Published online 18 October 2010 *Nucleic Acids Research, 2011, Vol. 39, No. 1* **e4**
doi:10.1093/nar/gkq909

Allele dynamics plots for the study of evolutionary dynamics in viral populations

Lars Steinbrück[1] **and Alice Carolyn McHardy**[1,2,*]

[1]Max-Planck Research Group for Computational Genomics and Epidemiology, Max-Planck Institute for Informatics, University Campus E1 4, 66123 Saarbrücken and [2]Department for Algorithmic Bioinformatics, Heinrich-Heine-University, Universitätsstr. 1, 40225 Düsseldorf, Germany

Received June 16, 2010; Revised August 30, 2010; Accepted September 24, 2010

ABSTRACT

Phylodynamic techniques combine epidemiological and genetic information to analyze the evolutionary and spatiotemporal dynamics of rapidly evolving pathogens, such as influenza A or human immunodeficiency viruses. We introduce 'allele dynamics plots' (AD plots) as a method for visualizing the evolutionary dynamics of a gene in a population. Using AD plots, we propose how to identify the alleles that are likely to be subject to directional selection. We analyze the method's merits with a detailed study of the evolutionary dynamics of seasonal influenza A viruses. AD plots for the major surface protein of seasonal influenza A (H3N2) and the 2009 swine-origin influenza A (H1N1) viruses show the succession of substitutions that became fixed in the evolution of the two viral populations. They also allow the early identification of those viral strains that later rise to predominance, which is important for the problem of vaccine strain selection. In summary, we describe a technique that reveals the evolutionary dynamics of a rapidly evolving population and allows us to identify alleles and associated genetic changes that might be under directional selection. The method can be applied for the study of influenza A viruses and other rapidly evolving species or viruses.

INTRODUCTION

Phylogenetic analysis allows the inference of evolutionary relationships from a set of genetic sequences, which may represent a distinct species or a genetic region of individuals of a population. For populations of rapidly evolving organisms, the evolutionary and epidemiological processes may occur on similar timescales. Newly developed analytical methods, known as phylodynamic techniques, allow the joint analysis of the genetic and epidemiological relationships of the underlying data (1,2). Based on epidemiological information, such as sampling locations or sampling times, phylodynamic methods enable the geographic migration patterns of individuals of a population to be studied, tracking viral spread across host tissues, searching for genetic sites subject to purifying or positive selection associated with adaptation, dating past evolutionary events and gaining insights into population-level processes using coalescence analysis. In (3), for example, the migration paths of the highly pathogenic avian influenza A (H5N1) virus across Asia are inferred with a 'phylogeographic' approach from genetic sequences and geographic sampling locations. Other studies revealed that chimpanzees serve as a natural reservoir for pandemic and nonpandemic HIV type 1 (4), based on 'phylogeographic' clustering, and identified the epidemic history and geographic source of HIV type 2 based on a molecular clock analysis of dated genetic sequences (5).

We describe a method for analyzing the population-level phylodynamics of a gene, which we call allele dynamics plots (AD plots). AD plots combine information from phylogenetic inference and ancestral character state reconstruction with isolate sampling times for the analysis of population-level evolutionary dynamics. Furthermore, we use the AD plot of a population-level sequence sample to identify the alleles that might be associated with a selective advantage. Based on this, we demonstrate how AD plots can be used to study evolutionary dynamics and to identify emerging viral strains with the example of two influenza A viruses: the human influenza A (H3N2) and the 2009 swine-origin influenza A (H1N1) viruses.

In research into the evolution of the influenza virus, a method that enables the identification of alleles under selection is to count the number of amino acid changes within a protein at sites under selection, which, in turn,

[*]To whom correspondence should be addressed. Tel: +49 211 81 10 591; Fax: +49 211 81 13 464; Email: mchardy@mpi-inf.mpg.de

can be identified based on the ratio of non-synonymous-to-synonymous mutations (dN/dS) (6). A recent study suggests, however, that dN/dS ratios may not always be informative with regards to detecting selection within a population. Moreover, the method is lacking in sensitivity when applied to individual sequence sites (7). A different approach was proposed by Pond *et al.* who introduced a phylogenetic maximum likelihood test based on a protein evolution model to test for directional evolution at individual sites of an alignment (8,9). Further related methods quantify the impact of 'key innovations' in species trees, e.g. what would happen if lineages that have acquired a beneficial feature were able to spread faster than others. These methods incorporate clade sizes and shifts in diversification rates identified from the phylogenetic tree based on likelihood estimators in the analysis. For an overview, see (10). However, these methods were conceived for species-level and not population-level analysis, and to evaluate macro-evolution. The method we describe here does not use dN/dS information and is designed for the analysis of longitudinally sampled population-level sequence data. In this sense, it complements the existing approaches.

Background on influenza A viruses

The influenza virus is a rapidly evolving pathogen that is suited for the application of phylodynamic techniques. The single-stranded negative-sense RNA viruses of the family *Orthomyxoviridae* are a major health risk in modern life, responsible for up to 500 000 deaths annually (11). Three distinct genera (types A, B and C) are endemic in the human population. Types B and C evolve slowly and circulate at low levels. However, through rapid evolution of the antibody-binding (epitope) sites of the surface proteins, influenza A continuously evades host immunity from previous infection or vaccination, and regularly causes large epidemics. Influenza A viruses can furthermore be distinguished based on the surface proteins hemagglutinin (HA) and neuraminidase (NA). For type A viruses, 16 known subtypes of HA and nine of NA occur in various combinations in aquatic birds (12). In the human population, influenza A viruses of the subtypes H3N2 and H1N1 currently circulate. Of these, the swine-origin influenza A (H1N1) virus ('swine flu'), which entered the human population in 2009, is currently responsible for the majority of infections (13,14).

Human influenza A viruses continuously change antigenically in a process known as antigenic drift. This refers to the successive fixation of mutations that affect viral fitness by increasing a virus' ability to circumvent host immunity and protective antibodies elicited by previously circulating viral variants (6,15). Antigenically relevant changes are located mainly in the epitope sites of the viral HA (16–19). Influenza viruses also have a segmented genome composed of eight distinct segments and can evolve by means of reassortment. In segment reassortment, new viral strains are generated, which can inherit genomic segments from two distinct viruses simultaneously infecting the same host cell. This mechanism can affect antigenic evolution, as segments encoding antigenically novel surface proteins, but which are harbored by viruses with low overall fitness due to other reasons, and can thus be transferred into a more favorable genetic context and subsequently rise to predominance (20–25).

Antigenically novel strains of influenza A appear and become predominant in worldwide epidemics on a regular basis, which requires frequent adaptation of the influenza vaccine composition. The World Health Organization (WHO) monitors the genetic and antigenic characteristics of the circulating influenza A virus population and searches for antigenically novel emerging strains in a global surveillance program (26,27). The gathered surveillance information, combined with human serological data, is evaluated by a panel of experts. The panel meets twice a year to decide if an update of the vaccine composition for the next winter season for both the Northern and Southern hemispheres is necessary. This approach results in a well-matched vaccine in most years, and significantly reduces the morbidity and mortality of seasonal influenza epidemics. However, a decreased vaccine efficacy can be caused by a new antigenic variant if it is identified too late to reformulate the vaccine composition.

A large body of work exists on computational studies of influenza A virus evolution. Phylogenetic reconstruction plays a key role here, since it was successfully used to unravel the global migration of human influenza A (H3N2) viruses (28) and to identify East and Southeast Asia as a global evolutionary reservoir of seasonal influenza A (H3N2) viruses (29). Furthermore, genome-wide phylogenetic analysis of all eight viral segments determined that the evolutionary dynamics of influenza A (H3N2) virus are shaped by a complex interplay between genetic and epidemiological factors, such as mutation, reassortment, natural selection and gene flow (30).

Besides these analytical studies, further computational methods have been applied to study and predict the evolution of human influenza A (H3N2) viruses. Changes within the hemagglutinin HA1 subunit sequence composition over time were visualized and analyzed by Shih *et al.* using amino acid frequency diagrams (31). However, this procedure does not take the underlying evolutionary relationships and structure of the data into account, as isolate sequences and individual sites are treated independently. Plotkin *et al.* used agglomerative single-linkage clustering on hemagglutinin HA1 genetic sequences for decomposing the data into disjoint clusters, finding that influenza evolution is characterized by a succession of predominant clusters or 'swarms' of similar strains (32). This pattern is also reflected by a narrow phylogenetic tree topology with one surviving viral lineage over time and a viral diversity that is periodically diminished by selective sweeps of a novel viral strain throughout the population (11,30). Analyzing the cluster size–time relation, Plotkin *et al.* suggested using a representative of the largest cluster as the vaccine strain for the following winter season (32). Du *et al.* constructed a co-occurrence network from co-occurring nucleotides across the whole genome (33).

