Clonal interference in the evolution of influenza

Natalja Strelkowa\textsuperscript{a} and Michael Lässig\textsuperscript{b}

(a) Department of Bioengineering, Imperial College London, South Kensington Campus, UK
(b) Institute for Theoretical Physics, University of Cologne, Zülpicherstr. 77, 50937 Köln, Germany

July 21, 2012

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 haemagglutinin. This process involves recurrent selective sweeps, in which clusters of simultaneous nucleotide fixations in the haemagglutinin 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.
Introduction

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 haemagglutinin (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., 1999). 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; Shi 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; 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, where 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 (Shi et al., 2007) and clusters of their fixations are caused primarily by fitness interactions (epistasis) between epitope sites (Shi et al., 2007; Kryazhimskiy et al., 2011).

In this paper, we focus on genomic determinants of influenza’s evolutionary process. As we will 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 factor of its epidemiology (Holmes et al., 2005; Rambaut et al., 2008). They key point of this paper is to show that rapid adaptation under linkage produces clonal interference, a specific mode of evolution by recurrent selective sweeps, within the haemagglutinin 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.,
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). Fig. 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, is incompatible with 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 aminoacid fixations (Plotkin et al., 2002; Wolf et al, 2006; Shi et al., 2007). Our results for neutral polymorphisms, in particular, show that hitchhiking effects are strong, in contrast to a previous analysis based only on nonsynonymous polymorphisms (Shi et al., 2007). In the third and central part of the paper, 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 Fig. S1.
Each 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 Fig. S2). 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 Supporting Methods (File S1).

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 co-existing clones (Shi 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 Supporting Methods). Fig. 2 and Fig. S3 show scaled haplotype frequencies $y_{12} \equiv \frac{x_{12}}{\min(x_1, x_2)}$ 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

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

where $D(x_{12}, x_1, x_2) \equiv x_{12} - x_1 x_2$ is the linkage disequilibrium and $\xi_{12}$ 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 $C = 0$ indicates statistical independence of the two polymorphisms (i.e., linkage equilibrium, $x_{12} = x_1 x_2$), the maximum $C = 1$ complete genetic association between the mutant alleles (i.e., $x_{12} = \min(x_1, x_2)$ or $x_{12} = 0$). Compared to the familiar $D'$ (Lewontin, 1964), the allele frequency correlation $C$ is a stricter measure of association, which is appropriate for the analysis of haplotypes on influenza strain trees (for details, see Supporting Methods). For pairs of HA1 sequence polymorphisms, we find an average frequency correlation $\bar{C} = 0.96$ (for details, see Fig. 2 and Fig. 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 (Fig. 2 and Fig. 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).

Recruent selective sweeps

To understand how linkage affects the evolution of influenza haemagglutinin, 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. Fig. 3(a) 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 $1 \pm 0.25$ for a finite sample of Poisson-distributed values. Defining a fixation cluster as a period with at least 8 nucleotide fixations per year in the HA1 domain, we obtain a total of 8 major clusters, which are marked by dashed lines in Fig. 3(a). 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 more than 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. Non-synonymous 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 (Shi et al., 2007; Kryazhimskiy et al., 2011).

Dips in sequence diversity correlate with fixation clusters. Fig. 3(b) 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 one 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 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. Fig. 3(d) 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 ge-
Netic drift and compete by clonal interference? We will 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 is not confounded by clonal interference and is robust to sampling biases in our dataset. We define the frequency propagator $G(x)$ as the likelihood that a new allele appearing in our sequence sample reaches a frequency larger than $x$ at some later point. We evaluate the ratio

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

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

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

for each class of nonsynonymous mutations 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 origins 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 Supporting Methods). 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 dataset, 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 Fig. 4(a). 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 Fig. 4(a). 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 Fig. 3(c), 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 (Fig. 3(a), 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 Fig. 4(b) takes values \(h(x) > 1\) for intermediate frequencies, signaling positive selection acting on lost mutations. The mutations in this class reach intermediate frequencies with higher probability 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 will 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 Fig. 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
likely 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 non-recombining sequences. In contrast to previous model-based studies of influenza’s epidemiological and spatial dynamics (Gog et al, 2003; Ferguson et al., 2003; Tria et al., 2005; Koelle et al., 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. (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 adaption, and to explore its biological consequences. Details of our model and simulations are described in Supporting Methods.

