Sweep Dynamics (SD) plots: Computational identification of selective sweeps to monitor the adaptation of influenza A viruses

Thorsten R. Klingen[1,*], Susanne Reimering[1*], Jens Loers[1], Kyra Mooren[1], Frank Klawonn[2,3], Thomas Krey[4,5], Gülsah Gabriel[6,7] and Alice C. McHardy[1,5 †]

[1]Department for Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany

[2]Biostatistics Group, Helmholtz Center for Infection Research, Braunschweig, Germany

[3]Department of Computer Science, Ostfalia University of Applied Sciences, Wolfenbüttel, Germany

[4]Institute of Virology, Hannover Medical School, Hannover, Germany

[5]German Center for Infection Research (DZIF)

[6]Viral Zoonoses and Adaptation, Heinrich Pette Institute, Leibniz Institute for Experimental Virology, Hamburg, Germany

[7]University of Lübeck, Germany

[†]Correspondence to: Alice.McHardy@helmholtz-hzi.de

*Co-first authors.

Keywords: influenza A virus, selective sweep, amino acid changes, directional selection, population genomics

Abstract

Monitoring changes in the genome of influenza A viruses is crucial to understand its rapid evolution. These changes reflect how the virus adapts to changing environmental conditions such as establishment within a novel host species. Selective sweeps represent a rapid mode of adaptation and are typically observed in the evolution of human influenza A viruses. Here we describe a computational method named Sweep Dynamics (SD) plots. These combine phylogenetic algorithms with statistical techniques, allowing to characterize the molecular adaptation of rapidly evolving viruses from longitudinal sequence data. To our knowledge, it is the first method that allows not only to identify selective sweeps, but also the time periods in which these occurred and the associated amino acid changes that may have provided a selective advantage to the virus. Using SD plots, we studied the past genome-wide adaptation of the 2009 pandemic H1N1 influenza A (pH1N1) and seasonal H3N2 influenza A (sH3N2) viruses. The pH1N1 influenza virus showed simultaneous amino acid changes in various proteins, particularly in seasons of high pH1N1 activity. Some of these adaptive changes resulted in functional alterations that facilitated virus transmission by respiratory droplets, which is key for sustained human-to-human transmission, directly after pandemic emergence. In the evolution of sH3N2 influenza viruses since 1999, we detected a large number of amino acid changes characterizing vaccine strains. Amino acid changes found in antigenically novel strains rising to predominance were occasionally revealed in a selective sweep one season prior to the recommendation of the WHO, suggesting the value of the technique for the vaccine strain selection problem. Taken together, our results show that SD plots allow to monitor and characterize the adaptive evolution of influenza A viruses by identifying selective sweeps and their associated signatures.

Introduction

Influenza A viruses are rapidly evolving pathogens causing respiratory infections with high morbidity and mortality in the human population (WHO 2014a). Annual influenza epidemics result in 3 to 5 million reported infections and up to 250,000–500,000 cases of death (WHO 2014a). Currently, the viral subtypes sH3N2 and pH1N1 are circulating in the human population. The H3N2 virus was introduced into the human population in 1968 and is endemic ever since. The swine-origin H1N1 subtype emerged in the 2009 influenza pandemic and was subsequently referred to as the 2009 pH1N1 virus. It replaced the formerly circulating seasonal H1N1 subtype (WHO 2009, 2011a). The negative-sense RNA genome consists of eight segments that encode for 14 viral proteins (Wise et al. 2012). A constant arms-race between the human immune system and the virus results in continuous adaptation of the viral genome. These changes facilitate the virus to escape the host's immune response elicited through vaccination or previous influenza infections (McHardy and Adams 2009). Alterations in the major glycoproteins − hemagglutinin (HA) and neuraminidase (NA) − and genomic reassortment that change the viral antigenicity are defined as antigenic drift and result in re-occurring epidemics of seasonal influenza viruses (Medina and García-Sastre 2011). Continuous antigenic changes of circulating strains require a re-evaluation of antigenically predominant strains by the WHO twice a year at the end of each season, leading to a recommendation for the vaccine composition for the following year (Smith et al. 2004; Russell et al. 2008). The establishment of an antigenically new virus strain into an immunologically naive human population causes pandemics, mostly due to alterations in the receptor binding protein HA (Bouvier and Palese 2008).

Measuring the impact of natural selection plays a crucial role in molecular evolution, as it determines the genomic constitution and diversity of a population (Vitti et al. 2013). Directional evolutionary processes boost the fitness of individuals by introducing advantageous amino acids. The respective alleles rise in frequency in the viral population, consequentially reducing variation. This process is called a 'selective

3

sweep' (Kosakovsky Pond et al. 2008). Hereafter, we use the term 'sweep-related change' when referring to an amino acid exchange rising in frequency due to a selective sweep. Due to linkage within genomic segments, the advantageous amino acid exchange promoting the sweep and other changes in close genomic proximity jointly rise in frequency, leading to an increase of linkage disequilibrium (i.e. nonrandom associations between genomic regions). Thus, selective sweeps are of profound interest since they represent a rapid shift of the whole genotype carrying the selected amino acid change. The increasing number of available viral genomes that are provided by modern sequencing techniques allows us to conduct a genome-wide analysis for recent selective sweeps to gain insight into the within-host evolution and adaptation of human influenza A viruses.

Several methods were developed to detect and measure the effect of directional evolution for viral populations. We have previously described Allele Dynamics (AD) plots, which characterize the evolutionary dynamics of sets of amino acid changes, indicating those most likely to be under positive selection using population level time-series data sets of genetic sequences (Steinbrück and McHardy 2011). Other methods to analyze natural selection are based on non-synonymous to synonymous mutation rates (dN/dS) (Yang and Bielawski 2000). Synonymous changes are assumed to be neutral, while a relative excess of non-synonymous changes, i.e. $dN/dS > 1$, indicates positive selection (Yang and Bielawski 2000). dN/dS is either calculated by counting synonymous and non-synonymous substitutions (Nei and Gojobori 1986; Yang and Nielsen 2000) or by estimation using maximum likelihood models (Nielsen and Yang 1998; Yang and Nielsen 2002). This statistic is not applicable to detect selective sweeps, as a large number of synonymous substitutions could occur after a substitution rose to fixation in a selective sweep, which would result into a dN/dS ratio smaller than 1 and the conclusion that this site is not evolving under positive selection (Kosakovsky Pond et al. 2008). Moreover, dN/dS was originally developed for the analysis of divergent species and the interpretation of $dN/dS > 1$ as positive selection may not be correct for single populations (Kryazhimskiy and Plotkin 2008a). To elude the limitations of

dN/dS, (Bhatt et al. 2011) developed a statistic based on site frequencies to calculate site- or segment-specific adaptation rates. Furthermore, several other approaches have been developed to identify selective sweeps using accelerated substitution rates (Kosakovsky Pond et al. 2008), a skew in the allele frequency spectrum (Nielsen et al. 2005) or an excess of linkage disequilibrium (Kim and Nielsen 2004). Detecting the season or the time period in which a specific amino acid change involved in a selective sweep emerged, would improve comparative molecular studies of natural selection and indicate changes with potential effect on the viral fitness. This could aid to uncover the drivers of adaptive evolution in viral populations.

Here, we describe Sweep Dynamics (SD) plots, which allow analyzing the population-level phylodynamics of influenza virus proteins or proteins of other rapidly evolving organisms from longitudinal samples of genetic sequences. A statistical evaluation reveals selective sweeps, and in addition the season in which they occurred and the associated individual amino acid changes. We used the SD plots for a genome-wide characterization of directional selection in pH1N1 influenza viruses since their introduction into the human host. In all proteins under consideration, we inferred sweep-related changes that indicate human-adaptive changes after its emergence in 2009, several of which were in structural proximity to known mammalian host adaptation sites. Furthermore we detected sweep-related changes in antigenicity- and avidity-changing sites of the sH3N2 influenza virus hemagglutinin that correlate with newly emerging antigenic variants in the human population and show the value of SD plots for vaccine strain selection.

Data & Methods

Data download and sampling

Protein and nucleotide coding sequences were downloaded from the NCBI flu database (Bao et al. 2008). For both the pH1N1 influenza and sH3N2 influenza subtype, we downloaded sequences of the proteins HA, NA, M1, M2, NS1, NS2, NP, PA, PB1 and PB2. The pH1N1 influenza virus was analyzed from the emergence of the virus in early 2009 until the end of September 2015, while sH3N2 influenza was studied from October 1998 until the end of September 2015. We downloaded sequences with the full date available, to properly assign sequences to influenza seasons. We used the standardized definition for the Northern hemisphere season (N) that begins on 1st October of the previous year and ends on 31st March and the Southern hemisphere season (S) that begins on 1st April and ends on 30th September in the same year, as before (Steinbrück and McHardy 2011).

