

Limited predictability of amino acid substitutions in seasonal influenza viruses

Pierre Barrat-Charlaix,^{1,2} John Huddleston,³ Trevor Bedford,⁴ and Richard A. Neher^{1,2,*}

¹*Biozentrum, Universität Basel, Switzerland*

²*Swiss Institute of Bioinformatics, Basel, Switzerland*

³*Molecular Cell Biology, University of Washington, Seattle, WA, USA*

⁴*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA*

(Dated: July 31, 2020)

Seasonal influenza viruses repeatedly infect humans in part because they rapidly change their antigenic properties and evade host immune responses, necessitating frequent updates of the vaccine composition. Accurate predictions of strains circulating in the future could therefore improve the vaccine match. Here, we studied the predictability of frequency dynamics and fixation of amino acid substitutions. Current frequency was the strongest predictor of eventual fixation, as expected in neutral evolution. Other properties, such as occurrence in previously characterized epitopes or high *Local Branching Index* (LBI) had little predictive power. Parallel evolution was found to be moderately predictive of fixation. While the LBI had little power to predict frequency dynamics, it was still successful at picking strains representative of future populations. The latter is due to a tendency of the LBI to be high for consensus-like sequences that are closer to the future than the average sequence. Simulations of models of adapting populations, in contrast, show clear signals of predictability. This indicates that the evolution of influenza HA and NA, while driven by strong selection pressure to change, is poorly described by common models of directional selection such as travelling fitness waves.

INTRODUCTION

Seasonal influenza A viruses (IAV) infect about 10% of the global population every year, resulting in hundreds of thousands of deaths [1, 2]. Vaccination is the primary measure to reduce influenza morbidity. However, the surface proteins hemagglutinin (HA) and neuraminidase (NA) continuously accumulate mutations at a high rate, leading to frequent antigenic changes [2–5]. While a vaccine targeting a particular strain may be efficient for some time, antigenic drift will sooner or later render it obsolete. The World Health Organization (WHO) regularly updates influenza vaccine recommendations to best match the circulating strains. Since developing, manufacturing, and distributing the vaccine takes many months, forecasting the evolution of influenza is of essential interest to public health [6, 7].

The number of available high quality HA and NA sequences has increased rapidly over the last 20 years [8, 9] and virus evolution and dynamics can be now be tracked at high temporal and spatial resolution [10]. This wealth of data has given rise to an active field of predicting influenza virus evolution [6, 7]. These models predict the future population of influenza viruses by estimating strain fitness or proxies of fitness. Luksza and Lässig [11], for example, train a fitness model to capture antigenic drift and protein stability on patterns of epitope and non-epitope mutations. Other approaches by Steinbrück et al. [12], Neher et al. [13] predict fitness by using hemagglutination inhibition (HI) data to determine possible

antigenic drift of clades in the genealogy of the HA protein. Finally, Neher et al. [14] use branching patterns of HA phylogenies as a proxy for fitness. These branching patterns are summarized by the Local Branching Index (LBI), which was shown to be a proxy of relative fitness in mathematical models of rapidly adapting populations [14].

The underlying assumption of all these methods is that (i) differences in growth rate between strains can be estimated from sequence or antigenic data and (ii) that these growth rate differences persist for long enough to be predictive of future success. Specific positions in surface proteins are of particular interest in this context. The surface proteins are under a strong positive selection and change their amino acid sequence much more rapidly than other IAV proteins or than expected under neutral evolution [4, 15]. Epitope positions, i.e., positions targeted by human antibodies, are expected to change particularly often since viruses with altered epitopes can evade existing immune responses [3, 5, 16]. It therefore seems plausible that mutations at these positions have a tendency to increase fitness and a higher probability of fixation [15]. But one has to be careful to account for the fact that these positions are often ascertained post-hoc [3] and human immune responses are diverse with substantial inter-individual variation [17].

In this work, we use HA and NA sequences of A/H3N2 and A/H1N1pdm influenza from year 2000 to 2019 to perform a retrospective analysis of frequency trajectories of amino acid mutations. We quantify how rapidly mutations at different frequencies are lost or fixed and how rapidly they spread through the population. We further investigate whether any properties or statistics are predictive of whether a particular mutation fixes or not. To our surprise, we find that the predictability of

* Correspondence to: Richard Neher, Biozentrum, Klingelbergstrasse 70, 4056, Basel, Switzerland.
richard.neher@unibas.ch

these trajectories is very limited: The probability that a mutation fixes differs little from its current frequency, as would be expected if fixation happened purely by chance. This observation holds for many different categories of mutations, including mutations at epitope positions. This weak predictability is not attributable solely to clonal interference and genetic linkage, as simulation of models including even strong interference retain clear signatures of predictability. Consistent with these observations, we show that a simple predictor uninformed by fitness, the consensus sequence, performs as well as the Local Branching Index (LBI), the growth measure based on the genealogy used in [14]. This suggests that although LBI has predictive power, the reason for its success may not be related to it approximating fitness of strains.

RESULTS

The main underlying question asked in this work is the following: given a mutation X in the genome of influenza that we observe at a frequency f in the population at a given date, what can we say about the future of X ? The trajectory of a mutation will depend on its own effect on fitness, the contribution of the genetic background on the same segment, and the effect of the remaining seven segments. Here, we investigate properties of broad categories of mutations effectively averaging over different genetic backgrounds to isolate the effects intrinsic to the mutation.

First, we ask whether we can quantitatively predict the frequency of X at future times $f(t)$. In other words, having observed a mutation at frequencies (f_1, f_2, \dots, f_n) at dates (t_1, t_2, \dots, t_n) , what can we say about its frequency at future dates $(t_{n+1}, t_{n+2}, \dots)$? A simpler, more qualitative question, is to ask whether X will fix in the population, will disappear, or whether the site will stay polymorphic.

We use amino-acid sequences of the HA and NA genes of A/H3N2 (since the year 2000) and A/H1N1pdm (since the year 2009) influenza available in GISAID [9] (see supplementary materials for an acknowledgment of all data contributors). This amounts to 44 976 HA and 36 300 NA sequences for A/H3N2 and 45 350 HA and 40 412 NA sequences for A/H1N1, with a minimum of 100 per year. These sequences are binned in non-overlapping intervals of one month. Each single-month time bin and the sequences that it contains represent a (noisy) snapshot of the influenza population at a given date. The number of sequences per time bin varies strongly both with year and according to the season, with earlier time bins containing around 10 sequences while more recent bins contain several hundreds (see figures S5 and S6 in SM for details).

The central quantities that we derived from this data are *frequency trajectories* of amino acids at each position in the sequences. If an amino acid X_i is found at position i at a frequency between 5% and 95% in the population of a given time bin t , then the population is considered

polymorphic at position i and at time t . This polymorphism is characterized by the frequency of X_i , $f_{X_i}(t)$, and also by frequencies of other amino acids at i . The series of values $f_{X_i}(t)$ for contiguous time bins constitutes the frequency trajectory of X_i . A trajectory is terminated if the corresponding frequency is measured above 95% (resp. below 5%) for two time bins in a row, in which case amino acid X_i is considered as *fixed* (resp. *absent*) in the population. Otherwise, the trajectory is considered *active*. Examples of trajectories can be seen in figure S7 of the Supplement.

In the rest of this work, we will focus on frequency trajectories that are starting at a zero (low) frequency, i.e. $f(t=0) = 0$. These represent new amino acid variants which were absent in the population at the time bin when the trajectory started and are currently rising in the population (see Methods). Such distinction in novel and ancestral variants is necessary to meaningfully interrogate predictability. Each rising trajectory of a new mutation implies the existence of another decreasing one at the same position, since frequencies of all amino acids at a given position must sum to one. If novel variants arise by selection, we expect to see a stronger signal of selection after conditioning on these novel variants. In classic models of population genetics, strongly advantageous variants undergo rapid selective *sweeps*, i.e., the rapid rise and fixation. The sweep of a mutation can be due to its own fitness effect, to the genetic background or to the effect of the seven other segments. By considering the ensemble of novel variants that are rising in frequency, we effectively average over backgrounds, obtaining a set of mutations that we expect to be beneficial on average. If such sweeps are common in the evolution of HA and NA, the restriction to trajectories that start at low frequency should thus enrich for mutations that are positively selected and on their way to fixation.

Predicting future frequencies

Having observed the frequency trajectory $f(t)$ of a mutation until a given date t_0 , how much can we say about the future values of f after t_0 ? We consider the idealized case sketched in panel **A** of figure 1: given the trajectory of a *new* mutation, i.e. that started at a frequency of 0, and that we observe at frequency f_0 at time t_0 , what is the probability $P_{\Delta t}(f)$ of observing it at a value f at time $t_0 + \Delta t$?

To answer this question retrospectively, we use all frequency trajectories extracted from HA and NA sequences that satisfy these conditions for a given f_0 . The number of trajectories is limited and the frequency estimates themselves are based on a finite sample and are hence imprecise. Therefore, we consider trajectories in an interval $[f_0 - \delta f, f_0 + \delta f]$ with $\delta f = 0.05$.

For $f_0 = 0.3$, we found 120 such trajectories in the case of A/H3N2 influenza, represented on the panel **B** of figure 1, where time is shifted such that $t_0 = 0$. The

same analysis was performed for A/H1N1pdm, with the 89₂₅₄ found trajectories displayed in figure S9. Some trajectories fall in the frequency bin around f_0 while decreasing, even₂₅₅ though they crossed that bin at an earlier time. This₂₅₆ is due to the fact that some trajectories “skipped” the₂₅₇ interval f_0 in question on their initial rise due to sparse₂₅₈ sampling. These trajectories are nevertheless rising in₂₅₉ the sense that they start at frequency 0 for $t \rightarrow -\infty$ ₂₆₀. Removing them does not change results significantly.₂₆₁

Since rapid sequence evolution of influenza HA and₂₆₂ NA mediates immune evasion, one could expect that₂₆₃ a significant fraction of new amino acid mutations on₂₆₄ rising trajectories in figure 1 are *adaptive*. We could thus₂₆₅ expect that most of these trajectories continue to rise after₂₆₆ reaching frequency f_0 , at least for some time. A fraction₂₆₇ of those would then sweep through the population and₂₆₈ fix.₂₆₉

To quantify the extent to which this preconception of₂₇₀ sweeping adaptive mutations is true, we estimated the₂₇₁ probability distribution $P_{\Delta t}(f|f_0)$ of finding a trajectory₂₇₂ at frequency f after a time Δt given that it was observed₂₇₃ at f_0 at time 0. The results for different Δt are shown in₂₇₄ figure 1C. Initially, *i.e.* at time $t_0 = 0$, this distribution is₂₇₅ by construction peaked around f_0 . If a large fraction of₂₇₆ the trajectories keep increasing after this time, we should₂₇₇ see the “mass” of $P_{\Delta t}(f|f_0)$ move to the right towards₂₇₈ higher frequencies as time progresses.₂₇₉

However, future distributions for $\Delta t > 0$ do not seem₂₈₀ to follow a pattern compatible with selective sweeps. The₂₈₁ thick black line in Figure 1B shows the average frequency₂₈₂ of all trajectories. This average makes a sharp turn at₂₈₃ $t = 0$ and is essentially flat for $t > 0$ in the case of₂₈₄ A/H3N2, and slightly increasing for A/H1N1pdm (see₂₈₅ supplement). Hence, the fact that this average rose for₂₈₆ $t < 0$ gives little information for $t > 0$, and is due to the₂₈₇ conditions by which these trajectories were selected. This₂₈₈ shows that sweep-like trajectories rising steadily from₂₈₉ frequency 0 to 1 are not common enough to dominate the₂₉₀ average trajectory.₂₉₁

Consistent with the average, the frequency distribution₂₉₂ of the selected trajectories broadens in time without a₂₉₃ significant shift of the mean as time passes. After 60 days,₂₉₄ the distribution is rather symmetrical around the initial₂₉₅ $f_0 = 0.3$ value, suggesting that the knowledge that the₂₉₆ trajectories were rising is lost after two months. On a₂₉₇ timescale of 60 to 120 days, the only possible prediction₂₉₈ is that trajectories are likely to be found in a broad₂₉₉ interval around the initial frequency f_0 . After one year₃₀₀ the distribution becomes almost flat (excluding mutations₃₀₁ that have disappeared or fixed), and the initial peak at f_0 ₃₀₂ is not visible anymore. The only information remaining₃₀₃ from the initial frequency is the fraction that fixed or was₃₀₄ lost (see below). This behavior is expected in neutral₃₀₅ models of evolution [18] but incompatible with a dynamic₃₀₆ dominated by sweeps taking over the population.₃₀₇

While this observation does not rule out that signal₃₀₈ tures exist that predict future frequency dynamics, past₃₀₉ dynamics alone is weakly informative.₃₁₀

Prediction of fixation or loss

Instead of predicting future frequency, let’s consider the long-term goal of predicting the probability that a mutation fixes in the population. We first estimate the fraction of frequency trajectories that either fix in the population or are lost, as well as the time it takes for one or the other to happen. Panels **A** and **B** of figure 2 shows the fraction of frequency trajectories in HA and NA that either have fixed, were lost or remained active as a function of the time elapsed since they were first seen above 25% frequency. Most mutations are either lost or become fixed after 2-3 years, with very few trajectories remaining active after 5 years. This time scale of 2-3 years is consistent with the typical coalescence time observed in phylogenetic trees of A/H3N2 influenza [10, 19]. We also note that the fraction of lost trajectories increases sharply at small times with 40% of mutations observed above 25% frequency being lost within one year for A/H3N2, while it takes longer to fix a mutation in the whole population.

We then examined the probability of mutations to fix in the population as a function of the frequency at which they are seen. For different values of frequency f , we consider all trajectories that started at a null frequency and are seen in the interval $[f - 7.5\%, f + 7.5\%]$ at any given time. The probability of a mutation fixing given that it is seen at frequency f , $P_{fix}(f)$, is then estimated by the fraction of those trajectories which terminate at a frequency larger than 95%, *i.e.* our fixation threshold. Panels **C** and **D** of figure 2 show $P_{fix}(f)$ as a function of f for NA and HA. For both proteins, the probability of fixation of a new mutation at frequency f is close to f itself, that is $P_{fix}(f) \simeq f$. This result is exactly what is expected in a population evolving in the *absence* of selection. A mutation or trait appearing at frequency f is shared by $f \cdot N$ individuals, and the probability for one of them to become the ancestor of all the future population is $f \cdot N/N = f$. Thus, the probability of this mutation or trait to fix in the population is equal to its current frequency, a case which we will refer to as the neutral expectation. Panel **C** of figure 2 indicates that mutations in the surface proteins of A/H3N2 influenza are in good agreement with the neutral expectation, while those in A/H1N1pdm show only small deviations from it. In both cases, the probability of fixation seems to be mainly dictated by the current frequency f at which the mutation is observed.

This dynamics is in apparent contradiction with evidence that influenza surface proteins are under strong selective pressure to evade human immune responses [4]. If strong selection was present, we would expect rising amino acid mutations to fix at a distinctively higher frequency than the one at which they are measured. In an extreme case where most trajectories would be clean sweeps, $P_{fix}(f)$ should be close to 1 for all but very small values of f .

Next, we searched for features of mutations that allow prediction of fixation beyond frequency by dividing

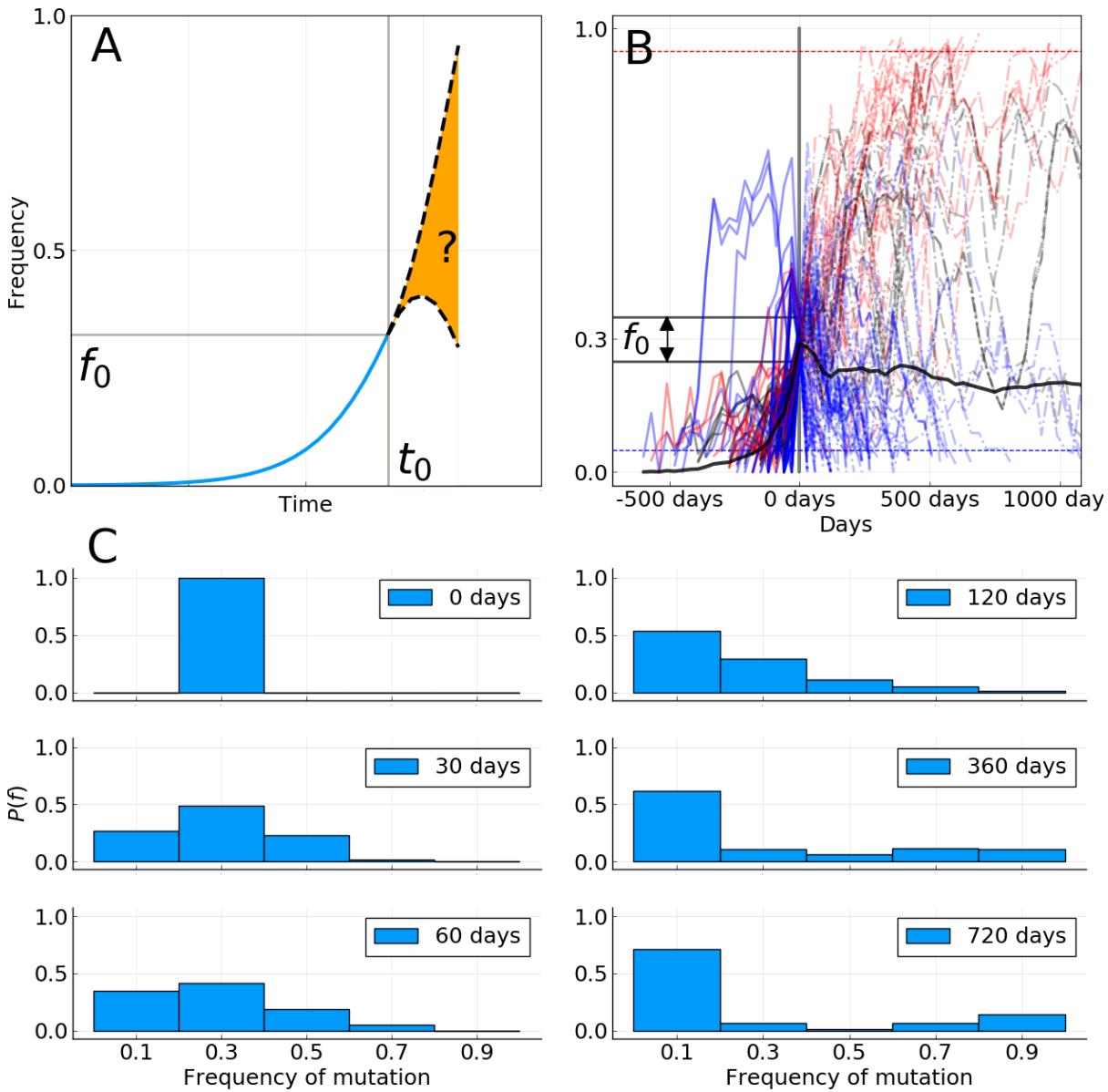


FIG. 1. **A:** Sketch of the idea behind the short term prediction of frequency trajectories. Given a mutation that we have seen increasing in frequency and that we “catch” at frequency f_0 at time t_0 , what can we say about the distribution of future frequencies $P_{\Delta t}(f|f_0)$? **B:** All frequency trajectories of amino acid mutations in the A/H3N2 HA and NA genes that were absent in the past, are seen around $f_0 = 30\%$ frequency at time $t_0 = 0$, and are based on more than 10 sequences at each time point. Red curves represent mutations that will ultimately fix, blue the ones that will be lost, and black the ones for which we do not know the final status. Dashed horizontal lines (blue and red) represent loss and fixation thresholds. The thick black line is the average of all trajectories, counting those that fix (resp. disappear) as being at frequency 1 (resp. 0). Figure S8 shows equivalent figures for other values of f_0 . **C:** Distribution of future frequencies $P_{\Delta t}(f|f_0)$ for the trajectories shown in panel **B** and for specific values of Δt .

