PLOS | PATHOGENS

RESEARCH ARTICLE

Fitness cost of reassortment in human influenza

Mara Villa, Michael Lässig*

Institute for Theoretical Physics, University of Cologne, Cologne, Germany

* lassig@thp.uni-koeln.de

OPEN ACCESS

Citation: Villa M, Lässig M (2017) Fitness cost of reassortment in human influenza. PLoS Pathog 13(11): e1006685. https://doi.org/10.1371/journal.ppat.1006685

Editor: Claus O. Wilke, University of Texas at Austin, UNITED STATES

Received: February 7, 2017

Accepted: October 9, 2017

Published: November 7, 2017

Data Availability Statement: All A/H3N2 sequence data used in this work are available from GISAID EpiFlu database (http://platform.gisaid.org/). Free registration to the platform is required in order to view the data.

Funding: This work has been supported by Deutsche Forschungsgemeinschaft (http://www.dfg.de/index.jsp), grant SFB680. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Reassortment, which is the exchange of genome sequence between viruses co-infecting a host cell, plays an important role in the evolution of segmented viruses. In the human influenza virus, reassortment happens most frequently between co-existing variants within the same lineage. This process breaks genetic linkage and fitness correlations between viral genome segments, but the resulting net effect on viral fitness has remained unclear. In this paper, we determine rate and average selective effect of reassortment processes in the human influenza lineage A/H3N2. For the surface proteins hemagglutinin and neuraminidase, reassortant variants with a mean distance of at least 3 nucleotides to their parent strains get established at a rate of about 10^{-2} in units of the neutral point mutation rate. Our inference is based on a new method to map reassortment events from joint genealogies of multiple genome segments, which is tested by extensive simulations. We show that intra-lineage reassortment processes are, on average, under substantial negative selection that increases in strength with increasing sequence distance between the parent strains. The deleterious effects of reassortment manifest themselves in two ways: there are fewer reassortment events than expected from a null model of neutral reassortment, and reassortant strains have fewer descendants than their non-reassortant counterparts. Our results suggest that influenza evolves under ubiquitous epistasis across proteins, which produces fitness barriers against reassortment even between co-circulating strains within one lineage.

Author summary

The genome of the human influenza virus consists of 8 disjoint RNA polymer segments. These segments can undergo reassortment: when two viruses co-infect a host cell, they can produce viral offspring with a new combination of segments. In this paper, we show that reassortment within a given influenza lineage induces a fitness cost that increases in strength with increasing genetic distance of the parent viruses. Our finding suggests that evolution continuously produces viral proteins whose fitness depends on each other; reassortment reduces fitness by breaking up successful combinations of proteins. Thus, selection across proteins constrains viral evolution within a given lineage, and it may be an important factor in defining a viral species.

Introduction

Influenza virus is a negative-sense single strand RNA virus. Humans can be infected by three phylogenetically and antigenically distinct influenza lineages—A, B and C—that co-circulate globally. Among these lineages, influenza A shows the fastest rate of evolution [1–4]. The genome of the virus is segmented into 8 RNA filaments that encode 11 different proteins. Within each segment, genomic evolution is a purely asexual process carried by point mutations [5], which are subject to genetic drift and natural selection. In particular, positive selective pressure by host immunity plays an important role in the evolution of the surface glycoproteins haemagglutinin (HA), which governs viral binding and entry into host cells, and neuraminidase (NA), which drives the release and escape of new virions from the cell [6, 7]. The gradual accumulation of adaptive mutations in these two proteins maintains the ability of the virus to continually evade host immunity [4, 8]; this phenotypic process has been called *antigenic drift* [9, 10].

In parallel to point mutations within single proteins, the genome of the influenza virus changes by so-called reassortment processes. If the same host cell is co-infected by two or more viruses carrying distinct genomes, mixing of genomic segments within that cell may produce a hybrid genotype carrying segments from different parental strains. The evolutionary implications of these dynamics are quite complex. On the one hand, in rare cases, reassortment can lead to *antigenic shifts*, which are new combinations of haemagglutinin and neuraminidase that strongly enhance fitness [11] by escape from host immunity [12, 13]. The acquisition of new HA and NA variants by human influenza A through reassortment with avian strains, for example, has been shown to cause global pandemics in 1957 and 1968, known as the "asian" and the "Hong Kong" flu, respectively [14, 15]. Many reassortments, however, have negligible antigenic effects but may have other fitness effects. Specifically, fitness interactions between segments across lineages are observed as biases in observed pairings [16–22]. By partly randomizing such pairings, reassortment generates a fitness cost and a resulting increase of subsequent compensatory mutations [23]. Broad negative selection has been postulated for reassortment between well distinct influenza B lineages [24], but the overall selective effects of intra-subtype ressortment have not been systematically analyzed so far.

In this paper, we infer a comprehensive map of intra-lineage reassortment between the surface proteins HA and NA of influenza A/H3N2, and we provide evidence that most of these events are under negative selection increasing with distance between parental strains. An important methodological basis for our analysis is a faithful inference of intra-subtype reassortment events from sequence data. Although these dynamics have long been recognized as a potentially important mechanism for evolution [25, 26], the detection of events within the same subtype is notoriously difficult due to their weak phylogenetic signal. There is a number of current methods to infer reassortment events from a data set of viral sequences. These methods can be roughly divided into two groups: distance-based methods [27, 28] and methods based on the phylogeny [13, 26, 29–35]. As recently pointed out [36], these approaches coherently report some fraction of the reassortment events but show a substantial degree of discrepancy between their results, which can be traced to method-specific differences in sensitivity. Distance-based methods rely on the assumption that, for reassortant strains, high similarity between the sequences of one segment goes along with large differences in the other segment. These methods do not pass through the step of inferring the phylogeny of the virus. They are fast, can be efficiently applied to large alignments, and are insensitive to errors in tree reconstruction. Without a viral phylogeny, however, it can be hard to determine if two or more inferred events are independent. Hence, distance-based methods can generate multiple representations of the same original event between unobserved ancestral strains. Resolving false

positives constitutes a major issue for this kind of algorithms. Phylogenetic methods, on the other hand, are based on the observation that reassortant strains are located in different clades of the coalescent trees built for different segments. These approaches are usually successful in detecting reassortment across different lineages of the virus, i.e., between strains with substantial genetic differences. The incompatibility between tree topologies, however, can also be a result of phylogenetic errors, so that inconsistencies in the evolutionary histories of different segments are a necessary but not sufficient condition for reassortment. Successful attempts to overcome this issue [33] have produced algorithms which are applicable only to small datasets. Since the scaling of the number of inferred events with sample sizes has not been investigated, it is not clear if the rate of reassortment is independent on the size of the trees. Hence, even if the gain in information coming from the phylogenetic trees constitutes an advantage over the distance-based algorithms, the limited resolution constrains these methods in the detection of intra-lineage events [8]. In order to fill this methodological gap, we propose a new genealogical inference method for reassortment in fast-evolving populations with segmented genomes, such as influenza virus. We analyze the mutations arising in joint genealogies built from pairs of segments and set a simple criterion to identify reassortment events. An important part of our method is the exclusion of false positive events generated by ambiguities in tree reconstruction, which can be estimated from a statistical null model of non-reassorting sequences. In order to reduce the list of putative reassortments to a minimal set of independent events, we include internal inferred nodes as possible candidates for reassortment and cluster events with similar patterns of mutations. The overall number of reassortment events reported by our method turns out to be in broad agreement with the results of previous studies [37–39]. Furthermore, we have extensively tested our method on simulated data for evolving influenza-like strain populations under mutations, genetic drift, selection, and reassortment. We find that a large fraction of the simulated reassortment events are recovered by our algorithm, and these events outweigh the rate of false positives.

In the second part of the paper, we turn to our main biological objective of mapping selection on reassortment within an influenza lineage. We apply two independent and complementary selection inference methods to a set of intra-subtype reassortments in the A/H3N2 lineage inferred by our genealogy-based algorithm. First, we compare the distributions of the RNA distances between actual reassortant strains with a suitable background distribution of co-circulating strains, which quantifies the neutral opportunities for reassortment. Second, we compare the total population sizes between the reassortant clade and the non-reassortant clade defined by individual reassortment events; these sizes are estimated from the number of strains in the sequence sample. We find a consistent signal of broad negative selection on intra-lineage reassortment by both methods. We interpret this signal in terms of ubiquitous cross-protein epistasis and discuss evolutionary consequences.

