UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN BIOSCIENZE E BIOTECNOLOGIE
INDIRIZZO: BIOTECNOLOGIE
CICLO XXVII

GENOMIC AND BIOINFORMATIC APPROACH TO AVIAN INFLUENZA VIRUS EVOLUTION

Direttore della Scuola: Ch.mo Prof. Paolo Bernardi
Coordinatore d'indirizzo: Ch.ma Prof.ssa Fiorella Lo Schiavo
Supervisore: Ch.mo Prof. Francesco Filippini
Co-Supervisore: Dott. Giovanni Cattoli

Dottoranda: Adelaide Milani

Alla mia famiglia

TABLE OF CONTENTS

RIASSUNTO

I virus zoonotici, cioè in grado di infettare l'uomo e alcune specie animali, hanno un impatto significativo e costituiscono una costante, potenziale minaccia sia per la salute pubblica umana che per quella animale. Ecosistemi dagli equilibri modificati, una crescente urbanizzazione e connessioni facilitate hanno influenzato sempre piu' il rapporto tra patogeni e specie ospiti affini. Negli ultimi anni la fonte della maggior parte dei virus potenzialmente pericolosi e in grado di causare malattie emergenti sembra derivi da ospiti di origine animale; si tratta prevalentemente di virus a RNA che, grazie alla possibilità di moltiplicarsi in breve tempo all'interno di una popolazione ampia ed all'alto tasso di mutazione, permettono una rapida evoluzione, un'elevata variabilità genetica e la selezione di nuove varianti. Un adeguato e costante programma di sorveglianza, la condivisione di conoscenze e una collaborazione tra diverse competenze professionali sono fondamentali e necessarie per seguire l'evoluzione virale e per formulare politiche di sanità pubblica efficienti (Howard e Fletcher, 2012).

L' Influenza virus di tipo A è considerato uno dei virus a RNA più importanti, tanto per il suo potenziale ruolo zoonotico nell'interfaccia animale-umano, quanto per la salute globale e l'impatto economico. Quasi ogni anno epidemie di influenza provocano morbilità e mortalità nell'uomo e talvolta gli stessi virus possono essere associati a pandemie.

Il serbatoio naturale dei virus influenzali di tipo A è rappresentato dagli uccelli, sia selvatici che domestici (influenza aviaria) (http://www.cdc.gov/flu/about/viruses/transmission.htm); in particolare gli uccelli selvatici sembrano costituire la fonte dell'influenza A virus tutte le altre specie animali. Diverse tecniche sono disponibili per studiare i virus e caratterizzarli geneticamente al fine di capirne il loro comportamento, le dinamiche evolutive, il loro rapporto con l'ospite e la loro origine e per sviluppare profilassi e terapie adeguate creando un valido supporto durante la fasi di sorveglianza e diagnosi di un'eventuale epidemia .

Nell'ambito del mio dottorato è stato utilizzato un approccio integrato, sia genomico che strutturale, per studiare l'evoluzione dell'influenza aviaria; particolare interesse è stato rivolto allo studio dell'emoagglutinina virale, la principale glicoproteina di superficie, appartenente ai sottotipi H5, H7 e H9 (i principali sottotipi "aviari" responsabili di infezione nell'uomo).

Le analisi mediante *Next Generation Sequencing* (NGS) hanno favorito lo studio e la caratterizzazione della complessità nella popolazione virale, consentendo di monitorare finemente l'evoluzione delle varianti geneticamente correlate presenti all'interno della popolazione virale tramite l'identificazione delle mutazioni a bassa frequenza. Per confrontare ed analizzare i dati genetici, l'approccio filogenetico si è rivelato un utile strumento per l'analisi dell'evoluzione virale; è stato usato per spiegare l'epidemiologia molecolare, la trasmissione e l'evoluzione virale. Al fine di ottenere una visione più completa in termini di 'evoluzione funzionale', l'analisi filogenetica è stata integrata con le informazioni provenienti dal confronto strutturale. L'approccio strutturale, considerando lo spazio tridimensionale dell'emoagglutinina, ha dimostrato di poter essere uno strumento utile per evidenziare eventuali somiglianze e per ispezionare e valutare quei motivi il cui ruolo non può essere correttamente interpretato utilizzando le sole sequenze primarie. Infatti, nelle sequenze primarie il peso delle mutazioni non tiene conto dell'effetto sul fold o sulle proprietà di superficie, mentre nelle strutture tridimensionali, quanto ciascuna mutazione sia in grado di influenzare le caratteristiche strutturali e le interazioni, è direttamente rilevabile. Questo approccio ha inoltre

portato un ulteriore contributo all'analisi filogenetica. In particolare lo studio si è concentrato sull'analisi delle dinamiche evolutive e delle strategie adattative dei sottotipi H7N1 ed H7N3 dell'influenza aviaria circolanti nel Nord Italia per periodi di tempo analoghi e in condizioni epidemiologiche simili. Inoltre è stato utilizzato il *deep sequencing* per studiare le dinamiche evolutive e di trasmissione intra- e inter-ospiti del virus aviario sottotipo H7N7 che colpì alcuni allevamenti italiani nel 2013. L'analisi NGS è stata utilizzata per caratterizzare la complessità della popolazione virale in due gruppi di animali sperimentalmente infetti con lo stesso virus ad alta patogenicità (HPAI) H5N1 ed immunizzati con distinti vaccini. E' stato inoltre eseguito un ampio confronto strutturale su domini e sub-regioni dell'emoagglutinina di diversi sottotipi del virus dell'influenza, con particolare interesse per i diversi clades di HPAI H5N1 circolanti in Egitto (ove l'influenza aviaria è endemica nei volatili), per indagare eventuali variazioni dominio-specifiche. I virus influenzali del sottotipo H9 sono stati analizzati da un punto di vista sia filogenetico che strutturale, per rilevare caratteristiche tipo specifiche e verificare se la variazione delle proprietà di superficie possa essere un marcatore di 'evoluzione funzionale' dei determinanti di superficie virali, come dimostrato nel sottotipo H5N1. Questo lavoro suggerisce che il confronto e l'integrazione tra analisi genomica, filogenetica e strutturale può aiutare a capire l' 'evoluzione funzionale' del virus dell'influenza aviaria di tipo A.

ABSTRACT

Viral zoonotic agents have a significant impact both on human and veterinary public health. Ecosystems changes, increasing urbanization and easy connection have influenced the balance between pathogen and related host species. In recent years most threatening viruses, originated from animal hosts causing emerging diseases; most of them are RNA viruses that thanks to a large population sizes, high mutation rate and short generation time allow rapid evolution, genetic variability and the selection of new variants. A constant and adequate surveillance program and the sharing of different professional expertise are necessaries to follow viral evolution and to formulate efficient public health policy (Howard and Fletcher, 2012).

Influenza A virus is considered one of the most challenging RNA viruses for its zoonotic potential role in the animal-human interface, for global health and economic impact; almost every year influenza epidemics cause morbidity and mortality in the human and is also associated with influenza virus pandemics.

Both wild and domestic birds are considered the primary natural reservoir of influenza A virus and in particular wild birds are thought to be the source of influenza A viruses in all other animals (http://www.cdc.gov/flu/about/viruses/transmission.htm). Different techniques are available to genetically characterize and study viruses in order to understand their behavior, the evolutionary dynamics, the host-virus interactions and their origin; the aim is to develop a valid support with appropriate treatments during the phases of surveillance and diagnosis of possible epidemics.

During my PhD it was used an integrated approach, both genomic and structural, to study the evolution of avian influenza A virus in particular focusing on the hemagglutinin, the major surface glycoprotein, belonging to the H5, H7 and H9 (the major "avian" subtypes responsible for human infection).

Next-generation sequencing (NGS) was used to investigate and characterize the complexity of the viral population to detect low-frequency mutations and to follow the evolution of the genetically related variants present in a viral population. To compare and inspect genetic data, phylogenetic approach has shown to be a useful tools in the analysis of viral evolution. It has been used to explain the molecular epidemiology, transmission and viral evolution. In order to obtain a more complete view of the 'functional evolution', phylogenetic analyses based on sequence comparison and resulting in trees, was integrated taking into account information from structural comparison. Three-dimensional structural approach have shown to be a useful tool to display similarities and to inspect motifs that cannot be discovered analyzing primary sequences alone. Indeed, in the primary sequences the introduction of a mutation does not take into account the effect on the protein folding or on the surface properties, while in the three-dimensional structures, since each mutation is able to influence the structural characteristics and interactions, is directly detectable. This approach has also brought a further contribution to the phylogenetic analysis. In particular the study has focused on the evolutionary dynamics and the adaptive strategies of avian influenza H7N1 and H7N3 subtypes that circulated in Northern Italy for similar periods of time under similar epidemiological conditions. Within and between host population dynamics of Avian HPAI H7N7 viruses, that affected Italy during 2013, were investigated using next generation technology. NGS analysis was used to characterize viral population complexity into two groups of animals challenged with the same virus H5N1 HPAIvirus but vaccinated with vaccine conferring different protection levels. An extensive comparison of structural domains and sub-regions was performed on the hemagglutinin of different subtypes of influenza A virus, with

particular interest to different clades of HPAI H5N1 circulating in Egypt (where bird flu is endemic in poultry), to investigate any domain-specific changes.

Influenza A viruses belonging to H9 subtype were inspected from a phylogenetic and a structural point of view to infer type-specific characteristic and confirm if surface properties could be associated to 'functional evolution' of viral surface determinants as seen in H5N1 subtype. This work suggests that integrating genomic, phylogenetic, and structural comparison can help in understanding the 'functional evolution' of avian influenza A virus.

PREFACE, OUTLINE AND LIST OF MANUSCRIPTS INCLUDED IN THIS THESIS

During my PhD, both genomic and structural approaches to the viral genome and the hemagglutinin (HA) protein have been followed to shed light on - and infer trends in - the evolution and circulation of influenza A viruses. Structural analysis were mainly carried out at the Molecular Biology and Bioinformatics research laboratory, Department of Biology, University of Padua, whereas the genomic and bioinformatics studies were performed at the Research & Innovation Department, Division of Biomedical Science, Istituto Zooprofilattico Sperimentale delle Venezie, Padova, Italy.

The thesis work is presented as follows: an introduction section (with its own references) briefly outlines the characteristic of the influenza A virus and the multiple approaches used to study its evolutionary dynamics, then a short section presents the overall aims of the thesis. Results from each workpackage are presented and discussed as chapters each corresponding to a manuscript, either already published, submitted or going to be submitted for publication. Based on such outline, discussion of results is presented within each individual chapter/article. An aggregate reference list is not presented to avoid redundancy with the reference lists already included in the five manuscripts. Then, a final section presents concluding remarks for the overall thesis work and results, including ongoing research and open perspectives.

For readers convenience, the list of manuscripts included in this thesis is presented hereafter together with short summary of each chapter/article:

CHAPTER 1

manuscript: Fusaro A, Tassoni L, Hughes J, Milani A, Salviato A, Schivo A, Murcia PR, Bonfanti L, Cattoli G, Monne I. Evolutionary trajectories of two distinct avian influenza epidemics: Parallelisms and divergences. Infect Genet Evol. 2015 Aug;34:457-66. doi: 10.1016/j.meegid.2015.05.020

summary: this work describes the comparison of two distinct avian influenza epidemics caused by the H7N1 and H7N3 subtypes that circulated under similar epidemiological conditions. The aim was to study the evolutionary dynamics and the adaptive strategies of distinct avian influenza lineages in response to environmental and host factors, considering the same domestic species reared in the same densely populated poultry area for similar periods of time. The two strains appear to have experienced largely divergent evolution: the H7N1 viruses evolved into a highly pathogenic form, while those from H7N3 subtype did not. A detailed molecular and evolutionary analysis revealed several common features: (i) the independent acquisition of some identical mutations, (ii) the evolution and persistence of two sole genetic groups with similar genetic characteristics; (iii) a comparable pattern of amino acid variability of the HA proteins during the low pathogenic epidemics; and (iv) similar rates of nucleotide substitutions. These findings suggest that the evolutionary trajectories of viruses with originally the same pathogenicity circulating in analogous epidemiological conditions may be similar. In addition, the Next Generation Sequencing (NGS)

analysis revealed parallel mutations already present at the beginning of the two epidemics, suggesting that their fixation may have occurred with different mechanisms, dependent on the fitness gain provided by each mutation. This highlighted the difficulties in predicting the acquisition of mutations possibly correlated to changes in virus virulence.

CHAPTER 2

manuscript: Fusaro A, Tassoni L, Milani A, Hughes J, Salviato A, Murcia PR, Massi P, Bonfanti L, Marangon S, Cattoli G, Monne I. Ultra-Deep Sequencing Data Reveal Unexpected Inter-farm Transmission Dynamics During a Highly Pathogenic Avian Influenza Epidemic. Submitted to Journal of Virology

summary: this work focuses on the study within and between host population dynamics of the highly pathogenic avian influenza H7N7 epidemic, which hadaffected five industrial holdings and a backyard in Italy in 2013. NGS technology was performed on clinical samples to inspect the virus population diversity, the evolution of virus pathogenicity and the pathways of viral inter-farm transmission. This study revealed several viral introductions from multiple sources, genetic heterogeneity of the viruses and a co-circulation of two viral strains with a different amino acid insertion length in the cleavage site of the index case. This work has demonstrated the importance to support epidemiological investigations with genetic data during the control activities so that the transmission dynamics of the viruses and the within and between farms genetic diversity of the viral population during an outbreak may be assessed.

CHAPTER 3

manuscript: Milani A, Fusaro A, Bonfante F, Tassoni L, Salviato A, Mancin M, Mastrorilli E, Hussein A, Hassan M, Cattoli G, Monne I. Vaccine immune pressure influences viral population complexity of avian influenza virus during infection. Going to be submitted by February 2016

summary: this work describes the NGS analyses performed to evaluate the viral population complexity on two groups of animals challenged with the same highly pathogenic avian influenza (HPAI) H5N1 virus but vaccinated with vaccines conferring different protection levels. Previous studies have shown that a suboptimal and/or inadequate vaccine protection with a consequent moderate immune pressure can favour viral spreading and production of heterogenic viral populations in infected animals. Thanks to NGS technologies and assuming that each viral population consists of a mixture of genetically related variants, we managed to characterize viral population diversity, to detect low-frequency mutations and follow their evolution. The results obtained from our preliminary study gave us an overview of the depth of viral genetic diversity, subject to a different immune pressure, that classical methods cannot provide.

CHAPTER 4

manuscript: Righetto I, Milani A, Cattoli G, Filippini F. Comparative structural analysis of haemagglutinin proteins from type A influenza viruses: conserved and variable features. BMC Bioinformatics. 2014 Dec 10;15:363. doi: 10.1186/s12859-014-0363-5.

summary: in this work extensive structural comparison of influenza virus HAs, their domains and subregions was performed to investigate type and/or domain-specific variation. We found that structural closeness and primary sequence similarity are not always tightly related; moreover, type-specific features could be inferred when comparing surface properties of HA subregions, monomers and trimers, in terms of electrostatics and hydropathy. Focusing on H5N1, we found that variation at the receptor binding domain (RBD) surface related to branching of still circulating clades from those ones that are no longer circulating. This work suggests that integrating phylogenetic and serological analyses by an extensive structural comparison can help us understand the 'functional evolution' of viral surface determinants. In particular, variation in electrostatic and hydropathy patches can provide molecular evolution markers: intriguingly, surface charge redistribution characterizing the HA receptor binding domain (RBD) from circulating H5N1 clades 2 and 7 might have contributed to antigenic escape, hence to their evolutionary success and spreading.

CHAPTER 5

manuscript: Milani A^, Heidari A^, Fusaro A, Righetto R, Cattoli G, Monne I, Filippini F. Phylogenetic, phylogeographic and structural bioinformatic approach to the evolution and spreading of H9N2 avian influenza virus. Going to be submitted by February 2016

^A.M. and A.H. contributed equally to this work

summary: in this work, genetic diversity of H9N2 subtype was assessed through large-scale phylogenetic analysis, providing a novel classification scheme of H9N2 classes and clades that is based on both phylogenetic topology and evolutionary distances. Starting from this dataset, viruses representative for each clade were selected to infer type-specific structural features and to confirm whether surface properties could be associated to 'functional evolution' and spreading of H9N2, as observed for H5N1. In particular, variation in the electrostatic properties of HA1 and RBD subregions confirmed evidence from the previous work on H5N1 and unveiled possible fingerprints of the H9N2 evolution. Furthermore, we investigated and compared surface properties of H7 HA proteins belonging to HPAI and LPAI viruses to highlight specific features ascribable to pathogenicity.

INTRODUCTION

Influenza is a global public health disease, caused by RNA viruses belonging to *Orthomyxoviridae* family; antigenic differences in their nucleoprotein (NP) and matrix protein (M1) allow influenza viruses to be classified as types A, B, C or D (Webster et al., 1992; Ducatez et al., 2015). Influenza viruses are characterized by their capability to be highly adaptable, evade the host immune response and infect new host species (Vandegrift et al., 2010); these properties are the result of an error-prone RNA-dependent RNA polymerase, a lack of error correction during replication and a segmented genome.

Influenza type A virus is one of the most important from an epidemiological point of view and it has been involved in recent pandemics and severe epidemics. It is considered one of the most challenging viruses which poses a threat both for human and animal health. Wild birds, like Charadriiformes and Anseriformes, have long been considered a source of influenza A virus capable to infect domestic avian and/or mammal hosts (Webster et al., 1992), furthermore, recent studies in bats have suggested the possible existence of additional reservoir species (Tong et al., 2012; Tong et al., 2013). Influenza A virus is further subtyped considering the antigenic properties of the surface glycoprotein hemagglutinin (HA) and neuraminidase (NA). To date, 16 HA (H1-H16) and 9 NA (N1-N9) subtypes (Fouchier et al., 2005) have been found in wild aquatic birds and only two subtypes (H17N10 and H18N11) have been described in bats (Tong et al.,2013). H5, H7 and H9 influenza A virus subtypes circulating and isolated in avian species have aroused most interest for their role in the human - animal interface; in particular, H5N1, H7N2, H7N3, H7N7 and H9N2 viral subtypes have infected humans even thought a human-to-human host transmission has yet to occurr.

The Avian influenza A virus devastatingly impacted on the poultry industry worldwide, mainly from the late 1990s; the H5N1 panzootic virus, the H7N1 epidemic in Italy (1999-2001), the H7N7 epidemic in the Netherlands (2003) and the H7N3 in Canada (2004) caused huge losses for the poultry industry and aroused serious human health concerns, given the capacity of the viruses to cross the species barrier.

Direct control strategies (i.e. movement restrictions, culling of infected or possibly infected birds) and vaccination are considered the most important tools for the eradication or the containment of the virus in animals. Constant surveillance in the animal reservoirs to monitor viral circulation, evolution and host adaptation becomes pivotal in the study of potential pandemics viruses, such as influenza A virus.

ETIOLOGY, VIRUS GENOME AND PROTEINS

Influenza A viruses are pleomorphic enveloped single stranded negative sense RNA viruses; they can assume spherical/ovoid (80-120 nm diameters) or filamentous shapes (80-120 nm diameters up to 2000 nm in length). Their genome consists of 8 segments (from 0.9 to 2.3 Kb) coding for both structural and non-structural proteins (fig. 1)

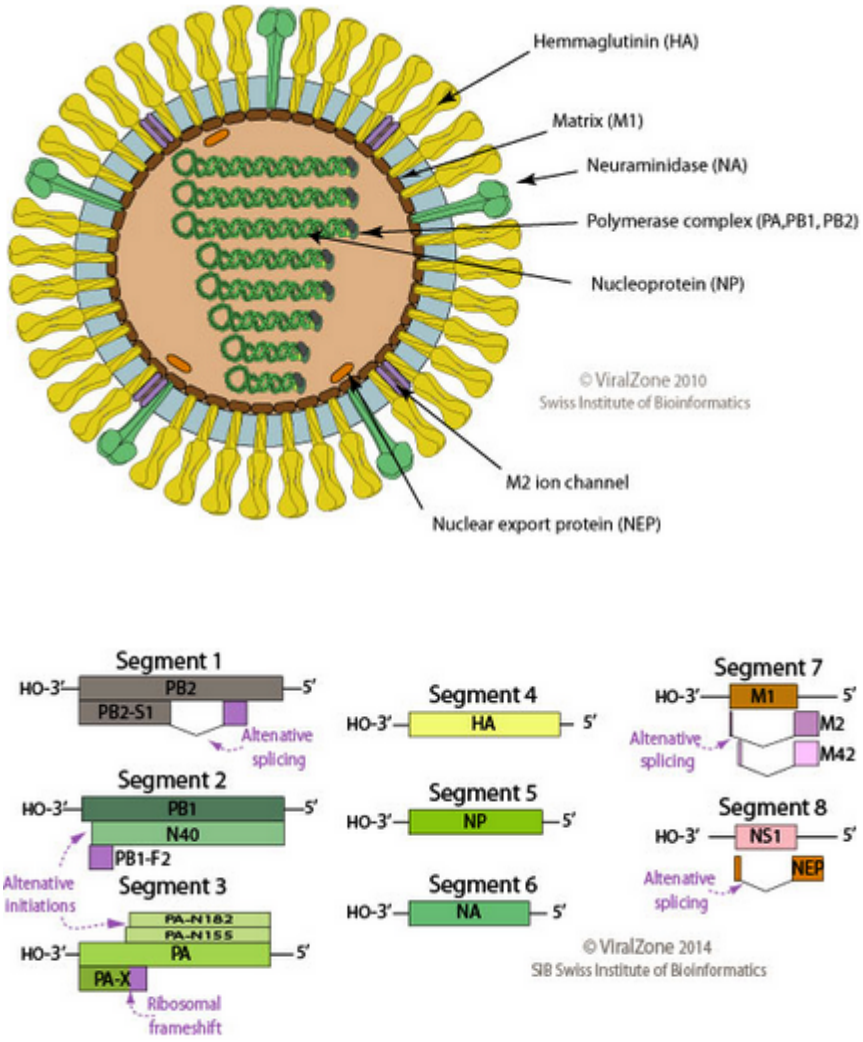


Fig. 1. Graphical representation of spherical shape of influenza A virus and its genome. The ssRNA(-) genome is encapsidated by nucleoprotein, it consists of 8 segments with a size range between 890 and 2341 nt and codes for 12-14 proteins (depending on strains). From ViralZone, SIB Swiss Institute of Bioinformatic (http://viralzone.expasy.org/).

Each gene segment is associated to the ribonucleoprotein complex (vRNP) formed by the nucleoproteins (NP) and the polymerase complex (PA, PB1 and PB2), which is responsible for viral replication and transcription. The complex is then immersed in a protein matrix, consisting of the M1 protein, which is covered by a lipid envelope (the viral membrane) where the transmembrane proteins are inserted: hemagglutinin (HA) and neuraminidase (NA) and the M2 protein. Viral proteins can be divided into three main categories:

- internal: M1, PB1, PB1-F2, PB1-N40, PB2, PA, PA-X, PA-N155, PA-N182, NP
- non-structural: NS1, NS2, N40
- surface: HA, NA, M2, M42

Internal proteins

The M1 protein is a polypeptide of about 28 kDa, abundantly present in the virion, which constitutes the inner viral membrane protein below the lipid bilayer; the layer formed by the M1 represents a bridge between the internal components and the virus surface proteins. It is associated to the RNP and the NS2; probably it interacts with the plasma membrane with HA, NA, and M2 proteins, playing an important role in the budding of virus particles from the host cell. During acidification, through the work done by the M2 ion channel, M1 proteins are separated from the vRPN complex (PA, PB1, PB2, NP), which is then transported into the nucleus of the host cell via a nuclear localization signal present in the nucleoprotein (NP). It also seems that during the late stages of viral replication proteins, the newly synthesized M1 proteins inhibit further viral RNA transcription. It is also the only component sufficient for the formation of vesicles, because it has all the structural information necessary for viral self-assembly, and the morphology, the interaction with the cell membrane and the process of budding; it could also be involved, along with the NS2, in the export of viral RNA from the nucleus (Gomez-Puertas, 2000).

The nucleoprotein (NP) and the polymerase proteins (PA, PB1, PB2) are associated to the genome and are involved in the replication and transcription of the viral RNA. The nucleoprotein consists of about 500 amino acids (56 kDa), the most abundant of which is the arginine, a positively charged aa; the length of viral RNA complexed to each NP molecule was estimated to be in the order of 20 nucleotides. Their basic characteristics are to encapsulate the viral genome for the transcription, replication and packaging and to mediate the transport of newly synthesized viral RNPs from the cytoplasm to the nucleus. NP protein interacts with vRNA, 2 subunits of the vRNA dependent RNA polymerase, the matrix proteins and also with several cellular proteins such as actin, some components of the import/export of the core and with a viral RNA helicase (Portela and Digard, 2002).

The PB1, PB2, PA polymerase proteins have a weight of about 80-90 KDa, form a complex which has an RNA dependent RNA polymerase activity and is located in a small amount into the virion (about 50 units per viral particle). PB2 (polymerase basic 2) is a cap-binding protein essential for the synthesis of mRNA (Plotch et al., 1979); at the beginning of the transcript it recognizes and binds the cap to the 5 'RNA end of the host, and it is partially involved in cap snatching so that it can be used as a primer for the viral RNA synthesis.

PB1 (polymerase basic 1) is the main component of the polymerase complex; it is responsible for the catalysis reaction indispensable for the beginning and the elongation of the newly synthesized viral RNA; it is involved in stretching viral mRNA and vRNA and is localized in the nucleus of infected cells (Samji et al., 2009) The same genetic segment encodes also for PB1-F2 and N40 proteins. PB1-F2 is transcribed starting from an alternative open reading frame (ORF) and is involved in the apoptosis induction (Zamarin et al., 2005); the function of N40 remains unclear (Wise et al, 2009).

The PA (polymerase acidic) protein is involved in viral replication and its expression in infected cells is probably associated to the proteolytic activity. It is involved in the transcription, replication and transport into the nucleus. PA-X is translated from the same genetic segment of PA thanks to a ribosomal frameshift; it is shorter than PA. PA-X is used to repress gene expression of the cell and its loss causes increased apoptosis and inflammation (Jagger et al., 2012). The newly identified proteins, PA-N155 and PA-N182, are N-terminally truncated forms of PA; these proteins were detected in cells infected with various influenza A

viruses isolated from different host species and seems to possess important functions in the replication cycle of influenza A virus (Muramoto et al. 2013).

Non structural proteins

The colinear transcript of viral genetic segment 8 encodes for the NS1 protein (about 26 kDa) while the spliced mRNA for the protein NS2 (11 kDa).

NS1 is a poly (A) binding protein expressed in huge quantities both in the cytoplasm and the nucleus of infected cells, but has not been detected in the mature virus particle. In the nucleus it appears to inhibit various processing stages, such as cellular mRNA polyadenylation, splicing and transport mechanisms of RNA; in the cytoplasm it appears to increase the rate of viral mRNA translation. Its main functions are export inhibition of the RNA from the nucleus, block of dsRNA counteracting the production of IFN-beta by the host cell and splicing inhibition of pre-mRNA (Lu et al., 1994; Qiu and Krug, 1994).

NS2 is also called nuclear export protein (NEP) because it carries the newly synthesized RNPs from the nucleus to the cytoplasm, essential prerequisite for viral budding (Robb et al., 2010). This protein probably has a nuclear localization signal. It has been demonstrated that proteins NS2 and M1 interact in vitro: probably the RNP nuclear export of the molecules is facilitated by the association of these proteins.

Surface proteins

The membrane of influenza A virus contains a transmembrane protein, the M2; which forms a proton channel of 97 amino acids and consists of an extracellular N-terminal portion, a transmembrane segment and a C-terminal intracellular part. It is essential for viral replication and its main function is to equilibrate the pH during virus entry into the host cell. After the viral infection and before the membrane fusion, the M2 channel is activated by the low pH endosome, protons pass through the viral envelope and the interior of the virus starts to acidify. This step triggers membrane fusion, release of uncoated RNPS into the cytosol and later into the nucleus, where mRNA and vRNA are synthesized. M2 carries out another function by preventing the Golgi pH to decrease too much and therefore avoiding the potential conformational changing of the hemagglutinin during its transport to the viral envelope (Pielak and Chou , 2010). Mutations to obtain an alternative splicing or the introduction of a stop codons allowed the generation of defective highly attenuated M2 viruses. This novel M2-related protein, M42, showed differences in its cell localizationmostly accumulated in the Golgi apparatus (Wise et al., 2012).

The surface glycoproteins are the antigens mostly involved in the induction of a protective humoral and cell mediated immune response; they are the most abundant viral proteins for diagnostics and vaccine prophylaxis. After the identification of antigenic differences among the surface proteins, it was possible to classify 16 subtypes of hemagglutinin and 9 of neuraminidase. Up to now all possible combinations of HA and NA have been isolated from several avian species and only two subtypes (H17N10 and H18N11) described in bats (Tong et al.,2013): this indicates the extreme antigenic variability characteristic of these viruses.

The neuraminidase (NA), an integral membrane glycoprotein is a homotetramer of 220 kDa consisting of a head, which is an enzymatically active domain, and a tail which allows to anchor the protein to the membrane (Hausmann et al., 1997). Its main functions are i) the hydrolysis of sialic acid on the cellular receptor for the hemagglutinin, which allows the release of the virus from the cell surface, and ii) the removal

of sialic acid residues from the viral particles to prevent aggregation. The inhibitors developed for this enzymatic protein are the main antiviral drugs. In the absence of this enzyme the virus remains attached to the cellular receptors, thus inhibiting the spread of the virus progenies in the host tissues. Therefore, antibodies directed against the neuraminidase protein do not prevent infection but reduce the spread of the virus in tissues.

The hemagglutinin (HA) is the most abundant antigenic glycoprotein of the viral surface encoded by the fourth viral segment of the influenza A genome. It is a type I membrane glycoprotein responsible for the binding of the virus to the receptors present in the host cell surface, for virus internalization and fusion with the endosomal host membrane. In the viral particles, each mature HA is a homotrimer which projects onto the viral envelope to form a rod-shaped structure; in infected cells, this protein is synthesized as a precursor polypeptide, called HA0, which must be cleaved into two subunits (HA1 and HA2, of 36 and 27 kDa, respectively) by host trypsin-like proteases (Copeland et al., 1986). After the proteolitic cleavage of the precursor, the two subunits are covalently linked to each other by a disulfide bond, whereas each trimer is associated to the others by non covalent bonds (Klenk et al., 1975). This processing is necessary for the virus infectivity because it activates the fusion of HA and it is a determinant of pathogenicity (Hamilton et al., 2012). The cleavage site is a prominent surface loop near a cavity in HA0 (Chen et al., 1998); the results of the cleavage process is a structural rearrangement in which the fusion peptide, formed by nonpolar N-terminus amino acids of HA2,is relocated into the interior of the trimer and buries ionizable residues involved in the acidification-induced conformational changes in the endosome. At the membrane-distal tip of each HA1 we find the receptor binding domain (RBD), formed by the 130-loop, 190-helix, and 220-loop and four highly conserved sites (Weis et al., 1988, Martín et al., 1998); it forms the sialic acid binding pocket and contains most of the antigenic regions recognized by neutralizing Abs.The stem-like structure HA2 has a membrane fusion activity. The 2 loops and the helix all contain amino acids that interact either with sialic acids or with internal sugars of the glycan chain associated with glycoproteins and glycolipids on the surface of epithelial cells; the base of the site contains several highly conserved residues that form an extensive hydrogen bond network . The three-dimensional structure of few HA subtypes has been resolved and characterized with respect to the localization and structure of their antigenic sites. In the H3 subtype five antigenic sites (A, B, C, D, E) have been mapped (Wiley et al., 1991), and the same structure was used to map antigenic sites on the H1 and H2. The H1N1 subtype shows five antigenic immunodominant antigenic sites (Sa, Sb, Ca1, Ca2, Cb) comparable to the H3N2 virus (Caton et al.,1982). When the 3D structures of H5 and H9 HA were resolved, the H5 HA molecule was antigenically mapped. For H5 structure, the localization of two antigenic sites has been described. Site 1 includes residues 140 to 145 in HA1 (H3 numbering), which corresponds to antigenic sites A of H3 and Ca2 of H1. Site 2 comprises one site on residues 156 and 157 in HA1, corresponding to site B in the H3 subtype, and one from residue 129 to 133 in HA1 that corresponds to site Sa in the H1 subtype (Peng et al., 2014). Changes in the HA and NA antigenic combinations of a virus (antigenic shift) may derive from the genomics segmentation. When the same cell is co-infected by distinct viral subtypes, the viral progeny may originate from a reassortment of parental genes from different viruses.

Infection and replication

The influenza A virus, takes advantage of the host cell machinery in almost all the phases of its replication. Influenza A virus attachment and entry starts with the binding of viral hemagglutinin to the N-acetylneuraminic (sialic) acid present on the host cellular surface. After binding to the host cell, the virus is internalized through the *endosomal* pathway; the acidification of the vesicle causes a conformational change of the hemagglutinin, which exposes the N-terminal hydrophobic domain of the HA2 chain (site of fusion) and causes its interaction with the endosomal membrane. The viral envelope and the endosomal membrane are then merged and the viral nucleocapsid is released into the cytoplasm of the host cell.

The virus is rapidly internalized into clathrin coated vesicles and begins endocytic internalization; starting from the early endosomes (located under the plasma membrane) to the later ones (close to the Golgi apparatus and the nucleus) the pH decreases to a value of about 5.5. At this pH value, the process of viral fusion is activated instead of being transferred from the endosomes to lysosomes, where it would be degraded : the virus escapes by virtue of the properties of the hemagglutinin. The HA glycoprotein undergoes a conformational change to form α supercoiled helices and exposes the hydrophobic fusion peptide HA2 that inserts into the endosomal membrane. The endosome acidification has another crucial function in the entry of the virus into the cell: the M2 ionic channel present in the viral envelope allows the viral components, located inside the virus (M1 and vRNPs), to be exposed to the low pH of the endosome. This latter is a necessary prerequisite for breaking the interactions between M1-vRNP and removes the coat proteins from the virus. Ribonucleoproteins (RNPs), once released into the cytoplasm, are transported into the nucleus of the host cell (through the nuclear pore complex) thanks to the presence of a localization signal.

In the host cell nucleus, the viral RNP is used as a template by the RNA polymerase complex (PB1, PB2, PA) to produce two different types of single-stranded positive RNA segments: cRNA (complementary RNA) used by RNA polymerase to produce more copies of vRNA (viral RNA) and mRNA (messenger RNA). The synthesis of viral RNA always starts at the 5' end of the new molecule of viral RNA and proceeds in the 5'-3' direction up to the 3 'end. There is no mechanism for error correction during RNA synthesis and error frequencies are similar to those of DNA transcription (1 error every 104 nt synthesized). Although both in cellular and viral mRNA cap structures at the 5' end (necessary for the ribosome attack) and poly (A) tail at the 3' end (necessary as they protect the mRNA from degradation by cytoplasmic ribonuclease) are present,the acquisition mechanisms of these sequences are different. The cap structures at the 5' end of cellular mRNA are synthesized *de novo* by cellular enzymes (Shuman, S. 1995), while the cap structures at the 5' end of influenza viruses are obtained from the fragmentation of cellular pre-mRNA during viral mRNA synthesis. An endonuclease intrinsic to the viral polymerase cuts the cellular capped pre-mRNA to produce fragments of 10-13 nt with a capped 5' end which are used as primers by viral RNA polymerase for the viral mRNA synthesis (this procedure is required because the viral RNA polymerase doesn't have a catalytic activity to produce capped primers). The poly(A) tail at the 3' end of the viral mRNA is synthesized by viral polymerase.

In Influenza virus infected cells, the nuclear export of cellular mRNAs is blocked; the cellular pre-mRNA and mRNA are degraded in the nucleus. This nuclear export block is selective: all viral mRNAs are efficiently exported; as a consequence, these newly synthesized mRNAs prevail in reaching the translation machine into the cytoplasm of infected cells, thus helping to permanently stop the gene expression of the

host cell and selectively synthesize viral proteins. The nuclear export of the viral mRNA uses the host cell machinery but it is selective: it is controlled by a viral non-structural protein (NS1) (Gary R. Whittaker, 2001) that inhibits the expression of mRNA synthesized after the cell infection. Furthermore, the pre-mRNA and mRNA retained in the cell nucleus are available to the cap dependent viral endonuclease for the production of the capped RNA primers required for viral mRNA synthesis. Viral mRNAs, after being transported into the cytoplasm, are translated to produce the corresponding proteins. Membrane proteins (HA, NA and M2) are transported across the RER and the Golgi apparatus to the plasma membrane. When nucleocapsid and viral coat proteins reach the plasma membrane, they form a bud whose coating contains coating proteins immersed in the lipids bilayer of the host cell. Subsequently the bud separates and virus particles are released outside the cell. During viral budding the host proteins present in the plasma membrane are excluded from the final viral particle. The viral proteins possessing a nuclear localization signal are transported into the nucleus (PB1, PB2, PA, NP, M1, NS1 and NS2). (Gary R. Whittaker 2001; Palese, Garcia-Sastre 1999).

Regulation of gene expression in virus infected cells

The influenza virus infection can be divided into an early and a late gene expression; specific vRNA, viral mRNA and viral proteins are synthesized during the early phase of synthesis. After a first transcription, an RNA template is synthesized starting from a parental vRNA. Subsequently specific RNA templates are selectively transcribed in vRNA; in particular, RNA molecules encoding for protein NS and NP are a priority, while the RNA synthesis of matrix proteins (M) is delayed. The NP protein is synthesized in this first phase probably because it is necessary for the synthesis of vRNA and RNA templates; the NS1 protein is required for the functions previously described. The synthesis of the M1 protein is delayed because this protein permanently interrupts the viral RNA transcription in the corresponding mRNA and it is also involved in the transport of nucleocapsids containing vRNA from the nucleus to the cytoplasm. The rate of synthesis of a particular vRNA is correlated with the rate of synthesis of the corresponding mRNA and of the corresponding protein. During the final phase, the relation between vRNA and mRNA viral synthesis and related proteins changes drastically. In this stage the synthesis of all viral mRNA reaches its highest rate; also the pattern of the proteins synthesized during the first phase differs from the second one, where M1 and HA (Lamb, Krug 2001) are the two most synthesized proteins

EVOLUTION

Antigenic drift and shift

As for all RNA viruses, the evolutionary dynamics of influenza A virus are complex and are the results of the combination of high mutation rate, rapid replication and infection of large population size. Two are the mechanisms that allow influenza viruses to rapidly evolve and adapt: the high mutation rate and the reassortment of the segmented genome.

The lack of a proofreading mechanism during viral replication results in a high mutation rate that provides the opportunity to change; many changes may be deleterious and can be lost during the selection process so as to maintain the fittest virus in a population. The error-prone RNA polymerases produce complex populations

of genetically related but non-identical variants called quasi-species, which interact and cooperatively contribute to characterize the whole population, and are subjected to a continuous genetic variation, competition among variants, and selection in a given environment (Domingo et al., 2012). This mechanism can provide a broad range of viral subpopulations able to adapt to the action of multiple factors; in each viral population minority variants are always present and can be selected out of the majority population as a consequence of a given environmental pressure . As an example, subpopulations of viruses already resistant to antiviral drugs, previously used for the treatment of influenza infections, have been found in circulating avian influenza viruses belonging to the H5 subtype (Wainright et al., 1991). From a practical point of view, the consensus sequence obtained with Sanger sequencing corresponds to the most represented nucleotide at each genomic position; consensus sequence may or may not exist in the population. The genome characterization of a viral population is a great challenge; recent next generation technologies have allowed deep sequencing and investigation of the genetic composition, even if sometimes the identification of true variants from sequencing error is still a real problem. The eight genetic segments have a similar mutation rate, although HA and NA genes seem to have more changes because of positive selection. When unpredictable point mutations cause minor gradual variations in the two main surface glycoproteins, neuraminidase and hemagglutinin, new virus strains with differerent antigenic properties may be produced. This results in a decreased power of antibody binding that will reduce any possible acquired immunity of the host and facilitate the spread of the epidemic. Antibody pressure against the hemagglutinin in previously immunized or vaccinated hosts is considered one of the major selective factors (Plotkin and Dushoff, 2003). Five antigenic regions have been identified on the globular head of the human H3 protein close to the receptor binding site; antibodies against these antigenic sites can have a neutralizing effect blocking the access to the receptor binding site and preventing the virus from binding to host receptors and infect host cells (Webster and Laver, 1980). Amino acid changes are tolerated in these regions and when they occur, neutralizing epitopes may be modified so that this viral escape mutant may be able to escape the host's immune response, with increased replication possibilities and transmission. The antibodies to the HA molecule mainly belong to the IgG and IgA classes; they are able to neutralize the viral infectivity and are the major causes for resistance to infection (which means that they constitute the basis of vaccination against viral strains). The response of Ig to the hemagglutinin is subtype specific, but the accumulation of point mutations (*antigenic drift*) allows infectious viruses to escape from antibody-mediated destruction. Vaccination stimulates the production of neutralizing antibodies and is presently considered the most effective prophylactic measure against influenza virus; several studies have shown that Ab binding in proximity to the receptor-binding domain can block virus attachment to the sialic receptors on host cells.

Viral recombination during evolution is another aspect to take into consideration due to its abilityo generate new strains, which may have new acquired properties and enhanced virulence (Andino and Domingo, 2015). Influenza A viruses have a segmented genome that allows gene rearrangement during infection; the new recombinant viruses may be characterized by high infectivity and virulence when transmitted between individuals of the same species and are also able to cause pandemics. A drastic reassortment of sections of the viral segmented genome is referred to as *antigenic shift.* As a matter of fact, a consequence of reassortment could be the production of novel influenza A virus subtypes, containing some genes from strains that normally infect birds and some others from strains which normally infect humans: thismay cause influenza pandemics in humans, as observed in the case of the 1957 and 1968

outbreaks (Clancy et al., 2008). Today the World Health Organization (WHO) has included H5, H7 and H9 avian influenza subtypes as those with the greatest pandemic potential. The spread of HPAI and LPAI viruses of the H5, H7 or H9 subtypes amongst birds and sporadic infections in humans continue to pose a threat to public health (Lin et al., 2000).

Avian influenza and pathotypes

Avian influenza viruses can be classified into two groups based on their difference in virulence: low pathogenic (LPAI) or highly pathogenic (HPAI) form. To date, only H5 and H7 subtypes of influenza A viruses have shown to be able to evolve from a LPAI to a HPAI form after the introduction into poultry from the wild bird reservoir; however, there are rare examples of other viruses that could technically be considered HPAI. LPAI viruses in poultry can be asymptomatic or cause mucosal infections, with mild to severe respiratory diseases, water and feed decrease and drops in egg production; usually they do not result in high mortality of the infected hosts. In poultry and domestic birds, HPAI viruses usually cause systemic infections and are associated with severe disease and high mortality; these viruses do not usually cause illness or death in wild birds. The two pathotypes posses a different ability to cause disease in intravenously inoculated experimentally infected young chickens. From a molecular point of view, H5 and H7 highly pathogenic viruses contain a polybasic sequence at the cleavage site allowing intracellular cleavage by ubiquitous, subtilisin-like serine endoproteases, such as furin (Garten et al., 2008), which causesa systemic infection. Low pathogenic influenza viruses cause an anatomically localized infection in the hosts as a consequence of the restricted range of extracellular trypsin-like proteases, which can recognize and cleave the cleavage site where the linker consists of a single R/K.

The released fusion peptide obtained by the cleavage of the HA glycoprotein is mandatory for the initiation of viral infection. The low pathogenicity H5 subtype usually hasa well conserved hemagglutinin cleavage site, even though some exceptions are allowed, in position 321-330 with the amino acid sequence: . ..QRETR/GLFG . . .; the cleavage point is situated between the R (arginine) and G (glycine). In most low pathogenicity isolates of the H7 subtype the protein cleavage site consists of a 11 amino acids conserved region with the sequence ...PEXPKXR/GLFG... where X can be a neutral or a basic amino acid (Perdue et al., 1997). In bird hosts, the proteases able to cleave the hemagglutinin of the LPAI subtype are mainly situated in the respiratory and intestinal tracts, but enzymes involved in cleavage have not been fully identified. Mutations and/or insertions of amino acids within the cleavage sites ,resulting in an increased number of basic amino acids (R and/or K), allow the recognition and cleavage by widely distributed furin-like or subtilisin-like endoproteases (Garten et al., 1981; Horimoto and Kawaoka, 1995) . This is the reason why highly pathogenic viruses are able to replicate in a wider range of host tissues with broader replication possibilities.

Up to now four are the mechanisms by which AI viruses acquire basic amino acids at the cleavage site: single nucleotide changes, accumulated nucleotide insertion; tandem duplications of stretches of purines and RNA/RNA recombination events.

Single Site Mutations can occur naturally in RNA viruses, without affecting viral fitness and becoming fixed in the population; an example is the H5N2 outbreak that affected the U.S. poultry industry in 1983, resulting in the culling of 17 million birds and the loss of $63 milllion. (Kawaoka and Webster, 1985; Webster et al., 1986)

Accumulated nucleotide insertions can result in a functional codon when three successive nucleotide are added; the HPAI outbreak in turkeys in Ontario in 1966 and several H7 turkey virus isolates in England before 1963 had probably originated from this mechanism. Another example assessing the ability of LPAI to evolve into HPAI form are the two H7 avian outbreaks that affected Northern Italy between 1999 and 2001. The addition of 12 nucleotides resulted in a longer cleavage site with multiple basic amino acids, although it still remains unclear how such an insertion had occurred and in what way the RNA/RNA recombination event happened. Epidemiological information, phylogenetic analysis, and deep sequencing approaches revealed the evolution of LPAI to HPAI pathotype and a common ancestor among strains (Monne et al., 2014).

An example of tandem duplications of stretches of purines at the cleavage site is the Mexican outbreak of avian influenza in 1994-1995, where in a few weeks viruses evolved into the HPAI pathotype (Garcia et al., 1996). Molecular analysis at the cleavage site of isolated samples showed one site mutation and successive tandem duplications of the sequence AAAGAA, resulting in 6 amino acid insertion (R-K-R-K-R-K). In many cases, H5 and H7 subtypes with multiple basic amino acid insertions at the cleavage site of highly pathogenic strains may be the result of tandem duplication events (Perdue et al., 1997).

Another example is the characterization of an HPAI H7 subtype isolated from seal showingan insertion of 60 nucleotide at the cleavage site, probably originated from the nucleoprotein gene of the same strain. The outbreak of H7N3 that affected poultry in Chile in May 2002, demonstrated the lengthening with 30-nucleotide inserts of the hemagglutinin cleavage site by the recombination with the nucleoprotein RNA of the same virus, which lead to the evolution of the LPAI pathotype into a HPAI one. The same scenario happened in an H7N3 AI outbreak in British Columbia: an insertion of 21 nucleotide at the protein cleavage site, probably from the M gene, resulted in a highly pathogenic strain.

Because of the risk of a H5 or H7 virus of low pathogenicity becoming highly pathogenic by mutation, these subtypes and all high pathogenicity viruses from birds are notifiable to the Office International des Epizooties (OIE). H7 LPAI virus usually causes mild respiratory disease and a production decrease in infected poultry; its evolution into a HPAI form results in the generation of a virus able to cause severe disease and death in the poultry population.
(http://www.oie.int/fileadmin/Home/eng/Health_standards/tahm/2.03.04_AI.pdf).

Despite not possessing a multi-basic cleavage site and their inability to cause systemic infections, some viruses belonging to the H10 subtypes have characteristics of HPAI pathotype. When these viruses are administered intravenously into poultry, they cause the death of the host by impairing the function of the kidney (Swayne and Alexander, 1994). In a previously study (Bonfante et al., 2014) a nephritic H10N1 avian influenza virus was genetically and phenotipically characterized; it did not display multiple basic amino acids at the cleavage site although it did show an intravenous pathogenicity index (IVIP) of 1.9. Furthermore, when administered by a natural route (intranasal), this H10N1 virus could cause mortality.

HOST SPECIFICITY

To date, the factors determining the viral to host restriction are probably associated to several molecular determinants within the viral genome that involve either the viral genes encoding for surface glycoproteins (HA and NA) or genes encoding for internal proteins such as the NP and the PB2 genes.

Hemagglutinin membrane glycoprotein plays a critical role in influenza A virus infection (Gamblin et al. 2010). Binding preference of viruses to host cells is considered to be one of the most relevant determinant that prevents crossing the species barrier and influence viral tropism. Sialic acids (SA) are present on the surface of many cell types and animal species; in general, carbon-2 of the terminal sialic acid is bound to the carbon-3 or carbon-6 of an adjacent galactose, which allows the formation of an α-2,3 or α2,6-linkages configuration. The hemagglutinin of avian and equine influenza viruses usually has a binding preference for α2,3 linked sialic acid; human isolates exhibit α2,6 linkage whereas viruses from swine bind to both. More recently it has been demonstrated that the human epithelial cells of the lower respiratory tract harbour SA with both α2,6 and α-2,3 linkages; furthermore, the finding of α-2,3 linked SA in the human airway epithelium can explain the ability of viruses of avian origin to infect and replicate in humans, although not sufficiently enough to inducean efficient human to human virus transmission (Matrosovich et al., 1999; Matrosovich et al., 2004). Neuraminidase is also involved in the interaction with host receptors; in particular, neuraminidase of avian origin prefers to hydrolyze the α-2,3 linked SA rather than the α-2,6. The equilibrium between the two main surface glycoproteins HA and NA must be balanced to allow a perfect enzymatic activity and a functional viral replication (Wagner et al., 2002).

Occasionally, infection of humans and other mammals by avian influenza viruses may occur (Capua and Alexander, 2008), which proves that the species barrier is not insurmountable. Some avian or swine influenza A virus subtypes (e.g. swine H1, swine H3, H5, H6, H7, H9 and H10) have occasionally infected humans, without however establishing in humans. In 1997 in Hong Kong the first detection event of an H5N1 avian influenza virus able to infect humans causing serious disease was reported. Six years later, in the same location, an H5N1 HPAI virus of avian origin infected humans again (two cases were reported) (Peiris et al., 2004). Subsequently, the same subtype started to circulate in the avian population of South East China and in wild and domestic birds throughout Asia, Europe and Africa. The number of human cases infected by HPAI H5N1 influenza A virus steadily increased over the years, mainly due to a direct or indirect contact with infected poultry (de Graaf and Fouchier, 2014). Although human-to-human transmission of avian influenza viruses has not been established yet, monitoring the evolution of circulating viruses is an issue of great importance. Occasionally, viruses of the H7 subtype can also cross the species barrier. Subtypes H7N2, H7N3 and H7N7 were indeed detected in humans, where they mainly caused mild symptoms such as conjunctivitis. On the contrary, a more severe illness associated to severe pulmonary and acute respiratory symptoms and caused by the H7N9 avian influenza virus infection was reported in China (Gao HN et al., 2013; Gao R et al.,2013). Previous studies have shown that the H7N9 subtype is probably the result of several reassortment events involving H9N2 and subtypes H7 and N9, all of avian origin (Lam et al., 2013). The avian H9 influenza virus subtype can also infect humans and cause a mild influenza-like illness (Butt et al., 2010).

DIAGNOSTIC METHODS FOR AI

Diagnosis of AI virus infection requires laboratory testing, such as virus isolation in embryonated chicken eggs or the detection of viral nucleic acid, viral protein, or antibodies against AI virus. Diagnostic tests can identify any type of influenza A viruses or can be subtype specific; in the latter case the main

targets are the H5, H7 and H9 hemagglutinin subtypes for their potential of being highly pathogenic in domestic poultry. The application of molecular methods, such as RT-PCR and Real-time RT-PCR, has become an important tool for the rapid detection and typing of AI viruses. Advantages are their high sensitivity and high specificity, as well as their ability to analyse a wide range of sample type, to process inactivated viruses and to rapidly providing results. Further characterization may be required; these may include the chicken pathogenicity test to assess virulence, sequencing the hemagglutinin cleavage site to differentiate LP from HP pathotypes, whole genome sequencing and phylogenetic analysis. Currently there exist manuals, such as the the OIE Terrestrial Animal Health Code (Terrestrial Code), which provide a detailed definition on the HPAI of the H5 and H7 subtypes; furthermore, there are reviews and texts available from the OIE, FAO and WHO describing methods, recommendations and laboratory procedures for diagnosis and detection of avian influenza (OIE, 2004, http://www.oie.int/fileadmin/Home/fr/Health_standards/tahm/2.03.04_AI.pdf; WHO, 2007, http://www.who.int/influenza/resources/documents/RecAIlabtestsAug07.pdf; FAO, 2004, http://www.fao.org/docs/eims/upload/200354/hpai_manual.pdf)

CONTROL STRATEGIES

Strategies developed to control avian influenza must take into consideration several factors, i.e.virus circulation in the country, pathogenicity, subtype, species of birds at risk or infected, type of ecosystem, availability of appropriate financial resources, veterinary medical infrastructure. The presence of LPAI viruses in wild aquatic birds seems to be part of ecosystems worldwide, thus the control of LPAI should be addressed to prevent the virus introduction from wild bird to domestic hosts. The major sources of bird-to-human infection are caused by the direct handling of infected animals with HPAI viruses, or else by indirect contact with contaminated environments through the respiratory and gastrointestinal tract or the conjunctiva.
International organization, such as the WHO, the OIE and the FAO have shared their competences with a network of expertise and have made great efforts to establish and provide guidelines to face and combat H5N1 epidemics (www.offlu.net). To date, the evolution from a LPAI to an HPAI phenotype has interested only H5 and H7 influenza A viruses; it is important for LPAI viruses to be managed in an appropriate manner and for official veterinary services to adopt suitable measures (Capua and Alexander, 2006). Biosecurity is the first,most important and efficient step to limit and prevent a further spread of the disease into the domestic poultry (Capua and Alexander, 2006). Biosecurity involves some basic steps aimed at improving physical barriers and/or working activities, with particular attention to all the possible preventative measures which should be implemented to to avoid or limit viral introduction and spread in susceptible avian species. Bioexclusion measures are employed to prevent direct or indirect contactto contrast first introduction, and if necessary they may be reinforced, as in the case of an outbreak. Particular attention must be addressed to preventive measures: personnel in direct contact whith the animals must wear appropriate personal protective equipment (PPE), change clothes, keep clean and disinfect the contaminated areas; furthermore, it is important to clean and disinfect materials and vehicles entering or leaving infected farms.Restriction on movement and isolation of the infected or exposed birds culling of infected animals, disposal of carcases,

eggs and any contaminated material are the measures presently implemented (Capua and Alexander, 2009). A capillary and efficient activity of veterinary services, the work carried out by diagnostic laboratories, restriction policies, biosecurity, and stamping out policies are only some of the the valid tools which have contributed to cope against the influenza threat (Capua and Cattoli, 2013).

VACCINATION

Vaccination and depopulation are important tools that can be used to control influenza viruses; depopulation measures are used when an epidemics of HPAI influenza virus occurs in poultry, whereas vaccination is a preventative approach. Up to now there have been two different types of available vaccines: inactivated and live recombinant ones. Inactivated vaccines contain a killed field isolate or a virus obtained by reverse genetics techniques. To date, live recombinant vaccines are vector-based vaccines expressing the major virus protecting-antigen, the hemagglutinin protein (HA) of avian influenza viruses. It is essential forvaccines to be cross-protective in order to reduce clinical signs, mortalityand viral replication and to increase the resistance to challenge with infectious virus. Furthermore, vaccination should be properly adopted within a well defined control strategy. Countries like China, Vietnam, Pakistan, Indonesia, Bangladesh and Egypt have already applied or are currently performing poultry vaccination campaigns against H5N1 viruses, which have become enzootic in these areas (Webster et al., 2006).

GENETICS AND BIOINFORMATICS APPROACHES TO STUDY VIRAL EVOLUTION

Different techniques are available to genetically characterize and study viruses. One of the most widely used method is the classical Sanger sequencing, which allows to generate a consensus sequence starting from an amplified target (a specific gene or the whole genome). The limitations of this method include the information given by the consensus sequence, which is a summary of the most abundant variants present in the heterogenic viral population but unable to detect the low abundant ones, the so-called minority variants. Until recent years the only available method to isolate and amplify minority variants was the biological clone, which was howevertime consuming, at times inefficient and expensive (Vignuzzi et al. 2006). Next-generation sequencing (NGS) has revolutionized research byperforming parallel sequencing massively, during which process millions of DNA fragments from single or multiple samples are sequenced simultaneously. Results are high speed, throughput and have alow cost per sequenced base, in addition, NGS also offers the possibility to complete analyses in a few weeksrather than in years, as it would happen by using the first generation sequencing. The advantages of this technology is the possibility to obtain DNA sequences by amplifying fragments, without therefore the need of cloning and sequencing entire genomes extremely rapidly. This approach is used to investigate and characterize the complexity of the viral population to detect low-frequency mutations and to follow the evolution of the genetically related variants present in a viral population. The phylogenetic approach has proved to be a useful tool to compare and inspect genetic data in viral evolution analyses. It has been used to explore the molecular epidemiology, transmission and evolution of different viruses such as HIV, SARS, CoV and, more recently, the evolving

epidemiology of both avian and human influenza viruses. Advances of phylogenetic methods have led to the identification of recombinant and reassorted viruses, their origin and the spread dynamics within a specific population, geographical region, period of time and hosts (Lam et al. 2010). In order to obtain a more complete view of the 'functional evolution', phylogenetic analyses based on sequence comparison and resulting in trees might be integrated taking into account information from structural comparison. A three-dimensional structural approach could be useful to display similarities and to inspect motifs that cannot be discovered by analyzing sequences alone and help to infer phylogeny data(Ravantti et al. 2013). During my PhD, genomic and structural approaches have been used to study the hemagglutinin (HA) protein during the evolution of the influenza A virus; in particular, I focused the attention on H5, H7 and H9 avian influenza virus A subtypes.

SEQUENCING

Automated Sanger sequencing

Sanger sequencing was developed by Edward Sanger in 1975 (Sanger et al. 1977) and was considered the gold standard nucleic acid sequencing method and the most widely used until the more recentdeep sequencing approach. In Sanger sequencing oligonucleotide primers anneal to a target denatured DNA molecule and later a DNA polymerase starts the extension step by incorporating nucleotide triphosphates. A mixture of deoxynucleotide triphosphates (dNTPs) and chain-terminating dideoxynucleotide triphosphates (ddNTPs), lacking the hydroxyl group on the 3' carbon, is used. Without the 3' OH, no more nucleotides can be added and DNA polymerase irreversibly stops the incorporation of nucleotide in the new chain. The newly synthesized DNA chains will turn out to be a mixture of lengths as a result of randomly incorporated ddNTPs. Each dideoxynucleotide is labelled with a dye with a different emission spectra; all four dye-labelled terminators when excited by an argon ion laser at 488nm produce a peak emission that could be distinguished by the detector; thus, the sequencing reaction can be carried out in a single reaction tube and prepared for loading once the reaction reagents have been filtered out. The capillary system is set up a) to deliver a new polymer to the capillary, b) to load the sequencing reaction into it, c) to apply a constant electrical current through the capillary and d) to have the resolved fragments migrate past an optical window where a laser excites the dye terminator. A detector collects the fluorescence emission wavelengths, and software would interpret the emission wavelengths as nucleotides (França et al. 2002).

Next generation sequencing

NGS technologies can be grouped into two different categories: second generation sequencing, where a DNA synthesis chemistry is used as in the traditional Sanger's sequencing, and third generation sequencing (single molecule sequencing), which does not require amplification of the template molecules prior sequencing reaction. NGS technologies involve a biological side, which includes template preparation and sequencing procedures, as well as bioinformatics tools to process and analyze post sequencing data output. The key steps are the same as those implemented in the different Second-generation sequencing (SGS) technologies currently available: preparation and amplification of template DNA, anchoring of templates on a support, sequencing, imaging, base calling, quality control and analysis of the output data.

During pre-sequencing steps, NGS libraries are prepared by fragmenting the DNA (or cDNA) sample using physical, chemical or enzymatic methods, and ligating primers adaptors (synthetic oligonucleotides of a known sequence) onto the ends of the DNA fragments. Once constructed, libraries are clonally amplified for sequencing; after the sequencing reaction is performed, billions of reads are generated. In the post-sequencing bioinformatics analysis sequence images are processed to generate base sequences, and converted to readable files (fastq); sequences are mapped on a reference sequence, variants identified and and annotated (Buermans et al. 2014).

Illumina technology

In Illumina technology all the enzymatic processes, bridge amplification for polony generation, sequencing and imaging steps take place in a flow cell. Miseq is a second generation sequencing platform based on sequencing-by-synthesis (SBS) reaction; RNA and DNA samples must be in sufficient quantity and of high-quality to obtain excellent sequencing results.

During the first step, DNA is fragmented using an enzymatic (or mechanic) method and index-adapters are ligated to the fragments obtained; these oligos are necessary, both to capture the template DNA in the solid surface and as primers for subsequent amplification. After size selection, purification step and quantity normalization, libraries are ready to be processed into the Illumina platform. Bridge amplification occurs after the single-stranded sequences anchor to a flow-cell pre-coated with two different types of oligos, randomly distributed, complementary to the adapters; the amplification by PCR bridge is needed to make the signal strong enough to be captured by the CCD Camera. Optimal cluster density on the flow-cell is approximately 800–1000 K clusters per mm2 for an Illumina Miseq V2 (250x2) whereas 1000-1400 K for Illumina Miseq V3 (300x2); yield are influenced by library concentration and molecules length. After amplified molecules have been generated, thanks to successive rounds of PCR, a sequencing-by-synthesis process follows the linearization of DNA molecules within each cluster; polonies generated on the flow cell are read one nucleotide at a time in repetitive cycles. The sequencing can be single-end if only one end is involved, or paired-end where both ends are involved. The advantages of the paired ends are a more accurate alignment to the reference sequence, an improvement during assembly in *de novo* sequencing and of help to resolve repeats. Starting from the primer, a DNA polymerase begins its activity by incorporating fluorescently labelled dNTPs into the growing DNA chain; the results will be a new chain complementary to the template. All four different fluorescent-labelled nucleotides (A, C, T, G), containing a 3' reversible terminator, are simultaneously added and the nucleotides are incorporated base-by-base. A laser is used to excite the incorporated labelled dNTPs which generate fluorescence at four wavelengths, and a high-resolution image is recorded and used to determine which cluster has incorporated the nucleotide; subsequently, the fluorophore along with the 3' reversible terminator is chemically removed allowing the next round of sequencing to occur. Intensities of light signals from the sequencing reactions are converted to bases, and quality score assigned to each base as a measure of the probability of an error in the call. Ideally, all bases within a cluster will be extended in phase. However, it may happens that a small portion of molecules does not extend properly and falls either behind (phasing) or advance a base (pre-phasing); considering that these type of errors will accumulate after many cycles, the result is a quality decrease at each end of the reads (Buermans et al. 2014).

PHYLOGENETIC ANALYSIS

The study of molecular evolution is based on comparative methods and normally uses the phylogeny approach to determine and analyze evolutionary relationships between organisms. Output results are shown by a phylogenetic tree, a graph made of nodes and branches displaying and representing a hypothesis on evolutionary events. The nodes represent the taxonomic unit, a group of individuals that are distinguished from others by their molecular characteristics, while the branches define the relationship in terms of evolutionary descent between individual taxa. The length of the branches in a phylogenetic tree is proportional to the difference between the gene sequences of contiguous species. Trees are classified as rooted and unrooted; a rooted tree has a particular node (the root), which is the common ancestor of all the nodes represented in the tree, and the branches of the tree are oriented in function of time. Only rooted trees allow to determine the direction of an evolutionary process. An unrooted tree describes exclusively the evolutionary relationships between taxa, without providing any information about the evolutionary process as a function of time; in this case it is only possible to know how much a species is distant from another one in terms of evolution. Different mathematical models can be used to build a phylogenetic tree; the currently used ones can be divided into four types: methods based on distance matrix, maximum parsimony, maximum likelihood and Bayesian (Harrison and Langdale 2006; O'Halloran D. 2014).

Methods based on the distances are heuristic and take into consideration the evolutionary distances between species and mostly similar group sequences; the distance is calculated by the number of mutations needed to switch from one species to another. The distance approaches do not allow to analyse which are the characters involved in a particular grouping. Among the methods that use the matrix of distances, the most widely used are the UPGMA (Unweighted Pairs Group Method with Arithmetic mean) and the Neighbor-Joining (NJ). In the first method it is assumed that the evolutionary rates are almost constant in the different evolutionary lines; looking at the distance matrix, the taxa with higher degree of similarity are taken into consideration. These taxa are linked together in the tree and are regarded as single new taxa; the matrix is rebuilt and will contain one element less. The process continues iteratively until we are left with only two taxa linked to one another.; the midpoint between these two indicates the root of the tree. The Neighbour Joining (NJ) calculates the distances between all possible pairs of sequences, and the tree is built considering the relationships between these distances. The algorithm at the basis of this method has the task of identifying the tree without root that minimizes the sum of the lengths of the various branches.

The maximum parsimony methods assume that shared characters among different sequences result from common ancestors. As it is impossible to know the ancestral sequences and if we consider each mutation a mere assumption, we realize that this method tries to make as fewer assumptions as possible. The advantages of the method are the absence of other assumptions in addition to the maximum parsimony. The tree building therefore requires the construction of the possible topologies and the calculation of the minimum number of substitutions for each topology. From this calculation positions that do not show those substitutions and replacements that appear only once are obviously excluded. The method of maximum parsimony (MP), despite being sufficiently satisfactory, has a high computational cost. The method of Maximum Parsimony assumes that the best tree corresponds to that requiring the least number of substitutions to explain the initial data; it is necessary that an algorithm determines which among all the generated trees has the least number of developing steps.

Maximum Likelihood (ML) is a statistical method that computes the probability for a given tree to fit into specific dataset, given a specified model of sequence evolution (Harrison and Langdale 2006). Most of the models assume that the evolution of different sites and different branches are independent; unfortunately this method is extremely expensive from a computational point of view and is not usable for a number of higher taxa, but allows anyway to test statistically different hypotheses.

The Bayesian Inference (BI) method is based on the posterior probability, that is the probability that is estimated on a model of a priori expectations, knowing something in most of the data (Huelsenbeck JP, Ronquist F., 2001).

STRUCTURAL APPROACH

Proteins are essential components of an organism and possess extremely diversified functions:catalytic, structural, transport and storage, immunological and regulative.

Solved structure available in Protein Data Bank (PDB), sequences present in the NCBI sequence databases (http://www.ncbi.nlm.nih.gov/) and the knowledge obtained from in-vitro biochemical studies are all important elements that help us study proteins and their involvement in biological process. Methods based on sequence homology are essential for protein identification and classification; comparison among protein sequences can be used to highlight evolutionary relations of proteins, and sequence similarity can be considered a measure of evolutionary distance among organisms. Proteins with similar sequences starting from the same ancestor have evolved, which implies that proteins could carry out the same functions. It is important to highlight that the relation between sequence and protein functions is not biunique: proteins with similar sequences have similar structures but this does not necessarily imply that proteins with a different sequence always have different structures. On the contrary, it seems that evolution uses the same structures to obtain different functions. Changes on protein structure are the most conserved; protein evolution usually happens in ways that do not modify the folding of protein structure. During evolution most conserved regions are localized inside the protein structure formed by the core and the secondary structure elements, whereas the most evident differences usually appear in regions close to the protein surface, such as the loop regions, where the physical-chemical properties of amino acid side chains frequently change (Illergård et al. 2009). Protein structure prediction starting from amino acid sequences is one of the biggest challenges in bioinformatics and computational biology. Nowadays the most successful computational methods for protein structure prediction rely on a comparative approach called *homology modelling*, which predicts the 3D structure of a target protein sequence, based on a template (resolved protein structure) (Fiser et al., 2010). The method consists of five steps: availability of 3D structure of a protein related to the target sequence, identification and choice of the best structure that will be used as template, amino acid sequence alignment between target and template proteins, building and model evaluation (Fiser et al., 2010). First of all it is important to verify if resolved proteins similar to the target protein are present in Protein Data Bank and how many there are; it is possible to do so by using BLAST software (Altschul S.F. et al., 1990). Numerous algorithms allow to measure the identity or similarity between two sequences in a manner sufficiently accurate to evaluate whether a protein can be the right candidate to be a template to outline an homology modelling approach. It happens that the alignment with the higher identity does not always correspond to the

one that gives the best structural superimposition among proteins. This is why the alignment is obtained using not only the amino acid sequences of the two proteins, but also the structure of the protein template, the prediction of the secondary structure and sometimes a structural superimposition of resolved proteins homolog to the template. The structure of the template can be used to verify if insertions or deletions fall into secondary structure elements or on the surface, whereas the superimposition is used to identify which regions are more structurally conserved in the family and those that probably will be conserved in the protein target too. So the quality of the model obtained will depend on the degree of similarity among target-template sequences and quality of the alignment.