311 frequencies into categories that deviate from the diagonal
 312 in panels **C** and **D** of figure 2. We first turn to
 313 the *Local Branching Index* (LBI), a quantity calculated
 314 for each node in a phylogenetic tree that indicates how
 315 dense the branching of the tree is around that node. LBI
 316 has previously been successfully used as a predictor of
 317 the future population of influenza [14], and was shown
 318 to be a proxy for fitness of leaves or ancestral nodes in
 319 mathematical models of evolution. Here, we define the
 320 LBI of a mutation at date t as the average LBI of strains
 321 that carry this mutation and that were sampled in the
 322 time bin corresponding to t . Panel A of figure 3 shows
 323 fixation probability for HA mutations with LBI in the
 324 top or bottom half of the distribution. Both groups have

to be a proxy for fitness of leaves or ancestral nodes in
 mathematical models of evolution. Here, we define the
 LBI of a mutation at date t as the average LBI of strains
 that carry this mutation and that were sampled in the
 time bin corresponding to t . Panel A of figure 3 shows
 fixation probability for HA mutations with LBI in the
 top or bottom half of the distribution. Both groups have

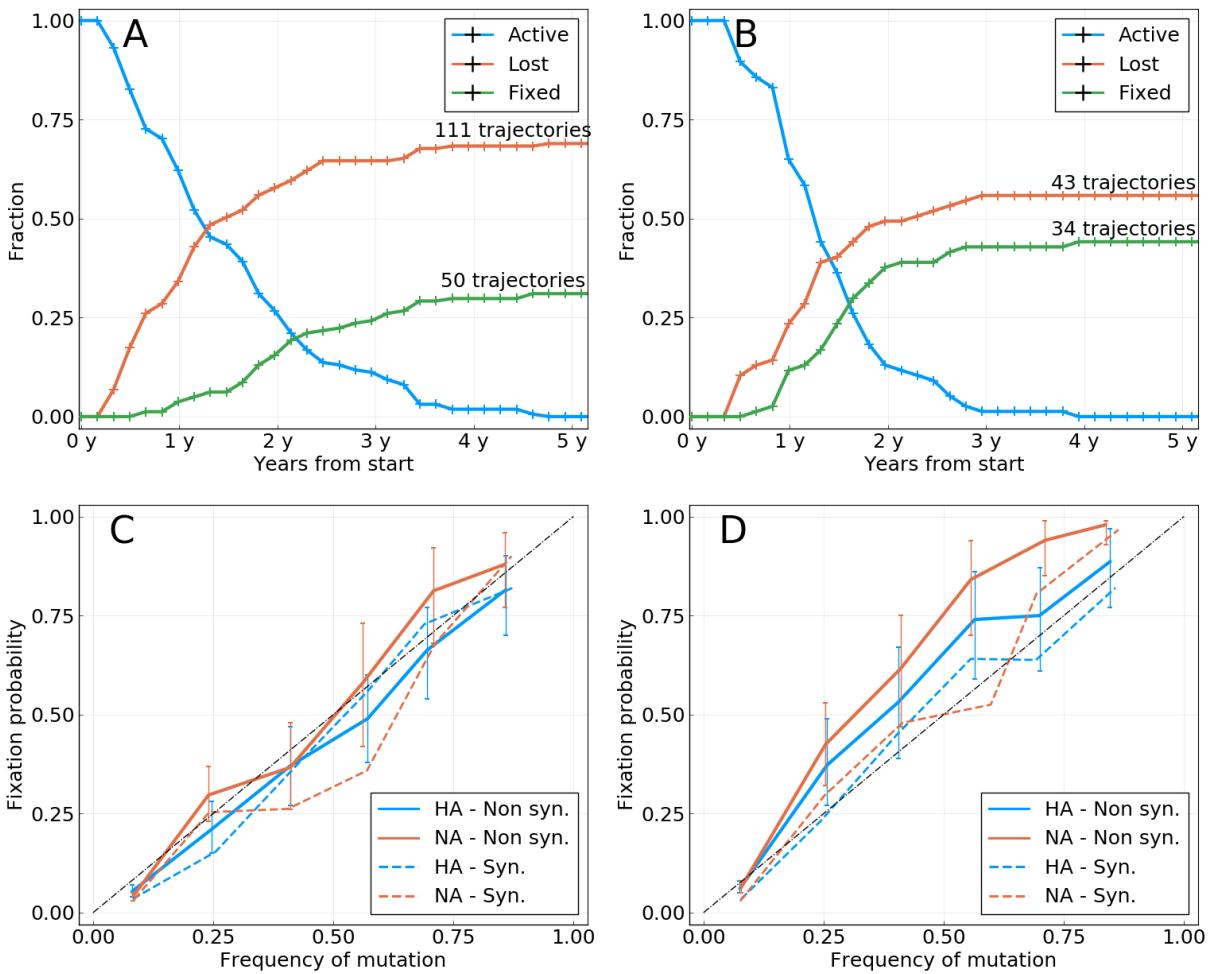


FIG. 2. **A:** Activity of all rising frequency trajectories seen above 25% frequency for A/H3N2 HA and NA. **B:** Same as **A** for A/H1N1. **C:** Probability of fixation of a mutation (amino acid or synonymous) $P_{fix}(f)$ as a function of the frequency f at which it is measured, for A/H3N2 HA and NA. Only new mutations are considered, *i.e.* mutations that were absent in the past. The diagonal dashed line is the expectation from a neutrally evolving population. Colored dashed lines represent synonymous mutations. Colored solid lines represent amino acid mutations. Error bars represent a 95% confidence interval. **D:** Same as **C** for A/H1N1.

identical probability of fixation, suggesting that LBI carries very little information on the probability of fixation of a mutation.

Next, we focused on previously reported antigenic sites in the A/H3N2 HA protein, referred to as *epitope positions*. Mutations at these positions might mediate immune escape and are therefore likely under strong selection and show sweep-like behavior. We used four lists of relevant epitope positions from different sources comprising from 7 to 129 positions in the sequence of the HA1 protein [3, 5, 11, 16]. Panel Fig. 3B shows fixation probability as a function of frequency for the four lists of epitopes. Only mutations at the 7 epitope sites reported in [5] have higher chances of fixation than expected by chance. No clear difference is found for the lists by Luksza and Lässig [11], while positions from Shih et al. [3] show lower chances of fixation. One should also note that

many of these positions were determined post-hoc and might be enriched for positions that experienced rapid substitutions before the publication of the respective studies.

Two ways of categorising mutations, however, suggest some power to predict fixation. In panel Fig. 3C, we split trajectories into those occurring at binary positions where only two amino acid variants co-circulate and non-binary positions with more than two variants. Novel variants at non-binary positions, *i.e.* ones for which competition between three amino acids or more has occurred at least once, have a higher chance of fixation. In panel D, we separated mutations that appear more than once or only once in the reconstructed tree (see methods), and found that the former fix more often. Panels C and D show that it is possible to gain some information on the chance of fixation of a particular mutation, as was done in panel

359 **B.** However, the predictive power remains small, with⁴¹⁴
360 the “top” curves in panels **C&D** being very close to the⁴¹⁵
361 diagonal.⁴¹⁶

362 We conduct the same analysis on A/H1N1pdm⁴¹⁷
363 influenza, with results shown in figure S11. Results are⁴¹⁸
364 qualitatively similar to those obtained for A/H3N2, with⁴¹⁹
365 LBI giving little information and mutations at non-binary⁴²⁰
366 positions having a higher chance of fixation. Panel⁴²¹
367 **D** differs between figures 3 and S11, with convergent⁴²²
368 evolution giving less information on fixation in the latter⁴²³
369 case. However, this could be due to the shorter time⁴²⁴
370 period over which A/H1N1pdm evolved, resulting in a⁴²⁵
371 shorter tree and less possibilities of convergent evolution⁴²⁶
372 Indeed, error bars for mutations appearing multiple times⁴²⁷
373 in **D** of figure S11 are relatively large, indicating a lower⁴²⁸
374 amount of trajectories.⁴²⁹

375 Since influenza is seasonal in temperate regions, geo⁴³¹
376 graphic spread and persistence might be predictive of the⁴³²
377 success of mutations. We quantify geographic spread of⁴³³
378 a mutation by the entropy of its frequency distribution⁴³⁴
379 across regions (see methods) and its persistence by the⁴³⁵
380 age of the trajectory by the time it reaches frequency f ⁴³⁶
381 Figures S12 and S13 show the fixation probabilities as⁴³⁷
382 a function of observed frequency for mutations classified⁴³⁸
383 according to these scores. The two scores also allow a⁴³⁹
384 quantitatively moderate distinction between mutations⁴⁴⁰
385 for a given frequency f , mutations found in many regions⁴⁴¹
386 or those that are older (in the sense that they have taken⁴⁴²
387 more time to reach frequency f) tend to fix more often⁴⁴³
388 than geographically localized mutations or more recent⁴⁴⁴
389 ones, but the effect is small. These two scores are in⁴⁴⁵
390 fact correlated, with older trajectories representing mu⁴⁴⁶
391 tations that are more geographically spread, as can be⁴⁴⁷
392 seen in figure S14 of SM. However, it is important to note⁴⁴⁸
393 that sampling biases and heterogeneity across time and⁴⁴⁹
394 space (see supplementary figures S5 and S6) make answer⁴⁵⁰
395 ing such specific hypothesis challenging. Frequency of⁴⁵¹
396 mutations might thus be amplified through different sam⁴⁵²
397 pling biases, making the connection between geographic⁴⁵³
398 spread, seasonality and mutation frequency non-trivial to⁴⁵⁴
399 measure.⁴⁵⁵

401 Simulations of models of adaptation⁴⁵⁶

402 The results shown in figures 2 and 3 are difficult to⁴⁶⁰
403 reconcile with the idea that seasonal influenza virus evo⁴⁶¹
404 lution is driven by rapid directed positive selection. One⁴⁶²
405 possible explanation for the weakly predictable behaviour⁴⁶³
406 of mutations (beyond their current frequency) might be⁴⁶⁴
407 tight genetic linkage inside each segment and strong com⁴⁶⁵
408 petition between different adaptive mutations [15, 20].⁴⁶⁶
409 We design a simple model of population evolution based⁴⁶⁷
410 on the `ffpopsim` simulation software to test this hypoth⁴⁶⁸
411 esis [21]. The model represents a population of binary⁴⁶⁹
412 genomes of length $L = 200$ evolving in a fitness landscape⁴⁷⁰
413 that changes through time.⁴⁷¹

First, we use an additive fitness function, with sequence $(x_1 \dots x_L)$ having a fitness $\sum_i h_i x_i$. This implies that for a given genome position i , the trait $x_i = 1$ is favored if $h_i > 0$ whereas $x_i = -1$ is favored if $h_i < 0$. All h_i 's have the same magnitude, and only their signs matter. Every Δt generations, we randomly choose a position i and flip the sign of h_i , effectively changing the fitness landscape. Individuals in the population now have the opportunity to make an adaptive mutation at site i giving them a fitness advantage $2|h|$. A “flip” at position i of the fitness landscape will decrease fitness of all individuals that carried the adapted variant at position i and increases the fitness of those that happened to carry a deleterious variant.

To increase competition between genomes, we designed a second model that includes epistasis. Once again, the baseline fitness of a genome is an additive function, this time with values of h_i that do not change through time. In addition, we added a component that mimics immune selection. Every Δt generation, we now introduce “antibodies” that target a specific sub-sequence of length $l = 5$, noted $(x_{i_1}^{ab}, \dots, x_{i_l}^{ab})$. The positions $(i_1 \dots i_l)$ are chosen at random, while the targeted sub-sequence is the dominant state at each position. Genomes that include the *exact* sub-sequence targeted by the antibody suffer a strong fitness penalty. However, a single mutation away from that sub-sequence removes this penalty completely, resulting in a fitness landscape with very strong epistasis. This has the effect of triggering a strong competition between adaptive mutations: for a given antibody, $l = 5$ possible mutations are now adaptive, but combinations of these mutations do not bring any fitness advantage.

Having simulated populations in these two fitness landscapes, we perform the same analysis of frequency trajectories as for the real influenza data. Figure S16 of the SM shows the $P_{fix}(f)$ as a function of f for the two models and for different values of the inverse rate of change Δt of the fitness landscape. For all models, this curve deviates significantly from the diagonal. This is most evident for the case of a simple additive fitness landscape that changes rarely $\Delta t = 1000$: rising mutations almost always fix in the population, with $P_{fix}(f) \simeq 1$ for any f larger than a few percent. This is corroborated by visual inspection of the trajectories, which shows that evolution in this regime is driven by regular selective sweeps that take a typical time of ~ 400 generations. In other regimes, with smaller Δt or with strong epistatic competition, $P_{fix}(f)$ is reduced and closer to the diagonal. However, it takes an extremely fast changing fitness landscape to push P_{fix} close to the diagonal: with $\Delta t = 10$, that is about 40 changes to the fitness landscape in the time it would take a selective sweep to go from 0% to fixation, $P_{fix}(f)$ differs from f in a way that is comparable to what is observed in A/H1N1pdm influenza.

These models are not meant to be accurate models of influenza viruses evolution. But figure S16 does show is that the patterns observed in influenza virus evolution are only reproduced by models of adapting populations when push-

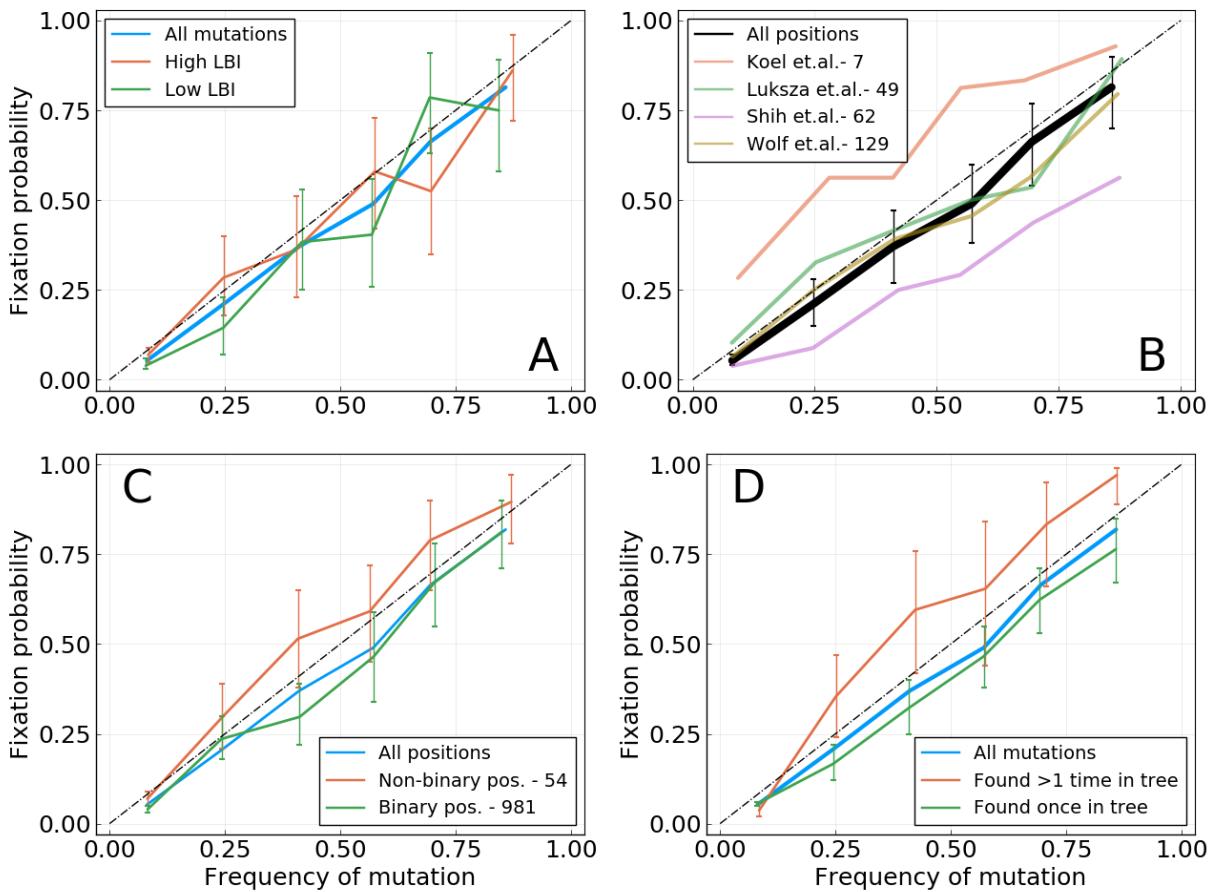


FIG. 3. Fixation probability $P_{fix}(f)$ as a function of frequency, for A/H3N2 influenza. Figure S11 shows the same analysis for A/H1N1. **A:** HA mutations with higher or lower LBI values, based on their position with respect to the median LBI value. **B:** Different lists of epitope positions in the HA protein. The authors and the number of positions is indicated in the legend. **C:** HA and NA mutations for binary positions, *i.e.* positions for which we never see more than two amino acids in the same time bin. **D:** HA and NA mutations that appear once or more than once in the tree for a given time bin.

ing clonal competition to extreme values. We conclude₄₉₀ that the pattern in figure 2 may not be a straightforward₄₉₁ manifestation of genetic linkage and clonal interference,₄₉₂ but that some more intricate interplay of epidemiology,₄₉₃ seasonality, human immunity and chance gives rise to₄₉₄ the weakly predictable yet strongly selected evolutionary₄₉₅ dynamics of IAVs.₄₉₆

497 Why do predictions work?₄₉₈

The statistics of frequency trajectories seem to be in₅₀₁ conflict with the notion that influenza evolution is pre-₅₀₂ dictable. Likewise, the LBI, a quantity that correlates₅₀₃ with fitness in mathematical models and is used to predict₅₀₄ future influenza populations [14], does not seem to contain₅₀₅ any information on whether a specific mutation is going₅₀₆ to fix or not, see figure 3. To resolve this conundrum, we₅₀₇ first note that the criterion by which predictive power for₅₀₈ influenza was measured in [14] was the distance between₅₀₉ the strain with the highest LBI and the future popula-₅₁₀

tion, not the ability of the LBI to predict dynamics. The distance was compared to the average distance between the present and future population, as well as the post-hoc optimal representative and the future.

