JVI Journals.ASM.org

Computational Prediction of Vaccine Strains for Human Influenza A (H3N2) Viruses

L. Steinbrück,[a] T. R. Klingen,[a,b] A. C. McHardy[a,b]

Department for Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, Germany[a]; Department for Computational Biology of Infection Research, Helmholz Centre for Infection Research, Braunschweig, Germany[b]

ABSTRACT

Human influenza A viruses are rapidly evolving pathogens that cause substantial morbidity and mortality in seasonal epidemics around the globe. To ensure continued protection, the strains used for the production of the seasonal influenza vaccine have to be regularly updated, which involves data collection and analysis by numerous experts worldwide. Computer-guided analysis is becoming increasingly important in this problem due to the vast amounts of generated data. We here describe a computational method for selecting a suitable strain for production of the human influenza A virus vaccine. It interprets available antigenic and genomic sequence data based on measures of antigenic novelty and rate of propagation of the viral strains throughout the population. For viral isolates sampled between 2002 and 2007, we used this method to predict the antigenic evolution of the H3N2 viruses in retrospective testing scenarios. When seasons were scored as true or false predictions, our method returned six true positives, three false negatives, eight true negatives, and one false positive, or 78% accuracy overall. In comparison to the recommendations by the WHO, we identified the correct antigenic variant once at the same time and twice one season ahead. Even though it cannot be ruled out that practical reasons such as lack of a sufficiently well-growing candidate strain may in some cases have prevented recommendation of the best-matching strain by the WHO, our computational decision procedure allows quantitative interpretation of the growing amounts of data and may help to match the vaccine better to predominating strains in seasonal influenza epidemics.

IMPORTANCE

Human influenza A viruses continuously change antigenically to circumvent the immune protection evoked by vaccination or previously circulating viral strains. To maintain vaccine protection and thereby reduce the mortality and morbidity caused by infections, regular updates of the vaccine strains are required. We have developed a data-driven framework for vaccine strain prediction which facilitates the computational analysis of genetic and antigenic data and does not rely on explicit evolutionary models. Our computational decision procedure generated good matches of the vaccine strain to the circulating predominant strain for most seasons and could be used to support the expert-guided prediction made by the WHO; it thus may allow an increase in vaccine efficacy.

In addition to influenza pandemics that have caused up to 50 million deaths (1), human influenza A viruses are responsible for substantial morbidity and mortality worldwide (2). Of the three distinct genera (A, B, and C), type A viruses evolve the most rapidly and cause the majority of infections (3, 4). The influenza A virus genome consists of eight single-stranded negative-sense RNA molecules that encode one or more proteins each (5–7). The viruses are further classified into subtypes based on the composition of the surface glycoproteins hemagglutinin (H or HA serotypes 1 to 18) and neuraminidase (N or NA serotypes 1 to 11) that occur in various combinations in viruses of different hosts (8–10). Currently, influenza A viruses of the subtypes H1N1, referred to as influenza A (H1N1) pdm09, and H3N2 are endemic in the human population (11).

Human influenza A viruses continuously change antigenically to circumvent the immune protection elicited by vaccination or previously circulating viral strains. This antigenic drift is caused by amino acid changes, mainly in the antibody-binding (epitope) sites of HA and NA (12–14), and results in the regular appearance of novel antigenic variants, against which cross-protective immunity in the human population is reduced (14). To track the genetic and antigenic composition of the globally circulating viral population, the World Health Organization (WHO) runs the Global Influenza Surveillance and Response System (GISRS) (15). The collected information is continuously evaluated by a panel of experts, who decide twice a year on the composition of the influenza vaccine. Currently, four strains are included, one strain each of the influenza A (H1N1)pdm09, influenza A (H3N2), influenza B (Yamagata lineage), and influenza B (Victoria lineage) viruses (16). This decision is made in February for the following year's Northern Hemisphere (NH) winter season and in September for the following year's Southern Hemisphere (SH) winter season, to allow for sufficient time for vaccine production. In general, this approach results in a good match of the vaccine strain to the cir-

Received 14 July 2014 **Accepted** 5 August 2014
Published ahead of print 13 August 2014
Editor: T. S. Dermody
Address correspondence to A. C. McHardy, mchardy@hhu.de.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.01861-14.

doi:10.1128/JVI.01861-14

culating predominant strain and significantly reduces mortality and morbidity (17).

Several predictive properties for the genetic and antigenic evolution of influenza A viruses are known, such as variation at specific HA positions or changes in charge on the protein surface, in particular within the antibody-binding sites of HA (13, 18–25). Some methods incorporate both genetic and antigenic data to predict the antigenic novelty of viral strains (26–28). The antigenic phenotype of viral strains can be quantified with the hemagglutination inhibition (HI) assay, which measures the antigenic similarity of two viral isolates based on the inhibition of the agglutination of red blood cells (caused by a viral antigen) by an antiserum (29). Note that changes in HI data not only reflect antiserum-based hemagglutination inhibition but also may be influenced by alterations in virus receptor avidity resulting, for instance, from an acquired capability of neuraminidase to agglutinate red blood cells (30). Smith et al. developed "antigenic cartography," a method which is based on multidimensional scaling and allows one to visualize and quantify the antigenic differences between different antigens from HI assay data in a two-dimensional map (14). Applying this to influenza A (H3N2) virus isolates sampled over 35 years showed that the antigenic evolution of the virus is clustered, with an antigenic cluster being predominant for 3.3 years, on average, before being replaced by a novel antigenic cluster. The amount of available surveillance data has increased in recent years, and expert evaluation of the epidemiological, antigenic, and genetic data is now guided by phylogenetic analysis and antigenic cartography (31), resulting in the proposal of mostly well-suited vaccine strains. For the first year when an antigenically novel strain rises to predominance, the selection of the best-matching viral strain for production of the influenza A virus vaccine remains a challenge. In a recent study, we compared the WHO's vaccine strain recommendations to the reported predominant viral strains in seasonal epidemics (32). This showed that following WHO recommendations, the vaccine composition was in many cases updated only after a novel antigenic strain became predominant, resulting in a vaccine strain mismatch and reduced vaccine efficacy for the first one or two seasons.

We have previously described allele dynamics plots (AD plots), which visualize the evolutionary dynamics of the different alleles of a gene in a population over time and indicate the alleles that are most likely to be subject to directional selection (32). The merits of this technique for the identification of sets of coding changes conferring a selective advantage on a viral strain were demonstrated in a study of the hemagglutinin of influenza A (H3N2) virus isolates sampled between 1998 and 2008. In four out of five test seasons, the AD plots allowed correct identification of the alleles and their associated viral strains that subsequently became predominant in the viral population. A limitation of AD plots is that a particular allele scores best in every season, regardless of whether its antigenic characteristics are distinct from those of the current predominant strain or not. Evaluation of the antigenic impact of the selected allele can help to resolve this issue.

Recently, we developed a method for the inference of "antigenic trees" (33). Using nonnegative least-squares optimization, we mapped pairwise antigenic distances onto the branches of a phylogenetic tree. This resulted in the inference of antigenic weights for the individual branches of the tree and allowed antigenic weights to be determined for sets of coding changes in HA. In this work, we combined AD plots and antigenic trees to identify antigenically distinct HA alleles and the associated viral strains that are on the rise to predominance in the viral population. Using genetic and antigenic data for influenza A (H3N2) virus isolates sampled between 2002 and 2007, we demonstrate how this allows us to predict the genetic and antigenic evolution of the virus, which enables a straightforward application of our method in the annual vaccine strain recommendation process (17).

