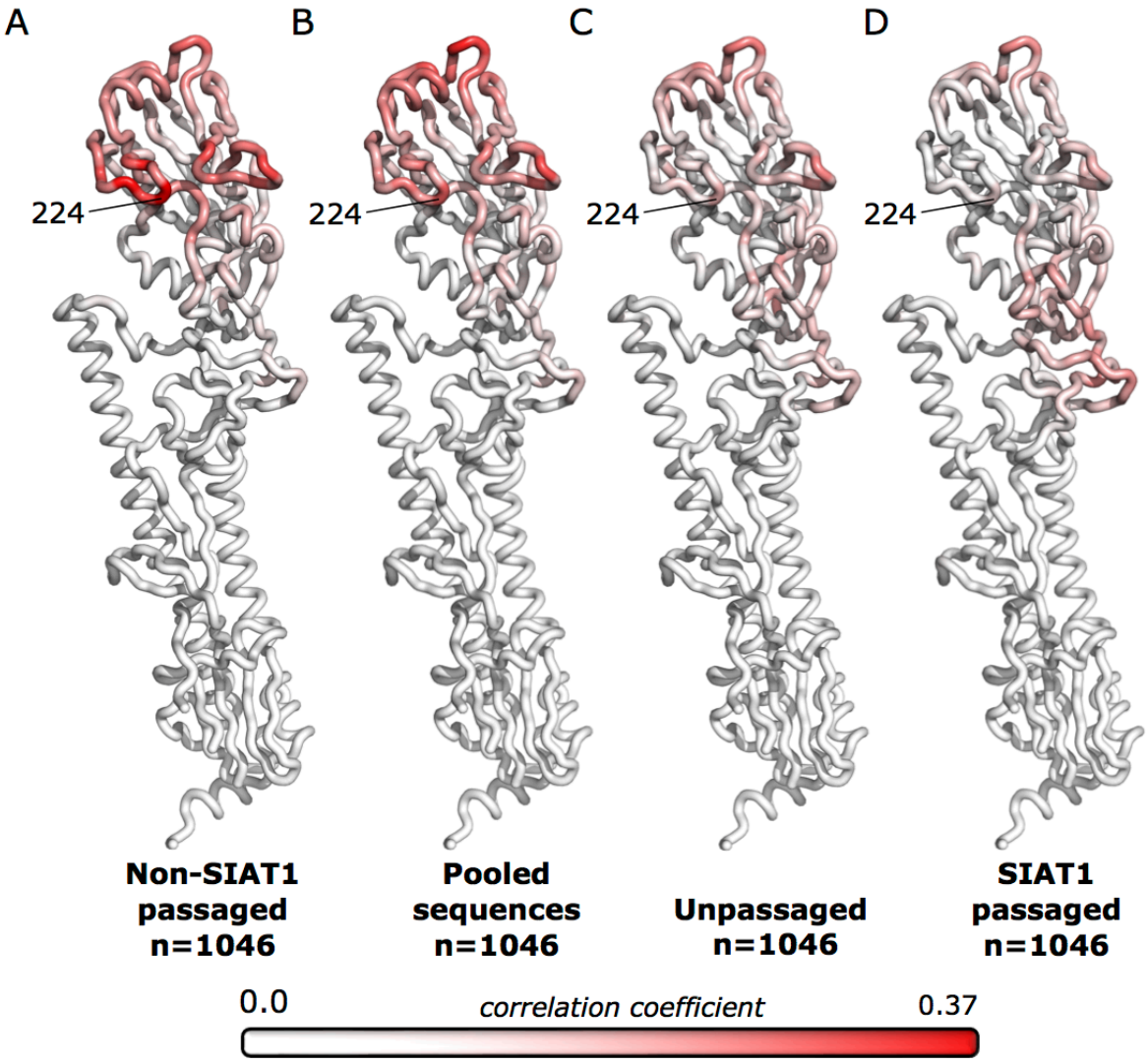


Figure 5. Accumulation of passaging artifacts with increasing numbers of serial passages in non-SIAT1 cell culture. (A) Sitewise *dN/dS* values for virus which was not passaged, passaged once, passaged twice, and passaged three to five times in non-SIAT1 cell culture (*n* = 304 for each group). Arrows 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

911 hemagglutinin head structure demonstrates the accumulation of passage adaptations
912 with increasing rounds of passages. Labeled sites correspond to regions denoted with
913 arrows in (A). Data used to generate this figure are available in Supplementary Data 6.