Homology modelling cannot be used if a fitting protein template structure is not available or if the sequence similarity between protein and template is not high enough. In this case another approach could be followed using the *ab initio* method that does not use structures already resolved. Considering the limited number of possible protein folding in nature, the three dimensional structure is more conserved than the amino acid sequence and so it will be possible to identify homology on structures starting from homology in sequences. Homology modelling allows to build 3D structures starting from a single sequence template, but the higher the number of structurally similar sequences available more accurate the analysis will be. In comparative modelling most errors arise during the choice of the best fitting model and in the alignment between target and query sequences. The homology modelling approach has proved to be reliable only if sequence homology between the resolved protein structure and protein to be model is more than 50% (Chothia C. and Lesk AM, 1986). If sequence identity is not higher than 30%, the *homology modelling* approach is not feasible because we are in the so called midnight zone (Rost et al., 1997), but it is suggested to use other approaches, such as *ab initio* or *fold recognition* methods. Homology modelling is based on the observation that proteins having a good level of sequence similarity show a good level of structure similarity; so there is a relation between the similarity of two protein sequences and the similarity between the corresponding three dimensional structures. Proteins with a sequence similarity higher than 30% usually maintain a similar structure, allowing to use homology modelling approach. For the purpose of this work, it was possible to use a homology modelling approach, considering that the sequence similarity of the proteins we took into consideration are in the range between 70%-98%. There are several computer programs and web servers that automate the homology modeling process; in this thesis SwissModel server (http://swissmodel.expasy.org/workspace/) was used to perform an automated modelling pipeline. For a quality evaluation of each model four parameters were considered: QMEANscore, Z-score, Dfire enerty and Ramachandran plot. (Arnold et al., 2006; Kiefer et al., 2009; Schwede et al., 2003; Guex et al., 1997; Peitsch et al., 1995).

Swiss Model

Swiss model is a server dedicated to the automated modelling of proteins using an homology modelling approach and can be used to look for template via tools like PSI BLAST; it is used for the first step in building of models and for their final validation. For models that have a good homology template it is possible to use the Automatic Mode; the user can check all the steps: identify functional domains, secondary structures and disorderly regions; the alignment can be refined manually and it is possible to evaluate the quality of the model. Multiple analyses are done to enable the user to obtain the greatest quantity of data byprecisely estimating the quality of the model obtained (Arnold et al., 2005). To evaluate the potential

energy of the obtained model many tools are used (i.e., ANOLEA, GROMOS, QMEAN, DFire and Z-SCORE). ANOLEA is used to evaluate the quality of the model packaging and to identify regions with high energy and steric clashes (Melo and Feytmans, 1998). GROMOS considers the energy of each amino acid singularly (Gunsteren et al., 1996), whereas the entire energy of the model is evaluated by QMEAN and DFire tools. QMEAN is a scoring function that combines six different parameters, each one with a range value between 0 and 1. The measure of the quality for each model is given by QMEAN Z-SCORE; models with a low quality shows a low Z-SCORE (Benkert et al., 2009;Zhou et a., 2002).

Electrostatic potential

The weak interactions play an important role because they stabilize the three dimensional structure of a protein and are involved in the interactions with other molecules. In a protein the contribution given by Van der Waals interactions, hydrogen bonds, electrostatic and hydrophobic interactions are essential to achieve the final three-dimensional structure of a protein. In a macromolecule there are many charged amino acids: some positive and others negative; charges are involved in the recognition enzyme-substrate or protein-protein via the formation of specific salt bridges. The Coulomb law is useful to calculate the potential around a protein; moreover, both the charges distribution on the protein and their distance allow to calculate the electrostatic potential. A more accurate measurement is obtained through the Poisson-Boltzmann equation, which takes into consideration two dielectrics constants: one of the water with a high value (80) and another one of the protein with a low value (3-4).

In a protein,20%-30% of the total amino acids is given by charged amino acids present mainly on its surface; they must interact with water and other molecules to make the protein soluble. In general, the HA1 subunit of hemagglutinin is positively charged (Arinaminpathy and Grenfell, 2010); positive and negative charges situated near or within the receptor binding pocket seem to influence the affinity to host receptors which are negatively charged and enhance or reduce the avidity to cell membrane(Hensley et al., 2009). In hemagglutinin glycoprotein it seems that the net-charge of HA1 domain could evolve to compensate for the effect of the gain and loss of NGS, probably through changing the avidity; moreover, it seems that the net-charge variation in HA1 has a compensatory effect on the NA activity for keeping the HA-NA balance (Kobayashi et al. 2012).

It is possible to calculate the protein electrostatic potential by using software available on line; APBS and PDB2PQR are among the most widely used free software packages. This web service has been developed to provide users with the necessary amount of computational capabilities (http://www.poissonboltzmann.org/) and to allow working on portable computing platforms (Unni et al., 2011). To compare electrostatic interaction properties of proteins, the webPIPSA service has proved to be a valid tool whichallows to classify proteins according to their interaction properties. (Richter et al., 2008; Blomberg et al., 1999; Wade et al. 2001; Gabdouille et al., 2007). For the similarity analysis it is necessary to upload a set of related protein structures in PDB format. After calculatingthe protein electrostatic potentials, the server proceeds with the calculation of similarity indices based on the electrostatic properties for all pairs of proteins; similarity indices obtained are then converted to electrostatic 'distances'. The electrostatic potential

distance matrix is represented as a heat map, displayed in color coded form and as epogram in a tree format.

AIMS OF THE THESIS

Relevance and impact of the Influenza A virus with respect to public health and poultry industry has been highlighted in the introduction section; however, it can be shortly reminded here that this zoonotic agent is responsible for infectious and contagious diseases in humans and animals resulting in serious and sometimes huge economic losses worldwide. Therefore, in addition representing a topic of interest to basic and health science, the study of the AI viruses also concerns with biotechnological strategies to improve vaccination strategies and surveillance as well as to infer evolutionary trends. The evolutionary dynamics of influenza A virus, like other RNA viruses, are complex in that depending on the combination of high mutation rates, rapid replication and infection of large population size. The resulting viral populations are formed by a mixture of quasi-species variants genetically related but non-identical, which interact and cooperatively contribute to characterize the whole population, and are subjected to a continuous genetic variation, competition, and selection by environmental pressure. Main characteristics of this etiologic agent are a marked genetic and antigenic variation resulting in the ability of viral genomes to tolerate the introduction of mutations or recombinations, which facilitate the timely emergence of new variants against which the traditional vaccines are ineffective. Nowadays, the development of effective vaccines, able to confer improved cross protection and to reduce the risk of emergence of viral mutants, is probably the most effective strategy to control the spread of these viruses.

A specific objective of this study is to combine and integrate genomic, phylogenetic and structural approaches to understanding both the genetic variability and functional evolutionary dynamics of avian influenza A virus. In fact, while genomic mutation is indeed the source for variation, structural analyses allow for properly "weighting" the effect of mutations, as mutations at DNA level altogether contribute to genetic diversity, but they quite differently contribute to "functional" variation, depending on they are either silent mutations (no variation in the protein product), roughly compatible mutations (changes to quite similar residues) or mutations resulting in dramatic effects on protein fold and/or interaction surface features. Deep sequencing technologies can be of help to investigate and characterize the viral population complexity, to reveal low-frequency mutations and to follow the genetic evolution of the quasi-species variants present in a viral population. Current enhancement of next-generation techniques allows to obtain a large number of sequences and thus to improve robustness of the phylogenetic investigations. Phylogenetic analysis is a standard and essential tool to compare molecular sequences of viruses in several environments and under multiple selection pressures, as well as to study their genetic relationships and their evolutionary dynamics. In the AI virus system, the hemagglutinin protein, which modulates the antigenicity, can be responsible for antigenic drift; in order to unveil major determinants in modulation, the recognition of epitopes conserved among different variants is thus to be studied. Given that protein structures are more conserved than corresponding coding sequences and that an increasing number of solved 3D structures are available for proteins we are interested in, it is possible to obtain structural model to perform wide comparisons.

CHAPTER 1

Evolutionary trajectories of two distinct avian influenza epidemics: Parallelisms and divergences

Fusaro A, Tassoni L, Hughes J, Milani A, Salviato A, Schivo A, Murcia PR, Bonfanti L, Cattoli G, Monne I.

Infect Genet Evol. 20 15 Aug;34:457-66

Infection, Genetics and Evolution 34 (2015) 457–466

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Evolutionary trajectories of two distinct avian influenza epidemics: Parallelisms and divergences

Alice Fusaro [a,*], Luca Tassoni [a], Joseph Hughes [b], Adelaide Milani [a], Annalisa Salviato [a], Alessia Schivo [a], Pablo R. Murcia [b], Lebana Bonfanti [a], Giovanni Cattoli [a], Isabella Monne [a]

[a] Istituto Zooprofilattico Sperimentale delle Venezie, viale dell'Università, 10, Legnaro (PD), Italy
[b] MRC-University of Glasgow Center for Virus Research, 464 Bearsden Road, Glasgow, United Kingdom

ARTICLE INFO

Article history:
Received 13 March 2015
Received in revised form 5 May 2015
Accepted 19 May 2015
Available online 20 May 2015

Keywords:
Avian influenza virus
H7 subtype
Parallel evolution
Deep sequencing
Evolutionary dynamics
Molecular analysis

ABSTRACT

Influenza A virus can quickly acquire genetic mutations that may be associated with increased virulence, host switching or antigenic changes. To provide new insights into the evolutionary dynamics and the adaptive strategies of distinct avian influenza lineages in response to environmental and host factors, we compared two distinct avian influenza epidemics caused by the H7N1 and H7N3 subtypes that circulated under similar epidemiological conditions, including the same domestic species reared in the same densely populated poultry area for similar periods of time.

The two strains appear to have experienced largely divergent evolution: the H7N1 viruses evolved into a highly pathogenic form, while the H7N3 did not. However, a more detailed molecular and evolutionary analysis revealed several common features: (i) the independent acquisition of 32 identical mutations throughout the entire genome; (ii) the evolution and persistence of two sole genetic groups with similar genetic characteristics; (iii) a comparable pattern of amino acid variability of the HA proteins during the low pathogenic epidemics; and (iv) similar rates of nucleotide substitutions. These findings suggest that the evolutionary trajectories of viruses with the same virulence level circulating in analogous epidemiological conditions may be similar. In addition, our deep sequencing analysis of 15 samples revealed that 17 of the 32 parallel mutations were already present at the beginning of the two epidemics, suggesting that fixation of these mutations may occur with different mechanisms, which may depend on the fitness gain provided by each mutation. This highlighted the difficulties in predicting the acquisition of mutations that can be correlated to viral adaptation to specific epidemiological conditions or to changes in virus virulence.

1. Introduction

Since the 90s, outbreaks caused by avian influenza virus of the H7 subtype have been frequently reported in domestic poultry throughout the world, causing not only important damage to the poultry industry, but also a great concern for human health, as demonstrated by the recent H7N9 epidemic in China (Chen et al., 2013). Once in poultry, this subtype can evolve into a highly pathogenic form. While the low pathogenic avian influenza (LPAI) virus causes only a mild, primarily respiratory disease in the infected domestic fowl along with production drops, the highly pathogenic avian influenza virus (HPAI) produces an extremely serious disease that can devastate the poultry population.

As shown in previous studies (Campitelli et al., 2004; Lebarbenchon and Stallknecht, 2011), the H7 viruses collected from poultry are genetically related to the viruses from wild birds, suggesting relative frequent interspecies transmissions. Similarly, the distribution of the HPAI H7 strains throughout the phylogenetic trees indicates the evolution of multiple independent highly pathogenic forms from the low pathogenic progenitors (Röhm et al., 1995; Lebarbenchon and Stallknecht, 2011; Abdelwhab et al., 2014).

Following the transmission from wild to domestic birds the virus can experience an accelerated fixation of beneficial mutations to adapt to new species and new environmental conditions. Sequence adaptations to land-based avian species, such as the acquisition of new additional glycosylation sites near the hemagglutinin (HA) receptor binding site (RBS), deletions at the

* Corresponding author at: Division of Comparative Biomedical Sciences, OIE and National Reference Laboratory for avian influenza & Newcastle disease, FAO Reference Centre for animal influenza and Newcastle disease, OIE Collaborating Centre for Diseases at the Human-Animal Interface, Istituto Zooprofilattico Sperimentale delle Venezie, viale dell'Università, 10, 35020 Legnaro (PD), Italy.
E-mail address: afusaro@izsvenezie.it (A. Fusaro).

http://dx.doi.org/10.1016/j.meegid.2015.05.020
1567-1348/

neuraminidase (NA) stalk region or the C-terminal truncation of the non-structural protein 1 (NS1) have been observed in H7 field outbreaks in poultry (Banks et al., 2001; Campitelli et al., 2004; Iqbal et al., 2009; Dundon et al., 2006; Spackman et al., 2003; Bataille et al., 2011) as well as in experimental studies (Giannecchini et al., 2010).

Besides selective pressure applied to the virus by the host, many other changes of ecological conditions can drive the evolutionary dynamics of avian influenza viruses. Vaccination, for example, can determine an increase in the rate of mutations in the antigenic sites of the surface glycoproteins (Beato et al., 2014; Cattoli et al., 2011a,b). However, since epidemiological conditions can be different from one epidemic to the next, their influence on virus evolution can be difficult to establish. This paper aims to provide new insights into the evolutionary dynamics of different viruses experiencing similar host and ecological selective pressures.

Between 1999 and 2004 the densely poultry populated area of Northern Italy experienced two distinct H7 epidemics: one in 1999–2001 caused by an H7N1 virus, the other in 2002–2004 originated by an H7N3 virus. The H7N1 epidemic (1999–2001) was caused by a LPAI strain, which mutated into a highly pathogenic form after circulating in the industrial poultry population for approximately 9 months (from the end of March to December 1999) and causing 199 outbreaks. The HPAI strain provoked the death or culling of over 16 million poultry, as well as substantial economic losses to the industry before its eradication in April 2000. Four months later, the LPAI H7N1 re-emerged, affecting other 78 flocks. To reduce the economic impact of this second wave of LPAI viruses, a DIVA (Differentiating Infected from Vaccinated Animals) vaccination campaign was initiated in November 2000 (Capua and Marangon, 2007; Mulatti et al., 2010). The second epidemic started in October 2002 and was caused by an H7N3 LPAI strain, which, according to previous phylogenetic analyses, was probably introduced from the wild bird reservoir into the domestic poultry (Campitelli et al., 2004). To contain the rapid spread of the infection, from January 2003 a DIVA vaccination campaign was carried out in layers, capons and meat turkeys. The virus managed to circulate for 1 year (until October 2003) and to infect a total of 388 poultry holdings. Similarly to the H7N1 strain, the LPAI H7N3 subtype re-emerged 1 year later, in September 2004. However, thanks to the ongoing vaccination program, this time it caused only 28 new outbreaks (Capua and Marangon, 2007).

Using a Bayesian phylogenetic approach, in our previous study (Monne et al., 2014) we compared the evolutionary dynamics of the H7N1 HPAI viruses with those of low pathogenicity collected during the 1999–2001 epidemic and provided evidence of the origin of the HPAI strain from the LPAI viruses. Starting from these results, here we compared the evolutionary trajectories of two distinct naturally occurring epidemics (the 1999–2001 H7N1 and the 2002–2004 H7N3), which had affected the same domestic species (mainly turkeys and chickens) reared in the same geographic area (Veneto, Lombardia and Emilia Romagna regions) for similar periods of time (about 2 years).

2. Materials and methods

2.1. Viruses included in this study

In this study, we generated the complete genome sequences of 35 H7N3 avian influenza A viruses collected from poultry in Northern Italy from October 2002 to December 2004. In addition, we sequenced the partial genomes of five samples from which whole genome sequences could not be obtained. Sequences of 37 H7N3 viruses from the 2002–2004 epidemic publicly available in the Influenza Virus Resource at GenBank were also included in the analysis.

These data were compared to the HA sequences of 144 samples and to the complete genome of 109 isolates collected during the 1999–2001 LPAI/HPAI H7N1 epidemic, sequenced and analyzed in our previous study (Monne et al., 2014).

Epidemiological information (collection date and province of collection) for all the H7N3 viruses included in this study is available in the Supplementary material (Table S1).

2.2. Sanger sequencing

Viral RNA was extracted from the infected allantoic fluid of specific-pathogen-free fowls' eggs using the Nucleospin RNA kit (Macherey–Nagel, Duren, Germany) and reverse transcribed with the SuperScript III Reverse Transcriptase kit (Invitrogen, Carlsbad, CA). PCR amplifications were performed by using specific primers (sequences are available on request). Amplicons were subsequently purified with ExoSAP-IT (USB Corporation, Cleveland, OH) and sequenced using the BigDye Terminator v3.1 cycle sequencing kit (Applied Biosystems, Foster City, CA). The products of the sequencing reactions were cleaned-up using the PERFORMA DTR Ultra 96-Well kit (Edge BioSystems, Gaithersburg, MD) and analyzed on a 16-capillary ABI PRISM 3130xl genetic analyzer (Applied Biosystems, Foster City, CA).

2.3. Library preparation, Illumina sequencing and data analysis

To assess virus population diversity, next-generation sequencing (NGS) was performed on all the H7N3 clinical samples available in our repository (eight tracheas and one pool of organs). These nine samples were collected during the first 3 months of the epidemic. Unfortunately, no clinical samples collected after January 2003 were available. Full sample details are described in Table S1. These newly generated sequences were analyzed together with the NGS data generated by Monne et al. (2014) for six LPAI H7N1 clinical samples.

Viral RNA was extracted directly from the infected clinical samples using the Nucleospin RNA kit (Macherey–Nagel, Duren, Germany) and processed as described by Monne et al. (2014). In summary, the complete influenza A genomes were amplified with the SuperScript III One-Step RT-PCR system with Platinum®Taq High Fidelity (Invitrogen, Carlsbad, CA) (Zhou et al., 2009). Sequencing libraries were obtained using Nextera DNA XT Sample preparation kit (Illumina). Finally the indexed libraries were pooled in equimolar concentrations and sequenced in multiplex for 250 bp paired-end on Illumina MiSeq, according to the manufacturer's instructions.

Raw sequence reads were inspected using FASTQC to assess the quality of data. Fastq files were cleaned with PRINSEQ and Trim Galore to remove low quality bases at the 5′ and 3′-end of each read and to exclude reads with a Phred quality score below 30 and shorter than 80 nucleotides. Reads were aligned to A/turkey/Italy/8535/2002 (H7N3) reference sequences using Stampy (Lunter and Goodson, 2011). The BAM alignment files were parsed using the diversiTools program (http://josephhughes.github.io/btctools/) to determine the average base-calling error probability and to identify the frequency of polymorphisms at each site relative to the reference used for the alignment. Only polymorphisms with a frequency above 2% were considered.

2.4. Phylogenetic and molecular analyses

Sequences of the HA gene and the gene segments coding for the six internal proteins of the H7N1 and H7N3 viruses were aligned and compared with the most related sequences available in

GenBank. GISAID or GenBank accession numbers of all the Italian H7 sequences analyzed here are reported in Table S1 of the Supplementary material.

The likelihood mapping analysis available in the TREE PUZZLE program (Schmidt et al., 2003) was adopted to visualize the phylogenetic content of the eight datasets. In particular, we investigated the phylogenetic signals for (a) all codon positions, (b) first codon position, (c) second codon position, (d) third codon position, (e) first and second codon positions of the alignments. Since the results obtained for the three codon positions showed the highest percentage of resolved quartets, the following analyses were performed using all the alignment positions.

Since the NA gene segment of the H7N1 and H7N3 viruses belong to two different subtypes, we focused our analyses on the remaining segments (segments 1–5, 7 and 8).

To characterize the presence of recombination in the Italian H7 viruses, the seven gene segments were analyzed for recombination by using two different software packages. We employed the RDP (Martin and Rybicki, 2000), GENECONV (Padidam et al., 1999), 3Seq (Boni et al., 2007), MaxChi (Maynard Smith, 1992) and BootScan methods implemented in the RDP3 program version 3.44 (Martin et al., 2010). We also used the Single Breakpoint Recombination and Genetic Algorithm Recombination Detection (GARD) methods (Kosakovsky Pond, 2006) within the HyPhy package (Pond and Frost, 2005) or on the Datamonkey server (Delport et al., 2010). None of these methods showed evidence of recombination. In addition, to exclude the presence of reassortant viruses, the concatenated gene segments were analyzed using the statistical methods RDP, GENECONV, 3Seq, MazChi and BootScan available in the RDP3.44 package.

Maximum likelihood (ML) phylogenetic trees were constructed using the best-fit general time-reversible (GTR) model of nucleotide substitution with gamma-distributed rate variation among sites (with four rate categories, $\Gamma 4$) and a heuristic SPR branch-swapping search (Guindon and Gascuel, 2003) available in the PhyML program version 3.0. To assess the robustness of individual nodes of the phylogeny, one hundred bootstrap replicates were performed. Topologies of phylogenetic trees were confirmed using the ML method available in the RAXML program (Stamatakis, 2006a), incorporating the GTR model of nucleotide substitution with the CAT model of rate heterogeneity among sites (data not shown) (Stamatakis, 2006b). Phylogenetic trees were visualized with the program FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/).

The history of the character evolution along the branches of the phylogenies was graphically visualized using the parsimony algorithm available in the Mesquite program (Maddison and Maddison, 2014). All the residue numbering reported throughout the text and in the figures will be according to the mature H7 protein. Negative numbers will correspond to the positions located in the signal peptide.

The parallel mutations identified in the HA protein were mapped on the three-dimensional structure of A/turkey/Italy/214845/2002 (H7N3) obtained from Protein Databank (PDB ID: 4BSG) (Xiong et al., 2013).

We explored the possible adaptive role to domestic birds of the parallel mutations identified in the HA protein. To this aim, the sequences of all H7 viruses of avian origin available in GenBank were downloaded and grouped according to the domestic (chicken, guinea fowl, turkey, quail and ostrich) or wild status (*Anseriformes*, *Charadriiformes*, *Columbiformes*, *Passeriformes*, *Psittaciformes*) of the host, for a total of 1143 sequences (723 from wild and 420 from domestic birds). All the sequences, for which the host origin was not clearly defined, were excluded from the analysis. For each position the relative frequency of every amino acid was calculated for both viruses from domestic and wild birds and compared. In addition, the entropy difference between the two groups was calculated using the Entropy-Two tool available at http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy.html.

To determine the change in the amino acid variability over time, samples were grouped according to the month of collection and the within group mean *p*-distance was calculated for the HA protein of each group containing at least three sequences using the program MEGA 5 (Tamura et al., 2011).

2.5. Analysis of selection pressures

Gene- and site-specific selection pressures for all segments of the Italian H7N1 and H7N3 viruses were measured as the ratio of nonsynonymous (d_N) to synonymous (d_S) nucleotide substitutions per site. In all cases, d_N/d_S ratios were estimated using the fixed-effects likelihood (FEL) and Fast Unconstrained Bayesian AppRoximation (FUBAR) methods (Pond and Frost, 2005; Murrell et al., 2013) available at the Datamonkey online version of the Hy-Phy package (Delport et al., 2010). All analyses utilized the GTR model of nucleotide substitution and ML phylogenetic trees.

2.6. Evolutionary dynamics

For each gene segment of the H7N3 viruses, rates of nucleotide substitution per site per year (subs/site/year) of the sampled data were estimated using the Bayesian Markov chain Monte Carlo (MCMC) approach available in the BEAST program, version 1.7.5 (Drummond and Rambaut, 2007). For each analysis, we employed a relaxed (uncorrelated lognormal) molecular clock, a flexible Bayesian skyline coalescent tree prior (10 piece-wise constant groups), a HKY85 + Γ_4 model of nucleotide substitution and the SRD06 codon position model with two data partitions of codon positions (1st + 2nd positions, 3rd position) with base frequencies unlinked across all codon positions. Default prior distributions were used for all the parameters, except for the nucleotide substitution rate prior, for which a gamma distribution (initial value 0.001, shape 0.001, scale 1000) was set. In all cases, statistical uncertainty is reflected in values of the 95% highest probability density (HPD) for each parameter estimate. For each analysis chain lengths were run for sufficient time to achieve convergence as assessed using Tracer v1.5 program (Drummond and Rambaut, 2007).

3. Results

3.1. Parallel evolution of the H7N1 and H7N3 viruses

Our maximum likelihood (ML) phylogenetic analyses of the seven gene segments (segments 1–5, 7 and 8) of 181 (109 H7N1 and 72 H7N3) H7 viruses representative of the 1999–2001 and 2002–2004 Italian epidemics (Figs. 1 and S1–S6 in the Supplemental material) indicate that the two subtypes fall within two well-supported monophyletic clades, defined by both high bootstrap values (>80%) and long branch length, as exemplified by the HA phylogeny (Fig. 1 – H7N1 clade in blue and H7N3 clade in green). This finding indicates that these clades (H7N1 and H7N3) are likely to represent two separate introductions of the virus into the northern Italian regions. As observed in our previous study (Monne et al., 2014), the H7N1 HPAI viruses (marked in yellow in Fig. 1) form a separate cluster within the H7N1 clade in all the phylogenies and are characterized by 19 unique amino acid changes acquired across the entire genome.

The phylogenetic tree inferred for the HA gene for a total of 144 H7N1 and 77 H7N3 viruses identifies in both clades two main groups of LPAI viruses, namely H7N1 LPAI-I and H7N3 LPAI-I (marked in gray in Fig. 1). The H7N1 LPAI-I group is characterized by a total of

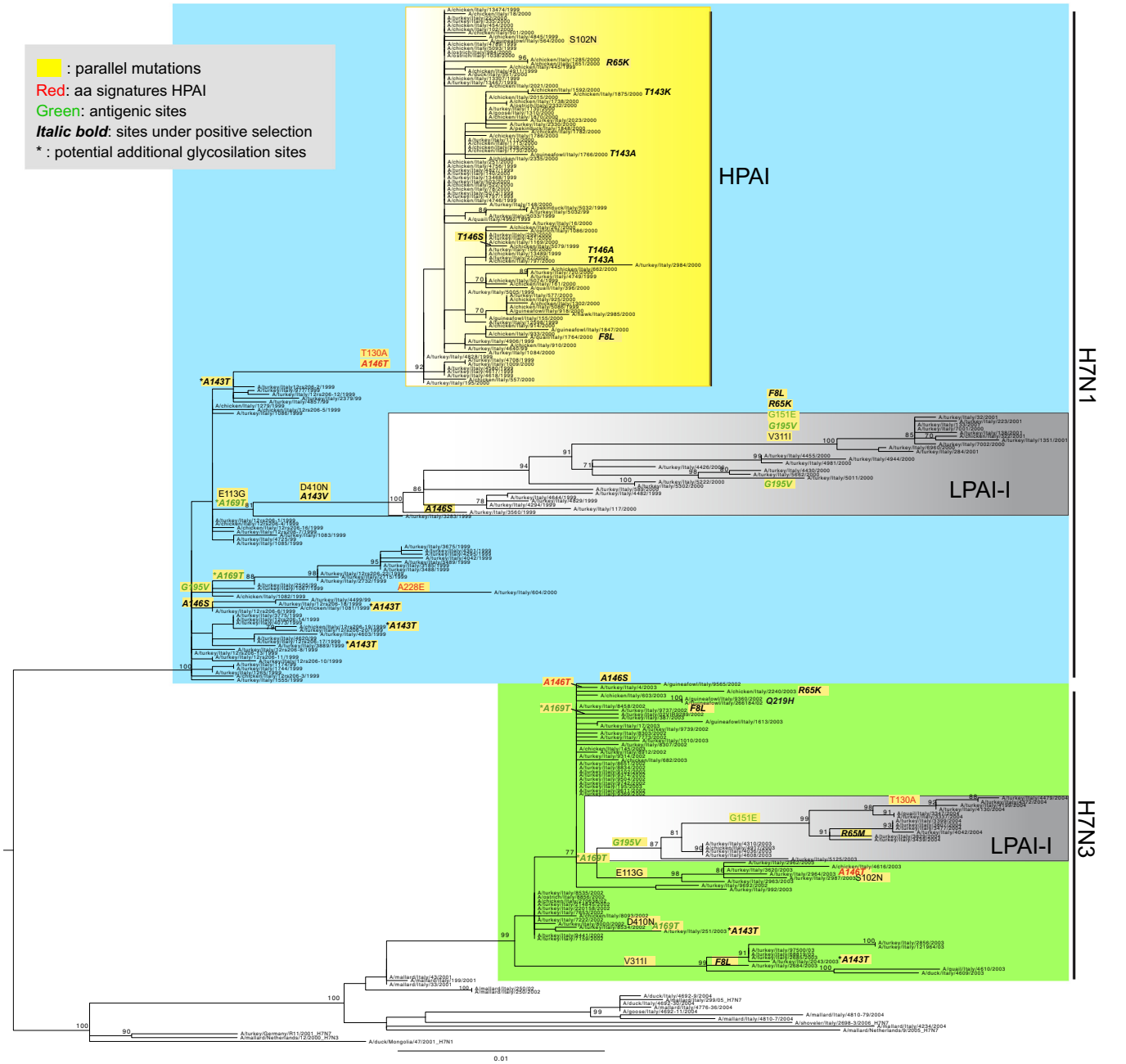


Fig. 1. ML phylogenetic tree of the HA gene segment of Italian H7 avian influenza viruses. Genetic groups are colored as follows: blue for H7N1 viruses collected during the 1999–2001 Italian epidemic, green for the H7N3 viruses collected during the 2002–2004 Italian epidemic. The HP H7N1 viruses are marked in yellow, while gray represents the two LPAI-I groups identified during the H7N1 and H7N3 epidemics. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. Sequences obtained using the NGS platform are marked in red. Parallel mutations are highlighted in yellow. The tree is mid-point rooted for clarity only. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

16 amino acid signatures (Monne et al., 2014) and is defined by high bootstrap values (>70%) and long branches in the HA, PB2, PA, NS phylogenies (Fig. 1 and Supplementary Figs. S1, S3 and S6), while the H7N3 LPAI-I viruses form a distinct cluster with high bootstrap supports (>87%) only in the HA and PA phylogenies (Fig. 1 and Supplementary Fig. S3) and showed only three amino acid signatures, all located in the HA gene (A151T, Q201H and G177V).

Interestingly, these LPAI-I groups are characterized by several common features: (a) emerged 7 months after the first outbreak, (b) persisted until the end of the epidemic, and (c) contain an additional glycosylation site at position 149–151 of the HA gene.

This parallel evolution of the H7N1 and H7N3 epidemics can be observed in many other aspects, which go beyond the emergence of the LPAI-I groups. We identified a total of 32 identical amino acid substitutions across the viral genomes, which were independently acquired by both subtypes during their evolution. Thirteen (41%) of these mutations are positioned within the HA protein (mutations highlighted in yellow in Fig. 1; Table 1) and 9 of them (positions 47, 84, 95, 112, 125, 128, 133, 151, 177) localize on the trimer surface (Fig. 2). In particular, three of these amino acid changes are located within antigenic sites, six are positively selected sites and two create new additional glycosylation sites

Table 1
Characteristics of the 13 amino acid changes of the HA protein acquired independently by both H7N1 and H7N3 viruses (parallel mutations). Positive selection = mutations located in sites under positive selection; HPAI = mutations characteristic of the HPAI H7N1 viruses; antigenic site = mutations positioned in antigenic sites; AGS = additional glycosylation site.

Site	Positive selection	HPAI	Antigenic site	AGS
F-11L	√			
R47K	√			
S84N				
E95G				
T112A		√		
A125T	√			√
A128T	√	√		
A128S	√			
G133E			√[b]	
A151T	√		√[c]	√
G177V	√		√[d]	
V293I				
D67N[a]				
TOT	7	2	3	2

[a] Position in the HA2.
[b] (Stevens et al., 2006; Bush et al., 1999)(position 142 in H3 numbering.)
[c] (Kaverin et al., 2007; Li et al., 2009)(position 160 in H3 numbering.)
[d] (Bush et al., 1999)(position 186 in H3 numbering.)

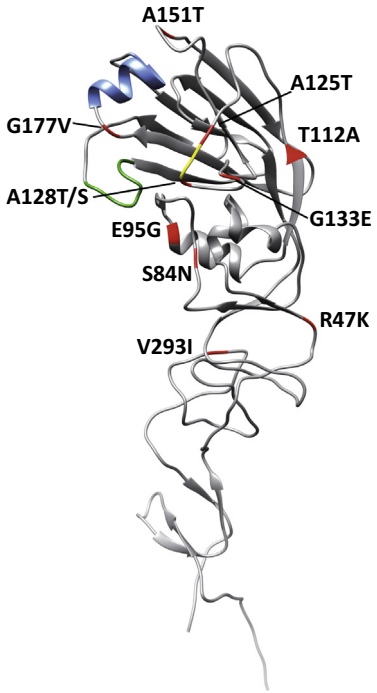


Fig. 2. A/turkey/Italy/214845/2002 (H7N3) HA1 monomer. Positions of the parallel mutations in the HA1 monomer are shown in red; the 130 and 220 loops and the 190 helix are colored in yellow, green and blue respectively (image drawn with UCSF Chimera software). The positions 125 and 128 are part of the 130 loop; the mutations E95G and V293I are buried in the trimer; the mutation F-11L is not included in the three-dimensional structure. The protein model was obtained from the Protein Data Bank (PDB identification 4BSG, Xiong et al., 2013). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(Table 1). Some of these parallel mutations caused significant changes in the chemical properties of the involved amino acids; noteworthy 5 mutations determine the loss (T112A) or the acquisition (A125T, A128T, A128S, A151T) of a hydroxyl group, potentially involved in hydrogen bonds, which could stabilize the structure of the HA conformation. One mutation, G133E, determines the acquisition of a negative charge, while E95G causes a loss of a negative charge.

The other 19 parallel substitutions are distributed across the internal proteins. In particular, three are situated in the PA (V127I, R213K, K609R), two in the PA-X (G213S, R221Q), five in the PB2 (R62K, D195N, R508K, M570I and D740N), four in the PB1 (A14V, T39A, I667V, K586R), two in the NS1 (M/I79V, S228P), one in the NP (K77R) and two in the M1 protein (R134K, G228R). Moreover, one H7N3 virus and 10 H7N1 samples possessed a truncated PB1-F2 protein of respectively 8 and 11 amino acids.

Most of these parallel mutations (26/32) occurred along the internal branches of the phylogenies. Of note are the A151T, G177V, G133E substitutions in the HA protein, which were acquired during the evolution of both H7N1 LPAI-I and H7N3 LPAI-I groups (Fig. 1). These three mutations were located close to the RBS and within the corresponding H3 antigenic sites, suggesting a possible role in escaping the immune response (Fig. 2). In particular, the position 177 is situated between the 190 helix and the 220 loop, while the site 133 localize near the 130 loop.

On the other hand, five parallel mutations – A128S in the HA gene (Fig. 1), G213S and K609R in the PA, G213S in the PA-X and M/I79V in the NS1 – are specific to single individuals and are not transmitted, at least at the consensus level, to other individuals included in this study. In addition, 11 of the 32 parallel mutations (F-11L, R47K, S84N, A125T, A128T, A128S, A151T, G177V in the HA, R221Q in the PA-X and D195N, D740N in the PB2) were fixed along multiple branches of the tree, suggesting that these positions were subjected to positive selection. This has been confirmed also by our analysis of selection pressure described below.

To investigate the frequency of the occurrence of the parallel substitutions which became fixed at a population level, we compared the number of amino acid variants between one of the first isolated samples from each epidemic, which fall near the root of each lineage (A/turkey/Italy/1555/1999 (H7N1) and (A/turkey/Italy/7159/2002 (H7N3)), and the most recent isolates of the H7N1 and H7N3 epidemics, which lied on the external branches of the trees (A/turkey/Italy/1351/2001 (H7N1) and A/turkey/Italy/4479/2004 (H7N3)). Comparison of the PB2, PB1, PB1-F2, PA, PA-X, HA, NP, M1 and M2 proteins revealed a total of 47 amino acid differences between the 1999 and 2001 H7N1 viruses, 28% (13/47) of which were parallel substitutions, and 35 mutations between the 2002 and 2004 H7N3 samples, 20% (7/35) of which were parallel. The NA and NS gene segments, which respectively belong to different subtypes and alleles, were excluded from these calculations. Interestingly, limiting the comparison to the only HA protein, the percentage of parallel substitutions on the total number of amino acid changes increased to 56% for the H7N1 and 44% for the H7N3 viruses.

In addition, we assessed the amino acid differences between the HA protein of the two most recent isolates of both epidemics (A/turkey/Italy/1351/2001 (H7N1) and A/turkey/Italy/4479/2004 (H7N3)). Parallel amino acid replacements were observed at 7 (33%) of 21 positions showing variations.

3.2. Intra-host genetic variability

To better understand the mechanism behind the acquisition of the parallel mutations we used an ultra-deep sequencing approach. Specifically, we examined whether the 32 parallel mutations were present at low level in viruses collected during the first wave of the two epidemics. Unfortunately, the likelihood to detect subpopulations containing the parallel mutations that emerged in later stages of the epidemic was reduced due to the lack of available H7N3 samples after the first 3 months of the epidemic.

The complete genomes of eight H7N3 samples and the HA, NP, M, NS of A/turkey/Italy/9289/02 were sequenced. These data were

analyzed together with the reads obtained in our previous study for six LPAI H7N1 clinical samples (Monne et al., 2014). The mean depth of coverage ranged from 1909 reads for sample A/turkey/Italy/1744/99 to 13762 reads for A/turkey/Italy/4/03 (Table S1). In particular, the depth of coverage of the polymerase genes (PB2, PB1 and PA) was lower than for the other five segments (ranging approximately between 20 and 6000 reads).

We assessed the genetic variability within each viral population for the 32 sites where parallel mutations between the H7N1 and the H7N3 viruses were identified. We detected eleven sites where the parallel substitutions have already become the prevalent viral population in at least one sample, with a frequency ranging from 51.6% to 100% (Fig. 3). In addition, we identified six sites showing parallel mutations with a frequency from 2% to 35.4% (minority variants). In particular, three substitutions (A128T and A128S of the HA1 and D67N of the HA2 protein) were observed both as minority and fixed variants in different samples of the same subtype (Fig. 3).

3.3. Positively selected sites in the H7N1 and H7N3 viruses

We identified in the HA gene five sites under positive selection with the FUBAR method (posterior probabilities $\geqslant 0.9$), four of which were also detected with the FEL method (p-value <0.05, Table 2). As expected, all these sites (positions 47, 125, 128, 151 and 177) coincide with the positions of the parallel mutations detected in multiple branches of the HA phylogeny. Interestingly, two of them (151 and 177) are located in the H3-corresponding B antigenic site (Kaverin et al., 2007; Li et al., 2009; Bush et al., 1999). Moreover, mutation at position 151 introduces a potential additional glycosylation site, which may have a strong antigenic effect. Analyses of selection pressure on the remaining gene segments identify one positively selected site in the PB2 and nine in the PB1-F2 protein only with the FUBAR method (Table 2). The latter protein is the only one showing evidence of diversifying selection (d_N/d_S = 5.07), although this is likely to be an artefact due to its encoding in an overlapping reading frame (Holmes et al., 2006).

Table 2
Amino acid sites under putative positive selection (PSS, positive selected sites) detected using different models (FUBAR and FEL) and mean d_N/d_S ratio for each gene.

Gene	d_N/d_S (95% CI)	FUBAR		FEL	
		PSS	Post. prob	PSS	p-Value
HA	0.33 (0.29–0.38)	47	0.96	47	0.04
		125	0.99	125	0.02
		128	0.99	128	0.01
		151	0.98	151	0.04
		177	0.90		
PB2	0.13 (0.11–0.15)	398	0.92		
PB1	0.10 (0.08–0.12)	–	–	–	–
PA	0.15 (0.12–0.18)	–	–	–	–
NP	0.06 (0.04–0.08)	–	–	–	–
M1	0.22 (0.15–0.29)	–	–	–	–
M2	0.81 (0.49–1.13)	–	–	–	–
NS1	0.27 (0.23–0.32)	–	–	–	–
NS2	0.25 (0.19–0.32)	–	–	–	–
PB1-F2	5.07 (3.30–6.84)	23	0.96		
		30	0.99		
		37	0.93		
		38	0.90		
		43	0.90		
		54	0.97		
		66	0.97		
		69	0.90		
		82	0.98		

3.4. Frequency of the parallel mutations in the H7 viruses from wild and domestic birds

To explore whether the parallel mutations identified in the HA protein of the Italian H7 viruses may be associated with molecular adaptation to domestic hosts, we analyzed the relative frequency of each amino acid in the H7 viruses collected from poultry and

gene	mutation	H7N1						H7N3								
		1744	2732	3283	3675	4295	4829	7773	8093	8303	9289	9369	9565	9611	4	17
HA1	F-11L	na		na		na				na						
	R47K	2,2														
	S84N															
	E95G			100,0			99,9									
	T112A											51,6				
	A125T															
	A128T											7,8			83,8	
	A128S					6,5					7,9		99,9		2,6	
	G133E					2,4										
	A151T		99,0	98,7	98,4	98,7	99,5				80,2					
	G177V		99,0		99,3	99,0										
	V293I															
HA2	D67N			2,0			99,8									
PB1	A14V	na			99,0	na		99,8	99,7	99,8	na	na	99,8	100,0	100,0	na
	T39A	na				na					na					na
	K586R										na					
	I667V										na					
PB2	R62K	na									na					
	D195N	na									na					
	R508K	na									na					
	M570I	na									na					
	D740N	na									na					
PA	V127I	na								35,4	na					
	R213K	na									na					
	K609R	na									na					
NP	K77R							100,0	99,9	100,0	na	100,0	100,0	99,9	99,9	100,0
M	R134K										na					
	G228R										na					
NS	I79V								99,8		na					
	S228P		99,5		99,7	99,7					na					

Fig. 3. Heat-map of the percentage of reads showing the parallel mutations in the H7N1 and H7N3 viruses. Only mutations with a frequency higher than 2% are reported and highlighted with a grayscale according to the frequency of the mutation (white <2%, black 100%).

Table 3
Amino acid frequencies in the parallel sites of the HA protein in H7 viruses collected from wild and domestic hosts. AA = amino acid of the progenitor viruses of the H7 Italian epidemics; sub = amino acid substitution acquired independently by both subtypes (H7N1 and H7N3) during the two Italian epidemics (parallel mutations).

Protein	Pos	Parallel mutations		H7 from wild birds			H7 from domestic birds		
		AA	Sub	% AA	% Sub	% Other AA	% AA	% Sub	% Other AA
HA1	−11	F	L	66.8	32.7	0.6	81.8	15.9	2.3
	47	R	K	93.6	6.1	0.3	92.1	6.0	1.9
	84	S	N	41.4	11.8	46.9	22.1	16.9	61.0
	95	E	G	96.7	3.0	0.3	96.4	3.6	0.0
	112	T	A	42.2	4.3	53.5	19.3	18.6	62.1
	125	A	T	95.4	1.5	3.0	77.9	12.1	10.0
	128	A	T	97.9	1.5	0.6	91.9	1.2	6.9
	128	A	S	97.9	0.4	1.7	91.9	6.9	1.2
	133	G	E	98.6	0.8	0.6	98.6	0.7	0.7
	151	A	T	98.5	0.4	1.1	93.8	4.8	1.4
	177	G	V	93.8	2.5	3.7	46.0	12.4	41.7
	293	V	I	99.0	0.7	0.3	94.0	5.0	1.0
HA2	67	D	N	99.0	0.8	0.1	95.2	4.8	0.0

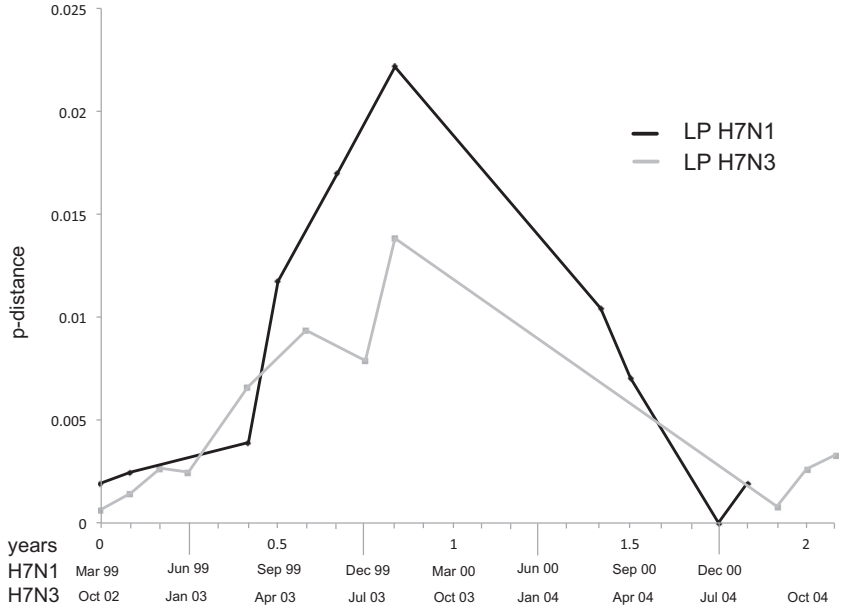


Fig. 4. Pattern of amino acid variability. The graph represents the trend of the *p*-distance among the amino acid sequences of the HA gene of the H7N3 (gray) and H7N1 (black) viruses collected during the same month.

wild birds available in GenBank. The amino acid changes acquired independently by both subtypes appear to be at low frequency in both H7 viruses from wild and domestic birds. However, most of them (8 out 13) seem to be acquired more often in viruses circulating in poultry and show a significantly higher entropy (*p*-value <0.05) compared to sequences of viruses from wild birds. This is particularly evident for positions 84, 112, 125 and 177, where the frequency of the mutated amino acids is higher (5–14%) in the viruses from domestic animals (Table 3).

3.5. Comparison of the dynamics of the amino acid sequence variability through time

To further explore the pattern of amino acid sequence variability of the HA protein during the H7N1 and H7N3 epidemics and the possible effect of bottleneck events, we grouped the viruses according to the month of collection and calculated the *p*-distances within each group containing at least three sequences (Fig. 4). Surprisingly, the LPAI viruses belonging to the two subtypes show a similar pattern of genetic variability, with a rapid increase in the genetic diversity during the first 10 months of each epidemic (corresponding to the first epidemic wave for both the LPAI H7N1 and H7N3) with a steeper increase for H7N1, followed by a drastic reduction of the amino acid variability, that reached the lowest value in both epidemics about 1 year after the last outbreak of the first wave (March 1999–December 1999 for the H7N1, October 2002–October 2003 for the H7N3). For the H7N3 epidemic, this date (September 2004) corresponds to the re-occurrence, after almost a year with no reported cases, of the infection caused by viruses of the H7N3 LPAI-I group, thus explaining this low amino acid variability. Similarly, all the LPAI H7N1 viruses that re-emerged in August 2000, approximately 8 months after the last H7N1 LPAI outbreak, belong to the H7N1-LPAI-I group. However, during the H7N1 epidemic the re-emerged viruses collected between August and September 2000 showed an intermediate variability, which progressively decreased reaching the lowest

Table 4
Estimated rates of nucleotide substitution for the H7N1 and H7N3 viruses collected during the two Italian epidemics.

Gene	Genetic group	Evolutionary rates (sub/site/year)		Comments
		Mean ($\times 10^{-3}$)	95% HPD ($\times 10^{-3}$)	
HA	H7N1	10.15	8.5–11.9	Monne et al.
	H7N3	8.46	6.46–10.65	This study
NA	H7N1	9.86	8.01–11.79	Monne et al.
	H7N3	3.78	2.56–5.23	This study
PA	H7N1	5.84	4.79–6.91	Monne et al.
	H7N3	4.65	3.17–6.19	This study
M	H7N1	7.21	5.02–9.51	Monne et al.
	H7N3	4.61	2.94–6.4	This study
PB1	H7N1	5.57	4.54–6.68	Monne et al.
	H7N3	4.56	3.24–5.99	This study
PB2	H7N1	6.81	5.54–8.04	Monne et al.
	H7N3	5.08	2.78–7.76	This study
NP	H7N1	5.42	3.81–7.15	Monne et al.
	H7N3	5.65	3.71–7.62	This study
NS	H7N1	7.72	5.76–9.91	Monne et al.
	H7N3	4.59	3.01–6.35	This study

value in December 2000, 12 months after the last LPAI outbreak of the first wave (Fig. 4).

3.6. Comparison of the rate of nucleotide substitution of the two epidemics

To further confirm the similar evolutionary dynamics of the H7 viruses collected during the 1999–2001 and 2002–2004 epidemics, we also compared the evolutionary rates of the eight gene segments of the H7N1 (Monne et al., 2014) and H7N3 viruses, both inferred using a Bayesian coalescent approach (Drummond and Rambaut, 2007). Of note, the evolutionary rate appears to be very similar between the two subtypes for most of the eight gene segments (7 out 8). Both viruses show high nucleotide evolutionary rate for the HA glycoprotein: mean rate of 10.15×10^{-3} substitutions per site per year (sub/site/year) (95% HPD, $8.5–11.9 \times 10^{-3}$) for the H7N1 viruses, and 8.46×10^{-3} sub/site/year (95% HPD, $6.46–10.65 \times 10^{-3}$) for the H7N3 viruses (Table 4). In contrast, different evolutionary rates have been observed for the NA genes belonging to the two different subtypes (N1 and N3), with the N1 subtype evolving faster (mean rate of 9.86 sub/site/year, 95% HPD, $8.01–11.79 \times 10^{-3}$ sub/site/year) than the N3 (mean rate of 3.78 sub/site/year, 95% HPD, $2.56–5.23 \times 10^{-3}$ sub/site/year) (Table 4).

4. Discussion

Evolutionary rates of influenza A viruses allow them to quickly acquire genetic mutations that may be associated with increased virulence, host switching or antigenic changes. Here we compared and contrasted the evolutionary trajectories of two avian influenza strains circulating in similar epidemiological conditions, providing a better understanding of the adaptive solutions of avian influenza viruses during naturally occurring epidemics.

The phylogenetic analyses of the avian influenza viruses collected during the 1999–2001 H7N1 and 2002–2004 H7N3 Italian epidemics support the occurrence of the independent introduction of these two subtypes in the Italian poultry population, which form two separate lineages. Despite circulating under similar epidemiological conditions, at first glance they appear to have experienced largely divergent evolution, given that the H7N1 viruses evolved into a highly pathogenic form, while the H7N3 did not.

However, a deeper insight into the evolution of these two strains has revealed many similarities, as detailed below.

4.1. Emergence of parallel mutations

During the two epidemics, both subtypes have independently acquired 32 identical amino acid mutations throughout the entire genome, 13 of them (40%) found in the HA segment. It is worth mentioning that three mutations located within the antigenic sites of the HA protein (G133E, A151T and G177V) are of particular interest: having been acquired and subsequently maintained until the end of the epidemics by both subtypes suggest that those antigenic changes could be advantageous for the virus. Of note, the three mutations had been previously described as associated to the antigenic drift of the Italian H7N3 viruses (Beato et al., 2014), as a consequence of the vaccination program implemented since January 2003. Interestingly, these mutations were also present in all the H7N1 viruses collected after vaccination was introduced in November 2000, 4 months before the eradication of the H7N1 infection and, in particular, two of them (G133E and G177V) had risen only after the implementation of this control strategy, although the role of vaccination on the emergence of these substitutions cannot be assessed. Additionally, the order in which these beneficial mutations were fixed in the viral population was similar. In particular, substitution A151T, which introduces a potential additional glycosylation site, emerged earlier and seemed to be independent from the use of vaccination, given that the H7N1 viruses acquired it before vaccine implementation. However, the use of a 1999 H7N1 LPAI virus as a vaccine strain during the H7N3 vaccination campaign might have favoured the evolution of mutants harboring substitutions at the globular head of the HA1 protein in the H7N3 viruses, similar to the ones acquired by the H7N1 strain.

The high percentage (56% for the H7N1 and 44% for the H7N3) of parallel substitutions on the total number of amino acid changes of the HA protein, which became fixed during the viral evolution, suggests that they may not be caused only by the stochastic forces of mutation, but instead may support the hypothesis that selection on the HA has resulted in parallel evolution of independent lineages circulating in similar environmental conditions. Further studies would be necessary to better characterize the specific role of this set of parallel mutations, although some of them may be related to the adaptation of the virus to the poultry population. The highest frequency of eight mutations of the HA gene in the H7 sequences obtained from viruses circulating in poultry rather than in wild birds may support this hypothesis. In particular, the acquisition of two potential additional glycosylation sites at position 123–125 and 149–151 in our samples is a typical change observed during adaptation of aquatic bird viruses to the domestic population (Giannecchini et al., 2010; Aamir et al., 2009; Lebarbenchon and Stallknecht, 2011; Bataille et al., 2011).

Our deep sequencing analysis revealed that 14 of the 32 parallel mutations were already present in at least one of the 15 analyzed samples collected at the beginning of the two epidemics. In particular, three of them were identified as minority variants (frequency <50%), eight as fixed variants (frequency >50%) and three showed a variable frequency ranging from 2% to 99.9%. This finding suggests that there may be additional benefits to the virus population by maintaining these variants. The final acquisition through reassortment and/or fixation of these mutations may occur with different mechanisms – i.e. bottleneck events, positive selection, random genetic drift or antigenic drift – which may depend on the increased replication capacity, pathogenicity or adaptive

advantage to specific selection pressures (i.e. vaccination) provided by each mutation or a combination of these.

4.2. Similarities between the LPAI-I groups

The HA phylogeny showed that only two LPAI genetic groups (H7N1 LPAI-I and H7N3 LPAI-I) persisted until the end of the two epidemics, probably due to their higher reproductive success. These clusters showed similar characteristics: they were detected 7 months after the first outbreak and possessed a potential additional glycosylation site (AGS) in the HA protein.

4.3. Similar pattern of amino acid variability

The change in the heterogeneity of HA proteins during the epidemic is comparable, indicating that the evolution of the HA follows a similar pattern for both viruses. In particular, the rapid increase in the amino acid variability during the first epidemic wave suggests a fast increase in the population size, which had consequently led to the emergence of multiple competing variants successively depleted by bottleneck events, which caused the survival of the sole LPAI-I group in both epidemics, which may have circulated undetected for months in a restricted population. Moreover, the amino acid distance among the viruses collected from each epidemic ranged from 0% to 1.8% for both the H7N1 and the H7N3 viruses, suggesting that the two strains accrued similar amino acid variability.

4.4. Similar evolutionary rates

Except for the NA gene, which belongs to different subtypes, the remaining seven segments displayed similar rates of nucleotide substitutions, thus suggesting that these two strains have a similar predisposition to acquire mutations.

4.5. Divergences of viral evolution

This parallel evolution may be observed only with viruses with similar virulence. Indeed, our study suggests that the stochastic appearance of amino acid substitutions affecting the virus virulence may completely change the direction of viral evolution, as observed during the H7N1 epidemic. According to Galvani (2003) virulence may arise as a consequence of within-host competition among variants or as a strategy to maximize the transmission rate, although the host survival and, hence, the persistence of the virus in the host will be reduced. This may also explain why the H7N3 viruses did not manage to evolve into a highly pathogenic form. Indeed, this theory suggests a trade-off between the rate of transmission, which increases with the virulence, and the persistence of the virus in the host (Galvani, 2003). According to this idea, vaccination during the H7N3 epidemic may have reduced the opportunities of the virus to be transmitted, thus selecting the virus with a lower virulence but a higher capacity to persist in the host. Moreover, at the time of the H7N1 epidemic, the LPAI strains were not subject to compulsory control policy, while during the H7N3 epidemic control measures were immediately adopted in accordance with the legislative decree 28 September 2000 of the Italian Ministry of Health. This may have further contributed to reduce the probability of the H7N3 virus to evolve into a more virulent strain.

5. Conclusions

The unprecedented opportunity offered by the H7N1 and the H7N3 datasets to explore the evolutionary strategies of distinct epidemics, which occurred in similar circumstances, has shown how different strains may adopt similar evolutionary strategies within the constraints of similar ecological conditions. Although they are capable of acquiring a large amount of different mutations during their evolution, the few changes that provide the highest fitness advantage in a specific environment have the highest probability of reaching fixation independently of the genetic lineages. However, this study has also highlighted the difficulties in predicting either the acquisition of specific mutations or changes in the virus virulence, which depend on several ecological factors and strain characteristics. This underlies the need to constantly implement sustainable surveillance programmes in poultry to promptly identify viruses with the potential to acquire highly pathogenic properties and to control their spread in a timely manner.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under Grant Agreement n°278433-PREDEMICS and Epi-SEQ (research project supported under the 2nd joint call for transnational research projects by EMIDA ERA-NET [FP7 project nr 219235]). Alice Fusaro acknowledges the receipt of a fellowship from the OECD Co-operative Research Programme: Biological Resource Management for Sustainable Agricultural Systems in 2013.

The authors would like to acknowledge Enrico Massimiliano Negrisolo (University of Padua, Padua, Italy) for his support with phylogenetic analyses, Viviana Valastro for her technical support and Francesca Ellero for providing language help.

This study was conducted in the framework of the Doctoral school in Veterinary Science at the University of Padua (Alice Fusaro).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2015.05.020.

References

Aamir, U.B., Naeem, K., Ahmed, Z., Obert, C.A., Franks, J., Krauss, S., Seiler, P., Webster, R.G., 2009. Zoonotic potential of highly pathogenic avian H7N3 influenza viruses from Pakistan. Virology 390, 212–220. http://dx.doi.org/10.1016/j.virol.2009.05.008.

Abdelwhab, E.M., Veits, J., Mettenleiter, T.C., 2014. Prevalence and control of H7 avian influenza viruses in birds and humans. Epidemiol. Infect. 142, 896–920. http://dx.doi.org/10.1017/S0950268813003324.

Banks, J., Speidel, E.S., Moore, E., Plowright, L., Piccirillo, A., Capua, I., Cordioli, P., Fioretti, A., Alexander, D.J., 2001. Changes in the haemagglutinin and the neuraminidase genes prior to the emergence of highly pathogenic H7N1 avian influenza viruses in Italy. Arch. Virol. 146, 963–973.

Bataille, A., van der Meer, F., Stegeman, A., Koch, G., 2011. Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic. PLoS Pathog. 7, e1002094. http://dx.doi.org/10.1371/journal.ppat.1002094.

Beato, M.S., Xu, Y., Long, L.-P., Capua, I., Wan, X.-F., 2014. Antigenic and genetic evolution of low-pathogenicity avian influenza viruses of subtype H7N3 following heterologous vaccination. Clin. Vaccine Immunol. 21, 603–612. http://dx.doi.org/10.1128/CVI.00647-13.

Boni, M.F., Posada, D., Feldman, M.W., 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176, 1035–1047. http://dx.doi.org/10.1534/genetics.106.068874.

Bush, R.M., Bender, C.A., Subbarao, K., Cox, N.J., Fitch, W.M., 1999. Predicting the evolution of human influenza A. Science 286, 1921–1925. http://dx.doi.org/10.1126/science.286.5446.1921.

Campitelli, L., Mogavero, E., De Marco, M.A., Delogu, M., Puzelli, S., Frezza, F., Facchini, M., Chiapponi, C., Foni, E., Cordioli, P., Webby, R., Barigazzi, G., Webster, R.G., Donatelli, I., 2004. Interspecies transmission of an H7N3 influenza virus from wild birds to intensively reared domestic poultry in Italy. Virology 323, 24–36. http://dx.doi.org/10.1016/j.virol.2004.02.015.

Capua, I., Marangon, S., 2007. The use of vaccination to combat multiple introductions of notifiable avian influenza viruses of the H5 and H7 subtypes between 2000 and 2006 in Italy. Vaccine 25, 4987–4995. http://dx.doi.org/10.1016/j.vaccine.2007.01.113.

Cattoli, G., Fusaro, A., Monne, I., Coven, F., Joannis, T., El-Hamid, H.S.A., Hussein, A.A., Cornelius, C., Amarin, N.M., Mancin, M., Holmes, E.C., Capua, I., 2011a. Evidence for differing evolutionary dynamics of A/H5N1 viruses among countries applying or not applying avian influenza vaccination in poultry. Vaccine 29, 9368–9375. http://dx.doi.org/10.1016/j.vaccine.2011.09.127.

Cattoli, G., Milani, A., Temperton, N., Zecchin, B., Buratin, A., Molesti, E., Aly, M.M., Arafa, A., Capua, I., 2011b. Antigenic drift in H5N1 avian influenza virus in poultry is driven by mutations in major antigenic sites of the hemagglutinin molecule analogous to those for human influenza virus. J. Virol. 85, 8718–8724. http://dx.doi.org/10.1128/JVI.02403-10.

Chen, Y., Liang, W., Yang, S., Wu, N., Gao, H., Sheng, J., Yao, H., Wo, J., Fang, Q., Cui, D., Li, Y., Yao, X., Zhang, Y., Wu, H., Zheng, S., Diao, H., Xia, S., Zhang, Y., Chan, K.-H., Tsoi, H.-W., Teng, J.L.-L., Song, W., Wang, P., Lau, S.-Y., Zheng, M., Chan, J.F.-W., To, K.K.-W., Chen, H., Li, L., Yuen, K.-Y., 2013. Human infections with the emerging avian influenza A H7N9 virus from wet market poultry: clinical analysis and characterisation of viral genome. Lancet 381, 1916–1925. http://dx.doi.org/10.1016/S0140-6736(13)60903-4.

Delport, W., Poon, A.F.Y., Frost, S.D.W., Kosakovsky Pond, S.L., 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26, 2455–2457. http://dx.doi.org/10.1093/bioinformatics/btq429.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214. http://dx.doi.org/10.1186/1471-2148-7-214.

Dundon, W.G., Milani, A., Cattoli, G., Capua, I., 2006. Progressive truncation of the non-structural 1 gene of H7N1 avian influenza viruses following extensive circulation in poultry. Virus Res. 119, 171–176. http://dx.doi.org/10.1016/j.virusres.2006.01.005.

Galvani, A.P., 2003. Epidemiology meets evolutionary ecology. Trends Ecol. Evol. 18, 132–139. http://dx.doi.org/10.1016/S0169-5347(02)00050-2.

Giannecchini, S., Clausi, V., Di Trani, L., Falcone, E., Terregino, C., Toffan, A., Cilloni, F., Matrosovich, M., Gambaryan, A.S., Bovin, N.V., Delogu, M., Capua, I., Donatelli, I., Azzi, A., 2010. Molecular adaptation of an H7N3 wild duck influenza virus following experimental multiple passages in quail and turkey. Virology 408, 167–173. http://dx.doi.org/10.1016/j.virol.2010.09.011.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704. http://dx.doi.org/10.1080/10635150390235520.

Holmes, E.C., Lipman, D.J., Zamarin, D., Yewdell, J.W., 2006. Comment on "large-scale sequence analysis of avian influenza isolates". Science 313. http://dx.doi.org/10.1126/science.1131729, 1573b-1573b.

Iqbal, M., Yaqub, T., Reddy, K., McCauley, J.W., 2009. Novel genotypes of h9n2 influenza A viruses isolated from poultry in Pakistan containing NS genes similar to highly pathogenic H7N3 and H5N1 viruses. PLoS One 4, e5788. http://dx.doi.org/10.1371/journal.pone.0005788.

Kaverin, N.V., Rudneva, I.A., Govorkova, E.A., Timofeeva, T.A., Shilov, A.A., Kochergin-Nikitsky, K.S., Krylov, P.S., Webster, R.G., 2007. Epitope mapping of the hemagglutinin molecule of a highly pathogenic H5N1 influenza virus by using monoclonal antibodies. J. Virol. 81, 12911–12917. http://dx.doi.org/10.1128/JVI.01522-07.

Kosakovsky Pond, S.L., 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol. Evol. 23, 1891–1901. http://dx.doi.org/10.1093/molbev/msl051.

Lebarbenchon, C., Stallknecht, D.E., 2011. Host shifts and molecular evolution of H7 avian influenza virus hemagglutinin. Virol. J. 8, 328. http://dx.doi.org/10.1186/1743-422X-8-328.

Li, J., Wang, Y., Liang, Y., Ni, B., Wan, Y., Liao, Z., Chan, K., Yuen, K., Fu, X., Shang, X., Wang, S., Yi, D., Guo, B., Di, B., Wang, M., Che, X., Wu, Y., 2009. Fine antigenic variation within H5N1 influenza virus hemagglutinin's antigenic sites defined by yeast cell surface display. Eur. J. Immunol. 39, 3498–3510. http://dx.doi.org/10.1002/eji.200939532.

Lunter, G., Goodson, M., 2011. Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. Genome Res. 21, 936–939. http://dx.doi.org/10.1101/gr.111120.110.

Maddison, W.P., Maddison, D.R., 2014. Mesquite: A Modular System for Evolutionary Analysis. Version 3.0 <http://mesquiteproject.org..

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P., 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26, 2462–2463. http://dx.doi.org/10.1093/bioinformatics/btq467.

Martin, D., Rybicki, E., 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16, 562–563. http://dx.doi.org/10.1093/bioinformatics/16.6.562.

Monne, I., Fusaro, A., Nelson, M.I., Bonfanti, L., Mulatti, P., Hughes, J., Murcia, P.R., Schivo, A., Valastro, V., Moreno, A., Holmes, E.C., Cattoli, G., 2014. Emergence of a highly pathogenic avian influenza virus from a low-pathogenic progenitor. J. Virol. 88, 4375–4388. http://dx.doi.org/10.1128/JVI.03181-13.

Mulatti, P., Bos, M.E.H., Busani, L., Nielen, M., Marangon, S., 2010. Evaluation of interventions and vaccination strategies for low pathogenicity avian influenza: spatial and space–time analyses and quantification of the spread of infection. Epidemiol. Infect. 138, 813. http://dx.doi.org/10.1017/S0950268809991038.

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S.L., Scheffler, K., 2013. FUBAR: a fast, unconstrained Bayesian AppRoximation for inferring selection. Mol. Biol. Evol. 30, 1196–1205. http://dx.doi.org/10.1093/molbev/mst030.

Padidam, M., Sawyer, S., Fauquet, C.M., 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265, 218–225. http://dx.doi.org/10.1006/viro.1999.0056.

Pond, S.L.K., Frost, S.D.W., 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21, 2531–2533. http://dx.doi.org/10.1093/bioinformatics/bti320.

Röhm, C., Horimoto, T., Kawaoka, Y., Süss, J., Webster, R.G., 1995. Do hemagglutinin genes of highly pathogenic avian influenza viruses constitute unique phylogenetic lineages? Virology 209, 664–670. http://dx.doi.org/10.1006/viro.1995.1301.

Schmidt, H.A., Petzold, E., Vingron, M., von Haeseler, A., 2003. Molecular phylogenetics: parallelized parameter estimation and quartet puzzling. J. Parallel Distributed. Comput. 63, 719–727. http://dx.doi.org/10.1016/S0743-7315(03)00129-1.

Smith, J., 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. 34. http://dx.doi.org/10.1007/BF00182389.

