The Genomic Rate of Molecular Adaptation of the Human Influenza A Virus

Samir Bhatt[1,2] Edward C. Holmes,[3,4] and Oliver G. Pybus*[,1]

[1]Department of Zoology, University of Oxford, Oxford, United Kingdom

[2]Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom

[3]Center for Infectious Disease Dynamics, The Pennsylvania State University, Mueller Laboratory

[4]Fogarty International Center, National Institutes of Health, Bethesda, Maryland

***Corresponding author:** E-mail: oliver.pybus@zoo.ox.ac.uk.

Associate editor: Daniel Falush

Research article

Abstract

Quantifying adaptive evolution at the genomic scale is an essential yet challenging aspect of evolutionary biology. Here, we develop a method that extends and generalizes previous approaches to estimate the rate of genomic adaptation in rapidly evolving populations and apply it to a large data set of complete human influenza A virus genome sequences. In accord with previous studies, we observe particularly high rates of adaptive evolution in domain 1 of the viral hemagglutinin (HA1). However, our novel approach also reveals previously unseen adaptation in other viral genes. Notably, we find that the rate of adaptation (per codon per year) is higher in surface residues of the viral neuraminidase than in HA1, indicating strong antibody-mediated selection on the former. We also observed high rates of adaptive evolution in several nonstructural proteins, which may relate to viral evasion of T-cell and innate immune responses. Furthermore, our analysis provides strong quantitative support for the hypothesis that human H1N1 influenza experiences weaker antigenic selection than H3N2. As well as shedding new light on the dynamics and determinants of positive Darwinian selection in influenza viruses, the approach introduced here is applicable to other pathogens for which densely sampled genome sequences are available, and hence is ideally suited to the interpretation of next-generation genome sequencing data.

Key words: influenza, McDonald–Kreitman test, neutrality tests, rate of adaptation.

Introduction

Human influenza A viruses (negative-sense RNA, family *Orthomyxoviridae*) are among the most rapidly evolving of all pathogens, a behavior that arises from the complex interplay of rapid mutation and replication, strong natural selection, and frequent genome segment reassortment. The molecular adaptation of influenza A virus is of great interest from both an evolutionary and a public health perspective: Abundant genome sequence data make influenza a good model system for evolutionary study, while the continual adaptation of the virus to host immune responses is a key determinant of its epidemic behavior and a challenge to effective vaccination.

Most previous studies of human influenza A virus adaptation have focused on the hemagglutinin (HA) gene and to a lesser extent the neuraminidase (NA) gene, which encode envelope glycoproteins that contain many antigenic sites targeted by humoral immune responses, and have concentrated on the H3N2 subtype that has been the most common cause of seasonal influenza since 1968. Positive selection has been detected at numerous codons within HA1 (the major immunogenic subdomain of HA) using methods that analyze the ratio (d_N/d_S) of nonsynonymous-to-synonymous nucleotide substitutions per site (e.g., Ina and Gojobori 1994; Yang 2000). In addition, HA1 phylogenies exhibit a characteristic asymmetrical shape, comprising a main "trunk" lineage that represents the pathway of fixed mutations through time, and terminal branches that represent nonpersistent lineages (Fitch et al. 1991). The surviving trunk lineage experiences strong positive selection at codons that mostly coincide with known antigenic sites (Fitch et al. 1997), giving rise to a process of continual viral adaptation commonly known as "antigenic drift." Other studies have used d_N/d_S methods to investigate the nature of positive selection in HA. For example, Wolf et al. (2006) suggested that the evolution of seasonal influenza is characterized by periods of "antigenic stasis" followed by rapid bursts of antigenic evolution, whereas Shih et al. (2007) concluded that antigenic drift is a continual rather than punctuated process and the result of multiple mutations at antigenic sites.

Despite the new abundance of complete human influenza A virus genomes, little is known about adaptation outside the antigenic proteins mentioned above. Suzuki (2006) applied a d_N/d_S approach to 100 complete H3N2 genomes and concluded that negative selection dominated in all proteins; significant positive selection was observed only in a handful of codons in the HA, NA, and nucleoprotein (NP) genes. Pond et al. (2008) analyzed the same data using a technique more sensitive to the detection of individual selective sweeps and reported more evidence for adaptation, in five genes: PB2 (polymerase basic 2 - two codons), PB1 (polymerase basic 1 - five codons), PA (polymerase acidic - three codons), HA (ten codons), and NA (four codons).

Almost all these studies have employed various forms of phylogenetic d_N/d_S methods (e.g., Nielsen and Yang 1998).

Although such methods are effective in detecting selection among sequences sampled from divergent species (Yang and Bielawski 2000), they possess some disadvantages when applied to viral data sets. First, they are too computationally intensive to apply to alignments comprising many hundreds of complete viral genomes. Second, when applied to within-population sequences that have been sampled serially through time, the relationship between the strength of selection and d_N/d_S can become difficult to discern (Sheridan et al. 2004; Rocha et al. 2006; Kryazhimskiy and Plotkin 2008). Although Nielsen and Yang (2003) have shown that it is possible to convert d_N/d_S values into estimates of population selection coefficients, the interpretation of specific d_N/d_S estimates depends on the mode of selection operating in the population, which in many cases will not be known with certainty (Nielsen and Yang 2003). Third, d_N/d_S methods, by definition, are more sensitive to recurrent selection (i.e., repeated amino acid changes at the same codon) than to selected mutations that occur only once (i.e., solitary selective sweeps). Lastly, although d_N/d_S methods can indicate which sites have been positively selected, they do not provide a direct estimate of the rate of adaptive fixation.

Given these limitations, it is clearly important to develop computationally efficient methods for detecting positive selection in viral populations and to explore alternatives to d_N/d_S methods that are based on different sets of assumptions. Previously, Williamson (2003) demonstrated that population genetic methods based on the McDonald–Kreitman (MK) test (McDonald and Kreitman 1991) can be practically applied to serially sampled viral sequences to quantify a rate of adaptive fixation through time. We have shown that, with appropriate modifications, MK-based methods have good statistical properties when applied to RNA viruses (Bhatt et al. 2010).