They identified co-occurring inter- and intra-segment changes, and used these co-occurrence modules for sequence clustering. This results in a grouping similar to the structure inferred by phylogenetic reconstruction. Xia *et al*. used mutual information to identify and visualize co-occurring mutations in a 'site transition network' (34). They also used this network to predict future mutations, resulting in 70% sensitivity but also in a rather high false positive rate. However, it should be noted that, although the term 'predicting mutations' may convey that mutations are introduced independently in viral isolates in the following season, the effect that a particular genetic change increases in frequency over two consecutive seasons is often due to a previously low-abundance mutant circulating at higher prevalence.

Most of the abovementioned studies assess the underlying evolutionary relationships and structure for the population-level sequence sample in some way. However, the standard way to estimate evolutionary relationships is by phylogenetic inference. As described above, Bush *et al*. identified 18 sites under positive selection by analyzing the ratio of dN/dS on the trunk of a phylogenetic tree of hemagglutinin HA1 subunit sequences (6). They subsequently used these sites to predict the direction of evolution for a phylogenetic tree of influenza A (H3N2) virus HA by identifying the strains within the phylogenetic tree that had the most pronounced evidence for positive selection (35). However, the dN/dS ratio lacks sensitivity if applied to individual sites, as substantial evidence is required for a site to be considered informative. Not all relevant sites may thus be detectable and, furthermore, the most relevant sites may change over time (15). In a more recent study, Pond *et al*. identified nine sites as being under directional selection in the HA segment of the influenza A (H3N2) virus, using a model-based phylogenetic maximum likelihood test. Seven of these sites are not detected with the traditional dN/dS ratio test (9). Nevertheless, this method depends on the baseline amino-acid-substitution matrix and failed to identify adaptive sites when applied to dim-light and color-vision genes in vertebrates (36).

To analyze the antigenic evolution of influenza A viruses, Smith *et al*. introduced a novel method known as antigenic cartography, which is based on multidimensional scaling of assay data on hemagglutination inhibition (15,37). This technique revealed that antigenic evolution is more clustered than genetic evolution, depending on the antigenic impact of individual amino acid exchanges, and that major changes (cluster jumps) occur every 3–4 years on average (15). Accordingly, including both antigenic and genetic data within evolutionary models enables the most accurate analysis of influenza A virus evolution. Some studies try to incorporate antigenic data (38–40); however, because of limited publicly available data, the results have to be approached with caution. To account for this lack of antigenic information for the respective isolate sequences in our evaluation, we identified all predominant antigenic variants over the analyzed time period based on the genetic changes reported in the literature.

MATERIALS AND METHODS

Phylogenetic inference

HA sequences from 4913 seasonal human influenza A (H3N2) virus isolates sampled from 1988 to 2008, and from 1516 swine-origin influenza A (H1N1) virus isolates with exact sampling times (year and month) were downloaded from the influenza virus resource (41) (Supplementary Tables S1 and S2). Alignments of DNA and protein sequences were created with Muscle (42) and manually curated. Phylogenetic trees were inferred with PhyML v3.0 (43) 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 (44).

Allele dynamics plots

We describe AD plots for visualizing the evolutionary dynamics of a gene in a population and for identifying the alleles that are potentially under directional selection. In a nutshell, AD plots visualize gene alleles and their frequencies over time and thus enable a detailed analysis of a gene in a population. The basic idea involves the following four steps: (i) Inference of the evolutionary relationships for a sequence sample of a population. (ii) Ancestral character state reconstruction and inference of evolutionary intermediates based on the reconstructed evolutionary relationships. (iii) Mapping genetic changes to branches of the tree topology and defining the prevalence of distinct alleles of a gene at different points in time. (iv) Finally, evaluating how fast new alleles or genetic variants propagate throughout the population.

Population genetics theory posits that, in a population of constant size, genetic drift will result in variation in allele frequencies and the continuous fixation of variants even in the absence of selection (45–47). However, given that selection acts on an allele and confers a fitness advantage to the individual organism, this will allow such alleles to rise faster in frequency than alleles without a selective advantage. Hence, alleles that increase in frequency most rapidly over time are more likely to be subject to directional selection than other alleles. This criterion can be applied to identify those alleles that might be associated with a selective advantage from AD plots.

Following the phylogenetic inference of a tree topology using any standard method [maximum likelihood, Neighbor-Joining or a consensus tree constructed from a posterior sample of trees inferred with a Bayesian method (48,49)], substitution events in the evolutionary history are reconstructed using ancestral character state reconstruction and assigned to individual tree branches. In detail, substitution events are assigned to the tree branches based on the evolutionary intermediates reconstructed as ancestral characters. We use the parsimony method of Fitch *et al*. (50) for ancestral character state reconstruction; however, in principle, any available method can be applied (51,52). In our analysis, we chose the isolate with the earliest sampling date as an outgroup and used accelerated transformation (AccTran) (51) to resolve ambiguities in character state reconstruction. This procedure results in changes being mapped preferentially closer to the root of the phylogenetic tree.

We define each branch that is associated with a non-empty set of substitutions to represent an individual allele. The number of alleles thus equals the number of branches with non-empty sets of substitutions in the phylogenetic tree. We define the frequency of an allele within a specific period as the ratio of the number of isolates in the subtree of the allele relative to the number of all isolates within the designated period. An allele that occurs later on the path from the root to the most recent isolates includes the substitutions of the alleles that occurred earlier on this path and thus is more specific. Allelic frequencies are subsequently adjusted in case multiple related alleles emerge within the same period. Isolates located in the subtrees of a newly defined allele within a period are counted only once for the most closely placed parental allele in the phylogenetic tree. This means that, for calculating the allele frequency of all less specific alleles, isolates that occur in the subtree below the more specific allele are not considered. Alleles and the relevant substitutions are discussed using the following nomenclature: *allele substitutions* **substitutions of parental alleles from the same period** (Figure 1).

Construction of AD plots for human influenza A viruses

In analyzing the evolution of human influenza A viruses, we are particularly interested in those changes that affect the antigenic properties of a virus. To identify viral variants with increased fitness for propagation through the host population, non-synonymous genetic changes of HA are of particular interest. To this end, we constructed AD plots from the substitutions for the complete viral HA of the influenza A (H1N1) virus. Secondly, we constructed AD plots for the seasonal influenza A (H3N2) virus based on the changes in the five epitope regions of HA (16,17).

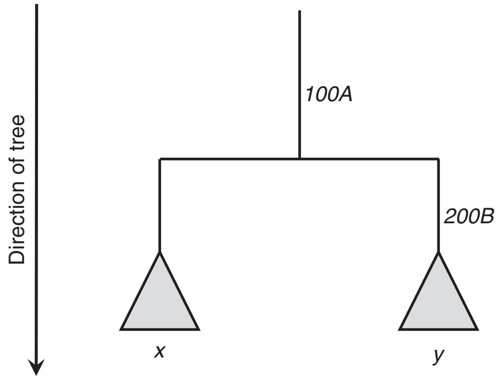

Figure 1. A tree demonstrating the concepts of alleles and allele frequency correction. For allele *100A*, only the isolates of subtree *x* are counted, whereas for allele *200B *100A**, the isolates in subtree *y* are considered.

Influenza infections in the human population show a pattern of seasonality. Peaks of activity occur mainly in the winter months in temperate regions of each hemisphere (53). We use the standard definitions for the influenza season for the Northern and Southern hemispheres in our analysis. For the Northern hemisphere, the influenza season begins on 1 October and ends on 31 March in the following year. For the Southern hemisphere, the influenza season begins on 1 April and ends on 30 September in the same year. For a comparison with the WHO vaccine strain recommendation, we restricted our analysis to sequences sampled up to the end of January for the Northern hemisphere season and to the end of August for the Southern hemisphere season, which is when the WHO decides on the vaccine composition.

To identify the alleles corresponding to the viral strains with antigenically novel HA variants, we used the literature to determine the genetic changes reported for every predominant antigenic variant over the analysis period. These appear, on average, every 3.3 years and then predominate worldwide in seasonal epidemics (15). The changes in these strains for the five HA epitopes are given in Table 1.