To account for variable numbers of sequences per season, we sampled the data with replacement, resulting in the same number of sequences per season for our analysis. We further guaranteed that our data comprised a nucleotide and a corresponding amino acid sequence for each isolate. For HA, the numbering of amino acids is based on the mature protein without the signal peptides of seventeen and sixteen amino acids for pH1N1 influenza viruses and sH3N2 influenza viruses, respectively (Burke and Smith 2014).

In the following, we describe the analysis workflow that was applied to all influenza proteins.

Alignment & Phylogenetic inference

We generated multiple sequence alignments with nucleotide and amino acid sequences using MUSCLE (Edgar 2004), separately. To avoid shifts in the numbering of both alignments, positions with gaps in more than 80% of the sequences are removed from the alignment with TrimAl (Capella-Gutiérrez et al.

2009). Amino acid positions are numbered in increasing order as made by Hay et al. (2009). A phylogenetic tree is inferred based on the nucleotide alignment using fasttree (Price et al. 2010) with the GTR-model, which enables a quick tree computation for a large number of sequences while retaining a good accuracy compared to slower methods. For the pH1N1 influenza virus subtype, the A/California/05/2009 strain and for the sH3N2 influenza virus subtype the A/Moscow/10/1999 strain were used as an outgroup to root the trees. Both viruses were predominantly circulating in the viral population in the first season of our analyses (WHO 1999, 2009). We resolve multifurcations by adding further nodes and zero length edges into the outgoing edges to obtain a binary representation of the tree and subsequently apply the parsimony model of Fitch's algorithm with accelerated transformation (ACCTRAN) (Fitch 1971) to reconstruct amino acid sequences for the internal nodes of the phylogenetic trees. Amino acid changes are then inferred from the node-associated amino acid sequences and mapped to branches of the tree.

Protein structure analyses

To investigate the structural properties of sweep-related changes, we mapped all sweep-related changes in HA1, NA, NP and the polymerase (PB1, PB2 and PA) onto the respective structures. For NA, the crystal structure of the A/California/07/2009(H1N1) strain (PDB 4B7Q) was used that covers residues 83-469 of NA. For HA1 and NP, we generated homology models based on structures from the RSCB database and the amino acid sequence of the A/California/07/2009(H1N1) strain using MODELLER (Sali and Blundell 1993; Webb and Sali 2014). Multiple sequence alignments containing protein templates and target sequences for modeling were calculated with MUSCLE (Edgar 2004). Two structures (PDB 3M6S and 3LZG) from the A/Darwin/2001/2009(H1N1) strain displaying a sequence identity of 99.07% within residues 1-322 compared to the A/California/07/2009(H1N1) sequence were used as templates for composite modeling. For NP, we used the crystal structure of NP from the A/Wilson-Smith/1933(H1N1) strain (PDB 3RO5) with an identity of 91.9% compared to the A/California/07/2009(H1N1) strain. To

7

obtain a three dimensional model of the viral polymerase, we generated individual models of PA, PB1 and PB2 for the A/California/05/2009(H1N1) strain using the Phyre II server (Kelley et al. 2015). These models were subsequently superposed onto the structure of the intact influenza A polymerase of Reich et al. (2014) (PDB 4WSB).

We determined the exposure of each site with the tool SURFACE from the CCP4 toolkit (Lee and Richards 1971; Winn et al. 2011) and calculated the accessible surface area (ASA), using a 1.4 Å radius for the probe sphere equal to the radius of water and the van der Waals radii for different types of atoms defined by (Chothia 1976), as well as a step variable of 0.1. To calculate the relative solvent accessibility (RSA), we normalized ASA values with a theoretical value of the maximum possible solvent accessibility and classified residues as buried when the RSA is ≤ 0.05 and as exposed otherwise (Tien et al. 2013).

Results

Sweep Dynamic (SD) Plots

Sweep Dynamics (SD) plots are an extension of the AD plot technique that we previously described (Steinbrück and McHardy 2011). SD plots analyze the evolutionary dynamics of alleles, representing individual amino acid changes within the viral population. The dynamics of reconstructed amino acid changes (alleles) in an homogeneous, constant-sized viral population can be described by a Fisher model (Fisher 1922a; Wright 1931). This model is also considered valid with changing population sizes (Wright 1933; Wright 1938; Charlesworth 2009) and seasonal influenza virus populations have been modelled as homogenous due to their rapid spread around the globe (Kryazhimskiy and Plotkin 2008b). Genetic drift and selection are acting on alleles, resulting in changes of allele frequencies (ratio of copies of one genetic variant relative to the population size) (Hein et al. 2005; Templeton 2006). In this population, an allele under directional selection rises faster in frequency than alleles without a selective advantage. Amino acid changes (alleles) that increase in frequency swiftly over time are thus more likely to be under directional selection than other alleles with lower frequencies. We apply this criterion in the SD plots method together with a statistical evaluation, to pinpoint those amino acid changes that increase significantly faster in frequency than others and thus might provide a selective advantage. Changes are inferred under consideration of the evolutionary relationships, to consider their relatedness and avoid overcounting changes resulting from denser sampling of specific parts of the viral population.

From the sequences for a particular protein from a viral population sample, a phylogenetic tree is inferred and amino acid changes in its evolution are reconstructed that map to the individual branches in the tree (Material & Methods). The frequencies of circulating isolates that have inherited a particular amino acid change are deduced from the phylogenetic tree by counting the number of viral isolates descending from the branch where a particular amino acid change was introduced. This is done separately for each season

(Figure 1A&B). Other than the AD-plots, which combine amino acid changes at different positions of the coding sequence into alleles when they share a branch or occur in close proximity in the phylogeny, SD plots analyze individual amino acid changes for their potential effect on viral fitness. The visualization of SD plots highlights these identified sweep-related changes and their variation in frequency over time. Specifically, the following procedure is carried out:

Using a phylogenetic tree T and a set of reconstructed amino acid changes A_T for all internal branches, the frequency of isolates in the population that carry a specific amino acid change $a \in A_T$ is calculated as follows: let L_T represent the set of viral isolates assigned to the leaves in T, with each viral isolate $l \in L_T$ being labeled with a time stamp s indicating the season in which it was sampled and L_T, s representing the subset of sequences of L_T labeled with time stamp s. Starting from the root, we perform a level order traversal, visiting each edge e in the tree T, and calculate the frequency for each individual amino acid change $a \in A_T$ that is located on the edge e as follows: Let T_a be the subtree that is rooted at the node with the in-edge e. Let A_{T_a} be the set of amino acid changes in the subtree T_a, with $A_{T_a} \subset A_T$ and $a \in A_T \wedge a \notin A_{T_a}$. The frequency $f(a, s)$ of an individual amino acid change a in season s is defined as:

$f(a, s) = \frac{|L_{T_a,s}|}{|L_{T,s}|}$. It represents the ratio of all isolates in the subtree that have acquired the amino acid change a within season s relative to the number of all isolates within the designated season s.

As we consider individual positions, the prevalence of one amino acid change per position is affected by the emergence of a more recent amino acid change at the same position, making it necessary to adjust the frequency $f(a, s)$ at position p in the alignment. The frequency $f(a, s)$ of a occurring at position p is adjusted when another amino acid change β occurs at p in the subtree T_a − a process referred to as frequency correction, as follows: let there be n amino acid changes at p in the subtree T_a, i.e. $|A_{T_a}| = n$. Each subtree τ_β is rooted at the node with the in-edge that represents the amino acid change $\beta \in A_{T_a}$. The set $\{\tau_\beta \mid \beta \in A_{T_a}\} = X_{T_a}$ represents the n subtrees that are contained in the tree T_a. Let $L_{X_{T_a},s} =$

$\bigcup_{\tau \in X_{T_a}} L_{\tau,s}$ be the set of all leaves with the time stamp s in all subtrees in X_{T_a}. Note that $L_{X_{T_a},s} \subseteq L_{T_a,s}$.

