THE UNIVERSITY
of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

The evolution of drug resistance in influenza A viruses

Mojca Zelnikar

Doctor of Philosophy

The University of Edinburgh

2014

Abstract

Influenza A viruses are important pathogens of humans, other mammals and birds. Swine are considered to be the 'mixing vessel' for influenza viruses because of their susceptibility to infection with not only swine influenza viruses but also human and avian influenza viruses. After infection of pigs with different influenza viruses, reassortment events between genomic RNA segments and point mutations can take place which can result in novel influenza virus strains capable of causing human pandemics. To combat infections, vaccination is available in many countries for humans, but not typically used in pigs. However, anti-influenza drugs have been used to treat livestock, and mutations conferring drug resistance occur in circulating strains. The mechanisms responsible for the emergence and spread of drug resistant mutations against amantadine and oseltamivir have been studied previously but often gave conflicting results. Therefore, this PhD thesis focused on resolving the mechanisms responsible for this rapid drug resistance spread.

In chapter one I examine the extent of reassortment events in swine influenza A viruses by analysing within subtype reassortment and extrapolating the results for the between subtype reassortment. Reassortment is one of the mechanisms that can be responsible for mutations, conferring resistance to drugs, to spread between strains, and thus spread in the host population. The findings of this chapter show that the genomic segments most prone to reassortment code for a polymerase (PB1) and both glycoproteins, within all three subtypes studied. Since particular mutations in the matrix protein (MP) segment cause resistance to amantadine, my study focused on MP compared to other segments and revealed moderate level of reassortment. MP reassorts well with polymerases, both within and between subtype, while non-structural (NS) is least likely to reassort.

Chapter two of this thesis aimed at resolving the origin and spread of the most common drug resistance conferring mutation in swine influenza viruses which causes amantadine resistance. I show first that this mutation occurred in swine influenza viruses and was therefore not transmitted from the recently ancestral avian influenza

strains, and second that the prevalence of resistance in swine influenza viruses is due to functional linkage of mutations at other sites and not by direct drug pressure.

In chapter three I examine the mechanisms responsible for the rapid rise and spread of oseltamivir resistance in human influenza H1N1 viruses which arose in the absence of drug use. The primary mutation lies in the neuraminidase glycoprotein but because of the close functional interaction I focus on changes in haemagglutinin that occurred in association with resistance. The results showed several mutations in haemagglutinin were associated with resistance suggesting selection acting on haemagglutinin in order to balance the activity of both glycoproteins.

Overall these results show the importance of functional linkage between segments as a mechanism for the occurrence of drug resistance conferring mutations, and reassortment as a means of spreading these mutations into newly emerging strains.

Lay summary

Influenza A viruses are important pathogens of humans, other mammals and birds. The genetic material of these viruses is separated into eight molecules, called segments, and may change very rapidly. Swine are considered to be crucial for generating genetically novel influenza viruses because they are susceptible to infection with influenza viruses infecting swine, humans and birds. After pigs are infected with different influenza viruses at the same time, viral segments can be mixed into different combinations within swine generating novel influenza viruses. This process is called reassortment. These novel influenza A viruses are potentially capable infecting humans worldwide in a relatively short time causing severe illness. Two anti-influenza drugs have been widely used and resistance to these drugs has arisen: amantadine in both humans and livestock, and oseltamivir in humans. The mechanisms of the virus for the global spread of drug resistant influenza A viruses have been studied previously but often gave conflicting results. Therefore, this PhD thesis focused on resolving the mechanisms responsible for this rapid drug resistance spread.

Firstly, I investigated how much reassortment occurred in pigs as it can be responsible for generating drug resistant viruses. In this chapter I show that some of the eight segments are very prone to reassortment and some are unlikely to reassort. The segments that reassort very frequently encode two proteins found on the surface of the virus and one protein essential for replication of viral genetic information during infection. The segments that reassort very infrequently encode four proteins, two of them are structural components of the virion, the other two proteins are nonstructural, regulatory proteins.

Next the origin of genetic changes (mutations) rendering the swine influenza A viruses amantadine resistant was studied. Results of this chapter show that these mutations occurred in the swine host and not in the bird host before this virus was transmitted from birds to swine. The results also showed that the influenza A virus did not acquire the amantadine drug resistance multiple times. The mechanism for

acquiring amantadine drug-resistant mutations lies in the functional connection of advantageous mutations for the virus to the amantadine drug-resistant mutations.

Finally, influenza A viruses of H1N1 subtype infecting humans were examined. These viruses acquired mutations that make this virus oseltamivir resistant. This drug binds to the neuraminidase, found on viral surface, and blocks its function. The function of another surface protein, called haemagglutinin, is balanced with the function of the neuraminidase. The results of this chapter show that several mutations in the haemagglutinin occurred around the time of oseltamivir drug-resistant mutation in the neuraminidase. So, mutations in the haemagglutinin are associated with the resistance mutation in the neuraminidase.

Overall, these results show the importance of 'functional linkage' between different viral segments for the occurrence of mutations making the influenza A viruses drug resistant. Further, reassortment is an important mechanism for spreading these mutations into newly emerging influenza A viruses.

Acknowledgements

I would like to express my great appreciation and thanks to my supervisor Prof Andrew Leigh Brown, who was the first to accept me to the University of Edinburgh to work in his group of excellent scientists, guided me on my scientific path with patience and kindness and always believed in me. My deep gratitude goes also to my second supervisor Dr Samantha Lycett for encouraging me in my research and for endless effort put toward my growth as a scientist. Your scientific and computer expertise not only enhanced my work but also inspired me to continue learning and discovering the fields of bioinformatics and phylogenetics. I would also like to thank my colleague Emma Hodcroft for immense help with programming and other computer issues. Without your input my work would not be as successful and enjoyable as it has been. My thank you also goes to Lu Lu for all your advice, support and the nice time we spent together as PhD students. My big thank you goes to Gytis Dudas for many times I sought your scientific advice and help. I am grateful to Manon Ragonnet-Cronin for helpful discussions and to Dr Melissa Ward for many kind suggestions, advice and great help with my PhD work but also for welcoming me warmly to your home and showing me around Edinburgh upon my first arrival. My appreciation also goes to Dr Jessica Hedge and Dr Jayna Raghwani for helping me when I first started my PhD.

My scientific endeavour as a PhD student would not have been possible without my work being funded. Therefore, I would like to thank the Slovene Human Resources Development and Scholarship Fund for rewarding me with their scholarship.

A huge thank you goes to my parents and brother for love and support during my 4 years study in UK, far from home. I would like to thank my dear friends Metod Krivec, Sabina Miklavcic and Natasa Sumi from my homeland Slovenia. Thank you Owolabi Badejo for loving me and being proud of me in the country we now both call home.

Declaration

I declare that this thesis has been composed by me and that the work described herein is my own. This work has not been submitted for any other degree or professional qualification except as specified.

Mojca Zelnikar

22 October 2014

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction to influenza viruses

1 Introduction to influenza viruses

In this chapter, an overview of the influenza virus is provided. In particular, the focus is on human and porcine influenza viruses, their structure, replication, transmission and evolution.

1.1 Background to influenza

Influenza is an infectious disease of birds and mammals that is caused by influenza viruses. Influenza of humans spreads around the world in seasonal epidemics with three to five million cases of severe illness and between 250,000 to 500,000 deaths per year (WHO, 2009a). In pandemic years, loses of human lives can be in millions. New influenza strains of humans often appear when an existing influenza virus spreads from some animal species to humans, or when an existing human influenza strain obtains new genes from another influenza virus that commonly infects birds or pigs. Therefore, vaccination of humans and farmed poultry is available in developed countries (WHO, 2005; Villegas, 1998). Also, antiviral drugs have been used for decades (WHO, 2009b). However, influenza viruses evolve rapidly and vaccines against new influenza strains have to be produced on a yearly basis. In the last decade, we have also seen a rapid rise in drug resistant strains. Therefore, influenza viruses pose a threat to human health in either developed or developing countries.

In this PhD, the focus is on studying the origin and mechanisms for the spread of drug resistant influenza strains in humans (oseltamivir resistance) and pigs (amantadine resistance). In both cases, drug resistance rose and spread globally in a very short time period and the reasons for that are not fully understood or they are contradictory. One of the mechanisms for this could be the reassortment occurring in swine therefore this PhD focused also on the study of the amount or the quantity of this genetic process in porcine host, which is also one of the areas of influenza study not fully elucidated.

1.2 Influenza A virus taxonomy and nomenclature

Influenza viruses belong to the *Orthomyxoviridae* virus family which includes six genera: Influenzavirus A, Influenzavirus B, Influenzavirus C, Isavirus, Thogotovirus and Quaranjavirus (Fauquet et al., 2005; Bussetti et al., 2012). Influenza viruses are classified into types A, B and C on the basis of their ribonucleoprotein antigens. Influenza A viruses only are classified further into subtypes according to the combination of two surface glycoproteins; haemagglutinin and neuraminidase (WHO Expert Committee on Influenza, 1953; Anon, 1979). Haemagglutinin (HA) and neuraminidase (NA) are morphologically and immunologically distinct and are known to undergo antigenic variation. Therefore, all three antigens must be considered when describing influenza virus strains. A uniform code was proposed for designating influenza virus strains. Their name consists of a viral genus (type), the species from which the virus was isolated (omitted if human), isolate location, isolate number, the year of isolation, and only in the case of influenza A viruses, the subtypes of HA and NA. For example, the name A/England/1/1953 refers to the type A, which was isolated from a human host in England in 1953. Number 1 indicates the isolate number. Currently there are 18 different hemagglutinin and 11 different neuraminidase subtypes known (Fouchier et al., 2005; Webster et al., 1992; Tong et al., 2012; Tong et al., 2013).

1.3 Genome structure

The genome of influenza A viruses is approximately 13.5 kb long and segmented into eight negative sense single stranded RNA molecules which are numbered 1 to 8 with segment 1 being the longest and segment 8 the shortest. Negative sense RNA is complementary to mRNA and must be copied to positive sense RNA by viral RNA polymerase prior to translation. The three largest segments code for RNA polymerase complex proteins PB2, PB1 and PA. In some strains of influenza A virus the segment 2 which codes for PB1 polymerase subunit also codes for the accessory protein PB1-F2, a small, 87-amino acid protein with pro-apoptotic activity, in a +1

alternate reading frame (Chen et al., 2001). Segment 4 carries a gene for HA and segment 5 a gene for nucleoprotein (NP). Segment 6 of influenza A virus encodes the NA protein and segment 7 codes for the M1 matrix protein. In the influenza A genome, the M2 ion channel is also expressed from segment 7 by RNA splicing (Lamb et al., 1981). From segment 8 influenza A expresses the interferon-antagonist NS1 protein (Dauber et al., 2004; Garcia-Sastre, 2001; Kochs et al., 2007) and, by mRNA splicing, the NEP/NS2 (Briedis and Lamb, 1982; Lamb et al., 1980), which is involved in viral RNP export from the host cell nucleus (Table 1.1). Other proteins have also been discovered but their functions remain unclear. These proteins are PA-X (Jagger et al., 2012), PA-N155, PA-N182 (Muramoto et al., 2013) and PB1-N40 (Wise et al., 2009).

Table 1.1

Influenza A genome segments, the main proteins they encode and function of these proteins.

No OF SEGMENT	SEGMENT	SEGMENT LENGTH (nt)	PROTEIN/S IT ENCODES	FUNCTION OF PROTEIN/S	
1	PB2	2,277	PB2 polymerase subunit	binds to a 5' cap of host mRNAs	viral RNA synthesis
2	PB1	2,271	PB1 polymerase subunit	transcription and replication	
3	PA	2,148	PA polymerase subunit	endonuclease, removes 5' cap from host mRNAs	
4	HA	1,698	haemagglutinin	attachment and entrance into the host cell	
5	NP	1,494	nucleoprotein	structural and regulatory role	
6	NA	1,407	neuraminidase	removes sialic acid from receptors on cell surface, from HA and NA to facilitate virus release	
7	MP	979	M1 matrix protein	virus structure: forms a matrix holding viral RNPs	
			M2 matrix protein	proton-selective ion channel protein	
8	NS	835	nonstructural NS1 protein	inhibits host innate response (interferon antagonist)	
			nonstructural NS2 protein	export of vRNPs from the nucleus	

RNA segments share a common structure (Figure 1.1). Both ends of each segment are untranslated regions (UTRs) and constitute promoter sequences which contain a universal conserved region (UCR) and a segment specific region (SSR). UCR is 13 nucleotides long at 5' end and 12 nucleotides at 3' end (Desselberger et al., 1980; Robertson, 1979; Skehel and Hay, 1978). However, the fourth nucleotide at 3' end is variable carrying either C (cytosine) or U (uracil). UCR regions are highly conserved across all influenza segments and due to partial complementarity these ends base pair to form a helical hairpin which binds proteins of RNA polymerase complex (Bouvier and Palese, 2008). UCRs are followed by segment specific regions (SSRs) which vary in length across segments (19 to 58 nucleotides). The rest of the segment is coding region (open reading frame – ORF). SSRs and coding regions are covered by positively charged NP due to arginine residues (Baudin et al., 1994; Compans et al., 1972; Murti et al., 1988). Influenza virus segments also contain packaging signals which are crucial for the incorporation of the whole viral genome into newly synthesized virions. Packaging signals span the entire UTR and terminal parts of a coding region. Alltogether, this structure of viral RNA and proteins represents a ribonucleoprotein (RNP) (Hutchinson et al., 2010) (Figure 1.1).

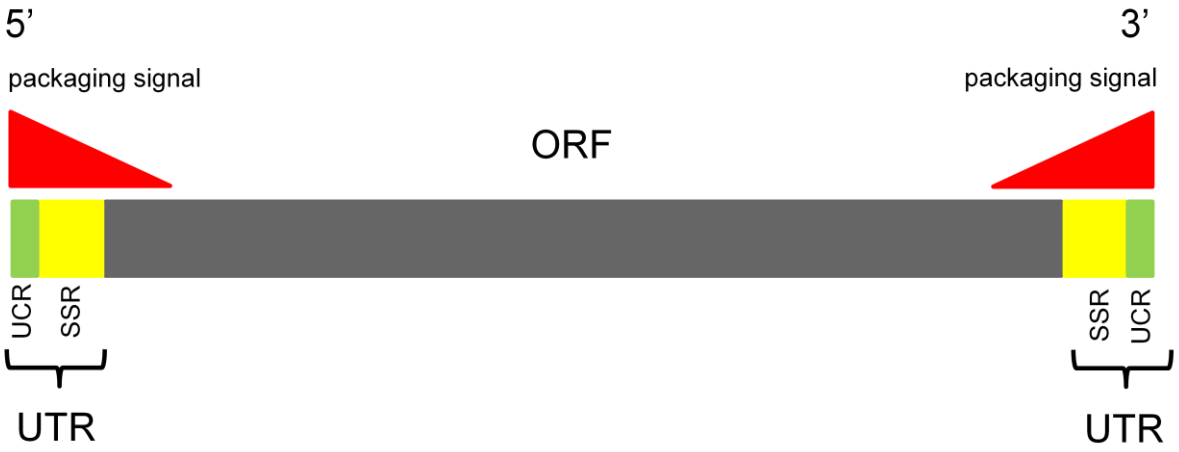

Figure 1.1

Schematic diagram of the structure of one influenza genome segment. Untranslated regions (UTRs) at 5' and 3' segment ends constitute a promoter and are composed of a universal conserved region (UCR) (green box) and a segment specific region (SSR) (yellow box). Each segment contains packaging signals (red box). Open reading frame of a protein coding region (ORF) is presented as a grey box. Figure made by the author.

1.4 Virion structure

When outside the host cell, the infectious particle of influenza is called a virion. The shape of influenza virions is usually spherical or elliptical with 80-120 nm in diameter or filamentous reaching more than 20 μm in length (Bourmakina and Garcia-Sastre, 2003). The surface of influenza virions is covered by a membrane which is host derived (Nayak et al., 2004) and contains two glycoproteins, HA and NA as well as M2 ion channel protein, protruding from the membrane (Figure 1.2). Each virion contains approximately 500 HA and 100 NA molecules. Beneath the lipid membrane is the M1 matrix protein that forms a coat for the encapsidation of the viral ribonucleoprotein complexes (Sha and Luo, 1997) which are the genetic material of the virus. Each ribonucleoprotein complex contains one RNA segment bound by PB2, PB1, PA and NP proteins. Virions also contain NEP/NS2 protein (Richardson and Akkina, 1991) (Figure 1.2).

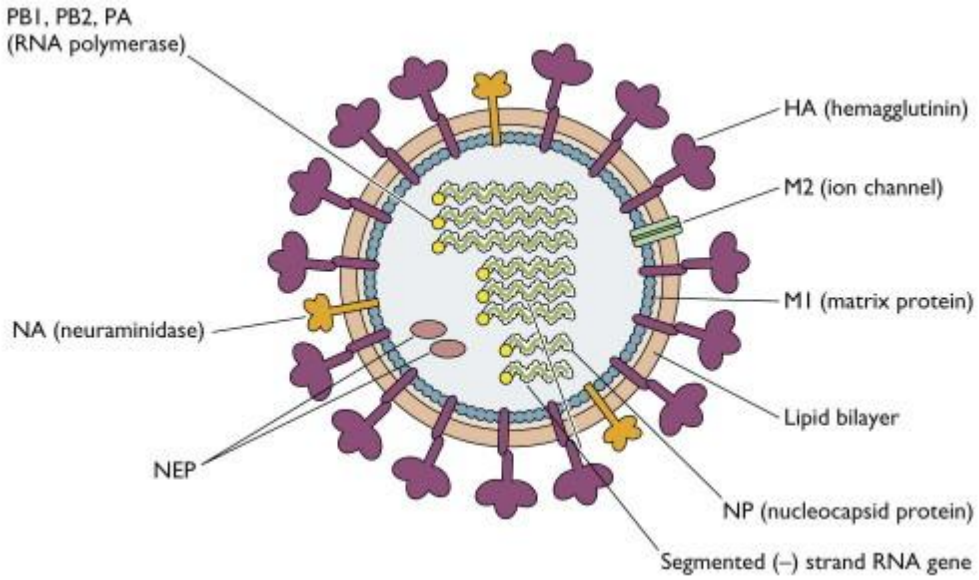

Figure 1.2

Schematic diagram of the influenza A virion structure (from Racaniello, 2009).

1.5 Characterization of the HA and NA antigens into subtypes

The haemagglutinin and neuraminidase subtypes are characterized separately (Anon, 1971). It is not advised to use sera containing antibodies to both antigens because antibodies to one antigen can sterically interfere with antibodies to another antigen. By using the haemagglutination - inhibition test, the haemagglutinin antigen subtype is recognized. The neuraminidase antigens are divided into subtypes according to the results of the neuraminidase - inhibition test. For both, haemagglutinin and neuraminidase, immunoprecipitation test can be used, which reveals antigenic variation within a given subtype (Anon, 1971).

1.5.1 Characterization of the HA subtype

The haemagglutination-inhibition assay uses reference antisera and is the standard method for subtyping HA of influenza viruses (Office International des Epizooties, 2005; Swayne et al., 1998). The basis for this test is that influenza viruses will bind to the sialic acid on red blood cells. One influenza virion can attach to more than one red blood cell (erythrocyte) at a time, therefore cross-linking or clumping of erythrocytes occurs. Haemagglutination can be inhibited by antibodies directed against the HA protein, which are subtype specific and so antibodies against one subtype will not typically react with the HA of another subtype. Therefore, the haemagglutination-inhibition assay has been used as the primary and classical method of identifying the HA subtype of an unknown influenza virus.

1.5.2 Characterization of the NA subtype

The basis for this assay is the inhibition of viral neuraminidase enzymatic activity by specific antibodies. The neuraminidase-inhibition assay can be performed on cell culture fluid or amnioallantoic fluid with subtype-specific antisera for the determination of NA subtype. Using reference sera from viruses with the same HA

subtype as the test materials will improve specificity by allowing the NA subtype to be the only variable in the test (Van Deusen et al., 1983).

1.6 Structures of HA

There are three structures of haemagglutinin ectodomain known that have been determined by X-ray crystallography for 1968 Hong Kong influenza: (i) HA0, which is a single-chain precursor; (ii) the conformation found on virus and (iii) the fusion conformation (reviewed in Skehel and Wiley, 2000). These structures enabled the insight in the activity of haemagglutinin when binding to a receptor and membrane fusion during viral entry.

1.6.1 Roles of different HA structures in influenza virus replication process

Haemagglutinin is a type I membrane glycoprotein with a signal sequence that is removed post-translationally. It also encompasses a membrane anchor domain near the C-terminus and a short N-terminus cytoplasmic tail (reviewed in Steinhauer DA, 1999). In the infected cell, haemagglutinin is synthesized as a precursor HA0 of approximately 75 kDa and homotrimerized noncovalently with the aid of chaperones in the endoplasmic reticulum and then transferred to the cell surface through the Golgi apparatus (Gething et al., 1986; Copeland et al., 1988; Hebert et al., 1995). Cleavage of HA0 into HA1 and HA2 subunits of human and most non-pathogenic avian strains occurs at a conserved arginine residue and is carried out by extracellular host proteases after the release of newly synthesized virions from the host cell. Cleavage of HA0 of highly pathogenic avian strains occurs at polybasic residues intracellularly by ubiquitous proteases (Steinhauer, 2010). HA1 and HA2 are linked by a single disulfide bond. This activates the membrane fusion potential of the haemagglutinin (Maeda and Ohnishi, 1980; Huang et al., 1980; White et al., 1981) and thus enables virus infectivity (Klenk et al., 1975; Lazarowicz et al., 1975). Cleavage of the precursor HA0 liberates the hydrophobic N-terminus of HA2 (the fusion peptide), which is inserted into target membranes during fusion. After

cleavage, the HA2 undergoes molecular rearrangements required for fusion, triggered by the acidic pH of endosomes (Bullough et al., 1994). The cleavage properties of HA0 and the distribution of infectivity-activating proteases in the host are major factors for virus tropism and the capacity for systemic spread. HA1 subunit contains oligosaccharide attachment sites, receptor binding site which determines host specificity of the virus, and antigenic sites which are prone to mutations causing structure changes against which neutralizing antibodies are produced (reviewed in Skehel and Wiley, 2000).

1.7 HA cleavage site as a critical determinant of pathogenicity

Factors contributing to the pathogenicity of influenza viruses have been extensively studied (reviewed in Steinhauer, 1999). Relevant factors for the influenza virus pathogenicity are tissue tropism and host range specificity, the availability of a functional receptor on the host cell and the presence of an appropriate gene constellation of the virus, which determines growth characteristics and the pathogenicity (Steinhauer, 1999). The human pandemics viruses (1957 and 1968), which were generated due to reassortment of avian and human influenza viruses, derived the surface antigens from an avian source (Kawaoka et al., 1989; Scholtissek et al., 1978b, Smith et al., 2009b). For naturally occurring avian influenza A viruses, the most important determinant of pathogenicity is the cleavage site structure of the HA precursor. It is present at the HA1 – HA2 junction and contains basic residues in pathogenic strains (Bosch et al., 1981; Porter et al., 1979). Viruses with basic residues at the cleavage site have been isolated from different geographical regions and different host species at different times (Wood et al., 1993; Senne et al., 1996; Purdue et al., 1997). Table 1.2 shows a number of examples of these (adapted from Steinhauer, 1999). In addition to the presence of the polybasic residues at the cleavage site, most pathogenic strains also contain amino acid insertions compared to non-pathogenic strains. All of the highly pathogenic natural isolates characterized to date are either H5 or H7 subtype viruses (Suenaga and Kumar, 2014).

Table 1.2

Examples of cleavage site sequences from natural isolates of H5 and H7 influenza A viruses (adapted from Steinhauer DA, 1999).

ISOLATE	CLEAVAGE SITE	PATHOGENICITY
H5 subtypes		
A/chicken/Mexico/31381/94	P Q R E - - - - T R ↓ G	-
A/chicken/Pueblo/94	P Q R K R K - - T R ↓ G	+
A/chicken/Queretaro/20/95	P Q R K R K R K T R ↓ G	+
A/duck/Ireland/113/83	P Q R K R K - - K R ↓ G	+
A/turkey/Ireland/1378/83	P Q R K R K - - K R ↓ G	+
A/chicken/Pennsylvania/1/83 (CHO+)	P Q K K - - - - K R ↓ G	-
A/chicken/Pennsylvania/1370/83 (CHO-)	P Q K K - - - - K R ↓ G	+
A/duck/Singapore/645/97	P Q R E - - - - T R ↓ G	-
A/chicken/Hong Kong/990/97	P Q R E R R R K K R ↓ G	+
A/Hong Kong/156/97 – (human)	P Q R E T R R K K R ↓ G	+
A/ Hong Kong/486/97 – (human)	P Q R E R R R K K R ↓ G	+
H7 subtypes		
A/tern/Potsdam/79	P E I P K - - - - G R ↓ G	-
A/chicken/Leipzig/79	P E I P K K K - - G R ↓ G	+
A/goose/Leipzig/137/79	P E I P K R K - - G R ↓ G	+
A/goose/Leipzig/187/79	P E I P K K K K - G R ↓ G	+
A/goose/Leipzig/192/79	P E I P K K K K K G R ↓ G	+
A/duck/Victoria/76	P E I P K - - - - K R ↓ G	-
A/chicken/Victoria/76	P E I P K K K E - K R ↓ G	+
A/chicken/Victoria/1/85	P E I P K K R E - K R ↓ G	+
A/starling/Victoria/5156/85	P E I P K K R E - K R ↓ G	+

The cleavage site differs among the HA subtypes and determines the pathogenicity of the strain. Highly pathogenic strains (HPAI) contain polybasic residues at the HA cleavage site and HA is cleaved by intracellular subtilisin-like proteases, such as furin, found ubiquitously in the host organism thus enabling systemic infection of the host (Garten and Klenk, 2008). Low pathogenicity avian influenza (LPAI) contain monobasic cleavage site that is cleaved by trypsin-like serine proteases, found

mainly in the respiratory tract and either excreted from cells or bound to the plasma membrane, causing localized infection.

1.8 Influenza virus replication process

Influenza virus infection is initiated by the binding of viral haemagglutinin to the host receptor that contains sialic acid residues, which are present on glycoproteins and glycolipids (Skehel and Wiley, 2000). These sialic acid residues are found in different conformations (either α-2,3 or α-2,6 sialic acid linkage) and the ability of a virus to bind to one or both of them determines the tropism of the virus. Influenza viruses infecting avian species preferentially bind α-2,3 linked sialic acid, humans are infected by a virus binding to α-2,6 linked sialic acid, while swine influenza viruses recognize both (Steinhauer and Wharton, 1998; Suzuki et al., 2000). Mutations of the haemagglutinin receptor binding domain in as little as one amino acid can change the viral tropism and so enable the transmission of the virus to a new host (Matrosovich et al., 2000; Stevens et al., 2006).

After binding of the viral haemagglutinin to the host receptor, a virus is internalized by endocytosis, either clathrin-dependent or clathrin-independent (Lakadamyali et al., 2004; Lakadamyali et al., 2006; Sieczkarski and Whittaker, 2002; Chen and Zhuang, 2008), or macropinocytosis (De Vries et al., 2011; De Conto et al., 2011). Viral entry may require co-receptors that have not been identified yet (Chu and Whittaker, 2004; Stray and Cummings, 2000), and a signalling pathway including receptor tyrosine kinases (Sieczkarski et al., 2003; Eierhoff et al., 2010; Ehrhardt et al., 2006). Influenza viruses enter the host epithelial cells via the apical surface (Gottlieb et al., 1993) with the human influenza viruses entering non-ciliated airway epithelial cells with high levels of α-2,6-linked sialic acid, while avian influenza viruses show preference for ciliated cells with α-2,3-linked sialic acid (Matrosovich et al., 2004).

After the entry, the virus is trafficked through the endosomal network (Lakadamyali et al., 2006; Sieczkarski and Whittaker, 2003; Lakadamyali et al., 2003), where H+ ions are transferred into the virion through the M2 ion channels penetrating the viral membrane (Bui et al., 1996; Pinto et al., 1992). Due to the low pH, conformational changes in HA occur that expose the HA2 (fusion peptide), which enables the fusion of viral and endosomal membranes (Wiley and Skehel., 1987). The acidic environment of endosomes also triggers the release of viral ribonucleoproteins (vRNPs) from matrix M1 into the cytosol and their travel to the nucleus where viral transcription and replication take place (Kemler et al., 1994; Babcock et al., 2004). All proteins in the RNP structure (PB2, PB1, PA and NP) contain a nuclear localization signal (NLS) (Boulo et al., 2007) which bind to the proteins of cell's nuclear import machinery and enter the nucleus.

Because the genome of influenza viruses is made of negative sense strands of RNA it must first be converted to positive sense strands of RNA, which serve as a template to produce new viral negative strands of RNA. Viral RNA polymerase starts the replication of influenza genome without a primer and by utilizing cell's replication machinery. Viral mRNAs have a 3' poly(A) tail but no 5' cap when in the virion. However, viral mRNAs found in the host cell do have a 5' cap (Plotch et al., 1978; Krug et al., 1976). This 5' cap is cleaved off the cellular mRNAs by a "cap-snatching" mechanism of PA (Dias et al., 2009; Yuan et al., 2009). PA contains an endonuclease activity that cleaves the 5' cap from cellular mRNAs, which is used as a primer for viral transcription (Li et al., 2001). The products of viral transcription are positive sense mRNAs which are transported from the nucleus into the cytoplasm for translation on host ribosomes into viral proteins. Influenza glycoproteins HA and NA as well as M2 are transported to endoplasmic reticulum and Golgi apparatus for protein folding and glycosylation (Shors, 2013) followed by the transport to the cell membrane area where viral components are assembled into virions and released from host cell by budding. NA cleaves the glycosidic linkages of sialic acids from the cell receptors and newly synthesized virions in order to enable virions the exit from host cell by preventing viral particles to aggregate on the cell surface (Palese and Compans, 1976; Palese et al., 1974).

1.9 Influenza A viruses in swine

In 1918, the world experienced a Spanish influenza pandemic which first appeared in humans in March 1918 in Kansas, USA. In October of the same year, farmers in Iowa, USA noticed respiratory symptoms in pigs that were similar to human influenza disease (Table 1.3). Koen concluded that this was the same disease that he called 'swine influenza' (Koen, 1919). Koen's theory was rejected by most scientists at that time but in 1930 Shope proved this experimentally to be correct by extracting the mucus from a sick pig and using it to infect a healthy pig which later developed the same disease. This disease was developed also by pigs that Shope infected with the filtrate of the mucus from sick pigs suggesting that a pathogen causing the disease was not a bacterium but a virus (Shope, 1931). This first influenza virus isolate was A/swine/Iowa/15/30[H1N1]. Decades later many experiments have been conducted that proved the 1918 swine and 1930 swine viruses to be the same. In 2009 a study with several animal models was carried out by Weintgartl and colleagues which supported Koen's and Shope's observations of a human pandemic 1918 influenza virus being introduced into a susceptible swine population with the ability to infect pigs, replicate in them and spread to new porcine hosts (Weingartl et al., 2009). This swine influenza virus that first appeared in America was known as Classical swine H1N1 influenza virus. It remained antigenically very stable for approximately 80 years in sharp contrast to influenza viruses infecting humans (reviewed in Lee and Krilov, 2009) (Table 1.3).

Classical H1N1 virus spread from USA to European pigs where it became endemic until its replacement by Eurasian avian-like H1N1 virus in 1979 which caused the extinction of Classical H1N1 virus in Europe (Brown, 2000). Years before, in 1970, European swine experienced the introduction of human H3N2 virus and its establishment in pigs. The same year avian H3N2 virus crossed over to pigs in Asia. In 1980s and especially in 1990s novel swine influenza viruses appeared in pigs in North America, Europe and Asia due to reassortment of human, avian and swine influenza viruses (reviewed in Lee and Krilov, 2009).

In March 2009, a novel pandemic H1N1 virus appeared in humans and later spread to pigs (Table 1.3). Phylogenetic analysis showed this virus to be the reassortant between the 1998 triple reassortant swine H1N2 circulating in North American pigs, and Eurasian avian-like swine H1N1 virus endemic in European swine. The origin of gene segments of the progenitor triple reassortant H1N2 virus was avian (PB2 and PA), human (PB1) and swine (HA, NP, NS). The progenitor Eurasian avian-like H1N1 contributed NA and MP gene segments (Smith et al., 2009a).

Table 1.3

Comparison of evolutionary timeline of human and swine influenza viruses, 1918 – 2009 (adapted from Lee and Krilov, 2009).

YEAR	HUMAN VIRUSES	SWINE VIRUSES
1918 Spanish flu pandemic	1st known appearance and initial epidemic of human H1N1 in Kansas, March 1918. 2nd wave and true pandemic begins in August 1918.	1st known appearance of Classical swine H1N1 in Cedar Rapids, Iowa, October 1918.
1930		Classical swine H1N1 1st isolated by Shope and shown to be antigenically similar to human H1N1.
1933	Human H1N1 1st isolated by Smith, Andrews, and Laidlaw.	
1957 Asian flu pandemic	Human H1N1 disappears. Human H2N2 first appears (a human H1N1 and avian H2N2 reassortant).	
1968 Hong Kong flu pandemic	Human H3N2 appears (a human H2N2 and avian H3N2 reassortant).	
1970		Human H3N2 1st appears and becomes established in European swine. Avian H3N2 1st appears and becomes established in asian swine.
1974	1st documented case of swine influenza in a human – Classical swine H1N1 - Minnesota	
1976	A novel swine H1N1 outbreak in New Jersey, Jan.-Feb. only.	
1977 Russian flu	Human H1N1 reappears after 20 years, genetically close to a 1950 H1N1. Human H2N2 disappears.	
1979		Avian H1N1 1st appears and becomes established in European and Asian swine as swine (avian) H1N1.
1984		Swine H3N2 1st appears (a human H3N2 and swine (avian) H1N1 reassortment).
1994		Swine H1N2 1st appears (a human H1N1 + swine H3N2 reassortant) and becomes established in European and Asian swine.
1998		After 80 years, 1st new North American swine viruses appear: Triple reassortant swine H3N2 (Classical swine H1N1 + human H3N2 + unknown avian A virus). Reassortant swine H3N2 (Classical swine H1N1 + human H3N2). Reassortant swine H1N2 (Classical swine H1N1 + triple reassortant swine H3N2). Reassortant swine H1N1 (Classical swine H1N1 + triple reassortant swine H3N2). Reassortant swine H1N1 (Classical swine H1N1 + reassortant swine H1N2).
2009 Novel H1N1 pandemic	Novel H1N1 1st appears (a reassortant of reassortant swine H1N2 and a swine (avian) H1N1).	Pandemic H1N1 2009 virus spreads from people to pigs.

1.10 Influenza A viruses in humans

Early descriptions of influenza epidemics and pandemics (that were caused by influenza A type viruses) date back to 12th century (reviewed in Cunha, 2004). Historical records reveal that influenza epidemics appeared in 1173 (England, Germany, Italy), 1323 (Italy, France), 1387 (Florence), 1411 (Paris), 1414 (Paris, Italy), 1427 (Paris). The whole of Europe experienced influenza epidemics in 1510, 1557, 1580, 1761 and 1788 (reviewed in Cunha, 2004).

The first known pandemic outbreak occurred in 1580 in Europe which followed the 1580 epidemic, and spread to Asia and Africa. The subsequent pandemics appeared in 1729, 1732 and 1781. The 1781 pandemic was very severe occurring first in North America and spreading to South America and most of Europe. The pandemics of the 19th century appeared in 1830, 1833 and 1889. The 1889 pandemic originated in Russia and is therefore known as the Russian flu. Later this pandemic spread to Europe, North America, further to Latin America, Asia, New Zealand, Australia and Africa. One million people died as the result of the infection with this influenza strain (reviewed in Cunha, 2004).

The first influenza pandemic of the 20th century occurred in 1918 in North America and then spread to Europe with American soldiers during the World War I, and later to the rest of the world (Table 1.3). It is generally known as the Spanish flu due to public and detailed follow up of the disease in Spain (reviewed in Cunha, 2004). This pandemic was unusually severe with older sources estimating the total number of deaths between 40 and 50 million (Patterson and Pyle, 1991). The current estimation of lives lost is between 50 and 100 million (Barry, 2005). Phylogenetic analyses of influenza virus sequences obtained from fixed and frozen lung tissues of people who died during the 1918 pandemic (Taubenberger et al, 1997; Smith et al., 2009) suggest that the 1918 virus was generated by reassortment between human, swine and avian viruses but the exact origin remains ambiguous (Smith et al., 2009b).

In 1957, a novel pandemic strain H2N2 emerged in humans causing the first outbreak in China (Table 1.3). This virus was the result of a reasssortment between human influenza viruses and Eurasain avian viruses (Smith et al., 2009b). The origin of PB2, PA, NP, MP and NS segments was from human influenza viruses circulating prior to 1957 while the HA, NA and PB1 segments were from an avian source (Reid and Taubenberger, 2003; Smith et al., 2009b). The total death toll of this pandemic, also called Asian flu, was between one and four millions.

A decade later, in 1968, the world experienced a new pandemic caused by an H3N2 influenza virus (Table 1.3). This was a reassortant virus with HA and PB1 genes acquired from avian source while all the other gene segments were from the previous human H2N2 virus (Smith et al., 2009b). This pandemic is known as the Hong Kong flu and it killed around one million people worldwide.

In 1974, a first case of swine influenza in humans was documented proving that swine influenza viruses can infect humans (reviewed in Myers et al., 2007). Two years later, In January 1976, several soldiers at Fort Dix, USA complained of respiratory illness and in the beginning of February one of them died (Table 1.3). The cause of death was confirmed to be an infection with the H1N1 swine influenza virus closely related to the 1918 pandemic virus. As a result, a mass vaccination of people was carried out (Shellenbarger, 2009) which resulted in over 500 cases of Guillain-Barre syndrome affecting the peripheral nervous system and 25 deaths due to this complications. Consequently, the vaccination program was aborted. By that time, 24% of the USA population was vaccinated (Shellenbarger, 2009).

The next large outbreak of influenza in humans occurred in 1977 in northern China and Russia (Table 1.3). It affected mostly young people under the age of 25. Antigenic and molecular studies showed that this pandemic was caused by an H1N1 virus, very similar to those circulating in humans between 1947 and 1957 (Nakajima et al., 1977). Because influenza A viruses accumulate genetic changes rapidly, the above finding means, that the virus remained practically unchanged for 20 years which implies that the virus was most likely frozen and probably accidently or

intentionally released from a laboratory, although both Chinese and Russian scientists denied that (Nakajima et al., 1979).

H1N1 and H3N2 viruses have been co-circulating in a human population since 1977, causing seasonal epidemics. With the emergence of the pandemic H1N1 influenza virus in 2009, seasonal H1N1 lineage disappeared (Medina and Garcia-Sastre, 2011).

The first influenza pandemic of the 21st century began with the outbreak of influenza like illness in Mexico in March 2009 (Table 1.3). The virus was H1N1 and originated from a reassortment between human, avian and two swine lineages (Smith et al., 2009a). This pandemic caused death of approximately 285,500 people worldwide (Dawood et al., 2012).

1.11 Influenza A viruses in birds

Wild aquatic birds, particularly orders *Anseriformes* (ducks, geese, swan) and *Charadriiformes* (gulls, terns, surfbirds, sandpipers) are considered to be a natural reservoir of influenza A viruses. So far, sixteen HA subtypes and nine NA subtypes have been isolated from wild aquatic birds, allowing 144 possible HA-NA combinations (Horimoto and Kawaoka, 2005). However, only 103 different HA-NA combinations have been observed, some of them very rarely. In wild birds, influenza A viruses replicate in the epithelial cells of lower intestinal tract (Taubenberger and Kash, 2010). Transmission of influenza A viruses between wild birds and to domestic birds occurs via a fecal-oral route and through contaminated water (reviewed in Gibbs, 2010; Taubenberger and Kash, 2010). Influenza A viruses have been isolated from domestic birds such as chicken, turkey, goose, duck, quail and pheasant. In domestic poultry, influenza A viruses replicate in both, the respiratory and intestinal tract. While wild birds usually remain asymptomatic, domestic birds often develop respiratory illness (Taubenberger and Kash, 2010). Based on their virulence, avian influenza A viruses can be classified as highly pathogenic avian influenza viruses (HPAI) or low pathogenic avian influenza viruses (LPAI) (Kalthoff

et al., 2010; Klenk and Garten, 1994). High pathogenicity is due to cleavage site in HA between HA1 and HA2, rich in basic amino acid residues (Klenk and Garten, 1994; Stienekegrober et al., 1992). HPAI strains are usually arising in subtypes H5N1, H7N3, H7N7 and H7N9 (reviewed in Brown et al., 2006).

1.12 Influenza A viruses in horses

Horse influenza occurs globally and is caused by two main strains of influenza A: equine-1 (H7N7) and equine-2 (H3N8). Equine-2 (H3N8) causes more severe disease that can affect the heart muscle. The first equine influenza A virus isolate was obtained in 1956 (A/equine/Prague/56 [H7N7]) during an influenza pandemic in horses in East Europe (Webster, 1993). In 1963, an A/equine/Miami/63 [H3N8] was isolated in USA responsible for a large influenza epidemic among horses. In 1989, China experienced a severe epidemic caused by an H3N8 virus which was shown to be of avian origin and unrelated to the American H3N8 virus (Webster and Guo, 1991). Phylogenetic analyses of HA sequences have shown that equine influenza forms two distinct lineages, the North American and European (Daly et al., 1996) and the American lineage evolved into two sublineages (Lai et al., 2001). Due to transportation of horses, the viruses causing epidemic in one geographical area spread to another part of the world causing influenza disease outbreaks. For example, the outbreaks of equine influenza in Australia and Japan, both in 2007, were due to an infection with viruses circulating in USA (Elton and Bryant, 2011).

1.13 Influenza A viruses of other hosts

Influenza A viruses also infect marine mammals (seals and whales), minks, dogs and cats. Between 1979 and 1980, avian origin H7N7 influenza A virus caused a severe epidemic in seals in USA (Webster et al., 1981). Another outbreak of influenza in seals in the USA occurred between June 1982 and March 1983, which was caused by an avian origin H4N5 virus (Hinshaw et al., 1984; Wright et al., 2007).

H13N2 and H13N9 subtypes were isolated from a stranded whale (Hinshaw et al., 1986) and might have caused or contributed to its stranding (Wright et al., 2007).

Minks can be infected with human and avian influenza viruses (Klingeborn et al., 1985; Okazaki et al., 1983). H10N4 subtype was isolated from a dead mink in Sweden (Klingeborn et al., 1985; Wright et al., 2007).

Domestic cats and zoo tigers died of H5N1 after eating infected poultry (Enserink and Kaiser, 2004; Wright et al., 2007).

An outbreak of influenza caused by an equine H3N8 virus occurred in racing dogs in Florida in 2004 and spread to the general dog population (Crawford et al., 2005).

1.14 Transmission of influenza viruses between different hosts

Successful transmission of influenza A viruses between different hosts depends on several viral and host factors.

1.14.1 The role of viral surface glycoproteins

The HA protein is the major surface glycoprotein. It mediates two important functions in virus replication: binding to the appropriate receptor on the host cell and fusion between viral and endosomal membranes. HA of avian viruses preferentially binds to the N-acetylneuraminic acid α2,3-galactose (NeuAcα2,3Gal) form of sialic acid receptors and human viruses preferentially bind to NeuAcα2,6Gal sialic acid receptors. Different HA subtypes require various mutations at the receptor-binding domain to switch binding between α2,3 and α2,6-linked sialic acid receptors. For H2 and H3, mutations Gln226Leu and Gly228Ser are required to switch from avian to human binding specificity. In H1, the Glu190Asp is critical for adaptation to human-like receptor binding (Matrosovich et al., 2006b). For H5, contradictory findings have been recorded depending on the type of strain used in the study. However, similar mutations than those in H3 might play a role in higher binding ability to α2,6-

linked sialic acid and still retaining the ability to bind to avian-like receptors (Shinya et al., 2005; Stevens et al., 2006; Yamada et al., 2006). Avian-swine viruses isolated until 1984 from European pigs recognized both types of receptors, while viruses isolated after 1985 were strictly binding α2,6-linked sialic acid receptors. A Ser142Leu substitution in HA contributed to the loss of binding to α2,3-linked receptors (Ito et al., 1998) and mutations Glu190Asp and Gly225Glu were probably responsible for initial changes in receptor-binding specificity (Matrosovich et al., 2000).

Receptor affinity can also be modulated by glycosylation and sialyation of the HA head domain. Compared to viruses from aquatic birds and pigs, human H1 viruses are more glycosylated (Inkster et al., 1993). It has been postulated that this hyperglycosylation of the HA in combination with compensating NA stalk deletion, modifies the aquatic bird virus prior to developing virulence in chickens (Perdue et al., 1995; Matrosovich et al., 1999; Banks et al., 2000, 2001; Baigent and McCauley, 2003). The NA plays a minor role in determining host-range specificity, compared to HA. However, it is critical that NA's cleavage specificity is functionally compatible with HA's receptor-binding properties, along with the stalk length of the NA. Like HA's specificity, also NA can cleave sialic acid specifically in the form of α2,3 and/or α2,6 linkage. This specificity still needs to be characterized but it was shown that the specificity of NA (N8) is associated with amino acid sites 275 and 144 (Saito and Kawano, 1997). It is also known that when HA shows a smaller degree of affinity to its receptor, also NA's specific activity is reduced (Baigent and McCauley, 2003).

1.14.2 The role of viral ribonucleoproteins (RNP)

RNP's contribute to determining host-range specificity through three main mechanisms: (i) increasing genetic diversity of variants with high adaptation and transmission potential in the new host; (ii) interaction with host proteins that enables a suitable environment for efficient virus replication; (iii) generation of escape mutants that avode the innate immune system of the new host. Adaptation of the viral

variant to a new host through RNP occurs either directly in the new host, or in an intermediate host (Naffakh et al., 2008), same as for the surface glycoproteins. Due to strong physical and functional interaction of PB2, PB1, PA and NP, these proteins seem to share similar evolutionary pathways, which suggests co-evolution (Naffakh et al., 2008).

52 host-associated signatures were identified throughout the influenza viral genome (Chen et al., 2006) and 35 of them are found in the proteins of the RNP (2 in PB1, 8 in PB2, 10 in PA and 15 in NP). It was shown that not only HA and NA of the human origin are needed for a sufficient replication of a virus in mammals but also the internal segments (Naffakh et al., 2008). The most significant genetic signature for host adaptation is located at residue 627 of PB2.

Most human viruses harbour Lys at this position, while avian viruses usually contain Glu. In 1918 H1N1 virus, the PB2 was of avian origin but the amino acid at position 627 was Lys (Taubenberger et al., 2005). This substitution was also observed in many of the human H5N1 isolates and human H7N7 isolates from the fatal pneumonia cases in the 2003 Netherlands outbreak (Fouchier et al., 2004). However, swine and equine isolates of avian-like lineages retained the Glu at this position, suggesting less selective pressure for this residue in both species (Shinya et al., 2007). Also, the 2009 pandemic H1N1 virus which is thought to have originated from the swine influenza virus, contains Glu at this position (Chen and Shih, 2009). Beside its role in virulence and host range adaptation, PB2-627 plays a role in tissue tropism of the virus.

Human viruses that contain Lys at position 627 in PB2 replicate well at 33°C but avian viruses with the same substitution do not. Considering that the body temperature of birds is 41°C where avian viruses replicate, cold sensitivity of the polymerase proteins limits the replication of avian viruses in the upper respiratory tract of humans and thus limits human-to-human transmission (Naffakh et al., 2008).

The amino acid site 375 of PB1 seems important for the adaptation of avian viruses to a human host. Asn is present at this position in most avian species, while most human viruses have Ser at this same position. However, this position is not classified as a genetic signature because some human H3N2 viruses have Asn at this position and many avian isolates have Ser (Taubenberger et al., 2005).

In combination with Glu627Lys, other amino acid residues are associated with high replication and pathogenicity of one strain of H7N7 and many H5N1 isolates in mice and humans. These residues are PB2-701, PB2-714, PA-97, PA-615, PA-624 and NP-319 (Naffakh et al., 2008).

1.14.3 The role of host receptors

Infection of a host cell starts with the interaction of viral HA and host cell receptors. Specific receptors for influenza viruses are still unknown, however, sialic acid receptors are recognized as receptor determinants for those viruses. Sialic acids are nine-carbon acid sugars that are bound to the termini of oligosaccharide chains of glycolipids and glycoproteins, forming α-glycosidic linkage (Matrosovich et al., 2006b). Two main species are: N-acetylneuraminic acid (Neu5Ac) and N-glycolylneuraminic acid (Neu5Gc), which are bound to sugars in the form of α2,3- or α2,6-linkage (Matrosovich et al., 2006b).

The distribution of these receptors differs between hosts and even organs of the same species. Pekin ducks express α2,3-linked sialic acid in the intestine and in the tracheal epithelial cells. Both types of receptors are equally expressed in vascular endothelium and tubular cells in the kidney, endocardium and alveolar cells. Chickens also express mainly α2,3-linked sialic acid in the intestine, however, tracheal epithelial cells express 10 times as much α2,6-linked sialic acid as α2,3-linked sialic acid. Both types of receptors are expressed in different organs with the dominance of α2,6-linkage in the kidney vascular epithelial cells (Kuchipudi et al., 2009). Turkeys express both types of receptors on tracheal cells in approximately equal amounts, whereas intestinal walls harbour mostly the α2,3-linkage (Pillai et al.,

2009). Quails show the abundance of both types of receptors in both the trachea and the intestine (Wan and Perez, 2006).

Humans express α2,6-linkages on tracheal epithelial cells, while α2,3-linkages are expressed on alveolar pneumocytes and on the ocular and lacrimal duct epithelial cells, which could explain infections of the lungs with H5N1 HPAI viruses (Nelli et al., 2010; Nicholls et al., 2007; Chan et al., 2009) and conjunctivitis because of the infection with avian influenza viruses, especially H7 and H9 subtypes (Olofsson et al., 2005).

Pigs express large amounts of α2,6- and α2,3-linked sialic acid receptors on their tracheal epithelial cells (Nelli et al., 2010) and since they are considered as a mixing vessel, both avian and human as well as swine influenza viruses can infect epithelial cells, which can result in the generation of reassortant viruses.

1.14.4 Host immune system

Immunity against influenza is dependent on the presence of natural barriers that prevent virus replication, vaccination or pre-infection of the host with homologous and/or heterologous virus and genetic makeup of the host (Belser et al., 2010).

1.14.5 Environmental factors

Beside host- and virus-related factors, the transmission of influenza viruses depends strongly on environmental factors such as weather conditions, social organizations, host behaviour, stability of the virus in nature, mode of transmission and virus load (Weber and Stilianakis, 2008b).

1.15 The epidemiology of influenza A viruses

Influenza A viruses infect a wide variety of warm-blooded animals and humans. Human viruses circulate yearly in epidemics with a peak in winter months. Occasionally (every few decades), an antigenically novel strain appears as a pandemic virus. Influenza is estimated to kill 30,000 people annually in the United States alone. Every few years, additional 10,000-15,000 deaths occur due to influenza. When a pandemic occurs, it infects 20% to 40% of the population in a single year and usually raises death rates dramatically above normal levels. Human pandemic influenza A viruses emerged four times in the last 100 years: in 1918 (H1N1 subtype), in 1957 (H2N2), in 1968 (H3N2) and in 2009 (H1N1). How and when novel influenza emerge as pandemic strains and their precise mechanism of pathogenesis are still not understood (reviewed in Clark and Lynch, 2011).

Within influenza A virus type there are currently 18 different hemagglutinin and 11 different neuraminidase subtypes (Fouchier et al., 2005; Webster et al., 1992; Tong et al., 2012; Tong et al., 2013) but only three HA (H1, H2, H3) and two NA (N1, N2) have caused human epidemics, as defined by sustained, widespread, person-to-person transmission (Palese and Shaw, 2007).

1.16 The evolution of influenza A viruses

Influenza viruses evolve by three different mechanisms: (1) antigenic drift is enabled by viral RNA polymerase lacking the ‘proofreading’ mechanism therefore not being able to correct errors during replication (reviewed in Forrest and Webster, 2010). As a result, point mutations arise in the newly replicating viral genome which consist of insertions, deletions and substitutions. New viruses evolve relatively slowly by stepwise mutation and selection, (2) antigenic shift is a larger change of a viral genome and is associated with two different processes: (i) reassortment can occur when the same host cell is infected with at least two different influenza A viruses that exchange gene segments leading to the generation of a novel virus with a different

combination of gene segments (McHardy and Adams, 2009), (ii) direct transmission of an influenza virus from one host to another and its establishment in the new host, (3) recombination which can be non-homologous and homologous. Large amount of evidence exists for non-homologous recombination in influenza A viruses but less so for homologous recombination (reviewed in Forrest and Webster, 2010; reviewed in Cox and Subbarao, 2000).

Evolutionary rates are highest for influenza A virus compared to influenza virus types B and C. Within A type, avian influenza viruses evolve significantly more slowly than influenza A viruses of other hosts, both at the nucleotide and amino acid levels. The reason for that is attributed to wild birds being the reservoir of these viruses. The evolutionary rate is at equilibrium in wild aquatic birds, suggesting the optimal adaptation of viruses to these hosts. Even if amino acid substitutions occur they do not provide selective advantage so mutations do not always result in amino acid changes. Mammals and land-based poultry exhibit accumulation of mutations resulting in amino acid substitutions (Webster et al., 1992).

Evolutionary rate varies among segments in influenza A viruses. H3 HA are evolving faster than PA, PB1, PB2, NP and M1 genes. 57% of mutations in H3 HA are silent, whereas in PB2 gene more than 90% are silent (Webster et al., 1992). H3 HA evolves with a rate of 4 x 10-3 nucleotide substitutions per site per year and HA1 part of the H3 HA evolves with the rate of 5 x 10-3 nucleotide substitutions per site per year (Webster et al., 1992; Hay et al., 2001). Mutations in HA are very important for influenza viruses because they confer antigenic variation and enable the virus to escape the host's immune system. The rates of mutation accumulation in H1N1 viruses seem to be similar to evolutionary rates of H3 HA (Hay et al., 2001).

PB2 segment codes for an internal protein, therefore only a limited level of immune pressure is imposed on this protein. There are also structural constraints limiting PB2 from genetic change. M1 and M2 proteins are both encoded in the same gene segment and under different selective pressures: in M1, 96% of mutations are silent, in M2 only 66% (Webster et al., 1992). The M2 is therefore under strong selective

pressure, while M1 is well adapted to mammalian hosts. The segment encoding NS1 and NS2 proteins also exhibits differences in evolutionary rate among these two genes (Webster et al., 1992).

Chapter 2

Methods

2 Methods

In this work, influenza evolution was studied by employing bioinformatic and phylogenetic methods.

2.1 Bioinformatics

Bioinformatics is an interdisciplinary scientific field, combining life sciences with computational biology with the aim of developing methods for retrieving, storing, organizing and analyzing biological data (Ouzounis, 2012).

Bioinformatics is a young scientific discipline reaching back only approximately 20 years. Between 1996 and 2001, databases with biological data (nucleotide sequences, molecular structures, expression profiles, full genomes) and key computational algorithms were in place and increasing in size (Ouzounis and Valencia, 2003). Therefore, the scientific community and industry saw opportunities in combining experimental work with computational research (Hatzimanikatis, 2000; Palsson, 2000; O'Donnell, 2000; Reed, 2000). The expectations of this new field were great and as a result funding of bioinformatics substantially increased (Aldridge, 2001) and also was covered extensively by media (Ouzounis, 2012). Between 2002 and 2006, bioinformatics was making deeper and intertwined connections with science, for example, it introduced the terms of personalized medicine (Gurwitz et al., 2006) and synthetic biology (de Lorenzo et al., 2006) and paved the way of robotics and automation platforms into medicine (Ilyin et al., 2004; Ritchie, 2005). Because of the latter, bioinformatics became a matter of discussion from the standpoint of public health, ethics, law, education and social issues (Maojo and Martin-Sanchez, 2004; da Fontoura Costa, 2004). Since 2007, bioinformatics became a part of life sciences, such as biodiversity conservation planning (Faith and Baker, 2007), network biology (gene and protein interaction networks) (Gatenby and Frieden, 2007), cancer studies (Hanauer et al., 2007), genomic medicine (Butte, 2008), drug discovery (Chen and Chen, 2008), biomarker discovery (Simpson et al., 2008), literature mining

(Krallinger et al., 2008), genome assembly (Pop, 2009), protein design (Suarez and Jaramillo, 2009), metagenomics (Kyrpides, 2009), infectious diseases (Berglund et al., 2009), phenotyping (Thorisson et al., 2009) and others. Future challenges for bioinformatics remain in the fields of food production, fuels and energy sources, materials, health and environment (Ouzounis, 2012).

2.1.1 R programming language

R is a programming language that can be freely downloaded and used for statistical computing and graphics production. It is widely used for developing statistical software (Fox and Andersen, 2005; Vance, 2009) and data analysis (Vance, 2009). R can be used with a command line interface; but several graphical user interfaces are also available.

The source code for R is written in C, Fortran, S and R. Users with advanced knowledge can use code written in different programming languages such as C, C++ (Eddelbuettel and Francois, 2011), Java (Temple Lang, 2010) and Pyton for manipulating R objects directly. But also many of R's functions are written in R itself. Many user-written packages for specific functions are submitted to the R project (Anonim, 2013) and are subsequently available for download as add on packages. Advantages of R are its inbuilt functions and its static graphics that produces high quality figures, suitable for publication.

Throughout this work, R has been used to perform initial sequence processing, statistical analyses, and generation of selected figures. Additional packages including the ape phylogenetics package (Paradis et al., 2004) and random forest machine learning package (Liaw and Wiener, 2002) have been employed for more specialised analyses and their use is described in chapters Amantadine resistance in swine influenza A viruses and Oseltamivir resistance in human seasonal A/H1N1 influenza viruses.

2.2 Phylogenetics

Phylogenetics is a scientific field that studies the evolutionary relationships among organisms or genes (Maser et al., 2001), demographic changes and migration patterns of species (Edwards, 2009), the origin, transmission and epidemiology of pathogens (Marra et al., 2003; Grenfell et al., 2004) and other (Gray et al., 2009; Paten et al., 2008; Ma, 2011). The input for phylogenetic studies is most often molecular sequencing data but also morphological data matrices. The result of phylogenetic studies is a phylogeny, which is a hypothesis about the evolutionary history of taxonomic units. Phylogeny is represented graphically by a phylogenetic tree (Strimmer and von Haeseler, 2009).

2.2.1 Phylogenetic tree

A phylogenetic tree is a diagram, showing the inferred evolutionary relationships among different biological species or among different strains of the same species and it is based upon similarities and differences in their physical or most commonly their genetic characteristics. This diagram is called a phylogenetic tree because it resembles the structure of a tree and also for referring to the different parts of the diagram as a root, branch, node and a leaf. The phylogenetic tree has several parts. The **taxa** represent biological species and **nodes** represent the most recent common ancestor of those taxa. The taxa joined together are descended from a common ancestor. Taxa are usually represented as tips of the **branch**. Branches connect nodes and represent genetic relationships with the length of the branch usually representing some measure of evolutionary change. The branches determine tree's **topology**. A group of taxa with the same common ancestor (i.e. A, B, C on Figure 2.1) are called a **cluster** or **clade** and have a **monophyletic origin**. Internal nodes are generally called hypothetical taxonomic units, as they cannot be directly observed. External nodes are called **tips** or **leaves** and represent actual viruses that have been sampled and their sequences determined by sequencing. The numbers associated with each node, represent a measure for the support for the node, which is usually between 0 and 1 (but also given as percentages) where 1 means maximal support. These values

are calculated by using bootstrapping (section 2.2.5) and Bayesian posterior probabilities (2.2.6) (Rambaut, 2013, Strimmer and von Haeseler, 2009).

Phylogenetic trees can be rooted or unrooted. Rooted trees have a node, called a **root**, which represents a common ancestor of all taxa shown in a tree. The root node joins the outgroup (i.e. H and I in Figure 2.1) and ingroup taxa (all taxa but the outgroup) and therefore represents their common ancestor. Rooted trees are therefore directional, since all taxa evolved from the root. This outgroup is phylogenetically close enough to all the taxa but far enough to be a clear outgroup. Unrooted trees show the relatedness of the taxa without making assumptions about ancestry. While unrooted trees can always be generated from rooted ones by simply omitting the root, a root cannot be inferred from an unrooted tree without some means of identifying ancestry; this is normally done by including an outgroup in the input data (Theobald, 2012, Strimmer and von Haeseler, 2009).

Figure 2.1 shows a scaled phylogenetic tree, where the length of horizontal lines gives the amount of genetic change. The longer the horizontal line (branch), the larger the amount of change. The **scale** for the amount of change is found at the bottom of the phylogenetic tree (Figure 2.1), where the number 0.04 and the bar underneath show the length of the branch that represents an amount of genetic change of 0.04. The unit for this value, which is also the unit of the branch length, are usually nucleotide substitutions per site. This value is calculated from dividing the number of changes (substitutions) by the length of the sequence. The vertical lines carry no information, they are just connecting the horizontal lines (Rambaut, 2013).

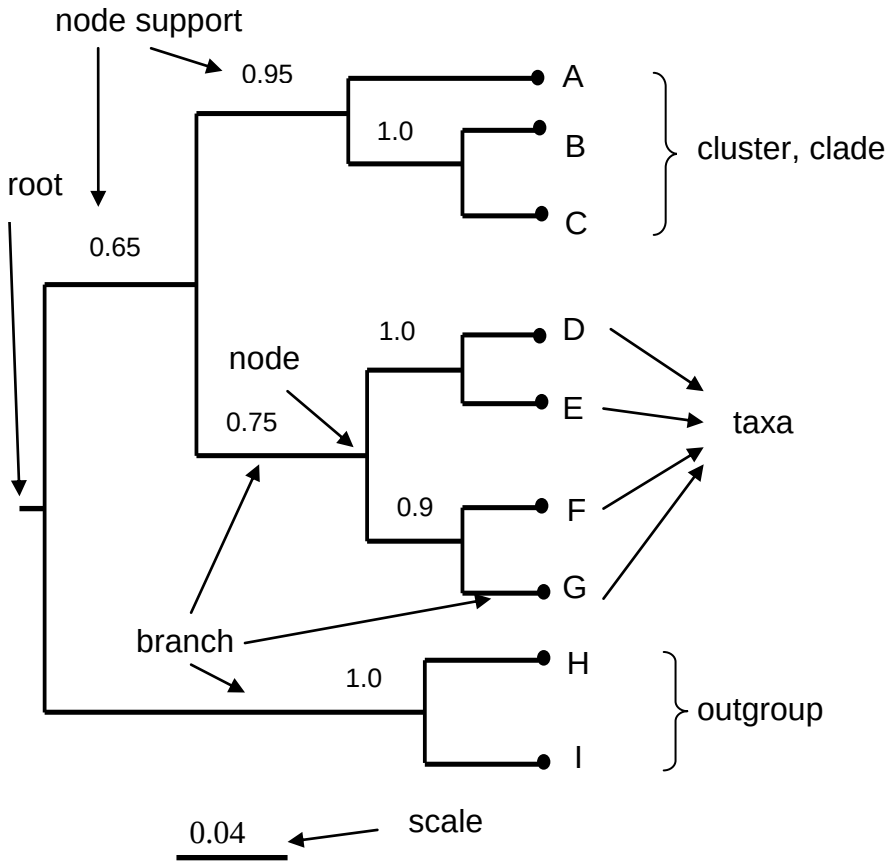

Figure 2.1

A structure of a phylogenetic tree. Taxa are the biological species or their genes and are represented as tips of the branches. Branches show the amount of genetic change and are connected with nodes, which present the common ancestor of a group of taxa, called a (monophyletic) clade or cluster. The root is a common ancestor of every taxonomic unit in the tree. The outgroup is a taxonomic unit for inferring rooted trees. The scale is used to measure the amount of genetic change in branches. The numbers at nodes represent support for nodes. Figure made by the author.

2.2.2 The coalescent theory

Coalescent theory is a model of the distribution of alleles that are shared by every individual within the population with the attempt to trace the origin of these alleles retrospectively (back through time) to a single ancestral copy. A phylogenetic tree helps trace the alleles back through different lineages to see where they come back together (coalesce). The point where alleles coalesce is called the most recent common ancestor. For simplicity, the coalescent theory assumes that there is no gene

flow, no natural selection and no recombination. However, there are advanced methods available that incorporate coalescent theory and include recombination, selection and population demographic models (Arenas and Posada, 2007).

2.2.3 A measure of evolutionary pressure, dN/dS versus dN-dS

Evolutionary pressure on nucleotide sites encoding proteins can be quantified by dN/dS, which is the ratio of the substitution rate at silent sites (dS) and the ratio of substitution rate at non-silent sites (dN). Silent sites are the ones where substitution does not result in a change of an amino acid in a protein and non-silent sites are the ones where the substitution causes the change in amino acid sequence. dN/dS can be larger than 1 and is a hallmark of positive selection acting upon site(s). dN/dS can be equal to 1 and that indicates neutral selection or in other words, no selection. If dN/dS is smaller than 1, then that is a sign of a purifying (stabilizing) selection (Kryazhimskiy and Plotkin, 2008).

Another measure of the amount of selective pressure acting upon nucleotide sites is dN-dS. This is the difference between the rate of non-synonymous substitutions at non-silent sites and the rate of synonymous substitutions at silent sites. dN-dS value can be positive or negative. If positive, it means that there are more non-synonymous mutations than synonymous at a particular site. If negative, the reverse is true.

2.2.4 Models of sequence evolution

Different models of DNA sequence evolution have been proposed. They differ in the number of free parameters they use to describe the rate of one type of nucleotide substituting another during evolution. Earliest models that are also the simplest have none or a few free parameters. It is desirable to reduce the number of free parameters because they are often unknown and have to be estimated from the data (Strimmer and von Haeseler, 2009).

Nucleotide substitutions can be divided into two groups: transversions (Tv) and transitions (Ts). Transversions are substitutions where a purine is substituted by a pyrimidine or vice versa (A↔C, A↔T, G↔C, G↔T) (Fig 2.2). Transitions are substitutions of a purine with a purine (A↔G, Ts_R) and a pyrimidine with a pyrimidine (C↔T, Ts_Y) (Figure 2.2).

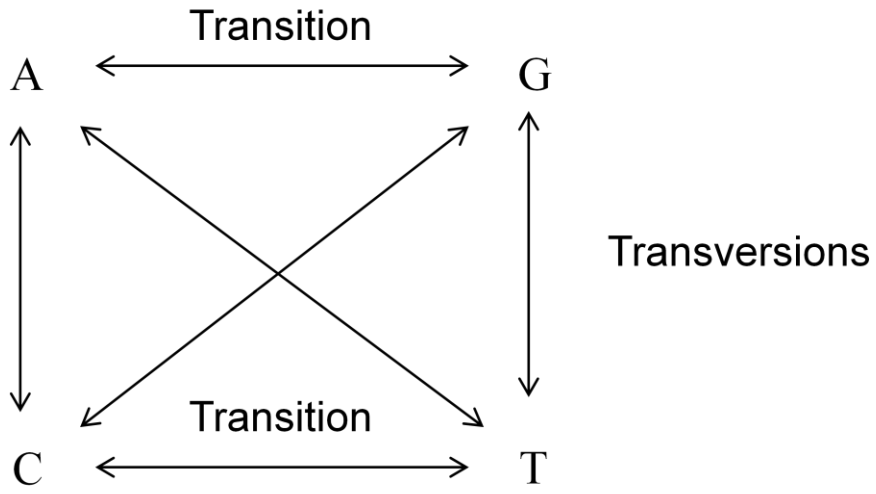

Figure 2.2

The six possible nucleotide substitution processes used in models of DNA evolution: transitions (substitutions of a purine with a purine and substitutions of a pyrimidine with a pyrimidine) and transversions (substitutions of a purine with a pyrimidine and substitutions of a pyrimidine with a purine) (adapted from Strimmer and von Haeseler, 2009).

The simplest substitution model is **JC69 model** proposed by Jukes and Cantor in 1969 with no free parameters. It assumes equal nucleotide frequencies and equal mutation rates (Strimmer and von Haeseler, 2009).

Nucleotide frequencies (π):

$$\pi A = \pi G = \pi C = \pi T = \frac{1}{4}$$

Mutation rates (μ):

$$\mu A = \mu G = \mu C = \mu T = \frac{1}{4}$$

K80 model was proposed by Kimura in 1980 and it assumes that frequencies of nucleotides are equal just like in the JC69 model.

Nucleotide frequencies (π):

$$\pi A = \pi G = \pi C = \pi T = \frac{1}{4}$$

However, K80 model distinguishes between transitions and transversions rates that were first denoted as α and β, respectively. Later, the rate of transversions was set to 1 and the ratio of transition/transversion rate was denoted as $_K$ which is the only free parameter of the K80 model (Strimmer and von Haeseler, 2009).

In 1981 Felsenstein proposed the **F81 model**. This model is an extension of the JC69 model and it assumes that nucleotide frequencies (π) are not equal.

$$\pi A \neq \pi G \neq \pi C \neq \pi T$$

F81 model has three free parameters (π_A, π_G, π_C) (Strimmer and von Haeseler, 2009).

Hasegawa, Kishino and Yano proposed the **HKY85 model** in 1985. This model combines some features of K80 model and F81 model. It recognizes the ratio of transitions and transversions as $_K$ (like the K80 model) and it assumes unequal nucleotide frequencies (like F81 model). The number of free parameters is four (transition/tansversion rate ratio and three nucleotide frequencies) (Strimmer and von Haeseler, 2009).

Tamura and Nei proposed the **TN93 model** in 1993. This model assumes that transitions A↔G have a different rate compared to transitions C↔T. All types of transversions are assumed to occur at the same rate, which is allowed to be different to the rates of transitions. This model also assumes unequal nucleotide frequencies. This model has five free parameters (transition/transversion rate, purine transition rate/pyrimidine transition rate, three nucleotide frequencies) (Strimmer and von Haeseler, 2009).

GTR model (General time-reversible model) was proposed by Tavaré in 1986. This model assumes that all six types of substitutions occur at a different rate.

$$\mu A \leftrightarrow G \neq \mu C \leftrightarrow T \neq \mu A \leftrightarrow C \neq \mu G \leftrightarrow T \neq \mu A \leftrightarrow T \neq \mu G \leftrightarrow C$$

GTR model also assumes that nucleotide frequencies (π) are different.

$$\pi A \neq \pi G \neq \pi C \neq \pi T$$

This model has eight free parameters (five substitution rates and three nucleotide frequencies) (Strimmer and von Haeseler, 2009).

Shapiro, Rambaut and Drummond proposed the **SRD06 model** in 2006. This model refers to a particular combination of models used for analyzing nucleotide sequences, encoding proteins. SRD06 partitions the alignment into 1^{st}+2^{nd} codon positions and 3^{rd} codon positions and then applies independent HKY85+gamma substitution models to those two partitions. SRD06 inherited different nucleotide frequencies and a transitions/transversions ratio parameter from the HKY85 model (Shapiro et al., 2006). Gamma distribution is used to model the nucleotide substitution rate heterogeneity over sites.

2.2.5 Methods for constructing phylogenetic trees

The **maximum parsimony** tries to find a topology of a tree that can be explained by smallest amount of character changes (i.e. substitutions), where each position in the aligned sequences is a character and the nucleotide or amino acid at that position is a state (Vandamme, 2009). Therefore, the parsimony criterion states that the best tree describing the data is the tree that supposes the least evolutionary change to explain observed data (hence maximally parsimonious).

The disadvantage of maximum parsimony is that parsimony only works well if rates of evolution are slow and branches are short (Felsenstein 2004; Felsenstein 1981a; Li 1997). When rates of evolution are high, or when some branches are very long, or

when the number of possible character states is limited, the output tree is not a true tree. This is often true for nucleotide sequences, which have only four possible character states (A, C, T, or G). In cases such as these, other phylogenetic methods can be more accurate than parsimony (Theobald, 2012).

Another commonly used method for determining phylogenetic trees is **maximum likelihood** (ML) (Edwards and Cavalli-Sforza 19643; Felsenstein 1981b; Fisher 1912). ML finds the tree and evolutionary parameters that produce the observed data with the highest probability. In other words, ML calculates the probability of a tree on a given set of data, using a specific evolutionary model. The supposition is that a topology with a higher probability of reaching the observed state is preferred to a topology with a lower probability. The method searches for the tree with the highest probability or likelihood. The disadvantage of ML is that with the increasing number of taxa included in the dataset, the number of possible trees as well as the computing time grow exponentially, therefore an exhaustive search of all possible tree topologies is not possible (Theobald, 2012,Vandamme, 2009).

PhyML is a software for inferring maximum likelihood phylogenies from nucleotide or amino acid datasets. Since its first introduction in 2003 (Guindon and Gascuel, 2003), this software has been developed further to improve its performance and reduce the computation time needed. Older version of Phyml used NNI (nearest neighbour interchange) which exchanges two subtrees that are connected by a single edge (Guindon and Gascuel, 2003). However, NNI tree searches sometimes become trapped in suboptimal maxima of the likelihood function, therefore SPR (sub tree prune and regraft) moves were introduced to PhyML (Hordijk and Gascuel, 2005) that explore the space of tree topologies. The second implementation of PhyML was the introduction of the approximate likelihood-ratio test (aLRT) (Anisimova and Gascuel, 2006). This approach is combined with NNI moves and compares each subtree configuration around the internal branch of interest to two alternative configurations that are defined by NNI moves around that branch. Then it computes the likelihood values of these two alternative topologies and performs an aLRT test, which is based on the logarithm of the ratio between the likelihood value of that tree

and the likelihood value of the best alternative tree (Guindon et al., 2010). With this upgrade, PhyML can resolve trees' phylogenies very accurately (depending on various parameters chosen by the user) but can be time consuming and also, the input file is preferred to be between 3 and 500 sequences of less than 2,000 characters long.

In this PhD, PhyML was used in only one research chapter (Amantadine resistance in swine influenza A viruses) to analyze a relatively large dataset of MP sequences' alignment from three different hosts.

RAxML (Randomized Accelerated Maximum Likelihood) is a software suitable for inferring phylogenies of very large datasets (several thousand sequences) under maximum likelihood. This program encompasses four different ways for obtaining bootstrap support and also a bootstopping option (Pattengale et al., 2010) that enables RAxML to automatically determine the number of bootstrap replicates needed for a stable support value (Stamatakis, 2014). RAxML is able to correctly analyze not only nucleotide and amino acid datasets but also binary data, morphological data with multiple states, RNA secondary structure data and data containing single nucleotide polymorphisms (SNPs). RAxML can also be used to assess which regions of a gene in question show a strong phylogenetic signal and these regions can therefore be used for amplification (Stamatakis, 2014). There are also other improvements of PhyML (parallel versions, post-analysis of trees, vector intrinsics, memory saving) (Stamatakis et al., 2014).

In this PhD, RAxML was used only in one research chapter (Amantadine resistance in swine influenza A viruses) for analysis of a very large dataset of MP sequences.

All **distance methods** transform character data into a matrix of pairwise distances, one distance for each possible pairing of the taxa under study. Distance methods are used almost exclusively with molecular data. Several of the distance methods are guaranteed mathematically to converge on the correct tree as more data is included

(Theobald, 2012). The most simple distance matrix is merely the number of character differences between two taxa, such as the number of nucleotide differences between two DNA sequences (Theobald, 2012).

The ***neighbor-joining*** (NJ) algorithm takes as input a distance matrix specifying the distance between each pair of taxa. The algorithm starts with a completely unresolved tree, whose topology corresponds to that of a star network, and iterates over the following steps until the tree is completely resolved and all branch lengths are known. NJ is joining the pairs of neighbouring taxa connected by a single interior node until every possible taxa pair is connected to yield the shortest tree. So, NJ joins two taxa by introducing the shortest possible internal branch. Optimal topology of the tree is the one that minimizes the tree length, when summing the branch lengths together (Theobald, 2012; Vandamme, 2009).

In this PhD, NJ was used in the chapter Reassortment in swine influenza A viruses for constructing phylogenies of all eight segments of swine influenza A viruses separately and comparing them against each other in order to assess for reassortment among segments.

Bootstrapping is a statistical method for assessing the reliability of the branches in a phylogenetic tree (Felsenstein 1985). In other words, bootstrapping analysis gives a way to judge the strength of support for clades on phylogenetic trees. In a bootstrap analysis, a fictional dataset is created by sampling data randomly with replacement from the real dataset until a new dataset is created of the same size. That means that some data will not be included in the bootstrap sample at all, while others will be included once or more. This process is done repeatedly (hundreds or thousands of times), and the parameter of interest is estimated from each fictional dataset. Then, for each bootstrap data set, a tree is constructed and the percentage of a certain clade in a tree is calculated (Felsenstein 1985). These bootstrapped phylogenies will likely have different topologies. The parts of the bootstrapped trees that are in common are ascribed a high confidence, while the parts that vary extensively are assigned a low confidence. A number is written by a node, which reflects the percentage of

bootstrap trees which also resolve a particular clade (Efron et al., 1996, Theobald, 2012,Vandamme, 2009).

2.2.6 Bayesian inference methods

Markov chain is a collection of random variables with the probability distribution of states where the state of the next random variable in the sequence, given the past and current states, depends only on the current state. **Monte Carlo** methods are computational algorithms (simply sets of instructions) for randomly sampling from a probability distribution. They are a form of computer simulation of some mathematical or physical process. The Monte Carlo expression comes from the analogy between a casino and random number generation. These two concepts have been put together (MCMC) for solving problems in Bayesian inference and elsewhere. MCMC first constructs a Markov chain of random variables which converges to the desired probability distribution after a certain (large) number of steps. The state of the chain is then used as a sample from this distribution (Ronquist et al., 2009).

Using MCMC, parameters of models are estimated and the likelihood of the model given the data is calculated. Then new parameters are proposed and the likelihood is calculated again. If the new likelihood is better (lower value), the new parameters are accepted but if the new likelihood is worse, then the new parameters are also accepted with a probability dependent on the ratio of the new and old likelihoods. This is how a chain of accepted parameters and their corresponding likelihoods is made. At the start of the chain, the likelihood of the models and its parameter estimates are still growing (improving) therefore are not in the equilibrium yet and are therefore discarded as burn-in. When the chain reaches the equilibrium, the distribution of parameter estimates reaches approximately the true distributions. So taking a sample of the parameters and posterior likelihoods after removing the burn-in period, gives an approximate of the true values.

BEAST (Bayesian Evolutionary Analysis Sampling Trees) is an advanced method for testing evolutionary hypotheses or constructing phylogenies. It incorporates the coalescent theory. BEAST includes several program packages for Bayesian MCMC analysis of molecular sequences. It enables the construction of rooted, time-measured phylogenetic trees with an inferred molecular clock (strict or relaxed). It also enables the user to choose the appropriate nucleotide substitution model and population demographic model (Drummond and Rambaut, 2007).

BEAST is based on Bayesian Markov chain Monte Carlo (MCMC) method for phylogenetic reconstruction. MCMC is a stochastic algorithm that produces estimates of a target distribution of choice, based on a sample. Stochastic process is a process, which is independent of a previous process (e.g. DNA substitution) and is in a particular state for a random amount of time before transiting to another state. In BEAST, the target distribution is the posterior distribution of a set of evolutionary parameters on a given set of molecular sequences. The most noted feature of BEAST is a resulting set of rooted trees incorporating a time scale. This is achieved by modelling the rate of molecular evolution on each branch of the tree. In the simplest case, this rate can be uniform across the tree, where the rate is known in advance or estimated from calibration information. BEAST was the first to introduce the relaxed molecular clock that does not assume a constant rate across lineages (Sanderson, 1997; Thorne et al., 1998; Rambaut and Bronham, 1998; Yoder and Yang, 2000; Kishino et al., 2001; Sanderson, 2002; Thorne and Kishino, 2002; Aris-Brosou and Yang, 2003), which is one of the biggest advances in molecular phylogenetics (Drummond and Rambaut, 2007).

Another big advantage of BEAST is a single coherent framework of different evolutionary models that complement each other. These models are nucleotide substitution models, demographic models, tree shape priors, molecular clock models and others. This means that simpler model components have been brought together to construct a complex evolutionary model in order to avoid oversimplifying, which is still present in many evolutionary analysis packages. It also enables the user to tailor

the analysis according to its own specific set of questions by choosing model specifications (Drummond and Rambaut, 2007).

In this PhD, BEAST was used in all three research chapters. In the chapter Reassortment in swine influenza A viruses, BEAST was used to construct a set of 1000 trees for each segment for each subtype studied and the topology of those trees was further compared in R to calculate the topology distance and so estimate the reassortment among segments. In the chapter Amantadine resistance in swine influenza A viruses, BEAST was used to determine the origin (time and host) of this resistance in swine influenza viruses. In the chapter Oseltamivir resistance in human seasonal influenza A/H1N1 viruses, BEAST was used to infer the time of oseltamivir resistant strains appearance and the relationship among resistant strains.

MrBayes is a program for constructing phylogenetic trees based on Bayesian phylogeny using a variant of MCMC called Metropolis coupled Markov chain Monte Carlo (MC^3) (Geyer, 1991). It has first been introduced in 2001 and implemented since then. MrBayes has a command-line interface and reads in an alignment of nucleotides or amino acids. The user can choose prior information such as the nucleotide substitution model and molecular clock. The output options are samples of ancestral states, site rates, site dN/dS ratios, branch rates and node dates. Many of the statistics can be output for visualization in FigTree (Rambaut, 2007) or other software (Huelsenbeck and Ronquist, 2001; Ronquist et al., 2012).

In this PhD, MrBayes has been used initially to construct trees based on alignments of influenza segments' DNA. However, MrBayes failed to construct trees for several datasets therefore it has not been used further.

2.2.7 Methods for detecting reassortment

Reassortment can be detected based on the topology of taxa on a phylogenetic tree. For example, if one taxon from dataset X has reassorted, it is found elsewhere on a tree for the dataset Y (Figure 2.3).

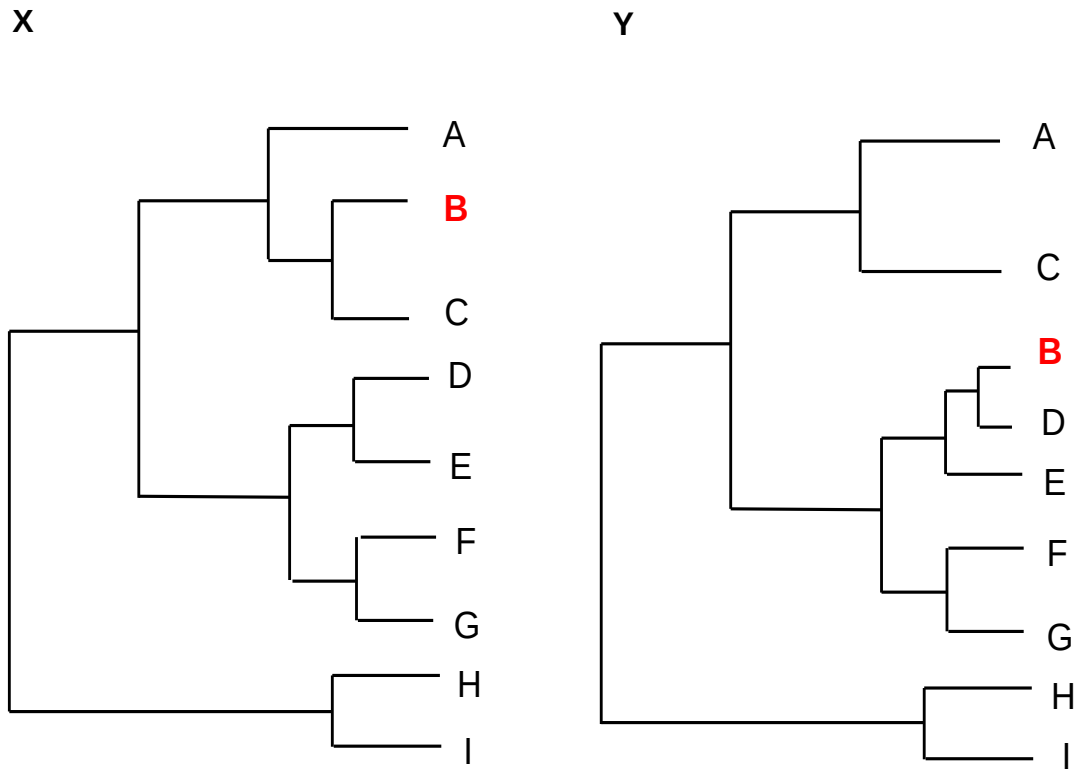

Figure 2.3

Two phylogenetic trees (one for dataset X and one for dataset Y) with identical topologies except for one taxon (B, in red) where reassortment occurred. Figure made by the author.

Kosakovsky Pond et al (2006) developed **GARD** (Genetic Algorithm Recombination Detection) which is a novel method for recombination detection in either biological or simulated sequences. The method is likelihood-based and uses a genetic algorithm that searches multiple sequence alignments for evidence of recombination. GARD is a robust method that screens multiple sequence alignments in order to find the evidence of phylogenetic incongruence. It also identifies the number and location of recombination breakpoints and sequences involved in putative recombination events. The advantage of this method is that it can be run in parallel on a cluster of computers and so it can screen for recombination in large datasets, it has better power and accuracy compared to other methods for recombination detection and it also gives good statistical properties for methods aimed at detecting positive selection. Because recombination is quite widespread in certain viruses (HIV-1, influenza) it can substantially affect the power and accuracy of molecular

evolutionary analyses tools, such as phylogenetic reconstruction (Posada and Crandall, 2002), molecular clock inference (Schierup and Hein, 2000) and the detection of positively selected sites (Shriner et al., 2003). Therefore, methods that can detect recombination events are an important part of every phylogenetic analysis. GARD has been implemented as language scripts in HyPhy, enabled to run in an MPI environment (Kosakovsky Pond et al., 2005). GARD can be freely accessed on: http://www.datamonkey.org/GARD/ and enables the user to: (i) upload an alignment of sequences to screen. The number of sequences and their length is periodically increased; (ii) select an appropriate model of nucleotide evolution (Kosakovsky Pond and Frost, 2005) and the distribution for modelling site-to-site variation in substitution rates; (iii) run GARD that screens for recombination; (iv) visualize and download the results of analysis, which include the number and location of breakpoints, improvement in AIC score, phylogenetic trees based on each non-recombinant breakpoint and other; (v) result files and also scripts for running the analyses can be downloaded and run locally (Kosakovsky Pond et al., 2006).

