ARTICLE

doi:10.1038/nature13087

A predictive fitness model for influenza

Marta Łuksza[1,2] & Michael Lässig[1]

The seasonal human influenza A/H3N2 virus undergoes rapid evolution, which produces significant year-to-year sequence turnover in the population of circulating strains. Adaptive mutations respond to human immune challenge and occur primarily in antigenic epitopes, the antibody-binding domains of the viral surface protein haemagglutinin. Here we develop a fitness model for haemagglutinin that predicts the evolution of the viral population from one year to the next. Two factors are shown to determine the fitness of a strain: adaptive epitope changes and deleterious mutations outside the epitopes. We infer both fitness components for the strains circulating in a given year, using population-genetic data of all previous strains. From fitness and frequency of each strain, we predict the frequency of its descendent strains in the following year. This fitness model maps the adaptive history of influenza A and suggests a principled method for vaccine selection. Our results call for a more comprehensive epidemiology of influenza and other fast-evolving pathogens that integrates antigenic phenotypes with other viral functions coupled by genetic linkage.

The evolution of influenza A/H3N2 is well documented by sequence data of several thousand strains since 1968[1]. Most of these data contain the gene sequence of haemagglutinin (HA), which covers one of eight segments of the influenza genome and is the primary locus of interaction with the human immune system[2]. Consistent with this functional role, antigenic changes in the HA epitopes carry the adaptive evolution of the pathogen[3–11].

Evolutionary analysis has a particular role for influenza: it serves not only to reconstruct the dynamical process and its causes, but to predict future changes[3,4]. Any prediction of evolution is essentially an estimate of fitness differences between strains. It is these differences that lead to deterministic changes in population frequency, which are predictable if we know how fitness depends on genotype and host environment. Predictability is limited by stochastic events, which range from mutations in individual viral sequences to sampling in host-to-host transmission. Predictions of influenza HA evolution can inform vaccine selection if, despite this limitation, they are sufficiently accurate from one year to the next. Currently, the selection of vaccine strains is based primarily on haemagglutination inhibition assays, which are used to map antigenic changes between viral strains[12]. But the fitness of a strain is a complex phenotype, which integrates antigenic properties with multiple other molecular functions, one of which is simply the thermodynamic stability of proteins[13,14]. Because there is no recombination, the evolution of these functions is strongly coupled, at least within each genomic segment[9,10] (whereas genetic linkage between segments is reduced by reassortment[15]). Here we show that this coupled dynamics can be captured by a fitness model that predicts the evolution of influenza from genomic data.

Clades as units of prediction

Our analysis is based on a sample of 3,944 unique HA coding sequences obtained from influenza A/H3N2 isolates between 1968 and 2012 (ref. 1), partitioned into half-year seasons (Methods). The HA sequences of a given season differ from each other by several epitope and non-epitope nonsynonymous point mutations. To quantify this diversity, we can estimate the population frequencies of mutant alleles at individual RNA sites, of combinations of mutant alleles at two or more sites, and of individual strains. From an epidemiological point of view, the frequency of a strain is simply the fraction of the infected host individuals corresponding to that strain[16]. We infer the genealogy of these strains by an ensemble of trees; see Methods for details of frequency estimation and tree reconstruction. We can then trace the evolution of strain lineages or clades, which are defined as sets of strains descending from a common ancestor (Fig. 1). Whereas strains are typically observed only in a single season, clades have an evolutionary history that extends up to about 5 years and ends with fixation or loss[3]. Clades destined for fixation originate on the so-called trunk of the tree; all other clades are destined

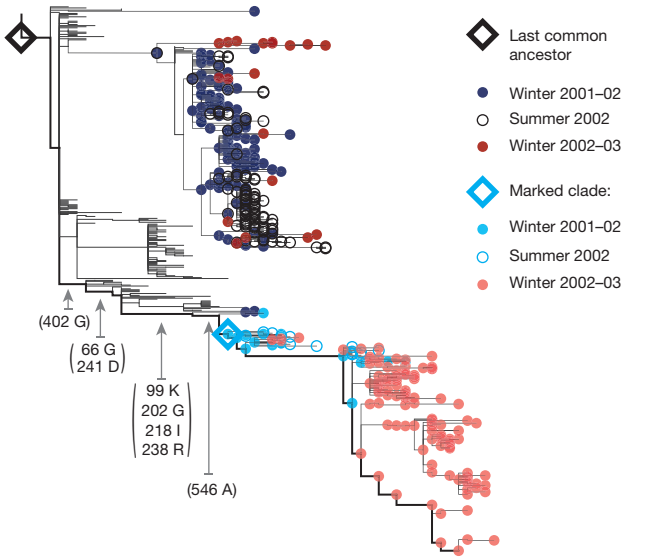


Figure 1 | Evolution of influenza clades. The figure shows a partial influenza strain tree, which is based on strains observed in years 2002 and 2003 (bullets and circles). Each strain i has a frequency x_i in its season's strain population. Our units of prediction are clades, which are defined as sets of strains descending from recent last common ancestors. For one of these clades, we mark its strain content in winter seasons $t = 2002$ and $t + 1$ (light-colour bullets) and its last common ancestor (blue diamond). Each clade is linked by a set of mutations to the last common ancestor of all strains in year t (black diamond); codon position and target amino acid of these mutations are indicated for the marked clade. A clade v observed in season t has a frequency $X_v(t)$, which is the sum of the frequencies of its strains in season t. The marked clade grows substantially from $X_v(t) = 0.08$ to $X_v(t+1) = 0.86$.

[1]Institute for Theoretical Physics, University of Cologne, Zülpicher Strasse 77, 50937 Köln, Germany. [2]Biological Sciences, Columbia University, 607D Fairchild Center, New York, New York 10027, USA.

for loss (Fig. 1). The evolution of these clades is what we want to predict from one year to the next. A successful clade diversifies from its ancestor strain through subsequent mutations during its expansion in the population. At the same time, the same mutation often originates independently in different clades. It is specific combinations of mutations that distinguish each clade from the other coexisting clades. We make predictions for these clades by averaging over the ensemble of equiprobable trees, which minimizes the effects of tree reconstruction ambiguities[3] (Methods).

Our prediction is based on frequency and fitness data that depend only on information actually available at a given point in time. Consider a clade v containing a set of strains i with frequencies x_i in a given season t. The observed frequency of that clade in season t, which is denoted as $X_v(t)$, is simply the sum of these strain frequencies, $X_v(t) = \sum_{i:v,t} x_i$. This sum is defined as an average over strain trees, as detailed in Methods. Each strain has a Malthusian fitness or growth rate f_i (measured in units of 1/year), which is to be specified by our model. Given these initial data, we predict the frequency of that clade in the season 1 year later,

$$\hat{X}_v(t+1) = \sum_{i:v,t} x_i \exp(f_i) \qquad (1)$$

as illustrated in Fig. 1 (for details, see Methods). Equation (1) describes the large-scale population dynamics averaged over many transmission cycles and over the yearly epidemic cycle. We restrict predictions to clades with frequencies $X_v(t) > 0.15$, which are large enough for reliable estimation. These clades are geographically well-mixed[17,18] (93% of them cover two or more continents), whereas smaller clades are dominated by sampling noise and geographical bias (94% are observed on a single continent only). We can check the quality of our method a posteriori by comparing predicted and actual clade evolution, using the observed frequencies $X_v(t+1)$.

Fitness model

Our fitness model has two components, which describe the selection on epitope and non-epitope HA genotypes, respectively. Epitope changes are predominantly under positive selection[7–11], because they affect the antigenic characteristics of a strain. Antigenic selection is contained in multi-strain epidemiological models, which describe a susceptible–infected–recovered (SIR) dynamics[19–23]. In this type of model, host individuals acquire partial immunity against infections with all strains of similar antigenic characteristics. Therefore, the strain growth rates f_i depend on the population history of previously circulating strains. We use an SIR model to derive our minimal epitope fitness model (Methods): a given strain i incurs a cross-immunity load generated by all previous strains j, each of which generates a fitness cost proportional to its frequency x_j and to the cross-immunity amplitude $\mathcal{C}(\mathbf{a}_i,\mathbf{a}_j)$. These amplitudes depend on the antigenic similarity of the strains i and j, which is encoded in the epitope segments of their HA sequences $\mathbf{a}_i$ and $\mathbf{a}_j$. We neglect higher-order antigenic interactions involving more than two strains[19] and the birth–death turnover of the host population, which can be argued to produce only subleading effects in the epidemiology of influenza A/H3N2.

Non-epitope mutations are predominantly under negative selection[9], because they affect protein stability and other conserved molecular functions[13,14]. Here we describe these effects by a simple mutational-load model: each strain incurs a fitness cost $\mathcal{L}(\mathbf{a}_i)$ that is the cumulative effect of recent non-epitope amino acid changes, which occur in its ancestral lineage in the current season (Methods).

Together we obtain a strain fitness of the form

$$f_i = f_0 - \mathcal{L}(\mathbf{a}_i) - \sum_{j:\, t_j < t_i} x_j \mathcal{C}(\mathbf{a}_i,\mathbf{a}_j) \qquad (2)$$

with a constant f_0 ensuring the correct normalization of strain frequencies (Methods). Importantly, this strain-based model goes beyond a fitness model for individual mutations: it counts each new beneficial or deleterious mutation together with the previous changes in its ancestral lineage.

The simplest fitness model of this form has uniform selective effects: each non-epitope mutation generates a fitness cost σ_{ne}, and each epitope mutation reduces the cross-immunity amplitude by an amount σ_{ep}. However, the biology of cross-immunity and protein stability deviates from this model. Both phenotypes are non-uniform and non-linear functions of genetic distance[4,12–14,24]; that is, the effect of a mutation depends on its sequence position, on the amino acids involved, and on its background of previous mutations. Our full fitness model uses nonlinear cross-immunity amplitudes $\mathcal{C}(\mathbf{a}_i,\mathbf{a}_j)$, and it includes position-specific effects and nonlinear fitness terms that are inferred from observed clade histories (Methods; see also a related allele-based inference scheme[10]). Importantly, this model has only four fit parameters, which can be inferred from our data set without compromising predictive power.

Frequency predictions for clades

The winter-to-winter prediction for the Northern Hemisphere obtained from our full, clade-based fitness model is shown in Fig. 2. To determine the accuracy of this model, we compare the predicted frequency ratio, or Wrightian fitness, $\hat{W}_v = \hat{X}_v(t+1)/X_v(t)$ with the posterior observed ratio $W_v = X_v(t+1)/X_v(t)$ for all clades with frequencies $X_v(t) > 0.15$ in a given season. The data points $(W_v, \hat{W}_v)$ in Fig. 2a are distributed around the diagonal of correct prediction; some scatter can be explained by statistical errors in frequency estimation due to tree reconstruction and sampling (Extended Data Fig. 1). As discussed below, these predictions can be improved further by broadening the data basis of our model. The direction of frequency evolution is predicted with remarkable accuracy. There are 121 clades with observed growth ($W_v > 1$), which we predict correctly in 93% of the cases ($\hat{W}_v > 1$). For the 67 clades with observed decline ($W_v < 1$), we correctly predict decline in 76% of the cases. Importantly, the fitness amplitude W_v is predicted accurately for the clades destined for fixation, which have $\hat{X}_v(t+1) \approx 1$ and appear close to the diagonal in Fig. 2a.

The fixation of a clade implies the fixation of all mutations that appear in its ancestor strain. As shown in Fig. 2b, the yearly numbers of nucleotide fixations between 1994 and 2012 are also well predicted by our model. This pattern is well known to be clustered (80% of the nucleotide fixations occur in a subset of 11 years), which reflects recurrent selective sweeps in the evolution of influenza HA[4–7,9].

Figure 2c maps our prediction onto the strain tree. Each clade, represented by its ancestor strain, is coloured according to the maximum of the predicted frequency changes $\hat{X}_v(t+1) - X_v(t)$ over its history. We find clade expansion predominantly close to the trunk and decline far away from the trunk, which is consistent with the observed shape of the influenza tree.

Tagging clades by their point of origination (Methods), allows us to analyse correlations between fitness and geographical location. For clades originating in east and southeast Asia, we predict growth ($\hat{W}_v > 1$) in 77% of the cases, compared to 54% of the cases for clades originating elsewhere; the corresponding fractions with observed growth ($W_v > 1$) are 77% and 49%. This is consistent with the particular role of east and southeast Asia in seeding antigenic variants, which has been established previously[18]. Thus, our analysis captures broad spatial patterns in the evolution of influenza A/H3N2, although the underlying fitness model is geographically neutral (this point will be discussed further below).

We can quantify the statistical information gain due to our prediction by comparing distributions of predicted and posterior next-year frequencies (Methods). We find the observed frequency evolution to be more likely by a factor $>10^{250}$ under our fitness model compared to a null model with constant frequencies (that is, zero fitness) for all strains. We emphasize that our prediction works only from one year to the next, because it cannot predict the new mutations that arise after its base year and shape the course of evolution over longer periods.

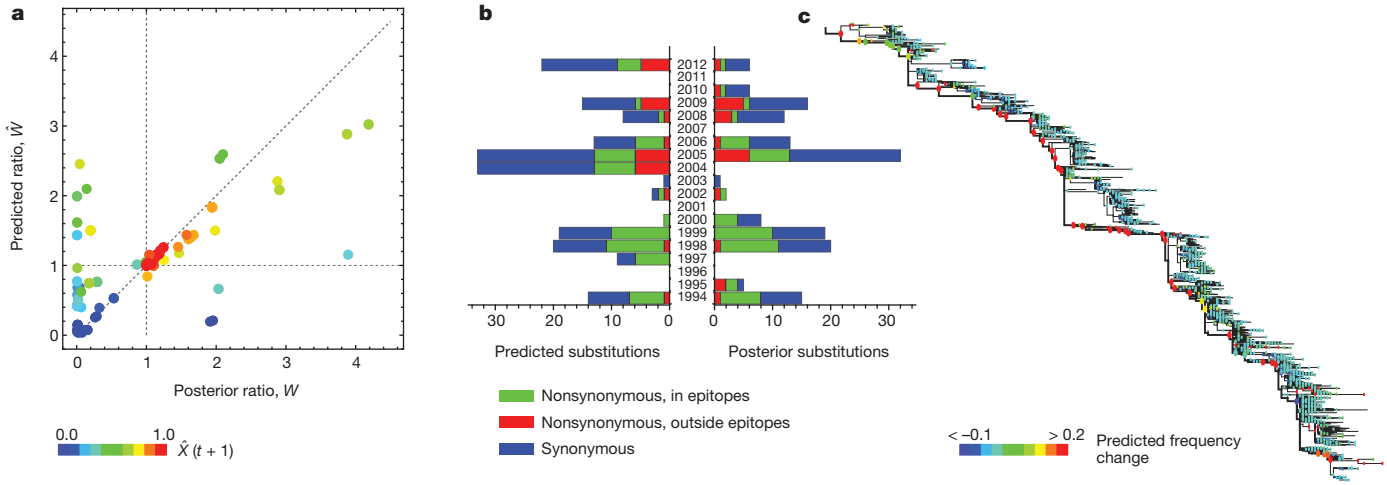


Figure 2 | Year-to-year predictions of HA evolution. **a**, Wrightian fitness: the predicted frequency ratio $\hat{W}_v = \hat{X}_v(t+1)/X_v(t)$ is plotted against the posterior ratio $W_v = X_v(t+1)/X_v(t)$ for 188 influenza HA clades with initial frequency $X_v(t) > 0.15$ observed since 1993 (error bars due to tree reconstruction and sampling are given in Extended Data Fig. 1). The predicted frequency $\hat{X}(t+1)$ is indicated by colour; clades destined for fixation are shown in red. Clade growth ($W_v > 1$) is correctly predicted in 113 of 121 cases, clade decline in 51 of 67 cases. **b**, Yearly numbers of HA nucleotide fixations: predicted numbers are compared to posterior numbers. **c**, Dynamics on the influenza strain tree: for each clade originating between 1993 and 2010, the ancestor node is coloured according to the maximum of the predicted frequency changes, $\max_t[\hat{X}_v(t+1) - X_v(t)]$. Our model correctly predicts expansion along the trunk (thick line) and loss of branches off trunk.