To quantify the ability of the LBI and other measures to pick good representatives of the future, we construct a large tree of HA sequences with 100 sequences in non-overlapping time bins of 4 months from year 2003 to 2019 (a total of 4402 as some 4 month intervals contain less than 100 sequences). Each time bin is considered as a snapshot of the A/H3N2 influenza population and we will refer to sequences in time bin t as the population of the *present*. From this present population, we predict *future* populations in time bin $t + \Delta t$, using only sequences in time bin t and before.

To assess the ability of the LBI to pick a close representative of the future, we compute the LBI of each node of one time bin in the tree using only the leaves that belong to that time bin. The top panel in figure 4 shows the hamming distance of the strain with the highest LBI to future populations at different Δt along with the same

511 distance for a randomly chosen strain. The figure shows₅₆₆
512 the distance averaged over all possible values of t for Δt_{567}
513 between 0 and 32 months, giving us an average efficiency₅₆₈
514 of a predictor over 16 years of influenza evolution.₅₆₉

515 The strain with the highest LBI is consistently closer₅₇₀
516 to the future than the average strain by about 1-2 amino₅₇₁
517 acids, while the overall distance increases linearly due to₅₇₂
518 the continuous evolution of the population. We hence₅₇₃
519 reproduce previous results showing that the LBI picks₅₇₄
520 closer than average representatives [14]. To investigate₅₇₅
521 whether this apparent success is due to the ability of the₅₇₆
522 LBI to predict fitness or not, we explored a different pre₅₇₇
523 dictor: the amino acid consensus sequence of the present₅₇₈
524 population (see Methods for a definition of the consensus₅₇₉
525 sequence). The choice is motivated by the fact that it₅₈₀
526 can be shown to be the best possible long term predictor₅₈₁
527 for a neutrally evolving population in terms of Hamming₅₈₂
528 distance (see SM section 1). Figure 4 shows that the con₅₈₃
529 sensus sequence is in fact a equally good or even slightly₅₈₄
530 better representative of the future than the sequence with₅₈₅
531 highest LBI (note that the consensus sequence does *not*₅₈₆
532 necessarily exist in the population).₅₈₇

533 This near equivalence of the consensus and the strain₅₈₈
534 with highest LBI can be explained as follows: The LBI₅₈₉
535 tends to be high for nodes in a tree that are close to₅₉₀
536 the root of a dense and large clade. A typical sample₅₉₁
537 of influenza HA sequences fall into a small number of₅₉₂
538 recognizable clades, and the strains with maximal LBI₅₉₃
539 will often be close to the root of the largest of those clades₅₉₄
540 This root of the largest clade will often be close to the₅₉₅
541 consensus of the whole population, explaining the similar₅₉₆
542 distance patterns. To test that hypothesis, we measure the₅₉₇
543 hamming distance from the sequence of the top LBI strain₅₉₈
544 to the consensus sequence for populations of all time bins.₅₉₉
545 Panel **B** of figure 4 shows these distances, scaled with₆₀₀
546 respect to an average strain (details in caption). It clearly₆₀₁
547 shows that the top-LBI strain and the consensus sequence₆₀₂
548 are indeed quite similar: out of 48 time bins, only once₆₀₃
549 is the sequence of the top-LBI strain farther away from₆₀₄
550 the consensus than the average sequence is. Moreover,₆₀₅
551 the sequence of the top-LBI strain *exactly* matches the₆₀₆
552 consensus in 19 cases.₆₀₇

517 predict influenza evolution work primarily because they
518 pick strains that represent the future well, not because
519 they predict future dynamics.

520 The primary focus in this work was the investigation
521 of frequency trajectories of new amino acid mutations. In
522 the short term, we found that on average the direction of
523 trajectory does not persist for longer than a few months.
524 Indeed, the average trajectory in figure 1 takes a sharp
525 turn when going from $t < 0$ to $t > 0$, instead of showing
526 “inertia”. This suggests that selective sweeps are not
527 representative of typical trajectories.

528 On a longer timescale, we investigated the probability
529 that a novel mutation observed at frequency f fixes. In
530 neutral models of evolution this probability equals f , while
531 it should be higher or lower than f for mutations with
532 a beneficial or deleterious effect on fitness, respectively.
533 However, in the case of influenza, this probability differs
534 little from f , making current frequency the best predictor
535 for fixation. In figure 3, we split trajectories into groups
536 for which we expected P_{fix} to deviate from f . Many of
537 these splits, such as high/low LBI or epitope/non-epitope
538 positions, did not result in an increased predictability,
539 while others gave limited information on fixation. Despite
540 the lack of predictability of mutation frequency trajectories,
541 influenza surface proteins show strong signatures of
542 selection [4, 15].

543 Methods for predicting the future evolution of influenza
544 either construct explicit fitness models [11, 22], use historical
545 patterns of evolution [11, 23], phenotypic assays
546 [13, 24], or dynamic or phylogenetic patterns [14, 25]. The
547 goal of these methods is to pick strains that are good
548 representatives of future populations and could serve as
549 vaccine candidates [6].

550 The low power to predict frequency dynamics or fixation
551 naturally triggers the question why the above methods
552 have been found to work. Picking representatives of the
553 future and predicting frequency dynamics are distinct
554 objectives and success at the former (as compared to ran-
555 dom picks) is not necessarily inconsistent with a lack of
556 predictable dynamics. In fact, [22] reports that the rate
557 at which the frequency of a strain changes is often a poor
558 predictor – consistent with our observations here. But
559 despite the fact that future frequencies are not predicted
560 by the LBI, the strain with the highest LBI in the pop-
561 ulation is a better predictor of the future population than
562 a randomly picked one. While the LBI was shown to be a
563 correlate of relative fitness and be predictive of fixation in
564 mathematical models of evolution [14], it does not seem
565 to be predict influenza evolution because it measures fit-
566 ness from genealogical structure. Instead, we believe it
567 picks closer than average strains simply because it has the
568 tendency to be maximal at the base of large and dense
569 clades. These basal genotypes are closer to the future
570 populations than the current tips of the tree and hence
571 a better predictor on average. The consensus sequence
572 of all present strains performs slightly but consistently
573 better than picking the strain with the highest LBI. The
574 consensus sequence is the best possible predictor for a

DISCUSSION

554 Predicting the trajectory of a mutation requires (i)₆₁₂
555 significant fitness difference between genomes carrying₆₁₃
556 different variants at the site and (ii) a selection pressure₆₁₄
557 that changes slowly over time. Under such conditions, it is₆₁₅
558 expected that frequency trajectories will show a persistent₆₁₆
559 behavior which would make them predictable for some₆₁₇
560 time. However, we could find only limited evidence for₆₁₈
561 such persistent behavior in the past 19 years of IAV₆₁₉
562 evolution. This lead us to conclude that (i) influenza₆₂₀
563 virus evolution is qualitatively different from models of₆₂₁
564 rapidly adapting population (despite clear evidence for₆₂₂
565 frequent positive selection), and (ii) previous methods to₆₂₃

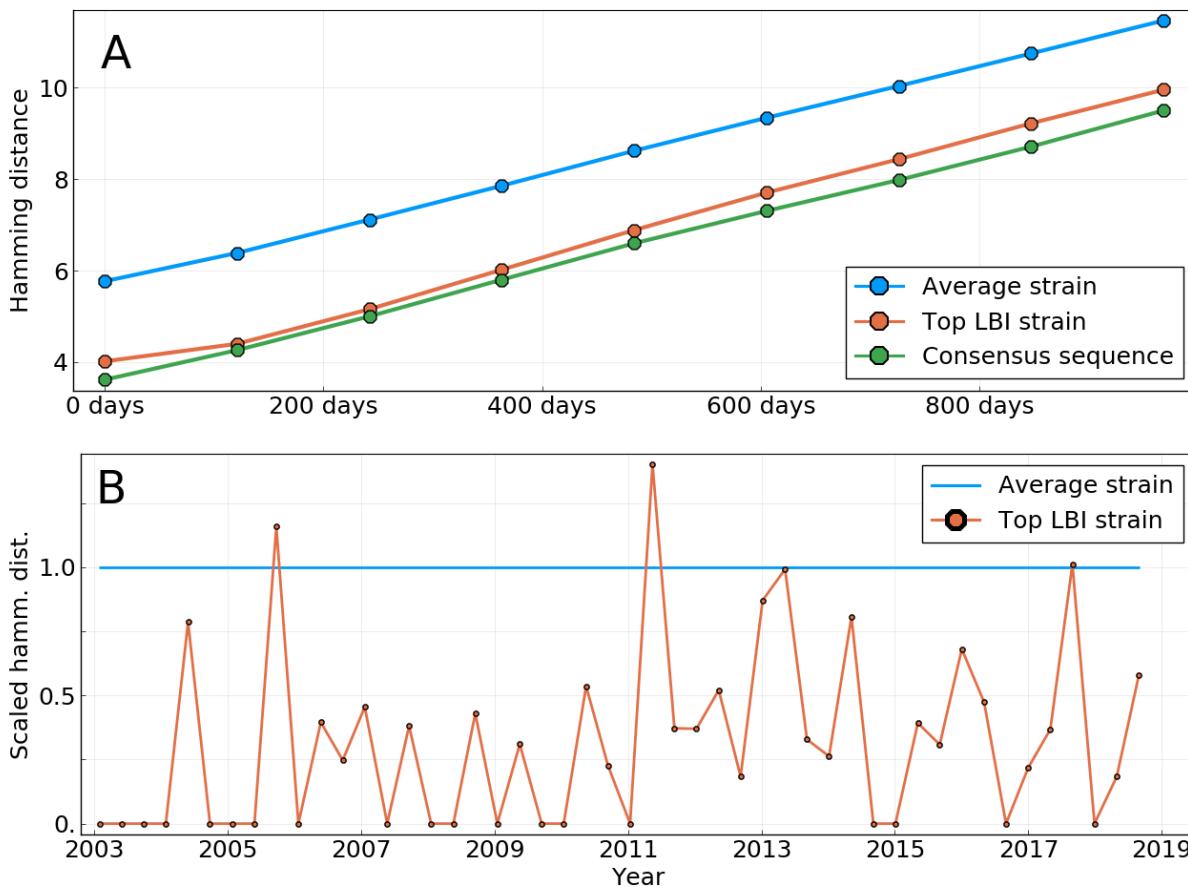


FIG. 4. **A:** Average Hamming distance of the sequences of different predictors to HA sequences of future influenza populations, themselves averaged over all “present” populations from year 2003 to 2019. Predictors are: a randomly picked sequence in the present population; the sequence of the strain with the highest LBI in the present population; the consensus sequence of the present population. **B:** Scaled Hamming distance between the sequence of the top LBI strain and the consensus sequence for populations at different dates. The scaling is such that for each date, the Hamming distance between a strain from the population and the consensus is on average 1. The strain with the highest LBI is almost always closer to the consensus sequence than the average strain.

neutrally evolving population, and does not attempt to model fitness in any way.

At the same time, influenza virus phylogenies show clear deviations from those expected from the neutral Kingman coalescent, similar to those expected under Bolthausen-Sznitman coalescent (BSC) processes that are generated by traveling wave models of rapid evolution [26, 27]. The correspondence between the BSC and traveling wave models comes from transient exponential amplification of fit strains before these fitness differences are wiped out by further mutation. This exponential amplification generates long-tailed effective offspring distributions which in turn can lead to genealogies described by the BSC [26, 28]. Many processes other than selection, including seasonality and spatio-temporal heterogeneity, can generate effective long tailed offspring distributions even in absence of bona-fide fitness differences, which might explain ladder-like non-Kingman phylogenetic trees.

A recent preprint proposed that influenza virus evo-

lution is primarily limited by an asynchrony between population level selection and generation of new variants within infected hosts [29]. Along these lines, it is possible that the A/H3N2 population readily responds once population level selection is high enough by giving rise to essentially equivalent variants. Furthermore, selection might cause the rapid rise of a novel variant to macroscopic frequencies (observable in a global sample) but its benefit rapidly “expires” because competing variants catch up and/or it mediates immune escape only to a small fraction of the population. These considerations might explain the disconnect between models of rapid adaptation and the frequency dynamics observed in influenza virus populations.

657

METHODS

658

Data and code availability

659

The sequences used are obtained from the GISAID⁷¹⁰ database [9]. Strain names and accession numbers are⁷¹¹ given as tables in two supplementary files. Outliers⁷¹² strains listed at <https://github.com/PierreBarrat/FluPredictability/src/config>⁷¹³ were removed.⁷¹⁴ The code used to generate the figures presented here⁷¹⁵ is available at <https://github.com/PierreBarrat/FluPredictability>.⁷¹⁶

667

Frequency trajectories

668

For a set of sequences in a given time bin, we compute frequencies of amino acids at each position by simple counting. We make the choice of not applying any smoothing method in an attempt to be as close to the data and “model-less” as possible. This is especially important for the short term prediction of frequency trajectories, as estimations of the “persistence time” of a trajectory might be biased by a smoothing method.

669

We compute frequency trajectories based on the frequencies of amino acids. A trajectory begins at time t if an amino acid is seen under the lower frequency threshold of 5% (resp. above the higher threshold of 95%) for the two time bins preceding t , and above this lower threshold (resp. below the higher threshold) for time bin t . It ends in the reciprocal situation, that is when the frequency is measured below the lower threshold (resp. above the higher threshold) for two time bins in a row.

670

In order to avoid estimates of frequencies that are too noisy, we only keep trajectories that are based on a population of at least 10 sequences for each time bin. As said in the Results section, we also restrict the analysis to trajectories that begin at a 0 frequency, in part to avoid double counting. We find a total of 460 such trajectories. However, only 106 reach a frequency of 20%, on which figure 2 is based for instance.

671

Note that the fact that we use samples of relatively small sizes – at least for some time bins – leads to biases in the estimation of frequencies. We show in Supplementary Material that these biases are generally small and do not induce any qualitative changes to results presented here.

698

Local Branching Index

699

LBI was introduced in [14] as an approximation of fitness in populations evolving under persistent selective pressure that is fully based on a phylogenetic tree. It relies on the intuition that the tree below high-fitness individuals will show dense branching events, whereas absence of branching is a sign of low-fitness individuals. Quantitatively, the LBI $\lambda_i(\tau)$ of a node i is the integral

706

of all of the tree’s branch length around i , with an exponentially decreasing weight $e^{-t/\tau}$ with t being the branch length. When considering a time binned population, the LBI is computed once for each time bin by considering only the leaves of the tree that belong to the time bin. This means that only branches that ultimately lead to a leaf that belongs to the time bin are considered in the integration.

707

τ is the time scale for which the tree is informative of the fitness of a particular node. Here, we use a value of τ equal to a tenth of $T_C \simeq 6$ years, the coalescence time for influenza A/H3N2 strains, converted to units of tree branch length through the average nucleotide substitution rate ($\simeq 4 \cdot 10^{-3}$ substitutions per site per year for HA). We have observed that given our method to predict the future from present populations corresponding to time bins of 4 months, changing the value of τ has little effect on the pick of the top LBI strain. By retrospectively optimizing its value, it is possible to reduce the average distance to the population 2 years ahead by ~ 0.25 amino acids on average, making the LBI method almost as good as the consensus on figure 4.

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

Measuring the geographical spread of a mutation

For a mutation X we define its regional distribution using the numbers $n_r(X)$ that represent the number of sequences sampled in region r that carry X . Regional weights are then defined as

$$w_r(X) = \frac{n_r(X)}{\sum_r n_r(X)}.$$

We can then measure the geographical spread $G(X)$ of X by using the Shannon entropy of the probability distribution $w_r(X)$:

$$G(X) = \sum_r w_r(X) \log(w_r(X)).$$

$G(X)$ is a positive quantity that is larger when X is equally present in many regions, and equal to zero when X is concentrated in only one region.

