Serial passaging causes extensive positive selection in seasonal influenza A 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 rarely sequenced directly. Instead, a majority of these isolates (~70% in 2015) are first subjected to serial 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, and it confounds evolutionary analyses of the viral sequences. We find distinct patterns of adaptation to generic (MDCK) and monkey cell culture. These patterns also dominate pooled data sets not separated by passaging type. 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 all future influenza evolutionary analyses must appropriately control for potentially confounding effects of passaging adaptations.

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 (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 serial passaging of influenza virus before sequencing.

Clinical isolates are generally passaged in culture 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 cell culture (Balish et al., 2005; Bogner et al., 2006). Influenza virus may also

67 be passaged through monkey kidney (RhMK or TMK) cell culture or injected directly into
68 egg amniotes. Alternatively, complete influenza genomes can be obtained from PCR-
69 amplified influenza samples without intermediate passaging (Katz et al., 1990; Lee et
70 al., 2013a).
71
72 Several experiments have demonstrated that influenza hemagglutinin (HA) accumulates
73 mutations following rounds of serial passaging in both cell (Ilyushina et al., 2012; Lee et
74 al., 2013b; Wyde et al., 1977) and egg culture (Robertson et al., 1993). The decreased
75 number of mutations in MDCK-based cell culture is the main argument for use of this
76 system over egg amniotes in vaccine production (Katz and Webster, 1989), with MDCK
77 cells expressing human SIAT1 having the highest fidelity to the original sequence and
78 reduced host adaptation (Hamamoto et al., 2013). Viral adaptations to eggs have
79 recently been linked to reduced vaccine efficacy (Skowronski et al., 2014; Xie et al.,
80 2015) and were implicated as potentially contributing to reduced efficacy of 2014-2015
81 seasonal H3N2 influenza vaccination in the World Health Organization's
82 recommendations for 2015-2016 vaccine strains (The World Health Organization,
83 2015). As the majority of influenza vaccines worldwide are produced in eggs, vaccine
84 strain selection is limited to virus with the ability to replicate rapidly in this system (World
85 Health Organization Global influenza surveillance network, 2011).
86
87 Although egg-passaged sequences are increasingly excluded from influenza
88 phylogenetic analysis (see e.g. the NextFlu tracker (Neher and Bedford, 2015)), due to
89 the known high host-specific substitution rates, cell culture is generally not thought to be

90 sufficiently selective to produce a discernable evolutionary signal. One of few existing
91 evolutionary analyses of passaging effects on influenza (Bush et al., 2000)
92 demonstrated that passaging causes no major changes in clade structure between egg
93 and cell passaged viruses. Moreover, several studies have recommended the use of
94 internal branches in the phylogenetic tree to reduce passaging effects in evolutionary
95 analysis of Influenza A (Bush et al., 2001; Suzuki, 2006). Another study discovered egg
96 culture to be the cause of misidentification of several sites under positive selection in
97 Influenza B (Gatherer, 2010), but this study was limited to comparing egg-cultured to
98 cell-cultured virus. As the availability of unpassaged influenza sequences has
99 dramatically increased over the past ten years, we can now perform a direct comparison
100 of passaged to circulating virus.
101
102 Here, we compare patterns of adaptation in North American seasonal H3N2 influenza
103 HA sequences derived from passaged and unpassaged virus. We divide viral
104 sequences by their passaging history, distinguishing between unpassaged clinical
105 samples, egg amniotes, RhMK (monkey) cell culture, and generic/MDCK-based cell
106 culture. For the latter, we also distinguish between virus passaged in MDCK-SIAT1 cell
107 culture (SIAT1) and in unmodified MDCK or unspecified cell culture (non-SIAT1). We
108 find clear signals of adaptation to the various passaging conditions. These signals are
109 strongly present in the tip branches of the phylogenetic trees but can also be detected in
110 internal branches. Finally, we demonstrate that the identification of antigenic escape
111 sites from sequence data has been confounded by passaging adaptations, and that the

112 exclusion of passaged sequences allows us to use sequence and structural data to
113 highlight regions involved in antigenic escape.
114
115 **RESULTS**
116 Most influenza-virus samples collected from patients are first serially passaged through
117 one or more culturing systems, prior to PCR amplification and sequencing (Figure 1A).
118 Reconstructed trees of influenza evolution contain a mixture of passage histories at
119 their tips (Figure 1B). During serial passaging, influenza genomes accumulate adaptive
120 mutations, and the effect of these mutations on evolutionary analyses of influenza
121 sequences is not well understood.
122
123 **Sitewise evolutionary rate patterns differ between passage groups**
124 To quantify any evolutionary signal that may be introduced by passaging, we
125 assembled, from the GISAID database (Bogner et al., 2006), a set of North American
126 human influenza H3N2 hemagglutinin sequences collected between 2005 and 2015.
127 We initially sorted these sequences into groups by their passage history: (1)
128 unpassaged, (2) egg-passaged, (3) generic cell-passaged, and (4) monkey cell-
129 passaged (Table 1). To assess evolutionary variation at individual sites, we calculated
130 site-specific *dN/dS* (Echave et al., 2016), using Single Likelihood Ancestor Counting
131 (SLAC). Specifically, we calculated one-rate *dN/dS* estimates, i.e., site-specific *dN*
132 values normalized by a global *dS* value (see Methods for details). In addition to
133 considering groups of sequences with specific passage histories, we also calculated
134 *dN/dS* values by pooling all sequences into one combined analysis. This pooled group

135 corresponds to a typical influenza evolutionary analysis in which passage history has
136 not been accounted for.
137
138 We first correlated the sitewise *dN/dS* values we obtained for virus sequences derived
139 from different passage histories. If passage history did not matter, then the *dN/dS*
140 values obtained from different sources should correlate strongly with each other, with *r*
141 approaching 1. Instead, we found that correlation coefficients ranged from 0.68 to 0.88,
142 depending on which specific comparison we made (Figure 2A). (In this analysis, and
143 throughout this work, we down-sampled alignments to the smallest number of
144 sequences available for any of the conditions compared, to keep the samples as
145 comparable as possible overall. The analysis of Figure 2 used n = 917 randomly drawn
146 sequences for each condition.) Unpassaged *dN/dS* correlated more strongly with cell
147 and pooled *dN/dS* (correlations of 0.77 and 0.79, respectively) than with monkey-cell
148 *dN/dS (*0.68*)*. Note that the *dN/dS* values from the pooled group, which corresponds to
149 a typical data set used in a phylogenetic analysis of influenza, more closely correlated
150 with the *dN/dS* values from the generic cell group (r = 0.87) than from the unpassaged
151 group (r = 0.79). Egg-derived sequences were excluded from this analysis due to low
152 sequence numbers (n = 79), however evolutionary rates from this condition correlated
153 particularly poorly with those of random draws of 79 unpassaged sequences
154 (Supplementary Figure 1). This result is consistent with the conclusions of (Bush et al.,
155 2000), (Suzuki, 2006), and (Gatherer, 2010) that egg-derived sequences show specific
156 adaptations not found otherwise in influenza sequences.
157

158 Because the common ancestor of any two passaged influenza viruses is a virus that
159 replicated in humans, we would expect that any adaptations introduced during
160 passaging should not extend into the internal branches of a reconstructed tree.
161 Therefore, we additionally subdivided phylogenetic trees into internal branches and tip
162 branches, and calculated site-specific *dN/dS* values separately for these two sets of
163 branches. In fact, (Bush et al., 2000) had recommended the use of internal branches to
164 reduce variation seen between egg and non-egg passaged virus. As expected, we
165 found that when *dN/dS* calculations were restricted to the internal branches, the
166 correlations between the passage groups overall increased (Figure 2B), even though
167 distinct differences between the passage groups remained. Conversely, when only
168 considering tip branches, correlations among most groups were relatively low (Figure
169 2C), with the exception of cell-passaged sequences compared to the pooled
170 sequences. This finding emphasizes once again that the pooled sample is most similar
171 to the cell-passaged sample. We conclude that different passaging histories leave
172 distinct, evolutionary signatures of adaptation to the passaging environment.
173
174 To further investigate the apparent discrepancies between *dN/dS* derived from
175 unpassaged sequences, monkey-cell passaged sequences, cell-passaged sequences,
176 and the pooled set, we compared the magnitude of the site-wise rates (Figure 2D). Cell-
177 passaged and pooled sequences had, on average, significantly inflated *dN/dS* values
178 compared to unpassaged and monkey-cell-passaged sequences in the full phylogenetic
179 tree (paired *t* test, $P = 1.5 \times 10^{-05}$ and $P = 9.1 \times 10^{-05}$, respectively) and along tip
180 branches (paired *t* test, $P = 1.8 \times 10^{-06}$ and $P = 6.3 \times 10^{-05}$, respectively). By contrast,

181 there were no significant differences between cell-passaged and pooled sequences in
182 all three cases (paired *t* test, $P = 0.26$, $P = 0.24$, and $P = 0.26$, respectively, for the full
183 tree, internal branches, and tip branches). *dN/dS* values were generally more similar
184 along internal branches, however a significant difference of *dN/dS* from cell-passaged
185 and pooled sequences relative to monkey-cell-passaged sequences remained. These
186 results demonstrate that both cell-passaged and pooled sequences show artificially
187 inflated *dN/dS* values compared to unpassaged sequences.
188
189 In aggregate, these results show that while both generic-cell-passaged sequences and
190 monkey-cell-passaged sequences yield different sitewise *dN/dS* patterns relative to
191 unpassaged sequences (Fig. 2A-C), cell passaging additionally creates inflated *dN/dS*
192 values (Fig. 2D), indicating positive adaptation to the passaging condition. At the same
193 time, *dN/dS* values derived from monkey-cell-passaged sequences are the least similar
194 to *dN/dS* from unpassaged sequences (Fig. 2A–C). The pooled group of sequences,
195 which corresponds to a typical data set used in evolutionary analyses of influenza virus,
196 describes evolutionary rates of specifically cell passaged virus and poorly matches
197 evolutionary rates of circulating influenza virus.
198
199 **Adaptations to cell and monkey-cell passage display characteristic patterns of**
200 **site variation**
201 We next asked whether adaptations to passage history were located in specific regions
202 of the HA protein. To address this question, we employed the geometric model of HA
203 evolution we recently introduced (Meyer and Wilke, 2015). For H3N2 HA, this model