914

915 **Figure 6. Correlations of *dN/dS* with inverse distances, mapped onto the**
916 **hemagglutinin structure for non-SIAT1-passaged, pooled, unpassaged, and SIAT1**
917 **passaged sequences.** The correlation between *dN/dS* and inverse distance for each
918 reference site was mapped onto the hemagglutinin structure for (A) non-SIAT1
919 sequences, (B) pooled sequences, (C) unpassaged sequences, and (D) SIAT1
920 passaged sequences. Sequences analyzed were collected between 2005 and 2015.
921 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 unpassaged/SIAT1. In particular, the loop containing site 224 lights up strongly for non-SIAT1 and pooled sequences but not for unpassaged and SIAT1 sequences. Data used to generate this figure are available in Supplementary Data 7.

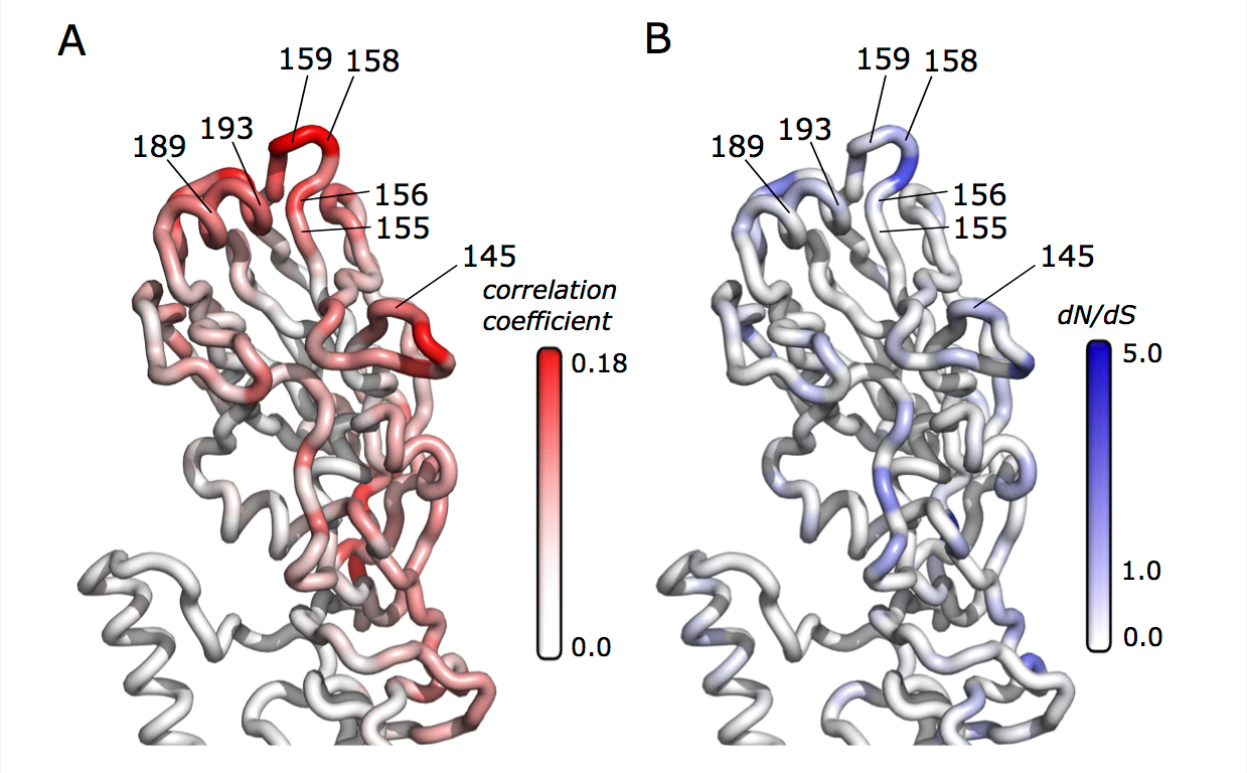


Figure 7. **Unpassaged sequences allow recovery of antigenic regions from positive-selection analysis.** For each site, the correlation between *dN/dS* and inverse distance (A) or *dN/dS* directly (B) were mapped onto the hemagglutinin structure, for *dN/dS* derived from unpassaged sequences collected 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.