MATERIALS AND METHODS

Genomic data. HA1 domain sequences of the hemagglutinin segment for 1,431 seasonal human influenza A (H3N2) virus isolates sampled between 1995 and 2007 and used by Russell et al. (34) were downloaded from the Influenza Virus Resource (35) (see Table S1 in the supplemental material). Of these, 54 sequences that were represented as antigens in the data (see below) and/or had partial sampling information (missing month and day of sampling) were excluded.

Antigenic data. HI assay data from Russell et al. (34) were normalized according to the procedure used by Smith et al. (14). For each antigen i, antiserum j and the corresponding HI titer $h_{i,j}$, the distance was set as $d_{i,j} = \log_2[max(h_j)/h_{i,j}]$, where $max(h_j)$ is the maximum entry for antiserum j. Antigens and reference sera for which no HA sequence was available were excluded from the analysis. Additionally, threshold values (e.g., <40; these values indicate the lower bound in the HI assay, which was the lowest tested dilution) were excluded to avoid bias introduced by setting these entries to fixed values. Multiple measurements of antigen-antiserum distances were available when antigens and antisera raised against a viral strain were tested in multiple laboratories or at several time points, or when multiple antisera were raised against the same strain. For multiple measurements, the median of these distances was used. The resulting antigenic data set comprised 11,564 distances between 1,377 antigens and 82 reference sera.

Predicting suitable HA alleles for the influenza A (H3N2) virus vaccine. We developed a method to predict the most suitable strain for production of the seasonal influenza A (H3N2) virus vaccine by identifying antigenically novel HA alleles that are on the rise to future predominance. The method involves (i) reconstructing a phylogenetic tree, (ii) constructing an AD plot from this tree and using isolate sampling times to identify the three HA alleles that are most likely to become predominant in the future (32), (iii) constructing an antigenic tree from the phylogenetic tree and HI distances and identifying the antigenic impact for the three HA alleles (33), and (iv) if an antigenically novel HA allele (with an antigenic weight of at least 0.5 antigenic unit) was identified as being likely to become predominant, proposal of the corresponding strain for inclusion into the vaccine for the influenza season in the following year. We chose an average antigenic weight of 0.5, as this gave us a good trade-off in detecting type-defining branches that indicate a true antigenic transition (33) (see below for a more detailed explanation). If no antigenically novel HA allele is identified as being likely to become predominant, we predict that no update of the vaccine should be undertaken. Steps ii and iii of our method were performed as described previously and are summarized below (32, 33). To simulate realistic testing conditions, we applied our method in a retrospective testing scenario to the data available until the end of each individual influenza season and, like the WHO, made predictions based on recent available information for the future influenza season 1 year ahead.

Data preprocessing. For reference sera generated from viral isolates without complete sampling information, 1 year was added to the specified sampling time to prevent including these data in the retrospective testing analysis earlier than they may have been actually sampled. For instance, the timestamp of a viral isolate with a sampling time of 2004-00-00 was set to 2005-00-00, and thus, the viral isolate was used as early as possible for inference of the phylogenetic tree for the 2004-2005 NH influenza season. Because of the uncertainty, sequences of these reference sera were excluded from the allele frequency calculation and antigenic analysis for the corresponding HA alleles. Influenza seasons were defined as the Northern

Hemisphere influenza season (from 1 October to 31 January) and the SH influenza season (from 1 April to 31 August) as before (32). The WHO decides on the composition of the next influenza vaccine to be produced and whether updates of the vaccine strains are necessary at the end of the respective hemisphere's winter season (17). Data from after this point (February and March in the NH winter season and September in the SH winter season) were excluded, to obtain a data set similar to the one used by the WHO for its decision on the vaccine strains. Note that this does not exclude the possibility that a few strains sampled late during this period were not available for the analysis in reality, due to the time required for sample shipping and processing.

Tree inference. For each season from 2002 to 2007, starting with the 2002-2003 NH season, we reconstructed a phylogenetic tree based on the viral sequences available until the end of the respective season (as defined under data preprocessing) and not sampled earlier than 2 years prior to that season. Additionally, HA sequences from strains used to generate reference sera in the past were included. Alignments of RNA and protein sequences were created with Muscle (36) and manually curated. Phylogenetic trees were inferred with PhyML v3.0 (37) under the general time reversal GTR+I+Γ_4 model, with the frequency of each substitution type, the proportion of invariant sites (I), and the Gamma distribution of among-site rate variation (with four rate categories [Γ_4]) estimated from the data. Subsequently, the tree topology and branch lengths of the maximum likelihood tree inferred with PhyML were optimized for 200,000 generations with Garli v0.96b8 (38). Isolate A/Wuhan/359/1995 was used as an outgroup to root the phylogenetic tree.

AD plots. Ancestral character states for the HA tree were reconstructed under the parsimony model using Fitch's algorithm (39). Any other available method for ancestral character state reconstruction can be applied (e.g., maximum likelihood or Bayesian inference [40, 41]); however, previously, we found few differences for H3N2 ancestral character states reconstructed with different techniques (33). Based on the differences in ancestral character states between each pair of parental and descendant nodes, synonymous and nonsynonymous mutations were mapped to the edges of the tree. From this, HA alleles were defined, each corresponding to a nonempty set of mutations associated with an individual branch. We restricted our analysis to nonsynonymous mutations causing amino acid changes in the antibody-binding (epitope) sites (42, 43), as changes in these regions cause the largest antigenic change (44, 45), are under positive selection, and are most relevant for the adaptive evolution of human influenza A viruses (20). We applied AD plots as described in reference 32 to analyze variations in epitope sites only, as changes in these sites are most relevant for changing the antigenic properties of a given isolate and allowed us to predict newly emerging antigenic variants accurately 1 year in advance. We use the following nomenclature for an allele: *allele substitution *substitutions of parental alleles from the same time period**, e.g.*156H *75Q, 155T**. The allele frequency for a specific season was estimated based on the ratio of the number of isolates in the subtree belonging to the allele-associated tree branch relative to the number of all isolates sampled within the season. Accordingly, the increase in allele frequency indicates that the affected allele is more likely to provide a selective advantage compared to others. For each season, we identified the three candidate alleles that were most likely to become predominant in the future, corresponding to those alleles rising most rapidly in frequency in comparison to the preceding season with an increase in frequency of at least 5%. This threshold was applied to remove low-abundance alleles with larger stochastic fluctuations in abundance from further consideration. Furthermore, we required that these alleles were not predominant before (frequency $<$ 50%). For identified candidate alleles, we determined the overall antigenic impact of the allele-associated nonsynonymous changes, as described below.

Antigenic trees. Antigenic trees were inferred by mapping antigenic distances to the branches of a phylogenetic tree that had been reconstructed from the associated genetic sequences of HA for the respective viral isolates with nonnegative least-squares optimization. This resulted in the inference of antigenic weights for the individual branches of the phylogenetic tree. Antigenic weights are represented as two independent weights for each branch to account for the asymmetric nature of the antigenic distances (the HI titer for an antigen of viral strain A to the antiserum raised against strain B may be different from the titer of the antigen of strain B to the antiserum raised against strain A). For each season, the three top-ranking HA alleles in terms of their increase in prevalence within consecutive seasons were considered for antigenic validation. In case parental edge substitutions were included in an allele's definition, antigenic weights for an allele only were used, not those for the parental edge, as these are the only ones specific to a particular allele, whereas parental substitutions may be shared. The threshold for the detection of antigenically relevant HA alleles was set to 0.5 antigenic unit and used to predict HA alleles for future vaccine strain construction. This threshold allows the detection of HA alleles that define antigenic variants as well as HA alleles that account for minor antigenic changes that still necessitate a vaccine update (33). Note that 0.5 is lower than the threshold of 2.0 (4-fold dilution) used by the WHO, as it indicates individual edges of antigenic relevance, while the latter is similar to the sum of multiple edge weights between pairs of antigenically distinct isolates in our tree. Alleles were linked to antigenic strains described in the literature by genetic changes as in reference 32. Antigenic strains are denoted by their commonly used abbreviations, namely, MO99, FU02, WE04, CA04, WI05, and BR07 (32).