Results

Reassortment events and their detection

A reassortant strain reassembles the genome segments of two parent strains that co-infect a host cell. In this study, we focus on reassortment between the HA and NA genes of influenza, because the evolution of both proteins has been linked to immune escape and functional epistasis between them affects vaccine efficacy [40, 41]. Hence, we restrict the genomic analysis to the two segments carrying the HA and NA genes; each parent strain contributes exactly one of these segments to the reassortant strain (Fig 1). Our inference of reassortment is based on genealogical trees constructed from linked sequence of these two segments (Materials and methods). A tree representation of a joint genealogy with a reassortment process is shown in

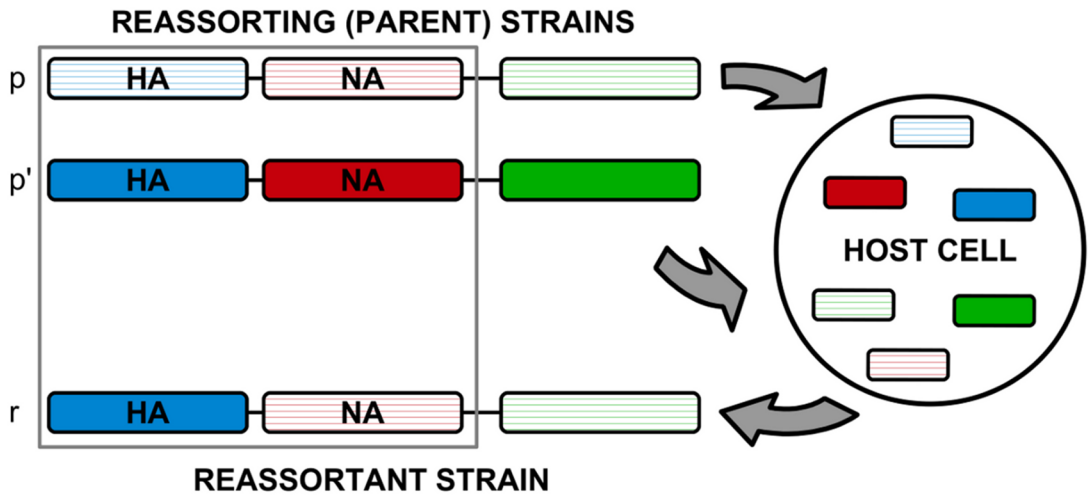


Fig 1. Schematic of a reassortment process. Two parent strains, p and p', co-infect a host cell and produce a reassortant strain r. Here we focus on reassortment of the two surface proteins HA (blue segments) and NA (red segments); the reassortant strain r inherits one of these segments from each parent.

https://doi.org/10.1371/journal.ppat.1006685.g001

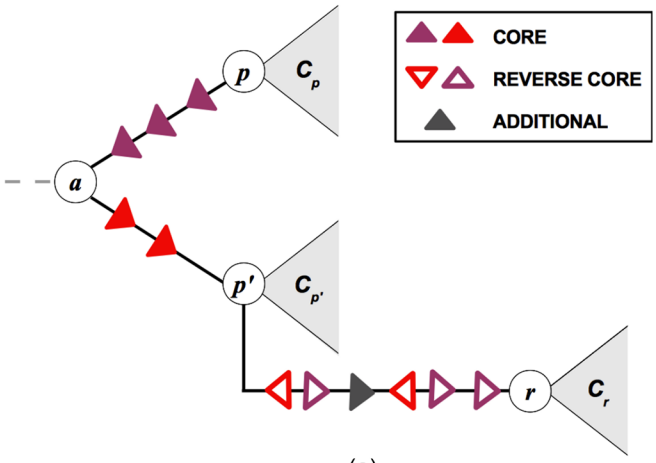
Fig 2a. The two parental strains p and p' appear in different sublineages, and the reassortant strain r is shown as a descendant of one of these parents (here p'). These strains define three distinct clades of descendant strains, C_p, $C_{p'}$, and C_r (grey areas in Fig 2a); the numbers of strains in these clades are denoted by n_p, $n_{p'}$, and n_r, respectively. We note that the "direction" of the reassortment event (here from p to p') is merely a property of the tree representation, and there is an equivalent tree with the roles of p and p' exchanged. This reassortment pattern can be readily identified in two-segment trees. Fig 2b shows an example of a HA-NA reassortment event in the genealogy of influenza A/H3N2, another example using tree data from simulated evolution of a population in a regime of clonal interference is shown in Fig 2c.

We identify candidate reassortment events (p, p', r) from their signal in a two-segment genealogical tree: a set of core mutations, $\mathcal{A}_{pp'}$, appears on the branches between the nodes p and p', and their reverse mutations appear on the branch between p' and r. These mutations are in the segment that r inherits from p (Fig 2a). The resulting list of events must undergo further statistical analysis: false positives must be excluded and candidates representing the same reassortment event must be clustered. In Materials and Methods, we detail our inference scheme and show that its fidelity strongly depends on the number of core mutations, $\delta = |\mathcal{A}_{pp'}|$. Inferred events with $\delta \geq 5$ are very likely to be true reassortments, while events with $\delta \leq 4$ are enriched with false positive counts reflecting alignment ambiguities (Fig 3).

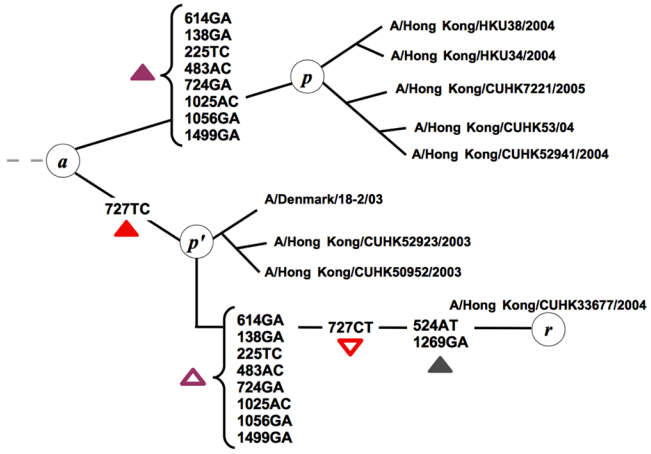
To characterize the span of a reassortment event, we use the mean genetic distance between the parent strains p and p' in both segments, $d = \frac{1}{2}(d_{\mathrm{HA}} + d_{\mathrm{NA}})$ (we evaluate these distances for nucleotides and for amino acids). The quantity d is also the mean genetic distance of the reassortant strain r from its parents; below, we report evidence for negative epistasis between these mutations.

Rate and genealogy of reassortment for influenza A/H3N2

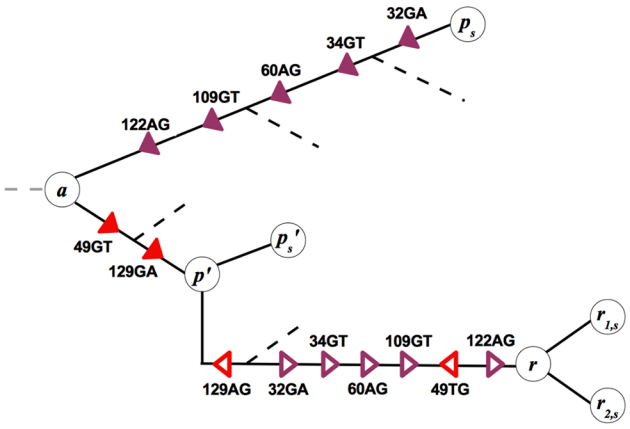
We apply our reassortment inference method to a sequence dataset of human influenza A/H3N2 collected from 1968 to 2015. In each two-segment tree, we map HA-NA reassortments as detailed in Methods: we identify candidate events (p, p', r) by the criterion (Eq 5), we keep only events with core distance $\delta \geq 5$, we prune events with strongly overlapping core sets



(a)



(b)



(c)