204 explains over 30% of the variation in *dN/dS* using two simple physical measures, the
205 relative solvent accessibility (RSA) of individual residues in the structure (Tien et al.,
206 2013) and the inverse linear distance in 3D space from each residue to protein site 224
207 in the hemagglutinin monomer. Notably, the geometric model was previously applied to
208 a pooled sequence set including sequences of various passaging histories. To what
209 extent it carries over to sequences with specific passaging histories is not known.
210
211 We first considered the correlation between *dN/dS* and RSA (Figure 3A). We found that
212 for all passage groups, R^2 values ranged from 0.10 to 0.16 in the full tree, consistent
213 with our earlier work (Meyer and Wilke, 2015). The high congruence among R^2 values
214 for internal branches and all branches suggests that RSA imposes a pervasive selection
215 pressure on HA, independent of passaging adaptations. Thus, RSA represents a useful
216 structural measure of a persistent effect of *dN/dS* with stronger correlations in the full
217 tree and internal branches than in tip branches.
218
219 Next we considered the correlation between *dN/dS* and the inverse distance to site 224
220 (Figure 3B). In contrast to RSA, correlations here were systematically higher in tip
221 branches, suggesting a recent adaptive signal. We found virtually no correlation for
222 unpassaged sequences, while a low correlation existed for monkey-cell cultured
223 sequences and a higher correlation for cell-passaged and pooled sequences.
224 Correlations from pooled sequences mirrored cell culture correlations and persisted
225 through internal branches. Thus, the correlation of *dN/dS* with the inverse distance to
226 site 224 seems to be primarily an artifact of cell passage, even though its effect can be

227 seen along internal branches as well. As the majority of the available HA sequences are
228 cell-derived, this cell-specific signal dominates the pooled data set. Further, this cell-
229 specific signal is partially attenuated along internal branches and amplified along tip
230 branches, as we would expect from a signal caused by recent host-specific adaptation.
231 Even though this signal is a true predictor of influenza evolutionary rates for virus grown
232 in cell culture, it does not transfer to unpassaged sequences and therefore has no
233 relevance for the circulating virus. This finding serves as a strong demonstration of
234 passage history as a confounder in evolutionary analysis of hemagglutinin evolution, not
235 just for egg passage as previously demonstrated, but also for cell and monkey-cell
236 passage.
237
238 Surprisingly, the correlation we found here between *dN/dS* and inverse distance to site
239 224 for pooled sequences (R^2 = 0.067) was less than half of the value reported by
240 (Meyer and Wilke, 2015) (Fig. 3B). However, using a dataset of sequences more
241 temporally matched to that paper's dataset (2005–2014 instead of 2005–2015), we
242 recovered the previously seen higher correlation. This finding suggests that there is
243 some feature in the additional 2015 sequences that changes the pooled dataset's
244 relationship with inverse distance to site 224. In 2015, unpassaged and SIAT1
245 sequenced each doubled in number compared to in 2014, while the number of non-
246 SIAT1 cell cultured sequences dropped dramatically (Table 2). Therefore, we next
247 investigated whether the drop in correlation from 2014 to 2015 could be attributed to the
248 recent reduction in cell culture using non-SIAT1 cells.
249

250 **There is little signal of adaptation to passage in SIAT1 cells**

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

257

258 The SIAT1 cell system, which overexpresses human-like 6-linked sialic acids over
259 native 3-linked sialic acids (Matrosovich et al., 2003), is currently the dominant system
260 for serial passaging of influenza virus. Approximately half of the 2015 influenza
261 sequences currently available from GISAID derive from serial passaging through SIAT1
262 cells. Experimental analysis of SIAT1 demonstrates improved sequence fidelity and
263 reduced positive selection over unmodified MDCK cell culture (Hamamoto et al., 2013;
264 Oh et al., 2008). We sought to determine if the apparently cell-culture-specific
265 correlation of site-wise evolutionary rates and inverse distance to site 224 extended to
266 SIAT1 cell culture. To compare cell-culture varieties, we created sample-size matched
267 groups of non-SIAT1 cell culture, SIAT1 cell culture, and unpassaged sequences
268 collected between 2005 and 2015 (n = 1046), excluding sequences that had been
269 passaged through both a non-SIAT1 and a SIAT1 cell culture.

270

271 All groups showed similar correlations between *dN/dS* and RSA, regardless of whether
272 *dN/dS* was calculated for the entire tree, for internal branches only, or for tip branches

273 only (Figure 4A). By contrast, inverse distance to site 224 uniquely correlated with
274 *dN/dS* from non-SIAT1-cultured virus (Figure 4B). This effect was the strongest along
275 tip branches (R^2 = 0.139), but it was almost as strong along the entire tree (R^2 = 0.129).
276 The correlation was reduced, though still significant, among internal branches (R^2 =
277 0.075). Thus, we conclude that the correlation between *dN/dS* and the inverse distance
278 to site 224 (Meyer and Wilke, 2015) represents a unique signal of adaptation to
279 passaging in non-SIAT1 cells. In our previous analysis (Meyer and Wilke, 2015), a non-
280 SIAT1-specific signal completely dominated our evolutionary rate models, due to use of
281 a standard, pooled data set mainly composed of sequences passaged in non-SIAT1
282 cells. In our new analysis (Figure 3B), the high correlation of non-SIAT1 cell *dN/dS* with
283 inverse distance to site 224 is suppressed in the pooled condition, because the number
284 of unpassaged and SIAT1-passaged sequences grew substantially in 2015. This
285 difference in sample composition explains the lower than expected correlations in
286 Figure 3B for pooled *dN/dS*.
287
288 When considering all branches in the phylogenetic tree, we found that *dN/dS* values
289 were significantly inflated in sequences passaged in non-SIAT1 cells compared to both
290 unpassaged and SIAT1-passaged sequences (paired *t* test, $P = 5.05 \times 10^6$ and $P = 6.94$
291 $\times 10^8$, respectively, Figure 4C), whereas unpassaged and SIAT1-passaged sequences
292 showed no significant increase (Figure 4C). Unpassaged and non-SIAT1-passaged
293 sequences showed significant differences along internal branches (paired *t* test, P =
294 0.036) and tip branches as well (paired *t* test, $P = 2.03 \times 10^6$, Figure 4C). Thus, virus

295 amplified in non-SIAT1 cell culture measurably adapts to this non-human host, and
296 these adaptations can significantly confound downstream evolutionary analyses.
297
298 As these three conditions are somewhat temporally separated (most non-SIAT1 cell
299 culture sequences are pre-2015, and most unpassaged and SIAT1 culture sequences
300 are post-2014), we controlled for season-to-season variation by drawing 249 sequences
301 from each group from 2014. First, we again considered site-wise *dN/dS* correlations
302 among passaging groups, and we found that overall, unpassaged and SIAT1-passaged
303 sequences appeared the most similar (Supplementary Figure 2A–C). However, both
304 SIAT1 and non-SIAT1 showed *dN/dS* values that were inflated over *dN/dS* in
305 unpassaged sequences when considering the full tree (paired *t* test, $P = 0.029$ and $P =$
306 0.0005, respectively, Supplementary Figure 2D), although only non-SIAT1 *dN/dS* was
307 significantly inflated in tip branches (paired *t* test, $P = 0.0008$, Supplementary Figure
308 2D). (No significant difference was seen along internal branches.) Notably, in this more
309 controlled comparison of SIAT1 cell culture to unpassaged sequences from the same
310 year, we observed a significant difference in *dN/dS* between these conditions,
311 suggesting that at least minor passaging artifacts remain after SIAT1 passaging.
312
313 **Evolutionary variation in sequences from unpassaged virus predicts regions**
314 **involved in antigenic escape**
315 The preceding results might suggest that the inverse distance metric we previously
316 proposed (Meyer and Wilke, 2015) only captures effects of adaptation to non-SIAT1 cell
317 culture. However, this is not necessarily the case. Importantly, inverse distance needs

318 to be calculated relative to a specific reference point. We previously used site 224 as
319 the reference point because it yielded the highest correlation for the data set we
320 analyzed then. For a different data set, one that doesn't carry the signal of adaptation to
321 non-SIAT1 cell culture, a different reference point may be more appropriate.
322
323 We thus repeated the analysis of (Meyer and Wilke, 2015) for a size matched sample of
324 1703 sequences from both non-SIAT1 cell passaged and unpassaged virus collected
325 between 2005 and 2015 (Figure 5). In brief, for each possible reference site in the
326 hemagglutinin structure, we measured the inverse distance in 3D space from that site to
327 every other site in the structure. We then correlated the inverse distances with the
328 *dN/dS* values at each site, resulting in one correlation coefficient per reference site.
329 Finally, we mapped these correlation coefficients onto the HA structure, coloring each
330 reference site by its associated correlation coefficient. If inverse distances measured
331 from a particular reference amino acid have higher correlation with the sitewise *dN/dS*
332 values, then this reference site will appear highlighted on the structure.
333
334 For non-SIAT1-passaged virus, this analysis recovered the finding of (Meyer and Wilke,
335 2015) that the loop containing site 224 appeared strongly highlighted (Figure 5A).
336 However, this signal was entirely absent in unpassaged virus (Figure 5B), with no sites
337 in that loop working well as a reference point. These results suggest that this loop is
338 specifically involved in adaptation of hemagglutinin to non-SIAT1 cell culture, explaining
339 the non-SIAT1-specific signal shown in Figure 4A. Thus, the inverse distance metric is
340 useful for differentiating regions of selection particular to different experimental groups.