RESULTS

We predicted the most suitable strains of the seasonal influenza A (H3N2) viruses to include in the seasonal influenza vaccine based on our estimates of their antigenic novelty and whether they would rise to predominance within 1 year with a retrospective testing scenario (see Materials and Methods). Our data set comprised genetic sequences for the HA gene and antigenic information in the form of HI titers for 1,377 viral isolates, as used by Russell et al. (34). This is a representative sample of the viral population worldwide for the study period. Starting with the 2002-2003 NH season, we inferred maximum likelihood phylogenetic trees (Fig. 1) for each season from 2002 to 2007 using the data collected within the 2 years preceding that particular season. Analysis of HA allele mutations was restricted to those resulting in amino acid changes in the antibody-binding sites, as in reference 32, as these are the most relevant for antigenic evolution (43–46) and are under positive selection (20). The phylogenetic trees were used to construct AD plots and to identify the HA alleles which had the largest increase in prevalence relative to the previous season and were not predominant ($<$50%) before. Assuming that alleles with a selective advantage rise faster in frequency than those without a selective advantage, those that increase the fastest in frequency of all sampled alleles are most likely to be subject to directional selection and to become predominant in the future (32). We determined the antigenic impact of the allele-associated amino acid changes for the three top-ranking HA alleles using antigenic trees. The alleles most likely to be on the rise to predominance with an estimated antigenic impact sufficient to warrant a vaccine strain update were proposed as vaccine strain components for the influenza season of the following year (Table 1). If no such allele was identified, we predicted that the vaccine composition should be left unchanged.

For performance evaluation, we applied standard methodology for evaluating predictive performance in binary classification problems: If an antigenically novel viral clade did become predominant 1 year later, we considered the associated HA allele to be an example of the "positive class," representing strains suitable to

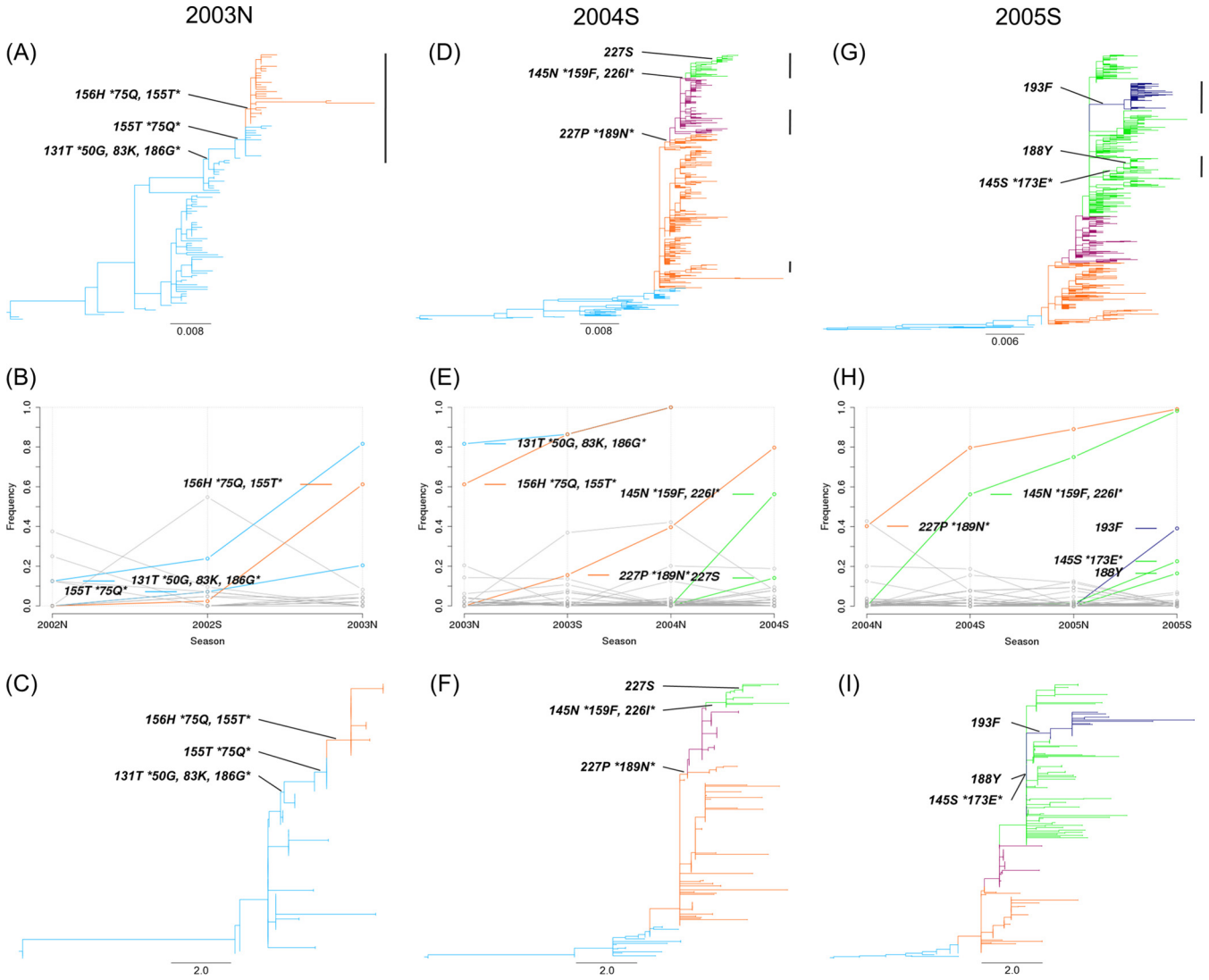

FIG 1 Data analysis in influenza seasons with replacement of the predominant antigenic strain. Columns represent results for the 2002-2003 Northern Hemisphere influenza season (2003N), the 2004 Southern Hemisphere influenza season (2004S), and the 2005 Southern Hemisphere influenza season (2005S). Panels A, D, and G give the maximum likelihood phylogenetic trees inferred for the 2003N, 2004S, and 2005S influenza seasons, respectively. Colors represent known antigenic strains identified by key amino acid substitutions reported in the literature and used before (32): SY97/MO99/PA99 (light blue), FU02 (orange), WE04 (violet), CA04 (green), WI05 (dark blue), and BR07 (yellow). Horizontal bars indicate clades that contain viral isolates sampled in the relevant influenza season. Panels B, E, and H depict AD plots computed for the 2003N, 2004S, and 2005S influenza seasons, respectively. Alleles with a frequency of >90% or with a frequency increase of ≥10% in the relevant influenza season are shown in color (colors were arbitrarily chosen). Panels C, F, and I show the antigenic trees inferred for the 2003N, 2004S, and 2005S influenza seasons, respectively. Colors are as in panels A, D, and G. Branch lengths represent the maximum of the two branch weights (up- and down-weights). Weights for terminal branches and branches leading to subtrees without an isolate used as antiserum are set to 0 antigenic units for the sake of clarity.

