Clonal Interference in the Evolution of Influenza

Natalja Strelkowa* and Michael Lässig†,1

*Department of Bioengineering, Imperial College London, South Kensington SW7 2AZ, United Kingdom and †Institute for Theoretical Physics, University of Cologne, 50937 Köln, Germany

ABSTRACT The seasonal influenza A virus undergoes rapid evolution to escape human immune response. Adaptive changes occur primarily in antigenic epitopes, the antibody-binding domains of the viral hemagglutinin. This process involves recurrent selective sweeps, in which clusters of simultaneous nucleotide fixations in the hemagglutinin coding sequence are observed about every 4 years. Here, we show that influenza A (H3N2) evolves by strong clonal interference. This mode of evolution is a red queen race between viral strains with different beneficial mutations. Clonal interference explains and quantifies the observed sweep pattern: we find an average of at least one strongly beneficial amino acid substitution per year, and a given selective sweep has three to four driving mutations on average. The inference of selection and clonal interference is based on frequency time series of single-nucleotide polymorphisms, which are obtained from a sample of influenza genome sequences over 39 years. Our results imply that mode and speed of influenza evolution are governed not only by positive selection within, but also by background selection outside antigenic epitopes: immune adaptation and conservation of other viral functions interfere with each other. Hence, adapting viral proteins are predicted to be particularly brittle. We conclude that a quantitative understanding of influenza's evolutionary and epidemiological dynamics must be based on all genomic domains and functions coupled by clonal interference.

INFLUENZA is one of the major infectious diseases in humans. Seasonal strains of the influenza A (H3N2) virus circulating in the human population account for about half a million deaths per year. Due to its impact on health, influenza has become a uniquely well-documented system of molecular evolution. The viral genome contains eight segments, one of which encodes the surface protein hemagglutinin (HA). The HA1 domain of this protein contains antigenic epitopes, which are the primary loci of interaction with the human immune system (Wiley *et al.* 1981). Its gene sequence is now available for several thousand strains (Bao *et al.* 2008) and is used to construct strain trees spanning several decades of influenza evolution (Bush *et al.* 1999).

A striking and extensively studied feature of this process is its punctuated pattern, which is particularly visible in antigen–antibody binding data: periods of relative stasis (called antigenic clusters) are separated by cluster transitions, which occur every few years and produce most of the antigenic adaptation (Smith *et al.* 2004). Clustering has also been observed in the temporal distribution of amino acid fixations (Plotkin *et al.* 2002; Wolf *et al.* 2006; Shih *et al.* 2007) and in simulation studies of epidemiological models (Ferguson *et al.* 2003; Gog *et al.* 2003; Tria *et al.* 2005; Koelle *et al.* 2006; Strelkowa, 2006; Minayev and Ferguson 2009). However, the evolutionary cause of this pattern remains controversial (Holmes and Grenfell 2009). Clustering has been described by a model of episodic evolution, in which antigenic clusters correspond to periods of neutral evolution and positive selection is restricted to cluster transitions (Koelle *et al.* 2006; Wolf *et al.* 2006; Koelle *et al.* 2010). Other recent studies argue that most amino acid substitutions in the viral epitopes are under positive selection (Shih *et al.* 2007) and clusters of their fixations are caused primarily by fitness interactions (epistasis) between epitope sites (Shih *et al.* 2007; Kryazhimskiy *et al.* 2011).

In this article, we focus on genomic determinants of influenza's evolutionary process. As we show below, there is no significant recombination between mutations *within* the HA1 domain. Hence, epitope and non-epitope sites of the HA1 sequence evolve in a genuinely asexual way, that is, under almost complete genetic linkage. At the same time, the population dynamics of influenza involves reassortment *between* genomic segments, which is known to be a major

doi: 10.1534/genetics.112.143396
Manuscript received April 23, 2012; accepted for publication July 17, 2012
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.143396/-/DC1.
[1]Corresponding author: University of Cologne, Institute of Theoretical Physics, University of Cologne, Zülpicher Str. 77, 50937 Köln, Germany. E-mail: lassig@thp.uni-koeln.de

factor of its epidemiology (Holmes *et al.* 2005; Rambaut *et al.* 2008). The key point of this article is to show that rapid adaptation under linkage produces *clonal interference*, a specific mode of evolution by recurrent selective sweeps, within the hemagglutinin gene. Clonal interference is characteristic of asexual organisms evolving at high mutation rates and is well documented by evolution experiments with bacterial and viral laboratory populations (Gerrish and Lenski 1998; de Visser *et al.* 1999; Miralles *et al.* 1999; Perfeito *et al.* 2007; Miller *et al.* 2011). Here, we use genome analysis to obtain the first evidence of clonal interference in a wild system. This mode of evolution produces temporal clustering of fixations as a generic consequence of genetic linkage and high supply of beneficial mutations, which does not depend on details of the fitness landscape (Park and Krug 2007). Thus, it provides a new, parsimonious explanation for influenza's punctuated genome evolution.

Theoretical studies of asexual evolution show that clonal interference emerges whenever there is a sufficiently high supply of beneficial mutations to trigger competition between mutant clones (Gerrish and Lenski 1998; Wilke 2004; Schiffels *et al.* 2011). A clone is a set of strains with similar sequences and a recent common ancestor, which is distinguished from its background by the new mutations that appear in its ancestor sequence. For any set of competing clones, only lineages descending from a single high-fitness clone will survive, while all other clones will eventually become extinct. The expansion of successful clones is driven by strongly beneficial mutations, which fix in the population rapidly. We call these events selective sweeps. Neutral and moderately deleterious changes are frequently carried to fixation by hitchhiking, that is, as passenger mutations within sweeps. At the same time, sweeps drive other moderately beneficial mutations to loss, if they are harbored in outcompeted clones. Note that we have defined the terms "clone" and "sweep" in a broad way, which is adequate for the high mutation rate of influenza. Clones consist of strains that are genetically similar but often not identical. While successful clones expand in the population, subsequent mutations continue to produce sequence and fitness variation; that is, new clones originate nested within previous clones (Park and Krug 2007; Desai and Fisher 2007). As it has become clear from recent studies, the competition between beneficial mutations in disjoint clones and their mutual reinforcement in nested clones are two sides of the same dynamics: interference interactions can be positive or negative (Schiffels *et al.* 2011; Good *et al.* 2012; Lässig 2012). In other words, the recurrent selective sweeps in the clonal interference mode reduce but do not remove diversity, and the population always remains multiclonal. Thus, clonal interference differs from a regime of *episodic selective sweeps*, which has been suggested as a model for influenza (Koelle *et al.* 2006). Episodic sweeps occur if there is a low supply of strongly beneficial mutations, that is, for sufficiently low mutation rates or small populations (Gillespie 1991, 1993). Every such sweep removes all fitness variation from the population, and there are extended periods of neutral evolution between consecutive sweeps. This mode of evolution is also referred to as *periodic selection* (Atwood *et al.* 1951) (which is somewhat misleading, because no time dependence of selection is implied). A given system can cross over from periodic selection to clonal interference if its population size, its mutation rate, or the time dependence of selection is increased; the dependence on population size has been observed in recent evolution experiments (Perfeito *et al.* 2007; Miller *et al.* 2011). Figure 1 contrasts the two modes of evolution by simulations of influenza-like strain trees (the model used for the simulations is explained in detail below). Our genomic analysis provides evidence that the actual evolutionary process of influenza is governed by clonal interference and, thus, generates more beneficial mutations than episodic sweeps.

Our analysis proceeds in several steps. First, we show that the HA1 domain of influenza evolves under almost complete genetic linkage, which can be quantified by allele frequency correlations between polymorphic sequence sites. Second, we provide a quantitative analysis of influenza's recurrent selective sweeps. This pattern manifests itself in a number of characteristics: nucleotides fix in temporal clusters, dips in sequence diversity are correlated with these clusters, and lifetimes to fixation follow a similar distribution for different classes of polymorphisms. The sweep pattern is consistent with previous results on punctuated antigenic evolution (Smith *et al.* 2004) and on clustering of amino acid fixations (Plotkin *et al.* 2002; Wolf *et al.* 2006; Shih *et al.* 2007). Our results for neutral polymorphisms, in particular, show that hitchhiking effects are strong, in contrast to a previous analysis based on only nonsynonymous polymorphisms (Shih *et al.* 2007). In the third and central part of the article, we provide evidence that influenza evolves under a sufficiently high supply of beneficial mutations to trigger clonal interference: on average, more than one beneficial mutation that has overcome genetic drift is present in the population. We use a new *frequency propagator* method to infer selection from polymorphism time series, which is applicable to recurrent selective sweeps in linked genomes. Our analysis shows that the influenza strain tree emerges from a particular coalescent process under positive selection. The inference of clonal interference by the propagator method is corroborated by a minimal model for influenza genome evolution, which reproduces the characteristics of polymorphism time series observed for influenza. In the final part, we use this evolutionary model to derive consequences of clonal interference for biological functions of the influenza A virus, and we discuss how this mode of evolution may arise from the underlying host–pathogen immune interactions.

Results

Evolution under genetic linkage

Our study is based on a sequence sample of 1971 influenza A (H3N2) strains occurring between 1969 and 2007 (Bao *et al.* 2008). We build an ensemble of equiprobable strain trees from these sequences by maximum parsimony; a typical tree is shown in Supporting Information, Figure S1. Each

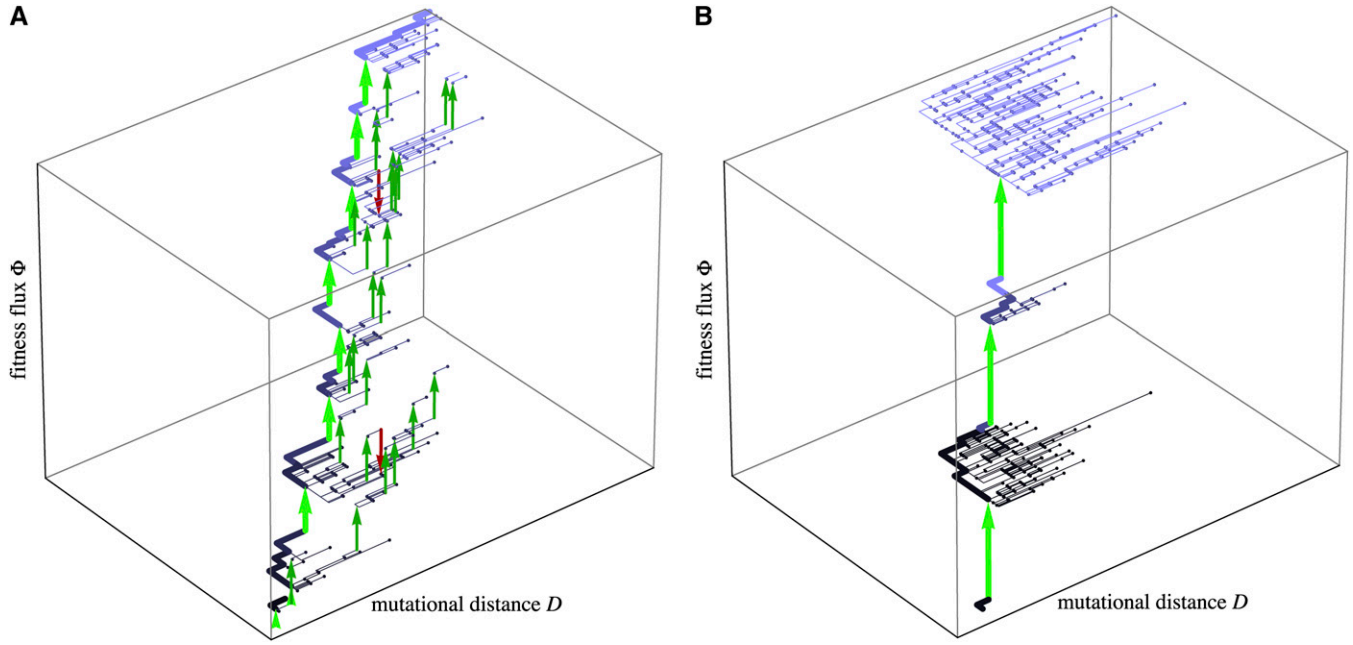

Figure 1 Modes of evolution under linkage: Clonal interference *vs.* episodic selective sweeps. The figure shows strain trees of the influenza evolution model (see text). Nodes of the tree represent strains with distinct HA sequences. Mutations are mapped on individual branches of the tree, all fixed changes appear on the trunk of the tree (thick line). For each node, the horizontal coordinate D counts the number of mutations from the root to its strain sequence, and the vertical coordinate Φ is the sum of their selection coefficients (the so-called cumulative fitness flux (Mustonen and Lässig 2007, 2009, 2010)). Upward (green) and downward (red) arrows indicate individual branches with positive and negative fitness flux, respectively. (A) Clonal interference. In this mode, high supply of beneficial mutations generates competition between coexisting clones: many beneficial changes reach substantial frequencies, but only a fraction of them are fixed (thick green arrows on the trunk), while others are eventually outcompeted (thin green arrows off the trunk). Neutral evolution (represented by planar subtrees) occurs for limited periods within subpopulations. (B) Episodic sweeps. In this mode, low supply of beneficial mutations generates selective sweeps interspersed with extended periods of neutral evolution. Interference interactions are negligible; *i.e.*, all beneficial mutations reaching substantial frequencies are fixed (all green arrows are on the trunk). We show that the evolution of influenza A (H3N2) is governed by clonal interference and not by episodic sweeps; see text and Figure 4.

unique HA1 sequence observed in a given year is represented by an external node; unobserved strains are represented by internal nodes. The trees predict sequence and year of internal nodes and map point mutations between directly related strains onto specific branches (see Figure S1). A sequence clone is uniquely associated with a subtree and is distinguished from its background by the mutations that are mapped onto the branch to its common ancestor. Here, we use strain trees to estimate yearly population frequencies of clones and of the mutations they carry. Details of strain selection, tree construction, and strain frequency estimation are given in File S1; a list of the strains used in this study appears in File S2.