341

342 (Meyer and Wilke, 2015) had concluded that sites under positive selection differed from
343 sites involved in immune escape. Here, we have found that the origin of this positive
344 selection is adaptation to the non-human passaging host, not immune escape in or
345 adaptation to humans. Therefore, we next asked what residual patterns of positive
346 selection remained once the adaptation to non-SIAT1 cells was removed. Even though
347 site-wise correlations are relatively low for unpassaged virus compared to the ones
348 observed for non-SIAT1-passaged virus, we could still recover relevant patterns of HA
349 adaptation after rescaling our coloring. In particular, we found that sites opposite to the
350 loop containing site 224 lit up in our analysis of unpassaged sequences (Figure 6A).
351 Sites in this region are known to be involved in antigenic escape. In fact, many of the
352 highlighted regions contain experimentally determined antigenic sites (Koel et al., 2013)
353 and/or the sites determined to be responsible for the antigenic shift in the 2014/2015
354 seasonal flu (Chambers et al., 2015) (Table 2). We found a similar pattern of
355 concordance with antigenic sites when mapping *dN/dS* values directly onto the structure
356 (Figure 6B). The inverse-distance correlations, however, performed better at identifying
357 antigenic sites than did raw *dN/dS* values. When considering the 90th percentile (top
358 10% highest scored sites) by either metric, the inverse-distance correlations recovered
359 7 of 8 sites while *dN/dS* alone recovered only 2 of 8 sites (Table 2).

360

361 **DISCUSSION**

362 We have found that serial passaging of influenza virus introduces a measurable signal
363 of adaptation into the evolutionary analysis of natural influenza sequences. There are

364 unique, characteristic patterns of adaptation to egg passage, monkey cell passage, and
365 non-SIAT1 cell passage. Monkey cell-derived sequences show different molecule-wide
366 evolutionary rate patterns, even though they show no *dN/dS* inflation when compared
367 with unpassaged sequences. Non-SIAT1 cell-derived sequences instead display both
368 *dN/dS* inflation and a hotspot of positive selection in a loop underneath the sialic-acid
369 binding region. This hotspot has been previously noted (Meyer and Wilke, 2015) but no
370 explanation for its origin was available. Further, we have found that virus passaged in
371 SIAT1 cells seems to accumulate only minor passaging artifacts. Throughout our
372 analyses, we have found limited utility to subdividing phylogenetic trees into internal and
373 terminal branches. While signals of passage adaptation are consistently elevated along
374 terminal branches and attenuated along internal branches, evolutionary rates along
375 internal branches remain confounded by passaging artifacts. Finally, we could
376 accurately recover the experimentally determined antigenic regions of hemagglutinin
377 from evolutionary-rate analysis by using a data set consisting of only unpassaged viral
378 sequences.
379
380 Previous studies (Bush et al., 2001; Suzuki, 2006) have suggested the use of internal
381 branches to alleviate passage adaptations. However, we have found here that this
382 strategy is insufficient, because the evolutionary signal of passage adaptations can
383 often be detected along internal branches. This finding seems counterintuitive, as
384 internal nodes should exclusively represent human-adapted virus. We suggest that
385 passaging adaptations in internal branches may be caused by convergent evolution; if
386 different clinical isolates converge onto the same adaptive mutations under passaging,

387 then these mutations may incorrectly be placed along internal branches under
388 phylogenetic tree reconstruction. Additionally, although the use of only internal branches
389 removes some differences between the passage groups, the exclusion of terminal
390 sequences can obscure recent natural adaptations and thus obscure actual sites under
391 positive selection. Therefore, analysis of internal branches is not only insufficient for
392 eliminating artifacts from passaging adaptations but also suboptimal for detecting
393 positive selection in seasonal H3N2 influenza.
394
395 The safest route to avoid passaging artifacts is to limit sequence data sets to only
396 unpassaged virus, although this approach limits sequence numbers. The human-like 6-
397 linked sialic acids in SIAT1 (Matrosovich et al., 2003) greatly reduce observed cell
398 culture-specific adaptations, particularly in the loop of hemagglutinin which contains site
399 224. This lack of selection concords with multiple experiments finding low levels of
400 adaptation in this cell line (Hamamoto et al., 2013; Oh et al., 2008). As our analysis only
401 detected minor differences between unpassaged and SIAT1 passaged virus, we posit
402 that this passage condition is an acceptable substitute for unpassaged clinical samples.
403 Even so, our findings do not preclude the existence of SIAT1-specific adaptations that
404 may confound specific analyses.
405
406 Although the majority of the sequences from the year 2015 are SIAT1-passaged or
407 unpassaged, several hundred sequences from that year derive from monkey cell
408 culture. The use of monkey cell culture has surged in 2014 and 2015 compared to
409 previous years. We recommend that these recently collected sequences be excluded

410 from influenza rate analysis, in favor of the majority of unpassaged and SIAT1-
411 passaged sequences. As passaging is a useful and cost effective method for
412 amplification of clinically collected virus, unpassaged viral sequences are unlikely to
413 completely dominate influenza sequence databases in the near future. However, new
414 human epithelial cell culture systems for influenza passaging, as in (Ilyushina et al.,
415 2012), could soon provide an ideal system that both amplifies virus and protects it from
416 non-human selective pressures.
417
418 Passage history should routinely be considered as a potential confounding variable in
419 future analyses of influenza evolutionary rates. Future studies should be checked
420 against unpassaged samples to ensure that conclusions are not based on adaptation to
421 non-human hosts. We recommend the exclusion of viral sequences which derive from
422 serial passage in egg amniotes, monkey kidney cell culture, and any unspecified cell
423 culture. Prior work that did not consider passaging history may likely have been
424 confounded by passaging adaptations. In particular, we suggest that the evolutionary
425 markers of influenza virus determined by (Belanov et al., 2015) be reevaluated to
426 ensure these sites are not artifacts of viral passaging. Similarly, many of the earlier
427 studies performing site-specific evolutionary analysis of HA, such as (Bush et al., 1999;
428 Meyer and Wilke, 2015, 2013; Pan and Deem, 2011; Shih et al., 2007; Suzuki, 2008,
429 2006; Tusche et al., 2012), likely contain some conclusions that can be traced back to
430 passaging artifacts. Additionally, even though passage artifacts do not appear to be
431 sufficiently strong to affect clade-structure reconstruction (Bush et al., 2000), they do
432 have the potential to cause artificially long branch lengths, due to *dN/dS* inflation, or

433 misplaced branches, due to convergent evolution under passaging. Thus, future
434 phylogenetic predictive models of influenza fitness and antigenicity, as in (Łuksza and
435 Lässig, 2014), (Neher et al., 2014), and (Bedford et al., 2014), should too be checked
436 for the presence of passage-related signals. Finally, while it is beyond the scope of this
437 work to investigate passage history effects in other viruses, we suspect that passage-
438 derived artifacts could be a factor in their phylogenetic analyses as well. The use of data
439 sets free of passage adaptations will likely bring computational predictions of influenza
440 positive selection more in line with corresponding experimental results.
441
442 Sequences without passage annotations are inadequate for reliable evolutionary
443 analysis of influenza virus. Yet, passage annotations are often completely missing from
444 strain information, and, when present, are often inconsistent; there is currently no
445 standardized language to represent number and type of serial passage. We note,
446 however, that passage annotations from the 2015 season are greatly improved when
447 compared to previous seasons. Several major influenza repositories, including the
448 Influenza Research Database (Squires et al., 2012) and the NCBI Influenza Virus
449 Resource (Bao et al., 2008), do not provide any passaging annotations at all.
450 Additionally, passage history is not required for new sequence submissions to the NCBI
451 Genbank (Benson et al., 2012). The EpiFlu database maintained by the Global Initiative
452 for Sharing Avian Influenza Data (GISAID) (Bogner et al., 2006) and OpenFluDB
453 (Liechti et al., 2010), however, stand apart by providing passage history annotations for
454 the majority of their sequences. Of these, only the OpenFluDB repository allows filtering
455 of sequences by passage history during data download. Our results demonstrate the

456 strength of passaging artifacts in evolutionary analysis of influenza. The lack of a
457 universal standard for annotation of viral passage histories and a universal standard for
458 serial passage experimental conditions complicate the analysis and mitigation of
459 passaging effects.

460 **METHODS**

461 **Influenza sequence data**

462 Non-laboratory strain H3N2 hemagglutinin (HA) sequences collected in North America
463 were downloaded from The Global Initiative for Sharing Avian Influenza Data (GISAID)
464 (Bogner et al., 2006) for the 1968–2015 influenza seasons. Non-complete HA
465 sequences were excluded. Sequences were trimmed to open reading frames, filtered to
466 remove redundancies, and aligned by translation-alignment-back-translation and
467 MAFFT (Katoh and Standley, 2013). Sequence headers of FASTA files were
468 standardized into an uppercase text format with non-alphanumeric characters replaced
469 by underscores. As H3N2 strains have experienced no persistent insertion or deletion
470 events, we deleted sequences which introduced gaps to the alignment. To ascertain
471 overall data quality, we built a phylogenetic tree of the entire sequence set (using
472 FastTree 2.0 (Price et al., 2010)) and checked for any abnormal clades or other
473 unexpected tree features. We found one abnormal clade of approximately 20
474 sequences with an exceptionally long branch length (> 0.01) and removed the
475 sequences in that clade from further analysis. Our final data set consisted of 6873
476 sequences from 2005-2015 as well as an outgroup of 45 sequences from 1968–1977
477 (not considered for further analysis). We did not consider sequences collected from
478 1978-2004.

479

480 **Identification of passage history and evolutionary-rate calculations**

481 We divided sequences into groups by their passage history annotation and collection
482 year, determining passage history by parsing with regular expressions for key words in
483 FASTA headers (Table 1). We classified 1133 sequences with indeterminate or missing
484 passage histories, or passage through multiple categories of hosts (i.e. both egg and
485 cell), as “other”. The final data sets for individual passage groups contained between 79
486 and 3041 sequences (Table 1).

487

488 We next constructed phylogenetic trees for each passage group as well as one tree for
489 a pooled data set combining all individual passage groups and other sequences. All
490 phylogenetic trees were constructed using FastTree 2.0 (Price et al., 2010). We
491 calculated site-specific *dN/dS* values using a one-rate SLAC (Single-Likelihood
492 Ancestor Counting) model implemented in HyPhy (Pond et al., 2005). One rate models,
493 which fit a site-specific *dN* and a global *dS*, yield more accurate estimates than two-rate
494 models and hence are preferred (Spielman et al., 2015). Among different one-rate, site-
495 specific models, SLAC performs nearly identical to other approaches, and it was chosen
496 here due to its speed and ease of extracting *dN/dS* estimates along internal and tip
497 branches. To obtain branch-specific estimates, we extracted the *dN/dS* values
498 calculated by the SLAC algorithm at internal and tip branches.