Here, we extend and generalize Williamson's (2003) method for estimating rates of adaptive evolution and apply this to thousands of complete genome sequences of influenza A virus. One key advantage of our approach is that it corrects for the transient presence of segregating deleterious mutations, which are a common feature of most RNA populations (Pybus et al. 2007; Holmes 2009). In doing so, we provide the first estimates of the rate of molecular adaptation of human influenza A virus across the whole viral genome. In addition, we directly test the hypothesis that the adaptive dynamics of the H1N1 subtype are significantly slower than those of H3N2 (Rambaut et al. 2008). Because our framework can be applied to many thousands of complete viral genomes, it presents a practical solution to the study of pathogen adaptive evolution once next-generation sequencing technologies become commonplace in molecular epidemiology.

Methodological Framework

Our approach extends and generalizes that applied to HIV-1 by Williamson (2003), which was in turn based on earlier work by Smith and Eyre-Walker (2002) and McDonald and Kreitman (1991). These methods require a "main alignment" of nucleotide sequences plus a homologous "outgroup" that is used to determine whether sites in the main alignment are ancestral or derived. The temporally structured evolution of influenza A virus (and other rapidly evolving populations) means that we can use sequences from an earlier time point as an outgroup rather than a sister species or population (see Bhatt et al. 2010 for details).

As with Smith and Eyre-Walker (2002) and Williamson (2003), our approach aims to estimate a rate of adaptive change by defining a class of nucleotide sites that are assumed to be neutral, then using that class to detect an excess of replacement fixations (or polymorphisms) in other classes of sites. Specifically, Smith and Eyre-Walker (2002) assumed that all polymorphic sites were neutral, as might occur under a scenario of strong directional selection and no weak selection, and then estimated the proportion of among-species nucleotide fixations that were adaptive. Williamson (2003) suggested that for viruses, in addition to the process of adaptive fixation, some high-frequency polymorphisms may be adaptive as a result of fluctuating or frequency-dependent selection. Williamson (2003) therefore assumed that only low-frequency polymorphisms (observed frequencies <0.5) were neutral and subsequently estimated two adaptive variables—the number of adaptive fixations and the number of adaptive high-frequency polymorphic sites. Williamson's (2003) premise that low-frequency polymorphisms are neutral seems reasonable if effective population sizes are large (as is almost certainly the case for most viral populations) and if segregating deleterious polymorphisms are absent. We believe that the latter condition may not be met: even when selection is strong and N_e is large, many rare deleterious polymorphisms in RNA virus populations will be observed as a consequence of the exceptional mutation pressure in such populations (e.g., Pybus et al. 2007). The presence of such deleterious changes could bias estimated levels of adaptive evolution (Fay et al. 2001; Charlesworth and Eyre-Walker 2008).

Therefore, we propose that selectively neutral sites in human influenza A viruses are mostly likely found among those polymorphisms that are neither rare nor common. In many instances, such polymorphisms will be too infrequent to permit analysis, but this is not an issue here due to the very large number of sequences available. In addition to assuming large effective population sizes, our approach assumes that recurrent or frequent-dependent selection is not so prevalent that many mid-frequency polymorphisms are adaptive. However, we find evidence that this is indeed the case for two regions of the influenza genome, and we show how this can be resolved by partitioning genes into functional domains (see Materials and Methods). To explore sensible values for the upper and lower bounds of the "mid-frequency" class of sites, we investigated a simple Wright–Fisher model of directional selection using selection coefficients and mutation rates typical of RNA viruses (Supplementary Material online). Under conservative estimates of the global effective population size of influenza, this model suggests that polymorphisms with frequencies between ~0.15 and ~0.75 represent the most neutral class of sites for RNA virus data. Much more importantly, we show that the results generated by this method are robust to the exact site-frequency threshold values chosen (see Results and fig. 4).

Calculation of our method begins by classifying each nucleotide site in the main alignment as either a 1) silent fixation, 2) replacement fixation, 3) silent high-frequency polymorphism, 4) replacement high-frequency polymorphism, 5) silent mid-frequency polymorphism, 6) replacement mid-frequency polymorphism, 7) silent low-frequency polymorphism, or 8) replacement low-frequency polymorphism. The number of sites in each class is denoted s_f, r_f, s_h, r_h, s_m, r_m, s_l, and r_l, respectively. High-frequency, mid-frequency, and low-frequency polymorphisms are defined as having observed frequencies >0.75, 0.75–0.15, and <0.15, respectively. If it is assumed that all silent mutations and all mid-frequency polymorphisms are neutral, then the expected numbers of nonneutral sites in each frequency class (a_l, a_h, and a_f) are:

$$a_l = r_l - s_l\left(\frac{r_m}{s_m}\right),\ a_h = r_h - s_h\left(\frac{r_m}{s_m}\right),\ a_f = r_f - s_f\left(\frac{r_m}{s_m}\right) \quad (1)$$

In each case, the a value represents the excess number of replacement sites over and above that expected by neutral evolution. Hence, a_f equals the number of adaptive fixations (sensu Smith and Eyre-Walker 2002), and a_h represents the number of adaptive polymorphisms that are at high frequency but kept from fixation by recurrent or fluctuating selection (sensu Williamson 2003). By analogy, a_l can be interpreted as the excess number of low-frequency polymorphisms, whose value will be determined by mutation pressure, and is not considered further here. The total number of sites that have undergone adaptive fixation or spread in the time between the outgroup and the main alignment is thus $a = a_h + a_f$. The proportion of high-frequency and fixed replacement sites that have undergone adaptive change are therefore a_h/r_h and a_f/r_f, respectively. We combine these values to calculate the total proportion of fixed or high-frequency sites that have undergone adaptive change, specifically, $(a_h + a_f)/(r_h + r_f)$. To avoid bias (Welch 2006), the ratio (r_m/s_m) is calculated separately for each gene (see Materials and Methods). A computer program (Adapt-A-Rate) to perform these methods is available from http://evolve.zoo.ox.ac.uk.