RESULTS

Evolutionary dynamics of influenza A (H3N2)

We analyze the evolutionary dynamics of the seasonal influenza A (H3N2) virus with AD plots generated using a maximum likelihood tree (Figure 2) from available HA sequences. The H3N2 subtype has been circulating since 1968, but here we focus on the time from 1998 until the end of 2008. For this more recent period, there is considerably more sequence data available and the bias of sequences toward isolates with unusual virulence or other atypical properties is reduced (54) (Supplementary Figure S3).

The AD plot for HA of the human H3N2 virus (Figure 3, Supplementary Figure S1) shows several alleles that rise to predominance and reach fixation (their frequency in subsequent periods equals one) between 1998 and 2008, such as *57Q *137S**, *156H *75Q, 155T** and *193F*. Other alleles reach high frequencies and subsequently vanish, such as *160R* in the 1999 Southern season, *273S* in the 2000/01 Northern season or *126D* in the 2003 Southern season. Furthermore, a lot of minor-frequency allelic variation is evident within each period.

Alleles becoming predominant and rising to fixation in the surviving lineage correspond to substitutions that map to the trunk of the phylogenetic tree of HA from the human influenza A (H3N2) virus. Besides such changes, the observable variation of alleles that do not become fixed (gray-colored alleles) is rather high within each time interval in the analyzed sample. Although some alleles transiently reach high frequencies, they are only present over a short period. Notably, many of these alleles appear during times when an antigenic variant has been predominant for several years, such as the time from 2000 to 2003, when the A/Panama/2007/1999 (PA99)

Table 1. Antigenically novel viral variants of influenza A (H3N2) that emerged and rose to predominance in worldwide epidemics between 1998 and 2008, and the corresponding substitutions reported in the literature in the five epitope sites of HA

Antigenic cluster	Substitutions	Reference
A/Sydney/5/1997 (SY95)	62E, 156Q, 158K, 196A, 276K	(59)
A/Moscow/10/1999 (MO99)	57Q, 137S	(59)
A/Panama/2007/1999 (PA99)	144N, 172E, 192I	(59)
A/Fujian/411/2002 (FU02)	50G, 75Q, 83K, 131T, 155T, 156H, 186G	(60)
A/California/07/2004 (CA04)	145N, 159F, 189N, 226I, 227P	(61)
A/Wisconsin/67/2005 (WI05)	193F	(62)
A/Brisbane/10/2007 (BR07)	50E, 140I	(63)

Note that PA99 is antigenically similar to MO99 and was used as the vaccine candidate strain for MO99 (56).

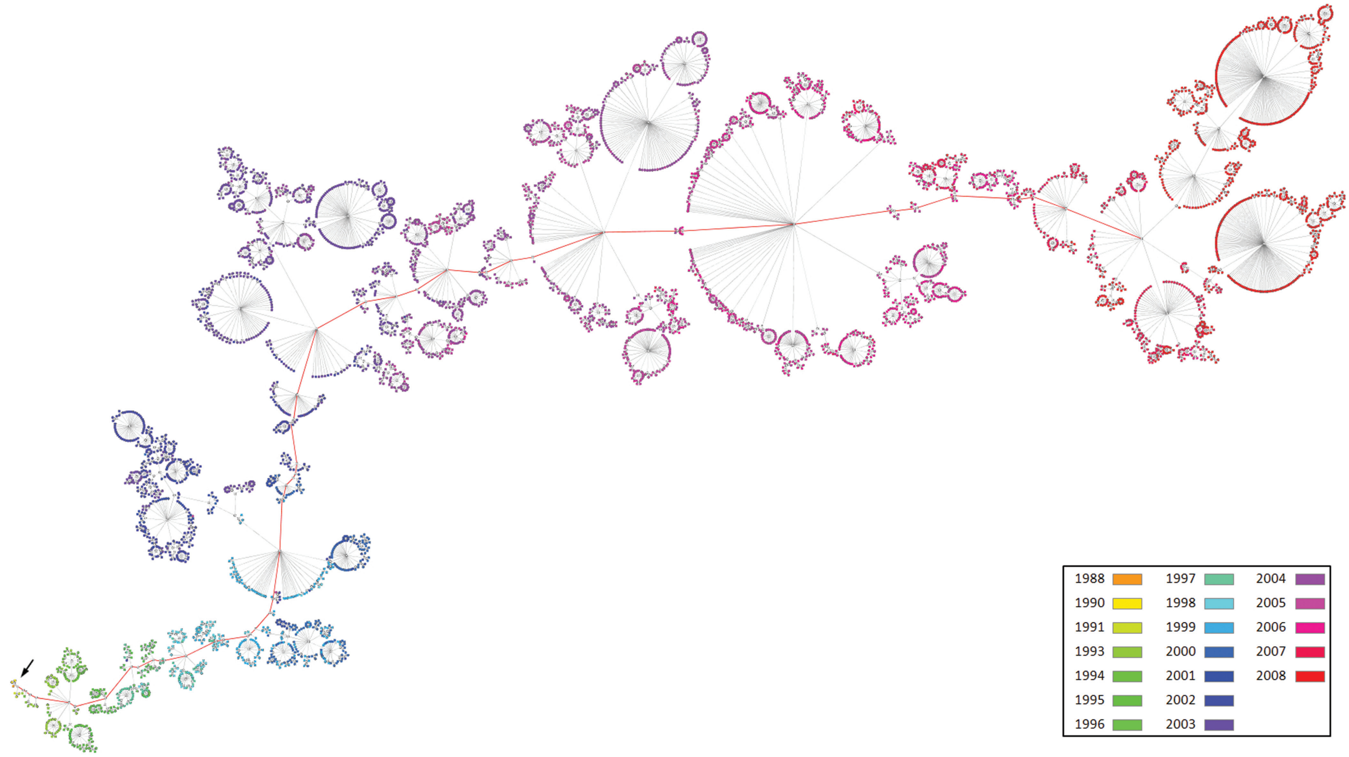

Figure 2. Maximum likelihood tree topology inferred for 4913 hemagglutinin sequences of seasonal human influenza A (H3N2). Leaf nodes are color-coded according to the sampling dates of the viral isolates. The first sampled isolate, A/Siena/3/1988, is indicated with an arrow. The trunk of the tree (i.e. the path from the root to the most recent clade) is colored in red.

variant was predominant. In these years, several new alleles with similar antigenic properties, such as *160R* in the 1999 Southern season, *92T* in the 1999/2000 Northern season, *273S* and *50G, 247C* in the 2000/01 Northern season, and *144D* **186G** in the 2001/02 Northern season, (55–58) appeared successively and rose to high frequencies without reaching fixation.

Most of the alleles rising to fixation (colored in Figure 3) are associated with substitutions reported in the literature (59–63) for the five distinct strains that represent predominant antigenic variants in the analysis period (Table 1). Note that the substitutions of a particular antigenic variant are not necessarily all part of the same allele (i.e. they do not map to the same branch on the trunk of the phylogenetic tree). Instead, they often follow each other in immediate succession in the AD plot and are located on consecutive trunk branches of the phylogenetic tree. The earliest antigenic variant of the analysis period (PA99) is an exception, in this sense, as a single allele represents multiple substitutions. This reveals the limitations of the dataset for the earlier years (Supplementary Figure S3), which does not allow the order in which the PA99 substitutions were acquired by H3N2 to be resolved. For all subsequent antigenic variants, the order of the acquired substitutions is resolved and a set of multiple alleles becoming fixed within an interval are evident from the AD plot. Thus, the evolutionary path and the order in which these changes were acquired in the evolution of antigenically new strains of H3N2 are revealed in the AD plot. For instance, for the antigenic variant BR07, which was predominant from 2006 to 2009, the HA plot shows that, of the two relevant substitutions, 140I was acquired first, followed by 50E.