As a first step of the analysis, we evaluate the amount of genetic association (linkage disequilibrium) within the HA1 domain. Although strong association is to be expected for an asexually reproducing virus, its actual degree requires analysis. This is because the high mutation rate of influenza sometimes causes the same mutation to originate independently in coexisting clones (Shih *et al.* 2007; Kryazhimskiy *et al.* 2008), an effect reducing genetic association even in the absence of recombination. We evaluate HA1 haplotypes containing mutant alleles at pairs of simultaneously polymorphic sequence sites. For a given pair, we compare the double-mutant haplotype frequency x_{12} with the (marginal) frequencies x_1 and x_2 of the single-nucleotide mutant alleles at site 1 and 2 (for details, see File S1). Figure 2 and Figure S3 show scaled haplotype frequencies $y_{12} \equiv x_{12}/\min(x_1, x_2)$ for pairs of simultaneous polymorphisms in different mutation classes of the HA1 domain. In the vast majority of cases, we find values $y_{12} = 0$ or $y_{12} = 1$, which is indicative of complete genetic association of the mutant alleles. We measure the degree of association for a given pair of mutations by the allele frequency correlation

$$\mathcal{C}(x_{12}, x_1, x_2) = \frac{\mathcal{D}(x_{12}, x_1, x_2)}{\mathcal{D}(\xi_{12}, x_1, x_2)}, \qquad (1)$$

where $\mathcal{D}(x_{12}, x_1, x_2) \equiv x_{12} - x_1 x_2$ is the linkage disequilibrium and ξ_{12} is the maximal or minimal double-mutant haplotype frequency consistent with given allele frequencies; *i.e.*, $\xi_{12} = \min(x_1, x_2)$ if $x_{12} \geq x_1 x_2$ and $\xi_{12} = 0$ if $x_{12} < x_1 x_2$. The minimum value $\mathcal{C} = 0$ indicates statistical independence of the two polymorphisms (*i.e.*, linkage equilibrium, $x_{12} = x_1 x_2$), the maximum $\mathcal{C} = 1$ complete genetic association between the mutant alleles (*i.e.*, $x_{12} = \min(x_1, x_2)$ or $x_{12} = 0$). Compared to the familiar $\mathcal{D}'$ (Lewontin 1964), the allele frequency correlation $\mathcal{C}$ is a stricter measure of association, which is appropriate for the analysis of haplotypes on influenza strain trees (for details, see File S1). For pairs of HA1 sequence

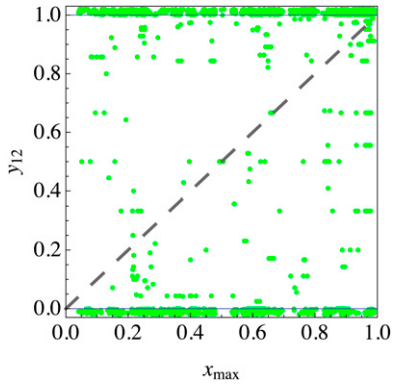

Figure 2 Genetic linkage in the influenza HA1 domain. For pairs of mutations with haplotype frequency x_{12} and marginal (allele) frequencies x_1 and x_2, the scaled haplotype frequency $y_{12} = x_{12}/\min(x_1, x_2)$ is plotted against the larger allele frequency, $x_{max} = \max(x_1, x_2)$. Yearly frequency data are shown for 934 pairs of nonsynonymous epitope polymorphisms (1969 green points), which have an average frequency correlation $\bar{C} = 0.948$. Most points show maximum linkage disequilibrium characteristic of complete genetic linkage; *i.e.*, $y = 1$ for polymorphisms in nested clones and $y = 0$ for polymorphisms in disjoint clones (these points are shown with random y values in the interval (1, 1.02) and (−0.02, 0), respectively, to make a larger number of points visible). Some mutations originate in multiple clones and break complete linkage, as shown by values $0 < y_{12} < 1$. However, the overall pattern is far from linkage equilibrium ($y_{12} = x_{max}$, dashed line). Analogous data for other polymorphism classes are shown in Figure S3.

polymorphisms, we find an average frequency correlation $\bar{C} = 0.96$ (for details, see Figure 2 and Figure S3). Furthermore, we do not find any dependence of $\bar{C}$ on the distance between sequence sites, which indicates that the small deviations from complete association are generated by independent originations of the same point mutation in competing clones and not by recombination of alleles between sites. Such multiple originations can be observed for some alleles (Figure 2 and Figure S3), but their effect is too weak to reduce frequency correlations significantly. This result implies that selection acts on genotypes and not on individual mutations, which is a prerequisite for clonal interference (Neher and Shraiman 2009).

Recurrent selective sweeps

To understand how linkage affects the evolution of influenza hemagglutinin, we record the histories of all single-nucleotide polymorphisms in the sequence sample, starting with the entry of a new allele into the population and ending with fixation or loss of that allele. We classify these polymorphisms according to their sequence position: nonsynonymous epitope changes, nonsynonymous changes outside the epitopes, and synonymous changes. This rather broad classification of genomic changes reflects the aim of this study, which is to derive influenza's mode of evolution and its selective cause, but not the role of individual codons or amino acid changes in this process. Without too much biophysical *a priori* information, our analysis will produce a quantitative inference of heterogeneous selection in the HA1 domain: nonsynonymous epitope changes are predominantly under positive selection, nonsynonymous changes outside the epitopes are predominantly under negative selection, and synonymous changes evolve near neutrality. This inference is in accordance with a number of previous studies (Bush *et al.* 1999; Plotkin *et al.* 2002; Wolf *et al.* 2006; Bhatt *et al.* 2011) and will be detailed in the next section. We first discuss the genomic evidence that influenza evolves by recurrent selective sweeps, a pattern that is consistent with clonal interference:

Nucleotides fix in temporal clusters: Figure 3A shows the fixation year for all 160 fixed HA1 polymorphisms. The resulting distribution of the number of yearly fixations deviates strongly from the form expected for independently evolving sites, *i.e.*, a Poisson distribution with the same mean value of 4.1 substitutions per year (dashed line). This deviation defines the amount of clustering, that is, the accumulation of fixation events in some years and the corresponding depletion in others. It can be measured by the ratio of variance and mean of yearly fixation numbers; we find a ratio of 6.7 in the data, which is much larger than the range 1 ± 0.25 for a finite sample of Poisson-distributed values. Defining a fixation cluster as a period with at least eight nucleotide fixations per year in the HA1 domain, we obtain a total of eight major clusters, which are marked by dashed lines in Figure 3A. This definition of fixation clusters is clearly not unique, but our conclusions are robust under variations of the threshold number of fixations. The fixation clusters cover 24% of the time span, but contain >64% of fixations in each of the three polymorphism classes. In particular, the clustering of (near-neutral) synonymous changes signals pervasive hitchhiking in selective sweeps. Nonsynonymous non-epitope substitutions still occur preferentially in these clusters, although their number is reduced by negative selection (see below). The observed clustering of epitope amino acid fixations is not stronger than that of neutral changes. Hence, it can also be explained by hitchhiking, but intra-epitope fitness interactions are likely to contribute to this effect (Shih *et al.* 2007; Kryazhimskiy *et al.* 2011).

Dips in sequence diversity correlate with fixation clusters: Figure 3B shows the yearly diversity in all three polymorphism classes. There are recurrent dips in sequence diversity, which occur close in time to the fixation clusters (dashed lines). These dips have also been associated with antigenic cluster transitions (Smith *et al.* 2004). A dip occurs when beneficial alleles in a successful clone remove the ancestral sequence diversity in competing clones; the subsequent rebound of diversity is caused by new mutations within the successful clone. The minimum diversity is observed when a sweep has driven most competing ancestor strains to low frequency. This occurs sometimes 1 year before the fixation cluster, which marks the extinction of all but one of these clones. A closer look at the strain tree reveals a complex sweep dynamics. A single, putatively beneficial epitope allele is often observed to originate and rise to intermediate

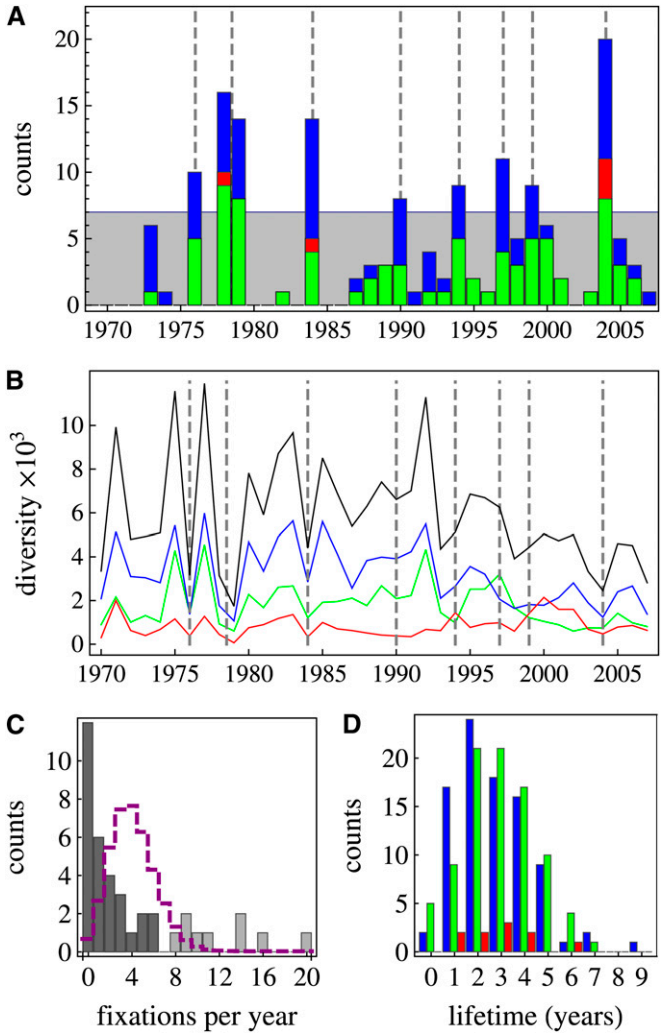

Figure 3 Influenza evolves by recurrent selective sweeps. The histories of 160 fixed polymorphisms in the influenza HA1 domain signal recurrent selective sweeps consistent with clonal interference: (A) Histogram of fixation years between 1969 and 2007 in three polymorphism classes (blue, synonymous; red, nonsynonymous non-epitope; green, nonsynonymous epitope). About 70% of all fixations occur in eight major fixation clusters containing eight or more mutations (columns reaching above shaded area, dashed lines). (B) Sequence diversity *vs.* year of occurrence, contributions of the three polymorphism classes (blue, red, and green line), and total divergence (black line). Dips in diversity are correlated with major fixation clusters. Diversity is measured by the expected number of pairwise nucleotide differences per unit sequence length between strains of the same year. (C) Histogram of the number of yearly nucleotide fixation events (bars); major fixation clusters are highlighted (light bars). The data distribution deviates strongly from a Poisson distribution with the same mean value of 4.1 substitutions per year (dashed line). (D) Histogram of polymorphism lifetimes between entry and fixation of the new allele. The corresponding normalized distributions are similar in all three mutation classes, with average lifetimes between 2.9 years and 3.1 years.

frequencies within two or more disjoint contemporary clones, a pattern similar to so-called soft selective sweeps (Pennings and Hermisson 2006). When its frequency reaches one, however, all but one of these clones have been lost. That is, the fixation of HA1 alleles always occurs within a single successful clone: ultimately, all sweeps in the influenza HA1 domain are hard. This dynamics reflects the high rate of beneficial and deleterious mutations. Clones harbor multiple selected alleles, which leads to fitness differences even between clones sharing a given beneficial allele.

Polymorphism lifetimes are similar: Figure 3D shows the distributions of lifetimes for fixed polymorphisms in all three classes. The average times are similar, which is consistent with recurrent selective sweeps. In this mode, fixation times are determined by the total selection on sweeping clones rather than by selection coefficients of individual nucleotide changes. For unlinked sites, nonsynonymous mutations would have much shorter lifetimes to fixation than synonymous changes, given their substantial level of (positive or negative) selection inferred below.

Inference of clonal interference

How many beneficial mutations drive these selective sweeps? Is their supply sufficient to generate, on average, two or more coexisting beneficial alleles that have overcome genetic drift and compete by clonal interference? We now answer this question by a more detailed analysis of polymorphism histories, which produces quantitative estimates of selection acting on influenza. The analysis has to address an important caveat: genetic linkage and interference interactions themselves confound the inference of selection, because correlations between polymorphism histories reduce the statistical differences between sites evolving under selection and neutral sites. This caveat applies to all standard population-genetic selection tests based on polymorphism frequency distributions, as well as to methods based on time-series analysis for independent loci (Nielsen 2005). Here, we infer selection by a new method, which is not confounded by clonal interference and is robust to sampling biases in our data set. We define the *frequency propagator* $G(x)$ as the likelihood that a new allele appearing in our sequence sample reaches a frequency $>x$ at some later point. We evaluate the ratio

$$g(x) = \frac{G(x)}{G_0(x)} \tag{2}$$

of the propagator $G(x)$ for nonsynonymous mutations (either within or outside the epitopes) and its counterpart $G_0(x)$ for synonymous changes. In a similar way, we analyze polymorphisms whose new allele reaches frequencies exceeding a given threshold x at some intermediate point of its lifetime but is eventually lost. The likelihood of this process is given by the *loss propagator* $H(x)$, and we define the propagator ratio

$$h(x) = \frac{H(x)}{H_0(x)} \tag{3}$$

for each class of nonsynonymous mutations with respect to the synonymous reference class.

In the limit $x = 1$, the propagator ratio $g(x)$ reduces to the ratio of fixation probabilities $g = (d/n)/(d_0/n_0)$, where d, d_0

are the numbers of fixed polymorphisms and n, n_0 are the total numbers of polymorphisms in the two mutation classes. Hence, g is a history-based measure of selection that is conceptually and computationally related to the McDonald–Kreitman test of selection (McDonald and Kreitman 1991). At the same time, the propagator method fundamentally differs from the popular D_n/D_s test (Li *et al.* 1985), which has been used in previous influenza studies (Bush *et al.* 1999; Wolf *et al.* 2006). The D_n/D_s test is often used on phylogenetic trees across species, where it counts nonsynonymous and synonymous *substitutions* on individual branches. The influenza tree, however, is a genealogical tree, which describes the coalescence process between strains under selection. If applied to a genealogical tree, the D_n/D_s test counts *originations* of nonsynonymous and synonymous polymorphisms, which provide only a dilute signal of selection. The propagator method, however, is based on entire polymorphism histories, which include frequency and time information. This is related to an intuitive picture of the method, which we discuss below: propagators not only count mutations, but evaluate their position on the strain tree.

Propagator ratios are insensitive to uncertainties in entry frequency and timing of polymorphism histories, as well as to frequency-dependent bias in polymorphism numbers, as long as this bias does not depend on mutation class (see File S1). Most importantly, the propagator method measures selection in a way not confounded by clonal interference. In particular, a propagator ratio $g < 1$ signals evolutionary constraint, from which we infer that at least a fraction $(1 - g)$ of the nonsynonymous changes are under negative selection. Similarly, a propagator ratio $g > 1$ signals an increase in substitution probability of nonsynonymous over synonymous mutations, from which we infer that at least a fraction $(g - 1)/g$ of the nonsynonymous changes are beneficial (McDonald and Kreitman 1991; Smith and Eyre-Walker 2002). These standard estimates of constraint and adaptation become more stringent under conditions of genetic linkage. Clonal interference implies the existence of a characteristic selection strength $\tilde{\sigma}$, such that mutations with selection coefficient $\sigma > \tilde{\sigma}$ are mostly driving mutations (*i.e.*, independent of interference) and mutations with $\sigma < \tilde{\sigma}$ are mostly passenger mutations (*i.e.*, subject to interference). Moderately beneficial or deleterious passenger mutations (with selection coefficients $-\tilde{\sigma} < \sigma < \tilde{\sigma}$) are reduced to near-neutral fixation probabilities, and only strongly deleterious mutations (with $\sigma < \tilde{\sigma}$) are under significant evolutionary constraint (Schiffels *et al.* 2011). Hence, the above tests infer the fraction of strongly beneficial driving mutations and the fraction of strongly deleterious passenger mutations, respectively.

Applying the propagator method to the polymorphism time series of the influenza HA1 data set, we obtain estimates of heterogeneous selection. For nonsynonymous non-epitope mutations, we find a strongly reduced fixation probability ($g = 0.3 \pm 0.15$); see Figure 4A. Thus, amino acid changes outside the epitopes evolve under substantial evolutionary constraint, indicating that at least 70% of these changes are under negative selection strong enough to suppress passenger substitutions. For epitope sites, the propagator method produces strong evidence of clonal interference:

Multiple beneficial mutations occur simultaneously: Nonsynonymous epitope mutations have a substantially increased fixation probability ($g = 2.3 \pm 0.6$); see Figure 4A. Hence, of the 80 epitope amino acid substitutions, only a fraction $80/g = 35$ would be expected under neutrality, and at least $80(g - 1)/g = 45$ are strongly beneficial mutations driving adaptation. A similar number of beneficial substitutions has recently been estimated by Bhatt *et al.* (2011). Given a mean lifetime to fixation of 2.9 years as shown in Figure 3C, we conclude that the population contains at least three simultaneous adaptive mutations on average. This supply is too high for sequential fixation by episodic sweeps, but is consistent with clonal interference. Temporally overlapping beneficial mutations reinforce each other if they occur in nested clones, and they compete with each other if they occur in disjoint clones (Schiffels *et al.* 2011; Good *et al.* 2012). Clonal interference implies that beneficial mutations are always present in the population and not just in cluster years, as assumed in the scenario of episodic sweeps (Koelle *et al.* 2006; Wolf *et al.* 2006; Koelle *et al.* 2010). It is their fixation events that are clustered: assuming that every epitope change is equally likely to be adaptive, we infer at least 29 driving mutations among the 51 epitope changes contained in the 8 major sweep fixation clusters (Figure 3A, dashed lines). Hence, a given sweep involves an average of 3.6 driving mutations.

