Computational Methods for the Study of Influenza A Virus Phylodynamics

Kumulative Dissertation

zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Lars Steinbrück
aus Weida

Düsseldorf, Juli 2012

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathemathisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent:	Prof. Dr. Alice C. McHardy
Koreferent:	Prof. Dr. Martin J. Lercher
Koreferent:	Prof. Dr. David A. Liberles
Tag der mündlichen Prüfung:	04.12.2012

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation eigenständig und ohne fremde Hilfe angefertig habe. Arbeiten Dritter wurden entsprechend zitiert. Diese Dissertation wurde bisher in dieser oder ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den ....

(Lars Steinbrück)

Statement of authorship

I hereby certify that this dissertation is the result of my own work. No other person's work has been used without due acknowledgement. This dissertation has not been submitted in the same or similar form to other institutions. I have not previously failed a doctoral examination procedure.

Summary

Influenza is a contagious respiratory viral infection that has been endemic in humans for centuries causing substantial morbidity and mortality. Despite of comprehensive vaccination campaigns influenza is annually responsible for approximately 41,000 deaths in the USA alone and, thus, results in an enormous health and economy burden. Three distinct types are endemic in humans with type A viruses evolving most rapidly and being commonly associated with the most influenza infections. In a process known as antigenic drift, the virus continuously alters the sequence composition of the two surface antigens, hemagglutinin and neuraminidase, to evade recognition by the host immune system. Therefore, the composition of the influenza vaccine has to be updated on a regular basis. Based on this, the correct and accurately timed identification of strains that are on the rise to predominance is of utmost importance to ensure sufficient vaccine efficiency.

The aim of this work was to develop computational methods for the analysis of the phylodynamics of seasonal influenza A viruses. This means that the developed methods should allow the analysis of the genetic, antigenic and epidemiological dynamics of influenza and give insights into how their interplay shapes the evolution of the virus. Two different strategies were developed to tackle the problem from (i) a population genetics point and (ii) a molecular genetics point.

The rapid evolution of influenza A viruses results in the evolutionary processes to proceed on a similar scale to the epidemiological processes. This allows for a joint analysis of the genetic and spatiotemporal dynamics of the virus. In this context, we developed allele dynamics plots (AD plots), for the visualization of the evolutionary dynamics of a gene in a population. Based on a sample of dated genetic sequences AD plots visualize gene alleles, i.e. non-empty sets of amino acid changes mapped to individual branches of a phylogeny, and their frequency over time. The method's merits are demonstrated with a 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. Furthermore, AD plots enable the identification of those alleles that are likely to be subject to directional selection. Identification of alleles with the largest frequency increase between consecutive influenza seasons resulted in the early detection of those influenza A (H3N2) virus strains that later rise to predominance.

A selective advantage of individual alleles implies their novelty in the antigenic phenotype with respect to alleles already circulating at high frequency. However, not all substitutions associated with an allele contribute equally to the change in antigenicity. In this sense, distinguishing substitutions in the hemagglutinin of human influenza A viruses that have a significant impact on the antigenic phenotype from (near-) neutral 'hitchhikers' is of high relevance. We therefore developed a method that allows for the inference of antigenic trees for the major viral surface protein hemagglutinin. Antigenic trees enable the determination of antigenic branch lengths for a given tree topology using least-squares optimization. Thus, it allows to resolve the antigenic impact of branch-associated amino acid changes. The accuracy of our technique to predict antigenic distances is comparable to antigenic cartography. However, the inference of antigenic trees allows for a more detailed study of the antigenic evolution of influenza A (H3N2) viruses. Besides the identification of antigenic types, i.e. groups of viruses with similar genetic and antigenic properties, we identified seven sites and five amino acid changes with high antigenic impact in the evolution of influenza A (H3N2) viruses from 1968 to 2003.

In summary, the developed methods are useful tools for the analysis of the phylodynamics of influenza A viruses with a potential application for the biannual vaccine strain selection process. However, application is not limited to this pathogen. With AD plots any organism/pathogen where homologous genetic sequence data and associated sampling times are available can be analyzed. For phenotype trees, application is possible if pair-wise phenotype distances and according homologous genetic sequence data are available. Thus, the developed methods have value for a broad scientific community.

Zusammenfassung

Influenza ist eine ansteckende virale Erkrankung der Atemwege, die seit Jahrhunderten für beträchtliche Morbidität und Mortaliät in der menschlichen Bevölkerung verantwortlich ist. Allein in den USA verursacht das Influenzavirus jährlich trotz weitreichender Impfstrategien etwa 41.000 Todesfälle. Damit stellt Influenza eine erhebliche Belastung für das Gesundheitswesen und die Wirtschaft dar. Es zirkulieren drei verschiedene Typen in der menschlichen Bevölkerung, wobei sich Viren vom Typ A am schnellsten weiterentwickeln und gewöhnlich für die meisten Infektionen verantwortlich sind. Durch einen Prozess, der als "*antigenic drift*" bezeichnet wird, verändert das Virus fortlaufend die genetische Zusammensetzung seiner zwei Oberflächenproteine Hämagglutinin und Neuraminidase, um einer Erkennung und Neutralisierung durch das menschliche Immunsystem zu entkommen. Dementsprechend muss die Zusammensetzung des Influenzaimpfstoffes regelmäßig erneuert werden. In diesem Zusammenhang ist eine genaue und rechtzeitige Erkennung von viralen Stämmen, welche das Potential zur Dominanz in der viralen Population besitzen, von größter Bedeutung, um eine hinreichende Wirksamkeit des Impfstoffes zu gewährleisten.

Ziel dieser Arbeit war die Entwicklung von Methoden für die Analyse der "Phylodynamiken" von Influenza-A-Viren. Die Hauptaufgabe bestand somit darin, die evolutionären und epidemiologischen Vorgänge von Influenza zu untersuchen und darzustellen, wie deren Zusammenspiel die Evolution des Virus beeinflusst. Zwei verschiedene Strategien wurden entwickelt, welche diese Fragestellung von (i) einem populationsgenetischen Punkt und von (ii) einem molekulargenetischen Blickwinkel aus betrachten. Influenza-A-Viren zeichnen sich durch eine hohe Mutationsrate und Populationsgröße aus. Dadurch laufen die zugehörigen evolutionären Prozesse zeitlich auf einer ähnlichen Skala ab wie epidemiologische Prozesse. Diese Gegebenheit ermöglicht eine gemeinsame Analyse der genetischen und räumlich-zeitlichen Dynamik des Virus. Wir haben "*allele dynamics plots*" (AD-plots) entwickelt, welche die evolutionären Veränderungen eines Genes in einer Population grafisch darstellen. AD-plots visualisieren Gen-Allele und deren Häufigkeiten über die Zeit unter Verwendung von datierten genetischen Sequenzen. Allele sind in diesem Zusammenhang als nicht-leere Mengen von Aminosäureaustauschen definiert, die für einzelne Äste eines phylogenetischen Baumes inferiert wurden. Die Vorzüge dieser Methode wurden anhand einer Untersuchung der evolutionären Dynamiken saisonaler Influenza-A-Viren demonstriert. AD-plots des Hämagglutinin von saisonalen Influenza-A-Viren (Subtyp H3N2) und 2009 pandemischen Influenza-A-Viren (Subtyp H1N1) zeigen die Abfolge von Substitutionen, die in der Evolution

der zwei viralen Populationen fixiert wurden. Des Weiteren erlauben AD-plots die Identifizierung von Allelen, die wahrscheinlich einer gerichteten Selektion unterliegen. Die Identifizierung der Allele mit dem größten Häufigkeitanstieg zwischen aufeinanderfolgenden Saisons erlaubte die zeitige Erkennung solcher Influenza-A-Viren (Subtyp H3N2), die zu einem späteren Zeitpunkt dominant in der viralen Population wurden. Der selektive Vorteil individueller Allele impliziert eine signifikante Veränderung des antigenischen Phänotyps bezüglich anderer Allele, die bereits mit einer hohen Häufigkeit in der viralen Population zirkulieren. Allerdings tragen nicht alle Substitutionen, die mit einem Allel assoziiert sind, gleichermaßen zu der veränderten Antigenizität bei. Daher ist die Unterscheidung von Substitutionen, die einen signifikanten Einfluss auf den antigenischen Phänotyp haben, von (fast) neutralen "*hitchhicker*-Mutationen" von entscheidender Bedeutung. Wir haben eine Methode für die Inferenz von antigenischen Bäumen für Influenza-A-Viren (Subtyp H3N2) entwickelt. In antigenischen Bäumen werden paarweise antigenische Distanzen auf eine gegebene Baumtopologie mittels Optimierung der kleinsten Quadrate abgebildet. Dies ermöglicht die Inferenz von antigenischen Astlängen und somit die Aufdeckung des antigenischen Einflusses der Ast-assoziierten Aminosäureaustausche. Die Genauigkeit dieser Methode mit Bezug auf die Vorhersage antigenischer Distanzen ist vergleichbar mit derer von "*antigenic cartography*". Neben der Identifizierung antigenischer Typen, d.h. Gruppen von Viren mit ähnlichen genetischen und antigenischen Charakteristika, konnten wir sieben Positionen und fünf Aminosäureaustausche bestimmen, die einen großen antigenischen Einfluss in der Evolution von Influenza-A-Viren (Subtyp H3N2) zwischen 1968 und 2003 hatten. Zusammenfassend stellen die entwickelten Methoden nützliche Werkzeuge in der Analyse der "Phylodynamik" und der antigenischen Evolution von Influenza-A-Viren mit einer potentiellen Anwendungmöglichkeit für die halbjährlich stattfindende Auswahl geeigneter Impfstämme dar. Die Anwendung dieser Verfahren ist allerdings nicht auf diesen Erreger beschränkt. Im Prinzip kann jeder Organismus oder Erreger mittels AD-plots analysiert werden, für den homologe genetische Sequenzen und entsprechende Datierungsinformationen vorhanden sind. Phänotyp-Bäume können inferiert werden, wenn homologe genetische Sequenzen und paarweise phänotypische Distanzen vorhanden sind. Daher sind die entwickelten Methoden auf viele wissenschaftliche Fragestellungen übertragbar.

Acknowledgements

Writing acknowledgements is always a hard task. Your mind is constantly driven by the question “Did I forget anyone?” Limiting the amount of people I would like to thank to only a few and putting all the remaining people into one big box seems to be a reasonable solution. Therefore, I would like to thank everyone! All of you made the last five years of my life an incredibly exiting experience. In particular, I would like to thank Alice McHardy for giving me supervision and motivation as well as confidence making this thesis possible. I would like to thank all those people who spent hours of their life time with discussions about my work and proofreading this thesis. Here, special thanks goes to Christina Tusche. Further thanks goes to all members of the Algorithmic Bioinformatics group at HHU and the Computational Biology and Applied Algorithmics department at MPI Saarbrücken for creating a nice and inspiring work environment. I am greatly indebted to my mother for all her love, support and encouragement. Finally, I would like to thank all the poor souls for letting me sing my songs on the way to cafeteria. Rock on!

Contents

List of Figures

List of Tables

List of Abbreviations

Influenza A virus

Epitope sites	Antibody-binding sites
H / HA	Hemagglutinin
M1	Matrix protein 1
M2	Matrix protein 2
N / NA	Neuraminidase
NEP / NS2	Nuclear export protein
NP	Nucleoprotein
NS1	Nonstructural protein 1
PA	Polymerase acidic protein
PB1	Polymerase basic protein 1
PB2	Polymerase basic protein 2
vRNP	Viral ribonucleoprotein

Influenza A virus strains

BA79	A/Bangkok/1/1979
BE89	A/Beijing/353/1989
BE92	A/Beijing/32/1992
BR07	A/Brisbane/10/2007

CA04	A/California/07/2004
CC85	A/Christchurch/4/1985
EN72	A/England/42/1972
FU02	A/Fujian/411/2002
GU89	A/Guizhou/54/1989
HK02	A/Hong Kong/1143/2002
HK68	A/Hong Kong/1/1968
JO94	A/Johannesburg/33/1994
LE86	A/Leningrad/360/1986
MO99	A/Moscow/10/1999
PA99	A/Panama/2007/1999
SH93	A/Shangdong/9/1993
SI87	A/Sichuan/2/1987
SY97	A/Sydney/5/1997
TX77	A/Texas/1/1977
WE04	A/Wellington/1/2004
WI05	A/Wisconsin/67/2005
WU95	A/Wuhan/359/1995

Methodology

AD plots	Allele dynamics plots
AccTran	Accelerated transformation
CC	Pearson's correlation coefficient
DEPS	Directional evolution of protein sequences
EBF	Empirical Bayes factor
GTR	General time reversal substitution model
JTT	Jones-Taylor-Thornton substitution model
LCA	Least common ancestor
LSO	Least-squares optimization
MAE	Mean absolute error
RMSE	Root mean squared error
SD	Standard deviation
dN	Number of non-synonymous mutations
dS	Number of synonymous mutations

Miscellaneous

GISN	Global influenza surveillance network
HI assay	Hemagglutinition inhibition assay
IVR	Influenza virus ressource
WHO	World Health Organization

CHAPTER 1

Introduction

Influenza is a contagious respiratory viral infection causing substantial morbidity and mortality in annual epidemics around the globe. Being responsible for up to 500,000 deaths and up to five million infections annually (WHO, 2009a), influenza results in an enormous health and economic burden (Molinari *et al.*, 2007). Therefore, unraveling the different evolutionary as well as epidemiological aspects of influenza remains a crucial task for science.

1.1 Motivation and research aim

Influenza is the topic of many research projects and is studied by a large scientific community ($>$ 6000 articles in the PubMed database[1] published in 2011). These publications focus on different aspects of influenza pathology and biology, ranging from medical questions on the symptoms of an infection and underlying physical causes (Eccles, 2005; Roxas and Jurenka, 2007) to abstract theoretical modeling of the course of an epidemic (Adams and McHardy, 2011). With the recent advent of new sequencing technologies massive amounts of sequential information became available advancing re-

[1] PubMed [Internet]. Bethesda (MD): National Library of Medicine (US). [1946]. Available from: http://pubmed.gov/.

search on the genetic and phenotypic evolution of influenza viruses (Bao *et al.*, 2008). Numerous studies used these novel data sets and increased the general understanding of how genetic and phenotypic variations in combination with epidemiological processes shape the evolution of the influenza virus. For instance, Smith *et al.* showed that the antigenic evolution of influenza A (H3N2) viruses is clustered by multi-dimensional scaling of antigenic distances, calculated from hemagglutinition inhibition titers, between isolates sampled over a 35 year period (Smith *et al.*, 2004). Furthermore, Russell *et al.* identified East-Southeast Asia as potential reservoir from which annual epidemics of the influenza A (H3N2) virus are seeded based on phylogenetic inference in combination with the analysis of antigenic information and sampling times (Russell *et al.*, 2008b). These two studies are exemplary for the improvement of knowledge gained in recent years. However, we are far away from having answered all questions regarding influenza A evolution. For instance, global dynamics of influenza A (H3N2) viruses are much more complex with other geographical regions playing important roles in seeding and migration events, too (Bedford *et al.*, 2010; Bahl *et al.*, 2011).
The central aim of this work was to develop computational methods for the analysis of the phylodynamics of seasonal influenza A viruses, i.e. to analyze how genetic and consequently phenotypic variation is modulated by epidemiological processes and immune pressure (Grenfell *et al.*, 2004). In more detail, the developed methods should aid in resolving questions such as which mutations shape the evolutionary population structure of the virus and which mutations have a phenotypic impact in terms of antigenicity. This is important for the identification of novel viral strains that are on the rise to predominance. To warrant sufficient vaccine efficacy, the identification of mutations resulting in antigenically distinct viral strains with epidemic potential is, therefore, of utmost importance.

1.2 Outline

The present work is a cumulative dissertation based on peer-reviewed articles published in different international journals in the field of natural science. It is composed of five chapters that place the articles into a larger scientific context (chapter 1), present the main articles of the author (chapters 2 to 4) as well as a synopsis on the presented publications (chapter 5). For each article the author's contribution as well as the state of publication is addressed.
The articles are presented in chronological order in the author's manuscript version as accepted by the respective journal allowing for a uniform appearance. However,

the content (text, figures and tables) of each article is identical to the edited journal version. The published articles are additionally provided in the appendix.

1.3 Influenza viruses

The scope of application of the developed methods in this thesis is on the influenza virus, in particular genera A subtype H3N2, as this rapidly evolving pathogen poses a substantial threat to public health and economy (Dushoff *et al.*, 2006; Molinari *et al.*, 2007; WHO, 2009a). Furthermore, influenza A (H3N2) viruses are well studied and a large amount of publicly available data records (genome sequences) exists (Bao *et al.*, 2008). In the following section the epidemiological and molecular dynamics of influenza viruses with a focus on influenza A viruses will be explained.

1.3.1 Disease patterns and global activity

Influenza viruses cause mild to severe illness that has a sudden onset and is naturally overcome within one week without requiring any medical treatment (WHO, 2009a). Exceptions are risk groups like the very young, the elderly, pregnant women and people suffering from medical conditions, for which an infection may lead to life-threatening complications. Viral transmission from person to person is carried out through air or via direct skin-to-skin contact (WHO, 2009a) and local epidemics easily spread within the population via crowded places, such as schools or work places (Viboud *et al.*, 2006; Cauchemez *et al.*, 2008).
Global influenza activity is characterized by seasonal epidemics occurring annually during winter seasons in temperate regions, whereas in tropical regions influenza is prevalent year-round (**Figure 1.1**). Reasons for the pattern of seasonality in temperate regions are still unclear and various theories exist (Lipsitch and Viboud, 2009). One possible explanation was given by Shaman and Kohn who showed based on experiments with guinea pigs that virus transmission is most efficient at low vapor pressure (Shaman and Kohn, 2009).

1.3.2 Influenza A virus - genomic structure

The influenza virus is a single-stranded, segmented, negative-sense RNA virus of the family *Orthomyxoviridae*. Three distinct genera or types (A, B and C), typed on the basis of the nucleoprotein and matrix protein antigens (see below), circulate in nature and highly differ in host range and pathogenicity (Webster *et al.*, 1982; Taubenberger

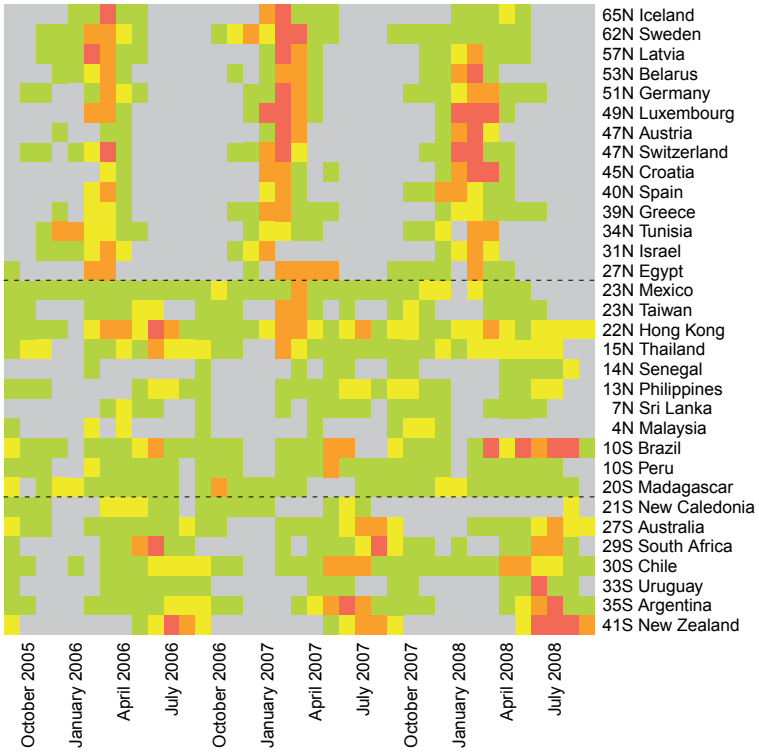


Figure 1.1: Overall influenza (subtypes A (H3N2), A (H1N1) and B) activity for selected countries between autumn 2005 and summer 2008 sorted by latitude (WHO, 2006a,b, 2007a,b, 2008a,b). The World Health Organization classifies influenza activity into five distinct categories: **No activity** (gray), **sporadic activity** (green), **local activity** (yellow), **regional outbreaks** (orange) and **widespread outbreaks** (red). In temperate regions of the northern and southern hemisphere (separated by dashed lines) influenza causes annual epidemics in winter seasons, whereas in tropical regions influenza is prevalent year-round. Motivated by Nelson and Holmes (2007).

and Kash, 2010). Of these, influenza A viruses evolve most rapidly and are commonly associated with most influenza infections in humans (Lin *et al.*, 2004). Furthermore, type A viruses can cause zoonotic infections, host switch events and create pandemic viruses (Webster *et al.*, 1992; Steinhauer and Skehel, 2002).

The influenza A virus genome is composed of eight distinct segments encoding eleven or twelve proteins (Medina and García-Sastre, 2011, **Figure 1.2**): The viral RNA polymerase complex is formed by a heterotrimer consisting of polymerase basic protein 2 (PB2), polymerase basic protein 1 (PB1) and polymerase acidic protein (PA), each encoded by a different genome segment. The PB1 segment further encodes the pro-apoptotic protein PB1-F2, which is expressed only by some viruses, and the newly

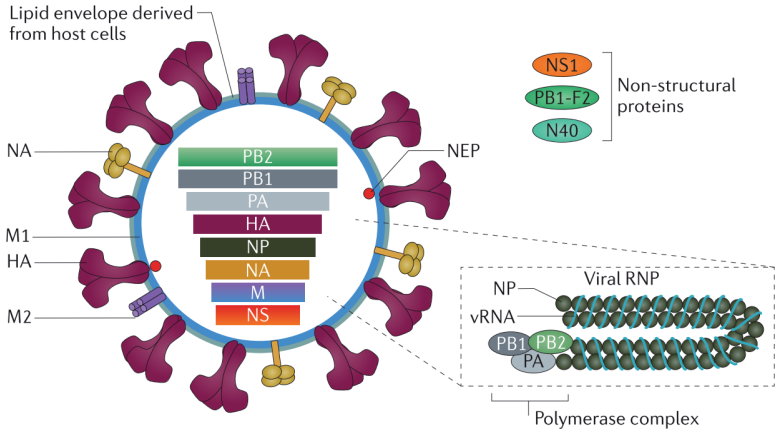


Figure 1.2: Schematic representation of the influenza A virion. Adapted from Medina and García-Sastre (2011) with permission from nature publishing group.

identified protein N40 of unknown function (Wise *et al.*, 2009). The receptor-binding and fusion protein hemagglutinin (H or HA), neuraminidase (N or NA), which is involved in the release of new virions from infected cells, and nucleoprotein (NP) are all encoded by individual genome segments. HA and NA serve as viral antigens. The M segment encodes the matrix protein 1 (M1), which is located below the lipid membrane, and the trans-membrane matrix protein 2 (M2) that acts as an ion channel. Finally, the NS segment encodes the nuclear export protein (NEP or NS2) and the nonstructural protein 1 (NS1). NS1 is involved in cellular RNA transport, splicing, translation, as well as host mediated antiviral response.

The genome segments exist in the form of viral ribonucleoproteins (vRNP), with the viral RNA wrapped around NP monomers. The polymerase complex is bound to a short hairpin structure formed by the partially complementary 5' and 3' untranslated regions of each RNA segment. In total, the influenza A virus genome is of 13.5 kb in length with genomic segments ranging from 2.3 kb to 0.9 kb (Ghedin *et al.*, 2005). The virion itself is composed of a lipid bilayer derived from a host cell and a layer of M1, which interacts with the vRNPs, as well as with the cytoplasmic part of the surface proteins. The viral surface is covered with the two integral membrane proteins HA (homotrimeric) and NA (homotetrameric) and is penetrated by M2 ion channels. Besides the vRNPs, NEP is also present in the virion, whereas NS1, PB1-F2 and N40 are only expressed in the host cell. For a more detailed overview on the genomic structure of influenza A viruses and the role of each encoded protein in the viral life cycle, see Lamb and Krug (2001) or Medina and García-Sastre (2011).

Influenza A viruses are further divided into different serotypes based on genetic and antigenic properties of the surface proteins HA and NA (Webster *et al.*, 1982). In nature, 17 distinct HA subtypes (H1-H17) and nine distinct NA subtypes (N1-N9) exist and circulate in various combinations in wild aquatic birds (Fouchier *et al.*, 2005; Liu *et al.*, 2009; Tong *et al.*, 2012). In the human population, currently the subtypes H1N1 and H3N2 are endemic (WHO, 2012). However, sporadic cases of infections with influenza A viruses of subtypes H5N1, H1N2 and H9N2 are observed from time to time (WHO, 2011b,a, 2012).

1.3.3 Evolutionary mechanisms - antigenic drift and shift

Human influenza viruses are in a constant race with the host immune system to circumvent host immunity. Immunity is achieved by protective antibodies against the viral antigens HA and NA and is elicited by previous infections or vaccination. The virus has to constantly alter the sequence composition of the two surface glycoproteins and, thus, their physico-chemical properties, to evade recognition by the host immune system (Webster *et al.*, 1982). This process is known as *antigenic drift* and results in the accumulation of advantageous mutations mainly in the antibody-binding (epitope) sites of HA, which increase viral fitness by changing its antigenic properties (Bush *et al.*, 1999b; Smith *et al.*, 2004). However, not all changes in the HA are advantageous, such that the according viral isolates fail to survive in the viral population. For subtype H3N2, this dynamic results in a "cactus-like" (also known as "ladder-shaped") phylogenetic tree of the HA segment: the trunk represents the surviving viral lineage over time and short side branches that stem from the trunk represent extinct lineages (Bush *et al.*, 1999b; Ferguson *et al.*, 2003; Nelson and Holmes, 2007; Holmes, 2010, **Figure 1.3**). However, other studies have shown that this behavior is unique to HA (Holmes *et al.*, 2005). Genetic diversity in the viral population is much higher if all genomic segments are included in the analysis.

The segmented nature of the influenza A virus genome furthermore allows for genome reassortment events. Reassortment events denote the interchange of genomic segments between different viral isolates that co-infect a host cell (Nelson and Holmes, 2007). The resulting viral isolates have an altered genome composition, where different segments are inherited from the co-infecting viral isolates. Consequentially, the viral phenotype may be altered with respect to, for instance, replicability, pathogenicity and transmissibility. Reassortment events play a crucial role in influenza A virus evolution, as single proteins with advantageous properties but detained by the remaining genome composi-

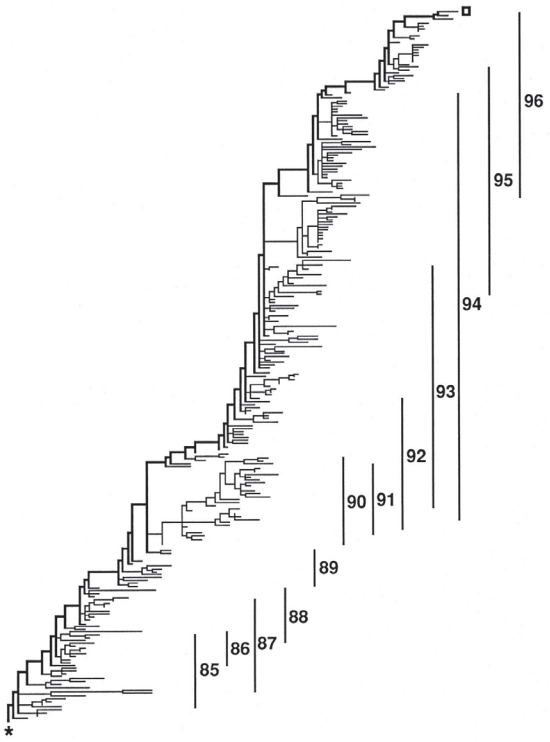


Figure 1.3: Maximum parsimony phylogeny of HA sequences of seasonal influenza A (H3N2) viruses isolated between 1983 and 1996. The thick line drawn from the root (asterisk) to the evolutionary most distant viral isolate (square) indicates the trunk of the tree, representing the surviving lineage over time. Years of isolation are indicated by vertical lines. Adapted from Fitch *et al.* (1997) with permission from the National Academy of Sciences of the United States of America.

tion can be put into a more favorable genetic context laying the path to predominance. If the major viral antigen HA is involved in such a reassortment event, novel serotypes may arise against which the human population is immunologically naive (Webster *et al.*, 1982; Cox and Subbarao, 2000). These *antigenic shifts* are the major cause of severe pandemics and happened four times in the last 100 years, accounting for millions of deaths worldwide (Tognotti, 2009; WHO, 2010a). Besides reassortment events, the introduction of an animal influenza A virus into the human population can cause antigenic shifts, too (Cox and Subbarao, 2000). However, the mechanisms of how a viral strain becomes a pandemic strain and the determinants of pathogenicity are still unclear and are the subject of ongoing research. Besides antigenic shifts, reassortment events are common in the evolution of endemic influenza, too (Barr *et al.*, 2005; Holmes *et al.*, 2005), although the extent is still unclear.

1.3.4 Preventing infections - influenza vaccination

Due to antigenic drift, antigenically novel strains of influenza A virus appear on a regular basis and rise to predominance in world-wide epidemics. Therefore, the composition of the influenza vaccine, consisting of one viral strain of each influenza A (H1N1), A (H3N2) and B virus, has to be adapted regularly to match the currently predominant viral strains (Russell *et al.*, 2008a). The World Health Organization (WHO) maintains a global influenza surveillance network (GISN), consisting of 4 WHO collaborating centers and >120 national centers, to monitor the genetic and antigenic properties of the circulating influenza virus population (Cox *et al.*, 1994; Russell *et al.*, 2008a). The gathered information is evaluated twice a year to assess whether an vaccine update is necessary or not. This evaluation is made in January or February for the Northern hemisphere winter season and in September for the Southern hemisphere winter season. Various serological tests and computational methods guide the analysis. If a viral strain is detected that is antigenically distinct enough to the current vaccine strain, i.e. antibodies induced against the vaccine strain do not suppress the binding of the novel strain, an update of the vaccine composition is recommended. In most cases this strategy results in a well-matched vaccine significantly decreasing morbidity and mortality in the human population (Karlsson Hedestam *et al.*, 2008). The downside of this "predict and produce" approach is that the decision has to be made almost one year in advance of the actual usage, to assure sufficient time for the vaccine manufactures. Thus, if a novel antigenic variant is identified too late to be included into the influenza vaccine, vaccine efficiency will be decreased due to an insufficient match to the predominant viral strain (de Jong *et al.*, 2000; Gupta *et al.*, 2006).
Besides the seasonal influenza vaccine, ongoing research is focusing on universal vaccines (Du *et al.*, 2010). The aim is to circumvent frequent reformulations and to enable cross-protection across different serotypes. The cross-protection also bears the possibility of an universal vaccine to be applicable to pandemic viral strains. Given the ongoing threat of the highly pathogenic influenza A (H5N1) virus suddenly gaining the characteristics necessary for sustained human-to-human transmission, the importance of universal vaccines becomes clear (Abdel-Ghafar *et al.*, 2008; Sui *et al.*, 2009). Popular approaches for the development of an universal influenza vaccine target conserved regions of the less variable stem of the HA (Sui *et al.*, 2009), the ectodomain of the matrix protein 2 (Huleatt *et al.*, 2008) or the internal proteins of influenza A viruses (Berthoud *et al.*, 2011).

1.4 Studying the antigenic evolution of influenza A viruses

The molecular genetic evolution of the influenza A virus is shaped by both mutations and reassortment events. But even though reassortment events are frequent in the evolution of seasonal influenza A viruses (Barr *et al.*, 2005; Holmes *et al.*, 2005), antigenic drift is the evolutionary mechanism that has the highest impact. Evaluation of sampled viral isolates with phylogenetic analysis and serological tests is the standard approach to assess the antigenic evolution of the virus (WHO, 2012). However, understanding the course of antigenic drift and the underlying mechanisms that drive single antigenic variants to predominance is the key to ensure high vaccine efficacy (Carrat and Flahault, 2007). In the following section, we will cover computational methods that aid in the understanding of the antigenic evolution of influenza A viruses.

1.4.1 Phylogenetic inference

A common way to analyze and visualize the genetic structure of and evolutionary relationships within viral populations is the use of phylogenetic inference methods (Felsenstein, 2004; Grenfell *et al.*, 2004; Yang and Rannala, 2012). These methods infer a phylogenetic tree or phylogeny, "the evolutionary history of an organism or group of related organisms" (Cammack *et al.*, 2006). More precisely, phylogenetic trees illustrate the ancestral relationships between single genes or sets of species, represented by molecular sequences (Felsenstein, 2004). In a phylogenetic tree, the input sequences are called operational taxonomic units or taxa and are mapped to the terminal nodes of a connected acyclic graph (**Figure 1.3**). This allows for an easy visualization and interpretation of the evolutionary relationships.
Several different methods exist to infer a phylogenetic tree based on different optimality criteria. One distinguishes maximum parsimony, maximum likelihood and Bayesian methods. Each has advantages and disadvantages with respect to runtime/complexity, accuracy and interpretability (Yang and Rannala, 2012). However, all have in common that the underlying search space increases super exponentially with increasing number of taxa (Harding, 1971) and, thus, heuristic search techniques are usually implemented for which there is no guarantee that the optimal solution can be found. An alternative approach are distance-based methods, such as neighbor joining (Saitou and Nei, 1987), that apply clustering techniques to infer a phylogenetic tree from pair-wise evolutionary distances. Application of ancestral character state reconstruction techniques on phylogenetic trees furthermore allows the reconstruction of mutational paths in the evolutionary history of the underlying data (Fitch, 1971; Yang *et al.*, 1995; Pagel *et al.*, 2004).

In influenza A virus research phylogenetic inference of single or concatenated genomic segments is a common technique to study the evolutionary history of the viral population. Basically, all inference techniques are applied and there is no clear preference of one technique over the others, although maximum likelihood and Bayesian inference gained more popularity due to increased computing power in recent years. Phylogenetic inference can be used to study different questions regarding influenza A evolution. For example, Holmes *et al.* constructed phylogenies of all genomic segments of influenza A (H3N2) virus isolates to detect multiple reassortment events among viruses circulating in New York State, USA, between 1999 and 2004 (Holmes *et al.*, 2005). Another study used phylogenies of the HA and NA segments of influenza A viruses (subtypes H1N1 and H3N2) to detect pairs of sites with putative epistatic interactions (Kryazhimskiy *et al.*, 2011). These examples highlight the easy applicability and the value of phylogenetic inference in the research of influenza virus evolution.

1.4.2 Detection of natural selection

Natural selection as introduced by Darwin (1859) is one of the basic concepts in molecular biology and describes the survivability among phenotypes in a certain environment (Clifford, 1976; Hurst, 2009). In more detail, natural selection acts in populations with multiple forms of a phenotypical trait (discrete or continuous), where individuals of a specific trait tend to survive and reproduce more successfully due to better adjustment to the given environment and, thus, ensure the perpetuation of the phenotypical characteristics in succeeding generations[2]. This varying reproductive success is measured in terms of fitness that is dependent on the given environmental conditions (Clifford, 1976; Orr, 2009). In order of natural selection to act, a selective pressure is needed, i.e. a specific environmental constraint that results in differential fitness evaluation of given phenotypic traits.

Three modes of natural selection exist (Brodie *et al.*, 1995; Hurst, 2009; Oleksyk *et al.*, 2010, **Figure 1.4**): (i) positive selection, (ii) purifying selection and (iii) balancing selection. Positive selection favors advantageous alleles, i.e. alleles with increased fitness, that results in their fixation in the population. This mode also represents the form of natural selection originally introduced by Darwin (1859). Positive selection is also known as Darwinian or directional selection (if associated with a quantitative trait). Purifying selection eliminates deleterious mutations. This results in favoring alleles that are already present at high frequencies in the population by reducing the

[2] "Natural selection" [Internet]. Merriam-Webster.com. [2011]. Available from: http://www.merriam-webster.com (8 June 2012).

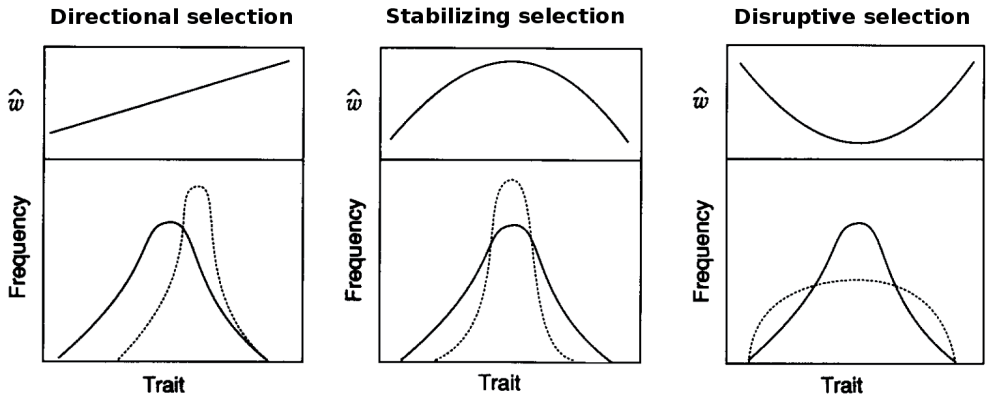


Figure 1.4: Modes of natural selection. Illustrations show the expected relative fitness ($\hat{\omega}$) as a function of a phenotypic trait and phenotypic distributions before (solid curves) and after selection (dashed curves). Adapted from Brodie *et al.* (1995) with permission from Elsevier.

spectrum of the phenotypic trait. In other words, purifying selection increases the frequency of already well-adapted alleles in the population and, thus, is expected to be the common mode of selection in nature. Purifying selection is also known as negative or stabilizing selection. The third mode of natural selection, balancing selection, maintains multiple distinct phenotypes at high frequencies favoring diversity. This mode of selection acts when an organism is present in multiple environments with different environmental conditions, such that single alleles are advantageous in the different environments. Therefore, an allele can never become fixed in the population. Balancing selection is also known as disruptive selection.

In seasonal influenza A virus evolution natural selection is a key source of molecular variation. The virus is subject to directional selection, i.e. it has to continuously alter the composition of its surface antigens to evade host recognition and immunity gained by previous infections or vaccination. Thus, viral isolates expressing surface antigens with unknown sequence compositions to the human immune system have a higher fitness in comparison to viral isolates whose surface antigens are identical to those included in the seasonal influenza vaccine. Therefore, frequent updates of the influenza vaccine composition are inevitable (Hay *et al.*, 2001).

On the molecular level, different studies have focused on the major surface antigen hemagglutinin to search for evidence of positive selection. A main method to detect positive selection acting on specific protein sites is the use of the ratio of non-synonymous (dN) to synonymous (dS) mutations observed in a phylogenetic tree. Depending on

this ratio (dN/dS) either positive selection (> 1), negative selection (< 1) or no selection ($= 1$) is assumed to act on the given protein site. However, interpretation of this test should be done carefully if applied to samples drawn from a single population, as statistical power may be reduced in this setup (Kryazhimskiy and Plotkin, 2008). Application to HA sequences from influenza A (H3N2) viruses isolated between 1983 and 1997 detected 18 protein sites to be under positive selection (Bush *et al.*, 1999b). These codons were subsequently used to identify viral strains that are most likely to rise to predominance (Bush *et al.*, 1999a). However, results are biased as retrospective tests were made on the data set that was used to detect the codons under positive selection. An alternative approach to detect directional evolution at individual sites of an protein alignment (DEPS) was given by Pond *et al.*. Based on a specific protein evolution model they performed a phylogenetic maximum likelihood test and identified 20 sites subject to directional evolution in the complete genome of influenza A (H3N2) virus (Pond *et al.*, 2005, 2008). However, a disadvantage of this approach is that it relies on the baseline evolutionary model and application to dim-light and color vision genes in vertebrates did not result in the correct identification of adaptive sites (Nozawa *et al.*, 2009).

1.4.3 Advanced approaches

Many studies analyze the genetic evolution and possible means towards the antigenic evolution of influenza A (H3N2) viruses. For instance, Plotkin *et al.* (2002) used agglomerative single-linkage clustering to group hemagglutinin HA1 sequences into disjoint clusters that were successively predominant, and Xia *et al.* (2009) proposed a "site transition network" to visualize co-occurring amino acid changes in the HA on the basis of a mutual information approach. Other studies that include phylogenetic inference in their analysis take advantage of the nature of the underlying data and are categorized as *phylodynamic* techniques.

Phylodynamic techniques

The term "*phylodynamics*" was coined by Grenfell *et al.* (2004) and applies to the analysis of the evolutionary and epidemiological patterns of rapidly evolving pathogens such as RNA viruses (Grenfell *et al.*, 2004). RNA viruses are characterized by large population sizes, short generation times and a high mutation rate that allows them to efficiently react to the strong selective pressure imposed by the host immune system (Duffy *et al.*, 2008). For influenza A viruses, this is approximately 7.6×10^{-5} mutations

per site per genome replication (Drake, 1993), which is several magnitudes higher than for double-stranded DNA viruses, such as the herpes simplex virus type 1 (7.7×10^{-8} mutations per site per genome replication, Drake and Hwang, 2005). These dynamics of rapidly evolving pathogens allow the evolutionary processes to occur on a time scale similar to the epidemiological processes. If homologous sequences and appropriate epidemiological data are available, so called phylodynamic techniques allow for a joint analysis (Grenfell *et al.*, 2004; Pybus and Rambaut, 2009). In more detail, these methods try to unify the interacting epidemiological and genomic dynamics on the basis of phylogenetic inference. Using epidemiological information such as sampling locations or sampling times, phylodynamic techniques facilitate, for instance, the inference of geographic migration patterns or to date past evolutionary events. For instance, Wallace *et al.* used a "phylogeographic" approach to infer migration paths of highly pathogenic avian influenza A (H5N1) viruses across Asia (Wallace *et al.*, 2007) and Smith *et al.* dated the genomic most recent common ancestor of the 2009 pandemic influenza A (H1N1) virus (Smith *et al.*, 2009).