In this PhD, GARD was used in the chapter Reassortment in swine influenza A viruses in order to detect if there is any reassortment among all pairs of segments (concatenated sequences) and also for quantifying the amount of reassortment found.

Nagarajan and Kingsford (2011) presented a method, called **GiRaF** (Graph-incompatibility-based Reassortment Finder), that can find reassortments in a given collection of sequences by using data-mining techniques. GiRaF can identify the set of isolates arising from a reassortment. The method is based on comparing distributions of trees by constructing an ‘incompatibility graph’ and mining it for phylogenetic differences. Results of the authors of Giraf show that GiRaF can identify precisely recent reassortments as well as complex reassortment histories. GiRaF can efficiently analyze large datasets (Nagarajan and Kingsford, 2011).

In this PhD, GiRaF was used initially for detecting reassortment in influenza sequences. However, the results of this method were difficult to interpret therefore they were not included in my thesis.

2.2.8 Machine learning methods

CART (Classification and Regression Tree Analysis) methodology was first introduced in 1984 by Breiman et al. The term CART refers to two types of decision trees that can be constructed with this method. First, **classification trees** where the variables are categorical and the tree helps to identify a class within which a target variable falls into. Second, **regression trees** where the variables are continuous and the tree helps to predict their values. CART is used for constructing prediction models from data. These models are obtained by partitioning the data in a binary recursive way. The term binary refers to the fact that each group of variables, represented by a **node** in a decision tree, can only be split into two groups. So, each node (called parental node) can be split into two further nodes (called child nodes). The term recursive refers to the fact that this process of binary partitioning can be repeated over and over again. The data is split into two parts, one part has samples with a particular variable larger or equal to a threshold, and the other has samples which have values smaller than a threshold. Thus, a parental node gives rise to two child nodes and these two child nodes each give rise to two further child nodes. The larger the distance between the nodes (depth) the more significant is the split (Figure 2.4). The term partitioning refers to the dataset being split into parts or partitioned (Lewis, 2000). The partitioning is completed when is no longer possible to continue the tree building process. That is usually when there is only one value in each of the child nodes. Each value of the target variable at the end of the node represents a **leaf** (Fig 2.4). The tree that is created is usually overfit. That means that this tree fits every value of the variable in the dataset used to create that tree. So, this tree would not fit other independent datasets (Lewis, 2000). The parts of the decision tree that are more likely to represent an overfit are later parts of the tree. Therefore, tree pruning has been introduced into the CART methodology. It begins at the last level (the terminal nodes) which is pruned away. As more and more nodes are pruned away, the decision tree is becoming simpler and simpler (Lewis, 2000).

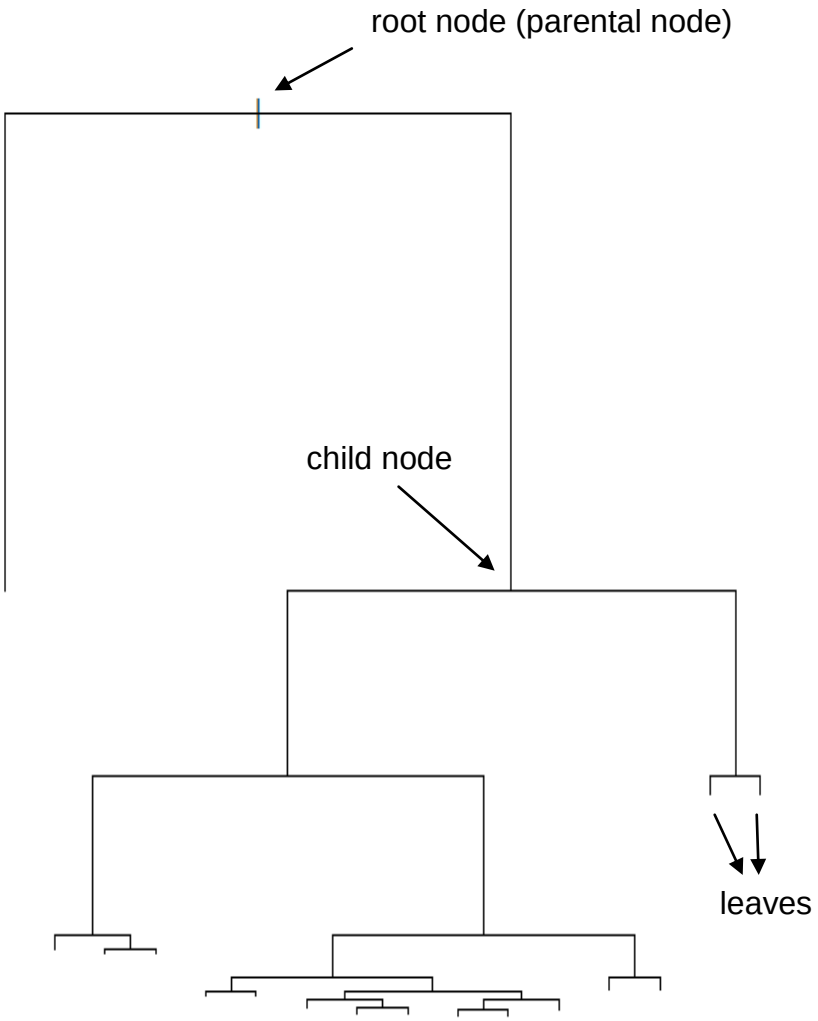

Figure 2.4

A decision tree constructed by using CART. Each group of variables in the dataset (parental node) is split into two parts (child nodes). Values of the variables are called leaves. Nodes names and leaves names are omitted for clarity. The figure made by the author.

In this PhD, CART was used in the chapter Oseltamivir resistance in human seasonal A/H1N1 influenza viruses, to determine the association between amino acid sites from two different proteins.

Random Forest is an extension of the tree learning paradigm. This method is robust to over-fitting and gives better predictive performance than CART (Breiman, 2001). Individual trees are constructed similarly to CART with the difference that splits are determined on a random subset of the available amino acids at each leaf. Because the number of trees is created by growing them on bootstrap samples of the training data this method includes the word 'forest'. So, this method constructs a defined number of decision trees where splits are determined on the selected subset of variables at each leaf. To assess the predictive power of the models tested, it is important to use 10-fold crossvalidation. That means that sequences (the database) are randomly split into 10 parts with one part used as a test dataset and the 9 remaining parts of sequences used as a training dataset. This is repeated 10 times. To interpret the random forest models, the 'permutation accuracy importance' (PAI) measure is applied, which shows the importance of each amino acid site to predict phenotype (Breiman, 2001). To calculate the PAI, the values of amino acids at each site are randomly permutated as to break any link with the phenotype response. Next, the accuracy of the model to predict phenotype using this permutated and all other unpermutated amino acid sites is measured. The greater the change to the model's accuracy, the more important the amino acid site to predict the phenotype. Amino acid sites are then ranked according to their impact on the accuracy of the model in order to identify the most important amino acid sites for the prediction of the phenotype.

In this PhD, random forest was used in the chapters Oseltamivir resistance in human seasonal A/H1N1 influenza viruses, and Amantadine resistance in swine influenza A viruses. In both chapters this method was used to determine which amino acid sites are associated with amino acid site(s) where drug resistance conferring mutations can occur.

2.2.9 Method for detecting co-evolving sites

A **Bayesian graphical model** (BGM) is a directed acyclic graph with nodes and edges (arrows) that represent variables and the relationships among the variables, respectively. So, edges (arrows) represent dependencies between variables (nodes). The degree of dependency of a certain node on the other is expressed by a probability value. Variables represented by nodes can be observable quantities or even unknown parameters or hypotheses. Nodes that are connected are dependent and the direction of the arrow indicates the mode of dependency (Figure 2.5, for example, C depends on A). Each node is associated with a probability value for each dependency (Lee and Abbott, 2003). A Bayesian graphical model is constructed on the basis of nucleotide sequences and a phylogenetic tree of those sequences. The phylogenetic tree can be provided by the user or constructed by the software for BGM analysis.

In this PhD, Bayesian graphical models were used in the chapter Amantadine resistance in swine influenza A viruses to assess which amino acid sites from different genome segments, shown to be functionally associated, are co-evolving and what is the probability of that event.

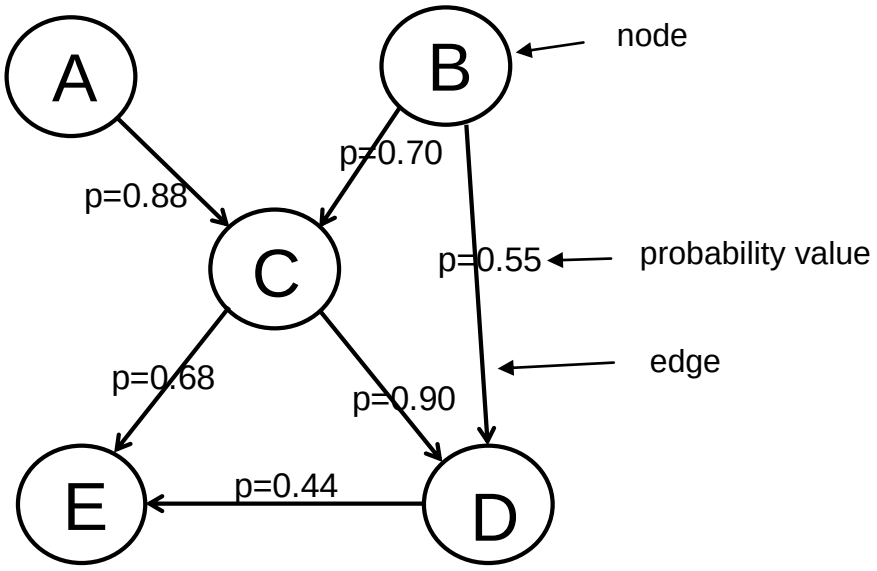

Figure 2.5

A structure of a Bayesian graphical model with nodes representing variables (i.e. amino acid sites), edges representing relationships among the variables, and probability values for each dependency. Figure made by the author.

2.2.10 Methods for detecting amino acid sites under selection

Methods for identifying sites under selection on a site-by-site basis have been proposed. These methods can be divided into three classes and are complementary. First class includes the counting methods that count the number of non-synonymous and synonymous substitutions in the dataset. Second class of methods are called random effects models and they calculate the distribution of evolutionary rates across sites and then infer the rate for each individual site based on that distribution. Third class are fixed effects models and estimate the ratio of non-synonymous to synonymous substitutions for every site (Kosakovsky Pond and Frost, 2005).

The first class is a counting method **SLAC** (Single Likelihood Ancestor Counting) and is used to estimate the number of nonsynonymous and synonymous mutations that have occurred at each codon in the dataset. This method is a modification of the approach first described by Suzuki and Gojobori (1999), and can reconstruct the ancestral sequences by using parsimony (Suzuki and Gojobori, 1999) or methods based on likelihood (Nielsen 2002; Nielsen and Huelsenbeck 2002; Suzuki 2004). SLAC counts the number of non-synonymous and synonymous substitutions occurring on the branches of the tree (i.e. changes between nodes and their immediate ancestors). SLAC is fast computationally and estimates site specific rates directly from each site. The disadvantage of SLAC is underestimation of the number of substitutions that have taken place in a dataset if a dataset is small. So, SLAC can calculate wrongly the rate at which sites are evolving. However, SLAC can correctly detect non-neutral evolution in large datasets (Kosakovsky Pond and Frost, 2005).

REL assumes that both dN and dS vary across sites independently. REL calculates a distribution of substitution rates across sites and then infers the rate at which individual sites evolve (Kosakovsky Pond and Frost, 2005). REL is the most powerful of these three methods but is susceptible for type I errors (false positive result), particularly on small datasets (Kosakovsky Pond and Frost, 2005).

FEL is a derivate of the model first proposed by Yang and Swanson (2002), who suggested that sites evolve under different dN/dS**.** FEL is similar to SLAC in terms of not making any assumption regarding the distribution of rates across sites. On the other hand, FEL is slower than SLAC and can be difficult to use due to the large number of parameters involved, for example models of nucleotide substitution (Kosakovsky Pond and Frost, 2005; Nielsen, 1997).

In this PhD, only SLAC method was used. It was used in the chapter Amantadine resistance in swine influenza A viruses in order to detect the amount of evolutionary changes on codons where adamantine resistance can occur in avian sequences (wild and domestic separately) and swine sequences (North American and Eurasian avian-like separately).

Chapter 3

Reassortment in swine influenza A viruses

3 Reassortment in swine influenza A viruses

3.1 Reassortment

Reassortment is the mixing of genetic material into new combinations. The prerequisite for the reassortment event of influenza viruses is that a host cell is infected with at least two different strains of influenza viruses of the same type (the subtypes can differ, also the host origin can be different). Influenza viruses have a segmented genome, which makes the exchange of viral genetic segments within the infected cell possible. Therefore, some segments come from one influenza strain and some from another. The new reassortant strain carries properties of both parental lineages. Reassortment is responsible for the major genetic shifts in influenza viruses, causing the appearance of all pandemic influenza strains of the 20th and 21st century. The 1957 and 1968 pandemic influenza strains were caused by reassortment between an avian and a human influenza virus. The 2009 pandemic H1N1 virus resulted from the reassortment between avian, human and swine influenza viruses (Hood, 2006).

3.2 Host restriction

Avian hosts, especially waterfowl, present the major reservoir for influenza A viruses. Occasionally, certain influenza A strains cross the species barrier from the avian host into a mammalian host, such as swine, humans and horses (Webster et al., 1992). The consequence can be the epizootic and endemic outbreaks and establishment of these new lineages in the recipient host populations. Out of 18 haemagglutinin subtypes and 11 neuraminidase subtypes known so far, only 3 HA and 2 NA subtypes have become established in the human population and caused epidemics and occasional pandemics. One preventing factor for other viral subtypes' establishment in humans is the receptor specificity of the virus (either α-2,3-linked or α-2,6-linked sialic acid receptors) and the availability of the appropriate receptor in the host (Connor et al., 1994; Ito et al., 1998a; Matrosovich et al., 2000; Rogers and

D'Souza, 1989). The α-2,3-linked sialic acid receptors are found only in small amounts and only in the lower respiratory tract of humans, therefore transmission of avian influenza A viruses is not frequent although direct infection can occur (CDC, 2008; CDC, 2014). Swine influenza A viruses can adhere to both types of receptors and both types of receptors are found in equal abundance in swine lungs. Therefore, as human and avian influenza A viruses can infect pigs, consequently, pigs have been suggested to play a role in facilitating the adaptation of avian viruses to human hosts (Ito et al., 1998a; Scholtissek et al., 1985).

3.3 Pigs as mixing vessels - mixing vessel hypothesis

The "mixing vessel" hypothesis was first introduced by Scholtissek et al. (1985) who reasoned that swine can be infected with either swine, human or avian influenza virus at the same time. This multiple infection could produce reassortant viruses between swine and avian/human viruses, which have the potential to be transmitted back to humans and so introduce unique viruses into the human population. The reassortant viruses could then cause human pandemics. The basis for this hypothesis was the finding that antigens and genome segments between swine, human and avian influenza seem similar. Also, the susceptibility of swine to infection by human and avian viruses was proven. Support for this hypothesis was found at the molecular level with the discovery that avian and human influenza viruses have preferential binding to specific receptor types. There are three parts to the mixing vessel hypothesis: 1) swine are susceptible to avian and human influenza A viruses; 2) reassortment of swine/avian/human viruses occurs in the pig and 3) pigs can transmit reassortant influenza viruses to people.

3.4 A brief history of reassortment events in swine

In influenza A viruses, segments usually correspond to host type, however, there have been reassortments among segments as well as various cross-species transmissions, the outcome of which are the new genotypes or clades (dos Reis et al., 2009; Dunham et al., 2009; Schultz et al., 1991; Smith et al., 2009). These events led to: (i) the transmission of H1N1 from birds to swine, creating the H1N1 Eurasian avian-like swine lineage clade found predominantly in Europe and Asia; (ii) reassortment between Eurasian avian-like swine H1N1 and human viruses, resulting in the establishment of H1N2 and H3N2 reassortants in swine in Europe; (iii) the complex reassortment of swine, avian and human H1N1, H1N2 and H3N2 viruses that created the swine triple-reassortant viruses, circulating predominantly in North American swine from 1990s onwards; (iv) reassortment between the H1N1 Eurasian avian-like swine influenza viruses (i) and swine triple-reassortant viruses (iii), creating the 2009 H1N1 human pandemic strain (Brown et al., 1998; Castrucci et al., 1993; Garten et al., 2009; Olsen, 2002; Pensaert et al., 1981; Smith et al., 2009; Webby et al., 2000; Zhou et al., 1999).

3.5 Antigenic drift in swine influenza versus human influenza viruses

Antigenic drift is the accumulation of point mutations in the influenza virus' genome due to the error prone RNA polymerase of influenza viruses. Consequently, these mutations cause the viruses to be antigenically different from their parent viruses (Gerhard and Webster, 1978). In commercial swine herds, there is less pressure to select for antigenic changes compared to humans for two main reasons: (i) a high proportion of susceptible individuals at any given time because of a high birth rate and a lifespan of no more than six months; (ii) no widespread use of vaccines for swine in Europe. This means that evolutionary pressures on HA and NA are not expected to be the same in swine and humans – for swine H3 antigenic drift rate is much lower (de Jong et al., 2007; Lorusso et al., 2011). For seasonal influenza viruses in humans we see the correlated lineages in all segments being driven by the

HA and NA subtype, however, in swine the persistence and transmission of the internal polymerase coding segments regardless of the HA and NA subtype was observed. Therefore, for swine, the polymerase segments are the primary heritable units which acquire different surface glycoproteins (HA and NA) through reassortment. To counteract herd immunity in humans, the antigenic drift (antigenic novelty) is driving the dynamics of influenza strains (Garten et al., 2009; Boni, 2008; Russell et al., 2008; Smith et al., 2004). There is very little reassortment in humans, so the external HA and NA subtypes broadly correspond to the internal gene segments.

3.6 Swine influenza A viruses in Europe

Swine influenza is considered to be an important disease of swine in Europe. The characteristics of the viruses circulating in Europe share some similarities with strains circulating in North America and Asia but there are also a number of differences. The epidemiology and etiology of swine influenza in Europe have historically been different from other parts of the world and therefore also approaches for its' control remain different (Zell et al., 2013).

3.6.1 Classical H1N1

This virus was introduced into pig populations around 1918 (Smith et al., 2009) and was endemic in European swine until the emergence of avian-like H1N1 virus when it disappeared (Brown, 2000). It was antigenically relatively stable with no evidence of involvement in virus reassortment (Brown et al. 1997).

3.6.2 Human-like H3N2 viruses

Under natural conditions pigs can become infected with prevailing human subtypes. In the 1970s H3N2 viruses were isolated regularly from European pigs (Tumova et al. 1976; Ottis et al. 1982). Since 1984 these viruses caused frequent outbreaks of

clinical influenza in pigs in Europe (Aymard et al. 1985; Haesebrouck et al. 1985; Pritchard et al. 1987). Although related closely to the prototype human viruses, swine human-like H3N2 viruses have antigenic differences in the surface glycoproteins (Haesebrouck and Pensaert 1988; Brown et al. 1995). The HA gene of the human-like H3N2 virus has experienced considerable antigenic variation due to genetic drift and this led to an increase in the epizootics of this virus (DeJong et al. 1999). These recent viruses are only distantly related to the early strains.

3.6.3 Avian-like H1N1 viruses

These viruses have dominated pig populations in Europe since 1979. They are antigenically distinct from the Classical swine H1N1 viruses but related to avian H1N1 isolated from ducks (Pensaert et al. 1981; Scholtissek et al. 1983). Because these avian-like H1N1 viruses replaced the Classical H1N1 lineage in Europe, it suggests they had the selective advantage (Campitelli et al. 1997; Brown, 2000) because of their evolutionary distance from the viruses previously found in swine. This lineage came into pigs from birds and became established in European pigs where it remains the dominant strain in many European countries.

3.6.4 H1N2 viruses

These viruses were first isolated in the late 1980s and are derived from the Classical swine H1N1 and human-like swine H3N2 (Gourreau et al. 1994). Their HA gene segment was inherited from the Classical swine H1N1 lineage and the NA segment was derived from the human-like swine H3N2 virus. At first these viruses did not spread widely but in 1994 H1N2 viruses related antigenically to human and human-like swine viruses became endemic in pigs (Brown et al. 1995; Van Reeth et al. 2000; Marozin et al. 2002). Occasionally H1N2 viruses with different HA appear but do not persist. Also H1N2 viruses with a novel NA appeared in Germany in Dotlingen and Cloppenburg in 2005 (Zell et al. 2008).

3.6.5 H1N1pdm09

In March 2009, a novel H1N1 virus was detected in humans in Central America. It contained a unique gene constellation not previously reported in humans or pigs (Garten et al. 2009). Two gene segments were derived from avian strains of North American lineage, one gene from human influenza, three gene segments from the Classical swine lineage and two segments from the Eurasian swine lineage (NA and MP). The precise origin of this virus is uncertain as there is a gap in its ancestry. However, the most likely origin are pigs as at least seven of eight segments are similar to those of viruses circulating in pigs (Smith et al. 2009). This virus had not been detected in European pigs before its discovery in the human population (Kyriakis et al. 2011).

3.6.6 Reassortment in European swine influenza A viruses

Lycett et al. (2012) studied the co-circulation of influenza strains and how this can lead to the generation of reassortants. They estimated the rate of virus reassortment in European swine and they found that after the formation of the Eurasian avian-like swine clade, there were two introductions of human seasonal lineage H1 HA and two of classical swine lineage H1 HA. They also found two introductions of human H3 HA and N2 NA seasonal lineages, respectively. The estimated median TMRCA of the N2 NA introductions was 1983 (H1N2 into England, Continental Europe and Asia) and 1986 (H3N2 into Continental Europe and Asia). For both of these introductions, they also estimated the TMRCA of the European sequences only, since these formed separate subclades from the Asian isolates. The first introduction of H1N2 into Europe had the median TMRCA of 1988-1989 (95 % highest posterior density (HPD) interval: 1984–1991). The predominantly H3N2 introduction into Europe occurred around 5 years later, with a median TMRCA of 1994 (95 % HPD: 1990–1998). These multiple introductions of HA and NA segments resulted in a mix of subtypes in the Eurasian swine clade (as defined by the internal protein-coding segments).

Lycett et al. (2012) found that the median rate at which the Eurasian swine N1 switch to human seasonal-origin N2 on the polymerase backbone is 0.13 exchanges/year. The human seasonal-origin N2 switching to Eurasian swine N1 is 0.17 exchanges/year, which is not significantly different. For HA switching, the most significant rates were between Eurasian swine H1 and human seasonal-origin H1 (0.37 exchanges/year for the H1 human seasonal-origin to H1 swine Eurasian direction; for the opposite direction the median rate was 0.16 exchanges/year). Lower rates – but still significant – were found for H1 swine Eurasian and H3 human seasonal-origin, and H3 human seasonal-origin to H1 swine Classical. The rates between the two swine lineages (H1 swine Eurasian and H1 swine Classical) and two human seasonal-origin lineages (H1 and H3) were not significant. The rates for H1N1 Eurasian swine to/from H1N2 human seasonal-origin were almost the same as those for H1 swine Eurasian to/from H1 human seasonal-origin. Especially the rate for H1N2 changing to H1N1 Eurasian was high (0.4 exchanges $year^{-1}$), whilst the reverse rate was significantly lower (0.16 exchanges $year^{-1}$). The rates between swine lineages H1N1 swine Eurasian and H1 swine Classical N1 swine Eurasian, and the two human seasonal-origin lineages (H1N2 human seasonal-origin and H3N2 human seasonal-origin) were not significant.

3.7 Swine influenza A viruses in Asia

The circulation of swine influenza A viruses is more complex in Asia than in North America and Europe because of constant importation of pigs from these two continents. Apart from these viruses, there are also several lineages found only in Asia (Zhu et al., 2013).

3.7.1 H1N1 viruses

The Classical swine virus that first appeared in 1918 in USA was first observed in Chinese pigs that same year (Chun, 1919). Later it spread to the rest of Asia (Yip, 1976; Yamane et al. 1978; Das et al. 1981; Shortridge and Webster, 1979;

Kupradinun et al. 1991). With the spread of H1N1 Russian influenza viruses there was also one case of human to pig transmission in Thailand (Nerome et al. 1982). After avian-like H1N1 appeared in European swine, it was also isolated from Chinese pigs and circulated with the Classical swine H1N1 (Guan et al. 1996). It was phylogenetically distinct from the European viruses and formed a separate lineage (Brown, 2000). There were sporadic appearances of the European avian-like swine H1N1 viruses in China in 2007 (Liu et al. 2009), introductions of non-pandemic human-like H1N1 in Japan and China (Katsuda et al. 1995; Yu et al. 2007, 2009a) and also reassortants of North American and European lineages (Choi et al. 2004; Song et al. 2003; Pascua et al. 2008; Chutinimitkul et al. 2008; Takemae et al. 2008). In southern China and southeast Asia, the Classical swine H1N1 lineage continues to be endemic (Guan et al. 1996; Peiris et al. 2009; Qi et al. 2009).

3.7.2 H3N2 viruses

Human-like H3N2 was first isolated in swine in Taiwan in 1969 after the Hong Kong pandemic (Kundin, 1970). After that the evolution of this lineage has been quite complex, especially in Southeast Asia. In the early 1980s the H3N2 virus was isolated in China with the human-like HA and NA and Classical internal gene segments (Shu et al. 1994; Nerome et al. 1995). Around the same time, two triple-reassortant H3N2 viruses, unrelated to the reassortants of the North America or Europe, were isolated from Chinese pigs (Yu et al. 2008). In 1999, the European reassortant human-like swine H3N2 viruses emerged in chinese pigs. Just two years later, further reassortment with the Classical swine H1N1 virus was observed and that formed triple reassortant strains (Yu et al. 2008). European reassortant human-like swine H3N2 viruses also recombined with human-like H1 or H3 strains (Takemae et al. 2008; Chutinimitkul et al. 2008).

3.7.3 H1N2 viruses

These viruses were derived from the genetic reassortment of human-like swine H3N2 and Classical swine H1N1. In Asia, they were first isolated in Japan in 1978

(Sugimura et al. 1980) and became established in Japan and Taiwan (Yoneyama et al. 2009; Tsai and Pan, 2003). However, in Korea, the North American reassortant H1N2 with the Classical swine HA spread widely (Choi et al. 2002a; Pascua et al. 2008). Also reassortant H1N2 viruses with the Classical H1 and the remaining gene segments from the Eurasian avian-like swine lineage were isolated from pigs in Thailand (Chutinimitkul et al. 2008; Takemae et al. 2008). Different reassortant H1N2 viruses have been observed in China, most of them are the product of genetic reassortment between Classical swine viruses and European reassortant or North American triple reassortant viruses (Smith et al. 2009; Yu et al. 2009b).

3.8 Swine influenza A viruses in North American swine

Coincidently with the 1918 human pandemic that killed from 50 to 100 million people worldwide (Murphy and Webster, 1996; Barry, 2005) influenza was clinically recognized in pigs during the late summer and autumn of 1918 in the Midwestern United States (Koen, 1919; Easterday and Hinshaw, 1992). The first swine isolates were collected and studied by Shope in 1930 (Shope, 1931). Those were the ancestral lineages of the Classical swine lineage circulating till today. Other subtypes of swine influenza have been isolated in Europe and Japan, mostly H3N2 and H1N2 (Sugimura et al., 1980, Nakajima et al., 1982, Ottis et al., 1982, Yasuhara et al., 1983, Mancini et al., 1985, Castrucci et al., 1994 and Campitelli et al., 1997; Nerome et al., 1985, Gourreau et al., 1994, Brown et al., 1995, Brown et al., 1998, Ouchi et al., 1996, Ito et al., 1998b; Van Reeth et al., 2000). From 1930 through the mid-1990s, influenza in North American pigs was caused almost exclusively by infection with Classical H1N1 swine viruses. The studies between 1997–1998 (Olsen et al., 2000) revealed an unexpected and substantial increase in H3 seropositivity, and H3N2 viruses began to be isolated from pigs in both the US and Canada during this time (Karasin et al., 2000a; Zhou et al., 1999). Subsequently, reassortment between H3N2 viruses and classical H1N1 swine viruses led to the appearance of second generation H1N2 reassortant viruses (Karasin et al., 2000b and Karasin et al., 2002).

3.8.1 Classical H1N1 viruses

This virus strain was first recognized in pigs in 1918 in USA and that coincided with the human influenza pandemic. It was not until 1930 that the first influenza was isolated (Shope, 1931) and shown to cause the respiratory disease in pigs similar to human influenza. This strain was subsequently recognized as a Classical H1N1 influenza virus and remained relatively stable genetically and antigenically until the late 1990s.

3.8.2 H3N2 viruses

Although the Classical H1N1 lineage was predominant in pigs in the USA until mid-1990s, sporadic appearances of H3 were also observed between 1988 and 1989 but lacked to form a stably transmissable lineage (Chambers et al., 1991). In 1998, there were severe outbreaks of swine influenza disease observed in several US regions. The causative agent for these outbreaks was recognized as the H3N2 subtype and formed two genetically different lineages. The initial virus was a double reassortant; its gene segments were similar to those of the Classical lineage (PB2, PA, NP, MP, NS) and human seasonal H3N2 virus circulating in 1995 (PB1, HA, NA). The other lineage of this H3N2 virus was a triple reassortant and contained gene segments from the Classical swine viruses (NP, MP, NS), human seasonal H3N2 (PB1, HA, NA) in combination with the avian virus (PB2, PA) (Zhou et al., 1999). The double reassortant did not become established but the triple reassortant became widespread in the pig population in the USA (Webby et al., 2000). This constellation of human seasonal H3N2 PB1, avian lineage (PB2, PA) and swine lineage (NP, MP and NS) is conferred to as the triple reassortant internal gene (TRIG) cassette (Vincent et al., 2008). Since 1998 there were at least three introductions of human H3 into swine population, leading to phylogenetic clusters I, II and III (Richt et al., 2003; Webby et al., 2004). The cluster III became dominant in North America and continued to evolve into clade IV (Olsen et al., 2006). The H3N2 viruses further reassorted with the extant Classical H1N1 viruses resulting mostly in reassortants that possess the TRIG. The H1N1 viruses containing the HA and NA from the Classical H1N1 and

the TRIG from the triple reassortant H3N2 viruses are referred to as reassortant H1N1. The viruses that contain the HA from the Classical H1N1 and NA and TRIG from the triple reassortant H3N2 are H1N2 viruses (Karasin et al., 2002; Webby et al., 2004). There were also further genetic drift variants of these reassortant viruses observed (Webby et al., 2000, 2004; Richt et al., 2003; Olsen et al., 2006; Choi et al. 2002b; Karasin et al. 2002; Webby et al., 2004). Further, H3N1 and triple reassortant with TRIG and avian H2 were observed; the later forming the novel triple reassortant swine H2N3 in 2006 (Ma et al., 2007). Introduction of H1 viruses with the human H1N2 HA was observed in Canada (Karasin et al. 2006). Since 2005, H1N1 and H1N2 viruses that have HA and NA segments similar to human seasonal H1N1 and H1N2 viruses from around 2003 have emerged (Vincent et al. 2009).

3.8.3 H1N1pan09 viruses

In spring 2009 an outbreak of severe human influenza in Mexico was associated with a novel pandemic H1N1 virus that was isolated from humans in California. It possessed a unique genome with six gene segments (PB2, PB1, PA, HA, NP and NS) most closely related to the triple reassortant viruses from North America and with the MP and NA genes derived from the Eurasian swine lineage (Dawood et al. 2009). This novel virus became known as ‘swine flu’ due to the phylogenetic origin of its genome segments. However, infection of humans was not connected to exposure to pigs (Dawood et al. 2009). Immediately after the onset of infection with this pandemic virus in humans, cases of infection of pigs were reported in different areas of the world (Wahid, 2013). It was later shown that the H1N1pan09 spread from humans to pigs with subsequent sustained pig-to-pig transmission. Phylogenetic analyses of all eight gene segments of this pandemic virus revealed that none of these gene segments cluster tightly with the genes of swine influenza viruses circulating in the USA prior to the outbreak of 2009 pandemic in humans (Lorusso et al. 2011; Smith et al. 2009). So, neither the 2009 pandemic H1N1 nor closely related progenitor viral genes were present in U.S. swine influenza viruses prior to 2009 (Lorusso et al. 2011). However, there is a gap in genetic surveillance of swine influenza viruses and therefore the ancestor lineages of the pandemic 2009 virus

were circulating undetected in pigs for approximately a decade (Smith et al., 2009). A closely related progenitor virus with the same eight gene constellation has yet to be identified in swine or other species, although a 2004 swine virus with 7/8 of the 2009 pandemic H1N1 genome was identified in Hong Kong, China (Smith et al. 2009). The 2009 pandemic virus was found to cause sustained human-to-human transmission but also transmission from human-to-swine (Howden et al. 2009), swine-to-swine (Lange et al. 2009; Brookes et al. 2010) and swine-to-human (Weingartl et al. 2010). This virus was shared between people and pigs and has therefore the potential to further change the epidemiology of influenza viruses in human and swine populations.

3.8.4 Reassortment in the North American swine influenza

Nelson et al. (2012) estimated the extent of genomic reassortment in influenza A viruses circulating in North American swine by performing a phylogenetic analysis of whole genome viral sequences sampled during 1998-2011. The highest amounts of reassortment were detected between the H3 and the internal gene segments (PB2, PB1, PA, NP, MP and NS) and the lowest reassortment frequencies were observed among the H1pdm and neuraminidase segments, particularly N1. They also observed less reassortment among specific HA-NA combinations that were more prevalent in swine, which might suggest that some genome constellations may be evolutionary more stable. More than 100 human cases of novel reassortant swine-origin H3N2 influenza viruses were identified since July 2011 and this reassortant continues to be monitored closely for pandemic potential (CDC, 2012; Lindstrom et al., 2012). The combination of six triple reassortant internal genes (TRIG) is conserved in North American swine and has remained prevalent since 1998. Overall, their data indicate that the HA reassorts more frequently with internal gene segments than with NA, particularly N1.

3.9 Study on reassortment in swine influenza viruses by Khiabanian et al (2009)

Khiabanian et al (2009) used the publicly available full-genome sequences of swine H1N1, H1N2, H3N2 deposited in the Influenza Virus Resource of the NCBI to study the reassortment patterns in swine influenza A viruses. Their approach used several statistical methods. First, for each segment the sequences were aligned using the Smith-Waterman algorithm and the normalized Hamming distances only at the third codon positions were calculated, to eliminate the effects of evolutionary pressure due to positive selection. Second, in order to find the possible reassortant strains the method first introduced by Rabadan et al. (2008) was used. In this method, the number of nucleotide differences between the segments of any two strains is calculated. Assuming that the segments have proportional substitution rates at the third codon positions, the differences between two segments of two strains should be proportional if the two segments have a common origin. If not proportional, there has been a reassortment event – the histories of the two segments are different. Therefore, when the distances between two segments of different strains are plotted against each other, the points corresponding to possible reassortment events lie off the diagonal.

They confirmed that HA and NA reassort more frequently than the other segments. They also found that one of the polymerase segments, PB1, reassorts quite frequently. However, the use of Hamming distances rather than a statistical model of nucleotide substitution (see Methods) means the distances would have been underestimated and the power of these studies to detect reassortment therefore reduced.

3.10 The aim of study

In this thesis chapter swine influenza virus evolution with particular reference to reassortment was studied. The focus was on interactions between different swine influenza A segments and extent of their co-evolution. The first question to answer was whether some segments associate more strongly. Does reassortment between segments coding for internal proteins occurs less often in swine than in the segments that code for surface proteins haemagglutinin and neuraminidase? Based on the observation that an A/H1N1pan09 virus had Eurasian avian-like MP and NA segments, this chapter tried to answer the question if these two segments always reassort together. Also, within and between subtypes reassortment were compared.

3.11 Methods and data

3.11.1 Dataset

A dataset of full genome 507 H1N1 swine sequences, 115 H1N2 swine sequences and 226 H3N2 swine sequences was downloaded separately from two databases: NCBI FLU (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) and GISAID (platform.gisaid.org/epi3/frontend). The sequences were isolated from January 1st 1970 to December 31st 2012.

3.11.2 Phylogenetic analysis

H1N1, H1N2, H3N2 sets of sequences (all 8 segments separately) were analysed in BEAST v1.6.1 with the GTR (for MP and NS segments) or SRD06 (for all other segments) nucleotide substitution model, a relaxed clock (uncorrelated lognormal) and a constant population size demographic model with an MCMC chain length of 200,000,000 and 10% burn-in. These BEAST posterior samples were further subsampled to create sets of 1000 post burn-in trees for each segment, MCC (maximum clade credibility) trees were also created from the post burn-in samples.

Taxa in the MP MCC tree were manually annotated using different colours for different clades in FigTree (Rambaut, 2007), and then this same colouring scheme was applied to all the other segment trees, enabling a visual tree comparison and identification of reassorting clades.

3.11.3 Distance in topology

To compare the topology of two trees (i.e. the branching order) a distance topology measure can be used (Penny and Hendy, 1985; Rzhetsky and Nei, 1992). In particular, the dist.topo function within R package ape (Paradis, 2009) using the PH85 measure was used. Here the topological distance is defined as twice the number of internal branches defining different bipartitions of the tips (Penny and Hendy, 1985). Rzhetsky and Nei (1992) proposed a modification of the original formula to take multifurcations into account which is also implemented with the dist.topo function.

The topological distance measure was used as a means to assess for stronger/weaker association among the segments. The lower the value for the topological distance between two segments, the closer their topologies of trees – the segments share a common phylogenetic history. In the case of a higher value of the distance in topology, this can reflect the reassortment event(s) between the segments.

Since the distance in topology score value depends upon the number of taxa in the trees as well as the difference in tree structure, the scores of trees with randomised tip labels were calculated in order to provide a suitable comparative measure. In particular, 5%, 10%, 15% and 20% of tips were randomised and scored against the unrandomised sets.

The topology distance was used to compare two sets of 1000 post burn-in trees (the first tree from set 1 against the first tree from set 2, the second tree from set 1 against the second tree from set 2 etc).

Also, the distance in topology values for each segment were normalised according to the formula:

$$\frac{S1\ vs\ S2}{\left[0.5 * \left((S1\ vs\ S1) + (S2\ vs\ S2)\right)\right]}$$

Where S1 is segment 1 and S2 is segment 2.

So, the self-values for each segment (i.e. HA-HA) were set at a value 1 and values of other segment pairs were comparable to the self-values. This was done because segments have different lengths and their self-values differ (i.e. MP-MP and NS-NS self-values are higher as these two segments are shorter compared to other virus' segments).

Wilcoxon test was used to test the pairs of segments if the difference in topology score is statistically supported.

Two questions were adressed: (1) if there is evidence that e.g. segment1-segment2 is reassorted compared to segment1-segment1 (self) by distance topology score – if the distance in topology score for segment1-not segment1 is always higher than segment1-segment1(self), and the confidence intervals do not overlap, the Wilcoxon statistical test should show that the distributions are different, so there is reassortment.

(2) Is e.g. segment1-segment2 more or less reassorted than segment1-segment3; if the value from the Wilcoxon test for segment1-segment2 is significantly higher than segment1-segment3 then there is less correlation between segment 1 and segment 2 than segment 1 and segment 3. For the Wilcoxon test, all values were divided by the value of segment(x)-segment(x) compared to all other pairs of segments and the normalised values obtained were considered as the p-value. The p-value of ≤ 0.05 was considered as significant. The self comparisons (on the diagonal) after normalisation are 0.5 (i.e. half of the samples are higher than the other half).

Because the same test (Wilcoxon test) was performed on a large number of different samples (each segment pair compared to another segment pair), the p-value 0.05 can no longer be considered as a threshold for significance because of the problem of multiplicity. This means that when increasing the number of hypotheses in a test, the likelihood of different events is also increasing (Streiner and Norman, 2011). One of the approaches to deal with multiplicity is the Bonferroni correction which needs to be applied in order to counteract the problem of multiple comparisons. Therefore, the p-value 0.05 has to be divided by the number of tests performed (Streiner and Norman, 2011) and this new value has to be considered as a measure of significance. In the case of this study, there are two different ways to present results of the Wilcoxon test. The first is on the segment level where for each segment studied, there are 64 possible comparisons of two segment pairs, therefore 0.05/64 equals 0.0008 (the p-value). The second approach is to present results on a genome level where the number of tests is 64x8, therefore 0.05/(64x8) equals 0.0001. For the purpose of this study, the results were interpreted on a segment level.

Next, simulated sequences were generated based on the real influenza virus sequences in a java simulator (adapted by Dr S Lycett) utilising the SeqGen functionality within BEAST (Drummond and Rambaut, 2007) by choosing the length for simulated sequences which was the same as for the real sequences (without stop codons). The nucleotide substitution model chosen was HKY and gamma was the same as for the real sequences. BEAST analysis of simulated sequences was done in BEAST v1.6.1 using 500,000,000 generations, GTR nucleotide substitution model, a relaxed molecular clock and a constant demographic model. For each set of simulated sequences corresponding to the length of real influenza genomic segments, the BEAST trees sets (1000 in each set) were compared against each other to assess their topology distance. Shorter segments have higher distance topology scores because the tree inference is more uncertain for shorter segments. Hence to show that this was indeed the case, sequences of different lengths (but using the same true tree) were simulated and the distance topologies done.

R environment v2.14.1 was used. For determining the distance in topology score between sets of trees for pairs of segments, a script written by Dr Lycett was used. This reads the trees in the BEAST output files, compares sets of 1000 BEAST trees and outputs the distance in topology measure for each comparison utilising the dist.topo function within package ape. For obtaining boxplots of the distance in topology measure for each segment pair, the boxplot function was used from grDevices library. For performing the tip perturbations, an all-in-one script (written by Dr Lycett) was used which reads in BEAST MCC tree (nexus format), performs tip perturbation, writes results and image to file (as default).

3.11.4 GARD

GARD analysis was run for each pair of segments for swine H1N1, H1N2, H3N2 in order to detect reassortment, which is shown by improvement in c-AIC. The majority of sequences of H1N1 subtype were isolated from Asia (252/507), followed by North America (186/507) and Europe (65/507). For H1N1 only, the dataset of 507 sequences was randomly split in 10 files and the first 5 files were joined together and realigned, so obtaining 230 sequences. The subsampled H1N1 sequences followed the same distribution as the original H1N1 dataset: the majority of sequences were from Asia (116/230), followed by North America (81/230) and Europe (31/230). The reason for making the dataset smaller was the problem (with the c-AIC error) where the number of sites vs number of sequences (i.e. number of observations vs number of parameters to infer) was too large. For H1N2 and H3N2 where the datasets were smaller, the number of sites is about the same but the number of sequences is fewer, so there are fewer parameters to estimate from the same number of sites.

A positive c-AIC score indicates reassortment (recombination); and if only one breakpoint is detected in the concatenated file, and if the location of the breakpoint is near the join, then this is interpreted as evidence of reassortment.

GARD analyses were submitted to a server cluster Eddie of the University of Edinburgh, which runs as a command-line interface.

To prepare the sequence files for GARD analysis, R environment v2.14.1 was used. For concatenating two pairs of segment sequences, base, ape and seqinr libraries were used. For H1N1 only, 230 sequences out of 507 were extracted randomly with: (i) the use of libraries ape and igraph for extracting identical sequences and leaving a file with unique sequences for MP segment, (ii) matching unique sequences for MP to other segments with ape library, (iii) concatenating sequences as described above. These scripts were written by Dr Lycett and Emma Hodcroft.

3.12 Results

To assess how much reassortment there is in swine H1N1, H1N2 and H3N2 sequences separately, two lines of evidence were constructed: (i) the results from distance in topology measure and (ii) the results from GARD.

The first compares each segment to another segment in all possible combinations in order to see which segments reassort more strongly than others. In support of the distance in topology measure, tip randomisations have been conducted to be able to indicate the equivalent percentage of tip randomisation corresponding to a particular numerical score (e.g. a score of 300 might be equivalent to 10% randomisation in a particular dataset).

Also, since trees inferred from shorter sequences are likely to be less accurate, larger distance in topology scores from segments with shorter sequences are expected. Consequently, a normalisation scheme was also applied because segments differ in length and that affects the distance in topology measure. Simulations were also performed to verify this point.

For the second line of evidence, concatenated sequences of two segments in all combinations were made and analysed in GARD and the resulting c-AIC score was recorded. The higher this score the more reassortment there is between two

segments. Also, the Δc-AIC value obtained from GARD analysis was recorded. The higher this value, the more reassortment there is between two segments.

3.12.1 Within subtypes reassortment

3.12.1.1 Large scale clade structure and reassortments

First, phylogenetic trees of each segment were constructed for each subtype, coloured according to different clades and compared against each other (Figures 3.1–3.9). The MP tree of swine H1N1 sequences (Figure 3.1) shows clear separation of two clades. The red clade contains the majority of North American sequences and also Asian sequences, while the blue clade contains Eurasian sequences. The MCC trees for other 7 segments were compared to the MCC tree of MP visually to assess for reassortment. The HA tree (Figure 3.2) also forms two separate clades; one clade is coloured only blue – Eurasian sequences, while the other is mostly red – North American and some Asian sequences. Similarly, the NA tree (Figure A1) is almost entirely separated into two clades, not well mixed in colour. Therefore, MP of swine H1N1 does not reassort frequently with HA and NA. The NP segment (Figure A2) is also well separated in two differently coloured clades but with some blue coloured strains in the red clade. Therefore, there has been more reassortment of NP with MP compared to HA or NA. The situation is similar in the tree for NS (Figure A3) where the colours are slightly mixed. The trees for polymerase segments PA (Figure A4), PB1 (Figure 3.3) and PB2 (Figure A5) are similar to each other with partial mixing in one clade while there is none in the other. Swine H1N2 MP tree (Figure 3.4) also separates in two well defined clades with a similar reassortment pattern than swine H1N1. One clade stays mostly un-reassorted, while the other mixes up. The two segments with most of reassortment seen are HA (Figure 3.5) and NA (Figure A6) for swine H1N2 – compared against the MP tree. Other segments reassort less with MP (Figures A7-A10). The majority of colour mixing (reassortment) is obvious from trees of swine H3N2 where all 7 segments reassort with the MP (Figures 3.7–3.9, A11–A15).

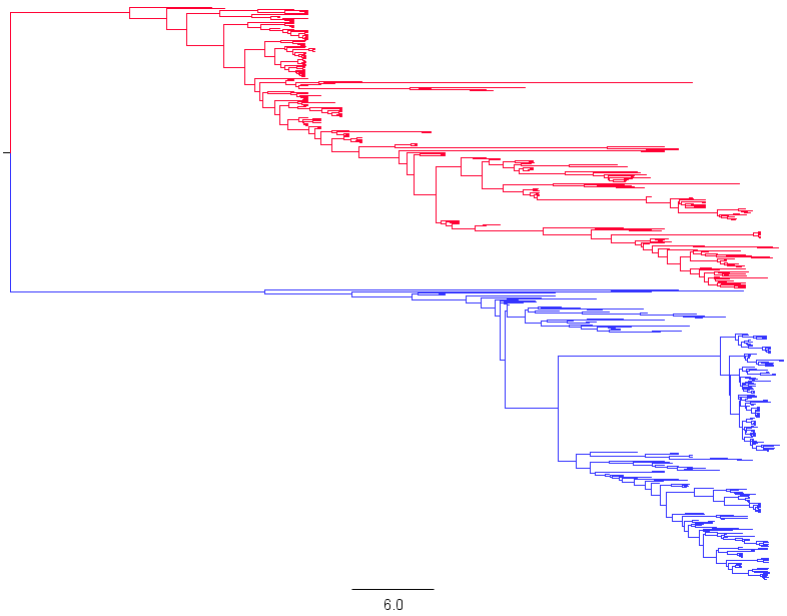

Figure 3.1

The MCC tree of MP for swine H1N1. Strains in the two major clades are shown by the same colour for each of the 3 figures (3.1 – 3.3). Departures from this pattern in trees for the other segments indicate reassortment.

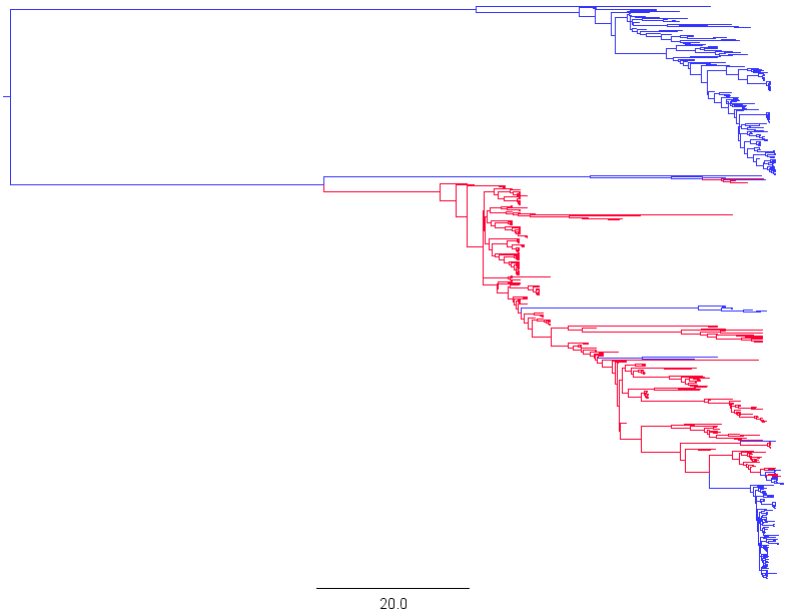

Figure 3.2

The MCC tree of HA for swine H1N1, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

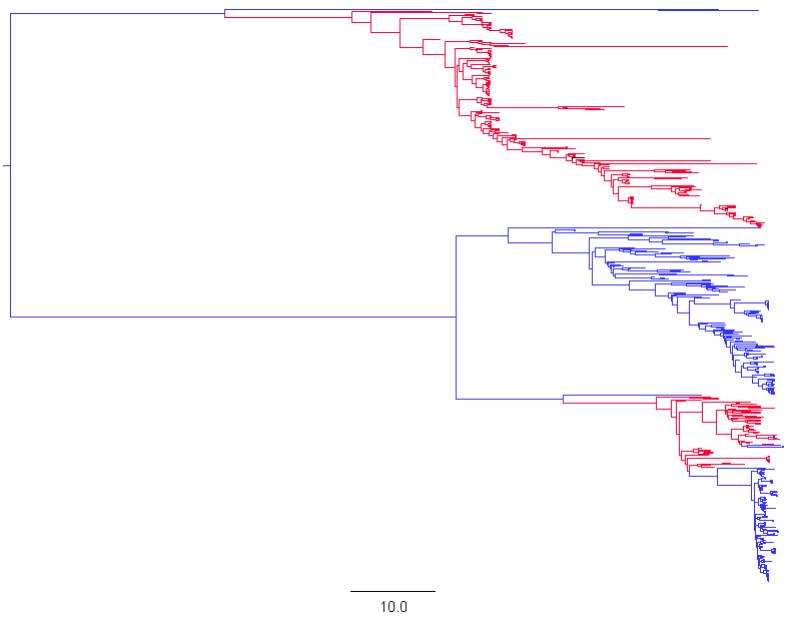

Figure 3.3

The MCC tree of PB1 for swine H1N1, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

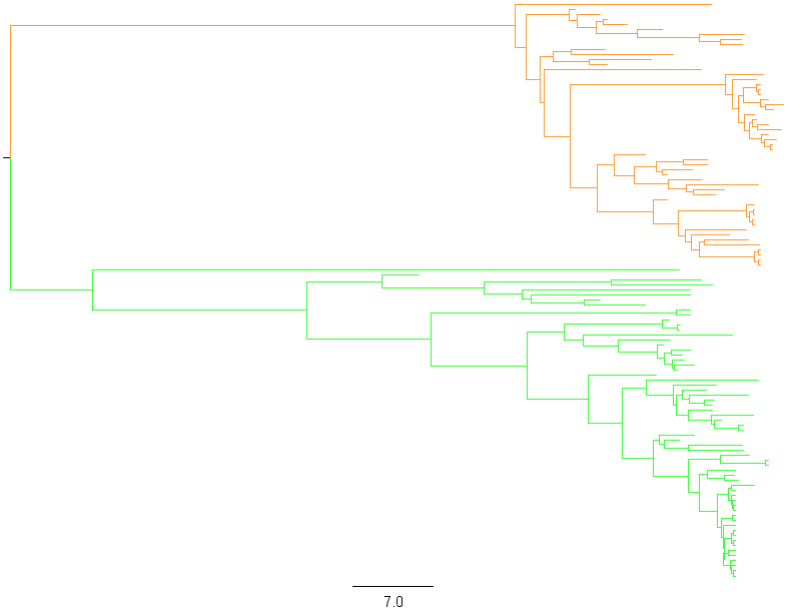

Figure 3.4

The MCC tree of MP for swine H1N2. Strains in the two major clades are shown by the same colour for each of the 3 figures (3.4 – 3.6). Departures from this pattern in trees for the other segments indicate reassortment.

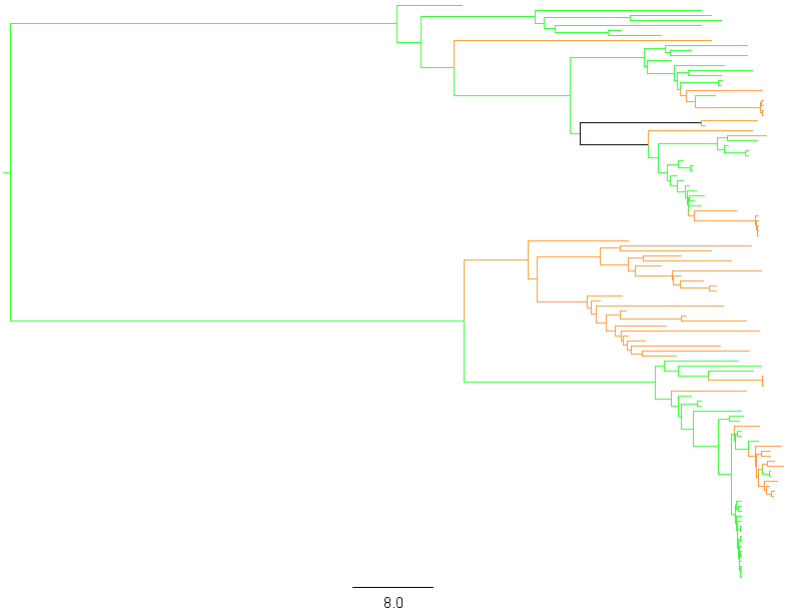

Figure 3.5

The MCC tree of HA for swine H1N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

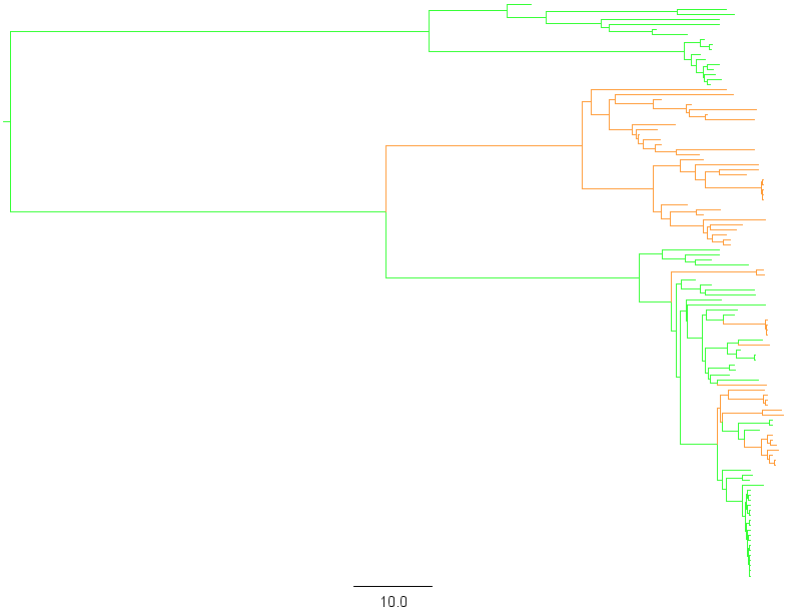

Figure 3.6

The MCC tree of PB1 for swine H1N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

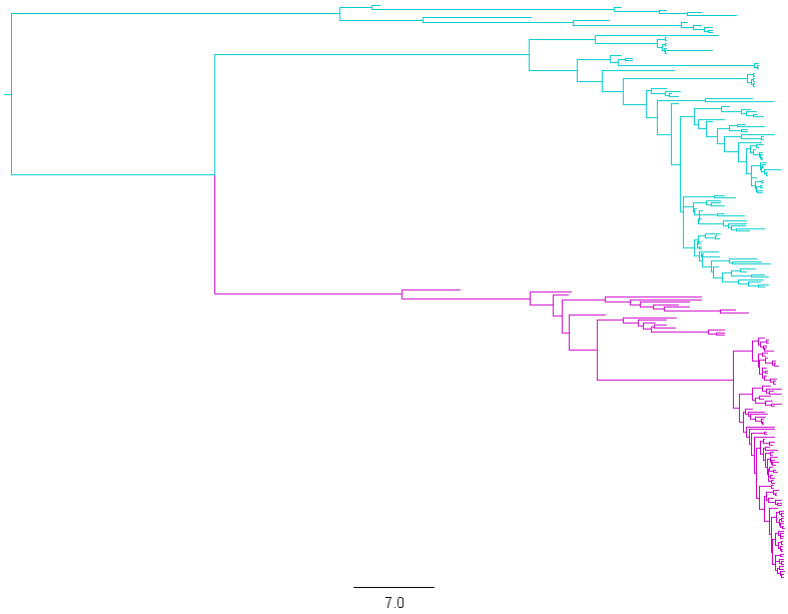

Figure 3.7

The MCC tree of MP for swine H3N2. Strains in the two major clades are shown by the same colour for each of the 3 figures (3.7 – 3.9). Departures from this pattern in trees for the other segments indicate reassortment.

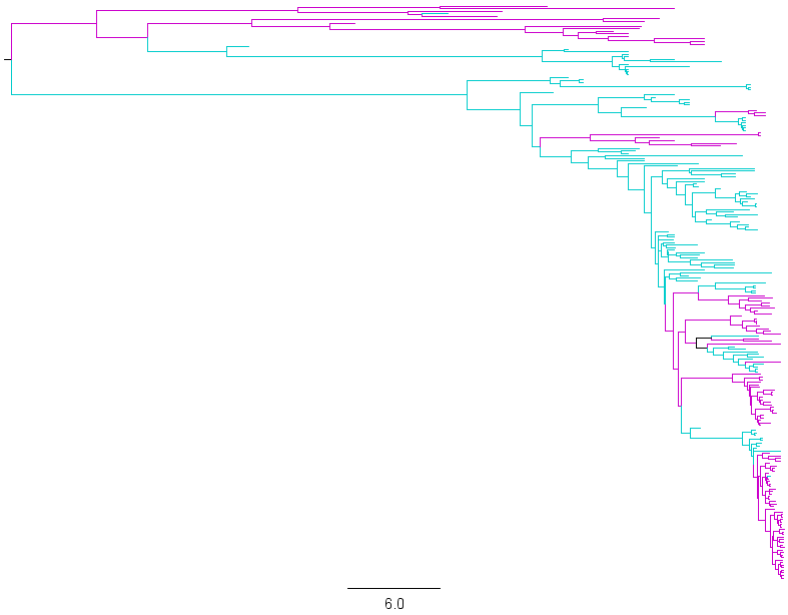

Figure 3.8

The MCC tree of HA for swine H3N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

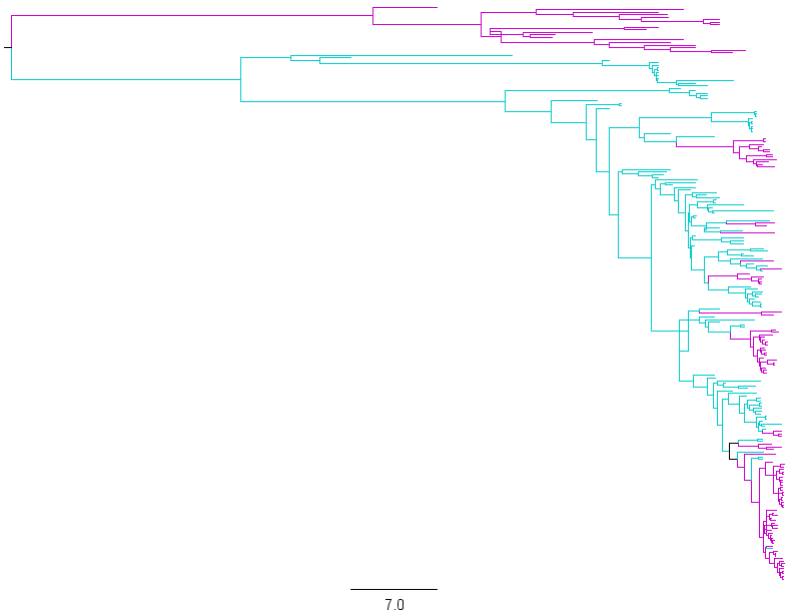

Figure 3.9

The MCC tree of PB1 for swine H3N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

3.12.1.2 Distance in topology results

The trees for all pairs of segments were compared and the original and normalised distance in topology measures were calculated (Tables B1-B3, 3.1–3.3, Figures B1-B48). These results show that the comparisons of coloured trees for each segment for the detection of reassortment are reflected in the distance in topology analysis. For H1N1, the largest values (so the most of reassortment) are seen for all other segments with polymerase segments (PB1, PB2, PA) and both glycoproteins (HA, NA). Similarly, for H1N2 polymerase segments reassort frequently with HA, NA and NS. Other segments reassort mostly with all three polymerase segments and HA. For H3N2 the segments reassorting the most are polymerase segments and both glycoproteins (similar to H1N1).

Table 3.1

Normalised distance in topology scores for swine H1N1 using dist.topo (PH85).

Segment	PB2	PB1	PA	HA	NP	NA	MP	NS
PB2	1.00	1.92	1.78	1.97	1.79	1.99	1.70	1.75
PB1		1.00	1.90	2.09	1.86	2.06	1.74	1.81
PA			1.00	1.94	1.74	1.92	1.65	1.72
HA				1.00	1.85	1.85	1.70	1.76
NP					1.00	1.83	1.62	1.65
NA						1.00	1.71	1.73
MP							1.00	1.54
NS								1.00

Table 3.2

Normalised distance in topology scores for swine H1N2 using dist.topo (PH85).

Segment	PB2	PB1	PA	HA	NP	NA	MP	NS
PB2	1.00	2.48	2.47	3.05	2.46	2.77	2.38	2.57
PB1		1.00	2.54	2.97	2.49	2.79	2.44	2.59
PA			1.00	3.00	2.54	2.73	2.54	2.70
HA				1.00	2.76	2.34	2.50	2.60
NP					1.00	2.58	2.32	2.42
NA						1.00	2.41	2.41
MP							1.00	2.25
NS								1.00

Table 3.3

Normalised distance in topology scores for swine H3N2 using dist.topo (PH85).

Segment	PB2	PB1	PA	HA	NP	NA	MP	NS
PB2	1.00	1.98	1.89	2.16	1.94	1.98	1.86	1.81
PB1		1.00	1.86	2.13	1.98	1.95	1.84	1.84
PA			1.00	2.12	1.92	1.94	1.84	1.80
HA				1.00	2.10	1.85	1.82	1.98
NP					1.00	1.94	1.81	1.77
NA						1.00	1.74	1.81
MP							1.00	1.64
NS								1.00

The distance in topology measure score was also calculated on randomized tips for comparison (Figures 3.10–3.12, B49-B69). For H1N1, the distance in topology measure score is between 700 and 800, which is the score we would get if 25% of tips were randomised. For H1N2, this score is between 150 and 200, which is again the score if 25% of tips were randomised on a tree. For H3N2, the score for distance

in topology is between 300 and 400, which is the same as 25% of the tips on a tree were randomised. The percentage of randomised tips can be thought of an approximation for reassortment.

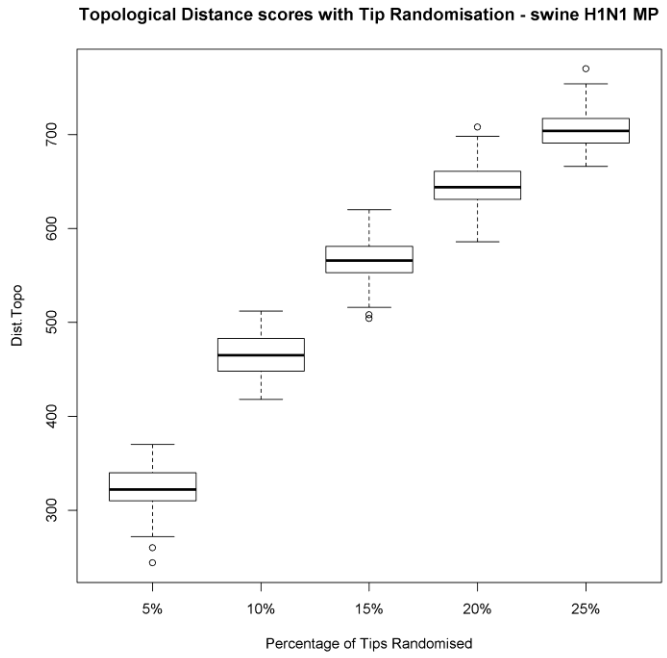

Figure 3.10

Distance in topology for swine H1N1 MP if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

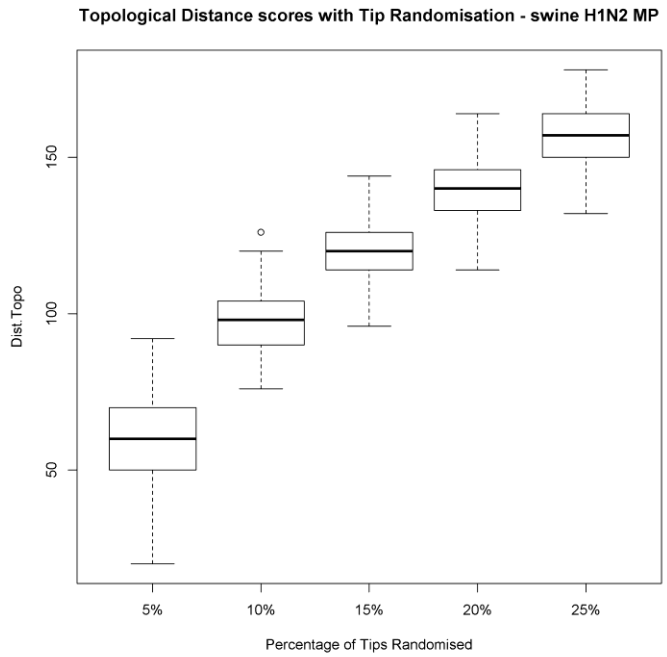

Figure 3.11

Distance in topology for swine H1N2 MP if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

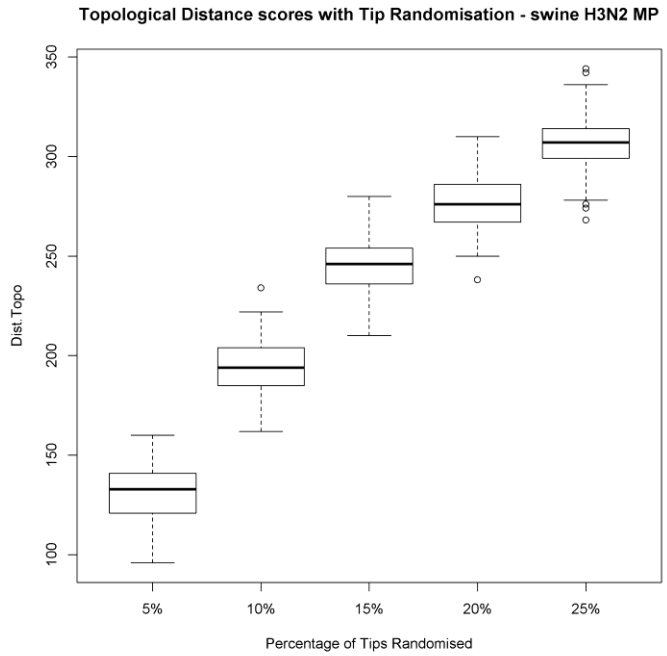

Figure 3.12

Distance in topology for swine H3N2 MP if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

Third, for Wilcoxon test, all the numbers in the table were divided by 810000 because the segment self-comparisons (i.e. MP-MP) to any other segment pair are 810000 regardless of the dataset (H1N1, H1N2, H3N2). Then the self-comparisons (on the diagonal) are 0.5 (i.e. half of the samples are higher than the other half). Tables with results from the Wilcoxon test (B4–B27) and tables with normalised values from the Wilcoxon test (Tables 3.4–3.6 and B28–B48) are asymmetric, therefore the direction of reading them is important. The segment pair comparisons in these tables should be read as a row against a column.

After applying the Bonferroni correction, the results become significant if their p-value is ≤0.0008. In this study, there are no results with a p-value ≤0.0008, therefore, the p-value 0.05 was kept as a threshold of significance. These results are still valid because of a very conservative nature of a Bonferroni correction but are indicative rather than conclusive.

Looking at the normalised MP-X table for swine H1N1 (Table 3.4), where X is any segment, it can be seen that for example MP-NS compared to MP-NA is 807088 (Table B5), so the p-value of this reassortment event is: 807088/810000=0.996 (Table 3.4). MP-NA compared to MP-NS (Table B5) is 2912, so the p-value of this reassortment occurring is: 2912/810000=0.004 (Table 3.4). That means that the segment MP reassorts more frequently with the NA segment than with the NS segment because the p-value for the MP-NA reassortment event is lower and also significant compared to the MP-NS pair, where the p-value is insignificant.

Considering the same three segments (MP, NA, NS) for swine H1N1 in another order, it can be seen that NA-MP versus NA-NS (Table B29) gives a probability 0.649 and for NA-NS versus NA-MP the p-value is 0.350 (both p-values insignificant). Therefore, when evaluating if one segment reassorts with the other, it is important to consider the direction of reassortment, where the segment is either a donor or a recipient.

Normalised values from the Wilcoxon test for MP pairs for H1N2 and H3N2 are given in tables 3.5 and 3.6, respectively.

Table 3.4

Normalised values from the Wilcoxon test for MP H1N1 pairs.

p-value	MP-HA	MP-MP	MP-NA	MP-NP	MP-NS	MP-PA	MP-PB1	MP-PB2
MP-HA	0.5	1	0.462	0.883	0.988	0.760	0.264	0.508
MP-MP	0	0.5	0	0	0	0	0	0
MP-NA	0.538	1	0.5	0.917	0.996	0.803	0.291	0.123
MP-NP	0.117	1	0.083	0.5	0.903	0.274	0.030	0.123
MP-NS	0.012	1	0.004	0.097	0.5	0.028	0.0003	0.011
MP-PA	0.240	1	0.197	0.726	0.972	0.5	0.078	0.256
MP-PB1	0.736	1	0.709	0.970	0.999	0.922	0.5	0.733
MP-PB2	0.492	1	0.454	0.877	0.989	0.744	0.267	0.5

The segment pair comparisons in the table are to be read as a row against a column.

The results of Wilcoxon test with significant p-values are indicative of reassortment. Only p-values lower than 0.10 are listed here for H1N1 due to a large number of segment pairs with a p-value lower than or equal to 0.050: HA-PA vs HA-MP (0.005), HA-PB1 vs HA-MP (0.001), HA-PB1 vs HA-NS (0.003), MP-NA vs MP-NS (0.004), MP-PB1 vs MP-NS (0.0003), NA-PA vs NA-MP (0.001), NA-PA vs NA-NS (0.007), NA-PB2 vs NA-MP (0.003), NA-PB2 vs NA-NS (0.005), NP-HA vs NP-MP (0.007), NP-NA vs NP-MP (0.005), NP-PB1 vs NP-MP (0.004), NP-PB1 vs NP-NS (0.008), NS-HA vs NS-MP (0.004), NS-NA vs NS-MP (0.002), NS-PA vs NS-MP (0.002), NS-PB2 vs NS-MP (0.003), PA-HA vs PA-NS (0.006), PA-NA vs PA-NS (0.004), PA-PB1 vs PA-NS (0.007), PB1-HA vs PB1-NS (0.003), PB1-NA vs PB1-NS (0.002), PB2-NA vs PB2-MP (0.004), PB2-NA vs PB2-NS (0.008), PB2-PB1 vs PB2-MP (0.007) (Tables 3.4 and B28-B34).

That means that the first pair of segments in the comparison reassorts significantly more frequently than the second pair of segments (Figure 3.13). Most frequently reassorting segments in H1N1 (second segment in the first pair of segments, red circle in Figure 3.13) are NA, PB1, HA, PA, PB2, respectively. Least frequently reassorting segments in this subtype (second segment in the second pair of segments, blue circle in Figure 3.13) are MP and NS.

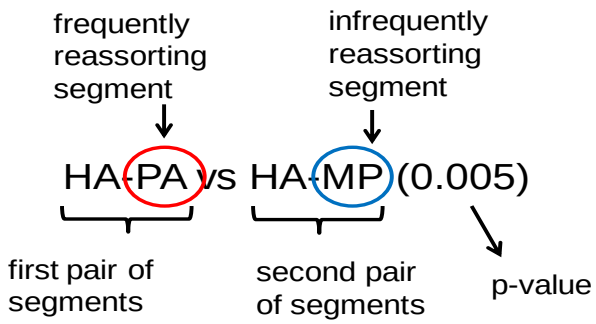

Figure 3.13

Two segment pairs compared in terms of reassortment with the p-value showing the significance of the first segment pair reassorting more frequently than the second pair of segments. The red circle indicates the segment reassorting frequently and the blue circle indicates the segment reassorting infrequently.

Table 3.5

Normalised values from the Wilcoxon test for MP H1N2 pairs.

p-value	MP-HA	MP-MP	MP-NA	MP-NP	MP-NS	MP-PA	MP-PB1	MP-PB2
MP-HA	0.5	1	0.655	0.780	0.872	0.435	0.589	0.696
MP-MP	0	0.5	0	0	0	0	0	0
MP-NA	0.345	1	0.5	0.643	0.756	0.295	0.440	0.554
MP-NP	0.220	1	0.357	0.5	0.625	0.187	0.311	0.417
MP-NS	0.128	1	0.244	0.375	0.5	0.109	0.209	0.304
MP-PA	0.565	1	0.705	0.813	0.891	0.5	0.643	0.738
MP-PB1	0.411	1	0.560	0.689	0.791	0.357	0.5	0.606
MP-PB2	0.304	1	0.446	0.583	0.696	0.262	0.394	0.5

The segment pair comparisons in the table are to be read as a row against a column.

For H1N2 segment pairs only two such pairs had a p-value lower than 0.10: HA-PA vs HA-NA (0.007) and HA-PB2 vs HA-NA (0.005) (Table B35). Other segment pairs with a p-value between or equal to 0.010 and 0.050 are: none for MP segment, NA-PB1 vs NA-HA (0.050), NP-HA vs NP-MP (0.036), NS-PA vs NS-MP (0.028), PA-HA vs PA-MP (0.049), PA-HA vs PA-PB2 (0.044), PB1-HA vs PB1-MP (0.034), PB2-HA vs PB2-MP (0.011), PB2-HA vs PB2-NP (0.024), PB2-HA vs PB2-PA (0.035), PB2-HA vs PB2-PB1 (0.040) (Tables 3.5 and B35-B41).

Looking at the list above to determine the most frequently and infrequently reassorting segments according to figure 3.13 it can be seen that the segment reassorting most in H1N2 is HA, followed by PB1, PB2, PA. The segment reassorting least is MP.

Table 3.6

Normalised values from the Wilcoxon test for MP H3N2 pairs.

p-value	MP-HA	MP-MP	MP-NA	MP-NP	MP-NS	MP-PA	MP-PB1	MP-PB2
MP-HA	0.5	1	0.796	0.524	0.976	0.424	0.254	0.354
MP-MP	0	0.5	0	0	0	0	0	0
MP-NA	0.204	1	0.5	0.208	0.875	0.146	0.063	0.108
MP-NP	0.476	1	0.792	0.5	0.977	0.393	0.222	0.321
MP-NS	0.024	1	0.125	0.023	0.5	0.015	0.003	0.008
MP-PA	0.576	1	0.854	0.607	0.985	0.5	0.305	0.422
MP-PB1	0.746	1	0.937	0.778	0.997	0.695	0.5	0.621
MP-PB2	0.646	1	0.892	0.679	0.992	0.578	0.379	0.5

The segment pair comparisons in the table are to be read as a row against a column.

In H3N2 the pairs of segments with significant p-values lower than 0.010 are: HA-NP vs HA-MP (0.003), HA-PA vs HA-MP (0.003), HA-PB1 vs HA-MP (0.002), HA-PB2 vs HA-MP (0.001), HA-PB2 vs HA-NA (0.007), MP-PB1 vs MP-NS

(0.003), MP-PB2 vs MP-NS (0.008), NP-HA vs NP-MP (0.001), PA-HA vs PA-MP (0.003), PA-HA vs PA-NS (0.002), PB1-HA vs PB1-MP (0.002), PB1-HA vs PB1-NS (0.002), PB1-HA vs PB1-PA (0.006), none for NA, NS and PB2 segments (Tables 3.6, B42-B48).

The list above of segment pairs with significant p-values shows that the most frequently reassorting segment in H3N2 (according to Figure 3.13) is HA and least frequently reassorting segments are MP and NS.

Because the distance topology score (not normalised) depends on the segment length (the shorter the segment length, the larger the distance topology score), simulated sequences were generated based on the real influenza virus sequences in SeqGen (adapted by Dr S Lycett). Shorter segments could have higher distance topology scores because the tree inference is more uncertain for shorter segments. To test this, sequences of different lengths were simulated using the same true tree and then the distance topologies done. When comparing real versus simulated sequences (Figures B70 – B72), we see the largest values of distance in topology for MP-MP and NS-NS pairs for H1N1, H1N2 and H3N2, which supports the above findings of MP and NS being the segments with lowest tree accuracy.

3.12.1.3 Reassortment detection using GARD

GARD is a method for recombination detection on a set of sequence alignments. It is based on likelihood and uses a genetic algorithm for identifying recombinant sequences (Kosakovsky Pond et al., 2006). GARD analysis has been conducted for each pair of segments for swine H1N1, H1N2, H3N2 separately in order to further quantify the extent of reassortment in swine influenza A viruses. c-AIC (not shown) and Δc-AIC values (Tables 3.7–3.9) have been extracted. c-AIC values are dependent on the length of the segment but Δc-AIC values are not and are therefore easier to interpret (personal obsevation). For H1N1 (Table 3.7) the highest Δc-AIC values – therefore most reassortment - are for H1N1 as follows: HA-PB2, HA-PA, HA-PB1, HA-NP, NA-PA, NA-PB2, NA-PB1. For H1N2 (Table 3.8) the highest Δc-

AIC values are: HA-PA, HA-PB2, HA-PB1, HA-NP, NA-PB2, NA-PA, HA-MP, NA-PB1, NA-NP. For H3N2 (Table 3.9) the highest Δc-AIC values are: HA-NP, HA-PB1, HA-PB2, HA-PA, NA-PB2, NA-NP, NA-PA, NA-PB1. For all three subtypes there is a similar pattern of segments reassorting most strongly. Most frequent reassortment events are between HA and polymerase segments, between HA and NP, and between NA and polymerase segments.

Focusing on polymerase segments only, GARD reveals that the most frequent reassortment events are between the PB2 and HA (H1N1), PA and HA (H1N2), PB1 and HA (H3N2). Comparing the Δc-AIC values for polymerase segments against all other segments, it is evident that out of the polymerase segments, it is the PB1 segment that displays highest values (most reassortment) for H1N2 and H3N2. For H1N1, PB1 and PB2 perform equally well and better than PA.

These results support the results of a Wilcoxon test and the distance in topology score results.

Table 3.7

The results for GARD H1N1 pairs of sequences (Δc-AIC).

Δc-AIC	HA	MP	NA	NP	NS	PA	PB1	PB2
HA	0	4257.76	4709.34	6995.99	4818.44	8574.56	8165.21	10073.2
MP		0	3228.59	2346.82	1814.6	2764.35	2134.51	1994.64
NA			0	4301.25	3283.14	6909.35	6388.67	6751.57
NP				0	1933.41	2828.41	2925.71	2388.99
NS					0	1982.97	1301.14	2576.46
PA						0	3294.23	2188.77
PB1							0	4656.17
PB2								0

Table 3.8

The results for GARD H1N2 pairs of sequences (Δc-AIC).

Δc-AIC	HA	MP	NA	NP	NS	PA	PB1	PB2
HA	0	6867.09	3947.14	7744.54	5157.88	10997.4	8244.31	10946.6
MP		0	4564.51	2211.65	2198.05	3332.19	3548.18	2424.09
NA			0	5494.45	4029.06	7156.75	5574.91	7289.85
NP				0	2603.72	2951.54	3943.11	2695.61
NS					0	3215.44	3765.65	3109.01
PA						0	3521.81	2455.7
PB1							0	3693.31
PB2								0

Table 3.9

The results for GARD H3N2 pairs of sequences (Δc-AIC).

Δc-AIC	HA	MP	NA	NP	NS	PA	PB1	PB2
HA	0	6994.29	5589.76	12206.7	7413.91	9233.99	10933.1	9770.62
MP		0	5402.13	3450.16	2519.93	4254.79	5854.09	5214.91
NA			0	9592.46	6612.76	9458.07	9024.56	9815.32
NP				0	4844.93	5412.69	7858.96	7459.45
NS					0	4280.26	5080.84	3443.92
PA						0	5627.42	6742.35
PB1							0	7116.92
PB2								0

3.12.2 Between subtypes reassortment

3.12.2.1 Large scale clade structure and reassortments

To visually assess for the recombination events of influenza virus segments between different subtypes, neighbour-joining (NJ) trees were constructed for H1, H3, N1, N2 separately (Figures C1-C4) and for each internal segment separately with all three subtypes combined (Figures 3.14-3.19). The tree topologies were visualised in FigTree v1.3.1 (Rambaut, 2007). Taxa were coloured according to joined HA and NA subtype. Tree topologies and colour mixing were compared to assess for between subtype reassortment for glycoproteins (Figures 3.14–3.19). Green colour indicates H1N1 subtype, orange colour is for H1N2, violet for H3N2. All internal segments that have sequences from all three subtypes included, reveal good mixing in colour, meaning there is reassortment among internal segments and HA and NA from

different subtypes. Tree topologies differ when compared against each other. MP and NP trees are more similar to each other and trees for polymerase segments show some similarities too. This indicates that segments have a common evolutionary history but also a history of reassortment events.

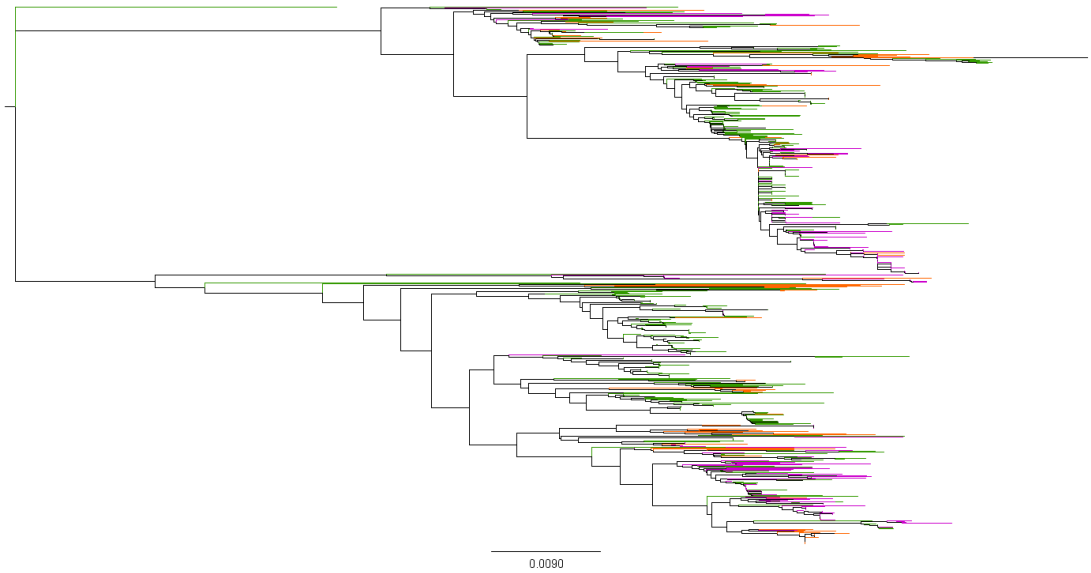

Figure 3.14

NJ tree for MP (H1N1, H1N2 and H3N2 subtypes sequences). Green colour is for H1N1 subtype, orange colour is for H1N2 subtype, violet colour is for H3N2 subtype.