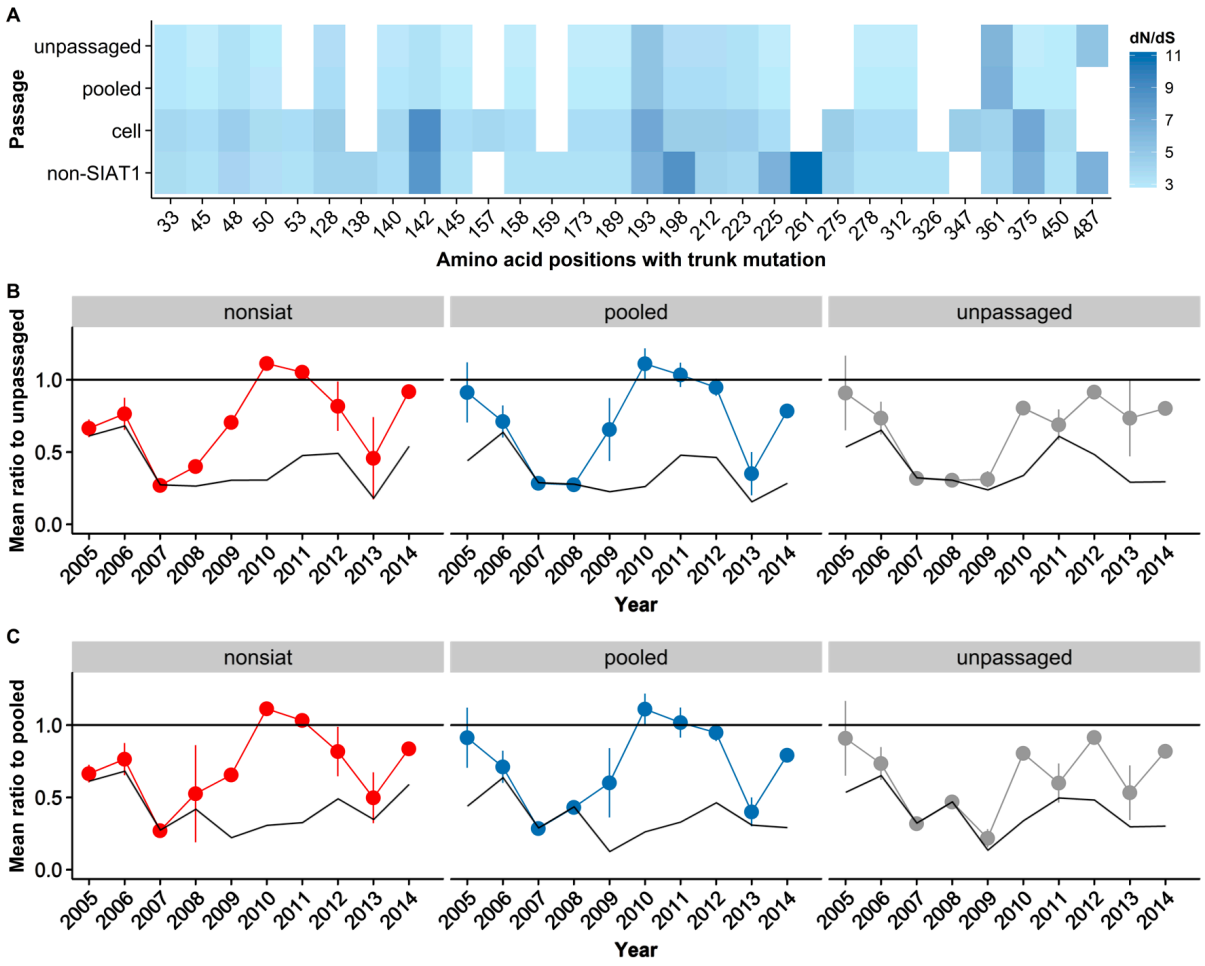
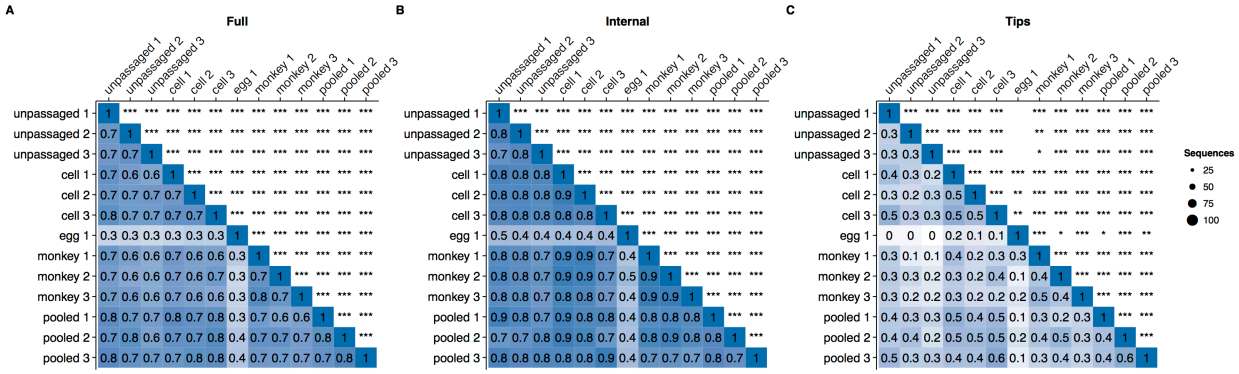