To test our method on a related system, we obtain clade fitness predictions for seasonal influenza A/H1N1. This lineage has re-entered the human population in 1977 and evolved in a way broadly similar to H3N2 until 2009, when the pandemic H1N1 lineage emerged. Compared to the H3N2 data set, the H1N1 strain sample[1,25] has larger regional and seasonal biases, potentially weaker antigenic selection[8], and larger uncertainty about the exact epitope positions[26] (Methods). Our predictions for H1N1 are comparable to H3N2 but somewhat more noisy, as expected from their less informative strain sample (Extended Data Fig. 2). This establishes a proof of principle for the applicability of our model to other influenza strains.

Vaccine strain selection

Our model provides a principled method to select strains for influenza vaccines. By equation (2), vaccination based on a strain v reduces the fitness of each circulating strain i proportionally to the cross-immunity amplitude $\mathcal{C}(\mathbf{a}_i,\mathbf{a}_v)$. This causes a reduction in the total number of infections that is proportional to the average cross-immunity between the vaccine strain and the circulating strains in a given season, $\mathcal{C}_v(t) = \sum_{i:t} x_i \mathcal{C}(\mathbf{a}_i,\mathbf{a}_v)$ (Methods). The optimal vaccine maximizes this reduction, which defines the cross-immunity centre of mass of the circulating strains. Equation (1) predicts next-year cross-immunity amplitudes $\hat{\mathcal{C}}_v(t+1) = \sum_{i:t} x_i \exp(f_i)\mathcal{C}(\mathbf{a}_i,\mathbf{a}_v)$, which can be compared a posteriori with the observed amplitudes $\mathcal{C}_v(t+1)$.

In particular, we can compare the optimal vaccine strains predicted by our model and actual vaccine strains used in the Northern Hemisphere[27] to the posterior centre-of-mass strains observed in the following year (the established procedure of vaccine strain selection is described in Methods). Figure 3 shows this comparison for influenza A/H3N2 in the winter seasons from 1994 to 2012. In all years, the model-selected vaccine strains have a smaller amino acid distance from the cross-immunity centre of mass of the same season than the actual vaccine strains (insert of Fig. 3). This can be explained in part by differences between our sequence-based cross-immunity measure $\mathcal{C}(\mathbf{a}_i,\mathbf{a}_j)$ and the haemagglutination-inhibition-based antigenic distances currently used for vaccine selection. The latter are known to evolve in a more punctuated way[12], but we observe distance differences even in years when vaccine strains have been updated. These results suggest that a fitness-model-based prediction of influenza evolution can contribute to vaccine strain selection; however, we caution against premature conclusions before our prediction scheme is carefully tested with haemagglutination inhibition data. Our model can also be used to estimate how vaccination affects the course of influenza evolution (Methods).

Mapping the adaptive process

The fitness effects underlying our predictions can be displayed in a quantitative map of influenza's adaptive history. As key quantity we use the cumulative fitness flux[28,29], which measures the total amount of adaptation up to a given clade; this quantity is defined in Methods and illustrated in Extended Data Fig. 3. The map of Fig. 4 shows the fitness flux for 234 influenza A/H3N2 clades on a tree between 2003 and 2008 (see Extended Data Fig. 3 for fitness flux over a longer period). It displays clades with multiple different values of fitness and fitness flux in each year. The evolution of this distribution generates a travelling fitness flux wave, which links influenza to recent theoretical models of asexual evolution[30–34]. The advance of the wave is measured by the population mean fitness flux, which is shown as a black dashed line in Fig. 4. This quantity measures correlations between fitness and actual frequency changes of clades. It can be used to compare the predictive power of different fitness models. The best epitope-only fitness model captures about 63%, and the best model with uniform selective effects about 57% of the cumulative fitness flux given by the full model (Extended Data Table 1). An information-theoretic comparison of fitness models shows the same ranking (Methods, Extended Data Table 1). These results indicate that non-epitope changes and nonlinear selective effects have an important role in the adaptive process of influenza and its successful prediction.

The underlying mode of evolution is revealed by individual flux genealogies shown in Fig. 4. We observe that high-fitness clades seed future high-fitness clades by beneficial mutations. In particular, the clades on the trunk of the tree have consistently high predicted fitness values and are driven to fixation by multiple beneficial mutations during their expansion. At the same time, many high-fitness clades are eventually driven to loss, because they are overtaken in fitness by other competing clades. Individual beneficial alleles are lost if they arise in a low-fitness clade or if they are outcompeted by subsequent beneficial mutations in disjoint clades. These observations provide direct evidence of clonal interference in the evolution of influenza[9] with pervasive effects of

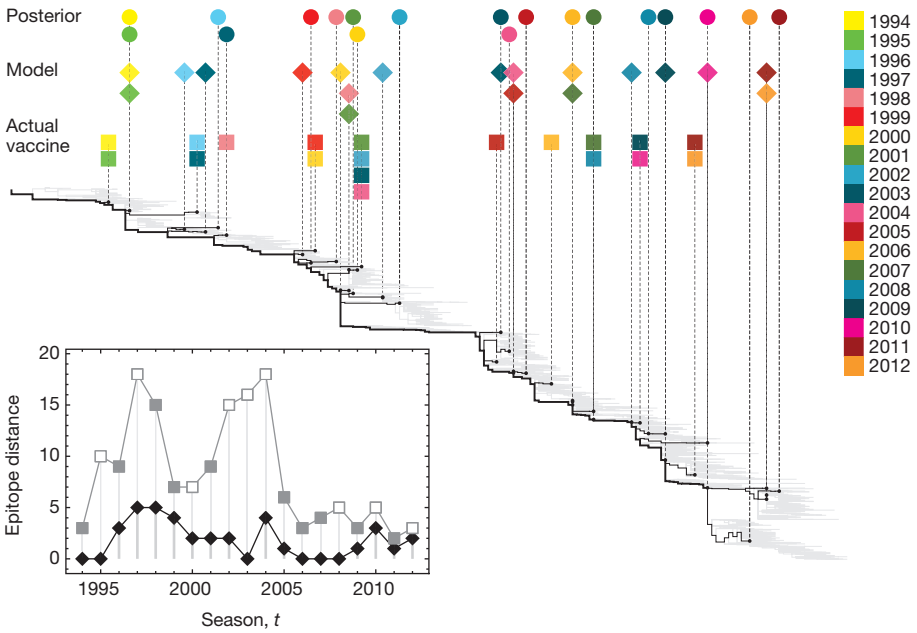


Figure 3 | Vaccine selection. Optimal vaccine strains predicted by our model (diamonds) and actual vaccine strains used in the Northern Hemisphere[27] (squares, listed in Supplementary Information) are compared to posterior cross-immunity centre-of-mass strains (bullets) for the winter seasons from 1994 to 2012. Model predictions are obtained by maximizing the predicted cross-immunity overlap between the vaccine strain and the circulating strains, which amounts to maximizing the predicted reduction of infections (see text and Methods). Insert: yearly epitope amino acid distances of the model-selected vaccine strain (diamonds) and the actual vaccine strain (squares, update years marked by filled squares) to the posterior cross-immunity centre-of-mass strain.

genetic linkage on individual alleles[10]. This mode of evolution is well known from laboratory evolution experiments with microbial and viral populations[35,36].

Clonal interference can explain the observed regional fitness differences between influenza A/H3N2 clades as an effect of multiple beneficial mutations coexisting in a population: individual antigenic mutations originating in east and southeast Asia have the same average effect as mutations originating elsewhere, but they occur in lineages that have accumulated more previous beneficial mutations in their recent past.

Discussion

We have developed a dynamical model that successfully predicts the year-to-year evolution of individual influenza clades, based on epitope and non-epitope characteristics of their HA gene. Our general model is applicable whenever host–pathogen interactions—in particular, antigenic selection—generate continual adaptive evolution of a predominantly asexual population. Our results highlight the determinants of predictive power: we need sufficient information on the genotypic and phenotypic basis of antigenic and mutational-load fitness components, and model training requires a sufficiently deep and unbiased strain sample. This suggests that predictions can be improved by integrating diverse genotypic and phenotypic data, which include free-energy effects of specific mutations[13], haemagglutination inhibition data[12], the genomics of neuraminidase[8] and the geographical distribution of strains[18]. Furthermore, the prediction scheme can be extended from population frequencies to absolute growth rates and population numbers, which includes the dynamics of yearly incidence rates[23]. Together, we expect an improved understanding of selective effects for specific mutations from limited strain data. This is key to evolutionary predictions for other influenza variants, including the potentially pandemic avian A/H5N1 and A/H7N9 lineages.

In a broader context, our model establishes a direct link between population genetics and epidemiology that is to be explored more comprehensively in future work. This link is the strain-specific fitness function of equation (2), which governs the dynamics of infected host individuals in an SIR framework. Strain fitness depends not only on antigenic characteristics, but also on other phenotypes encoded in genetically linked sequence. We expect that this coupling between antigenic adaptation and conservation of other functions is not limited to influenza, but is a generic feature of fast-adapting pathogens. Therefore, the epidemiology of such systems should be based on the ensemble of phenotypes linked to the adaptive process.

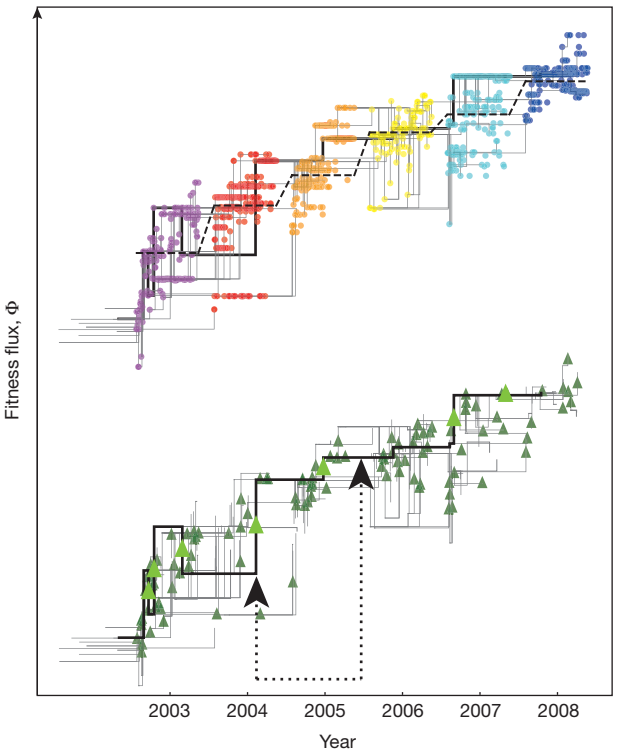


Figure 4 | Adaptation map of influenza. The fitness flux $\Phi_v(t)$, computed from the fitness model (2) and observed frequency changes, is shown for 234 clades on a tree between 2003 and 2008; see Methods for the definition and Extended Data Fig. 3a for an illustration of fitness flux. Top graph: strains within these clades are ordered by year and, within each year, by mutational distance to the last common ancestor. The mean cumulative fitness flux $\Phi(t)$ is shown as dashed line; see also Extended Data Fig. 3b. This map displays a travelling fitness flux wave. Bottom graph: the same map is shown with nonsynonymous epitope mutations marked by green triangles; these mutations are mostly beneficial[7,9,10]. This gives evidence of clonal interference: successful clades are driven to fixation by multiple beneficial mutations (large green triangles; origination and fixation of one such clade are marked by arrows), whereas other beneficial mutations are driven to loss (small green triangles).

Beyond pathogens, this work touches upon the fundamental question of how predictable evolution is. Although there is clearly no general answer to this question, our analysis shows under what auspices limited predictions may be successful.

METHODS SUMMARY

We partition our strain sample into seasons; $t = y$ labels the period from October of year $y - 1$ to April of year y. For the prediction from season t to $t + 1$, we use maximum-likelihood HA sequence trees of all strains up to season t; validation is based on trees for the full period. Epitope and non-epitope mutations are mapped onto the branches of these trees (Extended Data Fig. 4). Clades are defined as sets of all descendants of a given HA sequence $\mathbf{a}_\nu$; clade frequencies are estimated by averaging over equiprobable trees.

Our fitness model has three components. The epitope fitness component $f_i^{\mathrm{ep}} = \sigma_{\mathrm{ep}} \sum_{j:t_j<t_i} x_j c\big(D_{\mathrm{ep}}(\mathbf{a}_i,\mathbf{a}_j)\big)$ is computed from linear amino acid distances D_{ep} between pairs of sequences in epitope codons[7], using a nonlinear cross-immunity amplitude $c(D_{\mathrm{ep}}) = \exp(-D_{\mathrm{ep}}/D_0)$. The non-epitope fitness component $f_i^{\mathrm{ne}} = -\sigma_{\mathrm{ne}} D_{\mathrm{ne}}(\mathbf{a}_i,\mathbf{a}_i^*)$ depends on the non-epitope amino acid distance D_{ne} between $\mathbf{a}_i$ and the last ancestor of strain i in a previous season, $\mathbf{a}_i^*$. We show that these recent non-epitope mutations are under substantial negative selection[9] (Extended Data Fig. 4). The full fitness model (equation (2)) exploits an additional feature of the population history: synonymous mutations hitchhiking in selective sweeps[9] reinforce the inference of positive selection by a term $f_i^{\mathrm{nl}} = \lambda D_0^{\mathrm{ave}}(\nu(i),\mathbf{a}^*(t))$ proportional to the average neutral mutational distance of all strains in clade $\nu(i)$ from the last common ancestor $\mathbf{a}^*(t)$. The full model takes the form of equation (2), $f_i = f_i^{\mathrm{ep}} + f_i^{\mathrm{ne}} + f_i^{\mathrm{nl}} + f_0$, with a constant f_0 given by the normalization conditions $\sum_{i:\,t} x_i = \sum_{i:\,t} x_i \exp(f_i) = 1$.

To quantify the quality of clade frequency predictions, we partition the strains of each season t into non-overlapping clades $\nu \in K(t)$ with frequencies $\mathbf{Y}_t = (Y_\nu(t))_{\nu\in K(t)}$. We define the relative information of observed versus predicted next-year frequencies, $H\big(\mathbf{Y}_{t+1}|\hat{\mathbf{Y}}_{t+1}\big) = \sum_{\nu\in K(t)} Y_\nu(t+1)\log\big[Y_\nu(t+1)/\hat{Y}_\nu(t)\big]$, which determines the likelihood p that the observed strains are a sample of the predicted frequency distributions. We use this likelihood to rank fitness model variants, which include alternative epitopes and glycosylation effects, and to infer optimal parameters from strain data of 4–8 years before the base year of prediction (Extended Data Table 1 and Extended Data Fig. 6).