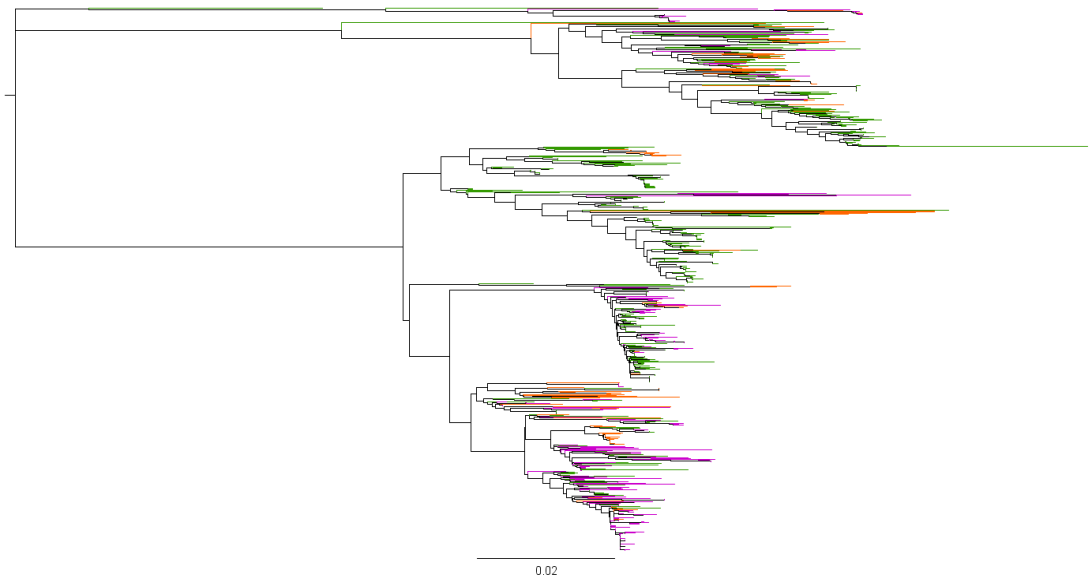

Figure 3.15

NJ tree for NP (H1N1, H1N2 and H3N2 subtypes sequences). Green colour is for H1N1 subtype, orange colour is for H1N2 subtype, violet colour is for H3N2 subtype.

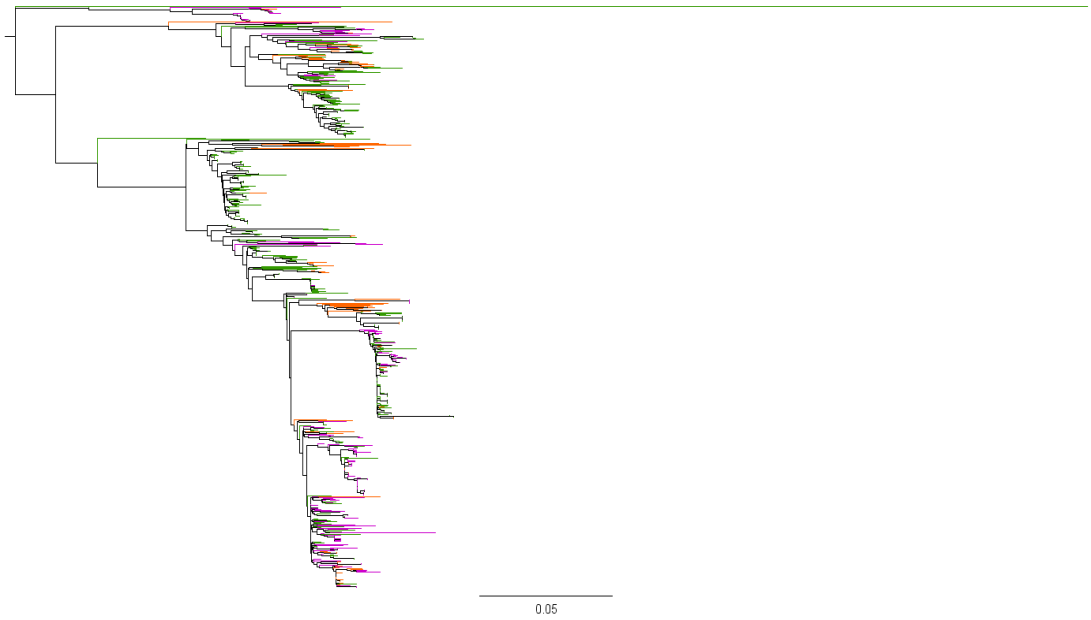

Figure 3.16

NJ tree for NS (H1N1, H1N2 and H3N2 subtypes sequences). Green colour is for H1N1 subtype, orange colour is for H1N2 subtype, violet colour is for H3N2 subtype.

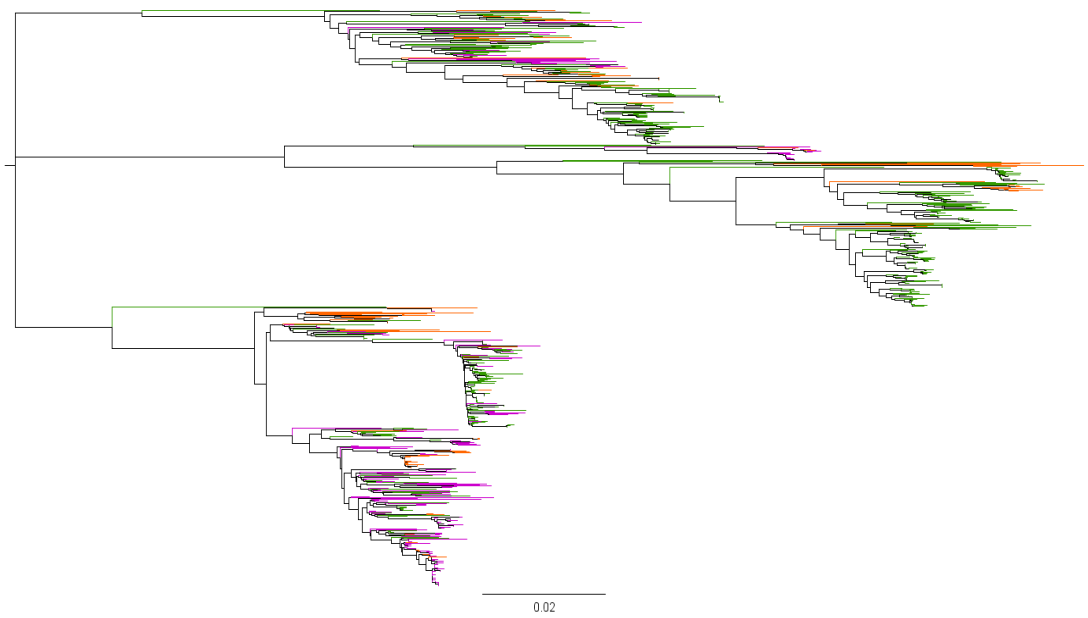

Figure 3.17

NJ tree for PA (H1N1, H1N2 and H3N2 subtypes sequences). Green colour is for H1N1 subtype, orange colour is for H1N2 subtype, violet colour is for H3N2 subtype.

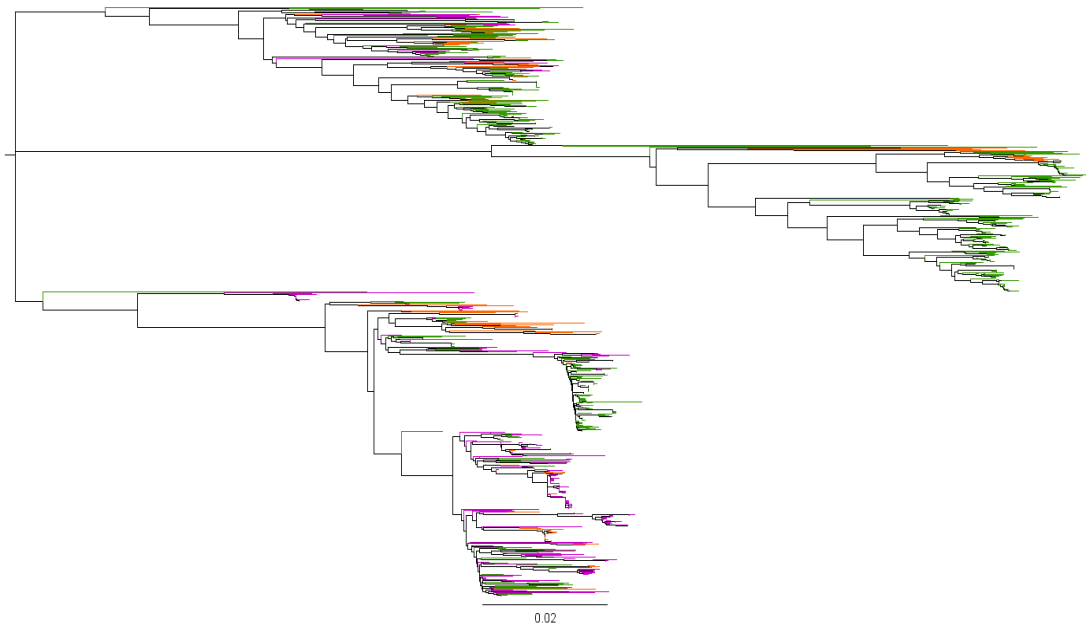

Figure 3.18

NJ tree for PB1 (H1N1, H1N2 and H3N2 subtypes sequences). Green colour is for H1N1 subtype, orange colour is for H1N2 subtype, violet colour is for H3N2 subtype.

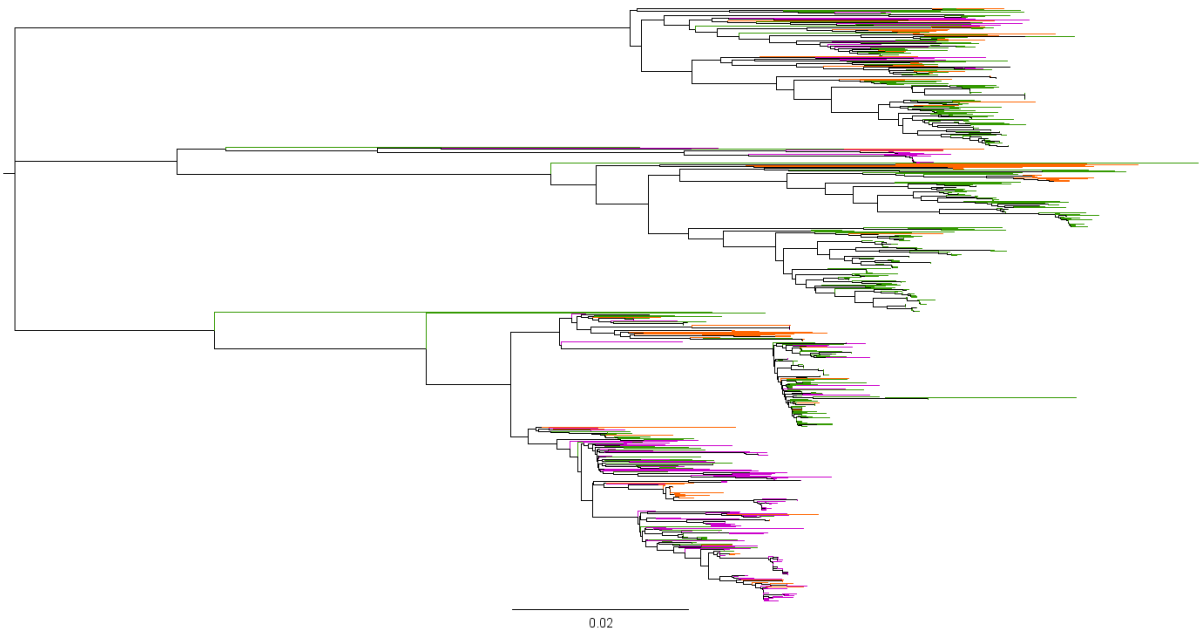

Figure 3.19

NJ tree for PB2 (H1N1, H1N2 and H3N2 subtypes sequences). Green colour is for H1N1 subtype, orange colour is for H1N2 subtype, violet colour is for H3N2 subtype.

3.13 Discussion

The aim of this chapter was to identify the reassortment patterns between different segments of swine influenza A viruses within three separate subtypes, and to quantify the amount of reassortment found. The questions asked were: Is there more reassortment among surface glycoproteins HA and NA compared to internal segments? Which segment pairs are most likely to reassort and what is the probability of this reassortment? The dataset comprised 507 H1N1 sequences, 115 H1N2 sequences and 226 H3N2 sequences (all full genome). The methods used were the measure of distance in topology of trees of different segments, and GARD. The first method was supported by a statistical test (Wilcoxon test).

First, MCC trees for each segment were constructed and coloured according to different clades in MP. MP segment was chosen because mutations conferring drug resistance occur in this segment. Unlike NA where drug resistance mutations appear

also, matrix proteins M1 and M2 encoded by the MP segment are highly conserved among influenza A subtypes (Tompkins et al., 2007). This approach was the first step in determining if there is any reassortment present and it revealed there is reassortment happening within different subtypes. To get a better insight on the reassortment event, it would be better to colour initially not just the MP tree and compare other trees against it but also initially colour trees based on other segments and compare the rest to them (all possible combinations).

Second, to quantity the reassortment found, the sets of 1000 BEAST trees were compared against each other to estimate the distance in topology of these trees for all segment pairs. For H1N1 and also H3N2 the results were not overlapping due to large datasets but for H1N2 the confidence intervals were overlapping due to smaller dataset. However, this lower number of sequences represents all of the sequences submitted to both NCBI and GISAID databases on a date of downloading. The study presented in this chapter focused mainly on within subtype reassortment which is much more difficult to quantify than between subtype reassortment because there are no subtype labels to use in case of within subtype reassortment, so a discrete trait analysis is not possible. That is why the dist.topo measure approach is appropriate. For example, the study by Lu et al (2014), which investigated the reassortment patterns of internal segments between different subtypes of avian influenza, was based on a discrete traits analysis with HA, NA or joined HA-NA subtypes being the discrete traits on MCC trees for six internal segments. In this case, the discrete trait analysis was possible because of a large number of subtypes that were used as discrete traits. Similarly, Lu et al (2014) assessed the reassortment rates of six internal segments based on HA, NA and joined HA-NA subtypes. In contrast, the analysis of reassortment described in this chapter compared all the segments against each other.

Third, GARD analysis was used to detect evidence for reassortment. This type of analysis was used in the study by Lycett et al (2012) but not in the study of Nelson et al (2012) and Khiabanian et al (2009). Lycett et al (2012) estimated reassortment rates between swine H1N1, H1N2 and H3N2 influenza viruses' neuraminidase and

haemagglutinin subtypes on a polymerase backbone, in terms of exchanges/year focusing on Eurasian swine influenza viruses only. Nelson et al (2012) studied reassortment in North American swine influenza viruses by performing the phylogenetic analysis on whole genome sequences and using different subtypes of swine influenza as discrete traits. The study by Khiabanian et al (2009) focused on all three subtypes from different geographic regions. However, most of the strains in the study by Khiabanian et al (2009) were isolated in USA. The study described in this chapter encompassed swine influenza viruses of subtypes H1N1, H1N2 and H3N2 (like in the study by Lycett et al (2012) and Khiabanian et al (2009)), however the sequences were isolated from different parts of the world. Also, this study focused largely on intrasubtype reassortment. The reason for this is that GARD is slow and unable to analyse larger (more than few hundred sequences) datasets.

The results presented here show that the segments most prone to reassortment are HA and NA as well as all three polymerase segments for all three subtypes. This is consistent with the results by Khiabanian et al (2009) who found that HA and NA reassort more frequently than internal segments but also PB1 reassorts quite frequently. The disadvantage of the study described in this chapter was its focus largely on reassortment within three subtypes separately, which showed very similar results. Because swine are considered to be a mixing vessel for reassortment, therefore the host with the largest probability of harbouring reassortment events, the results of this study may not reflect the overall reassortment patterns and their extent. However, the methods used for this study would not be appropriate for the study of between subtype reassortment. Namely, the sets of 1000 BEAST trees can be constructed for a joined dataset of all three subtypes but when used for the distance in topology measure calculation, only sequences of six internal segments could be compared against each other. The reason for that is when comparing a set of trees for an internal segment against an external segment (e.g. MP vs HA-H1) a set of phylogenetic trees made on a set of MP sequences from all three subtypes would be compared against a set of phylogenetic trees made on a set of HA-H1 sequences, which would include only subtypes H1N1 and H1N2 but not H3N2. Next problem when analysing the between subtype reassortment with methods used in this study

would be a small dataset (less than 300 sequences) that is required for the GARD analysis. Therefore, the existing datasets of 507 sequences for H1N1, 115 sequences for H1N2 and 226 sequences for H3N2 would have to be subsampled with more than half of the sequences not included in the joined dataset. That would mean a huge loss of information.

This study of the reassortment focused on the MP tree for each subtype, which revealed moderate level of reassortment. If the colouring was to be done according to an HA tree it would probably show different results because in swine influenza viruses the primary heritable units are (internal) polymerase segments which acquire external HA and NA through reassortment. Therefore, different swine influenza viruses with the same polymerase unit can have different HA and NA molecules. Because MP is the most infrequently reassorting segment (together with NS), different results for MP coloured on the basis of HA tree can be expected as HA is the most frequently reassorting segment. For between subtype reassortment, MP and other internal segments' trees were constructed and coloured according to joined HA-NA subtype. These trees revealed moderate levels of reassortment of internal segments between subtypes based on colour mixing. Therefore, a higher level of between subtype reassortment can be expected than within subtype reassortment.

Comparing results of MP vs polymerases shows that MP reassorts infrequently with polymerases within subtype but between subtype this particular comparison (MP vs polymerases) is not possible based on the coloured phylogenetic trees for internal segments (Figures 3.14 – 3.19) because trees are coloured against HA-NA subtypes.

The high susceptibility of swine for infection with different influenza viruses due to high levels of appropriate receptors for infection, poor vaccination of swine and large numbers of swine at farms all contribute to swine being the perfect environment for reassortment events to occur. Influenza virus infections are so highly prevalent in swine causing endemic infections. Also infections of wild birds with influenza A viruses are endemic due to birds being the natural reservoir for the huge majority of influenza A virus subtypes. Therefore, the evolution of influenza viruses infecting

swine or wild birds is different from the evolution of influenza viruses infecting humans or domestic birds. In humans, HA and NA are major heritability units and because there is not much reassortment happening in human influenza viruses, HA and NA evolution corresponds largely to the evolution of internal segments. In birds, there is also a vast amount of reassortment happening among internal segments with the exception of low pathogenic H5N1 and H9N2 subtypes, which do not reassort very much, probably due to their high prevalence in domestic birds, which are more physically isolated from other birds than in natural environment offering therefore less opportunity for reassortment (Lu et al., 2014).

The results of the study described in this chapter showed that HA segment is the most frequently reassorting segment and is reassorting mostly with polymerase segments. The second segment reassorting very frequently is NA and it is also reassorting mostly with polymerase segments. The study also showed that the segments reassorting very infrequently are MP and NS.

The contribution of work described in this thesis chapter is in the quantification of reassortment events in swine influenza A viruses with some of the methods not used before for this purpose. Also, most of the studies on reassortment in swine influenza A viruses focused only on a certain geographical region and used smaller datasets with sequences spanning shorter time range while the study described in this chapter considered global swine influenza A sequences with a large dataset isolated over 43 years. The importance of this study lies in the identification of segments frequently reassorting together, in the identification of segments reassorting least and in the probability of these events. Namely, some of other studies only identified segments with highest rates of reassortment but not the segments least likely to reassort. Overall, this study confirmed previous studies which found HA and NA segments to be the most frequently reassorting segments and also provided evidence that these two segments do not reassort together but with the polymerase segments which some other studies did not point out. This study also provided evidence that MP and NA segments which were found reassorting together when the A/H1N1pan09 was formed, reassort very infrequently therefore the reassortment between these two

segments is not a general event. So, this study used some novel methods and provided more detailed insight in the reassortment patterns within swine influenza A viruses.

Chapter 4

Amantadine resistance in swine influenza A viruses

4 Amantadine resistance in swine influenza A viruses

4.1 Adamantane drug resistance in influenza A viruses

Anti-influenza drugs are used in humans for the treatment and prevention of influenza disease caused by seasonal epidemic strains and occasionally by pandemic ones (Sheu et al., 2008; Maltezou, 2008; Hota and McGear, 2007). There are two classes of antiviral drugs available: the adamantanes amantadine and rimantadine (blocking the ion channels of M2 in the viral envelope and preventing virion uncoating and the release of genome segments into the cytoplasm) and neuraminidase inhibitors oseltamivir and zanamivir (blocking the neuraminidase and preventing budding of newly synthesized virions from the host cell). However, H3N2 resistance to adamantanes rose dramatically worldwide during the 2005-2006 influenza season thus influenza prophylaxis and treatment predominantly relied upon neuraminidase inhibitors only (Maltezou, 2008; Hota and McGear, 2007; Tang et al., 2008; Deyde et al., 2007; Kiso et al., 2004; Bright et al., 2005). The use of adamantanes is restricted by a rapid emergence of resistant strains that retain full virulence and transmissibility. Drug-resistance conferring mutations in M2 can occur at five different amino acid sites. One out of these 5 mutations is enough for the virus to develop adamantane resistance (Boivin et al., 2002; Hay et al., 1986). Amantadine and rimantadine resistant mutants are cross-resistant with no evidence of fitness impairment (Belshe et al., 1988).

The primary way for the prevention of influenza infections in humans is vaccination. When vaccination is unavailable or in cases of individuals who have not been or cannot be vaccinated, antiviral agents serve as an alternative. For many years the only antiviral agents available were amantadine and rimantadine (Belshe et al., 1988; Dolin et al., 1982; Tominack and Hayden, 1987). Despite the long and widespread use of adamantanes in humans, the global frequency of resistant strains in A/H3N2, isolated between 1991 and 1995, remained very low (0.8%) (Ziegler et al., 1999). However, in the year 2004 the amantadine resistance in A/H3N2 rose to 12.3% (Bright et al., 2005). By next year, different countries reported higher incidence of

resistant strains: 14.5% in the United States, 72% in South Korea and 96% in China (Bright et al., 2005). On the other hand, amantadine resistance in A/H1N1 remained much lower: on a global scale it reached 15.5% of all strains isolated (Deyde et al., 2007).

4.2 Structure and mechanism of M2 ion channel of influenza A virus

M2 protein contains 97 residues with its N- and C-termini directed toward the outside and inside the virion, respectively. It contains one internal hydrophobic transmembrane domain of 19 residues that spans the membrane once, 23 N-terminal extracellular residues – without the N-terminal methionine residue which is removed (Tobler et al., 1999), and 54 residues that form the cytoplasmic tail (Lamb et al., 1985). M2 is a homotetramer in its native state (Sugrue and Hay, 1991; Holsinger and Lamb, 1991). It is an integral membrane protein which functions as a pH-regulated selective proton channel sensitive to the pH environments (Lamb et al., 1994; Holsinger et al., 1994). The channel is formed by the four transmembrane helices where the His37 residue is the pH sensor and Trp41 is the gate (Pinto et al., 1992a; Tang et al., 2002; Pinto et al., 1997b).

There are three different sites where drug resistance mutations are known to appear in the M2 protein: (i) mutations that face the interior of the channel pore and are all located on the N-terminus and close to pore-binding site (V27A, A30T, S31N, G34E); (ii) mutations in spaces between the helices that are located at the N-terminus and are close to the pore-binding site (L26F); (iii) mutations in spaces between the helices that are located at the C-terminus which is outside the pore-binding site but may affect it allosterically (L38F, D44A). There are also mutations of pore-facing residues His37 and Trp 41 occurring but are deleterious to the virus because these two residues are crucial for proton conductance and channel gating (Figure 4.1) (Gu et al., 2013).

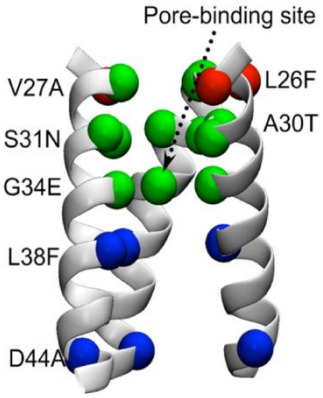

Figure 4.1

Drug-resistance mutations on different positions of the M2 protein. In green, the mutations of residues at the pore-facing positions (V27A, A30T, S31N, G34E). In red, mutations in spaces between the helices that are located at the N-terminus and are close to the pore-binding site (L26F). In blue, mutations at the C-terminal site. The pore-binding site is indicated by an arrow. For clarity, only three helices of M2 are shown (from Gu et al., 2013).

4.3 M2 ion channel inhibitors: amantadine and rimantadine

Two M2 ion channel inhibitors, amantadine (1-adamantanamine hydrochloride, Symmetrel™, Lysovir™, Symadine™) and its close analog rimantadine (α-methyl-1-adamantanemethylamine hydrochloride, Flumadine™, Roflual™) are in use for influenza indications. The structure of amantadine is a tricyclic 10-carbon ring with a primary amine group on the superior pole (Figure 4.2). Amantadine was first shown to possess an anti-influenza activity in 1964/1965 in cell culture and in ferret and mouse model (Davies et al., 1964; Tsunoda et al., 1965). In 1966 it was approved for the prophylaxis and treatment of Asian pandemic influenza A/H2N2 virus infections in the USA. In 1976 it was approved for all influenza A infections after further clinical evaluations.

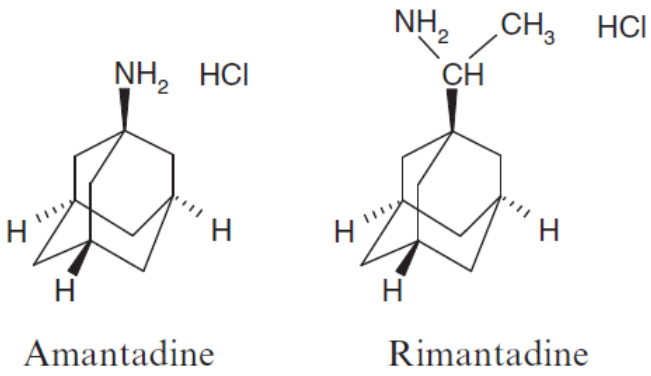

Figure 4.2

Chemical structures of amantadine and rimantadine (M2 ion channel inhibitors) (from Tisdale, 2009).

Rimantadine is a closely related derivative that shares the same hydrocarbon ring structure. The difference from amantadine structure is in incorporation of a carbon with a methyl group between the nitrogen and adamantine ring. Rimantadine was approved for clinical use in 1993 in USA. Both amantadine and rimantadine are very stable after long term storage (over 25 years at ambient temperature) and retain full antiviral activity (Scholtissek and Webster, 1998).

4.4 Agricultural amantadine usage

Influenza viruses are important pathogens of not only humans but also poultry and swine. Swine influenza virus infection causes an acute and highly contagious respiratory disease and is common in pig populations globally due to the incomplete vaccination in pigs as well as the absence of the use of the drug in pigs. The anti-influenza drug amantadine has however been widely used for agricultural purposes in domestic poultry. After the 1983 outbreak of avian influenza in chickens in Pennsylvania, amantadine was routinely added to water supply for commercial poultry as a means of preventing the avian influenza infections in USA (Webster et al., 1985). Amantadine is licensed only for human use, however, its use in Chinese poultry production has been reported (WHO, 2013). It is believed that the amantadine resistance of influenza viruses appeared in avian influenza because of

wide use of this drug by Chinese poultry farmers in late 1990's in order to limit avian influenza infections. Chinese companies producing this drug were selling the amantadine to Chinese farmers for 10 dollars per pound of the drug (Ayaz, 2012). Researchers from the US Department of Agriculture Laboratory showed that avian influenza viruses develop drug resistance in a few days after receiving amantadine. The amantadine has been available and used for the treatment as well as prevention of infection with avian influenza viruses. Amantadine resistance of H5N1 was reported in 1997 and the use of this drug in Chinese poultry farming was spread for years prior to that (Ayaz, 2012).

Influenza A viruses can infect many species of mammals and birds. The wild birds are the reservoir for all influenza A subtypes known except for H17N10 and H18N11 which were found in bats (Tong et al., 2012; Tong et al., 2013). Novel influenza viruses can emerge from birds and infect mammalian species including humans. It has been speculated that the domestic pigs serve as the intermediate 'mixing vessel' for some human and avian viruses because the respiratory epithelium of pigs possesses sialic acid receptors for both avian and human viruses. This way novel reassortant influenza viruses are created in pigs that have the potential to be transmitted to humans. Along with acquiring new genetic segments, the reassortant viruses can acquire drug-resistance conferring mutations also.

4.5 Proposed mechanisms of amantadine drug resistance acquisition and spread

Adamantane resistance has become widely spread in the human A/H3N2 influenza population in the 2005-2006 influenza season. In nearly every case of influenza in US studied, the cause for resistance was a single substitution of serine with asparagine at site 31 (S31N) in the M2 ion channel protein. S31N is one of 5 amino acid replacements in the M2 protein known to be linked to adamantane resistance. Theoretically, two evolutionary mechanisms could be responsible for the global spread of adamantane resistance in human and/or swine influenza virus. First, the

high level of amantadine use in some countries could confer sufficient individual selective advantage to facilitate the spread of S31N mutation (Regoes and Bonhoeffer 2006). Alternatively, the spread of S31N may be unrelated to drug selection pressure and instead result from its interaction with advantageous mutations located elsewhere in the viral genome. Such interactions could take the form of either genetic 'hitch-hiking', which is the process by which an allele may increase in frequency by virtue of being linked to a gene that is positively selected (Barton, 2000). In this case the S31N mutation is pulled to fixation because of its physical linkage to beneficial mutations, such as those in the haemagglutinin protein (HA) that facilitate immune escape (Bush et al. 1999; Smith et al. 2004). Another mechanism is genetic epistasis. Epistasis is a phenomenon in which the fitness of an allele at one locus depends on the presence of an allele at another (see section 5.6). In epistasis natural selection for a characteristic other than adamantane resistance could favour a specific combination of mutations which include S31N.

4.6 Studies of amantadine drug resistance in influenza A viruses

4.6.1 The study by Krumbholz et al (2009)

In the study of Krumbholz et al. (2009), a molecular epidemiological investigation of European swine Eurasian avian-like influenza viruses, analysis of their amantadine resistance phenotype and a genetic analysis of the MP segment was performed. A phylogenetic analysis encompassed human, swine, avian and equine MP sequences. A phylogenetic tree was inferred using Bayesian MCMC analysis. The resulting tree (Figure 4.3) indicates that the MP sequences of the European avian-like clade constitute a monophyletic branch. Krumbholz et al. (2009) found it surprising that all amantadine-resistant swine strains (as deduced from the S31N substitution) cluster in a single clade with A/sw/Schwerin/103/89 (H1N1) being closest to the root. The respective node is supported with high posterior probability (1.0). Beside the S31N mutation, additional substitutions (A26F, L26I, V27A, V27I, A27T) could be traced. None of them emerged independently of the S31N substitution.

Krumbholz et al (2009) investigated the susceptibility of Eurasian avian-like swine influenza A viruses to amantadine. They found that all resistant strains harboured the S31N substitution, which prevents binding of amantadine to the M2 protein. Several isolates contained an additional mutation, V27A, which enables resistance by another mechanism: amantadine-binding is retained but M2 still functions as a proton channel (Astrahan et al., 2004). Krumbholz et al. (2009) made two observations: the emergence of the amantadine resistance in the M2 gene was a single event as demonstrated by a monophyletic resistant clade. So, the susceptible H1N1 strains were replaced by resistant H1N1 strains. They also concluded that a reassortment event took place, which introduced resistance to H3N2 and H1N2 strains. They discuss that pigs in Europe have most likely not been treated with amantadine as there is no documentation of that, so the amantadine resistance rise is probably due to a natural cause. In humans, this resistance has been increasing rapidly in the absence of amantadine use for treatment. Two mechanisms may be responsible: (i) spontaneous mutations with low incidence (Ziegler et al., 1999) and their fixation in the presence of drug treatment (Shiraishi et al., 2003), (ii) hitch-hiking where amantadine-resistance conferring mutation is linked to a mutation at another genomic site that enhances viral fitness. Krumbholz et al (2009) turned down both options as a cause for amantadine resistance in Eurasian swine influenza A viruses but left the possibility of a bottleneck effect (Domingo and Holland, 1997) open for discussion. The study of Krumbholz et al. (2009) also indicated negative selection on codons 30, 31, 34 which are known to confer amantadine resistance. On site 31, only two codons were observed, either coding for serine (amantadine susceptible) or asparagine (amantadine resistant) in human or swine strains. That is probably because S31N is the only mutation that maintains both structural constraints (i.e. formation of a proper transmembrane helix which is able to tetramerize) and functional constraints of the M2 protein (i.e. proton channel activity even in the presence of amantadine). Substitutions of alanine 30 and glycine 34 were not observed in the Eurasian swine influenza viruses.

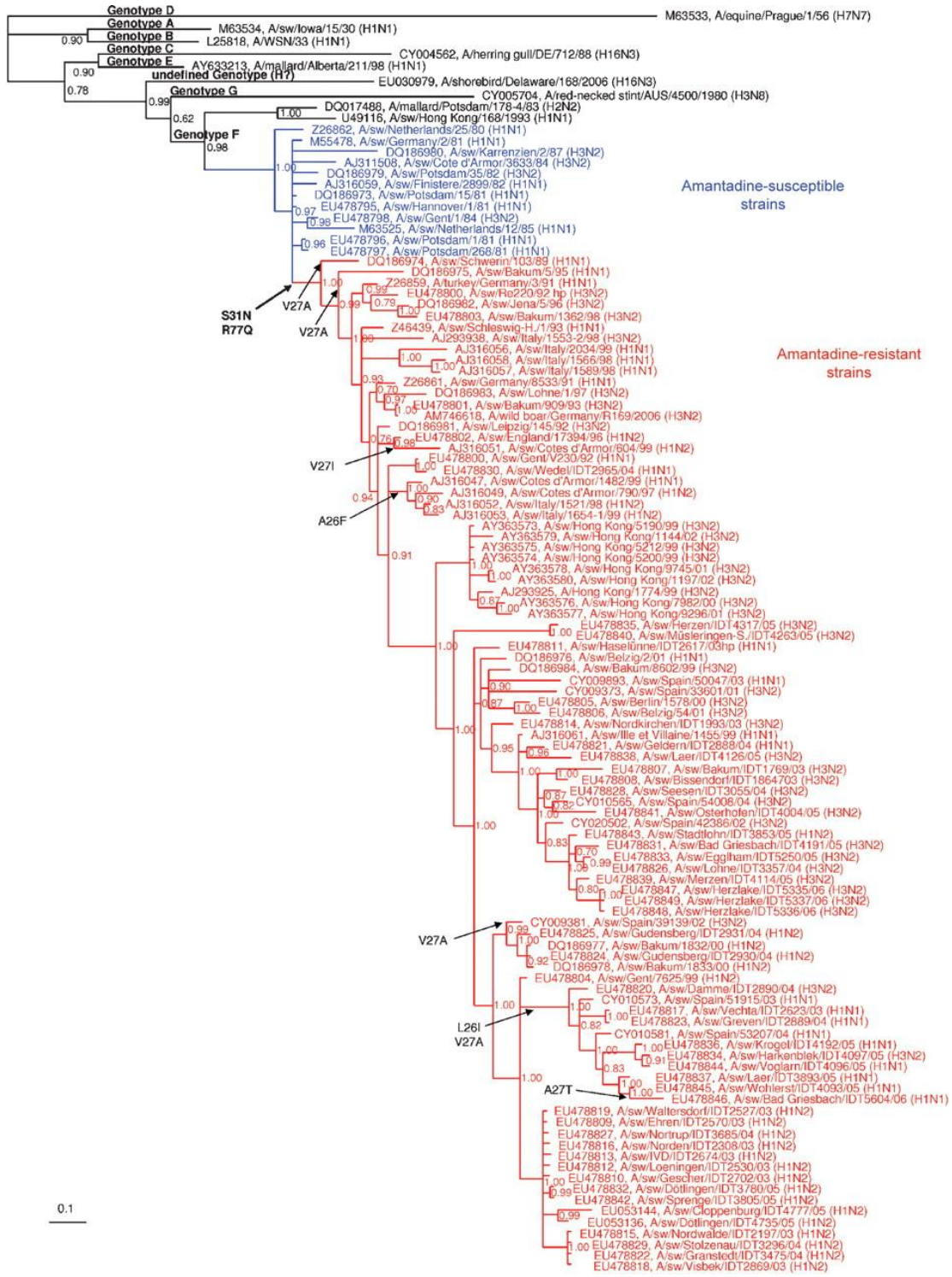

Figure 4.3

Phylogenetic analysis of 113 MP sequences of influenza A virus isolated from Eurasia. Sequences were isolated from different hosts (swine, avian, equine and human) and were of different subtypes. Blue and red colours indicate amantadine susceptible and amantadine resistant strains, respectively (Krumbholz et al., 2009).

The study indicated that Eurasian swine lineage appeared in 1979 [node age: 26.83 years before 2006, 95% highest posterior density (HPD) interval: 26.14−33.39 years before 2006) and amantadine-resistance (S31N substitution) may have emerged in 1988 (node age: 17.8 years before 2006, range of the 95% HPD interval: 17.09−22.55 years before 2006). 1979 as the year of emergence of European swine influenza A viruses was also inferred by Ludwig et al (1995).

Codon-based tests were performed to identify natural selection at single amino acid sites. Positive (diversifying) selection, i.e. excess of non-synonymous substitutions, and negative (purifying) selection, i.e. excess of synonymous substitutions, were investigated using different maximum-likelihood methods (SLAC, FEL and IFEL). Several codons were suggested to be restrained by negative selection: (i) codons 30, 31 and 34 which confer amantadine resistance and (ii) codons 32, 50, 51, 58, 67 and 71. Positive selection was suggested for codon 19 in the IFEL method. All three methods showed a similar pattern of dN−dS differences. However, the SLAC and IFEL methods were found to be more conservative, resulting in fewer suggestions for natural selection. 46 of 97 codons (47.4%) were invariant. The mean substitution rate of this dataset was 0.54 substitutions/site and the mean dN/dS was 0.60. They found that diverse selective forces act on the M2 proteins of porcine and human influenza A viruses.

4.6.2 The study by Furuse et al (2009)

The study by Furuse et al. (2009) investigated the mechanisms for emergence and spread of amantadine resistance by analysing the MP gene of influenza A viruses from different host species. They detected drug resistance-associated mutations on positions 26, 27, 30 and 31 but not at position 34. Also double mutants were found (positions 27 and 31). Most frequent mutation was at position 31 (S31N). The drug amantadine has been available since 1966, however the drug-resistant mutations spread after the year 2000. They were found in different subtypes, isolated from different hosts at different geographical locations. Strains with the mutations at positions 26, 27 and 30 were found only sporadically – they did not spread and

eventually disappeared. The same study also evaluated dN/dS ratios for influenza A viruses isolated from different hosts. The value above 1 means positive selection on the codon and the value below 1 means negative selection (Kosakovsky Pond et al., 2007; Suzuki and Gojobori, 1999). The value of 1 is for neutral selection. dN/dS was above 1 for amino acid sites 26 (1.35) and 27 (4.42) and below 1 for sites 30 (0.21) and 31 (0.83). Furuse et al. (2009) investigated the differences in selective pressures between hosts and found no significant differences between human and swine viruses. Human viruses experienced higher selection pressure at sites 26 and 31 than viruses from other hosts. On sites 27 and 30, human influenza viruses were under smaller selection pressure than avian viruses but the difference was not significant. Furuse et al. (2009) also estimated the change in selective pressure in

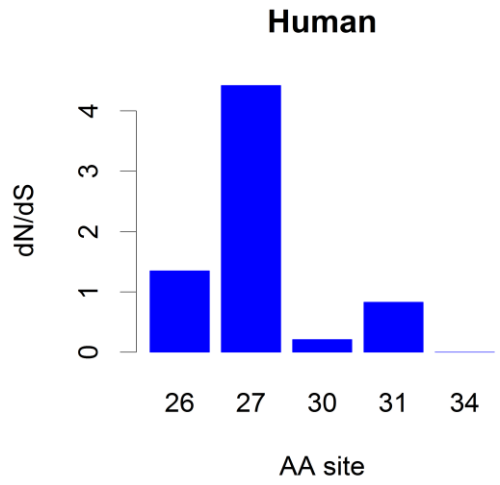

Figure 4.4

dN/dS ratio for human influenza A virus in M2 sequences on amino acid positions where mutations that confer resistance to adamantanes can occur. dN/dS values larger than 1 mean positive selection, dN/dS values smaller than 1 mean negative selection, dN/dS value equal to 1 means neutral selection (adapted from data by Furuse et al., 2009).

human influenza A viruses with time and found that the entire selection pressure became smaller with time for the entire MP segment. The dN/dS value increased since 1966 (when amantadine became available) for amino acid sites 26, 27 and 30, where resistance-conferring mutations can occur (Figure 4.4). For the site 31, the dN/dS value decreased. Furuse et al (2009) also constructed phylogenetic trees of influenza A viruses from different hosts for both lineages – North American clade

and Eurasian avian-like clade which suggested that the acquisition of the drug resistance-conferring mutations is due to point mutations and not by obtaining the MP segment by reassortment from other hosts or subtypes (Figure 4.5). This finding is contradictory to the conclusions of Schmidtke et al (2006) who suggested the reassortment between human and swine viruses to be the cause of the rise and spread of amantadine resistance in humans. Schmidtke et al (2006) also suggested further rise in the amantadine resistance due to further reassortment between swine and human viruses but Furuse et al (2009) found amantadine resistance in human and avian viruses did increase after the year 2000 but it did not come from swine. However, Furuse et al (2009) could not explain the rapid increase in the amantadine resistance, especially after 2000 when the usage of amantadine dropped because of the availability of a new drug oseltamivir.

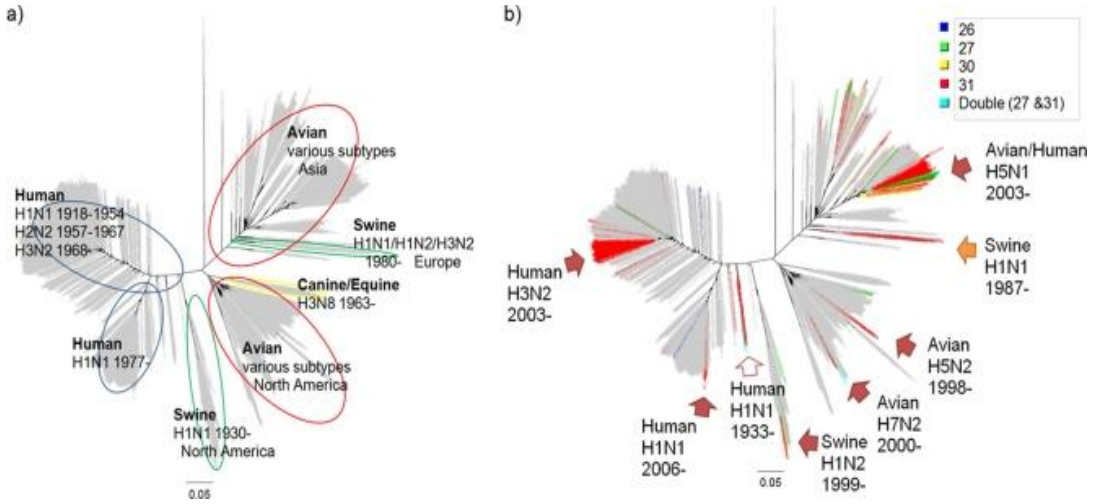

Figure 4.5

Phylogenetic trees for the MP gene segment constructed using RAxML are shown. The trees are marked with host-specific lineages and their profiles (a) and amantadine resistance mutations shaded in colors by mutation positions (b). The arrows in panel b indicate major clusters of amantadine-resistant strains (Furuse et al (2009)).

4.7 The aim of study

In this study the origin of amantadine resistance in swine influenza A viruses was investigated. The initial hypothesis was that resistance came into swine influenza virus from avian influenza viruses infecting wild birds when the avian influenza virus switched host from wild birds to swine host forming the Eurasian avian-like swine clade. We hypothesized that amantadine drug resistance could have been transmitted from avian influenza viruses into swine host once or multiple times due to the agricultural use of the drug in poultry especially in China where wild birds can come into contact with domestic poultry and become infected with the resistant avian virus circulating in poultry. Alternative hypothesis was that resistance appeared in swine influenza virus spontaneously due to potential use of the drug in swine. With the progression of the work, the study focused on the possibility of the spread of amantadine drug resistance-conferring mutations being unrelated to agricultural drug use in swine and/or poultry but being due to epistatic linkage to some other sites in the viral genome that are positively selected for. This way, drug resistance could spread in influenza A virus population with genetic hitch-hiking. Considering that a few studies on this subject have already been made but are contradictory, decision was made to do the study with phylogenetic and bioinformatics methods to elucidate the mechanism of amantadine drug resistance appearance and rise within a swine host.

4.8 Methods and data

Four datasets were used in this study to address different research questions.

4.8.1 Dataset 1

This dataset was created to focus on the prevalence of amantadine drug resistance mutation S31N in the M2 gene of human and swine influenza viruses, isolated in different time periods.

Influenza A sequences were obtained from the NCBI flu database (www.ncbi.nlm.nih.gov/genomes/FLU) using a set of predefined parameters: host (swine or human), country/region (any), segment 7 (MP), subtype (any), sequences length (full-length plus to allow for sequences without a start and/or stop codon); additional filters were: pandemic (H1N1) 2009 viruses (include), the FLU project (include), lab strains (exclude), vaccine strains (exclude), lineage defining strains (include). The FLU project includes sequences that come mostly from large scale flu genome sequencing projects and usually contain complete genomes and detailed source information (http://www.ncbi.nlm.nih.gov/genomes/FLU/help.html#project). Sequences were exported as txt files and automatically aligned using function Muscle. The sequences were then manually checked, nucleotides before the start codon (ATG) and after the stop codon (either TAA, TAG, TGA) were removed so that segments were of appropriate length (982 nucleotides).

On February 1st 2011, all available full-length sequences for matrix proteins (MP) of swine H1N1 (369 sequences), swine H1N2 (93 sequences) and swine H3N2 (106 sequences) were downloaded from the NCBI FLU database. For human influenza virus MP, there were 1.050 sequences of pandemic H1N1, 1.430 sequences of seasonal H1N1 and 2.808 sequences of seasonal H3N2. After aligning and trimming the sequences (so that they start with a start codon and stop with a stop codon) the M2 sequences were extracted from MP and resistance at amino acid position 31 was determined. The sequences were divided into three groups: from 1950 to 1966 (before the availability of amantadine), from 1967 to 1999 (from the availability of amantadine and before the availability of oseltamivir) and from 2000 to 2011 (after the availability of oseltamivir). Sequences from before 1950 were not analysed for prevalence of resistance mutations in M2.

4.8.2 Dataset 2

This dataset was used to answer the question: when and where (which host) did the amantadine drug resistance appear first?

Swine MP sequences were searched for in the NCBI FLU database as described for the Dataset 1. Avian and human influenza A sequences were subsequently manually added to swine influenza A sequences according to the geographic location, avian species, susceptibility or resistance to amantadine and subtype. Alltogether there were 389 sequences, of which 72 were avian, 8 human and 309 were swine. They were also searched for in the NCBI FLU database with BLAST and were collected from 1931 to 2010. With BioEdit resistance codons in segment 7 (MP) were removed to avoid sequences appearing more similar to each other because of potential resistance. Those codons were in M2 gene of the MP segment (which also codes for M1 protein) on positions 26, 27, 30, 31, 34. The resistance mutations on these positions are L26F, V27A, A30T, S31N, G34E.

4.8.2.1 Phylogenetic analysis

Maximum likelihood phylogenetic analysis was conducted on the 389 MP nucleotide sequences (with all five resistance-conferring codons removed) in PhyML on the Haldane cluster with 1000 bootstraps, gamma distributed rates among sites and gamma parameter 4. The output tree file was opened in FigTree v1.3.1 (Rambaut, 2007). Sequences that clustered within the Eurasian avian-like swine clade and North American swine clade were then separated from the other MP sequences, XML files were created via BEAUTi v1.6.1 (Drummond and Rambaut, 2007) and BEAST analysis with discrete traits (strains susceptible or resistant to adamantanes) and Markov jumps was run. Discrete trait analysis enables to determine which was first, susceptibility or resistance on a certain branch and Markov jumps analysis enables to determine where on the branch the change occurred. The parameters used were: GTR nucleotide substitution model, relaxed clock (uncorrelated lognormal) to allow potentially different rates for the human, avian and swine strains and because it was preferred over a strict clock by Bayes factor test (larger than 55), constant demographic model, 200 million generations (the chain length). Two independent runs were then combined in LogCombiner (Drummond and Rambaut, 2007) with removed burn-in of 10%, further down-sampled to 1900 trees and by using

TreeAnnotator (Drummond and Rambaut, 2007) the maximum clade credibility tree (MCC) was annotated.

4.8.3 Dataset 3

This dataset served for the estimation of dN/dS ratio on codons where adamantine resistance can occur. Also overall dN/dS ratio was estimated.

Avian and swine MP nucleotide sequences were searched for in the NCBI FLU database as described for Dataset 1. The downloaded sequences encompassed the entire database as on the date June 4th 2013. The number of sequences was 7,023 of avian MP and 2,817 sequences of swine MP. Maximum likelihood analysis was performed in RAxML on the Haldane cluster for avian and swine sequences separately. Avian sequences were further separated into domestic and wild, and swine sequences were separated into Classical and Eurasian avian-like swine clades. dN/dS was estimated in HyPhy (SLAC) for those four sequence groups.

4.8.4 Dataset 4

The purpose of this dataset was to find the mechanism responsible for the acquisition and rapid spread of amantadine drug resistance mutation S31N. This dataset served for determining the association between different amino acids in different proteins, which could potentially present epistatic interactions between those segments, and for determining the extent of co-evolution between the associated amino acid sites.

A new dataset of 236 sequences for matched HA, NA, MP was downloaded for swine H1N1, isolated from Eurasia. Sequences were isolated from January 1st 1960 to December 31st 2008. The date of downloading was 1st June 2013. According to the Neighbour-joining tree of MP with 1000 bootstraps these sequences fall into Eurasian and Classical lineages. The aim of this study is to look for interactions between mutations in pairs of 3 segments with machine learning (the methods of machine learning are described in chapters Methods and Oseltamivir resistance in

human seasonal influenza A/H1N1 viruses) to answer the question: Are there any amino acid sites in M2, M1, HA and NA that have an evolutionary interaction to the amantadine drug resistance/susceptibility in M2? Next, BGM in HyPhy (www.datamonkey.org) were used to elucidate which amino acid sites, determined to be associated, are co-evolving. More precisely, the purpose of this analysis was to find out which amino acid site is influencing another amino acid site to mutate. For this purpose, maximum likelihood trees were constructed in MEGA v4 (Tamura et al., 2007) for HA, NA and MP segments, separately. Next, a file with concatenated sequences of M2, M1, HA, NA was constructed and saved as a nexus file, which included also a maximum likelihood tree for either, HA, NA or MP set of sequences. For comparison, also a file with concatenated sequences but no maximum likelihood tree was constructed and analysed. The files of concatenated sequences with or without a maximum likelihood tree were then subjected to a BGM analysis in HyPhy separately for each pair of amino acids, shown by Random Forest analysis to be associated. The result was a posterior probability of one amino acid site influencing the other amino acid site in terms of evolution. These results of posterior probabilities for each amino acid pair were joined together in an excel file and imported into Cytoscape v3.0.1 (Su et al., 2014). The output from Cytoscape was a network showing evolutionary dependencies between amino acid sites.

4.9 Results

4.9.1 Dataset 1

This dataset was used to estimate the prevalence of amantadine drug resistance mutation S31N in influenza A viruses, isolated from two different hosts in different time periods. Resistant mutations were most prevalent on position 31; estimated prevalence for swine H1N1 ranged from 16.2% (1967 – 1999) to 74.6% (2000 – 2011), for swine H1N2 from 41.9% (2000 – 2011) to 85.7% (1967 – 1999) and for swine H3N2 from 6.9% (1967 – 1999) to 44.2% (2000 – 2011) (Table 4.1, Figure 4.6). While human seasonal H1N1 influenza A virus shows a low prevalence of

resistance at position 31 (3.9%), seasonal H3N2, which alternates with H1N1, that also predominates in the human population until 2009 shows high prevalence (47.2%) at that position (Table 4.1, Figure 4.6).

Table 4.1

Prevalence of S31N mutation in M2 gene for swine and human sequences studied.

	prevalence of S31N mutation in M2 (%)		
	1950 - 1966	1967 - 1999	2000 - 2011
swine			
H1N1	22.2	16.2	74.6
H1N2	/	85.7	41.9
H3N2	/	6.9	44.2
human			
H1N1 seasonal	0	0.6	3.9
H1N1 pandemic	/	/	100
H3N2 seasonal	/	1.6	47.2

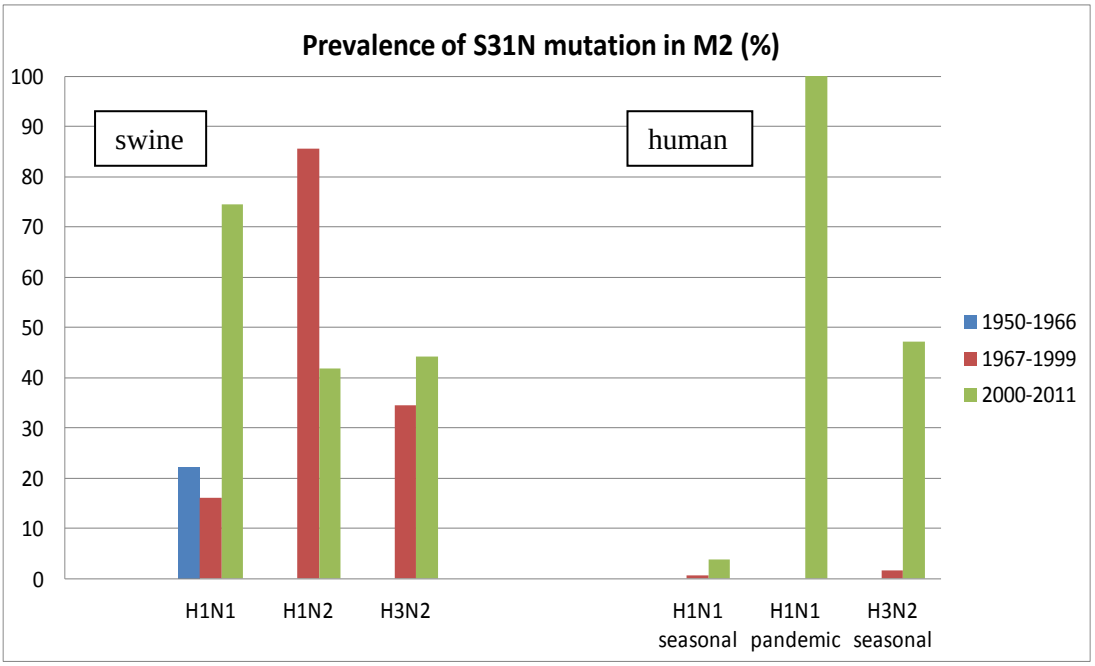

Figure 4.6

The prevalence of S31N mutation in M2 divided into host, subtypes, and time period of isolation: 1950-1966 (before the availability of amantadine), 1967-1999 (after the availability of amantadine and before the availability of oseltamivir), 2000-2011 (after the availability of oseltamivir).

The human pandemic H1N1 virus that originated in swine and has the potential to transmit back to swine and for reassorting with other influenza A viruses was 100% resistant at position 31 (Table 4.1, Figure 4.6). There was no continential bias regarding the origin of sequences as for all groups of sequences, the majority of them were isolated from North America, Asia and Europe. Human seasonal H1N1 and H3N2 also had a large number of sequences from Australia.

4.9.2 Dataset 2

This dataset served for determining the origin of amantadine drug resistance (which host) and for estimating the time of the amantadine drug resistance appearance.

4.9.2.1 Maximum likelihood analysis

PhyML was used to conduct a phylogenetic analysis of MP coding region for swine influenza A sequences (H1N1, H1N2, H3N2 subtypes) with avian influenza A sequences and human influenza A sequences of different subtypes added to see an adamantane resistance in this virus. From Figure 4.7A two major clades are seen: (i) viral strains of resistant Eurasian swine influenza A sequences with the resistant H5N1 avian viral lineage and some other avian and human sequences, and (ii) susceptible North American swine influenza A viruses with some resistant swine and susceptible avian viral sequences. Resistance was assigned by presence of resistant mutations. Figure 4.7B shows a susceptible avian strain (A/avian/goose/Leipzig Germany/1877/1979 H7N7 suscep) sharing a common ancestor with a susceptible swine strain (A/swine/France/WVL4/1985 H1N1 suscep). This susceptible swine strain shares the common ancestor with the first resistant swine strain (A/swine/Italy 670/1987 H1N1 resist).

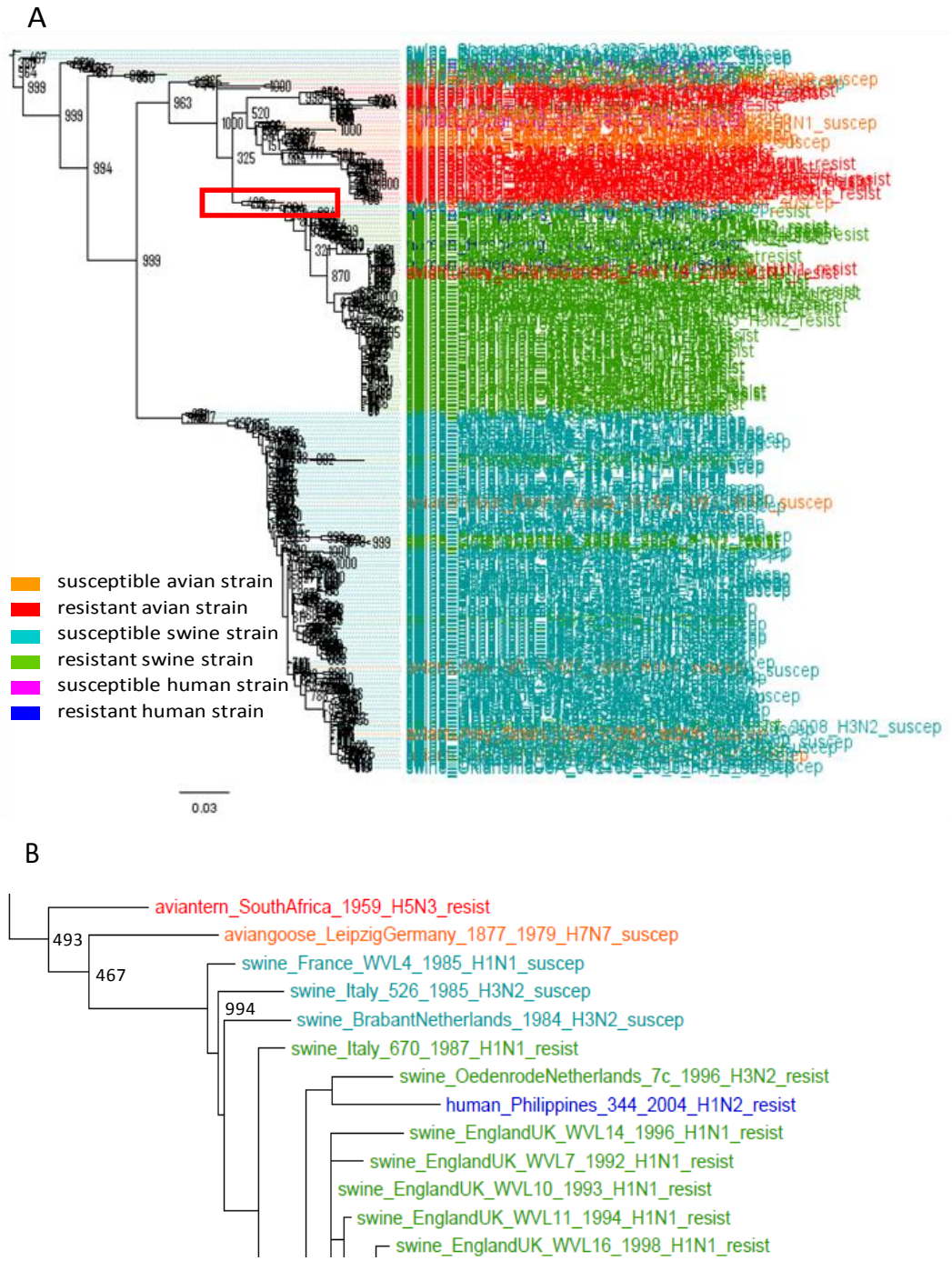

Figure 4.7

PhyML tree for avian, swine and human MP sequences from Dataset 2. (A) Avian, swine and human sequences of MP segment either resistant or susceptible to adamantanes, with bootstrap values. Earliest resistant swine sequences are captured within a red box. (B) The cut-out of a part of a tree from figure 4.7A (red box), where first swine resistant strains appeared, with bootstrap values.

4.9.2.2 BEAST

In order to estimate when the first resistant swine sequence emerged and to which sequences does it cluster closest to, Bayesian MCMC approach implemented in BEAST was used. This was done in order to answer questions: Is the phylogenetically closest sequence to first resistant swine sequence, avian or swine influenza A? Is this closest sequence resistant or susceptible?

First, BEAST analysis with discrete traits was done for the North American clade. Most of the sequences in this clade are susceptible to adamantanes with the most recent strain dating to 2010.7. The first resistant sequence is from 1996 (A/swine/Saitama (Japan)/1996/H1N2), which shares a common ancestor with the susceptible sequence A/swine/Ehime (Japan)/1/1980/H1N2. This MRCA dates back to 1979 (HPD = 30.3 – 32.0 => 1978.7 – 1980.4) (Figure 4.8). Clearly the resistance occurred in swine in Japan and was not transmitted from avian influenza A, as there are no resistant avian influenza A sequences in a North American clade in this dataset. Second cluster of resistant swine influenza A strains is from Canada, all 2004 and they also share a MRCA with a susceptible strain from the same country; in this case Canada and with a resistant strain from Japan. Also, there are three separate introductions of resistance with just one strain each time (an H1N2 strain from Hong Kong, an H1N2 strain from South Dakota and an H3N2 strain from Oklahoma. They seem to have all occurred in swine spontaneously.

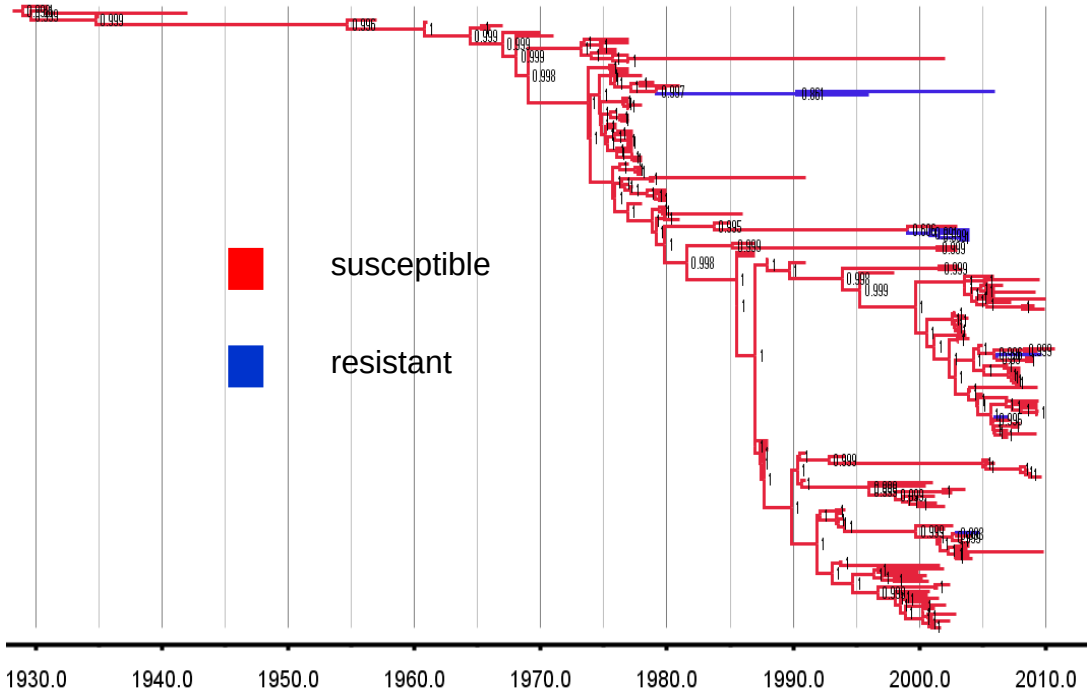

Figure 4.8

BEAST phylogenetic tree of North American swine clade with some avian sequences coloured according to susceptibility (red) or resistance (blue) to adamantanes, with discrete trait analysis and Markov jumps, with posterior support.

Next, BEAST analysis with discrete traits was conducted for the Eurasian clade. In order to investigate which was first: a susceptible strain developing into a resistant one or vice versa, an analysis with discrete trait resistance or susceptibility to adamantanes was performed. At the node, representing the most recent common ancestor of an H7N7 strain isolated from goose in 1979 (Leipzig, Germany) which is susceptible, and an H5N3 strain isolated from tern in 1959 (South Africa) which is resistant, was most likely susceptibility (p=0.8488) (Figure 4.9). The most recent common ancestor of an H1N1 swine strains isolated from Italy in 1987 which is resistant, of other resistant strains, and of a swine susceptible clade is most likely susceptible (p=0.9511) and then mutated into a resistant strain (Figure 4.9).

A susceptible avian goose strain from Germany, isolated in 1979 shares a common ancestor with three susceptible swine strains in Eurasian swine clade from France, the Netherlands and Italy. They share a common ancestor with a completely resistant

clade of swine sequences. The most recent sequence in this dataset dates back to 2010.685 and the TMRCA of resistance appearance in swine sequences is 1979.915 (HPD = 17.196, 30.64 – 25.1792, 30.91 => 1979.78 – 1980.05) (Figure 4.10). The first resistant swine strain from this dataset was isolated in 1987 in Italy with TMRCA to the rest of the Eurasian clade in 1982.9067 (HPD = 27.73 – 27.82 => 1982.87 – 1982.96) (all of the strains are resistant). From the tree (Figure 4.9) it can be seen that resistance appeared multiple times in avian species but only once in swine.

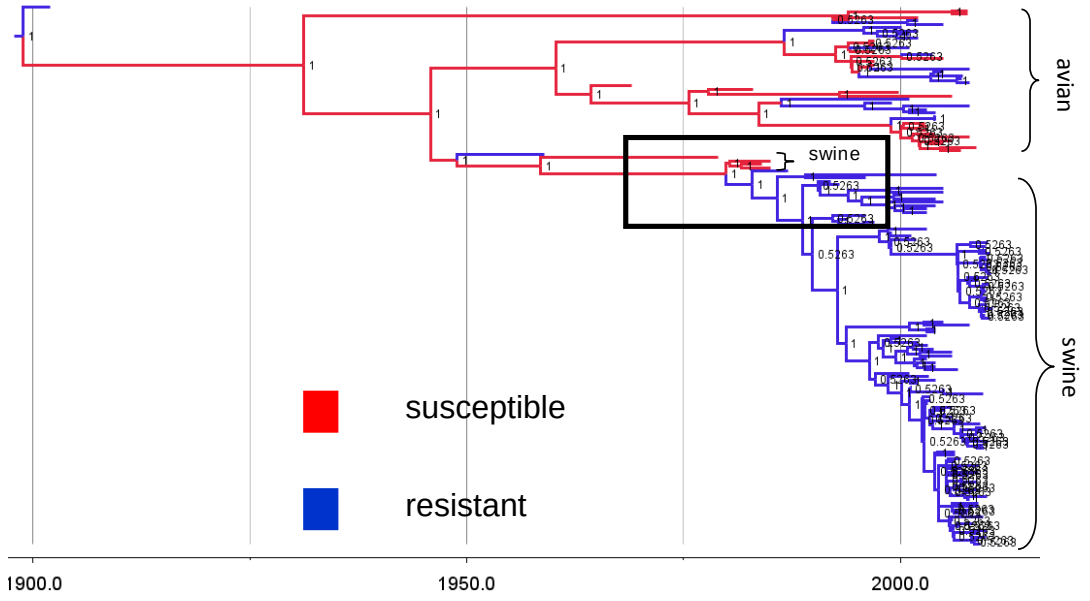

Figure 4.9

BEAST phylogenetic tree of avian and Eurasian swine clades and some human sequences coloured according to susceptibility (red) or resistance (blue) to adamantanes, with discrete trait analysis and Markov jumps, with posterior support. Earliest resistant swine sequences are captured within a black box.

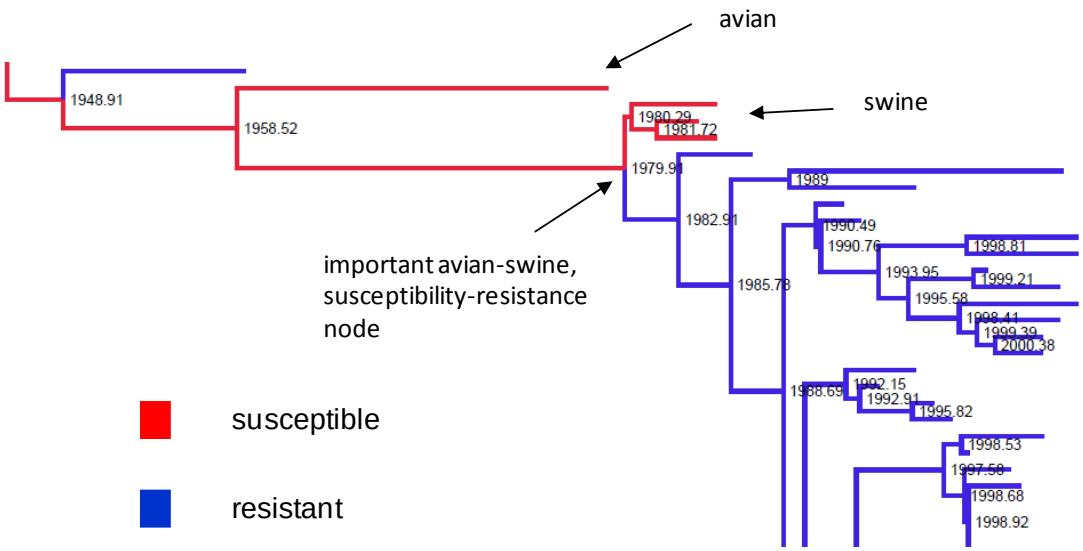

Figure 4.10

A cut-out from the black box in figure 4.9 with time (in years in decimal) of appearance of avian-swine node (the establishment of avian influenza A virus in a swine host) and susceptibility-resistance node (the establishment of amantadine resistant swine influenza A virus).

4.9.3 Dataset 3

This dataset served to determine the dN/dS ratio on each of the five drug resistance conferring mutation sites in M2 for four subgroups of sequences: Eurasian swine clade, North American swine clade, domestic avian and wild avian influenza M2 sequences. First, the maximum likelihood trees were constructed for avian (Figure 4.11) and swine (Figure 4.12) MP sequences not separated into clades (swine) and avian host species. The reason for including avian sequences in this analysis is in the origin of the Eurasian avian-like swine clade, which is avian. The aim of this analysis was to compare dN/dS ratios of these four groups to each other and to published results.

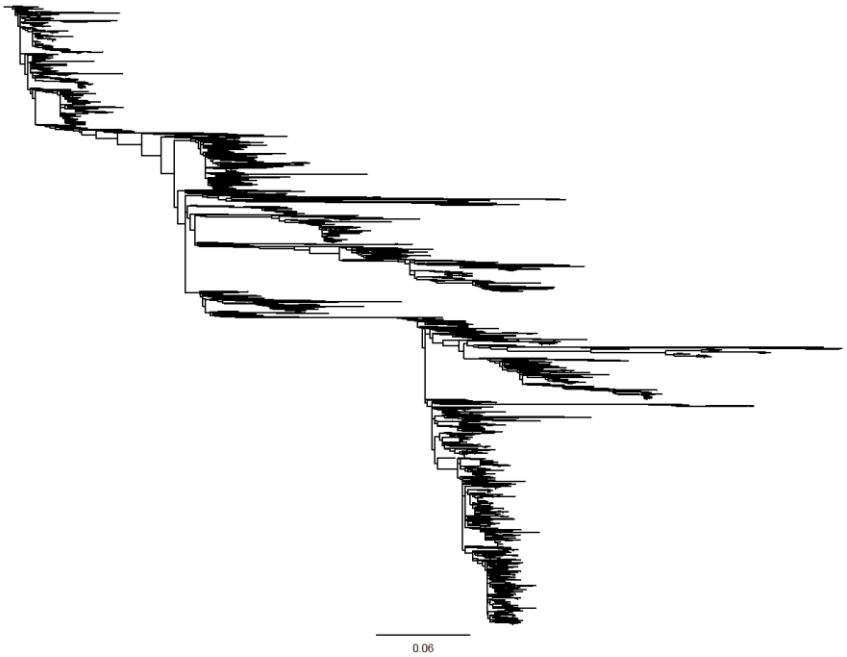

Figure 4.11

RAxML tree of avian MP sequences for both, strains isolated from domestic or wild birds.

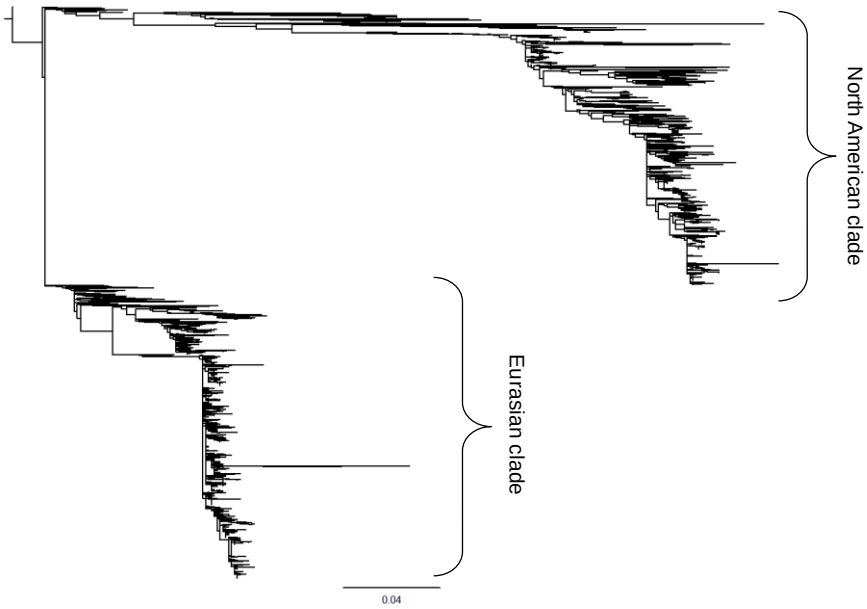

Figure 4.12

RAxML tree of swine MP sequences, which form two distinct clusters: North American and Eurasian.

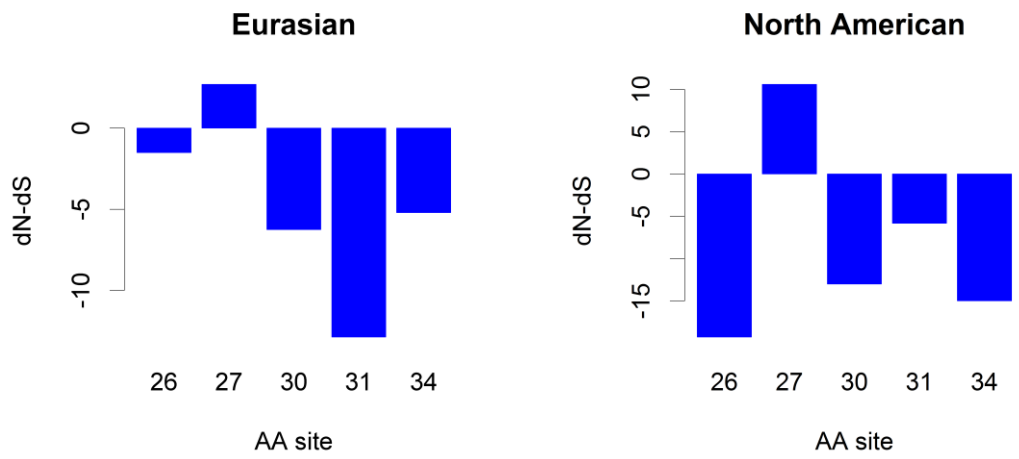

Figure 4.13

dN-dS values for amino acid sites in swine M2 where resistance to adamantanes can occur in Eurasian clade and North American clade separately. dN-dS values larger than 0 mean positive selection, dN-dS values smaller than 0 mean negative selection and dN-dS value equal to 0 mean neutral selection.

Table 4.2

dN-dS values for swine Eurasian and North American clades and avian strains, separated into two groups, isolated from domestic and wild birds.

AA site in M2	swine		avian	
	Eurasian	North American	domestic	wild
26	-1.51	-19.27	-17.01	-31.62
27	2.69	10.59	17.87	21.33
30	-6.26	-13.00	-6.13	-4.50
31	-12.87	-5.84	-3.16	-25.07
34	-5.21	-15.00	-14.90	-20.00
overall dN/dS	0.581	0.532	0.546	0.471

AA=amino acid

All five codons in human versus swine and avian influenza A viruses where resistance conferring mutations can occur exhibit differences in the patterns of substitution rates suggesting different selection modes. In swine and avian influenza A viruses only the codon 27 in M2 is under positive selection and all other under strong negative (purifying) selection (Table 4.2, Figures 4.13 and 4.14). In human influenza A viruses (according to the study by Furuse et al., 2009; Figure 4.4) codons

26, 27, 30 and 31 are under positive selection forces and at codon 34 there is no selection pressure. Codon 27 in human influenza A viruses showed greater positive selection than other codons, however it was still weaker than in North American swine clade and both avian groups, domestic and wild, on that codon. Overall dN/dS is comparable for both swine clades and for domestic avian influenza. Influenza A viruses isolated from wild birds showed lower overall dN/dS (Table 4.2), which is most likely the consequence of the influenza A viruses being more adapted to wild birds, which are the viruses' natural reservoir.

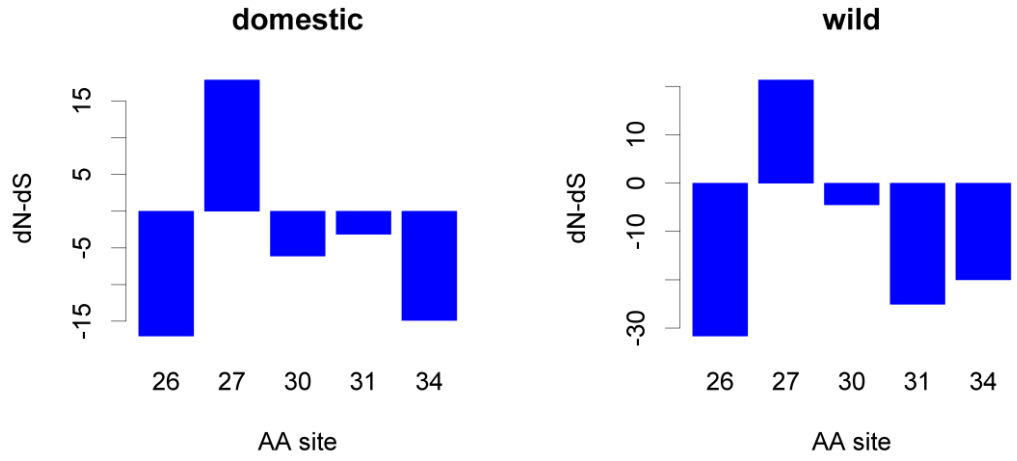

Figure 4.14

dN-dS values for amino acid sites in avian M2 where resistance to adamantanes can occur in influenza A viruses, isolated from domestic and wild birds, separately. dN-dS values larger than 0 mean positive selection, dN-dS values smaller than 0 mean negative selection and dN-dS value equal to 0 mean neutral selection.

4.9.4 Dataset 4

This dataset was used to determine if there are any evolutionary associations between amino acid substitutions at different sites in M2, M1, HA and NA. Further, the analysis was conducted in order to determine if these associated sites are also epistatically linked.

4.9.4.1 Phylogenetic analysis

The Neighbor-Joining (NJ) tree of the MP segment shows clear separation of influenza strains into two clades: Eurasian avian-like swine clade and Classical swine clade. Resistance to amantadine is restricted to the Eurasian avian-like swine clade (Figure 4.15).

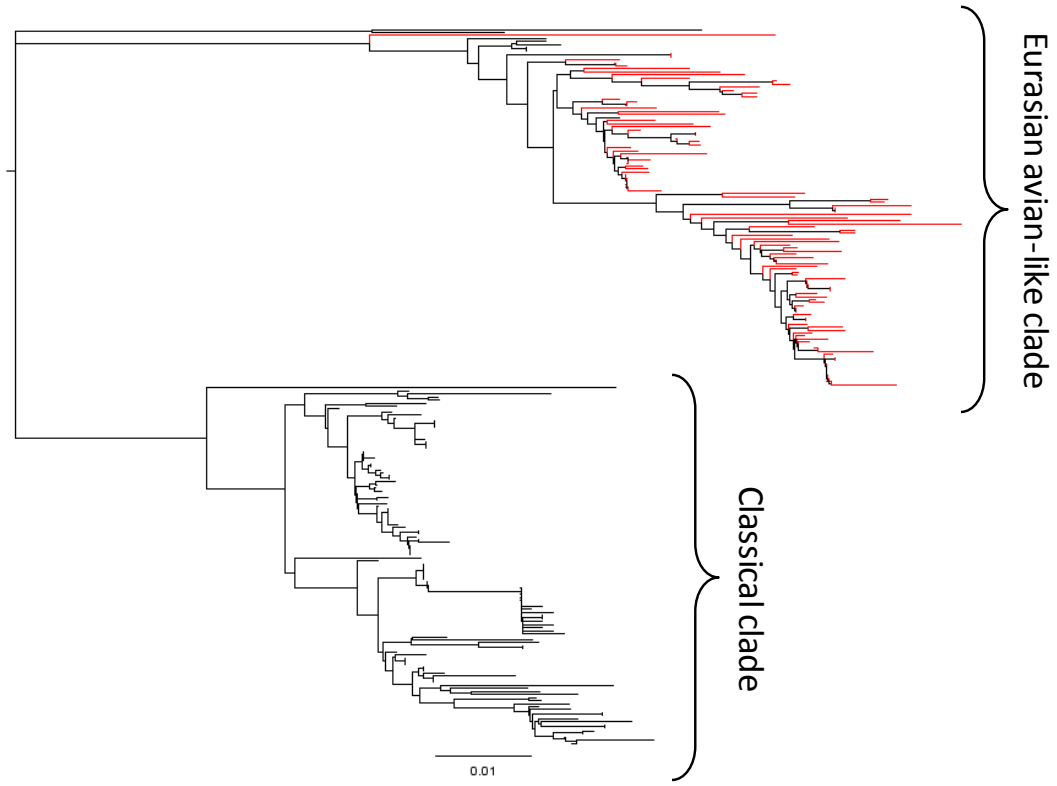

Figure 4.15

NJ tree of MP segment of influenza A swine sequences, isolated from Eurasia, coloured according to resistance to amantadine (red = resistance, black = susceptibility). Sequences fall into two distinct clades: Eurasian avian-like and Classical clades.

4.9.4.2 Random Forest analysis

The results of machine learning – Random Forest analysis of M2 amino acids associated with susceptibility/resistance to amantadine (encoded on site 31 in M2) showed very high sensitivity, specificity and area under the ROC curve (Figure 4.16) – AUC values. The sensitivity was 0.990, the specificity was 0.992 and AUC was

0.999. There were 14 sites in M2 identified as significantly associated with amantadine resistance or susceptibility, which is encoded on site 31 in M2 (Figure 4.17). The amino acid site in M2 with the highest mean increase in mean square error was site 77, which is one of the eight sites (70-77) which are responsible for efficient interaction with M1 and viral filament formation (morphology of the virus) (McCown and Pekosz, 2006).

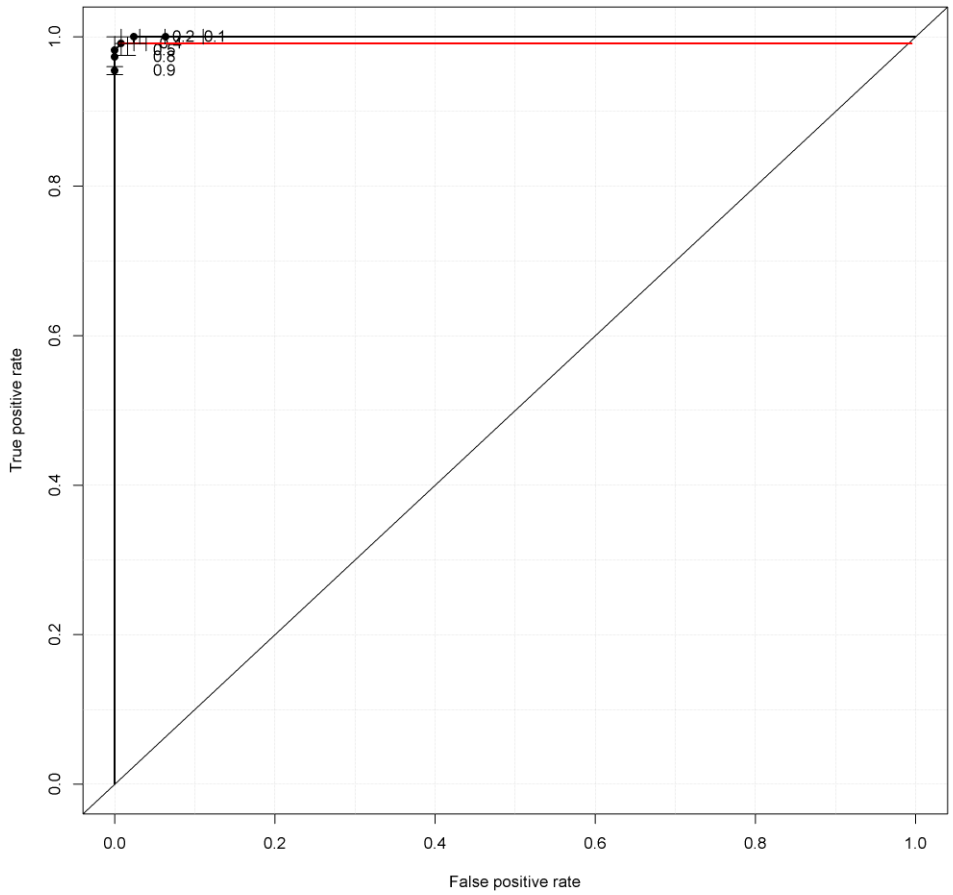

Figure 4.16

ROC curve for M2-M2 pair of segments. The X axis represents the false positive rate which equals 1-specificity. The Y axis represents the true positive rate which equals sensitivity. The ROC curve defines the area under the ROC curve-AUC- which can take value between 0 and 1. The larger the AUC value the better the model. The AUC for this model is 0.999.