Materials and Methods

Sequence Collection

We obtained all available whole human influenza A genome sequences belonging to subtypes H1N1 and H3N2 with known sampling dates between 1977–2009 (http://www.ncbi.nlm.nih.gov/genomes/FLU). This range corresponds to the period during which these two subtypes co-circulated in human populations. (Note: H1N1 refers to the lineage reintroduced to humans in 1977 not to the 2009 swine-origin pandemic). Recombinant or laboratory generated strains, plus those containing genome segments from nonhuman sources, were excluded. Separate alignments were constructed for each influenza virus gene (PB2, PB1, PA, HA, NP, NA, M1, NS1, and NS2) and subtype (H3N2, H1N1). Noncoding regions and overlapping reading frames were removed. Because the majority of NS2 consists of overlapping codons, this gene was removed from analysis. For each alignment, Neighbor-Joining phylogenies were constructed using QuickTree (Howe et al. 2002) and inspected to identify and remove incorrectly labeled sequences. To obtain adequate sample sizes, sequences were collated into time points each representing 3 contiguous years. Time points containing <10 sequences were removed. The final data set comprises 775 H1N1 and 1674 H3N2 genomes (supplementary table S1, Supplementary Material online). We created two HA sub-alignments, representing the HA1 (globular, antigenically variable) and HA2 (proximal membrane) subdomains of HA. We similarly sought to partition the NA alignment: In the absence of defined subdomains, we partitioned NA codons into those that represent solvent-accessible surface residues (the "NA surface" sub-alignment) and those that represent solvent-inaccessible or internal residues (the "NA internal" sub-alignment; supplementary table S3, Supplementary Material online). The structural properties of NA residues were approximated from previously published structures using ESpript (Gouet et al. 1999; PDB accession numbers 2HTY and 1NN2).

Estimating Adaptation Rates

We computed the estimators defined in equation 1 using the proportional site counting method described in Bhatt et al. (2010), which is optimized for the analysis of highly variable RNA virus data. To be conservative, negative a -values were set to zero. Following Williamson (2003), the consensus sequence of the first time point was used as the outgroup. We investigated the accumulation of adaptations through time by calculating a for each subsequent time point, generating a time series from which we obtained a rate of adaptation, R, using linear regression (R equals the number of adaptive substitutions per year). Confidence limits for R cannot be calculated using parametric regression because the a are autocorrelated. We therefore use a standard bootstrap approach to assess statistical uncertainty for each gene and subtype, as follows: 1) Codons are sampled with replacement from the empirical ancestral sequence to create a bootstrap ancestral sequence of equal length. 2) For each time point, a bootstrap main alignment (equal in size to the empirical alignment) is created by sampling from the empirical alignment. Sampling proceeds according to the codon order defined in step 1. 3) Using the bootstrap data created in steps 1 and 2, we calculate a bootstrap time series of a values. A bootstrap rate of adaptation, R^*, is then calculated from the bootstrap time series, as above. 4) Steps 1–3 are repeated 1,000 times, thereby generating bootstrap distributions for R^* and for the time series of a values. 5) The 95%

percentiles of the bootstrap distributions created in step 4 are calculated.

The *M* Ratio

Each estimator in equation 1 includes $M = (r_m/s_m)$, which is the ratio of replacement-to-silent polymorphisms in the site frequency range that is assumed to be neutral. Although *M* can be calculated separately for each time point and gene, this will result in unacceptably high variances for some calculations due to the small number of available sites. Previous studies have reduced variance by combining site counts across genes (e.g., Smith and Eyre-Walker 2002), but this may introduce bias if *M* varies significantly among genes (see Welch 2006). However, the temporal structure of our data provides a solution: For any given gene, *M* is not expected to vary through time provided that long-term effective population sizes remain sufficiently large (a condition almost certainly met in this case). Our data support this conclusion: For both subtypes, *M* varies significantly among genes but not among time points (two-way analysis of variance; $P < 0.01$). Therefore, for each subtype and gene, we calculate *M* by combining site counts among time points. For the PB2, PB1, PA, NP, M1, M2, and NS1 genes, we calculated *M* using all codons in the alignment. However, the exceptional number of selective sweeps and the likelihood of frequency-dependent selection in the antigenic regions of HA and NA means that mid-frequency polymorphisms in these will not be entirely neutral (supplementary table S2, Supplementary Material online). Thus, for the HA gene analyses, *M* was calculated from the HA2 sub-alignment. Similarly, for the NA analyses, *M* was calculated from the "NA internal" sub-alignment, which, like HA2, is unlikely to be subject to antigenic selection. Consequently, for both HA and NA, three *a* estimates were calculated: one for the whole gene and one for each of the two sub-alignments.

Results

Figure 1 shows the estimated accumulation of adaptive substitutions in H3N2 human influenza A virus over the last three decades, whereas the equivalent results for H1N1 are shown in figure 2. The gradient of each plot represents the rate of molecular adaptation of each influenza A virus gene. As expected, rates of adaptation vary significantly among genes, with the highest rates observed in the surface glycoproteins HA and NA. The adaptation rates of HA and NA are higher for the H3N2 subtype (1.52 and 1.23 adaptive substitutions per gene per year, respectively) than for H1N1 (1.02 and 1.06, respectively); the between-subtype difference in adaptation rate is greater for HA than NA. Importantly, all genes have accumulated adaptive substitutions at an approximately constant rate; we found no evidence for periods of selective activity and stasis in either subtype (figs. 1 and 2). The number of adaptations in the HA and NA genes of the H1N1 subtype do dip slightly in 2008, but this is not significant given the range of statistical uncertainty.

Figures 1 and 2 display the accumulation of adaptive substitutions "per gene," whose alignments vary in length from 706 (PB2) to 98 (M2) codons. To directly compare the intensity of adaptation among genes, we calculated the rate of adaptive substitution per codon per year (hereafter, adaptations/codon/year) with appropriate confidence limits derived using bootstrapping (fig. 3). As before, the HA and NA genes of H3N2 adapt faster than those of H1N1. If we define statistical significance as being achieved if the point estimate for one population falls outside the confidence limits of another, then the difference between H1N1 and H3N2 is statistically significant for HA but-marginally nonsignificant for NA. The range of statistical uncertainty for each gene is negatively correlated ($P < 0.001$) with the length of its alignment, as expected.