be selected for a vaccine strain update. Positive examples are viral strains that rise to predominance in the next season. All remaining viral isolates and associated HA alleles, which do not represent viral strains that become predominant in the following year, represent examples of the negative class, i.e., strains that are not suitable for a vaccine strain update. In general, a vaccine update should be recommended only if in the next season an antigenically novel strain becomes predominant. In our prediction, we considered alleles with a predicted antigenic impact of more than 0.5 antigenic unit, for either their up- or down-weight, as sufficiently antigenically novel to be predicted positive, i.e., recommended for a vaccine strain update. We chose 0.5 antigenic unit, as in the tree of the current data set antigenic changes are well resolved to individual branches, due to the large number of available sequences and antigenic distances. If considering the joint impact of multiple successive branches, higher thresholds for antigenic units might be sensible to use. If parental edge substitutions were included in an allele's definition, antigenic weights for an allele were used only, not those for the parental edge, as parental substitutions are shared with other alleles. All other alleles were predicted as negatives. A comparison of these predictions to the underlying truth results in four categories for allele assignments: true positives (positive alleles predicted as being positive), true negatives (negative alleles predicted as being negative), false positives (negative

TABLE 1 Genetic and antigenic properties of HA alleles increasing in prevalence for influenza seasons between 2002 and 2007[a]

Season	HA allele(s)	Frequency increase	Antigenic wt (up/down)	Comment	WHO	Predominant 1 year later
2002/03 NH	*156H *75Q, 155T**	0.59	0.82/0.28	FU02, match	FU02 (57)	FU02 (51)
	*131T *186G**	0.58	0.12/0.00			
	*155T *75Q**	0.13	0.00/0.41			
2003 SH	*126D*	0.39	0.00/0.20		FU02 (47)	FU02 (52)
	227P	0.16	0.00/0.00			
	193N	0.08	0.00/0.00			
2003-2004 NH	*227P *189N**	0.24	0.00/0.08		FU02 (51)	CA04 (48)
	159F	0.20	0.38/0.00			
	140E	0.13	2.97/0.00	False positive		
2004 SH	*145N *159F, 226I**	0.56	0.67/0.12	CA04, match	False positive WE04 (52)	CA04 (62)
	*227P *189N**	0.40	0.00/0.25			
	227S	0.14	0.00/0.27			
2004-2005 NH	*188N *159F, 226I, 145N**	0.09	0.00/0.30		CA04 (48)	CA04 (53)
	278K	0.08	0.00/0.37			
	*226V*216S**	0.05	0.00/—			
2005 SH	*193F*	0.39	0.79/0.44	WI05, match	CA04 (62)	WI05 (49)
	*145S *173E**	0.21	0.00/—			
	188Y	0.17	0.00/—			
2005-2006 NH	*193F*	0.18	0.00/1.23	Selected before	WI05 (53)	WI05 (54)
	50E	0.18	0.00/0.00			
	198T, 310R	0.12	0.00/0.00			
2006 SH	*50E*	0.24	0.00/0.00		WI05 (49)	BR07 (50)
	144D	0.06	0.00/—			
	*50E, 157S *128A, 142G, 173E**	0.05	0.00/0.00			
2006-2007 NH	*144D*	0.22	0.00/—		WI05 (54)	BR07 (63)
	142G	0.22	0.00/—			
	*128A *157S, 142G, 173E**	0.14	0.00/0.00			
	[*50E, 140I*]	0.04	0.90/0.00	BR07, match		

[a] For each season, the three top-ranking HA alleles are shown ordered by their increase in frequency relative to the preceding influenza season. For each allele, the respective antigenic weights (up and down [33]) are given. The columns "WHO" and "Predominant" give the recommended vaccine strain and the predominant viral strain for the influenza season 1 year later. HA alleles with high antigenic weights (up or down) of at least 0.5 antigenic unit are indicated by gray shading. "Match" indicates a true positive. NH, Northern Hemisphere; SH, Southern Hemisphere. —, no down-weight. Numbers in parentheses are reference numbers.

alleles predicted as being positive), and false negatives (positive alleles predicted as being negative). Performance is optimal if no false positives or false negatives are obtained. For this particular problem, false positives, resulting in production of a mismatching vaccine, would have a more negative cost than false negatives, in which the vaccine is not updated. This is because in addition to the vaccine mismatch obtained in both cases, an inefficient vaccine would be produced and distributed in the case of a false-positive prediction.

Within the nine influenza seasons from 2002 to 2007, four antigenically distinct strains successively became predominant, known as FU02, CA04, WI05, and BR07 (47–50). The HA alleles of these viruses represent the positive examples. All other HA alleles represent negative examples. We identified alleles representing three of four positive examples correctly, namely, for the MO99-FU02 transition in the 2002-2003 NH season, the FU02-CA04 transition in the 2004 SH season, and the CA04-WI05 transition in the 2005 SH season (Fig. 1 and 2). Overall, for the 27 HA alleles (the three top-ranked alleles in the AD plots for the nine influenza seasons tested) (Table 1), only one false positive (*140E*) and one false negative (*50E*, *140I*) were predicted, resulting in an accuracy of 93%. When seasons were scored as true or false predictions according to the predicted predominant strain, our method resulted in six true positives, three false negatives, eight true negatives, and one false positive (78% accuracy). In compar-

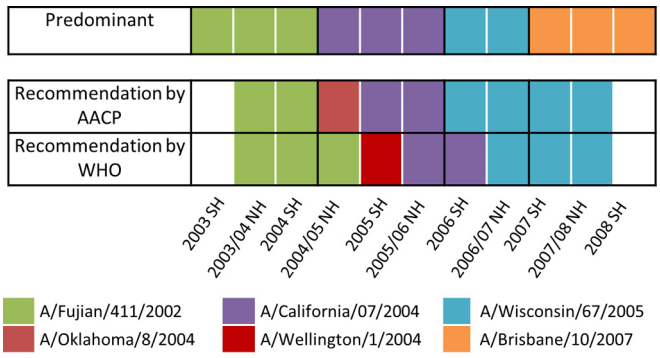

FIG 2 Performance evaluation of antigenic allele-based computational prediction of vaccine strains (AACP) for human influenza A (H3N2) based on the combination of AD plots and antigenic trees and comparison with recommendations by the World Health Organization. In contrast to Table 1, now antigenic strains are shown for the seasons in which they were predominant; thus, for both methods, predictions are shown for the year in which the vaccine was made available, not for the year before (when it was to be produced). The top row shows a succession of predominant and antigenically distinct strains. The second row shows the recommendations made by the AACP, while the third row shows the recommendations made by the WHO, both for the 2003-2004 NH influenza season and for the season 2007-2008 NH influenza season. This figure illustrates that the predominant strains were predicted correctly in six out of nine seasons by the AACP and the recommendation made by the WHO matched four out of nine seasons. Both recommendations included one false positive, as the recommended *140E* HA allele (A/Oklahoma/8/2004) predicted by AACP and the WE04 strain recommended by the WHO did not belong to the positive sample. Note that a vaccine update was necessary only if in the previous season a different strain was predominant.

ison to the recommendations by the WHO, we identified the correct antigenic variant once at the same time (in the 2002-2003 NH season) and twice one season ahead of the WHO (in the 2004 SH season and in the 2005 SH season) (Fig. 2). Overall, four true-positive, five false-negative, eight true-negative, and one false-positive prediction (66% accuracy) were made by the WHO. Previously, we found that based on the available data, antigenically novel strains rose to predominance within a single year or even a single season from the time when they were first observed (32). Therefore, for cases where identification of the correct strain was delayed for one season with our method, we believe that it is unlikely that predictions could be further improved.

FU02 was predominant from 2003 to 2004-2005 (47, 51, 52). The associated HA allele with coding changes *156H *75Q, 155T** ranked first in the 2002-2003 NH season and had antigenic weights of 0.82 (up) and 0.28 (down) and thus was correctly predicted by our method as a suitable candidate for a vaccine strain update for the 2003-2004 NH season. Previously, based on our predictions of future predominant HA alleles by using AD plots only, we could not correctly identify FU02; this shows the additional value of antigenic information (32). The FU02 strain was recommended by the WHO as a vaccine strain in the same season as with our method.