Spackman, E., Senne, D.A., Davison, S., Suarez, D.L., 2003. Sequence analysis of recent H7 avian influenza viruses associated with three different outbreaks in commercial poultry in the United States. J. Virol. 77, 13399–13402. http://dx.doi.org/10.1128/JVI.77.24.13399-13402.2003.

Stamatakis, A., 2006a. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690. http://dx.doi.org/10.1093/bioinformatics/btl446.

Stamatakis, A., 2006b. Phylogenetic models of rate heterogeneity: a high performance computing perspective. IEEE, 8. http://dx.doi.org/10.1109/IPDPS.2006.1639535.

Stevens, J., Blixt, O., Tumpey, T.M., Taubenberger, J.K., Paulson, J.C., Wilson, I.A., 2006. Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. Science 312, 404–410. http://dx.doi.org/10.1126/science.1124513.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739. http://dx.doi.org/10.1093/molbev/msr121.

Xiong, X., Martin, S.R., Haire, L.F., Wharton, S.A., Daniels, R.S., Bennett, M.S., McCauley, J.W., Collins, P.J., Walker, P.A., Skehel, J.J., Gamblin, S.J., 2013. Receptor binding by an H7N9 influenza virus from humans. Nature 499, 496–499. http://dx.doi.org/10.1038/nature12372.

Zhou, B., Donnelly, M.E., Scholes, D.T., St. George, K., Hatta, M., Kawaoka, Y., Wentworth, D.E., 2009. Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and swine origin human influenza a viruses. J. Virol. 83, 10309–10313. http://dx.doi.org/10.1128/JVI.01109-09.

CHAPTER 2

Ultra-Deep Sequencing Data Reveal Unexpected Inter-farm Transmission Dynamics During a Highly Pathogenic Avian Influenza Epidemic

Fusaro A, Tassoni L, Milani A, Hughes J, Salviato A, Murcia PR, Massi P, Bonfanti L, Marangon S, Cattoli G, Monne I.

Submitted to Journal of Virology

ABSTRACT

Next Generation Sequencing technology is now being increasingly applied to study the within and between host population dynamics of viruses. However, information on avian influenza virus evolution and transmission during a naturally occurring epidemic is still limited. Here, we use deep sequencing data obtained from clinical samples collected from five industrial holdings and a backyard farm infected during the 2013 highly pathogenic avian influenza (HPAI) H7N7 epidemic in Italy to unravel i) the epidemic virus population diversity, ii) the evolution of virus pathogenicity, and iii) the pathways of viral transmission between different holdings and sheds. We show a high level of genetic diversity of the HPAI H7N7 viruses within a single farm as a consequence of separate bottlenecks and founder effects. In particular, we identified the co-circulation in the index case of two viral strains showing a different insertion at the Hemagglutinin cleavage site, as well as nine nucleotide differences at the consensus level and 92 minority variants. To assess inter-farm transmission, we combined epidemiological and genetic data and identified the index case as the major source of the virus, suggesting the occurrence of multiple routes of spread of different viral haplotypes from the index farm to the other industrial holdings, probably at different time points and with different transmission modes. Our results revealed inter-farm transmission dynamics that the epidemiological data alone could not unravel and demonstrated that delay in the disease detection and stamping out was the major cause of the emergence and the spread of the HPAI strain.

IMPORTANCE

The within and between host evolutionary dynamics of a highly pathogenic avian influenza (HPAI) strain during a naturally occurring epidemic is currently poorly understood. Here, we perform for the first time an in-depth sequence analysis of all the samples collected during a HPAI epidemic and demonstrate the importance to complement outbreak investigations with genetic data to reconstruct the transmission dynamics of the viruses and to evaluate the within and between farms genetic diversity of the viral population. We show that the evolutionary transition from the low pathogenic to the highly pathogenic form occurred within the first infected flock where we identified haplotypes with hemagglutinin cleavage site of different lengths. We also identify the index case as the major source of virus, indicating that prompt application of depopulation measures is essential to limit virus spread to other farms.

INTRODUCTION

Today, Next Generation Sequencing (NGS) techniques allow the investigation of viral population dynamics at any level (from within host to the epidemiological scale) with high resolution. In addition, NGS, can be used to identify low frequency variants which may be selected for and transmitted to other hosts. Avian influenza viruses (AIVs) exist in the host as populations of genetically related variants (1). The rate at which genetic diversity is generated within the host, the competitive replication ability of each variant, and the occurrence of genetic drift and of bottleneck events are some of the processes that drive virus evolution.

NGS has been applied on avian influenza virus to i) characterize the emergence of mutations in the viral subpopulations associated to an increased virulence (2, 3) or to adaptation to new hosts, (4, 5) ii) to study

genetic bottlenecks upon transmission events (6, 7); iii) to investigate the dynamics of virus evolution during outbreaks in poultry (8) ; and iv) to identify co-infection with different subtypes (9). However, application of high throughput sequencing for the exploration of avian influenza virus evolution and transmission during a naturally occurring epidemic is still limited, making the interpretation of genomic data collected from outbreaks far from straightforward.

Between August 14th and September 3rd of 2013, thirteen years after the last highly pathogenic avian influenza (HPAI) outbreak, Italy experienced a new avian influenza epidemic caused by a HPAI virus of the H7N7 subtype, which infected five industrial poultry holdings, four of which belonged to a large vertically integrated layer company, and one backyard flock (10). Detailed information on these outbreaks has been provided in a previous study (10). The epidemiological investigation indicated that the contact between free-range hens and wild waterfowl in the first affected holding may have favoured the introduction of a low pathogenic avian influenza (LPAI) virus, which rapidly mutated into a HP form within the infected sheds (10) through the acquisition of multiple basic amino acids at the hemagglutinin (HA) cleavage site, which is considered as being the major molecular determinant of an HPAI virus (11).

Here we used NGS to unravel the virus population diversity and the evolution of virus pathogenicity within the affected poultry farms. We also determined the transmission pathways of the H7N7 virus between different holdings and sheds during the course of the epidemic by combining deep sequencing and epidemiological data.

MATERIALS AND METHODS

Viruses

Fourteen positive clinical samples (organs and swabs) were collected between August 13th and September 3rd 2013 from each infected shed of the five industrial farms and a backyard flock, counting for all the cases detected during the epidemic. Epidemiological information, including collection date, sample type (swabs, organs), farm and shed of origin, number of birds present in each farm at the time of the forfeiture and depopulation date, is available in Table 1.

Generation of viral sequence data

Total RNA was purified from the 14 infected clinical samples using the Nucleospin RNA kit (Macherey–Nagel, Duren, Germany). Complete influenza A virus genomes were amplified with the SuperScript III One-Step RT-PCR system with Platinum Taq High Fidelity (Invitrogen, Carlsbad, CA) using one pair of primers complementary to the conserved elements of the influenza A virus promoter as described in (12). PCR products were visualized on a 0.7% agarose gel. Sequencing libraries were obtained using Nextera DNA XT Sample preparation kit (Illumina) following the manufacturer's instructions and quantified using the Qubit dsDNA High Sensitivity kit (Invitrogen, USA). The average fragment length was determined using the Agilent High Sensitivity Bioanalyzer Kit. Finally the indexed libraries were pooled in equimolar concentrations and sequenced in multiplex for 250 bp paired-end on Illumina MiSeq, according to the manufacturer's instructions.

Quality trimming, assembly and SNP detection

Illumina MiSeq reads were inspected using FASTQC to assess the quality of data. Fastq files were cleaned with PRINSEQ and Trim Galore to remove low quality bases at the 5' and 3'end of each read and to exclude reads with a Phred quality score below 30 and shorter than 80 nucleotides. The filtered, trimmed reads were aligned to the eight gene segments of A/chicken/Italy/13VIR4727-11/2013, for which the consensus genome were previously obtained using Sanger method (data not shown), using BWA-MEM, an accurate aligner for paired-end reads longer than 70 (http://arxiv.org/abs/1303.3997v2). The BAM alignment files were parsed using the diversiTools program (http://josephhughes.github.io/btctools/) to determine the average base-calling error probability and to identify the frequency of polymorphisms at each site relative to the reference used for the alignment. In order to minimize artefacts introduced through RT-PCR and sequencing errors, for all the analysis conducted throughout this study we considered only polymorphisms with a frequency above 2% identified in positions with a minimum coverage of 500. This choice was based on the comparison of data obtained from two technical replicates of three samples (4541-8, 4541-9, 4541-34), sequenced on two different Illumina sequencing machines (MiSeq), starting from two separate libraries obtained from the same extracted RNA. This threshold should guarantee the exclusion of 99.6% of the errors from our deep sequencing data (S1 Fig).

For each replicate, only the assembled genome with the highest coverage was used in the following analyses.

Genetic distance, entropy and transmission tree

We computed the genetic distance between the complete genome of all pairs of individuals (S1 and S2) using the following formula: $d = \frac{1}{N}\sum_{i=1}^{N}(|f A_{iS1} - fA_{iS2}| + |fC_{iS1} - fC_{iS2}| + |fT_{iS1} - fT_{iS2}| + |fG_{iS1} - fG_{iS2}|)^2$,

where fA_{iS1}, fC_{iS1}, fT_{iS1}, fG_{iS1} are the frequencies of nucleotide A, C, T and G at position *i* in the two samples and N is the length of the sequence. This matrix was used to compute a neighbour-joining phylogenetic tree using the web server T-REX (13). In addition, we combined the distance matrix and the collection dates to reconstruct the transmission tree of the H7N7 during the Italian outbreak, using SeqTrack (14), a graph-based approach particularly suitable to infer maximum parsimony genealogies of viruses in densely sampled disease outbreak. The *adegenet* (15) and *igraph* packages (16) for the R software were used to perform the analysis and to draw the network.

To measure the complexity of the viral populations within a sample, we calculated the Shannon entropy of each sample using the following equation:

$$E = -\frac{1}{N}\sum_{i=1}^{N}(f_{iA}\ln f_{iA} + f_{iG}\ln f_{iG} + f_{iT}\ln f_{iT} + f_{iC}\ln f_{iC})$$

where f_i is the frequency of the nucleotide A, T, G or C at position *i* and N is the total length of the genome.

Only nucleotides with a frequency above the 2% threshold identified in positions with a minimum coverage of 500 were included in this calculation.

Phylogenetic analyses

Consensus sequences of the complete genome of the 14 samples were aligned using MAFFT v. 7 (17) and compared with the most related sequences available in GenBank and in GISAID. Maximum likelihood (ML) phylogenetic trees were obtained for each gene segment using the best-fit general time reversible (GTR) model of nucleotide substitution with gamma-distributed rate variation among sites (with four rate categories, $\Gamma 4$) available in RAxML-MPI v.8.1.7 (18). To assess the robustness of individual nodes of the phylogeny, one hundred bootstrap replicates were performed. Phylogenetic trees were visualized with the program FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/).

The eight gene segments of the influenza virus genome were manually concatenated and the alignment was used to construct a phylogenetic network using the Median Joining method implemented in the program NETWORK 4.5 (http://www.fluxus-engineering.com) (19). This method uses a parsimony approach to reconstruct the relationships between highly similar sequences, and allows the creation of "median vectors", which represents unsampled sequences, that are used to connect the existing genotypes in the most parsimonius way. The parameter *epsilon* was set to 10 and the transition to transversion ratio to 3:1.

Nucleotide sequence accession numbers

MiSeq sequences were submitted to the NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra/) under accession numbers SRR3036850, SRR3036852, SRR3036854, SRR3036856, SRR3036860, SRR3036864, SRR3036910, SRR3036911, SRR3036914, SRR3036916, SRR3036917, SRR3036919, SRR3036920, SRR3036945. Consensus sequences of the 14 H7N7 viruses were submitted to GISAID under accession numbers EPI677984 to EPI678095.

RESULTS

Phylogenetic analysis of consensus sequences

To investigate influenza virus variation during the HPAI H7N7 epidemic, we sequenced the eight genomic segments for all the clinical samples received from each infected farm. The highest number of positive samples (8) was submitted from the three infected sheds (shed 2, 4 and 5) of the index case, while only one sample per infected shed was received from the remaining five outbreak sites, for a total of one or two samples per farm. Farms are labelled from 1 to 6, according to the collection date of the samples. Details of location, date of sample collection, farm characteristics, sample type and mean depth of coverage are reported in Table 1.

Our maximum likelihood phylogenetic analyses of the consensus sequences show that the fourteen HPAI H7N7 viruses form a distinct genetic group, defined by high bootstrap values (>96%) and long branches in all the eight phylogenies, suggesting the occurrence of a single viral introduction (Fig 1 and S1 to S7 Figs). In the HA and NA phylogenetic trees, they cluster with H7 viruses collected in Europe between 2009 and 2014. In particular, the HA gene segment of the Italian samples show the highest similarity (99.1-99.3%) with an LPAI H7N7 virus collected from a wild bird in Italy in 2014, for which only the HA sequence is available (Fig 1), while the NA gene segment display the highest identity (99-99.1%) with an H7N7 virus collected from

chicken in Netherlands (S1 Fig). In the phylogenies of the internal gene segments the Italian samples group with viruses of different subtypes circulating mainly among wild birds in Eurasian countries (S2 to S7 Figs).

High genetic variability of the first infected flock

Surprisingly, molecular analysis of the eight viruses collected from the index case shows the co-circulation of two highly pathogenic strains with a different insertion at the HA cleavage site compared to a H7 LP virus. Specifically, sequences of the two viruses from shed 5 (4541-9 and 4541-34) show an insertion of 6 nucleotides, while the remaining samples identified in sheds 2 and 4 possess a longer cleavage site with a nine nucleotide insertion.

To better understand the evolution of the pathogenicity of the H7N7 viruses within the first infected flock, we focused our analysis on the deep sequencing data of the HA cleavage site. The sequencing coverage in this genetic region ranges from 4445 for the sample 4541-7 to 23511 for the sample 4541-34. We did not identify any reads showing the cleavage site typical of a LPAI strain. 99.9% of the reads of the two samples from shed 5 (named for clarity V+6) posses a cleavage site with an insertion of six nucleotides, with only a few reads containing an insertion of three, five and nine nucleotides (Table 2). While, 99.7% to 99.9% of the reads of viruses from shed 2 (named V+9) have an insertion of nine nucleotides, with only a few minority variants showing an insertion of six, seven or eight nucleotides (Table 2). On the other hand, in one of the samples from shed 4 (4541-33) we identified a mixed population with both type of cleavage sites displaying an insertion of nine (95.7%) and six (4.1%) nucleotides.

Similarly to the samples from shed 2, the majority (from 99.9% to 100%) of the viral population of the subsequent outbreaks possesses the longer cleavage site, suggesting that this variant (V+9) may have a higher fitness advantage.

Besides the cleavage site, the samples V+9 collected from shed 2 and 4 of the first infected farm can be distinguished from the two samples from shed 5 by nine nucleotide signatures, which resulted in three amino acid changes (PA Q116R, PA C631G, NS1 R118K). These signatures are maintained in all samples identified in the subsequent outbreaks, suggesting that only viruses from sheds 2 and 4 of the index case were transmitted to the other five farms (Fig 2). In addition, we identified one non-synonymous mutation at position 130 of the M2 gene, responsible of the amino acid substitution D44N, which is shared between the V+6 viruses and the samples 4527-11 from shed 2 of farm 1, 4603 from farm 2, 4678 from farm 3 and 5091 from farm 5 (Fig 2). However, whether this mutation emerged by chance in the shed 2 virus of the index case and was then transmitted to the other outbreaks or whether it was acquired by the V+9 samples through a reassortment event cannot be assessed.

To determine whether the shed 5 viruses (V+6) were the progenitors of the variant V+9, we examined the presence of the nine signature mutations as minority variants in the analysed samples. None of the mutations typical of the V+9 viruses were already present in shed 5 viruses (V+6) with a frequency higher than 2% (the frequency threshold used in this study, see the Materials and Methods section for details). Similarly, none of the mutations characteristic of V+6 was identified in the subpopulations of the V+9 samples, except for the virus from shed 4 of the index case (4541-33), which, besides the shorter cleavage site, possessed subpopulations containing all the mutations distinctive of V+6 variant, with a frequency ranging from 3% to 9%, confirming the presence of a mixed population (V+6 and V+9) (Fig 3).

Genetic diversity of H7N7 viruses

Overall, we observed mutations at 185 sites (excluding the HA cleavage site) distributed among the eight gene segments, of which 111 are non-synonymous and 74 synonymous. Specifically, a total of 35 consensus-level nucleotide substitutions are recovered along the entire genome, defining 11 different genomes, five of which identified within the first infected farm (Fig 2). The PB2 gene, with a total of ten nucleotide variants (8 synonymous and 2 non-synonymous), is the most polymorphic segment at the consensus level. Notably, 13 out of 35 mutations distributed along twelve proteins (HA, NA, PB2, PB1, PB1-F2, PA, PA-X, NP, M1, M2, NS1 and NS2) are non-synonymous, with the PA protein showing the highest number of amino acid variations (4) (Fig 2).

Besides these consensus-level variant sites, our deep sequencing analysis identifies 209 minority variants in 151 sites (97 non-synonymous and 54 synonymous) with a frequency ranging from 2% to 49.8% (Fig 3). The virus collected from shed 4 of the index case (4541-33), which displayed a mixed population of V+6 and V+9, and the sample 4603 collected from farm 2, comprise the highest number of minority variants (respectively, 40 and 41). On the contrary, we did not detect any subpopulations in the samples 4541-34 and 4541-9. No correlation between the number of variants and the type of samples used for the analysis (pool, organs or swab) was observed.

We measured the complexity of the viral population of each sample using Shannon entropy (represented by the size of the circles in Fig 4). In the first infected flock, entropy measures fluctuate considerably: the lowest values are observed for viruses from the shed 5 (V+6), suggesting that these samples (4541-9 and 4541-34) had recently experienced a narrow bottleneck and had not recovered from the loss of complexity. Conversely, viruses from shed 2 show intermediate values of entropy, while samples 4541-33 from shed 4 of the index case and 4603 from farm 2 displayed the highest entropy level, consistent with the high genetic diversity observed across their genomes. This finding suggests that a viral strain can evolve independently within separate sheds, going through bottlenecks of different intensity.

Minority variants transmitted between sheds and farms

Focusing our analysis of the first infected flock, we observed that only a few mutations were shared at a shed and farm level, while the majority of the minor changes were unique to individual samples. Specifically, at the shed level we detected 44 minority changes in viruses from shed 2, of which 22 are found in individual samples and not shared with others, and 48 in viruses from shed 4, of which 37 are identified in single samples. Similarly, at the farm scale we counted 92 mutations, of which 12 are shared between 2-4 samples, while 59 minority variants were identified in single individuals (Fig 3).

Interestingly, five of these variants change markedly in frequency within the first infected farm becoming the majority viral population (fixed variants, highlighted with black arrows in Fig 3) and three of them were also transmitted or independently acquired by viruses collected from the other premises. Interestingly, four of these are non-synonymous mutations and cause changes at the protein level (NS M119T, M2 D44N, PA V100I, PB2 K574R).

We detected only seven minority variants (HA 1351A, M 942G, M 955G, PA 1251 G, PA 1748A, PB2 981G and NA 390A) transmitted between two or three farms. Interestingly, five of them result in amino acid

mutations (HA D451N, M2 D85G, PA R583Q, PB2 G327G, NA M130I), suggesting that most of the transmitted variants are associated with changes in viral fitness.

Transmission dynamics of the H7N7 virus

To assess the inter-farm transmission, a Median Joining phylogenetic network was inferred using the concatenated consensus sequences of the eight gene segments of the 14 analysed viruses (Fig 5). Within the first infected farm we identified five sequence genotypes (grey circles): one within shed 5, two within shed 2, and two in shed 4. Viruses from sheds 2 and 4 appear to be at the origin of the infection to the other farms, although one or two median vectors (red circles), which represent the lost ancestral sequences, separate them from viruses of the other holdings, except for the sample 5051-3 from farm 6, which appears to be a direct descendant of shed 2 viruses. Sequences from farms 2 to 6 grouped within two main clusters which shared a common ancestor (c1 and c2): c1 includes viruses collected from farms 2 (4603), 3 (4778) and 5 (5091), while c2 contains virus sequences from farms 4 (4774) and 6 (5051-1). Sequences within these two clusters are separated by 6 to 10 nucleotide differences, whereas 9-13 differences are observed between viruses of the two clusters. Therefore, the high number of mutations and median vectors identified between the analysed samples makes the relationship between sequences hard to determine. Our deep sequencing data may contribute to better understand this relationship. To this aim, first we inferred a neighbour-joining phylogenetic tree based on the distance matrix calculated from our NGS data, which confirmed the clustering identified by our network analysis (Fig 3). Then we used the distance matrix and the collection dates to reconstruct a transmission tree using the graph-based algorithm SeqTrack. This approach, which considers the sampled viruses as a fraction of the genealogy, is particularly suitable to infer the transmission pathway during disease outbreaks, where one strain can be the ancestor of another strain (Fig 4).

Despite 21 days passing from the first to the last outbreak, the inferred genealogy suggests that all but one of the outbreaks descend directly from shed 2 (V+9) of the index case. The only exception is represented by the virus (5091) collected from the backyard farm on September 2 (farm 5), which appears to have been infected directly by farm 3.

However, based on the number of shared mutations between the analysed sequences, we may speculate further scenarios. For example, sample 4603 from farm 2 shared two fixed mutations with samples 4678 (farm 3) and 5091 (farm 5) (group c2 of the network analysis), thus a transmission event from farm 2 to farm 3 cannot be excluded. Similarly, viruses 4774 and 5051-1 share 3 unique minority variants and 1 unique fixed mutation, making a transmission event between these two farms highly plausible. In addition, samples 4774 and 5051-1 share 1 fixed mutations and 3 minority variants (group c1 of the network analysis), and in turn they share 2 fixed mutations with the sample 5051-3. Although viruses 5051-1 and 5051-3 were collected from two different sheds of farm 6, we observed a relatively high nucleotide distance between them. Specifically, they show 7 and 14 nucleotide differences at the population and subpopulation level, respectively, although all the consensus level mutations were present as minority variants in the other sample (Fig 3). Thus, the occurrence of two separate introductions in farm 6 from the index case and/or farm 4 cannot be excluded.

Overall, these analyses indicate that shed 2 of the index case is the major source of the virus. An early strain (c1) appears to have spread from the first infected flock to farm 2 (19 August) and 3 (21 August) and then from farm 3 to the backyard farm 5 (2 September). Since farms 2 and 3 belong to different companies (circle colour in Fig 4) and are located 50 Km apart (map in Fig 5), it is more plausible that viruses with similar genetic characteristics were transmitted from the index case to both holdings. A later spread with a slightly different strain (c2) may have occurred from the first infected flock to farm 4 (27 August) and 6 (3 September). These two farms are located in the same area, with a distance of 3 Km, and belong to the same layer company as the first infected holding, thus an exchange of virus between them cannot be ruled out.

DISCUSSION

Acquisition of a virulent phenotype by H7 avian influenza viruses may have devastating consequences to the poultry industry and in some instance can create major human health issues, including the risk of generating a new pandemic strain (20). Despite the identification of multiple basic amino acids at the HA cleavage site as one of the most important molecular markers of virus pathogenicity, the mechanisms underlying the emergence, spread and evolution of HPAI during an epidemic are poorly understood and limited to few studies (3, 21). Here we performed for the first time a deep sequencing analysis of all the samples collected during a HPAI epidemic to evaluate the transmission dynamics and the within and between farms genetic diversity of the viral population.

We showed that the fourteen H7N7 Italian samples collected from six different farms form a cluster distinct to other Eurasian sequences for all the eight gene segments, suggesting the occurrence in the poultry population of a single viral introduction. The high similarity of the HA gene segment with a virus collected from a wild bird in Italy and the contact between free-range hens and wild waterfowl in the first infected farm (10), indicates that the LPAI progenitor strain may have been introduced from the wild bird population into the first infected holding, where it rapidly mutated into a HP form.

Despite our phylogenies suggesting a single viral introduction, we observed a high genetic variability of H7N7 between the different sheds of the first infected flock. In particular, at the consensus level, viruses collected from shed 5 possessed a shorter HA cleavage site and nine nucleotide differences compared to the viruses from sheds 2. This number of nucleotide substitutions is not compatible with the occurrence of different introductions, when usually a higher number of mutations are observed (22), but it can be explained by i) a rapid evolution of the virus following some bottleneck events, ii) independent evolution of the same virus within two separate sheds, or iii) the establishment of the infection starting from two different seeding variants of the same progenitor viral population. Nevertheless, our analysis of the mutation spectra of viral populations suggests that the two variants arose as a consequence of a founder event or a narrow population bottleneck. Indeed, the haplotype V+6, circulating in shed 5, was not identified in the viral subpopulations of shed 2 and similarly haplotype V+9, identified in shed 2, was not detected as a minority population in shed 5 animals. In addition, at the HA cleavage site of viruses from sheds 2 (V+9) and 5 (V+6) we identified only a total of 16 and 3 reads with an insertion respectively of six and nine nucleotides.

Entropy values obtained for the two viruses from shed 5 further supports this hypothesis. Samples founded by few viral particles should have low entropy, since the strong bottleneck drastically reduce the

diversity of the viral population. On the other hand, samples that experienced relatively loose bottlenecks should display higher entropy. Therefore, the low entropy values of the viruses from shed 5 indicate that they had recently experienced a narrow bottleneck. Conversely, viruses from shed 2 show intermediate entropy values, suggesting that i) they were founded by a larger seeding population, ii) they experienced a high-level of replication, or that iii) they had circulated within the shed for a longer period of time. This last option is supported by the identification of H7-specific antibodies in animals from this shed, but not in animals from sheds 4 and 5 (10), while the second hypothesis may be supported by the high number of dead birds found in shed 2 compared to the other sheds, considering the virulence of the two variants were equal (intravenous pathogenicity index of 3 for both variants, data not shown).

Surely, sequences of early viruses, might have helped us to provide a better characterization of the evolution of this strain within the index case. Indeed, the identification of H7-specific antibodies in animals from shed 2 and from the outer sheds 1 and 7, where no viruses were isolated (10), indicates that the virus had been circulating undetected within the farm before its identification, likely with a low pathogenic phenotype.

Our analysis of the transmission dynamics indicates that only one of the two variants (V+9), probably the one with the highest fitness advantage, was transmitted from the index case to the other farms. Four out of the six infected farms (farms 1, 2, 4, 6) belong to one large vertically integrated layer company (Fig 4), therefore virus dissemination might have occurred through shared equipment, human-mediated mechanical transport, and also through infected workers, as H7N7 virus was diagnosed for three humans involved in the control of the epidemic (23). The low number of shared mutations between farms (seven) suggests that the transmission depended on the dissemination of a few viral particles. However, the high frequency threshold (2%) used in this study to identify the minority variants and the scarce number of analysed samples for each farm need to be taken into consideration.

In the farms for which it was possible to sequence more than one sample (eight for farm 1 and two for farm 6), we identified the co-circulation in the same premise of different related variants and the possible occurrence of multiple introductions in the same holdings (ie. farm 6), which can be detected only through the sequencing of a larger number of samples. Moreover, the high number of median vectors identified between the analysed samples in our phylogenetic network reveals missing ancestral sequences from our analyses, which might have been detected with increased sampling. As a consequence, increasing the number of viruses sampled from each farm and also from the environment could increase the resolution of our inter-farm transmission dynamic.

We identify farm 1 as the major source for the spread of the virus to the other four industrial holdings, while the rural farm (farm 5) appears to have received the virus from the turkey farm (farm 3). Interestingly, this finding allowed the National authorities to demonstrate the occurrence of uncontrolled movements of birds from the infected turkey flock (farm 3), underlining the importance of genetic data to complement the outbreak investigation. Despite 21 days elapsing from the index case (August 13) to the last outbreak (September 3), the late depopulation date of the first infected flock (August 27) and the ability of the avian virus to persist in the environment (24), might explain the virus spread between these two holdings (1 and 6). In addition, results of our analysis of the transmission dynamics suggests that, despite farm 2 being located in close proximity to farms 4 and 6, transmission links are absent between these two premises. On the

contrary, the virus sampled from this farm appears to be more related to the virus from farms 1 and 3, located, respectively, 38 km and 36 km from farm 2.

This finding suggests that multiple introductions of different viral haplotypes occurred from farm 1 to the other farms, probably at different time points and with different transmission modes, ie. neighbourhood spread (i.e. farm 1 and 3), human-mediated transport among farms of the same company (ie. farm 1 and 2 or farm 1 and 4). These different means of viral diffusion have been observed also during other HPAI epidemics (21, 25) suggesting that long distance transmission events may play an important role for the virus dissemination into new areas.

Overall this study shows that analysis of deep sequencing data can complement epidemiological investigations, providing important insights and revealing unexpected dynamics on the inter-farm transmission network. Specifically, we demonstrated that the delay in the disease detection and stamping out in the index case might have been the major cause of the emergence and the spread of the HPAI strain. Epidemiological investigations did not recognize the central role of the first infected flock in the diffusion of the virus to most of the farms, and suggested an epidemiological link between farms 2, 5 and 6, which has not been confirmed by our data. In addition the epidemiological data alone was not sufficient to trace back the source of the virus detected in the rural farm (farm 5), which we demonstrated to be linked to the turkey farm (farm 3).

Moreover, we show that a farm can harbour a high level of heterogeneity, potentially caused either by separate bottlenecks and founder effects in the different sheds, or by multiple viral introductions from different sources. Hence, the importance during the control activities to collect and analyse several samples from each infected farm to provide a complete picture of the evolutionary process during an avian influenza epidemic.

FUNDING INFORMATION

This work was financially supported by the European projects Epi-SEQ (research project supported under the 2nd Joint Call for Transnational Research Projects by EMIDA ERA-NET [FP7 project no. 219235]) and PREDEMICS (research project supported by the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement n. 278433). PRM and JH are supported by the Medical Research Council of the United Kingdom (grant number G0801822).

ACKNOWLEDGEMENTS

The authors would like to acknowledge Gianpiero Zamperin for his technical support.

REFERENCES

1. **Manrubia SC, Escarmís C, Domingo E, Lázaro E.** 2005. High mutation rates, bottlenecks, and robustness of RNA viral quasispecies. Gene **347**(2):273–282.

2. **Iqbal M, Reddy KB, Brookes SM, Essen SC, Brown IH, McCauley JW**. 2014. Virus Pathotype and Deep Sequencing of the HA Gene of a Low Pathogenicity H7N1 Avian Influenza Virus Causing Mortality in Turkeys. PLoS ONE **9**(1):e87076. doi: 10.1371/journal.pone.0087076.

3. **Monne I, Fusaro A, Nelson MI, Bonfanti L, Mulatti P, Hughes J, Murcia PR, Schivo A, Valastro V, Moreno A, Holmes EC, Cattoli G**. 2014. Emergence of a Highly Pathogenic Avian Influenza Virus from a Low-Pathogenic Progenitor. J. Virol. **88**(8):4375–4388.

4. **Jonges M, Welkers MR, Jeeninga RE, Meijer A, Schneeberger P, R. Fouchier AM, de Jong MD, Koopmans M**. 2014. Emergence of the virulence-associated PB2 E627K substitution in a fatal human case of highly pathogenic avian influenza virus A(H7N7) infection as determined by Illumina ultra-deep sequencing. J. Virol. **88**(3):1694–1702.

5. **Poole DS, S. Yú, Y. Caì, Dinis JM, M. A. Müller, Jordan I, Friedrich TC, Kuhn JH, Mehle A**. 2014. Influenza A virus polymerase is a site for adaptive changes during experimental evolution in bat cells. J. Virol. **88**(21):12572–12585.

6. **Varble A, Albrecht RA, Backes S, Crumiller M, Bouvier NM, Sachs D, A. García-Sastre, tenOever BR**. 2014. Influenza A Virus Transmission Bottlenecks Are Defined by Infection Route and Recipient Host. Cell Host Microbe **16**(5):691–700.

7. **Wilker PR, Dinis JM, Starrett G, Imai M, Hatta M, Nelson CW, D. H. O'Connor, Hughes AL, Neumann G, Kawaoka Y, Friedrich TC**. 2013. Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. Nat Commun **4**:2636.

8. **Fusaro A, Tassoni L, Hughes J, Milani A, Salviato A, Schivo A, Murcia PR, Bonfanti L, Cattoli G, Monne I**. 2015. Evolutionary trajectories of two distinct avian influenza epidemics: Parallelisms and divergences. Infect Genet Evol **34**:457–466.

9. **Yu X, Jin T, Cui Y, Pu X, Li J, Xu J, Liu G, Jia H, Liu D, Song S, Yu Y, Xie L, Huang R, Ding H, Kou Y, Zhou Y, Wang Y, Xu X, Yin Y, Wang J, Guo C, Yang X, Hu L, Wu X, Wang H, Liu J, Zhao G, Zhou J, Pan J, Gao GF, Yang R, and Wang J**. 2014. Influenza H7N9 and H9N2 viruses: coexistence in poultry linked to human H7N9 infection and genome characteristics. J Virol. **88**(6):3423–3431.

10. **Bonfanti L, Monne I, Tamba M, Santucci U, Massi P, Patregnani T, L. Loli Piccolomini, Natalini S, Ferri G, Cattoli G, Marangon S**. 2014. Highly pathogenic H7N7 avian influenza in Italy. Vet. Rec. **174**(15):382–382.

11. **Bosch FX, Garten W, Klenk HD, Rott R**. 1981. Proteolytic cleavage of influenza virus hemagglutinins: primary structure of the connecting peptide between HA1 and HA2 determines proteolytic cleavability and pathogenicity of Avian influenza viruses. Virology **113**(2):725–735.

12. **Zhou B, Donnelly ME, Scholes DT, St George K, Hatta M, Kawaoka Y ,Wentworth DE**. 2009. Single-Reaction Genomic Amplification Accelerates Sequencing and Vaccine Production for Classical and Swine Origin Human Influenza A Viruses. J Virol. **83**(19):10309–10313.

13. **Boc A, Diallo AB, Makarenkov V**. 2012. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Res **40**:W573–W579. doi: 10.1093/nar/gks485.

14. **Jombart T, Eggo RM, Dodd PJ, Balloux F**. 2011. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity **106**(2):383–390.

15. **Jombart T**. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics **24**(11):1403–1405.

16. **Gabor C, Tamas N**. 2006. The igraph software package for complex network research. InterJournal, Complex Systems 1695 http://igraph.org

17. **Katoh K, Standley DM**. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol **30**(4):772–780.

18. **Stamatakis A**. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9):1312–1313.

19. **Bandelt HJ, Forster P, Röhl A**. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol **16**(1):37–48.

20. **Capua I, Marangon S**. 2006. Control of Avian Influenza in Poultry. Emerg. Infect. Dis. **12**(7):1319–1324.

21. **Bataille A, van der Meer F, Stegeman A, Koch G**. 2011. Evolutionary Analysis of Inter-Farm Transmission Dynamics in a Highly Pathogenic Avian Influenza Epidemic. PLoS Pathog. **7**(6):e1002094. doi: 10.1371/journal.ppat.1002094.

22. **Bouwstra R, Koch G, Heutink R, Harders F, van der Spek A, Elbers A, Bossers A**. 2015. Phylogenetic analysis of highly pathogenic avian influenza A(H5N8) virus outbreak strains provides evidence for four separate introductions and one between-poultry farm transmission in the Netherlands, November 2014. Euro Surveill **20**(26):21174.

23. **Puzelli S, Rossini G, Facchini M, Vaccari G, Di Trani L, Di Martino A, Gaibani P, Vocale C, Cattoli G, Bennett M, McCauley JW, Rezza G, Moro ML, Rangoni R, Finarelli AC, Landini MP, Castrucci MR, Donatelli I, Influenza Task Force**. 2014. Human infection with highly pathogenic A(H7N7) avian influenza virus, Italy, 2013. Emerg Infect Dis **20**(10):1745–1749.

24. **Brown JD, Swayne DE, Cooper RJ, Burns RE, Stallknecht DE**. 2007. Persistence of H5 and H7 avian influenza viruses in water. Avian Dis **51**(1 Suppl):285–289.

25. **Souris M, Gonzalez JP, Shanmugasundaram J, Corvest V, Kittayapong P**. 2010. Retrospective space-time analysis of H5N1 Avian Influenza emergence in Thailand. Int J Health Geogr **9**:3.

Table 1. Epidemiological information of the 14 samples collected during the HPAI H7N7 outbreak (TS= tracheal swabs).

Farm	Sample	Mean depth of coverage	Sample type	Farm type	Collection date	Province	Number of birds	Depopulation date
1 shed 2	4527-11	19354	Pool of 10 TS	Laying hen (industrial farm)	13 Aug 2013	Ferrara	128000	27 Aug 2013
	4527-12	36772	Pool of 10 TS					
	4541-7	24696	Organ pool					
	4541-32	53292	Kidney					
1 shed 4	4541-8	34018	Organ pool					
	4541-33	42661	Kidney					
1 shed 5	4541-9	23390	Organ pool					
	4541-34	58893	Kidney					
2	4603-1	43810	Pool of 10 TS	Laying hen (industrial farm)	19 Aug 2013	Bologna	584900	8 Sept 2013
3	4678	19893	Organ pool	Meat turkey (industrial farm)	21 Aug 2013	Ferrara	19850	27 Aug 2013
4	4774	31804	Organ pool	Laying hen (industrial farm)	27 Aug 2013	Bologna	121705	8 Sept 2013
5	5091	24510	Organ pool	Backyard flock	2 Sept 2013	Ferrara	3	5 Sept 2013
6	5051-1	46615	Trachea	Pullets (industrial farm)	3 Sept 2013	Bologna	98200	8 Sept 2013
	5051-3	48562	Trachea					

Table 2. Number of reads showing an insertion from 0 to 9 nucleotides at the HA cleavage site of the eight samples collected from three different sheds of the index case

N. nt insertion	SHED 5		SHED 4		SHED 2			
	4541-34	4541-9	4541-8	4541-33	4527-11	4527-12	4541-7	4541-32
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0
5	0	1	0	0	0	0	0	0
6	23509	14861	0	591	11	0	1	4
7	0	0	4	18	5	3	5	27
8	0	0	0	0	1	0	0	5
9	2	1	16587	13700	6660	13725	4439	22929

Fig 1. ML phylogenetic tree of the HA gene segment of H7 avian influenza viruses. HPAI H7N7 viruses collected during Italian epidemic are coloured according to the farm of collection: grey for farm 1, purple for farm 2, light blue for farm 3, yellow for farm 4, green for farm 5 and orange for farm 6. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only.

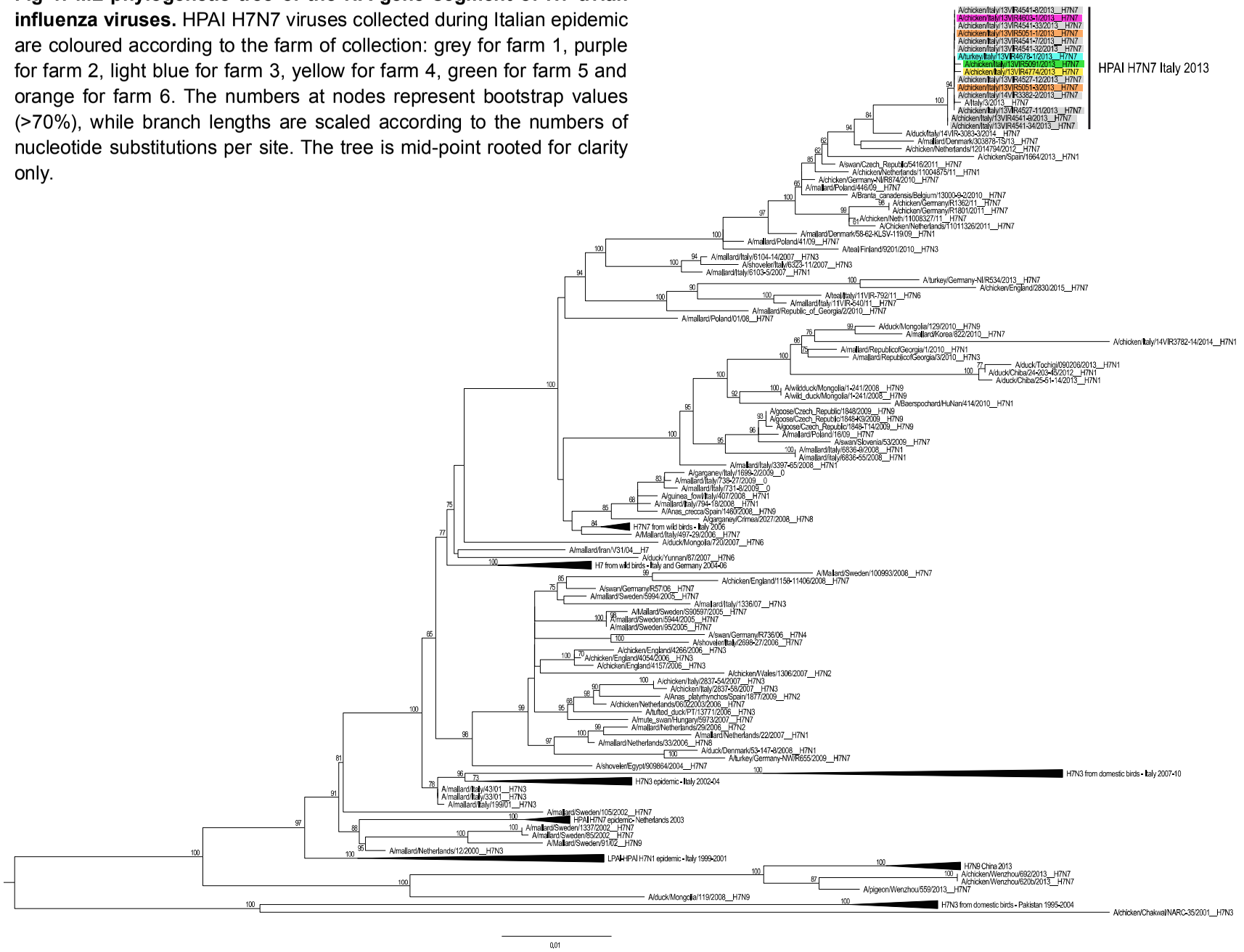


Fig 2. Consensus level nucleotide and amino acid differences among the complete genome of the 14 Italian H7N7 viruses. Each sample (column) is coloured according to the farm of collection: grey for farm 1, purple for farm 2, light blue for farm 3, yellow for farm 4, green for farm 5 and orange for farm 6. The farm and shed of belonging is indicated above the sample name. The nucleotide (NT) differences identified between each sample and the viruses from shed 5 of the index case (samples 4541-9 and 4541-34, column 1 and 2) are reported. Amino acid mutations (AA) are highlighted in red, while silent mutations are in black. *cleavage site

COLL. DATE	13 Aug 2013																19 Aug 2013		21 Aug 2013		27 Aug 2013		2 Sept 2013		3 Sept 2013			
GENE	4541-9 1-shed 5		4541-34 1-shed 5		4541-33 1-shed 4		4541-8 1-shed 4		4527-11 1-shed 2		4527-12 1-shed 2		4541-7 1-shed 2		4541-32 1-shed 2		4603 2		4678 3		4774 4		5091 5		5051-1 6-shed 16		5051-3 6-shed 15	
	NT	AA	NT	AA	NT	AA	NT	AA	NT	AA	NT	AA	NT	AA	NT	AA	NT	AA	NT	AA	NT	AA	NT	AA	NT	AA	NT	AA
HA	AA–AG	R/K/R	AA–AG	R/K/R	G471A AAGAGAG	E157E R/E/R	G471A AAGAGAG	E157E R/E/R	G471A AAGAGAG	E157E R/E/R	G471A AAGAGAG	E157E R/E/R	G471A AAGAGAG	E157E R/E/R	G471A AAGAGAG	E157E R/E/R	G471A G813A AAGAGAG	E157E G271G R/E/R	C149C/T G471A AAGAGAG	T50T/M E157E R/E/R	G471A AAGAGAG G1410A	E157E R/E/R E470E	C149 T G471A AAGAGAG	T50M E157E R/E/R	G471A AAGAGAG	E157E R/E/R	G471A AAGAGAG	E157E R/E/R
NA																	A1347G	L449L	A1347G	L449L			A1347G	L449L			C462T	F154F
PB2					A1721A/G A1851G T1860A	K574K/R K617K P620P	A1721G A1851G T1860A	K574R K617K P620P	A1851G T1860A	K617K P620P	A1851G T1860A	K617K P620P	A1851G T1860A	K617K P620P	A1851G T1860A	K617K P620P	A556G A1851G T1860A	T186A K617K P620P	C1044T T1542C/T A1851G T1860A	N348N S514S K617K P620P	T333C A1851G T1860A	Y111Y K617K P620P	T279C C1044T A1851G T1860A	P93P N348N K617K P620P	T333C G1281A/G A1851G T1860A	Y111Y R427R K617K P620P	C1793T A1851G T1860A	T598I K617K P620P
PB1																					T1215G C1882T	P405P L628L	C183T G1674T/G	T61T T558T				
PB1-F2																							C89T	P30L				
PA					G298G/A A347G G1251G/A T1891G	V100V/I Q116R L417L C631G	G298A A347G T1891G	V100I Q116R C631G	A347G T1891G	Q116R C631G	A347G G1251A/G T1891G	Q116R L417L C631G	A347G G1251A/G T1891G	Q116R L417L C631G	A347G G1251A/G T1891G	Q116R L417L C631G	A347G T1891G	Q116R C631G	A347G T1891G	Q116R C631G	A347G T1891G	Q116R C631G	A347G T1891G	Q116R C631G	A347G G767A T1891G	Q116R R255K C631G	A347G G1251A T1891G	Q116R L417L C631G
NP					G219A C316A	E73E R106R	G219A C316A	E73E R106R	G219A C316A	E73E R106R	G219A C316A	E73E R106R	G219A C316A	E73E R106R	G219A C316A	E73E R106R	G219A C316A	E73E R106R	G72A G219A C316A	E24E E73E R106R	G219A C316A	E73E R106R	G72A G219A C316A	E24E E73E R106R	G77A G219A C316A	R28K E73E R106R	G219A C316A	E73E R106R
M1																	G573A	Q191Q										
M2					G130A	D44N	G130A	D44N	G130G/A	D44D/N	G130A	D44N	G130A	D44N	G130A	D44N					G130A	D44N	G196A	E66K	G130A	D44N	G130A	D44N
NS1					G353A T356T/C G378A	R118K M119M/T K126K	G353A T356C G378A	R118K M119T K126K	G353A G378A	R118K K126K	G353A G378A	R118K K126K	G353A G378A	R118K K126K	G353A G378A	R118K K126K	G353A G378A	R118K K126K	G353A G378A	R118K K126K	G353A G378A	R118K K126K	G353A G378A	R118K K126K	G353A G378A	R118K K126K	G353A G378A	R118K K126K

Fig 3. Heat-map of the nucleotide frequency. Samples, coloured according to the farm of collection, are in column and positions showing nucleotide differences among the complete genomes of the 14 samples are in rows. The colour scale represents the nucleotide frequency according to the scale bar at the top of the figure. Black dots represent positions for which deep sequencing data were not available (coverage <500). Black arrows indicate the variants that change markedly in frequency within the first infected farm becoming the majority viral population. The dendrogram above the heatmap represents the neighbour-joining tree obtained from the distance matrix calculated from the deep sequencing data.

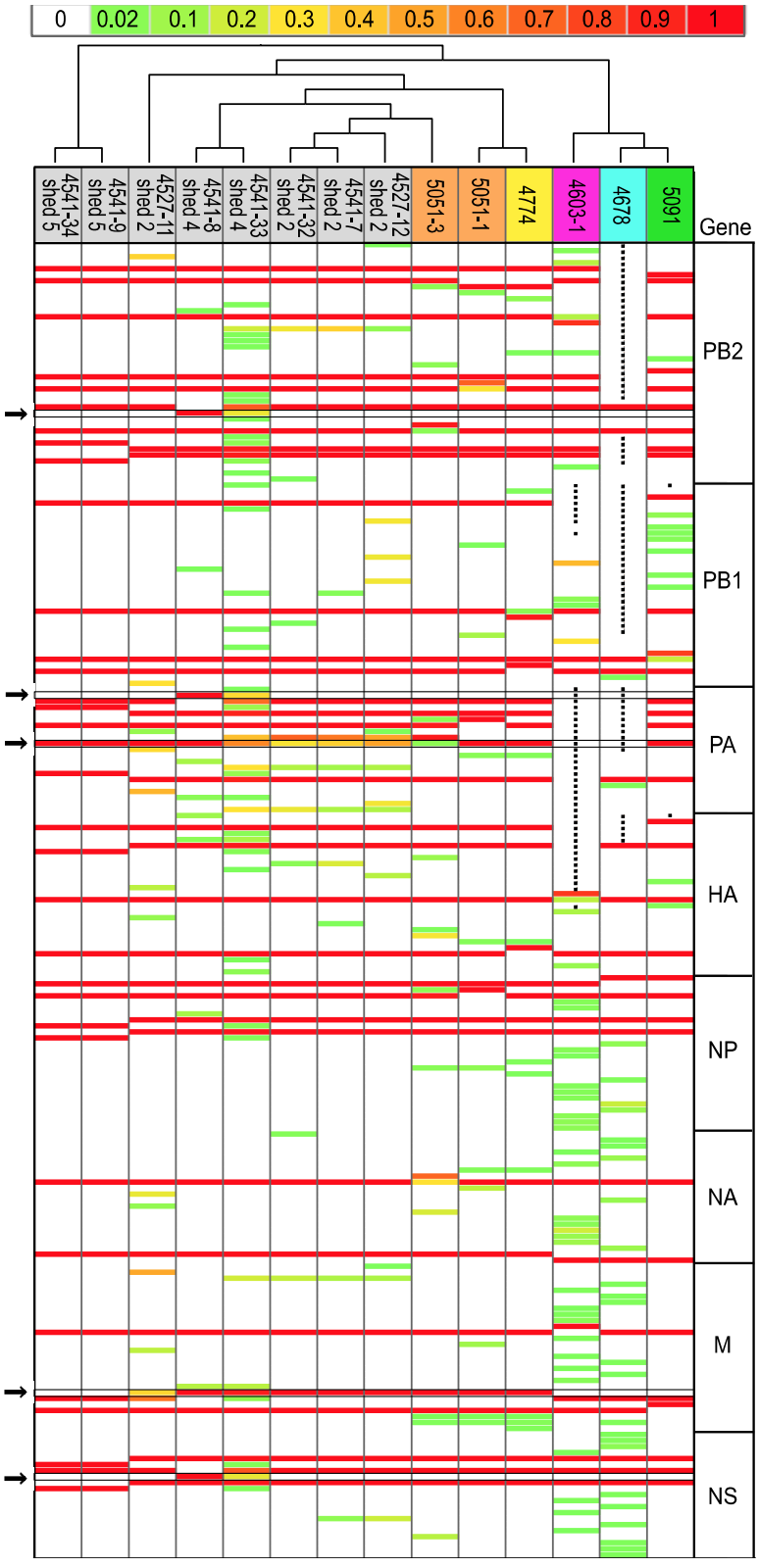


Fig 4. Transmission tree obtained from deep sequencing data. Each circle represents an individual sample. The size of the circles is proportional to the mean entropy value. The vertical axis represents the time of collection of each sample (samples in the same row belong to the same farm) and the numbers within the circle correspond to the farm number (1 to 6). Circle colours are assigned accordingly to the owner of the farm: farm 1, 2, 4 and 6 (green) belong to the same layer company, while the turkey farm 3 (purple) and the backyard farm (violet) belong to two different owners. Connecting arrows correspond to the results obtained from SeqTrack, while dashed lines are alternative hypotheses of transmission events formulated based on the number of shared mutations. Numbers over the lines are the genetic distance calculated from the deep sequencing data between the samples. Coloured area represents genetic groups identified based on the number of shared mutations and the results of both the neighbour-joining phylogenetic tree (Fig 3) and the network analysis (Fig 5).

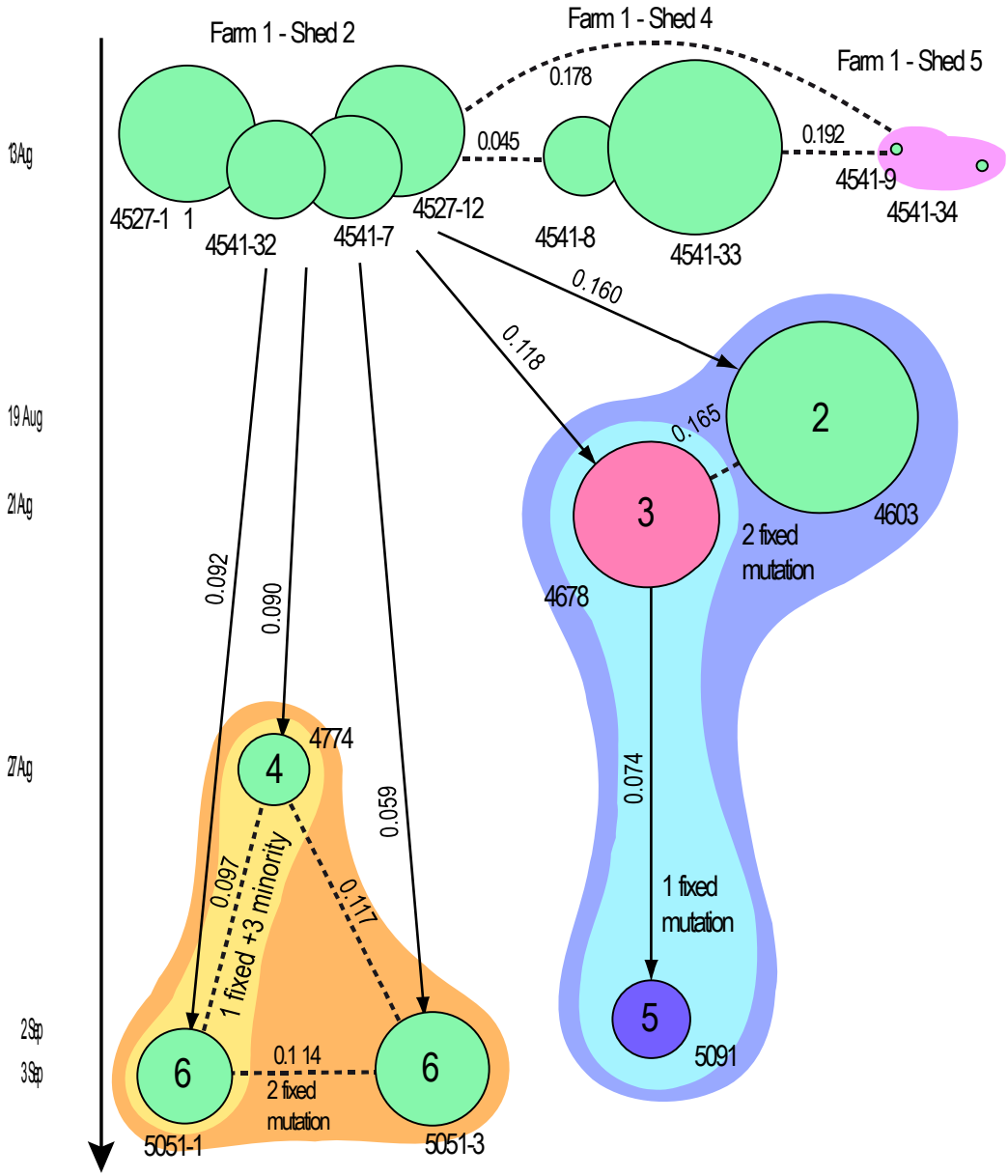
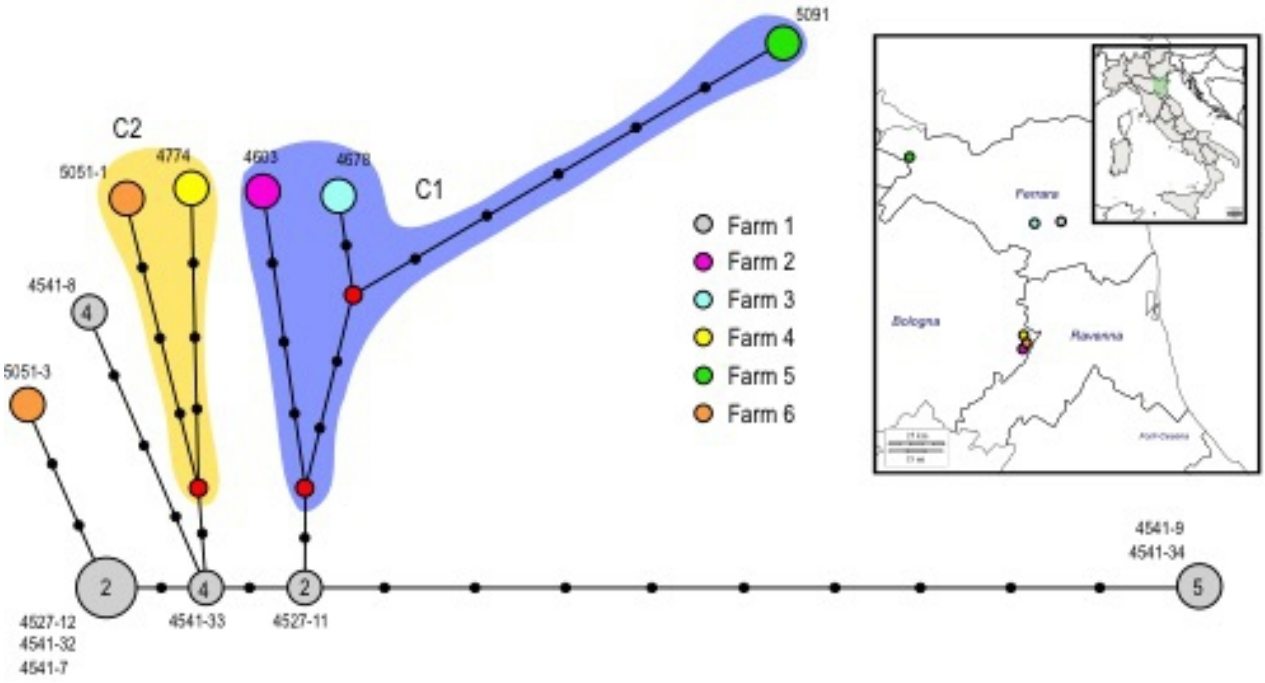


Fig 5. Median-joining phylogenetic network. A) The network was constructed from the consensus sequences of the eight concatenated gene segments. Each unique sequence genotype is represented by a circle sized relatively to its frequency in the dataset. Numbers next to the circles correspond to the samples showing that particular genotype, while the number within the circle represents the shed where the genotype was identified. Genotypes are coloured according to farm. Branches represent the shortest trees and black circles represent the number of nucleotide mutations that separate each node. Median vectors are shown as red circles. The violet and yellow shading represent the two identified genetic groups C1 and C2. B) The map shows the geographic position of the six infected farms.



CHAPTER 3

Vaccine immune pressure influences viral population complexity of avian influenza virus during infection

Milani A, Fusaro A, Bonfante F, Tassoni L, Salviato A, Mancin M, Mastrorilli E, Hussein A, Hassan M, Cattoli G, Monne I.

Going to be submitted by February 2016

Abstract

Vaccines are useful tools to control influenza A virus infection in poultry, but they need to be periodically reformulated to guarantee appropriate protection from infection and to avoid extensive viral circulation and replication, which could favour the emergence of new variants. In this study, an ultra-deep sequencing approach was used to inspect, characterize and follow the evolution of the viral population in infected vaccinated animals. High resolution hemagglutinin sequence data of H5N1 highly pathogenic avian influenza virus in chickens infected with two vaccines conferring different protection levels were analysed to examine the fine-scale genetic changes of viral populations. Our preliminary results suggested that the evolution of the viral population, as well as the abundance and minority variants heterogeneity could be influenced by the immune pressure conferred by vaccination.

Introduction

Influenza A virus is a zoonotic agent with a significant impact on both public health and poultry industry. Vaccination is a useful tool used worldwide to support intervention strategies, such as stamping out and biosecurity policies, in order to keep the infection under control and prevent the diffusion of avian influenza viruses in poultry (Lee and Suarez, 2005). However, as demonstrated in previous studies the use of a vaccine strain antigenically different from the circulating viruses or application of inadequate vaccine protocols may favour the antigenic drift and cause vaccination failure (Lee et al., 2004; Cattoli et al., 2011; Swayne DE 2012). A more extensive knowledge of the mechanisms underlying intra-host evolution of avian influenza viruses circulating in vaccinated poultry populations could be of help to formulate and adopt more adequate vaccine strategies.

Previous studies conducted in partially immune pigs indicated that the variability in immune response may influence the overall diversity of swine influenza virus during infection (Diaz et al., 2015) and showed that the hemagglutinin gene displayed nucleotide mutations at the very beginning of viral infection (Diaz et al., 2013, Murcia et al., 2012). However, to date there is no information on the intra-host evolution of HPAI avian influenza viruses circulating in vaccinated poultry populations. The surface glycoprotein, hemagglutinin (HA) is involved in the induction of a protective humoral and cell mediated immune response, and represent one of the major antigenic determinants of type A influenza viruses. In poultry, antigenic drift is driven primarily by multiple amino acid substitutions within major antigenic sites.

In this preliminary study, we analysed swabs sampled from vaccinated and challenged chickens with different levels of clinical and virological protection. Next-generation sequencing was performed to compare nucleotide and amino acid diversity at the level of the hemagglutinin among groups.

Material and methods

A deep sequencing analysis on the HA gene segment was performed on samples collected in a previous challenge study that aimed at assessing the protective efficacy of two avian influenza vaccines, against a HPAI H5N1 virus. Briefly, two groups of ten Specific Pathogen Free (SPF) day-old chicks were vaccinated twice at a 10-day interval by the sub-cutaneous route, using two distinct influenza inactivated vaccines (here named A and B). The birds were challenged with 106 50% Embryo Infectious Dose (EID50) HPAI H5N1 A/chicken/Egypt/4453-7/2011 virus (clade 2.2.1) (WHO/OIE/FAO H5N1 Evolution Working

(2012)) 21 days from boosting. Antibody responses were assessed by means of hemagglutinin inhibition test (HI) ten days from priming, on the day of boosting and 2 weeks after the challenge. Tracheal swabs (TS) were collected on days 2, 4, 6, 8 and 10 post challenge (p.c.) to evaluate viral shedding by quantitative real-time RT-PCR (qRRT-PCR) and calculate the EID50 equivalents.

The Egyptian HPAI H5N1 virus used for the challenge, as well as all TS positive by qRRT-PCR (six samples from group A and fourteen samples from group B), were processed as described below. Total RNA was isolated from tracheal swabs using Nucleospin RNA kit (Macherey-Nagel, Duren, Germany). Viral RNA encoding the HA gene segment was retro-transcribed and amplified using SuperScript III one-step reverse transcription-PCR (RT-PCR) system with PlatinumTaq High Fidelity (Invitrogen, Carlsbad, CA) using H5 specific primers. Sequencing libraries were prepared using Nextera XT DNA Sample preparation kit (Illumina) and processed as described by Monne et al. (2014) on Illumina Miseq desktop sequencer.

FASTQC software was used to inspect quality score of raw sequence files and post processing data coming from the high-throughput sequencing pipelines. Fastq files were cleaned with Trimmomatic (Bolger et al. 2014), using a 4-base-pair sliding-window algorithm with a quality score cut-off of 20; only reads longer than 80 nucleotides were considered and mapped to the hemagglutinin H5 reference sequence using bwa-mem (Li et al., 2010; Li H., 2013). The BAM alignment files obtained were parsed using the diversiTools program (http://josephhuges.github.io/btctools/) to determine the average base-calling error probability and to identify the frequency of single nucleotide polymorphisms (SNP). A 500x coverage and a 1.0% frequency were the minimum threshold parameters chosen according to the data obtained from deep-sequenced plasmid DNA internal control. A statistical strand bias test (Holm-Bonferroni), implemented in the LoFreq software (Wilm et al. 2012) was used to confirm diversiTools SNP results. Shannon entropy (SE) was calculated to measure the complexity of viral populations in each sample belonging to group A and B, using the following formula:

$$E = -\frac{1}{N}\sum_{i=1}^{N}(f_{iA}\ln f_{iA} + f_{iG}\ln f_{iG} + f_{iT}\ln f_{iT} + f_{iC}\ln f_{iC})$$

where f_i is the frequency of the nucleotide A, T, G or C at position i and N is the total length of the hemagglutinin gene.

The Wilcoxon Mann–Whitney rank-sum test was used to verify whether the distribution of EID50, Entropy and polymorphism were identical in both groups.

Amino acid changes situated near or within previously identified antigenic sites (Kaverin et al.,2007; Kaverin et al., 2002) and to the receptor binding site (Kovácová et al., 2002; Cattoli et al., 2011) were mapped on an hemagglutinin structure obtained by homology modelling using the Swissmodel server (http://swissmodel.expasy.org/) (Bordoli et al., 2009); UCSF Chimera (Pettersen et al., 2004) v.1.10.2 software was used for viewing.

Results

None of the two vaccines conferred either full clinical or virological protection. Nevertheless, all of the birds that received vaccine A survived the challenge, whereas vaccine B prevented death in only 70% of the birds (Tab. 1). Moreover, the vaccines differed in terms of suppression of viral shedding, as at each time p.i. fewer birds in group A shed viral RNA from the trachea ($p<0.10$), and the amount of shed virus was significantly lower than in group B on day 2 p.i. ($p=0.072$) (Tab. 1, Fig. 1). Ten days from priming, animals in

both groups showed no detectable levels of HI antibodies against the challenge virus. After boosting, birds in group A recorded a 2,6 log2 HI geometric mean titre (GMT), whereas in group B all birds resulted either negative or recorded HI titres of 1 log2 (GMT of 0,2). After the challenge, seroconversion, expressed as an HI GMT increase equal to or higher than 2 log2, was observed in all of the survived birds in group B (GMT of 2,2), as opposed to 50% of the animals in group A (HI GMT of 3,0 log2).

Sufficient RNA for deep sequencing analysis was recovered from TS only on days 2 and 4 p.c. Data were obtained for a total of twenty-one positive samples, specifically: a) the challenge virus (4453/11), b) six samples belonging to group A, five of which at 2 days p.c. (34A2, 35A2, 37A2, 47A2, 59A2) and one at 4 days p.c. (34A4), and c) fourteen samples belonging to group B, nine of which at 2 days p.c. (72B2, 73B2, 75B2, 79B2, 80B2, 81B2, 83B2, 86B2, 88B2) and five at 4 days p.c. (73B4, 79B4, 81B4, 86B4, 88B4). Each sample was identified as follows: the first two digits refer to the animal identification code, the alphabetic characters (A or B) identify the group and the last digit indicates the number of days p.c.

To characterize the complexity of the viral population of the 20 clinical samples from the vaccinated birds, the per-site Shannon entropy was calculated, considering the frequencies of nucleotide substitutions across the hemagglutinin gene. The entropy measures fluctuated considerably: the samples with the lowest values belonged to group A (0 and 0.000130), while the ones with the highest values (0.00076 and 0.00085) belonged to group B.

The analysis of the nucleotide sequence diversity of the hemagglutinin gene showed several synonymous and non-synonymous polymorphisms distributed on the HA gene of all samples. However, a comparison between the two groups revealed a great variability in the number of polymorphisms among samples. Five out of six samples belonging to group A showed from one to six minority variants per sample (tab.2), randomly distributed across eleven nucleotide positions, with a frequency ranging from 1.05% to 6.88%. Only two of the identified polymorphic sites (929 and 1071) were shared among two or more samples. Differently, group B displayed a higher number of polymorphisms (tab. 3), from three to thirteen per sample, distributed in sixty-three positions and showing a frequency ranging from 1.01% to 68.70%. Six of these polymorphisms (residues 258, 470, 929, 1032, 1071 and 1379) were acquired independently by two or more samples. The positions which displayed the highest frequency values (2.38% to 32.74%) were position 258, shared among 3 out of 14 samples, and position 1032 (1.25% to 68.70%), revealed to be the one shared by the largest number of samples (6 out of 14) belonging to group B. Among the synonymous minority variants identified, only positions 1071 was shared between group A and B with frequency values slightly higher in samples belonging to group A (tab.2 and tab.3).

