Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins

Sergey Kryazhimskiy[1¤], Jonathan Dushoff[2], Georgii A. Bazykin[3], Joshua B. Plotkin[1,4*]

[1] Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, [2] Department of Biology, McMaster University, Hamilton, Canada, [3] Institute for Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences, Moscow, Russia, [4] Program in Applied Mathematics and Computational Science, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Abstract

The surface proteins of human influenza A viruses experience positive selection to escape both human immunity and, more recently, antiviral drug treatments. In bacteria and viruses, immune-escape and drug-resistant phenotypes often appear through a combination of several mutations that have epistatic effects on pathogen fitness. However, the extent and structure of epistasis in influenza viral proteins have not been systematically investigated. Here, we develop a novel statistical method to detect positive epistasis between pairs of sites in a protein, based on the observed temporal patterns of sequence evolution. The method rests on the simple idea that a substitution at one site should rapidly follow a substitution at another site if the sites are positively epistatic. We apply this method to the surface proteins hemagglutinin and neuraminidase of influenza A virus subtypes H3N2 and H1N1. Compared to a non-epistatic null distribution, we detect substantial amounts of epistasis and determine the identities of putatively epistatic pairs of sites. In particular, using sequence data alone, our method identifies epistatic interactions between specific sites in neuraminidase that have recently been demonstrated, in vitro, to confer resistance to the drug oseltamivir; these epistatic interactions are responsible for widespread drug resistance among H1N1 viruses circulating today. This experimental validation demonstrates the predictive power of our method to identify epistatic sites of importance for viral adaptation and public health. We conclude that epistasis plays a large role in shaping the molecular evolution of influenza viruses. In particular, sites with $\mathrm{dN}/\mathrm{dS}<1$, which would normally not be identified as positively selected, can facilitate viral adaptation through epistatic interactions with their partner sites. The knowledge of specific interactions among sites in influenza proteins may help us to predict the course of antigenic evolution and, consequently, to select more appropriate vaccines and drugs.

Citation: Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB (2011) Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins. PLoS Genet 7(2): e1001301. doi:10.1371/journal.pgen.1001301

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, United States of America

Received August 4, 2010; **Accepted** January 7, 2011; **Published** February 17, 2011

Funding: JBP acknowledges funding from the Burroughs Wellcome Fund, the David and Lucile Packard Foundation, the James S. McDonnell Foundation, the Alfred P. Sloan Foundation, the Defense Advanced Research Projects Agency (HR0011-09-1-0055), and the US National Institute of Allergy and Infectious Diseases (2U54AI057168). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jplotkin@sas.upenn.edu

¤ Current address: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America

Introduction

Influenza A is among the most extensively studied viruses, owing to its importance as a human pathogen [1–6]. With a large, public database of genetic sequences, influenza viruses also offer a model system for studying molecular evolution in general. The evolution of influenza viruses is characterized by frequent reassortment events within subtypes [3,7] as well as high rates of amino-acid substitutions in the viral surface proteins hemagglutinin (HA) and neuraminidase (NA) [8–10]. Such high evolutionary rates reflect both the poor fidelity of the viral polymerase [10], and the strong selection pressures to evade the human immunity [8,9,11–13] and, more recently, to develop drug resistance [14–16].

Numerous experimental studies and statistical analyses of genetic and antigenic data have identified sets of residues in HA and NA proteins, the so called epitopes, that are bound by human antibodies [17–20]. As a consequence, the epitopic sites tend to evolve especially quickly, in order to evade immunity [8,21,22]. Moreover, several recent studies have suggested lists of amino acids at specific residues in HA that evolved under positive selection over the past 40 years [23–25].

In addition to escaping human antibodies, several other selective forces act on hemagglutinin. As with any functional protein, HA must maintain its stability and its function – namely, to bind the sialic acid receptor of host cells and subsequently mediate membrane fusion [17,18,20,26]. Thus, antibody escape mutations must not compromise these properties. Yet, numerous studies of protein evolution in vitro [27–29] as well as studies in bacteria [30] and viruses [20,31,32] have shown that beneficial mutations are often pleiotropic: in addition to their original beneficial effect, they cause some, usually negative, side effects on other protein properties, such as stability [28,33]. These negative effects can typically be alleviated or compensated by other mutations, making certain combinations of mutations substantially more beneficial than single mutations alone [34,35]. This phenomenon is known as positive epistasis between mutations [36]. Epistasis can also be negative, if a combination of mutations confers a smaller fitness gain than would be expected under additive effects of the individual mutations [36].

Author Summary

Epistasis describes non-additive interactions among genetic sites: the consequence of a mutation at one site may depend on the status of the genome at other sites. In an extreme case, a mutation may have no effect if it arises on one genetic background, but a strong effect on another background. Epistatic mutations in viruses and bacteria that live under severe conditions, such as antibiotic treatments or immune pressure, often allow pathogens to develop drug resistance or escape the immune system. In this paper we develop a new phylogenetic method for detecting epistasis, and we apply this method to the surface proteins of the influenza A virus, which are important targets of the immune system and drug treatments. The authors identify and characterize hundreds of epistatic mutations in these proteins. Among those identified, we find the specific epistatic mutations that were recently shown, experimentally, to confer resistance to the drug Tamiflu. The results of this study may help to predict the course of influenza's antigenic evolution and to select more appropriate vaccines and drugs.

Epistasis is commonplace in eukaryotes [37–39], bacteria [30,40,41], and viruses [31,34,35,42], and it plays an important role in the evolution of immune escape and drug resistance in various pathogens [35,43–46] including influenza [16,32]. Surprisingly, the extent of epistatic interactions in influenza proteins has not been systematically quantified or utilized. Yet, the knowledge of such interactions might provide a powerful tool for predicting future antigenically important substitutions and, consequently, for selecting better vaccine strains.

Numerous methods have been developed for detecting epistasis between mutations, based on sampled genetic sequences [47]. Early methods were based on the idea that co-evolving pairs of sites in a protein should leave a typical signature in a sequence alignment, which can be detected using quantities such as mutual information [48–50]. However, such methods ignore the phylogenetic relationships among sequences and so are justified only if the divergence times between samples are very large [51]. Various corrections for the phylogenetic non-independence have been proposed [52–54], and their performance has been shown to be satisfactory in some cases [55–57]. Nevertheless, methods that explicitly take account of the phylogeny are preferable [58]. Several such methods have been proposed recently [42,59–66]. Most of them attempt to detect unusually frequent co-occurrences of substitutions at pairs of sites on individual branches of the phylogeny. This approach is conservative since it detects only those positively epistatic pairs of sites for which a mutation at one site increases the beneficial effect of a mutation at the second site so dramatically that one mutation could not fix without the other one [42]. Such strong epistasis can occur, for example, when one mutation confers a strongly deleterious effect that is compensated by a second mutation. However, a mutation at one site in a protein may lead to only a moderate increase in the beneficial effect of a mutation at another site, so that the latter substitution occurs at an accelerated rate, but it does not necessarily appear exclusively on the same branch of the phylogeny [59,62,64,65]. In other words, substitutions at positively epistatic pairs of sites are likely to be temporally clustered [67]. In this paper, we exploit this idea to design an "epistasis statistic" that allows us to detect a broad class of epistatically interacting pairs of sites.

In essence, for each ordered pair of sites in a protein we measure the amount of phylogenetic time that typically elapses between a substitution at the first site and a subsequent substitution at its partner site. The epistasis statistic is defined as a decreasing function of this time interval. Thus, pairs in which the substitution rate at the second site tends to be increased after a substitution at the first site will have a larger value of the statistic. We obtain the null distribution of this statistic for all pairs simultaneously, by randomly shuffling the identities of substitutions on the phylogeny. We show that the number of site pairs in the surface proteins of the human influenza A/H3N2 virus with large values of the epistasis statistic significantly exceeds the null expectation—thus, influenza surface proteins evolve under substantial positive epistasis. We characterize the epistatically interacting sites we have inferred in terms of their overall patterns of evolution, protein locations, and functional significance. For type-1 neuraminidase, we compare the identities of the epistatic sites we have inferred with those that have been experimentally verified. We discuss the implications of our results both for practical issues surrounding influenza's antigenic drift and drug resistance, and for broader issues surrounding protein evolution in general.