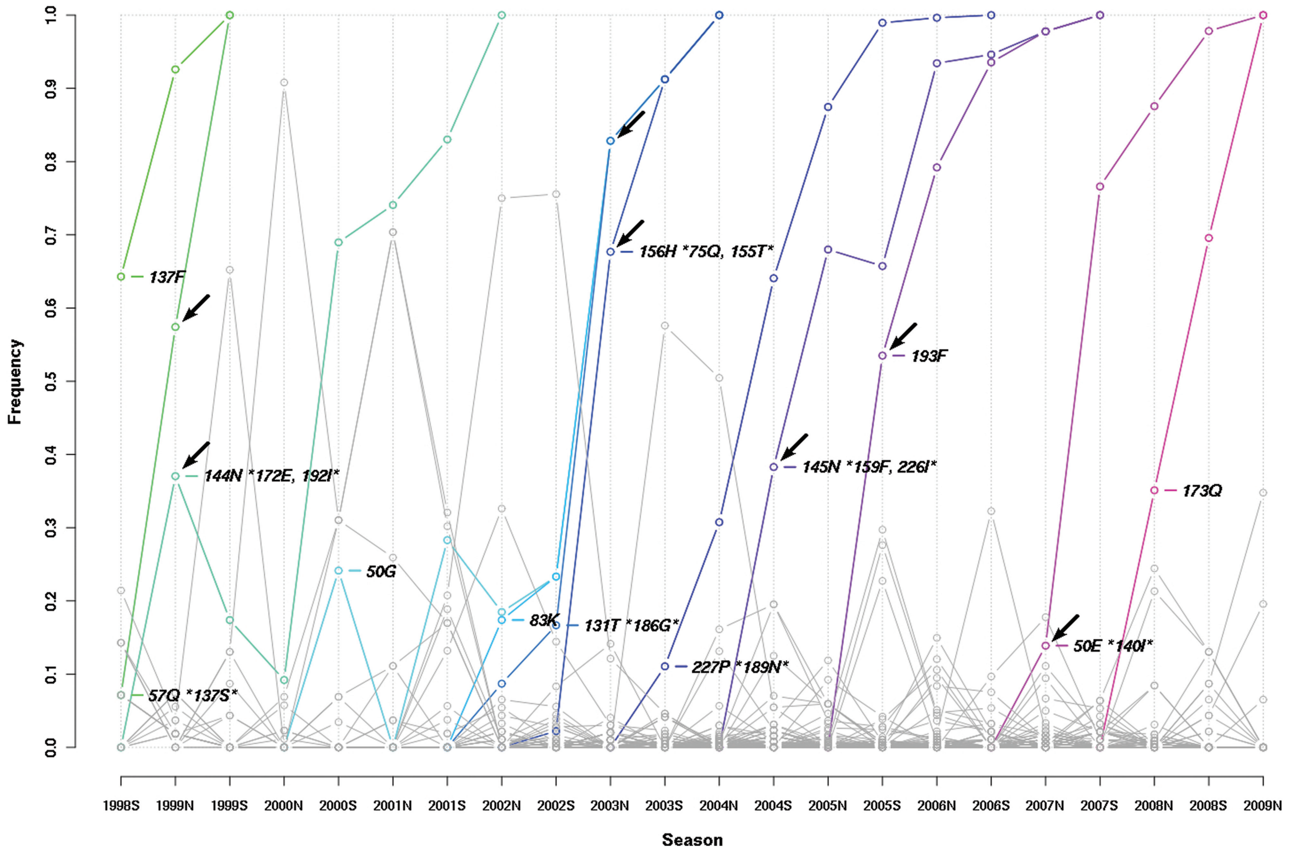

Figure 3. Allele dynamics plot for the major surface protein and antigenic determinant of the seasonal influenza A (H3N2) virus. The Northern and Southern influenza seasons from 1998 to 2008 are shown. Alleles that reach a prevalence of more than 95% and are subsequently fixed are shown in color; all other alleles are shown in gray. Substitutions are restricted to those that occur in the five epitope regions and are enumerated according to HA1 numbering (86). Alleles that rise most quickly in frequency and are of interest with respect to vaccine strain selection are indicated by arrows.

Identification of alleles under directional selection in influenza A (H3N2)

The AD plot, which visualizes the changes in frequencies of individual alleles in a sequence sample, enables us to easily identify those alleles that increase in prevalence most rapidly over two consecutive influenza seasons. The corresponding viral strains are likely candidates to be under the influence of directional selection and to have an advantage relative to other alleles. We identified the alleles with the largest increase in frequency between consecutive seasons that do not represent >50% of the sequences in the first season (otherwise they would already be predominant; Table 1). Of the strains of the five antigenically distinct predominant variants (MO99/PA99, FU02, CA04, WI05 and BR07), four can be correctly identified by this criterion (Table 2). Thus, this measure allows us to use the AD plots to easily identify the strains that are most relevant when deciding the composition of the influenza A (H3N2) vaccine.

In the 1998/99 Northern season, the allele that scores best is *57Q *137S**, which represents the MO99 variant that was predominant from the 1999 Southern season to the 2002–03 Northern season (55–58,64–67). The allele *144N *172E, 192I**, which represents the antigenically very similar strain PA99, ranks second best. In agreement with the AD plot observations, the WHO also recommended MO99 as the vaccine strain for the 2000 Southern season (55). As no suitable well-growing candidate strain could be produced, the previously predominant SY97 strain was used in this season for the vaccine. PA99 was subsequently included as a vaccine component starting from the 1999–2000 Northern season (56). Thus, for the SY97-PA99 antigenic cluster transition, the AD plot allows the timely identification of a suitable strain that is in agreement with the original recommendation of the WHO.

The FU02 variant, which predominated from 2003 to 2004/05 (68–71), is associated with seven distinct substitutions: 50G, 75Q, 83K, 131T, 155T, 156H and 186G. The 155T and 156H define the FU02 antigenic phenotype (72). In the AD plot, the seven FU02 substitutions are associated with seven distinct alleles, each with a single substitution. In the 2002–03 Northern season, alleles with the substitutions *131T *186G** and *156H *75Q, 155T** score first and second best, respectively. The best scoring allele for the 2002/03 Northern season lacks the relevant substitutions 155T and 156H described for FU02. Here, the frequency indicator does not directly reveal the best candidate strain based on the available data. Antigenic information would probably allow a more detailed analysis. The second high-scoring allele would presumably be a good choice as a vaccine strain, as it

Table 2. Alleles and their associated antigenic phenotypes with the steepest slopes in the seasons when they are predicted to become predominant

Season	Alleles	Slope	Antigenic variant	WHO	Predominant
1998/99 North	*57Q *137S**	0.5027	MO99	SY97 (80)	MO99/PA99 (56)
	*144N *172E, 192I**	0.3704	PA99		
2002 South	*155T *75Q**	0.0833	FU02	MO99 (58)	FU02 (68)
	*131T *186G**	0.0797	FU02		
	83K	0.0594	HK02/FU02		
	50G	0.0485	HK02/FU02		
2002/03 North	*131T *186G**	0.6616	FU02	FU02 (67)	FU02 (69)
	*156H *75Q, 155T**	0.6546	FU02		
	83K	0.5950	HK02/FU02		
	50G	0.5950	HK02/FU02		
2004 South	*145N *159F, 226I**	0.3828	WE04/CA04	WE04 (69)	CA04 (73)
	*227P *189N**	0.3331	WE04/CA04		
2005 South	*193F*	0.5350	WI05	CA04 (73)	WI05 (74)
2006/07 North	*50E *140I**	0.1389	BR07	WI05 (75)	BR07 (78)

Alleles in one season are ordered by decreasing slope. Further comparisons show the recommended reference strain for the use in the next year's vaccine by the WHO and the predominant antigenic variant in the next year's influenza season for the same hemisphere. Note that A/Hong Kong/1143/2002 (HK02, [50G, 83K, 186G]) is a PA99-like sublineage present before FU02 and A/Wellington/1/2004 (WE04, [159F, 189N, 227P]) was directly replaced by CA04 in 2004/05 Northern season before becoming predominant.

has other antigenically relevant changes and shows a rapid increase in prevalence during the season. In agreement with this conjecture, the corresponding strain (A/Fujian/411/2002) was recommended by the WHO as the vaccine strain for the 2003–04 Northern season (67). However, as no suitable well-growing candidate strain could be produced, the MO99/PA99 strain was used for the vaccine. In the 2002 Southern season, the *155T *75Q** allele ranks first, but the correct allele (*156H *75Q, 155T**), which features all necessary substitutions, increases only a little in frequency and is thus not selected.

Interestingly, an additional substitution (186G) found in the highest scoring allele for the 2002–03 Northern season appears independently in another frequent allele in the preceding season. This seems a general aspect of H3N2 evolution—the repeated appearance of the same substitution in multiple different alleles. Often, the respective alleles have different phylogenetic histories, in that they occur in different parts of the tree, and the substitutions are occasionally encoded by different codons. Such repeated changes can either reflect neutral changes at highly variable sequence positions or they can be the result of directional selection against a certain residue at a given position at this time. The AD plot allows us to identify such changes easily for further analysis.

The CA04 variant was predominant from 2004–05 to 2005–06 (73,74) and was recommended as vaccine strain for the 2005–06 Northern season in the spring of 2005 (71). The HA allele of this strain scores highest in the 2004 Southern season. Here, the two alleles featuring the substitutions *145N *159F, 226I** and *227P *189N**, respectively, rank first and second. Both of these alleles contain substitutions of the CA04 variant, but only the top-ranking one possesses all relevant substitutions and thus is the correct choice.