499

500 We chose sequences from 2005-2015 as our sample set due the low number of
501 available sequences prior to this period. As *dN/dS* estimates can be confounded by

502 sample size (Spielman et al., 2015), we sought to limit this effect by down-sampling
503 each experimental set to match the number of sequences in the smallest group being
504 considered in a particular analysis (Table 1). To reduce season-to-season variation in
505 the comparison of unpassaged, SIAT1, and non-SIAT1 cell culture, we performed one
506 analysis with sequences from only 2014, which is the year that maximizes sequences
507 available from all three conditions (n = 249 each).
508

509 **Geometric analysis of *dN/dS* distributions**

510 For each site i in HA, we computed the correlation of *dN/dS* at every site $j \neq i$ with the
511 inverse Euclidian distance between j and i in the 3D crystal structure of the protein. This
512 method is discussed in detail in (Meyer and Wilke, 2015). This correlation is then color-
513 mapped onto the reference site. Sites spatially closest to positively selected regions in
514 the protein have the highest correlation in this analysis. Thus, this approach allows us to
515 visualize regions of increased positive selection. We processed the HA PDB structure
516 as discussed in (Meyer and Wilke, 2015), and we provide a renumbered and formatted
517 H3N2 structure derived from PDB ID 2YP7 (Lin et al., 2012) with our data analysis code
518 (see below).
519

520 **Statistical analysis and data availability**

521 Raw influenza sequences used in this analysis are available for download from GISAID
522 (http://gisaid.org) using the parameters "North America", "H3N2", "1976 – 2015.
523 Acknowledgements for sequences used in this study are available in Supplementary
524 File 1. The complete, processed data set used in our statistical analysis is available in

525 Supplementary Dataset 6, including protein and gene numbering, computed
526 evolutionary rates, relative solvent accessibility for the hemagglutinin trimer, and
527 sitewise distance to protein site 224. Relative solvent accessibility of the hemagglutinin
528 trimer was taken from (Meyer and Wilke, 2015). Site-wise distances between all amino
529 acids in the HA structure PDBID:2YP7 were recalculated as in (Meyer and Wilke, 2015).
530 Statistical analysis was performed using R (Ihaka and Gentleman, 1996), and all graph
531 figures drawn with the R package ggplot2 (Wickham, 2009). Throughout this work, *
532 denotes a significance of $0.01 \leq P < 0.05$, ** denotes a significance of $0.01 \leq P < 0.05$,
533 and *** denotes a significance of $P < 0.001$.
534
535 Linear models between sitewise *dN/dS* and RSA or inverse distance were fit using the
536 lm() function in R. Correlations were calculated using the R function cor() and
537 significance determined using cor.test().
538
539 Our entire analysis pipeline, instructions for running analyses and raw data (except
540 initial sequence data per the GISAID user agreement) are available at the following
541 Github project repository:
542 https://github.com/wilkelab/influenza_H3N2_passaging.
543
544
545
546
547

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 COW.

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.

569 **REFERENCES**

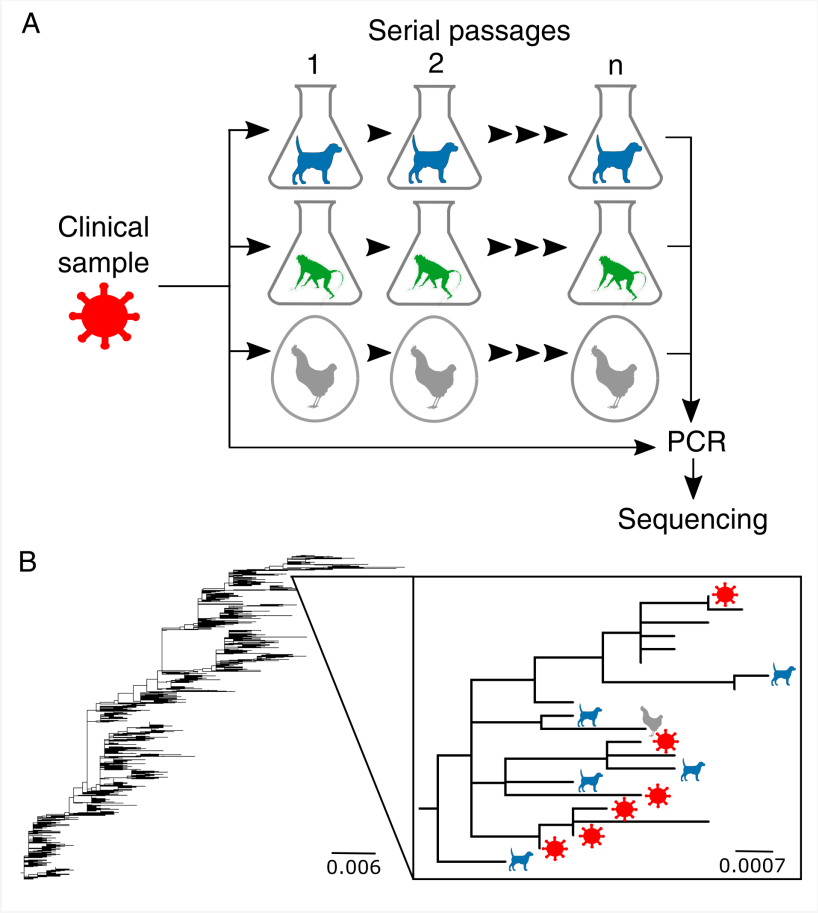
570 Balish, A.L., Katz, J.M., Klimov, A.I., 2005. Influenza: Propagation, Quantification, and Storage,
571 in: Current Protocols in Microbiology. John Wiley & Sons, Inc.
572 Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., Lipman,
573 D., 2008. The Influenza Virus Resource at the National Center for Biotechnology
574 Information. J. Virol. 82, 596–601. doi:10.1128/JVI.02005-07
575 Bedford, T., Suchard, M.A., Lemey, P., Dudas, G., Gregory, V., Hay, A.J., McCauley, J.W.,
576 Russell, C.A., Smith, D.J., Rambaut, A., 2014. Integrating influenza antigenic dynamics
577 with molecular evolution. eLife 3, e01914. doi:10.7554/eLife.01914
578 Belanov, S.S., Bychkov, D., Benner, C., Ripatti, S., Ojala, T., Kankainen, M., Lee, H.K., Tang,
579 J.W.-T., Kainov, D.E., 2015. Genome-wide analysis of evolutionary markers of human
580 influenza A(H1N1)pdm09 and A(H3N2) viruses may guide selection of vaccine strain
581 candidates. Genome Biol. Evol. evv240. doi:10.1093/gbe/evv240
582 Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., Sayers, E.W., 2012.
583 GenBank. Nucleic Acids Res. 40, D48–D53. doi:10.1093/nar/gkr1202
584 Blackburne, B.P., Hay, A.J., Goldstein, R.A., 2008. Changing Selective Pressure during
585 Antigenic Changes in Human Influenza H3. PLoS Pathog 4, e1000058.
586 doi:10.1371/journal.ppat.1000058
587 Bogner, P., Capua, I., Lipman, D.J., Cox, N.J., others, 2006. A global initiative on sharing avian
588 flu data. Nature 442, 981–981. doi:10.1038/442981a
589 Bush, R.M., Fitch, W.M., Bender, C.A., Cox, N.J., 1999. Positive selection on the H3
590 hemagglutinin gene of human influenza virus A. Mol. Biol. Evol. 16, 1457–1465.
591 Bush, R.M., Fitch, W.M., Smith, C.B., Cox, N.J., 2001. Predicting influenza evolution: the impact
592 of terminal and egg-adapted mutations. Int. Congr. Ser. 1219, 147–153.
593 doi:10.1016/S0531-5131(01)00643-4
594 Bush, R.M., Smith, C.B., Cox, N.J., Fitch, W.M., 2000. Effects of passage history and sampling
595 bias on phylogenetic reconstruction of human influenza A evolution. Proc. Natl. Acad.
596 Sci. 97, 6974–6980. doi:10.1073/pnas.97.13.6974
597 Chambers, B.S., Parkhouse, K., Ross, T.M., Alby, K., Hensley, S.E., 2015. Identification of
598 Hemagglutinin Residues Responsible for H3N2 Antigenic Drift during the 2014–2015
599 Influenza Season. Cell Rep. 12, 1–6. doi:10.1016/j.celrep.2015.06.005
600 Echave, J., Spielman, S.J., Wilke, C.O., 2016. Causes of evolutionary rate variation among
601 protein sites. Nat. Rev. Genet. doi:10.1038/nrg.2015.18

602 Gatherer, D., 2010. Passage in egg culture is a major cause of apparent positive selection in
603 influenza B hemagglutinin. J. Med. Virol. 82, 123–127. doi:10.1002/jmv.21648
604 Hamamoto, I., Takaku, H., Tashiro, M., Yamamoto, N., 2013. High Yield Production of Influenza
605 Virus in Madin Darby Canine Kidney (MDCK) Cells with Stable Knockdown of IRF7.
606 PLoS ONE 8, e59892. doi:10.1371/journal.pone.0059892
607 Ihaka, R., Gentleman, R., 1996. R: A Language for Data Analysis and Graphics. J. Comput.
608 Graph. Stat. 5, 299–314. doi:10.2307/1390807
609 Ilyushina, N.A., Ikizler, M.R., Kawaoka, Y., Rudenko, L.G., Treanor, J.J., Subbarao, K., Wright,
610 P.F., 2012. Comparative study of influenza virus replication in MDCK cells and in
611 primary cells derived from adenoids and airway epithelium. J. Virol. 86, 11725–11734.
612 doi:10.1128/JVI.01477-12
613 Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7:
614 Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780.
615 doi:10.1093/molbev/mst010
616 Katz, J.M., Wang, M., Webster, R.G., 1990. Direct sequencing of the HA gene of influenza
617 (H3N2) virus in original clinical samples reveals sequence identity with mammalian cell-
618 grown virus. J. Virol. 64, 1808–1811.
619 Katz, J.M., Webster, R.G., 1989. Efficacy of Inactivated Influenza A Virus (H3N2) Vaccines
620 Grown in Mammalian Cells or Embryonated Eggs. J. Infect. Dis. 160, 191–198.
621 doi:10.1093/infdis/160.2.191
622 Koel, B.F., Burke, D.F., Bestebroer, T.M., van der Vliet, S., Zondag, G.C.M., Vervaet, G.,
623 Skepner, E., Lewis, N.S., Spronken, M.I.J., Russell, C.A., Eropkin, M.Y., Hurt, A.C., Barr,
624 I.G., de Jong, J.C., Rimmelzwaan, G.F., Osterhaus, A.D.M.E., Fouchier, R.A.M., Smith,
625 D.J., 2013. Substitutions near the receptor binding site determine major antigenic
626 change during influenza virus evolution. Science 342, 976–979.
627 doi:10.1126/science.1244730
628 Koelle, K., Cobey, S., Grenfell, B., Pascual, M., 2006. Epochal evolution shapes the
629 phylodynamics of interpandemic influenza A (H3N2) in humans. Science 314, 1898–
630 1903. doi:10.1126/science.1132745
631 Kumar, S., Henrickson, K.J., 2012. Update on Influenza Diagnostics: Lessons from the Novel
632 H1N1 Influenza A Pandemic. Clin. Microbiol. Rev. 25, 344–361.
633 doi:10.1128/CMR.05016-11