Results

Identifying epistatic pairs of sites

We reconstructed the phylogenetic trees for HA and NA proteins (subtypes H3N2 and H1N1) and inferred the nucleotide sequences at internal nodes by maximum likelihood as described in "Materials and Methods". In order to detect pairs of codon sites in a protein that have evolved under positive epistasis we used the "epistasis statistic" described in "Materials and Methods".

Briefly, the epistasis statistic considers an ordered pair of sites, the first of which is called the "leading site" and the second is called the "trailing site". The epistasis statistic tends to be large for pairs of sites (i,j) in which a non-synonymous substitution at site j tends to quickly follow a non-synonymous substitution at site i, and for which substitutions at the trailing site occur in multiple lineages (see schematic in Figure 1). We measure time between a pair of non-synonymous substitutions as the number of synonymous substitutions that occur between them. Since we are interested in positive epistasis and would like to detect only those pairs of substitutions in which the second substitution is beneficial, we excluded all substitutions at terminal branches, because many such substitutions are likely to be deleterious. We also discarded all sites that experienced fewer than two substitutions at the internal branches (Table 1).

The epistasis statistic depends on the parameter τ that sets the timescale over which substitutions contribute information. Pairs of substitutions that are separated by times much shorter than τ contribute significantly to the epistasis statistic, whereas pairs of substitutions that are separated by times much longer than τ do not. We set τ to be equal to the average time $\langle t_{\mathrm{ns}} \rangle$, measured in the number of synonymous substitutions, that elapses between two non-synonymous substitutions randomly sampled from a phylogeny (see Text S1 and Table 1). For each phylogeny (H1, N1, H3, and N2) and its corresponding value of τ, we computed the epistasis statistic for all qualifying ordered pairs of sites and, for each such pair, we computed the distribution of the epistasis statistic under the non-epistatic null hypothesis (see "Materials and Methods" for details). We then selected all pairs of sites whose nominal P-value was smaller or equal to 0.01. In H3, we identified 333 site pairs with a nominally significant epistasis statistic; we identified 225 such pairs in H1, 205 such pairs in N1, and 188

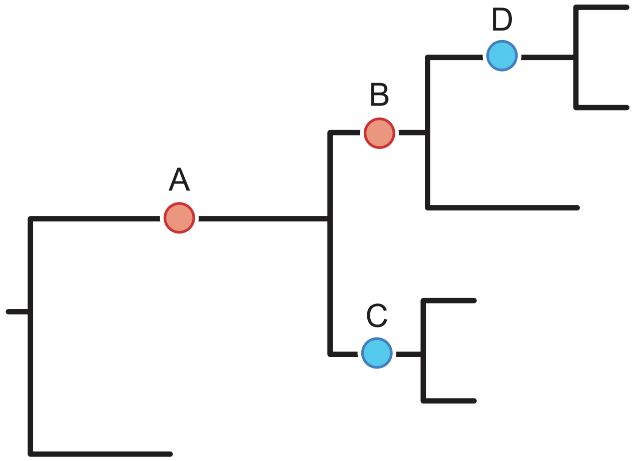

Figure 1. Detecting positive epistasis between mutations at two sites, *i* and *j*. The epistasis statistic is defined in terms of the total time elapsed between all pairs of consecutive substitutions at sites i and j (see "Materials and Methods"). In this schematic figure, substitutions at sites i and j are denoted by red and blue circles, respectively. Substitutions (A,C) and (B,D) form consecutive pairs. Substitutions (B,C) are not consecutive because they occur on different lineages. Substitutions (A,D) are not consecutive because substitution B at site i occurs between them.
doi:10.1371/journal.pgen.1001301.g001

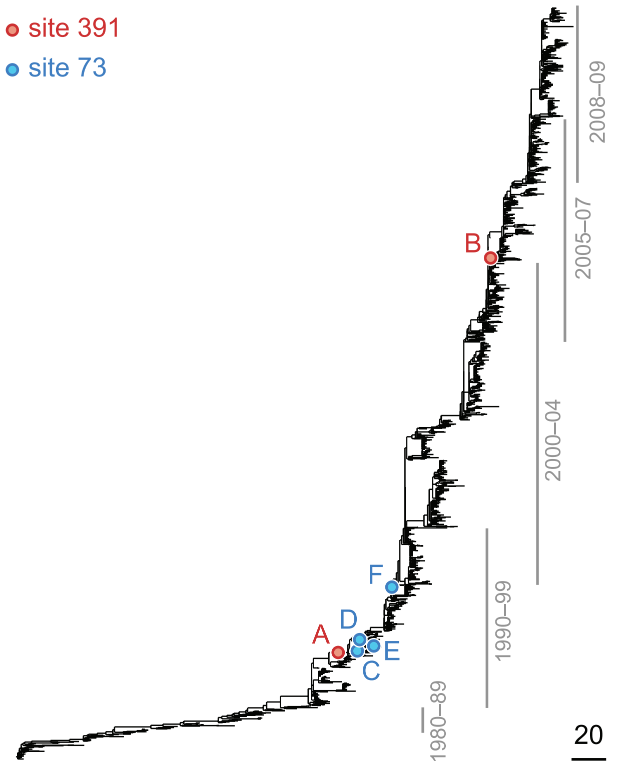

Figure 2. Phylogenetic tree of HA (subtype H3) illustrating a putatively epistatic interaction between sites 391 (red circles) and 73 (blue circles). Site 391 is not in an epitope, $\mathrm{dN/dS} = 0.47$; site 73 is epitope E, $\mathrm{dN/dS} = 0.72$. Only substitutions at internal nodes are displayed. Branch lengths are equal to the total number of substitutions across all sites. Vertical bars show the approximate years in which the sequences were isolated. Substitutions C, D, E, and F at site 73 closely follow substitution A at site 391, leading to a highly significant value of the epistasis statistic ($E_{63}(391,73) = 3.48$, nominal $P\text{-value} < 10^{-4}$). As a result, the ordered pair of sites (391, 73) is detected as epistatic by our method. At the same time, only a single substitution, B, at site 391 follows substitution F at site 73 – and only after a long period of time – resulting in a low value of the epistasis statistic for the inverse pair ($E_{63}(73,391) = 1.20$, nominal $P\text{-value} = 0.18$). Therefore, the ordered pair of sites (73,391) is not detected as epistatic by our method.
doi:10.1371/journal.pgen.1001302.g002

such pairs in N2 (see Table 1, and Table S1). Examples of epistatic site pairs in HA and NA are shown in Figure 2 and Figure 3.

We computed the false discovery rate (FDR) as well as the overall P-value for the observed number of significant pairs (see "Materials and Methods"). Although the FDR in all proteins was high, around 60%, the observed number of nominally significant pairs was much larger than would be expected by chance ($p < 0.015$, see Table 1). Reducing the nominal P-value cutoff somewhat reduced the FDR but also disproportionately reduced the number of inferred positives (see Figure S1).

We tested the sensitivity of our method with respect to the choice of the timescale parameter, τ in the range from $\tau = 1$ to $\tau = 200$, as well as to uncertainty in phylogeny and internal node reconstruction. The results remained qualitatively similar to those reported here (see Text S1, Figure S2, and Table S4). As a negative control, we performed 100 simulations in which sites evolved independently (i.e. without epistasis) along a given phylogenetic tree (see "Materials and Methods"). In 52 out of 100 simulations, the number of significant pairs at the P-value cutoff 0.01 was smaller than expected, in 47 cases this number was larger that expected, but not significantly so. In only one simulation out of 100 was this number larger than expected and significant ($P < 0.01$). We are therefore confident that our method indeed detects epistatic pairs of sites, and it does not systematically report more false positives than our FDR computation indicates.

Table 1. Summary of data used in our analysis.