Fig 2. Tree representation of reassortment. (a) Representation of reassortment in a two-segment genealogical tree. The parent strains p and p' are in different sublineages of the tree; the reassortant strain r appears as a descendant of one of these parents (here p'; there is an equivalent tree in which r appears as a descendant of p). The strains p, p', and r are the focal nodes of the clades C_p, $C_{p'}$, and C_r, respectively (grey areas). We identify the reassortment event by its set of core mutations, $\mathcal{A}_{pp'}$, which appear on the segment that r inherits from p and generate the genetic distance between the parent strains in that segment. The core

mutations appear on the branches between the nodes p and p' (filled red triangles: mutations between p and the last common ancestor a, filled purple triangles: mutations between a and p'). Their reverse mutations appear on the branch between p' and r (empty red and purple triangles), which can also contain additional mutations (grey triangles). (b) A true event (nr 1 in S2 Table) detected by our algorithm on the joint HA-NA tree. Each mutation on HA segment is labeled with a number between 1 and 1701 that indicates the site. The pattern of repeated and reversed mutations (filled and empty triangles) follows the scheme in Fig. 2a: the reassortant strain A/Hong Kong/CUHK33677/2004 is generated by an event with $\delta = 9$ between p and p' clades. (c) The result of a simulated reassortment event on the reconstructed genealogical tree, correctly detected by the algorithm. The internal node r is inferred as the reassortant ancestor of $r_{1/2,s}$, i.e. the strains evolved from the sequence that was actually generated by reassortment between p_s and p'_s.

https://doi.org/10.1371/journal.ppat.1006685.g002

$\mathcal{A}_{pp'}$, and we eliminate double-counting of events. At the same time, the total number of detected events does not depend on the number of inferred trees, indicating that our mapping exhausts the events occurring in the original dataset. Furthermore, we have verified that passaging mutations do not confound our inference (Methods).

This procedure produces a list of 103 bona fide reliable and independent HA-NA reassortments in our data set. These events have a mean genetic distance $d \geq 3$ between reassortant

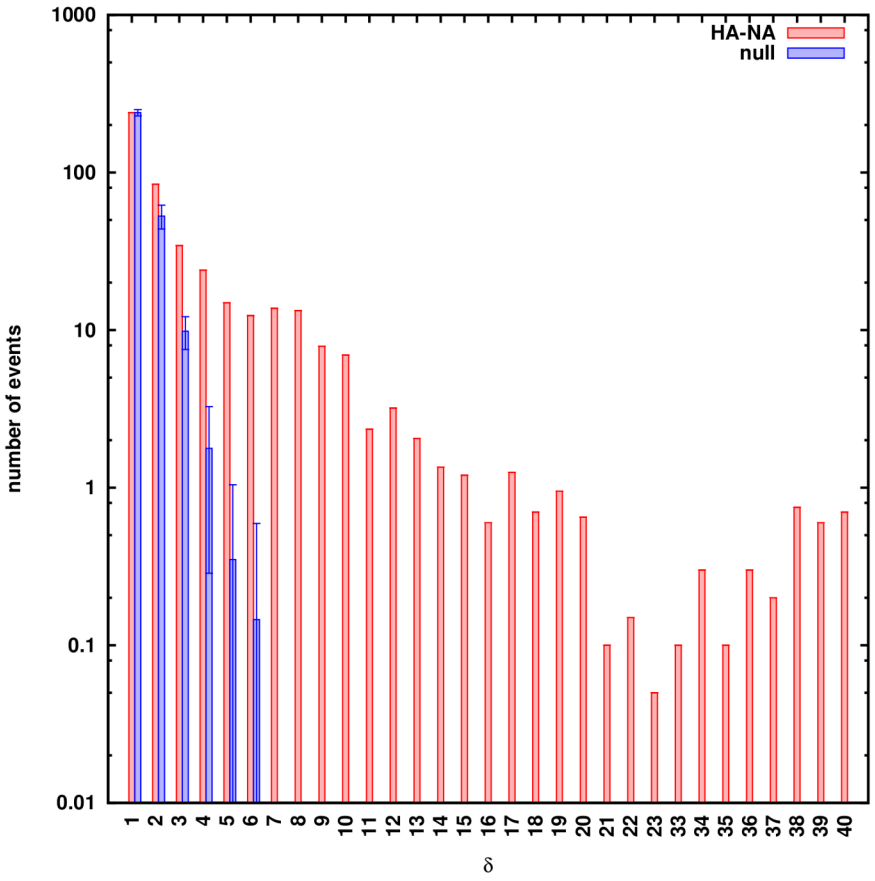


Fig 3. Fidelity of reassortment inference. Histograms of reported HA-NA reassortment events for different core distances δ (red bars) are compared to expected number of false positives due to ambiguities in tree reconstruction, $n_0(\delta)$, from a null model of non-reassorting sequences (blue bars; error bars reflect the statistics over different realizations of the null model). The function $n_0(\delta)$ decreases exponentially with increasing δ (cf. Eq 6 in Materials and methods); the overall amplitude is set by the conservative assumption that all counts at $\delta = 1$ are false positives. The resulting total number of false positives with $\delta \geq 5$ is below 1.

https://doi.org/10.1371/journal.ppat.1006685.g003

and parent strains in both segments and an average $d_{ave} = 10$, which sets them clearly apart from individual point mutations and provides the genetic basis for potentially strong and epistatic selection (see below). In S2 Table, we report the genetic distance d, as well as representative strains from the reassortant clade C_r and the parent clades C_p, $C_{p'}$. From the year 2000 on, we find an average of 6 unique reassortment events per year, which is in overall agreement with other studies [37–39]. Furthermore, events detected by the majority of these studies are well represented in our list (starred events in S2 Table); these include reassortments between New York strains isolated between 2000 and 2005 [13, 28, 32, 34] (see S1 Text for more details). The clean statistical test used here, however, addresses the over-counting of events in a more objective way.

Fig 4 shows the inferred reassortments since 2000 mapped on a joint HA-NA tree. These events cover the entire time interval of the tree with a slight increase in frequency in recent years, which is likely due to increased depth of the tree. In all cases, the parent strains were collected at close times, which is consistent with the fast evolutionary speed and the resulting short sojourn periods of specific genotypes in the population of circulating strains. By comparing the number of reassortment events with the number of synonymous nucleotide changes on the same tree, we estimate that reassortant variants get established at a rate of order 10^{-2} in units of the neutral point mutation rate. This establishment rate refers to observed variants in a strain sample, which clearly depends on the sampling depth (our data set has a detection threshold frequency of order 10^{-3}). Our finding of broad negative selection on reassortment, which is reported below, suggests that the reassortment rate of individual virions is higher, but many reassortant variants are rapidly lost in the population of circulating strains.

Reassortment is under broad negative selection

As shown in Fig 4, the majority of observed reassortment events are on peripheral positions of the joint HA-NA tree. This observation is broadly consistent with a neutral or, on average, deleterious process. We now turn to measuring selection on reassortment in a more quantitative way.

First, we compare the distribution of distances in detected reassortment events, $P(d)$, with the corresponding background distribution sequence distances between all pairs of strains circulating in a given influenza season, $P_0(d)$ (both distributions are defined in the regime $d \geq 3$). The latter distribution represents the background pool of a priori equiprobable opportunities for reassortment. In the absence of selection, reassortment should occur with equal probability between these pairs, regardless of their genetic distance, and the distribution $P(d)$ should be similar to the background distribution $P_0(d)$. However, Fig 5a shows significant differences between these distributions: there are far fewer actual events with larger values of d than in the background distribution. We measure the statistical significance of these differences by the Kullback-Leibler (KL) divergence $D_{KL} = \Sigma_{d\geq 3} P_0(d) \log(P_0(d)/P(d))$ and by the Kolmogorov-Smirnov statistics, $D_{KS} = \max |F(d) - F_0(d)|$, where $F(d)$ and $F_0(d)$ are the corresponding cumulative distributions. We find the suppression of high-d reassortment events to be significant by both tests, with $D_{KL} = 0.56$ (compared to a 5% error threshold at $D_{KL} = 0.1$) and $D_{KS} = 0.34$ (giving a probability $p < 10^{-23}$ to find a larger distance by chance). As shown by Fig 5b, the ratio of reassortment to background counts with $d \geq d_{min}$ decreases as a function of the lower cutoff d_{min}. We attribute this effect to distance dependent deviations from neutrality: negative selection on reassortment increases in strength with distance d. The same analysis based on amino acid distances, which provide a more coarse-grained measure of genetic differences, also shows a significant suppression of large-d reassortment events (S3 Fig).