We define the cumulative fitness flux for each non-overlapping clade $\nu \in K(t)$,

$$\Phi_\nu(t) = F_\nu(t) - \bar{F}(t) + \sum_{t'=t_0}^{t-1} \sum_{\nu\in K(t')} [F_\nu(t') - \bar{F}(t')][Y_\nu(t'+1) - Y_\nu(t')]$$

where $F_\nu(t) = \log\big[\hat{Y}_\nu(t+1)/Y_\nu(t)\big]$ is the fitness of clade ν and $\bar{F}(t) = \sum_{\nu\in K(t)} F_\nu(t) Y_\nu(t)$ is the mean population fitness in year t. The mean cumulative flux $\bar{\Phi}(t) = \sum_{\nu\in K(t-1)} \Phi_\nu(t-1) Y_\nu(t)$ is a measure of adaptation satisfying the fitness flux theorem[29]. Deviations from this mean reflect fitness differences between clades, $\Phi_\nu(t) - \bar{\Phi}(t) = F_\nu(t) - \bar{F}(t)$.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 April 2013; accepted 29 January 2014.
Published online 26 February 2014.

1. Bao, Y. *et al.* The Influenza Virus Resource at the National Center for Biotechnology Information. *J. Virol.* **82,** 596–601 (2008).
2. Wiley, D. C., Wilson, I. A. & Skehel, J. J. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* **289,** 373–378 (1981).
3. Bush, R. M., Bender, C. A., Subbarao, K., Fox, N. J. & Fitch, W. M. Predicting the evolution of human influenza A. *Science* **286,** 1921–1925 (1999).
4. Plotkin, J. B., Dushoff, J. & Levin, S. A. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proc. Natl Acad. Sci. USA* **99,** 6263–6268 (2002).
5. Koelle, K., Cobey, S., Grenfell, B. & Pascual, M. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science* **314,** 1898–1903 (2006).
6. Wolf, Y. I., Viboud, C., Holmes, E. C., Koonin, E. V. & Lipman, D. J. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biol. Direct* **1,** 34 (2006).
7. Shih, A. C.-C., Hsiao, T.-C., Ho, M.-S. & Li, W.-H. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc. Natl Acad. Sci. USA* **104,** 6283–6288 (2007).
8. Bhatt, S., Holmes, E. C. & Pybus, O. G. The genomic rate of molecular adaptation of the human influenza A virus. *Mol. Biol. Evol.* **28,** 2443–2451 (2011).
9. Strelkowa, N. & Lässig, M. Clonal interference in the evolution of influenza. *Genetics* **192,** 671–682 (2012).
10. Illingworth, C. J. R. & Mustonen, V. Components of selection in the evolution of the influence virus: linkage effects beat inherent selection. *PLoS Pathog.* **8,** e1003091 (2012).
11. Meyer, A. G., Dawson, E. T. & Wilke, C. O. Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin. *Phil. Trans. R. Soc. B* **368,** 20120334 (2013).
12. Smith, D. J. *et al.* Mapping the antigenic and genetic evolution of Influenza virus. *Science* **305,** 371–376 (2004).
13. Bloom, J. D. & Glassman, M. J. Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. *PLOS Comput. Biol.* **5,** e1000349 (2009).
14. Wylie, C. S. & Shakhnovich, E. I. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc. Natl Acad. Sci. USA* **108,** 9916–9921 (2011).
15. Holmes, E. C. *et al.* Whole-genome analysis of human inuenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol.* **3,** e300 (2005).
16. Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303,** 327–332 (2004).
17. Rambaut, A. *et al.* The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453,** 615–619 (2008).
18. Russell, C. A. *et al.* The global circulation of seasonal influenza A (H3N2) viruses. *Science* **320,** 340–346 (2008).
19. Gog, J. R. & Grenfell, B. T. Dynamics and selection of many-strain pathogens. *Proc. Natl Acad. Sci. USA* **99,** 17209–17214 (2002).
20. Tria, F., Lässig, M., Peliti, L. & Franz, S. A minimal stochastic model for influenza evolution. *J. Stat. Mech.* P07008 (2005).
21. Kryazhimskiy, S., Dieckmann, U., Levin, S. A. & Dushoff, J. On state-space reduction in multi-strain pathogen models, with an application to antigenic drift in influenza A. *PLOS Comput. Biol.* **3,** e159 (2007).
22. Minayev, P. & Ferguson, N. Improving the realism of deterministic multi-strain models: implications for modelling influenza A. *J. R. Soc. Interface* **6,** 509–518 (2009).
23. Rasmussen, D. A., Ratmann, O. & Koelle, K. Inference for nonlinear epidemiological models using genealogies and time series. *PLOS Comput. Biol.* **7,** e1002136 (2011).
24. Kryazhimskiy, S., Dushoff, J., Bazykin, G. A. & Plotkin, J. B. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet.* **7,** e1001301 (2011).
25. Bogner, P. *et al.* A global initiative on sharing avian flu data. *Nature* **442,** 981 (2006).
26. Huang, J.-W., Lin, W.-F. & Yang, J.-M. Antigenic sites of H1N1 influenza virus hemagglutinin revealed by natural isolates and inhibition assays. *Vaccine* **30,** 6327–6337 (2012).
27. WHO Recommendations for Influenza Vaccine Composition. Retrieved from http://www.who.int/influenza/vaccines/vaccinerecommendations1/en.
28. Mustonen, V. & Lässig, M. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet.* **25,** 111–119 (2009).
29. Mustonen, V. & Lässig, M. Fitness flux and ubiquity of adaptive evolution. *Proc. Natl Acad. Sci. USA* **107,** 4248–4253 (2010).
30. Desai, M. M. & Fisher, D. S. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* **176,** 1759–1798 (2007).
31. Rouzine, I. M., Brunet, E. & Wilke, C. O. The traveling-wave approach to asexual evolution: Mullers ratchet and speed of adaptation. *Theor. Popul. Biol.* **73,** 24–46 (2008).
32. Schiffels, S., Szollosi, G. J., Mustonen, V. & Lässig, M. Emergent neutrality in adaptive asexual evolution. *Genetics* **189,** 1361–1375 (2011).
33. Good, B. H., Rouzine, I. M., Balick, D. J., Hallatschek, O. & Desai, M. M. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc. Natl Acad. Sci. USA* **109,** 4950–4955 (2012).
34. Neher, R. A. & Hallatschek, O. Genealogies of rapidly adapting populations. *Proc. Natl Acad. Sci. USA* **110,** 437–442 (2013).
35. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102–103,** 127–144 (1998).
36. Miralles, R., Gerrish, P. J., Moya, A. & Elena, S. F. Clonal interference and the evolution of RNA viruses. *Science* **285,** 1745–1747 (1999).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge discussions with B. D. Greenbaum, B. Grenfell, C. Illingworth, A. Levine, J. W. McCauley, V. Mustonen, S. Pompei and R. Rabadan. This work has been partially supported by Deutsche Forschungsgemeinschaft grant SFB 680 and by German Federal Ministry of Education and Research grant 0315893-Sybacol. Part of this work was performed at the Kavli Institute of Theoretical Physics (Santa Barbara), which has been supported by National Science Foundation grant PHY05-51164.

Author Contributions Both authors designed research, developed methods, analysed data, interpreted results and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M. Lässig (lassig@thp.uni-koeln.de).

METHODS

1. Strain trees and clades. In this section, we first discuss the structure of our sequence data set and the inference of genealogical trees from these data. We then define the key notion of sequence clades, which are the units of our prediction scheme, and we detail the procedure to estimate clade frequencies.

Sequence data set. Our study is based on a data set of 5,306 human influenza A/H3N2 sequences available from the NCBI database[1], which contain 3,944 unique HA strains (as defined below). We include only sequences that contain the full HA domain with at most 5 nucleotides missing and are annotated by time and location of observation. We exclude 17 sequences that are clearly recognizable as outliers in the reconstructed strain trees. The GenBank accession numbers of all strains used in this study are provided as Supplementary Data.

The available influenza sequences are clearly not a randomly sampled data set, which would be ideal for population-genetic analysis. Known systematic biases of the sequence data include: (a) yearly variations in sampling depth. Far fewer sequences are available for earlier years (1968–1992) than for later years (1993–2012); (b) regional variations in sampling depth. For example, the New York sequence project[37] leads to an over-representation of US sequences in some years; (c) passage history effects. Egg-cultured strains show additional mutations, which are not present in the wild-type sequences[38].

To ensure a sufficient sampling depth, we restrict our predictions to base years t starting from winter 1992–93 in the Northern Hemisphere. To reduce systematic biases, lab strains and marked egg isolates are excluded, and strains with identical sequences, location and year of observations are counted only once. Furthermore, we verify that partial exclusion of data from local projects does not affect our results.

Estimation of strain frequencies. We partition the period of our data set into half-year seasons labelled by an index t. The Northern winter from October of year $y-1$ to April of year y is season $t=y$, the Southern winter from April to September of year t is season $t=y+\frac{1}{2}$. We refer to a unique HA genotype tagged by its season of observation, $(\mathbf{a}, t)$, as a HA strain. Most sequences of the data set are annotated by their month of observation; for some sequences, their season has been retrieved from the literature. Identical HA sequences in our data set are counted as one strain if they are observed in the same season and as different strains otherwise. We label all observed strains by an index i; the index set of all strains observed in a given season t is denoted by $\mathcal{S}(t)$. Each strain has a multiplicity m_i in our data set, which counts the number of sequences with HA genotype $\mathbf{a}_i$ and season of observation t_i (sequences observed in the same season and in the same location are counted only once). The strain frequency is then defined as

$$x_i = \frac{m_i}{m(t)} \tag{3}$$

where $m(t)=\sum_{i\in\mathcal{S}(t)} m_i$ is the total number of sequences observed in season t. These frequencies are to be interpreted as fractions of the infected host population corresponding to a particular strain. Clearly, the frequencies of individual strains are dominated by sampling noise and regional bias. Hence, we predict frequencies for clades, which are defined below. Clades are larger population units, and their frequencies can be estimated more reliably. The robustness of our predictions with respect to regional sampling bias is further discussed in section 4 below.

Strain tree reconstruction. Our analysis is based on an ensemble of strain trees obtained from the HA sequence data set. Such trees describe the genealogy of influenza strains resulting from an evolutionary process under selection[9]. The tree ensemble is obtained with FastTree[39], which very time-efficiently reconstructs maximum-likelihood phylogenies. We use a general time-reversible model. We further use PAUP[40] to reconstruct maximum-likelihood sequences of internal nodes, given the FastTree output topology. Trees are rooted using the strain A/Albany/17/1968 (GenBank accession number ABO52357). Both FastTree and PAUP impose that the reconstructed trees are binary and resolve higher-degree nodes by arbitrary subtrees. We process this output to restore higher-degree nodes, which reduces the arbitrariness of tree reconstruction.

In these trees, each observed strain is mapped onto a unique node. The remaining (internal) nodes represent unobserved HA genotypes inferred by maximum likelihood. Each tree node is assigned a season of occurrence: nodes representing observed HA strains are assigned their season of observation; nodes representing inferred HA genotypes are assigned the season of their oldest observed descendant. Similarly, each node is assigned a geographical region of occurrence (east and southeast Asia, remainder of Asia, Europe, Africa, North America, South America, Oceania). Nodes representing observed HA strains are assigned their reported location, nodes representing inferred HA genotypes are assigned the location of the observed same-season descendant that has minimal mutational distance.

Tree statistics. To reduce fluctuations in tree reconstruction, our analysis is based on averages over an ensemble of K maximum-likelihood trees, which are labelled by an index $\kappa = 1,\ldots, n$. These trees come from 10 independent runs of FastTree, followed by PAUP (these runs differ in the order of sequences added). We obtain typical values of n ranging from 10 to 30. Variation between maximum-likelihood trees occurs only on peripheral branches; the large-scale tree structure and the tree-based statistical observables are well conserved. We construct separate tree ensembles for the strains up to a given season t and for the full period of the data set, as detailed below.

Mapping of mutations. Maximum likelihood 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, that is, the appearance of a nucleotide difference between the strains descending from that branch and its ancestral lineage. Extended Data Fig. 4 shows an example of a reconstructed strain tree with all mapped mutations, which are partitioned into three classes: (a) synonymous mutations, (b) nonsynonymous HA epitope mutations and (c) nonsynonymous mutations outside the antigenic HA epitopes (see section 2). These mutations are the basis of our fitness model.

Definition of clades. Consider an ensemble of sequence trees constructed from all strains up to season t. Each of these trees contains a subtree that links the strains $i\in\mathcal{S}(t)$ to their last common ancestor sequence (Fig. 1). We collect all HA genotypes that occur in at least one of these subtrees in a set $\bar{K}(t)$ and label the elements of this set by an index v. For each genotype $\mathbf{a}_v$ ($v\in\bar{K}(t)$), we construct its clade as the set of all strains that are descendants of $\mathbf{a}_v$. This notion is intuitive and well-defined for a given tree (Fig. 1). Because we do not know the true genealogy of strains, we use the following probabilistic definition of clades based on the ensemble of inferred trees: (a) We refer all HA genotypes $\mathbf{a}_v$ ($v\in\bar{K}(t)$) to the last common ancestor sequence $\mathbf{a}^*(t)$ of the tree ensemble. Each genotype $\mathbf{a}_v$ is then in one-to-one correspondence with a mutant haplotype, which is defined as the minimal set of mutations (each specified by its sequence position and its mutant allele) from $\mathbf{a}^*(t)$ to $\mathbf{a}_v$. An example of a mutant haplotype is shown in Fig. 1. (b) In a given tree κ, a given mutant haplotype defines a unique minimal set of disjoint subtrees, $\mathcal{T}_v^{(\kappa)}$, that contains all strains carrying the mutant alleles of $\mathbf{a}_v$. In most cases, this set consists of a single subtree. (c) The subtree sets $\mathcal{T}_v^{(1)},\ldots,\mathcal{T}_v^{(n)}$ define an ensemble of descendant strain sets in season t,

$$\mathcal{S}_v^{(1)}\equiv\mathcal{T}_v^{(1)}\cap\mathcal{S}(t),\ldots,\mathcal{S}_v^{(n)}\equiv\mathcal{T}_v^{(n)}\cap\mathcal{S}(t)\quad\text{for } v\in\bar{K}(t) \tag{4}$$

(d) This ensemble defines the probability that strain i is contained in clade v,

$$\rho_{i,v}=\frac{1}{n}\sum_{\kappa=1}^{n}\epsilon_{i,v}^{(\kappa)}\quad\text{with }\epsilon_{i,v}^{(\kappa)}=\begin{cases}1 & \text{if } i\in\mathcal{S}_v^{(\kappa)}\\ 0 & \text{otherwise}\end{cases} \tag{5}$$

Estimation of clade frequencies. Given the strain frequencies (equation (3)) and the probabilities (equation (5)), it is straightforward to evaluate the frequency of a clade v in the population of season t,

$$X_v(t)=\sum_{i\in\mathcal{S}(t)}\rho_{i,v}x_i\equiv\sum_{i:v,t}x_i\quad\text{for } v\in\bar{K}(t) \tag{6}$$

which defines the abbreviated notation used in the main text. On the basis of the frequencies (equation (6)), we define a smaller set $K(t)\subset\bar{K}(t)$, which consists of all clades $v\in\bar{K}(t)$ with $X_v(t)>0.15$. As explained in the main text, these clades are used for frequency predictions, because they are geographically well-mixed[17,18] and less affected by sampling noise than smaller clades. Details of the prediction method are given in section 3.