	H3N2		H1N1	
	H3	N2	H1	N1
number of sequences	2149	2339	1219	1836
PDB accession number	2VIU	1NN2	1RUZ	3BEQ
protein length	566	469	565	470
number of epitopic sites[1]	131	45	–	–
sites considered	141	111	115	113
pairs considered	19740	12210	13110	12656
τ used	63	54	78	60
pairs significant at 0.01 (exp)	196	122	132	122
pairs significant at 0.01 (obs)	333	188	225	205
FDR, %	58.8	64.8	58.7	59.3
P-value[2]	<0.01	0.015	<0.01	<0.01

[1] epitopic sites are taken from [8,17] for H3, from [8,19,20] for N2;
[2] P-value is for the number of nominally significant pairs. FDR stands for "false discovery rate" (see "Materials and Methods").
doi:10.1371/journal.pgen.1001301.t001

Characterizing epistatic pairs of sites

Having obtained a list of pairs of sites with putative epistatic interactions (Table S1, Figure 2 and Figure 3), we inspected the properties of these pairs, compared to an appropriate null set. In particular, we compared the true, epistatic pairs to the pairs that had appeared as nominally significant in the 400 "fake data sets"

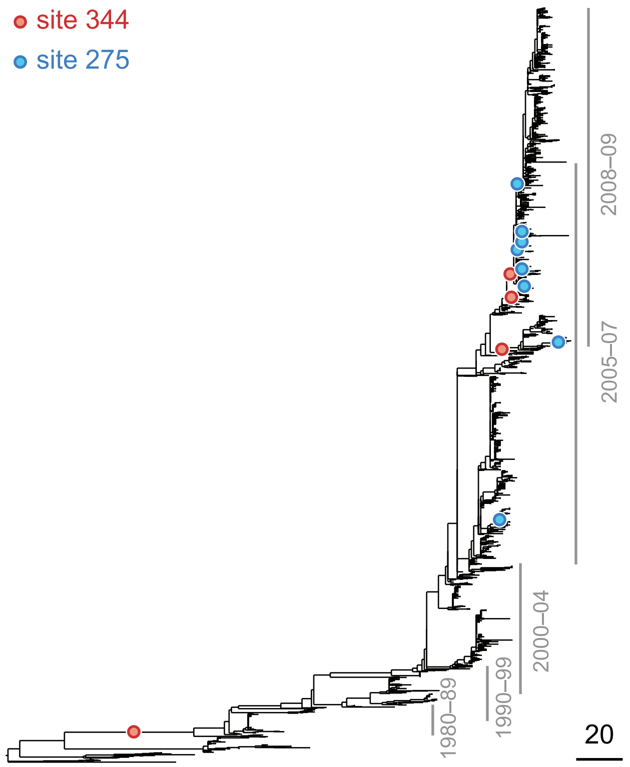

Figure 3. Phylogenetic tree of NA (subtype N1) illustrating a putatively epistatic interaction between the leading site 344 (red circles) and the trailing site 275 (blue circles). Other notations are as in Figure 2.
doi:10.1371/journal.pgen.1001301.g003

produced by permutation (see "Materials and Methods"). Thus, we asked whether the pairs that we detected as epistatic differed systematically from the false positive pairs. We investigated three types of properties: the average dN/dS value at epistatically interacting sites, their location in the protein with respect to known epitopes (for H3N2 only), and the distances between interacting sites. For comparison of physical and linear distances we also excluded sites that were not present in the resolved crystal structure (see "Materials and Methods"). The results are summarized in Table 2 and Table 3, and discussed below.

The dN/dS values at the leading sites among the putative epistatic pairs were not significantly larger, on average, than at the leading sites among pairs identified under permutation. However, the average dN/dS value at the leading sites was less than one, which is usually interpreted as evidence of purifying selection. Therefore, without an analysis of epistasis, many of the leading sites would not have been identified as experiencing positive selection, even though they may play a critical role in facilitating adaptation in co-ordination with substitutions at their partner sites.

By contrast to the leading sites, the dN/dS values at the trailing sites were significantly larger than the null expectation, and they exceeded one on average (with the exception of N1). Thus, the trailing sites exhibit the characteristic signature of positive selection [68], even though the positive selection they experience was likely made possible (or, at least, more likely) by preceding substitutions at their corresponding leading sites. In other words, many of the positively selected substitutions that have occurred in HA and NA may have been facilitated by previous substitutions at epistatically interacting sites. In previous work, we have identified 25 sites in H3 at which certain specific amino acids evolved under directional positive selection [25]. Interestingly, 22 of these sites appear to be involved in epistatic interactions: 10 sites appear as leading sites, 7 sites appear as trailing, and 5 sites appear as both.

We also studied the location of epistatic sites with respect to the known antigenic regions of the influenza surface proteins, for H3 and N2. In both proteins, the leading site in an epistatic pair was more likely to fall within an antigenic epitope than under the null expectation (and significantly so in H3). The trailing site was slightly more likely to fall outside of known epitopes, despite the fact that the dN/dS ratio was typically greater than 1 at such sites. Thus, pairs of sites in which the leading site was in an epitope and the trailing site was not in an epitope, were typically overrepresented (although not significantly so). This suggests that the leading sites may often be directly involved in antigenic escape and the trailing sites may subsequently compensate for deleterious (e.g. destabilizing) side effects of the initial mutation. In some cases, however, both the leading and trailing sites of an epistatic pair fall within epitopes (47% of pairs in H3, 9% of pairs in N2). In such cases, for H3, the leading and trailing sites were significantly more likely to fall in *different* epitopes from each other, than expected – suggesting that substitutions across multiple epitopes may be particularly important for antigenic escape, at least in hemagglutinin of the H3 subtype. This observation reflects the widely held belief that antigenic change in hemagglutinin typically requires multiple substitutions spread across multiple epitopes [21,26].

How far apart are the leading and trailing sites of epistatic pairs? Surprisingly, neither the average linear (sequence) distance nor the physical distance between the leading and the trailing sites in an epistatically interacting pair was significantly smaller than would be expected among false positive pairs.

Finally, we investigated the timing of consecutive substitutions at the leading and trailing sites in epistatic pairs. On average, both across pairs and across consecutive substitutions, a substitution at a leading site in H3 was followed by a consecutive substitution at its corresponding trailing site 3.7 years later (see Text S1 for details). Similarly, in H1 the mean time between consecutive substitutions was 5.8 years; 4.4 years in N1; and 4.2 years in N2. In all cases, the mean time between consecutive substitutions exceeds two years – which suggests that the observation of a substitution at the leading site of a known epistatic pair may provide useful predictive value for anticipating a subsequent substitution at its corresponding trailing site, within the time-frame required for selecting a seasonal vaccine strain [69].

Epistasis and the evolution of oseltamivir resistance

Our analyses of substitution patterns suggest that positive epistasis is prevalent among sites in HA and NA. However, it is important to verify experimentally that the identified pairs of sites indeed show non-additive fitness effects. Fortunately, such verification has recently been performed for two specific pairs of sites in type-1 neuraminidase.

Currently circulating variants of the seasonal H1N1 subtype are resistant to the drug oseltamivir, which inhibits neuraminidase [15]. Resistance to this drug is conferred by the mutation His→Tyr at site 275, which is referred to as the "H274Y" mutation in the literature [15]. However, this mutation is known to be strongly deleterious in the absence of the drug [70]. Recently, Bloom et al. demonstrated that mutations R222Q and V234M restore the drug-resistant mutant's fitness in vitro [16], for seasonal H1N1. They also observed that mutations R222Q and V234M were fixed in the seasonal H1N1 population prior to the emergence of the H275Y mutation, and thus they likely acted as epistatic "permissive mutations" for the emergence of drug resistance in competent viruses.

Table 2. Characterization of epistatic pairs in subtype H3N2 surface proteins, compared to expectations for randomly chosen pairs.