The WI05 variant predominated from 2006 to 2006–07 (74,75) and was recommended one season too late as the vaccine strain for the 2006–07 Northern season (76). In the 2005 Southern season, the *193F* allele associated with the WI05 variant scores highest. The second substitution associated with WI05, 225N, is not evident from this plot, as it is not part of the epitope regions. If non-epitope sites are included in the analysis, both substitutions appear on subsequent branches, corresponding to two consecutive emerging alleles in the plot (data not shown). In this plot, the allele *225N *193F** scores highest. The AD plot thus allows us to identify the WI05 variant from the available data one season before the WHO's official recommendation.

Finally, the antigenic variant BR07, which predominated from 2007 onwards (13,77–79), scores highest in the 2006–07 Northern season and is represented by an allele with the substitutions *50E *140I**. A matching strain was recommended for the vaccine of the 2008 Southern season (77). The AD plot allows us to identify this emerging variant for the 2007–08 Northern season.

Applying a maximum likelihood test for directional evolution of protein sequences (DEPS) (9) to the HA data of H3N2 from 1988 to 2008 revealed 42 sites in the HA epitopes. Nine of these sites are also under positive selection according to a dN/dS ratio test (8) (data not shown). However, of the 20 epitope sites where changes rise to fixation over the analysis period (Figure 2), only 12 are detected by the DEPS method (Supplementary Table S3). This highlights that such rapidly fixed changes cannot all be identified by common selection tests.

Retrospectively, our approach allows the identification of the CA04/WI05 antigenic cluster transition in the 2005 Southern season, one year before it rises to predominance in the 2006 season (Figure 6). In all other cases, our method allows us to identify the correct strain one season before the respective antigenic variant becomes predominant: The SY97/MO99 transition is detected in the 1998–99 Northern hemisphere season, while the MO99 variant became predominant in the 1999 Southern hemisphere season. The FU02/CA04 transition is predicted in the 2004 Southern hemisphere season, while CA04 became predominant in the 2004–05 Northern season. Finally, the WI05/BR07 transition is identified in the 2006–07 Northern season, while the BR07 antigenic variant became predominant in the 2007 Southern season. In comparison to the WHO recommendations (13,14,55–58,64–71,73–80), this approach identifies the newly emerging variants one season earlier. This may be because the WHO tends to be conservative in recommendations, to avoid suggesting an antigenic variant that may never actually rise to predominance in the future. However, in general, new variants reach predominance very rapidly, if the time from the first appearance in the available genetic sequences is measured. In all three cases mentioned above, the new variant rose to predominance after its first appearance within a single year. Thus, given the available data, predicting this event one year ahead of time would be impossible. Fortunately, in some cases the antigenic changes between successive variants are not that large (15,37). For instance, MO99 was antigenically similar to SY97. Thus, even though most isolates sampled in the 1999 Southern season reacted to a higher titer with the ferret antisera raised against MO99 (55), recommending SY97 for the vaccine composition thus did not result in a dramatically lower vaccine efficacy.

Influence of timing on antigenic variant identification

Twice a year, in February and September, vaccine strains are recommended for influenza B, influenza A (H3N2) and influenza A (H1N1) to the manufacturers of the seasonal influenza vaccine. This recommendation is made approximately one year before the vaccine will be used in the Northern or Southern seasons, respectively (27). Above, we analyzed the data available only up to that point. If we use all available data until the end of the influenza seasons, emerging alleles appear at high frequencies in the respective AD plot. For example, this happened for the BR07 allele in the 2006–07 Northern hemisphere season (Figure 3, Supplementary Figure S2). Previously circulating strains, on the other hand, occur at lower frequencies in comparison, as newly emerging antigenic variants increase in prevalence typically toward the end of a season. This effect is more pronounced for the Northern hemisphere than for the Southern hemisphere, possibly because after the vaccine meeting in the Northern hemisphere, two months of the winter season are still to follow, whereas only one month of winter still remains in the Southern hemisphere. However, overall the picture remains very similar. Based on all available data, all five antigenic variants can be identified based on their rapid increase in prevalence. A noteworthy difference is evident only for the 2002–03 Northern season, where the *156H *75Q, 155T** allele of the emerging FU02 antigenic variant now ranks first. In summary, limiting the data to what is available by the time of the WHO vaccine meetings, reduces the frequency of alleles associated with newly emerging variants in the AD plot, but the ability to identify viral strains that subsequently rise to predominance is preserved in four out of five cases.

Evolutionary dynamics of the influenza A (H1N1) virus

We next studied the evolutionary dynamics of the 2009 influenza A (H1N1) virus, using 1516 available, exactly dated HA sequences (Figure 4). The virus has circulated in the human population only since April 2009 (81–83). Therefore, we have studied the evolutionary dynamics in monthly intervals (Figure 5, Supplementary Figure S4). As isolate A/California/05/2009 was the only one sampled in March, it was assigned to 1 April to avoid errors introduced through the small sample size for March 2009. The AD plots show that one non-synonymous and another synonymous change become fixed over the analysis period. The corresponding substitutions, T658A [encoding the S206T change (H3 HA1 numbering)] and C1408T (encoding a synonymous substitution for leucine), have already been reported to divide the sequenced isolates into two distinct clusters (84), but have no known antigenic impact (81). Furthermore, Pan *et al.* have already reported an increase in allele frequency for the S206T substitution among new H1N1 sequence isolates (85).

Besides these changes, the plot also reveals the existence of several other alleles, which, so far, appear only at low frequencies and did not become fixed until December of 2009. Despite the fact that the data currently is very limited, at this point, the plots do not reveal any alleles or associated substitutions that seem to be on the rise. Thus, based on the available data, the virus currently seems stable in terms of antigenicity, indicating that no update of the vaccine strain for this virus will be required for the 2010–11 season [also reported by the WHO (14)]. However, some caution is warranted in this interpretation, as different months are represented very unevenly, with lots of data from April and May of 2009 and much less from the following months (Supplementary Figure S5).

DEPS analysis of the H1N1 data identifies five sites in HA with evidence for directional evolution. Three of these sites are also predicted to be under positive selection based on a dN/dS ratio test (Supplementary Table S4). This includes position 206, where a non-synonymous change has become fixed within the analysis period (220 in H1 sequence numbering). This indicates that this site might have been under positive selection and that several further sites could be of relevance for the future evolution of H1N1. However, overall, these results should be taken with care, as the analysis period of 1 year, during which extensive sampling has taken place, is rather short, and the data might be more enriched than samples obtained over longer periods, with many neutral or slightly deleterious mutations.

CONCLUSIONS

AD plots provide a simple and easy to interpret visualization of the evolutionary dynamics of a gene within a population from a sample of dated genetic sequences. This is particularly helpful for the analysis of large-scale sequence datasets, where a standard visualization such as a phylogenetic tree topology is difficult to interpret