2. Fitness models. The fitness models used in this study specify the growth rates f_i of individual influenza strains based on their HA genotype. These rates govern the large-scale dynamics of population frequencies, as given by equation (1). Our full, haplotype-based fitness model has two components: immune selection acting on mutations in antigenic HA epitopes, and negative selection acting on mutations outside the epitopes. For both fitness components, we discuss their genetic basis, as well as their general functional form and its biological interpretation. The inference of selection from our data set requires a parametrization of the fitness model in terms of few selection coefficients. To this end, we partition the model into epitope and non-epitope components based on linear genetic distances and an additional nonlinear component, which is inferred for clades.

Identification of antigenic codons. The antigenic epitopes of HA are influenza's primary locus of interaction with the human immune system. There are five known epitopes (labelled A–E) containing a total of 62 amino acids[2,3,41,42]. The set of epitope codons includes known sialic acid receptor binding sites[41,43]. Over the last decades, the receptor binding properties of influenza A/H3N2 have evolved towards reduced binding of both avian and human receptors[44].

Amino acid changes in epitope codons are known to be under predominantly positive selection[7–9,11]. However, positive selection is not limited to changes in

these codons[7,11], and it can be correlated with specific structural features of the HA protein[11]. To quantify selection on a class of mutations, we use the frequency propagator ratio[9]

$$g(X)=\frac{G(X)}{G_0(X)} \tag{7}$$

where $G(X)$ is the likelihood that a mutation in a given class reaches frequency X, and $G_0(X)$ is the corresponding likelihood for synonymous mutations, which evolve near neutrality.

Our analysis is based on a set of codons in epitopes A–E reported by Shih *et al.*[7], which aggregates the results of previous studies[2,3,41,42]. As shown in Extended Data Fig. 5, amino acid changes in these codons have significantly increased propagator ratios, which is indicative of substantial positive selection[9]. Increased propagator ratios are also observed for the subset of mutations in epitopic receptor binding sites. However, as detailed below, the predictive contributions are heterogeneous: epitopes A–D are found to be informative, epitope E not. This is consistent with selective differences between epitopes: epitope E has a reduced propagator ratio compared to the other epitopes (Extended Data Fig. 5); the ratio between non-synonymous and synonymous changes per codon (dN/dS) reported in Meyer *et al.*[11] has an average 0.47 in epitopes A–D, and 0.18 in epitope E. Hence, our predictions in the main text use a set of 49 codons in epitopes A–D.

To quantify the dependence of our method on the choice of antigenic sequence sites, we compare the predictions from this set of epitope codons with those from a number of alternative choices. These include (a) a subset of epitope codons excluding receptor binding sites. This set gives a poorer prediction than the full set, which is consistent with the role of receptor binding in the adaptive process of influenza[43,44]; (b) a subset of epitope codons reported to be under positive selection by Bush *et al.*[3] This set gives a significantly poorer prediction than the full set; (c) an extended set including the 13 additional codons with the highest dN/dS values according to Meyer *et al.*[11]. The extended set also produces a prediction of reduced quality. A possible explanation is that the additional codons contain compensatory mutations, whose positive fitness effects depend on previous changes in epitope codons. Because epistatic interactions are not explicitly included in the model, including additional codons may generate noise and reduce the predictive power of the additional codons. In our strain-based fitness model, epistasis is taken into account in a more summaric way by a nonlinear fitness component that will be introduced below; (d) a null model of 49 codons randomly chosen from HA1 domain, which is evaluated by averaging over 10 independent codon sets. As expected, this model gives no significant prediction.

Our comparative analysis of fitness models is detailed in section 4; the information-theoretic ranking of these model variants is given in Extended Data Table 1.

Antigenic selection model. Antigenic selection is linked to the epidemiology of influenza: after infection, host individuals acquire long-term immunity against all strains with sufficiently similar antigenic characteristics. This effect curbs the fitness of existing strains and favours antigenic innovation. Here we use the minimal antigenic fitness model

$$f_i^{\mathrm{agen}}=f_0-\sum_{j:t_j<t_i} x_j\mathcal{C}(\mathbf{a}_i,\mathbf{a}_j) \tag{8}$$

In this model, a given strain i incurs a fitness cost caused by cross-immunity interactions with all strains j occurring at times $t_j < t_i$; we refer to this cost as cross-immunity load. Each strain j generates a load component that is proportional to its frequency x_j and to the cross-immunity amplitude $\mathcal{C}(\mathbf{a}_i,\mathbf{a}_j)$. Hence, the fitness f_i^{agen} depends on the frequency history of the population up to the time t_i, which can be estimated from our data set (see section 1). The cross-immunity amplitudes $\mathcal{C}(\mathbf{a}_i,\mathbf{a}_j)$ depend on the antigenic similarity between strains i and j, which is encoded in their HA sequences $\mathbf{a}_i$ and $\mathbf{a}_j$; the inference of these amplitudes from our data set will be discussed below. The constant f_0 is determined by the normalization of strain frequencies, as given by equation (19) below. This model describes continual antigenic adaptation. According to equation (8), the fitness of any given genotype is a monotonically decreasing function of time. Successful strains have fitness values $f_i^{\mathrm{agen}}>0$, but they will always cross over to negative fitness and be replaced by mutant strains with reduced cross-immunity load.

A generic antigenic fitness model of the form (8) contains multiple nonlinearities, which involve the dependence of antigenic interaction phenotypes[12] on HA sequences and the dependence of fitness on these phenotypes. To infer antigenic selection from our data set, we write the antigenic fitness component (8) in a specific form,

$$f_i^{\mathrm{ep}}=f_0-\sigma_{\mathrm{ep}}\sum_{j:t_j<t_i} x_j c\big(D_{\mathrm{ep}}(\mathbf{a}_i,\mathbf{a}_j)\big) \tag{9}$$

This form depends on a linear amino acid distance between pairs of HA sequences, $D_{\mathrm{ep}}(\mathbf{a}_i,\ \mathbf{a}_j)$, which counts amino acid differences in epitopes A–D. The fitness component f_{ep} retains an essential nonlinearity in the cross-immunity amplitude $c(D_{\mathrm{ep}})$. We use an exponentially decreasing function,

$$c\big(D_{\mathrm{ep}}\big)=-\exp\big(-D_{\mathrm{ep}}/D_0\big) \tag{10}$$

This form reflects a property of influenza A/H3N2 that is well-known from haemagglutination inhibition data[12]: cross-immunity between strains decays with a characteristic time of a few years. Together, the fitness component f_i^{ep} depends on the two parameters σ_{ep} and D_0. The inference and optimization of these parameters are discussed in section 4.

Epidemiological basis. The dynamics of host populations can be described by a multi-strain SIR model with cross-immunity interactions[19,45,46]. This well-known class of epidemiological models specifies the dynamics of susceptible, infected and recovered host populations that interact with multiple antigenically related pathogen strains. Our antigenic fitness model (equation (9)) can be derived from a status-based[46] multi-strain SIR dynamics similar to the model of Gog and Grenfell[19],

$$\frac{dS_i}{dt}=-\nu\sum_j c_{ij}I_j \tag{11}$$

$$\frac{dI_i}{dt}=\frac{\beta}{N}S_iI_i-\nu I_i \tag{12}$$

$$\frac{dR_i}{dt}=\nu I_i \tag{13}$$

Equations (12) and (13) describe the dynamics of infections: a host infected with strain i generates $\beta S_i/N$ new infected hosts per unit time. This rate is proportional to the transmission rate β and to the density of hosts susceptible to that strain, S_i/N, where N is the host population size. Infected hosts recover at a rate ν. Equation (11) describes the dynamics of cross-immunity: infection with and recovery from strain j confer immunity—that is, loss of susceptibility—to strain i with probability c_{ij}. These probabilities are maximal for identical strains, $c_{ii}=1$, and decrease with increasing antigenic distance between the strains i and j. Integrating equations (11) and (13) with the initial conditions $S_i=N$ and $R_i=0$ produces an infection dynamics of the form

$$\frac{dI_i}{dt}=\left[\frac{\beta}{N}\left(1-\nu\sum_j c_{ij}R_j\right)-\nu\right]I_i\equiv F_iI_i \tag{14}$$

which defines the growth rates or absolute fitness values F_i of individual strains. For any given strain, R_i is a monotonically increasing function, and F_i is a monotonically decreasing function of time. Next, we coarse-grain the dynamics (equation (14)) to discrete seasons and approximate the recovered populations $R_j(t)$ as the total infected populations in previous seasons: $R_j(t)=I_j(t_j)$ if $t_j<t$ and $R_j(t)=0$ otherwise. Here we use our convention that an individual strain j occurs only in a single season t_j (strains found in more than one year are counted with different labels j; see section 1). Equation (14) then predicts next-year strain populations

$$I_i(t_i+1)=I_i(t_i)\exp(F_i) \tag{15}$$

where F_i is measured in units of 1/year. Finally, we change to frequency variables $x_i=I_i/I_{\mathrm{tot}}$, where I_{tot} is the total infected host population in a given year. This leads to a strain frequency dynamics $\hat{x}_i=x_i\exp(f_i)$ as in equation (1), with an antigenic fitness component $f_i=F_i-\log\sum_{i:t}x_i\exp(F_i)$ satisfying the normalization condition (19). This component is of the form (9) with the rescaled cross-immunity coefficients $\mathcal{C}(\mathbf{a}_i,\mathbf{a}_j)=(\beta\nu I_{\mathrm{tot}}/N)c_{ij}$.

Compared to general multi-strain SIR models, our minimal model contains a number of simplifications (related model simplifications have been discussed in the recent literature[19–23,46,47]): (a) The model is truncated to cross-immunity interactions between pairs of strains[19]; higher-order interactions are neglected. This is justified for influenza, because the decay time of cross-immunity (around 5 years[12]) is shorter than the average interval between infections of an individual host (above 10 years, as estimated from yearly incidence rates[48]). (b) Similarly, the turnover of the host population by birth and death is neglected, because the decay time of cross-immunity is short compared to the average human lifetime. (c) The model equates the cross-immunity load to a sum over frequencies of hosts infected in previous seasons. Neglecting the infected and recovered populations from the current season produces only a small error, because the cross-immunity load builds up over several seasons. An alternative approximation, $R_j(t)=I_j(t)$ if $t_j\leq t$ and $R_j(t)=0$ otherwise, can be shown to produce equivalent results. (d) The model is reduced from absolute population numbers to population frequencies, and we neglect year-to-year variations of the total number of infections, I_{tot}. This reflects our primary interest in predicting the relative prevalence of strains, which depends mainly on the frequency history of the population. However, it is straightforward

to extend our prediction framework to absolute population numbers and absolute growth rates. This extension requires estimating incidence numbers for specific strains and seasons.

These reductions in model complexity are important, because they make the minimal model inferable from data (see below).

Identification of informative non-epitope mutations. Amino acid changes outside the antigenic epitopes are predominantly under negative selection[9], because they affect protein stability and other conserved molecular functions[13,14]. The level of nonsynonymous sequence diversity, as well as strength and direction of selection, are heterogeneous across the HA sequence. About 36% of the non-epitope codons show no diversity throughout the period of the data set, which is indicative of strong purifying selection. Our fitness model is based on observed amino acid changes in the remaining codons. These changes are in part deleterious[9]; that is, they carry a mutational load that is to be included in our non-epitope fitness model. Other non-epitope changes are approximately neutral, have compensatory fitness effects or may contribute to the adaptive process[11,13,49]. Including such effects would require a comprehensive biophysical model of HA protein structure, which is beyond the scope of this paper. To delineate a set of deleterious mutations, we use the following heuristic: a non-epitope nonsynonymous mutation is counted as informative for the fitness model in the season of its origination, but as non-informative in all subsequent seasons until fixation or loss. Hence, the informative non-epitope mutations of a strain i are the differences between its sequence $\mathbf{a}_i$ and that of its closest ancestor from a previous season, which we denote by $\mathbf{a}_i^*$.

To show differential selection on informative and non-informative mutations, we use the frequency propagator ratio (equation (7)). This a posteriori analysis is reported in Extended Data Fig. 5. It indicates that the above definition of informative non-epitope mutations serves its purpose: negative selection occurs predominantly in the class of informative mutations, whereas non-informative mutations have a small average level of selection. In other words, if a non-epitope mutation has been observed already in the previous season, it is less likely to be deleterious (because many deleterious alleles occur only in a single season). This does not imply that any given mutation changes its selective effect between the year of origination and the subsequent years.

Mutational load of non-epitope mutations. We describe selection on non-epitope HA sequence by a mutational-load model,

$$f_i^{\mathrm{load}} = -\mathcal{L}(\mathbf{a}_i) \tag{16}$$

In this model, a given strain i from season t incurs a fitness cost that is the cumulative effect of its informative non-epitope mutations, as defined in the previous paragraph. A generic model of mutational load contains multiple non-linearities, for example in the dependence of HA protein fold stability on sequence and in the dependence of fitness on fold stability. To infer mutational load from our data set, we use the linearized fitness component

$$f_i^{\mathrm{ne}} = -\sigma_{\mathrm{ne}} D_{\mathrm{ne}}(\mathbf{a}_i, \mathbf{a}_i^*) \tag{17}$$

which is given in terms of the non-epitope amino acid distance between the strain sequence $\mathbf{a}_i$ and the sequence of its closest ancestor from a previous season, $\mathbf{a}_i^*$. This component depends on a single selection coefficient σ_{ne} (see section 4 for details on parameter inference).