634 Lee, H.K., Tang, J.W.-T., Kong, D.H.-L., Koay, E.S.-C., 2013a. Simplified Large-Scale Sanger
635 Genome Sequencing for Influenza A/H3N2 Virus. PLoS ONE 8.
636 doi:10.1371/journal.pone.0064785
637 Lee, H.K., Tang, J.W.-T., Kong, D.H.-L., Loh, T.P., Chiang, D.K.-L., Lam, T.T.-Y., Koay, E.S.-C.,
638 2013b. Comparison of Mutation Patterns in Full-Genome A/H3N2 Influenza Sequences
639 Obtained Directly from Clinical Samples and the Same Samples after a Single MDCK
640 Passage. PLoS ONE 8, e79252. doi:10.1371/journal.pone.0079252
641 Liechti, R., Gleizes, A., Kuznetsov, D., Bougueleret, L., Mercier, P.L., Bairoch, A., Xenarios, I.,
642 2010. OpenFluDB, a database for human and animal influenza virus. Database 2010,
643 baq004. doi:10.1093/database/baq004
644 Lin, Y.P., Xiong, X., Wharton, S.A., Martin, S.R., Coombs, P.J., Vachieri, S.G., Christodoulou,
645 E., Walker, P.A., Liu, J., Skehel, J.J., Gamblin, S.J., Hay, A.J., Daniels, R.S., McCauley,
646 J.W., 2012. Evolution of the receptor binding properties of the influenza A(H3N2)
647 hemagglutinin. Proc. Natl. Acad. Sci. 109, 21474–21479. doi:10.1073/pnas.1218841110
648 Łuksza, M., Lässig, M., 2014. A predictive fitness model for influenza. Nature 507, 57–61.
649 doi:10.1038/nature13087
650 Matrosovich, M., Matrosovich, T., Carr, J., Roberts, N.A., Klenk, H.-D., 2003. Overexpression of
651 the alpha-2,6-sialyltransferase in MDCK cells increases influenza virus sensitivity to
652 neuraminidase inhibitors. J. Virol. 77, 8418–8425.
653 Meyer, A.G., Wilke, C.O., 2015. Geometric Constraints Dominate the Antigenic Evolution of
654 Influenza H3N2 Hemagglutinin. PLoS Pathog. 11. doi:10.1371/journal.ppat.1004940
655 Meyer, A.G., Wilke, C.O., 2013. Integrating Sequence Variation and Protein Structure to Identify
656 Sites under Selection. Mol. Biol. Evol. 30, 36–44. doi:10.1093/molbev/mss217
657 Neher, R.A., Bedford, T., 2015. nextflu: real-time tracking of seasonal influenza virus evolution
658 in humans. Bioinformatics btv381. doi:10.1093/bioinformatics/btv381
659 Neher, R.A., Russell, C.A., Shraiman, B.I., 2014. Predicting evolution from the shape of
660 genealogical trees. eLife 3, e03568. doi:10.7554/eLife.03568
661 Nelson, M.I., Simonsen, L., Viboud, C., Miller, M.A., Taylor, J., George, K.S., Griesemer, S.B.,
662 Ghedin, E., Sengamalay, N.A., Spiro, D.J., Volkov, I., Grenfell, B.T., Lipman, D.J.,
663 Taubenberger, J.K., Holmes, E.C., 2006. Stochastic Processes Are Key Determinants of
664 Short-Term Evolution in Influenza A Virus. PLoS Pathog 2, e125.
665 doi:10.1371/journal.ppat.0020125

666 Oh, D.Y., Barr, I.G., Mosse, J.A., Laurie, K.L., 2008. MDCK-SIAT1 cells show improved isolation
667 rates for recent human influenza viruses compared to conventional MDCK cells. J. Clin.
668 Microbiol. 46, 2189–2194. doi:10.1128/JCM.00398-08
669 Pan, K., Deem, M.W., 2011. Quantifying selection and diversity in viruses by entropy methods,
670 with application to the haemagglutinin of H3N2 influenza. J. R. Soc. Interface 8, 1644–
671 1653. doi:10.1098/rsif.2011.0105
672 Pond, S.L.K., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies.
673 Bioinformatics 21, 676–679. doi:10.1093/bioinformatics/bti079
674 Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood
675 Trees for Large Alignments. PLoS ONE 5, e9490. doi:10.1371/journal.pone.0009490
676 Robertson, J.S., Nicolson, C., Major, D., Robertson, E.W., Wood, J.M., 1993. The role of
677 amniotic passage in the egg-adaptation of human influenza virus is revealed by
678 haemagglutinin sequence analyses. J. Gen. Virol. 74 (Pt 10), 2047–2051.
679 doi:10.1099/0022-1317-74-10-2047
680 Shih, A.C.-C., Hsiao, T.-C., Ho, M.-S., Li, W.-H., 2007. Simultaneous amino acid substitutions at
681 antigenic sites drive influenza A hemagglutinin evolution. Proc. Natl. Acad. Sci. 104,
682 6283–6288. doi:10.1073/pnas.0701396104
683 Skowronski, D.M., Janjua, N.Z., De Serres, G., Sabaiduc, S., Eshaghi, A., Dickinson, J.A.,
684 Fonseca, K., Winter, A.-L., Gubbay, J.B., Krajden, M., Petric, M., Charest, H., Bastien,
685 N., Kwindt, T.L., Mahmud, S.M., Van Caeseele, P., Li, Y., 2014. Low 2012–13 Influenza
686 Vaccine Effectiveness Associated with Mutation in the Egg-Adapted H3N2 Vaccine
687 Strain Not Antigenic Drift in Circulating Viruses. PLoS ONE 9, e92153.
688 doi:10.1371/journal.pone.0092153
689 Spielman, S., Wan, S., Wilke, C.O., 2015. One-rate models outperform two-rate models in site-
690 specific dN/dS estimation. bioRxiv 032805. doi:10.1101/032805
691 Squires, R.B., Noronha, J., Hunt, V., García-Sastre, A., Macken, C., Baumgarth, N., Suarez, D.,
692 Pickett, B.E., Zhang, Y., Larsen, C.N., Ramsey, A., Zhou, L., Zaremba, S., Kumar, S.,
693 Deitrich, J., Klem, E., Scheuermann, R.H., 2012. Influenza Research Database: an
694 integrated bioinformatics resource for influenza research and surveillance. Influenza
695 Other Respir. Viruses 6, 404–416. doi:10.1111/j.1750-2659.2011.00331.x
696 Stöhr, K., Bucher, D., Colgate, T., Wood, J., 2012. Influenza Virus Surveillance, Vaccine Strain
697 Selection, and Manufacture, in: Kawaoka, Y., Neumann, G. (Eds.), Influenza Virus,
698 Methods in Molecular Biology. Humana Press, pp. 147–162.

699 Suzuki, Y., 2008. Positive selection operates continuously on hemagglutinin during evolution of
700 H3N2 human influenza A virus. Gene 427, 111–116. doi:10.1016/j.gene.2008.09.012
701 Suzuki, Y., 2006. Natural selection on the influenza virus genome. Mol. Biol. Evol. 23, 1902–
702 1911. doi:10.1093/molbev/msl050
703 The World Health Organization, 2015. Recommended composition of influenza virus vaccines
704 for use in the 2015- 2016 northern hemisphere influenza season.
705 Tien, M.Z., Meyer, A.G., Sydykova, D.K., Spielman, S.J., Wilke, C.O., 2013. Maximum allowed
706 solvent accessibilites of residues in proteins. PloS One 8, e80635.
707 doi:10.1371/journal.pone.0080635
708 Tusche, C., Steinbrück, L., McHardy, A.C., 2012. Detecting patches of protein sites of influenza
709 A viruses under positive selection. Mol. Biol. Evol. 29, 2063–2071.
710 doi:10.1093/molbev/mss095
711 WHO Writing Group, Ampofo, W.K., Baylor, N., Cobey, S., Cox, N.J., Daves, S., Edwards, S.,
712 Ferguson, N., Grohmann, G., Hay, A., Katz, J., Kullabutr, K., Lambert, L., Levandowski,
713 R., Mishra, A.C., Monto, A., Siqueira, M., Tashiro, M., Waddell, A.L., Wairagkar, N.,
714 Wood, J., Zambon, M., Zhang, W., 2012. Improving influenza vaccine virus
715 selectionReport of a WHO informal consultation held at WHO headquarters, Geneva,
716 Switzerland, 14–16 June 2010. Influenza Other Respir. Viruses 6, 142–152.
717 doi:10.1111/j.1750-2659.2011.00277.x
718 Wickham, H., 2009. ggplot2: elegant graphics for data analysis. Springer New York.
719 Wolf, Y.I., Viboud, C., Holmes, E.C., Koonin, E.V., Lipman, D.J., 2006. Long intervals of stasis
720 punctuated by bursts of positive selection in the seasonal evolution of influenza A virus.
721 Biol. Direct 1, 34. doi:10.1186/1745-6150-1-34
722 World Health Organization Global influenza surveillance network, 2011. Manual for the
723 laboratory diagnosis and virological surveillance of influenza. Geneva, Switzerland.
724 Wyde, P.R., Couch, R.B., Mackler, B.F., Cate, T.R., Levy, B.M., 1977. Effects of Low- and High-
725 Passage Influenza Virus Infection in Normal and Nude Mice. Infect. Immun. 15, 221–
726 229.
727 Xie, H., Wan, X.-F., Ye, Z., Plant, E.P., Zhao, Y., Xu, Y., Li, X., Finch, C., Zhao, N., Kawano, T.,
728 Zoueva, O., Chiang, M.-J., Jing, X., Lin, Z., Zhang, A., Zhu, Y., 2015. H3N2 Mismatch of
729 2014–15 Northern Hemisphere Influenza Vaccines and Head-to-head Comparison
730 between Human and Ferret Antisera derived Antigenic Maps. Sci. Rep. 5.
731 doi:10.1038/srep15279
732