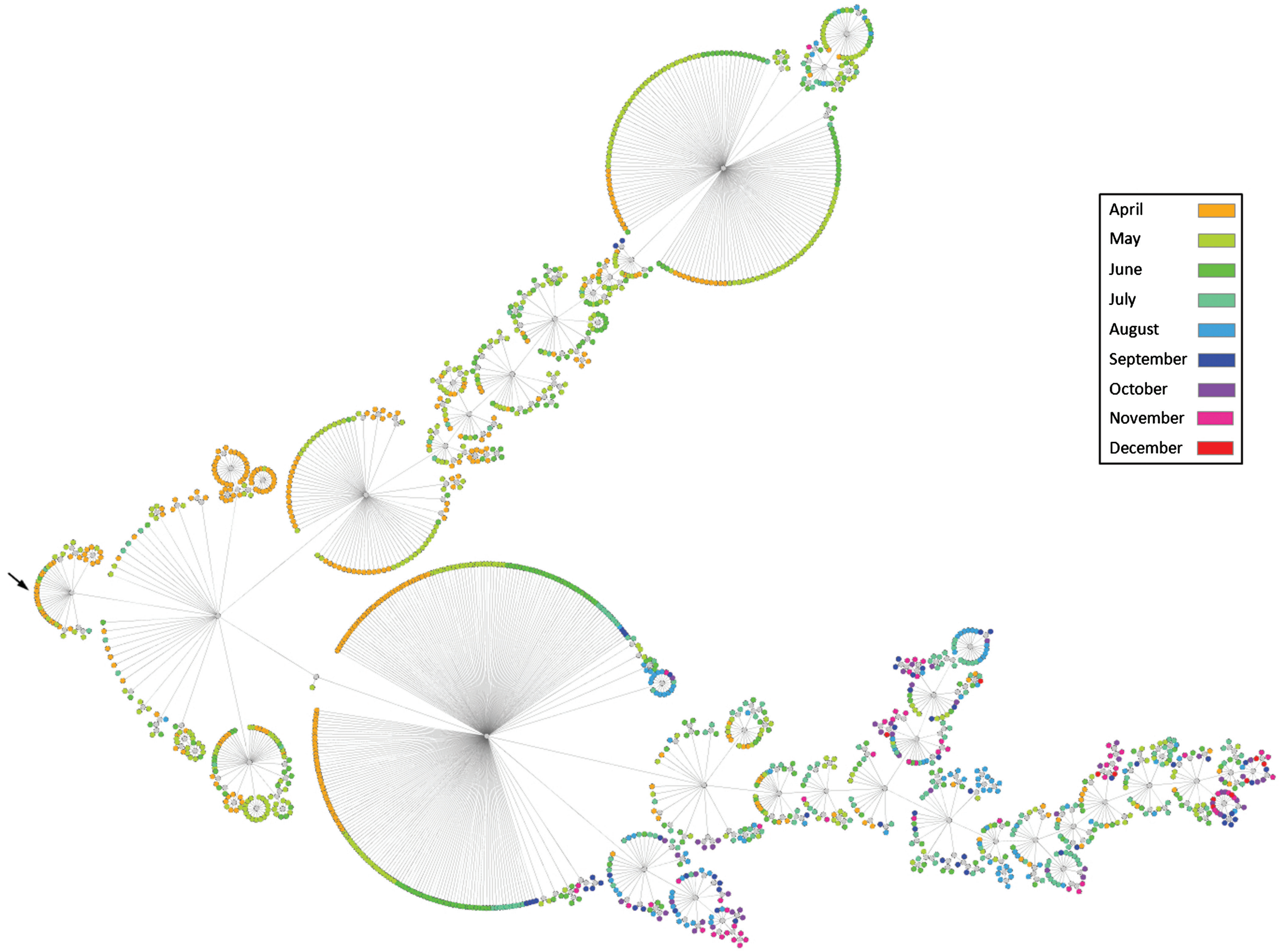

Figure 4. Maximum likelihood tree topology inferred from 1516 2009 swine-origin influenza A (H1N1) hemagglutinin sequences. Leaf nodes are color-coded according to the sampling dates of the viral isolates. The first sampled isolate, A/California/05/2009, is indicated with an arrow.

manually and does not directly display sampling times. Here, we have applied our method to investigate the evolutionary dynamics of seasonal influenza A H3N2 and H1N1 viruses, for which available sequence data is abundant.

From the AD plot for influenza A (H3N2), one can easily determine the order in which substitutions of the surviving lineage became fixed over the analysis period, and one can identify the predominant antigenic variants between 1998 and 2008. Furthermore, we propose a novel indicator for directional selection, which allows us to identify the alleles and corresponding substitutions that might have a selective advantage. We demonstrate this approach for identifying future predominant and novel viral strains. With this method, strains for four out of five antigenic phenotype transitions in influenza A (H3N2) evolution can be identified, based on the data available up to the time of the WHO vaccine strain meeting. One limitation for this application is the fact that a particular allele may score best for every time period, with no information on whether it is antigenically similar or different from the current vaccine strain. Hence, antigenic information also has to be considered to decide whether a vaccine update is warranted. In summary, AD plots enable a sensitive and timely method for detecting emerging viral strains that rise to high frequencies in subsequent seasons. In our analysis, we find that AD plots permit us to accurately identify those alleles that subsequently rise to predominance and become fixed in the course of viral evolution. In combination with antigenic information on the individual strains, AD plots thus present a new tool for the detailed analysis of influenza surveillance data that could be used in the selection of strains for the seasonal influenza A virus vaccine.

Secondly, we used AD plots to analyze the evolutionary dynamics of the 2009 influenza A (H1N1) virus. The AD plot for this virus reveals several new variants with unique genetic composition that circulate at low levels in the human population and two genetic changes that became fixed in the period from April to December 2009. At this point, the plot does not allow identification of any further genetic changes that may become fixed in the near future, indicating that the virus currently is evolutionarily stable, even though data is limited.

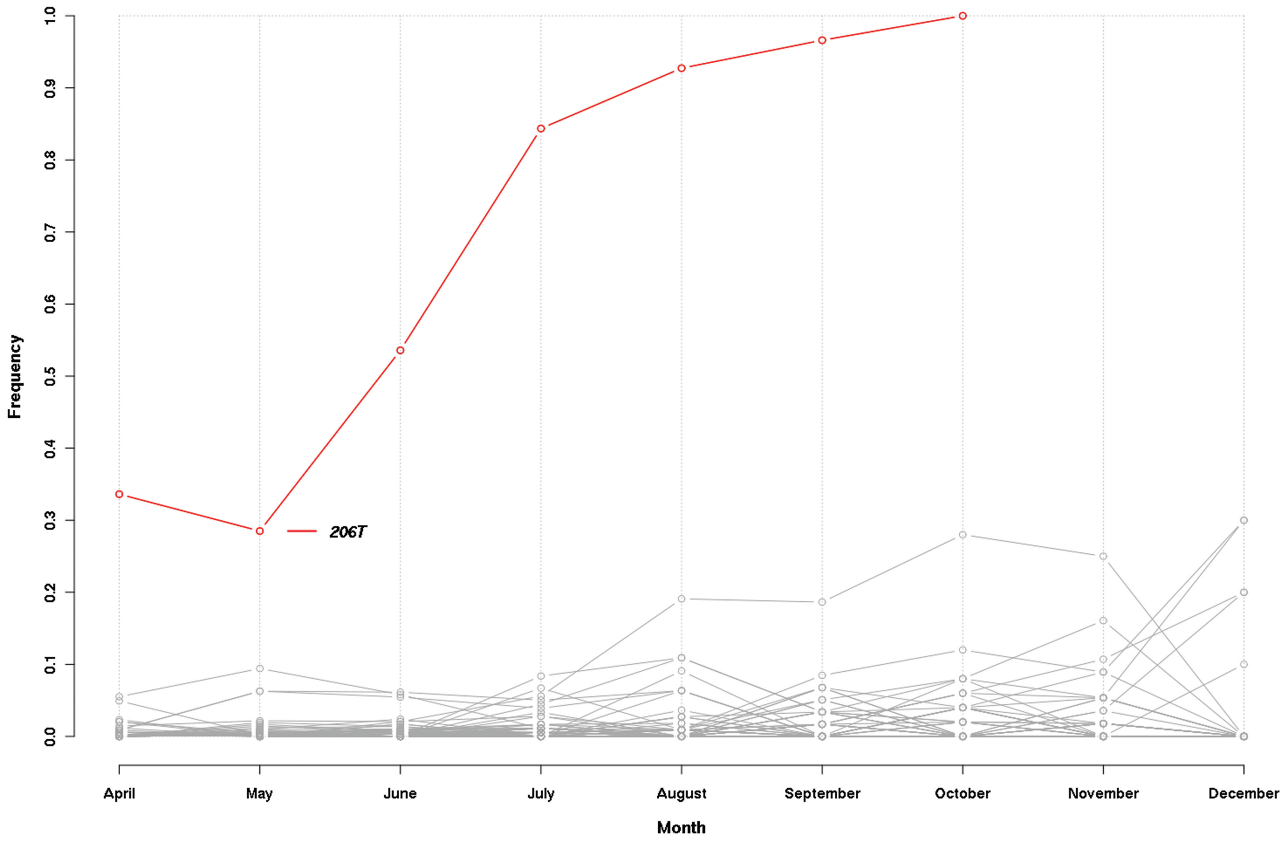

Figure 5. Allele dynamics plot for the major surface protein and antigenic determinant of the new influenza A (H1N1) based on sequences sampled between April and December of 2009 without allele frequency correction. Alleles that reach a prevalence of more than 95% and are subsequently fixed are shown in color; all other alleles are shown in gray. Substitutions are enumerated according to H3 HA1 numbering (86).