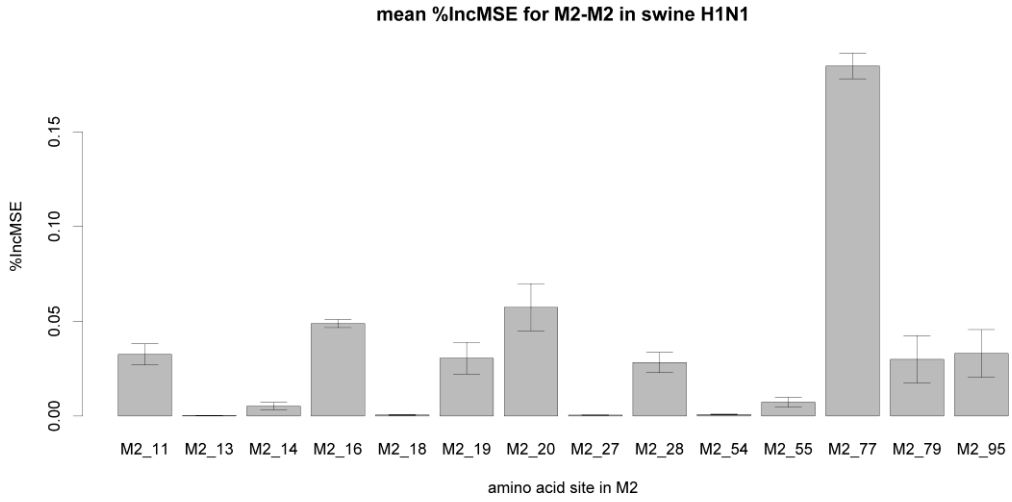

Figure 4.17

Mean increase in mean square error (%) for M2-M2 pair. The larger this value the more significant the association between the amino acid sites evaluated. There are 14 amino acid sites in M2, determined as significantly associated with amantadine resistance/susceptibility encoded on site 31 in M2. The amino acid residue in M2 associated most with the residue 31 in M2 is at position 77.

Next, the analysis was done in order to check if there are any amino acid sites in M1, associated with susceptibility or resistance to amantadine (encoded in M2). The results of Random Forest analysis showed very high sensitivity, specificity and AUC. The sensitivity was 0.990, specificity was 0.992 and AUC was 0.991 (data not shown). There were 11 sites in M1 determined as associated with susceptibility/resistance to amantadine in M2 (Figure 4.18). The highest mean increase in mean square error in M1 was displayed by site 101, which is one of 5 residues (101-105) that form the nuclear localization signal (Burleigh et al., 2005). Also, these and the surrounding residues in M1 may be important for RNA and RNP binding (Burleigh et al., 2005).

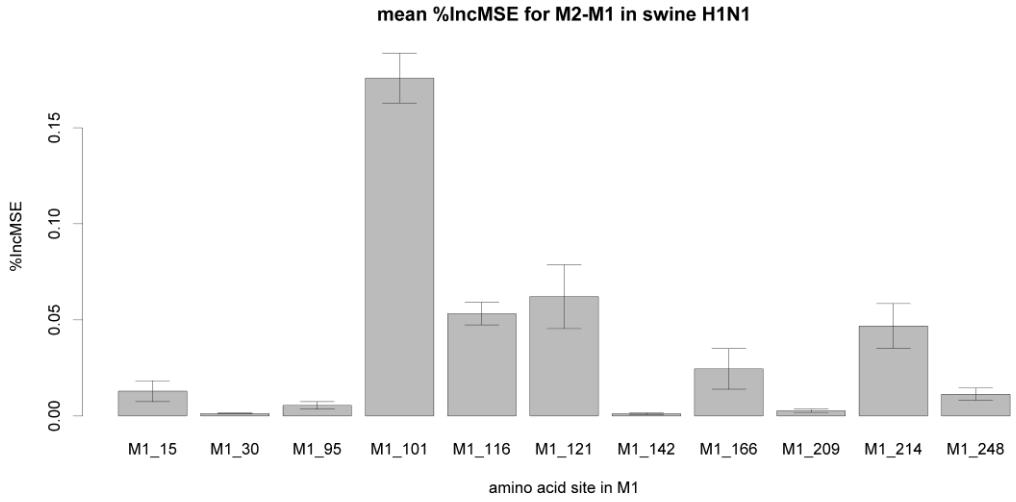

Figure 4.18

Mean increase in mean square error (%) for M2-M1 pair. The larger this value the more significant the association between the amino acid sites evaluated. There are 11 amino acid sites in M1, determined as significantly associated with amantadine resistance/susceptibility encoded on site 31 in M2. The amino acid residue in M1 associated most with the residue 31 in M2 is at position 101.

One site in M2 (77) and one in M1 (101) were found to be highly associated with amantadine drug resistance/susceptibility. From the literature it can be seen that amino acid 77 in M2 is one of the residues that are responsible for efficient interaction with M1 and viral filament formation (McCown and Pekosz, 2006). Site 101 in M1 is one of the residues where nuclear localization signal is contained and is also important for RNA and RNP binding (Burleigh et al., 2005).

So, the conclusion is that there are epistatic interactions between these sites and that site 77 in M2 is binding to site M1 and also influences viral morphology, which is filamentous. It has been shown that viruses with an Eurasian avian-like MP segment form 60% of filamentous virions and the rest are spherical, they are highly transmissible. On the other hand, viruses with a Classical MP segment are predominantly spherical and have lower transmission capability (Lakdawala et al., 2011).

In short, the conclusion is that the mechanism for the wide spread of S31N mutation

in M2 protein lies in its epistatic linkage to site 77 in the same protein which is one of the sites that bind to M1 protein and are responsible for the filamentous virion morphology that is highly transmissible.

It has also been reported that M2 may bind to HA in an M1-independent manner (Rossman and Lamb, 2011). Random Forest analysis was conducted in order to detect associations between amino acid sites in M2-HA pair.

There is one site in HA (276 in H1 numbering) that is highly associated with the susceptibility or resistance to amantadine, encoded in M2 (Figure 4.19). This site is a part of the sequence which encodes the esterase of HA1, located in the stem domain of HA (Sriwilaijaroen and Suzuki, 2012). It is also quite near two receptor-binding sites (Figure 4.20).

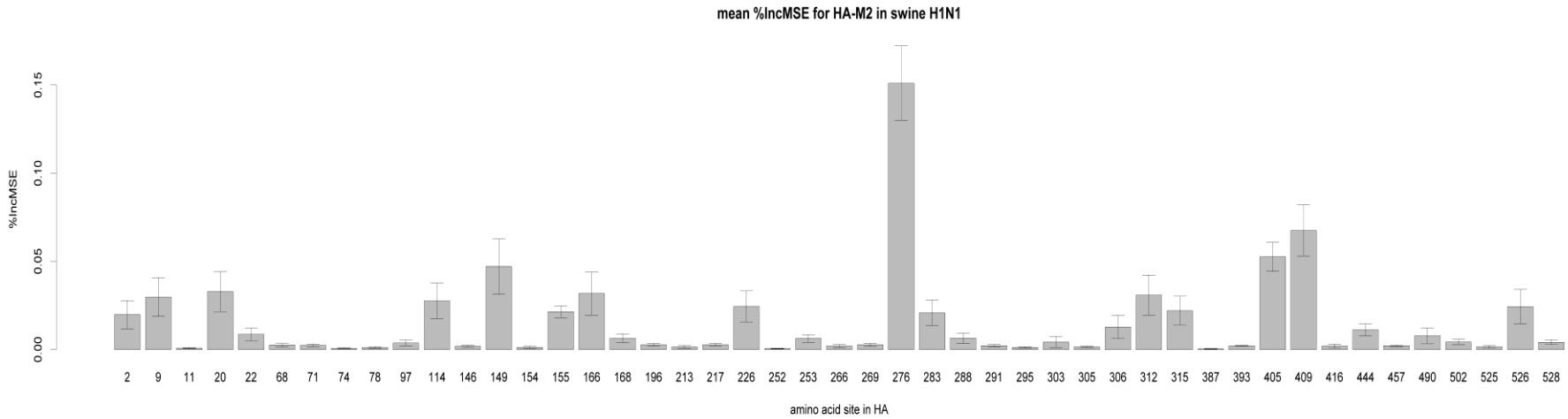

Figure 4.19

Mean increase in mean square error (%) for HA-M2 gene pair. The larger this value the more significant the association between the amino acid sites evaluated. There are 47 amino acid sites in HA, determined as significantly associated with M2 amantadine resistance/susceptibility. The amino acid residue in HA associated most with the residue 31 in M2 is at position 276.

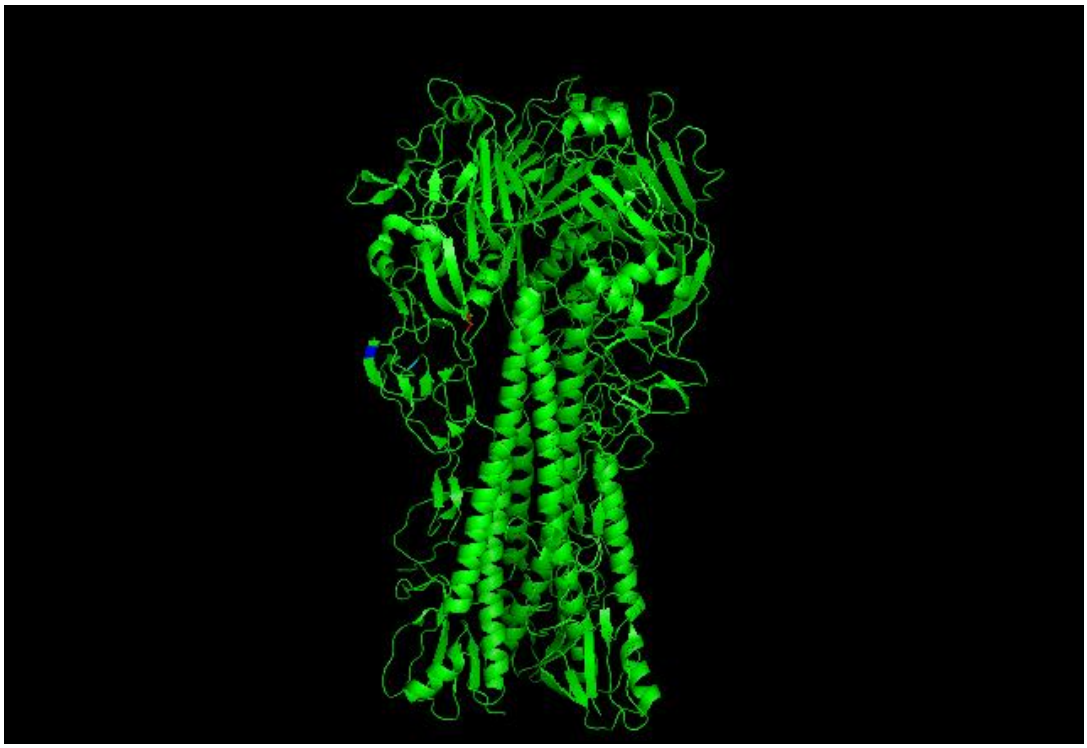

Figure 4.20

The tetramer of HA from 1930 swine influenza A. Coloured are amino acid sites 276 (red), 290 (blue) and 300 (cyan). Amino acid site 276 is associated with the amino acid site 31 in M2 as shown by Random Forest analysis. Sites 290 and 300 are receptor-binding sites and are located in the vicinity of site 276. PDB ID: 1RUY. Figure made by the author in PyMol v1.5.0.4.

Random Forest analysis was also performed to see if there are any sites in NA, associated with resistance or susceptibility to amantadine. Twentysix amino acid sites in NA displayed association with amantadine resistance or susceptibility encoded in M2. One site displayed high association; that was the site 395 (N1 numbering) (Figure 4.21), which is located on the surface and near the 380 loop (residues 380 – 392) that coincides with the putative calcium binding site (Figure 4.22). It is also near the glycosylation site (Xu et al., 2008).

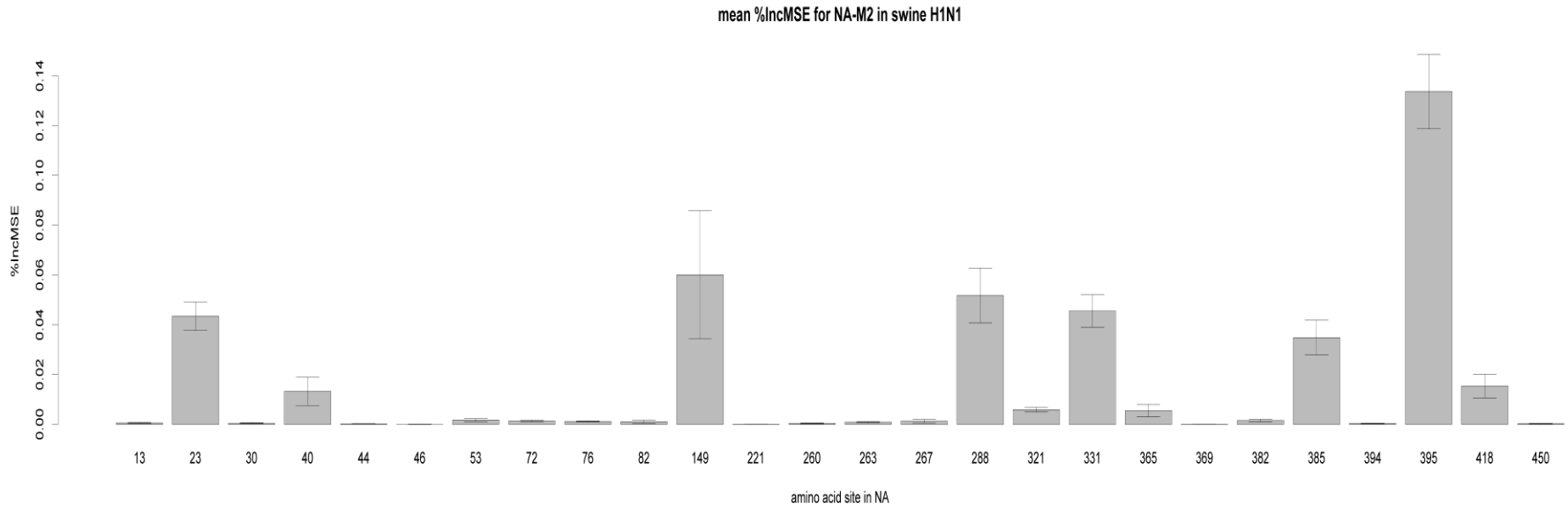

Figure 4.21

Mean increase in mean square error (%) for NA-M2 gene pair. The larger this value the more significant the association between the amino acid sites evaluated.There are 26 amino acid sites in NA, determined as significantly associated with M2 amantadine resistance/susceptibility. The amino acid residue in NA associated most with the residue 31 in M2 is at position 395.

It can be concluded that there are epistatic interactions between segments M2 and NA because one residue on the surface domain of NA changes according to amantadine drug susceptibility/resistance pattern.

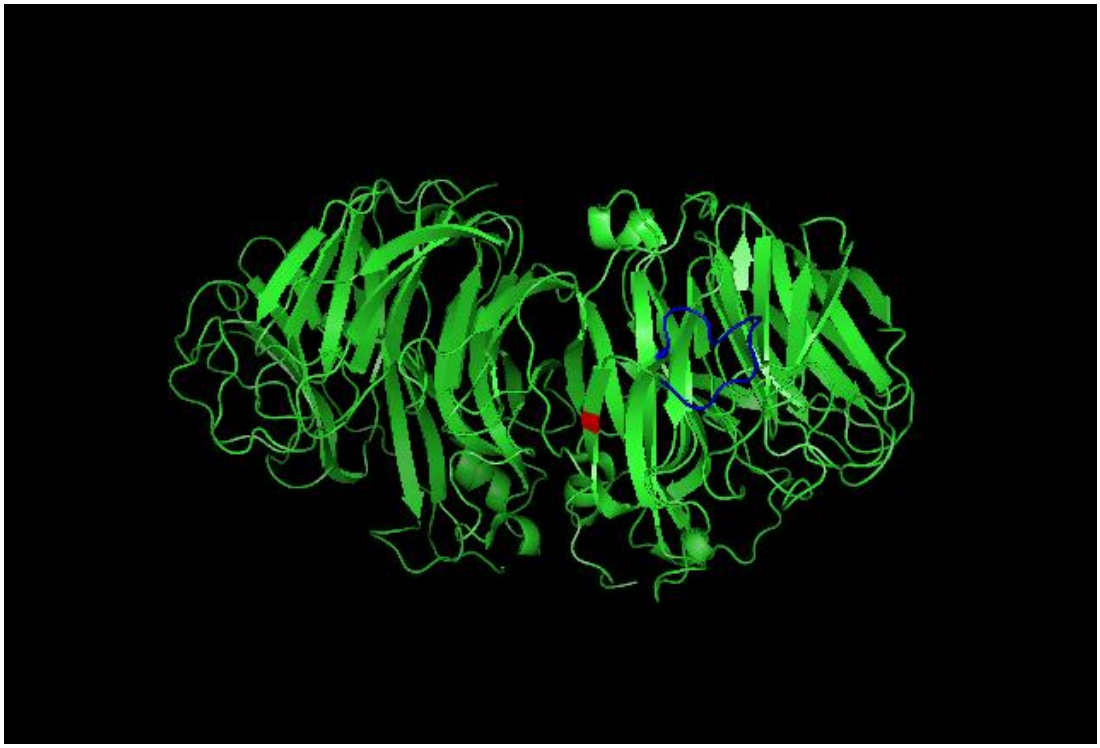

Figure 4.22

A part of a tetramer of NA polypeptides, isolated from human 1918 influenza A virus. Coloured is the site 395 (red) and 380 loop (blue) that contains calcium-binding site and is located in the vicinity of the site 395. PDB ID: 3BEQ. Figure made by the author in PyMol v1.5.0.4.

4.9.4.3 BGM analysis

BGM analysis was conducted in order to distinguish functional association between amino acid sites from phylogenetic linkage. The dataset was the same as for the above Random Forest analysis. Initially, three sets of files were submitted to BGM analysis: the first with concatenated sequences M2, M1, HA, NA with a maximum likelihood tree based on MP sequences. The second file contained the concatenated sequences with an HA tree and third file contained concatenated sequences with an

NA tree. Also, the file with concatenated sequences and no maximum likelihood tree was analysed for comparison.

The BGM analysis of a file with a tree based on HA sequences showed that the amino acid site 276 in HA is changing under the influence of amino acid sites 395 in NA and 31 in M2 (Figure 4.23) with the posterior probability values 0.88 and 0.34, respectively. HA-276 is in turn influencing M2-31 (0.34), M2-77 (0.92) and M1-101 (0.40).

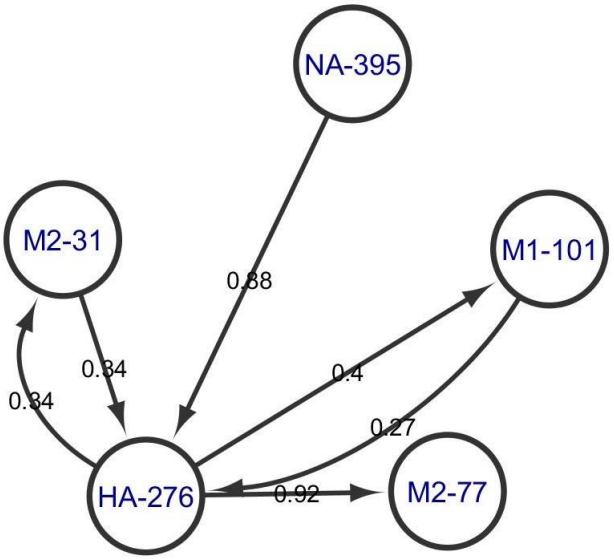

Figure 4.23

The result of a BGM analysis for swine H1N1 with a maximum likelihood tree based on HA sequences. Nodes represent amino acid sites from four different proteins and edges (arrows) represent dependencies between nodes. The degree of dependency is expressed by a probability value. Figure made in Cytoscape v3.0.1.

The file with an NA tree gave different results (Figure 4.24). M2-31 and HA-276 are both influencing NA-395 with high probability (0.80 and 0.78, respectively). M2-77 and M1-101 are changing according to NA-395 which corresponds to results based on an HA tree (Figure 4.23). NA-395 is influencing M1-101 and M2-77 to change; these two interactions are absent when considering results based on an HA tree (Figure 4.23).

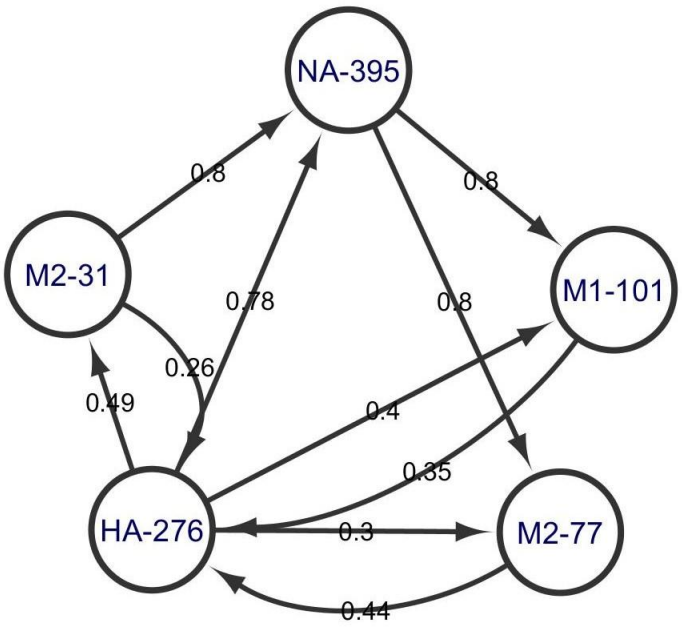

Figure 4.24

The result of a BGM analysis for swine H1N1 with a maximum likelihood tree based on NA sequences. Nodes represent amino acid sites from four different proteins and edges (arrows) represent dependencies between nodes. The degree of dependency is expressed by a probability value. Figure made in Cytoscape v3.0.1.

The results with an MP tree included in the nexus file of concatenated sequences show that the amino acid site 395 in NA influenced the amino acid site, where resistance to amantadine can occur (31 in M2) to change (Figure 4.25). The posterior probability for this event is very high (0.96). The amino acid site 31 in M2 further influenced the HA-276, which influenced M2-77, causing these sites to mutate. M2-77 then further influenced the NA-395 amino acid site.

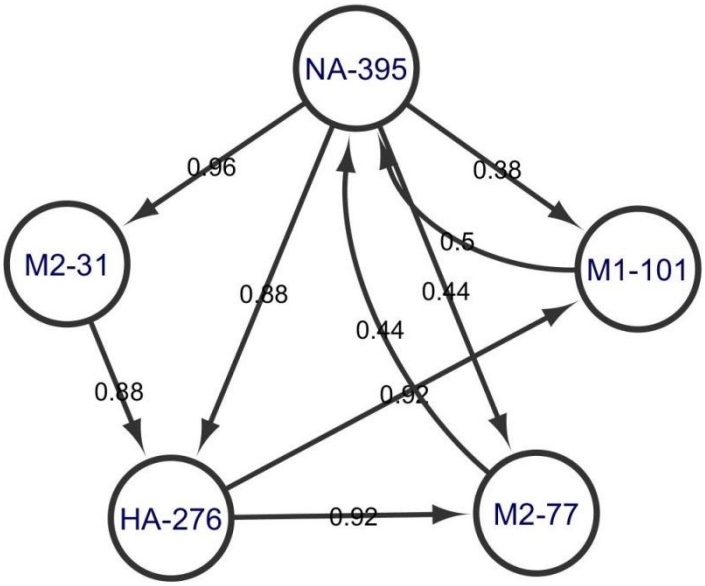

Figure 4.25

The result of a BGM analysis for swine H1N1 with a maximum likelihood tree based on MP sequences. Nodes represent amino acid sites from four different proteins and edges (arrows) represent dependencies between nodes. The degree of dependency is expressed by a probability value. Figure made in Cytoscape v3.0.1.

However, when submitting the file of concatenated sequences with an included maximum likelihood tree based on MP sequences for a BGM analysis, the software did not display the option for using the tree provided in the file. This option was displayed and used for the files with an HA and NA based tree. Therefore, the tree used for BGM analysis was most likely the neighbour-joining tree of concatenated sequences which is constructed and used during BGM analysis unless specified otherwise. To test this, a file with concatenated sequences but with no maximum likelihood tree was subjected to BGM analysis. The result (Figure 4.26) was identical to the results obtained using a file with an MP tree (apart from some of the values for posterior probability) which confirms that the BGM results for a file with an MP tree were constructed on a tree for concatenated sequences and are therefore not optimal for the interpretation of co-evolving amino acid sites from four different influenza virus proteins.

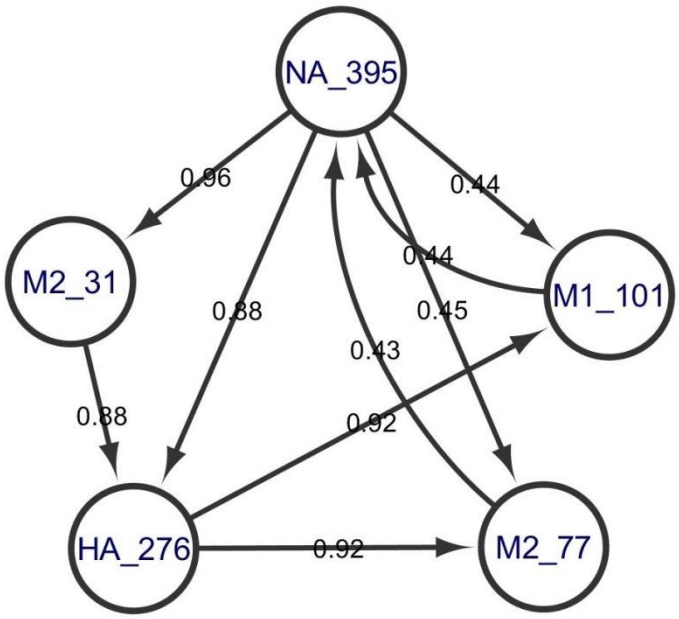

Figure 4.26

The result of a BGM analysis for swine H1N1 with no maximum likelihood tree included. Nodes represent amino acid sites from four different proteins and edges (arrows) represent dependencies between nodes. The degree of dependency is expressed by a probability value. Figure made in Cytoscape v3.0.1.

The co-evolving amino acid sites identified in this study are in the cytoplasmic domains of M1 and M2 (M1-101 and M2-77) and external parts of the virion (M2-31, HA-276 and NA-395). The importance of these physical and functional linkages is probably in maintaining high transmissibility of the virus due to its filamentous morphology (influenced by M2-77) and evading host's immune response (HA-276 and NA-395).

It can be concluded that the substitution from susceptible to resistant amino acid on site 31 in M2 is due to a hitch-hiking event, where changes in the frequency of an allele occur because of linkage with a positively or negatively selected allele at another locus (Futuyma and Douglas, 2013). These mutations are also due to epistasis between four influenza virus proteins studied here.

4.10 Discussion

In this chapter, four datasets were constructed in order to address different scientific questions. Also, different methods were used to answer those questions. Evolutionary analysis of full length swine influenza MP sequences available from the NCBI Influenza Virus Resource was one of the approaches used. Phylogenies were obtained which showed evolutionary relationships between swine MP sequences isolated from different time periods, and analysis with BEAST allowed visualisation of trees on an explicit time scale. Second approach used was Random Forest analysis, which is one of the machine learning methods. Its advantage is in high accuracy of predicted result values. This method was used to detect any associations between amino acid sites in different segments of influenza genome that might imply the cause of certain events, such as drug resistance mutation spread in the absence of a drug use.

Overall, a large number of swine influenza virus MP sequences were available but were used in only one of four datasets constructed. The reason for this lies in the limitation of one of the phylogenetics methods used (BEAST), which can process a rather small number of sequences to ensure valid results. The second reason is in the decreasing number of MP sequences available when matched to sequences of other segments from the same isolate. Although a large number of swine influenza virus MP sequences have been deposited to the online database, there are only a few sequences of early (before 1990) Eurasian avian-like clade available. The reason is deficient sampling of swine isolates at that time. So, it was hard to estimate the time of Eurasian avian-like clade emergence and amantadine drug resistance emergence, both of which occurred before 1990.

Phylogenetic analyses using maximum likelihood as well as Bayesian coalescent-based inference in BEAST, confirmed the time of the Eurasian avian-like swine clade emergence to be in the 1979. However, due to a small number of MP sequences available before the year 1990, the estimation of a time of an amantadine drug resistance emergence in this clade of swine influenza viruses was different –

estimated to be around 1979 also, however, according to Krumbholz et al. (2009) that happened in 1988.

According to Schmidtke et al. (2006) the cause of this amantadine drug resistance appearance and its wide spread lies in the reassortment of segments but according to Furuse et al. (2009) it is the point mutation. However, none of the studies conducted so far was able to explain the mechanism behind the rapid rise of the amantadine drug resistance after the year 2000 when this drug was no longer in use in humans. This reinforced the hypothesis, that the wide spread of S31N mutation responsible for amantadine drug resistance is due to genetic epistasis between different amino acids in different proteins as well as hitch-hiking (Futuyma and Douglas, 2013). As amino acid at site 77 in M2 is changing in order to efficiently interact with M1 protein and to form filamentous virions, which are highly transmissible, the epistatically linked amino acid at site 31 is also changing to a conformation that still allows the M2 ion channel to perform its function and it just happens to also confer amantadine drug resistance. The observation that also HA and NA change each at one amino acid site (276 and 395, respectively) indicates a net of epistatically linked sites and confirms the proteins' interactions in order to perform their functions. This was further supported by the BGM analysis, which revealed the four proteins studied to be changing by mutations in a dependent manner.

The results of a BGM analysis are reliable for files which included HA and NA trees but not for a file with an MP tree. However, the results based on an HA tree differ from results obtained for a file with an NA tree in the direction of amino acid sites' interactions. In chapter 3 (Reassortment of swine influenza A viruses) it is shown that swine HA and NA segments are frequently reassorting therefore the discrepancy in the results of a BGM analysis is most likely due to reassortment. Nonetheless, BGM analysis results show that there are co-evolving amino acid sites found in four different influenza A proteins (M2, M1, HA, NA) with some of the proteins coming in a physical contact therefore, there are clearly epistatic effects among M2, M1, HA and NA proteins. Namely, epistatic effects can take place when proteins are in a functional interaction which may be direct or indirect. In this case, the four proteins

studied are in a direct contact with one another which has been proven in laboratory based studies (McCown and Pekosz, 2006; Burleigh et al., 2005; Lakdawala et al., 2011; Rossman and Lamb, 2011; Bilsel et al., 1993; Mitnaul et al., 1996; Chen et al., 2005). However, these interactions are all between cytoplasmic domains of M2, M1, HA and NA. The results of the study described in this chapter show also co-evolving sites which are not in a physical interaction. Amino acid sites on HA and NA that are epistatically associated with the M2-31 site are on the surface of these two glycoproteins. HA-276 is near the receptor binding sites and NA-395 is near the calcium-binding site and one glycosylation site. Also, the M2-31 amino acid site is at the pore-facing side of the ion channel M2. So, HA-276, NA-395 and M2-31 are not found inside the virion. Hence, the epistatic interaction between HA-276 and M2-31 as well as the epistatic interaction between NA-395 and M2-31 probably enable the virus to evade host's immune response. Amino acid site M2-77 is one of the amino acids in this protein that enable the virus to take a filamentous form, which is highly transmissible. Therefore, the substitution from susceptible to resistant amino acid on site 31 in M2 is due to its linkage to amino acid sites in M2, M1, HA and NA proteins, that all enable the influenza virus to escape host's immune system as well as spread efficiently to different individuals of the host species.

The importance of this study is in the further elucidation of the mechanisms responsible for a rapid spread of an amantadine conferring mutation that took place in the time when amantadine usage was in decline or absent. The study offers a novel insight on this topic since it uses advanced methods of phylogenetics combined with bioinformatics as well as a machine learning method not used so far for influenza virus, the highly sensitive and specific method Random Forest, which can explicitly identify amino acid sites involved to explore the association between different proteins, coupled with BGM analysis showing the direction of association among amino acid sites.

Chapter 5

Oseltamivir resistance in human seasonal A/H1N1 influenza viruses

5 Oseltamivir resistance in human seasonal A/H1N1 influenza viruses

5.1 Oseltamivir drug resistance in human influenza A viruses

Antiviral agents constitute a means for the prevention and treatment of seasonal and pandemic influenza (Sheu et al., 2008; Maltezou, 2008; Hota and McGear, 2007). Currently, there are two classes of antiviral drugs for preventing and treating influenza infections: M2 ion channel blockers and neuraminidase inhibitors. Since the rapid spread of resistance to M2 ion channel blockers that began increasing in the 2004-2005 influenza season (5.8% of resistant viruses in Asia) and reached 15.5% globally in the 2005-2006 influenza season for A (H1N1) while for A (H3N2) the proportion of resistant viruses was 90.6% in the 2005-2006 season (Bright et al., 2005; Deyde et al., 2007), this class of drugs is no longer recommended for treatment. Oseltamivir-resistant seasonal influenza A/H1N1 was detected in Europe, first in France, Spain, Switzerland, and the United Kingdom in week 40 of 2007 and spread globally in the absence of drug selective pressure in the season of 2007-2008 (Hauge et al., 2009; Hurt et al., 2009; Meijer et al., 2009). At the end of this season in September 2008, the mean percentage of oseltamivir-resistant isolates was 24.3% in Europe, 12.3% in USA, 26% in Canada, 12% in Hong Kong and 3% in Japan (Casalegno et al., 2010; Lackenby et al., 2008; Dharan et al., 2009). By the 2008-2009 season, most seasonal influenza A/H1N1 viruses were resistant to oseltamivir (WHO, 2009).

5.2 Structure and mechanism of influenza A virus neuraminidase

The influenza neuraminidase is found on the viral surface and composed of four identical subunits. Each subunit contains about 470 amino acids and is further arranged in four domains: (i) a N-terminal cytoplasmic sequence which is followed by (ii) a membrane-penetrating hydrophobic transmembrane domain and a (iii) thin

stalk of variable length, ending in (iv) a globular head domain with the enzyme active site (Air, 2012).

The symmetry of four subunits is stabilized by a carbohydrate side chain facing inwards (Colman, 1991) and by metal (calcium) ions. Two calcium binding sites have been identified, one of them near the binding pocket (Colman et al., 1983). It is believed that calcium ions positively modulate NA enzyme activity during the release of newly synthesized virions from the cell (Chong et al., 1991).

The cytoplasmic domain contains six conserved amino acids (MNPNQK), found in almost all influenza A subtypes. However, its function remains unknown. Viruses with mutations in the cytoplasmic tail show reduced budding and changed morphology (Barman et al., 2004; Jin et al., 1997).

The transmembrane domain is variable among influenza A subtypes. It is a helix containing amino acids 7-29 (Krogh et al., 2001; Moller et al., 2001) with two functions. The first is directing the NA through the endoplasmic reticulum during protein folding, and the second is anchoring the NA in the membrane.

The stalk is located between the transmembrane sequence and the globular head domain. Its structure has not been fully elucidated yet. However, it is known that it is of variable length and that all NA subtypes have Cys residues in the stalk and/or transmembrane domains, which may assist in tetramer formation. All NA stalks contain glycosylation sites, shorter stalks usually contain one site, full length stalks contain three or more. Most NAs have stalks of approximately 50 amino acids, but deletions of up to 18 amino acids have been found in N1 and N2 NA stalks (Blok and Air, 1982; Els et al., 1985). Several studies have associated deletions in the stalks of N1 NA of avian viruses with transmission from ducks to land-based poultry (Munier et al., 2010), but the mechanism remains unknown. Deletion of the stalk does not change the activity of NA for small substrates but activity with complex substrates is reduced (Els et al., 1985; Munier et al., 2010), which suggests that lower

accessibility of the substrate and therefore lower activity of NA is an advantage for viruses replicating in chickens compared to viruses of waterfowl (Air, 2012).

The globular head domain carries the active site located at the center of each subunit. The active site contains several conserved charged amino acid residues (Burmeister et al., 1992; Colman et al., 1983). The natural ligand for NA is the sialic acid and as NA inhibitors are analogs of sialic acid, the mechanism of their antiviral activity is through competition with the natural ligand and blocking the enzyme active site (Figure 5.1). NA inhibitors are effective against influenza A and B viruses and even different NA subtypes due to the conserved active site with only minor structural differences (Russell et al., 2006). Targeted viral function is the release of progeny virions from the infected cell and their spread to neighboring cells.

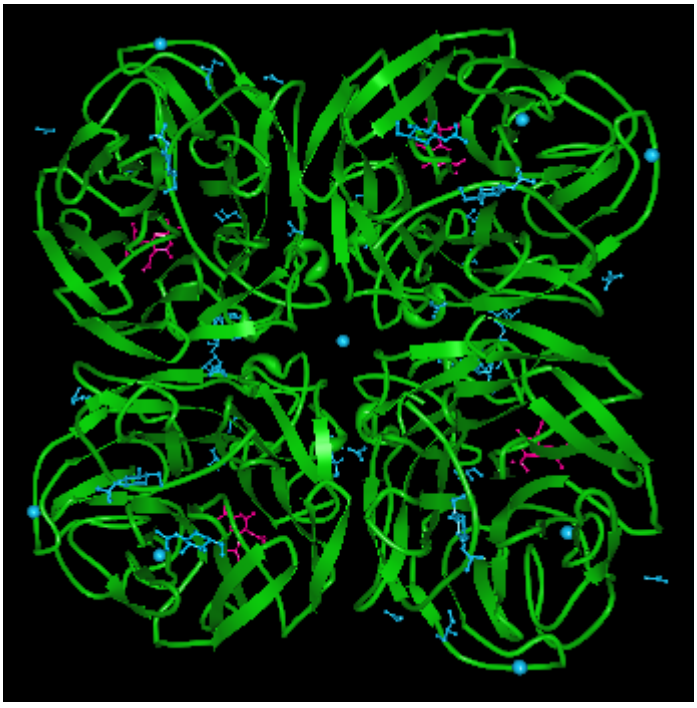

Figure 5.1

A neuraminidase tetramer, isolated from human pandemic 2009 influenza A virus, in complex with oseltamivir. Oseltamivir is coloured in magenta. Blue colour presents six different molecules or atoms: acetate ion, beta-D-mannose, calcium ion, glycerol, alpha-d-mannose, N-acetyl-D-glucosamine. Glycerol is a part of NA inhibitors and sialic acid, acetate is used for obtaining crystallized NA, and calcium ions are believed to positively modulate NA activity. PDB ID: 3TI6. Figure made by the author in RCSB PDB Protein Workshop 4.1.0.

5.3 Neuraminidase inhibitors: oseltamivir and zanamivir

Two neuraminidase inhibitors, oseltamivir (ethyl (3R,4R,5S)-5-amino-4-acetamido-3-(pentan-3-yloxy)-cyclohex-1-ene-1-carboxylate, Tamiflu^{TM}) and zanamivir ((2R,3R,4S) – 4 – guanidine – 3 - (prop – 1 – en – 2 - ylamino) – 2 - ((1R,2R) - 1,2,3 - trihydroxypropyl) -3,4-dihydro-2H-pyran-6-carboxylic acid, Relenza^{TM}) are used in the treatment and prophylaxis of influenza caused by influenza A virus and influenza B virus (Figure 5.2). Oseltamivir and zanamivir have been available since 1999. Oseltamivir is a prodrug, a (relatively) inactive chemical, which is converted into its active form (oseltamivir carboxylate) by metabolic process after it is taken into the body. Oseltamivir is taken orally, while zanamivir is inhaled (Burch et al., 2009).

Figure 5.2

Chemical structures of oseltamivir carboxylate and zanamivir (neuraminidase inhibitors) (Gubareva et al., 2004).

The primary mutations that confer resistance to oseltamivir are H275Y (N1 numbering) (Dharon et al., 2009; Hauge et al., 2009; Meijer et al., 2009) and N295S (N1 numbering) in the neuraminidase protein in that order of importance. H275Y is found near but not directly in the substrate-binding pocket (Russell et al., 2006).

5.4 Oseltamivir usage in humans

Oseltamivir has been available as an anti-influenza drug since 1999 and wide spread resistance to it appeared in the 2007-2008 influenza season (Hauge et al., 2009; Hurt et al., 2009; Meijer et al., 2009). The number of oseltamivir prescriptions per 1,000

inhabitants per country prior to the 2007-2008 influenza season does not always match the proportion of resistant strains isolated in that country. For example, Japan reported low levels of oseltamivir resistant strains but the stockpiling of this drug was high (Yasui et al., 2007). On the other hand, European countries experienced high proportions of resistant strains but reported moderate or low usage of this drug (Hauge et al., 2009; Kramarz et al., 2009). So the question of what led to the appearance and global spread of oseltamivir resistant strains in 2007-2008 influenza season remains.

5.5 Functional association of neuraminidase and haemagglutinin

The two glycoproteins, haemagglutinin (HA) and neuraminidase (NA), embedded into the viral membrane of influenza have complementary functions. Numerous studies have shown a functional balance between them is required for productive infections. While HA is a receptor-binding glycoprotein (Hirst, 1941), required for cell entry, NA is receptor-destroying (Burnet & Stone, 1947) and active on viral release from the cell. The balance between them can be destroyed by various factors such as drug inhibition of NA. Mutations in NA can occur that allow the virus to persist in the presence of a drug, although with a lower replication ability *in vivo* (Tai et al., 1998; McKimm-Breschkin et al., 1998; Gubareva et al., 1997). Because of the functional association between NA and HA, mutations in HA have been observed which consequently resulted in a decreased HA activity to balance the decrease in NA activity (Bantia et al., 1998; Blick et al., 1998; McKimm-Breschkin et al., 1996). This allows the virus to exit infected cells in the presence of a drug.

5.6 Epistasis

Epistasis describes non-additive interactions among genetic sites: the consequence of a mutation at one site may depend on the status of the genome at other sites. Epistatic mutations in viruses and bacteria that live under severe conditions, such as drug

treatment or immune pressure, allow pathogens to develop drug resistance or escape the immune system. These beneficial mutations are often pleiotropic: they have a beneficial effect but in addition to that they cause some side effects (usually negative) on other protein properties, such as stability (Bloom and Arnold, 2009; DePristo et al., 2005). These negative effects can usually be compensated by other mutations, making a certain combination of mutations more beneficial to the pathogen than single mutations alone (Sanjuan et al., 2005; Mateo and Mateu, 2007). This phenomenon is known as positive epistasis between mutations (de Visser and Elena, 2007). Epistasis can also be negative if a combination of mutations confer smaller benefit to the pathogen than what would be expected under additive effects of the individual mutations (de Visser and Elena, 2007).

5.7 Methods for exploring the association between (viral) genotype and phenotype

In the last few years, many different methods have been used to explore associations between genotype and phenotype using datasets of viral nucleotide sequences, predominantly in HIV. These methods include linear regression (Precious et al., 2000; Wang et al., 2004), decision trees (Beerenwinkel et al., 2002), random forests (Murray et al., 2008), support vector machines (Beerenwinkel et al., 2001), neural networks (Wang & Larder, 2003) and others (Sevin et al., 2000; Beerenwinkel et al., 2003). Many of these are based on machine learning approaches which can process large volumes of sequence data, incorporate multiple variants at individual amino acid sites and search for explicit models identifying the amino acid sites involved.

Beerenwinkel et al. (2002) used decision trees to predict drug resistance to 14 antiretroviral drugs from a genotype. The predictive error was in the range of 9.6% to 15.5% and 25.4% to 32.0%, depending on the drug. The sensitivity ranged from 58.2% to 92.5% and specificity between 62.5% and 97.2% according to drug. This approach was developed further by Murray et al. (2008) who introduced the case of

HIV-1 drug resistance determination for tenofovir random forest, yielding a specificity of 77%, which was higher than for decision trees (72%).

5.8 Studies on mutations on different sites that are epistatically linked and confer resistance to oseltamivir

The study by Kryazhimskiy et al. (2011) described a novel statistical method to detect positive epistasis between pairs of sites in a protein, based on the observed temporal patterns of sequence evolution. The method is based on the idea that a substitution at one site should rapidly follow a substitution at another site if the sites are positively epistatic. Kryazhimskiy et al. (2011) applied this method for detecting such pairs in either HA or NA of H3N2 and H1N1. They found substantial amounts of epistasis and determined the identities of putatively epistatic pairs of sites.

In the study by Baranovich et al. (2010) the authors collected samples from influenza patients from Japan in two subsequent influenza seasons: 2007-2008 and 2008-2009. They looked for mutations in NA at site 274 (N2 numbering) and in M2 at site 31. They found that 0.4% influenza A(H1N1) isolates from the 2007–2008 season and 100% of influenza A(H1N1) isolates from the 2008–2009 season possessed the H274Y substitution in the NA, which confers oseltamivir-resistance. Amantadine-resistance S31N substitution in the M2 gene was detected in 62.7% influenza A(H1N1) isolates from the 2007–2008 season but in none of the influenza A(H1N1) isolates from the 2008–2009 season. All of the isolates from 2008-2009 season plus one strain from the 2007-2008 season were amantadine sensitive but oseltamivir resistant and possessed the H274Y and D357G substitutions in the NA. Viruses from the 2008-2009 season also possessed an additional A193T (H3 numbering) substitution in the receptor-binding domain of HA. In addition, isolates from 2008-2009 season, had polymorphisms at positions 189 (G189A and G189V), 145 (S145N), and 196 (H196R) in the HA (all in H3 numbering).

5.9 The aim of study

The surface of HA and NA, where antibodies bind (epitopic sites) tends to evolve more quickly than other parts of the protein and especially more quickly than other viral proteins, in order to evade immunity (Suzuki, 2006; Bush et al., 1999; Wolf et al., 2006). Studies (for example the one by Kryazhimskiy et al. (2011)) have shown epistatically linked sites in HA and NA separately. Because of a functional balance between those two glycoproteins, we investigated the functional interaction between NA and HA genotype using machine learning techniques to look for correlation of amino acid variants in HA with oseltamivir resistance encoded in NA.

5.10 Methods and data

5.10.1 The dataset

Full-length sequences of NA and matching HA were downloaded from the GISAID database for human seasonal influenza A/H1N1 virus. Sequences were isolated from October 1st 2006 to December 31st 2008 and submitted by CDC. After excluding sequences from Africa, Europe, Oceania due to low numbers, 542 sequences remained in the dataset which were isolated from North, South and Central America and Asia (including Russia).

5.10.2 Bivariate analysis

The first 17 amino acids were excluded from the analysis as they represent the signalling sequence. The remaining 549 amino acids of HA segment were analysed in R to determine frequency count of amino acids on each site. Each amino acid site was tested for association with the susceptibility/resistance in NA using Fisher's exact test and recorded the p-values. Amino acid sites of a p-value of less than 10^{-6} were considered further resulting in 26 amino acid sites putatively associated with susceptibility/resistance data.

5.10.3 CART analysis

Decision trees were created by splitting the genotype-phenotype dataset until no further splits improved the accuracy. To avoid overfitting, trees were pruned. To assess the performance of the models to predict phenotype in unseen genotypes standard 10-fold crossvalidation was applied. For the classification models the generalization error is measured by the percentage of genotypes misclassified and the model's sensitivity and specificity. To select an optimal model, a ROC curve was calculated using 10-fold crossvalidation and the model which produced the best trade-off between sensitivity and specificity was chosen (Murray et al., 2008). Trees were pruned or extended to test the model fit on the basis of the ROC curve.

5.10.4 Random Forest analysis

This method constructs a number of decision trees where splits are determined on the selected subset of amino acid sites at each leaf. To interpret the Random Forest model, the importance of each amino acid site selected to predict phenotype was determined by the permutation accuracy importance (PAI) measure. PAI permutes the amino acid state in each of the sequences randomly which breaks the association with the phenotype on that amino acid site and then tests the impact on the accuracy of the model. Amino acid sites are ranked according to their impact.

5.10.5 Crossvalidation

10-fold crossvalidation was used to assess the predictive power of the models tested. The 542 HA sequences were randomly split into 10 parts (8 parts of 54 sequences each and 2 parts of 55 sequences each) with one part used as a test dataset and the 9 remaining parts of sequences used as a training dataset. This was repeated 10 times. True negative, true positive, false negative and false positive values and susceptibility and sensitivity were estimated.

5.10.6 Implementation

All models were created in the R environment (version 2.14.1). Decision trees were obtained by the tree and predict.tree functions in the tree library (Ripley, 2012). Pruned and extended trees were created by the prune.tree and tree.control functions respectively in the same library. Random forest trees were created by using randomForest and predict functions in the randomForest library (Liaw and Wiener, 2002). The ROC curves were calculated using prediction, performance and verify functions in the ROCR and verification libraries (Sing et al., 2005).

5.10.7 Phylogenetic analysis

Maximum likelihood phylogenetic analysis was conducted on 542 HA nucleotide sequences in RAxML v7.2.8 with 1000 bootstraps, the GTR (General Time Reversible) model of nucleotide substitution with the gamma rate heterogeneity and 4 discrete rate categories (Stamatakis et al., 2008). The output tree file was opened in FigTree v1.3.1 (Rambaut, 2007). The clade with the majority of resistant sequences was then separated from the other HA sequences and analysed in BEAST (Drummond and Rambaut, 2007) using the SRD06 nucleotide substitution model, a relaxed clock (uncorrelated lognormal) and a constant population size demographic model. The NA sequences for the same subset of strains were analysed in BEAST using the same settings.

5.11 Results

Full-length sequences of HA and matching NA sequences of human seasonal influenza A/H1N1 virus isolated from October 1st 2006 to December 31st 2008 were downloaded from the GISAID database. After excluding poorly represented regions, a dataset of 542 sequences from North, South and Central America and Asia (including Russia) was used in the analysis. Table 5.1 shows the geographical and temporal distribution of isolated sequences, divided into oseltamivir susceptible or

resistant according to mutations encoded in the NA. Resistance-conferring mutations in NA were present at only 1 amino acid site (of 2): 275 (N1 numbering). Out of 542 sequences, 354 were susceptible to oseltamivir and 188 were resistant. Zanamivir-resistant sequences or sequences resistant to both drugs, oseltamivir and zanamivir, were excluded from the study.

Table 5.1

Temporal and geographical distribution of sequences, divided into oseltamivir-susceptible (S) or oseltamivir-resistant (R).

	2006		2007		2008	
continent	S	R	S	R	S	R
North and Central America	5	/	101	18	66	103
South America	5	/	2	/	22	18
Asia (including Russia)	11	/	73	2	69	47

S=susceptible, R=resistant

Decision trees were used in initial analyses of the association between oseltamivir resistance and haemagglutinin genotype. Initial Decision Tree analyses resulted in relatively simple models with a depth of 4 for the unpruned tree and depth of 3 for the pruned one (Figure 5.3). The number of leaves was 4 for the pruned tree. The amino acid sites identified were 202, 205 and 206 (H1 numbering).

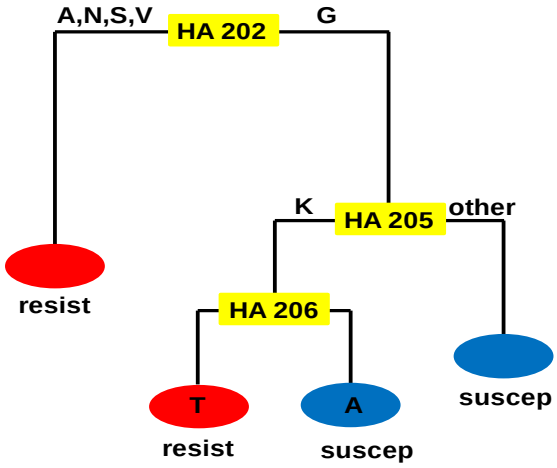

Figure 5.3

Classification tree from the CART analysis, showing three most important amino acid sites in HA and respective amino acids, associated with oseltamivir resistance in NA.The tree displayed has been pruned to 4 nodes only. Numbers indicate amino acid sites in the HA1 region in H1 numbering. Letters A, N, S, V, K, T indicate the amino acid variants observed in branches and leaves below that node. The amino acid site in HA that is associated most significantly with the oseltamivir - susceptible or –resistant phenotype (as encoded in NA) is 202. The second most significant amino acid site is 205 and the third most significant amino acid site is 206.

From the decision tree it can be seen that all mutations at amino acid site 202 in haemagglutinin are associated with resistance in NA – 92 strains had a mutation at this site (from G to any of A, N, S or V) all of which were resistant. However, 95 strains that were wild type (G) at 202 were also resistant to oseltamivir. Including amino acid sites 205 and 206 in the analysis allowed a further 42 of these to be identified. The percentage of samples misclassified using these 3 sites was 10.7% (58 samples). The percentage of correctly predicted resistant samples (sensitivity) was 71.3% and specificity (correctly predicting susceptibility in cases labelled as susceptible) was 98.6%. The area under the curve (AUC) was 0.882 (Table 5.2). CART analysis of susceptible versus resistant phenotype led to a correct classification of 89.1% of sequences on 10-fold crossvalidation.

Table 5.2

Specificity, sensitivity and AUC (area under the ROC curve) for five different models tested for determining association of HA genotype and oseltamivir resistance encoded in NA. The larger the values, the more significant the model.

model	SPECIFICITY	SENSITIVITY	AUC
CART unpruned, 7 leaves	0.986	0.713	0.899
CART pruned, 4 leaves	0.986	0.713	0.882
CART extended, 55 leaves	0.972	0.718	0.883
CART extended, 16 leaves	0.972	0.718	0.882
Random Forest	0.994	0.734	0.916

Random forest analysis is an extension of the decision tree learning where trees are created by growing individual trees on bootstrap samples of the data (Breiman, 2001). By using 10-fold crossvalidation the number of trees in the forest, the total number of amino acid sites randomly selected at each leaf and the maximum tree depth are optimized. To identify the key sites in the model, the permutation accuracy importance is calculated for each site by randomizing the amino acid allocations at the site and calculating the % change in the mean-square error of the model. The Random Forest model was a more accurate classifier than the Decision Trees with a higher specificity (99.4%), higher sensitivity (73.4%) and higher AUC (0.916) on the receiver operator characteristic (ROC) curve (Table 5.2 & Figure 5.4). While both models identified the same five amino acid sites: 202, 205, 206, 99, 52 (H1 numbering) they were not in the same order of importance (Figure 5.5). Other mutations are of marginal importance.

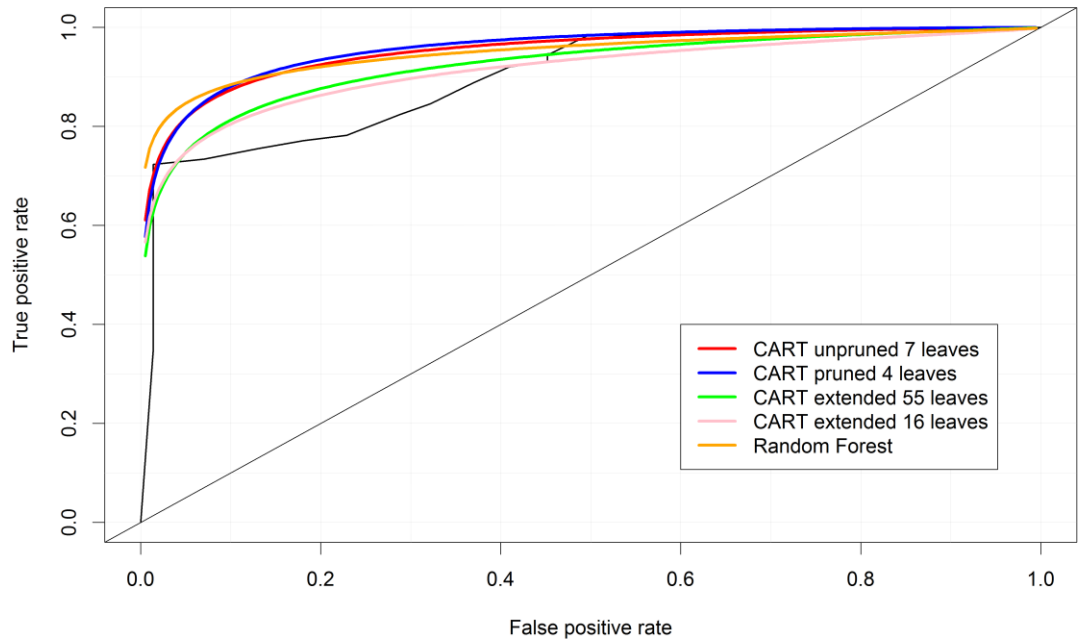

Figure 5.4

ROC curves for 4 different CART analyses and one Random Forest analysis. Red curve: CART model (unpruned, 7 leaves), blue curve: CART model (pruned, 4 leaves), green curve: CART model (extended, 55 leaves), pink curve: CART model (extended, 16 leaves), orange curve: Random Forest model, black curve: unfitted data for the CART model (unpruned). The X axis represents the false positive rate which equals 1-specificity. The Y axis represents the true positive rate which equals sensitivity. The ROC curve defines the area under the ROC curve-AUC-which can take value between 0 and 1. The larger the AUC value the better the model.

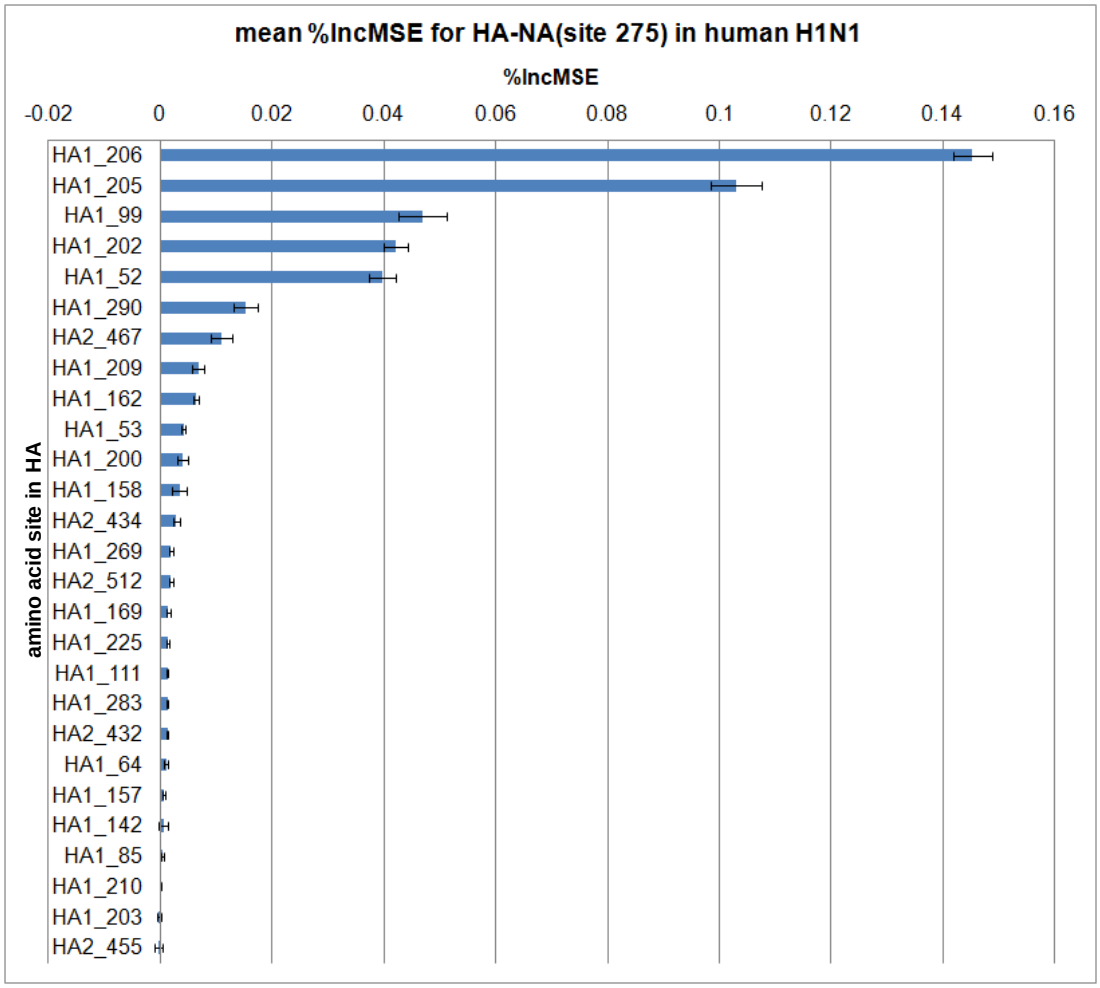

Figure 5.5

Mean increase in mean square error (%) (values on the horizontal axis) for amino acid sites in HA (values on the vertical axis) compared to amino acid site 275 in NA where susceptibility or resistance to oseltamivir is encoded. The larger the %IncMSE value the more significant the association between the amino acid sites evaluated. There are 27 amino acid sites in HA, determined as significantly associated with NA oseltamivir resistance/susceptibility. Amino acid sites are ordered according to their impact on the performance of the Random Forest model. The amino acid residue in HA associated most with the residue 275 in NA is at position 206 (and has the largest impact on the Random Forest model). HA1 indicates the codon is in the HA1 region of HA. HA2 indicates the codon is in the HA2 region of the HA. Numbers indicate amino acid sites in H1 numbering.

In order to understand better how the association between HA genotype and oseltamivir resistance came about the timeline of appearance of the associated variants was first examined in the context of the appearance of resistance. Figure 5.6 shows the presence of variant amino acids at sites 52, 99, 202, 205, 206 (H1

numbering) in HA according to the time the sequences were isolated. Most of the resistance-associated HA variants appeared sporadically from January 2007 onwards but it was the middle of November 2007 when genotypic resistance in NA took over and spread.

The phylogenetic relationships of the HA variants and resistance in the full dataset are indicated in Maximum likelihood trees coloured according to oseltamivir resistance in NA and different amino acids at positions 202, 205, 206 (H1 numbering) separately (Figures 5.7A, 5.7B, 5.7C, 5.7D, respectively). A large majority of HA sequences, belonging to oseltamivir resistant strains, fall within one clade (Figure 5.7A). Four sequences (2 of them found in other two clades) are cases of early sporadic appearances of resistance. There are 6 different HA variants at amino acid site 202 (H1 numbering; Figure 5.7B) with glycine being most common. This corresponds mostly to the susceptible strains but also to many resistant ones (95/542; 17.5%). Alanine (A), asparagine (N), serine (S) and valine (V), found in resistant strains only (see above), are mostly seen from November and December 2008 (Figure 5.6). The second most important amino acid site in HA (as recognized by CART) is 205 with 5 different amino acid variants, three of which are common and each found in the three phylogenetic clades (Figure 5.7C). Figure 5.7D shows 3 genotypes on amino acid site 206, with amino acids alanine (A) and threonine (T) prevailing. Those two amino acids are distributed approximately equally among susceptible and resistant strains.

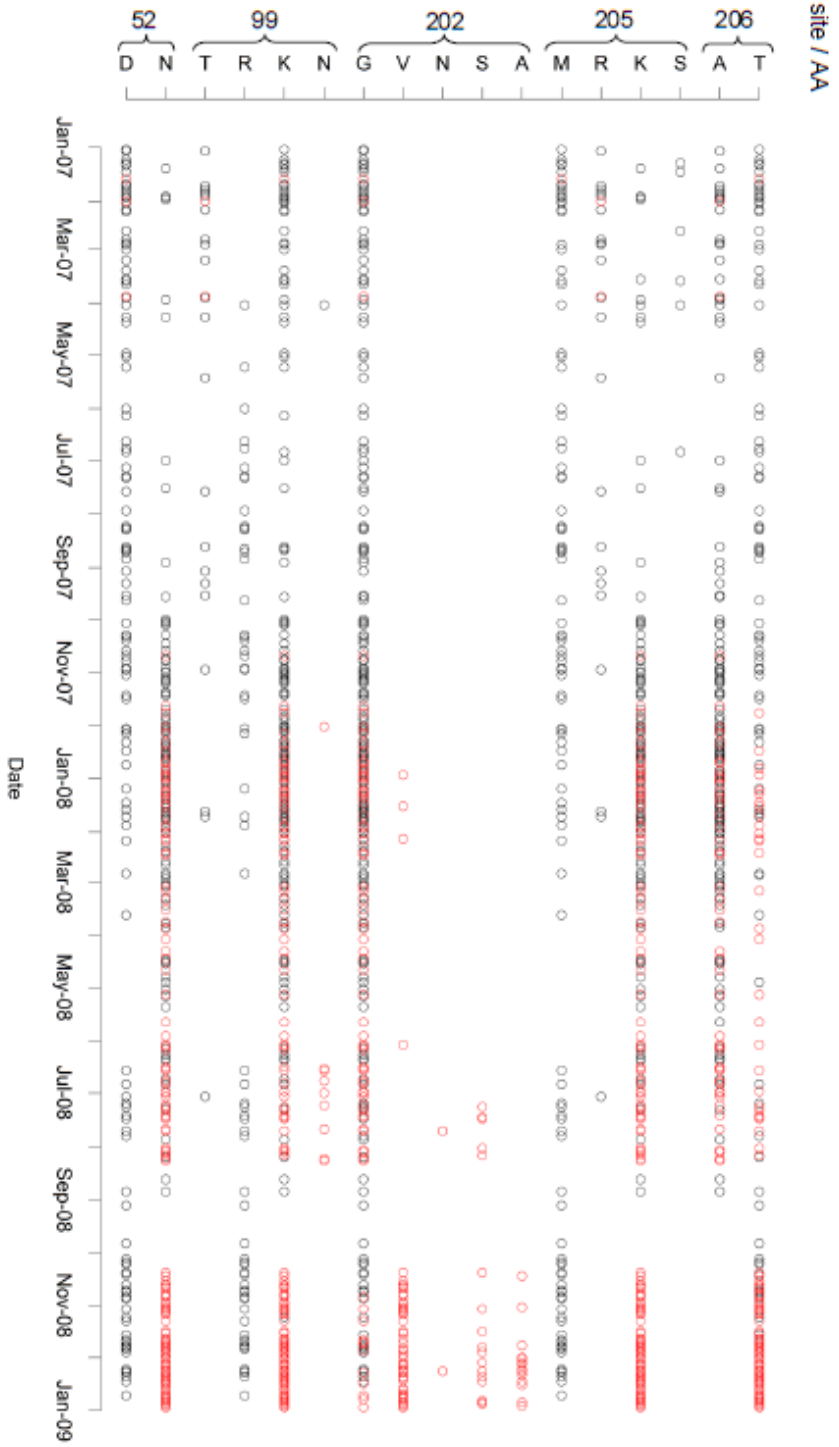

Figure 5.6

The most frequent amino acids in HA at five amino acid sites most strongly associated with oseltamivir susceptibility/resistance as identified with CART and RF models: 52, 99, 202, 205, 206 (H1 numbering) according to the isolation date of sequences. Red colour indicates resistance to oseltamivir, black colour indicates susceptibility to oseltamivir.

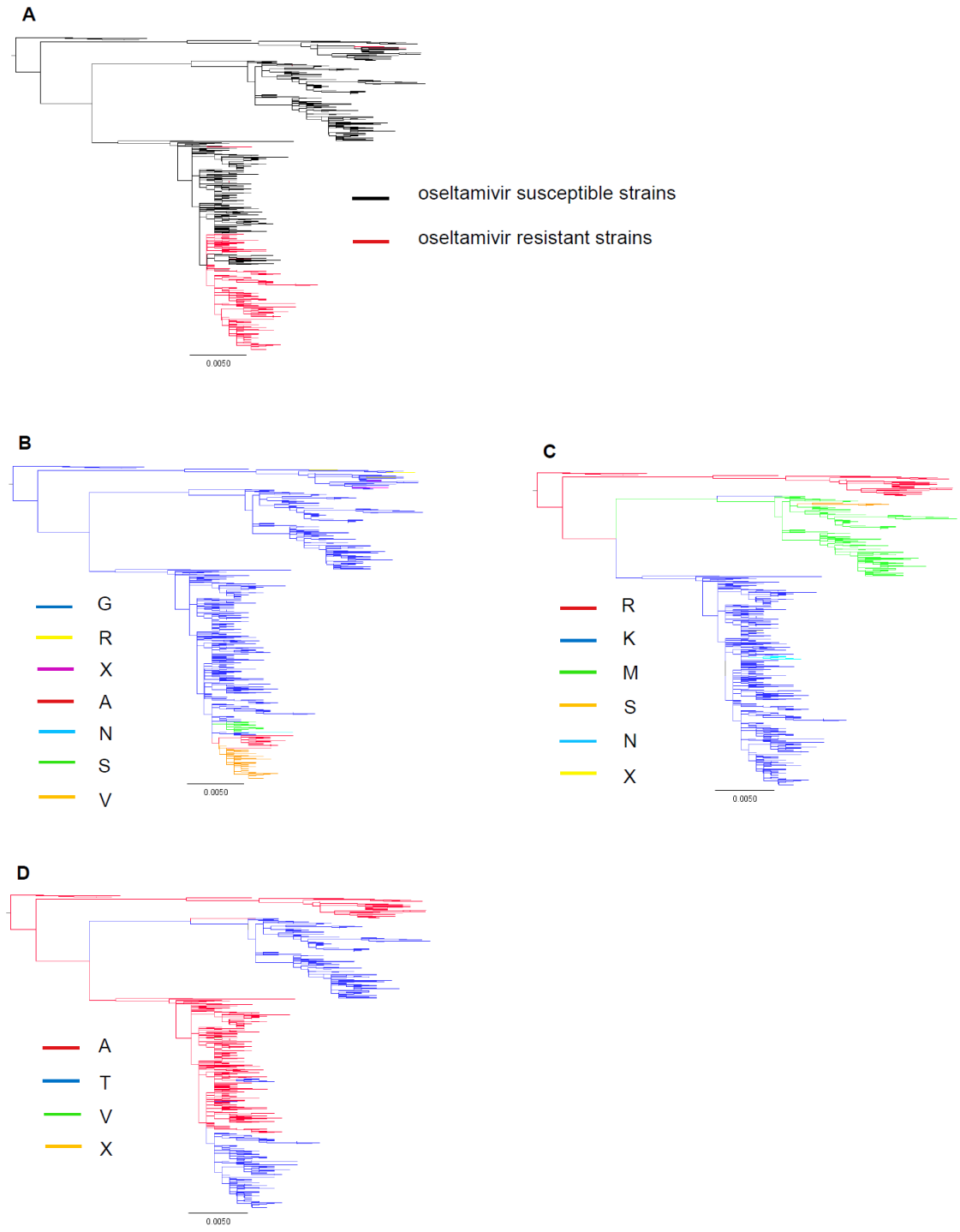

Figure 5.7

Maximum likelihood tree of 542 HA sequences, coloured according to: (A) oseltamivir resistance in NA, red colour indicates oseltamivir-resistant sequences; (B) different amino acids at site 202 (H1 numbering), blue: G, yellow: R, pink: X, red: A, turquoise: N, green: S, orange: V; (C) different amino acids at site 205 (H1 numbering), red: R, blue: K, green: M, orange: S, turquoise: N, yellow: X; (D) different amino acids at site 206 (H1 numbering), red: A, blue: T, green: V, orange: X.

The detailed relationships among these strains are shown in time-resolved trees in figures 5.8 (HA) and 5.9 (NA). From both trees it can be seen that sustained resistance appeared in a common ancestor in mid-2007. However, from figure 5.6 it is obvious that resistance started spreading in the middle of November 2007 as mentioned above. This discrepancy probably reflects the lack of influenza transmission in the Northern hemisphere summer. While the HA and NA trees are similar, particularly in regards to the time estimate of sustained resistance, however, they differ in one important detail. The HA tree (Figure 5.8) contains a clade of susceptible strains (coloured blue) among two resistant clades (red). In the NA tree (Figure 5.9) this susceptible clade (coloured blue) is found earlier and shares a common ancestor with the resistant sequences. It can be concluded that a reassortment event occurred between HA and NA lineages in association concurrent with the spread of oseltamivir resistance.

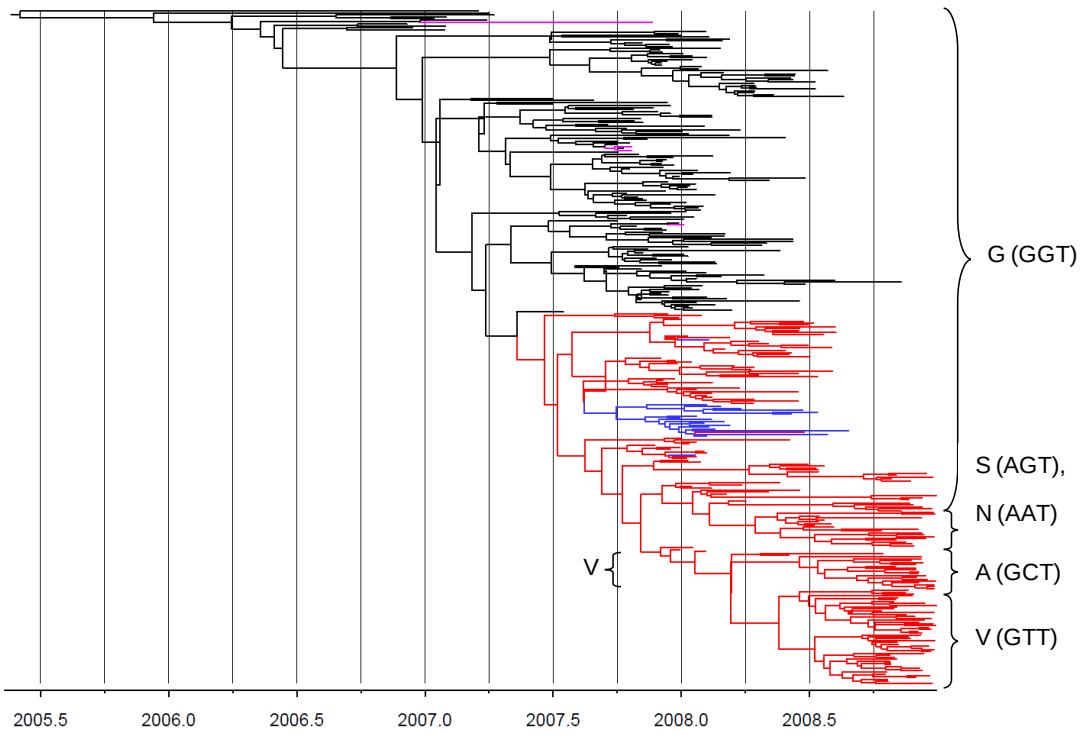

Figure 5.8

BEAST tree for a partial dataset of HA sequences forming a clade on the maximum likelihood tree with the majority of oseltamivir-resistant strains. This tree shows the evolutionary relationship among strains and the amino acid variants in HA at position 202 (H1 numbering). Black: susceptible strains, violet: sporadic resistant strains, blue: susceptible strains, found between two resistant HA clades, red: resistant strains

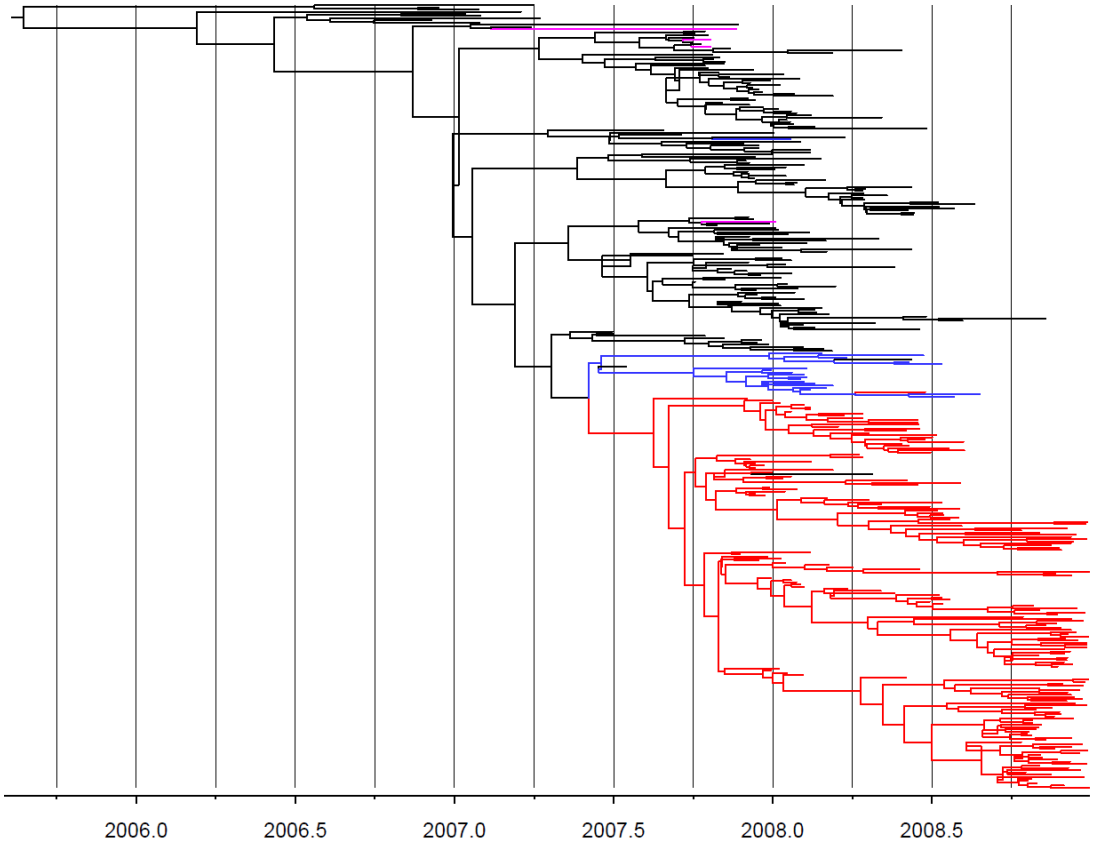

Figure 5.9

BEAST tree for a partial dataset of NA sequences matching the HA sequences that form a clade on the maximum likelihood tree with the majority of oseltamivir-resistant strains. Black: susceptible strains, violet: sporadic resistant strains, blue: susceptible strains, found between two resistant HA clades, red: resistant strains

The amino acids observed at position 202 in HA are plotted as tip labels on Figure 5.8. This shows that while among the susceptible strains glycine is almost universal, within the resistant clade there is substantial variability at this site, with glycine and 3 alternative amino acids, serine, alanine and valine, appearing in different resistant subclades. Each of these changes involved a single nucleotide substitution, with asparagines evolving from serine in 2 cases by one further single change. The asparagine arises from serine and not directly from glycine, while alanine and valine come directly from glycine. So the wild-type codon is GGT as expected but not GGA or GGG (Figure 5.8). From the time-resolved tree it can be seen that the wild-type glycine at HA 202 persisted at least 6 months after resistance arose in NA and

indeed, in a minor subclade continued until the end of 2008. The first variant to appear was valine (late 2007), followed in a different subclade by serine (early 2008) and alanine in a third subclade in mid-2008. There is therefore no simple phylogenetic explanation for the association of HA variation with oseltamivir resistance determined by NA.

5.12 Discussion

In this study evidence has been found which shows complex pattern of associations between genetic variants in HA and the appearance of the NA clade within which resistance arose, at sites in HA close to the receptor binding site.

The main mutation responsible for the oseltamivir resistant phenotype is H275Y (N1 numbering), however it is strongly deleterious to the virus in the absence of the drug (Aoki et al., 2007). Evidence that a balance of activity between NA and HA is required has been found in studies done by Xu et al (2012), Richard et al (2012), Lu et al (2005) and Baigent and McCauley (2001). Epistatic interactions between genetic variants at different loci can have a significant effect on fitness. Kryazhimskiy et al. (2011) identified 225 epistatic pairs of sites in H1 and 205 of them in N1. However, they did not look explicitly at associations between HA and NA. In this study a dataset comprising 542 human seasonal A/H1N1 sequences of HA and NA was obtained. Sequences were isolated from the end of 2006 and the whole of 2007 and 2008 – just before and just after the emergence of oseltamivir resistance in this group of influenza virus.

Associations between oseltamivir resistance (determined by the presence of H275Y in NA) and genetic variants across HA were analysed using machine learning models including Classification and Regression Trees analysis (CART) and Random Forest which both permit the identification of contributing amino acid sites (unlike neural networks or support vector machines). The results showed that three amino acid sites in HA are important for determining drug-resistant phenotype in influenza (202, 205,

206; H1 numbering) which were identified by both models used, although not in the same order of importance.

Amino acids at sites 205 and 206 are two of the amino acids that form receptor-binding sites (Sriwilaijaroen and Suzuki, 2012). Amino acid site 206 is also one of the antigenic sites. Amino acid site 202 is near the 190-helix, which is a receptor-binding site and amino acid site 99 is near a conserved glycosylation site (Sriwilaijaroen and Suzuki, 2012). Taking that in consideration it is not surprising that the mutations at sites 202, 205 and 206 are so diverse and consequently enabling the existence of different HA genotypes over a short time period. Looking at the frequency of the resistance-conferring mutations at position 202 in HA it can be seen, that all except one are found in strains, isolated in 2008.

Rameix-Welti et al. (2008) studied enzymatic properties of the neuraminidase of seasonal A/H1N1 and found that resistant viruses that harboured the H275Y mutation had a slightly higher activity and affinity for the substrate than sensitive viruses. They concluded that these features could contribute to the overall fitness of resistant strains, however, they did not exclude the possibility of the contribution from other genes.

Co-occurrence of mutations at different amino acid sites is expected, even in different genes because of shared phylogenetic history (Poon et al., 2007). However, the pattern observed in HA 202 is complex and cannot be explained on that basis. First, the wild-type glycine persists into several resistant subclades, then different allelic forms arise in 3 different resistant subclades at different times. It can be suggested that an optimum configuration was disrupted, which led to a number of variants having similar short-term fitness. Secondly, a reassortment event occurred in early to mid-2007 which led to a “resistance-adapted” HA clade becoming associated with a sensitive NA. This subclade lasted about a year but then failed. However, the long-term outcome of this set of alternative configurations cannot now be known as seasonal H1N1 was displaced by the appearance of pandemic H1N1 in 2009 and has not returned since (WHO, 2013). Nevertheless, the recurrent observation of

oseltamivir resistance in pandemic H1N1 (Ghedin et al., 2012; Storms et al., 2012; Hurt et al., 2012) show the importance of understanding the fitness interactions that may affect their long term fate (Kelso and Hurt, 2012).

Chapter 6

Thesis summary and discussion

6 Thesis summary and discussion

Influenza infections can be combated by two different strategies: vaccination and chemotherapy. The latter has been used for many decades because of the low price of anti-influenza drugs and limited vaccination effectiveness due to antigenic drifts and shifts that the influenza virus undergoes from year to year. However, viral resistance to both classes of anti-influenza drugs has emerged and spread widely (Lackenby et al., 2008).

In my PhD thesis I have focused on the origin and mechanisms of amantadine resistance in swine influenza A viruses and oseltamivir resistance in human seasonal influenza A viruses as well as reassortment in swine influenza A viruses.

My first research chapter focused on the quantification of reassortment in swine influenza viruses. Namely, reassortment is not only the mechanism for the emergence of novel influenza viruses (such as A/H1N1pan09) but also a mechanism for the acquisition and spread of drug resistant mutations in influenza viruses. As shown in the study by Simonsen et al (2007) the amantadine resistant A/H3N2 viruses circulating in 2005-2006 influenza season were generated by reassortment and genetic drift. Also, the study by Yang et al (2011) found that the same mechanisms are responsible for the emergence of oseltamivir resistant A/H1N1 viruses. Additionally, the amantadine resistant mutation S31N in M2 of A/H3N2 viruses and oseltamivir resistant mutation H275Y (N1 numbering) in NA of A/H1N1 were spread due to a hitch-hiking event (Simonsen et al., 2007; Yang et al, 2011).