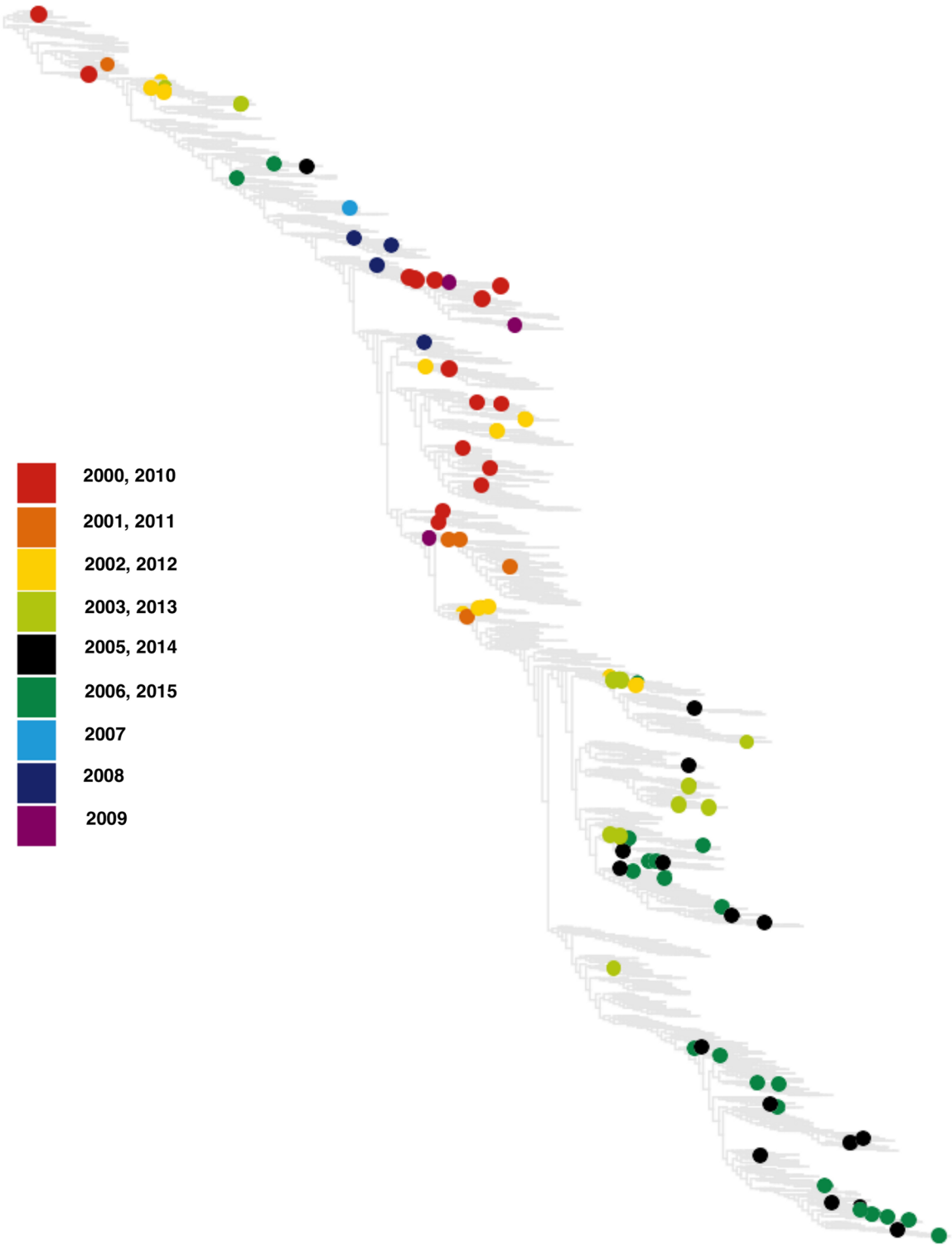


Fig 4. Reassortment of HA and NA in human influenza A/H3N2 from 2000 to 2015. The 95 inferred events are mapped on a joint HA-NA tree. The reassortant strain *r* of each event is represented by a filled circle (color-coded by year of occurrence). The events are homogeneously distributed over the tree and the reassortant clades are predominantly at peripheral positions of the tree.

https://doi.org/10.1371/journal.ppat.1006685.g004

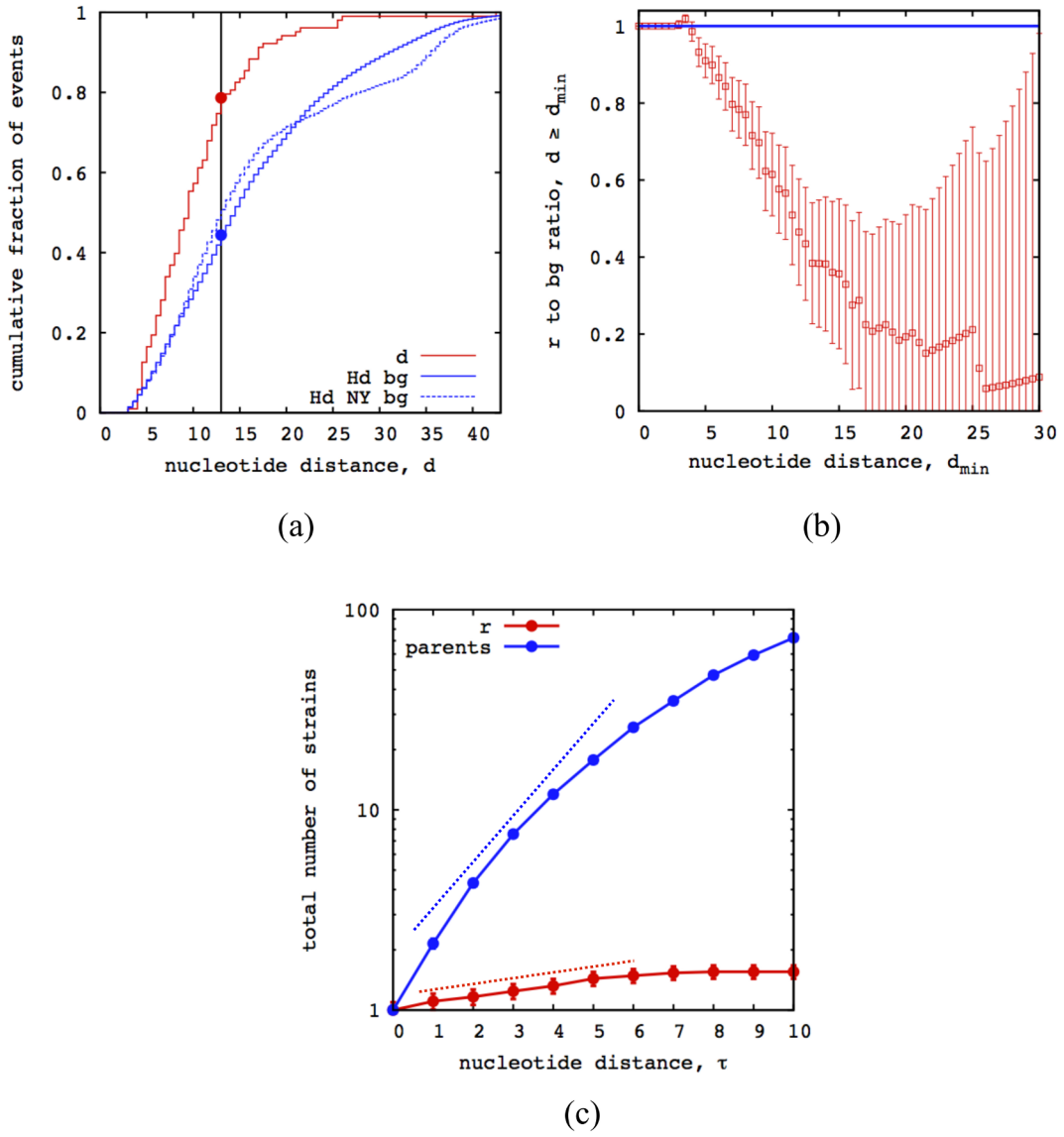


Fig 5. Negative selection on reassortment. (a) The cumulative distribution of mean nucleotide distances d between reassortant and parent strains for the HA-NA reassortments in influenza A/H3N2 (red line) is compared to the corresponding distribution of distances for co-circulating strains in the same influenza season (solid blue line) and from the New York area only [28] (dashed blue line). (b) The ratio of reassortment counts to background counts in the interval $d \geq d_{\min}$ (red circles) decreases with increasing lower threshold $d_{\min}$ and drops significantly below 1 (blue line). The suppression of reassortment at larger values of d signals distance-dependent negative selection. Bars show statistical errors due to the finite number of inferred reassortments. (c) The average number of strains in the reassortant clades with nucleotide distance $\leq \tau$ from the focal node, $\langle N_r \rangle(\tau)$ (red line) is compared to the corresponding average number of strains in the parent clades, $\langle N_0 \rangle(\tau)$. For $\tau \lesssim 6$, both functions increase with τ in an approximately exponential way; we estimate growth rates $f_r \approx 0.07$ and $f_0 \approx 0.5$, respectively (dashed lines; cf. Eq 1). The growth rate difference $\bar{s} \equiv f_0 - f_r \approx 0.4$ measures the average fitness cost of reassortment. Bars represent statistical errors due to the finite number of counts (not shown when these errors are smaller than the dot size). See S3 Fig for an analogous inference based on amino acid distances.