We also calculated separate rates of adaptation for the HA1 and HA2 subdomains of HA (fig. 3). The estimated adaptation rate of the conserved structural HA2 domain (which is not thought to be subjected to repeated immune selection) was very low for the H3N2 subtype (0.0005 adaptations/codon/year) and not significantly different from zero for H1N1. In contrast, for both subtypes, the adaptation intensity for the antigenic HA1 domain was greater that that of the whole HA gene (0.0040 and 0.0029 adaptations/codon/year for H3N2 and H1N1, respectively). This is not simply a reflection of a greater number of nonsynonymous fixations and polymorphisms in HA1: the proportion of such sites that are estimated to be positively selected is higher for HA1 than HA2. Specifically, we estimate ~80% of fixed or high-frequency nonsynonymous sites in HA1 are adaptive, whereas at most only ~40% of such sites in HA2 are adaptive (supplementary table S4, Supplementary Material online). The analogous results for different regions of the NA glycoprotein were even more striking: The adaptation rate of codons that encode solvent-accessible/surface NA residues was significantly higher than that observed for HA1 (0.0054 and 0.0045 adaptations/codon/year for H3N2 and H1N1, respectively; fig. 3). Correspondingly, we estimate that ~85% of nonsynonymous substitutions in surface residues of NA are the result of adaptive selection (supplementary table S4, Supplementary Material online). The estimated selection intensity for solvent-inaccessible/internal NA residues (0.0002 and 0.0004 adaptations/codon/year for H3N2 and H1N1) was minimal compared with that estimated for the surface residues (fig. 3).

Aside from HA and NA, we observed lower (but significantly nonzero) rates of adaptation in other influenza genes (fig 3). Although the polymerase complex genes (PA, PB1, and PB2) often contained more adaptive substitutions than NP, M1, M2, and NS1 (figs. 1 and 2), this appears to be an artifact of their greater length: per codon adaptation rates for the former are lower, indicating that positively selected sites in the polymerase genes are sparse (fig. 3). The rates of adaptation observed in the M2 and NS1 genes of both subtypes (0.0006–0.0016 adaptations/codon/year) and in the NP gene of H3N2 (0.0009 adaptations/codon/year) were higher than those observed in

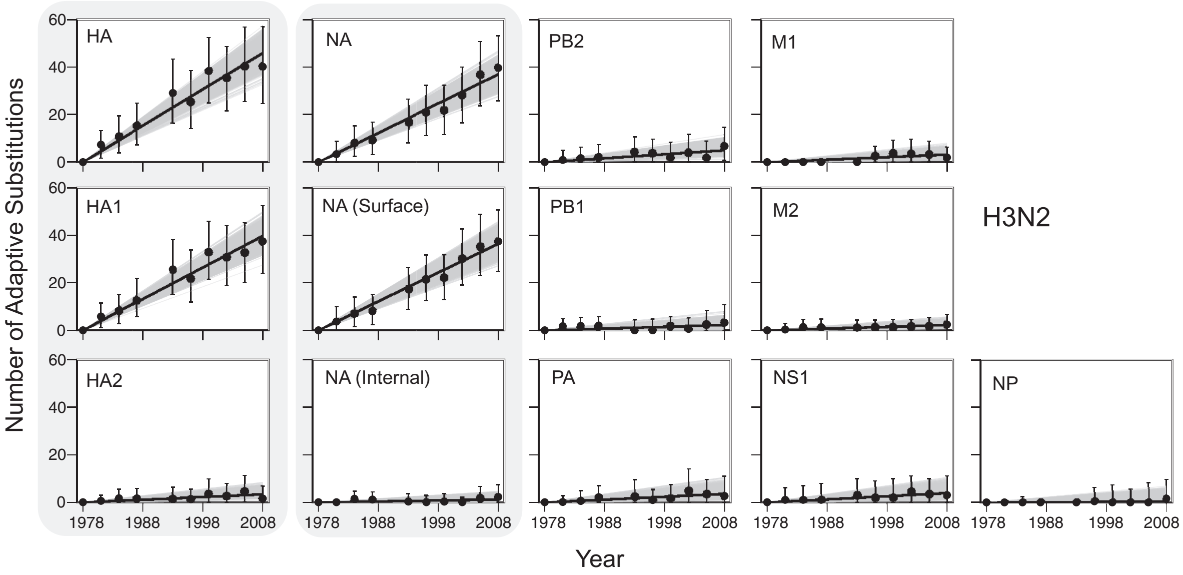

Fig. 1. The accumulation of adaptive substitutions in each gene of human influenza A subtype H3N2 during the period 1977–2009. Three plots are shown for each of the envelope proteins HA and NA (grey boxes): One for the whole gene and one for each of the two structural partitions (see Materials and Methods). In each panel, black circles show the estimated number of adaptive substitutions at each time point to which a linear regression model is fitted (black line). Gray lines show the bootstrap distribution of regression lines (1,000 replicates). Error bars represent the 95% bootstrap percentiles at each time point.

the polymerase complex genes and perhaps higher than expected given the results of previous studies of d_N/d_S (see Introduction). Further, in direct contrast to HA and NA, adaptation rates in the NP and M2 genes were significantly higher for H1N1 than for H3N2 (fig. 3). For genes other HA and NA, the percentage of nonsynonymous substitutions that are estimated to be adaptive is typically 10–55%, with one notable exception: We find that ~70% of such changes in the NP gene of H1N1 are estimated to be adaptive (supplementary table S4, Supplementary Material online), which matches the observation that H1N1 NP genes have adapted faster than those of H3N2.