Because reassortment is a frequent event appearing in pigs, contributing both to novel influenza virus emergence as well as the spread of drug resistant mutations, my study focused on determining the extent of this event in pigs. There are some studies published that researched the extent of reassortment in swine influenza viruses, however, none of them was as extensive as my study. Nelson et al (2012) based their paper on coloured phylogenetic trees but did also heat-maps. Khiabanian et al (2009) also used heat-maps and Lycett et al (2012) used coloured phylogenetic trees and

GARD. My study also used coloured phylogenetic trees and GARD but not heat-maps. In addition, my study used the distance in topology measure, which has not been described in published papers. The limitation of using coloured phylogenetic trees is that the trees usually represent only a few viral segments because of the large number of all possible trees. That was the case also in my study, where trees were coloured only according to one segment (MP). The results of my study support the results of previous studies. Namely, the segments most prone to reassortment are HA and NA that are reassorting most frequently with the polymerase segments. Pigs are considered to be mixing vessels for reassortment, where reassortment of viral segments from different hosts (swine, avian, human) can occur and spread to these hosts. Therefore, also established swine viral lineages contain segments from not only swine influenza viruses but also avian and human influenza viruses. The study by Furuse et al (2010) revealed the mutations in six internal segments of different hosts that appeared due to reassortment in order to break down the host barrier of the virus and enable its spread to new hosts. For swine host, Furuse et al (2010) looked for mutations in segments NP, MP and NS as these segments were retained in swine classical and swine triple reassortant lineages after reassortment. They found some mutations in all three of these segments and these mutations were similar to amino acids in avian and human host. These mutations probably appeared in order to accommodate avian PB2 and PA segments and human PB1 segment after reassortment (Furuse et al., 2010). My study also confirms that it is these three polymerase segments that are very prone to reassortment.

My second research chapter focused on the amantadine resistance in swine flu. The first thing I noticed was a wide spread of amantadine resistance in pigs and second, the appearance of this resistance (in swine H1N1) even before amantadine was available and after it was no longer widely used. Third, by far the most prevalent mutation in all types and both hosts studied (swine, human) was the mutation S31N in M2. It has been mentioned before (Hay et al., 1985) that the site 31 in M2 includes only one mutation, while on other resistance-associated sites, the variability of mutations is greater. It was also noted that in human influenza A viruses the S31N mutation is closely linked to significant outbreaks or epidemics of human influenza

or shows the high selective pressure of amantadine because of its high usage (Cheung et al., 2006; Lee et al., 2008). In swine, however, I believe that is not the case because as mentioned above, resistance to this drug appeared before amantadine was available and persisted strongly even after its usage declined. Another set of studies propose the appearance of S31N mutation to be linked to the emergence of viruses from the same lineage (Simonsen et al., 2007; Krumbholz et al., 2009). However, again I do not agree with that because in my study as mentioned above, the S31N mutation was the most prevalent in different subtypes and different hosts. Some authors (Saito et al., 2003; Abed et al., 2005; Astrahan et al., 2004) try to justify the widespread appearance of the S31N mutation in M2 with its high degree of resistance to amantadine compared to the V27A mutation which renders the virus resistant but impairs the M2's channel activity by 33% (Pinto et al., 1992). Other authors noted that the same mutation can have different impact on different M2 types (at different pH) (Balannik et al., 2010; Pinto et al., 1992; Holsinger et al., 1994). It is known that the S31N mutation causes impaired growth of the virus (Grambas et al., 1992) therefore it has been shown (Grambas et al., 1992) that other mutations in M2 (such as I27T and I27S) occur that cause an increase in M2 activity. However, in my study, those two mutations appeared rarely and were not always associated with the resistant S31N mutation. Taken together, majority of the studies focused only on the M2 ion channel but neglected other viral proteins.

However, some studies (Grambas et al., 1992; Steinhauer et al., 1991; Ilyushina et al., 2007) noticed that along with mutations in M2, the amantadine also causes mutations in haemagglutinin protein. These studies showed that in some cases the mutated hæmeagglutinin that arose by the growth in the presence of amantadine, is sufficient to provide amantadine tolerance. My study went even further and checked the amino acid changes in not only M2 but also M1, HA and NA proteins that are associated with the serine to asparagines change at site 31 in M2. I noticed that there are several mutations in each of these proteins that coincide with the S31N mutation, with one mutation in each protein being strongly associated with the resistance-conferring mutation in M2. I also show co-evolution of these amino acid sites in four different proteins. Considering that these four proteins come in a physical contact

inside a virion and that these physical interactions are important for virus' physiology (McCown and Pekosz, 2006; Burleigh et al., 2005; Lakdawala et al., 2011; Rossman and Lamb, 2011; Bilsel et al., 1993; Mitnaul et al., 1996; Chen et al., 2005) my results indicate there are epistatic interactions among these four proteins which are responsible for their co-evolution. Due to hitch-hiking these mutations spread in the population, one of them being the amantadine resistance-conferring mutation S31N. So, the reason for the wide spread of this mutation lies in the benefits for the virus other than viability in the presence of amantadine (filamentous morphology of the virions associated with their capacity for efficient transmissibility, evading host's immune response by changing epitopes in HA and NA).

The third research chapter of my thesis studied the correlation of amino acid variants in HA with oseltamivir resistance encoded in NA. Namely, it is known that the functions of HA and NA are balanced (Bantia et al., 1998; Blick et al., 1998; McKimm-Breschkin et al., 1996) therefore if the function of one of these two glycoproteins is decreased the other glycoprotein will also acquire mutations that decrease its function to enable the virus to still be infectious. One of the reasons for lower activity of NA is chemotherapy with neuraminidase inhibitors such as oseltamivir which binds in the active site of NA enzyme and prevents the NA from cleaving sialic acid residues (Moscona, 2005). The consequence is the inhibition of release of newly formed virions from infected cells (Liu et al., 1995), viral aggregation on the cell surface (Liu et al., 1995) and reduced viral infectivity (Suzuki et al., 2005; Matrosovich et al., 2004). Oseltamivir was first available in 1999 and until 2007-2008 influenza season it was very efficient in combating influenza infections. However, in 2007-2008 influenza season the oseltamivir resistance-conferring mutation H274Y (N2 numbering) appeared worldwide in seasonal H1N1 (Moscona, 2009) and within a year it was present in most seasonal H1N1 (Moscona, 2009). The H274Y mutation is found near but not directly in the substrate-binding pocket (Russell et al., 2006) and causes structural alterations of the NA that weaken oseltamivir binding (Russell et al., 2006; Collins et al., 2008) but has also a negative effect on viral growth and is therefore quickly replaced with a wild type NA in the absence of the drug. However, the resistant NA from 2007-2008 influenza season

showed no obvious attenuation relative to earlier viruses carrying tyrosine at position 274 (Rameix-Welti et al., 2008). This feature sparkled interest of scientific community that tried to find the mechanism(s) responsible for that. For example, Bloom et al (2010) hypothesized that the evolution of oseltamivir resistance was enabled by permissive mutations that alleviated the deleterious effects of subsequent occurrences of H274Y. They also further hypothesized that these permissive mutations obliterated previously unobserved deficiencies in NA folding or stability caused by H274Y (Bloom et al., 2010). Using a computational method (Bloom and Glassman, 2009), Bloom et al (2010) found the mutation R194G to restore the total surface-expressed activity of NA with H274Y to approximately wild-type levels. They also found two further mutations, R222Q and V234M, to decrease the magnitude of the defect caused by H274Y. However, Bloom et al (2010) did not elucidate the mechanism(s) responsible for the spread of these mutations in seasonal H1N1. They speculated that the permissive mutations could be the result of a drift, genetic hitchhiking (Simonsen et al., 2007), selection for antigenic change or balancing the NA/HA functions (Wagner et al., 2002). Unlike the study by Bloom et al (2010), my study focused on the HA and its mutations that followed the H274Y acquisition. Using two machine learning methods, we found several mutations in HA that appeared as the consequence of influenza virus acquiring the resistant NA. We concluded that these mutations were due to epistatic interactions between NA and HA and enhanced viral fitness. However, my study lacked to identify the substitutions in NA that might have also been associated with the appearance of H274Y and this is something that remains to be done as future work. On the other hand, Kryazhimskiy et al (2011) identified 225 pairs of sites in H1 and 205 in N1 of seasonal H1N1 that were epistatically linked. For N1 NA, the two leading sites were 344 and 275 (N1 numbering). Site 275 was associated with six other mutations that most likely enhance the fitness of the virus with H274Y mutation and therefore enable the virus to survive. So, the appearance of the resistance-conferring mutation in N1 is either predating or following other mutations in the same gene as also described by Bloom et al (2010).

When analysing influenza sequence data, it is imperative to search for sequence data from and/or upload sequence data to the appropriate database. One of the most important databases for influenza is GenBank, which plays a vital role in sharing and archiving influenza virus sequences (Schnirring, 2009). Another pivotal sequence repository of genetic data for influenza is GISAID's EpiFlu database, established in 2006, which provides a more complete picture of the influenza data because it includes some previously unpublished sequences and also because it permits scientists to submit extra information, such as clinical features and viral passage history. The advantage of using this database is also in the availability of sequences immediately after their submission (Schnirring, 2009).

The global spread of some drug resistant influenza strains, such as now extinct human seasonal H1N1 resistant to oseltamivir, and swine influenza circulating in Eurasia resistant to amantadine, as well as the pandemic H1N1 2009 infecting humans, has stressed the importance of fast, inexpensive and accurate sequencing methods for creating a broader array of sequences available. Apart from increasing the number of influenza strains sequenced it is also important to sequence the whole viral genome, not just heamagglutinin and neuraminidase segments as was the case years ago. Having whole influenza genomes available facilitates better understanding of the evolution of influenza A viruses in humans as well as other hosts, which is important for the surveillance of this important pathogen and also for selecting the appropriate influenza vaccine strain (Holmes et al., 2005).

Because multiple influenza lineages can co-circulate, persist and reassort to cause epidemics and occasional pandemics it is very important to obtain full genomic data of this viruses for future influenza surveillance (Holmes et al., 2005).

The influenza surveillance is crucial in order to be able to predict the emergence of new circulating influenza strains for developing influenza vaccine on an annual basis (Gensheimer et al., 1999). Today, the global influenza surveillance network is coordinated by the World Health Organization, which every year selects the appropriate strains of influenza A and B viruses for the production of vaccine for the

northern and southern hemispheres. For making the best recommendation on candidate vaccine strains, antigenic, genetic and epidemiological data are examined (Holmes et al., 2005) which again underlies the importance of maintaining sequence databases and their extent regarding the amount of information available for each strain.

Because the number of influenza virus sequences in the past was small and the genetic information usually available just for HA and NA segments, it is unclear how the entire genome of influenza A viruses evolves during epidemics. Sequencing of influenza A virus genome was in the past limited to HA and NA segments only because it was known that antigenic drift of these two proteins is vital for the survival of influenza viruses (Colman et al., 1983). Today we know that other factors such as binding specificity of HA (Nobusawa et al., 2000), matched activity between HA and NA (Kaverin et al., 2000; Mitnaul et al., 2000; Wagner et al., 2002) and the interaction of other influenza virus' proteins with each other and with the proteins of their host cells are also affecting viral fitness (Holmes et al., 2005). Because of this lack of molecular data for segments other than HA and NA, it is also unclear how many lineages of influenza A viruses are circulating between epidemics (Holmes et al., 2005). Due to lack of molecular data it was also believed that influenza genetic diversity is more restricted and due only to genetic drift. However, with the accumulation of whole genome sequences of influenza viruses from different geographical locations, it became clear that reassortment is a very important event that has the potential to cause sudden and significant changes to viral antigenic structure (Holmes et al., 2005). Therefore, collecting whole genome sequences from different geographic locations is very important for understanding the mechanisms of viral evolution and its rates, for understanding the pathogenesis and transmission patterns of influenza viruses between humans and animals and for planning surveillance and control measures (Holmes et al., 2005).

For surveillance purposes a list of notifiable diseases of humans and animals, separately, has been set at the national level and renewed annually. Notifiable disease is considered one which has to be reported in a timely manner in order to prevent and

control the disease and therefore protect the public's health. Influenza infections are considered a notifiable disease of both, humans and animals. For humans, deaths due to influenza infection and infections with a novel influenza A virus have to be reported. For animals, infection of birds with highly pathogenic avian influenza virus has to be reported as well as an infection of humans with swine influenza viruses (CDC, 2012; Gov.UK, 2014).

Swine influenza A viruses that infect swine do not normally infect humans and human influenza A viruses that infect humans do not normally infect pigs. However, sporadic infections of humans with swine influenza viruses and infections of pigs with human influenza viruses have been reported. Transmission of swine influenza viruses from human to human and transmission of human influenza viruses from pig to pig are rare but do occur (CDC, 2014). As pigs are considered to be a mixing vessel for generating novel influenza strains due to reassortment, these transmissions of influenza viruses from pigs to humans and vice versa as well as the capability of pigs to be equally susceptible to infection with not only pig and human influenza viruses but also avian influenza viruses, open a biosecurity issue for the health of humans and animals. Namely, novel influenza strains unseen before can be generated and spread in the susceptible hosts but they can also be drug resistant due to reassortment which included a resistant virus.

So, my thesis which studied the mechanisms responsible for influenza virus' drug resistance and spread is relevant for understanding how to better protect human health in terms of a reasonable drug usage but can also be applied to vaccination of both, humans and pigs. It also underlies the importance of influenza surveillance in humans, pigs and birds in order to detect early the occurrence of novel influenza strains that can potentially be dangerous for human and animal health.

Chapter 7

Bibliography

7 Bibliography

A revised system of nomenclature for influenza viruses (1971). *Bull World Health Organ* **45**: 119-124.

Abed Y, Goyette N, Boivin G (2005). Generation and characterization of recombinant influenza A (H1N1) viruses harbouring amantadine resistance mutations. *Antimicrob Agents Chemother* **49:** 556-559.

Air, G.M. (2012). Influenza neuraminidase. *Influenza Other Respir Viruses* **6**, 245-256.

Aldridge S (2001). New era for the European Bioinformatics Institute. *Genome Biol* **2:** spotlight-20010919-20010901.

Anisimova M, Gascuel O (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol* **55:** 539-552.

Anon (1979). Reconsideration of influenza A virus nomenclature: a WHO memorandum. *Bull World Health Organ* **57**: 227-233.

Anonim (2013). The R project for statistical computing. http://www.r-project.org/ [cited on 13 Aug 2014]

Aoki, F.Y., Boivin, G., Roberts, N. (2007). Influenza virus susceptibility and resistance to oseltamivir. *Antiviral Therapy* **12**, 603-616.

Arenas M, Posada D (2007). Recodon: Coalescent simulation of coding DNA sequences with recombination, migration and demography. *BMC Bioinformatics* **8:** 458.

Aris-Brosou S, Yang Z (2003). Bayesian models of episodic evolution support a late Precambrian explosive diversification of the Metazoa. *Mol Biol Evol* **20:** 1947-1954.

Astrahan P, Kass I, Cooper MA, Arkin IT (2004). A novel method of resistance for influenza against a channel-blocking antiviral drug. *Proteins* **55:** 251-257.

Ayaz M (2012). Use of antiviral drug (amantadine) in commercial poultry – an alarm for future strategy on control of avian influenza in the country. (www.agricorner.com) [cited 14th February 2013]

Aymard M, Gourreau JM, Kaiser C, Fontaine M, Madec F, Tillon JP (1985). Les marqueurs immunovirologiques du risque d'influenza A H3N2 chez les porcs. *Revue Epidemiologie Medical Societe Sante Publique* **33:** 283–291.

Babcock HP, Chen C, Zhuang X. (2004). Using single-particle tracking to study nuclear trafficking of viral genes. *Biophys J* **87**: 2749–2758.

Baigent SJ, McCauley JW. (2003). Influenza type A in humans, mammals and birds: determinants of virus virulence, host-range and interspecies transmission. *Bioessays* **25**: 657-671.

Baigent, S.J., McCauley, J.W. (2001). Glycosylation of haemagglutinin and stalk-length of neuraminidase combine to regulate the growth of avian influenza viruses in tissue culture. *Virus Res* **79**, 177-185.

Balannik V, Obrdlik P, Inayat S, Steensen C, Wang J, Rausch JM, DeGrado WF, Kelety B, Pinto LH (2010). Solid-supported membrane technology for the investigation of the influenza A virus M2 channel activity. *Pflugers Arch* **459:** 593-605.

Banks J, Speidel ES, Moore E, Plowright L, Piccirillo A, Capua I, Cordioli P, Fioretti A, Alexander DJ. (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.

Banks J, Speidel EC, McCauley JW, Alexander DJ. (2000). Phylogenetic analysis of H7 haemagglutinin subtype influenza A viruses. *Arch Virol* **145**: 1047-1058.

Bantia, S., Ghate, A.A., Ananth, S.L., Babu, Y.S., Air, G.M., Walsh, G.M. (1998). Generation and characterization of a mutant of influenza A virus selected with the neuraminidase inhibitor BCX-140. *Antimicrob Agents Chemother* **42**, 801-807.

Baranovich, T., Saito, R., Suzuki, Y., Zaraket, H., Dapat, C., Caperig-Dapat, I., Oguma, T., Shabana, I.I., Saito, T., Suzuki, H., the Japanese Influenza Collaborative Study Group 1 (2010). Emergence of H274Y oseltamivir-resistant A(H1N1) influenza viruses in Japan during the 2008-2009 season. *Clin Virol* **47**, 23.

Barman S, Adhikary L, Chakrabarti AK, Bernas C, Kawaoka Y, Nayak DP. (2004). Role of transmembrane domain and cytoplasmic tail amino acid sequences of influenza a virus neuraminidase in raft association and virus budding. *J Virol* **78**: 5258–5269.

Barry JM (2005). The story of influenza: 1918 revisited: lessons and suggestions for further inquiry. In: The threat of pandemic influenza: are we ready? Workshop summary (Knobler SL, Mack A, Mahmoud A, Lemon SM; Eds). The National academic press, pp 60-61.

Barton, N H (2000). Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **355** (1403): 1553–1562.

Baudin F, Bach C, Cusack S, Ruigrok RW (1994). Structure of influenza virus nucleoprotein melts secondary structure in panhandle RNA and exposes the bases to the solvent. *EMBO J* **13:** 3158-3165.

Baum LG, Paulson JC. (1991). The N2 neuraminidase of human influenza virus has acquired a substrate specificity complementary to the hemagglutinin receptor specificity. *Virology* **180**: 10-15.

Beare AS, Webster RG (1991). Replication of avian influenza viruses in humans. *Arch Virol* **119**: 37-42.

Beerenwinkel, N., Daumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., Lengauer, T., Selbig, J., Walter, H. (2003). Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res* **31**, 3850-3855.

Beerenwinkel, N., Lengauer, T., Selbig, J., Schmidt, B., Walter, H., Korn, K., Kaiser, R., Hoffmann, D. (2001). Geno2pheno: Interpreting Genotypic HIV Drug Resistance Tests. *Intelligent systems in biology* **16**, 35-41.

Beerenwinkel, N., Schmidt, B., Walter, H., Kaiser, R., Lengauer, T., Hoffmann, D., Korn, K., Selbig, J. (2002). Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype *Proc Natl Acad Sci U S A* **99**, 8271-8276.

Belser JA, Maines TR, Tumpey TM, Katz JM (2010). Influenza A virus transmission: contributing factors and clinical implications. *Expert Rev Mol Med* **12:** e39.

Belshe RB, Smith MH, Hall CB, Betts R, Hay AJ (1988). Genetic basis of resistance to rimantadine emerging during treatment of influenza virus infection. *J Virol* **62** (5): 1508-1512.

Berglund EC, Nystedt B, Andersson SG (2009). Computational resources in infectious disease: limitations and challenges. *PLoS Comput Biol* **5:** e1000481.

Bilsel P, Castrucci MR, Kawaoka Y (1993). Mutations in the cytoplasmic tail of influenza A virus neuraminidase affect incorporation into virions. *J Virol* **67:** 6762-6767.

Blick, T. J., Sahasrabudhe, A., McDonald, M., Owens, I.J., Morley, P.J., Fenton, R.J., McKimm-Breschkin, J.L. (1998). The interaction of neuraminidase and haemagglutinin mutations in influenza virus in resistance to 4-guanidino-Neu5Ac2en. *Virology* **246**, 95-103.

Blok J, Air GM. (1982). Block deletions in the neuraminidase genes from some influenza A viruses of the N1 subtype. *Virology* **118**: 229–234.

Bloom JD, Gong LI, Baltimore D (2010). Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* **328:** 1272-1275.

Bloom JD, Glassman MJ (2009a). Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. *PLoS Comput Biol* **5:** e1000349.

Bloom, J.D., Arnold, F.H. (2009b). In the light of directed evolution: Pathways of adaptive protein evolution. *Proc Natl Acad Sci USA* **106**, 9995–10000.

Boivin G, Goyette N, Bernatchez H (2002). Prolonged excretion of amantadine-resistant influenza a virus quasi species after cessation of antiviral therapy in an immunocompromised patient. *Clin Infect Dis* **34** (5): E23-E25.

Boni MF (2008). Vaccination and antigenic drift in influenza. *Vaccine* **26** (Suppl. 3): C8–C14.

Bosch FX, Garten W, Klenk HD, Rott R. (1981). Cleavage of influenza virus hemagglutinins: Primary structure of the connecting peptide between HA1 and HA2 determines proteolytic cleavability and pathogenicity of avian influenza virus. *Virology* **113**: 725-735.

Boulay F, Doms RW, Wilson I, Helenius A. (1987). The influenza hemagglutinin precursor as an acid-sensitive probe of the biosynthetic pathway. *EMBO J* **6**: 2643–2650.

Boulo S, Akarsu H, Ruigrok RW, Baudin F. (2007). Nuclear traffic of influenza virus proteins and ribonucleoprotein complexes. *Virus Res* **124:** 12-21.

Bouloy M, Plotch SJ, Krug RM. (1980). Both the 7-methyl and the 2'-O-methyl groups in the cap of mRNA strongly influence its ability to act as primer for influenza virus RNA transcription. *Proc Natl Acad Sci U S A* **77:** 3952-3956.

Bouloy M, Morgan MA, Shatkin AJ, Krug RM. (1979). Cap and internal nucleotides of reovirus mRNA primers are incorporated into influenza viral complementary RNA during transcription in vitro. *J Virol* **32**: 895-904.

Bouloy M, Plotch SJ, Krug RM. (1978). Globin mRNAs are primers for the transcription of influenza viral RNA in vitro. *Proc Natl Acad Sci U S A* **75**: 4886-4890.

Bourmakina SV, Garcia-Sastre A (2003). Reverse genetics studies on the filamentous morphology of influenza A virus. *J Gen Virol* **84:** 517-527.

Bouvier NM, Palese P (2008). The biology of influenza viruses. *Vaccine* **26 (Suppl):** D49-D53.

Breiman L (2001). Random Forests. *Machine Learning* 45: 5-32.

Breiman L, Friedman JH, Olshen RA, Stone cj (1984). Classification and regression trees (2nd Ed.). Pacific Grove, CA; Wadsworth.

Briedis DJ, Lamb RA. (1982). Influenza B virus genome: sequences and structural organization of RNA segment 8 and the mRNAs coding for the NS1 and NS2 proteins. *J Virol* **42**: 186-193.

Bright RA, Medina MJ, Xu X, Perez-Oronoz G, Wallis TR, Davis XM, Povinelli L, Cox NJ, Klimov AI (2005). Incidence of adamantine resistance among influenza A (H3N2) viruses isolated worldwide from from 1994 to 2005: a cause for concern. *Lancet* **366:** 1175-1181.

Brookes SM, Núñez A, Choudhury B, Matrosovich M, Essen SC, Clifford D, Slomka MJ, Kuntz-Simon G, Garcon F, Nash B, Hanna A, Heegaard PM, Quéguiner S, Chiapponi C, Bublot M, Garcia JM, Gardner R, Foni E, Loeffen W, Larsen L, Van Reeth K, Banks J, Irvine RM, IH B (2010). Replication, pathogenesis and transmission of pandemic (H1N1) 2009 virus in non-immune pigs. *PLoS One* **5:** e9068.

Brown IH, Banks J, Manvell RJ, Essen SC, Shell W, Slomka M, Londt B, Alexander DJ (2006). Recent epidemiology and ecology of influenza A viruses in avian species in Europe and the Middle East. *Dev Biol (Basel)* **124:** 45-50.

Brown IH (2000). The epidemiology and evolution of influenza viruses in pigs. *Vet Microbiol* **74:** 29–46.

Brown IH, Harris PA, McCauley JW, Alexander DJ (1998). Multiple genetic reassortment of avian and human influenza A viruses in European pigs, resulting in the emergence of an H1N2 virus of novel genotype. *J Gen Virol* **79**: 2947–2955.

Brown IH, Ludwig S, Olsen CW, Hannoun C, Scholtissek C, Hinshaw VS, Harris PA, McCauley JW, Strong I, Alexander DJ (1997). Antigenic and genetic analyses of H1N1 influenza A viruses from European pigs. *J Gen Virol* **78:** 553–562.

Brown IH, Chakraverty P, Harris PA, Alexander DJ (1995). Disease outbreaks in pigs in Great Britain due to an influenza A virus of H1N2 subtype. *Vet Rec* **136:** 328–329.

Bui M, Whittaker G, Helenius A. (1996). Effect of M1 protein and low pH on nuclear transport of influenza virus ribonucleoproteins. *J Virol* **70**: 8391–8401.

Bullough P, Hughson FM, Skehel JJ, Wiley DC. (1994). The structure of influenza hemagglutinin at the pH of membrane fusion. *Nature* **371**: 37-43.

Burch, J., Corbett, M., Stock, C., Nicholson, K., Elliot, A.J., Duffy, S., Westwood, M., Palmer, S., Stewart, L. (2009). Prescription of anti-influenza drugs for healthy adults: a systematic review and meta-analysis. *Lancet Infect Dis* **9**, 537–545.

Burleigh LM, Calder LJ, Skehel JJ, Steinhauer DA (2005). Influenza A viruses with mutations in the M1 helix six domain display a wide variety of morphological phenotypes. *J Virol* **79:** 1262-1270.

Burmeister WP, Ruigrok RW, Cusack S. (1992). The 2.2 A resolution crystal structure of influenza B neuraminidase and its complex with sialic acid. *EMBO J* **11**: 49–56.

Burnet, F. M., Stone, J.D. (1947). The receptor-destroying enzyme of V. cholerae. *Aust J Exp Biol Med* **25**, 227-233.

Bush RM, Fitch WM, Bender CA, Cox NJ (1999a). Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol* **16** (11):1457–1465.

Bush, R.M., Bender, C.A., Subbarao, K., Cox, N.J., Fitch, W.M. (1999b). Predicting the evolution of human influenza A. *Science* **286**, 1921–1925.

Bussetti AV, Palacios G, Travassos da Rosa A, Savji N, Jain K, Guzman H, Hutchinson S, Popov VL, Tesh RB, Lipkin WI (2012). Genomic and antigenic characterization of Jos virus. *J Gen Virol* **93(Pt 2):** 293-298.

Butte AJ (2008). Translational bioinformatics: coming of age. *J Am Med Inform Assoc* **15:** 709–714.

Campitelli L, Donatelli I, Foni E, Castrucci MR, Fabiani C, Kawaoka Y, Krauss S, Webster RG (1997). Continued evolution of H1N1 and H3N2 influenza viruses in pigs in Italy. *Virology* **232:** 310–318.

Casalegno, J. S., Bouscambert-Duchamp, M., Caro, V., Schuffenecker, I., Sabatier, M., Traversier, A., Valette, M., Lina, B., Ferraris, O., Escuret, V. (2010). Oseltamivir-resistant influenza A(H1N1) viruses in south of France, 2007/2009. *Antiviral Res* **87**, 242-248.

Castrucci MR, Campitelli L, Ruggieri A, Barigazzi G, Sidoli L, Daniels R, Oxford JS, Donatelli I (1994). Antigenic and sequence analysis of H3 influenza virus haemagglutinins from pigs in Italy. *J Gen Virol* **7:** 371–379.

Castrucci MR, Donatelli I, Sidoli L, Barigazzi G, Kawaoka Y,
Webster RG (1993). Genetic reassortment between avian and human influenza A viruses in Italian pigs. *Virology* **193**: 503–506.

CDC (2014a). Key facts about human infections with variant viruses (swine origin influenza viruses in humans). http://www.cdc.gov/flu/swineflu/keyfacts-variant.htm [cited on 20 Sep 2014]

CDC (2014b). Highly pathogenic avian influenza A (H5N1) in people. http://www.cdc.gov/flu/avianflu/h5n1-people.htm [cited on 21 Sep 2014]

CDC (2012a). Summary of notifiable diseases – United States, 2010. Morbidity and mortality weekly report 59: 1-111. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5953a1.htm [cited on 20 Sep 2014]

CDC (2012b). Evaluation of rapid influenza diagnostic tests for influenza A (H3N2)v virus and updated case count – United States, 2012. *MMWR Morb Mortal Wkly Rep* **61:** 619–621.

CDC (2008). Avian influenza A virus infections of humans. http://www.cdc.gov/flu/avian/gen-info/avian-flu-humans.htm [cited on 21 Sep 2014]

Chambers TM, Hinshaw VS, Kawaoka Y, Easterday BC, Webster RG (1991). Influenza viral infection of swine in the United States 1988–1989. *Arch Virol* **116:** 261–265.

Chan MC, Chan RW, Yu WC, Ho CC, Chui WH, Lo CK, Yuen KM, Guan YI, Nicholls JM, Peiris JS. (2009). Influenza H5N1 virus infection of polarized human alveolar epithelial cells and lung microvascular endothelial cells. *Respir Res* **10**: 102.

Chen GW, Shih SR. (2009). Genomic signatures of influenza A pandemic (H1N1) 2009 virus. *Emerg Infect Dis* **15**: 1897-1903.

Chen C, Zhuang X. (2008). Epsin 1 is a cargo-specific adaptor for the clathrin-mediated endocytosis of the influenza virus. *Proc Natl Acad Sci USA* **105:** 11790–11795.

Chen YP, Chen F (2008). Identifying targets for drug discovery using bioinformatics. *Expert Opin Ther Targets* **12:** 383–389.

Chen GW, Chang SC, Mok CK, Lo YL, Kung YN, Huang JH, Shih YH, Wang JY, Chiang C, Chen CJ, Shih SR. (2006). Genomic signatures of human versus avian influenza A viruses. *Emerg Infect Dis* **12**: 1353-1360.

Chen BJ, Takeda M, Lamb RA (2005). Influenza virus hemagglutinin (H3 subtype) requires palmitoylation of its cytoplasmic tail for assembly: M1 proteins of two subtypes differ in their ability to support assembly. *J Virol* **79:** 13673-13684.

Chen W, Calvo PA, Malide D, Gibbs J, Schubert U, Bacik I, Basta S, O'Neill R, Schickli J, Palese P, Henklein P, Bennink JR, Yewdell JW. (2001). A novel influenza A virus mitochondrial protein that induces cell death. *Nat Med* **7:** 1306-1312.

Chen J, Lee KH, Steinhauer DA, Stevens DJ, Skehel JJ, Wiley DC. (1998). Structure of the hemagglutinin precursor cleavage site, a determinant of influenza pathogenicity and the origin of the labile conformation. *Cell* **95**: 409–417.

Cheung CL, Rayner JM, Smith GJD, Wang P, Naipospos TSP, Zhang J, Yuen KY, Webster RG, Peiris JSM, Guan Y, Chen H (2006). Distribution of amantadine-resistant H5N1 avian influenza variants in Asia. *J Infect Dis* **193:** 1626-1629.

Choi C, Ha SK, Chae C (2004). Detection and isolation of H1N1 influenza virus from pigs in Korea. *Vet Rec* **154:** 274–275.

Choi YK, Goyal SM, Farnham MW, Joo HS (2002b). Phylogenetic analysis of H1N2 isolates of influenza A virus from pigs in the United States. *Virus Res* **87:** 173–179.

Choi YK, Kim SM, Kim HS et al (2002a). Isolation and genetic characterization of H1N2 subtype of influenza A virus from pigs in South Korea. Direct submission. GenBank Accession numbers: AY129156-AY129163.

Chong AKJ, Pegg MS, von Itzstein M. (1991). Influenza virus sialidase: effect of calcium on steady-state kinetic parameters. *Biochim Biophys Acta* **1077**: 65-71.

Chu VC, Whittaker GR. (2004). Influenza virus entry and infection require host cell n-linked glycoprotein. *Proc Natl Acad Sci USA* **101**: 18153–18158.

Chun JWH (1919). Influenza, including its infection among pigs. *Natl Med J China* **5:** 34–44.

Chutinimitkul S, Thippamom N, Damrongwatanapokin S, Payungporn S, Thanawongnuwech R, Amonsin A, Boonsuk P, Sreta D, Bunpong N, Tantilertcharoen R, Chamnanpood P, Parchariyanon S, Theamboonlers A, Poovorawan Y (2008). Genetic characterization of H1N1, H1N2 and H3N2 swine influenza virus in Thailand. *Arch Virol* **153:** 1049–1056.

Clark NM, Lynch JP 3rd. (2011). Influenza: epidemiology, clinical features, therapy, and prevention. *Semin Respir Crit Care Med* **32**: 373-392.

Collins PJ, Haire LF, Lin YP, Liu J, Russell RJ, Walker PA, Skehel JJ, Martin SR, Hay AJ, Gamblin SJ (2008). Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants. *Nature* **453:** 1258-1261.

Colman PM, Varghese JN, Laver WG (1983). Structure of the catalytic and antigenic sites in influenza virus neuraminidase. *Nature* **303:** 41-44.

Colman PM, Varghese JN, Laver WG. (1983). Structure of the catalytic and antigenic sites in influenza virus neuraminidase. *Nature* **303**: 41–44.

Compans RW, Content J, Duesberg PH (1972). Structure of the ribonucleoprotein of influenza virus. *J Virol* **10:** 795-800.

Connor RJ, Kawaoka Y, Webster RG, Paulson JC (1994). Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates. *Virology* **205**: 17–23.

Copeland C, Zimmer K, Wagner K, Healey G, Mellman I, Helenius Al. (1988). Folding, trimerization, and transport are sequential events in the biogenesis of influenza hemagglutinin. *Cell* **53**: 197–209.

Cox NJ, Subbarao K. (2000). Global epidemiology of influenza: past and present. *Ann Rev Med* **51**: 407-421.

Crawford PC, Dubovi EJ, Castleman WL, Stephenson I, Gibbs EP, Chen L, Smith C, Hill RC, Ferro P, Pompey J, Bright RA, Medina MJ, Johnson CM, Olsen CW, Cox NJ, Klimov AI, Katz JM, Donis RO (2005). Transmission of equine influenza virus to dogs. *Science* **310:** 482-485.

Cunha BA (2004). Influenza: historical aspects of epidemics and pandemics. *Infect Dis Clin N Am* **18:** 141-155.

da Fontoura Costa L (2004). Bioinformatics: perspectives for the future. *Genet Mol Res* **3:** 564–574.

Daly JM, Lai AC, Binns MM, Chambers TM, Barrandeguy M, Mumford JA (1996). Antigenic and genetic evolution of equine H3N8 influenza A viruses. *J Gen Virol* **77(Pt 4):** 661-671.

Das KP, Mallick BB, Das K (1981). Note on the prevalence of influenza antibodies in swine. *Ind J Anim Sci* **51:** 907–908.

Dauber, B., G. Heins, Wolff T. (2004). The influenza B virus nonstructural NS1 protein is essential for efficient viral growth and antagonizes beta interferon induction. *J Virol* **78**: 1865-1872.

David A. Steinhauer (1999). Role of Hemagglutinin Cleavage for the Pathogenicity of Influenza Virus. *Virology* **258**: 1-20.

Davies WL, Grunert RR, Haff RF, McGahen JW, Neumayer EM, Paulshock M, Watts JC, Wood TR, Hermann EC, Hoffmann CE (1964). Antiviral activity of 1-adamantanamine (Amantadine). *Science* **144** (3620): 862-863.

Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng PY, Bandaranayake D, Breiman RF, Brooks WA, Buchy P, Feikin DR, Fowler KB, Gordon A, Hien NT, Horby P, Huang QS, Katz MA, Krishnan A, Lal R, Montgomery JM, Mølbak K, Pebody R, Presanis AM, Razuri H, Steens A, Tinoco YO, Wallinga J, Yu H, Vong S, Bresee J, Widdowson MA (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *Lancet Infect Dis* **12:** 687-695.

Dawood FS, Jain S, Finelli L, Shaw MW, Lindstrom S, Garten RJ, Gubareva LV, Xu X, Bridges CB, Uyeki TM (2009). Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N Engl J Med* **360:** 2605–2615.

De Conto F, Covan S, Arcangeletti MC, Orlandini G, Gatti R, Dettori G, Chezzi C. (2011). Differential infectious entry of human influenza A/WSN/33 virus (H1N1) in mammalian kidney cells. *Virus Res* **155**: 221–230.

de Jong JC, Smith DJ, Lapedes AS, Donatelli I, Campitelli L, Barigazzi G, Van Reeth K, Jones TC, Rimmelzwaan GF & other authors (2007). Antigenic and genetic evolution of swine influenza A (H3N2) viruses in Europe. *J Virol* **81:** 4315–4322.

De Jong JC, van Nieuwstadt AP, Kimman TG, Loeffen WLA, Bestebroer TM, Bijlsma K, Verweij C, Osterhaus ADME, Claas ECJ (1999). Antigenic drift in swine influenza H3 haemagglutinins with implications for vaccination policy. *Vaccine* **17:** 1321-1328.

de Lorenzo V, Serrano L, Valencia A (2006). Synthetic biology: challenges ahead. *Bioinformatics* **22:** 127–128.

De Pristo, M.A., Weinreich, D.M., Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* **6**, 678–687.

Desselberger U, Racaniello VR, Zazra JJ, Palese P (1980). The 3’ and 5’-terminal sequences of influenza A, B and C virus RNA segments are highly conserved and show partial inverted complementarity. *Gene* **8:** 315-328.

de Visser, J.A.G.M., Elena, S.F. (2007). The evolution of sex: empirical insights into the roles of epistasis and drift. *Nat Rev Genet* **8**, 139–149.

De Vries E, Tscherne DM, Wienholts MJ, Cobos-Jimenez V, Scholte F, Garcia-Sastre A, Rottier PJ, de Haan CA. (2011). Dissection of the influenza a virus endocytic routes reveals macropinocytosis as an alternative entry pathway. *PLoS Pathog* **7**: e1001329.

Deyde VM, Xu X, Bright RA, Shaw M, Smith CB, Zhang Y, Shu Y, Gubareva LV, Cox NJ, Klimov AI (2007). Surveillance of resistance to adamantanes among influenza A(H3N2) and A(H1N1) viruses isolated worldwide. J Infect Dis **196** (2): 249-257.

Dhar R, Chanock RM, Lai CJ. (1980). Nonviral oligonucleotides at the 5' terminus of cytoplasmic influenza viral mRNA deduced from cloned complete genomic sequences. *Cell* **21**: 495-500.

Dharan, N. J., Gubareva, L.V., Meyer, J.J., Okomo-Adhiambo, M., McClinton, R.C., Marshall, S.A., St George, K., Epperson, S., Brammer, L. & other authors (2009). Infections with oseltamivir-resistant influenza A(H1N1) virus in the United States. *JAMA* **301**, 1034-1041.

Dias A, Bouvier D, Crepin T, McCarthy AA, Hart DJ, Baudin F, Cusack S, Ruigrok RW (2009). The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature* **458:** 914-918.

Di Trani L, Bedini B, Cordioli P, Muscillo M, Vignolo E, Moreno A, Tollis M (2004). Molecular characterization of low pathogenicity H7N3 avian influenza viruses isolated in Italy. *Avian Dis* **48**: 376-383.

Dolin R, Reichman RC, Madore HP, Maynard R, Linton PN, Webber-Jones J (1982). A controlled trial of amantadine and rimantadine in the prophylaxis of influenza A infection. *N Engl J Med* **307** (10): 580-584.

dos Reis M, Hay AJ, Goldstein RA (2009). Using non-homogeneous models of nucleotide substitution to identify host shift events: application to the origin of the 1918 'Spanish' influenza pandemic virus. *J Mol Evol* **69**: 333–345.

Drummond AJ and Rambaut A (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**: 24.

Dunham EJ, Dugan VG, Kaser EK, Perkins SE, Brown IH, Holmes EC, Taubenberger JK (2009). Different evolutionary trajectories of European avian-like and classical swine H1N1 influenza A viruses. *J Virol* **83**: 5485–5494.

Easterday BC, Hinshaw VS (1992). Swine influenza. Leman AD, Straw BE, Mengeling WL, D'Allaire SD, D.J. Taylor DJ Jr (Eds.), Diseases of Swine, Iowa State Press, Ames, IA, pp. 349–357.

Eddelbuettel D, Francois R (2011). Rcpp: seamless R and C++ integration. *J Stat Software* **40**.

Edwards SV (2009). Is a new and general theory of molecular systematics emerging? *Evolution* **63**: 1–19.

Edwards AWF, Cavalli-Sforza LL (1963). The reconstruction of evolution. *Ann Human Genet* **27:** 105-106.

Efron B, Halloran E, Holmes S (1996). Bootstrap confidence levels for phylogenetic trees. *PNAS* **93**: 13429-13434.

Efron B and Gong G (1983). A leisurely look at the bootstrap, the jackknife, and cross validation. *American Statistician* **37:** 36-48.

Efron B (1979). Bootstrap methods: Another look at the jackknife. *Ann Statistics* **7:** 1-26.

Ehrhardt C, Marjuki H, Wolff T, Nurnberg B, Planz O, Pleschka S, Ludwig S. (2006). Bivalent role of the phosphatidylinositol-3-kinase (PI3K) during influenza virus infection and host cell defence. *Cell Microbiol* **8**: 1336–1348.

Eierhoff T, Hrincius ER, Rescher U, Ludwig S, Ehrhardt C. (2010). The epidermal growth factor receptor (EGFR) promotes uptake of influenza a viruses (IAV) into host cells. *PLoS Pathog* **6**: e1001099.

Els MC, Air GM, Murti KG, Webster RG, Laver WG. (1985). An 18-amino acid deletion in an influenza neuraminidase. *Virology* **142**: 241–247.

Elton D, Bryant N. (2011). Facing the threat of equine influenza. *Equine Vet J* **43**: 250-258.

Enserink M, Kaiser J (2004). Virology. Avian flu finds new mammal hosts. *Science* **305:** 1385.

Faith DP, Baker AM (2007). Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evol Bioinform Online* **2:** 121-128.

Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (Eds.) (2005). Orthomyxoviridae. Virus Taxonomy: VIIIth Report of the International Committee on Taxonomy of Viruses. London, Elsevier Academic Press.

Felsenstein J (2004). Inferring Phylogenies, chapters 9, 11, 13. Sunderland, MA: Sinauer Associates.

Felsenstein J (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39:** 783-791.

Felsenstein J (1981a). A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biol J Linn Soc Lond* **16:** 183-196.

Felsenstein J (1981b). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* **17:** 368-376.

Fisher RA (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41:** 155-160.

Forrest HL, Webster RG (2010). Perspectives on influenza evolution and the role of research. *Anim Health Res Rev* **11:** 3-18.

Fouchier RA, Munster V, Wallensten A, Bestebroer TM, Herfst S, Smith D, Rimmelzwaan GF, Olsen B, Osterhaus AD. (2005). Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. *J Virol* **79**: 2814-2822.

Fouchier RA, Schneeberger PM, Rozendaal FW, Broekman JM, Kemink SA, Munster V, Kuiken T, Rimmelzwaan GF, Schutten M, Van Doornum GJ, Koch G, Bosman A, Koopmans M, Osterhaus AD. (2004). Avian influenza A virus (H7N7) associated with human conjunctivitis and a fatal case of acute respiratory distress syndrome. *Proc Nat Acad Sci USA* **101**: 1356-1361.

Fox J, Andersen R (2005). Using the R statistical computing environment to teach social statistics courses. Department of Sociology, McMaster University. [cited on 03 Aug 2006]

Furuse Y, Suzuki A, Oshitani H (2010). Reassortment between swine influenza A viruses increased their adaptation to humans in pandemic H1N1/09. *Infect Genet Evol* **10:** 569-574.

Furuse Y, Suzuki A, Oshitani H (2009). Large-scale analysis of M gene of influenza A viruses from different species: mechanisms for emergence and spread of amantadine resistance. *Antimicrob Agents Chemother* **53:** 4457-4463.

Garcia-Sastre A. (2001). Inhibition of interferon-mediated antiviral responses by influenza A viruses and other negative-strand RNA viruses. *Virology* **279**: 375-384.

Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A,
Sessions WM, Xu X, Skepner E & other authors (2009). Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* **325:** 197–201.

Garten W, Klenk HD. (2008). Cleavage Activation of the Influenza Virus Hemagglutinin and Its Role in Pathogenesis. In *Avian Influenza*; Klenk HD, Matrosovich MN, Stech J, Eds, Karger: Basel, Switzerland.

Gatenby RA, Frieden BR (2007). Information theory in living systems, methods, applications, and challenges. *Bull Math Biol* **69:** 635–657.

Gensheimer KF, Fukuda K, Brammer L, Cox N, Patriarca PA (1999). Preparing for pandemic influenza: the need for enhanced surveillance. *Emerg Infect Dis* **5:** 297-299.

Gerhard W, Webster RG (1978). Antigenic drift in influenza A viruses. I. Selection and characterization of antigenic variants of A/PR/8/34 (HON1) influenza virus with monoclonal antibodies. *J Exp Med* **148:** 383-392.

Gething MJ, Doms RW, York D, White J. (1986). Studies on the mechanism of membrane fusion: site-specific mutagenesis of the hemagglutinin of influenza virus. *J Cell Biol* **102:** 11-23.

Geyer CJ (1991). Markov chain Monte Carlo maximum likelihood. In Keramidas (ed.), Computing science and statistics: Proceedings of the 23rd symposium on the interface. Interface foundation, Fairfax station, pp156-163.

Ghedin, E., Holmes, E.C., DePasse, J.V., Pinilla, L.T., Fitch, A., Hamelin, M.E., Papenburg, J., Boivin, G. (2012). Presence of oseltamivir-resistant pandemic A/H1N1 minor variants before drug therapy with subsequent selection and transmission. *J Infect Dis* **206**, 1504-1511.

Gibbs SEJ. (2010). Avian biology, the human influence on global avian influenza transmission, and performing surveillance in wild birds. *Anim Health Res Rev* **11**: 35-41.

Gotlieb TA, Ivanov IE, Adesnik M, Sabatini DD (1993). Actin microfilaments play a critical role in endocytosis at the apical but not the basolateral surface of polarized epithelial cells. *J Cell Biol* **120**: 695-710.

Gourreau JM, Kaiser C, Valette M, Douglas AR, Labie J, Aymard M (1994). Isolation of two H1N2 influenza viruses from swine in France. *Arch Virol* **135:** 365–382.

Gov.UK (2014). Notifiable diseases in animals. Department for environment, food and rural affairs and Animal health and veterinary laboratory agency. https://www.gov.uk/government/collections/notifiable-diseases-in-animals [cited on 20 Sep 2014]

Grambas S, Bennett MS, Hay AJ (1992). Influence of amantadine resistance mutations on the pH regulatory function of the M2 protein of influenza A viruses. *Virology* **191:** 541-549.

Gray RD, Drummond AJ, Greenhill SJ (2009). Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science* **323:** 479-483.

Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303:** 327-332.

Gu RX, Liu LA, Wang YH, Xu Q, Wei DQ (2013). Structural comparison of the wild-type and drug-resistant mutants of the influenza A M2 proton channel by molecular dynamics simulations. *J Phys Chem B* **117**: 6042-6051.

Guan Y, Shortridge KF, Krauss S, Li PH, Kawaoka Y, Webster RG (1996). Emergence of avian H1N1 influenza viruses in pigs in China. *J Virol* **70:** 8041–8046.

Gubareva, L.V. (2004). Molecular mechanisms of influenza virus resistance to neuraminidase inhibitors. *Virus Research* **103**, 199-203

Gubareva, L. V., Robinson, M.J., Bethell, R.C., Webster, R.G. (1997). Catalytic and framework mutations in the neuraminidase active site of influenza viruses that are resistant to 4-guanidino-Neu5Ac2en. *J Virol* **71**, 3385-3390.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML3.0. *Syst Biol* **59:** 307-321.

Gurwitz D, Lunshof JE, Altman RB (2006). A call for the creation of personalized medicine databases. *Nat Rev Drug Discov* **5:** 23–26.

Haesebrouck F, Pensaert M (1988). Influenza in swine in Belgium (1969–1986): epizootiologic aspects. *Comp Immunol Microbiol Infect Dis* **11:** 215–222.

Haesebrouck F, Biront P, Pensaert MB, Leunen J (1985). Epizootics of respiratory tract disease in swine in Belgium due to H3N2 influenza virus and experimental reproduction of disease. *Am J Vet Res* **46:**1926–1928.

Hanauer DA, Rhodes DR, Sinha-Kumar C, Chinnaiyan AM (2007). Bioinformatics approaches in the study of cancer. *Curr Mol Med* **7:** 133–141.

Hatzimanikatis V (2000). Bioinformatics and functional genomics: challenges and opportunities. *AIChE J* **46:** 2340–2343.

Hauge SH, Dudman S, et al. (2009). Oseltamivir-resistant influenza viruses A (H1N1), Norway, 2007-08. *Emerg Infect Dis* **15** (2): 155-162.

Hay AJ, Gregory V, Douglas AR, Lin YP (2001). The evolution of human influenza viruses. *Philos Trans R Soc Lond B Biol Sci* **356:** 1861-1870.

Hay AJ, Wolstenholme AJ, Skehel JJ, Smith MH (1985). The molecular basis of the specific anti-influenza action of amantadine. *EMBO J* **4:** 3021-3024.

Hay AJ, Zambon MC, Wolstenholme AJ, Shekel JJ, Smith MH. (1986). Molecular basis of resistance of influenza A viruses to amantadine. *J Antimicrob Chemother* **18**: 19-29.

Hebert D, Foellmer B, Helenius A. (1995). Glucose trimming and reglucosylation determine glycoprotein association with calnexin in the endoplasmic reticulum. *Cell* **81**: 425–33.

Hinshaw VS, Bean WJ, Geraci J, Fiorelli P, Early G, Webster RG (1986). Characterization of two influenza A viruses from a pilot whale. *J Virol* **58:** 655-656.

Hinshaw VS, Bean WJ, Webster RG, Rehg JE, Fiorelli P, Early G, Geraci JR, St Aubin DJ (1984). Are seals frequently infected with avian influenza viruses? *J Virol* **51:** 863-865.

Hirst, G. K. (1941). Agglutination of red cells by allantoic fluid of chick embryos infected with influenza virus. *Science* **94**, 22-23.

Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, St George K, Grenfell BT, Salzberg SL, Fraser CM, Lipman DJ, Taubenberger JK (2005). Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol* **3:** e300.

Holsinger LJ, Nichani D, Pinto LH, Lamb RA (1994). Influenza A virus M2 ion channel protein: a structure-function analysis. *J Virol* **68**: 1551-1563.

Holsinger LJ, Lamb RA (1991). Influenza virus M2 integral membrane protein is a homotetramer stabilised by formation of disulfide bonds. *Virology* **183**: 32-43.

Hood E (2006). Flu vaccine production gets a shot in the arm. *Environ Health Perspect* **114:** A108-A111.

Hordijk W, Gascuel O (2005). Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics* **21:** 4338-4347.

Horimoto T, Kawaoka Y (2005). Influenza: lessons from past pandemics, warnings from current incidents. *Nat Rev Microbiol* **3:** 591-600.

Hota S. and McGeer A (2007). Antivirals and the control of influenza outbreaks. *Clin Infect Dis* **45** (10): 1362-1368.

Howden KJ, Brockhoff EJ, Caya FD, McLeod LJ, Lavoie M, Ing JD, Bystrom JM, Alexandersen S, Pasick JM, Berhane Y, Morrison ME, Keenliside JM, Laurendeau S, Rohonczy EB (2009). An investigation into human pandemic influenza virus (H1N1) 2009 on an Alberta swine farm. *Can Vet J* **50:** 1153–1161.

Huang RTC, Wahn K, Klenk HD, Rott R. (1980). Fusion between cell membranes and liposomes containing the glycoprotein of influenza virus. *Virology* **104**: 294-302.

Huelsenbeck JP, Dyer KA (2004). Bayesian estimation of positively selected sites. *J Mol Evol* **58:** 661–672.

Huelsenbeck JP, Ronquist F (2001). MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17:** 754-755.

Hughson FM. (1995). Structural characterization of viral fusion proteins. *Curr Biol* **5**: 265-274.

Hurt, A.C., Chotpitayasunondh, T., Cox, N.J., Daniels, R., Fry, A.M., Gubareva, L.V., Hayden, F.G., Hui, D.S., Hungnes, O. & other authors (2012). Antiviral resistance during the 2009 influenza A H1N1 pandemic public health, laboratory, and clinical perspectives. *Lancet Infect Dis* **12**, 240-248.

Hurt, A. C., Ernest, J., Deng, Y.M., Iannello, P., Besselaar, T.G., Birch, C., Buchy, P., Chittaganpitch, M., Chiu, S.C. & other authors (2009). Emergence and spread of oseltamivir-resistant A(H1N1) influenza viruses in Oceania, South East Asia and South Africa. *Antiviral Res* **83**, 90-93.

Hutchinson EC, von Kirchbach JC, Gog JR, Digard P (2010). Genome packaging in influenza A virus. *J Gen Virol* **91:** 313-328.

Ilyin SE, Bernal A, Horowitz D, Derian CK, Xin H (2004). Functional informatics: convergence and integration of automation and bioinformatics. *Pharmacogenomics* **5:** 721–730.

Ilyushina NA, Govorkova EA, Russell CJ, Hoffmann E, Webster RG (2007). Contribution of H7 haemagglutinin to amantadine resistance and infectivity of influenza virus. *J Gen Virol* **88:** 1266-1274.

Inkster MD, Hinshaw VS, Schulze IT. (1993). The hemagglutinins of duck and human H1 influenza viruses differ in sequence conservation and in glycosylation. *J Virol* **67**: 7436-7443.

Ito T, Couceiro JN, Kelm S, Baum LG, Krauss S, Castrucci MR, Donatelli I, Kida H, Paulson JC, Webster RG, Kawaoka Y. (1998a). Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. *J Virol* **72**: 7367-7373.

Ito T, Kawaoka Y, Vines A, Ishikawa H, Asai T, Kida H (1998b). Continued circulation of reassortant H1N2 influenza viruses in pigs in Japan. *Arch Virol* **143:** 1773–1782.

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 (2012). An overlapping protein-coding region in influenza A virus segment 3 modulates the host response. *Science* **337:** 199-204.

Jin H, Leser GP, Zhang J, Lamb RA. (1997). Influenza virus hemagglutinin and neuraminidase cytoplasmic tails control particle shape. *EMBO J* **16**:1236–1247.

Kalthoff D, Globig A, Beer M (2010). (Highly pathogenic) avian influenza as a zoonotic agent. *Vet Microbiol* **140:** 237-245.

Karasin AI, Carman S, Olsen CW (2006). Identification of human H1N2 and human-swine reassortant H1N2 and H1N1 influenza A viruses among pigs in Ontario, Canada (2003 to 2005). *J Clin Microbiol* **44:** 1123–1126.

Karasin AI, Landgraf J, Swenson S, Erickson G, Goyal SM, Woodruff M, Scherba G, Anderson GA, Olsen CW (2002). Genetic characterization of H1N2 influenza A viruses isolated from pigs throughout the United States. *J Clin Microbiol* **40:**1073-1079.

Karasin AI, Schutten MM, Cooper LA, Smith CB, Subbarao K, Anderson GA, Carman S, Olsen CW (2000a). Genetic characterization of H3N2 influenza viruses isolated from pigs in North America, 1977–1999: evidence for wholly human and reassortant virus genotypes. *Virus Res* **68:** 71–85.

Karasin AI, G.A. Anderson GA, Olsen CW (2000b). Genetic characterization of an H1N2 influenza virus isolated from a pig in Indiana. *J Clin Microbiol* **38:** 2453–2456.

Katsuda K, Sato S, Shirahata T, Lindstrom S, Nerome R, Ishida M, Nerome K, Goto H (1995). Antigenic and genetic characteristics of H1N1 human influenza virus isolated from pigs in Japan. *J Gen Virol* **76:** 1247–1249.

Kaverin NV, Matrosovich MN, Gambaryan AS, Rudneva IA, Shilov AA (2000). Intergenic HA-NA interactions in influenza A virus: Postreassortment substitutions of charged amino acid in the hemagglutinin of different subtypes. *Virus Res* **66:** 123-129.

Kawaoka Y, Krauss S, Webster RG. (1989). Avian-to-human transmission of the PB1 gene of influenza A virus in the 1957 and 1968 pandemics. *J Virol* **63**: 4603-4608.

Kelso, A., Hurt, A.C. (2012). The ongoing battle against influenza: Drug-resistant influenza viruses: why fitness matters. *Nature Med* **18**, 1470-1471.

Kemler I, Whittaker G, Helenius A. (1994). Nuclear import of microinjected influenza virus ribonucleoproteins. *Virology* **202**: 1028–1033.

Khiabanian H, Trifonov V, Rabadan R (2009). Reassortment patterns in Swine influenza viruses. *PLoS One* **4:** e7366.

Kishino H, Thorne J, Bruno W (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol Biol Evol* **18:** 352-361.

Kiso M, Mitamura K, et al. (2004). Resistant influenza A viruses in children treated with oseltamivir: descriptive study. *Lancet* **364** (9436): 759-765.

Kitching IJ, Forey PL, Humphries CJ, Williams DM (1998). Cladistics: the theory and practice of parsimony analysis, p 17. Second Edition. The Systematics Association Publication No. 11. Oxford: Oxford University Press.

Klenk HD, Garten W (1994). Host cell proteases controlling virus pathogenicity. *Trends Microbiol* **2:** 39-43.

Klenk HD, Rott R, Orlich M, Blodorn J. (1975). Activation of influenza A viruses by trypsin treatment. *Virology* **68**: 426-439.

Klingeborn B, Englund L, Rott R, Juntti N, Rockborn G (1985). An avian influenza A virus killing a mammalian species – the mink. Brief report. *Arch Virol* **86:** 347-351.

Kochs G, Garcia-Sastre A, Martínez-Sobrido L. (2007). Multiple anti-interferon actions of the influenza A virus NS1 protein. *J Virol* **81**: 7011-7021.

Koen JS (1919). A practical method for field diagnosis of swine diseases. *Am J Vet Med* **14:** 468–470.

Kosakovsky Pond SL, Poon AFY, Frost SDW (2007). Estimating selection pressures on alignments of coding sequences analyses using HyPhy. http://www.hyphy.org/pubs/hyphybook2007.pdf [cited 14th May 2013]

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006). GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22:** 3096-3098.

Kosakovsky Pond SL, Frost SWD (2005a). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* **21:** 2531-2533.

Kosakovsky Pond SL (2005b). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21:** 676-679.

Krallinger M, Valencia A, Hirschman L (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol* **9:** Suppl 2S8.

Kramarz, P., Monnet, D., Nicoll, A., Yilmaz, C., Ciancio, B. (2009). Use of oseltamivir in 12 European countries between 2002 and 2007 – lack of association with the appearance of oseltamivir-resistant influenza A(H1N1) viruses. *Euro Surveill* **14**, pii=19112.

Krogh A, Larsson B, von Heijne G, Sonnhammer E. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol* **305**: 567–580.

Krug RM, Broni BA, Bouloy M. (1979). Are the 5' ends of influenza viral mRNAs synthesized in vivo donated by host mRNAs? *Cell* **18**: 329-334.

Krug RM, Morgan MA, Shatkin AJ. (1976). Influenza viral mRNA contains internal N6-methyladenosine and 5'-terminal 7-methylguanosine in cap structures. *J Virol* **20**: 45-53.

Krumbholz A, Schmidtke M, Bergmann S, Motzke S, Bauer K, Stech J, Durrwald R, Wutzler P, Zell R (2009). High prevalence of amantadine resistance among circulating European porcine influenza A viruses. *J Gen Virol* **90:** 900-908.

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

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

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

Kuchipudi SV, Nelli R, White GA, Bain M, Chang KC, Dunham S. (2009). Differences in influenza virus receptors in chickens and ducks: implications for interspecies transmission. *J Mol Genet Med* **3**: 143-151.

Kundin WD (1970). Hong Kong A2 influenza virus infection among swine during a human epidemic in Taiwan. *Nature* **228:** 857.

Kupradinun S, Peanpijit P, Bhodhikosoom C, Yoshioka Y, Endo A, Nerome K (1991). The first isolation of swine H1N1 influenza viruses from pigs in Thailand. *Arch Virol* **118:** 289–297.

Kyriakis CS, Brown IH, Foni E, Kuntz-Simon G, Maldonado J, Madec F, Essen SC, Chiapponi C, Van Reeth K (2011). Virological surveillance and preliminary antigenic characterization of influenza viruses in pigs in five European countries from 2006 to 2008. *Zoonoses Public Health* **58:** 93–101.

Kyrpides NC (2009). Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat Biotechnol* **27:** 627–632.

Lackenby A, Thompson CI, Democratis J (2008a). The potential impact of neuraminidase inhibitor resistant influenza. *Curr Opin Infect Dis* **21:** 626-638.

Lackenby, A., Hungnes, O., Dudman, S.G., Meijer, A., Paget, W.J., Hay, A.J., Zambon, M.C. (2008b). Emergence of resistance to oseltamivir among influenza A(H1N1) viruses in Europe. *Euro Surveill* **13**, pii=8026.

Lai AC, Chambers TM, Holland RE Jr, Morley PS, Haines DM, Townsend HG, Barrandeguy M (2001). Diverged evolution of recent equine-2 influenza (H3N8) viruses in the Western Hemisphere. *Arch Virol* **146:** 1063-1074.

Lakadamyali M, Rust MJ, Zhuang X. (2006). Ligands for clathrin-mediated endocytosis are differentially sorted into distinct populations of early endosomes. *Cell* **124**: 997–1009.

Lakadamyali M, Rust MJ, Zhuang X. (2004). Endocytosis of influenza viruses. *Microbes Infect* **6**: 929–936.

Lakadamyali M, Rust MJ, Babcock HP, Zhuang X. (2003). Visualizing infection of individual influenza viruses. *Proc Natl Acad Sci USA* **100**: 9280–9285.

Lakdawala SS, Lamirande EW, Suguitan AL Jr, Wang W, Santos CP, Vogel L, Matsuoka Y, Lindsley WG, Jin H, Subbarao K (2011). Eurasian-origin gene segments contribute to the transmissibility, aerosol release, and morphology of the 2009 pandemic H1N1 influenza virus. *PLoS Pathog* **7:** e1002443.

Lamb RA, Holsinger LJ, Pinto LH (1994). Receptor-mediated virus entry into cells. Wimmer E, editor. Cold Spring Harbor Laboratory Press; Cold Spring Harbor, NY.

Lamb RA, Zebedee SL, Richardson CD (1985). Influenza virus M2 protein is an integral membrane protein expressed on the infected-cell surface. *Cell* **40**: 627-633.

Lamb RA, Lai CJ, Choppin PW. (1981). Sequences of mRNAs derived from genome RNA segment 7 of influenza virus: colinear and interrupted mRNAs code for overlapping proteins. *Proc Natl Acad Sci U S A* **78**: 4170-4174.

Lamb RA, Choppin PW, Chanock RM, Lai CJ. (1980). Mapping of the two overlapping genes for polypeptides NS1 and NS2 on RNA segment 8 of influenza virus genome. Proc Natl Acad Sci U S A **77**: 1857-1861.

Lange E, Kalthoff D, Blohm U, Teifke JP, Breithaup tA, Maresch C, Starick E, Fereidouni S, Hoffmann B, Mettenleiter TC, Beer M, Vahlenkamp TW (2009). Pathogenesis and transmission of the novel swine-origin influenza virus A/H1N1 after experimental infection of pigs. *J Gen Virol* **90:** 2119–2123.

Lazarowitz SG, Choppin PW. (1975). Enhancement of the infectivity of influenza A and B viruses by proteolytic cleavage of the hemagglutinin polypeptide. *Virology* **68**: 440-454.

Lee PJ and Krilov LR. (2009). Swine influenza viruses and their pandemic potential. *Ped Ann* **38**:12.

Lee J, Song YJ, Park JH, Lee JH, Baek YH, Song MS, Oh TK, Han HS, Pascua PNQ, Choi YK (2008). Emergence of amantadine-resistant H3N2 avian influenza A virus in South Korea. *J Clin Microbiol* **46:** 3788-3790.

Lee SM, Abbott PA (2003). Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers. *J Biomed Inform* **36:** 389-399.

Lewis RJ (2000). An introduction to Classification and Regression Tree (CART) Analysis. Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California.

Li J, zu Dohna H, Anchell NL, Adams SC, Dao NT, Xing Z, Cardona CJ. (2010). Adaptation and transmission of a duck-origin avian influenza virus in poultry species. *Virus Res* **147**: 40-46.

Li ML, Rao P, Krug RM. (2001). The active sites of the influenza cap-dependent endonuclease are on different polymerase subunits. *EMBO J* **20**: 2078-2086.

Li W-H (1997). Molecular Evolution, p 128, chapters 3, 4, 5. Sunderland, MA: Sinauer Associates.

Liaw A, Wiener M (2002). Classification and regression by randomForest. R News 2: 18-22.

Lindstrom S, Garten R, Balish A, Shu B, Emery S, Berman L, Barnes N,
Sleeman K, Gubareva L, Villanueva J, Klimov A (2012). Human infections with novel reassortant influenza A(H3N2)v viruses, United States, 2011. *Emerg Infect Dis* **18:** 834–837.

Liu J, Bi Y, Qin K, Fu G, Yang J, Peng J, Ma G, Liu Q, Pu J, Tian F (2009). Emergence of European avian influenza virus-like H1N1 swine influenza A viruses in China. *J Clin Microbiol* **47:** 2643–2646.

Liu J, Okazaki K, Ozaki H, Sakoda Y, Wu Q, Chen F, Kida H. (2003). H9N2 influenza viruses prevalent in poultry in China are phylogenetically distinct from A/quail/Hong Kong/G1/97 presumed to be the donor of the internal protein genes of the H5N1 Hong Kong/97 virus. *Avian Pathol* **32**: 551-560.

Liu C, Eichelberger MC, Compans RW, Air GM. (1995). Influenza type A virus neuraminidase does not play a role in viral entry, replication, assembly, or budding. *J Virol* **69**: 1099-1106.

Lorusso A, Vincent AL, Harland ML, Alt D, Bayles DO, Swenson SL, Gramer MR, Russell CA, Smith DJ, Lager KM, Lewis NS (2011). Genetic and antigenic characterization of H1 influenza viruses from United States swine from 2008. *J Gen Virol* **92:** 919–930.

Lu L, Lycett SJ, Leigh Brown AJ (2014). Reassortment patterns of avian influenza virus internal segments among different subtypes. *BMC Evol Biol* **14:** 16.

Lu B, Zhou H, Ye D, Kemble G, Jin H. (2005). Improvement of influenza A/Fujian/411/02 (H3N2) virus growth in embryonated chicken eggs by balancing the hemagglutinin and neuraminidase activities, using reverse genetics. *J Virol* **79**: 6763-6771.

Lycett SJ, Baillie G, Coulter E, Bhatt S, Kellam P, McCauley JW, Wood JL, Brown IH, Pybus OG, Leigh Brown AJ; Combating Swine Influenza Initiative-COSI Consortium (2012). Estimating reassortment rates in co-circulating Eurasian swine influenza viruses. *J Gen Virol* **93:** 2326-2336.

Ma J (2011). Reconstructing the history of large-scale genomic changes: biological questions and computational challenges. *J Comput Biol* **18:** 879-893.

Ma W, Vincent AL, Gramer MR, Brockwell CB, Lager KM, Janke BH, Gauger PC, Patnayak DP, Webby RJ, Richt JA (2007). Identification of H2N3 influenza A viruses from swine in the United States. *Proc Natl Acad Sci USA* **104:** 20949–20954.

Maeda T, and Ohnishi S. (1980). Activation of influenza virus by acidic media causes hemolysis and fusion of erythrocytes. *FEBS Lett* **122**: 283-287.

Maltezou H.C (2008). Nosocomial influenza: new concepts and practice. *Curr Opin Infect Dis* **21** (4): 337-343.

Mancini G, Donatelli I, Rozera C, Ruiz GA, Butto S (1985). Antigenic and biochemical analysis of influenza A H3N2 viruses isolated from pigs. *Arch Virol* **83:** 157–167.

Maojo V, Martin-Sanchez F (2004). Bioinformatics: towards new directions for public health. *Methods Inf Med* **43:** 208–214.

Marozin S, Gregory V, Cameron K, Bennett M, Valette M, Aymard M, Foni E, Barigazzi G, Lin Y, Hay A (2002). Antigenic and genetic diversity among swine influenza A H1N1 and H1N2 viruses in Europe. *J Gen Virol* **83:** 735–745.

Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, Khattra J, Asano JK, Barber SA, Chan SY, Cloutier A, Coughlin SM, Freeman D, Girn N, Griffith OL, Leach SR, Mayo M, McDonald H, Montgomery SB, Pandoh PK, Petrescu AS, Robertson AG, Schein JE, Siddiqui A, Smailus DE, Stott JM, Yang GS,Plummer F, Andonov A, Artsob H, Bastien N, Bernard K, Booth TF, Bowness D, Czub M, Drebot M, Fernando L, Flick R, Garbutt M, Gray M, Grolla A, Jones S,Feldmann H, Meyers A, Kabani A, Li Y, Normand S, Stroher U, Tipples GA, Tyler S, Vogrig R, Ward D, Watson B, Brunham RC, Krajden M, Petric M, Skowronski DM, Upton C, Roper RL (2003). The Genome sequence of the SARS-associated coronavirus. *Science* **300:** 1399-1404.

Maser P, Thomine S, Schroeder JI, Ward JM, Hirschi K, Sze H, Talke IN, Amtmann A, Maathuis FJ, Sanders D, Harper JF, Tchieu J, Gribskov M, Persans MW, Salt DE, Kim SA, Guerinot ML (2001). Phylogenetic relationships within cation transporter families of Arabidopsis. *Plant Physiol* **26:** 1646-1667.

Mateo, R., Mateu, M.G. (2007). Deterministic, compensatory mutational events in the capsid of foot-and-mouth disease virus in response to the introduction of mutations found in viruses from persistent infections. *J Virol* **81**, 1879–1887.

Matrosovich MN, Klenk HD, Kawaoka Y. (2006). Influenza Virology, Current Topics: receptor specificity, host range, and pathogenicity of influenza viruses. Norfolk, England: Caister Academic Press.

Matrosovich MN, Matrosovich TY, Gray T, Roberts NA, Klenk HD (2004a). Human and avian influenza viruses target different cell types in cultures of human airway epithelium. *Proc Natl Acad Sci USA* **101**: 4620-4624.

Matrosovich MN, Matrosovich TY, Gray T, Roberts NA, Klenk HD. (2004b). Neuraminidase is important for the initiation of influenza virus infection in human airway epithelium. *J Virol* **78**: 12665-12667.

Matrosovich M, Tuzikov A, Bovin N, Gambaryan A, Klimov A, Castrucci MR, Donatelli I, Kawaoka Y. (2000). Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. *J Virol* **74**: 8502-8512.

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

McCown MF, Pekosz (2006). Distinct domains of the influenza A virus M2 protein cytoplasmic tail mediate binding to the M1 protein and facilitate infectious virus production. *J Virol* **80:** 8178-8189.

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

McKimm-Breschkin JL, Sahasrabudhe A, Blick TJ, McDonald M, Colman PM, Hart GJ, Bethell RC, Varghese JN. (1998). Mutations in a conserved residue in the influenza virus neuraminidase active site decreases sensitivity to Neu5Ac2en-derived inhibitors. *J Virol* **72**: 2456-2462.

McKimm-Breschkin JL, Blick TJ, Sahasrabudhe A, Tiong T, Marshall D, Hart GJ, Bethell RC, Penn CR. (1996). Generation and characterization of variants of NWS/G70C influenza virus after in vitro passage in 4-amino-Neu5Ac2en and 4-guanidino-Neu5Ac2en. *Antimicrob Agents Chemother* **40**: 40-46.

Meijer, A., Lackenby, A., Hungnes, O., Lina, B., van-der-Werf, S., Schweiger, B., Opp, M., Paget, J., van-de-Kassteele, J. & other authors (2009). Oseltamivir-resistant influenza virus A (H1N1), Europe, 2007-08 season. *Emerg Infect Dis* **15**, 552-560.

Mitnaul LJ, Matrosovich MN, Castrucci MR, Tuzikov AB, Bovin NV (2000). Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza A virus. *J Virol* **74:** 6015-6020.

Mitnaul LJ, Castrucci MR, Murti KG, Kawaoka Y (1996). The cytoplasmic tail of influenza A virus neuraminidase (NA) affects NA incorporation into virions, virion morphology, and virulence in mice but is not essential for virus replication. *J Virol* **70:** 873-879.

Moller S, Croning M, Apweiler R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**: 646–653.

Moscona A (2009). Global transmission of oseltamivir-resistant influenza. *N Engl J Med* **360:** 953-956.

Moscona A (2005). Neuraminidase inhibitors for influenza. *N Engl J Med* **353:** 1363-1373.

Munier S, Larcher T, Cormier-Aline F, Soubieux D, Su B, Guigand L, Labrosse B, Cherel Y, Quéré P, Marc D, Naffakh N. (2010). A genetically engineered waterfowl influenza virus with a deletion in the stalk of the neuraminidase has increased virulence for chickens. *J Virol* **84**: 940–952.

Munster VJ, Fouchier RA (2009). Avian influenza virus: of virus and bird ecology. *Vaccine* **27**: 6340-6344.

Muramoto Y, Noda T, Kawakami E, Akkina R, Kawaoka Y (2013). Identification of novel influenza A virus proteins translated from PA mRNA. *J Virol* **87:** 2455-2462.

Murphy BR, Webster RG (1996). Orthomyxoviruses. Fields BN, Knipe DM, Howley PM, Chanock RM, Melnick JL, Monath TP, Roizman B, Straus SE (Eds.), Field's Virology, Lippincott-Raven Publishers, Philadelphia, PA, pp. 1397–1445.

Murphy BR, Hinshaw VS, Sly DL, London WT, Hosier NT, Wood FT, Webster RG, Chanock RM. (1982). Virulence of avian influenza A viruses for squirrel monkeys. *Infect Immun* **37**: 1119-1126.

Murray, R.J., Lewis, F.I., Miller, M.D., Leigh Brown, A.J. (2008). Genetic basis of variation in tenofovir drug susceptibility in HIV-1. *AIDS* **22**, 1113-1123.

Murti KG, Webster RG, Jones IM (1988). Localization of RNA polymerases on influenza viral ribonucleoproteins by immunogold labeling. *Virology* **164:** 562-566.

Myers KP, Olsen CW, Gray GC (2007). Cases of swine influenza in humans: a review of the literature. *Clin Infect Dis* **44:** 1084-1088.

Naffakh N, Tomoiu A, Rameix-Welti MA, van der Werf S. (2008). Host restriction of avian influenza viruses at the level of the ribonucleoproteins. *Ann Rev Microbiol* **62**: 403-424.

Nagarajan N, Kingsford C (2011). GiRaF: robust, computational identification of influenza reassortments via graph mining. *Nucleic Acids Res* **39:** e34.

Nagarajan N, Kingsford C (2008). Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine. Washington DC: IEEE Computer Society; 2008. Uncovering genomic reassortments among Influenza strains by enumerating maximal bicliques.

Nakajima K, Nakajima S, Shortridge KF, Kendal AP (1982). Further genetic evidence for maintenance of early Hong Kong-like influenza A (H3N2) strains in swine until 1976. *Virology* **116:** 562–572.

Nakajima K, Desselberger U, Palese P (1978). Recent human influenza A (H1N1) viruses are closely related genetically to strains isolated in 1950. *Nature* **274:** 334-339.

Nayak D, Hui E, Barman S (2004). Assembly and budding of influenza virus. *Virus Res* **106:** 147-165.

Nei M, Kumar S (2000). Molecular Evolution and Phylogenetics, chapters 2, 3, 6. New York, NY: Oxford University Press.

Nelli RK, Kuchipudi SV, White GA, Perez BB, Dunham SP, Chang KC. (2010). Comparative distribution of human and avian type sialic acid influenza receptors in the pig. *BMC Vet Res* **6**: 4.

Nerome K, Kanegae Y, Shortridge KF et al (1995). Genetic analysis of porcine H3N2 viruses originating in southern China. *J Gen Virol* **76:** 613–624.

Nerome K, Yoshioka Y, Sakamoto S, Yasuhara M, Oya A (1985). Characterization of a 1980-swine recombinant influenza virus possessing H1 haemagglutinin and N2 neuraminidase similar to that of the earliest Hong Kong (H3N2) virus. *Arch Virol* **86:** 197–211.

Nerome K, Ishida M, Oya A, Kanai C, Suwicha K (1982). Isolation of an influenza H1N1 virus from a pig. *Virology* **117:** 485–489.

Nicholls JM, Chan MC, Chan WY, Wong HK, Cheung CY, Kwong DL, Wong MP, Chui WH, Poon LL, Tsao SW, Guan Y, Peiris JS. (2007). Tropism of avian influenza A (H5N1) in the upper and lower respiratory tract. *Nature Med* **13**: 147-149.

Nielsen R (2002a). Mapping mutations on phylogenies. *Syst Biol **51*:** *729*–739.

Nielsen R, Huelsenbeck J (2002b). Detecting positively selected amino acid sites using posterior predictive p-values. *Pac Symp Biocomput **7*:** *576*–588.

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

Nielsen R (1997). Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA. *Syst Biol* **46:** 346–353.

Nobusawa E, Ishihara H, Morishita T, Sato K, Nakajima K (2000). Change in receptor-binding specificity of recent human influenza A viruses (H3N2): a single amino acid change in hemagglutinin altered its recognition of sialyloligosaccharides. *Virology* **278:** 587-596.

O'Donnell CJ (2000). Cardiovascular genomics: recent progress, current challenges, future promise. *Genome Biol* **1:** 401-409.

Office International des Epizooties (2005). Avian influenza. In Manual of diagnostic tests and vaccines for terrestrial animals, 5th edition. Office International des Epizooties, Paris, France.

Okamatsu M, Saito T, Yamamoto Y, Mase M, Tsuduku S, Nakamura K, Tsukamoto K, Yamaguchi S. (2007). Low pathogenicity H5N2 avian influenza outbreak in Japan during the 2005-2006. *Vet Microbiol* **124**: 35-46.

Okazaki K, Yanagawa R, Kida H, Noda H (1983). Human influenza virus infection in mink: serological evidence of infection in summer and autumn. *Vet Microbiol* **8:** 251-257.

Oloffson S, Kumlin U, Dimock K, Arnberg N. (2005). Avian influenza and sialic acid receptors: more than meets the eye? *Lancet Infect Dis* **5**: 184-188.

Olsen CW, Karasin AI, Carman S, Li Y, Bastien N, Ojkic D, Alves D, Charbonneau G, Henning BM, Low DE, Burton L, Broukhanski G (2006a). Triple reassortant H3N2 influenza A viruses, Canada, 2005. *Emerg Infect Dis* **12:** 1132–1135.

Olsen B, Munster VJ, Wallensten A, Waldenström J, Osterhaus AD, Fouchier RA. (2006b). Global patterns of influenza a virus in wild birds. *Science* **312**: 384-388.

Olsen CW (2002). The emergence of novel swine influenza viruses in North America. *Virus Res* **85:** 199–210.

Olsen CW, Carey S, Hinshaw L, Karasin AI (2000). Virologic and serologic surveillance for human, swine and avian influenza virus infections among pigs in the north-central United States. *Arch Virol* **145:** 1399–1419.

Ottis K, Sidoli L, Bachmann PA, Webster RG, Kaplan MM (1982). Human influenza A viruses in pigs: isolation of a H3N2 strain antigenically related to A/England/42/72 and evidence for continuous circulation of human viruses in the pig population. *Arch Virol* **73:** 103–108.

Ouchi A, Nerome K, Kanegae Y, Ishida M, Nerome R, Hayashi K, Hashimoto T, Kaji M, Kaji Y, Inaba Y (1996). Large outbreak of swine influenza in southern Japan caused by reassortant (H1N2) influenza viruses: its epizootic background and characterization of the causative viruses. *J Gen Virol* **77:** 1751–1759.

Ouzounis CA (2012). Rise and demise of bioinformatics? Promise and progress. *PLoS Comput Biol* **8:** e1002487.

Ouzounis CA, Valencia A (2003). Early bioinformatics: the birth of a discipline–a personal view. *Bioinformatics* **19:** 2176–2190.

Palese P, Compans RW (1976). Inhibition of influenza virus replication in tissue culture by 2-deoxy-2,3-dehydro-N-trifluoroacetylneuraminic acid (FANA): mechanism of action. *J Gen Virol* **33**: 159-163.

Palese P, Tobita K, Ueda M, Compans RW. (1974). Characterization of temperature sensitive influenza virus mutants defective in neuraminidase. *Virology* **61**: 397-410.

Palsson B (2000). The challenges of in silico biology. *Nat Biotechnol* **18:** 1147–1150.

Paradis E (2009). dist.topo.R. Topological distances, tree bipartitions, consensus trees, and bootstrapping phylogenies.
http://www.math.ncu.edu.tw/~chenwc/R/bin/ape/R/dist.topo.R [cited on 06 Aug 2014]

Paradis E, Claude J, Strimmer K (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20:** 289-290.

Pascua PN, Song MS, Lee JH et al (2008). Seroprevalence and genetic evolutions of swine influenza viruses under vaccination pressure in Korean swine herds. Virus Res **138:** 43–49.

Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, Birney E (2008). Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res* **18:** 1829-1843.

Pattengale ND, Alipour M, Bininda-Emonds OR, Moret BM, Stamatakis A (2010). How many bootstrap replicates are necessary? *J Comput Biol* **17:** 337-354.

Patterson KD, Pyle GF (1991). The geography and mortality of the 1918 influenza pandemic. *Bull Hist Med* **65:** 4-21.

Peiris JS, Poon LL, Guan Y (2009). Emergence of a novel swine-origin influenza A virus (S-OIV) H1N1 virus in humans. *J Clin Virol* **45:** 169–173.

Penny D, Hendy MD (1985). The use of tree comparison metrics. *Syst Zool* **34:** 75-82.

Pensaert M, Ottis K, Vandeputte J, Kaplan MM, Bachmann PA (1981). Evidence for the natural transmission of influenza A virus from wild ducts to swine and its potential importance for man. *Bull World Health Organ* **59:** 75–78.

Perdue ML, Garcia M, Senne D, Fraire M. (1997). Virulence associated sequence duplication at the hemagglutinin cleavage site of avian influenza viruses. *Virus Res* **49**: 173-186.

Perdue ML, Latimer JW, Crawford JM. (1995). A novel carbohydrate addition site on the hemagglutinin protein of a highly pathogenic H7 subtype avian influenza virus. *Virology* **213**: 276-281.

Pillai SP, Suarez DL, Pantin-Jackwood M, Lee CW. (2009). The high susceptibility of turkeys to low pathogenic avian influenza viruses of different origins imply their importance as intermediate host. In: 7th International symposium on avian influenza: avian influenza in poultry and wild birds.

Pinto LH, Lamb RA. (2006). The M2 proton channels of influenza A and B viruses. *J Biol Chem* **281**: 8997-9000.

Pinto LH, Dieckmann GR, Gandhi CS, Papworth CG, Braman J, Shaughnessy MA, Lear JD, Lamb RA, DeGrado WF (1997a). A functionally defined model for the M2 proton channel of influenza A virus suggests a mechanism for its ion selectivity. *Proc Natl Acad Sci USA* **94**: 11301-11306.

Pinto LH, et al (1997b). A functionally defined model for the M2 proton channel of influenza A virus suggests a mechanism for its ion selectivity. *Proc Natl Acad Sci USA* **94**: 11301-11306.

Pinto LH, Holsinger LJ, Lamb RA (1992). Influenza virus M2 protein has ion channel activity. *Cell* **69**: 517-528.

Plotch SJ, Bouloy M, Ulmanen I, Krug RM. (1981). A unique cap(m7GpppXm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription. *Cell* **23**: 847-858.

Plotch SJ, Bouloy M, Krug RM. (1979). Transfer of 5'-terminal cap of globin mRNA to influenza viral complementary RNA during transcription in vitro. *Proc Natl Acad Sci U S A* **76**: 1618-1622.

Plotch SJ, Tomasz J, Krug RMl. (1978). Absence of detectable capping and methylating enzymes in influenza virions. *J Virol* **28**: 75-83.

Poon, A.F., Lewis, F.I., Pond, S.D. (2007). An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Comput Biol* **3**, e231.

Pop M (2009). Genome assembly reborn: recent computational challenges. *Brief Bioinform* **10:** 354–366.

Porter AG, Barber C, Carey NH, Hallewell RA, Threlfall G, Emtage JS. (1979). Complete nucleotide sequence of an influenza virus haemagglutinin gene from cloned DNA. *Nature* **282**: 471-477.

Posada D, Crandall KA (2002). The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol* **54:** 396-402.

Precious, H.M., Gunthard, H.F., Wong, J.K., D'Aquila, R.T., Johnson, V.A., Kuritzkes, D.R., Richman, D.D., Leigh Brown, A.J. (2000). Multiple sites in HIV-1 reverse transcriptase associated with virological response to combination therapy. *AIDS* **14**, 31-36.

Pritchard GC, Dick IGC, Roberts DH, Wibberley G (1987). Porcine influenza outbreak in East Anglia due to influenza A virus (H3N2). *Vet Rec* **121:** 548.

Qi X, Pang B, Lu CP (2009). Genetic characterization of H1N1 swine influenza A viruses isolated in eastern China. *Virus Genes* **39:** 193–199.

Rabadan R, Levin AJ, Krasnitz M (2008). Non-random reassortment in human influenza A viruses. Influenza Other Respir Viruses 2: 9–22.