In the present work we analyzed the impact of amino acid changes on the evolution of influenza A viruses over time. Different approaches exist that, for instance, use amino acid frequency diagrams (Shih *et al.*, 2007) or analyze patterns of co-occurrence either within or between genomic segments (Du *et al.*, 2008; Xia *et al.*, 2009). In contrast, our method to infer allele dynamics plots (AD plots) takes advantage of the evolutionary structure of the data inferred by phylogenetic inference (Steinbrück and McHardy, 2011). AD plots visualize the frequency of different gene alleles over time. Gene alleles are defined as non-empty sets of amino acid changes mapped to individual branches of a phylogeny and their frequencies are calculated based on the number of viral isolates in a subtree. This allows for an accurate representation of the underlying population structure and, furthermore, enables the study of the according phylodynamics of the analyzed gene. Application to approximately 5,000 influenza A (H3N2) virus HA sequences visualizes the evolutionary dynamics between 1998 and 2009 with several alleles being present at low frequencies for short time periods and multiple alleles rising to fixation. Additionally, the AD plot shows several alleles that are present for a short time, only, and temporarily rise to high frequency. These alleles are mainly observed during times when an antigenic variant has been predominant in worldwide epidemics over several years. Overall, this picture reflects the "cactus-like" structure of the phylogeny, i.e. one surviving lineage over time and short side branches accounting for extinct lineages, at a higher degree of detail. For instance, one can easily determine the order in which the substitutions of the surviving lineage became fixed. Further ap-

plication to HA sequences of early 2009 pandemic influenza A (H1N1) viruses showed that many different alleles co-circulated at low levels but only one allele became fixed. This indicates that the virus was rather stable in terms of antigenicity during that time. On the other hand, caution is warranted as only a short time period is covered, with unbalanced sampling of viral isolates in different months.
A key benefit of AD plots is the early identification of antigenic variants that became predominant in the analyzed time period based on a novel measure for directional selection. If we assume that selection acts on individual alleles, those alleles with a fitness advantage will rise in frequency faster than those alleles without a selective advantage. Hence, gene alleles with the largest frequency increase between two time intervals are more likely to be subject to directional selection than others. Application to the afore mentioned AD plot of HA sequences of seasonal influenza A (H3N2) viruses resulted in the detection of those changes with a beneficial impact on the viral population. In four out of five cases the AD plot identified the sets of amino acid changes that were associated with the viral lineage that later became predominant in the viral population. A comparison to DEPS and dN/dS ratio tests, which identify sites under directional selection, revealed that the classic tests cannot detect all rapidly fixed amino acid changes. In this sense our method complements the classical tests for directional selection.

Detecting positive selection - from sites to patches

The classical dN/dS ratio test to detect positive selection treats single protein sites independently. Due to structural constraints within the protein, large-scale protein-protein interactions and protein interactions with other macro-molecules this independence usually doesn't hold in nature. In a different study with minor contribution of the author of this thesis we extended the basic method of dN/dS ratio tests for the detection of positive selection to identify groups of protein sites under positive selection on the surface of a protein (Tusche *et al.*, 2012). We use a graph-cut algorithm to cluster sites, located in spatial proximity on a protein surface, that are assumed to be under positive selection. Positive selection is measured based on a significant deviation of dN/dS ratios from the protein-wide average. Application to HA of seasonal influenza A viruses of subtypes H3N2 and H1N1 identified several patches on the protein surface that are mainly composed of known antibody-binding sites. Comparison to the standard dN/dS ratio test revealed that our method had a higher accuracy in detection of known antibody-binding sites. Further application to the PB2 protein of the 2009 pandemic influenza A (H1N1) virus identified patches that included sites

known to be relevant for successful replication in mammalian hosts. For the HA of the 2009 pandemic influenza A (H1N1) virus several patches were identified. Due to the rather young evolutionary history of the virus, the evolutionary importance of these patches needs further experimental validation. In summary, joint identification of sites in spatial proximity under positive selection is more informative than the identification of single sites alone.

From genotype to phenotype

AD plots and the way how they detect alleles under directional selection identify the top-ranking alleles in every season. This is a limitation, as novel antigenic clusters only become predominant in global epidemics every three to four years (Smith *et al.*, 2004). Additional antigenic information is necessary to evaluate the potential influence of the identified alleles on the antigenic evolution of the virus, which was not available for the used data set. All methods that use genetic data only to make statements about antigenic properties of influenza A viruses, rely on single reference strains, whose antigenic characteristics in comparison to other viral strains are known (for instance Plotkin *et al.*, 2002). These reference strains, usually identified by the WHO, are used to represent groups of viruses with similar antigenic characteristics that were predominant at a given point in time (for instance WHO, 2007a). Therefore, conclusions about the identified properties of the antigenic evolution of influenza A viruses can only be made at a discrete or clustered level with such methods.
In general, the antigenic characteristics of specific viral strains are measured with means of hemagglutinition inhibition (HI) assays (Hirst, 1943). This is a binding assay that quantifies the ability of an antiserum to inhibit the natural ability of HA to agglutinate red blood cells. The HI titer represents the degree of dilution of an antiserum needed, such that this inhibition is no longer possible. This makes it possible to compare sampled viral isolates with respect to prepared antisera. Additionally, this test is also one of the standard tests to evaluate the antigenic characteristics of currently circulating viral strains done by the WHO and, thus, is important in the vaccine strain selection process (Russell *et al.*, 2008a).
In 2004, Smith *et al.* introduced antigenic cartography to study the antigenic evolution of influenza A (H3N2) viruses (Smith *et al.*, 2004). They used log-transformed HI distances in a multi-dimensional scaling approach to visualize the antigenic evolution of the virus in a two-dimensional map. Based on data sampled over a 35 year time period they showed that although the genetic evolution of the virus is continuous, the

antigenic evolution is clustered, with single antigenic clusters being predominant in world-wide epidemics for ~ 3.5 years. However, the downside of this approach is that the antigenic and the genetic evolution are only linked by means of prototype viruses of the individual antigenic clusters. Other methods use both genetic and antigenic data in a classification task to predict antigenically novel variants (Lee *et al.*, 2007; Liao *et al.*, 2008; Huang *et al.*, 2009). Antigenic information is used to decide whether two viral isolates are antigenically distinct or not. Hence, a qualitative evaluation of the degree of antigenic impact of genetic variations is not performed with such methods. Currently, such information is gained by time- and cost-expensive experimental characterization of mutant viruses only (Smith *et al.*, 2004). We introduced a method to infer genotype-phenotype relationships for the influenza A virus by means of antigenic trees (Steinbrück and McHardy, 2012). In antigenic trees pair-wise log-transformed antigenic distances are fitted to single branches of a phylogeny inferred from associated HA sequences using least-squares optimization. This allows for the inference of antigenic weights for individual branches and, thus, for the mapped genetic changes, which were reconstructed using ancestral character state reconstruction. For sufficiently resolved data the inference of antigenic weights could be tracked down to individual changes. Evaluation of the accuracy of predicting unseen antigenic distances for an antigenic tree of HA sequences of 258 influenza A (H3N2) viruses showed a good performance similar to antigenic cartography (0.86, 0.72 and 0.86 versus 0.83, 0.67 and 0.8; absolute prediction error, standard deviation and Pearson's correlation coefficient). This showed that the antigenic evolution can be accurately correlated to the genetic evolution of the influenza A (H3N2) virus. Identification of branches with high antigenic weights allowed to define antigenic types and associated amino acid changes. Although antigenic types and the antigenic clusters found by antigenic cartography largely overlap in terms of assigned viral isolates, both concepts differ. Antigenic clusters are characterized by similar antigenic properties only, whereas antigenic types are characterized by similar antigenic characteristics and evolutionary history. Additionally, the k-means approach used in antigenic cartography relies on a well-defined cluster structure of the data. In other scenarios, where clusters are less distinctive, robust cluster assignments and the identification of phenotype-associated amino acid changes are hard to achieve by the k-means approach. In contrast, our method would likely be able to resolve genotype-phenotype relationships that are supported by the data. We furthermore identified protein sites and individual amino acid changes of high relevance in the antigenic evolution of influenza A (H3N2) viruses. Comparison with experimentally validated positions from an unpublished study of Koel *et al.* showed large overlap with

the identified positions and amino acid changes further highlighting their antigenic importance (*Koel et al.; Antigenic evolution of influenza A (H3N2) virus is dictated by 7 residues in the hemagglutinin protein; 2nd International Influenza Meeting, Münster; 2011*).

1.5 Outlook

In this thesis new methods for analyzing the phylodynamics of influenza A viruses were developed and evaluated. Both methods, AD plots and antigenic trees, provide further insights into the antigenic evolution of influenza A viruses and, thus, may assist in the biannual vaccine strain selection process. Evaluation of the performance of the vaccine strain selection process was demonstrated by a comparison of the recommendations made by the WHO and the actually predominant antigenic variants. Correct identification of novel antigenic variants by the WHO was delayed by at least one season resulting in a mismatch of the vaccine strain and the circulating viral strain in the selected seasons. Both methods can aid in the evaluation of the genetic and antigenic composition of the currently circulating viral population. For AD plots we showed their potential to predict future predominant lineages.

A possible continuation of the two projects is their combination, i.e. the antigenic evaluation of the top ranking alleles. Russell *et al.* provide a valuable data set for this purpose that comprises a huge set of HI titers and sequenced HA data for viral isolates sampled worldwide between 2002 and 2007 (Russell *et al.*, 2008b). They used these data to identify East-Southeast Asia as potential reservoir from which seasonal epidemics of influenza A (H3N2) viruses are seeded. These data are particularly suited for our analysis as they are a representative sample of the globally circulating viral population over the analyzed years. Our initial results are promising and further evaluation may lead to a valuable tool.

Another future research direction is the extension of the antigenic tree concept to further resolve the antigenic impact of individual changes. A possible approach is to impose weights on individual changes that map to single branches in the phylogeny rather than on single branches alone. In this setup, branch weights would result from a combination of weights inferred for the different changes that map to the according branches. However, this requires a high resolution of the data to guarantee confident estimates for the single weights.

Note that although the application focus of both methods was on the analysis of influenza A viruses, their application is not limited to this pathogen. For AD plots, any

organism/pathogen where homologous sequence data and associated sampling times are available can serve as input. It could, for instance, be used to study the intra-host evolution of HIV infections. For antigenic trees, application is possible to research questions where pair-wise phenotype distances and homologous sequence data are available.

CHAPTER 2

Personal bibliography

Publications of the thesis

- Steinbrück, L. and A. C. McHardy (2011). Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic Acids Res*, **39** (1): e4. http://nar.oxfordjournals.org/content/early/2010/10/18/nar.gkq909
- Steinbrück, L. and A. C. McHardy (2012). Inference of genotype-phenotype relationships in the antigenic evolution of human influenza A (H3N2) viruses. *PLoS Comput Biol*, **8** (4): e1002492. http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002492

Other publications

- Gilbert, J. A., J. A. Steele, J. G. Caporaso, L. Steinbrück, J. Reeder *et al.* (2011). Defining seasonal marine microbial community dynamics. *ISME J*, **6** (2): 298–308. http://www.nature.com/ismej/journal/v6/n2/full/ismej2011107a.html
- Tusche, C., L. Steinbrück and A. C. McHardy (2012). Detecting patches of protein sites of influenza A viruses under positive selection. *Mol Biol Evol*, (in press). http://mbe.oxfordjournals.org/content/early/2012/03/16/molbev.mss095.abstract

CHAPTER 3

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

Status	**published**
Journal	Nucleic Acids Research (Impact factor 7.836)
Citation	Steinbrück, L. and A. C. McHardy (2011). Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic Acids Res*, **39** (1): e4.
URL	http://nar.oxfordjournals.org/content/early/2010/10/18/nar.gkq909
Own contribution	75% Performed the experiments Analyzed the data (with co-authors) Wrote the manuscript (with co-authors)

3.1 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 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.

3.2 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 (Grenfell *et al.*, 2004; Pybus and Rambaut, 2009). 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 (Wallace *et al.*, 2007), 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 (Keele *et al.*, 2006) based on 'phylogeo-

graphic' clustering, and identified the epidemic history and geographic source of HIV type 2 based on a molecular clock analysis of dated genetic sequences (Lemey *et al.*, 2003).

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, can be identified based on the ratio of non-synonymous to synonymous mutations (dN/dS) (Bush *et al.*, 1999b). 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 (Kryazhimskiy and Plotkin, 2008). 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 (Pond *et al.*, 2005, 2008). 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 (Ricklefs, 2007). 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.

3.2.1 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 (Koelle *et al.*, 2006). 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 9 of NA occur in various combinations in aquatic birds (Fouchier *et al.*, 2005). 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 (WHO, 2009b, 2010b).

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 (Bush *et al.*, 1999b; Smith *et al.*, 2004). Antigenically relevant changes are located mainly in the epitope sites of the viral HA (Wiley *et al.*, 1981; Wiley and Skehel, 1987; Wilson and Cox, 1990; Skehel and Wiley, 2000). 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, can thus be transferred into a more favorable genetic context and subsequently rise to predominance (Webster *et al.*, 1992; Kuiken *et al.*, 2006; Lowen and Palese, 2007; Morens *et al.*, 2009; Neumann *et al.*, 2009; Zimmer and Burke, 2009).

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 (Cox *et al.*, 1994; Russell *et al.*, 2008a). 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 (Nelson *et al.*, 2007) and to identify East and Southeast Asia as a global evolutionary reservoir of seasonal influenza A (H3N2) viruses (Russell *et al.*, 2008c). 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 (Rambaut *et al.*, 2008).

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 (Shih *et al.*, 2007). 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 (Plotkin *et al.*, 2002). 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 (Koelle *et al.*, 2006; Rambaut *et al.*, 2008). 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 (Plotkin *et al.*, 2002). Du *et al.* constructed a co-occurrence network from co-occurring nucleotides across the whole genome (Du *et al.*, 2008). 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' (Xia *et al.*, 2009). 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 non-synonymous to synonymous nucleotide substitutions (dN/dS) on the trunk of a phylogenetic tree of hemagglutinin HA1 subunit sequences (Bush *et al.*, 1999b). 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 (Bush *et al.*, 1999a). 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 (Smith *et al.*, 2004). 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 (Pond *et al.*, 2008). 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 (Nozawa *et al.*, 2009).

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 (Smith *et al.*, 2004; Fouchier and Smith, 2010). 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 three to four years on average (Smith *et al.*, 2004). 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 (Lee *et al.*, 2007; Liao *et al.*, 2008; Huang *et al.*, 2009); 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.

3.3 Methods

3.3.1 Phylogenetic inference

HA sequences from 4,913 seasonal human influenza A (H3N2) virus isolates sampled from 1988 to 2008, and from 1,516 swine-origin influenza A (H1N1) virus isolates with exact sampling times (year and month) were downloaded from the influenza virus resource (Bao *et al.*, 2008). Alignments of DNA and protein sequences were created with Muscle (Edgar, 2004a) and manually curated. Phylogenetic trees were inferred with PhyML v3.0 (Guindon and Gascuel, 2003) 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 (Zwickl, 2006).

3.3.2 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 (Futuyma, 1997; Hein *et al.*, 2005; Templeton, 2006). 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 (Huelsenbeck and Ronquist, 2001; Drummond and Rambaut, 2007)), 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 (Fitch, 1971) for ancestral character state reconstruction; however, in principle, any available method can be applied (Felsenstein, 2004; Pagel *et al.*, 2004). In our analysis, we chose the isolate with the earliest sampling date as an outgroup and used accelerated transformation (AccTran) (Felsenstein, 2004) 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. Allele 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 3.1**).

3.3.3 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

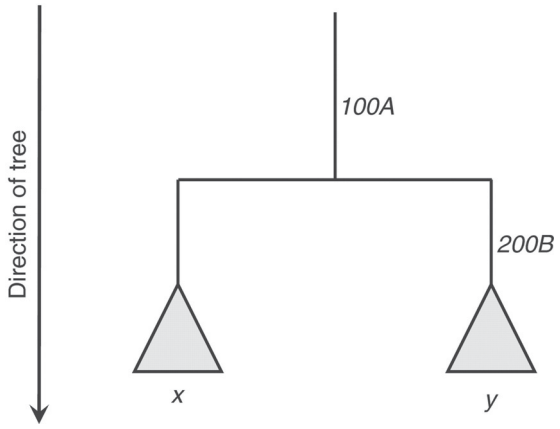


Figure 3.1: A tree demonstrating the concepts of alleles and allele frequency correction. For allele *100A*, only the isolates of subtree x are counted, whereas for allele *200B *100A**, the isolates in subtree y are considered.

genetic changes of hemagglutinin 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 (Wiley *et al.*, 1981; Wiley and Skehel, 1987).

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 (Nelson and Holmes, 2007). 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 October 1st and ends on the 31st of March in the following year. For the Southern hemisphere, the influenza season begins on the 1st of April and ends on the 30th of 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 (Smith *et al.*, 2004). The changes in these strains for the five HA epitopes are given in **Table 3.1**.

Table 3.1: 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 hemagglutinin. Note that PA99 is antigenically similar to MO99 and was used as the vaccine candidate strain for MO99 (WHO, 2002a).

Antigenic cluster	Substitutions	Reference
A/Sydney/5/1997 (SY95)	62E, 156Q, 158K, 196A, 276K	(Lin *et al.*, 2004)
A/Moscow/10/1999 (MO99)	57Q, 137S	(Lin *et al.*, 2004)
A/Panama/2007/1999 (PA99)	144N, 172E, 192I	(Lin *et al.*, 2004)
A/Fujian/411/2002 (FU02)	50G, 75Q, 83K, 131T, 155T, 156H, 186G	(Hay *et al.*, 2003)
A/California/07/2004 (CA04)	145N, 159F, 189N, 226I, 227P	(Hay *et al.*, 2005)
A/Wisconsin/67/2005 (WI05)	193F	(Hay *et al.*, 2006)
A/Brisbane/10/2007 (BR07)	50E, 140I	(Hay *et al.*, 2007)

3.4 Results

3.4.1 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 3.2**) 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 towards isolates with unusual virulence or other atypical properties is reduced (Ghedin *et al.*, 2005) (**Supplementary Figure 3.9**).

The AD plot for HA of the human H3N2 virus (**Figure 3.3, Supplementary Figure 3.7**) 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 allele 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 hemagglutinin 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) variant was

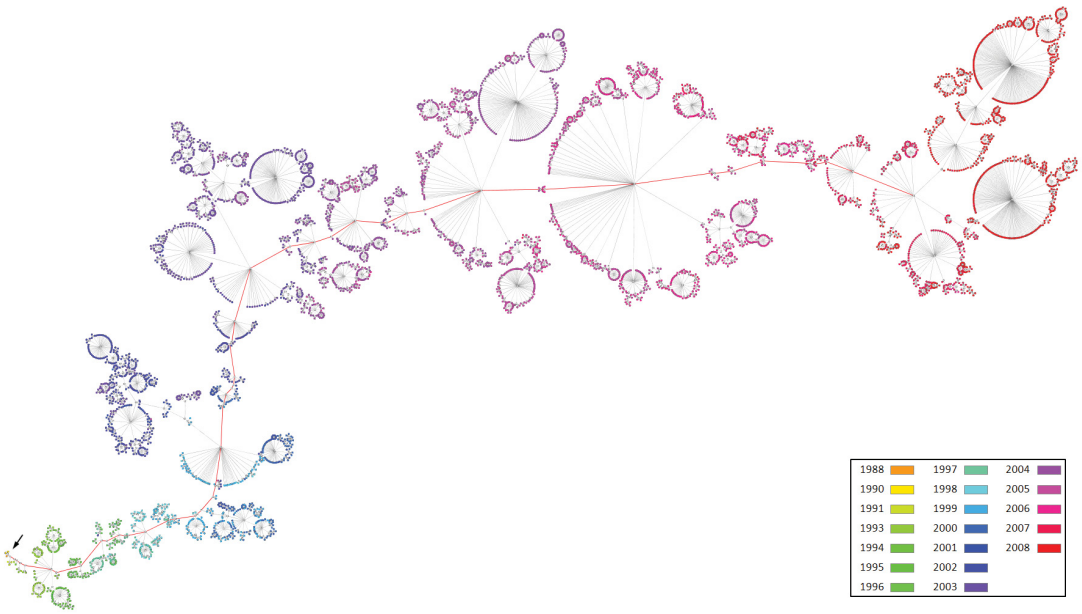


Figure 3.2: Maximum likelihood tree topology inferred for 4,913 hemagglutinin sequences of seasonal human influenza A (H3N2). Leaf nodes are color-coded according to the sampling dates of the viral isolates. The first sampled isolate, A/Siena/3/1988, is indicated with an arrow. The trunk of the tree (i.e. the path from the root to the most recent clade) is colored in red.

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, (WHO, 1999b, 2000a, 2001a, 2002a) appeared successively and rose to high frequencies without reaching fixation.

Most of the alleles rising to fixation (colored in **Figure 3.3**) are associated with substitutions reported in the literature (Lin *et al.*, 2004; Hay *et al.*, 2003, 2005, 2006, 2007) for the five distinct strains that represent predominant antigenic variants in the analysis period (**Table 3.1**). 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 3.9**), which does not allow the order in which the PA99 substitutions were acquired by

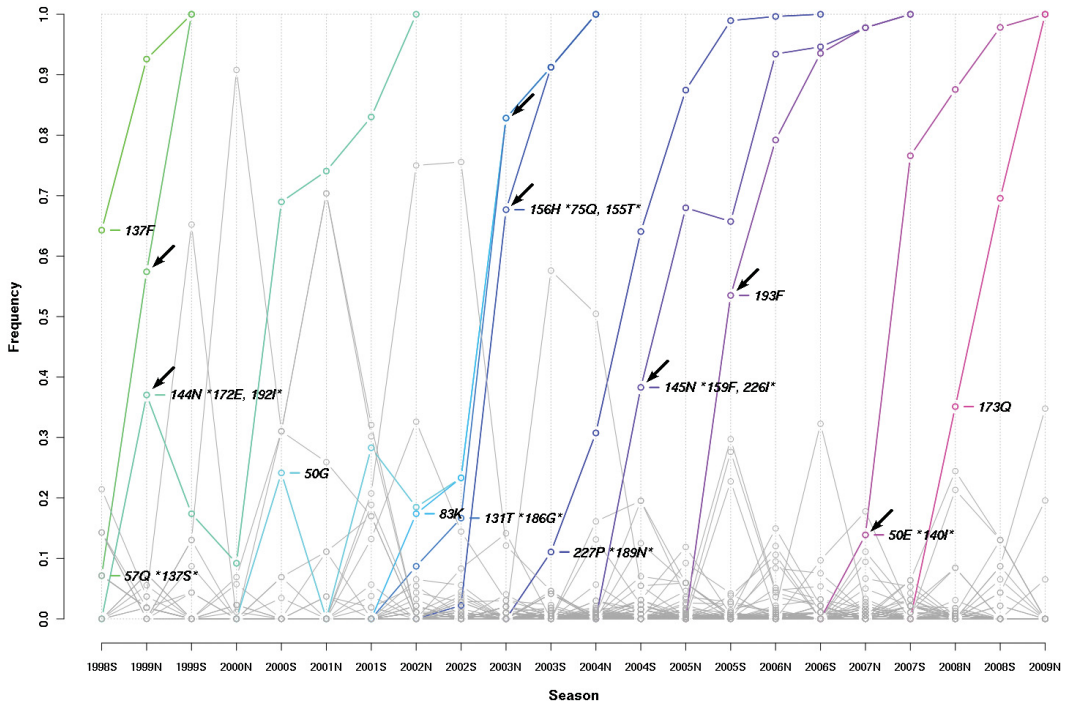


Figure 3.3: Allele dynamics plot for the major surface protein and antigenic determinant of the seasonal influenza A (H3N2) virus. The Northern and Southern influenza seasons from 1998 to 2008 are shown. Alleles that reach a prevalence of more than 95% and are subsequently fixed are shown in color; all 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 (Nobusawa *et al.*, 1991). Alleles that rise most quickly in frequency and are of interest with respect to vaccine strain selection are indicated by arrows.

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 hemagglutinin plot shows that of the two relevant substitutions, 140I was acquired first, followed by 50E.

Table 3.2: Alleles and their associated antigenic phenotypes with the steepest slopes in the seasons when they are predicted to become predominant. 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.

Season	Alleles	Slope	Antigenic variant	WHO	Predominant
1998/99 North	*57Q *137S**	0.5027	MO99	SY97 (WHO, 1999a)	MO99/PA99 (WHO, 2000a)
	*144N *172E, 192I**	0.3704	PA99		
2002 South	*155T *75Q**	0.0833	FU02	MO99 (WHO, 2002a)	FU02 (WHO, 2003b)
	*131T *186G**	0.0797	FU02		
	83K	0.0594	HK02/FU02		
	50G	0.0485	HK02/FU02		
2002/03 North	*131T *186G**	0.6616	FU02	FU02 (WHO, 2003a)	FU02 (WHO, 2004a)
	*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 (WHO, 2004a)	CA04 (WHO, 2005b)
	*227P *189N**	0.3331	WE04/CA04		
2005 South	*193F*	0.5350	WI05	CA04 (WHO, 2005b)	WI05 (WHO, 2006b)
2006/07 North	*50E *140I**	0.1389	BR07	WI05 (WHO, 2007a)	BR07 (WHO, 2008a)

3.4.2 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 3.1**). 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 3.2**). 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 (WHO, 1999b, 2000a,b, 2001a,b, 2002a,b, 2003a). 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 (WHO, 1999b). 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 1999/2000 Northern season (WHO, 2000a). 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 (WHO, 2003b, 2004a,b, 2005a), is associated with seven distinct substitutions: 50G, 75Q, 83K, 131T, 155T, 156H and 186G. 155T and 156H define the FU02 antigenic phenotype (Jin *et al.*, 2005). 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 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 (WHO, 2003a). 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 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 (WHO, 2005b, 2006a)

and was recommended as vaccine strain for the 2005/06 Northern season in the spring of 2005 (WHO, 2005a). 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 (WHO, 2006b, 2007a) and was recommended one season too late as the vaccine strain for the 2006/07 Northern season (WHO, 2006a). 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 (WHO, 2009b, 2007b, 2008a,b), 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 (WHO, 2007b). 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) (Pond *et al.*, 2008) 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 (Pond *et al.*, 2005) (data not shown). However, of the 20 epitope sites where changes rise to fixation over the analysis period (**Figure 3.2**), only 12 are detected by the DEPS method (**Supplementary Table 3.3**). 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 3.4**). 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 vari-

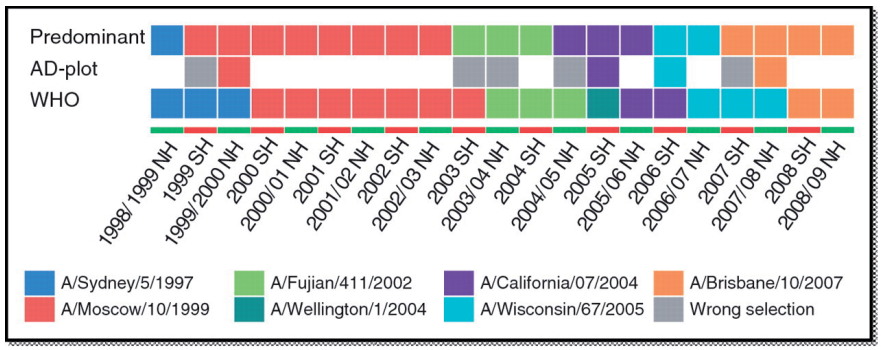


Figure 3.4: Comparison of predominant influenza A (H3N2) strains, WHO vaccine strain recommendations and strains identified by AD plot analysis. For the AD plot analysis, seasons with antigenic cluster transitions are shown in color. The information shown for the AD plot and the WHO recommendation represents the selection made one year earlier.

ant became predominant in the 2007 Southern season. In comparison to the WHO recommendations (WHO, 1999a,b, 2000a,b, 2001a,b, 2002a,b, 2003a,b, 2004a,b, 2005a,b, 2006a,b, 2007a,b, 2008a,b, 2009b, 2010b), 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 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 (Smith *et al.*, 2004; Fouchier and Smith, 2010). 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 (WHO, 1999b), recommending SY97 for the vaccine composition thus did not result in a dramatically lower vaccine efficacy.

3.4.3 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 (Russell *et al.*, 2008a). Above, we analyzed the data available only up to that point. If using 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 3.3, Supplementary Figure 3.8**). Previously circulating strains, on the other hand, occur at lower frequencies in comparison, as newly emerging antigenic variants increase in prevalence typically towards 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.

3.4.4 Evolutionary dynamics of the influenza A (H1N1) virus

We next studied the evolutionary dynamics of the 2009 influenza A (H1N1) virus, using 1,516 available exactly dated HA sequences (**Figure 3.5**). The virus has circulated in the human population only since April 2009 Garten *et al.* (2009); Smith *et al.* (2009); Dawood *et al.* (2009). Therefore, we have studied the evolutionary dynamics in monthly intervals (**Figure 3.6, Supplementary Figure 3.11**). As isolate A/California/05/2009 was the only one sampled in March, it was assigned to April 1st 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 (Fereidouni *et al.*, 2009), but have no known antigenic impact (Garten *et al.*, 2009). Furthermore, Pan *et al.* has already reported an increase in allele frequency for the S206T substitution among new H1N1 sequence isolates (Pan *et al.*, 2010).

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

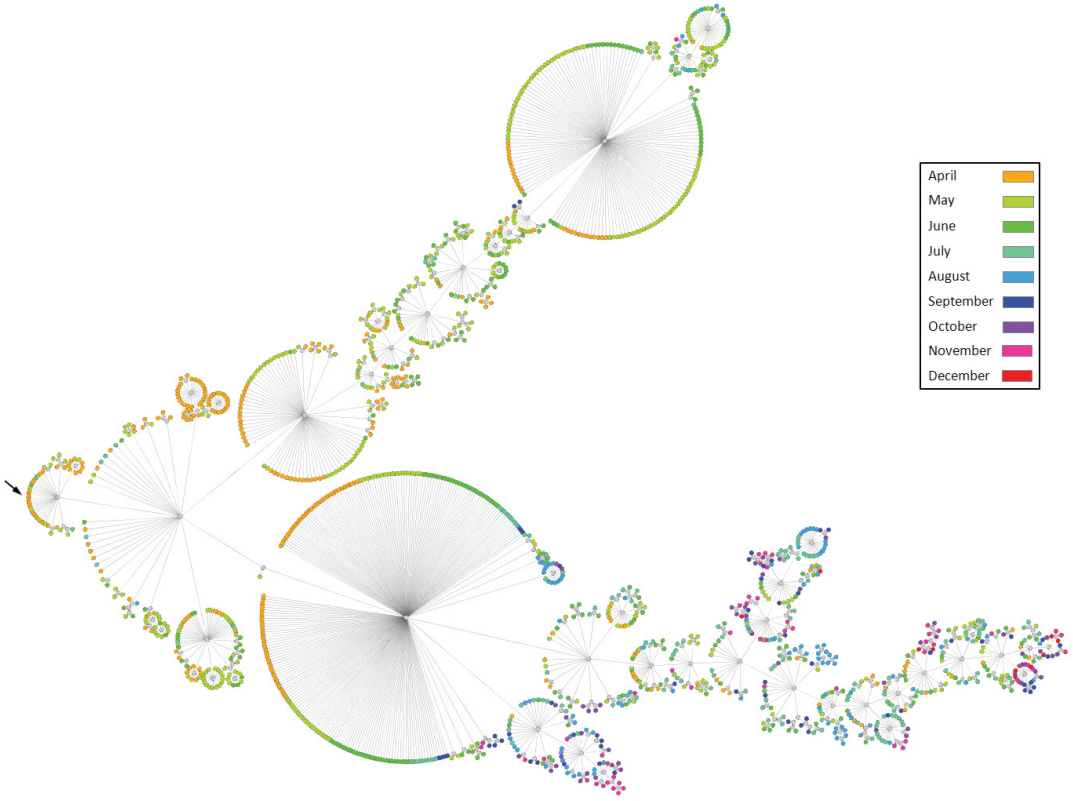


Figure 3.5: Maximum likelihood tree topology inferred from 1,516 2009 swine-origin influenza A (H1N1) hemagglutinin sequences. Leaf nodes are color-coded according to the sampling dates of the viral isolates. The first sampled isolate, A/California/05/2009, is indicated with an arrow.

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 (WHO, 2010b)]. 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 3.10**).

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 3.4**). This includes position 206, where a non-synonymous change has become fixed within the analysis period (220 in

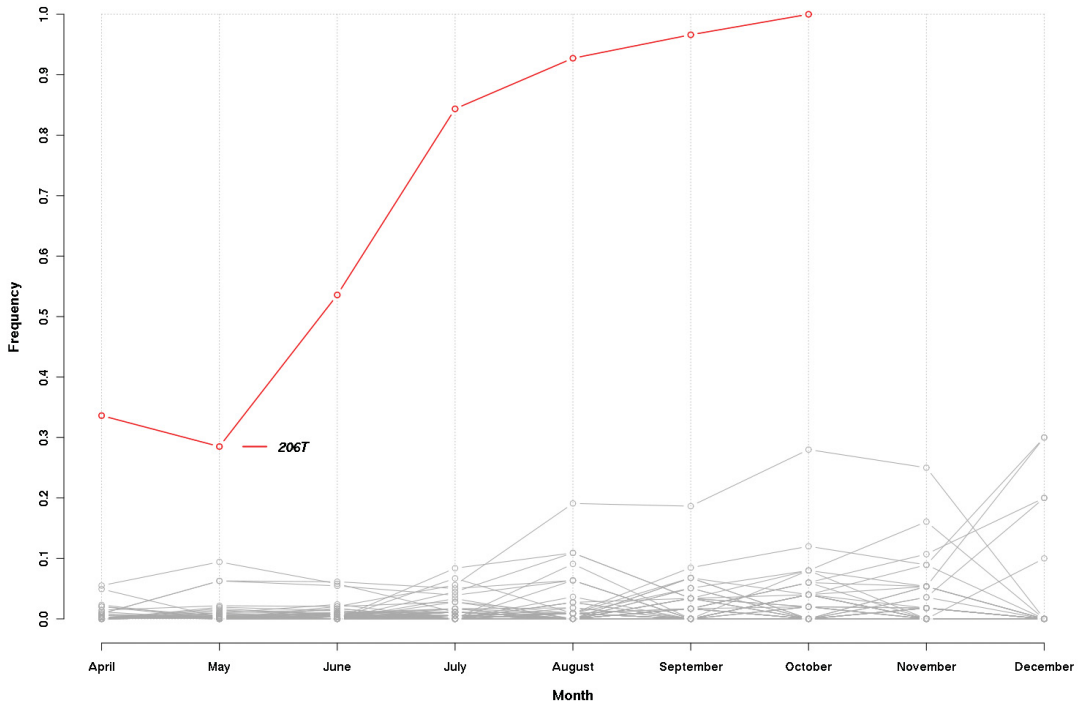


Figure 3.6: Allele dynamics plot for the major surface protein and antigenic determinant of the new influenza A (H1N1) based on sequences sampled between April and December of 2009 without allele frequency correction. Alleles that reach a prevalence of more than 95% and are subsequently fixed are shown in color; all other alleles are shown in gray. Substitutions are enumerated according to H3 HA1 numbering (Nobusawa *et al.*, 1991).

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 one 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.

3.5 Conclusion

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 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 evolutionary stable, even though data is limited.

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 viruses 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.

3.6 Supporting material

The supporting material comprises tables 3.3 and 3.4 and figures 3.7 to 3.11. For the sake of limited space and dimensionality the supporting tables S1 and S2 from the published article are not included in this thesis and can be accessed via the online version of the article (http://nar.oxfordjournals.org/content/early/2010/10/18/nar.gkq909).

Table 3.3: Test for directional evolution of protein sequences (DEPS) and traditional dN/dS test (FEL) on the influenza A (H3N2) hemagglutinin sequences and according phylogenetic tree under the influenza A substitution model. Only positions associated with substitutions that rise to predominance in Figure 2 are shown (HA1 numbering). If the empirical Bayes factor (DEPS EBF) is high enough (> 20) the position shows evidence for directional selection. Positions also detected to be under positive selection are shown in bold.

Site	Associated antigenic cluster	DEPS EBF	FEL dN/dS	FEL p-value
50	FU02, BR07	9.3 e+17	3.8346	0.0006
57	MO99	3.8 e+02	0.5690	0.1609
75	FU02	-	0.2954	0.0134
83	FU02	1.6 e+03	0.8736	0.8157
131	FU02	2.5 e+02	0.4001	0.0262
137	MO99	-	2.3345	0.2390
140	BR07	3.6 e+03	1.9657	0.3263
144	PA99	1.9 e+04	1.2223	0.5560
145	CA04	1.8 e+02	3.1626	0.0276
155	FU02	1.3 e+02	0.4341	0.0882
156	FU02	-	0.4491	0.0152
159	CA04	-	0.9656	0.9484
172	PA99	-	0.4315	0.1588
173	-	1.0 e+08	0.9277	0.8290
186	FU02	1.0 e+26	1.2644	0.4093
189	CA04	-	1.7099	0.3031
192	PA99	-	3.8748	0.0096
193	WI05	1.5 e+04	1.6220	0.2277
226	CA04	9.4 e+14	1.5711	0.0801
227	CA04	-	0.4398	0.0832

Table 3.4: Test for directional evolution of protein sequences (DEPS) and traditional dN/dS test (FEL) on the 2009 swine-origin influenza A (H1N1) hemagglutinin sequences and according phylogenetic tree under the influenza A substitution model. If the empirical Bayes factor (DEPS EBF) is high enough (> 20) the position shows evidence for directional selection. Positions also detected to be under positive selection are shown in bold. Positions 2, 8, 559, 561, 562 and 565 that were also detected by DEPS are neglected due to low sequence coverage. Positions are enumerated in H1 sequence numbering.

Site	DEPS EBF	FEL dN/dS	FEL p-value
106	3.5 e+02	1.1680	0.8983
220	3.5 e+04	inf	0.0028
239	2.4 e+06	1.2567	0.7142
240	1.0 e+09	inf	0.0231
278	2.2 e+07	inf	0.0159

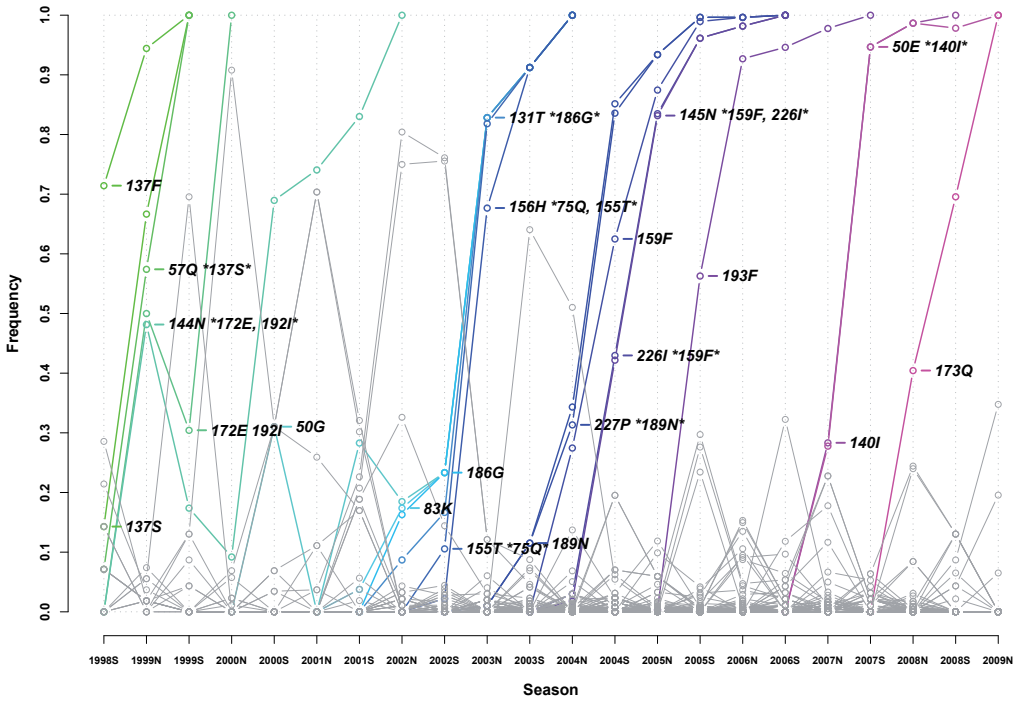


Figure 3.7: AD plot for major surface protein and antigenic determinant of seasonal influenza A (H3N2) virus in the Northern and Southern hemisphere influenza seasons from 1998 to 2008 without allele frequency correction. Alleles that reach a prevalence of more than 95% and are subsequently fixed are shown in color; all other alleles are shown in gray. Substitutions are restricted to occur in the five epitope sites and are enumerated according to HA1 numbering.

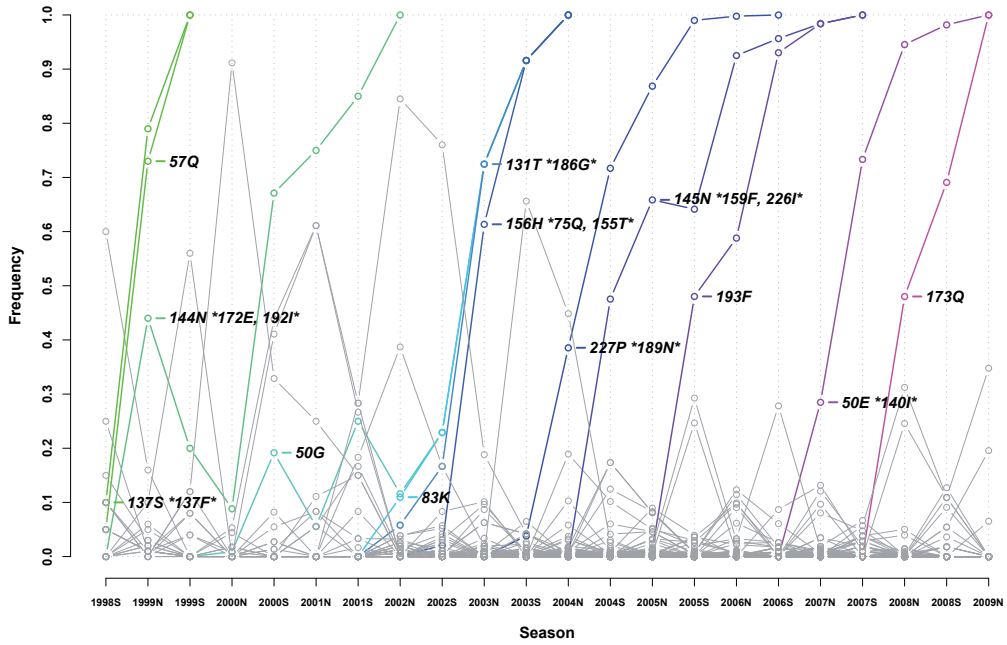


Figure 3.8: AD plot for major surface protein and antigenic determinant of seasonal influenza A (H3N2) virus in the Northern and Southern hemisphere influenza seasons from 1998 to 2008 with relaxed seasonal cutoff. Alleles that reach a prevalence of more than 95% and are subsequently fixed are shown in color; all other alleles are shown in gray. Substitutions are restricted to occur in the five epitope sites and are enumerated according to HA1 numbering. Alleles that rise most quickly in frequency and are of interest at a certain point of time are indicated by arrows.