The minority variants identified at position 1032 of the HA gene of six samples of group B was already present in the challenge strain, like polymorphisms in position 164 and 1395 displayed separately in only two samples of this group; all the other variants appear to have emerged only after the viral introduction in the host. None of the polymorphism already present within the HA gene of the challenge virus were later identified among samples belonging to group A.

We performed a non parametric Wilcoxon Mann–Whitney test to evaluate whether the EID50 equivalents, the entropy values and the number of polymorphisms of the samples from the two groups were significantly different. The samples collected at 4 days p.c. were only six, five from group B and one from group A, therefore they were excluded from the statistical comparison. The test indicated that the distribution of values

of EID_{50}, entropy and polymorphism were different in the two vaccination groups, with group B showing the highest values of EID_{50}, entropy and number of polymorphisms (fig 1).

Non-synonymous substitutions represented respectively 56% and 64% of the total polymorphisms in group A and B and were randomly distributed across the HA gene. Amino acid positions in the HA protein refer to H5 numbering; challenge virus A/chicken/Egypt/4453-7/2011 used in this study displayed a deletion in position 129. Among samples belonging to group A, seven nucleotide positions, across the whole HA gene, showed non-synonymous minority variants with a frequency ranging from 1.05% to 6.88%. Only one non-synonymous mutation, leading to amino acid substitution H295L was in common in 3 out of 6 samples. None of the six samples belonging to group A showed non- synonymous minority variants located at the globular head of the HA1 protein near or within the three secondary structural elements of the receptor binding domain (RBD), formed by the 130-loop, 190-helix, and 220-loop, and/or in antigenic sites previously identified. Compared to group A, group B showed a higher number of nucleotide positions (forty-nine) involved in non-synonymous minority variants randomly distributed across the hemagglutinin gene. The minority variants shared among two or more samples were identified at six nucleotide positions, three of which, 470, 941 and 1379 led to non synonymous polymorphism (frequency values are shown in table 3. Interestingly, non synonymous SNPs at positions 1018 and 1019 led to the mutations R325K and R325G situated in the cleavage site of the hemagglutinin proteins in two samples, respectively at 2 and 4 days p.c. 2 out of fourteen (2/14) samples displayed a common non synonymous mutation at nucleotide position 941 that lead to amino acid mutation H295L, as previously found in group A. Six out of nine samples (6/9) at 2 days p.c. and one out of five samples (1/5) at 4 days p.c. displayed from one to two non-synonymous polymorphisms within the receptor binding cavity (from 130 to 225 amino acid position), where also antigenic sites A, B and partially D are present, for a total of eleven minority variants. Four mutations near or within antigenic site A were found randomly in three samples at 2 days p.c. and in one sample at 4 days p.c.; in particular, C135F, and S141F mutations determined a change of the physical chemical properties, whereas minority variant S142Yand S142F , identified separately in two samples, displayed not-charged and polar properties for both amino acids. Minority variant Y157C close to antigenic site B was detected in one sample at 2 days p.c.; however, on day 4 p.c the same sample did not show the same minority variant. Non-synonymous mutations observed within the RBD were C135F, I213V and K218E. In particular, C135F was positioned within the 130 loop, whereas, S141F and S142Y/F were close to this secondary structural domain. I213V and A214D appeared near the 220 loop, whereas K218E was close to antigenic site D situated in the 220 loop. One sample belonging to group B displayed minority variant S141F previously shown to be involved in antigenic drift of Egyptian H5N1 HPAI viruses (Cattoli et al., 2011).

Discussion

For many influenza subtypes, such as HPAI H5N1 virus, vaccination programmes are currently underway in attempt to control and eradicate these diseases. However, influenza A viruses evolve rapidly in response to selection pressures generated through vaccine protection, and the emergence of virus strains for which existing vaccines are not well matched and offer little protection continuously challenges the effectiveness of vaccines in the field. Deep sequencing technologies are used to investigate and characterize the complexity of the viral population, to detect low-frequency mutations and to follow the

evolution of the genetically related variants present in a viral population. Samples collected in the framework of a previous experimental study aiming at assessing the protection efficacy of two distinct vaccines against HPAI H5N1 virus, offered the unique opportunity to compare viral population diversity in two distinct immune status background. In particular, the two experimental challenge groups (A and B) allowed to mimic different level of immunity and than to explore how viruses evolve within hosts that have received only partial vaccination with influenza inactivated vaccines. The HI test conducted on sera collected prior to challenge indicated that the HA protein of vaccine A seemed to be antigenically similar to the HA of the viral challenge strain, whereas HA protein of vaccine B seemed antigenically different from the viral challenge strain. All of the birds that received vaccine A survived the challenge, whereas vaccine B prevented death in only 70% of the birds. Results obtained applying an ultra-deep sequencing approach to samples collected from A and B experimental groups suggest that a suboptimal level of antibody protection may be considered a factor involved in the generation of a viral population with the highest genetic heterogeneity. We identified a total of 16 polymorphisms (56% non-synonymous) in group A and of 76 polymorphisms (64% non-synonymous) in group B. Interestingly, two samples belonging to group B displayed minority variants R325G and R325K situated within the cleavage site in the hemagglutinin protein; a previous study carried out on Egyptian (HPAI) H5N1 viruses, showed that position 325 significantly reduced pathogenicity without altering the transmission efficiency (Yoon et al., 2013). Eleven out of forty-nine (12/49) non-synonymous polymorphisms identified in the group B fall within or close to a previously identified receptor binding cavity; seven of them were near or within antigenic sites A, B and partially D, whereas none of the samples belonging to group A showed non-synonymous minority variants in the same area. As previously shown (Cattoli et al. 2011), reverse genetics mutants on the hemagglutinin of a H5N1 highly pathogenic avian virus demonstrated that five amino acid positions (74, 140, 141, 144, and 162) could be involved in the antigenic drift observed in the HPAI H5N1 field strain circulating in Egypt in 2008. Interestingly, one sample belonging to group B displayed the minority variants S141F. These observations suggest that a suboptimal immune protection may induce an increase of the complexity in the viral population and promote the selection of minority variants, some of which could be involved in antigenic drift. Our study highlights that viral evolution and appearance of amino acid mutations within interesting antigenic sites can be observed from the early stages of infection. None of the samples at 4 days p.c. showed a fixation of non-synonymous substitutions; this could be due to a bias during the sampling, or to the deep sequencing procedure that needs an improvement, or simply the reason could be ascribed to the unknown time required for minority variant fixation. Considering that our sample size was rather limited, this preliminary study should be further confirmed by making an assessment on a greater number of samples, and samples collected in a wider range of time should be selected, as well. Further studies on the whole influenza A virus genome could provide us with an overview on the effect of suboptimal vaccine protection in the evolution of viral populations. The higher entropy and number of polymorphisms were associated with two concomitant features of group B. In fact, not only more birds in this group shed viral RNA by the tracheal route, but they were also found to shed a higher amount of virus compared to birds in group A. Moreover the humoral immunity in group B proved to match the challenge virus in a less efficient way then the one elicited by the A vaccine. The combination of a higher viral replication and a different immune pressure might be held responsible for increased viral diversity and favoured an increase in the number of non-synonymous minority variants.

Conclusions

Deep sequencing analysis proved to be a valid tool to explore and characterize differences among heterogenic viral populations present in vaccinated animals during infection; furthermore, it highlighted the presence of minority variants at the very beginning of the infectious phase, which could not be revealed by using the classical sequencing method. In addition, this work highlights the need to further explore the results that can be generated applying NGS approach to other experimental models and influenza subtypes to confirm, as it seems from these data, that it could be a suitable method to understand the mechanisms that underpin how viruses escape vaccine protection and have early indication of threats to the effectiveness of vaccine control programmes.

Acknowledgment

This work was financially supported by the European projects Epi-SEQ (research project supported under the 2nd Joint Call for Transnational Research Projects by EMIDA ERA-NET [FP7 project no. 219235]). The authors would like to acknowledge Francesca Ellero for providing language help. This study was conducted in the framework of the Doctoral school in Bioscience and Biotechnology at the University of Padua (Adelaide Milani).

References

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PubMed PMID: 24695404; PubMed Central PMCID: PMC4103590.

Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. Protein structure homology odeling using SWISS-MODEL workspace. Nat Protoc. 2009;4(1):1–13. doi: 10.1038/nprot.2008.197.

Cattoli G, Milani A, Temperton N, Zecchin B, Buratin A, Molesti E, Aly MM, Arafa A, Capua I. Antigenic drift in H5N1 avian influenza virus in poultry is driven by mutations in major antigenic sites of the hemagglutinin molecule analogous to those for human influenza virus. J Virol. 2011 Sep;85(17):8718-24. doi: 10.1128/JVI.02403-10. Epub 2011 Jul 6. PubMed PMID: 21734057; PubMed Central PMCID: PMC3165837.

Connie Leung YH, Luk G, Sia SF, Wu YO, Ho CK, Chow KC, Tang SC, Guan Y, Malik Peiris JS. Experimental challenge of chicken vaccinated with commercially available H5 vaccines reveals loss of protection to some highly pathogenic avian influenza H5N1 strains circulating in Hong Kong/China. Vaccine. 2013 Aug 2;31(35):3536-42. doi: 10.1016/j.vaccine.2013.05.076. Epub 2013 Jun 19. PubMed PMID: 23791547.

Diaz A, Allerson M, Culhane M, Sreevatsan S, Torremorell M. Antigenic drift of H1N1 influenza A virus in pigs with and without passive immunity. Influenza Other Respir Viruses. 2013 Dec;7 Suppl 4:52-60. doi: 10.1111/irv.12190. PubMed PMID: 24224820.

Diaz A, Enomoto S, Romagosa A, Sreevatsan S, Nelson M, Culhane M, Torremorell M. Genome plasticity of triple-reassortant H1N1 influenza A virus during infection of vaccinated pigs. J Gen Virol. 2015 Oct;96(10):2982-93. doi:10.1099/jgv.0.000258

Kaverin NV, Rudneva IA, Ilyushina NA, Varich NL, Lipatov AS, Smirnov YA, Govorkova EA, Gitelman AK, Lvov DK, Webster RG. Structure of antigenic sites on the haemagglutinin molecule of H5 avian influenza virus and phenotypic variation of escape mutants. J Gen Virol. 2002 Oct;83(Pt 10):2497-505. PubMed PMID: 12237433.

Kaverin NV, Rudneva IA, Govorkova EA, Timofeeva TA, Shilov AA, Kochergin-Nikitsky KS, Krylov PS, Webster RG. Epitope mapping of the hemagglutinin molecule of a highly pathogenic H5N1 influenza virus by using monoclonal antibodies. J Virol. 2007 Dec;81(23):12911-7. Epub 2007 Sep 19. PubMed PMID: 17881439; PubMed Central PMCID: PMC2169086.

Jeong-Ki Kim, Ghazi Kayali, David Walker, Heather L. Forrest, Ali H. Ellebedy, Yolanda S. Griffin, Adam Rubrum, Mahmoud M. Bahgat, M. A. Kutkat, M. A. A. Ali, Jerry R. Aldridge, Nicholas J. Negovetich, Scott Krauss, Richard J. Webby, Robert G. Webster. Puzzling inefficiency of H5N1 influenza vaccines in Egyptian poultry. Proc Natl Acad Sci U S A. 2010 June 15; 107(24): 11044–11049. Published online 2010 June 1. doi: 10.1073/pnas.1006419107 PMCID: PMC2890765

Kovácová A, Ruttkay-Nedecký G, Haverlík IK, Janecek S. Sequence similarities and evolutionary relationships of influenza virus A hemagglutinins. Virus Genes. 2002;24(1):57-63. PubMed PMID: 11928990.

Lee CW, Suarez DL. Avian influenza virus: prospects for prevention and control by vaccination. Anim Health Res Rev. 2005 Jun;6(1):1-15. Review. PubMed PMID: 16164006.

Lee CW, Senne DA, Suarez DL. Effect of vaccine use in the evolution of Mexican lineage H5N2 avian influenza virus. J Virol. 2004 Aug;78(15):8372-81. PubMed PMID: 15254209; PubMed Central PMCID: PMC446090.

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010 Mar 1;26(5):589-95. doi: 10.1093/bioinformatics/btp698. Epub 2010 Jan 15. PubMed PMID: 20080505; PubMed Central PMCID: PMC2828108.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 arXiv:1303.3997.

Monne I, Fusaro A, Nelson MI, Bonfanti L, Mulatti P, Hughes J, Murcia PR, Schivo A, Valastro V, Moreno A, Holmes EC, Cattoli G. Emergence of a highly pathogenic avian influenza virus from a low-pathogenic

progenitor. J Virol. 2014 Apr;88(8):4375-88. doi: 10.1128/JVI.03181-13. Epub 2014 Feb 5. PubMed PMID:24501401; PubMed Central PMCID: PMC3993777.

Murcia PR, Hughes J, Battista P, Lloyd L, Baillie GJ, Ramirez-Gonzalez RH, Ormond D, Oliver K, Elton D, Mumford JA, Caccamo M, Kellam P, Grenfell BT, Holmes EC, Wood JL. Evolution of an Eurasian avian-like influenza virus in naïve and vaccinated pigs. PLoS Pathog. 2012;8(5):e1002730. doi: 10.1371/journal.ppat.1002730. Epub 2012 May 31. PubMed PMID: 22693449; PubMed Central PMCID: PMC3364949.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–1612. doi: 10.1002/jcc.20084

Swayne DE. Impact of vaccines and vaccination on global control of avian influenza. Avian Dis. 2012 Dec;56(4 Suppl):818-28. Review. PubMed PMID: 23402099.

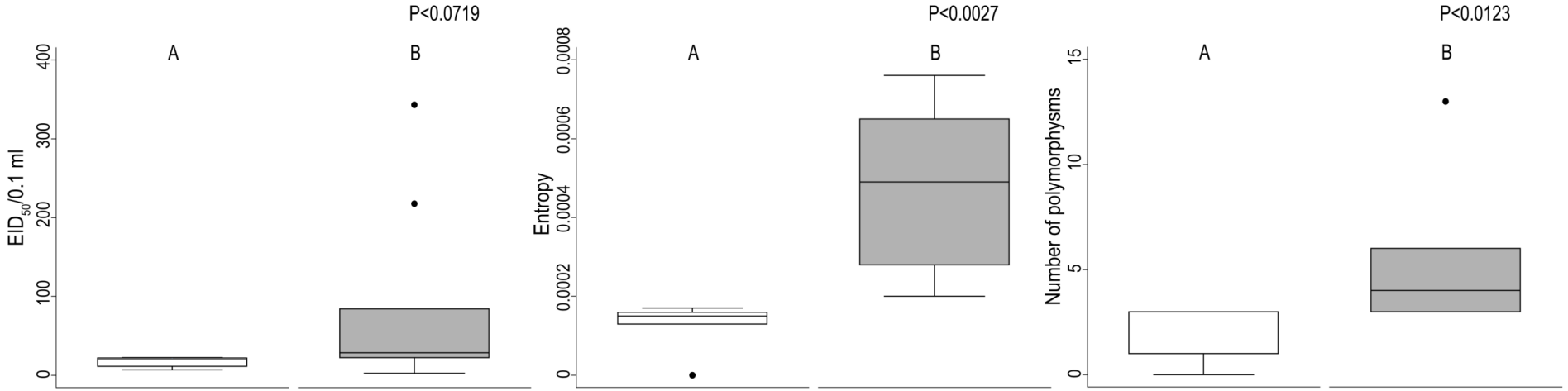
Toro H, van Santen VL, Jackwood MW. Genetic diversity and selection regulates evolution of infectious bronchitis virus. Avian Dis. 2012 Sep;56(3):449-55.Review. PubMed PMID: 23050459.

WHO/OIE/FAO H5N1 Evolution Working Group (2012), Continued evolution of highly pathogenic avian influenza A (H5N1): updated nomenclature. Influenza and Other Respiratory Viruses, 6: 1–5. doi: 10.1111/j.1750-2659.2011.00298

Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012 Dec;40(22):11189-201. doi: 10.1093/nar/gks918. Epub 2012 Oct 12. PubMed PMID: 23066108; PubMed Central PMCID: PMC3526318.

Yoon SW, Kayali G, Ali MA, Webster RG, Webby RJ, Ducatez MF. A single amino acid at the hemagglutinin cleavage site contributes to the pathogenicity but not the transmission of Egyptian highly pathogenic H5N1 influenza virus in chickens. J Virol. 2013 Apr;87(8):4786-8. doi: 10.1128/JVI.03551-12. Epub 2013 Feb 13. PubMed PMID: 23408622; PubMed Central PMCID: PMC3624353.

Fig. 1:Box plot of EID_{50}, entropy and mutation by vaccination group. Boxplot was drawn to display the distribution of the quantitative values of EID_{50}, entropy and mutation for each vaccination group. For both groups A and B distribution of values of EID_{50} , entropy and total number of polymorphisms were calculated. Group B displays higher values for each variables than group A.



Tab. 1. Viral shedding, survival and seroconversion rates for groups A and B.

	Shedding rate (%)						Survival rate (%)	Seroconversion Rate (%)
	Trachea			Cloaca				
	2 dpi	4 dpi	6 dpi	2 dpi	4 dpi	6 dpi		
Vaccine A	80	50	10	0	0	0	100	50
Vaccine B	100	80	25	10	10	12.5	70	100

Tab.2. SNP identified in samples belonging to group A; frequency values are in percentage

CDS	POLYM	34A2	35A2	37A2	47A2	59A2	34A4
229	L61F	1,34					
238	N64D	1,54					
489	***SIL***						1,13
929	H295L		1,05	1,24			1,32
1071	***SIL***	2,37	1,87	3,55			2,77
1190	V382A			1,12			
1222	F393L				6,88		
1236	***SIL***						2,01
1400	V452A						2,72
1625	L527P						2,35
1701	***SIL***		1,37				

Tab.3. SNP identified in samples belonging to group B; frequency values are in percentage

CDS	POLYM	72B2	73B2	75B2	79B2	80B2	81B2	83B2	86B2	88B2	73B4	79B4	81B4	86B4	88B4
15	***SIL***												1,70		
17	-L6P														3,02
20	-L7P									1,41					
149	E34G										7,93				
153	***SIL***														2,86
160	N38D						1,52								
164	G39E							38,84							
213	***SIL***										1,59				
258	***SIL***		2,38							16,68					32,74
268	***SIL***														2,72
357	H103Q		2,30												
410	S122F													3,49	
411	***SIL***		1,44												
414	W123*		1,49												
449	C135F		2,28												
458	***SIL***	1,25													
467	S141F									1,01					
470	S142Y/F		1,65											3,88	
492	W149*											1,75			
497	T151I				4,46										
515	Y157C					2,99									
536	Y164C								5,35						
555	D170E						5,29								
585	***SIL***													2,89	
682	I213V									1,50					
686	A214D				3,77										
697	K218E													1,05	
804	***SIL***												1,74		
828	***SIL***					1,39									
848	E268G							3,56							
880	Q279K								3,92						
929	H295L	1,04						1,27							
931	P296S				5,16										
979	V312I		1,30												
985	A314T											1,98			
1015	E324K							2,66							
1018	R325G												1,44		
1019	R325K		1,33												
1032	***SIL***		1,58	68,70		54,45					6,74	1,25			15,16
1048	A335T		1,27												
1071	***SIL***	2,01		1,46				1,66					1,76		
1107	***SIL***											1,20			
1119	N358D		1,40												
1150	E369K				10,51										
1163	K373R											1,02			
1199	I385T								2,26						
1264	I407L											1,29			
1290	E415D		1,42												
1327	L428F		2,13												
1349	E435G			2,92											
1364	F440S												2,71		
1365	***SIL***											1,84			
1379	V445A				5,67		1,54								
1394	D450G							3,96							
1395	***SIL***										6,62				
1416	***SIL***											1,12			
1429	E462K											1,61			
1504	Y487H						1,22								
1517	Q491L				2,50										
1522	S493P	2,53													
1575	***SIL***													2,67	
1598	S518*											1,05			
1628	A528V									1,25					

CHAPTER 4

Comparative structural analysis of haemagglutinin proteins from type A influenza viruses: conserved and variable features

Righetto I, Milani A, Cattoli G, Filippini F.

BMC Bioinformatics. 2014 Dec 10;15:363.

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Comparative structural analysis of haemagglutinin proteins from type A influenza viruses: conserved and variable features

Irene Righetto[1], Adelaide Milani[2], Giovanni Cattoli[2] and Francesco Filippini[1*]

Abstract

Background: Genome variation is very high in influenza A viruses. However, viral evolution and spreading is strongly influenced by immunogenic features and capacity to bind host cells, depending in turn on the two major capsidic proteins. Therefore, such viruses are classified based on haemagglutinin and neuraminidase types, e.g. H5N1. Current analyses of viral evolution are based on serological and primary sequence comparison; however, comparative structural analysis of capsidic proteins can provide functional insights on surface regions possibly crucial to antigenicity and cell binding.

Results: We performed extensive structural comparison of influenza virus haemagglutinins and of their domains and subregions to investigate type- and/or domain-specific variation. We found that structural closeness and primary sequence similarity are not always tightly related; moreover, type-specific features could be inferred when comparing surface properties of haemagglutinin subregions, monomers and trimers, in terms of electrostatics and hydropathy. Focusing on H5N1, we found that variation at the receptor binding domain surface intriguingly relates to branching of still circulating clades from those ones that are no longer circulating.

Conclusions: Evidence from this work suggests that integrating phylogenetic and serological analyses by extensive structural comparison can help in understanding the 'functional evolution' of viral surface determinants. In particular, variation in electrostatic and hydropathy patches can provide molecular evolution markers: intriguing surface charge redistribution characterizing the haemagglutinin receptor binding domains from circulating H5N1 clades 2 and 7 might have contributed to antigenic escape hence to their evolutionary success and spreading.

Keywords: Haemagglutinin, Avian influenza virus, Viral evolution, H5N1, Antigenic drift, Receptor binding domain, Homology modeling, Isopotential contour, Hydropathy analysis

Background

Influenza caused by influenza A viruses occurs in both birds and mammals. In humans, influenza A viruses infect hundreds of millions individuals, causing a high number of deaths per year. Indeed, influenza A outbreaks occurred in 1918, 1957 and 1968 resulted in death for ~100 million people worldwide [1]. However, seasonal epidemic outbreaks cause estimated 250.000 to 500.000 yearly deaths worldwide [2] (data from the World Health Organization (WHO) [3] and from the Center for Disease Control and prevention [4]). The largest reservoir of all subtypes of influenza A is found in wild water avian species and some viruses can infect different hosts [5,6]. Classification of influenza type A virus subtypes is based on antigenic and genetic differences in the two surface spike proteins: haemagglutinin (HA) and neuraminidase. For instance, H5N1 viruses combine the haemagglutinin of the H5 subtype with neuraminidase of the N1 subtype. A wide interest for haemagglutinin depends on evidence that this protein (i) is crucial to the attachment and penetration into the host cell, (ii) represents the main viral surface antigen, and (iii) is a major player in the stimulation of the neutralizing antibody response [7]. Haemagglutinin is synthesized as a precursor and then processed by cellular proteases to yield mature polypeptide subregions. In order to provide unambiguos information, hereafter acronyms

* Correspondence: francesco.filippini@unipd.it
[1]Molecular Biology and Bioinformatics Unit (MOLBINFO), Department of Biology, University of Padua, via U. Bassi 58/B, 35131 Padova, Italy
Full list of author information is available at the end of the article

for haemagglutinin are the followings: 'HA' for haemagglutinin in general; HA0 for the precursor; HA1 and HA2 for the two subregions and 'H' followed by progressive numbering (H1 to H16) for each haemagglutinin subtype. Influenza virus haemagglutinin is a type I transmembrane glycoprotein that is exposed at the viral surface as a homotrimer. Trimerization is possible once proteolytic cleavage of the unfolded HA0 precursor occurs hence allows for folding of monomers, each consisting of two mature chains: HA1 and HA2 [7]. Structurally, each monomer consists of a globular 'head' (part of chain HA1) and of a 'stem' region (contributed by both chains HA1 and HA2). The head includes a receptor-binding domain (RBD) and a vestigial esterase domain (VED), whereas the stem is structured as a mainly α helical, coiled coil region. Functionally, the RBD mediates docking to the host cell by binding sialic acids as cell entry receptors, whereas the stem domain mediates both tethering and membrane fusion once conformational change is occurred, caused by pH decreasing along the endosomal route. For several years, classification of HA from influenza viruses was mainly based upon serological and/or phylogenetic analysis [8]. However, structural genomics projects are providing the scientific community with an increasing number of structural templates, while contemporary reverse genetics, immunogenomics investigations and improved sequencing technologies are producing a high number of mutant sequences. Changes in serological specificity depend on variation of epitopes recognized by the specific antibody rather than on the extent of sequence divergence, meaning i.e. that (i) two proteins with highly similar sequences may show quite different properties when considering recognition of specific epitopes and (ii) two proteins may share antigenic properties even when having highly divergent sequences, if epitopes involved in the specific recognition were conserved. Variation of some protein properties sometimes may depend only on 'local and limited changes', e.g. mutation of a few - or even only one – residue(s) within linear or conformational motifs. In fact, even when local variation in sequence is seemingly poorly evident, it may result in 'locally dramatic' changes in accessible surface area, electrostatic potential, hydropathy or hydrophilicity features that can deeply change motif functionality. It is common knowledge that variation in surface features of a protein can modulate 'recognition' interactions of the protein itself. Since variation often depends on mutation of a number of residues and changes in side chains can vary multiple biochemical features, it is difficult or even nonsense trying to establish *a priori* which specific property (among e.g. surface area and shape, electrostatics or hydrophobicity) should be more relevant than others in modulating recognition interactions. In fact, changes in each specific property can result in such modulation, and this can be independent on variation of other features, or modulation can result from the aggregate or synergistic effect of multiple feature changes. So far, several sequence-based studies on variation could provide valuable phylogenetic evidence; however, such studies are of minor help in inferring variation at protein regions including amino acids that are far each other in the primary sequence and quite close within the 3D protein structure (conformational epitopes). In practice, while sequence-based investigation can be good in highlighting very evident changes at individual positions of a protein chain, in general they fail in highlighting meaningful 'group variation', i.e. in identifying - especially when the overall variation is relevant and spread - relationship of specific multiple changes to variation in conformational epitopes hence in interactions they mediate.

Once solved structures are available, presence of one or more structural templates allows for shifting to 'conformational epitope based' studies on variation and, in particular, to investigating on surface region variation. Stressing relevance of local surface variation is particularly important when considering special constraints addressing viruses evolution: keeping basic properties in simplified but complex pathogenic systems while simultaneously varying - as much as possible - all variable epitopes, in order to escape the immune responses of their hosts. Therefore, viral genome evolution runs along two parallel tracks, both of which, like in railways, must be followed: (i) mutations in sites crucial to protein machinery mediating basic functions (e.g. in motifs relevant to host recognition or cell entrance) are not allowed because they strongly impair viral fitness, and at the same time, (ii) hyper-variability is needed to escape recognition by neutralizing antibodies ('antigenic drift', [7]). Given that surface viral proteins do not interact only with antibodies (as their original function is to contact the host), in addition to determining antigenic drift, variation can also influence pathogenicity (because e.g. of modified interaction with cell receptors in different tissues and organ districts) or host specificity. Influenza viruses do not escape such a two-tracks rule, hence while global structure conservation ensures basic functions, limited or even subtle changes in local structural features may modulate interactions of the viral proteins with the host molecules/cells and thus mechanisms underlying antigenic drift, pathogenicity shifts and host specificity change. Phylogenetically and serologically, haemagglutinins are divided into either two supergroups or four groups: Group 1 (H1, 2, 5, 6, 11, 13 and 16); Group 2 (H8, 9 and 12); Group 3 (H3, 4 and 14) and Group 4 (H7, 10 and 15). The two supergroups consist of Groups 1 + 2 and 3 + 4, respectively [9,10]. Thanks to the availability of thousands of viral genomes/gene sequences and of several specific antibodies/vaccines, a large number of sequence-

based/phylogenetic and serological analyses of avian flu viruses have been performed and published so far. This notwithstanding, mechanisms in viral evolution are still elusive, as genome/proteome-wide analyses on sequence variation or antigenic features are able to only partially unveil a number of relevant changes, because of the overall mutational noise. Therefore, structural 'zoom in' is needed to integrate such analyses by identifying 'meaningful' variation. This prompted us to take advantage from availability of structural templates to perform structural comparison among different HA subtypes, in order to identify subtype- and subregion-specific feature variation suggestive for possible involvement in antigenic recognition, or pathogenicity and host specificity. Last but not least, evidence from structural comparison can check relationship among serological, phylogenetic and structural closeness.

We started our analyses using six currently available solved HA structures; then, in order to investigate structural variation possibly underlying H5N1 clades evolution and spreading, we also created clade models by homology modeling. The six HA structures solved so far: H1 [11], H2 [12], H3 [13], H5 [14], H7 [9], H9 [15], all concern mature proteins, consisting of the two HA1 and HA2 parts of haemagglutinin. Solved structure of H16 [16] was not considered for this analysis because it corresponds to the HA0 precursor. Comparative analysis of structural features unveiled that some discrepancy may occur with respect to a generally observed agreement between sequence and structural closeness, because of subregion local variation. Structural analysis was performed by comparison of secondary structure topology and surface analysis, in terms of both electrostatic and hydropathy analysis.

Results and discussion

Comparison among solved HA structures

Prior to creating models, preliminary analysis of the six available HA structures was performed in order to evaluate intra- and inter-group structural variation by superposition of all structure pairs and computation of their Root Mean Square Deviation (RMSD). Indeed, the RMSD of two superposed structures indicates their 'structural divergence' from one another. As both sequence mutation and conformational variation inflate the RMSD, values up to 2 Ångstrom indicate structural similarity [17]. Structural superposition of each possible combination of two different HA molecules (hereafter referred to as 'pairs') and RMSD computing were performed using Chimera 1.8.1 software [18]. Pair-wise method was chosen to calculate RMSD because all superpositions only compared pairs in order to properly relate a structural closeness index for a pair to identity/similarity values (commonly reported as an index to state closeness) from the corresponding aligned sequences. Fold comparison method based on sequence fragmentation and order-independent resorting was not considered because order-dependent global alignment is an established standard for comparing highly similar sequences in structural biology and the alignment of sequence blocks for phylogenetic analyses is also order-dependent.

In addition to superposing structures of HA monomers, also corresponding structures of their Receptor Binding domains (RBDs) were superposed. Results are summarized in Table 1. Evidence that RMSD values for monomer pairs are lower than those ones for corresponding HA1 or RBD regions is not surprising, because RBDs are major determinants in antigenic variation [9]. Moreover, HA2 'stem' region of the monomer is structurally less variable than HA1 [19], hence its contribution results in decreasing the overall monomer RMSD value. RMSD values for HA1 pairs are higher than corresponding RBDs because of unstructured regions connecting RBDs to stems. Group 1 is - at least to date - the only HA group in which multiple structures (in particular, H1, H2 and H5) are solved. Structural comparison within this group highlights some intriguing evidence. When comparing monomers amino acid sequences, H5 results to be closer to H2 than to H1, independently on identity (roughly 73% vs. 63%) or similarity (approximately 86% vs. 81%) is considered. Such relationship is confirmed for both HA1 and RBD sequences, as shown by identity and similarity values in Table 1. However, when comparing structures, H5 is closer to H1 than H2, as in all comparisons, H5:H1 superposition RMSD values are lower than H5:H2 ones. Commonly, % identity is taken into account as an index for relationship among proteins [20]. However, from a structural point of view, 'type' of mutations occurred - rather than the overall sequence divergence - is very important: a few mutations (or even a single one) to some specific residues in 'critical' regions can result in dramatic structural changes. Structural fold and architecture can be highly conserved even among proteins and protein domains showing no sequence homology because of either long evolutionary divergence or even convergent evolution [21]. At the same time, within such families, fold can be disrupted (resulting in loss of function and disease) by single or few specific mutation(s), which indeed result in keeping 99% or higher sequence identity values [22,23]. In the structural comparison of H5 to haemagglutinins from different groups (represented by H9, H3 and H7) further interesting points emerge. In the monomer comparison, % identity approximately ranges from 41 to 49%. The same 8% difference in % identity is retrieved in % similarity (ranging from 64 to 72%). However, RMSD for corresponding monomer pairs keep quite similar values, i.e. they are not impaired by lower %

Table 1 Structural and sequence closeness among pairs of haemagglutinin proteins with solved structures

	RBD				
	H2	**H5**	**H9**	**H3**	**H7**
H1	r:1.343 i:55.4 s:78.4	r:0.918 i:52.0 s:78.3	r:1.249 i:45.7 s:69.7	r:2.292 i:38.0 s:61.1	r:2.784 i:37.2 s:63.7
H2		r:1.130 i:65.6 s:83.7	r:1.636 i:41.4 s:66.8	r:2.083 i:36.8 s:57.3	r:1.772 i:33.5 s:60.7
H5			r:1.498 i:41.4 s:66.4	r:2.241 i:37.3 s:61.4	r:3.085 i:38.4 s:67.4
H9				r:1.983 i:36.9 s:60.4	r:2.069 i:33.9 s:58.4
H3					r:1.429 i:35.0 s:63.6
	HA1				
	H2	**H5**	**H9**	**H3**	**H7**
H1	r:1.476 i:56.7 s:78.7	r:1.065 i:56.6 s:79.2	r:1.563 i:46.4 s:69.4	r:2.548 i:37.1 s:62.9	r:2.941 i:36.1 s:63.3
H2		r:1.527 i:67.7 s:83.3	r:2.087 i:43.5 s:65.3	r:3.253 i:35.3 s:58.3	r:3.025 i:34.5 s:60.6
H5			r:1.680 i:43.5 s:67.0	r:3.043 i:37.2 s:61.9	r:2.755 i:36.9 s:66.7
H9				r:2.320 i:35.8 s:60.9	r:3.672 i:33.5 s:59.8
H3					r:1.631 i:37.8 s:64.0
	Monomer				
	H2	**H5**	**H9**	**H3**	**H7**
H1	r:1.180 i:64.2 s:82.9	r:0.98 i:62.8 s:81.5	r:1.350 i:50.4 s:71.3	r:1.710 i:40.0 s:61.6	r:1.780 i:42.4 s:67.1
H2		r:1.100 i:73.0 s:85.7	r:1.450 i:49.0 s:69.6	r:1.760 i:37.6 s:59.6	r:1.730 i:40.6 s:66.5
H5			r:1.686 i:48.7 s:72.0	r:1.680 i:40.2 s:63.9	r:1.620 i:42.3 s:69.9
H9				r:1.760 i:37.9 s:61.7	r:1.850 i:40.8 s:66.1
H3					r:1.250 i:44.0 s:66.2

Within each cell, the upper value is RMSD (r) for the superposed pair and lower values (in %) are identity (i) and similarity (s) for corresponding, aligned amino acid sequences.

identity or similarity values. This is not surprising, because - as shown by aforementioned example (and by many others in literature) - very ancient divergence or convergence can result in fold conservation among proteins without significant sequence similarity. Structural differences become clearly evident when comparison focuses on HA1 and RBD regions: H5 is quite closer to H9 than H3 and H7 (roughly doubled RMSD) and in this instance substantial agreement between structural and sequence divergence is found. Once again, a rationale for this is found when considering common properties of protein domains. Different subregions of the same protein are involved in different interactions and pathways. Therefore, molecular evolution can locally change subregion structures to modulate specific interactions and pathways, without affecting those ones mediated from other subregions of the same protein. In practice, only when structural variation analysis is performed at both overall and local level (i.e. focusing on individual domains and/or domain motifs), it is possible to boost subsequent experimental work. In fact, subregion analysis allows for shedding light on specific molecular properties that are likely to underlie different functions of the protein. In conclusion, agreement between sequence homology and structural closeness which is generally observed [20] has not to be strictly interpreted as 'a rule' to be followed. Values from Table 1 show that, in most instances, such an agreement is found. However, in several examples and depending on local variation, superimpositions between pairs with quite comparable % identity and similarity may show very different RMSD values and vice versa.

Comparative analysis of secondary structure elements

Available structures were superposed and then tiled using UCSF Chimera 1.8.1 to keep the same orientation and to avoid visual superposition. This way, variation of secondary structure elements among individual structures can be clearly distinguished and viewed. In order to exclude any artifact from modeling, only the six available solved structures were compared. In terms of secondary structure, three subregions can be distinguished within the HA2 stem [see Additional file 1, panel A]: an α subregion and two β subregions (being either proximal or distal to the VED). The former consists of α helices A-C-D and the B loop (that upon fusion becomes B helix [1]). No meaningful variation - in terms of secondary structure - is found in the α subregion of the stem, because structural changes only concern the B loop [see Additional file 1, panel B], which indeed is unfolded in the pre-fusion state. The B loop coordinates depend on crystallization conditions and in particular on pH [14]. The VED-proximal and distal β subregions are recognized by respectively antibodies CR6261 and CR8020 [24]. The VED-proximal β subregion shows a varying number (zero, two or four) of β strands [see Additional file 1, panel C] and such variation is not relevant to antibody recognition specificity. For instance, a four-strands structure is shared between H5 (recognized by CR6261) and H3 (not recognized); moreover, a two-strands structure is shared between H2 (recognized) and H7

(not recognized). Secondary structure variation is evident also in the distal β subregion [see Additional file 1, panel D], but once again it does not relate to antibody recognition: e.g., CR8020 recognizes subregion from H7 but not corresponding one from H5. Given that subregions recognized by each antibody are clearly different (CR6261 recognizes H1, H2, H5 and H9 independently on they are showing either zero, two or four β strands) such a preliminary analysis demonstrates that secondary structure variation as viewed by cartoon representation is not indicative for epitope variation. Secondary structure variation in the globular RBD-VED region is poorly evident, according to the aforementioned 'two-tracks' rule: mutations altering the overall backbone/fold of the RBD would impair binding to host cells hence conservation (track 1) is needed to keep such basic function. However, local variation (track 2) is needed to modulate surface features hence interactions. Therefore, we did not further investigate secondary structure variation and moved instead to surface analysis, considering both most relevant features: (i) electrostatic charge distribution and (ii) hydropathy/hydrophilicity patches.

Comparative analysis of electrostatic potentials

In order to perform analyses taking into account the influence of ionic strength (I), the spatial distribution of the electrostatic potential was calculated at both I = 0 mM (Coulombic interactions unscreened by counter-ions) and I = 150 mM (physiological), assuming +1/-1 charges for the counter-ions. Prior to electrostatic potential calculations, partial charges and van der Waals radii were assigned with PDB2PQR [25,26]; then, linear Poisson-Boltzmann (PB) equation calculations were carried out by using Adaptive PB Solver (APBS) [27] through Opal web service (see Methods). The spatial distribution of the electrostatic potential was determined for each HA subregion, monomers and trimers, comparing the six available HA structures to identify possible HA-specific signatures. In particular, we focused on the role of charge distribution as visualized by isopotential contours within the tertiary structure and on classifying conservation and divergence among the different HAs. In order to evaluate electrostatic distance (ED) also in a quantitative way, clustering of the spatial distributions of the electrostatic potentials was obtained by WebPIPSA (Protein Interaction Property Similarity Analysis; [28], having the use of Hodgkin and Carbo similarity index (SI) [29] (see Methods). The Carbo SI is sensitive to the shape of the potential being considered but not the magnitude, whereas the Hodgkin SI is sensitive to both shape and magnitude. Therefore, WebPIPSA results obtained using the Hodgkin SI are shown in Figures 1, 2, 3, 4 and 5, and evidence from analyses performed using the Carbo SI is cited to confirm parameter independent data.

Stem subregions

The electrostatic patches at ionic strength I = 0 mM clearly show for all six stems preferential side disposition (Figure 1, top left), as observed for SNAREs [30]. In particular, density of negative potential (red) at the 0° side is higher than at the 180° side; positive potential (blue) shows a reverse distribution, highest density being at the 180° side. At physiological ionic strength (Figure 1, top right), preferential distribution of the positive potential (180° side) is more evident, whereas higher density in negative potential (0° side) is less evident, because most Coulombic interactions are masked by counter-ions. When considering individual stem variation, net charge roughly doubles from the −8 *e* value of H1 and H9 to −15 *e* of H7. However, similar net charge does not necessarily correspond to similar distribution (along the stem) of the potential, that can preferentially locate at either the VED-distal stem subregion (left side in figure) or at the VED-proximal one (right side). This is the case for H1 and H9 stem, sharing net charge −8 *e*, and showing (more evident at I = 0 mM) preferential VED-distal and VED-proximal negative potential, respectively. Such preferential VED-distal location of the negative potential shown by H1 is conserved also in the other two stems from Group 1, in spite of their different net charge (−10 *e*). Positive potential is more homogeneously distributed along all stems. Heat maps and corresponding density plots (Figure 1, bottom) depict the overall similarity among HA stem electrostatic profiles. Comparison between the density plots at I = 0 mM and I = 150 mM highlights a general increase in distance, i.e. a peak shift from middle ED (green region) to high ED (cyan/blue region). When comparing Group 1 stems to those from other groups it can be noticed that - at both ionic concentrations - H3 is slightly closer to Group 1 than H7, while H9 is far apart. However, H9 distance is not homogeneous with respect to the three Group 1 stems, as it is closer to H2 than to H1 and H5. Indeed, H9 stem is also quite far from H7 because it shows the highest overall distance, with respect to other stem structures. When using WebPIPSA, the distance matrix of the electrostatic potential can also be displayed as a tree referred to as 'epogram' (electrostatic potential diagram). Epograms [see Additional file 2] further highlight at both ionic concentrations that: (i) H9 stem shows unique electrostatic features (i.e., the highest ED with respect to other stems) and (ii) H7 is closer to H3 than to other stems. This clustering is confirmed when using Carbo SI. The highest electrostatic distance shown by H9 might depend on its mammalian (swine) rather than avian origin. Therefore, structural models were obtained by homology modeling for avian H9 (A/Chicken/Jiangsu/H9/2010(H9N2), UniProtKb AC: G8IKB3) and horse H3 (A/Equine/Mongolia/56/2011(H3N8); UniProtKb AC: J9TJ60),

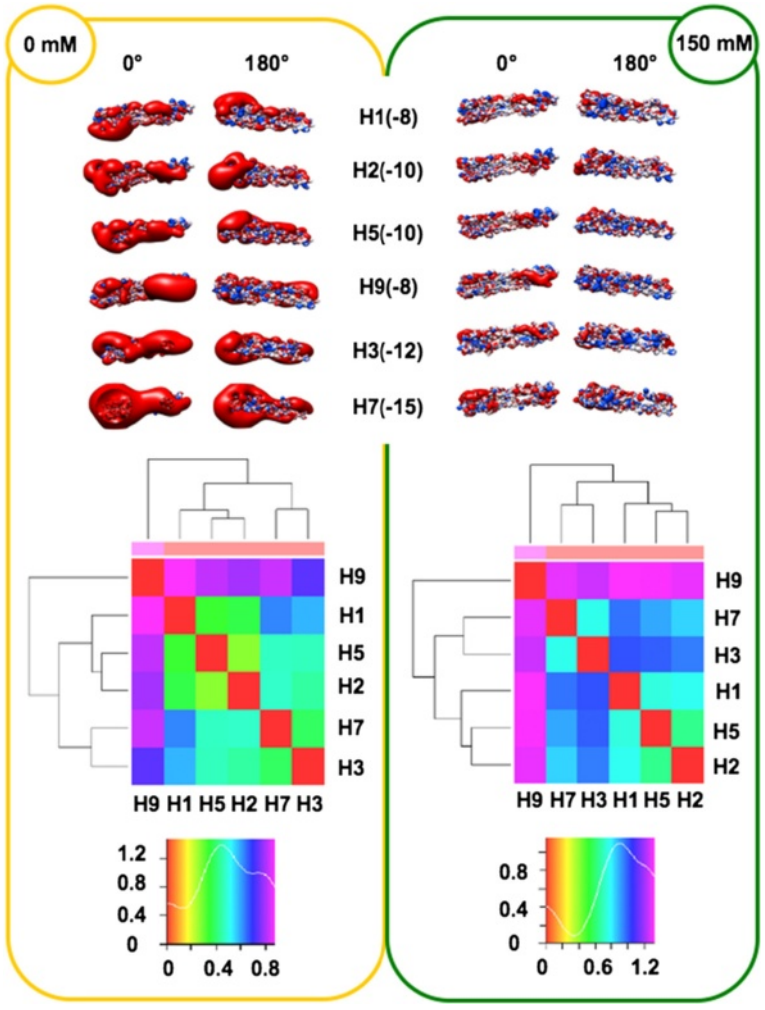


Figure 1 Isopotential contours (top), heat maps (middle) and density plots (bottom) of HA stems. Electrostatic features are shown at I = 0 mM (left panel, yellow) and I = 150 mM (right panel, green). Electrostatic potentials (blue for positive and red for negative) are presented in two orientations (0° and 180°). Isopotential contours are plotted at $\pm 3k_BT/e$. In heat maps, red, warm and cold colors correspond to identity, low and high Electrostatic Distances (ED), respectively. The number of hits (pairs of compared structures) is plotted and referred to as 'Density'.

using as structural templates 1JSD (H9) and 1MQL (H3), respectively and investigated using WebPIPSA. Comparison of epograms alternatively including either the avian H9 model or the swine template showed conservation of the highest distance observed for H9: at I = 0 mM, swine/avian epogram clustering was congruent; at I = 150 mM, avian H9 sorted with H3 and H7; this notwithstanding, highest distance of H9 from other HAs was anyway kept [see Additional file 3]. Concerning equine H3, it sorted like avian H3 at both I = 0 mM and I = 150 mM (congruent epograms see Additional file 3). In conclusion, electrostatic distance is not significantly influenced by taxonomy hence segregation depends on HA-specific features.

RBD subregions

As with the stem subregion, charge separation onto the RBD surface is more evident at I = 0 mM. Group 1 RBDs have an overall slightly negative (H1 and H2) or neutral (H5) net charge, which is positive (up to +3*e* in H3) in other groups. At large, the RBD net charge is less negative than stems (Figure 2, top). Side disposition in RBDs is not 'side preferential' as for stems, and no meaningful difference is observed when comparing the 0° and 180° views. However, preferential local distribution is clearly apparent also for RBDs, when a roughly orthogonal axis is considered: negative charges are densely distributed at the VED-proximal region (left side in figure), whereas charge of the VED-distal region (right side) is more positive. This is particularly evident for Group 1 RBDs at I = 0 mM. At physiological ionic strength, such preferential distribution is less evident, in particular for H3, where differently charged patches are interspersed. Peaks at the blue/purple regions in density plots (Figure 2, bottom) depict high electrostatic distances

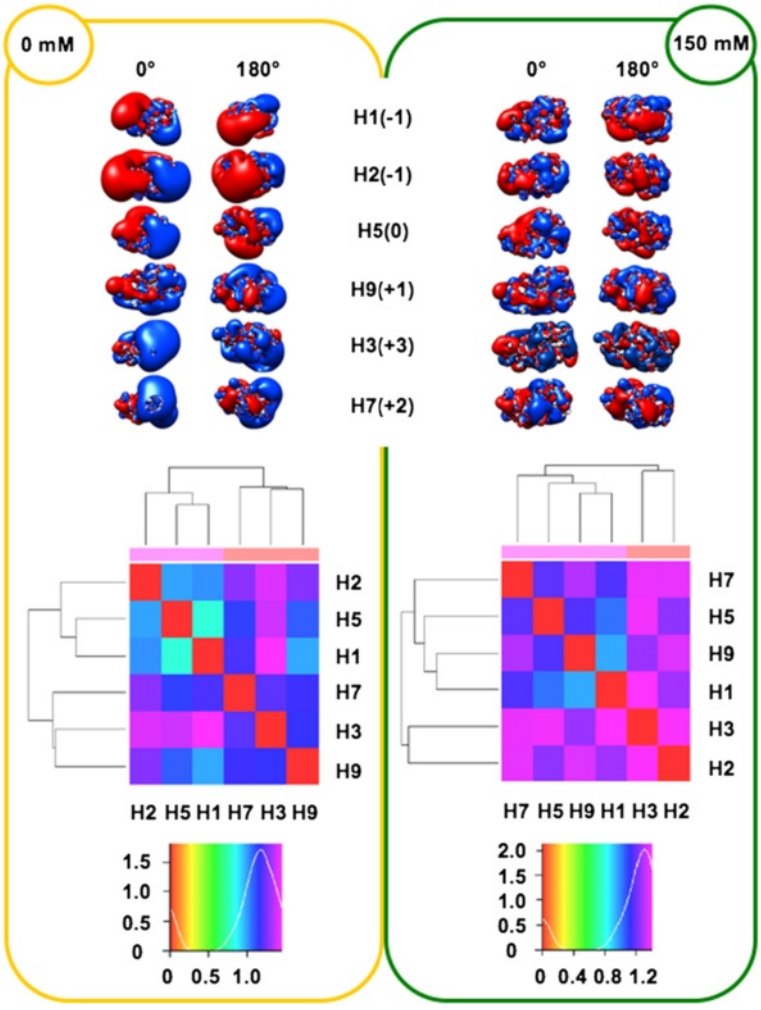


Figure 2 Isopotential contours (top), heat maps (middle) and density plots (bottom) of HA RBDs. See Figure 1 caption for color code and definitions. Isopotential contours are plotted at $\pm 1 k_B T/e$.

at both ionic strengths. Surprisingly - and independently on using either Hodgkin or Carbo SI - at I = 150 mM, the electrostatic potential of the H5 RBD is closer to H9 and H7 than to RBDs from H2, in spite H5 and H2 belong to the same Group. Splitting of Group 1 is confirmed by epogram [see Additional file 2] at I = 150 mM: H5 and H1 create a new cluster with H7 and H9.

HA1 subregions

Once the electrostatic analysis is repeated for the whole HA1 region, including the VED and F' subregions in addition to the RBD [14], the most evident difference is an overall shift towards net positive charge (see upper panels in Figures 2 and 3), according to the presence of basic patches in F' subregions [2,6]. Comparison of density plots (RBD vs. HA1) shows that peaks similarly locate at the high distance blue/purple regions (see lower panels in Figures 2 and 3) but, at I = 150 mM, Group 1 no longer splits, as H1, H2 and H5 form a cluster including H9. Resembling RBD distances, it also occurs with HA1 that members from Group 1 (H1 and H5) can be closer to an outgroup (H9) than to a member of the same group (H2) (see at I = 150 mM both heat map in Figure 3 and epogram in Additional file 2). This parameter independent evidence further highlights the relevance of counter-ions to shape the final electrostatic profile, as well as the possible disagreement between classic clustering (based on phylogenetic and serologic data) and electrostatics of the RBDs.

Monomers

The net charge is negative for all monomers, ranging -4*e* to -11*e* (Figure 4, top). Evidence that the net charge is quite negative for all stems (−8*e* to -15*e*) while being close to 0 for RBDs (−1*e* to +3*e*), stresses the total charge balancing by local basic patches in VED and F'

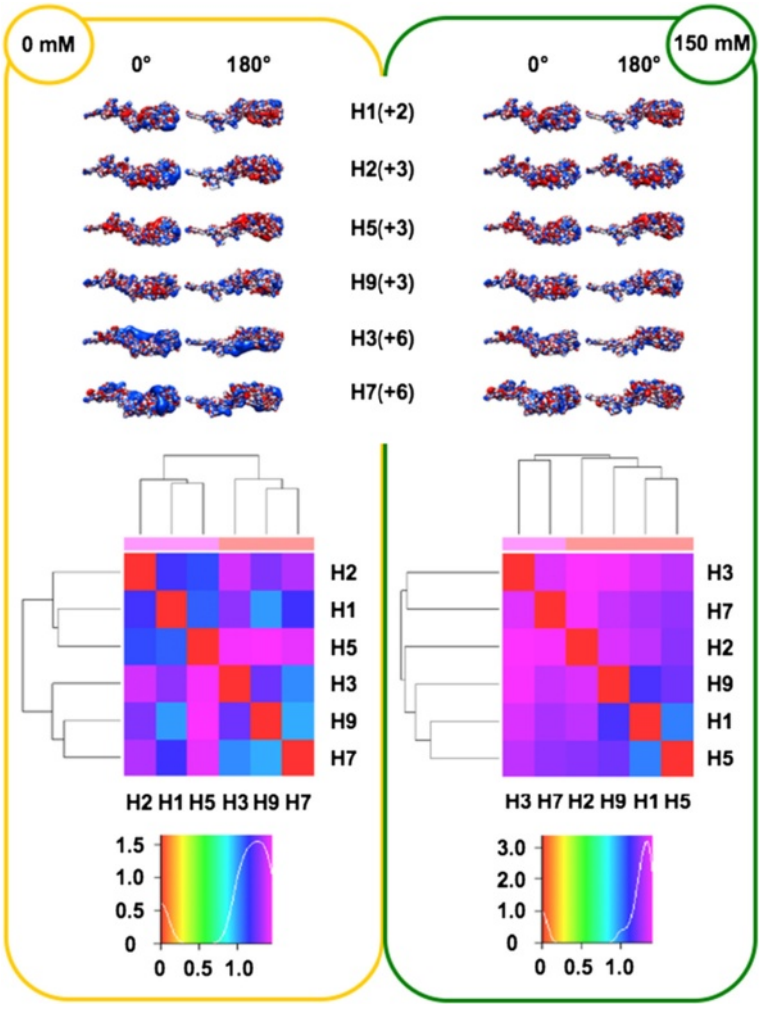


Figure 3 Isopotential contours (top), heat maps (middle) and density plots (bottom) of HA1 subregions. See Figure 1 caption for details. Isopotential contours are plotted at $\pm 2k_BT/e$.

subregions. Once again, peculiar electrostatic features are evident (and SI independent) for H9, characterized by the less negative net charge and forming its own branch at both I = 0 mM and I = 150 mM (heat maps in Figure 4, bottom, and epograms in Additional file 2). Disagreement with serological and phylogenetic data is less evident when performing electrostatic analysis with entire monomer structures, as shown by clustering of Group 1 members in Figure 4 and Additional file 2.

Trimers

Once the entire haemagglutinin functional unit is analyzed, disagreement with serological and phylogenetic clustering is highlighted again by Group 1 splitting; in particular (and independently on which SI is used) at I = 0 mM, H1 sorts separately from H2 and H5 (see Figure 5, trimer heat maps and Additional file 2, trimer epograms). Such splitting is also observed at I = 150 mM, as H5 and H1 sort with H9 and H7, whereas H2 sorts out with H3. Comparison of net charges from monomers and corresponding trimers unveils striking doubling vs. triplication mechanisms: trimer net charge values for H1 and H3 is roughly three-fold with respect to corresponding monomers, or even more (–37*e* vs. -11*e*) for H5. Instead, trimer values are only roughly twofold increased for H2, H7 and H9. Therefore, different orientations of monomers within corresponding trimers results in significant modulation of the trimer surface electrostatic charge and this in turn can be quite relevant to HA interactions. Different HA clustering at I = 0 mM and I = 150 mM may highlight the importance of ionic screening of coulombic interactions [31,32]. As a final remark, based on absence of net charge-based clustering in any executed electrostatic analyses, the spatial distribution of electrostatic potential is suggested to be more suitable than net charge alone for eventual use as a further 'signature' for protein/domain function.

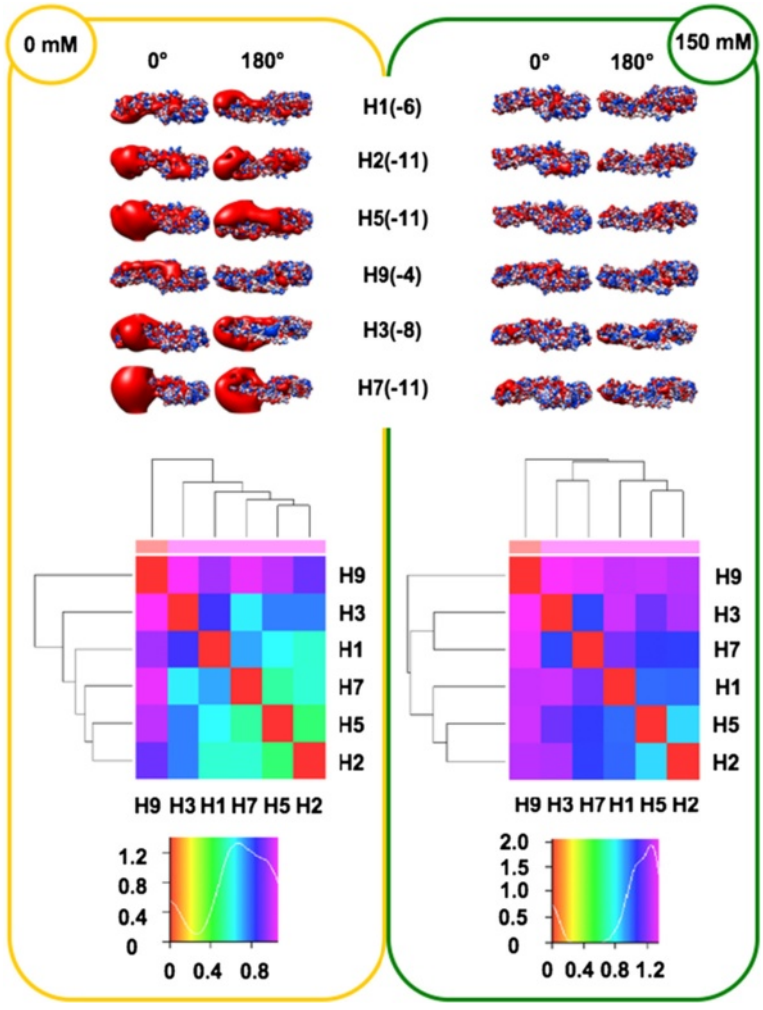


Figure 4 Isopotential contours (top), heat maps (middle) and density plots (bottom) of HA monomers. See Figure 1 caption for details. Isopotential contours are plotted at $\pm 2k_BT/e$.

Hydrophobicity analysis

Search for HA-specific motifs/signatures can be integrated by hydropathy analysis. Both electrostatics and hydrophobicity are key determinants in surface properties hence in regulating protein interactions. In particular, hydrophobic patches located at the protein surface create unstable areas. The identification of well-defined patches rather than a 'patchwork surface' of hydrophobic and hydrophilic areas can thus shed light on molecular evolution of haemagglutinin. Stem, RBD and HA1 profiles were obtained and compared using ProtScale [33] and Protein Hydrophobicity Plots [34]. Profiles from the stem subregions did not unveil any clearly meaningful difference and thus are not shown here.

RBD subregions

Figure 6 shows GRand AVerage hYdrophobicity (GRAVY) indexes, Kyte-Doolittle plots and 0° +180° surface hydropathy views for the RBDs from the six available HA structures. Similar to total electrostatic charges, GRAVY indexes are reported here for completeness of information; however, they are not suitable for use as evolutionary or functional fingerprint. In fact, variation of GRAVY values amongst the six RBDs does not correspond to high conservation and fine tuning of their surface patches as depicted in 0° and 180° views. However, comparison of Kyte-Doolittle plots could infer variation at specific positions. Plots in Figure 6 always start by residue 1 because the default numbering system from the software refers to analyzed sequence fragments (RBDs in this case); therefore, for Reader's convenience, hereafter we report both real numbers (referring to complete protein sequences) and software output numbers (between parentheses). Within Group 1, the highest intra-group hydrophilicity is shown by H1 positions Arg223 (160) of the 220-loop and by H2 at

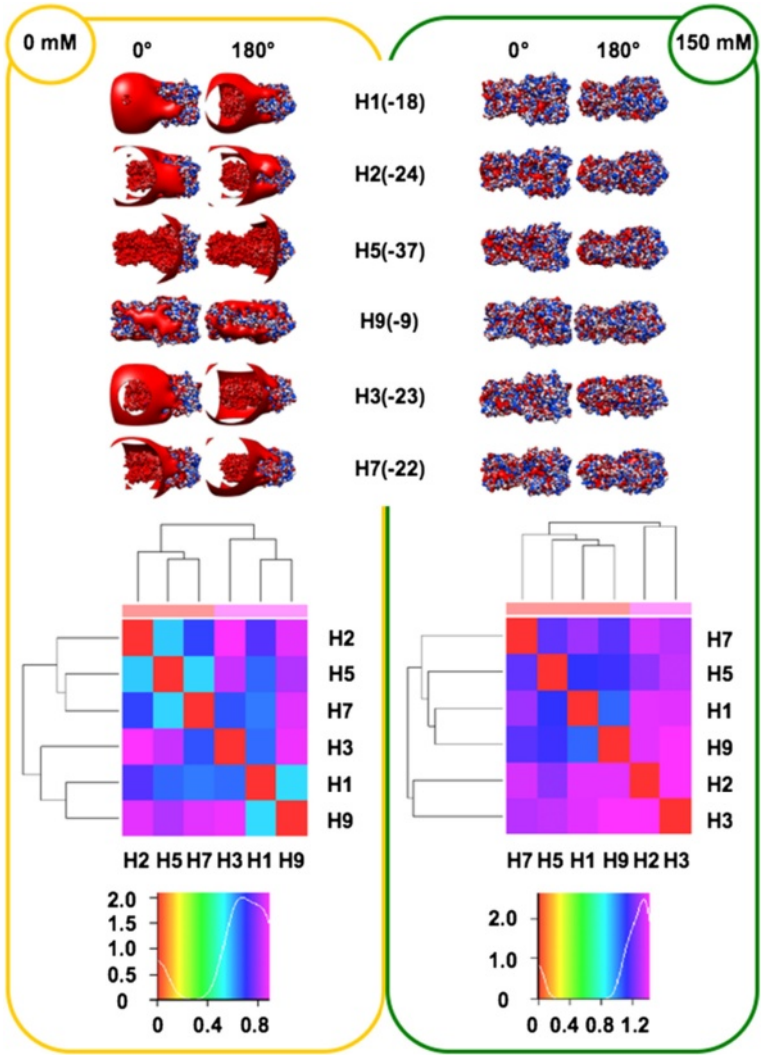


Figure 5 Isopotential contours (top), heat maps (middle) and density plots (bottom) of HA trimers. See Figure 1 caption for details. Isopotential contours are plotted at $\pm 3k_BT/e$.

positions Asn80, Ser136 and Glu202 (17, 73 and 139). At position 112 (49), H1 is significantly more hydrophobic (Ile) than H2 and H5 (Asn). Inter-group comparison highlights in H3 three hydrophilic peaks centered on residues Asp191, Thr208 and Gln227 (114, 135 and 154), as well as increased hydrophobicity of H7 in subregion 105–155 (50–100). Comparative analysis of surface patches unveiled possible HA-specific fingerprints. Within Group 1, variation concerns both the VED and RBD subregions. Such variation is even more evident when extending comparison to H9, H3 and H7. Hydrophobic patches (light and dark orange) are variable in terms of position and area. Comparison of 0° views highlights a large orange surface encompassing the VED-RBD border, specific to H9. Moreover, H5 and H7 show at the VED subregion a hydrophilic (violet) surface (green ovals) that in other HAs includes at least one small orange patch. Comparison of 0° views shows that H2 and H3 share three hydrophobic spots in an RBD subregion (blue circles) where other HAs can lack one, two or even all such spots. Further variation can be observed, and in general it seems to concern 'position-shifting' rather than significant difference in the total ratio of hydrophilic/hydrophobic surfaces. Therefore, combined variation in both electrostatic and hydropathy features is likely to fine tune local interaction properties of the different HA RBDs.

HA1 subregions

Apart from differences already observed in the RBD subregion, no further meaningful variation was found among HA1 hydropathy profiles. The only relevant evidence concerns the hydrophilicity peak at position 297 in H3 haemagglutinin (not shown).

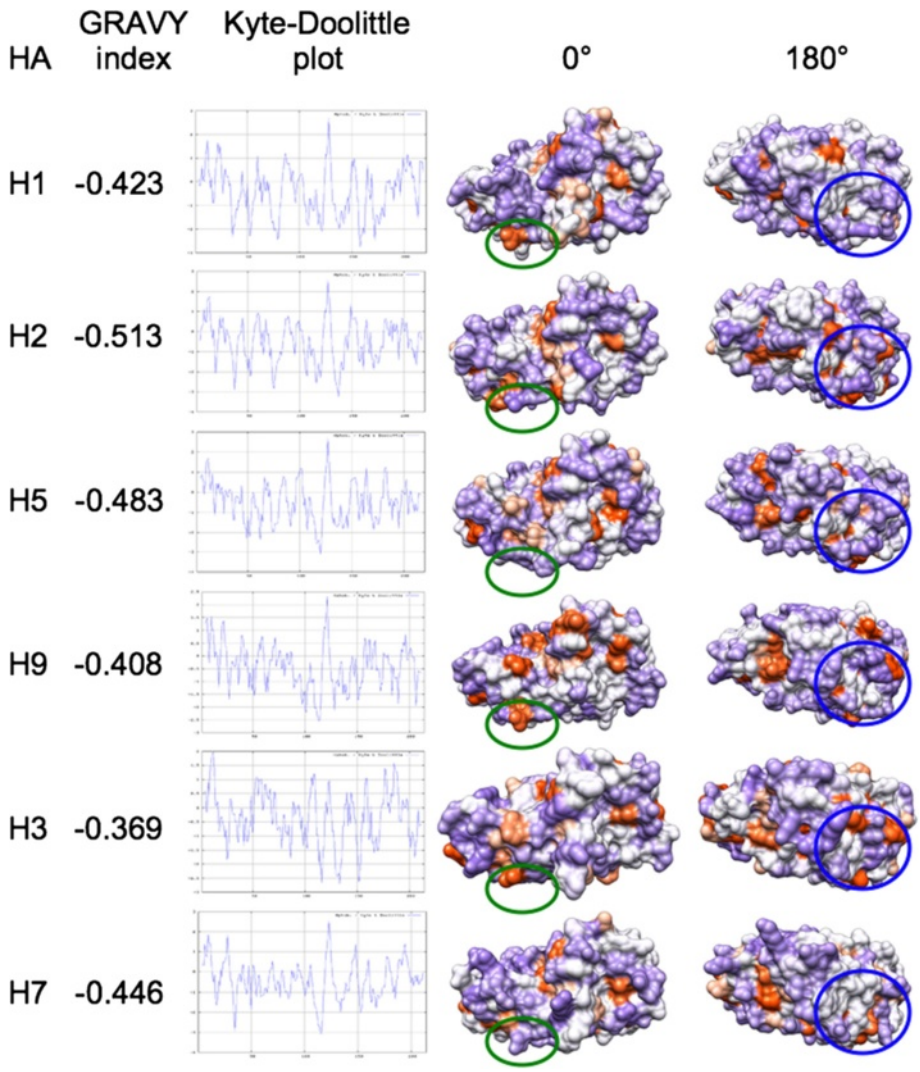


Figure 6 Hydrophobicity analysis of the RBD subregions from the six available HA structures. GRAVY Index, Kyte-Doolittle plots and surface hydrophobic (orange) and hydrophilic (violet) patches (as both 0° and 180° views) are depicted.

Structural modeling of H5N1 clades and electrostatic features comparison

Electrostatic features can vary among different types of haemagglutinins (see above). This prompted us to further investigate on differential electrostatic features as a possible fingerprint for monitoring viral evolution, i.e. as a tool to distinguish among circulating/spreading and extinguished H5N1 clades. Table 2 resumes relevant data concerning the ten clades used for this analysis; their geographical spread is shown in Figure 7. Spreading of no longer circulating clades (0, 3, 4, 5, 6, 8 and 9) is restricted to the eastern part of China and to Vietnam (see Figure 7, zoom in map); noticeably, all such clades share one or more outbreak areas with the most ancient clade (clade 0, black spots). Among circulating clades, clade 7 was also found in western China and clade 1 also spread towards India and Indochina countries (Thailand, Laos, Cambodia and Malaysia). The widest spreading concerns circulating clade 2 (red dots in the upper map of Figure 7), having reached Japan and Korea, Mongolia, Russia, several countries from Middle-East and Europe (including UK) as well as a number of African countries from the Northern hemisphere. So far, spreading of H5N1 viruses neither concerns Americas nor any country from the Southern hemisphere (Oceania and sub-equatorial Africa).