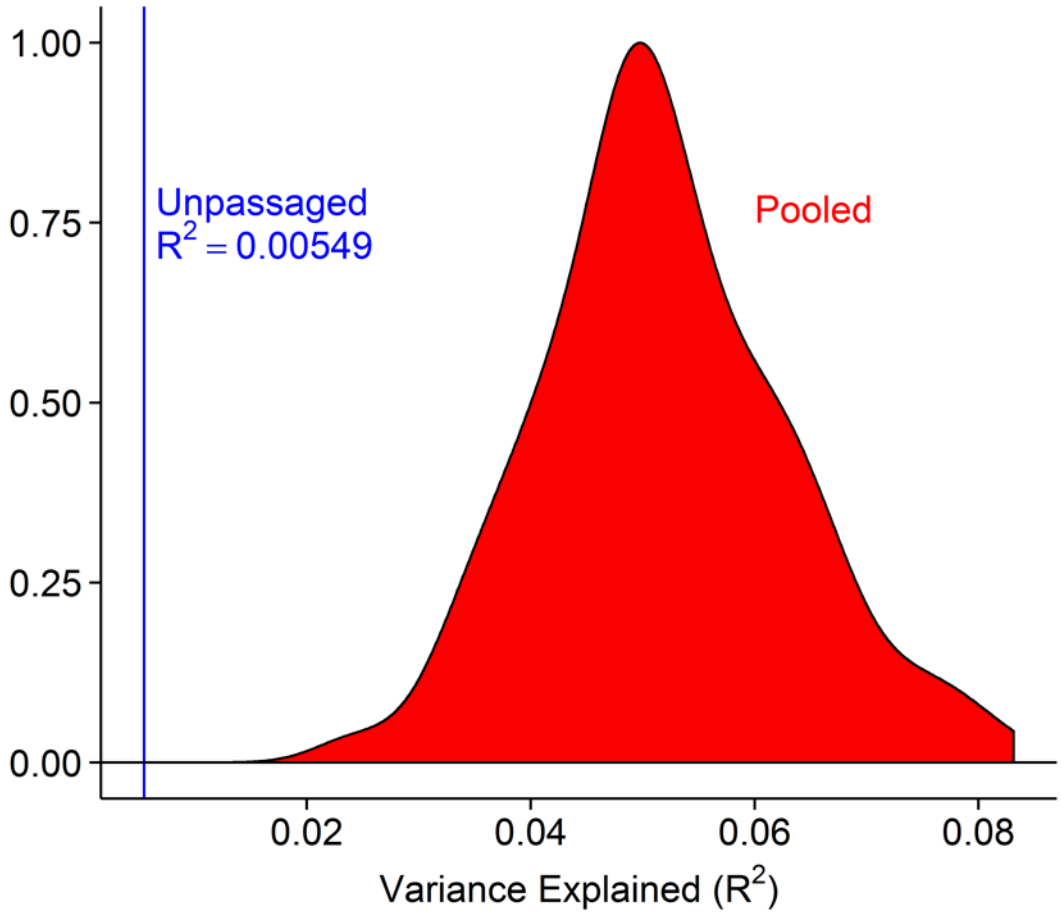
Figure 8. Passage artifacts affect trunk *dN/dS* and topology-based predictions. (A) Sitewise trunk *dN/dS* values for passage groups (*n* = 1703). Only sites with at least one non-synonymous mutation along the trunk were included. Many sites appear under positive selection only in the trunk reconstructed from passaged sequences, e.g. 159, 261, 275. (B, C) Prediction of future dominant clade by Local Branching Index (LBI) depends on passaging history. LBI was calculated for trees derived from non-SIAT1, pooled, and unpassaged 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