The corrected frequency $f_{corr}(a,s)$ of amino acid change a in season s is then defined as: $f_{corr}(a,s) = \frac{|L_{corr}(T_a,s)|}{|L_{T,s}|}$, with $L_{corr}(T_a,s) = L_{T_a,s} - L_{X_{T_a},s}$. The frequency of an amino acid change a is adjusted by

excluding isolates $L_{X_{T_a},s}$ that are subject to a more recent substitution (Figure 1). The SD plots always

apply $f_{corr}(a,s)$ because either the frequency needs to be corrected or whenever no additional amino acid

change occurs in the phylogenetic tree T after the amino acid change α, the term $\left|L_{X_{T_a},s}\right|$ is zero and

automatically results in the uncorrected case ($f(a,s)$). Thus, the equation $f_{corr}(a,s) \leq f(a,s)$ holds.

Note that this allows an individual analysis of the same amino acid exchange introduced several times in

the phylogeny on different branches (e.g. S1T (red & light green) in Figure 1) and their distinct

evolutionary trajectories.

To identify a selective sweep and the associated amino acid change, we define two criteria: an amino acid

change should show a significant increase in frequency relative to the previous season. For each season,

we only report changes that were not reported as significant before. For each amino acid change, we test

the null hypothesis that the number of viruses in the viral population carrying this change is equal or

lower than in the previous season. We evaluate the significance of frequency changes for each amino acid

change a using Fisher's exact test (Fisher 1922b), using $|L_{corr}(T_a,s)|$ and $|L_{corr}(T_a,s+1)|$ over

consecutive seasons s and $s+1$. A significant p-value ($p \leq 0.05$) indicates that an amino acid change

significantly increased in frequency in the current season relative to the previous one. To correct for

multiple testing, we adjusted the p-values with the Benjamini-Hochberg procedure controlling the false

discovery rate at level $\alpha = 0.05$ (Benjamini and Hochberg 1995).

For each dataset, the seasons from the second to the last one are tested, comparing each season to the

preceding one. We report amino acid changes for the season s of their first predominant occurrence

$(f_{corr}(a, s) > 0.5$; it occurs in more than 50% of the isolates within the designated time period) as in (Steinbrück and McHardy 2011) in combination with a significant p-value ($p \leq 0.05$).

The results are visualized in SD plots (Figure 2 & 3). These provide a detailed overview of the emergence of individual amino acid changes in the viral population and in which season a sweep took place. The changes in frequency of sweep-related changes are depicted as trajectories over consecutive seasons. The season of their first predominant occurrence is indicated with an asterisk. In each season, the amino acid exchanges of all sweep-related alterations are listed in a panel above the graph. Within the panel, they are bottom-up ordered with ascending frequency.

Selective sweeps in the evolution of pH1N1 influenza viruses

We applied the SD plots to detect selective sweeps in the past evolution of pH1N1 influenza viruses. We analyzed nucleotide and amino acid sequences of ten proteins (HA, NA, M1, M2, NS1, NS2, NP, PA, PB1 and PB2) collected since the appearance of the virus in the beginning of 2009 until the end of September 2015 (WHO 2009). Sequences were assigned to influenza seasons using the common definitions for seasons in the Northern and Southern hemisphere (Data & Methods). For pH1N1 influenza, the data covered fourteen seasons from 2009N to 2015S (Figure 2, Supplementary Table 1). To investigate the structural relationships of sweep-related changes in the proteins, we mapped sweep-related changes in HA1, NA, NP and the polymerase (PB1, PB2 and PA) onto the respective structures (Material & Methods).

The SD plots analysis indicated selective sweeps and associated changes for the 2009S, 2011N, 2012N, 2013N, 2013S, 2014N, 2014S and 2015N seasons. The most sweep-related changes (twelve and seventeen, respectively) were detected for the surface proteins HA and NA and the fewest were found for NS2 (two changes) over all thirteen seasons. Newly arising changes often occurred simultaneously, i.e. were detected in the same season, with up to five changes detected in a protein in one season (2013S in

12

PA and NA, 2013N in PB2), and with changes occurring simultaneously in multiple proteins (Figure 2). In addition to changes providing a selective advantage, some of these are likely hitchhikers without notable effect on fitness. With hitchhikers, we refer to (almost) neutral changes that are introduced into a sequence shortly before or after a change causing a selective sweep. Hitchhikers then rise in frequency together with beneficial changes due to genomic linkage and are thus detected as being sweep-related. Technically, the simultaneous occurrence of amino acid changes on the same branch prevents a computational distinction between their potential effects, as they are ancestral to the same set of leaf nodes. Besides linkage within a genomic segment, there is also a strong linkage across all influenza segments (Rambaut et al. 2008; Varich et al. 2008; Khiabanian et al. 2009; Wille et al. 2013), resulting in hitchhikers from other segments being carried along to higher frequencies by functionally relevant changes. The visualization of evolutionary dynamics in the SD plots can identify some of these 'sweep-related hitchhikers' as those decreasing in frequency after their emergence and not becoming fixed.

For the 2009S season, we detected a selective sweep at seven amino acid positions throughout the viral genome (in PA (L581M), HA (S203T) (position 202 in Otte et al. (2016)), NP (V100I, I373T), NA (V106I, N248D) and NS1 (I123V)) (Figure 2C, 2D, 2E, 2F and 2I, Supplementary Table 1), which were also described in (Otte et al. 2016). All of these changes rapidly became predominant in this season and continued to rise in frequency until they were close to fixation in 2010N, suggesting that the new amino acid was present in nearly all sampled isolates in this season. These findings are in agreement with the study of Elderfield et al. (2014), in which 2009S was identified as the time of the first wave of pH1N1 activity after its emergence. The pandemic waves were defined based on large viral prevalence with a phase of low viral prevalence in between. Five of seven detected sweep-related changes were also described in this study as emerging changes in the first wave of the pandemic, with frequencies close to 100% in the second wave in 2010N, similar to the frequencies we observed in our analysis (frequencies between 91% and 99%, with slight differences most likely due to differing time frames and datasets in both analyses). With the exception of the V106I and N248D changes in NA, all sweep-related changes

13

detected in 2009S were still fixed in the population in 2015. The frequency of the amino acid change V106I in NA decreased from 2011S onwards, as it was replaced by the I106V sweep-related change in 2013N. The frequency of V106I accordingly dropped rapidly from 78% to 11% between 2012S and 2013N and continued to decrease until it disappeared in 2014S. From 2011S onwards, the frequency of N248D also decreased to 0% in 2014S. This amino acid change re-emerged in 2013N and replaced the strain with the first change. This repeated emergence indicates a particular relevance of this change regarding viral fitness. All seven sweep-related changes detected in 2009S have been reported to induce functional changes that might have facilitated viral adaptation to the human population: we previously described that all changes with the exception of NS I123V mediated increased pH1N1 influenza virus pathogenicity in a mouse model (Otte et al. 2015). Moreover, all seven adaptive mutations were associated with enhanced respiratory droplet transmission in a ferret model, suggesting their crucial role to sustain with the mammalian host (Otte et al. 2016). The ability of influenza A viruses to transmit from human-to-human is considered to be key in pandemic spread. Hereby, particularly the receptor binding properties of HA play an important role. We have also previously shown that HA S203T increases binding to α2,6-linked sialic acids (corresponds to position 202 in Otte et al. (2016)) that are predominantly expressed in the upper respiratory tract of humans, a key site in virus transmission among humans (Shinya et al. 2006; van Riel et al. 2006; Otte et al. 2016).

Both residues I100 and T373 in NP are solvent exposed ($RSA = 0.32$ and $= 0.26$, respectively); T373 is in close proximity to N319 (10.2 Å measured from Cα to Cα, Figure 5A), at which an N to K change promotes adaptation to mammalian cells (Gabriel et al. 2008). Elevated virulence was also found for substitutions V106I in NA and I123V in NS1 (Sun et al. 2014). The V106I and N248D substitutions enhance viral stability at low pH, which confers replicative fitness and likely promotes virus spread (Takahashi et al. 2013). Notably, through the N248D change, the protein acquires a negative charge (at neutral pH) that is solvent-exposed ($RSA = 0.56$) and located within < 12 Å vicinity to residues

belonging to the three-dimensional structure of the active center of NA (Colman et al. 1993; Abed et al. 2006) (Figure 5B).