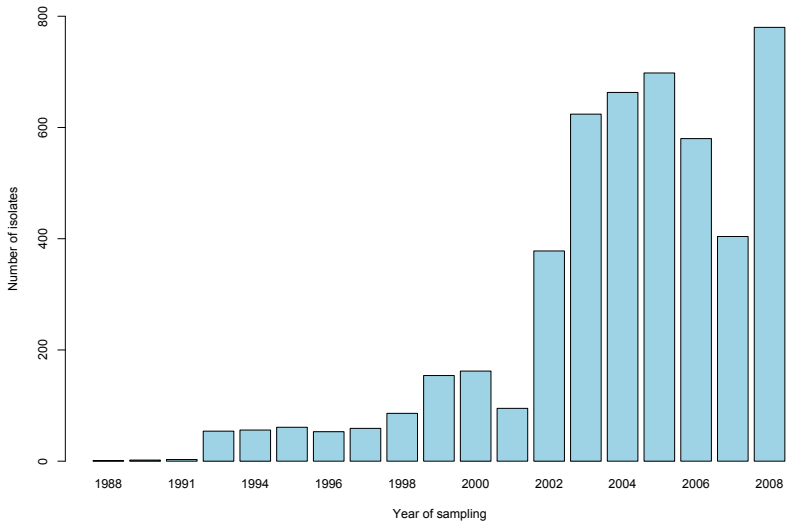


Figure 3.9: Sampling bias of the influenza A (H3N2) hemagglutinin sequences. Numbers of available isolate sequences by season from 1988 to 2008.

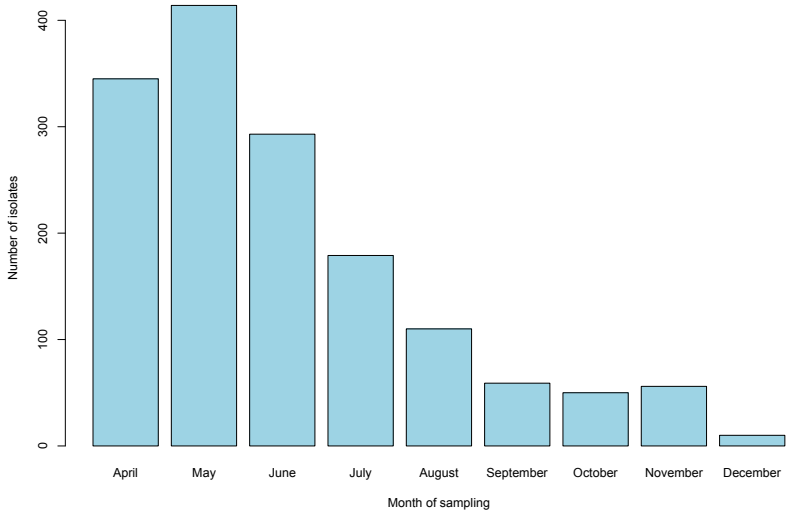


Figure 3.10: Sampling bias of the swine-origin influenza A (H1N1) hemagglutinin sequences. Numbers of available isolate sequences between April 2009 and December 2009.

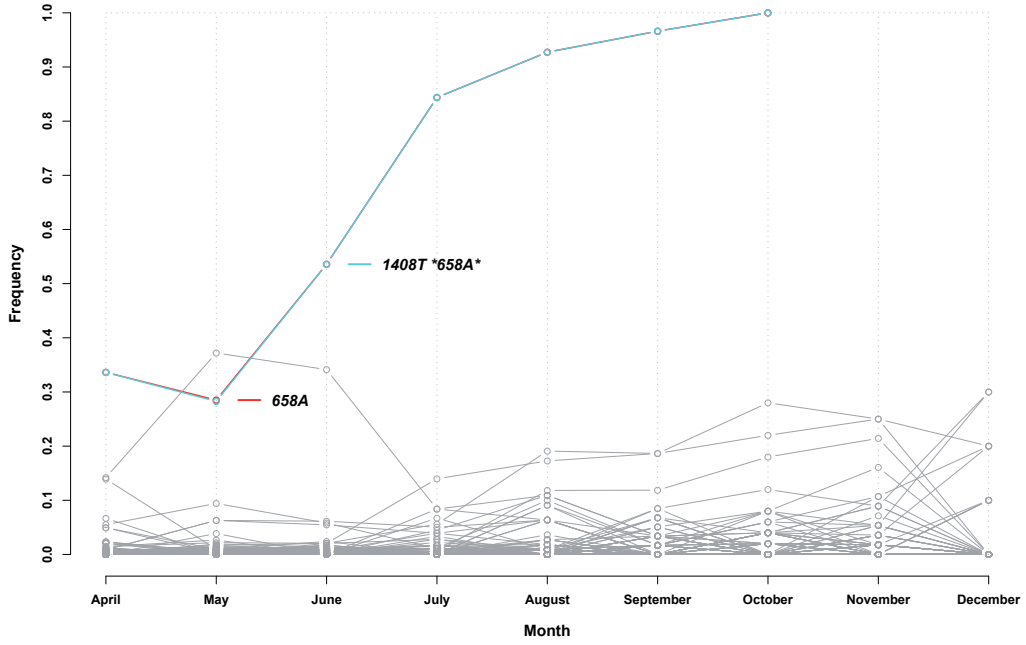


Figure 3.11: AD plot for the genomic segment of the major surface protein and antigenic determinant of the new influenza A (H1N1) based on genomic sequences sampled between April and December of 2009 without allele frequency correction. Alleles that reach a prevalence of more than 95% and are subsequently fixed are shown in color; all other alleles are shown in gray.

CHAPTER 4

Inference of genotype-phenotype relationships in the antigenic evolution of human influenza A (H3N2) viruses

Status	**published**
Journal	PLoS Computational Biology (Impact factor 5.515)
Citation	Steinbrück, L. and A. C. McHardy (2012). Inference of Genotype-Phenotype Relationships in the Antigenic Evolution of Human Influenza A (H3N2) Viruses. *PLoS Comput Biol*, **8** (4): e1002492.
URL	http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002492
Own contribution	75% Performed the experiments Analyzed the data (with co-authors) Wrote the manuscript (with co-authors)

4.1 Abstract

Distinguishing mutations that determine an organism's phenotype from (near-) neutral 'hitchhikers' is a fundamental challenge in genome research, and is relevant for numerous medical and biotechnological applications. For human influenza viruses, recognizing changes in the antigenic phenotype and a strains' capability to evade pre-existing host immunity is important for the production of efficient vaccines. We have developed a method for inferring 'antigenic trees' for the major viral surface protein hemagglutinin. In the antigenic tree, antigenic weights are assigned to all tree branches, which allows us to resolve the antigenic impact of the associated amino acid changes. Our technique predicted antigenic distances with comparable accuracy to antigenic cartography. Additionally, it identified both known and novel sites, and amino acid changes with antigenic impact in the evolution of influenza A (H3N2) viruses from 1968 to 2003. The technique can also be applied for inference of 'phenotype trees' and genotype-phenotype relationships from other types of pairwise phenotype distances.

4.2 Author summary

The molecular evolution of any organism is described by changes in the genotype resulting from genetic drift or selection to maintain or establish fitness under the given environmental conditions. Identification of phenotype-defining changes and their distinction from (near-) neutral ('hitchhikers') ones is a fundamental challenge in genome research. The standard approach involves time- and cost-intensive mutation experiments, which are typically low throughput, due to their experimental nature. We have developed a computational method for the inference of phenotypic impact of genotypic changes that is applicable to any system, within or across species, where homologous genetic sequences and associated pairwise phenotype distances are available. We demonstrate the accuracy of our method by application to the human influenza A (H3N2) virus. This exemplary system is of particular interest, as recognizing changes in the antigenic phenotype and a viral strains' capability to evade pre-existing host immunity is important for the production of efficient vaccines. We accurately identified known sites and amino acid changes with antigenic impact over 35 years of evolution, and provide further details on individual antigenically relevant changes in the evolution of influenza A (H3N2) viruses.

4.3 Introduction

Influenza viruses are responsible for ∼500,000 deaths annually and are a substantial threat to human health (WHO, 2009a). Besides seasonal infections caused by human viruses, four major pandemics over the last 100 years have resulted in ∼50 million deaths worldwide (Tognotti, 2009; Taubenberger and Morens, 2006; WHO, 2010a). The viruses are classified into three genera (A, B, C), all from the *Orthomyxoviridae* family, which comprises single-stranded, negative sense RNA viruses. Influenza A and B viruses evolve rapidly and continuously accumulate amino acid changes in the antibody-binding (epitope) sites of the surface proteins, resulting in changes in antigenicity. Thus, novel 'antigenic types' regularly appear and rise to predominance, causing worldwide epidemics despite existing vaccination programs (Smith *et al.*, 2004; Nelson *et al.*, 2007). Influenza A viruses are further categorized into subtypes based on the composition of their surface proteins, hemagglutinin (H or HA) and neuraminidase (N or NA). In the human population, the subtypes H1N1 and H3N2 are currently circulating (WHO, 2011c). Both global population structure and geographic migration patterns are known to influence the evolution of H3N2. Russell *et al.* suggested East-Southeast Asia to serve as a global reservoir, from which seasonal epidemics in temperate zones are seeded (Russell *et al.*, 2008c). Other regions, such as China or USA, might serve as seeding regions, too, and migration from and to other tropical regions than East-Southeast Asia is thought to have a significant influence on the global dynamics (Bedford *et al.*, 2010; Bahl *et al.*, 2011).

To monitor genetic and antigenic changes, the World Health Organization (WHO) runs a global surveillance program (Russell *et al.*, 2008a). Quantification of viral antigenic phenotypes is done with the hemagglutination inhibition (HI) assay, which measures the ability of an antiserum to inhibit the agglutination of red blood cells by a viral antigen (Hirst, 1943). Antigenic cartography, involving multidimensional scaling of log-normalized HI titers, subsequently generates an accurate low-dimensional representation of the antigenic distances between antigen-antiserum pairs (Lapedes and Farber, 2001; Smith *et al.*, 2004). If a novel antigenic type with increasing prevalence is detected, the vaccine composition, consisting of two strains of influenza A (H3N2 and H1N1) and one strain of influenza B, is updated to include an antigenically closer match.

Antigenic cartography of influenza A (H3N2) isolates from 1968 to 2003 revealed that antigenic types circulate for 3.3 years, on average, in worldwide epidemics before being replaced by a successor (Smith *et al.*, 2004). A comparison of antigenic and genetic

maps showed that, the antigenic impact of genetic changes varies, depending on the nature of the amino acids exchanged, their structural positioning and epistatic interactions with other sites. Subsequent studies have incorporated both antigenic and genetic data for predicting antigenically novel strains (Lee *et al.*, 2007; Liao *et al.*, 2008; Huang *et al.*, 2009). Additionally, many groups have investigated the influence of sequence positions and sequence variation on viral evolution, based on different computational criteria (Bush *et al.*, 1999b,a; Plotkin *et al.*, 2002; Shih *et al.*, 2007; Du *et al.*, 2008; Pond *et al.*, 2008; Xia *et al.*, 2009; Steinbrück and McHardy, 2011).

Even though the general principles governing the antigenic evolution of influenza A viruses are well studied, computational methods for directly determining the antigenic impact of individual amino acid exchanges do not yet exist. Such analyses currently require time- and cost-intensive experimental characterization of mutant viruses (Smith *et al.*, 2004). On the other end of the spectrum, antigenic cartography allows identification of 'cluster difference substitutions', comprising all near-conserved changes that distinguish consecutive antigenic clusters.

We describe a method for the inference of 'antigenic trees', which is based on a least-squares optimization (LSO) procedure of fitting pairwise antigenic distances onto an evolutionary tree for the major antigenic determinant of influenza A. It is a computational method allowing for a more fine-grained resolution of the antigenic impact of individual changes than antigenic cartography without time- and cost-intensive experiments. Application to HA sequences and serological data from human influenza A (H3N2) viral isolates from 1968 to 2003 determined the antigenic impact of all branch-associated amino acid changes for this time period. Our technique identified known antigenic types and the amino acid changes associated with the type transitions. For sufficiently resolved branches, the antigenic impact of individual exchanges could be quantified. The method furthermore found known and novel key HA sites and changes in antigenic evolution.

4.4 Results

We applied our method to infer an antigenic tree from genetic sequences of the hemagglutinin segment and serological data (HI titers of antigen-antiserum pairs) for 258 influenza A (H3N2) isolates sampled between 1968 and 2003 (Smith *et al.*, 2004). Antigenic branch lengths were determined by fitting the antigenic distances between viral isolates (the antigens) and antisera raised against reference strains to the branches of a maximum likelihood tree (see Materials and Methods). Antigenic branch lengths

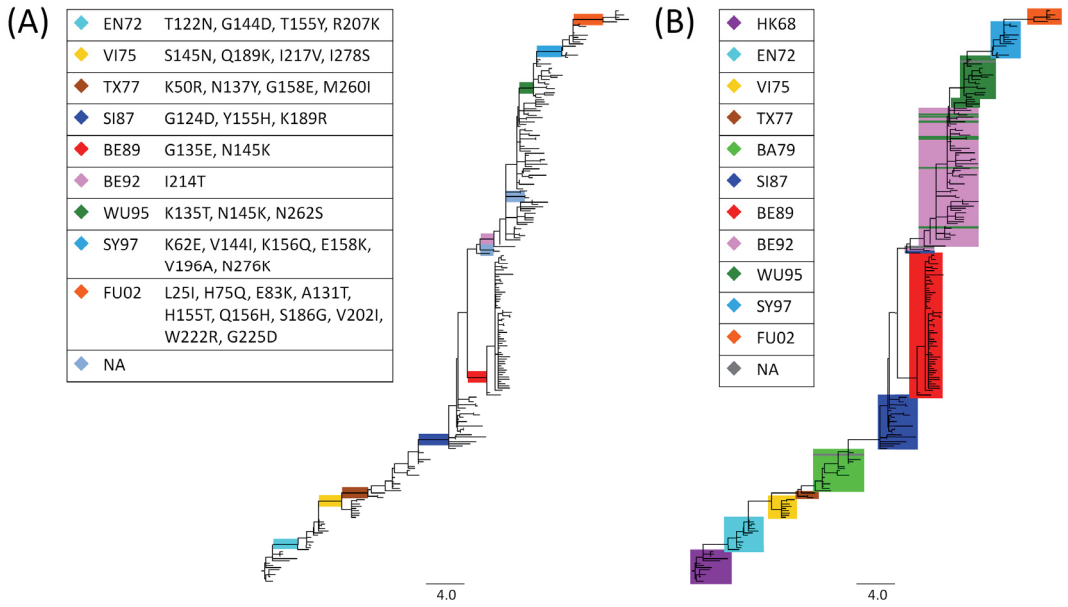


Figure 4.1: Antigenic tree for influenza A (H3N2) viruses. Branch lengths represent antigenic distances (maximum of up- and down-weights for each branch) inferred from a maximum likelihood tree of 258 hemagglutinin sequences of seasonal influenza A (H3N2) virus isolates and serological data. (A) Colored edges show antigenic type transitions, with internal branches with high average antigenic weights ($\geq$ 1.0 antigenic units). Gray-blue edges represent high weight branches leading to a subtree with three isolates or less, representing low abundance types. (B) Isolates are color-coded by antigenic clusters according to Smith *et al.* (2004). Three isolates (A/Christchurch/4/85, A/Hong Kong/34/90 and A/Netherlands/172/96) are only present as antisera and were not assigned a cluster label.

were realized as two independent weights (up and down) and represented the antigenic properties of antigens and antisera in the tree. The antigenic path length between two isolates, corresponding to the sum of the branch weights (either up- or down-weight, depending on the direction in the tree) for all connecting branches on the path between them in the tree, reflected their overall antigenic distance (**Figure 4.1**).

To investigate how accurately antigenic distances were fitted onto the tree, we evaluated its ability to predict unseen antigenic distances by leave-one-out cross validation (Hastie *et al.*, 2004). In this experiment, an antigenic tree is inferred from all but one antigenic distance and then is applied to predict the left out distance. A predicted distance corresponds to the antigenic path length between the two respective isolates in the tree (see Materials and Methods). This was repeated for every antigenic distance and the overall accuracy of predicting antigenic distances estimated by the absolute

prediction error and the root mean squared error (RSME) averaged over all leave-one-out experiments (see Materials and Methods). The leave-one-out absolute prediction error was 0.86 antigenic units (a two-fold dilution, SD 0.72) and the correlation measured by Pearson's correlation coefficient between predicted and measured values was 0.86. Using placement on an antigenic map estimated from the same data, Smith *et al.* report antigenic an average absolute prediction error of 0.83 antigenic units (SD 0.67) and a Pearson's correlation coefficient of 0.80 for 481 measurements of antigenic distances (Smith *et al.*, 2004). The root mean squared error (RMSE) penalizes large prediction errors more than small prediction errors, and is a well suited measure of predictive accuracy. For our method, the leave-one-out RMSE is 1.12 antigenic units, corresponding to approximately a two-fold dilution. This is comparable to the ten-fold cross validation RMSE of Cai *et al.* on this data set (1.05 antigenic units) (Cai *et al.*, 2010), who used a matrix completion algorithm prior to multi dimensional scaling. Our method therefore performs similarly to antigenic cartography in predicting antigenic distances, with a slightly larger error but also a slightly higher correlation between predicted and measured values. This is despite the fact that inferring antigenic branch lengths for an antigenic tree allows far fewer degrees of freedom than an antigenic map, where the data is not forced on a fixed structure. Note that for the prediction of antigenic distances, other well-suited methods also exist (Cai *et al.*, 2010; Ndifon, 2011).

As we infer a tree topology from nucleotide sequences, branches might be without any amino acid changes and thus lack explanatory power if they are assigned antigenic weights. This allows accommodating measurement errors in HI titers in antigenic branch weights or variation caused by changes in other viral antigens, such as the surface glycoprotein neuraminidase. HI titers are imprecise, as they reflect two-fold dilutions instead of quantitative estimates, and are often highly variable, with measurements varying between experiments and laboratories. For instance, the two isolates A/Finland/220/92 and A/Stockholm/20/91 have the same nucleotide sequence, and hence no changes on their respective tip branches (tips), but differ strongly in their HI values, where A/Finland/220/92 shows an antigenic distance from the same antisera that is, on average, 1.0 antigenic units (a two-fold dilution) larger than that of A/Stockholm/20/91. Note that, in general, even though neuraminidase may influence the HI titers, the WHO recommends application of the HI assay under conditions where its influence is negligible (Network, 2011). To incorporate a possible influence of neuraminidase activity one may use concatenated viral sequences (hemagglutinin and neuraminidase) and fit antigenic distances on a tree topology inferred from these se-

quences. If doing so, one should first ensure that reassortment events have not resulted in larger topological changes between the HA and NA genealogies during the analyzed time period (Nelson *et al.*, 2006, 2007). In case of larger topopological changes due to segment reassortment, a joint tree is inferred for data which cannot be described by a tree-like evolutionary history, overall, and the results are likely to be only partially informative.

On average, internal branches without amino acid changes have weights of 0.30 (up) and 0.21 (down), respectively. Less noise occurs on the tree trunk, which represents the viral lineage surviving over time, with 0.19 (up-weight) and 0.19 (down-weight) assigned, on average. Interestingly, the average antigenic weight of branches with amino acid changes is higher on the tree trunk than for all internal branches (up = 0.52, down = 0.61 vs. up = 0.44, down = 0.46). This is in agreement with an expected fitness advantage for viral isolates with larger antigenic changes, and therefore preferential fixation and establishment appear as changes on the tree trunk.

4.4.1 Antigenic types resolved in the tree

Antigenic types are clearly distinguished by high average weights ($\geq$ 1.0 antigenic units) in the antigenic tree (see Materials and Methods). Exclusion of branches leading to subtrees with three or less isolates, representing undersampled groups, identified nine branches defining type transitions (**Table 4.1**) and ten antigenic types. Abbreviations for these (HK68, EN72, VI75, TX77, SI87, BE89, BE92, WU95, SY97 and FU02) are used as in Smith *et al.* (Smith *et al.*, 2004). SY97, for instance, denotes antigenically similar A/Sydney/5/1997-like strains. The average antigenic distances of these branches range from 1.0 (SI87-BE89) to 2.6 antigenic units (WU95-SY97; **Table 4.1**, **Figure 4.1A**). Eight of the nine type transition branches are on the trunk of the tree, which represents the influenza A (H3N2) lineage surviving over time. An exception is BE89, which is located in a subtree that has become extinct.

Table 4.1: Internal branches with high average antigenic weights ($\geq$ 1.0 antigenic units) and according antigenic types in comparison to antigenic clusters identified by antigenic cartography (branches leading to three or less isolates are excluded). Branch amino acid changes indicate the corresponding branches, where changes in bold were also found by Smith *et al.*, and weights give the respective up, down and average branch weights. Multiple branches that can be mapped to a single antigenic type are separated by dashed lines. Additional amino acid changes indicate branches that carry further mutations found to be cluster transition substitutions by Smith *et al.*. For some branches, the down-weight was not defined, as no antiserum was in the respective subtree Branches that can be mapped to multiple type transitions are shown at the first mapping only. Smith *et al.* present average distances between consecutive antigenic clusters, whereas average antigenic branch weights give a minimum distance between consecutive antigenic types. Note that on branches with multiple changes not all changes have to contribute to the antigenic weight, though their individual impacts could not be resolved with the dataset (unsampled viral isolates).

Type transition	Branch amino acid changes	Weights (up/down/avg)	Trunk	Additional amino acid changes	Weights (up/down/avg)	Trunk Trunk	Smith et al.
HK68-EN72	**T122N, G144D, T155Y, R207K**	2.6/0.4/1.5	x	L3F, **N188D**	0.9/0.2/0.5	x	3.4
EN72-VI75	**S145N, Q189K, I217V, I278S**	0.6/2.4/1.5	x	**N53D, N137S, L164Q, F174S, N193D, R201K, I213V, I230V**	0.0/1.0/0.5		4.4
VI75-TX77	**K50R, N137Y, G158E, M260I**	0.6/2.8/1.7	x	**E82K**	1.0/-/0.5		3.4
TX77-BA79				**N133S, P143S, G146S, K156E, T160K, Q197R, V217I**	1.4/0.0/0.7	x	3.3
				D2N, **N53D, N54S, I62K, D172G, V244L**	0.0/0.3/0.2	x	
BA79-SI87	**G124D, Y155H, K189R**	0.2/3.3/1.7	x				4.9
SI87-BE89	G135E, **N145K**	2.0/0.0/1.0					4.6
BE89-BE92	I214T	1.4/1.1/1.3	x	**E156K, E190D**, N193S, L226Q, **T262N**	1.0/0.0/0.5	x	7.8
				S133D	0.0/0.4/0.2	x	
BE92-WU95	K135T, **N145K**, N262S	1.5/1.1/1.3	x				4.6
WU95-SY97	**K62E**, V144I, **K156Q, E158K V196A, N276K**	2.5/2.6/2.6	x				4.7
SY97-FU02	**L25I, R50G, H75Q, E83K, A131T, H155T, Q156H**, S186G, **V202I, W222R, G225D**	1.8/3.2/2.5	x				3.5

The setting of the threshold parameter for identification of antigenic types in the tree influences the performance of our method (**Supplementary Table 4.4**). The selected threshold of 1.0 antigenic unit identified nine of ten antigenic type transitions found by antigenic cartography (Smith *et al.*, 2004). The TX77-BA79 transition was not predicted with our method in this setting, as the weights of the corresponding branch were slightly below the threshold (up-weight 1.4, down-weight 0.0). Our method resolves antigenically relevant changes between successive antigenic types in several cases to several successive branches. Therefore, a higher threshold of 2.0 antigenic units for individual branches (a four-fold dilution), as suggested to distinguish antigenically diverse viral strains (Russell *et al.*, 2008a), does not allow distinction between different antigenic groups (only if the transition is not well resolved in the data and the antigenic impact of multiple changes is summarized on a single branch). On the other hand, choosing a lower threshold of 0.5 antigenic units selects twelve additional type-defining branches (**Supplementary Table 4.4, Supplementary Figure 4.3**). Among these is the TX77-BA79 type-defining branch that corresponds to an antigenic cluster transition according to antigenic cartography (Smith *et al.*, 2004). Furthermore, four of these additional branches define antigenic subtypes that were distinct enough to warrant a vaccine update. A more detailed discussion of type-defining branches at the threshold of 0.5 antigenic units can be found in the supporting material (**Section 4.7.1**). Note, that the choice of the threshold distance is equivalent to find a minimal antigenic distance to distinguish groups of antigenically and genetically similar viral isolates. This is different from the question whether two specific viral isolates are antigenically similar or not, although both tasks are related to each other.

For the nine jointly identified type transitions, seven agree 100% in terms of the assigned viral isolates. For the BE89-BE92 transition, the isolate A/Netherlands/938/1992 is placed within BE92 using antigenic cartography and as preceding BE92 by our technique. Isolate assignment differs the most for the BE92-WU95 transition. This is likely to be caused by multiple occurrences of N145K, which is, according to Smith *et al.* (Smith *et al.*, 2004), the change that defines the BE92-WU95 transition and has a major antigenic impact in that context (2.6 antigenic units). It was already noted by Smith *et al.* that isolates classified by antigenic cartography within WU95 are placed in the vicinity of BE92 in a tree. Our analysis agrees with these findings (**Figure 4.1B**). We found that for each branch adjacent to these disagreeing placements, N145K is present (isolates of the antigenic type WU95 located in the area of BE92), with large branch-associated antigenic weights (an average up-weight of 1.3), similar to the type-defining branch of WU95 (up-weight 1.5). This indicates that N145K has a large antigenic

impact for all these isolates and, interestingly, was evolutionary volatile during that period.
Analysis of up- and down-weights for type-defining branches allows us to determine a direction for antigenic impact. For example, the branch separating HK68 and EN72 has a weight of 2.6 (up)/0.4 (down), which means that isolates of HK68 are antigenically more similar to sera raised against EN72 than vice versa. The opposite example represents the SY97-FU02 transition, where the corresponding branch weight is 1.8 (up)/3.2 (down), which means that SY97 isolates are more distant from antisera raised against FU02 than vice versa. Both examples are in agreement with results published by the WHO (WHO, 1972, 2003a).
As influenza A evolution in the analyzed data set is characterized by an underlying cluster structure, both antigenic types and antigenic clusters allow determination of cluster-difference or antigenic type associated substitutions. However, antigenic types (inferred by our method) and antigenic clusters (inferred by antigenic cartography) have different interpretations. Antigenic types represent sets of viral isolates showing similar evolutionary (defined by the phylogenetic tree) and antigenic (defined by the antigenic branch lengths) patterns. Antigenic cluster are solely defined by antigenic patterns and are determined by a k-means clustering approach. In datasets with less well-defined cluster structure, the k-means approach would hardly result in robust clusters and identification of phenotype-associated changes would be more difficult, whereas our method would likely be able to resolve phenotype-genotype relationships up to the level of resolution supported by the data.

4.4.2 Substitutions in antigenic type transitions

Amino acid changes from eight of nine type transitions identified by both antigenic cartography and the antigenic tree include the cluster difference substitutions described in Smith *et al.* (Smith *et al.*, 2004) (**Table 4.1**). Smith *et al.* define 'cluster difference substitutions' as changes in conserved residues between two consecutive antigenic clusters (conserved meaning present in at least n - 1 isolates within a cluster of size n). For five transitions, all cluster difference substitutions are on the type-defining branch (BA79-SI87, SI87-BE89, BE92-WU95, WU95-SY97 and SY97-FU02). For three transitions (EN72-VI75, VI75-TX77 and HK68-EN72), the substitutions were resolved to several branches with different antigenic branch weights, which allows a more fine-grained distinction. The 12 substitutions of the EN72-VI75 transition were assigned to two consecutive branches, one with high and one with moderate antigenic weights. The

branch with S145N, Q189K, I217V and I278S has a high antigenic weight, indicating that one or several of these have a very large antigenic impact. For the HK68-EN72 and the VI75-TX77 transitions, the substitutions were resolved to two consecutive branches with high and moderate antigenic weights, too.

For BE89-BE92, the amino acid changes differ from cluster difference substitutions. Here, the cluster difference substitutions are found on branches that precede the type-defining branch. The type-defining branch carries the change I214T, while the cluster difference substitutions map to two preceding branches with lower antigenic weights. I214T has not been mentioned in the literature before and is reversed downwards in the tree on a branch without any assigned antigenic weight. Thus, either the measurements here were too noisy to resolve the correct branch, or this position has an antigenic impact as an epistatic effect, allowing for the preceding changes to become antigenically effective. Support for a potential epistatic effect of this change can be found by detailed analysis of individual HI measurements for two isolates (A/Hong Kong/34/1990 and A/Netherlands/938/1992), which already have the preceding branch changes for BE92 but not the I214T change. On average, all antigens labeled BE92 by Smith *et al.* have a large antigenic distance (greater 4.7 antigenic units) from the antiserum A/Hong Kong/34/1990. A/Netherlands/938/1992 is similar to A/Hong Kong/34/1990, with an antigenic distance of 0.7 to this antiserum.

Four branches with type transitions (SI87-BE89, BE92-WU95, WU95-SY97 and SY97-FU02) include additional changes besides the cluster difference substitutions. For SI87-BE89, the change G135E is present, in addition to N145K. G135E appears twice more in the tree, with an average up-weight of 0.64. This indicates that it may also have an antigenic effect in SI87-BE89. For BE92-WU95, the changes K135T and N262S are present on the type-defining branch, in addition to N145K. Both are located in the antibody binding sites (Wiley *et al.*, 1981) and became fixed following their appearance on this trunk branch.

In a recent (unpublished) study, Koel *et al.* (*Koel et al.; Antigenic evolution of influenza A (H3N2) virus is dictated by 7 residues in the hemagglutinin protein; 2nd International Influenza Meeting, Münster; 2011*) determined by site-directed mutagenesis changes at seven positions in the HA protein (145, 155, 156, 158, 159, 189 and 193) responsible for significant phenotypic diversity in the evolution of influenza A (H3N2). We also find that for eight of the nine identified type-defining branches changes occur at five of these positions (no changes at positions 159 and 193 are involved in antigenic type transitions), which further confirm the relevance of these sites for antigenic evolution (**Table 4.1**). Note that, besides these five residues changes at 23 other positions

map to the type-defining branches which not all have to contribute to the antigenic weight, though their individual impacts could not be further resolved with the dataset (unsampled viral isolates).

4.4.3 Antigenic impact of individual amino acid changes and sites

We examined amino acid changes with strong antigenic relevance according to (i) the impact of all changes at a specific site and (ii) the impact of a specific change. In the first case, we determined all positions where at least three changes occurred, and the mean and median of the branch weights (up- or down-weight) were not less than one antigenic unit. Missing weights, e.g. where down-weights were not defined because no antiserum was raised for the corresponding subtree, were excluded from the calculations. Seven positions, 112, 137, 144, 155, 156, 189 and 208, satisfy these criteria (**Supplementary Table 4.2**). All except position 112 are part of the antibody binding sites of HA1 (Wiley *et al.*, 1981). Positions 137, 155 and 156 are also part of the receptor binding site (Wilson *et al.*, 1981). Positions 155 and 189 may be particularly important, as all changes occur on the tree trunk and are part of type transitions. The importance of H155T and Q156H was also verified for the FU02 transition (Jin *et al.*, 2005). For positions 137, 144, and 156, several changes map to the tree trunk (three of six, four of nine, and one of three, respectively), indicating their antigenic relevance. Changes at position 112 explain single isolate variations, as all occur on tips. The antigenic impact of these changes may be due to hitchhiking effects, as they occur only in combination with other changes.

Next, we identified changes occurring at least three times in the tree with a mean and median antigenic weight (up- or down-weight) of more than one unit (**Supplementary Table 4.3**). Again, missing weights were excluded from the calculations. Five changes satisfy these conditions. Four of these (K62E, N145K, L226Q and T248I) occur at positions in antibody binding sites (Wiley *et al.*, 1981). N145K was experimentally verified to have a large antigenic impact (Smith *et al.*, 2004). K62E is part of the WU95-SY97 transition and has a high weight assigned on two further tips. Finally, of the eight occurrences of L226Q, seven appear between 1990 and 1996 for isolates of the BE92 type, indicative of a fitness effect for this antigenic type in particular. Interestingly, the reverse change, Q226L, is known to play a role in receptor binding specificity for the adaptation of bird viruses to the human host (Matrosovich *et al.*, 2000; Kawaoka, 2006; Bateman *et al.*, 2008; Wan *et al.*, 2008). T248I had a high weight only in combination with other changes, indicating a potential epistatic effect. Besides

these four changes, we identified V112I, which only appeared on tips and explains single isolate variations.

We searched for changes with moderate antigenic impact (more than 0.5 antigenic units) which identified seven further changes.G135E is part of the SI87-BE89 transition (see above) and E156K was shown to impact immune escape in mice (Hensley *et al.*, 2009). Both are located in the antibody binding sites (Wiley *et al.*, 1981). For several additional changes, the importance was not immediately obvious, as they (i) occurred only in combination with other changes, (ii) exhibited a high weight only in combination with other changes (Q80K), (iii) only appeared on tips (S186I, S199P and V226I) or (iv) had high weights assigned only on tips and low weights on internal branches (A138T). In cases (i) and (ii), this may be the result of epistatic or hitchhiking effects, where epistasis may be more likely for (ii). Case (iii) changes are rare and explain single isolate sequence variations. This also seems to be likely in case (iv), where the effect on the tips is amplified due to other effects or amino acid changes. Notably, all case (iii) changes are also categorized as case (i) changes. Of all changes, E156K occurs once on the tree trunk. All changes appear at several points in time for different antigenic types, which indicates a potential antigenic influence. Furthermore, for five changes (G135E, A138T, E156K, S186I and V226I), the respective site was identified as being under positive selection (Bush *et al.*, 1999b).

In a recent (unpublished) study, Koel *et al.* (*Koel et al.; Antigenic evolution of influenza A (H3N2) virus is dictated by 7 residues in the hemagglutinin protein; 2nd International Influenza Meeting, Münster; 2011*) showed by site-directed mutagenesis that changes at seven positions in the HA protein (145, 155, 156, 158, 159, 189 and 193) are responsible for large antigenic changes, all except two are part of antigenic cluster transitions, over the 35 year time period. Of these, 155, 156 and 189 are also identified as generally important by our default method. If single isolate variations are excluded from the analysis, position 158 is also identified. For the other two positions (145 and 159) we identified changes with high antigenic weights (e.g. N145K and S159Y). For position 193, evidence of antigenic importance could be found in our analysis if using ancestral character state reconstruction with maximum parsimony (see Supplement). Thus, our results also support the relevance of the sites proposed by Koel *et al.* (2011), even though they are not entirely comparable due to differences in experimental set up. Koel *et al.* analyzed prototype viruses with the amino acid consensus sequences of antigenic clusters and introduced only the specific changes between these prototype viruses, while our method also considers genetic and antigenic variations between other viral strains of the dataset.

4.5 Discussion

The antigenic impact of amino acid substitutions in the antigenic evolution of influenza A viruses can reliably be determined by time- and cost-intensive experimental analysis. As an alternative, we present a computational technique for inferring the antigenic impact of amino acid changes. Our method determines antigenic branch lengths for a given tree topology by fitting pairwise antigenic distances between isolates onto the tree with LSO. For inference of the tree, any state-of-the-art method can be used. A comparison between maximum likelihood, maximum parsimony and neighbor-joining trees showed that all resulted in similar prediction errors (leave-one-out absolute prediction error: 0.86, 0.87 and 0.87 antigenic units, respectively; correlation between predicted and measured by Pearson's correlation coefficient was 0.86 for all three methods). The antigenic impact of the branch-associated amino acid changes is determined by reconstructing the branch-associated amino acid changes with maximum likelihood (Yang *et al.*, 1995); other techniques, such as maximum parsimony or Bayesian reconstruction, could also be used (Fitch, 1971; Pagel *et al.*, 2004). A comparison between maximum likelihood and maximum parsimony ancestral character state reconstruction showed that these differed only in minor aspects, with the maximum likelihood reconstruction being an intermediate between accelerated and delayed transition in case of ties with maximum parsimony reconstruction. However, we did observe that more trunk branches were not assigned changes based on maximum likelihood reconstruction, which decreased the interpretability of antigenic weights in some cases.

We studied the antigenic evolution of the influenza A (H3N2) virus from 1968 to 2003 with antigenic trees inferred from data described in Smith *et al.* (Smith *et al.*, 2004). This allowed us to identify areas and branches in the tree corresponding to known antigenic types and transitions between these types. Analysis of antigenic weights identified seven sites in the HA1 domain of HA that were repeatedly associated with high antigenic impact. Additionally, our method identified five amino acid changes with high antigenic weights at several places in the antigenic tree. The sites and substitutions identified by our method may be of particular relevance for influenza A (H3N2) virus antigenic evolution, which has not been described before. For six of the seven positions found by site-directed mutagenesis to defining antigenic clusters for the 35 year time period (*Koel et al.; Antigenic evolution of influenza A (H3N2) virus is dictated by 7 residues in the hemagglutinin protein; 2nd International Influenza Meeting, Münster; 2011*), changes with high antigenic weights were identified with our technique, thus further supporting their relevance for influenza A (H3N2) revolution. The additional

sites detected by our method could be more relevant for genetic and antigenic variations between viral strains in our data set not resulting in antigenic cluster transitions. These were not analyzed by Koel *et al.*, who characterized antigenic differences of prototype viruses with the amino acid consensus sequences of the antigenic clusters.
As the dataset covers 35 years of viral evolution with a relatively small number of isolates, not all substitutions could be resolved to individual branches and their individual antigenic impacts inferred. A denser sampling of data points would allow a more precise decoding of the genotype-antigenicity relationships, as viral isolates were unevenly sampled across the 35 years. The median number of viral isolates available from between 1989 and 1997 was 15, whereas for the remaining years only three isolates per year were sampled (median). This unequal sampling is reflected in resolution of mutations to specific branches. Between 1989 and 1997, 19% of the branches with assigned changes carry three or more changes, whereas for the other years this is the case for 37% of the branches.
Our method allows inference of genotype to phenotype relationships from genetic sequences and associated pairwise phenotypic distances between individuals of a population or different taxa. We demonstrated the usefulness of this technique for analyzing the antigenic impact of amino acid changes in the evolution of human influenza A. An application of our method could be in influenza A virus surveillance. Here, it could be used to identify isolates and associated changes with large antigenic impact, which need to be identified for vaccine strain updates prior to an antigenic type transitions (McHardy and Adams, 2009). However, our method is not restricted to the analysis of influenza viruses or antigenic distance information but can be applied to the study of any system, be it within or across species, where homologous genetic sequences and associated pairwise phenotype distances are available. The software is available upon request from the authors.

4.6 Materials and Methods

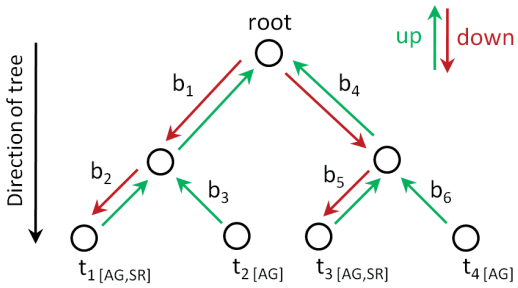
4.6.1 Inferring the phenotypic impact of amino acid changes in protein evolution

Our idea is to adapt the least-squares optimization (LSO) technique of Cavalli-Sforza and Edwards (Cavalli-Sforza and Edwards, 1967) for phylogenetic inference to the problem of identifying the phenotypic impact of amino acid changes in protein evolution. The original method of Cavalli-Sforza and Edwards (Cavalli-Sforza and Edwards, 1967)

identifies branch weights representing genetic distances according to the least-squares criterion for a tree topology. We applied this technique to infer 'antigenic trees', representing the antigenic evolution of the major surface protein of human influenza A virus (H3N2) over a 35-year period. In our adaptation, branch lengths represent antigenic distances inferred from HI assay data for human influenza A viruses and a maximum likelihood tree of the HA1 domain of hemagglutinin. Reconstruction of the amino acid changes associated with the branches of the tree allows us to infer the antigenic impact of the branch-associated amino acid changes. If sufficient data is available to resolve individual changes to individual branches, our method returns an estimate of the antigenic impact of the individual exchanges. In LSO, one minimizes the sum of squares between the given distances D and predicted distances d:

$$Q = \sum_{i=1}^{n} \sum_{j \neq i} w_{i,j}(D_{i,j} - d_{i,j})^2,$$

where W is the weight matrix for the different error terms, which were set to one here. The predicted distances $d_{i,j}$ are the sum of the branch weights on the path between leaf i and leaf j. Here, $d_{i,j} = \sum_k x_{i,j,k} v_k$, where $x_{i,j,k}$ equals one if branch k is on the path between leaves i and j in the phylogenetic tree and zero otherwise. Thus, we search for the best setting for the branch weights v_k. While evolutionary distances are usually used in this approach, here, we map antigenic distances to represent branch-specific weights. To restrict the branch weights to positive values, we used the Lawson-Hanson algorithm for non-negative LSO (Lawson and Hanson, 1995). Because the antigenic distances here are asymmetric (i.e. $d_{i,j} \neq d_{j,i}$) and because the antigen and antiserum raised against the same viral strain do not necessarily have the same position in the antigenic space (Lapedes and Farber, 2001), we introduce the concept of up-down trees. In up-down trees, viral strains are mapped to the leaves representing the corresponding antigen as well as the antiserum, and every branch is assigned two independent weights, the up- and the down-weight. Every path between two taxa i and j in the tree can be separated into the set of branches from taxon i to the least common ancestor (LCA) of i and j, and the branches from taxon j to the LCA. Now, the path between antigen i and antiserum j involves only the up-weights on branches from taxon i to the LCA and only the down-weights on branches from taxon j to the LCA (**Figure 4.2**).



	b_1	b_2	b_3	b_4	b_5	b_6
$t_1 \rightarrow t_3$	1/0	1/0	0/-	0/1	0/1	0/-
$t_2 \rightarrow t_1$	0/0	0/1	1/-	0/0	0/0	0/-
$t_1 \rightarrow t_1$	0/0	1/1	0/-	0/0	0/0	0/-

Figure 4.2: Schematic drawing demonstrating the up/down tree concept. For the two taxa t_2 and t_4, no antiserum is present, and thus, b_3 and b_6 only have up-weights. A path from t_1 to t_3 would use the up-weights of branch b_1 and b_2, and the down-weights of branch b_4 and b_5. Similarly, the path from t_2 to t_1 would use the up-weight of branch b_3 and the down-weight of branch b_2. Notably, the path from t_1 to t_1, namely the antigenic distance from antigen t_1 to the antiserum raised against strain t_1, would use the up-weight and the down-weight of branch b_1.