Inference of fitness components. The full strain fitness model given by (8) and (16),

$$f_i = f_0 - \mathcal{L}(\mathbf{a}_i) - \sum_{j:t_j<t_i} x_j \mathcal{C}(\mathbf{a}_i, \mathbf{a}_j) \tag{18}$$

depends on the nonlinear cost functions $\mathcal{L}(\mathbf{a}_i)$ and $\mathcal{C}(\mathbf{a}_i, \mathbf{a}_j)$. The strain-independent constant $f_0(t)$ is determined by the normalization of strain frequencies,

$$\sum_{i\in\mathcal{S}(t)} x_i = \sum_{i\in\mathcal{S}(t)} x_i \exp(f_i) = 1 \tag{19}$$

To infer strain fitness values f_i from our data set, we partition this model into the simpler, distance-based fitness components f_i^{ep} and f_i^{ne}, which are given by equations (9) and (17), and an additional nonlinear component f_i^{nl},

$$f_i = f_i^{\mathrm{ep}} + f_i^{\mathrm{ne}} + f_i^{\mathrm{nl}} \tag{20}$$

The fitness component f_i^{nl} is to measure additional epitope and non-epitope fitness differences between clades, which include epistatic interactions between the mutations carrying a clade. This term is treated differently from the other two terms: we infer it for individual clades, but we do not attempt to fit it to specific functional forms of epitope and non-epitope selection. We use the form $f_i^{\mathrm{nl}} = f_{v(i)}^{\mathrm{nl}}(t(i))$, where $v(i)$ is the clade and $t(i)$ is the season to which strain i belongs, and

$$f_v^{\mathrm{nl}}(t) = \log\left[\frac{1}{X_v(t)} \sum_{j:t,v} x_j \exp\left[\lambda D_0(\mathbf{a}_j, \mathbf{a}^*(t))\right]\right] \equiv D_0^{\mathrm{ave}}(v(i), \mathbf{a}^*(t)) \tag{21}$$

is the average exponential distance in the set of synonymous (that is, bona fide neutral) mutations between the strains of clade v in season t and the last common ancestor strain $\mathbf{a}^*(t)$ of all strains in season t. The clade-based fitness component $f_v^{\mathrm{nl}}(t)$ turns out to be correlated with the epitope fitness component of the linear model, $f_v^{\mathrm{ep}}(t) = \log\left[\sum_{i:v,t} x_i \exp\left[f_i^{\mathrm{lin}}\right] / X_v(t)\right]$, with a Pearson coefficient 0.36 ± 0.03. We conclude that synonymous mutations accumulating in a clade are markers of positive selection for that clade, which is consistent with their ubiquitous hitchhiking in selective sweeps[9]. It is not implied that these mutations are themselves under selection.

Together, the full fitness model has the four tunable parameters: σ_{ep}, D_0, σ_{ne} and λ. These parameters can be calibrated to our data set, as described in section 4.

Other fitness components. The interaction of influenza with host cells involves several complex biochemical processes, which implies that its functional and selective aspects are not limited to the fitness components described so far. For example, glycosylation can contribute to suppress the recognition of epitopes by antibodies[50] and has potential fitness effects[51,52]. The human influenza A/H3N2 lineage has acquired six new potential glycosylation sites since 1968[44,53]. As shown in Extended Data Fig. 6, this dynamics is closely linked to the epitopes: glycosylation sites with their H residue in an epitope account for the entire increase in number over our prediction period from 1993 to 2012, and this number shows substantial variation in the strain populations of some years. By contrast, the number of glycosylation sites with their H residue outside the epitopes has a constant value for strains in the trunk lineage over the same period. The number of epitopic glycosylation sites, n_{ep}, shows changes that are inhomogeneous in time. During our prediction period, the population mean of n_{ep} decreases between 1994 and 1995, which is followed by an increase between 1995 and 2001 and fluctuations in the range $n_{\mathrm{ep}} = 4$–5 in later years. The trunk lineage shows an even faster increase to $n_{\mathrm{ep}} = 5$ between 1995 and 1997, and it maintains this number in later years. These data suggest a possible adaptive evolution of n_{ep} after 1995 up to a saturation value $n_{\mathrm{ep}} = 5$. The temporal inhomogeneity of this process can be explained by epistasis between glycosylation and other traits of the evolving viral population.

To test the predictive value of glycosylation, we introduce an additional component in our fitness model,

$$f_i^{\mathrm{gl}} = \sigma_{\mathrm{gl}} \min(n_{\mathrm{ep}}, 5) \tag{22}$$

with a fixed selection coefficient $\sigma_{\mathrm{gl}}^* = 1$. As shown in Extended Data Table 1, the glycosylation model (equation (22)) has predictive value as a single-epitope fitness component. However, this component adds only a small predictive value to the full model. At the same time, the fitness component (22) involves an additional model parameter and the gain of glycosylation sites appears to be limited to the H3 protein[53]. Therefore, we omit glycosylation in the predictions of the main text. The non-additivity of information gain between the glycosylation component (22) and the full model suggests that the broader epistatic term f_i^{nl} accounts at least partially for the fitness effects of glycosylation. This does not exclude that a more detailed selective model of glycosylation and related processes[54] can affect predictions.

3. Model predictions. The fitness model of equation (20) predicts next-year clade frequencies, as well as the expected cross-immunity between circulating strains and a given vaccine strain. Here we detail our prediction method, discuss the resulting model-based criterion for vaccine strain selection, and compare with the established strain selection procedure[55].

Prediction of clade frequencies. Given a strain with fitness f_i (measured in units of 1/year) and frequency x_i in season t, our fitness model predicts the expected frequency $\hat{x}_i = x_i \exp(f_i)$ in season $t+1$. However, the evolution of individual strains is dominated by mutations and sampling noise, and meaningful predictions can only be made for suitable aggregate variables. For clades with frequency $X_v(t) > 0.15$ given by equation (6), we construct the frequency prediction of equation (1),

$$\hat{X}_v(t+1) = \sum_{i\in\mathcal{S}(t)} \rho_{i,v} x_i \exp(f_i) \equiv \sum_{i:\ v,t} x_i \exp(f_i) \quad \text{for } v \in K(t) \tag{23}$$

which involves averaging over the ensemble of strain trees according to equation (5). Importantly, the inference of strain trees does not introduce future information to this prediction, because the tree ensemble used in equation (23) is constructed only from strains observed up to season t. To estimate a posteriori the actual next-year frequencies $X_v(t+1)$, we again use equation (6) with an ensemble of trees constructed from all strains of our data set. Predicted and posterior frequencies are

reported from winter 1994 to winter 2011, which ensures sufficient sampling depth and stable strain trees at the validation step.

Prediction of cross-immunity. Another aggregate variable suitable for predictions is the average cross-immunity between a given strain v and the circulating strains in a given season,

$$\mathcal{C}_v(t) \equiv \sum_{i:t} x_i \mathcal{C}(\mathbf{a}_i,\mathbf{a}_v) \tag{24}$$

Our model predicts the expected average cross-immunity in season $t+1$,

$$\hat{\mathcal{C}}_v(t+1) = \sum_{i:t} x_i \exp(f_i)\mathcal{C}(\mathbf{a}_i,\mathbf{a}_v) \tag{25}$$

This quantity determines, to first order in perturbation theory, the expected change of the total number of infections through vaccination based on strain v,

$$\begin{aligned}\frac{\hat{N}_\epsilon(t+1)}{\hat{N}(t+1)} &= \frac{1}{\hat{N}(t+1)}\sum_{i:t} N_i(t)\exp[f_i - \epsilon\,\mathcal{C}(\mathbf{a}_i,\mathbf{a}_v)] \\ &= 1-\epsilon\sum_{i:t} x_i \exp(f_i)\mathcal{C}(\mathbf{a}_i,\mathbf{a}_v) + O(\epsilon^2) \\ &= 1-\epsilon\,\hat{\mathcal{C}}_v(t+1) + O(\epsilon^2)\end{aligned} \tag{26}$$

Here ϵ is an effective vaccination coverage (which is smaller than the actual coverage because vaccinations are distributed heterogeneously in the population), $\hat{N}_\epsilon(t+1)$ is the expected number of infections with vaccination, and $\hat{N}(t+1)$ is the corresponding number without vaccination. The predicted amplitudes $\hat{\mathcal{C}}_v(t+1)$ can be compared a posteriori with the observed amplitudes $\mathcal{C}_v(t+1)$.

Vaccine strain selection. We predict optimal vaccine strains by maximizing $\hat{\mathcal{C}}_v(t+1)$ over all candidate strains v in seasons $t' < t$. We compare these predictions with the posterior cross-immunity centre-of-mass strains defined by maximizing $\mathcal{C}_v(t+1)$ over all candidate strains in seasons $t' < t+1$, and with the actually used vaccine strains[27] listed in Supplementary Data. Our prediction of vaccine strains uses a strain sample restricted to the months from October to February in its base year t, similarly to the current vaccine strain selection procedure[55]. Predicted, actual and posterior vaccine strains are reported from winter 1994 to winter 2012. Our method for vaccine strain selection involves two assumptions: (a) the overall vaccination coverage is small. This assumption is justified, because substantial vaccine coverage is limited to some countries and vaccination programs cover primarily high-risk groups in the population[56]; (b) the cross-immunity profile in the vaccinated population is similar to that in the entire population. That is, the high-risk groups subject to vaccination have had similar exposure to previous infections, leading to similar cross-immunity amplitudes $\mathcal{C}(\mathbf{a}_i,\mathbf{a}_v)$, as the entire population.

We note that our method does not attempt to quantify the overall efficacy of influenza vaccines, which depends not only on vaccine strain selection, but also on population structure with regard to susceptibility and on group-specific vaccination coverage. Quantifying vaccine efficacy is also complicated by the heterogeneity of data sources[56].

Currently, the WHO recommendations on vaccine strain selection are established in biannual consultations according to the following principal criteria[55]: (a) emergence of an antigenically and genetically distinct variant among circulating viruses (including a novel influenza A virus with the potential to cause a pandemic); (b) evidence of the geographical spread of such a distinct variant and its association with outbreaks of disease, indicating its future epidemiological significance; (c) reduced ability of existing vaccine-induced antibodies to neutralize the emergent variant; and (d) availability of suitable candidate vaccine viruses.

In comparison, the WHO strain selection criteria and our method have a number of common features. In both schemes, the vaccine for a given winter season is based on the strains circulating in the previous winter. Second, both schemes require geographical spread of clades (in our case through the minimum condition on clade frequencies, which is correlated with cross-continental spread as discussed above). Third, both schemes use genetic data to map differences between strains. Our current method uses epitope sequence differences as a proxy for antigenic differences, but direct antigenic haemagglutination inhibition data can be incorporated in a straightforward way. The most important new feature of our method is that it quantifies differences between strains in terms of their fitness, which is computed from epitope and non-epitope sequence data. Hence, it can gauge individually small differences between strains that accumulate between major antigenic transitions, and it can offer a principled choice of vaccine strains in years with more than two distinct circulating viral clades.

Vaccination feedback. The evolutionary predictions of our model can readily be generalized to larger values of vaccination coverage. Vaccination with a strain v at effective coverage ϵ introduces a component of artificial selection on the circulating strains, which penalizes cross-immunity with the vaccine strain. This component combines with the natural selection component described in section 2 to a coverage-dependent fitness function

$$f_i(\epsilon) = f_i(\epsilon=0) - \epsilon\,\mathcal{C}(\mathbf{a}_i,\mathbf{a}_v) + \mathcal{C}_0(\epsilon) \tag{27}$$

where the constant $C_0(\epsilon) = -\log\sum_{i:t} x_i \exp[f_i(\epsilon=0) - \epsilon\,\mathcal{C}(\mathbf{a}_i,\mathbf{a}_v)]$ is given by the normalization condition (equation (19)). Equation (23) then produces coverage-dependent predictions of clade frequencies,

$$\hat{X}_v(t+1,\epsilon) = \sum_{i:v,t} x_i \exp[f_i(\epsilon=0) - \epsilon\,\mathcal{C}(\mathbf{a}_i,\mathbf{a}_v) + \mathcal{C}_0(\epsilon)] \tag{28}$$

These reveal a generic negative feedback effect: vaccination introduces positive selection on strains that have a small cross-immunity with the vaccine strain, that is, $\mathcal{C}(\mathbf{a}_i,\mathbf{a}_v) < \mathcal{C}_0(\epsilon)$. This, in turn, reduces the efficacy of vaccination.

4. Model validation and optimization. The clades of HA genotypes defined above overlap with each other: clades of substantial frequency $X_v(t)$ have multiple nested subclades, which originate through subsequent nonsynonymous mutations in their descendent lineage (Fig. 1). We use these nested clades as units of prediction, because they can map evolutionary histories from origination of new strains to fixation or loss of their descendants. We now introduce a modified set of non-overlapping clades, which are used to define the information gain of a fitness model. This quantity serves as an objective function for the comparison of fitness models and for the inference of optimal model parameters.

Model validation by non-overlapping clades. We map each original clade $v \in K(t)$ onto a non-overlapping clade with strain sets

$$\breve{\mathcal{S}}_v^{(1)}, \ldots, \breve{\mathcal{S}}_v^{(n)} \quad \text{for } v \in K(t) \tag{29}$$

such that for each season t and for each tree κ, the sets

$$\breve{\mathcal{S}}_v^{(\kappa)}(t) = \breve{\mathcal{S}}_v^{(\kappa)} \cap \mathcal{S}(t) \quad \text{for } v \in K(t) \tag{30}$$

form a partitioning of that season's strains; that is, any strain $i \in \mathcal{S}(t)$ is contained in exactly one clonal set $\breve{\mathcal{S}}_v^{(\kappa)}(t)$. We construct the set $\breve{\mathcal{S}}_v^{(\kappa)}$ by eliminating from $\mathcal{S}_v^{(\kappa)}$ all strains that are contained in any proper subset $\mathcal{S}_\mu^{(\kappa)}$

$$\breve{\mathcal{S}}_v^{(\kappa)} = \mathcal{S}_v^{(\kappa)} \setminus \left\{ i \,\middle|\, i \in \mathcal{S}_\mu^{(\kappa)} \subset \mathcal{S}_v^{(\kappa)} \right\} \tag{31}$$

In analogy to equation (5), this defines the probability that strain i is contained in the non-overlapping clade v,

$$\breve{\rho}_{i,v} = \frac{1}{n}\sum_{\kappa=1}^{n} \breve{\epsilon}_{i,v}^{(\kappa)} \quad \text{with } \breve{\epsilon}_{i,v}^{(\kappa)} = \begin{cases} 1 & \text{if } i \in \breve{\mathcal{S}}_v^{(\kappa)} \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

The frequencies of non-overlapping clades are constructed in analogy to equations (6) and (23). Each non-overlapping clade $v \in K(t)$ has the tree-averaged frequency

$$Y_v(t) = \sum_{i \in \mathcal{S}(t)} \breve{\rho}_{i,v} x_i \quad \text{for } v \in K(t) \tag{33}$$

in season t. By construction, the frequencies of the non-overlapping clades in a given season form a discrete distribution $\mathbf{Y}_t = (Y_v(t))_{v \in K(t)}$, which satisfies the normalization condition $\sum_{v \in K(t)} Y_v(t) = 1$. In the same way, we obtain the predicted next-year frequencies for non-overlapping clades,

$$\hat{Y}_v(t+1) = \sum_{i \in \mathcal{S}(t)} \breve{\rho}_{i,v} x_i \exp(f_i) \quad \text{for } v \in K(t) \tag{34}$$

which form a normalized distribution $\hat{\mathbf{Y}}_{t+1} = \left(\hat{Y}_v(t+1)\right)_{v \in K(t)}$. Compared to its counterpart for nested clades, equation (23), the prediction (34) differs only in the mapping of strains to clades. We evaluate the underlying genotype frequencies x_i according to equation (3) with the regularization

$$x_i = \frac{1}{z_\epsilon(t)}\left[\frac{m_i}{m(t)} + \frac{\epsilon}{m_{v(i,\kappa)}(t)}\right] \tag{35}$$

where $v(i, \kappa)$ is the non-overlapping clade to which genotype i belongs in tree κ,

$$m_{v(i,\kappa)}(t) = \sum_{j \in \mathcal{C}_{v(i,\kappa)}} m_j \tag{36}$$

is the number of strains in that clade, and the constants $z_\epsilon(t)$ ensure the correct normalization $\sum_{i \in \mathcal{S}(t)} x_i = 1$. According to equation (35), each non-overlapping clade is assigned a frequency pseudocount , which is distributed equally among its strains. We choose a pseudocount value $\epsilon = 10^{-3}$, which is within the sampling fluctuations of clade frequencies. These pseudocounts are negligible for the nested clades used for prediction, which have frequencies $X_v(t) \geq 0.15$. However, they are useful for the information-theoretic analysis of fitness models, which we describe

in the following section. The results of this analysis are insensitive to the exact choice of ϵ.