Importantly, our results are robust to the exact thresholds that are used to classify sites into different site frequency classes. As a demonstration, figure 4 shows the estimated per-codon rate of adaptation for each H1N1 gene as a function of the threshold used to define intermediate frequency sites (set to >0.15 in fig. 1–3; equivalent results for H3N2 are provided in Supplementary Material online). Provided that the threshold is >~0.1, our estimates of adaptation rate are largely invariant to the threshold value chosen. Although there is some numerical variation, this is insignificant in comparison with the observed variation in adaptation rate among genes and is

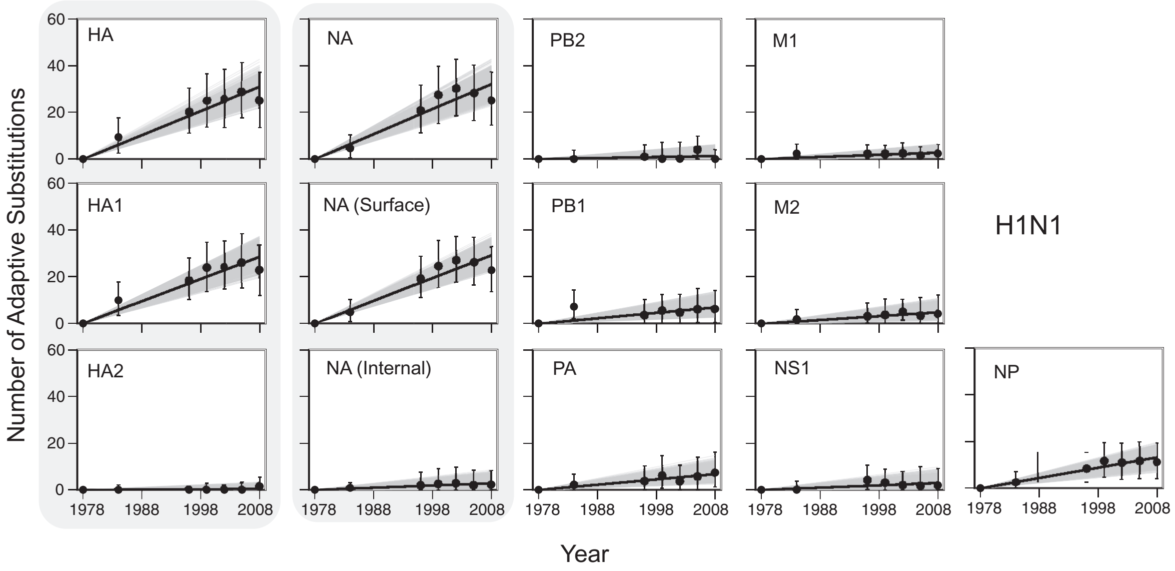

Fig. 2. The accumulation of adaptive substitutions in each gene of human influenza A subtype H1N1 during the period 1977–2009. See figure 1 legend for details.

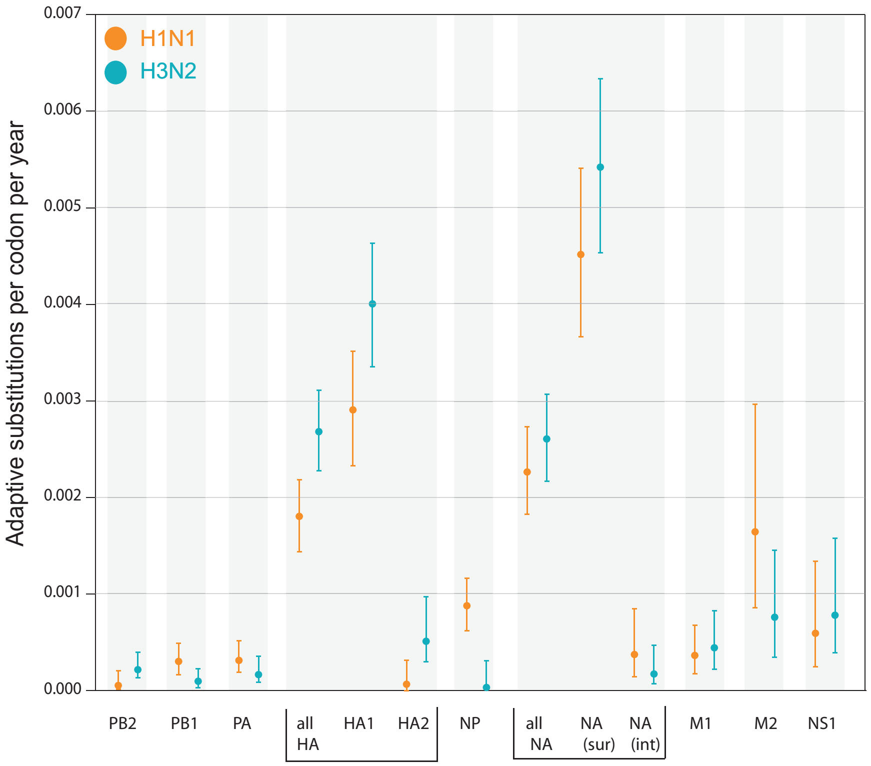

Fig. 3. The rate of adaptive substitutions "per codon" per year in human influenza A genes. Rates for subtype H1N1 are in orange and those for H3N2 in blue. Circles indicate the estimated adaptation rates, and error bars show the 95% bootstrap percentiles.

generally no greater than the statistical uncertainty for each estimate. These results also demonstrate that deleterious mutations are typically found at frequencies <0.1. When the threshold is <0.1, replacement deleterious mutations are incorrectly placed in the intermediate frequency class, resulting in an inflated r_m value and a corresponding underestimate of the rate of adaptation (see equation 1).