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 Fig. 1(a). 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 dataset. 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 Figs. S1, S5 and S6. Importantly, the model also reproduces the functional dependence of the propagator ratios: at intermediate frequencies, $g(x)$ saturates to values significantly above 1 and $h(x)$ raises to peak values significantly above 1; see Fig. 4(a,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, 2005; Koelle et al., 2006; Shi 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 Fig. 1(b). 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 Fig. 4(c,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 Supporting Methods. 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; Shi 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 Fig. S7(c,d). Thus, simple epistasis without clonal interference cannot explain the influenza data.

For completeness, Fig. S7(e,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 Supporting Methods). Depending on the sign of selection, the propagator ratio \( g(x) \) increases or decreases without saturation. The loss propagator ratio \( h(x) \) is always below 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

\[
\alpha = \frac{F - F_0}{F_{\text{max}} - F_0},
\]

where \( F \) is the mean Malthusian population fitness, \( F_{\text{max}} \) 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, $\alpha_{ep}$, and for non-epitope sequence, $\alpha_{ne}$. (ii) The speed of adaptation is measured by the mean fitness flux

$$\dot{\phi} = U_{ep} \Sigma_{ep},$$

where $U_{ep}$ is the rate and $\Sigma_{ep}$ the average selection coefficient of epitope substitutions (Mustonen and Lässig, 2007, 2009, 2010). The mean fitness flux $\phi$ is the time-derivative of the cumulative fitness flux $\Phi(t)$, averaged over the strains in a population and over time; cf. Fig. 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 $\alpha_{ep}$ and $\alpha_{ne}$ and sublinear increase of $\dot{\phi}$, as shown in Fig. 5(a). 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 $\alpha_{ep}$ and $\dot{\phi}$, as shown in Fig. 5(b). 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 haemagglutinin 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.

**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, the second indicates interference interactions: new beneficial alleles rise to substantial frequency, but are eventually driven to loss by a competing clone. Fig. 4(a,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 models, 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
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 are under overall positive directional selection and 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 Fig. 4(a,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 which 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, 2009), which is consistent with the observation that a large fraction of mutations are deleterious (Sanjuan et al., 2004) and with the recent inference of specific compensatory mutations affecting the stability of influenza haemagglutinin (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, all others are lethal (Zeldovich et al., 2007; Bloom and Glassman, 2009). This model has recently been extended to continuous fitness landscapes (Wylie et al., 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 Fig. 5). Hence, our analysis predicts that fast-adapting viral proteins, in particular haemagglutinin, 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 Fig. 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).

Supporting Information:

File S1: Supporting Methods
File S2: Supporting Data (NCBI accession numbers of the strains used in this study)
Supporting Figures S1 - S7
Atwood K C, Schneider L K, Ryan F J (1951) Periodic selection in Escherichia coli. *Proc Natl Acad Sci USA* 37:146-155.

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.

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

Bloom J B, Glassman M J (2009) Inferring stabilizing mutations from protein phylogenies: Application to influenza hemagglutinin. *PLoS Comput Biol* 5: e1000349.

Bolthausen E and Sznitman A S (1998) On Ruelle’s probability cascades and an abstract cavity method. *Commun. Math. Phys.* 197:247-276.

Brunet E, Derrida B, Simon D (2008) Universal tree structures in directed polymers and models of evolving populations. *Phys. Rev. E* 78:p061102.

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 and Fisher D S (2007) Beneficial mutation-selection balance and the effect of linkage on positive selection. *Genetics* 176:1759-1798.

Good B H, Rouzine I M, Balick D J, Hallatschek O, Desai M M (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, Galvani A P, Bush R M (2003) Ecological and Immunological Determinants of Influenza Evolution *Nature* 422:428-433.

Fraser, C *et al* (2009) Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings *Science* 324:1557-1561.

Gerrish P J & Lenski R E (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, 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.

Holmes, E. C. *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
Holmes EC, Grenfell BT (2009) Discovering the phylodynamics of RNA viruses. PloS Comp. Biol. 5:e1000505.

Kaiser VB, Charlesworth B (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. Prob. 19A:2743.

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.

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 WH, Wu CI, Luo CC (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, Andreasen V, Casagrandi R, Levin S A (2003) Traveling waves in a model of influenza A drift, J. Theor. Biol. 222:437-445.