We detected a second selective sweep in season 2011N corresponding to the third wave of the pandemic (Elderfield et al. 2014). We identified seven accompanying changes in PB2, PA, HA, NA and M1 (Figure 2A, 2C, 2D, 2F and 2G, Supplementary Table 1). The changes V244M and I354L (PB2), N321K (PA), S185T (HA) and N369K (NA) came close to fixation in later seasons, with frequencies of around 90% in 2012N. Notably, their rise in frequency was slower in comparison to the changes from 2009S, which reached a frequency of nearly 100% only one season after they were detected. The change S185T in HA was reported to enhance receptor binding avidity (de Vries et al. 2013), while N321K in PA increased polymerase activity (Elderfield et al. 2014). This residue is exposed on the structure ($RSA = 0.41$) and close to M336 (Figure 5C), which was reported to foster adaptation by increasing polymerase activity in mammalian cells (Gabriel et al. 2005). Furthermore, the amino acid changes N369K and V241I in NA were described as permissive mutations which enhanced viral fitness in oseltamivir resistant viruses (Butler et al. 2014). As the amino acid change V241I decreased in frequency in 2013N (Figure 2F) and is buried in the protein ($RSA = 0$), this could indicate that site N396K (exposed, $RSA = 0.1$; uncharged to positively charged) might be the key amino acid change that confers antiviral resistance.

After the third global wave in 2011N, the pH1N1 influenza virus continued to circulate in the human population with lower activity throughout 2012 in a wide range of countries, especially in Europe and Eastern Asia (He et al. 2015). The SD plots indicated a selective sweep in 2012N accompanied by eleven changes in the proteins PB1, PA, HA, NA, M2 and NS1 (Figure 2B, 2C, 2D, 2F, 2H and 2I, Supplementary Table 1). Eight of these changes (I435T in PB1 and all changes in HA, NA and M2) decreased in frequency right after this selective sweep and disappeared completely in 2013S. The A343T change in PA also was lost a year later in 2014S. This behavior is striking in comparison to the previously detected sweeps in 2009S and 2011N, where a majority of sweep-related changes became fixed (except

V241I (NA) and V80I (M1)). For the selective sweep in 2012N, only the I397M amino acid change in PB1 and the L90I amino acid change in NS1 continued to be present in the circulating viral population and became fixed in 2013S (Figure 2B and 2I), suggesting their relevance for the sweep while the remaining changes could have been hitchhikers.

In 2013N, the pH1N1 influenza virus re-emerged with more activity in the countries where activity was low in 2012 (He et al. 2015). Here the SD plots revealed twenty-two sweep-related changes in all proteins, except for PA (Figure 2, Supplementary Table 1). PB2 and NA had the most changes, with five amino acid changes each rising rapidly in frequency and coming close to fixation within one season. Seventeen of these twenty-two amino acid changes were still fixed at the end of 2015S, suggesting that some of these changes provide a selective advantage. The sweep-related change V731I (PB2) is located in the vicinity (9 Å) of residue 701 (Figure 5D), a position known to be involved in mammalian host adaptation (Gabriel et al. 2008; Gabriel et al. 2011). Further sweep-related changes not rising to fixation occurred in HA (V234I and K283E), M1 (V80I and M192V) and M2 (V21G).

We detected a sweep in 2013S with sixteen sweep-related changes in nine proteins, with six changes in PB2 (R251K), PB1 (K52R), PA (P560H), NP (E454D and N377S) and NS1 (T80A) decreasing quickly in frequency and showing frequencies close to 0% one season after their emergence. The other sweep-related changes were still fixed in the viral population in 2015S and thus could be of functional relevance, including V100I, I330V, K361R and R362K in PA, K163Q, K283E and A256T in HA and I321V in NA (Figure 2C, 2D and 2F). Especially the changes I330V, K361R and R362K in line with N321K (season 2011N) cluster significantly around the mammalian-adaptation site 336 on the structure of PA that increases polymerase activity (Figure 5C) (Bussey et al. 2011). This might suggest that these PA changes might also affect viral polymerase activity, as a hallmark of influenza disease severity in humans. We tested for enrichment of fixed sweep-related sites within a radius of < 13 Å in the vicinity of site 336 (hypergeometric distribution; $N = 716, K = 31, n = 6, k = 4; H_0$: fixed sweep-related changes occur

16

with the same probability in vicinity to site 336 and in the remaining protein region; $P = 4 \times 10^{-5}$; N being all sites in the protein model, K the median number of sites in a radius of 13 Å around a site in the protein model, n all detected sweep-related changes which remained fixed, k the number of fixed sweep related changes in the cluster). The change K283E in HA (site 283 is exposed; $RSA = 0.41$) results in a change of charge from positive to negative and is located in the stalk region that plays a key role in the induction of neutralizing antibodies (Laursen and Wilson 2013).

The 2014N and 2014S seasons showed sweep-related changes only for a small number of proteins. In 2014N, two changes were detected in both NA and NS1 (Figure 2F, 2I), which all rose to fixation and still circulated in 2015S. In 2014S, changes in the proteins PA, NP and M1 were identified subsequently. While I387V and S522G in PA disappeared in 2015S, M105T and A22T in NP were maintained at frequencies of around 40% (Supplementary Table 1). The V80I and M192V changes in M1 were already detected in 2013N, but disappeared due to back mutations and re-emerged, indicative of a functional relevance.

Similar to 2013S, many sweep-related changes detected in 2015N decreased in frequency in 2015S. Exceptions are L40I and N386K in NA, which remained close to fixation in 2015N and seem most likely to provide a selective advantage. Six sweep-related changes in PB2 (A184T), PB1 (R211K), NA (I117M), NP (V425I, V444I), M2 (G21V) decreased in frequency but still remained predominant, with frequencies around 50% or higher in 2015S (Figure 2A, 2B, 2D, 2E and 2H, Supplementary Table 1).

Selective sweeps in the evolution of sH3N2 influenza

To investigate the past evolution of the seasonal sH3N2 influenza virus, we analyzed sequence data of the 34 seasons from 1999N to 2015S. The sH3N2 influenza virus has been circulating in the human population since 1968, but before 1999N sequence data only few data are available. In the SD plots

analysis we therefore focused on more recent years, ensuring a sufficient data coverage for each season (Figure 3, Supplementary Table 2).

Overall, we identified fifteen seasons in the analyzed time period with sweep-related changes in the HA protein. For each detected selective sweep, we investigated whether it indicated the emergence of an antigenically novel strain that became predominant by matching the sweep-related changes to the amino acid changes that were reported for the viral strain by the WHO. If not stated directly for the predominant strain, the changes reported for the subclade including the predominant strain were used for matching to the WHO reported strains, as they distinguish a novel viral variant from previous ones. We further assessed the coherence between detected selective sweeps and novel antigenic variants by comparing the sweep-related sites to known antigenicity- and avidity-changing sites. We recently reported twenty-three antigenicity-altering sites organized in five patches on the structure of HA playing a role in the past antigenic evolution of sH3N2 viruses (Kratsch et al. 2016), a subset of which together with position 193 were also reported by (Koel et al. 2013). Of the sites influencing the receptor binding properties, i.e. avidity changing, we focused on the sites 193, 222 and 225 (Lin et al. 2012), as well as 145 (Li et al. 2013), which is also an antigenicity-altering site (Kratsch et al. 2016), bringing the set of considered sites to twenty-six.

Of the fifteen seasons in which sweep-related changes were detected in the HA protein, eleven included changes at these twenty-six sites (Figure 4). Of the eleven sweeps with relevant changes, three matched and indicated the emergence of a novel antigenic variant one season before the strain became predominant (A/Fujian/411/2002 strain, A/California/7/2004 strain and A/Perth/16/2009 strain) (Figure 4). In two cases, a newly emerging strain was matched and detected two seasons before predominance (A/Wisconsin/67/2005 and A/Brisbane/10/2007) (Figure 4). This is ideal for vaccine strain selection, as new strains to be included in the vaccine formulation are decided on one year in advance. Three cases matched and indicated a newly emerging strain in the season in which it was predominant

18

(A/Victoria/208/2009, A/Texas/50/2012 and A/Hong Kong/4801/2014). Interestingly, for two strains (A/Brisbane/10/2007 and A/Hong Kong/4801/2014) we identified the associated sweep-related changes over two consecutive seasons. Only the appearance of the A/Victoria/361/2011 strain coincided with a sweep without changes at known antigenicity or avidity sites (Figure 4), which indicates that our list of antigenicity or avidity altering sites likely is not complete. The four sweeps without changes in antigenically- and avidity-altering sites did not include changes observed for antigenically novel predominant strains reported by the WHO.