Information gain of fitness models. To measure the quality of a fitness model, we compare the predicted frequency distributions for non-overlapping clades, $\hat{\mathbf{Y}}_{t+1}$, with the observed distributions, $\mathbf{Y}_{t+1}$, over the prediction period. The likelihood that the observed distributions are samples of populations with the predicted clade frequencies takes the form

$$p \sim \exp\left[-\sum_t \mathcal{H}(t)\right] \tag{37}$$

with

$$\mathcal{H}(t) = \tilde{m}(t) H\left(\mathbf{Y}_{t+1} \middle| \hat{\mathbf{Y}}_{t+1}\right) \tag{38}$$

In this expression, $H\left(\mathbf{Y}_{t+1} \middle| \hat{\mathbf{Y}}_{t+1}\right)$ denotes the relative (Kullback–Leibler) entropy of the distributions $\mathbf{Y}_{t+1}$ and $\hat{\mathbf{Y}}_{t+1}$,

$$H\left(\mathbf{Y}_{t+1} \middle| \hat{\mathbf{Y}}_{t+1}\right) = \sum_{v \in K(t)} Y_v(t+1) \log\left[Y_v(t+1)/\hat{Y}_v(t+1)\right] \tag{39}$$

and $\tilde{m}(t)$ is an effective strain sample size. We use the conservative estimate $\tilde{m}(t) = 50$, which reflects the observation that frequency estimates for clades with $X_v(t)$ of order 0.1 and below are dominated by stochastic effects (this estimate is also used for the error analysis of Extended Data Fig. 1). We can also evaluate the likelihood p_0 under a null model of neutral evolution, which assigns fitness values $f_i^0 = 0$ to all strains and predicts clade frequencies $\hat{Y}_v^0(t+1) = Y_v(t)$. We compare the frequency prediction of these models by the relative likelihood

$$\frac{p}{p_0} \sim \exp\left[\sum_t \Delta\mathcal{H}(t)\right] \tag{40}$$

where

$$\Delta\mathcal{H}(t) = \tilde{m}(t)\left[-H\left(\mathbf{Y}_{t+1} \middle| \hat{\mathbf{Y}}_{t+1}\right) + H(\mathbf{Y}_{t+1} | \mathbf{Y}_t)\right] \tag{41}$$

measures the information gain of the fitness model compared to the neutral null model for the prediction in season t. If we define an effective fitness $F_v(t)$ for non-overlapping clades from equations (33) and (34),

$$F_v(t) = \log \frac{\hat{Y}_v(t+1)}{Y_v(t)} \tag{42}$$

we can express the information gain in terms of the mean fitness in the observed next-year population,

$$\Delta\mathcal{H}(t) = \tilde{m}(t) \sum_{v \in K(t)} Y_v(t+1) F_v(t) \tag{43}$$

We use this information gain in two distinct ways: to infer optimal model parameters from training data, and to quantify the quality of prediction for different model variants.

Fitness parameter inference. The predictions reported in the main text are obtained from the haplotype-based fitness model (equations (20) and (21), which has four independent parameters, (σ_{ep}, D_0, σ_{ne}, λ). We determine calibrated values of these parameters as follows: (a) The cross-immunity range and the non-epitope selection coefficient are set to fixed values $D_0^* = 14$ (which corresponds to 50% decay of cross-immunity over an average of 6–7 years and is consistent with the cross-immunity range known from haemagglutination inhibition data[12]) and $\sigma_{ne}^* = -0.5$ (which can be estimated from the propagator ratio for informative non-epitope mutations shown in Extended Data Fig. 5). These values correspond to broad maxima of the average information gain; that is, our predictions are only weakly sensitive to variations of these parameters. Specifically, we find a significant increase of information gain for $D_0 \lesssim D_0^*$, but smaller changes for $D_0 \gtrsim D_0^*$. Similarly, the information gain varies insignificantly (by less than 10 points) for $0.2 < \sigma_{ne} < 0.7$. The broadness of these maxima of information gain also implies that an inference of season-dependent optimal values would lead to overfitting. (b) The remaining two parameters, σ_{ep} and λ, determine the accuracy of predictions in a more sensitive way. We infer optimal values $(\sigma_{ep}^*(t), \lambda^*(t))$ for each season t by maximization of the cumulative information gain (equation (41)) over a training period of 8 years,

$$\left(\sigma_{ep}^*(t), \lambda^*(t)\right) = \arg\max_{\sigma_{ep},\lambda} \sum_{t'=t-8}^{t-1} \Delta\mathcal{H}\left(t'; \sigma_{ep}, D_0^*, \sigma_{ne}^*, \lambda\right) \tag{44}$$

This optimization is straightforward by steepest descent[57]. A training period of a few years is required, because the training data are noisy. However, our predictions are robust under changes of the training period. At the current quality of our strain sample, we obtain predictions of high information gain ($\Delta\mathcal{H}_{tot} > 500$) for training periods of 4 years and above; cf. equation (45) and Extended Data Table 1. Shorter training periods lead to overfitting and loss of information gain ($\Delta\mathcal{H}_{tot} = 213$ for a period of two years and $\Delta\mathcal{H}_{tot} = 448$ for a period of 3 years).

Comparison of fitness models. To measure the quality of different model variants, we use the cumulative information gain (equation (41)) of its predictions over the total period of the data set,

$$\Delta\mathcal{H}_{tot} = \sum_t \Delta\mathcal{H}\left(t; \sigma_{ep}^*(t), D_0^*, \sigma_{ne}^*, \lambda^*(t)\right) \tag{45}$$

We compare several classes of fitness models:

(a) Full, haplotype-based model: this model is defined in equations (20) and (21) and is used for our predictions in the main text. It has four tunable selection parameters, (σ_{ep}, D_0, σ_{ne}, λ), which parametrize the distance-based epitope and non-epitope fitness components and the clade-based nonlinear fitness component. (b) Linear model: to test the role of nonlinearities in our full fitness model, we introduce a model variant with the linearized fitness function

$$f_i^{lin} = \sigma_{ep} D_{ep}(\mathbf{a}_i, \mathbf{a}^*(t)) - \sigma_{ne} D_{ne}\left(\mathbf{a}_i, \mathbf{a}_i^*\right) \tag{46}$$

where $\mathbf{a}^*(t)$ is the sequence of the last common ancestor of all strains in season t, and $\mathbf{a}_i^*$ is the sequence of the closest previous-season ancestor of strain i (see section 2). The linear model has only two parameters, (σ_{ep}, σ_{ne}). Each new mutation has a fixed selection coefficient, which equals σ_{ep} for epitope mutations and $-\sigma_{ne}$ for informative non-epitope mutations. This model produces a significantly reduced total information gain, which shows that the nonlinearities in the full fitness function are important. (c) Epitope model variants: these models test alternative sets of antigenic HA sites discussed in previous studies[2,3,7,11,41,42] and the predictive role of glycosylation[44,50–53] (see section 3). The epitope model variants detailed above have the same four parameters (σ_{ep}, D_0, σ_{ne}, λ) as the full model and give substantially inferior predictions. The full model plus the glycosylation fitness component (equation (22)) has one additional parameter σ_{gl}, but does not produce significantly higher information gain. (d) Epitope-only models: to test the role of non-epitope fitness components in the full model, we introduce reduced models that contain only the epitope fitness component (equation (9)). These models have two parameters, (σ_{ep}, D_0), and produce a significantly reduced information gain compared to the full model. We also list a model based only on glycosylation, which has a single parameter σ_{gl}, and a null model with 49 random 'epitope' codons, which is uninformative as expected.

These parameter-optimized models are compared in Extended Data Table 1. We list the information gain $\Delta\mathcal{H}_{tot}$ together with two related measures, the cumulative mean fitness flux $\bar{\Phi}$, which is defined in equation (48) below, and the average fitness variance,

$$V = \frac{1}{\Delta t} \sum_t \sum_{v \in K(t)} Y_v(t)\left[F_v(t) - \bar{F}(t)\right]^2 \tag{47}$$

where $\bar{F}(t) = \sum_{v \in K(t)} Y_v(t) F_v(t)$ and Δt is the length of the prediction period. According to all three measures, the subleading models capture only substantially smaller fractions of the adaptive process than the full model. In particular, the non-epitope and nonlinear fitness components are important elements of our prediction.

Robustness of predictions. The prediction results depend only weakly on a number of model details:

(a) Duration of winter season: we report predictions for consecutive Northern winter periods from October to April, which cover most of the influenza isolates reported in the Northern Hemisphere. We have verified that our predictions are robust with respect to changes of these intervals. For example, the prediction from year t to $t + 1$ can be based on a strain sample restricted to the period from October to February in the base year t ($\Delta\mathcal{H}_{tot} = 514$ compared to $\Delta\mathcal{H}_{tot} = 541$ for the full period). The vaccine strain predictions reported in Fig. 4 use this restricted strain sample (see section 3). (b) Length of training period: our predictions are based on strain data from 8 consecutive years before the base year of prediction. Analogous prediction schemes with training periods of 4 years and above produce predictions of the same quality ($\Delta\mathcal{H}_{tot} > 500$). This is important for the applicability of our method to influenza strains with a shorter past circulation time in a given host population. (c) Regional sampling bias: our prediction method clearly depends on efficient geographical mixing[17,18] of the clades used for predictions. As discussed in the main text, 94% of these clades cover two or more continents. As an additional check, we have introduced an alternative procedure to evaluate strain frequencies, which includes a regional stratification: we partition the strains of each Northern winter season into geographical regions C (see section 2), and evaluate the regional frequencies $\chi_C = \sum_{i \in C} x_i$. In 15 of the 20 prediction seasons, one region accounts for the majority of the strains, $\chi_C > 1/2$. In these cases, we downweigh

the strain frequencies in that region ($x_i \to x_i/(2\chi_C)$ for $i \in C$) and upweigh the frequencies in the other regions, ($x_i \to x_i/(2(1-\chi_C))$ for $i \notin C$), which leads to $\chi_C \to 1/2$ and $\sum_{C' \neq C} \chi_{C'} \to 1/2$. Our predictions are robust under this change with a slight decrease in prediction quality, which is consistent with regional mixing. If clade frequencies were distorted by regional sampling bias in a significant way, this stratification should have improved the prediction results. (d) Choice of D_0 and σ_{ne}: as detailed above, our predictions are robust under variations of these parameters in a broad range. Hence, we choose fixed values D_0^* and σ_{ne}^* within this range, and we restrict the inference of time-dependent values to the parameters $\sigma_{ep}^*(t)$ and $\lambda^*(t)$.

These robustness properties provide consistency checks for our method.

Predictions for seasonal influenza A/H1N1. We test the broader applicability of our method by applying it to the seasonal human influenza A/H1N1. Our analysis is based on 2136 HA1 sequences from the years 1977–2009, which have been obtained from NCBI[1] and Gisaid[25] databases, accounting for redundancies in the two databases. Despite the similar overall number of sequences, this data set is less informative than its H3N2 counterpart: (a) there are only eight winter seasons t (1990, 1995–1997 and 2005–2008) with at least 12 unique HA1 genotypes in season t and $t+1$, and even fewer with the same number of complete HA sequences; (b) in the epitope codons identified by Huang *et al.*[26], we find evidence of positive antigenic selection by evaluating the frequency propagator ratio, equation (7). We obtain $g(1) = 1.6$, which is smaller than the corresponding value for H3N2 (see Extended Data Fig. 4). This may point to weaker antigenic selection[8], but also to larger uncertainty about the exact epitope positions[26].

To avoid overfitting, we restrict our analysis to the HA1 domain and to the eight winter seasons with at least 12 strains in base and target year of prediction. We use the simpler linear fitness model of equation (46), and we do not attempt to fit time-dependent model parameters. Informative epitope codons are defined by a maximum past nucleotide diversity of 0.15, which serves to delineate non-epitope mutations from false-negative antigenic changes in unknown epitope sites. We obtain the predictions reported in Extended Data Fig. 2, which correctly produce clade growth (decline) in 88% (63%) of the cases. A consistent information gain $\Delta\mathcal{H}_{tot} = 41$ is found for parameters $\sigma_{ep}^* = 0.3$ and $\sigma_{ne}^* < -1.5$, which is in line with the substantial negative selection on informative non-epitope mutations inferred from their propagator ratio $g(1) = 0.02$.

5. Fitness flux. Here we define fitness flux as a generic measure of adaptation and discuss the statistics of this measure. The cumulative mean fitness flux, which has been introduced in a number of previous publications[28,29,58], measures the total amount of adaptation in a population over a given interval of evolutionary time. We extend the theory of fitness flux to a clade-dependent flux, which measures the total amount of adaptation up to a given clade. We use the mean fitness flux to compare the prediction quality of fitness models; see Extended Data Table 1. The clade-dependent fitness flux serves to establish the adaptive map of influenza shown in Fig. 4.

Definitions of fitness flux. Consider an evolving population with observed genotypes (or non-overlapping genotype clades, as defined in section 4), which are labelled by an index $v \in K(t)$. We consider the evolutionary history of this population in a time interval $t' = t_0, t_0+1, \ldots, t$, which is described by observed frequencies $\mathbf{Y}(t') = (Y_v(t'))_{v \in K(t')}$ and predicted fitness values $F_v(t') = \log\left[\tilde{Y}_v(t'+1)/Y_v(t')\right]$ given by equations (33) and (34). The cumulative mean population fitness flux of this process is defined as[29]

$$\bar{\Phi}(t) = \sum_{t'=t_0}^{t-1} \sum_{v \in K(t')} [F_v(t') - \bar{F}(t')][Y_v(t'+1) - Y_v(t')] \tag{48}$$

where $\bar{F}(t) = \sum_{v \in K(t)} Y_v(t) F_v(t)$ denotes the mean population fitness. Here we extend this definition to a clade-specific flux,

$$\Phi_v(t) = F_v(t) - \bar{F}(t) + \bar{\Phi}(t) \tag{49}$$

The definition of fitness flux is illustrated in Extended Data Fig. 3a. Figure 4 shows the clade-specific fitness flux obtained from the full fitness model.

