File S1: Supporting Methods

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:

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

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,

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

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