		H3		N2	
		exp	**obs**	**exp**	**obs**
leading site	average dN/dS	0.88	0.91 (ns)	0.84	0.70 (ns)
	fraction ept	0.59	0.82 *	0.26	0.43 (ns)
trailing site	average dN/dS	0.86	1.12 **	0.83	1.29 **
	fraction ept	0.58	0.54 (ns)	0.27	0.20 (ns)
fraction of pairs	both npt	0.17	0.11 (ns)	0.54	0.46 (ns)
	(npt, ept)	0.24	0.07 *	0.20	0.12 (ns)
	(ept, npt)	0.25	0.35 (ns)	0.19	0.34 (ns)
	same ept	0.07	0.07 (ns)	0.03	0.03 (ns)
	diff. ept	0.27	0.40 *	0.04	0.06 (ns)
time btw. consec. subst. (in syn subst)	mean	N/A	14.71	N/A	13.82
	std	N/A	16.76	N/A	15.72
time btw. consec. subst. (in years)	mean	N/A	3.68	N/A	4.22
	std	N/A	4.20	N/A	4.80
average distance[1]	linear	135.5	114.3 (ns)	124.3	124.3 (ns)
	physical, Å	43.1	38.9 (ns)	28.6	28.2 (ns)

P-values are obtained from two-tailed tests, except for the last two rows which report one-sided tests regarding distances between sites. Single and double asterisks denote significance at 0.05 and 0.01 level, respectively; "ns", "ept", "npt", "obs", "exp", and "N/A" denote not significant, epitopic, non-epitopic, observed, expected, and not applicable respectively.

[1] Average distances are computed only over those significant pairs in which both residues are present in the crystal structure (see "Materials and Methods").

doi:10.1371/journal.pgen.1001301.t002

Our statistical analysis of epistasis in N1, based on patterns of sequence evolution alone, is remarkably concordant with the experimental findings of Bloom et al. In particular, our analysis indicates that sites 222 and 234 interact strongly with site 275 (see Table S1). Moreover, among the top 10 most significantly epistatic pairs in N1 there are 6 other pairs that involve the drug-resistance site 275 as the trailing site; the leading sites in these pairs are 214, 287, 329, 354, 382, and 344. In all cases the subsequent mutation at site 275 is His→Tyr. Therefore, aside from sites 222 and 234, our analysis predicts that these six additional sites may be permissive mutations that, in combination with H275Y, produce competent, drug-resistant viruses. Although no epistasis between two of these sites (214 and 382) and site 275 was found experimentally [16], one of the mutations (D344N) has subsequently been shown to help counteract the decrease in total surface-expressed activity associated with the mutant neuraminidase ([71] and Jesse Bloom, personal communication), and it, along with 224 and 234, may have played a role in the emergence of oseltamivir resistance in seasonal H1N1 viruses before 2009.

Although further experimental validation is required, the remarkable concordance between our statistical inferences and experimentally verified epistatic interactions [16] suggests that patterns of sequence evolution contain extremely useful information about a protein's fitness landscape. In the case of oseltamivir resistance, this information is highly specific and of significant import to public health.

Table 3. Characterization of epistatic pairs in subtype H1N1 surface proteins, compared to expectations for randomly chosen pairs.

		H1		N1	
		exp	**obs**	**exp**	**obs**
leading site	average dN/dS	0.94	0.89 (ns)	0.75	0.91 (ns)
trailing site	average dN/dS	0.93	1.90 **	0.73	0.76 (ns)
time btw. consec. subst. (in syn subst)	mean	N/A	25.56	N/A	14.75
	std	N/A	28.64	N/A	17.09
time btw. consec. subst. (in years)	mean	N/A	5.80	N/A	4.40
	std	N/A	6.50	N/A	5.10
average distance	linear	117.4	106.9 (ns)	130.1	108.6 (ns)
	physical, Å	37.5	37.6 (ns)	27.7	28.6 (ns)

Notations as in Table 2.

doi:10.1371/journal.pgen.1001301.t003

The oseltamivir resistance mutation, H275Y, was known in advance of the drug's widespread introduction. Moreover, this prior knowledge was used by Bloom et al. [16] to focus their experimental search for an epistatic partner to site 275. Nonetheless, our method of identifying epistatic pairs from sequence data implicates site 275 – without any prior knowledge of its role in drug resistance – as extremely important in the adaptive evolution of N1, especially in combination with sites 222, 234, and six other leading sites (Table S1). This demonstrates the practical, predictive power of our method for inferring the specific, epistatic interactions that shape viral adaptation. Thus, our method may, in the future, help us identify sites important for drug resistance or antigenic drift, even when no prior experimental data are available.

Finally, we note that our analysis implicates sites 222 and 234, which have been verified as important epistatic partners of the oseltamivir resistance site 275, as significant epistatic leading sites even when we restrict our data set to those viral isolates prior to the introduction of oseltamivir. In particular, based on sequence data prior to 2001, our method identifies sites 222 and 234 as participating in epistatic interactions with sites other than 275 (see Table S1). Thus, sites 222 and 234 may be structurally important and experience epistatic interactions even in the absence of selection for oseltamivir resistance.

Discussion

We have developed a statistical method to detect positive epistasis between pairs of sites in a protein, based on patterns of thoroughly-sampled sequence variation. The essential idea underlying this method is simple: a substitution at one site should rapidly follow a substitution at another site if the sites interact epistatically. We applied this method to identify putative epistatic pairs in the influenza surface proteins hemagglutinin and neuraminidase, and we found a highly significant number of interacting pairs. We characterized the properties of the leading and trailing sites identified as epistatic. Finally, we validated our approach by comparison to experimentally verified epistatic interactions in neuraminidase, with significant implications for public health.

This study sheds some light on methodological and empirical questions in molecular evolution generally, as well as practical questions about influenza viral evolution in particular. Methodologically, it is instructive to compare our approach to identifying epistasis with other techniques in the literature. Over very long timescales, interacting sites in a protein have been identified by inspecting multiple sequence alignments, ignoring the phylogenetic relationship among the sequences being compared. Such an approach is justifiable over timescales so long that each site may be treated independently, and indeed it has proven successful at identifying epistasis in proteins conserved across all domains of life [51]. However, such techniques are not justified for shorter timescales, because correlations between sites may arise simply as the result of linkage and shared ancestry [58]. Although techniques exist to control for phylogeny in such tests [52–54], it is preferable to leverage the phylogeny in the design of a more powerful statistic for epistasis – which is the approach we have taken here.

Even among the techniques that account for phylogeny, methods differ in their power to detect epistasis. Some methods will be more powerful in some contexts, and others in other contexts – depending upon the structure of epistasis among sites, the selection coefficients involved, and the density of sampling. Most existing methods that utilize phylogenetic information assume that epistatic substitutions will co-occur along the same branch of the phylogeny [42,60,63,64]. This assumption will not always be met, however, if the selective advantage conferred by a substitution at the trailing site is only moderate; in such cases, substitutions at trailing sites will occur at an accelerated rate but they may likely fall on subsequent branches in the phylogeny. To demonstrate this point, we applied the method of Poon et al [64], implemented in the HyPhy package [72], to the same data set of influenza sequences. That method detected 4 to 10 times fewer epistatically interacting pairs of sites than our method did, at the same false discovery rate (see Text S1 and Tables S2 and S3). Importantly, the method by Poon et al. failed to detect epistatic interactions between sites 222 and 234 and the drug-resistance site 275 in neuraminidase subtype N1, even though those pairs were highly ranked by our method and those epistatic interactions were confirmed experimentally. Although a thorough comparison between various existing methods is beyond the scope of this paper, we believe that the additional power of our method to detect epistasis in the influenza data arises because we allow for time lags between substitutions at interacting sites.

The epistasis statistic developed here is admittedly ad-hoc, compared to systematic, likelihood-based methods for jointly inferring phylogeny and epistasis under Markov substitution models [73–75]. At the same time, the vast dimensionality associated with substitution models incorporating pairwise epistasis, of order $(20L)^2$ for a sequence of length L, is daunting; whereas the frequentist statistic defined here seems to perform quite well. The strong performance of our approach likely arises from our ability to infer ancestral states reliably, due to the high-resolution sampling of influenza sequences.

Our method has several important shortcomings. One drawback is that it requires a large number of substitutions per site in order to discriminate between truly interacting site pairs and pairs that sustain substitutions in close succession just by chance. Moreover, even if the protein evolves rapidly, as influenza surface proteins do, the false discovery rate is still very high. Our method will likely perform much worse for proteins that evolve slowly or have been sampled sparsely.