4.6.2 Performance measures

To evaluate how accurately antigenic distances were fitted onto the tree, we used four performance measures in leave-one-out cross validation experiments: mean absolute error (MAE), root mean squared error (RMSE), standard deviation (SD) and Pearson's correlation coefficient (CC). In leave-one-out cross validation, an antigenic tree is inferred from all but one antigenic distances and then is applied to predict the left out distance. A predicted distance corresponds to the antigenic path length between the two respective isolates in the tree (see above). This was repeated for every antigenic distance. Given n observed distances $D_{i,j}$ and predicted distances $d_{i,j}$ the performance measures are defined as follows:

$$MAE = \frac{1}{n} \sum_{i,j} |D_{i,j} - d_{i,j}|,$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i,j} (D_{i,j} - d_{i,j})^2},$$

$$SD = \sqrt{\frac{1}{n}\sum_i (x_i - \mu)^2} \text{ with } \mu = \frac{1}{n}\sum_i x_i \text{ and } x_i = |D_{i,j} - d_{i,j}|,$$

$$CC = \frac{\sum_{i,j}(D_{i,j} - \mu_D)(d_{i,j} - \mu_d)}{\sqrt{\sum_{i,j}(D_{i,j} - \mu_D)^2}\sqrt{\sum_{i,j}(d_{i,j} - \mu_d)^2}}$$

$$\text{with } \mu_D = \frac{1}{n}\sum_{i,j} D_{i,j} \text{ and } \mu_d = \frac{1}{n}\sum_{i,j} d_{i,j}.$$

4.6.3 Up-weights and down-weights in the tree

Antigenic branch lengths are realized as two independent weights, allowing for a detailed analysis of the underlying structure of the antigenic data. Up-weights represent the antigenic distance from isolates below this branch to every other isolate outside of this subtree, whereas down-weights represent distances from isolates outside of the subtree to the isolates below this branch. Thus, the branch weight types reveal different properties of the subtree. Let e be the branch going upwards from the least common ancestor of an antigenically homogenous group of viruses (a type) in the tree. The up-weight of e defines the degree to which the antigenic type is separated from other antigenic types according to antisera in other parts of the tree, i.e. how well antigens of this type are neutralized by antisera raised against other types. The down-weight of e defines the degree to which the antigenic type is separated from other types based on antisera *within* this part of the tree, i.e. how well other antigenic types are neutralized by antisera of this type. The antigenic weights of two types often differ, which is not surprising, as antigenic distances are not symmetric. For tip branches, the two weights define the different behavior of the antiserum and antigen of a viral strain. The up-weight reflects the antigenic properties of the isolate, whereas the down-weight reflects the antigenic weight of the antiserum raised against the viral isolate. In case no antiserum is present in a subtree, down-weights are undefined and assignment of up-weights becomes ambiguous as they form linear combinations. To resolve this, optimization is done only on the up-weights leading to leaves in the according subtree. Afterwards, up-weights of the internal branches are set to the minimum of the up-weights on the branches leading to the respective child nodes (these up-weights are accordingly reduced by the minimum) in a bottom-up traversal. The rationale behind

this is that if no additional information is present antigenic weights should rather be a common feature of a subgroup of taxa than single isolate variation for every taxon in the subgroup.

4.6.4 Phylogenetic inference

Hemagglutinin (HA) sequences from 258 seasonal human influenza A (H3N2) virus isolates from 1968 to 2003 and that were used by Smith *et al.* (Smith *et al.*, 2004) were downloaded from the Influenza Virus Resource (IVR) (Bao *et al.*, 2008). Alignments of DNA and protein sequences, restricted to positions 1 to 363 (sites without missing data appeared in more than 80% of the sequences), were created with Muscle (Edgar, 2004b) and manually curated. Trees were inferred with PhyML v3.0 (Guindon and Gascuel, 2003) 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 (Zwickl, 2006). Isolate A/duck/33/1980 was used as outgroup to root the tree and subsequently removed from the further analysis.
For placement of amino acid changes on the tree branches, protein sequences for the HA1 domain of HA (excluding the additional sites used for a higher resolution of the tree during the tree inference step) were assigned to the leaves of the tree inferred from nucleotide sequences. Ancestral character states were reconstructed under the maximum likelihood criterion using PAML v4.5 (Yang, 2007) under the JTT+Γ_4+F model (Jones *et al.*, 1992), with the frequency of each amino acid and the Gamma distribution of among-site rate variation, with four rate categories (Γ_4), estimated from the data. Based on the reconstructed ancestral sequences for the internal nodes and leaf node sequences, amino acid changes were assigned to the individual tree branches.

4.6.5 Antigenic data

HI assay data from Smith *et al.* was used and normalized according to these researchers' methods (Smith *et al.*, 2004). For each antigen i, antiserum j and the corresponding HI titer $h_{i,j}$, the distance was set as $d_{i,j} = \log_2(\max(\frac{h_j}{h_{i,j}}))$, where $\max(h_j)$ is the maximum entry for antiserum j. The dataset comprises 4,215 measured values between 273 antigens and 79 reference sera. As not all strains were available in the IVR, 18 antigens and 9 reference sera could not be mapped to a genetic sequence and were

excluded from the analysis. Additionally, threshold values (e.g. < 10, indicating the lower bound in the HI assay below which dilutions are not measured) were excluded from the analysis, as these values define only long-distance relationships and we did not want to introduce a potential bias by setting these entries to fixed values. In case of multiple antisera raised to the same viral strain, median values of the distances were used.

4.6.6 Definition of antigenic types

Antigenic types in the antigenic tree can be distinguished by selecting type-defining branches according to a threshold distance. The threshold was set to 1.0 antigenic units for average weights (average of up- and down-weights), such that all branches are selected whose average weights are at least twice as high as the average weights of all internal branches. To exclude undersampled groups, all branches leading to subtrees with three or less isolates were excluded.

4.7 Supporting material

The supporting material comprises tables 4.2 to 4.4 and figure 4.3. For the sake of limited space and dimensionality the supporting tables S1 and S2 as well as supporting figures S1 and S2 from the published article are not included in this thesis and can be accessed via the online version of the article (http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002492).

Table 4.2: Positions with multiple changes in the phylogenetic tree and high antigenic weights (mean and median *geq*1 antigenic unit, highlighted in bold). 'Tip' indicates leaf branches.

Position	Up-weights (mean/median)	Down-weights (mean/median)	Trunk	Tip
112	**1.14/1.13**	0.30/0.18	0/4	4/4
137	0.21/0.15	**1.27/1.05**	3/6	3/6
144	**1.21/1.39**	0.95/0.41	4/9	4/9
155	**1.52/1.77**	**2.29/3.16**	3/3	0/3
156	**1.28/1.44**	0.79/0.00	1/3	2/3
189	0.45/0.57	**1.95/2.42**	3/3	0/3
208	**1.24/1.73**	0.59/0.59	0/3	2/3

Table 4.3: Changes with multiple occurrences in the phylogenetic tree and high antigenic weights (mean and median ≥ 1 antigenic unit). 'Tip' indicates leaf branches. Down-weights are omitted, as all changes were identified using up-weights.

Change	Up-weights (mean/median)	Trunk	Tip
K62E	1.52/1.42	1/3	2/3
V112I	1.14/1.13	0/4	4/4
N145K	1.36/1.52	1/9	5/9
L226Q	1.16/1.07	1/8	6/8
T248I	1.01/1.48	0/3	3/3

Table 4.4: Type-defining branches selected by different thresholds for average branch weights. Branches (1)-(9) were selected as type-defining branches at a threshold distance of 1.0 antigenic units. Branches (i)-(xii) reveal further subdivision of antigenic types at a threshold distance of 0.5 antigenic units. Asterisks mark branches whose sibling branch leads to a single isolate. Subscript $_2$ indicates that a branch is a direct successor of a type-defining branch (except for branch (i), which is a predecessor of the type-defining branch). Subscript $_{\text{sub}}$ indicates a subdivision of an antigenic type without a directly known reference strain.

Threshold	No.	Type transition	Branch amino acid changes	Weights (up/down/avg)	Trunk
	(1)	WU95-SY97	K62E, V144I, K156Q, E158K, V196A, N276K	2.5/2.6/2.6	x
2.0	(2)	SY97-FU02	L25I, R50G, H75Q, E83K, A131T, H155T, Q156H, S186G, V202I, W222R, G225D	1.8/3.2/2.5	x
	(3)	BA79-SI87	G124D, Y155H, K189R	0.2/3.3/1.7	x
	(4)	VI75-TX77	K50R, N137Y, G158E, M260I	0.6/2.8/1.7	x
	(5)	HK68-EN72	T122N, G144D, T155Y, R207K	2.6/0.4/1.5	x
1.5	(6)	EN72-VI75	S145N, Q189K, I217V, I278S	0.6/2.4/1.5	x
	(7)	BE92-WU95	K135T, N145K, N262S	1.5/1.1/1.3	x
	(8)	BE89-BE92	I214T	1.4/1.1/1.3	x
1.0	(9)	SI87-BE89	G135E, N145K	2.0/0.0/1.0	
	(i)	SI87-BE89$_2$		1.0/0.9/0.9	x
	(ii)	BA79-CC85/LE86	S159Y	1.1/0.7/0.9	x
	(iii)	BE92$_{\text{sub}}$	N145K	1.2/0.3/0.8	
	(iv)	BE92-SH93		0.4/1.2/0.8	
	(v)	BE92-JO94	S47P, D124N, N216D, S219Y	0.6/0.9/0.8	
	(vi)	TX77-BA79	N133S, P143S, G146S, K156E, T160K, Q197R, V217I	1.4/0.0/0.7	x
	(vii)	BA79$_{\text{sub}}$	N2K, D144V	1.4/0.0/0.7	x
	(viii)	BE92$_{\text{sub}}$	G135K	0.2/1.1/0.7	x
	(ix)	FU022*		0.0/1.3/0.7	
	(x)	SI87-GU89	E82K, K83E, T131A, K299R	0.8/0.3/0.6	x
	(xi)	EN72$_2$*	L3F, N188D	0.9/0.2/0.5	x
0.5	(xii)	EN72-VI75$_2$	N53D, N137S, L164Q, F174S, N193D, R201K, I213V, I230V	0.0/1.0/0.5	

4.7.1 Influence of threshold distance on type-defining branches

The definition of antigenic types depends on the choice of the threshold distance. An average weight of 1.0 antigenic units as threshold distance resulted in robust antigenic types, differing at least by a two-fold dilution from the preceding antigenic type. Our method resolves antigenically relevant changes between successive antigenic types in several cases to several successive branches. Therefore, a higher threshold of 2.0 antigenic units for individual branches (a four-fold dilution), as suggested to distinguish antigenically diverse viral strains (Russell *et al.*, 2008b), does not allow distinction between different antigenic groups (only if the transition is not well resolved in the data and the antigenic impact of multiple changes is summarized on a single branch). Choosing a lower threshold distance of 0.5 antigenic units selected twelve further type-defining branches (**Supplementary Table 4.4, Supplementary Figure 4.3**). Although a direct interpretation of the threshold distance of 0.5 antigenic units is difficult, as it is not directly representative of dilution steps, five of these branches represent further subdivisions of antigenic types described before in literature. Among these branches is the branch leading to the TX77-BA79 transition that was also identified by antigenic cartography as a cluster transition (Smith *et al.*, 2004). Different from antigenic cartography, only a subset of the cluster-difference substitutions is assigned to this branch. The remaining changes account for intra TX77 changes and are assigned to the preceding branch.

Branch (i) is a predecessor of the SI87-BE89 type-defining branch. Although no change was assigned to this branch with a maximum likelihood character state reconstruction, with maximum parsimony ancestral character state reconstruction the change N193S was assigned to it. This branch also precedes the following BE92 type on the trunk, indicating that this precursor circulated undetected in the evolutionary reservoir during this time period and that N193S conferred a significant fitness effect. N193S is located in the receptor binding site (Wilson *et al.*, 1981) and, thus, probably represents an antigenically important substitution for SI87-BE89, which was not revealed by antigenic cartography. Similarly, branches (vii) and (viii) account for antigenic type differentiations and define unsampled antigenically distinct intermediates between two antigenic types. For both branches, several of the identified changes (G135K, D144V) occur at positions under positive selection (Bush *et al.*, 1999b), which further supports their antigenic relevance.

Other branches, such as (ii), define antigenic types that include viral isolates with sufficient antigenic dissimilarities to preceding viral strains to warrant vaccine updates.

Here, branch (ii) refers to the antigenic type CC85/LE86, corresponding to two viral strains predominant from 1985 to 1987 (WHO, 1986, 1987b,a). With antigenic cartography, this type was placed within the BA79 cluster. Similarly, branch (iv) corresponds to the antigenic type SH93, branch (v) refers to the antigenic type JO94 and branch (x) refers to the antigenic type GU89. SH93 corresponds a viral strain predominant from 1993 to 1994 (WHO, 1994b,a) and was placed within the BE92 cluster with antigenic cartography. JO94 corresponds to a viral strain predominant from 1994 to 1996 (WHO, 1995b,a, 1996b,a) and was placed within the BE92 cluster with antigenic cartography. Finally, GU89 corresponds to a viral strain recommended for use in the influenza virus vaccine in the 1990/91 northern hemisphere season (WHO, 1990) and was placed within the SI87 cluster with antigenic cartography.

The remaining branches account for intra-type antigenic differentiations with different interpretations. Branch (iii) indicates either an evolutionary volatile change with high antigenic impact or discrepancies between phylogenetic inference and antigenic phenotype (see discussion on isolates of the WU95 antigenic cluster placed in the BE92 cluster in the phylogenetic tree and on N145K in the main article). Branch (xii) is a successor of the VI75 type-defining branch and further distinguishes the VI75 antigenic type. The assigned changes are of antigenic relevance for distinguishing VI75 from previous and successive antigenic types but were assigned less antigenic impact than the changes assigned to the type-defining branch. Finally, branches (ix) and (xi) directly follow two type-defining branches and therefore further resolve these antigenic type transitions. Single isolates are separated by these branches from the remaining isolates of an antigenic type, indicating the presence of precursors with high antigenic distance to previous antigenic types. However, these branches show that further antigenic change resulted in the antigenic types that rose to predominance.

Although, some branches have no changes assigned to them, possible changes at the HA2 domain of the hemagglutinin or in the according neuraminidase may account for these antigenic variations. Furthermore, using a fixed tree topology that might be wrong, introduces further bias for fitting antigenic distances resulting in false assignments on branches without changes.

(i)	$BE89_2$	
(ii)	CC85/LE86	S159Y
(iii)	$BE92_{sub}$	N145K
(iv)	SH93	
(v)	JO94	S47P, D124N, N216D, S219Y
(vi)	BA79	N133S, P143S, G146S, K156E, T160K, Q197R, V217I
(vii)	$BA79_{sub}$	N2K, D144V
(viii)	$BE92_{sub}$	G135K
(ix)	$FU02_2$	
(x)	GU89	E82K, K83E, T131A, K299R
(xi)	$EN72_2$	L3F, N188D
(xii)	$VI75_2$	N53D, N137S, L164Q, F174S, N193D, R201K, I213V, I230V

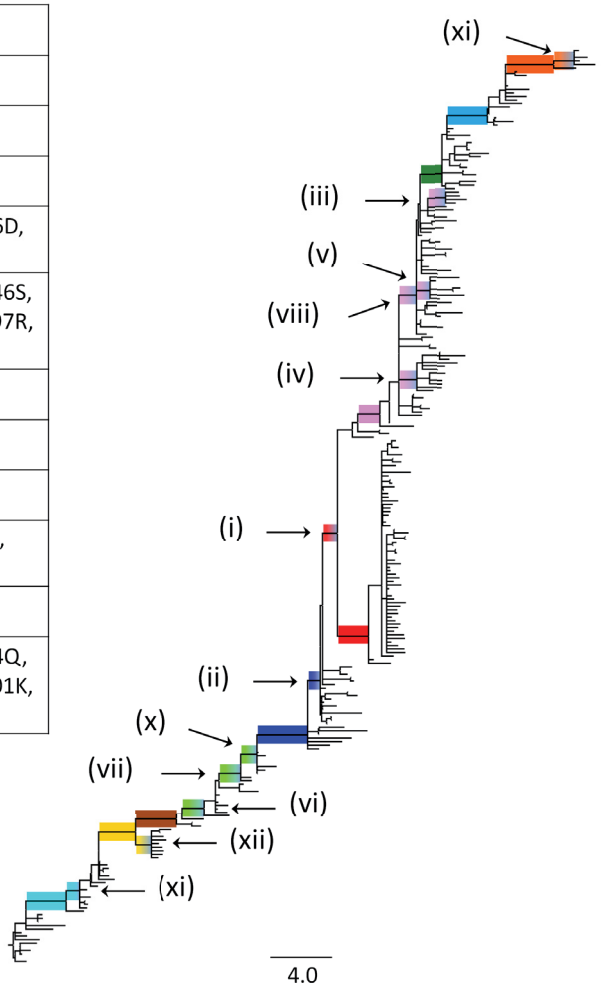


Figure 4.3: Branch lengths represent antigenic distances (maximum of up- and down-weights for each branch) inferred from a maximum likelihood tree of 258 hemagglutinin sequences of seasonal influenza A (H3N2) virus isolates and serological data. Colored edges show antigenic type transitions, with internal branches with high average antigenic weights (≥ 1.0 antigenic units, coloring according to Figure 4.1A) or moderate antigenic weights ≥ 0.5 antigenic units (coloring as gradient from the higher order antigenic type). Subscript $_2$ indicates that a branch was a direct successor of the according type-defining branch (except of branch (i), who is a predecessor of the according type-defining branch). Subscript $_{sub}$ indicates a subdivision of an antigenic type without a direct matching of a reference strain.

CHAPTER 5

Synopsis

The objective of this thesis was to develop computational methods for the analysis of the phylodynamics of rapidly evolving pathogens. We demonstrated the application of the developed methods with the influenza A virus, which has a large impact on public health. To study the virus on a population-genetics level we developed allele dynamics plots (AD plots) that allow to visualize the population-level dynamics of a gene over time (Steinbrück and McHardy, 2011). AD plots for the hemagglutinin gene of seasonal influenza A (H3N2) viruses showed the molecular dynamics of the gene over an eleven year period. Furthermore, we found that alleles with the highest frequency increase between two consecutive seasons show evidence for directional selection.

To study the antigenic evolution of seasonal influenza A (H3N2) viruses we developed antigenic trees, which allow to resolve genotpye-phenotype relationships with respect to the antigenic phenotype (Steinbrück and McHardy, 2012). Inference of antigenic branch weights on a phylogenetic tree of viral isolates sampled over 35 years allowed to resolve genotype-phenotype relationships with high accuracy. Furthermore, we identified both known and novel amino acid changes and protein sites that are of antigenic relevance for the influenza A (H3N2) virus.

In summary, both developed methods allow for a detailed analysis of the phylodynamics of influenza viruses and, thus, might aid in the biannual vaccine strain selection process.

References

Abdel-Ghafar, A.-N., T. Chotpitayasunondh, Z. Gao, F. G. Hayden, D. H. Nguyen *et al.* (2008). Update on avian influenza A (H5N1) virus infection in humans. *N Engl J Med*, **358** (3): 261–73. 8

Adams, B. and A. C. McHardy (2011). The impact of seasonal and year-round transmission regimes on the evolution of influenza A virus. *P Roy Soc B-Biol Sci*, **278** (1716): 2249–56. 1

Bahl, J., M. I. Nelson, K. H. Chan, R. Chen, D. Vijaykrishna *et al.* (2011). Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proc Natl Acad Sci USA*, **108** (48): 19359–64. 2, 49

Bao, Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky *et al.* (2008). The influenza virus resource at the National Center for Biotechnology Information. *J Virol*, **82** (2): 596–601. 2, 3, 27, 65

Barr, I. G., N. Komadina, a. C. Hurt, P. Iannello, C. Tomasov *et al.* (2005). An influenza A(H3) reassortant was epidemic in Australia and New Zealand in 2003. *J Med Virol*, **76** (3): 391–7. 7, 9

Bateman, A. C., M. G. Busch, A. I. Karasin, N. Bovin and C. W. Olsen (2008). Amino acid 226 in the hemagglutinin of H4N6 influenza virus determines binding affinity

for alpha2,6-linked sialic acid and infectivity levels in primary swine and human respiratory epithelial cells. *J Virol*, **82** (16): 8204–9. 58

Bedford, T., S. Cobey, P. Beerli and M. Pascual (2010). Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS Pathog*, **6** (5): e1000918. 2, 49

Berthoud, T. K., M. Hamill, P. J. Lillie, L. Hwenda, K. a. Collins *et al.* (2011). Potent CD8+ T-cell immunogenicity in humans of a novel heterosubtypic influenza A vaccine, MVA-NP+M1. *Clin Infect Dis*, **52** (1): 1–7. 8

Brodie, E., A. Moore and F. Janzen (1995). Visualizing and quantifying natural selection. *Trends Ecol Evol*, **10** (8): 313–8. 10, 11

Bush, R. M., C. A. Bender, K. Subbarao, N. J. Cox and W. M. Fitch (1999a). Predicting the evolution of human Influenza A. *Science*, **286** (5446): 1921–5. 12, 26, 50

Bush, R. M., W. M. Fitch, C. a. Bender and N. J. Cox (1999b). Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol*, **16** (11): 1457–65. 6, 12, 23, 24, 26, 50, 59, 69

Cai, Z., T. Zhang and X.-F. Wan (2010). A computational framework for influenza antigenic cartography. *PLoS Comput Biol*, **6** (10): e1000949. 52

Cammack, R., T. Atwood, P. Campbell, H. Parish, T. Smith *et al.* (2006). *Oxford Dictionary of Biochemistry and Molecular Biology*. Oxford University Press, UK, 2 edition. 9

Carrat, F. and a. Flahault (2007). Influenza vaccine: the challenge of antigenic drift. *Vaccine*, **25** (39-40): 6852–62. 9

Cauchemez, S., A.-J. Valleron, P.-Y. Boëlle, A. Flahault and N. M. Ferguson (2008). Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*, **452** (7188): 750–4. 3

Cavalli-Sforza, L. L. and a. W. Edwards (1967). Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet*, **19** (3 Pt 1): 233–57. 61

Clifford, J. (1976). *Introduction to natural selection*. University Park Press, Baltimore, 1 edition. 10

Cox, N. and K. Subbarao (2000). Global epidemiology of influenza: past and present. *Annu Rev Med*, **51**: 407–1. 7

Cox, N. J., T. L. Brammer and H. L. Regnery (1994). Influenza: global surveillance for epidemic and pandemic variants. *Eur J Epidemiol*, **10** (4): 467–70. 8, 24

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London, 1 edition. 10

Dawood, F. S., S. Jain, L. Finelli, M. W. Shaw, S. Lindstrom *et al.* (2009). Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N Engl J Med*, **360** (25): 2605–15. 37

de Jong, J. C., W. E. Beyer, a. M. Palache, G. F. Rimmelzwaan and a. D. Osterhaus (2000). Mismatch between the 1997/1998 influenza vaccine and the major epidemic A(H3N2) virus strain as the cause of an inadequate vaccine-induced antibody response to this strain in the elderly. *J Med Virol*, **61** (1): 94–9. 8

Drake, J. W. (1993). Rates of spontaneous mutation among RNA viruses. *Proc Natl Acad Sci USA*, **90** (9): 4171–5. 13

Drake, J. W. and C. B. C. Hwang (2005). On the mutation rate of herpes simplex virus type 1. *Genetics*, **170** (2): 969–70. 13

Drummond, A. J. and A. Rambaut (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, **7**: 214. 28

Du, L., Y. Zhou and S. Jiang (2010). Research and development of universal influenza vaccines. *Microbes Infect*, **12** (4): 280–6. 8

Du, X., Z. Wang, A. Wu, L. Song, Y. Cao *et al.* (2008). Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution. *Genome Res*, **18** (1): 178–87. 13, 25, 50

Duffy, S., L. A. Shackelton and E. C. Holmes (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet*, **9** (4): 267–76. 12

Dushoff, J., J. B. Plotkin, C. Viboud, D. J. D. Earn and L. Simonsen (2006). Mortality due to influenza in the United States–an annualized regression approach using multiple-cause mortality data. *Am J Epidemiol*, **163** (2): 181–7. 3

Eccles, R. (2005). Understanding the symptoms of the common cold and influenza. *Lancet Infect Dis*, **5** (11): 718–25. 1

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, **5** (1): 113. 27

Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32** (5): 1792–7. 65

Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer Associates Incorporated, Massachusetts. 9, 28

Fereidouni, S. R., M. Beer, T. Vahlenkamp, E. Starick and G.-i. Riems (2009). Differentiation of two distinct clusters among currently circulating influenza A(H1N1)v viruses, March-September 2009. *Eurosurveillance*, **14** (46): 1–3. 37

Ferguson, N., A. Galvani and R. Bush (2003). Ecological and immunological determinants of influenza evolution. *Nature*, **422** (6930): 428–33. 6

Fitch, W. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Zool*, **20** (4): 406–16. 9, 28, 60

Fitch, W. M., R. M. Bush, C. a. Bender and N. J. Cox (1997). Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc Natl Acad Sci USA*, **94** (15): 7712–8. 7

Fouchier, R., V. Munster, A. Wallensten, T. Bestebroer, S. Herfst *et al.* (2005). Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. *J Virol*, **79** (5): 2814–22. 6, 24

Fouchier, R. A. M. and D. J. Smith (2010). Use of antigenic cartography in vaccine seed strain selection. *Avian Dis*, **54**: 220–3. 26, 36

Futuyma, D. (1997). *Evolutionary biology*. Sinauer Associates Incorporated, Massachusetts, 3 edition. 27

Garten, R. J., C. T. Davis, C. a. Russell, B. Shu, S. Lindstrom *et al.* (2009). Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science*, **325** (5937): 197–201. 37

Ghedin, E., N. a. Sengamalay, M. Shumway, J. Zaborsky, T. Feldblyum *et al.* (2005). Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, **437** (7062): 1162–6. 5, 30

Gilbert, J. A., J. A. Steele, J. G. Caporaso, L. Steinbrück, J. Reeder *et al.* (2011). Defining seasonal marine microbial community dynamics. *ISME J*, **6** (2): 298–308.

Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly *et al.* (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, **303** (5656): 327–32. 2, 9, 12, 13, 22

Guindon, S. and O. Gascuel (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, **52** (5): 696–704. 27, 65

Gupta, V., D. J. Earl and M. W. Deem (2006). Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine*, **24** (18): 3881–8. 8

Harding, E. F. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Adv Appl Probab*, **3** (1): 44–77. 9

Hastie, T., R. Tibshirani and J. Friedman (2004). *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, 2 edition. 51

Hay, A., R. Daniels, Y. Lin, Z. Xiang, V. Gregory *et al.* (2007). *WHO Collaborating Centre for Reference and Research on Influenza, interim report September*. National Institute for Medical Research, London. 30, 31

Hay, A., Y. Lin, V. Gregory and M. Bennet (2003). *WHO Collaborating Centre for Reference and Research on Influenza, annual report*. National Institute for Medical Research, London. 30, 31

Hay, A., Y. Lin, V. Gregory and M. Bennet (2005). *WHO Collaborating Centre for Reference and Research on Influenza, interim report February*. National Institute for Medical Research, London. 30, 31

Hay, A., Y. Lin, V. Gregory and M. Bennet (2006). *WHO Collaborating Centre for Reference and Research on Influenza, interim report March*. National Institute for Medical Research, London. 30, 31

Hay, a. J., V. Gregory, a. R. Douglas and Y. P. Lin (2001). The evolution of human influenza viruses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **356** (1416): 1861–70. 11

Hein, J., M. Schierup and C. Wiuf (2005). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, Oxford, USA, 1 edition. 27

Hensley, S. E., S. R. Das, A. L. Bailey, L. M. Schmidt, H. D. Hickman *et al.* (2009). Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science*, **326** (5953): 734–6. 59

Hirst, G. (1943). Studies of antigenic differences among strains of influenza A by means of red cell agglutination. *J Exp Med*, **78** (10): 407–23. 15, 49

Holmes, E. C. (2010). The comparative genomics of viral emergence. *Proc Natl Acad Sci USA*, **107 Suppl** (8): 1742–6. 6

Holmes, E. C., E. Ghedin, N. Miller, J. Taylor, Y. Bao *et al.* (2005). Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol*, **3** (9): e300. 6, 7, 9, 10

Huang, J.-W., C.-C. King and J.-M. Yang (2009). Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses. *BMC bioinformatics*, **10 Suppl 1**: S41. 16, 26, 50

Huelsenbeck, J. P. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17** (8): 754–5. 28

Huleatt, J. W., V. Nakaar, P. Desai, Y. Huang, D. Hewitt *et al.* (2008). Potent immunogenicity and efficacy of a universal influenza vaccine candidate comprising a recombinant fusion protein linking influenza M2e to the TLR5 ligand flagellin. *Vaccine*, **26** (2): 201–14. 8

Hurst, L. D. (2009). Fundamental concepts in genetics: genetics and the understanding of selection. *Nat Rev Genet*, **10** (2): 83–93. 10

Jin, H., H. Zhou, H. Liu, W. Chan, L. Adhikary *et al.* (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** (1): 113–9. 34, 58

Jones, D. T., W. R. Taylor and J. M. Thornton (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, **8** (3): 275–82. 65

Karlsson Hedestam, G. B., R. a. M. Fouchier, S. Phogat, D. R. Burton, J. Sodroski *et al.* (2008). The challenges of eliciting neutralizing antibodies to HIV-1 and to influenza virus. *Nat Rev Microbiol*, **6** (2): 143–55. 8

Kawaoka, Y. (2006). *Influenza virology: current topics*. Caister Academic Press, London. 58

Keele, B. F., F. Van Heuverswyn, Y. Li, E. Bailes, J. Takehisa *et al.* (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*, **313** (5786): 523–6. 22

Koelle, K., S. Cobey, B. Grenfell and M. Pascual (2006). Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science*, **314** (5807): 1898–903. 24, 25

Kryazhimskiy, S., J. Dushoff, G. a. Bazykin and J. B. Plotkin (2011). Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet*, **7** (2): e1001301. 10

Kryazhimskiy, S. and J. B. Plotkin (2008). The population genetics of dN/dS. *PLoS Genet*, **4** (12): e1000304. 12, 23

Kuiken, T., E. C. Holmes, J. McCauley, G. F. Rimmelzwaan, C. S. Williams *et al.* (2006). Host species barriers to influenza virus infections. *Science*, **312** (5772): 394–7. 24

Lamb, R. A. and R. M. Krug (2001). *Orthomyxoviridae: the viruses and their replication*. In Fields Virology. Lippincott Williams & Wilkins, Philadelphia. 5

Lapedes, A. and R. Farber (2001). The geometry of shape space: application to influenza. *J Theor Biol*, **212** (1): 57–69. 49, 62

Lawson, C. and R. Hanson (1995). *Solving least squares problems*, volume 15. Society for Industrial Mathematics, Philadelphia. 62

Lee, M.-S., M.-C. Chen, Y.-C. Liao and C. A. Hsiung (2007). Identifying potential immunodominant positions and predicting antigenic variants of influenza A/H3N2 viruses. *Vaccine*, **25** (48): 8133–9. 16, 26, 50

Lemey, P., O. G. Pybus, B. Wang, N. K. Saksena, M. Salemi *et al.* (2003). Tracing the origin and history of the HIV-2 epidemic. *Proc Natl Acad Sci USA*, **100** (11): 6588–92. 23

Liao, Y.-C., M.-S. Lee, C.-Y. Ko and C. a. Hsiung (2008). Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics*, **24** (4): 505–12. 16, 26, 50

Lin, Y. P., V. Gregory, M. Bennett and a. Hay (2004). Recent changes among human influenza viruses. *Virus Res*, **103** (1-2): 47–52. 4, 30, 31

Lipsitch, M. and C. Viboud (2009). Influenza seasonality: lifting the fog. *Proc Natl Acad Sci USA*, **106** (10): 3645–6. 3

Liu, S., K. Ji, J. Chen, D. Tai, W. Jiang *et al.* (2009). Panorama phylogenetic diversity and distribution of Type A influenza virus. *PloS one*, **4** (3): e5022. 6

Lowen, A. C. and P. Palese (2007). Influenza virus transmission: basic science and implications for the use of antiviral drugs during a pandemic. *Infect Disord Drug Targets*, **7** (4): 318–28. 24

Matrosovich, M., a. Tuzikov, N. Bovin, a. Gambaryan, a. Klimov *et al.* (2000). Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. *J Virol*, **74** (18): 8502–12. 58

McHardy, A. C. and B. Adams (2009). The role of genomics in tracking the evolution of influenza A virus. *PLoS Pathog*, **5** (10): e1000566. 61

Medina, R. a. and A. García-Sastre (2011). Influenza A viruses: new research developments. *Nat Rev Microbiol*, **9** (8): 590–603. 4, 5

Molinari, N.-A. M., I. R. Ortega-Sanchez, M. L. Messonnier, W. W. Thompson, P. M. Wortley *et al.* (2007). The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine*, **25** (27): 5086–96. 1, 3

Morens, D., J. Taubenberger and A. Fauci (2009). The persistent legacy of the 1918 influenza virus. *N Engl J Med*, **361** (3): 225–9. 24

Ndifon, W. (2011). New methods for analyzing serological data with applications to influenza surveillance. *Influenza Other Respi Viruses*, **5** (3): 206–12. 52

Nelson, M. I. and E. C. Holmes (2007). The evolution of epidemic influenza. *Nat Rev Genet*, **8** (3): 196–205. 4, 6, 29

Nelson, M. I., L. Simonsen, C. Viboud, M. a. Miller and E. C. Holmes (2007). Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. *PLoS Pathog*, **3** (9): 1220–8. 25, 49, 53

Nelson, M. I., L. Simonsen, C. Viboud, M. a. Miller, J. Taylor *et al.* (2006). Stochastic processes are key determinants of short-term evolution in influenza a virus. *PLoS Pathog*, **2** (12): e125. 53

Network, W. G. I. S. (2011). *Manual for the laboratory diagnosis and virological surveillance of influenza*. World Health Organization, Geneva, 1 edition. 52

Neumann, G., T. Noda and Y. Kawaoka (2009). Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature*, **459** (7249): 931–9. 24

Nobusawa, E., T. Aoyama and H. Kato (1991). Comparison of complete amino acid sequences and receptor-binding properties among 13 serotypes of hemagglutinins of influenza A viruses. *J Virol*, **182** (2): 475–85. 32, 39

Nozawa, M., Y. Suzuki and M. Nei (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci USA*, **106** (16): 6700–5. 12, 26

Oleksyk, T. K., M. W. Smith and S. J. O'Brien (2010). Genome-wide scans for footprints of natural selection. *Phil Trans R Soc B*, **365** (1537): 185–205. 10

Orr, H. A. (2009). Fitness and its role in evolutionary genetics. *Nat Rev Genet*, **10** (8): 531–9. 10

Pagel, M., A. Meade and D. Barker (2004). Bayesian estimation of ancestral character states on phylogenies. *Syst Biol*, **53** (5): 673–84. 9, 28, 60

Pan, C., B. Cheung, S. Tan, C. Li, L. Li *et al.* (2010). Genomic signature and mutation trend analysis of pandemic (H1N1) 2009 influenza A virus. *PLoS one*, **5** (3): e9549. 37

Plotkin, J. B., J. Dushoff and S. a. Levin (2002). Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proc Natl Acad Sci USA*, **99** (9): 6263–8. 12, 15, 25, 50

Pond, S. L. K., S. D. W. Frost and S. V. Muse (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, **21** (5): 676–9. 12, 23, 35

Pond, S. L. K., A. F. Y. Poon, A. J. L. Brown and S. D. W. Frost (2008). A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol Biol Evol*, **25** (9): 1809–24. 12, 23, 26, 35, 50

Pybus, O. G. and A. Rambaut (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet*, **10** (8): 540–50. 13, 22

Rambaut, A., O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger *et al.* (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature*, **453** (7195): 615–9. 25

Ricklefs, R. E. (2007). Estimating diversification rates from phylogenetic information. *Trends Ecol Evol*, **22** (11): 601–10. 23

Roxas, M. and J. Jurenka (2007). Colds and influenza: a review of diagnosis and conventional, botanical, and nutritional considerations. *Altern Med Rev*, **12** (1): 25–48. 1

Russell, C., T. Jones, I. Barr, N. Cox, R. Garten *et al.* (2008a). Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*, **26 Suppl 4**: D31–4. 8, 15, 24, 36, 49, 55

Russell, C., T. Jones, I. Barr, N. Cox, R. Garten *et al.* (2008b). The global circulation of seasonal influenza A (H3N2) viruses. *Science*, **320** (5874): 340. 2, 17, 69

Russell, C. a., T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten *et al.* (2008c). The global circulation of seasonal influenza A (H3N2) viruses. *Science*, **320** (5874): 340–6. 25, 49

Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4** (4): 406–25. 9

Shaman, J. and M. Kohn (2009). Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc Natl Acad Sci USA*, **106** (9): 3243–8. 3

Shih, A. C.-C., T.-C. Hsiao, M.-S. Ho and W.-H. Li (2007). Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc Natl Acad Sci USA*, **104** (15): 6283–8. 13, 25, 50

Skehel, J. and D. C. Wiley (2000). Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu Rev Biochem*, **69**: 531–69. 24

Smith, D. J., A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan *et al.* (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science*, **305** (5682): 371–6. 2, 6, 15, 16, 24, 26, 29, 36, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 60, 65, 69

Smith, G. J. D., D. Vijaykrishna, J. Bahl, S. J. Lycett, M. Worobey *et al.* (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, **459** (7250): 1122–5. 13, 37

Steinbrück, L. and A. C. McHardy (2011). Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic Acids Res*, **39** (1): e4. 13, 50, 73

Steinbrück, L. and A. C. McHardy (2012). Inference of genotype-phenotype relationships in the antigenic evolution of human influenza A (H3N2) viruses. *PLoS Comput Biol*, **8** (4): e1002492. 16, 73

Steinhauer, D. a. and J. J. Skehel (2002). Genetics of influenza viruses. *Annu Rev Genet*, **36**: 305–32. 4

Sui, J., W. C. Hwang, S. Perez, G. Wei, D. Aird *et al.* (2009). Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nat Struct Mol Biol*, **16** (3): 265–73. 8

Taubenberger, J. K. and J. C. Kash (2010). Influenza virus evolution, host adaptation, and pandemic formation. *Cell Host Microbe*, **7** (6): 440–51. 3

Taubenberger, J. K. and D. M. Morens (2006). 1918 Influenza: the mother of all pandemics. *Emerg Infect Dis*, **12** (1): 15–22. 49

Templeton, A. (2006). *Population genetics and microevolutionary theory*. John Wiley & Sons Incorporated, Hoboken, 1 edition. 27

Tognotti, E. (2009). Influenza pandemics: a historical retrospect. *J Infect Dev Ctries*, **3** (5): 331–4. 7, 49

Tong, S., Y. Li, P. Rivailler, C. Conrardy, D. a. A. Castillo *et al.* (2012). A distinct lineage of influenza A virus from bats. *Proc Natl Acad Sci USA*, **eprint ahead of publication**: 1–6. 6

Tusche, C., L. Steinbrück and A. C. McHardy (2012). Detecting patches of protein sites of influenza A viruses under positive selection. *Mol Biol Evol*, (in press). 14

Viboud, C., O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. a. Miller *et al.* (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, **312** (5772): 447–51. 3

Wallace, R., H. HoDac, R. Lathrop and W. Fitch (2007). A statistical phylogeography of influenza A H5N1. *Proc Natl Acad Sci USA*, **104** (11): 4473. 13, 22

Wan, H., E. M. Sorrell, H. Song, M. J. Hossain, G. Ramirez-Nieto *et al.* (2008). Replication and transmission of H9N2 influenza viruses in ferrets: evaluation of pandemic potential. *PLoS one*, **3** (8): e2923. 58

Webster, R., W. Laver and G. Air (1982). Molecular mechanisms of variation in influenza viruses. *Nature*, **296** (5853): 115–21. 3, 6, 7

Webster, R. G., W. J. Bean, O. T. Gorman, T. M. Chambers and Y. Kawaoka (1992). Evolution and ecology of influenza A viruses. *Microbiol Rev*, **56** (1): 152–79. 4, 24

WHO (1972). Antigenic variation in influenza a viruses. *WHO WER*, **47** (40): 381–4. 56

WHO (1986). Recommended composition if influenza virus vaccines for the use in thes 1986-1987 season. *WHO WER*, **61** (9): 61–4. 70

WHO (1987a). Influenza. *WHO WER*, **62** (44): 335. 70

WHO (1987b). Recommended composition of influenza virus vaccines for use in the 1987-1988 season. *WHO WER*, **62** (9): 54–6. 70

WHO (1990). Recommended composition of influenza virus vaccines for use in the 1990-1991 season. *WHO WER*, **65** (8): 53–6. 70

WHO (1994a). Antigenic analysis of recent influenza virus isolates and influenza activity in the southern hemisphere. *WHO WER*, **69** (39): 291. 70

WHO (1994b). Recommended composition of influenza virus vaccines for use in the 1994-1995 season. *WHO WER*, **69** (8): 53–6. 70

WHO (1995a). Influenza - antigenic activity of recent influenza virus isolates and influenza activity in the southern hemisphere. *WHO WER*, **70** (39): 277. 70

WHO (1995b). Recommended composition of influenza virus vaccines for use in the 1995-1996 season. *WHO WER*, **70** (8): 53–6. 70

WHO (1996a). Influenza - antigenic analysis of recent influenza virus isolates and influenza activity in the southern hemisphere. *WHO WER*, **71** (39): 292–3. 70

WHO (1996b). Recommended composition of influenza virus vaccines for use in the 1996-1997 season. *WHO WER*, **71** (8): 57–61. 70

WHO (1999a). Recommended composition of influenza virus vaccines for use in the 1999-2000 season. *WHO WER*, **74** (8): 57–61. 33, 36

WHO (1999b). Recommended composition of influenza virus vaccines for use in the 2000 influenza season. *WHO WER*, **74** (39): 321–5. 31, 33, 34, 36

WHO (2000a). Recommended composition of influenza virus vaccines for use in the 2000-2001 season. *WHO WER*, **75** (8): 61–5. 31, 33, 34, 36

WHO (2000b). Recommended composition of influenza virus vaccines for use in the 2001 influenza season. *WHO WER*, **75** (41): 330–3. 33, 36

WHO (2001a). Recommended composition of influenza virus vaccines for use in the 2001-2002 influenza season. *WHO WER*, **76** (8): 58–61. 31, 33, 36

WHO (2001b). Recommended composition of influenza virus vaccines for use in the 2002 influenza season. *WHO WER*, **76** (40): 311–4. 33, 36

WHO (2002a). Recommended composition of influenza virus vaccines for use in the 2002-2003 influenza season. *WHO WER*, **77** (8): 62–6. 30, 31, 33, 36

WHO (2002b). Recommended composition of influenza virus vaccines for use in the 2003 influenza season. *WHO WER*, **77** (41): 344–8. 33, 36

WHO (2003a). Recommended composition of influenza virus vaccines for use in the 2003-2004 influenza season. *WHO WER*, **78** (9): 58–62. 33, 34, 36, 56

WHO (2003b). Recommended composition of influenza virus vaccines for use in the 2004 influenza season. *WHO WER*, **78** (43): 375–9. 33, 34, 36

WHO (2004a). Recommended composition of influenza virus vaccines for use in the 2004-2005 influenza season. *WHO WER*, **79** (9): 88–92. 33, 34, 36

WHO (2004b). Recommended composition of influenza virus vaccines for use in the 2005 influenza season. *WHO WER*, **79** (41): 369–73. 34, 36

WHO (2005a). Recommended composition of influenza virus vaccines for use in the 2005-2006 influenza season. *WHO WER*, **80** (8): 66–71. 34, 35, 36

WHO (2005b). Recommended composition of influenza virus vaccines for use in the 2006 influenza season. *WHO WER*, **80** (40): 342–7. 33, 34, 36

WHO (2006a). Recommended composition of influenza virus vaccines for use in the 2006-2007 influenza season. *WHO WER*, **81** (9): 82–6. 4, 34, 35, 36

WHO (2006b). Recommended composition of influenza virus vaccines for use in the 2007 influenza season. *WHO WER*, **81** (41): 390–5. 4, 33, 35, 36

WHO (2007a). Recommended composition of influenza virus vaccines for use in the 2007-2008 influenza season. *WHO WER*, **82** (9): 69–74. 4, 15, 33, 35, 36

WHO (2007b). Recommended composition of influenza virus vaccines for use in the 2008 influenza season. *WHO WER*, **82** (40): 351–6. 4, 35, 36

WHO (2008a). Recommended composition of influenza virus vaccines for use in the 2008-2009 influenza season. *WHO WER*, **83** (9): 81–7. 4, 33, 35, 36

WHO (2008b). Recommended composition of influenza virus vaccines for use in the 2009 southern hemisphere influenza season. *WHO WER*, **83** (41): 366–72. 4, 35, 36

WHO (2009a). Fact sheet no211. Accessed July 2011. 1, 3, 49

WHO (2009b). Recommended composition of influenza virus vaccines for use in 2009-2010 influenza season (northern hemisphere winter). *WHO WER*, **84** (9): 65–72. 24, 35, 36

WHO (2010a). Pandemic (h1n1) 2009 - update 112. *WHO Global Alert and Response*. 7, 49

WHO (2010b). Recommended viruses for influenza vaccines for use in the 2010-2011 northern hemisphere influenza season. *WHO WER*, **85** (10): 81–92. 24, 36, 38

WHO (2011a). Antigenic and genetic characteristics of zoonotic influenza viruses and development of candidate vaccine viruses for pandemic preparedness. *WHO WER*, **86** (43): 469–80. 6

WHO (2011b). Recommended composition of influenza vaccines for use in the 2012 southern hemisphere influenza season. *WHO WER*, **86** (42): 457–68. 6

WHO (2011c). Recommended composition of influenza virus vaccines for use in the 2011-2012 northern hemisphere influenza season. *WHO WER*, **86** (10): 81–91. 49

WHO (2012). Recommended composition of influenza virus vaccines for use in the 2012-2013 northern hemisphere influenza season. *WHO WER*, **87** (10): 83–96. 6, 9

Wiley, D. and J. Skehel (1987). The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annu Rev Biochem*, **56** (1): 365–94. 24, 29

Wiley, D., I. Wilson and J. Skehel (1981). Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*, **289** (5796): 373–8. 24, 29, 57, 58, 59

Wilson, I., J. Skehel and D. Wiley (1981). Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 angst resolution. *Nature*, **289** (5796): 366–73. 58, 69

Wilson, I. a. and N. J. Cox (1990). Structural basis of immune recognition of influenza virus hemagglutinin. *Annu Rev Immunol*, **8**: 737–71. 24

Wise, H. M., A. Foeglein, J. Sun, R. M. Dalton, S. Patel *et al.* (2009). A complicated message: identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA. *J Virol*, **83** (16): 8021–31. 5

Xia, Z., G. Jin, J. Zhu and R. Zhou (2009). Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus. *Bioinformatics*, **25** (18): 2309–17. 12, 13, 25, 50

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24** (8): 1586–91. 65

Yang, Z., S. Kumar and M. Nei (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141** (4): 1641–50. 9, 60

Yang, Z. and B. Rannala (2012). Molecular phylogenetics: principles and practice. *Nat Rev Genet*, **13** (May): 303–14. 9

Zimmer, S. M. and D. S. Burke (2009). Historical perspective–Emergence of influenza A (H1N1) viruses. *N Engl J Med*, **361** (3): 279–85. 24

Zwickl, D. (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.* Ph.D. thesis, The University of Texas at Austin. 27, 65

APPENDIX A

Journal versions of the published articles

Nucleic Acids Research Advance Access published October 18, 2010

Nucleic Acids Research, 2010, 1–12
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).