Racaniello V (2009). Structure of influenza virus. http://www.virology.ws/2009/04/30/structure-of-influenza-virus/ [cited on 10 Jun 2011]

Rambaut A (2013). How to read a phylogenetic tree. Epidemic: Molecular epidemiology and evolution of viral pathogens. http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny [cited on 21 Jun 2014]

Rambaut A (2007a). FigTree. Molecular evolution, phylogenetics and epidemiology. http://tree.bio.ed.ac.uk/software/figtree/ [cited on 06 Aug 2014]

Rambaut A (2007b). Molecular evolution, phylogenetics and epidemiology. FigTree. (http://tree.bio.ed.ac.uk/software/figtree).

Rambaut A, Bromham L (1998). Estimating divergence dates from molecular sequences. *Mol Biol Evol* **15:** 442-448.

Rameix-Welti MA, Enouf V, Cuvelier F, Jeannin P, van der Werf S (2008). Enzymatic properties of the neuraminidase of seasonal H1N1 influenza viruses provide insights for the emergence of natural resistance to oseltamivir. *PLoS Pathog* **4:** e1000103.

Reed J (2000). Trends in commercial bioinformatics. Oscar Gruss.

Regoes RR, Bonhoeffer S. (2006). Emergence of drug-resistant influenza virus: population dynamical considerations. *Science* **312**: 389–391.

Reid AH, Taubenberger JK (2003). The origin of the 1918 pandemic influenza virus: a continuing enigma. *J Gen Virol* **84(Pt 9):** 2285-2292.

Richard, M., Erny, A., Caré, B., Traversier, A., Barthélémy, M., Hay, A., Lin, Y.P., Ferraris, O., Lina, B. (2012). Rescue of a H3N2 influenza virus containing a deficient neuraminidase protein by a haemagglutinin with a low receptor-binding affinity. *PLoS One* 7: e33880.

Richardson JC, Akkina RK. (1991). NS2 protein of influenza virus is found in purified virus and phosphorylated in infected cells. *Arch Virol* **116**: 69-80.

Richt JA, Lager KM, Janke BH, Woods RD, Webster RG, Webby RJ (2003). Pathogenic and antigenic properties of phylogenetically distinct reassortant H3N2 swine influenza viruses cocirculating in the United States. *J Clin Microbiol* **41:** 3198–3205.

Ripley, B. (2012). http://cran.r-project.org/web/packages/tree/index.html [cited on 25 Feb 2013]

Ritchie MD (2005). Bioinformatics approaches for detecting gene-gene and gene-environment interactions in studies of human disease. *Neurosurg Focus* **19:** E2.

Robertson HD, Dickson E, Plotch SJ, Krug RM. (1980). Identification of the RNA region transferred from a representative primer, beta-globin mRNA, to influenza mRNA during in vitro transcription. *Nucleic Acids Res* **8**: 925-942.

Robertson JS (1979). 5' and 3' terminal nucleotide sequences of the RNA genome segments of influenza virus. *Nucleic Acids Res* **6:** 3745-3757.

Rogers GN, D'Souza BL (1989). Receptor binding properties of human and animal H1 influenza virus isolates. *Virology* **173**: 317–322.

Ronquist F, van der Mark P, Huelsenbeck JP (2009). Bayesian phylogenetic analysis using MrBayes: theory. In: The phylogenetic handbook. A practical approach to phylogenetic analysis and hypothesis testing. Lemey P, Salemi M, Vandamme A-M (ed.), p220-224.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61: 539-542.

Rossman JS, Lamb RA (2011). Influenza virus assembly and budding. *Virology* **411:** 229-236.

Rudneva IA, Sklyanskaya EI, Barulina OS, Yamnikova SS, Kovaleva VP, Tsvetkova IV, Kaverin NV. (1996). Phenotypic expression of HA-NA combinations in human-avian influenza A virus reassortants. *Arch Virol* **141**: 1091-1099.

Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V,
Gust ID, Hampson AW, Hay AJ & other authors (2008). The global circulation of seasonal influenza A (H3N2) viruses. *Science* **320:** 340–346.

Russell RJ, Haire LF, Stevens DJ, Collins PJ, Lin YP, Blackburn GM, Hay AJ, Gamblin SJ, Skehel JJ (2006). The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **443:** 45-49.

Rzhetsky A, Nei M (1992). A simple method for estimating and testing minimum-evolution trees. *MBE* **9:** 945-967.

Saito R, Sakai T, Sato I, Sano Y, Oshitani H, Sato M, Suzuki H (2003). Frequency of amantadine-resistant influenza A viruses during two seasons featuring cocirculation of H1N1 and H3N2. *J Clin Microbiol* **41:** 2164-2165.

Saito T, Kawano K. (1997). Loss of glycosylation at Asn144 alters the substrate preference of the N8 influenza A virus neuraminidase. *J Vet Med Sci* **59**: 923-926.

Sanderson M (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol Biol Evol* **19:** 101-109.

Sanderson M (1997). Nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol Biol Evol* **14:** 1218-1231.

Sanjuán, R., Cuevas, J.M., Moya, A., Elena, S.F. (2005). Epistasis and the adaptability of an RNA virus. *Genetics* **170**, 1001–1008.

Schierup M, Hein J (2000). Recombination and the molecular clock. *Mol Biol Evol* **17:** 1578-1579.

Schmidtke M, Zell R, Bauer K, Krumbholz A, Schrader C, Suess J, Wutzler P (2006). Amantadine resistance among porcine H1N1, H1N2, and H3N2 influenza A viruses isolated in Germany between 1981 and 2001. *Intervirology* **49**: 286-293.

Schnirring L (2009). Pandemic reveals strengths of new flu database. Avian influenza (bird flu) H1N1 2009 pandemic influenza. Center for Infectious Disease Research and Policy. University of Minnesota, Minneapolis, MN. http://www.cidrap.umn.edu/news-perspective/2009/06/pandemic-reveals-strengths-new-flu-database [cited on 18 Sep 2014]

Scholtissek C and Webster RG (1998). Long-term stability of the anti-influenza A compounds – amantadine and rimantadine. *Antiviral Res* **38**: 213-215.

Scholtissek C, Bürger H, Kistner O, Shortridge KF (1985). The nucleoprotein as a possible major factor in determining host specificity of influenza H3N2 viruses. *Virology* **147**: 287–294.

Scholtissek C, Burger H, Bachmann PA, Hannoun C (1983). Genetic relatedness of hemagglutinins of the H1 subtype of influenza A viruses isolated from swine and birds. *Virology* **129:** 521-523.

Scholtissek C, Rohde W, von Hoyningen V, Rott R. (1978). On the origin of the human influenza virus subtypes H2N2 and H3N2. *Virology* **87**: 13-20.

Schultz U, Fitch WM, Ludwig S, Mandler J, Scholtissek C (1991). Evolution of pig influenza viruses. *Virology* **183**: 61–73.

Senne DA, Panigraphy B, Kawaoka Y, Pearson JE, Suss J, Lipkind M, Kida H, Webster RG. (1996). Survey of the hemagglutinin (HA) cleavage site sequence of H5 and H7 avian influenza viruses: Amino acid sequence at the HA cleavage site as a marker of pathogenicity potential. *Avian Dis* **40**: 425-437.

Sevin, A. D., DeGruttola, V., Nijhuis, M., Schapiro, J.M., Foulkes, A.S., Para, M.F., Boucher, C.A. (2000). Methods for investigation of the relationship between drug-susceptibility phenotype and human immunodeficiency virus type 1 genotype with applications to AIDS clinical trials group 333. *J Infect Dis* **182**, 59-67.

Sha B, Luo M (1997). Structure of a bifunctional membrane-RNA binding protein, influenza virus matrix protein M1. *Nat Struct Biol* **4:** 239-244.

Shapiro B, Rambaut A, Drummond AJ (2006). Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences. *Mol Biol Evol* **23:** 7-9.

Shapiro GI, T. Gurney T Jr, Krug RM. (1987). Influenza virus gene expression: control mechanisms at early and late times of infection and nuclear-cytoplasmic transport of virus-specific RNAs. *J Virol* **61**: 764-773.

Shellenbarger P (2009). Local officials remember 1976 swine flu scare. President Ford's decision to order nationwide vaccinations. The Grand Rapids Press. http://www.mlive.com/news/grand-rapids/index.ssf/2009/04/local_officials_remember_1976.html [cited on 20 October 2014]

Sheu TG, Deyde VM, et al. (2008). Surveillance for neuraminidase inhibitor resistance among human influenza A and B viruses circulating worldwide from 2004 to 2008. *Antimicrob Agents Chemother* **52** (9): 3284-3292.

Shinya K, Watanabe S, Ito T, Kasai N, Kawaoka Y. (2007). Adaptation of an H7N7 equine influenza A virus in mice. *J Gen Virol* **88**: 547-553.

Shinya K, Hatta M, Yamada S, Takada A, Watanabe S, Halfmann P, Horimoto T, Neumann G, Kim JH, Lim W, Guan Y, Peiris M, Kiso M, Suzuki T, Suzuki Y, Kawaoka Y. (2005). Characterization of a human H5N1 influenza A virus isolated in 2003. *J Virol* **79**: 9926-9932.

Shope RE (1931). Swine influenza: III. Filtration experiments and etiology. *J Exp Med* **54:** 373-385.

Shors T (2013). Influenza viruses. In: Understanding viruses 2nd edition (Johnson M; Ed.), pp 345-397. Jones and Bartlett Learning, Burlington, Massachussetts, USA.

Shortridge KF, Webster RG (1979). Geographical distribution of swine (Hsw1N1) and Hong Kong (H3N2) influenza virus variants in pigs in Southeast Asia. *Intervirology* **11:** 9–15.

Shriner D, Nickle DC, Jensen MA, Mullins JI (2003). Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res* **81:** 115-121.

Shu LL, Lin YP, Wright SM, Shortridge KF, Webster RG (1994). Evidence for interspecies transmission and reassortment of influenza A viruses in pigs in southern China. *Virol* **202:** 825–833.

Sieczkarski SB, Brown HA, Whittaker GR. (2003a). The role of protein kinase C bii in influenza virus entry via late endosomes. *J Virol* **77**: 460–469.

Sieczkarski SB, Whittaker GR. (2003b). Differential requirements of Rab5 and Rab7 for endocytosis of influenza and other enveloped viruses. *Traffic* **4**: 333–343.

Sieczkarski SB, Whittaker GR. (2002). Influenza virus can enter and infect cells in the absence of clathrin-mediated endocytosis. *J Virol* **76**: 10455–10464.

Simonsen L, Viboud C, Grenfell BT, Dushoff J, Jennings L, Smit M, Macken C, Hata M, Gog J, Miller MA, Holmes EC (2007). The genesis and spread of reassortment human influenza A/H3N2 viruses conferring adamantine resistance. *Mol Biol Evol* **24:** 1811-1820.

Simpson RJ, Bernhard OK, Greening DW, Moritz RL (2008). Proteomics-driven cancer biomarker discovery: looking to the future. *Curr Opin Chem Biol* **12:** 72–77.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940-3941.

Skehel JJ and Wiley DC (2000). Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu Rev Biochem* **69**: 531-569.

Skehel JJ, Bizebard T, Bullough PA, Hughson FM, Knossow M, Steinhauer DA, Wharton SA, Wiley DC. (1995). Membrane fusion by influenza hemagglutinin. In Cold Spring Harbor Symposia on Quantitative Biology, Volume LX, pp. 573-580. Cold Spring Harbor Laboratory Press.

Skehel JJ, Hay AJ (1978). Nucleotide sequences at the 5’ termini of influenza virus RNAs and their transcripts. *Nucleic Acids Res* **5:** 1207-1219.

Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghwani J, Bhatt S, Peiris JS, Guan Y, Rambaut A (2009a). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**: 1122-1125.

Smith GJ, Bahl J, Vijaykrishna D, Zhang J, Poon LL, Chen H, Webster RG, Peiris JS, Guan Y (2009b). Dating the emergence of pandemic influenza viruses. *Proc Natl Acad Sci USA* **106:** 11709-11712.

Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus ADME, Fouchier Ram (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science* **305:** 371-376.

Song DS, Lee JY, Oh JS et al (2003). Isolation of H3N2 swine influenza virus in South Korea. *J Vet Diagn Invest* **15:** 30–34.

Sriwilaijaroen N, Suzuki Y (2012). Molecular basis of the structure and function of H1 hemagglutinin of influenza virus. *Proc Jpn Acad Ser B Phys Biol Sci* **88:** 226-249.

Sriwilaijaroen, N., Suzuki, Y. (2012). Molecular basis of the structure and function of H1 haemagglutinin of influenza virus. *Proc Jpn Acad Ser B Phys Biol Sci* **88**, 226-249.

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

Stamatakis A, Hoover P, Rougemont J (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Syst biology* **57:** 758–771.

Steinhauer DA (1999). Role of hemagglutinin cleavage for the pathogenicity of influenza virus. *Virology* **258:** 1-20.

Steinhauer DA (2010). Influenza A virus haemagglutinin glycoproteins. In: Influenza, Molecular Virology (Wang Q, Tao YJ, Eds.), pp. 69-108. Caister Academic Press, Norfolk, UK.

Steinhauer DA, Wharton SA. (1998). Structure and function of the haemagglutinin. In Textbook of Influenza (Nicholson KG, Webster RG, Hay AJ, Eds.), pp. 54-64. Blackwell Science Oxford Ltd., London.

Steinhauer DA, Wharton SA, Skehel JJ, Wiley DC, Hay AJ (1991). Amantadine selection of a mutant influenza virus containing an acid-stable hemagglutinin glycoprotein: evidence for virus-specific regulation of the pH of glycoprotein transport vesicles. *Prot Nat Acad Sci USA* **88:** 11525-11529.

Stevens J, Blixt O, Tumpey TM, Taubenberger JK, Paulson JC, Wilson IA. (2006). Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. *Science* **312**: 404-410.

Stienekegrober A, Vey M, Angliker H, Shaw E, Thomas G, Roberts C, Klenk HD, Garten W (1992). Influenza-virus hemagglutinin with multibasic cleavage site is activated by Furin, a subtilisin-like endoprotease. *Embo J* **11:** 2407-2414.

Stockwell TB, Lin X, Vincent AL, Gramer MR, Holmes EC (2012). Genomic reassortment of influenza A virus in North American swine, 1998-2011. *J Gen Virol* **93:** 2584-2589.

Storms, A.D., Gubareva, L.V., Su, S., Wheeling, J.T., Okomo-Adhiambo, M., Pan, C.Y., Reisdorf, E., St George, K., Myers, R. & other authors (2012). Oseltamivir-resistant pandemic (H1N1) 2009 virus infections, United States, 2010-11. *Emerg Infect Dis* **18**, 308-311.

Stray S, Cummings RD, Air GM. (2000). Influenza virus infection of desialylated cells. *Glycobiology* **10**: 649–658.

Streiner DL, Norman GR (2011). Correction for multiple testing: is there a resolution? *Chest* **140:** 16-18.

Strimmer K, von Haeseler A (2009). Genetic distances and nucleotide substitution models: theory. In: The phylogenetic handbook. A practical approach to phylogenetic analysis and hypothesis testing. Lemey P, Salemi M, Vandamme A-M (ed.), p121-125.

Su G, Morris JH, Demchak B, Bader GD (2014). Biological network exploration with cytoscape 3. *Curr Protoc Bioinformatics* **47:** 8.13.1-8.13.24.

Suarez M, Jaramillo A (2009). Challenges in the computational design of proteins. *J R Soc Interface* **6:** Suppl 4S477–491.

Suenaga E, Kumar PK (2014). An aptamer that binds efficiently to the hemagglutinins of highly pathogenic avian influenza viruses (H5N1 and H7N7) and inhibits hemagglutinin-glycan interactions. *Acta Biomater* **10:** 1314-1323.

Sugimura T, Yonemochi H, Ogawa T, Tanaka Y, Kumagai T (1980). Isolation of a recombinant influenza virus (Hsw1N2) from swine in Japan. *Arch Virol* **66:** 271–274.

Sugrue RJ, Hay AJ (1991). Structural characteristics of the M2 protein of influenza A viruses: evidence that it forms a tetrameric channel. *Virology* **180**: 617-624.

Suzuki, Y. (2006). Natural selection on the influenza virus genome. *Mol Biol Evol* **23:** 1902–1911.

Suzuki T, Takahashi T, Guo CT, Hidari KI, Miyamoto D, Goto H, Kawaoka Y, Suzuki Y (2005). Sialidase activity of influenza A virus in an endocytic pathway enhances viral replication. *J Virol* **79:** 11705-11715.

Suzuki Y (2004a). False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom Thalassiosira weissflogii and the tax gene of a human T-cell lymphotropic virus. *Mol Biol Evol* **21:** 914–921.

Suzuki Y (2004b). New methods for detecting positive selection at single amino acid sites. *J Mol Evol* **59:** 11–19.

Suzuki Y, Ito T, Suzuki T, Holland RE Jr, Chambers TM, Kiso M, Ishida H, Kawaoka Y (2000). Sialic acid species as a determinant of the host range of influenza A viruses. *J Virol* **74**: 11825-11831.

Suzuki Y, Gojobori T (1999). A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* **16**: 1315-1328.

Swayne DE, Senne DA, Beard CW (1998). Avian influenza, p. 150-155. In Swayne DE (ed.), A laboratory manual for the isolation and identification of avian pathogens. American Association of Avian Pathologists. Kennett Square, Pa.

Tai, C. Y., Escarpe, P.A., Sidwell. R.W., Williams, M.A., Lew, W., Wu, H., Kim, C.U., Mendel, D.B. (1998). Characterization of human influenza virus variants selected in vitro in the presence of the neuraminidase inhibitor GS 4071. *Antimicrob Agents Chemother* **42**, 3234-3241.

Takemae N, Parchariyanon S, Damrongwatanapokin S , Uchida Y, Ruttanapumma R, Watanabe C, Yamaguchi S, Saito T (2008). Genetic diversity of swine influenza viruses isolated from pigs during 2000 to 2005 in Thailand. *Influenza Other Respir Viruses* **2:** 181–189.

Tamura K, Dudley J, Nei M & Kumar S (2007). *MEGA4*: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24:1596-1599.

Tang JW, Ngai KL, et al. (2008). Emergence of adamantane-resistant influenza A(H3N2) viruses in Hong Kong between 1997 and 2006. *J Med Virol* **80** (5): 895-901.

Tang Y, Zaitseva F, Lamb RA, Pinto LH (2002). The gate of the influenza virus M2 proton channel is formed by a single tryptophan residue. *J Biol Chem* **277**: 39880-39886.

Taubenberger JK, Kash JC (2010). Influenza virus evolution, host adaptation, and pandemic formation. *Cell Host Microbe* **7:** 440-451.

Taubenberger JK, Reid AH, Lourens RM, Wang R, Jin G and Fanning TG. (2005). Characterization of the 1918 influenza virus polymerase genes. *Nature* **437**: 889-893.

Taubenberger JK, Reid AH, Krafft AE, Bijwaard KE, Fanning TG (1997). Initial genetic characterization of the 1918 'Spanish' influenza virus. *Science* **275:** 1793-1796.

Temple Lang D (2010). Calling R from Java. Nuiton [cited on 18 Sep 2013]

Theobald D (2012). Phylogenetics primer. 29+ evidences for macroevolution. http://www.talkorigins.org/faqs/comdesc/phylo.html [cited on 21 Jun 2014]

Thorisson GA, Muilu J, Brookes AJ (2009). Genotype-phenotype databases: challenges and solutions for the post-genomic era. *Nat Rev Genet* **10:** 9–18.

Thorne J, Kishino H (2002). Divergence time and evolutionary rate estimation with multilocus data. *Syst Biol* **51:** 689-702.

Thorne J, Kishino H, Painter I (1998). Estimating the rate of molecular evolution. *Mol Biol Evol* **15:** 1647-1657.

Tian SF, Buckler-White AJ, London WT, Reck LJ, Chanock RM, Murphy BR. (1985). Nucleoprotein and membrane protein genes are associated with restriction of replication of influenza A/Mallard/NY/78 virus and its reassortants in squirrel monkey respiratory tract. *J Virol* **53**: 771-775.

Tisdale M (2009). Influenza M2 ion-channel and neuraminidase inhibitors. In: Antimicrobial drug resistance, mechanisms of drug resistance. Mayers DL (Ed.) Humana Press, p421-447.

Tobler K, Kelly ML, Pinto LH, Lamb RA (1999). Effect of cytoplasmic tail truncations on the activity of the M2 ion channel of influenza A virus. *J Virol* **73**: 9695-9701.

Tominack RL, Hayden FG (1987). Rimantadine hydrochloride and amantadine hydrochloride use in influenza A virus infections. *Infect Dis Clin North Am* **1**: 459-478.

Tompkins SM, Zhao ZS, Lo CY, Misplon JA, Liu T, Ye Z, Hogan RJ, Wu Z, Benton KA, Tumpey TM, Epstein SL (2007). Matrix protein 2 vaccination and protection against influenza viruses, including subtype H5N1. *Emerg Infect Dis* **13:** 426-435.

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. (2012). A distinct lineage of influenza A virus from bats. *Proc Natl Acad Sci U S A* **109**:4269-4274.

Tsai CP, Pan MJ (2003). New H1N2 and H3N1 influenza viruses in Taiwanese pig herds. *Vet Rec* **153:** 408.

Tsunoda A, Maasab HF, Cochran KW, Eveland WC (1965). Antiviral activity of α-methyl-1-adamantanemethylamine hydrochloride. *Antimicrob Agents Chemother* **5**: 553-560.

Tumova B, Mensik J, Stumpa A, Fedova D, Pospisil Z (1976). Serological evidence and isolation of a virus closely related to the human A/Hong Kong/68 (H3N2) strain in swine populations in Czechoslovakia in 1969–1972. *Zentralblatt fur Veterinaermedizin* **B23:** 590–603.

Van Deusen RA, Hinshaw VS, Senne DS, Pellacani D (1983). Micro neuraminidase-inhibitions assay for classification of influenza A virus neuraminidase. *Avian Dis* **27:** 745-750.

Van Reeth K, Brown IH, Pensaert M (2000). Isolations of H1N2 influenza A virus from pigs in Belgium. *Vet Rec* **146:** 588–589.

Vance A (2009). Data analysts captivated by R's power. New York Times. [cited on 28 Apr 2009]

Vandamme A-M (2009).Basic concepts of molecular evolution. In: The phylogenetic handbook. A practical approach to phylogenetic analysis and hypothesis testing. Lemey P, Salemi M, Vandamme A-M (ed.), p23-28.

Varghese JN, Laver WG, Colman PM. (1983). Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 A resolution. *Nature* **303**: 35–40.

Villegas P (1998). Viral diseases of the respiratory system. Poult Sci **77:** 1143-1145.

Vincent AL, Ma W, Lager KM, Gramer MR, Richt JA, Janke BH (2009). Characterization of a newly emerged genetic cluster of H1N1 and H1N2 swine influenza virus in the United States. *Virus Genes* **39:** 176–185.

Vincent AL, Ma W, Lager KM, Janke BH, Richt JA (2008.) Swine influenza viruses a North American perspective. *Adv Virus Res* **72:** 127–154.

Wagner R, Matrosovich M, Klenk HD (2002). Functional balance between haemagglutinin and neuraminidase in influenza virus infections. *Rev Med Virol* **12:** 159-166.

Wagner R, Wolff T, Herwig A, Pleschka S, Klenk HD. (2000). Interdependence of hemagglutinin glycosylation and neuraminidase as regulators of influenza virus growth: a study by reverse genetics. *J Virol* **74**: 6316-6323.

Wahid (2013).
http://www.oie.int/wahis/public.php?page=weekly_report_index&admin=0 [cited on 14 May 2014]

Wan H, Perez DR. (2006). Quail carry sialic acid receptors compatible with binding of avian and human influenza viruses. *Virology* **346**: 278-286.

Wang, D., Larder, B. (2003). Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks. *J Infect Dis* **188**, 653-660.

Wang, K., Jenwitheesuk, E., Samudrala, R., Mittler, J.E. (2004). Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance. *Antivir Ther* **9**, 343-352.

Webby RJ, Rossow K, Erickson G, Sims Y, Webster R (2004). Multiple lineages of antigenically and genetically diverse influenza A virus co-circulate in the United States swine population. *Virus Res* **103:** 67–73.

Webby RJ, Swenson SL, Krauss SL, Gerrish PJ, Goyal SM,
Webster RG (2000). Evolution of swine H3N2 influenza viruses in the United States. *J Virol* **74:** 8243–8251.

Weber TP and Stilianakis NI. (2008). Inactivation of influenza A viruses in the environment and modes of transmission: a critical review. *J Infect* **57**: 361-373.

Webster RG (1993). Are equine 1 influenza viruses still present in horses? *Equine Vet J* **25:** 537-538.

Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. (1992). Evolution and ecology of influenza A viruses. *Microbiol Rev* **56**: 152-179.

Webster RG, Guo YJ (1991). New influenza virus in horses. *Nature* **351:** 527.

Webster RG, Kawaoka Y, Bean WJ, Beard CW, and Brugh M (1985). Chemotherapy and vaccination: a possible strategy for the control of highly virulent influenza virus. *Journal of Virology* **55**:173-6.

Webster RG, Hinshaw VS, Bean WJ, Van Wyke KL, Geraci JR, St Aubin DJ, Petursson G (1981). Characterization of an influenza A virus from seals. *Virology* **113:** 712-724.

Weingartl HM, Berhane Y, Hisanaga T, Neufeld J, Kehler H, Emburry-Hyatt C, Hooper-McGreevy KKS, Dalman B, Bystrom J, Alexandersen S, Li Y, Pasick J (2010). Genetic and pathobiologic characterization of pandemic H1N1 2009 influenza viruses from a naturally infected swine herd. *J Virol* **84:** 2245–2256.

Weingartl HM, Albrecht RA, Lager KM, Babiuk S, Marszal P, Neufeld J, Embury-Hyatt C, Lekcharoensuk P, Tumpey TM, Garcia-Sastre A, Richt JA (2009). Experimental infection of pigs with the human 1918 pandemic influenza virus. *J Virol* **83:** 4287-4296.

White JM. (1995). Membrane fusion: The influenza paradigm. In Cold Spring Harbor Symposia on Quantitative Biology, Volume LX, pp. 581-588. Cold Spring Harbor Laboratory Press.

White JM, Matlin K, Helenius A. (1981). Cell fusion by Semliki Forest, influenza, and vesicular stomatitis viruses. *J Cell Biol* **89**: 2887-2896.

WHO (2013a). http://www.who.int/influenza/vaccines/virus/recommendations/en/index.html [cited 2013 Jun 27]

WHO (2013b). Use of antiviral drugs in poultry, a threat to their effectiveness for the treatment of human avian influenza. (http://www.who.int/foodsafety/micro/avian_antiviral/en/index.html) [cited 14th February 2013].

WHO (2009a). Influenza (seasonal). Retrieved 13 February 2010.

WHO (2009b). Global alert and response (GAR). Antiviral drugs for pandemic (H1N1) 2009: definitions and use. Retrieved 22 December 2009.

WHO (2005). Influenza vaccines. WHO Weekly Epidemiological Record **80:** 277-288.

WHO Expert Committee on Influenza (1953). *Wld Hlth Org Techn Rep Ser*, **No. 64**

Wiley DC, Skehel JJ. (1987). The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Ann Rev Biochem* **56**: 365–394.

Wise HM, Foeglein A, Sun J, Dalton RM, Patel S, Howard W, Anderson EC, Barclay WS, Digard P (2009). A complicated message: identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA. *J Virol* **83:** 8021-8031.

Wolf, Y.I., Viboud, C., Holmes, E.C., Koonin, E.V., Lipman, D.J. (2006). Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biol Direct* **1:** 34.

Wood GW, McCauley JW, Bashiruddin JB, Alexander DJ. (1993). Deduced amino acid sequences at the haemagglutinin cleavage site of avian influenza A viruses of H5 and H7 subtypes. *Arch Virol* **130**: 209-217.

Wright PF, Neumann G, Kawaoka Y (2007). Orthomyxoviruses. In: Fields Virology, 5th edition, Volume 2 (Knipe DM, Griffin DE, Lamb RA, Straus SE, Howley PM, Martin MA, Roizman B, Eds.). Lippincott Williams & Wilkins, Philadelphia, USA, pp 1711-1712.

Xu X, Zhu X, Dwek RA, Stevens J, Wilson IA (2008). Structural characterization of the 1918 influenza virus H1N1 neuraminidase. *J Virol* **82:** 10493-10501.

Xu R, Zhu X, McBride R, Nycholat CM, Yu W, Paulson JC, Wilson IA (2012). Functional balance of the haemagglutinin and neuraminidase activities accompanies the emergence of the 2009 H1N1 influenza pandemic. *J Virol* **86**, 9221-9232.

Yamada S, Suzuki Y, Suzuki T, Le MQ, Nidom CA, Sakai-Tagawa Y, Muramoto Y, Ito M, Kiso M, Horimoto T, Shinya K, Sawada T, Kiso M, Usui T, Murata T, Lin Y, Hay A, Haire LF, Stevens DJ, Russell RJ, Gamblin SJ, Skehel JJ, Kawaoka Y. (2006). Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors. *Nature* **444**: 378-382.

Yamane N, Arikawa J, Odagiri T, Kumasaka M, Ishida N (1978). Distribution of antibodies against swine and Hong Kong influenza viruses among pigs in 1977. *Tohoku J Exp Med* **126:** 199–200.

Yang JR, Lin YC, Huang YP, Su CH, Lo J, Ho YL, Yao CY, Hsu LC, Wu HS, Liu MT (2011). Reassortment and mutations associated with emergence and spread of oseltamivir-resistant seasonal influenza A/H1N1 viruses in 2005-2009. *PLoS One* **6:** e18177.

Yang Z, Swanson W (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol* **19:** 49–57.

Yang ZH, Nielsen R, Goldman N, Pedersen AKM (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.

Yassine HM, Lee CW, Gourapura R, Saif YM. (2010). Interspecies and intraspecies transmission of influenza A viruses: viral, host and environmental factors. *Anim Health Res Rev* **11**: 53-72.

Yasuhara M, Hirahara T, Nakai M, Sasaki N, Kata J, Watanabe T, Morikawa M (1983). Further isolation of a recombinant virus (H1N2), formerly (Hsw1N2), from a pig in Japan in 1980. *Microbiol Immunol* **27:** 43–50.

Yasui K, Amano Y, Minami I, Nakamura S, Akazawa Y, Uchida N (2007). Recent changes in the trends of seasonal influenza outbreaks in the Nagano Prefectural area of Japan: an oseltamivir effect? *J Infect Chemother* **13**: 429-431.

Yip TKS (1976). Serological survey on the influenza antibody status in pigs of the Takwuling pig breeding centre. *Agric Hong Kong* **1:** 446–458.

Yoder A, Yang Z (2000). Estimation of primate speciation dates using local molecular clocks. *Mol Biol Evol* **17:** 1081-1090.

Yoneyama S, Hayashi T, Kojima H, Usami Y, Kubo M, Takemae N, Uchida Y, Saito T (2009). Occurrence of a pig respiratory disease associated with swine influenza A (H1N2) virus in tochigi prefecture, Japan. *J Vet Med Sci* **72:** 481-488.

Yu H, Zhou YJ, Li GX et al (2009a). Further evidence for infection of pigs with human-like H1N1 influenza viruses in China. *Virus Res* **140:** 85–90.

Yu H, Zhang PC, Zhou YJ et al (2009b). Isolation and genetic characterization of avian-like H1N1 and novel ressortant H1N2 influenza viruses from pigs in China. *Biochem Biophys Res Commun* **386:** 278–283.

Yu H, Hua RH, Zhang Q, Liu TQ, Liu HL, Li GX, Tong GZ (2008). Genetic evolution of swine influenza A (H3N2) viruses in China from 1970 to 2006. *J Clin Microbiol* **46:** 1067–1075.

Yu H, Zhang GH, Hua RH, Zhang Q, Liu TQ, Liao M, Tong GZ (2007). Isolation and genetic analysis of human origin H1N1 and H3N2 influenza viruses from pigs in China. *Biochem Biophys Res Commun* **356:** 91–96.

Yuan P, Bartlam M, Lou Z, Chen S, Zhou J, He X, Lv Z, Ge R, Li X, Deng T, Fodor E, Rao Z, Liu Y (2009). Crystal structure of an avian influenza polymerase PA(N) reveals an endonuclease active site. *Nature* **458:** 909-913.

Zell R, Scholtissek C, Ludwig S (2013). Genetics, evolution, and the zoonotic capacity of European swine influenza viruses. In: Swine influenza (Richt JA, Webby RJ, Eds.) Springer-Verlag Berlin Heidelberg, p28-55.

Zell R, Krumbholz S, Wutzler A, Volker PH, Durrwald R (2008). Novel reassortant of swine influenza H1N2 virus in Germany. *J Gen Virol* **89:** 271–276.

Zhou NN, Senne DA, Landgraf JS, Swenson SL, Erickson G, Rossow K, Liu L, Yoon K, Krauss S, Webster RG (1999). Genetic reassortment of avian, swine, and human influenza A viruses in American pigs. *J Virol* **73:** 8851-8856.

Zhu H, Webby R, Lam TTY, Smith DK, Peiris JSM, Guan Y (2013). History of swine influenza viruses in Asia. In: Swine influenza; Richt JA, Webby RJ (Eds). Springer-Verlag Berlin Heidelberg, p57-68.

Ziegler T, Hemphill ML, Ziegler ML, et al. (1999) Low incidence of rimantadine resistance in field isolates of influenza A viruses. *J Infect Dis* **180:** 935-939.

Chapter 8

Appendices

8 Appendices

8.1 Appendix A

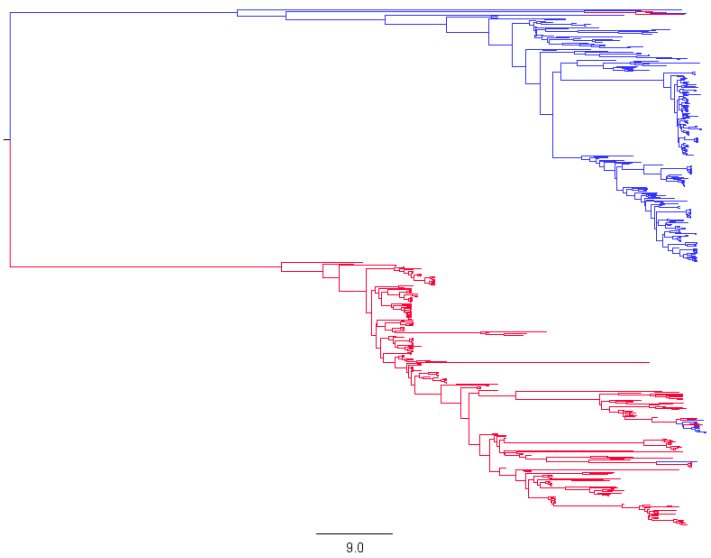

Figure A 1

The MCC tree of NA for swine H1N1, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

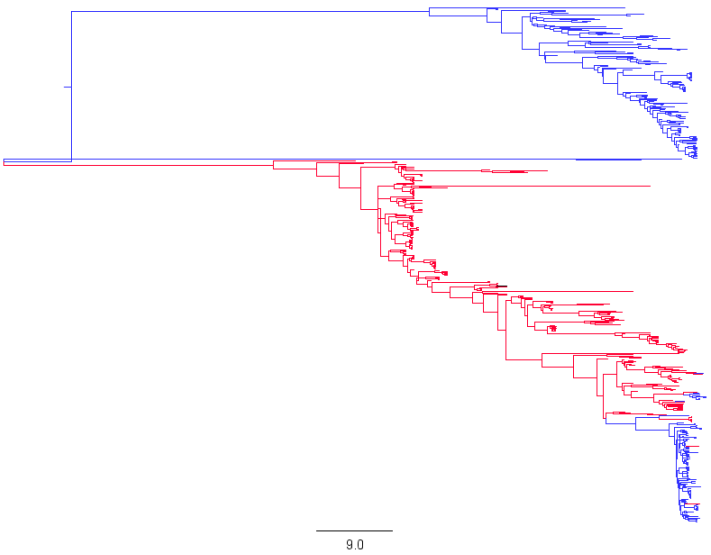

Figure A 2

The MCC tree of NP for swine H1N1, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

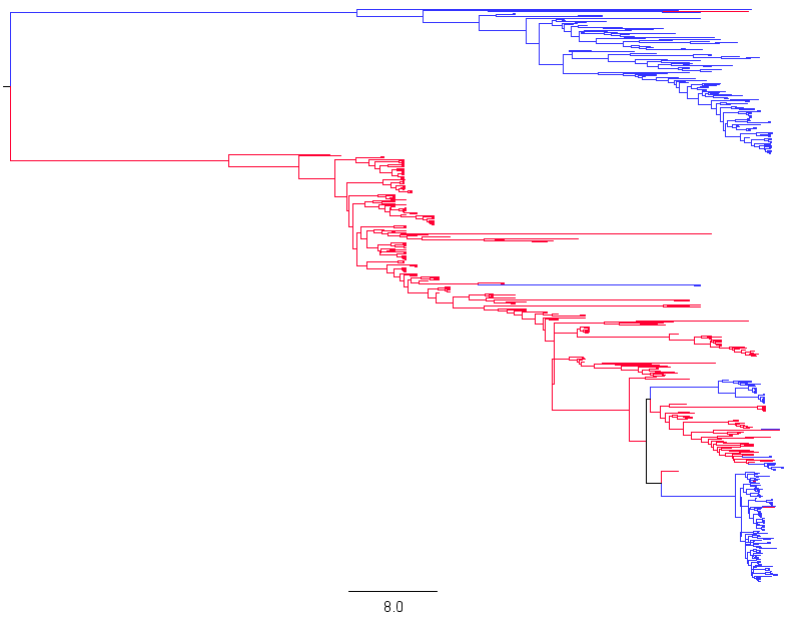

Figure A 3

The MCC tree of NS for swine H1N1, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

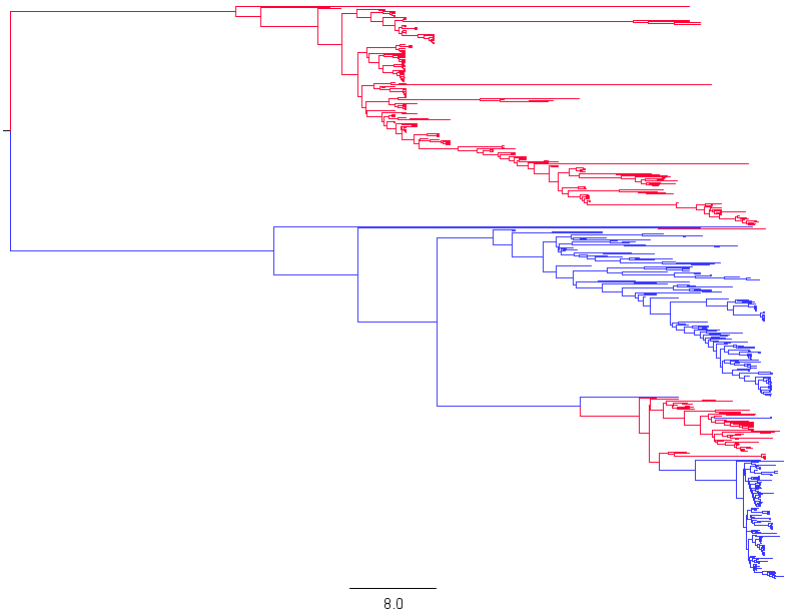

Figure A 4

The MCC tree of PA for swine H1N1, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

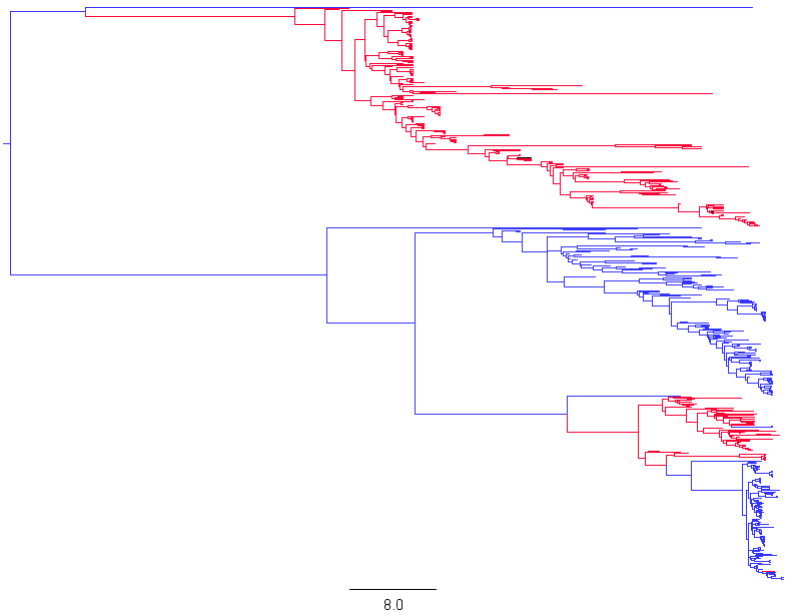

Figure A 5

The MCC tree of PB2 for swine H1N1, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

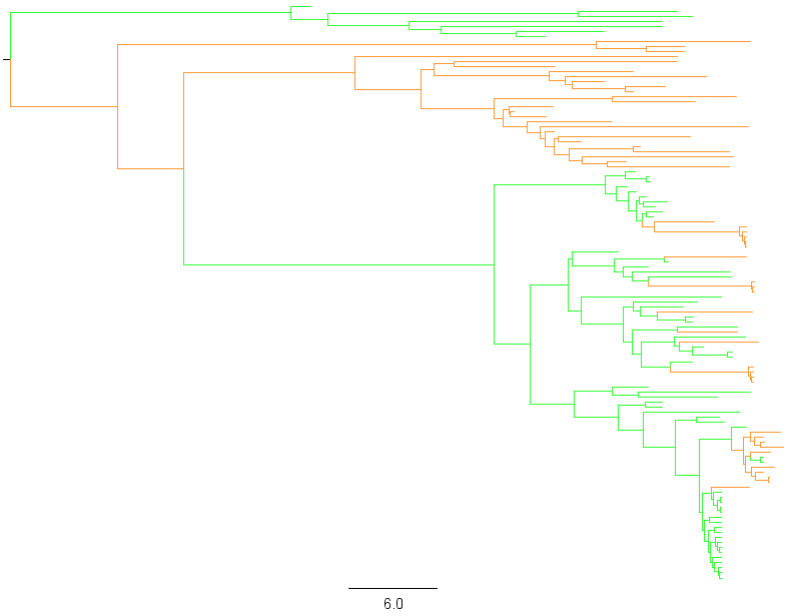

Figure A 6

The MCC tree of NA for swine H1N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

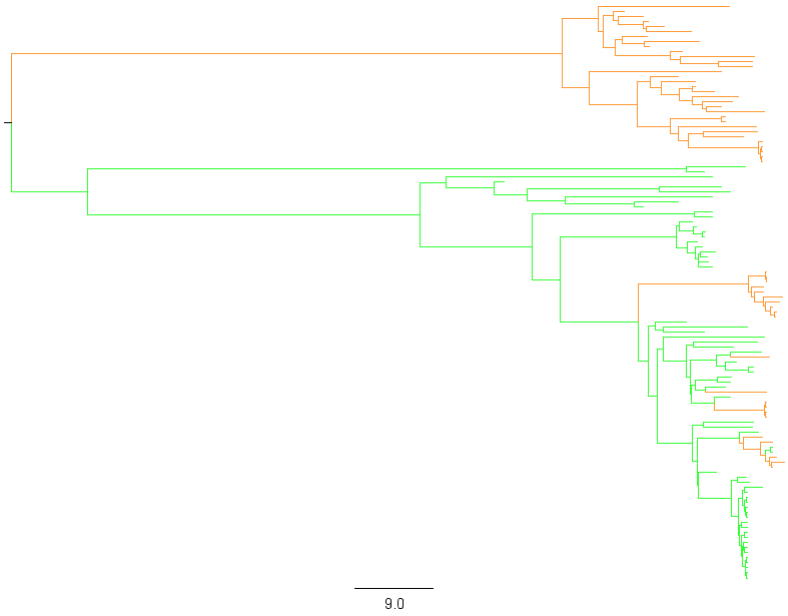

Figure A 7

The MCC tree of NP for swine H1N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

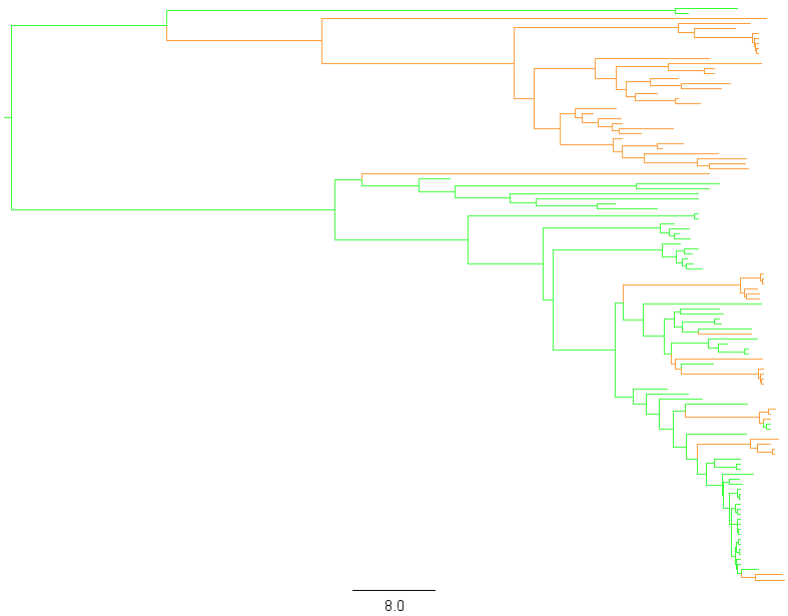

Figure A 8

The MCC tree of NS for swine H1N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

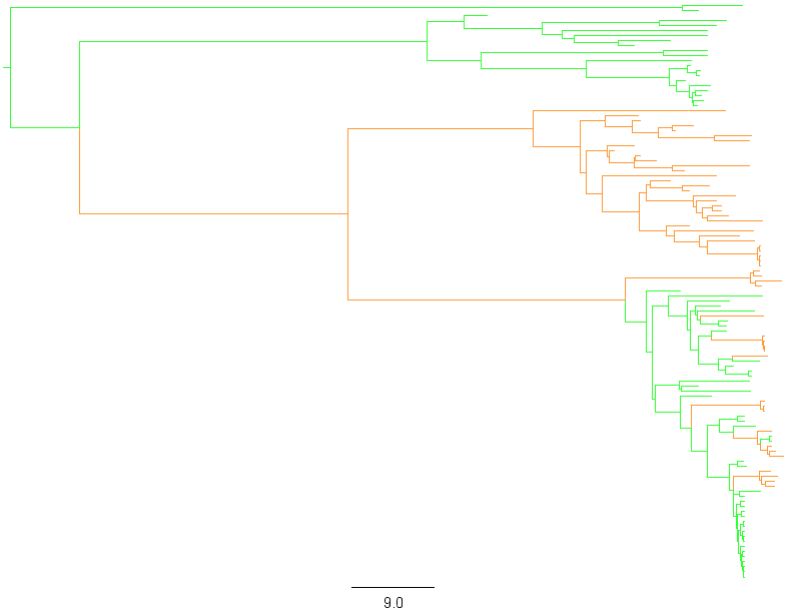

Figure A 9

The MCC tree of PA for swine H1N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

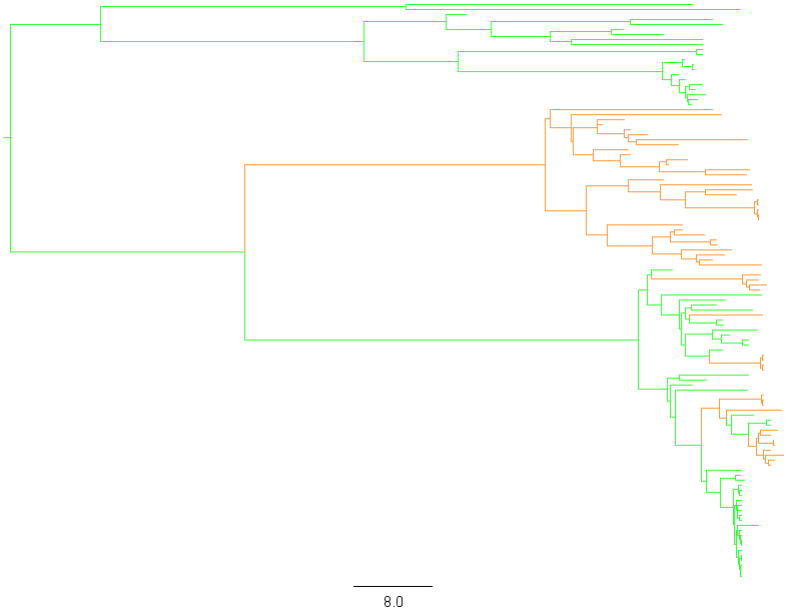

Figure A 10

The MCC tree of PB2 for swine H1N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

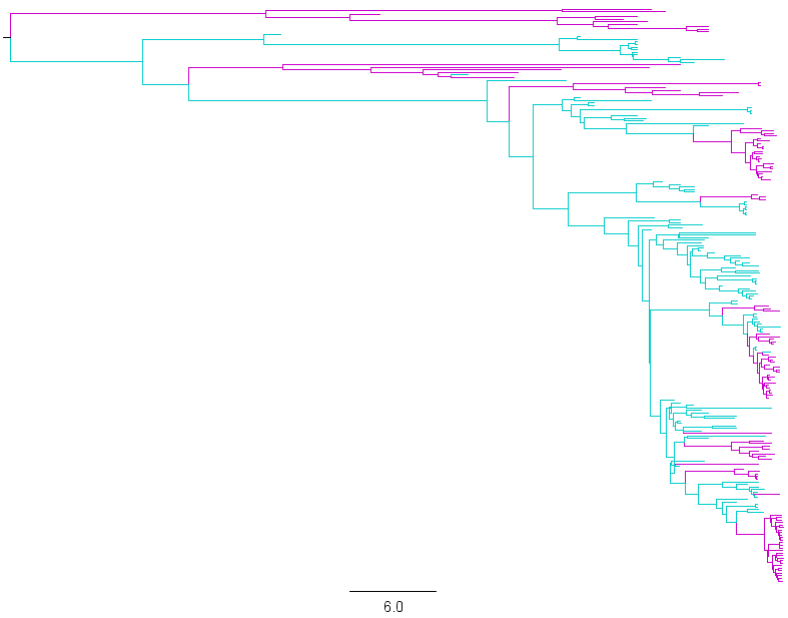

Figure A 11

The MCC tree of NA for swine H3N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

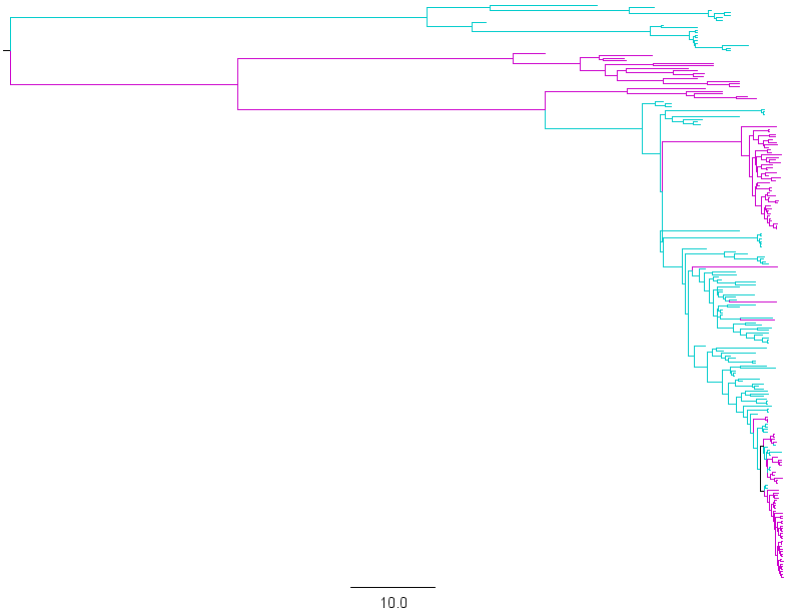

Figure A 12

The MCC tree of NP for swine H3N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

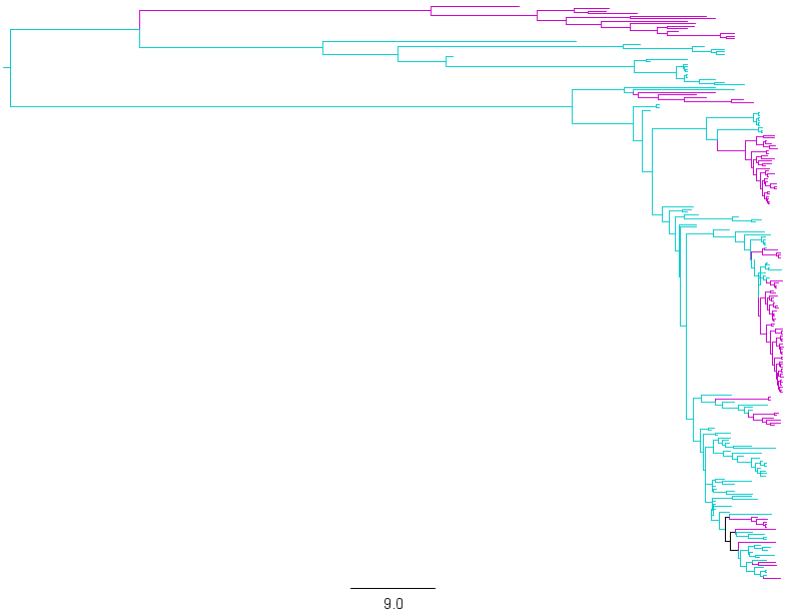

Figure A 13

The MCC tree of NS for swine H3N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

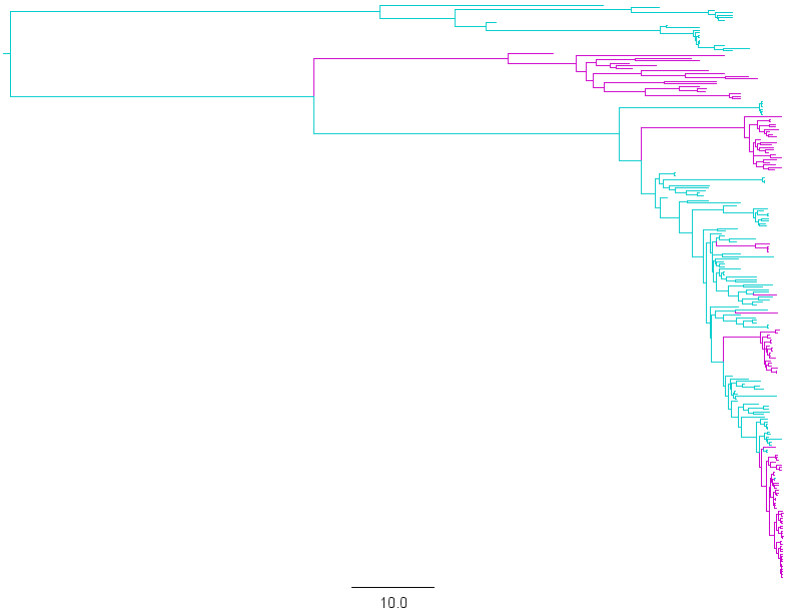

Figure A 14

The MCC tree of PA for swine H3N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

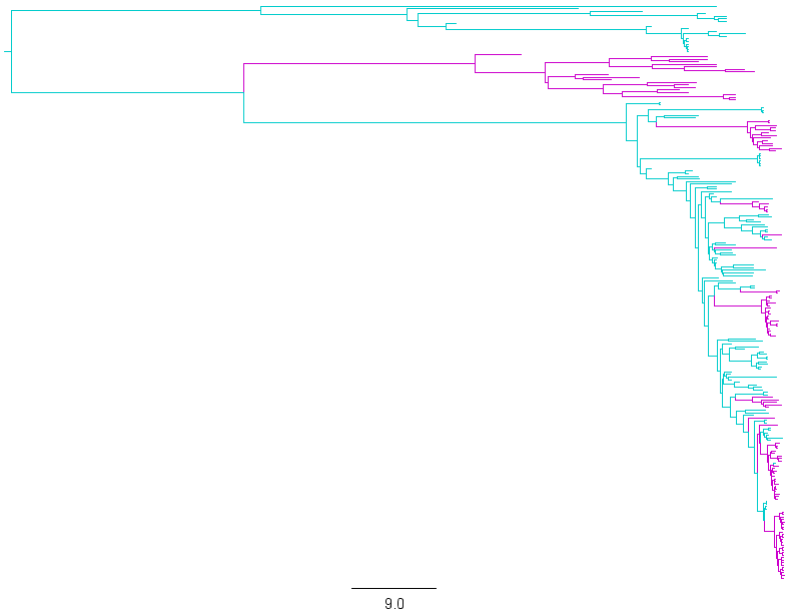

Figure A 15

The MCC tree of PB2 for swine H3N2, coloured according to two clades in MP. Departures of colour pattern of this tree compared to the MP tree indicate reassortment.

8.2 Appendix B

Table B 1

Distance in topology scores for swine H1N1 using dist.topo (PH85).

Segment	PB2	PB1	PA	HA	NP	NA	MP	NS
PB2	372.9	710.2	689.6	734.1	729.6	760.0	783.8	777.2
PB1		366.5	728.9	773.3	753.7	780.3	798.3	800.8
PA			401.5	752.7	733.6	761.1	787.7	789.8
HA				374.9	757.6	710.8	786.0	783.3
NP					444.3	767.4	806.2	794.5
NA						393.0	804.8	787.2
MP							551.2	824.7
NS								516.9
random	1007.6	1007.6	1007.6	1007.6	1007.6	1007.6	1007.6	1007.6

Table B 2

Distance in topology scores for swine H1N2 using dist.topo (PH85).

Segment	PB2	PB1	PA	HA	NP	NA	MP	NS
PB2	54.42	130.85	130.13	183.24	148.17	170.14	157.31	166.13
PB1		51.88	130.72	174.85	146.49	167.92	158.72	164.66
PA			51.68	176.43	149.22	164.23	164.66	170.95
HA				66.46	182.91	157.84	180.62	184.27
NP					66.60	174.35	168.06	171.64
NA						69.15	177.01	174.08
MP							78.61	173.35
NS								75.67
random	223.46	223.46	223.46	223.55	223.46	223.46	223.46	223.46

Table B 3

Distance in topology scores for swine H3N2 using dist.topo (PH85).

Segment	PB2	PB1	PA	HA	NP	NA	MP	NS
PB2	174.5	346.8	332.3	359.9	360.5	363.8	376.0	361.4
PB1		177.0	328.7	358.0	371.1	362.0	384.0	369.3
PA			176.8	356.4	359.3	360.0	374.2	361.3
HA				159.3	375.6	326.8	354.9	380.3
NP					198.6	380.2	389.2	374.3
NA						194.0	370.0	378.4
MP							231.3	375.2
NS								225.4
random	445.5	445.5	445.5	445.5	445.5	445.5	445.5	445.5

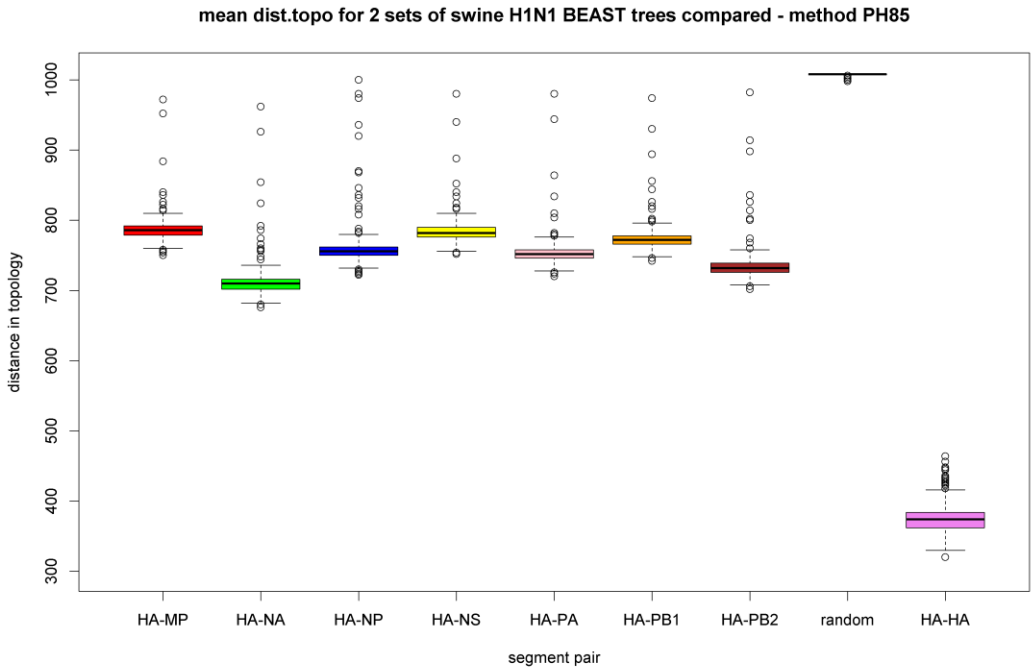

Figure B 1

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. HA compared to all other segments.

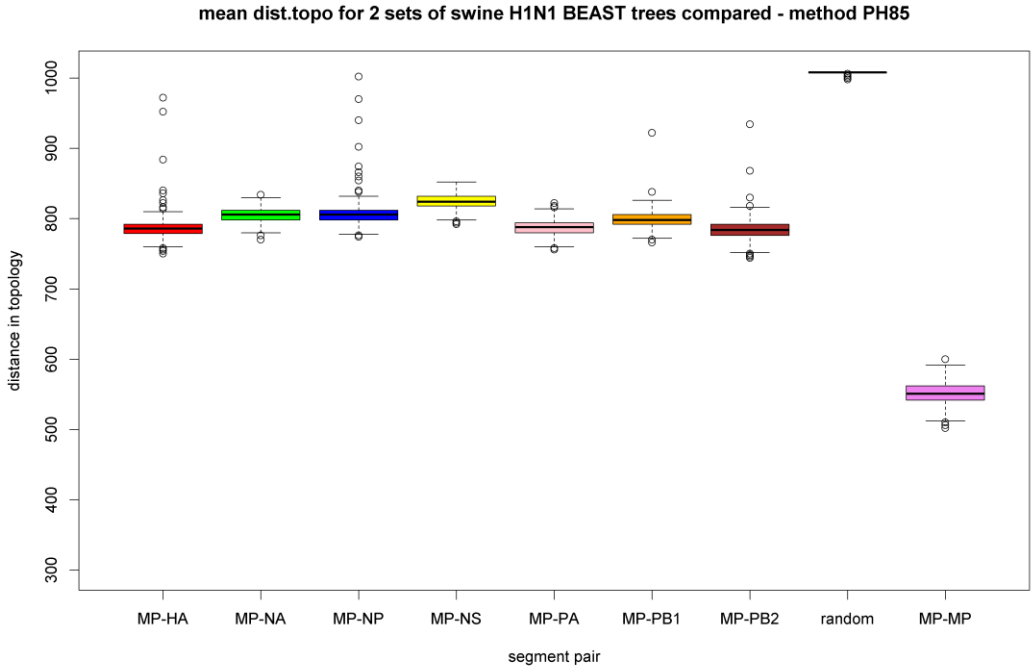

Figure B 2

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. MP compared to all other segments.

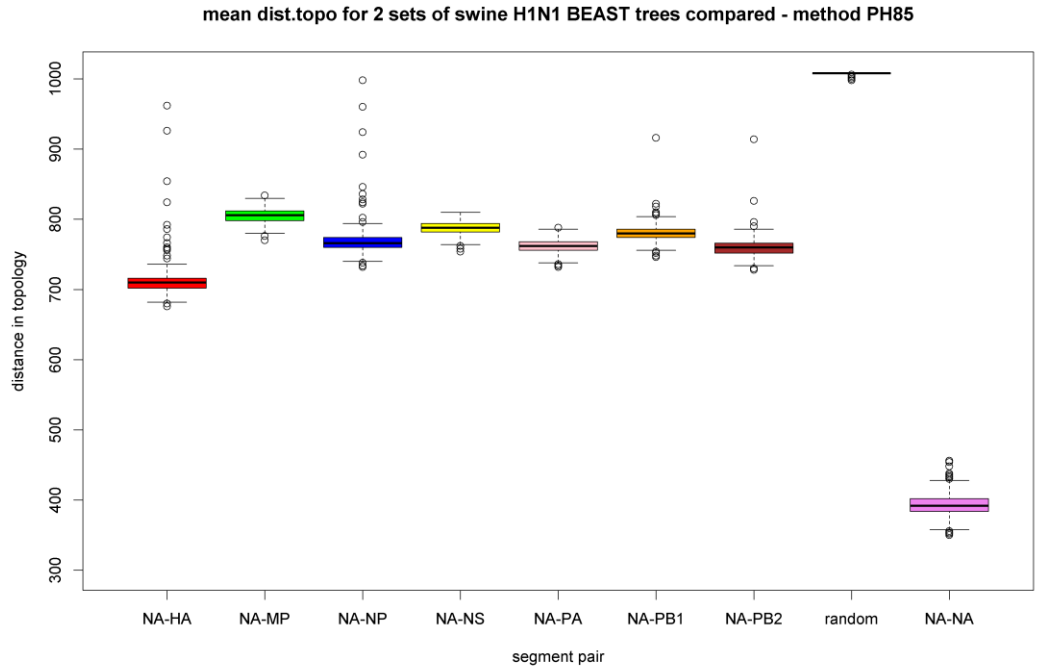

Figure B 3

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. NA compared to all other segments.

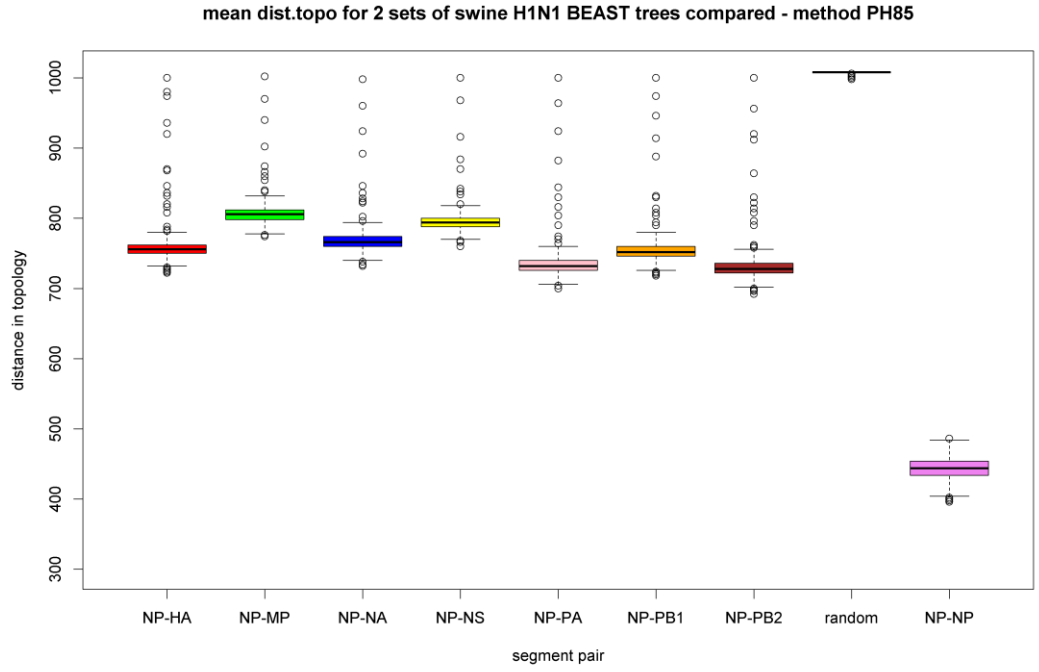

Figure B 4

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. NP compared to all other segments.

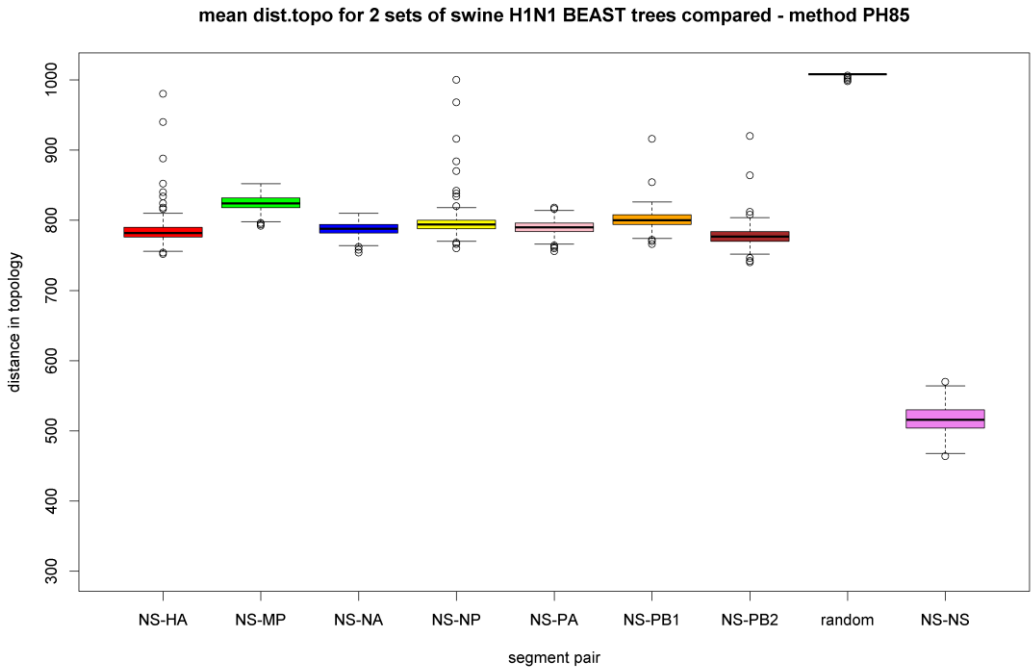

Figure B 5

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. NS compared to all other segments.

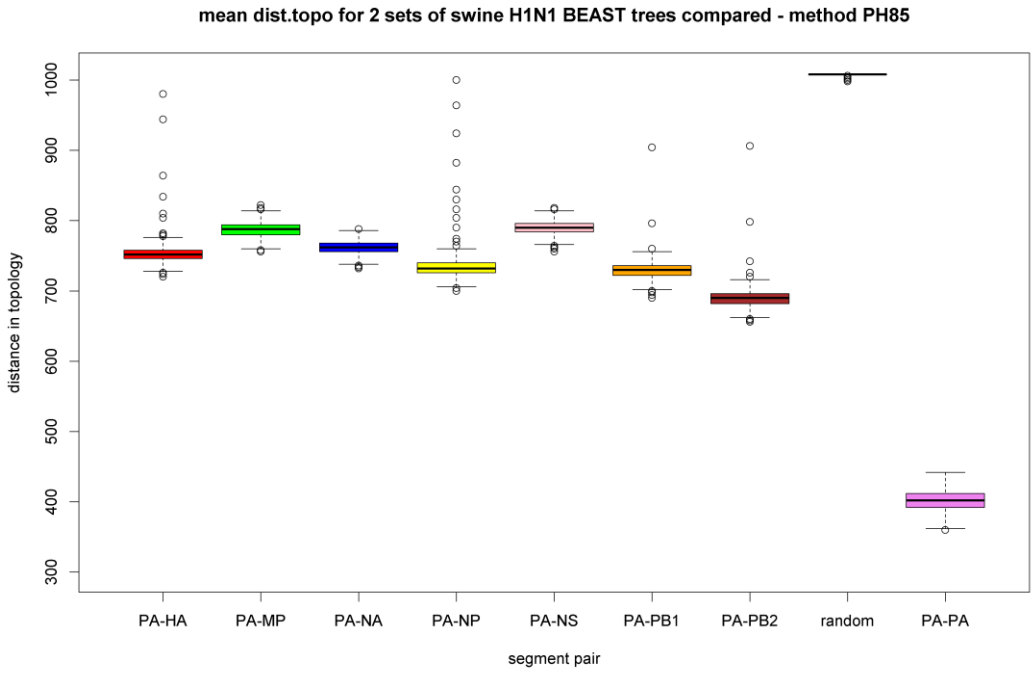

Figure B 6

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. PA compared to all other segments.

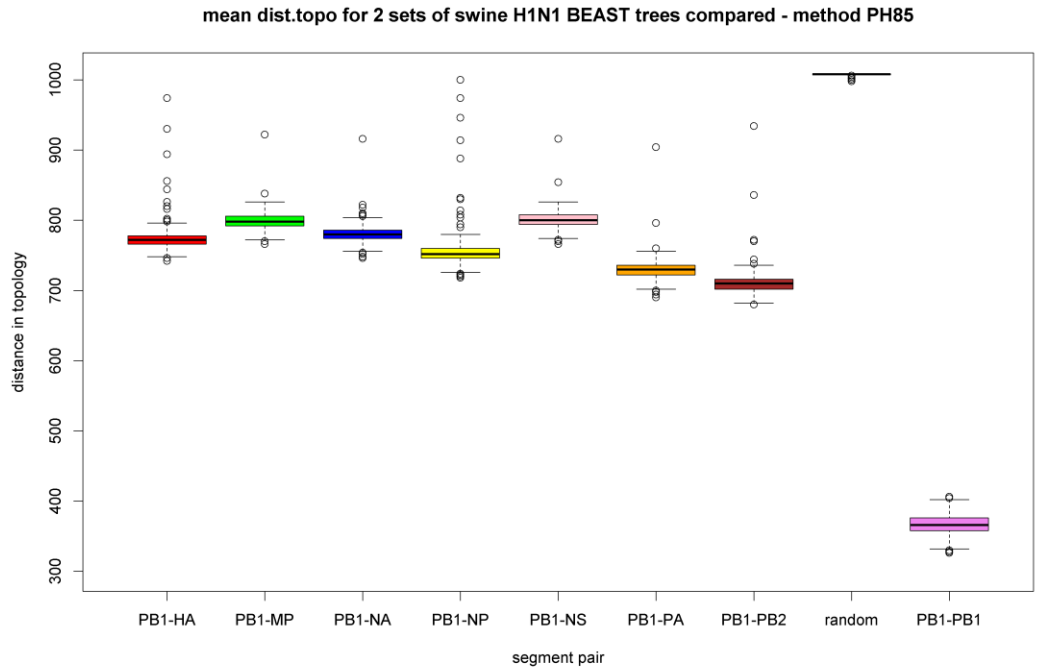

Figure B 7

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. PB1 compared to all other segments.

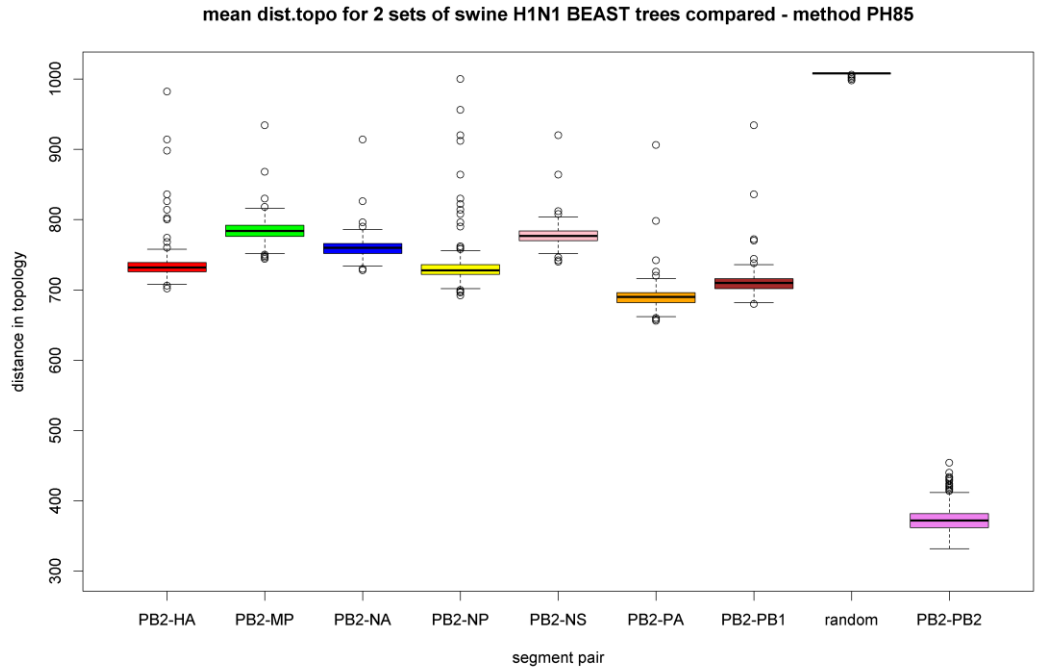

Figure B 8

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. PB2 compared to all other segments.

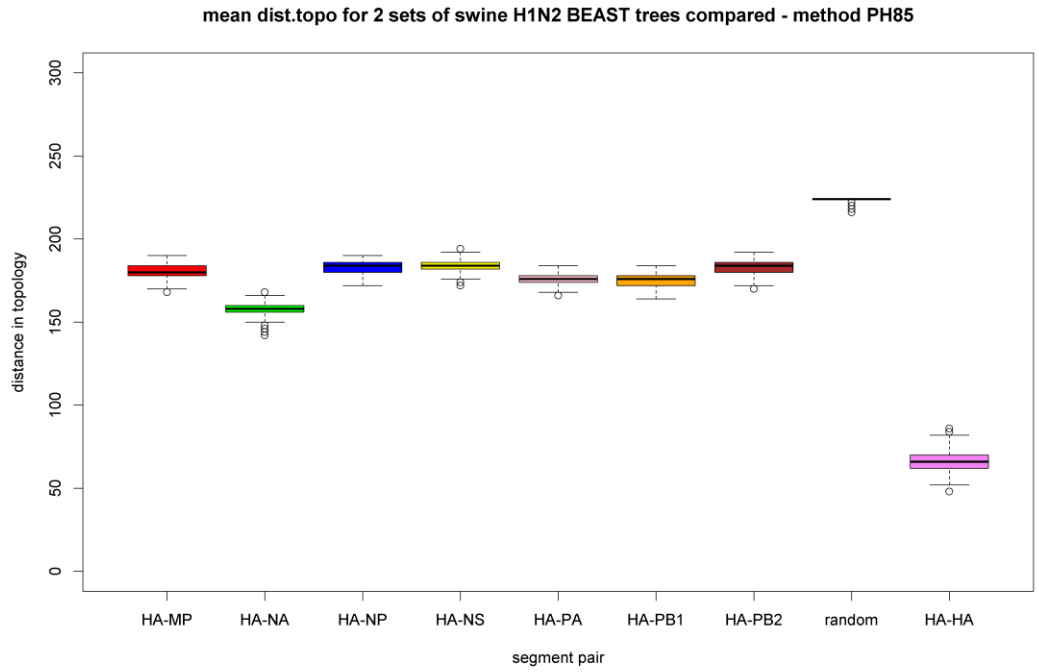

Figure B 9

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. HA compared to all other segments.

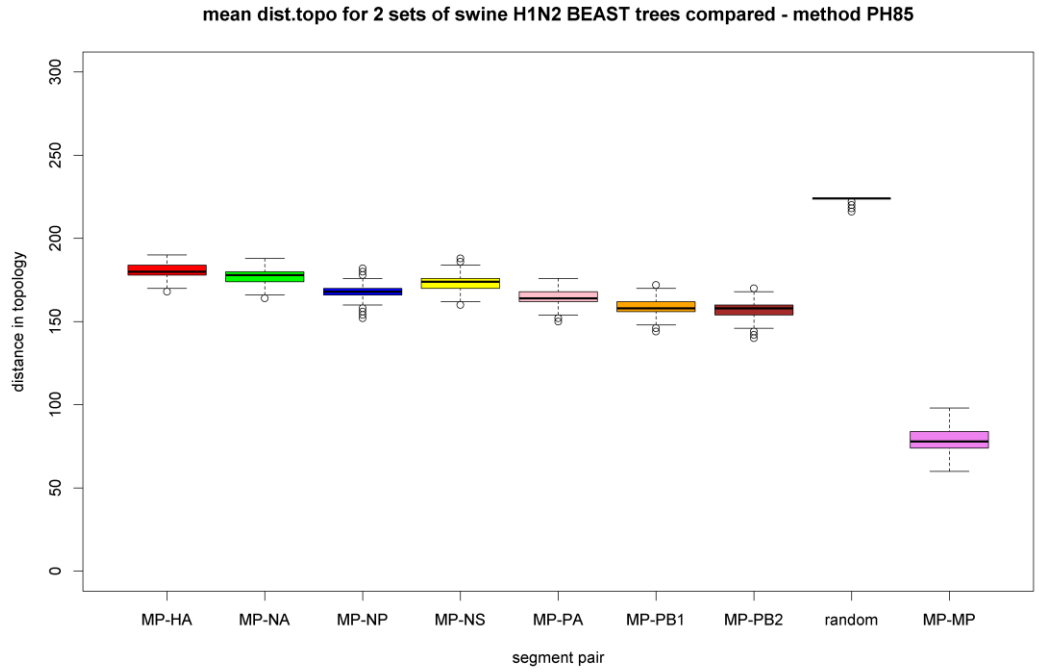

Figure B 10

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. MP compared to all other segments.

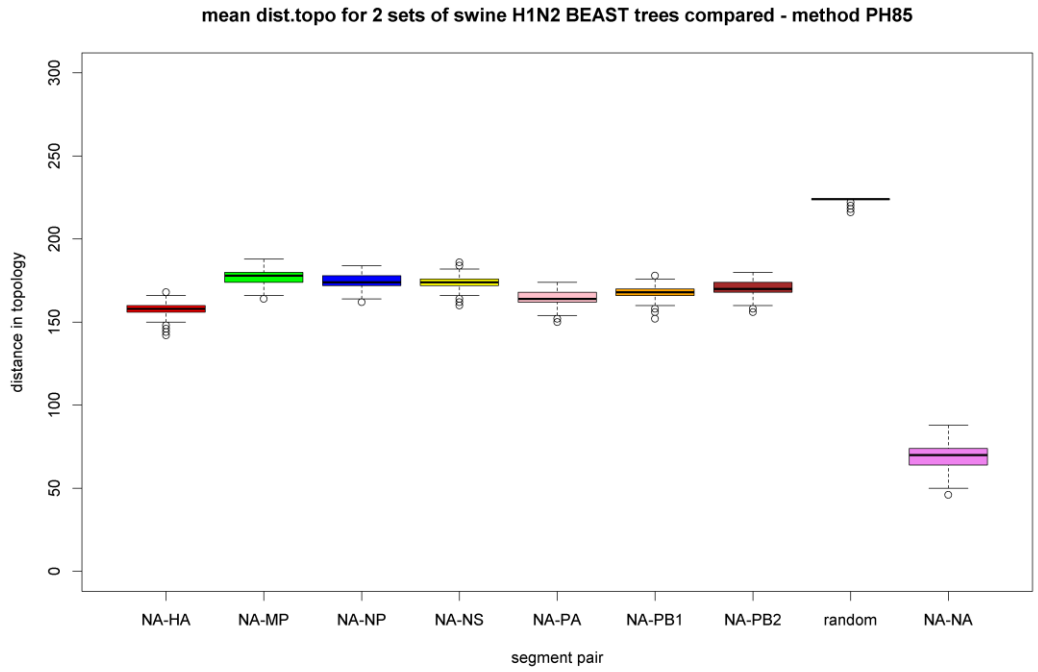

Figure B 11

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. NA compared to all other segments.

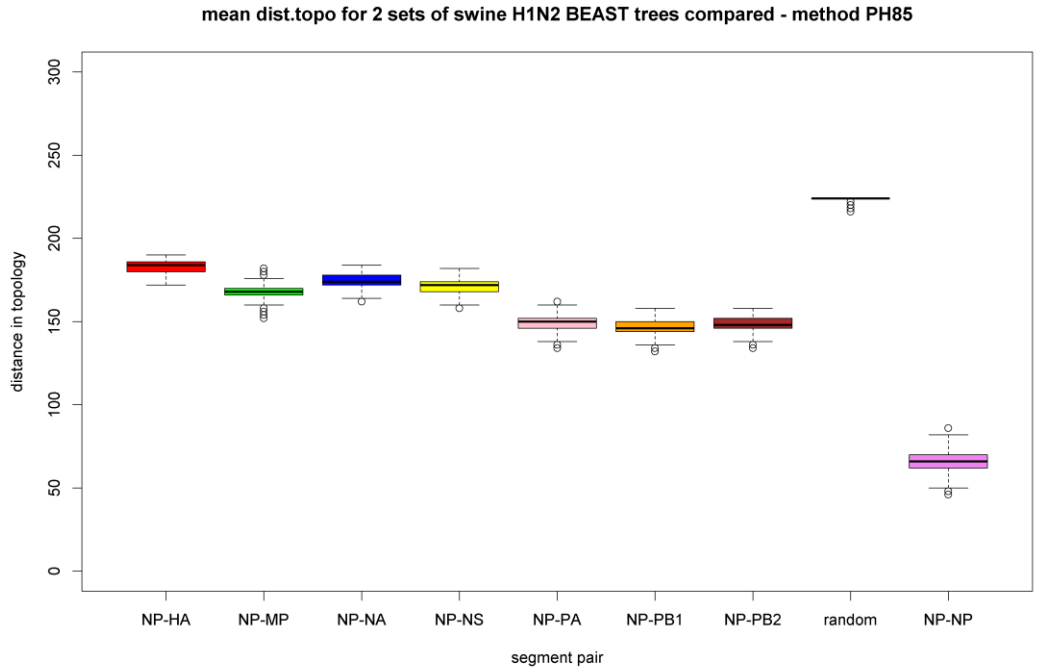

Figure B 12

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. NP compared to all other segments.