Two other concerns are problematic for our approach, as well as most other methods of detecting epistasis from phylogenetic data. Such approaches generally suffer from the inability to weed out spuriously epistatic pairs, which leads to high false discovery rates. There are at least two sources of spuriously significant pairs: hitchhiking and coordinated temporal variation in selection pressures across sites. Imagine, for example, that sites A and B interact epistatically and that multiple substitutions at site B in independent lineages rapidly follow a single substitution at site A. Then site pair (A,B) would be detected by our method. However, if the variant that carries the leading mutation at site A also, by chance, happens to carry a mutation at site C (which is not epistatic with B), then mutation C hitchhikes to fixation together with mutation A and so the site pair (C,B) may also be detected as epistatic. In fact, mutation at site C may be advantageous while mutation at site A may be a neutral or slightly deleterious mutation that hitchhikes to fixation together with C, but then "permits" the beneficial mutation at site B. It may be possible to reduce the false discovery rate by designing statistics that consider only those site pairs for which consecutive substitutions involve multiple independent substitutions at the leading site as well as the trailing site.

Coordinated temporal variation in selection pressures across sites is another source of potential false positives under this and other tests of epistasis. Consider, for example, sites 391 and 73 in H3 illustrated in Figure 2. Substitutions at site 73 appear to quickly follow substitutions at site 391 in the early 1990's. Apart from epistasis, the clustering of substitutions at these two sites could be explained if both sites independently experienced positive selection during this time period, and otherwise negative selection. However, if this explanation were the dominant one for the observed clustering of substitutions, then, for each nominally significant ordered pair of sites, we would expect its inverse pair also to be nominally significant, on average. Yet, we do not find a single nominally significant pair whose inverse pair is also nominally significant, even though consecutive substitutions do occur in the direct and reverse order (see for example, Figure 2 and Table S1). It is unlikely that this observation is caused by insufficient sampling. Indeed, in H3, there are typically more than 6 substitutions (at internal branches) at either a leading or trailing site in an identified epistatic pair, and similarly for the other proteins. Moreover, many sites appear in our lists as both leading and trailing. Thus, leading and trailing sites exhibit similar number and pattern of substitutions, and there is plenty of power to detect a significant epistasis statistic in both directions. This suggests that the excess of significant pairs we observe is likely caused by epistasis, rather than coordinated temporal variation in selection pressures.

Another shortcoming of our method is that it aims to detect epistasis only between pairs of sites, whereas interactions among residues in a protein are certainly more complex. This may be a cause of our large false discovery rate. Imagine, for example, three sites, A, B, and C, such that pairs (A,B) and (B,C) interact epistatically, but the pair (A,C) does not. If substitutions at site B quickly follow substitutions at site A and if substitutions at site C quickly follow substitutions at site B, our method may detect the pair (A,C) as epistatic, even though there is no direct epistatic interaction between these sites. Indeed, in our list of putatively epistatic pairs, we find 133 of such "circular" triplets in H3, 41 triplets in N2, 81 triplets in H1, and 71 triplets in N1. In order to discriminate truly epistatically interacting site pairs from spurious pairs, it may be possible to modify the Bayesian graphical models recently used for detecting epistasis in HIV [64,66] to incorporate a time lag between consecutive substitutions.

Finally, although our method detects epistatic interactions between pairs of sites, it does not determine which specific mutations at those sites were epistatic. Extending our permutation technique to incorporate the information about specific mutations may prove difficult, but in many cases it is unnecessary. Often we can *a posteriori* identify the specific mutations that led to a significant value of the epistasis statistic for a pair of sites. For example, the drug-resistance site 275 is identified as trailing with many leading sites in N1 (see Table S1), but the specific substitutions at site 275 are all in fact identical: H275Y.

Methodological issues aside, our results on epistasis in HA and NA have several important practical implications for our understanding of influenza evolution. We have demonstrated that our method reliably infers a critically important oseltamivir resistance site, as well as the associated leading sites at which initial mutations are required for the production of a viable, drug-resistant virus. Remarkably, we can identify some of the leading sites (222 and 234) even when restricted to sequence data prior to the introduction of the drug. This degree of specificity and accuracy may prove helpful in preparing for resistance to other drugs that may be developed, or in predicting the emergence of oseltamivir resistance in the recent type-1 swine neuraminidase responsible for the 2009-10 influenza pandemic.

In addition to NA, we have also detected substantial amounts of epistasis in HA, including in the known epitopes, likely associated with antigenic drift. Knowledge of specific pairs of sites that interact epistatically in HA may improve our ability to predict future antigenic variants, and thus to calibrate vaccine strain choices accordingly. Previous studies of HA antigenic evolution have focused almost exclusively on those sites with the strongest signatures of positive selection, e.g. elevated dN/dS ratios [8,21,22]. However, our results suggest that this approach will inevitably miss many sites of genuine importance to adaptation, and will implicate others that are not directly involved in antigenic escape. In particular, we have seen that the leading site of an epistatic pair often falls within an epitope, but it also often exhibits $\mathrm{dN/dS}<1$. In contrast, the trailing site typically falls outside of an epitope and it exhibits significantly elevated $\mathrm{dN/dS}>1$. This observation appears counter to our expectation that epitopic sites have elevated dN/dS values and non-epitopic sites have depressed dN/dS values. However, not all epitopic sites experience elevated dN/dS values at all times because different epitopes may be immunodominant at different times [76,77]. Thus, an average dN/dS value at a site may be well below 1 even if this site occasionally evolves under strong immune selection [78].

The patterns of epistatic interactions we have detected suggest the the following speculative model for the evolution of influenza surface proteins. If an epitope is immunodominant at a certain period of time, the pressure for antigenic escape is so strong that the leading site, of antigenic importance, substitutes despite a negative side-effect (e.g. diminished protein stability or function). This side-effect is subsequently compensated by a substitution at another, relatively unconstrained, site, in an epitope or not. It is possible that such unconstrained sites may act as "global suppressors", i.e., compensate the destabilizing effects of many mutations [79,80]. If there is a constant need for compensation (caused by antigenically important mutations), such compensating sites will continually evolve under positive selection and will exhibit $\mathrm{dN/dS}>1$. Under this scenario, the roles of the two sites would both be misinterpreted by an analysis based on dN/dS alone that neglects epistasis. In particular, our observations imply that mutations at sites with $\mathrm{dN/dS}<1$ may sometimes be extremely important for antigenic adaptation, even though they have been largely ignored in compilations of antigenically relevant sites [8,11,21]. Conversely, some mutations at sites with $\mathrm{dN/dS}>1$ may be unimportant for antigenicity per se, but are positively selected simply to compensate for prior antigenic escape mutations with deleterious side effects. Another potential mechanism of epistasis in influenza surface proteins could be the dynamic balance between mutations that simultaneously influence receptor binding avidity and antigenicity, as suggested recently by Hensley et al. [81].

Some of the epistatic pairs that we detect consist of an apparently neutral but "permissive" mutation at the leading site followed by a highly advantageous mutation at the trailing site, such as the pair of mutations V234M and H274Y in N1 previously identified by Bloom et al [16] and also detected by our method (Table S1). This observation is consistent with the idea that neutral or nearly neutral substitutions can facilitate adaptation at partner sites that might not otherwise have been available—a concept that has received much attention in theoretical studies of adaptation [82,83] and of influenza evolution in particular [84].

Finally, we have observed that epistatic residues do not tend to be significantly closer to each other in the folded protein structure than would be expected by chance – a result that we expect to hold generally, and which suggests that structural influences on epistasis are probably not as straightforward as simple proximity of residues. In the future, it will be important to investigate, by computation or experiment, whether epistatic partner sites are compensating for protein stability even if they are distant from each other in a folded protein structure.

Materials and Methods

Data

We downloaded all HA and NA coding region sequences of human influenza A virus subtypes H1N1 and H3N2 that were available in the NCBI's Influenza Virus Resource [85] in June 2010. The amino acid sequences were aligned using Clustal W ver. 1.83 [86] and the alignments were reverse translated using PAL2NAL [87]. Occasional gaps in the alignments were filled if more than 70 percent of sequences agreed on the nucleotide at the gap position; otherwise the sequence with a gap was excluded from further analysis. To test some aspects of our method, we used a smaller HA data set (subtype H3N2) which was downloaded in April 2009. To investigate whether our method detected any of the known epistatic site pairs in type-1 NA prior to the introduction of oseltamivir, we also used a truncated data set of N1 sequences with all sequences isolated subsequent to 2001 removed. All used alignments are available upon request.