Region used are the ones defined in the Nextstrain tool [30]. Those are North America, South America, Europe, China, Oceania, Southeast Asia, Japan & Korea, South Asia, West Asia, and Africa.

734

735

Assigning a fitness to trajectories

Consensus sequence

Given a set of N sequences $(\sigma^1, \dots, \sigma^N)$ based on an alphabet \mathcal{A} (e.g. \mathcal{A} has 20 elements for amino acids, 4 for nucleotides), we can define a profile distribution $p_i(a)$

by the following expression:

$$p_i(a) = \sum_{n=1}^N \delta_{\sigma_i^n, a}$$

where i is a position in the sequence, σ_i^n the character appearing at position i in sequence σ^n , a a character of the alphabet and δ the Kronecker delta. The profile $p_i(a)$ simply represents the fraction of sequences which have character a at position i .

We then simply define the consensus sequence σ^{cons} such that

$$\sigma_i^{cons} = \operatorname{argmax}_a p_i(a).$$

738 In other words, the consensus sequence is the one that
 739 has the dominant character of the initial set of sequences
 740 at each position.

741 Earth Mover's Distance

742 In order to measure the distance of several predictor
 743 sequences to the future population, we rely on the *Earth
 744 Mover's Distance* (EMD), a metric commonly applied
 745 in machine learning to compare collections of pixels or
 746 words [31, 32]. Here, we apply it to compute the dis-
 747 tance between the sequences of two populations, noted as
 750

$\mathcal{X} = \{(x^n, p^n)\}$ and $\mathcal{Y} = \{(y^m, q^m)\}$ with $n \in \{1 \dots N\}$ and $m \in \{1 \dots M\}$. In this notation, x^n and y^m are sequences, and p^n and q^m are the frequencies at which these sequences are found in their respective populations. For convenience, we also define $d_{nm} = H(x^n, y^m)$ as the Hamming distance between pairs of sequences in the two populations.

We now introduce the following functional

$$F(\mathbf{w}) = \sum_{n,m} d_{nm} w_{nm},$$

with $\mathbf{w} = \{w_{nm}\}$ being a matrix of positive weights. The EMD between the two populations \mathcal{X} and \mathcal{Y} is now defined as the minimum value of function F under the conditions

$$\sum_{n=1}^N w_{nm} = q^m, \quad \sum_{m=1}^M w_{nm} = p^n, \text{ and } w_{nm} \geq 0$$

742 Intuitively, the weight w_{nm} tells us how much of sequence
 743 x^n is “moved” to sequence y^m . The functional F sums
 744 all of these moves and attributes them a cost equal to
 745 the Hamming distance d_{nm} . The conditions on weights
 746 in \mathbf{w} ensure that all the weight p^n of x^n is “moved” to
 747 elements in \mathcal{Y} and vice versa.

The minimization is easily performed by standard linear optimization libraries. Here, we use the Julia library JuMP [33].

- 751 [1] World Health Organization, 2018. URL [https://www.who.int/fr/news-room/fact-sheets/detail/influenza-\(seasonal\)](https://www.who.int/fr/news-room/fact-sheets/detail/influenza-(seasonal)).
 752
 753
- 754 [2] Velislava N. Petrova and Colin A. Russell. The evolution
 755 of seasonal influenza viruses. *Nature Reviews Microbiology*,
 756 16(1):47–60, October 2017. ISSN 1740-1526, 1740-1534.
 757 doi:10.1038/nrmicro.2017.118. URL <http://www.nature.com/doifinder/10.1038/nrmicro.2017.118>.
 758
- 759 [3] Arthur Chun-Chieh Shih, Tzu-Chang Hsiao, Mei-Shang
 760 Ho, and Wen-Hsiung Li. Simultaneous amino acid sub-
 761 stitutions at antigenic sites drive influenza a hemagglu-
 762 tinin evolution. *Proceedings of the National Academy of Sciences*,
 763 104(15):6283–6288, 2007. ISSN 0027-8424.
 764 doi:10.1073/pnas.0701396104. URL <https://www.pnas.org/content/104/15/6283>.
 765
- 766 [4] Samir Bhatt, Edward C. Holmes, and Oliver G. Pybus
 767 The Genomic Rate of Molecular Adaptation of the Human
 768 Influenza A Virus. *Molecular Biology and Evolution*, 28:
 769 (9):2443–2451, September 2011. ISSN 0737-4038. doi:
 770 10.1093/molbev/msr044. URL <https://academic.oup.com/mbe/article/28/9/2443/1007907>.
 771
- 772 [5] Björn F. Koel, David F. Burke, Theo M. Bestebroer,
 773 Stefan van der Vliet, Gerben C. M. Zondag, Gaby Vervaet,
 774 Eugene Skepner, Nicola S. Lewis, Monique I. J. Spronken,
 775 Colin A. Russell, Mikhail Y. Eropkin, Aeron C. Hurt,
 776 Ian G. Barr, Jan C. de Jong, Guus F. Rimmelzwaan,
 777 Albert D. M. E. Osterhaus, Ron A. M. Fouchier, and
 778 Derek J. Smith. Substitutions near the receptor binding
 779 site determine major antigenic change during influenza
 780 virus evolution. *Science*, 342(6161):976–979, 2013. ISSN
 0036-8075. doi:10.1126/science.1244730. URL <https://science.sciencemag.org/content/342/6161/976>.
 781
- 782 [6] Dylan H Morris, Katelyn M Gostic, Simone Pompei,
 783 Trevor Bedford, Marta Luksza, Richard A Neher, Bryan T
 784 Grenfell, Michael Lässig, and John W McCauley. Predictive
 785 modeling of influenza shows the promise of applied
 786 evolutionary biology. *Trends in microbiology*, 26(2):102–
 787 118, 2018.
- 788 [7] Thorsten R. Klingen, Susanne Reimering, Carlos A.
 789 Guzmán, and Alice C. McHardy. In Silico Vaccine
 790 Strain Prediction for Human Influenza Viruses. *Trends
 791 in Microbiology*, 26(2):119–131, February 2018. ISSN
 0966-842X. doi:10.1016/j.tim.2017.09.001. URL
 792 <http://www.sciencedirect.com/science/article/pii/S0966842X17302068>.
 793
- 794 [8] Peter Bogner, Ilaria Capua, David J Lipman, and Nancy J
 795 Cox. A global initiative on sharing avian flu data. *Nature*,
 796 442(7106):981–981, 2006.
- 797 [9] Yuelong Shu and John McCauley. Gisaid: Global initia-
 798 tive on sharing all influenza data—from vision to reality.
 799 *Eurosurveillance*, 22(13), 2017.
- 800 [10] Andrew Rambaut, Oliver G. Pybus, Martha I. Nelson,
 801 Cecile Viboud, Jeffery K. Taubenberger, and Edward C.
 802 Holmes. The genomic and epidemiological dynamics of
 803 human influenza A virus. *Nature*, 453(7195):615–619, May
 804 2008. ISSN 1476-4687. doi:10.1038/nature06945. URL
 805 <https://www.nature.com/articles/nature06945>.
 806

- [11] Marta Luksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507(7490):57–61, March 2014. ISSN 1476-4687. doi:10.1038/nature13087. URL <https://www.nature.com/articles/nature13087>. Number: 7490. Publisher: Nature Publishing Group.
- [12] L. Steinbrück, T. R. Klingen, and A. C. McHardy. Computational prediction of vaccine strains for human influenza a (h3n2) viruses. *Journal of Virology*, 88(20):12123–12132, 2014. ISSN 0022-538X. doi:10.1128/JVI.01861-14. URL <https://jvi.asm.org/content/88/20/12123>.
- [13] Richard A. Neher, Trevor Bedford, Rodney S. Daniels, Colin A. Russell, and Boris I. Shraiman. Prediction dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 113(12):E1701–1709, March 2016. ISSN 1091-6490 0027-8424. doi:10.1073/pnas.1525578113.
- [14] Richard A Neher, Colin A Russell, and Boris I Shraiman. Predicting evolution from the shape of genealogical trees. *eLife*, 3, November 2014. ISSN 2050-084X. doi:10.7554/eLife.03568. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4227306/>.
- [15] Natalja Strelkowa and Michael Lässig. Clonal Interference in the Evolution of Influenza. *Genetics*, 192(2):671–682, October 2012. ISSN 0016-6731, 1943-2631. doi:10.1534/genetics.112.143396. URL <http://www.genetics.org/content/192/2/671>.
- [16] Yuri I Wolf, Cecile Viboud, Edward C Holmes, Eugene V Koonin, and David J Lipman. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biology Direct*, 1:34, October 2006. ISSN 1745-6150. doi:10.1186/1745-6150-1-34. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1647279/>.
- [17] Juhye M Lee, Rachel Eguia, Seth J Zost, Saket Choudhary, Patrick C Wilson, Trevor Bedford, Terry Stevens-Ayers, Michael Boeckh, Aeron C Hurt, Seema S Lakdawala, Scott E Hensley, and Jesse D Bloom. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *eLife*, 8:e49324, August 2019. ISSN 2050-084X. doi:10.7554/eLife.49324. URL <https://doi.org/10.7554/eLife.49324>. Publisher: eLife Sciences Publications, Ltd.
- [18] Motoo Kimura. Diffusion Models in Population Genetics. *Journal of Applied Probability*, 1(2):177–232, 1964. ISSN 0021-9002. doi:10.2307/3211856. URL <http://www.jstor.org/stable/3211856>.
- [19] Le Yan, Richard A Neher, and Boris I Shraiman. Phylogenetic theory of persistence, extinction and speciation of rapidly adapting pathogens. *eLife*, 8:e44205, September 2019. ISSN 2050-084X. doi:10.7554/eLife.44205. URL <https://doi.org/10.7554/eLife.44205>. Publisher: eLife Sciences Publications, Ltd.
- [20] R. A. Neher and B. I. Shraiman. Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics*, 188(4):975–996, August 2011. ISSN 1943-2631. doi:10.1534/genetics.111.128876.
- [21] Fabio Zanini and Richard A. Neher. FFPopSim: an efficient forward simulation package for the evolution of large populations. *Bioinformatics*, 28(24):3332–3333, 10 2012. ISSN 1367-4803. doi:10.1093/bioinformatics/bts633. URL <https://doi.org/10.1093/bioinformatics/bts633>.
- [22] John Huddleston, John R. Barnes, Thomas Rowe, Rebecca Kondor, David E. Wentworth, Lynne Whittaker. Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. *bioRxiv*, page 2020.06.12.145151, June 2020. doi:10.1101/2020.06.12.145151. URL <https://www.biorxiv.org/content/10.1101/2020.06.12.145151v1>. Publisher: Cold Spring Harbor Laboratory Section: New Results.
- [23] R. M. Bush, C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch. Predicting the evolution of human influenza A. *Science (New York, N.Y.)*, 286(5446):1921–1925, December 1999. ISSN 0036-8075.
- [24] Lars Steinbrück and Alice Carolyn McHardy. Inference of Genotype–Phenotype Relationships in the Antigenic Evolution of Human Influenza A (H3N2) Viruses. *PLOS Computational Biology*, 8(4):e1002492, April 2012. ISSN 1553-7358. doi:10.1371/journal.pcbi.1002492. URL <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002492>.
- [25] Thorsten R. Klingen, Susanne Reimering, Jens Loers, Kyra Mooren, Frank Klawonn, Thomas Krey, Gülsah Gabriel, and Alice C. McHardy. Sweep Dynamics (SD) plots: Computational identification of selective sweeps to monitor the adaptation of influenza A viruses. *Scientific Reports*, 8(1):373, January 2018. ISSN 2045-2322. doi:10.1038/s41598-017-18791-z. URL <https://www.nature.com/articles/s41598-017-18791-z>.
- [26] Richard A. Neher and Oskar Hallatschek. Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences of the United States of America*, 110(2):437–442, January 2013. ISSN 1091-6490 0027-8424. doi:10.1073/pnas.1213113110.
- [27] Michael M. Desai, Aleksandra M. Walczak, and Daniel S. Fisher. Genetic Diversity and the Structure of Genealogies in Rapidly Adapting Populations. *Genetics*, 193(2):565–585, February 2013. ISSN 0016-6731, 1943-2631. doi:10.1534/genetics.112.147157. URL <http://www.genetics.org/content/193/2/565>.
- [28] Jason Schweinsberg. Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Processes and their Applications*, 106(1):107–139, July 2003. ISSN 0304-4149. doi:10.1016/S0304-4149(03)00028-0. URL <http://www.sciencedirect.com/science/article/pii/S0304414903000280>.
- [29] Dylan H. Morris, Velislava Petrova, Fernando W. Rossine, Edyth Parker, Bryan Grenfell, Richard Neher, Simon Levin, and Colin Russell. Asynchrony between virus diversity and antibody selection limits influenza virus evolution. preprint, Open Science Framework, April 2020. URL <https://osf.io/847p2>.
- [30] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagunenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 05 2018. ISSN 1367-4803. doi:10.1093/bioinformatics/bty407. URL <https://doi.org/10.1093/bioinformatics/bty407>.
- [31] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, Jan 1998. doi:

- 936 10.1109/ICCV.1998.710701. 944
- 937 [32] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document 945
938 distances. In *Proceedings of the 32nd International Conference on Machine Learning* 947
940 - Volume 37, ICML'15, page 957–966. JMLR.org, 2015. 949
- 941 [33] Iain Dunning, Joey Huchette, and Miles Lubin. Jump: A 950
943 modeling language for mathematical optimization. *SIAM* 951
952
- Review, 59(2):295–320, 2017. doi:10.1137/15M1020575.
- [34] Julia Sigwart. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.—Jotun Hein, Mikkel H. Schierup, and Carsten Wiuf. 2004. Oxford University Press, Oxford. xiii + 276 pp. ISBN 0-19-852996-1, £29.95 (paperback); ISBN 0-19-852995-3, £65.00 (hardback). *Systematic Biology*, 54(6):986–987, 12 2005. ISSN 1063-5157. doi:10.1080/10635150500354860. URL <https://doi.org/10.1080/10635150500354860>.

SUPPLEMENTARY MATERIAL

1. Consensus sequence as a predictor for neutrally evolving populations

We consider the case of a neutrally evolving and structure-less population, such as the one in the Wright-Fisher model of evolution [34]. At an initial time $t = 0$, the population consists of N individuals with genomes $(\sigma^1 \dots \sigma^N)$ of length L (not necessarily distinct).

We make two hypotheses about this population. We first suppose that *no* mutations occur during the evolution of this population. This may seem surprising and is of course not true in the case of influenza. This assumption is however in line with the fact that the object of this work is to predict the outcome of *already existing* mutations in the influenza population. The prediction of mutations that we have not yet seen is not in its scope. Thus, assuming that no new mutations take place can be seen as a simple way to model the fact that we have no information about such events. The second assumption is that the population evolves in a completely neutral way, meaning that the average number of descendants of each genome σ^n is the same. Let us now consider the population after it has evolved for a long time $t \gg T$ where T is the typical coalescence time (for the Wright-Fisher model, $T = 2N$). At this point, all individuals in the future population will descend from a unique individual n_0 in the $t = 0$ population. Our two hypotheses now allow us to make two statements. First, since no new mutations are allowed, the population at $t \gg T$ will be clonal, with all individuals having genome σ^{n_0} . Second, since the evolution is neutral and does not favour any genome in particular, the probability that σ^{n_0} is equal to a given genome σ is $1/N$. In other words, the probability that a genome at $t = 0$ ultimately becomes the ancestor of all the future population is equal to its frequency in the $t = 0$ population.

We now try to find the genome σ that best predicts the future population on the long run, that is for $t \gg T$. Here, we take best to mean that the predictor minimizes $H(\sigma, \sigma^{n_0})$ where H is the Hamming distance defined by

$$H(\sigma^a, \sigma^b) = \sum_{i=1}^L (1 - \delta_{\sigma_i^a, \sigma_i^b}), \quad (1)$$

with σ_i being the character appearing at position i of genome σ and δ the Kronecker delta. Since we do not know n_0 , we have to average over all its possible values. σ must thus minimize the following quantity:

$$\begin{aligned} \langle H(\sigma, \sigma^{n_0}) \rangle_{n_0} &= \sum_{n=1}^N H(\sigma, \sigma^n) \\ &= \sum_{i=1}^L \sum_{n=1}^N (1 - \delta_{\sigma_i, \sigma_i^n}) \end{aligned} \quad (2)$$

by using the definition of the Hamming distance. We now assume that characters at each positions of the genomes can be indexed by an integer a running from 1 to q . For instance, if these were amino acid sequences, we could index the 20 amino acids by a running from 1 to $q = 20$. We rewrite the Kronecker delta in the previous expression using this indexation:

$$\delta_{\sigma_i, \sigma_i^n} = \sum_{a=1}^q \delta_{\sigma_i, a} \delta_{\sigma_i^n, a}.$$

We also introduce the *profile* frequencies $p_i(a)$ of the population at time $t = 0$:

$$p_i(a) = \sum_{n=1}^N \delta_{\sigma_i^n, a}. \quad (3)$$

977 $p_i(a)$ represents the frequency at which character a appears at position i in genomes of the initial population.
 978 Equation 2 now becomes

$$\begin{aligned}\langle H(\sigma, \sigma^{n_0}) \rangle_{n_0} &= \sum_{i=1}^L \sum_{n=1}^N \left(1 - \sum_{a=1}^q \delta_{\sigma_i, a} \delta_{\sigma_i^n, a} \right) \\ &= \sum_{i=1}^N \left(1 - \sum_{a=1}^q \delta_{\sigma_i, a} p_i(a) \right) \\ &= \sum_{i=1}^L (1 - p_i(\sigma_i))\end{aligned}\quad (4)$$

979 This means that the genome $\sigma = (\sigma_1 \dots \sigma_L)$ which best predicts the future population according to our definition is
 980 the one that minimizes the quantity $(1 - p_i(\sigma_i))$ for all positions i . This obviously implies that each σ_i must be chosen
 981 as to maximize $p_i(a)$, that is σ_i must be the character that appears the most frequently at position i . Thus, σ must be
 982 the *consensus* sequence of the initial population.