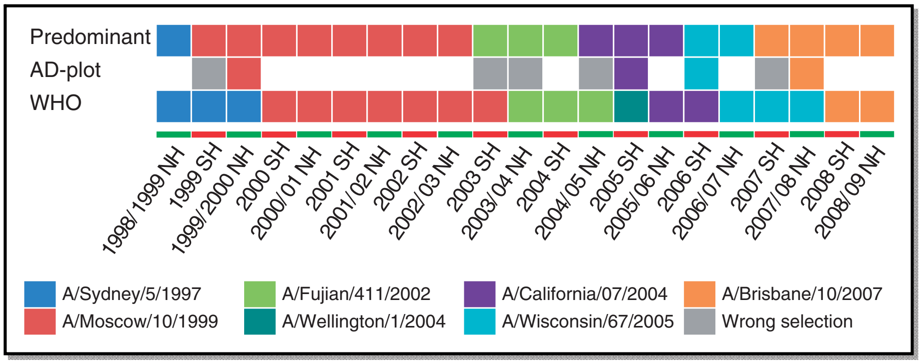

Figure 6. Comparison of predominant influenza A (H3N2) strains, WHO vaccine strain recommendations and strains identified by AD plot analysis. For the AD plot analysis, seasons with antigenic cluster transitions are shown in color. The information shown for the AD plot and the WHO recommendation represents the selection made 1 year earlier.

In summary, we present a novel visualization technique for the study of longitudinal population-level sequence samples and for the identification of alleles that are on the rise to predominance. The method allows us to investigate the evolutionary dynamics of rapidly evolving populations, under consideration of the inherent evolutionary relationships and structure of the data. It complements existing methods for detecting sites under directional and positive selection, such as dN/dS ratio tests or DEPS. Note that AD plots are not limited to the study of influenza A viruses, but can also be applied for the analysis of other fast-evolving populations, such as the intra-host evolution of human immunodeficiency or hepatitis C viruses. Generally, the best results are likely to be obtained if the analyzed sequence sample is representative for a constant-sized population without too much structure (e.g. geographic subdivisions). In this case, variations in frequencies can be taken as estimates for the evolutionary dynamics of the respective population. Finally, while many computational techniques have been applied to predict the evolutionary dynamics of influenza A viruses, our method integrates state-of-the-art phylogenetic inference, ancestral state reconstruction and a novel indicator of directional selection into the analysis, and thus provides a solution with extensive theoretical support.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

L.S. and A.C.M. were funded by the Max-Planck Society.

Conflict of interest statement. None declared.

REFERENCES

1. Grenfell,B.T., Pybus,O.G., Gog,J.R., Wood,J.L.N., Daly,J.M., Mumford,J.A. and Holmes,E.C. (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, **303**, 327–332.
2. Pybus,O.G. and Rambaut,A. (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.*, **10**, 540–550.
3. Wallace,R.G., HoDac,H.M., Lathrop,R.H. and Fitch,W.M. (2007) A statistical phylogeography of influenza A H5N1. *Proc. Natl Acad. Sci. USA*, **104**, 4473–4478.
4. Keele,B.F., Van Heuverswyn,F., Li,Y., Bailes,E., Takehisa,J., Santiago,M.L., Bibollet-Ruche,F., Chen,Y., Wain,L.V., Liegeois,F. *et al.* (2006) Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*, **313**, 523–526.
5. Lemey,P., Pybus,O.G., Wang,B., Saksena,N.K., Salemi,M. and Vandamme,A.-M. (2003) Tracing the origin and history of the HIV-2 epidemic. *Proc. Natl Acad. Sci. USA*, **100**, 6588–6592.
6. Bush,R. (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.*, **16**, 1457–1465.
7. Kryazhimskiy,S. and Plotkin,J.B. (2008) The population genetics of dN/dS. *PLoS Genet.*, **4**, e1000304.
8. Pond,S.L.K., Frost,S.D.W. and Muse,S.V. (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, **21**, 676–679.
9. Pond,K., Sergei,L., Poon,A.F.Y., Brown,L., Andrew,J. and Frost,S.D.W. (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.
10. Ricklefs,R.E. (2007) Estimating diversification rates from phylogenetic information. *Trends Ecol. Evol.*, **22**, 601–610.
11. Koelle,K., Cobey,S., Grenfell,B. and Pascual,M. (2006) Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science*, **314**, 1898–1903.
12. Fouchier,R.A.M., Munster,V., Wallensten,A., Bestebroer,T.M., Herfst,S., Smith,D., Rimmelzwaan,G.F., Olsen,B. and Osterhaus,A.D.M.E. (2005) Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. *J. Virol.*, **79**, 2814–2822.
13. WHO. (2009) Recommended composition of influenza virus vaccines for use in 2009–2010 influenza season (northern hemisphere winter). *WHO Wkly Epidemiol. Rec.*, **84**, 65–72.
14. WHO. (2010) Recommended viruses for influenza vaccines for use in the 2010–2011 northern hemisphere influenza season. *WHO Wkly Epidemiol. Rec.*, **85**, 81–92.
15. Smith,D.J., Lapedes,A.S., de Jong,J.C., Bestebroer,T.M., Rimmelzwaan,G.F., Osterhaus,A.D.M.E. and Fouchier,R.A.M. (2004) Mapping the antigenic and genetic evolution of influenza virus. *Science*, **305**, 371–376.
16. Wiley,D., Wilson,I. and Skehel,J. (1981) Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*, **289**, 373–378.
17. Wiley,D.C. and Skehel,J.J. (1987) The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annu. Rev. Biochem.*, **56**, 365–394.
18. Wilson,I.A. and Cox,N.J. (1990) Structural basis of immune recognition of influenza virus hemagglutinin. *Annu. Rev. Immunol.*, **8**, 737–771.
19. Skehel,J.J. and Wiley,D.C. (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu. Rev. Biochem.*, **69**, 531–569.
20. Kuiken,T., Holmes,E.C., McCauley,J., Rimmelzwaan,G.F., Williams,C.S. and Grenfell,B.T. (2006) Host species barriers to influenza virus infections. *Science*, **312**, 394–397.
21. Webster,R., Bean,W., Gorman,O., Chambers,T. and Kawaoka,Y. (1992) Evolution and ecology of influenza A viruses. *Microbiol. Mol. Biol. R.*, **56**, 152–179.
22. Lowen,A.C. and Palese,P. (2007) Influenza virus transmission: basic science and implications for the use of antiviral drugs during a pandemic. *Infect. Disord. – Drug Targets*, **7**, 318–328.
23. Morens,D.M., Taubenberger,J.K. and Fauci,A.S. (2009) The persistent legacy of the 1918 influenza virus. *N. Engl. J. Med.*, **361**, 225–229.
24. Neumann,G., Noda,T. and Kawaoka,Y. (2009) Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature*, **459**, 931–939.
25. Zimmer,S.M. and Burke,D.S. (2009) Historical perspective – emergence of influenza A (H1N1) viruses. *N. Engl. J. Med.*, **361**, 279–285.
26. Cox,N.J., Brammer,T.L. and Regnery,H.L. (1994) Influenza: global surveillance for epidemic and pandemic variants. *Eur. J. Epidemiol.*, **10**, 467–470.
27. Russell,C.A., Jones,T.C., Barr,I.G., Cox,N.J., Garten,R.J., Gregory,V., Gust,I.D., Hampson,A.W., Hay,A.J., Hurt,A.C. *et al.* (2008) Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*, **26**, 31–34.
28. Nelson,M.I., Simonsen,L., Viboud,C., Miller,M.A., Holmes,E.C. and Levin,B. (2007) Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. *PLoS Pathog.*, **3**, e131.
29. Russell,C.A., Jones,T.C., Barr,I.G., Cox,N.J., Garten,R.J., Gregory,V., Gust,I.D., Hampson,A.W., Hay,A.J., Hurt,A.C. *et al.* (2008) The global circulation of seasonal influenza A (H3N2) viruses. *Science*, **320**, 340–346.
30. Rambaut,A., Pybus,O.G., Nelson,M.I., Viboud,C., Taubenberger,J.K. and Holmes,E.C. (2008) The genomic and epidemiological dynamics of human influenza A virus. *Nature*, **453**, 615–619.
31. Shih,A.C.C., Hsiao,T.C., Ho,M.S. and Li,W.H. (2007) Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc. Natl Acad. Sci. USA*, **104**, 6283–6288.
32. Plotkin,J.B., Dushoff,J. and Levin,S.A. (2002) Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proc. Natl Acad. Sci. USA*, **99**, 6263–6268.
33. Du,X., Wang,Z., Wu,A., Song,L., Cao,Y., Hang,H. and Jiang,T. (2008) Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution. *Genome Res.*, **18**, 178–187.
34. Xia,Z., Jin,G., Zhu,J. and 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.
35. Bush,R.M., Bender,C.A., Subbarao,K., Cox,N.J. and Fitch,W.M. (1999) Predicting the evolution of human influenza A. *Science*, **286**, 1921–1925.
36. Nozawa,M., Suzuki,Y. and Nei,M. (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc. Natl Acad. Sci. USA*, **106**, 6700–6705.
37. Fouchier,R.A.M. and Smith,D.J. (2010) Use of antigenic cartography in vaccine seed strain selection. *Avian Dis.*, **54**, 220–223.
38. Huang,J.W., King,C.C. and Yang,J.M. (2009) Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses. *BMC Bioinformatics*, **10**, S41.
39. Lee,M.S., Chen,M.C., Liao,Y.C. and Hsiung,C.A. (2007) Identifying potential immunodominant positions and predicting antigenic variants of influenza A/H3N2 viruses. *Vaccine*, **25**, 8133–8139.