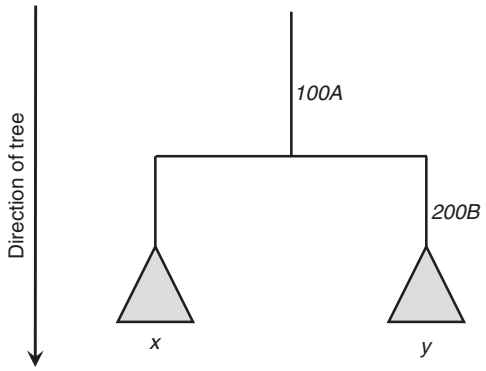


Figure 1. A tree demonstrating the concepts of alleles and allele frequency correction. For allele *100A*, only the isolates of subtree *x* are counted, whereas for allele *200B *100A**, the isolates in subtree *y* are considered.

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).

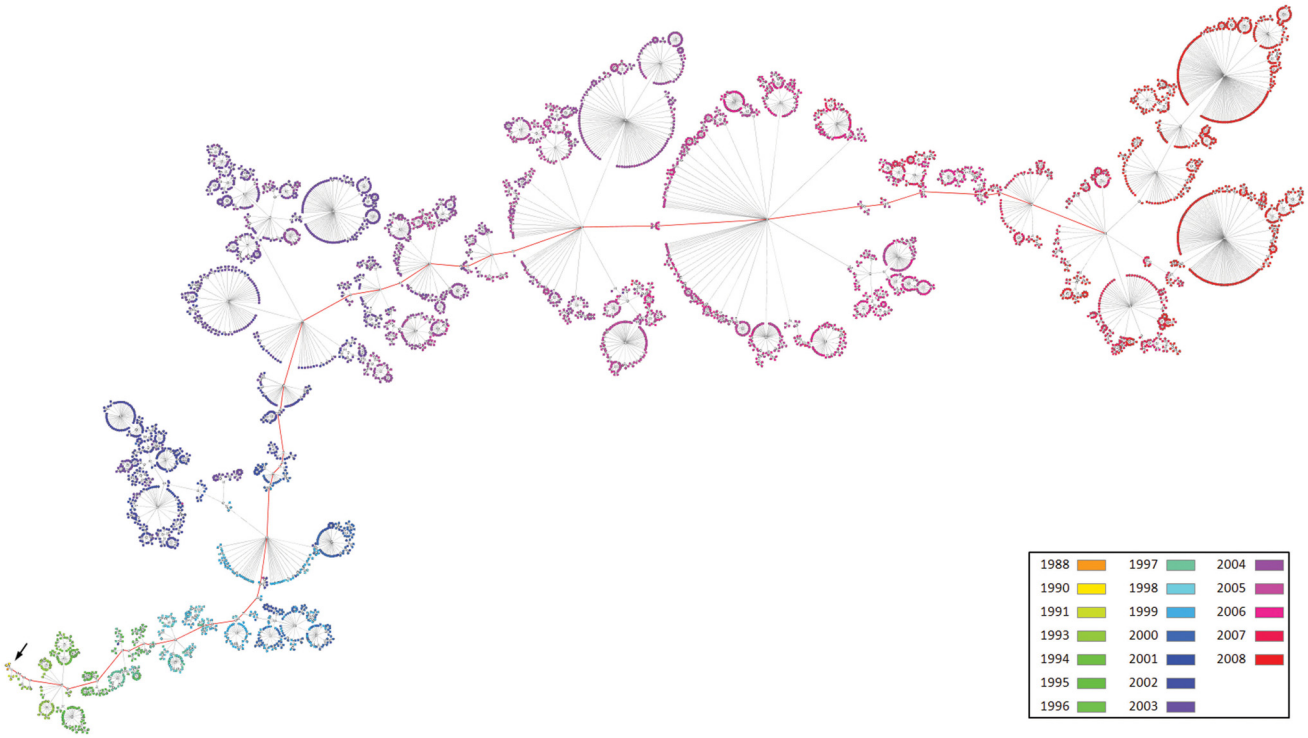


Figure 2. Maximum likelihood tree topology inferred for 4913 hemagglutinin sequences of seasonal human influenza A (H3N2). Leaf nodes are color-coded according to the sampling dates of the viral isolates. The first sampled isolate, A/Siena/3/1988, is indicated with an arrow. The trunk of the tree (i.e.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

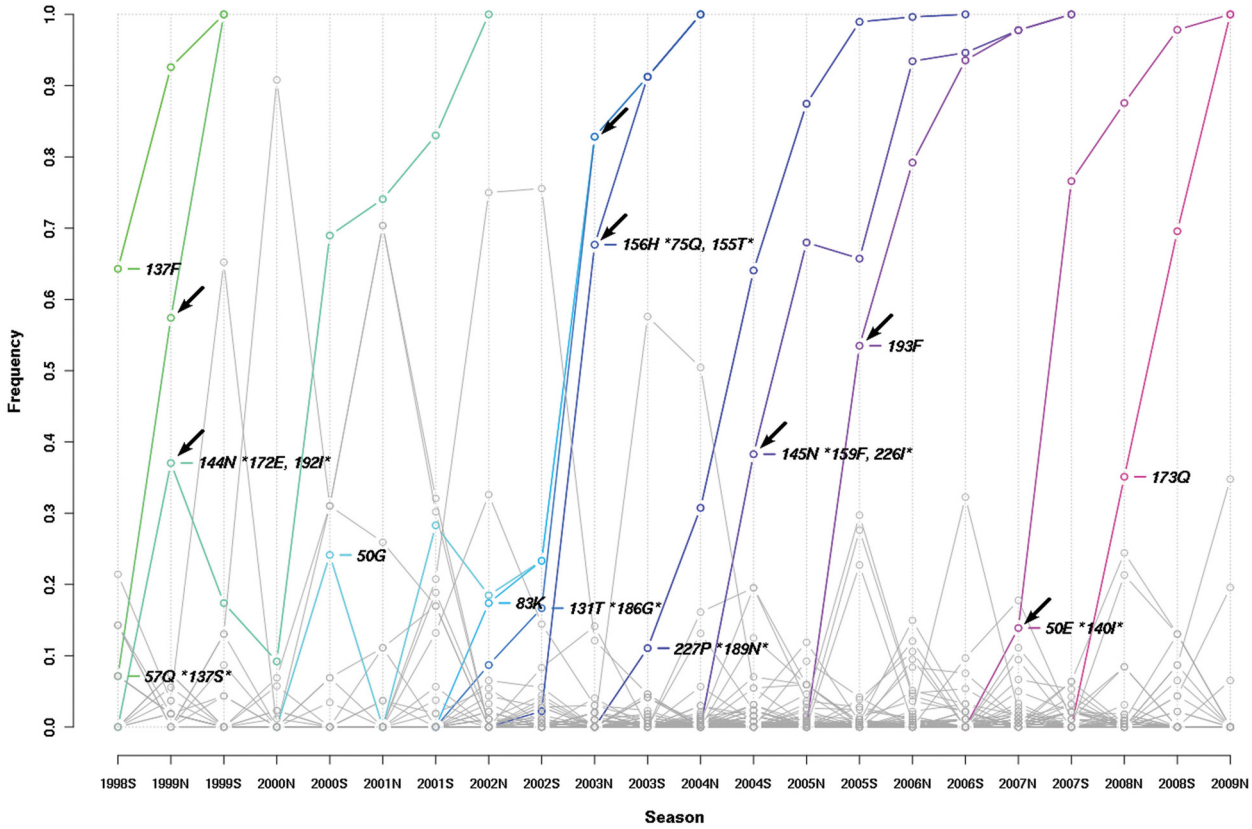


Figure 3. Allele dynamics plot for the major surface protein and antigenic determinant of the seasonal influenza A (H3N2) virus. The Northern and Southern influenza seasons from 1998 to 2008 are shown. Alleles that reach a prevalence of more than 95% and are subsequently fixed are shown in color; all 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

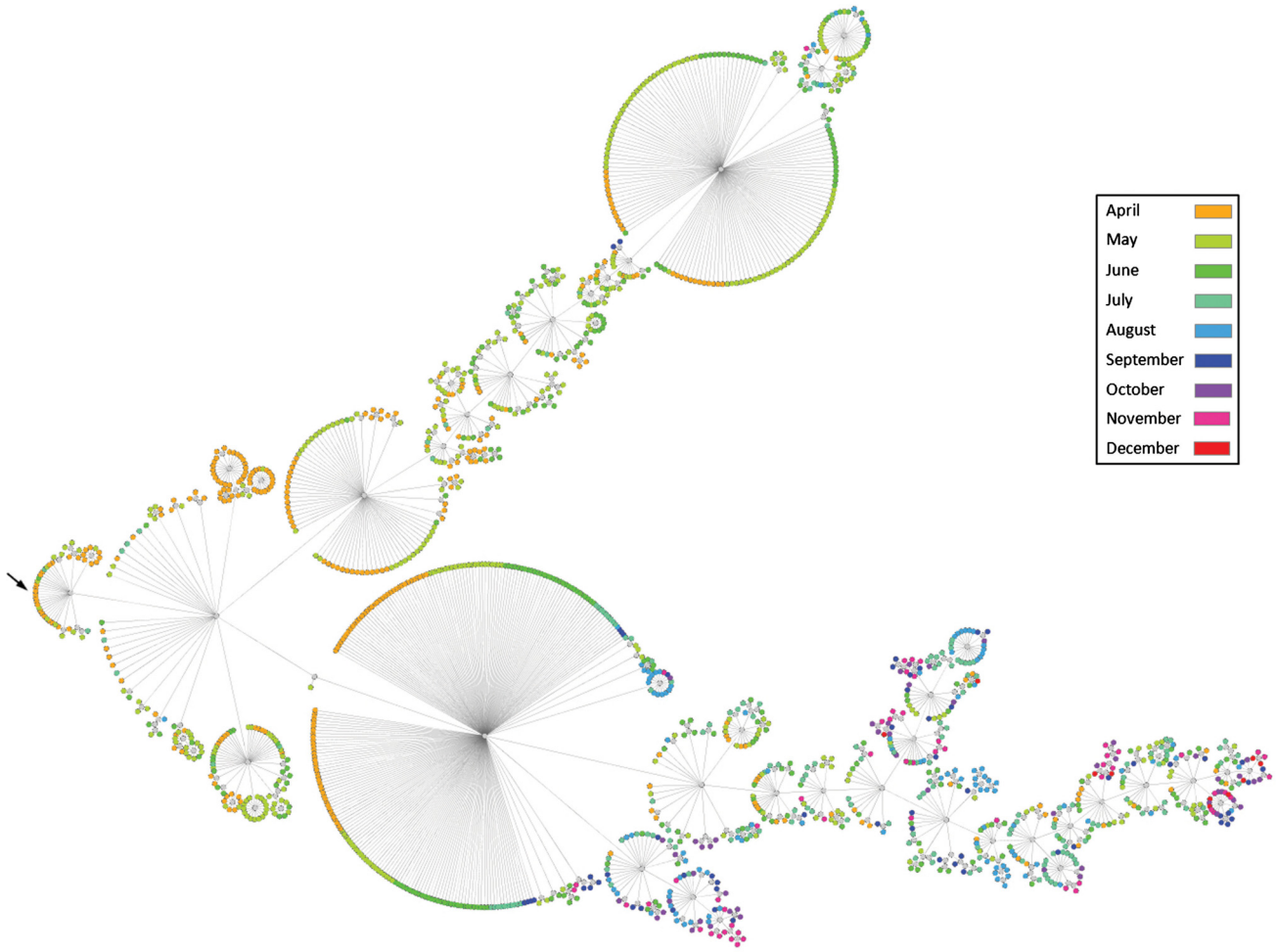


Figure 4. Maximum likelihood tree topology inferred from 1516 2009 swine-origin influenza A (H1N1) hemagglutinin sequences. Leaf nodes are color-coded according to the sampling dates of the viral isolates. The first sampled isolate, A/California/05/2009, is indicated with an arrow.

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.

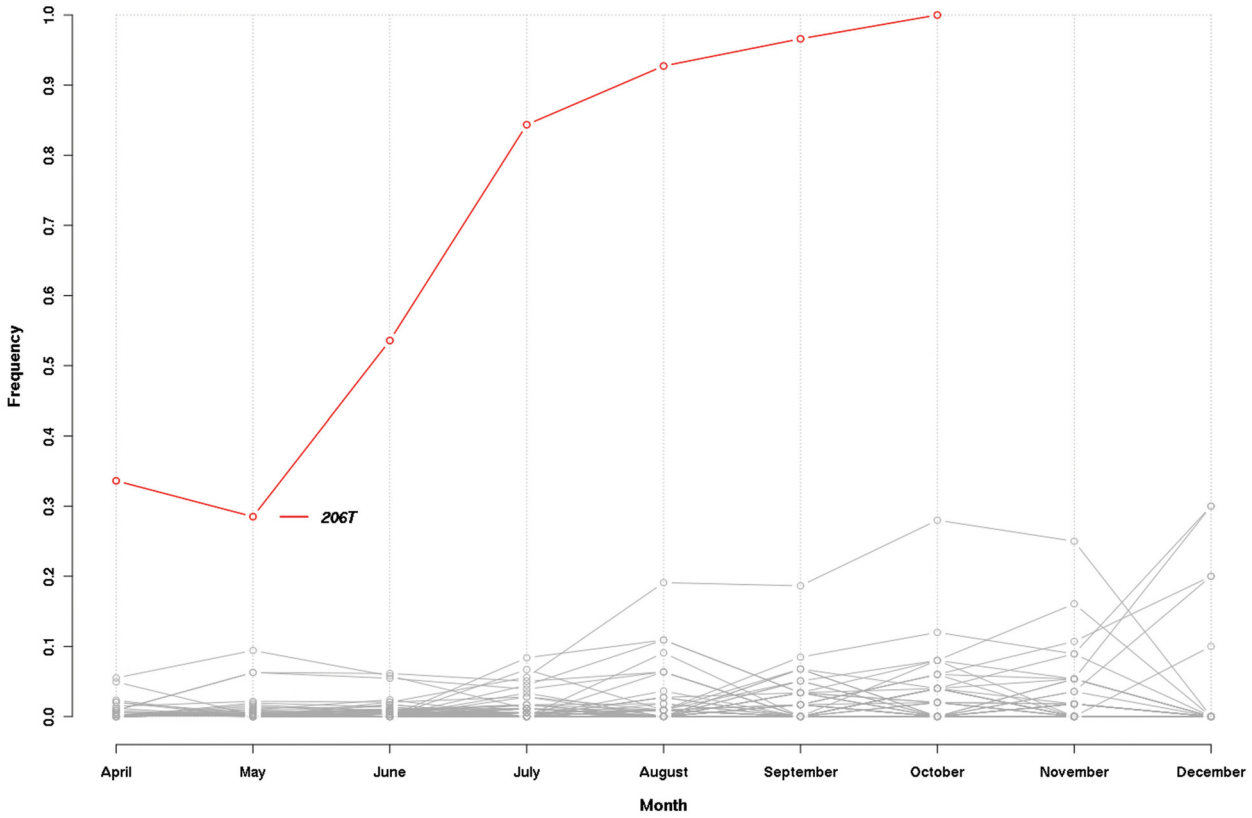


Figure 5. Allele dynamics plot for the major surface protein and antigenic determinant of the new influenza A (H1N1) based on sequences sampled between April and December of 2009 without allele frequency correction. Alleles that reach a prevalence of more than 95% and are subsequently fixed are shown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

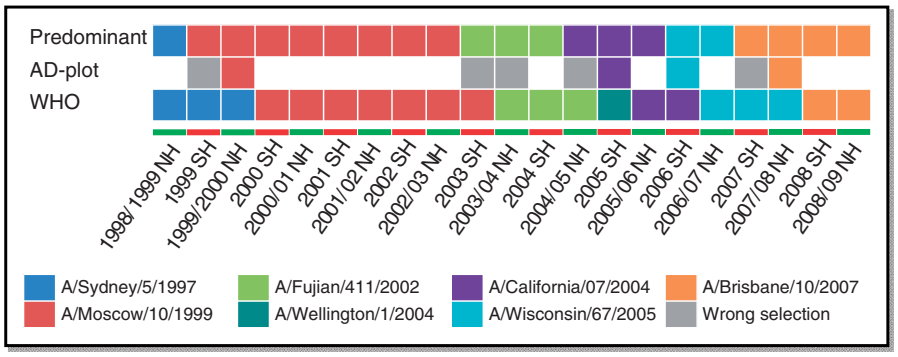


Figure 6. Comparison of predominant influenza A (H3N2) strains, WHO vaccine strain recommendations and strains identified by AD plot analysis. For the AD plot analysis, seasons with antigenic cluster transitions are shown in color. The information shown for the AD plot and the WHO recommendation represents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

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Inference of Genotype–Phenotype Relationships in the Antigenic Evolution of Human Influenza A (H3N2) Viruses

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

1 Department for Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, Germany, **2** Max-Planck Research Group for Computational Genomics and Epidemiology, Max-Planck Institute for Informatics, Saarbrücken, Germany

Abstract

Distinguishing mutations that determine an organism's phenotype from (near-) neutral 'hitchhikers' is a fundamental challenge in genome research, and is relevant for numerous medical and biotechnological applications. For human influenza viruses, recognizing changes in the antigenic phenotype and a strains' capability to evade pre-existing host immunity is important for the production of efficient vaccines. We have developed a method for inferring 'antigenic trees' for the major viral surface protein hemagglutinin. In the antigenic tree, antigenic weights are assigned to all tree branches, which allows us to resolve the antigenic impact of the associated amino acid changes. Our technique predicted antigenic distances with comparable accuracy to antigenic cartography. Additionally, it identified both known and novel sites, and amino acid changes with antigenic impact in the evolution of influenza A (H3N2) viruses from 1968 to 2003. The technique can also be applied for inference of 'phenotype trees' and genotype–phenotype relationships from other types of pairwise phenotype distances.

Citation: Steinbrück L, McHardy AC (2012) Inference of Genotype–Phenotype Relationships in the Antigenic Evolution of Human Influenza A (H3N2) Viruses. PLoS Comput Biol 8(4): e1002492. doi:10.1371/journal.pcbi.1002492

Editor: Neil Ferguson, Imperial College London, United Kingdom

Received November 10, 2011; **Accepted** March 9, 2012; **Published** April 19, 2012

Funding: A.C.M. and L.S. were funded by Max-Planck society and Heinrich Heine University Düsseldorf. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: alice.mchardy@uni-duesseldorf.de

Introduction

Influenza viruses are responsible for ~500,000 deaths annually and are a substantial threat to human health [1]. Besides seasonal infections caused by human viruses, four major pandemics over the last 100 years have resulted in ~50 million deaths worldwide [2–4]. The viruses are classified into three genera (A, B, C), all from the *Orthomyxoviridae* family, which comprises single-stranded, negative sense RNA viruses. Influenza A and B viruses evolve rapidly and continuously accumulate amino acid changes in the antibody-binding (epitope) sites of the surface proteins, resulting in changes in antigenicity. Thus, novel 'antigenic types' regularly appear and rise to predominance, causing worldwide epidemics despite existing vaccination programs [5,6]. Influenza A viruses are further categorized into subtypes based on the composition of their surface proteins, hemagglutinin (H or HA) and neuraminidase (N or NA). In the human population, the subtypes H1N1 and H3N2 are currently circulating [7]. Both global population structure and geographic migration patterns are known to influence the evolution of H3N2. Russell *et al.* suggested East–Southeast Asia to serve as a global reservoir, from which seasonal epidemics in temperate zones are seeded [8]. Other regions, such as China or USA, might serve as seeding regions, too, and migration from and to other tropical regions than East-Southeast Asia is thought to have a significant influence on the global dynamics [9,10].

To monitor genetic and antigenic changes, the World Health Organization (WHO) runs a global surveillance program [11]. Quantification of viral antigenic phenotypes is done with the hemagglutination inhibition (HI) assay, which measures the ability of an antiserum to inhibit the agglutination of red blood cells by a viral antigen [12]. Antigenic cartography, involving multidimensional scaling of log-normalized HI titers, subsequently generates an accurate low-dimensional representation of the antigenic distances between antigen–antiserum pairs [5,13]. If a novel antigenic type with increasing prevalence is detected, the vaccine composition, consisting of two strains of influenza A (H3N2 and H1N1) and one strain of influenza B, is updated to include an antigenically closer match.

Antigenic cartography of influenza A (H3N2) isolates from 1968 to 2003 revealed that antigenic types circulate for 3.3 years, on average, in worldwide epidemics before being replaced by a successor [5]. A comparison of antigenic and genetic maps showed that, the antigenic impact of genetic changes varies, depending on the nature of the amino acids exchanged, their structural positioning and epistatic interactions with other sites. Subsequent studies have incorporated both antigenic and genetic data for predicting antigenically novel strains [14–16]. Additionally, many groups have investigated the influence of sequence positions and sequence variation on viral evolution, based on different computational criteria [17–24].

Even though the general principles governing the antigenic evolution of influenza A viruses are well studied, computational methods for directly determining the antigenic impact of individual amino acid exchanges do not yet exist. Such analyses

Author Summary

The molecular evolution of any organism is described by changes in the genotype resulting from genetic drift or selection to maintain or establish fitness under the given environmental conditions. Identification of phenotype-defining changes and their distinction from (near-) neutral ('hitchhikers') ones is a fundamental challenge in genome research. The standard approach involves time- and cost-intensive mutation experiments, which are typically low throughput, due to their experimental nature. We have developed a computational method for the inference of phenotypic impact of genotypic changes that is applicable to any system, within or across species, where homologous genetic sequences and associated pairwise phenotype distances are available. We demonstrate the accuracy of our method by application to the human influenza A (H3N2) virus. This exemplary system is of particular interest, as recognizing changes in the antigenic phenotype and a viral strains' capability to evade pre-existing host immunity is important for the production of efficient vaccines. We accurately identified known sites and amino acid changes with antigenic impact over 35 years of evolution, and provide further details on individual antigenically relevant changes in the evolution of influenza A (H3N2) viruses.

currently require time- and cost-intensive experimental characterization of mutant viruses [5]. On the other end of the spectrum, antigenic cartography allows identification of 'cluster difference substitutions', comprising all near-conserved changes that distinguish consecutive antigenic clusters.

We describe a method for the inference of 'antigenic trees', which is based on a least-squares optimization (LSO) procedure of fitting pairwise antigenic distances onto an evolutionary tree for the major antigenic determinant of influenza A. It is a computational method allowing for a more fine-grained resolution of the antigenic impact of individual changes than antigenic cartography without time- and cost-intensive experiments. Application to HA sequences and serological data from human influenza A (H3N2) viral isolates from 1968 to 2003 determined the antigenic impact of all branch-associated amino acid changes for this time period. Our technique identified known antigenic types and the amino acid changes associated with the type transitions. For sufficiently resolved branches, the antigenic impact of individual exchanges could be quantified. The method furthermore found known and novel key HA sites and changes in antigenic evolution.

Results

We applied our method to infer an antigenic tree from genetic sequences of the hemagglutinin segment and serological data (HI titers of antigen-antiserum pairs) for 258 influenza A (H3N2) isolates sampled between 1968 and 2003 [5]. Antigenic branch lengths were determined by fitting the antigenic distances between viral isolates (the antigens) and antisera raised against reference strains to the branches of a maximum likelihood tree (see Materials & Methods). Antigenic branch lengths were realized as two independent weights (up and down) and represented the antigenic properties of antigens and antisera in the tree. The antigenic path length between two isolates, corresponding to the sum of the branch weights (either up- or down-weight, depending on the direction in the tree) for all connecting branches on the path between them in the tree, reflected their overall antigenic distance (**Figure 1**, high resolution **Figures S1 and S2**).

To investigate how accurately antigenic distances were fitted onto the tree, we evaluated its ability to predict unseen antigenic distances by leave-one-out cross validation [25]. In this experiment, an antigenic tree is inferred from all but one antigenic distance and then is applied to predict the left out distance. A predicted distance corresponds to the antigenic path length between the two respective isolates in the tree (see Materials & Methods). This was repeated for every antigenic distance and the overall accuracy of predicting antigenic distances estimated by the absolute prediction error and the root mean squared error (RSME) averaged over all leave-one-out experiments (see Materials & Methods). The leave-one-out absolute prediction error was 0.86 antigenic units ($\sim$ a two-fold dilution, SD 0.72) and the correlation measured by Pearson's correlation coefficient between predicted and measured values was 0.86. Using placement on an antigenic map estimated from the same data, Smith *et al.* report an average absolute prediction error of 0.83 antigenic units (SD 0.67) and a Pearson's correlation coefficient of 0.80 for 481 measurements of antigenic distances [5]. The RMSE penalizes large prediction errors more than small prediction errors, and is a well suited measure of predictive accuracy. For our method, the leave-one-out RMSE is 1.12 antigenic units, corresponding to approximately a two-fold dilution. This is comparable to the ten-fold cross validation RMSE of Cai *et al.* on this data set (1.05 antigenic units) [26], who used a matrix completion algorithm prior to multi dimensional scaling. Our method therefore performs similarly to antigenic cartography in predicting antigenic distances, with a slightly larger error but also a slightly higher correlation between predicted and measured values. This is despite the fact that inferring antigenic branch lengths for an antigenic tree allows far fewer degrees of freedom than an antigenic map, where the data is not forced on a fixed structure. Note that for the prediction of antigenic distances, other well-suited methods also exist [26,27].

As we infer a tree topology from nucleotide sequences, branches might be without any amino acid changes and thus lack explanatory power if they are assigned antigenic weights. This allows accommodating measurement errors in HI titers in antigenic branch weights or variation caused by changes in other viral antigens, such as the surface glycoprotein neuraminidase. HI titers are imprecise, as they reflect two-fold dilutions instead of quantitative estimates, and are often highly variable, with measurements varying between experiments and laboratories. For instance, the two isolates A/Finland/220/92 and A/Stockholm/20/91 have the same nucleotide sequence, and hence no changes on their respective tip branches (tips), but differ strongly in their HI values, where A/Finland/220/92 shows an antigenic distance from the same antisera that is, on average, $\sim$1.0 antigenic units (a two-fold dilution) larger than that of A/Stockholm/20/91. Note that, in general, even though neuraminidase may influence the HI titers, the WHO recommends application of the HI assay under conditions where its influence is negligible [28]. To incorporate a possible influence of neuraminidase activity one may use concatenated viral sequences (hemagglutinin and neuraminidase) and fit antigenic distances on a tree topology inferred from these sequences. If doing so, one should first ensure that reassortment events have not resulted in larger topological changes between the HA and NA genealogies during the analyzed time period [6,29]. In case of larger topopological changes due to segment reassortment, a joint tree is inferred for data which cannot be described by a tree-like evolutionary history, overall, and the results are likely to be only partially informative.

On average, internal branches *without* amino acid changes have weights of 0.30 (up) and 0.21 (down), respectively. Less noise

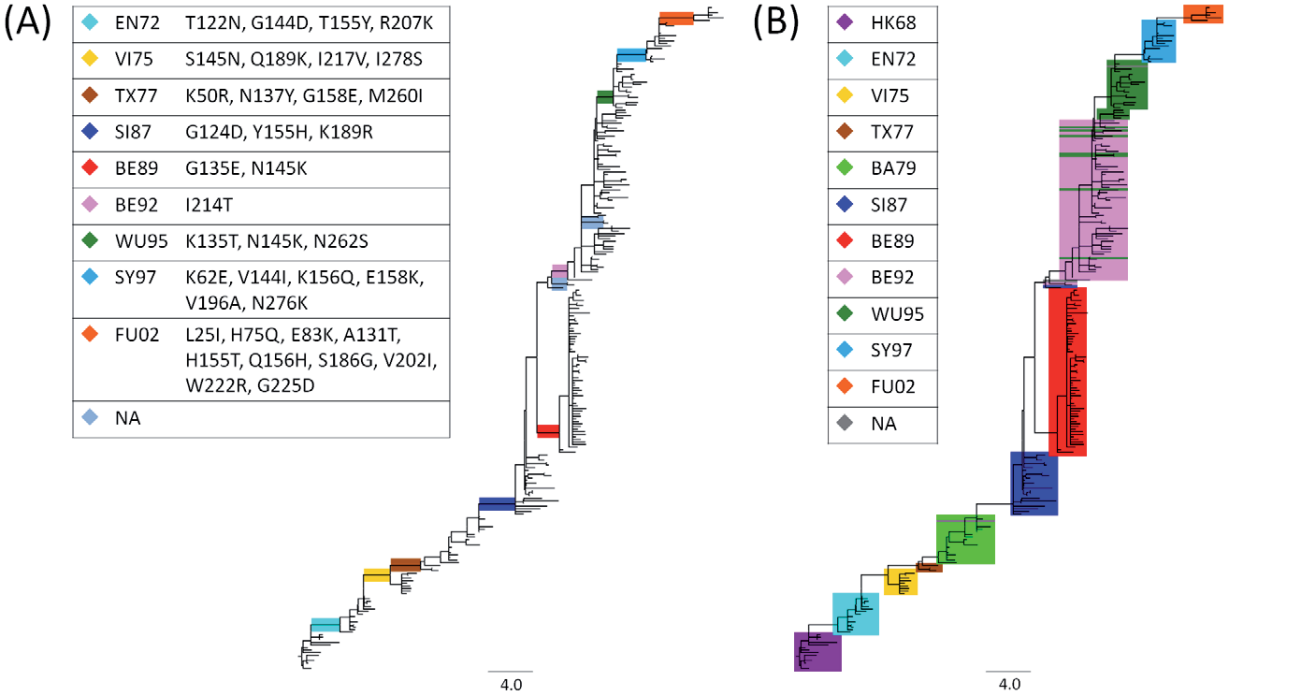


Figure 1. Antigenic tree for influenza A (H3N2) viruses. Branch lengths represent antigenic distances (maximum of up- and down-weights for each branch) inferred from a maximum likelihood tree of 258 hemagglutinin sequences of seasonal influenza A (H3N2) virus isolates and serological data. (A) Colored edges show antigenic type transitions, with internal branches with high average antigenic weights (≥1.0 antigenic units). Gray-blue edges represent high weight branches leading to a subtree with three isolates or less, representing low abundance types. (B) Isolates are color-coded by antigenic clusters according to Smith *et al.* (2004). Three isolates (A/Christchurch/4/85, A/Hong Kong/34/90 and A/Netherlands/172/96) are only present as antisera and were not assigned a cluster label.
doi:10.1371/journal.pcbi.1002492.g001

occurs on the tree trunk, which represents the viral lineage surviving over time, with 0.19 (up-weight) and 0.19 (down-weight) assigned, on average. Interestingly, the average antigenic weight of branches *with* amino acid changes is higher on the tree trunk than for all internal branches (up = 0.52, down = 0.61 vs. up = 0.44, down = 0.46). This is in agreement with an expected fitness advantage for viral isolates with larger antigenic changes, and therefore preferential fixation and establishment appear as changes on the tree trunk.

Antigenic types resolved in the tree

Antigenic types are clearly distinguished by high average weights (≥1.0 antigenic units) in the antigenic tree (see Materials & Methods). Exclusion of branches leading to subtrees with three or less isolates, representing undersampled groups, identified nine branches defining type transitions (**Table 1**) and ten antigenic types. Abbreviations for these (HK68, EN72, VI75, TX77, SI87, BE89, BE92, WU95, SY97 and FU02) are used as in Smith *et al.* (2004) [5]. SY97, for instance, denotes antigenically similar A/Sydney/5/1997-like strains. The average antigenic distances of these branches range from 1.0 (SI87–BE89) to 2.6 antigenic units (WU95–SY97; **Table 1**, **Figure 1A**). Eight of the nine type transition branches are on the trunk of the tree, which represents the influenza A (H3N2) lineage surviving over time. An exception is BE89, which is located in a subtree that has become extinct.

The setting of the threshold parameter for identification of antigenic types in the tree influences the performance of our method (**Table S5**). The selected threshold of 1.0 antigenic unit identified nine of ten antigenic type transitions found by antigenic cartography [5]. The TX77–BA79 transition was not predicted with our method in this setting, as the weights of the corresponding branch were slightly below the threshold (up-weight 1.4, down-weight 0.0). Our method resolves antigenically relevant changes between successive antigenic types in several cases to several successive branches. Therefore, a higher threshold of 2.0 antigenic units for individual branches (a four-fold dilution), as suggested to distinguish antigenically diverse viral strains [11], does not allow distinction between different antigenic groups (only if the transition is not well resolved in the data and the antigenic impact of multiple changes is summarized on a single branch). On the other hand, choosing a lower threshold of 0.5 antigenic units selects twelve additional type-defining branches (**Table S5**, **Figure S3**). Among these is the TX77-BA79 type-defining branch that corresponds to an antigenic cluster transition according to antigenic cartography [5]. Furthermore, four of these additional branches define antigenic subtypes that were distinct enough to warrant a vaccine update. A more detailed discussion of type-defining branches at the threshold of 0.5 antigenic units can be found in the supporting material (**Text S1**). Note, that the choice of the threshold distance is equivalent to find a minimal antigenic distance to distinguish groups of antigenically and genetically similar viral isolates. This is different from the question whether two specific viral isolates are antigenically similar or not, although both tasks are related to each other.

For the nine jointly identified type transitions, seven agree 100% in terms of the assigned viral isolates. For the BE89–BE92 transition, the isolate A/Netherlands/938/1992 is placed within BE92 using antigenic cartography and as preceding BE92 by our technique. Isolate assignment differs the most for the BE92–WU95 transition. This is likely to be caused by multiple occurrences of

Table 1. Internal branches with high average antigenic weights (≥1.0 antigenic units) and according antigenic types in comparison to antigenic clusters identified by antigenic cartography (branches leading to three or less isolates are excluded).

Type transition	Branch amino acid changes	Weights (up/down/avg)	Trunk	Additional amino acid changes	Weights (up/down/avg)	Trunk	Smith *et al.*
HK68–EN72	**T122N**, **G144D**, **T155Y**, **R207K**	2.6/0.4/1.5	x	L3F, **N188D**	0.9/0.2/0.5	x	3.4
EN72–VI75	**S145N**, **Q189K**, **I217V**, **I278S**	0.6/2.4/1.5	x	**N53D**, **N137S**, **L164Q**, **F174S**, **N193D**, **R201K**, **I213V**, **I230V**	0.0/1.0/0.5		4.4
VI75–TX77	**K50R**, **N137Y**, **G158E**, **M260I**	0.6/2.8/1.7	x	**E82K**	1.0/-/0.5		3.4
TX77–BA79				**N133S**, **P143S**, **G146S**, **K156E**, **T160K**, **Q197R**, **V217I**	1.4/0.0/0.7	x	3.3
				D2N, **N53D**, **N54S**, **I62K**, **D172G**, **V244L**	0.0/0.3/0.2	x	
BA79–SI87	**G124D**, **Y155H**, **K189R**	0.2/3.3/1.7	x				4.9
SI87–BE89	G135E, **N145K**	2.0/0.0/1.0					4.6
BE89–BE92	I214T	1.4/1.1/1.3	x	**E156K**, **E190D**, N193S, L226Q, **T262N**	1.0/0.0/0.5	x	7.8
				S133D	0.0/0.4/0.2	x	
BE92–WU95	K135T, **N145K**, N262S	1.5/1.1/1.3	x				4.6
WU95–SY97	**K62E**, V144I, **K156Q**, **E158K**, **V196A**, **N276K**	2.5/2.6/2.6	x				4.7
SY97–FU02	**L25I**, **R50G**, **H75Q**, **E83K**, **A131T**, **H155T**, **Q156H**, S186G, **V202I**, **W222R**, **G225D**	1.8/3.2/2.5	x				3.5

Branch amino acid changes indicate the corresponding branches, where changes in bold were also found by Smith *et al.* (2004), and weights give the respective up, down and average branch weights. Multiple branches that can be mapped to a single antigenic type are separated by dashed lines. Additional amino acid changes indicate branches that carry further mutations found to be cluster transition substitutions by Smith *et al.* (2004). For some branches, the down-weight was not defined, as no antiserum was in the respective subtree. Branches that can be mapped to multiple type transitions are shown at the first mapping only. Smith *et al.* (2004) present average distances between consecutive antigenic clusters, whereas average antigenic branch weights give a minimum distance between consecutive antigenic types. Note that on branches with multiple changes not all changes have to contribute to the antigenic weight, though their individual impacts could not be resolved with the dataset (unsampled viral isolates).

doi:10.1371/journal.pcbi.1002492.t001

N145K, which is, according to Smith *et al.* (2004) [5], the change that defines the BE92–WU95 transition and has a major antigenic impact in that context (2.6 antigenic units). It was already noted by Smith *et al.* that isolates classified by antigenic cartography within WU95 are placed in the vicinity of BE92 in a tree. Our analysis agrees with these findings (**Figure 1B**). We found that for each branch adjacent to these disagreeing placements, N145K is present (isolates of the antigenic type WU95 located in the area of BE92), with large branch-associated antigenic weights (an average up-weight of 1.3), similar to the type-defining branch of WU95 (up-weight 1.5). This indicates that N145K has a large antigenic impact for all these isolates and, interestingly, was evolutionary volatile during that period.