983 2. Predictor based on the local LBI maxima

In figure 15, we use several sequences as a predictor of the future population. Distance between two sets of sequences, *i.e.* the predictor sequences and the ones of the future population, is defined as the Earth Mover's Distance (EMD). Here, we show that for a population evolving under the same hypotheses as in section 1, the best *multiple* sequence long term predictor is again the consensus sequence with weight 1.

Let the predictor be a set of weighted sequences $\{(s^\alpha, q_\alpha)\}$. We again use the fact that in the long term, a unique sequence σ^{n_0} from the present will be the ancestor of the entire population. We want to compute the EMD from the predictor to σ^{n_0} , that is the EMD between the sets $\mathcal{X} = \{(s^\alpha, q_\alpha)\}$ and $\mathcal{Y} = \{\sigma^{n_0}, 1\}$. Applying the definition of the Methods section, it follows that the weights \mathbf{w} are in this case equal to the q_α s. By averaging over all values of n_0 , we now obtain

$$\langle \text{EMD}(\{(s^\alpha, q_\alpha)\}) \rangle_{n_0} = \sum_{n=1}^N \sum_{\alpha} H(s^\alpha, \sigma^n) \cdot q_\alpha.$$

By the same calculation procedure as in the previous section, this expression simplifies to

$$\langle \text{EMD}(\{(s^\alpha, q_\alpha)\}) \rangle_{n_0} = \sum_{i=1}^L \left(1 - \sum_{a=1}^q p_i(a) q_i(a) \right),$$

where the profile of the present population $p_i(a)$ has already been defined, and $q_i(a)$ stands for the profile of the predictor, that is

$$q_i(a) = \sum_{\alpha} \delta_{s_i^\alpha, a} q_\alpha.$$

984 To minimize this distance, we find a profile $q_i(a)$ that maximizes the quantity $\sum_{\alpha} \delta_{s_i^\alpha, a} q_\alpha$ for each position i . It is
 985 clear that this is done by assigning a value $q_i(a) = 1$ if a maximizes $p_i(a)$, and $q_i(a) = 0$ otherwise. Thus, the profile of
 986 the predictor must be that of the consensus sequence, which is only possible if the predictor becomes $\{\sigma^{cons}, 1\}$.

987 3. Biases in frequency estimations

988 The frequency of mutations in a given time-bin is simply performed by computing their frequency in sequences
 989 sampled in that time bin. This leads to potential biases in estimating frequencies, that arise for two reasons:

- 990 (i) A mutation present at frequency p in the population might be observed at another frequency $f \neq p$ if f is
 991 estimated using a sub-sample of the population.
- 992 (ii) For a neutrally evolving population, the distribution of frequencies of alleles is of the form $P(p) \propto 1/p$. This
 993 means that the amount of alleles at frequency p is lower when p is higher.

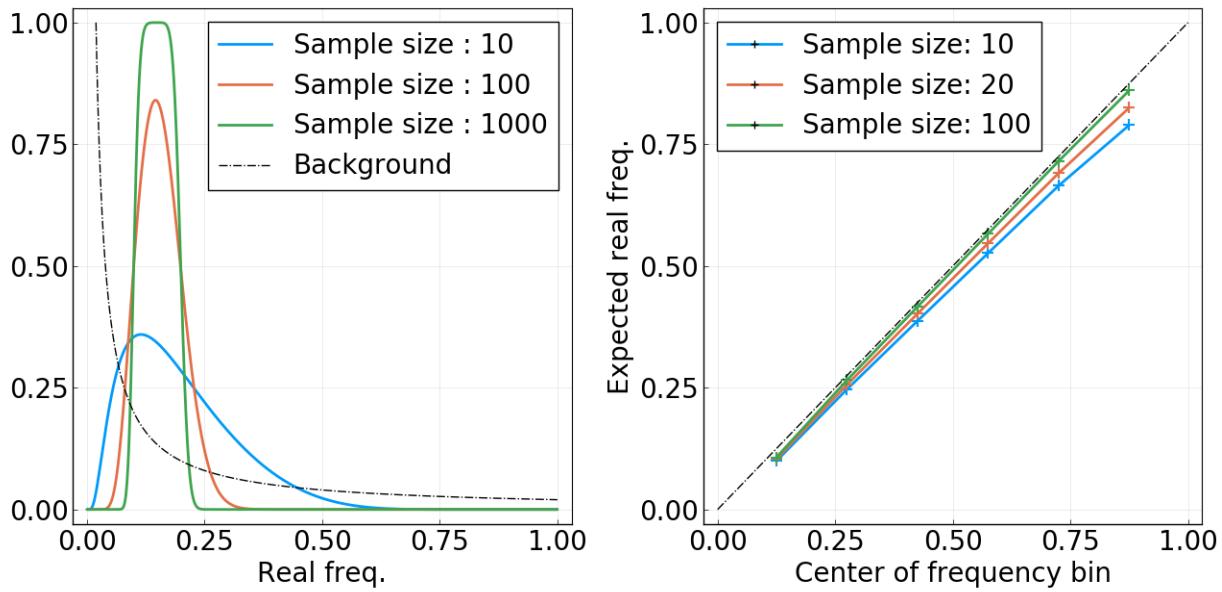


Figure S 1. **Left:** For a mutation present at frequency p in the population, probability of being observed in the frequency bin $[0.1, 0.2]$ as a function of p and for different sample sizes n . The dashed black line sketches the (non-normalized) background distribution $P_b(p)$. **Right:** Expected “real” average frequency of mutations found in frequency bin $[f_1, f_2]$ as a function of the centre of the bin $(f_1 + f_2)/2$, for different sample sizes.

994 To illustrate (i), let us compute the probability that a mutation present at “real” frequency p in the population is
995 found to be in a given frequency bin $[f_1, f_2]$ when p is estimated from a sample of size n . The sample consists of n
996 observations $\{x_i\}$ with $1 \leq i \leq n$, with $x_i = 1$ if sequence sequence i of the sample bears the mutation, and $x_i = 0$ if
997 not. If n is small with regard to the total population size, we can consider the x_i as random variables with a binomial
998 distribution, meaning that $P(x_i = 1) = p$ and $P(x_i = 0) = 1 - p$. The empirical frequency f is then estimated by
999 taking the average of the x_i variables, that is $f = (x_1 + \dots + x_n)/n$. If those are independently sampled and n is large
1000 enough, the probability of measuring value f is given by the Central Limit Theorem:

$$P_{n,p}(f) \propto e^{(f-p)^2/2\sigma^2}, \text{ where } \sigma^2 = \frac{p(1-p)}{n}. \quad (5)$$

1001 To compute the probability that this mutation is found in a given frequency bin $[f_1, f_2]$, we integrate this distribution:

$$P_{f_1, f_2}(p, n) = \int_{f_1}^{f_2} dx P_{n,p}(x). \quad (6)$$

1002 Function $P_{f_1, f_2}(p, n)$ is shown as a function of p for a fixed interval and for different values of n in the first panel of
1003 figure S1. Note the asymmetry of it: the variance of a binomial distribution of parameter p is small when p is close to
1004 0 or 1, and goes through a maximum at $p = 0.5$. For this reason, mutations present at frequency p close to 0.5 have a
1005 higher probability of being observed in other frequency bins. On the contrary, this is unlikely for very rare or very
1006 frequent mutations.

1007 We now try to estimate biases in frequency estimation due this phenomenon. Given a set of mutations that have
1008 been measured in frequency bin $[f_1, f_2]$, what is the average *real* frequency of these mutations? To compute this, we
1009 need to sum $P_{f_1, f_2}(p, n)$ over all possible real frequencies p , giving us the amount of mutations that are observed in
1010 interval $[f_1, f_2]$, and weigh this sum by the frequency value p as well as by the background distribution of frequencies
1011 $P_b(p) \propto 1/p$. This last quantity represents the expected amount of mutations that are present at frequency p in the
1012 population. Note that there is no divergence problem as the smallest non zero frequency is $1/N$, where N is the
1013 population size. This leads us to the following expression for the average of “real” frequencies:

$$\begin{aligned} \langle p \rangle(f_1, f_2, n) &= \int_{1/N}^{1-1/N} dp P_{f_1, f_2}(p, n) P_b(p) p \\ &= \int_{1/N}^{1-1/N} dp P_{f_1, f_2}(p, n). \end{aligned} \quad (7)$$

1016 We have not made normalization explicit in these equations. It is simply achieved by dividing the above expression by
1017 $\int dp P_{f_1, f_2}(p, n) P_b(p)$.

1018 In the second panel of figure S1, $\langle p \rangle(f_1, f_2, n)$ is plotted as a function of the centre of the interval $[f_1, f_2]$ and for
1019 different values of n . For sample sizes $n > 100$, the biases due to this effect are almost non existent. For smaller
1020 samples, for instance $n = 10$, they are small but non negligible. However, we argue that this is not a significant problem
1021 with respect to the main results presented in this article. First, figure S6 shows that sample sizes of the order of $n = 10$
1022 are only the case for a few months in the period going from year 2000 to 2018. From 2010 and onwards, more than a
1023 hundred sequences are available per month for most months. Secondly, even if most samples were in the $n = 10$ case,
1024 deviations shown in figure S1 are small enough that results shown in figures 2 and 3 would be *qualitatively* unchanged.
1025 Note that using the centre of the interval as a reference in figure S1, *i.e.* $(f_1 + f_2)/2$, would be correct in the case of
1026 a very large n and a flat background distribution $P_b(p)$. For figures 2 and 3 of the main text however, the average
1027 frequency of mutations found in an interval $[f_1, f_2]$ is computed by taking the average of the observed frequencies, and
1028 not the centre of the interval. This partially takes into account biases considered here, as the background distribution
1029 $P_b(p)$ is then accounted for, even though it is equivalent to assuming infinite sample sizes.

1030

4. Cutting off the HA1 159S branch



Figure S 2. Tree used for this study, based on a random selection of 100 strains per month from year 2002 to 2018. Nodes and branches are colored according to the amino acid found at position HA1:159. The HA1 159S mutation is visible as a thin but long light-greened color branch, coalescing with the “trunk” around 2013.

1031 The analysis of the main text is in a large part based on the probability of fixation of mutations. The motivation
1032 underlying this choice is the relatively short coalescence time of the A/H3N2 influenza population, typically around
1033 three years. This can be seen in figure 2 of the main text, which shows the typical lifetime of frequency trajectories,
1034 ending in fixation or loss after at most 3 years in most cases. The tree in figure S2 is another illustration of this: for
1035 the most part of it, a “trunk” is clearly identifiable, and lineages that depart from it have a relatively short lifetime.
1036 This is no longer the case since the year ~ 2013: two clades have been competing since then, with no definite way to
1037 identify a trunk in the tree. The clade defined by the HA1 159S mutation, colored in light green on figure S2, is one of
1038 these two competing lineages. Because of this particular situation, the number of mutations fixating in the population
1039 is strongly reduced, as a mutation must appear in both clades to reach a frequency of 1. This is a potential flaw in our
1040 analysis, which concentrates on mutations fixating.

1041 For this reason, we decided to re-run our analysis after having cut off the HA1 159S clade. In other words, we remove
1042 from the set of sequences those that carry the HA1 159S mutation. Results are shown in figures , equivalent to figures 2
1043 and 3 of the main text. It is clear that qualitative results are left unchanged when this competing clade is removed.
1044 This can be surprising, as almost no complete fixation of an amino acid mutation has occurred since 2013. Cutting off
1045 the HA1 159S branch should thus result in many new fixations, changing the analysis. The reason for the similarity
1046 of results can be explained: fixation (resp. loss) of a mutation are defined here as the frequency of this mutation
1047 being measured above 95% (resp. 5%) frequency for two months in a row. As the HA1 159S clade is rather sparsely
1048 populated, it reaches frequencies lower than 5% two times (in 2015 and 2017), allowing mutations in the competing
1049 clade to “fix” as defined here. Thus, removing strains carrying HA1 159S does not introduce a significant amount of
1050 “new” fixation events.

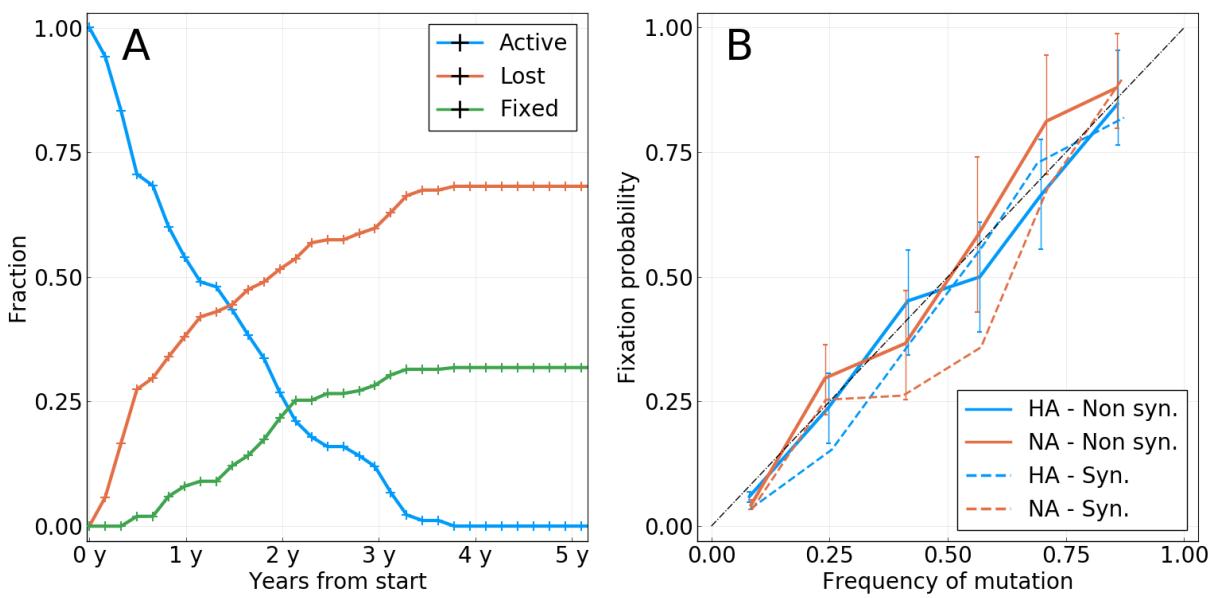


Figure S 3. Equivalent to figure 2 of the main text, but with strains carrying the HA1 159S mutation removed.

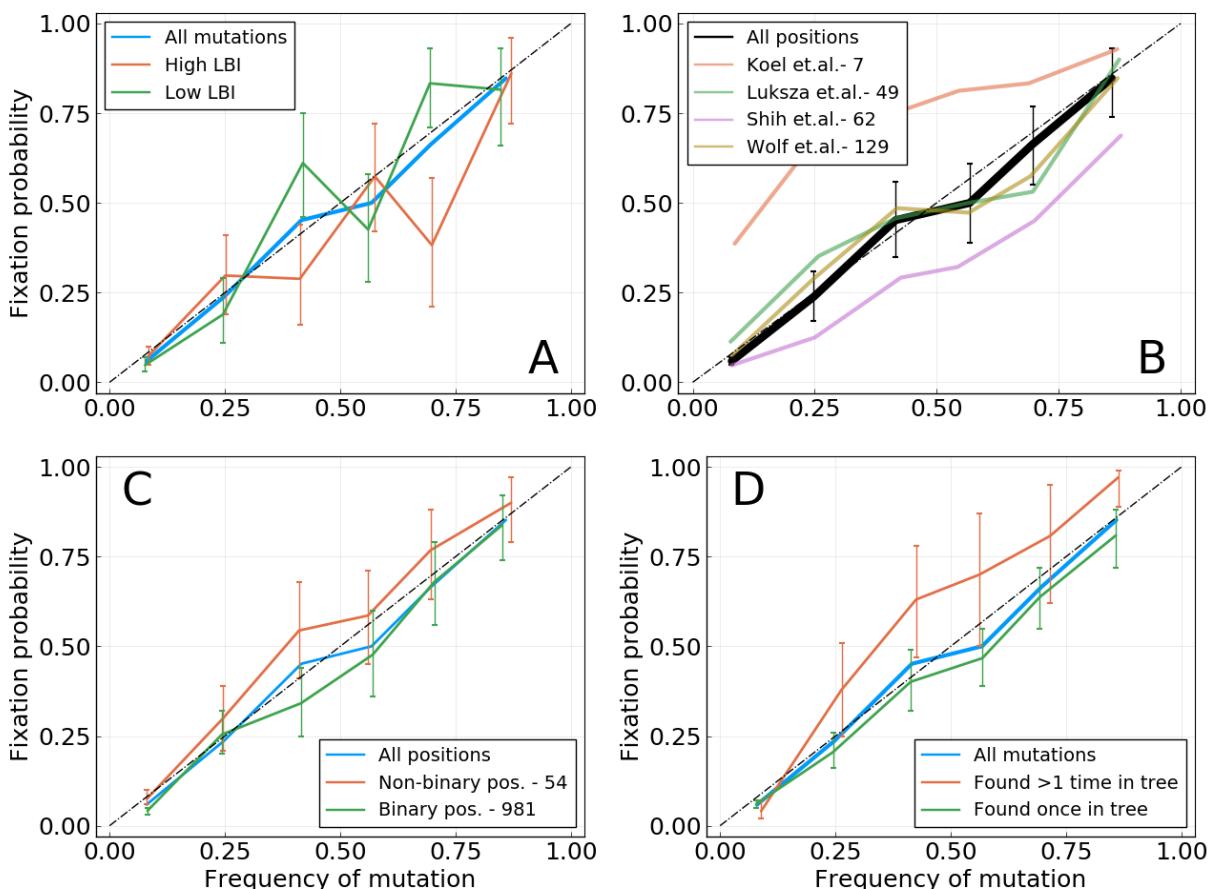


Figure S 4. Equivalent to figure 3 of the main text, but with strains carrying the HA1 159S mutation removed.