948 deviation of the mean, estimated from resampling (Methods). The solid black lines
949 represent the best possible prediction in each year. Predictions were evaluated relative
950 to the following year's ancestrally reconstructed root sequence obtained from (A)
951 unpassaged and (B) pooled sequences. Non-SIAT1, pooled, and unpassaged
952 sequences have divergent prediction quality across years, and on average, unpassaged
953 sequences seem to perform better than passaged or pooled sequences. (The distances
954 between dots and the solid black lines are, on average, the smallest for predictions
955 derived from unpassaged sequences.) Data used to generate this figure are available in
956 Supplementary Dataset 8 (A) and Supplementary Dataset 9 (B and C).
957

SUPPLEMENTARY FIGURES AND DATA



Supplementary Figure 1. Comparison of sitewise hemagglutinin *dN/dS* derived from size-matched samples of sequences with various passage histories, including egg amniotes. Pearson correlations between sitewise *dN/dS* values for HA sequences derived from passaged and unpassaged influenza virus collected between 2005 and 2015, randomly down-sampled to 79 sequences per passage group. (Since there were so few egg-derived sequences, each down-sampling was independently performed three times, resulting in replicates 1, 2, and 3 for each passage group.) Correlations were calculated separately for complete trees (A), internal branches only (B), and tip branches only (C). Asterisks denote significance of correlations (*0.01 ≤ *P* < 0.05, **0.001 ≤ *P* < 0.01, ****P* < 0.001). Data used to generate this figure are available in Supplementary Data 2.