Properties of fitness flux. Fitness flux measures the total amount of adaptation in a population history[29]. The flux measures (equations (48) and (49)) used in the main text have the following key properties. (a) The mean fitness flux $\bar{\Phi}(t)$ measures correlations between clade fitness values, which are model predictions in this study, and the actual clade frequency evolution in the following year; see equation (48). These correlations quantify prediction quality; a null model with scrambled fitness values has $\bar{\Phi}(t) \approx 0$. We use the mean fitness flux to rank different fitness models; see Extended Data Table 1. (b) The mean fitness flux $\bar{\Phi}(t)$ satisfies the fitness flux theorem, which implies that this quantity is an almost universally increasing function of time[29]. For a stationary non-equilibrium process, in particular, the expectation value of $\bar{\Phi}(t)$ increases linearly with time. This seems to be a reasonable approximation for the adaptive process of influenza, as shown by Fig. 4. (c) The clade-specific flux $\Phi_v(t)$ differs from the mean flux $\bar{\Phi}(t)$, reflecting deviations of clade fitness relative from mean population fitness,

$$\Phi_v(t) - \bar{\Phi}(t) = F_v(t) - \bar{F}(t) \tag{50}$$

Hence, the fitness flux variance equals the population fitness variance in season t,

$$\mathrm{Var}\ \Phi(t) = \sum_{v \in K(t)} Y_v(t)[\Phi_v(t) - \bar{\Phi}(t)]^2 = \mathrm{Var}\ F(t) \tag{51}$$

This quantity is shown in Extended Data Fig. 3b. (d) The time dependent distribution of fitness flux values,

$$\mathcal{Y}(\Phi, t) = \int d\Phi Y_v(t) \delta(\Phi - \Phi_v(t)) \tag{52}$$

defines a travelling fitness (flux) wave[31–33,59]. Figure 4 shows the fitness flux wave inferred for influenza A/H3N2; (e) the clade-specific flux can also be interpreted in terms of a variational calculus: the mean population flux for an evolutionary process in the time interval $t_0, \ldots, t+1$ with frequencies $\mathbf{Y}(t') = (Y_v(t'))_{v \in K(t')}$ for $t' = t_0, \ldots, t$ and hypothetical frequencies $\tilde{\mathbf{Y}} = (\tilde{Y}_v)_{v \in K(t)}$ in year $t+1$ is a linear form with coefficients $\Phi_v(t)$,

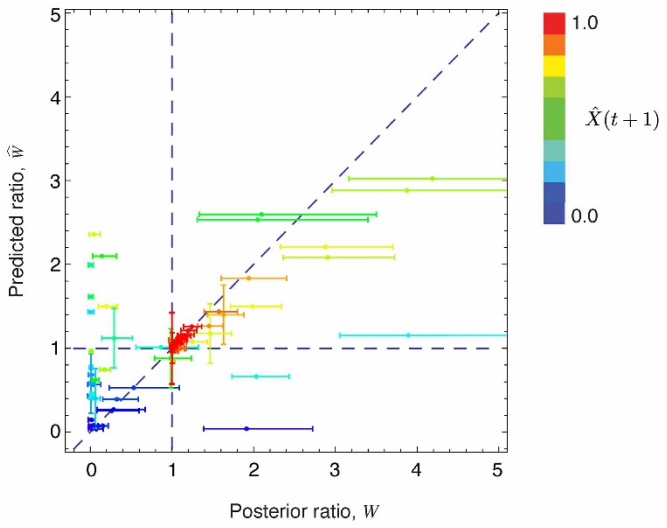
$$\bar{\Phi}(\tilde{\mathbf{Y}}, t+1) = \sum_{v \in K(t)} \Phi_v(t) \tilde{Y}_v \tag{53}$$

In particular, the mean population flux for the actual process is given by

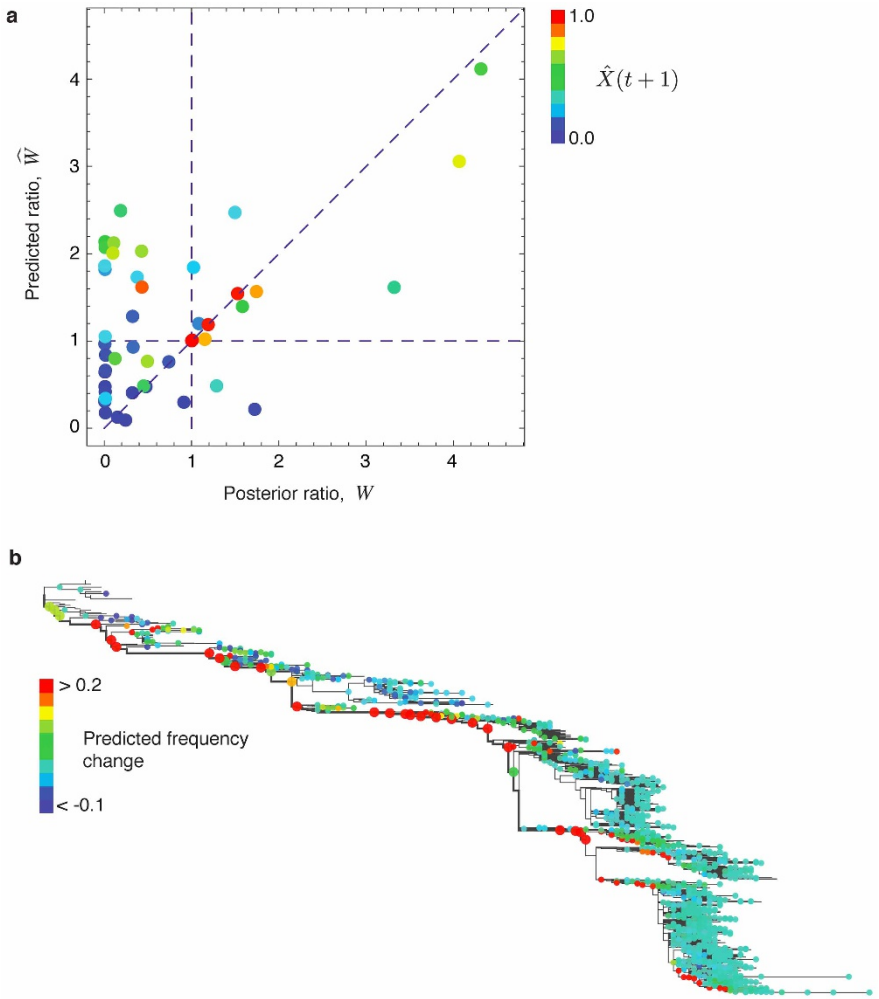
$$\bar{\Phi}(t+1) = \sum_{v \in K(t)} \Phi_v(t) Y_v(t+1) \tag{54}$$

37. Ghedin, E. *et al.* Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* **437,** 1162–1166 (2005).
38. Bush, R. M., Smith, C. B., Cox, N. J. & Fitch, W. M. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proc. Natl Acad. Sci. USA* **97,** 6974–6980 (2000).
39. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5,** e9490 (2010).
40. Swofford, D. L. *PAUP*: Phylogenetic Analysis Using likelihood (and Other Methods) 4.0 Beta* (Sinauer Associates, 2002).
41. Weis, W. *et al.* Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature* **333,** 426–431 (1988).
42. Wilson, I. A. & Cox, N. J. Structural basis of immune recognition of influenza virus hemagglutinin. *Annu. Rev. Immunol.* **8,** 737–787 (1990).
43. Skehel, J. J. & Wiley, D. C. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu. Rev. Biochem.* **69,** 531–569 (2000).
44. Lin, Y. P. *et al.* Evolution of the receptor binding properties of the influenza A(H3N2) hemagglutinin. *Proc. Natl Acad. Sci. USA* **109,** 21474–21479 (2012).
45. Andreasen, V., Lin, J. & Levin, S. A. The dynamics of cocirculating strains conferring partial cross-immunity. *J. Math. Biol.* **35,** 825–842 (1997).
46. Gog, J. R. & Swinton, J. A status-based approach to multiple strain dynamics. *J. Math. Biol.* **44,** 169–184 (2002).
47. Koelle, K., Khatri, P., Kamradt, M. & Kepler, T. A two-tiered model for simulating the ecological and evolutionary dynamics of rapidly evolving viruses, with an application to influenza. *J. R. Soc. Interface* **7,** 1257–1274 (2010).
48. Molinari, N. M. *et al.* The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine* **25,** 5086–5096 (2007).
49. Myers, J. L. *et al.* Compensatory hemagglutinin mutations alter antigenic properties of influenza viruses. *J. Virol.* **87,** 11168–11172 (2013).
50. Schulze, I. T. Effects of glycosylation on the properties and functions of influenza virus hemagglutinin. *J. Infect. Dis. Aug.* **176** (suppl. 1), S24–S28 (1997).
51. Blackburne, B. P., Hay, A. J. & Goldstein, R. A. Changing selective pressure during antigenic changes in human influenza H3. *PLoS Pathog.* **4,** e1000058 (2008).
52. Cui, J., Smith, T., Robbins, P. W. & Samuelson, J. Darwinian selection for sites of Asn-linked glycosylation in phylogenetically disparate eukaryotes and viruses. *Proc. Natl Acad. Sci. USA* **106,** 13421–13426 (2009).
53. Zhang, M. *et al.* Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology* **14,** 1229–1246 (2004).
54. Arinaminpathy, N. & Grenfell, B. Dynamics of glycoprotein charge in the evolutionary history of human influenza. *PLoS ONE* **5,** e15674 (2010).
55. Ampofo, W. K. *et al.* Improving influenza vaccine virus selection. Report of a WHO informal consultation held at WHO headquarters, Geneva, Switzerland, 14–16 June 2010. *Influenza Other Respir. Viruses* **6,** 147–152 (2010).

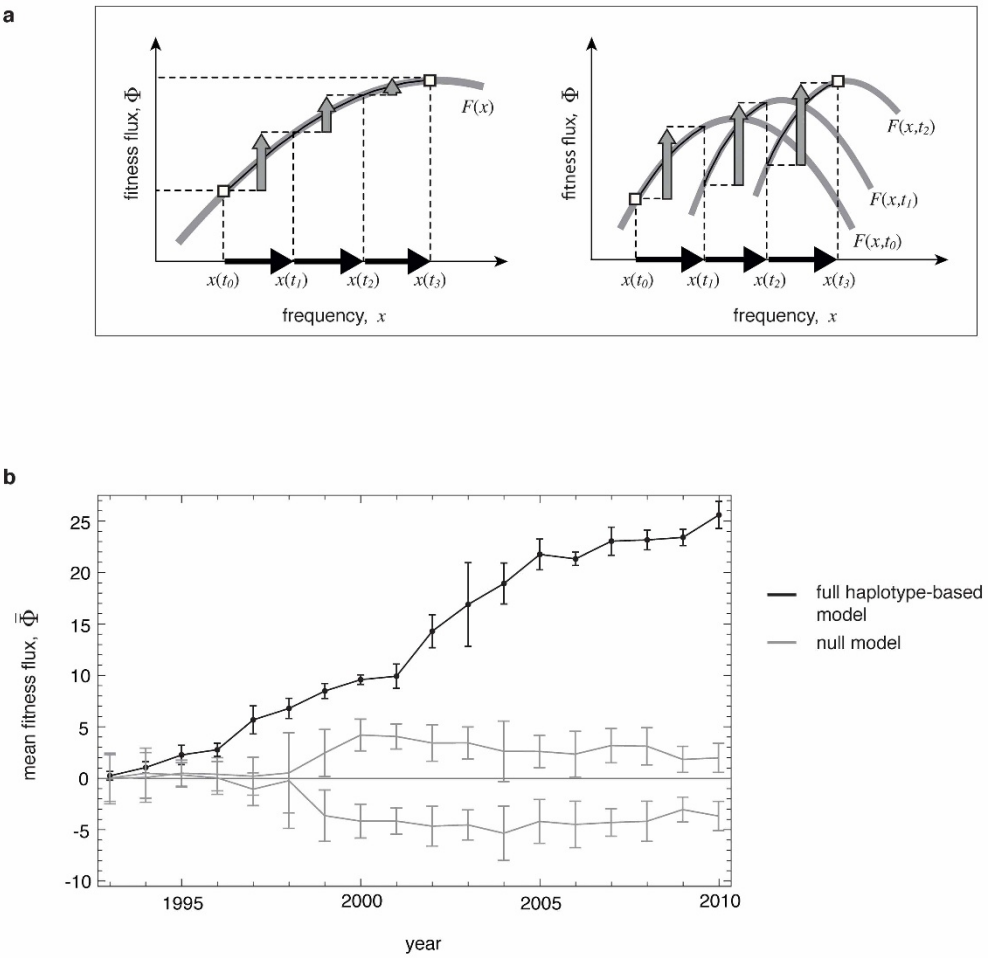
56. Osterholm, M. T., Kelley, N. S., Sommer, A. & Belongia, E. A. Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. *Lancet Infect. Dis.* **12,** 36–44 (2012).
57. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes: the Art of Scientific Computing* 3rd edn (Cambridge Univ. Press, 2007).
58. Mustonen, V. & Lässig, M. Adaptations to fluctuating selection in *Drosophila*. *Proc. Natl Acad. Sci. USA* **104,** 2277–2282 (2007).
59. Tsimring, L. S., Levine, H. & Kessler, D. A. RNA virus evolution via a fitness-space model. *Phys. Rev. Lett.* **76,** 4440–4443 (1996).



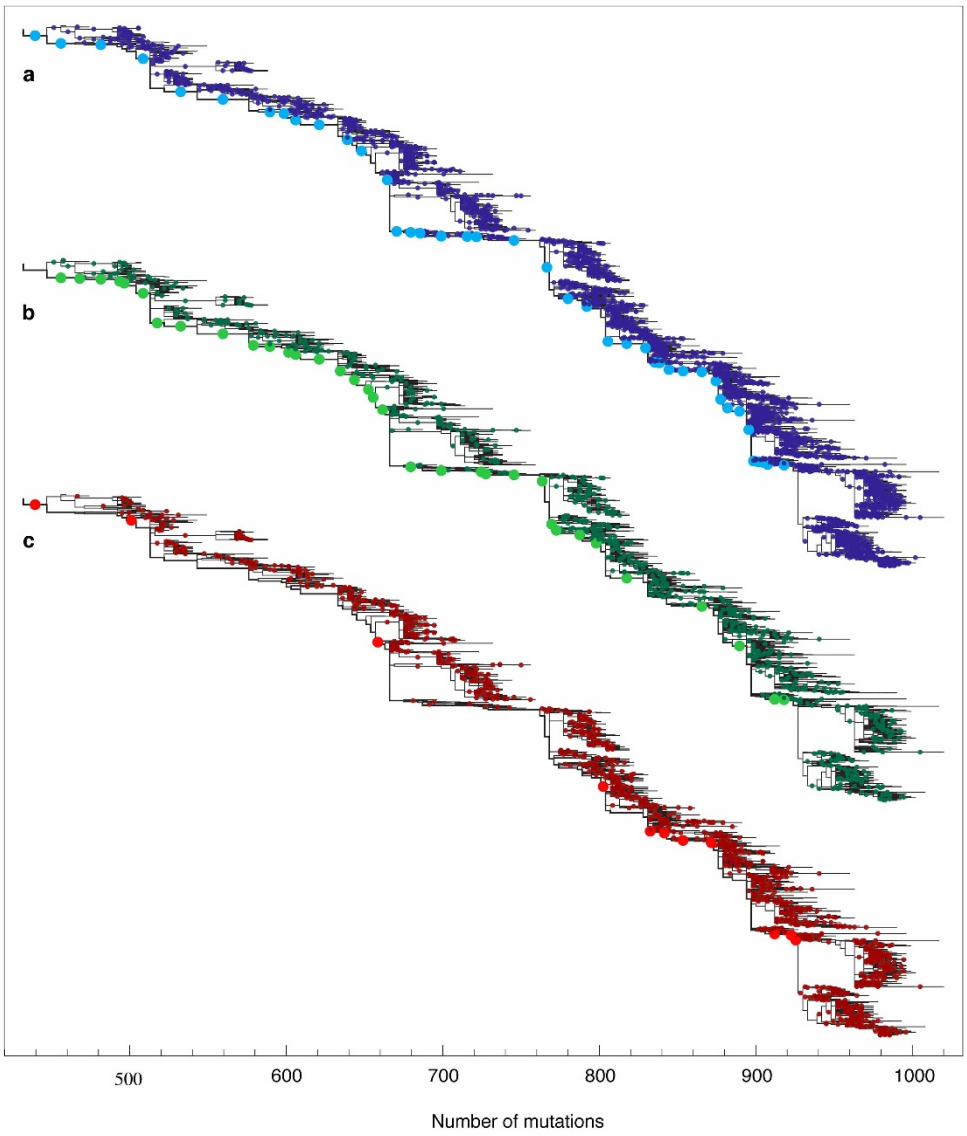
Extended Data Figure 1 | Statistical errors of predicted and posterior Wrightian fitness. The frequency ratio plot (W_v, $\hat{W}_v$) of Fig. 2a is shown together with the standard deviation of the predicted ratio $\hat{W}_v = \hat{X}_v(t+1)/X_v(t)$ in the ensemble of reconstructed trees (vertical bars) and the standard deviation of the posterior ratio $W_v = X_v(t+1)/X_v(t)$ due to sampling fluctuations of population frequencies (horizontal bars). See sections 1, 3 and 4 of Methods.