McDonald J H, Kreitman M (1991) Adaptive Protein Evolution at the Adh Locus in Drosophila Nature 351:652-654.

Miller C R, Joyce P, Wichman H A (2011) Mutational effects and population dynamics during viral adaptation challenge current models. Genetics 187:185-202.

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.

Miralles R, Gerrish PJ, Moya A, Elena S F (1999) Clonal Interference and the Evolution of RNA Viruses Science 285:1745-1747.

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

Mustonen V, and Lässig M (2008) Molecular Evolution under Fitness Fluctuations, Phys. Rev. Lett. 100:108-101

Mustonen V, Lässig M (2009) From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet. 25:111-119.
Mustonen V, Lässig M (2010) Fitness flux and ubiquity of adaptive evolution. *Proc Natl Acad Sci USA* 107:4248-4253.

Neher RA, Shraiman BI (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, Krug J (2007) Clonal interference in large populations *Proc Natl Acad Sci USA* 104:18135-18140.

Pennings PS, Hermisson J (2006) Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration *Mol Biol Evol.* 23:1076-84.

Perfeito L, Fernandes L, Mota C, Gordo I (2007) Adaptive mutations in bacteria: High rate and small effects. *Science* 317:813-815.

Plotkin J B, Dushoff J, Levin S A (2002) Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus *Proc Natl Acad Sci USA* 99(9):6263-6268.

Pybus O G, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease *Nat Rev Genet* 10:540 - 550.

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.

Rimmelzwaan G F, Berkhoff E G, Nieuwkoop N J, Smith D J, Fouchier R A M, Osterhaus A D M E (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 et al. (2008) The Global Circulation of Seasonal Influenza A (H3N2) Viruses *Science* 320:340-346.

Sanjun R, Moya A, Elena SF (2004) The distribution of fitness effects caused by single-nucleotid substitutions in an RNA virus. *Proc. Natl. Acad. Sci* 101:8396-401.

Schiffels S & Szollosi G J, Mustonen V, Lässig M (2011) Emergent Neutrality in Adaptive Asexual Evolution, *Genetics* 189:1361-1375.

Shih A C-C, Ilisiao 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 NG, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila *Nature* 415:1022-4.

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).

Tokuriki N, Oldfield CJ, Uversky VN, Berezovsky IN, Tawfik DS (2009) Do viral proteins possess
unique biophysical features? *Trends Biochem. Sci.* 34:53-59.

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

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

Wiley D C & Wilson I A, Skehel J J (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 (2007) The speed of adaptation in large asexual populations. *Genetics* 167:2045-2053.

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

Wylie C S, Shakhnovich E I (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, Chen P, Shakhnovich E I (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci USA* 104:16152-16157.
**Figure Captions**

**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, the vertical coordinate \( \Phi \) 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 co-existing 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 out-competed (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 Fig. 4.

**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, \( x_{\text{max}} = \max(x_1, x_2) \). 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 \((1.1, 0.02)\) 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 \( (y_{12} = x_{\text{max}}, \text{dashed line}) \). Analogous data for other polymorphism classes are shown in Fig. S3.

**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: non-synonymous non-epitope, green: non-synonymous epitope). About 70% of all fixations occur in 8 major fixation clusters containing 8 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) Distribution of polymorphism lifetimes between entry and fixation of the new allele. The distributions are similar in all three mutation classes, with average lifetimes between 2.9 years and 3.1 years.

**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 filled circles); cf. Fig. 1(a). 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 filled 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. Fig. 1(b). 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 \( L_{\text{ep}} = 120 \) (epitope sites), \( L_{\text{ne}} = 160 \) (non-epitope sites), mutation rate \( \mu = 5.8 \times 10^{-3} / \text{yr} \), average scaled selection strength \( \bar{\sigma}N = 100 \), selection flip rates \( \gamma = 3.3 \times 10^{-2} / \text{yr} \) (clonal interference) and \( \gamma = 3.6 \times 10^{-3} / \text{yr} \) (episodic sweeps). For model and simulation details, see Supporting Methods. Comparisons with further control models are shown in Fig. S7.

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

Figure 1

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Figure 1}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Figure 2}
\end{figure}
Figure 3
Figure 4
Figure 5