In comparison to the strain recommendations by the WHO, sweeps predicted with the SD plots and including changes at antigenicity- or avidity altering sites identified newly emerging strains at the same time in three cases (e.g. A/Fujian/411/2002) or up to two seasons before in six cases (e. g. A/Brisbane/10/2007; Figure 4). Notably, the SD plots analysis did not identify sweeps corresponding to A/Wellington/1/2004 and A/Switzerland/9715293/2013, which were false positive vaccine strain recommendations that did not match the dominantly circulating strain (Who 2005a; Steinbruck et al. 2014; Who 2015a). In some cases practical issues such as the lack of a vaccine strain with sufficient growth properties in eggs, which might have prevented recommendation of a suitable vaccine strain candidate by the WHO. SD plot analysis indicated suitable strains in the majority of cases using a fixed computational procedure, and therefore seems suitable for supporting the vaccine strain decisions of the WHO experts. Notably, different from the established WHO procedure, no interpretation of hemagglutination inhibition assay data was required for this result. This could be an additional benefit, as recently, HI assays fail to agglutinate many circulating sH3N2 viruses (Ampofo et al. 2013).

Discussion

We describe a new technique called SD plots that combines phylogenetics with a statistical analysis to detect selective sweeps and associated individual amino acid changes under directional selection from longitudinal samples of population-level sequence data. Due to the rapid evolution of influenza viruses, backmutations or repeated mutations at individual sites are common in certain protein regions, requiring elucidation of the evolutionary histories of the proteins for statistical analysis, to allow detection and consideration of such effects. To our knowledge, the method is the first to allow the suggestion of individual changes under directional selection from viral sequence data with statistical support, resulting in the suggestion of a clearly defined set of candidate changes for subsequent experimental studies. A characterization of the evolution of the two currently circulating human influenza A viruses, pH1N1 and sH3N2, demonstrated that the technique revealed most known changes of relevance for adaptation to the human host since the establishment of pH1N1 in 2009, or to host immune response in case of sH3N2. Several further candidate changes for experimental studies were identified for pH1N1. We demonstrate that SD plots enable the detection of newly emerging antigenic variants for the sH3N2 viruses, based on which suggested updates of the vaccine strain resulted in a slightly better match to future predominant strains than the suggestions provided by the WHO. SD plots could also be used to study the evolutionary dynamics of other rapidly evolving populations from longitudinal sequence samples, such as clinical intra-patient HIV data.

Acknowledgements

References

Abed Y, Baz M, Boivin G. 2006. Impact of neuraminidase mutations conferring influenza resistance to neuraminidase inhibitors in the N1 and N2 genetic backgrounds. Antivir Ther 11:971-976.

Ampofo WK, Al Busaidy S, Cox NJ, Giovanni M, Hay A, Huang S, Inglis S, Katz J, Mokhtari-Azad T, Peiris M, et al. 2013. Strengthening the influenza vaccine virus selection and development process: outcome of the 2nd WHO Informal Consultation for Improving Influenza Vaccine Virus Selection held at the Centre International de Conferences (CICG) Geneva, Switzerland, 7 to 9 December 2011. Vaccine 31:3209-3221.

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. Journal of virology 82:596-601.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57:289-300.

Bhatt S, Holmes EC, Pybus OG. 2011. The genomic Rate of Molecular Adaptation of the Human Influenza A Virus. Mol Biol Evol:1-9.

Bouvier NM, Palese P. 2008. The biology of influenza viruses. Vaccine 26.

Burke DF, Smith DJ. 2014. A Recommended Numbering Scheme for Influenza A HA Subtypes. PLoS ONE 9.

Bussey KA, Desmet EA, Mattiacio JL, Hamilton A, Bradel-Tretheway B, Bussey HE, Kim B, Dewhurst S, Takimoto T. 2011. PA residues in the 2009 H1N1 pandemic influenza virus enhance avian influenza virus polymerase activity in mammalian cells. J Virol 85:7020-7028.

Butler J, Hooper KA, Petrie S, Lee R, Maurer-Stroh S, Reh L, Guarnaccia T, Baas C, Xue L, Vitesnik S, et al. 2014. Estimating the Fitness Advantage Conferred by Permissive Neuraminidase Mutations in Recent Oseltamivir-Resistant A(H1N1)pdm09 Influenza Viruses. PLoS Pathogens 10.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195-205.

Chothia C. 1976. The nature of the accessible and buried surfaces in proteins. J Mol Biol 105:1-12.

Colman PM, Hoyne PA, Lawrence MC. 1993. Sequence and structure alignment of paramyxovirus hemagglutinin-neuraminidase with influenza virus neuraminidase. J Virol 67:2972-2980.

de Vries RP, de Vries E, Martínez-Romero C, McBride R, van Kuppeveld FJ, Rottier PJM, García-Sastre A, Paulson JC, de Haan CaM. 2013. Evolution of the hemagglutinin protein of the new pandemic H1N1 influenza virus: maintaining optimal receptor binding by compensatory substitutions. Journal of virology 87:13868-13877.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Elderfield Ra, Watson SJ, Godlee a, Adamson WE, Thompson CI, Dunning J, Fernandez-Alonso M, Blumenkrantz D, Hussell T, Zambon M, et al. 2014. Accumulation of Human-Adapting Mutations during Circulation of A(H1N1)pdm09 Influenza Virus in Humans in the United Kingdom. Journal of Virology 88:13269-13283.

Fisher RA. 1922a. On the dominance ratio. Proceedings of the Royal Society of Edinburgh 42:321-341.

Fisher RA. 1922b. On the Interpretation of χ2 from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society 85:87-94.

Fitch WM. 1971. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Systematic Biology 20 406-416.

Gabriel G, Dauber B, Wolff T, Planz O, Klenk HD, Stech J. 2005. The viral polymerase mediates adaptation of an avian influenza virus to a mammalian host. Proc Natl Acad Sci U S A 102:18590-18595.

Gabriel G, Herwig A, Klenk HD. 2008. Interaction of polymerase subunit PB2 and NP with importin alpha1 is a determinant of host range of influenza A virus. PLoS Pathog 4:e11.

Gabriel G, Klingel K, Otte A, Thiele S, Hudjetz B, Arman-Kalcek G, Sauter M, Shmidt T, Rother F, Baumgarte S, et al. 2011. Differential use of importin-alpha isoforms governs cell tropism and host adaptation of influenza virus. Nat Commun 2:156.

Hay AJ, Daniels R, Lin YP, Zheng X, Hou T, Gregory V, Whittaker L, Kloess J, Cattle N. 2009. Interim Report September 2009. In. London: WHO Collaborating Centre for Reference and Research on Influenza, National Institute for Medical Research.

He D, Lui R, Wang L, Tse CK, Yang L, Stone L. 2015. Global Spatio-temporal Patterns of Influenza in the Post-pandemic Era. Scientific reports 5:11013.

Hein J, Schierup MH, Wiuf C. 2005. Gene genealogies, variation and evolution: a primer in coalescent theory: Oxford University Press, USA.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. 10:845-858.

Khiabanian H, Trifonov V, Rabadan R. 2009. Reassortment patterns in Swine influenza viruses. PLoS Curr 1:RRN1008.

Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513-1524.

Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GC, Vervaet G, Skepner E, Lewis NS, Spronken MI, Russell CA, et al. 2013. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science 342:976-979.

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.

Kratsch C, Klingen TR, Mümken L, Steinbrück L, McHardy AC. 2016. Determination of antigenicity-altering patches on the major surface protein of human influenza A/H3N2 viruses. Virus Evolution 2.

Kryazhimskiy S, Plotkin JB. 2008a. The population genetics of dN/dS. PLoS Genetics 4.

Kryazhimskiy S, Plotkin JB. 2008b. The population genetics of dN/dS. PLoS Genet 4:e1000304-e1000304.

Laursen NS, Wilson IA. 2013. Broadly neutralizing antibodies against influenza viruses. Antiviral Res 98:476-483.

Lee B, Richards FM. 1971. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55:379-400.

Li Y, Bostick DL, Sullivan CB, Myers JL, Griesemer SB, Stgeorge K, Plotkin JB, Hensley SE. 2013. Single hemagglutinin mutations that alter both antigenicity and receptor binding avidity influence influenza virus antigenic clustering. J Virol 87:9904-9910.

Lin YP, Xiong X, Wharton SA, Martin SR, Coombs PJ, Vachieri SG, Christodoulou E, Walker PA, Liu J, Skehel JJ, et al. 2012. Evolution of the receptor binding properties of the influenza A(H3N2) hemagglutinin. Proc Natl Acad Sci USA 109:21474-21479.

McHardy AC, Adams B. 2009. The role of genomics in tracking the evolution of influenza A virus. In. PLoS Pathogens.