Discussion

Using current d_N/d_S methods, it has been established that positive selection in influenza virus occurs predominantly in the HA1 subdomain of HA, whereas the HA2 subdomain remains relatively conserved (Skehel and Wiley 2000). Our population genetic method recovers this expected pattern: For both subtypes, we observe a much higher rate of adaptive evolution in HA1. Similarly, we find that adaptation in the NA gene is concentrated at surface-accessible sites. That we detect little adaptation in HA and NA interior sites, despite these being completely linked to those sites undergoing the highest rate of adaptation (Boni et al. 2008), indicates that our approach does not generate significant false positives as a result of genetic hitchhiking. This result is perhaps unsurprising, as methods that use the same logic as the McDonald and Kreitman (1991) are known to be robust to assumptions about population demography and recombination rates (e.g., Sawyer and Hartl 1992; Nielsen 2005; Andolfatto 2008). Indeed, it is trivial to prove that our estimators (equation 1) are robust to the presence of any number of neutral mutations that have reached high frequencies due to hitchhiking (because the expected replacement-to-silent ratio of such mutations is r_m/s_m). The lower rates of adaptation estimated for genes other than HA and NA, particularly the polymerase complex, further indicate the effectiveness of our approach.

A key result from this study is that almost all adaptive evolution in NA is concentrated in residues on the surface of NA. Furthermore, the adaptation rate "per codon" and the percentage of adaptive nonsynonymous changes are higher for surface NA residues than for HA1 (although the latter contains many non-surface residues). Rambaut et al. (2008) reported that HA and NA have similar rates of nucleotide substitution (a measure that combines both neutral and adaptive evolution). The observation that "adaptive" evolution in NA occurs almost exclusively in solvent-accessible surface residues suggests that it is mainly due to antibody-mediated immune responses, rather than to changes associated with functional compatibility and coevolution with HA (e.g., for efficient virus replication; Mitnaul et al. 2000). In either case, these results reaffirm the importance of NA inhibition assays in vaccine selection and highlight the urgency of further research into NA evolutionary dynamics. In addition, our method of distinguishing between surface and non-accessible residues or between functional domains, could in future be used to investigate intragenic variation in adaptation rates for other viral proteins, provided that adequate models of their secondary structure are available.

We also noted a marked difference in HA and NA adaptation rates between subtypes, with higher rates of adaptation for H3N2 than H1N1 in both genes. Since the

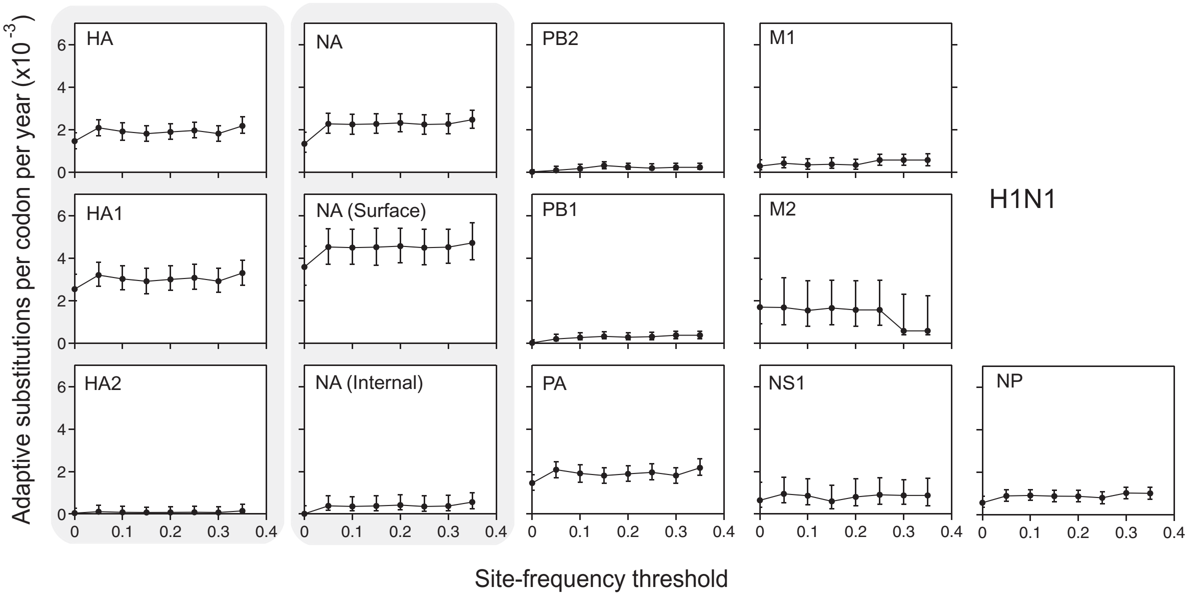

Fig. 4. The rate of adaptive substitution per codon for H1N1 genes as a function of the threshold value used to define the low and intermediate site-frequency classes (see Methodological Framework). The results in figures 1–3 correspond to a threshold value of 0.15. Error bars show the 95% bootstrap percentiles.

reintroduction of seasonal H1N1 into human populations in 1977, this subtype has exhibited a lower case-fatality rate than H3N2, particularly when these subtypes cocirculate (Wright et al. 1980; Kaji et al. 2003). In addition, H1N1 undergoes less severe seasonal genetic bottlenecks than H3N2, which has been suggested to reflect comparatively weaker immune selection on H1N1, resulting in less pronounced or frequent reductions in genetic diversity (Rambaut et al. 2008). Our subtype-specific differences in adaptation rate provide the first explicit support for this hypothesis (fig. 3).

Aside from the surface glycoproteins, the NP, NS1, and M2 genes also exhibit notable rates of adaptation per codon (fig. 3). Significant adaptation in these genes was not apparent in previous studies, possibly because our method is more sensitive than d_N/d_S approaches to sites that undergo adaptive substitution only once in their history. The causes of adaptation in NP, NS1, and M2 are unclear; one explanation is immune pressure mediated by cytotoxic T lymphocytes (CTLs), which these genes are known to experience (Berkhoff et al. 2005; Fernandez-Sesma et al. 2006). Indeed, the existence of CTL escape mutations in NP is well documented (McMichael et al. 1983; Townsend et al. 1986; Voeten et al. 2000) and responses against NP may play a role in partial cross-protection between subtypes (Epstein et al. 2005; Tumpey et al. 2005; Ahmed et al. 2007). Similarly, the adaptation we observe in NS1 could be explained by viral evasion of host innate immune responses, particularly as NS1 blocks the activation of interferon and protein kinase responses to double-stranded RNA (Baigent and McCauley 2003; Li et al. 2006). For the M2 gene, although rates of adaptation are higher in H1N1 viruses, the proportion of nonsynonymous substitutions that are adaptive is estimated to be greater for H3N2 (~55%) than H1N1 (~30%). This counterintuitive result could be explained if H3N2 M2 genes have experienced particularly strong selection at a restricted set of sites, as expected given the key role the M2 ion channel protein plays in the development of adamantane resistance in H3N2 (Bright et al. 2006, Simonsen et al. 2007**)**. In sum, it is clear that the contribution of NP, NS1, and M2 to influenza adaptation is of a sufficient magnitude that evolutionary change in these genes may have important applications for influenza vaccine and antiviral research.