https://doi.org/10.1371/journal.ppat.1006685.g005

A potential confounding factor for this analysis is the spatio-temporal population dynamics of the human influenza virus [12, 42]. Although influenza strains are known to travel rapidly, the local background distribution P_0 of genetic distances, which matters for reassortment, can in principle differ from its global counterpart used in our significance analysis. In order to

estimate this effect, we restrict the background distribution to strains that circulate in the same region; specifically, we calculate an alternative distribution P_0 using isolates from New York State only, which are available from a previous study focusing on that region [28] (as before, Hamming distances are computed only between strains reported in the same influenza season). We still find that the actual reassortment events differ significantly from the local distribution P_0 (Fig 5a and S3 Fig), while the global and the local background distributions are statistically indistinguishable. We conclude that the suppression of large-d reassortment is not a spurious demographic effect, but is indicative of selection.

To gain some insight on how reassortment constraint is distributed on the amino acid distances in individual segments, d_{HA} and d_{NA}, we evaluate the joint background distribution $P_0(d_{\mathrm{HA}}, d_{\mathrm{NA}})$ and compare it with the amino acid distance pairs $(d_{\mathrm{HA}}, d_{\mathrm{NA}})$ of the inferred reassortment events (S4 Fig). The joint statistics of $(d_{\mathrm{HA}}, d_{\mathrm{NA}})$ differs in the coordinates $d_{\mathrm{HA}} + d_{\mathrm{NA}}$ and $d_{\mathrm{HA}} - d_{\mathrm{NA}}$, indicating that selection is not a function of d only. In particular, the conditional distributions $P(d_{\mathrm{HA}} - d_{\mathrm{NA}}|d)$ differ between data and background (S4 Fig), which is consistent with the expectation that reassortants similar in one protein are less selected against, even if the distance in the other protein is larger.

To quantify negative selection on reassortment directly within the set of observed events, we compare the evolution of population sizes of reassortant clades and of parent clades as a function of their age τ. We evaluate, for each reassortment event, the number $N_r(\tau)$ of strains in the reassortant clade C_r with nucleotide distance $\leq \tau$ from the focal node r, together with the mean of the corresponding numbers of strains in the parent clades, $N_0(\tau) = (N_p(\tau) + N_{p'}(\tau))/2$. We obtain these functions in the joint HA-NA tree of the full data set, counting strains with the same sequence only once. Averaging over the set of reassortment events, we can measure the expected growth of reassortant and parent clades,

$$\langle N_r \rangle(\tau) \sim \exp(f_r \tau), \qquad \langle N_0 \rangle(\tau) \sim \exp(f_0 \tau); \tag{1}$$

similar inference methods for clade growth are discussed in refs. [43, 44]. The functions $\langle N_r \rangle(\tau)$ and $\langle N_0 \rangle(\tau)$ for influenza A/H3N2 indeed show approximately exponential growth in the regime $\tau \lesssim 6$, which corresponds to time intervals of order one year (Fig 5c). The fitted growth rate difference estimates the average fitness cost of reassortment in our set of events,

$$\bar{s} \equiv f_0 - f_r \approx 0.4 \tag{2}$$

in units of the total point mutation rate in both segments, which equals approximately 5×10^{-3} per day. The same analysis performed with aminoacid distances is reported in S3 Fig.

Epistasis across proteins

To further interpret the observed fitness cost of reassortment, we consider the simplest epistatic fitness model for combined (HA, NA) genotypes, $F_{\alpha\beta} \equiv f(\mathrm{HA}_\alpha, \mathrm{NA}_\beta)$, where $\alpha, \beta = +1$ denote the alleles of the parent strain p and $\alpha, \beta = -1$ the alleles of parent strain p'. The model takes the form

$$F_{\alpha\beta} = f^{\mathrm{HA}}_{\alpha} + f^{\mathrm{NA}}_{\beta} + \frac{\omega}{2}\alpha\beta, \tag{3}$$

where f^{HA} and f^{NA} denote single-protein fitness values and ω is the strength of cross-protein epistasis. In terms of this model, the mean fitness cost of a reassortant strain compared to its parent strains is

$$s = \frac{1}{2}(f_p + f_{p'}) - f_r = -\frac{1}{2}(f^{\mathrm{HA}}_{+} - f^{\mathrm{HA}}_{-}) - \frac{1}{2}(f^{\mathrm{NA}}_{-} - f^{\mathrm{NA}}_{+}) + \omega, \tag{4}$$

where we assume, without loss of generality, that the reassortant strain r inherits HA from parent p and NA from parent p'. If co-infection randomly mixes co-circulating strains, the single-protein fitness value of a reassortant strain is, on average, equal to the mean fitness of its parents. For strains observed in a sequence sample, this value can only be biased towards larger reassortant fitness; i.e., $\frac{1}{2}\langle f_+^{\mathrm{HA}} - f_-^{\mathrm{HA}}\rangle \geq 0$ and $\frac{1}{2}\langle f_-^{\mathrm{NA}} - f_+^{\mathrm{NA}}\rangle \geq 0$. Hence, the observed fitness cost (2) implies an average epistatic cost of reassortment, $\langle\omega\rangle > \bar{s} > 0$. Cross-protein epistasis in the observed reassortment events is of moderate strength but broadly distributed: reassortant variants are fit enough to reach population frequencies detectable in our sample, but they are, on average, less fit than their non-reassortant counterparts.

Discussion

We have developed a new method to map reassortment of genomic segments in an evolving viral population. We detect reassortment events based on their trace in the genealogy of the population. On a two-segment genealogical tree, a set of *core mutations* in one of the reassorted proteins appears twice: on the branches linking the reassorting (parent) strains and, in reverse direction, on the branch to the reassortant clade (Fig 2a). We are interested predominantly in reassortment events above a certain minimum genetic distance d from their parent strains (here at least 3 mutations), which are clearly set apart from the dynamics of point mutations. This is also the regime in which our method allows a reliable identification of events, which is not confounded by ambiguities in tree reconstruction (Fig 3).

The main biological result of this paper is that reassortment within human influenza A/H3N2 is under broad, distance-dependent negative selection. Specifically, there are fewer large-d reassortments in our sample than expected from the distribution of co-circulating strains, and reassortant strains have fewer descendants than their parent strains (Fig 5). These observations probe negative selection on reassortant genotypes at different scales of frequency and sojourn time in the population. The suppression of large-d reassortment signals purifying selection that prevents some reassortant variants from reaching sufficient frequencies to appear in our strain sample; the growth rate difference between reassortant and parent clades indicates moderate negative selection on the variants that do appear in the sample. Reassortment between very close sequences may well be approximately neutral, but sequence-based inference methods cannot distinguish such events reliably from point mutations. The inferred selective effects characterize the continuous evolution of a seasonal influenza lineage; they do not exclude rare large-effect reassortment events causing antigenic shifts and seeding new lineages. We stress that our results are based on statistical methods evaluating ensembles of inferred reassortments and background distributions. Any such method is subject to possible confounding factors and biases; for example, reassortment can be expected to occur preferentially in high-infection settings at the peak of seasonal epidemics. It is the consistent outcome of two distinct inference procedures that gives a credible signal of selection acting on reassortment.