Based on a very high, average % identity (over 90%) of the clade target sequences with the available structural H5 template (PDB: 3S11), structural models for clades 0 to 9 were obtained by homology. Given that distribution of surface charge is strongly influenced by the orientation of side chains, models refinement was performed using a number of tools based on different algorithms: SCWRL [35,36], ModRefiner [37] and SCit [38]. Then, QMEAN server was used to check model quality;

Table 2 H5N1 clades

Clade	Year	Strain name	Genomic Ac	Protein Ac
0	1996-2002	A/Goose/Guangdong/1/1996	AF144305.1	AAD51927.1
1 (c)	2002-2003	A/Quail/Shantou/3054/2002	CY028946.1	ACA47648.1
2 (c)	2005	A/Bar-headed Gooze/Qinghai/75/2005	DQ095619.1	AAZ16276.1
3	2000-2001	A/Duck/Hong Kong/2986.1/2000	AY059481.1	AAL31387.1
4	2002-2003 2005-2006	A/Duck/Shantou/700/2002	CY028943.1	ACA47615.1
5	2000-2003 2004	A/Duck/Zhejiang/52/2000	AY585377.1	AAT12042.1
6	2002-2004	A/Duck/Hubei/wg/2002	DQ997094.1	ABI94747.1
7 (c)	2002-2004 2005-2006	A/Chicken/Shanxi/2/2006	DQ914814.3	ABK34764.2
8	2001-2004	A/Chicken/Hong Kong/61.9/2002	AY575876.1	AAT39076.1
9	2003-2005	A/Duck/Guangxi/50/2001	AY585375.1	AAT12040.1

Periods (years) of circulation, strain names (based on year and location of identification) and accession numbers (for both genomic and protein data) are reported for each clade. Circulating clades are marked by (c).

QMEAN is a scoring function that measures multiple geometrical aspects of protein structure, ranging 0 to 1 with higher values indicating more reliable models [39]. QMEAN scores for each refined or not refined model (mQMEAN) and the average QMEAN score for each ten clades model series (aQMEAN) was calculated. Models refined by SCWRL showed the highest aQMEAN (0.734), with highest mQMEAN for clades 0, 1, 2, 3 and 5. However, quality was similarly good when models were not refined (aQMEAN: 0.724; highest mQMEAN for clades 6 and 7) or refined by ModRefiner (aQMEAN: 0.720; highest mQMEAN for clades 4, 8 and 9), confirming once again reliability and robustness of the SWISS-MODEL homology modeling method [40]. SCit refined models showed the lowest average quality (aQMEAN: 0.702). Therefore, electrostatic analyses were performed thrice, using the ten clades models: (i) refined by SCWRL, (ii) refined by ModRefiner and (iii) not refined.

Preliminary comparison at trimer and monomer level showed meaningful variation only at the VED-RBD subregion. In fact, direct comparison of stems did not allow for inferring any clade-specific signature as all clades were found to share - at both I = 0 mM and I = 150 mM - the typical isocontour of the H5 stem (see Figure 1, top). Moreover, apart from electrostatic differences in the VED-RBD subregion, no further meaningful variation was observed among HA1 isocontours. This prompted us to 'zooming in' variation analysis at the RBD subregion level.

Figure 8 illustrates local charge variation in RBD isocontours among H5N1 clades. Even though variation is more evident at I = 0 mM, meaningful difference is kept hence highlighted at physiological ionic strength. It is noteworthy that, independently on models are refined or not and on algorithm used for refinement, the same relevant local changes in RBD isopotential contours are found (see Figure 8, panels A to C). Early clades evolution is characterized by a charge shift event at the 220-loop: in the most ancient clade (clade 0), the side chain of amino acid 228 shows either negative (Glu: 50/89 and Asp: 1/89 sequences) or positive (Lys: 38/89 sequences) charge. The positive charge is 'fixed' in the most recent, and still circulating clades 2 (Lys: 308/310, Glu or Asp: 0/310 sequences) and 7 (Lys: 25/26; Glu: 1/26 sequences) (see Figure 8 and Table 3). Further loss of a negative residue (Asp) concerns the VED isocontour at the 110-helix region. Table 3 shows that in clade 0, position 110 is negatively charged (Glu or Asp: 67/89 sequences) or polar, non-charged (Asn: 22/89 sequences). This negative charge is almost completely lost in clade 2 (Asp: 3/310, Glu: 0/310), while being retained (Asp: 26/26) in clade 7; however, this latter clade shows ongoing loss of the negative charge at position 104 (Asp: 15/26; Gly: 11/26), that is positively charged in 100% of clade 0 and clade 2 sequences (Figure 8 and Table 3). In clades 2 and 7, such 'denegativization' of the VED isocontour is somehow counterbalanced by negativization (or depositivization) at the properly receptorial part of the RBD. In clade 2, this depends on Asn140Asp mutation (in 307/310 sequences) while in clade 7 both depositivization (Arg178Val in 8/26 sequences) and negativization (Ala200Glu in 12/26 sequences) mutations are observed (Figure 8 and Table 3). Intriguingly, when considering aforementioned replacements altogether, evolution of H5N1 still circulating clades seems having been characterized by an isocontour rearrangement based on a VED-to-RBD flow of negative charges; this process is 'partial' hence seemingly in progress in clade 7 (mutation arose in the clade and it is present, at least so far, in less than 50%

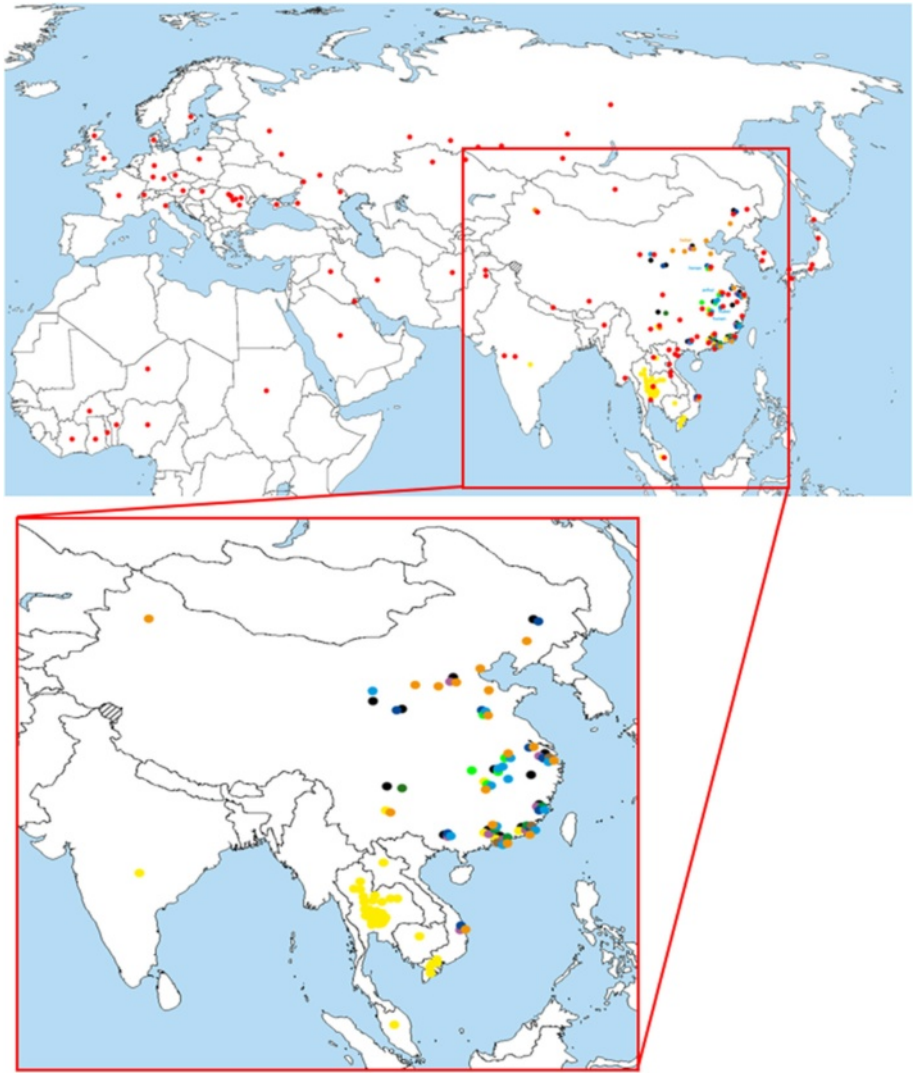


Figure 7 Geographical spread of H5N1 clades. Outbreak areas for each clade are color coded as follows: 0, black; 1, yellow; 2, red; 3, violet; 4, dark green; 5, dark blue; 6, light green; 7, orange; 8, brown; 9, cyan.

sequences) whereas it is complete and 'fixed' (99% sequences) in clade 2. Given that comparison of the six different HA structures identified HA-specific variation in both electrostatic and hydropathy features, and that specific electrostatic signatures of the RBD could also be associated to the ten H5N1 clades, clades analysis was integrated by comparison of the RBD surface hydropathy profiles (Figure 9). As for electrostatic analysis, the most ancient clade (clade 0) is the reference for tracking hydropathy profile variation along clades evolution. As previously explained, hereafter both real protein sequence numbering and (between parentheses) software output numbering is reported for Reader's convenience. Clade 3 shows no substantial difference with respect to clade 0, at least in terms of hydropathy plots. Instead, clade 4 shows increased hydrophilicity at position Asn211 (148). Clade 1 shows increased hydrophobicity around position Ser140 (77). Replacement at position 124 of a polar residue in clade 0 by Ile in all other clades results in increased hydrophobicity. Intriguingly, the hydropathy profile of clade 7 resembles the one of H3 haemagglutinin, including its aforementioned three hydrophilicity peaks. Please note that the apparent disagreement among positions of the three H3 peaks in Figure 6 and those from Clade 7 in Figure 9 is not confirmed in real numbering, as plot shift is determined by ten extra residues present in the really N-terminal region of H3. Apart from difference illustrated so far for the RBD, no further meaningful variation was observed when comparing other HA1 subregions or the stem profiles (not shown).

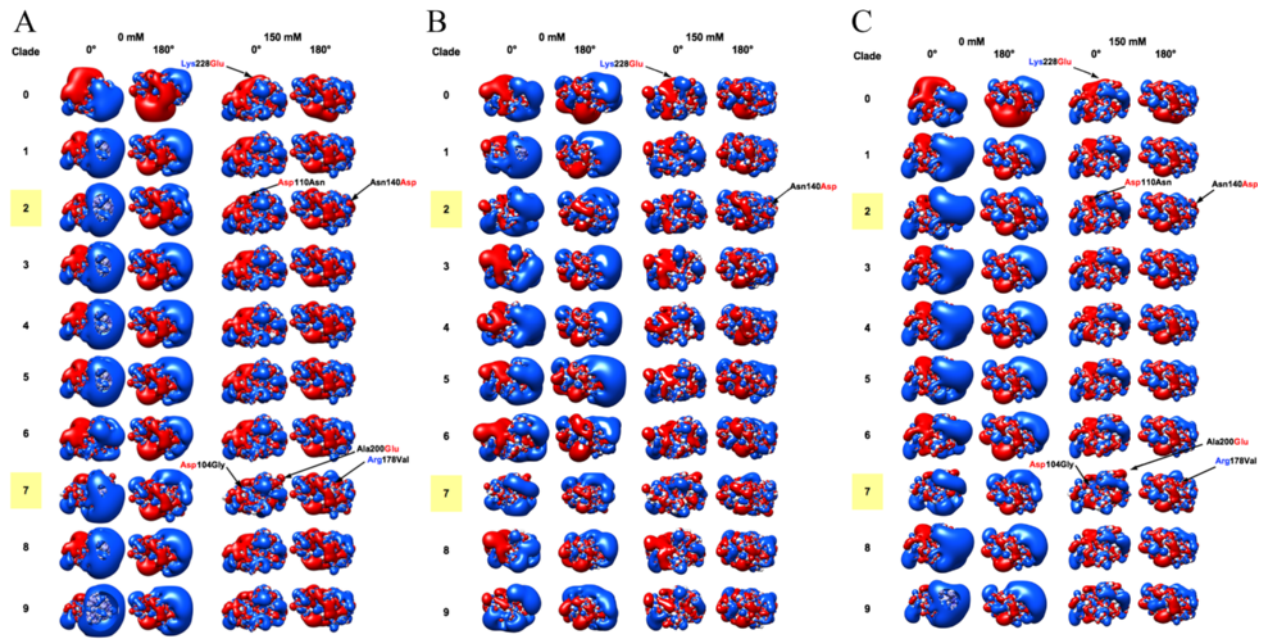


Figure 8 Isopotential contours of the RBD subregions from H5N1 clades 0 to 9. See Figure 1 caption for color code and definitions. Isopotential contours are plotted at $\pm 1 k_B T/e$. Specific mutations discussed in the text are highlighted. Positively charged residues are written in blue and negatively charged ones in red. Panels: models refined using SCWRL **(A)**, ModRefiner **(B)** or not refined **(C)**.

Conclusions

Evidence from this work shows that sequence homology is often, but not always, related to structural similarity and vice versa. In fact, in some instances, protein domains with less related sequences can show intriguing structural closeness. Therefore, in order to obtain a more complete view of the 'functional evolution', phylogenetic analyses based on sequence comparison and resulting in trees, might be integrated taking into account information from structural comparison. Dissimilarity in secondary structure elements does not always result in different antigenic properties. Sometimes, secondary structure is not prominent to the molecule antigenicity. Indeed, electrostatic features are crucial to interactions and in fact electrostatic profiles of the RBD subregion varies amongst different HAs. On the other hand, stems, HA1, monomers and trimers topology appears to be variable. As shown by H9 and H3 modeled structures, electrostatic profiles seem to depend on HA type rather than organism source. Hydrophobicity analysis reveals that local, 'spot' variation especially concerns the RBD subregion. No flow of hydrophobicity/hydrophilicity

Table 3 Mutations in H5N1 clades 0, 2 and 7

Clade	Sequences	Position					
		104	110	140	178	200	228
0	89	Asp = 89	Asp = 64	Asn = 86	Arg = 89	Ala = 89	Glu = 50
			Asn = 22	Asp = 3			Lys = 38
			Glu = 1				Asp = 1
2.2	310	Asp = 310	Asn = 302	Asp = 307	Arg = 284	Ala = 307	Lys = 308
			Lys = 4	Asn = 2	Ile = 26	Gly = 3	Asn = 1
			Asp = 3	Gly = 1			Gln = 1
			Ser = 1				
7	26	Asp = 15	Asp = 26	Asn = 24	Arg = 16	Ala = 14	Lys = 25
		Gly = 11		Asp = 2	Val = 8	Glu = 12	Glu = 1
					Gly = 2		

For each clade, the number of analyzed available sequence is shown. For each position (numbering refers to clade 0 sequence), the type of present residues and corresponding number of sequences showing that residue is shown.

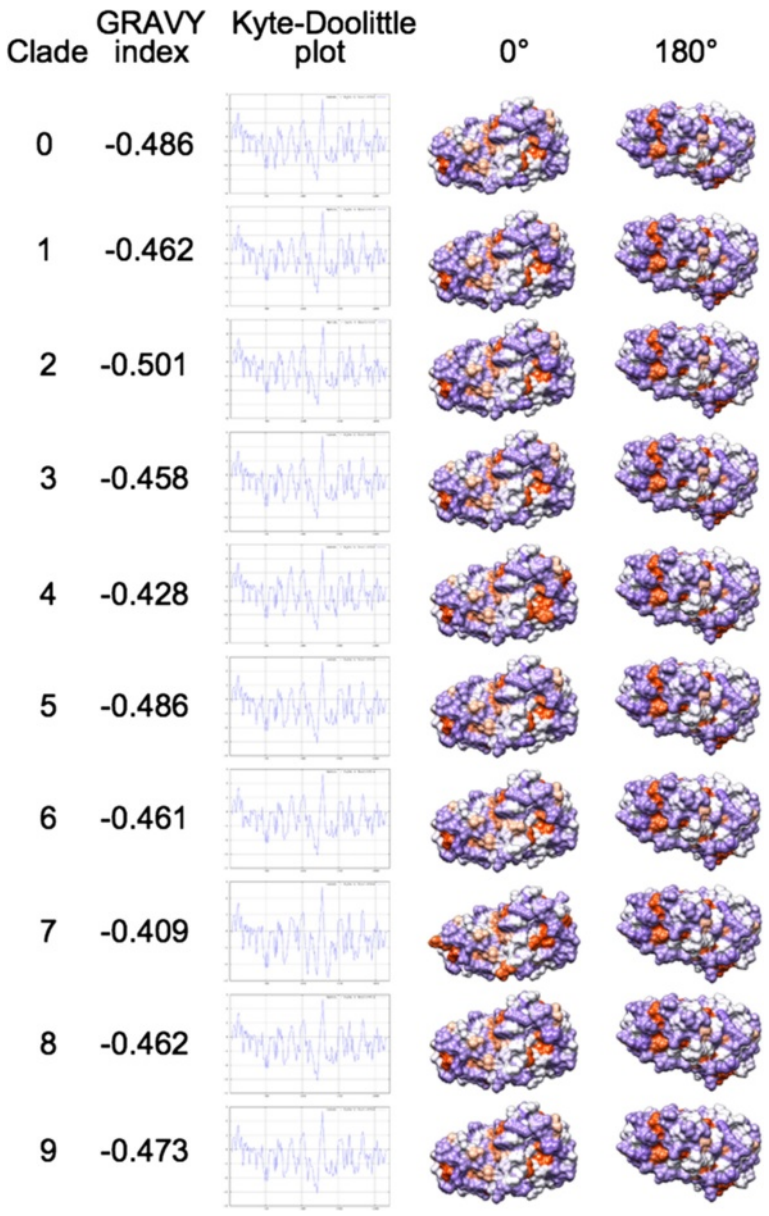


Figure 9 Hydrophobicity analysis of the RBD subregions from H5N1 clades 0 to 9. See Figure 6 caption for color code and definitions.

is observed as for charge flow in the electrostatic analysis. In H5N1 clades comparison, from an electrostatic point of view, meaningful variation concerns only the VED-RBD subregion. Intriguingly, a charge flow specifically concerns still circulating clades 2 and 7, where 'denegativization' of the VED isocontour is counterbalanced by negativization in the RBD. It is noteworthy (and a 'positive mark' for robustness of the observation) that the same specific differences are found when comparing refined or not refined clade models or models refined using different algorithmic strategies (as SCWRL is rotamer library-based [35,36] while ModRefiner is based on two-step atomic-level energy minimization [37]). Given that local charge concentration is typical for antigenic epitopes, it is tempting to speculate that charge redistribution in such clades might have contributed to antigenic escape hence to their evolutionary success and spreading. Indeed, such an hypothesis is in agreement with evidence that charge redistribution on the RBD characterizes the two clades (2 and 7) which were able to spread over the largest geographical distribution and that, in particular, such redistribution is fixed in sequences from clade 2, which is the world most spread clade. It is noteworthy that also variation in hydrophobic patches is especially observed in the RBD subregion.

Methods

Structural templates and target sequences

The following structures from the Protein Data Bank (PDB) were used as templates for modeling: H1, PDB 1RUZ, from viral strain A/South Carolina/1/1918(H1N1); H2, PDB 2WR5, from Asian pandemic influenza virus of 1957; H3, PDB 1MQL, from viral strain A/duck/Ukraine/1963 (H3N8); H5, PDB 3S11, from viral strain A/Goose/Guangdong/1/1996 (H5N1); H7, PDB 1TI8, from viral strain A/turkey/Italy/214845/2002(H7N3); H9, PDB 1JSD, from viral strain A/swine/Hong Kong/9/98(H9N2). UniProtKb accession codes (AC) of target sequences modeled by H.M. and corresponding viral strains (VS) are the followings: H4, AC F2NZ53, VS A/duck/Guangxi/912/2008(H4N2); H6, AC H8PBW2, VS A/duck/Fujian/6159/2007(H6N6); H8, AC D4NQL7, VS A/northern pintail/Alaska/44420-106/2008(H8); H10, AC P12581, VS A/Chicken/Germany/n/1949 (H10N7); H11, AC D5LPX8, VS A/turkey/Almaty/535/2004(H11N9); H12, AC E6XYK2, VS A/mallard/Interior Alaska/9BM1907R1/2009(H12); H13, AC P13101, VS A/Gull/Astrakhan/227/1984 (H13N6); H14, AC P26136, VS A/Mallard/Astrakhan/263/1982 (H14N5); H15, AC Q82565, VS A/duck/Australia/341/1983(H15N8); H16, AC Q5DL23, VS A/black-headed gull/Sweden/3/99(H16N3). Given that original UniProtKb sequences indeed correspond to H0 precursors, sequence fragments missing in mature chains were manually removed to avoid improper structural alignment.

Structural superpositions, Homology Modeling, model refinement and quality check

Structural superpositions were performed and viewed using UCSF Chimera [18] v. 1.8.1 (free download from [41]). Target protein sequences were modeled on best available structure templates using SWISS-MODEL [40]. Then, model structures were refined using SCWRL [35,36], ModRefiner [37] or SCit [38]. Model quality was checked via QMEAN server [39].

Electrostatic surface analysis

Isopotential contours were calculated using UCSF Chimera 1.8.1: the software utility allows for connecting - through Opal web server - to the Adaptive Poisson-Boltzmann Solver (APBS) server [42]. Isopotential contours were then plotted at $\pm 3k_BT/e$, $\pm 2k_BT/e$ and $\pm 1k_BT/e$ (RBDs). PDB2PQR was used to assign partial charges and van der Waals radii according to the PARSE force field [43]. Interior $\varepsilon_p = 2$ and $\varepsilon_s = 78.5$ were chosen for respectively the protein and the solvent [30,44,45], T = 298.15 K. Probe radius for dielectric surface and ion accessibility surface were set to be r = 1.4 Å and r = 2.0 Å, respectively. Electrostatic distance was calculated using the Hodgkin index and the Carbo index at the WebPIPSA server [46]. Rigid-body superposition was performed and electrostatic potential was computed using Chimera 1.8.1.

Hydropathy analysis

Hydropathy analysis was performed using the Kyte-Doolittle scale implemented in Protein Hydrophobicity Plots [34] and in ProtScale at the ExPASy server [47,48]. In order to highlight hydrophilic regions likely exposed on the surface, a seven amino acids window was chosen; regions with score >0 are hydrophobic [33]. Hydrophobic/hydrophilic patches were plotted onto structures through Chimera 1.8.1.

Additional files

Additional file 1: Two-pages figure relating HA stem secondary superstructures to immunogenic epitopes.

Additional file 2: Multi-page figure reporting epograms for each analyzed HA subregions (stem, RBD, HA1) and for HA monomers and trimers.

Additional file 3: Reports comparison amongst epograms for stem subregions obtained performing the WebPIPSA analyses with solved PDB structures or replacing either H9 or H3 templates by modeled structures.

Abbreviations

AC: Accession code; APBS: Adaptive PB Solver; ED: Electrostatic distance; Epogram: Electrostatic potential diagram; GRAVY: GRand AVerage hYdrophobicity; HA: Haemagglutinin; I: Ionic strength; N: Neuraminidase; PB: Poisson-Boltzmann; PDB: Protein data bank; PIPSA: Protein Interaction Property Similarity Analysis; RBD: Receptor-binding domain; RMSD: Root mean square deviation; SI: Similarity index; VED: Vestigial esterase domain; VS: Viral strain; WHO: World Health Organization.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FF and GC conceived the study. FF oversaw the study. IR performed most of bioinformatic analyses (modeling, electrostatics, hydropathy). IR and FF interpreted the data. AM performed part of the bioinformatic analyses on H5N1 clades and provided other authors with help in data interpretation. IR and FF wrote the paper with input from GC and AM. All authors read and approved the final manuscript.

Authors' information

IR is a PhD student and a bioinformatician; AM is a staff technician at the IZSVe, currently performing the PhD course, and a molecular virologist; GC is the Head of Research and Development Department, Division of Biomedical Science, OIE/FAO and National Reference Laboratory for Newcastle Disease and Avian Influenza, IZSVe; FF is Associate Professor of Molecular Biology and Bioinformatics and the PI of the MOLBINFO Unit at the Department of Biology, University of Padua.

Acknowledgements

We thank Stefan Richter for helpful information on WebPIPSA, Walter Rocchia and Sergio Decherchi for expert suggestions on electrostatic analyses, Stefano Vanin and Isabella Monne for useful discussions. This work was supported by basic funding ('ex 60%') from the Italian Ministry for University and Research (MIUR) to FF.

Author details

[1]Molecular Biology and Bioinformatics Unit (MOLBINFO), Department of Biology, University of Padua, via U. Bassi 58/B, 35131 Padova, Italy. [2]FAO-OIE and National Reference Laboratory for Newcastle Disease and Avian Influenza, Istituto Zooprofilattico delle Venezie (IZSVe), viale dell'Università 10, 35020 Legnaro, Italy.

Received: 14 July 2014 Accepted: 28 October 2014
Published online: 10 December 2014

References

1. Ekiert DC, Bhabha G, Elsliger MA, Friesen RH, Jongeneelen M, Throsby M, Goudsmit J, Wilson IA: **Antibody recognition of a highly conserved influenza virus epitope.** *Science* 2009, **324:**246–251.
2. Han T, Marasco WA: **Structural basis of influenza virus neutralization.** *Ann N Y Acad Sci* 2011, **1217:**178–190.
3. **World Health Organization** [http://www.who.int/research/en/]
4. **Center for Disease Control and prevention** [http://www.cdc.gov/datastatistics/]
5. Hamilton BS, Whittaker GR, Daniel S: **Influenza virus-mediated membrane fusion: determinants of hemagglutinin fusogenic activity and experimental approaches for assessing virus fusion.** *Viruses* 2012, **4:**1144–1168.
6. Sriwilaijaroen N, Suzuki Y: **Molecular basis of the structure and function of H1 hemagglutinin of influenza virus.** *Proc Jpn Acad Ser B Phys Biol Sci* 2012, **88:**226–249.
7. Velkov T, Ong C, Baker MA, Kim H, Li J, Nation RL, Huang JX, Cooper MA, Rockman S: **The antigenic architecture of the hemagglutinin of influenza H5N1 viruses.** *Mol Immunol* 2013, **56:**705–719.
8. Stanekova Z, Vareckova E: **Conserved epitopes of influenza A virus inducing protective immunity and their prospects for universal vaccine development.** *Virol J* 2010, **7:**351.
9. Russell RJ, Gamblin SJ, Haire LF, Stevens DJ, Xiao B, Ha Y, Skehel JJ: **H1 and H7 influenza haemagglutinin structures extend a structural classification of haemagglutinin subtypes.** *Virology* 2004, **325:**287–296.
10. Gamblin SJ, Skehel JJ: **Influenza haemagglutinin and neuraminidase membrane glycoproteins.** *J Biol Chem* 2010, **285:**28403–28409.
11. Gamblin SJ, Haire LF, Russell RJ, Stevens DJ, Xiao B, Ha Y, Vasisht N, Steinhauer DA, Daniels RS, Elliot A, Wiley DC, Skehel JJ: **The structure and receptor binding properties of the 1918 influenza haemagglutinin.** *Science* 2004, **303:**1838–1842.
12. Xu R, Wilson IA: **Structural characterization of an early fusion intermediate of influenza virus haemagglutinin.** *J Virol* 2011, **85:**5172–5182.
13. Sauter NK, Hanson JE, Glick GD, Brown JH, Crowther RL, Park SJ, Skehel JJ, Wiley DC: **Binding of influenza virus haemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and X-ray crystallography.** *Biochemistry* 1992, **31:**9609–9621.
14. DuBois RM, Zaraket H, Reddivari M, Heath RJ, White SW, Russell CJ: **Acid stability of the haemagglutinin protein regulates H5N1 influenza virus pathogenicity.** *PLoS Pathog* 2011, **7**(12):e1002398.
15. Ha Y, Stevens DJ, Skehel JJ, Wiley DC: **H5 avian and H9 swine influenza virus haemagglutinin structures: possible origin of influenza subtypes.** *EMBO J* 2002, **21:**865–875.
16. Lu X, Shi Y, Gao F, Xiao H, Wang M, Qi J, Gao GF: **Insights into avian influenza virus pathogenicity: the haemagglutinin precursor HA0 of subtype H16 has an alpha-helix structure in its cleavage site with inefficient HA1/HA2 cleavage.** *J Virol* 2012, **86:**12861–12870.
17. Carugo O, Pongor S: **A normalized root mean square distance for comparing protein three dimensional structures.** *Protein Sci* 2001, **10:**1470–1473.
18. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF Chimera–a visualization system for exploratory research and analysis.** *J Comput Chem* 2004, **25:**1605–1612.
19. Wang W, Anderson CM, De Feo CJ, Zhuang M, Yang H, Vassell R, Xie H, Ye Z, Scott D, Weiss CD: **Cross-neutralizing antibodies to pandemic 2009 H1N1 and recent seasonal H1N1 influenza A strains influenced by a mutation in haemagglutinin subunit 2.** *PLoS Pathog* 2011, **7**(6):e1002081.
20. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5:**823–826.
21. De Franceschi N, Wild K, Schlacht A, Dacks JB, Sinning I, Filippini F: **Longin and GAF domains: structural evolution and adaptation to the subcellular trafficking machinery.** *Traffic* 2014, **15:**104–121.
22. Jang SB, Kim YG, Cho YS, Suh PG, Kim KH, Oh BH: **Crystal structure of SEDL and its implications for a genetic disease spondyloepiphyseal displasia tarda.** *J Biol Chem* 2002, **277:**49863–49869.
23. Jeyabalan J, Nesbit MA, Galvanovskis J, Callaghan R, Rorsman P, Thakker RV: **SEDLIN forms omodimers: characterisation of SEDLIN mutations and their interactions with transcription factors MBP1, PITX1 and SF1.** *PLoS One* 2010, **5**(5):e10646.
24. Ekiert DC, Friesen RH, Bhabha G, Kwaks T, Jongeneelen M, Yu W, Ophorst C, Cox F, Korse HJ, Brandenburg B, Vogels R, Brakenhoff JP, Kompier R, Koldijk MH, Cornelissen LA, Poon LL, Peiris M, Koudstaal W, Wilson IA, Goudsmit J: **A highly conserved neutralizing epitope on group 2 influenza A viruses.** *Science* 2011, **333:**843–850.
25. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA: **PDB2PQR: an automated pipeline for the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculations.** *Nucleic Acids Res* 2004, **32**(Web server issue): W665–W667.
26. Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA: **PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations.** *Nucleic Acids Res* 2007, **35**(Web server issue):W522–W525.
27. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA: **Electrostatics of nanosystems: application to microtubules and the ribosome.** *Proc Natl Acad Sci U S A* 2001, **98:**10037–10041.
28. Richter S, Wenzel A, Stein M, Gabdoulline RR, Wade R: **WebPIPSA: a web server for the comparison of protein interaction properties.** *Nucleic Acid Res* 2008, **36**(Web Server Issue):W276–W280.
29. Hodgkin EE, Richards WG: **Molecular similarity based on electrostatic potential and electric field.** *Int J Quant Chem* 1987, **32**(Suppl 14):105–110.
30. Guo T, Gong LC, Sui SF: **An electrostatically preferred lateral orientation of SNARE complex suggests novel mechanisms for driving membrane fusion.** *PLoS One* 2010, **5**(1):e8900.
31. Lee KK, Fitch CA, Garcìa-Moreno EB: **Distance dependence and salt sensitivity of pairwise, coulombic interactions in a protein.** *Protein Sci* 2002, **11:**1004–1016.
32. López de Victoria A, Kieslich CA, Rizos AK, Krambovitis E, Morikis D: **Clustering of HIV-1 Subtypes Based on gp120 V3 Loop electrostatic properties.** *BMC Biophys* 2012, **5:**3.
33. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Mol Biol* 1982, **157:**105–132.
34. **Protein Hydrophobicity Plots** [http://arbl.cvmbs.colostate.edu/molkit/hydropathy/]
35. Bower M, Cohen FE, Dunbrack RL Jr: **Sidechain prediction from a backbone-dependent rotamer library: A new tool for homology modeling.** *J Mol Biol* 1997, **267:**1268–1282.
36. Canutescu AA, Shelenkov AA, Dunbrack RL Jr: **A graph theory algorithm for protein side-chain prediction.** *Protein Sci* 2003, **12:**2001–2014.
37. Xu D, Zhang Y: **Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization.** *Biophys J* 2011, **101:**2525–2534.
38. Gautier R, Camproux AC, Tufféry P: **SCit: web tools for protein side chain conformation analysis.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W508–W511.
39. Benkert P, Künzli M, Schwede T: **QMEAN Server for Protein Model Quality Estimation.** *Nucleic Acids Res* 2009, **37**(Web Server issue):W510–W514.
40. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T: **Protein structure homology modeling using SWISS-MODEL workspace.** *Nat Protoc* 2009, **4**(1):1–13.
41. **UCSF Chimera** [http://www.cgl.ucsf.edu/chimera/]
42. **APBS server** [http://www.poissonboltzmann.org]
43. Sitkoff D, Sharp K, Honig B: **Accurate calculation of hydration free energies using macroscopic solvent models.** *J Phys Chem* 1994, **98:**1978–1988.
44. Schutz CN, Warshel A: **What are the dielectric 'constants' of proteins and how to validate electrostatic models?** *Proteins* 2001, **44:**400–417.
45. Gorham RD Jr, Kieslich CA, Morikis D: **Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization.** *Ann Biomed Eng* 2011, **39:**1252–1263.
46. **WebPIPSA** [http://pipsa.eml.org/pipsa]
47. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A: **Protein identification and analysis tools on the ExPASy server.** In *The Proteomics Protocols Handbook.* Edited by Walker JM.: Humana Press; 2005:571–607.
48. **ExPASy server** [http://www.expasy.org]

doi:10.1186/s12859-014-0363-5
Cite this article as: Righetto *et al.*: Comparative structural analysis of haemagglutinin proteins from type A influenza viruses: conserved and variable features. *BMC Bioinformatics* 2014 **15**:363.

CHAPTER 5

Phylogenetic, phylogeographic and structural bioinformatic approach to the evolution and spreading of H9N2 avian influenza virus.

Milani A^, Heidari A^, Fusaro A, Righetto R, Cattoli G, Monne I, Filippini F.

Going to be submitted by February 2016

^A.M. and A.H. contributed equally to this work

ABSTRACT

Influenza A virus is a zoonotic agent with a significant impact both on public health and poultry industry and avian influenza H9N2 virus provided the first record of switch to human host. Therefore, surveillance and characterization are needed by the scientific community and public health systems and so far this was mainly based on intensive serological characterization and phylogenetic analyses aimed to infer evolutionary trends. In order to aid molecular epidemiologic assessment and support public health interventions, as well as to properly relate investigations worldwide thanks to shared nomenclature and robust guidelines, we developed a method for the clade nomenclature of all AI H9N2 hemagglutinin subtypes, based on the evolutionary dynamics of a large and non redudant viral strain dataset. This was combined to a phylogeographic analysis providing further information on the spatiotemporal evolution, correlation and spreading of H9N2 viruses from the beginning to current trends.

We found that H9N2 viruses can be clustered in five classes based on congruence of phylogenetic and phylogeographic data with structural comparison evidence. Structural analyses can properly depict structural closeness among proteins or protein domains and provide functional insights on surface regions possibly crucial to antigenicity and cell binding. Recent successful inference of surface feature fingerprints for H5N1 evolution prompted us to assess whether such fingerprints are peculiar to H5N1, or electrostatic variation could be associated to the evolution and spreading of other avian influenza viruses, e.g. the newly defined classes and clades from the H9N2 subtype.

Finally, surface feature fingerprints could be inferred that relate class and clade specific variation in electrostatic charges and isocontour to well-known hemagglutinin sites involved in modulation of immune escape and host specificity. Results from this work suggest the integration of up-to-date phylogenetic and phylogeographic analyses with sequence-based and structural investigation of surface features as a front-end strategy for inferring trends and relevant mechanisms in influenza virus evolution.

BACKGROUND/INTRODUCTION

Influenza A virus is a zoonotic agent with a significant impact both on public health and poultry industry. Wild water avian species represent the largest reservoir for influenza A virus subtypes that - in addition to birds - can infect mammalian hosts, such as e.g. humans and swine. This is suggestive for setting up a coordinated global surveillance network (Butler, 2012) as well as for studying viral evolution. Indeed, improving the capacity to monitoring viral genetic changes to predict 'evolutionary trends' can be crucial to boost surveillance, especially when considering those viral clades for which is reported - or likely to occur - avian to mammals/humans host switch (Al-Tawfiq et al., 2014). In addition to viral strains with well known potential to jump the host-species barrier (Nelson and Vincent, 2015), further risk for human and animal health depends on the emergence of novel reassortant viruses, especially in those regions where multiple strains and clades are known to co-circulate (Su et al., 2015). Avian influenza (AI) viruses from the H5N1 subtype are unique in their ecological success, showing extremely broad host range and geographical spreading (Guan and Smith, 2013). Therefore, based on intensive experimental characterization and phylogenetic analyses just for H5N1 clades and subclades a standard nomenclature was published (WHO/OIE/FAO H5N1 Evolution Working Group, 2008; Guan and Smith, 2013) and it is actually adopted by the scientific community. General concern for pandemic risk decreased after the peak of the H5N1 virus, but indeed novel reassorted subtypes (e.g.,

H7N9, H9N2, H10N8) jumped in recent years the host-species barrier and thus surveillance and characterization are needed by the scientific community and public health systems (Trombetta et al., 2015). In particular, the isolation of H9N2 virus from two Hong Kong children provided the first record of switch to human host (Peiris et al., 1999); then, these viruses could occasionally be transmitted from poultry to humans and other mammals (Lin et al., 2000; Butt et al., 2005; Sang et al., 2015). Recently, infection with the novel H10N8 virus in humans raised concerns about its pandemic potential worldwide (Hu et al., 2015).

Even though host jump and pandemic influenza phenomena raised much attention to AI viruses because of their potential impact on human health, studying virus variation related to either low to high-pathogenicity shift or to antigenic drift is needed as well, as it is quite relevant to animal health and of special impact on poultry industry and vaccine efficacy. In fact, outbreaks of high pathogenic AI (HPAI) can result in killing hundreds millions poultry and wild birds in tens of countries (Swayne, 2012). This makes proper vaccination strategies (at least in poultry animals) crucial to prevent wide mortality and in turn viral variation potentially resulting in antigenic drift becomes a factor to be monitored in that associated to risk of impairing vaccine efficacy.

Evolution and spread of low pathogenic AI (LPAI) and HPAI viruses, belonging to H5, H7 and H9 subtypes, amongst birds and their sporadic infection in humans continues to represent a great concern for public health (Lin et al., 2000). Therefore, these avian subtypes are included as top pandemic agents in the list from the World Health Organization (WHO). To date, only H5 and H7 subtypes of influenza A viruses were reported to evolve from a lpai to hpai form after their introduction into poultry from the wild bird reservoir (Alexander 2007). Both H5 and H7 viruses are notifiable to the Office International des Epizooties (OIE) because of the risk of LPAI becoming HPAI by mutation. H7 LPAI virus usually causes mild respiratory disease and a production decrease in infected poultry; its evolution into a HPAI form results in the generation of a virus able to cause severe disease and death in the poultry population (http://www.oie.int/fileadmin/Home/eng/Health_standards/tahm/2.03.04_AI.pdf).

One example assessing the ability of LPAI to evolve into HPAI form is the H7 avian outbreaks that affected Northern Italy between 1999 and 2001. Epidemiological information sustained by phylogenetic analysis, and deep sequencing approaches helped to reveal that HPAI strains evolved from the LPAI viruses and that both lineages shared a common ancestor (Monne et al., 2014).

Current AI vaccines are based upon the elicitation of a neutralizing antibody (Ab) response against the major epitope regions of the viral surface glycoprotein, hemagglutinin (HA). However, mutations in immune-dominant regions on the HA structure may result in antigenic drift allowing the virus to escape Ab neutralization (Velkov et al., 2013). Indeed, antigenic and genetic differences in HA and the other surface spike protein neuraminidase (NA) provide a rationale for classification of influenza type A virus subtypes: for instance, H9N2 viruses combine the H9 subtype HA with N2 subtype NA. Haemagglutinin plays a central role in influenza A virus evolution because it is crucial to the attachment and penetration into the host cell and - as the main viral surface antigen - it is also a major player in the stimulation of the neutralizing Ab response (Velkov et al., 2013).

In this work, genetic diversity of H9N2 subtype was assessed through large-scale phylogenetic analysis; this resulted in a novel and updated classification scheme based on the phylogenetic topology and evolutionary distances following the same WHO standards for virus classification as for H5N1. Classic, sequence based phylogenetic analyses can be integrated by structural bioinformatic investigations that more properly depict

structural closeness among proteins or protein domains. Indeed, we previously inferred molecular fingerprints for H5N1 evolution, as intriguing surface charge redistribution at the surface of the receptor binding domain (RBD) subregion of HA was found to relate to branching of still circulating clades 2 and 7 with respect to those ones that are no longer circulating (Righetto et al., 2014). This prompted us to assess whether such fingerprints are peculiar to H5N1, or electrostatic variation could be associated to the evolution and spreading of other AI viruses, e.g. the newly defined clades and subclades from the H9N2 subtype.

RESULTS

Phylogenetic analysis and novel classification scheme for AI H9N2 HA

Since the established classification system for AI virus subtypes is based on antigenic and genetic differences in the two surface spike proteins, we assessed the genetic diversity of the HA gene to develop a unified nomenclature and classification system for AI H9 subtype genetic groups. However, given that in the different classification there is a stronger correlation between the phylogenetic topology and the evolutionary distances within and between genetic groups, we used the genetic correlation as the base to develop objective criteria to classify strains and create a definitive unified nomenclature. The classification criteria were determined based on the phylogenetic topology and on specific evolutionary distances that reflect the diversity of the AI H9N2 subtype. In order to get robust validation for phylogenetic analyses, the evolutionary history of 1669 AI H9N2 strains was inferred for HA nucleotide sequences ≥ 1500 bp using three different algorithms: neighbor-joining (NJ), maximum likelihood (ML) and Bayesian; when the evolutionary history was inferred from a smaller dataset alignment (360 strains), this confirmed consistency of the proposed classification.

Figure 1 shows the nucleotide phylogenetic tree with the proposed grouping, while the full tree is depicted in Supplementary figure S1: the AI H9N2 strains clearly separate into five different monophyletic groups hereafter referred to as class A, B, C, D, and E. Within such classes, twenty-seven clades - identified by numbers - are separated based on inter-clade average distance ≥5% and intra-clade average distance <5% (Table 1) and separation for each identified clade is confirmed by C-value ≥ 1 (Table 2). Classes and clades were assigned when at least three isolates with different epidemiological history formed a distinct taxonomic group with bootstrap value at the defining node ≥60%. Clades separation based on distance value cut off was confirmed using two different calculation algorithms (see methods).

Fixed guidelines for classification are resumed in Table 3; circulation data and representative H9 subtype viruses used for different analyses performed in this work are listed in Supplementary table S1 to facilitate the interpretation of relationship to the proposed numbering system.

When comparing intra-class nucleotide distance values, class A shows the highest genetic heterogeneity with a value up to 18.3%, whereas this value is around 10-11% in classes B and C. Class D just contains three strains isolated in Malaysia and class E only a few strains isolated mostly in the USA. Therefore, separate clades were not defined for these two classes. Few different strains that are no longer circulating and do not group within any identified class are considered as ancestral. Such topology distribution and grouping was fully confirmed when performing phylogenetic analysis with corresponding protein sequences (not shown).

Phylogeographic analysis for AI H9N2 HA

In order to determine the worldwide dissemination of AI H9N2, the sequences of the HA gene were grouped into eigth geographic areas, namely: (i) North America, (ii) Europe, (iii) Oceania, (iv) China, (v) Meaddle East, (vi) South, (vii) South east and (viii) East Asia. The final dataset included 357 viruses that could be used for in-depth special analysis. Through Posterior distribution under the Bayesian framework, we reconstructed genealogical trees with time-scale and inferred ancestral locations of each branch using sequences' sampling collection dates and locations. The time-scaled phylogeographic maximum clade credibility (MCC) trees of HAs implemented in BEAST (see methods) and the root state posterior probability are illustrated in Figure 2 (phylogeography tree) and Figure 3 (phylogeography map), in which the most probable location of each branch is assigned different colours and the calibrating time-scale. Numbers at branch points in Figure 2 are reported where state probabilities with values ≥ 0.55 correspond to the most relevant events (i.e. to area-area transitions rather than to intra-areal ones). Such transition events are graphically depicted as arrows (with class-coded colors) in Figure 3 map. For graphical reasons (saving space to fit the one page format), names of the 357 individual viral strains are not reported in Figure 2; however, the same tree with all virus names is presented in Supplementary figure S2.

Our phylogeographic results suggest that the American strains are ancestral for all H9 subtypes. Those ones introduced in China then were spreading worldwide. In particular, American Class A strains reached China first and Chinese clades moved in turn to Europe, southeast/east Asia and Australia by migratory birds. Class B (mostly present in poultry), after introduction and circulation in Middle East moved to south Asia. Class C can also be referred to as China class because it evolved and expanded mostly in China; however, different viruses of class C were introduced and circulated in east and south Asia. Class D formed a separate class in southeast Asia, while class E evolved by back migration events of Chinese viruses to North America. The overall spatiotemporal representation of phylogeographic evolution and worldwide spreading of the H9 subtypes is presented in Supplementary visual animation S1.

Clustering by electrostatic features for AI H9N2: heat maps and epograms

A recent bioinformatic work has shown that the integration of sequence and structural analyses for HA (and especially its RBD) can shed more light on the evolution and spreading of AI H5N1 viruses by unveiling surface patches as possible evolutionary fingerprints (Righetto et al., 2014). Therefore, when considering findings emerged from comparative HA1 and RBD analysis, we decided to check whether variation in electrostatic features of H9N2 would relate to phylogenetic data, as observed for H5N1 (Righetto et al., 2014).

Representative strains for each clade of the five H9N2 classes identified in our phylogenetic analysis are summarized in Supplementary table S1. In order to quantitatively evaluate the electrostatic distance, clustering of the spatial distributions of the electrostatic potentials was obtained by WebPIPSA (Protein Interaction Property Similarity Analysis) (Richter et al., 2008). Figure 4 depicts the heat map and density plot for the RBD subregion of HAs from such representative strains. High electrostatic distance (dark blue, violet or magenta colors, see density plots) clearly separates classes A, D and E (typical of wild birds) from B and C (common to poultry birds), whereas the electrostatic distance between B and C is lower, as highlighted by prevalence of the light blue color. Therefore, clustering of H9N2 classes by electrostatic features shows substantial agreement to phylogenetic grouping, apart from a few exceptions: for instance, in terms of

electrostatic distance, B3 and B4 are closer to C2 (light blue) than to B2 strains. It can be noticed that B3 and B4 clades used for this analysis were both isolated from the same host bird (quail). In addition to heat maps and corresponding density plots, the distance matrices of the electrostatic potentials were also displayed as trees referred to as 'epograms' (electrostatic potential diagrams). The epogram for RBD confirms grouping of the A wild bird cluster as well as homogeneity of class C and B2 clades; moreover, once again B3 sorts with C clades rather than with B ones (Figure 5). Both heat map and epogram for HA1 subregion (not shown) confirmed clustering from the RBD analysis.

Variation in charge distribution among AI H9N2 classes and strains

In depth analysis of the distribution of charged residues, of their variation in number and position and of isocontours from the different HA subregions, further confirmed that variation especially concerns the RBD subregion and suggested possible electrostatic fingerprints are associated to different H9N2 classes.

Class-associated 'charge redistribution' is found to occur at RBD positions 135, 146 and 162: the net charge for these three positions is zero in all classes (as the sum of two opposite charges and a non charged residue), but the charge distribution pattern shared by the 'wild bird' classes A, D and E is different from the pattern conserved in viruses from the 'poultry' classes B and C. In fact, distribution at 135-146-162 is neutral-positive-negative in A-D-E, and negative-neutral-positive in B-C (Table 4). In particular, at position 135, almost all viruses from class A share a non charged residue (167/177 strains) with prevalence (127/177) of Asn, which is 100% conserved (16/16) in classes D-E; mutation to charged residue (N135D) only concerns 10/177 viruses from clades A5.3, A5.4 and A5.5. Instead, negativization at position 135 is most often observed in both classes B (311/364) and C (1011/1102), with prevalence of Asp/Glu over other amino acids in almost all B-C clades. A compensatory mechanism is observed for exceptions, i.e. for those clades that do not share a negative charge at position 135. For example, clade C1 lacks the negative charge of classes B-C and shares instead (31/31 sampled viruses) N135 with classes A-D-E; however, this is compensated as C1 is also the only B-C clade missing a positive charge at position 131. Similarly, B3 (showing prevalence of Gl35) is also the only B-C clade with a negative charge (Glu) instead of a non charged residue at position 180. Therefore, compensatory mutations in the RBD seem to keep class-specific fingerprints and net charge, while progressively 'sliding' positions of charged residues over RBD sites in the viral population. Residue 146 is His in 169/177 viruses from class A and in all D-E strains, while prevalence of Gln is observed in all clades from class B (363/364) and C (1075/1102). The only exception in class A is clade A.5.2, showing Q146 (like B-C) instead of H146 (common to A-D-E). However - like for example above - a counterbalancing unique mutation is observed: depositivization at position 146 of A.5.2 is compensated by peculiar denegativization at position 162 (E162N). In most (114/177) class A viruses and in all D-E strains, residue 162 is Glu except for clade A5.5, showing mutation E162W in 51/51 viruses. Intriguingly, the lost negative charge is rescued at the contiguous N-terminal position 161 by the equally conserved (51/51) and peculiar mutation N161D, suggesting the negative charge at position 162 (or 161) as a landmark for A-D-E viruses. Instead, in viral strains from classes B-C the major residues are Arg and Gln, with prevalence of the former over the latter in all clades but B2.4 and C2.2, where reverse prevalence is observed. Therefore, ongoing positivization of position 162 seems to be a landmark as well for viruses circulating in poultry. Altogether, counterbalancing mutation pairs observed at positions 131-135, 135-180, 146-162 and 161-162 seem to support the compensatory mechanism suggested above for mantaining the overall net charge of the RBD while sliding

charges over the RBD itself, i.e. for keeping class specific landmarks along with contemporary creation of novel fingerprints.

A second and different kind of variation in electrostatic features is observed at positions 180 and 186 of the RBD (Table 4). In all H9N2 classes, the net charge for this amino acid pair is zero. However, in A-D-E viruses this results from the sum of opposite charges (+1 -1 = 0), while in B-C viruses, it depends on replacement of both charged residues by neutral ones (0 + 0 = 0). Therefore in this case - differently from previous examples of charge redistribution - manteinance of the net charge is associated to a decrease in the percentage of charged residues in the RBD.

Class and sub-class specific variation in electrostatic and hydrophobicity features

Table 4 also reports intriguing variation at the contiguous positions 216 and 217. Clades from classes A-D-E share at position 216 a highly conserved (176/177 strains) polar residue (Gln), while polar to hydrophobic transition is clearly apparent in classes B and C. In particular, the 'original' Gln is replaced in the most of strains (1256/1466) by the hydrophobic residue Leu, showing prevalence in all clades (from classes B-C) but 'B' (the ancestral one without numbering) and C2. However, Gln is still present in 202/1466 B-C strains and the most represented residue in clades 'B' and C2. In the next position (217), sub-class variation is observed: only B.2.x viruses share a hydrophobic residue (Ile), while Gln is common to all other clades in classes A-B-C-D-E. When inspecting in more detail distribution of residues among individual clades and strains, the picture is meaningfully different from position 216. In addition to classes A-D-E (176/177 strains), the 'original' Gln is highly conserved also in class C (900/1102 strains) and in C1 the major residue is anyway a polar one (Thr, in 30/31 strains). Instead, a complex picture is displayed in class B: Gln is still 100% conserved in clades B, B1.1, B1.2 and B3, whereas polar to hydrophobic transition is ongoing in clade B4 (4 hydrophobic strains out of 7) and fully fixed (100% of strains) in the whole B.2.x subgroup. Such specific variation in sub-class B.2.x is only apparently restricted to hydrophobic patches, as 'charge sliding' is observed between positions 165 and 198. In particular, all H9 clades but B2.x share a 100% negatively charged residue at position 198, which is replaced in B2.x by a polar amino acid (mostly, Asn). Such a peculiar (with respect to other H9 viruses) denegativization is however compensated in B.2.x by an equally peculiar acquisition of a negative charge at position 165, where Asp is 100% conserved.

Residues involved in changes at the H9N2 RBD are surface exposed

The RBD from the solved structure of the H9 HA was viewed to highlight the nine amino acid positions involved in class or sub-class specific variation: as shown in Figure 6, all such positions are exposed at the RBD surface. The RBD subregions (130-loop, 190-helix and 220-loop) mediating SA binding are highlighted in yellow. The three residues 135, 146 and 162 involved in class specific 'charge redistribution' are highlighted in orange; in particular, 146 is close to 190-helix, 135 is part of 130-loop and 162 is surface exposed as well. Position pair 180-186 (mediating 'charge loss' in the A-D-E to B-C transition) is highlighted in purple and is part of 190-helix. Concerning the four positions involved in class and sub-class variation (highlighted in green), positions 216 and 217 are part of 220-loop, while 165 and 198 protrude at the other 'side'. Finally, positions 131 and 161 involved in compensatory variation (see Table 4) were also confirmed to be surface exposed (not shown).

Given that surface location is confirmed for all aforementioned positions, variation in H9N2 among different classes, subclasses and clades can be viewed and analyzed in more depth by comparison of the electrostatic isocontours, which were determined as previously reported (Righetto et al., 2014; see also the methods section). Prior to starting electrostatic analysis, whe checked the quality of models because shaping of the isopotential contour can be influenced by the orientation of side chains. However, model refinement proved to be unnecessary because of the very high average sequence identity (around 90%) of the H9N2 target sequences to the H9 template, as previously observed with H5N1 models (Righetto et al., 2014).

Figure 7 shows the isopotential contours from two viral strain that are well representative for electrostatic fingerprints from 'wild bird' viruses (classes A-D-E) and 'poultry' viruses (classes B-C). In fact, both strain A.1_AtkCA66 and C.2.3_AckHe07 match in all positions patterns in Table 4 typical for classes A-D-E or classes B-C, respectively. At position 162, the two strains clearly show the opposite charges; concerning position 135, the expected contours are found again, as A.1_AtkCA66 shows no charge while in C.2.3_AckHe07 a negative protrusion is found in the corresponding area. However, not always expectations are respected and comparison of the 180-186 amino acid pair clearly shows that the loss of both charged residue in the A-D-E to B-C transition does not just result in 'neutralization' of the corresponding surface area. In fact, in spite of the expected red(Glu)-to-neutral(Val) shift, position 180 shows in C.2.3_AckHe07 a seeming positivization (increased blue area), possibly because of the enlargement of electrostatic clouds from neighboring occurs.

For completeness of information, the isopotential contours of the RBDs from all representative strains used for creating the heat map and epogram in figures 4 and 5 are presented in Supplementary Figure S3.

DISCUSSION AND CONCLUSIONS

Haemagglutinin has the role of main viral surface antigen in the stimulation of the neutralizing antibody response (Velkov et al., 2013) and for many years, classification of viral HA has been based upon serological and phylogenetic analyses (Stanekova and Vareckova, 2010). However, comparative structural analysis of HA can provide functional insights on surface regions possibly crucial to antigenicity and binding to the host cells. In fact, recent work on H5N1 demonstrated that electrostatic and hydrophobicity variation at the RBD surface relates to both evolution and spreading of viral clades and is able to provide fingerprints and infer trends to complement phylogeny and functional analyses (Righetto et al., 2014). Intriguingly, when comparing RBD regions, H5 is structurally quite closer to H9 than H3 and H7, and when RBD electrostatic potential is considered, H5 is even closer to H9 (member of a different phylogenetic HA group) than to H2 (member with H5 of the same phylogenetic group) (Righetto et al., 2014). Therefore, the possibility that similar mechanisms might underlie H5 and H9 evolution and spreading further prompted us to investigate on H9 evolution and on surface features of the H9 HA.

We developed a method for the HA clade nomenclature of all AI H9N2 subtypes, based on the evolutionary dynamics of a large and non redudant viral strain dataset. Clade assignments were made by following several criteria (Table 3), collectively used to rationally name groups by clade numbering. Based on phylogenetic topology, five different genetic classes could be distinguished, which so far consist of twenty-five clades according to the different molecular analyses and fixed criteria. Once further clades arise along evolution of H9N2 viruses, such a nomenclature is ready to be expanded (by enlarged numbering) based on

already fixed criteria. Circulation and evolution of the H9N2 HA gene show a remarkable similarity to the H5 subtype and notable difference from the typical evolution of H3. The evolution of the human influenza viruses since 1968 is characterized by a limited diversity among circulating strains. This lack of diversity is likely the consequence of rapid extinction after the emergence of new clades and lineages. As expected, the evolutionary tree of human influenza HA genes has extended trunks and extremely short branches; conversely, AI H9N2 strains show extended branches as these viruses continue to co-circulate in different regions and host species and this allows the clades for further evolving and differentiating. Therefore, a standard nomenclature system for H9N2 classification is needed now in order (i) to provide a rationale for the AI H9N2 evolution, and (ii) to properly relate investigations worldwide thanks to robust guidelines. Moreover, while AI virus was a useful organism to study due to its rapid mutation rate and the wealth of surveillance data available, we are not limited to influenza. We believe that the rapid and accurate annotation of clades will aid molecular epidemiologic assessment and support public health interventions. Last but not least, shared nomenclature criteria can boost correlation analyses and favour proper naming of newly identified strains along epidemiological analyses.

As a fundamental component of modern biogeography (Riddle, 2009) and an approach of great impact on the most basic of biological questions, phylogeography is actually a hot research field boosting studies aimed at clarifying evolutionary dynamics in most life sciences disciplines (Turchetto-Zolet et al., 2013; Brown et al., 2014; Ni et al., 2014; Gräf et al., 2015; Pyron, 2015; Zhang et al., 2015; Maixner et al., 2016; Stacy et al., 2016) including analysis of influenza viruses (Lu et al., 2014; Bedford et al., 2015; Hill et al., 2015; Pollett et al., 2015; Tian et al., 2015). In addition to the classification scheme and as a complement to it, phylogeographic data can provide further information on the spatiotemporal evolution, correlation and spreading of AI H9N2 viruses from the beginning to current trends.

Relationship among variation in electrostatic features, viral evolution and clades spreading observed for H9N2 in this work confirms and further strengthen previous observations on H5N1 AI viruses, in which the evolution of circulating clades is accompanied by 'charge redistribution' at the RBD (Righetto et al., 2014). Moreover, most of changes occuring at the RBD in H9N2 viruses seems to concern sites known to play a special role in RBD interactions, immune escape and host specificity. In particular, three RBD subregions mediate binding to the sialic acid (SA) moieties from the host cell and they are major antigenic determinants hence involved in immune escape/antigenic drift: 130-loop (H3 residues 135-138), ı9̃-helix (H3 residues 190-198) and 220-loop (H3 residues 221-228) (Wilson et al., 1981; Kobayashi et al., 2012). For Readers' convenience, in addition to H9 and H3 numbering systems, Table 4 also shows HA mature chain numberings for subtypes H1, H5, and H7, all five numberings being based on the most recently published table of correspondence (Burke and Smith, 2014).

As described in the results section, changes at positions 162 and 217 of H9N2 HA result in either class or sub-class specific variation of charge or hydrophobicity features of the RBD. In particular, position 162 is involved - together with positions 135 and 146 - in class specific 'charge redistribution' and, when denegativization occurs at position 162 in two clades from class A, this is compensated by either depositivization of residue 146 (in A5.2) or negativization of residue 161 (in A5.5). Sub-class specific, polar to hydrophobic transition occurs instead at residue 217. The involvement in immune escape of mutations at both positions 162 and 217 in H9N2 has been recently reported (Peacock et al., 2016). Indeed, as part of the 220-loop, position 217 is also likely involved in increased virus binding to α2-6 SA and thus in improved

affinity to the human host; for instancet, residue 224 in H1N1 (217 in H9) mediates hydrogen bond interactions with α2,6-SA (Chutinimitkul et al., 2010). Evidence on the involvement of this RBD position in the modulation of host range is also based on previous studies on H5N1 (Gambaryan et al., 2005). Position 227 (H3 numbering) is located between amino acids 226 and 228, both being part of the 220-loop and playing a key role in receptor specificity and host range restriction of influenza A viruses (Vines et al., 1998). In this work, class-specific variation in H9N2 position 216 (226 in H3) is observed (Table 4). Positions 180-186 in H9N2 (where the conserved dual opposite charges pair in classes A-D-E shifts to a non-charged pair in classes B-C, see previous sections) both belong to the 190-helix of the RBD involved in binding to SA moieties from the host cell (Wilson et al., 1981; Kobayashi et al., 2012). Contemporary loss of the two opposite charges is somehow 'compensatory' in that saving the original net charge of the RBD. It is noteworthy that in influenza A viruses, amino acid substitutions increasing (charge+) and decreasing (charge-) the charge of the SA binding RBD region can modulate binding avidity and affinity, and thus contemporary charge+ and charge- compensatory substitutions are often observed and likely to compensate gain and loss effects and to ensure the HA-NA charge balance is kept (Kobayashi et al., 2012). However, in A-D-E to B-C class transition, compensation only keeps the net charge, while the overall loss of two charged residues from the RBD in B-C class viruses occurs. This in turn is likely to favour immune escape, because of both the location of the two residues and the well-known role of charged amino acids in modulating protein antigenicity and immunogenicity. In fact, it is well known and an established evidence that both positively and negatively charged residues improve the antigenic recognition (up to several folds, depending on their number in the antigenic site) by creating further salt bridges with the recognizing antibody complementary surface (Young CR, 1984; Farber et al., 2007).

As shown in Table 4, charge variation also occurs at position 146 (involved with 135 and 162 in 'charge redistribution' on the RBD, see results), which is exposed at the RBD surface close to 190-helix (Figure 6). Based on the aforementioned role of 190-helix in binding SA moieties from host cells, changes like the observed depositivization at position 146 are likely to influence binding affinity and specificity. Considering that chickens possess both α-2′3′ and α -2′6′ SA receptors (Gambaryan et al., 2002), it is tempting to speculate that such a changes could be linked to host adaptation and species specificity (Perez et al., 2003).

In addition to represent a valuable complement to integrate phylogenetic and serological studies, structural analyses are also of great help to improve sequence-based, functional comparison. Sequence comparison was able to infer class and sub-class specific fingerprints presented in Table 4 as sequence patterns; however, only once sequence analysis was complemented by the structural approach, a real-estate picture of the system emerged. Comparison of the electrostatic isocontours showed that identified mutations cannot be considered as just isolated 'point changes'. In fact, surface features - that are pivotal players in regulating molecular interactions e.g. immune escape and host specificity - are modulated by the direct change of any mutated residue, as well as by the effects that such a mutation may exert on the local equilibrium in the surrounding area (salt bridges or repulsions, hydrophobicity changes, decreased or increased charge density etc.), as shown by the unexpected variations observed in charge clouds.

In conclusion, although much further work is needed to clarify in details the complex network of equilibria that can be altered by specific mutations, evidence from this study supports the integration of up-to-date phylogenetic and phylogeographic analyses with sequence-based and structural investigation of surface features as a front-end strategy for inferring trends and relevant mechanisms in influenza virus evolution.

METHODS

Phylogenetic analyses

HA gene nucleotide sequences of H9N2 subtype were retrieved from the Global Initiative on Sharing Avian Influenza Data (GISAID) EpiFlu database (http://www.gisaid.org). Nucleotide sequences of at least 1500 bp length were selected. Multiple sequence alignment of HA sequences was performed with MAFFT version 7 (http://mafft.cbrc.jp/alignment/server). Redudant isolates with 100% sequence similarity (i.e., redundant sequences) were identified and removed, giving a final HA dataset and alignment of 1669 sequences that was subjected to phylogenetic trees reconstruction. The NJ, ML and Bayesian methods were used to construct three different phylogenetic trees for comparison. Analysis of the best-fit substitution model was performed using MEGA5 (Tamura et al., 2011), and the goodness-of-fit of each model was measured by Bayesian Information Criterion and corrected Akaine Information Criterion (AICc). The General Time Reversible (GTR) model with a discrete gamma distribution (+Γ) allowing for invariant sites (+I) was selected based on AICc and used in all data analyses. MEGA5 was also used to perform phylogenetic analysis and the evolutionary history was inferred by both NJ and ML methods (Tamura and Kumar, 2002), with standard errors being calculated based on 1000 bootstrap replicates.

Furthermore, PhyML (version 2.4.4) (Guindon et al., 2003) was used to create ML trees. The GTR + Γ + I model of nucleotide substitution was used for the analysis, with an estimated gamma shape parameter. Robustness of the groups was assessed using the bootstrap approach with 100 replicatest. Bayesian phylogenetic tree was inferred using MrBayes software (Ronquist and Huelsenbeck, 2003) and applied to generate the dendrograms as well as to assess statistical supports for the branches from the trees generated by the original dataset. For ease of display, and also to ensure that the clade topology would be maintained when fewer isolates are used, a small representative dataset of 360 H9N2 HA sequences was created and analyzed by the same aforementioned phylogenetic models. Phylogenetic trees were visualized using FigTree version 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

The largest HA gene dataset alignment (n = 1669; length ≥ 1500 bp) used for the phylogenetic reconstruction, was also used to infer evolutionary distances (within and between groups) by pair-wise analysis. The number of base substitutions per site was calculated by two different methods. The simplest one (uncorrected pairwise distance) was performed by averaging all sequence pairs between groups, while the second method followed the Maximum Composite Likelihood model. Variation rate among sites was modelled with a Γ distribution value = 9.4 (calculated by preliminary estimation from our dataset) and the differences in the composition bias among sequences were considered in the evolutionary comparisons. The C-value ratio used in the H9N2 clades partitioning - i.e. the ratio of the average pairwise distance between a particular taxon and its closest neighboring group divided by the average pairwise distance within that selected clade - was used to confirm the clades partitioning.

Bayesian phylogeography reconstruction

Time-scaled phylogenies of H9N2 HA were inferred by Bayesian Markov chain Monte Carlo (MCMC) method implemented in BEAST v1.8.0 (Drummond et al., 2005) using the SRD06 codon position model and the uncorrelated log-normal relaxed clock model under a Bayesian skyline coalescent tree prior to the MCMC simulations (Jin et al., 2014). Bayesian skyline plot with a Piecewise constant model was used to elucidate

the population dynamics of H9N2 viruses. Spatial location reconstruction and viral migration were estimated using the discrete Bayesian phylogeographic method that utilised a continuous time Markov Chain over discrete sampling locations, and applied a Bayesian stochastic search variable selection model (Lemey et al., 2009).

For our data set we performed four independent runs for 300 million generations with sampling every 30000 steps. Convergence and effective sampling size of estimates were assessed by visual inspection using Tracer v1.6 (http://beast.bio.ed.ac.uk/Tracer). Multiple chains were then combined after a 10% burn-in using LogCombiner v1.8.0 included in the BEAST package. The maximum clade credibility (MCC) trees with temporal and spatial annotation were summarized with a 10% burn-in removed using TreeAnnotator v1.8.0 in the BEAST package and presentation figures were generated using FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

We also conducted Bayes factor (BF) tests to provide statistical support for transmission routes between different geographic locations using SPREAD v1.0.6 with cutoff BF = 3 (Bielejec et al., 2011). BF values represent the difference between the posterior and prior probabilities that the rates between two locations are non-zero. Thus, routes with high BF have large odds that a migration exists between two locations.