For nonsynonymous epitope mutations at intermediate frequencies, we infer an even higher number of beneficial changes. The above estimate for the fraction of beneficial changes is not limited to substitutions ($x = 1$), but can be applied also at intermediate frequencies x. Given 118 observed epitope amino acid changes at frequency $x = 0.7$ and a propagator ratio $g(0.7) = 2.15$, we infer that at least $118[g(0.7) - 1]/g(0.7) = 63$ of these changes are beneficial. Importantly, this number is higher than the 45 beneficial epitope substitutions. This implies that at least 18 beneficial epitope changes are lost after they have reached frequencies $x > 0.7$, providing evidence of clonal interference. We note that all of these mutations have overcome genetic drift and would fix deterministically in the absence of clonal interference, because the inferred level of selection is strong compared to genetic drift. From our model-based analysis described below, we estimate products of selection coefficient and effective population size of order 100, so that only mutations with frequencies $x < 0.01$ are dominated by genetic drift.

Beneficial mutations are outcompeted: The frequency dependence of loss propagators provides the most direct evidence of clonal interference. The loss propagator ratio for epitope sites shown in Figure 4B takes values $h(x) > 1$ for intermediate frequencies, signaling *positive* selection acting on *lost* mutations. The mutations in this class reach intermediate

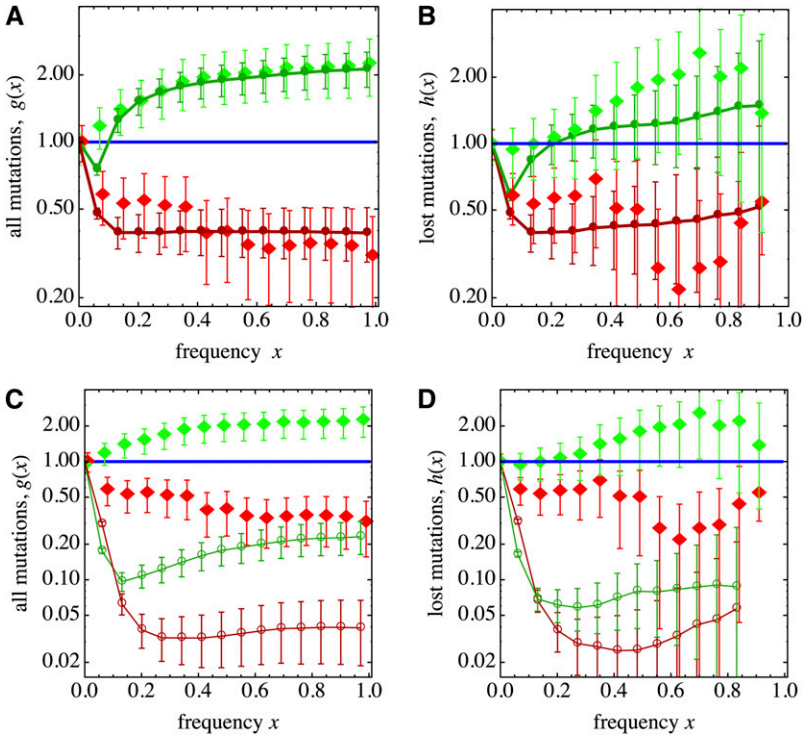

Figure 4 Inference of selection and clonal interference from polymorphism time series. The frequency propagator statistics $g(x)$ and $h(x)$, as defined by Equations 2 and 3, are evaluated for influenza HA1 and compared to simulated ratios for the minimal sequence evolution model. (A) Influenza frequency propagator ratio $g(x)$ for nonsynonymous non-epitope and epitope mutations (red and green diamonds, error bars are given by sampling fluctuations) with respect to the baseline of synonymous changes (blue line). These data are plotted together with simulations of $g(x)$ for the minimal model in the clonal interference mode (red and green circles); *cf.* Figure 1A. In the influenza data, the epitope frequency propagator ratio takes values $g(x) > 2$ for $x > 0.6$, signaling predominantly positive selection. For non-epitope sites, $g(x) < 1$ indicates predominantly negative selection. Both features of the influenza data are reproduced by the model results. (B) Influenza loss propagator ratio $h(x)$ for nonsynonymous non-epitope and epitope mutations (red and green diamonds), plotted together with simulations of $h(x)$ for the minimal model in the clonal interference mode (red and green circles). The epitope loss propagator ratio takes values $h(x) > 1$ for $x > 0.3$, signaling positive selection acting on mutations harbored in outcompeted clones. This is again reproduced by the model results. (C) Simulations of $g(x)$ in the mode of episodic sweeps (red and green open circles); *cf.* Figure 1B. The form of $g(x)$ does not match the influenza data (diamonds, same as in A). In the model dynamics, $g(x) < 1$ for epitope mutations signals a low rate of adaptation. (D) Simulations of $h(x)$ in the mode of episodic sweeps (red and green open circles). The form of $h(x)$ does not match the influenza data (diamonds, same as in B). In the model dynamics, $h(x) < 1$ for epitope mutations signals the absence of interference interactions. Model parameters: sequence length $L_{ep} = 120$ (epitope sites), $L_{ne} = 160$ (non-epitope sites), mutation rate $\mu = 5.8 \times 10^{-3}$/year, average scaled selection strength $\bar{\sigma}N = 100$, selection flip rates $\gamma = 3.3 \times 10^{-2}$/year (clonal interference), and $\gamma = 3.6 \times 10^{-3}$/year (episodic sweeps). For model and simulation details, see File S1. Comparisons with further control models are shown in Figure S7.

frequencies with probability higher than synonymous changes, before they are interfered with by a stronger competing clone. With a loss propagator ratio $h(0.7) \approx 2$, at least half the of changes that reach frequency $x = 0.7$ and are subsequently lost are inferred to be beneficial. As we discuss below, our inference of clonal interference is consistent with the underlying mechanisms of immune selection.

Clonal interference is global: In agreement with previous studies (Rambaut *et al.* 2008; Russell *et al.* 2008), we find strains with similar HA1 sequences occurring in the same year to be distributed over different geographical regions (see Figure S4). Migration of strains and occasional multiple originations may contribute to this mixing, which implies that the competition between strains takes place on a global scale. However, given the existence of a source population (Rambaut *et al.* 2008; Russell *et al.* 2008), this competition may well be fiercest in that population and some strains may be driven to extinction before migrating to other regions.

A minimal model of influenza evolution

Further insight into influenza's mode of evolution can be gained by comparing these population-genetic data to simulated evolution of a population of nonrecombining sequences. In contrast to previous model-based studies of influenza's epidemiological and spatial dynamics (Ferguson *et al.* 2003; Gog *et al.* 2003; Tria *et al.* 2005; Koelle *et al.* 2006; Strelkowa 2006; Fraser *et al.* 2009; Pybus and Rambaut 2009), we focus on genome evolution under mutations, genetic drift, and a minimal model of selection: (i) non-epitope sites have time-independent selection coefficients, so that most new mutations there are under negative selection, and (ii) epitope sites have selection with time-dependent direction: the preferred allele at any of these sites changes stochastically with a given rate, opening windows of positive selection and setting the supply of new beneficial mutations (Mustonen and Lässig 2007, 2008, 2009; Schiffels *et al.* 2011). The time dependence of selection describes the emergence of new beneficial epitopes resulting from immune escape, as well as selection changes due to reassortment with other genome segments (Holmes *et al.* 2005; Rambaut *et al.* 2008). Our minimal model is simpler than the actual process in two ways: the fitness of a strain is an additive function of its epitope and non-epitope alleles, and most selection coefficients at individual sites are constant over polymorphism lifetimes. The model does not introduce the epistatic interactions and the population history dependence of immune selection, so as to display the coupling of sequence sites by genetic linkage alone and to be independent of specific immune interaction mechanisms between strains (see *Discussion*). This is in tune with the scope of our simulations, which is to corroborate the inference of clonal interference, to contrast it with other modes of adaptation, and to explore its biological consequences. Details of our model and simulations are described in File S1.

The minimal clonal-interference model reproduces the influenza data: The minimal model dynamics depends on mutation rate, sequence length, strength and flip rate of selection, and population size. To calibrate the model with the actual process, we set mutation rate and sequence length to influenza values, and we fit the remaining three parameters by matching sequence diversity and epitope and non-epitope substitution rates. The calibrated model shows that selection on influenza is strong compared to genetic drift (with products of average selection coefficient and effective population size of order 100) and dynamic (an average epitope codon changes its preferred allele about every 30 years). In this regime, high supply of new beneficial epitope mutations generates clonal interference, as shown in Figure 1A. Beneficial mutations arise in different subpopulations after limited waiting times, and these changes compete for fixation. Clonal interference produces a dense pattern of selective sweeps, which are marked by clusters of nucleotide fixations. Despite its simplicity and few fit parameters, the calibrated minimal model matches several distinct characteristics of the influenza data set. These include the general pattern of recurrent selective sweeps, such as the shape of the strain tree and the strongly non-Poissonian distribution of yearly fixation numbers; see Figure S1, Figure S5, and Figure S6. Importantly, the model also reproduces the functional dependence of the propagator ratios: at intermediate frequencies, $g(x)$ saturates to values significantly >1 and $h(x)$ raises to peak values significantly >1; see Figure 4, A and B. These ratios are specific markers of clonal interference, as shown by the control models discussed below. We conclude that clonal interference is a parsimonious explanation of these data.

Clonal interference is compatible with epistasis: The biophysics of host–pathogen protein interactions generates a fitness landscape with epistasis between epitope changes, which is more complicated than our minimal model. Although single epitope mutations with large antigenic effect have been reported (Smith *et al.* 2004), combinations of several amino acid changes may often be required to produce new beneficial epitope variants (Rimmelzwaan *et al.* 2005; Koelle *et al.* 2006; Shih *et al.* 2007; Kryazhimskiy *et al.* 2011). Clonal interference is compatible with epistatic fitness landscapes, as long as the evolutionary process produces a sufficient supply of new beneficial mutations. In our minimal influenza model, the effects of epistasis are captured in an approximate way by selection flips at individual genomic sites. Given that the minimal clonal interference model matches the influenza data, we do not attempt to fit these data to an extended model with explicit immune interactions. Any such model would involve several more fit parameters compared to the minimal model and, thus, add little statistical significance to the analysis. Our results raise an important caveat for the analysis of evolutionary correlations between epitope sites: any inference of epistasis must carefully discount the effects of genetic linkage.

Control models without clonal interference do not match the influenza data: To test the specificity of our derivation of clonal interference, we introduce control models without clonal interference and show that they are incompatible with the influenza propagator ratios. The minimal model with a lower rate of epitope selection flips is shown in Figure 1B. In this regime, a low supply of new beneficial epitope mutations generates episodic selective sweeps, such that waiting periods between sweeps are longer than the fixation time of each individual sweep. We obtain propagator ratios $g(x) < 1$ and $h(x) < 1$, which are clearly incompatible with the influenza data; see Figure 4, C and D.

To test whether epistasis can produce a spurious signal of clonal interference in the propagator statistics, we introduce an escape mutant model, details of which are given in File S1. This model has strong synergistic epistasis, as expected for epitope adaptation (Ferguson *et al.* 2003; Gog *et al.* 2003; Tria *et al.* 2005; Koelle *et al.* 2006; Shih *et al.* 2007; Minayev and Ferguson 2009; Koelle *et al.* 2010; Kryazhimskiy *et al.* 2011). There is a parameter regime of episodic sweeps, which are interspersed with extended neutral search processes in epitope sequence space. In this regime, we find generic epitope propagator ratios $g(x) \approx 1$ and $h(x) \approx 1$; see Figure S7, C and D. Thus, simple epistasis without clonal interference cannot explain the influenza data.

For completeness, Figure S7, E and F, shows propagator ratios for independent sites evolving under positive or negative directional selection. This case can be solved analytically and serves as a useful illustration of the propagator method (see File S1). Depending on the sign of selection, the propagator ratio $g(x)$ increases or decreases without saturation. The loss propagator ratio $h(x)$ is always <1, reflecting the absence of interference interactions.

Our analysis shows that quantitative statistics of the punctuated sweep pattern and of polymorphism time series produces quite specific tests for models for influenza evolution and, in particular, for clonal interference. Propagator ratios, in particular, are more sensitive to the mode of evolution than the qualitative shape of strain trees, which is reproduced by any model with selective sweeps. Explaining the propagator data in the absence of clonal interference is likely to require a complicated model with fine tuning of several parameters, compared to the parsimonious explanation by the minimal clonal interference model.

Biological implications of clonal interference

How does influenza's adaptive dynamics depend on human immune challenge and viral genome architecture? In the model representation, we can probe how this process responds to changes of its input parameters and derive consequences of clonal interference for viral functions. To characterize the efficiency of the adaptive process, we use two quantities:

i. The *degree of adaptation* is defined by

$$\alpha = \frac{F - F_0}{F_{\max} - F_0}, \tag{4}$$

where F is the mean Malthusian population fitness, $F_{\max}$ is the fitness of a maximally adapted genotype (which carries the preferred nucleotide at all sites), and F_0 is the average fitness of random genotypes (which would arise from neutral evolution) (Mustonen and Lässig 2007). Hence, $1 - \alpha$ is a normalized measure of genetic load. We separately evaluate the degree of adaptation for epitope sequence, α_{ep}, and for non-epitope sequence, α_{ne}.

ii. The speed of adaptation is measured by the mean fitness flux

$$\phi = U_{ep}\Sigma_{ep}, \tag{5}$$

where U_{ep} is the rate and Σ_{ep} the average selection coefficient of epitope substitutions (Mustonen and Lässig 2007, 2009, 2010). The mean fitness flux φ is the time derivative of the cumulative fitness flux $\Phi(t)$, averaged over the strains in a population and over time; *cf*. Figure 1.

Our model predicts response patterns highlighting the interference of epitope and linked non-epitope loci with each other's evolution and function:

Functionality decreases with increasing rate of immune challenge: Clonal interference limits functionality and speed of adaptation, because beneficial mutations are continuously lost in the competition between strains (Gerrish and Lenski 1998). But it also limits the functionality of linked non-epitope loci by hitchhiking of deleterious mutations within sweeps. Both effects increase with increasing flip rate of epitope selection, leading to decrease of α_{ep} and α_{ne} and *sublinear* increase of φ, as shown in Figure 5A. For example, the load on non-epitope sites, $1 - \alpha_{ne}$, is an order of magnitude higher at influenza parameters than in the regime of episodic sweeps. These results suggest that an increase in immune challenge would strongly compromise the viability of influenza.

Functionality decreases with increasing genome size: Linked non-epitope loci limit epitope functionality and speed of adaptation by background selection (Kaiser and Charlesworth 2009), even if these loci have no functional connection to the adaptive process (as in our model). This effect increases proportionally to the length of linked non-epitope sequence, leading to decrease of α_{ep} and φ, as shown in Figure 5B. Indeed, the influenza genome is partitioned into short segments of linked sequence, suggesting that this genome architecture may have evolved partly to reduce the deleterious effects of background selection by reassortment between segments.

Discussion

As we have shown here, population-genetic analysis of time-dependent strain data opens a new avenue to understand influenza. We infer fitness and genetic constraints as determinants of adaptive evolution, and we predict speed and functional consequences of this process. We find that influenza evolves by clonal interference. That is, its adaptation is limited not by the supply of beneficial mutations, but by their competition. This mode of evolution explains the observed pattern of recurrent selective sweeps with clustering of nucleotide fixations in the viral hemagglutinin genome. Clonal interference is generated by genetic linkage and high supply of beneficial mutations. It is compatible with, but does not depend on, epistasis between antigenic epitope changes.