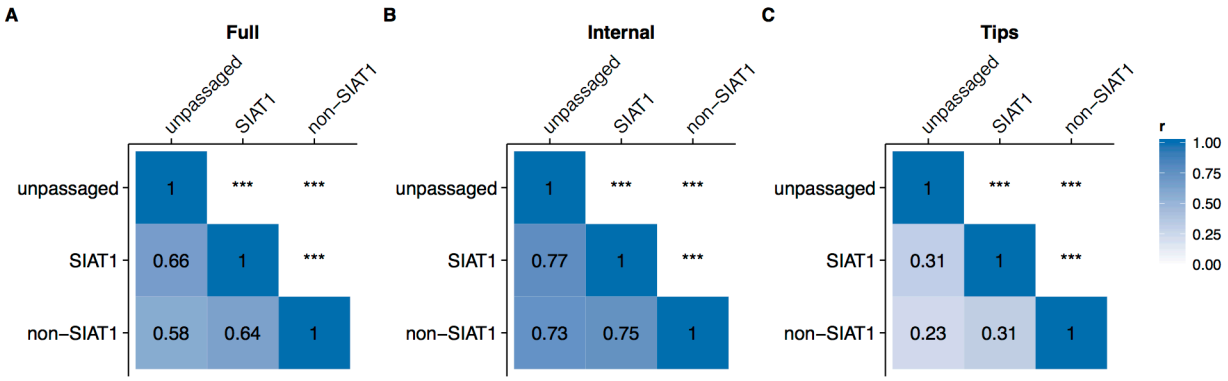


971

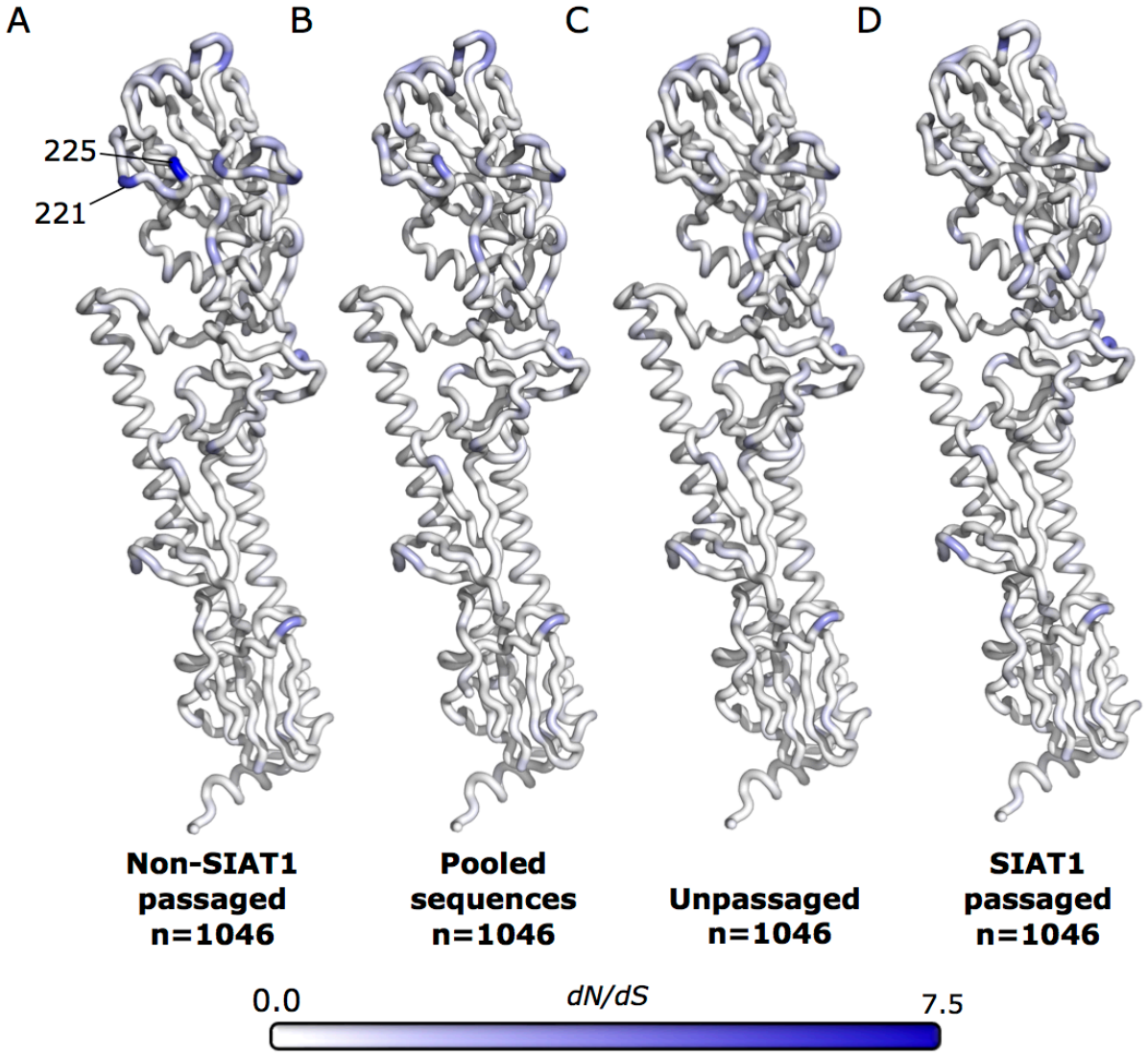
972 **Supplementary Figure 2. Differences in correlation strength not due to variance**
973 **from random sampling.** Percent variance in *dN/dS* explained by inverse distance to
974 site 224 for unpassaged and pooled sequences (*n* = 917). For pooled sequences, sets
975 of 917 sequences were randomly drawn 200 times, with replacement. For each set, we
976 calculated *dN/dS* and correlated with inverse distance. The R^2 for unpassaged
977 sequences falls well outside the distribution of R^2 values for pooled sequendes (z =
978 −4.22, $p = 1.2 \times 10^{-5}$). Data used to generate this figure are available in Supplementary
979 Data 3.

980

981



Supplementary Figure 3. Comparison of unpassaged, non-SIAT1-passaged, and SIAT1-passaged virus in 2014 only. (A–C) Pearson correlations between sitewise *dN/dS* values for the three passage groups (*n* = 249), for complete trees, internal branches only, and tip branches only, respectively. Data used to generate this figure are available in Supplementary Data 5.