FU02 was later replaced by CA04 in the 2004-2005 NH season (48). The CA04 HA allele with changes *145N *159F, 226I** ranked first in the AD plot in the 2004 SH season with antigenic weights of 0.67 (up) and 0.12 (down). It thus was predicted as a vaccine strain update for the 2005 SH season—one season late. This allele was not sampled in the 2003-2004 NH season, which would have been 1 year prior to its predominance, and thus accounts for a false negative of our method for the 2004-2005 NH season. Instead, the 140E allele was falsely predicted to be predominant for the 2004-2005 NH season. The WHO recommended the WE04 strain in the 2004 NH season. The WE04 strain recommended by the WHO for the 2004 SH season is distinct from CA04 (52) but could be considered an intermediate in terms of antigenicity between the previous FU02 strain and actual new CA04 strain. However, as WE04 is not the antigenically identical to CA04, we counted it as a false positive. The WHO predicted the CA04 strain for inclusion in the influenza vaccine in the 2004-2005 NH season—two seasons late—thus resulting in false negatives for the two preceding seasons.

In the 2006 SH season, WI05 became predominant and replaced CA04 (53). The associated HA allele with change *193F* ranked first in the AD plot for the 2005 SH season, with antigenic weights of 0.79 (up) and 0.44 (down), and thus was correctly predicted for the 2006 SH season by our method. The *193F* HA allele also ranked first, with high antigenic weights 0.0 (up) and 1.23 (down), in the following season, but we did not predict it (again) as a vaccine strain, as we had already selected it the season before. The WHO recommended a vaccine strain update for WI05 one season later, for the 2006-2007 NH season, thus making a false-negative call for the 2006 SH season.

Finally, BR07 became predominant in the 2007 SH season (50). The BR07 HA allele with changes *50E 140I* was first evident in the 2006-2007 NH season, with a small frequency increase in the AD plot (4%), and was not among the top-ranking HA alleles. Therefore, it was not recommended as a vaccine strain for the 2007 SH season or the 2007-2008 NH. These are two false negatives, which both our method and the WHO failed to identify (therefore, BR07 is not included among the three top-ranking alleles in Table 1). However, in the analyzed data, no viral samples are present from after December 2006, as this is the end of the period analyzed in our study and that by Russell et al. (54). This is usually the time where influenza activity peaks (34), and as BR07 appeared very late in the 2006-2007 NH season, it explains why BR07-like strains are underrepresented in our data set. Nevertheless, our method assigned to the BR07 HA allele the highest antigenic weight (0.90 [up] and 0.0 [down]) of all alleles increasing in frequency for the 2006-2007 NH season, well above the antigenic weight threshold (more than 0.5 antigenic unit) used for prediction. Thus, its antigenic impact was correctly revealed.

Antigenic weights for the negative examples were mostly low and resulted in correct identification of true-negative HA alleles for all but one season. In the 2003-2004 NH season, the *140E* HA allele, which never became predominant, ranked third, with a high antigenic weight, resulting in a false-positive prediction (Fig. 2). The respective clade, represented by the viral isolate A/Oklahoma/8/2004, has not been described in the literature. It became extinct after the 2004 SH season. As the antigenic weight of this HA allele in the antigenic trees for subsequent seasons was low (<0.5 antigenic unit), the high weight assigned in the 2003-2004 season might be an overestimate of its antigenic impact. The top-scoring HA allele of this season, with changes *227P *189N**, increased approximately twice as fast in frequency in comparison to the *140E* HA allele but had a low antigenic weight. This indicates that there was no novel antigenically distinct strain on the rise to predominance in the viral population in this season. In comparison, the WHO recommended the WE04 strain as a vaccine candidate in the 2004 SH season, which was immediately replaced by the CA04 strain (48, 52). This resulted in a false-positive assignment in this season.

Complementary strategy. Our basic strategy is to rank HA alleles based on their increase in frequency over two consecutive seasons and then assess their antigenic impact relevance based on their antigenic weights in the antigenic tree. In a complementary approach, we tested to first rank HA alleles by their antigenic weights and then selected those which increased in frequency the most over two consecutive seasons for a vaccine strain update (see Table S2 in the supplemental material). We restricted the analysis to alleles increasing in frequency and to those for which a reference serum was located in the respective subtree, with the aim of selecting a strain with a reference serum available. We do not know for certain whether these sera were available for analysis by the WHO Collaborating Centres during the earliest season in which this strain appeared, as for our analysis we only had the sampling dates of the respective viral isolates available. Overall, this resulted in similar results: for the positive examples, six were correctly predicted and one false-positive prediction was made. For each influenza season, we identified up to three alleles with an antigenic weight (up- or down-weight) of more than 0.5 antigenic unit. However, the frequency increase for most alleles in comparison to the preceding season was less than 5%, indicating no significant increase in prevalence. Exceptions were the HA alleles for which the associated antigenically distinct strains became predominant and which were described above, namely, the *156H *75Q, 155T** HA allele in the 2002-2003 NH season, the *145N *159F, 226I** HA allele in the 2004 SH season, and the *193F* HA allele in the 2005 SH season and in the 2005-2006 NH season. These three HA alleles ranked first in the AD plots for the respec-

tive seasons. Additionally, the *140E* and *144D *159F** HA alleles had antigenic weights of more than 0.5 antigenic unit and increased in frequency more than 5% in the 2004 NH season. The *140E* HA allele was also found using the method described above. The clade of the *144D *159F** HA allele represents the viral isolate A/Hiroshima/39/2004, which is not described in the literature and became extinct after the 2004 SH season. The antigenic weight of this HA allele in the antigenic trees for the following seasons was low, indicating an overestimation of the assigned antigenic weight in the 2004 NH season. The BR07 *50E 140I* HA allele ranked first in the 2006-2007 NH season but had only a low frequency increase (<5%), thus resulting in a false-negative assignment. In total, 29 HA alleles increased in frequency and had antigenic weights of more than 0.0 antigenic unit in the tested influenza seasons. One was a false-positive and one was a false-negative assignment, resulting in an overall assignment accuracy of 93%.

Robustness. To assess the robustness of our method, we repeated our experiments in a 10-fold cross-validation setup, repeated 10 times for every influenza season. For the antigenic trees, the average absolute error of the antigenic distance prediction for each influenza season was 0.83 (standard deviation, 0.07) and the average root mean squared error was 1.07 (standard deviation, 0.08). These results are comparable to the accuracy achieved on a different data set (33) and demonstrate that the inference of the antigenic tree model was stable for different seasons. The cross-validation setup also allowed us to calculate the average antigenic weights and standard deviations for individual branches. In general, the average antigenic weights for the individual alleles were similar to the final antigenic weights with low standard deviations, which indicates the robustness of the fitted weights (see Table S3 in the supplemental material). A notable exception was the weight of the *193F* HA allele of the WI05 clade in the 2005 SH season. For this HA allele, the difference between the final antigenic weights (0.79 [up-weight] and 0.44 [down-weight]) and average antigenic weights (0.42 [up-weight] with a standard deviation of 0.37 and 0.81 [down-weight] with a standard deviation of 0.39) was high. However, the average down-weight was above the prediction threshold (more than 0.5 antigenic unit) and correctly indicated the antigenic impact of the 193F change.