Reassortment can be seen as a natural "experiment" that continuously produces new combinations of viral proteins and probes their fitness in a fast-evolving population. Hence, the statistics of reassortment is informative of key selective forces governing the evolutionary dynamics. Specifically, our result of negative selection on reassortment signals ubiquitous fitness interactions (epistasis) between viral proteins; that is, the fitness of alleles of one protein depends on the genetic background of the other proteins in the same virion. The mutation-selection dynamics in non-reassortant sublineages produces favorable combinations of protein alleles, and reassortment introduces a fitness cost by randomizing these combinations. Importantly, this cost arises between genetic variants from co-circulating strains in a given viral

lineage; these variants are individually viable, differ by just a few mutations (of order 20 nucleotide changes in both reassorted segments together), and have a recent common ancestor (typically dating just one or two years back). This implies that new favorable protein combinations are continuously produced and selected for, while many random combinations incur a fitness cost. In other words, cross-protein epistasis constrains the adaptive evolutionary path of a given influenza lineage. This result could be tested, for example, by combining reassortment experiments [16–22] with in-vitro competitive fitness assays. Reassortant strains with substantial distances d should frequently be outcompeted by any of their parent strains.

The resulting negative selection on reassortment increases in strength with the genetic distance between reassortant and parent strains (Fig 5b). The mixing of genetic material by reassortment is somewhat similar to recombination in diploid populations or transformation in bacteria. However, recombination and transformation require physical splicing of genome segments; the rates of these enzymatic processes get strongly suppressed with increasing genetic distance of the parent sequences [45–47]. There are no corresponding physiological barriers against reassortment of viral segments. Instead, our results suggest that epistatic fitness barriers are already substantial between more distant co-circulating strains of the same lineage, which ties in with the observation of cross-lineage pairing constraints [22, 24]. Hence, on larger scales of genetic distance, such fitness barriers may be an important factor in delineating—and, thus, defining—viral species.

Materials and methods

Alignments and genealogical trees

A sample of HA and NA sequences is obtained by downloading all the available A/H3N2 human strains in the EpiFlu DATABASE (http://www.gisaid.org), regardless the geographical region, which were collected between January 1968 and October 2015. Only the strains with complete HA and NA sequences are taken into consideration. Acknowledgements for sequences used in this study are available in S1 File. From this sample, alignments of single segments are created year by year using BLAST [48]. After discarding the segments with more than three gaps, a first run of RAxML [49] is performed to reconstruct the genealogy for each segment and detect clear outliers sequences (i.e. sequences that are found clearly isolated from the rest of the tree and which were likely misreported) to be excluded from the subsequent analysis. We then obtain alignments of linked HA-NA segment pairs and construct maximum-likelihood two-segment genealogies by RAxML, choosing the best-scoring ML tree out of 10 RAxML runs. Potential ambiguities in the definition of the nucleotide at a certain site in a strain are resolved by assigning the orthologous nucleotide of the closest ancestral node with an unambiguous sequence. The subsequent reassortment inference is performed on trees of subsampled data with a maximum of 600 sequences per year. This step reduces computational time and avoids over-representation of recent viruses, which are the most abundant in the database. We note that these joint genealogies differ from single-segment phylogenetic trees, because the underlying process of reassortment violates the tree topology.

Some of the isolates we use in this study were subject to passaging for amplification in cell culture. To test the possible role of passaging adaptations as a confounding factor [50], we repeat our inference of reassortment on 20 trees built from a restricted alignment of 1053 unpassaged sequences. We observe the same patterns of mutations which characterize the reassortment events reported above, as well as a very similar distance dependence of the false positives counts (S2 Fig). The robustness of our inference is expected, because we only consider events with mean genetic distance $d \geq 3$ between reassortant and parent strains in the reassorting segments.

Primary inference of reassortment events

Consider the tree representation of a reassortment event with parental strains p and p' and the reassortant strain r (Fig 2a). This representation defines one of the two segments, referred to as the travelling segment, and a set of mutations which generate the genetic distance between the parent strains p and p' in that segment. These so-called core mutations appear on the branches between the nodes p and p', which pass through their last common ancestor a. We define the set $\mathcal{A}_{pp'}$ of core mutations by counting the mutations from p to a in upward direction on the tree (filled red triangles) and the mutations from a to p' in downward direction (filled purple triangles). The branch from p' to r contains a set of mutations $\mathcal{A}_{p'r}$ that includes the set of reverse core mutations, denoted by $\bar{\mathcal{A}}_{pp'}$ (open red and purple triangles), as well as additional mutations in the travelling segment (grey triangles). These mutations reflect insufficient sampling (i.e., one of the actual parent strains is not included in the tree) and noise in the reconstruction of the phylogeny (see below), or they are point mutations unrelated to the reassortment event. Together, we obtain a criterion to detect reassortment in a two-segment tree: we parse the tree for node triplets (p, p', r) with

$$\mathcal{A}_{p'r} \supseteq \bar{\mathcal{A}}_{pp'} \tag{5}$$

in a given travelling segment.

Pruning steps: Uniqueness and false positives

The detection of reassortment on genealogical trees overcounts the biological reassortment events. First, we exclude false positive events due to ambiguities in tree reconstruction by statistical comparison with a null model. Since recombination within segments does not occur in influenza, we use as null case a set of trees built from the alignments of the single segments. We decompose the sequence of the protein into two subsets of randomly chosen sites with lengths L_1 and L_2. These subsets have the appropriate ratio of lengths to mimic the segment structure in the original joint alignment, $L_1/L_2 = L_{\mathrm{HA}}/L_{\mathrm{NA}}$. In order to investigate the dependence of the number of false positive reassortments on the length of the chain, we run the detection algorithm on subsets of sites of increasing total length $L = L_1 + L_2$, maintaining a constant ratio L_1/L_2. We find an expected number $n_0(\delta)$ of false positive reassortment events that decays rapidly with increasing core distance,

$$n_0(\delta) = Ce^{-\gamma\delta} \tag{6}$$

with $\gamma = 1.6 \pm 0.1$. The decay exponent γ is approximately independent of the total sequence length L (S1 Fig). We can then evaluate the expected number of false positive reassortment events in the actual data (Fig 3 and S1 Table). Even if we assume that most of the counts at $\delta = 1$ are false positives (which sets the value of the constant C in Eq 6); the expected total number of false positive events with $\delta \geq 5$ drops below 1 (Fig 3 and S1 Table).

Second, two or more different reassortment events reported by our algorithm may represent the same biological reassortment event if they have similar core set. To address this source of overcounting, we compare the core sets $\mathcal{A}_{pp'}$ of putative events. If these sets differ by at least 30% of their mutations, the events are considered independent; otherwise we keep only the set with the largest core distance δ. The number of pruned events turns out to be insensitive to moderate changes of the threshold number of mutations. As third step, we cluster the reported events with different travelling segments that have very similar parent strains.

Testing the inference method by simulations

We simulate the genome evolution of a population of N individuals starting at a stationary state, under the effect of mutation, genetic drift and selection, based on the model used in [5]. Each strain is characterized by a sequence of epitope and non-epitope sites, flanked by neutral sites. Selection on epitope sites is time-dependent and its direction fluctuates randomly at a rate γ, while non-epitope sites are modeled with time-independent direction. To these basic steps we add reassortment, which occurs at each generation with probability λ: we select randomly two individuals (the parents) and divide their genome into two parts of fixed length L_1 and L_2, then mimic the process of reassortment by creating a new individual (r) with a mixed genome. We focus on events between strains at genetic distance $d_{1,2} \geq 5$ in each segment, discarding reassortment at lower distances. The results of each simulation are a set of sampled sequences, some of them involved in a reassortment event, that we use to build up the genealogical trees, as we would do with real observed strains.

We choose the parameters of the simulations as follows:

1. The evolution of $N = 1000$ individuals is simulated for 1500 generations. Each individual has a genome of length $L = L_{ep}+L_{non-ep}+L_{neut} = 560$ ($L_{ep} = 120$, $L_{non-ep} = 160$, $L_{neut} = 300$ number of epitope, non epitope and neutral sites, respectively), selection flips the direction at rate $\gamma = 0.033$ and the mutation rate is set to $\mu = 5.8 \times 10^{-3}$ per year. With these evolution parameters, the population turns out to be in a clonal interference regime comparable to influenza [5].