Medina RA, García-Sastre A. 2011. Influenza A viruses: new research developments. Nature reviews. Microbiology 9:590-603.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular biology and evolution 3:418-426.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. Genome Res 15:1566-1575.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Otte A, Marriott AC, Dreier C, Dove B, Mooren K, Klingen TR, Sauter M, Thompson KA, Bennett A, Klingel K, et al. 2016. Evolution of 2009 H1N1 influenza viruses during the pandemic correlates with increased viral pathogenicity and transmissibility in the ferret model. Sci Rep 6:28583.

Otte A, Sauter M, Daxer MA, McHardy AC, Klingel K, Gabriel G. 2015. Adaptive Mutations That Occurred during Circulation in Humans of H1N1 Influenza Virus in the 2009 Pandemic Enhance Virulence in Mice. Journal of virology 89:7329-7337.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS ONE 5.

Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. 2008. The genomic and epidemiological dynamics of human influenza A virus. Nature 453:615-619.

Reich S, Guilligay D, Pflug A, Malet H, Berger I, Crepin T, Hart D, Lunardi T, Nanao M, Ruigrok RW, et al. 2014. Structural insight into cap-snatching and RNA synthesis by influenza polymerase. Nature 516:361-366.

Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, Gust ID, Hampson AW, Hay AJ, Hurt AC, et al. 2008. Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. Vaccine 26 Suppl 4:D31-34.

Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779-815.

Shinya K, Ebina M, Yamada S, Ono M, Kasai N, Kawaoka Y. 2006. Avian flu: influenza virus receptors in the human airway. Nature 440:435-436.

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.

Steinbruck L, Klingen TR, McHardy AC. 2014. Computational prediction of vaccine strains for human influenza A (H3N2) viruses. J Virol 88:12123-12132.

Steinbrück L, McHardy AC. 2011. Allele dynamics plots for the study of evolutionary dynamics in viral populations. Nucleic Acids Res 39:e4-e4.

Sun Y, Xu Q, Shen Y, Liu L, Wei K, Sun H, Pu J, Chang K-C, Liu J. 2014. Naturally occurring mutations in the PA gene are key contributors to increased virulence of pandemic H1N1/09 influenza virus in mice. Journal of virology 88:4600-4604.

Takahashi T, Song J, Suzuki T, Kawaoka Y. 2013. Mutations in NA That Induced Low pH-Stability and Enhanced the Replication of Pandemic (H1N1) 2009 Influenza A Virus at an Early Stage of the Pandemic. PLoS ONE 8.

Templeton AR. 2006. Population genetics and microevolutionary theory: Wiley Online Library.

Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum allowed solvent accessibilites of residues in proteins. PLoS One 8:e80635.

van Riel D, Munster VJ, de Wit E, Rimmelzwaan GF, Fouchier RA, Osterhaus AD, Kuiken T. 2006. H5N1 Virus Attachment to Lower Respiratory Tract. Science 312:399.

Varich NL, Gitelman AK, Shilov AA, Smirnov YA, Kaverin NV. 2008. Deviation from the random distribution pattern of influenza A virus gene segments in reassortants produced under non-selective conditions. Arch Virol 153:1149-1154.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. Annu Rev Genet 47:97-120.

Webb B, Sali A. 2014. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics 47:5.6.1-32.

Fact sheet number 211 [Internet]. http://www.who.int/mediacentre/factsheets/fs211/en/2014a 16 July 2016].

WHO. 2011a. Recommended composition of influenza vaccines for use in the 2012 southern hemisphere influenza season. WHO Weekly Epidemiological Record 86:457-468.

WHO. 1999. Recommended composition of influenza virus vaccines for use in 2000. WHO Weekly Epidemiological Record 74:321-328.

WHO. 2003. Recommended composition of influenza virus vaccines for use in the 2004 influenza season. WHO Weeky Epidemiological Record 78:375-379.

Who. 2005a. Recommended composition of influenza virus vaccines for use in the 2005–2006 influenza season. WHO Weekly Epidemiological Record 80:66-71.

WHO. 2005b. Recommended composition of influenza virus vaccines for use in the 2005–2006 influenza season. WHO Weeky Epidemiological Record 80:65-76.

WHO. 2004. Recommended composition of influenza virus vaccines for use in the 2005 influenza season. WHO Weeky Epidemiological Record 79:369-376.

WHO. 2006. Recommended composition of influenza virus vaccines for use in the 2006–2007 influenza season. WHO Weeky Epidemiological Record 81:81-88.

WHO. 2007. Recommended composition of influenza virus vaccines for use in the 2008 influenza season. WHO Weeky Epidemiological Record 82:345-356.

WHO. 2009. Recommended composition of influenza virus vaccines for use in the 2010 influenza season. WHO Weeky Epidemiological Record 84:421-436.

WHO. 2011b. Recommended composition of influenza virus vaccines for use in the 2011-2012 northern hemisphere influenza season. WHO Weekly Epidemiological Record 86:81-91.

WHO. 2012. Recommended composition of influenza virus vaccines for use in the 2012–2013 northern hemisphere influenza season. WHO Weeky Epidemiological Record 87:81-96.

WHO. 2013a. Recommended composition of influenza virus vaccines for use in the 2013–2014 northern hemisphere influenza season. WHO Weeky Epidemiological Record 88:101-116.

Who. 2014b. Recommended composition of influenza virus vaccines for use in the 2014-2015 northern hemisphere influenza season. WHO Weekly Epidemiological Record 89:93-104.

WHO. 2013b. Recommended composition of influenza virus vaccines for use in the 2014 southern hemisphere influenza season. WHO Weekly Epidemiological Record 88:437-448.

Who. 2015a. Recommended composition of influenza virus vaccines for use in the 2015-2016 northern hemisphere influenza season. Wkly Epidemiol Rec 90:97-108.

WHO. 2014c. Recommended composition of influenza virus vaccines for use in the 2015 southern hemisphere influenza season. WHO Weeky Epidemiological Record 89:441-456.

WHO. 2015b. Recommended composition of influenza virus vaccines for use in the 2016 southern hemisphere influenza season. WHO Weeky Epidemiological Record 90:545-560.

WHO. 2010. Recommended viruses for influenza vaccines for use in the 2011 influenza season (southern hemisphere). WHO Weeky Epidemiological Record 85:402-412.

Wille M, Tolf C, Avril A, Latorre-Margalef N, Wallerstrom S, Olsen B, Waldenstrom J. 2013. Frequency and patterns of reassortment in natural influenza A virus infection in a reservoir host. Virology 443:150-160.

Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AG, McCoy A, et al. 2011. Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr 67:235-242.

Wise HM, Hutchinson EC, Jagger BW, Stuart AD, Kang ZH, Robb N, Schwartzman LM, Kash JC, Fodor E, Firth AE, et al. 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.

Wright S. 1931. Evolution in Mendelian Populations. Genetics 16:97-159.

Wright S. 1933. Inbreeding and Homozygosis. Proc Natl Acad Sci USA 19:411-420.

Wright S. 1938. Size of population and breeding structure in relation to evolution. Science 87:430-431.

Yang Z, Bielawski JR. 2000. Statistical methods for detecting molecular adaptation. In. Trends in Ecology and Evolution. p. 496-503.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular biology and evolution 19:908-917.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular biology and evolution 17:32-43.

Figure Legends

Figure 1: Phylogeny demonstrating the frequency correction, the corresponding SD plot and frequencies of amino acid changes.

This is a detailed illustration of how the frequency correction handles individual amino acid changes that occur at the same position. The tree T **(A)** consists of five main subtrees sampled from three seasons (colored in green, orange and purple). Each season s exhibits forty isolates ($\left|L_{T,s}\right| = 40 \ \forall s$). A_T contains four amino acid changes that occur on different internal branches: S1T (red), T1S (blue), S1A (pink) and S1T (light green). We consider S1T (red) and S1T (light green) as two separate amino acid changes with individual frequencies as they occur on different branches of the tree. Consequently, both are calculated separately and occur in the SD plot **(B)**. Amino acid changes S1T (red) and S1A (pink) rise in frequency in the second season and disappear in the last season while T1S (blue) and S1T (light green) arise in the third season in the corresponding SD plot. Note that the frequencies for change S1T (red) and for T1S (blue) are corrected while the frequencies for S1A (pink) and S1T (light green) do not require a correction **(C)**.

Note that this is an artificial example and does not necessarily reflect a parsimonious ancestral character state reconstruction for position one.

Figure 2: Sweep Dynamics (SD) plots for ten pandemic H1N1 influenza A proteins.