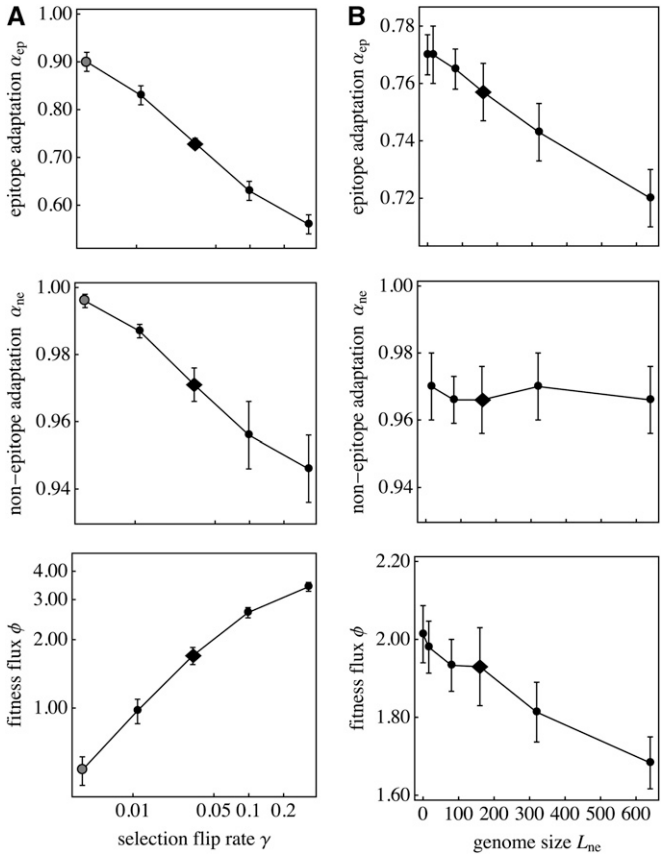

Figure 5 Genome functionality and speed of adaptation. The degree of adaptation, α, characterizes the functionality of a gene segment; the mean fitness flux, φ, measures the speed of adaptation (Mustonen and Lässig, 2007, 2009, 2010). (A) Model simulation results for α_{ep} (epitope sites), α_{ne} (non-epitope sites), and φ are plotted against the selection flip rate γ at epitope sites (solid diamonds, influenza calibration point $\gamma = 3.3 \times 10^{-2}$/year; shaded circles, episodic sweeps for $\gamma = 3.6 \times 10^{-3}$/year). All other model parameters are kept fixed to the influenza calibration point; see Figure 3. There is a γ-dependent *adaptive genetic load* ($1 - \alpha_{ep}$) on epitope sites and ($1 - \alpha_{ne}$) on linked non-epitope sites, and the fitness flux φ increases sublinearly with γ. (B) The same quantities are plotted against the non-epitope genome size L_{ne}, with all other model parameters kept fixed (solid diamonds, influenza calibration point $L_{ne} = 120$). The epitope genetic load ($1 - \alpha_{ep}$) increases and the fitness flux φ decreases with increasing length of linked sequence.

Frequency propagators measure adaptation and interference

Our main result rests on a new inference method for adaptive evolution in asexual populations, which uses polymorphism

frequency time-series data. Two summary statistics of these time series, the frequency propagator ratio $g(x)$ and the loss propagator ratio $h(x)$, are defined in Equations 2 and 3 for classes of mutations under selection compared to a neutral reference class. We infer clonal interference if two conditions are fulfilled: $g(x) > 1$ and $h(x) > 1$ for intermediate and large frequencies x. The first condition signals predominantly positive selection for a class of mutations, and the second indicates interference interactions: new beneficial alleles rise to substantial frequency, but are eventually driven to loss by a competing clone. Figure 4, A and B, shows this inference for nonsynonymous mutations in the influenza HA1 epitopes.

The frequency propagator method has a straightforward interpretation in terms of the distribution of mutations on the genealogical tree. The position of a mutation on a given branch marks the origination of a new allele in the population. Fixed mutations are mapped onto the trunk of the tree, lost mutations onto off-trunk branches. In particular, mutations reaching high intermediate frequencies before loss originate close to, but not on the trunk. Hence, the frequency propagator statistics under clonal interference implies that beneficial mutations are overrepresented, compared to neutral mutations, on the trunk as well as close to the trunk. This distribution of beneficial mutations can be understood as a fitness grading of the genealogical tree, which links influenza evolution and the propagator method to recent advances in the statistics of coalescent processes under selection (Brunet *et al.* 2008) and to directed polymers with quenched disorder (Bolthausen and Sznitman 1998). Because beneficial mutations seed high-fitness clones that expand in the population, two high-fitness strains sampled from the population at a given point in time have, on average, a more recent common ancestor than two random strains. This generates a statistics of *adaptive coalescent processes*, which differs from the familiar neutral coalescent (Kingman 1982). In particular, coalescence times between pairs of strains have a a distribution of different shape (Brunet *et al.* 2008) and a different overall scale, which is set by the characteristic sweep time instead of the effective population size (Schiffels *et al.* 2011).

Clonal interference and immune selection

The adaptive evolution of influenza is driven by host–pathogen interactions, which generate cross-immunity between strains: hosts infected with one strain become partially immune against infection by similar strains. These interactions lead to natural selection on the viral population, which has two key characteristics. First, partial escape from cross-immunity recurrently generates new strains with beneficial mutations, which carry new epidemics. At the same time, some residual competition between all coexisting strains maintains a bounded pool of susceptible hosts and suppresses "speciations" of influenza A into independent lineages. The source of this competition is debated; a possible mechanism is short-term unspecific immunity (Ferguson *et al.* 2003). The inference of clonal interference depends on both of these selection characteristics and imposes an additional constraint: in any model of the immune dynamics, compatibility with the genome data requires a rate of beneficial epitope mutations high enough to generate competition between coexisting mutant strains. Thus, our analysis addresses an important challenge: to establish a link between influenza's epidemiology and genome evolution (Holmes and Grenfell 2009). This link remains to be developed further in future work. Our current population-genetic model does build on a specific host–pathogen mechanism. Hence, it does not yet explain *how* beneficial mutations and a bounded host pool arise.

Influenza's host–pathogen interactions translate into a more complex fitness landscape than the simple picture of directional selection underlying our analysis—a similar caveat applies to selection inference in just about any wild population. However, our main result of clonal competition arises in a natural way from immune interactions. The selective effects of such interactions can be described by a standard susceptible-infected-recovered (SIR) model. In this type of model, each viral strain has a fitness (growth rate) that decreases monotonically with time, reflecting the buildup of specific host immunity. Hence, an epidemic caused by a single strain has a characteristic time course, with numbers of infected host individuals showing an initially exponential growth followed by a rapid decline. Now consider a second strain with an epitope mutation that substantially reduces its specific immunity. At its origination, this mutant has a positive fitness difference (selection coefficient) relative to the first strain. Modeling of the SIR dynamics shows that the epidemic caused by the mutant strain and the corresponding buildup of specific immunity will occur with a time delay relative to the first epidemic (Lin *et al.* 2003). In this process, the mutant will in general remain under positive directional selection and will displace the first strain in a selective sweep, as observed in the actual process of influenza (Plotkin *et al.* 2002). If at most two strains with different specific immunity coexist at any point in time, these sweeps are episodic and lead to propagator ratios $g(x) \leq 1$ and $h(x) \leq 1$, as shown in Figure 4, C and D, and Figure S7. If there are frequently three or more such strains, the SIR model is in the clonal interference regime. Because all strains have a monotonically declining fitness, strains originating later are more likely to be positively selected against earlier strains than vice versa. Thus, immune selection is a mechanism that fuels clonal interference by producing beneficial mutations.

In the clonal competition regime, all strains rise in frequency in the population of infected host individuals, reach a peak value often much smaller than one, and are subsequently lost. (This should not be confused with the rise and fall of total population numbers in an epidemic, which does not require clonal competition.) Alleles at individual epitope sites behave differently than strains. A mutant allele is subject to positive or negative interference, depending on

whether subsequent beneficial mutations occur predominantly in the mutant lineage or in the ancestral background. Thus, interference interactions have two effects: epitope alleles under overall positive directional selection have increased fixation rates; at the same time, some beneficial epitope alleles are lost. This is signaled by the joint occurrence of propagator ratios $g(x) > 1$ and $h(x) > 1$, as shown in Figure 4, A and B. It is conceivable that future studies go beyond the level of alleles and infer more specific characteristics of the influenza SIR dynamics from genomic data. This will require a larger and less biased strain sample than available at present. At the same time, the specific population genetics of the SIR fitness landscape will have to be developed.

Interference couples adaptation and conservation

Clonal interference has an important biological consequence: it tightly couples conservation and adaptation of viral functions that are encoded in linked genome sequence. Viral fitness crucially depends on antigenic adaptation, which takes place primarily by amino acid changes in antigenic epitope sites. However, it also depends on the conservation of protein stability and other functional traits, which are encoded in HA domains outside the epitopes and in other genome segments. Viral proteins have only marginally stable folds (Tokuriki *et al.* 2009), which is consistent with the observation that a large fraction of mutations are deleterious (Sanjuán *et al.* 2004) and with the recent inference of specific compensatory mutations affecting the stability of influenza hemagglutinin (Bloom and Glassman 2009). The simplest population-genetic model of stability-changing mutations is a mutation-selection equilibrium in a single-step fitness landscape: all protein states with a free energy below some threshold are viable folds, and all others are lethal (Zeldovich *et al.* 2007; Bloom and Glassman 2009). This model has recently been extended to continuous-fitness landscapes (Wylie and Shakhnovich 2011). Our analysis affects the population genetics of protein stability in two ways. First, we observe few nonsynonymous substitutions outside epitope sites, but a substantial number of polymorphisms at intermediate frequencies. This suggests that the dependence of fitness on free energy is described by a smooth landscape in which the deleterious effects of many non-epitope mutations are comparable in magnitude to the beneficial effects of epitope changes. Second, clonal interference drives the distribution of protein stabilities in the viral population far off equilibrium, because the frequencies of deleterious changes are enhanced by hitchhiking with beneficial changes in adaptive phenotypes (see Figure 5). Hence, our analysis predicts that fast-adapting viral proteins, in particular hemagglutinin, are more brittle than influenza proteins under less adaptive pressure. A similar argument links the adaptive process with other functional traits under stabilizing selection: clonal interference generates *adaptive genetic load* on conserved functions encoded in linked sequence.

The coupling between adaptation and conservation should be understood as a two-way effect. Not only are protein structure and other viral functions degraded by hitchhiking, the adaptive process itself is compromised by background selection in linked non-epitope sequence (see also Figure 5). Both deleterious effects increase with the length of linked genome segments. This introduces a selection pressure for short genome segments, which may help to explain influenza's genome architecture.

Together, our results imply that the course of influenza evolution is determined not only by antigenic changes. Successful viral strains are those that maximize the total fitness of antigen–antibody interactions and of other viral functions by a joint process of adaptation and conservation. Thus, while antigenic adaptation has been a focus of influenza research so far, this study suggests that we need to broaden our picture of viral function and fitness.

Acknowledgments

This work was partially supported by Wellcome Trust [080711/Z/06] (N.S.) and by Deutsche Forschungsgemeinschaft grant SFB 680 (to M.L.). This work was also supported in part by National Science Foundation under grant PHY05-51164 during a visit to the Kavli Institute of Theoretical Physics (University of California, Santa Barbara).

Literature Cited

Atwood, K. C., L. K. Schneider, and F. J. Ryan, 1951 Periodic selection in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 37: 146–155.

Bao, Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky *et al.*, 2008 The influenza virus resource at the National Center for Biotechnology Information. J. Virol. 82(2): 596–601.

Bhatt, S., E. C. Holmes, and O. G. Pybus, 2011 The genomic rate of molecular adaptation of the human influenza A virus. Mol. Biol. Evol. 28: 2443–2451.

Bloom, J. B., and M. J. Glassman, 2009 Inferring stabilizing nutations from protein phylogenies: application to influenza hemagglutinin. PLOS Comput. Biol. 5: e1000349.

Bolthausen, E., and A. S. Sznitman, 1998 On Ruelle's probability cascades and an abstract cavity method. Commun. Math. Phys. 197: 247–276.

Brunet, E., B. Derrida, and D. Simon, 2008 Universal tree structures in directed polymers and models of evolving populations. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. 78: 061102.

Bush, R. M., C. A. Bender, K. Subbarao, N. J. Fox, and W. M. Fitch, 1999 Predicting the evolution of human influenza A. Science 286: 1921–1925.

Desai, M. M., and D. S. Fisher, 2007 Beneficial mutation–selection balance and the effect of linkage on positive selection. Genetics 176: 1759–1798.

de Visser, J. A. G. M., C. W. Zeyl, P. J. Gerrish, J. L. Blanchard, and R. E. Lenski, 1999 Diminishing returns from mutation supply rate in asexual populations. Science 283: 404–406.

Good, B. H., I. M. Rouzine, D. J. Balick, O. Hallatschek, and M. M. Desai, 2012 Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proc. Natl. Acad. Sci. USA 109: 4950–4955.

Ferguson, N. M., A. P. Galvani, and R. M. Bush, 2003 Ecological and immunological determinants of influenza evolution. Nature 422: 428–433.

Fraser, C., C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove *et al.*, 2009 Pandemic potential of a strain of Influenza A (H1N1): early findings. Science 324: 1557–1561.
Gerrish, P. J., and R. E. Lenski, 1998 The fate of competing beneficial mutations in an asexual population. Genetica 102/103: 127–144.
Gillespie, J. H., 1991 *The Causes of Molecular Evolution*. Oxford University Press, Oxford.
Gillespie, J. H., 1993 Episodic evolution of RNA viruses. Proc. Natl. Acad. Sci. USA 90: 10411–10412.
Gog, J. R., G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and B. T. Grenfell, 2003 Population dynamics of rapid fixation in cytotoxic T lymphocyte escape mutants of influenza A. Proc. Natl. Acad. Sci. USA 100(19): 11143–11147.
Holmes, E. C., and B. T. Grenfell, 2009 Discovering the phylodynamics of RNA viruses. PloS Comp. Biol. 5: e1000505.
Holmes, E. C., E. Ghedin, N. Miller, J. Taylor, Y. Bao *et al.*, 2005 Whole-genome analysis of human inuenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS Biol. 3: e30.0
Kaiser, V. B., and B. Charlesworth, 2009 The effects of deleterious mutations on evolution in non-recombining genomes. Trends Genet. 25: 912.
Kingman, J. F. C., 1982 On the genealogy of large populations. J. Appl. Probab. 19A: 2743.
Koelle, K., S. Cobey, B. Grenfell, and M. Pascual, 2006 Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. Science 314: 1898–1903.
Koelle, K., P. Khatri, M. Kamradt, and T. Kepler, 2010 A two-tiered model for simulating the ecological and evolutionary dynamics of rapidly evolving viruses, with an application to influenza. J. R. Soc. Interface 7: 1257–1274.
Kryazhimskiy, S., G. A. Bazykin, J. B. Plotkin, and J. Dushoff, 2008 Directionality in the evolution of influenza A haemagglutinin. Proc. Biol. Sci. 275: 2455–2464.
Kryazhimskiy, S., J. Dushoff, G. A. Bazykin, and J. B. Plotkin, 2011 Prevalence of epistasis in the evolution of influenza A surface proteins. PLoS Genet. 7: e1001301.
Lässig, M., 2012 Chance and risk in adaptive evolution. Proc. Natl. Acad. Sci. USA 109: 4719–4720.
Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49: 49–67.
Li, W. H., C. I. Wu, and C. C. Luo, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2: 150–174.
Lin, J., V. Andreasen, R. Casagrandi, and S. A. Levin, 2003 Traveling waves in a model of influenza A drift. J. Theor. Biol. 222: 437–445.
McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654.
Miller, C. R., P. Joyce, and H. A. Wichman, 2011 Mutational effects and population dynamics during viral adaptation challenge current models. Genetics 187: 185–202.
Minayev, P., and N. Ferguson, 2009 Improving the realism of deterministic multi-strain models: implications for modelling influenza A. J. R. Soc. Interface 6: 509–518.
Miralles, R., P. J. Gerrish, A. Moya, and S. F. Elena, 1999 Clonal interference and the evolution of RNA viruses. Science 285: 1745–1747.
Mustonen, V., and M. Lässig, 2007 Adaptations to fluctuating selection in Drosophila. Proc. Natl. Acad. Sci. USA 104: 2277–2282.
Mustonen, V., and M. Lässig, 2008 Molecular evolution under fitness fluctuations. Phys. Rev. Lett. 100: 108101.
Mustonen, V., and M. Lässig, 2009 From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet. 25: 111–119.
Mustonen, V., and M. Lässig, 2010 Fitness flux and ubiquity of adaptive evolution. Proc. Natl. Acad. Sci. USA 107: 4248–4253.
Neher, R. A., and B. I. Shraiman, 2009 Competition between recombination and epistasis can cause a transition from allele to genotype selection. Proc. Natl. Acad. Sci. USA 106: 6866–6871.
Nielsen, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. 39: 197–218.
Park, S.-C., and J. Krug, 2007 Clonal interference in large populations. Proc. Natl. Acad. Sci. USA 104: 18135–18140.
Pennings, P. S., and J. Hermisson, 2006 Soft sweeps II: molecular population genetics of adaptation from recurrent mutation or migration. Mol. Biol. Evol. 23: 1076–1084.
Perfeito, L., L. Fernandes, C. Mota, and I. Gordo, 2007 Adaptive mutations in bacteria: high rate and small effects. Science 317: 813–815.
Plotkin, J. B., J. Dushoff, and S. A. Levin, 2002 Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc. Natl. Acad. Sci. USA 99(9): 6263–6268.
Pybus, O. G., and A. Rambaut, 2009 Evolutionary analysis of the dynamics of viral infectious disease. Nat. Rev. Genet. 10: 540–550.
Rambaut, A., O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger *et al.*, 2008 The genomic and epidemiological dynamics of human influenza A virus. Nature 453: 615–619.
Rimmelzwaan, G. F., E. G. Berkhoff, N. J. Nieuwkoop, D. J. Smith, R A M. Fouchier *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.
Russell, C. A., T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten *et al.*, 2008 The global circulation of seasonal influenza A (H3N2) viruses. Science 320: 340–346.
Sanjuán, R., A. Moya, and S. F. Elena, 2004 The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc. Natl. Acad. Sci. USA 101: 8396–8401.
Schiffels, S., G. J. Szollosi, V. Mustonen, and M. Lässig, 2011 Emergent neutrality in adaptive asexual evolution. Genetics 189: 1361–1375.
Shih, A. C.-C., T.-C. Hsiao, M.-S. Ho, and W.-H. Li, 2007 Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. Proc. Natl. Acad. Sci. USA 104(15): 6283–6288.
Smith, D. J., A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan *et al.*, 2004 Mapping the antigenic and genetic evolution of Influenza virus. Science 305: 371–376.
Smith, N. G., and A. Eyre-Walker, 2002 Adaptive protein evolution in Drosophila. Nature 415: 1022–1024.
Strelkowa, N., 2006 Influenza Dynamics, Diploma Thesis. University of Cologne, Cologne, Germany.
Tokuriki, N., C. J. Oldfield, V. N. Uversky, I. N. Berezovsky, and D. S. Tawfik, 2009 Do viral proteins possess unique biophysical features? Trends Biochem. Sci. 34: 53–59.
Tria, F., M. Lässig, L. Peliti, and S. Franz, 2005 A minimal stochastic model for influenza evolution. J. Stat. Mech. P07008.
Wiley, D. C., I. A. Wilson, and J. J. Skehel, 1981 Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature 289: 373–378.
Wilke, C. O., 2004 The speed of adaptation in large asexual populations. Genetics 167: 2045–2053.
Wolf, Y. I., C. Viboud, E. C. Holmes, E. V. Koonin, and D. J. Lipman, 2006 Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. Biol. Direct 1: 34.
Wylie, C. S., and E. I. Shakhnovich, 2011 A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc. Natl. Acad. Sci. USA 108: 9916–9921.
Zeldovich, K. B., P. Chen, and E. I. Shakhnovich, 2007 Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc. Natl. Acad. Sci. USA 104: 16152–16157.