Analysis of up- and down-weights for type-defining branches allows us to determine a direction for antigenic impact. For example, the branch separating HK68 and EN72 has a weight of 2.6 (up)/0.4 (down), which means that isolates of HK68 are antigenically more similar to sera raised against EN72 than vice versa. The opposite example represents the SY97–FU02 transition, where the corresponding branch weight is 1.8 (up)/3.2 (down), which means that SY97 isolates are more distant from antisera raised against FU02 than vice versa. Both examples are in agreement with results published by the WHO [30,31].

As influenza A evolution in the analyzed data set is characterized by an underlying cluster structure, both antigenic types and antigenic clusters allow determination of cluster-difference or antigenic type associated substitutions. However, antigenic types (inferred by our method) and antigenic clusters (inferred by antigenic cartography) have different interpretations. Antigenic types represent sets of viral isolates showing similar evolutionary (defined by the phylogenetic tree) and antigenic (defined by the antigenic branch lengths) patterns. Antigenic cluster are solely defined by antigenic patterns and are determined by a k-means clustering approach. In datasets with less well-defined cluster structure, the k-means approach would hardly result in robust clusters and identification of phenotype-associated changes would be more difficult, whereas our method would likely be able to resolve phenotype-genotype relationships up to the level of resolution supported by the data.

Substitutions in antigenic type transitions

Amino acid changes from eight of nine type transitions identified by both antigenic cartography and the antigenic tree include the cluster difference substitutions described in Smith *et al.* (2004) [5] (**Table 1**). Smith *et al.* define 'cluster difference substitutions' as changes in conserved residues between two consecutive antigenic clusters (conserved meaning present in at least $n-1$ isolates within a cluster of size n). For five transitions, all cluster difference substitutions are on the type-defining branch (BA79–SI87, SI87–BE89, BE92–WU95, WU95–SY97 and SY97–FU02). For three transitions (EN72–VI75, VI75–TX77 and HK68–EN72), the substitutions were resolved to several branches with different antigenic branch weights, which allows a more fine-grained distinction. The 12 substitutions of the EN72–VI75 transition were assigned to two consecutive branches, one with high and one with moderate antigenic weights. The branch with S145N, Q189K, I217V and I278S has a high antigenic weight, indicating that one or several of these have a very large antigenic impact. For the HK68–EN72 and the VI75–TX77 transitions, the substitutions were resolved to two consecutive branches with high and moderate antigenic weights, too.

For BE89–BE92, the amino acid changes differ from cluster difference substitutions. Here, the cluster difference substitutions

are found on branches that precede the type-defining branch. The type-defining branch carries the change I214T, while the cluster difference substitutions map to two preceding branches with lower antigenic weights. I214T has not been mentioned in the literature before and is reversed downwards in the tree on a branch without any assigned antigenic weight. Thus, either the measurements here were too noisy to resolve the correct branch, or this position has an antigenic impact as an epistatic effect, allowing for the preceding changes to become antigenically effective. Support for a potential epistatic effect of this change can be found by detailed analysis of individual HI measurements for two isolates (A/Hong Kong/34/1990 and A/Netherlands/938/1992), which already have the preceding branch changes for BE92 but not the I214T change. On average, all antigens labeled BE92 by Smith *et al.* have a large antigenic distance (greater 4.7 antigenic units) from the antiserum A/Hong Kong/34/1990. A/Netherlands/938/1992 is similar to A/Hong Kong/34/1990, with an antigenic distance of 0.7 to this antiserum.

Four branches with type transitions (SI87–BE89, BE92–WU95, WU95–SY97 and SY97–FU02) include additional changes besides the cluster difference substitutions. For SI87–BE89, the change G135E is present, in addition to N145K. G135E appears twice more in the tree, with an average up-weight of 0.64. This indicates that it may also have an antigenic effect in SI87–BE89. For BE92–WU95, the changes K135T and N262S are present on the type-defining branch, in addition to N145K. Both are located in the antibody binding sites [32] and became fixed following their appearance on this trunk branch.

In a recent (unpublished) study, Koel *et al.* (*Koel et al.; Antigenic evolution of influenza A (H3N2) virus is dictated by 7 residues in the hemagglutinin protein; 2nd International Influenza Meeting, Münster; 2011*) determined by site-directed mutagenesis changes at seven positions in the HA protein (145, 155, 156, 158, 159, 189 and 193) responsible for significant phenotypic diversity in the evolution of influenza A (H3N2). We also find that for eight of the nine identified type-defining branches changes occur at five of these positions (no changes at positions 159 and 193 are involved in antigenic type transitions), which further confirm the relevance of these sites for antigenic evolution (**Table 1**). Note that, besides these five residues changes at 23 other positions map to the type-defining branches which not all have to contribute to the antigenic weight, though their individual impacts could not be further resolved with the dataset (unsampled viral isolates).

Antigenic impact of individual amino acid changes and sites

We examined amino acid changes with strong antigenic relevance according to (i) the impact of all changes at a specific site and (ii) the impact of a specific change. In the first case, we determined all positions where at least three changes occurred, and the mean and median of the branch weights (up- or down-weight) were not less than one antigenic unit. Missing weights, e.g. where down-weights were not defined because no antiserum was raised for the corresponding subtree, were excluded from the calculations. Seven positions, 112, 137, 144, 155, 156, 189 and 208, satisfy these criteria (**Table S2 and S3**). All except position 112 are part of the antibody binding sites of HA1 [32]. Positions 137, 155 and 156 are also part of the receptor binding site [33]. Positions 155 and 189 may be particularly important, as all changes occur on the tree trunk and are part of type transitions. The importance of H155T and Q156H was also verified for the FU02 transition [34]. For positions 137, 144, and 156, several changes map to the tree trunk (three of six, four of nine, and one of three, respectively), indicating their antigenic relevance. Changes at position 112 explain single isolate variations, as all occur on tips. The antigenic impact of these changes may be due to hitchhiking effects, as they occur only in combination with other changes.

Next, we identified changes occurring at least three times in the tree with a mean and median antigenic weight (up- or down-weight) of more than one unit (**Table S4**). Again, missing weights were excluded from the calculations. Five changes satisfy these conditions. Four of these (K62E, N145K, L226Q and T248I) occur at positions in antibody binding sites [32]. N145K was experimentally verified to have a large antigenic impact [5]. K62E is part of the WU95–SY97 transition and has a high weight assigned on two further tips. Finally, of the eight occurrences of L226Q, seven appear between 1990 and 1996 for isolates of the BE92 type, indicative of a fitness effect for this antigenic type in particular. Interestingly, the reverse change, Q226L, is known to play a role in receptor binding specificity for the adaptation of bird viruses to the human host [35–38]. T248I had a high weight only in combination with other changes, indicating a potential epistatic effect. Besides these four changes, we identified V112I, which only appeared on tips and explains single isolate variations.

We searched for changes with moderate antigenic impact (more than 0.5 antigenic units) which identified seven further changes (**Table S2**). G135E is part of the SI87–BE89 transition (see above) and E156K was shown to impact immune escape in mice [39]. Both are located in the antibody binding sites [32]. For several additional changes, the importance was not immediately obvious, as they (i) occurred only in combination with other changes, (ii) exhibited a high weight only in combination with other changes (Q80K), (iii) only appeared on tips (S186I, S199P and V226I) or (iv) had high weights assigned only on tips and low weights on internal branches (A138T). In cases (i) and (ii), this may be the result of epistatic or hitchhiking effects, where epistasis may be more likely for (ii). Case (iii) changes are rare and explain single isolate sequence variations. This also seems to be likely in case (iv), where the effect on the tips is amplified due to other effects or amino acid changes. Notably, all case (iii) changes are also categorized as case (i) changes. Of all changes, E156K occurs once on the tree trunk. All changes appear at several points in time for different antigenic types, which indicates a potential antigenic influence. Furthermore, for five changes (G135E, A138T, E156K, S186I and V226I), the respective site was identified as being under positive selection [17].

In a recent (unpublished) study, Koel *et al.* (*Koel et al.; Antigenic evolution of influenza A (H3N2) virus is dictated by 7 residues in the hemagglutinin protein; 2nd International Influenza Meeting, Münster; 2011*) showed by site-directed mutagenesis that changes at seven positions in the HA protein (145, 155, 156, 158, 159, 189 and 193) are responsible for large antigenic changes, all except two are part of antigenic cluster transitions, over the 35 year time period. Of these, 155, 156 and 189 are also identified as generally important by our default method. If single isolate variations are excluded from the analysis, position 158 is also identified. For the other two positions (145 and 159) we identified changes with high antigenic weights (e.g. N145K and S159Y; **Table S2**). For position 193, evidence of antigenic importance could be found in our analysis if using ancestral character state reconstruction with maximum parsimony (see Supplement). Thus, our results also support the relevance of the sites proposed by Koel *et al.* (2011), even though they are not entirely comparable due to differences in experimental set up. Koel *et al.* analyzed prototype viruses with the amino acid consensus sequences of antigenic clusters and introduced only the specific changes between these prototype viruses, while our method also considers genetic and antigenic variations between other viral strains of the dataset.

Discussion

The antigenic impact of amino acid substitutions in the antigenic evolution of influenza A viruses can reliably be determined by time- and cost-intensive experimental analysis. As an alternative, we present a computational technique for inferring the antigenic impact of amino acid changes. Our method determines antigenic branch lengths for a given tree topology by fitting pairwise antigenic distances between isolates onto the tree with LSO. For inference of the tree, any state-of-the-art method can be used. A comparison between maximum likelihood, maximum parsimony and neighbor-joining trees showed that all resulted in similar prediction errors (leave-one-out absolute prediction error: 0.86, 0.87 and 0.87 antigenic units, respectively; correlation between predicted and measured by Pearson's correlation coefficient was 0.86 for all three methods). The antigenic impact of the branch-associated amino acid changes is determined by reconstructing the branch-associated amino acid changes with maximum likelihood [40]; other techniques, such as maximum parsimony or Bayesian reconstruction, could also be used [41,42]. A comparison between maximum likelihood and maximum parsimony ancestral character state reconstruction showed that these differed only in minor aspects, with the maximum likelihood reconstruction being an intermediate between accelerated and delayed transition in case of ties with maximum parsimony reconstruction. However, we did observe that more trunk branches were not assigned changes based on maximum likelihood reconstruction, which decreased the interpretability of antigenic weights in some cases.

We studied the antigenic evolution of the influenza A (H3N2) virus from 1968 to 2003 with antigenic trees inferred from data described in Smith *et al.* (2004) [5]. This allowed us to identify areas and branches in the tree corresponding to known antigenic types and transitions between these types. Analysis of antigenic weights identified seven sites in the HA1 domain of HA that were repeatedly associated with high antigenic impact. Additionally, our method identified five amino acid changes with high antigenic weights at several places in the antigenic tree. The sites and substitutions identified by our method may be of particular relevance for influenza A (H3N2) virus antigenic evolution, which has not been described before. For six of the seven positions found by site-directed mutagenesis to defining antigenic clusters for the 35 year time period (*Koel et al.; Antigenic evolution of influenza A (H3N2) virus is dictated by 7 residues in the hemagglutinin protein; 2nd International Influenza Meeting, Münster; 2011*), changes with high antigenic weights were identified with our technique, thus further supporting their relevance for influenza A (H3N2) evolution. The additional sites detected by our method could be more relevant for genetic and antigenic variations between viral strains in our data set not resulting in antigenic cluster transitions. These were not analyzed by Koel *et al.*, who characterized antigenic differences of prototype viruses with the amino acid consensus sequences of the antigenic clusters.

As the dataset covers 35 years of viral evolution with a relatively small number of isolates, not all substitutions could be resolved to individual branches and their individual antigenic impacts inferred. A denser sampling of data points would allow a more precise decoding of the genotype–antigenicity relationships, as viral isolates were unevenly sampled across the 35 years. The median number of viral isolates available per year between 1989 and 1997 was 15, whereas for the remaining years only three isolates per year were sampled (median). This unequal sampling is reflected in resolution of mutations to specific branches. Between 1989 and 1997, 19% of the branches with assigned changes carry three or more changes, whereas for the other years this is the case for 37% of the branches.

Our method allows inference of genotype to phenotype relationships from genetic sequences and associated pairwise phenotypic distances between individuals of a population or different taxa. We demonstrated the usefulness of this technique for analyzing the antigenic impact of amino acid changes in the evolution of human influenza A. An application of our method could be in influenza A virus surveillance. Here, it could be used to identify isolates and associated changes with large antigenic impact, which need to be identified for vaccine strain updates prior to an antigenic type transitions [43]. However, our method is not restricted to the analysis of influenza viruses or antigenic distance information but can be applied to the study of any system, be it within or across species, where homologous genetic sequences and associated pairwise phenotype distances are available. The software is available upon request from the authors.

Materials and Methods

Inferring the phenotypic impact of amino acid changes in protein evolution

Our idea is to adapt the least-squares optimization (LSO) technique of Cavalli-Sforza and Edwards [44] for phylogenetic inference to the problem of identifying the phenotypic impact of amino acid changes in protein evolution. The original method of Cavalli-Sforza and Edwards [44] identifies branch weights representing genetic distances according to the least-squares criterion for a tree topology. We applied this technique to infer 'antigenic trees', representing the antigenic evolution of the major surface protein of human influenza A virus (H3N2) over a 35-year period. In our adaptation, branch lengths represent antigenic distances inferred from HI assay data for human influenza A viruses and a maximum likelihood tree of the HA1 domain of hemagglutinin. Reconstruction of the amino acid changes associated with the branches of the tree allows us to infer the antigenic impact of the branch-associated amino acid changes. If sufficient data is available to resolve individual changes to individual branches, our method returns an estimate of the antigenic impact of the individual exchanges.

In LSO, one minimizes the sum of squares between the given distances D and predicted distances d:

$$Q = \sum_{i=1}^{n} \sum_{j:j \neq i} w_{i,j} \left(D_{i,j} - d_{i,j}\right)^2,$$

where W is the weight matrix for the different error terms, which were set to one here. The predicted distances $d_{i,j}$ are the sum of the branch weights on the path between leaf i and leaf j. Here, $d_{i,j} = \sum_k x_{i,j,k} v_k$, where $x_{i,j,k}$ equals one if branch k is on the path between leaves i and j in the phylogenetic tree and zero otherwise. Thus, we search for the best setting for the branch weights v_k. While evolutionary distances are usually used in this approach, here, we map antigenic distances to represent branch-specific weights. To restrict the branch weights to positive values, we used the Lawson–Hanson algorithm for non-negative LSO [45]. Because the antigenic distances here are asymmetric (i.e. $d_{i,j} \neq d_{j,i}$) and because the antigen and antiserum raised against the same viral strain do not necessarily have the same position in the antigenic space [13], we introduce the concept of up–down trees. In up–down trees, viral strains are mapped to the leaves representing the corresponding antigen as well as the antiserum, and every branch is assigned two independent weights, the up- and

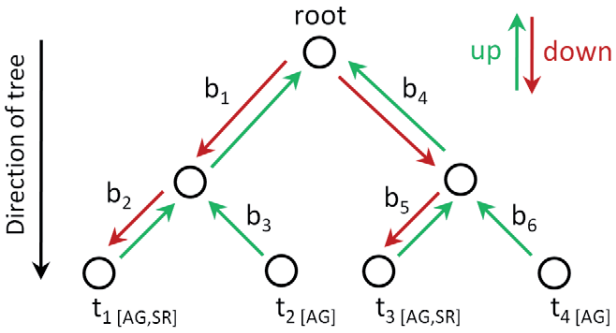
the down-weight. Every path between two taxa i and j in the tree can be separated into the set of branches from taxon i to the least common ancestor (LCA) of i and j, and the branches from taxon j to the LCA. Now, the path between antigen i and antiserum j involves only the up-weights on branches from taxon i to the LCA and only the down-weights on branches from taxon j to the LCA (**Figure 2**).

Performance measures

To evaluate how accurately antigenic distances were fitted onto the tree, we used four performance measures in leave-one-out cross validation experiments: mean absolute error (MAE), root mean squared error (RMSE), standard deviation (SD) and Pearson's correlation coefficient (CC). In leave-one-out cross validation, an antigenic tree is inferred from all but one antigenic distances and then is applied to predict the left out distance. A predicted distance corresponds to the antigenic path length between the two respective isolates in the tree (see above). This was repeated for every antigenic distance. Given n observed distances $D_{i,j}$ and predicted distances $d_{i,j}$ the performance measures are defined as follows:

$$MAE = \frac{1}{n}\sum_{i,j}|D_{i,j}-d_{i,j}|,$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i,j}(D_{i,j}-d_{i,j})^2},$$



	b_1	b_2	b_3	b_4	b_5	b_6
$t_1 \rightarrow t_3$	1/0	1/0	0/-	0/1	0/1	0/-
$t_2 \rightarrow t_1$	0/0	0/1	1/-	0/0	0/0	0/-
$t_1 \rightarrow t_1$	0/0	1/1	0/-	0/0	0/0	0/-

Figure 2. Schematic drawing demonstrating the up/down tree concept. For the two taxa t_2 and t_4, no antiserum is present, and thus, b_3 and b_6 only have up-weights. A path from t_1 to t_3 would use the up-weights of branch b_1 and b_2, and the down-weights of branch b_4 and b_5. Similarly, the path from t_2 to t_1 would use the up-weight of branch b_3 and the down-weight of branch b_2. Notably, the path from t_1 to t_1, namely the antigenic distance from antigen t_1 to the antiserum raised against strain t_1, would use the up-weight and the down-weight of branch b_1.
doi:10.1371/journal.pcbi.1002492.g002

$$SD = \sqrt{\frac{1}{n}\sum_{i}(x_i-\mu)^2} \text{ with } \mu = \frac{1}{n}\sum_{i}x_i \text{ and } x_i = |D_{i,j}-d_{i,j}|,$$

$$CC = \frac{\sum_{i,j}(D_{i,j}-\mu_D)(d_{i,j}-\mu_d)}{\sqrt{\sum_{i,j}(D_{i,j}-\mu_D)^2}\sqrt{\sum_{i,j}(d_{i,j}-\mu_d)^2}}$$

$$\text{with } \mu_D = \frac{1}{n}\sum_{i,j}D_{i,j} \text{ and } \mu_d = \frac{1}{n}\sum_{i,j}d_{i,j}.$$

Up-weights and down-weights in the tree

Antigenic branch lengths are realized as two independent weights, allowing for a detailed analysis of the underlying structure of the antigenic data. Up-weights represent the antigenic distance from isolates below this branch to every other isolate outside of this subtree, whereas down-weights represent distances from isolates outside of the subtree to the isolates below this branch. Thus, the branch weight types reveal different properties of the subtree. Let e be the branch going upwards from the least common ancestor of an antigenically homogenous group of viruses (a type) in the tree. The up-weight of e defines the degree to which the antigenic type is separated from other antigenic types according to antisera in other parts of the tree, i.e. how well antigens of this type are neutralized by antisera raised against other types. The down-weight of e defines the degree to which the antigenic type is separated from other types based on antisera *within* this part of the tree, i.e. how well other antigenic types are neutralized by antisera of this type. The antigenic weights of two types often differ, which is not surprising, as antigenic distances are not symmetric. For tip branches, the two weights define the different behavior of the antiserum and antigen of a viral strain. The up-weight reflects the antigenic properties of the isolate, whereas the down-weight reflects the antigenic weight of the antiserum raised against the viral isolate.

In case no antiserum is present in a subtree, down-weights are undefined and assignment of up-weights becomes ambiguous as they form linear combinations. To resolve this, optimization is done only on the up-weights leading to leaves in the according subtree. Afterwards, up-weights of the internal branches are set to the minimum of the up-weights on the branches leading to the respective child nodes (these up-weights are accordingly reduced by the minimum) in a bottom-up traversal. The rationale behind this is that if no additional information is present antigenic weights should rather be a common feature of a subgroup of taxa rather than single isolate variation for every taxon in the subgroup.

Phylogenetic inference

Hemagglutinin (HA) sequences from 258 seasonal human influenza A (H3N2) virus isolates from 1968 to 2003 and that were used by Smith *et al.* (2004) [5] were downloaded from the Influenza Virus Resource (IVR) [46] (**Table S1**). Alignments of DNA and protein sequences, restricted to positions 1 to 363 (sites without missing data that appeared in more than 80% of the sequences), were created with Muscle [47] and manually curated. Trees were inferred with PhyML v3.0 [48] 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 [49]. Isolate A/duck/33/1980 was used as outgroup to root the tree and subsequently removed from the further analysis.

For placement of amino acid changes on the tree branches, protein sequences for the HA1 domain of HA (excluding the additional sites used for a higher resolution of the tree during the tree inference step) were assigned to the leaves of the tree inferred from nucleotide sequences. Ancestral character states were reconstructed under the maximum likelihood criterion using PAML v4.5 [50] under the JTT+Γ_4+F model [51], with the frequency of each amino acid and the Gamma distribution of among-site rate variation, with four rate categories (Γ_4), estimated from the data. Based on the reconstructed ancestral sequences for the internal nodes and leaf node sequences, amino acid changes were assigned to the individual tree branches.

Antigenic data

HI assay data from Smith *et al.* (2004) was used and normalized according to these researchers' methods [5]. For each antigen i, antiserum j and the corresponding HI titer $h_{i,j}$, the distance was set as $d_{i,j} = log_2(max(h_j)/h_{i,j})$, where $max(h_j)$ is the maximum entry for antiserum j. The dataset comprises 4,215 measured values between 273 antigens and 79 reference sera. As not all strains were available in the IVR, 18 antigens and 9 reference sera could not be mapped to a genetic sequence and were excluded from the analysis. Additionally, threshold values (e.g. <10, indicating the lower bound in the HI assay below which dilutions are not measured) were excluded from the analysis, as these values define only long-distance relationships and we did not want to introduce a potential bias by setting these entries to fixed values. In case of multiple antisera raised to the same viral strain, median values of the distances were used.

Definition of antigenic types

Antigenic types in the antigenic tree can be distinguished by selecting type-defining branches according to a threshold distance. The threshold was set to 1.0 antigenic units for average weights (average of up- and down-weights), such that all branches are selected whose average weights are at least twice as high as the average weights of all internal branches. To exclude undersampled groups, all branches leading to subtrees with three or less isolates were excluded.

Supporting Information

Figure S1 Antigenic tree with branch lengths representing antigenic distances (maximum of up- and down weights for each branch) inferred from a maximum likelihood tree of 258 hemagglutinin sequences of seasonal influenza A (H3N2) virus isolates and serological data. Isolates are color-coded by antigenic clusters according to Smith *et al.* (2004). Three isolates (A/Christchurch/4/85, A/Hong Kong/34/90 and A/Netherlands/172/96) are only present as antiserum and were not assigned a cluster label. Changes on terminal branches are colored in black, whereas changes on internal branches are colored in blue.
(PDF)

Figure S2 Antigenic tree with branch lengths representing antigenic distances (maximum of up- and down weights for each branch) inferred from a maximum likelihood tree of 258 hemagglutinin sequences of seasonal influenza A (H3N2) virus isolates and serological data. Isolates are color-coded by antigenic clusters according to Smith *et al.* (2004). Three isolates (A/Christchurch/4/85, A/Hong Kong/34/90 and A/Netherlands/172/96) are only present as antiserum and were not assigned a cluster label. Branch labels depict assigned weights (up/down).
(PDF)

Figure S3 Antigenic tree for influenza A (H3N2) viruses. Branch lengths represent antigenic distances (maximum of up- and down-weights for each branch) inferred from a maximum likelihood tree of 258 hemagglutinin sequences of seasonal influenza A (H3N2) virus isolates and serological data. Colored edges show antigenic type transitions, with internal branches with high average antigenic weights ($\geq$1.0 antigenic units, coloring according to **Figure 1A**) or moderate antigenic weights $\geq$0.5 antigenic units (coloring as gradient from the higher order antigenic type). Subscript $_2$ indicates that a branch was a direct successor of the according type-defining branch (except of branch (i), who is a predecessor of the according type-defining branch). Subscript $_{sub}$ indicates a subdivision of an antigenic type without a direct matching of a reference strain.
(TIF)

Table S1 GenBank accession numbers of the used hemagglutinin sequences.
(DOC)

Table S2 Summary of changes in the phylogenetic tree. Branch amino acid changes refer to the set of changes mapped to a specific branch. For some branches, the down-weight was not defined, as no antiserum was in the respective subtree.
(DOC)

Table S3 Positions with multiple changes in the phylogenetic tree and high antigenic weights (mean and median $\geq$1 antigenic unit, highlighted in bold). 'Tip' indicates leaf branches.
(DOC)

Table S4 Changes with multiple occurrences in the phylogenetic tree and high antigenic weights (mean and median $\geq$1 antigenic unit). 'Tip' indicates leaf branches. Down-weights are omitted, as all changes were identified using up-weights.
(DOC)

Table S5 Type-defining branches selected by different thresholds for average branch weights. Branches (1)–(9) were selected as type-defining branches at a threshold distance of 1.0 antigenic units. Branches (i)–(xii) reveal further subdivision of antigenic types at a threshold distance of 0.5 antigenic units. Asterisks mark branches whose sibling branch leads to a single isolate. Subscript $_2$ indicates that a branch is a direct successor of a type-defining branch (except for branch (i), which is a predecessor of the type-defining branch). Subscript $_{sub}$ indicates a subdivision of an antigenic type without a directly known reference strain.
(DOC)

Text S1 Influence of threshold distance on type-defining branches.
(DOC)

Acknowledgments

We thank Derek Smith for providing antigenic data.

Author Contributions

Conceived and designed the experiments: ACM. Performed the experiments: LS. Analyzed the data: LS ACM. Wrote the paper: LS ACM.

References

1. WHO (2009) Influenza (seasonal). Fact sheet n°211. 3 p. Available from http://www.who.int/mediacentre/factsheets/fs211/en/.
2. Tognotti E (2009) Influenza pandemics: A historical retrospect. J Infect Dev Ctries 3: 331–334.
3. Taubenberger J, Morens D (2006) 1918 influenza: The mother of all pandemics. Emerg Infect Dis 17: 69–79.
4. WHO (2010) Pandemic (H1N1) 2009 - update 112. WHO Global Alert and Response 6 August 2010. 4 p. Available from http://www.who.int/csr/don/2010_08_06/en/.
5. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, et al. (2004) Mapping the antigenic and genetic evolution of influenza virus. Science 305: 371–376.
6. Nelson MI, Simonsen L, Viboud C, Miller MA, Holmes EC, et al. (2007) Phylogenetic analysis reveals the global migration of seasonal influenza a viruses. PLoS Pathog 3: e131.
7. WHO (2011) Recommended composition of influenza virus vaccines for use in the 2011–2012 northern hemisphere influenza season. WHO Wkly Epidemiol Rec 86: 81–91.
8. Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, et al. (2008) The global circulation of seasonal influenza a (h3n2) viruses. Science 320: 340–346.
9. Bedford T, Cobey S, Beerli P, Pascual M (2010) Global migration dynamics underlie evolution and persistence of human influenza a (h3n2). PLoS Pathog 6: e1000918.
10. Bahl J, Nelson MI, Chan KH, Chen R, Vijaykrishna D, et al. (2011) Temporally structured metapopulation dynamics and persistence of influenza a h3n2 virus in humans. Proc Natl Acad Sci U S A 108: 19359–19364.
11. Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, et al. (2008) Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. Vaccine 26: 31–34.
12. Hirst GK (1943) Studies of antigenic differences among strains of influenza a by means of red cell agglutination. J Exp Med 78: 407–423.
13. Lapedes A, Farber R (2001) The geometry of shape space: Application to influenza. J Theor Biol 212: 57–69.
14. Lee MS, Chen MC, Liao YC, Hsiung CA (2007) Identifying potential immunodominant positions and predicting antigenic variants of influenza a/h3n2 viruses. Vaccine 25: 8133–8139.
15. Liao YC, Lee MS, Ko CY, Hsiung CA (2008) Bioinformatics models for predicting antigenic variants of influenza a/h3n2 virus. Bioinformatics 24: 505–512.
16. Huang JW, King CC, Yang JM (2009) Co-evolution positions and rules for antigenic variants of human influenza a/h3n2 viruses. BMC Bioinformatics 10: S41.
17. Bush R (1999) Positive selection on the h3 hemagglutinin gene of human influenza virus a. Mol Biol Evol 16: 1457–1465.
18. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) Predicting the evolution of human influenza a. Science 286: 1921–1925.
19. Plotkin JB, Dushoff J, Levin SA (2002) Hemagglutinin sequence clusters and the antigenic evolution of influenza a virus. Proc Natl Acad Sci U S A 99: 6263–6268.
20. Shih ACC, Hsiao TC, Ho MS, Li WH (2007) Simultaneous amino acid substitutions at antigenic sites drive influenza a hemagglutinin evolution. Proc Natl Acad Sci U S A 104: 6283–6288.
21. Du X, Wang Z, Wu A, Song L, Cao Y, et al. (2008) Networks of genomic co-occurrence capture characteristics of human influenza a (h3n2) evolution. Genome Res 18: 178–187.
22. Pond K, Sergei L, Poon AFY, Brown L, Andrew J, et al. (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.
23. Xia Z, Jin G, Zhu J, Zhou R (2009) Using a mutual information-based site transition network to map the genetic evolution of influenza a/h3n2 virus. Bioinformatics 25: 2309–2317.
24. Steinbrück L, McHardy AC (2011) Allele dynamics plots for the study of evolutionary dynamics in viral populations. Nucleic Acids Res 39: e4.
25. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. New York: Springer. 763 p.
26. Cai Z, Zhang T, Wan X-F (2010) A computational framework for influenza antigenic cartography. PLoS Comp Biol 6: e1000949.
27. Ndifon W (2011) New methods for analyzing serological data with applications to influenza surveillance. Influenza Other Respi Viruses 5: 206–212.
28. WHO (2011) Manual for the laboratory diagnosis and virological surveillance of influenza. Geneva: WHO press. 151 p.
29. Nelson MI, Simonsen L, Viboud C, Miller MA, Taylor J, et al. (2006) Stochastic processes are key determinants of short-term evolution in influenza a virus. PLoS Pathog 2: e125.
30. WHO (1972) Antigenic variation in influenza a viruses. WHO Wkly Epidemiol Rec 47: 381–384.
31. WHO (2003) Recommended composition of influenza virus vaccines for use in the 2003–2004 influenza season. WHO Wkly Epidemiol Rec 78: 58–62.
32. Wiley D, Wilson I, 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.
33. Wilson IA, Skehel JJ, Wiley DC (1981) Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 [angst] resolution. Nature 289: 366–373.
34. Jin H, Zhou H, Liu H, Chan W, Adhikary L, et al. (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.
35. Matrosovich M, Tuzikov A, Bovin N, Gambaryan A, Klimov A, et al. (2000) Early alterations of the receptor-binding properties of h1, h2, and h3 avian influenza virus hemagglutinins after their introduction into mammals. J Virol 74: 8502–8512.
36. Kawaoka Y (2006) Influenza virology: Current topics. London: Caister Academic Press. pp 95–137.
37. Bateman AC, Busch MG, Karasin AI, Bovin N, Olsen CW (2008) Amino acid 226 in the hemagglutinin of h4n6 influenza virus determines binding affinity for {alpha}2,6-linked sialic acid and infectivity levels in primary swine and human respiratory epithelial cells. J Virol 82: 8204–8209.
38. Wan H, Sorrell EM, Song H, Hossain MJ, Ramirez-Nieto G, et al. (2008) Replication and transmission of h9n2 influenza viruses in ferrets: Evaluation of pandemic potential. PLoS ONE 3: e2923.
39. Hensley SE, Das SR, Bailey AL, Schmidt LM, Hickman HD, et al. (2009) Hemagglutinin receptor binding avidity drives influenza a virus antigenic drift. Science 326: 734–736.
40. Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141: 1641–1650.
41. Fitch WM (1971) Toward defining the course of evolution: Minimum change for a specific tree topology. Syst Zool 20: 406–416.
42. Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53: 673–684.
43. McHardy AC, Adams B (2009) The role of genomics in tracking the evolution of influenza a virus. PLoS Pathog 5: e1000566.
44. Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet 19: 233–257.
45. Lawson CL, Hanson RJ (1995) Solving least squares problems. Philadelphia: Society for Industrial and Applied Mathematics. 350 p.
46. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, et al. (2008) The influenza virus resource at the national center for biotechnology information. J Virol 82: 596–601.
47. Edgar RC (2004) Muscle: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
48. Guindon S, Gascuel O (2003) A simple, fast, and accurate method to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
49. Zwickl D (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [PhD dissertation]. Austin (Texas): Section of Integrative Biology, The University of Texas at Austin. 125 p. [http://www.zo.utexas.edu/faculty/antisense/zwicklDissertation.pdf].
50. Yang Z (2007) Paml 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.
51. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275–282.

Open

The ISME Journal (2011), 1–11

www.nature.com/ismej

ORIGINAL ARTICLE

Defining seasonal marine microbial community dynamics

Jack A Gilbert[1,2,3], Joshua A Steele[4], J Gregory Caporaso[5], Lars Steinbrück[6], Jens Reeder[5], Ben Temperton[1], Susan Huse[7], Alice C McHardy[6,8], Rob Knight[5,9], Ian Joint[1], Paul Somerfield[1], Jed A Fuhrman[4] and Dawn Field[10]

[1]*Plymouth Marine Laboratory, Prospect Place, Plymouth, UK;* [2]*Institute of Genomics and Systems Biology, Argonne National Laboratory, Argonne, IL, USA;* [3]*Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA;* [4]*University of Southern California, Department of Biological Sciences, Los Angeles, CA, USA;* [5]*Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, CO, USA;* [6]*Department of Algorithmic Bioinformatics, Heinrich-Heine University, Düsseldorf, Germany;* [7]*Josephine Bay Paul Centre for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA;* [8]*Max-Planck-Institut fur Informatik, Max-Planck Research Group for Computational Genomics and Epidemiology, Saarbrücken, Germany;* [9]*Howard Hughes Medical Institute, Boulder, CO, USA and* [10]*NERC Centre for Ecology and Hydrology, Wallingford, UK*

Here we describe, the longest microbial time-series analyzed to date using high-resolution 16S rRNA tag pyrosequencing of samples taken monthly over 6 years at a temperate marine coastal site off Plymouth, UK. Data treatment effected the estimation of community richness over a 6-year period, whereby 8794 operational taxonomic units (OTUs) were identified using single-linkage preclustering and 21 130 OTUs were identified by denoising the data. The *Alphaproteobacteria* were the most abundant Class, and the most frequently recorded OTUs were members of the *Rickettsiales* (SAR 11) and *Rhodobacteriales*. This near-surface ocean bacterial community showed strong repeatable seasonal patterns, which were defined by winter peaks in diversity across all years. Environmental variables explained far more variation in seasonally predictable bacteria than did data on protists or metazoan biomass. Change in day length alone explains >65% of the variance in community diversity. The results suggested that seasonal changes in environmental variables are more important than trophic interactions. Interestingly, microbial association network analysis showed that correlations in abundance were stronger within bacterial taxa rather than between bacteria and eukaryotes, or between bacteria and environmental variables.

The ISME Journal advance online publication, 18 August 2011; doi:10.1038/ismej.2011.107
Subject Category: microbial population and community ecology
Keywords: 16S rRNA; microbial; bacteria; community; diversity; model

Introduction

Only recently with the introduction of molecular techniques satisfactory descriptions of natural microbial assemblages have been generated (Fierer and Jackson, 2006; Rusch *et al.*, 2007; Costello *et al.*, 2009; Caporaso *et al.*, 2011). In this paper, we summarize a 6-year time series of 16S rRNA tag pyrosequencing of samples taken from a long-time series station in the English Channel. The aim was to understand seasonal variability and to try to determine which environmental factors might have the greatest influence on the varying diversity. In contrast to terrestrial environments that are essentially static, the marine environment has the added complication that the dispersion and movement of populations will be driven by hydrography. This adds to difficulties of interpretation of results, particularly if the sampling design is Eulurian (a fixed site) rather than Lagrangian (moving with the water flow). The Western English Channel has been studied intensively for more than 100 years (Southward *et al.*, 2005), and this wealth of data provide a robust context with which to explore temporal microbiological complexity. Inferences can be drawn regarding how bacterioplankton assemblages may potentially interact with the environment as well as with specific groups of organisms.

Previous efforts to determine which factors might affect microbial communities have largely focused on the relative importance of temperature and nutrient concentrations (Cullen, 1991; Kirchman *et al.*, 1995; Morris *et al.*, 2005; Fuhrman *et al.*,

Correspondence: JA Gilbert, Institute of Genomics and Systems Biology, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA.
E-mail: gilbertjack@anl.gov
Received 14 March 2011; revised 13 July 2011; accepted 14 July 2011

2006; Fuhrman, 2009; Gilbert *et al.*, 2009). These are obvious candidates because of the strong effect of temperature on biological processes (Nedwell and Rutter, 1994) and the fact that nutrient availability can drive niche structure through resource partitioning (Church, 2009). Of greatest relevance to the present study is the recent demonstration that bacterioplankton diversity followed a latitudinal gradient, with maximum potential richness being primarily driven by temperature, with many other factors modulating an intricate network of richness at any particular temperature (Fuhrman *et al.*, 2008).

The aim of the current study was to further characterize seasonal patterns of bacterioplankton diversity in the Western English Channel, beyond an initial 1-year study by Gilbert *et al.* (2009). Using these data, we tested three competing alternative hypotheses about potential drivers of diversity patterns, namely whether the observed seasonal patterns correlate with (1) varying concentrations of inorganic nutrients, (2) annual water–temperature cycle or (3) the population structure of the eukaryotic phytoplankton and zooplankton. The null hypothesis was that the seasonal patterns in microbial community composition in the Western English Channel showed no relationship with any of the physical or biological factors measured in this study.

Materials and methods

Sampling, DNA extraction, 16S rDNA V6 amplification and pyrosequencing

Seawater samples were collected on 72 instances from January 2003 to December 2008, from the L4 sampling site (50° 15.00′ N, 4° 13.02′) of the Western Channel Observatory (http://www.westernchannelobservatory.org.uk). Sampling, extraction, amplification, and sequencing protocols and environmental parameter analysis were performed simultaneously on the same samples as described previously by Gilbert *et al.* (2009); extensive information can be found in Supplementary Information (Supplementary Tables S1–S3). Bacterial diversity was examined in the context of the broad range of biotic and abiotic variables that are routinely measured at the Observatory. These included phytoplankton and zooplankton species abundance, the concentrations of ammonia, nitrate + nitrite, phosphate, silicate, total organic carbon and nitrogen, salinity, chlorophyll, photosynthetically active radiation, North Atlantic Oscillation data, day length, primary productivity and temperature. Statistical analyses used the routines of PRIMER (Clarke and Warwick, 2001; Clarke and Gorley, 2006).

Sequence data analysis

All sequence data were treated as reported previously (Gilbert *et al.*, 2010), using the same quality control that included random resampling to standardize the sequencing effort as described below, Sequence data noise reduction using Single-Linkage Preclustering (SLP; Huse *et al.*, 2010) and analysis (sample similarity derived from Bray–Curtis indices weighted on taxon abundance matrices) also followed previous protocols. In addition, several noise reduction strategies such as SLP (Huse *et al.*, 2010) and denoiser (Reeder and Knight, 2010) were compared to examine the impact of pyrosequencing errors on community diversity patterns observed in the data (see Supplementary Figure S1a). It is important to stress that both known and unknown biases associated with these techniques meant that these data could not be seen as quantitative, and hence all analyses are based on relative changes derived through comparison. As the same sequencing and sampling effort was applied to each sample, the operational taxonomic unit (OTU) richness (S) was used as a diversity metric, which showed a 97% correlation to two extrapolative estimators of diversity (Chao1 and Ace) over the 72 samples (Supplementary Figure S1b). Changes in community diversity and relationship to environmental parameters were examined using various nonparametric multivariate methods, discriminant function analysis (DFA), and association networks (see Supplementary Information).

To determine whether microbial communities in the Western English Channel demonstrated seasonal patterns over many years, 747 496 16S rDNA V6 sequences were analyzed, including those previously published for the year 2007 (Gilbert *et al.*, 2009). To compensate for potential overestimation in diversity resulting from pyrosequencing and amplification errors, a clustering technique was used. SLP grouped OTUs at 2% sequence identity and an average-linkage clustering followed, based on pair-wise alignments (Huse *et al.*, 2010), which resulted in 8794 OTUs. To remove sequencing effort bias, each sample was randomly resampled to the smallest individual sample sequencing effort (4505) as described before (Gilbert *et al.*, 2009). This resulted in a total of 4204 OTUs (for all 72 samples combined). Approximately, 53% of the OTUs were represented by only a single sequence (singletons). These results, in terms of relative abundance, were confirmed using a second denoising technique, Denoiser (Reeder and Knight, 2010), which generated greater total richness (21 130 OTUs). However, comparison between Denoiser, SLP and no-denoising/filtering indicated that overall, the same patterns of community diversity were evident with each technique (Supplementary Figure S1). SLP constituted by far the most conservative OTU predictions, and was therefore used for subsequent analysis.