In computing the epistasis statistic we excluded all substitutions at terminal branches, and we discarded all sites that experienced fewer than 2 substitutions at the internal branches.

We also downloaded the HA and NA crystal structures from the RCSB Protein Data Bank. In computing the linear (sequence) and physical distances between residues we excluded all residues that were not resolved in the crystal structures. We used the distance between the alpha-carbon atoms as a proxy for the physical distance between residues.

Phylogeny reconstruction and substitution mapping

We reconstructed the maximum likelihood phylogenetic trees for HA and NA using PHYML [88] under the GTR substitution model with the four-category discrete approximation of the gamma distribution for the substitution rates. We reconstructed the nucleotide sequences at the internal nodes of the phylogeny using maximum likelihood algorithm in PAUP* 4.0b10 [89]. For each codon site, we identified whether it experienced at least one synonymous and/or non-synonymous substitution on each branch of the reconstructed phylogeny. In those rare cases in which a codon experienced more than one substitution of the same kind (synonymous or non-synonymous) on a branch, we did not record the number of substitutions, in order to simplify computations.

The epistasis test statistic

Consider an ordered pair of sites (i,j) in the protein of interest. In order to detect a positive epistatic interaction for this pair, we designed a statistic that detects the acceleration of non-synonymous substitutions at site j, which we call the *trailing site*, after the occurrence of a non-synonymous substitution at site i, which we call the *leading site*.

First, we obtained a strict temporal order in which non-synonymous substitutions at sites i and j occurred on the phylogenetic tree. Such an order is not actually known if the phylogeny contains one or more branches on which both sites have experienced a non-synonymous substitution. We say that such branches are *temporally unresolved with respect to the pair* (i,j). Since we do not know in which order the sites in the pair have experienced substitutions on a temporally unresolved branch, we assume that both orders are equiprobable. If there are m_{ij} branches on the phylogenetic tree that are temporally unresolved with respect to the pair (i,j), there are a total of $2^{m_{ij}}$ equally likely distinct strict temporal orders of substitutions on the tree with respect to this pair.

Next, for each strict temporal order of substitutions $O_{ij}^{(k)}$, $k=1,\ldots,2^{m_{ij}}$, at sites i and j, we find all pairs of substitutions that are consecutive along the tree. Substitution B at the trailing site (in this case j) and substitution A at the leading site (in this case i) form a *consecutive pair* $\pi=(A,B)$ if A has occurred in the lineage ancestral to B and no other substitution at either site has occurred in the lineage between them. This notion is illustrated in Figure 1. If (A,B) is a consecutive pair, we also say that substitution B *is consecutive to* substitution A. For each consecutive pair (A,B), substitution A is called *initial* and substitution B is called *subsequent.*

Before defining the epistasis statistic we introduce some notation. We denote the fact that branch a is ancestral to branch b by $a\prec b$ ("a precedes b") or by $b\succ a$ ("b follows a"); if a and b denote the same branch, we naturally write $a=b$. We denote the number of synonymous substitutions that occurred on branch a by l_a. We measure time t_π between the initial substitution A and the subsequent substitution B of a consecutive pair $\pi=(A,B)$ as the expected number of synonymous substitutions that occurred between them. More precisely, if substitutions A and B occurred on branches a and b respectively, then

$$t_\pi\equiv t(a,b)=\begin{cases} l_a/3, & \text{if } a=b \\ l_a/2+\sum\limits_{\substack{c:\\ a\prec c\prec b}} l_c+l_b/2, & \text{if } a\prec b \end{cases} \tag{1}$$

The sum in this expression is taken over all branches on the lineage connecting branches a and b. Note that t_π can be zero if no synonymous substitutions occurred between substitutions A and B.

Let $S_{ij}^{(k)}$ denote the set of all consecutive substitution pairs at site pair (i,j) found on the phylogenetic tree with the order of substitutions $O_{ij}^{(k)}$. We define the epistasis statistic as

$$E_\tau(i,j)=\frac{1}{2^{m_{ij}}}\sum_{k=1}^{2^{m_{ij}}}\sum_{\pi\in S_{ij}^{(k)}}\exp\{-t_\pi/\tau\}, \tag{2}$$

where $\tau>0$ is a time-scale parameter that we specify in the "Results" section. The choice of an exponential function of t_π is arbitrary. We expect that any monotonic decreasing function of t_π would yield similar results. Note that if $m_{ij}=0$, i.e. if sites i and j never experience a non-synonymous substitution on the same branch, the strict temporal order of substitutions with respect to the ordered pair (i,j) is unique. In this simple case, the epistasis statistic is equivalent to

$$E_\tau(i,j)=\sum_{\pi\in S_{ij}^{(1)}}\exp\{-t_\pi/\tau\}.$$

If we take the time-scale parameter τ to be infinite, the statistic $E_\infty(i,j)$ simply equals the number of consecutive substitutions for the ordered pair (i,j). We also define $E_\tau(i,j)\equiv 0$ if all sets $S_{ij}^{(k)}$, $k=1,\ldots,2^{m_{ij}}$ are empty for the pair (i,j). In other words, the epistasis statistic is zero for pairs of sites that never experience consecutive substitutions.

The value of the epistasis statistic $E_\tau(i,j)$ is large if substitutions at the trailing site often follow substitutions at the leading site and if the time-lag between initial substitutions at the leading site and subsequent substitutions at the trailing site is typically small (compared to τ). We therefore expect that pairs of sites that evolve under positive epistasis will have a larger value of the epistasis statistic.

The non-epistatic null hypothesis

If there were no epistatic interactions between sites and no temporal variation in selection pressures then we would expect the non-synonymous substitutions at each site to be distributed randomly on the phylogenetic tree. In order to obtain the distributions of the epistasis statistic under this null hypothesis for all ordered pairs of sites, we utilize the following straightforward permutation procedure. We shuffle all non-synonymous substitutions on the phylogenetic tree while keeping two sets of marginal quantities preserved: (a) for each branch of the phylogeny, we preserve the number of non-synonymous substitutions that occurred on that branch and (b) for each site, we preserve the total number of non-synonymous substitutions that occurred at that site on the tree. Condition (a) ensures that any possible temporal biases in the sampling of viral isolates, which would apply equally to all sites, are preserved in the null distribution. Condition (b) ensures that the overall selective constraint on each site is preserved. Synonymous substitutions are unaffected by this shuffling procedure.

Although this permutation procedure is conceptually simple, its computational implementation is challenging. A priori, it is unclear how to efficiently sample the space of possible substitution configurations while preserving the aforementioned marginals. This problem can be rephrased as follows. We can represent the phylogenetic tree with non-synonymous substitutions as an $M \times N$ matrix, where M is the number of branches on the phylogeny and N is the number of sites, so that each cell in the matrix is either 1 or 0 depending on whether or not the given site experienced a non-synonymous substitution on the given branch. Thus, we would like to randomly permute the entries of this matrix while preserving the row and column sums. This problem is equivalent to the problem of obtaining the null distribution for the matrix of associations between individuals across a set of observations, which has been extensively studied in the ecology literature [90–92]. The method typically employed in ecology to sample the space of matrices that satisfy the constraint on the marginals is called the "swap method" and is based on the idea of swapping the entries of certain specific 2×2 submatrices in a way that does not violate the constraints. This method, although computationally efficient, generates matrices that are not independent [91]. An alternative "fill method" permutes the matrix entries and simply discards those resulting matrices that do not satisfy the constraints [92]. This method can be prohibitively computationally expensive if many matrices are discarded, but it guarantees independent sampling.

We employed the "fill method" and found that only about between 0.05% and 5% of matrices are accepted, yet this did not present a serious computational limitation. We generated 10^4 valid permutations per protein, which required about 10 minutes on a desktop computer.

Computing the false discovery rate

We compute the value of the epistasis statistic and its associated nominal P-value for many thousands of site pairs. We therefore need to quantify the fraction of false positives among the observed nominally significant pairs [93]. Because the values of the epistasis statistic for different site pairs are not independent, we estimate the distribution of the number of false positives in the data through bootstrap by designating 400 out of 10^4 permutations generated by the procedure described above as "fake data sets". For each such fake data set, which represents one draw from the null hypothesis, we computed the number of nominally significant site pairs. This allowed us to estimate the full distribution of the number of false positives in the data. In particular, we recorded the expected number of false positives in our data, which is typically referred to as "false discovery rate" (FDR), and overall P-value for the total number of positives actually observed.