To animate viral dispersal over the time, we converted annotated MCC trees into a keyhole markup language file using SPREAD v1.0.6, which can be visualized by Google Earth (http://earth.google.com) website platform.

Structural Modeling and surface analysis

Structural models for HA1 and RBD regions of target HA proteins were obtained by homology modelling as reported (Righetto et al., 2014) on best available structure templates using SWISS-MODEL (Bordoli et al., 2009). In particular, as a template for H9N2 HAs the currently available solved HA structure PDB 1JSH belonging to A/swine/Hong Kong/9/98 (H9N2) was used. Refinement of model structures was performed using three independent methods as reported (Righetto et al., 2014) and model quality was checked via QMEAN server (Benkert et al., 2009).

Protein structures were viewed using UCSF Chimera (Pettersen et al., 2004) v. 1.10.2 (free download from http://www.cgl.ucsf.edu/chimera/).

Comparative analysis of electrostatic potentials was performed as reported (Righetto et al., 2014), simulating physiological conditions, i.e. the spatial distribution of the electrostatic potential was calculated at ionic strength (I) = 150 mM, assuming +1/-1 charges for the counter-ions. Isopotential contours were calculated using UCSF Chimera, which allows for connecting - through Opal web server - to the Adaptive Poisson-Boltzmann Solver server (http://www.poissonboltzmann.org/apbs). PDB2PQR was used to assign partial charges and van der Waals radii according to the PARSE force field (Sitkoff et al., 1994). Electrostatic distance was calculated using the Carbo index at the WebPIPSA server (http://pipsa.eml.org/pipsa). Rigid-body superposition was performed and electrostatic potential was computed using UCSF Chimera 1.10.2.

For more details please see the methods section from Righetto et al., 2014.

LIST OF ABBREVIATIONS

Ab, antibody

AC, accession code

AI, Avian influenza

AICc, corrected Akaine Information Criterion

BF, Bayes factor

ED, electrostatic distance

Epogram, electrostatic potential diagram

GTR, General Time Reversible

GISAID, Global Initiative on Sharing Avian Influenza Data

HA, Haemagglutinin

High pathogenic Avian influenza (HPAI)

I, ionic strength

Low pathogenic Avian influenza (LPAI)

MCC, maximum clade credibility

MCMC, Markov Chain Monte Carlo

ML, maximum-likelihood

NA, Neuraminidase

NJ, neighbor-joining

OIE, Office International des Epizooties

PDB, Protein Data Bank

PIPSA, Protein Interaction Property Similarity Analysis

RBD, receptor-binding domain

SA; Sialic acid

WHO, World Health Organization

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

AM, AH, FF and GC conceived the study. FF oversaw the study. AM and AH performed most of analyses; AF and IR provided other authors with help in data interpretation. AM, AH and FF wrote the paper with inputs from AF, IM and GC. All authors read and approved the final manuscript.

AUTHORS INFORMATION

AM is a staff graduate technician at the IZSVe and a molecular biotechnologist, she is currently ending last year of her PhD course; AH is currently ending last year of his PhD course (phylogenetics) as well. AF is PhD and staff graduate technician at the IZSVe. IR is a PhD student and a bioinformatician; IM is PhD and Head of the Innovative Diagnostic Laboratory at the IZSVe. GC is the Head of Research and Development Department, Division of Biomedical Science, OIE/FAO and National Reference Laboratory for Newcastle Disease and Avian Influenza, IZSVe; FF is PhD and Associate Professor of Molecular Biology and

Bioinformatics, and the PI of the MOLBINFO Unit at the Department of Biology, University of Padua.

ACKNOWLEDGEMENTS

This work was supported by basic funding ('ex 60%') from the Italian Ministry for University and Research (MIUR) to FF.

REFERENCES

- Alexander DJ. An overview of the epidemiology of avian influenza. Vaccine. 2007;25(30):5637-44.
- Al-Tawfiq JA, Zumla A, Gautret P, Gray GC, Hui DS, Al-Rabeeah AA, Memish ZA. Surveillance for emerging respiratory viruses. Lancet Infect Dis. 2014;14(10):992-1000.
- Bedford T, Riley S, Barr IG, Broor S, Chadha M, Cox NJ, Daniels RS, Gunasekaran CP, Hurt AC, Kelso A, Klimov A, Lewis NS, Li X, McCauley JW, Odagiri T, Potdar V, Rambaut A, Shu Y, Skepner E, Smith DJ, Suchard MA, Tashiro M, Wang D, Xu X, Lemey P, Russell CA. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature 2015; 523(7559):217-20.
- Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W510-4.
- Bielejec F, Rambaut A, Suchard MA, Lemey P. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. Bioinformatics. 2011;27(20):2910-2.
- Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T: Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc 2009; 4(1):1-13.
- Brown MV, Ostrowski M, Grzymski JJ, Lauro FM. A trait based perspective on the biogeography of common and abundant marine bacterioplankton clades. Mar Genomics. 2014;15:17-28.
- Burke DF, Smith DJ. A recommended numbering scheme for influenza A HA subtypes. PLoS One. 2014;9(11):e112302.
- Butler D. Flu surveillance lacking. Nature. 2012 Mar 28;483(7391):520-2. doi:10.1038/483520a.
- Butt KM, Smith GJ, Chen H, Zhang LJ, Leung YH, Xu KM, Lim W, Webster RG, Yuen KY, Peiris JS, Guan Y. Human infection with an avian H9N2 influenza A virus inHong Kong in 2003. J Clin Microbiol. 2005;43(11):5760-7.
- Carugo O, Pongor S: A normalized root mean square distance for comparing protein three dimensional structures. Protein Sci 2001;10:1470-1473.
- Chutinimitkul S, Herfst S, Steel J, Lowen AC, Ye J, van Riel D, Schrauwen EJ, Bestebroer TM, Koel B, Burke DF, Sutherland-Cash KH, Whittleston CS, Russell CA, Wales DJ, Smith DJ, Jonges M, Meijer A, Koopmans M, Rimmelzwaan GF, Kuiken T, Osterhaus AD, Garcìa-Sastre A, Perez DR, Fouchier RA. Virulence-associated substitution D222G in the hemagglutinin of 2009 pandemic influenza A(H1N1) virus affects receptor binding. J Virol. 2010;84(22):11802-13.
- Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 2005;22(5):1185-92.
- Farber DL, Sleasman JW, Virella G. Immune response: Antigens, Lymphocytes and Accessory Cells. Medical Immunology, Sixth Edition, 2007; Chapter 4:35-54.
- Gambaryan, A., R. Webster, and M. Matrosovich. Differences between influenza virus receptors on target cells of duck and chicken. Arch. Virol. 2002;147:1197-1208.

- Gambaryan A, Tuzikov A, Pazynina G, Bovin N, Balish A, Klimov A. Evolution of the receptor binding phenotype of influenza A (H5) viruses. Virology. 2006;344(2):432-8.
- Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A: Protein Identification and Analysis Tools on the ExPASy Server. In The Proteomics Protocols Handbook Edited by Walker JM. Humana Press 2005:571-607.
- Gräf T, Vrancken B, Maletich Junqueira D, de Medeiros RM, Suchard MA, Lemey P, Esteves de Matos Almeida S, Pinto AR. Contribution of Epidemiological Predictors in Unraveling the Phylogeographic History of HIV-1 Subtype C in Brazil. J Virol. 2015;89(24):12341-8.
- Guan Y, Smith GJ. The emergence and diversification of panzootic H5N1 influenza viruses. Virus Res. 2013;178(1):35-43.
- Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52(5):696-704.
- Hill SC, Lee YJ, Song BM, Kang HM, Lee EK, Hanna A, Gilbert M, Brown IH, Pybus OG. Wild waterfowl migration and domestic duck density shape the epidemiology of highly pathogenic H5N8 influenza in the Republic of Korea. Infect Genet Evol. 2015;34:267-77.
- Hu M, Li X, Ni X, Wu J, Gao R, Xia W, Wang D, He F, Chen S, Liu Y, Guo S, Li H, Shu Y, Bethel JW, Liu M, Moore JB, Chen H. Coexistence of Avian Influenza Virus H10 and H9 Subtypes among Chickens in Live Poultry Markets during an Outbreak of Infection with a Novel H10N8 Virus in Humans in Nanchang, China. Jpn J Infect Dis. 2015;68(5):364-9.
- Jin Y, Yu D, Ren H, Yin Z, Huang Z, Hu M, Li B, Zhou W, Yue J, Liang L. Phylogeography of Avian influenza A H9N2 in China. BMC Genomics. 2014;15:1110.
- Kobayashi Y, Suzuki Y. Compensatory evolution of net-charge in influenza A virus hemagglutinin. PLoS One. 2012;7(7):e40422.
- Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5(9):e1000520.
- Lin YP, Shaw M, Gregory V, Cameron K, Lim W, Klimov A, Subbarao K, Guan Y, Krauss S, Shortridge K, Webster R, Cox N, Hay A. Avian-to-human transmission of H9N2 subtype influenza A viruses: relationship between H9N2 and H5N1 human isolates. Proc Natl Acad Sci U S A. 2000;97(17):9654-8.
- Lu L, Lycett SJ, Leigh Brown AJ. Determining the phylogenetic and phylogeographic origin of highly pathogenic avian influenza (H7N3) in Mexico. PLoS One. 2014;9(9):e107330.
- Maixner F, Krause-Kyora B, Turaev D, Herbig A, Hoopmann MR, Hallows JL, Kusebauch U, Vigl EE, Malfertheiner P, Megraud F, O'Sullivan N, Cipollini G, Coia V, Samadelli M, Engstrand L, Linz B, Moritz RL, Grimm R, Krause J, Nebel A, Moodley Y, Rattei T, Zink A. The 5300-year-old Helicobacter pylori genome of the Iceman. Science. 2016;351(6269):162-5.
- Monne I, Fusaro A, Nelson MI, Bonfanti L, Mulatti P, Hughes J, Murcia PR, Schivo A, Valastro V, Moreno A, Holmes EC, Cattoli G. Emergence of a highly pathogenic avian influenza virus from a low-pathogenic progenitor. J Virol. 2014;88(8):4375-88.
- Nelson MI, Vincent AL. Reverse zoonosis of influenza to swine: new perspectives on the human-animal interface. Trends Microbiol. 2015;23(3):142-53.
- Ni G, Li Q, Kong L, Yu H. Comparative phylogeography in marginal seas of the northwestern Pacific. Mol Ecol. 2014;23(3):534-48.

- Peacock T, Reddy K, James J, Adamiak B, Barclay W, Shelton H, Iqbal M. Antigenic mapping of an H9N2 avian influenza virus reveals two discrete antigenic sites and a novel mechanism of immune escape. Sci Rep. 2016;6:18745.
- Peiris M, Yuen KY, Leung CW, Chan KH, Ip PL, Lai RW, Orr WK, Shortridge KF. Human infection with influenza H9N2. Lancet. 1999;354(9182):916-7.
- Perez DR, Lim W, Seiler JP, Yi G, Peiris M, Shortridge KF, Webster RG. Role of quail in the interspecies transmission of H9 influenza A viruses: molecular changes on HA that correspond to adaptation from ducks to chickens. J Virol. 2003;77(5):3148-56.
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605-1612.
- Pollett S, Nelson MI, Kasper M, Tinoco Y, Simons M, Romero C, Silva M, Lin X, Halpin RA, Fedorova N, Stockwell TB, Wentworth D, Holmes EC, Bausch DG. Phylogeography of Influenza A(H3N2) Virus in Peru, 2010-2012. Emerg Infect Dis. 2015;21(8):1330-8.
- Pyron RA. Post-molecular systematics and the future of phylogenetics. Trends Ecol Evol. 2015;30(7):384-9.
- Richter S, Wenzel A, Stein M, Gabdoulline RR, Wade R. WebPIPSA: a web server for the comparison of protein interaction properties. Nucleic Acid Res 2008;36(Web Server Issue):W276-W280.
- Riddle BR. What is modern biogeography without phylogeography? J. Biogeogr 2009;36:1-2.
- Righetto I, Milani A, Cattoli G, Filippini F. Comparative structural analysis of haemagglutinin proteins from type A influenza viruses: conserved and variable features. BMC Bioinformatics. 2014;15:363.
- Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19(12):1572-4.
- Sang X, Wang A, Ding J, Kong H, Gao X, Li L, Chai T, Li Y, Zhang K, Wang C, Wan Z, Huang G, Wang T, Feng N, Zheng X, Wang H, Zhao Y, Yang S, Qian J, Hu G, Gao Y, Xia X. Adaptation of H9N2 AIV in guinea pigs enables efficient transmission by direct contact and inefficient transmission by respiratory droplets. Sci Rep. 2015;5:15928.
- Sitkoff D, Sharp K, Honig B: Accurate calculation of hydration free energies using macroscopic solvent models. J Phys Chem 1994;98:1978-1988.
- Stacy A, McNally L, Darch SE, Brown SP, Whiteley M. The biogeography of polymicrobial infection. Nat Rev Microbiol. 2016;14(2):93-105.
- Stanekova Z and Vareckova E: Conserved epitopes of influenza A virus inducing protective immunity and their prospects for universal vaccine development. Virol J. 2010;7:351.
- Su S, Bi Y, Wong G, Gray GC, Gao GF, Li S. Epidemiology, Evolution, and Recent Outbreaks of Avian Influenza Virus in China. J Virol. 2015;89(17):8671-6.
- Swayne DE. Impact of vaccines and vaccination on global control of avian influenza. Avian Dis. 2012;56(4 Suppl):818-28.
- Tamura K, Kumar S. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol Biol Evol. 2002;19(10):1727-36.
- Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28(10):2731-9.

- Tian H, Zhou S, Dong L, Van Boeckel TP, Cui Y, Newman SH, Takekawa JY, Prosser DJ, Xiao X, Wu Y, Cazelles B, Huang S, Yang R, Grenfell BT, Xu B. Avian influenza H5N1 viral and bird migration networks in Asia. Proc Natl Acad Sci U S A. 2015;112(1):172-7.
- Trombetta C, Piccirella S, Perini D, Kistner O, Montomoli E. Emerging Influenza Strains in the Last Two Decades: A Threat of a New Pandemic? Vaccines (Basel). 2015;3(1):172-85.
- Turchetto-Zolet AC, Pinheiro F, Salgueiro F, Palma-Silva C. Phylogeographical patterns shed light on evolutionary process in South America. Mol Ecol. 2013;22(5):1193-213.
- Velkov T, Ong C, Baker MA, Kim H, Li J, Nation RL, Huang JX, Cooper MA, Rockman S. The antigenic architecture of the hemagglutinin of influenza H5N1 viruses. Mol Immunol. 2013;56(4):705-19.
- Vines A, Wells K, Matrosovich M, Castrucci MR, Ito T, Kawaoka Y. The role of influenza A virus hemagglutinin residues 226 and 228 in receptor specificity and host range restriction. J Virol. 1998;72(9):7626-31.
- Wilson IA, Skehel JJ, Wiley DC. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution. Nature. 1981;289(5796):366-73.
- WHO/OIE/FAO H5N1 Evolution Working Group. Toward a unified nomenclature system for highly pathogenic avian influenza virus (H5N1). Emerg Infect Dis. 2008;14(7):e1.
- Young CR. Structural requirements for Immunogenicity and Antigenicity. in: Molecular Immunology - A Textbook. CRC Press, Atassi, Van Oss, Absolom eds. 1984;1-14.
- Zhang L, Li H, Li S, Zhang A, Kou F, Xun H, Wang P, Wang Y, Song F, Cui J, Cui J, Gouge DH, Cai W. Phylogeographic structure of cotton pest Adelphocoris suturalis (Hemiptera: Miridae): strong subdivision in China inferred from mtDNA and rDNA ITS markers. Sci Rep. 2015;5:14009.

clades	C.2.3	C.2.2	C.2.1	C.2	C.1	B.2.7	B.2.6	B.2.5	B.2.4	B.2.3	B.2.2	B.2.1	B.1.1	B.1.2	B.4	B.3	A.1	A.2	A.3	A.4	A.5.1	A.5.2	A.5.3	A.5.4	A.5.5	D.	E.
C.2.3	*		[0,5]	[0,4]	[0,6]	[0,7]	[0,6]	[0,6]	[0,7]	[0,7]	[0,7]	[0,7]	[0,6]	[0,6]	[0,6]	[0,6]	[0,8]	[0,8]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,7]	[0,7]	[0,8]	[0,7]
C.2.2	8,3	*	[0,5]	[0,4]	[0,6]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,6]	[0,6]	[0,6]	[0,6]	[0,7]	[0,8]	[0,6]	[0,6]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,6]
C.2.1	8,4	10	*	[0,4]	[0,6]	[0,7]	[0,6]	[0,6]	[0,7]	[0,7]	[0,7]	[0,7]	[0,6]	[0,6]	[0,7]	[0,6]	[0,8]	[0,8]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,7]
C.2	**6,7**	8,3	7,7	*	[0,5]	[0,6]	[0,6]	[0,6]	[0,6]	[0,6]	[0,6]	[0,6]	[0,5]	[0,5]	[0,5]	[0,5]	[0,7]	[0,7]	[0,6]	[0,6]	[0,7]	[0,7]	[0,7]	[0,6]	[0,7]	[0,7]	[0,6]
C.1	9,9	**10,9**	10,4	7,5	*	[0,7]	[0,6]	[0,6]	[0,7]	[0,7]	[0,7]	[0,7]	[0,6]	[0,6]	[0,6]	[0,6]	[0,8]	[0,8]	[0,7]	[0,7]	[0,8]	[0,8]	[0,8]	[0,8]	[0,8]	[0,8]	[0,7]
B.2.7	14,2	14,9	14,9	12,9	12,9	*	[0,4]	[0,5]	[0,5]	[0,5]	[0,5]	[0,5]	[0,5]	[0,5]	[0,6]	[0,5]	[0,8]	[0,8]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,7]
B.2.6	14,2	15	14,3	12,8	13	7,5	*	[0,4]	[0,4]	[0,5]	[0,5]	[0,4]	[0,5]	[0,5]	[0,6]	[0,5]	[0,8]	[0,8]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]
B.2.5	13,2	14,1	14,3	12	12,6	7,2	7,3	*	[0,5]	[0,5]	[0,5]	[0,4]	[0,5]	[0,5]	[0,5]	[0,5]	[0,8]	[0,8]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,7]
B.2.4	14,7	15,1	15,2	13,4	13,7	9,2	8,8	8,5	*	[0,5]	[0,4]	[0,4]	[0,6]	[0,6]	[0,6]	[0,5]	[0,8]	[0,8]	[0,7]	[0,6]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,7]
B.2.3	13,6	14,6	14,9	12,5	13,1	8	8	7,1	8,2	*	[0,5]	[0,4]	[0,5]	[0,6]	[0,6]	[0,5]	[0,8]	[0,8]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,7]
B.2.2	13,8	14,9	15	12,8	13,5	8,9	8,8	8,4	9,5	8,4	*	[0,4]	[0,5]	[0,5]	[0,6]	[0,5]	[0,8]	[0,8]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,7]
B.2.1	13,1	13,9	13,7	11,4	11,7	7,2	7	**6,6**	7,5	6,9	7,9	*	[0,5]	[0,5]	[0,5]	[0,5]	[0,8]	[0,8]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,8]	[0,7]
B.1.1	11,1	12,1	12,5	9,8	9,8	9,2	9,2	8,9	9,7	8,9	9,5	7,9	*	[0,3]	[0,4]	[0,4]	[0,8]	[0,8]	[0,6]	[0,7]	[0,6]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,7]
B.1.2	11,6	12,2	12,4	9,9	10,2	9,8	9,9	9,5	10,2	9,7	10,1	8,5	**5,4**	*	[0,5]	[0,4]	[0,8]	[0,7]	[0,6]	[0,6]	[0,6]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,7]
B.4	11,3	12,4	12,3	9,9	10,2	9,1	9,1	8,4	9,6	8,9	9,5	7,6	**5,4**	**6,1**	*	[0,5]	[0,8]	[0,8]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,8]	[0,7]	[0,8]	[0,8]
B.3	11,5	12,5	12,6	10	10,2	9,8	9,7	9,3	**10,3**	9,7	10,2	8,3	5,8	6,1	6,1	*	[0,8]	[0,8]	[0,7]	[0,7]	[0,7]	[0,7]	[0,8]	[0,8]	[0,7]	[0,8]	[0,7]
A.1	17,1	18,5	19,7	17,6	18,4	19,4	19	18,9	20,1	19,3	19,1	18,8	17,3	18	18,1	17,9	*	[0,8]	[0,7]	[0,7]	[0,8]	[0,8]	[0,8]	[0,8]	[0,8]	[0,8]	[0,7]
A.2	18,1	19	19,2	17,9	17,7	19,4	19,6	19,5	20,2	19,4	19,3	18,6	18	18,5	18,2	18,5	16,8	*	[0,7]	[0,7]	[0,8]	[0,7]	[0,7]	[0,7]	[0,8]	[0,8]	[0,7]
A.3	15,3	16,1	15,4	14,1	14	16,9	16,9	16,1	17,2	16,6	17	15,9	14,8	14,7	14,8	14,7	15,8	15,5	*	[0,5]	[0,6]	[0,6]	[0,6]	[0,6]	[0,7]	[0,7]	[0,6]
A.4	14	15,2	15	13,1	13	15,8	15,9	15,1	15,8	15,2	15,7	14,2	13,3	13,5	13,5	13,2	15,3	14,7	9,4	*	[0,5]	[0,6]	[0,6]	[0,6]	[0,7]	[0,7]	[0,6]
A.5.1	16,1	16,7	16,6	15,4	15,7	17,3	17,4	16,6	17,2	16,8	17	16	15,4	15	15,3	15,4	17,2	16	12,5	10,1	*	[0,6]	[0,5]	[0,5]	[0,6]	[0,7]	[0,6]
A.5.2	17,1	17,5	16,8	15,5	15,4	17,6	17,3	17,3	18	17,5	17,7	17	16,1	15,9	16,2	16,2	16,8	16,5	13,4	11,7	11,9	*	[0,6]	[0,6]	[0,7]	[0,8]	[0,6]
A.5.3	16,6	17	16,5	15,4	15,4	17,2	17	16,4	17,4	17,1	16,9	16,2	15,3	15,1	15,6	15,4	16,4	15,8	11,5	10,3	10,1	11,1	*	[0,4]	[0,6]	[0,7]	[0,6]
A.5.4	15,6	16,1	15,4	14,7	14,6	16,7	16,8	16,3	17,1	16,6	16,7	15,7	14,9	14,8	15,1	14,7	16	15,4	10,5	9	8,6	9,9	**6,9**	*	[0,6]	[0,7]	[0,6]
A.5.5	17,2	17,8	17,3	16,3	16,8	18,6	18,4	17,9	18,7	18	18,8	17,5	16,7	16,7	16,9	16,6	**18,3**	18,1	14,2	12,5	12,5	13,9	11,9	11	*	[0,8]	[0,7]
D.	15,2	16,1	16,2	14,6	14,9	17,1	17	16,4	16,4	16,2	16,6	15,7	14	14,2	14,3	14,5	18,4	17	13,5	11,6	14,3	15,4	14,2	13,5	16,6	*	[0,7]
E.	14,4	15,2	15,4	13,7	13,6	16	16	15,5	16,2	15,5	15,6	14,7	13,6	13,6	13,8	13,6	16,5	15,2	11,6	10,3	12,5	13,6	12,5	11,6	15	11,6	*

Table 1: Estimates of evolutionary distances between H9N2 identified clades. Evolutionary distances (calculated by p-distance) between identified clades were calculated using 1669 nucleotide HA sequences ≥ 1500 bp in lenght. Max/min evolutionary distance values within classes A, B and C are highlighted in respectively red and blue colors. Values between brackets are standard errors, obtained by a bootstrap procedure (500 replicates).

clades	APD Within	Closest clade	APD between clade & closest clade	C-Value
C.2.3	2,5 [0,3]	C.2	6,7	2,7
C.2.2	4,1 [0,2]	C.2.3	8,3	2,0
C.2.1	3,3 [0,2]	C.2	7,7	2,3
C.2	4,7 [0,2]	C.2.3	6,7	1,4
C.1	1,2 [0,1]	C.2	7,5	6,3
B.2.7	3,8 [0,2]	B.2.5	7,2	1,9
B.2.6	4,6 [0,3]	B.2.1	7,3	1,6
B.2.5	3,4 [0,2]	B.2.1	6,6	1,9
B.2.4	4,9 [0,3]	B.2.1	7,5	1,5
B.2.3	2,7 [0,2]	B.2	7,1	2,6
B.2.2	4,4 [0,2]	B.2.1	7,9	1,8
B.2.1	2,7 [0,2]	B.2.1	7,6	2,8
B.1.1	3,3 [0,2]	B.4	5,4	1,6
B.1.2	2,7 [0,2]	B.4	5,4	2,0
B.4	1,6 [0,2]	B.1.1	5,4	3,4
B.3	3,4 [0,2]	B.1.1	5,8	1,7
A.1	1,2 [0,2]	A.4	15,3	12,8
A.2	2,6 [0,3]	A.4	14,7	5,7
A.3	4,5 [0,3]	A.4	9,4	2,1
A.4	3,4 [0,3]	A.5.4	9	2,6
A.5.1	3,8 [0,3]	A.5.4	8,6	2,3
A.5.2	4 [0,4]	A.5.4	9,9	2,5
A.5.3	3,8 [0,2]	A.5.4	6,9	1,8
A.5.4	3,4 [0,2]	A.5.3	6,9	2,0
A.5.5	4,5 [0,3]	A.5.4	11	2,4
D.	0,8 [0,2]	A.4	12	15,0
E.	2,8 [0,2]	A.4	10,4	3,7

Table 2. Estimates of average pairwise distance and C-value within each identified H9N2 clade. The average pairwise distance was calculated using 1669 HA nucleotide sequences (≥ 1500 bp). Values between brackets are standard errors, obtained by a bootstrap procedure (500 replicates). The C-value is the ratio of distance between clade and its closest clade to the distance within clade.

	Criteria used for a class and clade designation
1	The nomenclature was established based on a non redudant dataset of 1669 HA sequences from H9 viruses with sequence lenght ≥ 1500 bp.
2	Classes were assigned based on phylogenetic topology distribution and confirmed by three different methods: Maximum Likelihood, NJ, Byesian.
3	Clades were assigned based on both phylogenetic topology distribution and evolutionary distances between different taxonomic branches. Clades separation was confirmed using the three aforementioned methods (see point 2).
4	New classes and clades were designated only when at least three independent isolates without a direct epidemiologic link (i.e. distinct outbreaks) were available.
5	Bootstrap values at the classes and clades defining node should be ≥60%.
6	Distinct clades should have ≥5% average distances between other clades. Distinct clades should have <5% average distances within the clade.
7	Cut-off value 5% with C value ≥ 1 was fixed to assign new clades in each class.

Table 3. H9N2 AI viruses, class and clade identification criteria. Text in the table cells is self-explaining

HA subtype numbering		HA (mature chain) position number										
H9Nx		**131**	**135**	**146**	**161**	**162**	**165**	**180**	**186**	**198**	**216**	**217**
H7Nx		127	134	145	162	163	166	181	187	199	217	218
H5Nx		133	141	152	167	168	171	186	192	204	222	223
H3Nx		137	145	156	171	172	175	190	196	208	226	227
H1Nx		134	142	153	168	169	172	187	193	205	223	224
Clade	**Strains**	**Fully / most conserved amino acid for each clade**										
A.1	4	R	N	H	N	E	N	E	K	D	Q	Q
A.2	3	K	N	H	T	E	N	E	K	D	Q	Q
A.3	33	K	N	H	N	E	N	E	K	D	Q	Q
A.4	8	K	N	H	N	E	N	E	K	D	Q	Q
A.5.1	15	A	N	H	N	E	N	E	K	D	Q	Q
A.5.2	3	A	N	Q	N	N	S	E	K	D	Q	Q
A.5.3	48	K	N	H	N	E	N	E	K	D	Q	Q
A.5.4	12	R	N	H	N	E	N	E	E	D	Q	Q
A.5.5	51	K	G	H	D	W	N	E	K	D	Q	Q
D	3	K	N	H	N	E	S	E	K	D	Q	Q
E	13	K	N	H	N	E	S	E	K	D	Q	Q
B	3	K	D	Q	N	R	S	A	V	D	Q	Q
B.1.1	30	K	D	Q	N	R	S	A	I	D	L	Q
B.1.2	44	K	D	Q	N	R	S	A	I	D	L	Q
B.3	7	R	G	Q	N	R	S	E	I	D	L	Q
B.4	7	K	D	Q	N	R	S	A	T	D	L	Q
B.2.1	33	K	D	Q	N	R	D	A	T	N	L	I
B.2.2	36	K	D	Q	N	R	D	A	T	N	L	I
B.2.3	9	K	N	Q	N	R	D	A	T	N	L	I
B.2.4	7	K	D	Q	N	Q	D	A	T	N	L	I
B.2.5	31	K	D	Q	N	R	D	A	T	N	L	I
B.2.6	33	K	E	Q	N	R	D	T	T	N	L	I
B.2.7	124	K	D	Q	N	R	D	A	T	N	L	I
C.1	31	N	N	Q	N	R	S	V	T	D	L	T
C.2	224	K	D	Q	N	R	N	T	T	D	Q	Q
C.2.1	150	K	D	Q	N	R	N	A	T	D	L	Q
C.2.2	694	K	D	Q	N	Q	N	T	T	D	L	Q
C.2.3	3	K	D	Q	N	R	N	V	T	D	L	Q

Table 4. Class and sub-class specific variation at the RBD in H9N2 viruses. The number of strain sequences and the most represented residue for each amino acid position (columns) are reported for each clade (rows). Negatively charged, positively charged and hydrophobic residues are highlighted by red, blue or yellow background, respectively. Position numbering for four HA subtypes is reported.

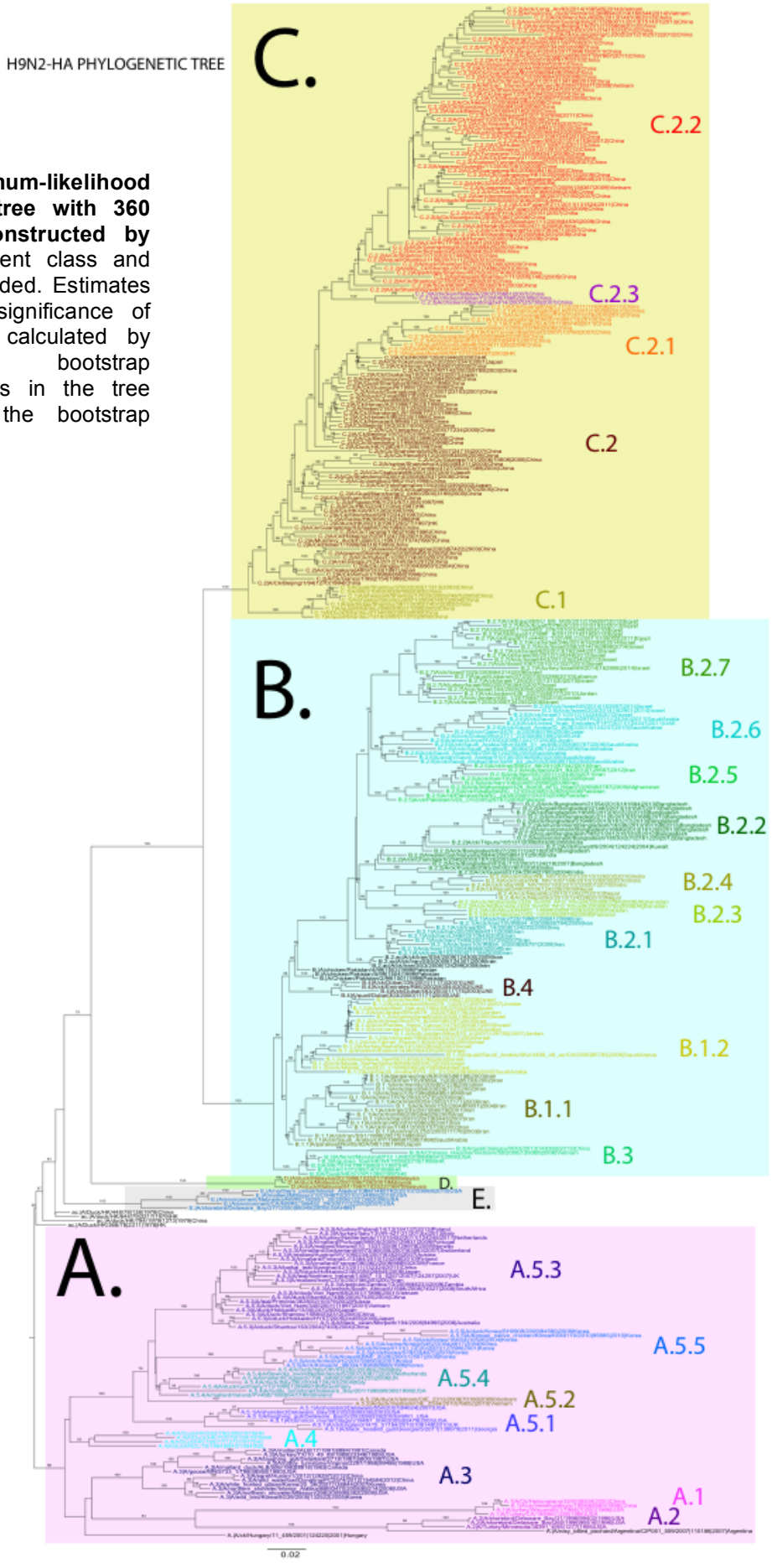


Figure 1. Maximum-likelihood short alignment tree with 360 H9N2 isolates constructed by PhyML. The different class and clades are color coded. Estimates of the statistical significance of phylogenies were calculated by performing 100 bootstrap replicates. Numbers in the tree nodes represent the bootstrap support (≥60).

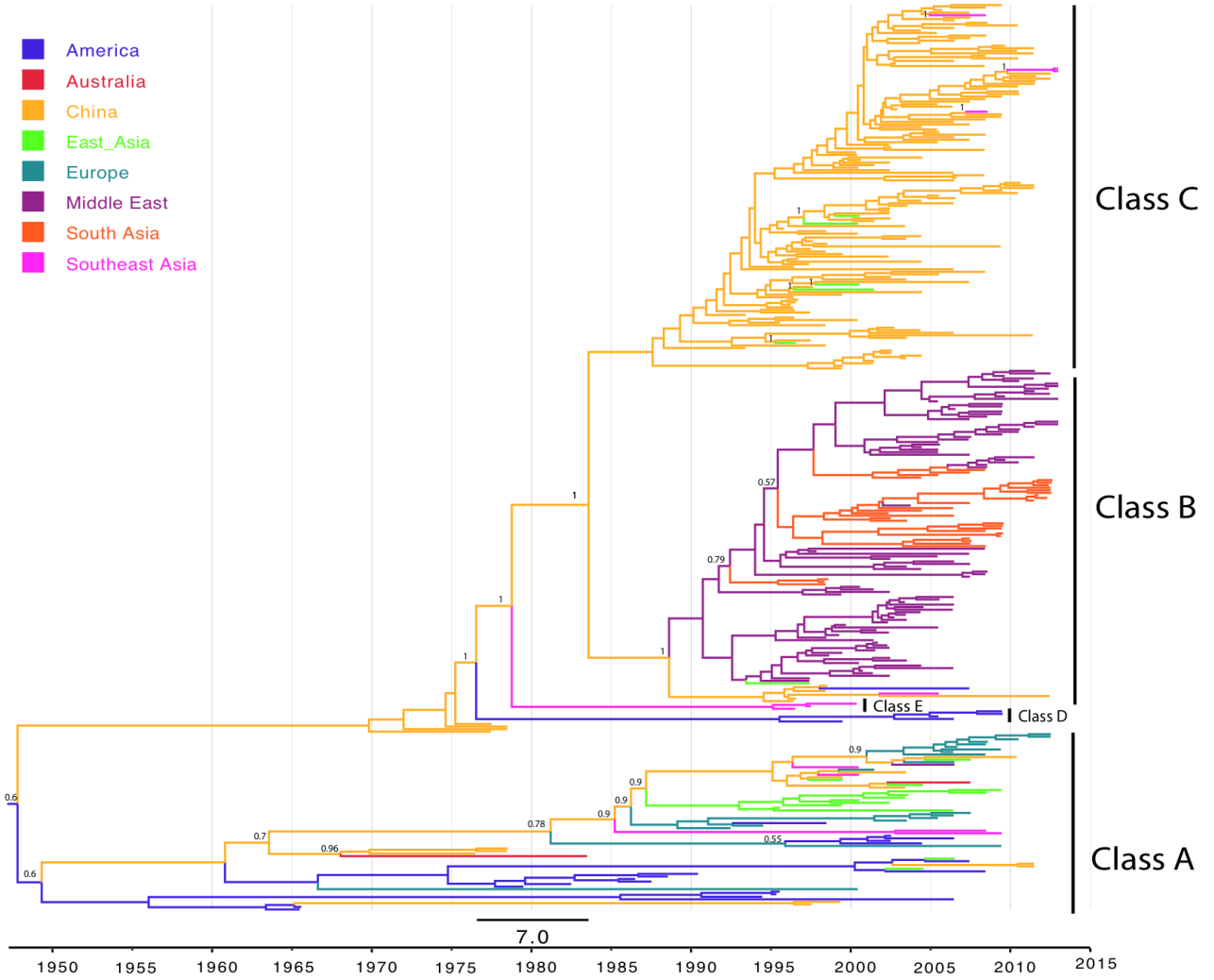


Figure 2. Maximum clade credibility (MCC) phylogenies inferred for the HA gene sequences of 357 viruses of AI H9 subtype. Branches are coloured according to the most probable ancestor location (in terms of geographic area) of their descendent nodes. Timeline at the bottom indicates the years before the most recent sampling time. Numbers are reported at branch points where state probabilities with values ≥ 0.55 correspond to geographic area transition events.

Figure 3. Phylogeography worldwide mapping and spreading of AI H9 subtype viruses. The eigth geographic areas in which H9 viruses are so far known to circulate are differently coloured. Viral transition from a geographic area to another one with the well supported Bayes factor > 3 and state probabilities values ≥ 0.55 are represented by arrows with colour code based on corresponding H9 classes.

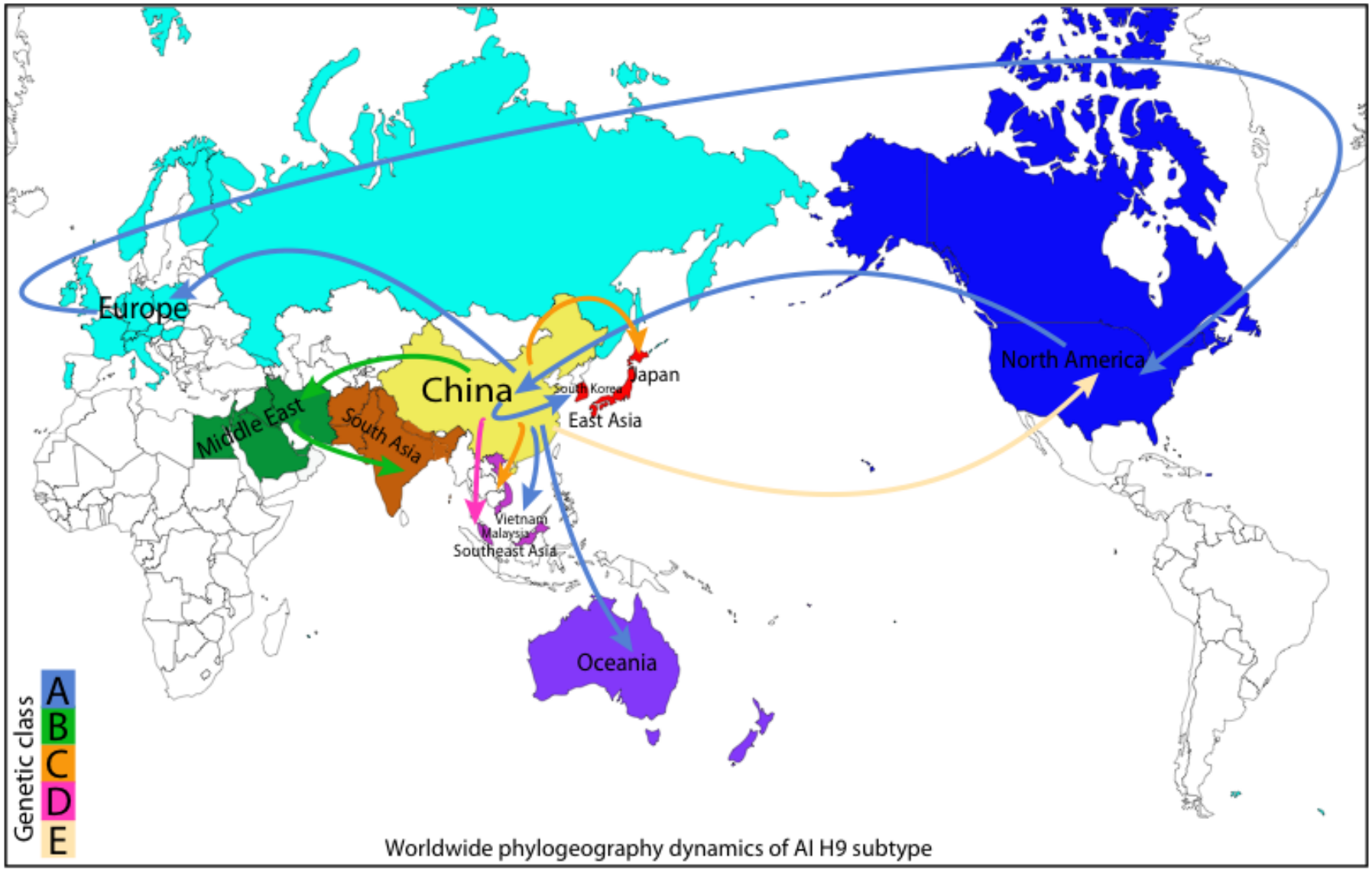


Figure 4. Heat map and density plot for the RBD subregion from representative H9N2 strains. The electrostatic distance formula is reported. In both density plot and heat map, warm to cold color shift corresponds to increasing electrostatic distance.

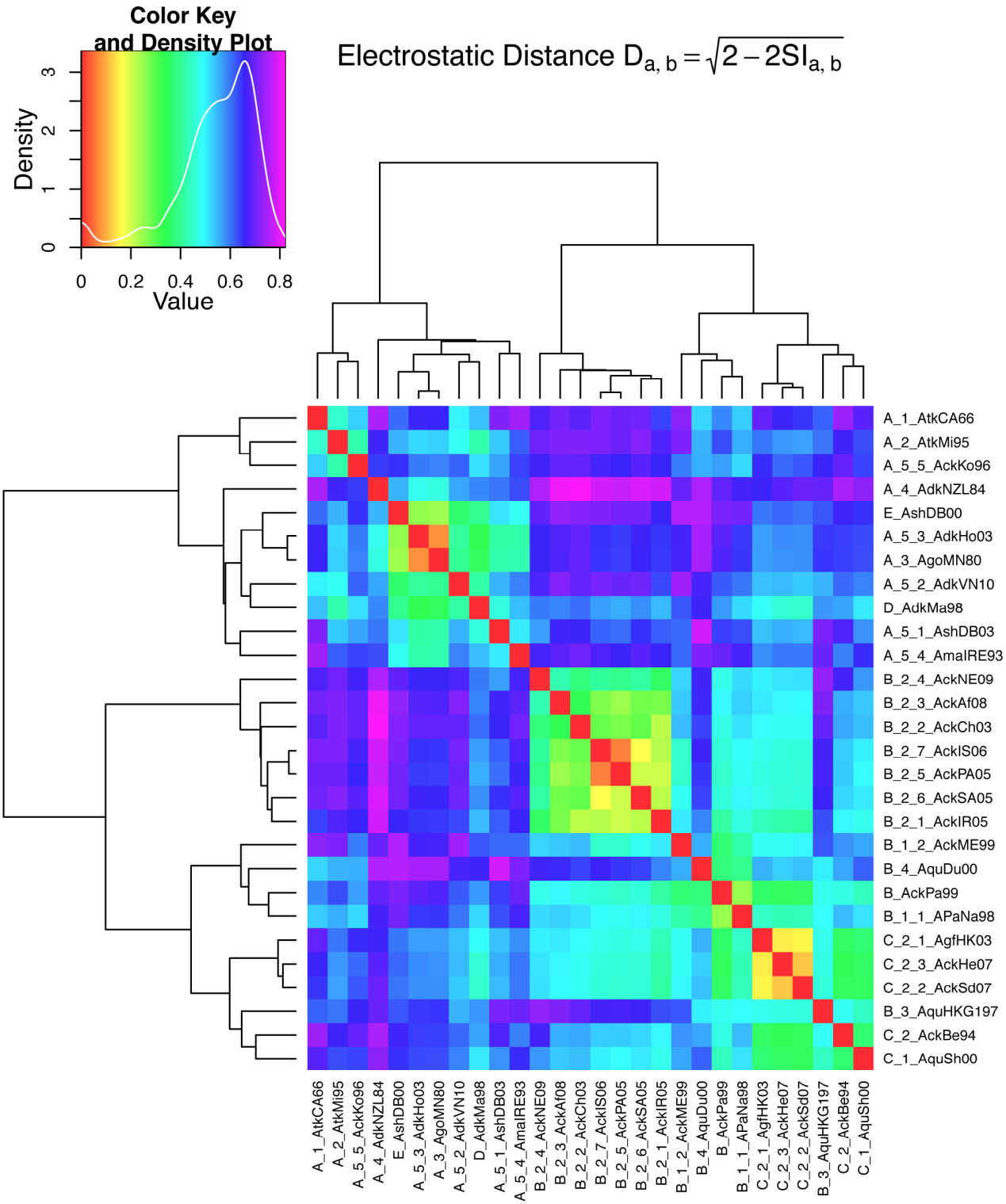


Figure 5. Electrostatic potential diagram (epogram) for the RBD subregion from representative H9N2 strains. The electrostatic distance formula is reported. The two main clusters are 3/4 bordered by red lines.

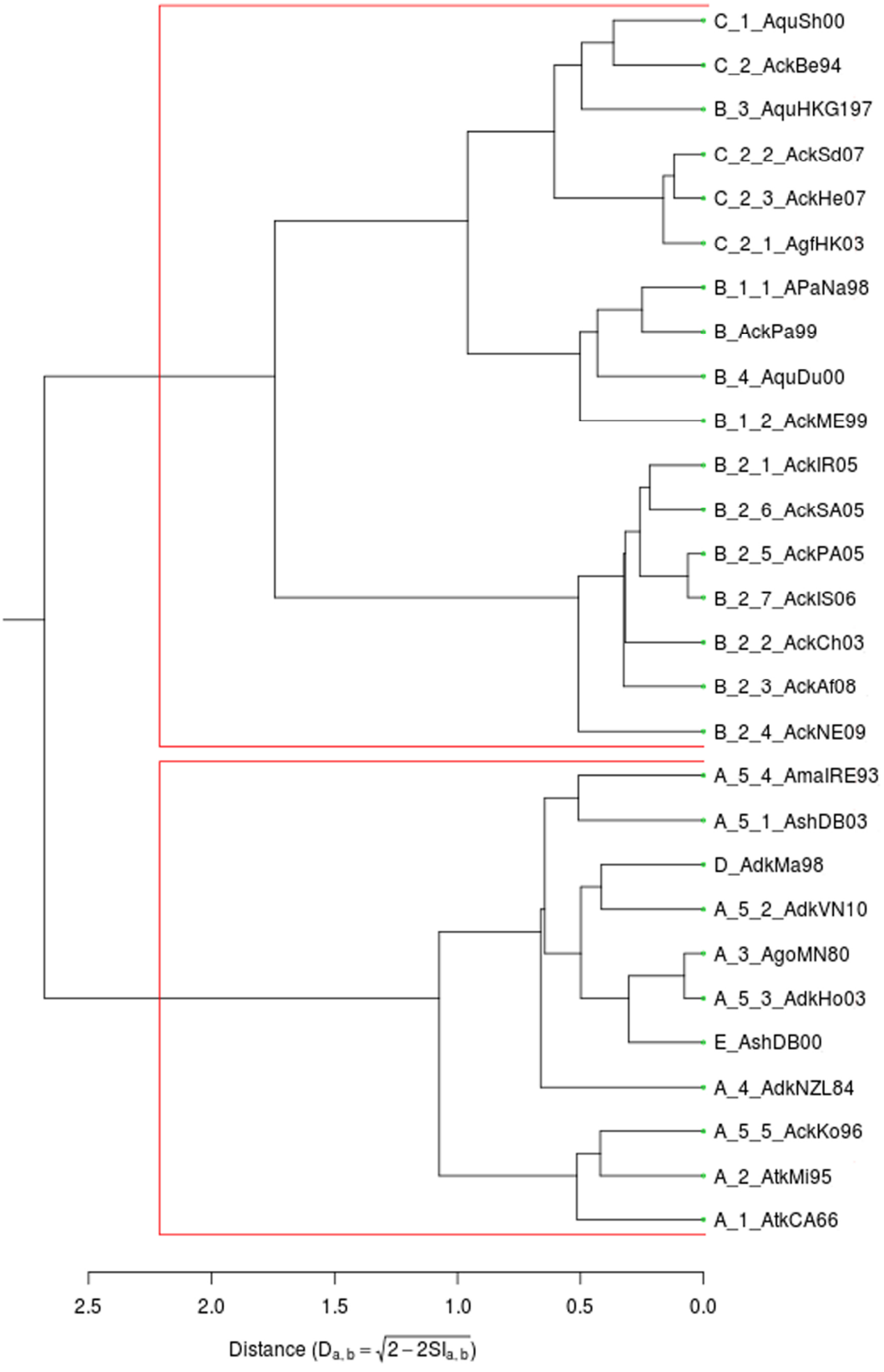
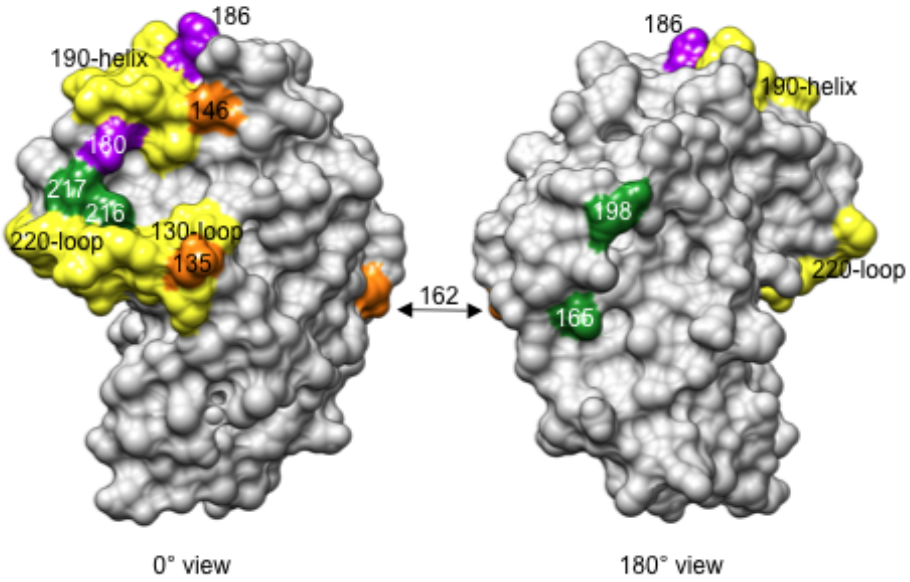


Figure 6. Location at the RBD surface of residues involved in class and sub-class specific variation in H9N2. Both 0° view (left image) and 180° view (right image) are shown. The three regions (130-loop, 190-helix and 220-loop) that mediate binding to the sialic acid (SA) moieties from the host cell (Wilson et al., 1981; Kobayashi et al., 2012) are highlighted in yellow. Color coding for amino acid positions is the following: 135, 146 and 162: orange; 180 and 186: purple; 165, 198, 216 and 217: green.



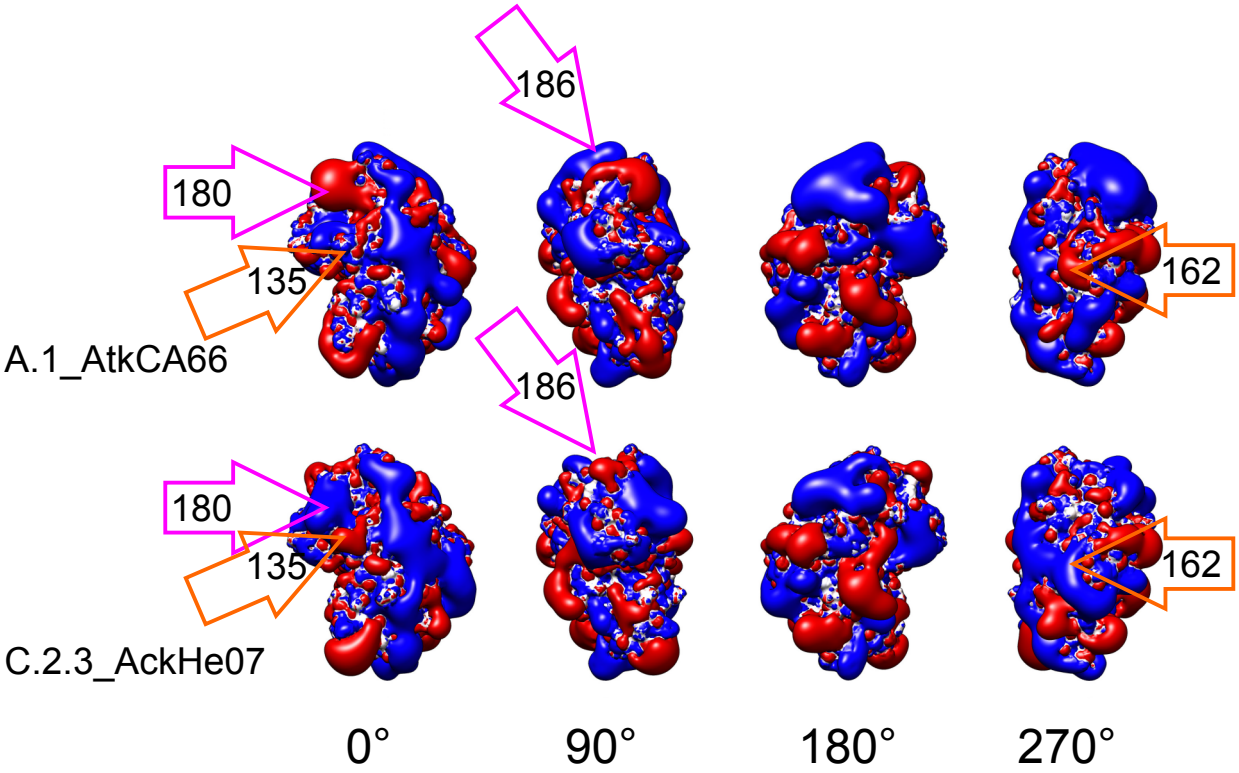


Figure 7: Comparison of representative profiles for the isopotential RBD contour in H9N2. Four 90° stepwise rotation view are presented for each representative RBD electrostatic isocontour. Names of the specific H9N2 virus strains are reported. Arrows for specific residues are color-coded according to Figure 6.

CONCLUDING REMARKS

The integration of genomic and bioinformatic analysis used during my PhD proved successful as an approach to explore and shed light on the complexity of the avian hemagglutinin evolution. Influenza A viruses, like other RNA viruses, possesses inherent characteristics, such as a rapid replication, high evolutionary rates, and the ability to infect a wide range of hosts, which altogether favour the formation of a heterogeneous population of variants, genetically related but not identical. Variables of different nature, such as the external environment, host adaptation, the use of vaccines and various other external forces, can exert a constant and selective pressure on the virus populations influencing their evolution, favouring the presence of minority variants and potentially the selection of new viruses with changed genetic and antigenic properties. The contribution given by deep sequencing, phylogeny and structural analysis allowed to obtain a complete overview on the viral evolution.

During my PhD study I focused my work on the application, improvement and refinement of new protocols in pre and post steps related to deep sequencing analysis, application of bioinformatic tools to clean raw data output files, to assemble reads and to investigate the heterogeneity viral populations and minority composition variants of each sample so as to infer information from data generated and application of structural analysis to implement phylogenetic and genomic result. Furthermore the consensus sequences generated by both Sanger and Next Generation Sequencing were investigated using phylogenetic analysis to study their genetic relationships and their evolutionary dynamics. The integrated use of the different approaches applied and specifically set up to study avian influenza viruses allowed to obtain interesting results helpful to investigate the acquisition of virulence determinants or host specificity, the detection of minority variants thus providing a better understanding to the influenza virus evolutionary dynamics.

Three Italian epidemics related to H7 subtype have been object of study. Chapter 1 describes the comparison between H7N1 and H7N3 avian influenza epidemics having interested Northern Italy during 1999-2001 and 2002-2004, respectively. Instead, chapter 2 focuses on virus population heterogeneity and virus pathogenicity evolution on clinical samples, collected during the HPAI H7N7 epidemic that interested five Italian industrial holdings and a backyard during the summer of 2013. Chapter 3 inspects the abundance of minority variants and the evolutionary effect on the viral population under the influence of vaccine immune pressure; samples analyzed in this study were obtained from an experimental study aiming at assessing the protection efficacy of two distinct vaccines against HPAI H5N1 virus. As illustrated in chapter 3, viral population heterogeneity and analysis of immune pressure on the viral evolution have been studied on H5 subtypes related to the avian host. As the experimental challenge, the avian host was immunized with vaccines conferring different levels of protection and then infected by an HPAI H5N1 virus to compare protection potentials. Given that antigenic recognition, immune escape and host specificity largely depend on variation in interaction/binding sites, comparative structural analyses performed in my thesis aimed at highlighting conserved and variable features amongst hemagglutinin proteins from different type A influenza viruses (Chapters 4 and 5). A special attention was addressed to H9N2 and HPAI H5N1 subtypes to find and display surface differences on the hemagglutinin protein possibly underying functional evolution. Sanger sequencing, deep sequencing and phylogenetic approaches revealed to be valid tools to study the evolutionary dynamics and the adaptive strategies of these two distinct avian influenza lineages and to highlight many similarities. In chapter 5 strains of avian influenza A virus belonging to H9N2 subtype were

analysed using phylogenetic and structural approaches. The first approach has allowed to obtain a novel classification scheme considering both phylogenetic topology and evolutionary distances; whereas the structural analysis on selected representative viruses for each clade allowed to inspect and confirm whether surface properties could be linked to 'functional evolution' and host adaptation of H9N2 subtype as already seen for HPAI H5N1 viruses.

The next generation sequencing (NGS) has shown to be a powerful and useful tool to study and characterize the complexity of the viral population, allowing to detect low frequency mutations both during the early stages of viral infection and the viral evolution itself by detecting quasi-species variants in avian samples. Despite it expensive cost, deep sequencing analysis allowed to process a huge amount of samples simultaneously providing the generation of an high amount of data in short time that with the classic Sanger sequencing would not be possible.

As described in Chapters 1, 2 and 5, the phylogenetic approach to avian influenza subtypes H7 and H9 proved to be an important tool to explain the molecular epidemiology as well as to shed light on viral transmission, evolutionary dynamics and adaptive strategies. Such phylogenetics approach was properly integrated by the structural one. Particularly in chapters 4 and 5, structural analysis successfully implemented phylogenetic data by inspecting hemagglutinin regions and sub-regions hence highlighting similarities among virus classes and clades that would be undetectable when only considering the primary sequences or the phylogenetic trees.

This study especially focused on the evolution of hemagglutinin from subtypes H5, H7 and H9, and stressed the importance of implementing genomic and structural approaches to analyze both genetic variability and functional evolution of avian influenza A virus. When using primary nucleotide or amino acid sequences, it is only possible to compare amount and kind of mutations, while structural analyses also allow to infer 'functional evolution' by 'weigthing' impact of amino acid substitutions, insertions or deletions on the 3D space of a protein.

Molecular characterization of the influenza A viruses analyzed in this thesis highlighted the heterogeneity of the HA sequences in all analysed subtypes (H5, H7 and H9). This study suggests that careful surveillance of genetic changes in the HA gene and protein during epidemics is needed as it may provide early information on the strains evolution, as well as useful epidemiologic inference.

Last but not least, this work provided a better and implemented understanding on the virus evolution, as well as essential information suggesting further studies to confirm the genetic and evolutionary characteristics of hemagglutinin influenza a viruses. Techniques like reverse genetics analysis could be used to test the results obtained with in silico work.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. PubMed PMID: 2231712.

Andino R, Domingo E. Viral quasispecies. Virology. 2015 May;479-480:46-51. doi: 10.1016/j.virol.2015.03.022. Epub 2015 Mar 29. Review. PubMed PMID: 25824477.

Arinaminpathy N, Grenfell B. Dynamics of glycoprotein charge in the evolutionary history of human influenza. PLoS One. 2010 Dec 30;5(12):e15674. doi: 10.1371/journal.pone.0015674. PubMed PMID: 21209885; PubMed Central PMCID: PMC3012697.

Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics. 2006 Jan 15;22(2):195-201. Epub 2005 Nov 13. PubMed PMID: 16301204.

Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics, 22,195-201.

Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W510-4. doi: 10.1093/nar/gkp322. Epub 2009 May 8. PubMed PMID: 19429685; PubMed Central PMCID: PMC2703985.

Blomberg N, Gabdoulline RR, Nilges M, and Wade RC. Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. Proteins: Str., Function and Genetics 1999, 37: 379-387.

Buermans HP, den Dunnen JT. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta. 2014 Oct;1842(10):1932-1941. doi: 10.1016/j.bbadis.2014.06.015.

Butt AM, Siddique S, Idrees M, Tong Y. Avian influenza A (H9N2): computational molecular analysis and phylogenetic characterization of viral surface proteins isolated between 1997 and 2009 from the human population. Virol J. 2010 Nov 15;7:319. doi: 10.1186/1743-422X-7

Capua I, Alexander DJ. Avian influenza vaccines and vaccination in birds. Vaccine. 2008 Sep 12;26 Suppl 4:D70-3. Review. PubMed PMID: 19230164.

Capua I, Alexander DJ. Ecology, epidemiology and human health implications of avian influenza viruses: why do we need to share genetic data? Zoonoses Public Health. 2008;55(1):2-15. doi: 10.1111/j.1863-2378.2007.01081.x. Review. PubMed PMID: 18201321.

Chen J, Lee KH, Steinhauer DA, Stevens DJ, Skehel JJ, Wiley DC. Structure of the hemagglutinin precursor cleavage site, a determinant of influenza pathogenicity and the origin of the labile conformation. Cell. 1998 Oct 30;95(3):409-17. PubMed PMID: 9814710.

Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986 Apr;5(4):823-6. PubMed PMID: 3709526; PubMed Central PMCID: PMC1166865.

Copeland CS, Doms RW, Bolzau EM, Webster RG, Helenius A. Assembly of influenza hemagglutinin trimers and its role in intracellular transport. J Cell Biol. 1986 Oct;103(4):1179-91. PubMed PMID: 2429970; PubMed Central PMCID: PMC2114319.

Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev. 2012 Jun;76(2):159-216. doi: 10.1128/MMBR.05023-11.

Ducatez MF, Pelletier C, Meyer G. Influenza D virus in cattle, France,2011-2014. Emerg Infect Dis. 2015 Feb;21(2):368-71. doi: 10.3201/eid2102.141449. PubMed PMID: 25628038; PubMed Central PMCID: PMC4313661.

Fiser A. Template-based protein structure modeling. Methods Mol Biol. 2010;673:73-94. doi: 10.1007/978-1-60761-842-3_6. Review. PubMed PMID: 20835794; PubMed Central PMCID: PMC4108304.

Fouchier RA, Munster V, Wallensten A, Bestebroer TM, Herfst S, Smith D, Rimmelzwaan GF, Olsen B, Osterhaus AD. Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. J Virol. 2005 Mar;79(5):2814-22. PubMed PMID: 15709000; PubMed Central PMCID: PMC548452.

França LT, Carrilho E, Kist TB. A review of DNA sequencing techniques. Q Rev Biophys. 2002 May;35(2):169-200. Review. PubMed PMID: 12197303.

Gabdoulline RR, Stein M, Wade RC. qPIPSA: Relating enzymatic kinetic parameters and interaction fields BMC Bioinformatics 2007, 8: 373

Garten W., Klenk H.-D. Cleavage Activation of the Influenza Virus Hemagglutinin and Its Role in Pathogenesis. In: Klenk H.-D., Matrosovich M.N., Stech J., editors. Avian Influenza. Karger; Basel, Switzerland: 2008.

Gómez-Puertas P, Albo C, Pérez-Pastrana E, Vivo A, Portela A. Influenza virus matrix protein is the major driving force in virus budding. J Virol. 2000 Dec;74(24):11538-47. PubMed PMID: 11090151; PubMed Central PMCID: PMC112434.

Guex, N. and Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. Electrophoresis 18: 2714-2723.

Gunsteren, Wilfred F.Van. Biomolecular Simulations: The GROMOS96 Manual and User Guide. 1996
Hall, B.G. (2001) Phylogenetic Trees Made Easy: A How-to Manual for Molecular Biologists. Sunderland Massachusetts: Sinauer Associates.

Hamilton BS, Whittaker GR, Daniel S. Influenza virus-mediated membrane fusion: determinants of hemagglutinin fusogenic activity and experimental approaches for assessing virus fusion. Viruses. 2012 Jul;4(7):1144-68. doi: 10.3390/v4071144. Epub 2012 Jul 24. Review. PubMed PMID: 22852045; PubMed Central PMCID: PMC3407899.

Harrison CJ, Langdale JA. A step by step guide to phylogeny reconstruction. Plant J. 2006 Feb;45(4):561-72. PubMed PMID: 16441349.

Hausmann J, Kretzschmar E, Garten W, Klenk HD. Biosynthesis, intracellular transport and enzymatic activity of an avian influenza A virus neuraminidase: role of unpaired cysteines and individual oligosaccharides. J Gen Virol. 1997 Dec;78 (Pt 12):3233-45. PubMed PMID: 9400974.

Hensley SE, Das SR, Bailey AL, Schmidt LM, Hickman HD, Jayaraman A, Viswanathan K, Raman R, Sasisekharan R, Bennink JR, Yewdell JW. Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. Science. 2009 Oct 30;326(5953):734-6. doi: 10.1126/science.1178258. PubMed PMID: 19900932; PubMed Central PMCID: PMC2784927.

Howard CR, Fletcher NF. Emerging virus diseases: can we ever expect the unexpected? Emerg Microbes Infect. 2012 Dec;1(12):e46. doi: 10.1038/emi.2012.47.

Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001 Aug;17(8):754-5. PubMed PMID: 11524383.

Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins. 2009 Nov 15;77(3):499-508. doi: 10.1002/prot.22458. PubMed PMID: 19507241.

Jagger BW, Wise HM, Kash JC, Walters KA, Wills NM, Xiao YL, Dunfee RL, Schwartzman LM, Ozinsky A, Bell GL, Dalton RM, Lo A, Efstathiou S, Atkins JF, Firth AE, Taubenberger JK, Digard P. An overlapping protein-coding region in influenza A virus segment 3 modulates the host response. Science. 2012 Jul 13;337(6091):199-204. doi: 10.1126/science.1222213. Epub 2012 Jun 28. PubMed PMID: 22745253; PubMed Central PMCID: PMC3552242.

Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Research. 37, D387-D392.

Klenk HD, Rott R, Orlich M, Blödorn J. Activation of influenza A viruses by trypsin treatment. Virology. 1975 Dec;68(2):426-39. PubMed PMID: 173078.