Communicating editor: J. J. Bull

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.143396/-/DC1

Clonal Interference in the Evolution of Influenza

Natalja Strelkowa and Michael Lässig

DOI: 10.1534/genetics.112.143396

Supporting Information
Supporting Methods and Figures S1 - S7

Here, we describe the influenza sequence dataset (Section 1), the reconstruction of strain trees from these data (Section 2), the estimation of population frequencies (Section 3), additional features of the propagator method (Section 4), and the models of sequence evolution used in this study (Section 5).

1. Sequence data

Our study is based on a dataset of 1971 sequences available from the NCBI database (Bao et. al, 2008). This dataset is well suited for the history-based inference of selection and evolutionary mode: it contains 160 substitutions in the HA1 domain distributed over a time span of 39 years, which is much larger than the average polymorphism lifetime of about 3 years; see Fig. 3. This allows for an accurate inference of substitution rates, whereas datasets with shorter observation periods would involve larger sampling errors (see Section 4).

The available influenza sequences are clearly not a randomly sampled dataset, which would be ideal for population-genetic analysis. Known systematic biases in the dataset include:

(i) Yearly variations in sampling depth. Far fewer strains are available for earlier years than for later years.

(ii) Regional variations in sampling depth. In particular, the New York sequence project (Ghedin et al., 2005) leads to an overrepresentation of US sequences.

(iii) Passage history effects. Egg-cultured strains show additional mutations, which may cause sampling bias (Bush et al., 2000).

Our analysis addresses these biases as follows:

(i) Pre-processing of the dataset: We include only sequences which contain the full HA1 domain (at least 987 bp) and are annotated by year and location of observation. Lab strains and marked egg isolates are excluded. Sequences from the New York project (Ghedin et al., 2005) are only partially included: for each year, we choose a random subset of these sequences, such that the fraction of US sequences is capped to a maximum percentage. However, we have checked that the propagator statistics does not change if all New York sequences are included.

(ii) Our conclusions are based on polymorphism time-series at substantial frequencies. For these data, the assumption of a geographically mixed population is justified. This is shown by Fig. S4, which confirms the results of previous studies (Rambaut et al., 2008; Russell et al., 2008). Furthermore, these frequencies are robust to variations of the sampling depth.

(iii) The propagator method is robust to variations in polymorphism entry time and entry frequency, which are expected to be particularly noisy in our dataset. Furthermore, propagator ratios do not depend frequency-dependent bias in polymorphism numbers, which can arise from our tree-based inference. For details, see Section 4.

The NCBI accession numbers of our strain sample are given in File S2. We obtain a gapless alignment of these sequences using MUSCLE (Edgar, 2004). Within the HA1 domain, we use a subset of codons as known antigenic epitope sites (Shi et al., 2007).

2. Reconstruction of strain trees

Tree structure and statistics. Our analysis of polymorphism histories is based on an ensemble of strain trees obtained from the HA1 sequence dataset. Such trees describe the genealogy of influenza strains resulting from a coalescent process under selection (Rosenberg and Nordborg, 2002). The tree ensemble is constructed with PAUP (Swofford, 2002) using a heuristic procedure to obtain globally optimized maximum-parsimony trees, which consists of random addition of branches followed by branch swapping. The procedure is used with an option to bias the mapping of mutations towards earlier years (ACCTRAN), following the procedure of previous studies (Fitch et al., 1997; Kryazhimskiy et al., 2008). Trees are rooted using the strain A/Bilthofen/16190/68 (NCBI accession number AY661039), which is closest to the avian outgroup of the HA1 domain (Smith et al, 2004).

In these trees, each node corresponds to a unique HA1 sequence in a given year, and each observed strain is mapped onto exactly one external node. Strains with the same HA1 sequence observed in the same year are mapped onto the same external node, and we count their number as multiplicity m of the node. Strains with the same HA1 sequence observed in different years are mapped onto different external nodes. A strain with descendants is represented by an external node and its internal father node, to which these descendants are linked (i.e., these two nodes have identical sequences). The remaining internal nodes represent unobserved sequences inferred by maximum parsimony.

Our tree statistics is built from 10 PAUP runs differing in the order of sequences added; each run produces 100 equiprobable trees. Variation between the trees occurs only on peripheral branches; the large-scale tree structure and the tree-based statistical observables are well conserved. Statistical errors in our tree-based selection inference are discussed in detail in Section 4.

An example of a maximum-parsimony tree is shown in Fig. S1. The overall consistency of the tree reconstruction procedure is supported by the correct timing of the observed strains (Fitch et al., 1997).

Mapping of mutations. Maximum parsimony maps point mutations between directly related strains onto the branches of the tree. A mutation on a given branch marks an origination event of a single nucleotide polymorphism, i.e., the appearance of a nucleotide difference between the clone of strains descending from the branch and its ancestral lineage. Fig. S2 shows these originations partitioned in the three classes used in our analysis: synonymous mutations, nonsynonymous mutations outside the epitope, and nonsynonymous epitope mutations. Tree-based inference can accurately disentangle synonymous and nonsynonymous changes, which become ambiguous in the raw sequence data if several changes in the same codon are observed in one year.

Timing of internal nodes. Each tree node is assigned a year of occurence as follows: Nodes representing observed HA1 sequences are assigned their year of observation, all other nodes are assigned the year for which the average D value of observed sequences is closest to the D value of the inferred node sequence. However, if any (external) descendant node occurs in an earlier year, the assigned year of the internal node is correspondingly advanced. Here D is the mutational distance of a node sequence to the sequence of the root node, i.e., the number of point mutations in the lineage between the two nodes (which can differ from their Hamming distance due to double mutations at the same sequence position).

3. Estimation of population frequencies

Strain frequencies. Each node of a timed strain tree is assigned a multiplicity m as approximate measure of the frequency of its HA1 sequence. For an external node, m is the number of occurrences of its sequence in the strains sampled in the corresponding year. For an internal node, m is the number of descendant nodes in the same year differing by a single point mutation in the HA1 sequence, which is seen to correlate well with population size in model simulations (Strelkowa, 2006).

Each HA1 sequence a occurring in a given year is assigned a multiplicity m_a, which is the sum of the m values of its (one or two) nodes, and a frequency $x_a = m_a / \sum_b m_b$ with the normalization given by all sequences b of the same year.

Polymorphism frequencies. The frequency x of a nucleotide allele in a given year is the sum of the frequencies x_a of all sequences in that year which carry the allele. The resulting allele frequency time-series are qualitatively similar to those of a previous study (Shi et al., 2007). However, our tree-based inference allows decomposing x into contributions of individual clones, which appear as descendant subtrees of a unique origination (see next paragraph). Furthermore, the entry point of an allele can be inferred on an internal node prior to its first observation.

Haplotype frequencies. For any pair of simultaneously polymorphic sequence sites in the HA1 domain, we consider the four possible haplotypes

$$(-,-), \quad (a,-), \quad (-,b), \quad (a,b), \tag{S1}$$

where a is the mutant allele at site 1, b is the mutant allele at site 2, and dashes denote the ancestral alleles. We compare the frequency x_{12} of the double-mutant haplotype (a,b) with the (marginal) frequencies x_1 and x_2 of the single-nucleotide mutant alleles a and b. To quantify genetic association in the HA1 domain, it is convenient to decompose these frequencies into clonal components. Assume that allele a has c_1 independent originations on different branches of the strain tree. These define a set of c_1 mutually disjoint, but temporally overlapping clones (i.e., subtrees) carrying the same mutant allele at site 1. In the same way, the mutant allele b is carried by a set of c_2 mutually disjoint clones. The allele frequencies are the sum of the clone frequencies x^α ($\alpha = 1, \ldots, c_1$) and x^β ($\beta = 1, \ldots, c_2$) at site 1 and site 2, respectively,

$$x_1 = \sum_{\alpha=1}^{c_1} x^\alpha \tag{S2}$$

and

$$x_2 = \sum_{\beta=1}^{c_2} x^\beta. \tag{S3}$$

The double-mutant haplotype frequency is given by

$$x_{12} = \sum_{\alpha=1}^{c_1} \sum_{\beta=1}^{c_2} x^{\alpha\beta} \tag{S4}$$

with

$$x^{\alpha\beta} = \begin{cases} x^\alpha & \text{if clone } \alpha \text{ is nested in clone } \beta, \\ x^\beta & \text{if clone } \beta \text{ is nested in clone } \alpha, \\ 0 & \text{if clones } \alpha \text{ and } \beta \text{ are disjoint.} \end{cases} \tag{S5}$$

We define the mutant alleles a and b to be under complete genetic association if only two of the three mutant haplotypes $(a,-)$, $(-,b)$, and (a,b) occur in the population. Complete genetic association signals that all originations of the mutant allele at one site occur on the same sequence background (ancestral or mutant) of the other site. More specifically, there are three distinct cases leading to complete association: (i) all originations of allele a occur nested in clones carrying allele b (i.e., $x_{12} = x_1 \leq x_2$), (ii) all originations of allele b occur nested in clones carrying allele a (i.e., $x_{12} = x_2 \leq x_1$), or (iii) all originations of both alleles occur in disjoint clones (i.e., $x_{12} = 0$). Pairs of mutations with unique originations on the tree ($c_1 = c_2 = 1$) are always under complete association. However, if at least one of the mutant alleles has multiple originations ($c_1 > 1$ or $c_2 > 1$), complete association can be broken, i.e., $0 < x_{12} < \min(x_1, x_2)$.

Fig. S3 shows scaled haplotype frequencies

$$y_{12} \equiv \frac{x_{12}}{\min(x_1, x_2)} \tag{S6}$$

for pairs of simultaneous polymorphisms in different mutation classes of the HA1 domain. In the vast majority of cases, we find values $y_{12} = 0$ or $y_{12} = 1$ indicative of complete genetic association between the mutant alleles. However, some haplotypes have scaled frequencies $0 < y_{12} < 1$, signaling originations of the mutant allele at one site on multiple sequence backgrounds at the other site (Shi et al., 2007; Kryazhimskiy et al., 2008).

Allele frequency correlation. We measure the effects of genetic linkage on the haplotype statistics of influenza HA1 by the frequency correlation $\mathcal{C}(x_{12}, x_1, x_2)$, which is defined in equation (2) of the main text. This correlation measures the degree of genetic association between mutant alleles and takes values between 0 and 1. The maximum $\mathcal{C} = 1$ signals that only two of the three mutant haplotypes $(a, -)$, $(-, b)$, and (a, b) occur in the population, which implies $x_{12} = \min(x_1, x_2)$ or $x_{12} = 0$.

We can compare $\mathcal{C}$ with Lewontin's

$$\mathcal{D}'(x_{12}, x_1, x_2) \equiv \frac{\mathcal{D}(x_{12}, x_1, x_2)}{\mathcal{D}_{\max}}, \tag{S7}$$

where $D_{\max}$ is the absolute value of the maximum or minimum linkage disequilibrium consistent with given allele frequencies, i.e., $D_{\max} = \min(x_1(1-x_2)), x_2(1-x_1))$ if $x_{12} \geq x_1 x_2$ and $D_{\max} = \min(x_1 x_2, (1-x_1)(1-x_2))$ if $x_{12} < x_1 x_2$ (Lewontin, 1964). This normalized measure of linkage disequilibrium takes values between -1 and 1. The maximum absolute value $|\mathcal{D}'| = 1$ signals that only three of the four haplotypes $(-, -)$, $(a, -)$, $(-, b)$, and (a, b) occur in the population. It is easy to show the inequality $\mathcal{C} \leq |\mathcal{D}'|$. As a consequence, our result of nearly complete genetic association between mutant alleles in the HA1 domain, $\bar{\mathcal{C}} = 0.96$, implies an equally strong average linkage disequilibrium in terms of Lewontin's measure, $\overline{|\mathcal{D}'|} \geq 0.96$. We find that $\mathcal{C}$ and $|\mathcal{D}'|$ take equal values for most HA1 haplotypes. The key difference between the two measures is that $\mathcal{C}$ distinguishes between ancestral and mutant alleles, which makes it a more specific measure of the haplotype origination statistics than $|\mathcal{D}'|$. The strict inequality $\mathcal{C} < |\mathcal{D}'|$ holds if and only if $x_1 + x_2 > 1$ and $x_{12} < x_1 x_2$. In particular, if all three mutant haplotypes $(a, -)$, $(-, b)$, and (a, b) occur in the population but the ancestral haplotype $(-, -)$ has been lost, we obtain $\mathcal{C} < 1$ and $|\mathcal{D}'| = 1$. The correlation $\mathcal{C}$ signals originations on mixed sequence backgrounds, while linkage equilibrium has become extremal by loss of the ancestral haplotype.