Simulations without epistasis

To ensure that our method does not detect epistatic interactions when there are none, we have performed detailed simulations of sequence evolution along a phylogenetic tree, with independent sites, as described in [25]. Briefly, the simulation algorithm takes as input a phylogenetic tree with branch lengths equal to the number of nucleotide substitutions and the nucleotide sequence at the root of the tree; it outputs the nucleotide sequences at all internal and terminal nodes. The sequence at a node is generated recursively, given that the sequence at the parental node is already known, using the following stochastic procedure. If the branch length connecting the focal node to the parental node is l, then l mutations are randomly distributed along the parental sequence proportionally to the entries in the 4×3 nucleotide mutation matrix and the codon-specific dN/dS values. We used the HA phylogenetic tree to perform this simulation as well as to infer the nucleotide mutation matrix and the codon-specific dN/dS values [25]. Using this simulation algorithm, we generated 100 independent sequence data sets. On each of them, we performed the analysis described above with 10^3 permutations, 100 of which were considered "fake data sets".

Supporting Information

Figure S1 Dependence of the results for HA (subtype H3N2) on the nominal P-value cutoff. Top panel shows the number of nominally significant pairs (solid line: data; dashed line: expectation). Middle panel shows the P-value for the observed number of significant pairs. Bottom panel shows the resulting FDR.
Found at: doi:10.1371/journal.pgen.1001301.s001 (0.02 MB EPS)

Figure S2 Dependence of the results for HA (subtype H3N2) on the scale parameter τ. Top panel shows the number of nominally significant pairs. Bottom panel shows the resulting FDR. Colors represent different nominal P-value cutoffs. The P-value for the observed number of significant pairs always stays below 0.05 and therefore not shown. The April 2009 version of the HA data set was used to generate this figure (see "Materials and Methods").
Found at: doi:10.1371/journal.pgen.1001301.s002 (0.17 MB EPS)

Table S1 Lists of putatively epistatic pairs identified by our analysis.
Found at: doi:10.1371/journal.pgen.1001301.s003 (0.21 MB XLSX)

Table S2 Posterior probabilities of co-evolution inferred by the BGM analysis (all branches). See Text S1 for details.
Found at: doi:10.1371/journal.pgen.1001301.s004 (5.92 MB XLSX)

Table S3 Posterior probabilities of co-evolution inferred by the BGM analysis (only interior branches). See Text S1 for details.
Found at: doi:10.1371/journal.pgen.1001301.s005 (1.36 MB XLSX)

Table S4 Summary of epistasis analyses for 10 bootstrap phylogenetic trees. "tree" is the bootstrap tree identifier; τ is the expected number of synonymous substitutions between a pair of non-synonymous substitutions randomly drawn from the given bootstrap tree; "sites" is the number of sites that experienced at least 2 substitutions at the internal nodes of the tree; "pairs" is the corresponding total number of ordered site pairs; "obs" is the number of site pairs with the nominal P-value of 0.01 (in this

column asterisk indicates that this number is significant at 0.05 level, and double asterisk indicates significance at 0.01 level); "FDR" is the false discovery rate and "overlap" is the number of pairs that are significant at the 0.01 level with the given bootstrap tree that are also in the list of pairs found in our original analysis; percentages in the parentheses are computed with respect to the 333 nominally significant pairs found in our original analysis (see Table 1 in the main text and Table S1).
Found at: doi:10.1371/journal.pgen.1001301.s006 (0.01 MB PDF)

Text S1 Supplementary information.
Found at: doi:10.1371/journal.pgen.1001301.s007 (0.15 MB PDF)

Acknowledgments

The authors are grateful to Jesse Bloom and Scott Hensley for fruitful discussions of epistasis in influenza and to three anonymous reviewers whose comments substantially improved the manuscript.

Author Contributions

Conceived and designed the experiments: SK JD GAB JBP. Performed the experiments: SK. Analyzed the data: SK JBP. Contributed reagents/materials/analysis tools: JD JBP. Wrote the paper: SK JD GAB JBP.

References

1. Earn DJD, Dushoff J, Levin SA (2002) Ecology and evolution of the flu. Trends Ecol Evol 17: 334–340.
2. 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.
3. Nelson MI, Holmes EC (2007) The evolution of epidemic influenza. Nat Rev Genet 8: 196–205.
4. Taubenberger JK, Morens DM (2008) The pathology of influenza virus infections. Annu Rev Pathol: Mech Dis 3: 499–522.
5. Arias CF, Escalera-Zamudio M, Soto-Del Río MD, Cobián-Güemes AG, Isa P, et al. (2009) Molecular anatomy of 2009 influenza virus A (H1N1). Arch Med Res 40: 643–654.
6. Das K, Aramini JM, Ma LC, Krug RM, Arnold E (2010) Structures of influenza A proteins and insights into antiviral drug targets. Nat Struct Mol Biol 17: 530–538.
7. Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, et al. (2005) Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS Biol 3: e300. doi:10.1371/journal.pbio.0030300.
8. Suzuki Y (2006) Natural selection on the influenza virus genome. Mol Biol Evol 23: 1902–1911.
9. Carrat F, Flahault A (2007) Influenza vaccine: the challenge of antigenic drift. Vaccine 25: 6852–6862.
10. Duffy S, Shackelton LA, Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9: 267–276.
11. 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.
12. Bush RM, Fitch WM, Bender CA, Cox NJ (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol 16: 1457–1465.
13. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, et al. (2004) Mapping the antigenic and genetic evolution of influenza virus. Science 305: 371–376.
14. Simonsen L, Viboud C, Grenfell BT, Dushoff J, Jennings L, et al. (2007) The genesis and spread of reassortment human influenza A/H3N2 viruses conferring adamantane resistance. Mol Biol Evol 24: 1811–1820.
15. Moscona A (2009) Global transmission of oseltamivir-resistant influenza. New Engl J Med 360: 953–956.
16. Bloom JD, Gong LI, Baltimore D (2010) Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. Science 328: 1272–1275.
17. Wiley DC, Wilson IA, Skehel JJ (1981) Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature 289: 373–378.
18. Laver WG, Air GM, Webster RG, Markoff LJ (1982) Amino acid sequence changes in antigenic variants of type A influenza virus N2 neuraminidase. Virology 122: 450–460.
19. Air GM, Els MC, Brown LE, Laver WG, Webster RG (1985) Location of antigenic sites on the three-dimensional structure of the influenza N2 virus neuraminidase. Virology 145: 237–248.
20. Gulati U, Hwang CC, Venkatramani L, Gulati S, Stray SJ, et al. (2002) Antibody epitopes on the neuraminidase of a recent H3N2 influenza virus (A/Memphis/31/98). J Virol 76: 12274–12280.
21. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) Predicting the evolution of human influenza A. Science 286: 1921–1925.
22. 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.
23. Blackburne BP, Hay AJ, Goldstein RA (2008) Changing selective pressure during antigenic changes in human influenza H3. PLoS Pathog 4: e1000058. doi:10.1371/journal.ppat.1000058.
24. Kosakovsky Pond SL, Poon AF, Brown AJL, Frost SDW (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.
25. Kryazhimskiy S, Bazykin GA, Plotkin JB, Dushoff J (2008) Directionality in the evolution of influenza A haemagglutinin. Proc R Soc B 275: 2455–2464.
26. Wilson IA, Cox NJ (1990) Structural basis of immune recognition of influenza virus hemagglutinin. Annu Rev Immunol 8: 737–771.
27. Guo HH, Choe J, Loeb LA (2004) Protein tolerance to random amino acid change. Proc Natl Acad Sci USA 101: 9205–9210.
28. Bloom JD, Arnold FH (2009) In the light of directed evolution: Pathways of adaptive protein evolution. Proc Natl Acad Sci USA 106: 9995–10000.
29. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10: 866–876.
30. Remold SK, Lenski RE (2004) Pervasive joint influence of epistasis and plasticity on mutational effects in *Escherichia coli*. Nat Genet 36: 423–426.
31. Sanjuán R, Moya A, Elena SF (2004) The contribution of epistasis to the architecture of fitness in an RNA virus. Proc Natl Acad Sci USA 101: 15376–15379.
32. Rimmelzwaan GF, Berkhoff EGM, Nieuwkoop NJ, Smith DJ, Fouchier RAM, et al. (2005) Full restoration of viral fitness by multiple compensatory co-mutations in the nucleoprotein of influenza A virus cytotoxic T-lymphocyte escape mutants. J Gen Virol 86: 1801–1805.
33. DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 6: 678–687.
34. Sanjuán R, Cuevas JM, Moya A, Elena SF (2005) Epistasis and the adaptability of an RNA virus. Genetics 170: 1001–1008.
35. Mateo R, Mateu MG (2007) Deterministic, compensatory mutational events in the capsid of foot-and-mouth disease virus in response to the introduction of mutations found in viruses from persistent infections. J Virol 81: 1879–1887.
36. de Visser JAGM, Elena SF (2007) The evolution of sex: empirical insights into the roles of epistasis and drift. Nat Rev Genet 8: 139–149.
37. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, et al. (2004) Global mapping of the yeast genetic interaction network. Science 303: 808–813.
38. de Visser JAGM, Park SC, Krug J (2009) Exploring the effect of sex on empirical fitness landscapes. Am Nat 174: S15–S30.
39. Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA (2010) Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. Nature 464: 279–282.
40. Blount ZD, Borland CZ, Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. Proc Natl Acad Sci USA 105: 7899–7906.
41. Kryazhimskiy S, Tkačik G, Plotkin JB (2009) The dynamics of adaptation on correlated fitness landscapes. Proc Natl Acad Sci 106: 18638–18643.
42. Shapiro B, Rambaut A, Pybus OG, Holmes EC (2006) A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. Mol Biol Evol 23: 1724–1730.
43. Weinreich DM, Delaney NF, DePristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312: 111–114.
44. Lozovsky ER, Chookajorn T, Brown KM, Imwong M, Shaw PJ, et al. (2009) Stepwise acquisition of pyrimethamine resistance in the malaria parasite. Proc Natl Acad Sci USA 106: 12025–12030.
45. Haq O, Levy RM, Morozov AV, Andrec M (2009) Pairwise and higher-order correlations among drug-resistance mutations in HIV-1 subtype B protease. BMC Bioinform 10: S10.
46. Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG, et al. (2009) Positive epistasis drives the acquisition of multidrug resistance. PLoS Genet 5: e1000578. doi:10.1371/journal.pgen.1000578.
47. Codoñer FM, Fares MA (2008) Why should we care about molecular coevolution? Evol Bioinform 4: 29–38.
48. Korber BTM, Farber RM, Wolpert DH, Lapedes AS (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. Proc Natl Acad Sci USA 90: 7176–7180.
49. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. Mol Biol Evol 17: 164–178.
50. Gloor GB, Martin LC, Wahl LM, Dunn SD (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 44: 7156–7165.
51. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286: 295–299.