733

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

744

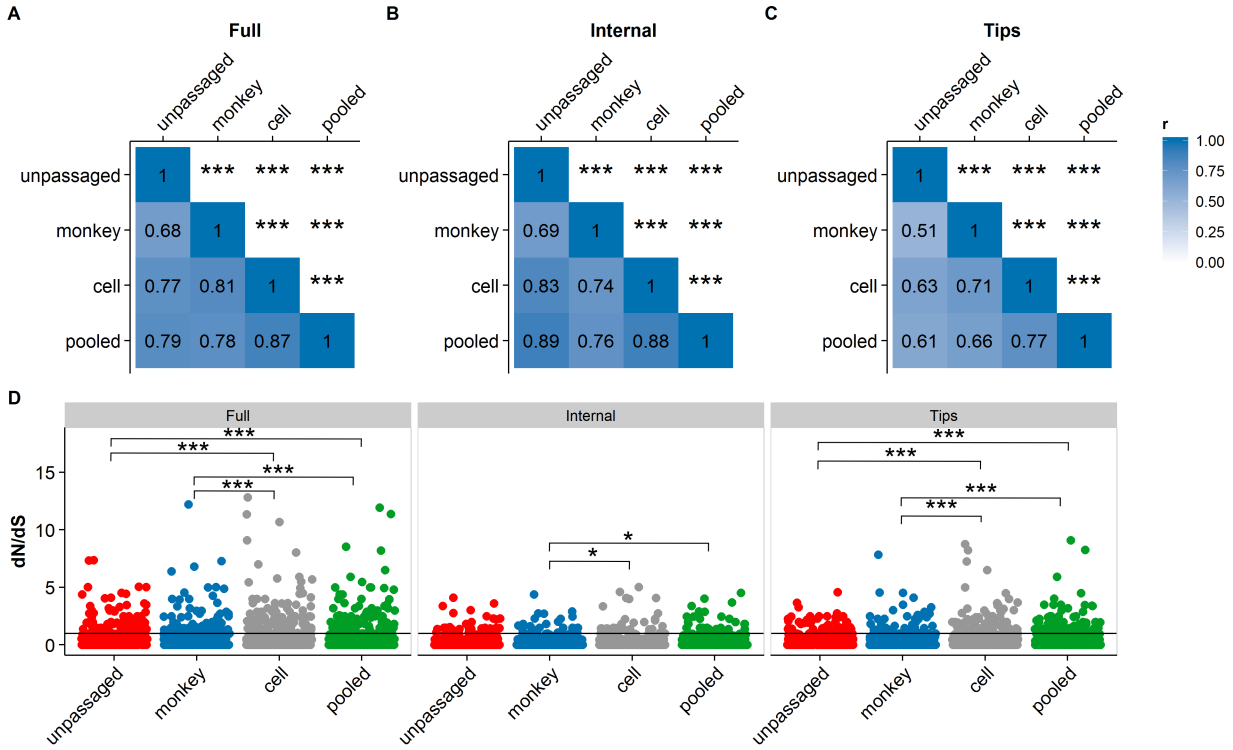
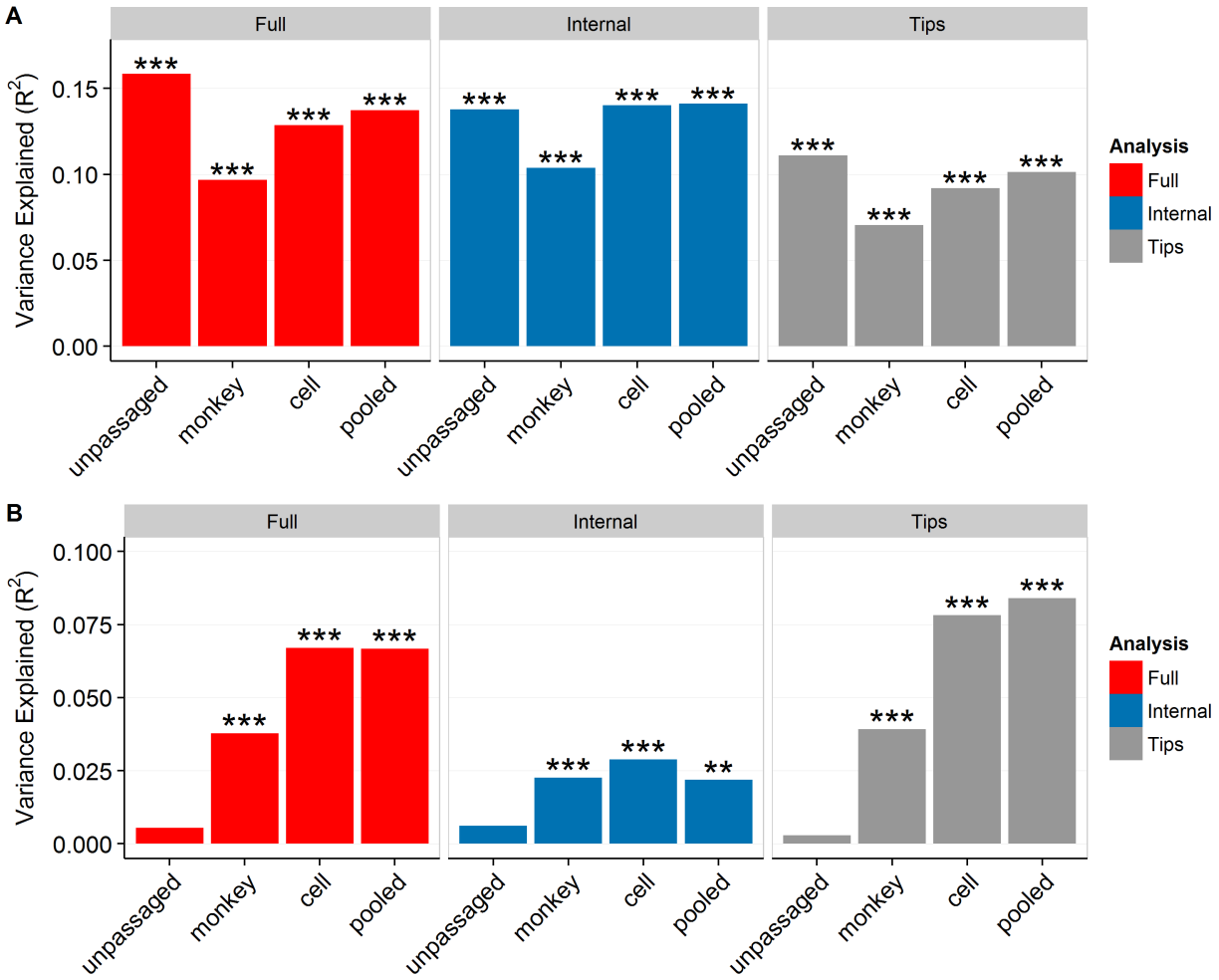


Figure 2. Comparison of sitewise *dN/dS* values among sequences with differing passage histories. (A–C) Pearson correlations between sitewise *dN/dS* values for HA sequences derived from passaged and unpassaged influenza virus collected between 2005 and 2015 (downsampled to n = 917 in all groups). Correlations 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 (*$0.01 \leq P < 0.05$, **$0.001 \leq P < 0.01$, ***$P < 0.001$). (D) Scatter plots show the raw sitewise *dN/dS* values used to calculate the correlations in parts A–C. We tested for systematic differences in *dN/dS* values with paired *t* tests, and significant differences are indicated with asterisks (*$0.01 \leq P < 0.05$, **$0.001 \leq P < 0.01$, ***$P < 0.001$). Data used to generate this figure are available in Supplementary Dataset 1.



758

759 **Figure 3. Percent variance in *dN/dS* explained by relative solvent accessibility (A)**

760 **and by inverse distance to protein site 224 (B).** (A) Relative solvent accessibility

761 (RSA) explains ~10%–16% of the variation in *dN/dS* for all sequences. (B) Inverse

762 distance to site 224 explains ~7% of the variation in *dN/dS* for cell-passaged sequences