2. We introduce reassortment at a rate rate $\lambda = 1 \times 10^{-6}$ per individual and per generation. This generates a density of reassortant variants at observable population frequencies that is comparable with the observed density in influenza A/H3N2.
With these parameters, we obtain trees that show $\sim$ 5 coalescent events on average, corresponding to approximately 10 years of influenza evolution. We apply our algorithm on each of the 100 reconstructed trees and check if the reassortment events recognizable in the sampled sequences (i.e. the ones with r and/or its offspring reaching a relevant frequency and therefore getting sampled) get detected. Out of the total 283 events generated in the simulations, 214 (76%) are correctly reported (see Fig 2c for an explicit example of a detected event), with 24 false positives signaled with small cores ($\delta \leq 5$).

Supporting information

S1 Fig. Distance dependence of spurious reassortment counts in non reassorting sequence. (a) Histograms of the number of events found in a HA tree as a function of δ, for sequences of total length L. Error bars represent the standard deviation obtained from 5 different random choices of the sites for each δ. (b) The decay exponent γ is shown as a function of L (cf. Materials and methods, Eq 6). The inferred values are stable for large values of L, allowing extrapolation to $L = L_{\mathrm{HA}} + L_{\mathrm{NA}}$.
(PDF)

S2 Fig. Reassortment inference between unpassaged sequences. Histograms of reported HA-NA reassortment events between unpassaged sequences for different core distances δ (red bars) are compared to expected number of false positives (blue bars), which decays exponentially with increasing δ. This result is qualitatively comparable with the distance dependence of real events and false positives that we find including in the analyses also strains subjected to passaging.
(PDF)

S3 Fig. Selection inference based on aminoacid distances. (a) The cumulative distribution of mean amino acid distances d between reassortant and parent strains for the HA-NA reassortments in influenza A/H3N2 (red line) is compared to the corresponding distribution of distances for co-circulating strains in the same influenza season (solid blue line) and from the New York area only (dashed blue line). (b) The ratio of reassortment counts to background counts in the interval $d \geq d_{\min}$ (red circles) decreases with increasing lower threshold $d_{\min}$ and drops significantly below 1 (blue line). The suppression of reassortment at larger values of d signals distance-dependent negative selection. Bars show statistical errors due to the finite number of inferred reassortments. See Fig 5 for the same analysis using nucleotide distances. (c) The average number of strains in the reassortant clades with aminoacid distance $\leq \tau_A$ from the focal node, $\langle N_r \rangle(\tau_A)$ (red line) is compared to the corresponding average number of strains in the parent clades, $\langle N_0 \rangle(\tau_A)$. For $\tau_A \lesssim 4$, both functions increase with τ_A in an approximately exponential way; we estimate growth rates $f_r(A) \approx 0.2$ and $f_0(A) \approx 0.7$, respectively (dashed lines; cf. Eq 1). The growth rate difference $\bar{s}_A \equiv f_0(A) - f_r(A) \approx 0.5$ inferred from distances in aminoacid units is similar to $\bar{s} \approx 0.4$ for nucleotide distances; cf. Fig 5c.
(PDF)

S4 Fig. Background distribution and reassortment events as a function of the amino acid distances d_{HA} and d_{NA} between strains. (a) The background distribution $P_0^{aa}(d_{\mathrm{HA}}, d_{\mathrm{NA}})$ (contour plot) is compared to reassortment counts (red dots). (b) Conditional background distributions $P_0^{aa}(d_{\mathrm{HA}} - d_{\mathrm{NA}} | d_{\mathrm{HA}} + d_{\mathrm{NA}})$ (whisker plots) are compared to reassortment counts (red dots). Whisker plots show the 0.25 quantile to the 0.75 quantile of the distribution (blue boxes); the white horizontal line represents the median, vertical bars span the dataset excluding outliers. The width of the bins is chosen to ensure a statistically relevant number of events for each bin. The reassortment data appear more spread in the coordinate $d_{\mathrm{HA}} - d_{\mathrm{NA}}$ compared to the background (red points are mainly placed outside or at the border of the blue boxes).
(PDF)

S1 Table. Number of expected false positive reassortment counts as a function of δ (cf. Fig 3).
(PDF)

S2 Table. List of inferred reassortment events from 1968 to 2015 between HA and NA segments in human influenza A/H3N2. Column 2: mean nucleotide distance d between reassortant strain and parent strains. Columns 3–5: representative observed strains in the clades of p, p' and r, respectively. Each isolate is identified by its number in the online EpiFlu DATABASE (http://www.gisaid.org) identifier (e.g. EPI_ISL_7064 is reported here as 7064). Stars indicate events which are reported in literature with large agreement (S1 Text).
(PDF)

S1 Text. Comparison with reassortment reported in literature.
(PDF)

S1 File. GISAID acknowledgement table.
(XLS)

S1 Code. Compressed code folder.
(ZIP)

Acknowledgments

We thank Simone Pompei, Marta Łuksza, and Daniel Klemmer for discussions and valuable comments on the manuscript.

Author Contributions

Writing – original draft: Mara Villa, Michael Lässig.

Writing – review & editing: Mara Villa, Michael Lässig.

References

1. Yamashita M, Krystal M, Fitch WM, Palese P (1988) Influenza B virus evolution: co-circulating lineages and comparison of evolutionary pattern with those of influenza A and C viruses. Virology 163(1): 112–122. https://doi.org/10.1016/0042-6822(88)90238-3 PMID: 3267218
2. Air GM, Gibbs AJ, Laver WG, Webster RG (1990) Evolutionary changes in influenza B are not primarily governed by antibody selection. Proc Natl Acad Sci USA 87(10): 3884–3888. https://doi.org/10.1073/pnas.87.10.3884 PMID: 2378639
3. Nobusawa E, Sato K (2006) Comparison of the Mutation Rates of Human Influenza A and B Viruses. J Virol 80(7): 3675–3678. https://doi.org/10.1128/JVI.80.7.3675-3678.2006 PMID: 16537638
4. Bedford T, et al (2014) Integrating influenza antigenic dynamics with molecular evolution. eLife 3: e01914. https://doi.org/10.7554/eLife.01914 PMID: 24497547
5. Strelkowa N, Lässig M (2012) Clonal interference in the evolution of influenza. Genetics 192(2) 671–682. https://doi.org/10.1534/genetics.112.143396 PMID: 22851649
6. Palese P, Tobita K, Ueda M, Compans RW (1974). Characterization of temperature sensitive influenza virus mutants defective in neuraminidase. Virology 61(2): 397–410. https://doi.org/10.1016/0042-6822(74)90276-1 PMID: 4472498
7. Liu C, Eichelberger MC, Compans RW, Air GM (1995). Influenza type A virus neuraminidase does not play a role in viral entry, replication, assembly, or budding. J Virol 69(2): 1099–1106. PMID: 7815489
8. Nelson MI, Holmes EC (2007) The evolution of epidemic influenza. Nature Reviews Genetics 8: 196–205. https://doi.org/10.1038/nrg2053 PMID: 17262054
9. Hampson AW (2002) Influenza virus antigens and 'antigenic drift'. Perspectives in Medical Virology 7: 49–85. https://doi.org/10.1016/S0168-7069(02)07004-0
10. Smith DJ, et al (2004) Mapping the antigenic and genetic evolution of influenza virus. Science 305 (5682): 371–376. https://doi.org/10.1126/science.1097211 PMID: 15218094
11. Li C, Hatta M, Nidom CA, Muramoto Y, Watanabe S, Neumann G, Kawaoka Y (2010) Reassortment between avian H5N1 and human H3N2 influenza viruses creates hybrid viruses with substantial virulence. PNAS 107(10): 4687–4692. https://doi.org/10.1073/pnas.0912807107 PMID: 20176961
12. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC (2008) The genomic and epidemiological dynamics of human influenza A virus. Nature 453: 615–619. https://doi.org/10.1038/nature06945 PMID: 18418375
13. Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, St George K, et al. (2005) Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS Biol 3(9): e300. https://doi.org/10.1371/journal.pbio.0030300 PMID: 16026181
14. Webster RG, Laver WG, Air GM, Schild GC (1982) Molecular mechanisms of variation in influenza viruses. Nature 296: 115–121. https://doi.org/10.1038/296115a0 PMID: 6174870
15. Belshe RB (2005) The Origins of Pandemic Influenza—Lessons from the 1918 Virus. N Engl J Med 353:2209–2211. https://doi.org/10.1056/NEJMp058281 PMID: 16306515
16. Lubeck MD, Palese P, Schulman JL (1979) Nonrandom association of parental genes in influenza A virus recombinants. Virology 95(1):269–274. https://doi.org/10.1016/0042-6822(79)90430-6 PMID: 442543
17. Hatta M et al (2002) Human influenza A viral genes responsible for the restriction of its replication in duck intestine. Virology 295(2): 250–255. https://doi.org/10.1006/viro.2002.1358 PMID: 12033783
18. Maines TR et al (2006) Lack of transmission of H5N1 avian–human reassortant influenza viruses in a ferret model. PNAS 103(32): 12121–12126. https://doi.org/10.1073/pnas.0605134103 PMID: 16880383