Results

Seasonal variations in diversity and persistence

Bacterioplankton were very diverse at this station and a total of 8794 different OTUs (defined using

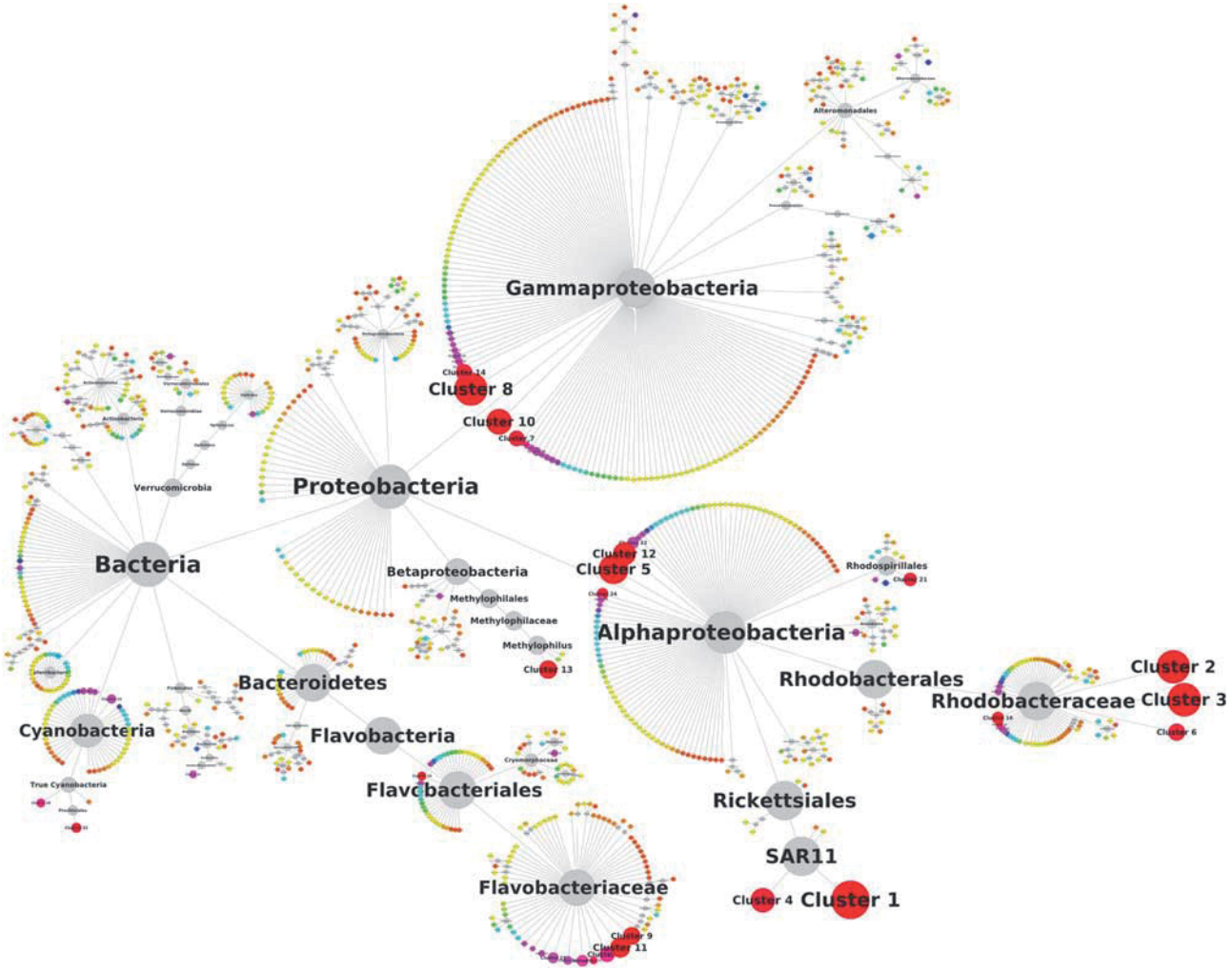


Figure 1 Persistence of OTUs in microbial communities at L4 over a 6-year time period. Median OTU abundance, calculated for all time points, over a 6-year period is set proportional to node size on a logarithmic scale. Only OTUs found in at least 5% of the time-series samples ($\geqslant 4$) are shown. This includes 22.53% of the OTUs, representing 97.48% of the sampled organisms. Node coloring shows the differences in persistence over time, with the color scale from orange (5%), yellow (16%), green (35%), blue (66%), red (100%) reflecting increasing persistence.

SLP) over a 6-year period were identified. Figure 1 summarizes the taxonomic identify of all the OTUs sequenced and also gives an indication of the persistence of OTUs in microbial communities at L4 over a 6-year time period. Although this study has shown high diversity of bacterioplankton in the English Channel, as with other studies of natural assemblages, the majority of sequences could not be identified to species. Indeed, only 6 of the 10 most abundant OTUs could be annotated below the level of Class and, of the top 100 most abundant OTUs, only 2% could be identified to the species level. The taxonomic level to which the OTUs could be identified was—Phylum (9%), Class (32%), Order (10%), Family (26%), Genus (21%). This was true using a number of different annotation strategies (that is, GAST (Sogin *et al.*, 2006); BLAST against Greengenes (DeSantis *et al.*, 2006), SILVA (Pruesse *et al.*, 2007) and RDP (Maidak *et al.*, 2001); RDP classifier (Maidak *et al.*, 2001); data not shown, references in Supplementary Information). These results suggest that a large fraction of as-of-yet uncharacterized lineages were present, even among the most abundant taxa, and highlights the difficulties associated with accurate annotation of short read-length tag sequences from hypervariable 16S rRNA regions (Wang *et al.*, 2007; Liu *et al.*, 2008).

Although there are significant seasonal variations in OTU frequency throughout a 6-year period (Figure 2), there are also strong repeating patterns. As other studies of marine microbial diversity have demonstrated, the *Alphaproteobacteria* were the most abundant Class. The OTUs most frequently recorded were members of the *Rickettsiales* and *Rhodobacteriales*. Other OTUs with high frequency were the *Flavobacteriales* (Class: *Bacteroidetes*) and there were also peaks in the *Gammaproteobacteria* (*Vibrionales* and *Pseudomonadales*).

Alpha diversity of the observed OTUs (S) was relatively constant across the time series, but

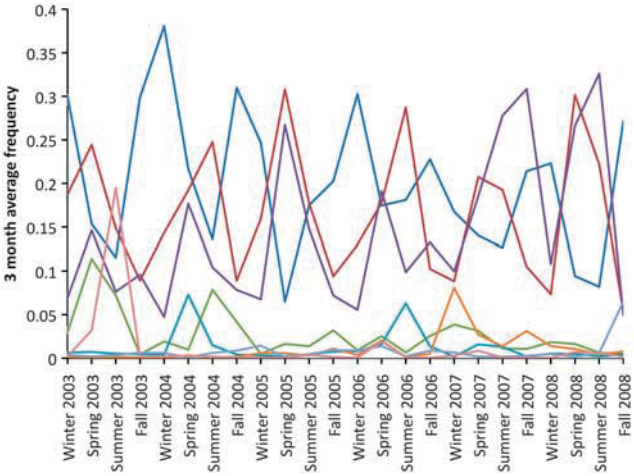


Figure 2 Plot representing the seasonal dynamics (grouped as an average of seasons; Winter: January–March; Spring: April–June; Summer: July–September; Fall: October–December) of taxa grouped at the taxonomic level of Order in the L4 6-year time series. Frequency is recorded based on abundances within a resampled abundance of 4101 sequences per sample. Only Orders whose average frequency peaked above 10% of the resampled community abundance were included.

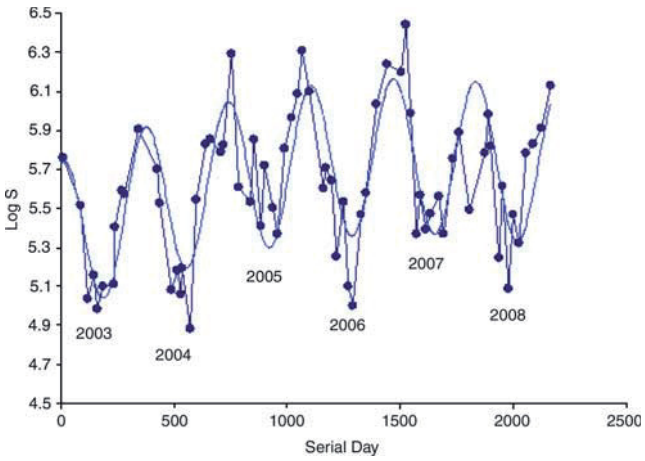


Figure 3 Alpha diversity (observed OTUs) plotted as the log of species richness (S) by month spanning 6 years of marine water sampling at the L4 site in the Western English Channel. A cyclic pattern is observed in alpha-diversity, with species richness peaking in the winter months.

showed distinct cyclical patterns with maxima in winter and minima in summer (Figure 3). The mean S per time point was 286, with an average minimum of 179 in summer and maximum of 437 in winter. This pattern was further confirmed by permutation-based analysis of variance (of S) for all taxa, and for a range of phyla (Supplementary Table S4). S was most similar when comparing the same time of year, and differences between seasons and among years were both highly significant. Seasonal differences tended to be greater than inter-annual (greater pseudo-*F* values although there were fewer d.f.). This lack of significant interaction terms suggested that the seasonal cycle was consistent across years. Overall persistence (Figure 1) was linked to abundance; OTUs that were present at more than three time points accounted for 97.48% of the sequences. In total, only 12 OTUs were found at every 1 of the 72 time-points, yet these were exceptionally abundant, comprising ~35% of all the sequence reads.

Seasonal trends in most abundant bacteria

The two most abundant Orders were *Rickettsiales* and *Rhodobacterales*, and they had different seasonal abundances. The *Rickettsiales* sequences were dominated by the SAR11 clade and tended to peak in winter (Figure 4). At this time, light and primary production were low, and inorganic nutrient concentrations were at their maximum. In contrast, the *Rhodobacterales*, which were dominated by the *Roseobacter* clade, tended to peak in Spring and Autumn, when nutrient concentrations were lower yet primary productivity was higher. This is consistent with what is known from single-strain-level studies; SAR11 are considered to be obligate oligotrophs, while the *Roseobacter* clade contains many genera whose cultured representatives tend to grow in organic nutrient-rich media, and may be likely to respond at times when rates of primary production are higher.

Rare taxa may dominate the assemblage

The largest bacterial 'bloom' occurred during August 2003, and this constituted a single *Vibrio* sp.,

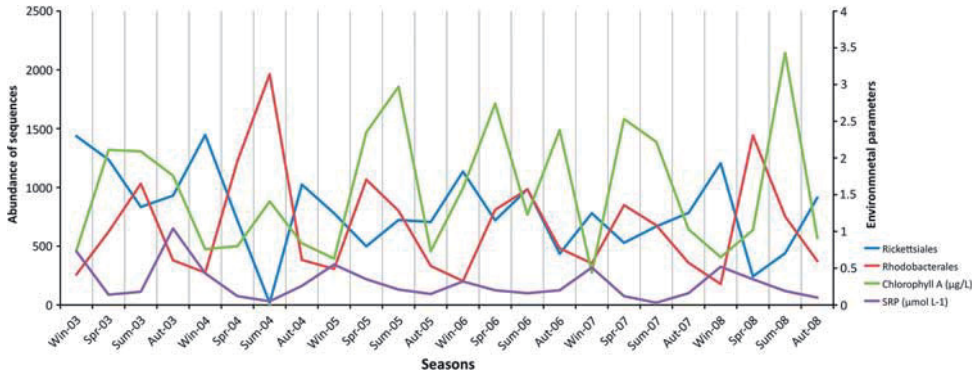


Figure 4 Plot representing the seasonal dynamics of the bacterial Orders, *Rickettsiales* and *Rhodobacterales*, and environmental parameters, chlorophyll a and soluble reactive phosphorus (SRP) in the L4 6-year time series. Frequency is recorded based on abundances (abundance of sequences per taxa) within a resampled abundance of 4505 sequences per sample.

which represented 54% of the sequences. Yet, for the rest of the time series, this taxon was relatively rare, having an abundance of 0–2%. Interestingly, this peak was correlated with an increase in the relative abundance of the diatom, *Chaetoceros compressus*. This diatom was also typically present at low abundance, between 0.002–0.2% of total phytoplankton biomass (Supplementary Table S2). However, in August 2003, *C. compressus* accounted for 1.2% of total eukaryotic phytoplankton. Our data do not distinguish between a causal relationship—a specific dependence of a bacterial species on a specific phytoplankton species—and simple co-occurrence, which might be a response to unusual environmental conditions. Certainly at this time point, the highest total organic nitrogen and carbon concentrations, and second highest chlorophyll a concentration were measured in the whole time series between 2003 and 2008 (Supplementary Table S1).

Seasonal succession in the community composition is robust

The dataset of environmental and biological variables was examined to investigate potential relationships between bacterioplankton and the environmental and eukaryotic abundance data. The community composition (rather than richness) was used, after determining whether seasonal patterns in community composition were as robust as those for species richness. Three different subsets of the bacterial OTUs, that is, the most abundant, most common and most variable (see Supplementary Materials) were defined. These definitions were robust across the different denoising strategies (that is, the same OTUs (based on sequence identity, with the same taxonomic inference defined). Using DFA, an eigenvector technique that, in this case, searches out the taxa which are best able to predict the month (Fuhrman *et al.*, 2006), we found that for each subset, the bacterial community could correctly predict the month with 100% accuracy, showed a clear repeating pattern (Figure 5), and was able to explain >60% of the variance in the community structure (Supplementary Table S5).

These patterns for most abundant, common and variable subsets are similar to those reported for similar subsets in a Californian near-surface bacterioplankton time series (Fuhrman *et al.*, 2006), suggesting that seasonal succession patterns of marine surface water bacterial communities in temperate regions may be conserved across different biomes. The Californian study was based on automated ribosomal RNA intergenic spacer analysis fingerprint technology, but the sequence-based annotation provided by this study allowed considerably better predictions for the bacterial taxa contributing most strongly to these signals. In this instance, these were members of the *Alphaproteobacteria* (for example, SAR11 and *Rhodobacteriaciae* groups), the *Gammaproteobacteria* (for example, *Pseudomonas, Pseudoalteromonas*, and *Vibrio* groups), the *Cyanobacteria*, and the *Bacteroidetes* (for example, *Flavobacteriaceae* group; Supplementary Table S6).

Seasonal variance in community composition

The relative significance of environmental versus biological factors in describing the seasonal variation in bacterioplankton assemblages was investigated using DFA. DFA, via multiple regression using environmental factors and eukaryotic counts, was used to predict the first discriminant function (DF1) from each subset of the community (that is, most abundant, most common and most variable). Environmental parameters explained 49–91% of the

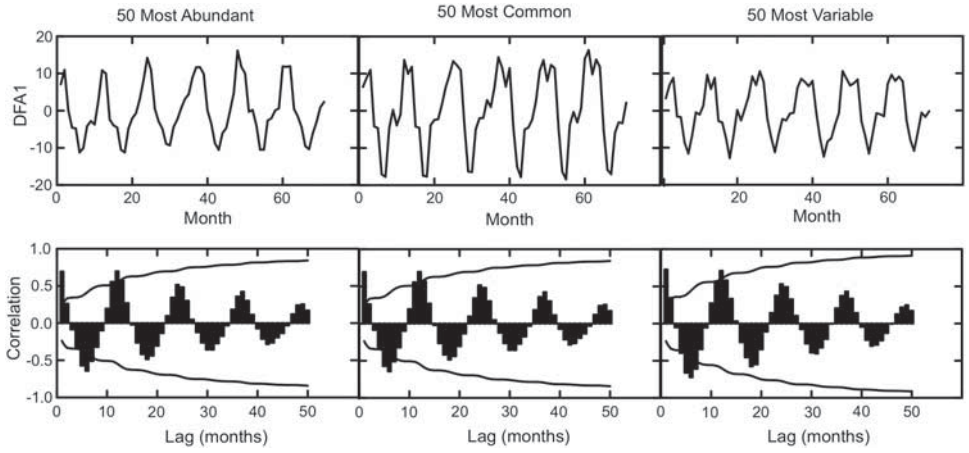


Figure 5 Annual repeating patterns from the bacterioplankton community sampled monthly from 2003–2008 in the English Channel determined by DFA where the model used the bacterioplankton community to predict the month. Upper row of graphs shows the time-series analysis of the first discriminant function (DFA1) over 72 months. The lower row shows the autocorrelation of the discriminant function with up to a 50-month lag. The lines in the lower row represent correlations with $P<0.05$.

variance in DF1, while eukaryotic variables explained 18–51% of the variance (Supplementary Table S6). This suggests that that the seasonally responsive members of the microbial community were responding to changing environmental factors, while interactions between the bacteria and the eukaryotes may have had a less comprehensive influence. Obviously, as shown for the *Vibrio* bloom in 2003, this trend is not absolutely uniform, and blooms of rare taxa can be influenced by the presence of eukaryotes. However, as defined by the robust annual cyclicity, the community recovers from these 'rare-bloom' events, suggesting an overall bottom-up influence on the community composition and structure. Essentially this suggests that nutrient concentrations, physical parameters and biology all demonstrate significant influence in an extraordinarily complex matrix.

Annual day length cycle explains most of the variability in the seasonal pattern of species diversity
To test whether changes in nutrients or temperature provided the best correlation with changes in community diversity, distance-based linear modelling was used (described in detail in Supplementary Material). This showed that, although a significant fit could be ascribed to a combination of temperature and photosynthetically active radiation and the richness of all OTUs, the most significant fit was always to the annual change in day length (Supplementary Table S7). This was best modelled by a cosine term (DX1) with the peak centered on December 22. When day length (DX1) was combined with serial day (D), it described 66.3% of the variance in OTU richness. However, when examining the phototrophic *Cyanobacteria* (Supplementary Table S7), the relationship of richness to day length was not always evident, for example, diversity peaked in spring but not in winter, and hence coincided with the lowest annual temperatures at L4. To account for the *Cyanobacteria* and to significantly improve the fit of our model ($\delta AIC > -2$), a second seasonal artificial term centered on March 22 (a sine-derived term—DX2) was added that closely tracked temperature. Also, because most of the taxa show subtle changes in their seasonal cyclicity over these years, it was possible to significantly improve the model further by adding a linear time trend term (D). However, this did not improve the fit for the cyanobacterial community diversity, which was remarkably stable over the 6 years. Strikingly, the *Cyanobacteria* were unique in that a combination of photosynthetically active radiation, temperature and nitrate/nitrite concentration provided as good a fit as the artificial descriptors (DX1, DX2 and D; Supplementary Table S7). Not unexpectedly, this suggests that, unlike other groups, the species diversity of these primary producers can be well defined by a combination of light availability, nitrogen availability and temperature, reflecting a different set of niches compared with the other potentially heterotrophic bacterioplankton.

Discussion

The repeating cycles in bacterioplankton diversity in this Eulerian study raise the question of whether unique water masses pass through the English Channel, and whether those water masses contain characteristic bacterioplankton assemblages. This is almost certainly not the case as the hydrography of the Western English Channel has been studied extensively (Southward *et al.*, 2005). From the earliest studies in the 1930s using drift-bottles, it was known that there was a strong flow through the English Channel from west to east. Later modelling and observational studies showed the importance of wind over a very wide shelf region (including the North Sea) in determining flow through the Western English Channel (Pingree and Griffiths, 1980). Southerly winds resulted in the greatest net transport of water along the English Channel through the Straits of Dover and into the southern North Sea; westerly winds were less effective.

It has recently been calculated that average residence time at the sampling site is on the order of 2 weeks (Lewis and Allen, 2009), although dispersion occurs continuously. The repeating annual patterns of bacterioplankton demonstrated in this study cannot be due to the repeated intrusion of water mass with an annual periodicity. We do not know how representative these robust annual patterns are of the entire English Channel. It may be that the observed patterns represent seasonal changes in bacterioplankton on the Celtic Sea Shelf, which is advected into the Western English Channel. Given that this advection will largely depend on wind conditions, it seems unlikely that such similar patterns would occur over a 6-year period. Clearly, further sampling on the European Shelf will be required to answer the question of the representativeness of this station.

The relationship between OTU richness and day length is interesting. To the best of our knowledge, this is the only example from a marine dataset where a single variable has such explanatory capacity (66.3% of the variance in OTU richness). There are examples from terrestrial systems; for example, tRFLP analysis identified an r^2 value of 0.7 between bacterial community richness and pH (Fierer and Jackson, 2006). Temperature would imply a clear mechanism; we can see no such direct mechanism that result in day length directly controlling bacterioplankton assemblages.

Other environmental factors that could suggest direct mechanisms did have significant relationships. They did not, however, apply to the most common and abundant taxa, but the composition of the most variable taxa could be significantly

predicted by nutrient concentrations (NH_4^+, total organic nitrogen (TON), soluble reactive phosphate, primary production and broad shifts in ocean currents indicated by the North Atlantic Oscillation (Supplementary Table S6). Overall, we conclude that the monthly pattern and response to broad seasonal changes indicate that the most common and most abundant bacterial OTUs have temporally defined niches. In contrast, the most variable OTUs have niches that can be defined temporally as well as by nutrient pulses and changes in currents. Temporal niche structure suggests taxa with a resilient seasonal pattern, for example, SAR11 and *Rhodobacteriaciae*, although tracking nutrient pulses and currents, are potentially less resilient to changing environmental conditions. However, the relationship is complex, and potentially a function of abundance, commonality and variability, as both SAR11 and *Rhodobacteriaciae* are in the most abundant, most common and most variable subset.

Interestingly, interactions were strongest within the bacterial and eukaryotic domains rather than between them, and relationships were stronger between bacterial taxa than with environmental variables. Association network analysis was employed in an attempt to deconvolute the complex network of relationships that were driving the observed DFA results. However, this revealed that the strongest correlations exist between bacterial OTUs (whether abundant, common or variable) and, to a slightly lesser extent among eukaryotes, compared with correlations between these two domains or between either bacteria or eukaryotes and environmental factors (Figure 6). Also, the integrity of these relationships was maintained across the three chosen subsets of OTUs (Figure 6). Even among the highly variable OTUs, which might be expected to respond to changing conditions enabling growth from rare to abundant, most significant correlations were still between bacteria (Supplementary Figure S2). Also, at a highly correlated ($r>0.7$, $P<0.001$, $q<0.0012$) level, there were many eukaryotic taxa in a loosely intercorrelated group (Supplementary Figure S2a), but there are still very few specific connections between the eukaryotes and the bacteria. Mostly the bacteria were correlated to one another and to the environmental factors, and the eukaryotes were also connected to one another and the environmental factors. The highly intercorrelated group (Supplementary Figure S2b) was almost completely devoid of eukaryotes, but was connected to an herbivorous, parasitic copepod (*Poescilostomatoida*), and to the seasonal factor DX1, NO_2+NO_3, and an interconnected cluster of *Gammaproteobacteria*, *Bacillus* and *Actinobacteria* OTUs.

Interactions between eukaryotes and bacteria became more apparent when moderate correlations were examined between different subsets of the eukaryotic community and the 300 most abundant bacterial OTUs. Mixotrophic eukaryotes (potential grazers on bacteria) and autotrophic eukaryotes both showed complex interactions with the prokaryotic community (Supplementary Figure S3). Although flagellates (when grouped by size) were correlated to each other ($r=0.59$, $P<0.001$, $q<0.0012$) and, naturally, to the total number of flagellates, only two bacterial OTUs (a single *Rhodobactereaceae* OTU and a single *Cyanobacteria* OTU) are correlated to all three groups (Supplementary Figure S3a). There were many bacterial and eukaryotic OTUs, which correlate to two of the flagellate subgroups, and a smaller number, which correlate to only one of the flagellate subgroups. The 5 μm flagellates were negatively correlated to a *Betaproteobacterial* and a *Gammaproteobacterial* OTU, and the diatom *Paralia*

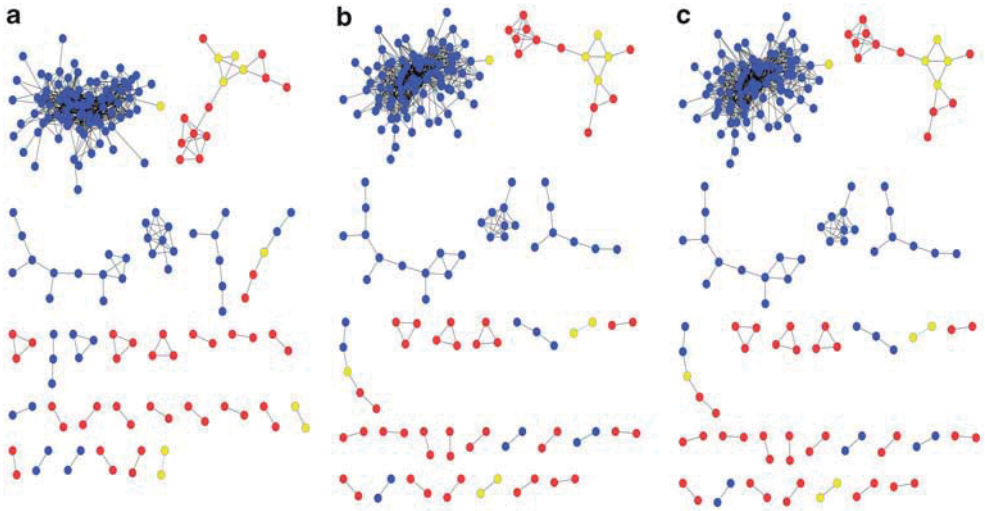


Figure 6 Broad view of correlation network for the microbial community and the environment at station L4. The network shows strong correlations ($r>0.8$, $P<0.001$, $q<0.002$) between microbial and environmental parameters for the 300 most abundant bacterial taxa (**a**), the 300 most common bacterial taxa (**b**), and the 300 most variable bacterial taxa (**c**). Bacteria are shown in blue, eukaryotes are shown in red and environmental variables are shown in yellow.

sulcata in samples with a 1-month lag, which reflects an increase in those abundant members of the community following a decrease in 5 μm-sized flagellates (Supplementary Figure S3a).

A similar situation was applied to correlations between autotrophic eukaryotes and abundant bacterial OTUs. The diatom, *P. sulcata*, correlated negatively to the total diatom counts with a 1-month time lag (Supplementary Figure S3b). This may indicate a situation where *P. sulcata* dominated the diatom community, while the total number of diatoms decreased. These two eukaryotic nodes shared 26 bacterial OTUs that correlated positively to *P. sulcata* and negatively with a 1-month time lag to the total diatom count (Supplementary Figure S3b). These bacterial OTUs may reflect a community shift indicated by the increase of *P. sulcata* and the 26 *Proteobacteria*, *Bacteroidetes* and *Verrucomicrobia* when the total number of diatoms decreased. The winter peak seasonal cycle, DX1, also positively correlated to *P. sulcata* and negatively correlated, with a 1-month lag, to total diatoms in the same way, possibly implying seasonal community succession. There were positive contemporaneous correlations between *P. sulcata* and $NO_3 + NO_2$, between silicate and mixed layer depth, and a negative 1-month lagged correlation between the North Atlantic Oscillation and total diatom counts; these results indicate that nutrient concentrations may be drivers of this succession (Supplementary Figure S3b). Interestingly, there were only positive correlations between bacterial OTUs and 2 μm flagellates (Supplementary Figure S3a), even though 2 μm flagellates might be expected to be the major grazers of bacterioplankton. Bacterial OTUs were also positively correlated to total flagellates, total phytoplankton, coccolithophores and *Emiliania huxleyi* (Supplementary Figure S3b).

Many environmental factors were highly correlated ($r > 0.7$, $P < 0.001$, $q < 0.0012$) with both eukaryotic OTUs and bacterial OTUs, when both the 300 most variable bacteria (Supplementary Figure S4a) and the 300 most common bacteria (Supplementary Figure S4b) were considered. Strikingly, the seasonal index peaking in winter (DX1) was correlated almost exclusively to bacterial OTUs, including *Proteobacteria* (for example, *Alphaproteobacteria*, *Gammaproteobacteria*, *Nitrospira*), unidentified bacteria, *Deferribacteres* and *Owenweeksia* in both the common and variable sub-networks (Supplementary Figure S4). *Cladocera* and *Echinodermata* were the only eukaryotes that connected to DX1 and they were negatively correlated with no lag and a 1-month lag, respectively. This suggests that seasonal factors (for example, day length, which is a proxy for DX1) may be more important for the bacterioplankton than for the eukaryotic community. The spring seasonal factor, DX2 was correlated with a 3-month lag to *Cladocera* (indicating a summer increase in abundance), and was negatively correlated to a *Bacteroidetes* OTU in the most variable subset (Supplementary Figure S4a).

Positive correlations were widespread in the microbe–environment network. Primary production (monthly average) was correlated to total diatoms, total ciliates, total microzooplankton and a *Rhodobacteriaceae* OTU (which also correlated to daily primary production and temperature). Daily primary production (ML primary production, calculated from observed chlorophyll values and integrated over the observed mixed layer depth) was also positively correlated to total diatoms, total phytoplankton, total ciliates and echinodermata (Supplementary Figure S4). This suggests that, as productivity and nutrients increased, these bacteria and eukaryotes also increased in abundance, that is, these taxa appear to perform best in a productive system. There was little correlation-based evidence for top-down effects in this system, although this may be a function of a lack of resolution of bacterivores among the eukaryotes or perhaps a limitation of this kind of analysis.

Local similarity analysis, with its ability to see time-lagged correlations, also provided insight into the relationships between environmental factors themselves. Although day length was not correlated to temperature at the 0.7 level, the Winter seasonal cycle (DX1) was negatively correlated to day length with no time delay, and to temperature and primary production with a 1-month time delay (Supplementary Figure S4); that is, day length changed seasonally, followed by a change in temperature. DX1 and day length (which was positively correlated to photosynthetically active radiation and primary production) may be serving as combinatory signals of seasonal environmental change, involving factors such as changes in input of energy into the system. These combinatory variables may more closely map the changes in the whole community of bacterioplankton as well as the individual bacterial OTUs connected to them. $NO_2 + NO_3$ were highly correlated with soluble reactive phosphate and silicate (Supplementary Figure S4). However, soluble reactive phosphate was correlated only to a *Gammaproteobacteria* OTU and a *Rhodobacteriaciae* OTU, while silicate was not highly correlated to any bacteria or eukaryotes. $NO_2 + NO_3$ was positively correlated to 12 bacterial OTUs, which were also positively correlated to DX1, and there were 10 bacterial taxa that were positively correlated solely to $NO_2 + NO_3$. The close coupling between these taxa and $NO_2 + NO_3$ (that is, these taxa were only abundant when there was an increased availability of nitrogen) suggests that these taxa may be seasonally nitrogen limited in this ecosystem.

Regardless of the subset of OTUs (for example, abundant, variable, or common) analysed, each subset was able to predict the month. In addition, each of the networks appeared to identify many of the same connections when we examined 300 taxa from any of the subsets (Figure 6, Supplementary Figures S2–S4). OTUs were ranked differently within each subset, but they produced similar patterns,

which were nearly identical at the $r>0.8$ correlation level (Figure 6). This was partly due to the stability of the bacterioplankton community at L4 and the depth of sampling into this community. It was also an effect of using statistical analyses that require a certain number of occurrences in order to detect a pattern; by design, these analyses would ignore the once-a-decade occurrence, for example, the spike in *Vibrio* spp. abundance in the summer of 2003. However, comparing these subsets allowed for a better sense of the ecology behind these bacterial OTUs. This is demonstrated most clearly when restricting the correlations to the 50 most common and most variable bacterial taxa, and their relationship to environmental factors (Figure 7). For instance, a SAR11 (Al*phaproteobacteria*_03_2), although common, changed abundance seasonally (it was the 6th most variable bacterial OTU) and increased in abundance when inorganic nutrient concentrations increased. A *Rhizobiales* member (*Alphaproteobacteria*_03_121) that correlated with NO_2+NO_3 (Figure 7a) was not as variable (Figure 7b), whereas the *Deferribacteres* member (*Deferribacteres*_03_12) that correlated with NO_2+NO_3 (Figure 7b) was not common (Figure 7a), but increased in abundance along with increased NO_2+NO_3 concentration. Among these observations of common influence, there were also hints

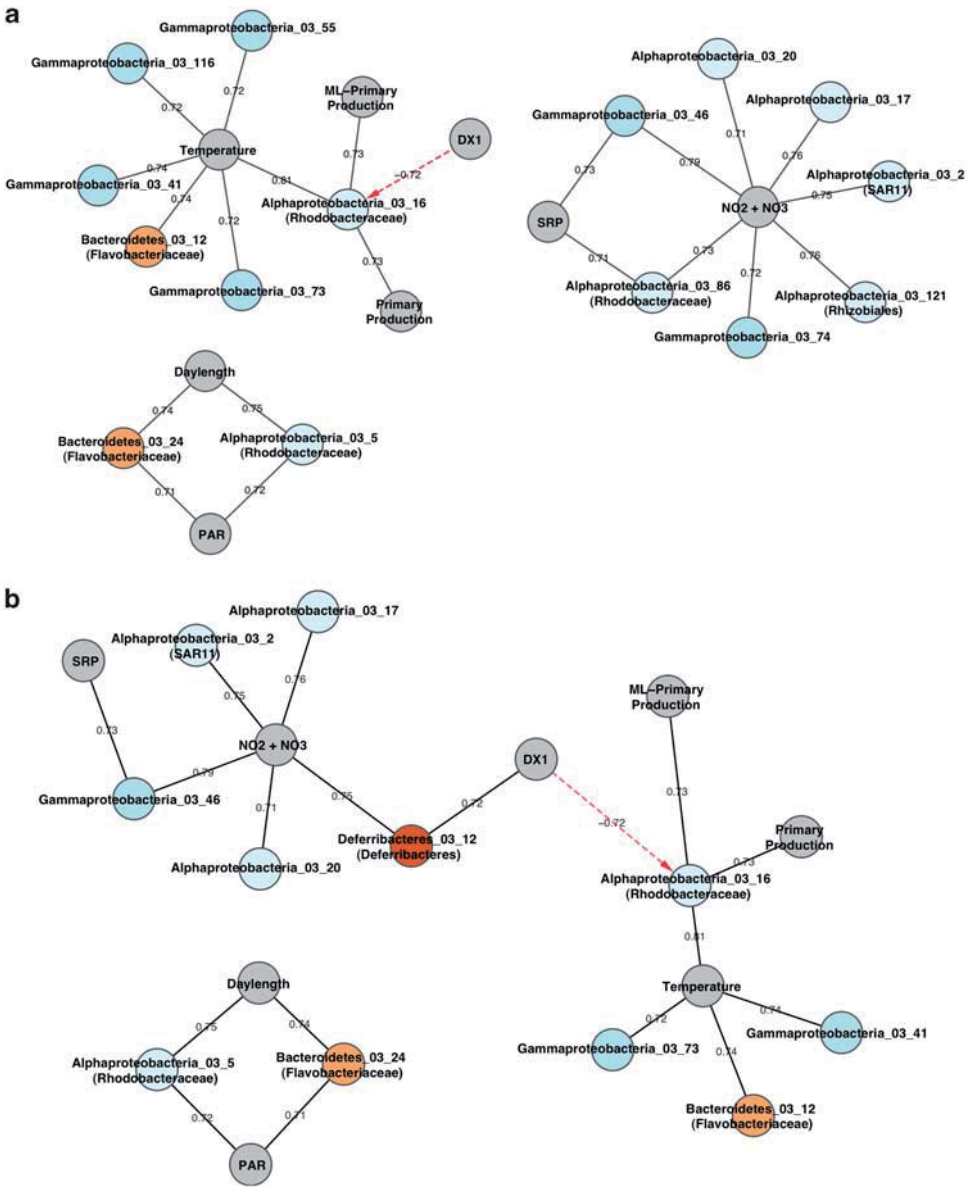


Figure 7 Sub-networks of highly correlated ($r>0.7$, $P<0.001$) variables built around environmental factors from the 50 most common (**a**) and 50 most variable (**b**) bacterial OTUs. Interactions between environmental variables and eukaryotic interactions with environmental variables have been removed for clarity. OTU identifications are from http://vampsarchive.mbl.edu/diversity/diversity_old.php. Identifications more specific than the taxonomic order are shown in parentheses. Solid lines represent positive correlations, dashed lines represent negative correlations. Black lines show no time delay while red arrows are delayed by 1 month.

at ecological differences between these OTUs. Although some taxa seemed to follow inorganic nutrient concentrations (for example, SAR11 and *Deferribacteres*), others followed system productivity (for example, *Rhodobacteriales*) or temperature (*Gammaproteobacteria* OTUs; Figure 7). These observations, made possible by extended studies of microbial assemblages, will lead to deeper understanding of microbial niches in the ocean and elsewhere.

This study has confirmed that strong seasonal patterns occur in this surface water microbial community and that potential drivers of this structure could be identified from the observatory data. Strikingly, the variable with most explanatory power for overall bacterial richness was day length, which appears to be as important for describing temporal community structure in coastal temperate seas as pH is for describing spatial microbial structure in terrestrial ecosystems. This study has highlighted the added value of much longer temporal observations of natural communities. Although the overall community succession was robust, subtle changes in the patterns of individual taxa were observed and were only detectable because of the long (6 years) time series. Examples of different taxa showing different seasonal cycles were SAR11 and *Roseobacter*, which had nearly exactly opposite peaks in richness. Additionally, blooms of rare OTUs may be linked to changes in eukaryotic species and environmental variables. Seasonal succession in the community composition was robust and the most variable OTUs were best at predicting the time of year. Environmental factors, rather than interactions with eukaryotes, were better at explaining seasonal variance in bacterial community composition. Meanwhile, interactions were strongest within domains rather than between them, and correlative relationships were stronger between taxa than with environmental variables. This may indicate that biological rather than physical factors can be more important in defining the fine-grain community structure. Finally, in making comparisons of the bacterial OTU subsets, a fundamental stability in the community has been shown, which suggests that the robust seasonal cyclicity noted for the alpha- and beta-diversity is also self-evident in the interactions between members of the community.

Acknowledgements

We would like to thank Dr KR Clarke for providing extensive expertize in statistical modelling, and Margaret Hughes for providing the pyrosequencing technical support. All sequencing data and environmental metadata can be found in the INSDC SRA under ERP000118 (http://www.ebi.ac.uk/ena/data/view/ERP000118). This work was supported in part by the US Deptartment of Energy under Contract DE-AC02-06CH11357. JAF and JAS were supported by NSF Grant 0703159 and JAS and SH by the Sloan Foundation (ICoMM).

Disclaimer

References

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ *et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA* **108**(Suppl 1): 4516–4522.

Church MJ. (2009). Resource control of bacterial dynamics in the Sea. In: Kirchman DL (ed). *Microbial Ecology of the Oceans*. Wiley & Sons Inc.: NJ, USA.

Clarke KR, Gorley RN. (2006). *PRIMER v6: User Manual/Tutorial*. Primer-E Ltd.: Plymouth, UK.

Clarke KR, Warwick RM. (2001). *Change in Marine Communities: An Approach to Statistical Analysis and Interpretation*, 2nd edn. Primer-E Ltd: Plymouth, UK.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. (2009). Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694–1697.

Cullen JJ. (1991). Hypothesis to explain high-nutrient conditions in the open sea. *Limnology Oceanography* **36**: 1578–1590.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Fierer N, Jackson RB. (2006). The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* **103**: 626–631.

Fuhrman JA. (2009). Microbial community structure and its functional implications. *Nature* **459**: 193–199.

Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S. (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci USA* **103**: 13104–13109.

Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL *et al.* (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci USA* **105**: 7774–7778.

Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T *et al.* (2009). The seasonal structure of microbial communities in the Western English Channel. *Environ Microbiol* **11**: 3132–3139.

Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B *et al.* (2010). The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS One* **5**: e15545.

Huse SM, Welch DM, Morrison HG, Sogin ML. (2010). Ironing out the wrinkles in the rare biosphere through

improved OTU clustering. *Environ Microbiol* **12**: 1889–1898.
Kirchman DL, Rich JH, Barber RT. (1995). Biomass and biomass production of heterotrophic bacteria along 140-degrees-W in the equatorial pacific - effect of temperature on the microbial loop. *Deep-Sea Res Part Ii-Topical Stud Oceanography* **42**: 603–619.
Lewis K, Allen JI. (2009). Validation of a hydrodynamic-ecosystem model simulation with time-series data collected in the Western English Channel. *Mar Syst* **77**: 296–311.
Liu Z, DeSantis TZ, Andersen GL, Knight R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* **36**: e120.
Maidak BL, Cole JR, Lilburn TG, Parker CT, Saxman PR, Farris RJ *et al.* (2001). The RDP-II (Ribosomal Database Project). *Nucleic Acids Res* **29**: 173–174.
Morris RM, Vergin KL, Cho JC, Rappe MS, Carlson CA, Giovannoni SJ. (2005). Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site. *Limnology Oceanography* **50**: 1687–1696.
Nedwell DB, Rutter M. (1994). Influence of temperature on growth-rate and competition between 2 psychrotolerant antarctic bacteria - low-temperature diminishes affinity for substrate uptake. *Appl Environ Microbiol* **60**: 1984–1992.
Pingree RD, Griffiths KD. (1980). Currents driven by a steady uniform wind stress on the shelf seas around the British Isles. *Oceanol Acta* **3**: 227–235.
Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.
Reeder J, Knight R. (2010). Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat Methods* **7**: 668–669.
Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.
Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.
Southward AJ, Langmead O, Hardman-Mountford NJ, Aiken J, Boalch GT, Dando PR *et al.* (2005). Long-term oceanographic and ecological research in the Western English Channel. *Adv Mar Biol* **47**: 1–105.
Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–5267.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)

MBE Advance Access published May 17, 2012

Detecting Patches of Protein Sites of Influenza A Viruses under Positive Selection

Christina Tusche,[1,2] Lars Steinbrück,[1,2] and Alice C. McHardy*[,1,2]

[1]Max Planck Research Group for Computational Genomics and Epidemiology, Max Planck Institute for Informatics, Saarbrücken, Germany

[2]Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Institute for Computer Science, Düsseldorf, Germany

***Corresponding author:** E-mail: mchardy@mpi-inf.mpg.de.

Associate editor: Helen Piontkivska

Abstract

Influenza A viruses are single-stranded RNA viruses capable of evolving rapidly to adapt to environmental conditions. Examples include the establishment of a virus in a novel host or an adaptation to increasing immunity within the host population due to prior infection or vaccination against a circulating strain. Knowledge of the viral protein regions under positive selection is therefore crucial for surveillance. We have developed a method for detecting positively selected patches of sites on the surface of viral proteins, which we assume to be relevant for adaptive evolution. We measure positive selection based on dN/dS ratios of genetic changes inferred by considering the phylogenetic structure of the data and suggest a graph-cut algorithm to identify such regions. Our algorithm searches for dense and spatially distinct clusters of sites under positive selection on the protein surface. For the hemagglutinin protein of human influenza A viruses of the subtypes H3N2 and H1N1, our predicted sites significantly overlap with known antigenic and receptor-binding sites. From the structure and sequence data of the 2009 swine-origin influenza A/H1N1 hemagglutinin and PB2 protein, we identified regions that provide evidence of evolution under positive selection since introduction of the virus into the human population. The changes in PB2 overlap with sites reported to be associated with mammalian adaptation of the influenza A virus. Application of our technique to the protein structures of viruses of yet unknown adaptive behavior could identify further candidate regions that are important for host–virus interaction.