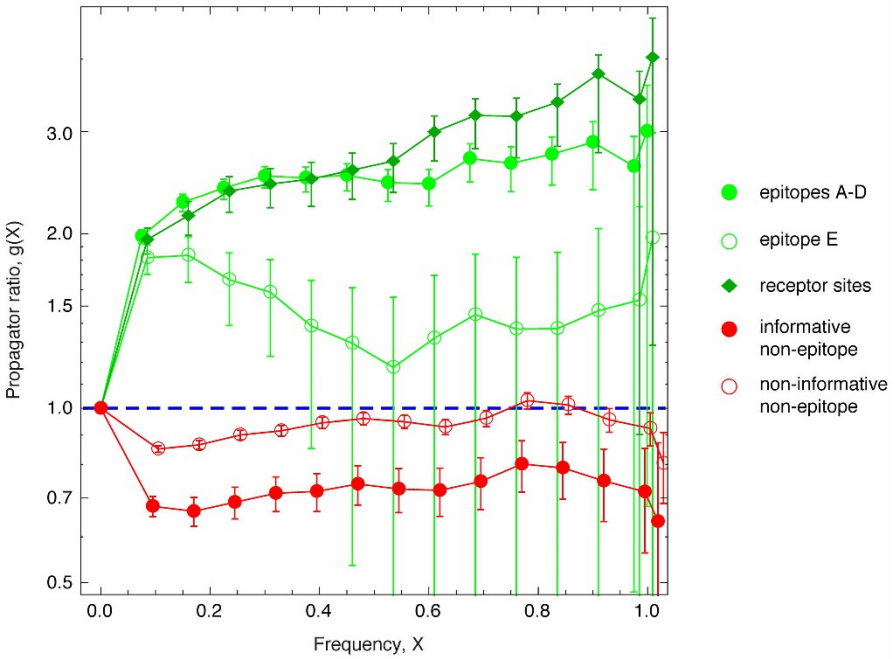
Extended Data Figure 2 | Fitness predictions for human influenza A/H1N1. **a**, Wrightian fitness: the predicted frequency ratio $\hat{W}_v = \hat{X}_v(t+1)/X_v(t)$ is plotted against the posterior ratio $W_v = X_v(t+1)/X_v(t)$ for 81 HA clades with initial frequency $X_v(t) > 0.1$. To be compared with Fig. 2a. **b**, Dynamics of the influenza strain tree: for each clade, the ancestor node is coloured according to the maximum of the predicted frequency changes, $\max_t[\hat{X}(t+1) - X(t)]$. To be compared with Fig. 2c. The predictions are based on a sample of 2,136 unique HA1 genotypes observed between 1977 and 2009. We restrict predictions to years when this sample contains at least 12 unique HA1 strains in the winter seasons t and $t + 1$, which are the years 1990, 1995–1998 and 2005–2008 (see Methods, section 4). These predictions are statistically significant ($P < 10^{-18}$) but somewhat more noisy than for influenza A/H3N2 (clade growth is correctly predicted in 88% of the cases, decline in 63% of the cases). The reasons include a significantly smaller and more biased strain sample and a less comprehensive knowledge of antigenic epitope sites[26]. The prediction of influenza A/H1N1 evolution illustrates the broader applicability of our method and highlights the determinants of predictive power.



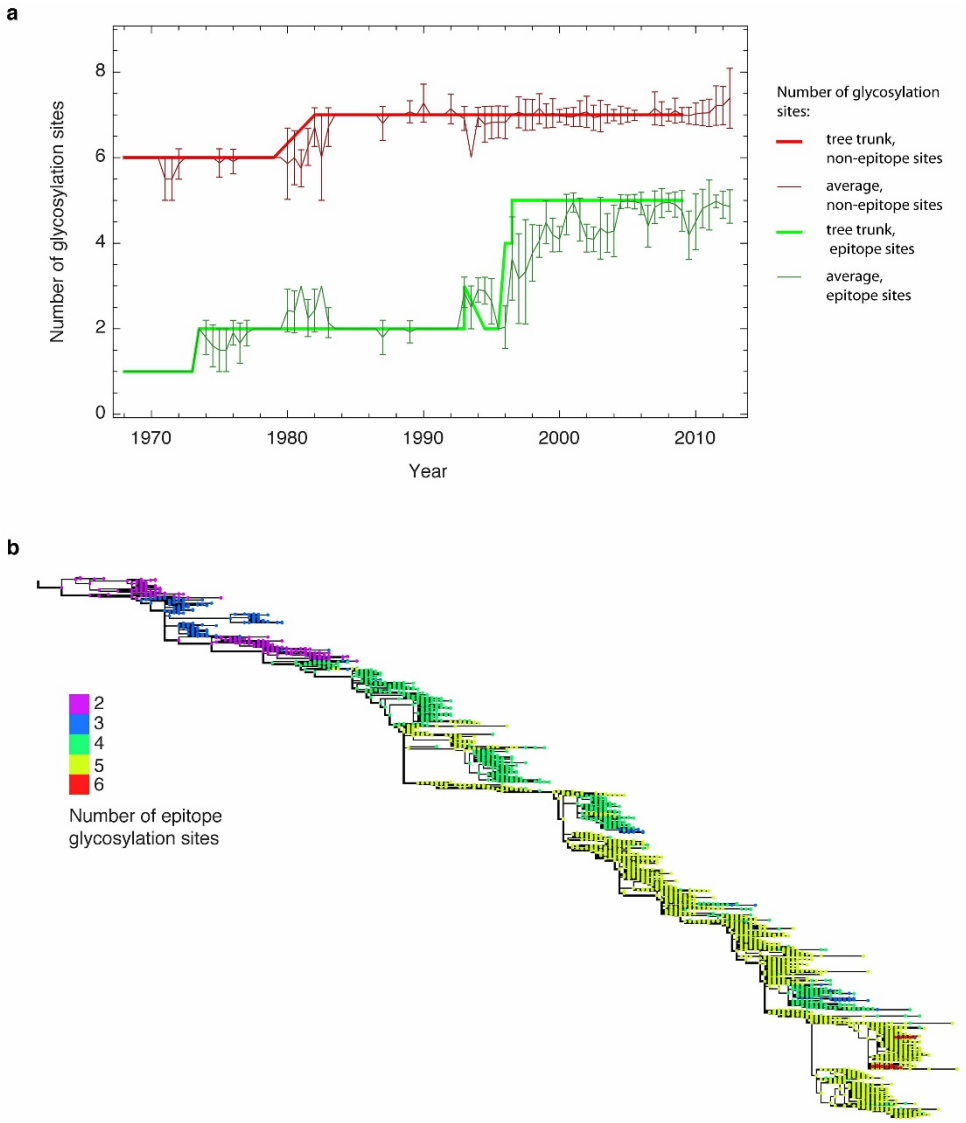
Extended Data Figure 3 | Fitness flux in the evolution of influenza. **a**, Fitness flux measures adaptation (schematic, adapted from ref. 29). The cumulative flux $\Phi(t)$, as defined in equations (48) and (49), is an aggregate measure of fitness changes due to frequency changes in a population's history up to a given clade v at a given time t (shown by uphill arrows)[28,29,58]. Left: in a static fitness landscape $F(x)$, the flux $\Phi(t)$ equals the fitness difference between the initial point and the final point of this history. Right: in a time-dependent fitness seascape $F(x, t)$, the flux $\Phi(t)$ is still a typically positive, increasing function of time, even if the population fitness does not increase. **b**, Mean cumulative fitness flux $\Phi(t)$ as given by equation (48) for influenza from 1993 up to season t. The mean flux inferred from our fitness model (black line) shows a continuous increase. The flux for a null model with scrambled clade fitness values (grey lines) fluctuates around 0. Vertical bars indicate the root mean squared fitness (flux) in each year's strain sample, $\sqrt{\mathrm{Var}\ \Phi(t)} = \sqrt{\mathrm{Var}\ F(t)}$, as given by equation (51).



Extended Data Figure 4 | Strain tree with mutations. **a**–**c**, Four classes of HA sequence mutations are mapped onto individual branches of the influenza strain tree: synonymous mutations (**a**, blue), nonsynonymous epitope mutations (**b**, green) and nonsynonymous non-epitope mutations (**c**, red). Each nonsynonymous mutation marks the origination of a clade in the population; each fixed clade (highlighted by bright colours) has an origination on the trunk of the tree (shown as thick line). The fixation probability, that is, the ratio of the number of fixations and the number of originations, is seen to be reduced for informative non-epitope changes and enhanced for nonsynonymous epitope changes compared to the baseline of synonymous changes; cf. Extended Data Fig. 5. The underlying tree (shown here from 1993 to 2012) is a sample from our ensemble of strain trees, which are constructed by maximum likelihood from the HA sequence of 3,944 strains (other equiprobable trees differ only in peripheral branches). The horizontal coordinate D of a node is its mutational distance from the root of the tree. The trunk of the tree (thick line) is the single lineage connecting past and future on timescales beyond the coalescence time.



Extended Data Figure 5 | Selection on epitope and non-epitope changes. The frequency propagator ratio[9] $g(X)$, as defined in equation (7), is shown for several classes of nonsynonymous HA mutations: mutations in epitopes A–D (green bullets), mutations in epitope E (green circles), mutations in sialic receptor binding sites (green diamonds), informative non-epitope mutations (red bullets) and non-informative non-epitope mutations (red circles). Error bars indicate sampling uncertainties. Mutations in epitopes A–D, including those in epitopic receptor binding sites, reach values $g(X) \geq 2.5$ for large frequencies, signalling substantial positive selection. Mutations in epitope E are under weaker positive selection, with $g(X) \approx 1.5$ for large frequencies. Informative non-epitope changes drop to $g(X) = 0.6$, signalling predominantly negative selection. Non-informative non-epitope changes evolve near the neutral baseline ($g = 1$, blue line), indicating weak or heterogeneous selection. See section 2 of Methods.



Extended Data Figure 6 | Evolution of glycosylation. **a**, Number of epitopic glycosylation sites, n_{ep}, in the influenza A/H3N2 strain sample between 1968 and 2012 (green lines): population mean value (thin line), root mean squared variation (error bars), and value for trunk lineages (thick line). The same data are shown for non-epitope glycosylation sites (red lines). Epitope sites show substantial changes with a net increase in number and substantial natural variation in some years, whereas non-epitope sites had only one fixation of a glycosylation site. **b**, Evolution of n_{ep} on the influenza strain tree between 1993 and 2012. Trunk strains show a rapid increase to $n_{ep} = 5$ between 1995 and 1997 and maintain this value in later years; the mean n_{ep} shows a slower increase between 1995 and 2001. Off-trunk clades drop below $n_{ep} = 5$ also in later years, and there are compensatory mutations back to $n_{ep} = 5$. The data suggest an adaptive increase of n_{ep} up to a saturation value $n_{ep} = 5$ after 1996. These observations inform the glycosylation fitness component (equation (22)), which is used to test the predictive value of glycosylation. See section 2 of Methods.

Extended Data Table 1 | Ranking of fitness models

model	σ^*_{ep}	D^*_0	σ^*_{ne}	λ^*	σ^*_{gl}	$\Delta\mathcal{H}_{tot}$	$\bar{\Phi}$	V
full haplotype-based	1.15 ± 0.29	14.	-0.5	0.31 ± 0.26	0	541.	26.	2.1
linear	0.52 ± 0.07	0	-0.5	0	0	326.	15.	0.9
epitope variants								
– excl. receptor binding sites [7]	1.06 ± 0.38	14.	-0.5	0.30 ± 0.25	0	451.	23.	1.9
– codon subset [3]	0.37 ± 0.36	14.	-0.5	0.39 ± 0.32	0	187.	14.	1.5
– extended codon set [11]	1.14 ± 0.29	14.	-0.5	0.32 ± 0.26	0	523.	25.	2.1
– incl. glycosylation [42, 48–51]	1.01 ± 0.32	14.	-0.5	0.29 ± 0.24	1.	549.	29.	2.2
epitope-only								
– all codons [7]	1.39 ± 0.50	14.	0	0	0	297.	16.	1.2
– excl. receptor binding sites [7]	1.20 ± 0.53	14.	0	0	0	201.	12.	0.8
– codon subset [3]	0.60 ± 0.44	14.	0	0	0	33.	2.	0.1
– extended codon set [11]	1.36 ± 0.49	14.	0	0	0	270.	15.	1.1
– only glycosylation [42, 48–51]	0	0	0	0	1.	136.	5.	0.2
– random codons	0.34 ± 0.36	14.	0	0	0	-69.	1.	.04

We compare the four classes of fitness models described in Methods, section 4: (1) Full haplotype-based model with four parameters (σ_{ep}, D_0, σ_{ne}, λ), which is used for the predictions reported in the main text. (2) Linearized model with two parameters (σ_{ep}, σ_{ne}). (3) Epitope model variants based on restricted[3,7] and extended[11] codon sets, and the full model plus glycosylation. (4) Epitope-only models with two parameters, (σ_{ep}, D_0), glycosylation-only model with one parameter, σ_{gl}, and a null model with 49 randomly chosen 'epitope' sites. For each model, optimal parameters $\sigma^*_{ep}(t)$, $\lambda^*(t)$ are obtained by maximizing the cumulative information gain over a training period of 8 years; see equation (44). We show mean and standard deviation of the optimized parameters over the entire prediction period, the total information gain $\Delta\mathcal{H}_{tot}$ given by equation (45), the total mean fitness flux Φ given by equation (48), and the average fitness variance V given by equation (47). According to all three measures, the five-parameter full model plus glycosylation captures a slightly larger fraction, all other subleading models capture only substantially smaller fractions of the adaptive process than the full model without glycosylation. The time-independent parameter values D^*_0, σ^*_{ne} and σ^*_{gl} correspond to broad maxima of the total information gain. See Methods, section 2 and equations (9), (17), (20), (21), (22) and (46) for model definitions.