Kobayashi Y, Suzuki Y. Compensatory evolution of net-charge in influenza A virus hemagglutinin. PLoS One. 2012;7(7):e40422. doi:10.1371/journal.pone.0040422. Epub 2012 Jul 12. PubMed PMID: 22808159; PubMed Central PMCID: PMC3395715.

Lam TT, Hon CC, Tang JW. Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. Crit Rev Clin Lab Sci. 2010 Jan-Feb;47(1):5-49. doi: 10.3109/10408361003633318. Review. PubMed PMID: 20367503.

LaRussa P. Pandemic novel 2009 H1N1 influenza: what have we learned? Semin Respir Crit Care Med. 2011 Aug;32(4):393-9. doi: 10.1055/s-0031-1283279. Epub 2011 Aug 19. Review. PubMed PMID: 21858744.

Lu Y, Qian XY, Krug RM. The influenza virus NS1 protein: a novel inhibitor of pre-mRNA splicing. Genes Dev. 1994 Aug 1;8(15):1817-28. PubMed PMID: 7958859. Qiu Y, Krug RM. The influenza virus NS1 protein is a poly(A)-binding protein that inhibits nuclear export of mRNAs containing poly(A). J Virol. 1994 Apr;68(4):2425-32. PubMed PMID: 7908060; PubMed Central PMCID: PMC236720.

Martín J, Wharton SA, Lin YP, Takemoto DK, Skehel JJ, Wiley DC, Steinhauer DA. Studies of the binding properties of influenza hemagglutinin receptor-site mutants. Virology. 1998 Feb 1;241(1):101-11. PubMed PMID: 9454721.

Matrosovich, M., N. Zhou, Y. Kawaoka, and R. Webster. 1999. The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties. Journal of Virology 73(2):1146–1155

Matrosovich, M.N., T.Y. Matrosovich, T. Gray, N.A. Roberts, and H.D. Klenk. 2004. Human and avian influenza viruses target different cell types in cultures of human airway epithelium. Proceedings of the National Academy of Sciences U S A 101(13):4620–4624

Melo F, Feytmans E. Assessing protein structures with a non-local atomic interaction energy. J Mol Biol. 1998 Apr 17;277(5):1141-52. PubMed PMID: 9571028.

Muramoto Y, Noda T, Kawakami E, Akkina R, Kawaoka Y. Identification of novel influenza A virus proteins translated from PA mRNA. J Virol. 2013 Mar;87(5):2455-62. doi: 10.1128/JVI.02656-12. Epub 2012 Dec 12. PubMed PMID: 23236060; PubMed Central PMCID: PMC3571384.

O'Halloran D. A practical guide to phylogenetics for nonexperts. J Vis Exp. 2014 Feb 5;(84):e50975.

Peitsch, M. C. (1995) Protein modeling by E-mail Bio/Technology 13: 658-660.

Plotch SJ, Bouloy M, Krug RM. Transfer of 5'-terminal cap of globin mRNA to influenza viral complementary RNA during transcription in vitro. Proc Natl Acad Sci U S A. 1979 Apr;76(4):1618-22. PubMed PMID: 287003; PubMed Central PMCID: PMC383441.

Plotkin, J.B., and J. Dushoff. 2003. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of infl uenza A virus. Proceedings of the National Academy of Sciences U S A 100:7152–7157

Portela A, Digard P. The influenza virus nucleoprotein: a multifunctional RNA-binding protein pivotal to virus replication. J Gen Virol. 2002 Apr;83(Pt 4):723-34. PubMed PMID: 11907320.

Ravantti J, Bamford D, Stuart DI. Automatic comparison and classification of protein structures. J Struct Biol. 2013 Jul;183(1):47-56. doi: 10.1016/j.jsb.2013.05.007. Epub 2013 May 21. PubMed PMID: 23707633.

Richter S, Wenzel A, Stein M, Gabdoulline RR, Wade RC. webPIPSA: a web server for the comparison of protein interaction properties. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W276-80. doi: 10.1093/nar/gkn181. Epub 2008 Apr 17. PubMed PMID: 18420653; PubMed Central PMCID: PMC2447742.

Robb NC, Jackson D, Vreede FT, Fodor E. Splicing of influenza A virus NS1 mRNA is independent of the viral NS1 protein. J Gen Virol. 2010 Sep;91(Pt 9):2331-40. doi: 10.1099/vir.0.022004-0. Epub 2010 Jun 2. PubMed PMID: 20519456.

Rost B, Schneider R, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol. 1997;270:471–480. doi: 10.1006/jmbi.1997.1101.

Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999 Feb;12(2):85-94. PubMed PMID: 10195279.

Samji T. Influenza A: understanding the viral life cycle. Yale J Biol Med. 2009 Dec;82(4):153-9. Review. PubMed PMID: 20027280; PubMed Central PMCID: PMC2794490.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. natn. Acad. Sci. USA 74, 5463–5467.

Schwede T, Kopp J, Guex N, and Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research 31: 3381-3385.

Tong S, Li Y, Rivailler P, Conrardy C, Castillo DA, Chen LM, Recuenco S, Ellison JA, Davis CT, York IA, Turmelle AS, Moran D, Rogers S, Shi M, Tao Y, Weil MR, Tang K, Rowe LA, Sammons S, Xu X, Frace M, Lindblade KA, Cox NJ, Anderson LJ, Rupprecht CE, Donis RO. A distinct lineage of influenza A virus from bats. Proc Natl Acad Sci U S A. 2012 Mar 13;109(11):4269-74. doi: 10.1073/pnas.1116200109. Epub 2012 Feb 27. PubMed PMID: 22371588; PubMed Central PMCID: PMC3306675.

Tong S, Zhu X, Li Y, Shi M, Zhang J, Bourgeois M, Yang H, Chen X, Recuenco S, Gomez J, Chen LM, Johnson A, Tao Y, Dreyfus C, Yu W, McBride R, Carney PJ,Gilbert AT, Chang J, Guo Z, Davis CT, Paulson JC, Stevens J, Rupprecht CE, Holmes EC, Wilson IA, Donis RO. New world bats harbor diverse influenza A viruses. PLoS Pathog. 2013;9(10):e1003657. doi: 10.1371/journal.ppat.1003657. Epub 2013 Oct 10. PubMed PMID: 24130481; PubMed Central PMCID: PMC3794996.

Unni S, Huang Y, Hanson RM, Tobias M, Krishnan S, Li WW, Nielsen JE, Baker NA. Web servers and services for electrostatics calculations with APBS and PDB2PQR. J Comput Chem. 2011 May;32(7):1488-91. doi: 10.1002/jcc.21720. Epub 2011 Feb 1. PubMed PMID: 21425296; PubMed Central PMCID: PMC3062090.

Vandegrift KJ, Sokolow SH, Daszak P, Kilpatrick AM. Ecology of avian influenza viruses in a changing world. Ann N Y Acad Sci. 2010 May;1195:113-28. doi: 10.1111/j.1749-6632.2010.05451.x. Review. PubMed PMID: 20536820; PubMed Central PMCID: PMC2981064

Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature. 2006 Jan 19;439(7074):344-8. Epub 2005 Dec 4. PubMed PMID: 16327776; PubMed Central PMCID: PMC1569948.

Wade RC, Gabdoulline RR and De Rienzo F. Protein Interaction Property Similarity Analysis. Intl. J. Quant. Chem. 2001, 83: 122-127.

Wainright, P.O., M.L. Perdue, M. Brugh, and C.W. Beard. 1991. Amantadine resistance among hemagglutinin subtype 5 strains of avian influenza virus. Avian Diseases 35:31–39.

Wagner, R., M. Matrosovich, and H. D. Klenk. 2002. Functional balance between haemagglutinin and neuraminidase in influenza virus infections. Rev. Med. Virol. 12:159–166.

Webster, R.G., and W.G. Laver. 1980. Determination of the number of nonoverlapping antigenic areas on Hong Kong (H3N2) influenza virus hemagglutinin with monoclonal antibodies and the selection of variants with potential epidemiological significance. Virology 104:139–148.

Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of influenza A viruses. Microbiol Rev. 1992 Mar;56(1):152-79. Review. PubMed PMID: 1579108; PubMed Central PMCID: PMC372859.

Webster R, Peiris M, Chen H, et al. H5N1 outbreaks and enzootic influenza. Emerg. Infect. Dis.2006; 12:3–8. [PubMed: 16494709]

Weis W, Brown JH, Cusack S, Paulson JC, Skehel JJ, Wiley DC. Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. Nature. 1988 Jun 2;333(6172):426-31. PubMed PMID: 3374584.

Wise HM, Foeglein A, Sun J, Dalton RM, Patel S, Howard W, Anderson EC, Barclay WS, Digard P. A complicated message: Identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA. J Virol. 2009 Aug;83(16):8021-31. doi: 10.1128/JVI.00826-09. Epub 2009 Jun 3. PubMed PMID: 19494001; PubMed Central PMCID: PMC2715786.

Zamarin D, García-Sastre A, Xiao X, Wang R, Palese P. Influenza virus PB1-F2 protein induces cell death through mitochondrial ANT3 and VDAC1. PLoS Pathog. 2005 Sep;1(1):e4. Epub 2005 Sep 30. PubMed PMID: 16201016; PubMed Central PMCID: PMC1238739.

Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 2002 Nov;11(11):2714-26. Erratum in: Protein Sci. 2003 Sep;12(9):2121. PubMed PMID: 12381853; PubMed Central PMCID: PMC2373736.

SUPPLEMENTARY MATERIAL CHAPTER 1

Table S1. List of samples included in the analysis. For each samples the available sequences are indicated; the number of mapped reads and the HA coverage are showed for those samples sequenced using NGS.

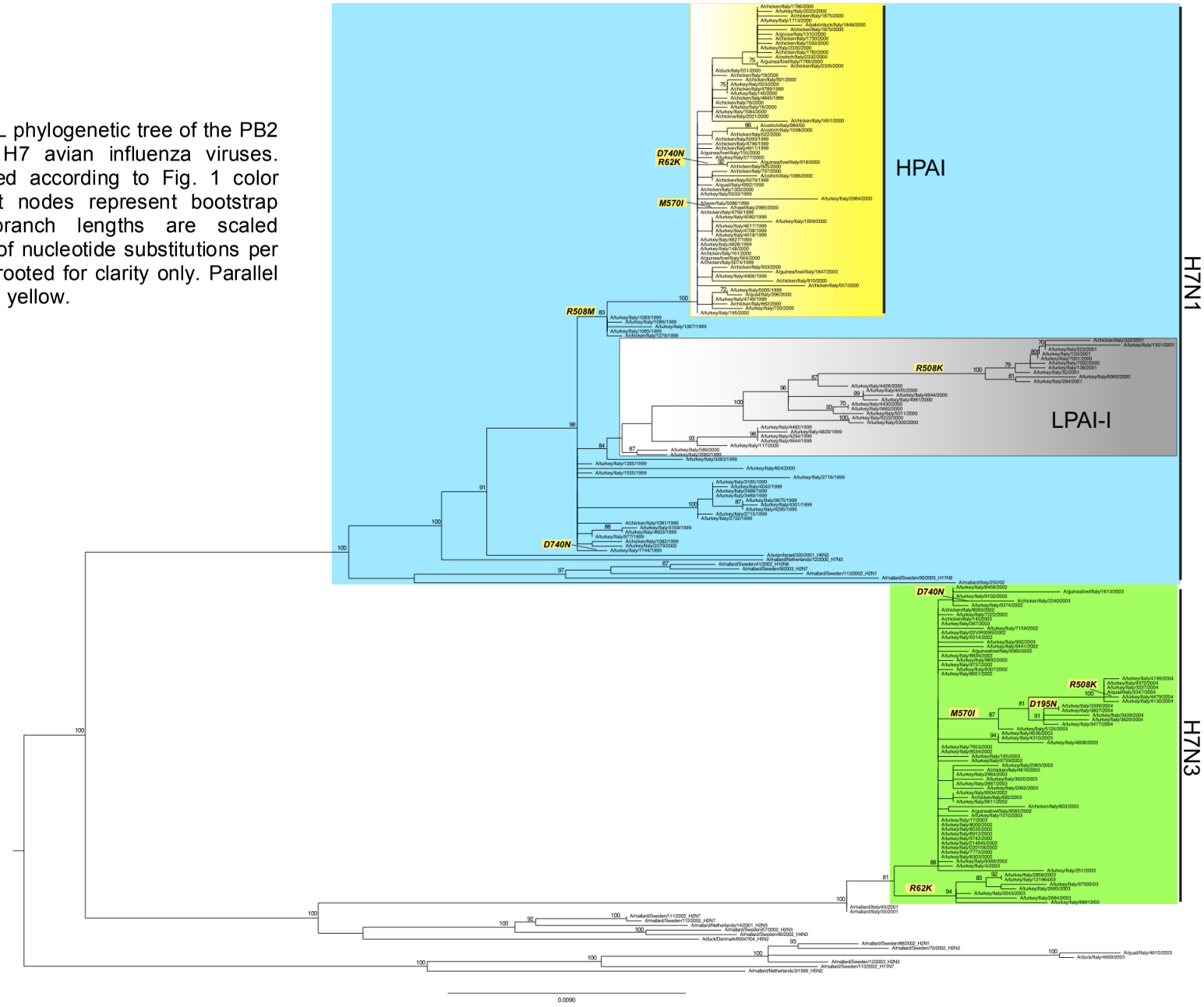
Virus	Collection date	Province	Region	Available sequences	Comments	Mapped reads	HA mean coverage
A/turkey/Italy/7159/2002	16/10/2002	BS	Lombardia	Complete genome	This study (EPI243276; EPI543954-EPI543960)		
A/turkey/Italy/7653/2002	29/10/2002	VR	Veneto	Complete genome	This study (EPI243277; EPI543961-EPI543967)		
A/turkey/Italy/8307/2002	07/11/2002	VR	Veneto	Complete genome	This study (EPI154967; EPI544023-EPI544029)		
A/chicken/Italy/8093/2002	09/11/2002	VR	Veneto	Complete genome	This study (EPI154966; EPI543982-EPI543988)	519352	5020
A/turkey/Italy/8651/2002	20/11/2002	VR	Veneto	Complete genome	This study (EPI154969; EPI544030-EPI544036)		
A/turkey/Italy/8834/2002	25/11/2002	VR	Veneto	Complete genome	This study (EPI154970; EPI544037-EPI544043)		
A/turkey/Italy/9102/2002	02/12/2002	VR	Veneto	Complete genome	This study (EPI154971; EPI544044-EPI544050)		
A/turkey/Italy/9314/2002	09/12/2002	VR	Veneto	Complete genome	This study (EPI154972; EPI544051-EPI544057)		
A/turkey/Italy/9369/2002	10/12/2002	VR	Veneto	Complete genome	This study (EPI154973; EPI544058-EPI544064)	450876	1373
A/turkey/Italy/9374/2002	11/12/2002	VR	Veneto	Complete genome	This study (EPI154974; EPI543968-EPI543974)		
A/guineafowl/Italy/1613/2003	12/03/2003	PD	Veneto	All genes except NP	This study (EPI154959; EPI543989-EPI543994)		
A/chicken/Italy/2240/2003	16/04/2003	VR	Veneto	Complete genome	This study (EPI154960; EPI543975-EPI543981)		
A/turkey/Italy/2963/2003	23/05/2003	VR	Veneto	Complete genome	This study (EPI243279; EPI543940-EPI543946)		
A/turkey/Italy/2856/2003	26/05/2003	MN	Lombardia	Complete genome	This study (EPI543996-EPI544002)		
A/turkey/Italy/4036/2003	17/07/2003	VR	Veneto	Complete genome	This study (EPI154963; EPI544016-EPI544022)		
A/turkey/Italy/3399/2004	22/09/2004	VR	Veneto	Complete genome	This study (EPI154961; EPI544003-EPI544009)		
A/turkey/Italy/3439/2004	22/09/2004	VR	Veneto	All genes except M	This study (EPI154962; EPI544010-EPI544015)		
A/turkey/Italy/4199/2004	05/11/2004	VR	Veneto	Complete genome	This study (EPI243280; EPI543947-EPI543953)		
A/turkey/Italy/7222/2002	18/10/2002	BS	Lombardia	Complete genome	This study (EPI543853-EPI543859)		
A/turkey/Italy/7773/2002	31/10/2002	VR	Veneto	Complete genome	This study (EPI543861-EPI543867)	868287	4855
A/turkey/Italy/8303/2002	13/11/2002	VI	Veneto	Complete genome	This study (EPI543869-EPI543875)	523301	228
A/ostrich/Italy/8856/2002	26/11/2002	BG	Lombardia	NP, HA, M, NS, NA	This study (EPI543792-EPI543795)		
A/turkey/Italy/02vir9289/2002	05/12/2002	VR	Veneto	Complete genome	This study (EPI543933-EPI543939)	320556	51483
A/guineafowl/Italy/9360/2002	11/12/2002	BS	Lombardia	Complete genome	This study (EPI543917-EPI543923)		

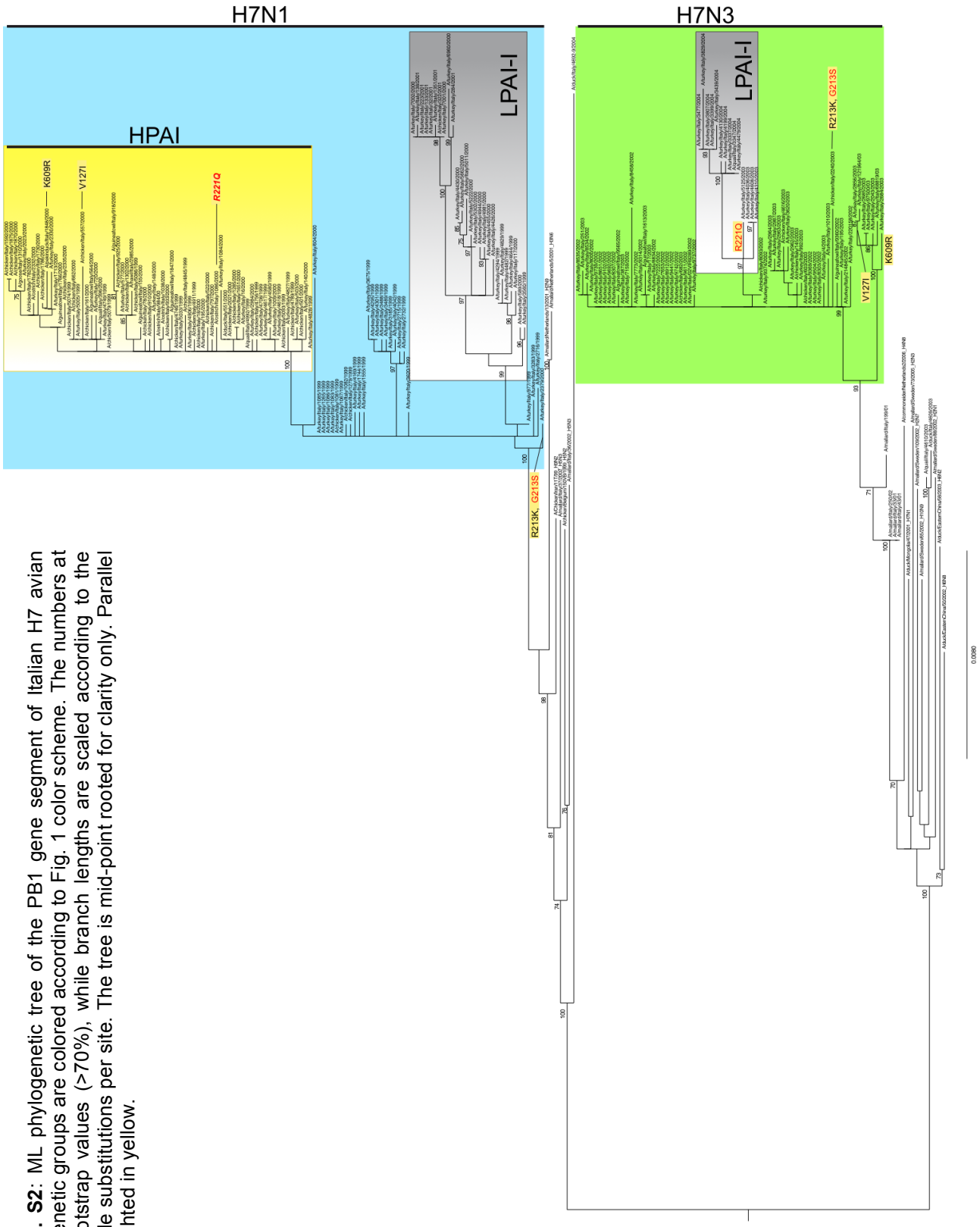
Virus	Collection date	Province	Region	Available sequences	Comments	Mapped reads	HA mean coverage
A/turkey/Italy/9441/2002	12/12/2002	BO	Emilia Romagna	Complete genome	This study (EPI543877-EPI543883)		
A/turkey/Italy/9504/2002	17/12/2002	PD	Veneto	Complete genome	This study (EPI543885-EPI543891)		
A/guineafowl/Italy/9565/2002	19/12/2002	VR	Veneto	Complete genome	This study (EPI543925-EPI543931)	466527	1428
A/turkey/Italy/9611/2002	20/12/2002	PD	Veneto	Complete genome	This study (EPI543893-EPI543899)	543221	2609
A/turkey/Italy/9692/2002	24/12/2002	VR	Veneto	Complete genome	This study (EPI543901-EPI543907)		
A/turkey/Italy/4/2003	03/01/2003	VR	Veneto	Complete genome	This study (EPI543829-EPI543835)	711224	4974
A/turkey/Italy/17/2003	07/01/2003	PD	Veneto	Complete genome	This study (EPI543805-EPI543811)	519973	1516
A/chicken/Italy/145/2003	09/01/2003	PD	Veneto	Complete genome	This study (EPI543770-EPI543776)		
A/turkey/Italy/195/2003	09/01/2003	VR	Veneto	Complete genome	This study (EPI543797-EPI543803)		
A/turkey/Italy/387/2003	17/01/2003	VR	Veneto	Complete genome	This study (EPI543821-EPI543827)		
A/chicken/Italy/603/2003	24/01/2003	BS	Lombardia	Complete genome	This study (EPI543784-EPI543790)		
A/turkey/Italy/992/2003	11/02/2003	PD	Veneto	Complete genome	This study (EPI543909-EPI543915)		
A/turkey/Italy/2964/2003	27/05/2003	VR	Veneto	Complete genome	This study (EPI543813-EPI543819)		
A/turkey/Italy/4310/2003	28/07/2003	VR	Veneto	Complete genome	This study (EPI543837-EPI543843)		
A/chicken/Italy/4917/2003	28/08/2003	VR	Veneto	PB1, HA, NP, NA, NS, M	This study (EPI543778-EPI543782)		
A/turkey/Italy/5125/2003	09/09/2003	VR	Veneto	Complete genome	This study (EPI543845-EPI543851)		
A/turkey/Italy/4042/2004	27/10/2004	VR	Veneto	HA	This study (EPI154964)		
A/turkey/Italy/214845/2002	15/10/2002	na	na	Complete genome	Campitelli et al. 2004		
A/turkey/Italy/220158/2002	15/10/2002	na	na	Complete genome	Campitelli et al. 2004		
A/turkey/Italy/8000/2002	06/11/2002	MN	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/8534/2002	13/11/2002	MN	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/8535/2002	13/11/2002	BS	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/8458/2002	14/11/2002	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/8912/2002	26/11/2002	VI	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/chicken/Italy/270638/02	15/12/2002	na	na	HA	Campitelli et al. 2008		
A/Guineafowl/Italy/266184/02	15/12/2002	na	na	HA	Campitelli et al. 2008		

Virus	Collection date	Province	Region	Available sequences	Comments	Mapped reads	HA mean coverage
A/turkey/Italy/9739/2002	23/12/2002	BS	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/9742/2002	23/12/2002	BS	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/9737/2002	30/12/2002	MN	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/251/2003	14/01/2003	BS	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/chicken/Italy/682/2003	21/01/2003	PD	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/1010/2003	10/02/2003	MN	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/68819/03	15/03/2003	na	na	Complete genome	Campitelli et al. 2008		
A/turkey/Italy/2043/2003	24/03/2003	BG	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/97500/03	15/04/2003	na	na	All genes except PB1	Campitelli et al. 2008		
A/turkey/Italy/121964/03	15/05/2003	na	na	Complete genome	Campitelli et al. 2008		
A/turkey/Italy/2684/2003	15/05/2003	CR	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/2685/2003	15/05/2003	BS	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/2962/2003	04/06/2003	VR	Veneto	Complete genome	Capua et al., 2013		
A/turkey/Italy/2987/2003	04/06/2003	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/3620/2003	01/07/2003	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/duck/Italy/4609/2003(H7N2)	06/08/2003	MN	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/quail/Italy/4610/2003(H7N2)	06/08/2003	BG	Lombardia	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/4608/2003	08/08/2003	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/chicken/Italy/4616/2003	11/08/2003	NO	Piemonte	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/3337/2004	20/09/2004	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/quail/Italy/3347/2004	21/09/2004	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/3477/2004	29/09/2004	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/3807/2004	13/10/2004	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/3829/2004	13/10/2004	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		

Virus	Collection date	Province	Region	Available sequences	Comments	Mapped reads	HA mean coverage
A/turkey/Italy/4130/2004	02/11/2004	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/4372/2004	18/11/2004	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		
A/turkey/Italy/4479/2004	23/11/2004	VR	Veneto	Complete genome	The NIAID Influenza Genome Sequencing Project		

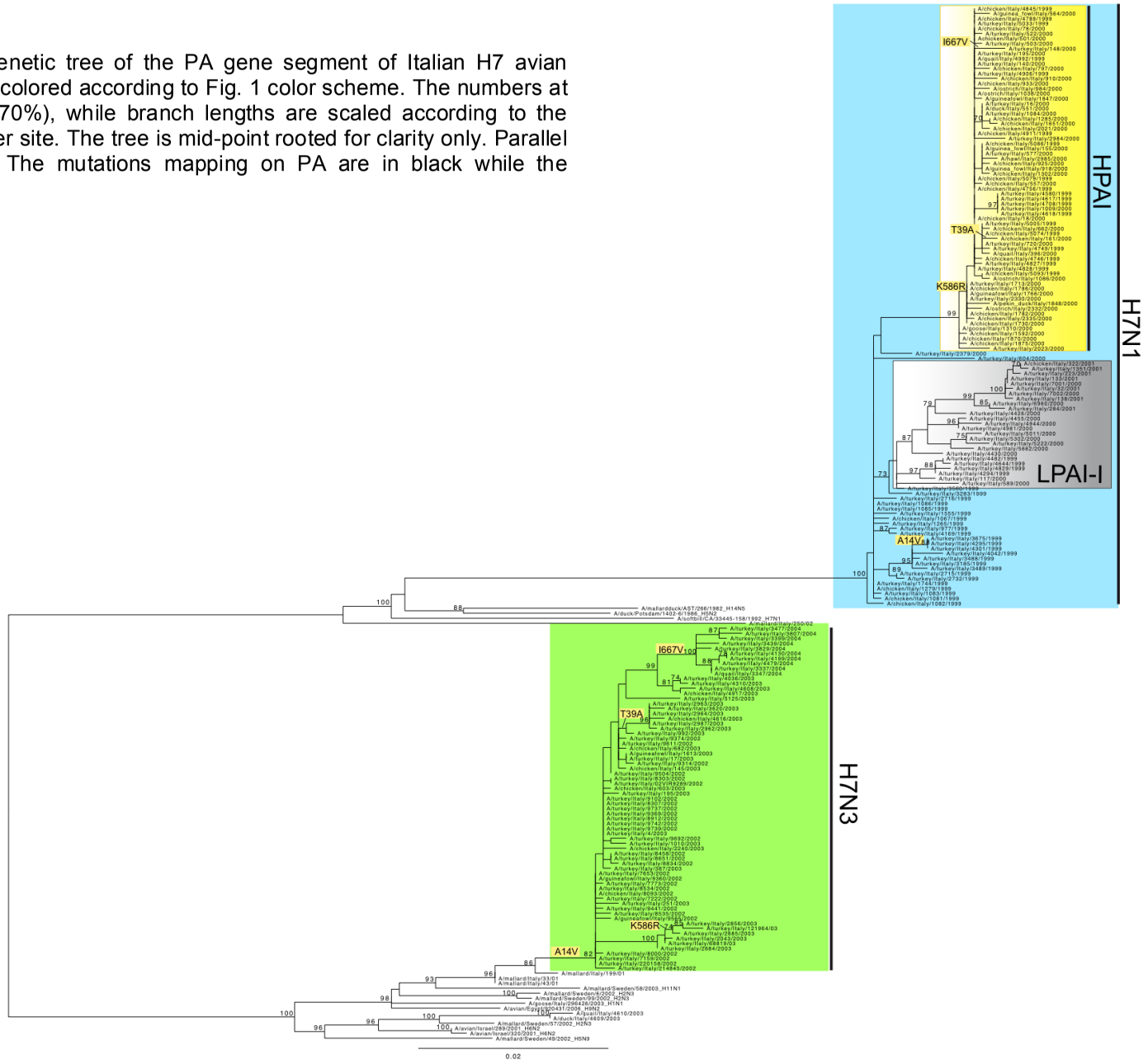
Supplementare Fig.S1: ML phylogenetic tree of the PB2 gene segment of Italian H7 avian influenza viruses. Genetic groups are colored according to Fig. 1 color scheme. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only. Parallel mutations are highlighted in yellow.



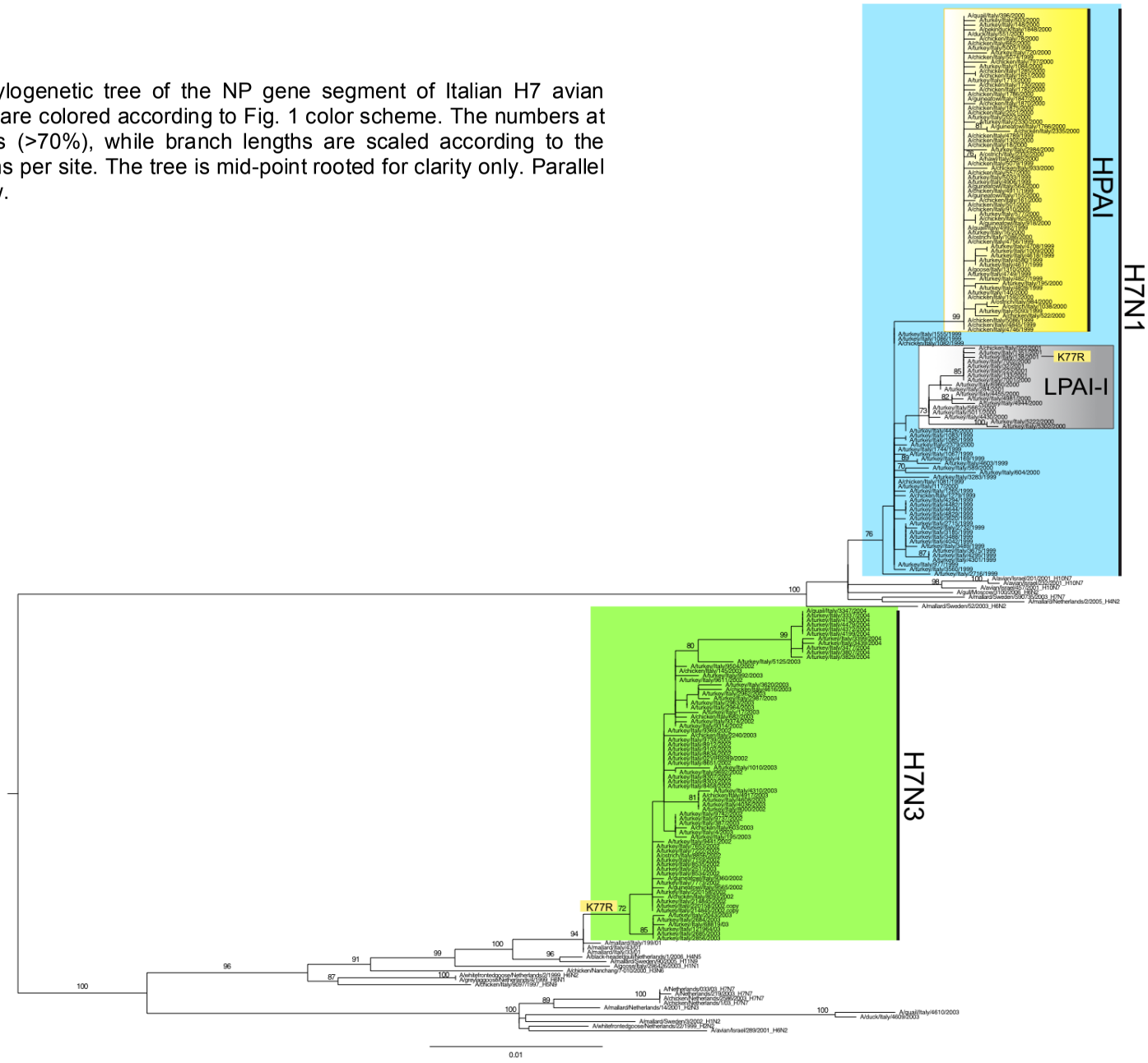


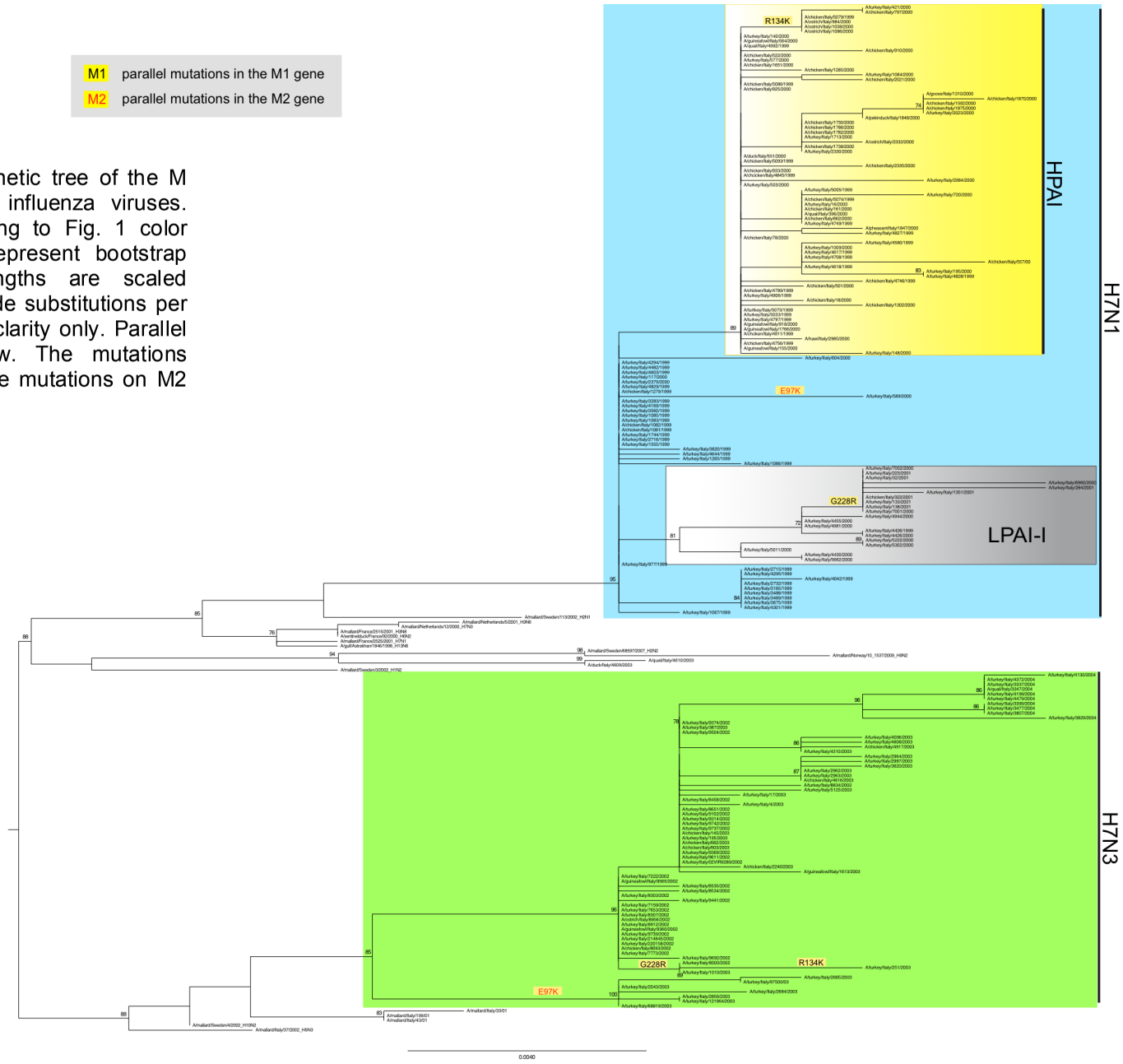
Supplementary Fig. S2: ML phylogenetic tree of the PB1 gene segment of Italian H7 avian influenza viruses. Genetic groups are colored according to Fig. 1 color scheme. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only. Parallel mutations are highlighted in yellow.

Supplementary Fig. S3: ML phylogenetic tree of the PA gene segment of Italian H7 avian influenza viruses. Genetic groups are colored according to Fig. 1 color scheme. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only. Parallel mutations are highlighted in yellow. The mutations mapping on PA are in black while the mutations on PA-X are colored in red.



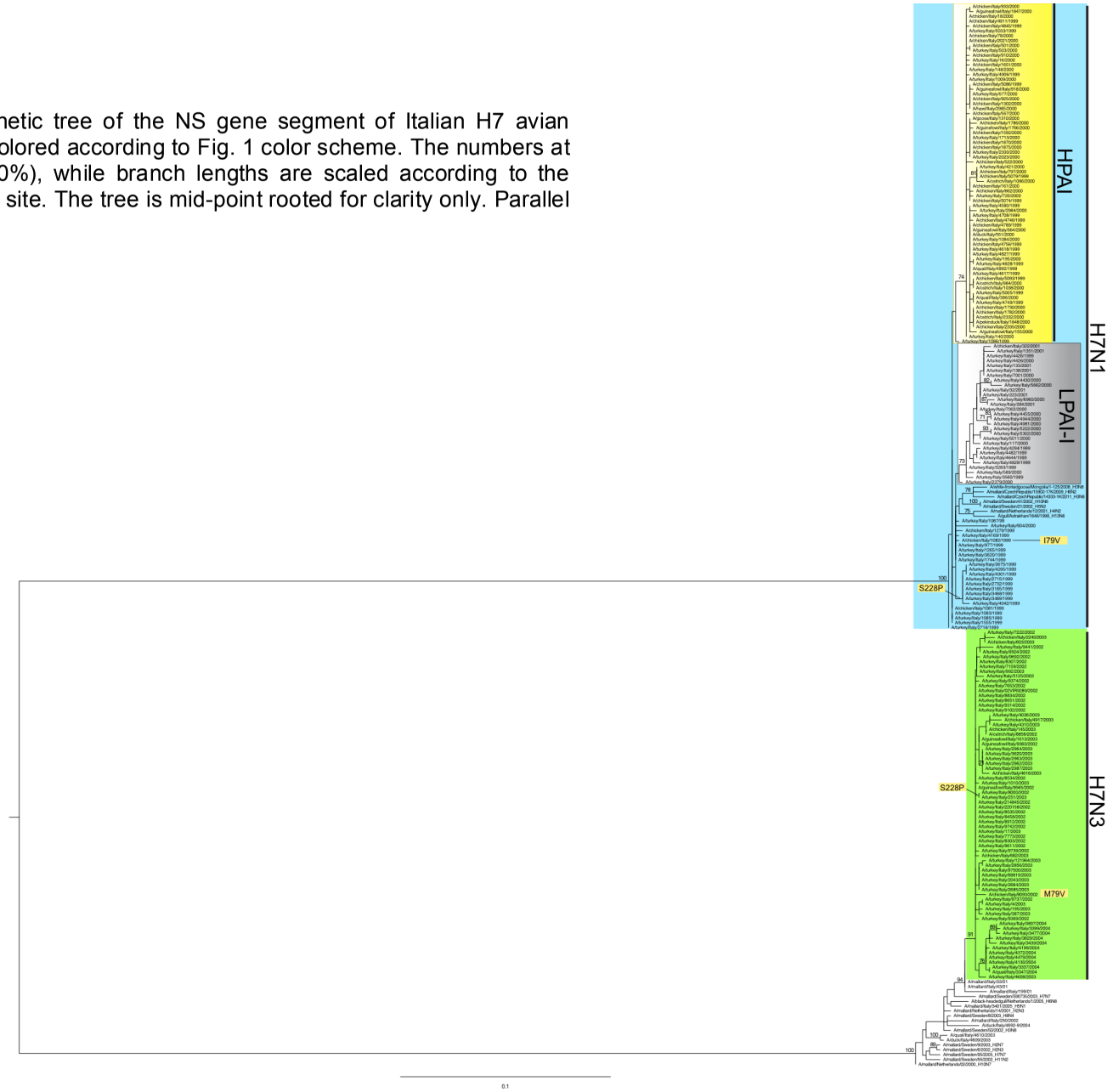
Supplementary Fig. S4: ML phylogenetic tree of the NP gene segment of Italian H7 avian influenza viruses. Genetic groups are colored according to Fig. 1 color scheme. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only. Parallel mutations are highlighted in yellow.



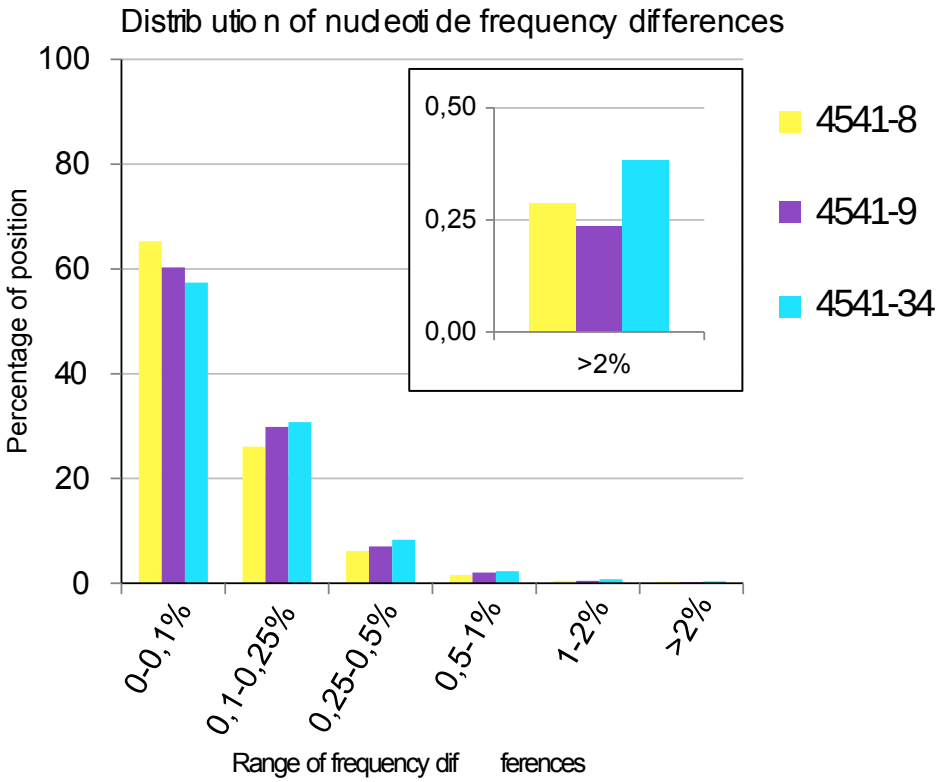


Supplementare Fig.S5: ML phylogenetic tree of the M gene segment of Italian H7 avian influenza viruses. Genetic groups are colored according to Fig. 1 color scheme. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only. Parallel mutations are highlighted in yellow. The mutations mapping on M1 are in black while the mutations on M2 are colored in red.

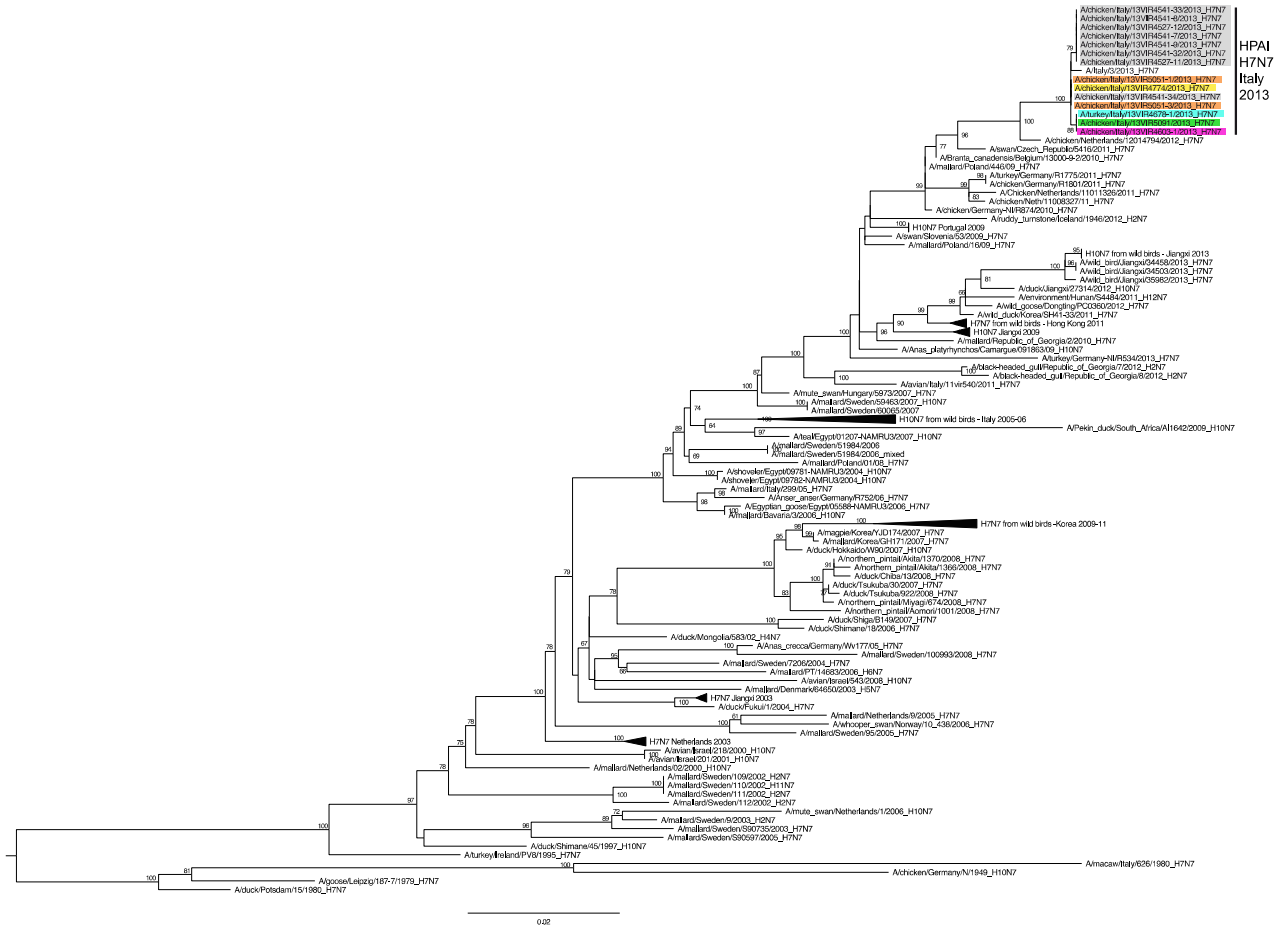
Supplementary Fig. S6: ML phylogenetic tree of the NS gene segment of Italian H7 avian influenza viruses. Genetic groups are colored according to Fig. 1 color scheme. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only. Parallel mutations are highlighted in yellow.



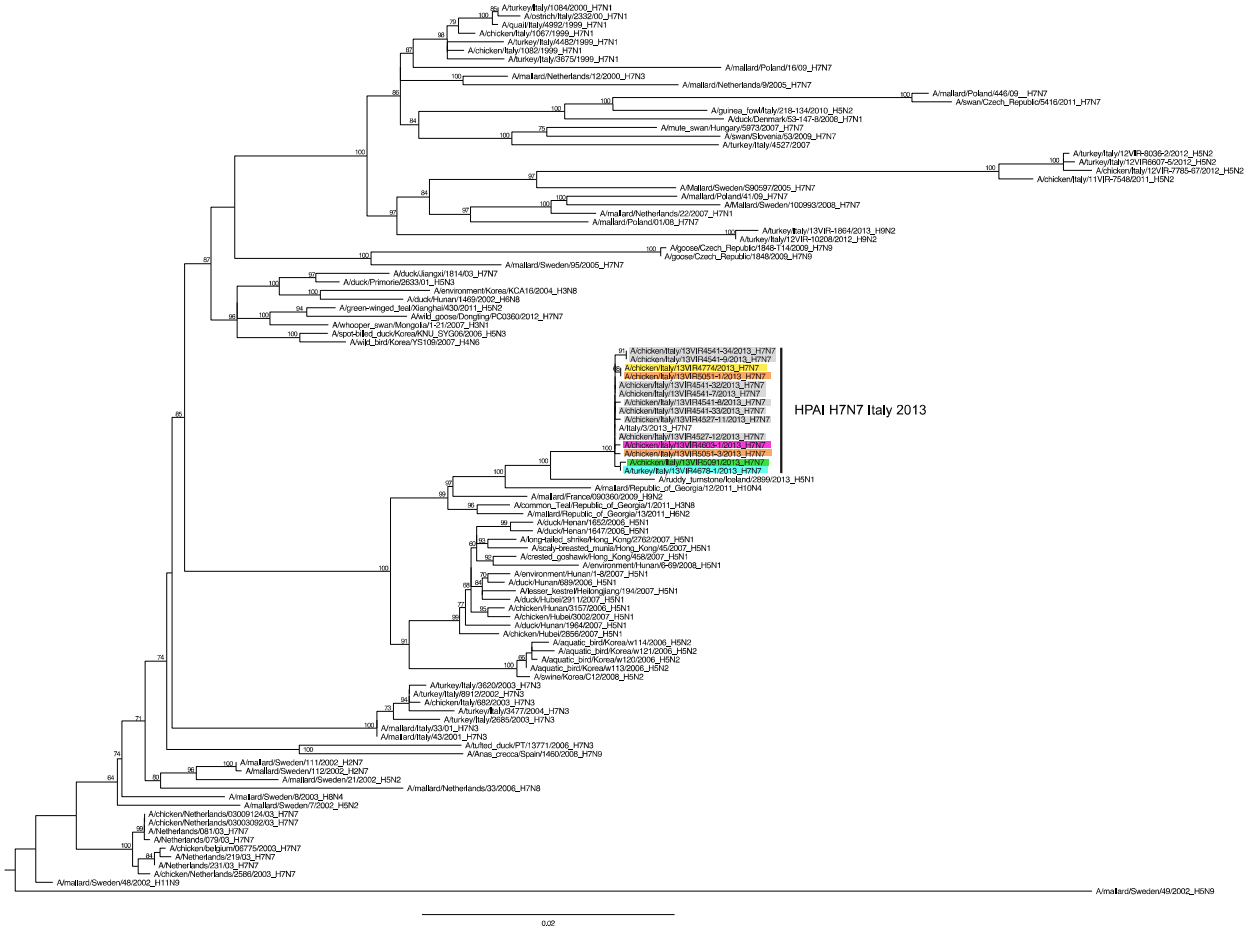
SUPPLEMENTARY MATERIAL CHAPTER 2



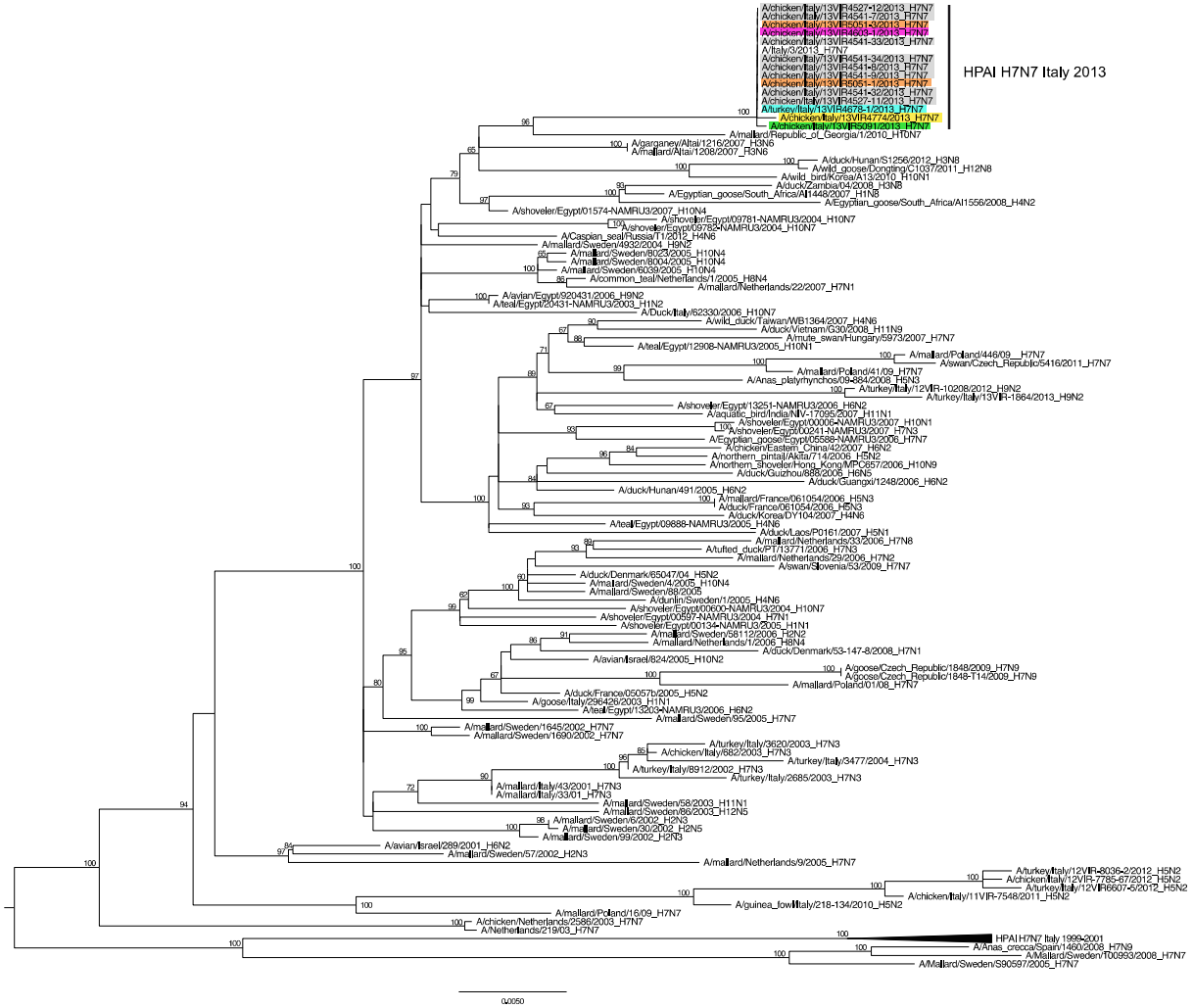
S1 Fig. Distribution of nucleotide frequency differences between three technical replicates. . For each genome position with a coverage >500 the frequency differences between the four bases (A, C, T and G) were obtained from the comparison of the replicates of the threes samples: 4541-8 in yellow, 4541-9 in violet, 4541-34 in blue. The y-axis represents the percentage of nucleotide positions where the highest frequency differences fall within the ranges 0-0.1%, 0.1-0.25%, 0.25-0.5%, 0.5-1%, 1-2% and >2% (x-axis). Frequency differences higher than 2% were observed in only 0.3%-0.4% of all the analysed positions (11501 to 13308) for all the replicates. Thus a 2% threshold allows the exclusion of 99.6% of the possible errors.



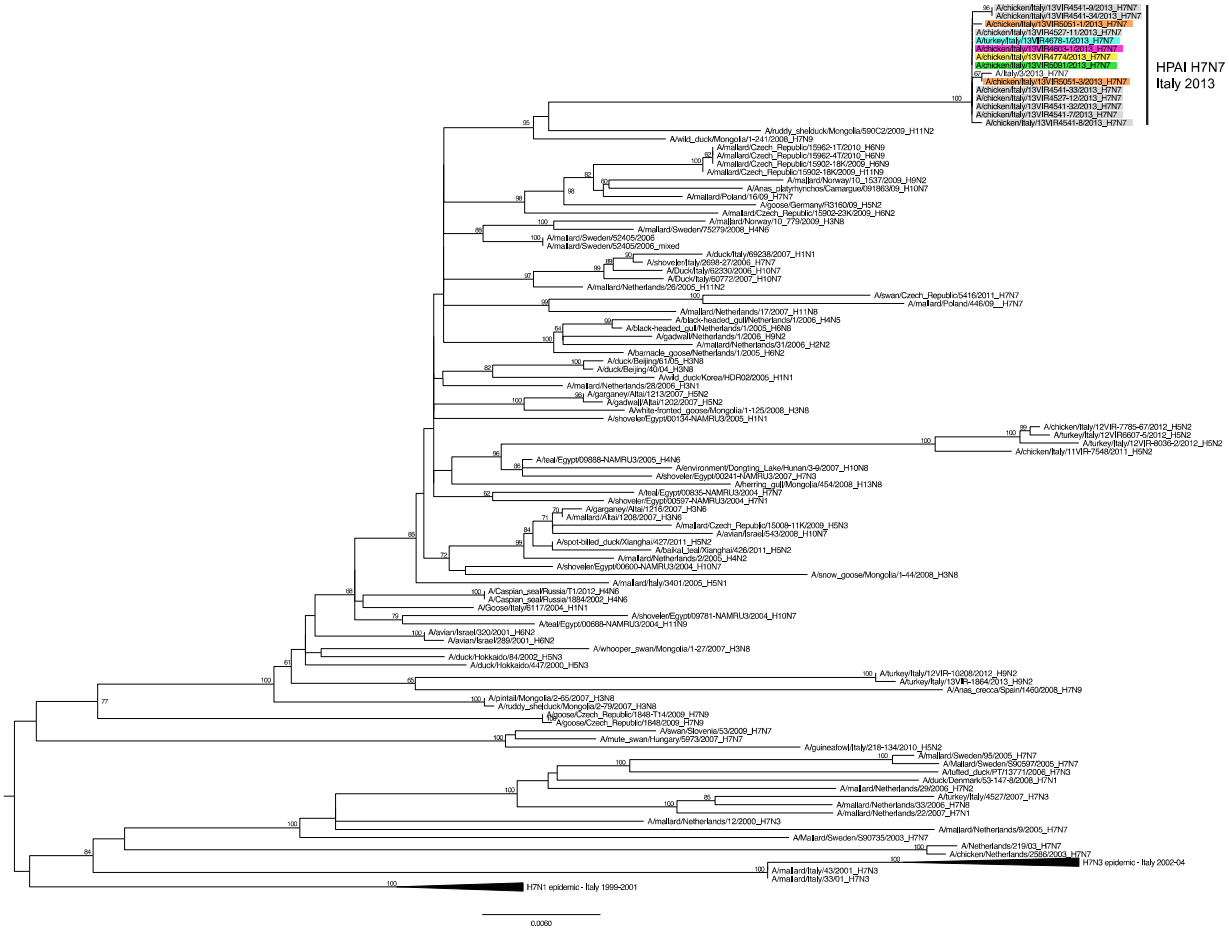
S2 Fig. ML phylogenetic tree of the NA gene segment of H7 avian influenza viruses. HPAI H7N7 viruses collected during Italian epidemic are coloured according to the farm of collection: grey for farm 1, purple for farm 2, light blue for farm 3, yellow for farm 4, green for farm 5 and orange for farm 6. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only.



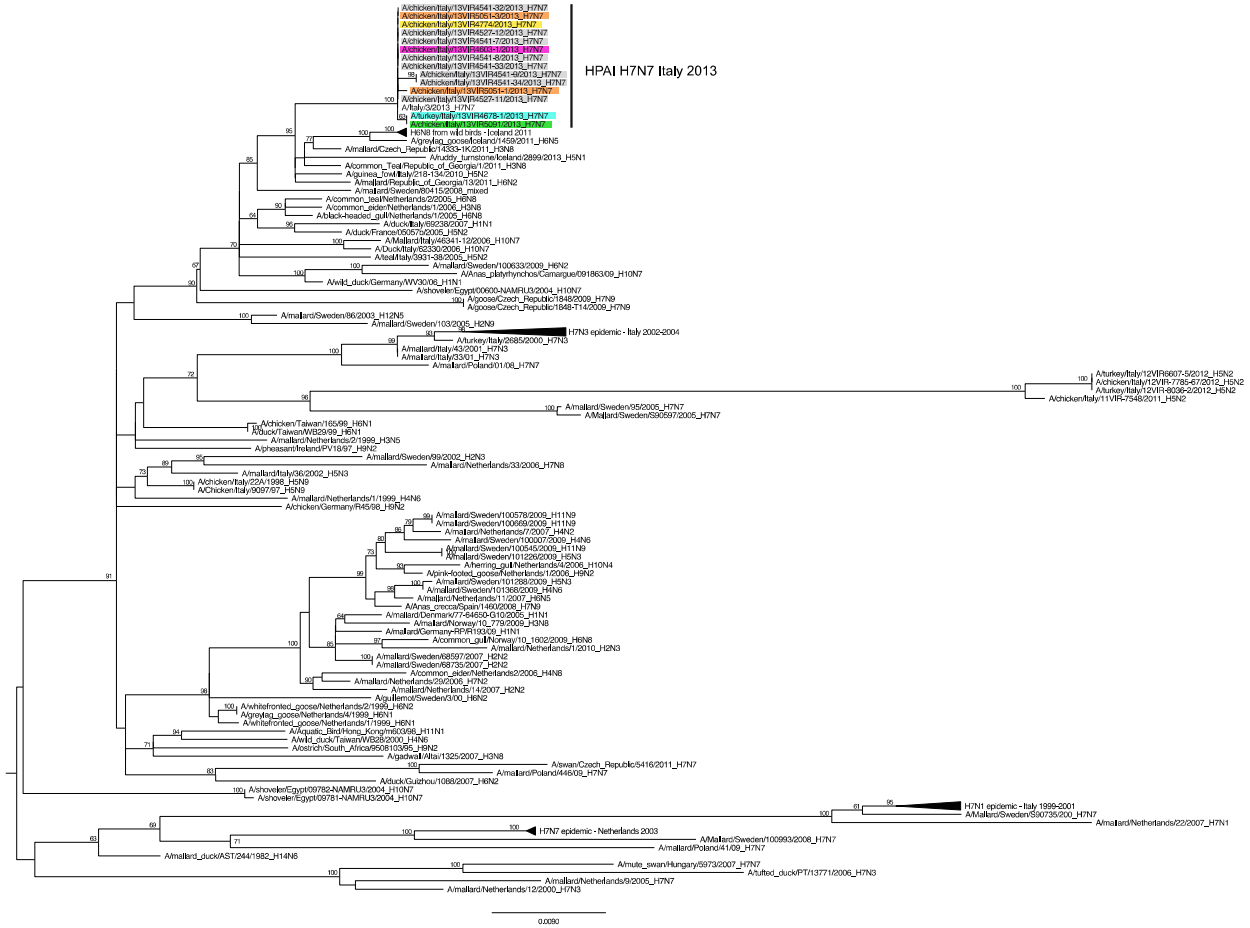
S3 Fig. ML phylogenetic tree of the PB2 gene segment of H7 avian influenza viruses. HPAI H7N7 viruses collected during Italian epidemic are coloured according to the farm of collection: grey for farm 1, purple for farm 2, light blue for farm 3, yellow for farm 4, green for farm 5 and orange for farm 6. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only.



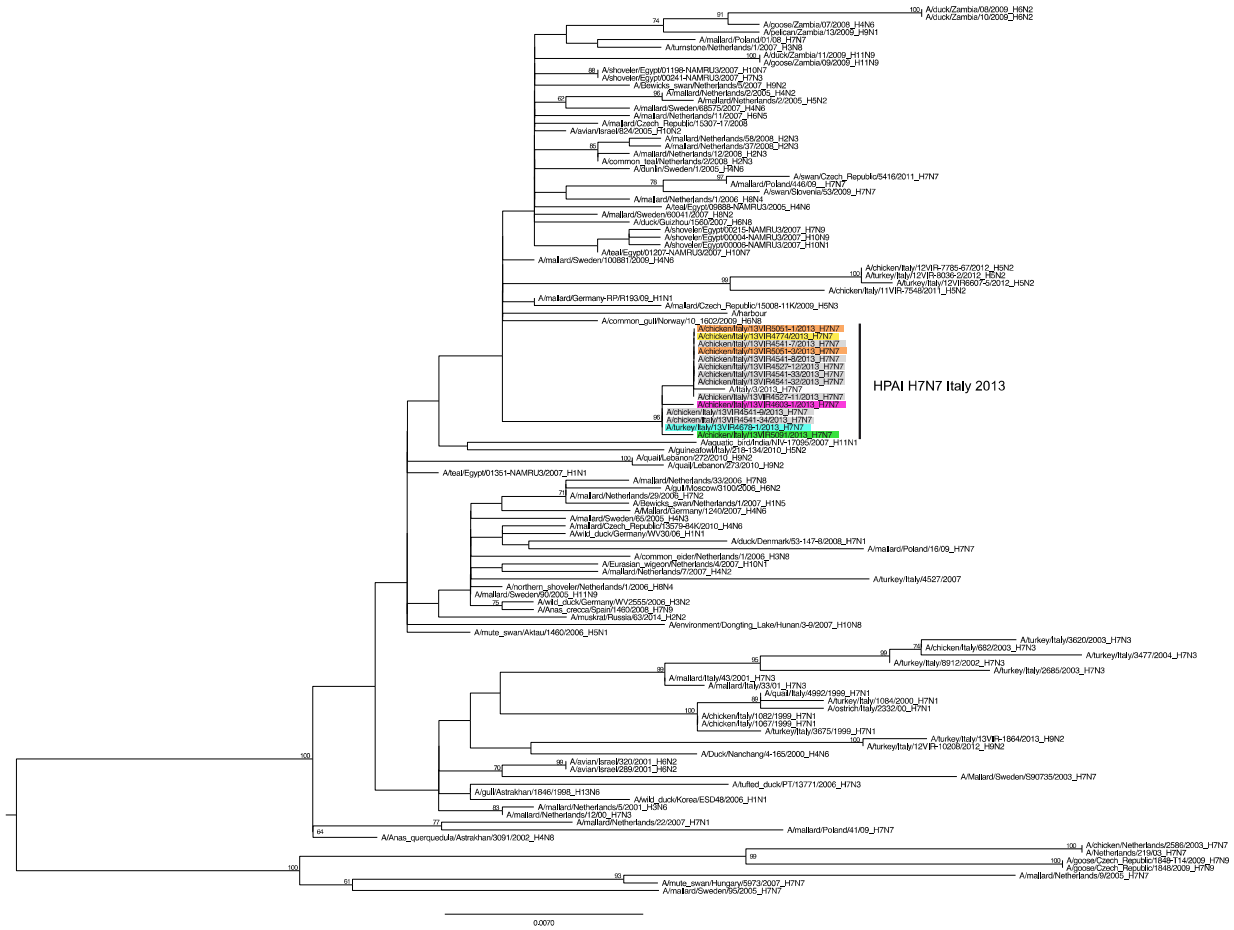
S4 Fig. ML phylogenetic tree of the PB1 gene segment of H7 avian influenza viruses. HPAI H7N7 viruses collected during Italian epidemic are coloured according to the farm of collection: grey for farm 1, purple for farm 2, light blue for farm 3, yellow for farm 4, green for farm 5 and orange for farm 6. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only.