4. Propagator method

Definition of propagators and propagator ratios. In this study, we use *frequency propagators* of polymorphism time-series as statistical measures of selection and as markers of clonal interference. The frequency propagator $G(x|x_i)$ is defined as the conditional probability that a polymorphism with frequency x_i at some first point of its history reaches a frequency $x > x_i$ at any later point. This observable is easily estimated from the frequency time-series in our dataset, $G(x|x_i) = n(x)/n(x_i)$, where $n(x)$ is the number of polymorphisms that reach frequency x. As a measure of selection, polymorphism histories are most informative if they are evaluated from their entry points observed in the sample. The resulting frequency propagator $G(x)$ is the average of $G(x|x_e)$ over the distribution of entry frequencies x_e in the sample; it is estimated in the dataset as $G(x) = n(x)/n$, where n is the total number of polymorphisms. The distribution of entry frequencies x_e depends on the sample size. Typical entry frequencies are quite variable in our dataset, because fewer data are available for early years. A more robust measure of selection is the ratio of propagators between a class of nonsynonymous polymorphisms and a neutral reference class of synonymous polymorphisms,

$$g(x) = \frac{G(x)}{G_0(x)}, \tag{S8}$$

which is largely independent of the entry frequencies, as long as they are sufficiently small (see below).

In a similar way, we evaluate polymorphisms whose new allele reaches frequencies exceeding a given threshold x at some intermediate point of its lifetime but is eventually lost. The likelihood of this process is given by the *loss propagator* $H(x) \equiv G(0|x)G(x)$, and we define the propagator ratio

$$h(x) = \frac{H(x)}{H_0(x)} \tag{S9}$$

with respect to the neutral reference class.

Systematic errors. The propagator method is designed to be applicable to the dataset of this study, because it is quite robust to various uncertainties and biases in the data:

(i) Propagator ratios are insensitive to variations in polymorphism entry frequencies (see above). Such variation is generated, for example, because our dataset contains fewer strains from earlier years and more from later years.

(ii) Propagator ratios are insensitive to frequency-dependent bias in polymorphism numbers $n(x)$, as long as it does not depend on mutation class. Such bias is generated, for example, by spurious mutations in egg isolates, which produce an excess number of low-frequency polymorphisms. Furthermore, we choose a conservative minimum entry frequency $x_{e,\min} = 0.01$, which excludes low-frequency polymorphisms located primarily on terminal branches of the coalescent tree.

(iii) Propagator ratios do not depend on the precise timing of polymorphism histories. Variations in entry times are generated, for example, by fluctuations between equiprobable trees.

Statistical errors. Selection inference by the propagator method is subject to two distinct sources of statistical error:

(i) Sampling fluctuations arise, because the system is observed over a finite period of time and, therefore, the absolute number of polymorphism histories reaching a given frequency is limited. These fluctuations turn out to be the dominant source of statistical error for frequency propagators (and prohibit the use of this method for other datasets with shorter observation spans). The error bars reported in Fig. 4 treat different data points as independent, which leads to an overestimation of sampling errors.

(ii) Fluctuations between equiprobable trees arise from the genealogy reconstruction process. We analyze these fluctuations using an ensemble of 1000 equiprobable trees obtained for our dataset (see Section 2 above). The resulting statistical errors for frequency propagators are found to be subleading to sampling errors, showing that our inference is robust to variations between trees.

Frequency propagators for independent sites, low-frequency asymptotics. Here, we analytically calculate the frequency propagators of an independent two-allele site evolving by mutations, genetic drift, and constant selection. This serves as an illustration of the propagator method and shows that frequency propagator ratios are asymptotically independent of entry frequencies.

The expression for G is obtained by generalizing the familiar calculation of the fixation probability (Kimura, 1983): $G(x|x_i)$ is the solution of the stationary backward diffusion equation

$$\frac{1}{2N}\frac{\partial^2}{\partial x_i^2}\,G + \sigma\frac{\partial}{\partial x_i}\,G = 0 \tag{S10}$$

with boundary conditions $G(x|x_i) = 1$ and $G(x|0) = 0$, where σ is the selection coefficient and N is the effective population size. Defining the scaled selection coefficient $s = 2N\sigma$, the solution reads

$$G(x|x_i) = \frac{1 - \mathrm{e}^{-sx_i}}{1 - \mathrm{e}^{-sx}}. \tag{S11}$$

For neutral evolution ($s = 0$), this expression reduces to $G_0(x|x_i) = x_i/x$.

In the limit $x_i \ll 1$, the propagator $G(x|x_i)$ has a linear asymptotic dependence on x_i. Hence, the propagator $G(x)$, which is defined as the average of $G(x|x_e)$ over entry frequencies x_e, can be written in the form

$$G(x) = G(x|\bar{x}_e) + O(\overline{x_e^2}), \tag{S12}$$

where $\bar{x}_e$ and $\overline{x_e^2}$ are mean and variance of entry frequencies. The ratio of propagators G/G_0 becomes asymptotically independent of entry frequencies in this limit:

$$\frac{G(x)}{G_0(x)} = g(x) + O(x_e). \tag{S13}$$

From eq. (S11), we obtain

$$g(x) = \frac{sx}{1 - \mathrm{e}^{-sx}} \tag{S14}$$

and as a special case the ratio of fixation probabilities

$$g \equiv g(1) = \frac{s}{1 - \mathrm{e}^{-s}}. \tag{S15}$$

In the same way, the the ratio of loss propagators becomes asymptotically independent of entry frequencies,

$$\frac{H(x)}{H_0(x)} \equiv \frac{G(x)}{G_0(x)} \frac{G(0|x)}{G_0(0|x)} = h(x) + O(x_e), \tag{S16}$$

where

$$h(x) = \frac{sx}{1 - \mathrm{e}^{-sx}} \frac{1 - \mathrm{e}^{s(1-x)}}{1 - e^{s}} \frac{1}{1 - x}. \tag{S17}$$

The loss propagator for independent sites has values $h(x) < 1$ for mutations under positive and under negative selection. This is because the function $H(x)$ decreases exponentially with increasing x under constant selection of any direction: alleles under negative selection are unlikely to reach substantial frequencies x, whereas alleles under positive selection are unlikely to be lost once they have reached such x. As an example, the functions $g(x)$ and $h(x)$ are plotted in Fig. S7(e,f) for the cases of neutral evolution ($s = 0$), moderate negative selection ($s = 2N\sigma = -6$), and moderate positive selection ($s = 2N\sigma = 6$). They are incompatible with the influenza data, which have $g(x)$ saturating at intermediate frequencies and $h(x) > 1$ due to clonal interference.

For linked sites, the propagator ratios $g(x)$ and $h(x)$ differ drastically from the form of eqs. (S14) and (S17), but they remain asymptotically independent of x_e as given by eqs. (S13) and (S16). This reflects the fact that the low-frequency dynamics of polymorphisms is always dominated by genetic drift.

5. Sequence evolution models

Model dynamics. In this study, we use simple models of sequence evolution under genetic linkage for two purposes: (i) A minimal model of clonal interference serves to infer evolutionary parameters of influenza's adaptive dynamics and to corroborate the propagator-based inference of clonal interference for this process. (ii) Control models which do not match the influenza data indicate the specificity of our evidence for clonal interference.

We consider a model population with a constant number N individuals. In the model dynamics, this parameter governs the relative importance of genetic drift compared to selection and mutations. We estimate numerical values of N from observed HA1 sequence diversity, as described below. In the actual evolutionary process, genetic drift is dominated by extreme bottlenecks during transmission between hosts, which involve a number of viral particles of order one. Therefore, the model parameter N should be associated with an effective number of infected hosts (and not with typical numbers of viral particles). Keeping N constant then reflects the well-known property that influenza A strain diversity does not proliferate and a bounded pool of susceptible and infected hosts is maintained (modulating N by seasonal changes does not affect our results).

The population is partitioned into subpopulations of N_β individuals infected by a given strain; different strains are distinguished by an index β. Each strain is characterized by its genotype $\mathbf{a}^\beta = (a_1^\beta, \dots, a_L^\beta)$, which is a sequence of length L partitioned into three classes of sites:

$$\mathbf{a}^\beta = (\underbrace{a_1^\beta, \dots, a_{L_{\rm ep}}^\beta}_{L_{\rm ep}\text{ epitope sites}}, \underbrace{a_{L_{\rm ep}+1}^\beta, \dots, a_{L_{\rm ep}+L_{\rm ne}}^\beta}_{L_{\rm ne}\text{ non-epitope sites}}, \underbrace{a_{L_{\rm ep}+L_{\rm ne}+1}^\beta, \dots, a_L^\beta}_{L-L_{\rm ep}-L_{\rm ne}\text{ neutral sites}}) \tag{S18}$$

Each sequence site has two nucleotide alleles $a_i = \pm 1$ ($i = 1, \dots, L$). The order of epitope, non-epitope, and neutral sites on the sequence is arbitrary, because genotypes evolve without recombination.

Strain content and population sizes evolve by selection, mutations, and genetic drift:

(i) *Selection*: In our minimal model, we use an additive, but explicitly time-dependent fitness function

$$F(\mathbf{a}, t) = F_{\rm ep}(\mathbf{a}, t) + F_{\rm ne}(\mathbf{a}, t) = \sum_{i=1}^{L} \frac{1}{2} \sigma_i \eta_i(t)\, a_i. \tag{S19}$$

Epitope and non-epitope sites have selection coefficients of magnitude $\sigma_i > 0$ independently drawn from a log-normal distribution with average $\bar{\sigma}$ and variance proportional to $\bar{\sigma}$ (the emergence of clonal interference is robust under changes of this distibution (Gerrish and Lenski, 1998)). For epitope sites, the direction of selection $\eta_i(t) = \pm 1$ fluctuates (Mustonen and Lässig, 2007) according to independent random processes with rate γ, non-epitope sites have a time-independent direction $\eta_i(t) = 1$, and neutral sites have $\sigma_i = 0$. Over a time interval Δt, selection generates a deterministic change in subpopulation sizes,

$$N_\beta(t) \to Z^{-1}(t)\, N_\beta(t) \exp[(\Delta t) F(\mathbf{a}^\beta, t)] \tag{S20}$$

with the normalization $Z(t) = \sum_\beta N_\beta(t) \exp[(\Delta t) F(\mathbf{a}^\beta, t)]/N$.

(ii) *Mutations:* For each strain β, we draw the number of mutant individuals from a Poisson distribution with mean $\mu L(\Delta t)$, choosing the time step Δt such that this mean is of order 1. Each mutant individual of strain β acquires a single point mutation $a_i^\beta \to -a_i^\beta$ at a randomly chosen site i and, thus, may belong to another existing strain or seed a new strain.

(iii) *Genetic drift:* After the selection and mutation steps, we define the population numbers $N_\beta(t + \Delta t)$ of the next generation by multinomial sampling, i.e., each individual is randomly assigned a single parent individual of the previous generation, which transmits its genotype. As discussed above, this sampling models the transmission between hosts.

Population observables. Following the evolution process over time, we can measure the following quantities:

(i) Sequence diversity

$$\pi(t) \equiv \frac{1}{2N^2} \sum_{\beta<\beta'} \sum_{i=1}^{L} N_\beta(t) N_{\beta'}(t) \big(1 - a_i^\beta a_i^{\beta'}\big). \tag{S21}$$

(ii) Epitope degree of adaptation

$$\alpha_{\rm ep}(t) = \frac{F_{\rm ep} - F_{\rm ep,0}}{F_{\rm ep,max} - F_{\rm ep,0}} = \frac{1}{N}\sum_{\beta} N_\beta(t)\,\frac{1}{\bar\sigma L_{\rm ep}}\sum_{i=1}^{L_{\rm ep}} \sigma_i \eta_i(t)\, a_i^\beta, \tag{S22}$$

where the second equality uses $F_{\rm ep,0} = 0$ and $F_{\rm ep,max} = \frac{1}{2}\bar\sigma L_{\rm ep}$.

(iii) Non-epitope degree of adaptation

$$\alpha_{\rm ne}(t) = \frac{F_{\rm ne} - F_{\rm ne,0}}{F_{\rm ne,max} - F_{\rm ne,0}} = \frac{1}{N}\sum_{\beta} N_\beta(t)\,\frac{1}{\bar\sigma L_{\rm ne}}\sum_{i=L_{\rm ep}+1}^{L_{\rm ep}+L_{\rm ne}} \sigma_i a_i^\beta. \tag{S23}$$

(iv) Total substitution rates of epitope and non-epitope sites, $U_{\rm ep}$ and $U_{\rm ne}$.

(v) Epitope fitness flux

$$\phi(t) = \frac{1}{N}\sum_{\beta,\beta'} N_{\beta\beta'}(t)\sum_{i=1}^{L_{\rm ep}} \frac{1}{2}\sigma_i\eta_i(t)\big(a_i^{\beta'} - a_i^\beta\big), \tag{S24}$$

where $N_{\beta\beta'}(t)$ is the number of individuals mutating from strain β to strain β' at time step t.

The process reaches a stationary state characterized by time-independent average values π^s, $\alpha^s_{\rm ep}$, $\alpha^s_{\rm ne}$, $U^s_{\rm ep}$, $U^s_{\rm ne}$, and $\phi^s = U^s_{\rm ep}\Sigma^s_{\rm ep}$, where $\Sigma^s_{\rm ep} > 0$ is the average selection coefficient of epitope substitutions. These observables depend on the model parameters L, $L_{\rm ep}$, $L_{\rm ne}$, $\bar\sigma$, γ, μ, and N.

Simulation procedures. The simulation is started at time t_0 with a population containing a single strain ($\beta = 1$) with a random epitope genotype and a perfectly adapted non-epitope genotype,

$$a_i^1 = \begin{cases} \pm\eta_i(t_0) & \text{for } i = 1,\dots,L_{\rm ep},\\ 1 & \text{for } i = L_{\rm ep}+1,\dots,L_{\rm ep}+L_{\rm ne}.\end{cases} \tag{S25}$$

This strain has epitope degree of adaptation $\alpha_{\rm ep} \approx 0$ and non-epitope degree of adaptation $\alpha_{\rm ne} = 1$.

After evolution over a few years, the population reaches a stationary state with stochastic fluctuations. This state has a few hundred coexisting strains, an adapted epitope ($\alpha_{\rm ep} > 0$), genetic load outside the epitope ($\alpha_{\rm ne} < 1$), and a finite speed of adaptation ($\phi > 0$), as shown in Fig. 5.

The data of Fig. 4 and of Fig. S7 are obtained by averaging over 10 runs with 400 years of stationary evolution in each run. The trees of Fig. 1 and Fig. S5 show single runs of stationary evolution, which are directly comparable to the data tree of Fig. S1. The distribution of yearly fixation numbers shown in Fig. S6 is obtained from 10 runs with 40 years of stationary evolution in each run.

Model parameters, evolutionary regimes. For a given set of model parameters, we record the above observables in the stationary state of the population dynamics. To compare the minimal model to the dynamics of influenza, a number of model parameters are chosen equal to their actual values:

(i) The point mutation rate is set to $\mu = 5.8 \times 10^{-3}$ per nucleotide and year. This value is inferred from the rate of neutral substitutions in the HA1 domain, confirming the result of a previous study (Fitch et al., 1999). Clonal interference strongly affects the polymorphism histories of neutral changes, but not their substitution rate: a new allele which has evolved neutrally up to population frequency x and is interfered with by a selective sweep, has a probability of fixation equal to x, the same value as for neutral evolution without the sweep. Hence, the substitution rate of neutral changes remains a measure of the mutation rate in an individual sequence. In our sample of the influenza HA1 domain, there are

about 75 synonymous substitutions over 39 years, 11 of which occur in the 62 epitope codons and 66 in the 267 non-epitope codons; these numbers are consistent with a uniform point mutation rate across the HA1 domain and produce the value of μ quoted above.

(ii) The sequence length parameters are set to $L_{\rm ep} = 120$, corresponding to 60 epitope codons in the HA1 domain, and $L_{\rm ne} = 160$ corresponding to about 80 codons under moderate negative selection. For definiteness, this number is chosen equal to the number of non-epitope codons where originations are observed in the HA1 domain. The actual number of non-epitope codons coupled to the epitope by linkage is larger. However, the evolutionary observables depend only weakly on $L_{\rm ne}$ (see Fig. 5(b)) and a substantial fraction of non-epitope mutations are expected to be under strong purifying selection ($\sigma \gg \bar{\sigma}$), for example, because they cause misfolds. These changes decouple from the clonal interference dynamics. The model sequences also contain $L - L_{\rm ep} - L_{\rm ne} = 300$ neutral sites, equal to the number of codons in the HA1 domain.