52. Wollenberg KR, Atchley WR (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci USA 97: 3288–3291.
53. Fares MA, Travers SAA (2006) A novel method for detecting intramolecular coevolution: Adding a further dimension to selective constraints analyses. Genetics 173: 9–23.
54. Wang Q, Lee C (2007) Distinguishing functional amino acid covariation from background linkage disequilibrium in HIV protease and reverse transcriptase. PLoS ONE 2: e814. doi:10.1371/journal.pone.0000814.
55. Caporaso JG, Smit S, Easton BC, Hunter L, Huttley GA, et al. (2008) Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics. BMC Evol Biol 8: 327.
56. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24: 333–340.
57. Buslje CM, Santos J, Delfino JM, Nielsen M (2009) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics 25: 1125–1131.
58. Govindarajan S, Ness JE, Kim S, Mundorff EC, Minshull J, et al. (2003) Systematic variation of amino acid substitutions for stringent assessment of pairwise covariation. J Mol Biol 328: 1061–1069.
59. Pollock DD, Taylor WR, Goldman N (1999) Coevolving protein residues: Maximum likelihood identification and relationship to structure. J Mol Biol 287: 187–198.
60. Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005) A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 22: 1919–1928.
61. Fukami-Kobayashi K, Schreiber DR, Benner SA (2002) Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. J Mol Biol 319: 729–743.
62. Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R (2005) Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics 21: i126–i135.
63. Dutheil J, Galtier N (2007) Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evol Biol 7: 242.
64. Poon AFY, Lewis FI, Kosakovsky Pond SL, Frost SDW (2007) An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. PLoS Comput Biol 3: e231. doi:10.1371/journal.pcbi.0030231.
65. Baussand J, Carbone A (2009) A combinatorial approach to detect coevolved amino acid networks in protein families of variable divergence. PLoS Comput Bio 5: e1000488. doi:10.1371/journal.pcbi.1000488.
66. Poon AFY, Swenson LC, Dong WWY, Deng W, Kosakovsky Pond SL, et al. (2010) Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the *nef* gene of HIV-1. Mol Biol Evol 27: 819–832.
67. Bazykin GA, Dushoff J, Levin SA, Kondrashov AS (2006) Bursts of nonsynonymous substitutions in HIV-1 evolution reveal instances of positive selection at conservative protein sites. Proc Natl Acad Sci USA 103: 19396–19401.
68. Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496–503.
69. Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, et al. (2008) Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. Vaccine 26S: D31–34.
70. Aoki FY, Boivin G, Roberts N (2007) Influenza virus susceptibility and resistance to oseltamivir. Antiviral Therapy 12: 603–616.
71. Collins P, Haire L, Lin Y, Liu J, Russell R, et al. (2009) Structural basis for oseltamivir resistance of influenza viruses. Vaccine 27: 6317–6323.
72. Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676–679.
73. Fornasari MS, Parisi G, Echave J (2002) Site-specic amino acid replacement matrices from structurally constrained protein evolution simulations. Mol Biol Evol 19: 352–356.
74. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 10: 1692–1704.
75. Rodrigue N, Lartillot N, Bryant D, Philippe H (2005) Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347: 207–217.
76. Plotkin JB, Dushoff J, Levin SA (2002) Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc Natl Acad Sci USA 99: 6263–6268.
77. Levin SA, Dushoff J, Plotkin JB (2004) Evolution and persistance of influenza A and other viruses. Math Biosci 188: 17–28.
78. Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. PLoS Genet 4: e1000304. doi:10.1371/journal.pgen.1000304.
79. Shortle D, Lin B (1985) Genetic analysis of staphylococcal nuclease: identification of three intragenic "global" suppressors of nuclease-minus mutations. Genetics 110: 539–555.
80. Poteete AR, Rennell D, Bouvier SE, Hardy LW (1997) Alteration of T4 lysozyme structure by secondsite reversion of deleterious mutations. Prot Sci 6: 2418–2425.
81. Hensley SE, Das SR, Bailey AL, Schmidt LM, Hickman HD, et al. (2009) Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. Science 326: 734–736.
82. Wagner A (2008) Neutralism and selectionism: a network-based reconciliation. Nat Rev Genet 9: 965–974.
83. Draghi JA, Parsons TL, Wagner GP, Plotkin JB (2010) Mutational robustness can facilitate adaptation. Nature 463: 353–355.
84. Koelle K, Cobey S, Grenfell B, Pascual M (2006) Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. Science 314: 1898–1903.
85. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, et al. (2008) The influenza virus resource at the National Center for Biotechnology Information. J Virol 82: 596–601.
86. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, et al. (2003) Multiple sequence alignment with the Clustal series of programs. Nucl Acids Res 31: 3497–3500.
87. Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucl Acids Res 34: W609–W612.
88. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
89. Swofford DL (2003) Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, Massachusetts: Sinauer Associates.
90. Connor EF, Simberloff D (1979) The assembly of species communities: chance or competition? Ecology 60: 1132–1140.
91. Manley BFJ (1995) A note on the analysis of species co-occurrences. Ecology 76: 1109–1115.
92. Sundaresan SR, Fischhoff IR, Dushoff J (2009) Avoiding spurious findings of nonrandom social structure in association data. Anim Behav 77: 1381–1385.
93. Sorić B (1989) "Discoveries" and effect-size estimation. J Am Stat Assoc 84: 608–610.