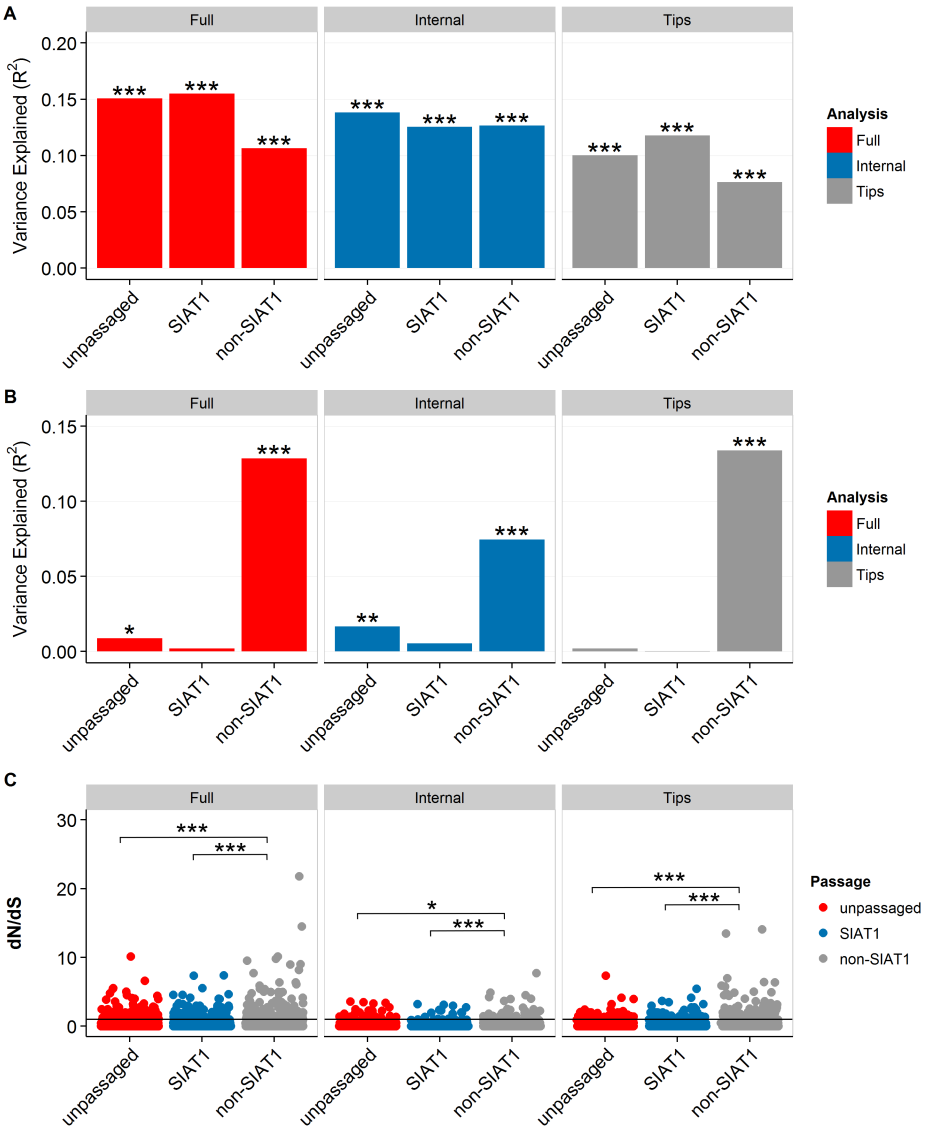
763 and for all sequences (pooled), however it explains virtually no variation for unpassaged

764 sequences. Asterisks denote significance of correlations (*0.01 ≤ P < 0.05, **0.001 ≤ P

765 < 0.01, ***P < 0.001). Data used to generate this figure are available in Supplementary

766 Dataset 1.

767



768

769 **Figure 4. Virus passaged in non-SIAT1 cells carries unique adaptations not**
770 **present in unpassaged or SIAT1-passaged virus.** For size matched groups of
771 sequences collected between 2005 and 2015 (*n* = 1046), (A) the correlation between
772 *dN/dS* and RSA is weakened for virus passaged in non-SIAT1 cells. (B) The correlation
773 between *dN/dS* and inverse distance to site 224, representing a positive-selection
774 hotspot in the vicinity of that site, is only present in virus passaged in non-SIAT1 cells.
775 (C) Scatter plots show individual *dN/dS* values obtained from the full phylogenetic tree,
776 internal branches only, and tip branches only. For the full tree, internal branches, and tip
777 branches, *dN/dS* in non-SIAT1-passaged virus is significantly elevated relative to
778 unpassaged and SIAT1-passaged virus (paired *t* test). Asterisks denote significance
779 levels (*$0.01 \leq P < 0.05$, **$0.001 \leq P < 0.01$, ***$P < 0.001$). Data used to generate this
780 figure are available in Supplementary Dataset 3.

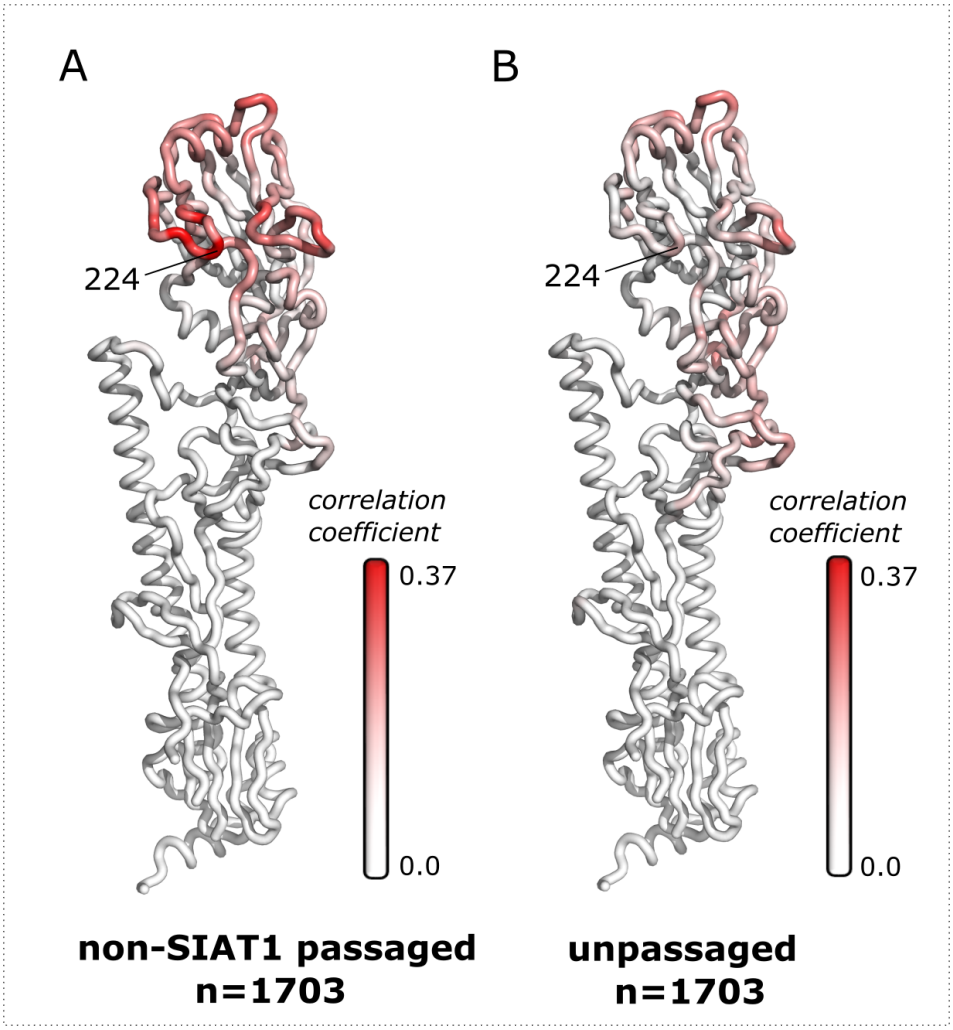


Figure 5. **Correlations mapped onto the hemagglutinin structure for non-SIAT1-passaged and unpassaged sequences.** The correlation between *dN/dS* and inverse distance for each reference site was mapped onto the hemagglutinin structure for (A) non-SIAT1 sequences and (B) unpassaged sequences collected between 2005 and 2015 (*n* = 1703). Red coloring represents positive correlations, while white represents zero or negative correlations. Non-SIAT1-passaged and unpassaged sequences yield distinct correlation patterns. In particular, the loop containing site 224 lights up strongly for non-SIAT1-passaged sequences but not for unpassaged sequences. Data used to generate this figure are available in Supplementary Dataset 5.