Although the methods used here have low type I statistical errors (Bhatt et al. 2010), there are factors that may reduce statistical power, resulting in underestimates of the rate of adaptive evolution. First, if positively selected sites are common at intermediate site-frequencies, then r_m/s_m will be biased upwards, and the number of adaptive changes correspondingly downwards. We attempted to correct for this effect in HA and NA by partitioning codons in these genes into separate sub-alignments according to their structural characteristics. Second, our method will underestimate the rate of adaptation if multiple selective sweeps occur at the same nucleotide site. If this effect was significant throughout our study, then we should have observed a declining rate of adaptation through time, due to "saturation" of adaptive change at such sites. However, no such decline was observed, even for the HA and NA genes. Indeed, unlike some previous studies (e.g., Nelson et al. 2006), we found no evidence that the adaptive process in HA1 is punctuated. This discrepancy is likely due to the different sample sizes and methods used: our duration of study is considerably longer than that of earlier studies (Nelson et al. 2006), such that we are unlikely to detect short-term fluctuations in selection. Our results are therefore best interpreted as a conservative estimate of the long-term influenza A adaptation rate. This is a "net" rate of adaptation, as positively selected fixations at earlier time points could be potentially lost at later time points due to further evolution. However, as discussed above, the constant accumulation of adaptive changes that observed in this study suggests that relatively few sites change in this manner

Clearly, all statistical frameworks for detecting the action of natural selection on gene sequences have advantages and disadvantages and provide different perspectives on this fundamental process (e.g., Nielsen 2005). For example, approaches based on d_N/d_S can detect specific codons under positive selection whereas ours cannot. Furthermore, our method does not explicitly study changes through time in the frequency of mutations at specific sites (see Steinbrück and McHardy (2011) for a recent approach that does). Our approach is complementary to that employed in parametric poisson random field methods (e.g., Bustamante et al. 2002): The latter can estimate mutational selection coefficients directly but are sensitive to assumptions concerning the dominant mode of selection. Our method (like those of Williamson (2003) and Smith and Eyre-Walker 2002) does not directly estimate population genetic parameters but instead quantifies a deviation from neutrality without specifying exactly what type of selection drives that deviation. However, in addition to the assumptions discussed in the Methodological Framework section, all three methods assume that silent sites are neutral, which may not be the case in those parts of RNA virus genomes that exhibit strong RNA secondary structure (Simmonds et al. 2004) or where synonymous sites have been shown to be of functional importance (Marsh et al. 2008). The dominant selective force on such silent sites is likely negative, the effect of which, in our case, will be to reduce the ratio (r_m/s_m), potentially leading to an overestimate of the number of adaptive sites. However, this bias will be partially but not wholly abrogated by concomitant reductions in s_h and s_f for the same reason. Careful interpretation of results is clearly warranted in such genome regions.

The methods used here provide a computationally tractable approach to the inference of positive Darwinian selection from exceptionally large genomic data sets. Indeed, our approach is only likely to be informative and accurate when applied to large volumes of highly polymorphic sequence data. Furthermore, to reliably estimate adaptation rates, we require sequences to be sampled serially through time as well as an outgroup sequence that is not too genetically divergent from the study sequences (Bhatt et al. 2010). Such data sets are poised to increase in number due to the rapid adoption of next-generation sequencing technologies by laboratories involved in infectious disease research (Pybus and Rambaut 2009). Although the influenza virus sequences analyzed here are strongly structured by time, our method does not require that sequences at each time point are monophyletic. Crucially, and in contrast to phylogenetic approaches, the computation time of our approach increases less than linearly with sample size.

Supplementary Material

Supplementary tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

S.B. was funded by the *Natural Environment Research Council* and in part by COSI UK, E.C.H. in part by grant R01 GM080533 from the National Institute of Health, and O.G.P. by The Royal Society. O.G.P. acknowledges the generous advice of Scott Williamson in 2005, to whom this work is dedicated. Many thanks to Thomas Bowden for assistance with gene structures. S.B. thanks Alexandros Vardoulakis, Udara Pieris, and Denise Xifara.

References

Ahmed R, Oldstone MBA, Palese P. 2007. Protective immunity and susceptibility to infectious diseases: lessons from the 1918 influenza pandemic. *Nat Immunol*. 8:1188–1193.

Andolfatto P. 2008. Controlling type-I error of the McDonald-Kreitman test in genomewide scans for selection on noncoding DNA. *Genetics* 180:1767–1771.

Baigent SJ, McCauley JW. 2003. Influenza type A in humans, mammals and birds: determinants of virus virulence, host-range and interspecies transmission. *Bioessays* 25:657–671.

Berkhoff EG, de Wit E, Geelhoed-Mieras MM, Boon AC, Symons J, Fouchier RA, Osterhaus AD, Rimmelzwaan GF. 2005. Functional constraints of influenza A virus epitopes limit escape from cytotoxic T lymphocytes. *J Virol*. 79:11239–11246.

Bhatt S, Katzourakis A, Pybus OG. 2010. Detecting natural selection in RNA virus populations using sequence summary statistics. *Infect Genet Evol*. 3:421–430.

Boni MF, Zhou Y, Taubenberger JK, Holmes EC. 2008. Homologous recombination is very rare or absent in human influenza A virus. *J Virol*. 82:4807–4811.