988

989 **Supplementary Figure 4. *dN/dS* mapped onto the hemagglutinin structure for non-**
990 **SIAT1-passaged, pooled, unpassaged, and SIAT1 passaged sequences.** *dN/dS*
991 was mapped onto the hemagglutinin structure for (A) non-SIAT1 sequences, (B) pooled
992 sequences, (C) unpassaged sequences, and (D) SIAT1 passaged sequences collected
993 between 2005 and 2015 (*n* = 1046). *dN/dS* from pooled sequences appears to be
994 intermediate between d*N/dS* from non-SIAT1-passaged and unpassaged sequences,

995 particularly at sites 221 and 225. Data used to generate this figure are available in

996 Supplementary Data 7.

997 **Supplementary Data 1.** Data used to generate Figures 2 and 3. This file includes 1)
998 sitewise *dN/dS* values of random draws of 917 unpassaged, generic cell cultured,
999 monkey cell cultured, and the pooled group sequences collected between 2005 and
1000 2015, 2) protein and gene numbering, 3) PDB:2YP7 sequence, 4) relative solvent
1001 accessibilities of the hemagglutinin trimer, 5) linear distances to protein site 224.
1002
1003 **Supplementary Data 2.** Data used to generate Supplementary Figure 1. This file
1004 includes 1) sitewise *dN/dS* values of random draws of 97 unpassaged, generic cell
1005 cultured, egg cultured, monkey cell cultured, and the pooled group sequences collected
1006 between 2005 and 2015, 2) protein and gene numbering, 3) PDB:2YP7 sequence.
1007
1008 **Supplementary Data 3.** Data used to generate Figure Supplementary Figure 2. This
1009 file includes 1) sitewise *dN/dS* values of 500 random draws of 917 pooled sequences
1010 and one column of sitewise *dN/dS* from 917 unpassaged sequences collected between
1011 2005 and 2015, 2) protein and gene numbering, 3) PDB:2YP7 sequence 4) linear
1012 distances to protein site 224.
1013
1014 **Supplementary Data 4.** Data used to generate Figure 4. This file includes 1) sitewise
1015 *dN/dS* values of random draws of 1046 unpassaged, SIAT1, and non-SIAT1 cell culture
1016 sequences collected between 2005 and 2015, 2) protein and gene numbering, 3)
1017 PDB:2YP7 sequence, 4) relative solvent accessibilities of the hemagglutinin trimer, and
1018 5) linear distances to protein site 224.
1019

Supplementary Data 5. Data used to generate Supplementary Figure 3. This file includes 1) sitewise *dN/dS* values of random draws of 249 unpassaged, SIAT1 cultured, and non-SIAT1 cell cultured sequences collected in 2014, 2) protein and gene numbering, 3) PDB:2YP7 sequence.

Supplementary Data 6 Data used to generate Figure 5. This file includes 1) sitewise *dN/dS* values of random draws of 304 unpassaged and 304 non-SIAT1 cell cultured sequences passaged once, twice, or 3-5 times collected between 2005 and 2015, 2) protein and gene numbering, 3) PDB:2YP7 sequence.

Supplementary Data 7. Data used to generate Figures 6, 7, and Supplementary Figure 4. This file includes 1) sitewise *dN/dS* values of random draws of 1703 unpassaged and non-SIAT1 cell cultured sequences collected between 2005 and 2015, 2) protein and gene numbering, 3) PDB:2YP7 sequence, 4) sitewise inverse distance correlations.

Supplementary Data 8. Data used to generate Figure 8A. This file includes 1) trunk sitewise *dN/dS* values of all available unpassaged, SIAT1 cultured, non-SIAT1 cell, generic cell culture, monkey cell cultured, and pooled sequences collected from 2005-2015, 2) protein and gene numbering, 3) PDB:2YP7 sequence.

Supplementary Data 9. Data used to generate Figure 8B and C. This file includes output of Local-Branching-Index analyses.

Supplementary Data 10. This file includes 1) all sitewise *dN/dS* values used to generate figures, 2) protein and gene numbering, 3) PDB:2YP7 sequence, 4) relative solvent accessibilities of the hemagglutinin trimer, and 5) linear distances to protein site 224.

Supplementary File 11. GISAID acknowledgements for hemagglutinin sequences collected between 1968 and 2015.