1051

5. Probability of fixation in single locus model of evolution

1052 In [18], Kimura investigates a simple model of evolution with a single locus and a population of size N . In this
1053 framework, a mutation at this locus with fitness effect s and observed at frequency f has the following probability of
1054 fixation:

$$P_{fix}(f|s, N) = \frac{1 - e^{-sNf}}{1 - e^{-sN}}. \quad (8)$$

1055 Expanding this formula for $sN \ll 1$, that is in the weak selection regime, yields at the first order

$$P_{fix}(f|s, N) = f + f(1 - f) \frac{sN}{2}. \quad (9)$$

1056 Equation 9 tells us two things. First, when the mutation is neutral, that is $s = 0$, we have $P_{fix}(f) = f$. This naturally
1057 confirms the result obtained for a neutral model of evolution. Seconds, when $sN \neq 0$, we can expect deviations from
1058 the diagonal in a P_{fix} against f plot. The sign of these deviations is determined by the sign of s , with beneficial
1059 mutations being found above diagonal while deleterious one are found below. The amplitude of these deviations
1060 depends on the strength of selection sN , as well as on the frequency through the $f(1 - f)$ term, making them larger
1061 for $f \sim 0.5$.

6. Mutation tables

Gene	Position	AA	Start date	End date	Shih	Luksza	Koel	Tree counts
HA1	144	D	2001-06-09	2002-02-04	true	true	false	0
HA1	189	N	2003-07-29	2004-05-24	false	true	true	2
HA1	159	F	2003-08-28	2004-05-24	false	true	true	2
HA1	226	I	2003-09-27	2004-09-21	true	true	false	3
HA1	145	N	2003-12-26	2004-11-20	false	true	true	2
HA1	227	P	2003-05-30	2005-04-19	false	true	false	2
HA2	32	I	2004-06-23	2005-07-18	false	false	false	1
HA1	193	F	2004-12-20	2006-03-15	false	true	true	1
HA2	46	D	2006-06-13	2007-05-09	false	false	false	2
HA2	121	K	2006-06-13	2007-06-08	false	false	false	1
HA1	50	E	2006-09-11	2007-06-08	false	true	false	2
HA1	140	I	2006-11-10	2007-11-05	true	false	false	1
HA1	173	Q	2007-07-08	2009-01-28	true	true	false	2
HA2	32	R	2007-07-08	2009-01-28	false	false	false	1
HA1	158	N	2009-01-28	2009-07-27	true	true	true	2
HA1	189	K	2009-01-28	2009-07-27	false	true	true	2
HA1	212	A	2009-03-29	2011-01-18	false	false	false	2
HA1	45	N	2010-03-24	2013-02-06	false	false	false	3
HA1	223	I	2010-12-19	2013-02-06	false	false	false	2
HA1	48	I	2011-03-19	2013-02-06	false	false	false	1
HA1	198	S	2011-03-19	2013-02-06	false	false	false	1
HA1	312	S	2009-08-26	2013-03-08	false	false	false	3
HA1	278	K	2011-06-17	2013-03-08	false	true	false	1
HA1	145	S	2011-04-18	2013-04-07	false	true	true	4
HA1	33	R	2011-06-17	2013-06-06	false	false	false	2
HA2	160	N	2012-07-11	2015-09-24	false	false	false	3
HA1	225	D	2013-08-05	2015-09-24	false	false	false	3
HA1	3	I	2013-08-05	2016-11-17	false	false	false	2
HA1	159	Y	2014-02-01	2016-11-17	false	true	true	2
HA1	160	T	2014-01-02	2017-07-15	false	true	false	2

Table S I. The 30 trajectories that took place between year 2000 and year 2018 and resulted in fixation. Columns Shih, Luksza and Koel respectively indicate whether the position is found in the epitopes lists in (respectively) [3], [11] and [5]. The Tree counts column indicates the number of times the mutation corresponding to the trajectory can be found in the phylogenetic tree. Note that a trajectory is only shown in the table if the sequenced population counts more than 10 strains at its time of fixation. This explains that only 30 trajectories are displayed, whereas more mutations did fix in this period of time.

Gene	Position	AA	Start date	End date	Fixation	Max. freq.
HA1	106	A	2001-02-09	2002-02-04	lost	1.0
HA1	144	D	2001-06-09	2002-02-04	fixed	1.0
HA1	105	H	2003-04-30	2003-10-27	lost	1.0
HA1	126	D	2003-04-30	2004-05-24	lost	1.0
HA1	140	Q	2004-01-25	2004-06-23	lost	0.31
HA1	226	I	2003-09-27	2004-09-21	fixed	1.0
HA1	173	E	2004-12-20	2006-03-15	lost	0.63
HA1	142	G	2006-06-13	2007-05-09	lost	0.71
HA1	144	D	2006-07-13	2007-05-09	lost	0.67
HA1	128	A	2006-09-11	2007-05-09	lost	0.25
HA1	157	S	2006-09-11	2007-05-09	lost	0.59
HA1	140	I	2006-11-10	2007-11-05	fixed	1.0
HA1	173	N	2007-12-05	2008-07-02	lost	0.3
HA1	157	S	2007-12-05	2008-09-30	lost	0.31
HA1	173	E	2006-06-13	2008-12-29	lost	0.67
HA1	173	Q	2007-07-08	2009-01-28	fixed	0.96
HA1	158	N	2009-01-28	2009-07-27	fixed	0.96
HA1	62	K	2009-01-28	2011-05-18	lost	0.73
HA1	144	K	2009-01-28	2011-05-18	lost	0.75
HA1	62	V	2011-04-18	2011-09-15	lost	0.34
HA1	157	S	2013-05-07	2015-09-24	lost	0.35
HA1	128	A	2012-08-10	2016-11-17	lost	0.81
HA1	197	K	2015-11-23	2016-11-17	lost	0.27
HA1	142	R	2018-05-11	2018-10-08	lost	0.38
HA1	142	G	2012-03-13		poly	0.86
HA1	144	S	2013-12-03		poly	0.96
HA1	121	K	2015-12-23		poly	0.82
HA1	142	K	2016-05-21		poly	0.77
HA1	62	G	2017-03-17		poly	0.75
HA1	128	A	2018-01-11		poly	0.56

Table S II. Trajectories of mutations at epitope positions in [3] (*Shih et. al.*) that have been observed at least once above frequency 0.25. The **Fixation** column indicates whether the mutation has fixed, disappeared, or is still polymorphic as of October 2018. The **Max.freq.** column indicates the maximum frequency reached by the trajectory. A maximum frequency of 1 for mutations that finally disappear is explained by trajectories reaching frequency 1 for one time bin and going back to lower values for following ones (a frequency above 0.95 for two time bins in a row defines fixation).

Gene	Position	AA	Start date	End date	Fixation	Max. freq.
HA1	50	G	2001-02-09	2002-02-04	lost	1.0
HA1	144	D	2001-06-09	2002-02-04	fixed	1.0
HA1	126	D	2003-04-30	2004-05-24	lost	1.0
HA1	189	N	2003-07-29	2004-05-24	fixed	1.0
HA1	159	F	2003-08-28	2004-05-24	fixed	1.0
HA1	226	I	2003-09-27	2004-09-21	fixed	1.0
HA1	145	N	2003-12-26	2004-11-20	fixed	1.0
HA1	188	N	2004-07-23	2005-02-18	lost	0.36
HA1	227	P	2003-05-30	2005-04-19	fixed	1.0
HA1	173	E	2004-12-20	2006-03-15	lost	0.63
HA1	193	F	2004-12-20	2006-03-15	fixed	0.97
HA1	142	G	2006-06-13	2007-05-09	lost	0.71
HA1	144	D	2006-07-13	2007-05-09	lost	0.67
HA1	157	S	2006-09-11	2007-05-09	lost	0.59
HA1	50	E	2006-09-11	2007-06-08	fixed	0.95
HA1	173	N	2007-12-05	2008-07-02	lost	0.3
HA1	157	S	2007-12-05	2008-09-30	lost	0.31
HA1	173	E	2006-06-13	2008-12-29	lost	0.67
HA1	173	Q	2007-07-08	2009-01-28	fixed	0.96
HA1	158	N	2009-01-28	2009-07-27	fixed	0.96
HA1	189	K	2009-01-28	2009-07-27	fixed	0.96
HA1	213	A	2009-01-28	2010-02-22	lost	0.68
HA1	144	K	2009-01-28	2011-05-18	lost	0.75
HA1	53	N	2009-11-24	2013-02-06	lost	0.72
HA1	278	K	2011-06-17	2013-03-08	fixed	0.98
HA1	145	S	2011-04-18	2013-04-07	fixed	0.99
HA1	159	S	2013-11-03	2015-08-25	lost	0.46
HA1	157	S	2013-05-07	2015-09-24	lost	0.35
HA1	159	Y	2014-02-01	2016-11-17	fixed	0.97
HA1	159	S	2015-10-24	2016-11-17	lost	0.4
HA1	197	K	2015-11-23	2016-11-17	lost	0.27
HA1	160	T	2014-01-02	2017-07-15	fixed	0.96
HA1	142	R	2018-05-11	2018-10-08	lost	0.38
HA1	135	N	2018-06-10	2018-10-08	lost	0.38
HA1	142	G	2012-03-13		poly	0.86
HA1	144	S	2013-12-03		poly	0.96
HA1	121	K	2015-12-23		poly	0.82
HA1	142	K	2016-05-21		poly	0.77
HA1	131	K	2016-09-18		poly	0.77
HA1	135	K	2016-11-17		poly	0.47

Table S III. Same as table SII, for [11] (*Luksza et. al.*).

Gene	Position	AA	Start date	End date	Fixation	Max. freq.
HA1	189	N	2003-07-29	2004-05-24	fixed	1.0
HA1	159	F	2003-08-28	2004-05-24	fixed	1.0
HA1	145	N	2003-12-26	2004-11-20	fixed	1.0
HA1	193	F	2004-12-20	2006-03-15	fixed	0.97
HA1	158	N	2009-01-28	2009-07-27	fixed	0.96
HA1	189	K	2009-01-28	2009-07-27	fixed	0.96
HA1	145	S	2011-04-18	2013-04-07	fixed	0.99
HA1	159	S	2013-11-03	2015-08-25	lost	0.46
HA1	159	Y	2014-02-01	2016-11-17	fixed	0.97
HA1	159	S	2015-10-24	2016-11-17	lost	0.4

Table S IV. Same as table SII, for [5] (*Koel et. al.*).

1063

7. Supplementary figures

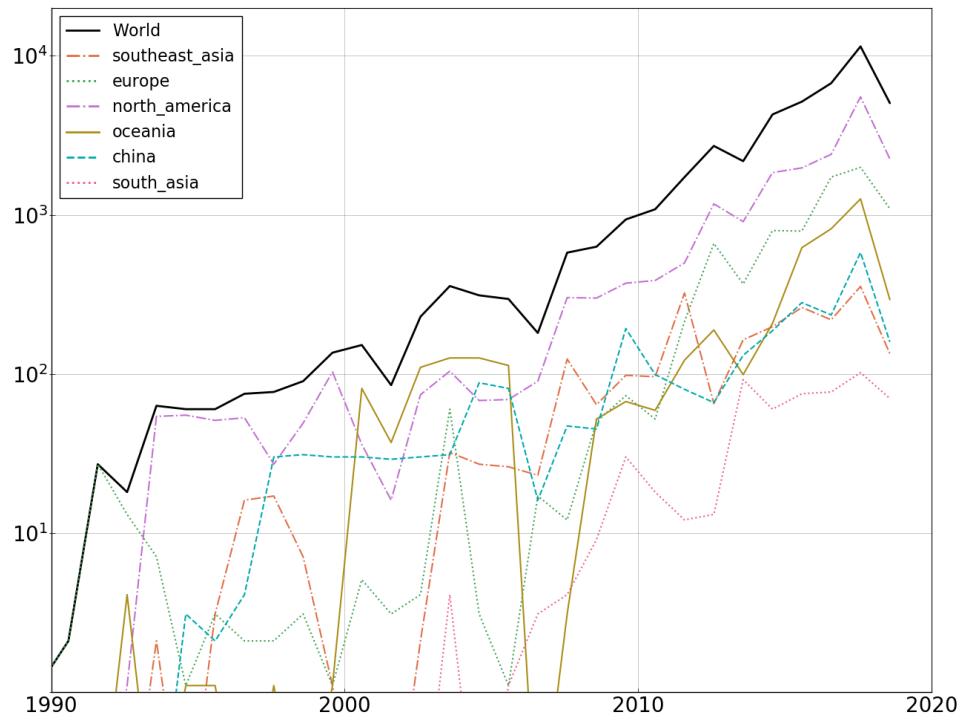


Figure S 5. Number of A/H3N2 HA sequences per year from year 1990.

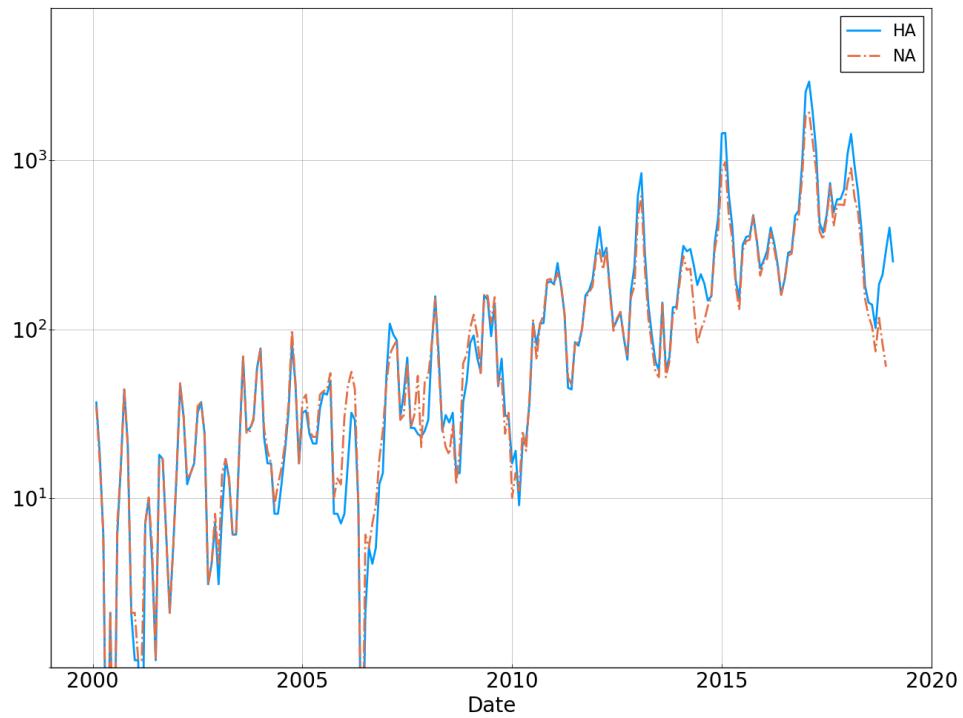


Figure S 6. Number of H3N2 HA and NA sequences per month from year 2000.

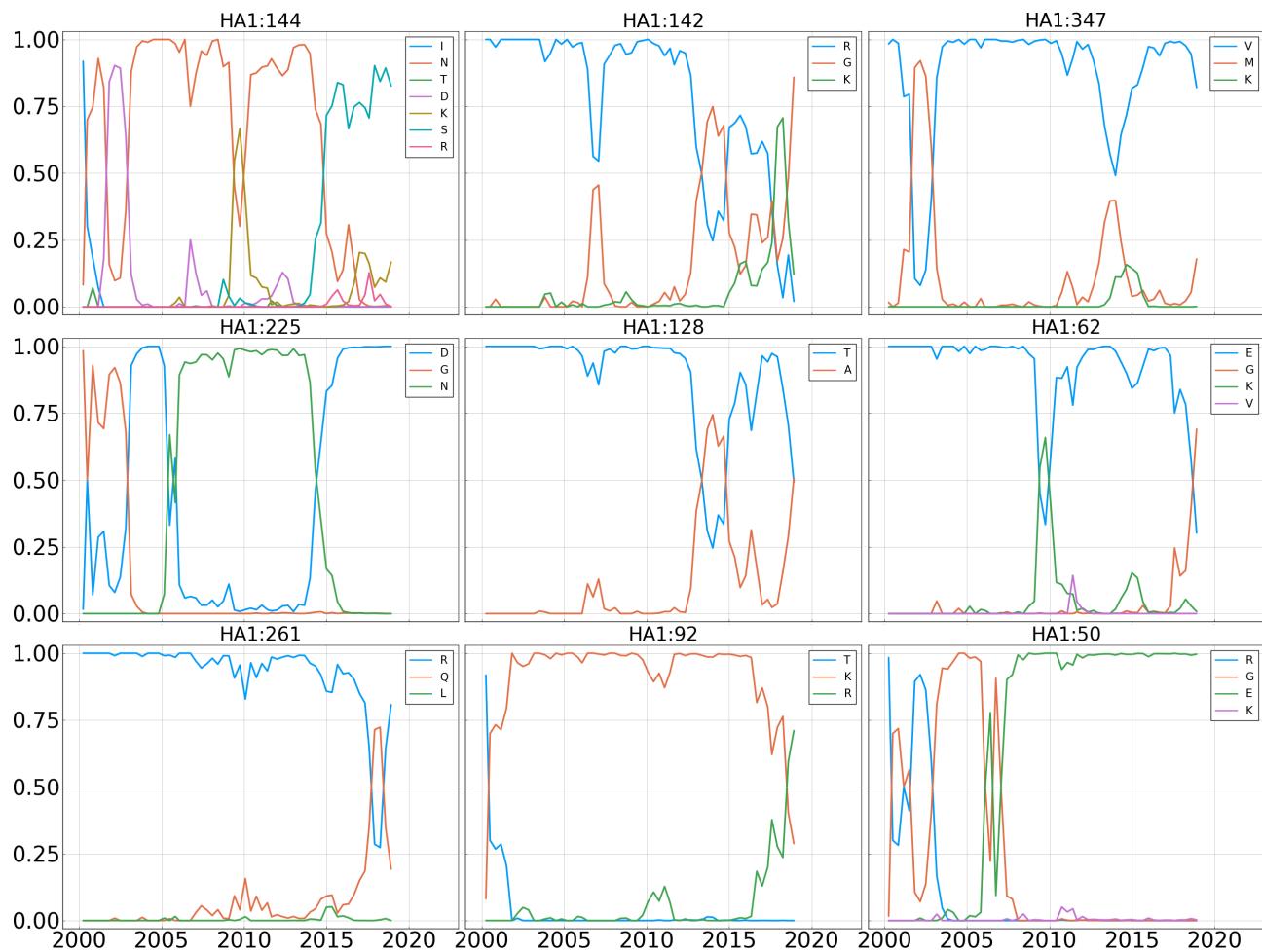


Figure S 7. Frequency trajectories for the 9 most entropic positions in the A/H3N2 HA protein.