We compared the antigenic weights of the true-positive HA alleles predicted for vaccine strain updates to the weights of these alleles in the antigenic trees inferred for the following seasons. In general, the antigenic weights for these alleles varied in subsequent seasons. Although the antigenic weights for these three HA alleles varied across influenza seasons, presumably due to differences in individual data sets, one of the two weights (the up- or the down-weight) of each HA allele was always above the threshold of 0.5 antigenic unit, supporting their antigenic relevance. The *156H *75Q, 155T** HA allele identified in the 2002-2003 NH season had antigenic weights of 0.82 and 0.28 (up- and down-weights). In the two following seasons, the allele weights were 0.61 and 0.46 and 0.35 and 0.80, respectively. The *145N *159F, 226I** HA allele identified in the 2004 SH season originally had antigenic weights of 0.68 and 0.12 (up- and down-weights). In the two following influenza seasons, the allele weights were 0.48 and 0.88 and 0.35 and 0.89, respectively. Finally, the HA allele *193F* identified in the 2005 SH season had antigenic weights of 0.79 and 0.44 (up- and down-weights). These weights were 0.0 and 1.23 and 0.0 and 1.36 in the two following seasons, respectively.

DISCUSSION

Deciding on the composition of the seasonal influenza vaccine involves data collection and analysis by experts at numerous institutes around the globe (17). Besides serological analysis, computer-guided analysis is becoming increasingly important in the interpretation of the large amounts of generated data. Previously, we developed methods for identifying the HA alleles that are most likely to become predominant in the future and for inferring set of amino acid changes in HA with larger antigenic weights in the evolution of human influenza A (H3N2) viruses (32, 33). In the present report, we describe how these techniques—inference of AD plots and inference of antigenic trees—can be combined to predict the antigenic evolution of the virus accurately. We used our method to predict, 1 year in advance, like the experts of the WHO, whether a viral strain with an antigenically novel HA allele would become predominant, i.e., whether it would be different enough to warrant a change in the strain used for vaccine production. To simulate realistic conditions, we performed all calculations with our method (tree inference, allele dynamics, and antigenic weight inference) for each influenza season based only on the part of data collected up to the month before the WHO decision. So, no data from after this point in time were used for predicting the vaccine strain for the influenza season 1 year later. As the different influenza seasons were sampled with various depths (49 to 194 viral samples), we used only viral isolates collected in the 2 years preceding each individual decision. This reduced the effects of various sample sizes for the different seasons and resulted in similar cross-validation errors for all influenza seasons.

Du et al. (55) used HI assay data to learn parameters for a sequence property-derived assessment of antigenic similarity of viral strains, showing that this allows determination of antigenically similar strains; however, they did not show a validation of their method in a realistic setting for determining suitable vaccine strains as we have done here, where strains available up to a year X are used to make predictions for the year $X + 1$. Instead, they based their predictions for season 2002-2003 to season 2008-2009 on strain abundances in their data set for the same period, which seems unrealistic, as strains that have become predominant are more abundant in this data set than in the time before they were predominant and when the decision by the WHO is required. Without using HI data, in a recent study, Łuksza and Lässig (56) described a fitness function calculating the growth rate of viral strains based on an adaptive evolutionary model and applied it for a year-to-year prediction as we did here. This dynamic model assesses epitope changes coupled with a susceptible-infected-recovered (SIR) model measuring pathogen-host interaction in combination with nonepitope alterations to determine suitable vaccine strains for the next year. Our framework is a data-driven alternative to such an approach which does not rely on an explicit evolutionary model and learns allele dynamics from the data. By combining allele dynamic estimates with inference of their antigenic impact, seasons are determined where antigenically altered strains occur that are on the rise to predominance. To our knowledge, this is the first successful demonstration of a computational approach which combines all relevant information in a realistic setting.

For the nine seasons within the period from 2002 to 2007 (34), we correctly predicted three out of four appearances of antigenically novel predominant strains. Only one false-positive predic-

tion was made. A fourth transition was not identified due to the low number of available samples in the preceding seasons. However, the relevant HA allele was assigned a high antigenic weight, which indicated the importance of the corresponding viral strain. For the positive examples (namely, the appearances of antigenically novel predominant strains requiring vaccine strain updates) as antigenically distinct, the magnitude of their estimated antigenic impacts varied. With antigenic cartography, only the antigenic change of the MO99-FU02 transition was described as sufficiently large to represent a true "jump" between distinct antigenic clusters (31). However, the WHO notes that the other three viral strains were sufficiently distinct in terms of their antigenicity to warrant a vaccine update, which is in line with our predictions (48, 50, 57).

In a recent study, Hensley et al. proposed that changes that alter the receptor-binding avidity drive antigenic drift in seasonal influenza A (H1N1) viruses (58). Similar patterns can be observed in the data analyzed here. Of the four antigenically distinct viral strains, three have changes in or close to the receptor-binding site of HA in their respective HA alleles (positions 155 and 156 for FU02, position 226 for CA04, and position 193 for WI05), which could be indicative of the relevance of receptor avidity also for the evolution of the H3N2 subtype (45, 59).

Although HA is the major viral antigen of the virus, NA also plays an important role in antigenic drift and immune evasion (60). For seasonal influenza A (H1N1) viruses, it was shown that changes in NA can have a significant impact on the antigenic characteristics of the virus, resulting in antigenic drift. Furthermore, low titers in HI assays, which are usually interpreted as effects of HA changes, can be misleading and may be caused by virus attachment via NA (30). Unfortunately, for the data used here, NA sequences were not available. Sandbulte et al. showed that based on HI data, the four antigenic strains (FU02, CA04, WI05, and BR07) which became predominant in the study period are antigenically similar (using HI assay data) but showed distinct antigenic characteristics based on neuraminidase inhibition assays (60). This effect may be seen in the AD plots, where an HA allele with only low antigenic impact rises very fast in frequency due to the associated changes in the NA segment of the viral lineage. In our analysis, the HA alleles linked to the four viral strains show distinct antigenic characteristics, but the strong increase in frequency of these HA alleles may also be accompanied by advantageous changes in the NA segment of the respective viruses, both of which are included in generating a novel vaccine strain. Overall, our method allowed us to accurately predict the antigenic evolution and suitable HA segments for the vaccine strain of human influenza A (H3N2) viruses. Inference of antigenic weights for individual HA alleles allowed us to accurately distinguish between alleles increasing in frequency in the viral population with and without antigenic impact. HA alleles with high antigenic weights but only slight increases in frequency turned out to be different from the prevailing antigenic strains but did not show the potential to rise to predominance in the viral population. This is in line with expectations from population genetics, which posits that most allelic diversity with altered fitness is usually present at low levels in a population and driven to extinction, with only a few alleles rising to predominance, with chances of fixation increasing along with their rise in frequencies (61). It is the combined consideration of allele epidemiological dynamics and estimates of their phenotype impact in terms of antigenicity which allowed us to predict future predominant alleles with altered antigenicity. Our method was more accurate than the WHO's recommendations for seasons in which antigenically novel strains that necessitate a vaccine strain update appear. Thus, our method may allow the production of a more efficacious vaccine for such seasons. Of course, in some cases, practicalities, such as availability of a fast-growing vaccine candidate strain, might have prevented the WHO from recommending a better-suited strain for vaccine production, even though this was not mentioned explicitly, to our knowledge, in the respective reports. Thus, our computational approach may have the potential to further improve the current procedure. Therefore, we propose that our method could be applied to the same data for several years in parallel to the currently used expert-based procedure and that its predictions could be recorded. If the high accuracy we observed in this study is further confirmed, our method could become part of the standard decision process. Some may be skeptical of computational approaches to deciding on vaccine strain composition. However, the complexity and amount of data generated in the Global Influenza Surveillance and Response System necessitate their use, and we should take advantage of the predictive power achievable with appropriate inference techniques.

ACKNOWLEDGMENTS

L.S., T.R.K., and A.C.M. gratefully acknowledge funding from Heinrich Heine University Düsseldorf.

We thank A. Hay for providing very helpful comments.

L.S. designed and performed the research, L.S. and A.C.M. analyzed the results, L.S, T.R.K., and A.C.M. wrote the paper, and A.C.M. planned the project.