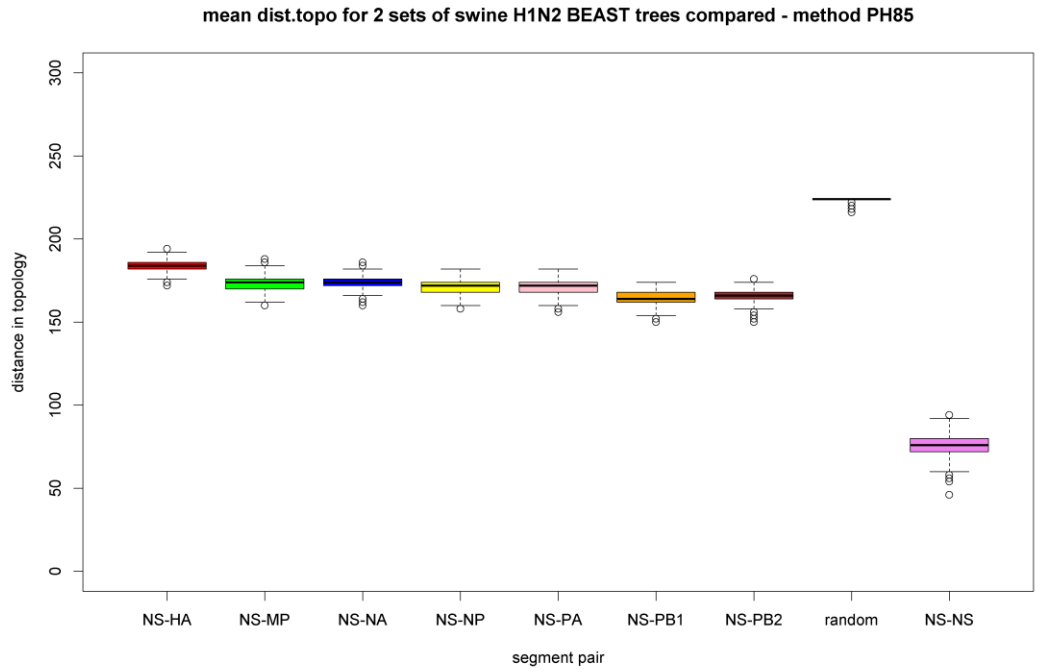

Figure B 13

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. NS compared to all other segments.

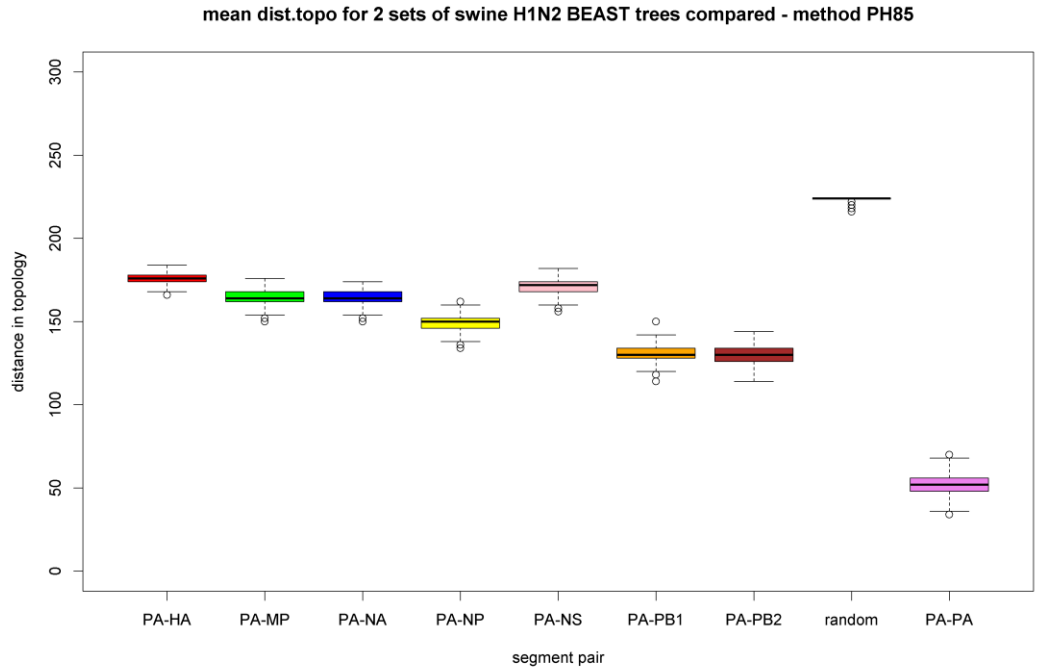

Figure B 14

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. PA compared to all other segments.

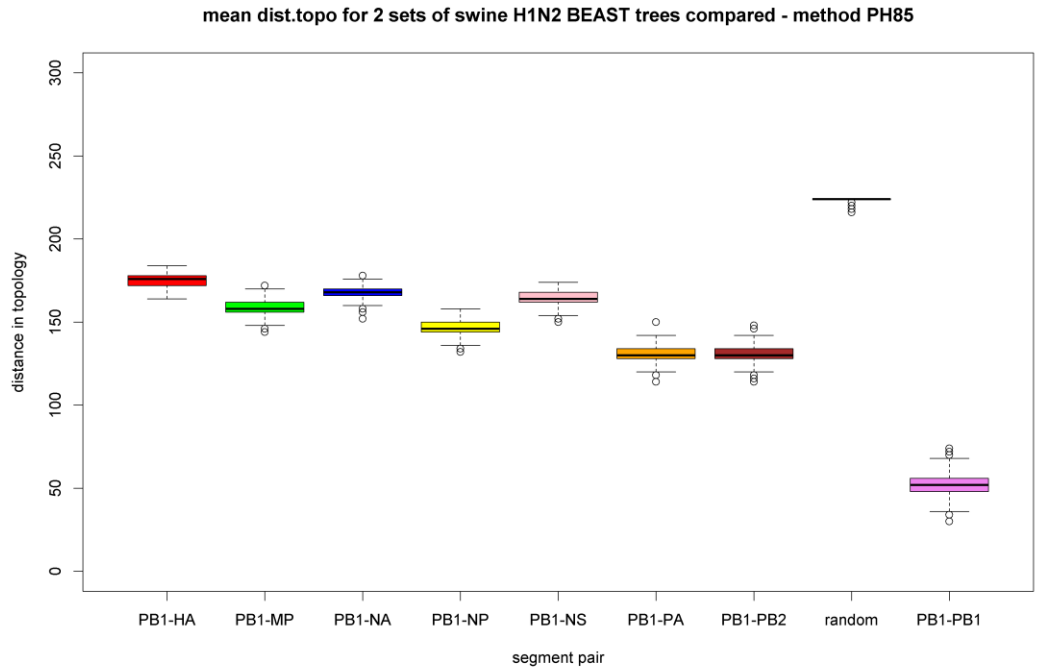

Figure B 15

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. PB1 compared to all other segments.

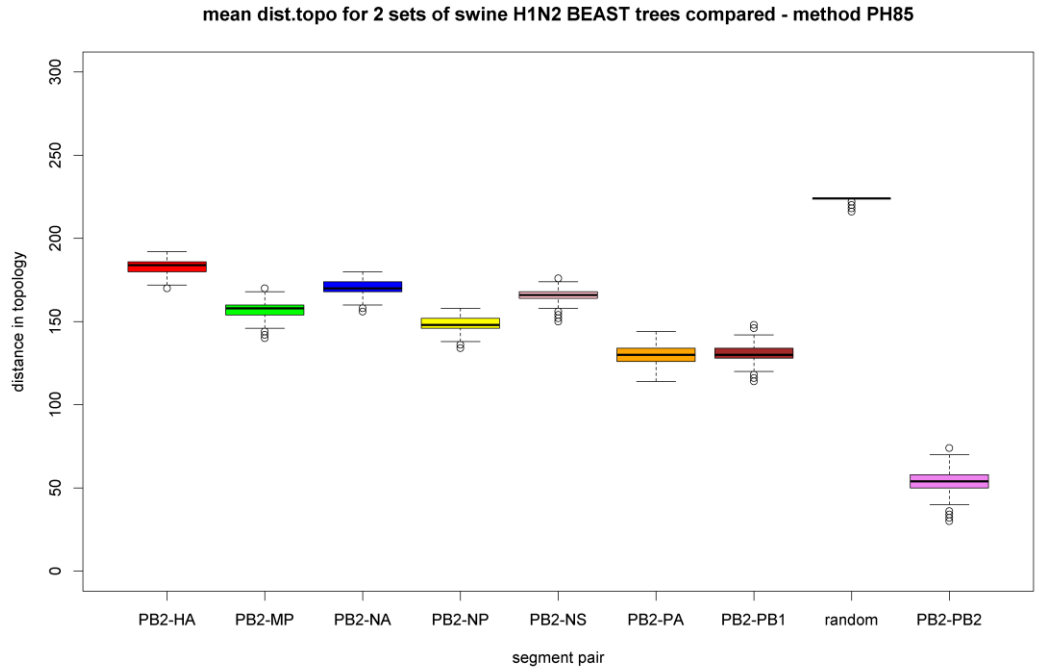

Figure B 16

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. PB2 compared to all other segments.

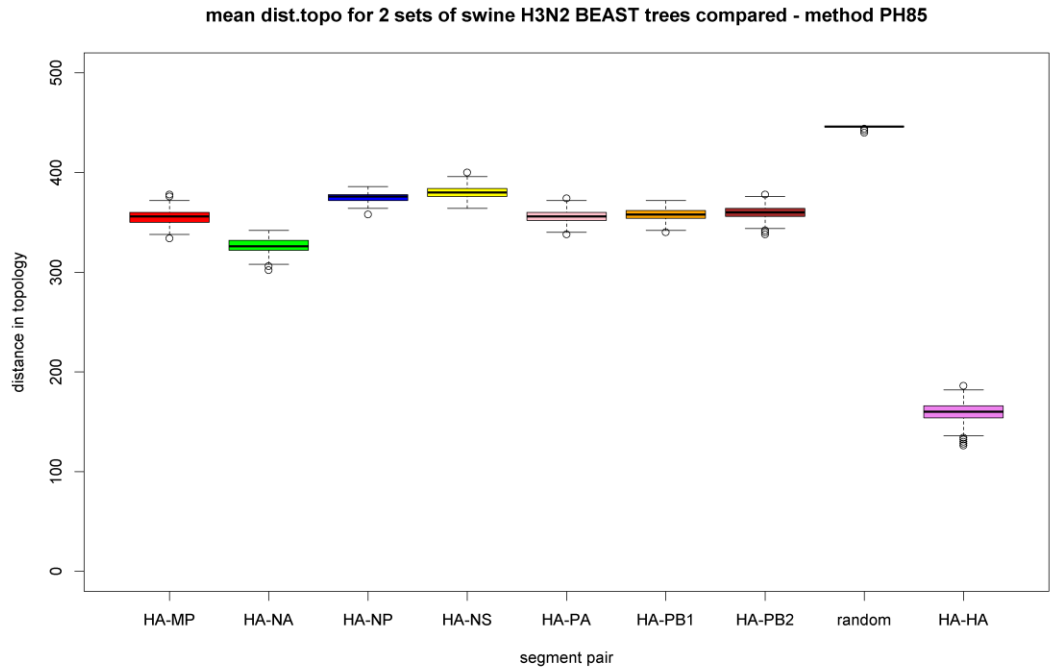

Figure B 17

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. HA compared to all other segments.

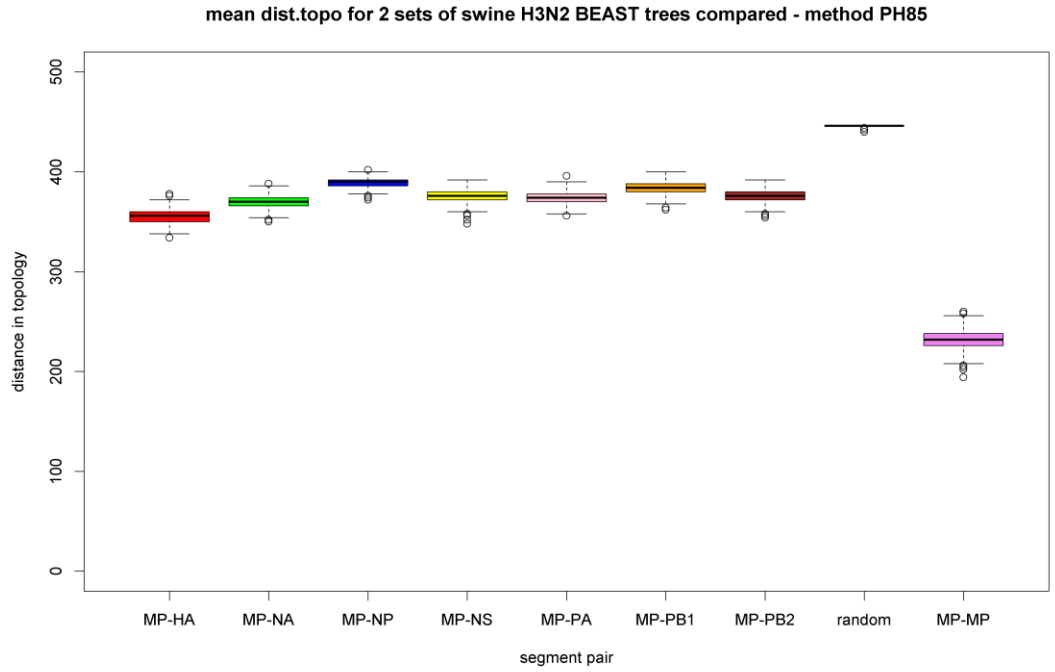

Figure B 18

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. MP compared to all other segments.

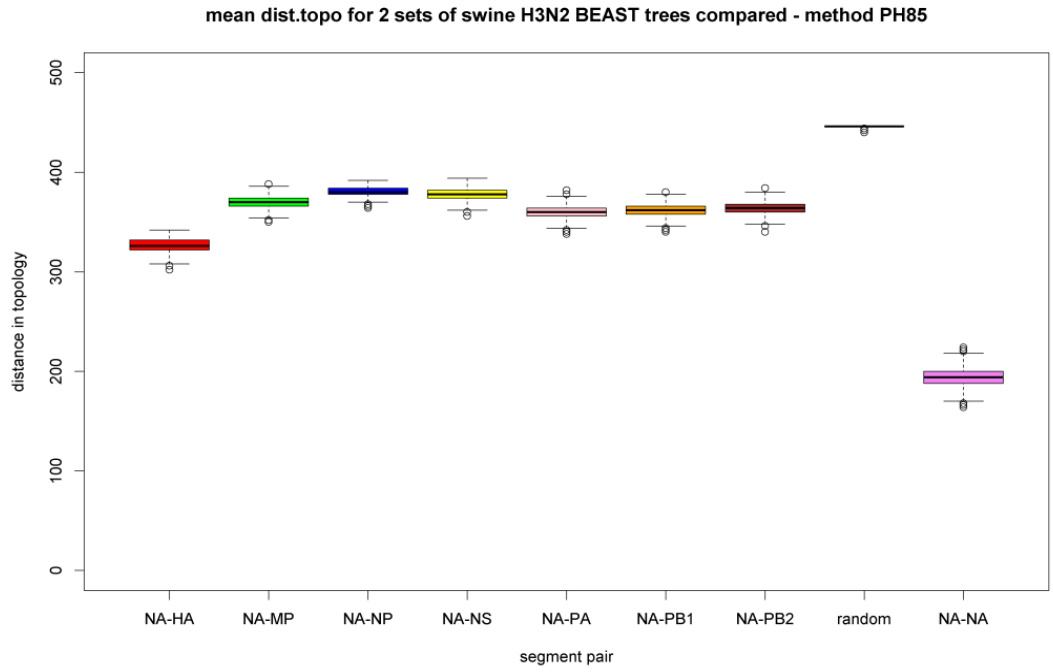

Figure B 19

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. NA compared to all other segments.

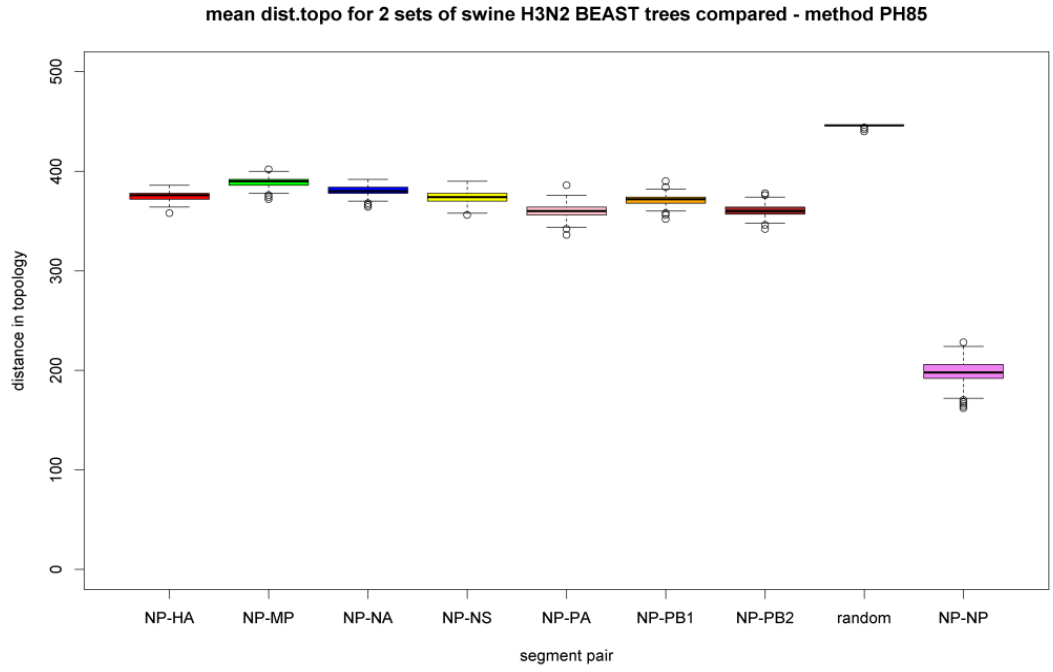

Figure B 20

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. NP compared to all other segments.

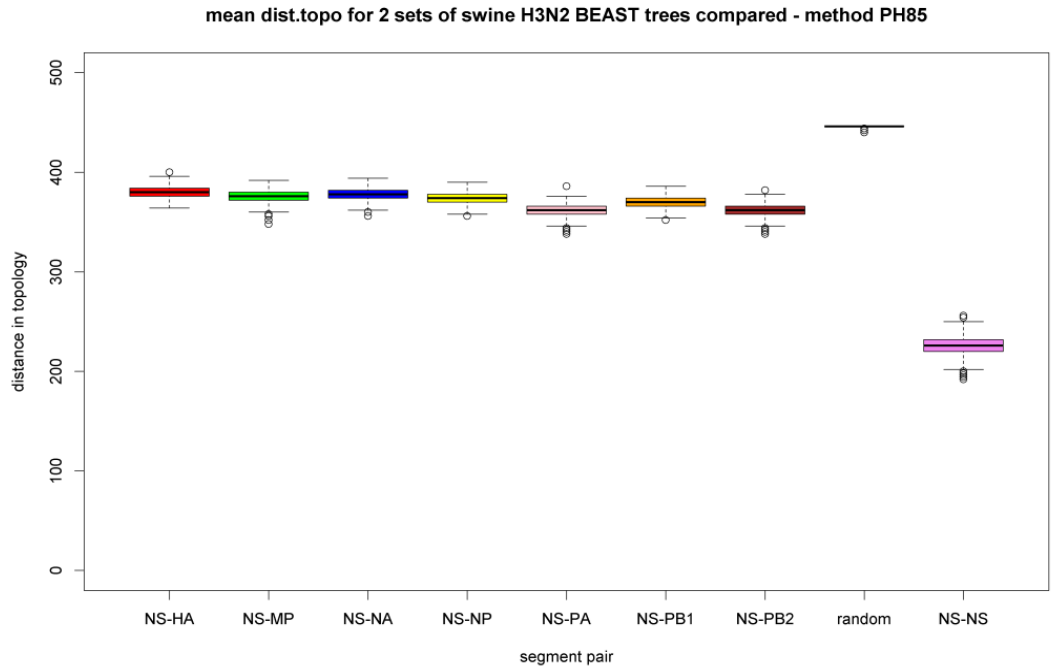

Figure B 21

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. NS compared to all other segments.

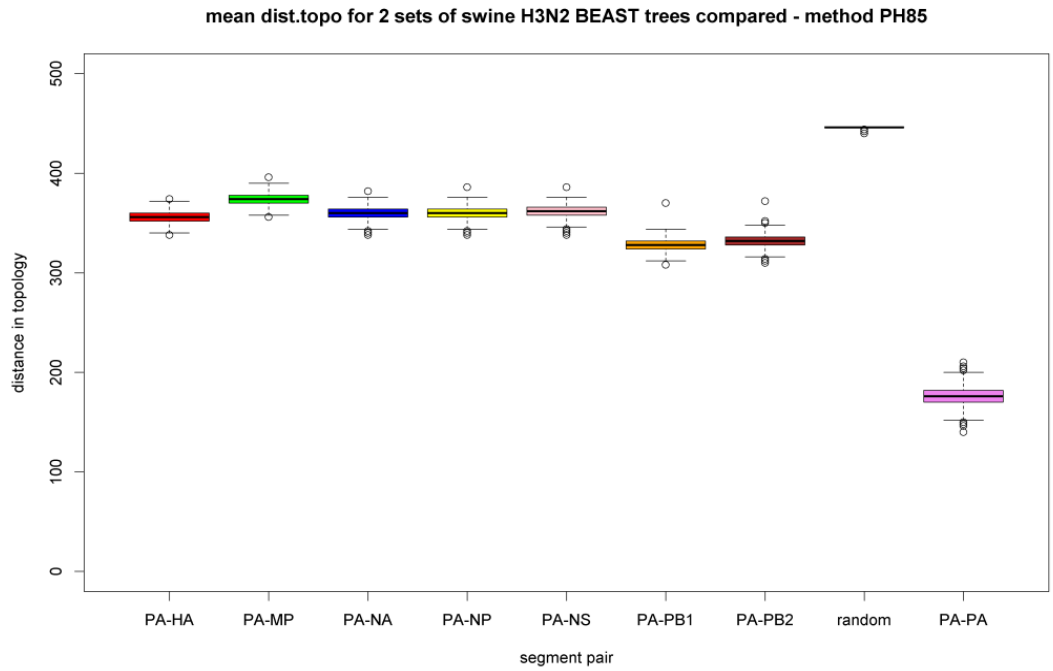

Figure B 22

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. PA compared to all other segments.

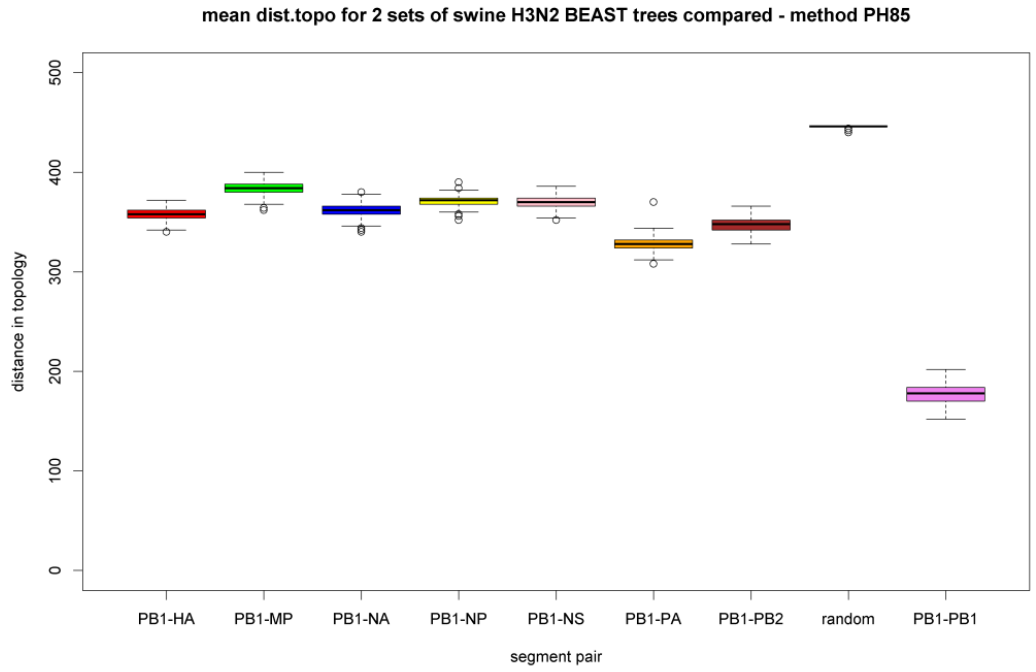

Figure B 23

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. PB1 compared to all other segments.

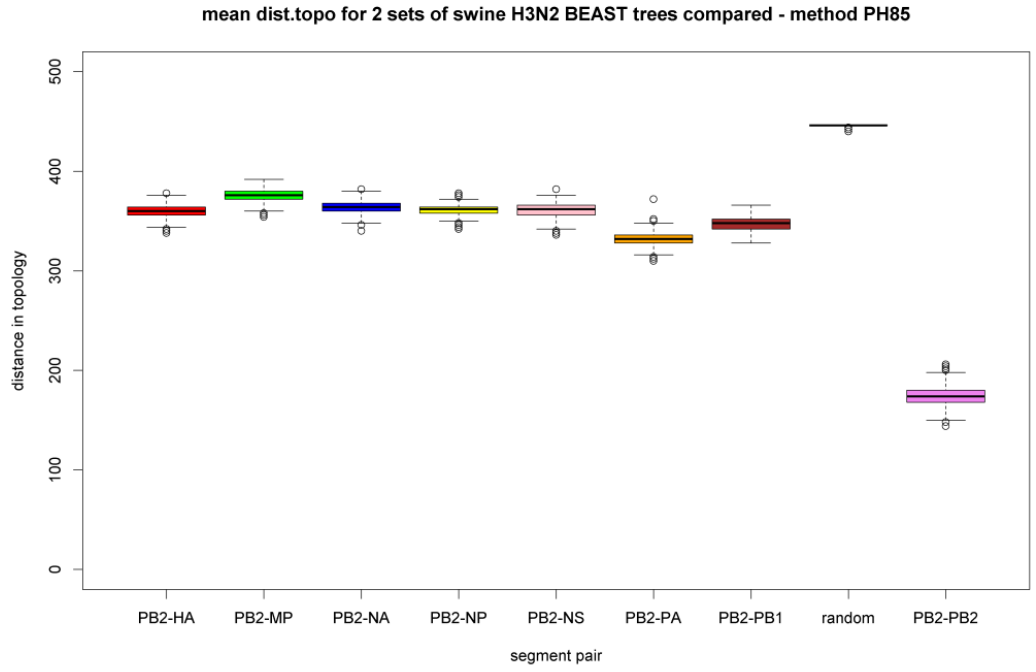

Figure B 24

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. PB2 compared to all other segments.

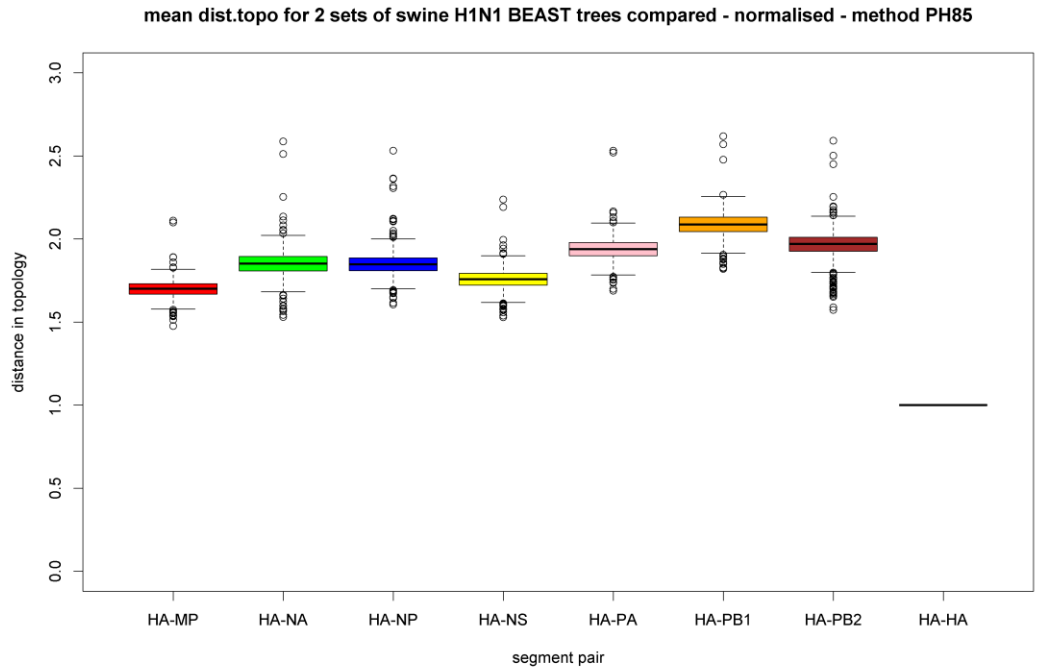

Figure B 25

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. HA compared to all other segments. The values are normalised to the value of HA.

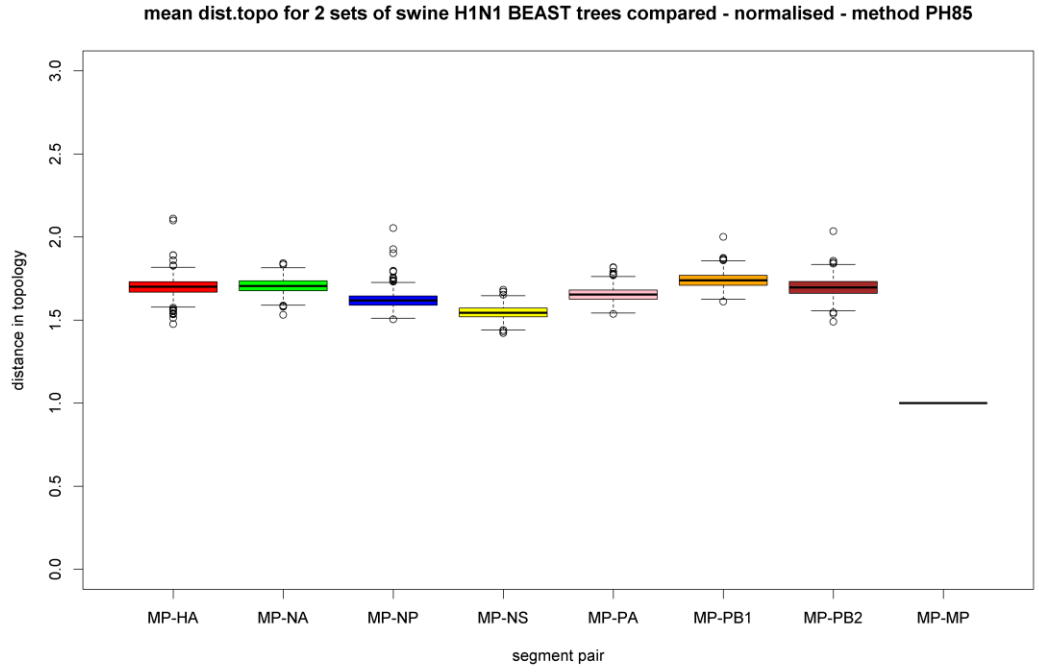

Figure B 26

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. MP compared to all other segments. The values are normalised to the value of MP.

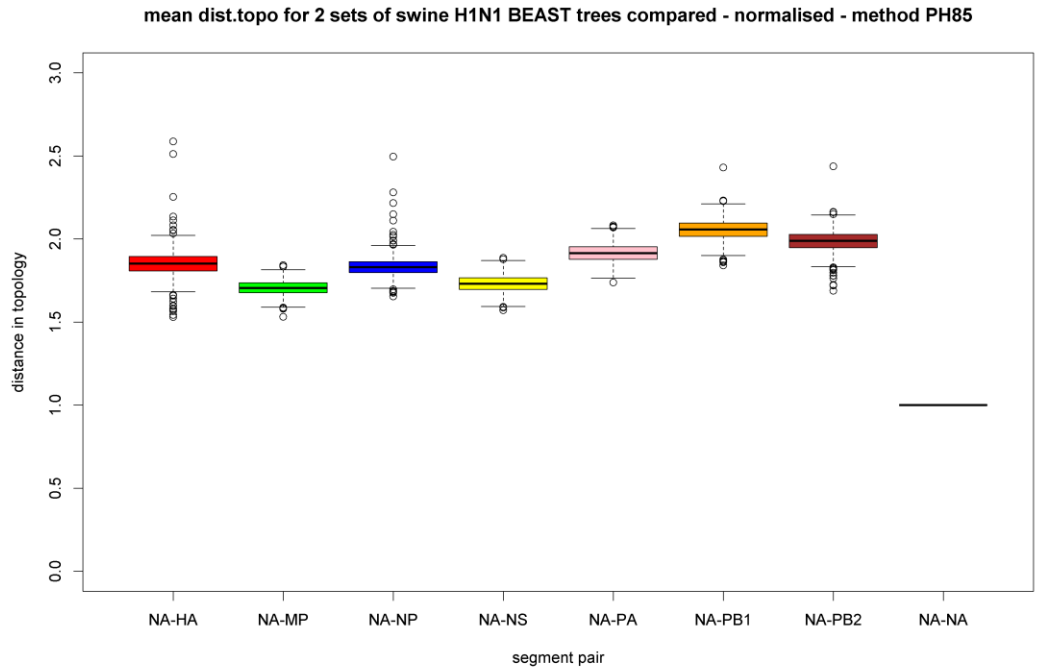

Figure B 27

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. NA compared to all other segments. The values are normalised to the value of NA.

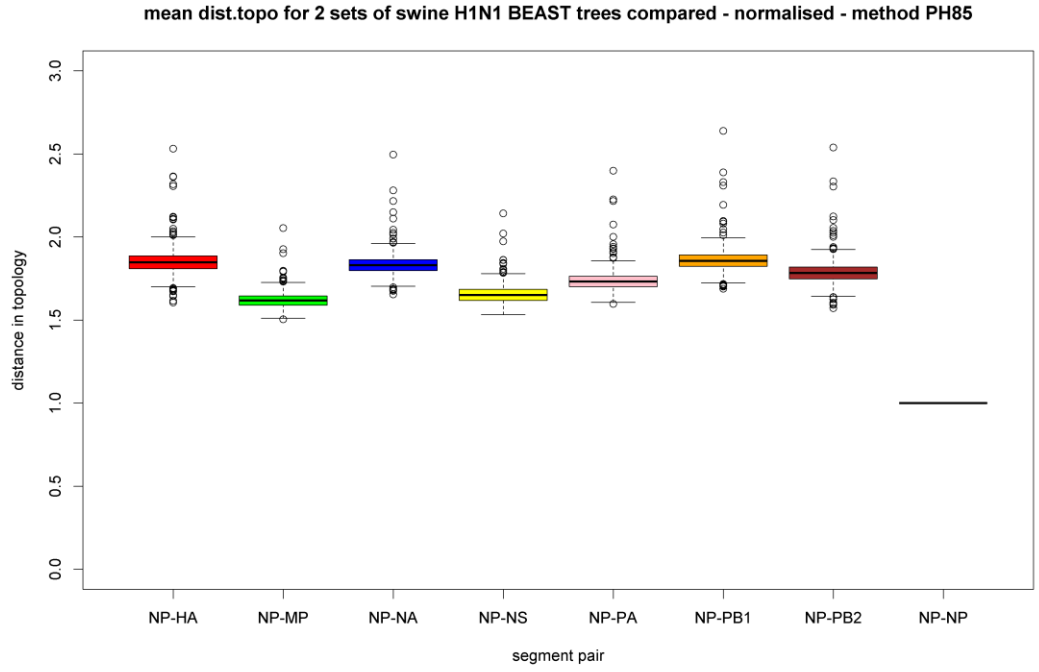

Figure B 28

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. NP compared to all other segments. The values are normalised to the value of NP.

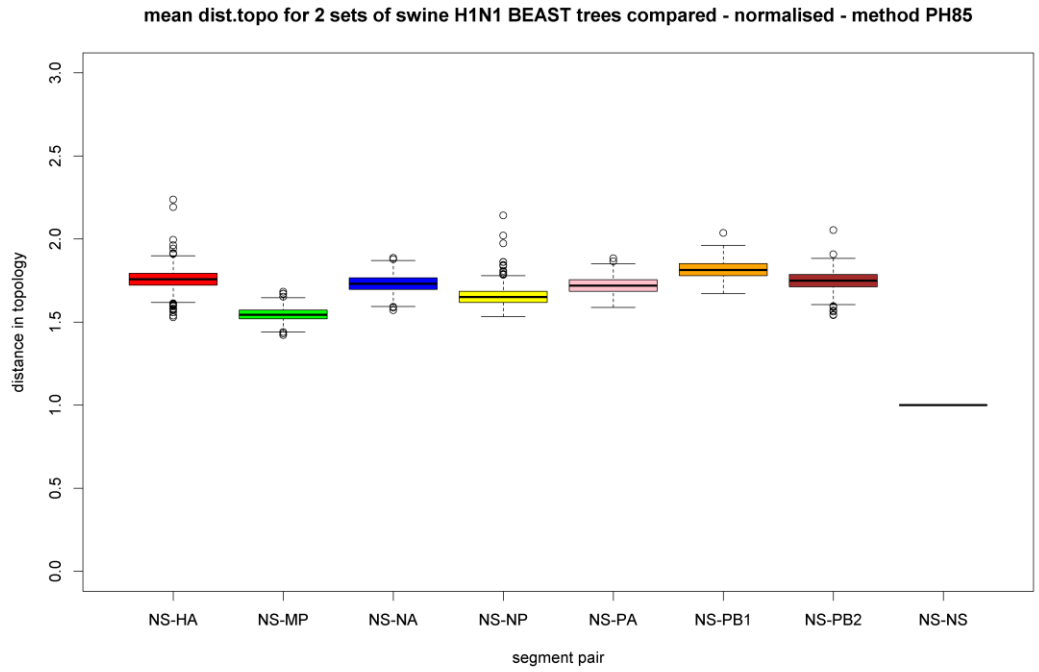

Figure B 29

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. NS compared to all other segments. The values are normalised to the value of NS.

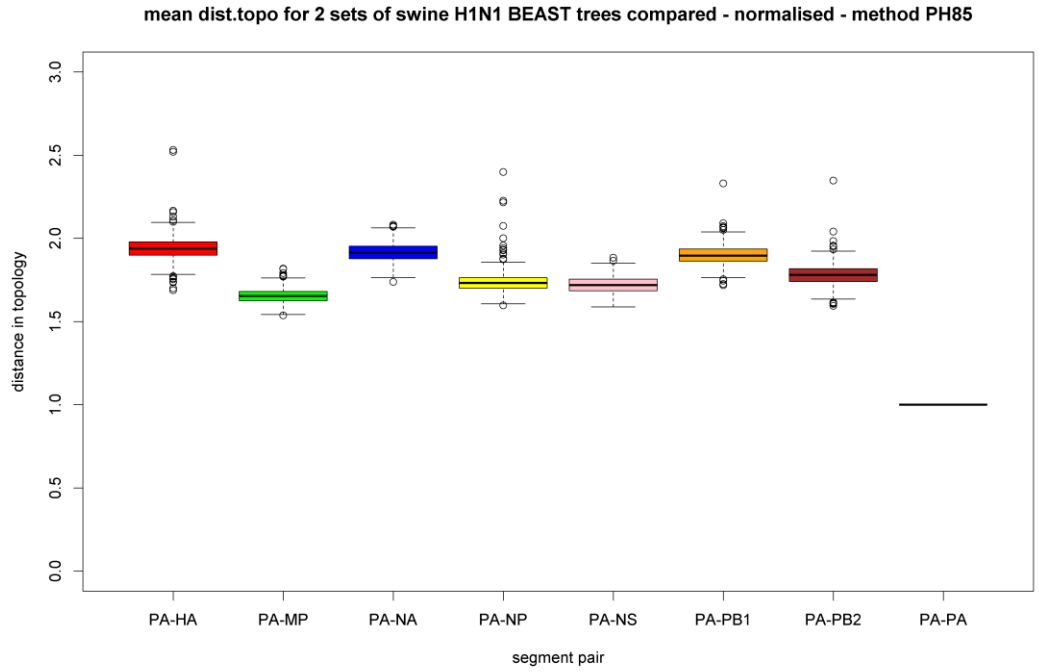

Figure B 30

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. PA compared to all other segments. The values are normalised to the value of PA.

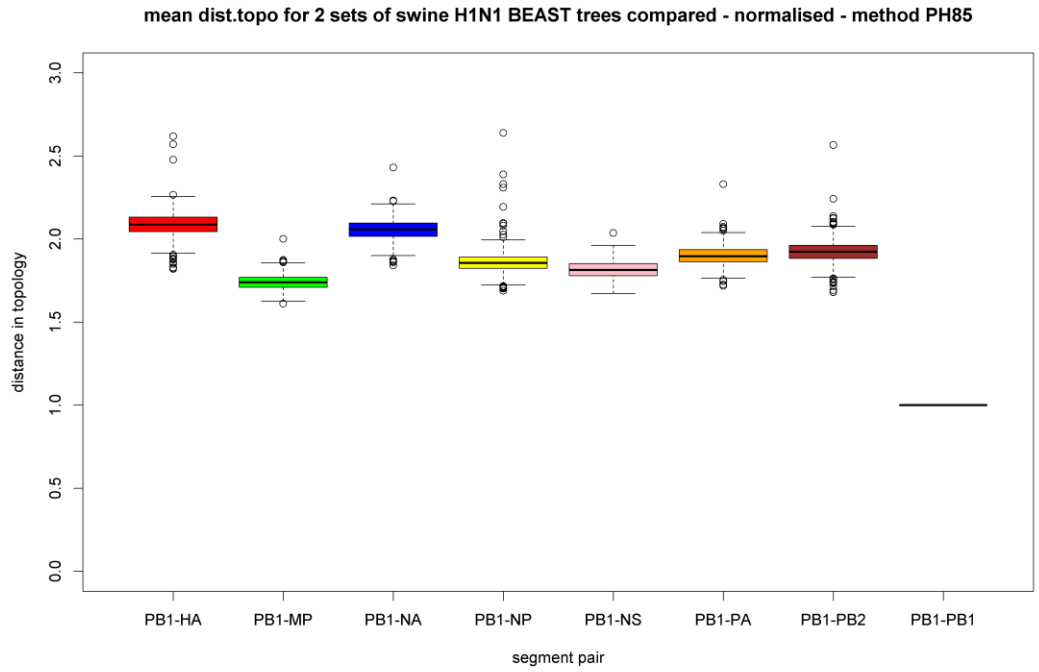

Figure B 31

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. PB1 compared to all other segments. The values are normalised to the value of PB1.

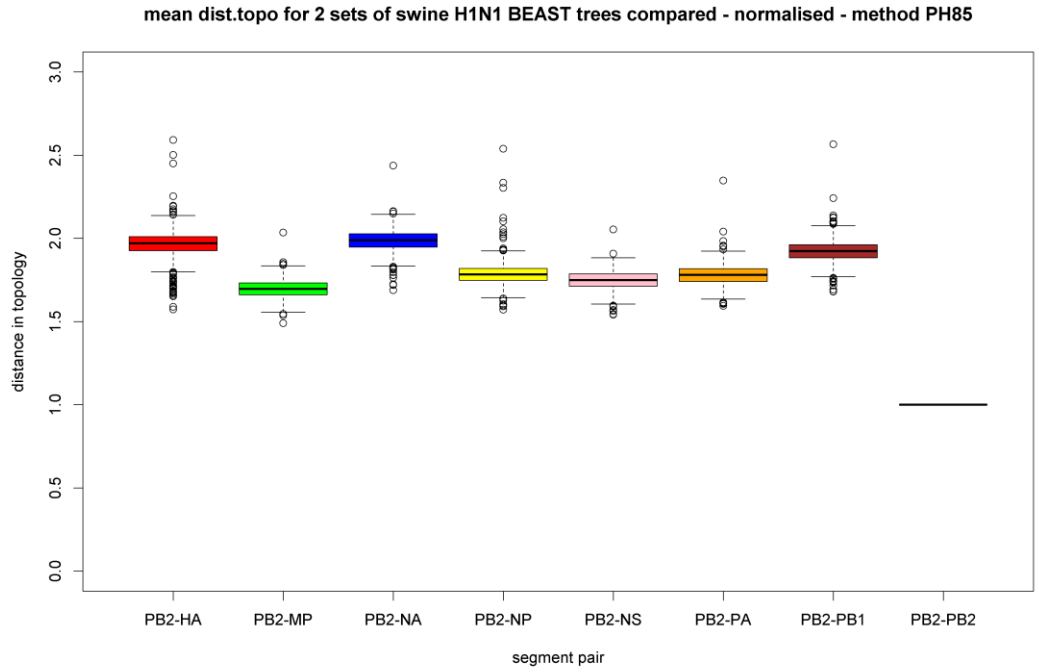

Figure B 32

Mean distance in topology for 2 sets of swine H1N1 BEAST trees compared. PB2 compared to all other segments. The values are normalised to the value of PB2.

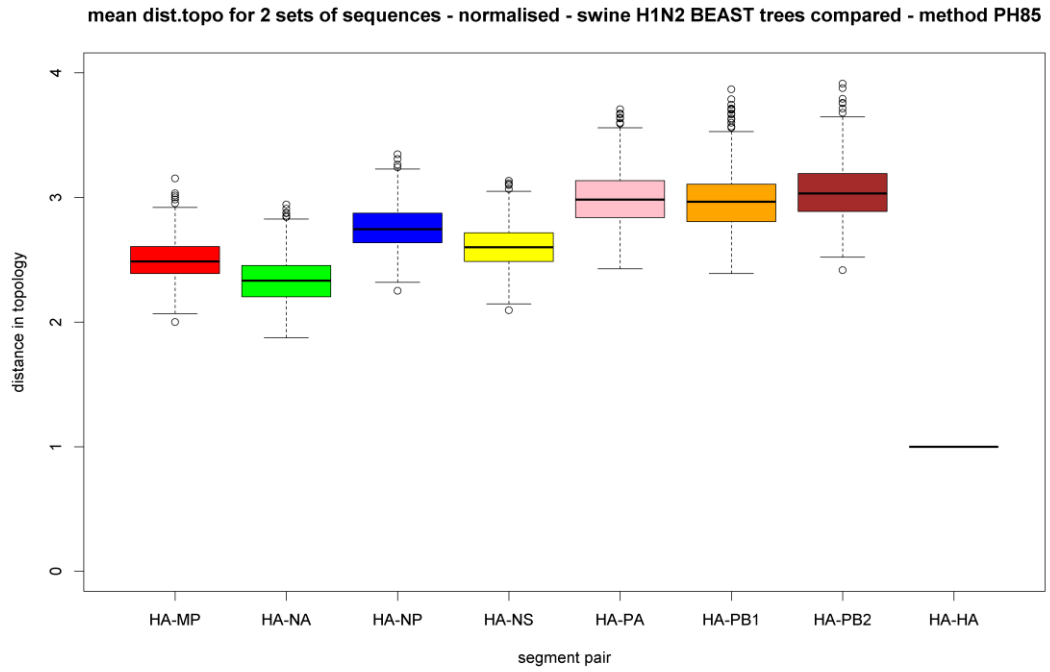

Figure B 33

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. HA compared to all other segments. The values are normalised to the value of HA.

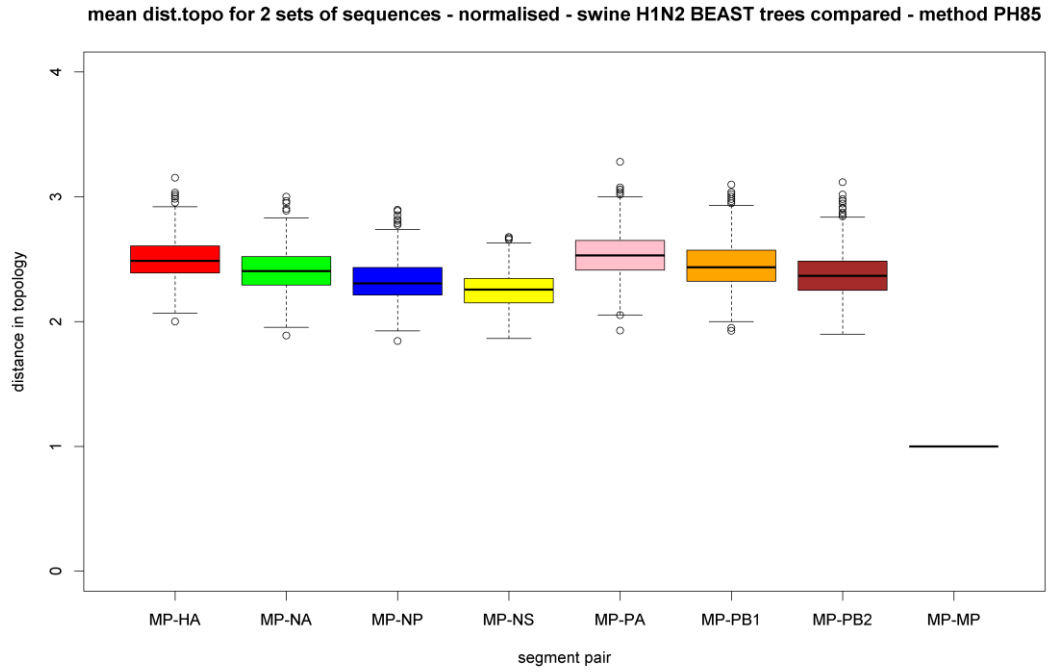

Figure B 34

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. MP compared to all other segments. The values are normalised to the value of MP.

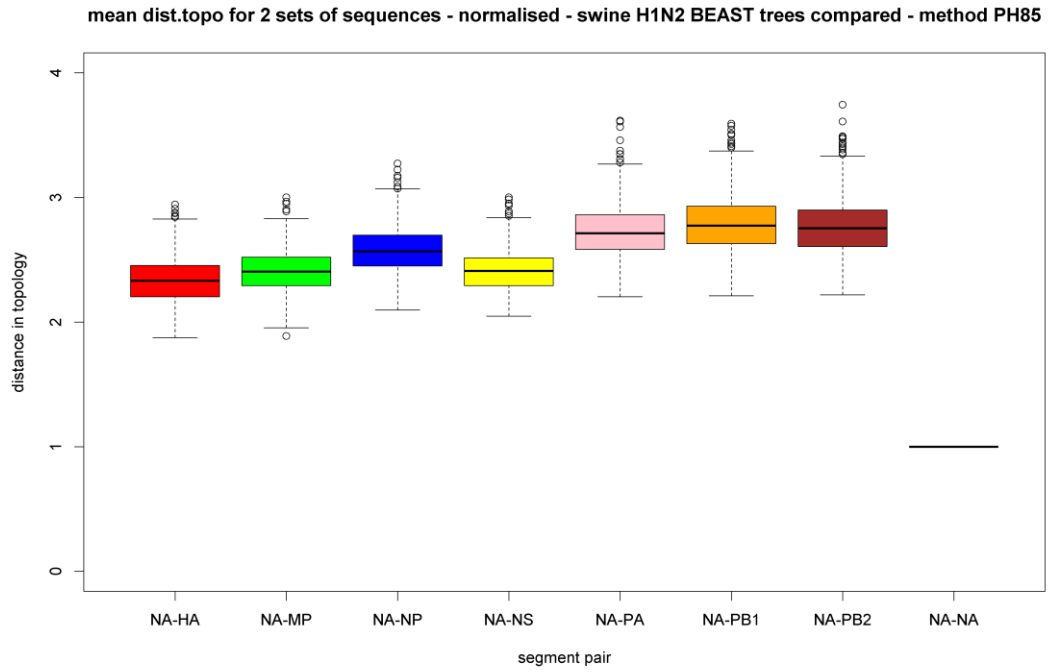

Figure B 35

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. NA compared to all other segments. The values are normalised to the value of NA.

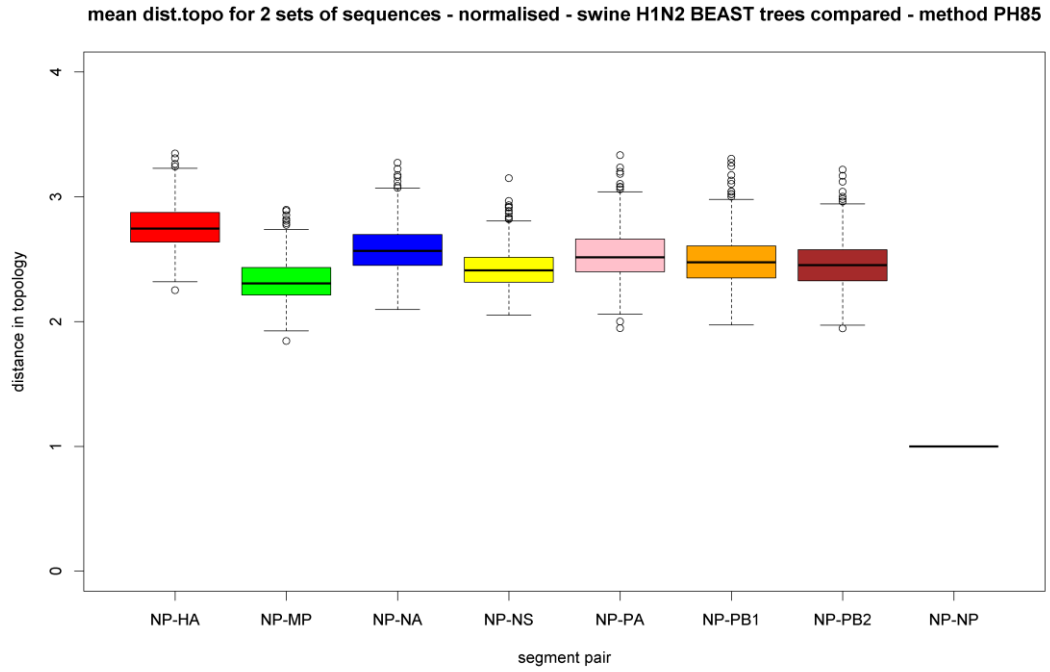

Figure B 36

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. NP compared to all other segments. The values are normalised to the value of NP.

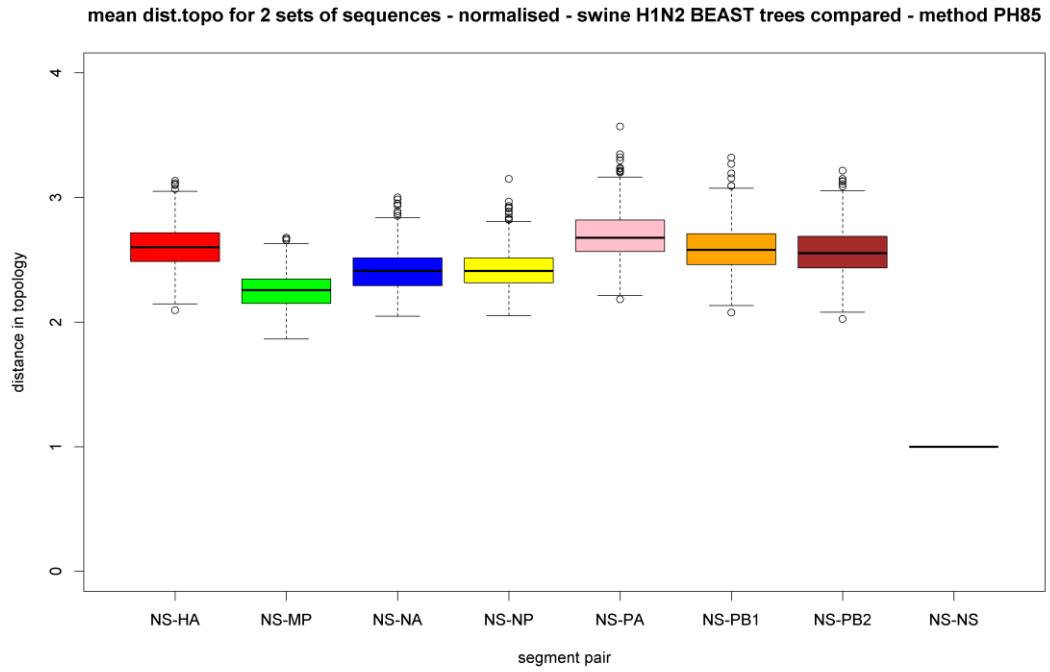

Figure B 37

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. NS compared to all other segments. The values are normalised to the value of NS.

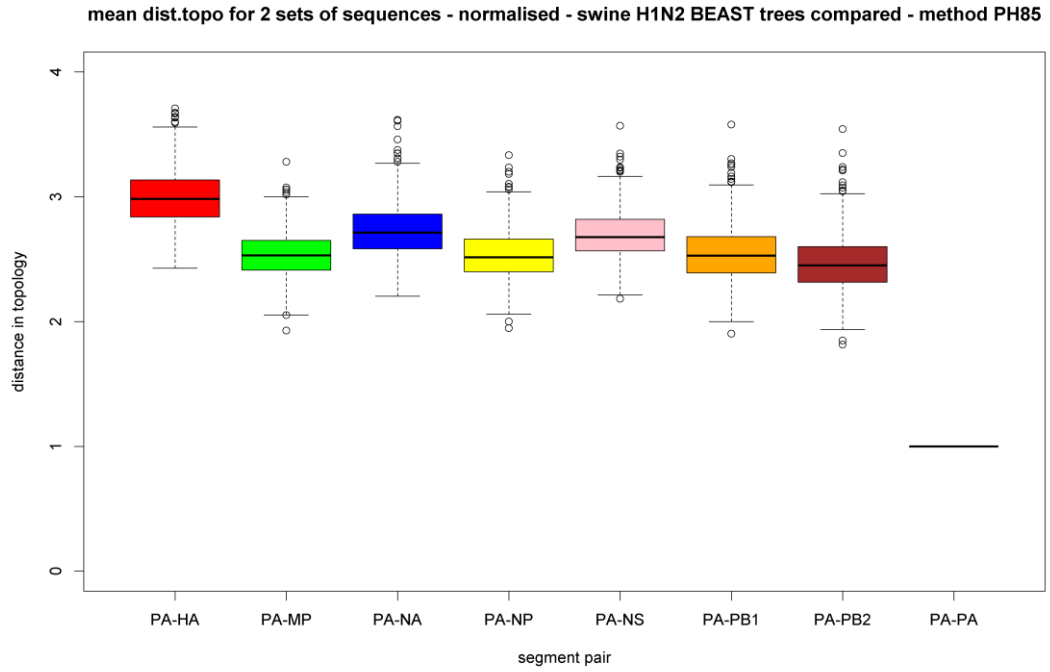

Figure B 38

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. PA compared to all other segments. The values are normalised to the value of PA.

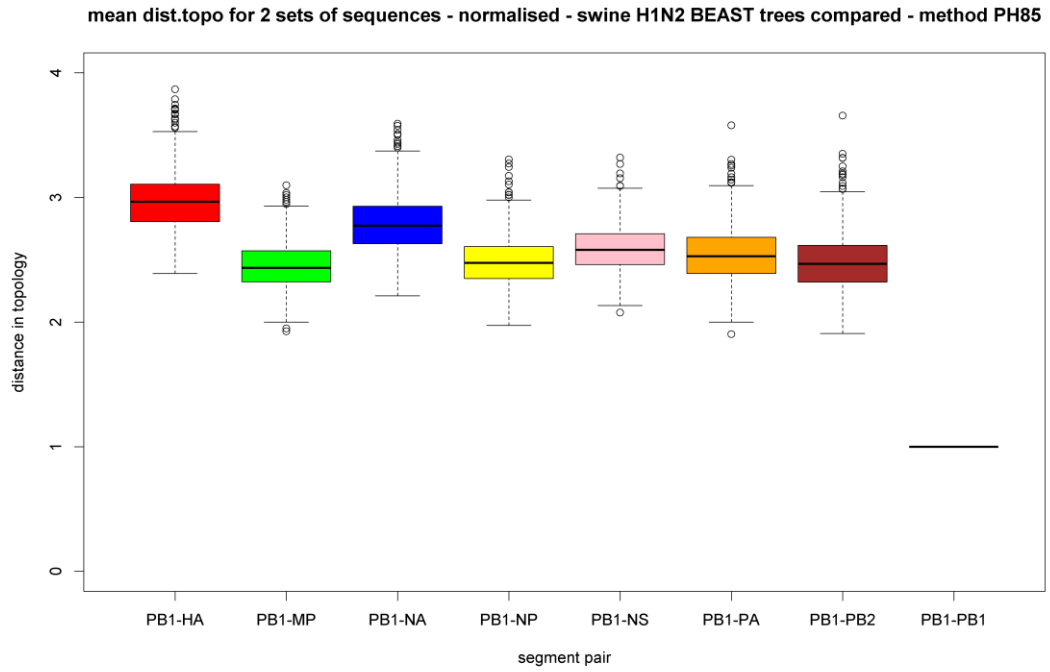

Figure B 39

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. PB1 compared to all other segments. The values are normalised to the value of PB1.

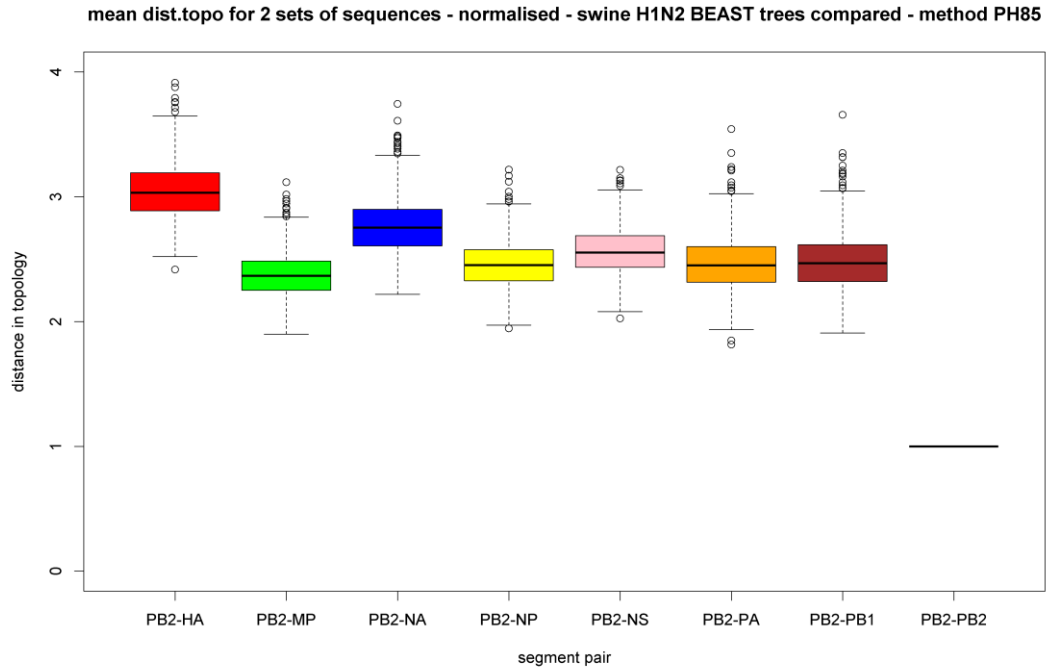

Figure B 40

Mean distance in topology for 2 sets of swine H1N2 BEAST trees compared. PB2 compared to all other segments. The values are normalised to the value of PB2.

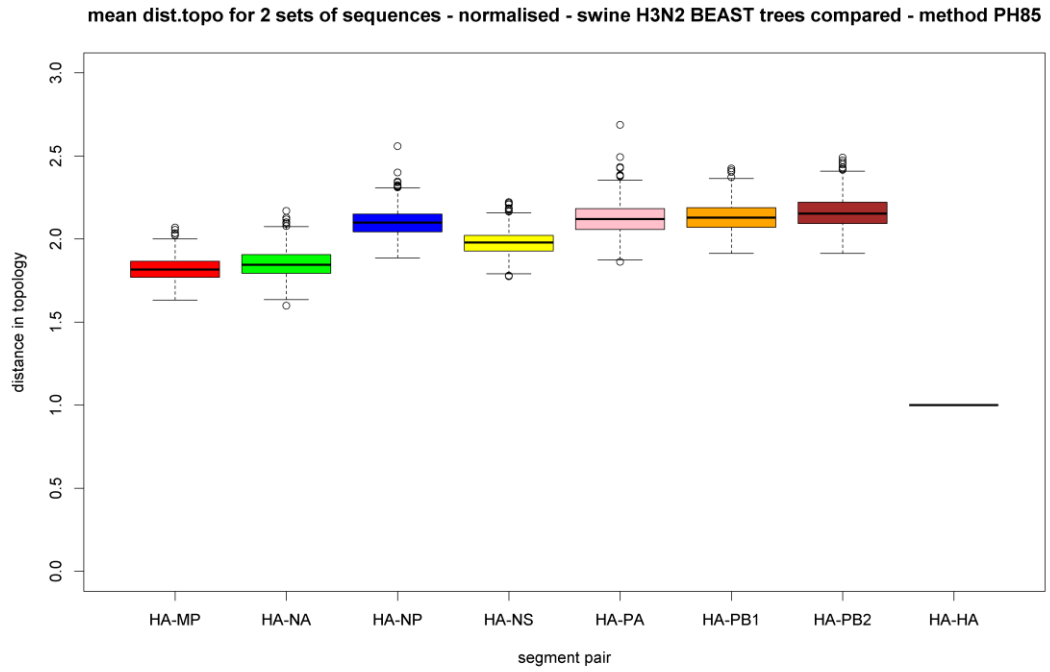

Figure B 41

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. HA compared to all other segments. The values are normalised to the value of HA.

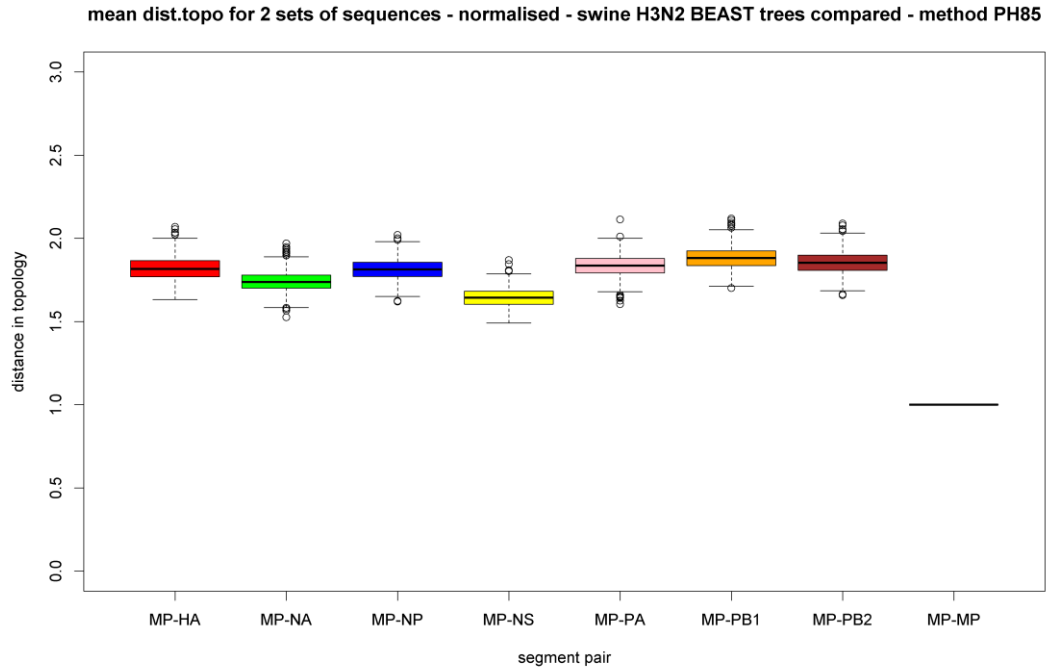

Figure B 42

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. MP compared to all other segments. The values are normalised to the value of MP.

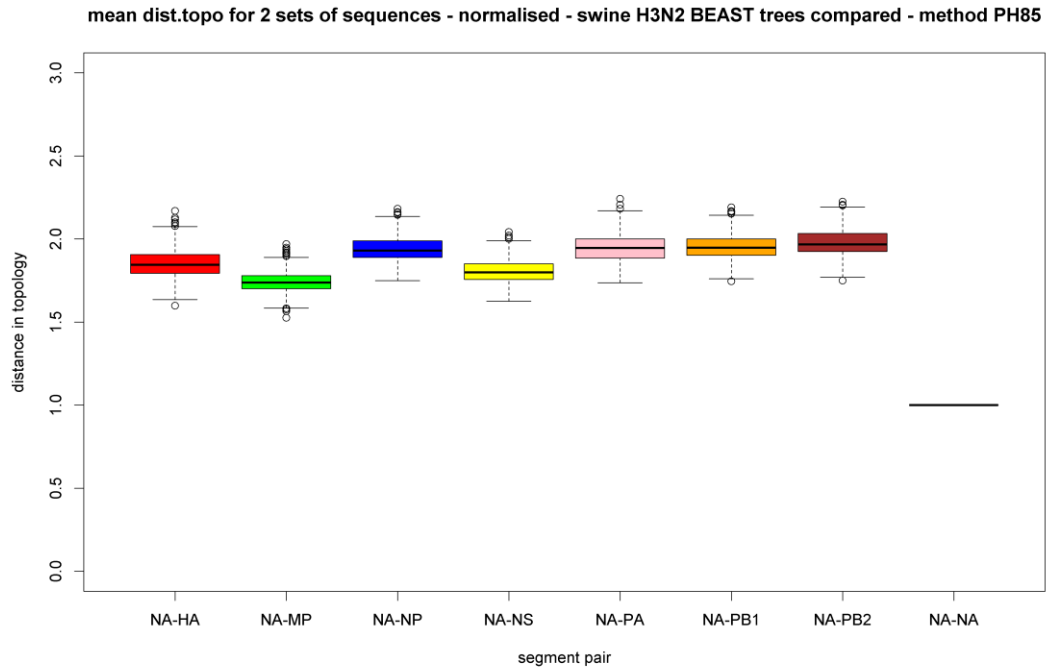

Figure B 43

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. NA compared to all other segments. The values are normalised to the value of NA.

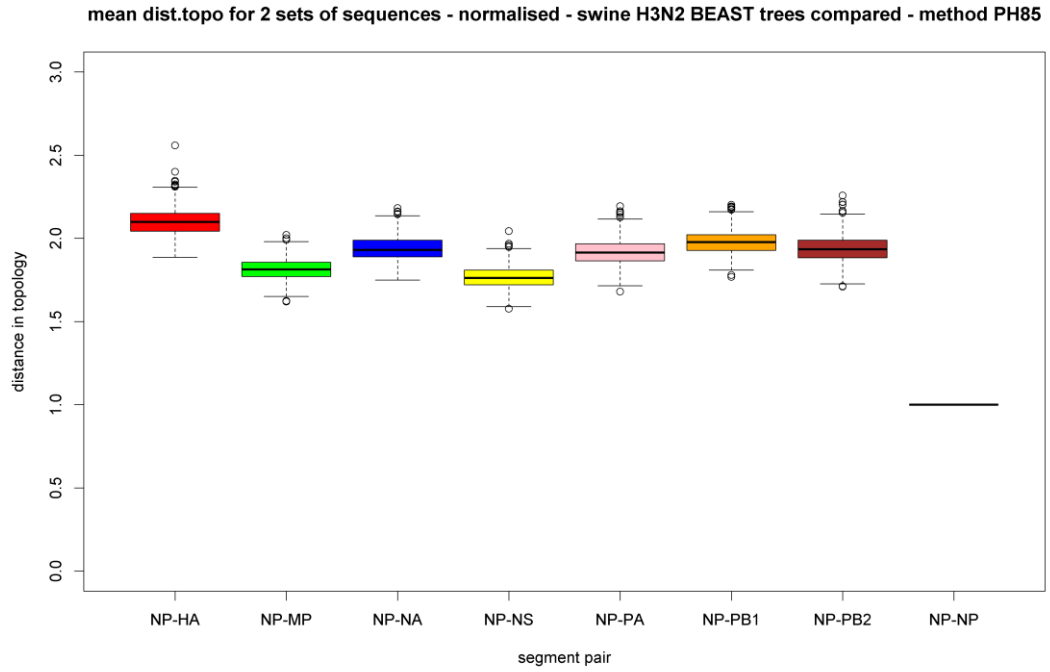

Figure B 44

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. NP compared to all other segments. The values are normalised to the value of NP.

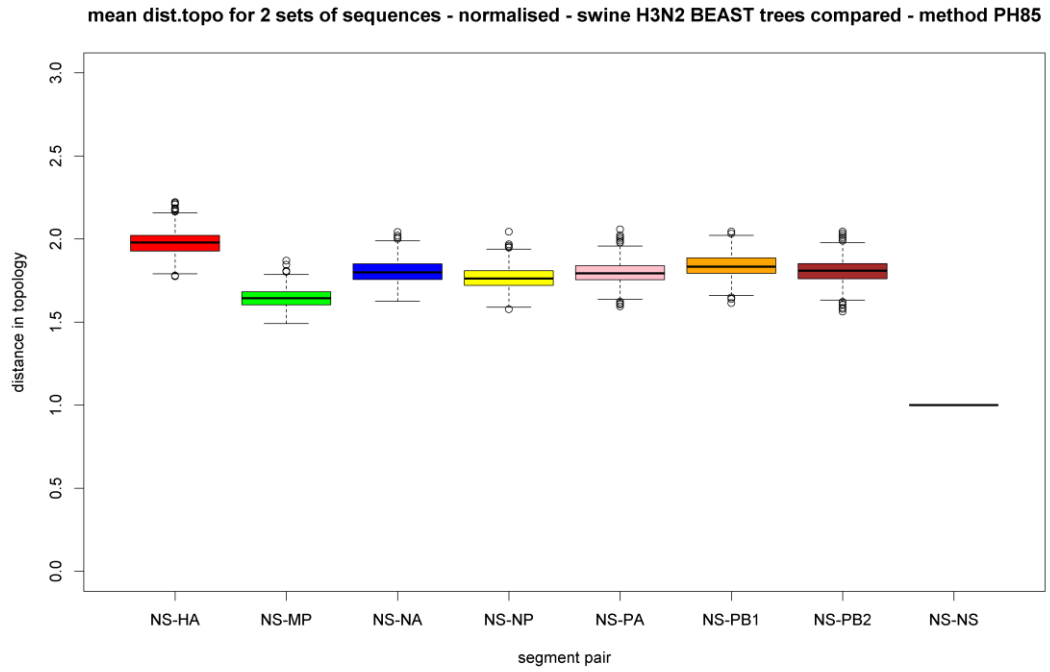

Figure B 45

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. NS compared to all other segments. The values are normalised to the value of NS.

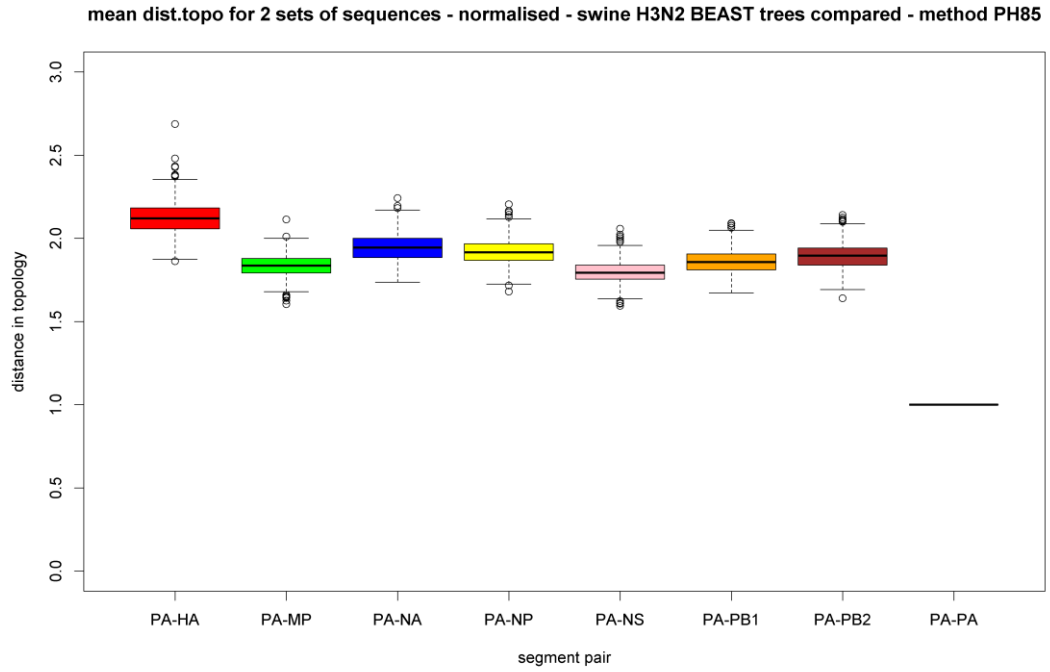

Figure B 46

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. PA compared to all other segments. The values are normalised to the value of PA.

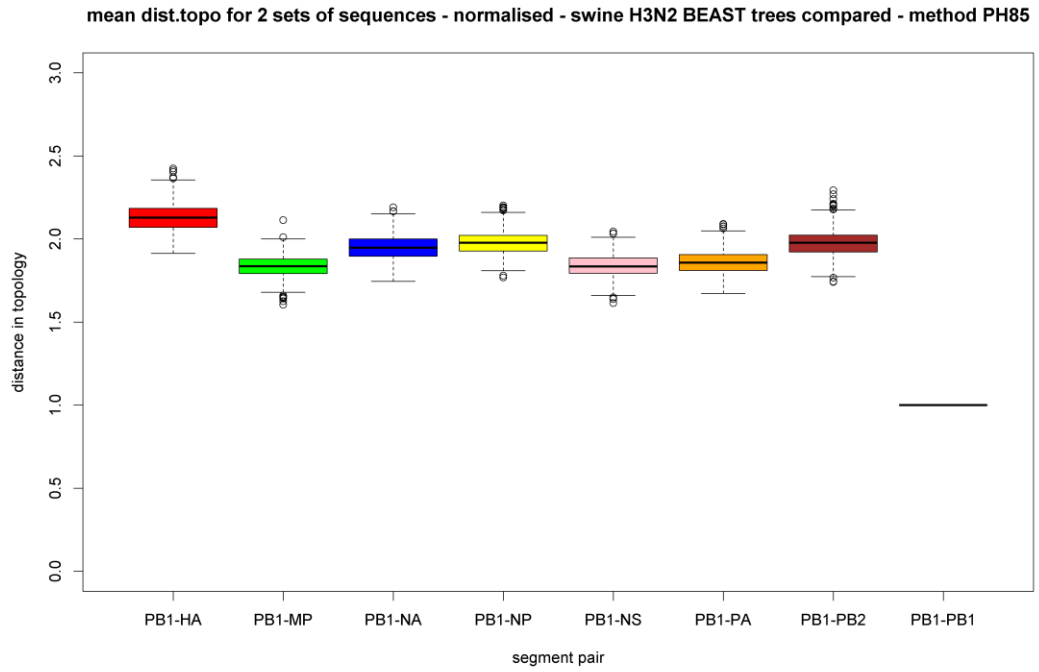

Figure B 47

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. PB1 compared to all other segments. The values are normalised to the value of PB1.

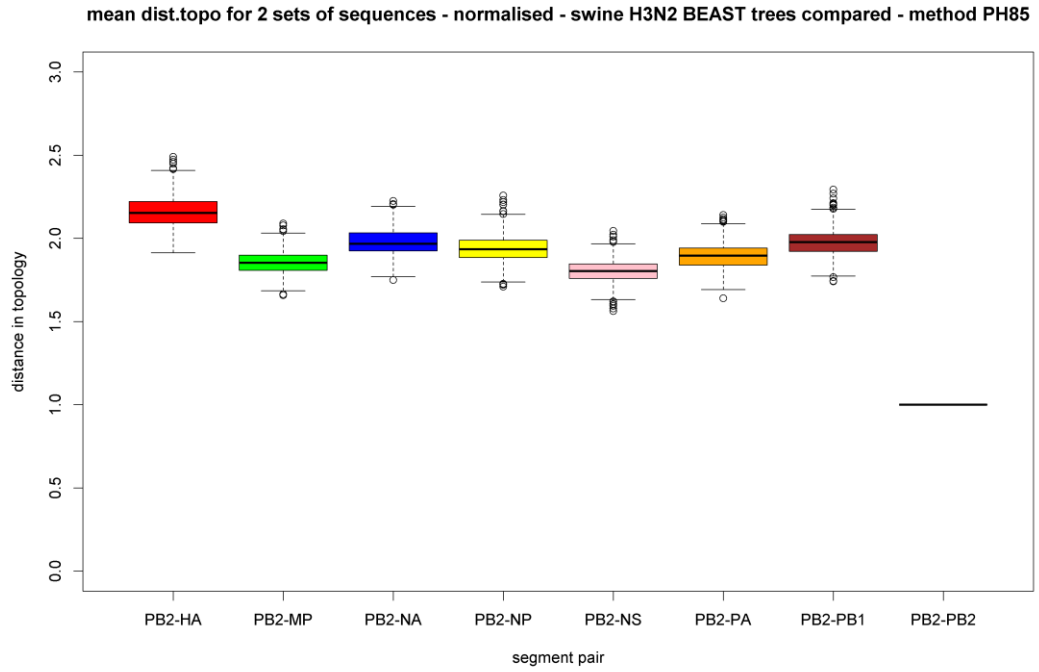

Figure B 48

Mean distance in topology for 2 sets of swine H3N2 BEAST trees compared. PB2 compared to all other segments. The values are normalised to the value of PB2.