40. Liao,Y.C., Lee,M.S., Ko,C.Y. and Hsiung,C.A. (2008) Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics*, **24**, 505–512.
41. Bao,Y., Bolotov,P., Dernovoy,D., Kiryutin,B., Zaslavsky,L., Tatusova,T., Ostell,J. and Lipman,D. (2008) The influenza virus resource at the National Center for Biotechnology Information. *J. Virol.*, **82**, 596–601.
42. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
43. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate method to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
44. Zwickl,D. (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. *Thesis*. The University of Texas at Austin.
45. Futuyma,D.J. (1998) *Evolutionary Biology*, 3rd edn. Sinauer Associates, Sunderland, MA.
46. Hein,J., Schierup,M. and Wiuf,C. (2005) *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford.
47. Templeton,A.R. (2006) *Population Genetics and Microevolutionary Theory*. Wiley-Liss, Hoboken, NJ.
48. Huelsenbeck,J.P. and Ronquist,F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.
49. Drummond,A. and Rambaut,A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, **7**, 214.
50. Fitch,W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406–416.
51. Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.
52. Pagel,M., Meade,A. and Barker,D. (2004) Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.*, **53**, 673–684.
53. Nelson,M.I. and Holmes,E.C. (2007) The evolution of epidemic influenza. *Nat. Rev. Genet.*, **8**, 196–205.
54. Ghedin,E., Sengamalay,N.A., Shumway,M., Zaborsky,J., Feldblyum,T., Subbu,V., Spiro,D.J., Sitz,J., Koo,H., Bolotov,P. *et al.* (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, **437**, 1162–1166.
55. WHO. (1999) Recommended composition of influenza virus vaccines for use in the 2000 influenza season. *WHO Wkly Epidemiol. Rec.*, **74**, 321–325.
56. WHO. (2000) Recommended composition of influenza virus vaccines for use in the 2000–2001 season. *WHO Wkly Epidemiol. Rec.*, **75**, 61–65.
57. WHO. (2001) Recommended composition of influenza virus vaccines for use in the 2001–2002 influenza season. *WHO Wkly Epidemiol. Rec.*, **76**, 58–61.
58. WHO. (2002) Recommended composition of influenza virus vaccines for use in the 2002–2003 influenza season. *WHO Wkly Epidemiol. Rec.*, **77**, 62–66.
59. Lin,Y., Gregory,V., Bennett,M. and Hay,A. (2004) Recent changes among human influenza viruses. *Virus Res.*, **103**, 47–52.
60. Hay,A.J., Lin,Y.P., Gregory,V. and Bennet,M. (2003) *WHO Collaborating Centre for Reference and Research on Influenza, Annual Report*. National Institute for Medical Research, London.
61. Hay,A.J., Lin,Y.P., Gregory,V. and Bennet,M. (2005) *WHO Collaborating Centre for Reference and Research on Influenza, Interim Report February*. National Institute for Medical Research, London.
62. Hay,A.J., Lin,Y.P., Gregory,V. and Bennet,M. (2006) *WHO Collaborating Centre for Reference and Research on Influenza, Interim Report March*. National Institute for Medical Research, London.
63. Hay,A.J., Daniels,R., Lin,Y.P., Xiang,Z., Gregory,V., Bennet,M. and Whittaker,L. (2007) *WHO Collaborating Centre for Reference and Research on Influenza, Interim Report September*. National Institute for Medical Research, London.
64. WHO. (2000) Recommended composition of influenza virus vaccines for use in the 2001 influenza season. *WHO Wkly Epidemiol. Rec.*, **75**, 330–333.
65. WHO. (2001) Recommended composition of influenza virus vaccines for use in the 2002 influenza season. *WHO Wkly Epidemiol. Rec.*, **76**, 311–314.
66. WHO. (2002) Recommended composition of influenza virus vaccines for use in the 2003 influenza season. *WHO Wkly Epidemiol. Rec.*, **77**, 344–348.
67. WHO. (2003) Recommended composition of influenza virus vaccines for use in the 2003–2004 influenza season. *WHO Wkly Epidemiol. Rec.*, **78**, 58–62.
68. WHO. (2003) Recommended composition of influenza virus vaccines for use in the 2004 influenza season. *WHO Wkly Epidemiol. Rec.*, **78**, 375–379.
69. WHO. (2004) Recommended composition of influenza virus vaccines for use in the 2004–2005 influenza season. *WHO Wkly Epidemiol. Rec.*, **79**, 88–92.
70. WHO. (2004) Recommended composition of influenza virus vaccines for use in the 2005 influenza season. *WHO Wkly Epidemiol. Rec.*, **79**, 369–373.
71. WHO. (2005) Recommended composition of influenza virus vaccines for use in the 2005–2006 influenza season. *WHO Wkly Epidemiol. Rec.*, **80**, 66–71.
72. Jin,H., Zhou,H., Liu,H., Chan,W., Adhikary,L., Mahmood,K., Lee,M.S. and Kemble,G. (2005) Two residues in the hemagglutinin of A/Fujian/411/02-like influenza viruses are responsible for antigenic drift from A/Panama/2007/99. *Virology*, **336**, 113–119.
73. WHO. (2005) Recommended composition of influenza virus vaccines for use in the 2006 influenza season. *WHO Wkly Epidemiol. Rec.*, **80**, 342–347.
74. WHO. (2006) Recommended composition of influenza virus vaccines for use in the 2007 influenza season. *WHO Wkly Epidemiol. Rec.*, **81**, 390–395.
75. WHO. (2007) Recommended composition of influenza virus vaccines for use in the 2007–2008 influenza season. *WHO Wkly Epidemiol. Rec.*, **82**, 69–74.
76. WHO. (2006) Recommended composition of influenza virus vaccines for use in the 2006–2007 influenza season. *WHO Wkly Epidemiol. Rec.*, **81**, 82–86.
77. WHO. (2007) Recommended composition of influenza virus vaccines for use in the 2008 influenza season. *WHO Wkly Epidemiol. Rec.*, **82**, 351–356.
78. WHO. (2008) Recommended composition of influenza virus vaccines for use in the 2008–2009 influenza season. *WHO Wkly Epidemiol. Rec.*, **83**, 81–87.
79. WHO. (2008) Recommended composition of influenza virus vaccines for use in the 2009 southern hemisphere influenza season. *WHO Wkly Epidemiol. Rec.*, **83**, 366–372.
80. WHO. (1999) Recommended composition of influenza virus vaccines for use in the 1999–2000 season. *WHO Wkly Epidemiol. Rec.*, **74**, 57–61.
81. Garten,R.J., Davis,C.T., Russell,C.A., Shu,B., Lindstrom,S., Balish,A., Sessions,W.M., Xu,X., Skepner,E., Deyde,V. *et al.* (2009) Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science*, **325**, 197–201.
82. Smith,G., Vijaykrishna,D., Bahl,J., Lycett,S., Worobey,M., Pybus,O., Ma,S., Cheung,C., Raghwani,J., Bhatt,S. *et al.* (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, **459**, 1122–1125.
83. Novel Swine-Origin Influenza,A. (H1N1) Virus Investigation Team. (2009) Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N. Engl. J. Med.*, **360**, 2605–2615.
84. Fereidouni,S.R., Beer,M., Vahlenkamp,T. and Starick,E. (2009) Differentiation of two distinct clusters among currently circulating influenza A(H1N1)v viruses, March–September 2009. *Euro Surveill.*, **14**, 19409–19411.
85. Pan,C., Cheung,B., Tan,S., Li,C., Li,L., Liu,S. and Jiang,S. (2010) Genomic signature and mutation trend analysis of pandemic (H1N1) 2009 influenza A virus. *PLoS ONE*, **5**, e9549.
86. Nobusawa,E., Aoyama,T., Kato,H., Suzuki,Y., Tateno,Y. and Nakajima,K. (1991) Comparison of complete amino acid sequences and receptor-binding properties among 13 serotypes of hemagglutinins of influenza A viruses. *Virology*, **182**, 475–485.