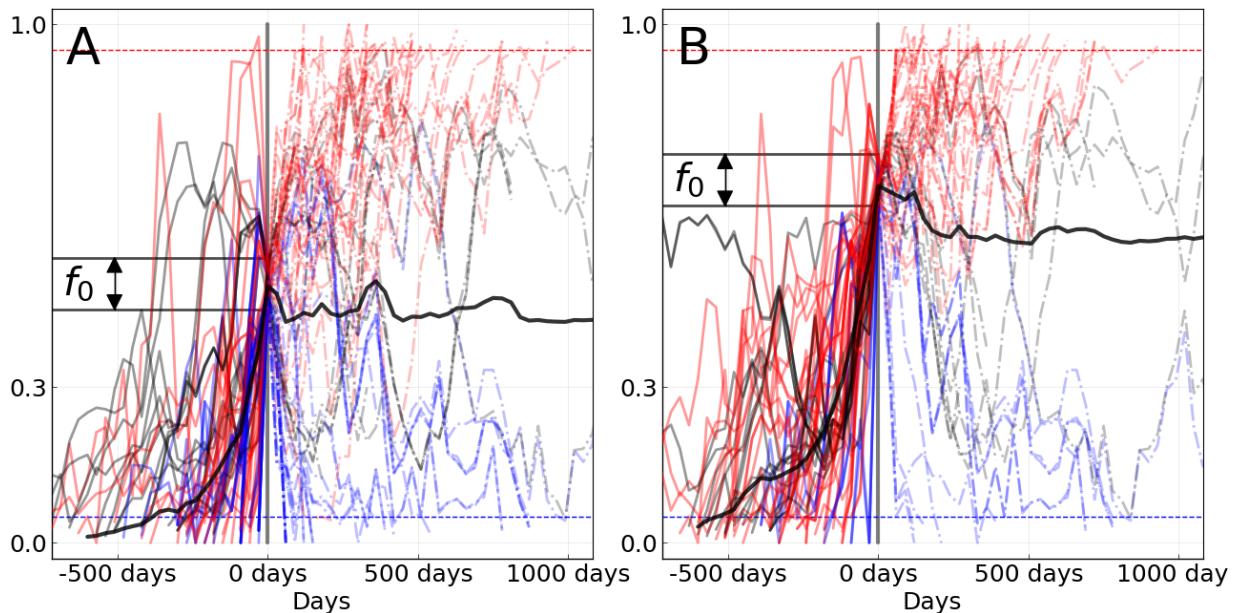


Figure S 8. Equivalent to panel **B** of figure 1 of the main text for A/H3N2, with f_0 equal to 0.5 in **A** (76 trajectories), and 0.7 in **B** (63 trajectories).

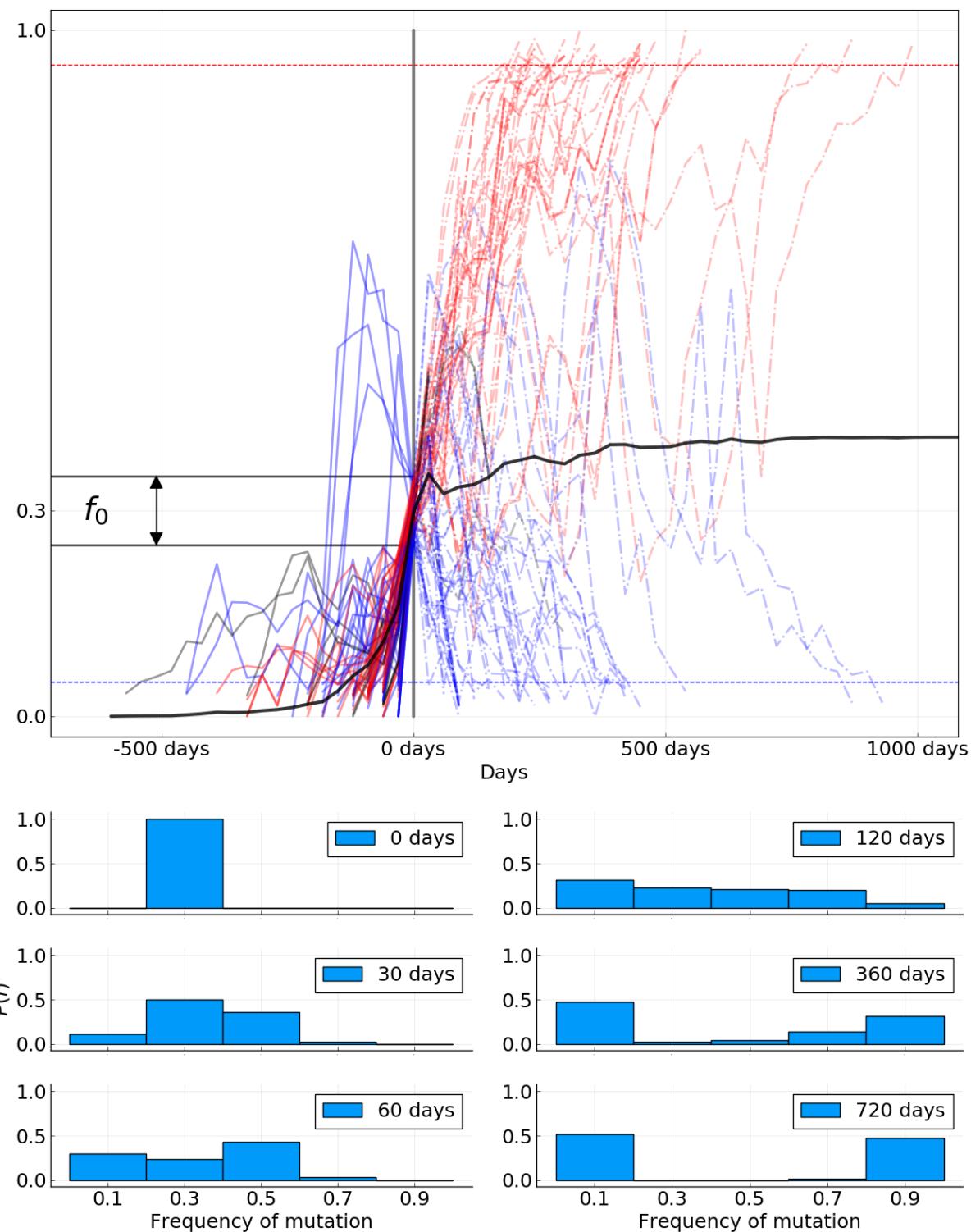


Figure S 9. Equivalent to panels **B** and **C** of figure 1 of the main text for A/H1N1pdm influenza. 89 trajectories are shown and participate to the mean (thick black line).

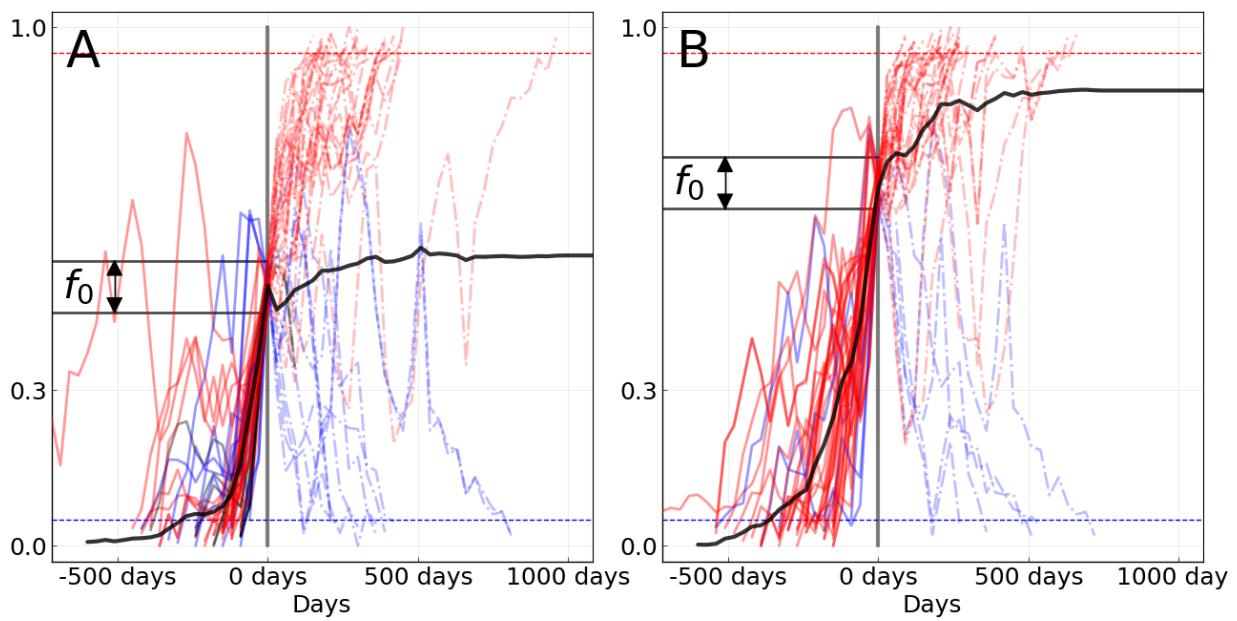


Figure S 10. Equivalent to panel **B** of figure 1 of the main text for A/H1N1pdm, with f_0 equal 0.5 in **A** (50 trajectories), and 0.7 in **B** (41 trajectories).

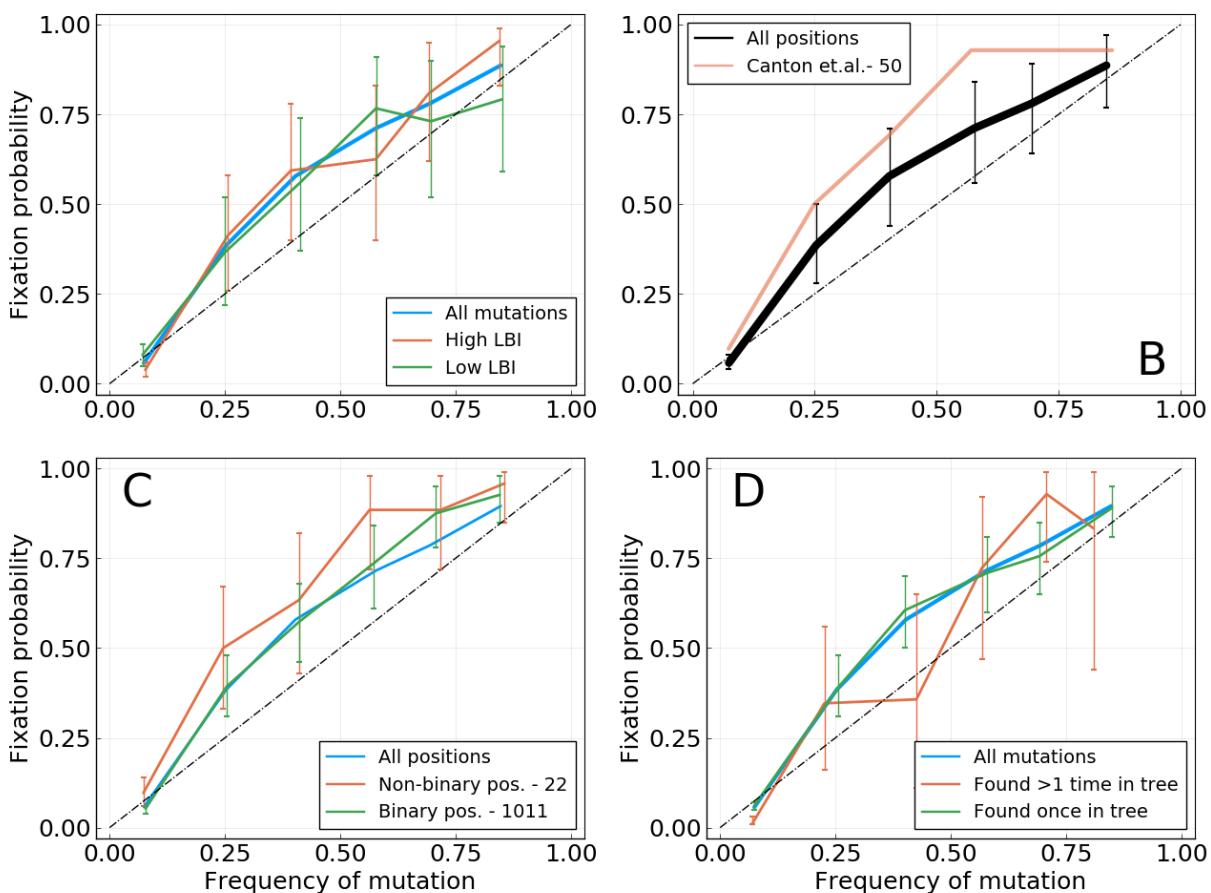


Figure S 11. Equivalent of figure 3 of the main text for the HA gene of A/H1N1pdm influenza. Fixation probability $P_{fix}(f)$ as a function of frequency. **A:** Mutation with higher or lower LBI values, based on their position with respect to the median LBI value. **B:** Different lists of epitope positions in the HA protein. The authors and the number of positions is indicated in the legend. **C:** Mutations for binary positions, *i.e.* positions for which we never see more than two amino acids in the same time bin. **D:** Mutations that appear once or more than once in the tree for a given time bin.

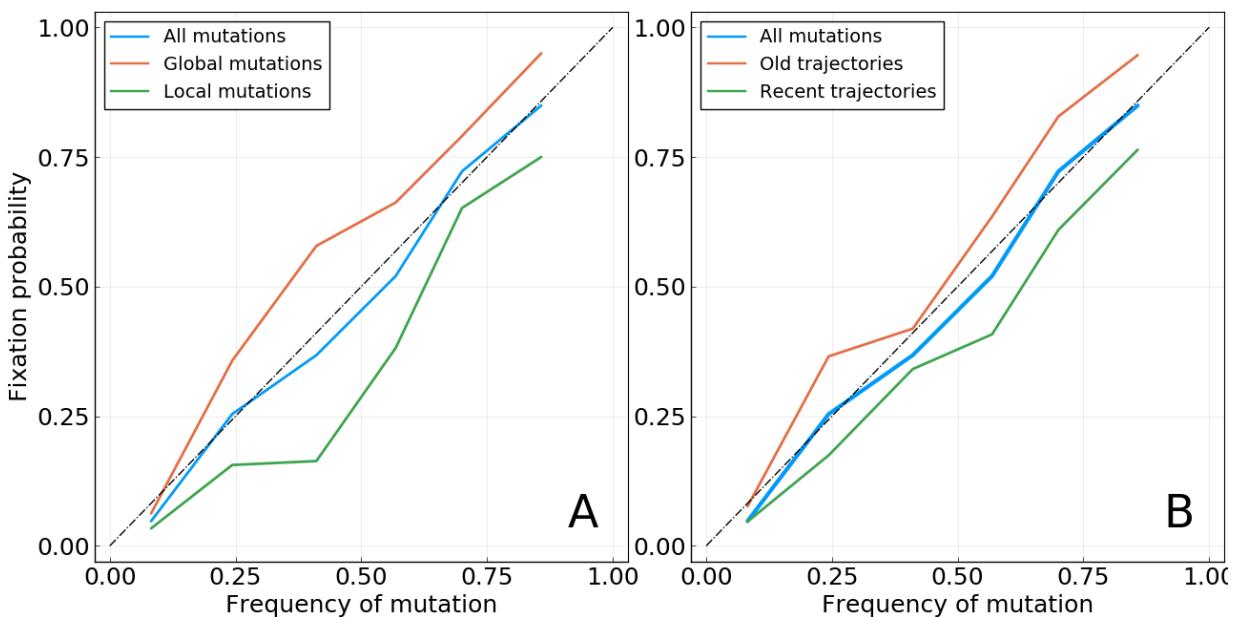


Figure S 12. Based on A/H3N2 HA and NA. **A:** Mutations with a higher or lower geographical spread, based on the median value of the score used (see Methods). *Note:* the words *local* and *global* only reflect the position of the geographic spread of the mutation relative to the median value computed for all mutations found at this frequency. As this median value may change with the considered frequency bin, so does the definition of local and global mutations. **B:** Mutations whose trajectories are older or more recent, based on the median age of trajectories when reaching the considered frequency f .