Key words: influenza, evolution, selection, adaptation, protein structure, pandemic.

Introduction

Influenza A viruses are single-stranded negative-sense RNA viruses typically causing short-term respiratory infections with considerable morbidity and mortality (WHO 2009). High mutation rates, swift spreading among individuals, and short replication times allow influenza A viruses to evolve and adapt rapidly to environmental conditions (Pybus and Rambaut 2009). Examples include the establishment of a virus in a novel host or an adaptation to escape increasing immunity of the host population to a circulating or a vaccine influenza strain (Dormitzer et al. 2011).

Past influenza pandemics resulted from the introduction into the human population of a transmissible virus with significantly different antigenicity from recent and currently circulating influenza strains. In all four pandemics that occurred within the last century, the respective influenza viruses carried hemagglutinin (HA) and several other genome segments of influenza A viruses from other host species, such as birds or swine (Webster et al. 1992; McHardy and Adams 2009). Configurational changes of multiple proteins of animal influenza A viruses are thought to be necessary to enable efficient replication and transmission in human hosts (Kuiken et al. 2006; Neumann and Kawaoka 2006). A region of particular importance for this process is the receptor-binding site of the viral hemagglutinin. It enables attachment to different types of host-specific glycosydic bonds on surface epithelial cells in the host respiratory and gastrointestinal tracts (Glaser et al. 2005; Neumann and Kawaoka 2006). Furthermore, certain areas of the viral polymerase complex determine host range (Neumann and Kawaoka 2006; Yamada et al. 2010). Following establishment of a virus within a novel host, additional adaptive changes are thought to optimize replication and dispersal rapidly within the population (Deem and Pan 2009; Hensley et al. 2009; Neumann et al. 2009; Smith et al. 2009).

Human influenza A viruses continuously change antigenically by accumulating changes in the antibody-binding sites of the viral surface proteins HA and neuraminidase (NA), (Bush et al. 1999; Smith et al. 2004; McHardy and Adams 2009; Weinstock and Zuccotti 2009). These changes allow reinfection of previously infected or vaccinated individuals. This requires the composition of the seasonal influenza A virus vaccine to be updated almost annually to ensure its continued effectiveness (Russell et al. 2008). Knowledge of the viral protein regions that are relevant for adaptation to a novel host or an increasingly immune population is therefore a crucial factor for the surveillance

and prevention of seasonal and pandemic influenza A virus infections.

Multiple methods allow identification of functional regions of proteins, for example, on the basis of evolutionary conservation ratios (Pupko et al. 2002; Glaser et al. 2003; Nimrod et al. 2005, 2008; Shazman et al. 2007; Ashkenazy et al. 2010). Regions under positive selection do not follow the assumption of strong conservation and can therefore not be detected by these methods. Other techniques predict the location of antibody-binding (epitope) sites based on structural and sequence information (Blythe and Flower 2005; El-Manzalawy et al. 2008; Rubinstein et al. 2008, 2009; Lacerda et al. 2010). However, besides epitope regions, receptor avidity–changing sites or host-specificity determinants can be subject to positive selection and might play a similarly important role for the adaptive evolution of influenza A viruses (Hensley et al. 2009). Furthermore, a part of the epitope regions is invariable due to functional and structural constraints.

Sites under positive selection indicate the relevance of a region within a protein for adaptation. Such sites can be identified based on the ratio of nonsynonymous to synonymous mutations (d*N*/d*S* ratio) (Bush et al. 1999). This has, for instance, identified regions of B- and T-cell epitopes which are under positive selection (Suzuki 2006). However, this measure is difficult to interpret directly when studying evolution within a population and lacks sensitivity when applied to individual sequence sites (Kryazhimskiy and Plotkin 2008). Other methods compare within-species with between-species substitution statistics or substitution rates at specific branches (Nei 2005; Nozawa et al. 2009). We have recently proposed how to identify individual alleles, or sets of mutations, instead of sites or genes, that might be under selection using a time series of sequence samples from human influenza A (H3N2) viruses (Steinbrück and McHardy 2011). Furthermore, maximum likelihood estimates of codon-based Markov models are used to detect sites under positive or directional selection (Yang 2000; Kosakovsky Pond et al. 2005, 2008) and can also consider the physiochemical properties of residues (Sainudiin et al. 2005). All these methods return statistics of positive selection for independent codons but do not consider protein structure and spatial information for sites. Other methods take the effects of solvent accessibility and pairwise interactions between amino acids into account in their evolutionary models (Robinson et al. 2003). In the method we describe here, we follow a similar approach but use a less complex evolutionary model and consider the spatial distribution of residues in a consecutive phase of our algorithm.

In contrast to this type of methods, we assume that not only mutations at individual sites but also of multiple sites within a certain region of a gene can cause adaptive protein conformation changes. Shape and charge modifications within larger patches of residues on the protein surface are important for viral adaptation to structural changes in the interacting proteins of the host (see e.g., Yamada et al. 2010). We therefore devised a method to detect dense patches showing a high average positive selection, using d*N*/d*S* estimates of positive selection for individual sites and information on the spatial distances between them. With this approach, we also included sites with a large, but not exceptionally large, d*N*/d*S* ratio. Such residues would be discarded by methods that rank sites based on a measure of selection and then cut the list below a certain threshold. With our method, such residues were included if their spatial position supported the continuity of a patch. By searching for clusters of sites that are close to each other in the protein structure and consistently exhibit elevated d*N*/d*S* values, one might have greater statistical power to detect adaptive evolution in genes compared to methods that test for elevated d*N*/d*S* ratios at individual sites.

As mentioned above, more advanced techniques can be used for estimating positive selection. We here rely on the d*N*/d*S* statistic to allow an easy understanding of the principles of our method. The d*N*/d*S* statistic used for clustering can easily be exchanged with other measures.

There are similar methods that search for clusters of positively selected sites (Suzuki 2004; Berglund et al. 2005; Zhou et al. 2008). These differ from ours in that they use a sliding window–based search for sphere-shaped clusters on the surface of the tertiary structure. Our approach does not require specification of a cluster radius nor does it restrict the geometrical form of the inferred clusters. We evaluated our method by applying it to HA data for human influenza A viruses of the subtypes H3N2 and H1N1. These are particularly suited for evaluation as large numbers of sequences are available and their interaction with the human host is very well studied. Additionally, we applied the method to HA and polymerase basic protein 2 (PB2) of swine-origin influenza virus (S-OIV) A/H1N1 to study the more recent development of the virus.

Materials and Methods

We implemented a graph-cut algorithm to cluster protein residues based on structural and evolutionary protein information. Our goal was to identify dense patches of spatially close residues on the protein surface that show significant signs of positive selection. Generally speaking, our algorithm includes residues in a patch if they show evidence for positive selection and are close to other patch residues. A patch is rated both by its average *P* value and the density of sites under selection. Individual sites can compensate for a weaker signal of positive selection by being close to neighbors with a strong signal. Structural protein models were used to identify the spatial coordinates of individual residues. To measure positive selection for individual sites, ancestral character states were inferred from phylogenetic trees constructed from available genetic sequences for a particular protein. Subsequently, d*N*/d*S* statistics for each site were calculated, according to the ratio of the number of synonymous and nonsynonymous changes mapping to the tree edges (Bush et al. 1999; Suzuki 2006). After clustering, the identified patches were visualized on the protein structure. The complete process is shown in figure 1.

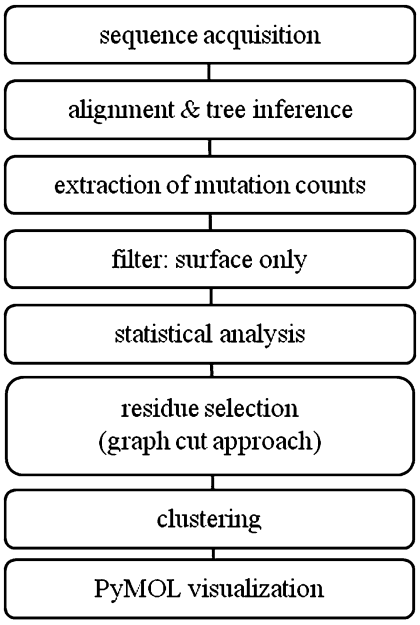


Fig. 1. Workflow for predicting patches under positive selection.

Structural Models

HA structures of the human influenza A/H3N2 virus, the human influenza A/H1N1, and S-OIV A/H1N1 were downloaded from the RSCB Protein Data Bank (PDB) (http://www.rcsb.org/) (for identifier codes of structures, sequences, and templates, see table 1). The analysis process was restricted to residues annotated in the PDB structure file and to sites found to be on the protein surface using the NetSurf software (Petersen et al. 2009). Structural models were generated for PB2 of the S-OIV isolate A/California/14/2009 (H1N1) based on the PB2 structures of PDB. To this end, the S-OIV PB2 sequence was compared with sequences of PB2 proteins with experimentally determined structure using Blast (Altschul et al. 1990). For PB2, there was no single structural template that covered all protein domains. Therefore, two models were generated from two templates, one for the PB2cap and one for the PB2c domain. The highest sequence identity, the largest coverage of the S-OIV protein, and the quality according to resolution and free R-factor values were used as criteria to select the best matching structural templates for the PB2cap and PB2c domains. The S-OIV sequences were aligned to the templates with MODELLER (version 9v6) (Sali and Blundell 1993). The alignments are expected to be reliable, given a sequence identity of 94% and a lack of insertions and deletions. Subsequently, the structural models were generated with MODELLER.

Sequence Data, Alignments, and Phylogenetic Tree Construction

Available HA sequences of the seasonal influenza A virus, subtypes H1N1 and H3N2, were downloaded from the GISAID EpiFlu database (http://platform.gisaid.org). Only sequences longer than 1,500 bp were selected, resulting in 1,734 and 3,221 sequences for H1 and H3, respectively (supplementary table S2, Supplementary Material online). Alignments of DNA and protein sequences were computed with MUSCLE (Edgar 2004), and manually curated. Phylogenetic trees were inferred with PhyML v3.0 (Guindon and Gascuel 2003) under the general time reversible (GTR) + I + Γ4 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 (Γ4) 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 (Zwickl 2006). Substitution events were inferred for the genome segment tree topologies from intermediates reconstructed with accelerated transformation (AccTran; Felsenstein 2004). The total number of substitutions occurring on all reconstructed internal branches was then calculated for each site independently. These numbers were used to compute the d*N*/d*S* ratio for each codon site (Bush et al. 1999; Suzuki 2006). The ratios were transformed to *P* values by a one-sided Fisher test for independence of the d*N* and d*S* values at an individual site and the mean values of the protein. *P* values were corrected for the ranking comparison with the false discovery rate (Benjamini and Yekutieli 2001) and used as a measure of selection for individual sites. Furthermore, 3,419 sequences of the PB2 protein and 7,373 sequences of the HA protein of the 2009 S-OIV A/H1N1 strains were downloaded from the GISAID EpiFlu database (supplementary table S2, Supplementary Material online). Phylogenetic trees were inferred using neighbor joining with PAUP (Swofford 2003) under the GTR model. Sequence alignment and residue statistics were inferred as described above.

Structural Clustering

Before clustering, all spatial coordinates were normalized to fit the protein structure into a hypercube of size 1. For

Table 1. Sequence Codes and PDB Codes of Selected Templates.

Protein	Query S-OIV Sequence	Template PDB Code and Chain	Template PDB Sequence	Query/Template Sequence Identity (%)
H1 (seas)	—	2wrgH,I	A/Brevig Mission/1/1918[1]	—
H3 (seas)	—	3hmgA,B	A/Aichi/2/1968[2]	—
H1 (swl)	—	3al4A,B	A/California/04/2009[3]	—
PB2cap	A/California/14/2009[3]	2vqzA	A/Victoria/3/1975[2]	94.00
PB2c	A/California/14/2009[3]	2vy6A	A/Victoria/3/1975[2]	94.00

Note.—Strains are of the subtypes [1]H1N1, [2]H3N2, and [3]H1N1swl.

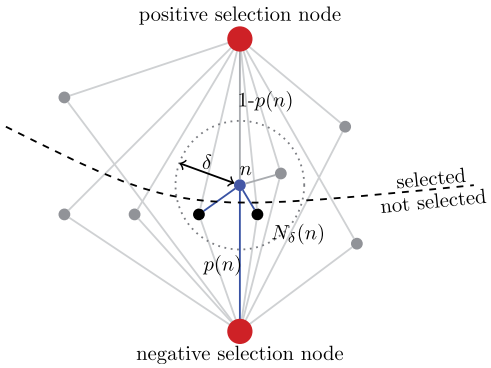


FIG. 2. Schematic drawing of the graph-cut approach. The minimum cut minimizes the sum of weights of all edges cut by the line separating the positive and negative selection nodes. For a single node *n*, these are the lines shown in blue: the scaled distances to the nonselected neighbors in $N_\delta(n)$ and the connection to the other side (i.e., the negative) selection node with the weight $P(n)$.

clustering with a graph-cut algorithm (Boykov et al. 2002), we constructed a graph in which each node represents a residue in the protein. Edges were added between all pairs of residues *m* and *n* for which the Euclidean distance dist(*m*, *n*) was below a threshold δ, and these edges were weighted according to their spatial distance (fig. 2). Weights were set to be in inverse exponential proportion to the Euclidean distance dist(*m*, *n*), that is, the closer the residues were located relative to each other on the protein structure, the larger the weight of the corresponding edge. Therefore, nodes that are close to each other have a strong connection to each other. We then augmented the graph with two additional nodes, which we call the "positive selection node" and the "negative selection node," corresponding to "source" and "sink" nodes in a standard graph-cut formulation. These two special nodes are connected to each residue node, with the weights equal to the *P* value $P(n)$ of the residue *n* in the case of the negative selection node or $1 - P(n)$ in the case of the positive selection node. Thus, residues that have high dN/dS ratios (large $1 - P(n)$) have a strong connection with the positive selection node, whereas nodes with low dN/dS values (large $P(n)$) have a strong connection with the negative selection node. The two types of edges and edge weights were added to the graph to represent the spatial information for each residue (by adding distances to close neighbors) and the evolutionary evidence for selection (by encoding the *P* value of the dN/dS ratios).

A "graph cut" will divide this graph in two halves, one containing the positive selection node and the other containing the negative one (fig. 2). A "minimum graph cut" is a graph cut that minimizes the sum *E* of the weights of the edges connecting these two halves:

$$E = \sum_{n\in\mathrm{Pos}} P(n) + \alpha \sum_{n\in\mathrm{Neg}} \bar{P}(n) + \beta \sum_{n\in\mathrm{Pos}} \sum_{\substack{m\in\mathrm{Neg},\\ m\in N_{\delta(n)}}} e^{-\mathrm{dist}(m,n)},$$

where $\bar{P}(n) = 1 - P(n)$, pos represents all nodes assigned to the positive selection half, Neg represents all nodes assigned to the negative one, and $N_\delta(n)$ represents all neighbors of residue *n* within a distance less than δ. This means that the minimum cut will select residues to be in Pos if they show strong signs of positive selection (i.e., a low *P* value) and if they separate well spatially from the residues in Neg. The distance δ defines how many sites of a single residue are considered to be neighbors. We set δ such that a residue has, on average, ten close neighbors. The factor β weighs this distance statistic. The smaller the β, the more likely the method is to balance the residue evenly between the positive and the negative selection set halves according to the ratio 1:α (we set $\alpha = 1$). The larger the β, the more expensive an even distribution becomes, and the more stringently the method searches for a small exclusive set of residues that spatially separate well from the rest. Since the total distance statistic is dependent on the number of residues in the protein, β has to be set manually (see supplemental text S1, Supplementary Material online). Finally, the selected residues were grouped into patches by merging all residues within a spatial distance *d* of each other into a set. The parameter *d* was set to represent the first quartile of all pairwise distances in the protein. Finally, we excluded outliers by filtering out all patches that contained two or less residues. Patches were identified for the H1 and H3 proteins of human influenza A viruses of the subtypes H1N1 and H3N2, respectively, and for the HA and PB2 proteins of the 2009 S-OIV of subtype H1N1. Subsequently, we analyzed their enrichment with known epitope sites (Caton et al. 1982; Wiley and Skehel 1987) and receptor avidity–changing sites (Hensley et al. 2009).

Evaluation and Visualization

For evaluation, we calculated the precision (ratio of selected epitope sites to all selected residues) and recall (ratio of selected epitope sites to all epitope sites) of the inferred patches based on the epitope regions defined for subtypes H1 (Caton et al. 1982) and H3 (Wiley et al. 1981; Wiley and Skehel 1987; Suzuki 2006). For a list of epitope sites used as a reference for evaluation, see supplementary table S1 (Supplementary Material online). The identified patches of all proteins were visualized with PyMOL software (Schrödinger 2012).

Results

We analyzed the merits of a clustering technique based on a graph-cut formalization for identification of patches of sites under selection on the surface of the HA and PB2 proteins of several influenza A viruses. Our goal was to rediscover regions known to play an important role in the interaction of the virus with the host's immune system and that comprise many important sites for adaptation. We therefore first considered known antigenic site regions on the HA of the human influenza A virus (Caton et al. 1982; Wiley and Skehel 1987) as our approximate reference for evaluation. The clustering algorithm identified dense patches of residues, which mostly consisted of sites with substantial deviation from the expected value of the protein-wide dN/dS. In comparison, a site ranking based

Table 2. Precision and Recall of Different Settings and Approaches When Put to the Task of Detecting Influenza Epitope Sites.

Setting	Recall (H1)	Precision (H1)	Recall (H3)	Precision (H3)
Graph cut	0.53	0.49	0.25	0.94
PV 0.05	0.19	0.4	0.15	0.86
PV 0.1	0.19	0.4	0.17	0.81

on P value alone resulted only in a low sensitivity for discovering relevant sites, with only 6 of 32 (H1) or 19 of 131 (H3) known antigenic or receptor avidity–changing sites exhibiting a significant ($P < 0.05$) signal. To compare this approach with our method, we calculated the precision and recall for sites selected by setting a P value ranking at $\theta = 0.05$ (PV 0.05) or $\theta = 0.1$ (PV 0.1) as a threshold as well as calculating these characteristics for the sites in patches identified with our graph-cut approach. Our evaluation (table 2) showed that including information on the spatial proximity of residues under selection and applying our clustering algorithm resulted in a significant improvement in recall (i.e., a larger number of epitope sites being identified) while maintaining similar or better precision (meaning that a similar or lower number of nonepitope sites were inferred). In the light of a recently proposed hypothesis on the relevance of receptor avidity–changing sites (Hensley et al. 2009), as opposed to the epitope sites of hemagglutinin in subtype H1 antigenic evolution, we also tested the value of these sites as a reference and compared these with the inferred patches of sites. The currently available data do not allow discrimination between these two hypotheses, as the reference sites of known receptor avidity–changing sites and antigenic sites overlap greatly (fig. 3). Still, residues 156 and 158, found to play the most significant role in receptor avidity, are included in the second patch identified for subtype H1.

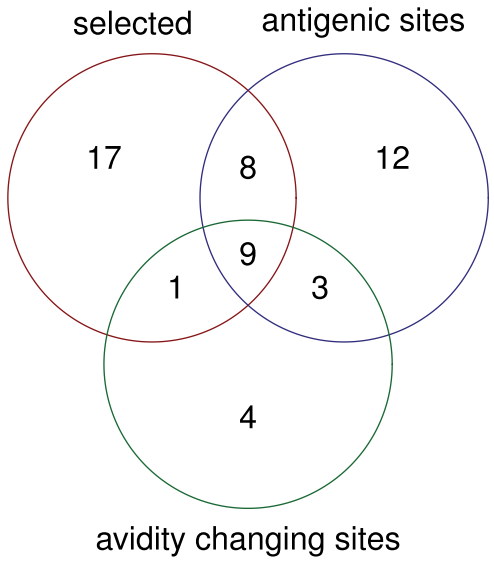


Fig. 3. Overlap between selected epitope and avidity-changing sites. Venn diagram showing the overlap between subtype H1 residues in patches selected by the dN/dS graph-cut approach (red), the influenza A H1 epitope sites according to Caton et al. (1982) (blue), and avidity-changing sites according to Hensley et al. (2009) (green).

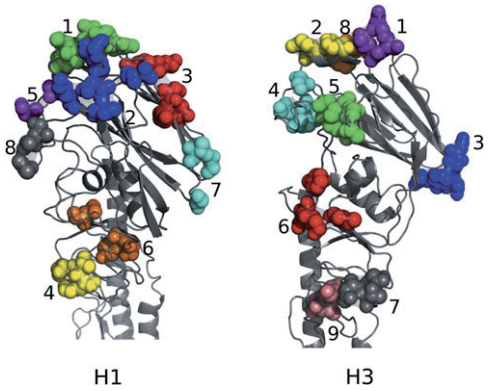


Fig. 4. Patches under positive selection on HA. Patches on the HA protein structure of subtype H1 and H3 selected by the graph-cut algorithm. Patches are numbered according to tables 3 and 4.

Of the detected patches on the HA protein surface (fig. 4 and tables 3 and 4), several include known epitope or receptor avidity–changing sites up to a fraction of 100%. The patches contain many sites that are relevant for antigenic evolution (Matrosovich et al. 1997; Hay et al. 2003; Lin et al. 2004; Yamada et al. 2010), including position 145, which has been shown experimentally to have a high antigenic impact (Smith et al. 2004).

We also compared our results with similar techniques for predicting the properties of sites under positive selection or relevant for adaptive evolution. Our predictions match 7 of 13 sites inferred to be under positive selection by a maximum likelihood approach (Yang 2000). However, 10 of these 13 sites are at least direct neighbors of those listed by our method, confirming its ability to find positively selected regions on the tertiary structure. Similar observations can be made for sites identified in Fitch et al. (1997), where five of six are matches or direct neighbors and the sites discussed in Bush et al. (1999) and Yang (2000) (10 of 13). Furthermore, several techniques combine biochemical and phylogenetic information to gain insights into the adaptive evolution of influenza A. It has recently been suggested that HA evolves by increasing the number of charged amino acids in regions recognized by the immune system, particularly in the dominant epitope (i.e., the one with the highest proportion of amino acid mutations, see Pan et al. 2011). We therefore compared the number of charged and uncharged amino acids in

Table 3. Patches and Residues Selected for the Influenza A Hemagglutinin Protein, Subtype H1.

Patch	Residues
1	187, 188, 189, 190, 192, 193, 196, 197, 198
2	131, 132, 133, 158, 156, 129
3	163, 165, 166, 244, 248
4	274, 275, 276
5	227, 225, 219
6	82, 81, 56
7	240, 169, 173
8	142, 144, 145

NOTE.—Underlined numbers refer to known epitope sites according to Caton et al. (1982) and supplementary table S1 (Supplementary Material online). All positions are given in H3 numbering (Aoyama et al. 1991).

Table 4. Patches and Residues Selected for the Influenza A Hemagglutinin Protein, Subtype H3.

Patch	Residues
1	156, 157, 158, 159
2	188, 189, 192, 193
3	171, 172, 173, 174, 175
4	186, 220, 229
5	137, 140, 142, 144, 145
6	62, 91, 92, 94
7	53, 275, 276
8	196, 197, 198, 199
9	47, 48, 50

NOTE.—Underlined numbers refer to known epitope sites (Wiley et al. 1981; Wiley and Skehel 1987; Suzuki 2006; see supplementary table S1, Supplementary Material online). All positions are given in H3 numbering.

the H1 and H3 consensus sequences for selected sites in the patches and sites lying outside the patches. Indeed, we found that the percentage of charged amino acids is much higher within patches (H1: 67%, H3: 67%) than outside patches (H1: 27%, H3: 28%). Finally, other authors suggest statistics based on rates of substitutions toward specific residues (Kosakovsky Pond et al. 2008; Kryazhimskiy and Plotkin 2008) or based on epistatic effects between pairs of sites (Kryazhimskiy et al. 2011). The overlap between the predictions by both methods and ours is not large, possibly due to the different nature of the measured quantities and statistics, and because, as Kryazhimskiy et al. 2011 discuss, hitchhiking changes without selective impact might comprise a fraction of identified epistatic pairs, particularly among the trailing change of a pair. However, our simple criterion for positive selection can easily be exchanged for more advanced estimates for adaptive evolution, allowing a search for clusters of residues that show significantly elevated statistics of such properties.

Additionally, we identified one patch in H1 without known epitope sites, but with similar evidence for positive selection as the other patches, which indicates its potential importance for antigenic evolution (table 3 and fig. 4, patch 4). For both subtypes, one patch in HA overlaps with the receptor-binding site of the protein. This could be due to the overlap of the antigenic and receptor-binding regions. However, the receptor-binding site, particularly position 189, is also known to be relevant for adaptation to avian and human hosts (Matrosovich et al. 1997; Sorrell et al. 2009). Both the H1 and H3 of human influenza A viruses show evidence of selection acting upon the receptor-binding region when grown in eggs, due to the effects of egg adaptation (Robertson et al. 1987; Gambaryan et al. 1999). Therefore, part of the signal in the receptor-binding sites could also be due to the effects of egg cultivation.

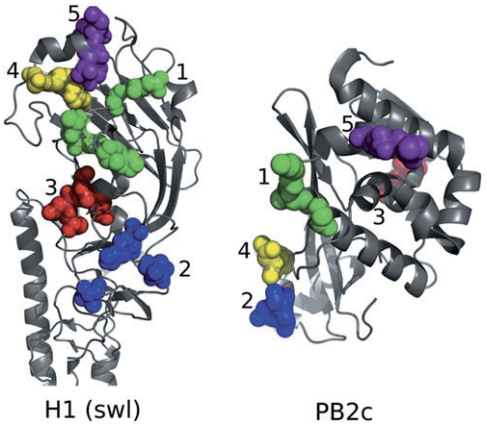


FIG. 5. Patches under positive selection on the HA and PB2 proteins of 2009 S-OIV. Patches on the 2009 swine-origin influenza A protein structures of HA and the c-terminal region of PB2, selected by the graph-cut algorithm. Patches are numbered according to tables 5 and 6.

As a second application, we analyzed data of 2009 S-OIV A/H1N1. The molecular basis of the successful establishment of the triple reassortant swine virus, which contains several recently acquired avian segments (Smith et al. 2009), in the human host is not fully understood. It has, in particular, been argued that lysine at position 627 of the PB2 protein of the viral polymerase complex, instead of the avian-like glutamic acid, is required for successful transmission and replication within mammals (Gabriel et al. 2005). However, the 2009 H1N1 virus still has lysine at position 627 in PB2, which it has maintained since its descent from an originally avian lineage. A change at residue 591 has been proposed to compensate for the lack of lysine in 627, allowing its efficient replication in mammals (Yamada et al. 2010). We searched for regions with evidence for positive selection and relevance for adaptation of PB2 since the introduction of the 2009 S-OIV into the human population. The virus might have acquired changes in PB2 to further optimize replication and transmission in the novel host. We identified five patches. The first (fig. 5 and table 5) is localized in a region around residue 591, which lends support to its relevance for mammalian and, in particular, human adaptation. To gain more insight, we allowed the method to also report patches containing only two residues. The resulting second patch surrounds residue 714, which is known to increase polymerase activity in mammals (Gabriel et al. 2005).

We furthermore analyzed the genetic sequences and protein structure of the HA protein of 2009 S-OIV A/H1N1. We identified five patches of sites under positive

Table 5. Patches and Residues Selected for the PB2 Protein of the 2009 Swine-Origin Influenza A/H1N1 Virus.

Patch	Residues
1	586, 588, 590
2	714, 715
3	660, 661
4	709, 711
5	575, 578

Table 6. Patches and Residues Selected for the HA Protein of the 2009 Swine-Origin Influenza A/H1N1 Virus.

Patch	Residues
1	135, 137, 140, 141, 142, 144, 145
2	53, 54, 56, 57, 276
3	63, 91, 92, 93, 94
4	186, 188, 189, 218
5	197, 198, 199, 200

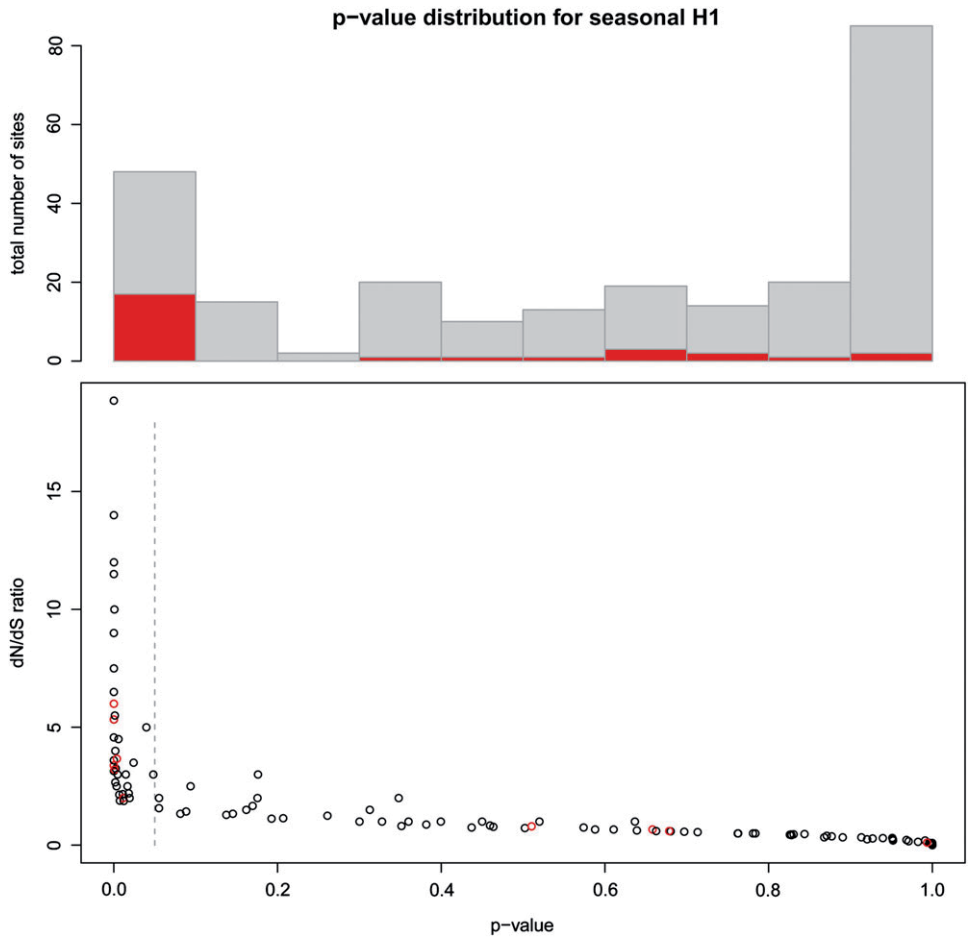


Fig. 6. Epitope sites not under positive selection. The histogram displays the ratio of residues within the corresponding *P* value intervals and demonstrates that many epitope sites feature insignificant *P* values resulting from an average d*N*/d*S* ratio. Epitopic sites are marked in red. The lower plot shows the distribution of the *P* values versus the d*N*/d*S* ratios for all residues of the H1 subtype.

selection. The first one (fig. 5 and table 6) overlaps with the Ca2 epitope site of seasonal H1 (Caton et al. 1982). The remaining ones cluster densely at the head of the protein, indicating emerging areas of relevance for adaptation and antigenic evolution of the 2009 H1N1 virus.

Our software, AdaPatch, is available online (http://www.cs.uni-duesseldorf.de/AG/AlgBio) and can also be applied to analyze other viral proteins.

Discussion

We have developed a technique for identification of candidate regions under positive selection in viral proteins. Our method utilizes a common measure of selection, and state-of-the-art techniques for phylogenetic tree inference, ancestral state reconstruction, and clustering or separation techniques. It requires only sequence information and a PDB structure file as input. We identified clusters of sites under positive selection based on information on the spatial proximity of sites. Although other methods search for functional importance, that is, conserved regions, or focus specifically on the detection of epitope sites, we aim to provide a fast and easy solution for identification of patches of arbitrary shape and size whose combined evolutionary signature indicates their importance for viral adaptive evolution. In addition to d*N*/d*S* statistics, other methods for evaluating positive selective pressure (e.g., Kosakovsky Pond et al. 2005) can easily be included.

Focusing on the HA of two subtypes of the seasonal influenza A virus and the HA and PB2 proteins of 2009 S-OIV A/H1N1, we searched for patches of sites under positive selection on their protein structures. The patches we identified for the HA of the seasonal influenza A viruses largely map to known epitope sites and sites associated with receptor binding. Among the patch sites, we identified for the PB2 protein of the 2009 S-OIV are sites with known relevance for successful replication in mammalian hosts. Our analysis showed that our approach increases the predictive accuracy relative to the commonly used approach of searching for individual sites with significantly deviating d*N*/d*S* statistics. This indicates that focusing on evolutionary change in larger regions, instead of individual sites, is helpful for revealing patches of residues that are important for adaptation, which together show a stronger signal of positive selection.

The precision and recall values for detecting known epitope sites based on patches under positive selection are rather low overall, mostly at or below 50%, indicating that not all sites in the epitope regions are under positive

selection and contributing to adaptation of the viral HA. Influenza A epitopes seem to be variable only in part (fig. 6) and probably change over time, thus diluting the overall signal of positive selection. Furthermore, receptor avidity–changing sites or host-specificity determinants may play a similarly important role in adaptive evolution, which lowers precision if one considers only the epitope sites that are predicted to be evolving under positive selection.

We evaluated our method using the influenza A viruses as they are very well studied and much is already known about the relevant sites for adaptive evolution. Still, our inferred patches might be more informative than individual sites for monitoring circulating viral strains for adaptive changes with relevance for transmission and spread in the human population. Our analyses of HA and PB2 identified many sites known to be relevant for antigenic drift or for the adaptation of influenza A to its host, improving its ability for infection, replication, and immune evasion. We therefore suggest analysis of the new patches identified in this study to determine the underlying causes of their consistent variability. We also suggest applying the method to other protein structures of rapidly evolving viruses with as yet unknown adaptive behavior in order to identify candidate regions that are important for virus–host interaction.

Supplementary Material

Supplementary text S1 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Dr Francisco Domingues for the provision of structural models of the PB2 protein of the A/California/04/2009 strain. We gratefully acknowledge funding by Max Planck Society and Heinrich Heine University Düsseldorf.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.

Aoyama T, Nobusawa E, Kato H. 1991. Comparison of complete amino acid sequences among 13 serotypes of hemagglutinins and receptor-binding properties of influenza A viruses indirect immunofluorescence. *Mutagenesis* 182:475–485.

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* 38:1–5.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann Stat.* 29:1165–1188.

Berglund AC, Wallner B, Elofsson A, Liberles DA. 2005. Tertiary windowing to detect positive diversifying selection. *J Mol Biol.* 60:499–504.

Blythe M, Flower D. 2005. Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.* 14:246–248.

Boykov Y, Veksler O, Zabih R. 2002. Fast approximate energy minimization via graph cuts. *Pattern Anal Mach Learn.* 23: 1222–1239.

Bush RM, Fitch WM, Bender CA, Cox NJ. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol.* 16:1457–1465.

Caton AJ, Brownlee GG, Yewdell JW, Gerhard W. 1982. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell* 31:417–427.

Deem MW, Pan K. 2009. The epitope regions of H1-subtype influenza A, with application to vaccine efficacy. *Protein Eng Des Sel.* 22:543–546.

Dormitzer PR, Galli G, Castellino F, Golding H, Khurana S, Del Giudice G, Rappuoli R. 2011. Influenza vaccine immunology. *Immunol Rev.* 239:167–177.

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

El-Manzalawy Y, Dobbs D, Honavar V. 2008. Predicting linear B-cell epitopes using string kernels. *J Mol Recognit.* 21:243–255.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates, Inc.

Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc Natl Acad Sci U S A.* 94:7712–7718.

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.

Gambaryan AS, Robertson JS, Matrosovich MN. 1999. Effects of egg-adaptation on the receptor-binding properties of human influenza A and B viruses. *Virology* 258:232–239.

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19:163–164.

Glaser L, Stevens J, Zamarin D, Wilson IA, García-Sastre A, Tumpey TM, Basler CF, Taubenberger JK, Palese P. 2005. A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. *Virology* 79:11533–11536.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.

Hay AJ, Lin Y, Gregory V, Bennet M. 2003. WHO collaborating centre for reference and research on influenza, Annual Report. Tech. Rep. London: National Institute for Medical Research.

Hensley SE, Das SR, Bailey AL, et al. (11 co-authors). 2009. Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science* 326:734–736.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.

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.

Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB. 2011. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet.* 7:e1001301.

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

Kuiken T, Holmes EC, McCauley J, Rimmelzwaan GF, Williams CS, Grenfell BT. 2006. Host species barriers to influenza virus infections. *Science* 312:394–397.

Lacerda M, Scheffler K, Seoighe C. 2010. Epitope discovery with phylogenetic hidden Markov models. *Mol Biol Evol.* 27: 1212–1220.

Lin YP, Gregory V, Bennett M, Hay A. 2004. Recent changes among human influenza viruses. *Virus Res.* 103:47–52.

Matrosovich MN, Gambaryan AS, Teneberg S, Piskarev VE, Yamnikova SS, Lvov DK, Robertson JS, Karlsson KA. 1997. Avian influenza A viruses differ from human viruses by recognition of sialyloligosaccharides and gangliosides and by

a higher conservation of the HA receptor-binding site. *Virology* 233:224–234.
McHardy AC, Adams B. 2009. The role of genomics in tracking the evolution of influenza A virus. *PLoS Pathog.* 5:e1000566.
Nei M. 2005. Selectionism and neutralism in molecular evolution. *Mol Biol Evol.* 22:2318–2342.
Neumann G, Kawaoka Y. 2006. Host range restriction and pathogenicity in the context of influenza pandemic. *Emerg Infect Dis.* 12:881–886.
Neumann G, Noda T, Kawaoka Y. 2009. Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature* 459:931–939.
Nimrod G, Glaser F, Steinberg D, Ben-Tal N, Pupko T. 2005. In silico identification of functional regions in proteins. *Bioinformatics* 21:i328–i337.
Nimrod G, Schushan M, Steinberg DM, Ben-Tal N. 2008. Detection of functionally important regions in "hypothetical proteins" of known structure. *Structure* 16:1755–1763.
Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci U S A.* 106:6700–6705.
Pan K, Long J, Sun H, Tobin GJ, Nara PL, Deem MW. 2011. Selective pressure to increase charge in immunodominant epitopes of the H3 hemagglutinin influenza protein. *J Mol Biol.* 72:90–103.
Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. 2009. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol.* 9:51.
Pupko T, Bell RE, Mayrose I, Glaser F, Ben-tal N. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18(Suppl 1):S71–S77.
Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet.* 10:540–550.
Robertson JS, Bootman JS, Newman R, Oxford JS, Daniels RS, Webster RG, Schild GC. 1987. Structural changes in the haemagglutinin which accompany egg adaptation of an influenza A(H1N1) virus. *Virology* 160:31–37.
Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol.* 20:1692–1704.
Rubinstein ND, Mayrose I, Halperin D, Yekutieli D, Gershoni JM, Pupko T. 2008. Computational characterization of B-cell epitopes. *Mol Immunol.* 45:3477–3489.
Rubinstein ND, Mayrose I, Pupko T. 2009. A machine-learning approach for predicting B-cell epitopes. *Mol Immunol.* 46: 840–847.
Russell CA, Jones TC, Barr IG, et al. (11 co-authors). 2008. The global circulation of seasonal influenza A (H3N2) viruses. *Science* 320:340–346.
Sainudiin R, Wong W, Yogeeswaran K, Nasrallah J, Yang Z, Nielsen R. 2005. Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J Mol Biol.* 60:315–326.
Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol.* 234:779–815.
Schrödinger. 2012. The PyMOL molecular graphics system, Version 1.4. New York: Schrödinger, LLC.
Shazman S, Celniker G, Haber O, Glaser F, Mandel-Gutfreund Y. 2007. Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. *Nucleic Acids Res.* 35:W526–W530.
Smith D, Lapedes A, de Jong J, Bestebroer T, Rimmelzwaan G, Osterhaus A, Fouchier R. 2004. Mapping the antigenic and genetic evolution of influenza virus. *Science* 305:371.
Smith GJD, Vijaykrishna D, Bahl J, et al. (11 co-authors). 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459:1122–1125.
Sorrell EM, Wan H, Araya Y, Song H, Perez DR. 2009. Minimal molecular constraints for respiratory droplet transmission of an avian-human H9N2 influenza A virus. *Proc Natl Acad Sci U S A.* 106:7565–7570.
Steinbrück L, McHardy AC. 2011. Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic Acids Res.* 39:e4.
Suzuki Y. 2004. Three-dimensional window analysis for detecting positive selection at structural regions of proteins. *Mol Biol Evol.* 21:2352–2359.
Suzuki Y. 2006. Natural selection on the influenza virus genome. *Mol Biol Evol.* 23:1902–1911.
Swofford D. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.
Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. 1992. Evolution and ecology of influenza A viruses. *Microbiol Rev.* 56:152–179.
Weinstock DM, Zuccotti G. 2009. The evolution of influenza resistance and treatment. *JAMA* 301:1066–1069.
WHO. 2009. Influenza Fact Sheet No 211. [cited 2011 Apr 12]. Available from: http://www.who.int/mediacentre/factsheets/fs211/en/index.html
Wiley D, Skehel J. 1987. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annu Rev Biochem.* 56:365–394.
Wiley D, Wilson I, Skehel J. 1981. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* 289:373.
Yamada S, Hatta M, Staker BL, et al. (11 co-authors). 2010. Biological and structural characterization of a host-adapting amino acid in influenza virus. *PLoS Pathog.* 6:e1001034.
Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J Mol Biol.* 51:423–432.
Zhou T, Enyeart PJ, Wilke CO. 2008. Detecting clusters of mutations. *PLoS One* 3:e3765.
Zwickl D. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [dissertation]. [Austin (TX)]: University of Texas.