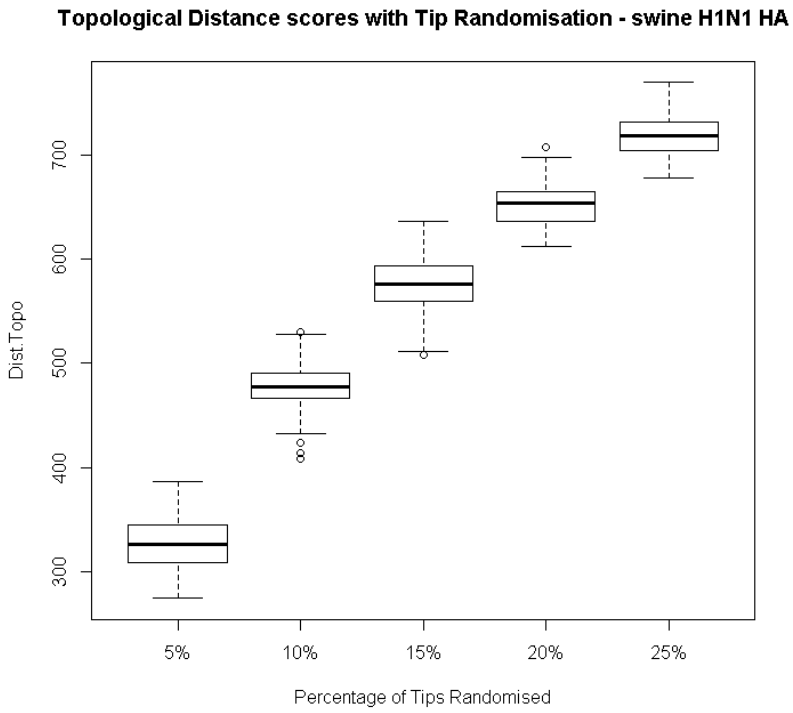

Figure B 49

Distance in topology for swine H1N1 HA if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

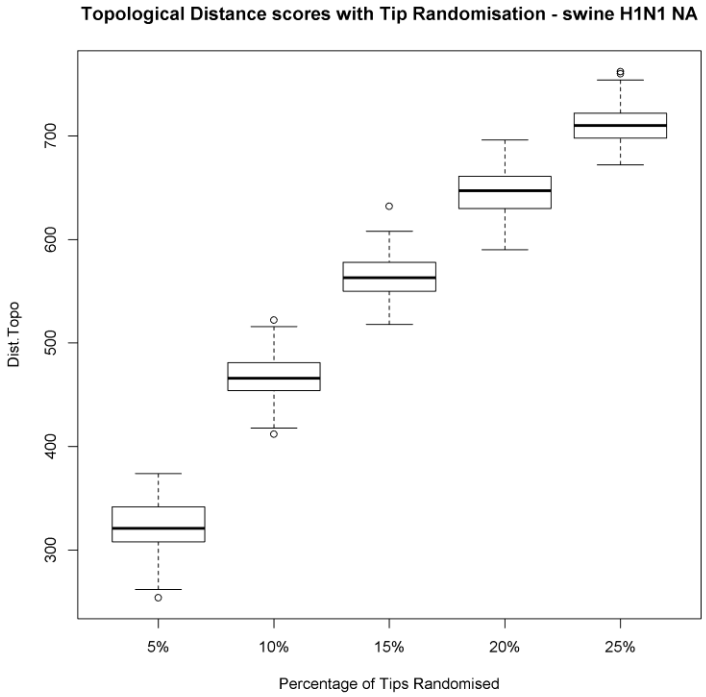

Figure B 50

Distance in topology for swine H1N1 NA if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

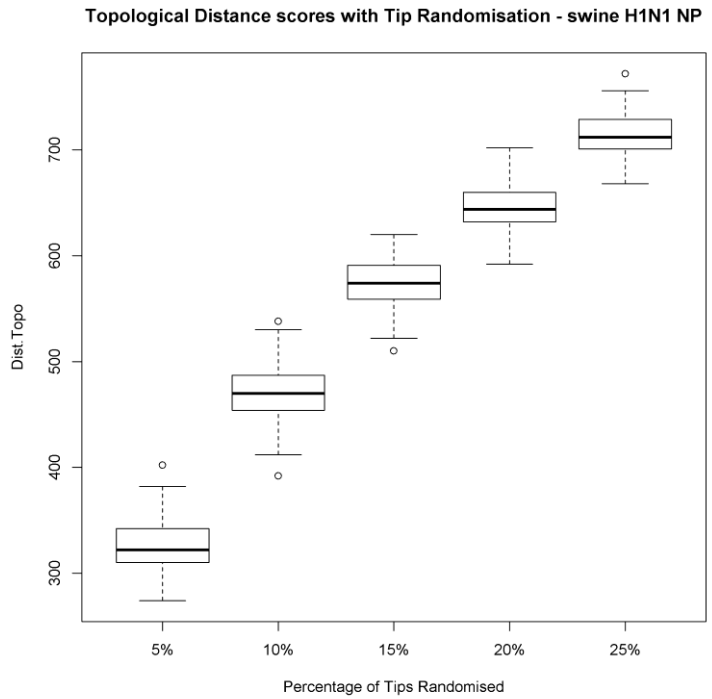

Figure B 51

Distance in topology for swine H1N1 NP if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

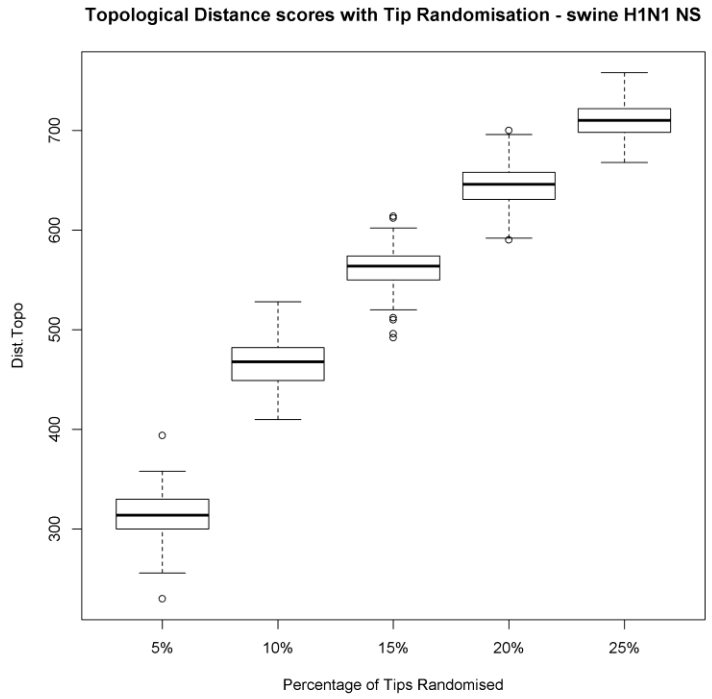

Figure B 52

Distance in topology for swine H1N1 NS if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

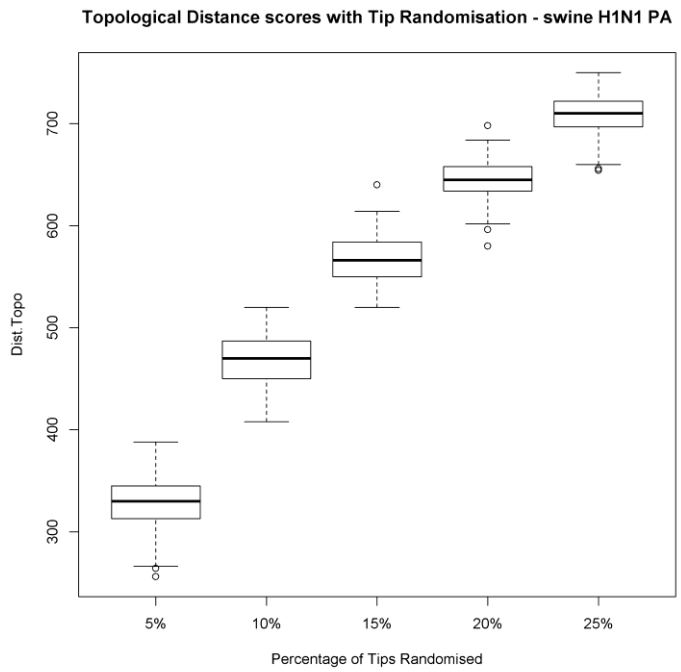

Figure B 53

Distance in topology for swine H1N1 PA if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

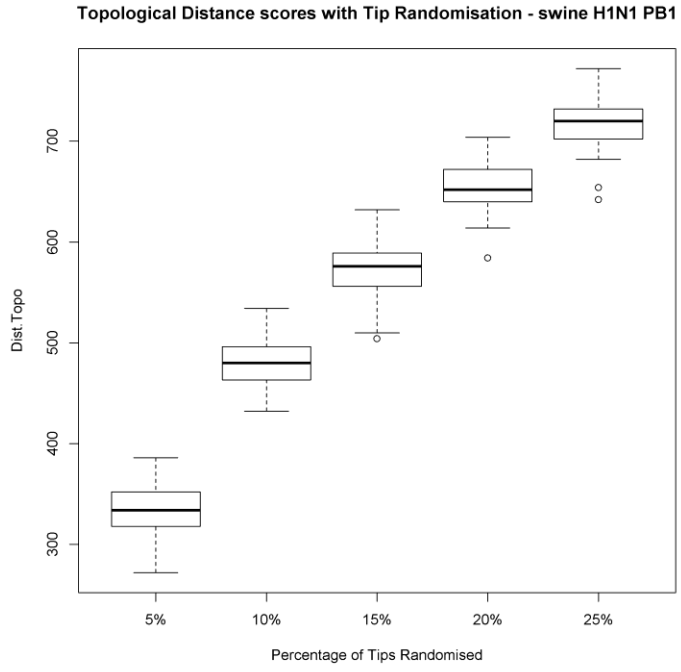

Figure B 54

Distance in topology for swine H1N1 PB1 if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

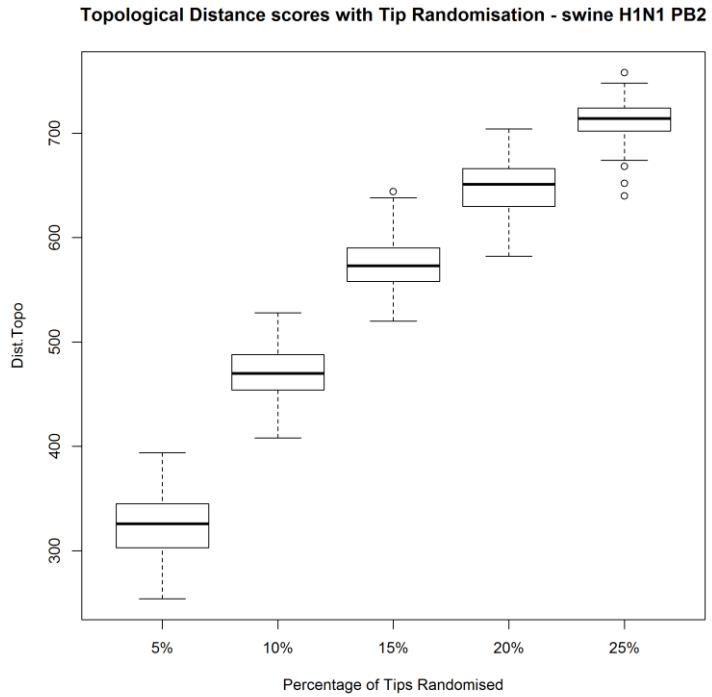

Figure B 55

Distance in topology for swine H1N1 PB2 if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

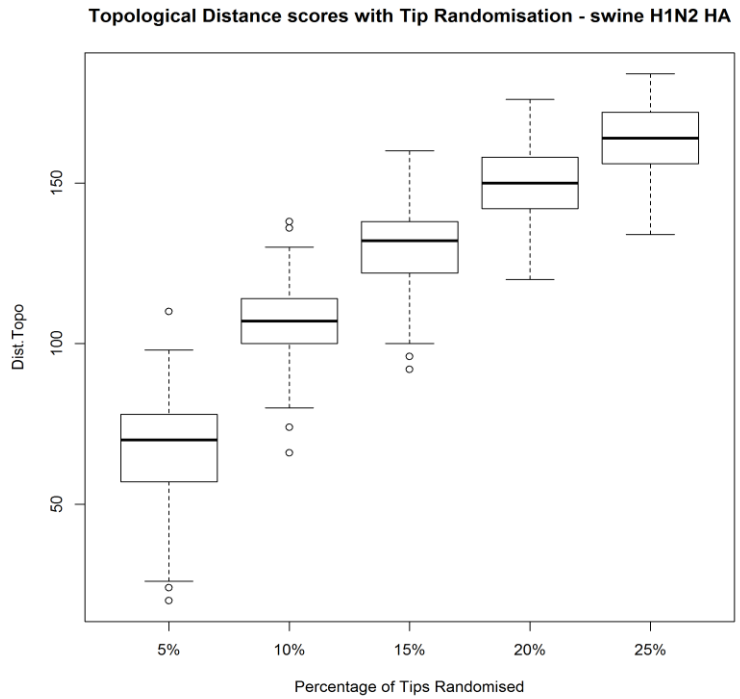

Figure B 56

Distance in topology for swine H1N2 HA if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

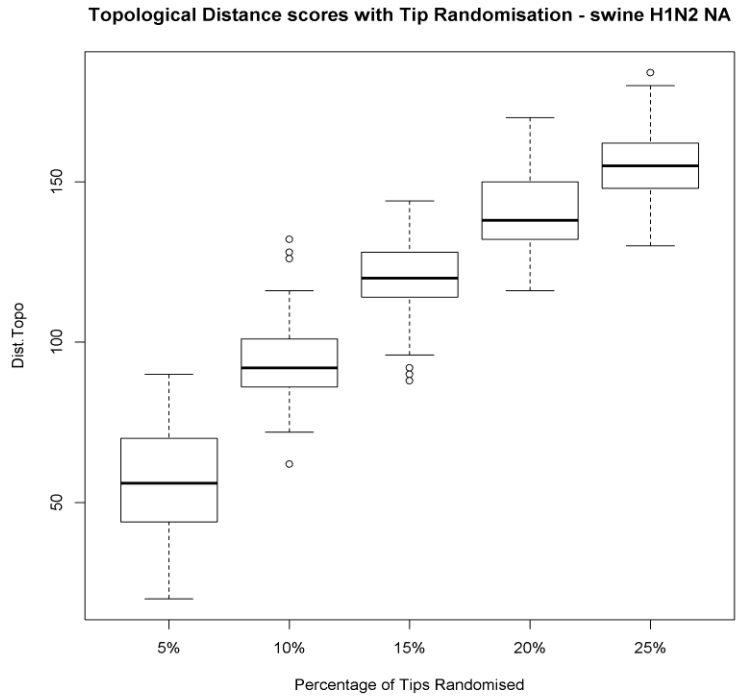

Figure B 57

Distance in topology for swine H1N2 NA if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

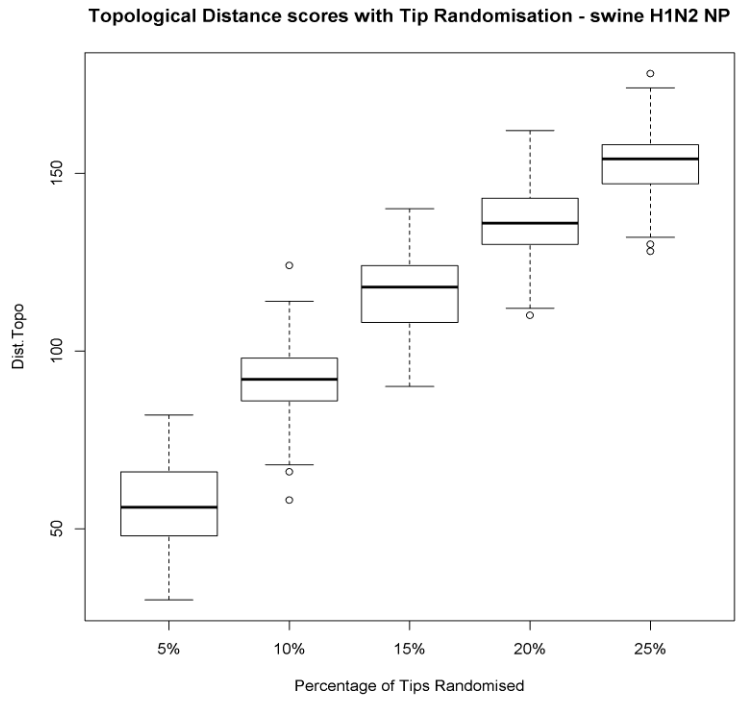

Figure B 58

Distance in topology for swine H1N2 NP if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

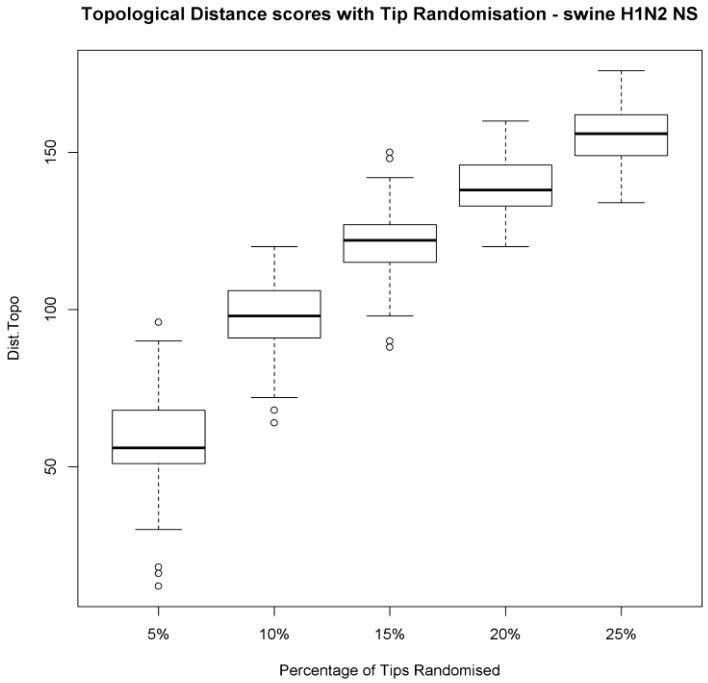

Figure B 59

Distance in topology for swine H1N2 NS if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

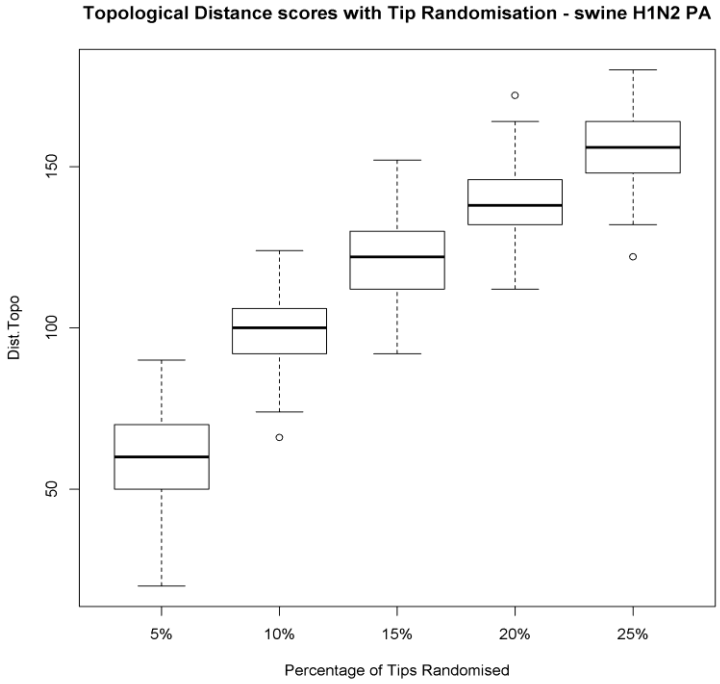

Figure B 60

Distance in topology for swine H1N2 PA if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

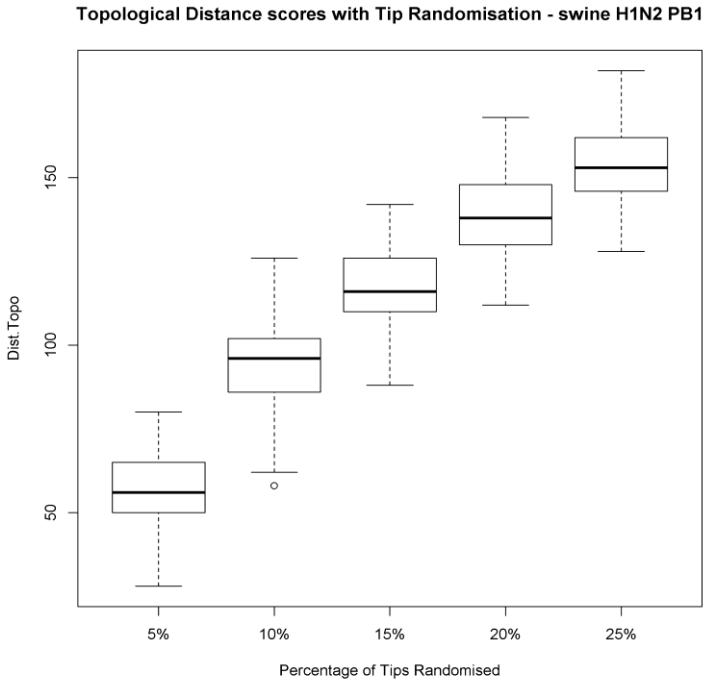

Figure B 61

Distance in topology for swine H1N2 PB1 if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

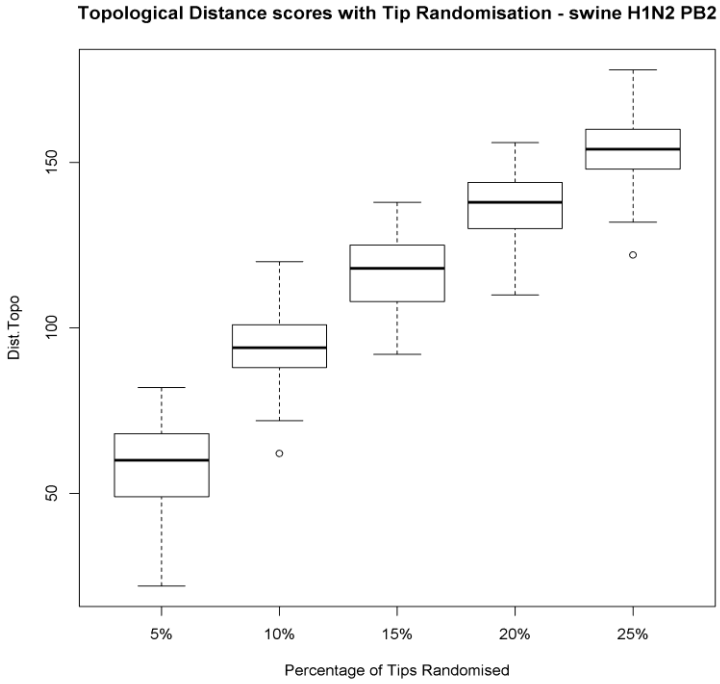

Figure B 62

Distance in topology for swine H1N2 PB2 if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

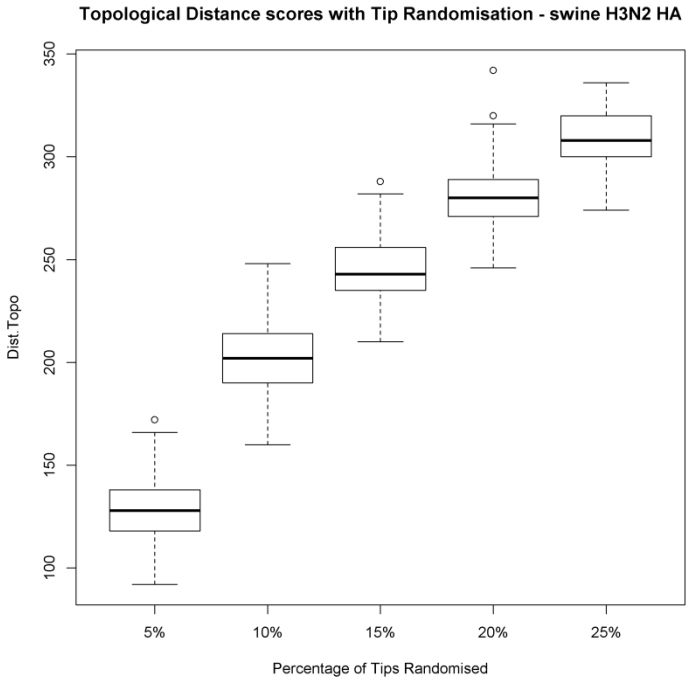

Figure B 63

Distance in topology for swine H3N2 HA if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

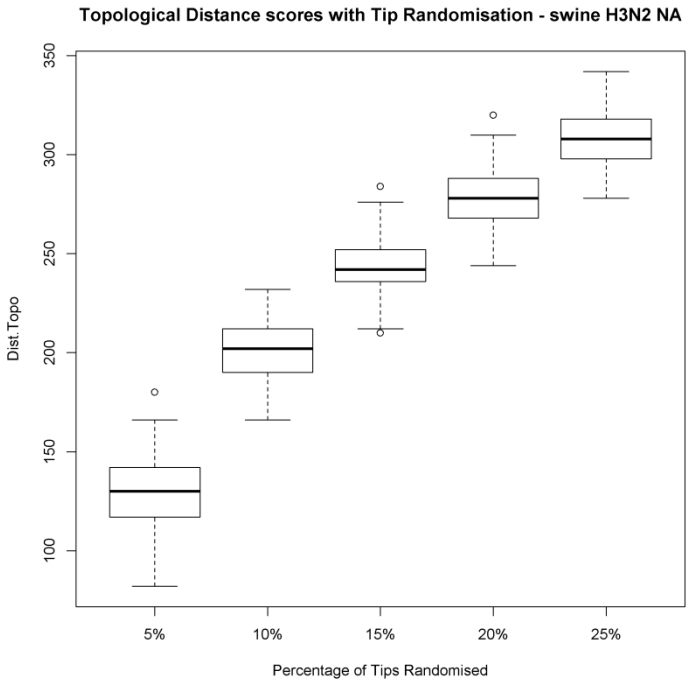

Figure B 64

Distance in topology for swine H3N2 NA if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

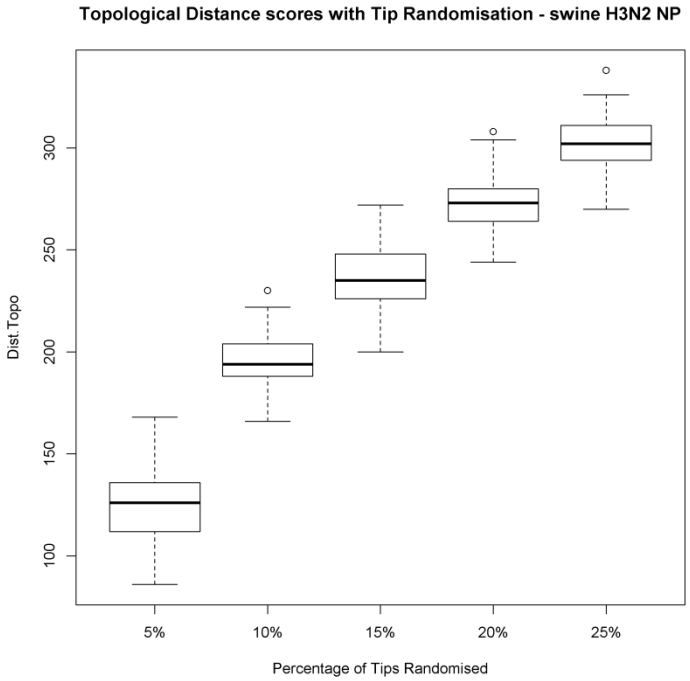

Figure B 65

Distance in topology for swine H3N2 NP if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

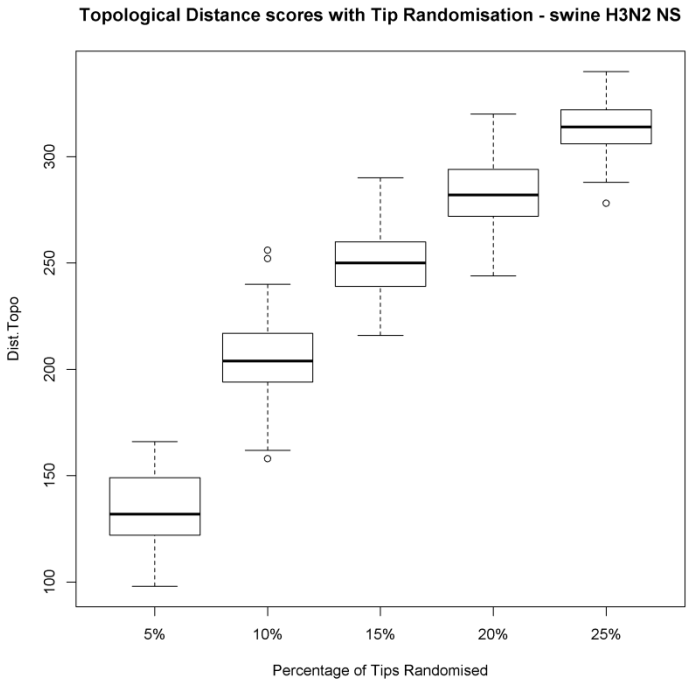

Figure B 66

Distance in topology for swine H3N2 NS if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

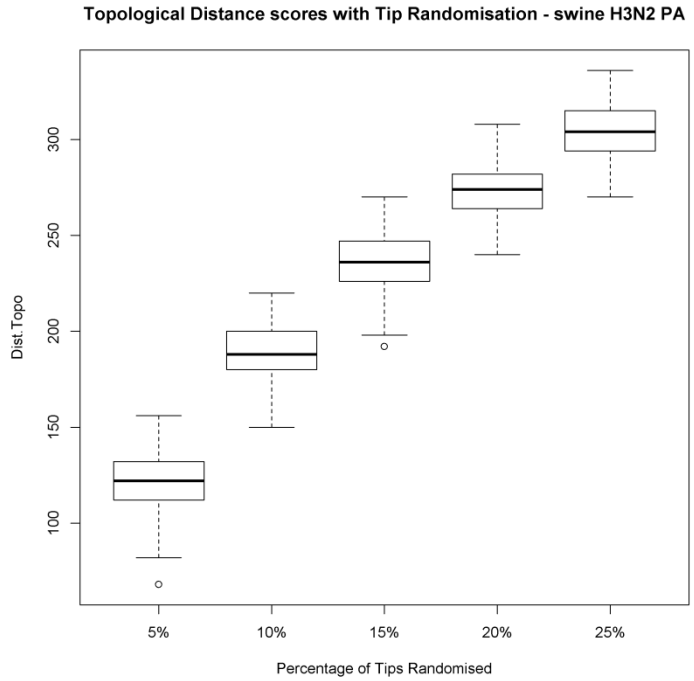

Figure B 67

Distance in topology for swine H3N2 PA if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

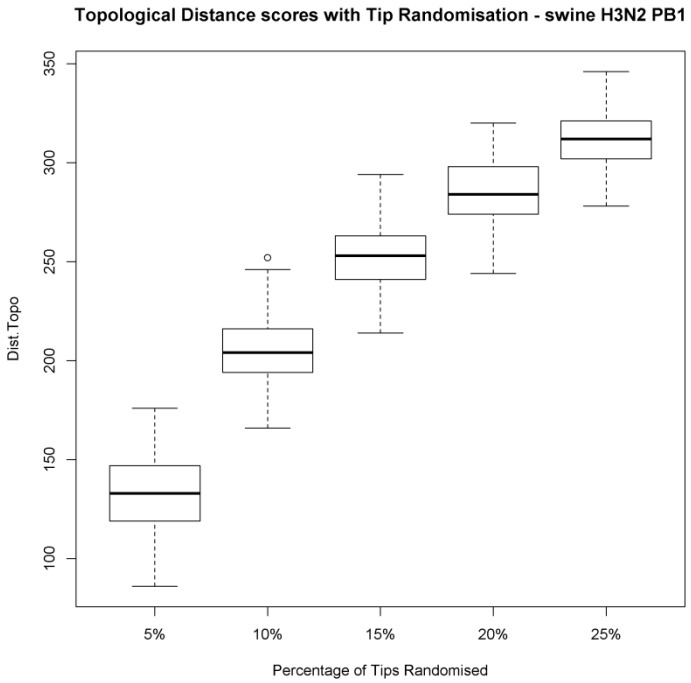

Figure B 68

Distance in topology for swine H3N2 PB1 if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

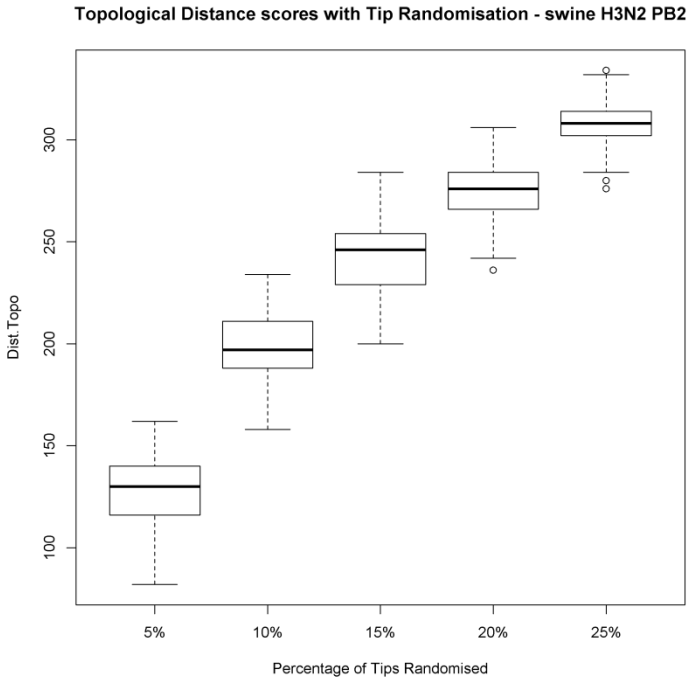

Figure B 69

Distance in topology for swine H3N2 PB2 if tips are randomized. The percentage of randomized tips is an approximation for reassortment.

Table B 4

Values from the Wilcoxon test for HA H1N1 pairs.

W	HA-HA	HA-MP	HA-NA	HA-NP	HA-NS	HA-PA	HA-PB1	HA-PB2
HA-HA	405000	0	0	0	0	0	0	0
HA-MP	810000	405000	29189.5	25010.5	167858.5	4378	1135	14036.5
HA-NA	810000	780810.5	405000	419967	699967	135416.5	13012.5	99730
HA-NP	810000	784989.5	390033	405000	699866	120562	14035	91117
HA-NS	810000	642141.5	110033	110134	405000	17918.5	2380.5	28193.5
HA-PA	810000	805622	674583.5	689438	792081.5	405000	50933.5	304333.5
HA-PB1	810000	808865	796987.5	795965	807619.5	759066.5	405000	717426
HA-PB2	810000	795963.5	710270	718883	781806.5	505666.5	92574	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 5

Values from the Wilcoxon test for MP H1N1 pairs.

W	MP-HA	MP-MP	MP-NA	MP-NP	MP-NS	MP-PA	MP-PB1	MP-PB2
MP-HA	405000	810000	374520	715369.5	800054.5	615784	213516	411706
MP-MP	0	405000	0	0	0	0	0	0
MP-NA	435480	810000	405000	742924.5	807088	650819.5	235364.5	442185.5
MP-NP	94630.5	810000	67075.5	405000	731708.5	222168	24120.5	99283
MP-NS	9945.5	810000	2912	78291.5	405000	22912.5	277	8681.5
MP-PA	194216	810000	159180.5	587832	787087.5	405000	63217	207179
MP-PB1	596484	810000	574635.5	785879.5	809723	746783	405000	594096
MP-PB2	398294	810000	367814.5	710717	801318.5	602821	215904	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 6

Values from the Wilcoxon test for NA H1N1 pairs.

W	NA-HA	NA-MP	NA-NA	NA-NP	NA-NS	NA-PA	NA-PB1	NA-PB2
NA-HA	405000	780768	810000	484557.5	751082.5	184474	14071.5	61208
NA-MP	29232	405000	810000	27050.5	283732.5	1157	1	2613.5
NA-NA	0	0	405000	0	0	0	0	0
NA-NP	325442.5	782949.5	810000	405000	742350.5	112924.5	7860	34673
NA-NS	58917.5	526267.5	810000	67649.5	405000	5477	35	4444
NA-PA	625526	808843	810000	697075.5	804523	405000	39507.5	163835
NA-PB1	795928.5	809999	810000	802140	809965	770492.5	405000	634230
NA-PB2	748792	807386.5	810000	775327	805556	646165	175770	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 7

Values from the Wilcoxon test for NP H1N1 pairs.

W	NP-HA	NP-MP	NP-NA	NP-NP	NP-NS	NP-PA	NP-PB1	NP-PB2
NP-HA	405000	804404.5	472328.5	810000	797992.5	747600	368906	635246.5
NP-MP	5595.5	405000	4123.5	810000	243366.5	36560	3019.5	13572.5
NP-NA	337671.5	805876.5	405000	810000	799397.5	736927.5	295353.5	595637.5
NP-NP	0	0	0	405000	0	0	0	0
NP-NS	12007.5	566633.5	10602.5	810000	405000	101492.5	6441	35800.5
NP-PA	62400	773440	73072.5	810000	708507.5	405000	40728.5	198374.5
NP-PB1	441094	806980.5	514646.5	810000	803559	769271.5	405000	672882
NP-PB2	174753.5	796427.5	214362.5	810000	774199.5	611625.5	137118	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 8

Values from the Wilcoxon test for NS H1N1 pairs.

W	NS-HA	NS-MP	NS-NA	NS-NP	NS-NS	NS-PA	NS-PB1	NS-PB2
NS-HA	405000	806872	519656.5	738592.5	810000	559944.5	186958.5	443634.5
NS-MP	3128	405000	1295	34118.5	810000	1548	4	2438
NS-NA	290343.5	808705	405000	702774	810000	451155.5	101967	329276.5
NS-NP	71407.5	775881.5	107226	405000	810000	135144.5	13777.5	84565
NS-NS	0	0	0	0	405000	0	0	0
NS-PA	250055.5	808452	358844.5	674855.5	810000	405000	78680	287444.5
NS-PB1	623041.5	809996	708033	796222.5	810000	731320	405000	652273.5
NS-PB2	366365.5	807562	480723.5	725435	810000	522555.5	157726.5	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 9

Values from the Wilcoxon test for PA H1N1 pairs.

W	PA-HA	PA-MP	PA-NA	PA-NP	PA-NS	PA-PA	PA-PB1	PA-PB2
PA-HA	405000	809543.5	491001	796919.5	805355	810000	564271	781449
PA-MP	456.5	405000	104	89114.5	124054	810000	278	28503.5
PA-NA	318999	809896	405000	797143	806632	810000	480844	772871
PA-NP	13080.5	720885.5	12857	405000	459527	810000	16950	209713.5
PA-NS	4645	685946	3368	350473	405000	810000	5976.5	170310.5
PA-PA	0	0	0	0	0	405000	0	0
PA-PB1	245729	809722	329156	793050	804023.5	810000	405000	755547.5
PA-PB2	28551	781496.5	37129	600286.5	639689.5	810000	54452.5	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 10

Values from the Wilcoxon test for PB1 H1N1 pairs.

W	PB1-HA	PB1-MP	PB1-NA	PB1-NP	PB1-NS	PB1-PA	PB1-PB1	PB1-PB2
PB1-HA	405000	809780	519265	796765	807379.5	790507	810000	773969.5
PB1-MP	220	405000	192.5	37701.5	113215.5	10664	810000	12675.5
PB1-NA	290735	809807.5	405000	797193.5	808550.5	785489.5	810000	758131
PB1-NP	13235	772298.5	12806.5	405000	585896	235471.5	810000	171169
PB1-NS	2620.5	696784.5	1449.5	224104	405000	101617	810000	72959.5
PB1-PA	19493	799336	24510.5	574528.5	708383	405000	810000	311334
PB1-PB1	0	0	0	0	0	0	405000	0
PB1-PB2	36030.5	797324.5	51869	638831	737040.5	498666	810000	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 11

Values from the Wilcoxon test for PB2 H1N1 pairs.

W	PB2-HA	PB2-MP	PB2-NA	PB2-NP	PB2-NS	PB2-PA	PB2-PB1	PB2-PB2
PB2-HA	405000	797148.5	339177.5	768698.5	786636.5	775099.5	566041	810000
PB2-MP	12851.5	405000	3115.5	100718	202488.5	108383.5	5429	810000
PB2-NA	470822.5	806884.5	405000	790670	803196.5	796939.5	626451.5	810000
PB2-NP	41301.5	709282	19330	405000	545306	414784	49257	810000
PB2-NS	23363.5	607511.5	6803.5	264694	405000	275589	19019	810000
PB2-PA	34900.5	701616.5	13060.5	395216	534411	405000	42948.5	810000
PB2-PB1	243959	804571	183548.5	760743	790981	767051.5	405000	810000
PB2-PB2	0	0	0	0	0	0	0	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 12

Values from the Wilcoxon test for HA H1N2 pairs.

W	HA-HA	HA-MP	HA-NA	HA-NP	HA-NS	HA-PA	HA-PB1	HA-PB2
HA-HA	405000	0	0	0	0	0	0	0
HA-MP	810000	405000	608813.5	114538.5	262773.5	24809	36152	19359
HA-NA	810000	201186.5	405000	35890.5	108092	5730	9433	4354.5
HA-NP	810000	695461.5	774109.5	405000	600058.5	161251	193211.5	130605
HA-NS	810000	547226.5	701908	209941.5	405000	55036.5	75251.5	43806.5
HA-PA	810000	785191	804270	648749	754963.5	405000	435521	354484
HA-PB1	810000	773848	800567	616788.5	734748.5	374479	405000	327217
HA-PB2	810000	790641	805645.5	679395	766193.5	455516	482783	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 13

Values from the Wilcoxon test for MP H1N2 pairs.

W	MP-HA	MP-MP	MP-NA	MP-NP	MP-NS	MP-PA	MP-PB1	MP-PB2
MP-HA	405000	810000	530751	631942	706201	352604	477231.5	564131
MP-MP	0	405000	0	0	0	0	0	0
MP-NA	279249	810000	405000	520487.5	612364	238922.5	356652	448516
MP-NP	178058	810000	289512.5	405000	506448.5	151165	251731	338020.5
MP-NS	103799	810000	197636	303551.5	405000	87929.5	169404.5	245989.5
MP-PA	457396	810000	571077.5	658835	722070.5	405000	520632	598159
MP-PB1	332768.5	810000	453348	558269	640595.5	289368	405000	490862
MP-PB2	245869	810000	361484	471979.5	564010.5	211841	319138	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 14

Values from the Wilcoxon test for NA H1N2 pairs.

W	NA-HA	NA-MP	NA-NA	NA-NP	NA-NS	NA-PA	NA-PB1	NA-PB2
NA-HA	405000	314011.5	810000	137194.5	306723.5	57651	40826.5	46692.5
NA-MP	495988.5	405000	810000	196662.5	397253.5	88188.5	63651	73332.5
NA-NA	0	0	405000	0	0	0	0	0
NA-NP	672805.5	613337.5	810000	405000	605778	240263	192882	211710
NA-NS	503276.5	412746.5	810000	204222	405000	95092	69784	79352.5
NA-PA	752349	721811.5	810000	569737	714908	405000	345658	369469
NA-PB1	769173.5	746349	810000	617118	740216	464342	405000	427877
NA-PB2	763307.5	736667.5	810000	598290	730647.5	440531	382123	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 15

Values from the Wilcoxon test for NP H1N2 pairs.

W	NP-HA	NP-MP	NP-NA	NP-NP	NP-NS	NP-PA	NP-PB1	NP-PB2
NP-HA	405000	780973	613237.5	810000	744064	651213.5	684687	708358
NP-MP	29027	405000	119654	810000	270118.5	161704	214693.5	236191.5
NP-NA	196762.5	690346	405000	810000	599248	461351	516450	547918
NP-NP	0	0	0	405000	0	0	0	0
NP-NS	65936	539881.5	210752	810000	405000	265921.5	328041	356785.5
NP-PA	158786.5	648296	348649	810000	544078.5	405000	462215	493231.5
NP-PB1	125313	595306.5	293550	810000	481959	347785	405000	434482.5
NP-PB2	101642	573808.5	262082	810000	453214.5	316768.5	375517.5	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 16

Values from the Wilcoxon test for NS H1N2 pairs.

W	NS-HA	NS-MP	NS-NA	NS-NP	NS-NS	NS-PA	NS-PB1	NS-PB2
NS-HA	405000	764627.5	640819	636385.5	810000	290162	417945	460257
NS-MP	45372.5	405000	190408	176664	810000	22749.5	61378	72418
NS-NA	169181	619592	405000	392348	810000	104056	192340.5	220492.5
NS-NP	173614.5	633336	417652	405000	810000	106110.5	197578	227063.5
NS-NS	0	0	0	0	405000	0	0	0
NS-PA	519838	787250.5	705944	703889.5	810000	405000	524924.5	561665.5
NS-PB1	392055	748622	617659.5	612422	810000	285075.5	405000	443936
NS-PB2	349743	737582	589507.5	582936.5	810000	248334.5	366064	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 17

Values from the Wilcoxon test for PA H1N2 pairs.

W	PA-HA	PA-MP	PA-NA	PA-NP	PA-NS	PA-PA	PA-PB1	PA-PB2
PA-HA	405000	770536.5	657342.5	765965.5	694328	810000	751668.5	774678.5
PA-MP	39463.5	405000	198640	413679.5	220998	810000	410090.5	495057.5
PA-NA	152657.5	611360	405000	610986.5	444882.5	810000	598036	658938
PA-NP	44034.5	396320.5	199013.5	405000	220459	810000	402155.5	486065
PA-NS	115672	589002	365117.5	589541	405000	810000	575988.5	643336
PA-PA	0	0	0	0	0	405000	0	0
PA-PB1	58331.5	399909.5	211964	407844.5	234011.5	810000	405000	483117.5
PA-PB2	35321.5	314942.5	151062	323935	166664	810000	326882.5	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 18

Values from the Wilcoxon test for PB1 H1N2 pairs.

W	PB1-HA	PB1-MP	PB1-NA	PB1-NP	PB1-NS	PB1-PA	PB1-PB1	PB1-PB2
PB1-HA	405000	782380	577643	763947	728154	736567.5	810000	758407
PB1-MP	27620	405000	92872.5	359238	231240.5	303847.5	810000	372895
PB1-NA	232357	717127.5	405000	683601.5	603780.5	638391	810000	680460.5
PB1-NP	46053	450762	126398.5	405000	278441	348660.5	810000	416191
PB1-NS	81846	578759.5	206219.5	531559	405000	470953	810000	534966
PB1-PA	73432.5	506152.5	171609	461339.5	339047	405000	810000	469208.5
PB1-PB1	0	0	0	0	0	0	405000	0
PB1-PB2	51593	437105	129539.5	393809	275034	340791.5	810000	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 19

Values from the Wilcoxon test for PB2 H1N2 pairs.

W	PB2-HA	PB2-MP	PB2-NA	PB2-NP	PB2-NS	PB2-PA	PB2-PB1	PB2-PB2
PB2-HA	405000	801383.5	655709.5	790615.5	767339	781361	777314.5	810000
PB2-MP	8616.5	405000	67675	303040	188558	305548.5	295664	810000
PB2-NA	154290.5	742325	405000	693656	612661	678569.5	667171	810000
PB2-NP	19384.5	506960	116344	405000	279098	402660.5	388745	810000
PB2-NS	42661	621442	197339	530902	405000	521028.5	504236	810000
PB2-PA	28639	504451.5	131430.5	407339.5	288971.5	405000	392644.5	810000
PB2-PB1	32685.5	514336	142829	421255	305764	417355.5	405000	810000
PB2-PB2	0	0	0	0	0	0	0	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 20

Values from the Wilcoxon test for HA H3N2 pairs.

W	HA-HA	HA-MP	HA-NA	HA-NP	HA-NS	HA-PA	HA-PB1	HA-PB2
HA-HA	405000	0	0	0	0	0	0	0
HA-MP	810000	405000	310591	2817	46173	2542	1413.5	1004
HA-NA	810000	499409	405000	13555	101772	10913.5	7870	5859.5
HA-NP	810000	807183	796445	405000	706183	344237.5	319336.5	260590
HA-NS	810000	763827	708228	103817	405000	84676.5	69049.5	51773.5
HA-PA	810000	807458	799086.5	465762.5	725323.5	405000	384362.5	321752
HA-PB1	810000	808586.5	802130	490663.5	740950.5	425637.5	405000	338847
HA-PB2	810000	808996	804140.5	549410	758226.5	488248	471153	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 21

Values from the Wilcoxon test for MP H3N2 pairs.

W	MP-HA	MP-MP	MP-NA	MP-NP	MP-NS	MP-PA	MP-PB1	MP-PB2
MP-HA	405000	810000	645103.5	424210	790235	343707.5	206090	286545
MP-MP	0	405000	0	0	0	0	0	0
MP-NA	164896.5	810000	405000	168493	708743.5	118496.5	50684	87587
MP-NP	385790	810000	641507	405000	791639	318526.5	179926	260314
MP-NS	19765	810000	101256.5	18361	405000	11863	2110.5	6323.5
MP-PA	466292.5	810000	691503.5	491473.5	798137	405000	247009.5	341619.5
MP-PB1	603910	810000	759316	630074	807889.5	562990.5	405000	502635
MP-PB2	523455	810000	722413	549686	803676.5	468380.5	307365	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 22

Values from the Wilcoxon test for NA H3N2 pairs.

W	NA-HA	NA-MP	NA-NA	NA-NP	NA-NS	NA-PA	NA-PB1	NA-PB2
NA-HA	405000	696871.5	810000	170563.5	540447.5	170915	144969.5	113089
NA-MP	113128.5	405000	810000	15577	196412.5	21117	12545.5	9376.5
NA-NA	0	0	405000	0	0	0	0	0
NA-NP	639436.5	794423	810000	405000	738205.5	385921	361565	296038
NA-NS	269552.5	613587.5	810000	71794.5	405000	78865.5	57888	43743.5
NA-PA	639085	788883	810000	424079	731134.5	405000	380684.5	317000
NA-PB1	665030.5	797454.5	810000	448435	752112	429315.5	405000	336961.5
NA-PB2	696911	800623.5	810000	513962	766256.5	493000	473038.5	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 23

Values from the Wilcoxon test for NP H3N2 pairs.

W	NP-HA	NP-MP	NP-NA	NP-NP	NP-NS	NP-PA	NP-PB1	NP-PB2
NP-HA	405000	808960	757470.5	810000	809574.5	768756.5	707629.5	755651.5
NP-MP	1040	405000	70736.5	810000	564784	118575.5	29306.5	84217
NP-NA	52529.5	739263.5	405000	810000	781081.5	471111.5	283034.5	418447.5
NP-NP	0	0	0	405000	0	0	0	0
NP-NS	425.5	245216	28918.5	810000	405000	57244.5	11754	38376.5
NP-PA	41243.5	691424.5	338888.5	810000	752755.5	405000	221446.5	352670
NP-PB1	102370.5	780693.5	526965.5	810000	798246	588553.5	405000	543010.5
NP-PB2	54348.5	725783	391552.5	810000	771623.5	457330	266989.5	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 24

Values from the Wilcoxon test for NS H3N2 pairs.

W	NS-HA	NS-MP	NS-NA	NS-NP	NS-NS	NS-PA	NS-PB1	NS-PB2
NS-HA	405000	809895.5	773793.5	797313.5	810000	780434	751906.5	770093.5
NS-MP	104.5	405000	24572	61318	810000	31154	10939	28809
NS-NA	36206.5	785428	405000	533696	810000	429716	296704	391008.5
NS-NP	12686.5	748682	276304	405000	810000	299120.5	181923	266007.5
NS-NS	0	0	0	0	405000	0	0	0
NS-PA	29566	778846	380284	510879.5	810000	405000	270837.5	366295.5
NS-PB1	58093.5	799061	513296	628077	810000	539162.5	405000	500294
NS-PB2	39906.5	781191	418991.5	543992.5	810000	443704.5	309706	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 25

Values from the Wilcoxon test for PA H3N2 pairs.

W	PA-HA	PA-MP	PA-NA	PA-NP	PA-NS	PA-PA	PA-PB1	PA-PB2
PA-HA	405000	807484	757793	775772	808603.5	810000	802296	788763.5
PA-MP	2516	405000	121374	166376.5	539463	810000	328170.5	232396.5
PA-NA	52207	688626	405000	474484	738728	810000	631373.5	541146
PA-NP	34228	643623.5	335516	405000	710458	810000	576455	476024
PA-NS	1396.5	270537	71272	99542	405000	810000	213038	145253
PA-PA	0	0	0	0	0	405000	0	0
PA-PB1	7704	481829.5	178626.5	233545	596962	810000	405000	302633
PA-PB2	21236.5	577603.5	268854	333976	664747	810000	507367	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 26

Values from the Wilcoxon test for PB1 H3N2 pairs.

W	PB1-HA	PB1-MP	PB1-NA	PB1-NP	PB1-NS	PB1-PA	PB1-PB1	PB1-PB2
PB1-HA	405000	808693	766160.5	746432.5	808751	805033	810000	740142
PB1-MP	1307	405000	93568	54453.5	403970	328170.5	810000	66275
PB1-NA	43839.5	716432	405000	318963	709850	660397.5	810000	338057.5
PB1-NP	63567.5	755546.5	491037	405000	752324	713084	810000	413745
PB1-NS	1249	406030	100150	57676	405000	332490	810000	70353
PB1-PA	4967	481829.5	149602.5	96916	477510	405000	810000	109873
PB1-PB1	0	0	0	0	0	0	405000	0
PB1-PB2	69858	743725	471942.5	396255	739647	700127	810000	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 27

Values from the Wilcoxon test for PB2 H3N2 pairs.

W	PB2-HA	PB2-MP	PB2-NA	PB2-NP	PB2-NS	PB2-PA	PB2-PB1	PB2-PB2
PB2-HA	405000	807678	755427	784131	809129.5	798920	757066	810000
PB2-MP	2322	405000	101211	171127.5	563667	286461	96901	810000
PB2-NA	54573	708789	405000	516677.5	762288.5	617867.5	404896	810000
PB2-NP	25869	638872.5	293322.5	405000	723322	525189.5	292902.5	810000
PB2-NS	870.5	246333	47711.5	86678	405000	165573.5	45982	810000
PB2-PA	11080	523539	192132.5	284810.5	644426.5	405000	186327	810000
PB2-PB1	52934	713099	405104	517097.5	764018	623673	405000	810000
PB2-PB2	0	0	0	0	0	0	0	405000

The segment pair comparisons in the table are to be read as a row against a column.

Table B 28

Normalised values from the Wilcoxon test for HA H1N1 pairs.

p-value	HA-HA	HA-MP	HA-NA	HA-NP	HA-NS	HA-PA	HA-PB1	HA-PB2
HA-HA	0.5	0	0	0	0	0	0	0
HA-MP	1	0.5	0.036	0.031	0.207	0.005	0.001	0.017
HA-NA	1	0.964	0.5	0.518	0.864	0.167	0.016	0.123
HA-NP	1	0.969	0.482	0.5	0.864	0.149	0.017	0.112
HA-NS	1	0.793	0.136	0.136	0.5	0.022	0.003	0.035
HA-PA	1	0.995	0.833	0.851	0.978	0.5	0.063	0.376
HA-PB1	1	0.999	0.984	0.983	0.997	0.937	0.5	0.886
HA-PB2	1	0.983	0.877	0.888	0.965	0.624	0.114	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 29

Normalised values from the Wilcoxon test for NA H1N1 pairs.

p-value	NA-HA	NA-MP	NA-NA	NA-NP	NA-NS	NA-PA	NA-PB1	NA-PB2
NA-HA	0.5	0.964	1	0.598	0.927	0.228	0.017	0.076
NA-MP	0.036	0.5	1	0.033	0.350	0.001	1	0.003
NA-NA	0	0	0.5	0	0	0	0	0
NA-NP	0.402	0.967	1	0.5	0.916	0.139	0.010	0.043
NA-NS	0.073	0.649	1	0.084	0.5	0.007	0	0.005
NA-PA	0.772	0.999	1	0.861	0.993	0.5	0.049	0.202
NA-PB1	0.983	0.999	1	0.990	0.999	0.951	0.5	0.783
NA-PB2	0.924	0.997	1	0.957	0.995	0.798	0.217	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 30

Normalised values from the Wilcoxon test for NP H1N1 pairs.

p-value	NP-HA	NP-MP	NP-NA	NP-NP	NP-NS	NP-PA	NP-PB1	NP-PB2
NP-HA	0.5	0.993	0.583	1	0.985	0.923	0.455	0.784
NP-MP	0.007	0.5	0.005	1	0.300	0.045	0.004	0.017
NP-NA	0.417	0.995	0.5	1	0.987	0.910	0.365	0.735
NP-NP	0	0	0	0.5	0	0	0	0
NP-NS	0.015	0.700	0.013	1	0.5	0.125	0.008	0.044
NP-PA	0.077	0.955	0.090	1	0.875	0.5	0.050	0.245
NP-PB1	0.545	0.996	0.635	1	0.992	0.950	0.5	0.831
NP-PB2	0.216	0.983	0.265	1	0.956	0.755	0.169	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 31

Normalised values from the Wilcoxon test for NS H1N1 pairs.

p-value	NS-HA	NS-MP	NS-NA	NS-NP	NS-NS	NS-PA	NS-PB1	NS-PB2
NS-HA	0.5	0.996	0.642	0.912	1	0.691	0.231	0.548
NS-MP	0.004	0.5	0.002	0.042	1	0.002	0	0.003
NS-NA	0.358	0.998	0.5	0.868	1	0.056	0.126	0.407
NS-NP	0.088	0.958	0.132	0.5	1	0.167	0.017	0.104
NS-NS	0	0	0	0	0.5	0	0	0
NS-PA	0.309	0.998	0.443	0.833	1	0.5	0.097	0.355
NS-PB1	0.769	1	0.874	0.983	1	0.903	0.5	0.805
NS-PB2	0.452	0.997	0.593	0.896	1	0.645	0.195	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 32

Normalised values from the Wilcoxon test for PA H1N1 pairs.

p-value	PA-HA	PA-MP	PA-NA	PA-NP	PA-NS	PA-PA	PA-PB1	PA-PB2
PA-HA	0.5	0.999	0.606	0.984	0.994	1	0.700	0.965
PA-MP	0	0.5	0	0.110	0.153	1	0	0.035
PA-NA	0.394	1	0.5	0.984	0.996	1	0.594	0.954
PA-NP	0.016	0.890	0.016	0.5	0.567	1	0.021	0.259
PA-NS	0.006	0.847	0.004	0.433	0.5	1	0.007	0.210
PA-PA	0	0	0	0	0	0.5	0	0
PA-PB1	0.303	1	0.406	0.979	0.993	1	0.5	0.933
PA-PB2	0.035	0.965	0.046	0.741	0.790	1	0.067	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 33

Normalised values from the Wilcoxon test for PB1 H1N1 pairs.

p-value	PB1-HA	PB1-MP	PB1-NA	PB1-NP	PB1-NS	PB1-PA	PB1-PB1	PB1-PB2
PB1-HA	0.5	1	0.641	0.984	0.997	0.976	1	0.956
PB1-MP	0	0.5	0	0.047	0.140	0.013	1	0.016
PB1-NA	0.359	1	0.5	0.984	0.998	0.970	1	0.936
PB1-NP	0.016	0.953	0.016	0.5	0.723	0.291	1	0.211
PB1-NS	0.003	0.860	0.002	0.277	0.5	0.125	1	0.090
PB1-PA	0.024	0.987	0.030	0.709	0.875	0.5	1	0.384
PB1-PB1	0	0	0	0	0	0	0.5	0
PB1-PB2	0.044	0.984	0.064	0.789	0.910	0.616	1	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 34

Normalised values from the Wilcoxon test for PB2 H1N1 pairs.

p-value	PB2-HA	PB2-MP	PB2-NA	PB2-NP	PB2-NS	PB2-PA	PB2-PB1	PB2-PB2
PB2-HA	0.5	0.984	0.419	0.949	0.971	0.957	0.700	1
PB2-MP	0.016	0.5	0.004	0.124	0.250	0.134	0.007	1
PB2-NA	0.581	0.996	0.5	0.976	0.992	0.984	0.773	1
PB2-NP	0.051	0.876	0.024	0.5	0.673	0.512	0.061	1
PB2-NS	0.029	0.750	0.008	0.327	0.5	0.340	0.023	1
PB2-PA	0.043	0.866	0.016	0.488	0.660	0.5	0.053	1
PB2-PB1	0.301	0.993	0.227	0.939	0.977	0.947	0.5	1
PB2-PB2	0	0	0	0	0	0	0	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 35

Normalised values from the Wilcoxon test for HA H1N2 pairs.

p-value	HA-HA	HA-MP	HA-NA	HA-NP	HA-NS	HA-PA	HA-PB1	HA-PB2
HA-HA	0.5	0	0	0	0	0	0	0
HA-MP	1	0.5	0.752	0.141	0.324	0.031	0.044	0.024
HA-NA	1	0.248	0.5	0.044	0.133	0.007	0.012	0.005
HA-NP	1	0.859	0.956	0.5	0.741	0.199	0.239	0.161
HA-NS	1	0.676	0.867	0.259	0.5	0.068	0.093	0.054
HA-PA	1	0.969	0.993	0.801	0.932	0.5	0.538	0.438
HA-PB1	1	0.955	0.988	0.761	0.907	0.462	0.5	0.404
HA-PB2	1	0.976	0.995	0.839	0.946	0.562	0.596	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 36

Normalised values from the Wilcoxon test for NA H1N2 pairs.

p-value	NA-HA	NA-MP	NA-NA	NA-NP	NA-NS	NA-PA	NA-PB1	NA-PB2
NA-HA	0.5	0.388	1	0.169	0.379	0.071	0.050	0.058
NA-MP	0.612	0.5	1	0.243	0.490	0.109	0.079	0.091
NA-NA	0	0	0.5	0	0	0	0	0
NA-NP	0.831	0.757	1	0.5	0.748	0.297	0.238	0.261
NA-NS	0.621	0.510	1	0.252	0.5	0.117	0.086	0.098
NA-PA	0.929	0.891	1	0.703	0.883	0.5	0.427	0.456
NA-PB1	0.946	0.921	1	0.762	0.914	0.573	0.5	0.528
NA-PB2	0.942	0.909	1	0.739	0.902	0.544	0.472	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 37

Normalised values from the Wilcoxon test for NP H1N2 pairs.

p-value	NP-HA	NP-MP	NP-NA	NP-NP	NP-NS	NP-PA	NP-PB1	NP-PB2
NP-HA	0.5	0.964	0.757	1	0.919	0.804	0.845	0.875
NP-MP	0.036	0.5	0.148	1	0.333	0.200	0.265	0.292
NP-NA	0.243	0.852	0.5	1	0.740	0.570	0.638	0.676
NP-NP	0	0	0	0.5	0	0	0	0
NP-NS	0.081	0.667	0.260	1	0.5	0.328	0.405	0.440
NP-PA	0.196	0.800	0.430	1	0.672	0.5	0.571	0.609
NP-PB1	0.155	0.735	0.362	1	0.595	0.429	0.5	0.536
NP-PB2	0.125	0.708	0.324	1	0.560	0.391	0.464	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 38

Normalised values from the Wilcoxon test for NS H1N2 pairs.

p-value	NS-HA	NS-MP	NS-NA	NS-NP	NS-NS	NS-PA	NS-PB1	NS-PB2
NS-HA	0.5	0.944	0.791	0.786	1	0.358	0.516	0.568
NS-MP	0.056	0.5	0.235	0.218	1	0.028	0.076	0.089
NS-NA	0.209	0.765	0.5	0.484	1	0.128	0.237	0.272
NS-NP	0.214	0.782	0.516	0.5	1	0.131	0.244	0.280
NS-NS	0	0	0	0	0.5	0	0	0
NS-PA	0.642	0.972	0.872	0.869	1	0.5	0.648	0.693
NS-PB1	0.484	0.924	0.763	0.756	1	0.352	0.5	0.548
NS-PB2	0.432	0.911	0.728	0.720	1	0.307	0.452	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 39

Normalised values from the Wilcoxon test for PA H1N2 pairs.

p-value	PA-HA	PA-MP	PA-NA	PA-NP	PA-NS	PA-PA	PA-PB1	PA-PB2
PA-HA	0.5	0.951	0.812	0.946	0.857	1	0.928	0.956
PA-MP	0.049	0.5	0.245	0.511	0.273	1	0.506	0.611
PA-NA	0.188	0.755	0.5	0.754	0.549	1	0.738	0.814
PA-NP	0.054	0.489	0.246	0.5	0.272	1	0.496	0.600
PA-NS	0.143	0.727	0.451	0.728	0.5	1	0.711	0.794
PA-PA	0	0	0	0	0	0.5	0	0
PA-PB1	0.072	0.494	0.262	0.504	0.289	1	0.5	0.596
PA-PB2	0.044	0.389	0.186	0.400	0.206	1	0.404	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 40

Normalised values from the Wilcoxon test for PB1 H1N2 pairs.

p-value	PB1-HA	PB1-MP	PB1-NA	PB1-NP	PB1-NS	PB1-PA	PB1-PB1	PB1-PB2
PB1-HA	0.5	0.966	0.713	0.943	0.899	0.909	1	0.936
PB1-MP	0.034	0.5	0.115	0.444	0.285	0.375	1	0.460
PB1-NA	0.287	0.885	0.5	0.844	0.745	0.788	1	0.840
PB1-NP	0.057	0.556	0.156	0.5	0.344	0.430	1	0.514
PB1-NS	0.101	0.714	0.255	0.656	0.5	0.581	1	0.660
PB1-PA	0.091	0.625	0.212	0.570	0.419	0.5	1	0.579
PB1-PB1	0	0	0	0	0	0	0.5	0
PB1-PB2	0.064	0.540	0.160	0.486	0.340	0.421	1	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 41

Normalised values from the Wilcoxon test for PB2 H1N2 pairs.

p-value	PB2-HA	PB2-MP	PB2-NA	PB2-NP	PB2-NS	PB2-PA	PB2-PB1	PB2-PB2
PB2-HA	0.5	0.989	0.810	0.976	0.947	0.965	0.960	1
PB2-MP	0.011	0.5	0.084	0.374	0.233	0.377	0.365	1
PB2-NA	0.190	0.916	0.5	0.856	0.756	0.838	0.824	1
PB2-NP	0.024	0.626	0.144	0.5	0.345	0.497	0.480	1
PB2-NS	0.053	0.767	0.244	0.655	0.5	0.643	0.623	1
PB2-PA	0.035	0.623	0.162	0.503	0.357	0.5	0.485	1
PB2-PB1	0.040	0.635	0.176	0.520	0.377	0.515	0.5	1
PB2-PB2	0	0	0	0	0	0	0	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 42

Normalised values from the Wilcoxon test for HA H3N2 pairs.

p-value	HA-HA	HA-MP	HA-NA	HA-NP	HA-NS	HA-PA	HA-PB1	HA-PB2
HA-HA	0.5	0	0	0	0	0	0	0
HA-MP	1	0.5	0.383	0.003	0.057	0.003	0.002	0.001
HA-NA	1	0.617	0.5	0.017	0.126	0.013	0.010	0.007
HA-NP	1	0.997	0.983	0.5	0.872	0.425	0.394	0.322
HA-NS	1	0.943	0.874	0.128	0.5	0.105	0.085	0.064
HA-PA	1	0.997	0.987	0.575	0.895	0.5	0.475	0.397
HA-PB1	1	0.998	0.990	0.606	0.915	0.525	0.5	0.418
HA-PB2	1	0.999	0.993	0.678	0.936	0.603	0.582	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 43

Normalised values from the Wilcoxon test for NA H3N2 pairs.

p-value	NA-HA	NA-MP	NA-NA	NA-NP	NA-NS	NA-PA	NA-PB1	NA-PB2
NA-HA	0.5	0.860	1	0.211	0.667	0.211	0.179	0.140
NA-MP	0.140	0.5	1	0.019	0.242	0.026	0.015	0.012
NA-NA	0	0	0.5	0	0	0	0	0
NA-NP	0.789	0.981	1	0.5	0.911	0.476	0.446	0.365
NA-NS	0.333	0.758	1	0.089	0.5	0.097	0.071	0.054
NA-PA	0.789	0.974	1	0.524	0.903	0.5	0.470	0.391
NA-PB1	0.821	0.985	1	0.554	0.929	0.530	0.5	0.416
NA-PB2	0.860	0.988	1	0.635	0.946	0.609	0.584	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 44

Normalised values from the Wilcoxon test for NP H3N2 pairs.

p-value	NP-HA	NP-MP	NP-NA	NP-NP	NP-NS	NP-PA	NP-PB1	NP-PB2
NP-HA	0.5	0.999	0.935	1	0.999	0.949	0.874	0.933
NP-MP	0.001	0.5	0.087	1	0.697	0.146	0.036	0.104
NP-NA	0.065	0.913	0.5	1	0.964	0.582	0.349	0.517
NP-NP	0	0	0	0.5	0	0	0	0
NP-NS	0	0.303	0.036	1	0.5	0.071	0.015	0.047
NP-PA	0.051	0.854	0.418	1	0.929	0.5	0.273	0.435
NP-PB1	0.126	0.964	0.651	1	0.985	0.727	0.5	0.670
NP-PB2	0.067	0.896	0.483	1	0.953	0.565	0.330	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 45

Normalised values from the Wilcoxon test for NS H3N2 pairs.

p-value	NS-HA	NS-MP	NS-NA	NS-NP	NS-NS	NS-PA	NS-PB1	NS-PB2
NS-HA	0.5	1	0.955	0.984	1	0.963	0.928	0.951
NS-MP	0	0.5	0.030	0.076	1	0.038	0.014	0.036
NS-NA	0.045	0.970	0.5	0.659	1	0.531	0.366	0.483
NS-NP	0.016	0.924	0.341	0.5	1	0.369	0.225	0.328
NS-NS	0	0	0	0	0.5	0	0	0
NS-PA	0.037	0.962	0.469	0.631	1	0.5	0.334	0.452
NS-PB1	0.072	0.986	0.634	0.775	1	0.666	0.5	0.618
NS-PB2	0.049	0.964	0.517	0.672	1	0.548	0.382	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 46

Normalised values from the Wilcoxon test for PA H3N2 pairs.

p-value	PA-HA	PA-MP	PA-NA	PA-NP	PA-NS	PA-PA	PA-PB1	PA-PB2
PA-HA	0.5	0.997	0.936	0.958	0.998	1	0.990	0.974
PA-MP	0.003	0.5	0.150	0.205	0.666	1	0.405	0.287
PA-NA	0.064	0.850	0.5	0.586	0.912	1	0.779	0.668
PA-NP	0.042	0.795	0.414	0.5	0.877	1	0.712	0.588
PA-NS	0.002	0.334	0.088	0.123	0.5	1	0.263	0.179
PA-PA	0	0	0	0	0	0.5	0	0
PA-PB1	0.010	0.595	0.221	0.288	0.737	1	0.5	0.374
PA-PB2	0.026	0.713	0.332	0.412	0.821	1	0.626	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 47

Normalised values from the Wilcoxon test for PB1 H3N2 pairs.

p-value	PB1-HA	PB1-MP	PB1-NA	PB1-NP	PB1-NS	PB1-PA	PB1-PB1	PB1-PB2
PB1-HA	0.5	0.998	0.946	0.922	0.998	0.994	1	0.914
PB1-MP	0.002	0.5	0.116	0.067	0.499	0.405	1	0.082
PB1-NA	0.054	0.884	0.5	0.394	0.876	0.815	1	0.417
PB1-NP	0.078	0.933	0.606	0.5	0.929	0.880	1	0.511
PB1-NS	0.002	0.501	0.124	0.071	0.5	0.410	1	0.087
PB1-PA	0.006	0.595	0.185	0.120	0.590	0.5	1	0.136
PB1-PB1	0	0	0	0	0	0	0.5	0
PB1-PB2	0.086	0.918	0.583	0.489	0.913	0.864	1	0.5

The segment pair comparisons in the table are to be read as a row against a column.

Table B 48

Normalised values from the Wilcoxon test for PB2 H3N2 pairs.

p-value	PB2-HA	PB2-MP	PB2-NA	PB2-NP	PB2-NS	PB2-PA	PB2-PB1	PB2-PB2
PB2-HA	0.5	0.997	0.933	0.968	0.999	0.986	0.935	1
PB2-MP	0.003	0.5	0.125	0.211	0.696	0.354	0.120	1
PB2-NA	0.067	0.875	0.5	0.638	0.941	0.763	0.500	1
PB2-NP	0.032	0.789	0.362	0.5	0.893	0.648	0.362	1
PB2-NS	0.001	0.304	0.059	0.107	0.5	0.204	0.057	1
PB2-PA	0.014	0.646	0.237	0.352	0.796	0.5	0.230	1
PB2-PB1	0.065	0.880	0.500	0.638	0.943	0.770	0.5	1
PB2-PB2	0	0	0	0	0	0	0	0.5

The segment pair comparisons in the table are to be read as a row against a column.

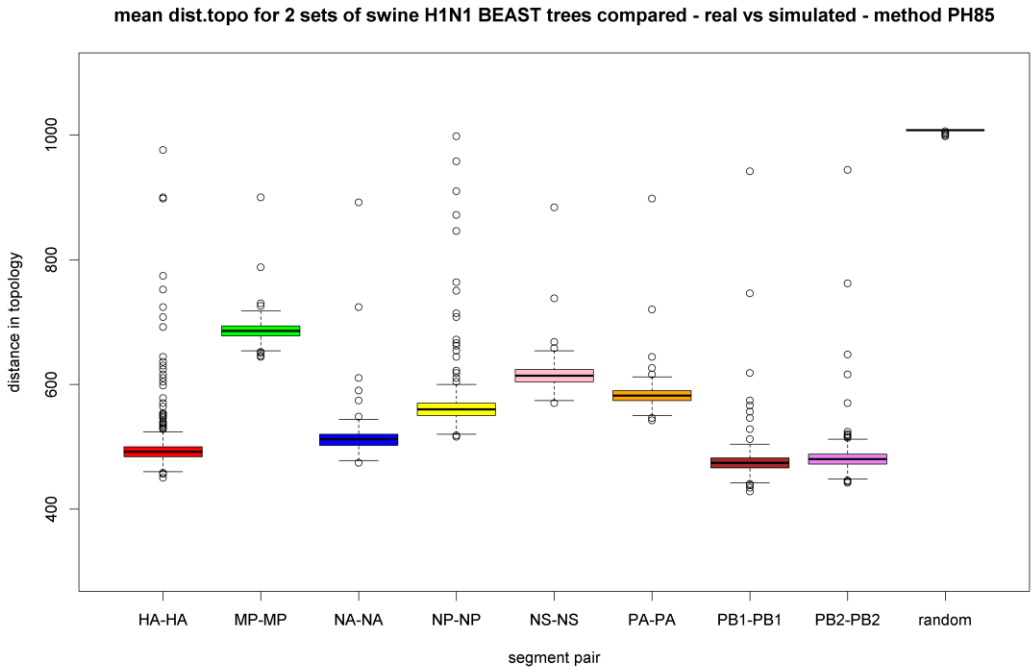

Figure B 70

Distance in topology for segment pairs of real and simulated sequences for H1N1.

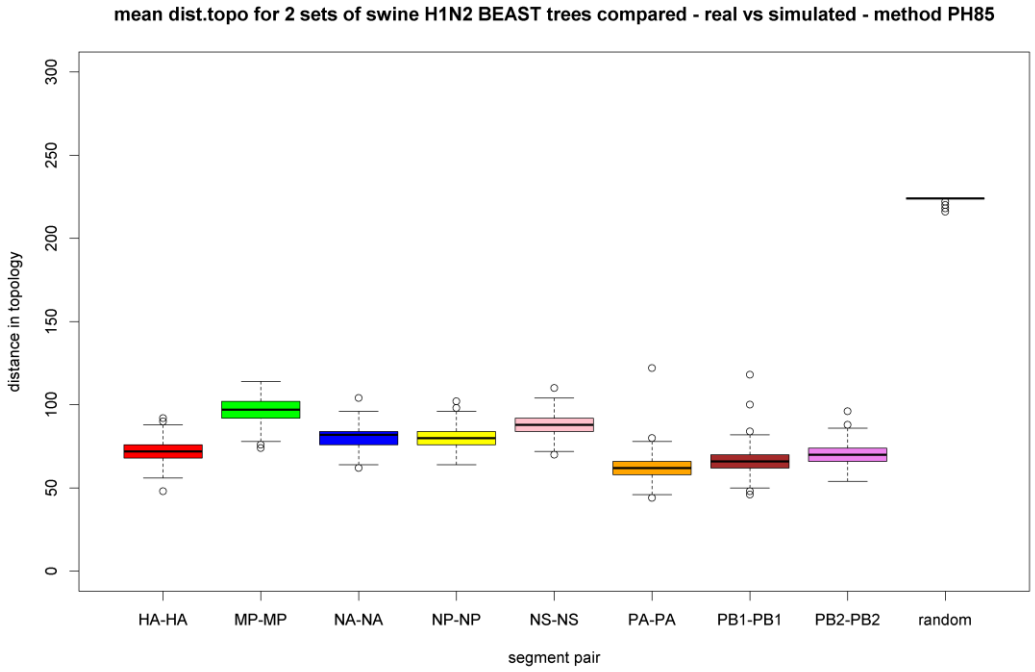

Figure B 71

Distance in topology for segment pairs of real and simulated sequences for H1N2.

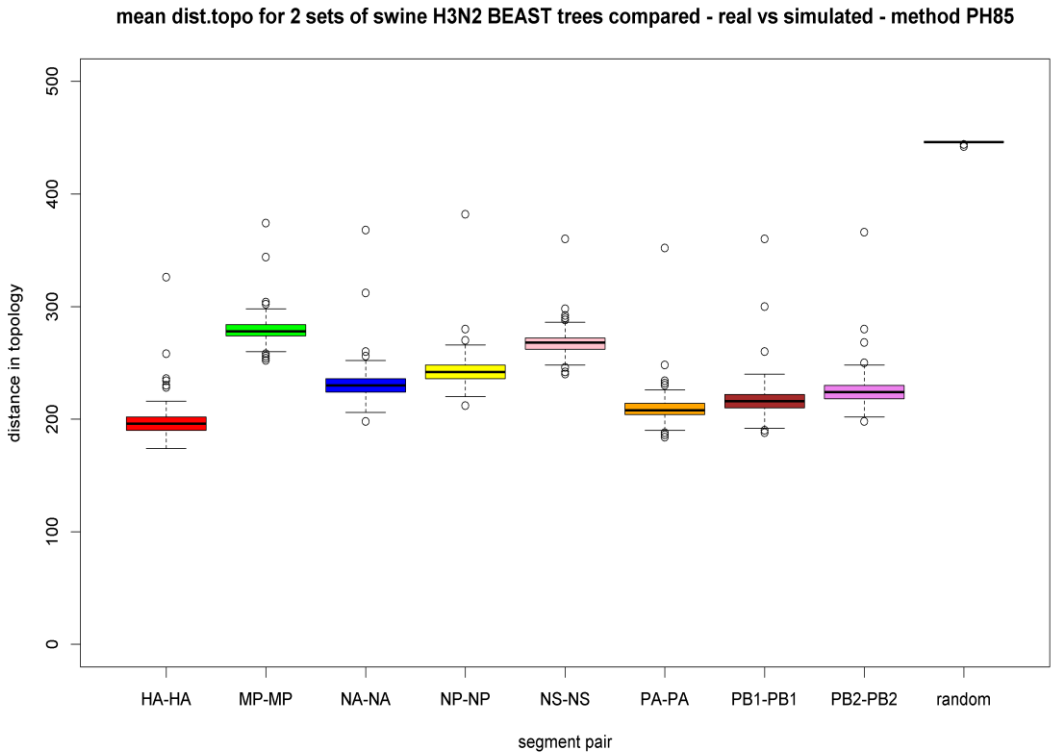

Figure B 72

Distance in topology for segment pairs of real and simulated sequences for H3N2.

8.3 Appendix C

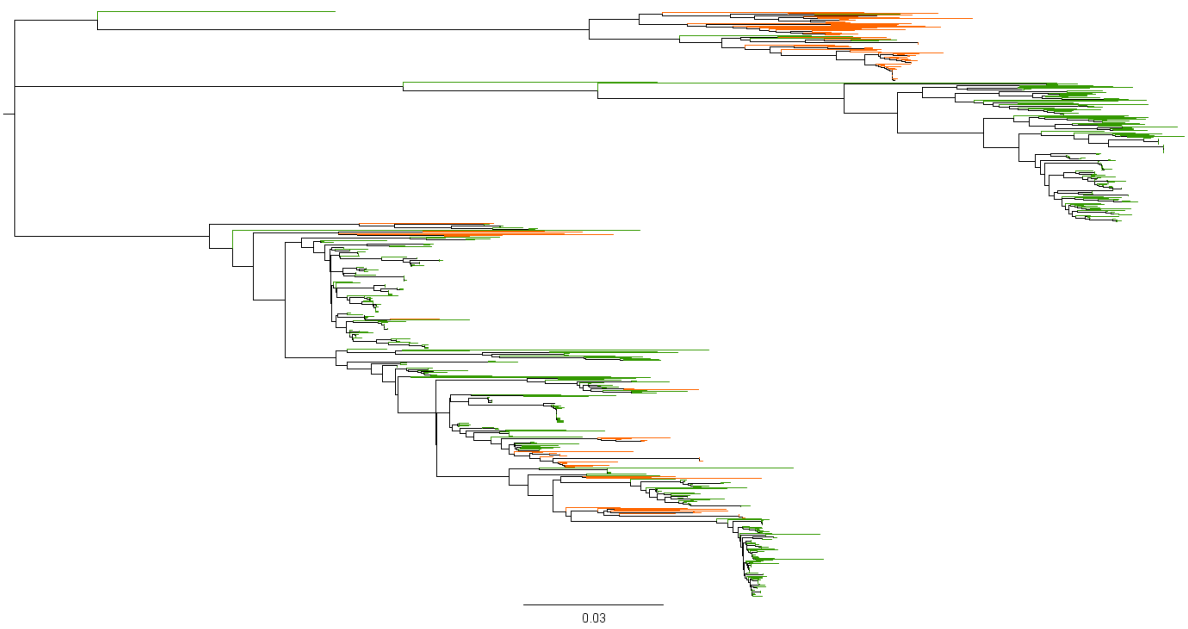

Figure C 1

NJ tree for H1 (H1N1 and H1N2 subtypes sequences). Green colour is for H1N1 subtype, orange colour is for H1N2 subtype.

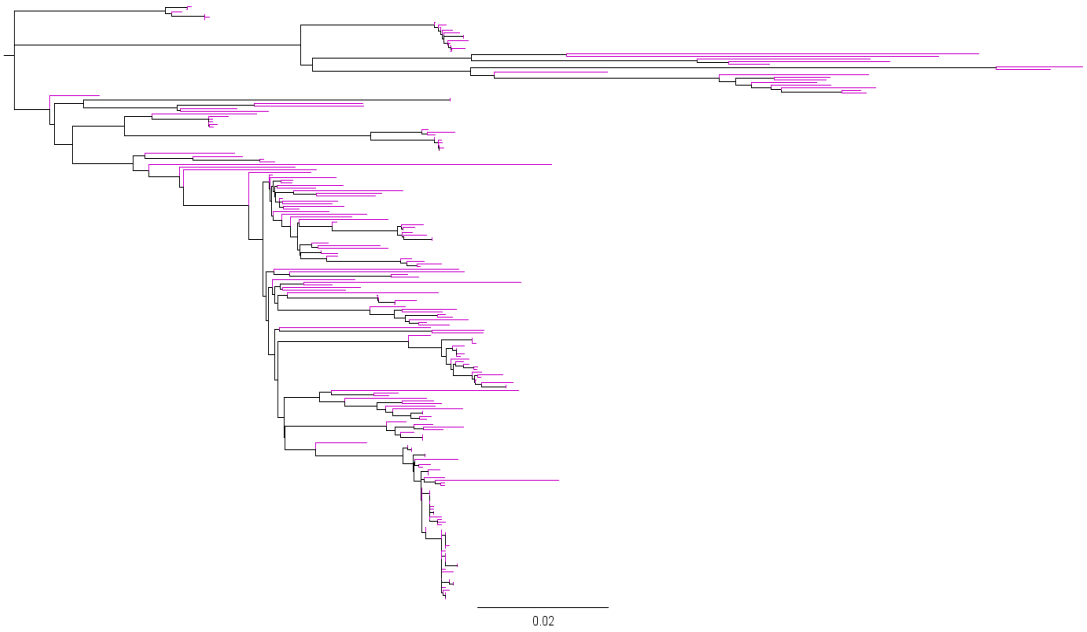

Figure C 2

NJ tree for H3 (H3N2 subtypes sequences). Violet colour is for H3N2 subtype.

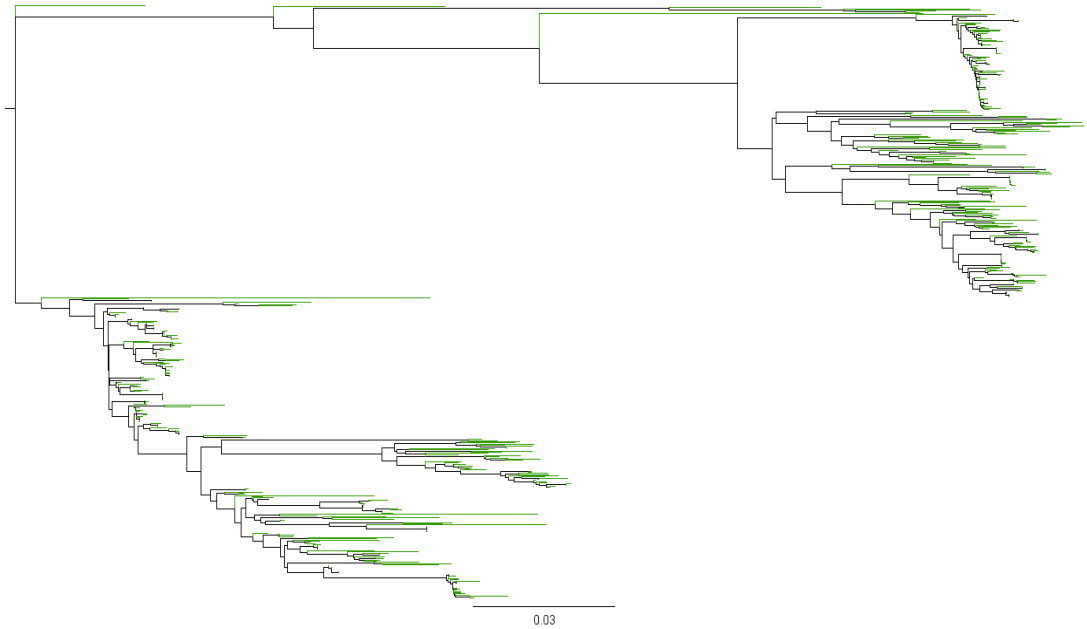

Figure C 3

NJ tree for N1 (H1N1 subtypes sequences). Green colour is for H1N1 subtype.

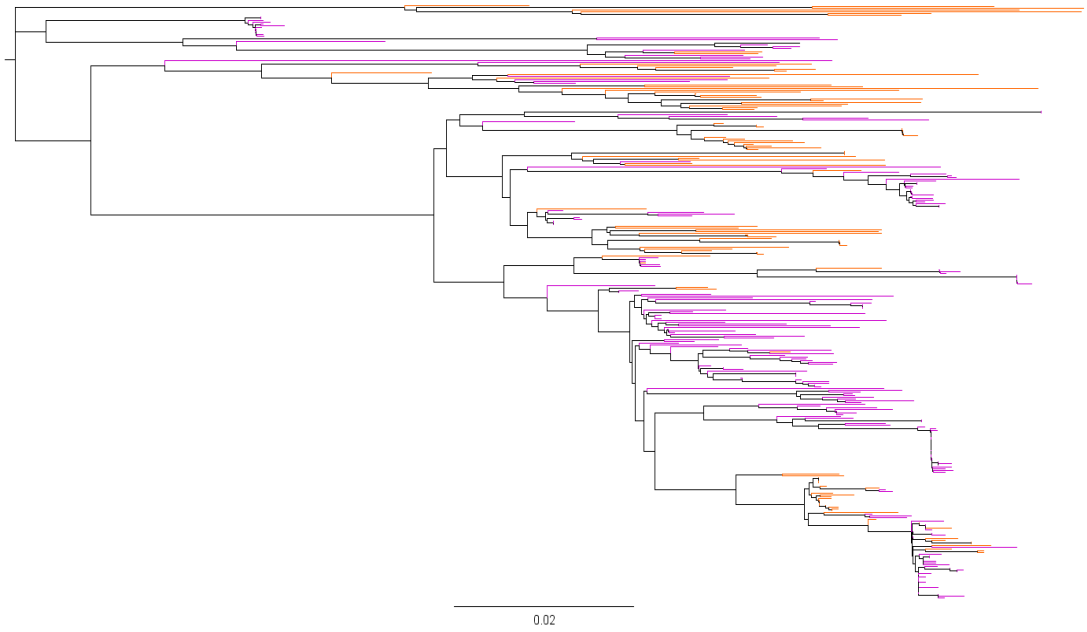

Figure C 4

NJ tree for N2 (H1N2 and H3N2 subtypes' sequences). Orange colour is for H1N2 subtype, violet colour is for H3N2 subtype.