REFERENCES

1. **Tognotti E.** 2009. Influenza pandemics: a historical retrospect. J. Infect. Dev. Ctries. **3:**331–334. http://dx.doi.org/10.3855/jidc.239.
2. **World Health Organization.** 2009. Fact sheet no. 211. World Health Organization, Geneva, Switzerland.
3. **Lin YP, Gregory V, Bennett M, Hay A.** 2004. Recent changes among human influenza viruses. Virus Res. **103:**47–52. http://dx.doi.org/10.1016/j.virusres.2004.02.011.
4. **Barr IG.** 2014. WHO recommendations for the viruses used in the 2013-2014 Northern Hemisphere influenza vaccine: epidemiology, antigenic and genetic characteristics of influenza A(H1N1)pdm09, A(H3N2) and B influenza viruses collected from October 2012 to January 2013. Vaccine **32:**4713–4725. http://dx.doi.org/10.1016/j.vaccine.2014.02.014.
5. **Medina RA, García-Sastre A.** 2011. Influenza A viruses: new research developments. Nat. Rev. Microbiol. **9:**590–603. http://dx.doi.org/10.1038/nrmicro2613.
6. **Muramoto Y, Noda T, Kawakami E, Akkina R, Kawaoka Y.** 2013. Identification of novel influenza A virus proteins translated from PA mRNA. J. Virol. **87:**2455–2462. http://dx.doi.org/10.1128/JVI.02656-12.
7. **Wise HM, Hutchinson EC, Jagger BW, Stuart AD, Kang ZH, Robb N, Schwartzman LM, Kash JC, Fodor E, Firth AE, Gog JR, Taubenberger JK, Digard P.** 2012. Identification of a novel splice variant form of the influenza A virus M2 ion channel with an antigenically distinct ectodomain. PLoS Pathog. **8:**e1002998. http://dx.doi.org/10.1371/journal.ppat.1002998.
8. **Fouchier RAM, Munster V, Wallensten A, Bestebroer TM, Herfst S, Smith D, Rimmelzwaan GF, Olsen B, Osterhaus ADME.** 2005. Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. J. Virol. **79:**2814–2822. http://dx.doi.org/10.1128/JVI.79.5.2814-2822.2005.
9. **Tong S, Li Y, Rivailler P, Conrardy C, Castillo DA, Chen L-M, Recuenco S, Ellison JA, Davis CT, York IA, Turmelle AS, Moran D, Rogers S, Shi M, Tao Y, Weil MR, Tang K, Rowe LA, Sammons S, Xu X, Frace M, Lindblade KA, Cox NJ, Anderson LJ, Rupprecht CE, Donis RO.** 2012. A distinct lineage of influenza A virus from bats. Proc. Natl. Acad. Sci. U. S. A. **109:**4269–4274. http://dx.doi.org/10.1073/pnas.1116200109.

10. **Tong S, Zhu X, Li Y, Shi M, Zhang J, Bourgeois M, Yang H, Chen X, Recuenco S, Gomez J, Chen LM, Johnson A, Tao Y, Dreyfus C, Yu W, McBride R, Carney PJ, Gilbert AT, Chang J, Guo Z, Davis CT, Paulson JC, Stevens J, Rupprecht CE, Holmes EC, Wilson IA, Donis RO.** 2013. New world bats harbor diverse influenza A viruses. PLoS Pathog. **9:**e1003657. http://dx.doi.org/10.1371/journal.ppat.1003657.
11. **WHO.** 2012. Recommended composition of influenza virus vaccines for use in the 2012-2013 northern hemisphere influenza season. Wkly. Epidemiol. Rec. **87:**83–96.
12. **Webster RG, Laver WG, Air GM.** 1982. Molecular mechanisms of variation in influenza viruses. Nature **296:**115–121. http://dx.doi.org/10.1038/296115a0.
13. **Bush RM, Fitch WM, Bender CA, Cox NJ.** 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol. Biol. Evol. **16:**1457–1465. http://dx.doi.org/10.1093/oxfordjournals.molbev.a026057.
14. **Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus ADME, Fouchier RAM.** 2004. Mapping the antigenic and genetic evolution of influenza virus. Science **305:**371–376. http://dx.doi.org/10.1126/science.1097211.
15. **WHO Writing Group, Ampofo WK, Baylor N, Cobey S, Cox NJ, Daves S, Edwards S, Ferguson N, Grohmann G, Hay A, Katz J, Kullabutr K, Lambert L, Levandowski R, Mishra AC, Monto A, Siqueira M, Tashiro M, Waddell AL, Wairagkar N, Wood J, Zambon M, Zhang W.** 2012. Improving influenza vaccine virus selection: report of a WHO informal consultation held at WHO headquarters, Geneva, Switzerland, 14–16 June 2010. Influenza Other Respir. Viruses **6:**142–152, e1–e5. http://dx.doi.org/10.1111/j.1750-2659.2011.00277.x.
16. **World Health Organization.** 2012. Recommended composition of influenza virus vaccines for use in the 2013 southern hemisphere influenza season. Wkly. Epidemiol. Rec. **87:**389–400.
17. **Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, Gust ID, Hampson AW, Hay AJ, Hurt AC, de Jong JC, Kelso A, Klimov AI, Kageyama T, Komadina N, Lapedes AS, Lin YP, Mosterin A, Obuchi M, Odagiri T, Osterhaus AD, Rimmelzwaan GF, Shaw MW, Skepner E, Stohr K, Tashiro M, Fouchier RA, Smith DJ.** 2008. Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. Vaccine **26:**D31–D34. http://dx.doi.org/10.1016/j.vaccine.2008.07.078.
18. **Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM.** 1999. Predicting the evolution of human influenza A. Science **286:**1921–1925. http://dx.doi.org/10.1126/science.286.5446.1921.
19. **Plotkin JB, Dushoff J, Levin SA.** 2002. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc. Natl. Acad. Sci. U. S. A. **99:**6263–6268. http://dx.doi.org/10.1073/pnas.082110799.
20. **Shih AC-C, Hsiao T-C, Ho M-S, Li W-H.** 2007. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. Proc. Natl. Acad. Sci. U. S. A. **104:**6283–6288. http://dx.doi.org/10.1073/pnas.0701396104.
21. **Du X, Wang Z, Wu A, Song L, Cao Y, Hang H, Jiang T.** 2008. Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution. Genome Res. **18:**178–187. http://dx.doi.org/10.1101/gr.6969007.
22. **Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, Frost SDW.** 2008. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. Mol. Biol. Evol. **25:**1809–1824. http://dx.doi.org/10.1093/molbev/msn123.
23. **Xia Z, Jin G, Zhu J, Zhou R.** 2009. Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus. Bioinformatics **25:**2309–2317. http://dx.doi.org/10.1093/bioinformatics/btp423.
24. **Tusche C, Steinbrück L, McHardy AC.** 2012. Detecting patches of protein sites of influenza A viruses under positive selection. Mol. Biol. Evol. **29:**2063–2071. http://dx.doi.org/10.1093/molbev/mss095.
25. **Lin YP, Xiong X, Wharton SA, Martin SR, Coombs PJ, Vachieri SG, Christodoulou E, Walker PA, Liu J, Skehel JJ, Gamblin SJ, Hay AJ, Daniels RS, McCauley JW.** 2012. Evolution of the receptor binding properties of the influenza A(H3N2) hemagglutinin. Proc. Natl. Acad. Sci. U. S. A. **109:**21474–21479. http://dx.doi.org/10.1073/pnas.1218841110.
26. **Lee M-S, Chen M-C, Liao Y-C, Hsiung CA.** 2007. Identifying potential immunodominant positions and predicting antigenic variants of influenza A/H3N2 viruses. Vaccine **25:**8133–8139.
27. **Liao Y-C, Lee M-S, Ko C-Y, Hsiung CA.** 2008. Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. Bioinformatics **24:**505–512. http://dx.doi.org/10.1093/bioinformatics/btm638.
28. **Huang J-W, King C-C, Yang J-M.** 2009. Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses. BMC Bioinformatics **10**(Suppl 1)**:**S41. http://dx.doi.org/10.1186/1471-2105-10-S1-S41.
29. **Hirst GK.** 1943. Studies of antigenic differences among strains of influenza A by means of red cell agglutination. J. Exp. Med. **78:**407–423. http://dx.doi.org/10.1084/jem.78.5.407.
30. **Lin YP, Gregory V, Collins P, Kloess J, Wharton S, Cattle N, Lackenby A, Daniels R, Hay A.** 2010. Neuraminidase receptor binding variants of human influenza A(H3N2) viruses resulting from substitution of aspartic acid 151 in the catalytic site: a role in virus attachment? J. Virol. **84:**6769–6781. http://dx.doi.org/10.1128/JVI.00458-10.
31. **Fouchier RAM, Smith DJ.** 2010. Use of antigenic cartography in vaccine seed strain selection. Avian Dis. **54:**220–223. http://dx.doi.org/10.1637/8740-032509-ResNote.1.
32. **Steinbrück L, McHardy AC.** 2011. Allele dynamics plots for the study of evolutionary dynamics in viral populations. Nucleic Acids Res. **39:**e4. http://dx.doi.org/10.1093/nar/gkq909.
33. **Steinbrück L, McHardy AC.** 2012. Inference of genotype-phenotype relationships in the antigenic evolution of human influenza A (H3N2) viruses. PLoS Comp. Biol. **8:**e1002492. http://dx.doi.org/10.1371/journal.pcbi.1002492.
34. **Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, Gust ID, Hampson AW, Hay AJ, Hurt AC, de Jong JC, Kelso A, Klimov AI, Kageyama T, Komadina N, Lapedes AS, Lin YP, Mosterin A, Obuchi M, Odagiri T, Osterhaus ADME, Rimmelzwaan GF, Shaw MW, Skepner E, Stohr K, Tashiro M, Fouchier RAM, Smith DJ.** 2008. The global circulation of seasonal influenza A (H3N2) viruses. Science **320:**340–346. http://dx.doi.org/10.1126/science.1154137.
35. **Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D.** 2008. The influenza virus resource at the National Center for Biotechnology Information. J. Virol. **82:**596–601. http://dx.doi.org/10.1128/JVI.02005-07.
36. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
37. **Guindon S, Gascuel O.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:**696–704. http://dx.doi.org/10.1080/10635150390235520.
38. **Zwickl DJ.** 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. thesis. The University of Texas at Austin, Austin, TX. http://repositories.lib.utexas.edu/handle/2152/2666?show=full.
39. **Fitch WM.** 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20:**406–416. http://dx.doi.org/10.2307/2412116.
40. **Yang Z, Kumar S, Nei M.** 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141:**1641–1650.
41. **Pagel M, Meade A, Barker D.** 2004. Bayesian estimation of ancestral character states on phylogenies. Syst. Biol. **53:**673–684. http://dx.doi.org/10.1080/10635150490522232.
42. **Wiley DC, Wilson IA, Skehel JJ.** 1981. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature **289:**373–378. http://dx.doi.org/10.1038/289373a0.
43. **Wiley DC, Skehel JJ.** 1987. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. Annu. Rev. Biochem. **56:**365–394. http://dx.doi.org/10.1146/annurev.bi.56.070187.002053.
44. **Wilson IA, Cox NJ.** 1990. Structural basis of immune recognition of influenza virus hemagglutinin. Annu. Rev. Immunol. **8:**737–771. http://dx.doi.org/10.1146/annurev.iy.08.040190.003513.
45. **Skehel J, Wiley DC.** 2000. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu. Rev. Biochem. **69:**531–569. http://dx.doi.org/10.1146/annurev.biochem.69.1.531.
46. **Wilson IA, Skehel JJ, Wiley DC.** 1981. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 angst resolution. Nature **289:**366–373. http://dx.doi.org/10.1038/289366a0.
47. **WHO.** 2003. Recommended composition of influenza virus vaccines for use in the 2004 influenza season. Wkly. Epidemiol. Rec. **78:**375–379.
48. **WHO.** 2005. Recommended composition of influenza virus vaccines for use in the 2005-2006 influenza season. Wkly. Epidemiol. Rec. **80:**66–71.