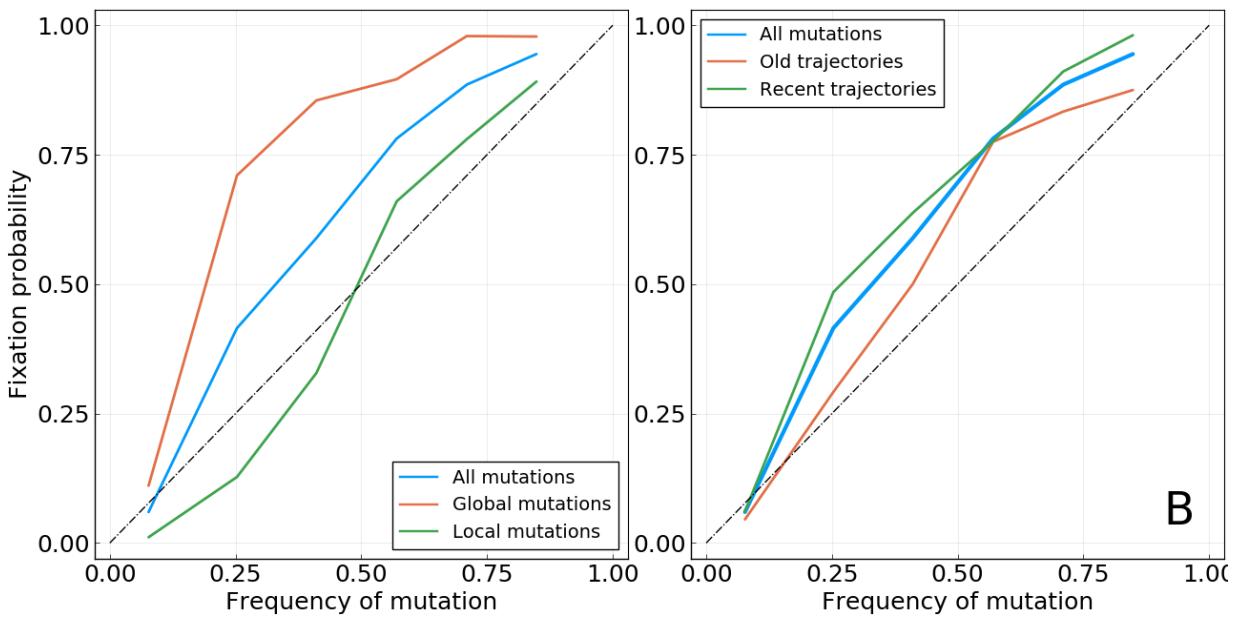


Figure S 13. Based on A/H1N1pdm HA and NA. **A:** Mutations with a higher or lower geographical spread, based on the median value of the score used (see Methods). *Note:* the words *local* and *global* only reflect the position of the geographic spread of the mutation relative to the median value computed for all mutations found at this frequency. As this median value may change with the considered frequency bin, so does the definition of local and global mutations. **B:** Mutations whose trajectories are older or more recent, based on the median age of trajectories when reaching the considered frequency f .

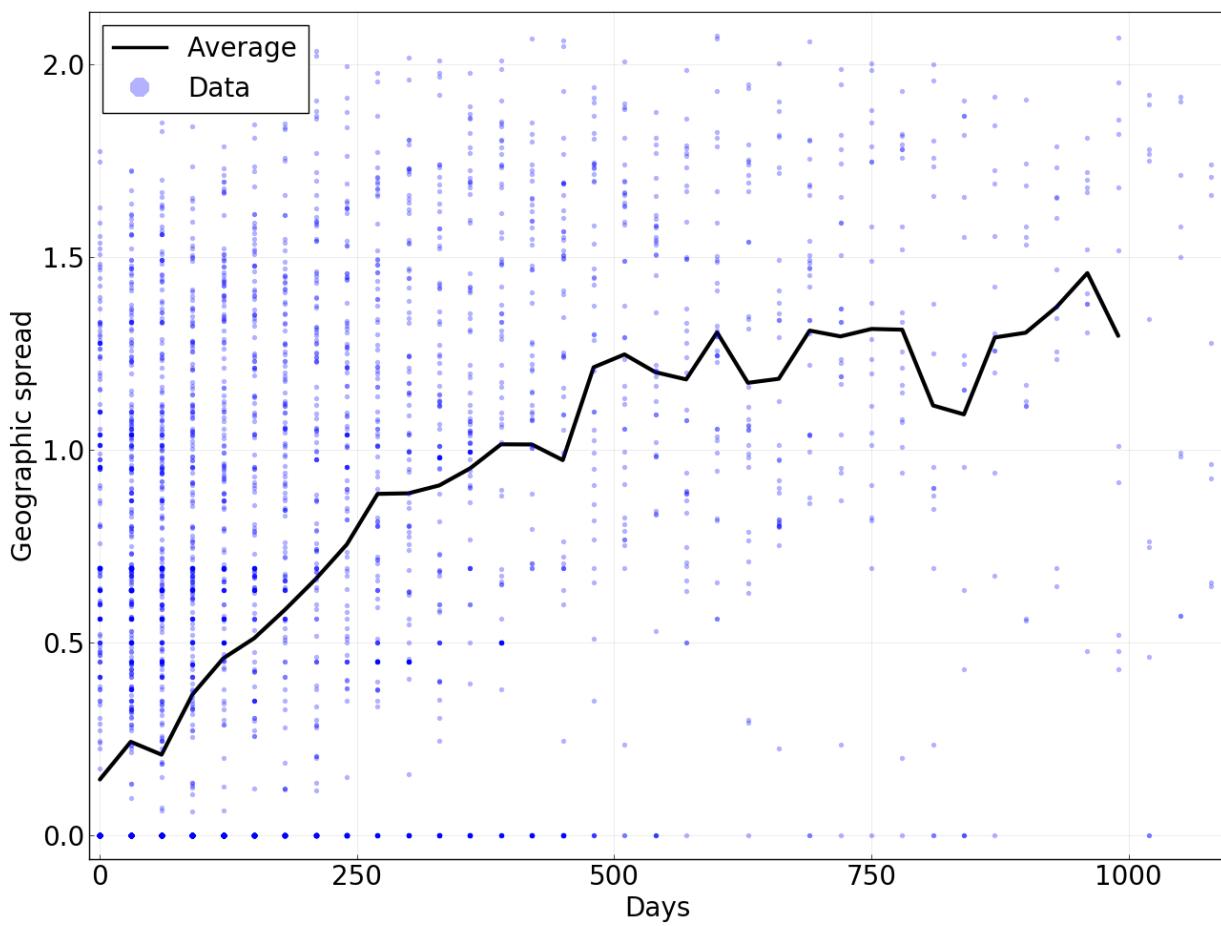


Figure S 14. Geographic spread of mutations as a function of the time for which they have been present in the population above a frequency of 5%. Points represent individual mutations and for a population in a given time bin. The line is the average of dots for a given value on the x -axis. Based on data for A/H3N2 HA.

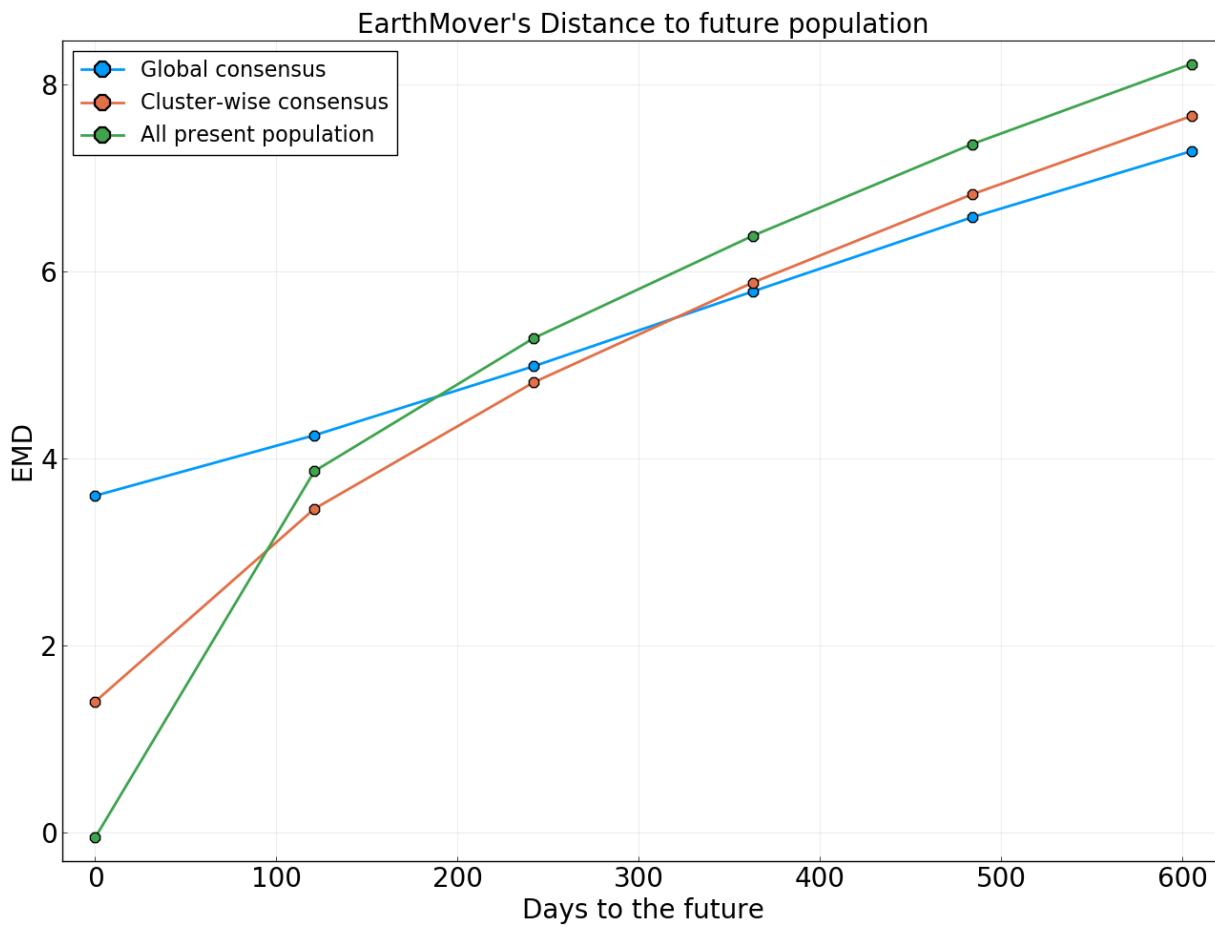


Figure S 15. Earth mover's distance to the future population for different predictors. A present population consists of all A/H3N2 HA sequences sampled in a 4 months time window. Quantities are averaged over all possible "present" populations from the year 2002. Predictors are: **Global consensus**: Consensus sequence of the present population. Best long-term predictor for a structure-less neutrally evolving population. **All present population**: All sequences in the present population. Perfect predictor if the population does not change at all through time. **Cluster-wise consensus**: Consensus sequence for each cluster in the present population. Clusters are based on local maxima of the LBI. Sequences are assigned to a given cluster based on their tree branch-length distance to the corresponding local maximum.

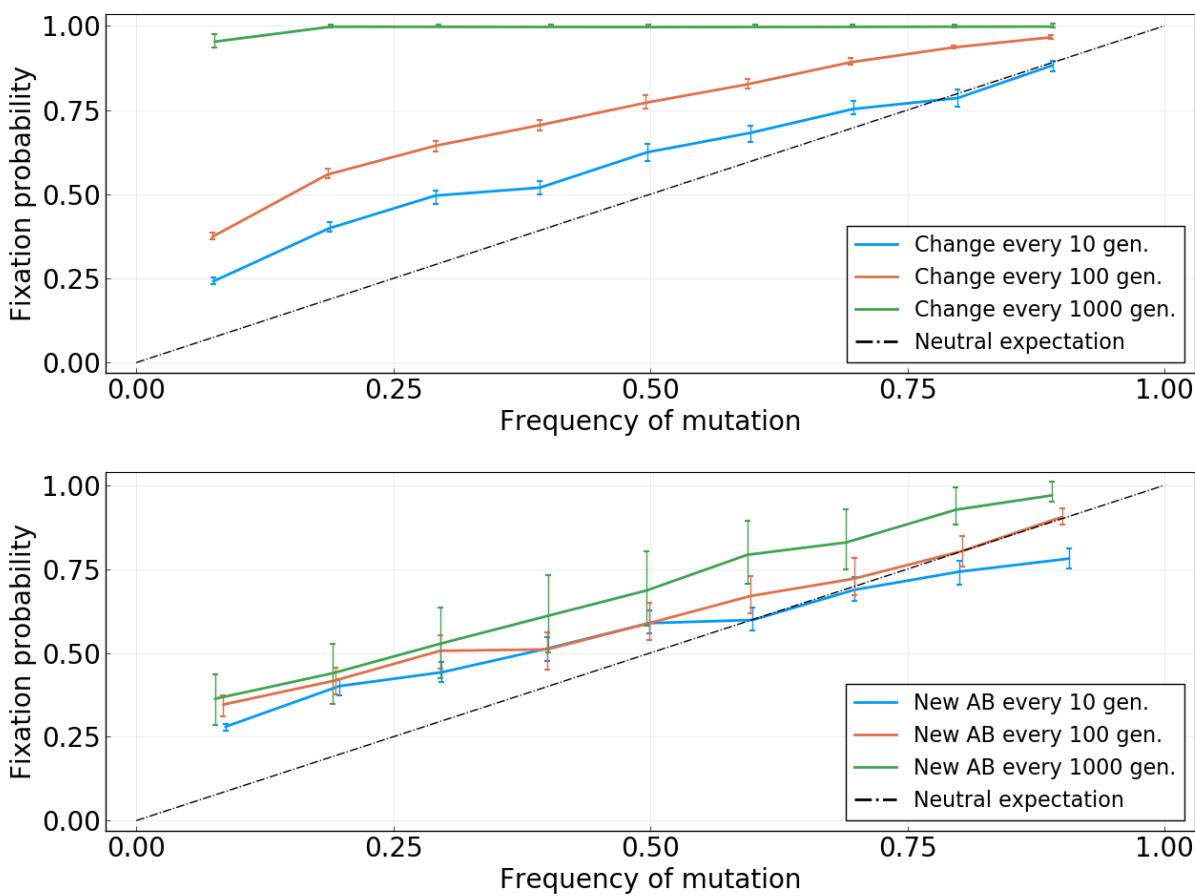


Figure S 16. Fixation probability as a function of frequency for the simulations discussed in the main text. **Top:** Simulation without antibodies. The three colored curves reflect different rate of change for the fitness landscape. Visual inspection of the frequency trajectories indicates a typical sweep time of ~ 400 generations. **Bottom:** Simulation with antibodies. The different colored curves indicate the rate at which antibodies are introduced.

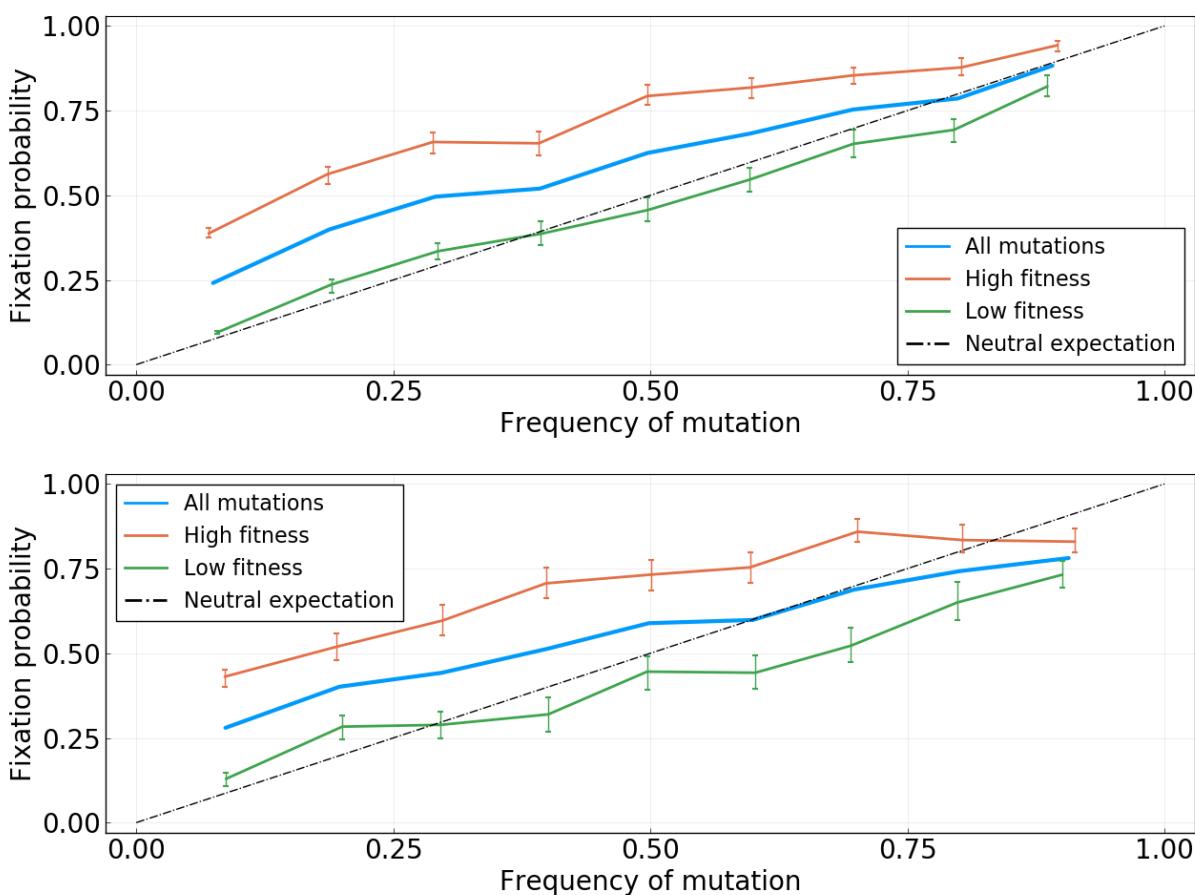


Figure S 17. Fixation probability as a function of frequency for the simulations discussed in the main text, with trajectories stratified according to real fitness values. “High” and “low” fitness classes are defined with respect to the median value. **Top:** Simulation without antibodies and with changes to the fitness landscape every $dt = 10$ generations. **Bottom:** Simulation with antibodies, with a new antibody every $dt = 10$ generations.