19. Li C, Hatta M, Watanabe S, Neumann G, Kawaoka Y (2008) Compatibility among polymerase subunit proteins is a restricting factor in reassortment between equine H7N7 and human H3N2 influenza viruses. J Virol 82(23): 11880–11888. https://doi.org/10.1128/JVI.01445-08 PMID: 18815312

20. Varich NL, Gitelman AK, Shilov AA, Smirnov YA, Kaverin NV (2008) Deviation from the random distribution pattern of influenza A virus gene segments in reassortants produced under non-selective conditions. Arch Virol 153(6): 1149–1154. https://doi.org/10.1007/s00705-008-0070-5 PMID: 18414973

21. Octaviani CP, Ozawa M, Yamada S, Goto H, Kawaoka Y (2010) High level of genetic compatibility between swine-origin H1N1 and highly pathogenic avian H5N1 influenza viruses. J Virol 84(20): 10918–10922. https://doi.org/10.1128/JVI.01140-10 PMID: 20686037

22. Greenbaum BD et al (2012) Viral reassortment as an information exchange between viral segments. PNAS 109(9): 3341–3346. https://doi.org/10.1073/pnas.1113300109 PMID: 22331898

23. Neverov AD, Lezhnina KV, Kondrashov AS, Bazykin GA (2014) Intrasubtype reassortments cause adaptive amino acid replacements in H3N2 influenza genes. PLoS Gen 10(1): e1004037. https://doi.org/10.1371/journal.pgen.1004037

24. Dudas G, Bedford T, Lycett S, Rambaut A (2015) Reassortment between influenza B lineages and the emergence of a coadapted PB1-PB2-HA gene complex. Mol Biol Evol 32(1): 162–72. https://doi.org/10.1093/molbev/msu287 PMID: 25323575

25. Schweiger B, Bruns L, Meixenberger K (2006) Reassortment between human A(H3N2) viruses is an important evolutionary mechanism. Vaccine 24(44–46): 6683–6690. https://doi.org/10.1016/j.vaccine.2006.05.105 PMID: 17030498

26. Lindstrom SE, Hiromoto Y, Nerome R, et al. (1998) Phylogenetic analysis of the entire genome of influenza A (H3N2) viruses from Japan: evidence for genetic reassortment of the six internal genes. J Virol 72(10): 8021–8031. PMID: 9733841

27. De Silva UC, Tanaka H, Nakamura S, Goto N, Yasunaga T (2012) A comprehensive analysis of reassortment in influenza A virus. Biol. Open 1: 385–390. https://doi.org/10.1242/bio.2012281 PMID: 23213428

28. Rabadan R, Levine AJ, Krasnitz M (2008) Non-random reassortment in human influenza A viruses. Influenza Other Respir Viruses 2(1): 9–22. https://doi.org/10.1111/j.1750-2659.2007.00030.x PMID: 19453489

29. Malijkovic Berry I, et al (2016) Frequency of influenza H3N2 intra-subtype reassortment: attributes and implications of reassortant spread. BMC Biology 14: 117. https://doi.org/10.1186/s12915-016-0337-3

30. Nelson MI, Simonsen L, Viboud C, Miller MA, Taylor J, St. George K, et al. (2006) Stochastic processes are key determinants of short-term evolution in influenza A virus. PLoS Pathog 2(12): e125. https://doi.org/10.1371/journal.ppat.0020125 PMID: 17140286

31. Nelson MI, Viboud C, Simonsen L, et al. (2008) Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. PLoS Pathog 4(2): e1000012. https://doi.org/10.1371/journal.ppat.1000012 PMID: 18463694

32. Nagarajan N, Kingsford C (2011) GiRaF: robust, computational identification of influenza reassortments via graph mining. Nucleic Acids Res 39(6): e34. https://doi.org/10.1093/nar/gkq1232 PMID: 21177643

33. Svinti V, Cotton JA, McInerney JO (2013) New approaches for unravelling reassortment pathways. BMC Evol Biol 13: 1. https://doi.org/10.1186/1471-2148-13-1 PMID: 23279962

34. Yurovsky A, Moret BME (2011) FluReF, an automated flu virus reassortment finder based on phylogenetic trees. BMC Genomics 12: S3. https://doi.org/10.1186/1471-2164-12-S2-S3 PMID: 21989112

35. Suzuki Y (2010) A phylogenetic approach to detecting reassortments in viruses with segmented viruses. Gene 464(1–2): 11–6. https://doi.org/10.1016/j.gene.2010.05.002 PMID: 20546849

36. Pinsent A, Fraser C, Ferguson NM, Riley S (2016) A systematic review of reported reassortant viral lineages of influenza A. BMC Infectious Diseases 16(1): 1–13 https://doi.org/10.1186/s12879-015-1298-9.

37. Westgeest KB et al. (2012) Genetic evolution of the neuraminidase of influenza A (H3N2) viruses from 1968 to 2009 and its correspondence to haemagglutinin evolution. J Gen Virol 93(9): 1996–2007. https://doi.org/10.1099/vir.0.043059-0 PMID: 22718569

38. Westgeest KB et al. (2014) Genomewide analysis of reassortment and evolution of human influenza A (H3N2) viruses circulating between 1968 and 2011. J Virol 88(5): 2844–2857. https://doi.org/10.1128/JVI.02163-13 PMID: 24371052

39. Neverov AD, Kryazhimskiy S, Plotkin JB, Bazykin GA (2015) Coordinated evolution of influenza A surface proteins. PLoS Gen 11(8): e1005404. https://doi.org/10.1371/journal.pgen.1005404

40. Monto AS et al. (2015) Antibody to influenza virus neuraminidase: an independent correlate of protection. J Infect Dis 212(8): 1191–1199. https://doi.org/10.1093/infdis/jiv195 PMID: 25858957

41. Couch RB et al. (2013) Antibody correlates and predictors of immunity to naturally occurring influenza in humans and the importance of antibody to the neuraminidase. J Infect Dis 207(6): 974–981. https://doi.org/10.1093/infdis/jis935 PMID: 23307936

42. Russell CA et al. (2008) The global circulation of seasonal influenza A (H3N2) viruses. Science 320(5874): 340–346. https://doi.org/10.1126/science.1154137 PMID: 18420927

43. Łuksza M, Lässig M (2014) A predictive fitness model for influenza. Nature 507: 57–61. https://doi.org/10.1038/nature13087 PMID: 24572367

44. Neher RA, Russell CA, Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. eLife 3: e03568. https://doi.org/10.7554/eLife.03568

45. Zawadzki P, Roberts MS, Cohan FM (1995) The log-linear relationship between sexual isolation and sequence divergence in Bacillus transformation is robust. Genetics 140(3): 917–32. PMID: 7672591

46. Ambur OH, Frye SA, Nilsen M, Hovland E, Tønjum T (2012) Restriction and sequence alterations affect DNA uptake sequence-dependent transformation in neisseria meningitidis. PLoS ONE 7(7): e39742. https://doi.org/10.1371/journal.pone.0039742 PMID: 22768309

47. Gangel H, Hepp C, Müller S, Oldewurtel ER, Aas FE, Koomey M, et al. (2014) Concerted spatio-temporal dynamics of imported DNA and ComE DNA uptake protein during gonococcal transformation. PLoS Pathog 10(4): e1004043. https://doi.org/10.1371/journal.ppat.1004043 PMID: 24763594

48. Altschul SF, et al (1990) Basic Local Alignment Search Tool. J Mol Biol 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

49. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu033

50. McWhite CD, Meyer AG, Wilke CO (2016) Sequence amplification via cell passaging creates spurious signals of positive adaptation in influenza virus H3N2 hemagglutinin. Virus Evol. 2(2): vew026. https://doi.org/10.1093/ve/vew026 PMID: 27713835