49. **WHO.** 2006. Recommended composition of influenza virus vaccines for use in the 2007 influenza season. Wkly. Epidemiol. Rec. **81:**390–395.
50. **WHO.** 2007. Recommended composition of influenza virus vaccines for use in the 2008 influenza season. Wkly. Epidemiol. Rec. **82:**351–356.
51. **WHO.** 2004. Recommended composition of influenza virus vaccines for use in the 2004-2005 influenza season. Wkly. Epidemiol. Rec. **79:**88–92.
52. **WHO.** 2004. Recommended composition of influenza virus vaccines for use in the 2005 influenza season. Wkly. Epidemiol. Rec. **79:**369–373.
53. **WHO.** 2006. Recommended composition of influenza virus vaccines for use in the 2006-2007 influenza season. Wkly. Epidemiol. Rec. **81:**82–86.
54. **WHO.** 2007. Recommended composition of influenza virus vaccines for use in the 2007-2008 influenza season. Wkly. Epidemiol. Rec. **82:**69–74.
55. **Du X, Dong L, Lan Y, Peng Y, Wu A, Zhang Y, Huang W, Wang D, Wang M, Guo Y, Shu Y, Jiang T.** 2012. Mapping of H3N2 influenza antigenic evolution in China reveals a strategy for vaccine strain recommendation. Nat. Commun. **3:**709. http://dx.doi.org/10.1038/ncomms1710.
56. **Łuksza M, Lässig M.** 2014. A predictive fitness model for influenza. Nature **507:**57–61. http://dx.doi.org/10.1038/nature13087.
57. **WHO.** 2003. Recommended composition of influenza virus vaccines for use in the 2003-2004 influenza season. Wkly. Epidemiol. Rec. **78:**58–62.
58. **Hensley SE, Das SR, Bailey AL, Schmidt LM, Hickman HD, Jayaraman A, Viswanathan K, Raman R, Sasisekharan R, Bennink JR, Yewdell JW.** 2009. Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. Science **326:**734–736. http://dx.doi.org/10.1126/science.1178258.
59. **Skehel J.** 2009. An overview of influenza haemagglutinin and neuraminidase. Biologicals **37:**177–178. http://dx.doi.org/10.1016/j.biologicals.2009.02.012.
60. **Sandbulte MR, Westgeest KB, Gao J, Xu X, Klimov AI, Russell CA, Burke DF, Smith DJ, Fouchier RA, Eichelberger MC.** 2011. Discordant antigenic drift of neuraminidase and hemagglutinin in H1N1 and H3N2 influenza viruses. Proc. Natl. Acad. Sci. U. S. A. **108:**20748–20753. http://dx.doi.org/10.1073/pnas.1113801108.
61. **Olson-Manning CF, Wagner MR, Mitchell-Olds T.** 2012. Adaptive evolution: evaluating empirical support for theoretical predictions. Nat. Rev. Genet. **13:**867–877. http://dx.doi.org/10.1038/nrg3322.
62. **WHO.** 2005. Recommended composition of influenza virus vaccines for use in the 2006 influenza season. Wkly. Epidemiol. Rec. **80:**342–347.
63. **WHO.** 2008. Recommended composition of influenza virus vaccines for use in the 2008-2009 influenza season. Wkly. Epidemiol. Rec. **83:**81–87.