With these choices, the minimal model has only three fit parameters: the average strength of selection, $\bar{\sigma}$, the fluctuation rate of selection, γ, and the population size N. We evaluate the model in the following parameter regimes:

(i) *Clonal interference regime.* This regime includes the influenza calibration point, which is determined by fitting the substitution rates $U^s_{\rm ep}$, $U^s_{\rm ne}$ and the diversity π^s to the values observed in the actual process. The fit values should be regarded as order-of-magnitude estimates, because clonal interference flattens the dependence of the evolutionary process on the population-genetic parameters (Gerrish and Lenski, 1998; Desai and Fisher, 2007). At these parameter values, clonal interference of the model dynamics manifests itself in a high supply of beneficial mutations at high frequencies (Fig. 4(a)), loss propagator values $h(x) > 1$ (Fig. 4(b)), the distribution of beneficial mutations on the strain tree (Fig. 1(a)), and in a sublinear increase of fitness flux with γ (Fig. 5(a)).

We probe the dependence of the minimal model dynamics on γ and on $L_{\rm ne}$ around the influenza calibration point $\gamma = 3.3 \times 10^{-2}$/yr, $L_{\rm ne} = 160$ with all other parameters kept fixed (Fig. 5).

(ii) *Episodic sweeps regime.* This regime is reached for substantially lower values of γ. The simulations shown in Figs. 1(b) and 4(c,d) have $\gamma = 3.6 \times 10^{-3}$/yr; see also the regime of low γ in Fig. 5(a). Episodic sweeps are characterized by a low number of beneficial mutations at high frequencies (Fig. 4(c)), loss propagator values $h(x) < 1$ (Fig. 4(d)), a distribution of beneficial mutations on the strain tree as shown in Fig. 1(b), and in a linear increase of fitness flux with γ (Fig. 5(a)).

Epistasis model. Because already the minimal model matches the influenza data, fitting a more complicated model with explicit fitness interactions between epitope sites would add little statistical significance to our analysis. However, we use a simple epistatic model to verify that such interactions are unlikely to produce a spurious signal of clonal interference in the frequency propagator statistics. This model describes neutral searches in epitope sequence space interspersed with selective sweeps triggered by beneficial *escape mutants* (Ferguson et. al, 2003; Gog et al., 2003; Tria et al., 2005; Koelle et al., 2006; Minayev and Ferguson, 2009; Koelle et al., 2010). Starting from an initial genotype with fitness F_0, new epitope mutations are neutral with probability $1 - p$ and lead to a genotype of higher fitness $F_1 = F_0 + \sigma$ with probability p (selection coefficients σ are drawn from a distribution as above). Following a sweep triggered by this mutant, a new search starts, until a second beneficial mutant with fitness $F_2 = F_1 + \sigma'$ occurs, etc. This model has strong synergistic epistasis: most individual mutations are neutral, and a positive fitness effect requires in most cases a combination of mutations away from the previous successful mutant. For low values of p, the model is in a regime of episodic sweeps, i.e., it does not produce clonal interference. In this regime, it shows propagator ratios $g(x) \approx 1$ and $h(x) \approx 1$ for epitope sites, which are characteristic of sparse sweeps and extended neutral evolution of epitope genotypes. These ratios do not match the influenza data; see Fig. S7(c,d).

References

Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D (2008) The Influenza Virus Resource at the National Center for Biotechnology Information. *J. Virol.* 82(2) 596-601

Bush R M, Smith C B, Cox N J, Fitch W M (2000) Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proc Natl Acad Sci USA* 97:69746980.

Bush R M, Bender C A, Subbarao K, Fox N J, Fitch W M (1999) Predicting the Evolution of Human Influenza A *Science* 286:1921-1925

Desai M M, Fisher D S (2007) Beneficial Mutation-Selection Balance and the Effect of Linkage on Positive Selection *Genetics* 176:1759-1798

Edgar R C (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Research* 32(5):1792-1797.

Ferguson N M, Galvani A P, Bush R M (2003) Ecological and Immunological Determinants of Influenza Evolution *Nature* 422:428-433

Fitch W M, Bush R M, Bender C A, COX N J (1997) Long term trends in the evolution of H(3) HA1 human influenza type A *Proc Natl Acad Sci USA* 94:7712-7718

Gerrish P J & Lenski R E (1998) The fate of competing beneficial mutations in an asexual population *Genetica* 102/103: 127-144

Ghedin E, Sengamalay N A, Shumway M, Zaborsky J, Feldblyum T, Subbu V, Spiro D J, Sitz J, Koo H, Bolotov P, Dernovoy D, Tatusova T, Bao Y, St George K, Taylor J, Lipman D J, Fraser C M, Taubenberger J K & Salzberg S L (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution *Nature* 437:1162-1166

Gog J R, Rimmelzwaan G F, Osterhaus A D M E, Grenfell B T (2003) Population dynamics of rapid fixation in cytotoxic T lymphocyte escape mutants of influenza A *Proc Natl Acad Sci USA* 100(19):11143-11147

Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge UK).

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

Koelle, K., Khatri, P., Kamradt, M., Kepler, T. (2010) A two-tiered model for simulating the ecological and evolutionary dynamics of rapidly evolving viruses, with an application to influenza. *J Roy Soc Interface* 7:1257-74

Kryazhimskiy S, Bazykin G A, Plotkin JB, Dushoff J (2008) Directionality in the evolution of influenza A haemagglutinin *Proc R Soc B* 275:2455-2464

Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB (2011) Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins. *PloS Genetics* 7:e1001301

Lewontin, R. C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49-67.

Minayev P and Ferguson N (2009) Improving the realism of deterministic multi-strain models: implications for modelling influenza A. *J Roy Soc Interface* 6:509-518

Mustonen V, Lässig M (2007) Adaptations to fluctuating selection in Drosophila *Proc Natl Acad Sci USA* 104:2277-2282

Rambaut A, Pybus O G, Nelson M I, Viboud C, Taubenberger J K, Holmes E S (2008) The genomic and epidemiological dynamics of human influenza A virus *Nature* 453:615-619

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms *Nat Rev Genet.* 3:380-90.

Russell C A *et al.* (2008) The Global Circulation of Seasonal Influenza A (H3N2) Viruses *Science* 320:340-346

Shih A C-C, Hsiao T-C, Ho M-S, Li W-H (2007) Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution *Proc Natl Acad Sci USA* 104(15):6283-6288

Smith D J, Lapedes A S, de Jong J C, Bestebroer T M, Rimmelzwaan G F, Osterhaus A D M E, Fouchier R A M (2004) Mapping the antigenic and genetic evolution of Influenza virus *Science* 305:371-376

Strelkowa N (2006) *Influenza Dynamics* (Diploma thesis, University of Cologne).

Swofford D L (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tria F, Lässig M, Peliti L, Franz S (2005) A minimal stochastic model for influenza evolution *J. Stat. Mech.* P07008

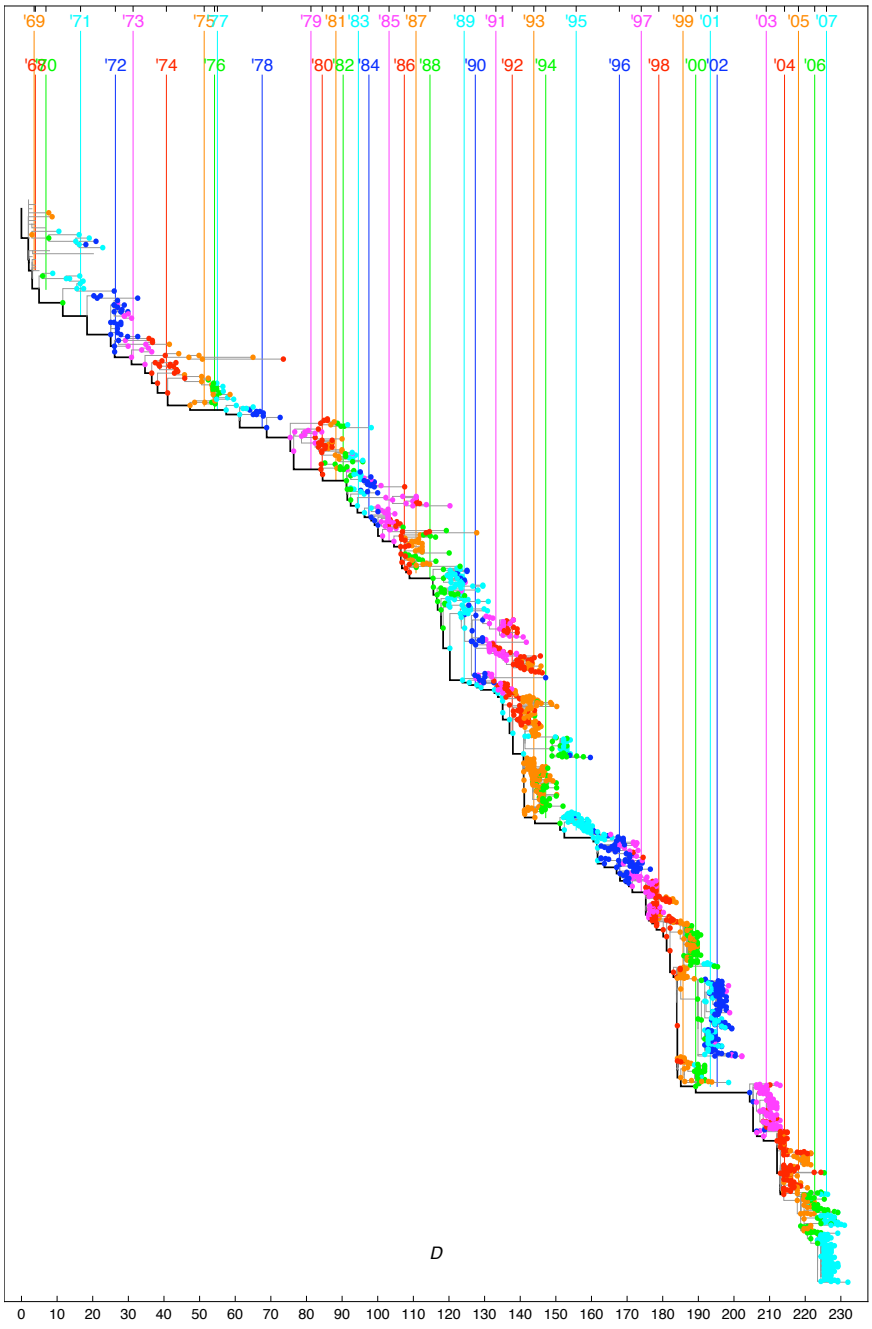

Figure S1: Strain tree of influenza A (H3N2). This sample tree is constructed by maximum parsimony from coding sequence of the HA1 domain of 1971 strains occurring between 1969 and 2007 (other equiprobable trees differ only in peripheral branches). The observed HA1 sequences appear as external nodes. The horizontal coordinate D of a node is its mutational distance from the root of the tree. The trunk of the tree, i.e., the single lineage connecting past and future on time scales beyond the coalescence time, is marked by a thick line. The year of occurrence of all sequences (colored dots) is estimated from their D value for inferred sequences on internal nodes (see *Methods*). The sequences of a given year are seen to be clustered around their average D value (colored lines), which increases by about 5.6 mutations per year.

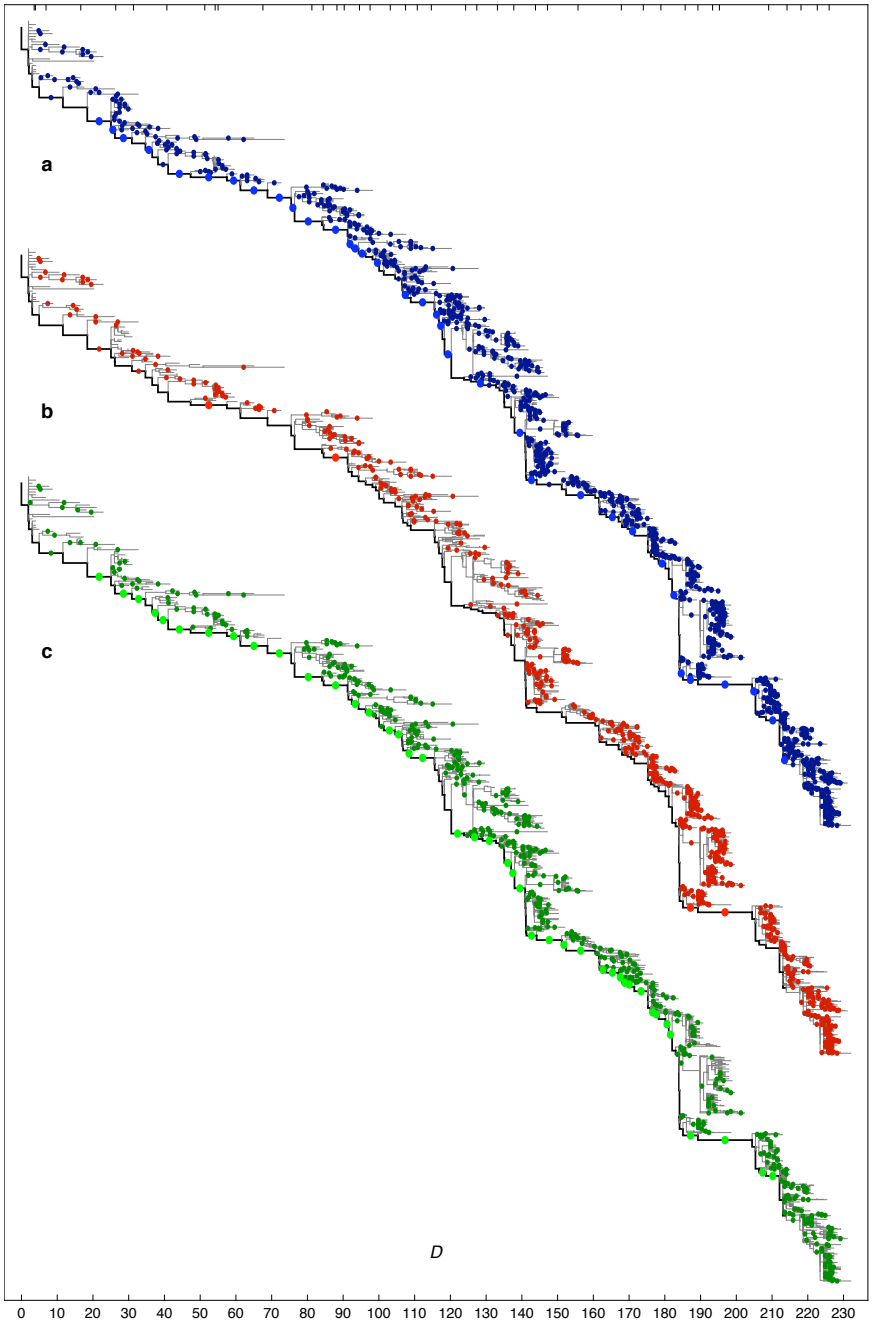

Figure S2: Tree map of mutations. (a) Synonymous (blue), (b) nonsynonymous non-epitope (red), and (c) nonsynonymous epitope changes (green). Each mutation marks an origination of a new allele in the population; each fixed allele has an origination on the trunk of the tree (highlighted by bright colors). The fixation probability, i.e., the ratio of the number of fixations and the number of originations, is seen to be reduced for nonsynonymous non-epitope changes and enhanced for nonsynonymous epitope changes compared to the baseline of synonymous changes.