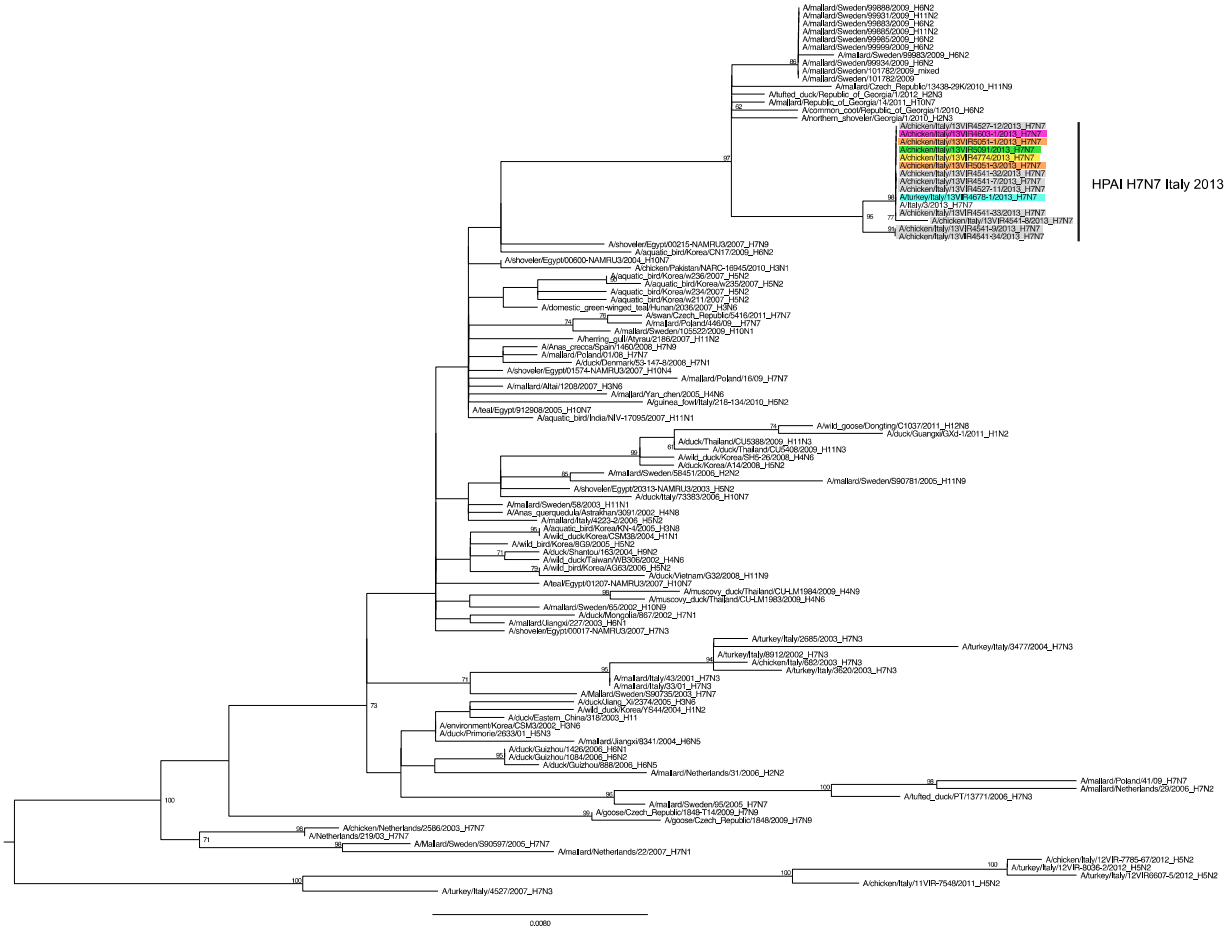
S5 Fig. ML phylogenetic tree of the PA gene segment of H7 avian influenza viruses. HPAI H7N7 viruses collected during Italian epidemic are coloured according to the farm of collection: grey for farm 1, purple for farm 2, light blue for farm 3, yellow for farm 4, green for farm 5 and orange for farm 6. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only.



S6 Fig. ML phylogenetic tree of the NP gene segment of H7 avian influenza viruses. HPAI H7N7 viruses collected during Italian epidemic are coloured according to the farm of collection: grey for farm 1, purple for farm 2, light blue for farm 3, yellow for farm 4, green for farm 5 and orange for farm 6. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only.

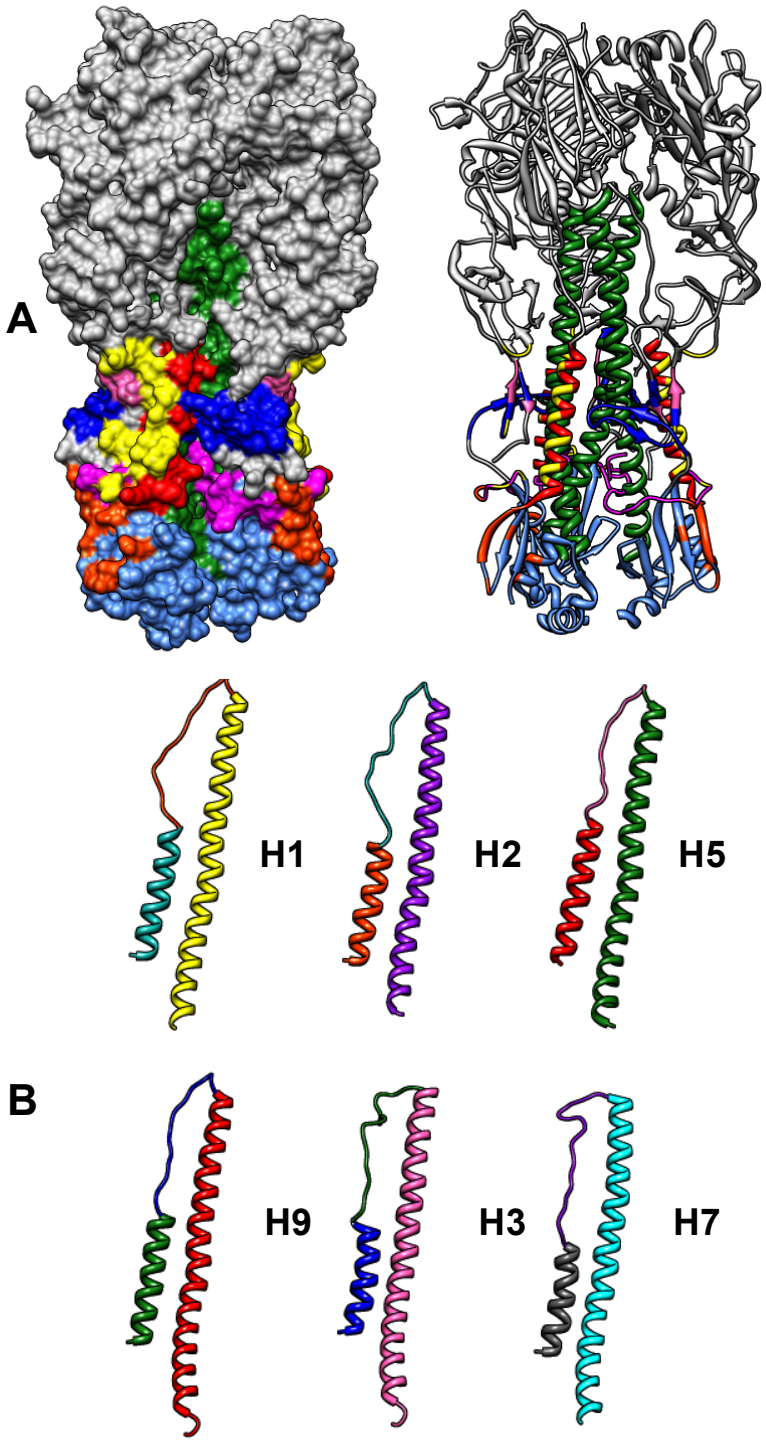


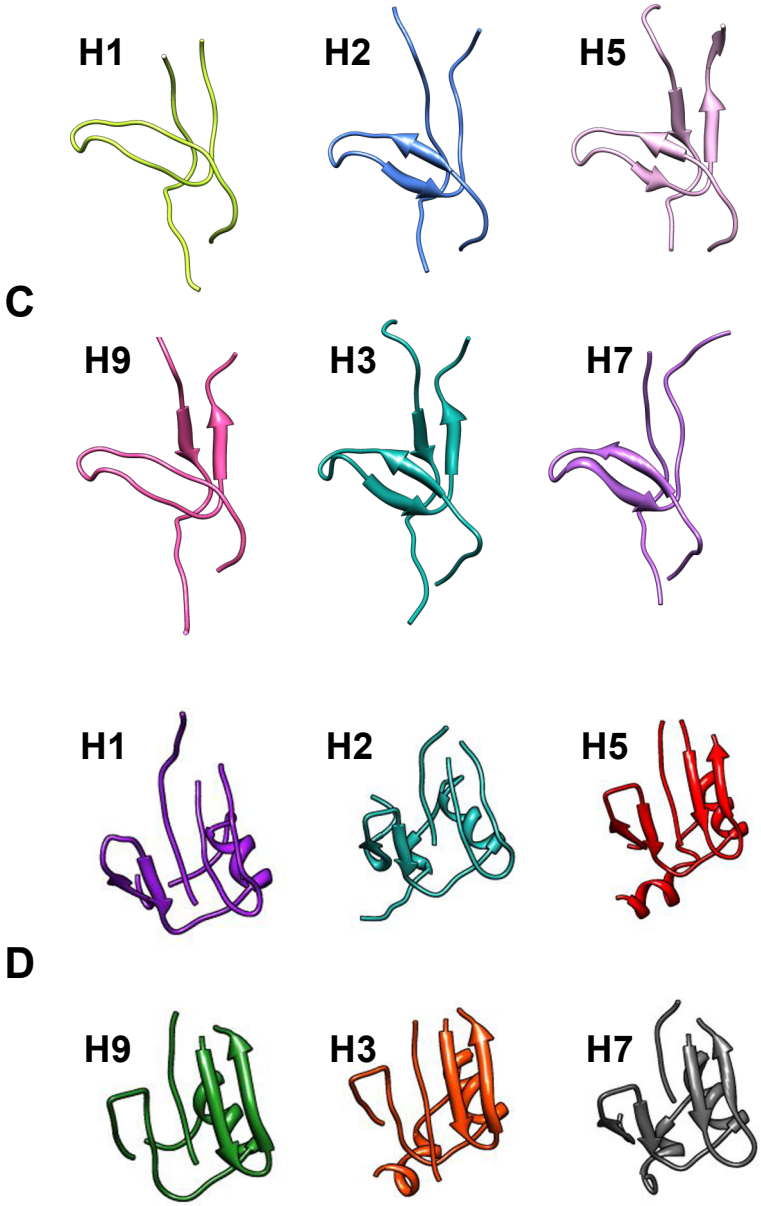
S7 Fig. ML phylogenetic tree of the M gene segment of H7 avian influenza viruses. HPAI H7N7 viruses collected during Italian epidemic are coloured according to the farm of collection: grey for farm 1, purple for farm 2, light blue for farm 3, yellow for farm 4, green for farm 5 and orange for farm 6. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only.



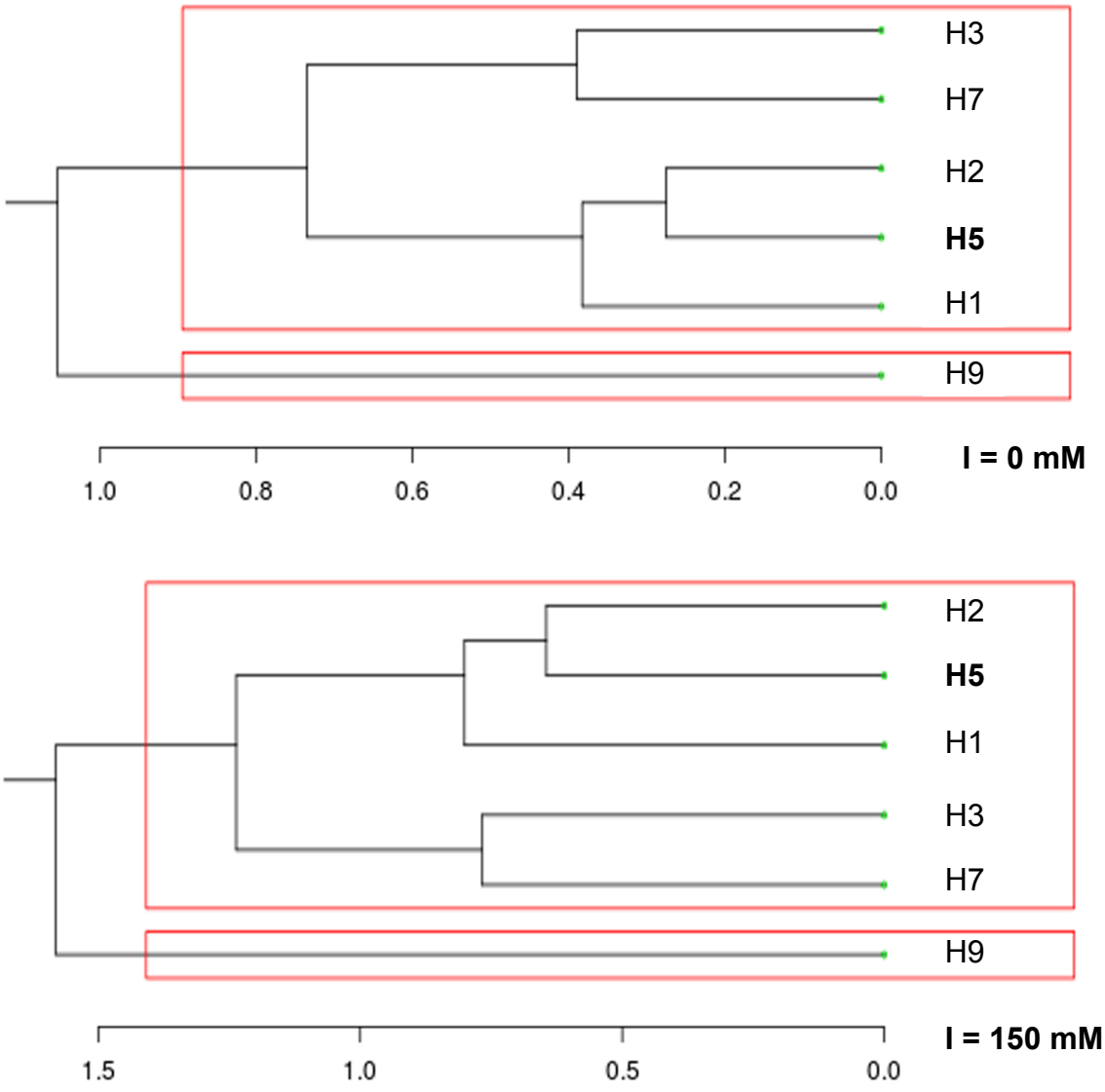
S8 Fig. ML phylogenetic tree of the NS gene segment of H7 avian influenza viruses. HPAI H7N7 viruses collected during Italian epidemic are coloured according to the farm of collection: grey for farm 1, purple for farm 2, light blue for farm 3, yellow for farm 4, green for farm 5 and orange for farm 6. The numbers at nodes represent bootstrap values (>70%), while branch lengths are scaled according to the numbers of nucleotide substitutions per site. The tree is mid-point rooted for clarity only.

SUPPLEMENTARY MATERIAL CHAPTER 4

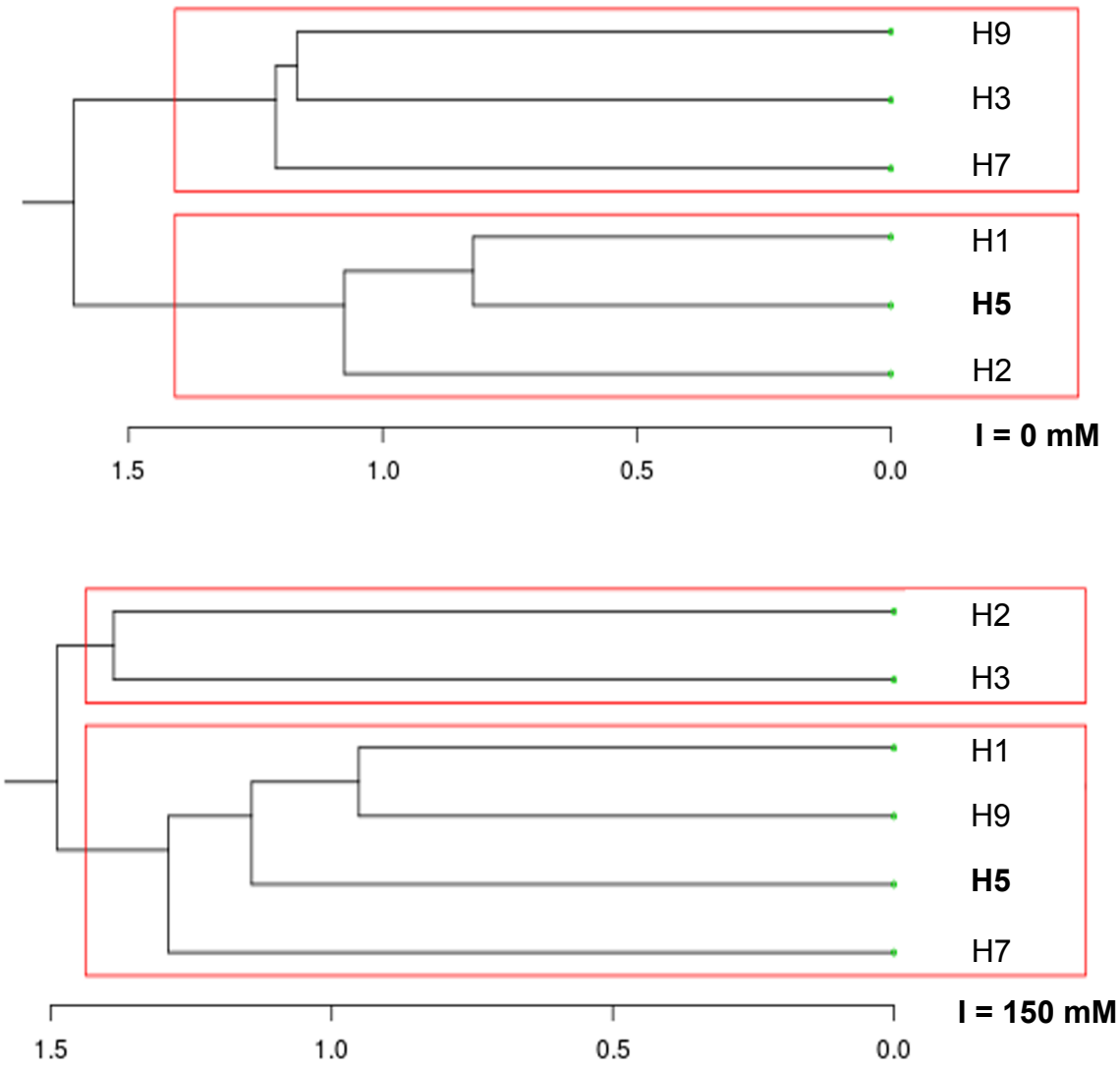

A
H1
H2
H5
B
H9
H3
H7



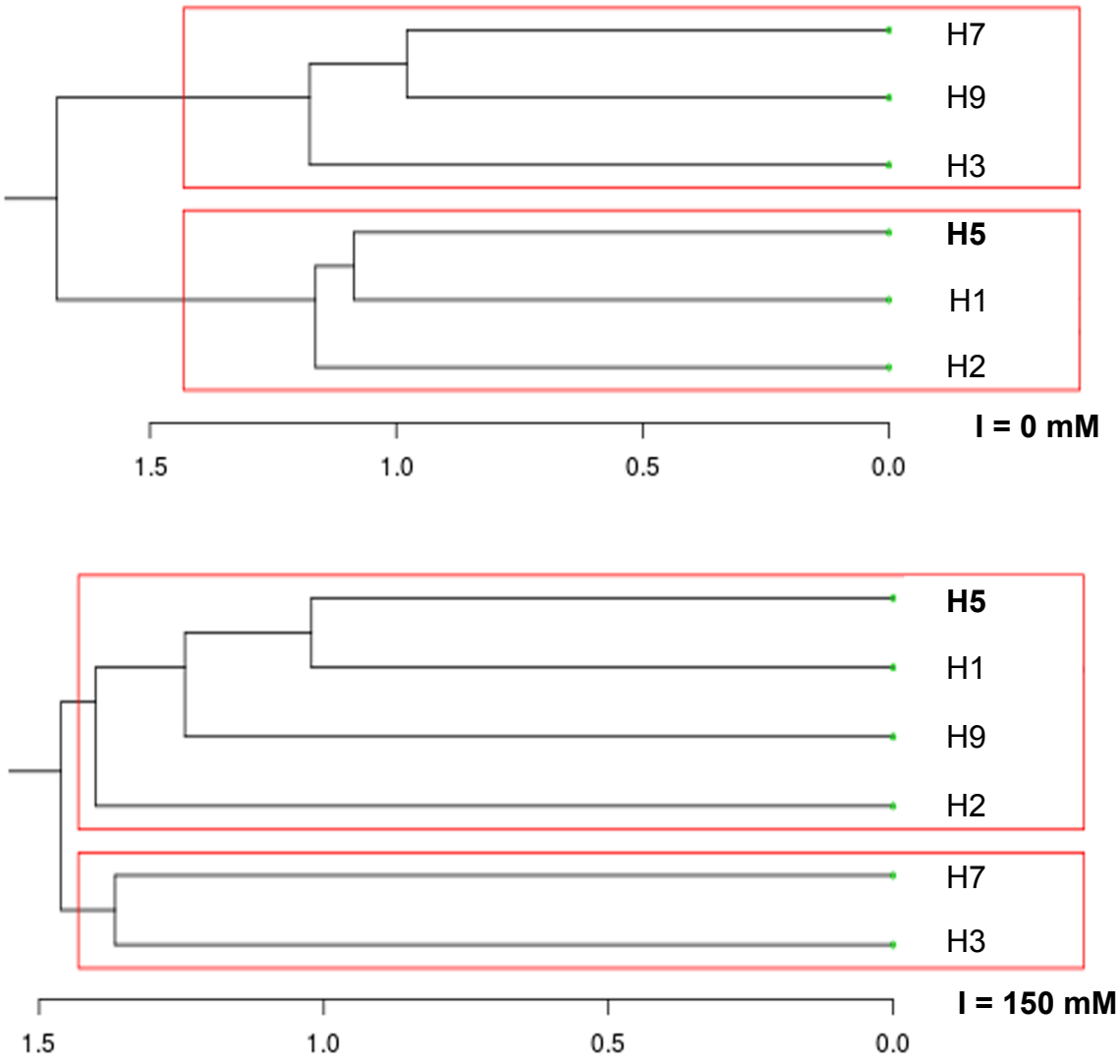
HA2 stem epitopes. Panel A: surface (left) and cartoon (right) representations of H5 trimers. Color code: RBD and VED, gray; A helix, red; C-D helices, green; fusion peptide, magenta; VED-proximal β region, blue; VED-distal β region, pale blue. Epitopes recognised by antibodies CR6261 and CR8020 are highlighted in yellow and orange, respectively. Panel B: comparison of the A-C-D α helices and B loop regions from the six available HA structures. Panels C and D focus, within the six available structures, on VED-proximal and VED-distal β regions, respectively.



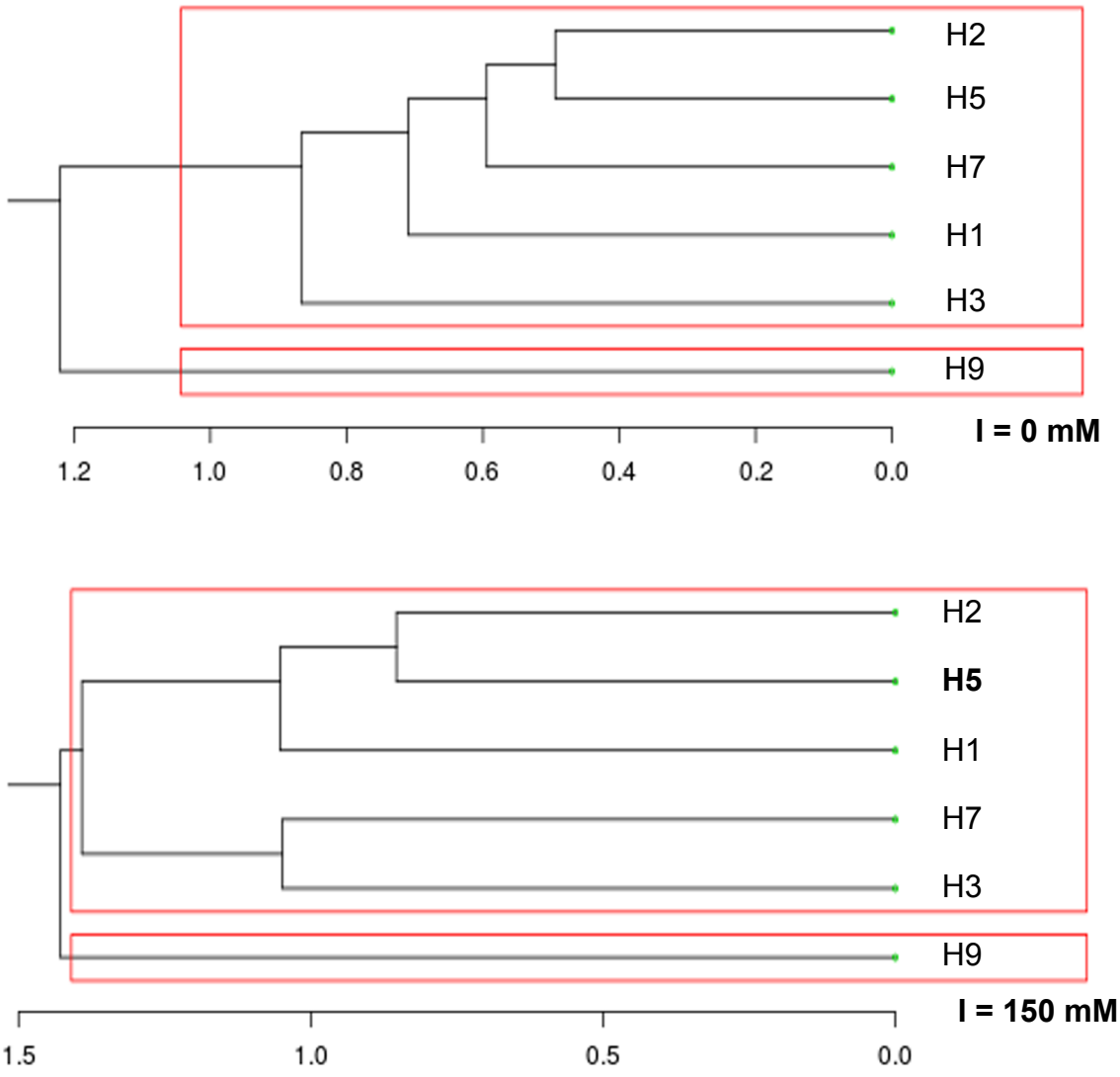
Epograms for the HA stem subregion. Epograms at I = 0 mM and I = 150 mM are shown. The horizontal axis of the epogram represents ED values.



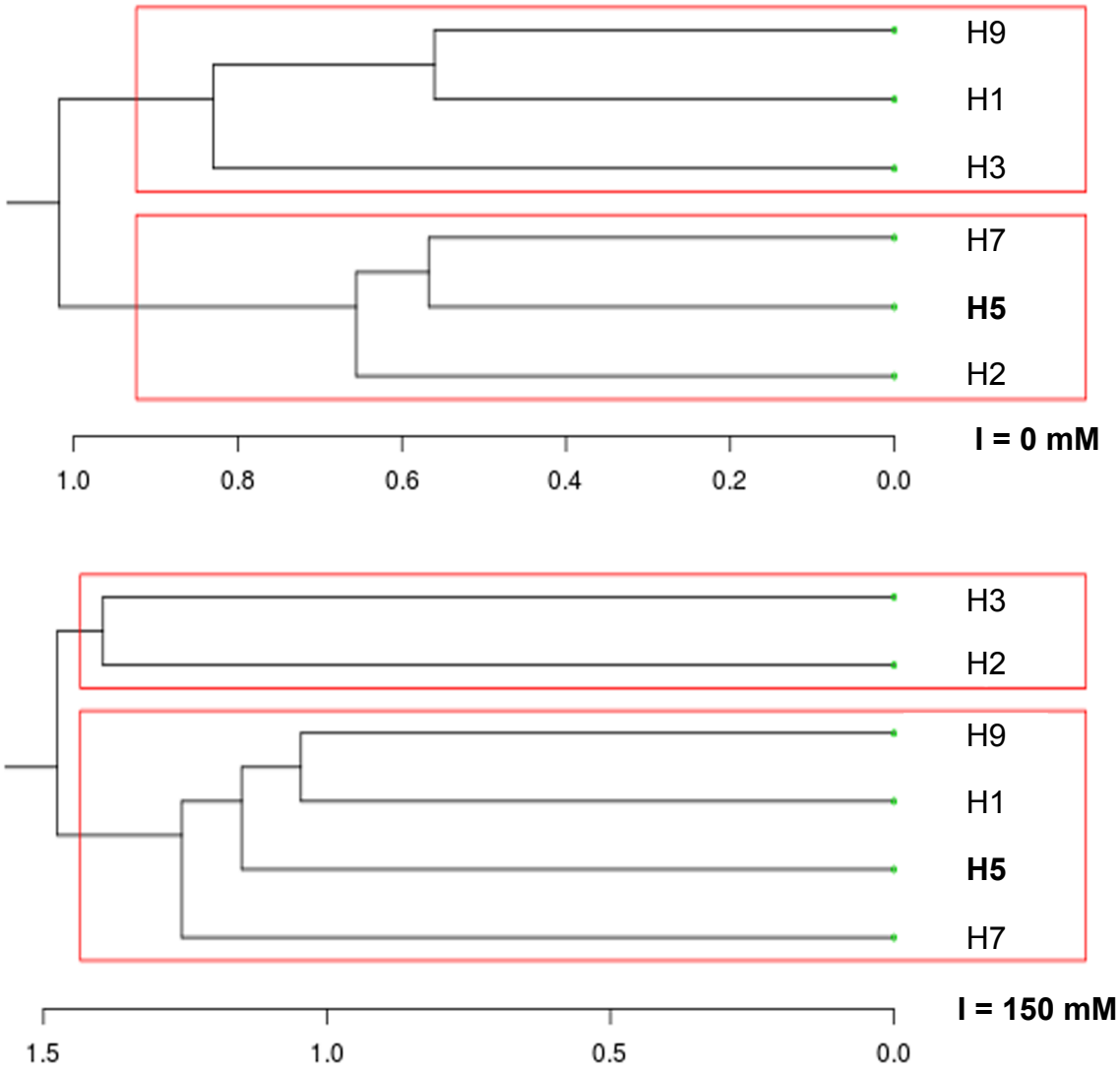
Epograms for the HA RBD subregion. Epograms at I = 0 mM and I = 150 mM are shown. The horizontal axis of the epogram represents ED values.



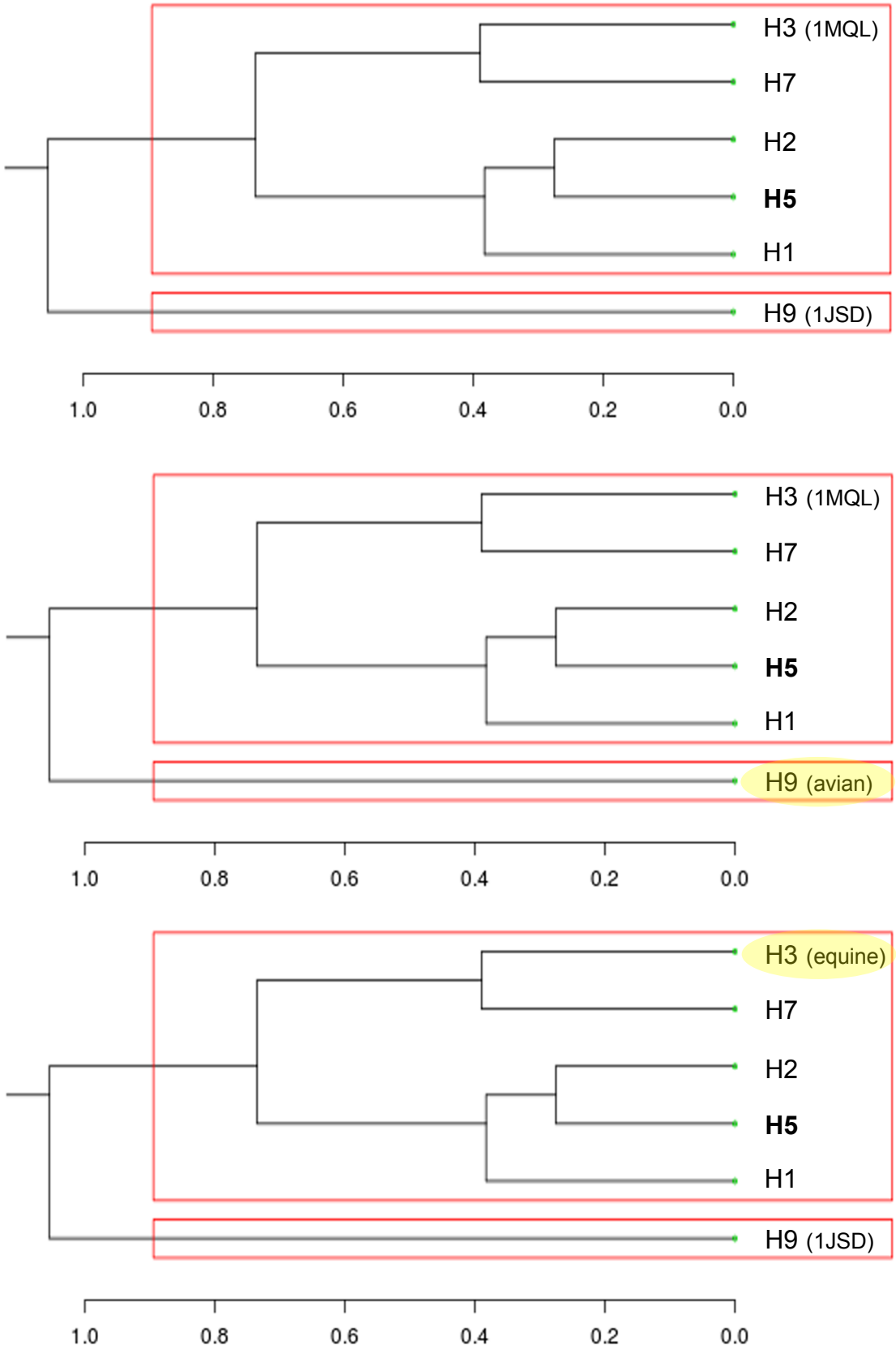
Epograms for the HA1 subregion. Epograms at I = 0 mM and I = 150 mM are shown. The horizontal axis of the epogram represents ED values.



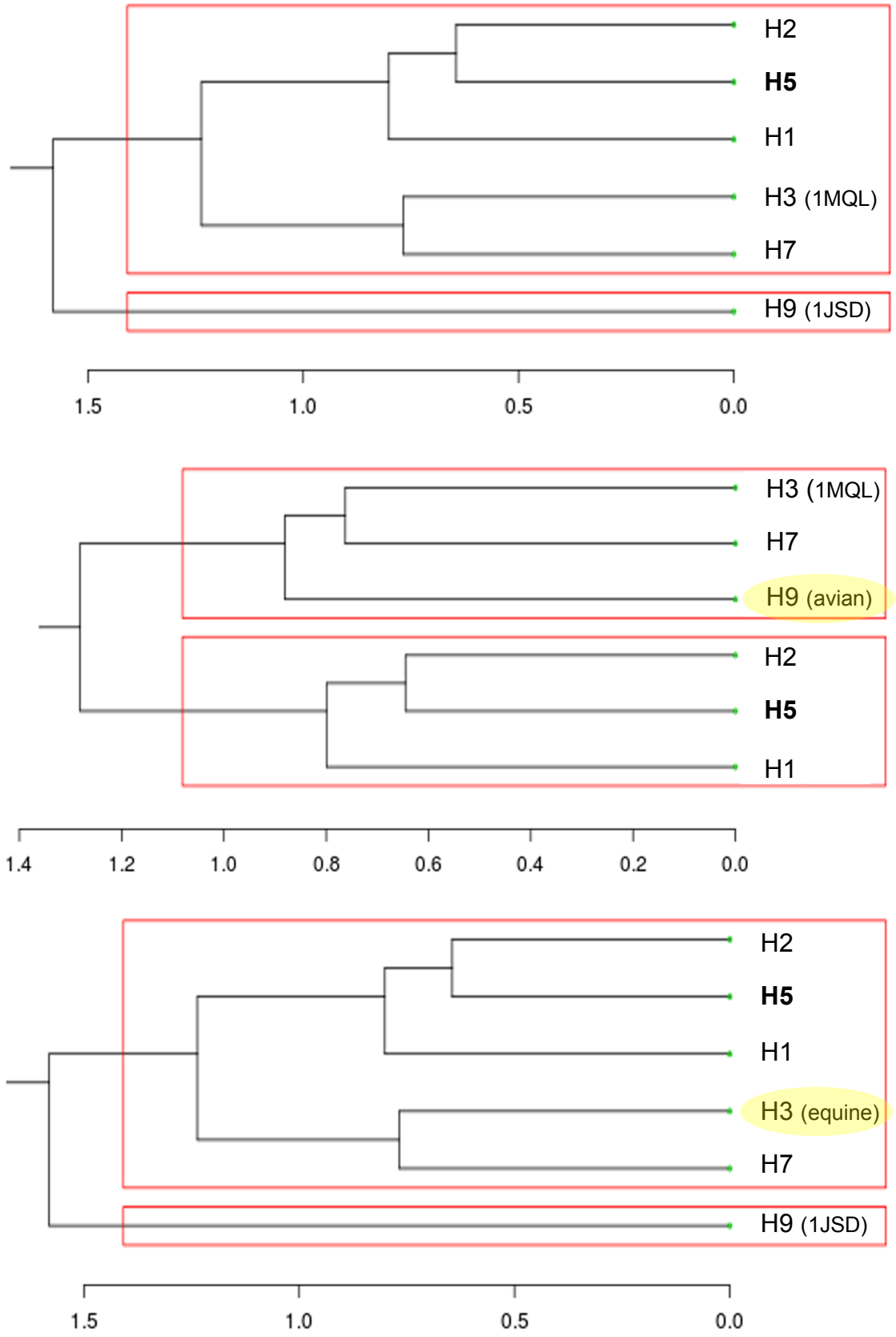
Epograms for the HA monomers. Epograms at I = 0 mM and I = 150 mM are shown. The horizontal axis of the epogram represents ED values.



Epograms for the HA trimers. Epograms at I = 0 mM and I = 150 mM are shown. The horizontal axis of the epogram represents ED values.



Epograms at I = 0 mM for the HA stem subregion. Modeled structures have yellow background. The horizontal axis of the epogram represents ED values.



Epograms at I = 150 mM for the HA stem subregion. Modeled structures have yellow background. The horizontal axis of the epogram represents ED values.

SUPPLEMENTARY MATERIAL CHAPTER 5

Table S1. H9N2 virus strains representative for each clade of the identified classes. Class and clade descriptions with isolation period, source, and geographic location are reported. (*) B ancestral strain for B class.

Class	Clade	Circulation data			Representative strain for each clade		
		Dates	Countries	Hosts	Full name	Short name	GISAID/NCBI
A	A.1	1966-2000	USA, China	Avian	A/turkey/CA/189/66	A.1_AtkCA66	EPI_ISL_1280/AF156390/AAD49000
	A.2	1995-1996	USA	Avian	A/turkey/Minnesota/38391-6/95	A.2_AtkMi95	EPI_ISL_1277/AF156387/AAD48997
	A.3	1980-2012	China, Korea, USA, Canada	Avian	A/goose/MN/5733-1/1980	A.3_AgoMN80	EPI_ISL_8950/CY006042/ABB88390
	A.4	1979-1984	Hong Kong, New Zealand	Avian	A/duck/NZL/76/1984	A.4_AdkNZL84	EPI_ISL_8942/CY005746/ABB20444
	A.5.1	2003-2011	USA, Georgia, UK	Avian	A/shorebird/Delaware Bay/283/2003	A.5.1_AshDB03	EPI_ISL_99336/CY102744/AET77176
	A.5.2	2009-2010	Vietnam	Avian	A/duck/Vietnam/OIE-2334/2010	A.5.2_AdkVN10	EPI_ISL_76652/AB569975/BAJ10561
	A.5.3	1999-2013	Japan, Russia, Australia, Chinam, Vietnam, UK, Netherlands, Italy, France, Finland, Austria, Switzerland, Norway, Portugal, South Africa, Zambia, Iran	Avian	A/duck/Hokkaido/13/00	A.5.3_AdkHo03	EPI_ISL_247/AY330340/AAQ97383
	A.5.4	1993-2008	Ireland, UK, Italy, Netherlands, Germany, USA	Avian	A/mallard/Ireland/PV46B/1993	A.5.4_AmaIRE93	EPI_ISL_647/AB303077/BAF62259
	A.5.5	1996-2009	Korea	Avian, Swine	A/chicken/Korea/AI-96004/1996	A.5.5_AckKo96	EPI_ISL_68688/GU053194/ACZ48629
D	D	1997-2001	Malaysia	Avian	A/duck/Malaysia/2001	D_AdkMa98	EPI_ISL_96739/CY073800/AEY75590
E	E	2006-2011	USA	Avian, environment	A/environment/California/NWRC186451-18/2007	E_AshDB00	EPI_ISL_132197/CY122538/AET77024

B	B (*)	1999	Pakistan	Avian	A/chicken/Pakistan/2/1999	B_AckPa99	EPI_ISL_146703/ KF188299/AGO17966
	B.1.1	1998-2007	Saudi Arabia, Japan, Iran	Avian	A/parakeet/Narita/92A/98	B.1.1_APaNa98	EPI_ISL_128/ AB049160/
	B.1.2	1998-2007	Lebanon, Iraq, Jordan, UAE, Israel, Saudi Arabia	Avian	A/chicken/Middle East/ED-1/1999	B.1.2_AckME99	EPI_ISL_68504/GU053201 /
	B.2.1	1998-2009	Iran, Iraq	Avian	A/chicken/Iran/B102/2005	B.2.1_AckIR05	EPI_ISL_11184/ EF063733/ABO09919
	B.2.2	2003-2008	India, Bangladesh, Kuwait	Avian	A/chicken/Chandigarh/2048/2003	B.2.2_AckCh03	EPI_ISL_78703/CY068643 /ADL64047
	B.2.3	2008-2014	Afghanistan, Pakistan, Iran	Avian	A/chicken/Afghanistan/329-6vir09-AFG-Khost9/2008	B.2.3_AckAf08	EPI_ISL_63785
	B.2.4	2009-2011	Nepal, India	Avian	A/chicken/Nepal/2490/2009	B.2.4_AckNE09	EPI_ISL_124228/ JX273549/AFO83282
	B.2.5	2005-2012	Pakistan, Afghanistan, Iran	Avian	A/chicken/Pakistan/UDL-01/2005	B.2.5_AckPA05	EPI_ISL_29787/ CY038410/ACP50642
	B.2.6	2005-2012	Saudi Arabia, UAE, Qatar, Israel, Libyan	Avian	A/chicken/Saudi Arabia/582/2005	B.2.6_AckSA05	EPI_ISL_124235/JX27355 6/AFO83289
	B.2.7	2006-2012	Israel, Egypt, Lebanon, Jordan	Avian	A/chicken/Israel/1525/2006	B.2.7_AckIS06	EPI_ISL_64314/FJ464728/ ACJ68774
	B.3	1999-2008	Hong Kong, USA, Vietnam	Avian, Human	A/Quail/Hong Kong/G1/97	B.3_AquHKG197	EPI_ISL_1268/ AF156378/ AAF00706
	B.4	2000-2002	UAE	Avian	A/quail/Dubai/303/2000	B.4_AquDu00	EPI_ISL_11171/ EF063512/ABM21877
C	C.1	2000-2005	Hong Kong, China	Avian/Human	A/quail/Shantou/1318/2000	C.1_AquSh00	EPI_ISL_11272/ EF154910/ABM46230
	C.2	1994-2008	China, Hong Kong, Japan	Avian, Swine, Human, Environment	A/chicken/Beijing/1/1994	C.2_AckBe94	EPI_ISL_146731/ KF188294/AGO17871
	C.2.1	2003-2011	China, Hong Kong, Japan	Avian	A/guinea fowl/Hong Kong/NT101/2003	C.2.1_AgfHK03	EPI_ISL_146694/ KF188382/ABB58955
	C.2.2	2001-2011	Hong Kong, Vietnam, China	Avian, Environment	A/chicken/Shandong/wd01/2007	C.2.2_AckSd07	EPI_ISL_24922/ FJ231868/ACI22608
	C.2.3	2007-2009	China	Avian	A/chicken/Hebei/A/2007	C.2.3_AckHe07	EPI_ISL_76661/ GQ202056/ACR56178

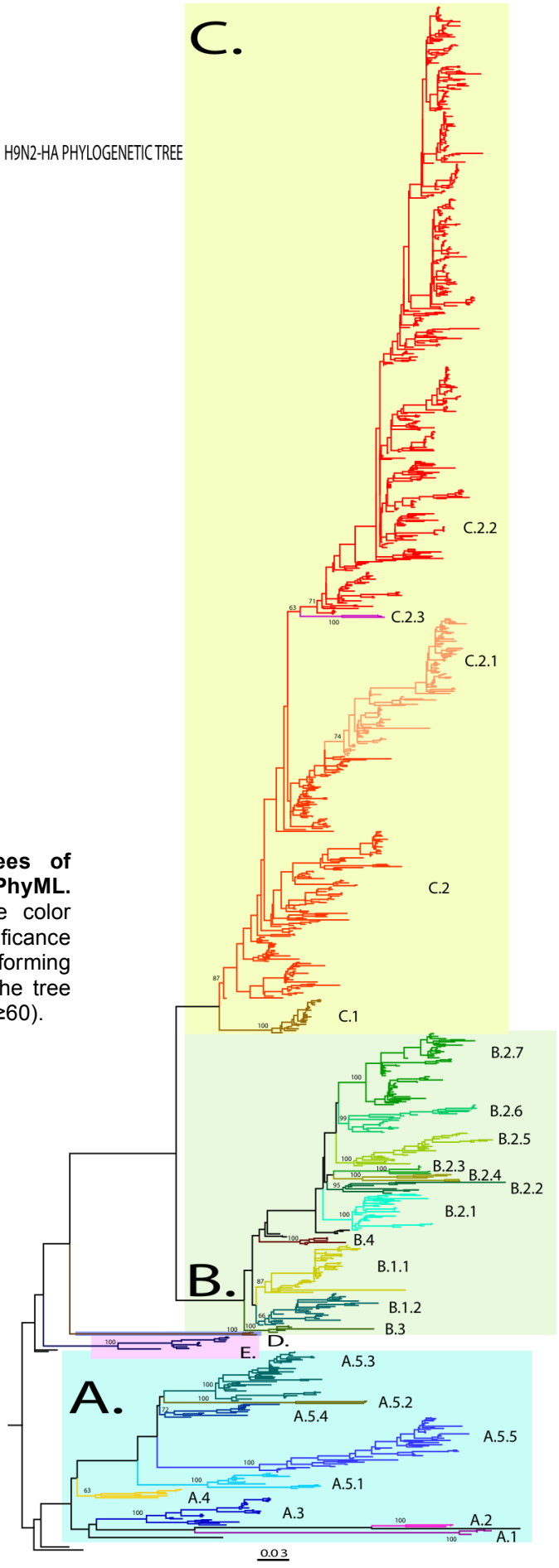


Figure S1: Maximum-likelihood trees of 1669 H9N2 isolates constructed by PhyML. The different classes and clades are color coded. Estimates of the statistical significance of phylogenies were calculated by performing 100 bootstrap replicates. Numbers in the tree nodes represent the bootstrap support (≥60).

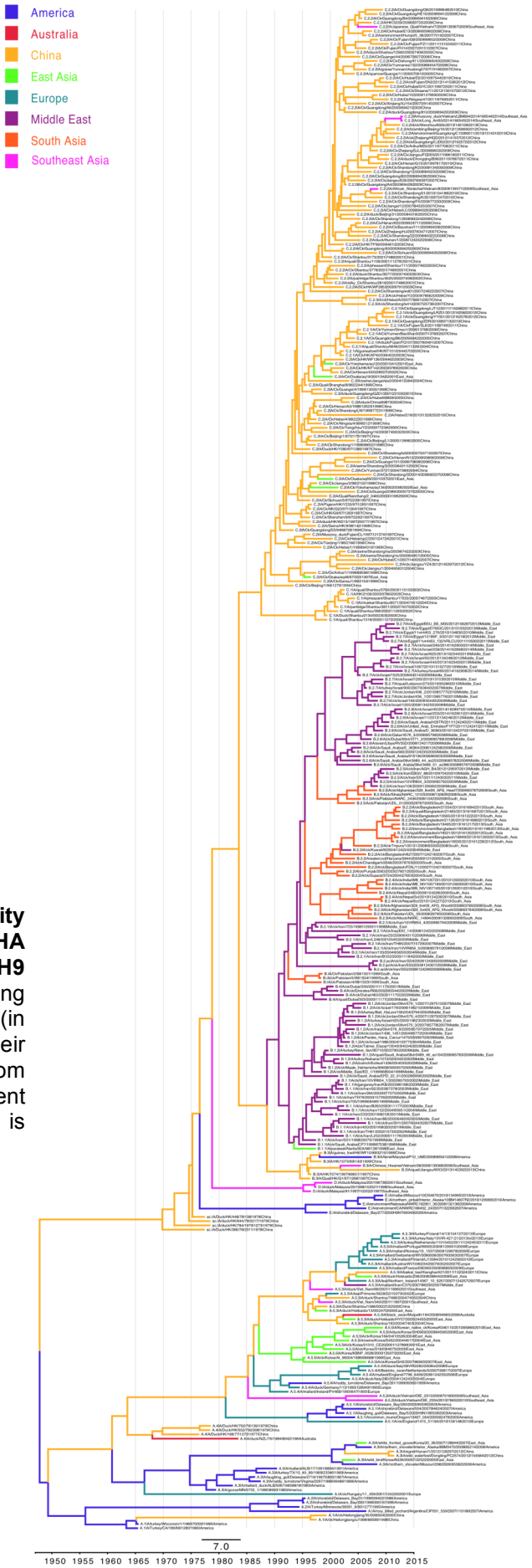
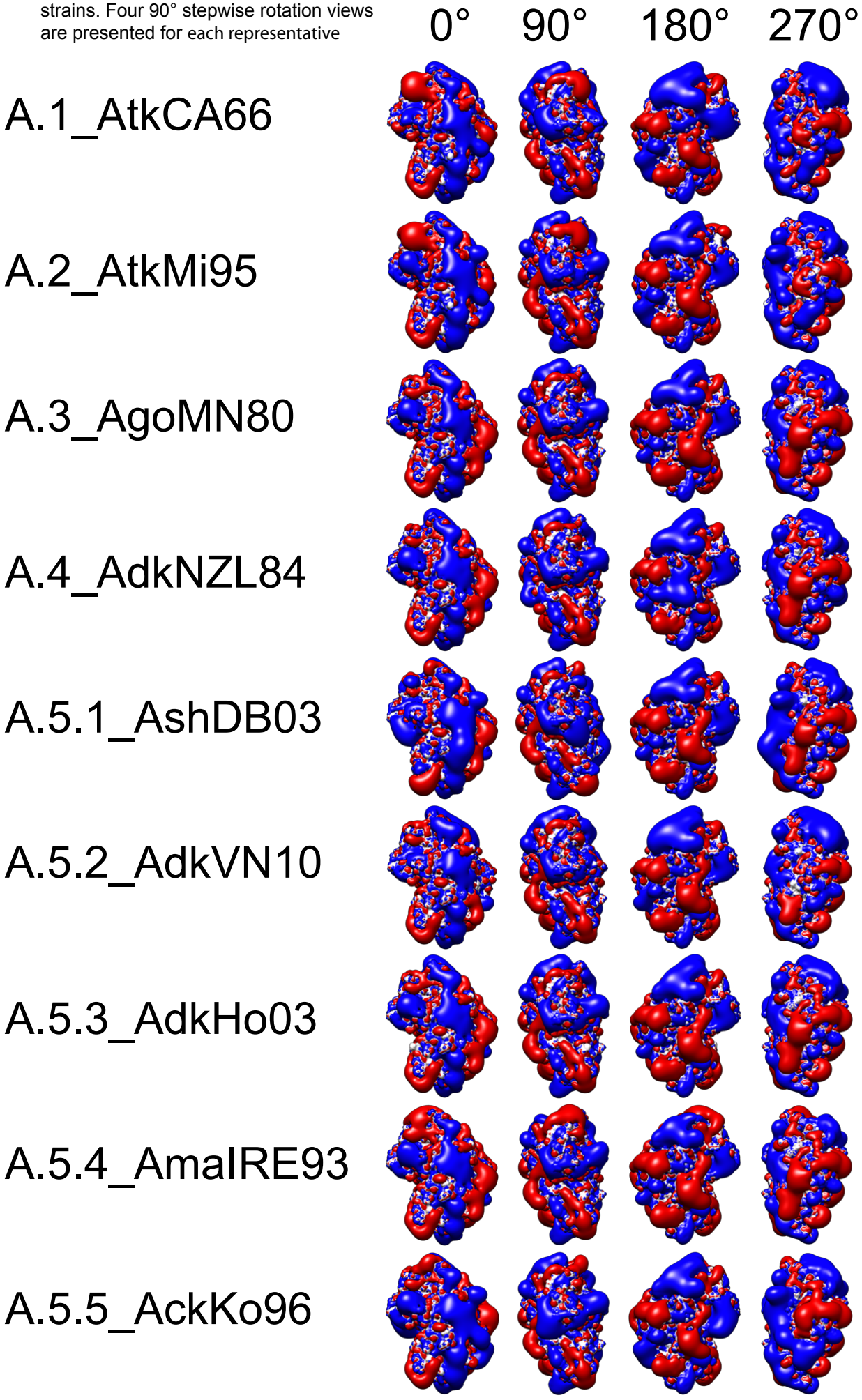
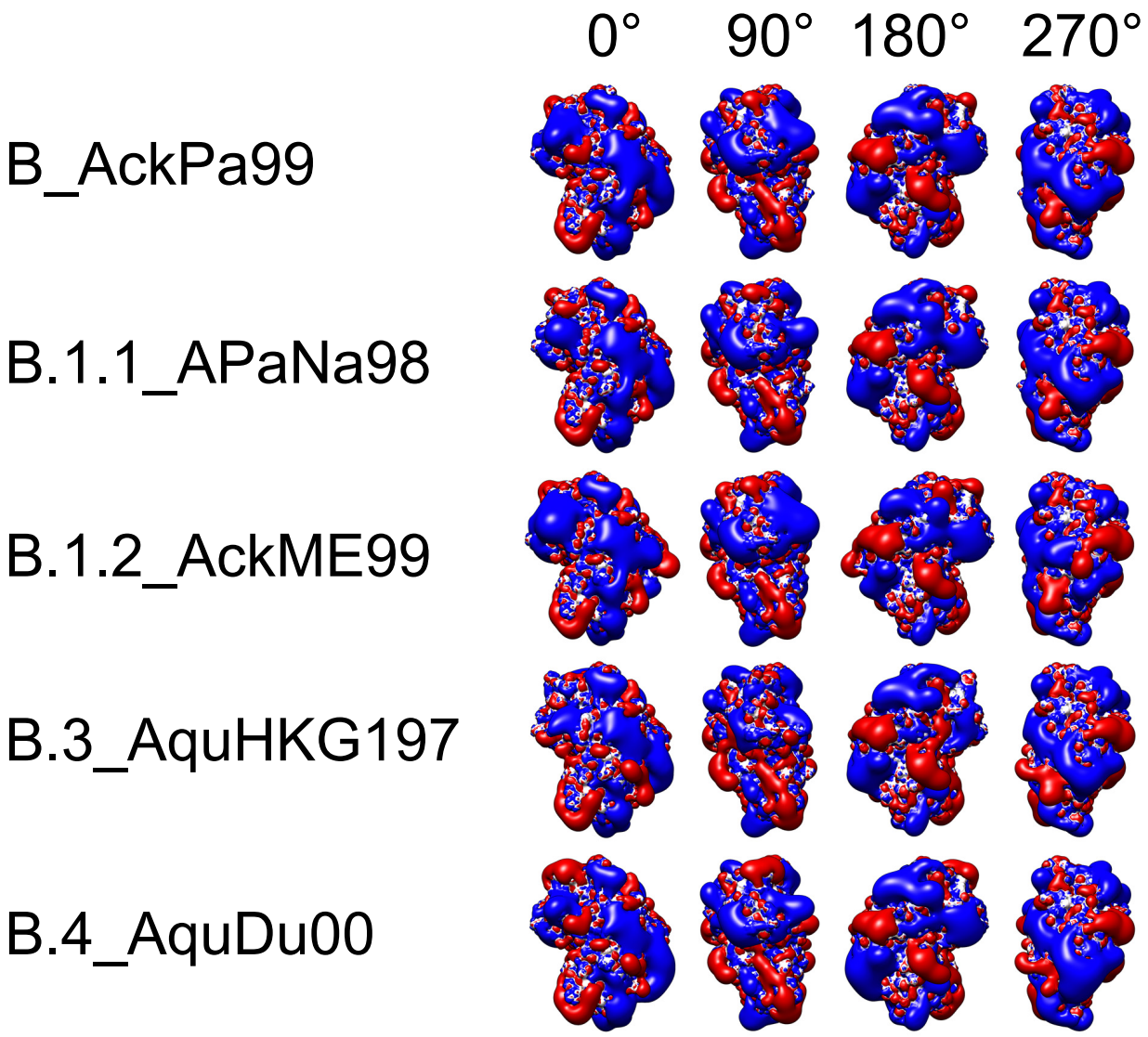
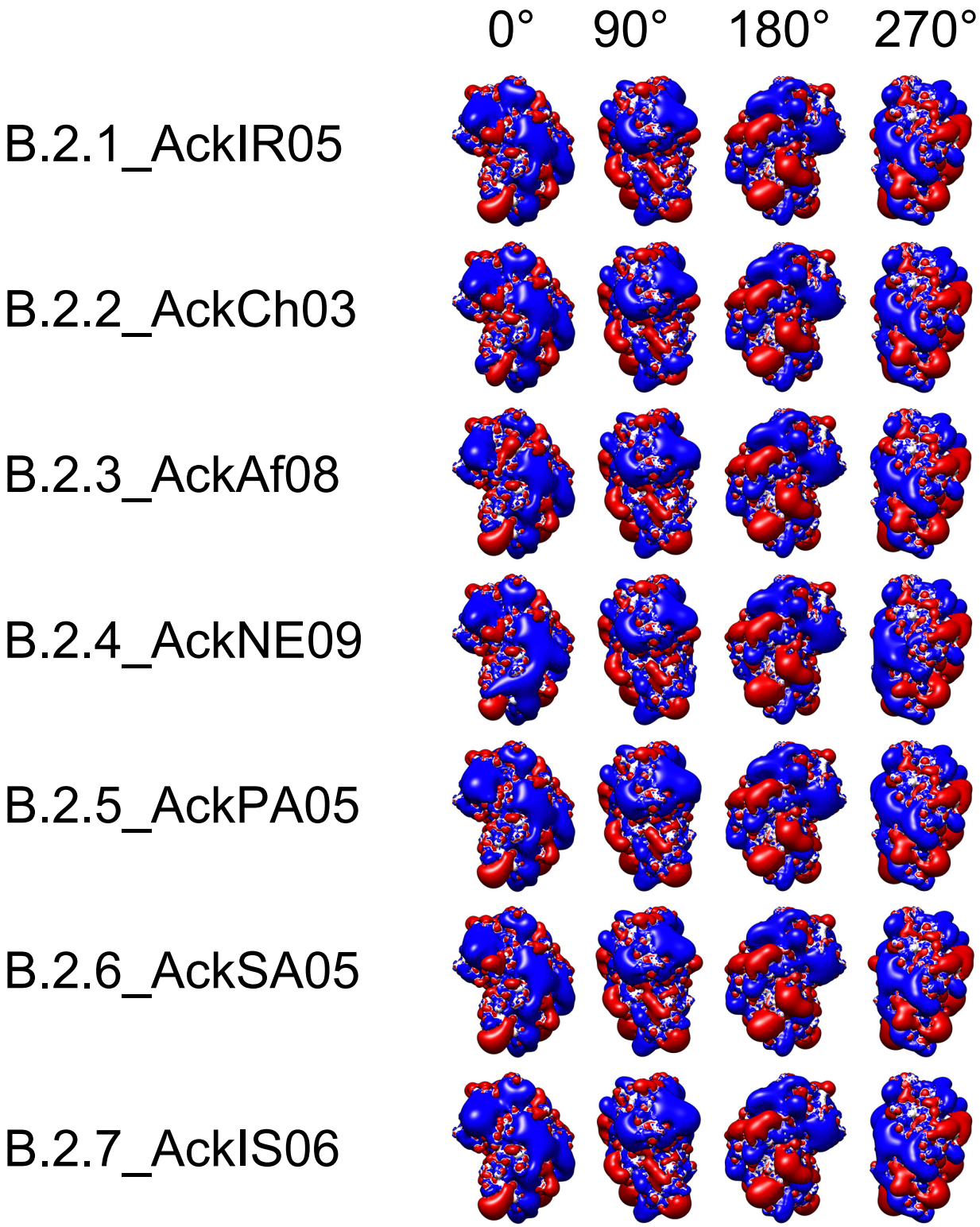


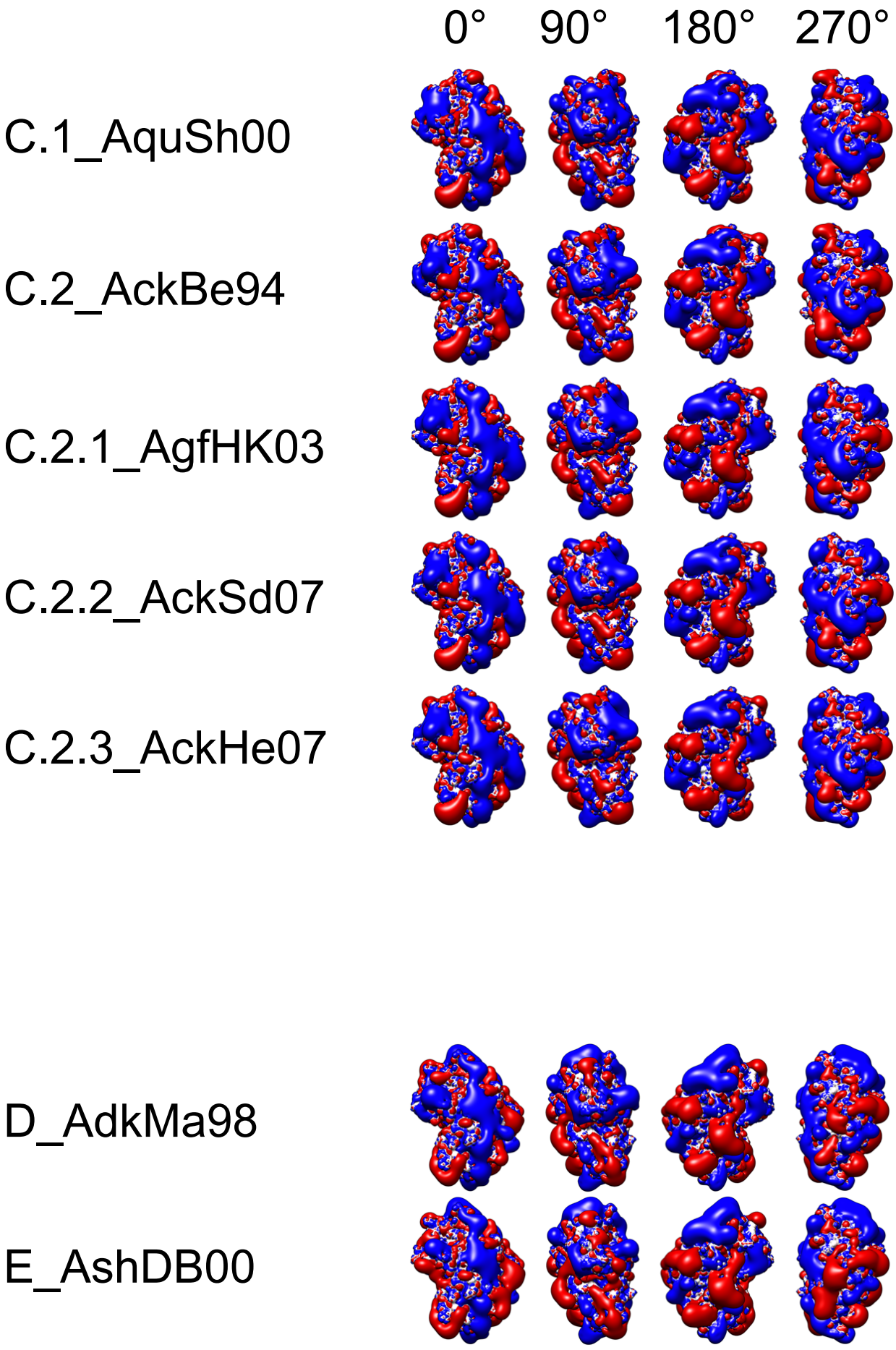
Figure S2. Maximum clade credibility (MCC) phylogenies inferred for the HA gene sequences of 357 viruses of AI H9 subtype. Branches are coloured according to the most probable ancestor location (in terms of geographic area) of their descendent nodes. Timeline at the bottom indicates the years before the most recent sampling time. Virus name information is reported in details.

Figure S3. Isopotential contours of the RBD from representative H9N2 virus strains. Four 90° stepwise rotation views are presented for each representative




0°
90°
180°
270°
B_AckPa99
B.1.1_APaNa98
B.1.2_AckME99
B.3_AquHKG197
B.4_AquDu00


0°
90°
180°
270°
B.2.1_AckIR05
B.2.2_AckCh03
B.2.3_AckAf08
B.2.4_AckNE09
B.2.5_AckPA05
B.2.6_AckSA05
B.2.7_AckIS06


0°
90°
180°
270°
C.1_AquSh00
C.2_AckBe94
C.2.1_AgfHK03
C.2.2_AckSd07
C.2.3_AckHe07
D_AdkMa98
E_AshDB00

ACKNOWLEDGEMENTS

I am greatly indebted to my enthusiastic supervisor prof. Francesco Filippini for being an important teacher who enlightened my interests on protein structural analysis and gave me invaluable direction and technical insight. Special mention goes to my co-supervisor Giovanni Cattoli for giving me so many wonderful opportunities and to have provided me consistent guidance and support throughout my study .

Similar, profound gratitude goes all my colleagues from IZSVe without whom this thesis would not be possible. In particular, I want to thank Isabella Monne and Alice Fusaro for their encouragements, advice and exchange of ideas. I also appreciate the crucial work of my colleagues Annalisa Salviato, Alessia Schivo, Luca Tassoni and many others who so generously contributed to the work presented in this thesis .

I am also hugely appreciative to Irene Righetto from Molecular Biology and Bioinformatics research laboratory, Department of Biology, especially for sharing her expertise on homology modelling.

Francesca Ellero is gratefully acknowledged for the proof reading of many parts of this thesis. A great thanks go to my family and to my friends for almost unbelievable support.