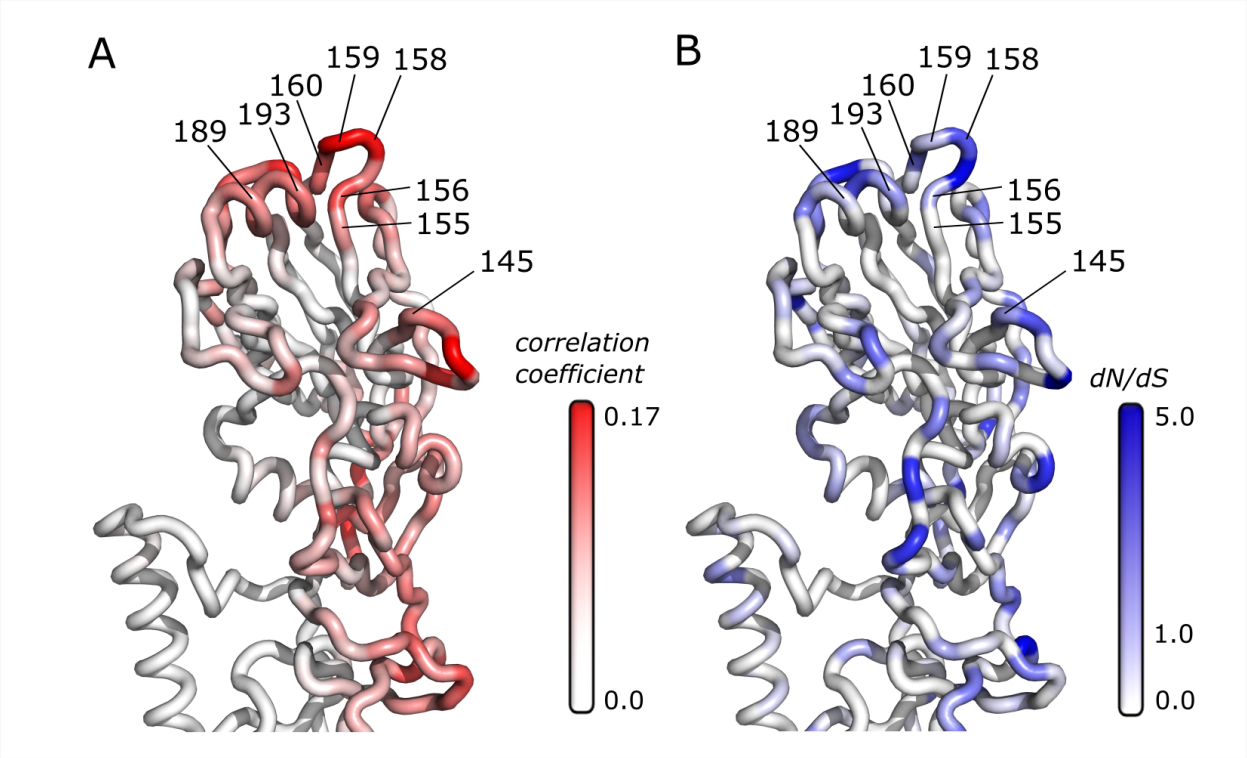
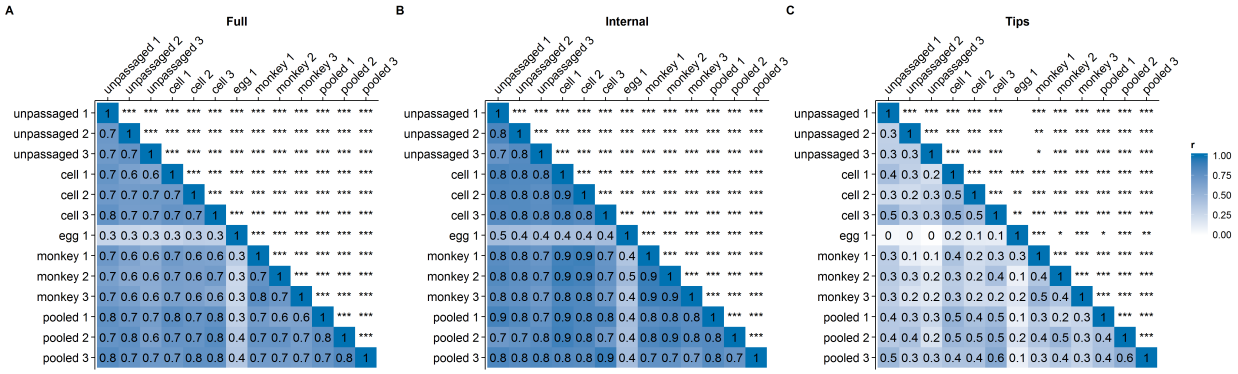
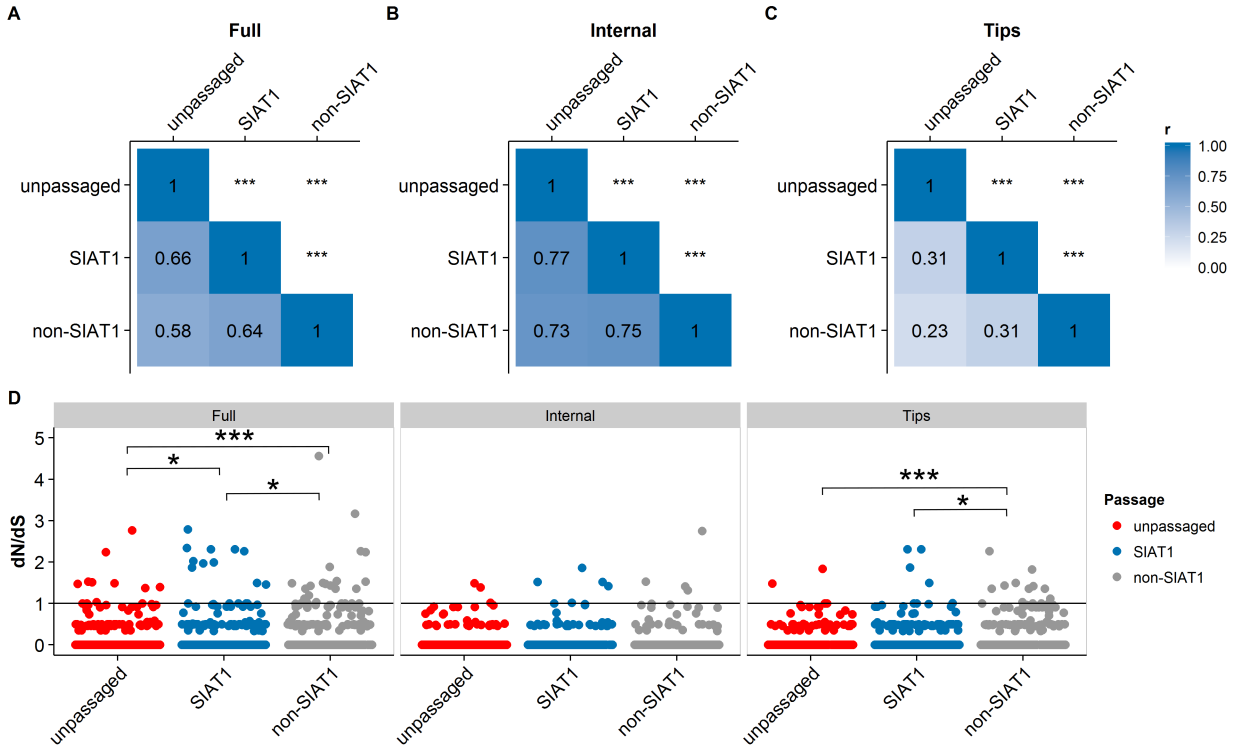


Figure 6. **Unpassaged sequences allow recovery of antigenic regions from positive-selection analysis.** For each site, the correlation between *dN/dS* and inverse distance (A) or *dN/dS* directly (B) were mapped onto the hemagglutinin structure, for *dN/dS* derived from unpassaged sequences collected between 2005 and 2015 (*n* = 1703). Red coloring represents higher correlation; blue coloring represents higher *dN/dS*. Highlighted regions contain sites (labeled with protein site number) which experimentally determined to be antigenic by (Koel et al., 2013) and sites experimentally determined by (Chambers et al., 2015) to be responsible for antigenic escape of the 2014-2015 influenza. Correlations and *dN/*dS for antigenic sites are given in Table 2. Data used to generate this figure are available in Supplementary Dataset 5.



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 Dataset 2.



Supplementary Figure 2. 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. (D) Scatter plots show the raw sitewise *dN/dS* values used to calculate the correlations in parts A–C. We tested for systematic differences in *dN/dS* values with paired *t* tests, and significant differences are indicated with asterisks (*$0.01 \leq P < 0.05$, **$0.001 \leq P < 0.01$, ***$P < 0.001$). For the full tree, *dN/dS* in both non-SIAT1-passaged virus and SIAT1-passaged virus is significantly elevated relative to unpassaged virus. For tip branches, *dN/dS* in non-SIAT1-passaged virus is significantly elevated relative to unpassaged virus. No significant difference was found for internal branches. Data used to generate this figure are available in Supplementary Dataset 4.

835 **Table 1. Parsing of passage-annotated FASTA sequences into passage history**
836 **groups.** For each passage group, we defined a regular expression that could reliably
837 identify sequences with that passage history. Regular expressions were applied through
838 the built-in python library "re". SIAT1 and non-SIAT1 cell culture regular expressions
839 were applied to the subset of sequences identified as generic cell culture sequences.
840 The right three columns list the number of sequences we identified for each passage
841 group, for years 2005–2015, 2014 only, and 2015 only.

842

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

843

844

845

846

847

848

849

850 **Table 2. Evolutionary rates and inverse distance correlations of antigenic sites.**
851 For each site, we determined *dN/dS* and the correlation between *dN/dS* and inverse
852 distance for unpassaged sequences collected between 2005 and 2015 (*n* = 1703). 7/8
853 antigenic sites have inverse-distance correlations above the 90th percentile, while only
854 2/8 antigenic sites have *dN/dS* values above the 90th percentile. Antigenic sites were
855 experimentally determined by (Chambers et al., 2015) and (Koel et al., 2013).
856

Antigenic site			**Raw *dN/dS***		***Inv.-dist. correlation***	
Gene	**Protein**	**Study**	***dN/dS***	**percentile**	***r***	**percentile**
161	145	Koel	1.828	0.863	0.071	0.910
171	155	Koel	0	0.002	0.061	0.871
172	156	Koel	1.371	0.805	0.116	0.982
174	158	Koel	3.244	0.965	0.157	0.994
175	159	Koel, Chambers	0.936	0.738	0.165	0.998
176	160	Chambers	2.922	0.953	0.133	0.980
205	189	Koel	0.965	0.748	0.076	0.928
209	193	Koel	1.829	0.879	0.084	0.947

857
858
859
860
861
862
863
864
865

866 **Supplementary File 1.** GISAID acknowledgements for hemagglutinin sequences
867 collected between 1968 and 2015.
868
869 **Supplementary Dataset 1.** Data used to generate Figures 2 and 3. This file includes 1)
870 sitewise *dN/dS* values of random draws of 917 unpassaged, generic cell cultured,
871 monkey cell cultured, and the pooled group sequences collected between 2005 and
872 2015, 2) protein and gene numbering, 3) PDB:2YP7 sequence, 4) relative solvent
873 accessibilities of the hemagglutinin trimer, and 5) linear distances to protein site 224.
874
875 **Supplementary Dataset 2.** Data used to generate Supplementary Figure 1. This file
876 includes 1) sitewise *dN/dS* values of random draws of 97 unpassaged, generic cell
877 cultured, egg cultured, monkey cell cultured, and the pooled group sequences collected
878 between 2005 and 2015, 2) protein and gene numbering, 3) PDB:2YP7 sequence.
879
880 **Supplementary Dataset 3.** Data used to generate Figure 4. This file includes 1)
881 sitewise *dN/dS* values of random draws of 1046 unpassaged, SIAT1, and non-SIAT1
882 cell culture sequences collected between 2005 and 2015, 2) protein and gene
883 numbering, 3) PDB:2YP7 sequence, 4) relative solvent accessibilities of the
884 hemagglutinin trimer, and 5) linear distances to protein site 224.
885
886 **Supplementary Dataset 4.** Data used to generate Supplementary Figure 2. This file
887 includes 1) sitewise *dN/dS* values of random draws of 249 unpassaged, SIAT1 cultured,

888 and non-SIAT1 cell cultured sequences collected in 2014, 2) protein and gene
889 numbering, 3) PDB:2YP7 sequence.
890
891 **Supplementary Dataset 5.** Data used to generate Figures 5 and 6. This file includes 1)
892 sitewise *dN/dS* values of random draws of 1703 unpassaged and non-SIAT1 cell
893 cultured sequences collected between 2005 and 2015, 2) protein and gene numbering,
894 3) PDB:2YP7 sequence, 4) sitewise inverse distance correlations.
895
896 **Supplementary Dataset 6.** This file includes 1) all sitewise *dN/dS* values used to
897 generate figures, 2) protein and gene numbering, 3) PDB:2YP7 sequence, 4) relative
898 solvent accessibilities of the hemagglutinin trimer, and 5) linear distances to protein site
899 224.
900