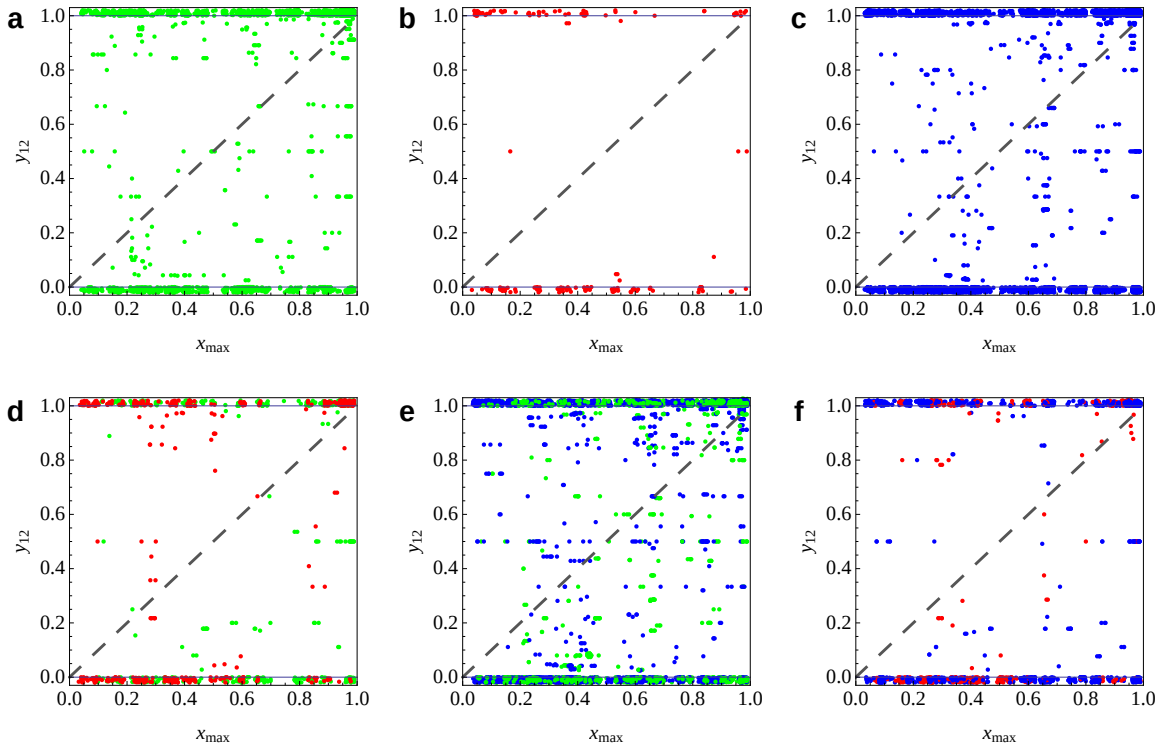

Figure S3: Genetic linkage in the influenza HA1 domain. For pairs of mutations with haplotype frequency x_{12} and marginal (allele) frequencies x_1 and x_2, the scaled haplotype frequency $y_{12} = x_{12}/\min(x_1, x_2)$ is plotted against the larger allele frequency, $x_{\max} = \max(x_1, x_2)$. Yearly frequency data for (a) 934 pairs of nonsynonymous epitope polymorphisms (1969 green points with average frequency correlation $\bar{\mathcal{C}} = 0.948$), (b) 75 pairs of nonsynonymous non-epitope polymorphisms (198 red points, $\bar{\mathcal{C}} = 0.987$), (c) 2022 pairs of synonymous polymorphisms (4118 blue points, $\bar{\mathcal{C}} = 0.964$), (d) 450 pairs of a nonsynonymous epitope polymorphism and a nonsynonymous non-epitope polymorphism (478 green points with larger frequency of the epitope mutant allele, 437 red points with larger frequency of the non-epitope mutant allele, $\bar{\mathcal{C}} = 0.957$), (e) 2738 pairs of a nonsynonymous epitope polymorphism and a synonymous polymorphism (2409 green points with larger frequency of the nonsynonymous mutant allele, 3170 blue points with larger frequency of the synonymous mutant allele, $\bar{\mathcal{C}} = 0.955$), (f) 723 pairs of a nonsynonymous non-epitope polymorphism and a synonymous polymorphism (514 red points with larger frequency of the nonsynonymous mutant allele, 1009 blue points with larger frequency of the synonymous mutant allele, $\bar{\mathcal{C}} = 0.973$). Most points show maximum linkage disequilibrium characteristic of complete genetic linkage, i.e., $y = 1$ for polymorphisms in nested clones and $y = 0$ for polymorphisms in disjoint clones (these points are shown with random y values in the interval $(1, 1.02)$ and $(-0.02, 0)$, respectively, in order to make a larger number of points visible). Some mutations originate in multiple clones and break complete linkage, as shown by values $0 < y_{12} < 1$. However, the overall pattern is far from linkage equilibrium ($y_{12} = x_{\max}$, dashed lines).

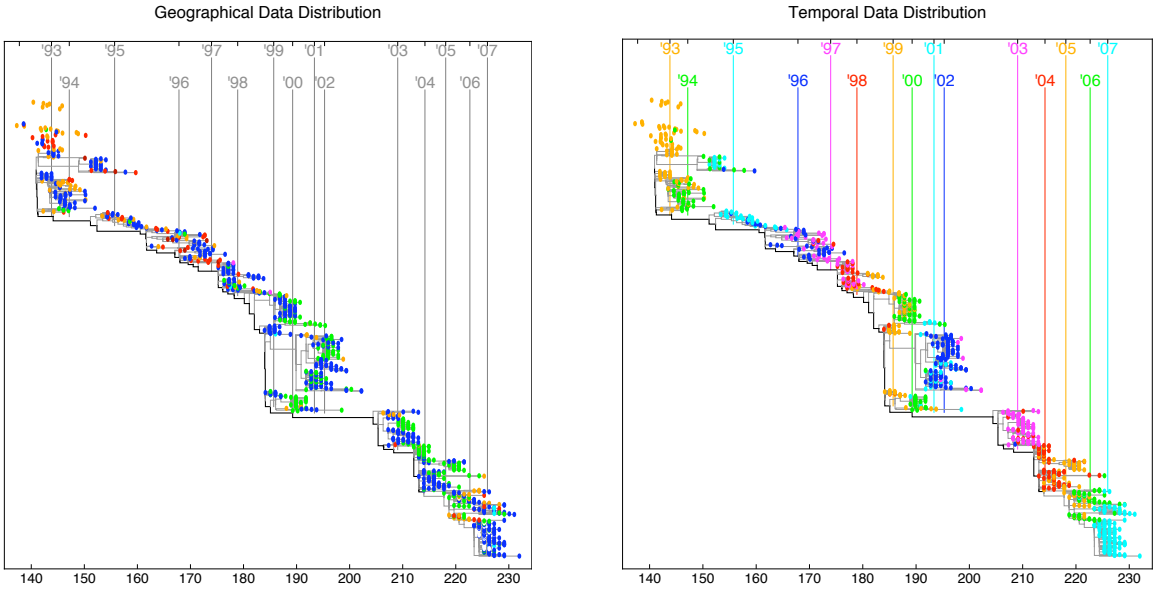

Figure S4: Geographical mixing vs. temporal differentiation of strains. Left panel: geographical location of observed strains (red: Asia, green: Australia, blue: North America, cyan: Central and South America, orange: Europe). The data indicate geographical mixing and confirm the results of previous studies (Rambaut et al., 2008; Russell et al., 2008): Strains with similar HA1 sequences occurring in the same year are distributed over different regions. Right panel: Year of occurence of observed strains (colors as in Fig. S1). The data fall into yearly clusters of increasing mutational distance D to the root node.

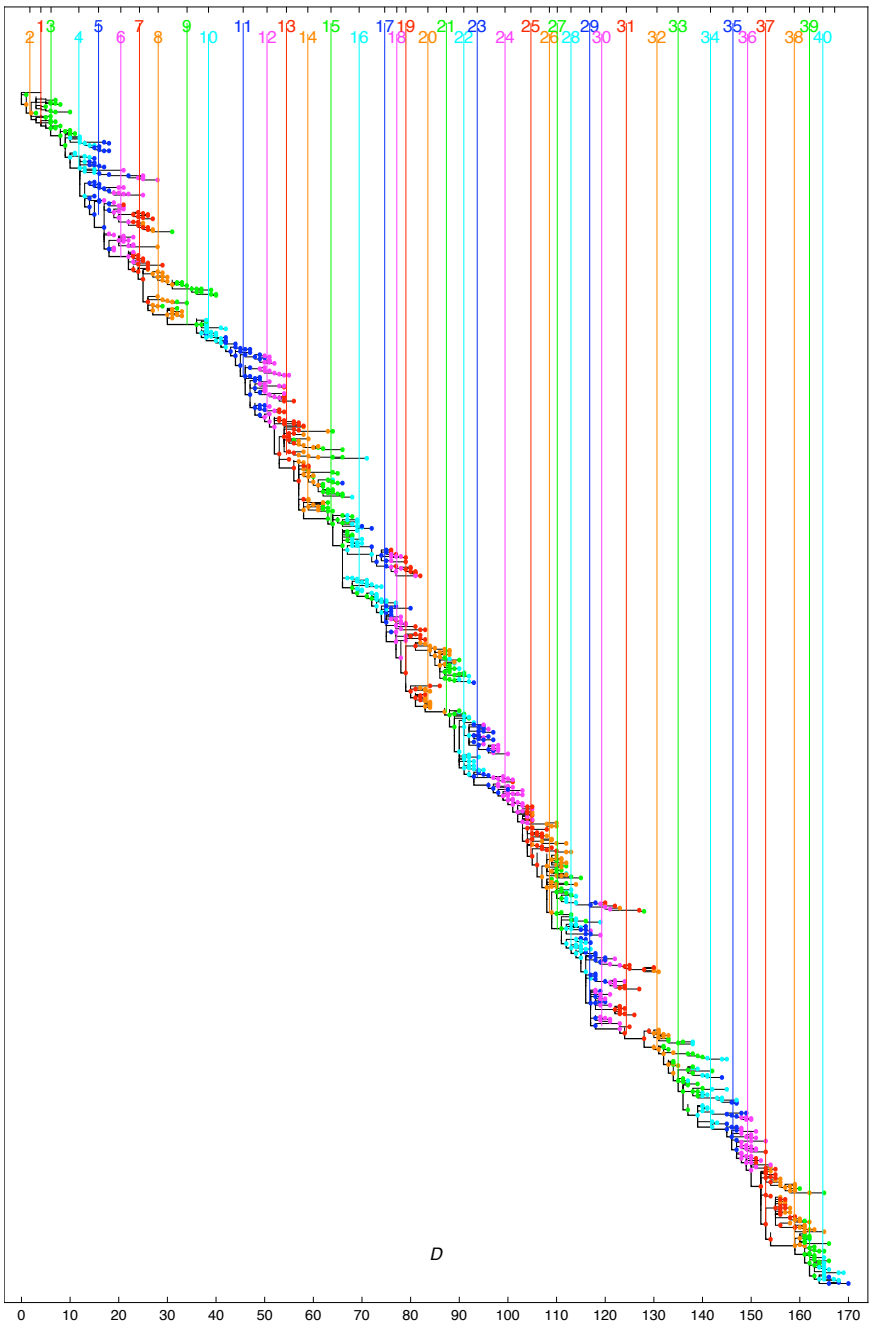

Figure S5: Typical sequence tree of the clonal interference model. The tree is obtained from the evolution of a single population in the stationary state of the minimal model over a time of 40 years (model parameters at the influenza calibration point). To be compared with Fig. S1.

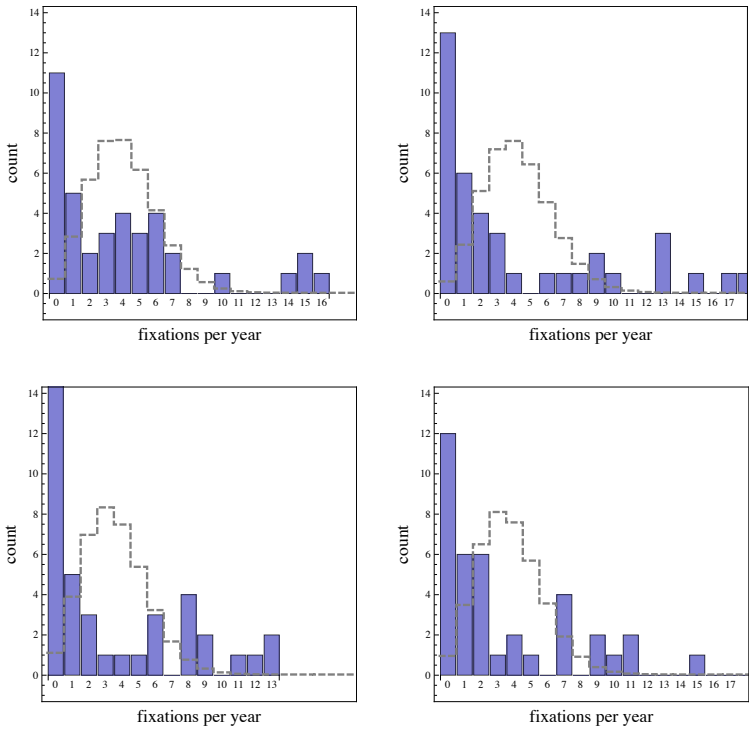

Figure S6: Clustering of fixations in the clonal interference model. Histograms of the number of yearly nucleotide fixation events (bars) obtained from the minimal evolution model in four simulation runs over 40 years (model parameters at the influenza calibration point). To be compared with Fig. 3(c). As in the actual process, these distributions deviate strongly from the Poisson form expected for independently evolving sites (dashed lines). The simulated distribution obtained from 10 runs has a ratio 5.0 ± 2.5 of variance and mean (error bars determined by sampling over a finite number of years). This value is compatible with the corresponding ratio 6.7 for the actual process, however, both ratios are much larger than the range 1 ± 0.25 for a finite sample of Poisson-distributed fixation numbers.

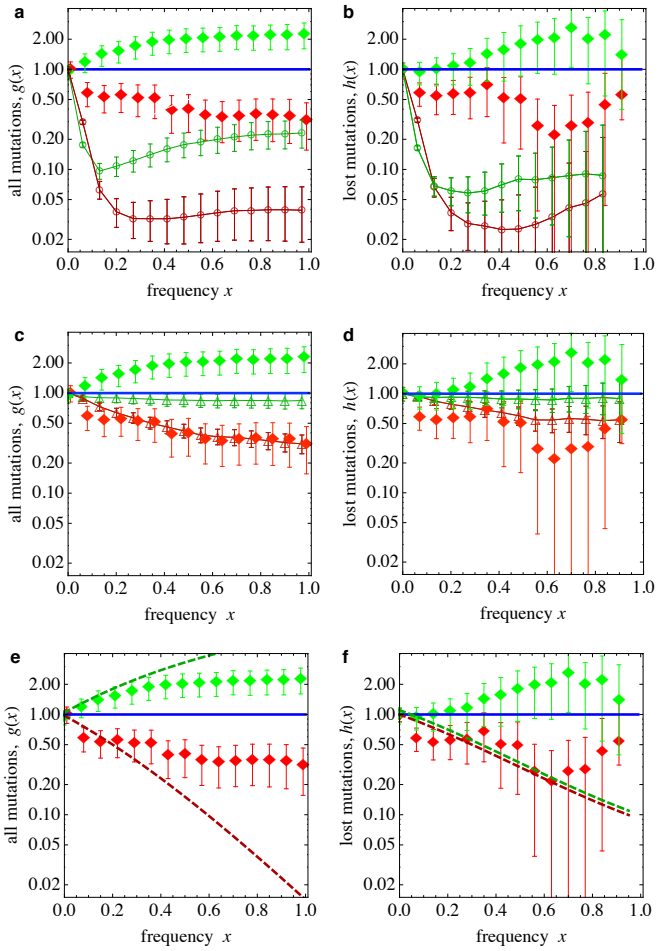

Figure S7: Control models without clonal interference do not match influenza data. (a,c,e) Frequency propagator ratio $g(x)$ and (b,d,f) loss propagator ratio $h(x)$ as defined in the text. Influenza data as in Fig. 3: Observed ratios for nonsynonymous non-epitope and epitope mutations (red and green diamonds, with error bars given by sampling fluctuations) with respect to the baseline of synonymous changes (blue line). The data match none of the following control models: (a,b) Episodic sweeps regime of the minimal model (red and green empty circles, as in Fig. 4). (c,d) Episodic sweeps regime of the epistasis model (red and green empty triangles); see Supporting Text, Section 5. In both models, $g(x) \leq 1$ reflects the low rate of adaptive epitope mutations and $h(x) \leq 1$ the absence of clonal competition. (e,f) Independent sites evolving under negative selection or positive selection (red and green dashed lines, analytical solutions given in Supporting Text, Section 4). In particular, $h(x) < 1$ reflects the absence of clonal competition for unlinked sites.