Bright RA, Shay DK, Shu B, Cox NJ, Klimov AI. 2006. Adamantane resistance among influenza A viruses isolated early during the 2005-2006 influenza season in the united states. *JAMA*. 295:891–894.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in Arabidopsis. *Nature* 416:531–534.

Charlesworth J, Eyre-Walker A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol*. 25:1007–1015.

Epstein SL, Kong WP, Misplon JA, Lo CY, Tumpey TM, Xu L, Nabel GJ. 2005. Protection against multiple influenza A subtypes by vaccination with highly conserved nucleoprotein. *Vaccine*. 23:5404–5410.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227–1234.

Fernandez-Sesma A, Marukian S, Ebersole BJ, Kaminski D, Park MS, Yuen T, Sealfon SC, García-Sastre A, Moran TM. 2006. Influenza virus evades innate and adaptive immunity via the NS1 protein. *J Virol*. 80:6295–6304.

Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc Natl Acad Sci U S A*. 94:7712–7718.

Fitch WM, Leiter JME, Li X, Palese P. 1991. Positive Darwinian evolution in human influenza A viruses. *Proc Natl Acad Sci U S A*. 88:4270–4274.

Gouet P, Courcelle E, Stuart DI, Metoz F. 1999. ESPript: multiple sequence alignments in PostScript. *Bioinformatics* 15:305–308.

Holmes EC. 2009. The evolution and emergence of RNA viruses. Oxford: Oxford University Press.

Howe K, Bateman A, Durbin R. 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* 18:1546–1547.

Ina Y, Gojobori T. 1994. Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses. *Proc Natl Acad Sci U S A*. 91:8388–8392.

Kaji M, Watanabe A, Aizawa H. 2003. Differences in clinical features between influenza A H1N1, A H3N2, and B in adult patients. *Respirology* 8:231–233.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genet*. 4:e1000304.

Li S, Min JY, Krug RM, Sen GC. 2006. Binding of the influenza A virus NS1 protein to PKR mediates the inhibition of its activation by either PACT or double stranded RNA. *Virology* 349:13–21.

Marsh GA, Rabadán R, Levine AJ, Palese P. 2008. Highly conserved regions of influenza a virus polymerase gene segments are critical for efficient viral RNA packaging. *J Virol*. 82:2295–2304.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the adh locus in Drosophila. *Nature* 351:652–654.

McMichael AJ, Gotch FM, Noble GR, Beare PA. 1983. Cytotoxic T-cell immunity to influenza. *N Engl J Med*. 309:13–17.

Mitnaul LJ, Matrosovich MN, Castrucci MR, Tuzikov AB, Bovin NV, Kobasa D, Kawaoka Y. 2000. Balanced Hemagglutinin and Neuraminidase activities are critical for efficient replication of influenza A virus. *J Virol*. 74:6015–6020.

Nelson MI, Simonsen L, Viboud C, et al. (15 co-authors). 2006. Stochastic processes are key determinants of short-term evolution in influenza A virus. *PLoS Pathog*. 2:e125.

Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet*. 39:197–218.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–36.

Nielsen R, Yang Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol*. 20:1231–1239.

Pond SLK, Poon AF, Leigh-Brown AJ, Frost SD. 2008. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol Biol Evol*. 25:1809–1824.

Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet*. 10:540–550.

Pybus OG, Rambaut A, Belshaw R, Freckleton RP, Drummond AJ, Holmes EC. 2007. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Mol Biol Evol*. 24:845–852.

Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. 2008. The genomic and epidemiological dynamics of human influenza A virus. *Nature* 453:615–619.

Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol*. 239:226–235.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. *Genetics*. 132:1161–1176.

Sheridan I, Pybus OG, Holmes EC, Klenerman P. 2004. High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. *J Virol*. 78:3447–3454.

Shih AC, Hsiao TC, Ho MS, Li WH. 2007. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc Natl Acad Sci U S A*. 104:6283–6288.

Simmonds P, Tuplin A, Evans DJ. 2004. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: implications for virus evolution and host persistence. *RNA* 10:1337–1351.

Simonsen L, Viboud C, Grenfell BT, et al. 2007. The genesis and spread of reassortment human influenza A/H3N2 viruses conferring adamantane resistance. *Mol Biol Evol*. 24: 1811–1820.

Skehel JJ, Wiley DC. 2000. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu Rev Biochem*. 69:531–569.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. *Nature* 415:1022–1024.

Steinbrück L, McHardy AC. 2011. Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic Acids Res*. 39:e4.

Suzuki Y. 2006. Natural selection on the influenza virus genome. *Mol Biol Evol*. 23:1902–11.

Townsend AR, Rothbard J, Gotch FM, Bahadur G, Wraith D, McMichael AJ. 1986. The epitopes of influenza nucleoprotein recognized by cytotoxic T lymphocytes can be defined with short synthetic peptides. *Cell* 44:959–968.

Tumpey TM, García-Sastre A, Taubenberger JK, et al. (11 co-authors). 2005. Pathogenicity of influenza viruses with genes from the 1918 pandemic virus: functional roles of alveolar macrophages and neutrophils in limiting virus replication and mortality in mice. *J Virol*. 79:14933–14944.

Voeten JT, Bestebroer TM, Nieuwkoop NJ, Fouchier RA, Osterhaus AD, Rimmelzwaan GF. 2000. Antigenic drift in the influenza A virus (H3N2) nucleoprotein and escape from recognition by cytotoxic T lymphocytes. *J Virol*. 74:6800–6807.

Welch JJ. 2006. Estimating the genome-wide rate of adaptive protein evolution in Drosophila. *Genetics* 173:821–837.

Williamson S. 2003. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Mol Biol Evol*. 20:1318–1325.

Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ. 2006. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biol Direct*. 1:34.

Wright PF, Thompson J, Karzon DT. 1980. Differing virulence of H1N1 and H3N2 influenza strains. *Am J Epidemiol*. 112:814–9.

Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J Mol Evol*. 51:423–432.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*. 15:496–503.