These plots show the alterations in frequency of sweep-related changes over consecutive time periods in the proteins PB2, PB1, PA, HA, NP, NA, M1, M2, NS1 and NS2. They are ordered from **(A) – (K)** based on decreasing segment size on which they are encoded. Each plot shows the seasons (from 2009S to 2015S) on the x-axis and the frequency on the y-axis. The initial emergence of a sweep-related amino acid change is indicated by an asterisk pinpointing the first season in which it significantly rises in frequency to a frequency of more than 50%. Sweep-related changes are listed in a panel above the graph

26

and are identically color coded as the curve in the plot. Within the panel, they are bottom-up ordered with ascending frequency.

Figure 3: Sweep Dynamics (SD) plot for the seasonal H3N2 influenza A hemagglutinin.

This plot shows alterations in frequency of sweep-related changes over consecutive time periods for the influenza A/H3N2 hemagglutinin. It depicts the seasons (from 1999S to 2015S) on the x-axis and the frequency on the y-axis. Sweep-related changes are listed in a panel above the graph and are identically color coded as the curve in the plot. Within the panel, they are bottom-up ordered with ascending frequency.

Figure 4: Comparison of predominant seasonal H3N2 influenza A strains, SD plots results and recommendations made by the WHO.

From 1999N until 2015S, eleven antigenically different sH3N2 influenza strains were selected for production of the seasonal influenza vaccine, named Moscow/10/1999, A/Fujian/411/2002, A/Wellington/1/2004, A/California/7/2004, A/Wisconsin/67/2005, A/Brisbane/10/2007, A/Perth/16/2009, A/Victoria/361/2011, A/Texas/50/2012, A/Switzerland/9715293/2013 and A/Hong Kong/4801/2014 (WHO 2003, 2004, 2005b, 2006, 2007, 2009, 2012, 2013a, 2014c, 2015b), which are indicated by colors as labeled in the legend. For the SD plots analysis, seasons are marked if sweep-related changes distinguish the respective vaccine strain from the previous one. False positive or false negative results are marked in grey. The selection of a vaccine strain takes place two seasons before the vaccine is available. Any prediction of newly arising antigenically novel strains should therefore not be compared to the current predominant strain, but to the predominant strain two seasons after the detection. This comparison of the WHO selections and the SD plots results with the predominant strain two seasons later is indicated by the diagonal lines in the upper part of the plot. [a]The strain A/Victoria/208/2009 was predominant in 10S, 11N and 11S but was antigenically indistinguishable from A/Perth/16/2009 (WHO 2010, 2011a, b).

27

[b]The strains A/Victoria/361/2011 and A/Texas/50/2012 were predominant in 13N, 13S and 14N but A/Texas/50/2012 was antigenically more effective (WHO 2013a, b, 2014b). [c]In 14S, two different clades containing the strains A/Switzerland/9715293/2013 and A/Hong Kong/4801/2014, respectively, were rising in frequency at the same time and antigenically differed from the previous vaccine strain A/Texas/50/2012 (WHO 2014c).

Figure 5: Representation of sweep-related changes on protein structures.

We mapped individual sweep-related changes on the corresponding models shown as cartoon with interesting residues shown as sticks and colored by atom-type (yellow and orange for carbon, red for oxygen and blue for nitrogen). In **(A)**, the structure of NP is shown in blue with sweep-related mutated residues T373 and I100 colored orange and mammalian-adaptation site N319 yellow. The distance between Cα-atoms of T373 and N319 is 10.2 Å (dotted line). The structure in **(B)** depicts NA (green) with residue N248 (orange) in proximity (<12 Å) to the active center (yellow) (Takahashi et al. 2013). In **(C)** and **(D)** we show views on the structural model of the polymerase with subunits PB2 (pink), PB1 (light green) and PA (light blue). Sweep-related changes at residues N321, I330, K361 and R362 (orange) in PA cluster around M336 (yellow) that increases polymerase activity in mammalian cells **(C)** (Gabriel et al. 2005). The structure of polymerase subunit PB2 (pink) is shown in **(D)** and highlights the sweep-related mutation V731 (orange), which is in close vicinity (≤9 Å) to position 701 that enhances mammalian host adaptation (Gabriel et al. 2008; Gabriel et al. 2011).

Figures

Figure 1: Phylogeny demonstrating the frequency correction, the corresponding SD plot and frequencies of amino acid changes.
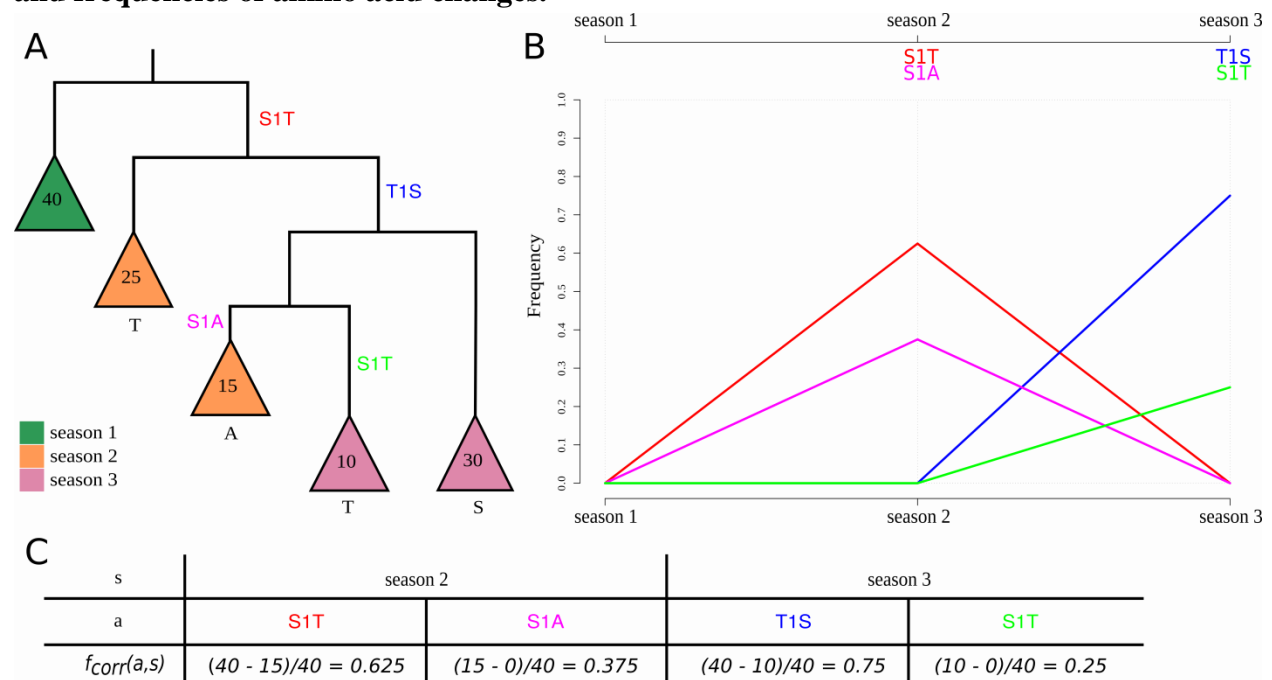
Figure 2: Sweep Dynamics (SD) plots for ten pandemic H1N1 influenza A proteins.
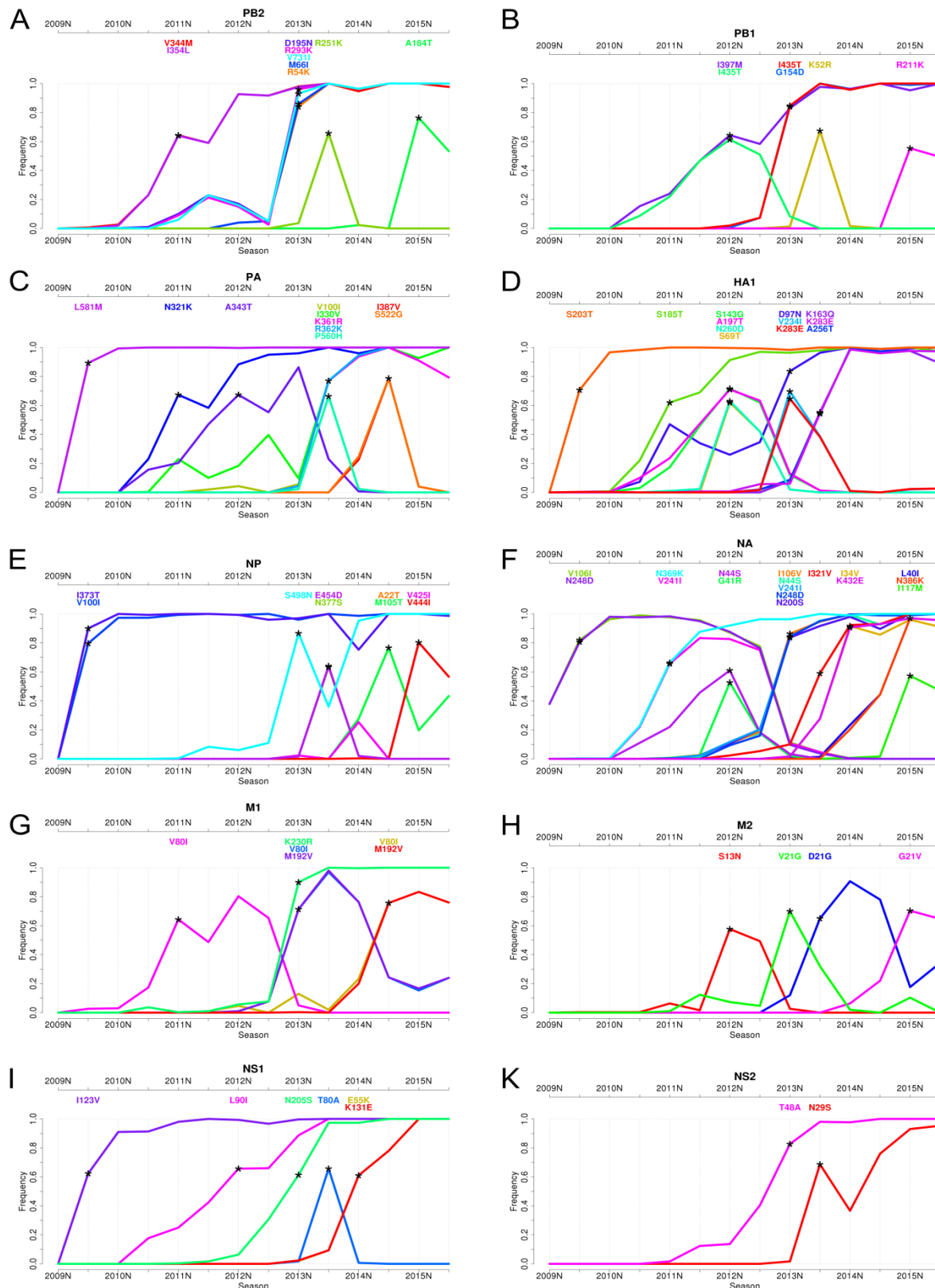
Figure 3: Sweep Dynamics (SD) plot for the seasonal H3N2 influenza A hemagglutinin.
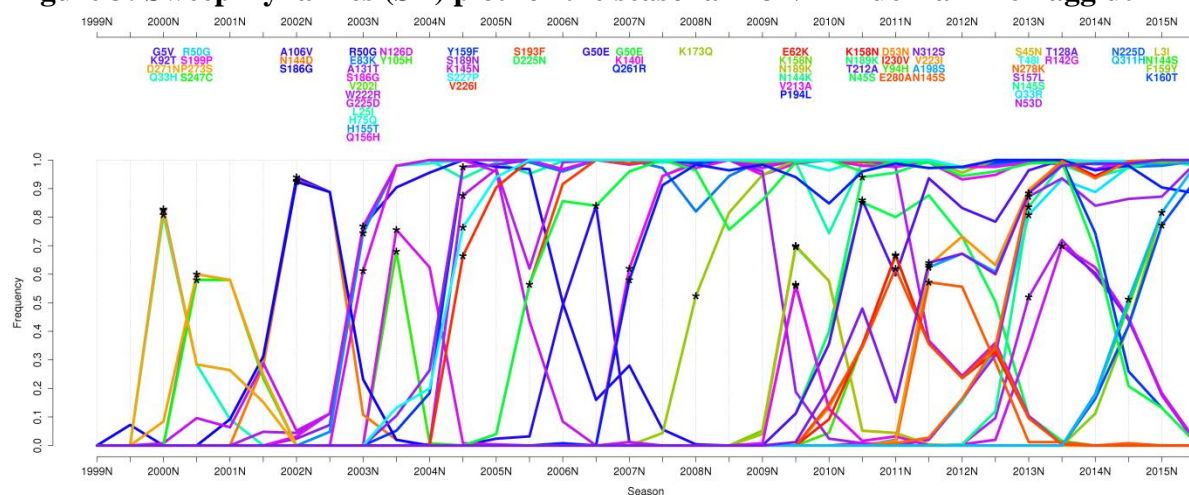

Figure 4: Comparison of predominant seasonal H3N2 influenza A strains, SD plots results and recommendations made by the WHO.
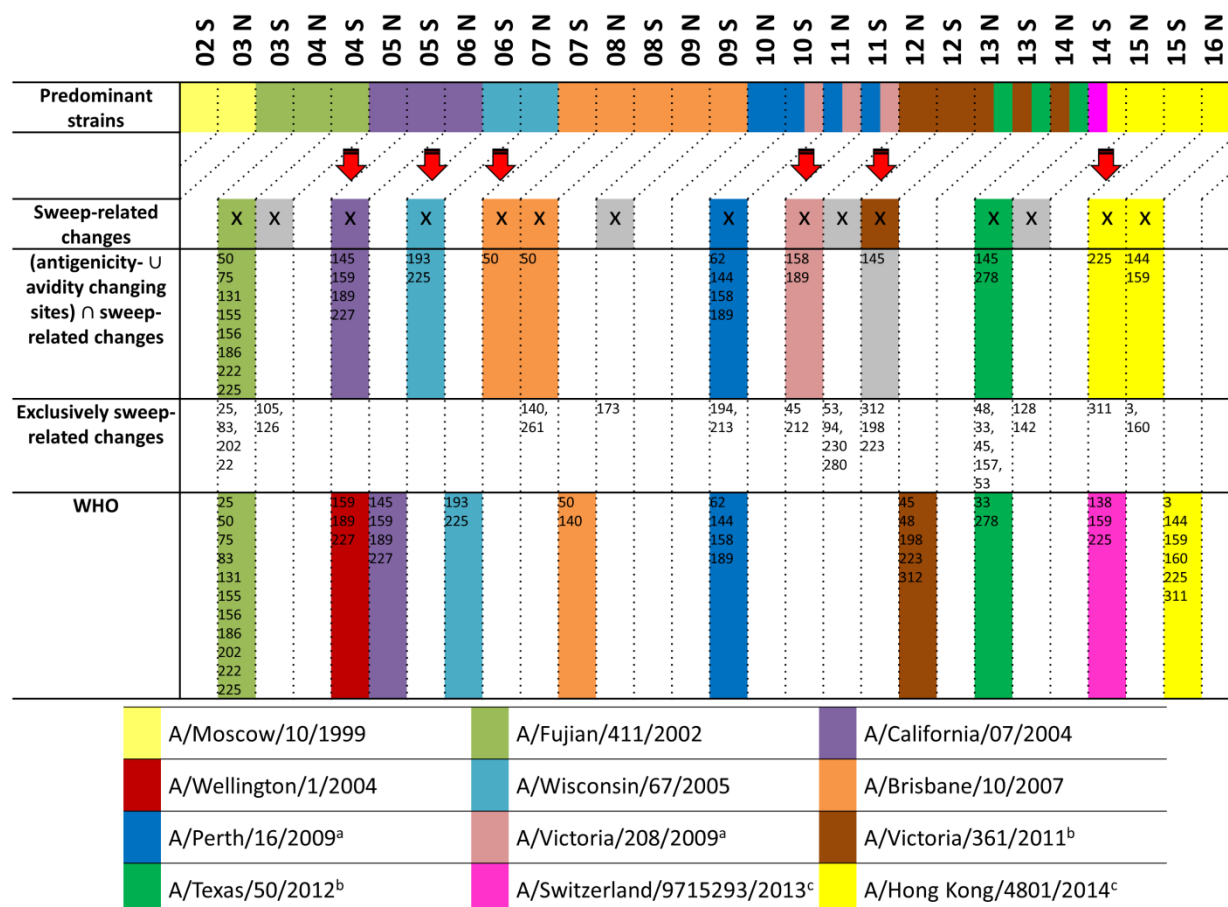
Figure 5: Representation of sweep-related changes on protein structures.